INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de **DOCTEUR DE L'INPG**
en **Sciences Cognitives**

Spécialité : **Informatique**

préparée à l'Institut de la Communication Parlée **ICP** - INPG
et à l'Institut Dalle Molle d'Intelligence Artificielle Perceptive **IDIAP** - EPFL
dans le cadre de l'École Doctorale
Ingénierie pour le vivant : Santé, Cognition, Environnement

présentée et soutenue publiquement
par
Hervé Glotin
le 13 juin 2001 à Grenoble

Élaboration et comparaison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole: incorporation des indices de voisement et de localisation

JURY

M. Jean-Luc Schwartz, Président
M. Jean-Paul Haton, Rapporteur
M. Henri Meloni, Rapporteur
M. James Crowley, Examinateur
M. Renato De Mori, Examinateur
M. Frédéric Berthommier, Codirecteur de thèse
M. Hervé Bourlard, Codirecteur de thèse

Elaboration et comparaison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation des indices de voisement et de localisation

Cette thèse effectuée à l'ICP et à l'IDIAP, dans le champ de la communication homme-machine et des projets EU. RESPITE & SPHEAR, contribue à augmenter la robustesse de reconnaisseur automatique de la parole dans le cadre original de l'analyse de scènes auditives. Deux voies sont traitées simultanément : (1) l'extraction d'indices fiables du signal et (2) la fusion de données dans le cadre multi-flux. (1) est fondée sur des mesures temps-fréquences de corrélations, relatives au taux de voisement ou aux localisations de sources. Nous montrons comment l'indice de voisement renforce le prétraitement de référence « Jrasta ». (2) est proposée via un modèle "combinaison complète" qui intègre par combinaisons de sous-bandes du spectre les redondances spectrales de la parole. Ce modèle est approximé avec une hypothèse faible d'indépendance des observations des sous-flux du spectre. La robustesse d'un système de reconnaissance hybride ANN/HMM, de chiffres téléphonés (NB95), est alors renforcée dans le cas de paroles simultanées (enregistrements stéréo), ou contre des bruits non stationnaires focalisés. Nous validons dans le cas de bruit de cafétéria, l'apport de l'indice de voisement pour la reconnaissance audiovisuelle grand vocabulaire (base Via Voice-IBM,MMG asynchrone). Nous proposons de plus un modèle de « Prédiction des Biais des Posteriors » guidé par les indices dont les premiers tests sont prometteurs. Nous comparons finalement ces différentes architectures, et en proposons une, dite « proactive », qui permet l'intégration d'informations complémentaires.
Mots-clés : reconnaissance automatique de la parole, IHM, robustesse au bruit, cocktail party, multi-bande, multi-flux, fusion de données, analyse de scène auditive, harmonicité, voisement, audiovisuel, Via Voice, IBM, MMG, localisation, HMM, modèle hybride, calcul bayesien, perceptron, ANN, réseau de neurones, Jrasta.

Comparative study and elaboration of robust adaptive multistream automatic speech recognition using voicing and localization cues

This thesis, taking part in European projects RESPITE and SPHEAR,between ICP-IDIAP, shows various means to reinforce automatic speech recognition (ASR), based on (1) Computational Auditory Scene Analysis and (2) multistream paradigm. In(1)we propose different reliability factors calculated from time-frequency correlations, related to voicing level or a localization cue in case of stereo data. We show how voicing cue reinforces the state of the art preprocessing « Jrasta ».In(2)we propose a multistream ASR model, called « Full Combination » (FC) that exploits spectral redundancy, while making minimum assumptions about noise type by considering every combination of data subbands as an independent data stream. Using our reliability factors and hybrid ANN/HMM ASR, we demonstrate under Numbers95 telephonic free digits data base, that FC is robust to non-stationary noise and to cocktail party effect. Furthermore, we develop a promising fusion model for the voicing cue in a multistream audiovisual large vocabulary ASR and efficient under cafeteria noise (tested on Via Voice IBM database, asynchronous GMM). After having analyzed FC errors we propose the Posteriors Bias Prediction model which gives an optimal fusion model for signal and estimates' reliabilities. First tests are promising. Then we compare these different ASR architectures, and we propose a « proactive » one, which allows integration of complementary information for robust ASR.

Key-words: automatic speech recognition, robustness, noise, cocktail party, multiband, multistream, data fusion, auditory scene analysis, harmonicity, voicing, audiovisual,Via Voice, IBM, localization, HMM, bayes, perceptron, ANN, GMM, neural network, Jrasta.

Spécialité : Informatique-Sciences Cognitives
Intitulés et adresses des laboratoires : Inst. de la Communication Parlée (ICP)-INPG-46 av. Viallet-38031 Grenoble Cedex & Inst. d'Intelligence Artificielle Perceptive (IDIAP)-Simplon 4-1920 Martigny-Suisse

"La réalité se présente à nous sous la forme de phénomènes, de formes, dont nous décelons la présence par leurs discontinuités qualitatives"

Modèles mathématiques de la morphogénèse

René Thom, 1974.

"Vous ne pouvez pas énoncer une idée nouvelle autrement qu'avec des mots anciens, ceux que vous avez à votre disposition. Il va donc falloir un temps de travail pour comprendre ce que vous venez de faire. C'est ce que Bachelard appelait la 'refonte épistémologique'."

La science contemporaine est-elle moderne ?

Jean-Marc Lévy-Leblond, 1999.

Remerciements

Un jour je traverse la brèche de Saint-Maurice et je rentre dans le Valais et ses 4000 majestueux, La Batiaz, les mayens, le Catogne, voici Martigny. Un autre jour je redescends le Rhône, je passe le long du château de Chillon qui se découpe sur la mer des montagnes, le héron cendré du lac du Bourget s'envole au bruit du train, et la bienveillante Chartreuse m'accompagne jusqu'à Grenoble. Ce chemin, emprunté pendant 3 ans et demi, presque 200 fois, ne m'a jamais paru identique, tant les questions qui édifiaient le pont IDIAP-ICP me passionnaient, tant cette fusion culturelle me motivait. Dans cette grande aventure scientifique et humaine qu'est une thèse, beaucoup d'impasses sont à éviter. Comme dans une randonnée alpine, certaines époques sont rudes, mais elles valent la peine d'être traversées.

Je remercie Frédéric Berthommier, chargé de recherche à l'Institut de la Communication Parlée (ICP) de l'INP Grenoble, qui a su éveiller ma curiosité avec son approche pluridisciplinaire. Grâce à sa ténacité, nous avons achevé un travail de recherche innovant.

Hervé Bourlard, professeur à l'EPF Lausanne, co-directeur de ma thèse, et directeur de l'Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP-Suisse), reçoit ma gratitude pour son encadrement de valeur de par son expertise en reconnaissance automatique de la parole. Depuis notre première discussion sur la reconnaissance multi-bande en Février 1997, par ses idées toujours claires, Hervé Bourlard m'a permis d'évoluer dans le monde de la reconnaissance automatique de la parole à la pointe des recherches actuelles.

L'IDIAP qui a financé une grande partie de ma thèse, m'a octroyé une grande liberté et confiance, en me permettant de travailler à temps partiel à l'ICP. Sans l'enthousiasme de son directeur, ce projet n'aurait pas abouti, car les dialogues électroniques ne remplacent pas le dialogue humain.
Cette recherche s'intègre et a été financée en grande partie par les projets européens COST, SPHEAR et RESPITE. Je remercie aussi la région Rhône-Alpes pour son soutien "Eurodoc".

Je remercie Pr. Frédérick Jelinek, directeur du *Center for Language and Speech*

Processing (CSLP) à Johns Hopkins University, Baltimore-USA, pour m'avoir invité invitation au *Whiting School of Engineering* durant l'été 2000.

Jean-Paul Haton, professeur au LORIA/INRIA de Nancy et titulaire de la chaire de modélisation informatique des processus perceptifs et cognitifs à l'*Institut Universitaire de France*, reçoit ici toute ma gratitude pour avoir soutenu notre thématique de recherche et nous avoir fait l'honneur d'être rapporteur de ma thèse. Son humanisme est une qualité qui enrichit ses communications scientifiques de grande valeur.

Je remercie le professeur Henri Méloni, président de l'*Université d'Avignon*, pour avoir chaleureusement accepté d'être rapporteur de ma thèse et en avoir tiré une fine analyse.

Pr. Renato De Mori, professeur et directeur du *Laboratoire d'Informatique d'Avignon* (LIA) est remercié pour avoir examiné ma thèse et avoir soulevé de très pertinentes réflexions lors de ma soutenance.

Je remercie Pr James Crowley de l'INRIA Alpes et professeur à l'INPG, pour son enthousiasme scientifique et l'examen qu'il a porté à ma thèse.

Merci à Jean-Luc Schwartz, HDR et directeur de notre équipe perception à l'ICP, d'avoir présidé la soutenance de ma thèse, et d'avoir toujours su nous faire bénéficier de sa puissance et clarté d'esprit, et de son humanité dans la gestion de notre équipe.

Je remercie vivement Pr. Chalapathy Neti et Pr. Gerasimos Potamianos de l'équipe Human Language Technologies d'IBM-Watson Research Team, USA, pour leur accueil chaleureux lors du workshop AVSR de l'été 2000 au CSLP, avec qui j'ai pu enrichir et valider certains aspects de notre recherche dans le cadre de la reconnaissance audiovisuelle grand vocabulaire sur le projet Via-Voice.

Merci à Pierre Escudier, directeur de l'ICP et de l'École doctorale de sciences cognitives de l'INPG, pour sa direction si efficace et également humaine.

Je remercie aussi mes collègues de travail, de l'IDIAP et de l'ICP. Merci à toi Juergen Luettin, pour m'avoir aidé avec ta grande maîtrise des systèmes HMM asynchrones. Merci Christopher Kermovan, ami du train, mais surtout des veillées IDIAPIENNES quotidiennes. Tu as été d'une aide précieuse pour mes travaux. Merci Emmanuel Tessier, avec qui j'ai pu travailler à l'ICP plus précisément le traitement de signal sur la base ST-Numbers. Merci Andrew Morris, Astrid Hagen, proches compagnons de travail à l'IDIAP.

Je remercie aussi mes collègues des projets européens SPHEAR et RESPITE, Stéphane Dupont, et Christophe Ris du TCTS de Mons, Belgique, pour leur aide et

leur excellent outil STRUT, ainsi que toute l'équipe du laboratoire de Sheffield, Dr. Barker, Pr. Martin Cooke et Pr. Phil Green, ainsi que P. Lockwood de Matra.

Je tiens également à remercier mes collègues de projet de l'*International Computer Science Institute* (ICSI-Berkeley) pour leur aide en reconnaissance hybride HMM/ANN. Tout particulièrement Nikki Mirghafori et Dan Ellis.

Un travail très intéressant de séparation aveugle de source sur ST-Numbers a été accompli avec Pr. Seugjin Choi de l'Université de Corée, merci pour son efficacité.

Merci à Chafic Mokbel qui a pu suivre et commenter nos travaux durant la période la plus effervescente.

Merci à toi Rafael Laboissière pour m'avoir fait plongé avec passion dans le domaine de la parole lors de mon DEA à l'ICP

Merci à Gordon Ramsay pour ces discussions si claires et ton archivage bibliographique astronomique mais non moins efficace !

Merci aux *System Managers* que j'admire pour leur compétences et leur patience, Olivier Bornet et son complice Frank Formaz de l'IDIAP, et toute ma gratitude à Christian Bulfone et Nino Mendez de l'ICP qui me supportent depuis mon DEA avec mes calculs gargantuesques d'AGORA ... Merci à vous sans qui ce travail serait mille fois plus harassant.

Pr. Gilbert Maître, merci à toi, Valaisan !, et à mes compagnons de travail et de montagne, Perry, Hans, Eddy, Doms, Katrin, Juergen, Mikko, Jean-Luc... pour avoir partagé la même trace.

Merci enfin à toi Pascale pour ta patience et ton aide, et à ma famille pour son soutien continuel. Je leur dédie ce travail.

Prologue

Cette thèse est le fruit d'une continuelle collaboration entre l'IDIAP et l'ICP, durant des séjours réguliers dans chacun des deux laboratoires.

Le sujet de cette thèse est ciblé sur la mise en place de détecteurs de fiabilité du signal et sur leur utilisation pour la fusion de reconnaissances phonétiques distribuées, par l'usage de modèles originaux de type multi-flux de reconnaissance automatique de la parole.

L'ICP regroupe les compétences en traitement du signal de parole, en sciences cognitives, et en système de perception.
L'IDIAP est expert en système de reconnaissance et en machine learning.

Nous espérons que le lecteur trouvera dans ce mémoire un intérêt pour cette fusion de deux écoles de pensée qui furent à tort trop longtemps éloignées.

Dans ce cadre, de fréquents échanges et études comparatives ont été réalisées durant les projets européens ESPRIT :
TMR SPHEAR[1] SPeech, HEaring and Recognition
et
LTR RESPITE[2] REcognition of Speech by Partial Information TEchniques

Les laboratoires partenaires laboratoires étaient Sheffield, Mons, Daimler, ICSI-Berkley et Babel, Keele, Matra. Notre thèse a profité des collaborations avec l'équipe de reconnaissance vocale d'IBM durant le workshop annuel 2000 d'ICLSP-Baltimore-USA. Nous avons également travaillé avec l'Université de Corée sur la séparation aveugle de sources.

Les bases de parole utilisées ont été partagées avec nos partenaires afin de bénéficier d'études comparatives. Ce sont des bases multilocuteurs de parole libre, des nombres téléphonés au grand vocabulaire de 15000 mots (IBM Via Voice), mono, stéréo ou augmentées de la modalité visuelle.

[1] http ://www.dcs.shef.ac.uk/ pdg/sphear/sphear.htm
[2] http ://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/

Etudes transversales : séparation aveugle de paroles simultanées

En marge de nos travaux de thèse nous avons mis en place et réalisé la double reconnaissance automatique de paroles simultanées sur les signaux doubles voies séparés par une extension récursive de l'algorithme Hérault-Jutten (Jutten & Hérault 1991). Cette méthode donne les meilleurs taux de reconnaissance connus sur cette base et a abouti à la publication Neurocomputing (Choi, Hong, Glotin & Berthommier Sept 2001), mais elle n'est pas traitée dans ce mémoire car en marge de notre étude.

Table des matières

Table des figures

Liste des tableaux

Notations

- $x_n = (x_{n1}, x_{n2}, \ldots, x_{nd})^T$: vecteur (acoustique) à l'instant n
- d : dimension of vecteurs acoustiques
- ω_k ou q_k : une classe (statistique)
- K : nombre de classes statistiques, états HMM, densités de probabilité ou sorties de réseau de neurones
- $X = \{x_1, \ldots, x_n, \ldots, x_N\}$: séquence de vecteurs acoustiques de longueur N
- f_j : l'événement statistique 'Seul le flux j est pris pour reconnaissance'
- X_j : '$X \cap f_j$'
- t_k : l'événement statistique 'le phonème à reconnaître est le phonème k'
- q_k : l'événement statistique 'le phonème reconnu est le phonème k'
- n : vecteur d'un bruit digitalisé
- M : Modèles de Markov (discrets ou cachés)
- $M = \{q_1, \ldots, q_\ell, \ldots, q_L\}$: modèle de Markov constitué de l'ensemble des états L
- L : nombre d'états d'un modèle HMM
- μ_k : vecteur moyen associé à la classe ω_k ou q_k dans la cas de distributions gaussiennes.
- q_k^n : événement $q^n = q_k$ (états q_k visité à l'instant n)
- q^n : état HMM observé à l'instant n
- $P(x_n|q_k)$: vraisemblance locale (probabilités d'émission)
- $P(q_k|x_n)$: probabilité a posteriori (probabilité conditionnelle) de la classe q_k étant donné l'observation x_n
- $P(q_k)$: probabilité a priori de la classe q_k
- $P(X|M)$: vraisemblance de la séquence X étant donné le modèle M
- $g(x_n) = \{g_1(x_n), \ldots, g_k(x_n), \ldots, g_K(x_n)\}^T$: vecteur de sortie d'un réseau de neurones étant donné x_n à son entrée
- $g_k(x_n)$: valeur observée à la k-ième sortie d'un réseau de neurones associé à la classe ω_k ou à un état HMM q_k

- E : fonction d'erreur (typiquement critère de moindres carrés) minimisée lors de l'entraînement de fonctions discriminantes ou d'un réseau de neurones
- $d(x_n) = \{d_1(x_n), \ldots d_k(x_n), \ldots, d_K(x_n)\}$: vecteur de sortie cible lors de l'entraînement de fonctions discriminantes ou d'un réseau de neurones
- $H(\cdot)$: entropie d'une distribution de probabilité
- ET : énergie totale de type parole
- EH : énergie harmonique de type parole
- EN : énergie de bruit propre à la parole
- EN' : énergie du bruit de l'environnement
- $F0$: fréquence du pitch
- $R_{xx} = R_x = AC(x)$ autocorrélogramme de x qui est un vecteur acoustique
- $R0$ = autocorrélation de x en délai nul soit l'énergie du signal
- $R1$: pic maximum de plus faible délai sur un corrélogramme
- Rp : pic maximum sur un corrélogramme dont le délai est compris dans le segment du pitch F0=[90 350]Hz
- $R = Rp/R0$: indice d'harmonicité (dit aussi de voisement)
- $R_{xy} = IC(xy)$ = intercorrélation de x et y (x et y sont des vecteurs acoustiques de même dimension)

Acronymes

- ASA : Analyse de Scène Auditive
- CASA : modèle computationnel d'ASA
- RAP (ou ASR) : Reconnaissance Automatique de la Parole
- RAV : Reconnaissance Automatique de la parole audio-Visuelle
- RSB (ou SNR) : Rapport signal sur bruit, mesuré en décibel (dB)
- WER : *Word Error Rate*, taux en % de l'erreur de reconnaissance des mots (formule en annexe)
- WEP : *Word Entrance Penalty*, Pénalité d'insertion de mot lors du décodage.
- IBB : Identification de la Bande Bruitée
- FC : '*Full Combination ASR model*' ou modèle 'combinaison complète'
- AFC : approximation du modèle FC
- PBP : modèle de Prédiction des Biais des Posteriors
- HMM : modèle de Markov caché (*Hidden Markov Model*)
- MAP : Maximum A Posteriori
- ML (ou MV) : Maximum de Vraisemblance (*Maximum Likelihood*)
- EM : Expectation-Maximisation
- DP : Programmation Dynamique (Dynamic Programming)
- PLP : Perceptual Linear Prediction
- EBP : algorithme de rétro-propagation de l'erreur (Error back-Propagation)
- KL : Kullback-Leibler divergence
- dKL : distance de Kullback Leibler
- CODES des BRUITS :
  FACT : bruit d'usine de la base Noisex
  CAR : bruit de voiture Daimler Benz, 120 km/h fenêtre fermée
  GWN : bruit gaussien blanc
  Bandx : bruit de bande 300 hz de large centré sur la sous-bande numéro x.
  Non-Stat : bruit en pavés temps-fréquence de 125 ms de Bandx, pris régulièrement de x = 1 2 3 4 4 3 2 1

Résumé

Les performances des systèmes actuels de reconnaissance automatique de la parole (RAP) sont mauvaises en milieu bruité. Pour les améliorer l'approche de RAP multi-flux propose de tirer parti de la redondance spectrale du signal de parole. Après un développement de cette technique de reconnaissance partielle, nous montrons comment l'Analyse de Scène Auditive Computationnelle (CASA), modélisant notre capacité à structurer notre environnement sonore, contribue à augmenter la robustesse de la RAP multi-flux. En étroite et constante collaboration entre deux écoles, l'une du traitement du signal et de la physiologie de la perception (ICP-Grenoble), et l'autre de la RAP multi-flux (IDIAP-Suisse), ce travail de thèse présente une étude complète et comparée du marquage du plan temps-fréquence et de son intégration dans un système adapté de RAP multi-flux robuste.

Tout d'abord, nous développons et formalisons un modèle probabiliste de fusion des flux dans le cadre de la reconnaissance multi-bandes HMM/ANN. Nous introduisons une variable latente qui indique à chaque trame la combinaison de sous-bandes la plus adéquate à la reconnaissance. Nous calculons alors la probabilité a posteriori de chaque phonème en intégrant sur toutes les positions possibles du meilleur estimateur de probabilité a posteriori. Nous montrons que les experts de combinaison de sous-bandes peuvent être approximés de façon fiable à partir des experts de chaque sous-bande initiale.

Puis nous développons les indices de marquage, associant à chaque pavé temps-fréquence une probabilité qui mesure sa fiabilité vis-à-vis du signal cible. L'estimation de cette probabilité est basée sur la mesure d'indices primitifs : (1) l'indice d'harmonicité mesuré sur l'autocorrélation des signaux monophoniques, ou (2) la différence interaurale de temps pour les signaux stéréophoniques. Ces deux attributs ont été originalement associés à la variable de référence adéquate de type A : rapport signal sur bruit ou B : niveau relatif des sources, et intégrés dynamiquement dans notre modèle "Full Combination". Nous comparons différents modèles de fusion de ces indices avant reconnaissance par rehaussement du spectre (gain de 4 dB par rapport au Jrasta seul) ou au niveau des probabilités générées par les experts phonétiques.

Les tests sont effectués sur des bases multilocuteur téléphoniques de référence, chiffres prononcés continûment monophoniques comme Numbers93 et Numbers95,

ou stéréophonique comme STNumbers95, et sur la base Via Voice multilocuteurs, grand vocabulaire 15000 mots augmentée de la modalité visuelle. Nous avons comparés les performances des modèles avec les techniques usuelles (Jrasta, soustraction spectrale ou séparation aveugle). Par rapport aux techniques usuelles, sur la base Numbers95, dans le cas A la robustesse face à un bruit coloré non-stationnaire peut atteindre en moyenne sur différent SNR jusqu'à 47% de gain relatif de reconnaissance de mot. Dans le cas B, le gain relatif est de 32% dans un contexte de type "cocktail party". Sur la base Via Voice augmentée de la modalité visuelle, le gain par fusion audiovisuelle atteind 57 % par rapport à la reconnaissance audio seule sur une interférence de type bruit de parole, et nous donnons une stratégie pour améliorer ce modèle en gérant la fusion par notre indice de voisement.

Cependant le modèle FC reste peut performant sur des bruits large bande. Nous analysons ses erreurs et présentons en détail un nouveau modèle de Prédiction des Biais des Posteriors (PBP) pour traiter le cas des bruits larges bandes. Ce modèle intègre les fiabilités contextuelles au niveau SNR des estimations de chaque expert, pour chaque phonème. Les résultats préliminaires sont donnés et nous discutons du potentiel du modèle PBP.

En perspective nous proposons l'architecture d'un reconnaisseur réactif, conçu comme un filtre adaptatif qui permettrait l'apport d'information de haut niveau spécifique à la parole en complément des indices traités dans ce mémoire.

Stratégie de recherche

Nous mettons en place des estimateurs temps-fréquence originaux de fiabilité du signal qui sont efficaces, et ce même en condition cocktail party. Leur pertinence comme indice de fiabilité du signal parole cible a été validée sur du signal monophonique ou stéréophonique, ainsi que dans le cadre de fusion multi-modale. Nous développons alors un nouveau modèle de fusion d'experts phonétiques compatibles, le modèle "Full Combination".

Nous montrons que le FC est performant sur une tâche de reconnaissance de deux sources simultanées de parole. Dans ce cas, la cible de la voie la plus énergétique est toujours plus ou moins atteinte. En paroles concurrentes, il n'y a donc pas d'état phonétique attracteur déviant le décodage des phonèmes cibles puisque les classes rivales sont cibles.
Cependant, les résultats sur des bruits large bande sur une seule voie sont un échec relatif, ce qui est corrélé aux travaux antérieurs.

Nous montrons alors que le modèle FC repose sur l'hypothèse que le détecteur phonétique est idéal (sensibilité=spécificité=1), sur tous les flux, tous les phonèmes,

ce qui est irréaliste. Ainsi nous proposons et développons un autre modèle de correction des biais de posteriors que nous couplons avec l'indice robuste de fiabilité du signal basé sur l'harmonicité du signal. Nous montrons que la complexité du modèle reste faible grâce aux fonctions combinatoires des estimateurs élémentaires et de l'indice de voisement.

Notre recherche de fusion optimale des estimations phonétiques et d'un indice primitif du signal débouche donc sur des fonctions originales de fiabilité du signal et sur un modèle au fort potentiel, intégrant de façon probabiliste les estimées phonétiques et la qualité de transmission des phonèmes dans les différents flux.

Chapitre 1

Introduction

1.1 Parole et audition

La parole est un véhicule stable de l'information en dépit de la très grande variabilité, en qualité et en production de cet assemblage de phonèmes. Cette propriété est sans doute due à l'extrême sophistication du récepteur et du décodeur de la parole c'est à dire du système auditif associé au cerveau. En effet ce dernier montre une étonnante robustesse aux interférences, tout à fait déroutante pour les ingénieurs qui tentent d'en approcher les performances en Reconnaissance Automatique de la Parole (RAP).

Notre thèse contribue à l'élaboration de tels modèles en les rendant plus robuste, en s'inspirant des propriétés du système auditif.
Nous démontrons tout d'abord quelques évidences de cette boucle perceptuo-motrice puis nous présentons un cas d'école de mécanisme de perception simple engendrant des comportements fort complexes et robustes aux interférences.

Comme tout système de communication, la transmission orale de l'information relève non seulement des mécanismes d'émission du son mais surtout des mécanismes de réception. Même si le locuteur parle la même langue que son auditeur, les caractéristiques spectrales de ses phonèmes, son timbre de voix, son intonation, sa vitesse d'élocution, son volume, et les nombreuses altérations subies dans le milieu de propagation, sont autant de variables qui "ouvrent" la boucle perceptuo-motrice parole. La seule solution pour une communication efficace réside donc en une très grande souplesse et adaptation du récepteur. La boucle perceptuo-motrice propre à l'ouïe et à l'appareil phonatoire est l'une des plus dynamiques de celles mises en jeu entre l'homme et son environnement, entre le faire et le percevoir des signes de l'intelligence. Cette boucle est très singulière sur plusieurs points dont les majeurs sont :

- Le contenu phonétique produit est intimement lié au système auditif. Les interactions mutuelles entre les appareils de production et de perception ont fait évoluer

le support phonétique des langues du monde vers quelques grands ensembles universels. Il est possible de simuler cette dynamique de boucle perceptuo-motrice en modélisant les contraintes sur les deux espaces concernés : acoustique et articulatoire (Glotin 1995, Glotin & Laboissière 1996), ou simplement via des contraintes limitées à l'espace acoustique seul (Berrah, Glotin, Laboissière, Bessière & Boë 1996).

- Une seconde singularité provient de la nature physique de l'information que la parole transporte : la parole est un signal "spatio-temporel" au contraire de l'écriture qui est figée dans le temps. Il faut donc capter et décoder ce message dans l'instant, avant qu'il ne disparaisse. Ceci amène à penser que le système perceptif pourrait peut être anticiper la scène qu'il perçoit afin d'avoir un temps d'avance pour un traitement optimal du signal. Nous reviendrons sur cette idée dans la perspective de la thèse.

Ceci démontre une grande capacité du système perceptif : adaptabilité, robustesse, rapidité. En fait les mécanismes de perception possédant de telles propriétés sont essentiellement basés sur des mécanismes simples mais nombreux et variés. Nous étayons cette réflexion dans la section suivante.

1.2 Des adaptions complexes émergeantes de perceptions élémentaires

Dans l'évolution du vivant, les capacités sensorielles complexes ont précédé tout contrôle volontaire. Les organismes primitifs unicellulaires démontrent que des mécanismes de perception élémentaires induisent un comportement fort complexe qui peut poser des problèmes d'interprétation durant des décennies à la communauté scientifique. Le cas d'une simple amibe, le Dictyostelium discoïdeum est révélateur, et il nous semble intéressant de le décrire rapidement (Glotin, Pinel, Cochard & Laboissière 1997), afin de convaincre le lecteur de nos propos d'introduction générale sur les mécanismes perception.

Le Dictyostelium discoïdeum est un unicellulaire de quelques microns vivant en colonie en forêt et digérant des bactéries. Dès qu'un individu α de la colonie manque d'aliment, il "alerte" ses congénères voisins en émettant une substance chimique qui se diffuse dans le milieu et à laquelle ses congénères sont sensibles (chémotactisme). En quelques heures, environ 10 000 individus migrent vers α, s'associent à lui en formant un agrégat pluricellulaire. Puis cet agrégat se déplace vers une région plus viable distante de quelques centimètres de la zone d'alerte initiale, et se transforme en une sorte de fruit d'un centimètre de haut d'où sont émises des spores pour former une nouvelle colonie dont les chances de survie sont augmentées.

Des spécialistes en Intelligence Artificielle (IA) tentèrent dans les années 1970 d'expliquer et de modéliser ce phénomène d'agrégation de milliers de cellules convergeant vers l'unique cellule centre α qui donna le signal initial d'alerte. Bien qu'il n'y

ait pas ici à proprement parler d'intelligence, ils mirent une dizaine d'année à trouver le bon mécanisme (Glotin et al. 1997).

Ils pensèrent tout d'abord à un mécanisme global de type "top down" : une cellule devient centre d'attraction, et dirige ses voisines vers elle en lui indiquant sa position absolue. Cette interprétation à commande globale a vite été rejetée car elle n'est pas plausible chimiquement étant données les distances de diffusion mises en jeu.

Il faut alors imaginer des relais pour propager l'alerte. Certes chaque cellule peut propager elle même le signal qu'elle reçoit, mais alors, comment se fait-il qu'elle ne devienne pas elle même un centre d'agrégation, et qu'au contraire α soit toujours l'unique centre ?
La conclusion était qu'aucun algorithme de type "top down" ne proposait de solution, et les progrès en biologie ne donnaient pas de nouvelles interprétations pour avancer de nouvelles hypothèses. Suite au courant d'intelligence artificielle distribuée né au milieu des années 1980, un modèle de seconde génération, basé sur le comportement local des amibes, fut proposé. Ce genre d'approche est classée dans la catégorie des algorithmes "bottom up" car elle décrit avec des règles locales un comportement émergent observable à une échelle supérieure (ici les migrations convergentes des amibes). Cette nouvelle approche permit de révéler le mécanisme d'agrégation centripète qui put être confirmé par les expériences des biologistes. Le mécanisme est basé sur le cycle de réponse suivant : à chaque perception de la substance P par une amibe λ, suit une émission de P par λ, et une inhibition des capteurs de λ à P, puis après une période réfractaire, une reprise de la sensibilité à la substance. Les simulations et les expériences prouvent que ce simple cycle est responsable de ce comportement complexe et robuste face aux propriétés très variables de diffusion de la substance dans le milieu (taux d'humidité, température, nature du substrat). Nous voyons donc que cet exemple de communication primitive, mais aux effets complexes, est basé sur quelques règles simples.

Plus tard des systèmes de perception robustes se développèrent chez les métazoaires, en olfaction, puis vision et audition, bien avant le contrôle volontaire et la pensée.

Pourquoi les performances de l'audition chez l'humain sont-elles actuellement inégalées par les meilleurs reconnaisseurs automatiques ? Cette capacité chez l'homme à séparer/identifier les objets auditifs qui se chevauchent est remarquable, cette opération est plus couramment nommée Analyse de Scène Auditive (ASA). La cause des échecs répétés de l'intelligence artificielle voulant modéliser les mêmes capacités réside peut-être dans le paradigme de la logique formelle. Il suffisait, pensait-on, de trouver les règles logiques et d'appliquer des techniques inductives ou déductives pour simuler les activités humaines de traitement de l'information. Dans la réalité, à cause de leur caractère chaotique, les systèmes experts se révèlent inutilisables en dehors d'applications très limitées proches des conditions idéales.

Une nouvelle vague de recherche vise plutôt à tirer partie de nos connaissances en neurophysiologie et psycho-acoustique pour comprendre une part des mécanismes de l'audition, et d'en tirer les grandes lignes de modèles intégrables dans les systèmes de reconnaissance automatique, ce courant est nommé CASA pour Analyse de Scène Auditive Computationnelle.

Nous décrivons dans les parties suivantes les grandes lignes des systèmes de reconnaissance automatique, en parole, puis les principes de l'ASA et les objectifs de notre thèse.

1.3 La reconnaissance de forme : principes de base

L'intelligence artificielle vise à reproduire les facultés humaines les plus élevées. Moins ambitieuse, la reconnaissance des formes se limite à la simulation des capacités humaines de perception, visuelles ou auditives. Mais les résultats ont longtemps tardé à sortir des laboratoires et à se manifester dans la vie quotidienne.

Aujourd'hui, nous voyons naître des modèles qui permettent d'expliquer ou de simuler des phénomènes naturels. Ils ont acquis droit de cité en biologie, en linguistique et, pour ce qui nous intéresse ici, en reconnaissance des formes.

Comment construire des opérateurs de reconnaissance ? Le principe est simple : recueillir les données d'un capteur, c'est-à-dire une représentation (ou le signifiant), et en obtenir une ou des interprétations (le ou les signifiés) par l'exécution d'algorithmes. Le but étant la reproduction des capacités humaines de perception, qu'elles soient visuelles ou auditives. Mais contrairement aux prévisions optimistes des pionniers de l'époque, les problèmes ainsi posés se sont révélés très délicats.

Dès le début l'intelligence artificielle avait affiché de hautes ambitions : reproduire toutes les facultés humaines, y compris les plus élevées, telles que le raisonnement logique et la compréhension du langage. Moins ambitieuse, la discipline scientifique de la reconnaissance des formes (RdF - Pattern Recognition) s'est développée depuis la fin des années 1960 au fur et à mesure de l'augmentation de la puissance et des capacités de mémoire des ordinateurs. Très tôt, dès le début des années 1970, les deux disciplines se sont séparées. Le fait était aussi regrettable qu'inévitable, étant donné les origines différentes des chercheurs de chaque domaine : les premiers (IA) provenaient des mathématiques tandis que les seconds (RdF) étaient en majorité des physiciens et des ingénieurs. Malgré de nombreux efforts des uns et des autres, les résultats des premiers programmes d'IA ou de RdF se sont révélés très éloignés des performances biologiques.

Une des raisons de cet échec est qu'un ordinateur est limité à opérer n'importe quelle transformation d'une donnée de départ, la représentation, en une image d'arrivée, une interprétation. Toutefois, cette généralité se paye par des contraintes importantes : le volume de mémoire, le temps de calcul. Tout algorithme, en particulier un opérateur de reconnaissance, est en effet caractérisé par une complexité de calcul

O(f(n)), nombre d'opérations à effectuer en fonction du volume n des données d'entrée. La plupart des algorithmes utilisés en RdF sont exponentiels, leur complexité varie en $O(e^n)$, ce qui les rend vite impraticables si n est grand, quelle que soit la puissance de calcul et de mémoire à disposition. Comment échapper à cette complexité ? On dispose de trois possibilités, non exclusives :
1. diminuer n, c'est-à-dire avoir affaire à des sous-images, des sous-problèmes.
2. décomposer le processus global de décision en niveaux multiples de décision.
3. utiliser des opérateurs qui ne sont pas exponentiels, mais polynomiaux, en particulier linéaires.

Ce faisant, on introduit nécessairement des erreurs mais on échange de la précision contre de la rapidité. Toute opération de reconnaissance de formes s'organise ainsi selon des niveaux successifs. A chaque niveau, une opération concrète permet de passer d'une représentation à une ou plusieurs interprétations, lesquelles vont constituer la représentation du niveau suivant. Chaque interprétation est plus abstraite, donc plus générale.

Les opérations de RdF apparaissent chaotiques, c'est-à-dire qu'une variation très petite de la donnée de départ peut faire basculer l'interprétation d'une décision à une autre. Ceci distingue particulièrement ces simulations, accomplies par ces opérateurs informatiques, de la perception humaine dont chacun sait, par expérience, qu'elle tolère des bruits et des déformations très importants.

Pour éviter ce caractère chaotique, il est courant d'introduire du "continu" là où l'algorithmique impose du "discret". On tente en effet de simuler des processus analogiques à l'aide de machines qui, par construction, sont d'essence binaire. Un caractère essentiel des décisions continues est de ne pas conclure tout de suite par " vrai ou faux", mais d'affecter à chaque possibilité de décision un coefficient de vraisemblance. Il s'agit d'entretenir l'ambiguïté sur l'interprétation et, autant que faire se peut, de retarder la décision définitive. Nous aurons l'occasion au cours de notre thèse d'en démontrer l'avantage, en particulier à travers les études de reconnaissance partielle binaire ou douce.

Pour obtenir un optimum, une autre leçon de l'expérience consiste à utiliser, pour une même interprétation, plusieurs opérateurs de RdF élaborés selon des principes différents. Chacun fournit un résultat qui, pris isolément, est le plus souvent peu informatif, mais leur combinaison devient pertinente. D'où d'intenses travaux visant à trouver la meilleure façon de combiner les données produites par ces opérateurs. En fait, il semble qu'une simple combinaison multiplicative de leurs résultats - l'addition des logarithmes - soit voisine de l'optimum.

1.4 Introduction à la Reconnaissance Automatique de la Parole (RAP)

La mise en oeuvre d'un reconnaisseur automatique de parole se découpe en plusieurs séquences.

La première tâche s'attaque au problème complexe du découpage de la parole continue en une suite de phonèmes, tout en tenant compte des contraintes phonétiques, lexicales et syntaxiques propres à la langue. Parfois, la segmentation est clairement définie : les objets à reconnaître sont connexes et séparés les uns des autres. Mais le plus souvent, ce n'est pas le cas. Il faut choisir entre différentes hypothèses de segmentation : elles sont évaluées en probabilités a priori et a posteriori par la reconnaissance de l'élément segmenté, y compris la non-reconnaissance : ce n'est pas un phonème. La segmentation implique la reconnaissance et inversement : les deux opérations ne sont pas séparables.

La seconde tâche s'attache à la voix d'un seul locuteur. Elle consiste à reconnaître quelques mots isolés en comparant le signal vocal avec des références acoustiques, préalablement stockées lors d'un processus d'apprentissage. Mais l'extension de cette approche à la reconnaissance indépendante du locuteur (n'importe qui peut alors utiliser le système sans apprentissage) exige un temps de traitement élevé et un espace mémoire considérable : il faut stocker l'ensemble des références acoustiques pour chacun des mots du vocabulaire choisi, avec les différentes prononciations.

1.4.1 Historique de la reconnaissance automatique de la parole

Pour remédier à ces problèmes, une modélisation statistique des différentes prononciations a été proposée au milieu des années 1970. Elle n'a vraiment été adoptée que dans la décennie suivante. L'idée est de remplacer l'ensemble des références acoustiques, représentant les différentes prononciations d'un même mot, par un modèle de ces prononciations. Chaque mot du vocabulaire choisi est alors représenté par ce que l'on appelle un modèle de Markov caché, et la reconnaissance d'un signal de parole consiste à trouver la séquence de modèles qui correspond le mieux au signal.

Ces modèles sont les plus utilisés aujourd'hui car ils offrent les meilleures performances en reconnaissance de la parole. Les modèles de Markov sont d'ailleurs également utilisés dans d'autres domaines, par exemple la reconnaissance des séquences du génome ou de l'écriture manuscrite.

Les années 1980 ont aussi été marquées par l'introduction des réseaux de neurones artificiels dans le traitement de la parole. Pour profiter des avantages et éviter les inconvénients des différentes approches, un ensemble de travaux sur des systèmes hybrides, combinant les méthodes markovienne et connexionniste, ont vu le jour au début des années 1990.

1.4.2 Les applications de la RAP

Mais malgré les gains théoriques et l'optimisme des chercheurs au début des années 1980, le marché des technologies vocales a tardé à se développer. Le problème vient du signal de parole lui-même, car il varie considérablement selon la vitesse de locution, le bruit ambiant, la prise de son, la coarticulation entre mots, etc.

Aujourd'hui, la recherche se poursuit notamment vers la modélisation du langage et la reconnaissance de parole dans des cas particuliers : pour de grands vocabulaires, dans un milieu bruyant, etc.

Un tournant dans la technologie vocale fut l'important succès, en 1992, du système de reconnaissance d'AT&T, qui a permis d'automatiser une partie du travail des téléphonistes. Malgré un petit vocabulaire (cinq mots isolés en mode multi-locuteurs "credit card", "collect-call", etc.), il permit d'accélérer sensiblement l'établissement des appels interurbains et d'économiser "plusieurs centaines de millions de dollars" d'après AT&T. En 1997, la société américaine UPS, qui offre un service mondial de distribution de courrier, a mis en service un serveur vocal pour faire le suivi d'acheminement. Il suffit à l'utilisateur d'appeler le serveur, de donner le code de son envoi, et le système lui indique l'état de l'acheminement. Le serveur peut gérer jusqu'à 100 000 appels par jour. Les bases de parole utilisées durant notre thèse sont similaires à ce type d'application.
Nous travaillerons sur ce type de tâche de reconnaissance avec la base Numbers95 qui contient une cinquantaine de chiffres énoncés continûment, à travers le téléphone par des enfants, des femmes et des hommes de tous âges, ce qui constitue une application aux débouchés industriels.
Pour le grand public, des systèmes de dictée vocale en continu sont maintenant disponibles, les plus connus étant ceux de Dragon Systems (Naturally Speaking) et d'IBM (Via Voice). Ils possèdent des vocabulaires de plusieurs dizaines de milliers de mots.
Nous présenterons dans cette thèse une de nos applications dans le cadre du produit Via Voice.
D'autres applications plus spécialisées permettent de traiter des vocabulaires adaptés aux domaines juridique, bancaire et médical. L'addition envisagée de la reconnaissance et de la synthèse de la parole aux ordinateurs personnels pourrait favoriser l'essor des technologies vocales mais converser de façon spontanée avec la machine ne sera pas réalisable avant longtemps.

Les vingt dernières années nous ont enseigné la prudence, mais certains champs prometteurs de recherche ont été sous-exploités, comme celui de l'Analyse de Scène Auditive que nous développons dans la section suivante.

1.4.3 Enjeu de la reconnaissance robuste de la parole

La parole est un phénomène très variable. Le problème de fond, sur lequel se porte l'essentiel des recherches, est la robustesse des systèmes de reconnaissance

vocale, c'est-à-dire leur résistance aux différentes formes de pollution sonore (bruits de toute nature, convolutifs ou additifs, soufflements, raclements de gorge ou encore bruits de la rue, prise de son dans une voiture qui roule...). Aucun algorithme ne sait aujourd'hui reconnaître sans erreur la parole au milieu d'un environnement bruyant.

1.5 L'Analyse de Scène Computationnelle au service de la RAP

L'Analyse de Scène Auditive Computationnelle (CASA) cherche à modéliser notre capacité à structurer notre environnement sonore.

Un des principes fondamentaux en analyse de scène est le principe de la régularité-singularité. Pour l'illustrer voici un exemple. Il est possible de décomposer un signal de parole en régions périodiques, qui correspondent aux voyelles. Celles-ci sont faciles à reconnaître : elles constituent les parties régulières du signal. En revanche, les consonnes, "accidents" des voyelles, sont les parties singulières, dont l'expérience montre qu'elles sont difficiles à localiser et identifier.

Connaissant la structure du signal de parole, fortement vocalique (donc harmonique), une idée d'application du principe de régularité-singularité serait de détecter en priorité les parties régulières et de les traiter en priorité. Il resterait ainsi les parties singulières, bien localisées, qui pourraient bénéficier de traitements spécifiques.

De même, pour l'écriture, un mot manuscrit, mal écrit, se réduit à une oscillation régulière, périodique, autour de l'horizontale. Les "accidents" de cette partie régulière sont les parties singulières de l'écriture : ascendants, descendants, boucles, ligatures. Une analyse de scène manuscrite tirerait son principe de la connaissance a priori de cette structure de l'écriture.

La CASA tire son principe de l'usage extensif de la structure connue de l'objet dans son environnement (sonore ou visuel). Il semble que le champ CASA ouvre des perspectives de modélisation constructive de l'audition et, plus largement, de la compréhension de nos mécanismes perceptifs, tout en restant conscient des différences considérables entre un ordinateur programmé et un cerveau. Jusqu'à présent, l'explication de la perception et des interprétations des sens était largement "philosophique". Si une immense quantité de recherches a été faite par les physiologistes de la vision et de l'ouïe, nous savons peu de choses sur le fonctionnement du cerveau en tant que système d'information. Ainsi exploiter intensivement la structure du signal, passe par une catégorisation des traits structurels les plus significatifs. Le moyen le plus efficace pour construire cette catégorisation est de comprendre les traits utilisés par nos systèmes sensoriels.

Inspirée par la physique et l'expérimentation psycho-acoustique plus que par les mathématiques, la CASA appliquée à la reconnaissance automatique de parole prend une voie complémentaire à la RdF classique.

1.6 Objectif de la thèse

Les objectifs de notre thèse sont d'apporter des contributions originales et compétitives au paradigme de la reconnaissance robuste de la parole.

Dans cette optique notre thème de recherche se décompose en deux parties :
1- Extraction d'indices de fiabilité du signal.
2- Fusion des flux de données.

Dans ce but nous mettons en oeuvre les approches d'extraction d'indices de confiance sur le signal parole en augmentant ainsi le paradigme de l'approche "CASA".

Deuxièmement nous développons la théorie des reconnaisseurs de type multi-bandes/multi-flux afin de pouvoir opérer dans un cadre stochastique la fusion des différents flux d'information, flux phonétiques propres au reconnaisseur, et flux des indices primitifs cités plus haut.

Nous avons pour objectif de prouver que la fusion de ces deux paradigmes, souvent éloignés de part leur appartenance à des écoles scientifiques malheureusement distantes aux niveaux culturel et conceptuel, peut être réalisée moyennant un certain effort d'ouverture et de synthèse. Mais surtout, nous nous attacherons à montrer qu'une telle approche peut ouvrir un champ de recherches originales et fructueuses et donner naissance à des implémentations de systèmes de reconnaissance de parole tout à fait compétitifs quant à leur robustesse aux conditions adverses.

Le premier thème se rapproche du traitement humain de la parole qui est abordé dans la partie I. Il est envisagé à la fois à la lumière des principes généraux de fonctionnement de systèmes biologiques de traitement de l'information et des contraintes propres des objets acoustiques produits par le système phonatoire humain. Il fait appel tant aux techniques de filtrage, qu'aux principes des cartes sensorielles et à la mise en place de détecteurs spécialisés visant à l'extraction des indices de pertinence sur le signal analysé. Nous développons deux de ces indices dans la partie I : indices d'harmonicité et de localisation.

Dans la seconde partie nous allons extraire de ces indices CASA une information qui pourra être cartographiée de façon systématique en relation avec un indice central en RAP : le rapport des énergies signal sur bruit (ou SNR) qui donne une mesure fiable de la qualité du signal. L'un des enjeux de notre thèse est de coupler ensuite ces cartes avec un système de reconnaissance automatique de parole et de montrer le gain de robustesse acquis. Le schéma général suivi est en figure 1.1. On y retrouve en entrée le signal de parole et le premier module qui extrait les traits et les indices de confiance. Cette information est donnée en parallèle à un reconnaisseur

partiel. La reconnaissance issue de cette fusion peut à nouveau générer un indice de confiance pour les fusions en court.

Le processus de fusion aura pour but de faire coopérer différentes techniques de classification complémentaires. Ce processus peut avoir lieu dans trois espaces (Andrews 1980) :
- 1/ L'espace des formes qui résulte de l'échantillonnage du capteur.
- 2/ L'espace des représentations qui doit orthogonaliser les flux d'information (par exemple la paramétrisation et la modélisation du signal de parole).
- 3/ L'espace des interprétations qui résulte d'une partition de l'espace de représentations en régions disjointes.

On parle de fusion de données dans le cas de fusion dans l'espace (1) ou (2), et de fusion de décision si elle a lieu dans (3).
Nous travaillerons sur les deux cas.

Enfin, dans la partie IV nous estimerons la contribution relative de chaque classificateur en vue de pondérer les scores qui en sont issus. Nous verrons en quoi ces poids sont équivalents à une estimation de la quantité d'information susceptible d'être exploitée par chaque reconnaisseur. Il nous faudra les combiner avec les estimateurs de fiabilité de chaque flux d'information. L'entraînement optimal de ces poids fera l'objet de la partie IV. Nous verrons qu'ils peuvent être entraînés suivant un critère discriminant bayesien.

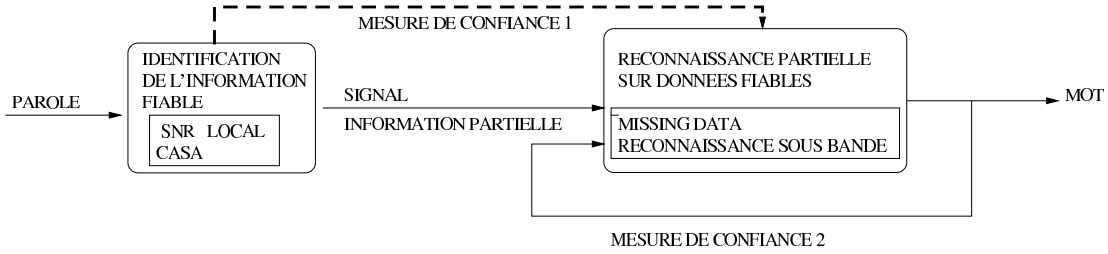


FIG. *1.1: Schéma général de système en RAP robuste. Les mesures de fiabilité peuvent provenir du signal (1), ou des sorties du reconnaisseur (2). Les mesures CASA sont du type (1) uniquement.*

Première partie

Analyse de Scène Auditive et extraction des indices de voisement et de localisation

Dans cette partie nous présentons le système auditif humain et nous décrivons les grandes lignes des modèles d'Analyse de Scène Auditive (CASA). Nous verons que la grande majorité du signal de parole est voisée. On peut tirer avantage de cette propriété fondamentale à travers des traitements simples type corrélation spatio-temporelle. En particulier, les fonctions de corrélation du signal apportent des mesures simples et efficaces des battements du signal. Nous verrons que de tels détecteurs de coincidence sont physiologiquement plausibles. Nous allons donc dans cette partie montrer comment extraire des indices de fiabilité du signal à partir des corrélogrammes.

Chapitre 2

Les mécanismes de l'analyse de scène auditive et sa modélisation

2.1 Objectif et introduction

Nous présentons les points principaux du système auditif humain intéressant l'analyse de scène auditive et la reconnaissance de parole automatique. Certaines propriétés vont inspirer dans les chapitres suivants les extractions d'indices qui seront soit utilisés pour rehausser le signal, soit fusionnés avec notre système RAP "Full Combination" (voir chapitres suivants).

Nous présentons brièvement les mécanismes clefs du système auditif humain, ce qu'est-ce que l'onde sonore et comment cette énergie vibratoire prendra finalement la forme d'une énergie électrique que le cerveau va pouvoir décrypter. Le but de ce chapitre est de comprendre dans les limites de nos connaissances ces mécanismes afin d'imaginer et d'élaborer de nouvelles stratégies d'extraction de traits. Nous montrons les points forts qui ont été exploités depuis des années (filtre Mel, bandes critiques), ou bien ceux dont nous tirerons parti dans nos recherches (les corrélateurs intra ou interaural).

2.2 Le son, les bruits

Le son peut être défini comme représentant la partie audible du spectre des vibrations acoustiques, de même que la lumière se définit comme la partie visible du spectre des vibrations électromagnétiques. L'audition prend essentiellement en compte deux paramètres des vibrations acoustiques : la fréquence ou nombre de vibrations par seconde (exprimés en Hertz = Hz) qui définit les sons aigus et graves et l'intensité ou amplitude de la vibration (décibel = dB) qui définit les sons forts ou faibles. L'oreille humaine perçoit des fréquences comprises entre 20 Hz (fréquence la plus grave) et 20 000 Hz (fréquence perçue la plus aiguë).

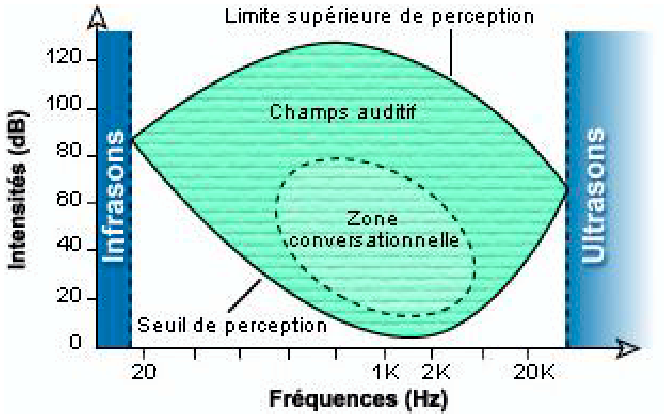


FIG. *2.1: Courbe audiométrique de l'oreille humaine (d'après Pujol).*

On voit que dans la courbe audiométrique de l'oreille humaine 2.1, pour chaque fréquence, le seuil de perception est différent : les fréquences les mieux perçues (la courbe avoisine le 0 dB) se situent dans la gamme moyenne entre 1 et 3 kHz. C'est aussi dans cette gamme que la dynamique de sensation est la plus grande (de 0 à 130 dB). La courbe supérieure représente la limite des intensités tolérables : au delà, il y a douleur et/ou destruction cellulaire dans l'oreille. La zone conversationnelle définit les sons utilisés pour la communication par la voix humaine : ce n'est que lorsque cette zone est affectée que handicap auditif apparaît vraiment.

2.2.1 Les grands traits des sons

Différents types de bruits interfèrent avec la parole, selon leurs caractéristiques spectrale et temporelle, différentes stratégies sont adoptées par les reconnaisseurs de parole.

Propriété	Attributs de la propriété
structure temporelle	continu / impusif / périodique
interaction avec la parole	additive / multiplicative / convolution
stationnarité	stationnaire / non-stationnaire
structure spectrale	large bande / confiné en bande
dépendance avec la parole	corrélé / décorrélé
spatialisation	cohérent / incohérent avec la source de parole

TAB. *2.1: Description des propriétés des nuisances sonores*

La situation la plus délicate à affronter en reconnaissance robuste de la parole est de traiter un signal détérioré par un bruit inconnu non stationnaire, avec cohérence temporelle, fréquentielle et spatiale à la source de parole. Mais ces conditions

extrêmes ne sont pas les plus courantes, et nous pouvons partager les propriétés des bruits des différentes classes décrites dans la table 2.1 et détaillées dans les sections suivantes.

2.2.2 Stationnarité

Une métaphore vis-à-vis de la stationnarité d'un signal est la suivante : imaginez vous en train de regarder un lac. Il est plat n'est-ce-pas ? Mais observez le en détail et vous verrez quantités de vaguelettes, le lac cependant a un comportement stationnaire.

Un bruit stationnaire, tel un bruit gaussien, est un bruit qui globalement ne fluctue pas dans le temps en fréquence. Mais tout est une question d'échelle. Ainsi une voyelle est stationnaire vis à vis d'une fricative, l'une étant stable en fréquence pendant 100 à 250 ms, alors qu'une fricative dure que quelque dizaine de ms. Cependant un bruit peut être qualifié de non stationnaire dès qu'il varie en fréquence toutes les 200 ms. Nous voyons donc que la stationnarité d'un signal est une notion relative à une échelle temporelle.

Si l'on se place à l'échelle de la parole, un bruit stationnaire (stable durant les syllabes au moins) est relativement aisément détectable (voir section dé bruitage classique) car ses énergies moyennes par canaux de fréquence sont stables dans le temps, ce qui le distingue de la parole. Dès lors les techniques de rehaussement peuvent être efficacement appliquées.
Mais une perturbation mouvante en fréquence à l'échelle de la parole, cf. plus haut, est difficilement décelable. L'homme est très habile à cette tâche, mais actuellement les meilleurs reconnaisseurs automatiques de parole sont très perturbés.

2.2.3 Les bruits additifs, bruit de fond

Soit un signal n de bruit, ce bruit est dit additif dès qu'interfère avec la parole ou le signal s cible suivant la loi simple : $y = s + n$

Un bruit de fond est bruit additif, continu, non corrélé à la parole et très courant. Par exemple le bruit de fond dans une voiture est généré par son moteur, par le bruit des roues, des courants d'air aux fenêtres ouvertes etc ... La caractéristique spectrale d'un tel bruit est aléatoire, mais le plus souvent dominante en basse fréquence. La caractéristique temporelle est très variable, de stationnaire (bruit de voiture) à non stationnaire (bruit de sirène).

2.2.4 Bruit impulsif

Ce type de bruit additif forme une classe à part. Le recouvrement spectral de ce bruit est presque entier. La caractéristique théorique de ce bruit est un Dirac.

Les exemples ne manquent pas : citons le bruit de marteau piqueur, mitrailleuse, claquement de porte.

2.2.5 Parole interférente

Ce bruit additif est composé d'un ou plusieurs autre(s) locuteur(s). Dans le cas de plusieurs locuteurs simultanés, l'interférence est connue sous le terme de "cocktail party". Les caractéristiques spectrales et temporelles sont proches de celle du signal de parole.

2.2.6 Distorsion de la source de parole et effet Lombard

Une pathologie de l'appareil phonatoire, une grippe ou autre, vont détériorer la qualité du signal original.
Le stress ou un milieu bruité amène le locuteur à augmenter volontairement ou non, le niveau de sa voix. Cet effet est répertorié sous le terme d'effet Lombard du nom de son premier analyste, une description plus précise se trouve dans (Junqua 1995). La parole avec effet Lombard a des propriétés spectrales différentes de la parole claire : la largeur de bande moyenne des formants décroît, l' amplitude et la durée des voyelles croît, la fréquence fondamentale croît.

Dans (Hansen & Clements 1989) il est montré que le taux de reconnaissance chute de 31% en condition de stress. Des études sont menées pour trouver un algorithme de compensation de ce phénomène. Ces études analysent les traits variables en condition de stress ou bruitée : pitch, intensité, durée, onde glottique, spectre de parole. Du fait de la variabilité inter locuteurs de l'effet Lombard, une compensation automatique est difficilement réalisable. Cependant en distinguant les différentes conditions de stress possibles (rapidité, volume sonore, colère, stress modéré ou fort) une méthode basée sur une moyenne pondérée des coefficients cepstraux est proposée dans (Hansen & Cairns 1995).

2.2.7 Détérioration de la source lors de son traitement

Tout signal traité est enregistré, transmis et digitalisé. Ces opérations malgré l'usage des meilleures techniques, causent des détériorations du signal original.

Par exemple, nos bases de données Numbers93 et Numbers95 sont des bases de données d'enregistrements téléphoniques. La bande passante téléphonique est de l'ordre de [300 3700] Hz, la fondamentale est donc perdue.
De plus, les microphones ont une fonction propre de transfert du signal. Bien qu'adaptés à la parole, le signal convolué de sortie ne restitue pas exactement le signal d'entrée.

2.2.8 Les bruits convolutifs

Les phénomènes convolutifs apparaissent surtout lorsqu'il y a des fonctions de transfert, des échos et des délais.

Ce type de bruit correspond à un mélange acoustique et non à une somme de signaux simplement décalés.

Le signal enregistré subit les fonctions de transfert de la source vers le micro. Ces fonctions de transferts sont liées aux caractéristiques acoustiques (positions relatives et environnement sonore).

Les échos, réverbérations du signal de parole forment une classe de bruit particulière. La thèse de Kingsbury par exemple (Kingsbury 1998) traite uniquement de la reconnaissance robuste de la parole sous ce type d'interférence, dont les traitements sont très spécifiques.

Ce genre de bruit ne peut pas être traité efficacement par les algorithmes classiques de type "Independant Composant Analysis (ICA) qui fonctionnent sur une hypothèse forte d'indépendance des signaux et de délais nuls ou constants entre les sources. Notons que des échos qui se présenteraient lors d'une analyse de l'harmonicité auraient un effet atténuateur (Culling, Summerfield & Marshall 1994).

L'enregistrement de NB95 ST (voir annexe) a été réalisé en chambre sourde mais n'étant pas totalement anéchoïque. Les fonctions de transfert des microphones ajoutent aux réverbérations des bruits convolutifs dans chacun des canaux.

2.2.9 Résumé des perturbations du signal

Le signal utile est donc susceptible d'être dégradé à plusieurs niveaux. La figure 2.2 résume l'ensemble des interactions possibles. Dans cette figure apparaissent n1(n) qui est le bruit de fond, qui à la fois induit l'effet Lombard (voir section effet Lombard) et s'ajoute au signal, puis n2(n) qui est le bruit additif du canal de transmission s'ajoute au bruit du récepteur n3(n). Durant sa réception et sa transmission, les distorsions de type convolutif sont subies par le signal utile.

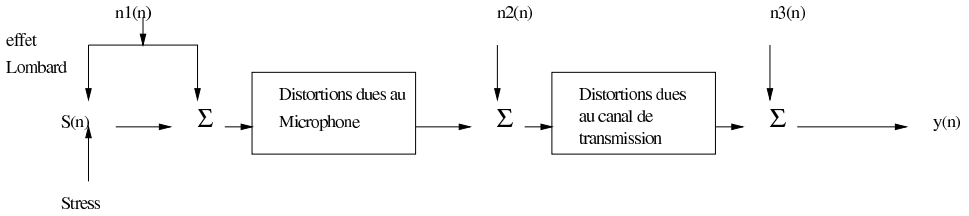


FIG. *2.2: Chaîne générale de bruitage, n1(n) est le bruit de fond, n2(n) est le bruit additif du canal de transmission, n3(n) est le bruit du récepteur.*

L'ensemble de cette chaîne est simplifiable en deux types de perturbations : convolutive et additive :

$y(n) = h(n) * s(n) + n(n)$

où
$h(n)$ est la réponse impulsionelle d'un filtre inconnu (microphone, canal de transmission, réverbération)
$n(n)$ est la somme des bruits additifs.

2.3 Les composants du système auditif

2.3.1 Du pavillon au tympan

En 1967, Wayne Batteau eut l'idée de combler les plis ou circonvolutions du pavillon de sujets humains avec de la cire. Ceux-ci montrèrent alors une certaine difficulté à localiser les sons dans l'espace. Le relief du pavillon est donc utile pour déterminer la localisation des sources sonores : les sons réfléchis accusant un retard par rapport aux sons directs le système auditif il interprète cette transformation des sons en terme de directionnalité.

Le rôle du conduit auditif externe est de recevoir les vibrations sonores et de les transmettre au tympan. Toute la cavité agit en caisse de résonance passive, augmentant la pression sonore transmise de 10 décibels, dans une bande de fréquence comprise entre 200 et 7000 Hertz.

Le tympan sépare le conduit auditif externe de la cavité de l'oreille moyenne qui est en relation avec la cavité buccale par la trompe d'Eustache. La fenêtre ovale, sur laquelle s'applique l'étrier, et la fenêtre ronde séparent oreille moyenne et oreille interne. L'oreille moyenne peut être considérée comme un adaptateur d'impédance sans lequel une très grande partie de l'énergie acoustique serait perdue.

Deux des osselets, le marteau et l'étrier, sont attachés à des muscles qui par commande réflexe peuvent modifier leur mobilité et réduire la fonction de transfert entre l'oreille externe et la cochlée. Ce réflexe ossiculaire protège la cochlée contre les sur-stimulations sonores ... mais avec des limites :

- il n'entre en jeu que pour des fréquences graves (ne dépassant guère 1 kHz)
- il n'intervient pas lors de bruits impulsifs (explosions, armes à feu, pétards, etc.).

Un autre réflexe ossiculaire est déclenché par la vocalisation. Il atténue la perception de la propre voix du locuteur (important pour les chanteurs).

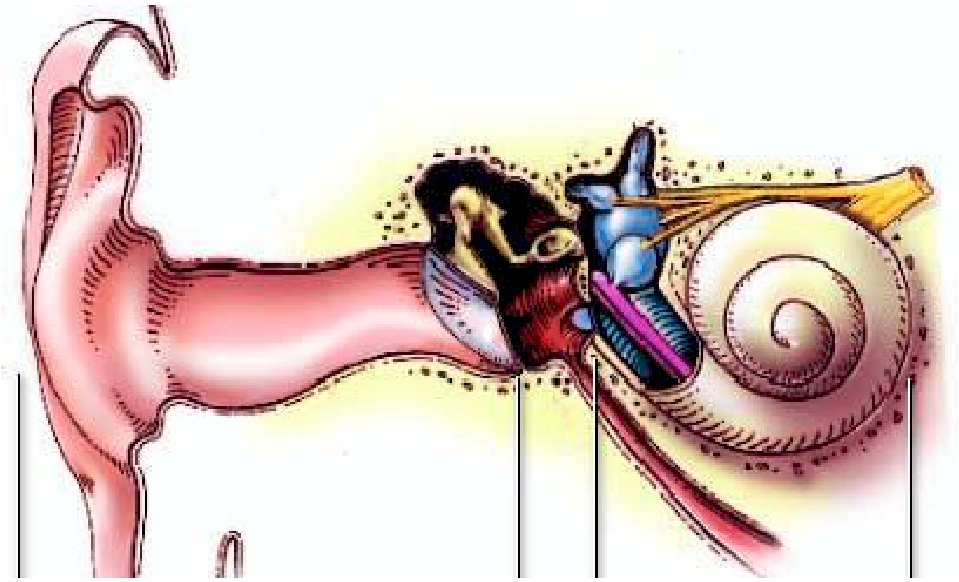


FIG. *2.3: Schéma de l'oreille externe, moyenne et interne. de gauche à droite : fenêtre ovale, marteau enclume, étrier, tympan, trompe d'eustache (Pujol 1999)*

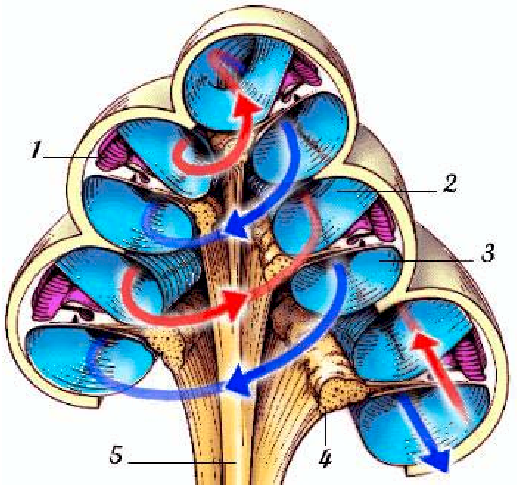


FIG. *2.4: Section axiale de la cochlée : enroulement du canal cochléaire (1) et celui des rampes vestibulaire (2) tympanique (3). Au centre le ganglion spiral (4) et les fibres du nerf cochléaire (5) (Kandel et al. 1991)*

2.3.2 La cochlée

La figure 2.4 schématise la cochlée. La membrane basilaire s'élargie depuis la base (sensible à 20 kHz) jusqu'à l'apex (sensible à 50 Hz).

L'onde transversale, perpendiculaire au plan de la membrane basilaire, change d'amplitude au cours de son déplacement, augmentant peu à peu pour atteindre son maximum dans sa région d'accord suivant les propriétés locales de la membrane avant de décroître rapidement. L'organe de Corti représenté en figure 2.5 est l'organe neurosensoriel de la cochlée. Il est composé des cellules sensorielles ou cellules ciliées, et des fibres nerveuses.

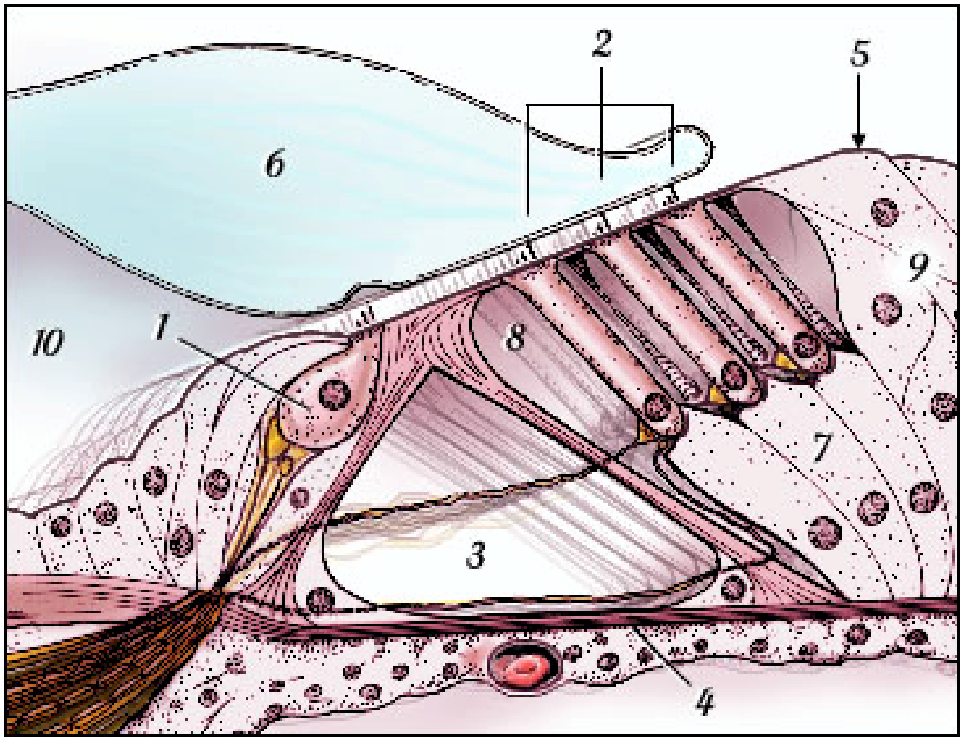


FIG. *2.5: Coupe schématique de l'organe de Corti : 1-Cellule ciliée interne (CCI) 2 :Groupe de cils de Cellules ciliées externes (CCEs) 3 :Tunnel de Corti 4-Membrane basilaire 5 :Lame réticulaire 6 :Membrane tectoriale (Pujol 1999)*

Les cils des cellules ciliées sont de longueur variable. Le déplacement vers le haut de la membrane basilaire et les mouvements liquidiens intra-cochléaire sont excitateurs. Seules les cellules ciliées internes (CCI) sont sensitives. Les cellules ciliées externes (CCE) semblent moduler l'excitabilité des CCI[1].

Chez l'homme, on dénombre environ 3.500 CCI et 12.500 CCE, nombre faible, si on le compare aux millions de photo- ou chémo-récepteurs! En outre, les cellules ciliées ont la propriété de faire leur mitose terminale avant de se différencier; en d'autres termes, leur nombre est fixé très tôt dans le développement et les cellules

[1]Cette remarque est intéressante puisqu'elle inspire des modèles de type rétro-contrôle.

ciliées endommagées au cours de la vie ne sont pas remplacées, ce qui entraîne une perte de sensibilité chez les personnes âgées en particulier.

2.3.3 Des amplificateurs sélectifs : les CCE

Les CCE sont des amplificateurs sélectifs. Les musiciens qui entendent des fractions de ton, auraient un nombre de CCE plus important (Aran, Dancer, Dolmazon, Pujol & Huy 1988). La perception d'acouphènes (sifflements d'oreille), provient de battements de CCE désaccordés, et ce bruit de fond est détecté par les CCI qui intègrent ces auto-variations de pression normalement nulles.

2.3.4 Les voies auditives afférentes

Comme dans tout système sensoriel, nous pouvons classer en deux catégories les voies auditives.

-Les voies afférentes, du capteur vers le système nerveux, ou encore voies centripète, d'information sensorielle, "bottom up".

-Les voies efférentes, du système central vers la périphérie et les capteurs, qui permettent un rétrocontrôle sur les capteurs, voies "top down".

Mécanoréception et transmission du signal auditif sur les voies afférentes

Chaque section de la membrane basilaire est innervée par les cellules ciliées qui sont des mécano-récepteurs. Leurs cils, solidaires de la membrane sont baignés dans un liquide et sont donc déplacés suivant les mouvements du liquide.

La transduction mécano-électrique résulte d'une intéraction mécanique directe entre le stimulus et les canaux membranaires du neurone. La stimulation (déplacement des cils) déforme localement la membrane du neurone et provoque l'ouverture d'un canal ionique et donc des flux d'ions K+ et Na+. Ces derniers déclenchent la dépolarisation locale qui est ensuite propagée sur l'axone.

Le groupement des axones, ici fibres afférentes, des cellules réceptrices des différentes régions fréquentielles de la membrane basilaire forme le nerf auditif à travers lequel les informations du signal auditif (intensité, durée, et suivant la position de la fibre nerveuse la fréquence) sont transmises vers le système nerveux central.

Une cascade de transductions non linéaires

Un transducteur est un dispositif recevant de l'énergie et fournissant une énergie correspondante, mais d'une autre nature. Une propriété importante des transducteurs liés au système auditif est la non-linéarité de leur réponse à l'excitation acoustique. Il n'est pas toujours possible d'établir théoriquement les caractéristiques de fonctions non linéaires. Comme nous le verrons, une fonction de redressement simple

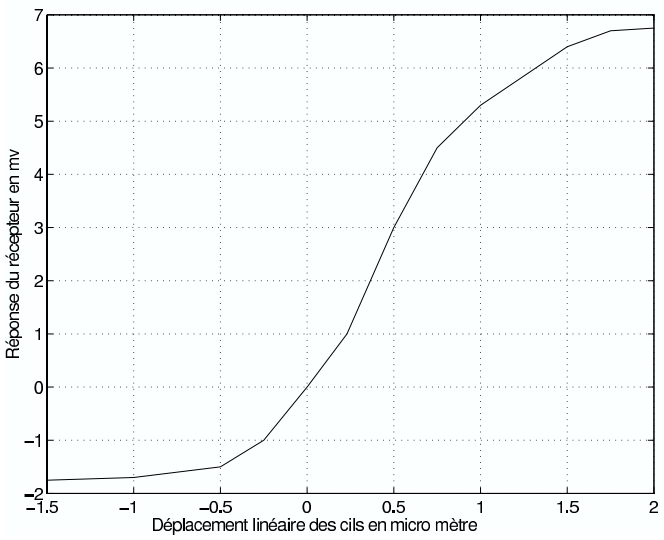


FIG. *2.6: Réponse en mV au déplacement linéaire des cils en micro-mètre par rapport à la position de repos (0). Calculs fait d'après l'angle du cil par rapport à sa position de repos (Aran et al. 1988). Cette réponse est analogue à une rectification simple alternance de l'onde acoustique.*

alternance est complexe à développer et nous ne pourrons guère en dresser une étude théorique que dans le cas de distribution gaussienne.

Après une première transduction de la membrane basilaire vers les CCI, ces dernières traduisent l'information mécanique en information électrique. Une CCI peut suivre la stimulation jusqu'à 20 kHz, mais on considère que c'est la limitation des techniques d'enregistrement de l'électromobilité qui ne permet pas de monter plus haut en fréquence. Dans tous les cas cela indique que la transduction peut se faire à ce rythme. [Communication personnelle Rémy Pujol].

Les voies auditives primaires

L'audition, comme toute autre modalité sensorielle, possède une voie et des centres primaires, c'est à dire totalement dédiés à cette fonction. Schématiquement, cette voie est courte (3 ou 4 relais), rapide (grosses fibres myélinisées) et aboutit au cortex auditif primaire (dans l'aire temporale chez l'homme).
Le premier relais de la voie auditive primaire est constitué par les noyaux cochléaires qui, au niveau du tronc cérébral, reçoivent les terminaisons axoniques des neurones du nerf auditif. Ces seconds neurones de la voie auditive effectuent un travail majeur dans le décodage de base du message : durée, intensité, fréquence.

Un deuxième relais majeur du tronc cérébral est le complexe olivaire supérieur : la plupart des fibres auditives y font synapse. Au départ de ce relais, un troisième

neurone fait monter le message au niveau du mésencéphale (colliculus supérieur). Ces deux relais jouent un rôle essentiel dans la localisation des sources.

Un dernier relais, avant le cortex, est effectué dans le thalamus (corps genouillé médian). C'est là que se fait un important travail d'intégration : préparation d'une réponse motrice (vocale par exemple).

Le dernier neurone de la voie auditive primaire relie le thalamus à cortex auditif où le message arrive largement décodé par le travail des neurones sous-jacents. Le message est reconnu, mémorisé...

Les voies auditives non primaires et les mécanismes d'attention auditive

Après le premier relais (noyaux cochléaires), qui est commun à toutes les voies auditives, des fibres rejoignent la voie réticulaire ascendante commune à toutes les modalités sensorielles. Après plusieurs relais dans la formation réticulée, puis dans le thalamus non spécifique, cette voie aboutit au cortex polysensoriel. Le rôle de cette voie, qui regroupe les différents messages sensoriels envoyés simultanément au cerveau, est de permettre une sélection du type d'information à traiter en priorité ; elle est reliée aux centres des motivations et d'éveil, ainsi qu'aux centres de la vie végétative.

L'intégralité et le bon fonctionnement des voies primaires et non primaires sont nécessaires à la perception consciente. Par exemple, au cours du sommeil la voie primaire fonctionne normalement (les sensations auditives sont décodées). De même, une pathologie affectant le cortex va supprimer la perception auditive tout en laissant s'exprimer les réactions réflexes et végétatives au son. La perception des sons fait donc appel à des mécanismes de bas et de haut niveau.

2.3.5 Voies auditives efférentes et adaptation des capteurs

Les boucles de rétro-contrôle

Nous avons décrit précédemment les capteurs d'où sont issues les voies afférentes du système sensoriel. Mais les propriétés des capteurs seraient mal abordées sans étudier le retour des voies efférentes du système nerveux (Kandel et al. 1991).

Ce schéma traduit cette architecture générale à tout capteur biologique et donne le principe de l'adaptation du système sensoriel.

Rôle du système efférent en analyse de scène

L'influence du système efférent sur le fonctionnement des récepteurs peut-être étudiée directement par activation artificielle ou suppression de ce système, ou par stimulation acoustique controlatérale. Les expériences décrites ci-dessous sont tirées de (Aran et al. 1988) pp. 125-126.

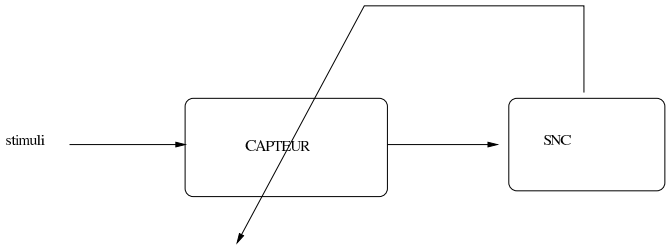


FIG. *2.7: Principes d'adaptation d'un capteur biologique : la réponse d'un capteur soumis à un stimulus prolongé s'atténue grâce à des processus d'adaptation qui peuvent provenir du capteur lui même, du SNC, ou des étapes intermédiaires du traitement de l'information sensorielle entre le capteur et le SNC.*

- Des stimulations électriques du 4 ième ventricule provoquent après une latence importante (50 ms à 100ms) une augmentation du potentiel microphonique cochléaire (c'est à dire l'activité moyenne locale d'une population de CCI ou CCE), une diminution du potentiel d'action composite du nerf auditif et une inhibition de l'activité évoquée sur les fibres sous stimulation auditive.

- La section du faisceau olivo-cochléaire n'entraine pas de diminution du Potentiel d'Action composite du nerf mais une augmentation du premier pic (ce qui correspond à une levée d'inhibition).

- Les stimulations controlatérales modifient la réponse des fibres afférentes ipsilatérales. Un son pur controlatéral (présenté à l'autre oreille) diminue la réponse à un son pur ipsilatéral, d'autant plus que les fréquences sont proches, et proches de la fréquence caractéristique.

Par contre, un bruit blanc controlatéral augmente la réponse à un son pur à la fréquence caractéristique. La latence des effets décrits étant de l'ordre de 100 ms, il est très probable que le système efférent en soit responsable.

L'ensemble de ces observations montre que le système efférent présente une spécificité fréquentielle, et favorise préférentiellement le message afférent de certaines fréquences. Ces interactions pourraient servir par exemple à améliorer la perception de différences de fréquences et d'intensité des sons stimulant les deux oreilles ou à favoriser certaines bandes de fréquences pour améliorer le rapport signal sur bruit.

2.4 Les composantes énergétiques du signal de parole et leurs mesures

Nous présentons dans ce chapitre une description élémentaire des éléments constituants la parole et de leur mesure.

2.4.1 Les éléments de la parole

La parole n'est pas un signal harmonique pur. Elle est constituée d'une partie non négligeable de bruit naturel dont la qualité et la quantité sont décrites dans les deux sections suivantes.

Le bruit perçu dans les segments du signal vocal est dû aux irrégularités des oscillations de la glotte et au bruit additif.

Même certaines parties voisées sont composées d'éléments aléatoires. C'est très marqué dans le cas des fricatives voisées /v/ /z/.

La parole est donc composée d'une part d'éléments harmoniques ou déterministes, périodiques, et d'autre part, d'éléments aléatoires ou stochastiques, qualifiés d'apériodiques.

Soit un signal de parole produit puis enregistré, digitalisé, dont le signal résultant à une énergie totale ET. Cette énergie totale est la somme de trois énergies :

$$ET = EH + EN + EN'$$

où :

EH est la composante harmonique produite par les composantes périodiques du signal.

EN est l'énergie du bruit produit par l'appareil phonatoire, nous l'appellerons énergie de "bruit naturel". Ce bruit naturel provient de deux sources que nous détaillons dans le paragraphe suivant.

EN' est l'énergie du bruit produit par l'environnement (bruit d'autre source et de réverbération), mais aussi par le bruit induit lors de l'enregistrement et la digitalisation. Généralement les amplitudes sont codées sur plus de 6 bits et le matériel est de haute fidélité, dès lors nous pouvons écrire la définition d'un milieu clair (ou encore de la parole claire) (Klingholz 1987).

La parole est dite claire si et seulement si $EN >> EN'$,
ce qui équivaut pour les voisées à $ET = EH$.

Origines du bruit naturel en parole

Deux phénomènes sont responsables de la présence du bruit naturel dans le signal de la parole claire.

La première source génère différents bruits roses ou du bruit blanc additifs à partir du jet d'air qui traverse la zone de constriction glottique alors que la fermeture est incomplète. Ce bruit est présent à toutes les fréquences et est distribué sur tout le spectre.

La seconde source est constituée par les fluctuations aléatoires de la forme de l'onde glottique et génère un bruit dit structurel (d'Allessandro, Darsinos & Yegnanarayana 1998, Klingholz 1987) Ces fluctuations existent à deux niveaux.

Le caractère aléatoire des modulations d'amplitudes induit le phénomène dit "shimmer". Ce caractère aléatoire se retrouve dans l'irrégularité des périodes du fondamental, phénomène connu sous le terme "jitter".

Proportion du bruit naturel en parole

La proportion de bruit naturel dans la parole claire peut être grossièrement approximée par la fréquence des phonèmes de type non voyelle.

Les voyelles étant considérées comme étant des harmoniques pures. Les autres phonèmes comportent plus ou moins du bruit naturel.

Une petite analyse ensembliste ci-jointe donne une vue rapide des trois cas possibles. Nous instancions cette analyse au cas de la base NB95 décrite dans la table A.1.

H = ensemble des phonèmes harmoniques purs, sans bruit naturel, soit l'ensemble des voyelles.

H = / iy ih eh ey ay ah ao ow uw /

NV = ensemble des phonèmes composés de bruit naturel, sans harmonique, soit l'ensemble des non voisés.

NV = / t k tcl kcl s f th hh /

M = ensemble des phonèmes mixtes : harmoniques plus bruit naturel.
M = V et non(H) = phonèmes voisés moins voyelles
M = / n w r l er d v z dcl /

Nous avons bien l'union H u NV u M égale à tous les phonèmes de Numbers95.

Nous obtenons les proportions respectives de ces 3 ensembles en sommant les probabilités a priori (qui sont estimées comme les fréquences d'occurrence sur la base d'entrainement) des phonèmes des ensembles respectifs et en les normalisant (le silence n'est pas pris en compte) :

$P(H) = 48\ \%$
$P(NV) = 24\ \%$
$P(M) = 28\ \%$

Nous voyons donc qu'a priori 76 % du signal de parole (de type chiffres continus) est harmonique.

2.4.2 Mesures de qualité du signal de parole

Il est usuel de décrire le niveau de bruit par le rapport signal sur bruit ou SNR. Le bruit pris en considération est uniquement le bruit ne provenant pas de la source de parole.

L'énergie du signal de parole S est :

$E_s = ET - EN' = EH + EN$

avec $E_s = \sum_{n=-\infty}^{\infty} s^2(n)$, énergie du signal de parole.

L'énergie du bruit non parole est :

$E_n = EN' = \sum_{n=-\infty}^{\infty} n^2(n)$.

Plusieurs définitions du SNR sont utilisées. Nous donnons ici la définition la plus générale et ses variantes : SNR segmental, SNR segmental pondéré en fréquence.

2.4.3 Définitions classiques du SNR

Le SNR est défini comme : $SNR = 10 * log10\frac{E_s}{E_n}$

Cette définition du SNR classique ramenée à l'équation de l'énergie totale est :
$SNR = 10.log_{10}((ET - EN')/(EN')) = 10.log_{10}((EH + EN)/(EN'))$

Cette mesure du SNR que nous utiliserons tout au long de notre étude est encore appelée SNR a priori par certains auteurs.

Notons que pour certaines analyses de la parole en milieu clair (Klingholz 1987), où nous avons $EN >> EN'$, il est nécessaire d'utiliser la mesure SNR_{clair} suivante :

$SNR_{clair} = 10.log_{10}((ET - EN')/(EN + EN')) \ = 10.log_{10}(ET/(ET - EH))$

Cette mesure est simple, mais elle n'intègre pas les variations dans le temps de l'énergie du bruit ou de la parole.

SNR Segmental

Cette nouvelle mesure intègre les variations dans le temps, par intervalles de M trames, soit de 15 à 25 ms en général.

$$SNR_{seg} = \frac{1}{M}\sum_{j=0}^{M-1} 10.log_{10}\left[\sum_{n=m_j-N+1}^{m_j}\frac{E_s}{E'_n}\right]$$

SNR segmental temporel et fréquentiel

Un raffinement de la mesure précédente consiste à faire cette mesure par bande de fréquence. Soit M le nombre de trames de parole, K le nombre de bandes de fréquence, l'énergie à court terme du signal contenue dans la K ième bande de fréquence est $E_{s,k}(m_j)$ et son équivallent pour le bruit est : $E_{n,k}(m_j)$. Cette mesure est affectée également par des poids suivant des critères perceptifs $w_{j,k}$. Ainsi :

$$SNR_{fw-seg} = \frac{1}{M}\sum_{j=0}^{M-1} 10log_{10}\left[\frac{\sum_{k=1}^{K} w_{j,k}10log_{10}\left[\frac{E_{s,k}(m_j)}{E_{\epsilon,k}(m_j)}\right]}{\sum_{k=1}^{K} w_{j,k}}\right]$$

Mesure XNR ou SNRpost

Nous retrouvons dans la bibliographie une définition annexe du SNR nommée XSNR ou "RSB croisé" XRSB ("Cross SNR ou XSNR") ou encore SNRpost pour SNR a posteriori. Suivant nos notations précédentes, le XNR est défini comme :

$XRSB = 10 * log(ET/EN')$
Suivant l'hypothèse du bruit additif ($ET = (EH + EN) + EN'$)
nous avons :

$XRSB = 10 * log(((EH + EN) + EN')/EN') = 10 * log((E_s/EN') + 1)$
d'où

$SNR = 10log(10^{(XNR/10)} - 1)$

2.4.4 Mesure HNR de l'énergie et de la qualité du signal de la parole

La proportion EH/EN est inconnue dans les mesures SNR ; or cette mesure est utile pour l'analyse des pathologies de l'appareil phonatoire le plus souvent dues à des pathologies du larynx (Klingholz 1987, Yumoto, Gould & Baer 1982) de parole (d'Allessandro et al. 1998) Notons de plus que dans le cas de segment de parole claire non voisée SNR = +inf et SNRclair = 0 dB.

Il serait cependant plus adéquat dans ce cas d'obtenir une valeur tendant vers moins l'infini pour rendre compte de la nature aléatoire de ce signal non voisé.

Une autre mesure est donc usuelle : le rapport Harmonique sur bruit défini comme (Klingholz 1987) :

$HNR = 10 * log_{10}(EH/(EN + EN')) = 10 * log_{10}(EH/(ET - EH))$

Nous montrerons dans notre thèse que cette mesure est appropriée à la fonction d'autocorrélation et qu'elle est utilisable dans le cadre du traitement de la parole bruitée, en l'associant statistiquement avec la mesure de type SNR.

2.5 Un bref état de l'art du CASA

Depuis plusieurs années la communauté scientifique s'intéresse à la modélisation du processus de décomposition de mélanges de sources sonores arbitraires chez l'homme, en composantes intelligibles parole musique par exemple, ou musique bruit, ou bruit parole etc ...

L'Analyse Computationnelle de Scène Auditive (CASA) propose des solutions implémentables d'extraction des traits des sources, et de reconstruction de ces sources.

Ce courant de recherche propose de modéliser notre capacité à structurer notre environnement sonore. Pour ce faire, l'une des approches envisagées consiste à considérer que cette capacité de notre système auditif résulte de l'utilisation, en coopération, de plusieurs images du plan temps fréquence. Construites à partir de l'extraction d'indices primitifs des signaux qui structurent la parole, l'approche CASA est une théorie proche du STRUCTURALISME : elle décrit la parole en tant que système dans lequel les éléments entretiennent des relations mutuelles de cohérence.

Typiquement, les systèmes CASA reposent sur des considérations neurosensorielles de la perception. Mais les modèles CASA sont par essence pluridisciplinaires, et se doivent d'intégrer les solutions apportées par des champs de recherche connexes comme la neurophysiologie, la psycho-acoustique, le traitement du signal, la reconnaissance automatique et l'IA.

Le paradigme CASA propose une approche générale du problème de l'extraction de l'information lorsqu'une source sonore est noyée dans un fond auditif bruité. Le fond auditif peut provenir d'un bruit environnemental, mais également d'autres interlocuteurs ou de la musique.

Un moyen d'aborder ce problème est de considérer la scène auditive comme un ensemble de plusieurs 'objets auditifs' et de concevoir un pré-traitement dans lequel les composants sont séparés et groupés objet par objet avant identification. Cette approche est généralement plus complexe que d'autres méthodes, cependant les données perceptibles tendent à montrer que c'est la stratégie utilisée par les auditeurs humains (Bregman 1990). L'attention porte principalement sur le domaine du 'groupement' des objets auditifs de bas niveau. Les propriétés acoustiques et modèles ASA disponibles à l'heure actuelle suggèrent que les primitives de début et fin d'événement parole ('onset-offset'), le taux d'harmonicité, la modulation d' l'ampli-

tude ou de fréquence, et la localisation spatiale sont les principales caractéristiques qui déterminent les décisions de groupement auditif.

Cependant, à l'heure actuelle, il n'existe aucune définition des objets auditifs de bas niveau et des règles déterminant leur formation. Par exemple le rôle de la FM comme indice primitif de ségrégation/fusion des flux auditifs est controversé dans la bibliographie, mais les publications les plus récentes montrent que la FM n'est pas une primitive.

2.5.1 La perception auditive et codage de l'information acoustique

Le codage de l'information acoustique est modulé par les voies efférentes qui sont résumées dans le schéma 2.8. Le Système Nerveux Central (SNC) donne un rétrocontrôle sur le Système Auditif Périphérique (S.A.P.), en particulier sur la membrane basilaire et sur les cellules ciliés. Ainsi il est possible que le système auditif fasse varier le poids qu'il attribue à chacun des indices phonétiques en fonction des conditions de stimulation. Cette fonction d'adaptation est une constante fondamentale des organismes vivants. Ce genre d'architecture est une source d'inspiration pour les modèles CASA.

2.5.2 Une analyse fine de signaux complexes

Le système auditif est capable d'analyser la composition d'un son complexe et d'en localiser la source, et ce même pour des énergies d'ondes sonores très faibles, et lorsque les sons sont des mélanges d'une multitude de fréquences différentes, noyées dans un environnement bruité.

Cette remarquable analyse du signal est accomplie par le système très sophistiqué de transduction électrique de l'oreille interne, en couplage avec le système nerveux central qui compare les signaux des deux oreilles.

Actuellement aucune technologie n'approche de telles performances en sensibilité sur une dynamique aussi vaste. L'humain est capable de détecter les sons dans une gamme de fréquence de 20 à 20 000 Hz (10 octaves) sur une gamme dynamique de 130 dB, et une résolution spatiale de l'ordre du degré d'arc.

2.5.3 Formation et contenu du codage sous-bande

Le codage digital de l'onde hydrodynamique basé sur les trains de potentiels d'action émis par les cellules ciliées, et parcourant les fibres du nerf auditif, donne un profil de l'onde stimulante. Ainsi sont transmis : son spectre, son amplitude et les

relations de phase entre les différentes composantes fréquentielles. Ce codage complexe n'est pas encore bien compris, mais il est clair que certaines fibres répondent spécifiquement à certaines bandes de fréquences assez étroites. Les expériences de Fletcher (Fletcher 1953,[1929]) repris par Allen (Allen 1994) ont inspiré les modèles de reconnaissance automatique multi-bandes. En effet ils montrent que les erreurs de reconnaissance produites sur des stimuli de types bandes spectrales se combinent pour former l'erreur globale du stimulus pleine bande. D'après ces résultats, le système auditif traiterait le signal en sous-bandes indépendantes. Malheureusement le système de combinaison, s'il existe, est inconnu. Nous proposerons dans ce mémoire des modèles qui pourraient rapidement évoluer aux cours des avancées en psychoacoustique ou neurophysiologie.

2.5.4 Le principe des indices primitifs et cartes sensorielles

Tous ces flux d'information montent en parallèle vers le cortex auditif où les dimensions temporelle, énergétique et fréquentielle des signaux sont cartographiées.

Les fibres du nerf auditif présentent un arrangement selon les fréquences, dit tonotopique, telles que celles provenant de l'apex de la cochlée (fréquences graves) soient au centre, entourées par celles issues de la base (fréquences aiguës). L'influx généré par le stimulus sonore va effectuer de nombreux relais dont le plus important se situe dans le thalamus. Le nerf auditif pénètre dans le tronc cérébral au niveau de la jonction bulbo-protubérantielle et y chemine jusqu'au cortex auditif. Au cours de ce trajet imposé, plusieurs nouveaux paramètres seront analysés, affinant le message et préparant son intégration par le cortex. Entre autres traitements, citons principalement ceux des informations liées à la localisation spatiale et à la binauralité (comparaison des messages provenant des deux cochlées). Le décodage complet des différentes variables sonores telles que la fréquence, l'intensité, la durée et la localisation spatiale, s'achève avant que le message n'atteigne le cortex.

Même si la manière dont le cerveau organise les messages sonores et les rythmes transmis par l'oreille est encore très mal connue, les cartes sensorielles sont des éléments de réflexion très intéressants. Tout comme on remarque que les informations qui vont vers ou viennent de zones adjacentes du corps sont adjacentes dans le cortex, il existe des cartes fonctionnelles du cortex pour l'information auditive. Chaque petite section fréquentielle de la membrane basilaire y est représentée de manière précise. L'information sonore est représentée dans le cerveau par cartographie ce qui peut permettre de concentrer l'information, d'amplifier les contrastes... Il est démontré que des cartes sensorielles reflètent la tonotopie iso-fréquentielle et iso-intensité, et il est avancé plus récemment qu'il en existe pour la périodotopie mais cela reste spéculatif (Langner & Schreiner 1988).

La diversité des zones spécialisées du cortex auditif reflètent la complexité de la perception des scènes auditives. Comme dans le cortex visuel où les formes, les

couleurs et la stéréoscopie sont traitées, cartographiées dans les zones distinctes, dans le cortex auditif différentes régions fonctionnelles décomposent la parole pour générer une perception de la :

- la hauteur de timbre (analogue à la couleur en vision)
- la localisation (analogue à la vision 3D, contour, superposition en vision)
- l'intensité (analogue à la luminosité en vision)

La présence d'une carte pour une dimension donnée du stimulus signifie que cette dimension est écologiquement importante pour cet organisme. C'est un principe d'organisation sensorielle. Cela assure une rapidité d'accès à l'information ainsi cartographiée.

2.5.5 Les grandes étapes d'un modèle CASA

Nous résumons ici les points forts d'une approche dite CASA couplée avec un système de reconnaissance automatique. Les points forts successifs sont : Processus amont, Etape de représentation, Etape de Groupement, Fusion et resynthèse et/ou Etiquetage dans l'espace de reconnaissance.

Processus amont

Le processus amont de CASA est dédié à extraire des traits acoustiques primitifs. Un point important et original par rapport aux système classiques, est que la variable SNR est une variable cachée.

En fait, les traits primitifs sont plus fins, et peuvent être catégorisés ainsi comme le montre la figure 2.9.

Etape de représentation

Les données provenant de l'analyse CASA sont de bas niveau. Une étape de représentation permet d'en tirer l'information de plus haut niveau, et de former l'information de base de l'analyse de scène.

Etape de Groupement

Ce traitement s'ajoute à l'étape de représentation. L'hypothèse de regroupement de Bregman (Bregman 1990) est reprise dans cette étape : regroupement des éléments qui semblent appartenir à une même source.

Fusion et resynthèse

Les sorties des analyses précédentes doivent être converties pour être compatibles avec le système de reconnaissance par exemple. Cette conversion peut être faite par resynthèse du signal rehaussé (Tessier, Berthommier, Glotin & Choi 1999, Berthommier & Glotin 2000)

2.5.6 Cartes de modulation d'amplitude

Berthommier et al. dans (Berthommier & Meyer 1995) propose la modélisation d'une cartographie de type Amplitude de Modulation. Cette approche est basée sur des expériences physiologiques qui suggèrent que les cellules du nucleus central du colliculus inférieur sont sensibles à des taux de modulation de l'amplitude choisis très précisément : les neurones pourraient être organisés selon des fréquences de 'best stimulus' et 'best modulation d'amplitude' (Langner & Schreiner 1988). En conséquence, en modélisant ces effets (fonctionnellement), on espère qu'en examinant une périodicité à long-terme, on pourra obtenir des mécanismes de groupement efficace, au moins pour ce qui est de la séparation des voyelles doubles. D'un point de vue fonctionnel, ce travail requiert une estimation précise des fréquences AM afin de produire une 'carte AM' du spectre, dans ce cas obtenue par démodulation (rectification simple alternance + filtrage passe bas) des filtres passe bande Gammatonne. La 'carte AM' résultante aboutit à une représentation similaire à celle de l'autocorrélogramme , mais avec un degré de plausibilité biologique supérieur. La prochaine étape utilise un filtre de type peignage harmonique pour produire un certain nombre d'estimation du ton, chacune de ces estimations étant ensuite mise en commun avec la carte AM pour produire une estimation de l'énergie spectrale liée à chaque candidat de pitch. Les spectres résultants sont alors identifiés en utilisant un Réseau Neural, soit un 'LVQ-Kohonen classifier', soit un Perceptron Multi-Couches. Lorsque ce système est testé avec le bruit blanc et AM, il fournit un taux de réussite de 60% à SNR -6 dB, pour la reconnaissance voyelle/voyelle environ 60% des stimuli conduisent à la reconnaissance des deux voyelles, avec au moins une voyelle détectée à 100%. En outre, ce système a montré une hiérarchie de la dominance spectrale des voyelles, ou certaines voyelles étaient de manière significative plus facilement détectables que d'autres. Cela correspond au concept de 'dominance' d'une source sur l'autre, débouchant sur la nécessité de suivre toutes les sources simultanément.

En ce qui concerne notre travail nous effectuons une "analyse de périodicité", mais nous ne faisons pas un mapping du pitch (comme avec les cartes AM).

Du point de vue de la tonotopie, 3 questions semblent se poser :

1- existe t-il des sélectivités neuronales à la fréquence de modulation d'amplitude ?

2- existe t-il des cartes organisant spatiallement ces sélectivités neuronales pour la fréquence de modulation ?

3- ces sélectivités et ces cartes, si elles existent, ont-elles une pertinence comportementale ? En d'autres termes, le système auditif se comporte t-il comme un banc de canaux de modulation sélectifs pour la fréquence de modulation ?

Ce type de question reste ouvert pour plusieurs analyses bas niveau du signal et font actuellement l'objet de recherches (Langner, Sams, Heil & Schulze 1997)

Notre approche peut être interprétée comme regroupant 3 types de cartes : tonotopie iso-fréquentielle, iso-ITD et iso-harmonique.

2.5.7 Corrélation spatio-temporelle et localisation de sources

Comme nous l'avons vu précédemment, le cerveau utilise le relief temporel (retard de réception des sons perçus entre les deux oreilles), relief amplifié et enrichi par la topologie des pavillons, pour déterminer la localiser spatialement des sources sonores. La traduction du relief temporel en information de localisation des sources se produit dans le cerveau : les signaux des deux oreilles sont combinés dans les voies ascendantes. Ces voies séparent les informations temporelles et d'intensité des signaux et aboutissent aux indices binauraux de localisation spatiale des sources.

Pour localiser une source sonore, le cortex auditif intègre les différences interaurales des intensités et des délais d'arrivée du son.

Les mécanismes neurobiologiques de l'écholocalisation chez la chauve souris donnent des pistes dans la compréhension des mécanismes de localisation chez l'homme basés sur les délais d'arrivée des sources et dont le schéma neurologique peut être interprété comme un corrélateur à retard ce qu'a proposé Licklider (Kandel et al. 1991).

L'information de localisation est très bénéfique en conditions bruitées. La technique qui fonctionne le mieux à l'heure actuelle, réside dans l'utilisation d'un vecteur de microphones ($4 * 2$ microphones par exemple fixés à un mur) destiné à la localisation spatiale basée sur la mesure des décalages de temps (Brandstein, Adcock & Silverman 1995). L'estimation des délais des sources cibles restent précise dans différentes conditions de SNR tout en étant suffisamment simple au niveau des calculs pour des systèmes en temps réel. L'information du délai du son peut alors être utilisée pour localiser la position des paroles et pour suivre les sources mobiles.
Sur les problèmes de localisation et l'approche CASA citons les travaux (Tessier 2001).

Il est intéressant de noter que dès la fin des années 1950 Licklider proposait un modèle neuronal de détecteur de coïncidence. Ce modèle est tout à fait apte à construire un code de type ITD. Nous montrons dans la figure 2.11 le schéma du modèle repris dans (Aran et al. 1988, Kandel et al. 1991). Ce modèle est basé sur

le fait que l'information circule sur les axones des neurones à une vitesse finie. Les retards du signal aux entrées des différents intégrateurs (neuronnes A,B,C,D ou E) sont dûs à la longueur des axones (dt entre chaque neurone dans cette figure), ou au retard des sources. Si le son arrive aux deux oreilles simultanément au temps Td=Tg, le neurone C donne la réponse maximum dans le code ABCDE puisque C est doublement excité et que les autres neurones n'ont pas d'excitations simultanées. Si le son est retardé dans l'oreille gauche, Tg=Td+T, chaque cellule présentera une excitation maximale pour une valeur donnée du délai. Le mot ABCDE forme directement un code de délai d'arrivée du signal entre les deux oreilles. De récents travaux montrent la présence de ce type de réseaux chez les mammifères (Kandel et al. 1991) et peut former une structure tonotopique ITD.

Il est intéressant de noter que le modèle de coïncidence de Licklider (Licklider 1959) est adaptable au calcul des autocorrélations dans le cas ou la même source est donnée aux mêmes entrées. En effet l'autocorrélateur neuronal de Licklider repose sur l'idée que la fonction d'autocorrélation d'une séquence temporelle de potentiels d'action pourrait être déterminée par un système neuronal simple schématisé en figure 2.12. Le message temporel véhiculé par les fibres du nerf auditif lors de l'émission d'un son pur est mis en relation avec plusieurs versions retardées de lui même, les retards résultant de délais de transmission synaptique. Le message est convoyé sur une voie directe et sur une seconde voies où intervient une série de délais. Les deux voies convergent au niveau d'un ensemble de neurones autocorrélateurs. Chacun de ces neurones reçoit le message original et sa version retardée d'un temps t qui dépend du nombre de synapses franchies sur la ligne à délais. Les neurones autocorrélateurs sont spatialement organisés selon une fonction monotone de t.
Pour un neurone qui reçoit le message retardé de t ms après le message original, les deux messages d'entrée coïncident temporellement à l'émission d'un son de période différente. Un tel neurone pourrait donc ne répondre que s'il lui parvient simultanément deux impulsions. "On obtiendrait alors un recodage tonotopique de l'information temporelle" (Aran et al. 1988).

Le modèle de Licklider peut servir de support en faveur de l'existence d'autocorrélateurs et d'intercorrélateur dans le système auditif génèrant des primitives lié aux harmoniques du signal ou à la localisation de sources

2.5.8 Autres travaux CASA couplés à la RAP

Un nouvel outil de marquage du plan temps-frequence par détection d'harmonicite exploitant une statistique de passages par zéro est décrit dans (Gaillard 1999, Gaillard, Berthommier, Feng & Schwartz 1997) donné une perspective en RAP. Son analyse est détaillée dans le chapitre suivant.

D'autres travaux reposent sur l'approche RAP couplée au CASA : (Ellis 1997, Green, Cooke & Crawford 1995*a*)

2.6 Conclusion sur l'approche CASA

En résumé, l'approche CASA consiste en la ségrégation puis fusion d'objets perceptuels. Ces opérations se ferait suivant des caractéristiques dites "primitives" car de traitement bas niveau. Parmi les traits primitifs ou structurels, notons la dimension la dimension d'auto-corrélation, de corrélation inter-aurale, et les transitions fréquentielles qui repère les changement (accidents) temporels pouvant signaler le début d'un son. Le CASA doit fournir une carte de fiabilité de la représentation temps fréquence au module de reconnaissance afin qu'il focalise son activité sur les régions où sont codés les objets perceptuels cibles (Green, Cooke & Crawford 1995*b*, Cooke, Morris & Green 1996).

Nous avons choisis d'étudier les traits primitifs d'harmonicité et de localisation (ITD) de source. Nous avons montré qualitativement que le taux d'harmonicité représente une mesure de qualité du signal). Une étude quantitative sera présentée dans le chapitre suivant. L'ITD sera appliqué comme indice de ségrégation dans le cas de sources simultanées.
Nous avons montré qu'ITD et R sont deux indices dont l'extraction est physiologiquement plausible car étant donné la simplicité du modèle de type de Licklider , on peut penser que des groupes neuronaux jouent le rôle d'auto- ou inter-corrélateur.

Si l'on admet ce mécanisme de structuration de l'environnement sonore, il peut être modélisé à deux niveaux après extraction de l'information primitive dans l'espace de perception : soit avant soit après reconnaissance.

Notre thèse repose sur le couplage d'indices primitifs de types CASA avant (rehaussement) et après reconnaissance phonétique. Nous pouvons faire un paralléle entre une opération précoce et les opérations de type ségrégation ou séparation des sources avant reconnaissance, et les fusions tardive servant de fusion au objets partiellement reconnus.

Les modèles de fusion directe (précoce) ou tardive sont courants en fusion audio-visuelle (Rogozan 1999, Teissier, Robert-Ribes & Schwartz 1999, Teissier, Robert-Ribes, Schwartz & Guérin-Dugué 1999).

Dans le cadre de la reconnaissance mono-modale, nous construirons un modèle de rehaussement du signal original, simple et efficace via les indices R ou ITD. Ce type de modèle avant est présenté en figure 2.10.
L'etiquetage des objets dans l'espace de reconnaissance, ou fusion tardive fera appel à la probabilité de bruitage extraite à partir de cartes des indices CASA construites a priori.
Cette approche de reconnaissance partielle puis fusion est proposée depuis Fletcher (Fletcher 1953,[1929]) puis reprise par Allen (Allen 1994). Ils montrent que les erreurs de reconnaissance produites sur des stimuli de types bandes spectrales se combinent pour former l'erreur globale du stimulus pleine bande. D'après ces résultats, le système auditif traiterait le signal en sous-bandes indépendantes. Nous

aborderons donc la reconnaissance partielle sous cet angle multi-bande.

Dans le principe d'étiquetage, nous pouvons considérer que les processus CASA agissent suivant un schéma direct indépendant de l'étape de reconnaissance, comme le soutient la thèse de Bregman (Bregman 1990). C'est ce que nous traiteront dans les premières parties de la thèse.

Dans (Glotin, Tessier, Bourlard & Berthommier 1998*a*, Glotin, Tessier, Bourlard & Berthommier 1998*b*, Glotin, Berthommier, Tessier & Bourlard 1998, Berthommier, Glotin, Tessier & Bourlard 1998) nous avons présenté un paradigme de reconnaissance partielle guidée par une mesure de type CASA 2.13 provenant d'indice d'harmonicité. Dans ce modèle représenté dans ce diagramme 2.13 chaque reconnaisseur est sélectionné (ici 1 parmi 4) et les autres sont exclus. C'est une décision booléenne de rejet ou non de l'estimation des probabilités qui est appliquée à chaque trame.

Ce modèle 2.13 est le prototype de fusion CASA/RAP qui fût développé dans nos premiers travaux, mais qui sera étendu dans les chapitres suivants à un modèle plus efficace dit "Full Combination".

Un schéma plus complexe est que le processus agit en parallèle et en interaction plus forte avec le processus de reconnaissance, en dépendance au classe phonétiques traitées, comme nous l'aborderont en partie IV.
De plus les arguments neurophysiologiques en faveur de rétrocontrôles, en provenance de noyau traitant des primitives acoustique, sur l'étape primtives de reconnaissance des objets propres à la parole ont été données dans les sections précédentes. Nous aborderont en perspective de la thèse de telles architectures dites "top-down".

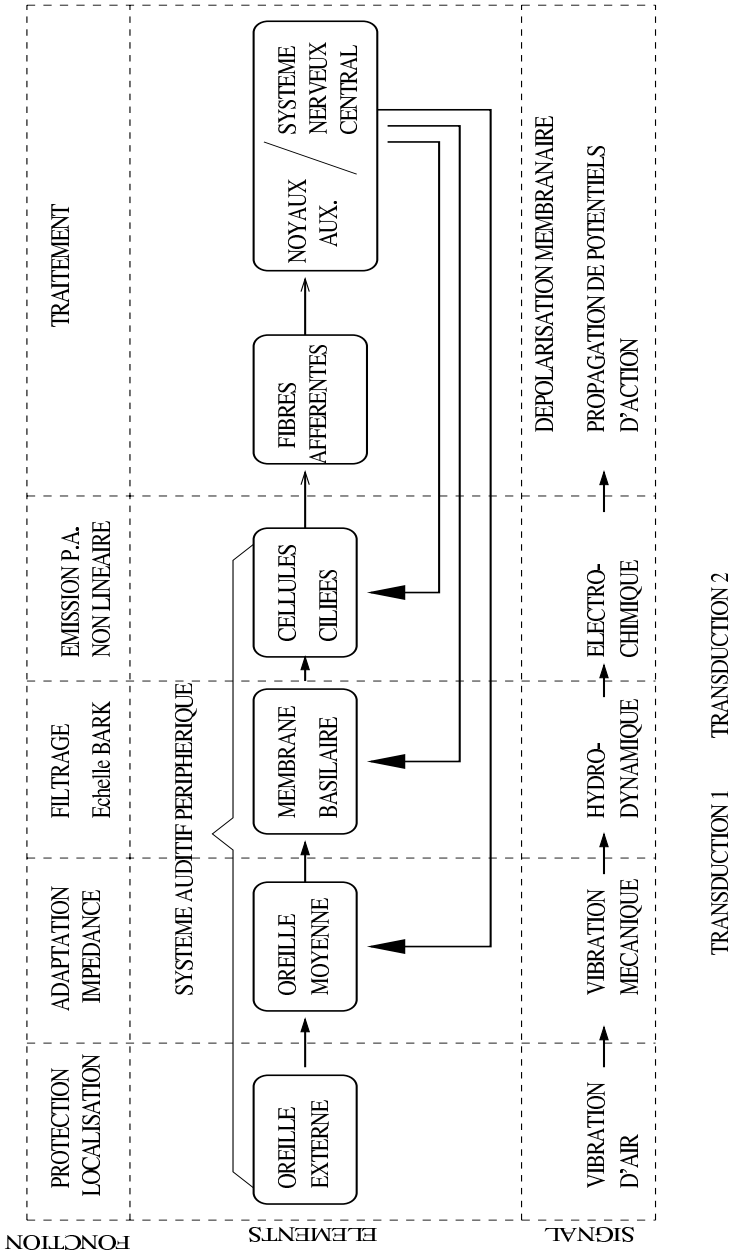


FIG. *2.8: Schéma synoptique du système auditif périphérique. De haut en bas, les fonctions remplies par les éléments anatomiques (centre), et les supports de l'information véhiculée. Noter le rétrocontrôle du Système Nerveux Central sur le Système Auditif Périphèrique (S.A.P.), en particulier sur la membrane basilaire et sur les Cellules ciliés. Ainsi il est possible que le système auditif fasse varier le poids qu'il attribue à chacun des indices phonétiques en fonction des conditions de stimulation. Cette fonction d'adaptation est une constante fondamentale des organismes vivants . Nous verrons plus loin le parallèle entre cette architecture et celle de notre CASA labelling.*

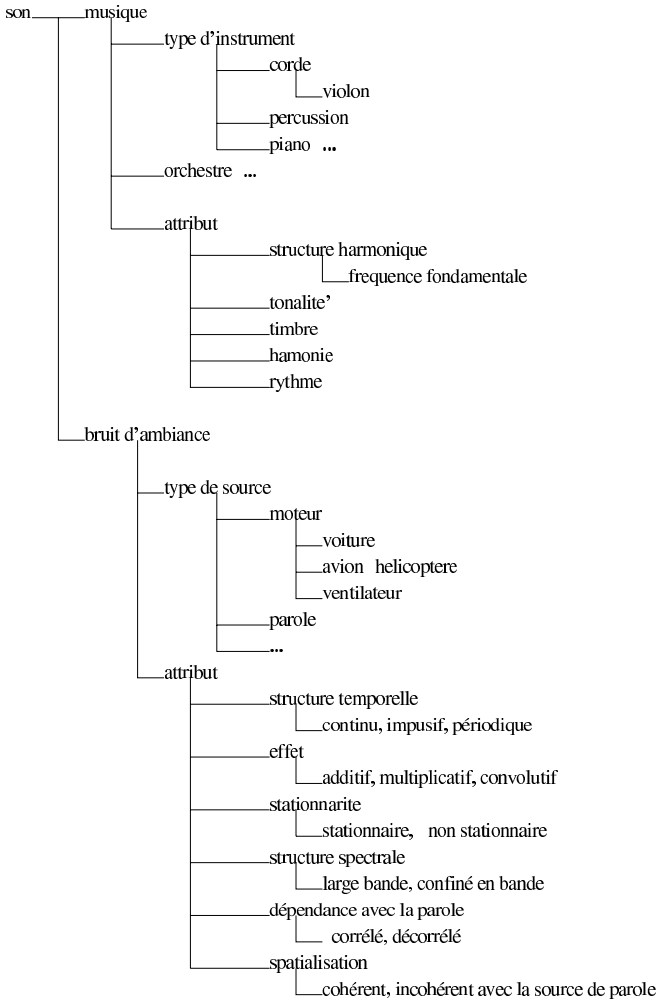


FIG. *2.9: Catégorisation des interférences non parole et leurs primitives ASA*

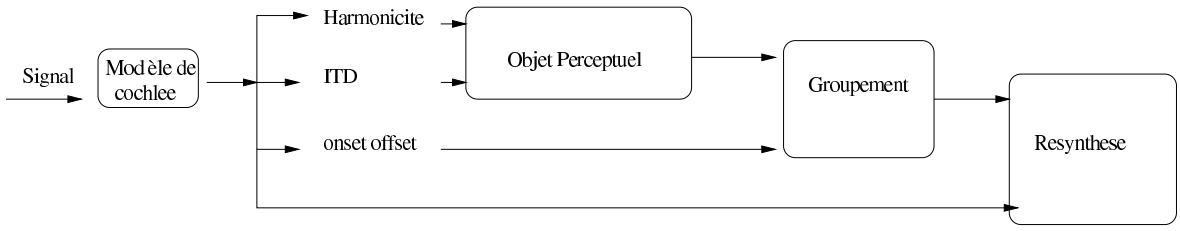


FIG. *2.10: Architecture classique d'analyse "data driven"*

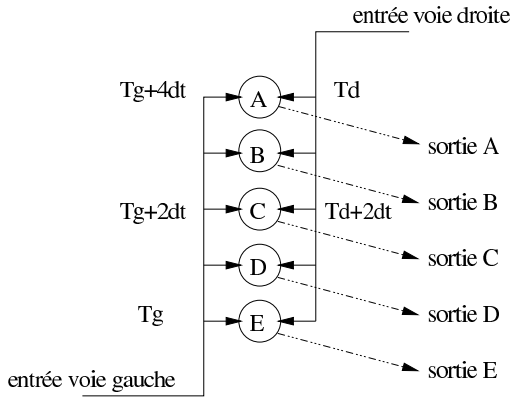


FIG. *2.11: Schéma du modèle du détecteur de coïncidences de Licklider repris dans (Kandel et al. 1991). Ce modèle est basé sur le fait que l'information circule sur les axones des neurones à une vitesse finie. Les retards du signal aux entrées des différents intégrateurs (neurones A,B,C,D ou E) sont dus à la longueur des axones (dt entre chaque neurone dans cette figure), ou au retard des sources. Si le son arrive aux deux oreilles simultanément au temps Td=Tg, le neurone C donne la réponse maximum dans le code ABCDE puisque C est doublement excité et que les autres neurones n'ont pas d'excitations simultanées. Si le son est retardé dans l'oreille gauche, Tg=Td+T, chaque cellule présentera une excitation maximale pour une valeur donnée du délai. Le mot ABCDE forme directement un code de délai d'arrivée du signal entre les deux oreilles. De récents travaux montrent la présence de ce type de réseaux chez les mammifères (Kandel et al. 1991).*

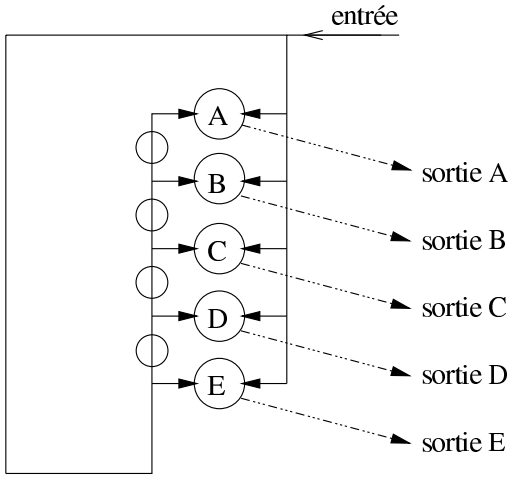


FIG. *2.12: Schéma du modèle d'autocorrélation tiré de (Licklider 1959). Ce modèle est basé sur le fait que l'information circule sur les axones des neurones à une vitesse finie et que la traversée des synapses retarde le signal. Ainsi le signal est propagé sur deux voies, dont l'une retarde le signal par passage synaptique ce qui permet le calcul de l'auto-corrélation (voir texte).*

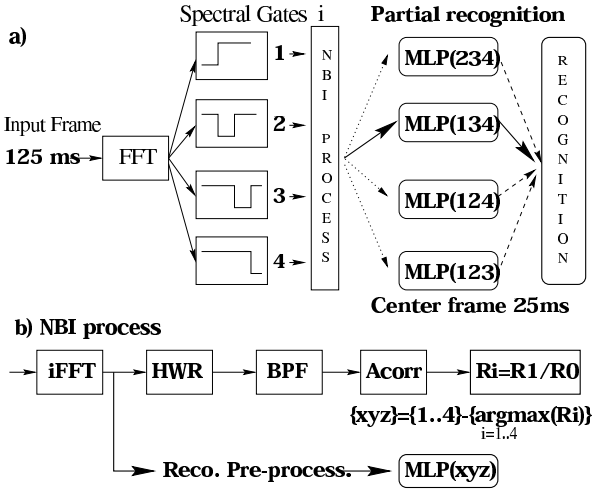


FIG. *2.13: Architecture d'un prototype CASA + reconnaissance partielle dynamique : chaque reconnaisseur est sélectionné (ici 1 parmi 4) et les autres sont exclus. C'est une décision booléenne de rejet de l'estimation des probabilités qui est appliquée à chaque trame, suivant une mesure du taux d'harmonicité effectuée en parallèle.*

Chapitre 3

Indices de fiabilité tirés de corrélogrammes

3.1 Introduction

Les méthodes non classiques d'estimation de la qualité du signal qui font partie du registre de l'Analyse de Scène Auditive et qui exploitent les propriétés du signal de parole font l'objet de recherches de plus en plus actives. Une des stratégies consiste à tirer partie du caractère fortement harmonique du signal de parole.

Récemment Gaillard (Gaillard 1999, Gaillard et al. 1997) son travail de thèse présente une méthode de marquage du plan temps fréquence basée sur les propriétés harmoniques des sons voisés : il utilise le principe d'un algorithme ancien d'extraction de pitch, l' " algorithme PPZ " (i.e. des passages par zéro), connu pour sa sensibilité a la présence d'interférence. L'analyse des passages par zéro en vue d'une extraction d'indice de confiance sur le signal est une analyse classique (Hess 1992), mais qui avait été délaissée car elle est très sensible au bruit. Gaillard montre que cette sensibilité peut être tournée en avantage pour la détection d'harmonicité en conditions interférentes : en effet, la statistique des passages par zéro fournit un indice de fiabilité permettant de classer chaque région du plan temps fréquence en deux catégories, selon qu'elle contient, ou non, une source harmonique et dominante en énergie. A partir de formalisations théoriques et de simulations, un modèle complet de marquage du plan temps fréquence est alors développe ; ce modèle est ensuite évalue en différents paradigmes d'interférences, incluant les paradigmes de doubles voyelles et de signaux bruites, puis sur des signaux à fortes variations prosodiques. Cependant cette étude plus poussée de l'utilisation des PPZ en milieu bruité ne convient que pour marquer les pavés temps fréquence des voisées, et seulement du label signal clair versus signal bruité. Il ne propose donc pas d'échelle de rapport signal sur bruit, ni de mesure sur les signaux non voisés. Le potentiel de cette mesure de fiabilité du signal est donc compromise dans une perspective probabiliste de

fusion d'experts en reconnaissance phonétique, et serait même délicate dans le cadre de la reconnaissance " missing data " tant les trames non voisées ne produisent pas d'indice fiable.

Une autre catégorie de traitement simple qui rendent compte de cette spécificité harmonique de la parole est la classe des fonctions de corrélation du signal. Elles apportent des mesures simples et efficaces des battements du signal. De plus nous avons vu dans le chapitre 3 que de tels détecteurs de coïncidence sont physiologiquement plausibles. Nous allons donc dans cette partie montrer comment extraire des indices de fiabilité du signal à partir des corrélogrammes.

A partir de représentations temps fréquence, trois type de corrélation sont extractables, elles sont représentées dans la figure 3.1 Nous explorons dans cette partie les cas (1) Autocorrélation des cellules (signal unidimensionnel) et (3) Intercorrélation des cellules dans le cas de signaux bidimensionnel.

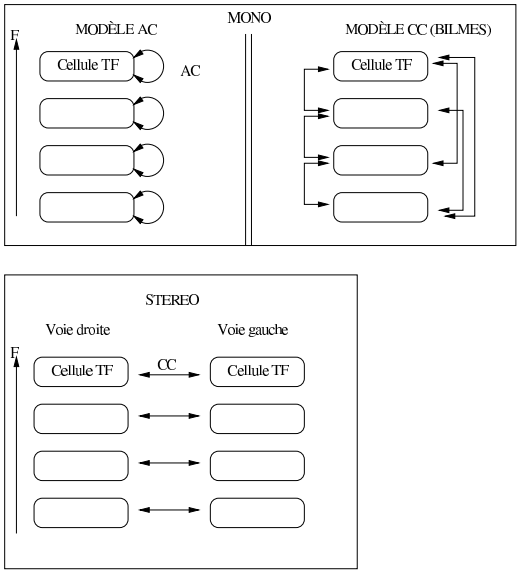


FIG. *3.1: Les analyses corrélationnelles possibles dans le cas de représentation Temps Fréquence : (1) autocorrélation des cellules (=pavés TF), (2) cross corrélation des cellules, (3) intercorrélation des cellules dans le cas de signaux bidimensionnel*

Nous montrons que le comportement de ce type de mesure en contexte temps fréquence étroit et en condition bruitée satisfait aux propriétés requises à une mesure de fiabilité du signal : peu coûteuse en temps de calcul, bonne dynamique, erreur de mesure peu dégradée à faible SNR.

3.2 Autocorrélation et mesures R0 et R1

L'AC donne la ressemblance entre le signal et la version décalée de ce même signal.

Dans la suite nous décrivons les mesures extraites de l'AC : $R0$ pic en délais nul, $R1$ second plus grand pic, Rp pic dans le domaine du pitch. Nous les relions aux grandeurs énergétiques du signal parole présentées dans la partie précédente.

En reprenant les notations de la partie précédente, il est connu (Reibhard 1996) que :

$$R0 = ET$$

En effet la fonction de AC d'un signal à énergie finie est le spectre de densité spectrale de la fft. Nous décrivons les autres mesures dans les sections suivantes.

3.2.1 Effet du bruit blanc

Dans le cas d'un bruit additif, de type gaussien, naturel ou non, la hauteur du pic R0 de délais nul de l'autocorrélation augmente. De plus les composantes de bruit gaussien blanc naturel ou non se cumulent seulement dans R0 et n'ont pas de répercussion ailleurs dans l' ACG.

L'effet du bruit blanc sur l'AC sous bande est illustré dans les figures suivantes 3.2 : sans bruit, l'AC de pavés voisés du signal est régulier. Avec du bruit non harmonique (ou harmonique sans fondamentale contenue dans le segment de mesure choisi), le second maximum mesuré est écrasé.

3.2.2 Effet des bruits périodiques

Les conclusions tirées ci dessus dans le cas de bruit additif Gaussien blanc ou le brut naturel ne sont pas généralisables dans le cas de bruits additifs périodiques ou bruits colorés.

Soit une interférence périodique de composante énergétique EN' confinée en sa fréquence centrale Fcb. Elle génère sur l'ACG des pics au délai multiples de 1/Fcb. Si leur fondamental est dans le domaine du pitch, alors la perturbation est gênante. Mais nous verrons que la démodulation suivi du filtrage passe bas (Voir annexe démodulation) permet d'atténuer les perturbation lié à ce genre de bruit tel de la parole concurrente.

3.2.3 Mesure R1 et estimation du pitch

Soit R1 le pic le plus élevé de l'autocorrélogramme après R0.

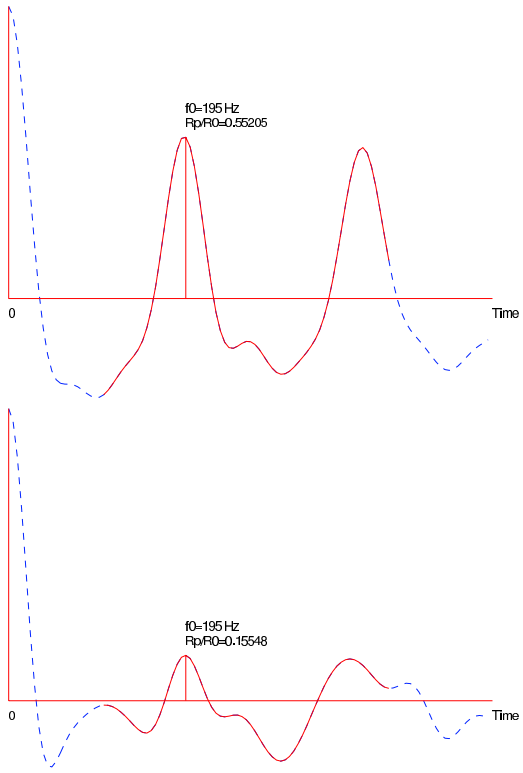


FIG. *3.2: Exemple de calcul d'autocorrélation de pavés de parole bruitée. De haut en bas, mesures sur des pavés sous-bandes de 128ms et de l'ordre de 1kHz de large, en basse fréquence, sur du signal claire puis bruité. Noter l'effet du bruit sur le rapport* $R = Rp/R0$.

Soit T1 le délai de ce pic R1 :

$R1$ = énergie cumulée de toutes les sources périodiques du signal de fréquences $= k/T1, k$ entier.

Dans le cas d'une trame non voisée, avec ou sans interférence de bruit blanc gaussien, R1 est proche de 0. Tout résidu mesuré est généré par le bruit naturel.

Par définition R1 est le pic le haut le plus proche de R0, c'est le pic de délais minimum accumulant le plus d'énergie harmonique du signal, nous avons donc a priori $1/T1 = F0$ fréquence fondamentale du signal de la trame analysée.

Cependant Hess dans (Hess 1992) note que l'autocorrélation est sensible au formant de forte amplitude ("rather sensitive to strong formants"). Le pic R1 peut se trouver décalé vers le délai nul, et le pitch peut être surestimé.

D'une façon générale la mesure de l'abscisse de $R1$ sur l'axe des délais est une technique d'estimation de pitch (Hess 1992, Rabiner 1977, Immerseel & Martens 1992) sensible à la présence de bruit qui provoque dans les meilleurs cas un doublement ou division par deux de sa valeur réelle. Il est alors intéressant de mesurer l'amplitude de $R1$ pour valider la mesure de pitch (Boersma 1993) (HNR) (Yumoto et al. 1982)

Une mesure aveugle de $R1$ second plus haut pic de l'ACG après $R0$ peut correspondre non plus à une mesure de EH mais à EN' ou une combinaison des deux. Il est donc nécessaire de se prémunir dans la mesure du possible de se genre de confusion. Toute mesure correcte de EH retombe dans le domaine des délais du pitch de la parole.

Nous voyons dans la figure 3.3 que si l'on mesure le $T1$ qui maximise R1 dans les 4 sous-bandes, on est capable de suivre de façon robuste le pitch du locuteur. Nous ne traiterons pas ce cas particulier, mais il renforce l'idée que le traitement sous-bande permet d'effectuer des mesures robustes sur le signal de parole.

3.2.4 Stabilité du domaine du pitch

Tous les locuteurs partagent le même domaine de pitch sauf cas extrêmes :
- l'intervalle du pitch varie suivant l'âge du locuteur, les bébés et les enfants en particulier.
- un chanteur virtuose.

Il est intéressant de se demander pourquoi ce domaine est aussi constant. Pour donner des hypothèses nous devons penser aux contraintes de la physiologie de notre boucle perceptuo-motrice, et aux contraintes extérieures qui sont de deux natures. La première découle du milieu aérien dans lequel se propage le signal et qui hérite des propriétés physiques du son dans ce milieu. La seconde découle des contraintes extérieures dont la plus forte a été de développer un système de communication universel et donc des traits phonétiques communs robustes aux perturbations extérieurs.

Ces traits phonétiques sont donc confinés dans un espace restreint , espace généré par des ondes glottiques dont les fréquences fondamentales (espace proximal) ont pu évoluer sous les contraintes de l'espace des traits phonétiques (espace distal) dont par exemple la fréquence du plus petit formant et la distance spectrale entre deux formants.

De plus l'homme a développé son langage depuis les temps les plus reculés dans un environnement bruité. Contraint par son appareil phonatoire et auditif, son domaine d'adaptation au bruit environnant était limité, mais nous pouvons penser que la tendance a été de dissocier le domaine du pitch à celui des fréquences des bruits environnants.
Remarquons surtout que l'homme modifie volontairement son pitch en milieu bruité

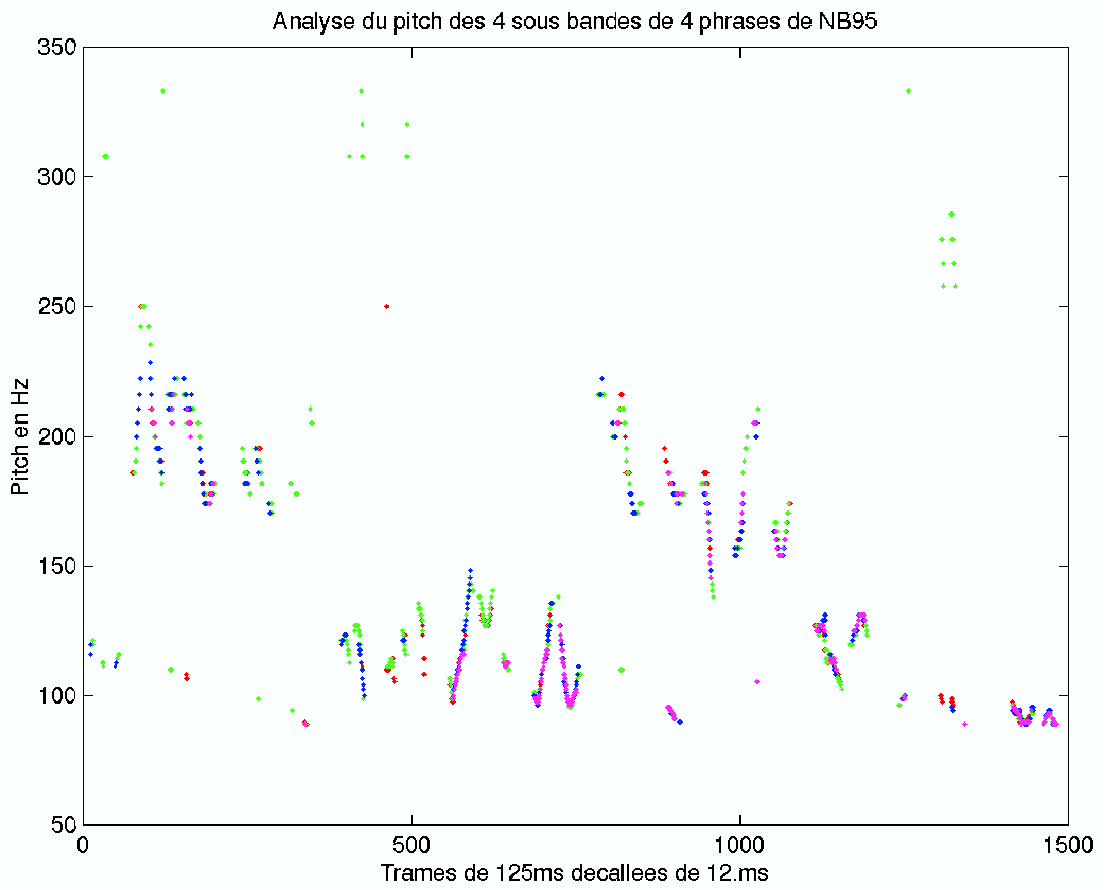


FIG. *3.3: Exemple de suivi de pitch rendu robuste par usage de R en sous bande. Figure faite sur 4 phrases de NB93, bruitée GWN 9dB.*

afin de l'écarter des perturbations. C'est une des propriétés de l'effet Lombard.

3.3 Mesure de R

La grande majorité des bruits périodiques qui nous entourent, mis à part les discussions simultanées entre plusieurs humains (cocktail party), ont une faible probabilité de posséder une Fcb qui appartienne au segment du pitch de la parole.

Toute mesure de $R1$ restreinte à un segment fixe propre au pitch permet de traquer les composantes propres au signal de parole et d'éviter de cumuler dans R1 une énergie d'un bruit périodique adverse. Nous notons cette mesure Rp.

Nous définissons :

$R = Rp/R0$

3.3.1 Détection en interférence parole : nécessité de la démodulation

Soit le détecteur défini sur R par :

H1= 'Il y a détection de parole dominante (SNR>0) si et seulement si' :

$SNR > 0$ ssi $R > K$

L'analyse ROC (voir annexe) de cet estimateur en tant que détecteur de la présence de bruit montre qu'il est performant, même dans le cas de bruit parole à $8.5dB$ si l'on démodule le signal comme le montre les courbes ROC en figure 3.4.

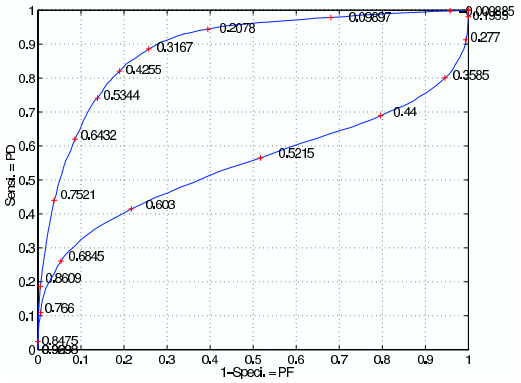


FIG. *3.4: Courbes ROC de R utilisé comme détecteur de SNR>0 dB pour 10000 trames du spectre entier et 128 ms. Calcul avec (dessus) et sans démodulation (dessous). Ces ROC ont été réalisées sur l'ensemble de développement de "Via Voice" bruité par de la parole à 8.5dB SNR. Elles montrent que le démodulateur permet de conserver une excellente qualité de détection de parole dominante même noyée parmi des paroles interférentes.*

Nous voyons là clairement l'importance de la démodulation, qui précèdera donc toujours nos mesures de R.

mod. \%	SEnsib.	SPecif.	K
SE=95%	95	60	21
SP=SE	81	81	42
SP=95%	55	95	71

TAB. *3.1: Valeur de sensibilité et spécificité pour le détecteur de parole suivant différents seuils*

3.3.2 Mesure de R sur de petits pavés temps fréquence

Nous montrons ici que le calcul de l'AC sur des pavés temps fréquence de 500 à 1kHz de large, et de 128 ms de long permet de mesurer suffisamment précisément le degré d'harmonicité des parties stationnaires du signal. De plus le comportement de R en condition bruitée satisfait aux propriétés requises à une mesure de fiabilité du signal : peu coûteuse en temps de calcul, bonne dynamique, erreur de mesure peu dégradée à faible SNR.

3.3.3 Effet du bruit sur l'AC en sous bande

Considérons maintenant que l'AC est faite après un filtrage en sous bande de fréquence centré en Fc.

Dans le cas d'un bruit additif N' coloré centré en Fc, le pic en délais 1/Fc est rehaussé. Cette propriété est utilisée dans (Kajita & Itakura 1995). En effet une autocorrélation en sous bande centrée en Fc du signal bruité par du bruit gaussien révèle le niveau du bruit par la hauteur du pic en 1/Fc. Mais cela n'est valable qu'en faisant l'hypothèse du bruit gaussien. De plus le pic en 1/Fc peut aussi cumuler l'énergie d'harmoniques.

En notant :
EH_{Fc} l'énergie des harmoniques multiples de Fc
et
EN_{Fc} l'énergie du bruit naturel, bruit coloré centré en Fc ou multiple de Fc
nous avons alors le pic en $1/Fc$:

$$R(1/Fc) = EN' + EH_{Fc} + EN_{Fc}.$$

On ne peut donc pas directement en déduire une valeur absolue de EN'. Ainsi nous montrons qu'il est nécessaire de passer par une étape statistique de calibrage pour corréler cette mesure avec le SNR.

Néanmoins, nous montrons dans les sections suivantes que les statistiques sur la distribution de R en sous bande permet d'inférer une mesure de type SNR.

L'effet du bruit sur l'AC sous bande est illustré dans la figure 3.2, et plus précisément dans la figure 3.5

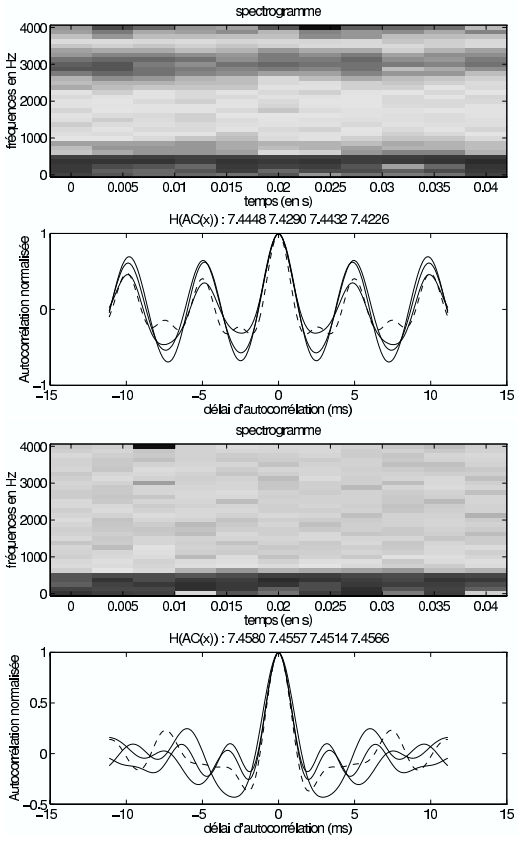


FIG. *3.5: Deux exemples avec spectrogramme de la fenêtre associée du calcul d'autocorrélation de pavés de parole bruitée en basse fréquence pour les 4 sous-bandes définies en partie I. En haut sur une fenêtre de voisée, en bas sur du silence. La réponse la plus distordue est dans les deux cas celle de la bande bruitée (BF) en pointillés. Nous donnons également la valeur de l'entropie de chaque AC qui pourrait servir de mesure (pour plus de détails voir (Glotin, Tessier, Bourlard & Berthommier 1998*a*, Glotin, Tessier, Bourlard & Berthommier 1998*b*))*

Nous illustrons dans la figure 3.6 R mesuré en sous bandes définies en partie I, sur des pavés temps fréquence de 128ms, en condition claire, ou bruité à 0 dB par du bruit en bande 1, ou par du bruit de voiture.

Nous voyons que les valeurs de R chutent en condition bruitée. Noter les différences d'amplitudes des R à travers les sous bandes. Ceci est du à la répartition des harmoniques à travers le spectre et à la distribution des formants de parole.

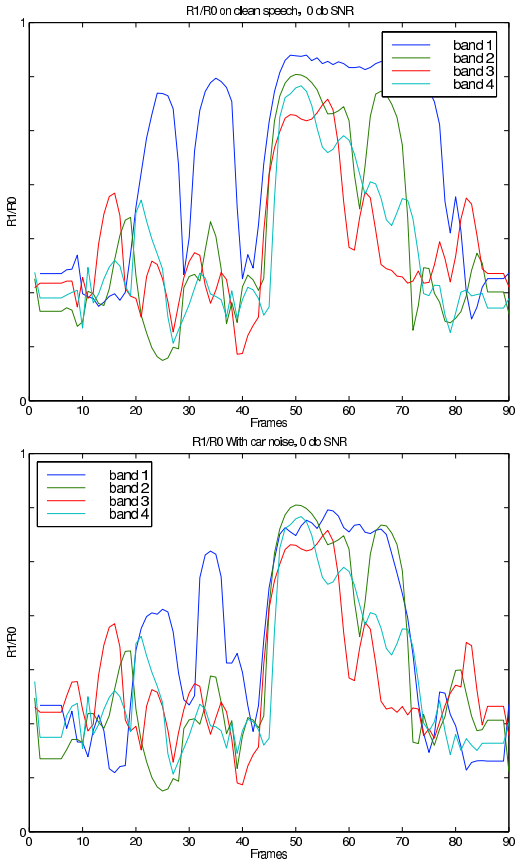


FIG. *3.6: De bas en haut pour une phrase de NB95 claire, ou de voiture, 0dB (tous les R chutent). Noter les différences d'amplitudes des R à travers les sous bandes. Ceci est du à la répartition des harmoniques à travers le spectre et à la distribution des formants de parole.*

3.3.4 Loi de combinaison de R et estimation pour flux de sous-bandes

Nous estimons ici pour les flux j composés de n sous bandes i, les $R(j)$ à partir d'une fonction des n R_i. Afin de ne pas augmenter de façon combinatoire le nombre de calcul d'autocorrélation. Il faut alors faire l'hypothèse que le pic maximum dans le domaine du pitch pour un flux ab est la somme des pics maximums dans le domaine du pitch des sous bandes le composant,

et comme l'énergie du flux ab est le cumul des énergies de ses sous bandes, on a alors :

$$R_{ab} =\simeq \frac{Rp_a+Rp_b}{R0_a+R0_b}$$

cette approximation est fiable à 95% minimum comme le montre la statistique faite sur une gamme de SNR allant de -18 à 18 dB pour 200 phrases du dev. test set de NB95. Nous illustrons cette loi dans la figure 3.7. Nous mesurons dans tous les cas, tous niveaux dB, que les estimations sont fiables à 95%.

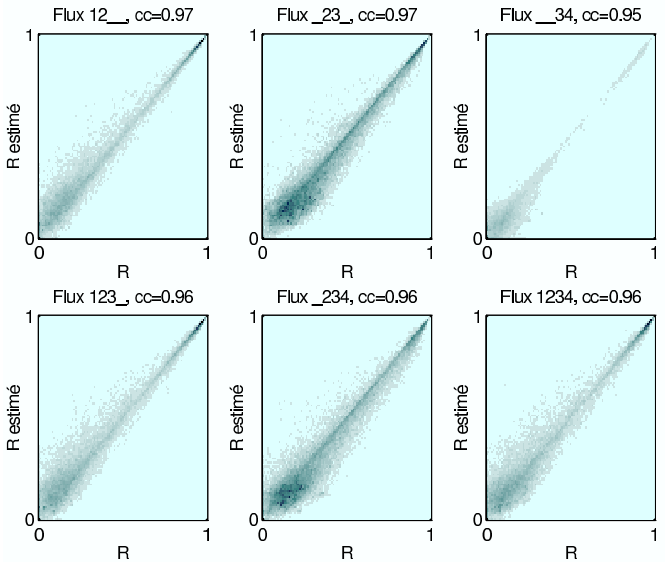


FIG. *3.7: Illustration de la loi de combinaison des 4 $R(i)$ de sous bandes pour estimation de 6 des flux j composés de sous-bandes. Distribution des $R(j)$ estimés par rapport au $R(j)$ réels, sur une gamme de SNR allant de -18 à 18 dB. La corrélation est partout supérieure à 0.95*

3.4 R comme indice de fiabilité

Plus de 70 % de la parole est composé de segments voisés. L'autocorrélogramme du signal démodulé peut donc servir comme localisation de parties voisées (d'Allessandro et al. 1998, Yumoto et al. 1982, Rabiner 1977, Boite, Bourlard, Dutoit, Hancq & Leich 2000, Hess 1983) dans un signal donné, parties assimilables à de la parole dans la plupart des cas d'interférence. Mais cette détection peut être plus finement construite en établissant la relation continue entre une mesure de fiabilité (SNR local, Maximum a posteriori) et les caractéristiques de l'autocorrélogramme.

Nous fixons les durées et décalages des analyses d'après les caractéristiques de notre reconnaisseur : 125 ms de long pour 12.5 ms de décalage.

Nous calculons alors l'indice R sur le signal démodulé après rectification simple alternance et filtrage passe bas dans le domaine du pitch ([90, 350] Hz) et nous

calculons pour chaque sous bande i l' indice $R_i = Rp/R0$, Rp étant le maximum local et $R0$ le pic en délais nul (énergie du pavé). Cette mesure est comparable à la mesure HNR : Harmonic to Noise Ratio proposée par (Yumoto et al. 1982)

Dans le cas d'analyse par groupement de 3 sous bande parmi 4, nous avons montré dans (Berthommier et al. 1998) qu'un seuillage sur les valeurs R (voir chapitre précédent) conduit à une bonne détection des sous bandes bruitées.

3.4.1 Détection de bruits colorés

Nous montrons ici que R calculé en sous bande de l'ordre de 500Hz à 1KHz peut servir comme détecteur efficace de pavé temps fréquence bruité. Nous définissons le détecteur de pavés bruités pour un instant t par :

H1= 'Le pavé p de la bande i est bruité ssi' :
$p = argmin_{i\in[1..4]}R(i)$

Ce détecteur est illustré en figure 3.8. Nous montrons que le taux de détection élevé, plus de détails sont dans (Glotin, Tessier, Bourlard & Berthommier 1998*a*, Glotin, Tessier, Bourlard & Berthommier 1998*b*).

L'enjeu est maintenant de passer à une fonction de détection continue.

Nous nous consacrons maintenant à la relation entre R et le SNR, la relation avec le MAP faisant l'objet de la dernière partie.

Nous construisons alors pour chaque pavé R_i la relation entre son R_i et son SNR local (voir figure 3.9).

Nous observons une forte corrélation non linéaire entre R et le SRN local des cellules voir figure 3.9

Nous dressons alors systématiquement les distributions des R_i suivant le SNR local pour 60 phrases de l'ensemble d'entraînement, bruitées de -21 à 39db par pas de 3 dB global (silence inclus) par pas de 6 dB avec du bruit blanc.

La distribution en figure 3.9 montre la forte corrélation entre le SNR et le R_i en bande 1, cette corrélation se retrouve dans les autres sous bandes (Berthommier & Glotin 1999) Elle sera à la base de la construction de probabilités de bruitage dans le chapitre suivant.

3.5 Intercorrélation et Indice ITD de localisation de la source dominante

Le principe de calcul de l'indice tiré de l'intercorrélation des cellules des deux voies (nous avons bien dans ce cas 2 micros) est semblable à celui tiré de l'autocorrélation. Quand le délai entre deux microphones est estimé en utilisant les signaux

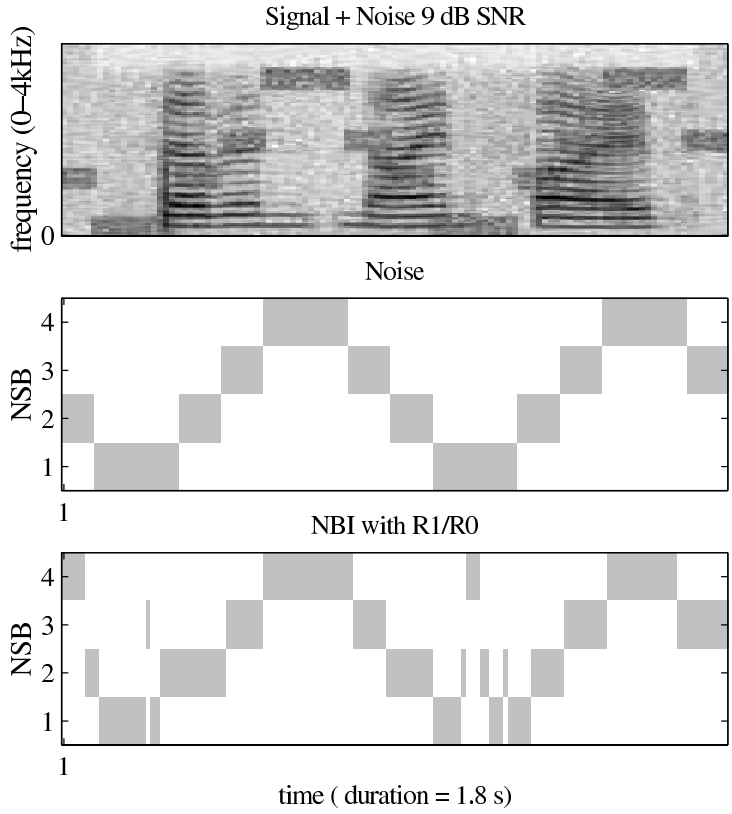


FIG. *3.8: Exemple de détection de pavés de parole bruitée tiré de nos travaux. De haut en bas, le spectre Temps Fréquence du signal bruité par un bruit synthétique de sirène - les pavés de bruits dans le signal - la détection de pavés de bruit détectés par notre module d'ASA 'Computationnelle' (" CASA "), tirant son principe de l'indice d'harmonicité des pavés sous-bandes de 128ms. Noter le fort taux de bonne détection des pavés bruités.*

enregistrés, chaque trame correspond à l'échantillonnage d'un intervalle de temps de 125ms et la représentation temps fréquence est faite en utilisant la FFT. Puis, la cross-corrélation est calculée et le TDOA (délai d'arrivée) ou "ITD" est estimé par la valeur maximale de la cross-corrélation. Le pic le plus grand a un abscisse de signe correspond à la source dominante. Par convention le signe - correspondra à la source gauche. On définit une mesure analogue au SNR mais applicable à deux sources concurrentes : le Relative Level (RL) (Glotin, Berthommier & Tessier 1999, Tessier et al. 1999)

RL est le logarithme décimal du rapport des contributions énergétiques des deux voies associées aux deux sources.

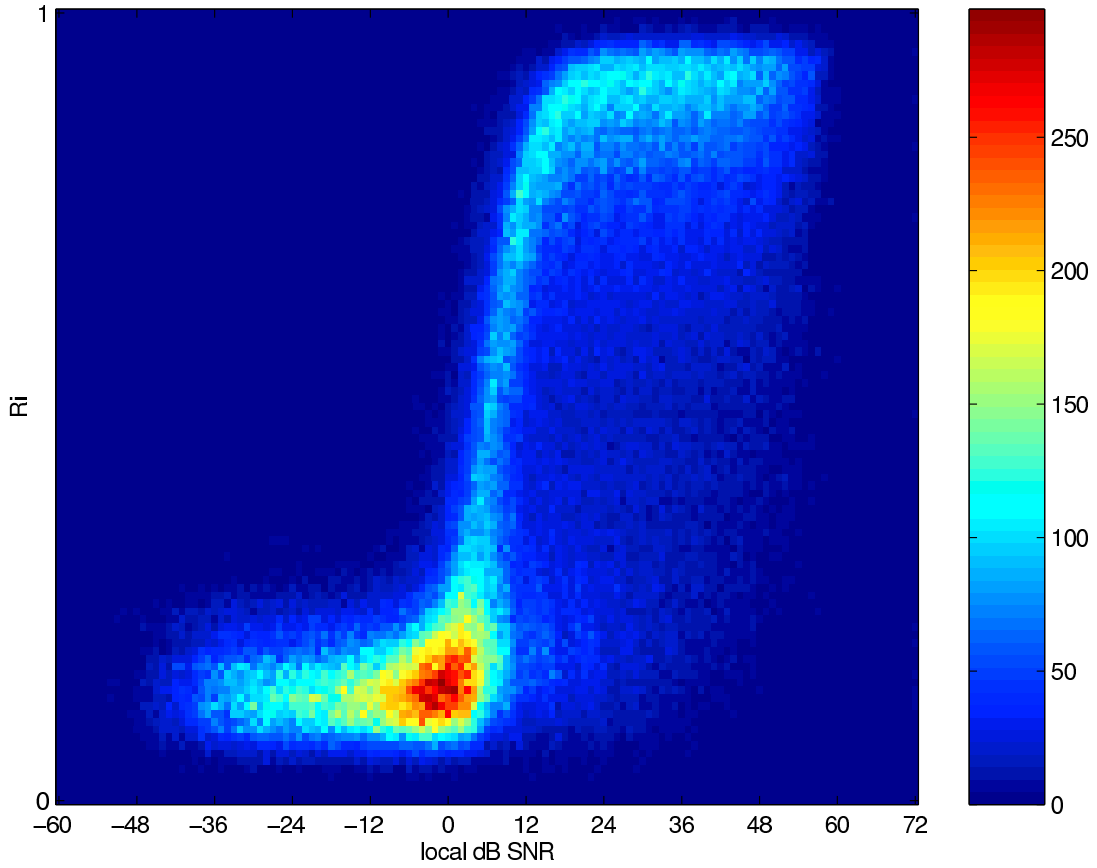


FIG. *3.9: Corrélation entre indice de voisement et SNR local : bi-histogramme de l'indice de voisement R issu de la bande 100-600Hz et 128 ms de signal bruité par du bruit blanc, et du SNR local associé montrant leur très forte corrélation non linéaire. pour 10000 trames du spectre entier et 128 m.*

On dresse sur la base STNB95 le bi-histogramme RL et ITD 3.10 On mesure encore une forte corrélation entre ITD et RL. Ceci peut servir à localiser la source ou a inférer un indice probabiliste d'appartenance d'une cellule à une source spécifique, ce que nous traiterons dans le chapitre suivant.

3.6 Conclusion

Physiologiquement plausibles, les fonctions de corrélations des cellules TF de l'ordre de 100 ms de long et 1kHz de large permettent de tirer des mesures for-

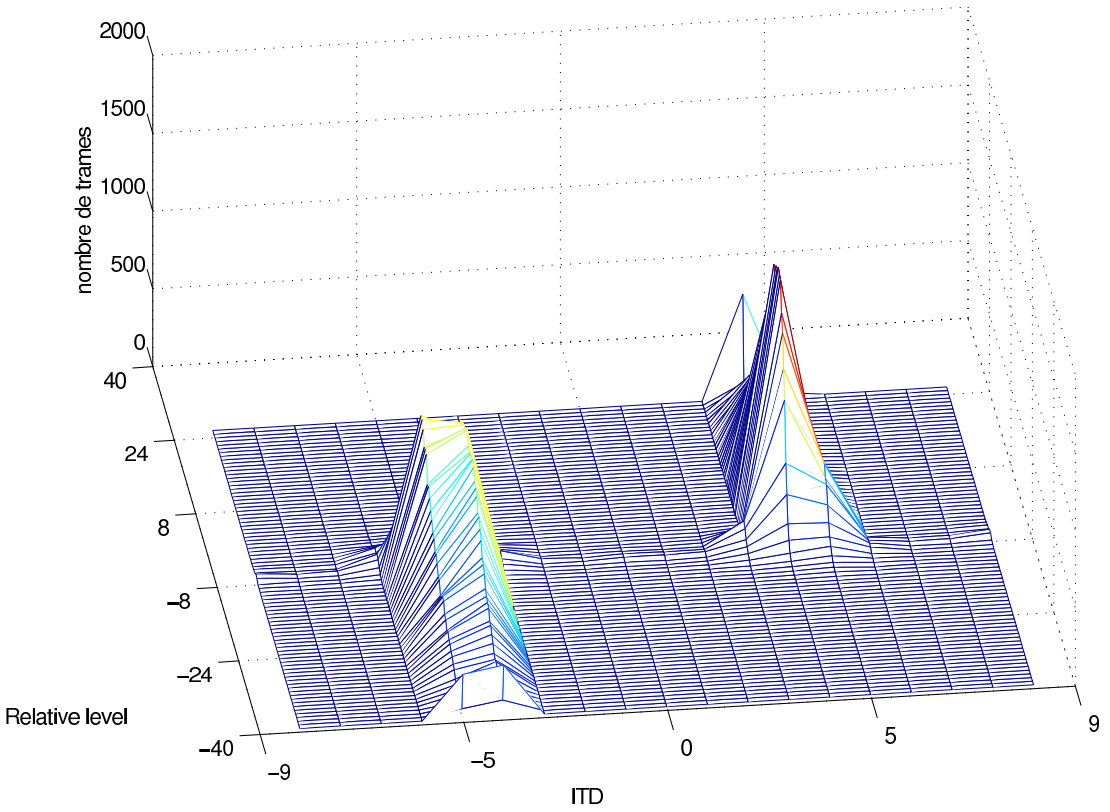


FIG. *3.10: Histogramme en bande 2 suivant les délais, sur 613 double phrases de NB95*

tement corrélées avec le SNR ou le niveau relatif de deux sources. Ceci se vérifie même dans le cas de parole concurrente, du moment qu'une étape de démodulation précède le calcul du corrélogramme. Nous allons dans le chapitre suivant présenter une technique de mesure de probabilité de bruitage des cellules temps-fréquences tirées à partir des distribution de nos mesures.

Nous disposons d'estimateurs de la fiabilité de cellules de 128 ms et 1kHz de large environ, dans le cas de sources concurrentes sur du signal stéréophonique ou bien monophonique, résistant aux interférences de type parole. Nous allons dans la partie suivante proposer différents modèles de fusion de ces mesures avec le signal ou avec les estimations phonétiques de reconnaisseur afin d'augmenter la robustesse des systèmes de reconnaissance automatique.

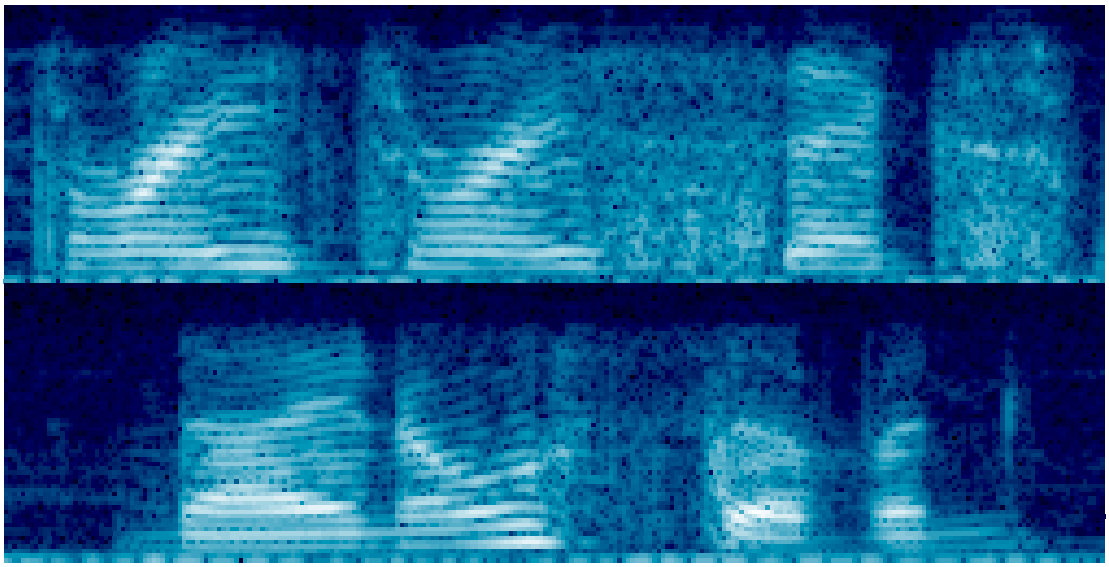


FIG. *3.11: Spectrogramme de deux phrases de Numbers95 qui seront mélangées dans STNB95 par émission simultanée par deux haut-parleurs distincts. En haut le spectre de la phrase émise dans le haut parleur gauche.*

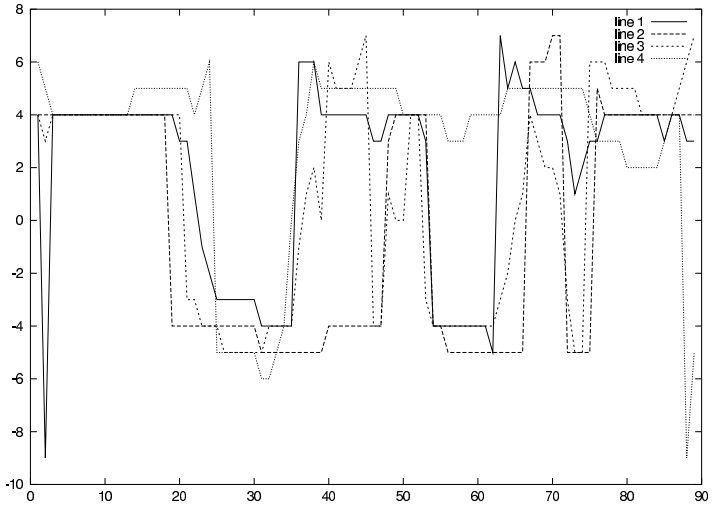


FIG. *3.12: Suivi des indices ITD pour la voie gauche (phrase du haut), noter la corrélation entre le spectre de la première phrase (émise sur le haut parleur gauche, donc dominante dans la voie gauche). Les ITD sous bandes sont corrélés mais portent une information relative à la présence locale de la source dans chaque sous bande.*

Deuxième partie

RAP robuste et l'approche multi-flux

Nous donnons dans cette partie les grandes lignes d'un reconnaisseur automatique de la parole. Nous présentons les modèles hybrides HMM/ANN. Nous donnons les méthodes classiques pour la RAP robuste. Nous exposons les grandes approches de type multi-flux. Nous présentons notre nouveau modèle de fusion dit "Full combination". Finalement nous fixons les paramètres de base de nos modèles.

Chapitre 4

Système de reconnaissance automatique de la parole et techniques classiques de robustesse

Ce chapitre présente les grandes lignes d'un système de reconnaissance, et l'état de l'art en système hybride HMM-ANN et systèmes multi-bandes. Il passe en revue les principales techniques de reconnaissance robuste.

4.1 Reconnaissance de la Parole par HMM

Les modèles statistiques sont maintenant très utilisés dans les problèmes de reconnaissance de séquences complexes telles que le signal de parole. L'introduction d'un formalisme statistique permet l'utilisation de plusieurs outils mathématiques très puissants pour déterminer les paramètres par entraînement, et pour effectuer la reconnaissance et la segmentation automatique de mots et de parole continue.

Pour la plupart de ces systèmes de reconnaissance, la parole est supposée avoir été générée selon un ensemble de distributions statistiques. Une distribution unique ne peut générer qu'un processus stationnaire. Il est donc nécessaire dans le cas de la parole de considérer plusieurs distributions, chacune modélisée par un ensemble de paramètres déterminés sur base d'un ensemble d'entraînement de façon à minimiser la probabilité d'erreur (solution bayesienne). Pendant la reconnaissance nous recherchons à travers l'espace de toutes les séquences de distributions possibles la séquence de modèles qui maximise la probabilité posteriors.

Une base de données reprenant tous les contextes acoustiques dans tous les environnements acoustiques n'existant pas, l'entraînement de ces distributions pour une reconnaissance robuste est une stratégie coûteuse, à moins de "blanchir" tous les bruits en les réduisant les analyses à des sous-bandes suffisamment étroites comme

le propose Dupont dans (Dupont 2000). Ainsi chaque reconnaisseur est entraîné sur une bande de bruit pour un certain niveau SNR. Mais cette technique reste assez lourde. Notre stratégie sera d'adapter les modèles entraînés en parole propre pour des conditions de décodage en milieu bruité.

4.2 Approche générale

Le schéma général d'un système de reconnaissance de parole est donné par la figure ci dessous :

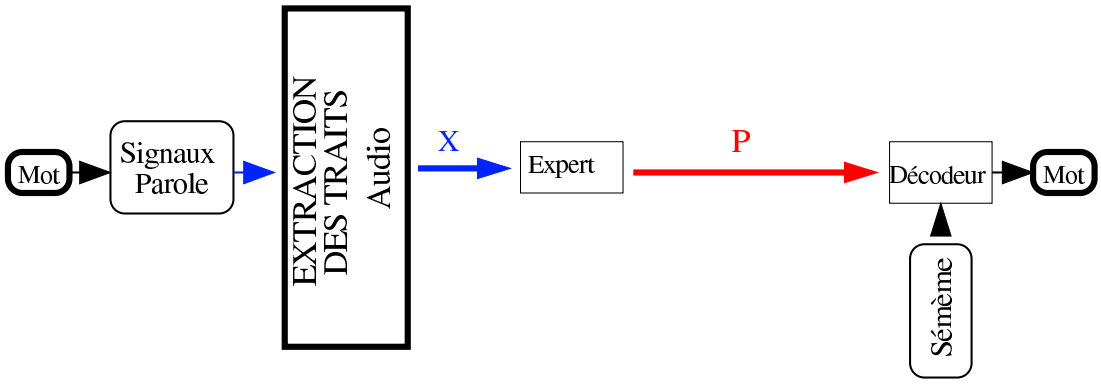


FIG. *4.1: Topologie d'un reconnaisseur de parole*

La loi de Bayes qui est à la base des système de reconnaissance HMM :

$$P(M_j|X) = \frac{p(X|M_j)P(M_j)}{p(X)} \tag{4.1}$$

où, dans notre cas, la classe M_j est le j-ième modèle statistique de phrase, avec $0 \leq j \leq J$, et X est la séquence de vecteurs acoustiques associée à cette séquence.

Selon cette loi de Bayes, la probabilité d'erreur minimum (de classer X dans la catégorie correcte M_c) est atteinte si on assigne X au modèle correspondant au maximum de probabilité posteriors $P(M_j|X)$, c'est-à-dire :

$$X \in M_k \quad \text{si} \quad k = \underset{j}{\text{argmax}}\ P(M_j|X) \tag{4.2}$$

L'entraînement des paramètres Θ devrait se faire selon cette loi de Bayes, ou critère de probabilité posteriors maximum :

$$\Theta^* = \underset{\Theta}{\text{argmax}} \prod_{j=1}^{J} P(M_j|X_j, \Theta) \tag{4.3}$$

où M_j représente le modèle HMM associé à la phrase d'entraînement X_j, et J le nombre total de phrases d'entraînement.

Chaque modèle M_j est décomposé en sous-modèles représentant des sous-unités linguistiques (phonèmes). Lorsque nous appliquons la loi de Bayes, il est donc nécessaire de faire apparaître la dépendance sur l'ensemble des paramètres Θ, ce qui donne :

$$P(M_j|X,\Theta) = \frac{p(X|M_j,\Theta)P(M_j|\Theta)}{p(X|\Theta)} \tag{4.4}$$

Dans le cas de la reconnaissance de la parole, $P(M_j|\Theta)$ dans (4.4) représente la probabilité à priori du modèle M, et reprend donc la contribution du modèle de langage.

La vraisemblance $p(X|M_j,\Theta)$ dans (4.4) représente la contribution acoustique et sera estimée par les modèles HMM. En supposant aussi que $p(X|\Theta)$ est constant et indépendant de j.

Nous discutons des modèles acoustiques estimant les valeurs $p(X|M_j,\Theta_A)$ pour tous les modèles M_j de phrases possibles. Nous ne traiterons pas le cas du modèle de langage.

4.3 Modèle acoustique HMM

Mis en place depuis plus de deux décennies, plusieurs ouvrages de références traitent de cet aspect, nous nous référons à (Boite et al. 2000).
Selon le formalisme des modèles de Markov cachés (HMM) (Rabiner 1989), le signal de parole est supposé être produit par un automate stochastique fini construit à partir d'un ensemble d'états stationnaires régis par des lois statistiques. Le formalisme des modèles HMM suppose que le signal de parole est formé d'une séquence de segments stationnaires, tous les vecteurs associés à un même segment stationnaire étant supposés avoir été générés par le même état HMM.

Chaque état de cet automate est caractérisé par une distribution de probabilité décrivant la probabilité d'observation des différents vecteurs acoustiques. Les transitions entre les états sont instantanées, caractérisées par une probabilité de transition. Chaque état du modèle permet de modéliser un segment de parole stationnaire, la séquence d'états permet quant à elle de modéliser la structure temporelle de la parole comme une succession d'états stationnaires(Boite et al. 2000). Les transitions permises par ce type de modèle sont soit des boucles sur un même état, soit le passage d'un état à l'état qui le suit directement. L'aspect séquentiel du signal de parole est ainsi modélisé. Les modèles sont dit *cachés* car la séquence d'états (par exemple, associée à une séquence de phonèmes) n'est pas directement observable; seule la séquence de vecteurs acoustiques est visible et est considérée comme une fonction statistique de la séquence d'états.

Chaque unité linguistique (par exemple, chaque phonème ou chaque mot) est donc modélisée par un ou plusieurs états stationnaires. Les mots sont ensuite construits en terme de séquences de phonèmes (à partir d'un **dictionnaire** et/ou de **règles phonologiques**) et les phrases en terme de séquences de mots (en utilisant une **syntaxe** et, éventuellement, des contraintes sémantiques). Chaque état stationnaire est représenté par les paramètres de fonctions statistiques invariables, par exemple la moyenne et la variance d'une distribution gaussienne.

Cette formulation étant admise, les trois problèmes fondamentaux en reconnaissance automatique de la parole sont les suivants :

- **Paramétrisation et estimation des probabilités** :$p(X|M, \Theta_A)$
- **Décodage (reconnaissance)** : étant donné un ensemble de modèles élémentaires M_j et une séquence d'observation X, comment déterminer la meilleure séquence de modèles élémentaires M_j de façon à maximiser la probabilité que cette séquence de modèles ait émis la séquence d'observation X ? Le problème de la reconnaissance est directement lié à celui de l'estimation des probabilités $p(X|M_j)$. Lors de la reconnaissance de parole continue, il faudra cependant y ajouter $P(M_j)$.
- **Entraînement** : étant donné un ensemble de séquences d'observations X_j et leurs modèles de Markov cachées respectifs M_j, comment estimer les paramètres des modèles de façon à maximiser la probabilité $p(X_j|M_j)$ que chaque modèle génère la séquence d'observation qui lui est associée ?

4.4 Modèle classique MG-HMM

Dans la plupart des systèmes, les vraisemblances sont estimées sur une base de **distributions gaussiennes**

$$\begin{aligned} p(x|q_k) &= N(x, \Theta_k) = N(x, \mu_k, \Sigma_k) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \end{aligned} \tag{4.5}$$

où μ_k et Σ_k représentent respectivement le vecteur moyen et la matrice de covariance associés à la classe q_k. Si les éléments diagonaux de la matrice de covariance sont représentés par σ_{ki}^2, et si l'on suppose que la matrice de covariance est diagonale (c'est-à-dire que les composantes des vecteurs caractéristiques sont supposées non corrélées), cette expression se réduit alors à :

$$p(x|q_k) = N(x, \mu_k, \Sigma_k) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right) \tag{4.6}$$

où μ_{ki} représente la i-ième composante de μ_k.

Les distributions gaussiennes sont une approximation raisonnable de beaucoup de distributions rencontrées dans les données réelles.

La gaussienne, malgré ses qualités, est une distribution uni modale, ne présentant qu'un seul maximum associé à la moyenne. Des distributions plus complexes sont alors approximées par une somme pondérée de distributions gaussiennes, appelée **distribution multi-gaussienne** et qui prend alors la forme suivante :

$$\begin{aligned} p(x|q_k) &= \sum_{j=1}^{J} p(x, j|q_k) \\ &= \sum_{j=1}^{J} P(j|q_k) p(x|j, q_k) \\ &= \sum_{j=1}^{J} c_{jk}\ N(x, \mu_j, \Sigma_j) \end{aligned} \tag{4.7}$$

où J est le nombre de gaussiennes et les paramètres $P(j|q_k) = c_{jk}$ représentent des facteurs de pondération (probabilités a priori de la gaussienne j pour la classe q_k) et sont des paramètres supplémentaires qui devront être déterminés pendant l'entraînement et qui vérifient les contraintes suivantes :

$$c_{jk} \geq 0,\ \forall j, k \tag{4.8}$$

et

$$\sum_{j=1}^{J} c_{jk} = 1,\ \forall\ k = 1, \ldots, K \tag{4.9}$$

Les multi-gaussiennes constituent un "approximateur universel" qui peut approcher n'importe quelle distribution, pour autant qu'il y ait suffisamment de gaussiennes et suffisamment de données d'entraînement.

4.5 Système hybride HMM/ANN

Une autre catégorie d'estimateurs de fonctions de vraisemblance repose sur les réseaux de neurones artificiels. Dans ce système, pour un vecteur acoustique donné, l'estimateur de sa probabilité d'appartenance à l'une des classes de sortie (*i.e.*, les états du modèle HMM) est un réseau de neurones de type perceptron multicouche (MLP) (Richard & Lippmann 1991, Bishop 1995).

Dans l'approche hybride HMM/ANN, les réseaux de neurones (MLP ou réseaux récurrents) sont donc utilisés comme estimateurs de probabilités d'émission locales. Ils permettent de calculer la probabilité (posteriors) que des vecteurs acoustiques appartiennent à différentes classes, chaque classe étant associée à un état stationnaire de l'ensemble $\Omega = \{\omega_1, ...\omega_K)$ décrivant l'ensemble des modèles HMM . Le réseau de neurones présente alors K nIJuds à sa sortie. Il a été montré que (Bourlard, Morgan

& Wellekens 1990, Bourlard & Morgan 1994, Morgan & Bourlard 1995) contrairement aux systèmes HMM standards, l'approche hybride permettait d'atteindre de bons résultats de reconnaissance en utilisant des modèles HMM de phonèmes à un seul état (avec contrainte de durée minimum) et que l'utilisation de plusieurs états n'améliore généralement pas les résultats. En fonction du problème, entre 40 et 60 classes (phonétiques) sont donc définies et représentées à la sortie du réseau.

Pour plus de détails nous renvoyons aux études récentes qui donnent des tables de comparaison entre les performances des systèmes HMM et HMM/ANN (Bengio, Mori, Flammia & Kompe 1992) (dans notre étude l'ANN est un Perceptron Multicouche (MLP)).

En pratique, il a également été observé (Boite et al. 2000) que l'utilisation de séquences de 9 vecteurs acoustiques à l'entrée du MLP conduisait souvent aux meilleurs résultats de reconnaissance. Chaque vecteur acoustique contenant typiquement les coefficients cepstraux (ou PLP) et leurs dérivées premières (et parfois secondes), ainsi que les valeurs relatives de l'énergie, ceci peut conduire à quelques centaines (300-400) unités d'entrée. Le nombre d'unités cachées peut atteindre plusieurs centaines (1000-2000), en fonction de la complexité de la tâche et de la taille de la base d'entraînement.

Cette approche, relativement simple, permet souvent d'atteindre de bons résultats de reconnaissance, souvent comparables aux systèmes standards beaucoup plus sophistiqués utilisant de nombreuses classes phonétiques et un plus grand nombre de paramètres

Aujourd'hui, ces modèles HMM/ANN représentent une alternative compétitive aux systèmes HMM standards. Conduisant à des performances comparables (Steeneken & Leeuwen 1995), cette approche HMM/ANN est aussi peu coûteuse en mémoire et en temps de calcul.

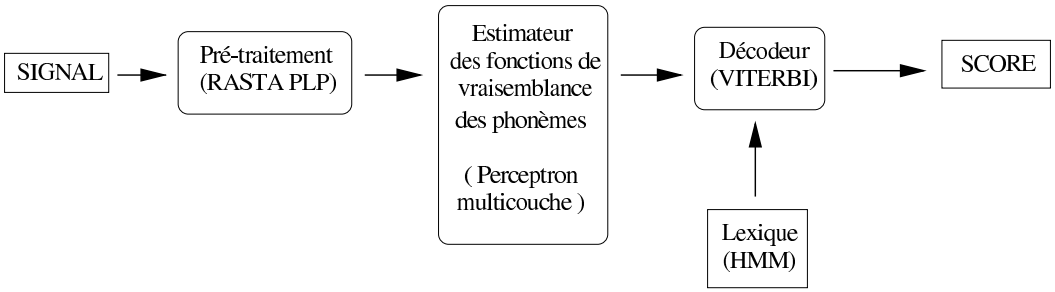
4.5.1 Estimation des fonctions de vraisemblances

Soit X une séquence de N vecteurs acoustiques $X = \{x_n\}_{n\in[1,N]}$. Chaque modèle HMM de mot, M_i, est construit à partir d'un ensemble de K classes $\Omega = \{q_k\}_{k\in[1,K]}$, une classe par phonème. Chaque topologie HMM M_i est définie comme un graphe orienté contenant K états ${q^k}_{k\in[1,K]}$ chacun associé à une classe ω_{q^k} de Ω. Il a été démontré que les perceptrons multicouches (MLP) sont de bons estimateurs de probabilités *posteriors* d'appartenance à une classe (Morgan & Bourlard 1995). Chaque MLP possède $C = K$ unités de sortie (une par classe q_k) et il est entraîné, puis testé sur les vecteurs acoustiques pour générer les probabilités *posteriors* $P(q_k|x_n,\Theta)_{\forall c\in[1,C]}$, où Θ représente l'ensemble des paramètres du MLP. Or, les probabilités *a priori* $P(q_k)$ sont connues et, d'après le théorème de Bayes nous avons :

$$\frac{P(q_k|x_n,\Theta)}{P(q_k)} = \frac{p(x_n|q_k,\Theta)}{p(x_n)}$$

avec $p(.)$ des densités de probabilité. Les $p(x_n)$ étant constants et indépendants de la classe, le terme de gauche est une grandeur proportionnelle à la vraisemblance $P(x_n|q_k, \Theta)$ (Bourlard 1996). Cette valeur sera utilisée comme probabilité d'appartenance à une classe dans l'algorithme de décodage.

En conclusion le schéma général du reconnaisseur HMM/ANN est en figure 4.2



Architecture complète du système de reconnaissance hybride ANN/HMM.

FIG. *4.2: Schéma général du reconnaisseur HMM/ANN*

4.5.2 Reconnaissance

La reconnaissance de la parole par système hybride HMM/ANN se pratique selon le même principe que la reconnaissance HMM. Le réseau de neurones est simplement utilisé comme estimateur de probabilités locales pour les modèles HMM. Après division par les estimateurs de probabilités a priori, le réseau ANN fournit donc les vraisemblances normalisées $p(x_n|q_k)/p(x_n)$ qui sont utilisées, soit dans un algorithme Viterbi, soit dans une récurrence "avant" afin d'estimer $P(M_j|X)$ pour tous les modèles M_j possibles et d'assigner la séquence X au modèle M_k conduisant au maximum de probabilité posteriors (après avoir inclu la contribution du modèle de langage).

4.6 Techniques classiques de reconnaissance robuste

4.6.1 Soustraction Spectrale

La soustraction spectrale est une technique de suppression du bruit utilisée pour diminuer les effets du bruit ajouté. Elle évalue la puissance de la parole claire en soustrayant explicitement la puissance du bruit à la puissance de la parole bruitée. Cela suppose évidemment que le bruit et le langage sont non corrélés et additifs

dans le domaine du temps. En outre, étant donné que les techniques de soustraction spectrale nécessitent une estimation du bruit pendant les pauses, on suppose que les caractéristiques du bruit changent lentement.

Cette technique simple et très utilisée est apparue dès fin des années 1970. Elle nécessite une estimation du spectre du bruit, qui est souvent obtenue par des mesures dans les zones de non parole.

Méthode

Soit un signal de parole $s(t)$ dégradé par un bruit additif, non corrélé, $n(t)$. Le signal résultant est alors

$$y(t) = s(t) + n(t) \tag{4.10}$$

Soient $Y(\omega), S(\omega)$ et $N(\omega)$ les transformées de Fourier de $y(t), s(t)$ et $n(t)$, nous avons :

$$\mid Y(\omega) \mid^2 = \mid S(\omega) \mid^2 + \mid N(\omega) \mid^2 + S(\omega) \cdot N^*(\omega) + N(\omega) \cdot S^*(\omega) \tag{4.11}$$

où $N^*(\omega)$ et $S^*(\omega)$ sont les complexes conjugués de $N(\omega)$ et $S(\omega)$.

Le but de cette méthode est d'estimer $\mid S(\omega) \mid^2$ qui est l'énergie du signal de parole de la trame considérée. Dans l'équation 4.11, $\mid Y(\omega) \mid^2$ est évidement directement mesurable sur le signal bruité. De plus les termes $\mid N(\omega) \mid^2, S(\omega) \cdot N^*(\omega)$ et $N(\omega) \cdot S^*(\omega)$ peuvent être tirés respectivement de $E[\mid N(\omega) \mid^2], E[S(\omega) \cdot N^*(\omega)]$ et $E[N(\omega) \cdot S^*(\omega)]$.
Si nous suivons l'hypothèse que le bruit $n(t)$ est décorrélé du signal de parole $s(t)$, alors nous avons :
$E[S(\omega) \cdot N^*(\omega)] = 0$
et $E[N(\omega) \cdot S^*(\omega)] = 0$
nous pouvons alors estimer $\mid S(\omega) \mid^2$.

Généralement $E[\mid N(\omega) \mid^2]$ est estimé durant les zones sans parole du signal $y(t)$, cependant une récente (Dupont, Bourlard & Ris 1997, Ris & Dupont 2001, Dupont 2000) Notre approche se situerait entre ces deux extrêmes, car nous n'avons aucune dépendance vis à vis des portions de voisement ou de non-parole, et notre méthode tire autant avantage des parties voisées des non voisées de part l'association SNR, R indifférenciée suivant les classes phonétiques, pour l'estimation de la fiabilité d'une fenêtre de 120ms (contexte phonétique large). Dans l'étude plus poussée du modèle PB (dernière partie), nous verrons que nous pourrons étudier le comportement joint de R et des catégorisations du reconnaisseur.

Il faut noter que Eq. 4.12 ne garantit pas $\mid \widehat{S}(\omega) \mid^2 > 0$. Les valeurs négatives de $E[\mid N(\omega) \mid^2]$ doivent être annulées ou mieux mises à une constante minimum. Cette non-linéarité induite dans le spectre conduit à l'incorporation de bruit dit 'musical', qui est assez perturbateur dans les reconnaisseurs. Deux méthodes sont couramment

utilisées pour y remédier : la méthode de rajout de la phase (OLA), ou simplement comme le propose Boll (Boll 1979) il est fréquent de remplacer $\mid Y(\omega) \mid^2$ dans la formule 4.12 par une moyenne sur trois trames : $\overline{\mid Y(\omega) \mid^2}$

Une autre technique utilisée (Kermorvant 1999) consiste à lisser $\mid \widehat{S}(\omega) \mid^2$ par un filtre passe bas pour filtrer les non linéarités induites par le seuillage. Le signal est initialisé par :

$\mid \bar{S}_{t=0}(\omega) \mid^2 = \mid \widehat{S}_{t=0}(\omega) \mid^2$
puis itérativement nous avons :

$$\mid \bar{S}_t(\omega) \mid^2 = \gamma * \mid \bar{S}_{t-1}(\omega) \mid^2 + (1-\gamma) * \mid \widehat{S}_t(\omega) \mid^2 \tag{4.12}$$

Il est commun de prendre alors $\gamma = 0.5$ donnant les meilleurs performances de reconnaissance (Kermorvant 1999).

4.6.2 Variante : Soustraction spectrale non linéaire

1) Berouti et *al* proposent dans (Berouti, Schwartz & Makhoul 1979) une généralisation de 4.12. Dans cette approche le spectre est élevé à la puissance a. Ils définissent alors $T(\omega)$ comme :

$$T(\omega) = \mid Y(\omega) \mid^{2a} - \alpha(SNR) * \mid \widehat{N}(\omega) \mid^{2a} \tag{4.13}$$

où $\alpha(\mid \widehat{SNR})$ est un facteur dépendant du SNR estimé :

$$\mid \widehat{SNR}(\omega) \mid^2 = \frac{\mid \widehat{S}(\omega) \mid^2}{\mid \widehat{N}(\omega) \mid^2} \tag{4.14}$$

Ce facteur α atténue l'impact du bruit musical. Si β est la valeur minimum après soustraction spectrale alors le signal propre est estimé par :

$$\mid \widehat{S}(\omega) \mid^2 = \begin{cases} T(\omega)^{1/a} & \text{si } T(\omega)^{1/a} > \beta \mid N(\omega) \mid^2 \\ \beta \mid N(\omega) \mid^2 & \text{sinon} \end{cases} \tag{4.15}$$

2) Une autre technique de débruitage a été développée par Lockwood et Boudy (Lockwood & Boudy 1991, Lockwood & Boudy 1992). L'idée est d'inverser les tendances de soustraction et de SNR : soustraire un minimum de bruit à fort SNR et inversement, soustraire un maximum de bruit à faible SNR.

Le facteur non linéaire ne dépend pas seulement du bruit estimé $E[\mid N(\omega) \mid^2]$ mais aussi du SNR estimé $\mid \widehat{S}(\omega) \mid^2$ et d'un facteur dépendant de la fréquence $\alpha(\omega)$. Dès lors Eq. 4.12 devient :

$$| \widehat{S}(\omega) |^2 = | Y(\omega) |^2 - \Phi(SNR, \alpha(\omega), E[| N(\omega) |^2]) \tag{4.16}$$

Plusieurs fonctions Φ ont été proposé dans (Lockwood & Boudy 1992), comme par exemple :

$$\Phi(\omega) = \alpha(\omega) - sigmoid(SNR(\omega))\Big(\alpha(\omega) - E[| N(\omega) |^2]\Big) \tag{4.17}$$

Cependant cette fonction nécessite de nombreux paramètres et réglages a priori.

4.6.3 Estimation du spectre de puissance du bruit

Comme nous l'avons vu la technique de débruitage par soustraction spectrale requiert une estimation du spectre du bruit $| \widehat{S}_{t=0}(\omega) |^2$.

L'estimation en dehors des temps de parole est la plus courante, généralement mise en oeuvre dans les parties de non parole. Pour déterminer les parties de non parole versus de parole, une mesure de distance est effectuée pour chaque trame entre le spectre du signal et la distribution du bruit comme le propose l'étude de C. Mokbel (Mokbel, Mauuary, Karray, Jouvet, Monne, Simonin & Bartkova 1997). Cette distance est comparée à un seuil pour décider si cette trame correspond à une période de parole ou non. Si la trame correspond à une période de non parole, les caractéristiques statistiques moyenne et variance du bruit sont mises à jour avec respectivement les facteurs α et $1 - \alpha$ dans un processus du premier ordre. (généralement $\alpha = 0.99$) ce qui correspond à une adaptation pour 100 trames (1 s de signal). L'initialisation du bruit estimé est faite sur les 10 premières trames dont on suppose qu'elle ne contiennent que du bruit.

Enfin on trouve l'estimation de "Hirsch", basée sur la distribution des énergies parole/bruit (Hirsch 1993).

4.6.4 Egalisation aveugle

L'égalisation aveugle est une technique de filtrage uniquement basée sur l'observation du signal traversant le canal. Les paramètres du filtre sont calculés via un critère d'erreur qui repose sur des statistiques connues a priori sur le signal transmis. Le filtre peut être adaptatif (Shynk 1992), comme cela été aussi proposé par Mokbel et *al.* (Mokbel, Jouvet & Monné 1996). Cette technique peut être appliquée autant dans le domaine spectral (Mauuary 1996) que cepstral (Mauuary 1998).

4.6.5 Rehaussement et transformations paramétriques

Dans l'espace paramétrique, un signal de la parole peut être représenté par un point qui se déplace en fonction des différents sons émis par le locuteur. Ainsi, une émission peut être représentée par une trajectoire de ces points à travers un domaine

choisi. La restitution du langage peut maintenant être vue comme la transformation de trajectoires bruitées et claires en une trajectoire de référence (normalisation).

Dans le domaine du rehaussement de la parole, c'est la transformations des paramètres de parole bruitée en paramètres de parole claire qui nous intéresse. Une des formes les plus simples de cartographie est la transformation linéaire, dans laquelle des exemples de parole claire et bruitée sont alignés selon l'algorithme de Dynamic Time Warping et aboutissent à une transformation utile dans les cas de bruits additifs et Lombard (Junqua 1995).
La méthode la plus simple pour trouver cette transformation est de minimiser l'erreur au carré moyenne (mean square error) entre les vecteurs paramétriques. La méthode la plus efficace est la régression linéaire qui suit une approche légèrement différente, celle d'une cartographie itérative d'une parole claire en une parole bruitée jusqu'à ce qu'un résultat satisfaisant soit obtenu. Cette méthode s'est révélée être efficace, supérieure à la cartographie spectrale et à la méthode 'd'ajustement des HMM state mean vectors' (Mokbel 1992)

Les perceptrons multicouches (multi-layer perceptrons) peuvent également être utilisés pour la transformation cartographique et ont été utilisés avec succès pour transformer les paramètres 'noisy cepstral' en leurs contreparties 'clean' Même pour des signaux qui diffèrent des données de 'training' en termes de données de la parole d'origine ou du type de bruit environnemental, le réseau neural a engendré des améliorations en robustesse du système de reconnaissance. Dans la mesure où les réseaux neuraux permettent d'obtenir arbitrairement des cartographies complexes, ils présentent de meilleurs résultats que les transformations linéaires (Mokbel 1992).

Malheureusement, alors que la performance peut être impressionnante en utilisant des ANN pour la cartographie spectrale dans la mesure où il n'y a pas de modèles paramétriques de parole ou de bruit, la performance est hautement dépendante de l'utilisation du système avec des types et des niveaux similaires de bruit. Cependant une analyse en sous-bande fine de l'ordre des bandes critiques homogénéise les variétés entre les bruits et donne alors une perspective intéressante. Un autre inconvénient de cette approche est qu'elle nécessite des versions claires de toutes les paroles bruitées, ce qui n'est pas toujours disponible dans des applications telles que la téléphonie.

4.6.6 Prétraitement PLP

Parmi les paramètres les plus couramment adoptés pour l'analyse acoustique, nous citons ici les cepstres et les paramètres PLP. L'**analyse cepstrale** est définie comme la transformée de Fourier du spectre logarithmique, et éventuellement calculés à partir d'un spectre non uniforme et espacé selon l'échelle "mel" (ou "bark") correspondant aux "bandes critiques" du système auditif (Davis & Mermelstein 1980)

(vecteurs mel-cepstraux). La motivation de cette représentation "mel" est de tenir compte de certaines propriétés de l'oreille humaine qui traite (et probablement perçoit) les sons selon une échelle de fréquence non uniforme. Les paramètres **PLP (Perceptual Linear Prediction)** (Hermansky & Junqua 1988, Hermansky 1990) sont calculés à partir d'un spectre représentant le contenu fréquentiel du signal suivant l'échelle des Bark (correspondant à l'échelle des bandes critiques du système auditif humain) et qui est ensuite lissé par un modèle autorégressif.

4.6.7 Prétraitement RASTA PLP

Pour rendre ces paramètres plus robustes aux variations linéaires de la fonction de transfert et aux variations spectrales à long-terme, il a été montré que des techniques de normalisation spectrale relativement simples pouvaient être assez efficaces. Ces techniques sont généralement appelées "égalisation spectrale", "égalisation aveugle" ou encore "déconvolution aveugle" et sont relativement efficaces si la phrase à reconnaître est suffisamment longue. Dans le cas des vecteurs cepstraux, ceci revient simplement à soustraire de chaque vecteur acoustique la moyenne de ces vecteurs calculée sur toute la phrase (ou calculée de façon adaptative). Cette méthode, appelée soustraction cepstrale, compense relativement bien les variations additives dans le domaine log-spectral (donc multiplicatives dans le domaine spectral). De même dans le cas des paramètres PLP, il a été montré qu'une méthode, appelée "RASTA-PLP" (Koehler, Morgan, Hermansky, Guenter & Tong 1994, Hermansky, Morgan, Bayya & Kohn 1992) temporel dans le domaine log-spectral avant la modélisation autorégressive permettait aussi de réduire l'effet du bruit de convolution.

Dans le cas de bruit additif (dans le domaine spectral), et en supposant que ce bruit est stationnaire, il est possible d'appliquer des méthodes de soustraction spectrale consistant à soustraire du spectre du signal bruité une estimation du spectre du bruit (obtenue en estimant le rapport signal/bruit dans chaque bande de fréquence). Malheureusement, dans la plupart des cas réels, le bruit n'est pas stationnaire et est constitué à la fois de bruit additif et de bruit de convolution.

4.6.8 Prétraitement J RASTA PLP

Cette technique (Hermansky & Morgan 1994, Hermansky, Morgan & Hirsch 1990) avantageusement les modes linéaire et logarithme à la fois, suivant un rapport J qui peut être fonction du SNR. Ce récent pré traitement est réputé pour sa robustesse aux bruits suffisamment stationnaires et constitue une référence pour la RAP robuste.

4.7 Conclusion

Nous avons décrit les mécanismes des systèmes RAP et les stratégies classiques de robustesse au bruit. Nous allons maintenant nous attacher à la présentation de l'approche multi-flux et à sa mise en oeuvre dans les chapitres suivants.

Nous avons vu dans ce chapitre que trois stratégies sont développées dans la littérature pour la mise en place de système de reconnaissance robuste.

- *Le filtrage* (Juang 1991), dont la soustraction spectrale est la technique la plus répandue (Boll 1979). L'inconvénient de cette technique est qu'elle supprime aussi une partie du spectre de la source cible.
- *L'Adaptation* du modèle de reconnaissance au bruit (Gales & Young 1992, Nadas, Nahamoo & Picheny 1989) Les effets des interférences indésirables sont apprises par les modèles. L'inconvénient est alors de ne pas pouvoir prédire tous les cas d'interférences possible.
- Enfin l'usage de *features* robustes au bruit (Ghitza 1986, Neti 1994) Dans ce type de méthode, les traits extraits du signal tirent partie de nos connaissances sur le système auditif humain et son traitement particulier du signal de parole. En particulier il s'agit d'incorporer les corrélations temporelles ou spectrales du signal (Ghitza 1986, Neti 1994).

Nous allons dans le chapitre suivant présenter de nouvelles techniques dites multi-flux.

Chapitre 5

Modèle Multi-Flux

5.1 Introduction

Les techniques présentées dans le chapitre précédent sont encore peu performantes dans le cas délicat de la réverbération, ou encore de la parole interférente.

La méthode multi-flux se veut de combler ces lacunes. Elle regroupe trois types de fusion :

- approche multi-classifieur ;
- approche multi-modale ;
- approche reconnaissance partielle ou multi-bande.

Elles sont représentées dans la figure 5.1.

5.2 Asynchronie des flux

Dans le cas de flux à information asynchrone, il est possible de construire des modèles gérant en partie les retards d'un flux sur l'autre dans la limite de la complexité du modèle. Ce type de modèle est dit "produit" et a été proposé par (Moore 1986, Varga & Moore 1990*a*) audiovisuel par (Dupont & Luettin 1998, Dupont & Luettin 2000) combinaisons des états à différents retards possibles de chaque modalité. Souvent le retard maximum est de 3 trames, et le modèle asynchrone est élagué des combinaisons d'état à plus grand retard afin d'alléger les ressources de calcul et de mémoire.

Ces modèles en figure 5.2 peuvent être appris par modalités séparés puis fusion, ou bien par un entraînement global.

5.3 Le cas multi-bande

Un cas particulier de l'approche multi-flux appliquée à la seule modalité auditive, sur un seul canal, est l'approche multi-bande.

La reconnaissance de parole Multi-Bandes est une méthode basée sur des travaux psycho-acoustiques menés par Fletcher dans les années 50, et plus récemment par J. B. Allen. Fletcher et Allen suggèrent qu'un traitement indépendant de différentes bandes fréquentielles est réalisé dans la première partie du cortex auditif, traitement suivi d'une phase pendant laquelle les résultats sont recombinés.

En dehors des motivations psycho-acoustiques originelles, l'intérêt de cette méthode réside essentiellement dans sa robustesse au bruit. En effet, un bruit n'affecte généralement qu'une zone fréquentielle limitée, et laisse donc les autres bandes intactes.

Dans ce cas chaque "flux" est une portion du spectre audio. Cette approche de reconnaissance partielle est issue pour une bonne part des recherches en "Missing Data"(Josifovski, Cooke, Green & Vizinho 1999, Cooke, Green, Josifovski & Vizinho 1999, Cooke, Green, Anderson & Abberley 1994)

et est apparue dès 1996 dans (Bourlard, Dupont & Ris 1996) puis suivie dans de

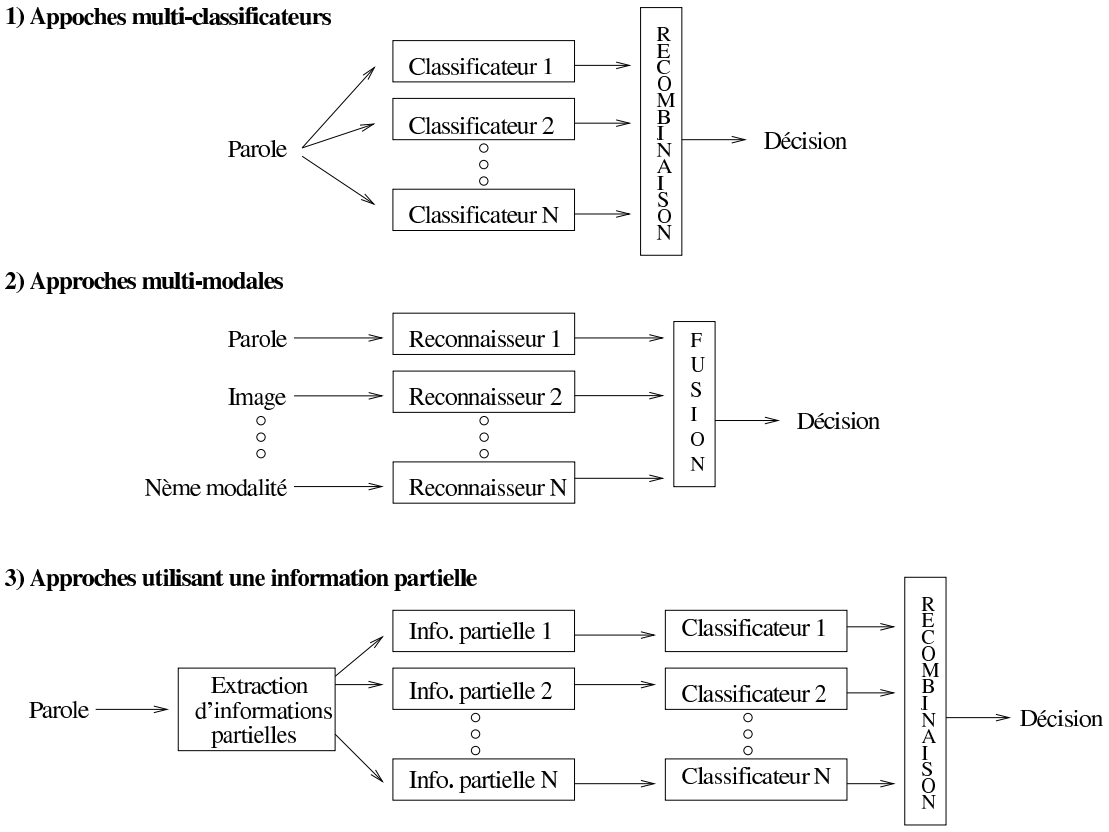


FIG. *5.1: La méthode multi-flux regroupe trois types de fusion : l'approche multi-classifieurs, l'approche multi-modale et l'approche reconnaissance partielle ou multi-bandes*

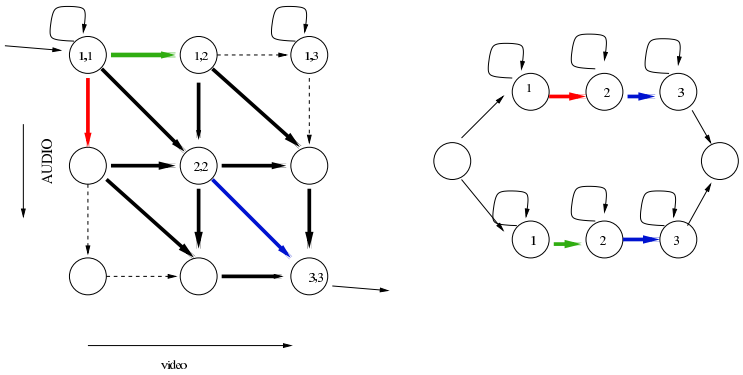


FIG. *5.2: Illustration des deux modèles, asynchrone à gauche, synchrone à droite. Dans le cas asynchrone, les états sont des états composites, combinaison des états à différent retard possible de chaque modalité. Ici le retard maximum est de 3 trames, et le modèle produit est élagué des combinaisons d'état à plus grand retard (1,3) et (3,1).*

nombreux travaux dont les notre et (Bourlard et al. 1996, Tibrewala & Hermansky 1997, Okawa, Bocchieri & Potamianos 1998, Cerisara 1999, Cerisara, Haton, Mari & Fohr 1997, Dupont 2000)

L'intérêt de cette approche réside dans le fait que les signaux de parole occupent une bande de fréquence utile qui s'étale de 100 Hz à 4 kHz et présentent une certaine redondance dans le domaine fréquentiel. Allen (Allen 1994) basé sur l'exploitation de cette redondance : entre 3 et 4 sous-bandes peuvent être traitées indépendamment avec une reconnaissance de leur contenu propre.

Dans ce cadre la méthode de reconnaissance partielle consiste à sélectionner pour la reconnaissance la meilleure combinaison de sous-bande dans le spectre, sans qu'interviennent des termes de dépendance inter flux.

5.3.1 La règle produit des erreurs

De plus, reprenant les travaux de Fletcher (Fletcher 1953,[1929]) montre que les produits des erreurs de ces sous-bandes forme l'erreur du flux recombiné.

Cette règle de l'erreur est une des motivation de l'approche sous-bande, car si le modèle de fusion respecte cette règle, alors il pourrait bénéficier d'une grande robustesse comme le montre la figure 5.3 tirée de (Morris, Hagen, Glotin & Bourlard 2001) dans le cas de deux flux. En effet la probabilité de reconnaissance correcte à 90% représente plus de 30% des cas possible.

A priori, ceci permet de mieux résister à la présence d'un bruit masquant complètement le contenu de l'une de ces bandes (*i.e.*, c'est une amputation de la représentation spectrale du signal) à condition de pouvoir sélectionner les sous-bandes dans lesquelles le signal est dominant. Les modèles classiques ne comportent pas

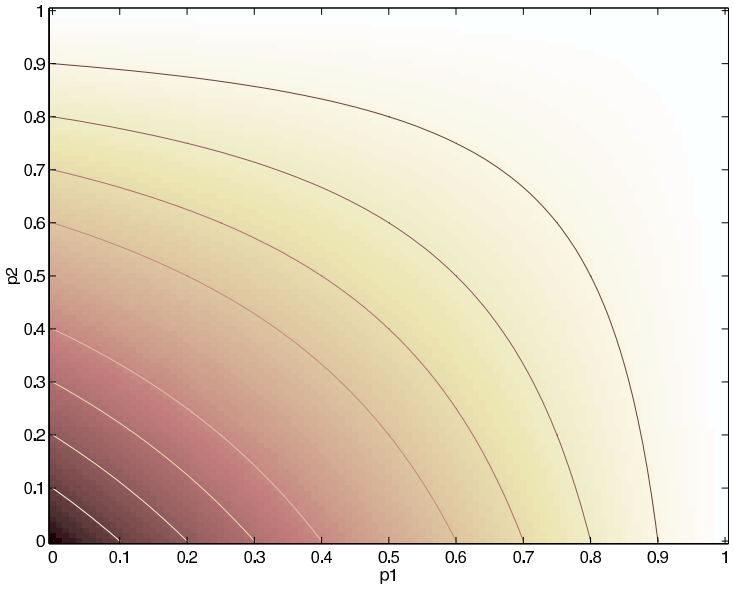


FIG. *5.3: Courbes de niveau de P(correct) suivant modèle produit des erreurs dans le cas de deux flux. P(correct) = 1-(1-p1).(1-p1) > 90% dans plus de 30% de cas.*

une telle étape, dite de segmentation primitive, au cours de laquelle on sépare les signaux à reconnaître des bruits parasites. Pour cela, nous utiliserons un système de sélection, dit AVANT, fondé sur une étape de traitement intermédiaire tenant compte du degré d'harmonicité du signal. Nous couplons ce processus avec un modèle de reconnaissance approprié, de type HMM/ANN (Morgan & Bourlard 1995). Pour comparaison, nous étudions aussi un mode de sélection dit APRES, fondé sur la sélection de la sous-bande bruitée à partir des distributions de sortie des ANN.

5.3.2 Choix des sous-bandes

Le modèle à 4 sous-bandes est intéressant vis à vis de l'espace formantique, mais inopérant face à certaines interférence trop large dans le spectre où les techniques de rehaussement classique qui restent prometteuses. Notons cependant que si le bruit est blanc, nous pouvons estimer qu'il existe tout de même toujours un formant dominant parmi les autres traits phonétiques. C'est cette marque saillante de parole dans le spectre que le reconnaisseur multi-bandes cherche intégrer indépendamment des zones où l'information est plus diffuse. La reconnaissance sous-bande agirait alors comme un rehausseur de signal pertinent : aux régions porteuses d'information sont générés des vecteurs de posteriors à plus faible entropie que dans les autres régions. Finalement la fusion globale des vecteurs de posteriors serait dominée par le vecteur d'entropie la plus faible.

5.4 Contrôle de la fusion

Le processus de fusion entre les deux modalités peut être géré par de nombreux facteurs, indépendants ou dépendants du stimulus.

En ce qui concerne les facteurs stimuli dépendants, on peut citer l'amplitude de la discordance audio visuelle représentée par les données audio visuelles incohérentes dans l'effet Mc Gurk (McGurk & MacDonald 1976).

D'une manière générale, les différents flux d'information devront être pondérés par un indice corrélé positivement avec leur fiabilité.

Le modèle FLMP de Massaro (Massaro 1987, Massaro & Stork 1998) c'est la sortie de chaque processus de décision qui gère leur fusion (Schwartz 2001). Des pondérateurs intrinsèques aux états de sortie sur chaque flux, indépendants du contexte des entrées pourraient donc suffire au contrôle de la fusion des flux.

Nous observerons ceci dans le cas de la fusion de sous-bande : les bandes qui fournissent des décisions ambiguës ne jouent aucun rôle dans la décision finale, mais cela n'est efficace comme nous le verrons que les cas extrêmes où un des experts de reconnaissance est fortement bruité. Dans le cas de fusion de flux modérément bruités, il faudra intégrer la fiabilité du signal d'entrée dans l'expert et tenir compte de l'ambiguïté interclasse à la sortie du classifieur, cela fera l'objet de la dernière partie de notre mémoire.

5.5 Les techniques de fusion Σ, Π et $\Sigma\Pi$ et leurs erreurs

Nous présentons les trois grandes classes de modèles. Nous montrons qu'elles ne sont pas toutes équivalentes vis-à-vis des erreurs générées. Un cas intéressant a été repris par (Besacier 1998) dans (Kittler, Li, Matas & Sanchez 1997) sur le comportement des modèles Σ et Π face à de petites erreurs, c'est à dire au cas où la reconnaissance est faite en condition suffisamment claire.
Nous étendons ces travaux au cas du modèle $\Sigma\Pi$ et montrons que ce modèle se comporte de façon équivalente que le meilleur des deux précédents suivant le critère de l'erreur du modèle résultante de faibles fluctuations des estimées de chaque flux.

Soit e_{jk} l'erreur commise sur l'estimation de $p(q_k|x_j)$:

$$\hat{p}(q_k|x_j) = p(q_k|x_j) + e_{kj} \tag{5.1}$$

5.5.1 Le modèle Σ

Pour le modèle Σ, $p(q_k|x)$ se calcule par : $s_j = p(q_k|x) = \sum_{j=1}^{n} \alpha_j p(q_k|x_j)$

Le modèle 'Σ' est un modèle dit indulgent car $\forall x \in [0,1]\Sigma(x,0) = x$.

De plus :

$$\begin{aligned}
\hat{s}_j &= \sum_{j=1}^{n} \alpha_j \hat{p}(q_k|x_j) && (5.2)\\
&= \sum_{j=1}^{n} \alpha_j (p(q_k|x_j) + e_{kj}) && (5.3)\\
&= [\sum_{j=1}^{n} \alpha_j p(q_k|x_j)] [1 + \frac{\sum_{j=1}^{n} e_{kj}}{\sum_{j=1}^{n} p(q_k|x_j)}] && (5.4)\\
&= s_j E_\Sigma \text{ où } E_\Sigma = 1 + \frac{\sum_{j=1}^{n} e_{kj}}{\sum_{j=1}^{n} p(q_k|x_j)} && (5.5)
\end{aligned}$$

5.5.2 Le modèle Π

Pour le modèle Π, $p(q_k|x)$ se calcule par : $s_j = p(q_k|x) = P(q_k)^{1-n} \prod_{j=1}^{n} p(q_k|x_j)$

Le modèle 'Π' est un modèle dit sévère car $\forall x \in [0,1] \Pi(x,1) = x$.
De plus avec un développement du 1er ordre :

$$\begin{aligned}
\hat{s}_j &= P(q_k)^{1-n} \prod_{j=1}^{n} [p(q_k|x_j) + e_{kj}] && (5.6)\\
&= P(q_k)^{1-n} \prod_{j=1}^{n} [p(q_k|x_j)(1 + \frac{e_{kj}}{p(q_k|x_j)})] && (5.7)\\
&\simeq P(q_k)^{1-n} [\prod_{j=1}^{n} p(q_k|x_j)] [1 + \sum_{j=1}^{n} \frac{e_{kj}}{p(q_k|x_j)}] \text{ pour } e_{kj} << p(q_k|x_j) && (5.8)\\
&\simeq s_j E_\Pi \text{ où } E_\Pi = 1 + \sum_{j=1}^{n} \frac{e_{kj}}{p(q_k|x_j)} && (5.9)
\end{aligned}$$

5.5.3 Le modèle $\Sigma\Pi$

Pour le modèle $\Sigma\Pi$, $p(q_k|x)$ se calcule par : $s_j = p(q_k|x) = \sum_{j=1}^{n} \alpha_j \prod_{i \in f_j} p(q_k|x_i)$

Le modèle '$\Sigma\Pi$' est un modèle dit compromis car
$\forall x \in [0,1] \Sigma\Pi(x,0) = x$ est faux
et
$\forall x \in [0,1] \Sigma\Pi(x,1) = x$ est faux

'$\Sigma\Pi$' est donc à mi-chemin de la dureté du mode de décision par rapport aux modèle Π et Σ. Nous allons montrer de plus que son erreur finale est équivalente à l'erreur finale du modèle Σ.

Pour ce modèle le développement du 1er ordre nous donne :

$$\hat{s}_j = \sum_{j=1}^{n} \alpha_j \prod_{i \in f_j} (p(q_k|x_i) + e_{ki}) \tag{5.10}$$

$$\simeq \sum_{j=1}^{n} \alpha_j \Big(\prod_{i \in f_j} p(q_k|x_i) \Big) \Big(1 + \sum_{i \in f_j} \frac{e_{ki}}{p(q_k|x_i)}\Big) \text{ pour } e_{ki} << p(q_k|x_i) \tag{5.11}$$

$$\simeq s_j E_{\Sigma\Pi} \text{ où } E_{\Sigma\Pi} = 1 + \frac{\sum_{j=1}^{n} \alpha_j \big[\prod_{i \in f_j} p(q_k|x_i)\big] \big[\sum_{i \in f_j} \frac{e_{ki}}{p(q_k|x_i)}\big]}{\sum_{j=1}^{n} \big[\alpha_j \prod_{i \in f_j} p(q_k|x_i)\big]} \tag{5.12}$$

5.6 Comparaison des facteurs d'erreur

Type de modèle	Expression de $p(q_k\|x)$	Expression du facteur d'erreur
Modèle Σ	$\sum_{j=1}^{n} \alpha_j p(q_k\|x_j)$	$E_\Sigma = 1 + \frac{\sum_{j=1}^{n} e_{kj}}{\sum_{j=1}^{n} p(q_k\|x_j)}$
Modèle Π	$P(q_k)^{1-n} \prod_{j=1}^{n} p(q_k\|x_j)$	$E_\Pi = 1 + \sum_{j=1}^{n} \frac{e_{kj}}{p(q_k\|x_j)}$
Modèle $\Sigma\Pi$	$\sum_{j=1}^{n} \alpha_j \prod_{i \in f_j} p(q_k\|x_i)$	$E_{\Sigma\Pi} = 1 + \frac{\sum_{j=1}^{n} \alpha_j \big[\prod_{i \in f_j} p(q_k\|x_i)\big] \big[\sum_{i \in f_j} \frac{e_{ki}}{p(q_k\|x_i)}\big]}{\sum_{j=1}^{n} \big[\alpha_j \prod_{i \in f_j} p(q_k\|x_i)\big]}$

On montre que pour tout $p(q_k|x_j) > 0$, $\frac{\sum_{j=1}^{n} e_{kj}}{\sum_{j=1}^{n} p(q_k|x_j)} \leq \sum_{j=1}^{n} \frac{e_{kj}}{p(q_k|x_j)}$,
donc $|E_\Sigma - 1| \leq |E_\Pi - 1|$

Le calcul montre que $E_{\Sigma\Pi} < E_\Pi$. De plus par simulation on trouve que $E_{\Sigma\Pi}$ est équivalente à E_Σ.

On a donc $E_{\Sigma\Pi} \simeq E_\Sigma < E_\Pi$.

Ces deux résultats sont en faveur du modèle $\Sigma\Pi$, modèle à décision compromise et faible fluctuation face aux petites erreurs.

On trouve depuis quelques années déjà différents modèles de type Σ ou Π. Les pionniers furent certainement Mirghafori, Dupont et Cerisara qui développèrent simultanément à nos travaux les premières applications de reconnaissance multi-bandes.

Le modèle Σ le plus courrant est la moyenne géométrique des posteriors. Moyenne géométrique, courante également est de type Π. Le vote par majorité quand à lui (décision dure) n'est pas très performant (Cerisara 1999). Nous avons comparé (Morris et al. 2001) ces modèles avec une instance du modèle $E_{\Sigma\Pi}$ que nous présentons dans

la section suivante et qui s'avère être souvent plus robuste (Morris et al. 2001) que les modèle Σ ou Π.

5.7 Le modèle "Full Combination" (FC)

Nous avons mis en lumière dans le chapitre précèdent l'avantage de travailler sur des fusions de groupes de sous-bandes. Nous proposons donc un nouveau modèle de fusion qui répond en divers points aux exigences d'un bon opérateur de fusion tout en conservant un bon compromis performance de WER et faible coût de calcul.

Nous formalisons dans ce chapitre un modèle probabiliste "Full Combination" (FC), de fusion des flux en introduisant une variable latente qui indique, à chaque trame, la combinaison de sous-bandes la plus adéquate à la reconnaissance. Mais ce modèle 'Full Combination' nécessite autant de sous-reconnaisseurs que de combinaisons de sous-bandes. Nous en développons alors une approximation (FCA) qui requiert seulement autant de reconnaisseurs que de sous-bandes plus le modèle pleine bande (FB) et une hypothèse faible d'indépendance. Testées sur des bruits colorés, avec des poids égaux, ces 2 méthodes utilisant des JRASTA sont plus robustes que le système spectre entier de référence JRASTA.

La RAP multi-bandes exploite la redondance spectrale dans le but d'augmenter sa robustesse à l'inadéquation des données tout en faisant un minimum d'hypothèses sur le bruit interférant (Dupont 2000, Cerisara 1999) Nous verrons que les experts des bandes non bruitées fournissent suffisamment d'information pour permettre un décodage robuste. Dans ce chapitre nous développons le modèle *Full Combination* (FC) qui s'inscrit dans le paradigme multi-bandes de la RAP et dont les relations avec la perception humaine de la parole sont reprises dans (Morris et al. 2001). Des expériences précédentes en RAP ont montré que le traitement indépendant des sous-bandes peut faire chuter les performances en parole claire. Une alternative consiste à travailler sur les estimations phonétiques des 2^d combinaisons (ou flux) des d sous-bandes, (y compris le flux vide correspondant aux probabilités *a priori*). Dans une première approche ces estimations ont été calculées pour chaque flux puis sélectionnées une à une (Hermansky, Tibrewala & Pavel 1996) estimations sont pondérées et sommées. Comme l'entraînement d'un expert par combinaison, soient 2^d experts pour d sous-bandes, est rapidement irréalisable, il est préférable de travailler avec des approximations de ces combinaisons. Nous montrons en fin de chapitre suivant qu'il est possible d'obtenir des résultats similaires ou meilleurs au modèle FC avec son approximation.

5.7.1 L'approche «Full Combination»

Les systèmes multi-bandes pour la RAP décomposent le domaine spectral en plusieurs sous-bandes, qui sont traitées indépendamment, et dont les paramètres

caractéristiques x sont passés aux reconnaisseurs correspondants. Les probabilités *posteriors* $P(q_k|x)$ des sous-bandes sont combinées dans le processus de reconnaissance. Dans notre approche les 2^d flux des combinaisons des d sous-bandes sont intégrés suivant les évènements j_{propre} collectivement exhaustifs et mutuellement exclusifs :

"la j^{ime} combinaison de sous-bande est le flux qui produit la meilleure reconnaissance parmi tous les flux possibles ".

Considérant que les données bruitées hors du j^{ime} flux propre sont négligeables dans l'estimation des probabilités *posteriors* (Lippmann & Carlson 1997, Morris et al. 2001), nous posons $P(q_k|x, j_{propre}) \simeq P(q_k|x_j)$, où x_j est le vecteur acoustique du j^{ime} flux. Nous avons alors suivant la loi des probabilités totales :

$$P(q_k|x) \simeq \sum_{j=1}^{2^d} P(j_{propre}|x) P(q_k|x_j) \tag{5.13}$$

Les probabilités dénotant les données claires $P(j_{propre}|x)$ dans (5.13) seront estimées dans les chapitres suivants de différentes façons comme cela est démontré dans (Berthommier & Glotin 1999, Hagen, Morris & Bourlard 1999, Morris et al. 2001) expériences présentées dans ce chapitre, les $P(j_{propre}|x)$ sont équiprobables, ce qui est déjà fort intéressant en terme de robustesse comme nous allons le montrer.

Dans l'approche FC les termes $P(q_k|x_j)$ sont donnés en sortie du réseau de neurone qui est entraîné et testé sur les paramètres acoustiques x_j. Dans l'approche FCA les termes $P(q_k|x_j)$ sont estimés à partir des sorties des ANN relatifs uniquement aux sous-bandes contenues dans la j^{ime} combinaison.

5.7.2 L'Approximation du FC : le FCA

Pour éviter l'entraînement de 2^d ANNs on peut estimer les probabilités $P(q_k|x_j)$ des combinaisons en utilisant uniquement les probabilités $P(q_k|x_i)$ issues des observations $x_{i,i\in\{1..d\}}$ des d sous-bandes qui composent cette combinaison (on notera J cet ensemble de sous-bandes, de cardinal $|J|$). Ce modèle ne requiert qu'une hypothèse d'indépendance des observations des sous-bandes conditionnellement à chaque classe phonétique, hypothèse plus faible que l'indépendance absolue (Morris et al. 2001). Nous avons alors $P(x_j|q_k) \simeq \prod_{i\in J} P(x_i|q_k)$, donc

$$P(q_k|x_j)\frac{p(x_j)}{p(q_k)} \simeq \prod_{i\in J} P(q_k|x_i)\frac{p(x_i)}{p(q_k)} \tag{5.14}$$

$$P(q_k|x_j) \simeq \frac{\prod_{i\in J} P(q_k|x_i)}{p^{|J|-1}(q_k)} \cdot \Theta \tag{5.15}$$

avec $\Theta = \frac{\prod_{i \in J} p(x_i)}{p(x_j)}$, qui disparaît par normalisation sur toutes les classes phonétiques pour obtenir des estimations telles que : $\sum_k P(q_k|x_j) = 1$.

Le calcul de l'approximation des probabilités $P(q_k|x_j)$ pour les 2^d combinaisons j peut se faire efficacement avec une procédure récursive qui réutilise les multiplications des composantes partagées par plusieurs combinaisons (Hagen & Glotin 2000).

5.8 Conclusion

Tout au long de ce mémoire nous développerons et comparerons les techniques présentées de cette partie avec de nouvelles architectures propices à l'intégration des indices développés dans la première partie.

On trouve dans la bibliographie des développements des différents modèles présentés dans ce chapitre dans les thèses de Cerisara, Dupont, Mirghafori. La tendance actuelle tend vers le modèle $\Sigma\Pi$. En effet le modèle Π pose souvent l'hypothèse forte de l'indépendance des sous-bandes, et nous avons vu comment le modèle FCA relaxe dans une certaine mesure cette contrainte.

Les modèles multi-flux seront traités à travers différents types de fusions : précoce ou tardive, et suivant les informations fusionnées : audio seule ou audio visuelle. Les α_j seront des mesures de corrélation spatio-temporelles, issues des auto- ou inter-corrélations (dans le cas de doubles voies) des cellules temps-fréquences.

La validité du modèle FC actuel sera discuté à travers les applications de la partie III. Puis nous analyserons les erreurs du FC en IV pour proposer un modèle tenant compte de ma qualité de la transmission de chaque phonème à travers chaque flux.

Chapitre 6

Mise en place des reconnaisseurs et scores de référence de l'approche HMM/ANN

6.1 Etude de l'effet des priors sur les systèmes Hybrides ANN-HMM

Il est évident que dans le cadre de ces approximations de flux par produit normalisé des sous bandes, les priors jouent un rôle important dans la précision de l'approximation. En effet un biais sur les priors peut induire une forte erreur sur les posteriors dans le cas du pleine bande ou du FCA. Nous avons donc veillé à ce que les occurrences des phonèmes de notre base de donnée soient suffisamment nombreuses pour générer des priors fiables. En fait, les phonèmes qui n'apparaissaient que quelque dizaines de fois dans les trames de l'ensemble d'entraînement de Numbers95 ont été simplement supprimés par l'équipe de l'ICSI, et remplacés par leur(s) voisin(s) droit et ou gauche. La base Numbers95 qui possède alors 27 phonèmes est donc déjà propre a cet égard. Par contre la base Numbers93 n'a pas été travaillée en ce sens et conserve des classes phonétiques représentées seulement par quelques unités ou dizaines de trames, ce qui constitue un ensemble total de 33 phonèmes, dont 6 non signifiants qui biaisent le vecteur de priors par des valeurs infinitésimales qui ne sont pas représentatives.

Dans cette section nous nous intéressons à la mise en place d' un modèle ANN HMM et des paramètres dont il dépend, en particulier, nous soulevons un problème délicat qui est le passage des posteriors aux fonctions de vraisemblances via les priors. Nous validons aussi dans ce chapitre un modèle dit FC ou FCA qui est décrit plus tard dans la thèse.

On suppose dans le modèle ANN/HMM que chaque classe q_k est caractérisée par une densité de probabilité $p(x|q_k)$ différente et qui peut être estimée séparément

et indépendamment des autres classes (ce qui n'est pas le cas des probabilités a posteriori $P(q_k|x)$ qui dépendent des paramètres de toutes les classes). Cette densité de probabilité $p(x|q_k)$ appelée "vraisemblance" peut s'exprimer en fonction de probabilités a posteriori par la loi de Bayes :

$$p(x|q_k) = \frac{P(q_k|x)p(x)}{P(q_k)} \tag{6.1}$$

où K représente le nombre de classes, le vecteur d'observation est x et la probabilité priors $P(q_k)$. Ce calcul des vraisemblance est théoriquement parfait, mais en pratique il requière une connaissance exacte des priors. Nous montrons dans cette section que cette approche n'est pas sans inconvénients car :

- Des hypothèses sont toujours requises concernant la forme du modèle paramétrique (des fonctions de vraisemblance) de $p(x|q_k)$.
- Les probabilités priors sont généralement très difficiles à estimer fidèlement.
- L'estimation des paramètres des modèles maximisant la fonction de vraisemblance n'optimise pas les propriétés discriminantes des modèles (Bourlard 1998). On peut montrer que la maximisation de la fonction de vraisemblance ne minimise le taux d'erreur que dans le cas où le modèle est correct (ce qui n'est généralement pas le cas) et où l'on possède un nombre suffisant (voire infini) de données d'entraînement.
  Ce problème est dû au fait que pendant l'entraînement des MLP ce sont les probabilités à posteriori qui sont optimisées et non les fonctions de vraisemblances contrairement aux modèles multigaussiens.

Cette étude montre que certains optimums sont atteints pour des valeurs et des usages variables des priors (en effet les priors sont difficilement estimables, mais on peut aussi choisir d'utiliser directement les posteriors pour la reconnaissance, ce qui revient à considérer les priors comme uniformes en dernière étape d'estimation des vraisemblances).

Les priors seront corrigés par un facteur de fiabilité sur les priors a tel que :
$priors' = a * priors + \frac{(1-a)}{K}$ avec K le nombre de classes phonétiques.

Dans cette expérience nous évaluons la sensibilité du reconnaisseur pleine bande et du FCA aux priors. L'intérêt est de déterminer si un biais sur les priors génère des erreurs importantes dans nos modèles. Inversement, si on estime qu' il y a un biais sur les priors, il serait alors utile de le corriger.

Les résultats ci dessous donnent les moyennes des WER "Jrasta pleine bande" pour les 11 bruits de notre base sur 200 phrases chacun (total 2000 phrases), du reconnaisseur pleine bande, et du FCA, RAP basée sur les posteriors ou les vraisemblances, avec la correction sur les priors. Pour le reconnaisseur pleine bande, la reconnaissance sur les posteriors est évidemment identique quels que soient les priors. Par contre, comme le FCA intègre ces priors, l'approche basée sur les posteriors varie tout de même en fonction des priors.

full band 1234 Jrasta	posteriors	vrais. a=1	vrais. a=0.5
parole propre	8.0	8.0	8.1
12 dB * 11 bruits	12.92	13.17	12.99
0 dB * 11 bruits	31.74	31.55	30.91
moyennes	17.55	17.57	17.33

TAB. *6.1: Reconnaissance sous posteriors versus vraisemblances, pour le reconnaisseur pleine bande, 200 phrases de l'ensemble de développement ("dev set").*

FCA(a), Jrasta	poster. a=1	poster. a=0.5	vrais. a=1	vrais. a=0.5
parole propre	12.5	11.5	11.2	11.9
12 dB * 11 bruits	17.75	15.8	14.85	16.9
0 dB * 11 bruits	38.59	35.9	34.43	27.44
moyennes	22.95	21.07	20.16	18.75

TAB. *6.2: Reconnaissance sous posteriors versus vraisemblances, pour le FCA, 200 phrases du dev.set.*

Nous constatons dans la table des performances du système de base plein spectre 6.1 que la reconnaissance n'est pas très affectée par les changements en vraisemblance ou non, et par les valeurs des priors. Dans le cas du modèle 'Full Combination' exposé dans la suite du mémoire, nous voyons que l'effet est plus marqué au tableau 6.2. Nous voyons que les meilleures performances sont données pour la reconnaissance effectuée par les 'scaled' vraisemblances.

Nous effectuerons tous nos calculs dans ce cadre. Cependant le résultat moyen diffère suivant les valeurs des priors. La validité des priors est donc remise en cause. Nous allons traiter de ce problème dans la section suivante.

6.2 Validation des priors

6.2.1 Etude fine des effets des priors dans le modèle FCA

Nous étudions ici l'influence des priors $P(q_k)$ sur nos modèles hybrides. En effet comme nous l'avons vu, il est délicat de les estimer de façon fiable, et pourtant elles interviennent à deux niveaux fondamentaux dans notre approche :
- au niveau des approximations de l'approche des flux combinés (FCA) qui utilisent les priors pour réduire les hypothèses d'indépendance des sous bandes en conditionnant par classe phonétique :
- au niveau du calcul des vraisemblances (voir figure 6.1).

]

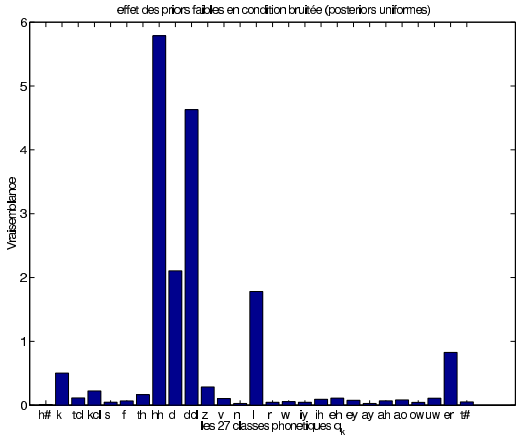


FIG. *6.1: Effet des priors faibles sur une distribution plate de posteriors : la loi de bayes, en considérant p(x) constant nous donne ces vraisemblances sur un vecteur de posteriors uniforme comme il est fréquent d'en mesurer en condition bruitée. Y a t-il surestimation des vraisemblances des classes /hh/ /d/ /dcl/ due à leur faibles priors ?*

6.2.2 Variation du WER en fonction des priors

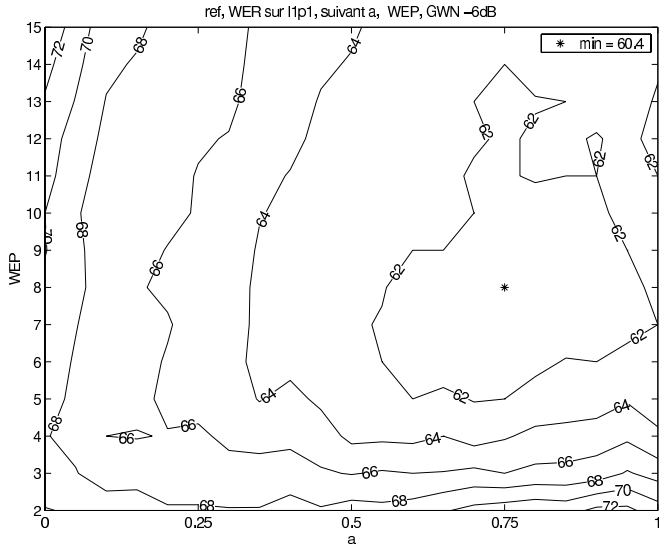


FIG. *6.2: Décodage sur vraisemblances, bruit blanc -6dB. Noter qu'alors l'optimum est pour a=0.75 quel que soit le WEP (Word Entrance Penalty, voir annexe)*

6.2.3 Estimation et fiabilité des priors

L'évaluation des priors $P(q_k)$ est basée sur la segmentation de la base d'entraînement des réseaux de neurones, soit environ 4000 phrases (500 000 trames). Les priors sont la fréquence de chacune des 27 classes sur cet ensemble.

Du fait de la représentation de quelque 0.001 % des trames de l'ensemble d'entraînement des phonèmes /hh dcl d/, alors que la moyenne est à $1/27 = 0.0370$ il est utile de mesurer le biais entre la distribution des priors sur l'ensemble d'entraînement, évalué sur près de 4000 phrases, par rapport à des plus petits ensembles.

Comme nous le voyons dans le cas limite dans la figure 6.1 en cas de mauvaise estimation des posteriors (condition bruitée et distribution plate des estimations de posteriors) ce sont bien évidement ces trois classes phonétiques qui vont 'dévier' le viterbi

Une solution consiste à pratiquer une "validation croisée". Dans ce cas, on garde une petite partie de l'ensemble d'entraînement, appelée ensemble de validation croisée[1] de côté de façon à tester les propriétés de généralisation des priors.

6.2.4 Validation des priors

Après partage de la base d'entraînement en deux sous-ensembles : l'ensemble d' entraînement et l'ensemble de test 10 fois plus petit. Nous mesurons les fréquences des classes dans chacun des 2 ensembles : nous obtenons les distributions p et q. Puis nous mesurons (voir plus loin) une distance $d(p, q)$ entre ces deux distributions. Et nous réévaluons cette distance après chaque application d'une correction de Dirichlet sur ces distributions. La correction de Dirichlet consiste à rajouter N observations de chaque classe :

$p^*(q_k) = \frac{(p(q_k)+N)}{\sum_k p(q_k)+N}$

La courbe qui à chaque itération associe les valeurs $d(p, q)$ corrigées possède son minimum pour N minimisant la divergence entre les distributions des priors de deux ensembles issus du training+cross validation set. Si N est faible, alors la correction est inutile et l'estimation des distributions est valide.

Les divergences et distances KL sont définies en annexe. Pour estimer la distance des priors $p(q_k)$de A et $r(q_k)$ de b, nous sommons les deux divergences : de A vis à vis de b et de b vis à vis de A.

$distance(p, r) = KL(p, r) + KL(r, p)$

Enfin la distance globale est la somme des distances de tous les couples (A, b) :

$distance_{globale} = \sum_{tousles A,b} distance(p, r)$

Nous utilisons alors une méthode inspirée de la régularisation : nous appliquons une distribution de Dirichlet aux priors, ce qui consiste simplement à dire que l'on a "vu" N fois supplémentaire(s) chaque phonème dans chaque ensemble.

[1]Le plus petit possible de façon à minimiser la perte d'exemples d'entraînement, tout en garantissant qu'il soit statistiquement représentatif du futur ensemble de test

$p_N = priors$ de A corrigé au sens de Dirichlet par N
$p_N(q_k) = \frac{card(q_k)+N}{\sum_k card(q_k)+N}$
On voit aisément que c'est une méthode de régularisation
- si N = 0, on ne change rien : $p_0(q_k) = p(q_k)$.
- si N = infini, on converge vers une distribution uniforme.
Reste le choix de N. Il peut venir d'informations extérieures (le bruit par exemple, N peut être choisi comme fonction du rapport bruit/signal).
Nous cherchons alors N* tel que :

$$N* = \underset{N}{\text{argmin}} \sum_{A,b}((KL(p_N, r_N) + KL(r_N, p_N)) \tag{6.2}$$

Un autre critère envisageable est de corriger p pour converger vers r, ou bien l'inverse, il n'est pas évident que N obtenu soit équivalent au précèdent :

$$N* = \underset{N}{\text{argmin}} \sum_{A,b}(KL(p_N, r) + KL(r, p_N)) \tag{6.3}$$

6.2.5 Résultats : Les priors sont généralisables

Nous tirons un gros ensemble de référence A et un plus petit b représentant une condition de test. A et b forment une partition aléatoire de l'ensemble de 540 443 trames. A est choisi 10 fois plus gros que b, rapport équivalent entre le training set et le cross validation set. Nous opérons alors 10 tirages de la sorte et obtenons 10 couples de priors : $p(q_k)$ pour A , et $r(q_k)$ pour b, dont nous calculons les distances respectives.
Les simulations répétées dix fois ont donné $N < 1$, ce qui signifie que les priors sont généralisables.
Les simulations (répétées dix fois) ont donné $N < 1$, ce qui signifie que les priors sont généralisables.
Les tests par validations croisées entre des priors calculés sur des sous ensembles de 200 phrases de l'ensemble d'entraînement et le reste de l'ensemble sont similaires (la correction de Dirichlet par validation croisée comparant les distances des deux distributions demande au pire l'ajout d'une observation supplémentaire par rapport au 500 000 trames observées). Cela signifie que les distributions des priors sont stables quel que soit le tirage des 200 phrases. L'estimation des priors n'est donc pas biaisée.

Il apparaît que l'usage direct des priors sans correction soit bénéfique. Ce débat est aussi ouvert dans (Hennebert, Ris, Bourlard, Renals & Morgan 1997, Bengio et al. 1992) Nous avons vérifié en mesurant les WER sous différentes valeurs des priors (plus ou moins uniformisées) sur différents bruits que le décodage sur les vraisemblances est plus efficace (voir tables précédentes). Nous avons constaté que les

priors légèrement biaisés peuvent être bénéfiques, sans que nous puissions l'expliquer par une erreur sur l'estimation des priors. Comme les priors estimées d'après la fréquence des phonèmes sur la base d'entraînement sont fiables, nous procéderons, comme la théorie l'indique, en utilisant les fonctions de vraisemblances pour nourrir le HMM.

6.3 Résultats de référence en soustraction spectrale

Nous avons établi les scores de référence en soustraction spectrale avec le laboratoire partenaire de Mons (Belgique) suivant la technique de Berouti, et avec les mêmes paramètres pour l'entraînement de l'ANN pleine bande. La méthode de Berouti ne fait que généraliser (Boll 1979) en travaillant sur le spectre à la puissance alpha, avec recherche du meilleur alpha. Le décodeur est identique. On estime le spectre du bruit sur les 100 premières millisecondes de chaque phrase (supposées silencieuses). Puis on suppose que le bruit est stationnaire pour toute la phrase et le spectre est remis à jour au début de la phrase suivante, ceci se répercute par de faibles performances dans le cas de bruit non stationnaire (voir figure 6.4).

Il est normal d'obtenir 10.5% WER qui est résultats moins bons que J-RASTA sur de la parole claire car on utilise des paramètres spectraux PLP sans filtrage rasta.

La comparaison avec les résultats Jrasta est disponible en annexe.

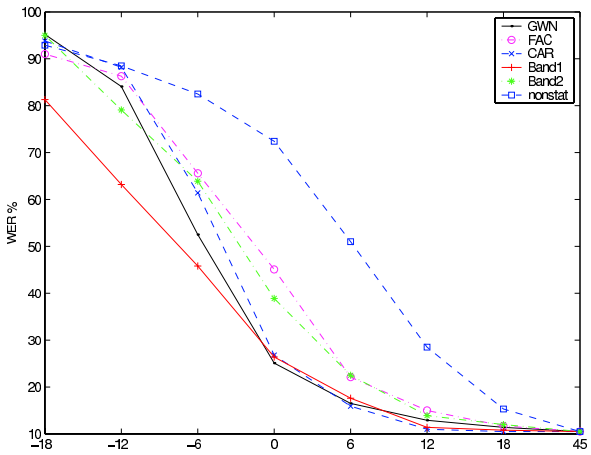


FIG. *6.3: Résultats en soustraction spectrale type Berouti, avec le reconnaisseur ANN HMM sur NB95, 200 ph. On suppose que le bruit est stationnaire pour toute la phrase et le spectre est remis a jour au début de la phrase suivante, ceci se répercute par de faibles performances dans le cas de bruit non stationnaire.*

6.4 Paramétrisation du système hybride

6.4.1 Le reconnaisseur spectre entier PLP et JRASTAPLP

Les paramètres d'entrée des ANNs sont du type PLP, avec un pré traitement J-RASTA (Hermansky & Morgan 1994) Les entrées de l'ANN comportent 9 vecteurs acoustiques consécutifs, fournissant une information contextuelle importante au système. Les sorties correspondent aux 27 phonèmes significatifs de la base de données. Les vecteurs acoustiques pour le spectre entier comprennent 12 coefficients PLP (ou J-RASTA-PLP) et l'énergie (ainsi que les dérivées premières et deuxièmes de ces paramètres). Pour les ANNs du spectre entier nous avons choisi 1750 unités cachées. Sur NB95, en clair, le taux d'erreur au niveau du mot est de 8.0 % pour le système spectre entier sur les J-RASTA-PLPs et de 7.1 % pour le système spectre entier sur les PLPs.

6.4.2 Les 4 reconnaisseurs sous-bandes ANN-HMM

Nous avons travaillé avec $d = 4$ sous-bandes. Pour ne pas rajouter des termes de dépendance inter bandes dans nos modèles FC et FCA, il est précieux de veiller à choisir des bandes spectrales du signal qui se recouvrent au minimum mais dont l'union représente le spectre entier. Dans ce but, nous avons redéfini les sous-bandes qui ne tenaient pas compte de ces contraintes d'indépendance dans des études précédentes (Berthommier & Glotin 1999, Hagen et al. 1999) De plus, nos expériences ont confirmé que la première bande critique n'est pas pertinente en parole téléphonique du fait de la définition même de la bande passante téléphonique, nous l'avons donc supprimée. Ainsi nous avons choisi un ensemble homogène de quatre sous-bandes décrit dans la table 6.3 et qui permet toujours de modéliser un formant par sous-bandes.

Le découpage en sous bande a donc été défini sous les contraintes suivantes :
1/ 4 sous bandes (une par formant)
2/ suppression des premières et dernières sous bandes critiques (perte en parole téléphonique, critère vérifié par expériences, voir annexes)
3/ le moins de recouvrement possible entre les sous bandes, afin de correspondre le plus possible à l'hypothèse d'indépendance entre sous bande.

Notons que dans la bibliographie actuelle en reconnaissance sous bande, ces règles n'avaient pas été suivies et les recouvrements étaient de l'ordre de 1 ou plusieurs bandes critiques (Mirghafori 1997).

Ayant fait une partie de notre étude sur des définitions de sous bandes sans contrainte d'indépendance forte (Morris et al. 2001) nous avons pu constater un gain significatif suite à se découpage, surtout pour le modèle FCA.

L'ordre des analyses LPC et le nombre de coefficients extraits ont été optimisés

sur plusieurs expériences. Dans le cas des fusions de sous-bandes i en une combinaison J, les ordres LPC ainsi que le nombre de coefficients extraits sont la somme de ceux des sous-bandes contenues dans J. Ainsi le nombre de paramètres dans le modèle FC et le modèle FCA sont identiques.

sous-bandes	en Hz	LPC	# coeff.
1	115-629 Hz	3	5
2	565 1370 Hz	3	5
3	1262 2292 Hz	2	3
4	2122 3769 Hz	2	3
134	115-629 Hz, 1262-3769 Hz	7	11

TAB. *6.3: Définition des 4 sous-bandes (coupure à 3dB) et des paramètres extraits. Exemple de combinaison : 134, montrant le calcul du nombre de paramètres des flux : somme de ceux des sous bandes, garantissant un nombre de paramètres constant entre FC et FCA. Le faible recouvrement fréquentiel entre sous-bandes est du aux filtres PLP des bandes critiques.*

Les ANN sous-bandes correspondent à l'ANN spectre entier de base, la seule différence étant le nombre d'entrées, le nombre d'unités cachées restant proportionnel au nombre d'entrée. Les ANNs des sous-bandes et des combinaisons de sous-bandes ont entre 666 et 1750 unités cachés.

Les transcriptions en phonèmes des mots du HMM sont les transcriptions canoniques utilisant un sous-groupe de phonèmes de l'IPA.

6.5 Comportement des reconnaisseurs

6.5.1 Robustesse des reconnaisseurs sous bandes

Nous montrons en figure 6.4 la différence des robustesses des différents modèles sous bandes sur du bruit blanc (qui est équivalent a du bruit coloré pour chaque sous bandes).

Nous voyons que l'amplitude de la dégradation entre 33 et -21 dB est plus forte pour les bandes en basses fréquences, les autres ayant des performances absolues plus faibles même en parole propre. Nous verrons que ces fonctions de dégradation sont précieuses pour étalonner des fonctions de fiabilité en fonction des tolérance relative au faible SNR.

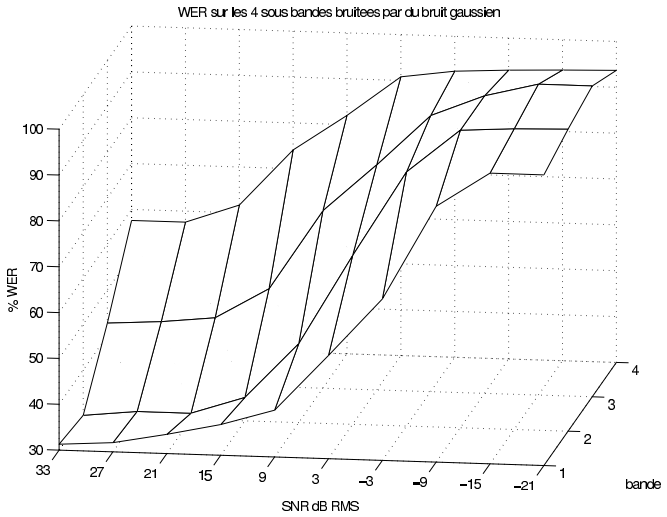


FIG. *6.4: Robustesse des reconnaisseurs sous bande Jrasta,ANN HMM sur NB95, 200 ph. bruit blanc. On voit que les différentes sous bandes on leur comportement propre.*

6.5.2 Comportement des fusion élémentaires des modèles sous-bandes

Nous avons dans nos expériences préliminaires (Glotin, Tessier, Bourlard & Berthommier 1998*a*, Glotin, Tessier, Bourlard & Berthommier 1998*b*, Glotin, Berthommier, Tessier & Bourlard 1998) proposé deux modèles :

1. l'un "associatif" est fondé sur la sélection et l'association des MLP(x) par produit,
2. l'autre "combinatoire" est basé sur la sélection des MLP(xyz).

Dans le cas d'un bruit de bande étroit stationnaire dans chaque fenêtre à court terme (*i.e.*, ne contaminant qu'une seule sous-bande à la fois), un sélecteur guidant le choix du meilleur MLP permet de conserver un taux de reconnaissance optimum. Tout d'abord, nous évaluons systématiquement les performances de tous les MLP(xyz) et MLP(x) avec et sans bruit dans la sous-bande 1 (voir Table 6.4). Le modèle associatif résulte du produit des réponses des MLP(x), fenêtre par fenêtre. Il est intéressant de noter que l'association des MLP(xyz) conduit à 11,5 % d'erreur en signal propre sur NB93, ce qui est équivalent au MLP(1234). Cela suggère qu'un modèle, également de type "associatif", mais avec des groupes de sous-bandes (i.e., à partir des MLP(xyz), plus robustes) est à envisager.

Nous voyons immédiatement table 6.4 que le modèle "associatif" conduit pour le moment à de mauvais résultats par rapport au modèle "combinatoire". Notre hypothèse est que cela traduit la perte des informations de covariance entre sous-

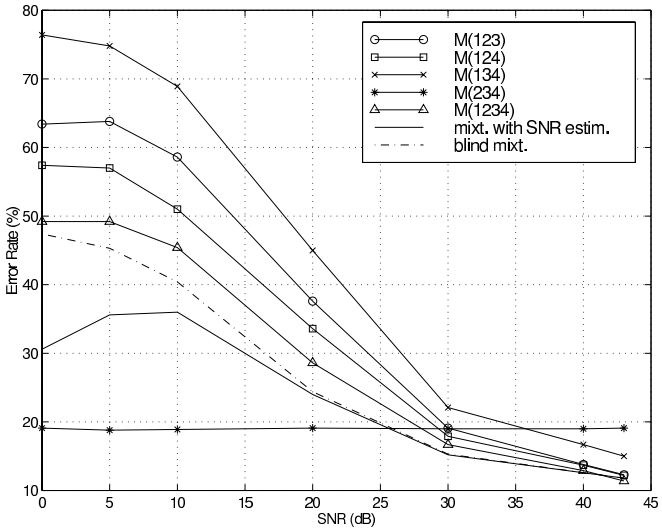


FIG. *6.5: Exemple de comportement des MLPs 'combinatoires' avec du bruit sinusoïdal à 400 Hz à différents SNR, et combinaison par SNR estimation classique voir plus bas.*

bandes. La différence avec les résultats sous Numbers'95 est explicable par le sous-échantillonnage et l'étiquetage phonétique peu sur de Numbers'93 (non labellisé et avec 6 fois moins d'échantillons.)

Les scores sont exprimés en terme de taux d'erreur. Le taux d'erreur est la proportion de mots incorrects en faisant la somme (délétions + insertions + substitutions). En pleine bande et sans bruit, il est de 11,2 %. Les résultats de tous les MLP, avec et sans bruit, sont dans la Table 6.4. Nous vérifions que le MLP(234) présente les meilleurs scores, avec environ 19 % d'erreur, que le signal soit bruité ou non. Ce score est une borne inférieure qui correspond à une identification toujours correcte de la bande bruitée (taux d'IBB=100%, voir chapitre 'Indices de fiabilité').

Ces fusions aux faibles performances face aux modèles entraînés globalement sont élémentaires et posent l'hypothèse forte d'indépendance des posteriors entre sous bandes. Nous verrons comment cette hypothèse peut être atténuée dans le chapitre FCA.

type du MLP	% err. signal propre	% err. signal bruité	MLP associés	% err. signal propre
(1)	38.9	84.8		
(2)	39.3	40.5		
(3)	55.3	56.8	Π_4 MLP(xyz)	11.5
(4)	65.2	65.7		
(1234)	11.2	55.6	(1)*(2)*(3)*(4)	27.6
(234)	19.0	19.2	(2)*(3)*(4)	38.6
(134)	14.9	55.9	(1)*(3)*(4)	32.7
(124)	12.2	48.6	(1)*(2)*(4)	27.8
(123)	12.3	50.5	(1)*(2)*(3)	24.9

TAB. *6.4: Taux d'erreur en pourcentage pour NB93 de mots continus reconnus : signal propre, signal bruité en sous-bande 1, pour les MLP(x), MLP(xyz), MLP(1234) et les produits de trois ou quatre MLP(x) (par ex. (1)*(2)*(3)). Le produit des 4 MLP(xyz) est noté* Π_4 *MLP(xyz). Noter le biais entre entraînement global (xyz) et produit des sous bandes (x)(y)(z).*

6.6 Performances comparées FC et FCA et autres fusion de référence

Nous présentons à la figure 6.8 les résultats des tests sur des bruits colorés dans les différentes sous-bandes 1 à 4, en utilisant les approches FC et FCA et des paramètres PLP. Pour comparaison, cette figure montre également les courbes correspondantes au système spectre entier testé dans les mêmes conditions. Nous constatons que non seulement l'approche FC mais aussi son approximation présentent de très bonnes propriétés de robustesse aux bruits colorés. En effet la reconnaissance peut s'effectuer de façon très fiable sur les composantes non bruitées et la contribution des flux bruités intégrés dans le FC ou FCA est peu perturbatrice car leur distribution de probabilités *a posteriori* a une forte entropie (distribution plus uniforme). Nous reviendrons sur cette analyse dans le cas de le FCA.

La figure 6.6 présente les résultats avec les paramètres caractéristiques J-RASTA-PLP. On voit que le filtre J-RASTA est capable de supprimer une partie des interférences dues au bruit coloré ce qui améliore les résultats sur tous les bruits colorés comparés aux résultats avec PLP seul. Mais là encore le FC en tout RSB est plus performant que le système spectre entier (ou égal en claire). Il en est de même pour le FCA sauf en parole claire où les performances sont moindres car l'expert spectre entier est estimé, la perte des termes de dépendance inter bande se faisant alors plus ressentir. De même les résultats FC ou FCA obtenus avec les PLPs et J-RASTA-PLPs dans le cas du bruit additif non-stationnaire (fig. 6.9) indiquent une robustesse

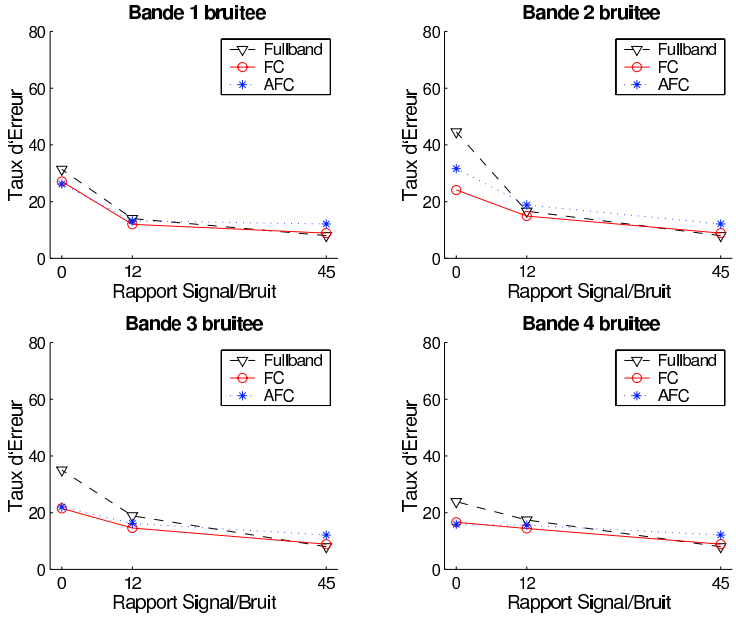


FIG. *6.6: Taux d'erreur des systèmes spectre entier, FC et FCA en utilisant des J-RASTA. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée stationnaire dans des sous-bandes 1 à 4 d'après (Hagen & Glotin 2000).*

plus élevée que celle du spectre entier, même si on note une hausse générale des taux d'erreur par rapport aux deux expériences précédentes (il y a en effet plus de trames bruitées par effet de bord). Dans ces conditions le processus J-RASTA est mis en défaut par rapport au PLP simple, ce qui montre l'inadéquation de J-RASTA à un bruit si non-stationnaire.

Une propriété intéressante de nos modèles est que le FCA montre une robustesse égale ou supérieure au FC, sauf dans le cas où la bande 2 est bruitée. Nous avons mesuré que la bande 2 est la plus performante en RAP propre parmi les 4 sous-bandes, le FCA est donc particulièrement pénalisée dans ce cas. Dans tous les autres cas, les bonnes performances de le FCA sont dues à une meilleure exploitation de la redondance du signal. En effet en condition bruitée les entropies des vecteurs de probabilités *a posteriori* augmentent autrement dit ces vecteurs ont une distribution plus aplatie : les probabilités *a posteriori* tendent vers l'équiprobabilité pour des SNR globaux moyens (pas moins de 0 dB global) [2]. Donc dans le modèle FCA qui procède par produits et normalisations, les sous-bandes bruitées affectent peu la distribution des vecteurs porteurs d'information correcte qui eux sont très

[2]Nous verrons par la suite que quelques classes de phonèmes peuvent saillantes dans cette distribution dans des conditions de faibles SNR.

discriminants. Le modèle FCA joue donc le rôle d'un filtre : l'information provenant d'une sous-bande de données claires est mieux conservée en sortie du modèle FCA, alors qu'elle est noyée avec les données bruitées dans le cas du modèle FC. En FC dès qu'un flux est partiellement bruité, les probabilités *a posteriori* issues directement de l'expert correspondant sont globalement détériorées et irrécupérables comme le montre (Lippmann & Carlson 1997), ce qui défavorise le FC comme le montre la figure 6.7.

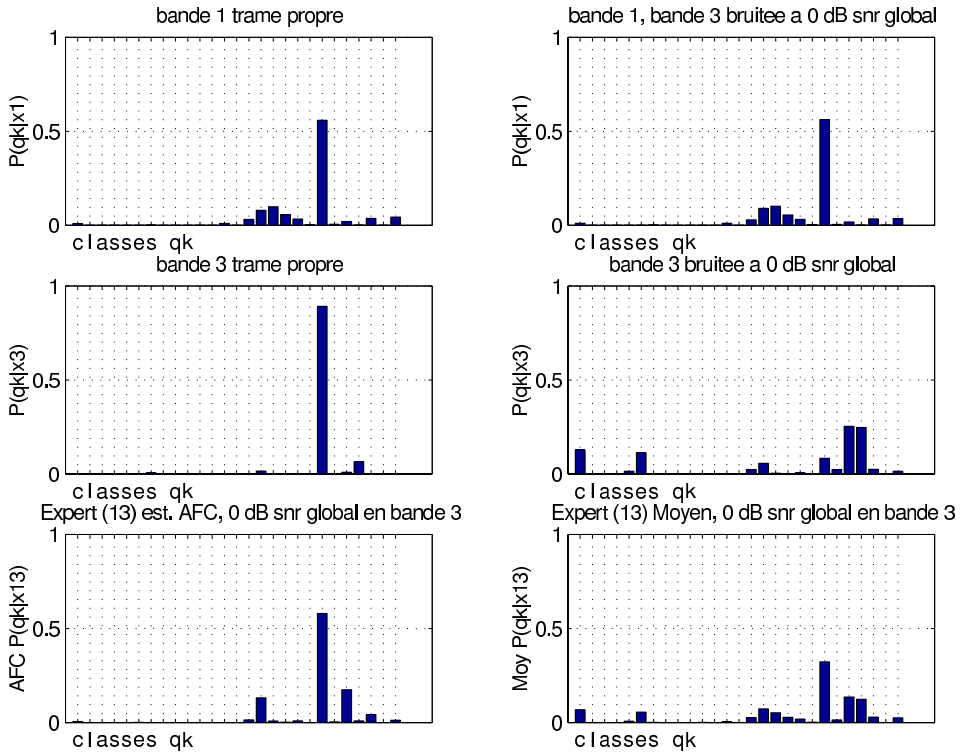


FIG. *6.7: Illustration du FC versus autres modèles pour une trame donnée, les bandes bruitées à 0 dB global ont une distribution aplatie alors que le modèle FCA conserve une distribution proche des références propres*

6.6.1 Discussion

Nous avons montré qu'en utilisant l'approche sous-bandes FC ou son approximation FCA, même dans le cadre le plus simple de la pondération équiprobable, la robustesse du RAP est plus élevée que celle d'un modèle spectre entier J-RASTA pour les bruits à bande limitée stationnaires et non-stationnaires. Avec des modèles à résolution spectrale supérieure des résultats identiques sont attendus en bruits naturels, ce qui est réalisable efficacement avec une procédure récursive. Une amélio-

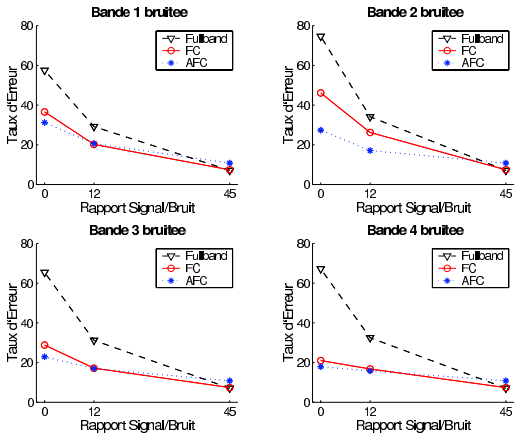


FIG. *6.8: Taux d'erreur des systèmes spectre entier, FC et FCA en utilisant des PLP. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée dans des sous-bandes 1 à 4. [JEP00]*

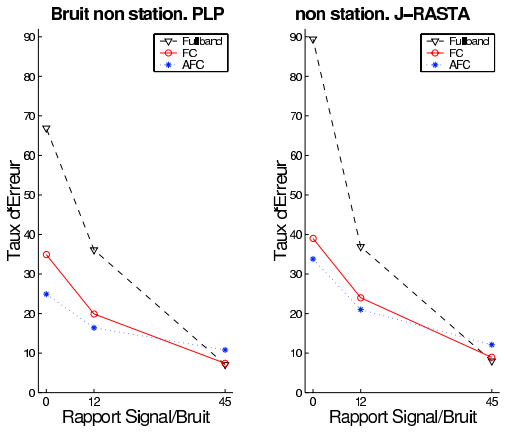


FIG. *6.9: Taux d'erreur des systèmes spectre entier, FC et FCA en utilisant des* **PLP** *(à gauche) et des* **J-RASTA** *(à droite). Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée non-stationnaire variant entre les sous-bandes 1 à 4. JEP00]*

ration en parole propre du modèle FCA est accessible en utilisant le vrai estimateur spectre entier, les performances en claires pour FCA et FC sont alors comparables suivant nos récentes expériences. Cette étude comparative entre FC et FCA a mis en évidence que le modèle FCA a des performances sensiblement égales en conditions bruitées, et parfois même supérieures au FC, ce qui a été discuté. Des études publiées ou en cours montrent un accroissement de robustesse des modèles lorsque les poids sont variables selon les performances relatives des flux en condition claire comme

les poids « Expectation Maximization » et « Least Mean Square Error » (Morris et al. 2001). Mais les gains en robustesse sont plus grands avec l'usage de poids adaptatifs au bruit. Ces derniers peuvent être basés sur un classique RSB (Hagen et al. 1999) suivant une approche «CASA» à partir d'indices d'harmonicité ou de localisation spatiale, poids développés et testés en FCA dans (Berthommier & Glotin 1999, Glotin et al. 1999)

6.7 Conclusion

A travers ces études préliminaires sous avons validé l'approche ANN HMM. Nous avons d'autres part fixé les paramètres de nos reconnaisseurs sous bande. Nous avons montré que des fusions simples de type produit peuvent amener à de gros biais. Nous avons soulevé le problème du passage des posteriors au vraisemblances. L'étude montre une légère perte de performance en pratique avec l'utilisation des priors issues des fréquences phonétiques sur la base d'apprentissage, mais un biais sur les priors n'est pas confirmé par une étude de validation croisée, nous effectuerons donc tous nos calculs avec les fréquences phonétiques.

Troisième partie

Mise en oeuvre des indices CASA en RAP robuste multi-flux

Nous verons deux grands types de modèles : un modèle à identification directe qui consiste en la fusion des informations des différents flux des observations, avant identification des classes, et deux modèles d'identification séparée (asynchrones ou synchrone). Nous allons dans tous les cas étudier la faisabilité de fusion avec nos indices de fiabilité CASA en les comparant aux techniques usuelles. Notre modèle "Full Combination" (FC) sera testé sur des interférences de genre monophoniques, ou bien dans le cadre "cocktail party".

Chapitre 7

Fusion directe : rehaussement par pondération acoustique CASA

7.1 Introduction

Dans ce chapitre, nous proposons un modèle d'extraction et d'usage de l'information CASA apparenté à un filtrage de Wiener. L'effet obtenu est une séparation incomplète du signal de parole et du bruit interférant. Le principe de renforcement est celui du filtrage de Wiener : dans le domaine fréquentiel, un avantage est donné aux composantes pour lesquelles le signal prédomine sur le bruit.

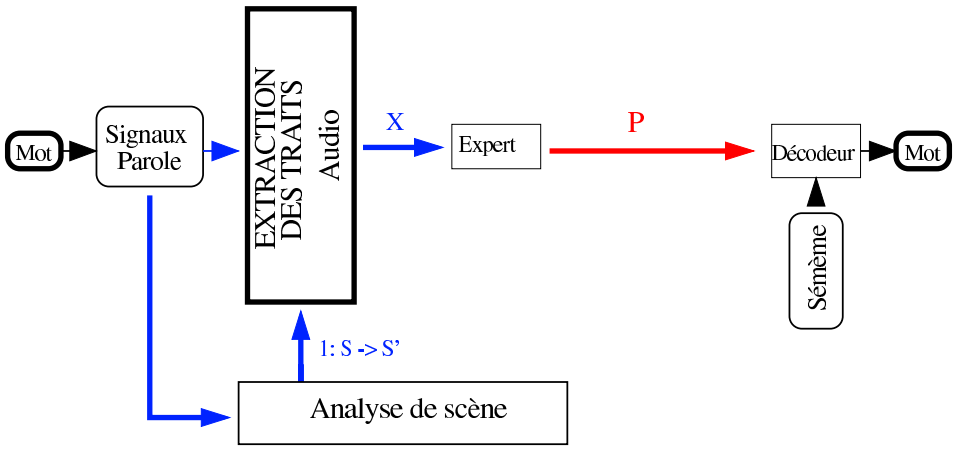


FIG. *7.1: Topologie du système de rehaussement*

Le schéma général du système de ce reconnaissance est donné par 7.1

7.2 Le modèle dans le cas de l'indice de voisement

Nous utilisons la quantification de l'harmonicité du signal liée au voisement : indice R présenté dans la partie I.

7.2.1 La représentation temps-fréquence

Nous appliquons sur le signal d'entrée un filtrage par banc de filtres. La fréquence d'échantillonnage est de 8 kHz. Afin de diviser le plan temps-fréquence en régions de tailles variables, nous faisons appel à la méthode de décomposition proposée par Tessier et coll. (Tessier 2001) . L'intérêt principal de cette méthode est de contrôler la taille des régions temps- fréquence en faisant varier le nombre de filtres nc=(4,8,16) et la durée de chaque trame temporelle entre (512, 1024, 2048) échantillons. Pour chaque filtre Fi, nous définissons une fenêtre de Hanning dans l'échelle Bark, centrée sur Fci, puis nous replaçons cette fenêtre dans l'échelle linéaire de la FFT.

$$F_{Bark} = 13 \arctan\left(\frac{0.76 F_{Hz}}{1000}\right) + 3.5 \arctan\left(\frac{F_{Hz}^2}{7500^2}\right) \tag{7.1}$$

Pour chaque trame du signal d'entrée n, les spectres sous-bandes obtenus sont :

$$|X_i(\omega)| = F_i(\omega).|X(\omega)| \tag{7.2}$$

Les fréquences centrales Fci sont calculées à partir des bornes Fmin et Fmax du domaine couvert par le banc et filtres et à partir du nombre de filtres nc. L'intervalle séparant deux filtres est égal à

$Fint = (Fmax - Fmin)/nc.$

Les deux filtres extrêmes ne couvrent que 1.5 intervalle tandis que les nc-2 filtres centraux s'étendent sur 2 intervalles. La fréquence centrale des filtres est $Fci = Fmin + (i - 0.5) * Fint$.

Nous avons donc veillé à ce que la somme des filtres soit égale à 1 en toute fréquence.

Il est intéressant de noter que l'usage de filtre PLP aurait conduit à un biais lors de la reconstruction du spectre car ce filtre n' a pas été étudié pour garantir la même propriété. Nous représentons dans la figure 7.3 la somme d'un filtre PLP.

7.2.2 Pondération du spectre par l'indice R

Nous comparons plusieurs versions de notre modèle. D'une part, nous utiliserons deux versions de l'algorithme d'évaluation de l'indice R, l'une faisant appel à la démodulation du signal en sous-bandes (appelée proc1) et l'autre pas (proc2). La

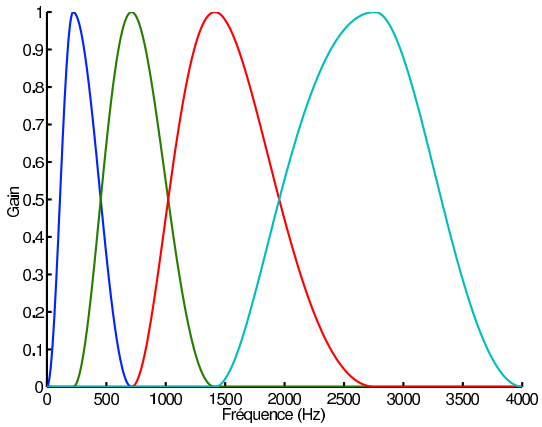


FIG. *7.2: Banc à 4 filtres. Entre 0 et 4000 Hz, nous avons Fint=4.31 Bark, et Fci=[2.16, 6.47, 10.79, 15.10]Bark*

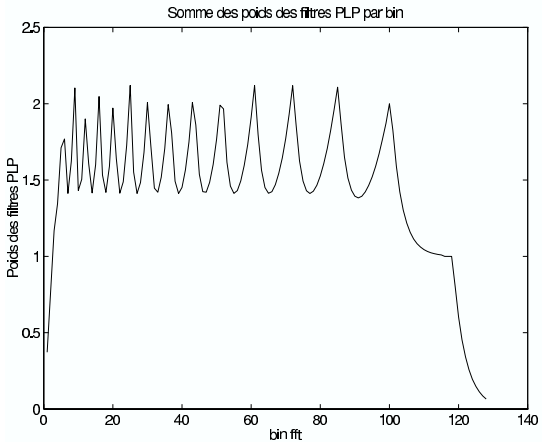


FIG. *7.3: Somme des poids des filtres PLP des 16 bandes critiques : la somme n'est pas constamment unitaire*

démodulation consiste en une rectification simple alternance suivie d'un filtrage passe-bande trapézoïdal [0,90,350,1000]Hz (voir partie I).

Pour chaque bande i c'est la racine carrée de R_i qui est utilisée pour calculer le facteur de pondération Wi :

$$W(R_i) = \sqrt{(max(0, R_i))} \tag{7.3}$$

Le choix de la puissance 0.5 est motivé par l'existence d'un compromis entre le niveau de renforcement et le niveau de distorsion des signaux résultants. Le facteur de pondération W permet d'évaluer le spectrogramme de la source c, de même que

celui de la source interférente n à partir de $(1 - W)$, ce qui correspond à la notion de "fonction de partage" et de ségrégation des sources (Tessier 2001). Le facteur W est appliqué sur le spectrogramme au niveau des régions temps-fréquence et

le spectre reconstruit de chaque fenêtre temporelle est égal à la somme des spectres sous-bandes :

$$|\widehat{X}(\omega)| = \sum_{i=1}^{nc} W(R_i).|X_{n,i}(\omega)| \tag{7.4}$$

7.2.3 Evaluation du modèle

Lorsque les signaux sont re-synthétisés, nous quantifions leur distorsion par rapport aux signaux clairs à l'aide d'une mesure de similarité entre les spectres. Le degré d'atténuation du bruit est aussi appréciable à l'écoute. Puis le gain du renforcement est évalué à partir des taux de reconnaissance automatique obtenus pour différents niveaux de bruit, par comparaison entre signaux non traités (proc0) et signaux traités (proc1 et proc2). Le modèle de reconnaissance utilisé comporte une étape de pré-traitement qui lui est propre (J-RASTA-PLP, [Her94]) et le gain que nous mesurons est cumulatif par rapport à celui conféré par cette méthode, qui est également efficace contre les bruits stationnaires que nous utilisons.

Nous utilisons le signal clair c de module |Xc| afin de mesurer la précision de la reconstruction obtenue à partir du signal bruité n, en faisant appel à une référence. Nous définissons le RA dans le domaine spectral. Nous complétons cette estimation à l'aide du SNRI, qui prend également en compte le signal bruité. Dans chaque trame temporelle, tous les spectres "pleine bande" sont normalisés, de telle sorte que la somme des modules soit égale à 1.

$$RA = 10\log\frac{\int_\Omega |X_c(\omega)|^2}{\int_\Omega \left(|X_c(\omega)| - |\widehat{X}(\omega)|\right)^2} \tag{7.5}$$

$$SNRI = 10\log\frac{\int_\Omega \left(|X_c(\omega)| - |X_n(\omega)|\right)^2}{\int_\Omega \left(|X_c(\omega)| - |\widehat{X}(\omega)|\right)^2} \tag{7.6}$$

où $\Omega/2\pi = [0, 4000]Hz$, où X_c est le spectre clair et X_n est le spectre bruité.

Une statistique de RA et SNRI est établie pour toutes les trames, silences inclus, des mêmes 100 phrases de la partie "test" de NB95 (base multilocuteur de "digits" téléphonés, à 8kHz). La durée de la trame d'analyse est fixée à 1024 échantillons, proche de celle utilisée pour les tests de reconnaissance (1000 éch.), avec recouvrement de moitié. L'effet des deux facteurs (1) nombre de sous- bandes (nc), et (2) longueur L de la fenêtre de traitement, est analysé en additionnant un bruit blanc gaussien (GWN) à 0dB (table 1). Le facteur nc a un petit effet négatif pour proc2,

et le facteur durée présente un petit effet positif pour les deux. Nous voyons que proc1 est légèrement meilleur que proc2, mais il n'y a pas de différence observée à 0 dB pour la condition principale de l'étude, qui correspond à celle des tests de reconnaissance (nc=4, 1024 éch., soit 128 ms).

nc/ L	512	1024	2048
4	6.1/6.0	6.1/6.0	6.5/6.4
8	6.5/6.4	6.5/6.2	6.5/6.2
16	6.0/5.3	6.3/5.6	6.3/5.6

TAB. *7.1: Moyenne de RA en dB pour proc1/proc2, sur toutes les trames de 100 phrases de la base de test. Colonne : variation de la durée. ligne : variation du nombre de sous-bandes (nc). Le SNRI moyen (non figuré) est bien corrélé avec le RA.*

Ensuite, pour cette condition, nous faisons varier le niveau du bruit blanc entre -18 et 21 dB (voir figure). Le SNR est ici exprimé en dB RMS silence inclus relativement au signal clair c. La figure 3 montre que proc1 est meilleur que proc2 lorsque le SNR est élevé, au dessus de 0dB, mais pas au dessous. De plus, nous observons que la courbe de SNRI est non monotone et présente un maximum à environ 6dB.

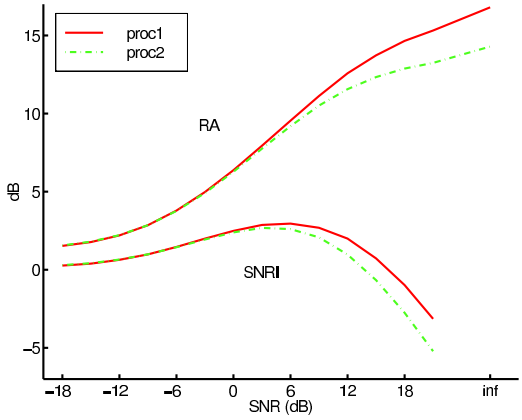


FIG. *7.4: Variation de RA et SNRI en fonction du SNR (en dB RMS, avec du bruit GWN). INF est obtenu avec le signal clair*

7.2.4 Tests de reconnaissance

La source désirée est reconstruite afin d'alimenter un système de reconnaissance pleine bande de type HMM/ANN. Nous testons la possibilité d'adapter le processus

de reconnaissance à ce mode de renforcement au cours de la phase d'apprentissage réalisée à partir des signaux clairs. Ceci est motivé par le fait que des distorsions sont introduites par ce traitement car il favorise les signaux voisés au détriment des signaux non voisés.

Nous entraînons le modèle dans 3 conditions différentes i=0,1,2 à partir des signaux d'origine de NB95 (proc0), ou bien à partir des signaux renforcés proc1 et proc2. Les paramètres sont fixés à nc=4 et durée=1024 d'après les résultats en RA (7.1) qui montrent que cela correspond au paramètres optimaux de notre expérience. Trois perceptrons multicouches (mlp0, mlp1, mlp2) résultent de cette phase d'adaptation. La méthode de pré-traitement des signaux d'entrée proci est toujours un J-RASTA-PLP afin de rechercher un effet coopératif entre l'extraction des caractéristiques propres au signal de parole et l'effet de renforcement lié à l'harmonicité.

Pour chaque point de la phase de test, nous ferons toujours appel aux mêmes 100 phrases de la base de test de NB95. Nous noterons proci>mlpj le test du reconnaisseur mlpj (entraîné avec des signaux procj) avec un signal d'entrée proci (voir table 7.2). Les performances sont établies en WER. Nous appliquons tout d'abord ce test sur les signaux obtenus en faisant varier les paramètres de durée et nc comme précédemment pour observer dans la table 7.5 de résultat que le RA n'est pas bien corrélé au WER dans plusieurs conditions. Nous voyons que proc2>mlp2 et proc1>mlp1 ne sont pas significativement différents.

Les paramètres nc=4 et durée=1024 sont logiquement les paramètres optimaux, car ils correspondent aux paramètres d'apprentissage des MLPs. Nous les choisissons pour établir une relation entre WER et SNR avec du bruit blanc gaussien (GWN). Le RA est corrélé négativement avec le SNR, et nous observons aussi que proc2>mlp2 est légèrement meilleur que proc1>mlp1 7.6. Nous évaluons le gain de performance des méthodes de renforcement proc1 et 2 par comparaison avec proc0>mlp0 faisant uniquement appel au J-RASTA-PLP comme méthode de pré-traitement. Ce gain correspond à une mesure du décalage exprimé en dB au point WER=65% placé au centre de l'intervalle [30-100]%, au niveau duquel nous observons une différence significative. Le gain est significatif au dessous de 3-6dB, sans perte importante pour le signal clair.

Un second test est réalisé à l'aide d'un bruit stationnaire de voiture (véhicule roulant à 80 km/h fenêtres fermées), dans des conditions strictement équivalentes à la figure 7.7. Les performances globales sont similaires à celles observées pour le bruit blanc GWN et les combinaisons proc1>mlp1 et proc2>mlp2 sont équivalentes. Mais nous constatons un gain plus faible et une petite dégradation des performances au dessus de 3dB pour proc1>mlp1 et proc2>mlp2. Cela est probablement dû à l'atténuation des phonèmes non voisés.

Enfin, nous établissons le gain pour les combinaisons proci>mlpj avec les deux types de bruit (7.2).

	mlp0	mlp1	mlp2
GWN	3.0/3.0	3.4/4.0	3.2/3.7
Bruit de voiture	1.9/1.9	3.1/3.1	3.0/3.0
Moyenne	2.5/2.5	3.2/3.6	3.1/3.3

TAB. *7.2: Gain en dB mesuré par la variation des SNR à WER constant=65%WERD. elta/WER65/ de proc1/proc2 en dB, pour le bruit blanc GWN et le bruit de voiture, obtenu avec les différent mlpi. Les points de référence appartiennent à la fonction proc0>mlp0 et ont pour valeurs -2.3dB (GWN) et -3.8dB (bruit de voiture) en WER=65%. On note que l'adaptation des MLP donne faible gain dB et que le maximum de gain est en moyenne pour proc2>mlp1 mais la différence n'est pas significative.*

.

7.3 Rehaussement par ITD

Ce modèle est similaire au modèle de rehaussement en monophonie décrit dans un chapitre précèdent et a fait l'objet de la thèse d'Emmanuel Tessier et de performances de reconnaissance en collaboration avec nos développements en HMM/ANN (Tessier et al. 1999, Tessier 2001) Les techniques basées sur les réseaux de microphones sont de plus en plus prometteuses car elles sont plus robustes en présence de bruit que les techniques basées sur un système d'acquisition avec un seul microphone. En plus, les réseaux de microphones ont un caractère directionnel très apprécié pour les applications mains-libres. Ces dernières deviennent nécessaires dans notre vie moderne pour le confort qu'elles offrent.

Un enjeu essentiel en reconnaissance de la parole et dans le courant CASA est la séparation de la parole par rapport à des sources interférentes.
Le problème est encore plus délicat si l'on considère le cas ou les interférences sont de même nature que la parole. Dans ce dernier cas nous sommes dans le paradigme dit "cocktail party" où les performances humaines de reconnaissance de la parole sont encore remarquables. Si la perception est audiovisuelle dans le cas réel, il reste que nous pouvons supposer qu'un traitement adéquat de l'information de localisation de la source à décoder est la clef de ce problème.

Il est montré (Tessier 2001) qu'une information sur la localisation de sources en concurrences permet d'augmenter les niveaux d'énergie de la cible, mais aussi d'améliorer l'intelligibilité du signal ainsi augmenté.

Face à ce courant aux inspirations psychoacoustiques, différentes approches ont été proposées ces dernières années, dont la séparation aveugle des sources. Cette dernière basée sur une minimisation de l'information mutuelle entre les 2 ou n voies a bénéficié d'importantes améliorations quand à la nature des problèmes solvables. En effet, le BS n'est théoriquement plus limitée au cas de signaux additifs (Choi

et al. Sept 2001). Une autre technique de filtrage est l'annulation adaptative de bruit et repose sur l'utilisation de deux microphones, l'un captant la parole bruitée et l'autre le bruit ambiant seul. Le principe consiste a calculer le filtre adaptatif permettant d'estimer le bruit perturbateur et de le supprimer du signal. Nous ne traiteront pas cette aspect trop contraignant quand à l'enregistrement du signal.

Nous avons aussi adopté une analyse en banc de filtres perceptifs que nous nous avons adaptés afin de changer facilement le nombre de filtres et d'avoir l'avantage d'un calcul de FFT. La différence avec d'autres bancs de filtres comme le MFCC est que nous utilisons une échelle en Barks (au lieu d'une échelle en Mels) et des fenêtres fréquentielles de Hanning (au lieu de triangles).

Nous garantissons toujours que la somme des poids vaut un (c'est-à-dire que le gain du filtre est unitaire). C'est le même principe pour les bancs de filtres de MFCC basés sur des triangles. Quelques modèles classiques utilisés dans l'identification (comme le PLP) ne préservent pas cette propriété fondamentale.

Les gains sur la base STNB95 sont significatifs : WER de l'ordre de 65% contre 75% sans rehaussement : voir (Tessier et al. 1999).

7.4 Conclusion

Dans les faibles SNR, nous montrons une amélioration significative des taux de reconnaissance dans des bruits stationnaires forts à partir d'une méthode de renforcement utilisée conjointement avec le pré-traitement de type J-RASTA-PLP. Les segments vocaliques sont les plus résistants aux bruits forts par leur intensité globale et parce que les trajectoires formantiques sont à la fois saillantes, redondantes et continues temporellement. L'effet du renforcement est cumulatif dans ces conditions. Par rapport à une méthode spectrale l'usage d'un algorithme d'évaluation temporelle de l'indice d'harmonicité est attrayant puisqu'il est (1) simple et rapide, (2) il fait appel à peu de connaissances a priori, et (3) il est plausible en tant que modèle auditif. Nous montrons aussi que la démodulation (proc1), ainsi que l'adaptation (mlp1 et 2) sont optionnelles pour le type de bruit testés (non harmonique). L'évaluation de l'index d'harmonicité étant locale spectralement et temporellement, il est concevable d'obtenir un gain important avec des bruits non stationnaires.

La méthode de rehaussement testée reste efficace dans le cas d'un rehaussement d'un spectre de double parole par ITD, ou d'interférences plus classique. L'étude valide donc ces indices.

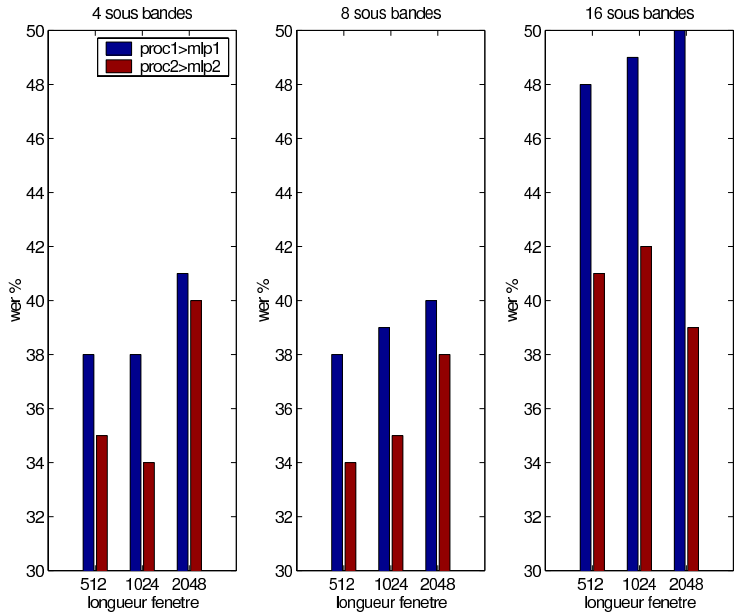


FIG. *7.5: WER moyen à 0dB pour proc1>mlp1/proc2>mlp2, sur 100 phrases de la base de test. colonne : variation de la durée de la trame (en échantillons) pour l'application de proci. ligne : variation du nombre de sous-bandes (nc). Le taux d'erreur (WER) de proc0>mlp0 est de 47% à 0dB.*

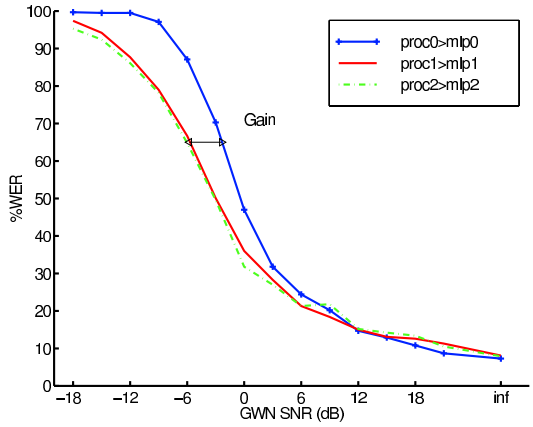


FIG. *7.6: Courbe de réponse du modèle en WER avec du bruit blanc GWN en comparaison avec la référence proc0>mlp0. JRASTA seul (proc0mlp0) ou précédé du rehaussement (procxmlpx, x=1 avec démodulation, x=2 sans). Calcul du gain delta WER65. Le WER de proc0>mlp0 en clair (INF) est de 7.3%*

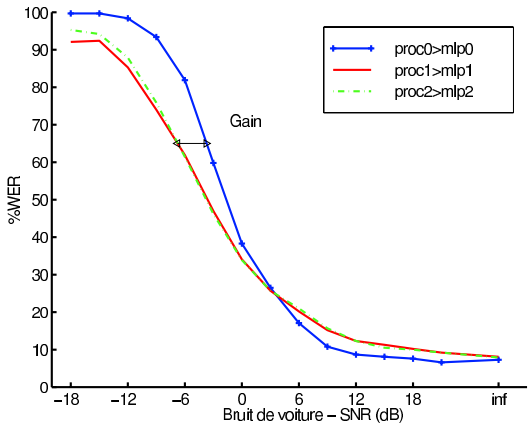


FIG. *7.7: WER obtenu en faisant varier le niveau d'un bruit stationnaire de voiture*

Chapitre 8

Application de l'indice de voisement en reconnaissance multi-flux audiovisuelle

Objectifs

Les flux audio et visuels sont à la fois cohérents (dus à une cause commune, donc jusqu'à un certain point liés) et autonomes, ce qui fait la richesse et la difficulté de l'intégration audiovisuelle.

Le but de ce chapitre est d'évaluer le gain en reconnaissance de parole continue, grand vocabulaire, apporté par notre indice de voisement gérant la fusion de ces deux modalités tant en condition claire qu'en condition d'interférences très perturbatrices pour un reconnaisseur automatique puisque nous nous sommes placés dans le paradigme "cocktail party" (paroles simultanées de cafétéria à 10 dB SNR). Nous nous sommes intéressés au problème de la fusion de l'estimateur, estimant le niveau de qualité audio globale de chaque phrase, puis nous nous sommes placés à l'échelle de la trame, intégrant dynamiquement ce même indice trame à trame.

Nous donnons les grandes lignes de l'état de l'art sur les recherches menées dans ce domaine (Schwartz 2001). Puis nous présentons les modèles développés dans le cadre de notre travail, et les gains apportés par notre estimateur de qualité de la modalité audio.

8.1 Introduction

Christian Benoît disait (Benoît 1996) :

"La bimodalité de la parole est au centre de nombreuses recherches fondamentales visant à mieux comprendre comment l'homme parle en coordonnant les commandes

d'une demi-douzaine d'articulateurs, la moitié d'entre eux étant entièrement ou partiellement visibles."

On comprend alors la complexité de la bimodalité de la parole. Mais rappelons aussi son efficacité : la lecture labiale à elle seule donne accès à environ 50 % des phonèmes d'une langue, et à 15 % des mots (Bernstein, Demorest & Tucker 1998). Dans le cas de parole bruitée, il est estimé que la lecture labiale apporte un bénéfice équivalent à un débruitage de 11 dB (MacLeod & Summerfield 1987).

Les travaux de l'équipe audiovisuelle de l'ICP (Schwartz 2001) ont montré que les scores d'intelligibilité sont toujours meilleurs avec deux entrées sensorielles qu'avec une, les gains produit par cette synergie se répercutent autant sur la reconnaissance des phrases, des mots que des phonèmes comme le montre les résultats de (Robert-Ribes 1995).

La reconnaissance audiovisuelle est orthogonale à toute les autres méthodes de fusion multi-flux, puisqu'elle propose d'ajouter une information dont les perturbations sont indépendantes du flux acoustique.

8.1.1 Architecture de reconnaissance audiovisuelle

Une classification des processus de décision multicapteurs peut se faire suivant 3 schémas, passant d'un ensemble de plusieurs flux de données à une décision globale. Les trois grandes classes de fusion ont été présenté précédemment.

De ces différents mécanismes découlent quatre architectures de base de la fusion audiovisuelle (Teissier 1999).

Le premier est dit Identification Directe (ID), il n 'y a pas d'étapes intermédiaire de mise en forme commune des données.

Le second est dit d'identification séparée (IS), il repose sur une classification phonétique séparée de l'image et du son, suivi d'un processus tardif de fusion.

Le troisième cas correspond à une intégration précoce. Il se partage en deux sous cas suivant que l'on considère une modalité dominante sur l'autre. Si l on considère que l'audition est pour la parole la modalité dominante (Schwartz 2001), est que l'entrée visuelle est recodée sous un format compatible avec celui des représentations auditives, c'est le modèle de Recodage dans la modalité Dominante (RD). Si l'on considère qu'il existe un recodage commun des deux entrées sensorielles vers la modalité motrice on abouti au modèle de Recodage Moteur (RM), sans dominance d'une modalité sur l'autre.

Le modèle IS est le plus courant en reconnaissance automatique car il permet de bénéficier des avancées théoriques en fusion de décisions. Le modèle RD apparaît peu efficace pour traiter la parole dans des applications de reconnaissance robuste car l'audition n'est pas systématiquement dominante.

Cette approche est depuis plusieurs années testée et surpasse la reconnaissance

audio seule (Adjoudani & Benoît 1996, Dupont & Luettin 2000, Potamianos & Graf 1998, Rogozan, Deléglise & Alissali 1997, Teissier, Robert-Ribes & Schwartz 1999)

Les gains en performance sont impressionnants comparés au système audio seul. Cependant il subsiste dans ces approches le problème de l'estimation de la qualité du flux audio. En effet les techniques de fusion des deux modalités font appel la plupart des cas à une estimation du SNR, qui peut s'avérer délicate dans le cas de parole interférente par exemple.
De plus toutes les études précédentes sont faites sur un petit vocabulaire (de l'ordre de la dizaine de mots), un nombre très restreint de locuteurs (Hennecke, Stork & Prasad 1996, Chibelushi, Deravi & Mason 1996) signal ne dépassant pas des dizaines d'heures ce qui rend la généralisation de leurs résultats délicats.

A ce jour deux problèmes restent encore ouverts pour la mise en place d'un modèle de reconnaissance audio-visuelle large vocabulaire (LVCSR) :
(a) Le choix des traits visuels.
(b) Le mode de fusion, et son contrôle dépendant du contexte.

En ce qui nous concerne le (a) est hors sujet. Mais notre indice de voisement et l'estimation de la probabilité de bruitage du flux audio en découlant est sujette à jouer un rôle intéressant dans le (b) qui devrait être validé sur un corpus grand vocabulaire multi-locuteurs.

Nous avons eu l'opportunité de nous confronter à cette problématique lors du workshop de reconnaissance audio-visuel de Johns Hopkins University-Baltimore USA durant l'été 2000. Organisée par F. Jelinek au CLSP, notre équipe invitée regroupait 6 chercheurs confirmés en reconnaissance grand vocabulaire, en extraction de traits visuels, en modèle de fusion audio-visuelle et en estimation de pondérateur de fusion. L'équipe comprenait 2 chercheurs d'IBM C.Neti et M.Potamianos, un chercheur de l'IDIAP J.Luettin, un Post doctorant du CMU, un docteur du CLSP et moi même.

8.2 Entraînement des modèles

8.2.1 Base visuelle et extraction des traits visuels

La base visuelle est décrite en annexe. Nous avons utilisé dans ce cadre les traits visuels générés par l'équipe d'IBM (Hennecke et al. 1996) transformations linéaires de la région représentant la bouche du locuteur (Potamianos, Verma, Neti, Iyengar & Basu 2000) et requiert un tracking efficace de cette région, en utilisant la méthode ROI (Senior 1999) Le vecteur de sortie de cette chaîne de traitement a 41 dimensions, et une fréquence d'échantillonnage équivalente au flux audio : 100 Hz.

8.2.2 Les techniques utilisées pour la fusion précoce des flux

Deux fusions ont été abordées : une fusion précoce, synchrone des traits audio-visuels par concaténation, et une fusion tardive par pondération des vraisemblances des deux modèles. Nous ne présentons ici que la seconde fusion, la première étant publiée dans (Glotin, Vergyri, Neti, Potamianos & Luettin 2000).

8.2.3 Modèle de fusion tardive asynchrone

Nous avons pour cette étude implémenté deux modèles multistream, l'un de type synchrone, l'autre de type asynchrone. Dans les deux cas, la vraisemblance finale conditionnée à une classe est le produit pondéré des vraisemblances des observations des flux audio et vidéo (voir fig. 8.1).

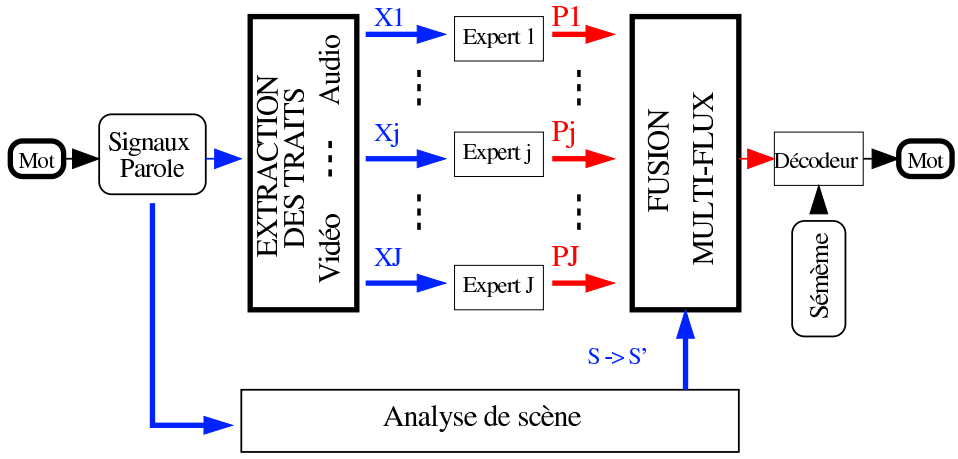


FIG. *8.1: Topologie du système de pondération audiovisuelle tardive*

Les exposants utilisés comme pondérateurs sont les facteurs de fiabilité attribués à chacune des modalités. De telles approches de pondération se trouvent dans (Teissier, Robert-Ribes, Schwartz & Guérin-Dugué 1999, Bourlard & Dupont 1996, Okawa, Nakajima & Shirai 1999, Dupont & Luettin 1998)

Nous n'utiliserons qu'un facteur de pondération déterminant les fiabilités relatives des deux modalités et sommant à un.

$\mathbf{o}^{(t)} = \{\mathbf{o}_{\mathrm{A}}^{(t)}, \mathbf{o}_{\mathrm{V}}^{(t)}\}$

L'émission de l'état du HMM multimodal pour chaque trame t est donnée par :

$$Pr[\mathbf{o}^{(t)}|\mathsf{c}] = \prod_{s\in\{\mathrm{A},\mathrm{V}\}} [\sum_{j=1}^{J_{sc}} w_{scj}\mathcal{N}_{D_s}(\mathbf{o}_s^{(t)}; \mathbf{m}_{scj}, \mathbf{s}_{scj})]^{\lambda_{sct}} \tag{8.1}$$

Les facteurs de pondération sont choisis tels que : $\lambda_{sct} > 0$

$\lambda_{\mathrm{V}ct} = 1 - \lambda_{\mathrm{A}ct}$.

Ce genre d'approche est courante sur du petit ou moyen vocabulaire (Potamianos & Graf 1998, Dupont & Luettin 1998, Dupont & Luettin 2000, Rogozan et al. 1997, Jourlin 1997)

De plus, nos modèles audiovisuels synchrones ou asynchrones ont été entraînés de façon globale sur un critère de maximum de vraisemblance, ce qui est rare dans la bibliographie. Nous montrons un gain de performance par rapport au cas de recombinaison de deux modèles unimodaux entraînés séparément (Glotin et al. 2000).

Le modèle asynchrone est du même type que ceux présentés dans (Brooke 1996, Dupont & Luettin 2000, Varga & Moore 1990*b*)

Le paramètre de pondération est central dans ce type d'approche. Afin de se concentrer sur son étude, nous nous placerons dans un cadre d'entraînement optimal des deux modèles : base d'entraînement claire, ou bruitée, afin de tester le seul paramètre de pondération dans ces deux cas.

Pour chaque entraînement global, les exposants fixes d'entraînement ont été choisis d'après les expériences avec même paramétrisation effectuées à IBM.

Les systèmes ont été entraînés étant donnés les traits acoustiques et visuels extraits par IBM, ainsi que les modèles de prononciation, avec la toolbox HTK (Rabiner & Juang 1993, Young, Kershaw, Odell, Ollason, Valtchev & Woodland 1999)

Outre l'entraînement de ces modèles, notre contribution majeure à cette étude pilote est de tester la pertinence notre indice R (Berthommier & Glotin 1999)

8.2.4 Réévaluation des treillis et calcul des erreurs mot

Etant donnée la charge de calcul (modèles de 600 MO, plus de 2h de signal de test...), nous avons procédé pour la tâche de reconnaissance à un Írescoringí des ÍlatticesÍ. C'est-à-dire qu'à partir de nos modèles, nous générons une fois pour toutes les N meilleurs réseaux de décodage possibles. Nous donnons ensuite un nouveau score à chacun de ces réseaux en considérant d'autres modèles, d'autres stratégies de fusion, etc, avec la fonction "HTK decoder" (HVite) (Young et al. 1999).

Les ensembles de réseaux qui ont été générés pour notre expérience correspondent aux traits clairs ou bruités, pour les modèles synchrones ou asynchrones. Dans chaque cas il s'agit de modèles pentaphoniques, ce qui résulte en près de 50000 mélanges de Gaussiennes.

8.2.5 Pondération par indice globale tirée de R

Dans cette section nous étudions le cas de la pondération de nos modèles suivant un exposant fixe pour chaque phrase, dérivé d'une moyenne sur les R de ses trames.

Notre idée est de donner au système une indication sur le taux de présence de parole cible dans le cas clair, ou le taux de dominance de la source audio par rapport

au fond de parole ajouté à 8.5 dB dans le cas bruité. D'après les caractéristiques tirées dans le chapitre sur les indices, nous dérivons le pondérateur comme étant la moyenne des R des trames dans lesquels $R > 1/2$.

Pour chaque phrase :

$\alpha = MEAN(R > 1/2)$

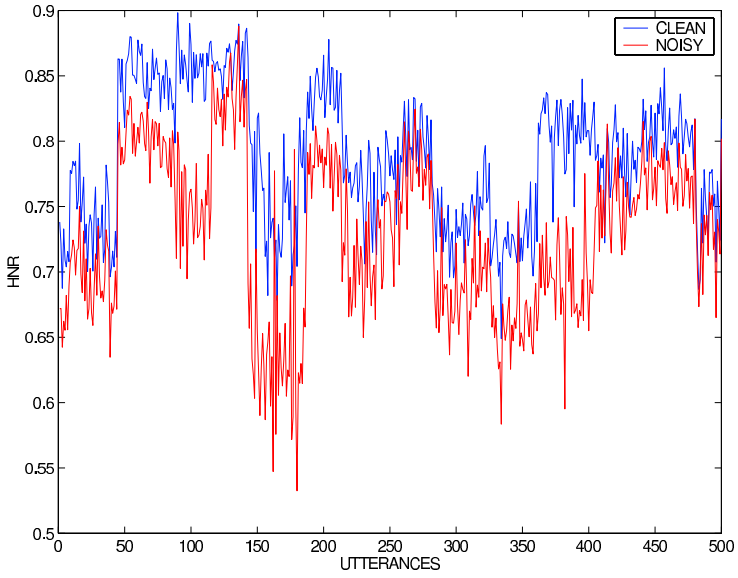


FIG. *8.2: α pour les 500 premières phrases de test, ce qui représente 14 locuteurs différents par paquets de 40 phrases environ. Valeurs en parole propre et bruitée par du bruit de parole. On note les différences des pondérateurs suivant les différents locuteurs, même en parole claire. L'usage de ce facteur permet d'ajuster les modèles en fonction de la qualité audio des enregistrement suivant notre critère $R > 1/2$.*

8.2.6 Résultats et discussion

Les WER sont présenté en Table 8.1.

Les détails de l'élaboration de chaque modèle sont donnés en (Glotin et al. 2000). Ces systèmes sont entraînés suivant les λ_A optimisés par IBM pour le cas propre (resp. bruité) : 0.7 (resp. 0.6) pour le modèle synchrone, et 0.6 (resp. 0.7) pour AV-PROD.

Le modèle AV-MS-UTTER (AV synchrone pondéré par indices globaux décrits dans cette section), donne un gain de 7% relatif WER par rapport au système AV non pondéré dynamiquement, ce qui renforce notre hypothèse de dépendance du WER à la "qualité" de parole du locuteur. Un gain plus faible est observé dans le cas de

	Clean audio WER% (relative)		Noisy audio WER% (relative)	
Audio-only	14.44	—	48.10	—
AV-MS	14.62	+1.2	36.61	-23.9
AV-MS-UTTER	13.47	-6.7	35.27	-26.7
AV-PROD	14.19	-1.7	35.21	-26.8
AV-PROD-UTTER	—		35.43	-26.3
AV-PROD-LOCAL	—		37.15	-22.8

TAB. *8.1: Scores de reconnaissance audiovisuelle, WER%. On compare les trois systèmes de base, MS étant le modèle multistream synchrone, et AV PROD le multistream asynchrone. Nous montrons les gains en WER relatif par rapport au système de référence audio seul.*

la parole bruitée (3%). Notre hypothèse est que le système de pondération globale n'est pas vraiment approprié dans le cas d'interférences non stationnaires comme la parole interférente. Le gain avec le modèle asynchrone en parole bruité est nul. Les détails sont publiés dans (Glotin, Vergyri, Neti, Potamianos & Luettin 2001).

8.3 Estimation dynamique de l'efficacité relative des deux modalités à base de R

Nous retrouvons à travers la bibliographie une pondération basée sur l'estimation du RSB (Meier, W.Hurst & Duchnowski 1996, Jourlin 1998, Teissier 1999, Rogozan 1999, Heckmann, Berthommier & Kroschel 2001).

Notre hypothèse de base est que l'information de voisement est transmise en majorité par la modalité auditive 8.3.

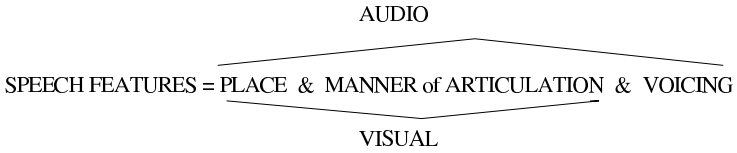


FIG. *8.3: Les trois composantes de la parole et leur appartenance aux différents flux.*

Une première fonction de pondération consiste alors à tirer l'exposant de fiabilité trame à trame suivant une fonction linéaire par partie (voir figure 8.4) de même type que celle de Meier (Meier et al. 1996), en utilisant la fonction identité entre les deux plateaux. Mais nous mesurons 8.3 qu'elle n'est pas efficace, même une fois optimisée sur l'ensemble de développement (c'est-à-dire les valeurs des plateaux ayant été

optimisées sur l'ensemble de développement).

Donc considérant que les indices de voisement sont convenablement estimés, il faut chercher une variante de cette fonction pour générer les poids.

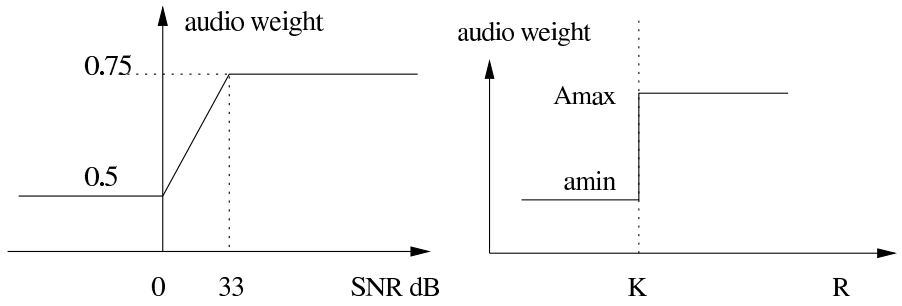


FIG. *8.4: Deux fonctions simples de pondérateurs locaux. A gauche estimation de poids Audio en fonction du SNR d'après (Meier et al. 1996). A droite notre fonction paramétrée d'après les caractéristiques ROC de notre détecteur de locuteur dominant qui donne le paramètre K, les paramètres amin, aMAX qui peuvent varier dans les limites* $0.5 < amin < aMAX < 1$. *En ordonnées est représenté l'exposant de la vraisemblance audio.*

	Err	
$\alpha = 0.7, \beta = 1 - \alpha$	35.21	
$minMax(0.5, 0.5, 0.7)$	37.15	

TAB. *8.2: WER pour alpha fixe et alpha d'après une fonction non optimisée de Meiers (test sur 1038 phrases, modèle asynchrone AV-PROD). Une dégradation est observée.*

$\lambda_{\mathrm{A}t} = min(max(R(t), 0.5), 0.7)$.

8.3.1 Pondération basée sur la détection du locuteur dominant

D'une façon générale, l'interprétation des poids sur les fonctions de vraisemblance est délicate. Au vu de nos résultats, et des fonctions usuelles, nous pensons que les fonctions de cartographie donnant la probabilité de bruitage pourraient répondre au problème. Nous proposons alors de revenir à la notion de probabilité de bruitage pour inférer les pondérateurs.

Les trois degrés de liberté sont le seuil K, et les amin, aMAX. On définit alors la fonction suivante :

$$\lambda_{At} = \text{amin si } R(t) < K \quad (8.2)$$
$$= \text{aMAX else} \quad (8.3)$$

Suivant les contraintes suivantes :

$0.5 < amin < 0.7$
et
$0.7 < aMAX < 1$

Nous choisirons K suivant les caractéristiques de spécificité (SP) et de sensibilité (SE) de la ROC présentée en Partie I.

On teste trois modes : hyper SE, hyper SP, et SE=SP.

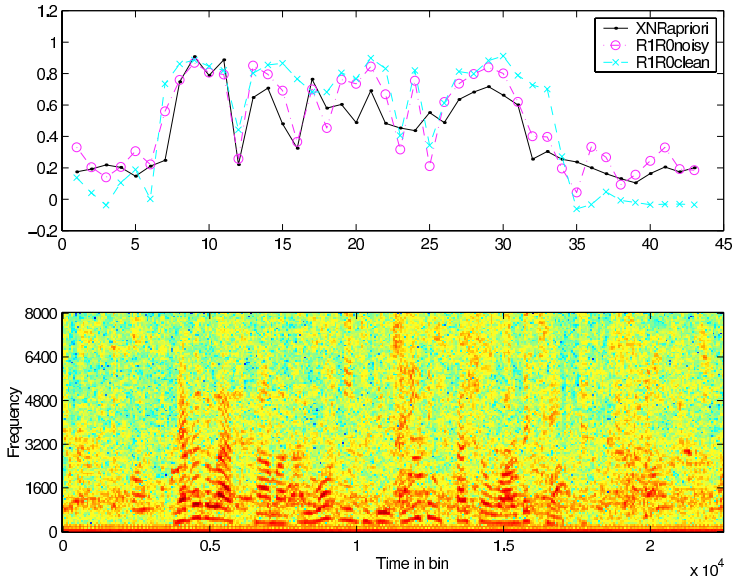


FIG. *8.5: En haut R sur parole bruitée (bruit de parole) ou claire. niveau de référence XNR en noir. Calculs faits sur trames de 128ms, décalées de 64ms. En bas le spectrogramme bruité par le bruit de parole à 8.5 dB.*

Les cartes en figure 8.6 montrent les variations de WER (delta WER%) par rapport au modèle de référence pour 20 phrases tirées au hasard de l'ensemble de test (soit environ 20 minutes, 300 mots). La performance de base est 36.84 % WER pour ces 20 phrases (Product model).

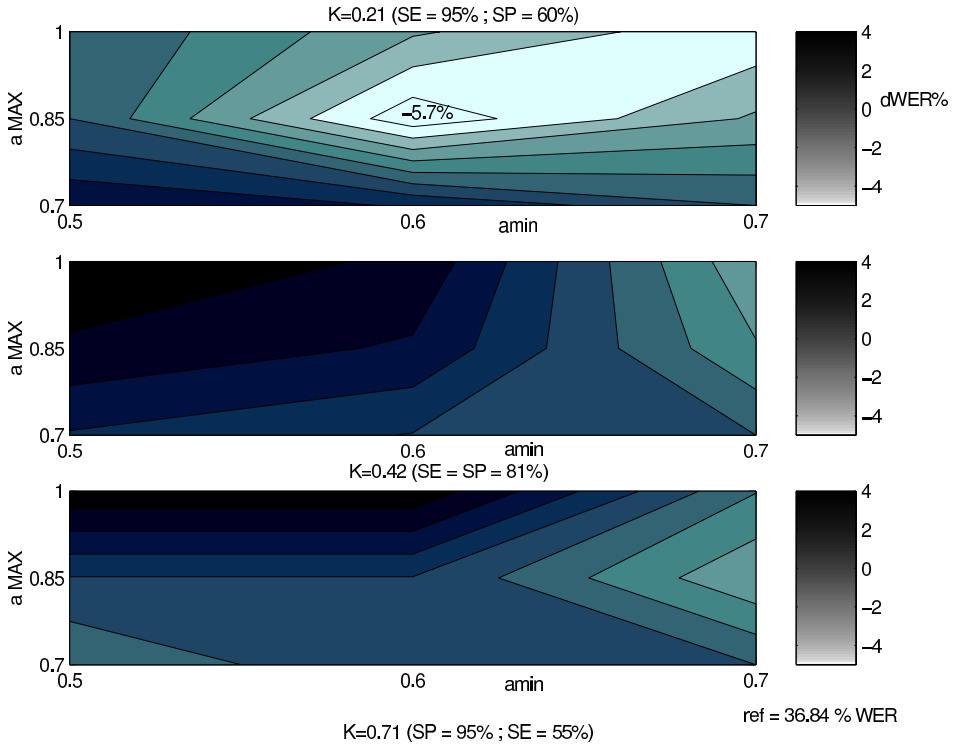


FIG. *8.6: Variations de WER (delta WER%) par rapport au modèle de référence pour 20 phrases tirées au hasard de l'ensemble de test (soit environ 20 minutes, 300 mots). La performance de base est 36.84 % WER pour le modèle produit pour ces 20 phrases (Product model avec amax=aMIN=0.7.). De haut en bas les 3 cas HSE, EQ, et HSP pour des couples variables (amin, aMAX) comparé aux performances du système le plus performant : le modèle produit : 36.84% WER.*

On voit sur la figure 8.6 que le meilleur gain est obtenu pour $K = 0.21$: delta WER de 5.7% (i.e. 34.74% WER) pour le couple amin aMAX (0.6, 0.85). Remarquablement le poids d'apprentissage 0.7 est au centre de ce couple.

8.4 Conclusion

Cette étude sur grand vocabulaire et interférence parole montre l'efficacité de la fusion par voisement des modalités audiovisuelles.

La validation de l'approche sur l'ensemble de test entier est en cours de traitement dans(Glotin 2001*a*).
Nous pensons que nous observerons un gain en robustesse de ce modèle guidé par

notre indice de voisement par rapport au modèle produit qui est le meilleur actuellement. En effet sur 10 minutes de signal aléatoire, la méthode permet de faire chuter significativement le WER, même contre du bruit de parole..

Cela montre que notre pondérateur apporte de l'information contextuelle de type signal audio, et que si cette information est convenablement fusionnée elle contribue à un gain de reconnaissance.

Nous allons maintenant essayer de généraliser ce résultat au cas du multi-bande, pour la même information de voisement ou pour un autre indice CASA de localisation.

Chapitre 9

Application des indices R et ITD au modèle 'Full Combination'

Après la description de la théorie et de l'implémentation des approches FC et FCA, les paramètres des systèmes hybrides HMM/ANN, nous testons ici le modèle «FCA» sur un large éventail de bruits et de SNR, en allant des bruits de bande(s) idéaux, stationnaires centrés sur une des d sous-bandes, ou changeant de fréquence centrale toutes les 125ms, ainsi que sur bruit de voiture et d'usine, ou dans le cas cocktail party. Nous testons ainsi les différents pondérateurs que nous avons présentés.

9.1 Pondération intrinsèque basée sur l'entropie des posteriors

Les posteriors peuvent être interprétés comme des facteurs de fiabilité. Nous avons dans ce qui suit pris des pondérations uniformes sur tous les flux afin de mesurer la robustesse íntrinsèque du modèle.

Ce type de modèle correspond a l'architecture représentée dans la figure 9.1.

Cet indice démontre encore que le SNR est couplé avec l'espace de reconnaissance, mais via des relations non triviales. Nous avons montré dans les expériences précédentes en reconnaissance partielle et dans (Glotin, Tessier, Bourlard & Berthommier 1998*a*) que l' entropie des posteriors est bien corrélée avec la présence de bruit en sous bande. Il est donc légitime de construire à partir de cette mesure un estimateur de fiabilité pour le FC.

Dans le cas de N classes phonétiques, nous proposons alors d'estimer :

$$P(C_i) = 1 - H(P(q_k|X_i))/log_2(N)$$

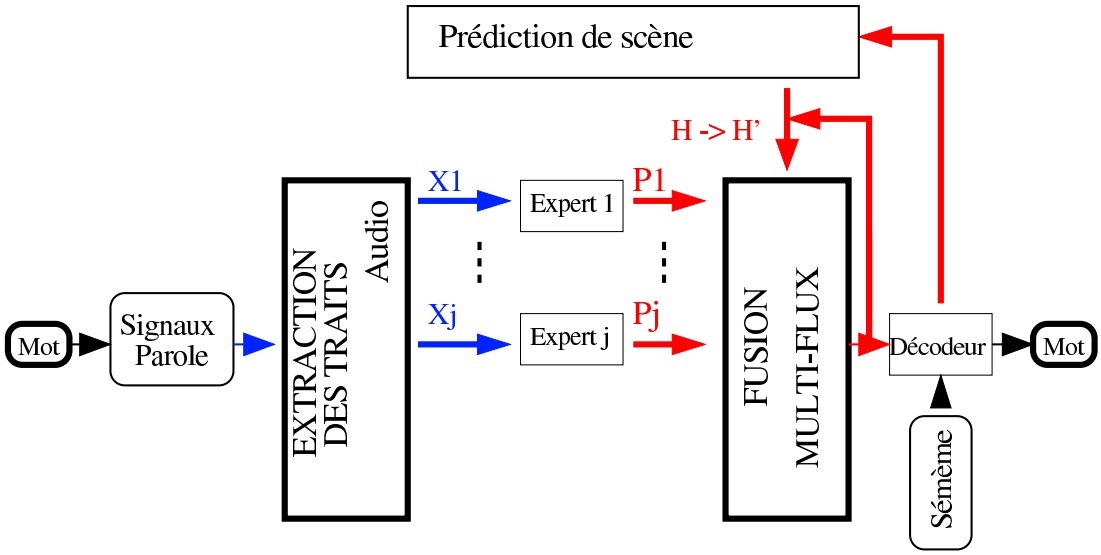


FIG. *9.1: Architecture de pondération intrinsèque : la fusion des flux d'hypothèses est contrôlée par les vecteurs d'hypothèses eux mêmes à travers une opération notée $H- > H'$ qui a lieu soit sans étape de décodage, soit après décodage et interprétation par un module de prédiction de scène*

où H est l'entropie log_2 du vecteur de sortie du MLP. $P(C_i)$ est l'efficacité du code.

Cependant nous voyons tout de suite que cette estimation n'apporte pas d'information extérieure au MLP, mais sa caractéristique temporelle (très locale) fait qu'elle reste intéressante en bruits fortement non stationnaires.

9.2 Indice de fiabilité basé sur une estimation du SNR

Le schéma général du système de reconnaissance par pondération extrinsèque est donné par la figure 9.2.

Plusieurs algorithmes d'estimation du SNR sont disponibles (Ris & Dupont 2001). Dès lors il est possible de dériver une estimation de l'indice de fiabilité par un centrage et une normalisation de la valeur SNR comme l'a proposé (Hirsch & Erlicher 1995) :

$$P(C_i|X) = \frac{SNR(X) - SNR_{min}(X)}{SNR_{max}(X) - SNR_{min}(X)}$$

SNR_{min} et SNR_{max} peuvent être fixés où bien estimés sur les 600 dernières ms du signal.

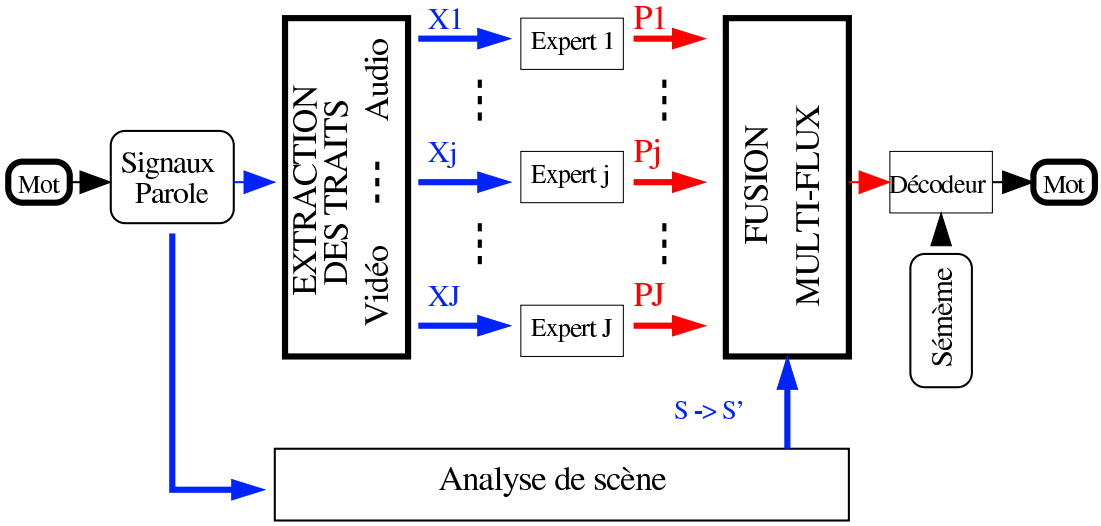


FIG. *9.2: Architecture de pondération intrinsèque : couplage FC+CASA*

Notre thèse ne portant pas sur l'élaboration d'un estimateur de SNR, nous nous plaçons dans le cas d'un estimateur idéal : nous utilisons les niveaux SNR connus a priori. Nous montrons que ces pondérations ne sont pas aussi efficaces qu'espéré.

Cette technique présente deux inconvénients. Tout d'abord l'estimation du SNR demande une analyse de l'ordre de 500 ms. Le second problème plus général est discuté dans la section suivante

9.3 Fonctions stochastiques de fiabilité du signal

Nous disposons (Voir partie I) des densités de probabilité (pdf) pour chacun de nos observables de la forme :

$$P(SNR_i = T_i | Observable)$$

Avec la règle de Bayes et après calcul des fonctions cumulatives (cdf), il nous est alors possible de tirer pour chaque sous bande i :

$$P(C_i) = Pclean = P(SNR_i > T_i | Observable)$$

le choix de T_i est un point clef de la méthode de cartographie (mapping).

Nous proposons et testons différentes manières d'optimiser ces seuils.

Le seuil T_i est à la charnière entre la détection (labelling) et l'identification (reconnaissance). Plus le seuil est bas, plus la détection est facile et l'identification difficile, et vice versa.

Notre cible étant le WER, nous choisirons une optimisation globale sur les différent modèles de reconnaissance .

9.3.1 Fonctions de l'indice d'harmonicité R

Un intervalle de R_i est assimilé à une valeur ponctuelle de R_i (une centaine de d'intervalles de même largeur sur [-0.2 1] garantissent une précision suffisante) : $P(SNR_i = T_i|R_i, X)$
avec X le vecteur acoustique contenant les traits du pavé de la sous bande i et SNR_i son SNR.

Nous représentons dans la figure 9.3 la carte normalisée associant la variable R et le SNR local en bande 2.

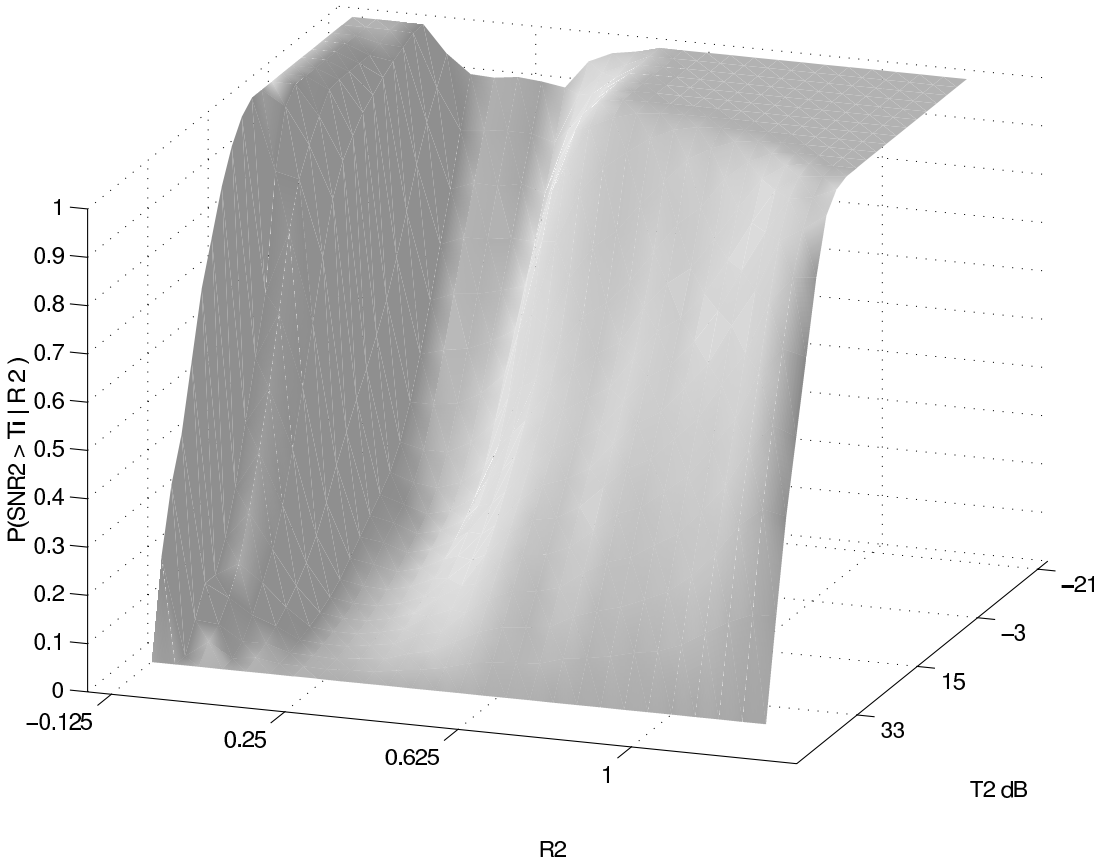


FIG. *9.3: Fonction de bruitage en bande 2 pour seuil variable. Cette carte normalisée associe la variable R et le SNR local en bande 2*

Puis nous calculons de ces pdf les cdf (cumulative density function) pour chaque

valeur seuil T_i de SNR_i :

$$P(C_i|X) = P(SNR_i > T_i|R_i, X)$$

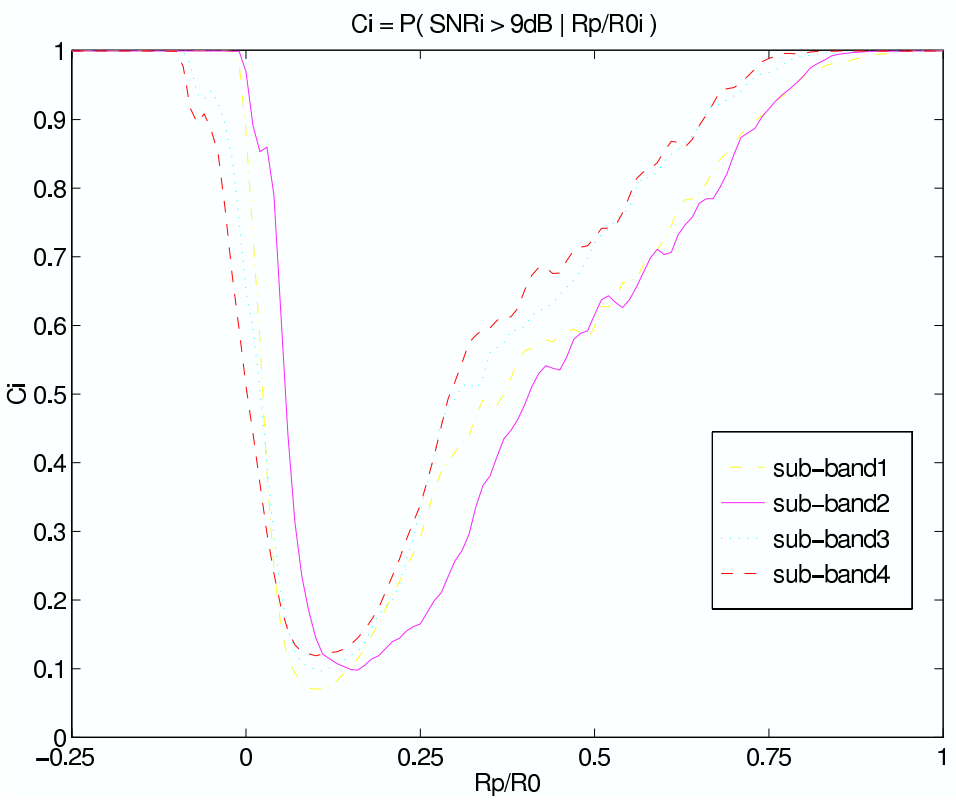


FIG. *9.4: $P(C_i)$ des 4 sous bandes pour Ti=9dB*

La figure 9.4 représente $P(C_i)$ des 4 sous bandes pour un seuil T=9dB. Chaque fonction de fiabilité ainsi obtenue est similaire à travers les sous bandes à une translation près.

En effet le minimum de la fonction est atteint pour la valeur R1R0 du bruit de bande (voir annexe), et cette valeur est légèrement décalée suivant les sous bandes.

Nous représentons en figures 9.5, 9.6 les fonctions de fiabilités pour les 2 premières sous bandes et différents seuils. Le cas des autres sous bandes sont similaires.

Une méthode pour la pose de seuil consisterait à établir l'entropie des posteriors et le SNRref - pour chaque sous bande - pour chaque fenêtre, et de comparer ces mesures avec

$$C_i = P(SNR_i > Ti|Obser)$$

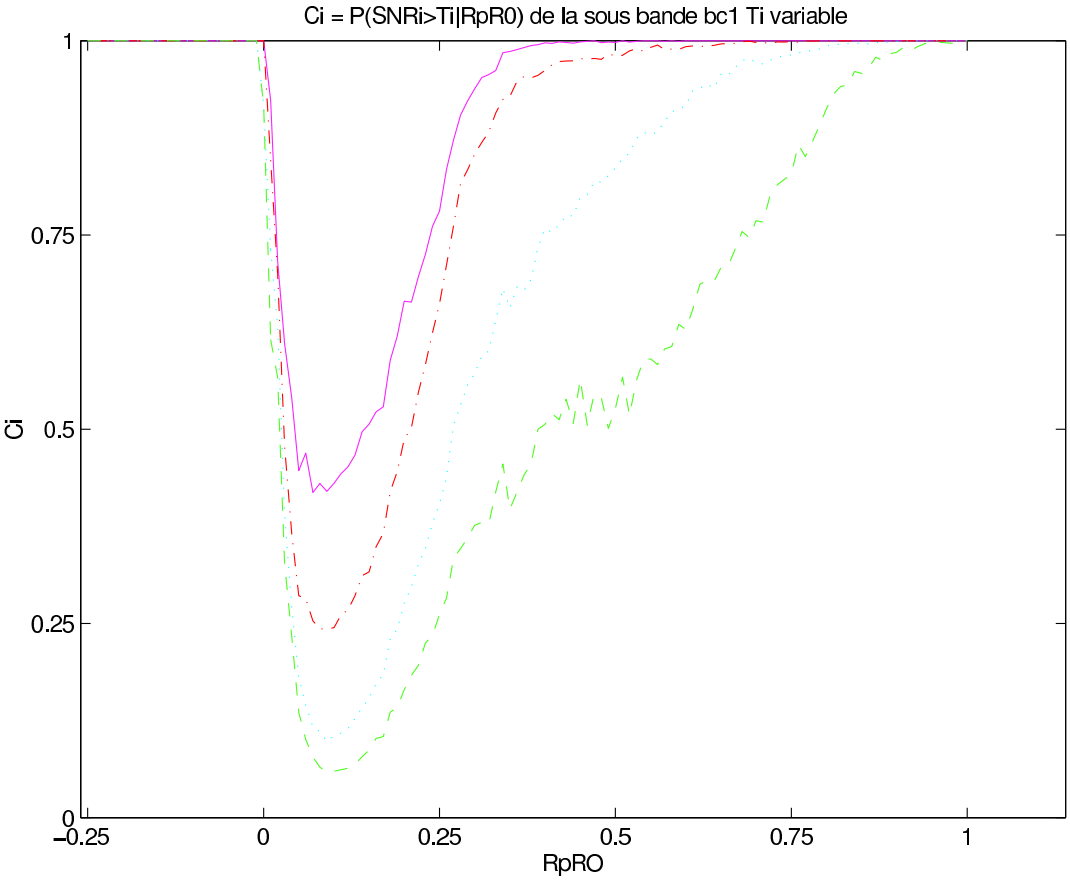


FIG. *9.5: Fonction de fiabilité pour sous bande 1, T_1 = -6,0,6,12 dB (de haut en bas)*

en fonction de T_i Cet indice mesure l'information disponible localement, mais le critère de WER n'est plus suivi, ce qui conduirait à un optimum local.

L'exploitation de ces fonctions et l'impact du choix des seuils seront détaillés dans les chapitres suivants.

La variabilité de la réponse R à différents bruits n'est pas forte étant donnée l'étape de démodulation avant extraction de l'indice qui a pour effet de 'normaliser' la réponse aux bruits non harmoniques/harmoniques (car leur fréquence fondamentale est hors du domaine du pitch).

Nous illustrons dans la figure 9.7 le comportement des C_i dans chaque sous bande en relation avec les $R(i)$.

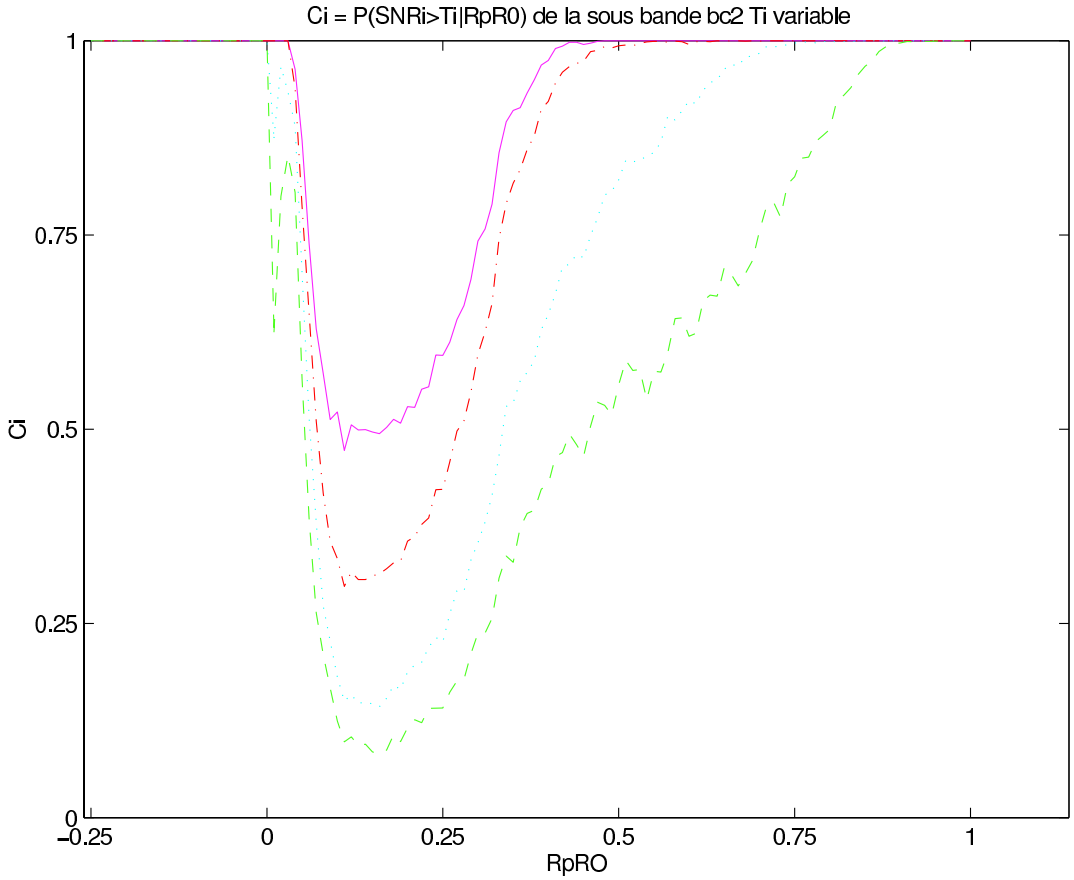


FIG. *9.6: Fonction de fiabilité pour sous bande 2, T_1 = -6,0,6,12 dB (de haut en bas)*

9.3.2 Fonctions de l'indice ITD

Les mêmes types de calculs sont faits.

Nous représentons l'histogramme en bande 2 et sa cdf d'où nous extrayons la fonction de confiance sur les pavés Temps fréquence, fonctions qui pondère le FC.

Nous posons les mêmes seuils de performance et dressons les fonctions de fiabilité correspondantes en figure 9.9.

9.3.3 Fusion avec un système sous bande

L'inconvénient des indices basés sur une estimation du SNR est qu'ils nécessitent une analyse sur des trames de l'ordre de 500 ms (Ris & Dupont 2001, Martin 1993) estimateur de fiabilité est qu'il intègre l'information à l'échelle de l'estimateur (dans notre cas un MLP avec un contexte de 125ms). Une analyse robuste d'harmonicité ne

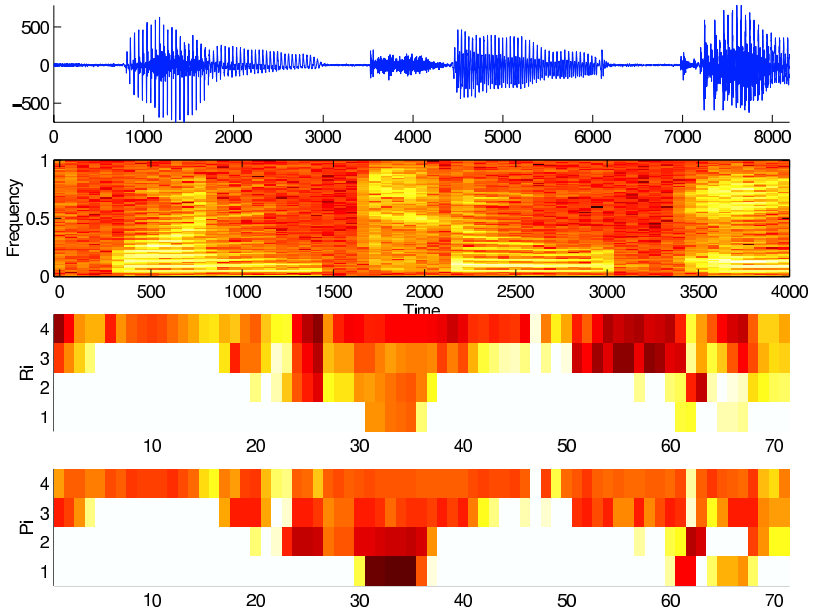


FIG. *9.7: Corrélation entre le SNR et R (et les facteurs de fiabilité dérivés)*

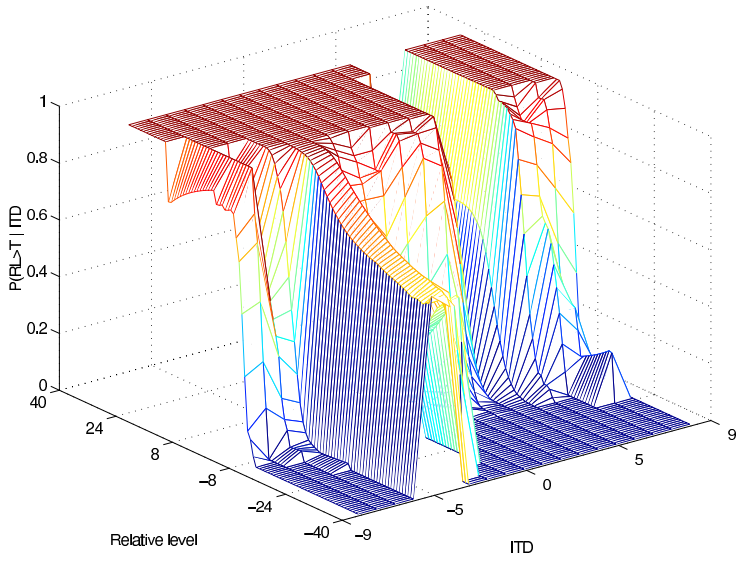


FIG. *9.8: Densité cumulative en bande 2*

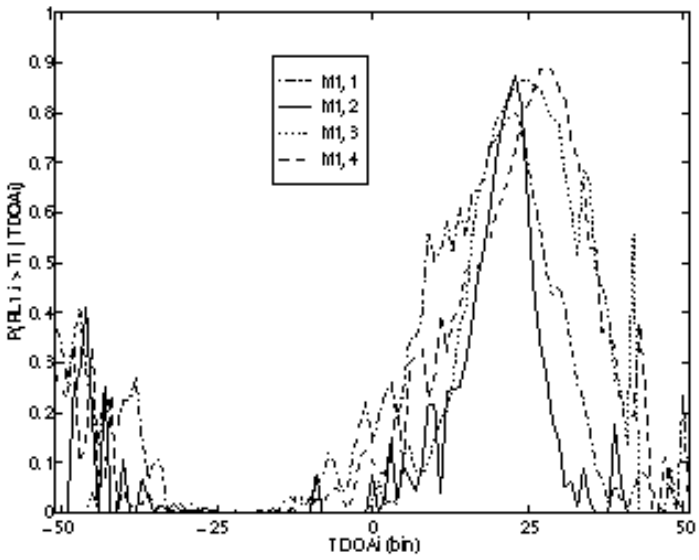


FIG. *9.9: Les 4 courbes de fiabilité dérivées de l'ITD dans chaque sous bande, pour les seuils respectifs [12,9,9,9] dB*

nécessite que quelques périodes de la fréquence fondamentale (Boite et al. 2000, Hess 1983). Avec une fréquence d'échantillonnage de 8Khz, et une fréquence fondamentale de 200Hz, 100 ms de signal permettent déjà d'intégrer 4 périodes du fondamental. Nous voyons donc qu'un indice basé sur cette mesures serait adapté au potentiel de la pondération dynamique du FC.

Nous noterons C_i l'événement 'le pavé temps fréquence de la sous bande i génère une estimation fiable du vecteur de posteriors'. Nous allons donc chercher à estimer les $P(C_i)$ à travers la mesure d'harmonique en relation avec le SNR local de chaque pavé temps fréquence. Nous comparerons cet indice avec des indices de fiabilité SNR estimés et issus de l'entropie normalisée des posteriors de chaque pavé temps fréquence.

La fusion de cette information avec le FC est immédiate à condition de faire l' hypothèse que l'événement C_i précédent est équivalent à l'événement L_s = 'Le s^{th} flux est le flux de parole claire le plus large parmi tous les flux possibles'

Cette approximation ne prend pas en compte les non linéarités générées par le pré traitement (dans notre cas Jrasta) et nous verrons plus tard qu'il sera avantageux de changer de référence.

Une seconde approximation est nécessaire pour évaluer les $P(C_i)$ à partir d' un ensemble de mesure restreint : nous admettons que les C_i sont indépendants à travers les sous bandes (ce qui est correct vis à vis d' interférences aléatoires, mais surtout de sous bandes à faibles recouvrement (Steeneken & Houtgast 1991)).

Nous avons alors pour un flux S, avec les sous bandes numérotées de 1 à 4 :

$P(L_s|X) = \prod_{i \in S} P(C_i|X).\prod_{i \notin S}(1 - P(C_i|X))$

Remarquons que ces fonctions de pondération ne nécessitent pas de détecteur de silence. Par contre le point sensible reste la pose du seuil T_i pour chaque sous bande. Dans les sections suivantes nous étudions les taux d'erreur du FC en fonction de différentes valeurs seuils.

9.4 Application de R au FC

Après avoir présenté le modèle FC, nous proposons diverses pondérations, classiques basées sur le SNR, ou bien basées sur une fonction de l' indice harmonique
.

L'inconvénient des indices basés sur une estimation du SNR est qu' ils nécessitent une analyse sur des trames de l' ordre de 500 ms (Ris & Dupont 2001, Martin 1993) estimateur de fiabilité est qu'il intègre l'information à l'échelle de l'estimateur (dans notre cas un MLP avec un contexte de 125ms). Une analyse robuste d'harmonicité ne nécessite que quelques périodes de la fréquence fondamentale (Boite et al. 2000, Hess 1983). Avec une fréquence d'échantillonnage de 8Khz, et une fréquence fondamentale de 200Hz, 100 ms de signal permettent déjà d'intégrer 4 périodes du fondamental. Nous voyons donc qu'un indice basé sur cette mesure serait adapté au potentiel de la pondération dynamique du FC.

9.4.1 Optimisation des fonctions de fiabilité

Le seuil Ti est à la charnière entre la détection (labelling) et l' identification (reconnaissance). Plus le seuil est bas, plus la détection est facile et l'identification difficile, et vice versa. les seuils étant en SNR local au pavé TF.

Seuillage sur bruit gaussien

Une première approche consiste à optimiser sur du bruit blanc les seuils T_i. Tout d'abord nous prenons les mêmes T_i pour les 4 sous-bandes. Nous varions le SNR de -18 à +18 dB par pas de 6dB, et les Ti de 12 à -12dB par pas de 3 dB. Nous avons donc 7 courbes. Nous observons qu'elles présentent toutes un minimum pour le seuil le plus faible 9.10.

Pour des seuils aussi faibles les probabilités pour chaque bande $P(C_i)$ tendent vers 1, avec C_i l'événement "la bande i est propre". Soit L_j l'événement : "le stream j est le plus large flux de données propres", nous avons alors $P(L_{1234})$ qui tend vers 1, et 0 pour les autres flux. Cela correspond en effet à l'optimum en reconnaissance sur du bruit blanc : le reconnaisseur pleine bande Jrasta est le meilleur expert dans ces conditions.

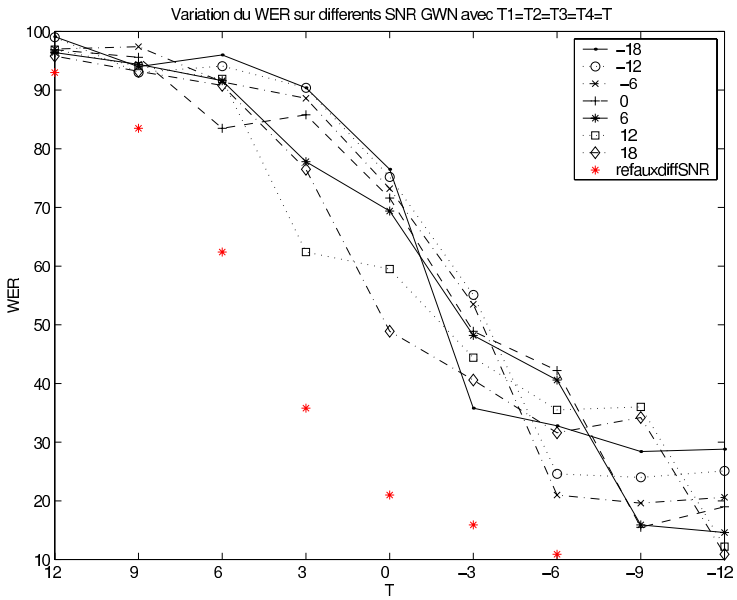


FIG. *9.10: WER suivant les seuils Ti uniformes sur du bruit blanc à différents SNR, * donne les scores de -18 à +18 dB du FCA uniforme*

Afin de ne plus dépendre de la structure du bruit pour la pose des seuils, nous avons dans les mêmes conditions fixer 3 des seuils à 1 et fait varier l'autre. les résultats sont en figure 9.11 pour la bande 1. Ils montrent que le même comportement que précédemment est observé. Ils se généralisent sur les autres sous-bandes.

Seuillage indépendant des fonctions sous bandes

Pour résoudre ce problème de convergence vers la pondération de l'expert pleine bande, nous avons fixé les poids de 3 bandes claires à 1 et étudié la dégradation à différents niveaux dB de la reconnaissance avec du bruit confiné dans la bande i, en variant T_i. Le bruit est un bruit blanc filtré, centré dans la bande considérée et coupé pour une largeur de 300 Hz. Cette méthode peut rendre dépendante les seuils à ce type de bruit de bande, mais notre seul critère est de rentre incompatibles les données d'entraînement et de test, ce que nous réalisons avec cette interférence, qui compte tenu de la largeur des sous-bandes correspond à une dégradation globale des pôles LPC.

Les comportements des différentes courbes de -18 à 18dB par pas de 6dB pour un même bruit de bande sont homogènes. En faible dB, nous remarquons que les seuils faibles sont favorisés, car ils favorisent le reconnaisseur pleine bande (full band) Jrasta est qui un des meilleurs experts pour ces niveau SNR. Nous avons représenté de gauche à droite de -18 à 18 dB, par des * les performances du FCA avec des poids uniformes, et par des + les performances du FCA pour un bruit en

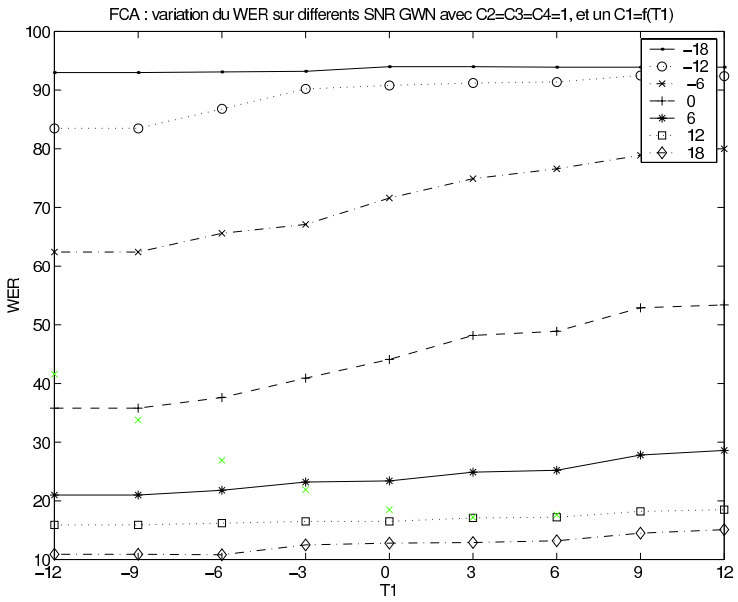


FIG. *9.11: WER suivant les seuil T1 variable les autres étant fixés à 1 sur du bruit blanc à différents SNR, * donne les scores de -18 à +18 dB du FCA uniforme*

sous bande i avec Ti=0 et les autres poids à 1 (condition idéale). Nous remarquons que dans nos conditions de test 'semi-guidée' notre modèle se situe entre le FCA et la condition idéale, et pour des seuils autour du seuil optimal, le modèle semi guidé est meilleur que le modèle guidé. Ce comportement confirme que la cartographie (mapping) apporte de l'information pertinente au modèle pour ce type de bruit.

Pour chaque sous bande, nous mesurons sur les courbes des moyennes des différents niveaux dB de -18 à 18 par pas de 6, un optimum en WER, qui est pour chaque sous bande respectivement :

T1 = 9 dB
T2 = 18 dB
T3 = 9 dB
T4 = 9 dB

Il est intéressant de noter que ces optimums sont corrélés avec les performances WER et de taux de reconnaissance des trames des experts sous bandes isolés en clair qui sont respectivement 61.5 % , 67.5 % , 63 % , 54.5 % .

Mais ces seuils ne sont valides que dans le cas des bruits colorés, autrement c'est un seuil de l'ordre de 0dB qui est optimal, favorisant le reconnaisseur pleine bande.

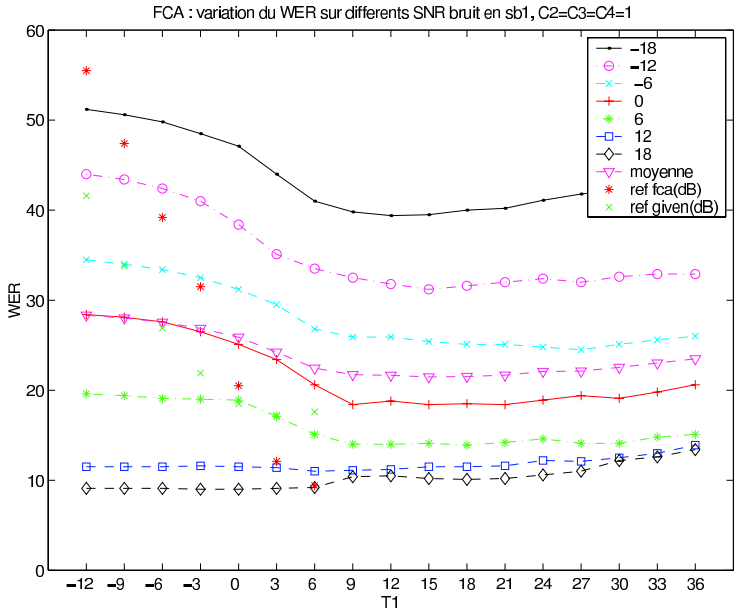


FIG. *9.12: WER suivant le seuil en band1, bande 1 bruitée à différents SNR,* $P(C1) = P(SNR1 > T1|obs)$ *, autres* $P(cj) = 1$*. * sont de gauche à droite les scores de -18 à +18 dB du FCA uniforme, + idem pour le FCA guidé :* $P(C1) = 0$ *autres* $P(cj) = 1$*.*

Nous choisirons donc les seuils suivant les tolérances aux bruits de chaque sous bande prise à 10%, ce qui donne [12 9 9 9] dB.

9.4.2 Discussion sur les seuillages des fonctions de fiabilité

La variation en fonction des seuils n'est pas bien marquée, ce qui invalide l'existence de seuil, et donc invaliderait cette fonction d'étiquetage. Les scores en bruits naturels sont moins bons comparées au modèle 'aveugle' avec des poids uniformes.

Mais un problème plus fondamental est mis en évidence ici : les performances du Jrasta pleine bande en bruit large bande est meilleur dans tout les cas que le FC. La pondération des experts sous bande doit tenir compte de la robustesse due au pré traitement, nos MLPs et notre modèle FC/FCA qui est par construction robuste au bruit de bande.

Malgré l'optimisation globale sur le critère du WER, chaque P(SNRi>Ti) est caractérisé par la réponse r1r0 du bruit blanc filtre en bande i. La technique de la cartographie ne généraliserait donc pas les réponses du reconnaisseur FC à d'autres bruits, large bande, de type bruits naturels. Il faudrait construire les fonctions de

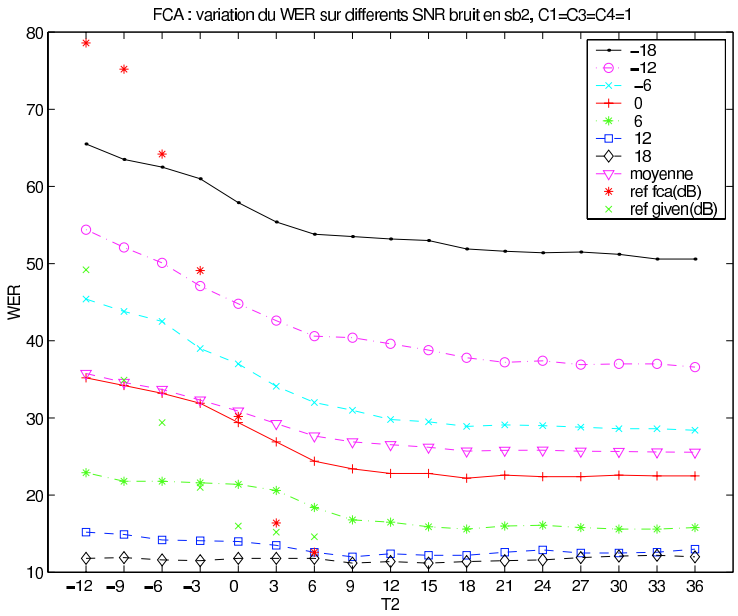


FIG. *9.13: WER suivant le seuil de bande 2, bande 2 bruitée à différents SNR,* $P(C2) = P(SNR2 > T2|obs)$ *, autres* $P(cj) = 1$. * *sont de gauche à droite les scores de -18 à +18 dB du FCA uniforme, + idem pour le FCA guidé :* $P(C2) = 0$ *autres* $P(cj) = 1$.

fiabilité pour le spectre pleine bande afin de donner une réponse cohérente au bruit large bande. Ou plus généralement construire les fonctions de fiabilités pour chaque combinaison du FC, en association avec un critère propre aux performances des posteriors (et non avec le SNR), afin d'intégrer les caractéristiques du modèle FCA et de son pré traitement. Nous reviendrons sur ce point plus tard.

Si la méthode par cartographie s'avère cependant dépendante de la variabilité de la réponse r1r0 au bruit, nous pouvons réduire la largeur des sous-bandes. En effet plus la sous bande est large, plus la méthode est spécifique au bruit utilisé pour dresser la cartographie. Une autre perspective du modèle FC en bruit naturel est donc vers une augmentation du nombre de sous bande[1].

La binarisation (discrétisation par seuillage à 1/2) des fonctions pour effectuer une reconnaissance partielle uniformise les mappings par rapport à la diversité des fonctions continues que l'on peut constater sur les figures à différents seuils. La reconnaissance partielle est donc moins sensible à la variabilité des seuils, elle ne peut donc pas être le support d'une pose des seuils T_i.

[1]Nous avons dressé les mappings pour chaque bandes critiques et il s'avère qu'ils restent pertinents jusqu'à 8 sous bandes au moins, surtout en HF où les bandes critiques sont plus larges.

9.5 Expériences de reconnaissance et comparaisons des pondérateurs

Nous reproduisons dans cette table les résultats fondamentaux de ce chapitre. Nous y retrouvons les méthodes binaires ou douces (continues) pour des pondérateurs dérivés du SNR relatif, de l'entropie.

Dans le cas des fonctions douces tirées des harmoniques, les fonctions pour les 4 sous bandes sont données par :

$P(C_i|x_i) = P(SNR_i > T_i|R_i, x_i)$

avec $T1 = 12; T2 = T3 = T4 = 9dB$

La fonction dure est facilement extraite avec :

si $P(C_i) > 1/2$ et $P(C_i) = 1$ sinon 0.

La fonction Srel a déjà été présentée. Les fonctions "sqR" consistent à utiliser $P(C_i) = \sqrt{}(max(R, 0))$ dans le cas "soft".

Tous les traits sont tirés de Jrasta. Nous vérifions que le système Jrasta est équivalent au système de référence de Soustraction spectrale mis au point par le laboratoire expert indépendant du LTS de Mons. Le Jrasta est dans le cas particulier d'interférence non stationnaire très peu performant par rapport à la soustraction spectral, en effet le Jrasta pose l'hypothèse de stationnarité des interférences.

Nous observons un gain en WER pour le seul cas de bruit en bande étroite, stationnaire ou non, mais toutes les méthodes échouent face au prétraitement Jrasta pour le spectre entier.

9.6 Reconnaissance robuste sur deux sources simultanées : Pondération/Etiquetage par indice ITD

9.6.1 Objectifs de l'étude

Nous généralisons le principe d'étiquetage CASA/RAP sur le paradigme de reconnaissance de deux paroles concurrentes, enregistrées en stéréo. L'indice le plus propice est alors la localisation de la source dominante à travers l'intercorrélation des sources. Nous validons tout d'abord l'usage de cet indice à travers un protocole de rehaussement direct sur les spectres, tout comme nous l'avions fait avec l'indice d'harmonicité. Puis nous utilisons la technique de base de cartographie pour générer les fonctions de fiabilité et exécuter notre algorithme FC. Dans ce cadre de travail

	gwn	fact	car	narb1	narb3	n.st
fband est.	38.2	37.8	33.7	26.6	30.8	90.6
SS est.	33.7	41.0	35.6	29.2	38.4	56.3
blind	46.9	45.6	44.2	24.5	21.7	49.9
SH	47.6	45.7	44.6	23.5	20.2	51.3
Srel	47.2	45.0	43.9	21.4	20.1	51.9
SsqR est.	47.7	45.6	44.8	28.9	20.4	49.6
Spro est.	47.3	45.0	45.0	27.1	19.3	59.8
BsqR est.	60.9	57.6	53.6	45.9	30.6	67.1
Bpro est.	58.1	54.8	51.7	34.8	23.5	66.1

TAB. *9.1: Word Error Rate (WER) pour méthodes de références fband = JRASTA pleine bande, SS :soustraction spectrale. Les méthodes de pondération qui suivent servent soit de référence avec les indices donnés à priori, soit servent à tester des modèles (estimés). Méthodes de pondération douce : blind=FC avec poids uniformes, SH=pondération par entropie, Srel=SNRrelatif priori, SsqR est.=racine de R estimée, Spro est.=P(ci) estimé. Fonctions binaires : BsqR est.=racine de R estimée, Bpro est.=P(ci) estimé.*
*Scores donnés en moyenne sur 200 phrases * 6 niveaux de bruits -12 à 18db par pas de 6dB. Col : Gaussien White Noise, factory, car, narrow banb 1 et 3, nonstation. noise. Intervalle de confiance = +/-1 à WER=20%. La reconnaissance partielle dans les cas de narb1 (ou narb3) donne 22.7 (ou 19.0) %WER +/- 1%. Voir annexe B.2 pour les intervalles de confiance.*

la bibliographie est toujours très riche, et nous comparons donc nos méthodes avec la technique de séparation aveugle qui à récemment été adaptée au cas de mélange convolutif (Taniguchi, Kajita, Takeda & Itakura 1998, Yen & Zhao May 1998, Yen & Zhao Mar. 1999*b*, Yen & Zhao 1999*a*, Choi, Lyu, Berthommier, Glotin & Cichocki 1999, Hong, Choi, Glotin & Berthommier 2000, Choi et al. Sept 2001).

Dans une optique de simplicité et de robustesse à l'utilisation, nous proposons une nouvelle architecture d'intégration de l'information de localisation, architecture de CASA labelling. Ce modèle sera testé et comparé aux deux précédemment cités qui ont été mis en place testés sur le même corpus.

Nous avons présenté les techniques de rehaussement au début de la partie II.

Les techniques de séparations aveugle, d'ICA, sont classiques dans ce domaine. Nous avons comparé nos résultats avec l'une d'entre elles (voir (Choi et al. 1999)).

Nous reprenons le même principe de pondération du modèle FC, mais ici les poids sont évalués suivant d'autres cartographies, découlant des délais mesurés dans

modèle	WER %
REFERENCE Jrasta	73
Blind separation	65
FC aveugle	76
FC + ITD	49
FC + non(ITD)	92
ITD* spectre	54

TAB. *9.2: Résultats de double reconnaissance pour différentes méthodes, Jrasta, ICA classique, divers FC, rehaussement spectral. La meilleure méthode reste le FC + CASA*

l'intercorrélogramme des 2 sources.

9.6.2 Généralisation des seuils en ITD

Le seuillage de la cartographie Parole/bruit est t-il généralisable en condition "cocktail party" ? Comme les sources sont symétriques les seuils pour chaque source doivent être opposés. Suivant la robustesse du reconnaisseur sous bande aux interférences due à une source concurrente, le seuil peut être plus ou moins éloigné de la condition de mélange uniforme (0dB). Des tests sont réalisés à différents seuils pour mesurer la sensibilité au seuil et déterminer le seuil optimal, voir section "cocktail party".

9.6.3 Résultats comparés

9.7 Conclusion

Dans le cas de 2 sources simultanées, enregistrées avec 2 micros à partir de Numbers95, le modèle FC et le modèle direct décrit dans le cas monophonique restent compatibles, via une cartographie ITD/relative level, ou une fonction de partage dépendant de l'ITD (dans notre étude une sigmoïde). Nous avons montré que face à des techniques de référence (séparation aveugle), l'usage de cet indice dans nos modèles est plus performant que la technique classique de séparation aveugle testée (elle même ayant des performances comparables à (Taniguchi et al. 1998))
La pondération de flux uni-modaux (audio) par l'indice de voisement contre des interférences confinées en fréquences, stationnaires ou non, est une réussite par rapport aux systèmes Jrasta ou soustraction spectrale. Cependant la pondération de flux uni-modaux (audio) par l'indice de voisement contre des interférences large bande échoue, nous allons traiter de ce problème dans les chapitres suivants.

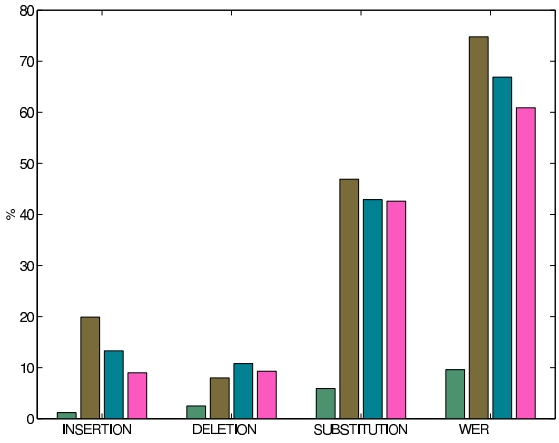


FIG. *9.14: Détail de résultats en séparation aveugle sous Jrasta, barres de gauche à droite : ST une seule voix (gauche), ST mélange sur voix gauche, ST mélange + BS, idem avec seuillage des énergies résiduelles*

Quatrième partie

Modèle multi-flux avec Prédiction des Biais de reconnaissance et fusion des indices de fiabilité du signal

A travers la partie III, nous avons montré que l'indice de voisement apporte une information pertinente au système pour le contrôle de la fusion précoce, ou tardive dans le cas de la pondération multi-modale. Le FC est bien corrigé dans le cadre du "cocktail party", mais nous constatons de faibles performances dans le cas de bruit large bande. Notre interprétation est que si deux cibles sont concurrentes, le système fait rarement de fausses détections positives. Au contraire, dans le cas de cible unique et d'interférences large bande, le système génère des fausses détections positives. Le but de cette partie est d'analyser les erreurs génériques des reconnaisseurs combinatoires dans le cas multi-bandes, et de proposer un nouveau modèle de fusion pouvant intégrer ces divers comportements dépendants du SNR, des classes phonétiques et de la région spectrale observée. Nous allons raffiner cette technique en réintégrant l'information de fiabilité du signal (liée au voisement par exemple) qui est alors traitée conjointement avec les taux de bonne classification des experts.

Chapitre 10

Analyse des erreurs du modèle FC multi-flux

10.1 Introduction

Nous avons montré que nous possédons des indices de fiabilité du signal qui ont un fort potentiel en terme de détection des pavés temps fréquence constitués de parole propre ou bruitée.
Nous avons vu qu'ils peuvent gérer efficacement le rehaussement du signal, ou la fusion de deux sources d'information parole (audio et visuelle ou bien information provenant de deux microphones distants).

Nous avons proposé un modèle de fusion intra-modal, ou "multi-bandes", dit "Full combination" ou combinaison complète. Ce modèle FC est peu coûteux en terme de calcul grâce à l'approximation de chaque reconnaisseur combinatoire qui est tout aussi performante que la version ne comprenant que des modèles entraînés. Les performances du FC sont meilleures que les modèles multi-bandes précédents, mais face à des bruits large bande, elles sont peu compétitives part rapport aux techniques éprouvées de soustraction spectrale. Ceci montre que l'approche multi-bandes en est encore à ses prémisses.

10.2 Analyse des erreurs du FC

L'étude précédente montre que le critère MAP ne permet de calculer les fonctions de pondération que lorsque le phonème est le phonème cible. L'inconvénient est donc de ne pas apprendre les comportements de phonèmes concurrents, qui peuvent se "discriminer".

Les analyses précédentes ont montré les points suivants

-l'indice R, simple à calculé, local (125 ms suffisent et bien corrélé avec le SNR, c'est donc une variable signal adéquate.
-chercher la cible dans l'espace des posteriors de chaque stream j et non le SNR du signal (du fait du filtrage JRASTA).
-La cible est la posteriors en parole propre pour le phonème considéré, et non l'unité, car nous n'entraînons pas un système mais nous en pondérons les sorties pour leur retirer le biais généré en condition bruitée.
-chercher les cibles pour chaque phonème, soit construire pour chaque phonème et chaque stream j les fonctions $B_{jk}(R)$.
-pour éventuellement regrouper les fonctions par groupes de phonèmes suivant les similitudes des B_{jk}

10.3 Mises en évidence des biais des estimations

Nous représentons tout d'abord les posteriors des phonèmes sur une phrase de test en parole claire, comparé au cas de la parole bruitée. Nous observons qu'à faible SNR local (rappelons que 16 dB peuvent séparer un SNR global du SNR d'un pavé Temps fréquence) certaines classes sont surestimés, fricatives, silence ...

Une analyse plus fine montre la distance euclidienne entre les posteriors de chaque classe dans le cas de la parole claire et de la parole bruitée à un certain SNR :

$$d = |P(qk|Xiclean) - P(qk|Xinoisy)|$$

Cette distance, comprise entre 0 et 1, est nulle dans le cas d'une reconnaissance idéale. En cas de non reconnaissance du phonème cible, elle est forte, et tend vers 1 dans le pire des cas. Notons que cette distance peu être forte même dans le cas où le phonème n'est pas cible (cas d'une fausse alarme).

Nous représentons ces distances obtenues sur du bruit blanc à différents dB (-12, 0, 12 et 24 dB global) sur 100 phrases de Numbers95, pour le flux de la bande 2 (les 4 autres flux élémentaires sont données en annexe H). On voit que la reconnaissance de la classe silence et de certaines fricatives ou autre phonèmes à spectre plat ont un comportement très singulier par rapport aux autres démontrant un très faible résistance au bruit. Au contraire la plupart des voyelles sont très robuste, la distance restant très faible pour tout SNR local.

10.4 Mises en évidence de la corrélation entre R et les Biais des posteriors.

Nous observons en détail qu'à fort SNR les distances définies précédemment sont faible. Comme notre indice R est fortement corrélé avec le SNR, il nous semble

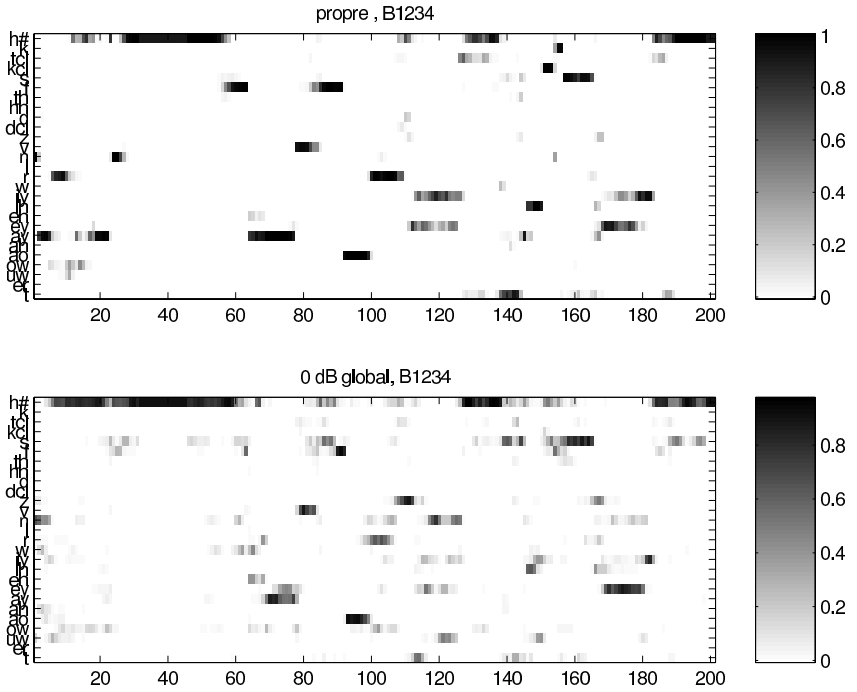


FIG. *10.1: Suivi des posteriors sur 3 phrases, propres ou bruitées : les classes silence, s et t sont surestimées, alors que les voisées sont sous-estimées*

intéressant d'observer la corrélation R et une distance de ce type. Nous modifions légèrement la définition de notre distance par la distance entre le maximum du vecteur des estimées en parole claire (= posterior du phonème cible) et la valeur obtenue pour l'estimation de ce phonème cible dans du bruit.

Dans la figure 10.3 nous montrons le bi-histogramme dans le cas de la sous-bande 2, entre R et cette distance obtenue sur du bruit blanc à différents dB (-12, 0, 12 et 24 dB global), sur 100 phrases de Numbers95. Noter la forte corrélation entre R et la distance.

Ceci confirme que R peut donner une indication sur le taux de biais de l'estimation d'une posterior.

10.5 Conclusion

Nous avons mis en évidence dans ce chapitre que l'estimation des posteriors de certains phonèmes sont robustes au bruit (/k,tcl,kcl, sont très robuste, la distance restant très faible pour tout SNR local. Ces résultats nous font penser que le relatif échec du FC provient du fait qu'il n'est pas assez fin, dans le sens où il n'intègre pas

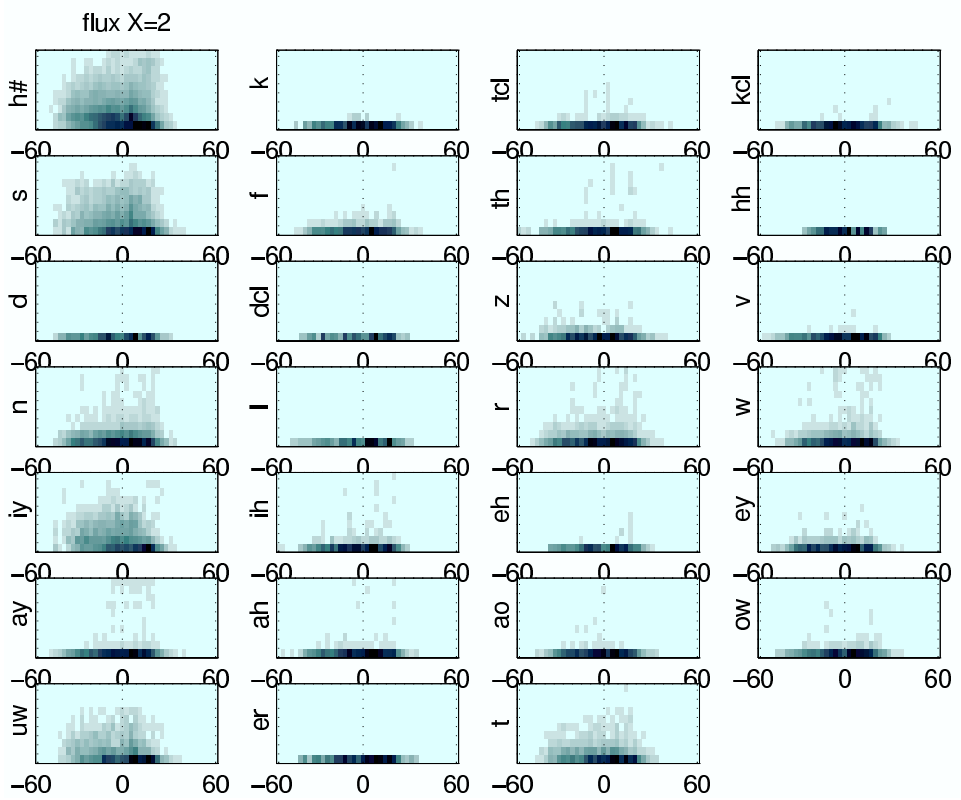


FIG. *10.2: Les distances* $|P(qk|X3clean) - P(qk|X3noisy)|$ *des posteriors sur signal propre et bruité à différents SNR, flux de bande 2. En abscisse le SNR local, en ordonnée la distance variant 0 à 1, obtenue sur du bruit blanc à différents dB (-12, 0, 12 et 24 dB global),sur 100 phrases de Numbers95*

les comportements des reconnaisseurs pour un SNR, une classe phonétique et une région spectrale observée. En effet, il est montré (Miller & Nicely 1955) que chaque région spectrale haute ou basse fréquence, est plus ou moins propice à la transmission de certaines classes phonétiques suivant le SNR. Il serait donc avantageux de tirer parti de cette information afin de sous pondérer les estimations phonétiques de nos reconnaisseurs lorsque a priori la fenêtre spectrale d'observation n'est pas propice à sa transmission. D'une façon générale la qualité d'une transmission dans un canal quelconque est corrélée à la présence de bruit. Il est donc intéressant de quantifier le niveau de détérioration de la transmission d'un phonème donné en fonction du niveau de bruit présent. C'est ce que nous allons faire dans ce qui suit, en nous attachant également à définir proprement comment de telles connaissances peuvent s'intégrer dans un reconnaisseur multi-bandes. En particulier, les mesures classiques de taux d'information transmises (Shannon 1948) ne sont pas aisément intégrables

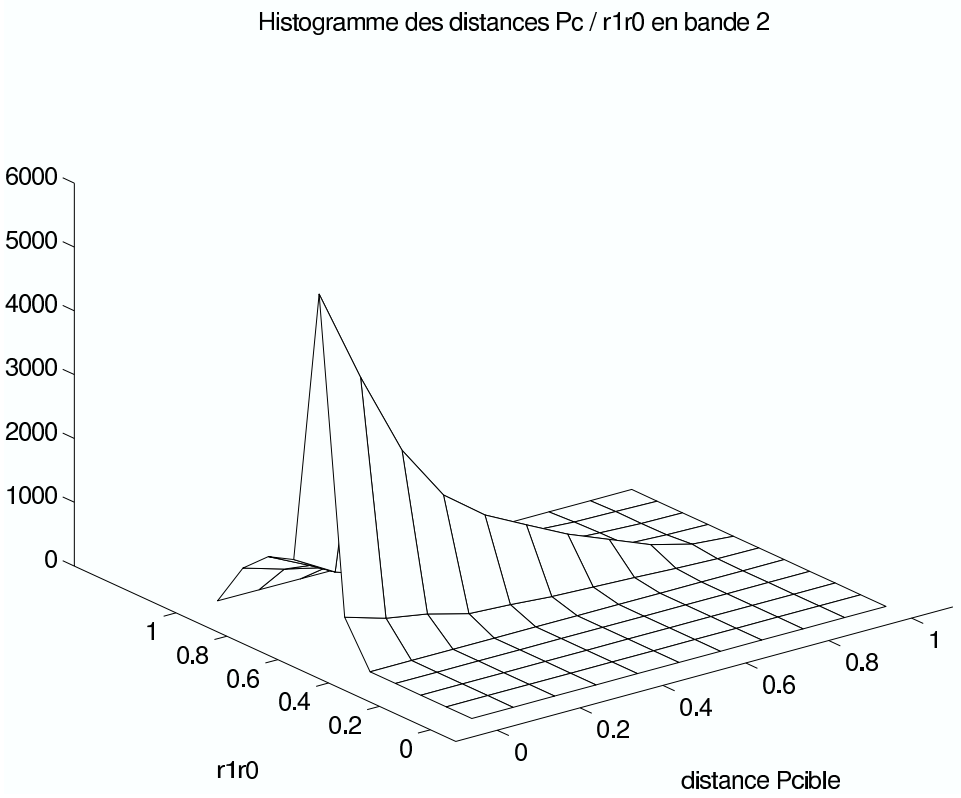


FIG. *10.3: Observation dans la sous-bande 2 de la corrélation entre R et la distance entre le Maximum a Posteriori du vecteur des estimées en parole claire (=par hypothèse posterior du phonème cible) et la valeur obtenue pour l'estimation du phonème cible dans du bruit blanc à différents dB (-12, 0, 12 et 24 dB global), sur 100 phrases de Numbers95. Noter la forte corrélation entre R et la distance.*

dans une équation de fusion de posteriors. Nous montrons comment une simple analyse des maximum de posteriors permet de définir une mesure de fiabilité de la transmission d'un phonème dans un flux pour un SNR donné, et nous montrons comment elle s'intègre simplement dans un nouveau modèle dit "Prédiction de Biais des posteriors" (PBP).

Chapitre 11

Modèle de Prédiction des Biais des Posteriors (PBP)

11.1 Système de correction de la transmission

Replaçons le problème qui nous intéresse dans le cadre idéal de la transmission d'information dans un canal discret bruité. C.E Shannon, le fondateur de la théorie de l'information après Nyquist et Hartley, décrit dans l'article majeur (Shannon 1948), les effets du bruit sur un canal de transmission, et les théorèmes dictant les règles limites théoriques de correction du code. En particulier il démontre que si le flux de données correctrices à une capacité de transmission suffisante[1], les erreurs de transmission peuvent pratiquement être toutes corrigées. Nous nous inspirons de son schéma de système correcteur pour nous ramener à notre problème 11.1. Nous y rajoutons les liens en pointillés qui ne figurent pas dans le schéma original, ils permettent la construction des fonctions de prédiction de biais en fonction du SNR S/N ou en fonction d'un observable ASA sur S'. Ces fonctions sont aussi les fonctions de comportement du reconnaisseur. Nous séparons les espaces apprentissage et décodage. L'union des deux permet l'apprentissage de la fonction de correction. Durant le décodage, seules les données de la partie "décodage" sont connues.

11.2 Dérivation du modèle PBP

Le but de notre estimateur multi-flux est d'estimer au plus juste la probabilité que le phonème k soit la cible.

Pour cela, nous proposons ici une nouvelle formule de fusion de type multi-flux basée sur l'architecture précédente. Nous limiterons la capacité du canal de correction de façon à ce que le nombre de paramètres ajoutés au modèles soit petit.

[1] En terme de bits par symbole

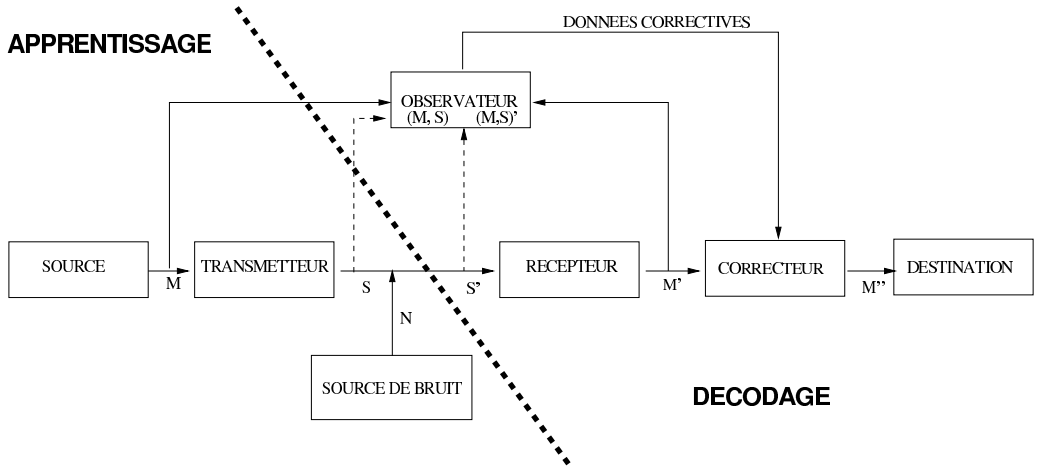


FIG. *11.1: Schéma de prédiction des biais inspiré de Shannon (Shannon 1948). Les erreurs de transmission de M dans M', du au bruit N ajouté au signal S dans S', peuvent pratiquement être toutes corrigées si la capacité en bits/symbole du canal de correction est suffisante. Nous avons rajouté les liens en pointillés vers l'observateur qui permettent la construction des fonctions de biais en fonction du SNR S/N ou en fonction d'un observable ASA sur S'. Durant le Décodage, seules les données de la partie "décodage" sont connues.*

Soit tk l'événement : 'le phonème cible est la classe k', où la cible désigne le phonème à reconnaître.

Soit qk l'événement : 'le phonème reconnu est la classe k',
où 'reconnu' suit le critère classique de maximisation des probabilités a posteriori du reconnaisseur $P(qk'|X)$ où X est le vecteur acoustique considéré de l'ensemble χ.

Il est essentiel de préciser ici que toutes les études précédentes confondent $P(tk|X)$ avec $P(qk|X)$. Notre nouvelle démarche est au contraire de bien les distinguer. Nous allons montrer qu'avec cette approche les nouveaux termes relatifs au comportement des reconnaisseurs apparaissent facilement, sans pour autant effacer les termes relatifs à la confiance du signal seul qui restent les termes de type analyse de scène auditive.

En effet, nous avons toujours :

$$P(tk|X) = P(tk, qk|X) + P(tk, \overline{qk}|X)$$

Et nous baserons notre reconnaissance sur l'estimation $P(tk|X)$ et non plus

directement sur $P(qk|X)$.

Soit fj l'événement : 'le reconnaisseur n'observe que le flux j à l'instant t.'

Puisque les fj sont mutuellement exclusifs et collectivement exhaustifs, ils forment une partition sur l'ensemble J des flux possibles.

Nous avons donc :

$$P(tk|X) = \sum_{j \in J}[P(tk, qk|X, fj) + P(tk, \overline{qk}|X, fj)]$$

Deux dérivations bayesiennes donnent :

$$P(tk|X) = \sum_{j \in J}[P(qk|X, fj).P(fj|X).P(tk|qk, X, fj) + \\ ...P(\overline{qk}|X, fj).P(fj|X).P(tk|\overline{qk}, X, fj)]$$

On note Xj l'événement : $fj \cap X$
et par définition : $P(\overline{qk}) = 1 - P(qk)$ donc :

$$P(tk|X) = \sum_{j \in J}[P(qk|Xj).P(fj|X).P(tk|qk, Xj) + \\ ...(1 - P(qk|Xj)).P(fj|X).P(tk|\overline{qk}, Xj)]$$

$$P(tk|X) = \sum_{j \in J}[P(qk|Xj).P(fj|X).[P(tk|qk, Xj) - \\ ...P(tk|\overline{qk}, Xj)] + P(fj|X).P(tk|\overline{qk}, Xj)]$$

$$P(tk|X) = \sum_{j \in J}[P(fj|X).[P(qk|Xj).(P(tk|qk, Xj) - \\ ...P(tk|\overline{qk}, Xj)) + P(tk|\overline{qk}, Xj)]$$

Nous noterons :
φ^- la fiabilité des estimation négatives
φ^+ la fiabilité des estimation positives
nous avons (voir Annexe ROC) :

$$1 - \varphi^-(k, Xj) = P(tk|\overline{qk}, Xj)$$

soit f la somme des fiabilités positives et négatives moins un :
$\Phi(k, Xj) = (\varphi^+ + \varphi^-)(k, Xj) - 1$

on vérifie que $\forall(k, Xj), 0 < \Phi(k, Xj) < 1$.

La formule de notre modèle est :

$$P(tk|X) = \sum_{j \in J}[P(fj|X).[\Phi(k, Xj).P(qk|Xj) + 1 - \varphi^-(k, Xj)]]$$

Nous nommerons ce modèle "Prédiction des Biais des posteriors" (PBP).

Nous avons donc deux termes facteurs de la fiabilité du signal $P(fj|X)$. Le premier est la probabilité a posteriori pondérée par la somme des fiabilités des estimations positives et négatives moins 1.
Le second terme est le complémentaire de la fiabilité des estimations négatives (c'est à dire le terme de "non fiabilité" des estimations négatives).

Nous allons donc dresser les fonctions de fiabilité des estimations négatives ou positives sur tous les flux en fonction d'un facteur de type SNR ou de tout autre mesure de fiabilité du signal.

11.3 Le FC est un cas particulier du PBP

Nous rappelons que le modèle FC est formulé comme suit :
$P(tk|X) = P(qk|X) = \sum_{j \in J} P(fj|X).P(qk|Xj)$

L'hypothèse de départ du modèle FC est bien de considérer que les événements tk et qk sont identiques. En effet dans ce cas le PB se réduit au FC comme nous le montrons :

Nous avons d'après Bayes :
en notant
SE(k) = Sensibilité(k)
et
SP(k) = Spécificité(k)

$$P(tk|qk) = SE(k).P(tk)/P(qk)$$

et de plus

$$1 - P(tk|\overline{qk}) = SE(k).P(\overline{tk})/P(\overline{qk})$$

Donc dans le cas où tk et qk sont identiques, on a :

$$P(tk|qk) = 1 = SE(k)$$

et

$$1 - P(tk|\overline{qk}) = 1 = SP(k)$$

Puisque dans ce cas

$P(tk|qk, Xj) = 1$

et

$P(tk|\overline{qk}, Xj) = 0)$
on montre donc que le FC est un sous cas de notre modèle PB. On remarque de plus que le FC pose donc deux hypothèses fortes :
1/ chaque reconnaisseur ne génère jamais de fausse alarme.
2/ le reconnaisseur ne génère pas de faux négatifs.

Le FC est une restriction du PBP dans le cas d'un reconnaisseur (ou détecteur de classe) idéal, ayant une sensibilité et une spécificité toutes deux égales à 1, et ce pour tout Xj, c'est à dire indépendamment du SNR du signal observé. Un tel reconnaisseur est inconcevable. Les faibles performances du modèle FC en bruit large bande peuvent être imputées à ce défaut. Alors qu'en bruit focalisé le modèle efface automatiquement les contributions des sous bandes bruitées, mais pour un bruit large bande ou la sélection du flux le plus propice à la reconnaissance est délicate, les confusions phonétiques ne sont pas corrigées.

11.4 Mise en évidence des biais des reconnaisseurs sous bandes en fonction du SNR local

11.4.1 Méthode

A partir de la segmentation de 100 phrases prises au hasard dans la base de développement, nous construisons ici les

$P(tk|qk, Xj) = (VP/(VP + FP))(k, Xj) =$
=fiabilité des estimations positives= $\varphi^+(k, Xj)$
et
$P(tk|\overline{qk}, Xj) = (FN/(FN + VN))(k, Xj) =$
$= 1-$fiabilité des estimations négatives$= 1 - \varphi^-(k, Xj)$

avec $FN(k, Xj)$ le total des faux négatifs[2], $VN(k, Xj)$ le total des vrais négatifs, $FP(k, Xj)$ le total des faux positifs et $VN(k, Xj)$ le total des vrais positifs de la matrice de confusion correspondant à la tranche de SNR souhaitée du flux Xj.

Les matrices de confusion phonétiques sont issues de chaque reconnaisseur combinatoire sur 100 phrases bruitées à -12, 12 et 24 dB SNR de bruit blanc Gaussien. Nous avons limité au minimum la taille de cette base, afin de montrer qu'il est

[2]Voir annexe ROC

possible de tirer les grands traits des comportements des reconnaisseurs à partir de quelques minutes de signal. Nous limitons nos échantillons à ceux d'un bruit blanc Gaussien filtré suivant la bande passante des reconnaisseurs analysés. Le critère de détection d'une classe est le critère MAP. Nous dressons alors les tables à trois entrées : (classe de référence, classe détectée, SNR local du pavé j considéré) et ce pour chaque flux j et son estimation associée et pour toutes les trames de notre petite base comprenant 80 locuteurs.

11.5 Calcul et prototypes des fonctions de fiabilité du PBP

Nous montrons tout au long de notre étude les moyennes des $\Phi(k)$ sur 4 groupes phonétiques choisis a priori suivant un critère phonétique et de robustesse au bruit afin de réduire la complexité de notre étude. Ainsi nous établissons les 4 groupes suivants :

S = l'état silence formera une super-classe à lui tout seul (il représente 30% des trames).

V = la super-classe de tous les phonèmes voisés, qui a priori sont les plus robuste au bruit :

V = /d,dcl,z,v,n,l,r,w,iy,ih,eh,ey,ay,ah,ao,ow,uw,er/

Restent les non voisées qui se partagent en deux super-classes suivant leurs caractéristiques :
Fnv = les fricatives non voisées qui a priori se noient facilement dans du bruit blanc = /s,f,th,hh/
Pnv = les plosives non voisées = /t,k,tcl,kcl/ qui ont la particularité d'être surtout présentes en haute fréquence.

Ces quatre groupes pourraient être choisis suivant une étude systématique de taux de transmission de l'information en milieu bruité comme dans (Miller & Nicely 1955), mais cela dépasse le cadre de notre thèse.

Le $1-\varphi^-$ d'un groupe de phonèmes K de taille $|K|$ est approché par la moyenne de ses éléments :
De même dans le cas de $\Phi(K)$.
$\varphi^-(K, Xj) = \sum_{k \in K} \varphi^-(k, Xj)/|K|$
et
$\Phi(K, Xj) = \sum_{k \in K} \Phi(k, Xj)/|K|$

Cette approximation néglige les compétition inter-phonèmes dans K, mais elle est intéressante pour simplifier la complexité du problème.

Nous mesurons dans la suite donc les comportements des fonctions de fiabilité, telles qu'elles apparaissent dans la formule du PBP, des reconnaisseurs multi-flux pour chacun des groupes.

La figure 11.3 donne donc les fonctions complémentaires de la fiabilité des estimations négatives pour tous les reconnaisseurs sous-bandes et pour les 3 grands groupes de phonèmes choisis, plus l'état silence, en fonction du SNR local.
La figure 11.7 donne les fonctions $\Phi(K, Xj)$, qui sont les sommes moins un des fiabilités des estimations positives et négatives du groupe K concerné.

On obtient pour tous les états "parole" P (tous sauf le silence) :
$1 - \varphi^{-}(P, Xj) \simeq 0$
et ce quelque soit Xj (c'est à dire quelque soit le pavé temps-fréquence considéré et son SNR).
Cela signifie que n'importe quel expert ANN est très fiable en terme d'estimation négative de n'importe quel phonème, même à très faible SNR. Le point $(0dB, \Phi = 0.5)$ est particulièrement intéressant sur ces 15 cadrans.

Par contre, quelque soit le flux considéré, nous mesurons que jusqu'à 0dB SNR les estimations négatives de l'état silence sont très fiables. La fiabilité des estimations négatives de l'état silence est sigmoïde croissante (une droite approximation bien la fonction en question). Nous mesurons que pour un SNR local supérieur à 0dB les estimations négatives de l'état silence sont très fiables ($\varphi^{-}(k, Xj) \simeq 1$). En dessous de 0 dB elles se dégradent très rapidement pour atteindre une fiabilité nulle vers -48 dB SNR.

D'autre part on observe que le facteur $\Phi = \varphi^{+} + \varphi^{-} - 1$, est variable suivant qu'on se situe en BF ou HF, sur des flux plus ou moins larges.
On voit que l'on pourrait regrouper certains flux suivant leurs comportements. Par exemple les plosives non voisées /t,k,tcl,kcl/ (en légende "Plos.") ont une fonction Φ similaire sur les flux : (3, 4, 14, 13, 23, 34, 24, 134, 234 et 1234). Cela pourrait être corrélé avec le fait que ces phonèmes ont en basse fréquence une caractéristique de fricative, leur fonction est alors similaire à celle des fricatives. En haute fréquence les plosives non voisées ont une caractéristique propre, différente des 2 autres grands groupes choisis qui sont voisées et fricatives non voisées.

11.6 Fusion des fiabilités et de R

De la même façon que dans la section précédente, nous construisons ici les fonctions de poids de PBP en fonction de matrices de confusions observées à différents intervalles de R.

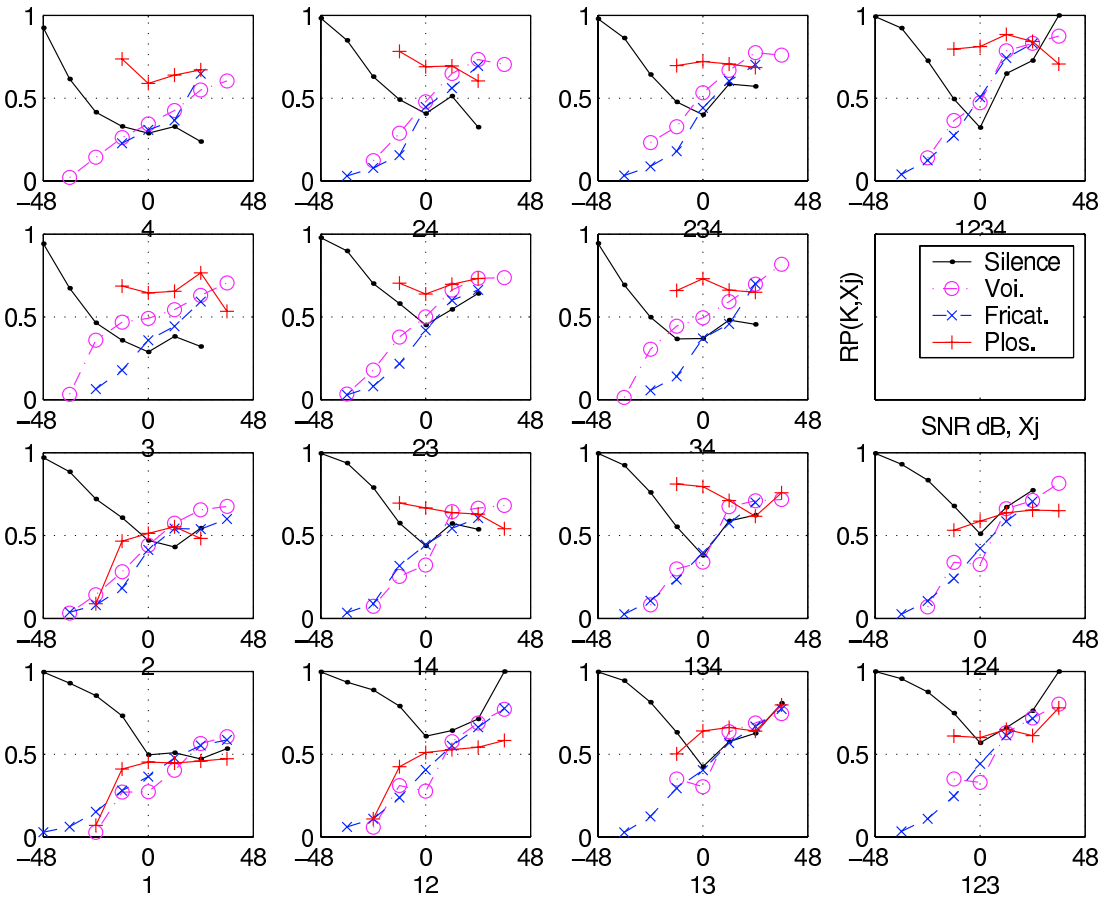


FIG. *11.2: Les fonctions de fiabilités des estimations positives φ^+ suivant le SNR local*

R étant corrélé avec le SNRlocal, on retrouve les mêmes formes pour φ^- que dans le cas en fonction du SNR local ou seul le silence se particularisait. Par contre pour φ^+ la nature même des phonèmes, plus ou moins harmonique, fait varier la forme des courbes de fiabilité par rapport à celle en SNR local. Ces fonctions sont le plus souvent linéaires en fonction de R. Ceci permet d'en extraire des prototypes de la même façon que dans la section précédente. A travers le groupe fricatif et état silence on peut distinguer encore les groupes BF et HF précédents comme permettant de regrouper cette fonction en deux grands groupes.

11.6.1 Test de reconnaissance du modèle PBP

A partir des fonctions de fiabilité précédentes, nous construisons le modèle PBP, en n'usant que des 4 variables R en sous-bande, et des 4 experts sous-bande plus

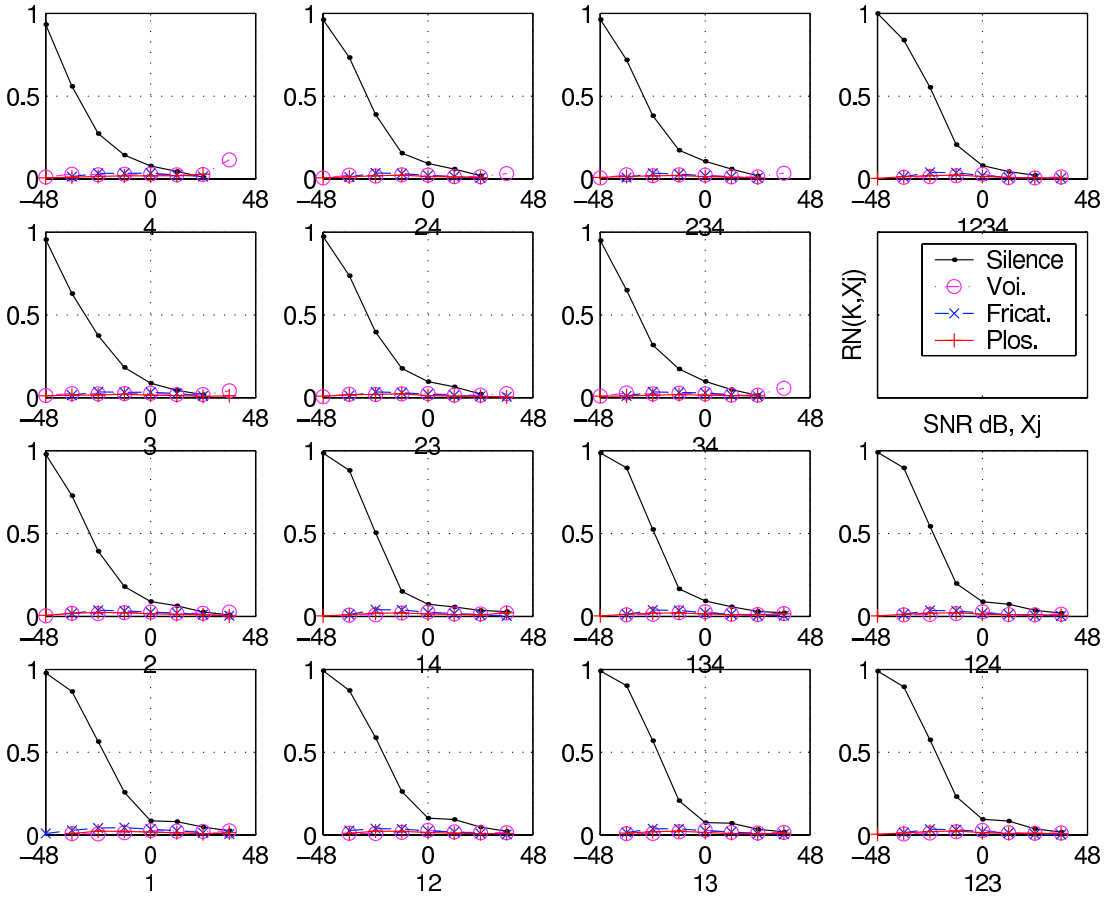


FIG. *11.3: Les* $1-\varphi^-$ *suivant le SNR local*

l'expert pleine bande. On montre que dans le cas du flux vide, $\varphi^+ = P(qk)$ et $\varphi^- = 1 - \varphi^+$. Les probabilités a priori $P(qk)$ sont réévaluées suivant les fréquences MAP en sorties des flux sous-bande et pleine bande, sur 100 phrases bruitées de l'ensemble de développement avec du bruit blanc de 0 à 18 dB. Le score de ce modèle est en moyenne meilleur que le modèle FC, sur les 6 bruits et les 6 niveaux de bruit de -12 à 18 dB, sans usage des probabilités de bruitage (mode blind, 37.1%WER contre 38.8% WER), et est meilleur en moyenne que les autres systèmes de référence (Jrasta pleine bande et Soustraction spectrale). Les tests comparatifs sont en annexe J et en table 11.6.1.
Nous voyons que le modèle PBP permet de compenser certaines erreurs du FC. Il est probable que des fonctions de fiabilité plus précises permettent un gain supplémentaire. Le PBP remplit une fonction très intéressante de démonstration des biais systématiques de certaines estimées des ANN en fonction de R et de la nature du flux.

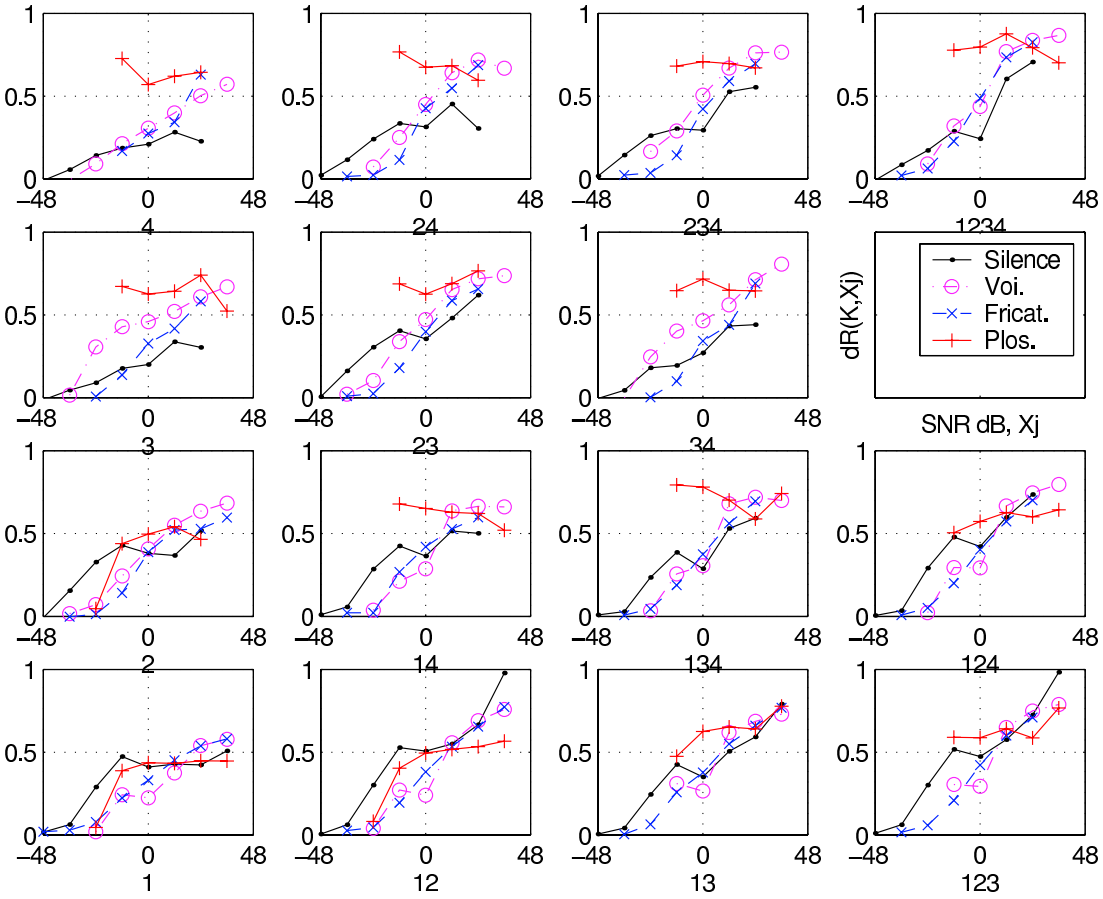


FIG. *11.4: Les* Φ *suivant le SNR local*

11.7 Conclusion

Le résultat central est la divergence de comportement inter-phonèmes ou inter superclasses pour un même flux. La simple divergence de comportement du silence suffirait à expliquer les erreurs engendrées par le modèle FC.

En ce qui concerne Φ chaque grand groupe a un comportement similaire sur les flux BF ou HF.

Mais tout ces comportements sont différents entre les grands groupes de classe et de flux, et suivant le SNR. Ils montrent que le système FC est fortement biaisé. Nous proposons donc d'utiliser nos fonctions et débiaiser les experts avec nos prédictions dans le cadre optimal du modèle PBP (Glotin 2001*b*). Les fonctions ayant en entrée R sont calculées de la même façon. L'architecture du modèle RAP est alors de la forme représentée en 11.8 et nous semble très prometteur.

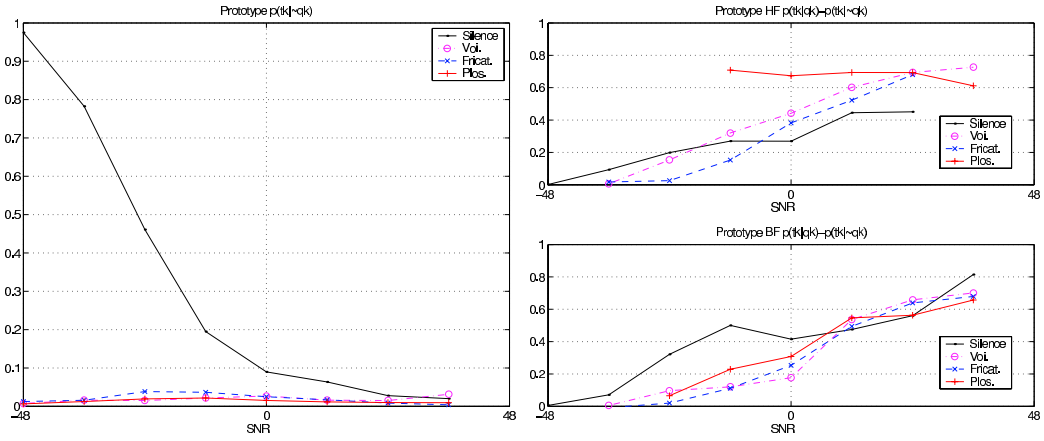


FIG. *11.5: Le prototype* $1 - \varphi^{-}(K)$ *à gauche, et à droite ceux de* $\Phi(K)$ *en HF en haut et en BF en bas. Ces variables sont exprimées en fonction du SNR local.*

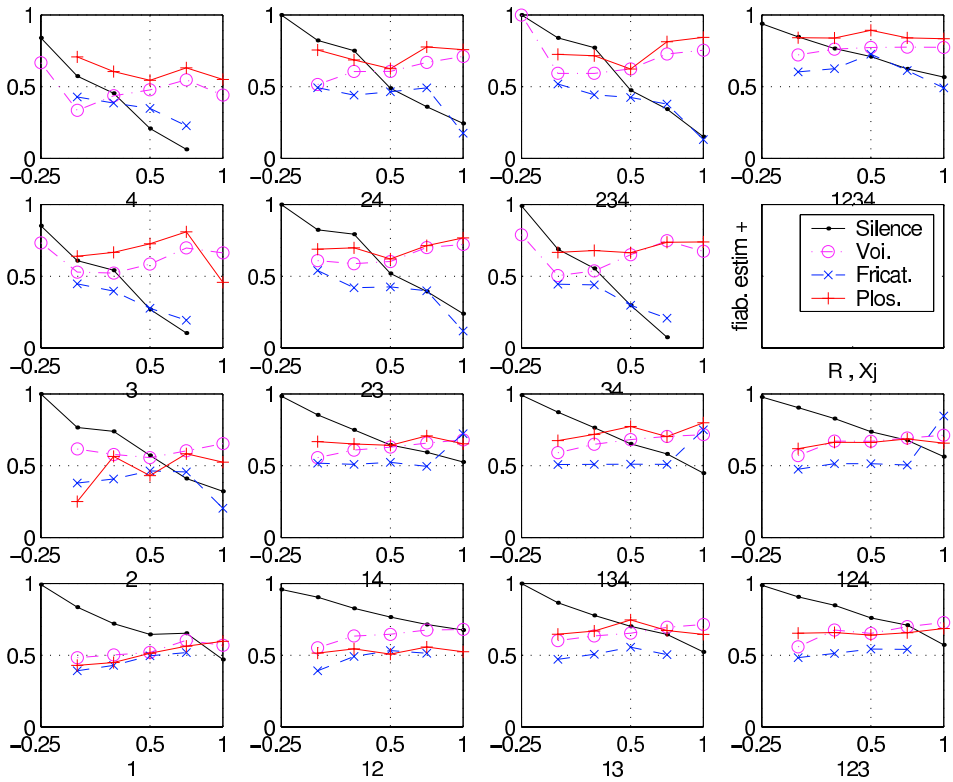


FIG. *11.6:* φ^{+} *suivant* R

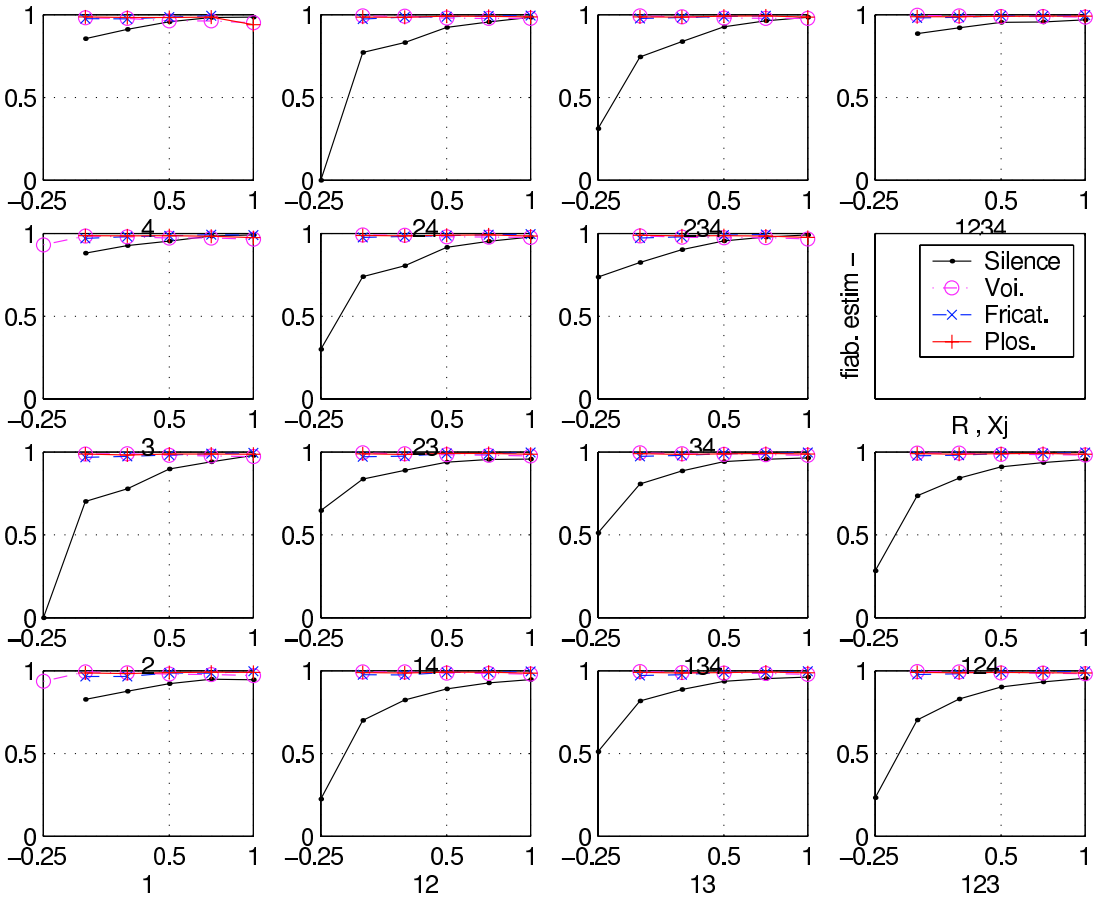


FIG. *11.7: Les φ^- suivant R*

	gwn	fact	car	narb1	narb3	n.st	MOYENNE
fband est.	38.2	37.8	33.7	26.6	30.8	90.6	42.9
SS est.	33.7	41.0	35.6	29.2	38.4	56.3	39.0
FC blind	46.9	45.6	44.2	24.5	21.7	49.9	38.8
FC Spro est.	47.3	45.0	45.0	27.1	19.3	59.8	40.6
PBP blind	46.1	43.7	42.4	25.3	21.1	44.3	37.1
PBP Spro est.	47.3	44.5	44.1	29.4	21.3	55.1	40.3

TAB. *11.1: pour méthodes de références fband = JRASTA pleine bande, SS :soustraction spectrale et pour FC et PBP. Les méthodes de pondération douce : blind=PBP ou FC avec poids $P(fj|X)$ uniformes, Spro est.= $P(fj|X)$ estimé.*
*Scores donnés en moyenne sur 200 phrases * 6 niveaux de bruits -12 à 18db par pas de 6dB. Col : Gaussien White Noise, factory, car, narrow banb 1 et 3, nonstation. noise. La reconnaissance partielle dans les cas de narb1 (ou narb3) donne 22.7 (ou 19.0) %. Intervalle de confiance = +/-1 à WER=20%, +/- 0.5 sur les moyennes finales. Voir annexe B.2 pour le détail des intervalles de confiance.*

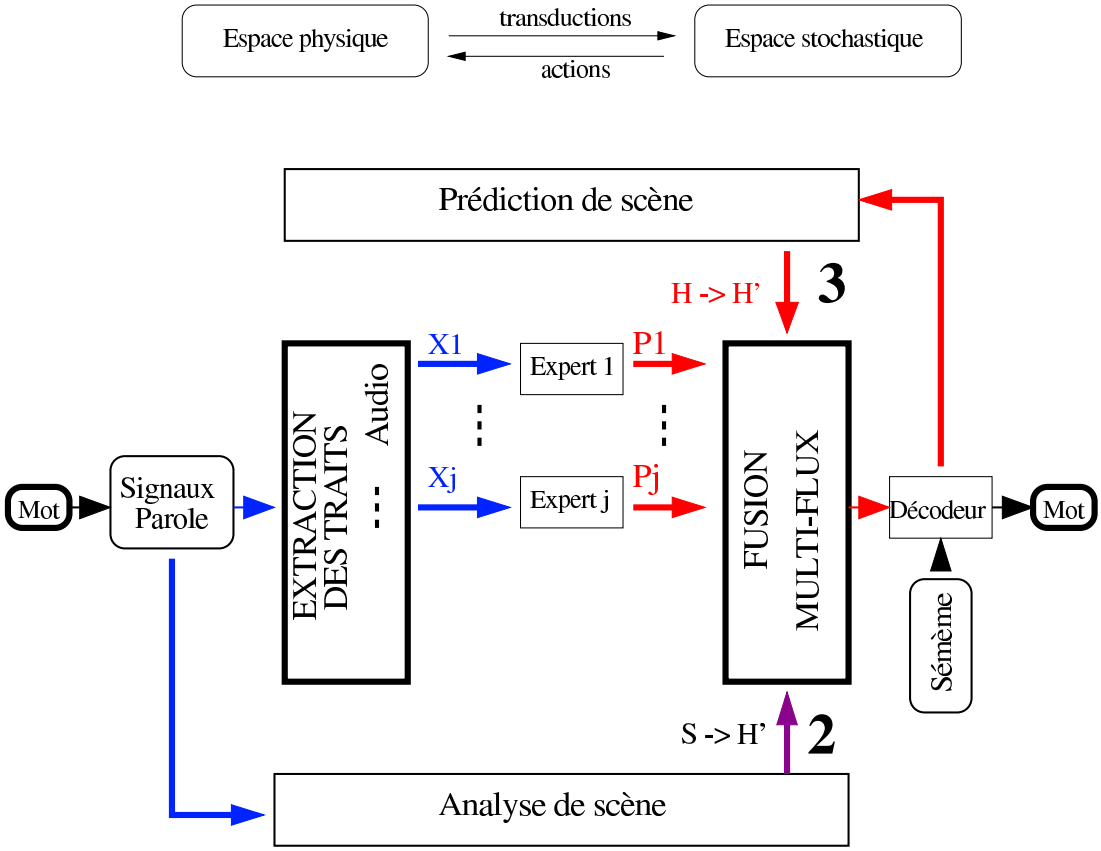


FIG. *11.8: Topologie du système de PBP : les sorties signal et estimées sont fusionnées pour gérer la fusion des reconnaissances partielles*

Cinquième partie

Conclusions et perspectives

Chapitre 12

Conclusion

12.1 Résumé des travaux effectués

Nous avons présenté des traitements du signal et des architectures de traitements innovantes et augmentant la robustesse du système. Cependant les performances en petit ou en gros vocabulaire sont encore faibles comparées aux performances humaines.

Nous traitons dans cette thèse le problème de la robustesse de la reconnaissance automatique de la parole dans le cadre original et peu traité de l'analyse de scènes auditives (CASA). Les deux voies traitées dans cette thèse ont été rarement envisagées simultanément : l'extraction d'indices fiables et la fusion de données dans un cadre multi-flux.

Les grands thèmes abordés suivent les points suivant
- Le traitement du signal visant à améliorer l'extraction des informations pertinentes pour la reconnaissance de la parole
- L'élaboration des modèles et algorithmes de reconnaissance multi-flux
- L'utilisation d'estimateur de qualité du flux audio en reconnaissance de parole sur signal monophonique, stéréophonique ou audiovisuel, en présence de bruits divers, dont le cas le bruits non stationnaires, de bruit de parole concurrente ou de la reconnaissance simultannée de 2 sources de parole.

Nous traitons l'extraction d'indices de fiabilité du signal pour la reconnaissance fondée sur une mesure du voisement obtenu par corrélation. Le calcul de cet indice est simple et permet une détection efficace de bruit confiné en sous-bande.

Dans le cadre de la reconnaissance multi-flux, nous présentons différents modes de fusion des reconnaisseurs sous-bandes. Nous détaillons leurs points forts et leurs faiblesses vis à vis de leur facteurs d'erreur. Puis nous paramètrons une système de référence de type hybride réseau neuronal/HMM, En particulier nous proposons le modèle "combinaison complète" et son approximation qui intègre toutes les combinaisons des reconnaissances partielles du signal, pondérées par leur fiabilité.

Les indices de fiabilité sont intégrés à ce modèle avec succès dans le cas de bruits focalisés. Dans le cas de bruit large bande, ces indices sont plus efficaces lors du prétraitement du signal en coopération avec l'algorithme Jrasta.

Cependant nous montrons dans le chapitre 8 que ce même indice est efficace pour renforcer la robustesse d'un reconnaisseur audiovisuel grand vocabulaire de référence (15 000 mots, système d'IBM Via Voice) dans le cas de bruit de parole. Nous proposons dans ce cadre une technique d'intégration de l'indice de fiabilité du signal reposant sur sa sensibilité et spécificité. Cette technique s'avère efficace sur un corpus limité à 10 minutes de signal mais apporte une contribution notable à la fusion d'indice de fiabilité de type voisement dans un contexte de reconnaissance audiovisuelle.

D'autre part, nous validons l'apport d'un indice de fiabilité de type corrélationnel

A la vue des résultats apportés par nos indices de fiabilité du signal , plus ou moins positifs suivant les architectures proposées, nous en tirons les conclusions dans la partie IV et nous augmentons l'architecture multi-flux par un module de correction par Prédiction des Biais des Posteriors (PBP).

Les premiers tests montrent alors que le reconnaisseur bénéficie alors d'une nouvelle source d'information, qui en renforce la robustesse au bruit en moyenne sur la base de test utilisée. Les perspectives d'un tel modèle sont discutées et suggèrent la possibilité d'augmenter la robustesse du système.

12.1.1 Estimation de la fiabilité du signal et rehaussement spectral

Nous avons proposé un estimateur de la fiabilité du signal basé sur une mesure du voisement issus des autocorrélogrammes des sous-bandes démodulées (Glotin et al. JEP 1998, Berthommier et Glotin ICPhS 1999). Nous avons montré que cet estimateur est fortement corrélé 3.4 avec le SNR local du pavé temps fréquence analysé de l'ordre de 500 Hz de large et 100 ms de durée. L'analyse ROC de cet estimateur en tant que détecteur de la présence de bruit montre qu'il est performant, même dans le cas de bruit parole à $8.5dB$, voir figure 3.4.

Nous avons tout d'abord appliqué cette mesure pour renforcer le prétraitement Jrasta. Le principe est apparenté à un filtrage de Wiener. Il consiste à pondérer les modules des sous-bandes spectrales par les indices de voisement. L'effet obtenu est une séparation du signal de parole et du bruit interférant. Puis on reconstitue le signal avant de poursuivre avec le traitement Jrasta classique. Malgré la simplicité du modèle, l'étude (Berthommier, Glotin & Tessier 2000) d'erreurs de mot de 4dB par rapport au Jrasta seul (voir figure 12.1.1).

Nous en avons démontré l' efficacité dans 3 modèles de reconnaissance robuste : rehaussement du spectre, reconnaissance multi-bande ou reconnaissance robuste audiovisuelle grand vocabulaire (15 000 mots) en condition cocktail party. Dans ce

dernier cadre de paroles simultanées, nous avons travaillé sur des signaux stéréo avec un autre indice de fiabilité du signal reposant sur la localisation des sources. Nous l'avons intégré à un reconnaisseur multi-flux. Cette étude de reconnaissance de 2 paroles simultanées à été comparée à la méthode ICA de séparation de sources, et à un nouvel algorithme de la même famille mais plus performant qui a été développé et testé dans nos travaux.

12.1.2 Elaboration de nouveaux modèles de reconnaissance multi-flux

Les reconnaisseurs automatiques multi-flux (dont sous-bandes) proposent une approche différente au problème de la reconnaissance robuste. Dans ce cadre nous avons proposé un nouveau modèle multi-flux dit "Full Combination" sommant les probabilités pondérées de toutes des combinaisons de $N(=4)$ flux spectraux élémentaires (ces derniers formant une partition du spectre pleine bande). Le poids affecté à chaque flux représente alors la probabilité pour que les estimées de ce flux soient les plus fiables parmi celles de tous les autres flux. Chacune de ces combinaisons est associée à un reconnaisseur hybride de type ANN [Artificial Neural Network] (Glotin et al. 1999, Morris et al. 2001).Ce modèle a pu être largement testé à l'échelle des laboratoires ICP, IDIAP et Mons sur la base commune NB95.

En particulier nous en avons validé une version peu coûteuse en calcul qui est construite à partir des N experts élémentaires (dont les produits estiment les experts associés aux combinaisons de sous-bandes). Cette version est tout aussi performante que le modèle usant des $2^N - 1$ experts entraînés en condition claire (équivalent également au modèle de référence) ou bruitée (Hagen & Glotin 2000). Nous avons aussi constaté que le modèle multi-flux de base usant de poids équiprobables ('pondération aveugle') était très performant[1] sur des bruits focalisés en fréquence (bruits colorés) ou même hautement non stationnaires (bruit de sirène par exemple), mais peu performant sur des bruits à large recouvrement spectral (de type bruit blanc, bruit d'usine, de voiture etc. ...)

Nous avons alors exploré divers pondérateurs de fusion des flux de reconnaissances phonétiques. Nous avons montré dans (Glotin & Berthommier 2000) que les pondérateurs basés sur un critère SNR font chuter la reconnaissance moyenne sur les 6 types de bruits variés. Nous avons étudié en profondeur ce problème très peu traité dans les thèses précédentes.

J'ai proposé une reconnaissance basée sur une pondération entropique des flux (chaque flux est pondéré à chaque instant par l'efficacité informative de son vecteur

[1] performance évaluée en moyenne sur 6 niveaux de bruits à caractéristiques spectro-temporelles différentes, et de -12 à 18 dB SNR

de probabilités phonétiques). Les résultats sont en moyenne pires que ceux obtenus par pondération aveugle.

Au vu des erreurs générées par le modèle FC en bruit large bande et son comportement intéressant en double parole (double cible), nous avons analysé en détails ses erreurs et nous en avons déduit un modèle de prédiction de Biais des Posteriors (PBP). Le PBP permet avant tout de discerner les différences de transmission phonétique à travers les différents flux et suivant certaines perturbations ce qui pourrait amener à une intégration phonétique plus robuste guidée par des indices primitifs (tel R ou TD).

12.1.3 Validation de notre indice en reconnaissance audiovisuelle sur grand vocabulaire

J'ai pu tester les techniques présentées dans cette thèse l'été 2000 lors de mon invitation au workshop "Audio-Visual Speech Recognition Engineering for Professionals Integrating Research and Education" pour travailler avec l'équipe IBM [2].
L'enjeu de cette étude effectuée avec IBM est triple : valider la Reconnaissance audiovisuelle (RAV) sur une tâche réelle de grand vocabulaire, puis implémenter et tester sur ce type de tâche les modèles de gestion de l'asynchronisme des flux audio et visuels, et optimiser leur fonction de pondération.

Nous avons premièrement spécifié la base de données audiovisuelle. 50 heures de données audiovisuelles ont été rajoutées par IBM sur la base du logiciel de reconnaissance continue "Via Voice".
Nous avons construit un modèle de reconnaissance de référence [3] égalant en performance le système de reconnaissance audio seul d'IBM, puis nous avons implémenté le modèle multi-flux de l'état de l'art gérant l'asynchronisme des flux audio et visuels (modèle "produit", développé en audiovisuel dans (Jourlin 1998, Dupont & Luettin 2000) avons validé le gain en robustesse d'un processus d'intégration asynchrone des modalités audio et visuelle dans une application réelle.

J'ai pu dans la seconde partie du workshop utiliser mon estimateur de qualité du flux audio basé sur le taux de voisement afin de relativiser les contributions des vraisemblances des flux audio et visuel dans l'estimation de la vraisemblance finale du modèle produit. Tout d'abord, une pondération globale phrase dépendante diminue de -8% le WER du modèle produit à pondération fixe [4], en condition parole propre et de -3% le WER en condition bruitée (parole interférente à 8.5dB) (Glotin

[2] voir www.clsp.jhu.edu/ws2000
[3] sous *HTK*
[4] Ce "modèle produit" fait déjà chuter de -26.8% le WER du modèle de référence audio seul

et al. 2001).

J'ai poursuivi cette étude en modifiant le RAP pour permettre d'y intégrer une pondération dynamique trame à trame (128ms) des vraisemblances, afin d'exploiter tout le potentiel de notre pondérateur dynamique dans cette application. Les résultats montrent une chute de -5.7% du WER en condition bruitée, soit une robustesse multipliée par deux comparée à une pondération statique phrase à phrase (Glotin 2000).

12.1.4 Reconnaissance de deux paroles simultanées

Nous avons travaillé sur une tâche de reconnaissance de 2 paroles simultanées d'énergie équivalentes[5] sur la base stéréo STNB95 qui est un ré-enregistrement de Numbers95 produit à l'ICP.

Tout d'abord le modèle multi-flux FC a été appliqué avec un indice de localisation des sources afin de réaliser la double reconnaissance automatique de paroles simultanées. Dans ces conditions le modèle multi-flux amène à 50% WER (Glotin et al. 1999), alors que le modèle de référence génère 70% WER (prétraitement Jrasta).

Puis nous avons collaboré avec l'université de Corée afin de comparer nos résultats avec les techniques de séparation aveugle de sources appliquées sur des paroles simultanées. Nous avons effectué la reconnaissance de parole sur les signaux séparés suivant un algorithme classique d'Indépendant Composante Analysis ICA de type Hérault-Jutten (1991) (Choi et al. 1999) et suivant un nouvel algorithme (Choi, Hong, Glotin & Berthommier 2000, Choi et al. Sept 2001). Cette dernière expérience a conduit aux meilleurs taux de reconnaissance connus sur cette base : 25% d'erreur de mot avec notre modèle.

12.2 Résumé des points forts

J'ai traité durant mon doctorat quatre grands piliers de la RAP robuste en développant des approches originales. Nous avons souvent obtenu des performances dépassant les systèmes de l'état de l'art. Nos travaux ont abouti à 3 journaux, 20 conférences internationales avec comité de lecture, 1 conférence internationale sans comité et 3 conférences nationales.

Notre thèse réunit deux aspects :
– l'analyse de la parole et l'extraction de traits robustes

[5]Cocktail party effect

– l'élaboration de nouveaux modèles de reconnaissance intégrant ces traits, notamment dans le cadre de modèle multi-bandes ou multi-stream.

Les deux points forts sont la mise en place de nouveaux estimateurs de fiabilité du signal qui sont efficaces même en conditions cocktail party. Ils ont été testé autant en signal monophonique ou stéréophonique ainsi que dans le cadre de fusion multi-modale.

La dérivation et la mise en oeuvre de nouveaux modèles de fusion de données qui tendent à être robustes en toutes conditions de bruit.

Nous avons démontré dans nos travaux l'intérêt de deux indices acoustiques en reconnaissance robuste de la parole, tirés (1) de l'autocorrélation ou (2) de l'intercorrélation pour le cas de signaux stéréophoniques. Inspirés des hypothèses de fonctionnement du système auditif humain, il s'avère que ces traits sont rapides à extraire et augmentent significativement la robustesse de différents systèmes de fusion proposés par rapport aux références classiques. J'ai développé deux modèles de fusion de ces indices : l'un (A) direct où les indices pondèrent le module des pavés temps fréquence (Berthommier & Glotin 2000) avant une synthèse du signal ainsi rehaussé. Le second (B) consiste en un nouveau modèle de fusion dans l'espace des probabilités de reconnaissance dans le cadre de l'approche multi-bandes et consiste en une intégration pondérée de toutes les combinaison de sous bandes (Glotin & Berthommier 2000, Glotin et al. 1999, Morris et al. 2001, Hagen & Glotin 2000) ci dessous un récapitulatif des performances des différents modèles.

Résultats sur bases monophoniques :
Les gains en dB dans le cas 1+A est de l'ordre de 4 dB sur Number95 comparé à la technique de référence Jrasta. 1+B est très performant en bruit non stationnaire (50%WER contre 90% WER en moyenne de -12 à 18 dB SNR par pas de 6dB * 200 phrases de Number95) ou en bruit de bande (21% contre 31% en moyenne pour un bruit centré à 1800 Hz). Le modèle PBP donne une piste dans le cas de bruit large bande.

De plus nous avons validé l'estimation et l'apport de l'information donnée par (1) dans le cas "cocktail party" (paroles de cafétéria en bruit de fond à $8.5dB$) sur la base de donnée grand vocabulaire Via Voice audiovisuelle (plus de 10 000 mots), conduisant à 5.7% de gain relatif de reconnaissance de mot par rapport au modèle de référence produit HMM (résultat sur 10 minutes de parole).

Résultats sur base stéréophonique en condition "cocktail party" :
2+A : conduit à 26% de gain relatif de reconnaissance de mot par rapport au Jrasta (Glotin et al. 1999), et 16% par rapport à une séparation aveugle de source classique (Choi et al. 1999, Tessier et al. 1999).
2+B : conduit à 32% de gain relatif de reconnaissance de mot par rapport au Jrasta, 24% par rapport à une séparation aveugle.

12.3 Discussions

Nous avons étudié et testé un premier exemple performant de couplage entre un niveau d'analyse primitif et le système multi-stream de reconnaissance de la parole, testé sur des bases de données de référence : Numbres93 et Numbers95.
Pour rappel, le système multi-bandes, consiste à faire travailler en parallèle sur les combinaisons de bandes de fréquence (4 bandes typiquement dans le spectre de la parole téléphonique) des reconnaisseurs de type perception multicouches (MLP), et d'en extraire l'information la plus pertinente au niveau de la reconnaissance de la parole. Le modèle multi-stream est plus large et intègre les modalités audio et visuelles.

Les résultats obtenus confirment la possibilité de séparer un signal de parole et une source interférente au cours d'une étape primitive (ici indice de degré d'harmonicité ou de localisation), et donc de sélectionner le reconnaisseur le plus adéquat.

Les autres solutions de ségrégation qui ont été testées, telle que la mesure de l'entropie du vecteur de sortie du MLP, semblent moins performantes dans le contexte que nous avons choisi mais elles pourraient apporter des solutions pour d'autres signaux interférents.
Pour progresser dans l'élaboration d'un modèle de reconnaissance robuste à des bruits de nature variée, nous avons montré que ces conclusions sont vérifiées en bruitant aléatoirement ou de façon régulière une des sous-bandes du spectre à chaque fenêtre d'analyse (typiquement 25 ms). Les résultats sont toujours performants avec ce type de bruit non stationnaire.

Nous avons développé l'approche 'full-combination' qui consiste à travailler sur la somme pondérée des probabilités de reconnaissance données par chacun des flux (ou streams) possibles, en incorporant la probabilité de bruitage de chaque flux. Nous avons testé avec notre indice de confiance des bruits de la base Noisex92. Les résultats montrent un gain de performance par rapport au multi-stream seul; prétraitements et MLP communs à d'autres expériences permettront bientôt des comparaisons vis à vis d'autres méthodes.

Nous avons étudié les distributions des valeurs de l'indice d'harmonicité en fonction du SNR et des classes phonétiques, ce qui nous donne une fonction de probabilité de présence du bruit que nous intégrons dans le système 'full combination'.
La méthode consiste à dresser les courbes de taux d'erreur de reconnaissance pour les 4 sous-bandes classiques sur Nb95 avec du bruit blanc de +33 à -21 db par pas de 6db. Ceci permet d'évaluer le seuil de bruitage au delà duquel plus de 10 % d'erreurs sont faites. Ce seuil est pour les bandes 1 à 4 respectivement 14, 15, 18 et 22 dB. Puis les fonctions de répartition de R1/R0 suivant le SNR ont été construites et nous en avons déduit les fonctions $P(SNR|R1R0)$. Avec l'étiquetage sur Nb95,

nous avons recalculé les fonctions de répartition de R en sous bande et par classe phonétique q_k. Nous possédons donc l'extension des probabilités de bruitage : $P(SNR < seuil|R1R0)$
aux $P(SNR < seuil|R1R0, q_k)$. L'information de bruitage suivant le phonème q_k permet de pondérer l'information de bruitage en fonction de la classe phonétique (ce qui a son intérêt dans le fait par exemple que les fricatives n'ont évidement pas les mêmes caractéristiques R que les voisées).

Le modèle FC a été falsifié, et nous proposons un nouveau modèle, de prédiction des biais PB qui intègre les biais connus a priori des posteriors des différentes classes. Ceci permet de révéler le comportement très singulier de l'état silence et de l'intégrer dans ce nouveau modèle prometteur.

12.4 Perspectives

Les perspectives sont les poursuites (1) des recherches de traits robustes et (2) des modèles de fusion de ces traits en reconnaissance de parole, afin d'améliorer les performances en parole propre et en conditions bruitées.

Une étude complète de fusion d'une part des indices actuels (harmonicité et localisation) et d'autre part des modèles 'direct' et 'FC' pourront faire l'objet d'une étude poussée et prometteuse étant donné que nous avons démontré qu'ils sont complémentaires.

Les approches de rehaussement et pondération tardives peuvent être fusionnées, ce qui amènerait à l'architecture de la figure 12.1.

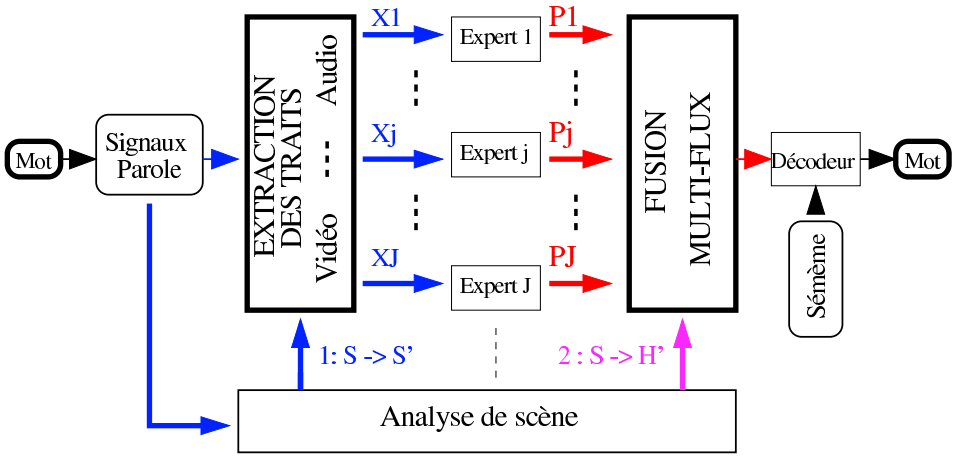


FIG. *12.1: Topologie du système de couplage fusion précoce+tardive.*

D'autres indices peuvent être développés : la modulation d'amplitude permettrait

de mieux couvrir les zones non voisées, l'entropie de l'autocorrélogramme, ou encore le nombre de PPZ sont autant d'indices à travailler.

Le modèle multiple est le cadre théorique et pratique idéal (Morris et al. 2001, Glotin et al. 1999) pour l'intégration de données multi-modale avec leur indice de fiabilité.

Nous nous proposons d'étendre ce principe à l'audio visuel en approfondissant le champ de recherche

Une autre perspective très intéressante est l'extension du modèle FC au cadre du multi-stream : en effet le modèle FC n'est pas limité au cas de combinaisons de n sous-bandes, mais peut très facilement s'appliquer au cas de flux de traits, qui sont alors des sorties d'experts phonétiques entraînés et testés sur des combinaisons de diverses modalités. Le FC se révèlerait alors comme un modèle de fusion idéal pour intégrer des traits de diverses modalités (acoustique, visuelle, articulatoire).

Chapitre 13

Vers une reconnaissance ProActive automatique de la parole : robustesse augmentée par rétro-contrôle et focalisation

13.1 Introduction

Au vu des architectures étudiées dans les chapitres précédents, nous pensons que les travaux menés par différentes équipes sur les mécanismes de perception humaine sont propices à de nouveaux traitements ou architectures dont nous explorons quelques idées dans ce chapitre.

Nous avons étudié la fusion de reconnaisseurs partiels en fonction du contexte du signal et des régions spectrales observées. L'architecture que nous proposons reste une architecture 'en avant'. Nous discutons dans cette partie de la capacité du système auditif à anticiper sur l'observation à venir. En effet pour comprendre les performances de notre système auditif, il est intéressent de s'attarder sur sa capacité d'attention sélective (Giard, Collet, Bouchet & Pernier 1994) Il s'agit du processus de perception selon lequel un stimulus cible est rehaussé par rapport à un stimulus concurrent. Une illustration de ce phénomène est le " cocktail party effect ", montrant que nous sommes capables de suivre sélectivement une conversation parmi plusieurs qui ont lieu simultanément.

Les psycho-acousticiens attribuent cette capacité à deux grands types de processus : primitifs et tardifs. Ils se distinguent par le niveau du traitement de l'information dans le système auditif. Les processus primitifs consistent en un filtrage des canaux de fréquences cibles d'après des attributs simples du signal. L'organisation des sons qui en découle conduit à la construction d'Analyse de Scène Auditive (ASA) (Bregman 1990). Les expériences récentes de psycho-acoustique mesurant l'activité cérébrale montrent que des effets attentionnels ont lieu dans le cortex au-

ditif primaire, avant ou pendant les premières phases du traitement cortical (Giard et al. 1994) chaîne de traitement du système auditif, et sont basés sur un complément de la perception par l'information descendante inhibant les afférences (Kandel et al. 1991) .

Les études en " Positron Emission Tomography " (O'Leary, Andreasen, Hurtig, Hichwa, Watkins, Ponto, Rogers & Kirchner 1996) sont mesurés dans le cortex auditif lors des tâches dichotiques, et que l'activité relative des deux hémisphères cérébraux est dépendante des phonèmes perçus, ce qui illustre un traitement tardif.

On peut faire l'hypothèse que les voies primitives et tardives travaillent simultanément. La voie primitive utiliserait les attributs du signal comme la structure temporelle ou spectrale, et la spatialisation. La voie tardive utiliserait les symboles reconnus par le système auditif ou prédits par amorçage sémantique. Cette architecture peut être vue en partie comme le schéma du "modèle interne" proposé par Lindblom (Lindblom & Lubker 1985) nous en inspirer dans notre projet.

Ces cinq dernières années ont vu évoluer une nouvelle architecture de RAP reposant sur les travaux de Fletcher (Fletcher 1953,[1929]) dans les années vingt les erreurs de reconnaissances produites par l'homme. Il a proposé un modèle d'intégration de l'information par sous-bandes spectrales. La technologie des RAP le permettant, ces travaux ont été repris soixante ans plus tard par Allen (Allen 1994) l'impulsion à un champ de recherche de RAP à architecture multi-bande, de type fusion d'experts qui est représentée par la voie (2) sur la figure 13.1. Cette architecture est confortée par l'organisation du système nerveux auditif. En effet chaque région de la cochlée répond spécifiquement à une fréquence du signal et le système central auditif qui reçoit les fibres des deux noyaux cochléaires est organisé lui aussi en zones tonotopiques d'isofréquence, on peut alors penser que le traitement de l'information est distribuée en régions spectrales.

L'architecture de la RAP multi-bande qui en découle utilise la redondance spectrale du signal de parole, et comparée à la représentation du signal Jrasta (technique classique de l'état de l'art de type soustraction spectrale reconnue pour sa robustesse (Hermansky & Morgan 1994) (Morris et al. 2001, Glotin et al. 1999, Bourlard et al. 1996, Bourlard 1996, Bourlard 1999, Mirghafori & Morgan 1998, Cerisara, Haton & Fohr Septembre 1999) compétitive en cas de bruits large bande en tenant compte des biais des estimations des posteriors connus a priori.

L'approche par fusion d'experts permet également de profiter de la multi-modalité de la parole. Dans le cas de l'audiovisuel, la redondance et la complémentarité de l'information est propice à une transmission robuste. Notons que l'architecture de la RAP multi-flux audio-visuel est du même type que celle de la RAP multi-bande.

13.2 Un processus primitif d'Analyse computationelle de scène pour une reconnaissance multi-flux

Nous proposons un modèle de segmentation du flux auditif que permet la représentation temps-fréquence du signal. Pour illustrer notre propos la figure 3.8 montre une application de l'indice tiré du taux d'harmonicité du signal sur un signal bruité.

Nous avons utilisé cet indice afin de rehausser le signal par pondération des pavés temps fréquences suivant leur taux de voisement ce qui amène (Berthommier et al. 2000) rapport au JRASTA.

Nous avons réalisé un système de reconnaissance, système hybride HMM/ANN couplé avec un module d'analyse de scène auditive (CASA pour "Computational Auditory Scene Analysis") qui fournit l'information complémentaire utile à la segmentation du flux auditif et qui sert à la sélection du bon estimateur des probabilités a posteriori des classes phonétiques.

Dans le cadre de voix simultanées, nous avons appliqué la même architecture et le même type de labélisation, et avons obtenu un chute relative des erreurs de mots de -28% sur Number95 (Glotin et al. 1999).
Dans le même cadre nous avons intégré le flux visuel en complément de l'audio, guidé par un indice ASA d'harmonicité ce qui amena un gain de significatif de quelques pourcents de reconnaissance sur la base "Via Voice" grand vocabulaire (Glotin et al. 2001).

13.3 Reconnaissance automatique ProActive de la parole

Nous pensons que la robustesse des systèmes de reconnaissance automatique de la parole (RAP) repose sur leur adaptabilité. Nous avons présenté dans l'introduction des hypothèses de traitement de l'information dans le cortex auditif. Nous nous en inspirons et proposons une nouvelle architecture en figure 13.1, tirant les bénéfices des processus primitifs et tardifs, pour extraire coopérativement les flux de parole bruités ou concurrents.

A chacun des traits Xj est associé un expert (type Multi-gaussian ou ANN), dont les estimées sont fusionnées dans le bloc qui suit. Nous avons montré qu'un processus global d'analyse de scène multi-modale peut intervenir efficacement en (1) et (2) afin d'augmenter la robustesse de la RAP. L'idée principale et l'originalité de notre schéma est l'intégration dans ces mêmes blocs des hypothèses phonétiques présentes ou passées par (3) et (4) lors des traitements suivants. C'est ce que nous appelons traitement 'proactif' de l'information, ou bloc de prédiction de scène.

Cette perspective s'inscrit dans le cadre du couplage entre les RAP et les données

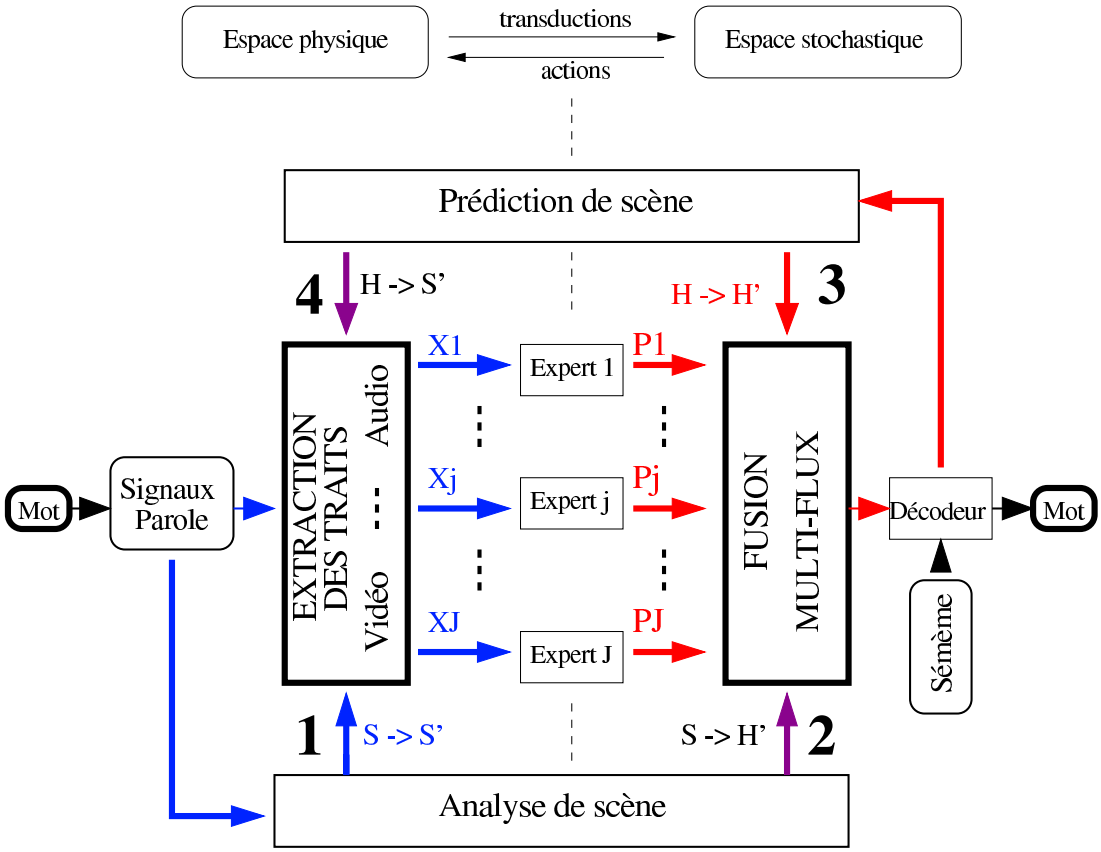


FIG. *13.1: Architecture de reconnaissance proactive. Les Pj sont les estimées des classes phonétiques pour chaque flux Xj. En (1) nous avons les techniques de rehaussement, en (2) les techniques de fusion des experts. L'estimateur de contexte signal est nommé 'analyse de scène' du terme générique consacré à cette approche qui traite des primitives physiques du signal de parole (harmonicité, localisation). Les voies (3) et (4) sont les processus proactifs. Pour chaque voie, figurent les types de transition, S pour Signal ou trait acoustique, H pour Hypothèse ou symbole.*

psychophysiques, elle est en adéquation avec l'idée selon laquelle : "le cerveau n'est pas une machine réactive, c'est une machine proactive qui projette sur le monde ses interrogations" (Berthoz & Petit 1996) de Gibson : "Information cannot be said to cause perception. Perception is not a response to a stimulus but an act of information pick-up. When stimulus energy is transformed into nervous impulses, they are said to be transmitted to the brain. But stimulus information is not anything that could possibly be sent up a nerve bundle and delivered to the brain, inasmuch as it has to be isolated and extracted from the ambient energy." (Gibson 1986) montrent que les phonèmes partiellement masqués peuvent être reconstruits par le système auditif

(Warren & Sherman 1974) suivant le contexte phonétique, ce qui est en faveur des processus tardifs (voies 3 et 4). L'architecture que nous proposons augmente donc l'architecture multi-flux par des mécanismes de rétro-contrôle ou " backward ". Les voies primitives et tardives, que l'on peut aussi dénommer " bottom up " et " top down " (Rabiner & Juang 1993) une même architecture.

13.4 Liens avec les travaux actuels

Notre thèse et nos participations au projet européens RESPITE et SPHEAR et à l'équipe de RAP audio-visuelle d'IBM nous a permis d'étudier les points (1) et (2) de la figure 13.1, petit ou grand vocabulaire. J'ai travaillé sur la séparation aveugle de sources, l'extraction de traits (harmonicité, localisation) et la fusion d'experts.

L'extraction des flux (notamment suivant leur caractéristique d'information transmise (Miller & Nicely 1955) et de rehaussement (1) (Berthommier et al. 2000) des travaux sur l'optimisation de la fusion des estimées (2).

Mais surtout, les perspectives ouvertes sont les projections des meilleures hypothèses phonétiques courantes dans les processus d'extraction ou de fusion (3) ou d'extraction (4). Avec l'expérience acquise sur les points (1) et (2), les (3) et (4) peuvent être réalisés et donner un contexte phonétique, ou une prédiction pour l'analyse de la trame suivante. La théorie du point (3) est celle du FC pondérée par l'efficacité des vecteurs de sorties en chaque flux.

Le point (4) est envisageable sous deux angles. Soit une synthèse du signal à partir des hypothèses les plus probables dans le HMM (Pratibha & Hermansky 2000), apportant ainsi un nouveau flux, ou une aide pour l'estimation du SNR. Soit (4) contrôlerait les flux suivant un critère de sensibilité et spécificité des capteurs des signaux audio-visuels de parole. Cette focalisation des traits a l'avantage de prendre en compte toute la chaîne de reconnaissance et un certain filtrage de l'information.

Je soutiens la thèse selon laquelle cette architecture devrait produire un renforcement de la robustesse de RAP : si l'hypothèse propagée dans (3) et (4) est fausse, sa vraisemblance chute, (inversement si elle est juste). On pourrait donc aussi concevoir ce système comme un testeur d'hypothèse.

L'architecture proposée est proche des systèmes réactifs. Elle repose sur le couplage entre perception et action, qui est propice à dynamiser l'interaction entre des signaux et leur concept. Cette gestion dynamique des flux d'information permet de s'affranchir des signaux masqués ou manquants.

Je propose d'utiliser le contexte phonétique, ou plutôt celui des hypothèses phonétiques courantes $H(t)$ en sortie du décodage pour guider l'extraction des $Xj(t+dt)$, avec dt de l'ordre de quelques trames (dt négatif si l'on envisage un système récursif).

Je propose aussi que la fusion des estimées $Pj(t+1)$ soit réalisée en sachant les $H(t)$, ce qui est présenté en voie (3). Cette architecture est celle du FC pondérée par l'efficacité des vecteurs de sorties en chaque flux.

L'idée de manipuler dynamiquement les traits pour adapter le système de reconnaissance on-line a été aussi présentée dans (Rathinavelu & Deng 1997) où il est montré qu'une extraction de traits dépendants des états HMM renforce la reconnaissance par rapport au MFCC. Nous proposons ici de travailler dans le cadre plus large de l'architecture multi-flux, et en considérant également des indices primitifs du signal comme guide d'extraction des traits ou fusion des flux.

13.5 Discussion et conclusion

Ce système pro-actif est similaire à une architecture de type maintenance de la vérité ou "Truth Maintenance System" TMS (Junqua & Haton 1996, Crowley & Demazeau 1993). Il permet la fusion de diverses connaissances sur la parole et la perception. J'adhère à la pensée soutenant que les techniques de rehaussement des signaux de parole peuvent bénéficier d'une orientation anthropomorphique, de même pour les mécanismes d'intégration des flux. J'ai contribué à une avancée significative dans ces deux grands domaines.

Durant ma thèse en cotutelle j'ai puisé dans le savoir faire de deux communautés ou plus particulièrement de deux laboratoires d'excellence dans le domaine. J'ai pu m'imprégner de la culture de ces deux laboratoires en y séjournant par périodes alternées de dizaines de jours tout au long de ma thèse. Le premier laboratoire est l'ICP spécialisé en traitement du signal et psychophysique de la perception. Le second est l'IDIAP reconnu pour ces travaux en technique de reconnaissance automatique de la parole (auditive, visuelle, audio-visuelle) et fusion d'experts. Ma thèse a contribué à l'élaboration de modèles s'inspirant des propriétés du système auditif, qui ont été couplés à la technique de reconnaissance partielle de pointe dans des modèles "multi-flux".

Cette architecture Pro-Active permet d'intensifier l'adaptabilité des RAP à travers une architecture pertinente permettant de contrôler l'activité des processus d'extraction des traits multi-modaux du signal de parole, et la fusion de reconnaisseurs partiels. Cette stratégie s'inscrit dans le cadre d'une boucle d'analyse/synthèse, (ou encore prédiction/correction) qui est une structure robuste pour un estimateur dynamique. Notre architecture peut aussi être analysée comme une tentative d'implémentation des propositions de Lindblom des "modèles internes" ou des cartes sensori-motrices (Lindblom & Lubker 1985)

L'originalité de cette architecture réside dans le fait que l'on force le système à travailler sur des traits multi-modaux prédéterminés, ou évoluant suivant des critères d'analyse de scène (SNR, localisation de source...) et de prédiction de scène (contexte phonétique...).

Pour cela nous proposons de renforcer nos recherches actuelles (1,2), et de conce-

voir une architecture symétrique à la précédente qui rend compte d'un traitement proactif de l'information en cours de décodage. L'idée centrale est d'intégrer un retour des hypothèses de reconnaissance vers les modules d'extraction des traits du signal ou vers le module de fusion des experts associés aux traits.

Au-delà de cet exemple, il s'ouvre des perspectives de modélisation constructive en reconnaissance audio visuelle et plus largement, des apports généraux sur la compréhension de nos mécanismes perceptifs, tout en restant conscient des différences considérables entre un ordinateur programmé et un cerveau. Jusqu'à présent, l'explication de la perception et des interprétations des sens était largement " philosophique ". Si une immense quantité de recherches a été faite par les physiologistes de l'ouïe et de la vision, nous savons peu de choses sur le fonctionnement du cerveau en tant que système d'information. D'où d'innombrables discussions sur ce qu'il faut entendre par le sens, c'est-à-dire l'interprétation d'un message extérieur. Sans prétendre que nos algorithmes " expliquent " l'audition humaine, on peut penser que ce type de démarches introduit la méthode expérimentale dans un domaine dont l'état empirique rappelle celui de la médecine au début du XIXe siècle, avant que Claude Bernard y introduise la méthode expérimentale. La science fait affecter des budgets gigantesques à des domaines tels que l'espace, la physique des particules, l'astrophysique, pour des retours que l'on peut parfois estimer minimes. En comparaison, les études qui concernent les mécanismes humains de compréhension et d'interprétation ne sont pratiquement pas soutenues alors que leur intérêt pour comprendre l'homme est extrême. Qu'est-ce que communiquer sinon interpréter des représentations ? Continuerons-nous à appréhender ces sujets avec la culture du passé ? En liaison avec les physiologistes, la voie de l'explication par la modélisation est ouverte. Comme l'a dit le prix Nobel Francis Crick, "il n'y a pas d'étude scientifique plus vitale pour l'avenir de l'homme que l'étude de son propre cerveau. Toute notre conception de l'Univers en dépend ".

Dans ce cadre nous pensons que les recherches les plus prometteuses porteront sur des architectures récursives, comme le présente remarquablement Hofstadter dans (Hofstadter 1985) :

> "Je suis convaincu que les explications des phénomènes "émergents" de nos cerveaux, comme les idées, les espoirs, les images, les analogies, et pour finir la conscience et le libre arbitre, reposent sur une sorte de Boucle Etrange une interaction entre des niveaux dans laquelle le niveau supérieur redescend vers le niveau inférieur et l'influence tout en étant lui-même déterminé par le niveau inférieur. Il y aurait donc une résonance auto-renforçante entre différents niveaux."

Sixième partie

Annexes

Annexe A

Bases de données

A.1 Les Bases de données Parole NB95 et NB93

Nos tests ont été réalisés sur la base de donnée NB93 et NB95 (Cole, Noel, Lander & Durham 1995). Ces bases sont composées de 3000 nombres, ordinaux et cardinaux, chaînes de chiffres continues ou isolées.

Les phrases ont été enregistré au laboratoire de CLSU à partir des appels téléphoniques de volontaires recrutés sans critère spécifique. La fréquence d'téchantillonnage est de 8Khz, les fichiers étant codés sur 16 bits. Les textes sont issus de numéros de téléphone, dates de naissance, numéros de rue, codes postaux. Numbers 95 représente un total de 15 000 fichiers (environ 15 heures de parole). Chaque fichier a une transcription orthographique, et pour la plupart une transcription phonétique également.

A.1.1 Les phonèmes

Les phonèmes composant les bases Numbers93 et Numbers95 sont classés dans la table ci jointe. Leur fréquence d'apparition dans la base NB95 est donnée dans la table jointe. Noter que Number93 a été utilisée avec 58 labels phonétiques incluant ceux décrit ci dessous.

Nous pouvons discriminer les phonèmes suivant un critère de voisement connu a priori (Rabiner & Juang 1993). Cette catégorisation est donnée dans la table A.2. Une catégorisation plus fine est donnée en table A.1.1

A.1.2 Probabilité a priori

Voir figure A.1.

Phonèmes de Numbers95, Symboles ICSI56	Exemple
iy	beat
ih	bit
eh	bet
ey	bait
er	bird
ay	buy
ah	but
ao	bought
ow	boat
uw	boot
d	debt
dcl	(fermeture d)
z	zoo
v	vote
n	net
l	like
r	right
w	wire
t	tea
k	key
tcl	(fermeture t)
kcl	(fermeture k)
s	sound
f	fish
th	thin
hh	hay
h (silence)	

TAB. *A.1: Liste de phonèmes de NB95*

Voisé / Non Voisé	Phonèmes de Numbers 95
Voisé	iy, ih, eh, ey, er, ay, ah, ao, ow, uw, d, dcl, z, v, n, l, r, w
Non Voisé	t, k, tcl, kcl, s, f, th, hh
Silence	h

TAB. *A.2: Classes phonologiques de Nb95*

A.1.3 Vocabulaire

Le petit vocabulaire de Numbers93 et Numbers95 est constitué des mots de la table A.1.3.

Catégorie	Phonèmes de Numbers95
Voyelle	iy, ih, eh, ey, ay, ah, ao, ow, uw
Nasale	n
Liquide	w, r, l, er
Plosive	d, t, k
Fricative	f, th, s, v, z, hh
Silence	dcl, tcl, kcl, h

TAB. *A.3: Classes phonétiques de NB95*

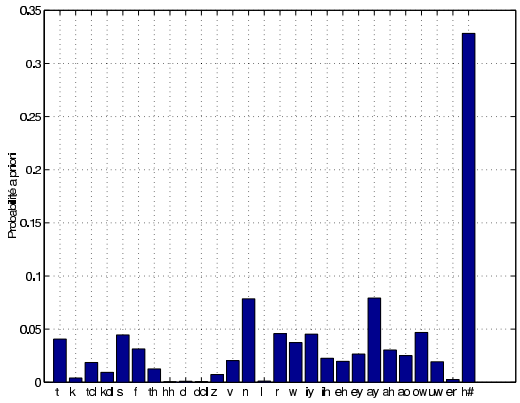


FIG. *A.1: Probabilité a priori des classes phonétiques, h# est représenté par 14 000 trames*

A.2 Bruitage des bases de données

A.2.1 Origines et description des bruits additifs

Les bruits pris pour test dans notre étude sont en partie tirés de la base NOISEX 92 (Varga, Steeneken, Tomlinson & Jones 1992).

zero	oh	ten
one	eleven	hundred
two	twelve	twenty
three	thirteen	thirty
four	fourteen	forty
five	fifteen	fifty
six	sixteen	sixty
seven	seventeen	seventy
eight	eighteen	eighty
nine	nineteen	ninety
uh	um	

Les bruits naturels ont une structure spectrale et temporelle très variés.

Pour représenter un échantillon représentatif des interférences possibles nous avons généré un certain nombre de bruit artificiels, et nous avons aussi recréer les conditions du 'cocktail party' dans NB95ST.

Les bruits utilisés dans nos travaux sont décrits en annexe J dans la table ci-contre.

type de bruit	caract. enregistrement	niveau dB
Speech noise	average speech spectrum	
M 109	30 km/h	110 dBA
Buccaneer	Pilot 190 Knots 1000 Feet	
Leopard 2	70 km/h	114 dBA
Wheel carrier	50-60 km/h	90 dBA
Buccaneer	450 Knots 300 Feet	106 dBA
Lynx	Platform	97 dBA
Leopard 1	70 km/h	104 dBA
Operation room	opsroom of destroyer	70 dBA
Destroyer	engine room	101 dBA
Machine gun	calibre 0.50 repeated	
HF radio	noise from HF radio channel	
STITEL	STI test signal	

A.2.2 Algorithme de bruitage

Chaque bruit a été ajouté au signal clair à différent niveaux SNR. Chaque niveau SNR est un niveau global sur la phrase bruitée, dont les trames de silence n'ont été exclues afin de ne pas augmenter le nombre de paramètres de l'expérience et une plus grande transparence des résultats (un détecteur de silence consiste en une décision basée sur un seuil d'énergie).

Le bruit n a été ajouté au SNR souhaité en ajustant le gain g tel que :

$SNR = 10 * log_{10}(\frac{\sum_t s^2(t)}{\sum_t g*n^2(t)})$

Les niveaux SNR choisis sont -18,-12,-6,0,6,12,18 dB SNR et niveau du signal original.

Cette base bruitée de NB95 a été ensuite diffusée à travers le réseau des projets Européens Sphear et Respite, et sert de test commun entre l'IDIAP et l'ICP, mais aussi Keel, Sheffields et Mons.

A.2.3 Définition des bruits idéaux

Afin de tester nos modèles dans des conditions idéales nous avons généré une série de quatre "bruits colorés" centrés sur les sous-bandes i (i=1..4) du modèle et

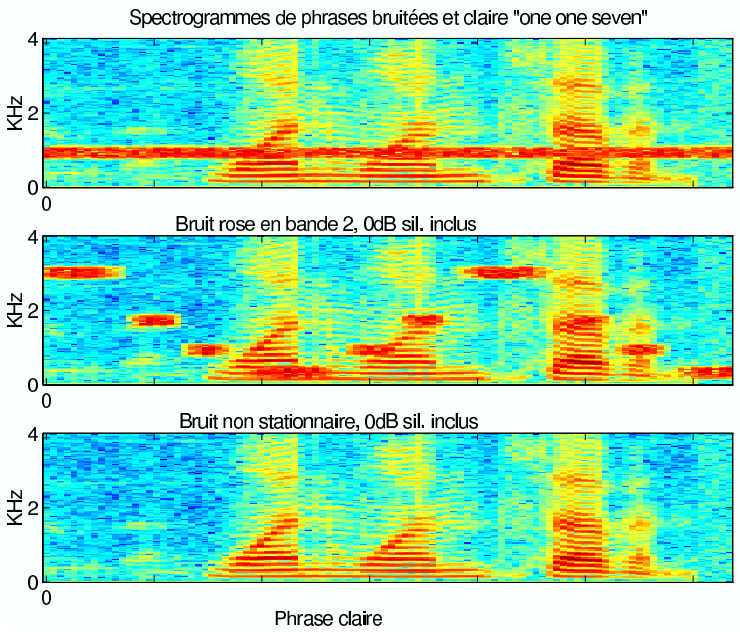


FIG. *A.2: Illustration de bruitage par "bruit coloré" dans la bande 2 (haut), du bruit non stationnaire (milieu), de la parole non-bruitée pour la même phrase «one one seven» (bas).*

n'affectant que cette sous-bande générés par des filtres trapézoidaux, 300 Hz de large, fréquence centrale égale à celle des sous-bandes du modèle hors recouvrement). Enfin pour tester l'effet de la robustesse des modèles FC et AFC suivant la distribution du bruit, nous construisons un bruit non stationnaire à partir des mêmes bruits colorés en conservant une répartition homogène du bruit sur les sous-bandes. Des pavés de 125 ms sont régulièrement tirés des sous-bandes 1, 2, 3, 4, 4, 3, 2 et 1 (voir figure A.2), comme dans (Berthommier et al. 1998).

A.3 La base StNumbers

Cette base a été enregistrée dans le cadre de la thèse de E. Tessier à l' ICP (Tessier 2001). Cet enregistrement a lieu en chambre sourde, avec l'émission simultanée de deux sources de NB95 (test set soit 2*613 phrases), couples choisis suivant leur durée (durée identique des deux phrases). Les sources sont fixes, comme le montre la figure A.3.

Initialement enregistrée à 44 KHz, la base ST numbers a été ré-échantillonée à 8 KHz car les microphones sont séparés par une distance de 40 cm, soit au moins deux fois la distance binaurale (env. 18 cm).

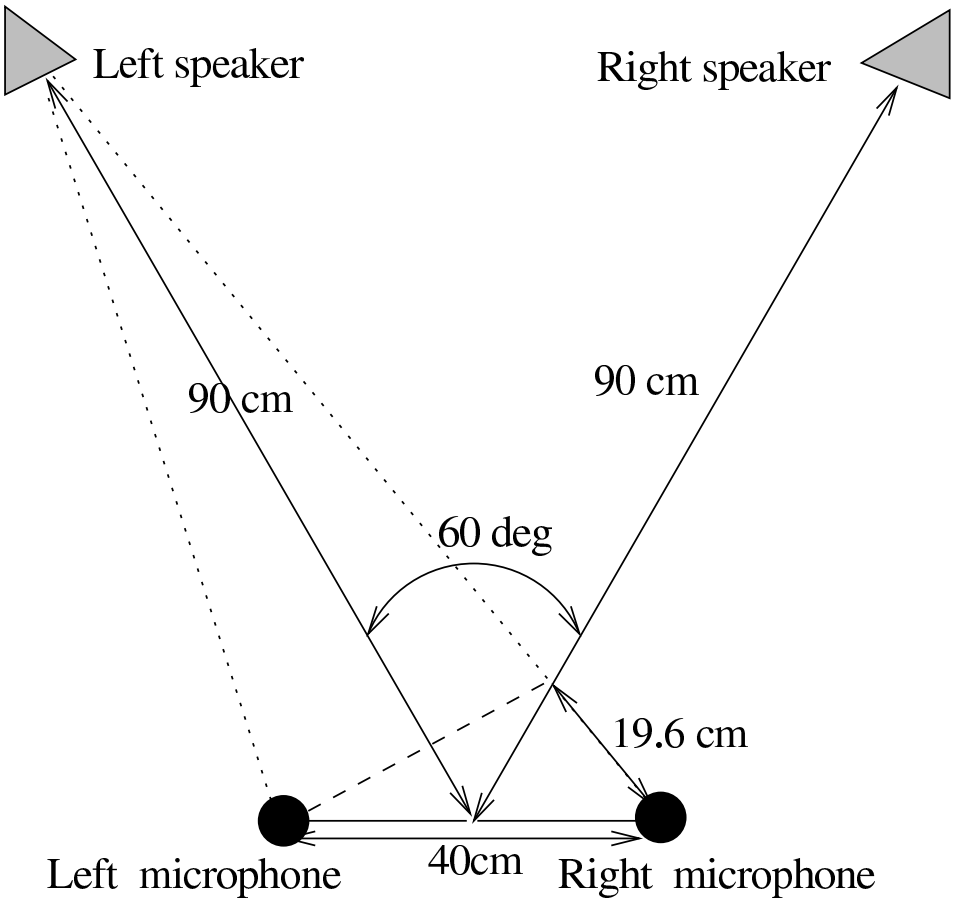


FIG. *A.3: Enregistrement de STNB95, position des 2 sources fixes, et 2 des micros fixes, en chambre sourde.*

Les sources sont aussi bien séparées dans ST, mais il faut noter que la résolution reste assez bonne meme à 8KHz (il y a 21 bins de fenêtre d'observation). (-> gain de vitesse dans un rapport 10 ou 20 au moins).

Cette caractéristique de ST nous permet d'augmenter le caractere "réaliste", car les temps de calcul sont maintenant très réduits (comparables à ceux de l'harmonicité). Un couplage des indices R+ITD devrait fonctionner en temps réel. C'est donc bien une application temps-réel qui est en vue.

Scenario	Set	Utter.	Duration	Subj.
SI/MS	Training	17111	34.9 hrs	239
SI	Held-out	2277	4.8 hrs	25
	Adaptation	855	2.1 hrs	26
	Test	1038	2.5 hrs	26
MS	Held-out	1944	4.0 hrs	239
	Test	1100	2.3 hrs	239
	Total	24325	50.6 hrs	290

TAB. *A.4: Partitions de la base Via Voice pour les expériencesspeaker independent et multi-speaker experiments*

A.4 Description de la base de donnée audio visuelle Via-Voice (IBM)

La base de donnée a été enregistrée a IBM Thomas J. Watson Research Center sur 290 sujets, la vidéo codée en MPEG2 étant cadrée sur leur visage A.4. L'extraction des paramètres labiaux est automatique et faite par l'algorithme ROI décrit en (Bregler & Konig 1994, Brooke & Scott 1994, Neti, Potamianos, Luettin, Matthews, Glotin, Vergyri, Sison & Mashari 2000)

Le vocabulaire est de l'ordre de 10500 mots.

La base est segmentée sur 24325 phrases. Le détail sur la composition de la base est donné dans la table .

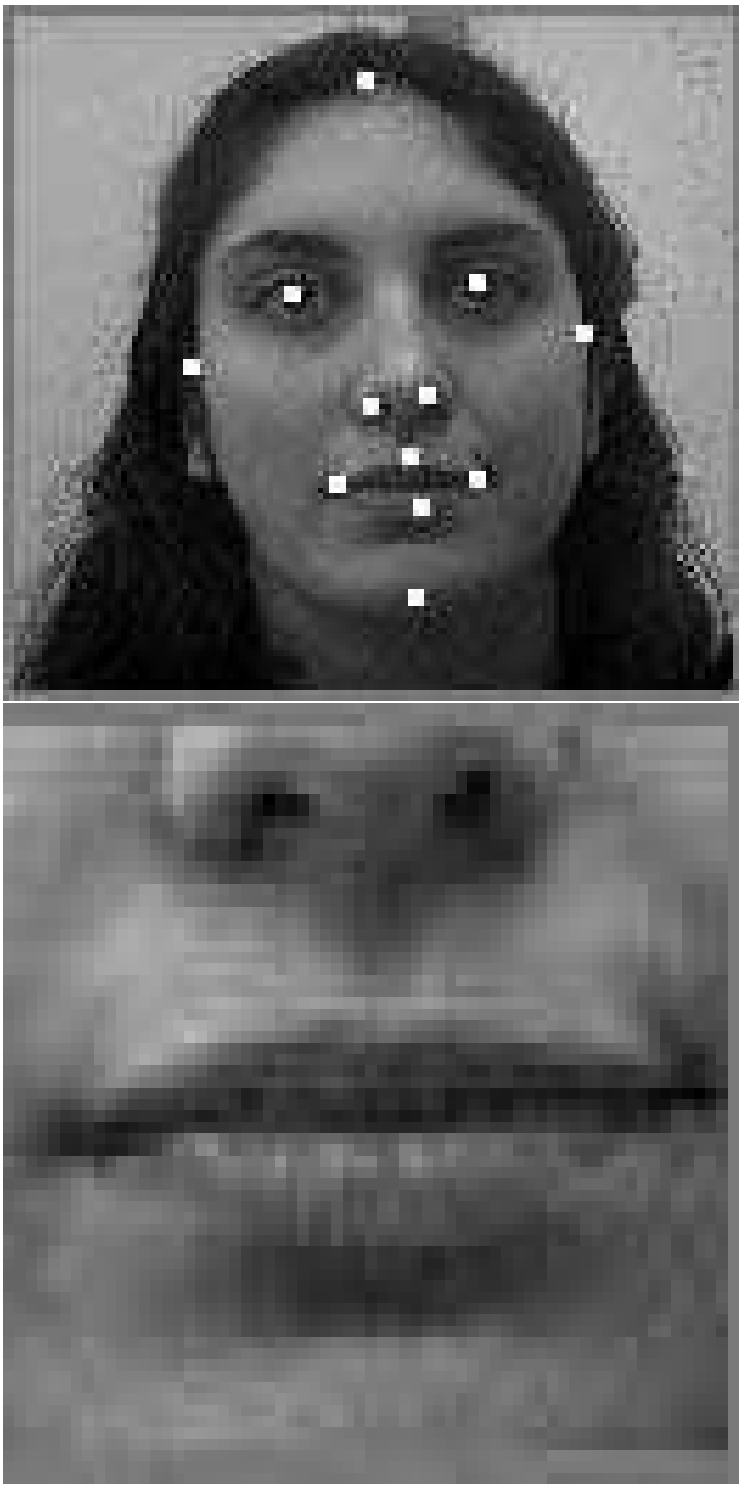


FIG. *A.4: Illustration des faces de la base Via Voice AV, avec extraction de la zone des lèvres par l'algorithme ROI.*

Annexe B

Calcul du taux d'erreur de reconnaissance de mot et leur intervalle de confiance

B.1 Calcul du taux d'erreur de reconnaissance de mot

Le calcul du taux d'erreur de mot reconnu (ou encore "Word Error Rate" en anglais) est donné par cette formule classique (dérivée de la formule de "Word accuracy") :

$WordErrorRate = 100 * (1 - \frac{N-D-S-I}{N})$

avec :

N le nombre de mots dans le fichier de transcription de référence.

D le nombre de mots omis dans le fichier résultat du reconnaisseur par rapport au fichier de transcription de référence.

S le nombre de mots substitués dans le fichier résultat du reconnaisseur par rapport au fichier de transcription de référence.

I le nombre de mots insérés dans le fichier résultat du reconnaisseur par rapport au fichier de transcription de référence.

B.2 Intervalle de confiance

En supposant que les erreurs de mot suivent une répartition binomiale, le calcul de l'intervalle I de confiance pour garantir que 95% des échantillons y sont contenus, est donné par :

$I = 1.96 * sqrt(P(1-P)/N)$

avec $N =$ nombre de mots à reconnaître : taille de l'échantillon à mesurer.

Nous montrons en figure B.1 la variation de I pour tous les WER mesurés dans nos expérience (200 phrases contenant $N = 800$ mots en tout).

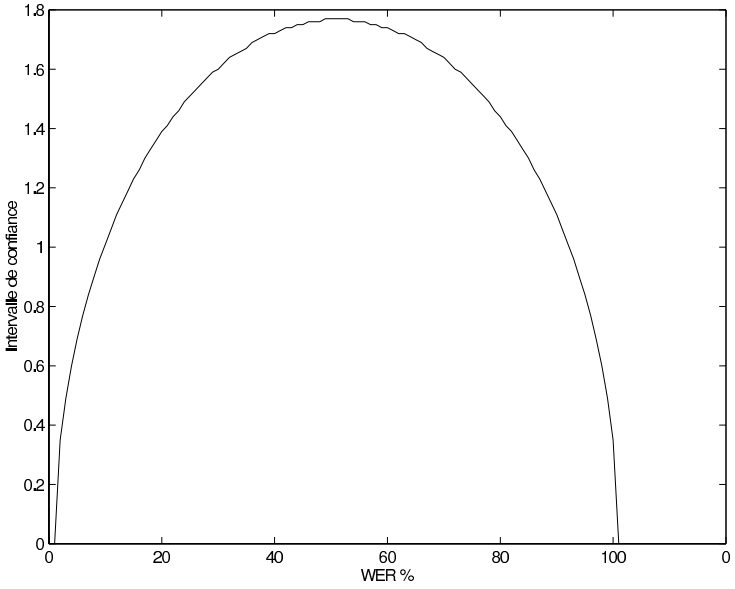


FIG. *B.1: Intervalle de confiance pour 800 mots suivant le WER*

Remarquons que pour les moyennes de WER sur différents bruits à différents SRN, nous pouvons faire l'hypothèse que le processus de processus de bruitage est identique et que tous les niveaux de bruits de SNR et tous les types de bruits sont équiprobables. Donc dans ce cas l'intervalle de confiance de la moyenne de M mesures est l'intervalle de confiance sur $N * M, N$ la étant la taille de l'échantillon mesuré pour chaque niveau de SNR différents et chaque type de bruit.

Annexe C

Algorithme de démodulation pour calcul de l'autocorrélogramme

1/ Fenêtre de hanning sur les 50 ms de signal (chevauchement de 25 ms)

2/ fft

3/ filtrage dans la sous bande considérée (ex : on prend le groupe b1b2b3)

4/ Calcul de H(AC(123)) par exemple

5/ ifft

6/ Rectification : on annule toutes les parties négatives du signal, ce qui a pour effet de renforcer la contribution des harmoniques (ajout des coupures). (nb : la rectification se fait dans le domaine temporel)

7/ fft

8/ filtrage passant [90 - 250 Hz]

9/ ifft

10/ calcul de l'autocorrélogramme.

Donc, dans un premier temps on se contente de faire des pré-traitements :
1/ filtrage par gammatone
2/ démodulation par HWR (half-wave rectification et filtrage [90-210] Hz)
Dans un deuxième temps on extrait les indices F0 avec des auto-corrélations dans le domaine temporel :

$ifft(fft(s1).conj(fft(s1)))$
On détecte le pic de la fréquence fondamentale F0 comme étant le maximum dans la fenêtre de délai 1/90Hz et 1/210Hz (le maximum absolu est situé au délai nul et il correspond au carré de l'intensité du signal).

Annexe D

Relation de Wiener-Khintchine

D.1 Transformée de Fourier

Pour tout signal $g(t)$ périodique donné, on obtient la transformée de Fourier $G(f) = TF[g(t)]$ par la formule : $G(f) = \int_{-\infty}^{\infty} g(t) \cdot e^{-i2\pi ft} dt$ Pour des signaux discrets nous avons la transformée de Fourier discrète suivante :
$G(f) = \sum_{k=-\infty}^{\infty} g(k\Delta t) \cdot e^{-i2\pi fk\Delta t}$

Le résultat est un vecteur complexe que l'on peut représenter par :
- son spectre d'amplitude $Sa(f) = \| G(f) \|$ qui est le module de $G(f)$ et qui nous donne l'énergie contenue dans ce signal pour chaque fréquence, et
- son spectre de phase qui donne le déphasage de chaque fréquence :
$Sp(f) = \arcsin(\frac{\Im[G(t)]}{\Re[G(t)]})$
La transformée de Fourier est inversible, c'est à dire qu'à partir de la transformée de Fourier $G(f)$, on peut retrouver le signal $g(t) = TFI[G(f)]$ avec
$g(t) = \int_{-\infty}^{\infty} G(f) \cdot e^{i2\pi ft} df$

D.2 Produit de convolution $(*)$

Le produit de convolution $h = f * g$ de deux fonctions f et g est défini par :
$h(x) = \int_{-\infty}^{\infty} f(t) \cdot g(x-t) dt$

D.3 Calcul du Cross corrélogramme (CC) et de l' Autocorrélogramme AC

Lorsque les signaux sont à valeurs continues, le CC $s_1 \mathrm{x} s_2$ est calculé de la façon suivante :

$CC(s_1 \text{x} s_2) = s'_1 * s_2$ où $\forall t, s'_1(t) = s_1(-t)$ et $*$ représente le produit de convolution. L'AC de s_1 se calcule comme le CC $s_1 \text{x} s_1$.

Lorsque l'on veut observer l'organisation temporelle des corrélogrammes, il peut être intéressant d'éliminer le biais dû au caractère fini des signaux, responsable de la décroissance du corrélogramme vers ses extrémités. Pour un signal de longueur M (donc un corrélogramme de longueur $2M-1$), le biais est éliminé en multipliant sa valeur en k par $M - |k|$.

D.4 Relation de Wiener-Khintchine

Une propriété des transformations de type Fourier (Fourier, Laplace, ...), est que l'opérateur de convolution en temporel est transformé en multiplication en fréquentiel, et inversement. Soit * l'opérateur de convolution et . la multiplication :

TEMPOREL	versus	FREQUENTIEL
s1(t), s2(t)	sera noté en fréquentiel par	S1(f), S2(f)
s1(t)*s2(t)	convolué équivaut en fréquentiel au produit	S1(f).S2(f)
s1(t). s2(t)	équivaut en fréquentiel à la convolution	S1(f)*S2(f)

Donc l'AC de s_1 se calcule en temporel comme :
$AC(s_1) = CC(s_1 \text{x} s_1) = (s_1(t) * s_1(-t)$
or
$S(f) = TF[s(t)] = \int_{-\infty}^{\infty} s(t) \cdot e^{-i2\pi ft} dt$
et
$s(-t) = conj(S(f)) = conj(fft(s(t)))$
donc nous avons en domaine fréquentiel :
$AC(S(f)) = S(f) \cdot conj(S(f))$
soit en repassant en temporel :
$AC(s(t)) = ifft(S(f) \cdot conj(S(f))))$

C'est la relation de Wiener-Khintchine.

Annexe E

Définition des reconnaisseurs sous bandes

Nous décrivons les paramètres essentiels de notre reconnaisseur HMM/ANN.

Les paramètres utilisées sont du type PLP. Les fenêtres sont de 25 ms de long, glissantes de 12.5 ms. Chaque ANN reçoit en entrée un contexte de 9 fenêtres d'après les expériences de (Mirghafori 1997). Les traits sont tous les coefficients du LPC plus l'énergie, plus leur dérivée première et seconde.

E.1 Définition des filtres PLP

Le Filtre PLP génère les 15 bandes critiques suivantes (coupure à -3 dB) :

bande critique	borne inférieure Hz	largeur en Hz	borne supérieure en Hz
1	17.2	144.1	161.3
2	115.3	120.5	264.6
3	216.364	303.7	374.985
4	323.152	417.289	495.227
5	438.465	541.886	628.533
6	565.34	680.778	778.419
7	707.138	837.628	948.84
8	867.586	1016.58	1144.29
9	1050.92	1222.34	1369.93
10	1261.98	1460.35	1631.7
11	1506.32	1736.88	1936.52
12	1790.41	2059.23	2292.43
13	2121.72	2435.9	2708.8
14	2509.01	2876.83	3196.64
15	2962.48	3393.65	3768.8

E.2 Paramètres des MLP sur Numbers

Nous donnons les scores d'entraînement des MLP en fonction de divers paramètres qui seront fixés pour nos expériences en fonction des performances.

Les notations sont les suivantes :
bi : les bandes critiques incluent dans la sous-bande considérée
cc : le nombre de coefficients extraits du LPC
lpc : le degré de l'analyse LPC
in : nombre de neurones d'entrée
hu : nombre de neurones cachés
out : nombre de neurones de sortie = 27
* : la configuration retenue
Pre-processing : Jrasta
param : nombre de paramètres du MLP (=poids à apprendre)
FRR train : taux de reconnaissance des trames sur la base d'entraînement
FRR cross : taux de reconnaissance des trames sur la base de validation

E.2.1 Paramètres du MLP spectre entier

bi	cc	lpc	in	hu	hu/in	# param	FRR train	FRR cross
MLP SPECTRE ENTIER								
1-15	12	10	351	1476	4.2		91.58%	87.25%
2-15	14	13	405	2000	4.9	864000	91.93%	85.81%
	13	13	378	1900	5	769500	91.78%	85.91%
	12	13	351	1750	4.9	661500	91.39%	86.10%
*	12	11	351	1750	4.9	661500	91.21%	86.15%
	id	id	351	3510	10	1326780	91.91%	86.16%
	12	10	351	1750	4.9	661500	91.02%	86.08%
3-15	12	10	351	1476	4.2		90.64%	85.69%

Malgré le meilleur résultat de FRR avec les filtres 1 à 15, nous choisissons le meilleur cas des filtres 2 à 15, car nos expériences en milieu bruité ont confirmé que la première bande critique d'un base téléphonique était trop sensible à l'ajout de bruit blanc, ce qui est aussi avancé dans ()

On voit donc que dans le spectre entier est plus performant avec moins de cc, et moins de hu (degré de liberté + petit).

Le score du MLP * est 6.9%WER sur les 1200 phrases de test de Numbers95.

E.2.2 Paramètres des MLP sous-bandes

Les contraintes que nous nous sommes imposées sont :
A/ Conserver pour chaque sous-bande le rapport obtenu pour le spectre plein : $hu/in \simeq 5$
B/ Avoir la somme des paramètres des 4 sous-bandes égale au nombre de paramètre du spectre entier
C/ Avoir un recouvrement minimal entre les sous-bandes (du uniquement à la coupure du filtre PLP)

bandes critiques	cc	lpc	in	hu	hu/in	# param	FRR train	FRR cross
sous bande 1 :								
[2,5]	5	3	162	1000	6.17	189000	64.14%	60.99%
*	5	id	162	1620	10	306180	64.465	61.4251
	6	id	189	1890	10	408240	64.90%	61.11%
	7	id	216	2160	10	524880	65.21%	61.19%
	8	id	243	2430	10	656100	65.28%	61.28%
sous bande 2 :								
[6,9]	5	3	162	1000	6.17	189000	70.66%	67.39%
*	id	id	id	1620	id	306180	70.9788	68.0645
sous bande 3 :								
[10,12]	3	2	108	1000	9.25	135000	67.04%	64.85%
	id	id	id	666	6.17	89910	66.51%	63.06%
*	id	id	108	1080	10	145800	67.1079	64.98%
sous bande 4 :								
[13,15]	3	2	108	666	6.17	89910	56.60%	54.06%
*	id	id	id	1080	10	145800	54.77%	54.77%
Somme des *						557 820		

E.2.3 Taux d'erreur de mot en clair

En parole clair, avec le prétraitement Jrasta, les MLP sélectionnés produisent les erreurs suivantes sur la base de test de Numbers95 :

MLP	WER %
1	32.3
2	28.1
3	38.4
4	53.5
(1234) (pleine bande)	6.9

Annexe F

Définition de la KL divergence et KL distance

Soit une variable aléatoire discrète Y, prenant différentes valeurs y dans un alphabet χ. Soit $d(y)$ et $g(y)$ deux distributions de probabilité pour tout $y \in \chi$. La divergence entre les deux distributions d et g est donnée par l'entropie de Kullback-Leibler $KL(d, g)$ (Kullback & Leibler 1951).

$KL(d, g) = \sum_{y \in \chi} d(y) \log \frac{d(y)}{g(y)}$

Remarques :
KL est aussi connue sous le terme d'entropie relative.

KL est bien définie car si $y \in \chi$ alors $g(y) \neq 0$.
Pour tous nos calculs KL, comme pour nos entropies nous prendrons un logarithme base 2.

Définition de la KL distance

Comme
$KL(d, g) \neq KL(g, d)$,
pour estimer la distance entre d et g nous sommons les deux divergences : de g envers d et de d envers g.

nous posons donc :
$dist(d, g) = KL(d, g) + KL(g, d)$

Annexe G

Résultats complémentaires sur différents bruits

Nous montrons ici des résultats génériques de dégradation des WER sur Numbers 93 bruitées par des bruits classiques. Noter les comportements différents des MLP(xyz) entraînés globalement après filtrage PLP. Ces résultats montrent que les experts large bande spectrale (ici couvrant toujours au moins 3 de nos sous-bandes), sont contaminés de façon assez homogène. La stratégie de reconnaissance robuste sera donc de découper les experts en sous-bande plus fine.

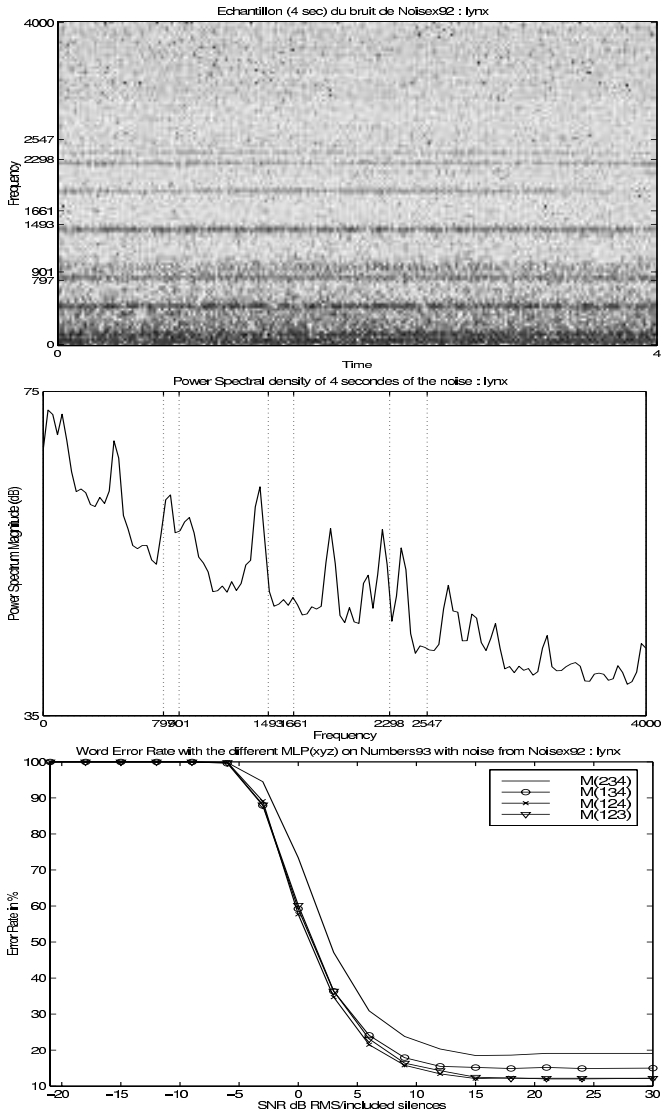


FIG. *G.1: Spectrogramme du bruit 'lynx', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

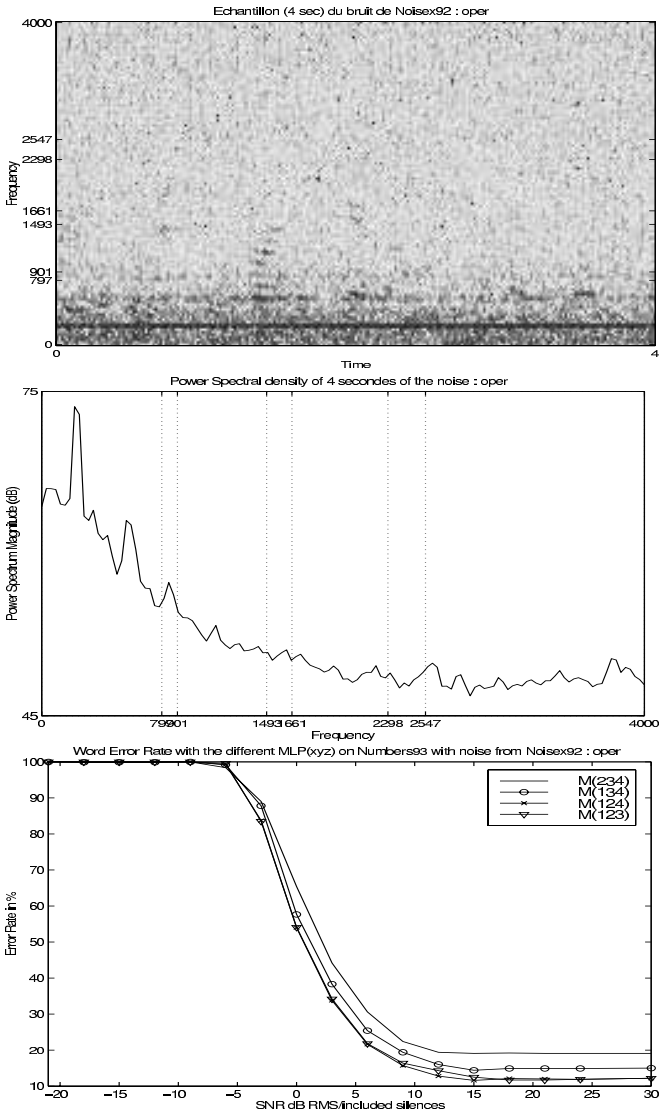


FIG. *G.2: Spectrogramme du bruit 'oper', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

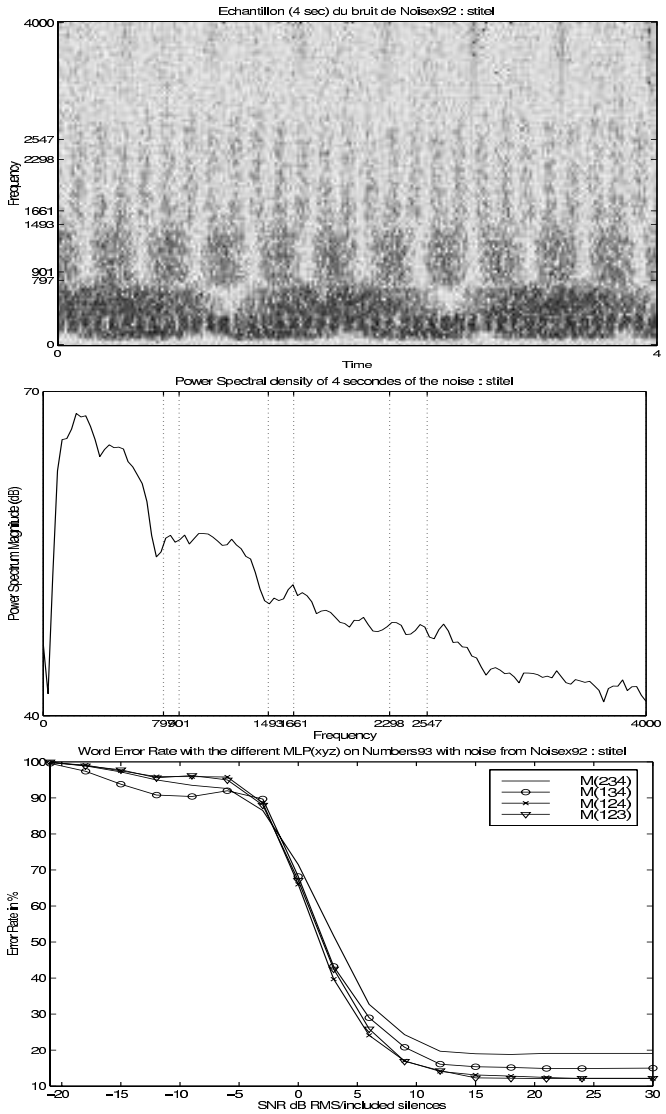


FIG. *G.3: Spectrogramme du bruit 'stitel', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

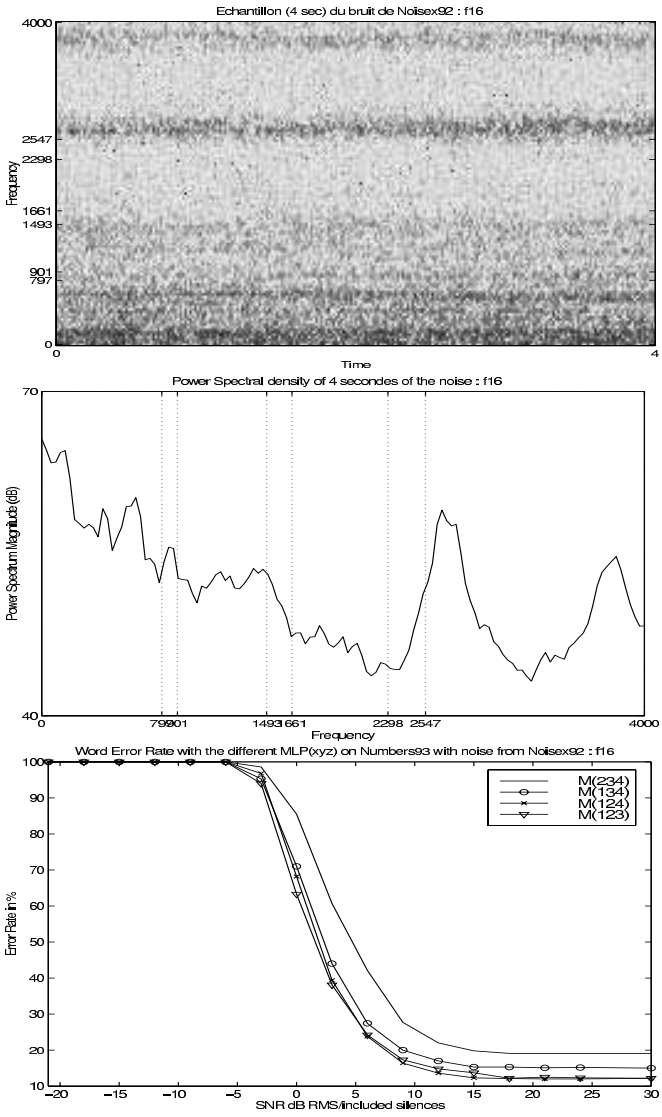


FIG. *G.4: Spectrogramme du bruit 'f16', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

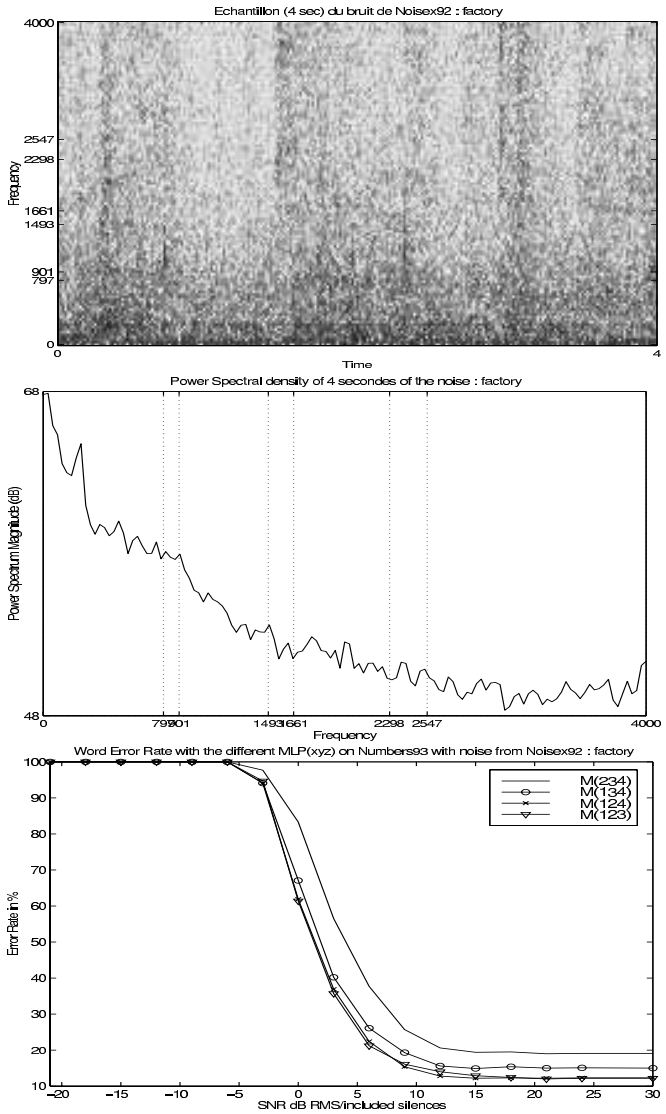


FIG. *G.5: Spectrogramme du bruit 'factory', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

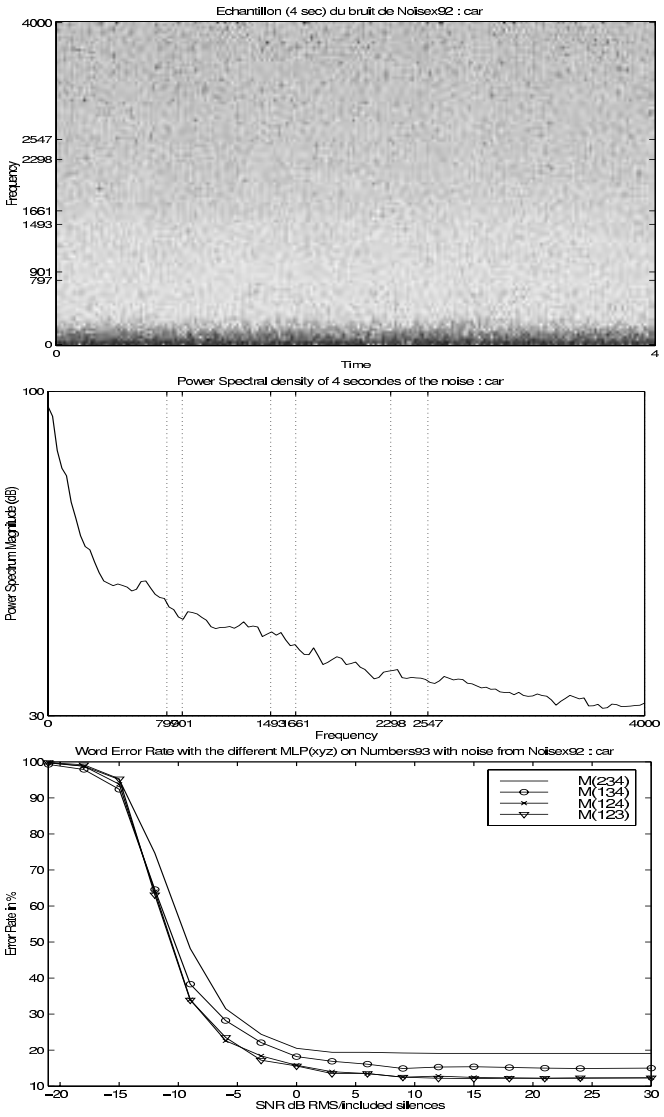


FIG. *G.6: Spectrogramme du bruit 'car', densité spectrale, et comportement des MLP(xyz) sous Numbers93 bruitée.*

Annexe H

Mise en évidence des biais par étude des distances des posteriors

Cette analyse montre la distance euclidienne entre les posteriors de chaque classe dans le cas de la parole claire et de la parole bruitée à un certain SNR. Nous représentons cette distance pour les flux bande 1,, 2, 3, 4, et le pleine bande. . On voit que la reconnaissance de la classe silence a un comportement très singulier par rapport aux autres.

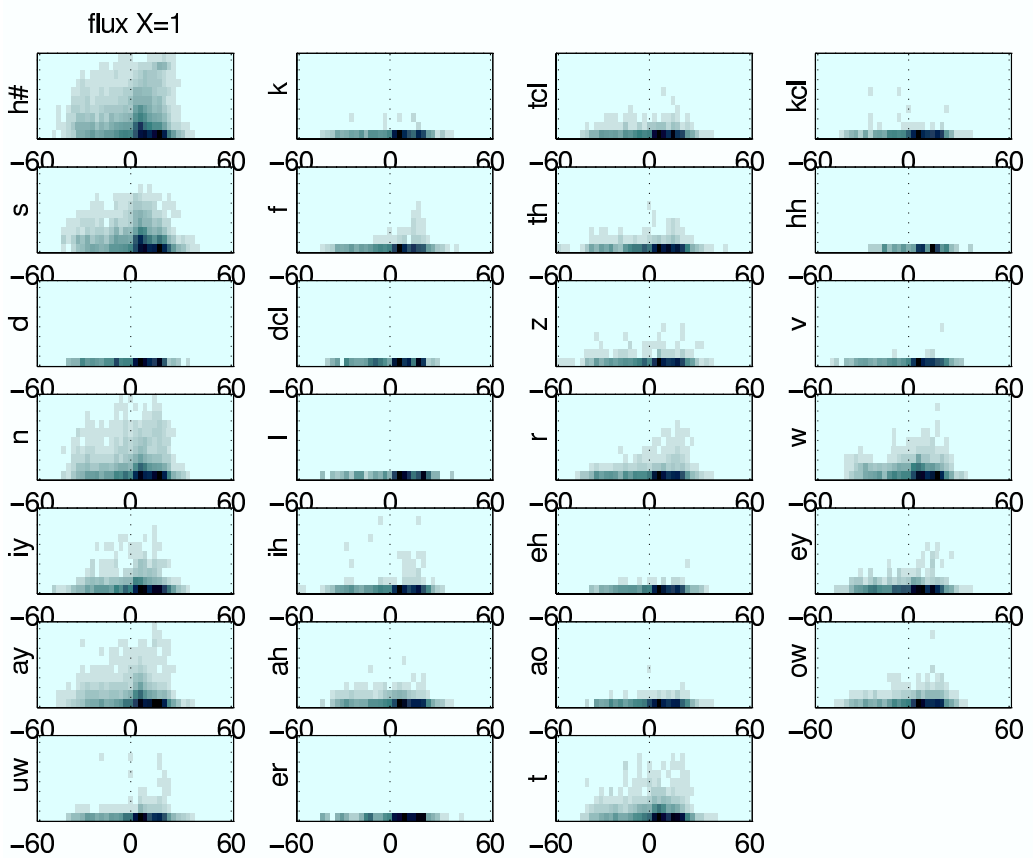


FIG. *H.1: Les distances des posteriors sur signal propre et bruité à différents SNR, flux 1*

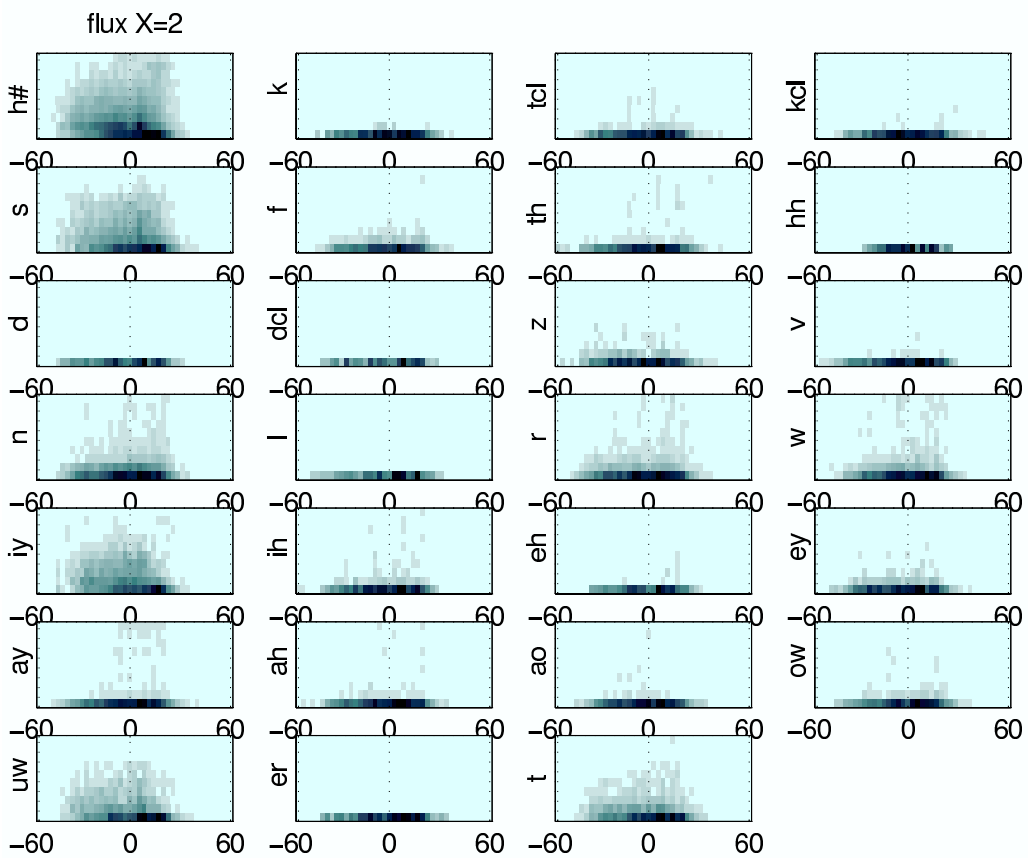


FIG. *H.2: Les distances des posteriors sur signal propre et bruité à différents SNR, flux 2*

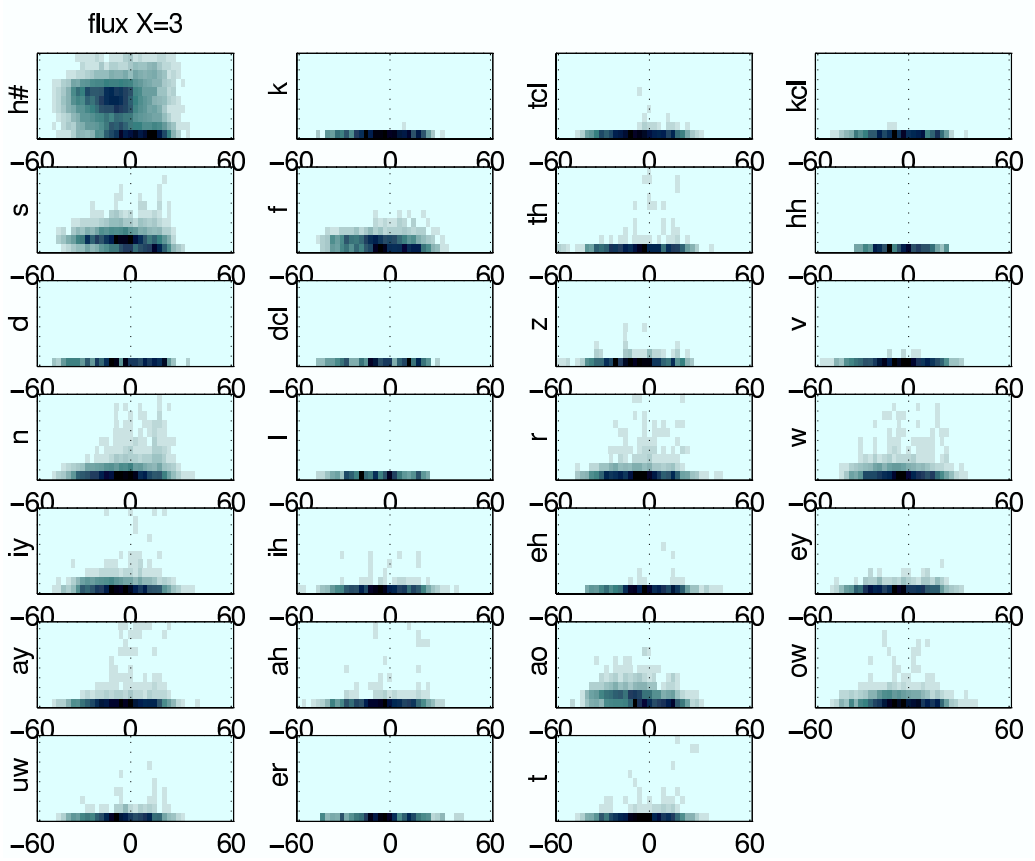


FIG. *H.3: Les distances des posteriors sur signal propre et bruité à différents SNR, flux 3*

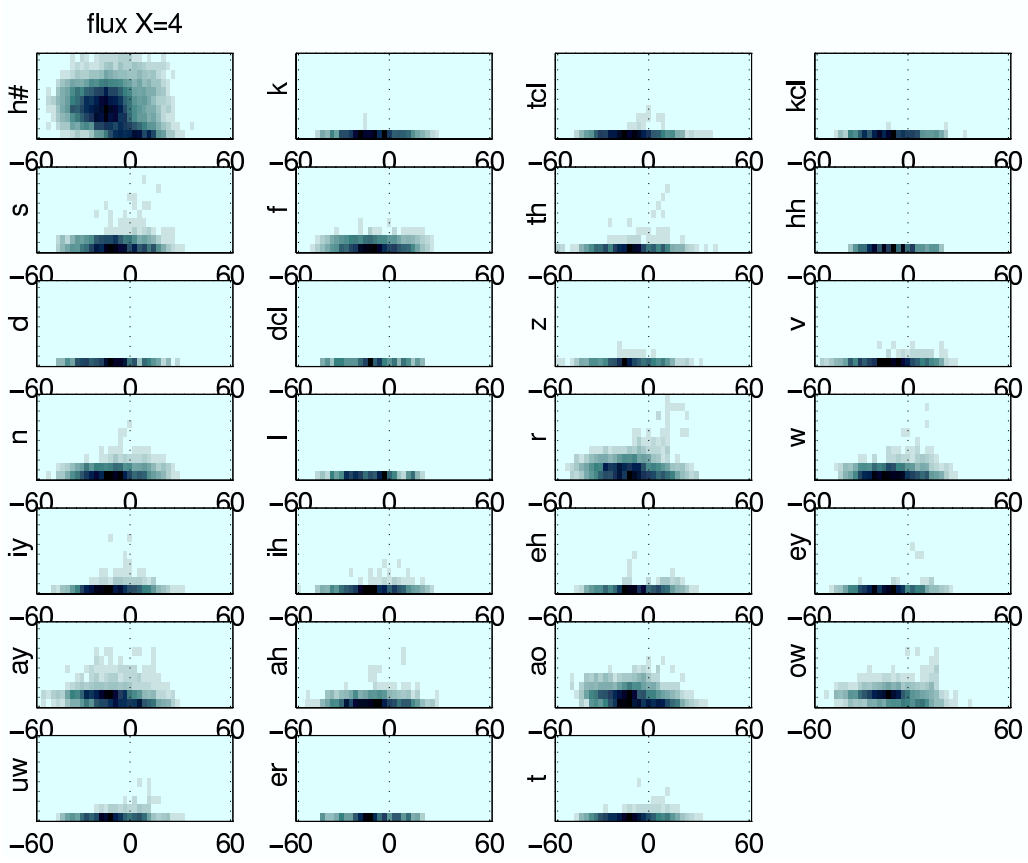


FIG. *H.4: Les distances des posteriors sur signal propre et bruité à différents SNR, flux 4*

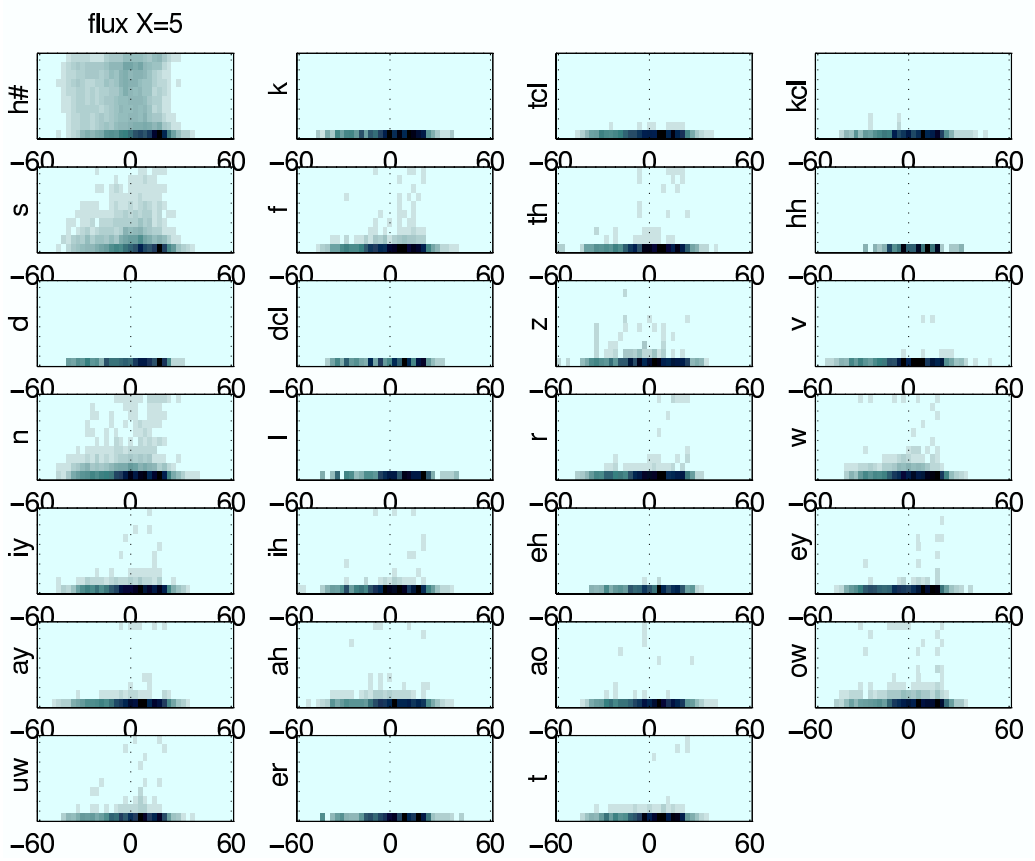


FIG. *H.5: Les distances des posteriors sur signal propre et bruité à différents SNR, flux pleine bande*

Annexe I

Analyse de détection du signal : sensibilité/spécificité

I.1 Définitions

Pour évaluer les décisions d'un test par rapport à une connaissance *a priori* ou par rapport aux résultats d'un autre test choisi comme référence, et les taux d'accord relatifs sont établis uniquement par rapport à la référence. On appelle alors sensibilité du test la proportion de réponses positives parmi les références positives, et spécificité la proportion de réponses négatives parmi les références négatives (*cf.* tableau I.1).

Résultat du test	Référence	
	Négatif	Positif
Négatif	Vrais-Négatifs (VN)	Faux-Négatifs (FN)
Positif	Faux-Positifs (FP)	Vrais-Positifs (VP)
Évaluation	Spécificité = VN/(VN+FP) = p(b|B)	Sensibilité = VP/(VP+FN) = p(s|S)

TAB. *I.1: Principe d'évaluation d'un test de décision*

L'évaluation des tests décrits se fera donc par l'intermédiaire de leur sensibilité et de leur spécificité par rapport à une référence.

I.2 Évaluation d'un détecteur à seuil variable

Lorsque le résultat du test à évaluer est obtenu par seuillage d'une valeur numérique, sensibilité et spécificité varient avec le seuil (*cf.* figure I.1).

L'évaluation du test sera alors basée sur l'évolution de la sensibilité et de la spécificité avec le seuil. La courbe ROC (Receiver Operating Characteristics) est la

représentation de la relation (non linéaire) entre sensibilité et spécificité pour tous les seuils possibles (*cf.* figure I.2).

Plus la courbe est proche de la diagonale principale, plus les résultats du test sont indépendants de la référence, et plus la courbe est incurvée vers le point (0,1), plus les résultats du test peuvent être corrélés avec la référence, et sensibilité et spécificité peuvent avoir des valeurs élevées pour un même seuil. L'aire sous la courbe ROC donne donc une indication globale sur l'adéquation du test aux valeurs de référence. Plus elle est proche de 1, mieux le test peut séparer les valeurs de référence.

Nous utiliserons l'aire sous la courbe ROC pour évaluer l'efficacité de différents paramètres à discriminer les motifs temporels desquels ils sont extraits.

I.3 Choix d'un seuil

Nous utilisons plusieurs tests dont la décision est générée par seuillage d'une valeur numérique. Nous devons donc trouver un compromis entre sensibilité et spécificité pour décider du seuil qui rend les réponses du test le plus en adéquation possible avec la référence.

Si nous n'avons aucun *a priori* sur les valeurs que doivent prendre sensibilité ou spécificité, nous choisirons le seuil de manière à maximiser le minimum de ces deux valeurs. Comme sensibilité et spécificité varient en sens opposés, ce seuil est la valeur pour laquelle la sensibilité est égale à la spécificité (intersection de la courbe ROC avec la diagonale en pointillé sur la figure I.2).

I.4 Calcul des fiabilités

Rappel : nous avons si qk est l'observable et tk la référence :
sensibilité=SE
spécificité=SP

Nous avons fiabilité des estimations positives :
$= P(tk|qk)$

et fiabilité des estimations négatives :
$= P(\overline{tk}|\overline{qk})$

nous rappelons que
$SE(k) = P(qk|tk) = VP/(VP + FN)$

$SP(k) = P(\overline{qk}|\overline{tk}) = VN/(VN + FP)$

et par définition

Par Bayes on a :

$p(tk|qk) = p(qk|tk).p(tk)/p(qk)$

Donc si $qk = tk$ (cas du détecteur idéal) alors se=sp=1 et RP=1, RN=0.

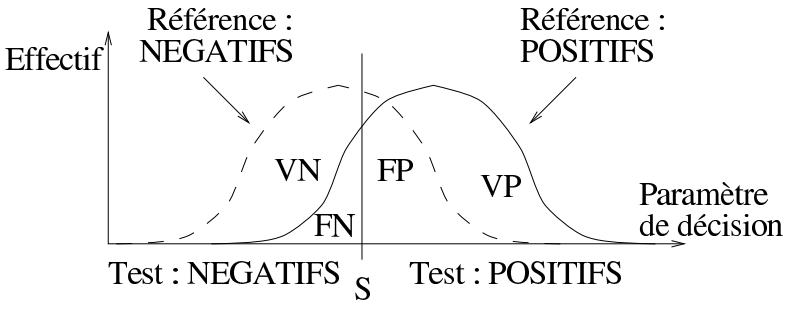


FIG. *I.1: Évolution conjointe de la sensibilité et de la spécificité avec le seuil de décision. Le résultat du test est obtenu par seuillage du paramètre de décision. Lorsque le seuil S est inférieur à la plus faible valeur que peut prendre ce paramètre, $FN = VN = 0$, donc la spécificité est nulle et la sensibilité maximale. Lorsque le seuil augmente, FN et VN augmentent tandis que FP et VP diminuent, donc la spécificité augmente et la sensibilité diminue. Lorsque le seuil est plus fort que la valeur maximale du paramètre de décision, la sensibilité est nulle et la spécificité maximale.*

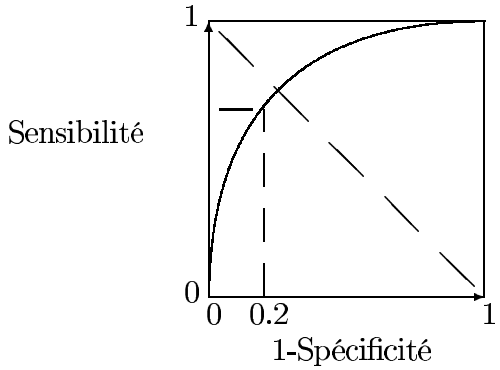


FIG. *I.2: La courbe ROC (Receiver Operating Characteristics) : principe de représentation de l'évolution conjointe de la sensibilité et de la spécificité avec le seuil. Lorsque le seuil augmente, la spécificité augmente et la sensibilité diminue (cf. figure I.1). Au niveau de l'intersection de la courbe ROC avec la diagonale en pointillé, la sensibilité est égale à la spécificité : le minimum de ces deux valeurs est donc maximal.*

Annexe J

Résultats WER comparatifs de PBP, FC, Jrasta et Soustraction spectrale

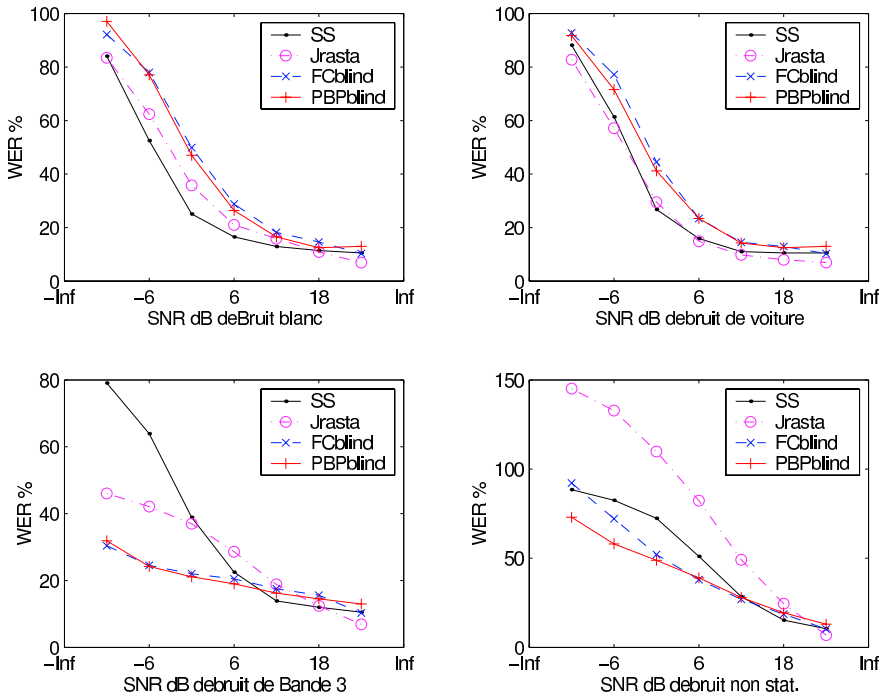


FIG. *J.1: WER pour 4 bruits, 2 large bande, 2 focalisés stationnaires ou non. 200 phrases de test pour chaque point. SS : Soustraction spectral, Jrasta : modèle pleine bande avec prétraitement Jrasta. FC et PBPblind : modèles FC ou PBP avec probabilité de bruitage $P(fj|X)$ uniformes. Intervalle de confiance +/- 1% à 20% WER.*

Les figures J.1 montrent dans le cas de bruits large bande que les méthodes

classiques restent les meilleures. Par contre, elles démontrent le fort gain des techniques sous-bandes par rapport aux techniques classiques de débruitage dans le cas de bruits focalisés stationnaires ou non. Elles montrent aussi le gain du modèle PBP par rapport au modèle FC dans le cas de faible SNR. Nous pouvons espérer de meilleures performances du modèle PBP avec des fonctions de pondération plus précises construites pour chaque phonème

Bibliographie

Adjoudani, A. & Benoît, C. (1996), On the integration of auditory and visual parameters in an hmm-based asr, pp. 461–471.

Allen, J. (1994), 'How do humans process and recognize speech ?', *IEEE Trans. on Speech and Audio Processing* **2**(4), 567–577.

Andrews, H. (1980), *Introduction to mathematical techniques in pattern recognition*, Wiley-intersciences.

Aran, J., Dancer, A., Dolmazon, J., Pujol, R. & Huy, P. T. B. (1988), *Physiologie de la cochlée*, INSERM, Paris.

Bengio, Y., Mori, R. D., Flammia, G. & Kompe, R. (1992), 'Global optimization of a neural network-hidden markov model hybrid', *IEEE Trans, on Neural Networks* **3**(2), 252–259.

Benoît, C. (1996), Intelligibilité audiovisuelle de la parole, pour l'homme et pour la machine, PhD thesis, Habilitation à diriger des recherches INP Grenoble.

Bernstein, L., Demorest, M. & Tucker, P. (1998), *What makes a good speechreader ? First you have to find one*, campbell edn, Hove(UK) : Psychology Press, chapter Hearing by eye, p. 211 228.

Berouti, N., Schwartz, R. & Makhoul, J. (1979), Enhancement of speech corrupted by acoustic noise, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', pp. 208–211.

Berrah, A., Glotin, H., Laboissière, R., Bessière, P. & Boë, J. (1996), From form to formation of phonetic structures : An evolutionary computing perspective, *in* 'Proceedings of the 13th International Congress on Machine Learning - Evolutionary computing and Machine Learning, ICML'96', MIT Press, Barni - Italy, pp. 23–29.

Berthommier, F. & Glotin, H. (1999), A new SNR-feature mapping for robust multistream speech recognition, *in* B. University Of California, ed., 'Proc. Int. Congress on Phonetic Sciences (ICPhS)', Vol. 1 of *XIV*, San Francisco, pp. 711–715.

Berthommier, F. & Glotin, H. (2000), Reconnaissance de la parole dans le bruit après renforcement fondé sur l'harmonicité, *in* 'Actes internationaux des Journées d'Etudes sur la Parole (JEP)', no IDIAP RR, see RESPITE www, Aussois.

Berthommier, F., Glotin, H. & Tessier, E. (2000), A front-end using the harmonicity cue for speech enhancement in loud noise, *in* 'Int. Conf. on Spoken Language Processing (ICSLP)'.

Berthommier, F., Glotin, H., Tessier, E. & Bourlard, H. (1998), Interfacing of casa and partial recognition based on a multistream technique, *in* 'Int. Conf. on Spoken Language Processing (ICSLP)', Vol. 4, Australian Speech Science and Technology Association, Incorporated (ASSTA), Sydney, pp. 1415–1419.

Berthommier, F. & Meyer, G. (1995), Source separation by a functional model of amplitude demodulation, *in* 'Eurospeech'95', pp. 135–138.

Berthoz, A. & Petit, L. (1996), 'Les mouvements du regard : une affaire de saccades', *La Recherche* **289**, 58–65.

Besacier, L. (1998), Un modèle parallèle pour la reconnaissance automatique du locuteur, PhD thesis, Université d'Avignon et des pays de Vaucluse.

Bishop, C. (1995), *Neural Network for Pattern Recognition*, Oxford University Press.

Boersma, P. (1993), Accurate short-term analysis of the fundamental frequency and the harmonics to noise ratio of the sampled sound, *in* 'Proc. IFA', Vol. 17, Amsterdam, pp. 97–110.

Boite, R., Bourlard, H., Dutoit, T., Hancq, J. & Leich, H. (2000), *Traitement de la parole*, collection electricité edn, Presses Polytechniques et universitaires Romandes.

Boll, S. (1979), 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Transactions on Acoustics, Speech and Signal Processing* **27**, 113–120.

Bourlard, H. (1996), Reconnaissance automatique de la parole : modélisation ou description ?, *in* 'Actes internationaux des Journées d'Etudes sur la Parole (JEP)', pp. 263–272.

Bourlard, H. (1998), Introduction à la reconnaissance de la parole et du locuteur, IDIAP-RR 13, IDIAP. Chapitre du livre Traitement de la Parole, à paraître aux Presses Polytechniques Universitaires Romandes.

Bourlard, H. (1999), 'Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR', *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions* .

Bourlard, H. & Dupont, S. (1996), A new ASR approach based on independent processing and recombination of partial frequency bands, *in* 'Proc. International Conference on Spoken Language Processing', Vol. 1, Philadelphia, pp. 426–429.

Bourlard, H., Dupont, S. & Ris, C. (1996), Multi-stream speech recognition, Technical report, IDIAP.

Bourlard, H. & Morgan, N. (1994), *Connectionist Speech Recognition. A Hybrid Approach*, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA.

Bourlard, H., Morgan, N. & Wellekens, C. (1990), Statistical inference in multilayer perceptrons and hidden markov models with applications in continuous speech recognition, *in* F. F. Soulié & J. Hérault, eds, 'Neurocomputing : Algorithms, Architectures and Applications', pp. 217–226.

Brandstein, M., Adcock, J. & Silverman, H. (1995), A closed-form method for finding source locations from microphone-array time-delay estimates, *in* 'Proceedings ICASSP95', pp. 3019–3022.

Bregler, C. & Konig, Y. (1994), 'Eigenlips' for robust speech recognition, *in* 'Proc. International Conference on Acoustics, Speech and Signal Processing', Adelaide, pp. 669–672.

Bregman, A. (1990), *Auditory scene analysis, the perceptual organization of sound*, MIT press, Cambridge.

Brooke, N. (1996), Talking heads and speech recognizers that can see : The computer processing of visual speech signals, pp. 351–371.

Brooke, N. M. & Scott, S. D. (1994), 'PCA image coding schemes and visual speech intelligibility', *Proc. Institute of Acoustics* **16**(5), 123–129.

Cerisara, C. (1999), Contribution de l'approche Multi-Bande à la reconnaissance automatique de la parole, PhD thesis, Doctorat de l'Institut National Polytechnique de Lorraine, Nancy, France.

Cerisara, C., Haton, J. & Fohr, D. (Septembre 1999), Towards a global optimization scheme for multi-band speech recognition, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', Budapest, Hongrie.

Cerisara, C., Haton, J., Mari, J. & Fohr, D. (1997), Multi-band continuous speech recognition, *in* 'Proc. Eurospeech '97', Rhodes, Greece, pp. 1235–1238.

Chibelushi, C. C., Deravi, F. & Mason, J. S. D. (1996), Survey of audio visual speech databases, Technical report, Department of Electrical and Electronic Engineering, University of Wales, Swansea.

Choi, S., Hong, H., Glotin, H. & Berthommier, F. (2000), Multichannel signal separation for cocktail party speech recognition : a dynamic recurrent network, *in* 'Int. Conf. on Spoken Language Processing (ICSLP)'.

Choi, S., Hong, H., Glotin, H. & Berthommier, F. (Sept 2001), 'Multichannel signal separation for cocktail party speech recognition : A dynamic recurrent network', *Neurocomputing, Special Issue, Blind Signal Separation and ICA* **invited.**

Choi, S., Lyu, Y., Berthommier, F., Glotin, H. & Cichocki, A. (1999), Blind separation of delayed and superimposed acoustic sources : learning algorithms an experimental study, *in* 'Proc. IEEE Int. Conference on Speech Processing (ICSP)', Seoul.

Cole, R., Noel, M., Lander, T. & Durham, T. (1995), New telephone speech corpora at cslu, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', Vol. 1, pp. 821–824.

Cooke, M., Green, P., Anderson, C. & Abberley, D. (1994), 'Recognition of occluded speech by hidden markov models'.

Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (1999), 'Robust ASR with unreliable data and minimal assumptions', *Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions* pp. 195–198.

Cooke, M., Morris, A. & Green, P. (1996), Recognising occluded speech, *in* 'ESCA Tutorial and Workshop on the Auditory Basis of Speech Perception', Keele University.

Crowley, J. & Demazeau, Y. (1993), Principles and techniques for sensor data fusion, *in* 'IEEE Signal Processing', Vol. 32, pp. 5–27.

Culling, J., Summerfield, Q. & Marshall, D. (1994), 'Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels', *SPEECHCOMM* **14**, 71–95.

d'Allessandro, C., Darsinos, V. & Yegnanarayana, B. (1998), 'Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources', *TransSAP* **6**(1), 12–23.

Davis, S. & Mermelstein, P. (1980), 'Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoustics, Speech, and Signal Processing* **28**(4), 357–366.

Dupont, S. (2000), Études et Développement de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole, PhD thesis, Doctorat de la Faculté Polytechnique de Mons.

Dupont, S., Bourlard, H. & Ris, C. (1997), 'Robust speech recognition based on multi-stream features', *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France* pp. 95–98.

Dupont, S. & Luettin, J. (1998), Using the multi-stream approach for continuous audio-visual speech recognition : Experiments on the M2VTS database, *in* 'Proc. International Conference on Spoken Language Processing', Sydney.

Dupont, S. & Luettin, J. (2000), 'Audio-visual speech modeling for continuous speech recognition', *IEEE Transactions on Multimedia* **2**(3), 141–151.

Ellis, D. (1997), Computational auditory scene analysis exploiting speech-recognition knowledge, *in* 'IEEE workshop on App. of Sig. Proc. to Aud. and Acous.', Mohonk.

Fletcher, H. (1953,[1929]), *Speech and Hearing in Communication*, Krieger, New York.

Gaillard, F. (1999), ANALYSE DE SCENES AUDITIVES COMPUTATIONNELLE (CASA) : Un nouvel outil de marquage du plan temps-frequence par detection d'harmonicite exploitant une statistique de passages par zero, PhD thesis, INPG, Grenoble.

Gaillard, F., Berthommier, F., Feng, G. & Schwartz, J. (1997), A modified zero-crossing method of pitch detection in interfering conditions, *in* 'proc. of Eurospeech'97', Rhodes.

Gales, M. & Young, S. (1992), An improved approach to hidden Markov model decomposition, *in* 'Proc. International Conference on Acoustics, Speech and Signal Processing', San Francisco, pp. 729–734.

Ghitza, O. (1986), 'Auditory nerve representation as a front end for speech recognition in noisy environments', *Computer, Speech and Language* **1**, 109–130.

Giard, M. H., Collet, L., Bouchet, P. & Pernier, J. (1994), 'Auditory selective attention in the human cochlea', *Brain Research* **633**, 353–356.

Gibson, J. (1986), *The Ecological Approach to Visual Perception*, lawrence erlbaum edn, Hillsdale.

Glotin, H. (1995), La vie artificielle d'une société de robots parlants : émergence et changements du code phonétique - AGORA, Mémoire de DEA sciences cognitives - master thesis, Institut National Polytechnique de Grenoble.

Glotin, H. (2000), Various adaptive weighting schemes for large vocabulary robust audio-visual ASR, with particular reference to the cocktail party effect, IDIAP-COM 4, IDIAP.

Glotin, H. (2001*a*), Dominant speaker detection in cocktail party noise for robust multi-stream speech recognition, *in* 'Adaptation methods in speech recognition, ISCA Workshop'.

Glotin, H. (2001*b*), Optimal fusion of expert's confidence and speech reliability for robust multi-stream asr : the bias prediction model, *in* 'IEEE International Workshop on Intelligent Signal Processing'.

Glotin, H. & Berthommier, F. (2000), Test of several external posterior weighting functions for multiband full combination ASR, *in* 'Int. Conf. on Spoken Language Processing (ICSLP)', Beijing-China.

Glotin, H., Berthommier, F. & Tessier, E. (1999), A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition, *in* 'EUROSPEECH', Vol. 5, pp. 2351–2354.

Glotin, H., Berthommier, F., Tessier, E. & Bourlard, H. (1998), Interfacing of CASA and multistream recognition, *in* P. Sojka, ed., 'TSD'98-Text, Speech and Dialog International Workshop', Masaryk University, Brno, BRNO-Czech Republic, pp. 207–212.

Glotin, H. & Laboissière, R. (1996), Emergence du code phonétique dans une société de robots parlants, *in* 'Actes de la conférence modélisation des systèmes naturels complexes ROCHEBRUNE'96', Ecole Nat. Supérieure des Télécommunications, Paris, pp. 113–125.

Glotin, H., Pinel, P., Cochard, J. & Laboissière, R. (1997), Dictyostelium discoideum : passage d'une identité individuelle à une identité collective, *in* 'Actes de la conférence modélisation des systèmes naturels complexes : Invariance, Interaction, Référence : l'identité en question ROCHEBRUNE'97', Ecole Nat. Supérieure des Télécommunications, Paris, pp. 49–68.

Glotin, H., Tessier, E., Bourlard, H. & Berthommier, F. (1998*a*), Reconnaissance multi-bandes de la parole bruitée par couplage entre niveaux primitif et d'identification, *in* 'Actes internationaux des Journées d'Etudes sur la Parole (JEP)', XXII, pp. 375–378.

Glotin, H., Tessier, E., Bourlard, H. & Berthommier, F. (1998*b*), Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes, *in* F. Alexandre & D. Kant, eds, 'IX ème Journées Neurosciences et Sciences de l'Ingénieur', LORIA/INRIA, Munster.

Glotin, H., Vergyri, D., Neti, C., Potamianos, G. & Luettin, J. (2000), Weighting schemes for audio-visual fusion in speech recognition, IDIAP-RR 44, IDIAP. published in IEEE International Conference on Acoustic, Speech, and Signal Processing.

Glotin, H., Vergyri, D., Neti, C., Potamianos, G. & Luettin, J. (2001), Weighting schemes for audio-visual fusion in speech recognition, *in* 'IEEE ICASSP'.

Green, P., Cooke, M. & Crawford, M. (1995*a*), 'Auditory scene analysis and hidden markov model recognition of speech in noise', *TransSP* pp. 401–404.

Green, P., Cooke, M. & Crawford, M. (1995*b*), Auditory scene analysis and hmm recognition of speech in noise, *in* 'ICASSP', pp. 401–404.

Hagen, A. & Glotin, H. (2000), Études comparatives des robustesses au bruit de l'approche 'full combination' et de son approximation, *in* 'Actes internationaux des Journées d'Etudes sur la Parole (JEP)', Aussois, France. IDIAP-RR 00-04.

Hagen, A., Morris, A. & Bourlard, H. (1999), Different weighting schemes in the full combination sub-bands approach in noise robust ASR, *in* 'Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions'.

Hansen & Cairns (1995), 'Icarius : Source generator based real-time recognition of speech in noisy stressful and lombard', *Speech Communication* **16**(4).

Hansen & Clements (1989), Stress compensation and noise reduction algorithms for robust speech recognition, *in* 'ICASSP', pp. 266–269.

Heckmann, M., Berthommier, F. & Kroschel, K. (2001), Optimal weighting of posteriors for audio-visual speech recognition, *in* 'ICASSP'.

Hennebert, J., Ris, C., Bourlard, H., Renals, S. & Morgan, N. (1997), Estimation of global posteriors and foward-backward training of hybrid hmm/ann systems, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', Vol. 4, pp. 1951–1954.

Hennecke, M. E., Stork, D. G. & Prasad, K. V. (1996), Visionary speech : Looking ahead to practical speechreading systems, *in* D. G. Stork & M. E. Hennecke, eds, 'Speechreading by Humans and Machines : Models, Systems and Applications', Vol. 150 of *NATO ASI Series F : Computer and Systems Sciences*, Springer-Verlag, Berlin, pp. 331–349.

Hermansky, H. (1990), 'Perceptual linear predictive (PLP) analysis of speech', *Journal of the Acoustical Society of America* **87**(4), 1738–1752.

Hermansky, H. & Junqua, J. (1988), Optimization of perceptually-based ASR front end, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', New York, pp. 219–223.

Hermansky, H. & Morgan, N. (1994), 'RASTA processing of speech', *IEEE Transactions on Speech and Audio Processing* **2**(4), 578–589.

Hermansky, H., Morgan, N., Bayya, A. & Kohn, P. (1992), 'Rasta-plp speech analysis technique', *IEEE Trans. on Signal Processing* **1**, 121–124.

Hermansky, H., Morgan, N. & Hirsch, H. (1990), Recognition of speech in additive and convolutional noise based on rasta spectral processing, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', Vol. 2, pp. 83–86.

Hermansky, H., Tibrewala, S. & Pavel, M. (1996), Towards ASR on partially corrupted speech, *in* 'Int. Conf. on Spoken Language Processing (ICSLP)', pp. 462–465.

Hess, W. (1983), *Algorithms and devices for pitch determination of speech signal*, Springer Verlag.

Hess, W. (1992), *Advances in speech Signal Processing*, furui and sondhi edn, Marcel Dekker, New York, chapter Pitch and voicing determination, pp. 3–48.

Hirsch, H. (1993), Estimation of noise spectrum and its application to SNR-estimation and speech enhancement, Technical Report TR-93-012, International Computer Science Institute, Berkeley, CA.

Hirsch, H. & Erlicher, C. (1995), Noise estimation techniques for robust speech recognition, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', Detroit MI, pp. 153–156.

Hofstadter, D. (1985), *Gödel Escher Bach : Les brins d'une guirlande éternelle*, InterEditions.

Hong, H., Choi, S., Glotin, H. & Berthommier, F. (2000), Blind acoustic source separation for cocktail party speech recognition, *in* IEEE, ed., 'ICONIP, 7th IEEE Int. Conf. on Neural Information Processing', Korea.

Immerseel, L. V. & Martens, J. (1992), 'Pitch and voiced/unvoiced determination with an auditory model', *JASA* **91**(6), 3511–3526.

Josifovski, L., Cooke, M., Green, P. & Vizinho, A. (1999), State base imputation of missing data for robust speech recognition and speech enhancement, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)'.

Jourlin, P. (1997), Word dependent acoustic-labial weights in HMM-based speech recognition, *in* 'Proc. European Tutorial Workshop on Audio-Visual Speech Processing (AVSP)', Rhodes, pp. 69–72.

Jourlin, P. (1998), Approche bimodale du traitement de la parole : application à la reconnaissance du message et du locuteur, PhD thesis, Université d'Avignon et des Pays du Vaucluse.

Juang, B. (1991), 'Speech recognition in adverse environments', *Computer, Speech and Language* **5**, 275–294.

Junqua, J. (1995), The influence of acoustics on speech production : a noise-induces stress phenomenon known as the lombard reflex, *in* 'ESCQ-NQTO Workshop on Speech under Stress', Lisbon, pp. 83–90.

Junqua, J. & Haton, J. (1996), *Robustness in automatic speech recognition*, Kluwer Academic Publishers.

Jutten, C. & Hérault, J. (1991), 'Blind separation of sources, part i : An adaptive algorithm based on neuromimetic architecture', *Signal Processing* (24), 1–10.

Kajita, S. & Itakura, F. (1995), Robust speech feature extraction using sbcor analysis, *in* 'ICASSP'.

Kandel, E., Schwartz, J. & Jessel, T. (1991), *Principles of Neural Science*, 3rd edn, Appleton and Lange.

Kermorvant, C. (1999), A comparison of noise reduction techniques for robust speech recognition, RR 10, IDIAP.

Kingsbury, B. (1998), Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments, PhD thesis, University of California at Berkeley.

Kittler, J., Li, Y., Matas, J. & Sanchez, M. R. (1997), *AVBPA*, Springers LNCS, chapter Combining evidence in multimodal personal identity recognition systems, pp. 327–334.

Klingholz, F. (1987), 'The measurement of the signal to noise ration (snr) in continuous speech', *Speech Communication* **6**, 15–26.

Koehler, J., Morgan, N., Hermansky, H., Guenter, H. G. & Tong, G. (1994), 'Integrating rasta-plp into speech recognition', *IEEE Trans. on Signal Processing* **1**, 421–424.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *Ann. Math. Stat.* **22**, 79–86.

Langner, G., Sams, M., Heil, P. & Schulze, H. (1997), 'Frequency and periodicity are represented in orthogonal maps in the human auditory cortex : Evidence from magnetoencephalography', *J.Comp.Physiol* .

Langner, G. & Schreiner, C. (1988), 'Periodicity coding in the inferior colliculus of the cat', *J.Neurophysiol* **60**, 1799–1822.

Licklider, J. (1959), *Psychology : a study of a science*, Vol. 1, mcgraw hill edn, New York, chapter Three auditory theories, pp. 41–144.

Lindblom, B. & Lubker, J. (1985), *Phonetics Linguistics,Essays in Honor to Peter Ladefoged*, v.a. fromkin edn, Academic Press, Orlando, chapter The speech Homonculus and a problem of Phonetics Linguistics, pp. 169–192.

Lippmann, R. & Carlson, B. (1997), Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise, *in* ESCA, ed., 'Eurospeech'97', pp. 37–40.

Lockwood, P. & Boudy, J. (1991), Experiments with a non-linear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in car, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', pp. 79–82.

Lockwood, P. & Boudy, J. (1992), 'Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars.', *Speech Communication* **11**, 215–228.

MacLeod, A. & Summerfield, Q. (1987), 'Quantifying the contribution of vision to speech perception in noise', *British Journal of Audiology* **21**, 131–141.

Martin, R. (1993), An efficient algorithm to estimate the instantaneous SNR of speech signals, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', pp. 1093–1096.

Massaro, D. (1987), *Speech perception by ear and eye : A paradigm for psychological inquiry*, laurence erlbaum associates edn, London.

Massaro, D. W. & Stork, D. G. (1998), 'Speech recognition and sensory integration', *American Scientist* **86**(3), 236–244.

Mauuary, L. (1996), Blind equalization for robust telephone based speech recognition, *in* 'Proc. European Signal Processing Conference (EUSIPCO)'.

Mauuary, L. (1998), Blind equalization in the cepstral domain for robust telephone based speech recognition, *in* 'Proc. European Signal Processing Conference (EUSIPCO)'.

McGurk, H. & MacDonald, J. (1976), 'Hearing lips and seeing voices', *Nature* **264**, 746–748.

Meier, U., W.Hurst & Duchnowski, P. (1996), 'Adaptive bimodal sensor fusion for automatic speechreading', *IEEE International conference on Acoustics, Speech and Signal Processing* .

Miller, G. & Nicely, P. (1955), 'An analysis of perceptual confusions among some english consonants', *Journal of the Acoustical Society of America* **27**, 338–352.

Mirghafori, N. (1997), Multi-band speech recognition : A summary of recent work at icsi, Technical Report TR-97-051, International Computer Science Institute, Berkeley, CA.

Mirghafori, N. & Morgan, N. (1998), Transmission and transitions : A study of two common assumptions in multi-band ASR, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)'.

Mokbel, C. (1992), Reconnaissance de la Parole dans le Bruit : Bruitage/Débruitage, PhD thesis, Ecole Nationale Supérieure des Télécommunications.

Mokbel, C., Jouvet, D. & Monné, J. (1996), 'Deconvolution of telephone line effects for speech recognition', *Speech Communication* **19**, 185–196.

Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monne, J., Simonin, J. & Bartkova, K. (1997), 'Towards improving ASR robustness for psn and gsm telephone applications', *Speech Communication* **23**, 141–159.

Moore, R. (1986), Signal decomposition using markov modeling techniques, Technical Report 3931, Royal Sig. Res. Estab.

Morgan, N. & Bourlard, H. (1995), 'Continuous speech recognition : An introduction to the hybrid hmm/connectionnist approach', *IEEE Signal Processing Magazine* **12**(3).

Morris, A., Hagen, A., Glotin, H. & Bourlard, H. (2001), 'Multi-stream adaptive evidence combination for noise robust ASR', *Speech Communication, Special Issue on Noise Robust* **34**(1-2).

Nadas, A., Nahamoo, D. & Picheny, M. (1989), 'Speech recognition using noise adaptive prototypes', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**, 1495–1503.

Neti, C. (1994), Neuromorphic speech processing for noisy environments, *in* 'Proc. IEEE International Conference on Neural Networks', Orlando, pp. 4425–4430.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J. & Mashari, A. (2000), Audio visual speech recognition, Technical report, Johns Hopkins University-CLSP.

Okawa, S., Bocchieri, E. & Potamianos, A. (1998), Multi-band speech recognition in noisy environment, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)'.

Okawa, S., Nakajima, T. & Shirai, K. (1999), A recombination strategy for multi-band speech recognition based on mutual information criterion, *in* 'Proc. European Conference on Speech Communication and Technology (EUROSPEECH)', Vol. 2, Budapest, pp. 603–606.

O'Leary, D., Andreasen, N., Hurtig, R., Hichwa, R., Watkins, G., Ponto, L., Rogers, M. & Kirchner, P. (1996), 'A positron emission tomography study of binaurally- and dichotically-presented stimuli : Effects of level of language and directed attention', *Brain and Language* **53**, 20–39.

Potamianos, G. & Graf, H. P. (1998), Discriminative training of HMM stream exponents for audio-visual speech recognition, *in* 'Proc. International Conference on Acoustics, Speech and Signal Processing', Vol. 6, Seattle, pp. 3733–3736.

Potamianos, G., Verma, A., Neti, C., Iyengar, G. & Basu, S. (2000), A cascade image transform for speaker independent automatic speechreading, *in* 'Proc. International Conference on Multimedia and Expo', Vol. II, New York, pp. 1097–1100.

Pratibha, J. & Hermansky, H. (2000), Temporal patterns of critical-band spectrum for text-to-speech, *in* 'ICSLP'.

Pujol, R. (1999), Promenade autour de la cochlée. www.iurc.montp.inserm.fr.

Rabiner, L. (1977), 'On the use of autocorrelation analysis for pitch detection', *TransASSP* **25**, 24–33.

Rabiner, L. (1989), A tutorial on hidden markov models and selected applications in speech recognition, *in* 'Proceedings of the IEEE', Vol. 77, pp. 257–285.

Rabiner, L. & Juang, B. (1993), *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, PTR Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632.

Rathinavelu, C. & Deng, L. (1997), 'Hmm-based speech recognition using state-dependent, discriminatively-derived transforms on mel-warped dft features', *IEEE Transactions on Speech and Audio Processing* **5**(3), pp. 243–256.

Reibhard, H. (1996), *Eléments de mathématiques du signal*, Dunod, Paris.

Richard, M. & Lippmann, R. (1991), 'Neural network classifiers estimate bayesian a posteriori probabilities', *Neural Computation* pp. 461–483.

Ris, C. & Dupont, S. (2001), 'Assessing local noise level estimation methods : application to noise robust ASR', *Speech Communication, Special Issue on Noise Robust* **34**. à paraître.

Robert-Ribes, J. (1995), Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles, PhD thesis, Institut Nationale Polytechnique de Grenoble.

Rogozan, A. (1999), Etude de la fusion des données hétérogènes pour la reconnaissande automatique de la parole audiovisuelle, PhD thesis, Orsay-Paris XI.

Rogozan, A., Deléglise, P. & Alissali, M. (1997), Adaptive determination of audio and visual weights for automatic speech recognition, *in* 'Proc. European Tutorial Workshop on Audio-Visual Speech Processing (AVSP)', Rhodes, pp. 61–64.

Schwartz, J. (2001), *Traitement automatique du langage parlé*, j. mariani edn, Hermes, chapter La parole multimodale :deux ou trois sens valent mieux qu'un.

Senior, A. W. (1999), Face and feature finding for a face recognition system, *in* 'Proc. 2nd International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)', Washington, pp. 154–159.

Shannon, C. (1948), A mathematical theory of communication, Technical Report 27, Bell System Technical Journal.

Shynk, J. (1992), 'Frequency-domain and multirate adaptive filtering', *IEEE Signal processing magazine* pp. 15–37.

Steeneken, H. & Houtgast, T. (1991), On the mutual dependency of octave-band-specific contribution to spech intelligibility, *in* 'Proc. European Conf. on Speech Communication and Technology (EUROSPEECH)', pp. 1133–1136.

Steeneken, J. & Leeuwen, D. V. (1995), Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems : the sqale project (speech recognition quality assessment for language engineering), *in* 'Eurospeech', pp. 1271–1274.

Taniguchi, T., Kajita, S., Takeda, K. & Itakura, F. (1998), 'Blind signal separation for recognizing overlapped speech', *JASA* **19-6**.

Teissier, P. (1999), Fusion de capteurs avec contrôle du contexte : application à la reconnaissance audiovisuelle de parole dans le bruit, PhD thesis, INPG.

Teissier, P., Robert-Ribes, J., Schwartz, J. & Guérin-Dugué, A. (1999), 'Comparing models for audiovisual fusion in a noisy-vowel recognition task', *IEEE Trans. Speech and Audio Processing* **7**, 629–642.

Teissier, P., Robert-Ribes, J. & Schwartz, J. L. (1999), 'Comparing models for audiovisual fusion in a noisy-vowel recognition task', *IEEE Transactions on Speech and Audio Processing* **7**(6), 629–642.

Tessier, E. (2001), Localisation de sources de parole interférentes et Analyse de Scène Auditive, PhD thesis, Doctorat de l'Institut National Polytechnique de Grenoble.

Tessier, E., Berthommier, F., Glotin, H. & Choi, S. (1999), A casa front-end using the localisation cue for segregation and then cocktail-party speech recognition, *in* 'Proc. IEEE Int. Conference on Speech Processing (ICSP)', IEEE, Seoul.

Tibrewala, S. & Hermansky, H. (1997), Multi-band and adaptation approaches to robust speech recognition, *in* 'Proc. Eurospeech '97', Rhodes, Greece, pp. 2619–2622.

Varga, A. & Moore, R. (1990*a*), Hidden markov model decomposition of speech and noise, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', pp. 845–848.

Varga, A., Steeneken, H., Tomlinson, M. & Jones, D. (1992), 'The noisex-92 study on the effect of additive noise on automatic speech recognition', *Technical Report, DRA Speech Research Unit* .

Varga, P. & Moore, R. K. (1990*b*), Hidden Markov model decomposition of speech and noise, *in* 'Proc. International Conference on Acoustics, Speech and Signal Processing', Albuquerque, pp. 845–848.

Warren, R. & Sherman, G. (1974), 'Phonemic restorations based on subsequent context', *Perception and Psychophysics* .

Yen, K. & Zhao, Y. (1999*a*), 'Adaptive co-channel speech separation and recognition', *IEEE Trans. Speech Audio Processing* **7**(2), 138–151.

Yen, K. & Zhao, Y. (Mar. 1999*b*), Adaptive decorrelation filtering for separation of co-channel speech signals from m $>$ 2 sources, *in* 'Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', Vol. 2, pp. 801–804.

Yen, K. & Zhao, Y. (May 1998), Improvements on co-channel speech separation using adf : Low complexity, fast convergence, and generatlization, in *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (), pp. 1025–1028.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (1999), *The HTK Book*, Entropic Ltd., Cambridge.

Yumoto, E., Gould, W. & Baer, T. (1982), 'Harmonic to noise ratio as an index of the degree of hoarseness', *Journal of the Acoustical Society of America* **1971**, 1544–1550.

(Cette thèse contient 176 références bibliographiques.)

Index

Elaboration et comparaison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole : incorporation des indices de voisement et de localisation

Cette thèse effectuée à l'ICP et à l'IDIAP, dans le champ de la communication homme-machine et des projets EU. RESPITE & SPHEAR, contribue à augmenter la robustesse de reconnaisseur automatique de la parole dans le cadre original de l'analyse de scènes auditives. Deux voies sont traitées simultanément : (1) l'extraction d'indices fiables du signal et (2) la fusion de données dans le cadre multi-flux. (1) est fondée sur des mesures temps-fréquences de corrélations, relatives au taux de voisement ou aux localisations de sources. Nous montrons comment l'indice de voisement renforce le prétraitement de référence « Jrasta ». (2) est proposée via un modèle "combinaison complète" qui intègre par combinaisons de sous-bandes du spectre les redondances spectrales de la parole. Ce modèle est approximé avec une hypothèse faible d'indépendance des observations des sous-flux du spectre. La robustesse d'un système de reconnaissance hybride ANN/HMM, de chiffres téléphonés (NB95), est alors renforcée dans le cas de paroles simultanées (enregistrements stéréo), ou contre des bruits non stationnaires focalisés. Nous validons dans le cas de bruit de cafétéria, l'apport de l'indice de voisement pour la reconnaissance audiovisuelle grand vocabulaire (base Via Voice-IBM,MMG asynchrone). Nous proposons de plus un modèle de « Prédiction des Biais des Posteriors » guidé par les indices dont les premiers tests sont prometteurs. Nous comparons finalement ces différentes architectures, et en proposons une, dite « proactive », qui permet l'intégration d'informations complémentaires.
Mots-clés : reconnaissance automatique de la parole, IHM, robustesse au bruit, cocktail party, multi-bande, multi-flux, fusion de données, analyse de scène auditive, harmonicité, voisement, audiovisuel, Via Voice, IBM, MMG, localisation, HMM, modèle hybride, calcul bayesien, perceptron, ANN, réseau de neurones, Jrasta.

Comparative study and elaboration of robust adaptive multistream automatic speech recognition using voicing and localization cues

This thesis, taking part in European projects RESPITE and SPHEAR,between ICP-IDIAP, shows various means to reinforce automatic speech recognition (ASR), based on (1) Computational Auditory Scene Analysis and (2) multistream paradigm. In(1)we propose different reliability factors calculated from time-frequency correlations, related to voicing level or a localization cue in case of stereo data. We show how voicing cue reinforces the state of the art preprocessing « Jrasta ».In(2)we propose a multistream ASR model, called « Full Combination » (FC) that exploits spectral redundancy, while making minimum assumptions about noise type by considering every combination of data subbands as an independent data stream. Using our reliability factors and hybrid ANN/HMM ASR, we demonstrate under Numbers95 telephonic free digits data base, that FC is robust to non-stationary noise and to cocktail party effect. Furthermore, we develop a promising fusion model for the voicing cue in a multistream audiovisual large vocabulary ASR and efficient under cafeteria noise (tested on Via Voice IBM database, asynchronous GMM). After having analyzed FC errors we propose the Posteriors Bias Prediction model which gives an optimal fusion model for signal and estimates' reliabilities. First tests are promising. Then we compare these different ASR architectures, and we propose a « proactive » one, which allows integration of complementary information for robust ASR.

Key-words: automatic speech recognition, robustness, noise, cocktail party, multiband, multistream, data fusion, auditory scene analysis, harmonicity, voicing, audiovisual,Via Voice, IBM, localization, HMM, bayes, perceptron, ANN, GMM, neural network, Jrasta.

Spécialité : Informatique-Sciences Cognitives
Intitulés et adresses des laboratoires : Inst. de la Communication Parlée (ICP)-INPG-46 av. Viallet-38031 Grenoble Cedex & Inst. d'Intelligence Artificielle Perceptive (IDIAP)-Simplon 4-1920 Martigny-Suisse