THÈSE

présentée par

NEAGU Adrian

pour obtenir le grade de DOCTEUR

de l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Spécialité : Signal - Image - Parole

Représentations phonétiques et identification des syllabes occlusive-voyelle en français

Date de soutenance :

30 avril 1998

Composition du jury :

Pascal PERRIER

Jacqueline VAISSIÈRE

Jean François MARI

Régine ANDRÉ OBRECHT

Gérard BAILLY

Institut de la Communication Parlée

46 av. Félix Viallet, 38031 Grenoble CEDEX 1 France

À mon père

Table des matières

Table des figures

Liste des tableaux

Remerciements

Mon travail à l'Institut de la Communication Parlée n'aurait pu aboutir sans la collaboration et le soutien de nombreuses personnes. Ces lignes essayent d'exprimer ma gratitude envers ceux qui m'ont permis de le mener à bien.

Avant tout, je tiens à remercier mon directeur de thèse, M. Gérard BAILLY. Je veux l'assurer ici de ma profonde reconnaissance. Dès le jour où il a accepté de diriger ce travail, et même après la dernière ligne ajoutée au mémoire, ses conseils, ses critiques, sa patience, son soutien ont été essentiels pour son avancement.

Je remercie vivement les membres du jury pour l'honneur qu'ils me font en y participant et pour avoir accepté de faire des rapports sur ce travail.

Je remercie également M. Jean-François SERIGNAT pour m'avoir amené à l'I.C.P. et pour la façon dont il m'a encadré pendant le D.E.A.

Je remercie mes collègues jeunes chercheurs Paul Munteanu, Bertrand Caillaud, Olivier Delemar, Yann Morlec, François Gaillard et les autres pour leur aide dans des situations divers allant de la correction d'une obscure erreur de compilation à un vif coup de main pour un déménagement.

Je remercie Nino MEDVES et les autres membres des équipes techniques et administratives de notre laboratoire pour le fait de nous assurer chaque jour un environnement lucratif et agréable.

Merci à tous ceux, trop nombreux pour être cités ici, qui, pendant ces années, m'ont supporté et ont fait les frais d'un long processus d'adaptation et d'intégration. Enfin, un grand merci à ma femme.

Chapitre 1

Introduction

1.1 Motivation. État du traitement automatique de la parole

La parole véhicule un flux continu d'informations. Par ce canal, le locuteur communique à son interlocuteur une suite d'unités discrètes, choisis dans un inventaire réduit d'items lexicaux combinables à l'infini dans les limites des constructions syntaxiques et semantiques possibles dans la langue. Par contraste, le signal sonore prend une grande variété de formes pour la transmission d'un même suite de symboles selon le locuteur, ses états mentaux, les conditions de communication, etc. Ce contraste entre la nature continue du signal et la nature discrète du message délivré, de même que l'importante variabilité du signal, a longuement empêché les machines de dépasser un niveau de performance ridicule comparé à l'aisance des humains sur des tâches comparables.

Ce manque de performances des systèmes artificiels a propulsé les activités de recherche fondamentale sur la parole, activités structurées comme un domaine d'investigation à part entière situé au croisement des sciences de l'ingénieur et des sciences de l'homme. Trois sources de progrès sont à la base des avancés les plus significatives de ce domaine :

- La maîtrise des algorithmes de traitement du signal, de reconnaissance des formes et de la modélisation du langage naturel
- Le progrès technologique des machines pour le traitement automatique de l'information.
- L'évolution de nos connaissances sur la perception et la production de la parole et de manière générale, sur la communication homme - homme.

Et les progrès réalisés sont sensibles. Pour mémoire, examinons les progrès accomplis en codage de la parole, en synthèse et reconnaissance de la parole par les machines. Aujourd'hui la parole garde son intelligibilité à un débit largement inférieur à celui de la ligne téléphonique classique (4,8 kbps par rapport à 33,6 kbps) et une qualité remarquable à 9,6 kbps (dans les GSM[1]). Les logiciels de commande vocale (reconnaissance de mots isolés dans un vocabulaire limité) et de synthèse à partir du texte (par unités stockées) sont devenu banals : ils accompagnent chaque ordinateur multimedia vendu dans les grandes surfaces commerciales. Plusieurs logiciels (IBM, Dragon Systems, Kurtz-

[1] GSM (Groupe Société Mobile) est le standard européen pour les appareils téléphoniques portables

weil, etc) de dictée vocale à grand vocabulaire (>20 000 mots) sont disponibles aujourd'hui pour moins de 800$, traitant toutes les langues de circulation internationale. Très récemment, ils se sont affranchis de la contrainte de séparation des mots et le temps d'initialisation nécessaire pour chaque locuteur est passé de plusieurs heures à une dizaine de minutes.

Vu cet état des applications, on peut se poser la question sur la nécessité de la recherche fondamentale sur la parole. Sans chercher à établir une hiérarchie entre les trois directions principales énumérées plus haut, nous voulons souligner par la suite le rôle joué par la dernière direction. Il y a eu, bien sûr, les retombés immédiates, comme l'étude de l'appareil auditif périphérique pour le codage à bas débit de la parole. Mais, plus important, l'étude du fonctionnement de l'humain dans le processus de la communication nous donne un cadre fertile pour attaquer une large classe de difficultés liées aux systèmes automatiques. Nous pensons, par exemple, au manque de rigueur dans la formulation initiale du problème de la reconnaissance de la parole. En effet, la communication en situation réelle comporte une négociation continue entre le locuteur et l'auditeur. Cette négociation fait intervenir des stratégies d'adaptation des deux côtés, stratégies que les premiers systèmes automatiques ignoraient complètement. De même, cette négociation utilise des canaux alternatifs comme la vision, elle aussi longuement négligée dans les systèmes automatiques.

Aujourd'hui les avions volent très vite employant une manière assez différente de celle utilisée par les oiseaux. Mais beaucoup de prototypes sont tombés avant que nous y arrivions. Dans la parole, les premiers systèmes automatiques sont "tombés" tellement vite qu'ils ont rappelés à leur concepteurs combien il est utile d'observer attentivement les oiseaux avant de faire de maquettes volantes. Et, si nous avons aujourd'hui ces systèmes capable de "voler", nous sommes encore loin des performances qu'on peut demander de manière justifiée de la part des systèmes automatiques. En effet, un auditeur moyen peut comprendre dans sa langue maternelle, sans trop de difficultés, une phrase qu'il n'a jamais entendue auparavant prononcée par un locuteur qu'il entend pour la première fois, phrase prononcée sur le fond d'autres conversations simultanées.

Le travail de cette thèse s'inscrit donc dans le cadre de cette recherche fondamentale qui vise, à court terme, une connaissance plus fine de la production et la perception de la parole et, à terme, l'amélioration de la robustesse et de la flexibilité des systèmes automatiques.

1.2 Pourquoi faire de la modélisation acoustique ?

Le décodage acoustico - phonétique doit faire face à toutes les sources de variabilité du signal de parole. Certaines sont liées aux caractéristiques anatomiques et physiologiques des locuteurs, d'autres sont liées au contexte phonétique ou bien aux contraintes d'environnement et à la négociation locuteur - auditeur et d'autres encore sont porteuses d'informations linguistiques et paralinguistiques. L'approche dominante aujourd'hui pour maîtriser cette variabilité dans la reconnaissance automatique de la parole est basée sur une représentation du signal par trames équidistantes et sur un algorithme de reconnaissance de séquences par chaînes de Markov. La force de cette approche est liée à la capacité d'apprentissage des modèles Markoviens. À condition de disposer d'une base d'exemples conséquente et d'accepter un nombre important de paramètres stockés, le modèle issu de l'apprentissage capte les caractéristiques des signaux à représenter. Cependant, il s'agit d'une modélisation statistique "aveugle" dans le sens où toutes les sources de variabilité sont traitées uniformément. Par conséquent, le système tolère mal une source absente de l'ensemble d'apprentissage.

Plusieurs chercheurs [Bou95, DRS96, RSS96] s'accordent à estimer que cette approche touche ses limites aujourd'hui. L'avancée suivante est attendue principalement d'une meilleure modélisation du signal de parole et moins d'une amélioration de l'appareil statistique en lui-même :

> " ... cette méthodologie de recherche a conduit à des systèmes de reconnaissance de la parole se focalisant surtout sur la description du signal de parole plutôt que sur sa véritable modélisation, résultant ainsi en des mauvaises propriétés de généralisation." [Bou96]
> " ... one key source of ideas should come from global speech production models which are capable of simulating key mechanisms of the human speech communication process ... " [DRS96]
> "The fields of phonetics and speech technology both have much to gain from a better interdisciplinary integration." [Nea96]

S'il est improbable que toute la variance observée dans le signal de parole soit modélisable de manière explicite, un consensus semble s'être établi sur la nécessité d'en expliquer une partie essentielle par des modèles spécifiques à la production et à la perception de la parole. Or les systèmes automatiques de reconnaissance (ASR) de l'approche décrite auparavant peuvent servir sans modification fondamentale pour apprendre, par exemple, le chant des oiseaux (même le modèle de langage par n-grams ne subira pas beaucoup de modifications, ce niveau de description étant pourtant une partie supposée très "spécifique"). Par contraste, pour la perception humaine :

> " ... evidence strongly suggests that phonetic percepts are never knowledge-innocent records of the raw signal. That conclusion ... applies also to the perception of elementary phonetic stimuli." [Lin96, page 1683]

Le but déclaré de cette thèse est donc de chercher les connaissances a priori qui rendront les ASR vraiment spécifiques à la parole. Le problème comporte deux aspects fondamentaux :

1. L'identification des connaissances pertinentes, des contraintes de perception et de production aptes à conditionner la sortie d'un algorithme d'extraction de paramètres du signal vocal.
2. La détermination d'un cadre computationel permettant d'effectuer cette opération.

Cette deuxième direction représente un problème typique d'algorithmique. Nous rappelons ici son défi principal : le réglage du compromis entre la confiance accordée aux connaissances a priori et celle accordée aux événements observés directement sur le signal. En effet, un contrôle trop serré de l'algorithme peut nous donner des sorties complètement rompues de la réalité du signal tandis qu'un contrôle trop faible permet des sorties non conformes au modèle, impossible à classer. Sans la négliger complètement (voir notre algorithme de suivi de formants), cette direction de recherche nous a moins préoccupé dans cette thèse.

1.3 Invariance acoustique. Quelques hypothèses de travail

Nous nous sommes concentrés sur la première direction de recherche. Il s'agit d'un travail de recherche et d'adaptation des modèles de production par le biais de réduction et sélection de paramètres.

Mais avant même de lancer cette investigation, il faut s'assurer qu'il existe au moins une chance minime de trouver là où nous voulons chercher. Un passage en revue des théories majeures dans la perception de la parole nous renseigne beaucoup sur la nature et la quantité de l'information portée par le signal de parole. Concernant la relation entre une représentation quasi-continue du signal, soit-elle articulatoire ou acoustique et une suite de symboles, on trouve deux théories opposées :

- "Strong articulatory relation"
- "Strong acoustic relation"

Dans la première catégorie, la théorie motrice de la perception de Liberman et Mattingly [LM85, Mat91] et le réalisme direct de Fowler [Fow86, Fow96], inspirée des travaux eu J.J. Gibson sur la perception humaine en général, postulent une relation forte entre les gestes articulatoires et les symboles. Les gestes sont les objets ultimes de la perception :

> "Putting together all the generalizations about the multiplicity and variety of acoustic cues, we should conclude that there is simply no way to define a phonetic category in purely acoustic terms." [LM85, page 12]

Une conséquence importante de cette théorie est le fait que l'information n'est pas toute contenue dans le signal. L'existence d'une plate-forme permettant l'inversion acoustico - articulatoire serait un pré-requis pour le décodage phonétique.

Par opposition, la deuxième catégorie des théories postulent l'invariance au niveau acoustique :

> " ... ultimately, it is to the auditory form of the utterance that the condition of sufficient discriminability applies." [Lin96, page 1690]
>
> " ... speech perception does not involve recovering the articulations that generated the acoustic signal." [Oha96, page 1718]
>
> "Our data stongly support a direct acoustic/auditory-based utilisation of the speech waveform to recover the phonetic segment." [SMM91, page 1321]

Donc, selon cette théorie, l'information utile est toute contenue dans le signal. L'échec concernant l'identification des invariants correspondant de manière biunivoque aux unités discrètes du code serait lié à une mauvaise méthode d'analyse et ne pourrait être que temporaire.

Mais il existe aussi des théories "Double strong relation" et des théories "Double weak relation". Il est intéressant à noter que même les partisans de ces dernières soulignent l'importance des modèles articulatoires :

> "However, even if [veridical speaker independent reconstruction of vocal tract gestures from acoustic signal] is unachievable, careful theoretical development and empirical evaluation of speech production models in speech recognition applications seems likely to lead to scientific progress and important technological advances." [Nea96, page 1717]

Finalement, une théorie "Strong acoustic and weak articulatory relation" proposée par B. Lindblom en 1990 semble remporter le plus de suffrages (K.N. Stevens, R.E. Remez, T.M. Nearey, D. O'Shaughnessy, etc.). Dans cette thèse, nous défendons cette position sur l'invariance acoustique. Nous allons donc la décrire plus en détail.

Appelée au départ la théorie de la hypo/hyper-articulation, on peut aussi la dénommer comme la théorie de l'acoustique "juste assez discriminante". Nous considérons que l'acoustique a la primauté mais que les gestes ont, eux aussi, un rôle important dans le codage de l'information selon les situations effectives de la communication. Plus précisément, pour la parole claire, notre théorie rejoint la position "Strong acoustic relation" : tous les indices distinctifs doivent être présents dans la réalisation acoustique et ceci indifféremment du locuteur et du type de phonème. Par contre, pour la parole "difficile", par exemple débit rapide en milieu bruité, nous nous attendons à ce que ces mêmes indices perdent une partie ou carrément toute leur capacité discriminante. Dans ces situations, l'auditeur est capable de récupérer le message linguistique par un mécanisme qui repose lourdement sur son expérience articulatoire de producteur de parole. Nous résumons donc ainsi notre position : il existeraient des corrélats acoustiques directs pour les traits ou les phonèmes mais ces corrélats ne sont invariants que dans un cadre restreint.

1.4 Du monde physique au monde mental. Comment modéliser ?

Au délà des divergences entre les théories sur la nature de l'invariance dans le signal de parole, il existe une source d'inspiration commune, un terrain propice à l'harmonisation des ces théories : la phénoménologie de Husserl [MP85]. La théorie quantique de l'invariance acoustique de K.N. Stevens [Ste72, Ste89] (théorie "strong acoutic relation") et le réalisme direct dans la perception de la parole de C.A. Fowler [Fow86, Fow96] (une théorie "Strong articulatory relation") suposent toutes les deux l'existence d'une structure marquée par des frontières naturelles dans l'ensemble des sons audibles et respectivement prononçables par l'être humain. Cette structure serait un reflet directe des lois physiques gouvernant l'appareillage que nous utilisons pour le transport du sens, i.e. notre système phonatoire et notre système perceptif.

> "All perceptual systems ... take advantage from the fact that certain media are lawfully and largely distinctively structured by objects and events in environment." [Fow86]

D'après Husserl, chaque objet doit se détacher en tant que phénomène, armé par une qualité de gestalt et aidé par des discontinuités limitrophes. Il est fort probable que ces principes ont présidé à la genèse des langues et ont dirigé le choix de l'inventaire réduit de sons que nous associons aujourd'hui (si bien que mal) aux phonèmes (voir la base UPSID [Val94]). Nous n'allons pas entrer dans le débat vif sur les quotas respectifs du système perceptif et du système de production de la parole dans la structuration de l'espace acoustique car il nous semble secondaire pour le traitement automatique de la parole. Par contre, pour ce but, il est de première importance de tenter la capture de ces tendances "naturelles", soit-elles d'origine perceptive ou articulatoire, afin d'affiner les corrélats acoustiques aux symboles du code d'une langue donnée.

Dans le schéma synoptique du passage du monde physique au monde mental (figure 1.1), Petitot-Concordat [Pet86] introduit un niveau qui précède le niveau sous-symbolique [Smo92, Smo86], niveau dit morphologique. Bien que d'autres niveaux peuvent bénéficier du conditionnement par la branche descendante du modèle, ce niveau est le réceptacle privilégié des qualités spécifiques du signal de parole. Il effectue la mise en forme des caractéristiques tonales et phasiques du signal selon nos attentes de ce que doit être un signal de parole, attentes issues du long apprentissage du langage. Les paramètres primaires véhiculés entre l'analyse périphérique et le traitement morphologique sont déductible pour tous les sons qu'on peut rencontrer (spectre, durée, etc.). Par contraste, les paramètres secondaires présents en sortie de ce dernier méritent le noms de "paramètres de parole" car lui étant spécifiques (formants, hauteur de voix, etc.). L'espace de l'ensemble de ces paramètres constitue **le monde sensible de la parole**, le "terrain de chasse" pour notre recherche de structures.

Comme nous avons mentionné, les systèmes automatiques actuels ne modélisent pas la "chaîne de traitement" de la parole à la manière de Petitot. Souvent, plusieurs niveaux du schéma 1.1 sont colapsés dans une "boite noire" apprise sur les données. Autrement dit, ces systèmes sont caractérisés, en général, par un **manque de représentations intermédiaires**. L'absence des ces représentations implique moins de "points de contrôle" sur la chaîne. La réduction des points de contrôle peut nuire à la performance intrinsèque (dans l'apprentissage sur les données) mais elle nuit sûrement à la modélisation, à la compréhension intime de cette chaîne de traitement. Par rapport à cette situation, l'idée centrale qui nous guide est de faire faire aux machines une **"lecture avertie"** du signal de parole. Dans cette thèse, nous recherchons donc des représentation du signal de parole, surtout au niveau morphologique. Nous proposons également des architectures de traitement, architectures que nous comparerons le plus souvent possible au schéma 1.1.

1.5 Organisation de la présentation

Je suivrai, à plusieurs reprises et à plusieurs niveaux, le même enchaînement :

1. Analyse phénoménologique
2. Modélisation
3. Validation

La première étape consiste dans la recherche de régularités et des spécificités dans des espaces de paramètres acoustiques calculés sur le signal de parole. C'est une étape d'exploration. La deuxième consiste dans la sélection des sous-ensembles ou le recodage de paramètres à la lumière des observations de l'étape précédente ou des connaissances accumulées dans le domaine. C'est une étape de recherche de jeux de paramètres. La capacité de discrimination sur une tache phonétique de ces nouvelles paramétrisations sera testé dans la troisième étape.

De manière plus pragmatique, mon travail se présente en deux parties successives. La première concerne **les voyelles orales** du français. C'est une partie de l'inventaire phonétique d'une caractérisation plus facile que les consonnes : pour la production des voyelles, nous disposons de modèles

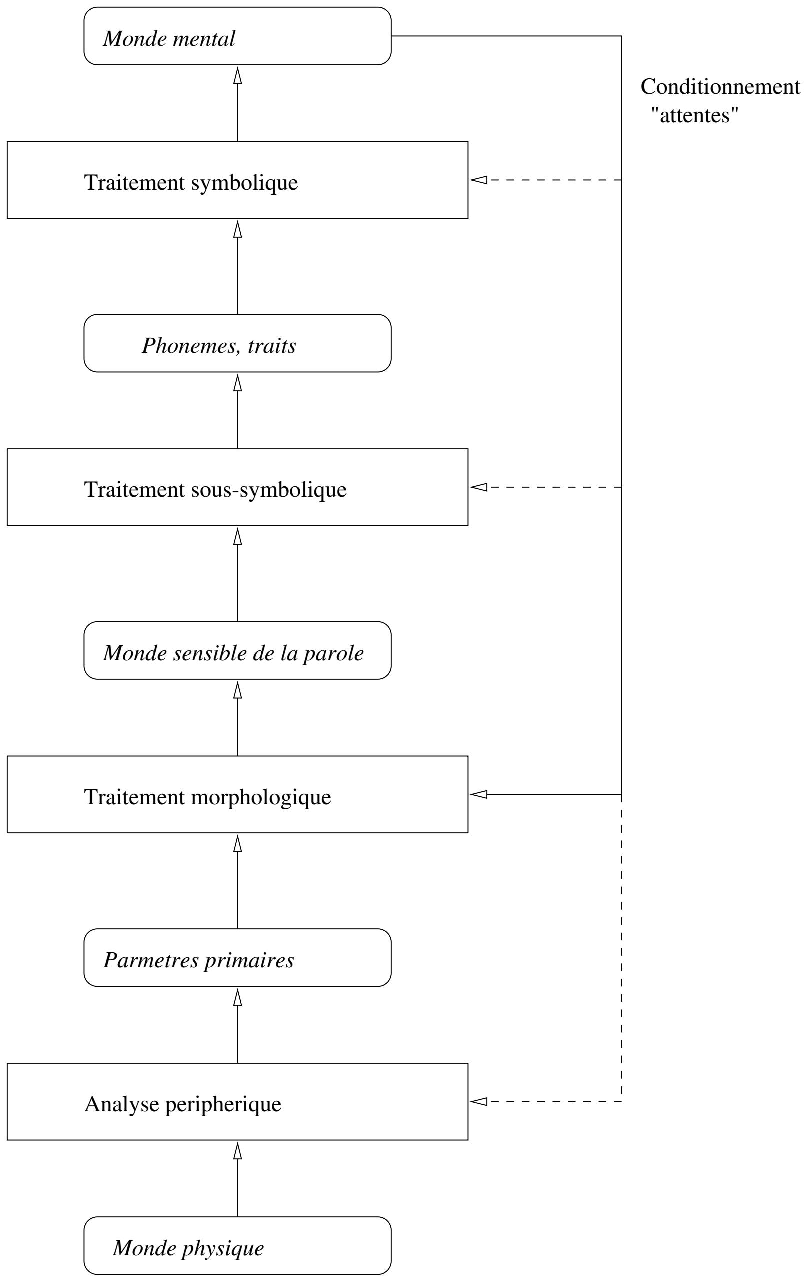


FIG. 1.1: Le paradigme morphologique : Un niveau situé entre l'analyse périphérique et le traitement sous-symbolique structure les signaux présents dans le monde physique en exploitant les régularités observées et/ou apprises.

acoustiques élaborés. Les connaissances que nous tenterons d'exploiter proviennent donc surtout de la production de la parole. Je commence par rappeler quelques résultats importants sur les **résonances du conduit vocal**. Je propose ensuite un jeu de paramètres pour les voyelles et je l'évalue par une méthode de passage interlocuteur. Je remarque certains obstacles pour la normalisation interlocuteur et je finis cette partie par une brève incursion sur l'extraction automatique des paramètres proposés.

La deuxième partie de mon travail, la plus importante, est dédiée à l'étude des **occlusives sourdes**. Les occlusives, partie du vocabulaire réputée difficile pour la reconnaissance de la parole, sont opposés aux voyelles. En effet, si la coarticulation affecte de manière négative l'identité des voyelles, au contraire, elle donne leur identité aux consonnes occlusives.

Ces phonèmes semblent donc représenter l'aspect purement dynamique de la parole. Les difficultés qui y sont liées ont concentré un effort de recherche important. Mon troisième chapitre est dédié à la littérature abondante qui traite de la nature de l'invariance pour ces phonèmes et de leur identification par les humains et par les machines. Malgré cette concentration d'efforts, il nous manque encore aujourd'hui des modèles acoustiques précis pour les occlusives. En conséquence, je me suis tourné vers la **perception de la parole** comme source de connaissances a priori pour la modélisation de ces consonnes. Je présente dans le quatrième chapitre une expérience de perception de monstres acoustiques consonne - voyelle "quasi-naturels". Cette approche est riche d'enseignements. Elle nous permet d'expliquer un nombre de résultats contradictoires rapportés dans la littérature. Elle nous permet aussi de nous poser les bonnes questions pour la suite des travaux.

Mon cinquième chapitre traite de la caractérisation des occlusives sourdes du français. Il est lui même structuré en trois sections principales :

– Une **analyse acoustique** détaillée est effectuée, analyse accompagnée d'essaies de décodage. Une importance particulière est donnée au test de différents **modèles de fusion** d'information.
– Une analyse des erreurs des classifieurs utilisés nous montre les voies pour améliorer les taux de reconnaissance. Je montre qu'une analyse morphologique est nécessaire pour expliquer les paradoxes observés.
– Une confrontation des algorithmes automatiques avec la perception des monstres acoustiques ouvre des nouvelles voies pour affiner notre modélisation.

Comme d'habitude, dans le dernier chapitre je rappelle les contribution originales de cette thèse et je passe en revue les pistes de recherche qui attendent encore leur exploration.

Chapitre 2

Les voyelles

Ce chapitre décrit notre étude des voyelles orales du français. Nous présenterons les données expérimentales en production et en perception de parole ainsi que les positions théoriques avancées dans la littérature qui plaident en faveur d'une lecture acoustique du spectre vocalique guidée par les liens articulatori-acoustiques ou par les représentations intégrées produites par le traitement perceptif.

Nous comparerons deux représentations, formants et résonances, dans des tâches d'identification de voyelles extraites de corpus multi-locuteurs. Ensuite, deux problématiques seront abordées de manière originale : l'évaluation des procédures de normalisation vocalique et la caractérisation des stratégies individuelles d'opposition d'aperture en français.

2.1 Production des voyelles. Résonances du conduit vocal

2.1.1 Des commandes musculaires au signal acoustique

Le signal de parole est le résultat de la propagation d'une onde acoustique dans un tuyau de forme variable. Ce tuyau est mis en forme par un ensemble de muscles agissant sur des structures rigides, semi-rigides ou des hydrostats comme la langue ou les lèvres. De même, les sources d'excitation sont produites par des compressions exercées sur les poumons, les affecteurs laryngés, les parois du conduit vocal dont la tension est en permanence régulée et contrôlée par des structures musculaires spécifiques. Le contrôle complexe de cette activité musculaire permet au locuteur de mettre en forme le contenu spectral et temporel du signal de parole. L'auditeur récupère la trace audible de cette suite de gestes - complétée occasionnellement par leur trace visible.

Comme nous l'avons mentionné, ce qui fait toute la difficulté du traitement automatique de la parole est la phénoménale variabilité des signaux que l'on peut enregistrer dans différents points dans de cette chaîne de communication (signaux articulatoires, acoustiques, activité neuronale). Cette variabilité puisse ses origines dans différentes sources indépendantes : les caractéristiques anatomiques, l'influence du contexte phonétique, les stratégies articulatoires propres à chaque locuteur, les états mentaux, les origines socioculturelles et géolinguistiques, etc.

Chaque auditeur sait, d'une part, s'affranchir de ces sources de variabilité afin d'associer au signal un contenu conventionnel (au sens du code linguistique). D'autre part, chaque auditeur sait identifier ces sources de variabilité afin d'acquérir l'habilité motrice nécessaire à une communication riche et robuste. Au contraire, les machines ne disposent pas aujourd'hui de moyens efficaces pour gérer cette variabilité.

Quelles sont donc les difficultés qui limitent la modélisation de cette variabilité dans une forme appropriée aux machines ? La principale raison provient d'un nombre de degré de liberté en excès dans le passage des commandes musculaires du locuteur au percept de l'auditeur. Nous essayons d'illustrer ce passage sous la forme d'une pyramide sur la figure 2.1. Cette non-unicité se manifeste aussi entre niveaux intermédiaires : le passage d'un niveau de caractérisation à un autre est une relation de type "plusieurs-à-un". Par exemple :

- Une même configuration articulatoire peut être obtenue par de multiples combinaisons d'activités musculaires [OFFM91].
- Il est mathématiquement établi qu'on ne peut pas récupérer de manière unique la forme d'un résonateur d'après son "empreinte" sonore [Cip92]. Les humains tire profit du fait que plusieurs formes de conduit peuvent générer un même son (voir figure 2.2)pour l'articulation compensatoire [GLL81, SPO95].
- Plusieurs sons différents donnent un même percept. Seulement certaines propriétés du signal sont "phonétiquement pertinentes".

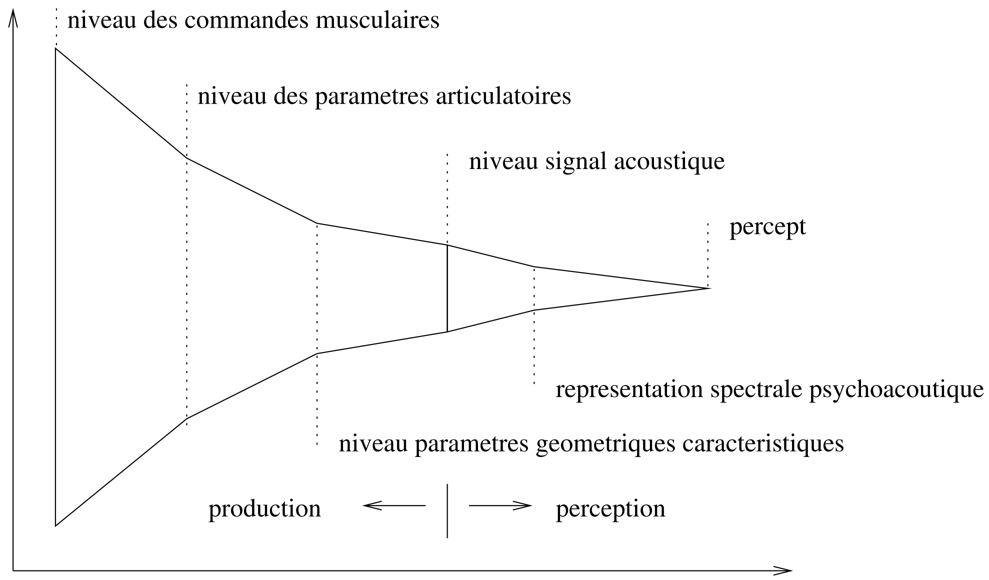


FIG. 2.1: Niveaux de représentation d'un son donné de sa production à sa perception

Dans une situation de communication courante, toutes les représentations sont accessibles au locuteur et lui donnent des retours actifs sur sa performance langagière. Les humains disposent ainsi de **cartes sensori-motrices** pour faire le lien entre gestes et percept lors de la production comme de la perception de la parole.

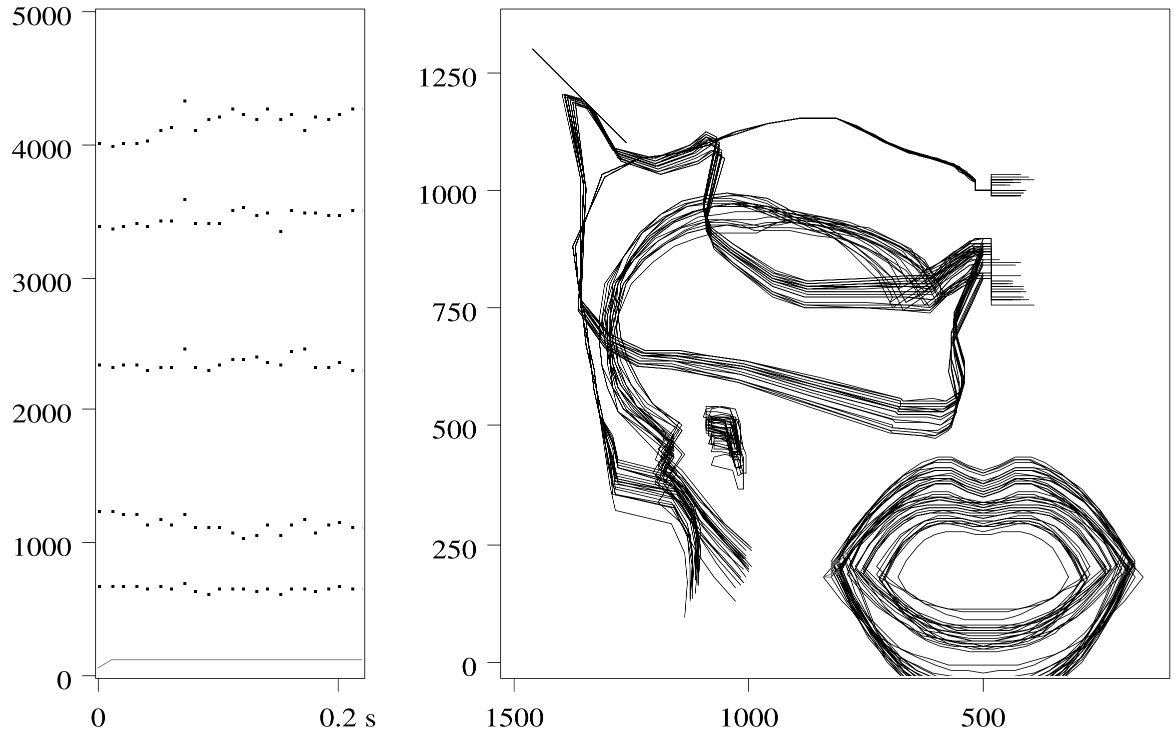


FIG. 2.2: Fibre articulatoire pour la voyelle /a/. Valeurs de formants (à gauche) et configurations du conduit (à droite). Simulation par Bergame [BBB+96], version du modèle articulatoire de Maeda développée à l'I.C.P.

Le niveau du signal acoustique est primordial et, par conséquent, les cartes qui le relie à d'autres niveaux, en aval ou en amont, le sont aussi. Comme énoncé par Morgan et Demuth dans leur récent livre consacré à l'acquisition du langage : "At the root [...] it is the signal that conveys words, phrases, sentences, indeed grammars to each new generation." [MD96, p.20]. Et, parmi toutes les représentations phonétiques, les formants ont une place privilégiée. Les cartes formantiques sont largement utilisées pour :

caractériser la production La représentation formantique est centrale pour prédire les espaces vocaliques [BSV94], pour caractériser les réalisations vocaliques ou consonantiques en fonction de divers paramètres contextuels [SMM91] ou prosodiques [vB95],

synthétiser On utilise les formants pour contrôler la production de sons artificiels et établir des règles de génération [Kla87]

inverser Les techniques d'inversion articulatori-acoustique prennent de manière privilégiée les formants comme caractérisation du spectre [ACMT78, BPB92, Soq95b]

étudier la perception la production de sons synthétiques en vue de l'étude des nos représentations perceptives a été principalement opérée par un contrôle effectif de la structure fréquentielle par les positions des formants.

Les résultats présentés dans la suite de cette section visent à poser les bases expérimentales de la restructuration de l'espace formantique que nous allons proposer. Cette restructuration vise à augmenter les propriétés de caractérisation et de discrimination de l'espace formantique original, à

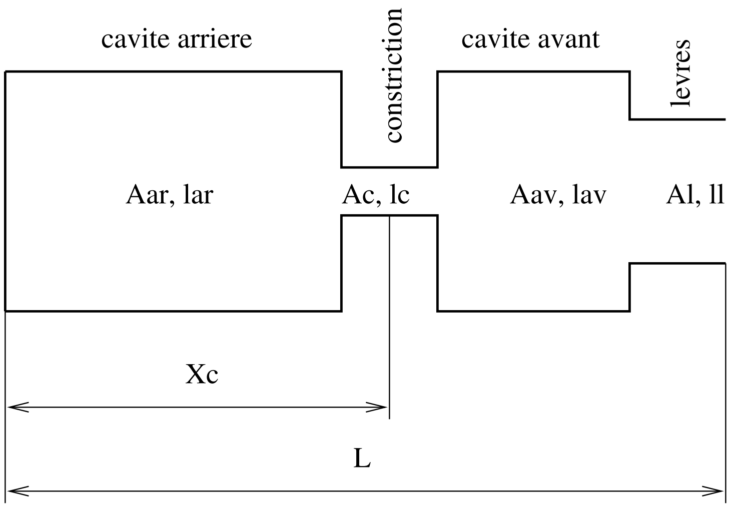


FIG. 2.3: Schéma du modèle à quatre tubes. Valeurs des paramètres : Aav=Aar=8cm^2, ll=1cm, lc=5cm, Al $\approx$ 0.16 - 4 cm^2, Ac $\approx$ 0.65 cm^2, L=16cm

mieux poser le problème de l'inversion articulatori-acoustique et à éclairer les données dont nous disposons sur la perception des voyelles.

2.1.2 Points focaux. Affiliation des formants

Le modèle à quatre tubes pour l'étude des voyelles a été proposé par Gunnar Fant en 1960 [Fan60]. Ce modèle simple (voir figure 2.3) a le mérite de mettre en évidence que, réduite à son expression la plus simple, la production d'une voyelle suppose de :

- positionner **une constriction** qui sépare le conduit en deux cavités (cavité arrière ou pharyngale et cavité avant ou buccale).
- contrôler le jeu labial (fermeture, ouverture, étirement ou protrusion).

Les paramètres de contrôle importants de ce modèle sont la position de la constriction (Xc) et l'aire au lèvres (Al). Xc et Al sont des bons candidats comme propriétés "phonétiques" du conduit car, en variant ces deux paramètres, on peut simuler l'ensemble des voyelles orales avec un modèle élémentaire.

Les propriétés de ce modèle ont été étudiées en détail dans plusieurs travaux [BBPA90, BBPA88]. Nous les présentons ici dans les grandes lignes en insistant sur les plus importants parmi les concepts auxquels ces études ont abouti.

2.1.2.1 Nomogrammes et points focaux

Nous nous intéressons à la variation de la fonction de transfert du conduit en fonction de la position de la constriction. On peut en obtenir une représentation graphique montrant les fréquences des pôles de cette fonction de transfert en fonction du Xc (représentation appelée "nomogramme de Fant").

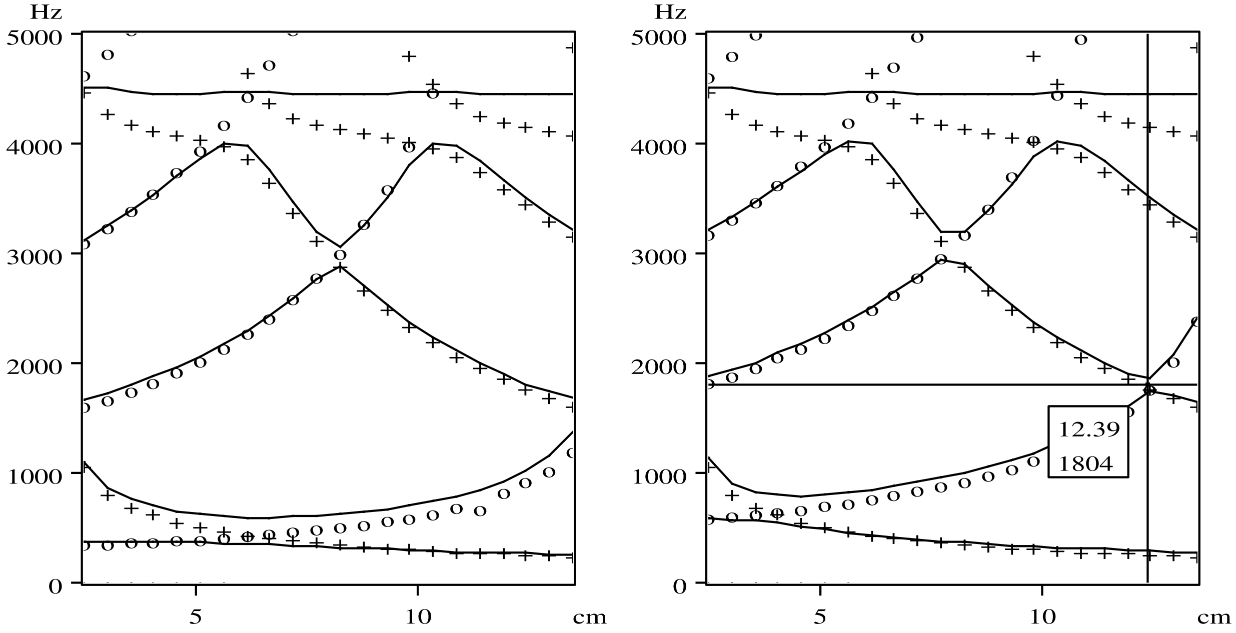


FIG. 2.4: Nomogrammes de Fant en fonction de Xc (de 1 à 15cm) pour faible aire aux lèvres (à gauche, Al=0.16cm^2) et aire moyen (à droite, Al=4cm^2). Trait plein : résonances du modèle complet, + : résonances cavité arrière, o : résonances de la cavité avant.

Nous observons sur cette représentation (voir figure 2.4) que les fréquences de résonance de l'ensemble du conduit (notés F1, F2, F3, etc.) balayent chacun un domaine de fréquence. Il en résulte ainsi des trajectoires qui se rencontrent. Ces points de convergence des trajectoires formantiques s'appellent des **point focaux** [BBPA88]. Un tel point est marqué sur le graphique de gauche sur la figure 2.4. Des simulations avec des modèles de fonction d'aire plus élaborés [Fan92] montrent un comportement identique. Il est important aussi de noter que ces points focaux, obtenu ici par simulation sur un modèle, sont présents également dans la parole naturelle dans les transitions voyelle-voyelle (voir figure 2.6).

Ce comportement des trajectoires s'explique si nous étudions séparément les fréquences de résonance des deux parties du modèle. La partie arrière (cavité pharyngale plus constriction) possède des résonances dont la fréquence diminue[1] en augmentant Xc. En même temps, la partie avant (cavité buccale et lèvres) présente une famille de résonances[2] dont les fréquences augmentent avec Xc (voir figure 2.5). Dans notre étude, nous négligeons volontairement les autres modes propres (par exemple, les résonances de la constriction ou du tube labial) en raison de leur fréquence élevée.

Nous précisons ici nos notations. Nous notons les résonances de la cavité arrière avec des indices impairs : R1, R3, etc. et celles de la cavité avant avec des indices pairs : R2, R4, etc. Les fréquences de résonance de l'ensemble du conduit sont notées traditionnellement F1, F2, F3, etc.

[1] la fréquence la plus basse qui est une résonance Helmholtz fait exception pour Xc extrême (voir figure 2.5)

[2] à la différence de la cavité arrière, la nature et donc la fréquence des résonances de la cavité avant dépendent de Al. Les résonances passent d'une famille multiples impairs de quart d'onde (pour Al maximum) à une famille Helmholtz plus multiples de la demi-onde (pour Al faible)

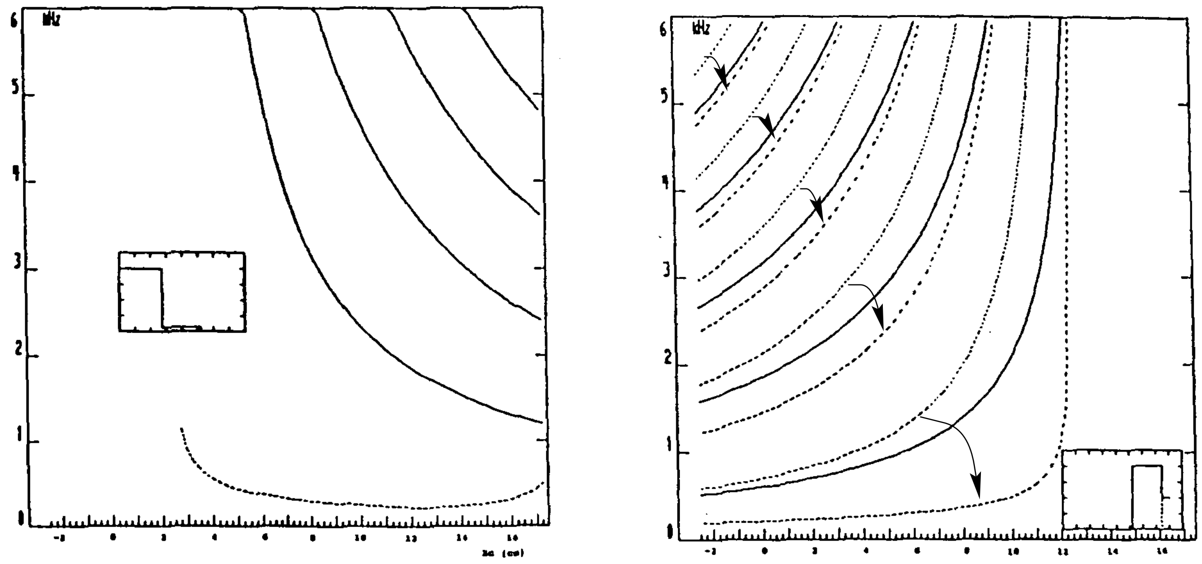


FIG. 2.5: Nomogrammes des cavités séparées. Résonances de la cavité arrière (à gauche) et famille de résonances de la cavité avant (à droite, pour Al=0.16cm2, 4cm2 et 8cm2) (Repris de [BBPA88]).

2.1.2.2 Affiliation des formants. Nomogrammes des cavités découplées

Plus l'aire de la constriction (Ac) est faible, plus le couplage plus entre les deux cavités est faible et, donc, plus la superposition des deux familles (R1, R3, ... et R2, R4, ...) approche bien les fréquence de résonance de l'ensemble (F1, F2, F3 ...). Nous pouvons maintenant définir **l'affiliation majoritaire** des formants. Pour les sons correspondant à Xc > 7cm, F2, la deuxième résonance de l'ensemble est due soit à une résonance de la cavité arrière (R3) soit à une résonance de la cavité avant (R2). Une trajectoire formantique de l'ensemble de tubes peut donc changer d'affiliation majoritaire au cours de son évolution temporelle (dans une transition voyelle-voyelle, par exemple). Ce changement d'affiliation sont marqués par les points focaux (cf. figure 2.4). Plus l'aire de la constriction est importante, plus le changement est masqué (la convergence des formants n'est plus complète, l'affiliation est moins nette, etc.). Néanmoins, la simple inspection des transitions entre voyelles dans la parole courante montre des évidents ruptures de dérivée des trajectoires formantiques (voir figure 2.6).

Pour des fonctions d'aire plus réalistes, par exemple celles produites par le modèle de Maeda [Mae79, Mae88], la définition du Xc n'est plus évidente. Une méthode de calcul a été mise au point à l'Institut de la Communication Parlée [Bai95, Rou92] : les nomogrammes des cavités découplées. Cette méthode calcule l'erreur de correspondance entre les résonances du conduit et les résonances des deux parties du conduit pour toutes les positions possibles d'une hypothétique constriction. Le minimum de cette erreur donne la partition optimale du conduit en cavité avant et arrière pour une fonction d'aire donnée. On peut définir ainsi un Xc "acoustique" pour une fonction d'aire beaucoup plus compliquée que le modèle à quatre tubes. Cette méthode permet d'établir l'affiliation majoritaire des formants, c'est à dire mettre en correspondance les Fi et les Ri. Pour cela, il suffit de regarder

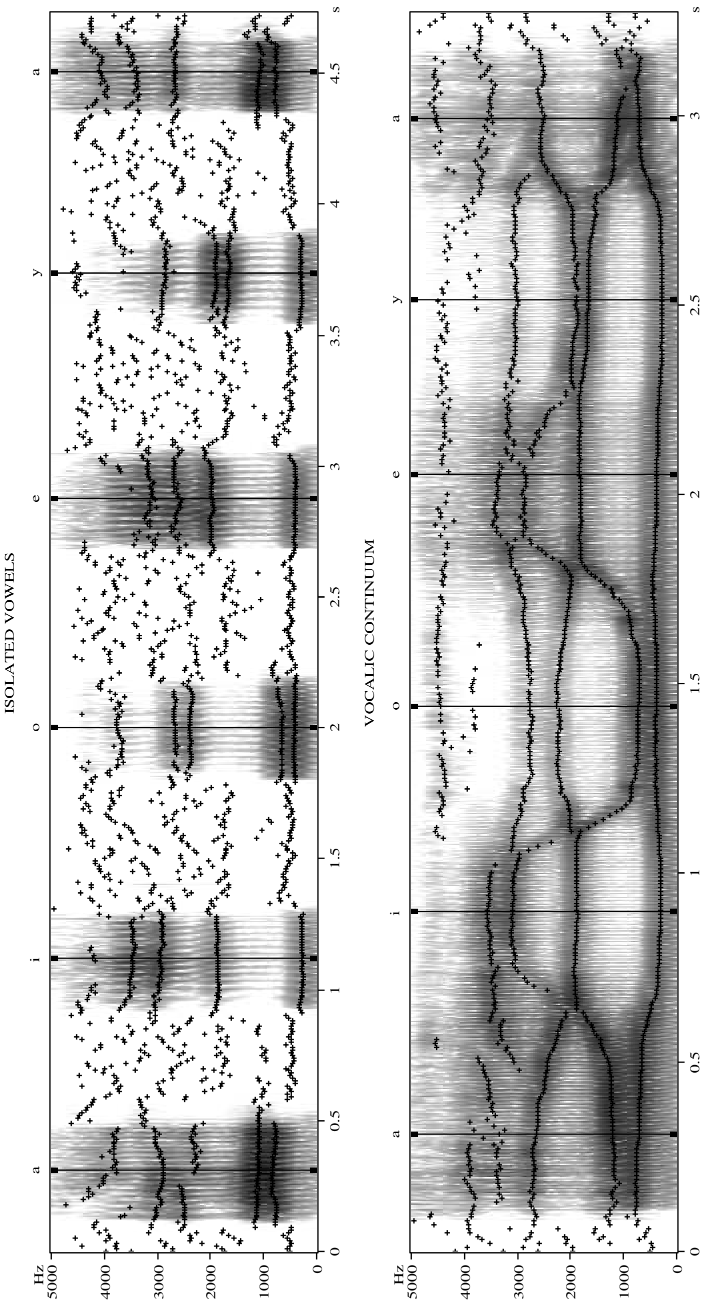


FIG. 2.6: Spectrogrammes des voyelles du français prononcées isolées (en haut) et en continu (en bas). On remarque les croisements des trajectoires formantiques.

la position relative des fréquences de résonances des deux cavités découplées pour le Xc trouvé. Par exemple, dans la figure 2.7, nous illustrons cette méthode sur la fonction d'aire d'une voyelle /i/.[3] Nous constatons que R2 est proche F2 pour toutes les voyelles orales du français sauf /i/ (où R2 est proche de F4) et /e/ (où R2 est proche de F3) [Rou92, Bea93].

2.2 Résonances en situation

Nous sommes maintenant en position de montrer l'intérêt du paradigme des résonances contre celui classique des formants. Si Xc et Al sont des bons paramètres candidats alors une paire de fréquences permettant de remonter à ces grandeurs le sera également. Traditionnellement[4] on considère F1 et F2. F1 indique l'aperture de la voyelle et F2 indique la position de la langue sur l'axe avant-arrière. L'inconvénient apparait pour les voyelle articulées à l'avant du conduit, comme /i/, où F2 est affilié comme F1 à la cavité arrière. Dans ce cas, F1 et F2 portent de l'information redondante.

Au contraire, la paire de fréquences R1 et R2 ne présente pas cet inconvénient. Nous disons que ce jeu de paramètres **"linéarise"** le passage entre l'acoustique et l'articulatoire par rapport au jeu F1 et F2. En effet, l'information porté par F2 "sature"[5] dès qu'on dépasse un point focal vers les voyelles avant. Il s'en suit que R1 et R2 peuvent constituer une meilleure base pour contraindre l'inversion acoustico-articulatoire [BPB92]. L'inconvénient est que R1 et R2 peuvent correspondre à F1 et F2 ou à F1 et F3 ou, parfois, à F1 et F4 donc ils nécessitent une étape supplémentaire de calcul par rapport aux F1 et F2.

Dans la suite de cette section, je me propose de présenter brièvement quelques résultats rapportés dans la littérature sur les avantages de la caractérisation des voyelles en termes résonances. Dans les autres sections de ce chapitre j'ajouterai les arguments en faveur des résonances issus de mes propres travaux.

2.2.1 Reproduction des continuum vocaliques

L'étiquetage des formants en terme d'affiliation permet d'expliquer des résultats des tests de reproduction des sons complexes par des locuteurs.

Un test de reproduction de cibles acoustiques situées sur un continuum vocalique a été conduit pour l'anglais en 1985 par B. Repp et collaborateurs pour étudier les effets de la perception catégorielle [RW85]. Dans ce tests, les stimuli synthétiques intermédiaires entre deux voyelles sont obtenus par simple interpolation linéaire des fréquences des formants aux deux extrémités du continuum. Les stimuli reproduits par les locuteurs ne suivent pas cette évolution linéaire. Ils montrent plutôt une évolution par paliers séparés par zones d'évolution rapide ce qui a été interprété par les auteurs comme une frontière de catégorie.

[3] Prototype de voyelle /i/ (simulée par le du modèle de Maeda)

[4] Le pouvoir de discrimination de F1 versus F2 compte parmi les connaissances les plus anciennes de la phonétique

[5] Une variation importante de Al, par exemple, ne modifie plus F2

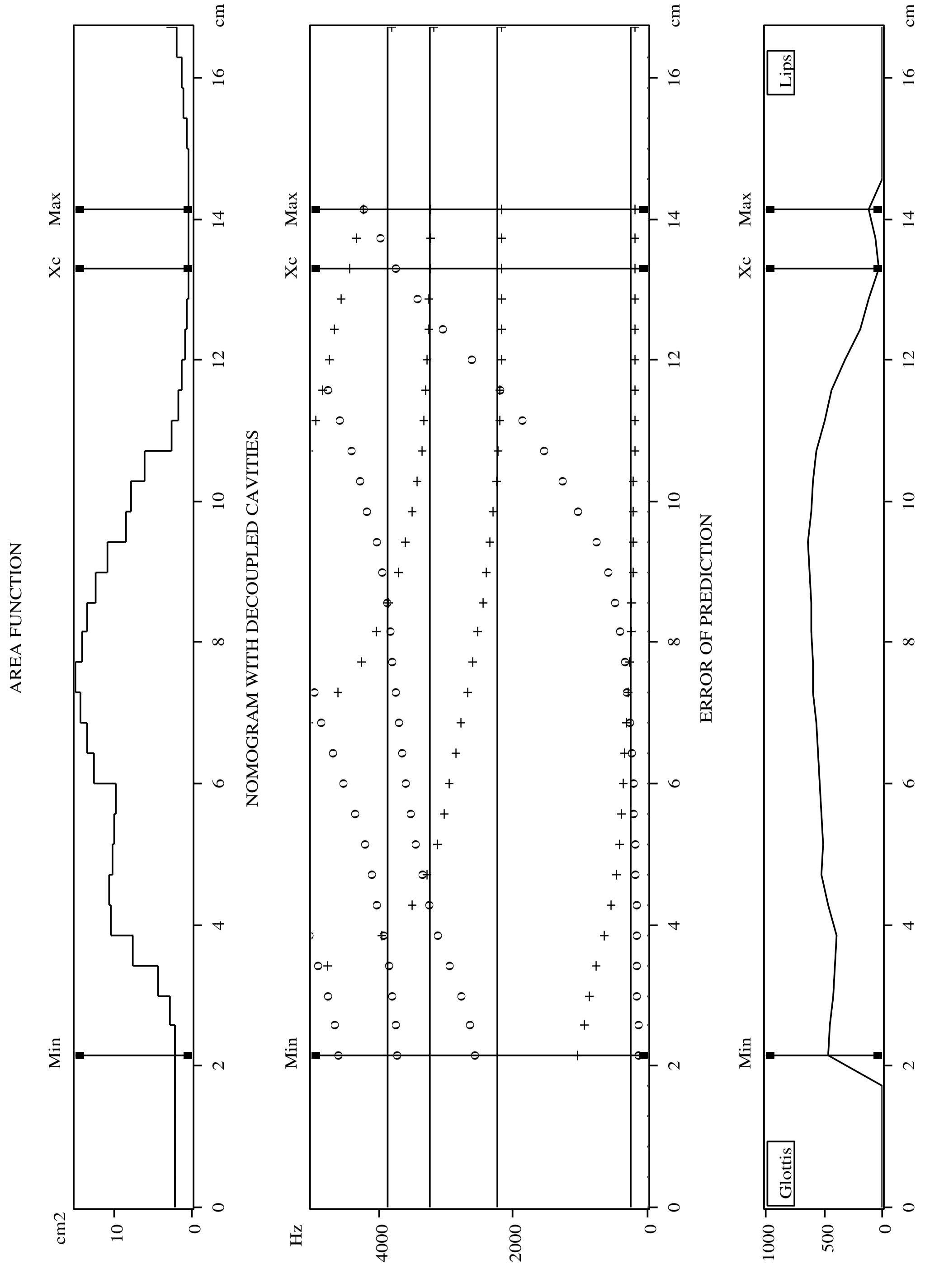


FIG. 2.7: Nomogrammes à cavité découplées pour la fonction d'aire du /i/. Le minimum de l'erreur de prédiction est marqué par Xc. On remarque que, pour cette position de la constriction, les premiers trois formants sont affiliés à la cavité arrière

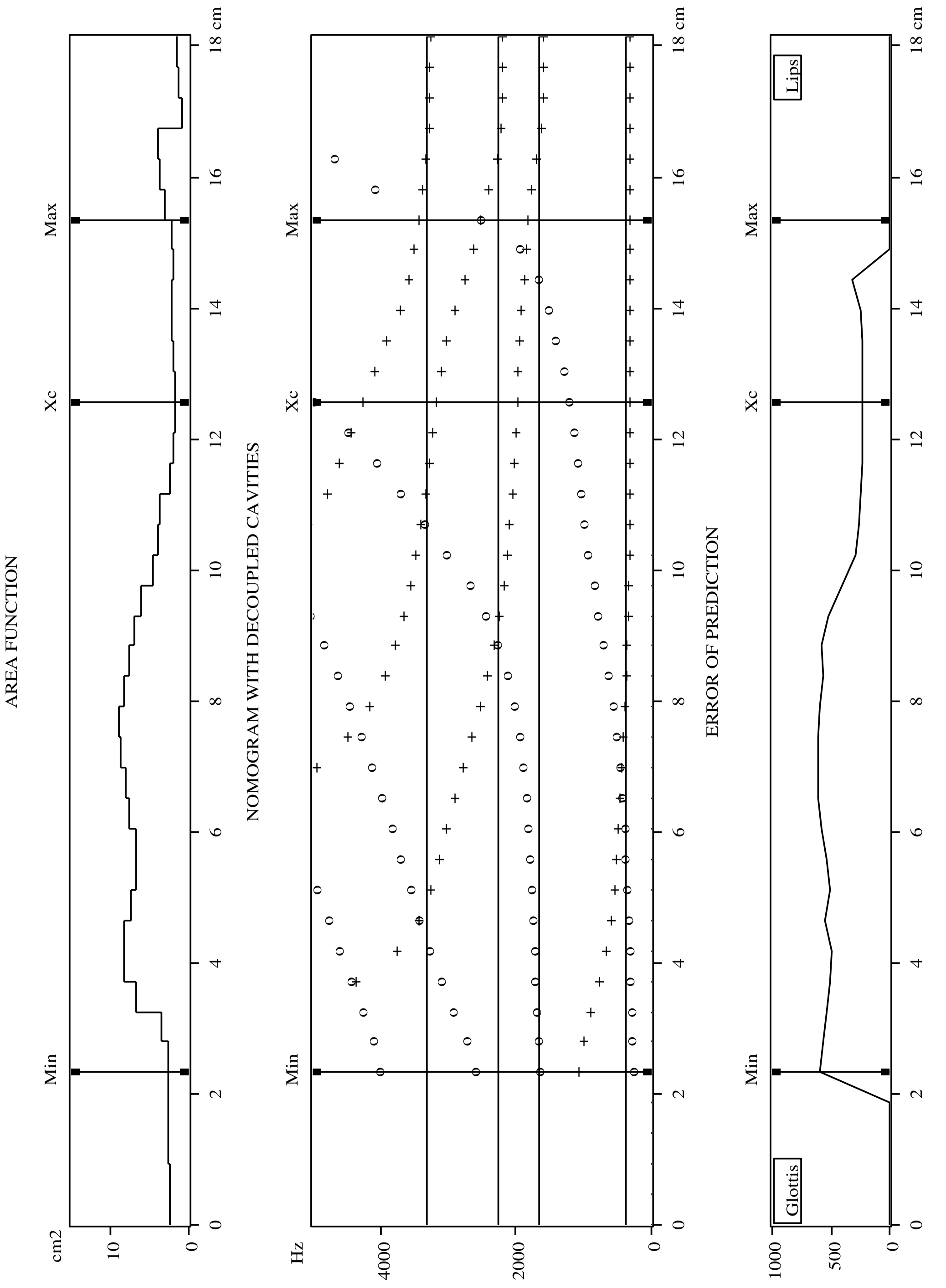


FIG. 2.8: Nomogrammes à cavité découplées pour la fonction d'aire du /ø/. F2 est affilié ici à la cavité avant, comme pour la majorité des voyelles orales du français

Ce test a été reproduit à l'I.C.P. par 12 locuteurs français [Bai93]. Nous présentons ici comparativement les résultats des reproductions pour la transition de /a/ à /i/ pour deux types de construction des signaux intermédiaires (voir figure 2.9). Nous appelons la première "interpolation des formants" car elle relie F2 du /a/ à F2 du /i/ et la deuxième "interpolation des résonances" car elle relie R2 du /a/ à R2 du /i/, c'est à dire F2 du /a/ à F4 du /i/. Nous observons que les locuteurs arrivent plus facilement à suivre les stimuli à reproduire pour l'interpolation des résonances. Nous observons également que, sur le continuum de l'interpolation des formants, les réponses suggèrent un point focal entre /a/ et /i/ malgré le fait que les stimuli à reproduire n'ont aucun croisement entre F2 et F3.

2.2.2 Rapprochement avec F'2

Une autre tentative de rendre compte des formants supérieurs avec un modèle simple est bien connue. Il s'agit des tests perceptifs du deuxième formant effectif (F'2). Le jeu de paramètres F1 et F'2 présente certaines caractéristiques similaires à celles de notre jeu R1 et R2, similitudes qu'il convient de discuter.

Les résultats sur F'2 proviennent d'une série des travaux [FR63, CGF70, CFG75, BF78, CFdSLT66], série initiée par Fant et par Chistovich. Ces tests établissent une correspondance entre un spectre vocalique à plusieurs formants et un spectre synthétique à deux formants, correspondance basée sur une distance perceptive. Ainsi, F'2 est une valeur ajustée de manière interactive par les auditeurs afin qu'un son à deux formants soit pour eux le plus proche possible d'un son de référence à plusieurs formants.

Afin de soutenir notre propos, nous groupons les résultats du F'2 dans deux catégories.[6]

1. Concernant d'abord le F'2 mesuré par rapport aux voyelles naturelles (voir figure 2.10), nous observons que, à l'exception des voyelles articulés à l'avant, F'2 prend des valeurs proches du F2, comportement semblable à celui R2, la première résonance de la cavité avant. Je retiens donc cette coïncidence remarquable : F'2 s'éloigne de F2 là où R2 est différent lui aussi de F2. Je rajoute l'observation empirique que les voyelles orales placent toujours un formant à proximité de 2200 Hz.[7]. La même coïncidence peut être lu aussi ainsi : F'2 est différent de F2 seulement là où F'2 dépasse cette fréquence de 2200 Hz, notamment pour les voyelles /i e/.

2. Concernant F'2 mesuré par rapport à une série de voyelles dans un continuum synthétique avec F1, F2 et F4 fixé et F3 variable, nous remarquons que la courbe de variation du F'2 ne suit pas parfaitement la courbe du F3 (voir figure 2.11). Elle reste proche de F2 et, respectivement, F4 et présente une zone intermédiaire de variation rapide. Ce comportement non-linéaire du F'2 peut être la manifestation de ce que nous avons appelé en introduction "une frontière naturelle".[8] Ma supposition est que les auditeurs ont "entendu" un changement d'affiliation.

[6]Nous n'abordons pas le débat sur les mécanismes qui expliquent ces résultats (intégration large bande, par exemple).

[7]Cette observation m'a inspiré pour notre algorithme d'étiquetage des formants en termes de résonances.

[8]en faveur de l'explication complémentaire et concurrente par l'intégration large bande, il a été démontré [Sch87, page 323] que ce comportement n'est pas seulement l'effet d'une frontière phonétique

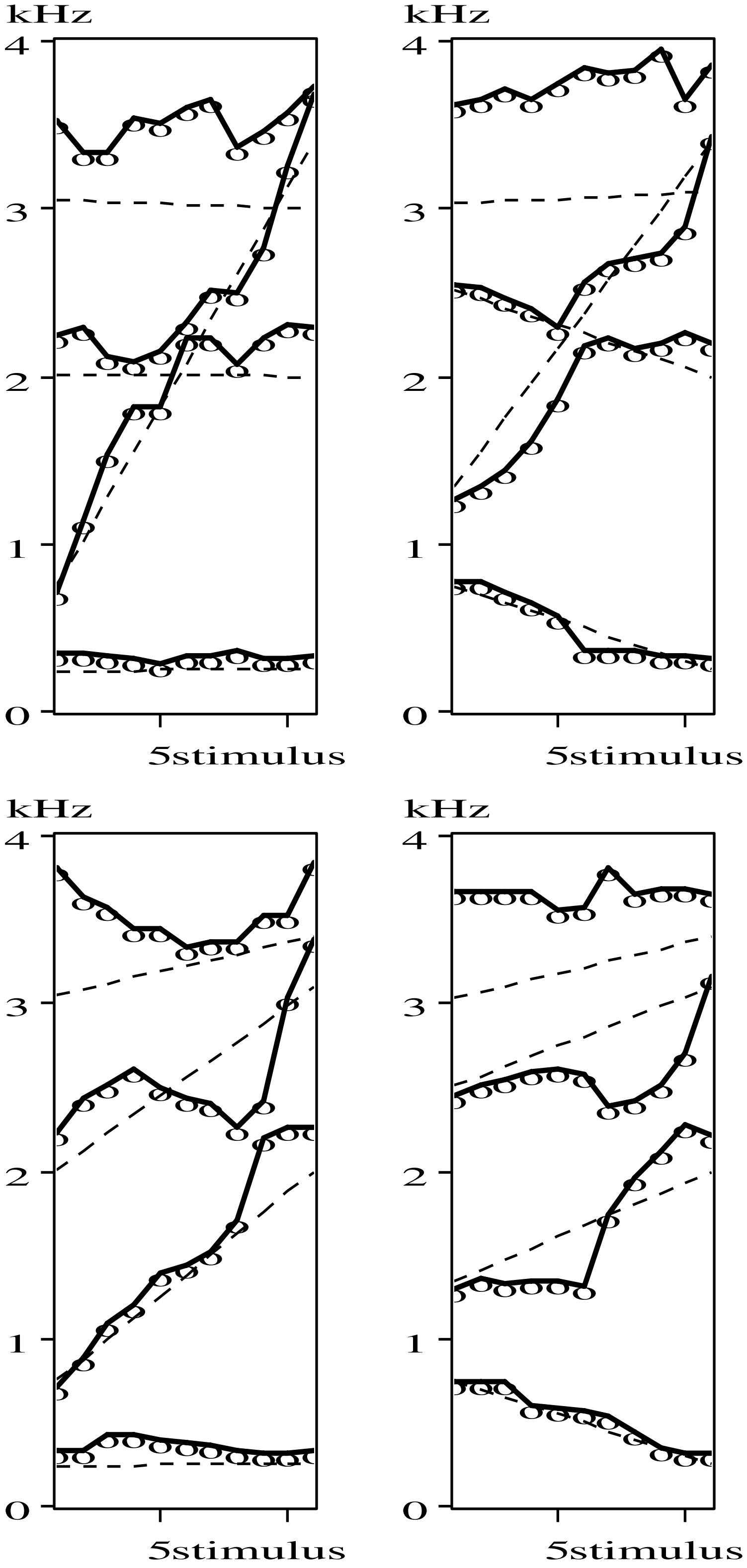


FIG. 2.9: Résultats d'imitation des continuum vocaliques /ui/ (à gauche) et /ai/ (à droite) pour le locuteur LJB. L'interpolation entre les valeurs extrêmes est faite entre formants de même rang (en bas) ou entre résonances du même rang (en haut). La superposition des formants des réalisations du locuteur au long du continuum montre un passage nécessaire par une focalisation.

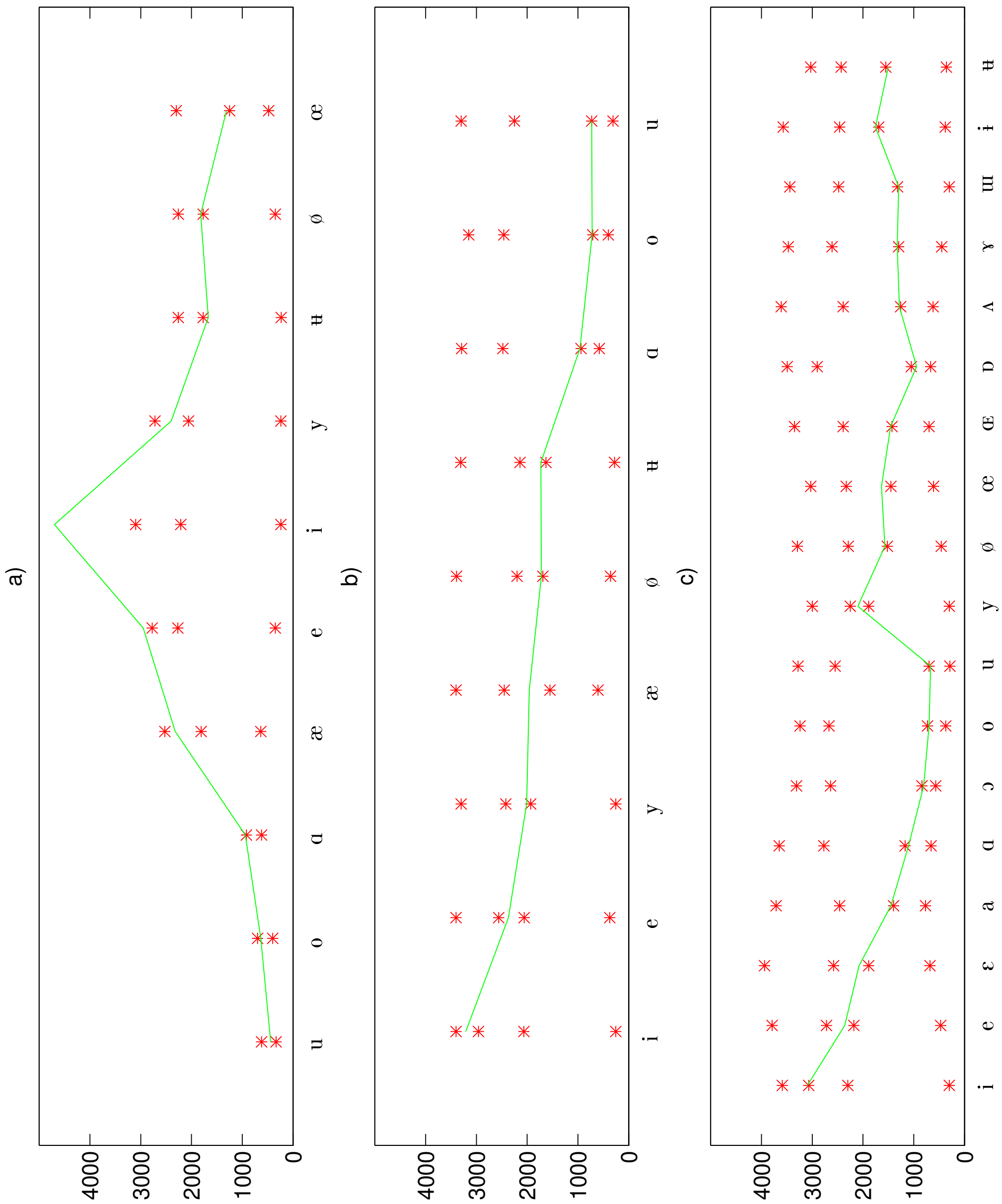


FIG. 2.10: Position de F'2 (trait plein) par rapport aux formants (*) dans tests rapportés dans la littérature : a) Fant et Risberg [FR63], b) Carlson et collaborateurs [CGF70], c) Bladon et Fant [BF78]. Il faut remarquer les situations où le F'2 n'est pas proche de F2 : [i e y æ] pour a), [i e æ] pour b) et [i y] pour c)

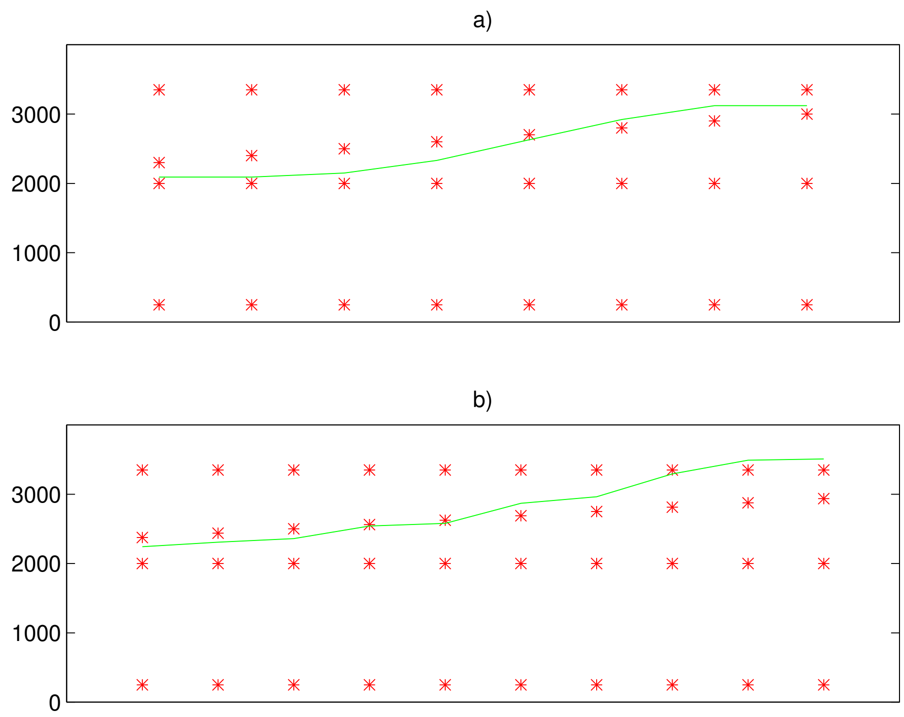


FIG. 2.11: Position de F'2 dans tests sur un continuum vocalique synthétique : a) Schwartz et Escudier [Sch87] b) Carlson et collaborateurs [CGF70]

2.2.3 Conditions particulières de locution

Un autre argument pour la prise en compte de R2 en lieu et place de F2 pour un modèle simplifié des voyelles nous provient des enregistrement de la parole produite sans vibration des cordes vocales (parole chuchotée, parole fricative).[9] Ces conditions d'excitation atypiques conduisent à un signal de nature forte différente mais qui garde une bonne intelligibilité pour les auditeurs humains.

La figure 2.12 montre les spectrogrammes d'un même mot prononcé normalement et avec un excitation fricative. Nous remarquons que le maxima spectral le plus proéminent gardé dans la deuxième condition n'est ni F2 ni F3 mais une combinaison de F2 et F3 évoluant selon la voyelle. Ceci s'explique par la nouvelle position de la source d'excitation, le signal répercute seulement la fonction de transfert de la cavité avant. Nous retenons que, en général, un signal qui garde la structure de résonance de la cavité avant conserve une bonne partie de l'intelligibilité. Cette observation a conduit Khun [Kuh75] et ensuite Hermansky [HB89] à formuler l'hypothèse que **la cavité avant** est le principal porteur du message linguistique. Ils postulent l'extraction des résonances de la cavité avant par le système perceptif. Ils vont jusqu'à identifier dans un seul concept ce que nous avons appelé R2 et le deuxième formant effectif F'2.

> "The size and shape of vocal tract's front cavity is the primary carrier of linguistic information. The back cavity geometry is its causal consequence and contribute mainly

[9] Dans la parole chuchotée, les cordes vocales sont en adduction de manière à produire une source fricative sans vibrer. Dans la parole fricative, moins naturelle, les cordes vocales sont suffisamment ouvertes et l'aire de la constriction linguale suffisamment réduite pour générer un bruit au niveau de cette constriction.

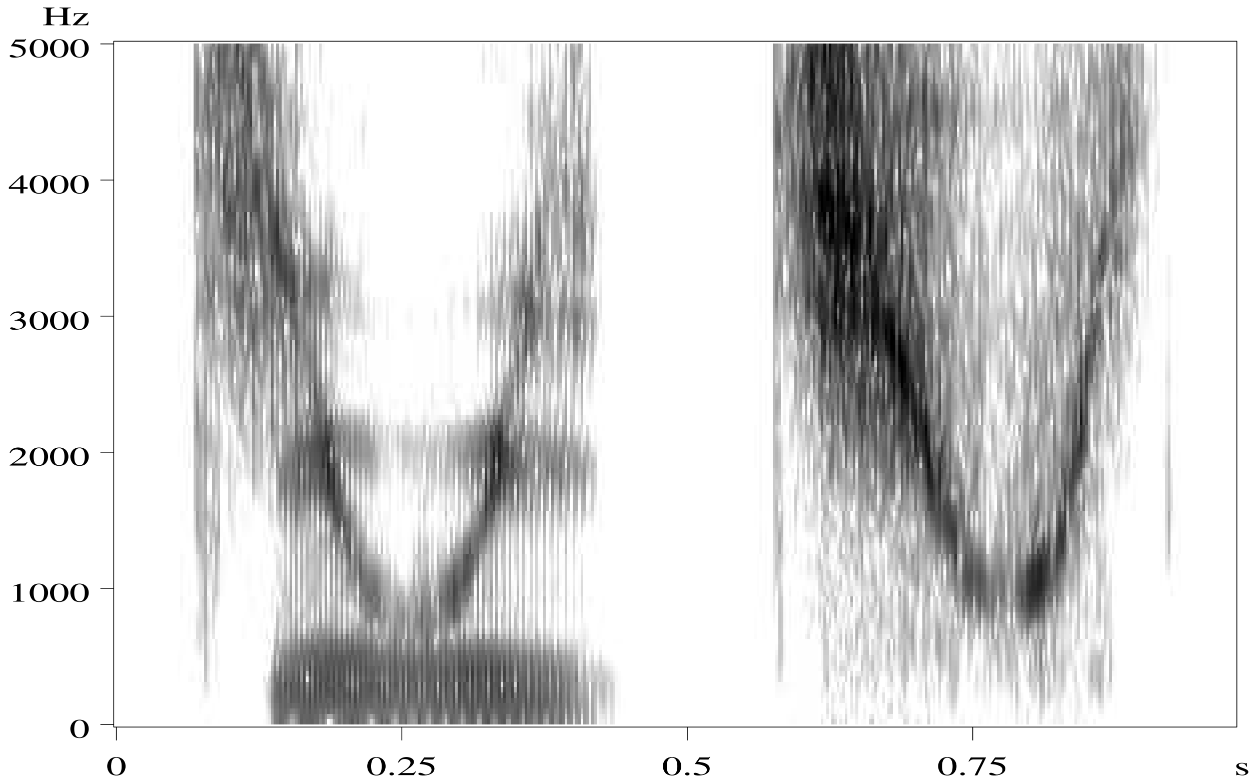


FIG. 2.12: Un mot (/tiwit/) prononcé dans deux conditions différentes : excitation voisée (à gauche) et excitation fricative à la constriction (à droite). On observe la disparition des formants qui varient le moins dans la transition. R2 est présent dans le deuxième signal.

> speaker-dependent information. [...] The human auditory system simplifies formant representation by extracting the effective second formant (F'2). F'2 carries in its frequency and bandwidth information about the front cavity [...]." [HB89, page 480]

Même si F'2 n'est qu'un artefact du à une expérimentation psycho-acoustique d'ajustement, nous retenons comme **conclusion** qu'un système automatique basé sur cette représentation devrait voir ses performances se dégrader progressivement en cas de changement de mode de phonation.

2.3 Première pause

Les voyelles sont un segment du vocabulaire qui a été longuement étudié, aussi bien sous l'angle de la perception que sous l'angle de la production. Comme nous l'avons vu, ces travaux suggèrent des possibles avantages de la lecture "articulatoire" des représentations acoustiques de la parole. L'association des étiquettes aux formants, étiquettes correspondant à leur affiliation majoritaire, "prépare" le signal pour l'identification phonétique. La suite de notre travail s'articule autour des trois axes suivants :

1. L'exploitation de ces résultats de production et perception. Nous désirons répondre à la question *"Trouve-t-on un codage des voyelles qui reflète les concepts présentés auparavant, par exemple, les points focaux ?"*
2. L'évaluation, la possibilité de chiffrer les avantages d'une telle représentation. *"Une représentation avertie, apporte-t-elle une amélioration significative sur une tâche phonétique ?"*
3. Le retour sur l'investissement : *"Quelles sont les incidences du succès ou insuccès d'une telle représentation sur la modélisation initiale ?"*

2.4 Le jeu de paramètres R1, R2, R3. Premières évaluations

Fidèles à notre approche basé principalement sur l'acoustique, nous n'allons pas chercher des paramètres géométriques. Nous allons chercher plutôt une interprétation des paramètres acoustiques éprouvés selon nos connaissances sur la production de la parole.

2.4.1 Corpus. Base de données

Nous avons considéré une base largement utilisé de signaux de parole numériques : BDSONS. Présentée sous forme de 7 CD-ROM, elle contiennent des signaux facilement exploitables. Nous avons sélectionné un sous-ensemble présentant une bonne fréquence d'occurrence pour tous les voyelles orales : entre 30 répétitions pour les voyelles extrêmes et 9 pour les voyelles centrales. Il s'agit des corpus nommés RAM01 et RAM10. Ils sont constitués chacun de 108 mots réels monosyllabiques isolés (CV et CVC). Les contextes consonantiques sont variés (fricatives, occlusives et liquides, par ordre de fréquence) et ne sont pas rigoureusement équilibrés d'une voyelle à l'autre. Ces corpus ont été prononcés par 6 locuteurs et 6 locutrices.

À partir de ces signaux, nous avons construit une banque de valeurs formantiques par expertise manuelle. Les étapes de l'extractions des valeurs formantiques étaient les suivantes :

1. Dans un premier passage, un expert humain a placé des marqueurs temporels au centre de la voyelle. La position de ce marqueur a été choisie par référence à l'oscilogramme et au spectrogramme du signal. Le code phonétique de la voyelle est extrait d'une transcription normative donnée sur CD-ROM. Pour certains réalisation vocaliques, cette transcription normative peut être inappropriée.

2. Deux méthodes automatiques d'extraction des formants (maxima du spectre FFT et racines du polynôme LPC) ont été appliquées sur plusieurs trames autour du marqueurs précédents.

3. Pendant le passage principal, un expert humain a ajusté la position de quatre marqueurs fréquenciels pour correspondre à la position des 4 premiers formants. Pour ce faire, l'expert disposait à l'écran de : 1) l'identité de la voyelle, 2) le spectrogramme du signal, 3) les sorties des algorithmes automatiques (superposées comme petites croix sur le spectrogramme) et, finalement, 4) un ou plusieurs spectres lissés calculés sur le signal à un endroit que l'expert pouvait choisir à volonté.

4. La durée d'une période fondamentale du signal a été mesurée (utilisant deux autres marqueurs temporels) à proximité de l'endroit où les formants ont été lus.

Un enregistrement de notre base contient donc un pointeur dans le signal, trois étiquettes phonétiques (la voyelle et le contexte) et cinq valeurs de fréquence (4 formants et une valeur de fréquence fondamentale). Cette base a été utilisé aussi par d'autres travaux [GR97]. Nous nous sommes concentrés d'abord sur la moitié (828 enregistrements) correspondant au locuteurs masculins.

2.4.2 Points de rencontre dans l'espace formantique

Le plan des deux premiers formants F1 F2 est bien connu pour sa similitude avec le plan des deux dimensions principales de l'articulation de la voyelles : l'ouverture de la mâchoire et l'avancement de la langue. Effectivement, en calculant la distributions de probabilité des voyelles dans ce plan (voir figure 2.13), nous trouvons une bonne séparabilité des voyelles si nous nous limitons aux locuteurs masculins. Cependant, dans une telle représentation les voyelles [i] et [y] sont très proches. La solution classique consiste à rajouter les formants supérieurs et à chercher une meilleure séparation dans un espace à dimension plus grande, en sacrifiant, bien sûr, l'élégance de l'analogie avec les dimensions articulatoires. Nous allons chercher un jeu de paramètres qui conserve cette analogie tout en assurant une meilleure séparabilité.

Nous remarquons sur les nomogrammes du modèle à quatre tubes que, finalement, seulement une seule résonance de chaque famille suffit à préciser les deux paramètres du modèle (Al et Xc). Les résonances supérieures sont redondantes de ce point de vue. Par contre, il est important de disposer d'un représentant de chaque famille. Afin d'éliminer la redondance de l'espace des formants, on pourra étudier la corrélation entre les formants (voir tableau 2.1). Un autre moyen est de chercher les points de rencontre dans l'espace formantique, les endroits où les voyelles ne se différencient pas

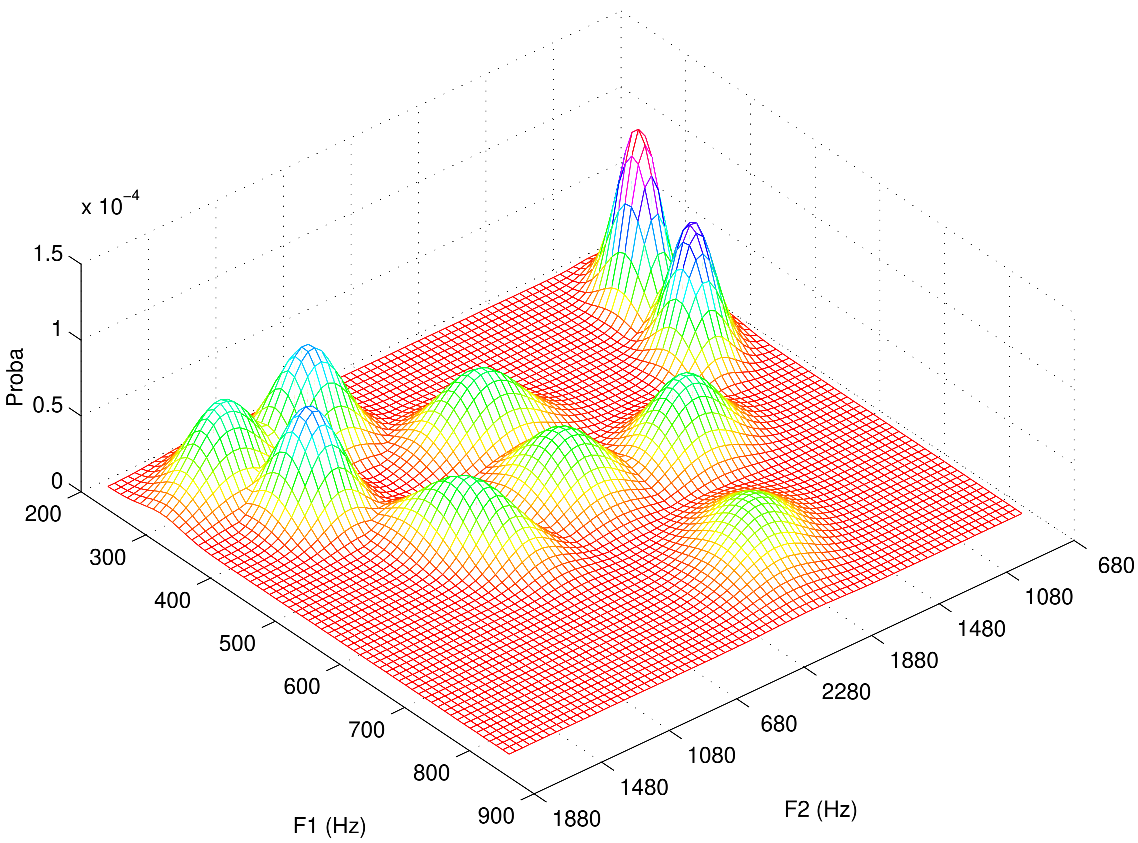


FIG. 2.13: Densité de probabilité des 10 voyelles orales du français dans le plan F1, F2. La hauteur maximale des lignes séparatrices est atteinte entre les voyelles /u/ et /o/ et entre /i/ et /y/. Cette densité est le résultats de l'apprentissage du classifieur baysien sur les locuteurs masculins de notre corpus

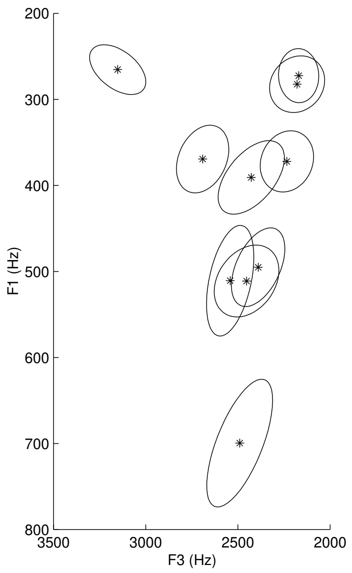


FIG. 2.14: Ellipses de dispersion des formants dans le plan F1 F3 pour 6 locuteurs. On remarque la superposition des [ɛ œ ɔ]. Cette façon d'orienter le plan F1 F3 nous facilitera la superposition avec le plan F1 F2.

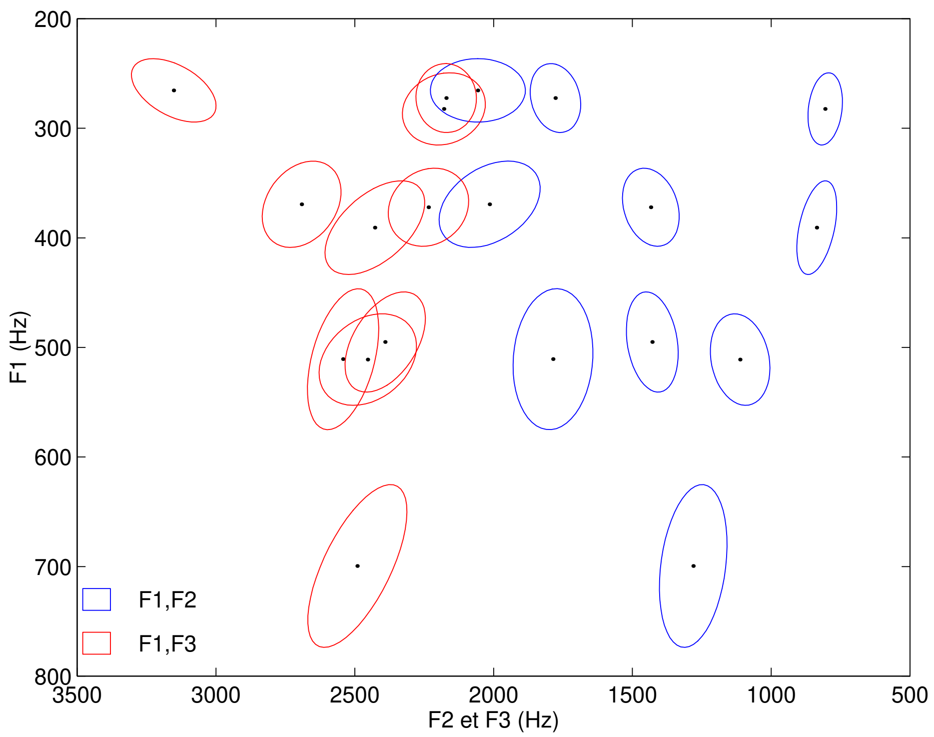


FIG. 2.15: Superposition des plans F1 F2 et F1 F3 pour les locuteurs masculins. On remarque le point de rencontre à [250Hz, 2100Hz] entre les points de ces deux espaces

selon la valeurs de leurs formants (constituant ainsi des sources de bruit pour une classification). Le plan F1 F3, celui qui présenté l'agglomérationn la plus forte (voir figure 2.14), présente un tel point de rencontre pour les trois voyelles demi-ouvertes [ɛ œ ɔ]. En plus du recouvrement des ellipses de dispersion, l'ordre de ces trois voyelles (par F3 moyen croissant) n'est pas la même d'un locuteur à l'autre.

Pour les voyelles fermées, deux se superposent bien ([u] et [y]) alors que le [i] est bien séparé. Cette situation change de manière radicale si nous superposons les plans F1 F2 et F1 F3. La figure 2.15) montré trois ellipses superposées pour les trois voyelles fermées, celle du [i] provenant du plan F1 F2. Il y a aussi un point de rencontre moins bien marqué entre les voyelles demi-fermées. C'est la voyelle [e] qui "rate" le plus cette troisième rencontre. En effet, sa position dans le plan F1 F2 et sa position dans le plan F1 F3 sont à égale distance de la position moyenne dans le plan F1 F3 des deux autres [o ø].

À partir de cette superposition des plans F1 F2 et F1 F3, nous allons opérer un regroupement différent des 20 ellipses (voir figure 2.16). La structure que nous dégageons dans l'espace formantique repose sur cette séparation. Nous allons appeler le premier plan issu de cette séparation le plan R1 R2 et le deuxième le plan R1 R3. On peut facilement remarquer que le plan R1 R3 est caractérisé par un groupement de ses points dans une bande de fréquence (largeur environ 600Hz) parallèle à la première bissectrice (droite de régression R3=1,08*R1+1818Hz pour les locuteurs masculins). Ceci nous suggère que le paramètre R3 n'apporte pas plus d'information utile que R1 et R2. On peut montrer que la transformation entre l'espace F1 F2 F3 et R1 R2 R3 est une rotation dépendante de la voyelle. La matrice de cette rotation est la matrice unité pour toutes les voyelles sauf [i] et [e]. Malgré la relative simplicité de cette transformation, nous montrerons qu'elle apporte des avantages pour la classification phonétique.

entre		c	entre		c
F1	F3	0.05	R1	R3	**0.71**
F1	F2	-0.17	R1	R2	-0.31
F2	F3	0.54	R2	R3	-0.46

TAB. 2.1: Corrélation entre les formants et entre les résonances. On remarque la correlation importante entre R1 et R3.

2.4.3 Prédiction de la position du R3

Notre jeu de paramètres R1 R2 R3 est défini comme suit :

- R1=F1, R2=F3, R3=F2 pour les voyelles [i] et [e]
- R1=F1, R2=F2, R3=F3 pour les autres voyelles.

L'utilisation de ce jeu de paramètres R1 R2 suppose donc une pré-décision de type avant/arrière sur les voyelles. Nous allons examiner ici la possibilité de la faire automatiquement.

Si R3 est corrélé avec R1 (alias F1) alors on devrait pouvoir récupérer sa position en connaissant la valeur de F1. Nous construisons donc un prédicteur, R3P, de la position du R3. Nous prendrons le décision avant/arrière en comparant les distances entre ce prédicteur et les formants F2 et F3 d'une voyelle donnée. En outre, nous disposons ainsi d'une méthode simple pour le passage de F1 F2 F3 à R1 R2 R3 : R1 vaut toujours F1 et R3 est le formant le plus proche de R3P.

Nous avons essayé plusieurs lieux du R3P dans le plan R1 R3 (un segment de droite, une sigmoïde approchée par plusieurs segments). Finalement, le meilleur prédicteur est représenté par deux paliers. La cassure entre les deux segments de droite semble nécessaire pour palier au manque de corrélation entre R3 et R1 dans la zone des voyelles demie-fermées. Nous avons calculé un prédicteur pour chaque locuteur selon l'équation :

- R3P = F3moyen(a), si F1 > (F1moyen(ø)+F1moyen(œ))/2
- R3P = (F3moyen(i)+F3moyen(u)+F3moyen(y))/3 sinon

L'algorithme qui utilise ces prédicteurs est très performant pour l'identification du R3 selon la définition donnée plus haut. Il atteint 99.3% de réussite sur les 7 locuteurs masculins (966 voyelles) et 98,9% sur les 6 locutrices (828 voyelles). Toutes les erreurs concernent seulement les voyelles [e] et [ɛ]. Nous pensons que les rares erreurs peuvent être attribuées à l'imprécision de l'étiquetage phonétique sur [e] versus [ɛ].[10] L'algorithme est robuste par rapport au variation des paramètres. Ainsi, pour un R3P indépendant du locuteur (R3P = 2150 Hz si F1 <= 420 Hz sinon R3P = 2750 Hz), le

[10]Nous n'avons pas réexaminé l'étiquetage

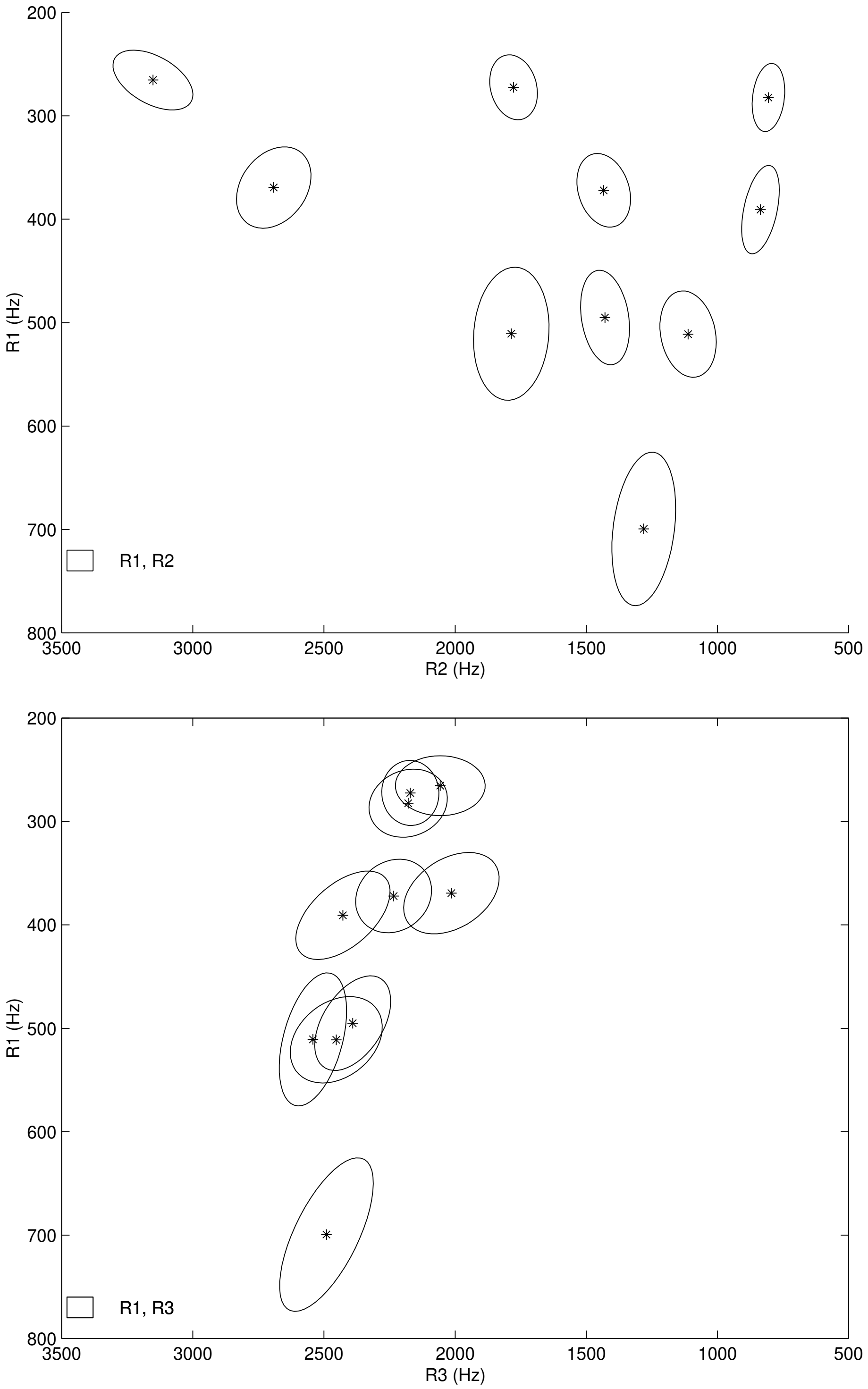


FIG. 2.16: Plan R1 R2 (a) et plan R1 R3 (b) pour 7 locuteurs masculins

taux de réussite est encore de 97,9% sur les 7 locuteurs masculins. Il est important à noter que la détermination correcte du R3 est sensible à la stratégie d'opposition d'ouverture propre au locuteur et non-prédictible à partir de son espace maximal. Cela nous a conduit à l'utilisation des moyennes des voyelles centrales dans le prédicteur adapté au locuteur en dépit de la volonté de garder le nombre de connaissances nécessaires a priori au minimum.

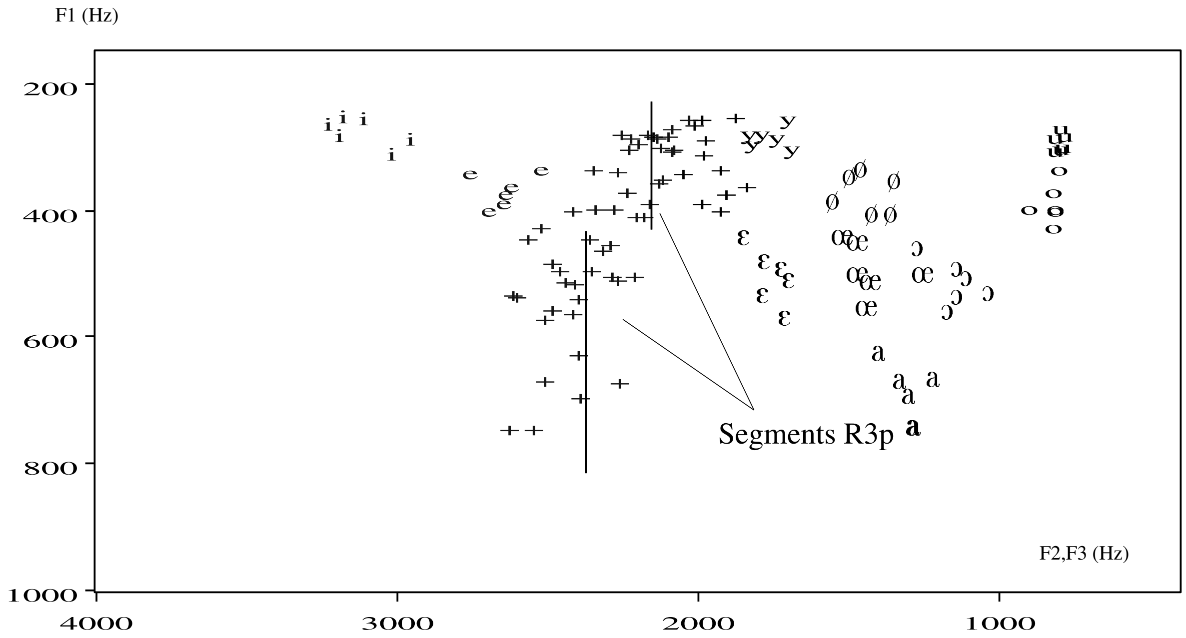


FIG. 2.17: Position du R3P dans l'espace vocalique des 6 locuteurs masculins. Les points identifiés comme appartenent au plan R1 R3 sont marqués par +.

D'un point de vue strictement algorithmique, notre transformations de l'espace F1 F2 F3 peut être expliquée par analogie avec une distribution bivariée de type XOR des exemples de deux classes (voir figure 2.18). Comme dans l'exemple présenté sur la figure, nous identifions un sous-ensemble des exemples pour lequel nous opérons une rotation des paramètres avant la classification. Une telle procédure ne vise pas particulièrement à améliorer la séparabilité (un classifieur adéquat comme le perceptron multi-couche aurait identifié les O et les X aussi bien avant qu'après la rotation) mais rend la distribution des paramètres apte à être représentée par des prototypes.

Finalement, nous souhaitons rappeler les simplifications qui différencient notre jeu de paramètres R1 R2 R3 et les vraies résonances du conduit vocal :

– Notre R1 vaut toujours F1 même si, pour certaines réalisations des voyelles [o] et [a] par exemple, F1 est associé à la cavité buccale totale.

– Pour la voyelle [i], R2 vaut F3 même si, pour certains locuteurs, le premier formant affilié à la cavité avant est souvent F4.

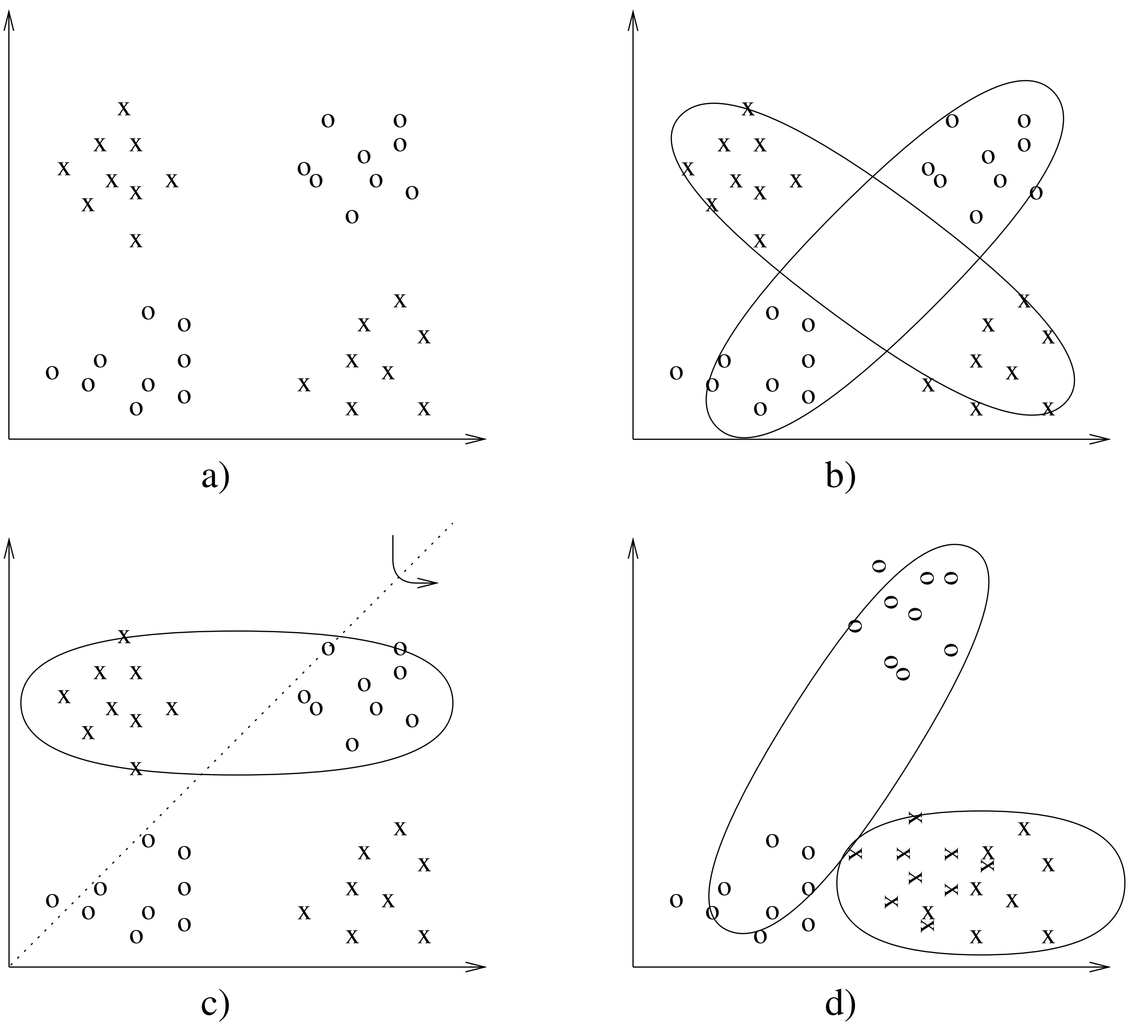


FIG. 2.18: Exemple de distribution bimodale. a) Disposition des exemples de deux classes (o et x), b) apprentissage des classes par un modèle gaussien sans transformation des paramètres, c) identification d'une modalité et rotation de l'espace des paramètres pour cette modalité (ici simple changement des axes) d) apprentissage des classes par un modèle gaussien après rotation.

2.4.4 Premières évaluations

Afin d'évaluer le gain maximum que le jeu de paramètres R1, R2, R3 peut apporter, nous allons par la suite supposer une extraction idéale selon la définition donnée précédemment. Nous confrontons les représentations formantiques et résonantiques sur une tâche de décodage acoustico-phonétique. Nous utilisons la classification bayesienne avec une fonction de probabilité gaussienne. La raison est que l'apprentissage du classifieur gaussien (CG) nous fournit des représentations facilement comparables avec les données de la littérature (moyennes et matrices de covariance).

Nous présentons les résultats (TR, pour taux de reconnaissance) d'un test de reconnaissance croisé[11], test effectué avec différents jeux de paramètres et pour données exprimées d'abord en hertz. Nous montrons aussi le taux de bonne identification en test sur le corpus d'apprentissage (TA pour taux d'apprentissage) afin de montrer que l'hypothèse de la distribution gaussienne est satisfaite. En outre, le taux d'apprentissage donne une borne supérieure pour la capacité des différents systèmes de paramètres à apprendre des prototypes vocaliques. Ces taux sont calculés sur 828 exemples de voyelles (tableau 2.2).

Paramètres	TR	TA	TA-TR
F1 F2	84,4%	90,7%	6,3%
F1 F2 F3	89,2%	94,8%	5,6%
R1 R2	90,3%	94,2%	3,9%
R1 R2 R3	91.0%	95,7%	4,7%
F1..3 - F1..2	4,8%	4,1%	-
R1..3 - R1..2	0,7%	1,5%	-

TAB. 2.2: Taux de reconnaissance et d'apprentissage pour 6 locuteurs.

Locuteur :	TR	TA
BP	89,9%	93,4%
FC	91,3%	94,1%
JB	93,5%	93,7%
JO	80,4%	97,0%
JS	94,2%	94,4%
LC	87,7%	94,4%
SL	99,2%	92,4%

TAB. 2.3: Exemples de taux de reconnaissance pour le jeu de paramètres R1 R2. TR calculé en testant les voyelles d'un locuteur, TA calculé sur les voyelles des autres locuteurs.

Nous n'attachons pas d'importance au fait que les TR sont meilleurs pour les résonances. C'est une simple conséquence de l'introduction des connaissances a priori avant la classification (par

[11] apprentissage sur cinq locuteurs et test sur le sixième par tour de rôle

exemple, toutes les confusions entre [e] et [ɛ] ont été ainsi supprimées). En revanche, nous remarquons deux autres propriétés qui nous démontrent que cette connaissance introduite est robuste :

– La dégradation entre TA et TR est plus petite pour les résonances (le minimum est obtenu pour le couple R1 R2).

– La dégradation pour le passage de trois à deux paramètres est moindre pour les résonances (5 fois plus petite). L'investigation en détail nous montre que le TR pour R1 R2 est meilleur que celui obtenu à partir de R1 R2 R3 pour 3 locuteurs sur 6 tandis que le TR pour F1 F2 sont systématiquement inférieurs aux résultat pour F1 F2 F3.

Nous avons reconduit le même test avec les données exprimées en barks. Nous avons calculé aussi pour tous les cas le nombre d'erreurs graves (les erreurs où même le deuxième candidat proposé par le classifieur ne correspondait pas à la référence). Tous les indices ont montré les mêmes tendances : meilleurs résultats et dégradation plus faibles pour les résonances.

2.4.5 Étude de la séparabilité

L'indice de séparabilité le plus simple à calculer et le rapport entre la distance moyenne interclasse et la distance moyenne intraclasse :

$$R = \frac{\sum_{i=1}^{M} N_i * D(m_i, mg)}{\sum_{i=1}^{M} \sum_{j=1}^{N_i} D(x(i,j), m_i)} \tag{2.1}$$

où M est le nombre de classes, N_i est le nombre d'exemples dans chaque classe, m_i est le centre de gravité de la classe i, mg est le centre global, $D(x, y)$ est la distance euclidienne entre x et y et $x(i, j)$ et le j-ème exemple de la classe i.

Nous calculons cet indice pour chaque locuteur et pour différents jeux de paramètres. Notre étude est similaire dans la méthode à celle de van Bergem dans [vB93]. Aux données utilisées dans le test précédent, nous ajoutons maintenant les six locutrices et le locuteur[12] qui nous a prêté sa voix pour le test de perception.

Les résultats sont présentés dans la figure 2.19. Nous observons que :

– La séparabilité est meilleure pour R1..3 par rapport à F1..3 et pour R1..2 par rapport à F1..2. Le test de Student montre que ces comparaisons sont hautement significatives (niveaux p inférieurs à 10^{-5}).

– La séparabilité est meilleure pour deux paramètres que pour trois paramètres (niveaux p inférieurs à 10^{-4}) aussi bien pour formants que pour résonances. Cet aspect est souvent caché dans les tests de reconnaissance. La raison est à chercher dans le plan F1 F3 où les voyelles demi-ouvertes ([ɛ œ ɔ]) sont pratiquement superposées.

– Les locuteurs montrent une meilleure séparabilité que les locutrices (voir tableau 2.4). La raison est partagé entre précision de prononciation de la part des locuteurs enregistrés et adé-

[12]Ce locuteur a une fréquence du R3 élevée : F2(i)=2350Hz par rapport au F2(i)≈2000Hz pour les six autres locuteurs

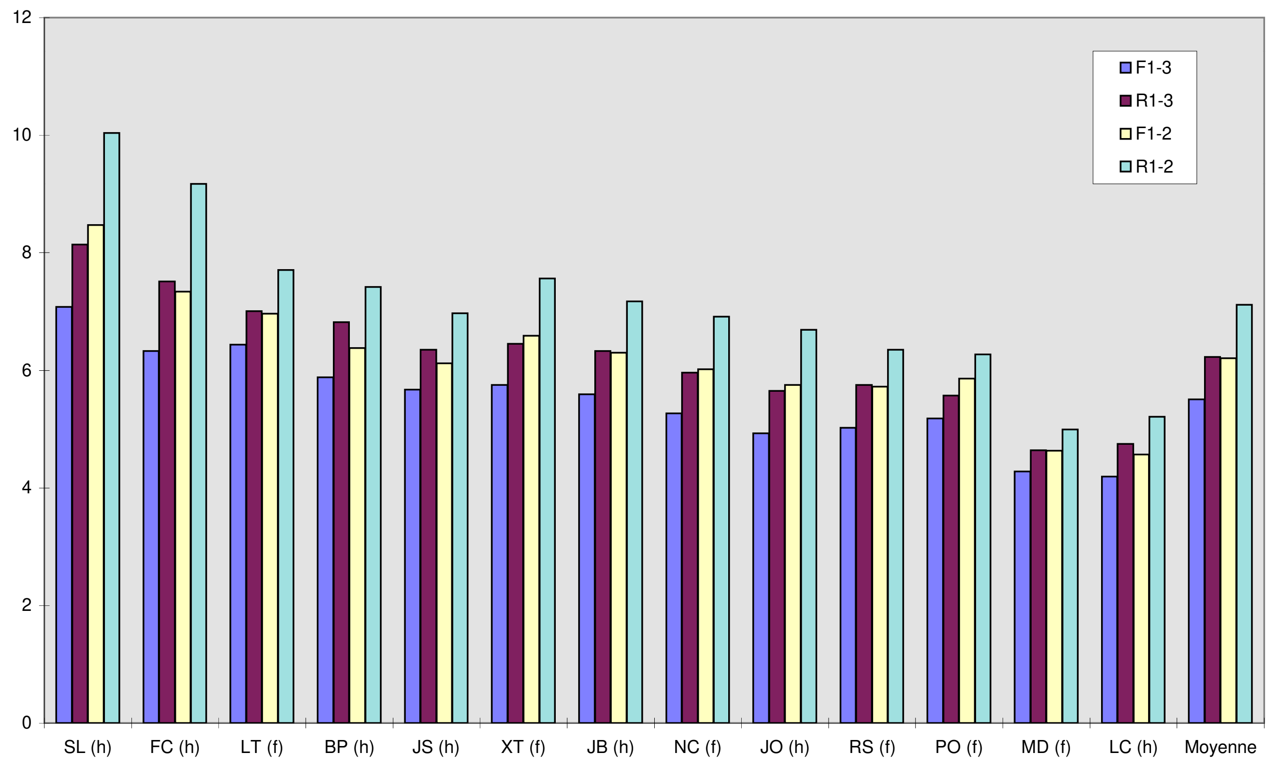


FIG. 2.19: Séparabilité des voyelles pour 13 locuteurs pour quatre jeux de paramètres. Fréquence des formants exprimée en Barks

Jeux :	F1..3	R1..3	F1..2	R1..2
hommes	5.67 ±0.93	6.51 ±1.13	6.42 ±1.23	7.52 ±1.61
femmes	5.32 ±0.73	5.90 ±0.81	5.46 ±1.44	6.63 ±1.00
mixtes	5.51 ±0.83	6.23 ±1.00	6.21 ±1.04	7.11 ±1.39

TAB. 2.4: Séparabilité des voyelles pour quatre jeux de paramètres. Valeurs moyennes et écarts type par groupe de locuteurs.

quation de la méthode d'analyse aux signaux.[13]

– Le locuteur choisi pour les tests des perception (JS) est dans la moyenne.

Nous obtenons le même type de résultats pour les données exprimées en Hertz et pour d'autres indices de séparabilité (λ de Wilks, R de Rao). Nous retenons que R1 R2 est un jeu de paramètres de dimension réduite qui garde l'essentiel de l'information phonétique de l'ensemble des voyelles orales.

[13]Nous rappelons que l'extraction des valeurs formantiques a été effectuée par expertise. Les outils mises à la disposition de l'expert permettent de s'affranchir de l'effet de la fréquence fondamentale F0. Cette fréquence, plus élevée pour les locutrices (voir tableau 2.7), a cependant une influence négative sur la précision de mesure : l'expert peut ne pas regarder toutes les coupes spectrales.

2.5 Évaluation par une méthode de passage interlocuteur

Dans le test de reconnaissance précédent les locuteurs avait des espaces maximaux relativement proches. La prise en compte de tout le corpus nous demande d'envisager une procédure de normalisation. En effet, nous pouvons observer facilement que, dans l'espace F1 F2, la position moyenne du [i] pour les hommes se superpose avec la position moyenne du [y] pour les femmes (voir figure 2.20).

2.5.1 Normalisation des espaces vocaliques. Un bref état de l'art

Plusieurs méthodes pour la normalisation des espaces vocaliques ont été proposées dans la littérature au cours des recherches sur les voyelles. Nous les classons dans deux catégories.

2.5.1.0.1 Méthodes basées sur des critères statistiques

- La procédure proposée par Gerstman en 1968 [Ger68], réduite à son expression la plus simple, consiste dans un recadrage de chaque paramètre. L'échelle est choisie arbitrairement (par exemple, entre 0 et 1). Il n'y a pas de relation entre la mise à l'echelle de chaque dimension de l'espace. Cette méthode suppose la connaissance des valeurs extrêmes pour chaque formant et chaque locuteur mais pas de connaissance explicite sur les voyelles.
- La procédure proposée par Lobanov en 1971 [Lob71] est une autre méthode classique dans la statistique : soustraction de la moyenne suivie de division par l'écart-type. Comme la méthode précédente, elle ne demande pas d'identification phonétique. Elle a besoin du même nombre de paramètres (2 par locuteur et par formant). La différence importante par rapport à la procedure précedente est que celle-ci fixe le centre de l'espace vocalique et pas ses extrêmes.
- La procédure proposée par Nearey en 1977 [Nea77] est un passage à une échelle logarithmique suivi de l'extraction d'un coefficient de correction spécifique au locuteur. Ce coefficient est basé sur une valeur moyenne de F1 et F2 mélangés du locuteur. L'idée sous-jacente à la méthode de Nearey est une relation de proportionalité entre les espaces vocaliques des deux locuteurs d'où la soustraction d'un coefficient dans un espace logarithmique. Une différence importante par rapport aux méthodes précedentes consiste dans le fait que, ce coefficient étant calculé sur les formants F1 et F2 confondus, il existe une relation explicite entre les transformations appliqués à chaque formant.
- Harshman et Papçun utilisent la méthode PARAFAC [Har76] pour la normalisation des voyelles [HP76]. Il s'agit d'une analyse factorielle. Ils utilisent trois poids sur chaque facteur : celui de la fréquence des formants, celui de la voyelle et celui du locuteur. Supposant une analyse à deux facteurs, cela revient à trouver une rotation du plan F1 F2 pour laquelle les contributions du locuteur et de la voyelle à la variance observée sur chaque axe soient indépendentes. Une fois cette rotation effectuée, les voyelles normalisées sont obtenues en remplacant, dans la décomposition linéaire des fréquences des formants, les poids du locuteur source par le poids d'un locuteur "standard".

- La procédure proposée par di Benedetto et Liénard en 1992 [BL92] est une méthode de passage entre deux locuteurs. C'est une transformation qui conserve les coordonnées baricentriques d'une voyelle donnée par rapport au repère formé par les trois voyelles extrêmes [a] [i] [u]. Cette méthode suppose donc le calcul d'une transformation du plan F1 F2, par exemple, de manière à superposer les voyelles extrêmes d'un locuteur sur celles d'un autre (superposition parfaite des moyennes). Elle nécessite donc 12 paramètres.[14]

2.5.1.0.2 Méthodes basées sur des critères anatomiques

- La première procédure a été proposée par Wakita en 1977 [Wak77, Wak79]. Il estime la longueur du conduit vocal du locuteur et calcule ensuite un coefficient de correction qu'il applique aux formants initiaux. Dans le même ordre d'idées, Nordström, en 1977, déforme la fonction d'aire des locuteurs masculin pour approcher celles des femmes et des enfants mais il dissocie la cavité buccale de celle pharyngale. Il calcule donc deux coefficients de passage [Nor77].
- La procédure proposée par Perrier et Payan en 1993 [PAP95, Pay92] et améliorée par Apostol en 1997 [Apo97] repose sur la notion d'affiliation majoritaire des formants à une des cavité du conduit vocal. La méthode suppose le calcul des coefficients de passage des cavités du locuteur aux cavités correspondantes d'un modèle de production (Bergame, voir [BBB$^+$96]). Cette méthode a l'avantage d'un nombre reduit de paramètres. Elle utilise comme connaissance sur le locuteur la fréquence des certains formants des voyelles [i], [a] et [u].
- La procédure proposée par Galvan en 1997 [GR97] repose sur une inversion acoustico-articulatoire. Elle utilise comme modèle de production une variante du modèle de Maeda [Mae79, Mae88] et comporte les étapes suivantes :
  1. En connaissant les formants des voyelles [i], [a], [y] et [u] du locuteur, les paramètres "anatomiques" du modèle sont ajustés pour qu'ils soit capable de produire les mêmes formants que le locuteur sur ces quatre voyelles.[15]
  2. Pour une voyelle donnée, sans toucher les paramètres anatomiques, on cherche les paramètres articulatoires qui font le modèle approcher au mieux cette voyelle.
  3. Finalement, sans toucher la configuration articulatoire, on effectue la transformation inverse des paramètres anatomiques ramenant ainsi le conduit du modèle à des dimensions "standard".
  4. Maintenant le modèle "produit" les formants normalisés. On les calcule de manière classique à partir de la fonction d'aire.

Les critères de performance d'une méthode de normalisation des espaces vocaliques tombent dans deux catégories :

- le degré de superposition des espaces normalisés, degré exprimé souvent par un coefficient de réduction de variance moyen pour la dispersion des voyelles. Un algorithme performant

[14] Moyennes de F1 et de F2 pour [a] [i] [u] pour le locuteur source et pour le locuteur cible.
[15] Cette méthode utilise donc 16 paramètres : F1..4 des quatres voyelles extrêmes pour le locuteur source.

obtient une réduction importante.

– le nombre des paramètres nécessaires à caractériser l'espace vocalique des locuteurs. Il est plus facile d'obtenir une reduction donnée en utilisant plus de connaissances sur l'espace vocalique qui doit être normalisé. Un algorithme performant devrait utiliser peu de paramètres.

Le critère de réduction de variance a été proposé par S. Ferrari-Disner [FD80]. Dans son étude comparative des différentes méthodes statistiques elle trouve que la méthode de Nearey est la plus performante pour la normalisation à l'intérieur d'une langue donnée. Ferrari-Disner met en garde sur l'impossibilité des comparaisons directes entre voyelles normalisées de langues différentes à l'aide des méthodes qu'elle évalue :

> "... it is shown that comparisons of normalized vowels of one language with the separately normalized vowels of another language are not valid if the vowel systems are different." [FD80, page 253]

Nous ne saurions donc étendre les résultats qui sont présentés par la suite à d'autres langues à part le français.

Nous n'avons pas utilisé ce critère de performance. Nous avons choisi comme critère des taux de reconnaissance, méthode d'evaluation plus proche de l'exploitation dans des sytèmes automatiques. En général, la réduction de variance est un bon critère pour une distribution de données déjà bien séparées (TA=100%). En revanche, pour le cas d'une mauvaise séparabilité, une réduction de variance ne vaut que si elle est accompagnée par une améliroation utile du taux de reconnaissance.

Nous remarquons que les travaux comparant les méthodes de normalisation [FD80, PAP95] se limitent presque toujours au plan F1 F2. Nous notons également que les méthodes "articulatoires" obtiennent une meilleure réduction de la dispersion des voyelles mais, souvent, au prix d'une connaissance a priori plus poussée sur les espaces vocaliques des locuteurs.

Nous finirons cette revue bibliographique en rappelant un domaine de recherche très proche. Il s'agit de l'exploitation de l'anamorphose fréquencielle comme prétraitement ("front-end processing") pour les systèmes de reconnaissance classiques. Si les critères de performance ne sont pas les mêmes (robustesse au bruit, coût de calcul, taux de mots erronés) les techniques de recherche et ajustement des paramètres sont les mêmes : statistiques sur les formants ou sur le spectre global. Une revue récente de ces recherches se trouve dans [GS97].

2.5.2 Méthode de recadrage des espaces vocaliques

À la différence des méthodes plus élaborées décrites plus haut, nous avons conçu une méthode simple pour rapprocher les espaces des deux locuteurs. Notre intention est de chercher les cas les plus difficiles pour la normalisation et de mesurer la performance dans ces cas par des taux de reconnaissance.

La représentation des valeurs moyennes des locutrices versus celles des locuteurs (voir figure 2.21) montre une corrélation importante. Cette linéarité nous a suggéré comme équation de passage une homothétie de l'espace formantique. Comme cette transformation vise uniquement à recaler les espaces maximaux, elle pourrait s'affranchir d'une identification de phonème a priori. Mais nous avons

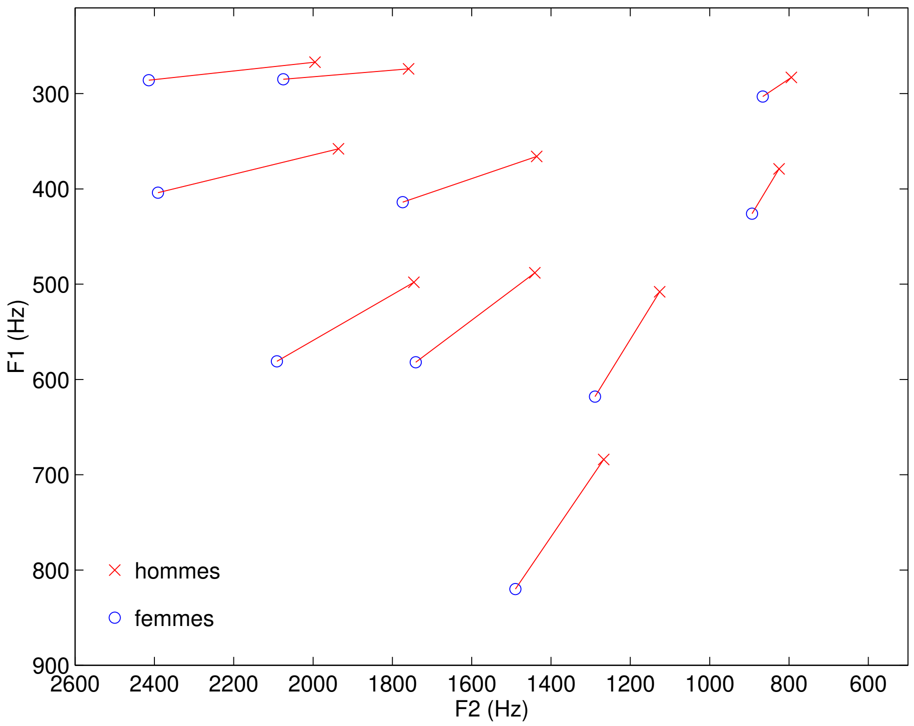


FIG. 2.20: Position relative des voyelles des locuteurs masculins et féminins dans le plan F1 F2. Le centre de la voyelle pour les hommes et pour les femmes ont été relié avec un trait pour chacune des 10 voyelles.

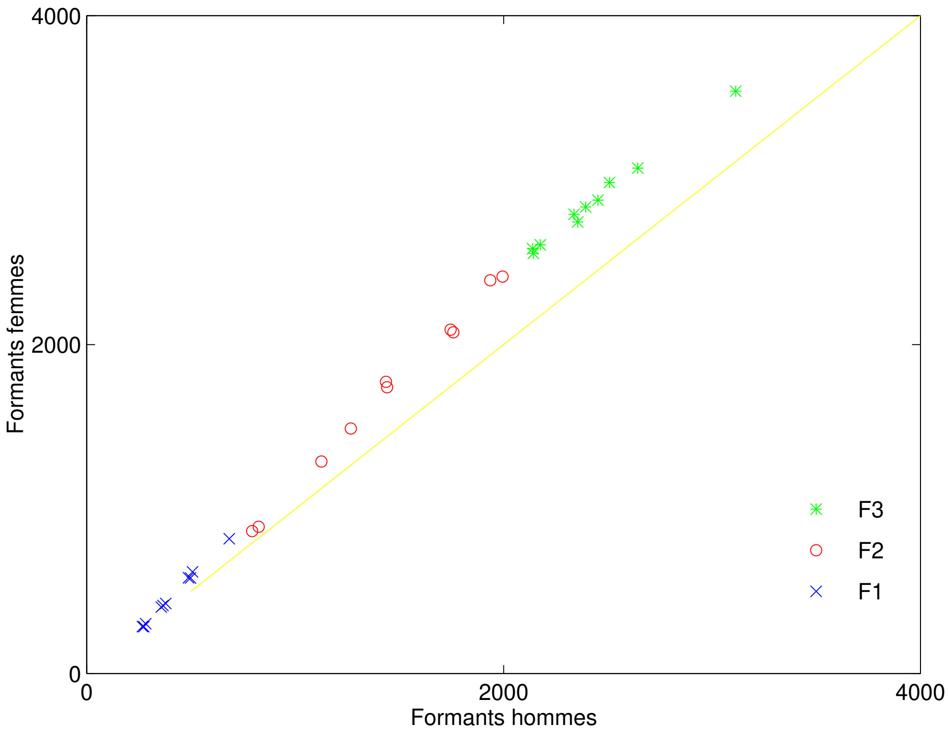


FIG. 2.21: Formants moyens des locutrices versus formants moyen des locuteurs pour les 10 voyelles. On remarque la dépendence linéaire. La première bissectrice est figurée comme référence.

utilisé comme valeurs extrêmes pour F1 le F1(a), pour F2 le F2(i) et pour F3 le F3(i). Le passage de F1, F2 et F3 avant la normalisation au F1, F2 et F3 après la normalisation est décrit par les 3 équations scalaires suivantes (en Hertz) :

$$F1_{norm} = \frac{F1(a)_{cible} - 230}{F1(a)_{source} - 230} * (F1 - 230) + 230 \quad (2.2)$$

$$F2_{norm} = \frac{F2(i)_{cible} - 230}{F2(i)_{source} - 230} * (F2 - 230) + 230 \quad (2.3)$$

$$F3_{norm} = \frac{F3(i)_{cible} - 230}{F3(i)_{source} - 230} * (F3 - 230) + 230 \quad (2.4)$$

Notre passage comporte donc six paramètres : la valeur moyenne du F1 de la voyelle [a] et la moyenne du F2 et du F3 pour la voyelle [i] pour le locuteur source et pour le locuteur cible. Selon notre expérience, ces fréquences formantiques, F2(i), F1(a) et F3(i) sont, dans cet ordre, facilement identifiables dans le signal car bien isolées sur le spectre. On peut douter de la fiabilité de F3(u), formant utilisé par d'autres méthodes [PAP95, Pay92] mais dont l'amplitude est faible [Nea94, page 47] [CM94, page90], alors que F2(i) est un paramètre robuste pour le lieu du R3 d'un locuteur.

Le centre d'homothétie a été choisi empiriquement à 230 Hertz.[16] On peut calculer un centre différent pour chaque formant (voir figure 2.21). Nous utilisons une valeur unique pour tous les formants par volonté de réduire le nombre de connaissance requises a priori. La même méthode de passage a été appliquée aussi sur l'espace tridimensionnel des résonances R1 R2 R3.

Par rapport aux autres méthodes rapportées, nous remarquerons que notre méthode s'apparente à celle de Nearey : une expansion/contraction linéaire de l'espace vocalique. Elle emprunte à celle de Gerstman l'idée de recadrage des maxima, pas des centres et aux méthode articulatoires la possibilité d'utiliser une information phonétique (certains formants de certaines voyelles). Ce dernier emprunt donne la différence fondamentale par rapport à la méthode de Nearey : les coefficients d'expansion différents sur chaque axe, calculés sur un sous ensemble précis des données. Le choix d'une méthode de passage linéaire est soutenu par les résultats de Gouvea et Stern [GS97]. En comparant un "warping" linéaire et non-linéaire, ils trouvent le deuxième à peine supérieur. En fait, selon le locuteur source, ils trouvent qu'une ou l'autre méthode est plus performante.

Notre manipulation consiste dans un ensemble de tests avec le même classifieur (CG) mais cette fois ci appris sur un locuteur (locuteur source) et testé sur un autre (locuteur cible) à tour de rôle. Nous présentons le taux moyen de bonne reconnaissance (baptisé ici TR) de l'ensemble de ces tests sans aucune normalisation. Nous comparons ce taux avec celui (baptisé ici TN) issu du même test dans lequel le locuteur cible a été rapproché du locuteur source par la méthode décrite plus haut avant la classification. Un test bilatéral de Student pour séries appariées a été effectué pour vérifier la signification statistique des résultats.

[16]Comme nous observons sur la figure 2.13, la voyelle [u] a une distribution déjà serrée. Il s'en suit le choix du centre d'homothétie dans sa proximité.

2.5.3 Résultats des tests de superposition des locuteurs

Nos premiers résultats comparent les différents jeux de paramètres pour les sept locuteurs masculins. Nous présentons la moyenne et l'écart type des taux d'identification pour les (6*7=42) combinaisons possibles de paires source-cible différentes (voir le tableau 2.5). Ces résultats appellent

Paramètres	TR	TN	p
F1 F2	69.1%	78.2%	0.0003
F1 F2 F3	70.7%	76.2%	0.0327
R1 R2	77.4%	84.3%	0.0245
R1 R2 R3	74.7%	78.3%	0.0838

TAB. 2.5: Taux de reconnaissance avant normalisation (TR) et après normalisation (TN) selon le jeu de paramètres employé. Moyennes de 42 tests (7 locuteurs masculins). Le niveau p (test de Student pour échantillons appariés) est donné.

plusieurs commentaires :

– Le jeu R1 R2 est le plus performant même avant la normalisation, ce qui donne aux résultats une certaine indépendance de notre méthode de passage interlocuteur.

– Comme dans l'étude de séparabilité, nous remarquons ici que le troisième paramètre introduit aussi du bruit dans la classification. Il faut préciser que les plans F1 F3 et R1 R3 ne présentent pas la même topologie d'un locuteur à l'autre dans la zone des voyelles demi-ouvertes ([ɛ œ ɔ]).

– L'augmentation du taux après normalisation est le résultat d'un jeu d'améliorations et dégradations. Les paires source-cible qui montraient un taux de 99% ont perdu cette performance après normalisation. Mais, plus importants, les paires montrant les pires scores avant normalisation se sont beaucoup améliorées.

Concernant la performance en réduction de variance, nous remarquons un comportement très varié selon la voyelle (voir figure 2.22). Ainsi, pour certaines voyelles la réduction est importante (par exemple, pour [ɛ] tandis que pour d'autres voyelles il peut y avoir une augmentation de variance. Nous devons faire mention ici de la difficulté de reproduction du contraste [o]/[ɔ], difficulté que les locuteurs nous ont signalé lors de nos enregistrements ultérieurs.

La suite de nos tests fait intervenir le jeu de paramètres le plus performant, R1 R2. Nous comparons les résultats des tests pour des groupes homogènes et hétérogènes de locuteurs. Nous attachons beaucoup d'importance au test qui fait intervenir toutes les combinaisons source-cible de sexe différent (voir tableau 2.6). C'est bien celui-ci le cas le plus difficile pour la normalisation. Nous avons repris les tests avec les valeurs des formants (et résonances) exprimées en barks et nous avons obtenu la même relation entre les différents jeux de paramètres.

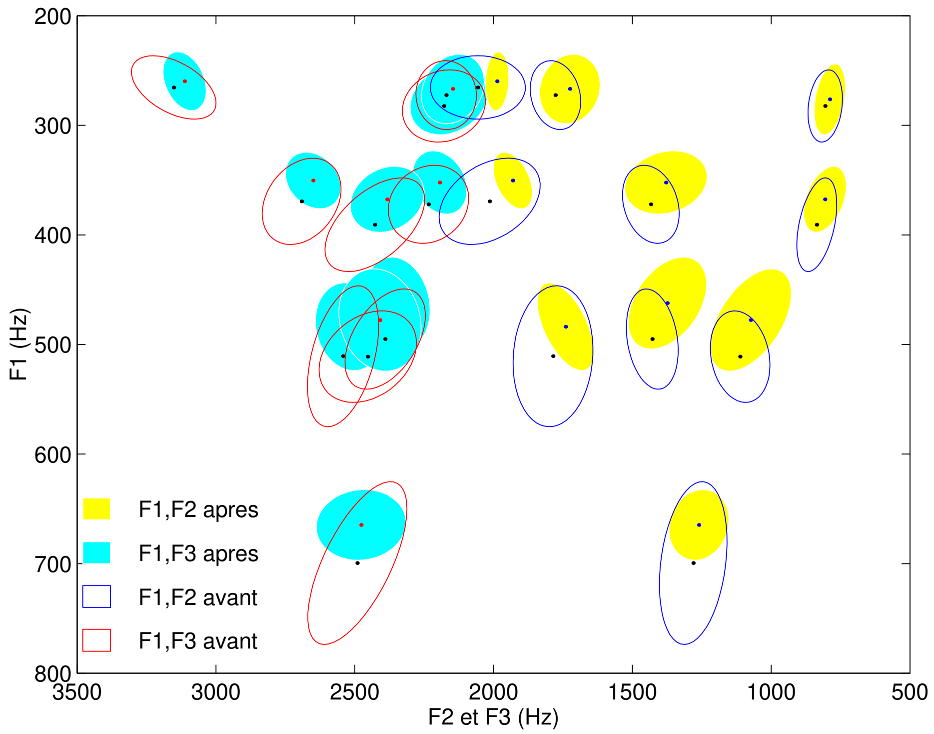


FIG. 2.22: Réduction de variance pour 7 locuteurs après normalisation (locuteur cible BP). On remarque que la réduction n'est pas uniforme sur les voyelles. Pour la voyelle [ɔ] notamment, il s'agit d'une augmentation de la variance

Type de passage	nb. tests	TR	TN
homme-homme	30	82.7%	87.1%
femme-femme	30	83.2%	84.9%
homme-femme	72	**55.7%**	**84.9%**
mixtes	132	65.5%	84.9%

TAB. 2.6: Taux de reconnaissance avant normalisation (TR) et après normalisation (TN) selon le type de passage. Moyennes des tests en aprentissage sur un locuteur et test sur un autre.

2.5.4 Conclusions

Nos résultats montrent que le recadrage linéaire des espaces vocaliques peut assurer en décodage acoustico-phonétique le même niveau de performance quelque soit le groupe de locuteurs considéré à condition de précéder la classification d'une décision avant/arrière. Ce niveau n'est pas particulièrement élevé (environ 85%) et laisse largement la place à des méthodes plus élaborées. Notre but est ici de montrer que :

1. La normalisation des locuteurs gagne à exploiter le paradigme des résonances du conduit vocal, surtout pour le passage homme-femme.
2. L'approximation linéaire pour le passage interlocuteur est loin d'être mauvaise : elle peut faire gagner près de 30% de taux de reconnaissance (ou réduire 3 fois le taux d'erreur) pour un groupe mixte de locuteurs.

2.6 De l'importance des stratégies d'articulation

Nos travaux, comme tout essai de normalisation, se sont heurtés au problème de la prédiction des positions des voyelles intérieures au triangle vocalique en partant de la position des voyelles extrêmes. Nos données semblent montrer que les locuteurs articulent les contrastes vocaliques de manière différente. Nous allons par la suite essayer de quantifier ce phénomène sur notre base de données.

Nous allons nous concentrer sur le contraste d'ouverture en français. F1 est le principal indice acoustique porteur de ce contraste. Nous rappelons que nos formants ont été mesurés dans des prononciations de mots du vocabulaire avec une signification réelle. Certaines locutrices montrent une organisation différentes des "lignes" du triangle vocalique correspondant aux voyelles avec un même degré d'ouverture (voir figure 2.23). Pour mesurer cette différence, nous allons prélever l'écart du F1 entre la première et la deuxième ligne du triangle vocalique, appelé ci-après D12. Nous utilisons comme unité de mesure l'écart maximal du F1, c'est-à-dire $F1(a) - F1(u)$ pour le locuteur respectif. Ce D12 présente une variation intralocuteur et une variation interlocuteur. Trois mesures du D12, une pour chaque lieu d'articulation(avant non arrondi, avant arrondi et arrière), nous renseignerons sur la variation intralocuteur. La moyenne et l'écart type de ces trois valeurs pour chaque locuteur sont présentés dans le tableau 2.7 à coté des valeurs correspondantes pour la distance D13.

De manière qualitative, nous remarquons tout de suite que les valeurs extrêmes des distances D12 et D13 sont présentées par quatre locuteurs différents. On observe aussi qu'il y a des locuteurs qui utilisent que la moitié basse fréquence de leur triangle, comme JO, ou qui utilisent l'écart maximal du F1 de manière équilibrée, comme RS.

De manière quantitative, nous allons comparer la moyenne des écart types avec l'écart type des moyennes présentés dans le tableau. La première, que nous appelons **Vintra**, rend compte de la précision de mesure de D12 et le deuxième, que nous appelons **Vinter**, rend compte des différences dans la façon d'articuler le contraste d'ouverture. Les données (voir tableau 2.8) montrent que Vintra12 est seulement 65% de Vinter12 (Vintra13 est 75% de Vinter13). Comme attendu, la différence

Locuteur		D12		D13		F0	
		moyenne	écart type	moyenne	écart type	moyenne	écart type
BP	h	0.215	0.029	0.545	0.073	112.5	6.2
FC	h	0.256	0.033	0.628	0.057	105.5	5.8
JB	h	0.268	0.034	0.595	0.094	139.6	17.5
JO	h	0.187	0.040	0.498	0.053	122.2	6.9
LC	h	0.166	0.093	**0.465**	0.069	120.2	8.0
SL	h	0.294	0.015	0.617	0.047	121.9	5.3
LT	f	**0.316**	0.042	0.472	0.052	240.4	13.7
MD	f	0.195	0.063	0.524	0.043	213.7	15.4
NC	f	0.227	0.030	**0.736**	0.042	210.3	15.9
PO	f	**0.142**	0.032	0.539	0.105	186.9	27.5
XT	f	0.263	0.011	0.637	0.016	226.7	9.0
RS	f	0.308	0.005	0.643	0.063	218.9	14.2

TAB. 2.7: Distance entre les voyelles [i y u] et [e ø o] (D12), distance entre les voyelles [i y u] et [ɛ œ ɔ] (D13) et fréquence fondamentale (F0) selon le locuteur. Les extrêmes de chaque distance en gras. Distances exprimées en fraction de la distance entre [u] et [a] (D14). Fréquence fondamentale exprimée en Hertz. F0 influence la précision de mesure des formants et, par conséquent, les distances D12, D13 et D14.

Unités	Vintra12	Vinter12	Vintra13	Vinter13
Hertz	16.2	30.8	27.7	50.2
Fractions D14	0.036	0.054	0.059	0.079

TAB. 2.8: Variation intralocuteur et interlocuteur du contraste d'ouverture exprimée en Hertz et en fractions de la distance entre [u] et [a] (D14)

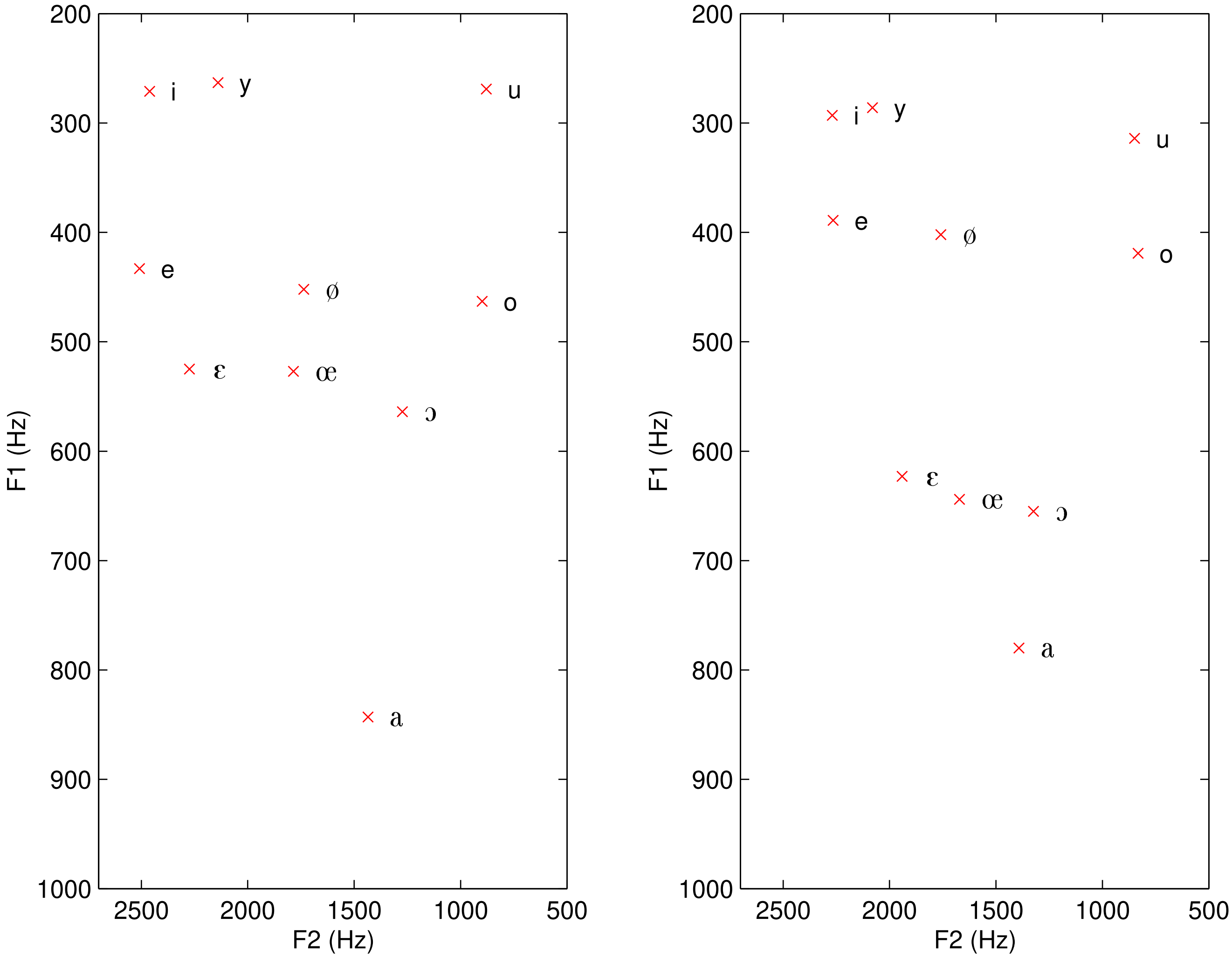


FIG. 2.23: Plan F1 F2 pour les locutrices LT (à gauche) et NC (à droite). On remarque l'occupation différente de l'écart maximal sur F1

se creuse encore si nous faisions le même calcul directement en Hertz.

Nous concluons que, malgré une structuration certaine de l'espace vocalique typique de la langue et indépendante du locuteur, une structuration secondaire apparaît, probablement au moment de l'acquisition du langage, pour chaque locuteur particulier. Deux conséquences concernent les futures travaux de normalisation de locuteur :

- L'existence de ces stratégies d'articulation compromet d'emblée l'idée de la conservation des coordonnées baricentriques.
- Plus important, ces stratégies individuelles de contraste (ici pour la machoire) semblent ne pas être déterminés par de simples différences anatomiques. Autrement posé, il serait impossible de récupérer ces différences acoustiques à partir seulement des gabarits maximaux (articulatoires ou acoustiques).

2.7 Vers l'extraction automatique

Nous allons décrire brièvement des essais de classification utilisant les voyelles du corpus que nous avons enregistré pour l'étude des consonnes. Il s'agit de 6624 réalisations des 10 voyelles orales prononcées par huit locuteurs masculins (plus de 20 réalisations par voyelle et par locuteur).

À la différence des tests précédents, ici les valeurs des formants sont détectées automatiquement. Nous utilisons une analyse LPC. Plusieurs longueurs de fenêtre temporelle (20 à 40 ms) et plusieurs ordres de l'analyse (16 - 20) ont été testés. Nous avons retenu comme méthode de calcul les maxima du spectre contre l'extraction des racines du polynôme LPC pour des raisons de fiabilité de détection. Ces valeurs brutes sont traités par deux paquets de règles. Le premier élimine les valeurs périphériques isolés ("outliers") par une série des comparaisons (le F2 ne peut pas être à 5000 Hz pour une voyelle orale, par exemple). Nous avons noté FNO (formants - no outlier) les paramètres que ce traitement nous fournit. Le deuxième paquet élimine ces valeurs mais, en plus, tente de prédire, en regardant la fréquence et l'amplitude relative des formants,[17] l'affiliation du F2. Nous appellerons les paramètres qu'il nous fournit R1, R2, R3, car ils sont l'analogue du jeu R1, R2, R3 proposé précédemment. Nous avons calculé, en parallèle, d'autres codages pour les mêmes spectres vocaliques : énergie en dix bandes de fréquence équidistantes en Hertz (LIN1..10), équidistantes en Barks (BRK1..10), transformé en cosinus discret (DCT1..10).

Nous allons comparer pour ces codages le taux d'apprentissage de la distribution des voyelles, TA, le pourcent des confusions graves (confusions entre voyelles qui ne sont pas voisines dans le triangle vocalique), TCG et le nombre d' erreurs, encore plus graves, entre les trois voyelles /a i u/, Naiu. Les résultats sont présentés dans le tableau 2.9. Nous faisons les remarques suivantes :

- Les scores sont beaucoup moins bons que dans les tests précédent. Les raisons sont le nombre

[17] Nous avons constaté que les décisions sur la fréquence sont plus fiables. Les prédicats des règles concernent les amplitudes dans deux situations : (a) pour lever l'ambiguïté possible entre un /u/ avec F1 et F2 focalisés et un /y/ avec F2 et F3 focalisés et (b) pour lever celle possible entre une voyelle antérieure fermée non arrondie et un /u/ focalisé ou un /y/ focalisé. Le reste des décisions sont prises sur les fréquences.

Codages	nb. par.	TA	TCG	Naiu
FNO1..3	3	87.8%	1.15%	8
FNO1..2	2	83.2%	1.78%	12
R1..3	3	86.6%	1.79%	8
R1..2	2	87.0%	1.74%	8
LIN1..10	10	78.2%	5.49%	4
BRK1..10	10	88.8%	1.31%	0
DCT1..10	10	84.3%	1.92%	1

TAB. 2.9: Taux d'apprentissage (TA), taux de confusions graves (TCG) et nombre de confusions très graves (Naiu) pour différents codages des 10 voyelles de 8 locuteurs masculins. 6624 réalisations (1 exemple = 0.015%)

plus important d'exemples, l'extraction automatique des paramètres (non plus par expertise) et la proportion plus importante de voyelles intermédiaires dans ce corpus.

- Contrairement à nos attentes, le jeu R1..3 n'apporte pas d'amélioration par rapport à FNO1..3.
- Comme attendu, le jeu R1..3 est plus robuste que FNO1..3 vis à vis du passage à deux paramètres.
- Les formants ou les résonances atteignent un TA comparable avec le codage de tout le spectre mais avec beaucoup moins de paramètres.
- Le codage de tout le spectre est nettement plus performant concernant les confusions entre les voyelles extrêmes. La fiabilité de la détection des formants est donc mis en cause.

Nous n'avons pas fait encore un test direct de la décision implicite avant/arrière qui est prise par le paquet de règles qui fournit R1..3. Nous avons cependant noté comme effet un rapprochement de la distribution normale de R3 par rapport à F3 : la statistique Kolmogorov-Smirnov passe de D=0.118 pour F3 à D=0.055 pour R3.

Le niveau de performance pour l'identification de la voyelle que nous avons obtenu ici est insuffisant. Afin de ne pas grever l'identification de la consonne, nous avons décidé d'employer la connaissance a priori de la voyelle pour l'initialisation du suivi de formants que nous avons implémenté. Cependant, la classification automatique peut être utilisée si on se contente de la reconnaissance de la classe de la voyelle. Plus de détails sont fournis dans la section concernant le suivi de formants (section 5.4.2.1.2).

2.8 Deuxième pause

Avant de passer à l'étude de consonnes, nous pouvons faire un court bilan. L'interprétation des formants en termes de résonances du conduit vocal est un paradigme qui améliore la normalisation des espaces vocaliques. Les voyelles [i e] semblent se différencier des autres : dans leur cas, l'affiliation des formants est spécifique. Nous avons maintenant une vision sur l'ensemble des voyelles

entre		c	entre		c
F1	F3	-0.01	R1	R3	0.52
F1	F2	-0.13	R1	R2	-0.26
F2	F3	0.49	R2	R3	-0.44

TAB. 2.10: Corrélation entre F1 F2 F3 et entre R1 R2 R3 pour 6624 voyelles, extraction automatique. Les valeurs sont proches des corrélations obtenus sur les paramètres extraits par expertise (voir tableau 2.1).

que nous pouvons tenter d'étendre aux consonnes.

D'un point de vue pratique, la clef de cette lecture articulatoire du signal est une décision avant/arrière visant à séparer les voyelles antérieures non-arrondies. Cette séparation est encore plus profitable quand la reconnaissance fait intervenir le passages homme/femme.

Accessoirement, nous avons montré, pour la première fois à notre connaissance, que, dans la structuration des espaces vocaliques, il existe une contribution des stratégies d'articulation des contrastes et, par conséquence, une information difficile voire impossible à récupérer à partir de critères anatomiques.

Chapitre 3

L'étude des consonnes occlusives. Un état de l'art

Dans ce chapitre, je me propose de systématiser les résultats et les propositions rapportées dans la littérature sur les syllabes occlusive - voyelle. Dans une première partie, je proposerai une classification des méthodes d'analyse utilisées. Dans la deuxième section, je présenterai les théories en place selon les principales dichotomies observables. Dans la troisième section, je passerai en revue les travaux qui prennent la même approche que la nôtre. Je conclurai sur un état d'avancement de nos connaissances sur les occlusives. Nous retiendrons de cet état un nombre de questions ouvertes pour guider la suite de notre travail.

3.1 Méthodes de recherche des indices acoustiques

La littérature ayant trait aux consonnes occlusives étant très riche, j'ai limité cette étude bibliographique de deux façons.

- Nous nous intéressons seulement à l'invariance caractérisant **le lieu d'articulation** (le français présente les trois lieux les plus répandus dans les langues du monde) et non pas les autres traits phonétiques (voisement, aspiration).
- Nous considérerons seulement les publications récentes, c'est-à-dire parues après l'année 1975, malgré le fait que les idées principales ont été énoncées bien avant. La raison est que les propositions antérieures ont été reprises après cette date et ont été soutenues par des expérimentations plus solides car affranchies du handicap technologique.

Il convient de noter une première dichotomie des approches selon le but final sous-jacent à la recherche des indices acoustiques :

La séparabilité Cette direction est caractéristique pour le décodage acoustico-phonétique par la machine. Il s'agit d'un but pragmatique, étroitement lié à la reconnaissance de la parole. En reprenant notre analogie aéronautique, ce but suppose qu'on veut construire des avions qui volent vite et tant pis s'ils ne sont pas capables d'autres comportements spécifiques aux oiseaux.

L'invariance Cette direction s'intéresse surtout à la matérialité des occlusives. Il s'agit ici d'un débat linguistique fondamental qui vise à extraire un ensemble d'invariants acoustiques permettant d'identifier un segment comme étant une occlusive. Dans l'analogie aéronautique, ce but suppose une étude du vol et des comportements des oiseaux sans référence aux autres engins volants.

Bien sûr, entre ces deux approche extrêmes, ils existe de nombreuses tentatives de relier ces domaines de recherche. Notre but premier est le décodage automatique mais, lorsque nos résultats nous le permettent, nous nous intéressons à la fonction d'unité sous-lexicale du segment occlusive-voyelle.

Avant de présenter les travaux réalisés sur l'identification des occlusives, nous devons passer en revue les méthodes d'inspection utilisées afin de mettre en évidence leurs qualités respectives et, surtout, leurs limites.

3.1.1 Tests de perception

Cette méthode est basée sur l'étude des réponses des auditeurs à des stimuli acoustiques. La tâche demandée aux auditeurs est, le plus souvent, une tâche d'identification phonétique. Les stimuli sont des signaux de parole manipulés, signaux que nous appelerons par la suite **monstres acoustiques**.

Une première catégorie de tests de perception cherche à répondre à la question *"Où se trouve l'information ?"*. Ces tests procèdent par troncature ou devoilement progressif du signal ("gating") pour localiser dans le temps de manière précise la zone pertinente. Cette méthode garantit seulement que la partie gardée est suffisante et que la partie coupée n'est pas nécessaire pour l'identification phonétique. Mais, comme la parole est redondante, la partie gardée peut ne pas être nécessaire et celle coupée peut être, elle aussi, suffisante. Or très peu de chercheurs utilisent aussi le découpage complémentaire [Fur86, StBC96a]. En outre, cette méthode n'informe pas sur la nature de l'invariant dans le segment retenu. Elle ne garantit même pas l'existence de l'invariance simple.

Une deuxième catégorie de tests, plus pertinente pour notre recherche, répondent à la question : *"Quels sont les indices discriminants ?"* à travers différentes manipulations du signal. Par conséquent, ces tests utilisent des signaux de synthèse plus ou moins proches de la qualité de la parole naturelle. Ces signaux non naturels impliquent deux types de problèmes :

– Les phénomènes post - perceptifs peuvent grever l'interprétation des résultats. En effet, les signaux reconnus comme manipulés par les auditeurs peuvent déclencher des raisonnements étrangers à la perception courante de la parole. On peut y palier par choix "rejet", choix rarement utilisé ou en forçant une réponse rapide (voir [Dup89]).
– Il y a un risque de ne pas atteindre les frontières qui structurent l'espace acoustique des signaux naturels avec des signaux artificiels. Les mêmes buts phonétiques peuvent être ainsi atteints reposant sur des mécanismes alternatifs de perception. Par exemple, en utilisant une étape d'accommodation des auditeurs trop importante, on peut se trouver dans la situation de tester la capacité des humains à apprendre des sons divers et non pas celle liée à la reconnaissance de la parole.

Nos solutions pour éviter ces problèmes, de même qu'une discussion plus en détail sur les protocoles des tests de perception, seront présentées dans le chapitre sur la perception des occlusives.

3.1.2 Analyses acoustiques

Ces méthodes reposent sur des analyses statistiques des paramètres calculés sur le signal de parole et, bien souvent, sur la simulation sur ordinateur des premiers niveaux du schéma conceptuel présenté en introduction (voir figure 1.1). L'avantage de ces méthodes est de pouvoir quantifier de manière précise la capacité d'un jeu de paramètres donné à séparer les occlusives. Un autre avantage est leurs facilité d'application sur la reconnaissance automatique de la parole. Nous pouvons distinguer deux étapes pour une telle expérience :

- **L'extraction des indices**. C'est le passage du signal continu à un jeu fini de paramètres discrets dans le temps (que nous appellerons paramètres primaires). Cette étape correspond à l'analyse périphérique du schéma de la figure 1.1.
- **La classification**. La décision d'appartenance à une classe phonétique est prise selon la position d'une réalisation acoustique dans l'espace de ces paramètres.

Il est important à noter que chacune de ces deux étapes peut être effectuée indépendamment de deux façons très différentes : d'un côté par algorithmes automatiques et de l'autre côté par sessions d'expertise.

Les analyses acoustiques partagent deux limitations souvent oubliées :

- Elles ne permettent pas de démontrer l'**absence de la séparabilité** (à la différence de tests de perception). En effet, si un mauvais taux d'identification est obtenu, cela peut être dû à un mauvais choix de paramètres ou bien à une méthode de classification non adaptée ou, de manière plus banale, à une fiabilité réduite d'une méthode automatique d'extraction d'indices.
- Elles ne renseignent pas sur **la plausibilité** psycho-acoustique des paramètres choisis. Par exemple, supposant qu'on trouve des indices très discriminants dans la bande de fréquence de 16 kHz à 20 kHz, on ne peut pas raisonnablement s'y appuyer pour "résoudre" l'identification des occlusives.

Étant données ces limitations, il est évident qu'une argumentation solide doit s'appuyer sur les deux types de méthodes. Notre recherche tire profit de cette combinaison. Nous tenons à citer ici d'autres travaux de thèse qui ont pris une approche semblable et qui ont traité le même sujet ou un sujet proche : Diana Krull en 1988 [Kru88], Linda Djezzar [Dje95a], Alain Soquet [Soq95b] et Roel Smits [Smi95] en 1995. A l'exception de Soquet, ces travaux combinent tests de perception et analyses acoustiques pour les occlusives mais sans les compléter avec une analyse des voyelles. Soquet part d'une étude des voyelles qu'il étend vers les consonnes mais sans faire de test perceptif. Smits se rapproche le plus de nos travaux, par l'esprit de la méthode et par ses conclusions. Il confronte les algorithmes automatiques (perceptron mono-couche) aux réponses des auditeurs pour les occlusives sourdes et sonores. Cependant, l'étendue de notre travail nous permet de compléter de manière importante ses résultats (gamme beaucoup plus riche de monstres et de tests sur items naturels) et de mieux les expliquer (voir le paradoxe du /i/ et notre interprétation dans les chapitres suivants).

3.2 La quête pour les indices acoustiques discriminants.

3.2.1 Invariance absolue contre invariance relative. Séparabilité

Le décodage spécifique des occlusives a été longuement un problème mal abordé. Précisément, la solution a été attendue sous la forme de prototypes. Les différentes sources de variabilité étaient censée associer à ces prototypes des nuages de points gardant cependant une distance intraclasse nettement inférieure à la distance interclasse, comme pour les voyelles. Or, dès les premiers essais de représentation des occlusives, les chercheurs se sont heurtés à plusieurs nuages de points par classe. Nous appellerons le cas où, pour au moins une classe, les exemples se répartissent dans plusieurs agglomérations non-connexes, un cas de **distribution multimodale** des exemples. Smits et ten Bosh rappellent quelques aspects de la théorie de la classification [StB95] qui ont une incidence directe sur les études phonétiques. Nous dégageons pour la suite les conclusions suivantes :

- Même dans le cas des distances intraclasses très supérieures aux distances interclasses (distribution multimodale) la reconnaissance des classes peut être possible. Il n'est pas nécessaire de disposer d'un nuage de points par classe, il suffit que les nuages soient localement séparés. C'est la **séparabilité** la condition incontournable pour porter de l'information et non pas l'invariance acoustique du phonème (voir figure 3.1).
- Pour retourner sur un terrain connu, la méthode la plus souvent employée est celle dite de **l'invariance relative** par opposition à l'invariance absolue. L'idée de base est d'effectuer une partition de la distribution des exemples en plusieurs zones pour lesquelles le principe de la variance autour d'un prototype unique soit respecté pour chaque classe. On appelle ces zones les modes de la distribution multimodale. Pour les occlusives, cette partition à pris la forme de la séparation par contexte vocalique et/ou par position de l'occlusive dans la syllabe. Assumer de manière constructive la variation avec le contexte, c'est-à-dire tolérer plusieurs prototypes pour chaque occlusive, est un avantage. On peut prendre en compte ainsi le fait que, au moment même de l'occlusion, les configurations articulatoires pour la réalisation d'une phonème sont fortement influencées par l'identité des phonèmes voisines. L'inconvénient de cette approche est le remplacement d'un processus de classification par deux processus successifs où les erreurs du premier (la classification qui identifie les modes) peuvent être fatales au second processus de clasification.
- La recherche d'un espace de paramètres présentent des prototypes uniques pour chaque occlusive n'est pas désespérée. En effet, une fois la séparabilité établie dans un espace, il est mathématiquement toujours possible de trouver une transformation[1] vers un espace cible où

[1] en général, une transformation non-continue, c'est-à-dire que l'image d'un compact peut ne pas être un compact

la même distribution sera monomodale. L'inconvénient est que ce nouveau espace risque de n'avoir aucun support psycho-acoustique.

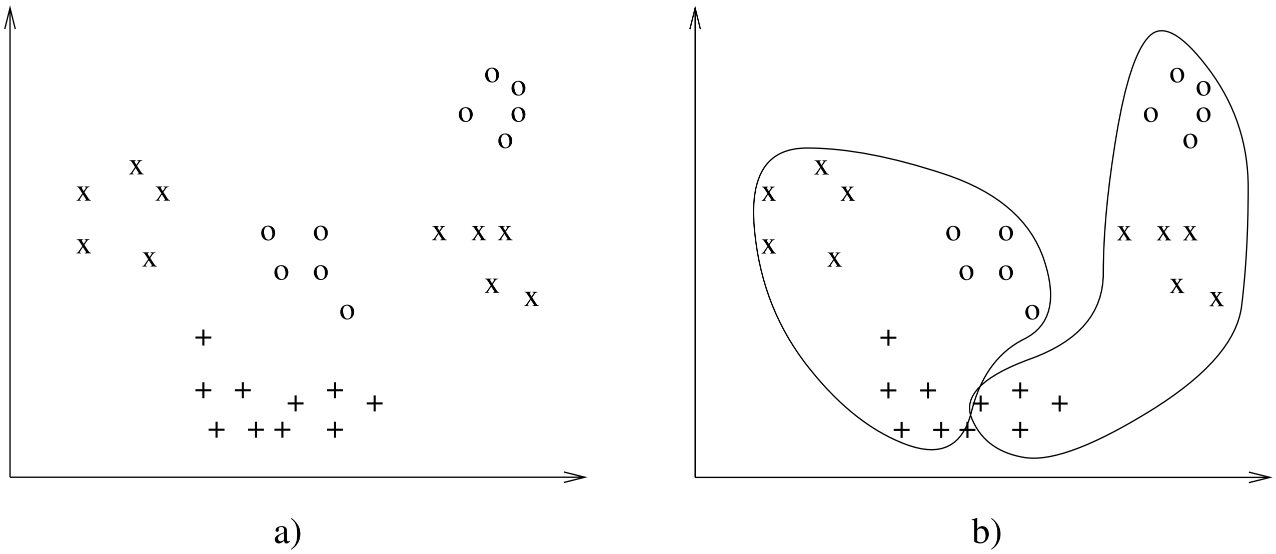


FIG. 3.1: a) Exemple de distribution multimodale pour trois classes. Les classes sont identifiables par des algorithmes basés sur le contraste (reseaux de neurones). b) Partition des exemples. Maitenant, dans chaque partition, les classes sont identifiables par des algorithmes basés sur une distance par rapport à un prototype.

Parmi les occlusives du français, le phonème /k/ présente un polymorphisme acoustique et articulatoire remarquable (palato-vélaire, vélaire). Revenant sur le rôle du contraste dans la délimitation des segments de parole, /k/ serait, peut être, mieux défini en français comme la catégorie qui est ni dentale ni bilabiale.

Le choix du contexte vocalique comme partition primaire de l'espace acoustique est naturel. La variation interlocuteur s'est avéré, dans nos tests, un facteur d'influence du deuxième ordre. En plus, selon des études de la partie sonore des syllabes VC du Hindi [Oha95], même les langues avec un inventaire consonantique riche (5 lieux d'articulation des occlusives) présentent une variation tout aussi large des indices acoustiques selon le contexte vocalique.

En conclusion, aujourd'hui il est généralement accepté que les paramètres primaires ne peuvent fournir qu'une représentation multi-prototype pour les occlusives. La recherche de l'invariance au sens classique a été abandonnée. Cependant, la recherche d'un espace de paramètres secondaires fournissant une représentation mono-prototype pour les occlusives garde son intérêt, aussi bien pour la reconnaissance automatique de la parole que pour les théories en place sur la perception des occlusives.

3.3 Indices acoustiques pour le lieu d'articulation

3.3.1 Indices du segment sourde contre indices du segment voisé

Afin de préciser les termes, avant même l'étude de notre corpus, nous allons présenter les événements acoustiques importants dans la réalisation d'une occlusive. Cette description est issue de la modélisation acoustique de la production des occlusives publiée par K. N. Stevens en 1993 [Ste93]. Il distingue quatre étapes temporelles successives dans réalisation d'une occlusive selon l'identité de la source d'excitation pour les vibrations acoustiques (voir figure 3.2) :

1. Le segment "transient" est la manifestation acoustique de l'onde de pression du au relâchement rapide de la constriction. Ce segment a une durée très courte : environ 1 ms.
2. Le segment "frication" est le segment créé par une turbulence au niveau de la constriction orale. Cette source atteint son maximum d'énergie immédiatement après le "transient" et décroit en intensité avec l'ouverture du conduit.
3. Le segment "aspiration" est excité par une source de même type mais située à la constriction formée par la glotte.
4. Le segment "voicing" est le segment caractérisé par la vibration des cordes vocales.

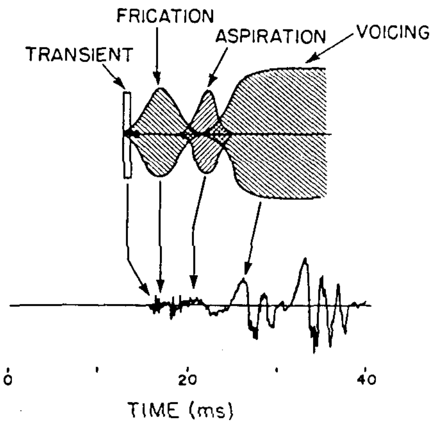


FIG. 3.2: Succession typique d'evenements lors du relachement d'une occlusive sourde d'après K. N. Stevens [Ste93].

En pratique, ces quatre segments se recouvrent plus ou moins selon la langue, le locuteur, le contexte, le débit de locution, etc. Par exemple, dans les occlusives sourdes du français, la période d'aspiration est généralement absente.[2] Pour le but de notre étude, nous allons séparer deux segments seulement (voir aussi figure 4.1 dans le chapitre suivant) :

– Le segment sourde, que nous appellerons **N-seg** (comme "Noise-like segment") dans la suite de cette thèse, est le segment qui regroupe les trois segments dont l'excitation a la nature d'un bruit.

[2] sauf pour les voyelles hautes

– Le segment voisé, que nous appellerons **V-seg** (comme "Voiced segment"), est le segment caractérisé par une composante de fréquence fondamentale plus intense que les composantes de haute fréquence.

La plus ancienne dichotomie que nous trouvons dans la littérature concerne le segment le plus approprié pour l'extraction des indices encodant le lieu d'articulation. Plusieurs arguments peuvent être cités en faveur de chaque segment. Par exemple, le V-seg par son énergie et par sa structure spectrale est plus robuste au bruit. Les indices qu'il porte seront en conséquence moins sensibles à ce type de perturbation. D'un autre côté, le N-seg bénéficie de son proximité par rapport au moment de l'occlusion. Ainsi, les indices qu'il porte seront moins sensibles au contexte vocalique, plus spécifiques à la configuration articulatoire de la consonne.

L'opposition entre les travaux sur ces deux segments est dû non seulement aux qualités intrinsèques des segments mais aussi, de manière un peu artificielle, aux méthodes de traitement de signal qui ont été utilisées. Ainsi, l'échec temporaire de la recherche d'une paramétrisation unitaire pour les deux segments de même que la relative facilité de l'extraction des indices à partir du V-seg ont influencé de manière décisive les premiers travaux. Ils ont étudié surtout la composition spectrale du V-seg.

Une partie des chercheurs se sont intéressé surtout à la structure phasique du signal et à des représentations psycho-acoustiques liées à cette structure. Ces travaux partent d'une modélisation du traitement neuronal dans système auditif périphérique [Del86]. Ces modèles ont été utilisé aussi pour l'identification du lieu d'articulation (récemment dans [MV97]). Cependant, l'application la plus prometeuse de ces modèles semble être la détection automatique des événements acoustico-articulatoires [WSE96, PSB$^+$96]. Ultérieurement, dans une tentative unificatrice, le codage du spectrogramme ("running spectra") des syllabes [NZ91] pourrait rendre compte à la fois de l'aspect phasique et tonique de la dynamique spectrale. La tendance actuelle est d'étudier et de travailler sur des représentations multirésolution du signal de parole [Bou96].

Nous allons regarder en détail ces différents approches. Nous passons en revue d'abord les indices utilisés classiquement pour le V-seg. Nous regardons ensuite les travaux sur le N-seg. Nos finirons cette section par les travaux qui essaient de traiter de manière unitaire les deux segments, travaux qui nous intéressent le plus pour la suite de nos recherches.

3.3.2 Indices du segment voisé. Équations du locus

La difficulté évidente pour l'identification du lieu d'articulation des occlusives est donné par l'absence du signal acoustique pendant la réalisation du geste consonantique. Aussi ancienne que l'observation de ce silence et l'observation d'une déformation spécifique des segments voisins. Sussman et Blumstein citent les travaux de Potter et collaborateurs en 1947, les premiers à noter une transition spécifique sur un spectrogramme de la consonne vers la voyelle. À notre connaissance, toutes les tentatives de capturer l'information contenue dans la zone voisé de la transition dans les syllabes CV se sont basées sur les formants.[3] Nous allons rappeler les deux concept importants :

[3] malgré le fait qu'aujourd'hui le systèmes de reconnaissance les plus performants utilisent les coefficients cepstraux.

– Le concept de **"locus point"** est issu des travaux classique à Haskins Laboratories dans les années 50 [DLC55]. Des tests de perception du lieu d'articulation basés sur la synthèse à deux formants ont montré l'importance de l'amplitude et de la direction de variation de la fréquence du F2. Pour réunir dans un modèle unique les trajectoires observées, il est supposé que le sens et l'amplitude de la variation de fréquence du F2 obéissent à la condition du passage de la trajectoire formantique par un point fixe au moment de l'occlusion. Ce point, appelé point du locus, serait spécifique à chaque lieu d'articulation. Le locus des dentales serait à 2 kHz, celui de bilabiales à 0.7 kHz et celui des velaires à 3 kHz. De nombreux travaux ont contesté ces valeurs, sauf pour les dentales. Ainsi, G. Fant, dans [Fan73], considère le point du locus comme une simplification sévère, non-généralisable au-delà de la synthèse à deux formants.
– Le concept de **"locus equation"** est une amélioration fondamentale du concept précédent. Cette amélioration a été proposé en 1963 par B. Lindblom [Lin63] avec l'introduction du concept d'invariance relative. Lindblom suppose que la fréquence des formants du côté de la consonne, FC, est liée à la fréquence dans la zone stable de la voyelle, FV, par une relation linéaire du type : $FC_i = k_i * FV_i + c_i$ où l'indice i est le nombre du formant. Ces sont les coefficients k_i et c_i qui vont spécifier la consonne. L'invariance est ici relative à la voyelle. Parmi les distances du point (FC_i, FV_i) aux trois hyper-plans donnés par les couples k_i, c_i pour les trois consonnes, la distance minimale identifie le lieu d'articulation. Une riche série de travaux ont confirmé cette hypothèse en multi-locuteurs et pour plusieurs langues. Nous citons les travaux bien connus de H. M. Sussman [SMM91, SHA93, SBDM97] pour l'anglais et ceux de M. Yeou [Yeo95] pour le français et l'arabe.

Les équations du locus démontrent donc la possibilité d'extraire à partir de l'acoustique du signal une caractéristique globale capable de séparer les lieux d'articulation. Le plan de la pente (k) contre l'abscisse à l'origine (c) présente des nuages bien séparés pour chaque classe d'occlusive. Cette uniformité, cette cohérence de chaque occlusive à travers les différents contextes vocaliques renouent les espoirs de succès de la théorie de l'invariance purement acoustique.

Cependant, plusieurs considérations limitent l'importance des équations du locus :

– Les droites des équations du locus se croisent. La conséquence en est que, pour le contexte correspondant au point de croisement, les fréquences des transitions formantiques sont identiques pour deux consonnes différentes. En fait, la bonne séparabilité montrée par le plan (k, c) repose sur certains contextes (par exemple, plus sur [a] et moins sur [y]).

  > "By themselves, locus equations cannot provide a complete and sufficient set of invariant cues for identification of stop place of articulation ..." [SHA93, page 1256]

  Pour palier à cet inconvénient, Sussman suggère l'utilisation d'autres paramètres mesurées sur les formants, comme la valeur du FC3 [Sus91]. Les distances d'un item aux droites du locus ont été retenues dans le jeu de paramètre le plus discriminant dans certaines comparaisons des paramètres [StBC96b]. Cepandant, en règle générale, ces distances sont rarement utilisées directement pour la reconnaissance des occlusives.
– Deuxième reproche fait à la théorie des équations du locus viennent des tests de perception.

Ainsi, Eek et Meisner [EM95], utilisant une synthèse LPC pour la manipulation des formants, montrent, entre autre, qu'une série de signaux sont perçus [p] malgré le fait que la pente (k) et l'abscisse (c) de la droite leur correspondant était complètement atypique (k=0, c=900 Hz). En conclusion, les équations du locus sont ni nécessaires, ni performantes pour la prédiction de la perception des signaux manipulés.

– Dernièrement, la classe de vélaires est moins bien décrite par une seule droite de régression. Ainsi, Sussman [SMM91, page 1315] sépare "palatal locus équation" du "velar locus équation" selon la voyelle suivante. Cette séparation est interprétable comme un changement d'affiliation majoritaire du F2, comme nous le montrons dans les simulations de production des occlusives.

Finalement, nous voulons répéter notre mise en garde contre les conclusions hâtives tiré à partir des échecs des analyses statistiques sur les formants. Ainsi, D. Kewley-Port affirme :

> "... although acoustic correlates for place of articulation may be distinctive in the formant transitions for some vowel contexts, it appears that parameters of formants transitions cannot generally serve as distinctive and reliable acoustic correlates of place of articulation in stops consonants" [KP82, page 387]

Cependant Kewley-Port ne prend en compte ni l'amplitude des formants ni l'évolution de celles-ci dans les transitions et n'étudie que 120 CV d'un seul locuteur. De plus, le plan FC2, FC3 est discriminant, à condition de connaître l'identité de la voyelle, selon ses propres résultats. Kewley-Port argumente que les tests perceptifs à partir du V-seg isolés n'ont jamais dépassé le taux de 80% de bonne identification du lieu d'articulation par les auditeurs humains [DSKR77, OS77, OS81, PS81]. Cependant, ces tests ont utilisé tous des signaux synthétiques et les influences des manipulations faites sur ces signaux font le sujet d'un vif débat (perception d'items à indices supprimés, présence ou absence des clics, phénomènes post-perceptifs, etc.).

3.3.3 Spectre du début de la syllabe

3.3.3.1 Expériences de troncation

L'étude du N-seg a ses racines dans les espoirs de chercheurs de pouvoir récupérer, à partir du signal proche de l'occlusion, une "empreinte" de la position des articulateurs. L'hypothèse de travail est qu'une "photographie" instantanée du conduit vocal prise tout de suite après le relâchement permet d'identifier le lieu d'articulation de manière univoque. Cet instantané a pris la forme d'un spectre calculé à proximité du relâchement (syllable onset spectrum).

Le mot "onset"(début) est utilisé souvent dans la littérature anglophone et de manière assez floue (parfois en se référant à un événement parfois à une période courte du signal). Nous allons adopter, dans cette thèse, la nomenclature suivante :

– Nous appelons **début de syllabe** (syllable onset) une période de durée constante, par exemple 25 ms, commençant au relâchement de l'occlusive. Conséquence de la durée fixe, cette période peut contenir soit le début du N-seg, soit le N-seg en entier plus le début du V-seg selon le lieu d'articulation, la langue, etc.

– Nous parlons aussi du **début du voisement** (vocalic onset or formant onset). Ces sont les premières périodes de fréquence fondamentale, le début du V-seg. C'est également l'endroit le plus proche du relâchement où on peut calculer les formants (au sens classique du terme).

L'argument principal en faveur du début de la syllabe comme segment discriminant a été apporté par les tests perceptif de dévoilement progressif des CV ("gating"). Ainsi, une série importante de travaux [BS80, KPPSK83, OHVM95, CG84] montrent que les auditeurs humains sont capables de récupérer le lieu d'articulation à partir de signaux aussi bref que 20 ms. Leur performance est au-dessus du niveau aléatoire même pour 10 ms [OHVM95]. Au-delà de l'identification d'un simple phonème, des modèles de préamorçage lexical [Dup89] et des modèles de pivot [DB88] ont été développés à partir de la perception d'un court segment du début de la syllabe.

Nous retenons que, de manière très claire, la présence des transitions formantiques complètes n'est pas une condition nécessaire à l'identification. Une brève période du début de la syllabe CV est suffisante. Cependant, comme la parole est redondante, l'inverse[4] peut être tout aussi vrai. Certains travaux ont procédé de manière plus systématique. Ils ont utilisé la troncation de la fin du signal mais aussi la troncation complémentaire, celle du début du signal. Ainsi S. Furui [Fur86], dans son étude des plusieurs catégories des consonnes du japonais, trouve que la partie la plus informative dans le continuum de la transition est la partie la plus abrupte :

> "A speech wave of approximately 10 ms in duration that includes the maximum spectral transition position bears the most important information for consonant and syllable perception." [Fur86, page 1106]

Nous résumons en deux points les conclusions de ce débat sur l'endroit qui concentre plus d'information sur la consonne :

– D'une part, l'information sur l'identité des plosives non-voisées est présente de manière redondante dans tout le continuum sonore de la syllabe.

– D'autre part, le segment de début de syllabe a conduit généralement à des meilleurs résultats d'identification que l'ensemble de la partie voisée de la syllabe.

Les raisons de cette soi-disant "supériorité" du début de la syllabe sur le reste de la transition sont, certes, très diverses et, parfois, discutables. Toujours est-il qu'elle s'est manifestée systématiquement pour la perception humaine comme pour la reconnaissance automatique.

3.3.3.2 Contributions dans le spectre du relâchement

Une fois la zone la plus informative établie, pour aller plus loin dans la recherche des indices, il convient de séparer de manière encore plus fine les approches qui tirent partie de cette information. Nous rappelons que "le début de la syllabe" signifie une partie du N-seg, tout le N-seg ou le N-seg plus le début du V-seg, au cas par cas. Selon l'importance accordée aux contributions des différents segments dans ce début de syllabe, nous trouvons trois approches :

[4]Cette situation inverse, "le début de la syllabe pas nécessaire et le V-seg suffisant", nous l'avons rencontré dans nos tests.

1. Approche **bruit de relâchement** ("release burst"). Une partie de chercheurs soutient que le tout début du son après l'occlusion est le plus informatif [RL89, BDL93, BDL96, Dje95a, KO87]. Ils calculent un spectre strictement non voisé, souvent sur une petite partie du N-seg ("release transient" chez Repp ou "barre de voisement" chez Bonneau et Djezzar). Des tests de perception et des test de reconnaissance automatique ont montré que les taux d'identification à partir du N-seg isolés sont corrects, même s'ils sont améliorables. Comme S. Blumstein a été parmi les premiers chercheurs à souligner l'importance de la présence du bruit dans les stimuli, ses travaux ont été souvent placé à tort dans cette catégorie. Pour le cas du français, Bonneau trouve que des segments de longueur fixe (25 ms) du bruit isolé des occlusives sourdes sont très bien reconnus (89% /p/, 86% /t/ et 87% /k/) par les auditeurs, après entraînement [BDL96].

2. Approche **indices intégrés** ("integrated cues"). Cette approche suppose le calcul d'un seul spectre sur toute la période de début de syllabe [SB78]. La fenêtre de calcul de ce spectre couvre le N-seg et une partie du V-seg plus ou moins importante selon la durée VOT. Les spectres que S. Blumstein calcule contiennent du voisement (notamment pour les consonnes labiales). Pour cette raison, ces spectres ne peuvent pas être donc directement comparés avec ceux de l'approche précédente (calculés sur N-seg isolé). Nous allons revenir sur cette approche dans un paragraphe dédié.

   > "This onset spectrum is determined both by burst spectrum and by the formant frequencies at the onset of voicing." [SB78, page 1359]

3. Approche **début de voisement** ("formants onset"). Une autre partie de chercheurs, après avoir étudiés des signaux avec et sans bruit de relâchement, décident que la forme du spectre est relativement peu influencé par ce bruit [BS80, page 660], [OHVM95]. Ils calculent donc un spectre sur la ou les premières périodes de la fréquence fondamentale. Ces travaux utilisent les occlusives voisées de l'américain. Même si les taux d'identification du lieu d'articulation par les auditeurs de différents âges à partir d'une seule période voisée sont importants, ces auteurs constatent que l'ajout du bruit influence toujours de manière favorable l'identification. L'étude du codage de ce segment temporel se réduit à un codage particulier du V-seg : on s'intéresse seulement à son début.

   > "... stimuli as short as 10-20 ms sampled from the onset of a consonant-vowel syllable were reliably identified for consonontal place of articulation, whether or not these stimuli contained moving formant transitions and whether or not an initial burst was present."
   > "The presence of a burst improved identification for the velar and alveolar places of articulation for all age groups ..." [OHVM95, page 3800]

Qu'ils s'agit d'une approche ou d'une autre, l'objet de l'étude est un spectre. À la différence des spectres vocaliques, pour ce spectre une palette très large de codages secondaires a été proposée, y compris les positions du maximas spectraux (formants) comme pour les voyelles. Un passage en revue de différents codages se trouve dans la section dédiée au N-seg.

Il est intéressant et important de noter que le N-seg, même s'il caractérise la consonne et s'il semble dépourvu des attributs vocaliques classiques (structure formantique claire), peut bien servir pour l'identification de la voyelle suivante. Plusieurs tests de perception et de reconnaissance montre des taux d'identification bien au-dessus du niveau de l'aléatoire pour la classe de la voyelle [WSR72, BS80, Kru88, OHM96, Bon96, Bon97]. Ces résultats prouvent que, bien qu'il soit informatif sur le lieu d'articulation, le N-seg n'est nullement invariant avec le contexte.

3.3.3.3 Indices statiques contre indices dynamiques

Les approches décrites ci-dessus cherchent généralement les indices du N-seg dans un spectre. Elles ont été dénommées **approches "statiques"** car elles supposent qu'un instantané pris au début de la transition suffit pour extraire l'invariance. Un débat les a opposé aux approches dites **"dynamiques"**. Ainsi, des analyses acoustiques (classification par session d'expertise) et des tests de perception ont montré que la façon de varier des indices acoustiques sur les premiers 40 ms de la syllabe sont plus importants que leur valeurs au début de cette transition [KP83, KPPSK83, LGB84]. Les défenseurs des approches dynamiques cherchent les indices discriminants dans une suite de spectres calculés par exemple tout les 5 ms ("running spectra"). En général, cette période couvre le N-seg et le début du V-seg. Cependant, plusieurs travaux ont tenté de mettre en évidence les changement spectraux à l'intérieur même du N-seg en utilisant, par exemple, la distribution Wigner-Ville [RD95, DR97, Dje95a].

Ce débat a été clos par l'adoption des indices dynamiques par les chercheurs qui ont initié les approches statiques :

> "... we would now agree that the gross shape of the spectrum is probably not the best means of capturing invariant properties for place of articulation. A metric incorporating a dynamic or time-varying parameter is probaly more optimal." [Blu86, page 201]

Il est très important de noter que cette **invariance dynamique** n'est qu'un autre nom pour l'**invariance relative**. Les défenseurs du "dynamique" comparent souvent le spectre calculé le plus proche possible du relâchement avec le spectre calculé au début du V-seg. Les défenseurs de l'invariance relative affirment que le spectre du relâchement doit être relativisé par rapport à la catégorie de la voyelle suivante [Suo85, Dje95a]. Puisqu'entre les deux extrêmes de la transition voisée, il n'y a pas généralement de changements de catégorie pour la voyelle, ces deux approches sont fondamentalement identiques.

En théorie, on devrait utiliser les mêmes métriques pour mesurer la variation entre deux instants de temps quelque soit la position de ces deux instants dans la transition (les deux dans N-seg, les deux dans V-seg ou un dans N-seg et un dans N-seg). En pratique, le changement de type d'excitation entre N-seg et V-seg pose divers problèmes. Cependant, une catégorie de métriques insensibles à l'excitation a été proposée par A. Lahiri en 1983 [LGB84]. L'idée de base est de comparer le spectre calculé en début du N-seg (spectre de relâchement) avec le spectre calculé en début du V-seg (formants onset) en faisant correspondre à chaque spectre (ou seulement à une partie de chaque spectre) une droite (droite de régression ou droite reliant les maxima spectraux). Quatre valeurs d'amplitude sont calculées à deux fréquences choisies sur ces deux droites. Les différences deux par deux entre ces quatre valeurs rendent compte du changement de la proportion des fréquences hautes et basses dans le spectre au cours de la transition et de vitesse de variation de l'énergie du signal (voir figure 3.3). Avec les repères fréquenciels proposées au début (1.5 kHz et 3.5 kHz), cette métrique s'est avérée efficace pour la séparation entre les bilabiales et les alvéolaires [LGB84, DL96, StBC96b].

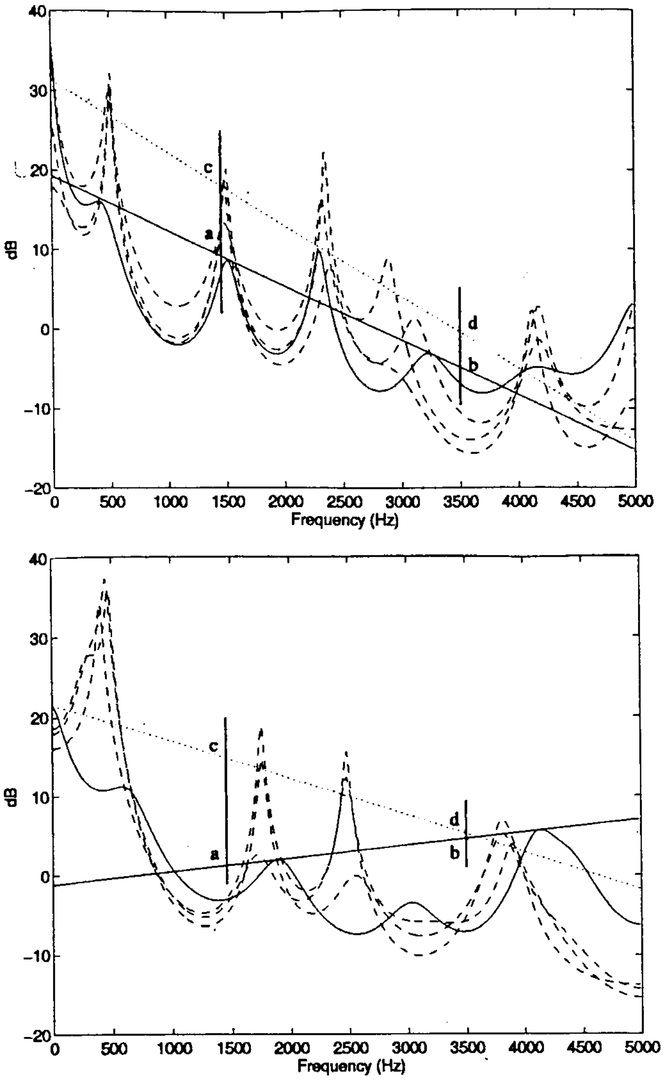


FIG. 3.3: Illustration de la metrique de Lahiri sur une syllabe [be] (en haut) et sur une syllabe [de] (en bas) d'après Dorman et Loizou [DL96]. On observe le changement de la proportion des fréquences hautes/basses entre le spectre du N-seg (trait plein) et le début du V-seg (trait interrompu) pour [de]. Ce changement n'est pas présent pour [be].

3.3.3.4 Indices intégrés contre indices fins. Niveaux de fusion

Une autre aspect qui a divisé la communauté scientifique est l'importance des indices acoustiques fins comparée à celle des indices intégrés. La différence entre ces indices tient à la résolution nécessaire pour leurs calcul :

> "We will define a **detailed cue** as the result of an acoustic measurement with high resolution in time domain (of the order of 1 ms) or in the frequency domain (of the order of 50 Hz). **Global spectro-temporal cues** are defined as the result of integrative acoustic measurements with resolutions in time and/or frequency above, for instance, 10 ms and 500Hz, respectively." [StBC96b, page 3865]

Ainsi, les fréquences des formants au début du V-seg sont des indices fins alors que la pente d'un spectre calculé sur une fenêtre à cheval sur le N-seg et V-seg est un indice intégré.

Les travaux de K. Stevens et S. Blumstein à la fin des années 70 au début des années 80 ont marqué la recherche sur les occlusives et, plus général, la recherche de l'invariance acoustique. Ce que nous retenons surtout de cette série de travaux [SB78, BS79, BS80, SB81, BIM82, Blu86, Ste80, Blu89] sont deux contributions importantes :

1. Ils ont été les premiers à souligner l'importance des **indices intégrés** (gross shape) alors que la fréquence des crêtes spectrales était, à l'époque, le seul type d'indice utilisé. Ils ont proposé la pente et la compacité spectrale.

2. Ils ont proposé un premier et élégant mécanisme pour la fusion des informations des deux segments N-seg and V-seg : le calcul spectral sur une durée fixe. La portion de la syllabe qu'ils ont étudié couvre ou pas le V-seg selon le lieu d'articulation (codant ainsi de manière implicite le VOT). Si la durée choisie pour les analyses (25.6 ms) est discutable (elle est spécifique aux réalisations des occlusives de l'anglo-américain), le principe de l'intégration sur une durée fixe est bien justifié par des considérations psycho-acoustiques.

Outre ces contributions, les travaux de K. Stevens et S. Blumstein sont devenus très populaires par les prototypes qu'ils ont proposés (voir figure 3.4) pour les spectres intégrés ("template of onset spectra"). Aujourd'hui nous savons que ces prototypes ne sont pas invariables : leur variabilité est largement démontrée par l'identification de la catégorie de la voyelle à partir de ces spectres. Mais les vrais inconvénients de ces prototypes sont liés à la difficulté de quantifier pour les systèmes automatiques les traits que les juges ont su reconnaître avec une certaine fiabilité. Ainsi, encoder la compacité de manière à ce que, en général, le spectre du /ki/ soit compact et celui du /ty/ ne le soit pas, est une tâche non-triviale.

Une alternative viable de codage à faible résolution a été proposée en 1991 par Nossair et Zahorian [NZ91]. Ils calculent le spectrogramme du début de la voyelle sur une durée fixe (60 ms). Ils codent la surface de ce spectrogramme d'une manière similaire à la compression de données pour les images (JPEG). Chaque spectre est décomposé dans une série de coefficients DCT ("Discrete Cosinus Tranform") et chaque trajectoire temporelle d'un tel coefficient est décomposé ensuite par la même procédure. Ainsi, à partir d'un jeu de quelque dizaines de paramètres, une version lissée du spectrogramme d'origine peut être reconstituée appliquant la transformation inverse (voir figure 3.5). Ce codage combine les avantages de plusieurs codages présentés précédemment :

– la durée fixe assure une pondération naturelle entre les poids du N-seg et V-seg,
– les caractéristiques spectrales globales au relâchement sont conservées,

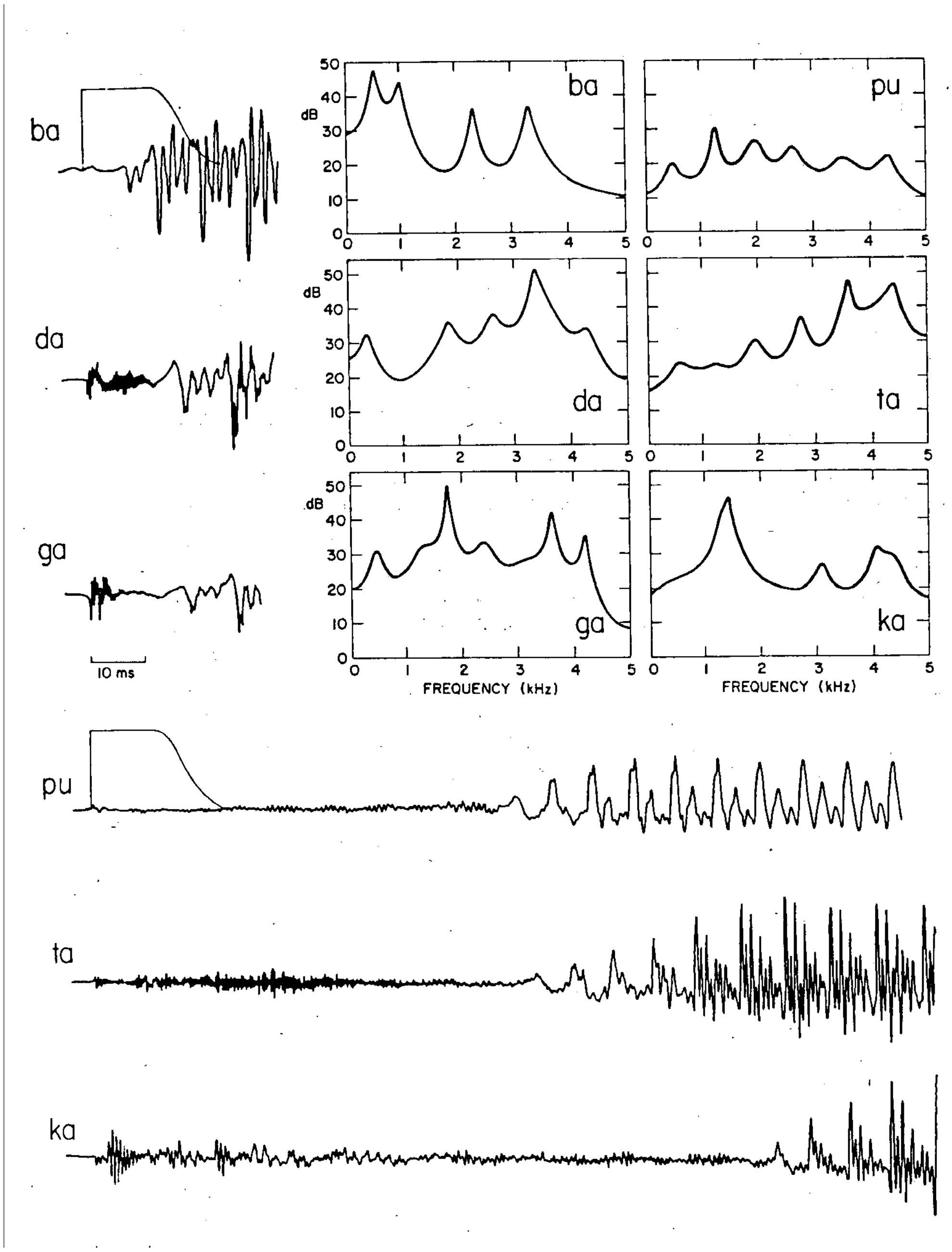


FIG. 3.4: Exemple de signaux satisfaisant les prototypes spectraux définis par Blumstein et Stevens (repris de [BS79]). La fenêtre de calcul du spectre est représentée. On note qu'il s'agit d'un spectre voisé pour [ba].

– la dynamique spectrale (changement de pente, maintien des pics spectraux,etc.) est bien rendue,
– finalement, ces paramètres sont facilement calculables.

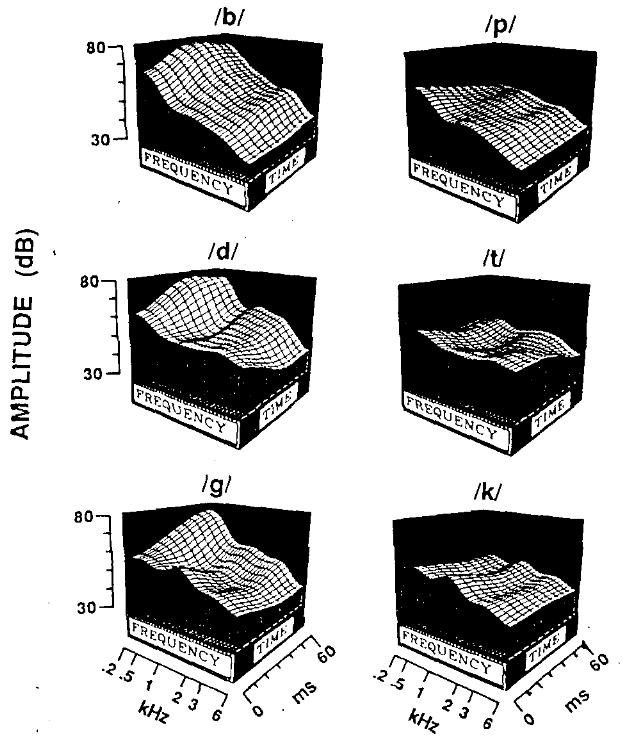


FIG. 3.5: Prototypes des spectrogrammes lisés pour les 6 occlusives de l'anglais calculés par Nossair et Zahorian [NZ91].

Ce type de codage suppose que la bonne collaboration entre les informations du segment N-seg et V-seg soit capturée pendant le calcul des paramètres. L'espace de ces paramètres reflète, de manière intrinsèque et non ajustable a posteriori, une contribution du N-seg et une contribution du V-seg. Nous appellerons ce type de fusion des informations des deux segments **fusion précoce** par rapport à sa position dans la chaîne ascendante du traitement (calcul du paramètres primaires).

Une autre approche possible est le calcul des paramètres secondaires séparément sur chaque segment pour tenter leur fusion ultérieurement. Par exemple, on peut calculer, d'un côté, la fréquence des formants au début du V-seg et, d'un autre coté, la durée et l'énergie du bruit et chercher la séparabilité dans un espace multi-dimensionnel regroupant ces paramètres. Nous appellerons ce type de fusion des données **fusion directe** car les paramètres du N-seg et ceux du V-seg sont directement présentés à l'entrée d'un classifieur. On calcule ici donc un ensemble d'**indices fins** qui sont censés porter l'information dans leurs relations réciproques.

La première étude à évaluer une gamme importante (14) d'indices acoustiques, à notre connaissance, est celui de T.J. Edwards en 1977 [Edw81]. Il montre l'importance des certains relations entre les deux segments (les rapports entre les fréquences des formants et la fréquence du bruit sont les paramètres les plus discriminants dans son classement de l'efficacité des indices).

Pour le français, A. Bonneau et S. Coste ont proposé un jeu de paramètres pour la classification des occlusives sourdes [CM94, BCDL92]. Ces travaux ont été continués par L. Djezzar [Dje95a, DP97]. Nous retenons le comportement varié des indices selon les contextes vocaliques et les bonnes performances en reconnaissance d'un système qui identifie à la fois avec la consonne la classe de la voyelle suivante (90% sur une base de données multi-locuteurs). Le classement suivant des contextes vocaliques par facilité de la reconnaissance automatique du lieu d'articulation a été trouvé : contexte postérieur (/u/), antérieur non arrondi (/i/), antérieur arrondi (/y/) et, finalement, contexte central (/a/).

Une étude récente de R. Smits et collègues [StBC96b, Smi96] comparent un jeu de 8 indices intégrés et 11 indices fins pour leur capacité à séparer le lieu d'articulation et pour leur capacité à reproduire les réponses des auditeurs humains. Sa conclusion en faveur des indices fins doit être pondérée par des considérations lié á la méthode d'extraction de paramètres (automatique pour les indices intégrés, bénéficiant d'expertise manuelle pour les indices fins). Nous retiendrons de son étude surtout la complémentarité des indices et la variation de leur capacité de discrimination avec le contexte vocalique.

Nous avons mentionné ici la fusion précoce et la fusion directe juste pour dire que la première est associée, d'habitude, à des indices intégrés et la deuxième à des indices fins. Nous traitons plus en détails la typologie des fusions dont nous donnons une taxinomie précise dans la section 5.2, avant de présenter les indices que nous avons calculés sur notre corpus.

3.4 Études comparatives des indices

Alors que les études concernant l'importance de différents indices pour le contraste de lieu d'articulation abondent, les études adressant directement la question de **l'importance relative** des indices sont peu nombreuses. En fait, l'idée sous-entendue était que l'indice ou le groupe d'indices que permet le meilleur score de reconnaissance (par les humains et/ou par les machines) est le plus important. Ainsi, la majorité des tests utilise des stimuli dont certaines indices ont été "effacés" ("**deleted-cue stimuli**") et peu de tests utilisent des stimuli où les indices sont en compétition pour la décision phonétique ("**conflicting-cue stimuli**"). Dans le chapitre sur la perception (chapitre 4), nous présentons les avantages et les inconvénients de chaque paradigme de test.

À notre connaissance, E. Fisher-Jørgensen est la première à utiliser les stimuli conflictuels [FJ72]. Elle est donc la première à constater la très forte dépendance du contexte de l'importance relative et à observer des effets de fusion de type McGurk (émergence du percept de la troisième consonne à partir des segments des deux autres). Malgré ses conclusions, les résultats des comparaisons sont souvent publiés tous contextes confondus.

M. F. Dorman et collègues [DSKR77, DR80] sont les premiers à employer l'expression "indices équivalents et complémentaires" de point de vue fonctionnel en se référant aux indices du V-seg et du N-seg ("functionally equivalent, context-dependent cues"). En utilisant des stimuli à indices supprimés, ils montrent (voir figure 3.6) que la capacité de discrimination du N-seg croit là où celle du V-seg baisse et réciproquement :

"... burts and [formants] transitions tended to be reciprocally related : Where the perceptual weight of one increased, the weight of the other declined." [DSKR77, page 109]

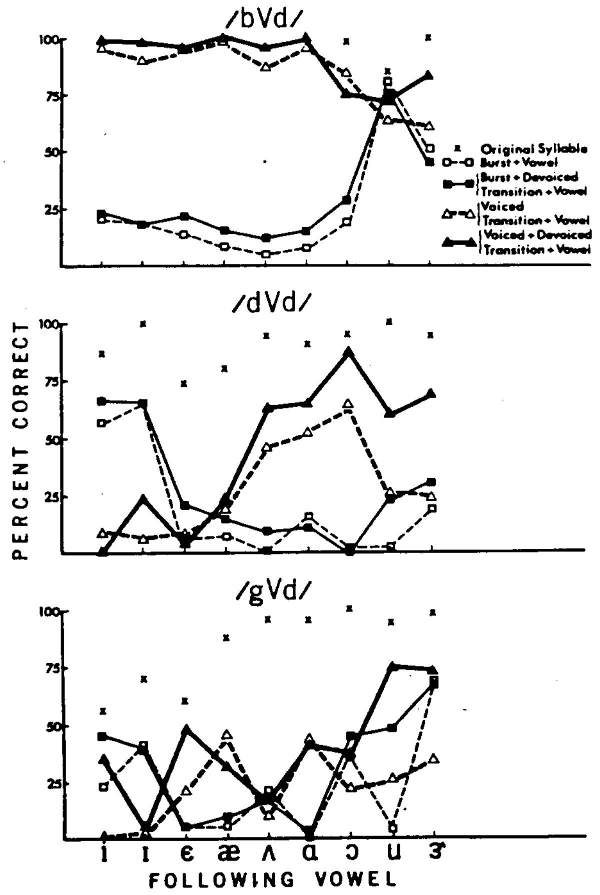


FIG. 3.6: Identification perceptive des occlusives voisées à partir des différents segments d'après Dorman et collaborateurs [DSKR77]. On note la complémentarité entre N-seg (petits carrés) et V-seg (petits triangles). On note également les mauvais résultats pour [g] en contexte antérieur.

Nous avons choisi deux articles pour illustrer la diversité des résultats rapporté dans la littérature sur la comparaisons des indices :

- M. E. H. Schouten et L. C. W. Pols ont conduit en 1983 une série de tests perceptifs avec des occlusives dans des séquences VCV [SP83a, SP83b]. Ils comparent l'importance des trois segments : la transition VC, le bruit et la transition CV. Ils remplacent avec du silence deux segments et mesurent le taux d'identification correcte à partir du segment présenté.

  "With few exceptions, **the burst elicits more correct responses** than either of the transitions." [SP83a]

  Dans leurs travaux, ils n'utilisent pas des stimuli conflictuels et les résultats sont présentés de manière globale, sans mise en garde sur l'effet du contexte.
- A. C. Walley et T. D. Carrell utilisent pour leur tests perceptifs en 1983 des stimuli conflictuels construits par synthèse à formants.

> "When adults and children were required to identify synthetic stimuli in which formant transitions conflicted with place of articulation as specified by the onset spectrum, their responses were ... **almost entirely determined by the formant transitions**." [WC83]

Cependant, les stimuli utilisés dans ces tests sont synthétiques et ne présentent que les voyelles /a/ et /u/.

De manière indirecte, les chercheurs qui ont étudié la fusion directe des indices fins font aussi une comparaison. Ainsi, ils choisissent, souvent en essayant toutes les combinaisons possibles, un "meilleur jeu" de paramètres pour la reconnaissance automatique. Nous présentons deux exemples de jeux gagnants :

1. le VOT, la fréquence de F2 et F3 au début des transitions (F2C, F3C), la fréquence et l'amplitude du maxima le plus proéminent d'un spectre lissé calculé au niveau du relâchement. Ce jeu a été utilisé pour prédire les réponses des auditeurs aux monstres acoustiques avec des occlusives sourdes [StBC96b, page 3875].
2. le VOT, l'intensité absolue du N-seg, la fréquence des deux maxima les plus importants dans le spectre du relâchement, la compacité de ce spectre et le rapport entre la fréquence du F2 au début du V-seg et celle du plus fort maxima dans le spectre du relâchement. Ce jeu a été utilisé pour la reconnaissance du lieu d'articulation des consonnes en contexte postérieur (/u o ɔ/) [Dje95a, page 118].

Nous remarquons que des indices des deux segments (voisé et sourd) se retrouvent souvent dans ces jeux choisis de paramètres.

Parmi les travaux sur l'importance relative des indices acoustiques, nous pouvons évoquer les recherches sur l'évolution de cette balance perceptive de l'enfance vers l'âge adulte [OHVM95, ELM$^+$81] et les recherches sur l'influence de la position de la voyelle dans la syllabe sur cette balance [OS77, SP83a, SBDM97]. Nous n'avons pas étudié l'influence de ces facteurs pendant nos propres recherches.

3.5 Conclusions. Chronologie du développement de la caractérisation des occlusives

3.5.1 Étape 1

La première étape de la recherche sur les occlusives a été dominée par la question :"Où sont cachés les indices du lieu d'articulation ?". Pour l'identification des occlusives, la performance des auditeurs paraissait disproportionnée par rapport à celle des systèmes automatiques. Les tests de troncation, entre autres, ont dressé un tableau plus clair de la distribution de l'information dans les différents segments de la syllabe. L'événement qui a marqué la fin de cette étape des recherches (étape "héroïque", vu les outils de traitement et d'analyse du signal par rapport à ceux d'aujourd'hui) a été la synthèse réussie des syllabes occlusive-voyelle. En effet, la mise au point du synthétiseur cascade - parallèle en 1980 par D. Klatt [Kla82] a fourni aux chercheurs un outil très puissant pour la manipulation des indices acoustiques.

3.5.2 Étape 2

La deuxième étape des recherches a été dominée par la question "Quels sont les indices acoustiques permettant d'obtenir le meilleur taux de reconnaissance du lieu d'articulation ?". L'exploration de l'espace de commande du synthétiseur a permis de redécouvrir la redondance de la parole : plusieurs paramètres concurrent à la perception (ou à l'illusion) du lieu. Nous avons assisté donc à une explosion d'indices proposés. Des travaux difficilement comparables ou soutenant des directions opposées sont la caractéristique de cette période. À notre avis, les travaux de Z. B. Nossair et S. A. Zoharian mettent fin à cette période en 1991. Ils démontrent que les consonnes occlusives peuvent être automatiquement reconnus avec suffisamment de précision. Ils utilisent 30 locuteurs (hommes, femmes, enfants, 15 pour apprentissage, 15 pour test), un corpus de 2481 syllabes CV et un codage par 20 coefficients du spectrogramme (60 ms). Ils obtiennent un score de bonne identification de 93.7% pour les six occlusives de l'anglais americain (/p t k b d g/). Il est important de noter que ce score correspond avec le résultat qu'ils ont obtenu sur les mêmes signaux pour des auditeurs humains [NZ91]. Ce niveau de bonne reconnaissance est comparable et même meilleur que ceux rapportés pour les voyelles. Cependant ces travaux laissent deux questions ouvertes :

1. Nossair et Zahorian ne font pas une analyse des erreurs de classification. Différents jeux de paramètres sont comparés mais les voies pour améliorer encore les résultats ne sont pas indiqués. Notamment, si tous les contextes vocaliques sont présents dans leur corpus, l'influence de la voyelle n'est pas étudiée.
2. Ils utilisent une segmentation manuelle des signaux. Ce bon résultat démontre donc seulement que nous sommes capables d'identifier les consonnes une fois celles-ci bien détectées. Vu la dualité fondamentale entre détection et classification, une recherche importante reste à faire pour garder le même taux en ajoutant la détection automatique des occlusives.

De cette étape nous retiendrons que le lieu d'articulation, entre autres traits phonétiques, est complètement caractérisé, au moins en théorie, par trois spectres clefs et les deux durées entre les instants de calcul de ces informations spectrales. Il s'agit du spectre du relâchement, du spectre à l'établissement du voisement, du spectre de la voyelle, du VOT et de la durée de la transition formantique. La détection de ces trois instants et le codage robuste de ces trois spectres pour l'identification du lieu d'articulation sont encore des questions ouvertes.

3.5.3 Étape 3

La troisième étape concerne l'étude des occlusives dans une perspective plus large que la classification proprement dite. En effet, la variabilité n'est qu'un sous-produit de la classification phonétique.

> "... il serait souhaitable de traiter sur le même plan tous les aspects perceptibles du signal, car il sont en interaction : aspects linguistiques mais aussi caractérisation du locuteur, de la position supposée du locuteur, des conditions d'élocution, de la prise de son. Négliger l'un ou plusieurs de ces aspects revient à créer de la variabilité sur celui auquel on s'intéresse." [Lié95]

Ainsi, dans l'ensemble des indices proposé, la question du "plus important indice" est devenu celle du "qui fait quoi ?". Autrement dit, la recherche pour éliminer à tout pris la redondance a cédé la place à une recherche de structure cohérente dans cette redondance. Les études comparatives des indices conflictuels et les études de l'évolution de la perception avec l'âge sont représentatives de cet axe de recherche sur les occlusives. Nous retenons de cette période la dépendance fondamentale de la perception avec le contexte vocalique et surtout le fait que les indices sont en collaboration et/ou compétition pour coder le lieu d'articulation, les mêmes indices peuvant modifier leur rôle selon plusieurs facteurs (contexte vocalique, condition d'élocution et d'environnement).

Chapitre 4

La perception des consonnes occlusives

Les études de perception et les études de production de la parole fournissent les connaissances a priori à intégrer dans les algorithmes automatiques. Après avoir décrit le paradigme et le protocole, nous présentons les résultats de nos tests de perception des occlusives. Nous insistons sur les aspects apparemment paradoxales des certaines conclusions. Nous finissons ce chapitre en présentant quelques résultats sur la production des occlusives.

4.1 Monstres acoustiques conflictuels. Choix de paradigme

Comme nous l'avons vu dans le chapitre précédant, plusieurs segments disjoints portent de l'information sur le lieu d'articulation. Comme chaque segment peut être à la fois suffisant mais pas nécessaire, la question la plus importante pour nous est "Où se trouve l'information *la plus prégnante pour le système perceptif*?". Nous avons mentionné deux façons de construire les monstres acoustiques typiques pour deux approches qui s'opposent :

Indices supprimés Ce paradigme suppose qu'on peut faire apparaître dans le signal un seul indice à la fois ou un groupe d'indices parmi plusieurs ensembles d'indices complémentaires. Les exemples ici sont le remplacement par du silence d'un bref segment temporel ou les fréquentes manipulations des transitions formantiques par synthèse de Klatt. Les stimuli complémentaires fournissent des taux de bonne identification dans de tests de perception et le taux le plus élevé est censé nous indiquer l'indice le plus pertinent.

Indices conflictuels Dans cette approche, chaque item porte deux indices ou deux groupes d'indices suggérant deux consonnes différentes. La consonne perçue par les auditeurs nous indique l'indice le plus important. L'exemple typique est la juxtaposition de deux segments provenant de syllabes différentes.

À notre avis, il est illusoire de tenter la suppression ou l'isolement des indices. Par exemple, il a été souvent supposé que la rectification des trajectoires formantiques "supprime" en quelque sorte l'information qu'elles contiennent. Or il s'agit dans ce cas d'indice modifié plutôt que d'indice supprimé. Plus généralement, un stimuli reconnaissable comme artificiel par les auditeurs sera moins susceptible de nous garder dans le cadre des processus perceptifs rapides, ceux qui sont utilisés dans la perception courante de la parole, ceux qui nous intéressent en premier lieu.

Nous avons donc choisi le paradigme des monstres conflictuels. Dans nos tests, les deux groupes d'indices sont produits à partir de parole naturelle comme dans les expérimentations de "tape splicing". On pourra s'attendre à ce que un signal contenant par définition une contradiction "sonne" plus artificiel qu'un autre. Mais, par un protocole expérimental suffisamment soigné, nos résultats montre exactement le contraire. Les auditeurs reconnaissent facilement les signaux tronqués mais sont incapables de distinguer les monstres conflictuels des signaux originaux non manipulés.

4.2 Tests de perception

Avant de décrire plus en détail les tests et les résultats, je donne un aperçu des toutes les tests de perception que nous avons effectués. Le but principal a été de quantifier l'apport relatif des deux segments : N-seg et V-seg. La raison principale est que la nature très différente de ces segments a conduit typiquement à des codages différents et difficilement conciliables. De plus, cette comparaison nous permettra de trancher entre plusieurs propositions contradictoires trouvées dans la littérature.

Une comparaison complète et efficace suppose, selon le "expanded factorial design" de D. Massaro [Mas87], quatre types de stimuli :

1. N-seg et V-seg cohérents
2. V-seg isolés
3. N-seg isolés
4. N-seg et V-seg conflictuels

Nous n'avons pas effectués des tests avec la troisième catégorie de signaux. Pour ceci, nous nous appuyons sur les tests très complets qui ont été effectués au CRIN [BDL93, BDL96, Dje95a] sur des N-seg isolés du français pour compléter nos propres résultats. Nous allons classer nos résultats en cinq catégories comme suit :

- Test 0 : N-seg et V-seg cohérents, V-seg isolés
- Test 1 : N-seg et V-seg conflictuels
- Test 2 : Amplitude du N-seg variable
- Test 3 : Durée du N-seg variable
- Test 4 : Illusions acoustiques

Le premier test reprend une manipulation classique et nous permet de souligner l'importance du protocole expérimental. C'est le Test 1 qui nous fournit les résultats fondamentaux qui vont guidés notre recherche. Les Test 2 et Test 3 ont été conduits afin de quantifier la contribution respective des indices temporels et des indices spectraux à la phénoménologie observée dans le Test 1. Le Test 4 veut être une vérification des connaissances que nous avons acquis par analyse acoustique. Il sera décrit en détail dans la dernière section du chapitre suivant.

4.2.1 Protocole de test

Si le paradigme des monstres conflictuels a été souvent exploité à l'époque de la bande magnétique, aujourd'hui il n'existe, à notre connaissance, qu'un seule autre expérimentation qui utilise la même approche et les mêmes méthodes [StBC96a]. Nos tests le dépassent quant à la richesse de l'inventaire vocalique et à l'étendue des manipulations (VOT, amplitude).

4.2.1.1 Les auditeurs

Les cinq tests ont été présentés aux auditeurs dans deux sessions séparées de quelques mois. La première session regroupe les items des tests 0 et 1. 34 auditeurs dont 11 auditrices ont participé à ces premiers tests. 15 auditeurs masculins dont 13 du groupe précédent ont passé la deuxième session, les tests 2, 3 et 4. Tous ces auditeurs ont le français comme langue maternelle. Ils couvrent une très grande plage d'âge. Environ deux tiers (dont les auditrices) sont naïfs par rapport à la phonétique, le reste étant des chercheurs dans la parole. À une exception près, ils ne connaissaient pas le sujet de notre investigation. Les auditeurs sont bénévoles. Ceux qui ont déclaré avoir des problèmes d'audition ont été écartés. Nous n'avons pas fait d'audiogramme pour nos auditeurs mais le très bon taux d'identification des items naturels justifie cette approche et valide notre protocole expérimental. Les réponses de tous nos auditeurs comptent dans nos statistiques finales. Chaque auditeur a passé une seule fois chaque session.

4.2.1.2 La tâche

La tâche démandée aux auditeurs est phonétique. Il s'agit d'un choix fermé entre quatre variantes possibles. Trois correspondent aux occlusives sourdes du français : [p], [t] ou [k]. Un quatrième choix était disponible : les auditeurs étaient instruits pour le choisir s'ils entendent "une absence de consonne ou aucune consonne identifiable" (ad literam). Cette possibilité de refus de la classification par les auditeurs nous démarque des autres travaux de même type. Nous ne pouvons cependant pas discerner parmi les items rejetés quels ont été perçus comme des aberrations acoustiques et quels l'ont été comme simple voyelle sans consonne. Par contre, nous pouvons être beaucoup plus sûrs des réponses qui n'ont pas été rejetés par rapport aux expérimentations qui nous ont précédées. Dans l'intention de capturer encore plus finement le degré de confiance des auditeurs, pendant la deuxième session, une choix binaire "sûr/pas sûr" a été démandée en plus pour chaque item.

4.2.1.3 Les stimuli

Un locuteur masculin phonéticien a prononcé un corpus formé de mots sans signification de type VCVCV dans des conditions d'articulation contrôlées. Il s'agit d'un locuteur dont la voix a été particulierement bien étudiée dans notre laboratoire. De ce corpus, nous avons extrait 30 syllabes CV (3 consonnes x 10 voyelles orales) accentuées. Pour chaque syllabe, nous avons placé des marqueurs temporels par expertise afin de séparer deux segments : N-seg, le segment sourd et V-seg, le segment voisé. Les critères utilisés pour le choix de l'endroit précis de cette coupure pour chaque stimuli sont communs pour tous les locuteurs enregistrés et sont décrits dans la sous-section "Segmentation" (5.3.1) du chapitre suivant. Nous montrons ici juste un seul exemple (voir figure 4.1).

Nous avons construit des monstres conflictuels en juxtaposant les segments des consonnes différentes mais gardant un même contexte vocalique. Par exemple, le V-seg de [pa] a reçu le N-seg de [ka] et celui de [ta]. Un soin particulier a été apporté à la séparation au passage par zero pour éviter la perception des "clics" dans les items recombinés. Nous n'avons pas utilisé un découpage avec recouvrement comme Smits et collaborateurs [StBC96a] car le VOT était souvent très faible (aussi bref que 5 ms pour certains [pa]). La première session a contenu donc 120 stimuli ordonnés aléatoirement (30 originaux + 30 V-seg isolés + 60(3x2x10) monstres conflictuels).

4.2.1.4 Le déroulement

Nous avons bénéficié pour les tests de la chambre sourde de notre laboratoire et de l'équipement afférent (carte son OROS AU32, casque haute qualité Beyer Dynamics, mixeur audio Yamaha MV802). Pour la première session, nous avons utilisé un niveau sonore fixe pour tous les auditeurs. Pour la deuxième, nous avons permis un ajustement en début de séance pour chaque auditeur.

Les auditeurs choisissaient une option parmi plusieurs dans un menu présenté à l'écran. Cet écran a été spécialement appauvri de tout autre élément d'interface. Chaque séance débutait par une courte séquence d'accommodation (4 stimuli) dont les réponses n'ont pas été gardées. Aucun autre apprentissage des auditeurs n'a été nécessaire. Les auditeurs ont progressé à leur rythme. Des pauses de conversation libre étaient prévues après 15 à 30 items. Ils ont mis entre 35 et 50 minutes pour une session complète. Les auditeurs ont pu réécouter chaque signal à volonté. La consigne demandait de ne pas abuser de cette facilité mais il n'y avait pas de limite ferme pour le nombre de répétitions demandé par l'auditeur. Un autre inconvénient de notre protocole est que ni temps de réponse ni nombre de réécoutes n'ont été conservées. Cela rend plus difficile de cerner l'influence des phénomènes post-perceptifs. Un autre point faible est le manque de précaution contre l'effet de mémoire qui peut apparaître d'un signal à l'autre (ni "beep" sonore entre les items, ni plusieurs ordres de précédence testés). Nous précisons que chaque séance comportait des stimuli naturels disséminés afin d'éviter un biais de l'espace perceptif des auditeurs à force d'écouter des monstres. Un autre point positif est le maintien des amplitudes naturelles (pas de normalisation d'amplitude entre les items d'une séance). À la fin du passage des auditeurs dans la chambre sourde, leurs remarques ont été recueillies et notées (sur la qualité des signaux, sur les difficultés qu'ils ont rencontrées).

4.2.2 Perception des items tronqués

Une preuve importante pour la qualité du protocole est le bon taux d'identification correcte pour les stimuli de contrôle dans le test 0 : 99,12% (26 auditeurs sur 34 obtiennent 100%). Les taux rapportés dans la littérature plafonnent autour de 96%. Seulement deux items sont affectés par des erreurs :

- La syllabe [py] a été confondu avec [ky] 6 fois

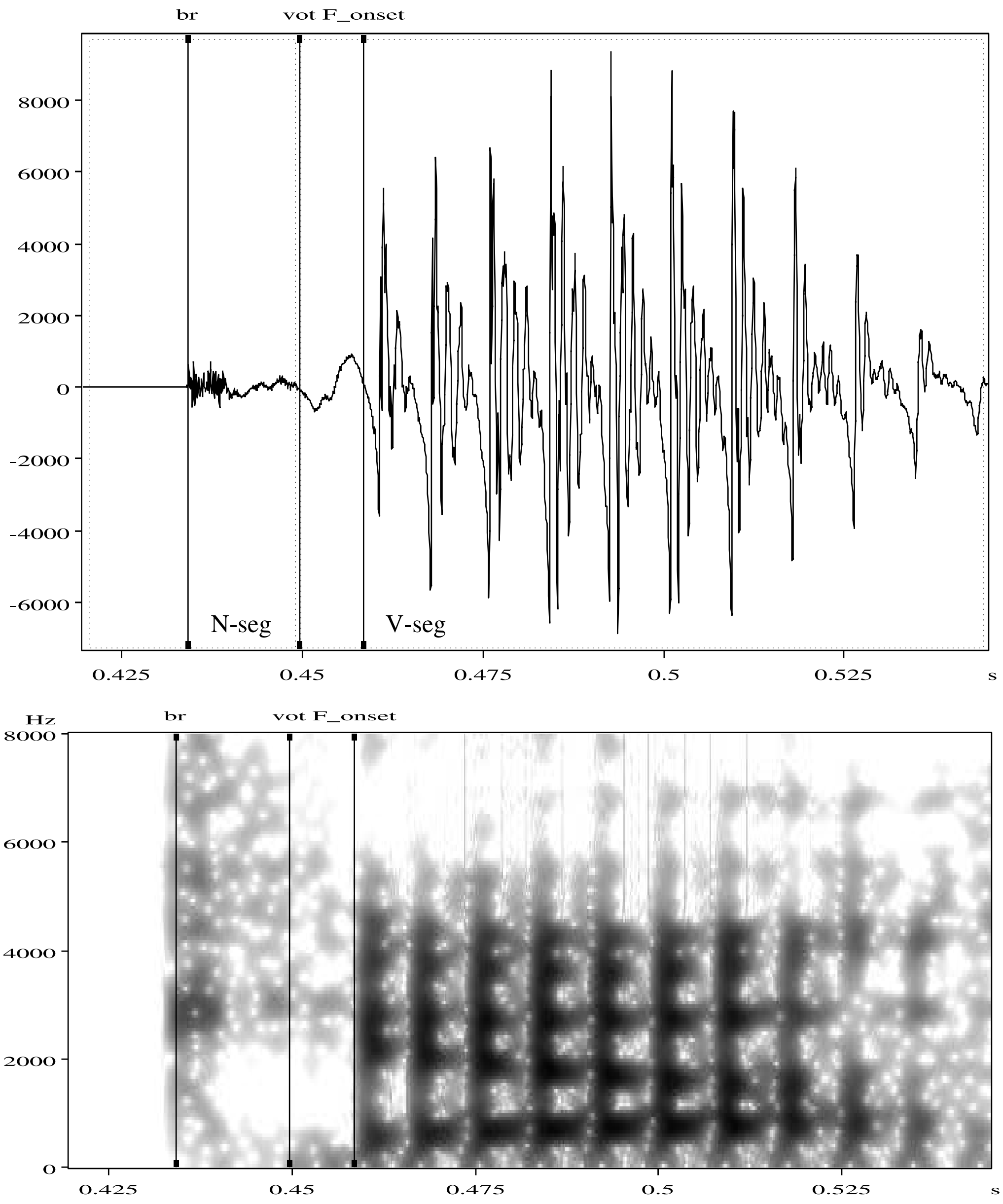


FIG. 4.1: Exemple de segmentation pour la syllabe [ka]. Les marqueurs "br" et "vot", mis sur l'oscilogramme, sont pour nous les lignes de démarcation du N-seg. Le marqueur "F_onset", mis sur spectrogramme, est utilisé par d'autre chercheurs [Kla75] (br=burst release, vot=voice onset time, F_onset=formant onset)

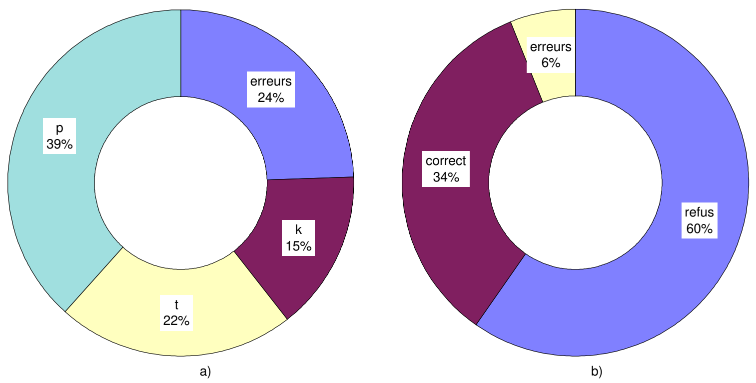


FIG. 4.2: Proportion des réponses pour les V-seg isolés b) toutes les réponses et a) distribution des réponses non rejetées.

– La syllabe [pi] a été donnée "sans consonne" 3 fois

Pour les V-seg isolés nous avons 1020 présentations (30 items x 34 auditeurs). Plus de la moitié (56,7%) de ces présentations ont été rejetées (classées comme "sans consonne ou consonne non identifiable"). Pour un peu plus d'un tiers (36,5%) de ces présentations, la consonne est correctement reconnue, malgré l'absence du N-seg. Il y a eu peu d'erreurs : 6,8% du total (voir figure 4.2). En fait, parmi les non rejetées, les bonnes identifications représentent une partie importante (84,5%). Il n'y a eu pratiquement qu'un seul type d'erreur : des consonnes [k] ou [t] ont été données pour [p] après suppression du N-seg.

Pour les 36,5% de réponses qui étaient en accord avec l'identité du V-seg, nous avons fait une statistique selon la consonnes et selon la voyelle. Donc, parmi ces réponses correctes, 47,1% étaient [p], 27,7% [t] et 25,2% [k]. Nous retrouvons un résultat plusieurs fois rapporté [StBC96a] : [p] est environ deux fois mieux reconnu et il n'y a pas de différence notable entre [k] et [t]. Pour les voyelles, nous observons un gradient (voir figure 4.3) où [a] arrive en premier (un cinquième des réponses correctes) et [i] arrive en bon dernier (une seule réponse correcte).

4.2.3 L'influence du contexte vocalique

Nous disposons de 2040 réponses pour les monstres conflictuels. Nous allons les classer en quatre catégories :

– réponses N quand la réponse indique la consonne qui a fourni le N-seg pour le monstre
– réponses V quand la réponse est imposée par la partie voisée du monstre
– réponses F (comme fusion) quand la réponse indique la troisième consonne donc ni celle du V-seg ni celle du N-seg.
– réponses R (comme rejetées) quand l'auditeur refuse de classifier en choisissant l'option "sans ou non identifiable".

Les résultats (voir figure 4.4 et tableau 4.2) nous montrent d'abord qu'il y a très peu de réponses rejetées (3,6%) dans le test 1 par rapport au test 0. Les monstres conflictuels paraissent donc, pour

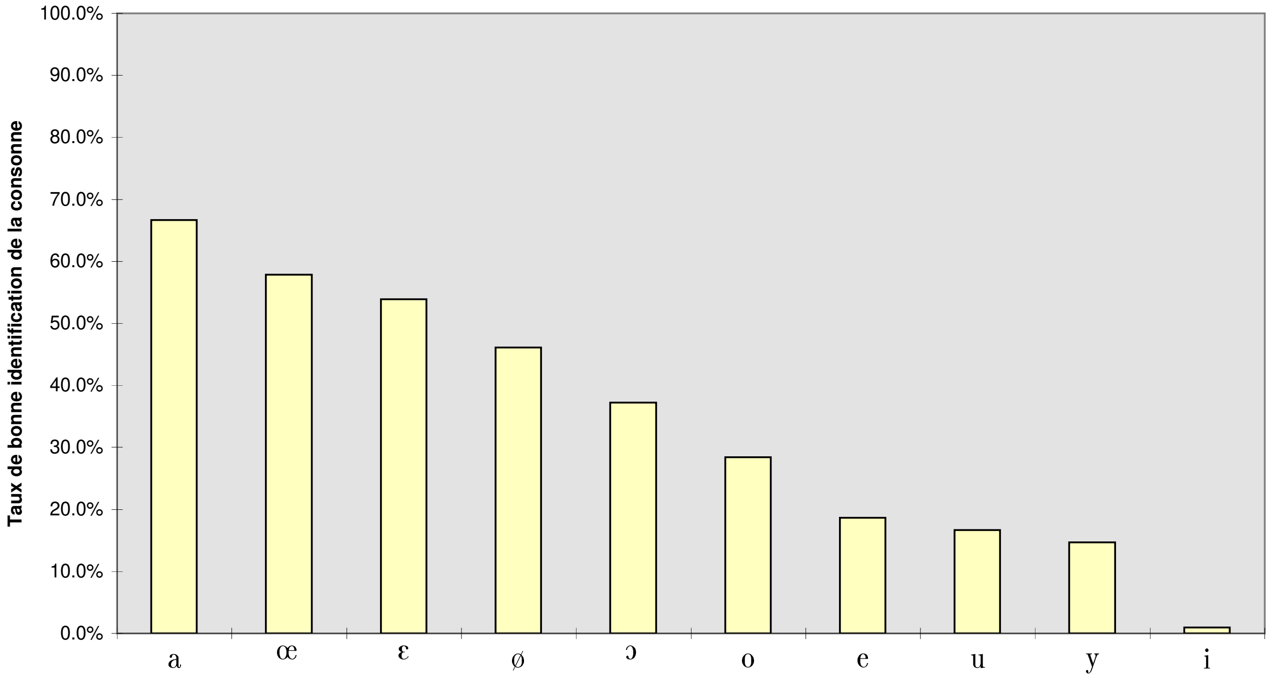


FIG. 4.3: Influence du contexte vocalique sur le taux d'identification de la consonne à partir du V-seg isolé des consonnes CV.

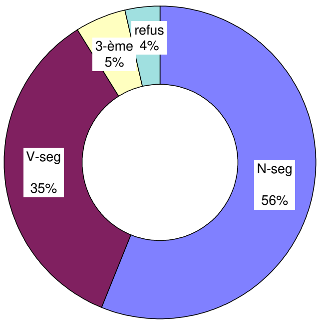


FIG. 4.4: Distribution des réponses pour les monstres conflictuels. Les quatre catégories sont celles indiquées dans le texte. Cette distribution nous montre les poids relatifs moyens des segments dans la perception des occlusives

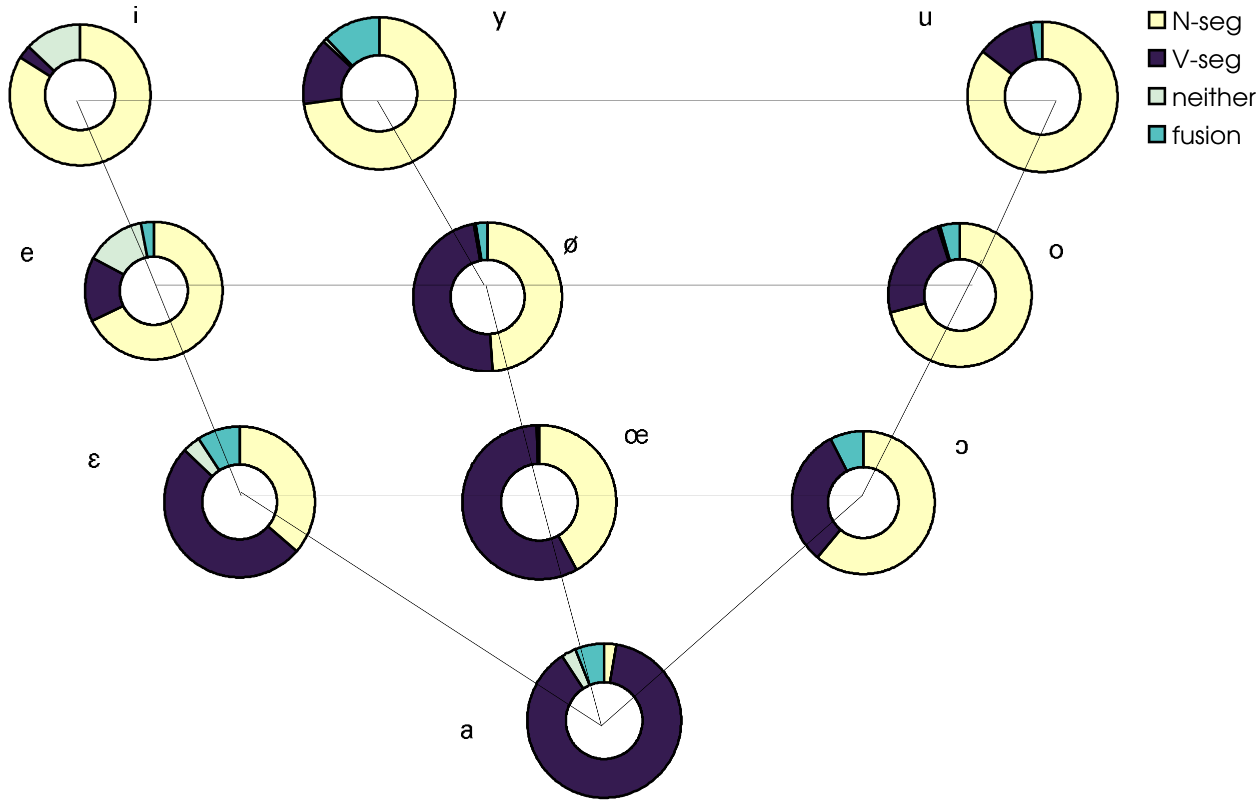


FIG. 4.5: Influence du contexte vocalique sur le poids relatif des segments. Même statistique que la figure précédente (4.4) mais séparément pour chaque voyelle

nos auditeurs, plus naturels que les items tronqués. Après quelques essais, nous avons constaté qu'ils ne sont pas capables de séparer les monstres conflictuels des items naturels mais qu'ils peuvent le faire généralement pour les V-seg isolés.

Les réponses N et V regroupent 91,2% des cas. Si globalement l'information du N-seg semble masquer celle du V-seg (56,1% par rapport au 35%), la proportion entre réponses N et V est très contrastée selon le contexte vocalique (voir figure 4.5). En effet, pour le contexte [a], la réponse est imposée presque toujours par l'identité du V-seg. Le contraire est valable dans le cas de voyelles fermées, [i], [y] et [u]. Ici la réponse suit le plus souvent l'identité du N-seg. En outre, un certain gradient est respecté sur chaque colonne du triangle vocalique : le N-seg est de plus en plus important selon le degré de fermeture de la voyelle. Les contextes protrus semblent plus influencés par le V-seg et les contextes arrières plus par le N-seg mais c'est un effet secondaire. La dépendance de l'ouverture est encore plus frappant si nous retenons que les monstres faisant intervenir [k] et [t] (voir tableau 4.1).

Nous sommes donc en présence d'un effet massif et systématique. Il nous montre que les statistiques globales (comme dans la figure 4.4) n'ont guère de signification. En effet, des taux variant de 0% à 100% peuvent être obtenus simplement en variant la proportion des contextes vocaliques dans les items présentés. Même s'il n'a jamais été mesuré sur autant de voyelles, nos recherches

voyelle	réponses N	réponses V	réponses F	réponses R
a	0.00%	100.00%	0.00%	0.00%
ɛ	25.00%	75.00%	0.00%	0.00%
œ	50.00%	48.53%	0.00%	1.47%
ɔ	64.71%	8.82%	25.00%	1.47%
ø	77.94%	22.06%	0.00%	0.00%
o	77.94%	2.94%	19.12%	0.00%
e	79.41%	20.59%	0.00%	0.00%
u	83.82%	1.47%	14.71%	0.00%
i	98.53%	1.47%	0.00%	0.00%
y	100.00%	0.00%	0.00%	0.00%

TAB. 4.1: Distribution des réponses pour les monstres conflictuels construits à partir des segments de [t] et [k] en fonction du contexte vocalique

Stimuli	Responses suggested by		third consonant responses	neither responses
	N-seg	V-seg		
N-seg = k	79.12%	15.88%	4.71%	0.29%
N-seg = p	34.26%	50.74%	5.15%	9.85%
N-seg = t	55.00%	38.53%	5.74%	0.74%
V-seg = k	46.18%	41.47%	6.62%	5.74%
V-seg = p	68.38%	26.32%	4.56%	0.74%
V-seg = t	53.82%	37.35%	4.41%	4.41%
overall	56.13%	35.05%	5.20%	3.63%

TAB. 4.2: Classification des réponses aux monstres conflictuels. Influence de la consonne.

bibliographiques montrent que cet effet a été observé déjà en 1972 par Eli Ficher-Jørgensen [FJ72]. Malgré l'ancienneté des premières observations, il a été malheureusement souvent ignoré. Ainsi, nous pouvons apprécier aujourd'hui que les résultats de Walley et Carrel [WC83] en faveur des formants comme indices dominants doivent beaucoup aux contextes qu'ils ont utilisés. De même, les résultats de Schouten et Pols [SP83a] en faveur du bruit comme indice dominant ont l'origine probablement dans la structure du système vocalique du danois (comme en français, plus de voyelles fermées que dans l'américain).

4.2.4 L'influence de la consonne

Le regroupement des réponses selon la consonne (voir le tableau 4.2) donne aussi des résultats contrastés. Nous trouvons un même ordre de prégnance perceptive parmi les consonnes (d'abord [k], après [t] et finalement [p]) aussi bien pour le N-seg que pour le V-seg (voir figure 4.6). Par exemple, la plupart des N-seg de type [k] ont imposé un percept [k] (colonne 1 ligne 1 du tableau 4.2) et les V-seg de type [k] ont mieux résisté au masquage par les N-seg (colonne 2 ligne 4 du tableau 4.2). Ce même ordre de préférence est confirmé par d'autres travaux sur des N-seg isolés du français [Dje95a]. Par

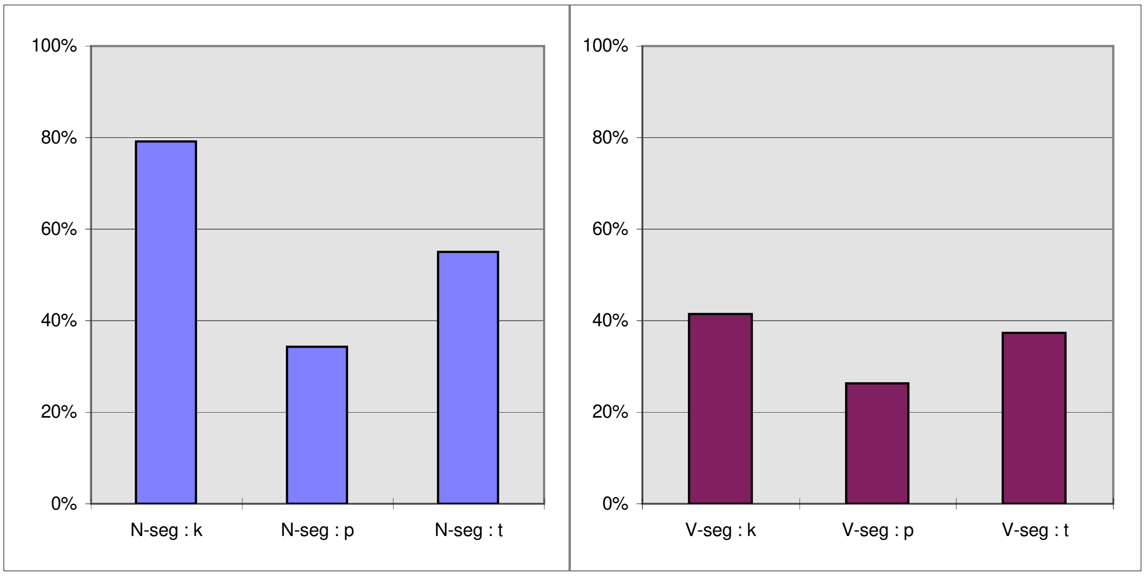


FIG. 4.6: Proportion des N-seg ayant masqué le V-seg (à gauche) et proportion des V-seg ayant masqué N-seg (à droite) selon l'identité consonantique du segment

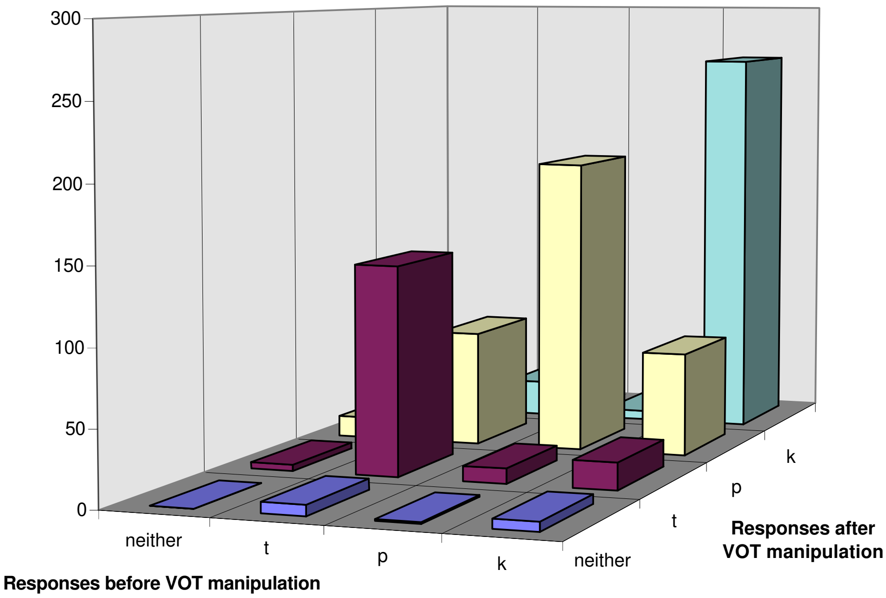
contre, les V-seg isolées ne confirment pas cet ordre. Nous reviendrons sur cette contradiction.

Une façon simplifiée de regarder les réponses est de considérer seulement la réponse majoritaire pour chaque stimuli (nous les appellerons réponses WTAP pour "winner take all policy"). Une liste des réponses WTAP pour les 60 monstres conflictuels avec le poids de cette réponse majoritaire est donnée en annexe.

Nous trouvons peu de cas où la concaténation du N-seg d'une consonne avec le V-seg d'une autre donne l'illusion d'une troisième consonne (5,2% du total). Concernant les réponses WTAP, seulement le N-seg de [kɛ] devant le V-seg de [pɛ] et le N-seg de [py] devant le V-seg de [ty] ont été perçus comme la troisième consonne. Deux monstres conflictuels sur 60 ont été majoritairement rejetés. Les deux sont construit avec le N-seg de [p] pris du contexte vocalique [i] ou [e]. Une vérification informelle pour plusieurs auditeurs nous a montré que ces deux items sont perçus comme "sans occlusive" et non pas comme "consonne non identifiable".

4.2.5 L'influence des indices acoustiques du N-seg

Une possible explication de la distribution de l'importance perceptive du N-seg est la variance de ses caractéristiques selon la voyelle qui précède. En effet, plus la voyelle est fermées, plus le N-seg est long, comme nous indique la statistique détaillée sur le VOT décrite dans la section dédiée aux indices du N-seg dans le chapitre suivant. En plus, plus la voyelle est ouverte, plus son énergie est importante, ce qui donne une amplitude relative plus grande du N-seg pour les contextes fermés. En somme, plus le N-seg est long et intense, plus forte est son influence sur la perception. Nous avons voulu cerner l'influence de ces deux indices temporels sur la distribution observée des réponses.



		Responses in Experiment III			
		/t/	/p/	/k/	neither
Responses	/t/	57.2%	31.3%	09.8%	01.7%
in	/p/	04.7%	92.0%	02.8%	00.5%
Experiment I	/k/	05.0%	19.2%	74.0%	01.8%

FIG. 4.7: Tableau de contingence montrant le changement de la distribution des réponses après modification du VOT (test 3 par rapport à test 1)

4.2.5.1 L'influence du VOT

La durée du segment non voisé (N-seg), tel que nous l'avons extrait, correspond bien avec le "voice onset time" (VOT) défini par Lisker [SB86]. Afin d'étudier son influence, nous avons modifié la durée du N-seg dans chacun des 60 monstres conflictuels du test 1. Dans les nouveaux monstres obtenus, le V-seg et le VOT suggère une consonne et seul le contenu spectral du N-seg suggère un autre consonne. Pour la manipulation, nous avons utilisé la technique TD-PSOLA (time domain pitch synchronous overlap and add) sur le résidu de l'analyse LPC [BBW92]. Les coefficients LPC ont été interpolés linéairement avant la resynthèse afin de reproduire au mieux les transitions spectrales observées dans le N-seg. Cependant, il n'a pas été possible de ramener tous les N-seg de type [p] au VOT désiré en conservant un spectrogramme identique à l'original à une anamorphose temporelle près. Ainsi nous avons choisi de remplacer 5 des 20 monstres au N-seg de type [p] avec des items de contrôle pour ce test.

Nous allons classer les réponses données à ce test en les comparant avec les réponses données par les mêmes auditeurs aux items correspondants dans le test 1 (voir figure 4.7). Les résultats nous montrent que :

- Le VOT n'est pas la raison principale de la distribution observée à l'issue du test 1. En effet, parmi les 55 monstres de ce test, seulement 16 donnent une réponse WTAP différente par rapport au test 1.

– Le VOT porte de l'information sur le lieu d'articulation en français. En effet, pour 12 sur 16 changements de perception majoritaire mentionnés plus haut, la consonne perçue a été celle suggéré par le VOT.

– Les monstres ayant un VOT de [p] ont tendance à être perçus comme des [p], quelque soit le contenu spectral de leur N-seg. C'est le cas pour 10 changements sur les 16 observés (voir figure 4.7).

Nous concluons que la durée du N-seg est un indice plus important que les indices spectraux pour les contrastes labial versus autre lieu d'articulation. Par contre, le contraste [t]/[k] est peu sensible à la manipulation de cet indice. Nous en déduisons qu'il doit être basé sur des indices spectraux.

4.2.5.2 L'influence de l'amplitude relative

Nous nous sommes limités pour ce test aux monstres construits à partir des segments de type [t] et [k]. Parmi ces 20 monstres, 15 ont reçu une réponse WTAP favorisant l'identité du N-seg. Nous avons diminué l'amplitude du N-seg de ces 15 monstres en trois étapes (-6dB, -12dB, -18dB) pour rendre ainsi le V-seg plus prégnant pour la perception de nos auditeurs. Les 5 autres monstres ont été perçus comme leur V-seg donc un rehaussement de leur N-seg a été effectué (+6dB, +12dB, +18dB) afin d'aider la consonne d'origine du N-seg emporter sur la perception. Nous avons donc 15*3 + 5*3 = 60 monstres modifiés dans ce test.

Le regroupement des réponses selon les catégories N, V, F et R est montré dans sur la figure 4.8 séparément pour les diminutions et amplifications du N-seg (pour ce test, réponse F = réponse [p]). Mais il nous semble plus intéressant de classer les réponses, ici aussi, selon leur différence ou ressemblance avec ceux qui leur correspondent dans le test 1. Les résultats sont regroupés dans le tableau de contingence 4.3.

– Trois quarts des réponses (76,8%) ne sont pas affectées par cette modification de l'amplitude relative.

– 11,1% des réponses ont changé entre les conditions 0dB et 6dB et n'ont plus changé après . Ces changements ont conduit à 6 réponses majoritaires (sur 20) différentes par rapport au test 1.

– 7,5% des réponses ont changé quelque part entre la condition 6dB et la condition 18dB

– 4,6% des réponses ont changé de manière incohérente ou non attendue et peuvent être raisonnablement considérées comme erreurs de perception.

Malgré ces deux dernières catégories (7,5 + 4,6 = 12,1%), il n'y a jamais eu un changement de réponse majoritaire entre la condition 6dB et les manipulations plus poussées. Nous pensons donc que seule la modification de 0 à 6 dB a changé de manière statistiquement significative les réponses.

Concernant les six changements de réponse WTAP observés, nous pouvons les classer en deux catégories :

– Les trois items avec un N-seg de [t] et un V-seg de [k] devant les voyelles arrières [u o ɔ] ont été perçus comme [t] dans le test 1. Après diminution du N-seg ils ont été perçus [p] (plus la

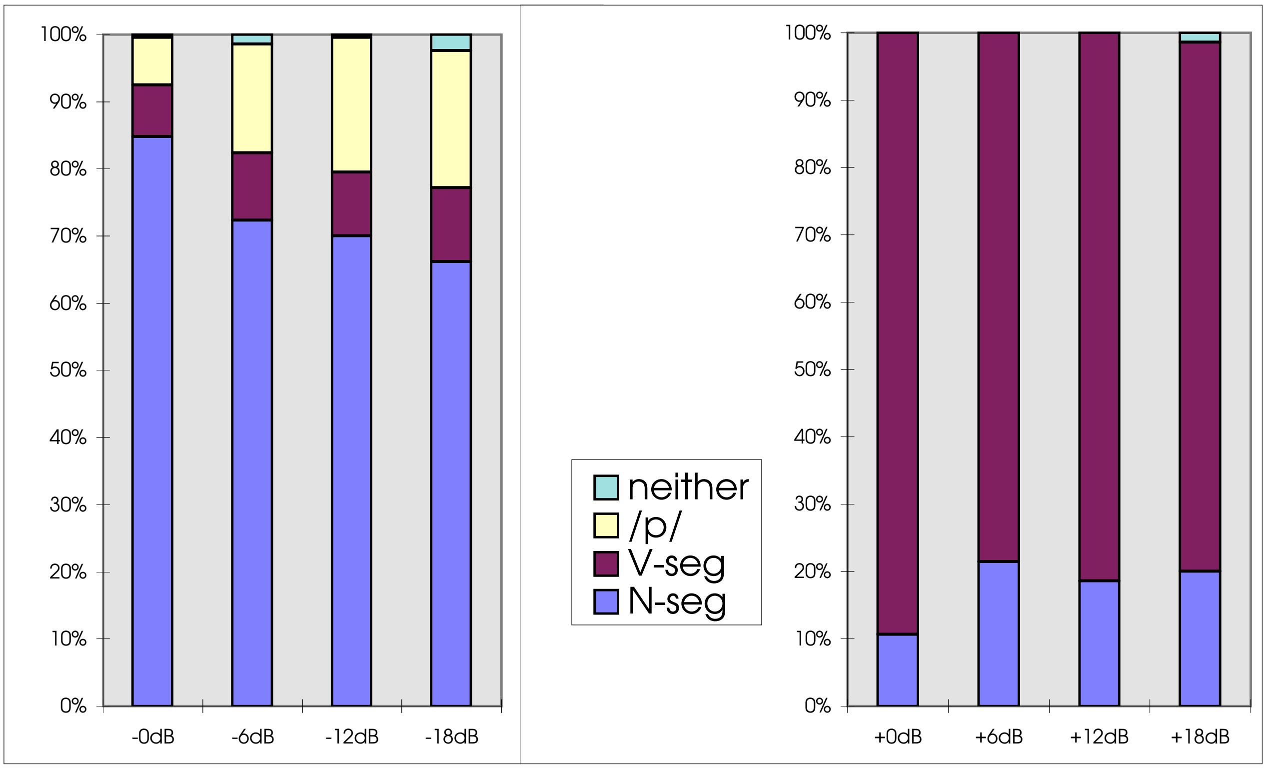


FIG. 4.8: Évolution de la proportion des réponses pour les monstres combinant [t] et [k] selon l'atténuation (à gauche) ou l'amplification (à droite) du N-seg. Ici, tous les réponses de type F sont, en fait, des réponses [p]

diminution est grande, plus nombreux sont les auditeurs à entendre [p]). Celui ci peut être un effet de la perception catégorielle, comme le biais en faveur du [p] dans le test 0.

– Pour trois items, nous avons effectivement réussi à changer la balance perceptive entre le N-seg et le V-seg. Le monstre N-seg [kɛ] suivi de V-seg [tɛ], perçu comme [t] dans le test 1, devient [k] après amplification du N-seg. Les monstres N-seg [te] suivi de V-seg [ke] et N-seg [tø] suivi de V-seg [kø], perçus [t] au départ, deviennent [k] après diminution du N-seg.

Nous remarquons que les changement dans le contraste [t]/[k] apparaissent seulement dans le contexte des voyelles [e ɛ ø], c'est à dire là où l'importance perceptive des deux segments semble s'équilibrer, jamais dans le contexte des voyelles extrêmes. En fait, nous constatons que, dans la plupart des cas, les items perdent leur naturalité (ou deviennent des [p], comme les items tronqués) avant que l'identité du segment masqué dans le test 1 soit révélé à la perception de nos auditeurs par manipulation d'amplitude. Nous concluons donc que, au moins pour les voyelles extrêmes du triangle vocalique, le contraste [t]/[k] est robuste par rapport aux variations dans l'amplitude relative des segments.

4.2.6 L'influence de l'auditeur

Nos auditeurs ont montré différents niveaux de préférence pour les indices de la partie sourde ou voisé des monstres (voir figure 4.9). Malgré le fait que l'auditeur est un facteur significatif dans

Responses in Exp. 2			Responses in Exp.1 (0dB)			
+/-6dB	+/-12dB	+/-18dB	neither	k (172)	p (23)	t (84)
k	k	k		155		14
p	p	p	1		21	14
t	t	t		2		39
k	k	neither		2		
k	t	t		1		
k	k	t		6		
t	t	neither				2
t	t	k				4
t	p	p				4
t	k	k				2
neither	neither	neither				1
neither	k	neither		1		
neither	k	k		1		
k	p	k				1
k	t	k		2		1
t	k	t		1		1
t	t	p			1	
t	p	p		1	1	
k	k	t				1

TAB. 4.3: Évolution des réponses lors de la modification de l'amplitude du N-seg. Les lignes 1 à 3 regroupent les réponses qui ne change plus après la conditions 6dB (76,8%). Les lignes 4 à 10 regroupent celles qui changent entre 6dB et 18dB (7,5%). Les lignes 11 à 19 regroupent celles qui changent de manière incohérente ou inattendue (4,6%)

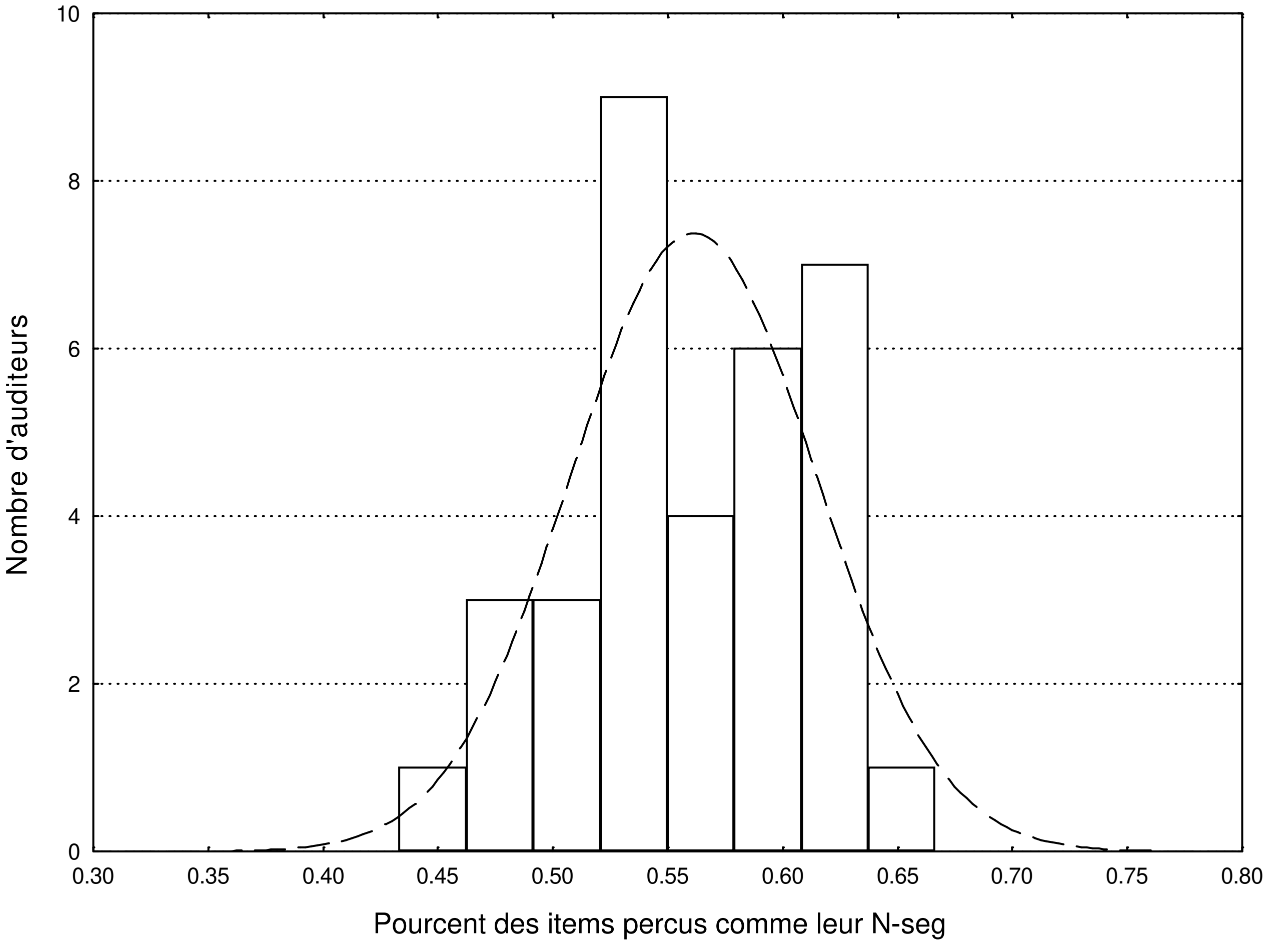


FIG. 4.9: Histogramme de auditeurs selon leur pourcent des réponses en faveur du N-seg pour les monstres conflictuels

la distribution des résultats, nous n'avons pas remarqué de groupement des auditeurs avec un comportement spécifique dans les test 0 et 1. Tout au plus, une légère différence peut être notée entre les auditeurs avec une expérience dans le traitement de la parole et ceux complètement naïfs. Les premiers sont plus exigeants, refusent plus facilement la classification (maximum 4 monstres conflictuels sur 60) par rapport aux derniers qui, assez souvent, ne rejettent aucun monstre conflictuel. Cette relative homogénéité des auditeurs, confirmée par d'autres travaux [StBC96a], justifie les statistiques à travers les auditeurs.

Pour les test 2 et 3, nous espérions tirer profit de la réponse "sûr/pas sûr" démandée en plus de la classification phonétique. Nous avons calculé le degré de confiance de chaque auditeur pour chaque test comme étant le pourcent des réponses données pour sûres (items de contrôle mis à part). Le niveau moyen de confiance est élevé et un peu plus faible pour le test qui a nécessité le traitement par PSOLA :

- test 2 : moyenne 84,9%, écart type 10%
- test 3 : moyenne 81,6%, écart type 16%

Mais, comme nous pouvons l'observer sur le tableau 4.4, nos auditeurs présentent des stratégies

Auditeur	Taux des réponses "pas sûr"	
	Test niveau	Test VOT
YP	1.67%	10.34%
MR	6.67%	0.00%
LA	6.67%	5.17%
PB	8.33%	13.79%
PA	10.00%	3.45%
NA	10.00%	17.24%
FS	11.67%	10.34%
PT	15.00%	13.79%
LV	15.00%	10.34%
JZ	15.00%	25.86%
GB	16.67%	13.79%
ET	16.67%	17.24%
HL	20.00%	27.59%
FG	33.33%	60.34%
CS	40.00%	46.55%
moyenne	15.11%	18.39%

TAB. 4.4: Proportion des réponses qualifiées "pas sûr" selon l'auditeur.

extrêmement différentes pour l'utilisation du choix "sûr/pas sûr". Ainsi, le niveau de confiance varie entre 40% et 100%. Les moins confiants ont classé "pas sûr" même plusieurs items à 100% naturels, items qu'ils identifient d'ailleurs toujours correctement, comme les autres auditeurs. Nous pensons que cette situation trahit un manque de rigueur lors du passage de la consigne, ce qui nous empêche d'interpréter plus loin cet aspect.

4.2.7 Le paradoxe du contexte avant

Dans les études comparant N-seg et V-seg, le cas des contextes des voyelles avant non-arrondies est particulier. Nos résultats sur les monstres conflictuels, comme ceux de Smits et al, montrent que le N-seg masque facilement l'identité du V-seg. De plus, les résultats sur les V-seg isolés place fermement le contexte [i] en dernière position comme porteur d'information sur le lieu d'articulation. En fait, les V-seg du contexte avant sont perçus comme "sans aucune consonne". Nous devrons en tirer la conclusion que, comme pour les autres contextes fermés, le N-seg est le support de l'informations discriminante dans le cas des voyelles avant.

Cependant, les tests avec N-seg isolés [BDL93, BDL96] et les tests de "gating" sur CV de l'anglais [KPPSK83] trouvent ces N-seg intrinsèquement peu discriminants. Les N-seg du contexte [i] est aussi (peu) performants que ceux du [a] et loin derrière ceux de [u], surtout pour le contraste [k] versus [t].

Il semble donc que, pour la perception humaine, les deux segments ne sont pas très prégnants. L'analyse acoustique, décrite dans le chapitre suivant, confirme largement cet état : la majorité des erreurs du classifieur automatique proviennent de ce contexte vocalique. En outre, les tentatives de prédictions des réponses des auditeurs se heurtent aussi à un obstacle :

> " ...for the vowel context /i/, the model is apparently not capable of reproducing the observed response behaviour. Calculation [...] shows that the fit for the vowel context /i/ is by far the worst (64% goodness of fit for 58% chance level)." [StBC96b, page 3977]

La solution à cet apparent manque de discrimination nous est suggérée par une ré-interprétation des résultats des tests perceptifs sur signaux tronqués publiés en 1983 par Kewley-Port et collaborateurs [KPPSK83, page 1783]. Elle observe qu'une durée plus grande des signaux est nécessaire afin d'éviter les confusions perceptives de /gi/ vers /di/. Vu le VOT de ces syllabes anglaises, il s'agit bien de la même confusion que celle rapportée au CRIN pour le français : /ki/ est confondu souvent avec /ti/. Le fait que ces confusions s'amenuisent de manière très significative en passant de 20 ms à 40 ms nous démontre que la présence physique du contexte vocalique joue un rôle crucial dans ce cas. Nous allons démontrer dans le chapitre suivant que le contexte vocalique physiquement présent, en facilitant l'analyse morphologique, est susceptible d'apporter une information beaucoup plus pertinente pour la perception par rapport à la connaissance a priori du contexte par les auditeurs, connaissance apportée d'une manière découplée de la modalité auditive.

L'asymétrie des confusions entre [ti] et [ki] est confirmée non seulement pour le français [Dje95a] mais également pour l'italien [DPRV95] et pour l'espagnol [PDO97]. Cependant, elle ne parait pas dans les études du Smits et collaborateurs sur le danois. Comme des confusions asymétriques semblables sont observables aussi dans la vision ("Q" confondu avec "O" beaucoup plus souvent que l'inverse), nous pouvons utiliser le cadre théorique plus général sur la perception humaine présenté par W. R. Garner dans [Gar78] :

> "This perceptual entropy can be explained as follows : if sound X possesses a cue that Y lacks, listeners are more likely to miss this all-in-one cue than to introduce it spuriously."

Nous allons revenir sur ces confusions dans le chapitre dédié à l'analyse acoustique. Nous montrerons quel est l'indice acoustique dans le spectre du bruit de relâchement du [ki] qui est susceptible d'être "raté" par les auditeurs pour arriver à un [ti].

Il conviennent de rappeler ici la répartition des syllabes /ti/ et /ki/ dans le français. Nos propres statistiques[1] sur le Grand Robert [Lir90] nous montrent (voir tableau 4.5) que les /ti/ sont beaucoup plus fréquents dans les mots du vocabulaire que les /ki/. En plus, cette asymétrie est spécifique au contexte antérieur, elle disparaît ou s'inverse pour les contextes /a/ et /u/. Le rapport entre l'asymétrie trouvée dans les fréquences d'occurrence dans le vocabulaire et l'asymétrie des confusions perceptives demande une étude plus approfondie : chacune peut être l'effet ou bien la cause de l'autre.

[1] Cette statistique est à prendre avec précaution car, pour la simplicité de la réalisation, nous avons compté toutes les entrées du dictionnaire. Par exemple, le préfixe /bi/ est compté autant de fois que des racines qu'il précède. Il aurait fallu tenir compte des compositions morphologiques et ne compter que les lemmes.

	[a]	[e]	[ø]	[o]	[i]	[y]	[u]	[ɛ]	[œ]	[ɔ]
[p]	3618	1781	107	652	2502	713	403	1581	138	2349
[t]	**5026**	5931	542	1566	**9177**	1712	**530**	2605	1662	3092
[k]	**5324**	692	128	1026	**945**	1790	**812**	470	165	3822
[b]	2061	765	45	289	2263	770	1045	664	74	1049
[d]	**1321**	5075	189	451	**3378**	743	**272**	632	184	1096
[g]	**1893**	214	58	341	**242**	367	**359**	173	112	931

TAB. 4.5: Fréquence d'occurrence de syllabes occlusive-voyelle dans l'inventaire lexical du français (d'après le Grand Robert). Il faut noter la forte asymétrie entre /ti/ et /ki/, de même qu'entre /di/ et /gi/.

4.2.8 Le paradoxe du V-seg de [p]

Nous avons déjà mentionné la contradiction entre le test 0 et le test 1 concernant la prégnance du V-seg du [p]. D'un coté, nous trouvons une forte préférence pour les réponses [p] dans le test 0, préférence confirmée dans la littérature (voir [StBC96a, page 3856]). [p] est deux fois mieux reconnu que [t] ou [k]. De l'autre coté, nous constatons une très faible prégnance du V-seg (et du N-seg) de [p]. Seuls un quart des V-seg de [p] n'ont pas été masqués. La solution de cette contradiction passe par l'identification de deux type de biais qui favorisent artificiellement les réponses [p] dans le test 0 :

- Le biais décisionnel. Le N-seg de [p] est habituellement bref et faible. En enlevant le N-seg nous rapprochons un stimuli du prototype du [p] et nous l'éloignons des deux autres.
- Le biais post-perceptif. Il s'agit de refus de la classification habituelle par la contestation des sons perçus par des niveaux supérieurs de traitement. Ces contestations consistent dans un raisonnement comme "Ce que j'ai bien entendu sonne tellement comme un [p] que j'ai du mal entendre ce qui n'y ressemblait pas".

Pour éviter le premier type de biais, il aurait fallu utiliser la méthode décrite par Schouten et Pols [PS78]. En précédant les V-seg d'une rafale de bruit blanc (la même pour tous les V-seg), ils ont observé une réduction nette du biais en faveur du [p]. Pour éviter le deuxième type de biais, nos auditeurs ont reçu l'instruction de rejeter les items sur lesquels "ils se posent trop de questions". Malheureusement, une autre moyen, comme par exemple un rythme imposé dans le test, aurait été plus efficace.

Il est important de noter que les tests à partir des N-seg isolés montrent un biais en défaveur du [p], biais complémentaire celui observée pour les V-seg isolés. Ainsi Djezzar [Dje95a] trouve les taux d'identification suivants : 95% pour [k], 91% pour [t] et seulement 62% pour [p] à partir de N-seg isolés. Ce comportement du [p] est confirmé pour l'anglais américain. Ainsi, en manipulant l'amplitude et le spectre du N-seg dans un continuum labial-alveolaire, Ohde et collaborateurs [OS83] trouvent que les réponses [p] apparaissent nettement plus tôt quand l'amplitude du N-seg est abaissé.

Si dans la perception humaine, la limite entre les deux types de biais est assez floue, il convient de les séparer pour les classifieurs automatiques :

– Le biais décisionnel est reproduit, même trop bien, dans les systèmes automatiques. Plusieurs travaux, dont le nôtre, trouvent pour [p] le meilleur taux global d'identification automatique parmi les trois phonèmes. Différents espaces de paramètres acoustiques, décrits dans le chapitre suivant, montrent que les réalisations de [p] sont les plus faciles à séparer. Cependant, il faut noter que nous n'avons pas de syllabes formées d'une simple voyelle dans notre corpus. Le silence étant le concurrent acoustiquement le plus proche de [p], un classifieur traitant toutes les quatre possibilités ne devrait plus montrer de préférence particulière pour la classe des labiales.

– Le biais post-perceptif, en toute logique, n'est pas reproduit dans nos décodages automatiques. Nous trouvons un exemple parlant dans le test de manipulation de niveau du N-seg. Les monstres construits du N-seg de [t] et V-seg de [k] dans le contexte des trois voyelles arrières sont perçus comme [t] par nos auditeurs et reconnus comme [t] par le classifieur avant manipulation. Après atténuation du N-seg, les auditeurs pour la plupart les donnent pour [p] tandis que le système automatique continue à les classer comme [t].

4.2.9 Conclusions sur la perception des occlusives

Nous avons montré dans ce chapitre que la dépendance de la voyelle est aussi importante, sinon plus, que la dépendance de la consonne pour la prégnance des indices acoustique de lieu d'articulation. Nous avons identifié et analysé trois types de balance perceptive entre le N-seg et le V-seg :

– Pour le contexte /a/, le V-seg est très discriminant et le N-seg ne l'est pas.

– Pour le contexte /u/, le V-seg est discriminant mais le N-seg l'est beaucoup plus.

– Pour le contexte /i/, le N-seg est à peine discriminant et le V-seg ne l'est pas du tout.

Nous trouvons que le comportement des CV avec voyelles intermédiaires peut être déduit avec suffisamment de précision par l'interpolation de ces trois cas extrêmes.

Nous sommes les premiers à isoler de cette manière le problème d'identification du lieu d'articulation posé par les syllabes avec voyelles avant non-arrondies. Pour les contextes /a/ et /u/, les résultats de perception peuvent s'expliquer par un simple **mécanisme de compétition**. La partie la plus discriminante emporte sur la perception. Par contre, pour le contexte /i/, il s'agit plus d'un **mécanisme de collaboration**. La jointure et seulement la jointure du N-seg et du V-seg garde la capacité de discrimination nécessaire. L'analyse acoustique ajoutera davantage d'arguments à cette affirmation.

Par conséquent, un décodage performant doit utiliser soit un jeu de paramètres pour chaque type de contexte vocalique soit un jeu unique insensible à la nature du contexte. Les résonances du conduit vocal nous semble le cadre le plus indiqué pour la recherche de ce jeu unique de paramètres.

4.3 Des voyelles vers les consonnes occlusives

Il existe peu d'études concernant directement les modèles de la production des consonnes occlusives. Les résultats que nous avons aujourd'hui proviennent des études d'aérodynamique [PJ96, PLT$^+$94, Mae86] et, surtout, des recherches en synthèse articulatoire de parole. Il s'agit de simulations qui mettent en oeuvre les modèles développés pour la production des voyelles. Dans cette partie, nous présentons brièvement quelques résultats d'expérimentation rapportés dans la littérature, résultats qui plaident en faveur de l'utilisation des résonances du conduit pour la caractérisation des occlusives.

4.3.1 Consonnes occlusives et points focaux

Nous examinons ici les fibres articulatoires produites en sélectionnant dans l'espace articulatoire du modèle de Maeda [Mae79, Mae88] les configurations produisant une constriction totale et unique sur un des points d'intersection de la grille semi-polaire avec le contour sagittal externe [Bai93]. Les figures 4.10 et 4.11 montrent les caractéristiques acoustiques et géométriques essentielles d'une telle fibre pour les lignes 23 et 21 de la grille, soit respectivement situées à environ 5 et 8 cm des dents. La superposition de l'ensemble des contours sagittaux obtenus montrent à l'évidence que la forme de la langue est fortement contrainte alors que les lèvres sont libres de coarticuler avec le contexte - et donc d'anticiper la voyelle suivante dans des stimuli CV. On peut donc s'attendre à ce que la fonction de transfert des cavités arrières de chaque fibre soit relativement stable alors que celle de la cavité avant varie fortement avec l'aire aux lèvres (Al).

Pour l'ensemble des configurations articulatoires de chaque fibre, le graphique reliant les maxima des fonctions de tranfert en fonction de l'aire aux lèvres (Al) montre effectivement que les résonances se répartissent clairement dans deux catégories :

Résonances affiliées à la cavité arrière : R1, R3 et R5. Comme la superposition des coupes sagitales le laissait entrevoir, ces résonances sont relativement fixes, centrées respectivement autour de 200, 2200 et 3500 Hz.

Résonances affiliées à la cavité avant : au contraire de la famille de résonances précédente, la première résonance de la cavité avant, R2, varie beaucoup avec Al (elle passe d'une résonance Helmholtz à une résonance quart d'onde). La variation du R2 est grossièrement logarithmique.

Dans l'articulation palattale présentée en 4.10, cette courbe coupe le lieu du R3 pour $Al \approx 2cm$ en ce que nous pourrons appeler un point focal. Notons que l'articulation palato-vélaire simulée en 4.11 ne comporte pas de point focal.

4.3.2 Reinterprétation des équations du locus

L'existence de changement d'affiliation pour F2 mis en évidence dans cette simulation fournit une possible explication aux phénomènes observés sur les équations du locus des consonnes /k g/. De nombreuses expériences montrent sur l'anglais que, pour ces consonnes, le nuage de points F2C

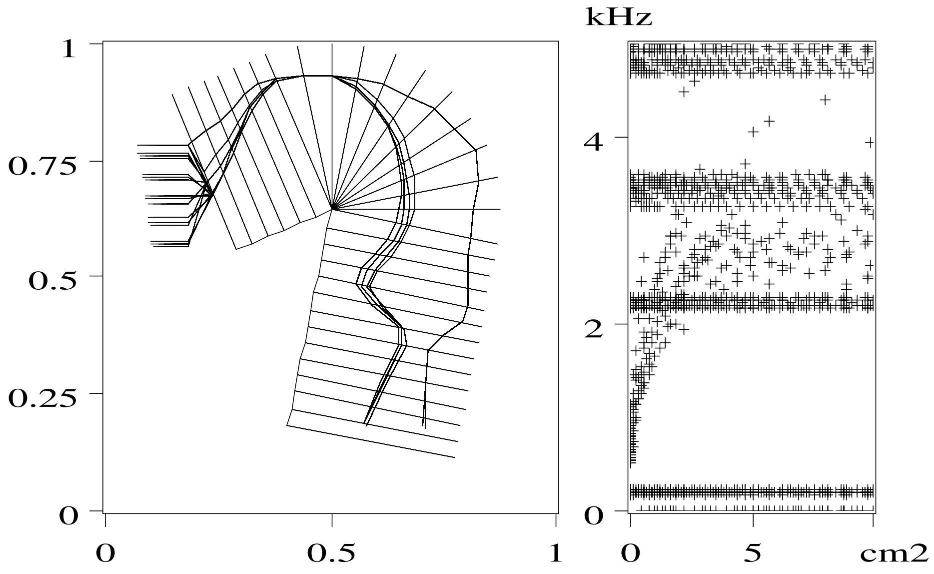


FIG. 4.10: Caractéristiques géométriques et acoustiques de la fibre articulatoire d'une occlusive palatale produite par le modèle de Maeda. La fréquence des résonances des configuration du conduit obtenues sont représentées en fonction de Al (à droite). Le conduit en coupe sagitale est représenté à gauche (une dizaine de coupes superposées seulement, par rapport aux 363 configurations articulatoires représentées à droite)

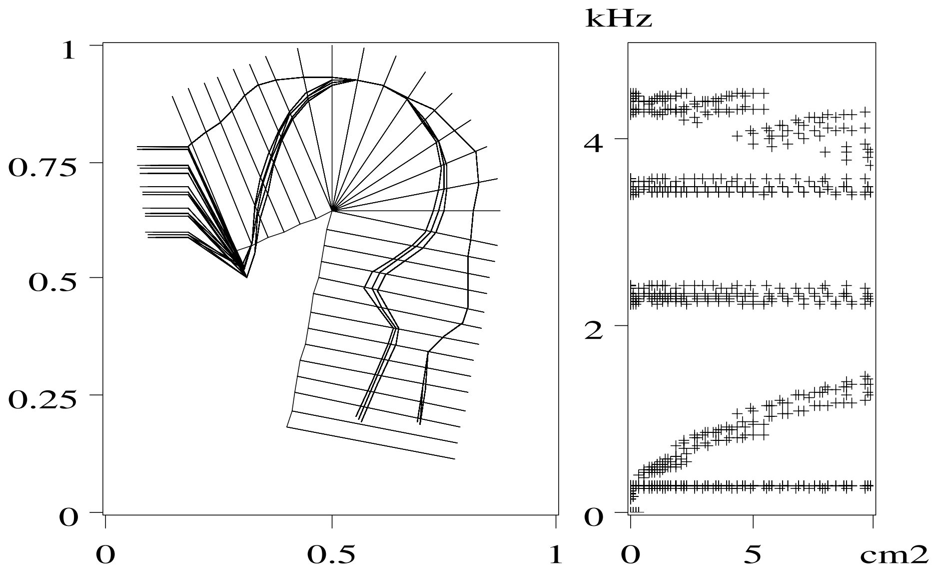


FIG. 4.11: Caractéristiques d'une occlusive palato-vélaire. Même représentation que la figure précédente 4.10.

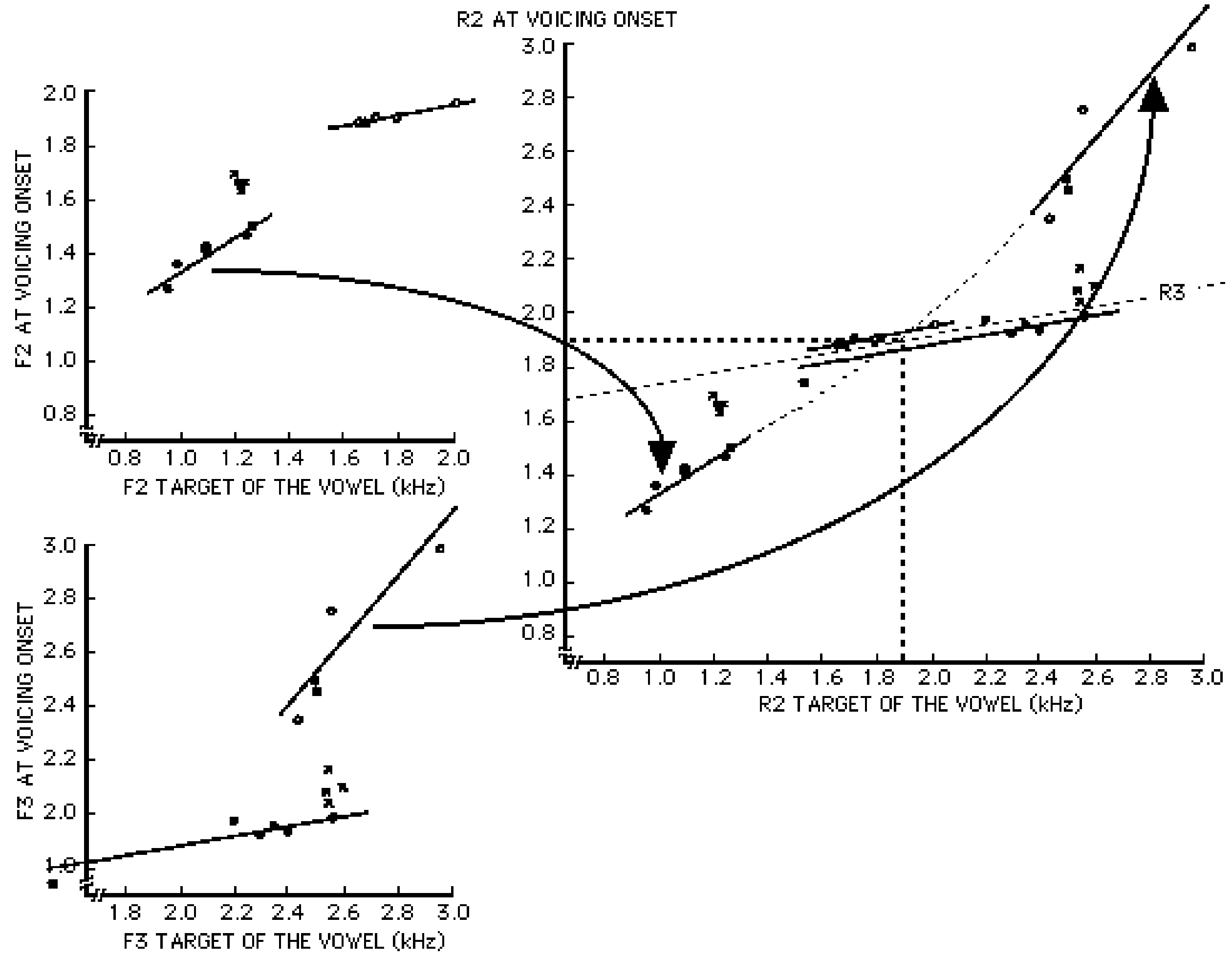


FIG. 4.12: Reinterprétation des équations du locus. Les deux comportements distincts de la consonne /g/ sont interprétés comme une simple conséquence d'un changement d'affiliation des formants.

versus F2V[2] est mieux représenté par deux droites de régression que par une seule droite comme pour les autres occlusives [SMM91, page 1315]. Une explication proposée est la variation du point de contact de la langue. On différencie ainsi un /g/ palatal dans un contexte antérieur et un /g/ velaire dans un contexte postérieur.

À la lumière de la simulation précédente, nous pouvons supposer que les deux droites correspondent en fait à deux affiliation différentes de F2C et F2V. Nous allons montrer une conformation de cette hypothèse sur des mesures publiées par Klatt en 1987 [Kla79, Kla82]. Pour ce faire, nous remarquerons que le plan F3C versus F3V pour la consonne /g/ présente aussi une rupture qui suggère deux droites de régression. Ainsi les plans F2C, F2V et F3C, F3V sont complémentaires. Leur superposition (voir figure 4.12) montre, sur ces données, une remarquable correspondance entre F2 des voyelles avant (/i e/) et F3 des voyelles arrière (/u o/). Une représentation R2C versus R2V (ou R3C versus R3V) montrerait une meilleure corrélation que les plans F2C, F2V ou F3C, F3V pour ces mesures.

[2]F2C est F2 mesuré au début et F2V est F2 mesuré à la fin du V-seg

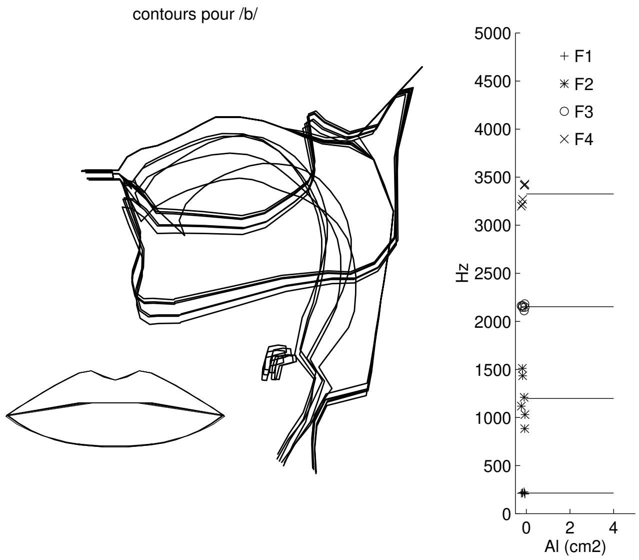


FIG. 4.13: Contours des articulateurs pour /b/ extraits à partir des données radiographiques. 6 réalisations du locuteur PB. Les formants mesurés sur la fonction d'aire sont représentés à droite en fonction du Al.

4.3.3 Données radiographiques

Finalement, nous allons présenter des profils sagitals obtenus par la mise en correspondance ("matching") des radiographies d'un sujet avec un modèle articulatoire [BGB+95, BBB+96]. Ces coupes représentent la position des articulateurs référencée au palais dur pour six réalisations de chaque occlusive voisée (/b d g/) par un locuteur français dans plusieurs contextes vocaliques.

Nous remarquons (voir figure 4.13, 4.14 et 4.15) la stabilité du point de contact (moins forte pour /g/) et la coarticulation pour les articulateurs libres, non-mobilisés par la consonne (plus forte pour /b/). Les fonctions de transfert de ces profils mesurés sur un locuteur sont en accord avec les simulation précédentes. Il est important de remarquer que la variation de la position de la langue pour /b/ ne conduit pas à des changements d'affiliation (les valeurs des formants sont groupées autour de la moyenne). Par contraste, pour /g/, même si la variation de la position semble moins importante, l'affiliation change comme la dispersion des formants F2 et F3 le montre sur la figure 4.15.

Nous retenons que la fonction de transfert du conduit vocal au moment du relâchement de l'occlusive est variable pour une même identité phonétique de la consonne en fonction du contexte vocalique. Comme cette variation est suffisamment forte pour changer l'affiliation du F2 dans le cas des /k g/, nous nous attendons à rencontrer des points focaux dans les transitions formantiques CV, comme dans les transitions VV.

4.4 Troisième pause

Dans ce stade de notre recherche nous pouvons dresser un court bilan. Depuis la dernière pause :

– nous avons passé en revue les propositions et les résultats sur la reconnaissance et la perception

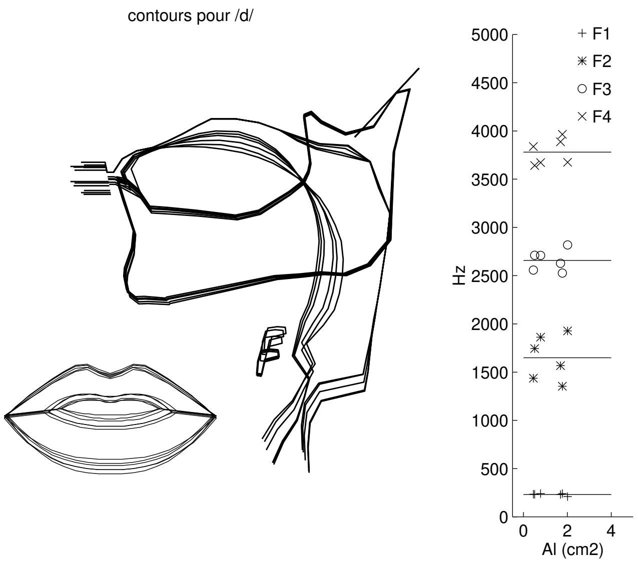


FIG. 4.14: Contours des articulateurs pour /d/ extraits à partir des données radiographiques. 6 réalisations du locuteur PB.

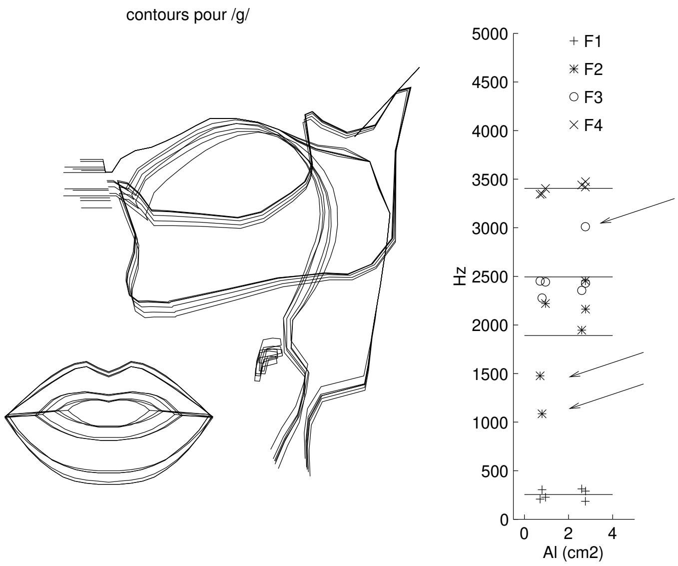


FIG. 4.15: Contours des articulateurs pour /g/ extraits à partir des données radiographiques. Les flèches indiquent les valeurs qui nous font penser que, à la différence des exemples sur les deux figures précédentes, pour ce phonème il y a changement d'affiliation.

des occlusives
- nous avons étudié la perception des occlusives à l'aide de monstres acoustiques conflictuels
- nous avons rappelé quelques résultats sur la production des occlusives.

Avant de procéder à l'analyse acoustique des stimuli, nous retenons plusieurs conclusions. Premièrement, nous constatons que la voyelle suivant la consonne est la source principale de variabilité.[3] En plus des variations des différents paramètres acoustiques selon le contexte vocalique rapportées par d'autres travaux, les tests de perception ont montré que les auditeurs humains mêmes réagissent très différemment aux monstres acoustiques en fonction de la voyelle.

Deuxièmement, nous avons établi une liaison entre notre étude des voyelles et les consonnes occlusives. À savoir, pour les occlusives, comme pour les voyelles, l'articulation à l'avant de la cavité buccale présente une spécificité particulière. Pour les voyelles, cette articulation fait changer l'affiliation majoritaire du F2. Pour les consonnes, cette articulation diminue la séparabilité entre les consonnes. Le contraste du lieu d'articulation semble différent dans le contexte antérieur par rapport à d'autres contextes car nécessitant la collaboration des différents segments de la syllabe. Nous essayons de guider notre analyse acoustique de manière à rendre explicite la relation entre la consonne et la voyelle qui la suit.

[3] En contexte V_1CV_2, la voyelle suivante influence beaucoup plus que la voyelle précédente, résultats en accord avec les conclusions de [Kru88, page 53].

Chapitre 5

Caractérisation et identification des consonnes occlusives

5.1 Analyse acoustique. Choix de la méthode d'investigation

Notre but premier n'est pas de chercher à tout prix le meilleur taux de reconnaissance des occlusives. Nous voulons plutôt chercher les raisons qui limitent les performances des systèmes actuels dont certains sont déjà assez performants [NZ91]. Nous poursuivons ce but en utilisant deux méthodes :

– Analyse des erreurs. Nous cherchons à établir une typologie des erreurs. Nous identifions les cas difficiles, ceux qui sont responsables pour la majorité des erreurs.

– Comparaison des architectures. Nous cherchons à comprendre quels sont les limitations que le choix d'une certaine architecture pour le système de reconnaissance implique.

Cependant, pour la comparaison des architectures, nous devons utiliser des critères de performance objectifs, comme les taux d'identification sur une tache phonétique. En outre, cette comparaison est tributaire de la qualité de l'implémentation de chaque module des différentes architectures possibles.

Nous partons du principe que les algorithmes actuels de classification et de reconnaissance des formes sont suffisamment performants pour notre tâche (voir plus loin les boites noires marquées "classification" dans la description des architectures). Nous croyons que seule une amélioration marginale pourra venir de l'optimisation de ces algorithmes. Par conséquent, nous utilisons des algorithmes classiques, spécialement le classifieur bayesien. Occasionnellement, nous avons utilisé des algorithmes qui construisent automatiquement des arbres de décision, algorithmes issues de l'intelligence artificielle classique. Un travail en informatique effectué en parallèle avec le nôtre dans le même laboratoire a traité de manière pointue des algorithmes de classification et d'apprentissage automatique [Mun96]. Une comparaison pour une gamme plus large d'algorithmes sur des données que nous avons partagées (plusieurs codages du N-seg et V-seg et fusion directe), peut être trouvé dans le rapport de cette thèse en informatique. Nous retenons que le classifieur gaussien est le plus performant suivi de près par les réseaux connexionistes (mis en oeuvre à l'aide du logiciel libre de droits NOPT[1]).

Beaucoup plus important que le choix du classifieur nous semble le choix des paramètres qui vont être présentés à l'entrée de celui-ci. Si un bon choix ne garantit pas une bonne classification, un mauvais choix à ce niveau est irrécupérable aux niveaux supérieurs. La majorité de nos efforts se sont portés sur le codage du signal acoustique en amont de la classification. L'espace de paramètres que nous cherchons doit absolument conserver la séparabilité des observations. De manière préférable, il devrait en plus montrer une distribution monomodale des observations (un seul nuage par classe).

Pour les distribution monomodales, la statistique nous met à disposition des outils puissants pour mesurer la variabilité intra-classe et la variabilité inter-classe comme l'analyse de la variance multivariée (MANOVA, [Abd87]). Nous avons choisi pour mesurer la qualité d'un jeu de paramètres deux indices :

1. le taux d'apprentissage (TA) d'une population d'exemples codés par ce jeu par un classifieur bayesien supposant une distribution de probabilité gaussienne. Nous précédons la classification par une analyse discriminante canonique incrémentale ce qui signifie qu'une partie des paramètres d'entrée peut ne pas être prise en compte pour la classification.
2. le coefficient lambda de Wilks (λ). Il est définie [Sta95] comme le rapport entre les déterminants des deux matrices de covariance (dans les groupes et entre les groupes). $\lambda = 0$ signifie une séparabilité parfaite et $\lambda = 1$ signifie des groupes totalement confondus. Lambda varie donc dans le même sens que le rapport entre la distance moyenne intra-classe et la distance moyenne inter-classe.

Le coefficient lambda est plus informatif que le TA pour les distributions déjà bien séparées (deux distributions avec TA=100% peuvent avoir des lambda différents d'un ordre de grandeur). Nous avons préféré le TA dans la présentation de nos résultats pour sa lisibilité immédiate. Ces deux mesures sont sensibles aux items périphériques isolés ("outlyers"). Nous nous servirons donc aussi de calculs d'information mutuelle, de calculs et études de prototypes pour chercher dans l'ensemble des paramétrisations possibles.

Finalement, nos critères de performance peuvent être séparés en deux catégories selon les données utilisés pour les tests :

1. D'une part, nous avons les taux d'apprentissage et les taux de reconnaissance sur une base de sons de parole enregistrée par nos soins spécialement pour ces tests.
2. D'autre part, nous testons la capacité d'un algorithme automatique de prédiction des réponses des humains aux monstres acoustiques.

Nous pouvons maintenant décrire l'organisation de ce chapitre :

– Nous commençons par une description de modèles utilisés pour la fusion de données.
– Nous décrivons le corpus de parole naturelle que nous avons enregistré et la procédure de segmentation.

[1]http ://www.cs.cmu.edu/afs/csproject/ai-repository/ai/areas/speech/systems/ogitools/ogitools.tgz

– Nous analysons les indices acoustiques des consonnes. Premièrement, nous étudions les deux segments, N-seg et V-seg, de manière individuelle. Deuxièmement, nous étudions les différents manières de fusionner les informations des deux segments.
– Nous analysons les sources des erreurs de classification. Nous regardons l'incidence du choix de l'architecture sur ces erreurs.
– Nous finirons par les tests des algorithmes sur les monstres acoustiques.

5.2 Modèles pour la fusion des données

Nous voulons introduire dans cette section le cadre général qui a guidé notre recherche de paramètres discriminants. Il s'agit du cadre théorique développé au départ pour la fusion des capteurs dans la parole audiovisuelle [RR95] mais généralisable pour la fusion des deux sources d'information. Malgré l'absence des capteurs séparés, ce cadre est approprié pour l'étude de N-seg et de V-seg. En effet, nos tests de perception et des nombreux autres travaux déjà mentionnés montrent que ces deux segments disjoints peuvent contribuer soit de manière complémentaire soit de manière indépendante à l'identification phonétique. L'analogie avec les rapports entre les indices acoustiques et visuels pour la parole est donc justifiée.

Nous sommes en presence de deux sources d'information de nature continue qui contribuent à une même décision d'appartenance à une classe discrète. Nous séparons trois grandes groupes de modèles pour la fusion de l'information provenant de deux sources (voir figure 5.1) plus le groupe des modéles hybrides :

Identification directe (ou fusion directe) Ce modèle simple utilise un seul classifieur. Les paramètres d'une source augmentent les dimensions de l'espace de paramètres de l'autre source à l'entrée du classifieur. L'intégration des deux sources est implicite. Les poids de chaque sources s'établissent dans le processus d'apprentissage et ne sont plus accessibles après (dans un réseau connexioniste, par exemple). Outre sa simplicité, ce modèle accepte des codages différents pour les deux sources. L'inconvénient de ce manque d'un espace commun est la difficulté d'obtenir en sortie une alarme si les deux sources sont contradictoires ou si une source est absente.

Identification séparée (ou fusion tardive) Ce modèle suppose deux classifieurs dont l'apprentissage est indépendant et un module d'intégration qui relie la sortie des deux. Le module d'intégration peut employer plusieurs stratégies, toutes basées essentiellement sur un ordonnancement des candidats par les classifieurs (nous y reviendrons plus tard). Cette architecture, de par la nature de l'apprentissage séparée, favorise plus la complémentarité ou la compétition des deux sources. Elle permet de comparer les entrées et s'accommode bien de l'absence d'une source.

Identification après recodage Ce modèle suppose un espace continu commun de paramètres (donc au moins un transcodeur) et un module de pondération qui combine les deux sources en amont du classifieur responsable de la décision discrete. On peut choisir différent espaces comme cible du recodage et différentes stratégies pour le module de pondération. Si cette stratégie est

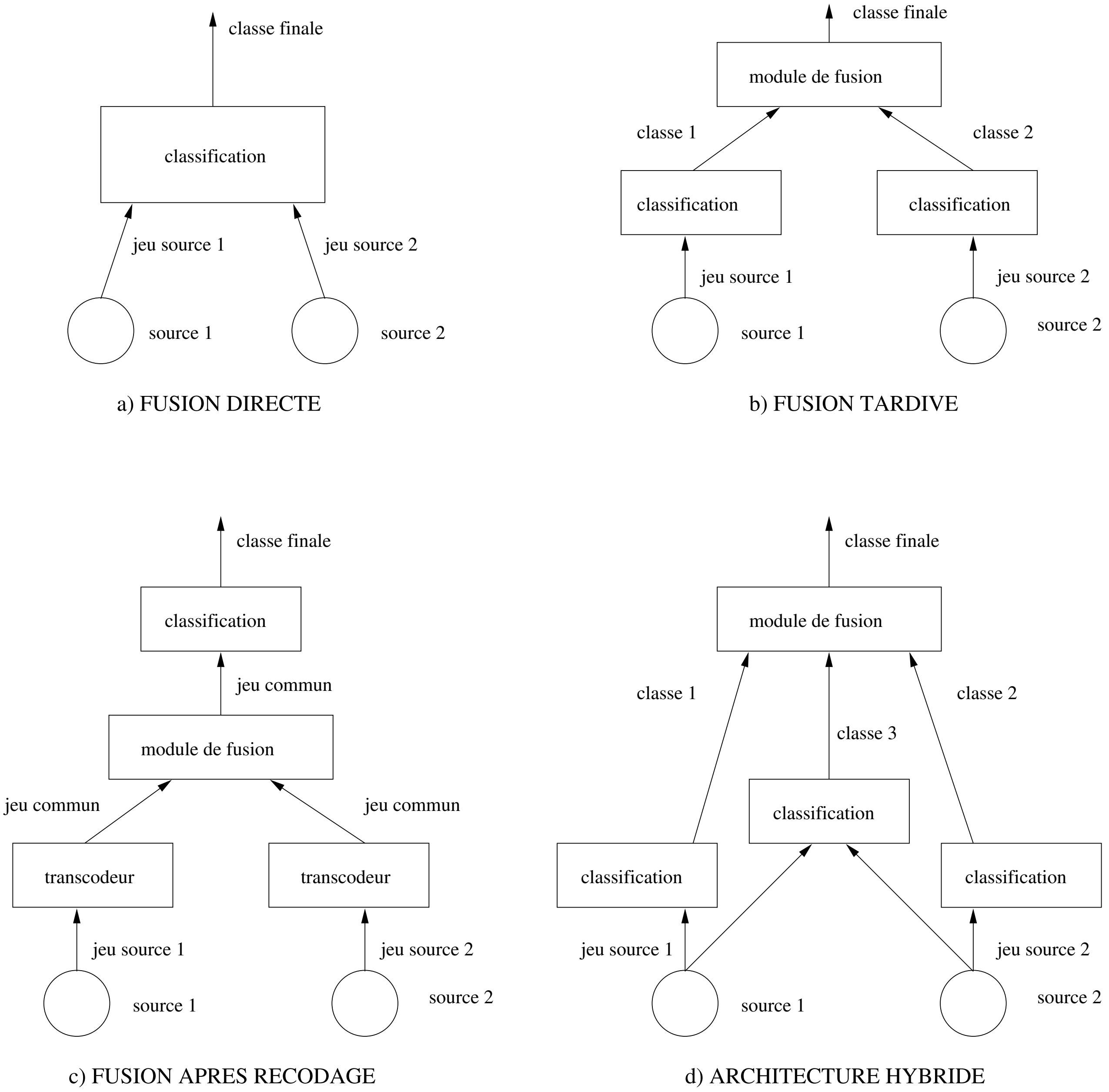


FIG. 5.1: Trois modèles pour la fusion des capteurs et une architecture hybride

apprise dans un processus joint avec l'apprentissage du classifieur, nous retrouvons le modèle à fusion directe.

Architectures hybrides Au prix d'une complexité accrue, elles palient aux inconvénients des architectures de base. Elles utilisent trois voies avant le module d'intégration (mais aussi parfois deux voix seulement). Une est identique à la fusion directe ou au recodage. En plus, cette architecture utilise un ou deux classifieurs dédiées à chaque source d'information, comme l'identification séparée. Ce type d'architecture permet de traiter de manière plus flexible une incohérence entre les deux sources.

Comme nous voulons exploiter et s'inspirer de la perception humaine, nous tenterons maintenant de mettre ces architectures en correspondance avec le modèle à quatre étages (analyse périférique, traitement morphologique, traitement sous-symbolique et traitement symbolique) que nous avons décrit en introduction. Par la même occasion, nous présentons les variantes des architectures que nous avons implémentées.

– Fusion directe simple. Cette implémentation comporte deux étages. Le premier fournit deux jeux disjoints de paramètres acoustiques, un pour le N-seg et un pour le V-seg. Le deuxième est un classifieur classique qui sépare des régions correspondant aux classes phonétiques dans l'espace de ces paramètres (figure 5.2).
– Fusion tardive. Cette version utilise un classifeur pour le N-seg et un autre pour le V-seg. La sortie de chacun de ces classifieurs donne les probabilités d'appartenance de l'entrée à une classe. Le module d'intégration les utilise pour prendre la décision finale. Nous avons essayé trois stratégies d'intégration et, en plus, deux architectures hybrides (figure 5.2).
– Fusion précoce. Pour cette version de recodage, les paramètres primaires sont calculés de manière plausible de point de vue psycho-acoustique à partir des deux segments. Cette architecture hérite de la fusion directe l'impossibilité de comparer les entrées et l'utilisation d'un seul classifieur. Cependant, la collaboration entre V-seg et N-seg n'est pas apprise mais figée par la méthode de calcul des paramètres (figure 5.3).
– Fusion explicite. Cette version particulière du recodage utilise un module de traitement intermédiaire avant le classifieur (figure 5.3). Ce module aspire à rendre compte de manière explicite de l'interaction entre les paramètres du N-seg et ceux du V-seg. Notre argumentation sur la nécéssité de l'analyse morphologique se base sur les performances de cette architecture.

5.3 Corpus

Comme nos tests de perception montrent une importance certaine du contexte vocalique, nous avons décidé d'enregistrer une base de signaux de parole spécifique pour notre analyse acoustique. Le but a été de disposer de syllabes combinant un nombre maximum de contextes possibles. Le corpus est formé des mots sans signification de la forme $V_1C_1V_2C_2V_2$ où V_1 est une de quatre voyelles extrêmes ([a], [i], [u] et [y]), C_1 est une des trois occlusives sourdes ([p], [t] et [k]), C_2 est toujours

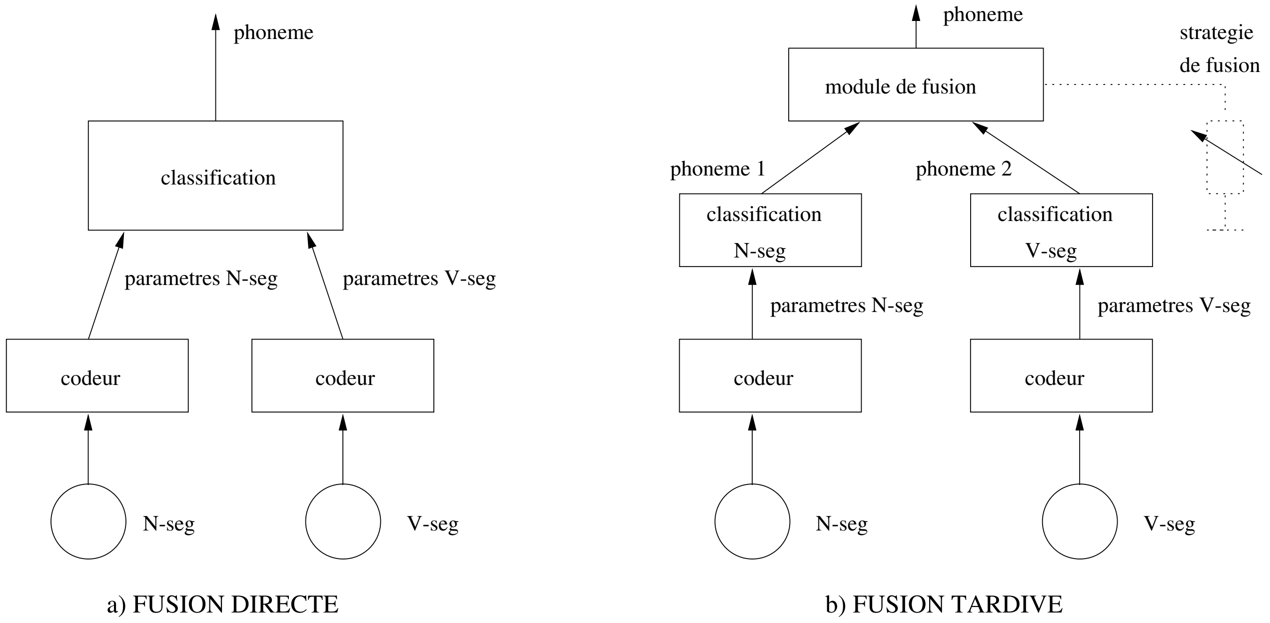


FIG. 5.2: Application des architectures à fusion directe et à fusion tardive à l'identification des occlusives

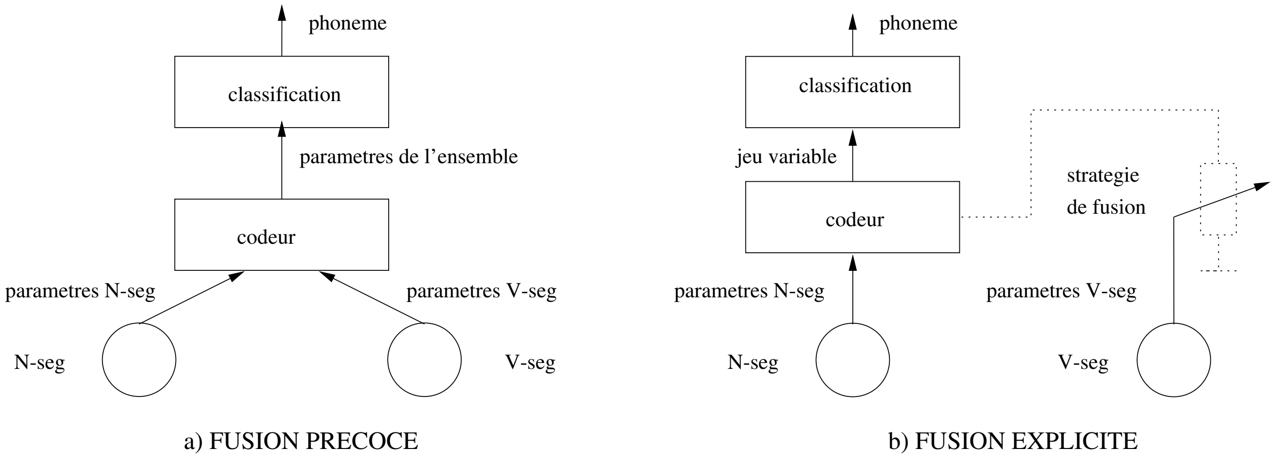


FIG. 5.3: Deux versions de fusion après recodage pour l'identification des occlusives

[s] et V_1 est une des dix voyelles orales du français [a e ɛ i u o ɔ y ø œ].[2] En outre, tous les contextes symétriques, $V_2C_1V_2sV_2$, ont été ajoutés. Nous avons ainsi 138 items différents.

L'enregistrement a été effectué dans la chambre sourde de l'Institut de la Communication Parlée. La même méthodologie et le même matériel [ZS90] a été utilisé lors de l'enregistrement de plusieurs base de données largement répandues comme EUROM1, BDBRUIT. Notamment, la numérisation a été effectué sur une carte OROS AU32 en 16 bits à 16 kHz. Cette fréquence d'échantillonage nous démarque des autres travaux dont la majorité ont étudié des signaux enregistrés à 10 kHz. Nous ne pensons pas que les indices acoustiques soient abondants dans la zone de fréquence ainsi récupérée (de 5 à 8 kHz). Cependant, une fréquence d'échantillonage élevée assure une meilleure résolution spectrale y compris dans les basses fréquences (nous disposons de plus d'échantillons dans le bruit de relâchement, par exemple).

Huit locuteurs masculins volontaires ont prononcé deux fois ce corpus dans deux sessions séparées d'au moins une journée. Chaque session a duré environ 40 minutes, pauses et passage de la consigne compris. Nous n'avons pas de raison valable pour ne pas avoir enregistré des locutrices. Les items qui nous ont servi pour les tests de perception constituent un petit sous-ensemble de cette base. Après contrôle des spectrogrammes, un nombre très réduit des mots a été réenregistrés. Nous disposons donc au total de 2208 VCVCV dans notre base (138*2*8). Les deux sessions comportaient une consigne différente. La première supposait la prononciation au débit d'une conversation courante et à l'intonation neutre (comme dans une énumération). La deuxième session demandait aux locuteurs de mettre l'accent sur la consonne occlusive. Cette intonation étant peu fréquente, un petit dialogue imaginaire était donné comme exemple : "C'est apasa, ça ? Non, c'est akasa !". Les locuteurs devaient copier l'intonation de la réponse négative. Chaque session comportait au début une petite séance non enregistrée pour l'accomodation des locuteurs avec les appareils et la prononciation demandée. A posteriori, cinq locuteurs ont rapporté la difficulté de produire de manière distincte les voyelles [o] et [ɔ] en dehors de mots avec une signification réelle. Un seul a rapporté la même difficulté pour [ø] et [œ].

5.3.1 Segmentation

En vue de l'extraction des paramètres acoustiques des occlusives sourdes, une segmentation manuelle de cette base a été effectuée. D'abord, il y a eu un premier passage, complètement automatique, utilisant un aligneur construit par des modèles phonétiques HMM implémentés avec HTK. Cet algorithme prend en entrée la chaîne phonémique et place des marqueurs sur le signal au centre et aux frontières entre les phonèmes [WBT90]. Le passage manuel a ajusté la position de ces marqueurs et a en ajouté deux autres pour délimiter le bruit de l'occlusive. Un soin particulier a été apporté au positionnement des marqueurs encadrant le N-seg :

1. afin de pouvoir mesurer le VOT avec précision (à quelques échantillons près)

[2] Après avoir analysé sous tous les angles cette base, le choix du C_2 s'est averré perfectible. Cette fricative a favorisé peut être trop le dévoisement des voyelles [i] après l'occlusive.

2. afin de rendre imperceptible le recollage des segments dans la construction des monstres acoustiques.

Le premier marqueur, appelé BR comme "burst release", marque l'ouverture du conduit vocal donc le début du signal après l'occlusion dans le cas des occlusives sourdes. Le deuxième marqueur que l'expert a mis sur le signal marque le délai d'établissement du voisement et nous l'avons appelé VOT (comme "voice onset time"). Nous avons souhaité que la durée entre ces deux marqueurs corresponde à l'indice acoustique VOT tel qu'il a été défini par Lisker [LA64] : la durée entre le relâchement de l'occlusive et la mise en vibration des cordes vocales. Cependant, dans la pratique, plusieures difficultés empêchent de mesurer avec précision ce paramètre :

- Le passage de l'excitation fricative à l'excitation glottique n'est pas instantané. En fait, il y a trois événements acoustiques distincts qui pourraient recevoir le marqueur VOT, cités ici par ordre chronologique :

  1. Le tout premier passage par zéro de la première oscillation glottique.
  2. La fin des hautes fréquences typique du bruit de friction.
  3. Début du formant F1 lu sur un spectrogramme à large bande [Kla75].

  Nous avons choisi de marquer systématiquement le premier événement mentionné : le passage par zéro. Et ceci même si le bruit débordait de quelques milisecondes dans la période voisée (voir aussi la figure 4.1). Autrement dit, nous avons voulu un N-seg complètement dépourvu d'oscillations glotiques quitte à tolérer un V-seg contenant du bruit de friction. Un choix semblable a été fait par Krull [Kru88, page 73]. Pour la description d'un cadre complet et plus général pour l'étiquetage des événements acoustico-articulatoire se reporter aux travaux de C. Abry et collaborateurs [ABBS85].[3]

- L'onde de choc du relâchement du [p] a une fréquence propre proche de la fréquence des vibrations des cordes vocales des nos locuteurs. La première partie de cette onde (la composante surpression) ressemble à une ouverture de glotte et la deuxième partie (la composante dépression) déborde souvent sur la première oscillation glottique. Nous avons cherché le passage par zéro entre les deux composantes de cette onde. Toutefois, à cause de la composante dépression, le marqueur VOT n'était pas mis à un passage par zéro. Cette onde basse fréquence caractéristique pour les labiales est d'autant plus importante que la distance du locuteur au microphone est faible.

- Un autre problème concerne la marqueur BR. Relâchements très peu marqués ou relâchements multiples dans la même consonne peuvent tromper l'expert sur la position correcte du BR. Ces situations sont beaucoup plus rares que celles qui affectent la position du VOT et sont éliminables par consensus entre les experts.

Nous présentons dans la figure 5.4 des cas typiques pour les trois consonnes étudiées devant les quatre voyelles extrêmes. Nous avons baptisé le segment entre les marqueurs BR et VOT N-seg

[3] Dans ce cadre, les trois événements mentionnés sont, dans le même ordre, VO, FT et VVO. Nous avons choisi pour VOT la durée entre CFO et VO.

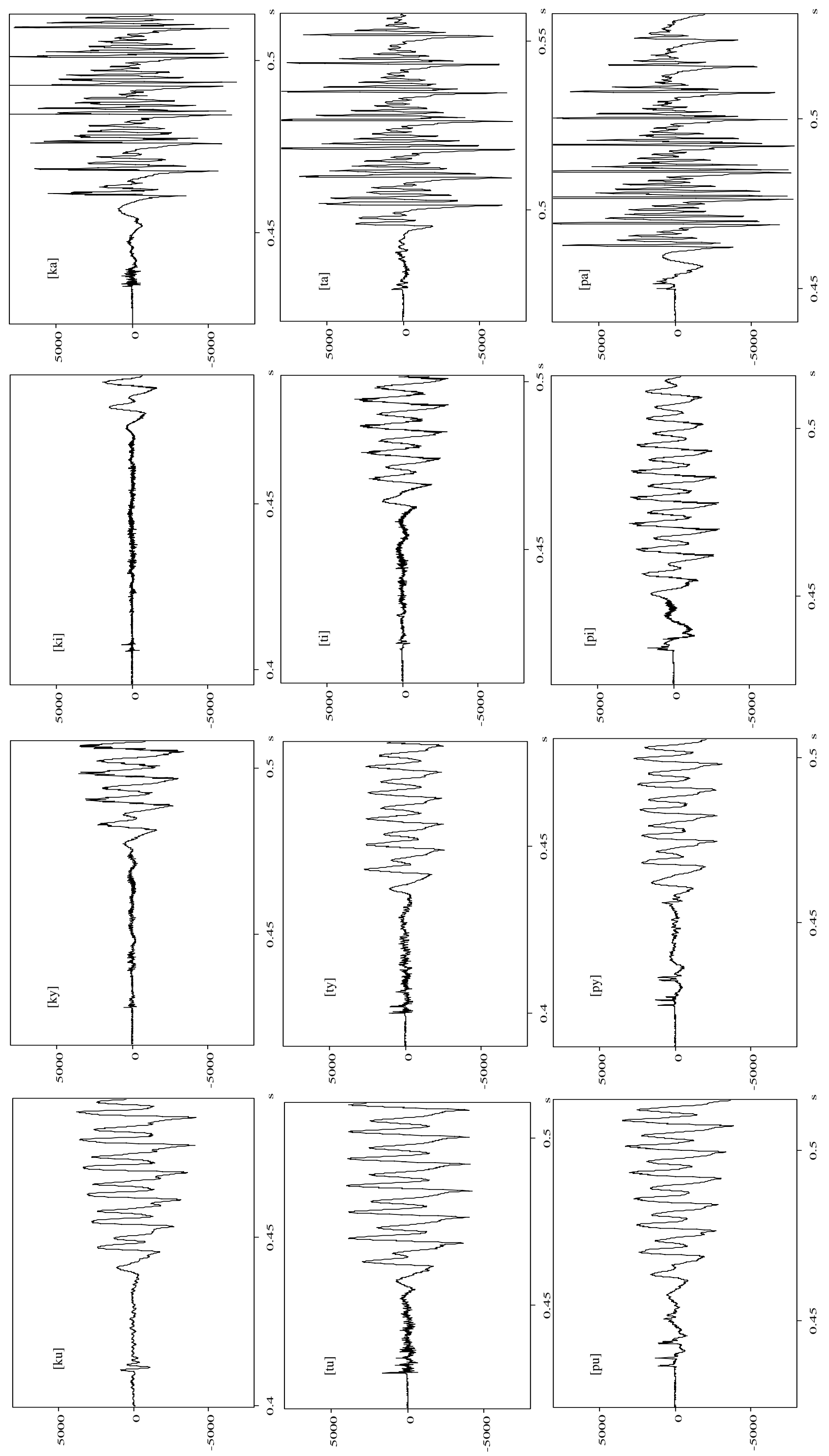


FIG. 5.4: Oscilogrammes des 12 CV du locuteur JS. 90 ms de chaque signal, alignement temporel sur l'instant du relâchement

(comme "noise-like segment") car, dans la quasi-totalité des cas, il renferme exactement la partie de type bruit de la consonne.

L'étape suivante aurait été de tenter la séparation dans le N-seg de la partie transitoire impulsionelle et de la partie excitée par la friction à la constriction, séparation possible par un algorithme automatique [Dje95a, page 26]. L'examen de nos signaux nous a convaincu de l'impossibilité de faire la séparation entre la barre d'explosion et le bruit de friction de manière systématique et cohérente pour tout le corpus. Si, pour les contextes fermés, le spectrogramme permet souvent de distinguer deux zones à caractéristiques spectrales homogènes, pour le [p] en contexte ouvert le bruit de friction est souvent trop faible ou simplement absent. Nous avons trouvé aussi l'inverse : des situations où le bruit de friction est bien plus intense, au point même que la barre d'explosion devient difficilement détectable.

Concernant la segmentation manuelle, nous savons qu'elle introduit une connaissance sur le signal qui ne sera pas présente dans une situation réelle d'exploitation. À notre connaissance, peu de travaux de recherche combinent la détection automatique de la position du bruit et la reconnaissance du lieu d'articulation avec des résultats acceptables [TI96], [CM94, page 66]. Souvent une élimination manuelle des plus grossières erreurs de la détection est utilisée. Des recherches parallèles dans notre laboratoire sur cette détection automatique [PSB$^+$96] ont donné de résultats qui aurait été trop limitatifs par rapport au but que nous poursuivons.

5.4 Étude de la capacité de séparation des indices

Dans cette section, nous allons procéder à l'analyse acoustique des signaux de notre base de données. Une sous-section sera dédiée à chacun des deux segments comparés dans l'étude de perception. Les buts seront, d'une part, de raffiner les prototypes des consonnes et, d'autre part, de chercher une bonne paramétrisation pour chaque segment. En outre, à l'occasion de la recherche de paramètres pour le V-seg, nous présenterons une méthode pour le suivi de formants.

Comme nous l'avons annoncé, notre conviction est que seul le codage de l'ensemble des deux segments pourrait assurer une robustesse à la reconnaissance automatique du lieu d'articulation. Donc une sous-section sera dédiée à la fusion des information des deux segments. Plus particulièrement, dans la partie "Fusion précoce", nous regarderons les codages globaux des segments sourd et sonore. Nous finirons par une relecture des résultats montrant comment l'information sur les identités de la consonne et de la voyelle est distribuée dans le continuum sonore des syllabes CV.

De cette section nous allons retenir donc les meilleurs indices dans différentes architectures, avec ou sans fusion. Nous n'étudions pas l'influence d'autres factures, notamment celle de la voyelle. Elle fera l'objet de la section suivante (5.5).

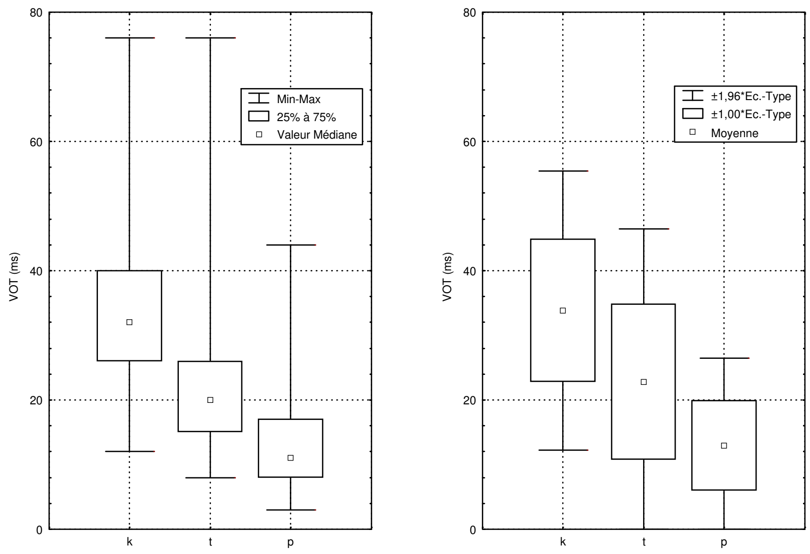


FIG. 5.5: Distribution du VOT par consonne. a) Médianne, quartile et étendue b) Moyenne et écart type.

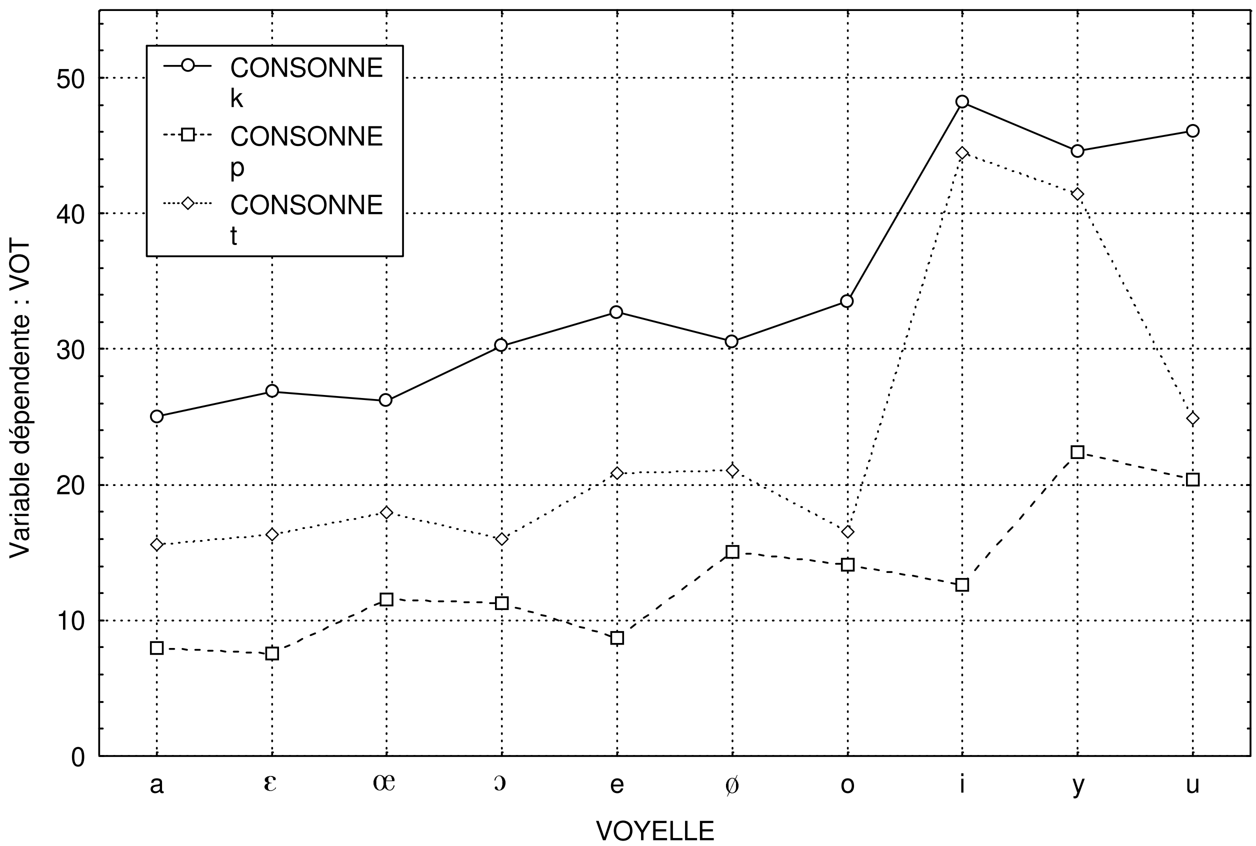


FIG. 5.6: VOT moyen fonction de la voyelle et de la consonne.

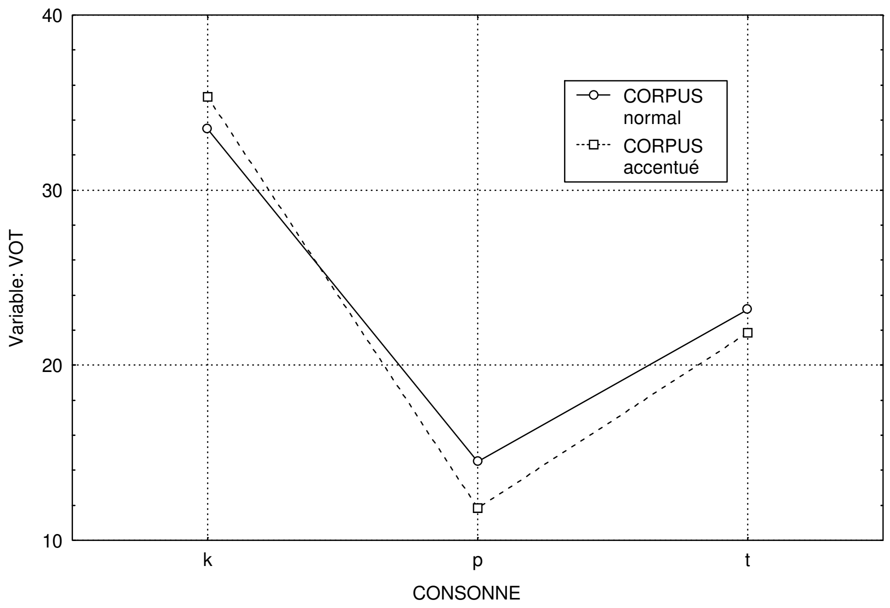


FIG. 5.7: VOT moyen fonction de la consonne et du degré d'articulation

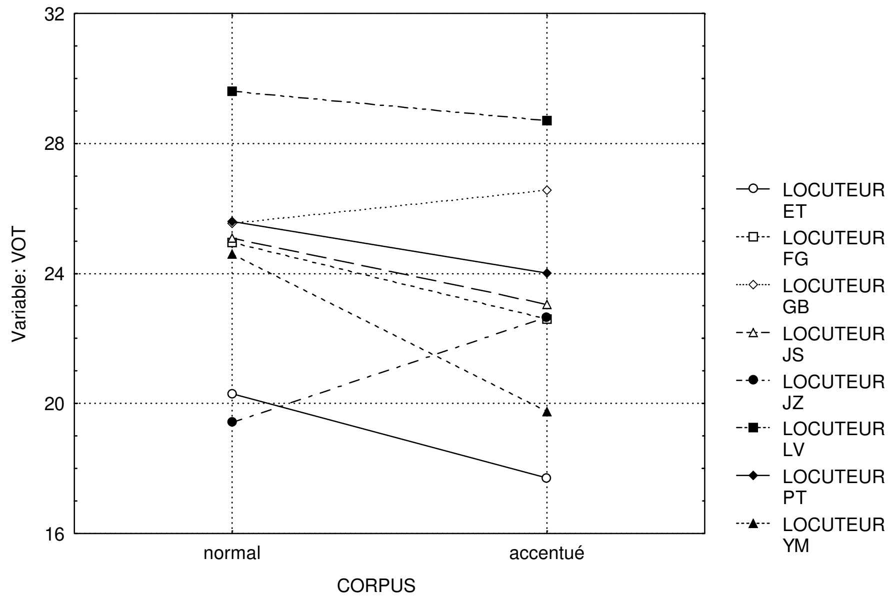


FIG. 5.8: Différents stratégies des locuteurs pour la parole claire, cas du VOT.

Consonne	M1	M2	M3	M4
p	13.0 ± 7.1	19 ± 10.5	21	17
t	22.8 ± 12.1	29 ± 16.0	29	26
k	33.8 ± 11.1	38 ± 13.6	38	45

TAB. 5.1: Moyennes et écart-types des occlusives sourdes en français en milisecondes. M1 sur notre corpus, M2 par Djezzar en 1995, M3 par Wajskop en 1979, M4 par Serniclaes en 1984.

5.4.1 Indices du N-seg

5.4.1.1 Étude du VOT

Nous présentons ici les résultats d'une analyse de variance répétée de la durée du segment N-seg (ANOVA effectuée par le logiciel Statistica 5.1). Cette analyse porte sur 2208 items. Nous étudions quatre facteurs : le locuteur (8 locuteurs), degré d'articulation (2 niveaux, normal et hyperarticulé), consonne (3 lieux d'articulation [p t k]) et voyelle (10 voyelles orales).

Cette analyse nous montre que l'influence des trois de ces facteurs est hautement significative (niveau $p < 10^{-6}$) : consonne, voyelle, locuteur (par ordre de signification). Cinq sur six interactions d'ordre deux sont significatives (toujours par ordre de signification) : articulation - consonne, articulation - locuteur, consonne - voyelle, locuteur - consonne et locuteur - voyelle. Trois des quatre interactions d'ordre 3 et l'interaction de tous les facteurs sont données significatives mais notre corpus n'est pas assez fourni pour étudier ces interactions.

Nous allons maintenant détailler et interpréter les influences statistiquement significatives. D'abord, la distribution du VOT selon la consonne correspond à celle rapportée maintes fois pour le français [Waj79, Ser84, Dje95a]. Nous avons moins de locuteurs mais deux à trois fois plus d'items que les études citées. Les valeurs moyennes sont plus faibles (voir tableau 5.1) ce qui s'explique facilement par les différences dans la procédure de segmentation (nous avons choisi un N-seg sans voisement). Concernant le pouvoir de discrimination, nous retrouvons le résultat bien connu (voir figure 5.5) : les labiales sont séparés des vélaires mais les dentales, au milieu, donnent un très important recouvrement entre les classes. En conclusion, le VOT ne peut pas être le seul paramètre pour la récupération du lieu d'articulation en français.

L'influence de la voyelle est très claire. Comme nous pouvons l'observer sur la figure 5.6 montrant l'interaction voyelle - consonne, en ordonnant les contextes vocalique par ordre croissant du VOT, nous obtenons en fait un arrangement des voyelles par degré d'ouverture. Nous remarquerons surtout les long VOT de [t k] en contexte [i y] du à un dévoisement de la voyelle, dévoisement facilité par la consonne fricative suivante ([s]). L'allongement du VOT avec la fermeture de la voyelle explique une partie de son impact sur la perception, comme nous l'avons déjà montré.

Malgré le fait que l'influence directe de l'articulation est négligeable, ce facteur donne des interactions très significatives. Le passage à la parole claire modifie la distribution du VOT dans le sens de l'augmentation du contraste : le [k] devient plus long et le [p] devient plus court (voir figure 5.7). Les locuteurs réagissent de manières très différentes au passage à l'hyper-articulation. Deux locuteurs ont augmenté le délai d'établissement du voisement et six l'ont diminué (voir figure 5.8) conduisant à une variation interlocuteur plus grande en parole claire.

Ces résultats sont à considérer avec la réserve des limites de validité de l'ANOVA : non seulement les variances ne sont pas égales pour chaque groupe mais les distributions sont asymétriques (donc non-gaussiennes) surtout pour [p], comme on peut remarquer sur la figure 5.5.

Nous n'avons pas accompagné la prise en compte de l'amplitude du N-seg dans le test de perception avec des statistiques sur celle-ci. D'autres travaux, voir [Dje95a] pour le français, montrent un pouvoir de discrimination comparable, sinon plus faible, que celui du VOT.

5.4.1.2 Spectres prototypiques

Pour l'analyse acoustique du N-seg, plusieurs méthodes ont été utilisées. D'abord, il est possible d'extraire le segment (l'entourer de silence). Cela évite l'influence sur le spectre des segments voisins, surtout le segment vocalique. Cependant, le calcul sans extraction est aussi envisageable, surtout d'un point de vue psycho-acoustique. Nous y reviendrons dans la sous-section "Fusion précoce".

Sachant que nous voulons un spectre d'un segment de durée variable, plusieurs alternatives sont possibles :

- utiliser une fenêtre de longueur variable (jamais exploitée)
- utiliser une fenêtre courte glissante et calculer un spectre moyen sur un nombre variable de spectres
- utiliser une fenêtre de longueur fixe égale ou supérieure à la longueur maximale (zéro padding)
- utiliser une fenêtre fixe courte sur une portion du segment (négliger le reste du segment) après avoir établi, automatiquement ou à la main la zone intéressante.

Les méthodes les plus élaborées utilisent une combinaison de ces solutions. Par exemple, Bonneau et Djezzar utilisent le spectre moyen d'une fenêtre glissant sur une zone (la barre d'explosion) délimitée de manière semi-automatique. Nous avons choisi la dernière solution : une fenêtre Hamming centrée sur le marqueur BR de longueur fixe. Comme nous étudions les occlusives sourdes, elle est assimilable à une fenêtre asymétrique démarrant au moment du relâchement. La durée de la fenêtre d'analyse est sujet à un compromis. Trop courte, la fenêtre n'assure pas une bonne résolution en fréquence et n'est pas robuste par rapport à l'étiquetage mais trop longue, elle s'écarte de l'événement acoustique important (l'occlusion) et le spectre devient plus une moyenne qu'un instantané. Vu la statistique sur la durée du N-seg (voir figure 5.5), nous avons choisi une durée de 10 milisecondes pour la fenêtre asymétrique. Cette durée représente, dans un [ki] par exemple, une petite partie du N-seg. En revanche, dans un [pa], elle déborde dans le V-seg, nos N-seg n'étant pas extraits du contexte. Trois exemples appuyant notre choix de durée d'analyse sont présentés sur la figure 5.9. Cette durée peut paraitre petite pour avoir une bonne résolution spectrale. Il convienent de rapeller ici l'étude de Kobatake et Ohtani en 1986 sur des CV japonnais. Ils trouvent que :

> "... the gross spectral shape which caracterizes /t/ and /p/ fade out very quickly and the time duration in which distinct spectral differences between them can be observed is as short as 5 ms after the burst [onset]." [KO87, page 1146]

Pour le lissage du spectre nous utilisons systématiquement la prédiction linéaire (LPC) précédée d'une pré-accentuation des haute fréquences. Nous avons essayé différents degrés de lissage, LPC d'ordre 12 à 20.

Nous présentons ici les spectres obtenus pour le premier locuteur. D'abord, nous montrons les items étudiés dans les tests de perception (voir figure 5.10). Pour avoir une bonne mesure de la variation intralocuteur, nous avons analysé avec la même méthode un autre corpus prononcé par le même locuteur, corpus conçu pour les études audiovisuelles [BLMA92]. Dans ce corpus nous avons moins de contextes (4) mais beaucoup plus de répétitions d'une même syllabe. La figure 5.11 montre

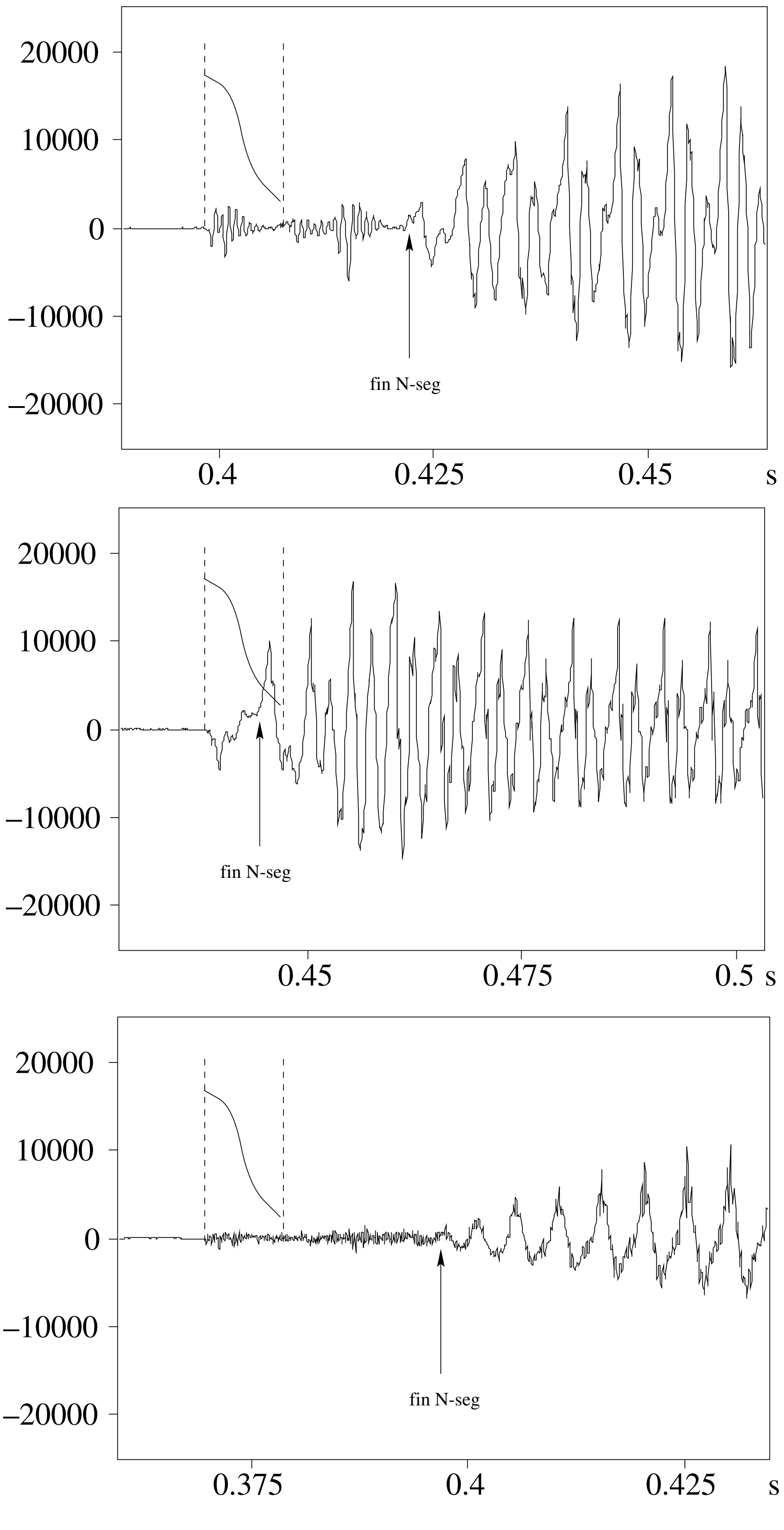


FIG. 5.9: Cas typique et deux cas extrêmes pour la position de la fenêtre d'analyse par rapport au N-seg : a) N-seg > 10 ms, [p] devant les voyelles fermées, [t] et [k] devant les autres, b) N-seg < 10 ms, [p] devant les voyelles ouvertes et c) N-seg » 10 ms, [t] et [k] devant les voyelles fermées. Les exemples sont [kɔ], [pɛ] et [ty] par ET, débit normal.

donc des spectres moyens à travers plusieurs réalisations du même CV et, en plus, l'écart type autour de ce spectre moyen. L'ensemble des résultats rejoignent les études du CRIN sur le français, malgré des petites différences dans la méthode [Dje95a, page 96]. Ces spectres appellent les remarques suivantes :

- Nous retrouvons les formes globales (gross shapes) décrites par Blumstein et Stevens [SB78, BS79] seulement pour les contextes [u y a] et si nous limitons nos spectres à 4 kHz. Ainsi, le spectre du [p] est plat ou descendant, le [t] est croissant vers les hautes fréquences et le [k] présente une masse spectrale compacte. Cette étude nous permet d'aller plus loin que cette description approximative.
- Nous retrouvons la dépendance du contexte de la forme globale rapportée par Suomi [Suo85]. Le plus influencé est le [k], puis le [t] et, finalement, le [p] semble presque insensible au contexte. Nous observons trois type de contraste entre les occlusives selon le contexte vocalique :
  1. En **contexte postérieur** le bruit de [k] a une forme bimodale très spécifique. Le premier maximum est situé à 1 kHz, au même niveau que la valeur de départ du F2 dans le segment vocalique. La consonne et la voyelle peuvent être bien reconnues à partir de ce bruit. Le [t] présente plus de hautes fréquences avec un maximum prononcé entre 3 et 4 kHz. Le [p] a moins d'énergie et moins de hautes fréquences que les deux autres dans tous les contextes.
  2. En **contexte labialisé** le bruit de [t] ne change pas de manière importante par rapport au contexte postérieur. Le bruit de [k] présente toujours un maximum important devant la valeur initiale du F2 dans le segment voisé suivant, cette fois à environ 2 kHz. Le deuxième maximum du [k] est nettement moins marqué. Mais le pouvoir de discrimination de la forme spectrale est conservé.
  3. En **contexte antérieur** de même qu'en contexte très ouvert le bruit du [t] change de forme. Il devient plus diffus que le [k] dans le même contexte. Ici c'est le [k] qui prend une forme proche de celle qu'avait [t] dans les autres contextes. Le pouvoir de discrimination est réduit. Le bruit de [k] et [t] sont différents dans une zone étroite de fréquence dont la position semble aussi corrélée avec la position du R2 de la voyelle suivante. Nous reviendrons sur cet aspect. À la différence des contextes ouverts, pour le [i], cette réduction de pouvoir de discrimination n'est pas compensée par l'augmentation de celui du segment voisé.
  4. Pour les voyelles intermédiaires, les formes des spectres peuvent être prédites par interpolation des formes des contextes extrêmes. Nous illustrons ce gradient par l'évolution du contraste [t]/[k] en contexte postérieur avec le degré d'ouverture de la voyelle (voir figure 5.13).
- Finalement, les différences entre les deux corpus sont concentrées au niveau du [p]. Les conditions d'enregistrement de deuxième corpus (débit de parole nettement plus rapide et microphone éloigné du locuteur et désaxé) ont fait disparaître la composante base fréquence bien

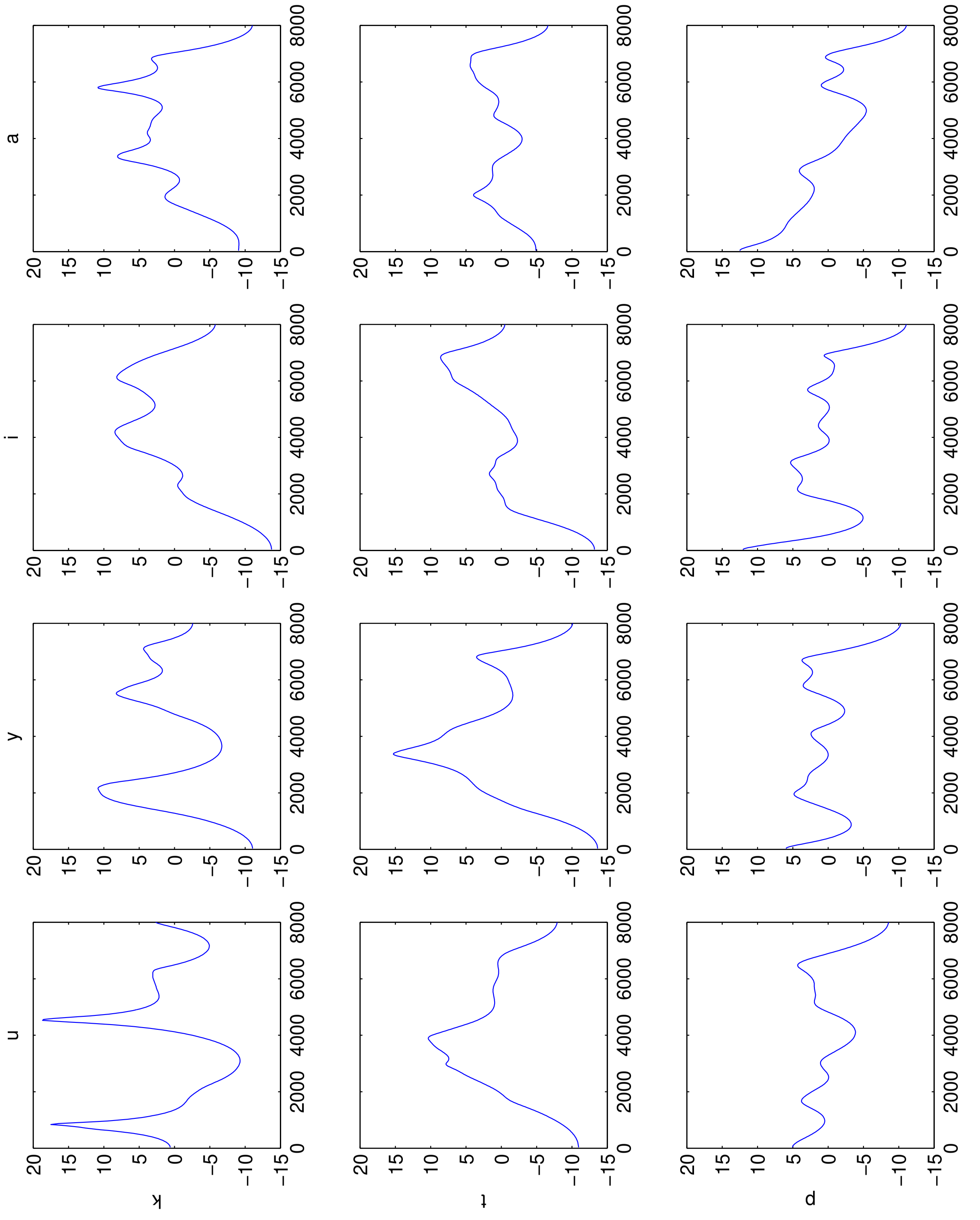


FIG. 5.10: Spectres du bruit de relâchement de CV du locuteur JLS. LPC ordre 20, Hamming asymétrique 10 ms, moyenne de 4 réalisations, 0-8 kHz.

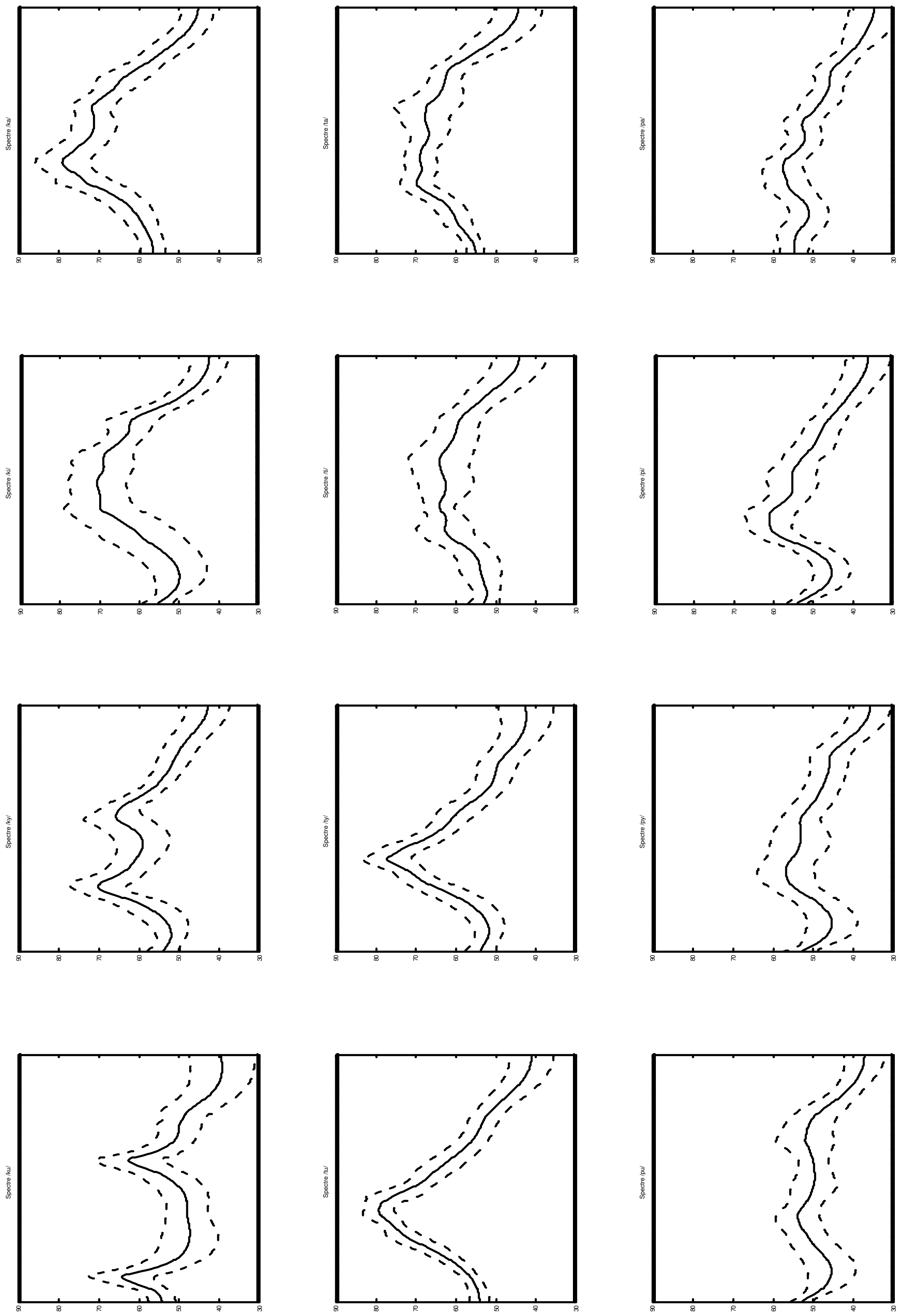


FIG. 5.11: Spectres du bruit de relâchement de CV du locuteur JLS. Même méthode de calcul mais représentation moyenne $\pm$ écart type. Corpus audiovisuel, 64 réalisations par case, 0-8 kHz.

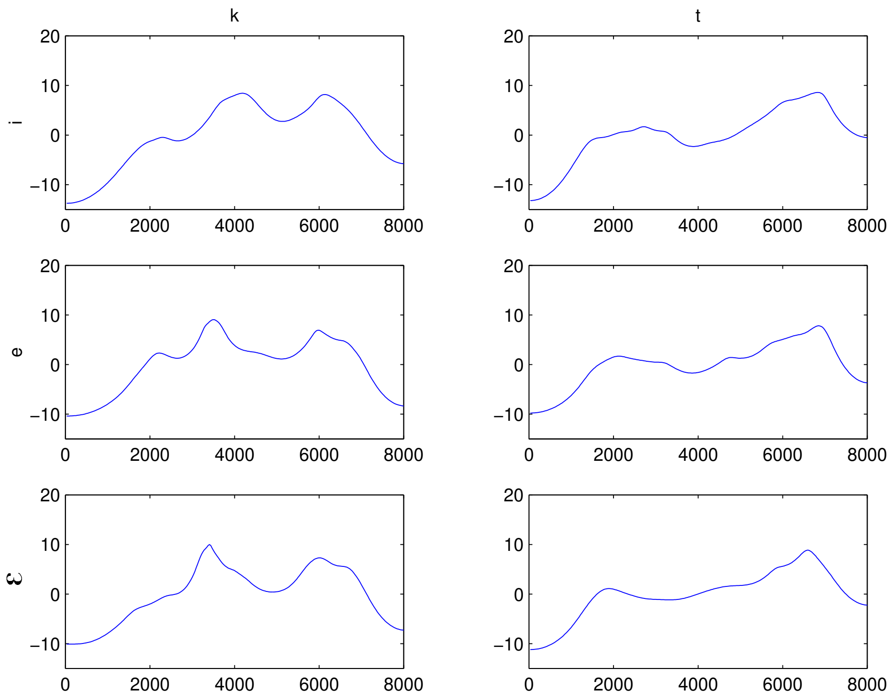


FIG. 5.12: Évolution du contraste spectrale [t]/[k] avec le degré d'ouverture de la voyelle suivante. Locuteur JLS, 5 réalisations par case. a) contexte antérieur

spécifique aux labiales dans notre corpus. Une augmentation des confusions [p]/[t] peut en résulter.

Pour identifier les spectres prototypiques, nous nous sommes servi d'un algorithme automatique pour la construction d'arbres de décision proposé par J. R. Quinlan : FFOIL[4]. À l'époque, nous n'avions enregistré que deux locuteurs. Ainsi, 276 CV du premier locuteur (celui étudié dans le test de perception) nous ont servi comme corpus d'apprentissage et 138 CV du deuxième locuteur ont constitué le corpus de test. Sur cette base réduite, nous avons mené un essai de reconnaissance avec et sans révision des règles produites à l'apprentissage par un expert phonéticien. L'intérêt de cette manipulation réside dans la capacité des algorithmes de construction des arbres à isoler de manière non-supervisée des prototypes et à sélectionner les paramètres les plus discriminants. Cette capacité est illustré par la figure 5.15. Ici les lignes correspondent aux trois consonnes et les colonnes aux 7 règles produites. Chaque règle est caractérisée par l'étiquette phonétique qu'elle mets sur un spectre satisfaisant ses prédicats. Les règles sont ordonnées de gauche à droite par l'importance du gain informationnel qu'elles apportent (ordre de construction par FFOIL).

Le fait que la première et seulement la première règle sépare les [p] signifie que ses spectres sont les plus facilement séparables. La deuxième règle sépare les [k] en contexte arrière et la troisième les [k] en contexte labialisé. La quatrième et la cinquième séparent les [t] en contextes avant ou ouvert et la sixième les [t] en contexte arrière et labialisé. Le fait que l'algorithme a spécialisé plus de règles avant d'arriver au contexte avant confirme qu'il y a multi-modalité et nous montre le contraste

[4]ftp ://ftp.cs.su.oz.au/pub/ffoil2.sh

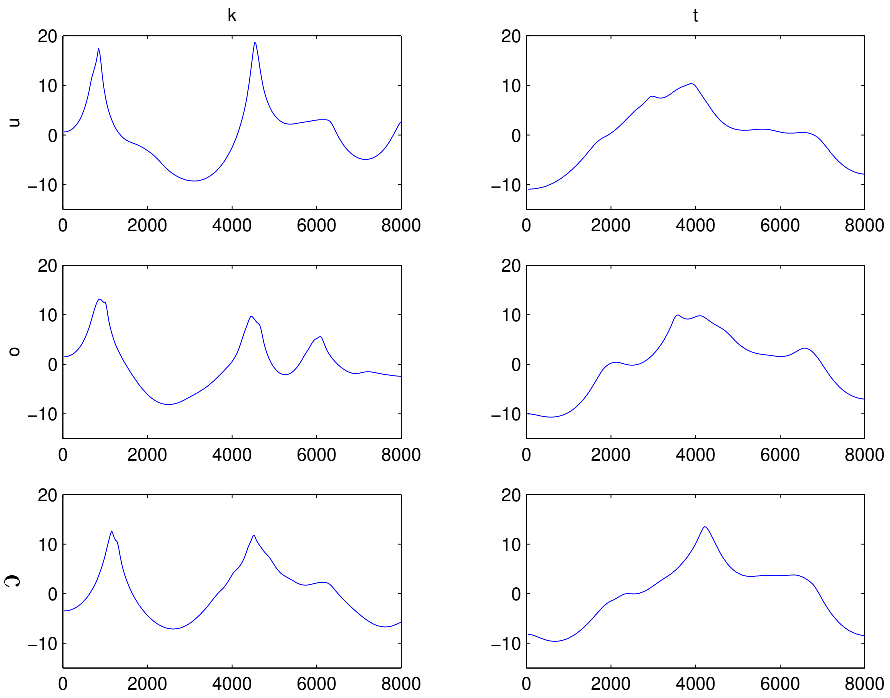


FIG. 5.13: Évolution du contraste spectrale [t]/[k] avec le degré d'ouverture de la voyelle suivante. Locuteur JLS, 5 réalisations par case. b) contexte postérieur

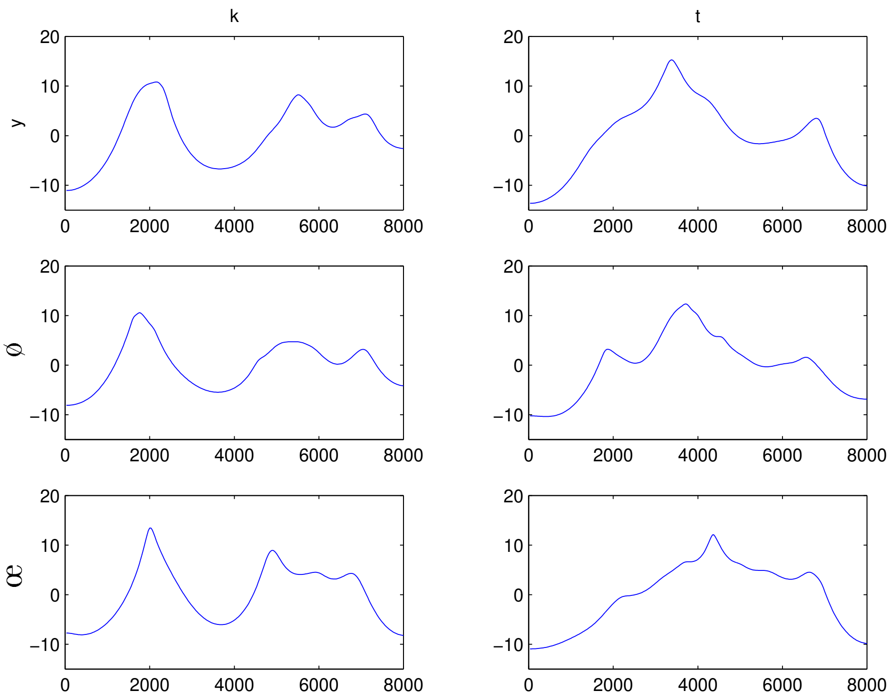


FIG. 5.14: Évolution du contraste spectrale [t]/[k] avec le degré d'ouverture de la voyelle suivante. Locuteur JLS, 5 réalisations par case. c) contexte labialisé

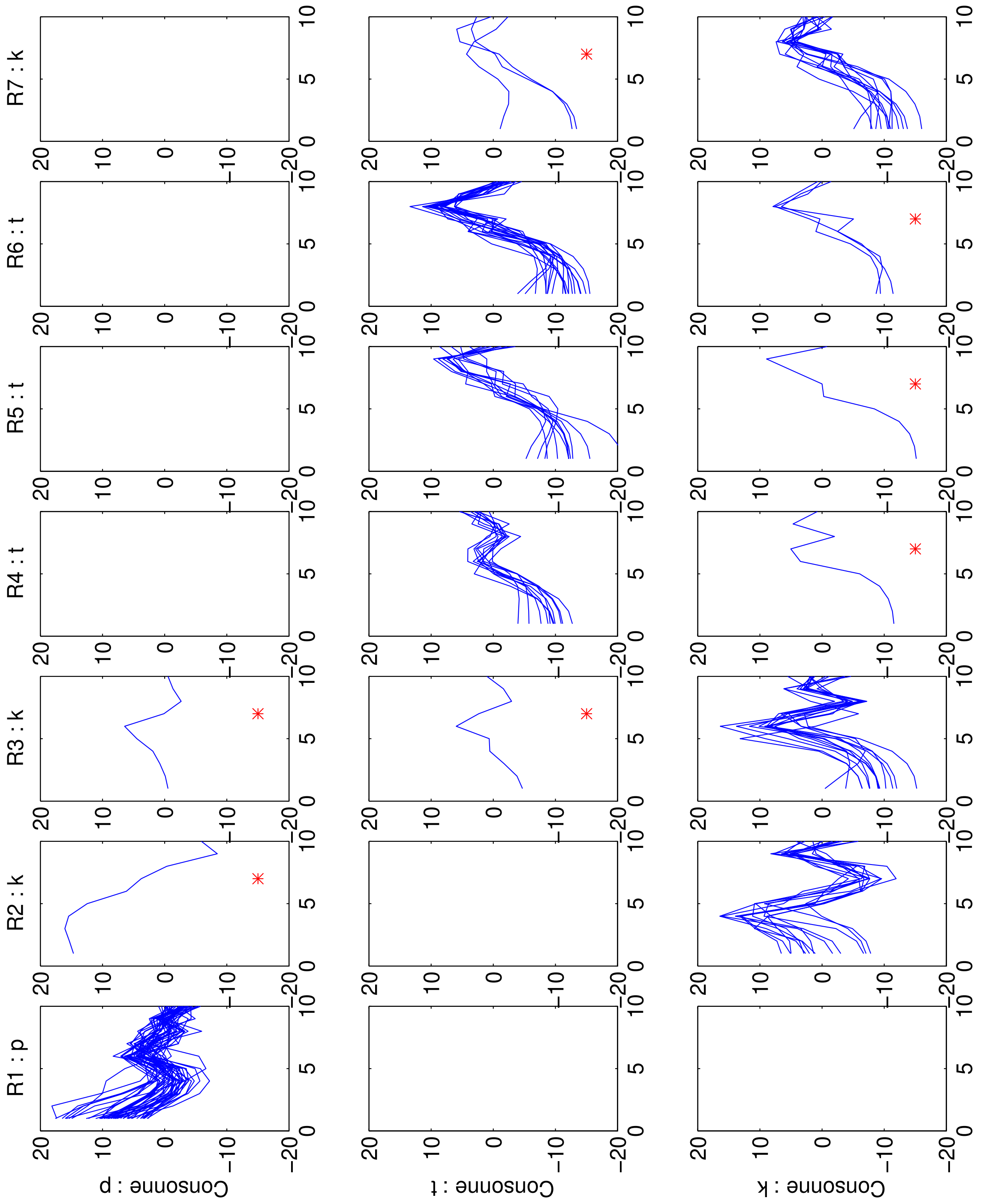


FIG. 5.15: Spectres du bruit des consonnes du corpus d'apprentissage. Les rangés correspondent aux 7 règles du classifieur respectivement au trois classes (p, t et k). Chaque spectre est représenté dans la case de sa classe et de la règle dont il satisfait les critères. Les cases marqués avec une étoile recèlent des erreurs de classification. Les spectres sont codés par l'énergie en 10 bandes de 2.2 barks

difficile pour ce codage ([t] versus [k] en contexte avant ou ouvert). Ce paquet de règles couvre bien les exemples d'apprentissage (taux d'apprentissage TA=92.0%) et donne un taux correct de reconnaissance du lieu sur le corpus de test : TR=83.3% (CG : 84.1% ; NOPT : 79.7%). Pour plus de détails sur les prédicats mêmes ou sur l'intérêt de la révision (dégradation du TA mais amélioration importante du TR, réduction du nombre de règles) voir la thèse de Munteanu [Mun96].

5.4.1.3 Paramétrisation du spectre

Comme nous l'avons déjà évoqué dans l'étude bibliographique, beaucoup de paramétrisations différentes ont été proposés dans la littérature. Le trait commun de ces travaux a été de transcender la variation contextuelle évidente sur la figure 5.11 tout en gardant un maximum de pouvoir de discrimination du lieu d'articulation.

Malgré l'utilisation fréquente de la valeur de fréquence du maxima les plus proéminent du spectre, également appelée BF pour "burst frequency" [Edw81, Dje95a, StBC96b], nous n'envisageons pas le calcul de cet indice pour les raisons suivantes :

- Premièrement, le [p] dans toutes les contextes et le [t] en contexte avant ou ouvert sont suffisamment diffus pour que la position du pic soit entaché de bruit : des spectres proches peuvent fournir des BF très différentes. En contexte arrière, le deuxième pic du [k] emporte parfois en amplitude et il est très proche de la zone où on attend le pic du [t]. Nous avons utilisé pour la détection de maxima les même deux méthodes que pour la détection des formants (seuilage à hytéresis et racine LPC). Comme pour [t] et [p] c'est le manque d'énergie à un endroit précis qui porte l'information, le BF n'est pas un paramètre pertinent.
- Une raison invoquée pour l'utilisation de BF est la facilité de l'étude de la relation avec les formants de la voyelle. Nos études confirment vivement la corrélation entre le R2 de la voyelle et le BF pour [k]. Ce BF de [k] peut donc servir pour l'identification de la voyelle suivante mais nous cherchons la paramétrisation la moins sensible au contexte. Un fort lissage et surtout la limitation a priori de la zone de recherche de BF sont les méthodes couramment utilisées pour rendre BF plus robuste. La section "Fusion explicite" décrit nos travaux pour la mise en correspondance des diverses paramétrisation du bruit avec les formants de la voyelle.

Nous allons décrire brièvement les méthodes utilisées pour le codage des spectres. Nous conservons la même fenêtre temporelle et la même méthode de calcul de spectre (10ms Hamming asymétrique, LPC ordre 20, Fe=16 kHz).

- *Moments statistiques du spectre.* Cette méthode a été proposée par K. Forrest et collaborateurs en 1988 [FWMD88]. Chaque spectre est codé par l'énergie totale (ET) et les premiers quatre moments de la courbe de l'énergie fonction de la fréquence, à savoir L1 : la fréquence moyenne du spectre, L2 : la dispersion de la fréquence, L3 : l'asymétrie (skewness) et L4 : la compacité (kurtosis). Comme les auteurs le recommandent, nous avons utilisé une échelle temporelle linéaire (et non pas en bark) et des valeurs normalisées pour L3 et L4. Nous avons vérifié que ET et L2 n'apportent pas beaucoup d'information. L'étude citée souffre d'ailleurs d'un nombre réduit de contextes vocaliques.

– *Centroides d'ordre 3*. Les centroides, proposés par A. S. Crowe en 1988 [Cro88], sont une généralisation élégante du centre de gravité de l'aire encadrée par un axe et un graphique. Étant donné le nombre des centroids à estimer, un processus d'optimisation (descente du gradient) se charge de chercher une partition de l'axe des fréquences. Les centroides sont les centres de gravité des parties ainsi séparées. Ils sont caractérisés chacun par une fréquence et une amplitude. Nous avons utilisé le spectre entre 0 et 6 kHz. La position des centroides dépend de l'énergie moyenne. Nous les avons calculés avec et sans normalisation entre 0 et 1 de l'amplitude du spectre.

– *La transformée en cosinus discret* (DCT). Cette méthode est bien connue pour la compression de l'information. Le spectre, vu comme un vecteur de 256 réels, passe, à travers la DCT, en 256 amplitudes de composantes cossinusoïdales. En retenant les premiers et en appliquant la transformation inverse, on obtient une bonne approximation du vecteur de départ. Les premiers coefficients de la transformée codent donc bien tout le vecteur. Les coefficients DCT sont de même nature que les coefficients cepstraux.

– *L'énergie dans N bandes de fréquence*. Il s'agit en fait d'un spectre hautement schématisé (réduction d'au moins un ordre de grandeur de la résolution fréquencielle). Nous avons étudié l'influence de plusieurs facteurs :

  1. le nombre de bandes
  2. le type de découpage en bandes sur l'axe des fréquences.
  3. l'influence de prelissage (par l'ordre de l'analyse LPC)
  4. l'apport des indices temporels (par ajout de l'indice VOT)

  Nous avons utilisé toujours des partitions équidistantes de la zone 0 à 8 kHz sur une échelle linéaire (Hertz) ou sur une échelle psycho-acoustique des fréquences (échelle de Bark, [ZT80]).

À part le calcul de kurtosis, nous n'avons pas tenté aucune autre mesure scalaire de la compacité car le spectre de [t] peut souvent être compact et le [k] en contexte avant ne l'est pas. Nous avouons en fait un penchant pour les codage de la forme globale, les codages où toutes les parties du spectre contribuent, contrairement à la caractérisation classique des voyelles par les maximas spectraux.

Nous présentons (voir tableau 5.2) comme indice synthétique pour la comparaison le taux d'apprentissage (noté TA par la suite) du lieu d'articulation. Nous rappelons que le TA est une borne supérieure pour le taux de reconnaissance. Le TA est calculé ici en sortie d'une analyse discriminante canonique (ADC), donc une rotation suivi d'une classification bayesienne dans ce nouveau espace de paramètres. Nous avons systématiquement utilisé l'ADC malgré sa relative sensibilité aux écarts de la normalité dans les distributions et malgré sa tendance gênante à capitaliser sur le simple nombre de paramètres.

Une analyse des performances par consonne, par contexte, une analyse de erreurs, une analyse de l'ordre d'inclusion dans le modèle de même que des tests croisés de reconnaissance ont été effectués. Nous retenons pour le moment que les jeux étudiés sont homogènes : pour la performance par consonne, tous reconnaissent mieux le [p] et donnent comme confusions une majorité de [t] versus

Codage N-seg	nb. par.	TA
Moments statistiques	5	58.1%
Centroides (ordre 3)	6	70.9%
DCT (CD1..10)	10	72.7%
DCT (CD1..6)	6	65.9%
10 bandes Hertz (LIN1..10)	10	70.0%
20 bandes Hertz (LIN1..10)	20	77.7%
10 bandes Bark (BRK1..10)	10	74.7%
idem mais VOT rajouté	11	80.3%
idem mais ordre LPC réduit (12)	10	72.7%
BRK1, BRK3	2	70.2%

TAB. 5.2: Taux d'apprentissage pour différents codages du spectre du N-seg. BRK1 et BRK3 donnent la meilleure séparabilité pour deux paramètres. 2208 CV, 8 locuteurs

Codage	TA1	TR1	TA2	TR2
10 bandes Hertz	92.0%	87.7%	100%	85.5%
10 bandes Bark	84.6%	82.8%	100%	86.9%
DCT de 1 à 6	90.6%	86.2%	100%	83.3%

TAB. 5.3: Test d'apprentissage (TA) et de reconnaissance (TR) pour le premier locuteur. TR selon la méthode "leave one item out", 276 CV. TA1, TR1 par CG et TA2, TR2 par RN à 7 neurones dans la couche cachée. Il faut noter que la dégradation diférente entre TA et TR pour les deux méthodes.

[k] pour le contexte avant (10 DCT : 90%[p], 63%[k], 64%[t] ; Centroides : 87%[p], 57%[k], 67%[t] ; 10 bandes Hertz : 89%[p], 60%[k], 61%[t]). Ces jeux de paramètres peuvent apprendre aussi, dans une certaine mesure, la voyelle (TA de 24% à 29%, niveau aléatoire 10%).

Le biais en faveur de [p] semble être une caractéristique des codages du spectre du N-seg ignorant le contexte. Les travaux qui qui cherchent par expertise les indices dépendants du contexte tendent à favoriser [k], la consonne la plus riche en indices de ce type. Par exemple, Djezzar et Pican [DP97] trouvent comme taux de bonne identification 87% [p], 80% [t] et 75% [k] sans et 84% [p], 87% [t] et 93% [k] avec la prise en compte du contexte. Ils ont utilisé des perceptrons multi-couche et un jeux de paramètres formé de : VOT de Klatt, intensité absolue et relative de la barre d'explosion, fréquence des deux maxima plus importants du spectre de celle-ci (BF1 et BF2), les rapports BF1/F2 et BF1/F3 (F2, F3 mesurés à la frontière du V-seg) et, finalement, un indice scalaire de compacité spectrale.

Nous précisons que ces pourcentages sont calculés sur les 2208 items, c'est-à-dire avec les deux principales sources de variabilité présentes (10 voyelles, 8 locuteurs). Évidemment, l'absence d'une de ces deux sources fait monter en flèche les TA et diminue l'écart ou renverse l'ordre des différentes paramétrisations (illustration dans le tableau 5.3).

En conclusion, nous retenons le codage par l'énergie en dix bandes de fréquence (largeur 2,2 barks chacune) pour la suite de nos expériences. Multiplier le nombre de bandes apporte moins d'informations que le rajout des autres indices. Nous reviendrons sur la robustesse des performances dans la section dédiée à l'analyse des erreurs après avoir passé en revue plus de codages et de classifieurs.

5.4.2 Indices du V-seg

Pour le codage des segments voisés de parole, les formants sont les paramètres les plus répandus. Comme nous l'avons rappelé dans la section "Équation du locus", les transitions des formants ont été des candidats favoris pour le codage du lieu d'articulation pour les occlusives. Nous présentons notre méthode pour le calcul des trajectoires formantiques. Ensuite, nous comparerons plusieurs façons d'encoder la dynamique de ces trajectoires.

Les transitions formantiques sont considérées souvent a priori comme le meilleur paramétrage du V-seg. Nous avons voulu vérifier en la comparant à une paramétrisation différente du spectrogramme pour le même segment temporel. Nous avons choisi un codage par coefficients DCT temps-fréquence de la surface du spectrogramme. Cette comparaison justifie l'utilisation par la suite des transitions formantiques.

5.4.2.1 Méthode pour le suivi des formants

Malgré la supériorité évidente de l'extraction manuelle des paramètres, nous avons décidé d'utiliser une méthode automatique pour le calcul des trajectoires formantiques, principalement en raison de la taille de notre corpus. Nous présenterons ici notre méthode car elle est hautement représentative de notre approche sur l'algorithmique du traitement de signal.

5.4.2.1.1 État de l'art

Il convient de rappeler que la définition même des formants est disputée. Les formants ont été définis au départ comme des maxima du spectre vocalique :

> "The spectral peaks of the sound spectrum are called formants." [Fan60, page 20].

Ainsi définis, leurs estimation exacte est impossible car dépendante de facteurs comme la position de la fenêtre d'analyse vis à vis des périodes de la fréquence fondamentale, le degré de lissage du spectre, etc. Une définition plus rigoureuse des formants est en tant que fréquences de résonance de la fonction de transfert acoustique du conduit vocal :

> "Traditionnellement, dans le domaine de la parole, les termes de formant et de pôle sont employés de manière interchangeable ..." [Cal89, page 80]

Il existe plusieurs méthodes pour la mesure directe de la fonction du transfert du conduit vocal et, en conséquence, de ses résonances. Cependant, ces méthodes sont complexes et demandent une mise en oeuvre spéciale (excitation extérieure). Même les plus rapides comme temps de mesure imposent de contraintes trop lourdes pour l'étude de la parole en situation. Pour une comparaison entre méthodes et un exemple précis (la méthode des impulsions corrélées), voir [DGBP91]. Pour notre part, nous nous limitons donc au calcul des résonances de manière classique (à partir du signal vocal pré-enregistré).

Les méthodes utilisées pour le suivi des formants se séparent selon deux approches principales :

- **Méthodes spectrales**. La caractéristique commune de ces méthodes est l'analyse fréquencielle par trame. Elles comportent deux étapes souvent découplées :
  1. *L'analyse locale* est l'étape qui aboutie au calcul des candidats pour les formants dans chaque trame.
  2. *L'analyse dynamique* ou le suivi proprement dit est l'étape pendant laquelle les candidats sont reliés pour former une trajectoire.

  Ce découplage est un avantage car l'analyse locale est généralement très précise. L'inconvénient tient à la complexité de la deuxième étape.
- **Méthodes directes.** Ces méthodes ne passent pas par une analyse spectrale typique. Elles évaluent les paramètres d'un modèle autorégressif directement à partir du signal. Étant des méthodes récurrentes, la recherche des candidats et la construction des trajectoires sont simultanées. Leurs inconvénients sont liés au coût de calcul et au manque de précision. Elles peuvent travailler aussi par trame mais alors les candidats choisis dans une trame sont employes comme valeurs initiales pour la recherche des candidats dans la trame suivante.

La première étape des méthodes spectrales, l'extraction des formants, a généré une riche littérature. Nous nous limiterons ici à une énumération non exhaustive des principales solutions adoptées car nous considérons que même les plus basiques ont des performances suffisantes pour le but que nous poursuivons :

- racines du polynôme LPC [McC74, MG76]
- maxima du spectre lissé par le calcul du cepstre
- zéros de la fonction de retard de groupe [DYM89]
- centroïdes [Cro88]
- "Multiband energy demodulation" [PM95, PM96]. Cette méthode récente s'affranchi du modèle autorégressif. Elle utilise un jeux de filtres de Gabor pour fournir les fréquences et les bandes passantes des formants.
- méthodes coopératives (un vote combinant plusieurs solutions) [Soq95a, Soq95b]

Par contraste, le problème du suivi proprement dit a connu moins de solutions originales :

- S. McCandless [McC74] a proposé en 1974 une première solution. Elle est basé sur un nombre de règles heuristiques pour choisir les candidats d'une trame en fonction des candidats retenus dans la trame précédente. Une description détaillée d'une version améliorée peut être trouvée dans [MG76]. Des versions de cette méthode ont été utilisé par la grande majorité des autres travaux.

- Une méthode significativement plus élaboré a été proposé au CRIN par Y. Laprie [Lap89, Lap90, LB94]. Les améliorations concernent la recherche des conflit (donc une segmentation plus fine de la zone du suivi) et une optimisation globale (sur toute la longueur du suivi) dans la phase du lissage. Cette méthode peut être décomposé en trois étapes :
  1. Proposition des trajectoires. Un nombre de trajectoires élémentaires continues sont construites par détection des contours dans le spectrogramme. Ces trajectoires élémentaires sont alors étiquetées en termes de formants et un ensemble de connections possibles sont évaluées. Pour gérer l'explosion combinatoire, une note est affectée à chaque connection et les n-meilleures solutions sont retenues.
  2. Une régulation des trajectoires est effectuée par la méthode des contours actifs [Ber91]. En donnant de la rigidité aux trajectoires, elle ajuste le compromis entre leurs degré de lissage et leurs proximité par rapport au maxima spectraux.
  3. La décision finale est prise en réévaluant les notes de chaque solution proposée auparavant.
- Une famille d'algorithmes de suivi a été proposée par G. Kopec en 1985 [Kop85, Kop86]. Dans sa forme la plus simple, l'algorithme associe à un spectre une probabilité de présenter un certain formant (étape de détection) et une série de probabilités pour que ce formant se trouve dans une suite d'intervalles de fréquence (étape d'estimation). Les deux étapes sont implémentées, séparément ou conjointement, par des chaînes de Markov cachées (HMM) pour un formant ou pour un jeux de formants. À la suite de l'apprentissage des HMM, l'algorithme est capable de choisir pour une suite de spectres voisés donnée la séquence des fréquences des formants la plus probable. L'optimum ici est global, sur toute la séquence des spectres, non pas local à une paire de trames. La méthode présente les avantages et les inconvénients de l'apprentissage des HMM. L'avantage important est qu'il élimine tout critère heuristique de décision tout en fournissant des statistiques intéressantes sur les formants dans le corpus d'apprentissage. Mais la performance est fort dépendante de la qualité et de la dimension de ce corpus. D'autant plus que, une fois ce coûteux apprentissage fini, il n'y a plus aucune possibilité de modifier le seuil de détection ou la précision de l'estimation. Mais la présence du corpus d'apprentissage permet de chiffrer exactement l'erreur moyenne sur les trajectoires.
- G. Rigoll a proposé une série de méthodes directes. La première utilise un estimateur de Kalman pour évaluer les paramètres d'un modèle proche du modèle autorégressif appelé FLPC [Rig86, NCH94]. La deuxième améliore le coût de calcul en utilisant un estimateur "quasilinéaire" [Rig88]. À chaque instant d'analyse (même pour chaque échantillon de signal) une correction des valeurs des formants obtenu à l'instant précédent est calculée. Ainsi, pour un signal stationnaire, les valeurs des paramètres convergent, dans un processus itératif, vers les valeurs réelles des formants. Nos essais avec cet algorithme ont montré que le compromis stabilité/vitesse de convergence est hautement dépendant du type d'excitation du signal. Malgré trois méthodes pour l'adaptation de l'estimateur de Kalman aux signaux impulsionnels, nous n'avons pas obtenu de résultats satisfaisants.

– Enfin, M.J. Hunt propose en 1985 un algorithme pour la comparaison des spectres [Hun85] basé sur une anamorphose fréquencielle (Dynamic Frequency Wraping). Cet algorithme peut être utilisé pour mettre en correspondance les formants d'une trame avec ceux de la trame suivante dans un suivi.

Outre l'utilisation à bon escient de connaissances a priori, autre problème fondamental du suivi des formants est l'évaluation objective des résultats. La comparaison simple avec des suivi de référence est rarement utilisée. La principale raison en est le coût du suivi effectué à la main. Si un corpus de test et un autre d'apprentissage de dimensions conséquentes doivent être passé en revue par l'expert, l'utilité même du suivi automatique est mise en doute. Le plus souvent, l'évaluation est subjective ou faite sur un petit corpus de test. Pendant la conception de notre algorithme, nous avons cherché des solutions pour que les traces de l'évolution fournies par l'algorithme soient les plus informatives pour l'évaluation.

5.4.2.1.2 Réalisation

Les problèmes rencontrés dans l'implémentation d'un suivi de formants se résument à deux compromis fondamentaux :

– Le choix des candidats pour la fréquence des formants. Plus la résolution spectrale est fine moins les omissions sont nombreuses mais plus grand est le risque des erreurs par insertion.
– La rigidité des trajectoires. Plus la réactivité du suivi aux candidats observés dans la trame courante est grande, plus grand est le risque de "dérailler" de la vraie trajectoire. Mais, si cette réactivité est trop faible, des variations utiles peuvent être gommées et les chances de récupérer après un "accident" de suivi sont réduites.

Le deux cas représentent donc des exemples typiques du problème plus général du réglage du compromis observations - modèle tel que nous l'avons énoncé dans l'introduction. Notre solution devant un tel obstacle consiste dans une décomposition en deux phases du chaînage avant de l'algorithme :

– La première phase donne la priorité à l'observation directe. La sortie de cette étape est une "ébauche" proche de la sortie finale. Cette partie de l'algorithme sera réglée de manière à ce que cette ébauche soit très proche de la sortie attendue pour une majorité d'exemples, surtout pour les exemples que l'expert phonéticien considère comme "évidents".
– La deuxième phase introduit la connaissance du modèle. Elle peut être vue comme un filtrage des sorties de l'étape précédente. Cette partie de l'algorithme sera réglée de manière à ce que le nombre de dégradations soit plus réduit que le nombre des améliorations pour les données filtrées.

Pour le compromis du choix des candidats, cette dichotomie s'est traduite par une partition de l'ensemble des valeurs de fréquence candidates en deux sous-ensembles. Le premier sous-ensemble est fourni par un extracteur pessimiste (qui préfère les omissions aux insertions). Le deuxième sous-ensemble et la différence entre la sortie d'un algorithme optimiste (qui ne fait pas d'omission quitte à faire des insertions) et le premier sous-ensemble. Nous allons donc choisir la fréquence des formants dans le premier sous-ensemble et, si la sortie ne satisfait le modèle et seulement en deuxième recours, nous allons chercher dans l'autre sous-ensemble. Pour l'extraction des candidats nous avons retenu deux méthodes. Par ordre de performance, elles sont :

1. Les maxima du spectre à court terme lissé par prédiction linéaire. Dans ce cas, l'ensemble "pessimiste" de candidats est constitué par les vrais maxima et l'ensemble "optimiste" est obtenu en ajoutant les épaulements.
2. Les racines du polynôme approximant la fonction des transfert du conduit. Dans ce cas, l'ensemble "optimiste" de candidats est obtenu par analyse LPC d'ordre élevé (20 ou 24) et l'ensemble "pessimiste" est obtenu en éliminant les racines à bandes passante plus large qu'un seuil donné (par exemple, 200 Hz).

Pour le suivi proprement dit, le découplage annoncé a pris la forme de deux passages de la voyelle vers la consonne :

– Le premier passage donne priorité à l'observation. Il nous donne des trajectoires ordonnées relativement bruitées. L'algorithme pour relier la trame courante aux précédentes donne l'originalité de notre méthode et va être décrit plus en détail plus loin. À la sortie de cet algorithme à chaque point choisi une note de confiance est associée.
– Le deuxième passage donne de la rigidité aux trajectoires. Les variations plus rapides que celles que le modèle pourra raisonnablement produire sont éliminées. Nous utilisons un lissage pondéré par les notes de confiance suivi de l'application d'un modèle de trajectoire. Cette modélisation consiste dans une décomposition en cosinus discret pour laquelle nous gardons un nombre réduit de coefficients (3 sur 10).

Un aspect qui nous singularise parmi les solutions proposées est le traitement des focalisations. Même si les croisement de formants sont typiques des séquences VV, nos résultats montrent qu'ils sont présents également dans les séquences CV. Pour l'instant, l'étiquetage des trajectoires au centre de la voyelle est simplement propagé au long du suivi, sans être remis en cause. Une recherche spéciale pour la distribution des candidats dans les canaux est cependant effectuée lorsqu'une focalisation est détectée.

Le passage inter-trame est inspiré de celui décrit en [MG76]. Il est basé aussi sur le partage en canaux de l'échelle des fréquences de la trame courante par les derniers points choisis sur chaque trajectoire (voire figure 5.16). Nous construisons quatre trajectoires (la limite supérieure du plus haut canal est figée à 5 kHz). Un paquet de règles[5] se charge de choisir les points dans l'ensemble pessimiste ou optimiste selon la charge en candidats de chaque canal. La note de confiance maximale (voir tableau 5.6) est accordée si un et un seul candidat de l'ensemble pessimiste est présent dans un canal et si tous les canaux de fréquence inférieure ont reçu cette note maximale. La note minimale (0) est accordée si aucun candidat n'a pas été trouvé (aucun canal voisin surchargé, pas de candidat optimiste). Dans ce cas, un point est ajouté dans la trame courante à la hauteur du point qui le précède sur la même trajectoire (une trajectoire ne s'arrête pas, seules les notes nous indiquent sa

[5]Je n'utilise pas de vrai moteur d'inférence mais un environnement de programmation procédurale (Matlab). Par conséquent, l'ordre de parcours des règles est figé.

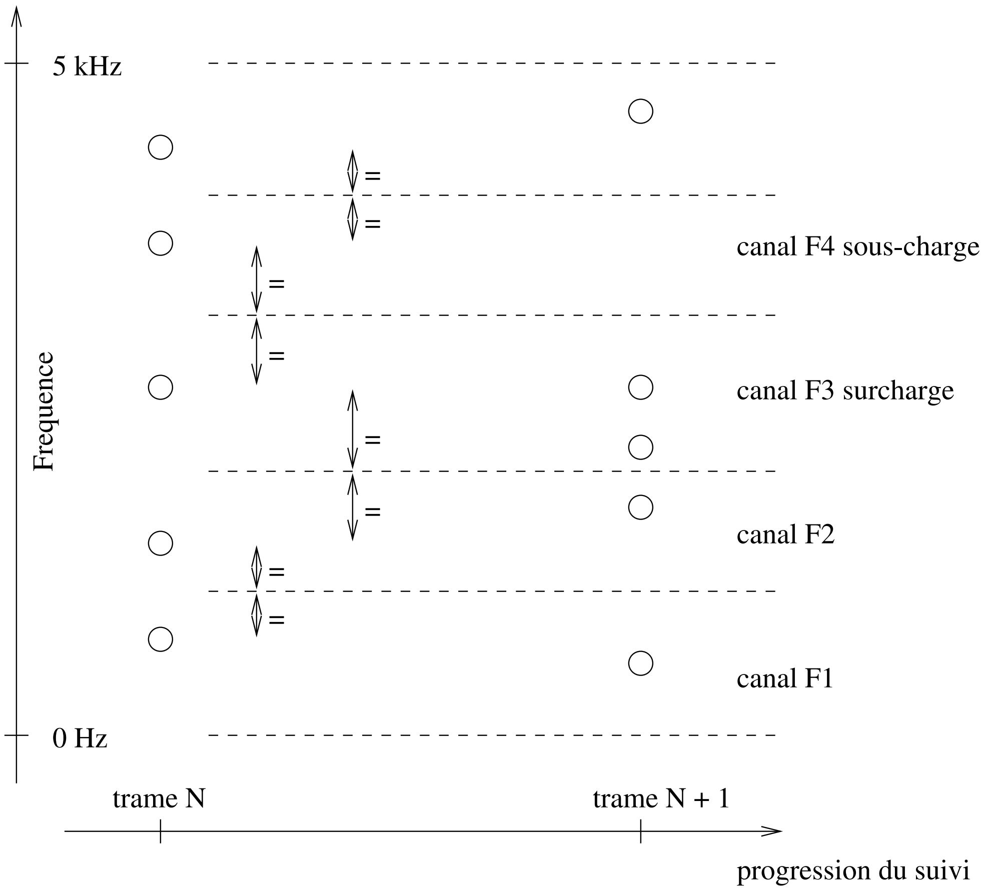


FIG. 5.16: Passage inter trame schématisé

disparition). Il faut préciser que les points qui génèrent la délimitation en canaux ne sont pas les points de la trame précédente mais le résultat d'une extrapolation linéaire des trajectoires calculées auparavant.

Une focalisation est détectée si ces extrapolations se croisent ou si deux trajectoires se rapprochent suffisamment. Le seuil pour ce rapprochement (62 Hz) et la durée prise en compte pour l'extrapolation (25 ms) sont les seuls paramètres du passage. Le suivi est cependant flexible : les canaux sont serrés quand les formants sont proches et sont larges quand les formants varient rapidement.

L'initialisation du suivi utilise le même algorithme de passage, mais ici les points de référence pour le partage en canaux sont les valeurs des premiers quatre formants du prototype de la voyelle (et la valeur fixe 5 kHz), valeurs stockées dans une base de données. Pour ce passage particulier, l'ordre de parcours des règles a été légèrement modifié pour favoriser la recherche en fréquence par rapport à la propagation des points de référence. En conséquence, c'est juste la connaissance de la classe de la voyelle qui est nécessaire à l'instant de l'initialisation. Par exemple, nous avons verifié que les trajectoires calculées sont identiques utilisant un seul patron formantique pour [u] [o] et [ɔ].

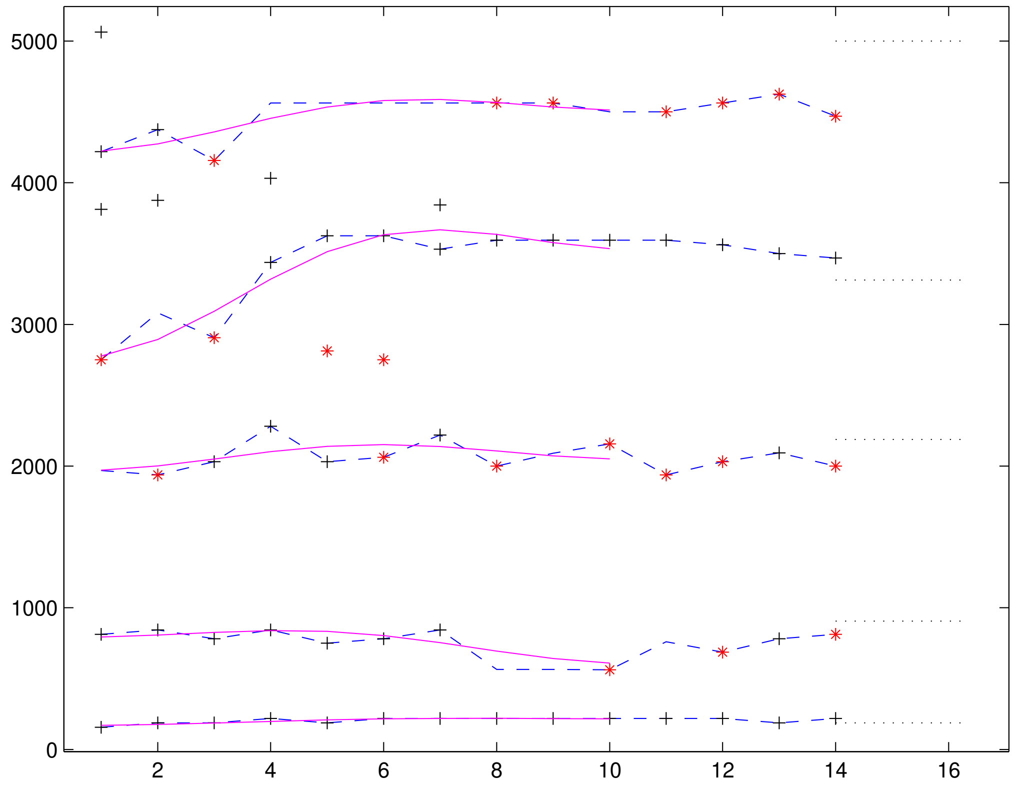


FIG. 5.17: Suivi de formants. Les candidats absents (ici sur F2) n'influence pas le suivi grâce à la prédiction linéaire. Syllabe [ku], locuteur PT, + : ensemble pessimiste de candidats formants, * : ensemble optimiste, trait interrompu : trajectoires des formants, trait plein : approximation par modèle de trajectoire, en pointillé : patron formantique de la voyelle (répété à droite pour visibilité)

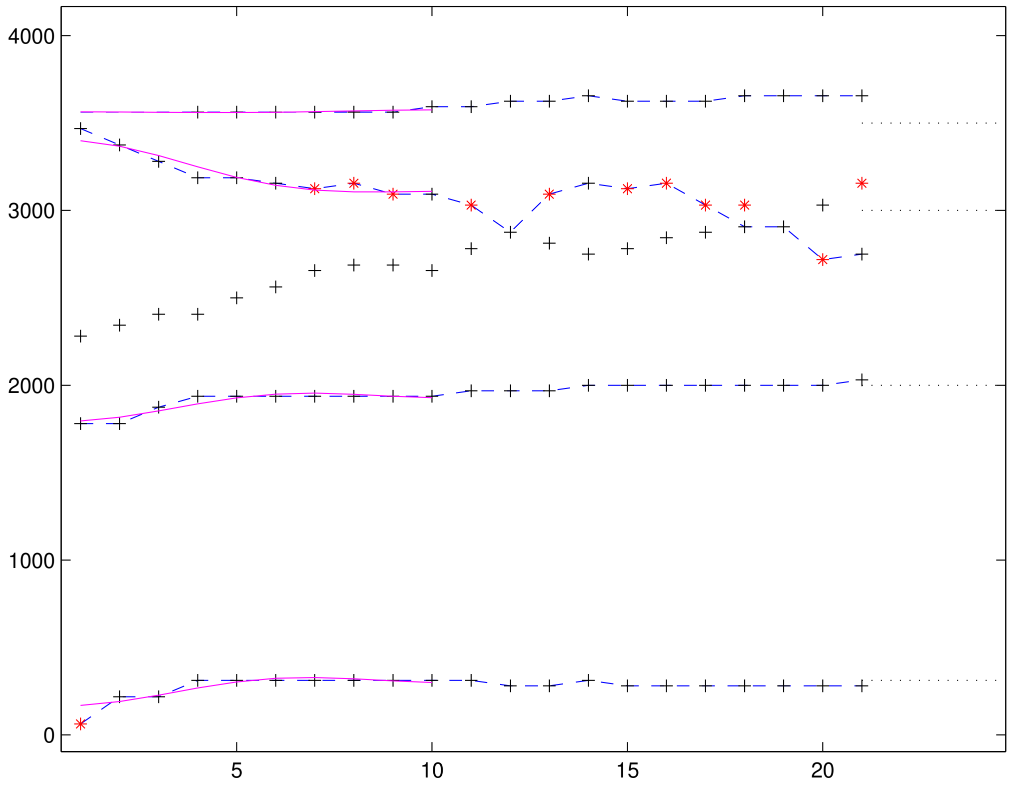


FIG. 5.18: Erreur de suivi. L'initialisation du suivi et l'abondance de candidats font rater la vraie trajectoire du F3. Syllabe [pi], locuteur PT, même représentation que la figure précédente. Nous aurions pu donner plus de poids au patron formantique de la voyelle pour éviter cette erreur. Cela aurait rendu l'algorithme plus sensible aux informations a priori sur la voyelle.

Nombre total de CV	2208
Nombre total de passages inter-trame	30356
Longeur suivi (nombre moyen de passages $\pm$ écart type)	13.75 ± 3.1
Focalisations rapportées	13
dont fausses alarmes ou mal traitées	4
Note initialisation (moyenne, min et max)	5.60 (de 3.5 à 6)
Initialisation ratées (une note 0)	7
dont récupérées au cours du suivi	7
CV avec alarmes de suivi (sur F1..3)	50 (2.2%)
dont vraies erreurs de suivi	10 (0.5%)
dont mauvaises initialisations non récupérées	2 (0.1%)
Note suivi par syllabe (moyenne)	5.76

TAB. 5.4: Statistiques sur le suivi automathique des formants

Statistique	F1	F2	F3	F4
Alarmes suivi (paires de notes 0 succesives)	4	18	28	73
Initialisation ratée (une note <=2)	0	3	1	3
Initialisation difficile (une note <=4)	16	35	73	963
Note par trajectoire (moyenne)	5.96	5.89	5.82	5.4
Note par trajectoire (minimum)	4.1	2.25	1.3	1.35
Note par trajectoire égale à 6	89.1%	67.3%	50.0%	11.1%
5<= note par trajectoire <6	10.7%	31.2%	47.0%	70.6%
4<= note par trajectoire <5	0.2%	1.3%	2.2%	14.7%
Trajectoires vraisemblablement ratées (note <4)	0%(0)	0.2%(4)	0.8%(18)	3.5%(77)

TAB. 5.5: Statistiques des notes de confiance par formant

5.4.2.1.3 Résultats

Parmi les avantages de notre algorithme, nous comptons sa complexité réduite. Il n'utilise pas un corpus d'apprentissage. Il satisfait au critère imposé lors de sa conception : ne pas faire d'erreurs dans les cas simples. Nous avons constaté qu'il est peu sensible aux erreurs d'initialisation. En effet, il est capable de se recaler au premier segment où les formants sont tous présents et réguliers.

Nous présentons comme indice de la performance une statistique sur les notes de confiance (voir tableaux 5.4 et 5.5). Une alarme de suivi est, pour nous, une paire de notes de confiance successives sur une trajectoire égales à zéro.[6] Nous avons vérifié en détail le suivi pour les 50 syllabes qui dechenchent des alarmes sur F1, F2 et F3.

[6]Les syllabes qui obtiennent les notes minimales (tableau 5.5, ligne 5) sont parmi celles qui dechenchent une telle alarme.

Candidat trouvé et tout candidat de fréquence inférieure trouvé	6
Candidat trouvé	5
idem mais candidat choisi dans l'ensemble optimiste	4
Candidat trouvé après réoccupation des canaux	3
idem mais candidat choisi dans l'ensemble optimiste	2
Focalisation détectée, candidat choisi par tentative de défocalisation	1
Candidat non trouvé, remplacé avec la valeur prédite	0

TAB. 5.6: Notes de confiance

En conclusion, l'algorithme est performant pour ce qui est des erreurs détectables. Les notes de confiance le confirment : les règles pour la réoccupation des canaux interviennent en moyenne une fois tous les cinq passages inter-trame. La note 0 apparaît dans des proportions négligeables sur le nombre total de passages inter-trames.[7].

En revanche, le manque d'informations sur les erreurs réelles (non détectables et/ou non détectées) est à déplorer. Notamment, nous estimons qu'un nombre important de focalisations ne soient pas détectées (surtout entre F1 et F2 dans les bilabiales en contextes postérieurs). D'un autre coté, toutes les alarmes de suivi ne sont pas des erreurs de suivi. Ainsi les 4 alarmes sur F1 n'en sont pas. Il s'agit de quatre situations où la crête du F1 apparait tard sur le spectrogramme par rapport à la position du marqueur VOT (>10 ms).

Les inconvénients de notre solution sont liés au modèle utilisé. L'impossibilité pour une trajectoire de s'arrêter ne permet pas des suivis au-delà de la frontière du noyau vocalique. L'algorithme est fonctionnel sur des segments courts autour du point d'initialisation. Nous n'avons pas testé l'algorithme avec de voyelles nasales et une dégradation sensible est à prévoir dans cette situation. En absence d'un corpus de référence, nous n'avons pas une mesure moyenne en Hertz de l'erreur de suivi.

Les améliorations de la méthode viendront, d'une part, de la prise en compte des autres informations sur les crêtes spectrales (amplitude, bande passante) et, d'autre part, de la comparaison avec une extraction manuelle des trajectoires sur un corpus comparable en variabilité.

5.4.2.2 Transitions formantiques prototypiques

L'analyse temps-fréquence de nos CV est basé sur un spectrogramme LPC avec des trames calculées toutes les 5 ms (description détaillée dans la section "Fusion précoce"). Le suivi des formants commence dans la trame la plus proche du marqueur du centre de la voyelle et s'arrête à la trame la plus proche du marqueur VOT qui est encore contenue dans le V-seg (progression dans le sens négatif du temps). Nous nous intéressons au 45 ms (10 trames) du début du V-seg. La trame du centre de la voyelle a été répétée (une ou deux fois) dans 25 cas (sur 2208) pour assurer cette longueur minimale. Le modèle de trajectoire a été appliqué sur cette fenêtre de 45 ms donc sur les 10 trames

[7] Proportions inférieures au pourcent sur chaque formant sauf sur F4

de la fin du suivi.

Nous présentons ici les prototypes des transitions formantiques pour les 45 ms suivant le déclenchement du voisement (voir figures 5.19 et 5.21) pour les trois consonnes et les quatre voyelles extrêmes. La moyenne a été faite pour les huit locuteurs, dans l'espace des paramètres du modèle de trajectoire. L'inspection des patrons des transitions soulève les remarques suivantes :

- Seule le contexte [a] possède des patrons spécifiques pour chaque consonne.
- Il n'y a pas de différence sensible entre [ti] et [ki] de même qu'entre les trois consonnes en contexte [y].
- Les erreurs de suivi (focalisations F1-F2 non détectées) ont donné un patron de la transition de F2 dans [pu] incorrect (le F2 aurait du continuer à descendre, de la voyelle vers la consonne, au point de se confondre avec F1). Il est possible que le même type d'erreur affecte aussi [py]. La focalisation reste donc un défi majeur pour les algorithmes de suivi.
- Malgré des écarts-type importants autour de ces prototypes (triangle vocalique et niveau sonore spécifique au locuteur), les différences relatives entre les patrons sont remarquablement bien conservés si la moyenne est faite sur les données d'un seul locuteur (voir figure 5.20)

Il est important de noter la comparaison entre les patrons des amplitudes des transitions dans les syllabes [ka] et [ta] (voir figure 5.22). Dans la trame de départ, l'ordre de l'amplitude des formants est : F1, F2, F3 et F4, comme pour toute voyelle centrale. Dans la trame voisine de la consonne, cet ordre est différent entre [t] et [k]. Si, pour le premier, l'ordre du centre de la voyelle est conservé, pour le second, l'ordre est F1, F3, F2 et F4 et même F1, F3, F4 et F2 pour le locuteur JS, ordre qui est typique du centre de la voyelle [i]. Ce changement d'amplitude dominante entre les formants est la preuve d'un changement d'affiliation majoritaire entre F2 et F3, même si les fréquences des formants ne sont pas très proches (conséquence d'une constriction faible). Six locuteurs (ET, FG, JS, LV, PT et YM) parmi les huit présentent ce changement dans [ka] et pas dans [ta] (voir figure 5.22). Un locuteur (JZ) présente un changement beaucoup plus marqué dans [ka] et un locuteur (GB) ne présente pas de différence. Comme les variations des fréquence ont le même sens dans [ta] et [ka], ce changement d'affiliation est un très bon candidat pour le trait binaire séparant les dentales des velaires. Il justifie aussi, dans le cas du locuteur JS, la bonne identification du lieu d'articulation par les auditeurs à partir du V-seg du [a], isolé ou non.

5.4.2.3 Équation du locus

Nous avons calculé pour ces 2208 transitions les paramètres du locus [SMM91], c'est-à-dire le profile de coarticulation pour chaque consonne à travers tous les contextes vocaliques dans l'espace des paramètres FC (fréquences des formants vers la consonne) et FV (fréquences des formants vers la voyelle). Malgré les différences de méthode (par exemple, 45 ms au lieu de 50 ms) et les rares erreurs du suivi (par exemple, [pu] peu coarticulé), nos résultats pour le français rejoignent ceux rapportés dans la littérature [SMM91, SHA93, Yeo95] :

- La consonne [t] est la plus forte concernant le pouvoir de coarticulation. C'est la consonne pour laquelle le concept antérieur du point du locus (point visé par la transition du F2) se

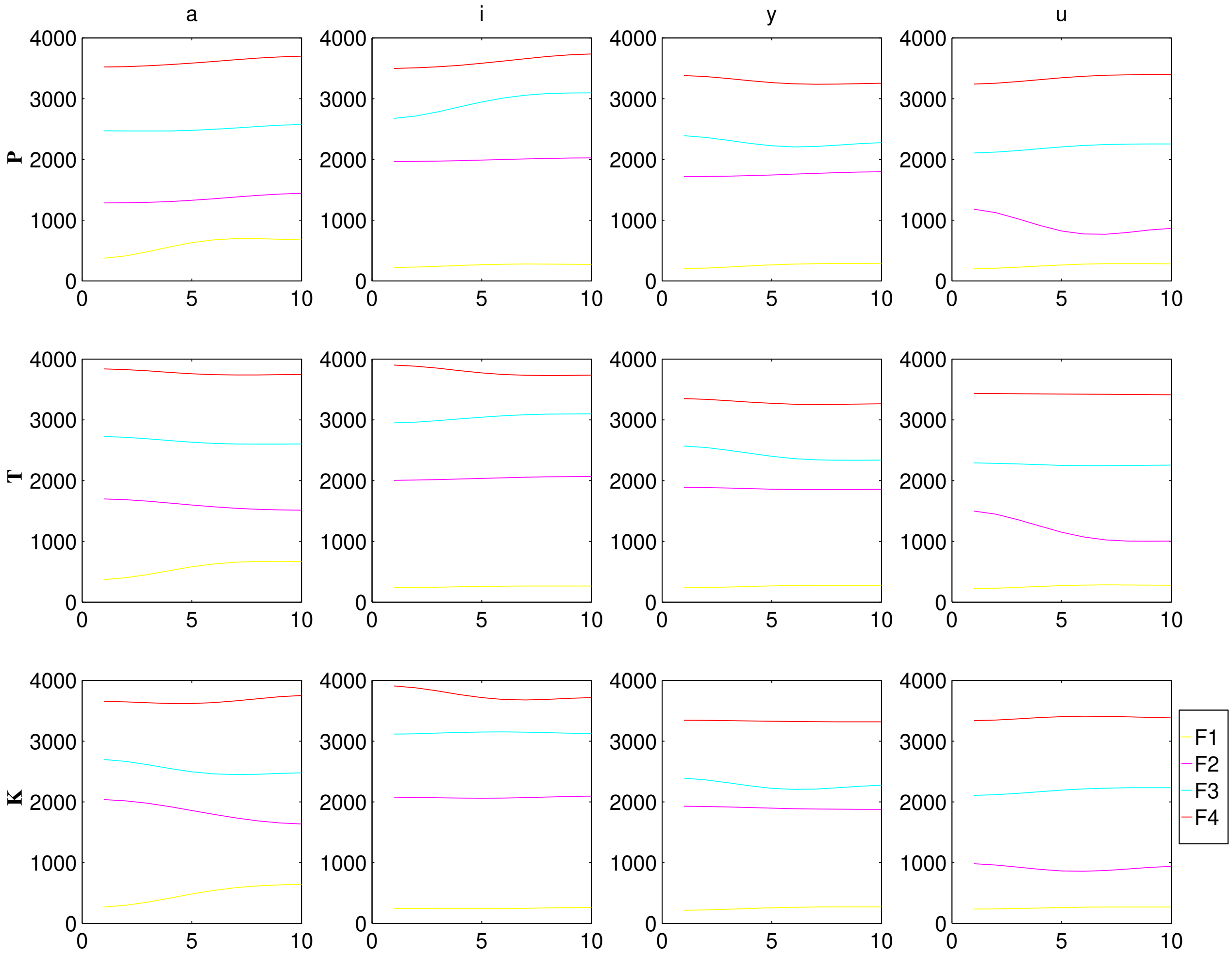


FIG. 5.19: Prototypes des fréquences des transitions formantiques. 8 locuteurs, 64 CV par case, longueur 10 trames (env. 50 ms)

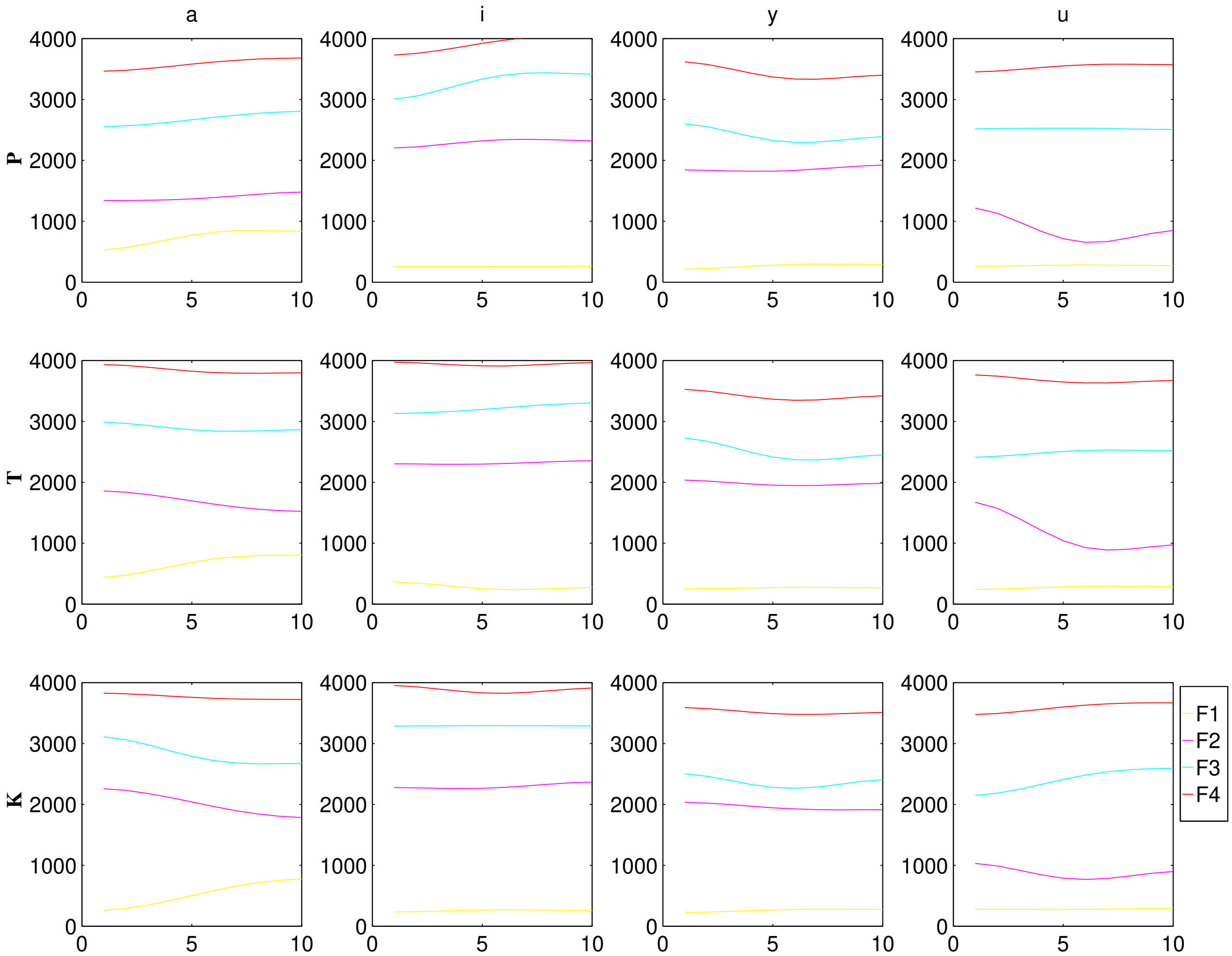


FIG. 5.20: Prototypes des fréquences des transitions formantiques. Locuteur JS

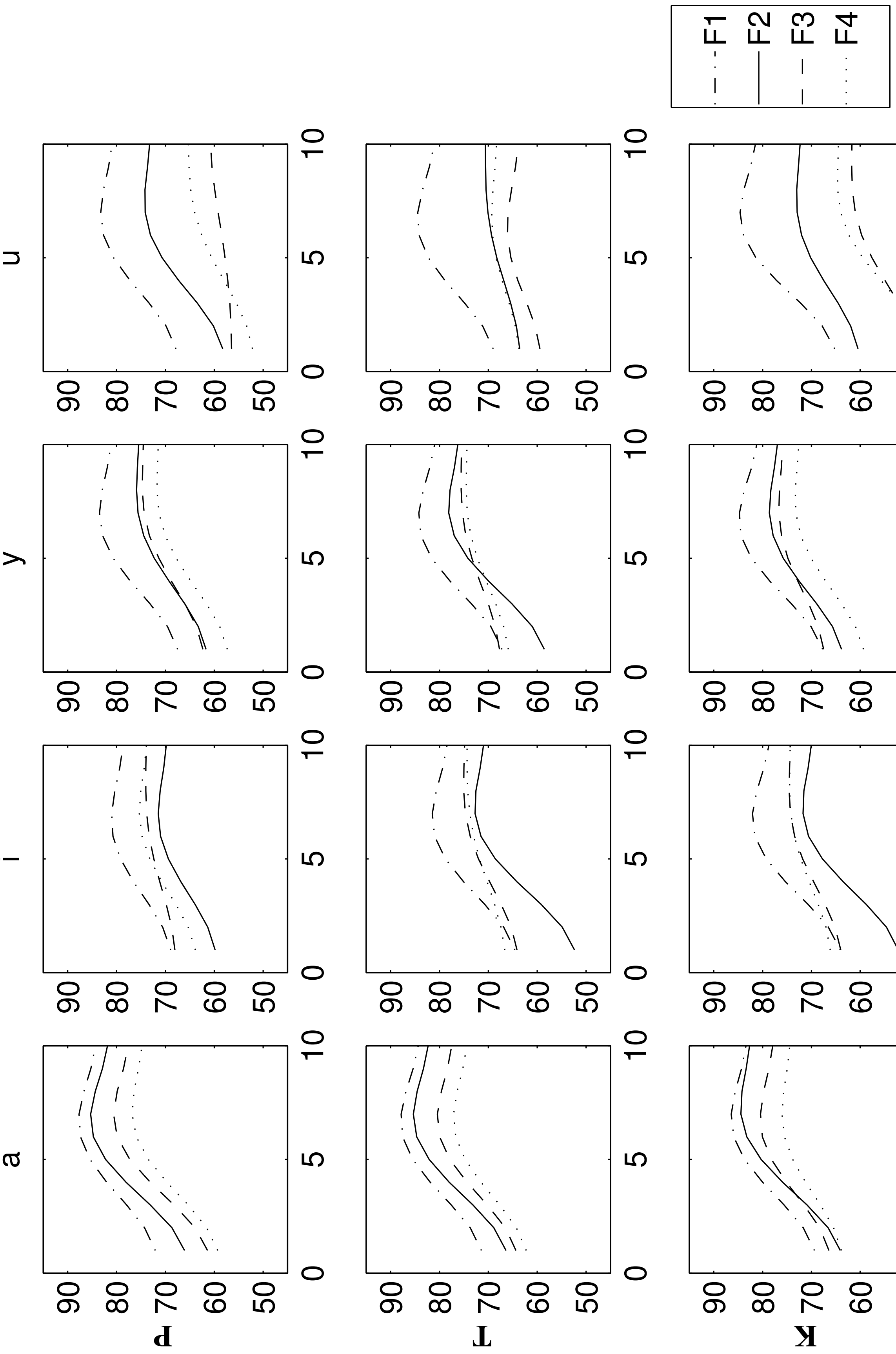


FIG. 5.21: Prototypes des amplitudes des transitions formantiques. 8 locuteurs, 64 syllabes par case, longueur 10 trames

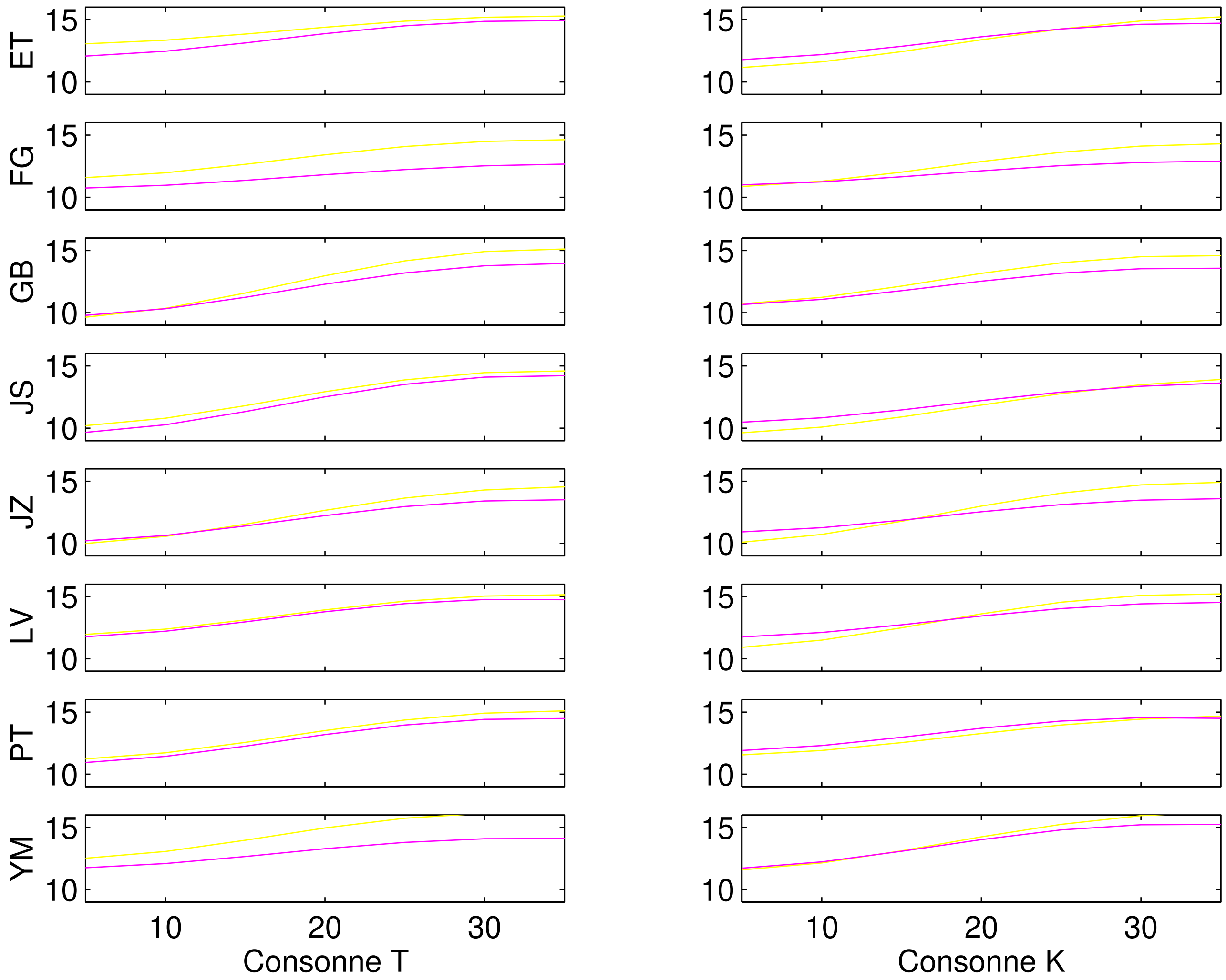


FIG. 5.22: Détail des prototypes d'amplitude pour les syllabes [ta] et [ka] faisant apparaître les changement d'amplitude dominante entre F2 et F3 pour 8 locuteurs. Abscise en milisecondes, F2 en pointillé.

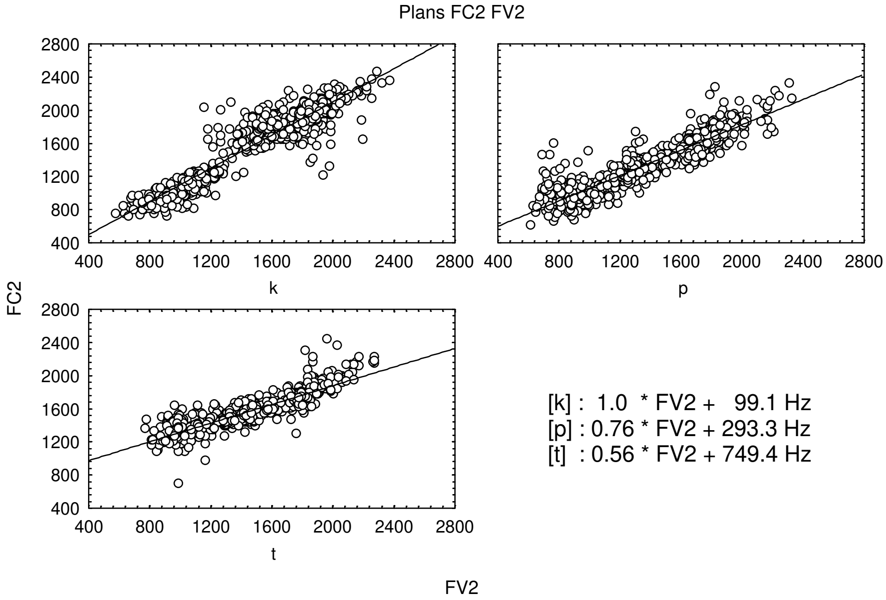


FIG. 5.23: Équation du locus pour les trois occlusives pour le deuxième formant (736 CV par case)

vérifie le mieux (voir figure 5.23).

- Comme dans les autres travaux, le locus du [k] présente une discontinuité évidente suggérant deux degrés de coarticulation selon la catégorie du contexte vocalique : faible influence de la consonne sur les voyelles postérieures et labialisées et forte influence sur les voyelles centrales. Cependant, le plan FC3, FV3 n'a pas présenté la rupture complémentaire dans nos résultats.

- À la différence des autres travaux, dans notre corpus, la droite de régression pour [p] a une pente moins prononcée que celle de la droite du [k].

- Les équations du locus présentent une variabilité interlocuteur relativement faible (voir figure 5.24) pour une méthodologie de calcul donnée.

- Les matrices de corrélation des FC et FV nous montrent des coefficients de corrélation remarquables pour les paires FC2, FV2 et FC3, FV3 (voir tableau 5.7). Ces coefficients justifient le choix de la caractérisation des occlusives dans les plans FCi, FVi.

En conclusion, les équations du locus apportent un argument de poids pour élaborer une théorie de l'invariance acoustique. Elles montrent une caractéristique cohérente à travers tous les contextes vocaliques du lieu d'articulation malgré la nature coarticulé des occlusives.

Néanmoins, leurs utilité pour le décodage acoustico-phonétique est réduite. Dans un système de reconnaissance, chaque réalisation d'une consonne est accompagnée d'un seul, tout au plus deux contextes. Dans leurs formulation classique, les locus se croisent (voir figure 5.25). Ils perdent leurs capacité de discrimination pour certaines contextes (voyelles antérieures, labialisées ou non).

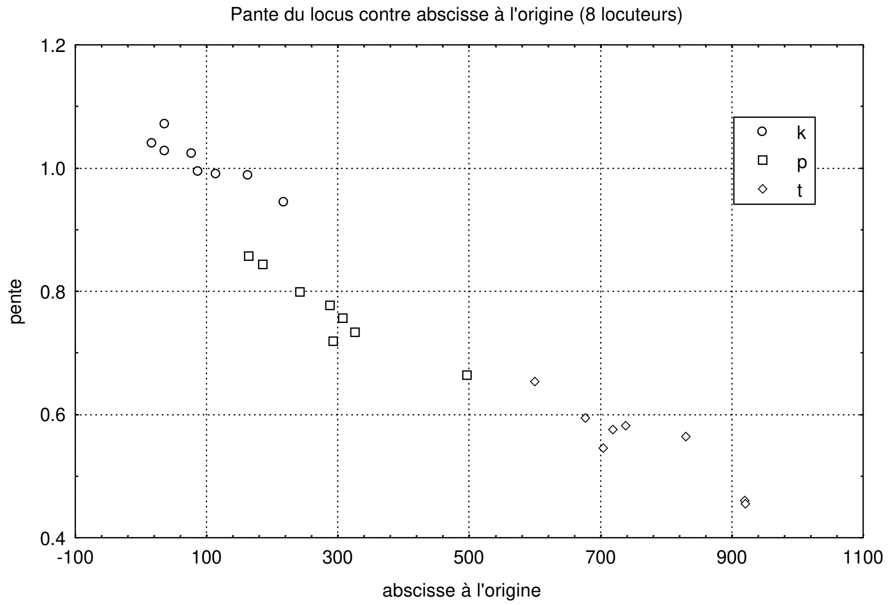


FIG. 5.24: Dispersion des droites de régression : 8 locuteurs, 3 occlusives.

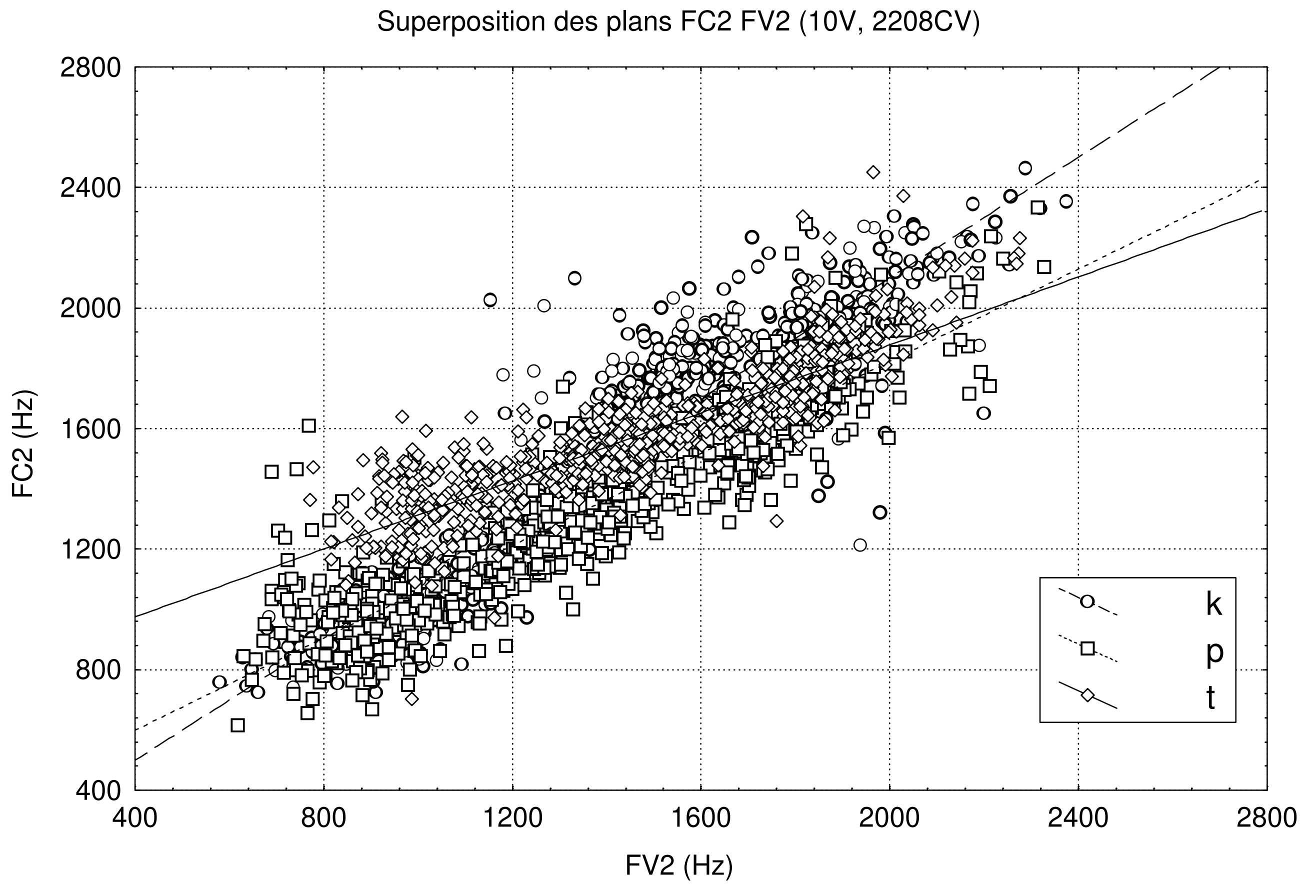


FIG. 5.25: Superposition des plans FC2 FV2 pour les trois occlusives

[p]	FC1	FC2	FC3	FC4
FV1	0.69	-0.24	0.10	0.23
FV2	0.16	**0.91**	0.39	0.19
FV3	0.19	0.51	**0.71**	0.40
FV4	0.35	0.28	0.43	0.57

[t]	FC1	FC2	FC3	FC4
FV1	0.78	-0.13	0.17	0.32
FV2	-0.06	**0.88**	0.60	0.42
FV3	0.11	0.60	**0.83**	0.66
FV4	0.32	0.37	0.52	0.77

[k]	FC1	FC2	FC3	FC4
FV1	0.58	0.13	0.05	0.15
FV2	-0.5	**0.93**	0.72	0.41
FV3	0.04	0.53	**0.87**	0.62
FV4	0.25	0.42	0.59	0.69

TAB. 5.7: Corrélations entre les fréquences des formants aux extrémités de la trajectoire

5.4.2.4 Paramétrisation des transitions formantiques

Dans ce paragraphe, nous nous proposons de chercher un schéma de codage pour les transitions formantiques (TF) et de comparer sa puissance de discrimination avec un codage alternatif du V-seg.

Pour le codage des TF, nous avons testé plusieurs paramètres possibles :

1. FC1..4, AC1..4 : les valeurs de fréquence et amplitude pour les formants à l'extrémité consonantique de la trajectoire
2. FV1..4, AV1..4 : les valeurs de fréquence et amplitude pour les formants 45 ms après le déclenchement du voisement.
3. DF1..4, DA1..4 : les différences entre FVi et FCi et respectivement entre AVi et ACi.
4. DAC2..4 : Les différences entre AC2..4 et AC1
5. DCT_F1..4 : Les paramètres du modèle (3 par trajectoire) pour la transition des fréquences des 4 formants
6. DCT_A1..4 : Les paramètres du modèle (3 par trajectoire) pour la transition des amplitudes

Une sélection réduite de taux d'apprentissage est présentée dans le tableau 5.8. Nous remarquons que :

1. FVs (fréquences des formants à 45 ms du VOT) contientent encore une (faible) information sur la consonne
2. L'apport du quatrième formant est réduit
3. Contrairement à nos attentes, les amplitudes (AC, AV, DA, DAC) n'apportent pas un gain sensible
4. DF présentent un bon TA par rapport au nombre de paramètres.
5. Le codage direct des TF (rassemblement des valeurs de début et fin) est comparable avec le codage par modèle de trajectoire pour un nombre de paramètres donné.

Nous savons que le codage par formants est un bon codage pour identifier le lieu d'articulation d'une voyelle. Nous avons voulu vérifier si ce type de codage reste aussi performant vis à vis de l'information consonantique. Pour cela nous avons utilisé un encodage global de toute la surface du spectrogramme qui nous a servi pour le calcul des TF. Le spectre de chaque trame (0 à 6 kHz) a été réduit à 7 coefficients par transformé en cosinus discret. L'évolution temporelle de chacun de ces 7 coefficients a été approximé par un modèle de trajectoire similaire au modèle appliqué aux formants. Nous obtenons un codage en 21 coefficients (7 fréquenciels fois 3 temporels). Ce codage, la DCT temps fréquence, est connu pour pouvoir restituer une version lissée du spectrogramme d'origine [NZ91]. Les résultats de la comparaison avec le codage du TF sont présentés dans le tableau 5.9. Nous retenons les conclusions suivantes :

1. les deux codages montrent des capacités proches pour la consonne
2. les formants sont mieux adaptés à l'identité de la voyelle (voir aussi [Soq95b]).
3. l'échelle de Bark agit différement sur les deux types de codage. Elle fait croître la capacité du codage global et décroître celle des formants, pour la consonne comme pour la voyelle, mais pas de manière significative.

Les algorithmes de construction d'arbres de décision ont donné des résultats comparables. Nous retenons qu'ils ont préféré F3C, DF2 et DF3 comme paramètres discriminants (choisis en tête des arbres dans cet ordre par les algorithmes décrits en [Mun96]).

Les taux sont en général moins bons que pour les codages du N-seg (sur le V-seg, environ un tiers des CV sont mal classées par rapport à un quart des CV, sur le N-seg). Pour les raisons exposées dans la section "Invariance absolue contre invariance relative", ce manque de séparation ne nous permet pas de tirer la conclusion qu'il y a moins d'information dans le V-seg. Soit cette information n'est présente que dans les contextes ouverts, comme le plaident nos tests de perception, soit, plus probablement, nous n'avons pas encore trouvé la bonne façon d'encoder le V-seg pour nos classifieurs. Une partie de l'échec est dû à l'extraction automatique des paramètres. Mais nous préférerons ces résultats aux taux, vraisemblablement meilleurs, qui auraient pu être fournis par l'extraction manuelle des trajectoires formantiques (par exemple, TR =90.6%, CG, "leave one item out", 6 paramètres, un seul locuteur).

En conclusion, nous retenons pour les expérimentations de fusion le codage des V-seg par les transitions formantiques de ses premiers 45 ms. Nous avons choisi pour chacun des premiers trois formants 3 coefficients pour décrire l'évolution de sa fréquence et autant pour son amplitude (DCT_F1..3 + DCT_A1..3 = 3*3+3*3=18 paramètres).

5.4.3 Fusion des informations du N-seg et du V-seg

Comme nous l'avons déjà précisé, la séparation du continuum de la syllabe dans deux segments, N-seg et V-seg, est due en partie au modalités typiquement différentes de codage pour ces signaux. Cependant, cela **ne signifie pas** que les auditeurs humains comparent des N-seg avec des prototypes de N-seg et des V-seg avec de prototypes de V-seg. Dans le processus courant de la perception de la parole, les auditeurs comparent des sons, des unités acoustiques où les deux segments contribuent

Paramètres	Nombre	TA consonne
Niveau aleatoire	0	33.3%
FV1..4	4	42.5%
FC1..4	4	49.6%
DF1..4	4	52.1%
FV2..3, DF2..3	4	56.5%
FC1..3, FV1..3	6	59.3%
FC1..4, FV1..4	8	61.5%
AC1..4, AV1..4	8	52.3%
FC1..4, DA2..4, DAC2..4	10	62.7%
FC1..4, FV1..4, AC1..4, AV1..4	16	67.0%
DCT_F1..3	9	61.1%
DCT_F1..3, DCT_A1..3	18	**69.8%**
DCT_F1..4, DCT_A1..4	24	71.9%

TAB. 5.8: Comparaison des codages des transitions formantiques. 2208 CV

Paramètres	Nombre	TA consonne	TA voyelle
DCT_Sona (Hertz)	21	71.9%	75.7%
DCT_Sona (Bark)	21	74.7%	82.0%
DCT_F1..3 (Hertz)	9	61.1%	81.6%
DCT_F1..3 (Bark)	9	60.4%	81.2%
DCT_F1..3, DCT_A1..3 (Hertz)	18	69.8%	85.1%

TAB. 5.9: Comparaison des paramétrisations globale et formantique du spectrogramme du segment voisé

à un percept unique. Il est donc naturel de tenter la récupération de ce percept à partir des deux segments.

Nous allons classer nos résultats sur les différents jeux de paramètres selon le niveau de la fusion dans le chaînage ascendant du traitement pour la classification phonétique :

– La fusion de plus bas niveau est obtenu quand les paramètres primaires du signal (un spectre, par exemple) sont calculés à travers les deux segments. C'est la fusion "naturelle" du point de vue perceptif. Nous l'appellerons **fusion précoce**.
– La fusion la plus simple, du point de vue des systèmes automatiques, est le calcul des paramètres sur les deux segments séparément et leur présentation conjointe à l'entrée du classifieur. Nous l'appellerons **fusion directe**.
– La fusion de plus haut niveau suppose qu'une décision phonétique est prise à partir de chaque segment. Les deux résultats de décision sont ensuite fusionnés pour aboutir à un seul phonème. Nous l'appellerons **fusion tardive**.

5.4.3.1 Fusion précoce. Spectre statique

Nous avons essayé deux méthodes de codage qui font fi de la distinction voisé/non voisé. Comme les codages à résolution grossière s'y prêtent mieux, il s'agit de codages de type intégré.

Le premier est un codage statique, il néglige les possibles évolutions à l'intérieur de la période analysée. Il s'agit, dans notre cas, d'un seul spectre calculé sur une fenêtre asymétrique à partir de l'instant de relâchement. La seule différence entre ce codage et le codage que nous avons étudié pour le N-seg est la durée de cette fenêtre. La représentation de chaque spectre par l'énergie en dix bandes de fréquence d'une largeur identique (sur une échelle psycho-acoustique) a été donc gardé. Nous avons utilisé la largeur de fenêtre proposé par S. Blumstein : 26 ms. Cependant, étant donné la différence entre le VOT moyen de /b d g/ en anglais et celui du /p t k/ en français, nous avons testé aussi une durée plus longue afin de préserver l'esprit de ce codage (spectre N-seg pur pour un nombre réduit d'items). Nous présentons donc aussi le taux d'apprentissage pour une durée de 40 ms (voir tableau 5.10).

Nous constatons que, en ce qui concerne la reconnaissance du lieu de la consonne, le taux pour 26 ms est meilleur que celui pour 40 ms et que les deux sont supérieurs à celui pour 10 ms. Même si les différences entre ces taux sont faibles (pour 26 ms, environ 110 syllabes correctement classés de plus que pour 10ms), le calcul du spectre du début de la voyelle en incluant une portion voisée améliore le pouvoir de discrimination de ce spectre. Nous retiendrons que l'ajout du paramètre VOT au spectre intégré sur les 10 premiers milisecondes apporte une amélioration toute à fait comparable.

5.4.3.2 Fusion précoce. Évolution spectrale

Pour rendre compte de la dynamique spectrale au début de la syllabe nous avons utilisé le principe de codage proposé par Z. B. Nossair et S. A. Zahorian en 1991 [NZ91]. Il s'agit, comme nous l'avons décrit déjà dans l'état de l'art, d'une méthode de lissage d'un ensemble de coupes spectrales

Codage	nb. par.	TA	par consonne
10 ms - 10 bandes (Bark)	10	74.7%	k :68 t :67 p :90
26 ms - 10 bandes (Bark)	10	79.2%	k :77 t :70 p :91
40 ms - 10 bandes (Bark)	10	78.3%	k :74 t :71 p :89
10 ms - 10 bandes (Bark) + VOT	11	80.3%	k :73 t :78 p :90

TAB. 5.10: Taux d'apprentissage pour différents codages du spectre intégré. 2208 CV, 8 locuteurs

Codage	nb. par.	TA
DCT_sona (Hertz)	21 (7*3)	80.8%
DCT_sona (Bark)	21 (7*3)	85.6%
DCT_sona (Hertz) + VOT	22	83.7%
DCT_sona (Bark) + VOT	22	86.2%
DCT_sona (Hertz)	45 (9*5)	89.4%
DCT_sona (Bark)	45 (9*5)	91.9%

TAB. 5.11: Taux d'apprentissage pour différents codages du spectrogramme. 2208 CV, 8 locuteurs

(calculées par LPC) par une double transformé en cosinus discret (DCT), d'abord sur l'axe des fréquences et ensuite sur l'axe du temps.

Par rapport à la méthode de Nossair et Zahorian, nous avons choisi une durée relativement courte (50 ms). Le choix de cette longueur d'analyse, plus courte que le plus longue des VOT, se justifie de deux façons. D'une part, cette longueur a demandé déjà la répétition de la trame situé au centre de la voyelle pour une dizaine de CV avec un noyau vocalique trop court. D'une autre part, selon les tests des auteurs [NZ91, page 2986], l'amélioration au-delà de cette durée est négligeable en anglais.

Ce codage a été clairement le plus performant sur notre corpus, même si nous n'avons pas atteint les scores rapportés par Z. B. Nossair et S. A. Zahorian. Nous avons voulu étudier l'influence du type d'anamorphose de l'échelle des fréquences et du rajout des indices fins (voir tableau 5.11). Nous concluons :

– L'échelle de Bark a une influence toujours positive sur la capacité de discrimination de la consonne pour ce type de codage.
– Malgré le fait que le spectrogramme code de manière implicite le VOT (par l'apparition du F1), le rajout du VOT apporte une amélioration (avec un meilleur rapport gain/nombre de paramètres qu'une résolution plus fine).

Nous avons essayé la capacité d'apprentissage pour différents résolutions en temps et en fréquence (voir figure 5.26). Malgré la croissance continue des taux, nous nous sommes arrêtés à 45 coefficients, un nombre déjà élevé considérant le rapport entre le nombre d'items dans le corpus et le nombre de paramètres mémorisé pendant l'apprentissage du classifieur. Comme attendu, plus l'ordre des coefficients est grand plus faible est leur apport pour la capacité de discrimination. Cependant, nous avons remarqué une exception : les coefficients du troisième ordre sont plus discriminants que

Codage	nb.par.	TA	TR
DCT_sona (Bark)	21	85.94% ± 0.6	83.23% ± 5.1
DCT_sona (Bark)	45	97.78% ± 1.5	87.95% ± 4.5

Appris sur	*Testé sur*	
	Corpus normal	Corpus hyper
Corpus normal	91.9%	86.1%
Corpus hyper	90.4%	93.3%

TAB. 5.12: Taux de reconnaissance et d'apprentissage sur 2208 syllabes. Partition par locuteur (en haut, moyenne ± écart type) et par degré d'articulation (en bas, codage DCT_sona1..45).

Nb. par.	5	10	15	20	25	30	35	40
TA	35.9	37.2	38.2	38.4	39.5	39.9	40.2	40.7

TAB. 5.13: Taux d'apprentissage selon le nombre de paramètres pour une distribution aléatoire des données (1 exemple = 0.05%)

les autres sur l'axe temporel (voir figure 5.27).

Nous avons mené des tests de reconnaissance utilisant plusieurs découpages tests/apprentissage. Nous avons utilisé une partition en huit parties correspondant aux locuteurs et une partition en deux parties correspondant aux degrés d'articulation (voir tableau 5.12). Notre meilleur résultat est TR=88%. Nous constatons que plus le TA est proche de la limite de 100% plus importante est la dégradation du TR par rapport à TA, syndrome de sur-apprentisage.[8] Nous trouvons que les consonnes hyper-articulées présentent moins de variance que les consonnes prononcées avec un débit normal, tout au contraire des voyelles du même corpus. Seuls deux locuteurs ont des performances en dessous de la moyenne (TR = 88%) : PT avec 80.1% et JS avec 82.6% ce qui met en doute le choix de ce dernier locuteur pour fournir les items des tests de perception.

Comme l'analyse discriminante incrémentale capitalise sur le nombre de paramètres, nous avons testé notre classifieur avec des données aléatoires. Pour ce faire, nous avons utilisé une matrice de 1998 lignes et 40 colonnes avec des nombres aléatoires fournis par un générateur respectant une distribution uniforme dans l'intervalle [0,1]. En utilisant les colonnes de cette matrice, nous avons calculé le taux d'apprentissage pour trois classes. Le TA ne dépasse jamais 41% (voir tableau 5.13). La croissance continue du TA sur nos données n'est pas artefactuelle.

Nous avons mesuré également le pouvoir de discrimination des métriques proposées par Lahiri. Pour cela, nous avons utilisé le spectre entre 200 Hz et 6 kHz et une méthode de régression pondérée (une régression favorisant les maxima spectraux, méthode suggérée par Smits [StBC96b]). Les paramètres que nous avons calculés sont :

– Les deux coordonnées du plan proposée par Lahiri [LGB84], c'est à dire C-A et D-B (voir

[8] pour le meilleur TR, le taux d'erreur augmente de plus de cinq fois entre TA et TR.

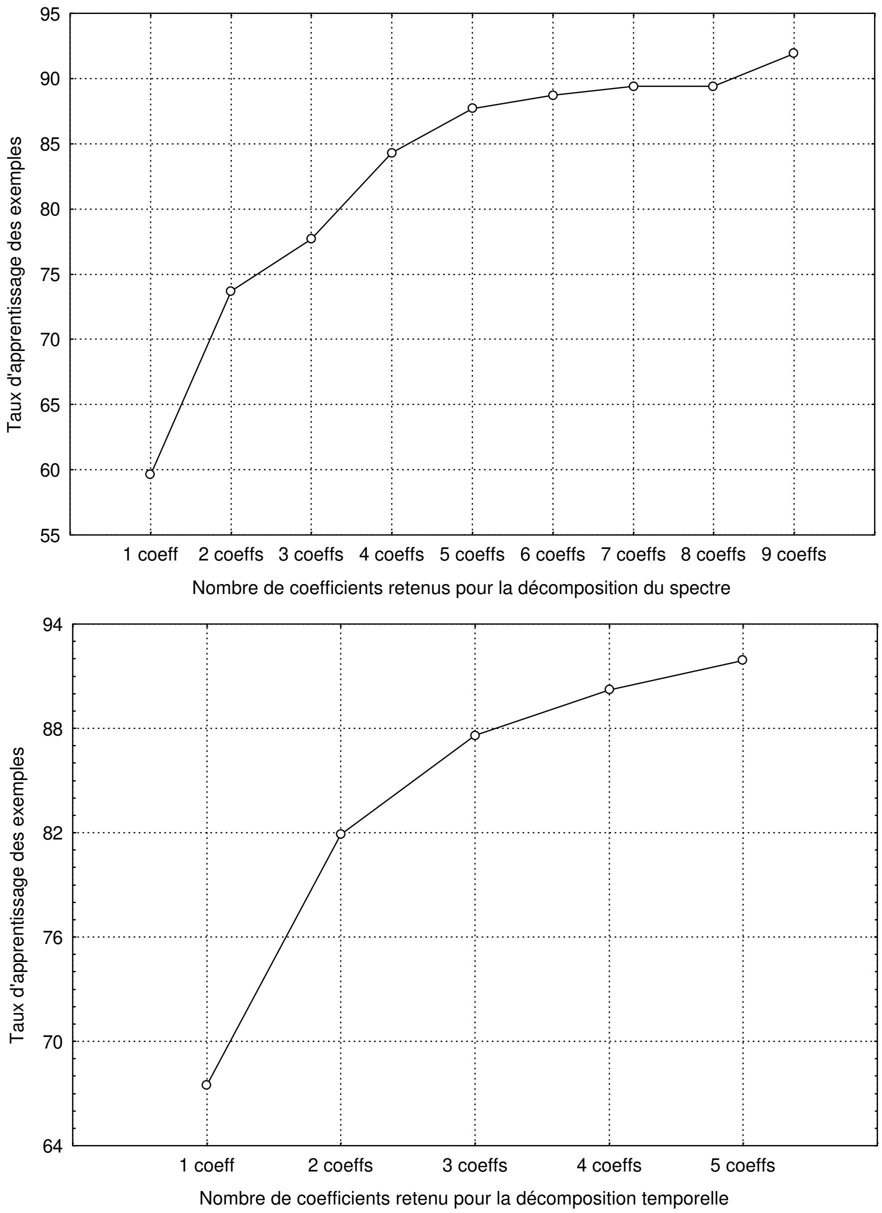


FIG. 5.26: Influence de la résolution a) en fréquence et b) en temps de la transformée DCT. (Taux d'apprentissage en fonction du nombre de coefficients choisi)

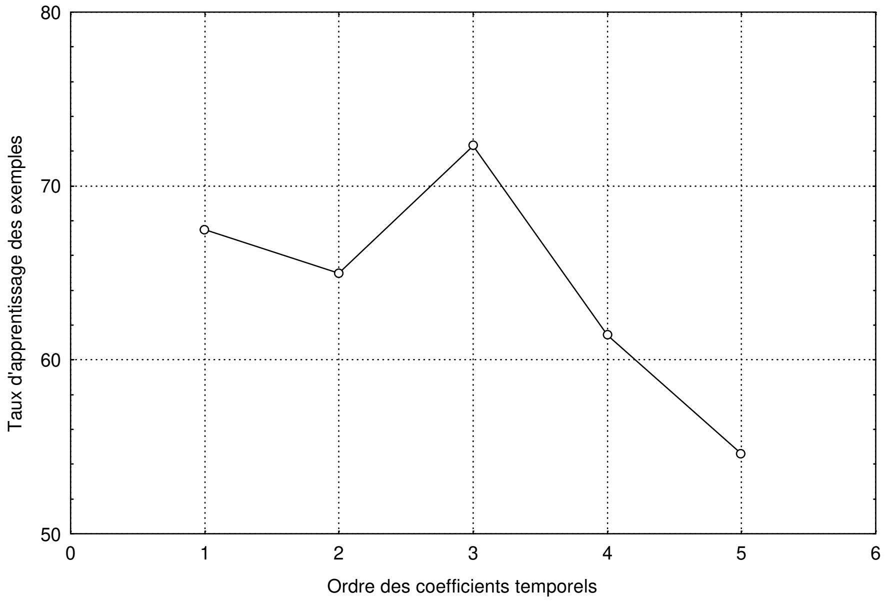


FIG. 5.27: Contribution individuelle des coefficients temporels de la transformation DCT (TA en fonction de l'ordre des coefficients retenus).

aussi la figure 3.3 dans le troisième chapitre).

- Pente et abscisse à l'origine de la droite de régression pour le spectre dans la première trame (P et Y0).
- Pente et abscisse à l'origine moyenne sur le spectrogramme (P_m et Y0_m).
- Variation de la pente au long d'une portion du spectrogramme (Delta_P).

La métrique de Lahiri suppose deux instants de temps pour le calcul de la variation spectrale : le moment du relâchement et un moment choisi dans le segment voisé. Nous avons considéré deux possibilités pour ce dernier :

- D'abord nous avons pris le deuxième repère temporel au début du V-seg, à la manière de Lahiri.
- Deuxièmement, nous avons pris le deuxième repère temporel à une distance fixée par rapport au moment du relâchement (Smits suggère 50 ms).

Les jeux de paramètres obtenus ont montré un même comportement : une bonne aptitude à séparer les labiales des dentales mais une difficulté certaine à apprendre les autres contrastes. Nous présentons comme exemple la matrice de confusion pour un jeu de paramètres dans le tableau 5.14. Malgré le fait que les confusions [t] versus [p] représentent seulement 4.6% des exemples dans ce tableau, les confusions de [k] et vers [k] sont trop nombreuses. Une amélioration de cette métrique est envisageable par la prise en compte de plusieurs repères en fréquence (pas seulement 1.5 kHz et 3.5 kHz, comme proposé au départ).

Étant donné les performances de la famille de codages par double DCT, nous retenons ce type de codage pour la fusion précoce. Par la suite, FP se référencera aux résultats obtenus par DCT_sona.

Référence :	TA	Appris comme : k	p	t
k	50.9%	375	118	243
p	72.1%	178	531	**27**
t	68.2%	160	**74**	502

TAB. 5.14: Matrice de confusion pour l'apprentissage à partir de 5 paramètres : P_m, Y0_m, Delta_P, C-A, D-B calculés entre le moment du relâchement et celui du VOT. TA global = 63.8%.

5.4.3.3 Fusion directe et fusion tardive

Nous présentons ici une première comparaison entre les architectures des fusion. Nous devons nous assurer que le TA après fusion est toujours plus grand que tous les TA sans fusion. Une deuxième comparaison est faite à l'occasion de l'analyse des erreurs.

Fusion directe (FD) Pour réaliser la FD, nous avons juxtaposé dans le vecteur d'entrée les paramètres du N-seg et du V-seg. Ce qui est spécifique à cette architecture est le fait que la compétition/collaboration des indices est apprise sur les données et non pas figé a priori. Nous nous attendons à ce qu'elle favorise la collaboration des indices.

Fusion tardive (FT) Pour la FT, plusieurs types de modules de fusion peuvent être utilisés. Les travaux d'Adjoudani [Adj97] montrent des versions possibles de réalisation. Un tel module reçoit en entrée un classement et des notes pour chaque candidat de chacun des deux classifieurs. Dans notre cas, les notes sont les probabilités a posteriori de chaque occlusive parmi /p/, /t/ et /k/ tels qu'ils sont fournies par les classifieurs : $P_{N-seg}(p)$, $P_{N-seg}(t)$, $P_{N-seg}(k)$, $P_{V-seg}(p)$, $P_{V-seg}(t)$ et $P_{V-seg}(k)$. Les trois stratégies que nous avons utilisé pour faire la décision à partir de ces 6 probabilités sont les suivantes :

1. **Maximum de probabilité** (FT_MP). Le maximum des trois probabilités P_{N-seg} et le maximum des trois P_{V-seg} sont comparés. Le candidat choisi est celui qui a la plus grande probabilité.

2. **Combinaison simple des probabilités** (FT_CP). Les probabilités pour chaque occlusive sont multipliées : $CP = P_{N-seg} * P_{V-seg}$. Nous obtenons un nouveau triplet de probabilités ($CP(p)$, $CP(t)$ et $CP(k)$). Le maximum dans ce triplet donne le candidat choisi.

3. **Probabilités pondérées** (FT_PP). Ici on calcule un facteur scalaire qui rend compte de la capacité de discrimination de chacun des deux classements. Nous les appellerons **coefficients de confiance** (CC). CC_{N-seg} est minimum (0) quand les trois candidats sont équiprobables ($P_{N-seg}(p) = P_{N-seg}(t) = P_{N-seg}(k) = 0.33$) et il est maximum (1) quand la décision est extrêmement sure (par exemple, $P_{N-seg}(p) = 1$ et $P_{N-seg}(t) = P_{N-seg}(k) = 0$). De même, on calcule un coefficient (CC_{V-seg}) sur le classement du segment voisé. Un nouveau triplet de probabilités ($PP(p)$, $PP(t)$ et $PP(k)$) est calculé à

partir des six probabilités d'entrée et les deux coefficients CC_{N-seg} et CC_{V-seg}. Comme précédemment, le maximum dans ce triplet donne le candidat choisi à la sortie.

$$
\begin{aligned}
CC_{N-seg} &= \frac{3 * max_{i=p,t,k}(P_{N-seg}(i)) - \sum_{i=p,t,k} P_{N-seg}(i)}{2} \\
CC_{V-seg} &= \frac{3 * max_{i=p,t,k}(P_{V-seg}(i)) - \sum_{i=p,t,k} P_{V-seg}(p)}{2} \\
PP(i)_{i=p,t,k} &= \exp(CC_{N-seg} * \log(P_{N-seg}(i)) + CC_{V-seg} * \log(P_{V-seg}(i)))
\end{aligned}
$$

Il faut remarquer que ces trois stratégies sont présentées dans l'ordre croissant de degré de collaboration entre les deux classements. Ainsi, pour la première manière de combiner les probabilités (MP), les deux classements sont en pure compétition : un des deux va l'emporter et va être propagé à l'identique à la sortie. Pour les deux autres les probabilités peuvent conspirer de manière à faire pencher la décision en faveur d'un candidat qui n'a pas la meilleur note à posteriori.

Soient, par **exemple**, les valeurs d'entrée suivantes : $P_{N-seg}(p) = 0.8$, $P_{N-seg}(t) = 0$, $P_{N-seg}(k) = 0.2$, $P_{V-seg}(p) = 0.1$, $P_{V-seg}(t) = 0.3$ et $P_{V-seg}(k) = 0.6$. Selon la stratégie MP le candidat gagnant est /p/, selon la CP c'est /k/ ($CP(p) = 0.08$, $CP(k) = 0.12$) et selon la PP c'est /p/ ($CC_{N-seg} = 0.7, CC_{V-seg} = 0.4$ => $PP(p) = 0.34, PP(k) = 0.26$).

Architectures de fusion hybride (FH) Nous avons testé aussi deux architectures hybrides :

- Architecture à deux voies (FH2). Nous utilisons le classifieur operant sur le N-seg (le plus performant sans fusion) et le classifieur de la FD.
- Architecture à trois voies (FH3). Nous utilisons ici les trois classifieurs, celui du N-seg, du V-seg et celui de la FD.

Les architectures hybrides utilisant un module de fusion des classements, les trois stratégies décrites plus haut leur sont appliquables.

Nous présentons une sélection de résultats dans le tableau 5.15. Nous retiendrons **les conclusions** suivantes :

- La prise en compte des deux segments apporte une amélioration et ceci indifféremment de l'architecture utilisé (voir tableau 5.16).
- Le TA global ne permet pas de départager les architectures. Par exemple, la FD et la FH3_CP obtiennent de meilleurs taux mais au prix d'un doublement du nombre des paramètres stockés à l'apprentissage.
- Nous retiendrons la stratégie CP pour la fusion tardive (le simple produit des probabilités). Les différences sont faibles, FT signifiera FT_CP par la suite.
- Les méthodes de fusion essayées ici sont moins performantes que la fusion précoce. Les TA obtenus sont comparables avec ceux du codage DCT_sona mais ce dernier utilise beaucoup moins de paramètres.

Architecture	nb. par.	TA
N-seg	10	74.7%
V-seg	24	71.9%
FD	34	83.9%
FT_MP	34	81.6%
FT_PP	34	82.5%
FT_CP	34	82.9%
FH2_CP	34	82.1%
FH3_CP	34	84.4%
FH3_MP	34	83.2%
DCT_sona (Bark)	21	85.6%

TAB. 5.15: Taux d'apprentissage pour différentes architectures de fusion. 2208 CV (1 exemple correspond à 0.042%)

Les coefficients de confiance (CC) à la sortie du classifieur pour différents jeux de paramètres informent sur les cas difficiles pour la reconnaissance. Des statistiques sur ces coefficients selon la consonne et le contexte vocalique seront présentées dans le chapitre suivant.

5.4.4 Localisation temporelle de l'information

Dans cette section nous voulons étudier la distribution de l'information tout au long de la transition entre la consonne et la voyelle. Pour cette étude, il faudrait utiliser une fenêtre d'analyse glissante du relâchement jusqu'au centre de la voyelle et un même codage tout au long de ce balayage. En pratique, nous allons reprendre une partie des résultats déjà mentionés. Nous avons choisi deux types de codage : le spectre du relâchement codé par l'énergie en 10 bandes de 2.2 Barks (SIRB, spectre intégré du relâchement en barks) et la fréquence de formants (F). Si les deux types de codage ne sont pas directement comparables, on peut néanmoins observer l'influence de la position du centre de la fenêtre temporelle d'analyse. Les trois instants choisis pour le calcul de formants sont : la première trame du spectrogramme contenue dans le V-seg (FC), la dixième trame dans V-seg donc à 45 ms du VOT (F45ms) et la trame la plus proche du marqueur du centre de la voyelle (Fcible). Généralement, la trame du calcul des Fcible est située après celle des F45ms. Seules une dizaines de syllabes ont un Fcible plus près du VOT que le F45ms.

Les résultats sont présentés dans le tableau 5.17 et la figure 5.28. Ils montrent que l'information discriminante pour la consonne et pour la voyelle est distribuée dans le continuum de la transition de manière complémentaire. En se déplassant vers la cible vocalique, il est plus facile de reconnaître la voyelle et réciproquement. Il est utile de noter que, même aux endroits les plus éloignés de la cible, l'information discriminante ne s'efface pas complètement, ni pour la voyelle, ni pour la consonne. Nous remarquons une faible asymétrie entre la voyelle et la consonne : le TA de la voyelle croit systématiquement en se rapprochant de la cible de la voyelle tandis que la fenêtre la plus serrée au

Architectures	*Erreurs*
N-seg	559
V-seg	620
FD	355
FT	381

Architectures		*Erreurs*		
Première	Deuxième	communes	spécifiques à la 1-ère	spécifiques à la 2-ème
N-seg	V-seg	184	375	436
FT	FD	235	146	120
FD	N-seg	252	103	307
FD	V-seg	202	153	418
FT	N-seg	298	83	261
FT	V-seg	259	122	361

TAB. 5.16: Nombre d'erreurs communes et spécifiques pour différents paires de classifieurs. 2208 CV

Codage	nb. par.	TA consonne	TA voyelle
SIRB 10 ms	10	74.7%	25.9%
SIRB 26 ms	10	79.7%	32.8%
SIRB 40 ms	10	78.3%	41.5%
FC1..3	3	49.6%	42.6%
F45ms1..3	3	42.5%	79.9%
Fcible1..3	3	40.7%	83.9%
FD	34	83.9%	86.3%
Niveau aléatoire	0	33.3%	10.0%

TAB. 5.17: Taux d'apprentissage de la consonne et de la voyelle pour différentes fenêtres temporelles. 1 exemple correspond à 0.042%

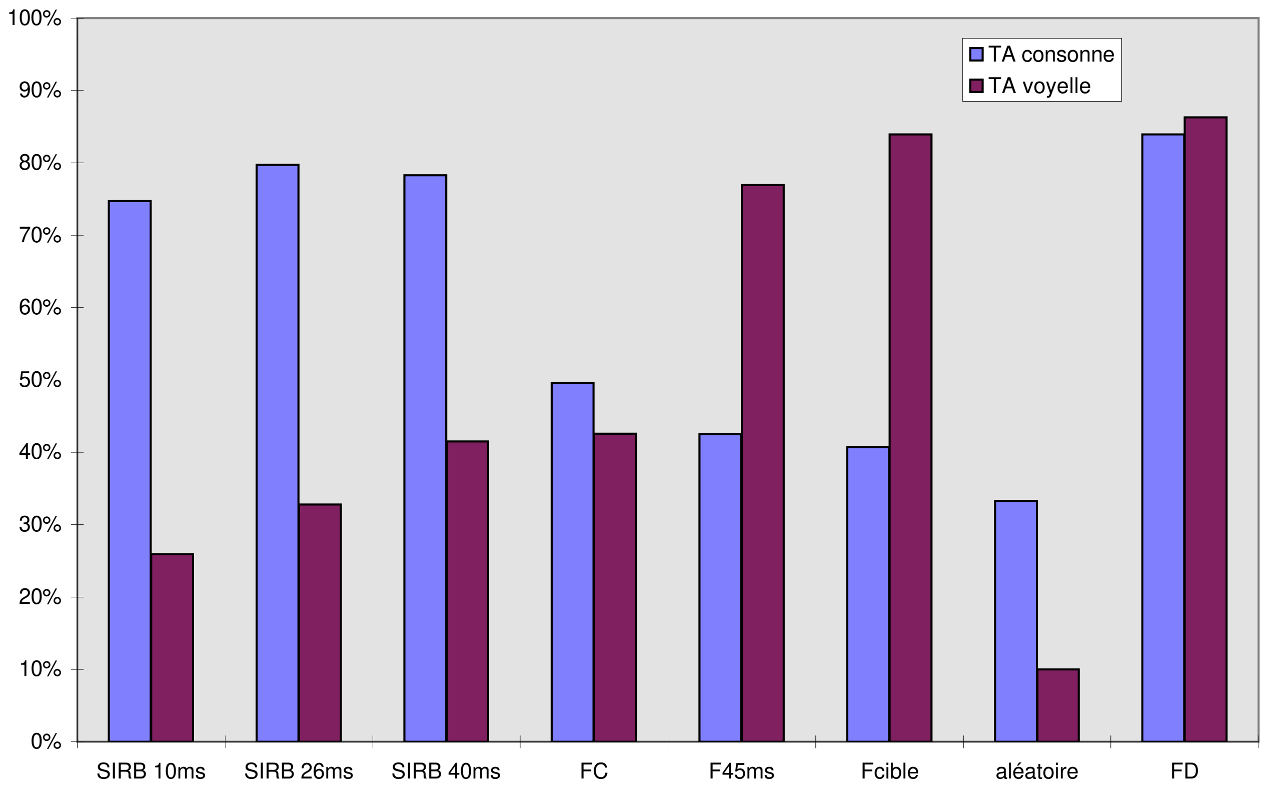


FIG. 5.28: Taux d'apprentissage de la consonne et de la voyelle pour fenêtres temporelles allant de la consonne vers la voyelle.

relâchement ne donne pas forcément le meilleur TA pour la consonne. Nous présentons, pour comparaison, le TA pour le jeu de paramètres de la fusion directe (FD). Il regroupe les paramètres du spectre de relâchement et des transitions formantiques pour un total de 28 à 34 paramètres. Comme attendu, les jeux de la FD apprennent mieux les deux tâches d'identification. Cependant, ils acquièrent ce résultat en utilisant beaucoup de paramètres. L'obtention d'une performance équivalente avec un nombre réduit de paramètres nous préoccupera dans la suite de nos travaux.

5.5 De l'analyse des erreurs

Nous allons procéder dans cette section à l'analyse des sourcess des erreurs faites à l'apprentissage des exemples de notre corpus par le classifieur baysien. Nous étudierons l'influence de différents facteurs comme la voyelle, la consonne, le type de fusion. Nous essayerons ensuite de réduire le nombre de paramètres par sélection afin de rendre la distribution des exemples plus facile à apprendre.

5.5.1 Performance par consonne

Tous nos tests ont conduit à des meilleurs performances pour [p] par rapport aux autres consonnes. Le tableau 5.18 montre la contribution de chaque consonne au taux d'apprentissage des exemples

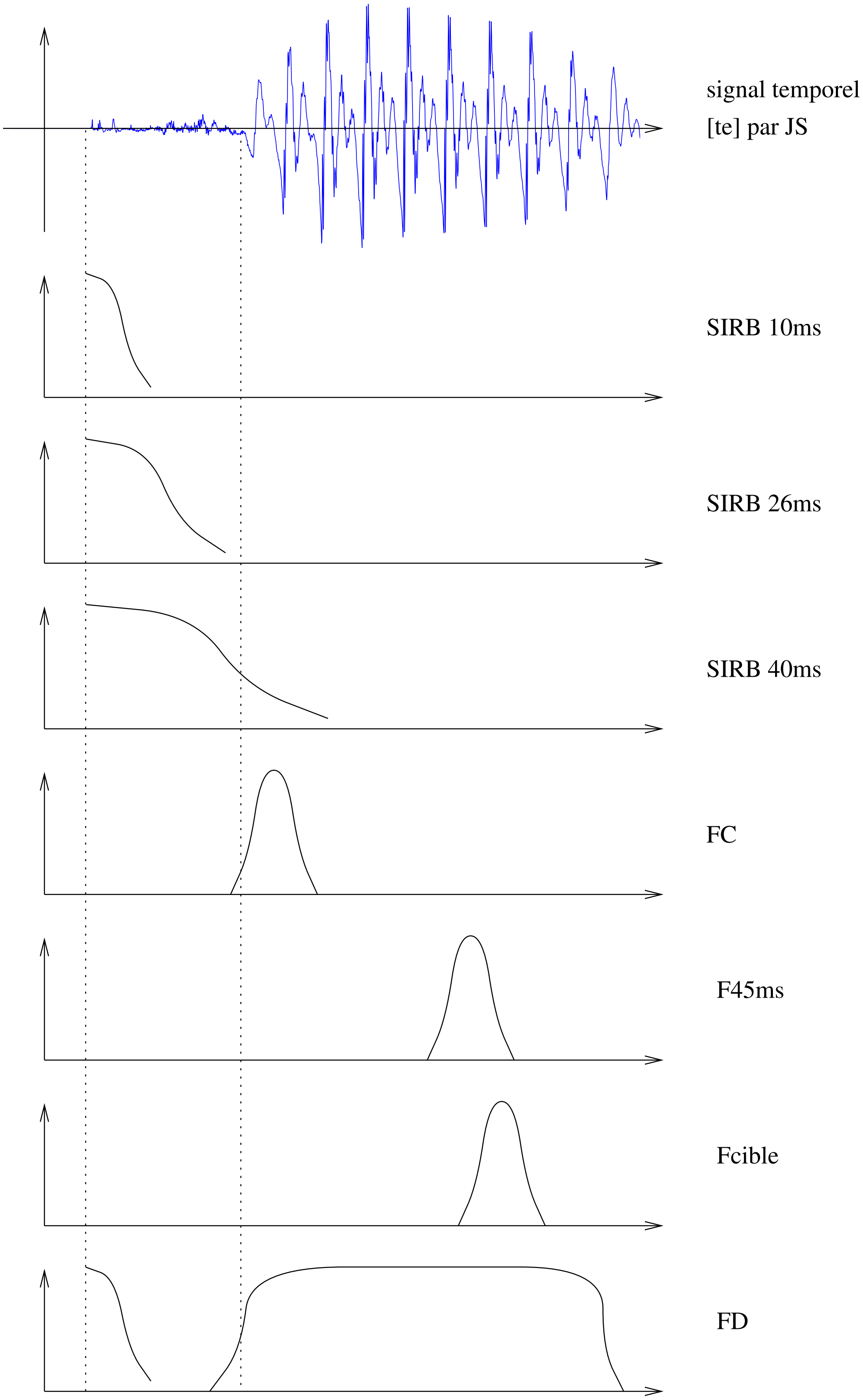


FIG. 5.29: Position relative des fenêtres d'analyse temporelle pour les codages présentés dans la figure précédente.

Architecture	k	p	t	TA global	nb. par.
N-seg	68.2%	90.1%	65.7%	74.7%	10
V-seg	62.6%	79.9%	73.2%	71.9%	24
FD	79.2%	93.7%	78.8%	83.9%	34
FT	76.5%	93.0%	78.7%	82.7%	34
FP	91.5%	96.7%	87.5%	91.9%	45

TAB. 5.18: Influence de la consonne sur la facilité d'apprentissage des exemples

Voyelle	TA par voyelle	k	t	p	Ordre		
i	78%	67%	67%	100%	p	t	k
a	96%	89%	98%	100%	p	t	k
y	81%	95%	56%	91%	k	p	t
u	96%	95%	56%	91%	k	p	t

TAB. 5.19: Interaction voyelle-consonne pour l'apprentissage des exemples par fusion précoce (FP)

pour plusieurs architectures. Comme nous l'avons souligné dans le chapitre sur la perception, ce résultats semble paradoxal vu le manque de traits spécifiques du [p] (Bonneau et collaborateurs ne définissent aucune "strong preference cue" pour les labiales dans le jeux de paramètres qu'ils proposent, [BCDL92]). Nous nous attendons à ce que cette préférence pour [p] disparaisse si des syllabes formées d'une voyelle sans consonne sont introduites dans le corpus. Dans l'absence d'une classification à quatre voix ([p], [k], [t] et silence), nous retenons que, pour notre corpus, le contraste le plus difficile est [t] versus [k].

Nos résultats sont à contraster avec les résultats obtenus par L. Djezzar en 1995 et 1997. Elle trouve un taux de reconnaissance de 95% pour [k], 88% pour [t] et 85% pour [p] [Dje95a, page 124] en utilisant des jeux de paramètres sélectionnés par contexte, notamment le trait de compacité. Cette différence d'ordre de préférence par le classifieur serait due, dans notre opinion, aux caractéristiques des jeux de paramètres. Les nôtres sont des paramètres intégrés, globaux, aptes à représenter le manque d'énergie du [p] par exemple. Les paramètres de Djezzar cherchent à capturer les indices comme un expert le ferait et le [k] présente le plus d'indices proéminents. Un argument pour cette explication peut être trouvé dans ses résultats de 1997 [DP97, page 164]. Les scores sont 92% pour [k], 89% pour [p] et 86% pour [t] avec la prise en compte du contexte mais l'ordre s'inverse (et les scores chutent, bien sûr) sans la prise en compte de celui-ci : 87% pour [p], 80% pour [t] et 75% pour [k].

Concernant l'interaction consonne-voyelle, nous ne présentons ici que les résultats pour la plus performante des architectures, la fusion précoce (FP). Nous observons que la facilité de reconnaissance de la consonne varie encore plus fortement avec la voyelle qu'avec l'architecture. La consonne [t] reste généralement la plus difficile à reconnaître (voir tableau 5.19).

Architecture	a	i	y	u	10 voyelles	nb. par.
N-seg	60.9%	60.4%	79.7%	93.7%	74.7%	10
V-seg	80.7%	51.0%	58.3%	64.4%	71.9%	24
FD	84.9%	66.7%	79.2%	95.3%	83.9%	34
FT	80.2%	65.5%	84.9%	95.3%	82.7%	34
FP	96.1%	78.3%	81.2%	96.1%	91.9%	45

TAB. 5.20: Influence de la voyelle sur la facilité d'apprentissage des exemplespar différentes architectures.

Architecture	Ordre de performance
N-seg	u y a $\approx$ i
V-seg	a u y i
FD	u a y i
FT	u y a i
FP	u $\approx$ a y i

TAB. 5.21: Ordre de performance des contextes vocalique pour différentes architectures du classifieur (le meilleur à gauche).

5.5.2 Performance selon les contextes vocaliques

Le contexte vocalique a été le facteur le plus influent sur la classification automatique des consonnes dans nos tests. Nous allons présenter donc plus en détail l'interaction voyelle-architecture (voir tableau 5.20 et figure 5.30). L'ordre de performance selon les contextes est reprise dans le tableau 5.21. Cet ordre est approximatif car non accompagné de tests de signification statistique. Nous retiendrons les conclusions suivantes concernant les contextes vocaliques :

– Le contexte difficile pour la reconnaissance est le contexte antérieur nonlabialisé ([i]).
– Le contexte ouvert ([a]) est le plus performant pour V-seg et un des moins performants pour N-seg. Les transitions formantiques sont plus discriminantes que le bruit de relâchement dans ce contexte.
– Le contexte postérieur labialisé ([u o ɔ]) est clairement le plus propice à la reconnaissance automatique du lieu d'articulation. Dans ce contexte, le bruit de relâchement est plus discriminant que les transitions formantiques.
– Plus la fusion est tardive, moins bonnes sont les performances en contexte antérieur (mais toujours meilleures que la meilleure des deux sources).
– La fusion directe en contexte [y] et la fusion tardive en contexte [a] n'arrivent pas au niveau de la meilleure des deux sources fusionnées.

Nos résultats sont à contraster avec ceux de Djezzar [Dje95a, page 124] qui trouve l'ordre [u]>[i]>[y]>[a]. Si la mauvaise position du [a] dans son classement est facilement explicable par le manque d'in-

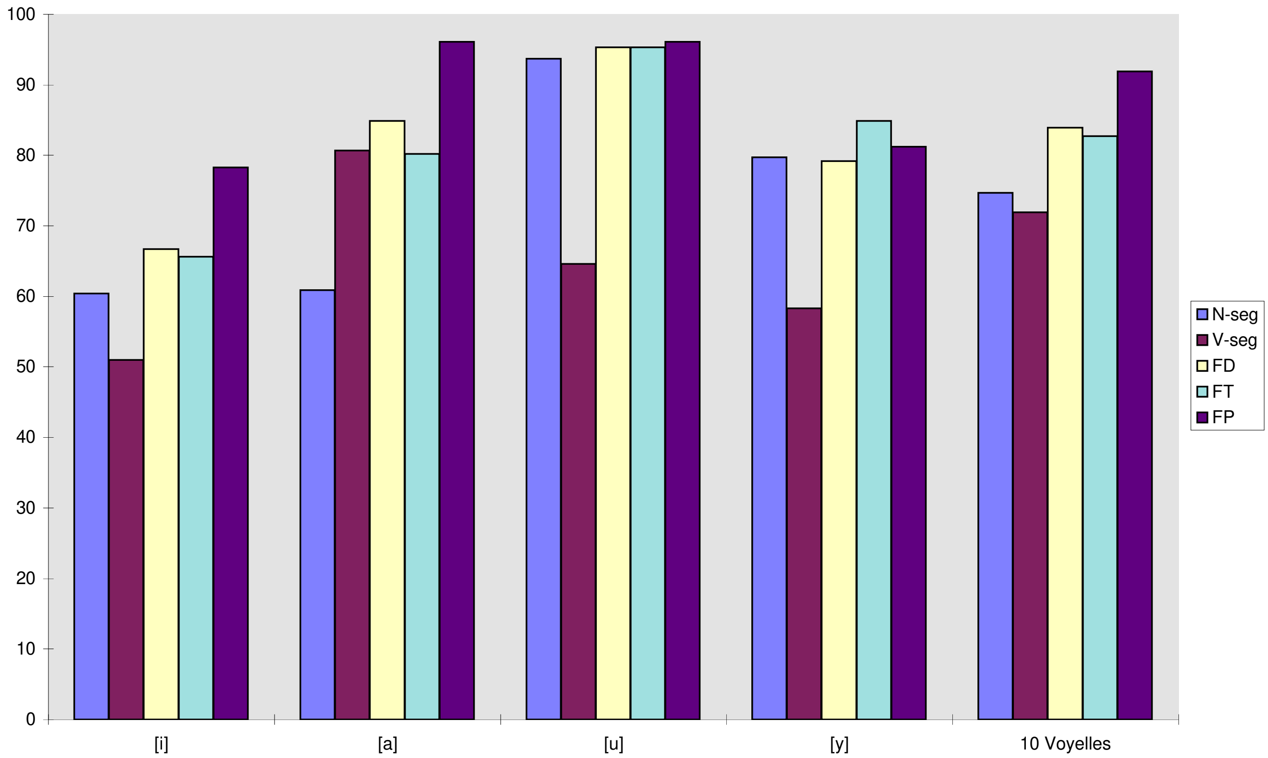


FIG. 5.30: Influence de la voyelle sur la facilité d'apprentissage des exemples. Regroupement par voyelle

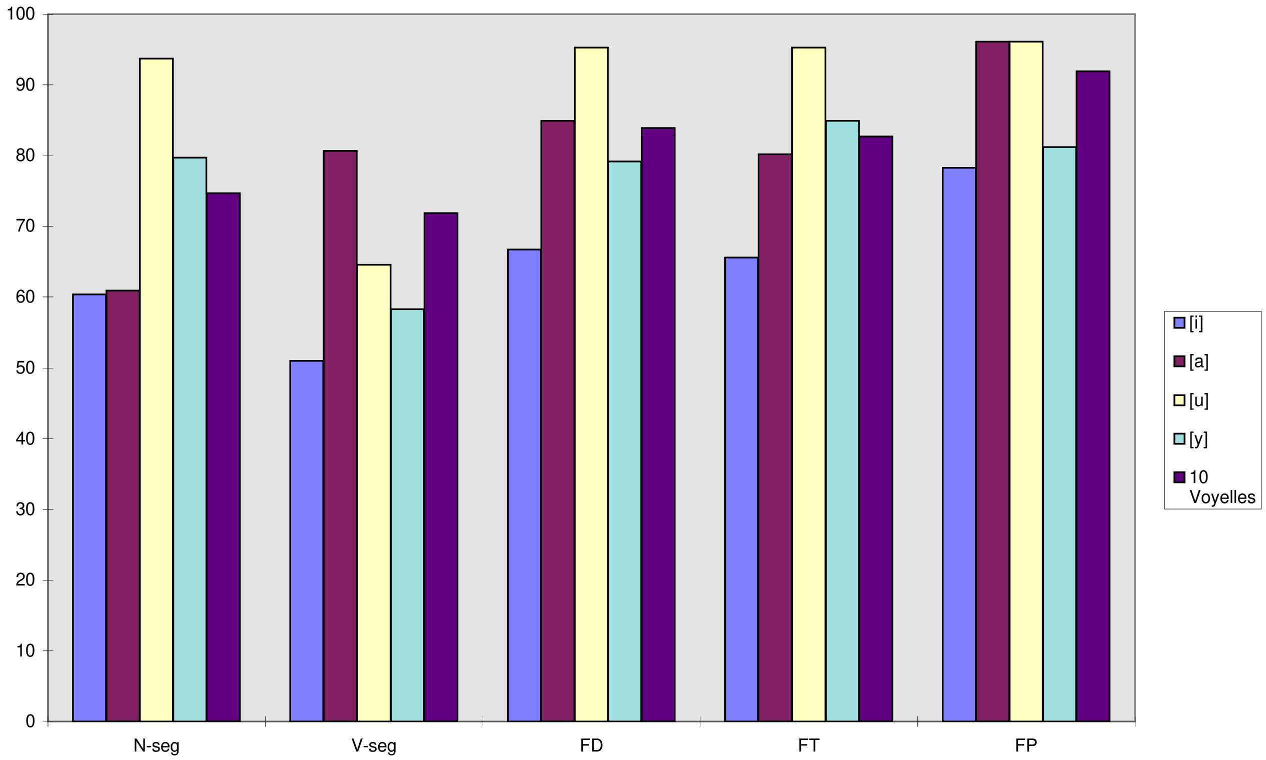


FIG. 5.31: Influence de la voyelle sur la facilité d'apprentissage des exemples. Regroupement par architecture

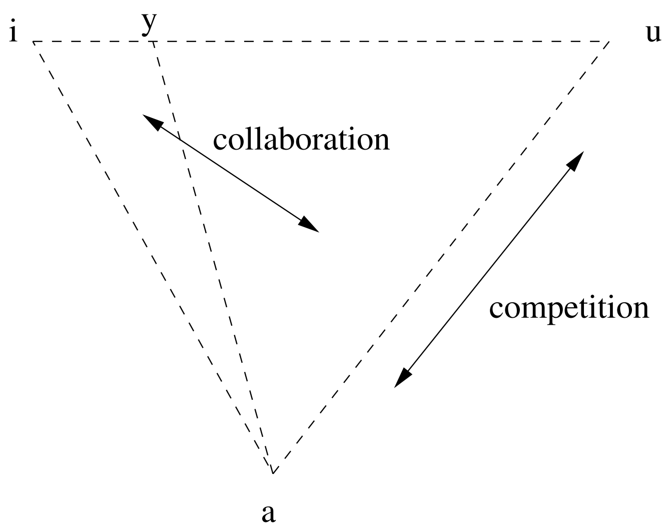


FIG. 5.32: Compétition et collaboration des indices du N-seg et V-seg selon les contextes vocaliques

dices sur le V-seg dans ses tests, nous ne pouvons pas expliquer le mauvais classement du [y]. En contrepartie, nos résultats sont en accord avec ceux de Smits [StBC96b] et ceux de Coste et Bonneau [BCDL92].

L'étude de la distribution des coefficients de confiance à la sortie du classifieur peut nous fournir une comparaison plus fine des contextes vocaliques et des architectures. Nous rappelons notre définition pour le coefficient de confiance (CC) : la différence entre la probabilité du candidat gagnant et la moyenne des deux autres probabilités (voir la section 5.4.3.3). Ainsi, un coefficient proche de 1 signale un candidat proche du prototype de sa classe et éloigné des prototypes des deux autres. Un histogramme des CC biaisé vers la valeur 1 trahit donc une distribution des candidats serrée autour de son prototype. De même, un histogramme biaisé vers la valeur 0 trahit une distribution où la majorité des candidats sont en proximité du point équidistant entre les prototypes. Nous avons fait des statistiques sur les coefficients de confiance en apprenant sur tout le corpus. Nous retenons que :

– Le nuage de points représentant le [p] est bien distincts des deux autres. Cette distinctivité est plus marqué sur N-seg que sur V-seg et elle est mieux conservée par FD que par FT (voir figure 5.37). Nous n'avons pas trouvé de différences importantes selon le contexte.
– Le contraste [k] versus [t] est fortement dépendant du contexte. Pour les paramètres du N-seg par exemple, les nuages de [t] et [k] se mélangent pour les contextes antérieur et central et sont mieux séparés pour les contextes postérieur et labialisé (voir figure 5.33).
– Les bons résultats obtenus sur V-seg dans le contexte central sont fragiles car l'histogramme des CC pour [k] et [t] dans ce contexte n'est pas visiblement différent de ceux des autres contextes (voir figure 5.34).
– La fusion directe est préférable à la fusion tardive. En effet, elle semble fournir des nuages mieux séparés, surtout pour le contexte antérieur (voir figures 5.35 et 5.36).

Notre conclusion peut se résumer au schéma présenté dans la figure 5.32. Nos observations nous amènent à considérer que les indices des deux segments sont à la fois en collaboration et en com-

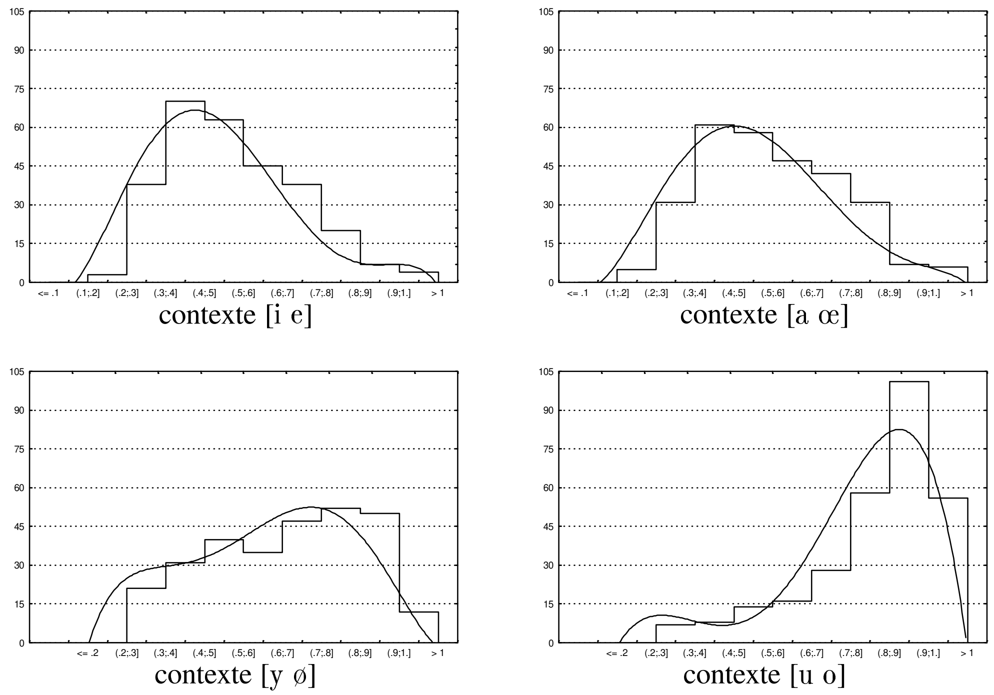


FIG. 5.33: Statistique sur les coefficients de confiance pour les exemples classés comme [t] ou [k] pour les paramètres du N-seg. Les courbes représentent des approximations polynômiales d'ordre 5 des histogrammes. Noter la bonne confiance pour les voyelles [u o] et la mauvaise confiance pour [i e].

pétition. Ainsi, sur l'axe du triangle vocalique qui va de [a] à [u], on peut noter plus facilement la compétition : pour [a] les transitions formantiques sont plus discriminantes alors que pour [u], c'est le bruit de relâchement. Nos tests de perception montrent que l'information de l'autre segment peut être masqué par le segment le plus discriminant. Au contraire, sur l'axe [i] - [u], on peut remarquer surtout l'aspect collaboration. Pour le contexte [i], comme nous l'avons constaté également dans les tests de perception, le N-seg et (encore plus) le V-seg tendent tous les deux à perdre leur pouvoir individuel de discrimination. Comme l'architecture à fusion tardive possède les voix indépendantes de traitement les plus longues, nous nous attendons à ce qu'elle favorise les aspects compétition des indices plus que leur collaboration. Cela explique ses piètres résultats sur le contexte antérieur (voir figure 5.36).

5.5.3 Multimodalité dans la distribution des indices acoustiques

Le fait que les indices des consonnes, à la différence de ceux des voyelles, présentent des distributions avec plusieurs nuages par classe phonétique a été souvent signalé (voir le chapitre sur l'état de l'art). Une méthode évidente mais peu rigoureuse de le montrer est l'étude des prototypes des syllabes. Nous présentons ici, comme exemple, les prototypes des syllabes pour les trois consonnes sourdes ([p], [t], [k]) et les quatre voyelles extrêmes ([a], [i], [y], [u]). Il s'agit (voir figure 5.38) des spectrogrammes lissés du premières 50 ms de signal après le relâchement. Ces prototypes sont

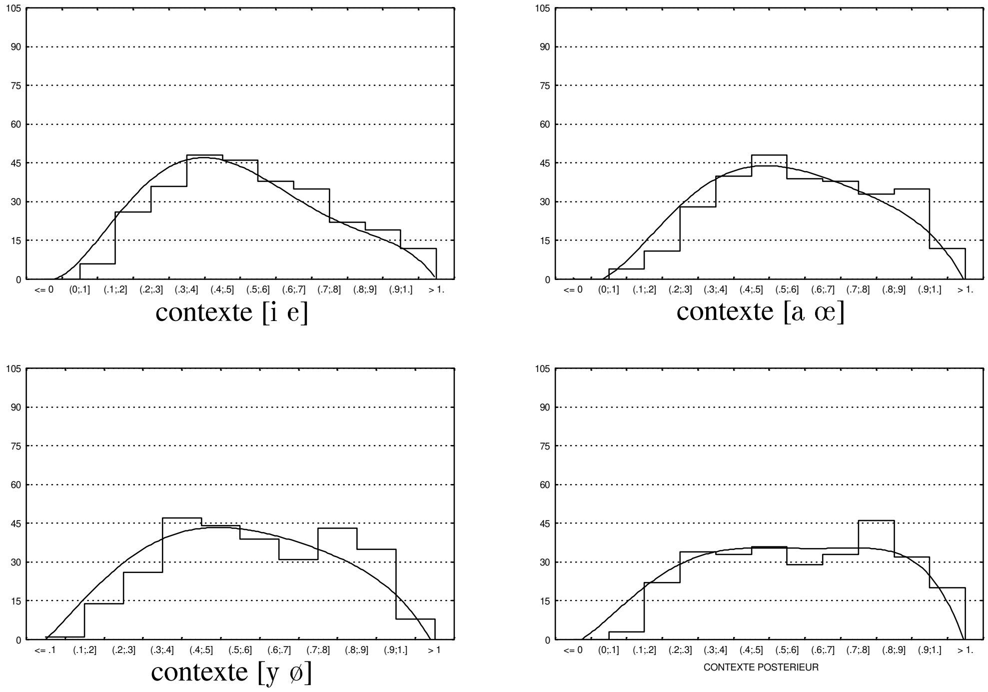


FIG. 5.34: Statistique sur les coefficients de confiance pour les exemples classés comme [t] ou [k] pour les paramètres du V-seg

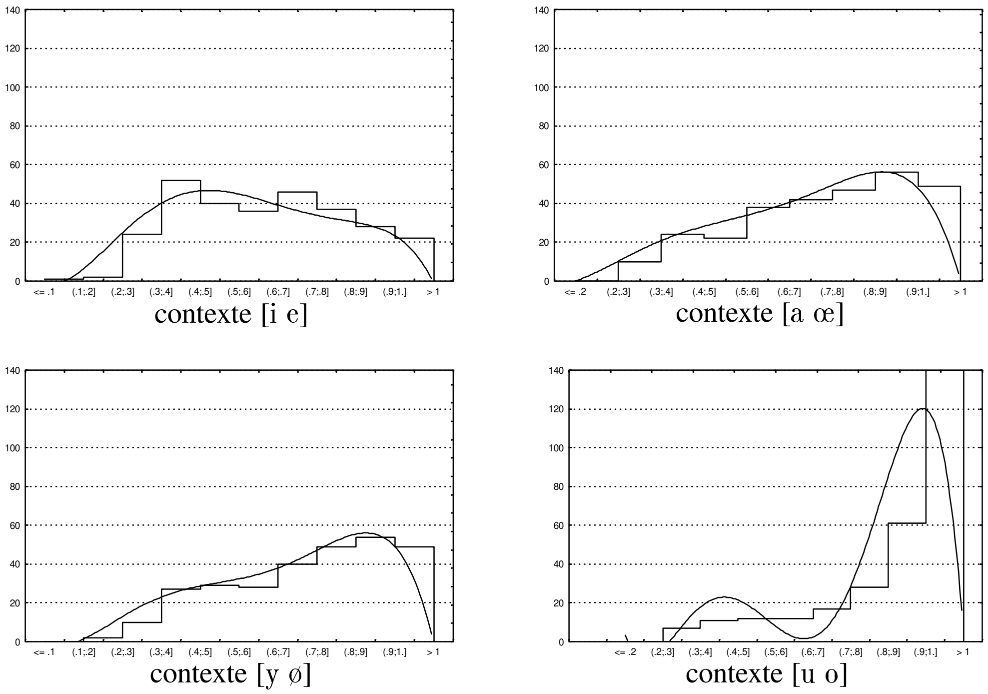


FIG. 5.35: Statistique sur les coefficients de confiance pour les exemples classés comme [t] ou [k] pour l'architecture à fusion directe.

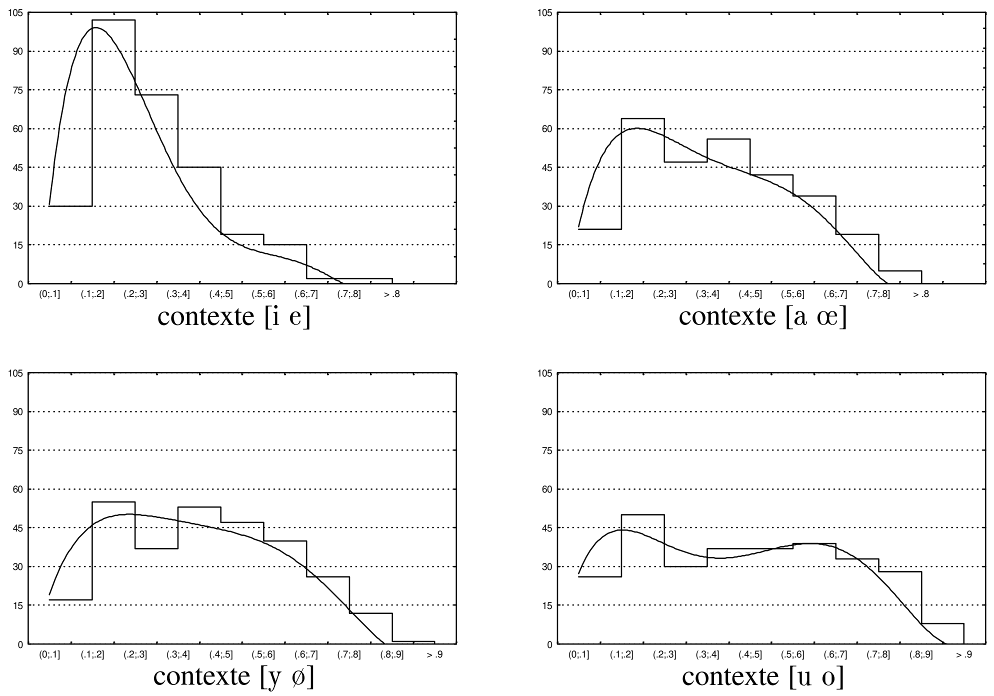


FIG. 5.36: Statistique sur les coefficients de confiance pour les exemples classés comme [t] ou [k] pour l'architecture à fusion tardive. Noter le faible rendement pour les voyelles antérieurs non-arrondies.

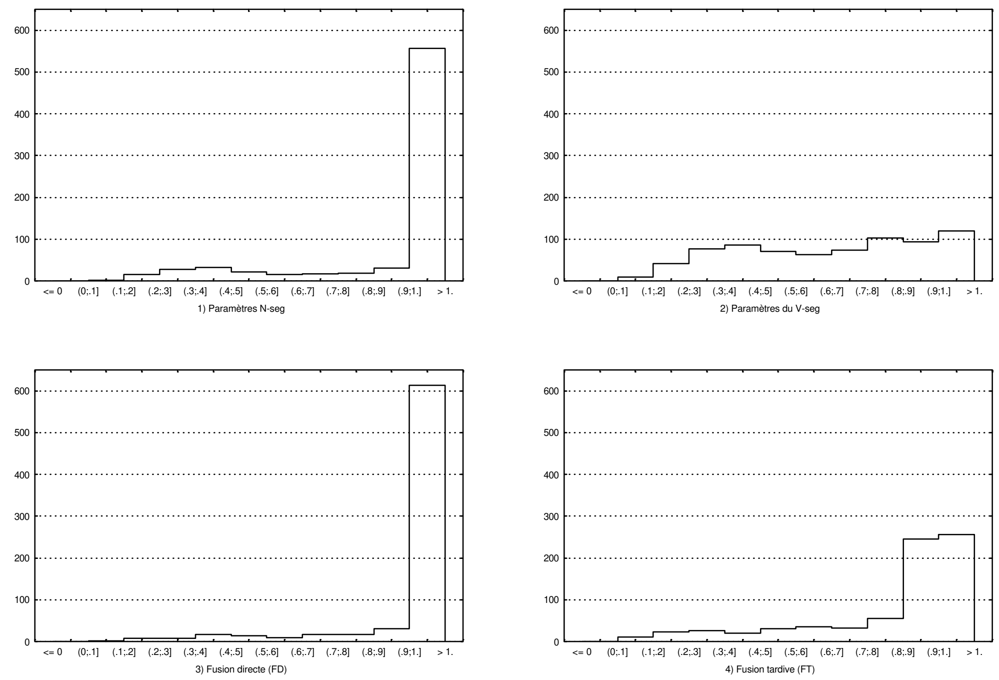


FIG. 5.37: Statistique sur les coefficients de confiance pour les exemples classés comme [p].

Jeu de paramètres	appris sur :	testé sur :	TR global	k	t	p
N-seg (10)	antérieur	postérieur	33%	**0.7%**	23%	75%
	postérieur	antérieur	55%	**0.7%**	67%	96%
V-seg (24)	antérieur	postérieur	50%	**0%**	87%	63%
	postérieur	antérieur	33%	**0%**	97%	**1.4%**
FP (45)	antérieur	postérieur	36%	19%	22%	67%
	postérieur	antérieur	44%	**0%**	40%	92%

TAB. 5.22: Taux de reconnaissance par consonne pour l'apprentissage sur un contexte vocalique et le test sur contexte différent. Contexte postérieur : [u o ɔ] et contexte antérieur : [i e ɛ].

issus de l'apprentissage du classifieur baysien sur tout le corpus sur les coefficients de la transformée en cosinus discret temps-fréquence. Les différences entre ces prototypes sont visibles. Cependant, ces différences semblent ne pas trop grever les performances d'un classifieur monomodal par excellence qui utilise ces prototypes (TR = 88%). Il est important de noter aussi que les contrastes sur la première trame de ces prototypes bidimensionnels multi-locuteurs correspondent tout à fait aux contrastes que nous avons remarqué sur les prototypes des spectres du relâchement d'un seul locuteur (dans la section 5.4.1.2 sur le paramétrage du N-seg).

Pour mesurer l'"étendue"de cette variabilité, nous avons mené un test spécifique. Nous avons entraîné nos classifieurs sur les syllabes d'un contexte et nous avons testé ces classifieurs sur les syllabes d'un autre contexte. Nous avons testé toutes les combinaisons pour plusieurs architectures. Nous présentons ici l'effet le plus saisissant, l'opposition contexte antérieur - contexte postérieur (voir tableau 5.22). Nous observons que, à la différence de la partition par locuteur, les prototypes par contexte ne sont pas du tout généralisables à d'autres contextes. Ainsi, pour certains codages, le prototype du [k] d'un contexte est un très bon prototype de [t] dans un autre contexte (le classifieur apprend les exemples à l'envers) ! L'architecture de la fusion précoce, la plus performante pour la reconnaissance, n'échappe pas à cette règle : tous les [k] de [ki] sont reconnus comme [t] en contexte postérieur et en contexte labialisé, tous les [t] de [ty] sont reconnus comme [p] en contexte antérieur non-labialisé, etc. Nos résultats montrent donc que le spectre du relâchement varie avec le contexte **au moins autant** que les transitions formantiques.

La conclusion de ces tests est que l'apprentissage de consonnes dans tous les contextes nuit à leur reconnaissance automatique. Cette conclusion s'applique à une gamme assez variée de jeux de paramètres (N-seg, V-seg, FP). Une solution possible est l'utilisation de classifieurs capables d'apprendre des distributions multi-modales. Cette solution ne nous semble pas attirante car elle ne permet pas de discerner entre éloignement permis d'une distribution gaussiene (selon la voyelle) et un éloignement non-permis (différents perturbations, caractéristiques du locuteur, etc.). Comme nous l'avons annoncé, nous désirons chercher une solution par un paramétrage apte à rendre la distribution la plus monomodale possible.

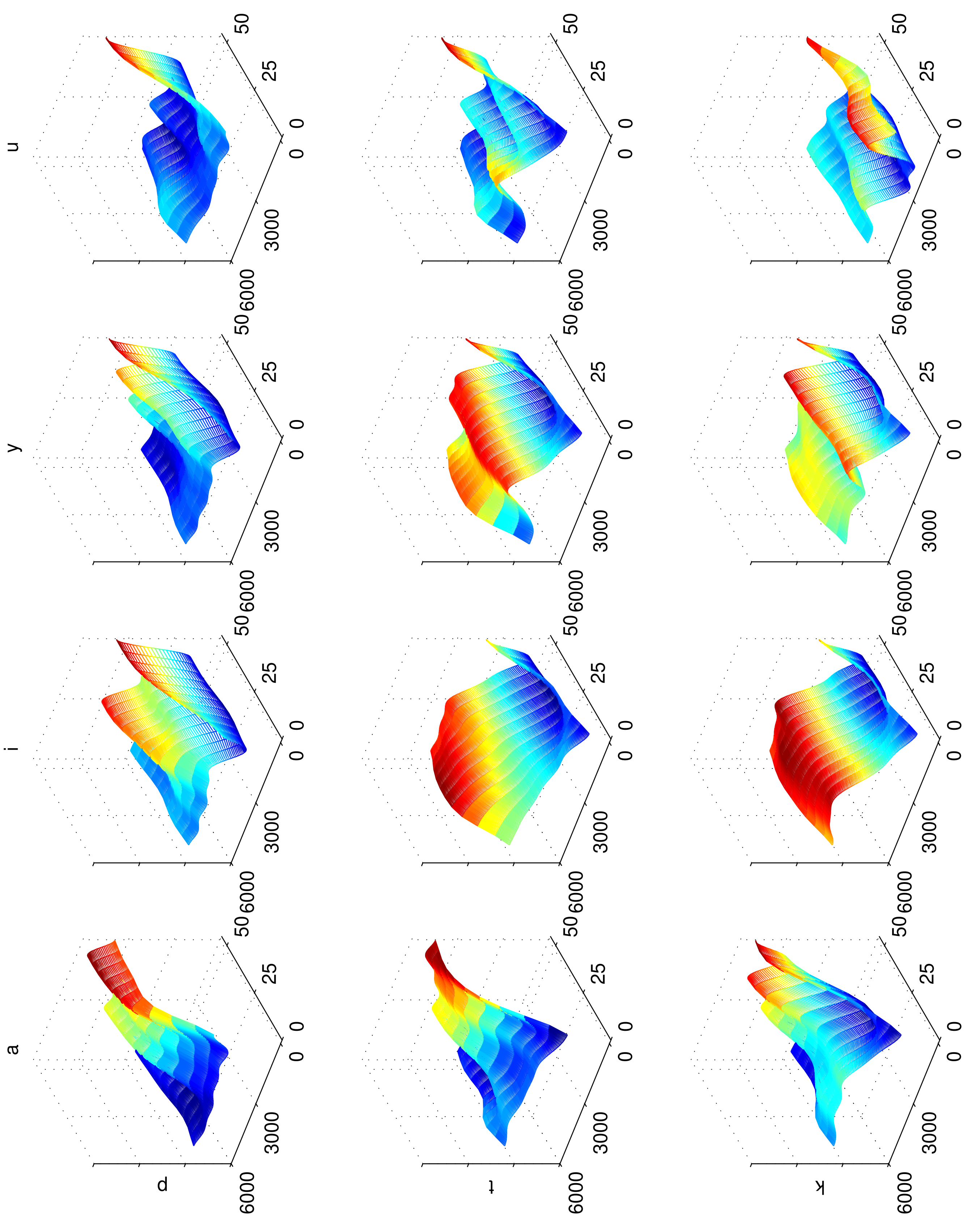


FIG. 5.38: Prototypes de syllabes CV pour les trois occlusives sourdes et les quatre voyelles extrêmes. Chaque case représente la moyenne de 64 réalisations (8 locuteurs masculins). Les spectrogrammes couvrent l'intervalle de 50 ms après le relâchement et les fréquences de 200 Hz à 6 kHz.

5.5.4 Améliorer la performance en contexte avant. Nécessité de l'analyse morphologique

Nos tests précédents nous permettent de dégager deux conclusions pour guider la suite de nos recherches :

- Quelle que soit l'architecture testée, c'est le contexte antérieur qui fourni la majorité des erreurs (voir tableau 5.20).
- Dans ce contexte vocalique, seul l'ensemble des deux segments, N-seg et V-seg, est discriminant. La collaboration des indices est ici maximale (voir figure 5.30).

Nous concluons que l'amélioration de la performance des algorithmes automatiques passe par l'étude du mécanisme de collaboration entre les indices dans ce contexte. La modélisation de ce mécanisme demande une nouvelle architecture. En effet, dans le cas de la fusion précoce ou de la fusion tardive, cette collaboration est figée. Dans le cas de la fusion directe, elle est apprise sur les données mais de manière aveugle dans le sens où nous n'avons pas de moyen de savoir si la collaboration ainsi apprise est appliquable à d'autres données que celles qui ont servi à l'apprentissage.[9]

Pour démontrer la nécessité de cette nouvelle architecture, sensible au contexte vocalique, nous allons comparer des indices de séparabilité[10] sur nos syllabes obtenus de plusieurs manières :

1. une première analyse discriminante canonique est appliquée sur toutes les syllabes. Nous obtenons ainsi une rotation de l'espace de paramètres, appelons-la R1.
2. $S_1 = \frac{1}{\lambda_1}$, la séparabilité de tous les exemples est mesurée dans ce nouveau espace.
3. $S_2 = \frac{1}{\lambda_2}$, la séparabilité des syllabes en contexte postérieur après la rotation R1 est mesurée.
4. $S_3 = \frac{1}{\lambda_3}$, la séparabilité des syllabes en contexte antérieur après la rotation R1 est mesurée.
5. les syllabes en contexte antérieur sont isolées. Une deuxième analyse discriminante canonique sur ces syllabes nous fournit la rotation R2. $S_4 = \frac{1}{\lambda_4}$, la séparabilité des syllabes en contexte antérieur après la rotation R2 est mesurée.

Nous présentons ici les résultats dans le plan des deux premières coordonnées canonique (le plan le plus discriminant qu'on peut obtenir par rotation à partir du jeu initial de paramètres) pour le locuteur JS pour le jeu de paramètres du N-seg (voir figure 5.39). Les résultats sont encore plus clairs sur le jeu de paramètres de la fusion précoce (voir figure 5.40). Nous constatons que :

- La consonne [p] est toujours la mieux séparée parmi les trois consonnes.
- Les consonnes en contexte postérieur sont mieux séparés que l'ensemble des consonnes ($S_2 > S_1$) quand la rotation est apprise sur tous les contextes (voir graphique 2 dans les figures).
- Les [k] et [t] en contexte antérieur sont moins bien séparés ($S_3 < S_1$) quand la rotation est apprise sur toutes les contextes (voir graphique 3 dans les figures).

[9] Supposant un échec sur des nouvelles données, serait-il du à une collaboration mal apprise ou à un banal changement de microphone ?

[10] Nous utilisons comme indice de séparabilité l'inverse du coefficient lambda de Wilks ($S = 1/\lambda$). C'est un indice de la nature du rapport entre la distance moyenne inter-classe et la distance moyenne intra-classe

– Les consonnes en contexte antérieur retrouvent le même niveau de séparabilité que les consonnes en contexte postérieur quand la rotation est apprise sur le contexte antérieur ($S_4 >> S_3$, voir graphique 4 dans les figures).

Nous concluons donc que la façon habituelle de faire l'apprentissage nuit à la séparabilité des consonnes en contexte antérieur. Les architectures que nous avons présentées jusqu'ici, essayent de gommer les différences entre les contextes. Elles essayent de rendre les jeux de paramètres pour le codage de la voyelle et de la consonne orthogonals. En s'appuyant sur l'identification de la catégorie de la voyelle à partir du N-seg par des auditeurs, nous avons toutes les raisons de penser que la perception humaine n'agit pas de la même manière. Elle ne tente pas d'"effacer" le contexte vocalique. Au contraire, il est fort probable qu'elle l'utilise pour s'adapter au différents types de contraste consonantique.

Nous pensons qu'il y a des "objets" visés par la perception dans le signal, comme les résonances du conduit vocal. Nous montrerons qu'il y a localisation temporelle et fréquencielle de ces objets (voir sous-section sur l'information mutuelle). Il n'y a pas de raison a priori pour penser que ce type de perception est limité seulement aux spectres voisés. Le mécanisme que nous proposons pour rendre explicite la relation entre V-seg et N-seg prend appui sur l'analyse que nous avons fait sur les voyelles. Précisément, nous montrerons qu'une zone du spectre de relâchement est plus discriminante que le reste du spectre et que cette zone peut être mise en relation avec R2. Il en résulte que l'analyse du spectre guidée par R2 est la bonne manière d'améliorer les performances en contexte antérieur. Cette analyse guidée par les résonances du conduit serait donc le mécanisme de perception spécifique à la parole dans le cas des consonnes occlusives.

5.5.5 Vers la fusion explicite

Nous voulons donc trouver une nouvelle architecture de classification permettant de modéliser la collaboration précoce entre N-seg et V-seg. Nous l'avons appelé architecture à fusion explicite (FE). Dans ce cas, la collaboration n'est ni figée (comme dans FT et FP) ni spécifique à un corpus d'apprentissage (comme dans FD). Dans la nouvelle architecture, l'interaction entre les paramètres du N-seg et du V-seg passe par l'évaluation explicite des paramètres du mécanisme de collaboration et de celui de compétition. Dans le cas des syllabes occlusive-voyelle, on pourrait chercher, par exemple, comme paramètre une mesure de la dissimilarité entre les formants au centre de la voyelle et les formants au début de la syllabe. Nous avons choisi de chercher la fréquence d'une bande discriminante pour le lieu d'articulation sur le spectre du début de syllabe. Les paramètres évalués dans une fusion explicite sont, dans notre opinion, le moyen le plus approprié pour l'étude de la variation des contrastes consonantiques avec le contexte vocalique où avec le locuteur.

Pour implementer et tester l'architecture à fusion explicite (FE) nous avons utilisé un corpus restreint. Nous avons choisi dans le corpus de base les syllabes contenant [t] ou [k] et [i], [y] ou [u]. Nous obtenons ainsi 384 syllabes. Ce sous-corpus garde l'essentiel de la variabilité et renferme la majorité des cas difficiles. Nous avons exclu le contexte central ouvert ([a]) car l'analyse du spectre de relâchement y semble moins importante.

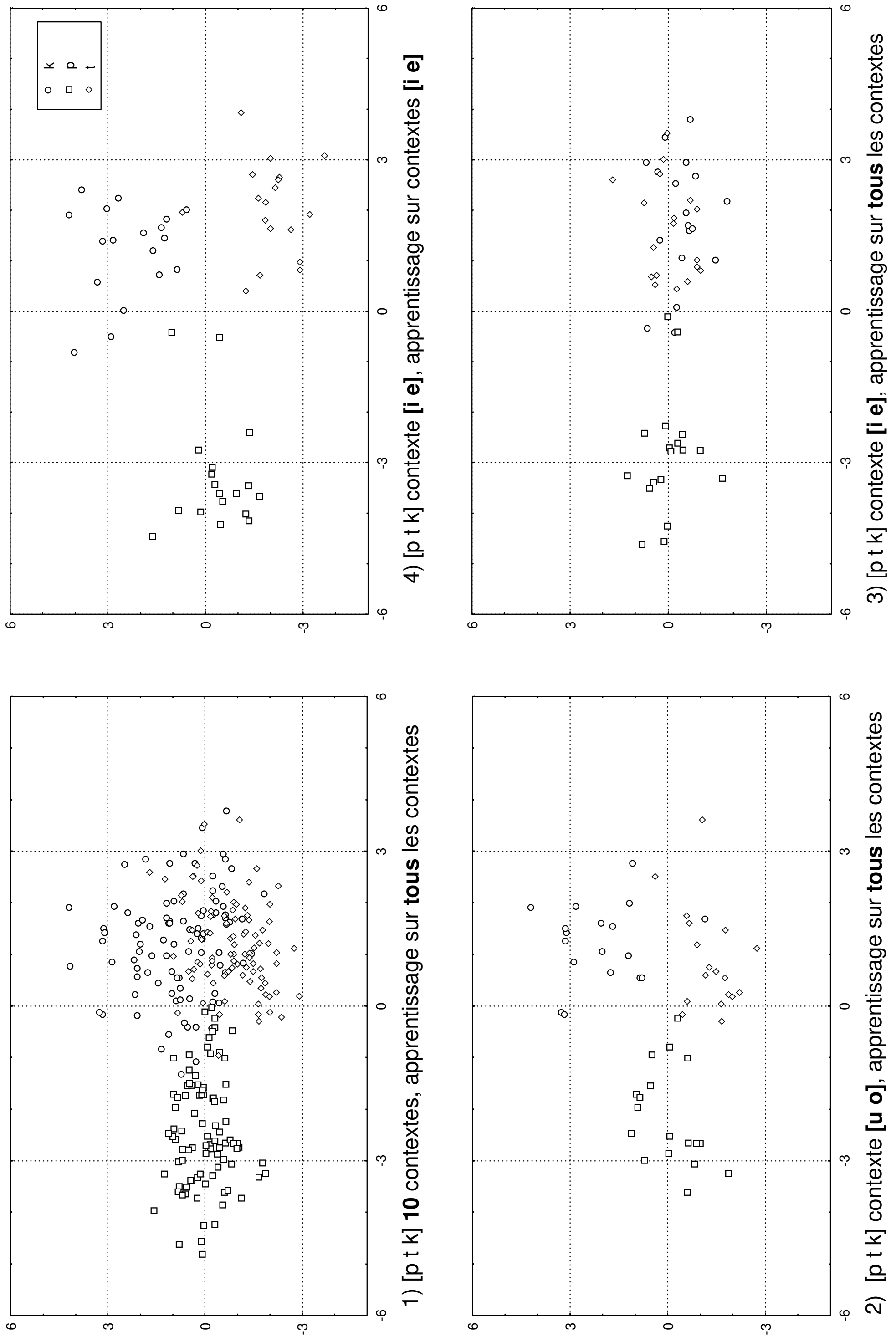


FIG. 5.39: Nécessité de l'apprentissage spécifique au contexte antérieur pour les paramètres du N-seg (BRK1..10). Les exemples sont présentés dans le plan des deux premières composantes canoniques. Syllabes du locuteur JS. Dans le sens trigonométrique : 1) tous les contextes (λ=0.172, TA=79%), 2) contextes [u o] (λ=0.083, TA=96%), 3) contextes [i e] pour rotation apprise sur tous les contextes (λ=0.195, TA=72%), 4) contextes [i e] pour rotation apprise sur les contextes [i e] (λ=0.033, TA=96%).

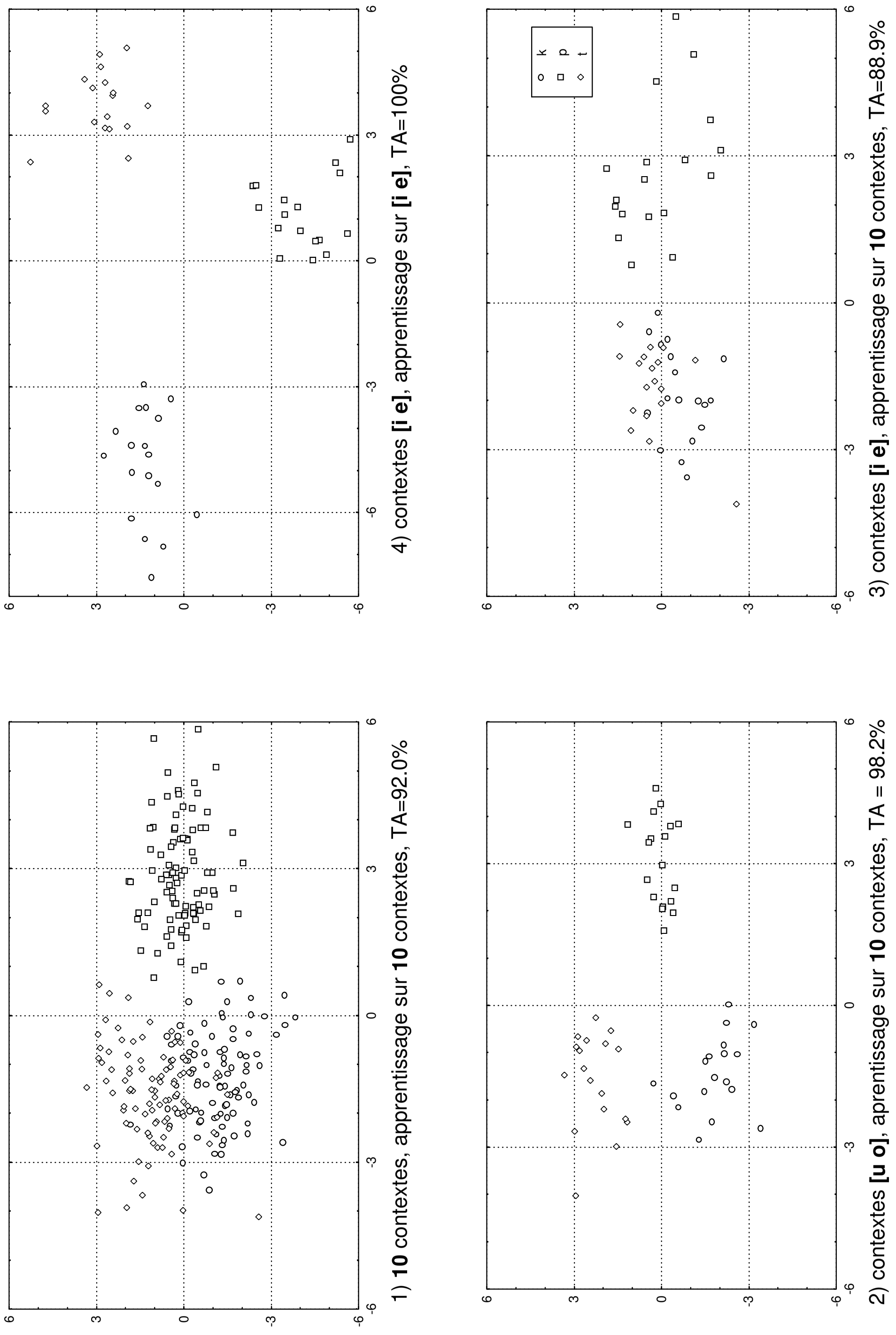


FIG. 5.40: Nécessité de l'apprentissage spécifique au contexte antérieur pour les paramètres de la fusion précoce (DCT_sona1..21). Syllabes du locuteur JS. Même disposition des graphiques que dans la figure précédente 5.39.

Paramètre	Fréquence (en Hertz)
BRK1	0 à 202
BRK2	202 à 423
BRK3	423 à 690
BRK4	690 à 1031
BRK5	1031 à 1464
BRK6	1464 à 2100
BRK7	2100 à 2945
BRK8	2945 à 4116
BRK9	4116 à 5740
BRK10	5740 à 8000

TAB. 5.23: Bandes de fréquence correspondant aux paramètres BRK1..10

5.5.5.1 Trois plus forts que dix

Premierement nous avons cherché à savoir si un sous-ensemble de paramètres sélectionnés peut garder la même séparabilité que l'ensemble de paramètres pour ce sous-corpus. Les dix paramètres de base utilisés pour caractériser le N-seg sont les énergies moyennes de 10 bandes spectrales (voir tableau 5.23). Nous notons ces énergies BRK1..10. Basé sur la dispersion des spectres de relâchement (voir figure 5.41), nous avons choisi trois paramètres (S1..3, S comme Sélectionné) comme suit :

$$S1..3 = \begin{cases} BRK3..5 & \text{si voyelle = [u]} \\ BRK5..7 & \text{si voyelle = [y]} \\ BRK7..9 & \text{si voyelle = [i]} \end{cases}$$

Nous avons comparé ce jeu de paramètres avec le jeu BRK1..10 pour sa capacité d'apprentissage phonétique et pour sa robustesse vis-à-vis du contexte vocalique. Les résultats (voir tableau 5.24) montre une très bonne capacité d'apprentissage. De même que R1..2 séparent mieux les voyelles que F1..2, S1..3 séparent mieux les occlusives que BRK1..10. L'ajout de l'information sur la voyelle introduit dans le jeu de paramètres une trace de la sélection faite. Mais, contrairement à notre attente, l'ajout des FC1..4 ou F_centreS2[11] n'a pas apporté une amélioration sensible des taux d'apprentissage.

Concernant les variations contextuelles, nous constatons (voir tableau 5.25) une robustesse largement meilleure pour S1..3. L'apprentissage sur [u] et le test sur [i] continue à poser problème : presque tous les [k] sont confondus avec de [t]. Finalement, pour S1..3, nous avons refait la manipulation en quatre étapes par analyse discriminante canonique. La figure 5.42 comparent les jeux BRK1..10 et S1..3. Elle montre l'évolution de la séparabilité pour différents corpus d'apprentissage. sur le premier axe discriminant.[12] Le gain des paramètres S1..3 est observable sur la troisième paire

[11]FC1..4 sont les fréquences de formants calculés au tout début du V-seg, F_centreS2 est la fréquence centrale du BRK4 ou BRK6 ou BRK8 selon la voyelle

[12]Représentation analogue à celle de la figure 5.39 mais pour deux classes seulement.

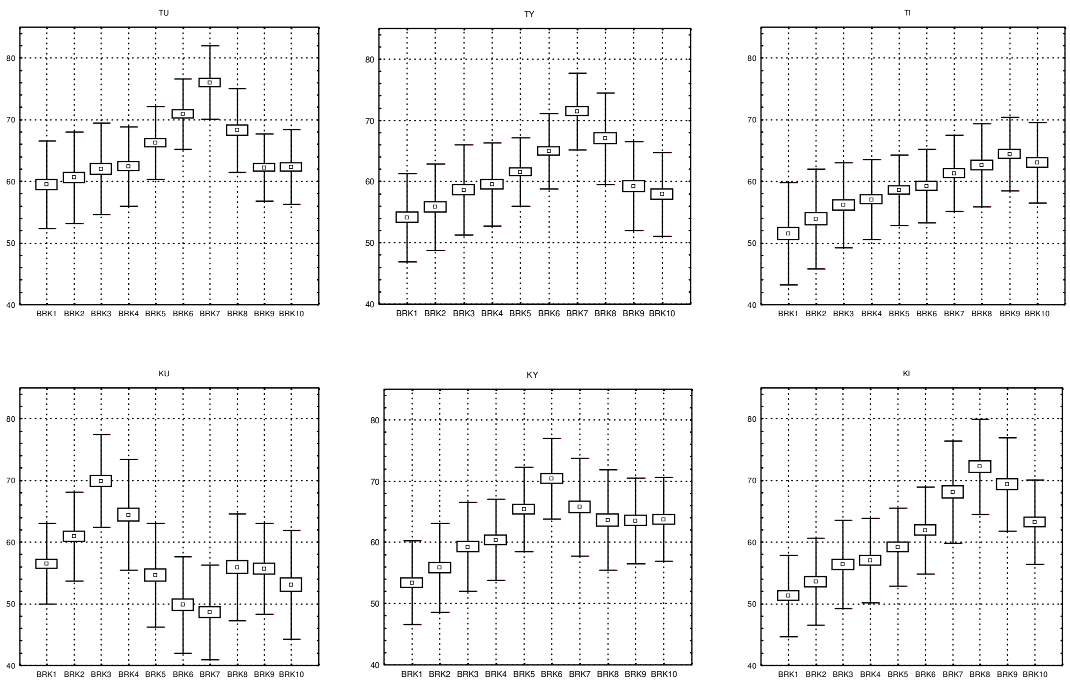


FIG. 5.41: Spectre moyen et écart type pour les 6 syllabes retenues dans le corpus restreint. (1 case représente 64 réalisations)

Jeu de paramètres	nb. par.	TA	i	y	u
BRK1..10	10	77.8%	53.1%	83.6%	96.9%
BRK1..10, FC1..4	14	78.1%	53.1%	85.1%	96.1%
S1..3	3	87.5%	78.9%	88.3%	95.3%
S1..3, F_centreS2	4	88.3%	82.0%	87.5%	95.3%

TAB. 5.24: Taux d'apprentissage sur le corpus restreint (1 exemple = 0.26%)

d'histogrammes. L'apprentissage sur tous les contextes a une influence négative plus faible sur la reconnaissance en contexte antérieur dans le cas des paramètres S1..3. Le prix à payer ici est visible sur la quatrième paire d'histogrammes : une moins bonne séparabilité quant l'apprentissage et le test sont faits sur un même contexte dans le cas des S1..3.

5.5.5.2 Du nouveau sur le paradoxe du N-seg en contexte [i]

Nous avançons donc qu'une bande relativement étroite de fréquence du bruit de relâchement porte de l'information importante pour la reconnaissance de la consonne. Comme nous l'avons vu, cette information est susceptible d'être moins variable avec le contexte que celle de toute la gamme des fréquences. Appelons cette bande de fréquence BFR.[13]

Nous remarquerons que ces résultats permettent d'expliquer le paradoxe du contexte [i] tel que

[13] "Burst frequency range".

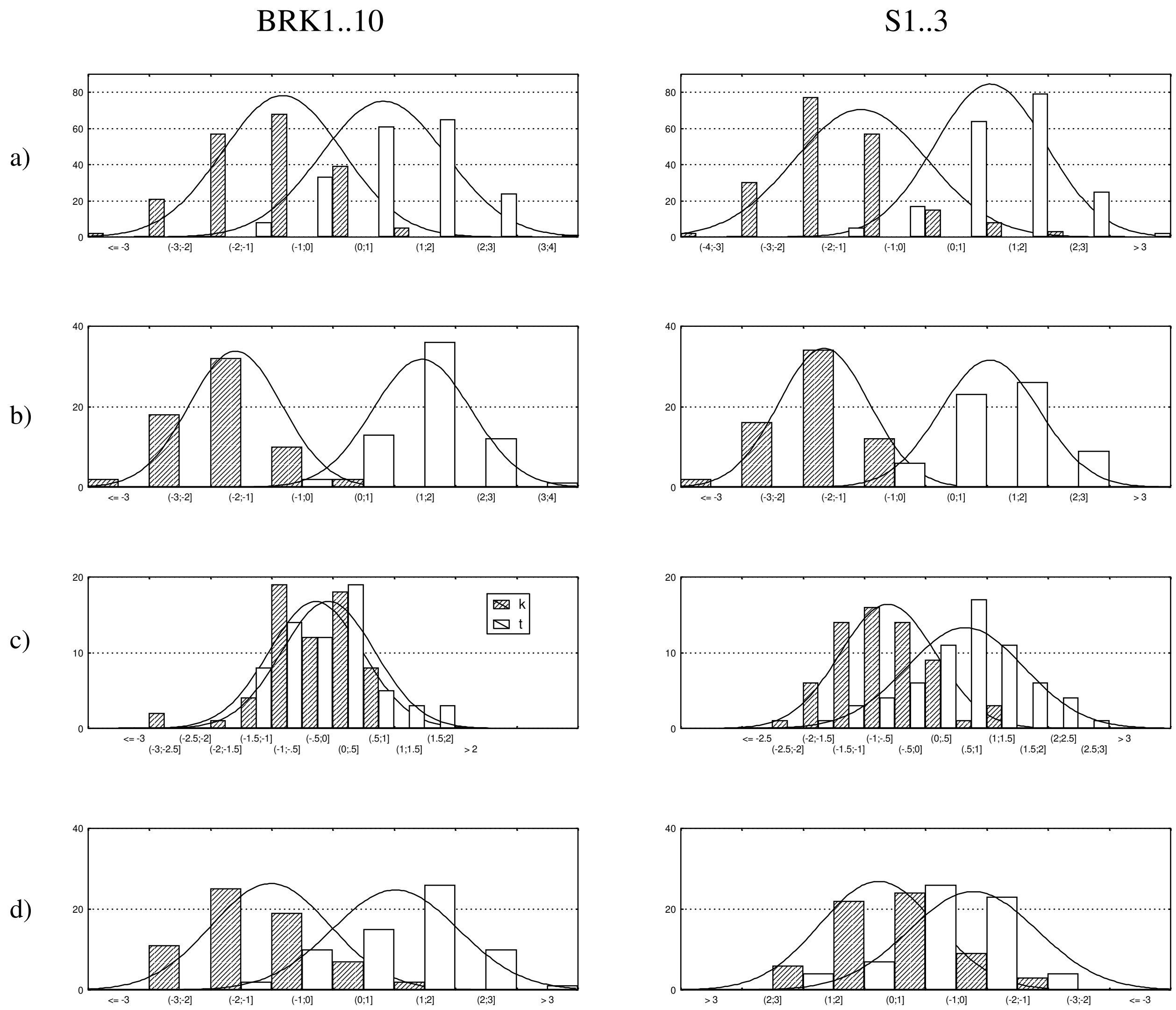


FIG. 5.42: Séparabilité des consonnes sur le corpus restreint (8 locuteurs, 6 syllabes) pour les paramètres BRK1..10 et pour les paramètres S1..3. Les histogrammes montrent le nombre de syllabes dans différents intervalles sur le premier axe canonique. Les courbes représentent l'approximation des histogrammes par des distribution gaussienes. Noter l'amélioration de la séparabilité si apprentissage sur [u i y] et test sur [i] (troisième ligne de graphiques). De haut en bas : a) apprentissage [u i y], test [u i y], b) apprentissage [u i y], test [u], c) apprentissage [u i y], test [i] et d) apprentissage [i], test [i].

Paramètres :		S1..3			BRK1..10		
appris sur	testé sur	taux global	k	t	taux global	k	t
u	u	99.2%	98.4%	100%	100%	100%	100%
y	y	89.8%	82.8%	96.9%	94.5%	90.6%	98.4%
i	i	82.0%	81.2%	82.8%	83.6%	85.9%	81.2%
Moyenne TA		**90.3%**	87.5%	93.2%	**92.7%**	92.2%	93.2%
u	i	58.6%	18.7%	98.4%	45.3%	1.6%	89.1%
u	y	75.0%	50.0%	100%	51.6%	3.1%	100%
y	u	92.9%	100%	85.9%	55.5%	23.4%	87.5%
y	i	81.2%	87.5%	75%	31.2%	25%	37.5%
i	u	87.5%	87.5%	87.5%	25.8%	0%	51.6%
i	y	86.7%	81.2%	92.2%	25.8%	14.1%	37.5%
Moyenne TR		**80.3%**	70.8%	89.8%	**39.2%**	11.2%	67.2%

TAB. 5.25: Taux de reconnaissance et d'apprentissage sur le corpus restreint (1 exemple = 0.78%, niveau aléatoire = 50%)

nous l'avons formulé dans le chapitre sur la perception. Nous rappelons que le N-seg a dominé la perception dans ce contexte sans qu'il soit très discriminant en soi. Notre étude des spectres (voir section "Indices du N-seg") et la sélection de paramètres présentée plus haut montrent que les spectres de [t] et [k] se différencient seulement dans la bande BFR devant [i]. Par contre, dans les autres contextes fermés, la différence n'est plus concentrée exclusivement dans BFR : les N-seg de [k] et [t] peuvent être donc reconnus indépendamment du V-seg devant [y] ou [u]. Si le V-seg apporte une information capitale par sa dynamique spectrale dans les contextes ouverts, il est aussi capital pour la perception en contexte [i] mais pour une autre raison : **indiquer la position de la BFR** dans le spectre de relâchement.

Bien sûr, le mécanisme de compétition, lié au pouvoir intrinsèque de discrimination de chaque segment, reste le plus important : son effet est massif et visible sur tous les contextes. D'ailleurs, Smits le considère comme unique base pour son interprétation :

> "... the direct recognition model[14] can indeed explain the observed seemingly vowel-dependent behaviour and explicit mechanisms, which fine-tune the stop-recognition to the vowel identity, are not required." [Smi96, page 2470]

Au contraire, nous montrons qu'un mécanisme fin de collaboration entre V-seg et N-seg est nécessaire pour éviter les erreurs d'identification en contexte antérieur et pour modéliser de manière précise la perception des occlusives dans ce contexte (voir les résultats des illusions acoustiques dans la section 5.7.1.3 et les difficultés rapportées par Smits pour le contexte [i]).

[14]Smits utilise une architecture similaire à notre fusion directe (FD).

5.5.5.3 Information mutuelle

Pour la suite de nos travaux, nous avons donc deux directions de recherche :

- La position en fréquence du BFR. Afin de l'identifier, nous avons fait recours au calculs d'information mutuelle. Ensuite nous comparons la position trouvé avec la position de formants au début du V-seg.
- Le codage du BFR. Nous avons étudié, à travers des tests d'apprentissage l'influence de différents facteurs pour le codage de cette portion du spectre de relâchement.

L'information mutuelle est une méthode qui peut servir pour étudier la corrélation entre deux variables aléatoires à valeurs discrètes. Introduite par Shannon en 1940, elle est définie [Jon79] comme la réduction de l'incertitude sur une variable en connaissant une autre variable. Dans cette définition, le degré d'incertitude est donné par l'entropie de la variable. L'entropie, introduite par L. Szilard en 1929, appelée également "self-information", est une mesure scalaire de la concentration ou de l'étalement de la distribution de probabilité d'une variable. Plus la distribution est étalée, plus faible est l'information intrinsèque de la variable et plus grande est son entropie. Soient X et Y deux variables discrètes et les intervalles $1..N_x$ et $1..N_y$ respectivement les domaines où elles peuvent prendre valeurs. L'entropie d'une variable est définie comme :

$$H(X) = -\sum_{i=1}^{N_x} p_i * log_2(p_i)$$

où p_i est la probabilité pour que X prenne la valeur i. Le minimum de l'entropie est atteint pour une variable constante (H(X)=0) et le maximum pour une variable équiprobable (H(X)=$log_2(N_x)$). On peut également définir l'entropie conditionnelle pour deux variables H(X|Y) et l'entropie mutuelle H(X,Y) :

$$H(X|Y) = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} p_{i,j} * log_2(p_{i,j}/q_j)$$

$$H(X,Y) = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} p_{i,j} * log_2(p_{i,j})$$

où q_j est la probabilité pour que Y prenne la valeur j et $p_{i,j}$ est la probabilité pour que (X,Y) = (i,j). L'information mutuelle est définie comme :

$$IM(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

Dans nos calcules, nous allons utiliser l'information mutuelle normalisée (avec valeurs dans l'intervalle de 0 à 1) :

$$IMN(X,Y) = \frac{IM(X,Y)}{log_2(min(N_x, N_y))}$$

Pour calculer l'information mutuelle entre l'occlusive et un spectre nous avons considéré Y={p, t, k}. Pour la variable X, nous prenons l'énergie dans une bande de fréquence du spectre de relâchement, par exemple BRK1. Comme cette énergie est une valeur continue, nous allons discrétiser l'intervalle de variation de BRK1 en N_x sous-intervalles égales. Nous obtenons ainsi une variable discrète, appelons la $B1_{N_x}$, variable qui prend la valeur i si BRK1 est dans le i-ème sous-intervalle.

Nous avons utilisé des discrétisations avec $N_x = 8$ et avec $N_x = 4$. Cette résolution très grossière pour l'amplitude nous est imposé par le nombre total réduit d'items dans notre corpus. En effet, le calcul de l'entropie est sensible à l'estimation des probabilités à partir des fréquences d'apparition, notamment pour les fréquences faibles. En pratique, plus le tableau de contingence est creux, plus grande est l'erreur systématique faite sur l'entropie. A. C. Morris a développé à l'Institut de la Communication Parlée [Mor92] une méthode pour compenser ce biais. Il montre que l'erreur est dans le sens de l'augmentation artificielle de l'information mutuelle. Basé sur ses simulations [Mor92, page 103], nous estimons que cette erreur est négligeable dans nos tests. En revanche, nous pouvons utiliser une séparation fine en bandes de l'échelle des fréquence. Cela revient à répéter le calcul pour différentes variables X.

Nous retenons de nos représentations des courbes d'information mutuelle, les **conclusions** suivantes :

– Le contraste entre le spectre du [p] et ceux des deux autres consonnes est concentré dans les basses fréquences (voir figure 5.43).
– Le contraste entre les spectres du [t] et du [k] est plus fort en contexte postérieur et diminue vers le contexte antérieur (voir figure 5.44). Ce contraste est basé sur deux zones de fréquence : une zone où le spectre du [k] est plus intense et une autre où le spectre du [t] est plus intense.
– La position de ces zones de fréquence covarie avec la hauteur de la voyelle (avant/arrière) et avec le degré d'ouverture (voir les figures 5.46, 5.47, 5.45). Cependant la zone du [t] est beaucoup plus stable.
– Le constat le plus important est qu'il n'y a plus qu'une seule zone intéressante pour les contextes i e ε a. Dans cette zone, le [k] est plus intense que le [t] malgré le fait que, dans les autres contextes, le [t] était le plus intense à cette fréquence.

La zone du [t] disparaît dans le contexte antérieur. Le maximum de l'information mutuelle situé dans les basses fréquences en contexte [u] et [y] est celui qui à la même nature que l'unique maxima trouvé pour le contexte avant. Si nous tenons à avoir un codage unique à travers les contextes, nous sommes amenés à négliger donc une zone du spectre beaucoup plus informative (maxima d'information mutuelle plus grand) des spectres en contexte [u] et [y].

Nous allons donc centrer la fenêtre fréquencielle d'analyse du spectre, BFR, sur la fréquence du maxima de l'information mutuelle (la fréquence la plus petite quand il y a deux maxima sur le graphique du IMN). Ce maxima correspond, de manière générale, avec le premier maxima spectral dans le spectre de relâchement du [k]. Nous retrouvons ainsi le concept de "fréquence du bruit" (BF) discuté dans la section dédiée au codage du N-seg. La différence importante est que, cette fois ci, BF est définie de manière statistique. En plus, elle ne correspond pas directement à un maxima du signal, mise à part pour la consonne [k].[15] Basé sur nos différents calculs d'information mutuelle,

[15] BF n'est pas ici un paramètre pour la classification des occlusives mais le support de l'information que le V-seg est censé fournir pour l'analyse du N-seg

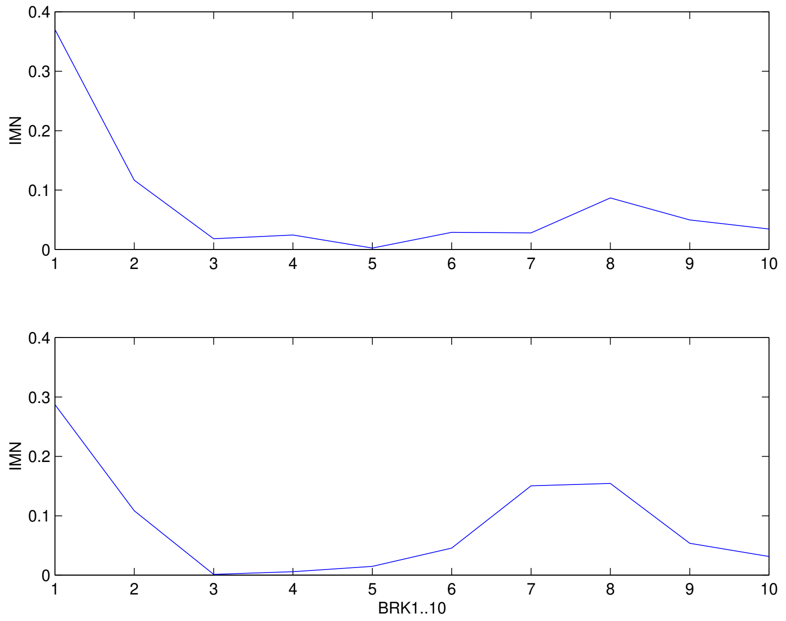


FIG. 5.43: Information mutuelle normalisé (IMN) entre les paramètres BRK et l'identité de la consonne. a) contraste [k]/[p]. b) contraste [t]/[p]. Les fréquences correspondant au paramètres BRK sont présentés dans le tableau 5.23.

nous proposons[16] pour BF les valeurs présentées dans la figure 5.49.

Le nombre pas assez important de CV dans notre corpus ne nous a pas permis d'étudier l'influence du locuteur sur la position du BF. Nous croyons que cette fréquence est spécifique au locuteur, à ses caractéristiques anatomiques comme à sa façon d'articuler les occlusives [t] et [k]. Nous avons la conviction que notre méthode pour le calcul du BF, appliquée à un corpus conséquent, ouvre une voie à la normalisation des consonnes.

5.5.5.4 Corrélation avec R2

Nous nous attendons à ce que la position du BF soit corrélée avec la valeur au début du V-seg de la première résonance de la cavité avant. En effet, le premier maximum du spectre de relâchement du [k] est aussi un reflet de cette même résonance. Notre méthode pour le calcul du BF n'est pas suffisamment précise pour pouvoir mesurer la relation exacte entre R2 et BF. Nous pouvons seulement observer que leur variation est semblable (voir tableau 5.26). Nous avons observé que la zone avec plus d'information mutuelle se place entre les valeurs du F3C et celles du F4C (plus proche de ces dernières) pour les contextes vocaliques [i] et [e] (voir figure 5.48).

[16]Notre BF est une proposition car elle est le résultat d'un choix sur un graphique non pas celui d'un calcul automatique d'indice.

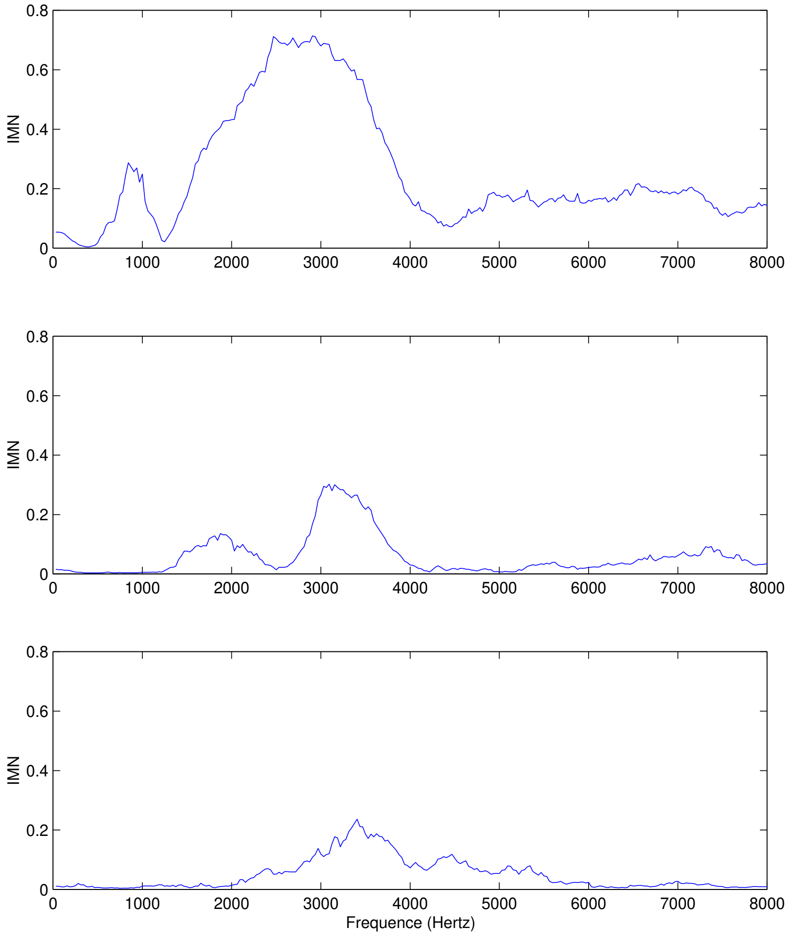


FIG. 5.44: IMN pour le contraste [k]/[t]. De haut en bas : a) contexte postérieur [u o], b) contexte labialisé [y ø], c) contexte antérieur [i e]. Nous remarquons la diminution de l'information du contexe postérieur vers le contexte antérieur. Nous remarquons aussi la séparation en deux zones (dans celle de gauche le [k] est plus intense).

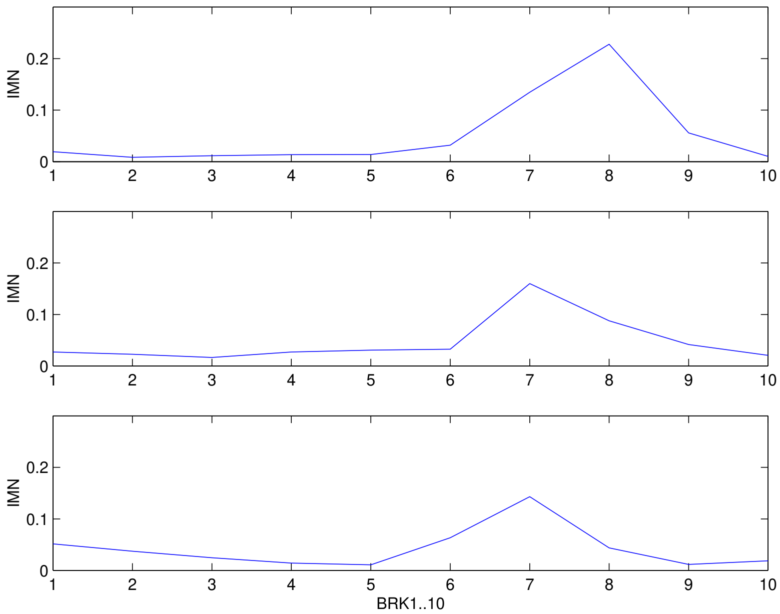


FIG. 5.45: IMN pour le contraste [k]/[t] en contexte antérieur et centrale. De haut en bas, les contextes [i], [e] et [a].

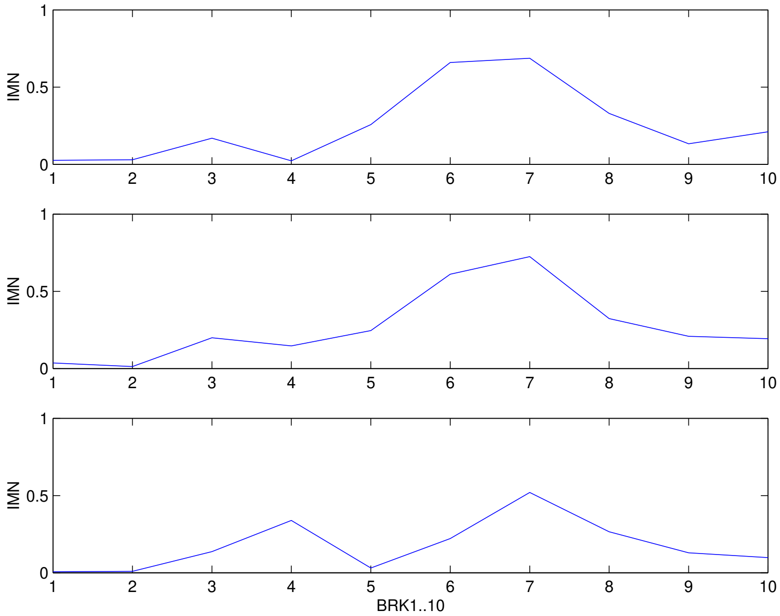


FIG. 5.46: IMN pour le contraste [k]/[t] en contexte postérieur. De haut en bas, les contextes [u], [o] et [ɔ].

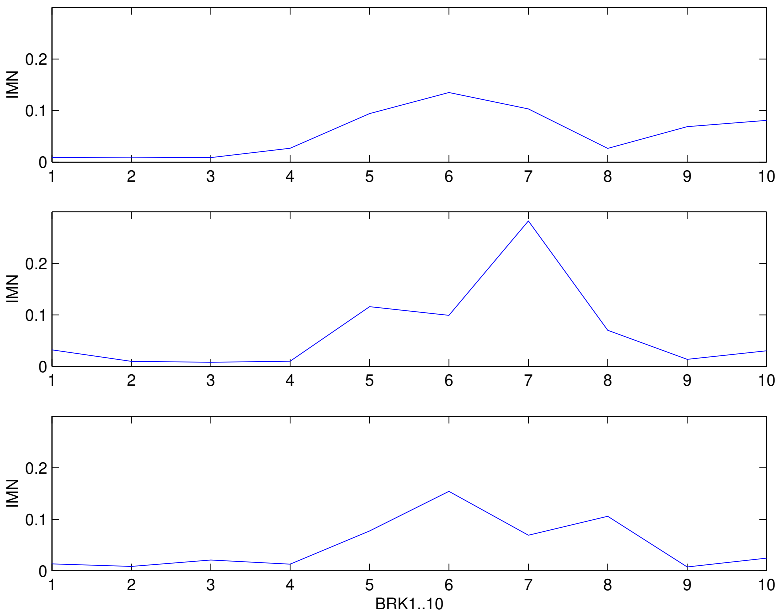


FIG. 5.47: IMN pour le contraste [k]/[t] en contexte antérieur labialisé. De haut en bas, les contextes [y], [ø] et [œ].

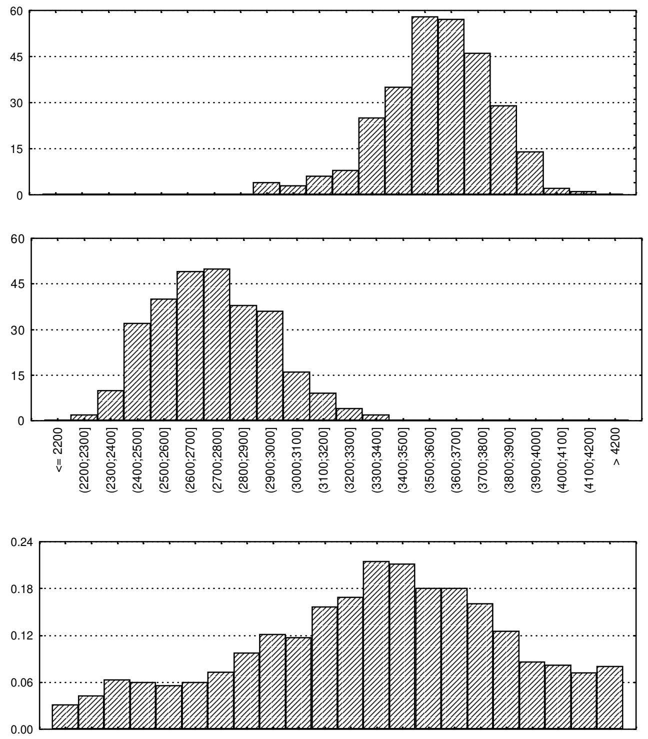


FIG. 5.48: Position relative en fréquence (Hertz) des formants au début du V-seg et du maximum IMN sur le spectre du relâchement pour les syllabes [ki], [ti], [ke] et [te] (256 items). De haut en bas : a) histogramme du F4C, b) histogramme du F3C, c) valeurs d'information mutuelle normalisée.

Voyelle	BF	R2C [k]	R2C [t]	R2C [p]	R2V
u	900	905 ±105	1325 ±160	1040 ±170	960 ±140
o	900	975 ±160	1335 ±95	945 ±130	980 ±120
ɔ	1170	1150 ±125	1420 ±95	1010 ±145	1160 ±140
a	2500	1865 ±155	1570 ±95	1230 ±160	1515 ±110
œ	2200	1720 ±190	1515 ±65	1290 ±115	1470 ±110
ø	1900	1680 ±90	1515 ±60	1340 ±115	1470 ±100
y	1900	1810 ±105	1775 ±140	1630 ±145	1830 ±100
ε	3150	2725 ±150	2520 ±115	2260 ±120	2620 ±120
e	3580	2780 ±150	2530 ±130	2260 ±130	2675 ±120
i	3800	2940 ±200	2805 ±135	2590 ±250	3060 ±150

TAB. 5.26: Comparaison entre BF, R2C et R2V selon la voyelle. Nous avons pris ici R2=F3 pour [i e ε] sinon R2=F2. Moyennes ± écart type, de 64 à 240 exemples par case. BF, R2C et R2V ont été testés comme possible centre de la fenêtre d'analyse du spectre de relâchement (voir tableau 5.28.

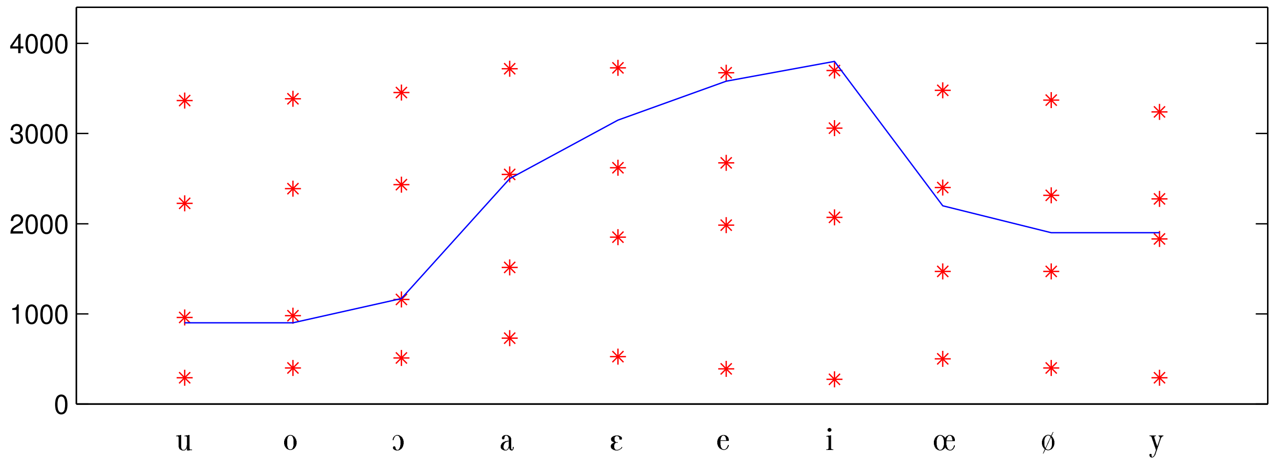


FIG. 5.49: Proposition pour BF (trait plein) basé sur les observations des maxima d'information mutuelle. Les fréquences moyennes sur notre corpus des premiers quatre formants (mesurés au centre de la voyelle) sont indiquées. Noter la variation dans le sens inverse de F2 pour les voyelles antérieures labialisées.

Architecture	codage	TA
N-seg	BRK1..10	77.8%
	LIN1..10	75.8%
	LIN1..20	79.7%
	CD1..10	75.5%
FP	DCT_sona1..21	89.3%
	DCT_sona1..45	94.0%

TAB. 5.27: Taux d'apprentissage sur le corpus restreint pour les codages mis au point sur le corpus entier. (1 exemple = 0.26%, niveau aléatoire = 50%)

5.5.5.5 Codage du BFR

Nous avons mené des tests d'identification d'abord sur le corpus réduit ([t] et [k] devant [i], [y] et [u]). Ensuite, nous avons testé les meilleurs jeux de paramètres sur le corpus entier. Nous avons essayé un nombre important de versions pour le codage d'une bande de fréquence. Ces codages tombent dans trois catégories :

1. Le codage par l'énergie dans plusieures bandes de fréquence (S1..n).
2. Le codage par centre de gravité spectral (F_G, A_G, D_G, R2C).
3. Le codage par transformé en cosinus discret (CD1..n).

Pour le premier type de codage, nous avons varié de manière **indépendante** le nombre de bandes (3 ou 5), la fréquence centrale de la bande du milieu (R2V, R2C, BF), la largeur de chaque bande (1, 2 et 3 barks), le degré de recouvrement (de 66% au -66%, le recouvrement négatif correspondant au bandes non-adjacentes) et l'ordre de l'analyse LPC (12 et 20). La largeur de la bande de fréquence analysée (BFR) a varié en conséquence de 3 à 10 barks. Concernant la fréquence centrale, BF sont les fréquences de la figure 5.49. R2V est F3 pour [i], [e] et [ɛ] et F2 pour les autres voyelles au centre de la voyelle. De même, R2C est le F3C ou le F2C, c'est à dire la fréquence des formants à la fin du suivi automatique de formants (sans aucune extrapolation).

Pour le deuxième type de codage, nous avons calculé la fréquence (F_G) et l'amplitude (A_G) du centre de gravité d'un segment du spectre de 4 barks centré sur R2C ou BF. Pour F_G nous avons retenu le calcul après normalisation entre 0 et 1 de ce segment de spectre. À ces deux paramètres nous avons rajouté D_G (la étendue en amplitude du segment du spectre) et R2C (le centre de la fenêtre analysée).

Pour le troisième type de codage, nous avons calculé la transformé en cosinus discret avec cinq coefficients (CD1..5) d'un segment de spectre de 4 barks centré soit sur BF soit sur R2C. Nous avons cherché si un sous ensemble de trois paramètre parmi CD1..5 garde autant d'information que S1..3.

Les résultats d'apprentissage que nous avons obtenus (dont une sélection succincte est présenté dans le tableau 5.28) appellent les commentaires suivants :

– Comme attendu,[17] centrer la bande d'analyse (BFR) sur BF a fourni les meilleurs taux.

[17]Le BF a été choisi, les bons taux autour de BF sont seulement une confirmation des calculs d'information mutuelle

Type codage	Centre du BFR	Caractéristique comparée	TA
S1..3	BF	bandes de 2 B, distance inter-bande 2 B	80.4%
	R2V		77.5%
	R2C		75.3%
S1..3	BF	LPC ordre 12	83.0%
	BF	LPC ordre 20	80.4%
S1..3	BF	bandes de 1 B, pas de recouvrement	71.6%
	BF	bandes de 2 B, pas de recouvrement	83.0%
	BF	bandes de 3 B, pas de recouvrement	87.8%
S1..3	BF	bandes de 2 B, distance inter-bande 1 B	76.0%
	BF	bandes de 2 B, distance inter-bande 2 B	83.0%
	BF	bandes de 2 B, distance inter-bande 3 B	87.5%
S1..n	BF	3 bandes de 2 B	83.0%
	BF	5 bandes de 1 B	83.8%
	BF	3 bandes de 3 B	87.8%
	BF	5 bandes de 2 B	86.7%
Centre de gravité	R2C	F_G, A_G, D_G, R2C	86.4%
	R2C	F_G, A_G, D_G	85.9%
	BF	F_G, A_G, D_G	83.8%
Transformé cosinus discret	BF	CD1..5	87.5%
	BF	CD2..4	83.0%
	R2C	CD2..5	82.3%
	R2C	CD2..5, R2C	81.2%

TAB. 5.28: Sélection de taux d'apprentissage sur le corpus restreint pour les codages inspirés par la fusion explicite. (1 exemple = 0.26%, niveau aléatoire = 50%)

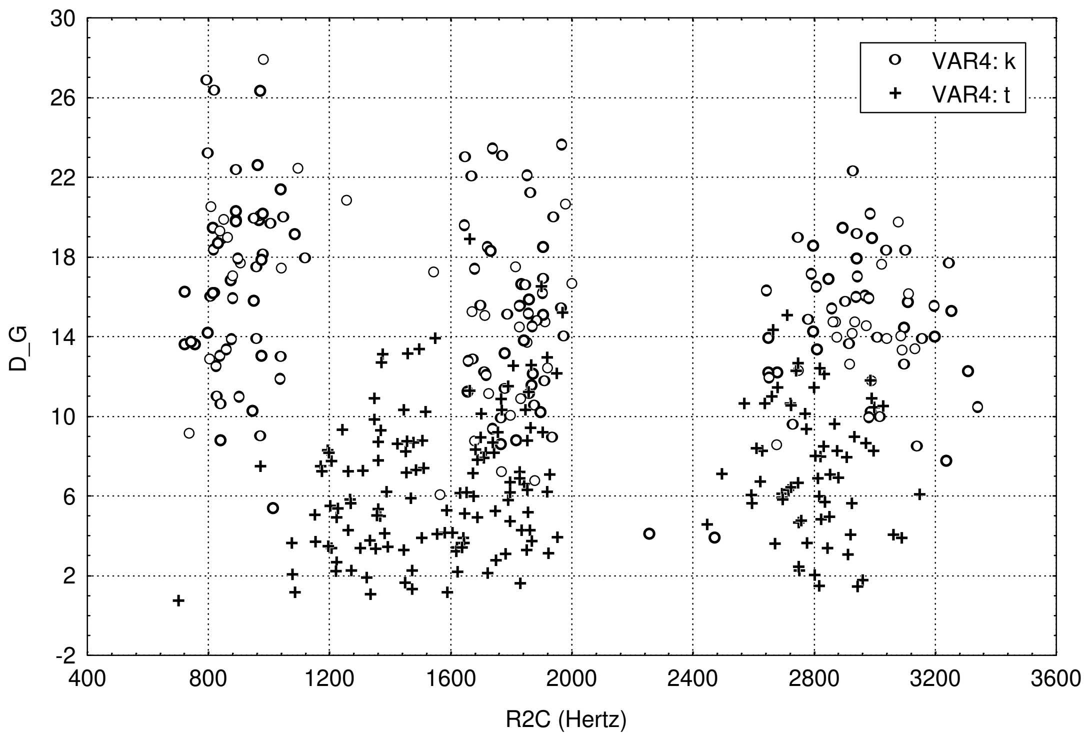


FIG. 5.50: Représentation des consonnes du corpus restreint dans le plan D_G, R2C (dynamique du spectre la bande analysée versus centre de la bande analysée). C'est la meilleure séparabilité sur le corpus restreint (384 syllabes) dans un espace à seulement deux paramètres.

- Contrairement à nos attentes, plus la position de BFR varie avec la consonne, moins bons sont les taux d'apprentissage. Les premières trois lignes du tableau 5.28 le demontre. En effet, BF ne varie pas avec la consonne, R2V est peu influencé par la consonne et R2C varie le plus avec la consonne.
- Le découpage en bandes plus fines n'apporte pas de gain.
- L'écartement des trois bandes (mêmes si elles sont étroites) a influencé toujours de manière positive la séparabilité.
- Le codage par centre de gravité est parmi les plus performants sur le corpus restreint. De manière surprenante, c'est le couple D_G, R2C qui assurent ces performances (voir figure 5.50). C'est le seul codage où le centrage sur BF dégrade le taux d'apprentissage.
- Pour le codage CD1..n, augmenter le nombre de paramètres améliore la capacité de séparation mais fait perdre l'avantage de ce type de codage.
- Le rajout de R2C ou BF, l'information sur la position centrale de la fenêtre d'analyse, n'apporte qu'une amélioration marginale.
- L'influence du pré-lissage des spectres (l'ordre de l'analyse LPC) est toujours présente sans que nous en trouvons une bonne interprétation.
- Le taux d'apprentissage ne dépasse pas le 90%, même sur le corpus restreint.

Codage	nb.par.	TA	k	t	p
S1..3	3	72.1%	74.5%	71.1%	70.7%
F_G, A_G, D_G, R2C	4	69.6%	82.6%	62.4%	63.8%
CD1..5	5	66.3%	74.0%	66.0%	58.7%

TAB. 5.29: Taux d'apprentissage sur le corpus entier selon la consonne pour les codages de la fusion explicite. (1 exemple = 0.042%, niveau aléatoire = 33.3%)

Un des meilleurs résultats sur le corpus restreint (TA = 87.5%) a été obtenu par l'énergie en trois bandes de 2 barks chacune située à 3 barks distance et ayant la bande S2 centrée sur BF (codage S1..3).

Les **conclusions** importantes que nous retenons de ces essais sont les suivantes :

1. La quasi-totalité des codages testés assurent sur le corpus restreint un taux meilleur que le codage de tout le spectre (parfois 10% en plus sur TA pour trois fois moins de paramètres). De même, ces codages approchent les taux de la fusion précoce mais avec cinq fois moins de paramètres (voir tableau 5.27).
2. Nous retenons BF comme vecteur pour la fusion explicite, paramètre à estimer à partir du V-seg et apte à filtrer l'information du N-seg.
3. Sur le corpus entier, la différence de comportement des codages FE par rapport aux autres est significative. Précisément, ici la consonne [k] est toujours la mieux séparée et [p] la plus confuse (voir tableau 5.29).
4. Malgré un rééquilibrage des performances entre les contextes vocaliques (voir tableau 5.24), l'ordre de performance des contextes demeure : [u] le meilleur, [i] le pire. Nous croyons donc qu'une spécialisation plus pointue de l'architecture de reconnaissance est nécessaire.

Le taux obtenu sur l'ensemble du corpus par S1..3, codage que nous retenons pour la FE, est faible (72,1%). Cependant, il convient de rappeler ici deux critères qui le placent en tête de classement :

1. Parmi tous les codages testés sur le corpus entier, S1..3 présente un des meilleurs rapports entre TA et nombre de paramètres.
2. S1..3 montre les meilleurs taux de reconnaissance en apprentissage sur un contexte et test sur un autre pour le contraste [t] versus [k].

Parmi les **inconvénients** de notre approche nous pouvons citer le fait qu'il s'agit plus d'une sélection explicite que d'une fusion explicite car l'information du V-seg n'est présente que de manière implicite. Un inconvénient important est le fait que notre BF, comme les équations du locus, n'est pas estimé à partir d'un seul item. Il est nécessaire donc d'approfondir les recherches sur la corrélation entre BF et le triangle vocalique des locuteurs, d'une part, et entre BF et la position des formants au début de la syllabe, d'une autre part. Un dernier inconvénient est lié à la quantité de données que nous avons utilisées pour nos études : nous n'avons pas de mesure de la signification statistique de nos dernières comparaisons.

Codage	nb.par.	TA	k	t	p
S1..3	3	72.1%	74.5%	71.1%	70.7%
S1..3 + VOT	4	77.3%	81.0%	70.8%	80.2%
CP_BRK1..2	2	74.4%	68.2%	65.2%	90.1%
S1..3 + BRK1..10	13	87.8%	86.0%	87.4%	90.1%
S1..3 + CP_BRK1..2	5	86.7%	81.4%	88.2%	90.5%

TAB. 5.30: Taux d'apprentissage sur le corpus entier selon la consonne pour codages complémentaires. (1 exemple = 0.042%, niveau aléatoire = 33.3%)

Codage	nb.par.	TA	TR
S1..3	3	72.37% ± 0.66	71.78% ± 4.69
BRK1..10	10	74.96% ± 1.02	72.74% ± 6.14
S1..3 + BRK1..10	13	87.78% ± 0.84	86.46% ± 6.37
S1..3 + CP_BRK1..2	5	86.76% ± 0.87	86.59% ± 5.94

TAB. 5.31: Taux de reconnaissance et d'apprentissage (moyenne ± écart type) pour codages complémentaires. Nous remarquons une faible dégradation entre TA et TR.

5.5.5.6 Complémentarité des indices. Tests de reconnaissance

Nous avons mentionné le fait que les codages que nous avons regroupé sous le nom "architecture à fusion explicite" sont complémentaires à ceux essayés précédemment. Les codages FE sont plus performants pour la séparation du [k] en contexte fermé, point faibles du codage de N-seg.

La façon la plus simple de palier le manque de performance du jeu S1..3 sur [p] est de lui rajouter le paramètre VOT. Ainsi 114 syllabes de plus sont correctement classés (voir tableau 5.30). Encore plus radicale est la combinaison avec le jeu BRK1..10. Afin de ne pas trop augmenter le nombre de paramètres nous avons combiné S1..3 avec CP_BRK1..2, les deux premières composantes canoniques du jeu BRK1..10 obtenus après une analyse discriminante canonique. Pour ces derniers codages, nous avons conduit un test de reconnaissance comme pour les jeux DCT_sona. Nous apprenons sur 7 locuteurs et nous testons sur le huitième par tour de rôle. Les résultats sont présentés dans le tableau 5.31. Nous notons des TR aussi bons que pour le codage DCT_sona et, surtout, une dégradation entre TA et TR nettement plus faible. Ces bonnes propriétés de généralisation sont dues au nombre réduit de paramètres qui élimine le danger de surapprentissage.

5.6 Quatrième pause

Notre analyse acoustique nous a conduit donc à un schéma pour décrire l'influence de la voyelle sur la consonne occlusive. Nous avons séparés, pour la première fois à notre connaissance, deux axes de comparaison :

1. L'axe /a/ - /u/. Les transition formantiques (donc un mouvement) pour /a/ et les spectre du relâchement pour /u/ sont les indices les plus discriminants et peuvent masquer les autres indices. L'information sur le lieu d'articulation est, entre /a/ et /u/, codée de manière physiquement différente. On peut modéliser cette différence comme un effet de compétition : là où la puissance de discrimination de certains indices baisse, d'autres indices prennent la relève.
2. L'axe /i/ - /u/. Pour toutes les voyelles fermées, le bruit de relâchement porte principalement l'information sur le lieu d'articulation. Si dans le cas du /u/ cette information est gardée en isolé, pour /i/ cette information se perd en dehors du contexte vocalique. Nous avons modélisé cette différence comme un effet de collaboration : là où le contraste s'affaibli le contexte est nécessaire pour préciser plus finement la zone d'analyse. Si la zone que nous avons identifiée n'est pas tout à fait celle du R2, nous avons encore toutes les raisons de croire qu'elle est déterminé par la cavité avant du conduit.

Nous devons imaginer maintenant les tests qui montreront si la théorie que nous avons présenté est valide. Si notre modélisation est correcte alors :

- Nous devrions être capables de produire des illusions acoustiques en manipulant des signaux naturels. Par exemple, il faudrait pouvoir obtenir un percept /k/ en partant d'un signal originel de /t/.
- Les algorithmes automatiques que nous avons retenus, quand ils sont appris sur les signaux naturels et testés sur les monstres, devraient montrer les mêmes réactions que les auditeurs humains.

5.7 Confrontation avec la perception des monstres acoustiques

5.7.1 Illusions perceptives

Nous nous sommes concentré sur le contraste /ti/ versus /ki/, le contraste le plus difficile dans nos tests de reconnaissance automatique. Nous avons voulu vérifier si des manipulations conforme à la morphologie que nous avons décrits peuvent changer le phonème perçu.

5.7.1.1 Construction des monstres

Nous avons construit deux catégories de monstres :

1. Nous avons manipulé le spectre de relâchement. Pour supprimer ou renforcer certaines fréquences nous avons appliqué la séquence : transformé de Fourier, pondération du spectre d'amplitude, transformé inverse. Nous appellerons donc ces monstres "ki_fft"et "ti_fft".
2. En utilisant la méthode de juxtaposition des N-seg et V-seg ("padding") nous avons construit des syllabes combinant le contexte /i/ et /y/. Nous les appellerons monstres bi-contexte. Nous avons présenté au auditeurs le monstre "ki_ky" (le N-seg de [ki] suivi du V-seg de [ky]) et le "ty_ti" (le N-seg de [ty] suivi du V-seg de [ti]).

Nous présenterons maintenant la construction des monstres "fft". Nous avons utilisé la technique OLA ("Overlapp and Add"). En utilisant deux fenêtres temporelles de longueur adaptée (voir figure 5.53), le N-seg a été décomposé dans un "signal gauche" et un "signal droite". Ainsi, la somme signal gauche plus signal droite restitue le signal d'origine. Pour obtenir les monstres le signal gauche à été modifié et ensuite additionné avec le signal droite non-modifié. Pour le monstre "ki_fft", le spectre de signal gauche a subi une atténuation avec une fenêtre de pondération de largeur effective 3,4 kHz centrée sur 4 kHz (voir figure 5.54), la fréquence que nous avons trouvé pour BF en contexte /i/ sur tous les locuteurs étant 3.8 kHz . Le spectre du monstre "ti_fft" a été pondéré par la fenêtre complémentaire à celle du "ki_fft" par rapport à 1. Les spectrogrammes des signaux originaux et transformés sont présentés sur les figures 5.51 et 5.52.

En plus de ces deux monstres, nous avons présenté aussi des syllabes [ki] avec une longueur du N-seg de la syllabe [ti]. Leur méthode de construction est identique à la précédente à deux exceptions près : les signaux gauche et droite sont plus courts et, avant l'addition, le signal gauche est déplacé à droite pour obtenir la longueur du N-seg de [ti]. Nous avons construit deux signaux, "ki_tronc" et "ki_fft_tronc", avec et sans pondération du spectre.

5.7.1.2 Prédit-on les monstres bi-contexte ?

Le premier test avec monstres naturels bi-contexte que nous connaissons est due au M. F. Dorman et collaborateurs [DSKR77]. Ils combinent le N-seg[18] d'une syllabe avec le V-seg des tous les autres syllabes avec la même consonne mais une voyelle différente. Dans leurs tests ([b d g], 9 voyelles de l'anglais, 1 locuteur, 8 auditeurs) ils constatent que les N-seg gardent généralement leur identité phonétique :

> "The rank order correlation [...] between performances elicited by transposed and untransposed burts is [...] a mesure of burst commutability or functional invariance." [DSKR77, page 118]

Selon leurs prédictions, nos monstres bi-contexte serons perçus comme la consonne qui a fournit les deux segments. Le monstre "ty_ti", par exemple, devrait être perçu comme [ti]. Suite à notre propre analyse, nous avons affirmé que le spectre de relâchement du [k] emprunte une forme typique de [t]. Autrement dit, un [t] d'un autre contexte devrait "faire l'affaire" comme N-seg de [ki], avec ou sans des transitions formantiques divergentes. Selon notre théorie, le monstre "ty_ti" devrait être perçu comme [ki].

Concernant les monstres "fft", on s'attend à ce que le VOT n'ait pas d'effet. On s'attend également à ce que ces monstres soient plus rejetés par rapport aux monstres bi-contexte car ceux-là utilisent des segment naturels. Pour récapituler, nous attendons les réponses données dans le tableau 5.32.

[18] N-seg avec aspiration remplacé avec silence.

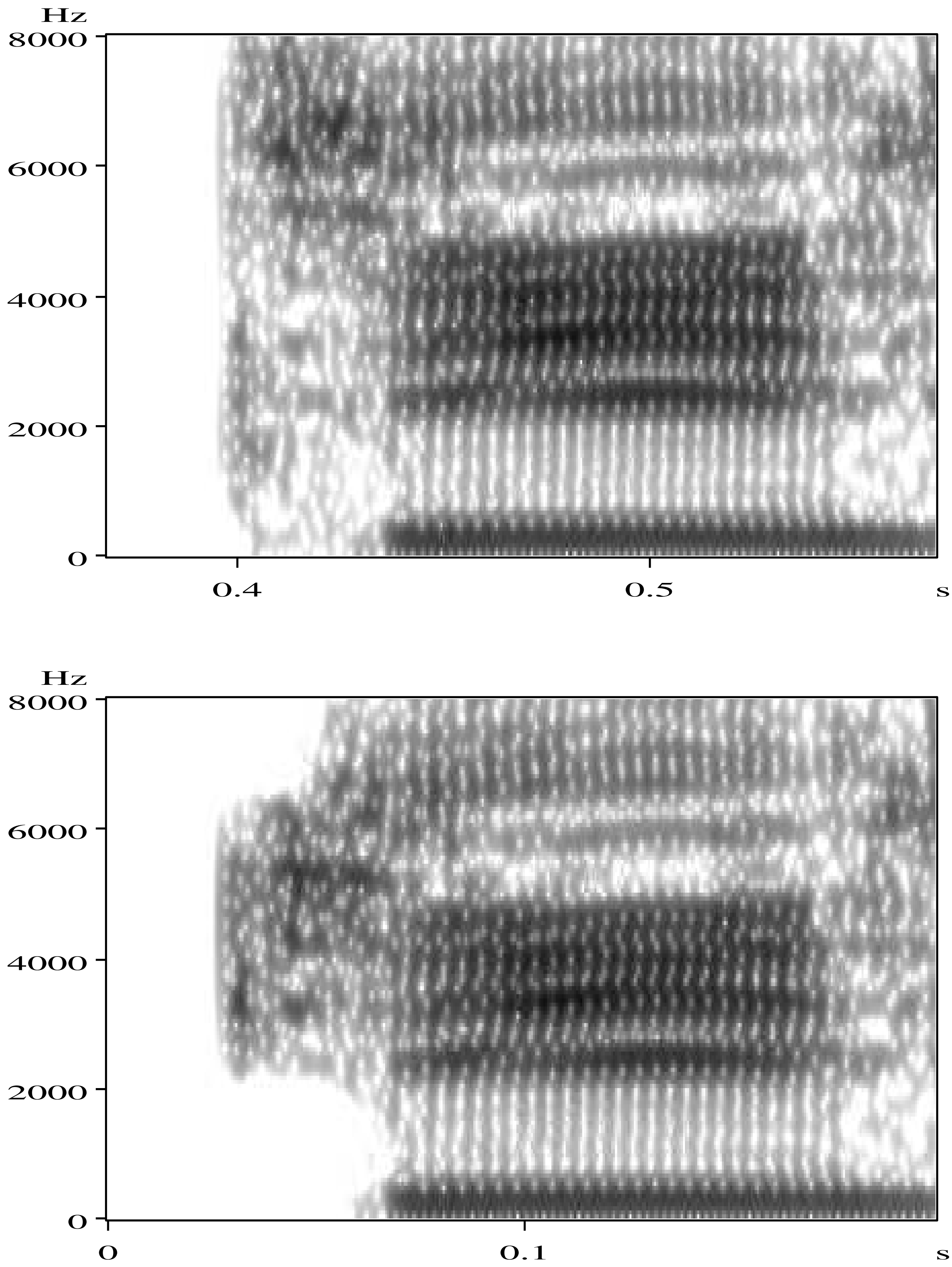


FIG. 5.51: Spectrogrammes des syllabes [ti] (en haut) et [ti] transformé (en bas). 200 ms de signal

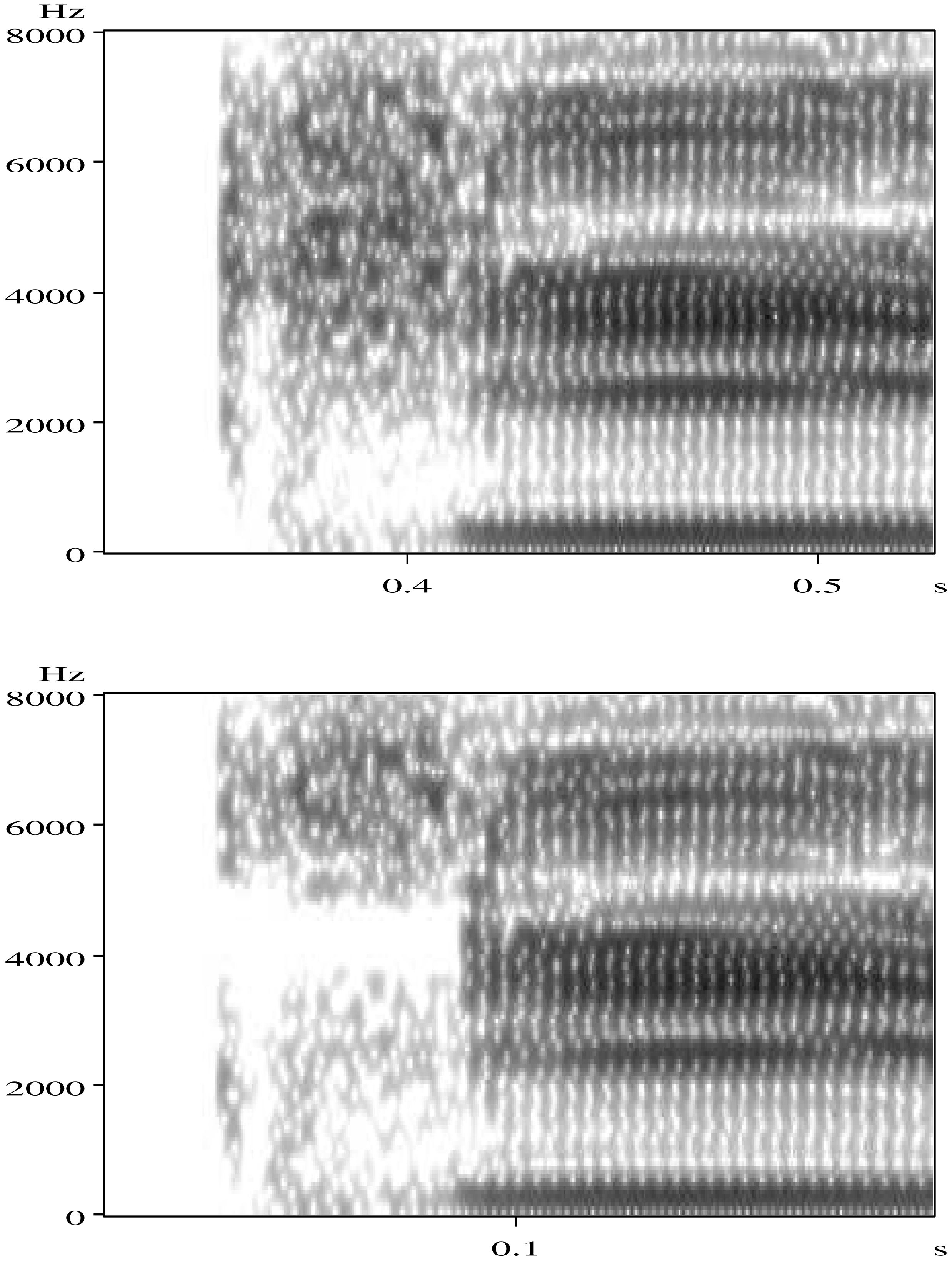


FIG. 5.52: Spectrogrammes des syllabes [ki] (en haut) et [ki] transformé (en bas). 200 ms de signal

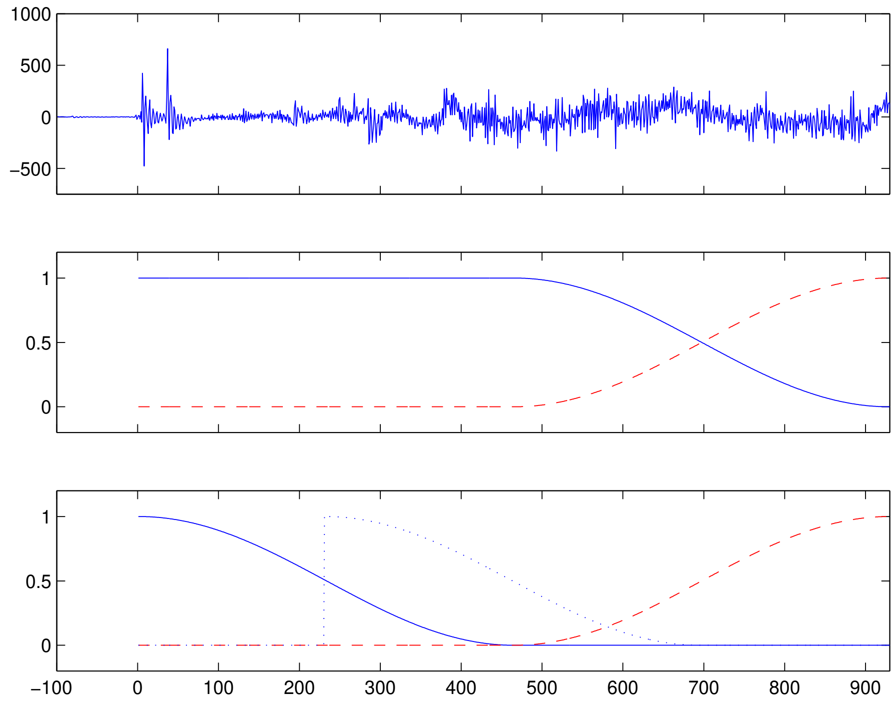


FIG. 5.53: OLA avant modification du spectre. De haut en bas : a) oscilogramme du N-seg du [ki], b) fenêtres pour extraction du signal de gauche (trait plein) et signal de droite (trait interrompu) et c) fenêtres pour extraction du signal de gauche et de droite dans le cas raccourciment. En pointillé la fenêtre de gauche après déplacement

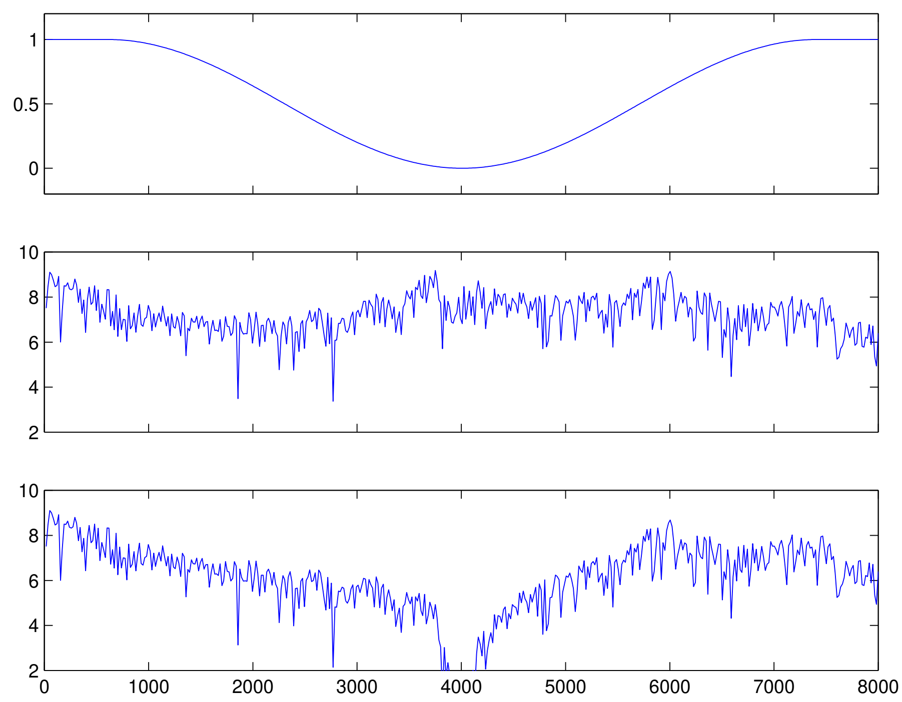


FIG. 5.54: Pondération spectrale du N-seg. De haut en bas : a) fenêtre de pondération (largeur effective 3.4 kHz), b) Spectre original et c) Spectre pondéré (en dB)

Monstre	Réponse
ki_ky	t
ty_ti	k
ti_fft	k
ki_fft	t
ki_fft_tronc	t
ki_tronc	k

TAB. 5.32: Réponses attendues dans le test perceptif des illusions acoustiques

Monstre	[k]	[t]	rejet
ki_ky	1	14	0
ty_ti	13	2	0
ti_fft	4	1	10
ki_fft	4	10	1
ki_fft_tronc	1	14	0
ki_tronc	15	0	0

TAB. 5.33: Réponses obtenues dans le test perceptif des illusions acoustiques

5.7.1.3 Résultats des tests

Nous avons présentés ces monstres aux 15 auditeurs qui ont participé aussi à nos derniers tests perceptifs. Ils doivent faire un choix parmi trois choix possibles : [k], [t] ou "ni [t] ni [k]" (refus de la classification). Les résultats sont présentés dans le tableau 5.33. Nous pouvons remarquer que le monstre "ti_fft" a été rejeté : nous l'écartons de nos conclusions. La majorité des auditeurs déclare qu'il est "incompréhensible".

Nous retenons les **conclusions** suivantes :

- Le N-seg n'est pas "commutable" entre les contextes.
- Le mécanisme de collaboration entre N-seg et V-seg reçoit ici une preuve perceptive additionnelle (en son absence il est impossible d'arriver à un percept [k] en partant de deux segment du [t]).
- Le contraste [ki]/[ti] pour le locuteur étudié ici est basé sur une zone de fréquence haute (à environ 4 kHz).

Nous avons demandé à nos architectures de classer quatre items de ce test (items qui n'ont jamais été rejetés). L'apprentissage a été effectué sur les 2208 syllabes naturelles. Les résultats sont présentés dans le tableau 5.35. Nous observons que l'architecture à fusion explicite (FE) est la seule à répondre comme les humains pour ces quatre signaux.[19] Cette performance conforte l'hypothèse

[19] Un test sur un item ne peut pas être une preuve définitive. Nous avons à l'esprit la reproduction du test sur un corpus plus important.

Architecture	item	réponse	$P(k)$	$P(p)$	$P(t)$
V-seg	ki_ky	k	0.75	0.04	0.21
	ty_ti	k	0.76	0.08	0.16
	kisi_tronc	k	0.49	0.06	0.45
	kisi_fft_tronc	k	0.85	0.15	0
N-seg	ki_ky	k	0.82	0	0.18
	ty_ti	t	0.17	0	0.83
	kisi_tronc	t	0.12	0.33	0.55
	kisi_fft_tronc	p	0	0.99	0.01
FT_PP	ki_ky	k	0.79	0	0.19
	ty_ti	t	0.35	0	0.37
	kisi_tronc	t	0.22	0.16	0.51
	kisi_fft_tronc	p	0	0.44	0
FD	ki_ky	k	0.93	0	0.07
	ty_ti	t	0.19	0	0.81
	kisi_tronc	t	0.33	0.2	0.46
	kisi_fft_tronc	p	0	1	0
FP	ki_ky	k	0.67	0	0.33
	ty_ti	t	0.32	0	0.68
	kisi_tronc	k	0.61	0.11	0.28
	kisi_fft_tronc	p	0	1	0
FE	ki_ky	t	0.2	0.34	0.46
	ty_ti	k	0.59	0.32	0.08
	kisi_tronc	k	0.5	0.15	0.35
	kisi_fft_tronc	t	0	0.25	0.75
auditeurs	ki_ky	t	0.07	-	0.93
	ty_ti	k	0.87	-	0.13
	kisi_tronc	k	1	-	0
	kisi_fft_tronc	t	0.07	-	0.93

TAB. 5.34: Prédiction des réponses par les algorithmes automatiques. Apprentissage sur 2208 CV, test sur 4 CV présentés dans le dernier test perceptif. Les décisions des l'algorithmes et les probabilités d'appartenance à chaque classe sont indiqués.

Architectures	ki_ky	ty_ti	kisi_tronc	kisi_fft_tronc
V-seg	k	k	k	k
N-seg (BRK1..10)	k	t	t	p
FT_PP	k	t	t	p
FD	k	t	t	p
FP (DCT_sona1..21)	k	t	k	p
FE (S1..3)	t	k	k	t
auditeurs	t	k	k	t

TAB. 5.35: Recapitulatif des résultats du tableau précédent 5.34 : seule la fusion explicite reproduit les réponses des auditeurs.

selon laquelle le jeu de paramètres S1..3 de la fusion explicite "lit" le spectre du relâchement de la même manière que les auditeurs.

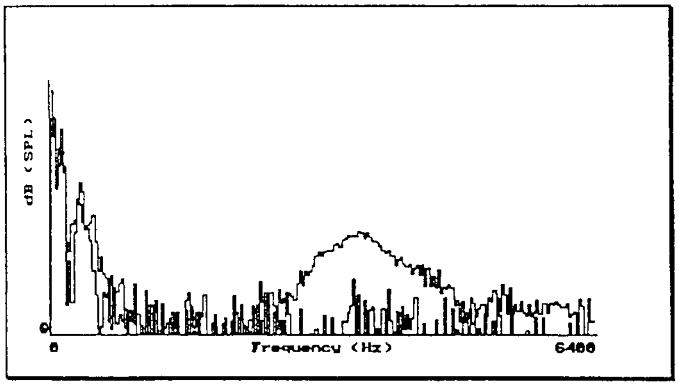


FIG. 5.55: Spectre de relâchement de [ti] et de [ki] (d'après Plauché et collaborateurs [PDO97]).

Les réponses de nos auditeurs doivent être mis en correspondance avec les résultats récents de M. Plauché, C. Delogu et J. J. Ohala qui ont monté une expérience similaire à Berkeley [PDO97]. Ils ont filtré la zone de 2,5 kHz à 5 kHz du N-seg de [ki] (voir figure 5.55) par un filtre coupe-bande Hanning d'ordre 10 et ils l'ont rajouté au V-seg de [ki] non-modifié. Leurs 15 auditeurs perçoivent ce signal comme un [ti] en unanimité. Ils ont montré aussi qu'un N-seg amplifié de [pi] conduit à un percept de [ti]. Leurs tests de perception ont été conduit sur les occlusives de l'espagnol et de l'italien donc des occlusives proches des [p t k] du français. Ces résultats rapportés par Plauché conforte notre théorie sur l'importance de la collaboration des indices des deux segments dans le contexte vocalique [i].

5.7.2 Importance perceptive des meilleurs indices acoustiques

Dans cette deuxième étape de notre confrontation, nous voulons chiffrer la capacité des algorithmes étudiés à prédire les réponses aux monstres. Dans cette partie, nous accordons plus d'at-

Monstres	[k]	[t]	[p]	total
conflit	25	21	14	60
niveau	42	9	9	60
vot	22	10	23	55
total	89	40	46	175

TAB. 5.36: Tableau de contingence des réponses de référence aux monstres acoustiques

tention à la fusion précoce.[20] Les faiblesses du codage par transformée temps-fréquence en cosinus discret seront mises en évidence.

5.7.2.1 Présentation des références

Les 175 stimuli que nous avons utilisés comme corpus de test par la suite seront groupés en trois catégories :

- Monstres conflictuels mono-contexte simples (N-seg et V-seg de consonnes différentes mais même voyelle).
- Monstres conflictuels avec amplitude du N-seg modifiée.
- Monstres conflictuels avec durée du N-seg modifiée.

Nous appellerons ces catégories respectivement "conflit", "niveau" et "VOT". D'autres facteurs qui influent sur la performance de prédiction, à part l'appartenance à une de ces trois catégories, n'ont pas été étudié en raison du nombre réduit d'items.[21]

Pour simplifier la comparaison entre series de réponses, nous allons retenir comme identité phonétique pour chaque monstre la réponse la plus fréquente parmi nos auditeurs ("winner takes all policy"). Nos algorithmes automatiques ne peuvent pas rejeter un item. Cela nous a obligé à prendre, pour les deux monstres rejetés, "pe_ke" et "pi_ti", la deuxième réponse la plus fréquente, en occurrence [p]. Le tableau 5.36 récapitule la proportion des identités phonétiques de référence des monstres dans chaque catégorie. Nous remarquons que des réponses [p] sont présentes dans le test "niveau" malgré le fait qu'il contient seulement des monstres en [t] et [k].

5.7.2.2 Tentative de prédiction des monstres acoustiques

Nous apprenons nos classifieurs soit sur les 2208 CV, soit sur les 276 CV du locuteur qui a prononcé les monstres et nous testons sur les 175 items qui ont été présentés effectivement dans les tests de perception. Pour les deux architectures de la fusion précoce, nous avons comparé les

[20] C'est l'architecture la plus performante dans l'absolu pour la reconnaissance des items naturels.

[21] Il serait intéressant de savoir si l'effet de fusion de type McGurk est reproduit et par quel algorithme. Par exemple, le monstre "kɛ_pɛ", perçu [t] par nos auditeurs, a été reconnu comme [t] par FD, les deux DCT_sona et par N-seg. Cet item a été reconnu comme [p] par V-seg et FT_MP et comme [k] par FT_CP et FT_PP. De même, il serait intéressant de savoir si les monstres d'un certain contexte vocalique sont mieux prédits que celui d'un autre et par quel algorithme.

réponses des algorithmes non seulement aux réponses des auditeurs mais aussi à l'identité du N-seg et à l'identité du V-seg.

À la suite des nos tests, nous faisons les remarques suivantes :

– Les algorithmes automatiques **ne prédisent pas** les réponses des auditeurs. Pour toutes les architectures étudiées, plus d'un tiers des monstres ne sont pas prédits (voir tableau 5.37). En plus, le meilleur codage (DCT_sona) paye sa performance par une sensibilité importante au locuteur.

– Cet échec n'est pas dû à une impossibilité d'apprentissage des exemples de test : la différence entre TA et TR est importante. Nous en déduisons que les exemples de test ne sont pas distribués de manière aléatoire mais seulement différemment par rapport à ceux d'apprentissage.

– Concernant la sensibilité au corpus d'apprentissage, nous nous attendons à ce que les codages ayant trait au V-seg (V-seg, DCT_sona, SIRB 40ms, FD, FD) soient plus performants pour l'apprentissage sur le locuteur JS, malgré la réduction du nombre des items appris. Seul le codage V-seg ne verifie pas cette hypothèse.

– Les codages basés principalement sur N-seg (BRK1..10, SIRB, S1..3) ont une difficulté certaine à apprendre les monstres (TA faible). Nous y voyons l'effet des monstres qui ont été perçus comme leur V-seg malgré un N-seg discriminant, comme certains monstres perçus [p] dans la catégorie "VOT".

– L'ordre de performance en prédiction des monstres montre des différences notables par rapport à celle sur items naturels : FT meilleure que FD, V-seg meilleur que N-seg, FE pas plus performante que N-seg.

– Sur les neuf changements perceptifs vers [p] constatés dans le test "niveau", aucun n'est prédit (voir tableau 5.38). Nous constatons que les algorithmes automatiques répondent [t], tout comme ce que les auditeurs avaient répondu pour ces monstres avant l'atténuation de l'amplitude. Pour reproduire cet effet, l'utilisation d'un classifieur non-linéaire est souhaitable. De même, la combinaison de la détection avec la classification est susceptible de mieux reproduire ces changements.

– Les monstres de type "VOT" sont moins bien prédits (surtout ceux perçus [k]). L'influence néfaste de la manipulation de la durée du N-seg nous conduit à la conclusion que les algorithmes de la FP sont **sensibles à un déphasage** entre les deux segments, N-seg et V-seg, dans les signaux de test par rapport à ceux d'apprentissage.

– L'architecture à fusion précoce donne un poids au N-seg plus grand que celui accordé par les auditeurs (voir tableau 5.39). Avec plus de syllabes en contexte ouvert, les résultats des prédictions seraient donc encore plus mauvais !

– Comme attendu, dans tout les cas, le poids du V-seg a augmenté quand le VOT a été modifié.[22].

[22]Nous rappelons que notre modification du VOT est dans le sens de sa mise en accord avec l'identité phonétique du V-seg

Architecture		TA	TR1	TR2
N-seg	BRK1..10	67.4%	53.2%	48.0%
V-seg		78.3%	56.0%	54.9%
FD		88.6%	53.7%	56.6%
FT_MP		77.7%	60.0%	61.2%
FT_CP		80.6%	58.9%	61.2%
FT_PP		78.9%	61.2%	61.7%
FE	S1..3	64.0%	56.0%	56.0%
FP	DCT_sona1..21	82.8%	53.2%	64.0%
	DCT_sona1..45	89.2%	57.2%	65.2%
	SIRB 26ms	65.2%	53.2%	52.6%
	SIRB 40ms	65.7%	49.7%	54.3%

TAB. 5.37: Taux de prédiction (TR) et taux d'apprentissage (TA) des monstres acoustiques. Pour TA, apprentissage sur les 175 monstres ; pour TR1, apprentissage sur 2208 syllabes des huit locuteurs ; pour TR2, apprentissage sur 276 syllabes du locuteur qui a fournit les monstres ; test sur les 175 monstres (1 exemple = 0.57%). Il faut noter la forte dégradation entre TA et TR pour toutes les architectures.

– Dans une tentative d'améliorer la prédiction des monstres de la catégorie "VOT", nous avons ajouté le paramètre VOT aux jeux de la FP. Ceci a apporté une amélioration notable dans un seul cas, le codage de type SIRB.

5.7.2.3 Comparaison qualitative

Dans une tentative de chercher les facteurs qui ont conduit à cette faible performance de prédiction, nous avons mené des tests sur des monstres simulés.[23] Nous retenons de ces tests les conclusions suivantes :

Dépendance de la voyelle : Comme nous l'avons montré, les pouvoirs de discrimination de N-seg et de V-seg varient fortement et de manière complémentaire avec le contexte vocalique. Nous constatons que les algorithmes appris sur les CV naturelles reproduisent bien cette compétition entre N-seg et V-seg selon la voyelle dès que des paramètres des deux segments sont présents en entrée. Malgré une certaine variabilité dans les proportions, nous pouvons noter pour les architectures à fusion que le N-seg domine les réponses en contexte [i u y] et V-seg domine celles en contexte [a œ], comme pour nos auditeurs. Le mécanisme de compétition doit être completé donc par d'autres pour rendre compte des résultats perceptifs.

Dépendance de la consonne : À la différence de la perception humaine, le N-seg de [p] est très

[23]Ces monstres sont obtenus par combinaison des paramètres N-seg d'une consonne et des paramètres V-seg d'une autre dans une base de donnée, sans construction effective des signaux acoustiques. Ils sont plus nombreux mais il ne permettent pas l'étude de toutes les architectures.

Codage	Type monstres	TR			
		[k]	[t]	[p]	total
DCT_sona1..21	conflit	64%	62%	64%	63%
	niveau	81%	67%	**0%**	67%
	VOT	50%	60%	65%	58%
DCT_sona1..45	conflit	60%	71%	64%	65%
	niveau	75%	55%	**0%**	62%
	VOT	41%	70%	70%	58%

TAB. 5.38: Influence de la consonne et du type de monstres sur la prédiction des réponses des auditeurs. Les réponses aux monstres "VOT" perçus [k] sont moins bien prédits. Les changement vers [p] dans la catégorie de monstres "niveau" ne sont pas prédits. 60 monstres "conflit", 60 "niveau" et 55 "VOT".

Codage	Type monstres	Pourcent de coïncidence avec :		
		N-seg	V-seg	auditeurs
DCT_sona1..45	conflit	**72%**	20%	65%
	niveau	**70%**	30%	62%
	VOT	31%	**67%**	58%
SIRB 26ms	conflit	83%	8%	65%
	niveau	57%	22%	53%
	VOT	58%	29%	42%
auditeurs	conflit	**56%**	35%	100%
	niveau	**57%**	28%	100%
	VOT	37%	**53%**	100%

TAB. 5.39: Poids du N-seg et du V-seg dans les réponses des algorithmes automatiques aux monstres acoustiques : soient 60 monstres "conflit", 60 "niveau" et 55 "VOT". Le pourcent de coïncidence est une mesure de corrélation entre les séries de réponses. Si le codage DCT_sona se rapproche le plus de la balance montrée par les auditeurs, sa sensibilité est pourtant exagérée. Les codages de type SIRB accordent toujours un poids trop important au N-seg.

prégnant pour les algorithmes testés, comme nous l'avons constaté aussi sur les syllabes naturelles. Parmi les monstres de la catégorie "conflit" ayant un N-seg de [p], plus de deux tiers sont donnés pour [p] par les algorithmes automatiques et un tiers seulement sont perçus comme [p] par les auditeurs. Cette prégnance trop grande dans la catégorie "conflit" est accompagnée par une prégnance trop faible du [p] dans la catégorie "niveau". Nous pensons que le fait de forcer le traitement à des instants données a priori (et non pas issus d'une détection) est une de causes qui affecte de manière négative la capacité de nos algorithmes à modéliser la perception du [p].

5.7.2.4 Monstres et apprentissage des systèmes automatiques

Les tests que nous avons présentés jusqu'ici partent de l'apprentissage d'items naturels. L'idée de base était de comparer le système automatique au système de perception humain. L'objectif est de changer l'architecture du système automatique pour la rendre plus performante. Mais il existe aussi un autre approche permettant de mettre en valeur les résultats de perception pour la reconnaissance automatique.

Cette approche est complémentaire à celle que nous avons utilisé plus haut. Il ne s'agit plus d'améliorer l'architecture de reconnaissance mais d'apporter une augmentation de performance par le contrôle des données d'apprentissage. À l'inverse de la situation précédente, les monstres acoustiques seront présentés à l'apprentissage et le test sera fait sur des items naturels. L'intérêt de cette approche est la promesse d'un meilleur pouvoir de généralisation à partir d'un corpus d'apprentissage réduit. En effet, les monstres acoustiques sont les cas extrêmes, très informatifs pour les contrastes, que le classifieur a peu de chances de rencontrer dans un corpus d'items naturels. Par exemple, dans le cas des occlusives, cela peut apporter l'information que les transitions formantiques en contexte /u/ peuvent être négligés, malgré leur puissance de discrimination non négligeable.

Sur nos données, nous n'avons pas obtenu une amélioration en utilisant divers combinaisons d'items (monstres seulement ou monstres plus items naturels) en apprentissage. Les raisons sont vraisemblablement liées à la disproportion entre monstres et items naturels (175 CV, un seul locuteur contre 2208 CV, 8 locuteurs) et au classifieur baysien (relativement peu influencé par les items périphériques isolés). Cependant, nous croyons qu'une amélioration pourrait être constatée sur un corpus conçu de manière plus equilibrée.

5.7.3 Bilan de la confrontation des architectures avec les monstres acoustiques

Nous allons, en guise de conclusion, présenter un état de l'art des recherches d'un point de vue spécial. Le fil conducteur de cet état est l'observation suivante : les meilleurs indices acoustiques, tels qu'ils ont été utilisés pour la reconnaissance automatique des occlusives, ont **systématiquement** montré, jusqu'à ce jour, un faible rendement explicatif des résultats des expériences de perception. En effet :

Forme du spectre du relâchement Stevens et Blumstein proposent en 1978 des prototypes pour la forme du spectre du relâchement [SB78]. Sur la base de cette forme, K. Suomi obtient des taux d'identification remarquables [Suo85].[24] Néanmoins, plusieurs tests perceptifs utilisant des monstres conflictuels ont montré que la forme de ce spectre peut être facilement masquée par d'autres indices acoustiques [WC83, BIM82]

> "These results suggest that although the invariant properties residing in the gross shape of the onset spectrum may serve as a classificatory framework for the phonetic dimensions of natural language, they may *not provide the primary perceptual attributes* for place of articulation in ongoing speech processing." [BIM82, page 43]

Changement de pente spectrale La métrique introduite par A. Lahiri en 1984 a montré une bonne capacité à séparer les labiales des dentales [LGB84]. Dans un article récent [DL96], M. F. Dorman et P. C. Loizou confirment que, dans 95% des cas, elle ne se trompe pas. Néanmoins, à l'aide de monstres conflictuels, ils découvrent que ce changement de pente spectrale peut être facilement masqué de point de vue perceptif :

> "Listeners' identification of [conflictiong] stimuli was determined principally by the information in the formant transitions. This outcome provide no support for the view that the relative spectral change is a significant perceptual cue to stop consonant place of articulation." [DL96, page 3825]

Équations du locus Les équation du locus réunissent dans une théorie unitaire les très variables (selon le contexte vocalique) transitions formantiques. Les caractéristiques de ces équations ont montré une stabilité remarquable (différents locuteurs, différentes langues) [SMM91, SHA93]. En 1995, A. Eek et E. Meisier montrent [EM95] que des syllabes synthétiques ayant un locus modifié (par exemple, [p] avec transitions beaucoup plus amples que celles prédites) sont perçues comme celles ayant un locus "normal" :

> "F2 transitions did not provide the primary context independent cues for place of articulation. Locus equation showed relational invariance in the production space but they have not the same relevant role in perception." [EM95, page 20]

"Running spectra" Comme nous avons vu, le codage de toute la surface du spectrogramme est la représentation la plus performante [NZ91]. La raison est sa capacité à capter de manière intégrée et précoce les caractéristiques statiques et dynamiques du début de la syllabe. Cependant, comme nous l'avons montré [NB97], ce codage ne rend pas compte des réponses des auditeurs aux monstres conflictuels. La raison est, selon nos résultats, une sensibilité grande au phasage de N-seg et V-seg.

Notre **conclusion** peut être donc formulée simplement : nous ne sommes pas capables de prédire de manière suffisamment précise la perception d'une syllabe donnée. Bien sûr, il n'est pas nécessaire

[24]Les scores vont de 70% à plus de 95%. Ses taux sont plus grands quand on accepte des prototypes variables avec le contexte et quand on prend en compte plus de trames vers la voyelle.

de prédire la perception des syllabes affectés de perturbation non-naturelles pour reconnaître les syllabes naturelles. Cependant, pouvoir le faire promet une meilleure reconnaissance des syllabes affectés par des perturbations naturelles (bruit de fond, débit variable, etc).

5.8 Synthèse de l'analyse acoustique des occlusives

Dans ce chapitre renfermant la partie la plus importante de notre travail, nous avons analysé les consonnes occlusives afin de trouver les voies pour l'amélioration de leur reconnaissance par les systèmes automatiques.

Nous avons établie une typologie des occlusives sourdes du français :

- /k/ est identifié par un bruit et des transitions très variables selon le contexte mais gardant une certaine continuité spectrale entre le N-seg et le V-seg.
- /t/ présente la cible articulatoire la plus forte dans le sens où son N-seg et son V-seg sont nettement plus résistant au contexte vocalique que ceux de /k/. Acoustiquement, /t/ présente un N-seg avec toujours plus de hautes fréquences que celui du /k/.
- /p/ est identifié par un bruit court et faible. Acoustiquement, le bruit de /p/ présente toujours une proportion moindre des hautes fréquences par rapport au /k/.

Ce tableau global est en accord avec la proposition récente par Eek et Meisier [EM95] d'un "espace perceptif" (voir figure 5.56) pour les consonnes. Cet espace sera basé sur la fréquence du F2 (pour identifier la voyelle) et sur une fréquence associé au bruit de relâchement (disons, par exemple, le centré de gravité d'une masse spectrale). Comme d'autres travaux, nous avons montré que, pour

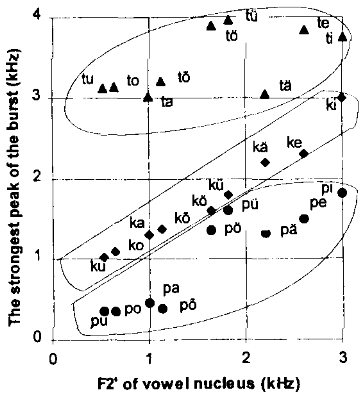


FIG. 5.56: Espace perceptif des occlusives proposé par Eek et Meisier [EM95].

aller au-delà d'une description sommaire, c'est à dire pour améliorer les performances déjà correctes des systèmes automatiques, la prise en compte du contexte vocalique s'impose. Nous avons montré également que, malgré la forte influence de ces contextes, un codage par prototype (donc a priori plus robuste) est possible.

Plusieurs autres travaux ont remarqué la difficulté de la reconnaissance des occlusives en contexte antérieur [i e ɛ]. Cependant, nous avons été les premiers, à notre connaissance, à formuler clairement le paradoxe du /ti/ versus /ki/. Nous avons été aussi les premiers à proposer une interprétation de ce paradoxe, interprétation basée sur l'étude des situations précises où le bruit de relâchement montre son pouvoir de discrimination. En deux mots, notre explication postule l'existance d'une continuité spectrale pour [ki] versus une absence de continuité pour [ti]. La difficulté de reconnaissance provient de la faible amplitude de ce contraste et de la largeur réduite de la bande de fréquence où il se manifeste. Notre interprétation est soutenue par la réussite des derniers tests perceptifs. Les auditeurs ont perçu une consonne différente quand nous avons modifié le spectre du bruit dans cette zone spécifique de fréquence.

Finalement, nous avons montré que les indices et les algorithmes que nous utilisons aujourd'hui n'arrivent pas à prédire dans une proportion acceptable les réponses des auditeurs aux monstres acoustiques. Cette incapacité de prédiction concerne plus l'aspect collaboration des indices que l'aspect compétition des indices.

Chapitre 6

Conclusions

6.1 Bilan général

Dans cette thèse nous avons étudié d'abord les voyelles et ensuite les consonnes occlusives du français. Pour récapituler, le travail a été organisé en quatre volets d'une étendue inégale : étude des voyelles, perception et production des consonnes, analyse acoustique des syllabes et confrontation avec les monstres acoustiques. Nous n'allons pas reprendre en détail les conclusions de notre travail. Elles sont disséminées dans des synthèses et bilans dans chaque chapitre. Nous allons rappeler ici juste le fil conducteur de nos recherches et les parties originales.

Nous avons cherché les situations difficiles pour la reconnaissance (pour les voyelles comme pour les consonnes) et nous avons proposé des solutions. Ces situations sont, pour les voyelles, le passage entre locuteurs de sexe différent et, pour les consonnes, la reconnaissance des occlusives précédant une voyelle antérieure non-arrondie. Les solutions que nous proposons soulignent l'importance pour l'analyse de la parole de l'"empreinte" laissée par les variations de la géométrie de la cavité buccale du conduit sur le signal vocal. Par exemple, nous montrons que le choix des paramètres liés à cette cavité facilite le passage entre les espaces vocaliques de locuteurs différents.

L'originalité de cette thèse vient de l'étude de l'ensemble consonne-voyelle. Nous avons mis en lumière les relations d'information mutuelle entre la consonne et la voyelle dans la syllabe. C'est cette relation qui est notre principal argument pour montrer le rôle de la cavité avant. Nous rappelons que le N-seg de [ki] est discriminant quand il est associé à un segment de [i] mais il est confondu avec [t] en isolé ou associé à un segment de [y]. Nous en déduisons que le système perceptif analyse le N-seg à la lumière d'une information qu'il extrait de la voyelle. Les auditeurs sont capables de focaliser leur attention sur une région précise (en temps et en fréquence). Nous avons démontré au cours de nos travaux que les systèmes automatiques peuvent être améliorés en utilisant une lecture "avisée" du N-seg. Ils gagnent à emprunter le même "faisceau de lecture" que celui qui permet la focalisation de la perception : les résonances de la cavité avant.

À coté des solutions proposées, nous accordons aussi de l'importance aux résultats dits "négatifs". Nous avons montré, par exemple, l'effet perturbateur des stratégies de locution sur la normalisation vocalique et l'incapacité de prédiction des réponses aux monstres acoustiques par les architectures courantes. Ces résultats négatifs sont une base de départ pour reformuler de manière adaptée les objectifs à poursuivre dans la recherche future.

Dans notre conception, un travail de thèse n'est pas seulement une collection des résultats mais aussi une occasion de reflexion, un regard critique porté sur le domaine du travail en général et sur son propre travail en particulier. Nous tenterons donc de répondre à la question *"Quelle est la portée des connaissances dégagées pendant ce travail pour notre domaine de recherche ?"*. Supposons qu'une implémentation parfaite d'un décodeur acoustico-phonétique selon les idées que nous avons développée soit faite. Même dans ce cas, l'amélioration brute du taux de reconnaissance sera marginale dans un corpus "normal".[1] Il est donc probable que, dans les produit commerciaux de dictée vocale, un gain plus rapide soit acquit par l'amélioration du modèle linguistique et non pas du modèle phonétique (c'est à dire par récupération a posteriori des erreurs de décodage acoustico-phonétique). Cependant, nous croyons avoir démontré la nécessité de la spécialisation du reconnaisseur au signal de parole et nous pensons avoir franchi un pas dans cette direction. Comme toujours, un changement radicale d'architecture du système de décodage peut être "douloureux". Dans un premier instant, il pourra même faire baisser les scores de reconnaissance. Mais, à long terme, il promet de dépasser les limites des systèmes actuels.

6.2 Perspectives

On ne peut pas considérer un bilan complet sans préciser ce qu'il reste à faire. Dans cette section se côtoient, d'habitude, les vrais ouvertures de nouvelles voies de recherche et les pistes laissées de coté en cours de route à cause de différentes contraintes pragmatiques. Faisons la convention d'appeler ces deuxièmes "perspectives à court terme" :

1. Une continuation nécessaire de notre travail concerne l'adaptation plus poussée de l'architecture afin d'arriver à un niveau acceptable de prédiction des monstres. L'architecture à fusion explicite, étant une sélection des paramètres du N-seg, accorde aussi une importance trop grande au N-seg au détriment du V-seg. L'amélioration viendra, à notre avis, d'un algorithme qui soit réellement dynamique. On peut considérer, par exemple, une évaluation cumulative des indices en partant du noyau vocalique vers le bruit de relâchement. Cet algorithme devrait non seulement calculer les probabilités d'une classe en explorant la transition mais, également, détecter les événements importants au long de celle-ci (début de voisement, relâchement de l'occlusion). Nous décrivons ici le comportement dont l'algorithme devrait être capable pour les occlusives sourdes devant trois voyelles extrêmes :

   Contexte /a/ - Effet de saturation L'algorithme devrait avoir départagé le candidat gagnant quand l'évaluation arrive au début du voisement. Un N-seg contradictoire avec le classement établi à cet instant ne devrait pas changer la décision finale. La partie la plus

[1] Nous parlons d'un corpus où les syllabes sont distribuées de manière équilibrée donc un corpus où les confusions /ti/ versus /ki/ ne seront jamais qu'une petite proportion des confusions possibles mais beaucoup moins difficiles. Cela même si /ti/ et /ki/ sont bien représentés dans le lexique (voir tableau 4.5)

importante de l'évaluation est dans V-seg.

Contexte /u/ - Report de décision Même si un candidat se détache au cours de l'évaluation du V-seg, l'algorithme devrait être capable de "changer d'avis" au cours de l'évaluation du N-seg. La partie la plus importante de l'évaluation est dans N-seg.

Contexte /i/ - Report de décision et guidage L'évaluation au cours du V-seg devrait fournir la position d'une bande de fréquence intéressante en plus d'une hiérarchie, probablement peu fiable, des candidats. L'évaluation du N-seg donnerait la décision finale.

Nous estimons qu'une telle architecture sera capable de prédire mieux les réponses des auditeurs. Entre autres, elle peut facilement être adaptée pour modéliser les réponses à des signaux tronqués. Par exemple, le biais en faveur du [p] pour V-seg isolés sera représenté comme une fonction décroissante à rajouter a priori à la probabilité du [p] dès que l'évaluation a dépassé l'instant VOT vers l'instant du relâchement. En revanche, il reste à trouver les "briques" pour combiner les trois comportements différents décrits plus haut afin d'implémenter cette architecture.

2. Une autre continuation importante concerne la validation des résultats de perception et d'identification sur un corpus plus varié. Il est nécessaire d'ajouter des voix féminines dans le corpus des items naturels et de rajouter plusieurs locuteurs des deux sexes dans le corpus d'items manipulés. Plus important encore à nos yeux, est l'étude de la structure de pondération perceptive N-seg/V-seg dans différents conditions : niveaux de bruit, canal de transmissions. Nous avons noté, de manière informelle, de légères différences, par exemple, entre l'écoute de certaines monstres au casque et l'écoute par haut-parleur.

3. Une autre continuation de notre travail est l'amélioration de la décision avant/arrière que nous préconisons en amont du décodage phonétique des voyelles. Pour mémoire, nos résultats ont montré une décision fiable à partir des valeurs formantiques "filtrées" par l'expert mais une performance moins bonne à partir directement du signal. Les voies pour acquérir cette amélioration sont liées à une étude plus approfondie des spectres vocaliques. Une méthode possible peut être un "codage populationel" pour passer des spectres aux formants.

Dans un cadre plus général, nous pensons que nos recherches ont soulevé des nouveaux problèmes auxquels il sera nécessaire de répondre à l'avenir :

1. Nous avons montré que les systèmes automatiques répondent différemment des auditeurs aux monstres acoustiques. Prenons comme hypothèse de travail[2] que les auditeurs humains ont la meilleure approche et donc la meilleure performance pour l'identification des occlusives. Ces différences dans les réponses aux monstres laissent supposer que les algorithmes automatiques pourrons gagner encore en performance. Pour ces raisons, nous pensons que des études plus poussés de la perception des occlusives se justifient.

[2] Dans la reconnaissance de la parole, le niveau de performance des auditeurs est considéré, parfois trop vite, comme la borne supérieure. Les systèmes actuels de dictée vocale demandent environ une demi-heure d'adaptation ce qui est jugé trop long. Cependant, nous serions tous contents si l'on pourrait apprendre une nouvelle langue, par exemple, ou quelque mille mots, après seulement "une demi-heure d'adaptation"...

2. Une direction nouvelle de recherche est la normalisation consonantique. Nous pensons que la liaison entre la voyelle et la consonne par les résonances de la cavité avant est la meilleure méthode, sinon la seule, pour y arriver. Concrètement, nous avons proposé un tableau de BF ("burst frequency") selon la voyelle. Cependant, il est raisonnable de supposer que plusieurs facteurs contribuent de manière orthogonale pour établir la position du BF. Les étude de normalisation devrons avoir pour but de séparer les deux facteurs les plus importants :
   (a) L'influence de la voyelle. Il faudrait chercher un algorithme capable de prédire le BF à partir de la voyelle, même à partir d'une seule réalisation.
   (b) L'influence du locuteur. La prédiction à partir de la voyelle pourra être raffinée à partir des connaissances synthétiques sur le locuteur (paramètres anatomiques, manière d'articulation).
3. Une autre direction nouvelle de recherche est la modélisation des stratégies d'articulation. Presque rien n'a été entreprit dans cette direction encore à notre connaissance.[3] Pour l'instant, la seule méthode dont on peut envisager l'utilisation est l'apprentissage à partir des données, méthode coûteuse en corpus de parole.
4. Nous avons argumenté en faveur de la syllabe comme unité fondamentale de la parole. Mais une unité, pour remplir sa fonction, doit posséder la qualité de détectabilité . Nous avons travaillé supposant une bonne détection sans vraiment la tester. À plusieurs reprises, nous avons observé qu'une bonne détection est critique (pour l'instant de relâchement des occlusives dans un corpus où des faux relâchements soient présents, pour les fréquences des formants dans un corpus où il n'y a pas que des voyelles orales). Dans ce contexte, l'étude de la segmentation et de la macro-segmentation (détection des segments et frontières) nous semble revêtir une nouvelle importance. La réponse à la question *"Est-ce que cette syllabe porte une occlusive ou non ?"* peut être aussi difficile que la question plus classique *"Quelle est cette occlusive ?"*. Nous avons décrit un modèle phonétique des occlusives sourdes mais ce modèle ne peut pas être complet avant que des syllabes V ne soient incluses à coté des syllabes CV dans notre corpus.

[3] Cependant, un travail de thèse a été démarré à l'I.C.P. sous la responsabilité de P. Perrier

Bibliographie

[ABBS85] Christian Abry, Christian Benoît, Louis-Jean Boë, and Rudolph Sock. Un choix d'évènements pour l'organisation temporelle du signal de parole. *Actes des XIVèmes Journées d'Etude sur la Parole*, pages 133–137, 1985.

[Abd87] Hervé Abdi. *Introduction au traitement statistique des données expérimentales*. Presses Universitaires de Grenoble, BP 47 X 38040 Grenoble France, 1987.

[ACMT78] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America*, 63 :1535–1555, 1978.

[Adj97] Ali Adjoudani. *Reconnaissance automatique de la parole audiovisuelle. Stratégies d'intégration et réalisation du LIPTRACK labiomètre en temps réel.* PhD thesis, Institut National Polytechnique, Grenoble - France, 1997. sous la direction de Christian Benoit.

[Apo97] Lian Apostol. Étude des caractéristiques individuelles par modélisation du processus de production de parole. Rapport d'activité à l'Institut de la Communication Parlée, 1997.

[Bai93] Gérard Bailly. Resonances as possible representation of speech in the auditory-to-articulatory transform. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 1511–1514, 1993.

[Bai95] Gérard Bailly. Caracterisation of formant trajectories by tracking vocal tract resonances. In Christel Sorin, Joseph Mariani, Henri Méloni, and Jean Schoentgen, editors, *Levels in speech communication : relations and interactions*, pages 91–102. Elsevier, Amsterdam, 1995.

[Bar84] William J. Barry. Place-of-articulation information in the closure voicing of plosives. *Journal of the Acoustical Society of America*, 76 :1245–1247, 1984.

[BBB+96] Denis Beautemps, Pierre Badin, Gérard Bailly, Arturo Galvàn, and Rafaël Laboissière. Evaluation of an articulatory-acoustic model based on a reference subject. In *ETRW on Speech Production : from Control Strategies to acoustics*, pages 45–48, Autrans - France, 1996.

[BBPA88] Pierre Badin, Louis-Jean Boe, Pascal Perrier, and Christian Abry. Vocalic nomograms : Acoustic considerations upon formant convergence. *Bulletin du Laboratoire de la Commmunication Parlée*, 2 :65–94, 1988.

[BBPA90] Pierre Badin, Louis-Jean Boë, Pascal Perrier, and Christian Abry. Acoustic considerations upon formant convergence. *Journal of the Acoustical Society of America*, 87(3) :1290–1300, 1990.

[BBW92] Gérard Bailly, Thierry Barbe, and Haïdong Wang. Automatic labelling of large prosodic databases : tools, methodology and links with a text-to-speech system. In G. Bailly and C. Benoît, editors, *Talking Machines : Theories, Models and Designs*, pages 323–333. Elsevier B.V., 1992.

[BCDL92] Anne Bonneau, Sylvie Coste, Linda Djezzar, and Yves Laprie. Two level acoustic cues for consistent stop identification. In *Proceedings of the International Conference on Speech and Language Processing*, pages 511–514, Banff, Alberta, Canada, 1992.

[BCML95] Anne Bonneau, Sylvie Coste-Marquis, and Yves Laprie. Strong cues identifying well-realized phonetic features. In *Proceedings of the International Congress of Phonetic Sciences*, pages 144–147, Stockholm - Sweden, 1995.

[BDL93] Anne Bonneau, Linda Djezzar, and Yves Laprie. Perception of French stop bursts. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 693–696, Berlin, Germany, 1993.

[BDL96] Anne Bonneau, Linda Djezzar, and Yves Laprie. Perception of place of articulation of French stop bursts. *Journal of Acoustical Society of America*, 100(1) :555–564, 1996.

[Bea93] D. Beautemps. *Récupération des gestes de la parole à partir de trajectoires formantiques : Identification de cibles vocaliques non-atteintes et modèles pour les profils sagittaux des consonnes fricatives*. PhD thesis, Institut National Polytechnique, Grenoble – France, 1993. sous la direction de Jean-Luc Schwartz.

[Ber91] Marie-Odile Berger. *Les contours actifs : Modélisation, comportement et convergence*. PhD thesis, Institut National Polytechnique de Lorraine, Nancy, France, 1991.

[BF78] R. A. W. Bladon and Gunnar Fant. A two-formant model and the cardinal vowels. Technical Report 1, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1978.

[BGB+95] Pierre Badin, Bernard Gabioud, Denis Beautemps, Tahar Lallouache, Gérard Bailly, Shinji Maeda, Jean-Pierre Zerling, and Gilbert Brock. Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model. In *International Congress on Acoustics*, pages 349–352, Trondheim - Norway, 1995.

[BIM82] Sheila. E. Blumstein, Ellyn Isaacs, and John Mertus. The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 72 :43–50, 1982.

[BL92] Maria-Gabriella Di Benedetto and Jean-Sylvain Liénard. Extrinsic normalization of vowel formant values based on cardinal vowels mapping. In *Proceedings of the International Conference on Speech and Language Processing*, pages 579–582, Banff - Alberta - Canada, 1992.

[BLMA92] Christian Benoît, Tahar Lallouache, Tayeb Mohamadi, and Christian Abry. A set of French visemes for visual speech synthesis. In G. Bailly and C. Benoît, editors, *Talking Machines : Theories, Models and Designs*, pages 485–501. Elsevier B.V., 1992.

[Blu86] S. E. Blumstein. On acoustic invariance in speech. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*, chapter 9, pages 178–193. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1986.

[Blu89] S. E. Blumstein. Theoretical implications of the quantal nature of speech : a commentary. *Journal of Phonetics*, 17(1/2) :55–61, 1989.

[Bon96] Anne Bonneau. Identification des voyelles à partir du bruit des occlusives. In *Journées d'Etudes sur la Parole*, pages 23–26, Avignon–France, 1996.

[Bon97] Anne Bonneau. Relevant spectral information for the identification of vowel features from bursts. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2139–2142, Rhodes, Greece, 1997.

[Bou95] Hervé Bourlard. Towards increasing speech recognition error rates. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 883–894, Madrid, Spain, 1995.

[Bou96] Hervé Bourlard. Reconnaissance automatique de la parole : modélisation ou description ? In *Journées d'Etudes sur la Parole*, pages 263–272, Avignon–France, 1996.

[BPB92] Louis-Jean Boë, Pascal Perrier, and Gérard Bailly. The geometric vocal tract variables controlled for vowel production : Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1) :27–38, 1992.

[BS79] Sheila. E. Blumstein and Kenneth N. Stevens. Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, 66 :1001–1017, 1979.

[BS80] S. E. Blumstein and Kenneth N. Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environment. *Journal of the Acoustical Society of America*, 67 :648–662, 1980.

[BSV94] Louis-Jean Boë, Jean-Luc Schwartz, and Nathalie Vallée. The prediction of vowel systems : perceptual contrast and stability. In Eric Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 185–214. John Wiley and Sons, Chichester, 1994.

[Cal89] Calliope. *La parole et son traitement automatique*. Masson, Paris, France, 1989.

[CFdSLT66] L. A. Chistovich, Gunnar Fant, A. de Serpa-Leitão, and P. Tjernlund. Mimicking and perception of synthetic vowels, part i. Technical Report 2, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1966.

[CFG75] Rolf Carlson, Gunnar Fant, and Bjorn Granström. Two-formant models, pitch and vowel perception. In Gunnar Fant and M. A. Tatham, editors, *Auditory Analysis and Perception of Speech*, pages 55–82. Academic Press, Inc., London, UK, 1975.

[CG84] Suzanne Cotton and François Grosjean. The gating paradigm : a comparison of successive and individual presentation formats. *Perception and Psychophysics*, 35(1) :41–48, 1984.

[CGF70] Rolf Carlson, Björn Granström, and Gunnar Fant. Some studies concerning perception of isolated vowels. Technical Report 2-3, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1970.

[Cip92] Barry Cipra. You can't hear the shape of a drum. *Science*, 255(3) :1642–1643, 1992.

[CM94] Sylvie Coste-Marquis. *Utilisation des techniques du raisonnement hypothétique pour la reconnaissance automatique de la parole*. PhD thesis, Université Henri Poincaré, Nancy, France, 1994. sous la direction de Jean Paul Haton.

[Cro88] A. S. Crowe. Generalised centroids, a new perspective on peak picking and formant estimation. In *Proceedings of 7th FASE Symposium*, pages 683–689, Edinburgh–Scotland, 1988.

[DB88] Grzegorz Dogil and Gunter Braun. *The PIVOT model of speech parsing*. Der Österreichischen Akademie des Wissenschaften, Wien – Österreich, 1988.

[Del86] Bertrand Delgutte. Analysis of French stops consonants using a model of the peripheral auditory system. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*, chapter 8, pages 163–177. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1986.

[DGBP91] A. Djeradi, B. Guérin, P. Badin, and P. Perrier. Measurement of the acoustic transfer function of the vocal tract : a fast and accurate method. *Journal of Phonetics*, 19 :387–395, 1991.

[Dje95a] Linda Djezzar. *Contribution à l'étude acoustico-perceptive des occlusives du français*. PhD thesis, Université Henri Poincaré, Nancy, France, 1995. sous la direction de Jean Paul Haton.

[Dje95b] Linda Djezzar. Effect of prior knowledge of the vowel on the perception of French stop bursts. In *Proceedings of the International Congress of Phonetic Sciences*, pages 262–265, Stockholm - Sweden, 1995.

[Dje95c] Linda Djezzar. Some new considerations about the spectral form of French stop bursts. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1389–1392, Madrid, Spain, 1995.

[DL96] Michael F. Dorman and Philipos C. Loizou. Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America*, 100(6) :3825–3830, 1996.

[DLC55] P. C. Delattre, Alvin M. Liberman, and F.S. Cooper. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27 :769–773, 1955.

[DP97] Linda Djezzar and Nicolas Pican. Phonetic knowledge embedded in a contexte sensitive MLP for French speaker-independent speech recognition. *Speech Communication*, 21 :155–167, 1997.

[DPRV95] C. Delogu, A. Paolini, P. Ridolfi, and K. Vagges. Intelligibility of speech produced by text-to-speech systems in good and telephonic conditions. *Acta Acoustica*, 3 :89–96, 1995.

[DR80] M. F. Dorman and L. J. Raphael. Distribution of acoutic cues for stop consonant place of articulation in VCV syllables. *Journal of the Acoustical Society of America*, 67(4) :1333–1335, 1980.

[DR97] Anastasios Delopoulos and Maria Rangoussi. The fractal behaviour of unvoived plosives : A means for classification. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 457–460, Rhodes - Greece, 1997.

[DRS96] Li Deng, Gordon Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. In *ETRW on Speech Production : from Control Strategies to acoustics*, pages 69–80, Autrans - France, 1996.

[DSKR77] Michael F. Dorman, M. Studdert-Kennedy, and L. J. Raphael. Stop-consonant recognition : release bursts and formant transitions as functionally equivalent, context-depending cues. *Perception and Psychophysics*, 22 :109–122, 1977.

[Dup89] Emmanuel Dupoux. Identification des mots parlés. Détection de phonèmes et unité prélexicale. Thèse de troisième cycle, Ecole des hautes études en sciences sociales, Paris, France, 1989. Sous la direction de J. Mehler.

[DYM89] G. Duncan, B. Yegnanarayana, and Hema A. Murthy. A non parametric method of formant estimation using group delay spectra. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 572–575, 1989.

[Edw81] Thomas J. Edwards. Multiple feature analysis of intervocalic English plosives. *Journal of the Acoustical Society of America*, 69 :535–547, 1981.

[EG96] H. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 346–348, 1996.

[ELM+81] Louis L. Elliot, Cheryl Longinotti, Dianne Meyer, Israel Raz, and Karen Zucker. Developmental differences in identifying and discriminating CV syllables. *Journal of the Acoustical Society of America*, 70(3) :669–677, 1981.

[EM95] A. Eek and E. Meister. The perception of stop consonants : Locus equation and spectral integration. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, pages 18–21, Stockholm - Sweden, 1995.

[Fan60] Gunnar Fant. *Acoustic theory of speech production*. The Hague : Mouton, 1960.

[Fan73] Gunnar Fant. Stops in CV-syllables. In Gunnar Fant, editor, *Speech Sound and Features*, chapter 7. MIT Press, Cambridge, MA - USA, 1973.

[Fan92] Gunnar Fant. Vocal tract area functions of Swedish vowels and a new three-parameter model. In *Proceedings of the International Conference on Speech and Language Processing*, volume 1, pages 807–810, Edmonton - Alberta, 1992.

[FD80] Sandra Ferrari-Disner. Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, 67 :253–261, 1980.

[FJ72] Eli Ficher-Jørgensen. Tape-cutting experiments with Danish stop consonants in initial position. Technical Report 6, Annual Repport, Institut of Phonetics, University of Copenhagen, Copenhagen, Denmark, 1972.

[Fow86] C. A. Fowler. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1) :3–28, 1986.

[Fow90] Carol A. Fowler. Listener-talker attunements in speech. *Haskins Laboratories Staturs Report on Speech research*, 1 :11–129, 1990.

[Fow96] Carol A. Fowler. Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3) :1730–1741, 1996.

[FR63] Gunnar Fant and A. Risberg. Auditory matching of vowels with two formant synthetic sound. Technical Report 4, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1963.

[Fur86] Sadaoki Furui. On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America*, 80 :1016–1025, 1986.

[FWMD88] Karen Forrest, Gary Weismer, Paul Milenkovic, and Ronald N. Dougall. Statistical analysis of word-initial voiceless obstruents : Preliminary data. *Journal of the Acoustical Society of America*, 84 :115–123, 1988.

[Gar78] W. R. Garner. Aspects of a stimulus : Features, dimensions and configurations. In E. Rosch and B. B. Lloyd, editors, *Cognition and Categorization*, pages 99–133. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1978.

[Ger68] L. H. Gerstman. Classification of self-normalised vowels. *IEEE Transactions on Audio and Electroacoustics*, AU-16 :78–80, 1968.

[GLL81] T. Gay, B. Lindblom, and J. Lubker. Production of bite-block vowels : Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America*, 69(3) :802–810, 1981.

[GR97] Arturo Galván-Rodríguez. *Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des plosives*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 1997. sous la direction de Rafael Laboissiere.

[GS97] Evandro B. Gouvêa and Richard M. Stern. Speaker normalisation through formant-based warping of the frequency scale. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1139–1142, Rhodes, Greece, 1997.

[Har76] Richard Harshman. *PARAFAC : Methods of three-way factor analysis and multidimensional scaling according to the principle of proportional profiles*. PhD thesis, UCLA, Los Angeles - California, 1976.

[HB89] H. Hermansky and D. J. Broad. The effective second formant f'2 and the vocal tract front-cavity. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 480–483, 1989.

[HP76] Richard Harshman and George Papçun. Vowel normalization by linear transformation for each speaker's acoustic space. *Journal of the Acoustical Society of America*, 59, 1976.

[HS95] V. Hazan and B. Shi. Individual variability in the perceptual weighting of cues to stop place and voicing contrasts. In *Proceedings of the International Congress of Phonetic Sciences*, pages 506–509, Stockholm - Sweden, 1995.

[Hun85] Melvin J. Hunt. A robust formant-based speech spectrum comparison measure. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1117–1120, 1985.

[JM91] Allard Jongman and James D. Miller. Method for the location of burst-onset spectra in the auditory-perceptual space : A study of place of articulation in voiceless stop consonants. *Journal of the Acoustical Society of America*, 89 :867–873, 1991.

[Jon79] D. S. Jones. Elementary information theory. In J. Crank, H. G. Martin, and D. M. Melluish, editors, *Oxford applied mathematics and computing science*. Clarendon Press, Oxford, 1979.

[Kla75] Dennis H. Klatt. Voice onset time, frication and aspiration in word-initial consonantal clusters. *Journal of Speech and Hearing Research*, 18 :686–706, 1975.

[Kla79] Dennis H. Klatt. Synthesis by rule of consonant-vowel syllables. *Working papers of the Speech Communication Group, MIT, Cambridge*, 3 :93–104, 1979.

[Kla82] Dennis H. Klatt. The klattalk text-to-speech conversion system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1589–1592, 1982.

[Kla87] Dennis H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3) :737–793, 1987.

[KO87] Hidefumi Kobatake and Shigeko Ohtani. Spectral transitions dynamics of voiceless stop consonants. *Journal of the Acoustical Society of America*, 81 :1146–1151, 1987.

[Kop85] Gary E. Kopec. Formant tracking using hidden Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1113–1116, 1985.

[Kop86] Gary E. Kopec. Formant tracking using hidden markov models and vector quantization. *IEEE Transactions on ASSP*, 34(4) :709–729, 1986.

[KP82] D. Kewley-Port. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, 72 :379–389, 1982.

[KP83] D. Kewley-Port. Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 73(1) :322–335, 1983.

[KPPSK83] Diane Kewley-Port, David B. Pisoni, and Michael Studdert-Kennedy. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 73(5) :1779–1793, 1983.

[Kru88] Diana Krull. *Acoustic properties as predictors of perceptual responses : a study of swedish voiced stops*. PhD thesis, Institute of Linguistics University of Stockholm, Stockholm, Sweden, 1988.

[Kru90] Diana Krull. Relating acoustic properties to perceptual responses : A study of Swedish voiced stops. *Journal of the Acoustical Society of America*, 88 :2557–2570, 1990.

[Kuh75] Gary M. Kuhn. On the front-cavity resonance and its possible role in speech perception. *Journal of the Acoustical Society of America*, 68 :578–585, 1975.

[LA64] L. Lisker and A. Abramson. A cross language study of voicing in initial stops : acoustical measurements. *Word*, 20 :384–422, 1964.

[Lap89] Yves Laprie. Formant tracking adapted to acoustic-phonetic decoding. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 669–672, Paris - France, 1989.

[Lap90] Yves Laprie. Optimum spectral peak track interpretation in terms of formants. In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 1261–1264, Kobe - Japan, 1990.

[LB82] Peter Ladefoged and Anthony Bladon. Attempts by human speakers to reproduce Fant's nomograms. *Speech Communication*, 1 :185–198, 1982.

[LB94] Yves Laprie and Marie-Odile Berger. A new paradigm for reliable automatic formant tracking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2 :201–204, 1994.

[LGB84] A. Lahiri, L. Gewirth, and S. E. Blumstein. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants : Evidence from a cross-language study. *Journal of the Acoustical Society of America*, 76 :391–404, 1984.

[Lié95] Jean-Sylvain Liénard. Perception et variabilité. In Henri Méloni, editor, *École Thématique : fondements et perspectives en traitement automatique de la parole*, pages 161–171. Université d'Avignon et des Pays de Vaucluse, Marseille-Luminy, 1995.

[Lin63] B. Lindblom. On vowel reduction. Technical Report 29, Speech Transmission Laboratory - Departement of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1963.

[Lin96] Bjorn Lindblom. Role of articulation in speech perception. *Journal of the Acoustical Society of America*, 99(3) :1683–1692, 1996.

[Lir90] Liris Interactive, 151, Rue Saint Honoré 75044 Paris. *Grand Robert Dictionnaire de la langue française*, 1990. CD ROM.

[LM85] A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revisited. *Cognition*, 21 :1–36, 1985.

[Lob71] B.M. Lobanov. Classification of russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, 49 :606–608, 1971.

[Mae79] Shinji Maeda. An articulatory model of the tongue based on a statistical analysis. *Journal of the Acoustical Society of America*, 65 :S22, 1979.

[Mae86] Shinji Maeda. Acoustique du relâchement de occlusives : une étude de simulation. In *Journées d'Etudes sur la Parole*, pages 317–320, Aix en Provence–France, 1986.

[Mae88] Shinji Maeda. Improved articulatory model. *Journal of the Acoustical Society of America*, 81(S1) :S146, 1988.

[Mas87] D. W. Massaro. *Speech Perception by Ear and Eye : A Paradigm for Psychological Inquiry*. Erlbaum, Hillsdale, NJ, USA, 1987.

[Mat91] I. G. Mattingly. In defense of the motor theory. *PERILUS XIV - Publication of the Department of Linguistics*, pages 167–172, 1991.

[McC74] S. McCandless. An algorithm for automatic formant extraction using linear prediction analysis. *IEEE Transaction on ASSP*, 22 :135–141, 1974.

[MD96] James L. Morgan and Katherine Demuth. *Signal to syntax : an overview*. Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1996.

[MG76] John D. Markel and Augustine H. Gray. *Linear prediction of speech*. Springer-Verlag, Berlin–Germany, 1976.

[Mor92] Andrew C. Morris. *An information-theoretical study of speech processing in the peripheral auditory system and cochlear nucleus*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 1992. sous la direction de Jean-Luc Schwartz et Pierre Escudier.

[MP85] Maurice Merleau-Ponty. *Phénoménologie de la perception*. Gallimard, Paris, France, 1985.

[MS95] S. Y. Manuel and K. N. Stevens. Formant transitions : Teasing apart consonant and vowel contribution. In *Proceedings of the International Congress of Phonetic Sciences*, pages 436–439, Stockholm - Sweden, 1995.

[Mun96] Paul Munteanu. *Extraction de connaissance dans les bases de données parole : Apport de l'apprentissage symbolique*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble - France, 1996. sous la direction de Jean François Serignat.

[MV97] Eduardo Sá Marta and Luis Vieira de Sá. Impact of "ascending sequence" cells on stop consonant perception. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2503–2506, Rhodes - Greece, 1997.

[NB97] Adrien Neagu and Gérard Bailly. Relative contributions of noise burst and vocalic transitions to the perceptual identification of stop consonants. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4, pages 2175–2178, Rhodes - Greece, 1997.

[NCH94] Mahesan Niranjan, Ingemar J. Cox, and Sunita Hingorani. Recursive tracking of formants in speech signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 205–208, 1994.

[Nea77] Terrance M. Nearey. *Phonetic feature systems for vowels*. PhD thesis, University of Connecticut, Storrs, Connecticut, 1977.

[Nea94] Adrian Neagu. Traitements distribués pour SIDOC - Application à la caractérisation des espaces vocaliques. Master's thesis, Institut National Polytechnique, Grenoble – France, 1994. DEA Signal Image Parole sous la direction de Jean François Serignat.

[Nea96] Terrance M. Nearey. Phonological contrast and articulatory perception. *Journal of the Acoustical Society of America*, 99(3) :1714–1717, 1996.

[Nor77] P. Nordström. Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics*, 5 :81–92, 1977.

[NZ91] Z. B. Nossair and S. A. Zahorian. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America*, 89 :2978–2991, 1991.

[OFFM91] D. J. Ostry, J. R. Flanagan, A. G. Feldman, and K. G. Munhall. Human jaw mouvement kinematics and control. In G. E. Stelmach and J. Requin, editors, *Tutorials in Motor Behavior II*. Holland Publishing Company, 1991.

[Oha95] Manjari Ohala. Acoustic study of VC transitions for Hindi stops. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, pages 22–25, Stockholm - Sweden, 1995.

[Oha96] John J. Ohala. Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3) :1718–1725, 1996.

[OHM96] Ralph N. Ohde, Katarina L. Haley, and Christine W. McMahon. A developmental study of vowel perception from brief synthetic consonant-vowel syllables. *Journal of the Acoustical Society of America*, 100(6) :3813–3824, 1996.

[OHVM95] Ralph N. Ohde, Katarina L. Haley, Houri K. Vorperian, and Christine W. McMahon. A developmental study of the perception of onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 97(6) :3800–3812, 1995.

[OS77] Ralph N. Ohde and Donald J. Sharf. Order effet of acoustic segments of VC and CV syllables on stop and vowel identification. *Journal of Speech and Hearing Research*, 20 :543–554, 1977.

[OS81] Ralph N. Ohde and Donald J. Sharf. Stop identification from vocalic transition plus vowel segments of CV and VC syllabes : A follow-up study. *Journal of the Acoustical Society of America*, 69(1) :297–300, 1981.

[OS83] Ralph N. Ohde and Kenneth N. Stevens. Effect of burst amplitude and the perception of stop consonant place of articulation. *Journal of the Acoustical Society of America*, 74(3) :706–714, 1983.

[PAP95] Pascal Perrier, Lian Apostol, and Yohan Payan. Evaluation of a vowel normalisation procedure based on speech production knowledge. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1925–1928, Madrid, Spain, 1995.

[Pay92] Y. Payan. Normalisation de l'espace de réalisation acoustique des locuteurs en vue de la synthèse de parole à l'aide d'un robot parlant. Rapport de DEA, I.N.P.G., Grenoble, 1992.

[PDO97] Madelaine Plauché, Cristina Delogu, and John J. Ohala. Asymmetries in consonant confusion. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 2187–2190, Rhodes - Greece, 1997.

[Pet86] Jean Petitot. *Le « morphological turn » de la phénoménologie. Chapitre I, II et III de morphogénèse du sens II.* Centre d'analyse et de Mathématique Sociales, EHESS-CNRS, 1986.

[PJ96] Xavier Pelorson and David Jorno. Fluid mechanics of plosive sounds. In *ETRW on Speech Production : from Control Strategies to acoustics*, pages 197–200, Autrans - France, 1996.

[PLT+94] Xavier Pelorson, M.T. Lallouache, S. Tourret, C. Bouffartigue, and Pierre Badin. Modeling the production of bilabial plosives : aerodynamical, geometrical and mechanical aspects. In *Proceedings of the International Conference on Speech and Language Processing*, volume 2, pages 599–602, 1994.

[PM95] Alexandros Potamianos and Petros Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 784–787, 1995.

[PM96] Alexandros Potamianos and Petros Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of the Acoustical Society of America*, 99 :3795–3806, 1996.

[PS78] Louis C. W. Pols and M. E. H. Schouten. Identification of deleted consonants. *Journal of the Acoustical Society of America*, 64(5) :1333–1337, 1978.

[PS81] L. C. W. Pols and M. E. H. Schouten. Identification of deleted plosives : The effect of adding noise or applying a time window. *Journal of the Acoustical Society of America*, 69(1) :301–303, 1981.

[PSB+96] Michel Piquemal, J. L. Schwartz, F. Berthommier, T. Lallouache, and P. Escudier. Détection et localisation auditive et visuelle d'explosions consonantiques dans des séquences VCV bruitées. In *Journées d'Etudes sur la Parole*, pages 355–358, Avignon–France, 1996.

[RD95] Maria Rangoussi and Anastasios Delopoulos. Recognition of unvoiced stops from their time-frequency representation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 792–795, 1995.

[Rep83] Bruno H. Repp. Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, 2 :341–361, 1983.

[Rig86] Gerhard Rigoll. A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1229–1232, 1986.

[Rig88] Gerhard Rigoll. Formant tracking with quasilinearization. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1 :307–310, 1988.

[RL89] Bruno H. Repp and Hwei-Bing Lin. Acoustic properties and perception of stop consonant release transient. *Journal of the Acoustical Society of America*, 85(1) :379–396, 1989.

[Rou92] Laurent Roussarie. Caractérisation des trajectoires formantiques par suivi de résonances. Master's thesis, Institut National Polytechnique de Grenoble, Grenoble - France, 1992.

[RR95] Jordi Robert-Ribes. *Modèles d'intégration audiovisuelle des signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble - France, 1995. sous la direction de Jean-Luc Schwartz.

[RSS96] R. C. Rose, J. Schroeter, and M. M. Sondhi. The potential role of speech production models in automatic speech recognition. *Journal of the Acoustical Society of America*, 99 :1699–1709, 1996.

[RW85] Bruno H. Repp and David R. Williams. Categorical trends in vowel imitation : preliminary observations from a replication experiment. *Speech Communication*, 4 :105–120, 1985.

[SB78] Kenneth N. Stevens and S. E. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64(5) :1358–1368, 1978.

[SB81] Kenneth N. Stevens and S. E. Blumstein. The search for invariant acoustic correlates of phonetic features. In P. Eimas and J. Miller, editors, *Perspectives on the study of speech*. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1981.

[SB86] Rudoplh Sock and Chritian Benoit. VOTs et VTT en français. In *Journées d'Etudes sur la Parole*, pages 307–310, Aix en Provence–France, 1986.

[SBDM97] Harvey M. Sussman, Nicola Bessell, Eileen Dalston, and Tivoli Majors. An investigation of stop place of articulation as a function of syllable position : A locus equations perspective. *Journal of the Acoustical Society of America*, 101(5) :2826–2838, 1997.

[Sch87] Jean-Luc Schwartz. *Représentations auditives de spectres vocaliques*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 1987.

[Ser84] W. Serniclaes. Fenêtre de prélèvement temporel des indices des occlusives. In *Journées d'Etudes sur la Parole*, pages 67–78, 1984.

[SHA93] Harvey M. Sussman, Kathryn A. Hoemeke, and Farhan S. Ahmed. A cross-liguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of the Acoustical Society of America*, 94(3) :1256–1268, 1993.

[Smi95] Roel Smits. *Detailed versus gross spectro-temporal cues for the perception of stop consonants*. PhD thesis, Institute of Perception Research (IPO), Eindhoven, Netherlands, 1995.

[Smi96] Roel Smits. Context-dependent relevance of burst and transitions for perceived place in stops : it's in production, not perception. In *Proceedings of the International Conference on Speech and Language Processing*, pages 2470–2473, Philadelphia - USA, 1996.

[SMM91] Harvey M. Sussman, Helen A. McCaffrey, and Sandra A. Matthews. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90(3) :1309–1325, 1991.

[Smo86] Paul Smolensky. Formal modeling of subsymbolic processes : an introduction to harmony theory. In N. E. Sharkey, editor, *Directions in the science of cognition*. Ellis Horwood, UK, 1986.

[Smo92] Paul Smolensky. IA connexionniste, IA symbolique et cerveau. In Daniel Andler, editor, *Introduction aux sciences cognitives*. Gallimard, France, 1992.

[Soq95a] Alain Soquet. A cooperative approach to formant extraction. In *Proceedings of the International Congress of Phonetic Sciences*, pages 448–451, Stockholm - Sweden, 1995.

[Soq95b] Alain Soquet. *Etude comparée de représentations acoustiques et articulatoires du signal de parole pour le décodage acoustico-phonétique*. PhD thesis, Université Libre de Bruxelles, Bruxelles – Belgique, 1995. sous la direction de Jean Schoentgen.

[SP83a] M. E. H. Schouten and L. C. W. Pols. Identification of intervocalic plosive consonants : the importance of plosive bursts versus vocalic transitions. In *Proceedings of the International Congress of Phonetic Sciences*, pages 464–468, Utrecht–Netherlands, 1983.

[SP83b] M. E. H. Schouten and L. C. W. Pols. Perception of plosive consonants - the relative contributions of bursts and vocalic transitions. In M.P.R. van den Broecke, V.J.J.P. van Heuven, and W. Zonneveld, editors, *Sound structures : Studies for Antonie Cohen*, pages 227–243. Foris, Dordrecht - The Netherlands, 1983.

[SPO95] Christophe Savariaux, Pascal Perrier, and Jean-Pierre Orliaguet. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube : A study of the control space in speech production. *Journal of the Acoustical Society of America*, 5 :2428–2442, 1995.

[Sta95] StatSoft, 2325 East 13th street, Tulsa OK 74104. *Statistica for Windows*, 1995. version 5.1.

[StB95] Roel Smits and Louis ten Bosch. A note on classification experiments in acoustic phonetics. *Journal of Phonetics*, 23 :477–485, 1995.

[StBC96a] Roel Smits, Louis ten Bosh, and René Collier. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. perception experiment. *Journal of the Acoustical Society of America*, 100(6) :3852–3864, 1996.

[StBC96b] Roel Smits, Louis ten Bosh, and René Collier. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. modeling and evaluation. *Journal of the Acoustical Society of America*, 100(6) :3865–3881, 1996.

[Ste72] Kenneth N. Stevens. The quantal nature of speech : Evidence from articulatory-acoustic data. In E. E. David Jr. and P. B. Denes, editors, *Human Communication : A unified view*, pages 51–66. McGraw-Hill, New York, 1972.

[Ste80] Kenneth N. Stevens. Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, 68 :836–868, 1980.

[Ste89] Kenneth N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17(1/2) :3–45, 1989.

[Ste93] Kenneth N. Stevens. Model for production and acoustic of stop consonant. *Speech Communication*, 13 :367–375, 1993.

[Suo85] Kari Suomi. The vowel dependence of gross spectral cues to place of articulation of stop consonants in CV syllables. *Journal of Phonetics*, 13 :267–285, 1985.

[Sus91] Harvey M. Sussman. The representation of stop consonants in three-dimensional acoustic space. *Phonetica*, 48(1) :18–31, 1991.

[TI96] M. I. Torres and P. Iparraguirre. Acoustic parameters for place of articulation identification and classification of Spanish unvoiced stops. *Speech Communication*, 18 :369–379, 1996.

[Val94] Nathalie Vallée. *Systèmes vocaliques : de la typologie auprédictions*. PhD thesis, Université Stendhal Grenoble III, 1994. sous la direction de Louis Jean Boë.

[vB93] Dick R. van Bergem. Acoustic vowel reduction as a function of sentence accent word stress, and word class. *Speech Communication*, 12(1) :1–23, 1993.

[vB95] Dick R. van Bergem. *Acoustic and lexical vowel reduction*. PhD thesis, University of Amsterdam, 1995.

[Waj79] M. Wajskop. Segmental duration of french intervocalic plosives. In B. Lindblom and S. Öhman, editors, *Frontiers of speech communication research*. Academic Press, London - UK, 1979.

[Wak77] H. Wakita. Normalisation of vowels by vocal tract length ans its application to vowel classification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-25 :183–192, 1977.

[Wak79] H. Wakita. Estimation of vocal tract shapes from acoustical analysis of the speech wave : the state of the art. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-27(3) :281–285, 1979.

[WBT90] Haïdong Wang, Gérard Bailly, and Denis Tuffelli. Automatic segmentation and alignment of continuous speech based on the temporal decomposition model. In *Proceedings of the International Conference on Speech and Language Processing*, volume 1, pages 457–460, Kobe - Japan, 1990.

[WC83] Amanda C. Walley and Thomas D. Carrell. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *Journal of the Acoustical Society America*, 73(3) :1011–1022, 1983.

[WSE96] Zong Liang Wu, Jean-Luc Schwartz, and Pierre Escudier. Physiologically plausible modules for the detection of articulatory-acoustic events. In W. A. Ainsworth, E. F. Evans, and C. M. Hackney, editors, *Advances in speech, hearing and language processing. Volume 3*, pages 479–495. JAI Press Inc., Greenwich, Connecticut - USA, 1996.

[WSR72] Harris Winitz, M. E. Scheib, and James A. Reeds. Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America*, 51(4) :1309–1317, 1972.

[Yeo95] Mohamed Yeou. An investigation of locus equations as a source of information for consonantal place. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1901–1904, Madrid - Spain, 1995.

[ZS90] Jérome Zeiliger and Jean-François Sérignat. Europec : a speech database recording system. Technical Report SAM-ICP/45, Institut de la Communication Parlée, Grenoble, France, 1990.

[ZT80] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, 68 :1523–1525, 1980.