Thèse de l'Institut National Polytechnique de Grenoble
Spécialité Sciences Cognitives
Avril 1997

présentée par

Arturo GALVÁN-RODRÍGUEZ

Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récuperation du lieu de constriction des plosives

A être soutenue le 4 avril devant le jury composé par :

Shinji MAEDA
Pascal PERRIER
Jean SCHOENTGEN
Rafael LABOISSIÈRE

Institut de la Communication Parlée
46 av. Félix Viallet, Grenoble CEDEX 1 France

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Qu'arrive-t-il lorsque un être humain répète ce qu'il entend ? Un signal en forme de perturbation de la pression de l'air est généré par le passage de l'air contenu dans les poumons à travers les cordes vocales tendues du locuteur. Cette perturbation passe ensuite par un tuyau de géométrie extrêmement variable (le conduit vocal) et est rayonnée dans l'atmosphère par la bouche. Captée par l'oreille de l'interlocuteur, elle traverse son conduit auditif et déclenche des signaux nerveux qui ensuite atteignent le cerveau. Ce cerveau les interprète et génère des commandes motrices telles qu'appliquées aux muscles du diaphragme, du larynx, de la langue, de la mâchoire et des lèvres génèrent un signal sonore servant (au moins) à mettre les membres d'une communauté linguistique d'accord : « Ils ont dit la même chose ».

Voilà la parole humaine. La nature humaine quant à elle est très ambitieuse et cherche à tout comprendre de ce phénomène.

Cette thèse est un humble essai de comprendre, dans la mesure des connaissances actuelles, certains aspects du phénomène décrit ci-dessus et plus précisément de donner réponse aux questions suivantes : un modèle de contrôle articulatoire du conduit vocal humain peut-il inférer des commandes telles qu'appliquées au modèle même produiront des effets acoustiques semblables à ceux qu'il « entend » ? Comment construire un tel modèle ?

La réponse à ces questions nécessite (1) une profonde connaissance des phénomènes de perception et de production de la parole et (2) des techniques très spécialisées et pluridisciplinaires :

- L'anatomie décrivant les articulateurs de la parole.
- La physique qui explique la génération et la transmission du son.
- Le traitement du signal, afin d'extraire des informations pertinentes du signal acoustique.

– La phonétique et la phonologie, qui décrivent les sons codés dans le signal et leurs interactions propres au langage.
– La cognition : Comment l'être humain perçoit-t-il et décode ces phonèmes, syllabes, mots, etc. ?

Cette thèse a été consacrée à l'élaboration d'un système de contrôle pour un modèle articulatoire capable de suivre (élégant synonyme de « répéter ») des trajectoires acoustiques présentées par un sujet quelconque.Bien évidemment, il s'agit d'une tâche d'envergure, étant donné qu'un tel système (contrôle + modèle articulatoire = « robot parlant ») est censé reproduire en détail les phénomènes de la communication parlée décrits dans le premier paragraphe de ce chapitre. D'une manière globale, les efforts de recherche liés à l'accomplissement de cette tâche ont été désignés ces dernières années par le terme d'« inversion ». La littérature récente dans le domaine de la production de la parole atteste d'une grande quantité de travaux réalisés sur ce thème (Abry, Badin & Scully 1994, Atal, Chang, Mathews & Tukey 1978, Bouabana 1995, Laboissière 1992, Lin & Fant 1989, Saerens, Soquet & Jospa 1991, Schoentgen & Ciocea 1995, Schroeter & Sondhi 1994, Sorokin 1992, entre autres).

Ces divers travaux ont adressé différents aspects du problème de l'inférence des commandes motrices à partir des consignes sonores à produire. Certes, ils nous montrent tous la faisabilité de l'approche, ce que nous conforte plutôt dans la certitude de nos choix. Cependant, plusieurs questions délicates sont laissées au passage sans réponse car, nous ne ferons que trop insister, la complexité de la tâche est énorme. Nous avons donc choisi une approche plus fondamentale au problème, sous le risque de ne pas avoir construit le « robot parlant » rêvé de tous, mais avec le bénéfice de la compréhension des problèmes de base. En effet, ayant posé comme but l'inversion des commandes articulatoires à partir d'une trajectoire formantique consonne-voyelle prononcée par un locuteur quelconque, nous nous sommes attaqués tout à tour au trois problèmes suivants :

1. *Amélioration du modèle articulatoire.* Le modèle articulatoire utilisé dans ce travail est originaire de celui développé par Maeda (1988), qui a fait ses preuves dans la modélisation articulatoire des séquences vocaliques (Boë, Perrier & Bailly 1992*a*, Laboissière 1992). Ayant remarqué certaines déficiences au niveau de la pointe de langue (partie essentielle pour une bonne description des consonnes alvéolaires) nous avons entrepris une étude de modélisation visant une meilleure description des mouvements apicaux. Pour ce faire, nous avons utilisé les mêmes données articulatoires qui ont donné origine au modèle de Maeda. Ce travail, ainsi que l'évaluation des résultats, est décrit dans le chapitre 2 de ce mémoire.Ce chapitre présente également les techniques qui ont été développées en vue de systématiser la génération des trajectoires que le modèle est censé reproduire.
2. *Normalisation de l'espace acoustique.* Toute méthode d'inversion acoustico-articulatoire à l'aide d'un modèle articulatoire doit se confronter aux problème des différences anatomiques entre le(s) locuteur(s) et le modèle en question. Ceci n'est pas forcément

vrai dans le cas où le modèle articulatoire est minutieusement obtenu pour un locuteur donné et l'on utilise les exemplaires de parole de ce même locuteur pour effectuer l'inversion. Or, ceci n'est pas le cas particulier traité dans cette thèse et ni même le cas générique des applications technologiques de l'inversion. Une procédure d'adaptation des dimensions longitudinales du modèle articulatoire a été donc développée et est décrite dans le chapitre 3. L'idée de base est d'associer deux paramètres de distorsion (reliés aux cavités pharyngale et orale) à chaque locuteur en fonction de leurs espaces vocaliques maximaux.Une évaluation de la méthode utilisée, dans un cadre indépendant de l'inversion consonne-voyelle y est incluse.

3. *Obtention des probabilités pour chaque lieu de constriction étant donnée une transition formantique.* Dans la dernière partie de notre travail, une nouvelle approche pour tester l'obtention du lieu de constriction des consonnes plosives à partir du son a été employée. Au contraire de la plupart des travaux suscités, nous avons opté pour une spécification minimale d'un modèle de contrôle qui imposerait des contraintes trop rigides à la forme des commandes articulatoires. Nous passons donc des systèmes « classiques » d'inversion où pour chaque séquence acoustique une et une seule trajectoire articulatoire est obtenue, vers une description probabiliste des toutes les possibilités articulatoires (dans notre cas, de tous les lieux de constrictions). Il nous semble que l'outil le plus adéquat pour mener à bien un tel travail soit l'Inférence Probabiliste Bayesienne, qui a la réputation de fournir des valeurs de probabilités *a posteriori* optimales et cohérentes à partir des connaissances *a priori* (pour une introduction convaincante à l'intérêt des techniques Bayesiennes, voir Jaynes 1997). Une présentation détaillé des principes et résultats est faite dans le chapitre 4

Chacun des travaux décrits ci-dessus peut être vu comme un développement indépendant des autres, mais leur interconnexion dans le cadre de notre but d'inversion est évidente. Ce travail de thèse a proportionné une excellente opportunité de développement des outils mathématiques appropriés pour le problème en question, en allant des techniques de régression et de projection géométriques, jusqu'à l'inférence probabiliste Bayesienne à l'aide des méthodes de Monte-Carlo. Pour ne pas encombrer l'exposition des idées dans ce mémoire, nous avons décrit certains détails des techniques mathématiques en appendice.

Cette thèse se veut une première dans l'exploration systématique et extensive des trois problèmes cités et, nous avons des raisons de le croire, les méthodes ici présentées pourront être utiles pour borner l'efficacité des méthodes d'inversion acoustico-articulatoire. Tel que le cycle de Carnot pour les machines thermiques ou le théorème de Shanon pour les canaux de transmission, nous pouvons envisager une extension logique de ce travail de thèse dans laquelle, les éléments de base étant mieux posés, les limites des méthodes d'inversion à l'aide d'un modèle articulatoire pourront être explicitées. Nous espérons, enfin, que le travail entrepris lors de cette thèse servira à mieux comprendre au moins une partie de ce complexe phénomène caractéristique des êtres humains que nous appelons la parole.

Chapitre 2

Le Modèle Articulatoire

Le but de notre travail étant d'inférer des patrons réalistes de commandes articulatoires à partir de séquences de signal de parole, il est clair que nous avons besoin d'un modèle articulatoire capable de produire ces séquences et cela de façon à ce que nous puissions en inférer les commandes. Ce chapitre est consacré au travail effectué dans le but de construire un tel modèle.

2.1 Introduction

Un modèle physique capable de produire un signal de parole à partir de commandes dites articulatoires est composé de plusieurs parties (Laboissière 1992) :

1. Modèle articulatoire capable de produire des coupes sagittales à partir de commandes articulatoires.
2. Modèle de passage de la coupe sagittale à la représentation en un ensemble de tuyaux équivalents au conduit vocal (fonction d'aire).
3. Modèle acoustique permettant de passer de la fonction d'aire aux formants associés.

Le modèle de passage de la coupe sagittale à la fonction d'aire à été l'objet de nombreuses études, dont la plupart furent inspirées par les résultats du travail de Heinz & Stevens (1965). Lors de ce travail, la fonction d'aire (c'est-à-dire, l'aire transversale à chaque section de la coupe sagittale) est calculée d'après l'équation 2.1, où d représente la distance entre la langue et les parois supérieure et arrière du conduit vocal (le palais et le pharynx) et les valeurs des coefficients α et β varient selon la région du conduit vocal.

$$A = \alpha\, d^{\beta} \tag{2.1}$$

Perrier, Boë & Sock (1992) utilisent cette équation avec $\beta = 1.5$ et α dépendant d'une part de la région du conduit vocal et d'une autre part de la valeur de la variable d. Le conduit vocal y a été divisé en sept sections avec deux valeurs de α associées à chacune de ces sections. Ces deux valeurs d'α reflètent les différences de géométrie du contour sagittal pour des faibles et grandes valeurs de d. Les aires prédites correspondent d'une manière satisfaisante à celles mesurées lors de leur étude[1].

Beautemps, Badin & Laboissière (1995) utilisent aussi l'équation 2.1 afin d'étudier le cas d'un locuteur produisant des consonnes fricatives et des voyelles. À la différence de Perrier et al. (1992), ils utilisent des valeurs de α qui varient de façon continue le long du conduit vocal :

> « This choice saves us from having to determine manually, and somehow arbitrarily, the limits between the different zones and from encountering discontinuities between regions. » (Beautemps et al. 1995)

Cette dernière méthode a été développée pour simuler des consonnes fricatives et la dépendance d'α en fonction de la région du conduit vocal a été obtenue à partir d'un corpus assez limité. Faute des études plus poussées concernant la proposition de Beautemps et al. (1995), c'est la méthode de Perrier et al. (1992) que nous utiliserons pour obtenir une estimation de la fonction d'aire à partir de la coupe sagittale.

Une partie importante de la fonction d'aire est celle correspondant aux lèvres. Abry & Boë (1986) ont fait une étude permettant d'estimer cette aire à partir des mesures d'ouverture et de protrusion. La méthode obtenue est en plus optimisée pour les voyelles du français. C'est cette méthode que nous utiliserons pour estimer la partie labiale de la fonction d'aire.

En ce qui concerne le modèle permettant de passer de la fonction d'aire aux formants, nous utiliserons la méthode consistant à obtenir un analogue électrique en ligne de transmission et à en calculer la fonction de transfert (cf. Badin & Fant (1984) pour les détails). Cette méthode prend en compte les pertes visco-termiques ainsi que les vibrations des parois. L'impédance de rayonnement est modélisée par un rallongement de la dernière section du conduit (Flanagan 1972). Les pôles de la fonction de transfert correspondent aux formants que le conduit vocal produit.

Bien que des interpolations polynômiales ou par des réseaux neuronaux aient été souvent utilisées (cf., par exemple, Bailly 1995*b*, Laboissière 1992) pour obtenir des approximations aux valeurs des formants directement à partir des paramètres articulatoires (ce qui évite d'avoir à calculer la fonction d'aire et ensuite la fonction de transfert correspondante), nous avons décidé d'éviter cette démarche. La raison de ce choix est double : d'une part, la procédure d'approximation analytique du passage commandes articulatoires $\rightarrow$ formants est utile pour l'étude des phénomènes d'apprentissage du contrôle moteur en production de la parole, mais introduit au passage des approximations dans le calcul des formants certes

[1] Ils ont testé la méthode sur leurs données et celles de Sanchez & Boë (1984).

tolérables, mais que nous voulons éviter dans notre travail. D'autre part, ces approximations analytiques sont construites pour un modèle articulatoire de dimensions fixes. Or la procédure utilisée pour normaliser les formants des locuteurs (voir chapitre 3) exige un calcul précis des résonances du conduit à partir de la fonction d'aire.

2.2 Choix du Modèle Articulatoire

Le choix du modèle articulatoire à utiliser représente un point crucial dans l'étude qui nous occupe. De nombreux modèles ont été créés dans le but de prédire les formes et caractéristiques acoustiques du conduit vocal associés aux phonèmes produits. Gabioud (1994) fait un résumé des modèles que l'on trouve dans la littérature et les classifie en *géométriques* (Öhman 1966, Öhman 1967, Coker & Fujimora 1966), *physiologiques* (Kiritani, Itioh & Fujimura 1975) et *statistiques* (Maeda 1979, Maeda 1988, Lindblom & Sundberg 1971).

Nous nous intéresserons notamment aux modèles dits « statistiques » et plus précisément à celui développé par Maeda (1988). Ce modèle articulatoire fut développé à partir de clichés cinéradiographiques d'un sujet féminin ayant prononcé dix phrases en français (Botherel, Simon, Wioland & Zerling 1986)[2]. Parmi les atouts de cette approche on trouve le fait que les commandes peuvent être interprétées comme des « commandes naturelles » jouant le rôle de la position de la mâchoire, corps et dos de la langue, hauteur du larynx, etc.

Majid (1986) a démontré que ce type de modèle (Maeda 1979) parvient à prédire des phénomènes de macro-sensibilité liés à la nature « quantique » de la parole (Stevens 1989, Perkell & Nelson 1985, Majid, Abry, Boë & Perrier 1987). Boë, Perrier & Bailly (1992*b*) ont utilisé avec succès le modele de Maeda (1988) afin de produire des voyelles du français.

Pour le cas des consonnes fricatives, le modèle s'est avéré limité et des modifications ont dû être mises en œuvre (Beautemps et al. 1995). D'autres limites ont aussi été remarquées par Gabioud (1994), notamment le manque de précision dans le calcul de la position de la pointe de la langue et de la hauteur du larynx.

Étant donné tous ses atouts, nous utiliserons le modèle de Maeda (1988) pour notre étude. Cela étant dit, nous sommes contraints d'y apporter quelques modifications afin de surmonter les limites que ce modèle présente.

[2]cf. Hassan (1988) pour une description plus approfondie des techniques cinéradiographiques.

2.3 Modifications au Modèle Articulatoire

La position de la pointe de la langue étant fort importante lors de la production de plosives apicales et rétroflexes (Sundberg & Lindblom 1990), nous avons décidé de remplacer le paramètre pointe de la langue du modèle original par d'autres paramètre permettant de positionner celle-ci de façon précise. Cette procédure a consisté en l'obtention (à la main) des coordonnées de la pointe de la langue sur les clichés cinéradiographiques afin d'observer la relation entre celle-ci et les autres paramètres du modèle. Nous appellerons ces coordonnées, ttp_x et ttp_y (selon les axes horizontaux et verticaux, respectivement). La figure 2.1 montre les différents cas que nous avons rencontré dans les clichés ainsi que les coordonnées de la pointe de la langue que nous avons choisies.

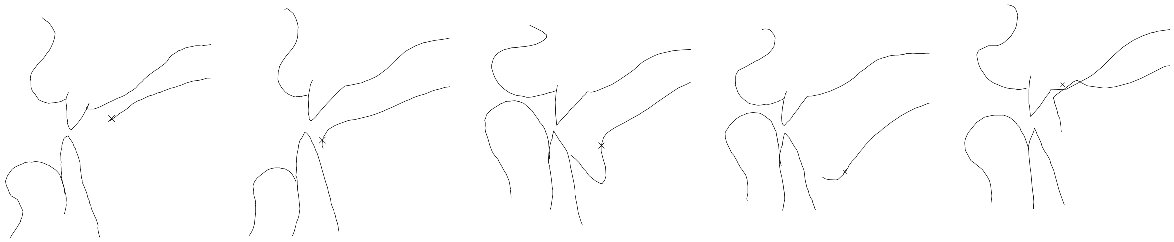

FIG. 2.1: Différents cas de clichés cinéradiographiques et la position de la pointe de la langue choisie pour chacun.

Nous avons ensuite éliminé l'influence des paramètres mâchoire, corps de la langue et dos de la langue (j, b, d respectivement) des points ttp_x, ttp_y en utilisant une technique de régression linéaire (Maeda 1979, Overall 1962).L'approche que nous suivons ici est de considérer les paramètres j, b et d du modèle original comme étant corrects pour la description des déformations linguales et de ajouter les nouveaux paramètres de la pointe à ceux-ci.

La figure 2.2 montre les points résultants de cette procédure. Puisque le nuage des points n'a pas une dimension intrinsèque unitaire, il est évident que deux paramètres sont nécessaires pour arriver à prédire avec précision la position de la pointe de la langue.

Étant donné que la position de la pointe de la langue est complètement décrite par deux mesures, nous avons choisi de relier les deux nouveaux paramètres t_x, t_y directement aux coordonnées ttp_x et ttp_y mesurées. Ce choix nous permettra de mieux visualiser les effets de ces paramètres sur la position finale de la pointe de la langue. L'équation 2.2 présente la façon d'obtenir cette position à partir du nouvel ensemble de paramètres articulatoires.[3]

[3]En toute rigueur, nous aurions du utiliser une matrice de covariance *pleine* pour les paramètres de la pointe de la langue, les zéros apparaissants dans la matrice de l'équation 2.2 étant remplacés par les valeurs de la covariance entre ttp_x et ttp_y. Nous avons opté pour ne pas prendre cette corrélation en considération parce que, d'une part, elle était faible et, d'autre part, nous gardons une interprétation directe des nouveaux paramètres de contrôle en termes de positionnement horizontal et vertical de la pointe de la langue dans la cavité orale.

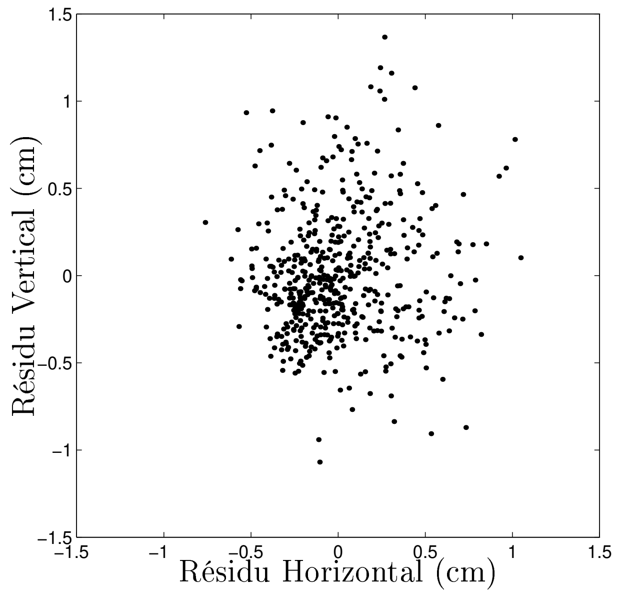

FIG. 2.2: Résidu de la position de la pointe de la langue.

$$\begin{bmatrix} ttp_x \\ ttp_y \end{bmatrix} = \begin{bmatrix} a_{tx} & 0 \\ 0 & a_{ty} \end{bmatrix} \times \begin{bmatrix} t_x \\ t_y \end{bmatrix} + A_{jbd} \times \begin{bmatrix} j \\ b \\ d \end{bmatrix} + B_t \tag{2.2}$$

Où : a_{tx} – variance du résidu en x.

a_{ty} – variance du résidu en y.

A_{jbd} – matrice correspondant à l'influence des paramètres j, b et d sur la position de la pointe de la langue.

B – moyenne de ttp_x, ttp_y.

Il est à remarquer que la position de la pointe de la langue est indépendante de la grille de référence définie par Maeda (1988). Les distances au long de cette grille se trouvant entre la pointe de la langue et les lèvres n'auront donc pas à être considérées lors de l'obtention de la coupe sagittale. Ces distances seront remplacées par les distance correspondantes jusqu'au « plancher » sous-lingual. Autrement dit, nous considérons que que la surface supérieure de langue s'« arrête » au point donné par les coordonnées t_x et t_y.

2.4 Obtention de la Coupe Sagittale

Nous présentons la façon d'obtenir la coupe sagittale à partir du nouvel ensemble de paramètres articulatoires. Nous commencerons par définir la ligne décrivant le palais et l'arrière du pharynx comme étant fixe. La position de cette ligne est obtenue en faisant les moyennes de toutes les positions obtenues des clichés cinéradiographiques.

Soit la grille montrée sur la figure 2.3. Nous définissons la position du contour de la langue traversant la ligne i de la grille comme une distance z_i au long de la ligne correspondante. Le vecteur Z contenant ces distances aura une dimension variable entre 22 et 25[4]. L'équation 2.3 montre ce fait, ainsi que la prise en compte de l'influence des paramètres t_x et t_y sur l'ensemble de la forme de la langue.

$$Z = \begin{bmatrix} z_{ini} \\ \vdots \\ z_{25} \end{bmatrix} = \begin{bmatrix} a_{ini,1} & \dots & a_{ini,5} \\ \vdots & \vdots & \vdots \\ a_{25,1} & \dots & a_{25,5} \end{bmatrix} \times \begin{bmatrix} j \\ b \\ d \\ t_x \\ t_y \end{bmatrix} + \begin{bmatrix} b_{ini} \\ \vdots \\ b_{25} \end{bmatrix} \tag{2.3}$$

Où : z_1 – distance au long de la ligne de la grille la plus proche des lèvres.

z_{25} – distance au long de la ligne de la grille la plus proche de la glotte.

La partie du contour correspondant au plancher sous-lingual à été obtenue d'un des clichés cinéradiographiques de la base de données originale. N'ayant pas trouvé suffisamment de clichés montrant ces données pour en inférer une relation statistiquement significative avec les paramètres du modèle, cette partie sera alors considérée non souple et bougera toujours d'une façon solidaire aux mouvements d'ouverture/fermeture de la mâchoire.

Afin de passer de la coupe sagittale à la fonction d'aire, nous représenterons la coupe sagittale comme un ensemble de rectangles de dimensions $d_{1,i}$, $d_{2,i}$. Pour ce faire, nous définirons le polygone D_i délimité par les points $a_1 \dots a_4$ (la partie droite de la figure 2.3, montre le polygone pour $i = 16$) correspondant d'une part au lignes i , $i+1$ de la grille de référence et d'autre part à l'approximation linéaire du contour de la langue et du palais (ou, le cas échéant, l'arrière du pharynx). La distance $d_{2,i}$ sera telle que le rectangle de dimensions $d_{1,i}$, $d_{2,i}$ ait la même aire que le polygone D_i correspondant, $d_{1,i}$ étant la distance mesurée entre les points $\bar{a}_1$ et $\bar{a}_2$ montrés sur la partie droite de la figure 2.3.

Nous utiliserons la distance $d_{2,i}$ de chacun des rectangles comme le paramètre d de l'équation 2.1 afin d'estimer l'aire transversale correspondante en utilisant la méthode de

[4]La dimension pour une configuration articulatoire donnée dépendra de la position de la pointe de la langue par rapport à la grille de référence (cf. équation 2.2).

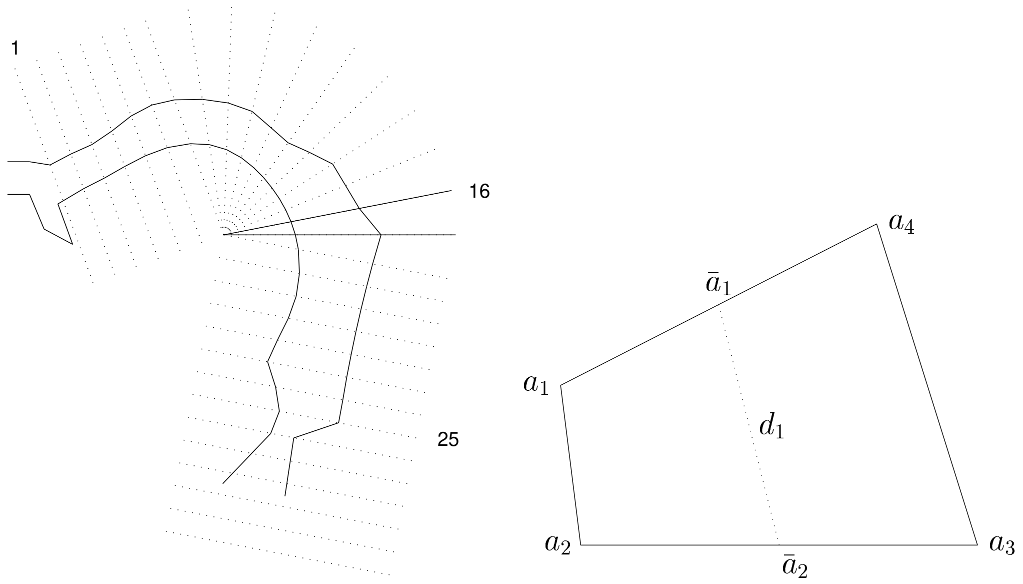

FIG. 2.3: Coupe sagittale avec grille de référence (gauche) et une section de cette coupe sagittale (droite).

Perrier et al. (1992). Cette méthode divise la coupe sagittale dont nous cherchons la fonction d'aire en six sections. La figure 2.4 montre les endroits où nous avons placé ces divisions.

Nous appellerons la cavité située entre la langue et les incisives « cavité avant »[5]. Nous avons défini cette cavité comme étant délimitée par les incisives, le palais et le plancher sous-lingual et par une ligne parallèle aux lignes de la grille de référence et passant par la pointe de la langue ttp_x, ttp_y.

La question naturelle qui se pose à ce stade de modélisation concerne l'effet acoustique de la cavité sous-lingual. En particulier, il est nécessaire d'obtenir des informations fiables sur la géométrie tri-dimmensionnelle de cette cavité pour une modélisation acoustique convenable. Sundberg & Lindblom (1990) ont estimé le volume de cette cavité en utilisant une méthode acoustique. Des estimations réalisées par ces méthodes ont l'avantage de ne prendre en compte que les volumes qui sont acoustiquement relevants (Fujimora & Lindqvist 1971). Bien que cette étude présente des résultats satisfaisants, il ne nous est pas possible d'obtenir une estimation de ce volume pour notre modèle. Ceci est dû au fait que notre modèle n'est capable de reculer la pointe de la langue que de 15mm derrière les incisives tandis que Sundberg & Lindblom utilisent des mesures allant jusqu'à 25mm.

N'ayant pas trouvé dans la littérature des données nous permettant de relier le volume de cette cavité à la distance dans le plan sagittal, nous avons utilisé la méthode de Perrier

[5]Bailly (1995*b*) a remarqué l'importance de cette cavité lors de l'inversion de séquences voyelle-consonne-voyelle.

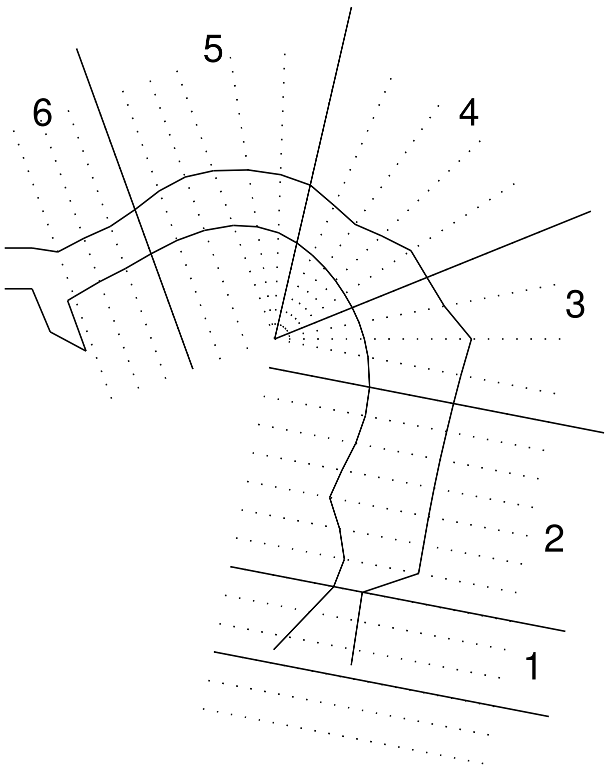

FIG. 2.4: Division de la coupe sagittale en six régions.

et al. (1992) pour estimer la partie de la fonction d'aire correspondant à cette cavité. Ce choix devra être vérifié ultérieurement (cf. chapitres suivants).

La figure 2.5 montre l'effet des nouvelles variables sur la coupe sagittale (la partie sous-linguale ayant été stylisée par une approximation en spline).

En ce qui concerne le paramètre du modèle correspondant à la hauteur du larynx, une variable l_h a été obtenue à partir des données cinéradiographiques. Cette variable (normalisée autour de sa valeur moyenne) sert à positionner le bas du larynx dans le plan vertical. La figure 2.6 illustre le mouvement du larynx lorsque l'on fait varier la valeur du paramètre correspondant entre -2 et 2 écarts-type.

Deux autres paramètres correspondant à la protrusion et à l'ouverture des lèvres, l_x, l_y respectivement, ont été obtenus à partir des données originales. La lèvre inférieure étant directement liée à la mâchoire, l'influence de celle-ci sur l'ouverture des lèvres a d'abord été éliminée.

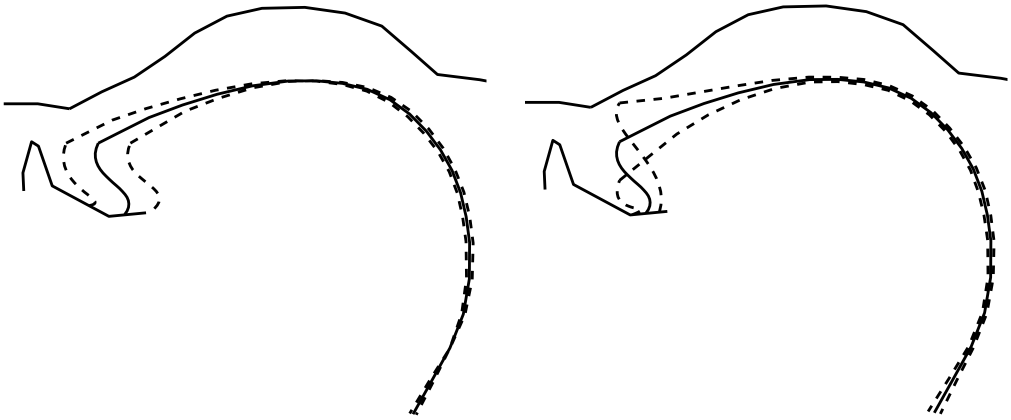

FIG. 2.5: Résultat sur la coupe sagittale lorsque l'on fait varier t_x, t_y entre ± 1.5.

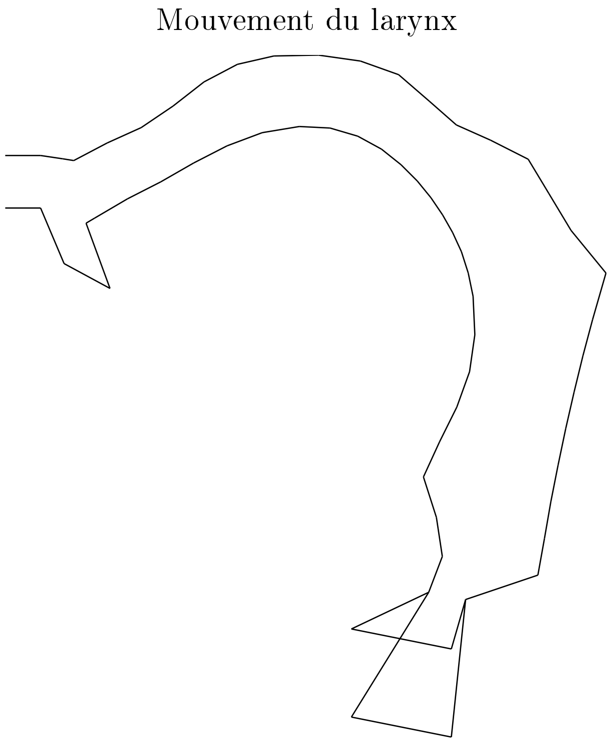

FIG. 2.6: Changement de la coupe sagittale lorsque l'on fait varier l_h entre ± 2.

2.5 Vérification et Exploitation du Modèle

Après la mise en œuvre du modèle, une première vérification a été faite. Celle-ci a consisté en l'application des configurations articulatoires ne produisant pas de constriction (aire minimale 0.1 cm^2) parmi les 519 ensembles de paramètres articulatoires correspondant à chacun des clichés dans la base de données originale. La figure 2.7 représente dans les plans F1-F2 et F2-F3 les résultats acoustiques de ces paramètres. Nous observons que l'espace maximal vocalique (Boë, Perrier, Guérin & Schwartz 1989) est couvert de façon satisfaisante.

Les cavités nasales n'ayant pas été prises en compte dans ce modèle, nous observons que la zone correspondant aux voyelles nasales (Maeda 1984, Feng, Abry & Guérin 1986) n'est pas couverte.

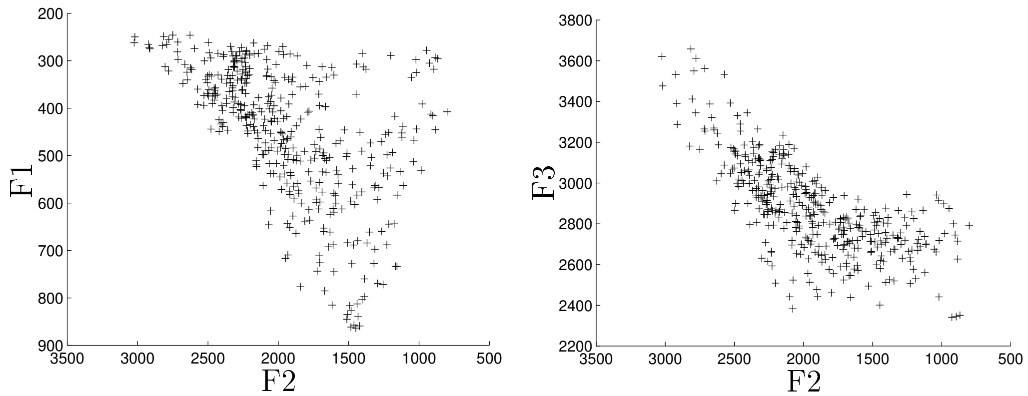

FIG. 2.7: Formants obtenus à partir des paramètres de la base de données.

2.5.1 Obtention d'occlusions

Notre objectif étant de trouver des commandes articulatoires telles que le robot parlant produise des séquences formantiques compatibles avec des séquences voyelle-consonne produites par un sujet quelconque, nous chercherons à faire en sorte que le modèle produise des occlusions d'une manière systématique.

L'équation 2.3 donne la forme de la langue à partir de commandes articulatoires données et l'équation 2.2 donne la position de la pointe de la langue à partir des mêmes commandes articulatoires. Soit c_i la distance entre l'origine de la $i^{ème}$ ligne de la grille de référence et le palais. Forcer un constriction à la $i^{ème}$ ligne de la grille de référence équivaut à contraindre la $i^{ème}$ ligne de l'équation 2.3 à être égale à c_i. Étant donné la façon d'obtenir la fonction d'aire (voir ci-dessus), nous sommes obligés de contraindre les lignes i , $i+1^{èmes}$ à être égales à c_i , c_{i+1}.

Une constriction à la $i^{ème}$ ligne de la grille peut alors être représentée par un ensemble de deux équations (cf. ci-dessous).

$$\begin{bmatrix} a_{i,1} & \dots & a_{i,5} \\ a_{i+1,1} & \dots & a_{i+1,5} \end{bmatrix} \times \begin{bmatrix} j \\ b \\ d \\ t_x \\ t_y \end{bmatrix} + \begin{bmatrix} b_i \\ b_{i+1} \end{bmatrix} - \begin{bmatrix} c_i \\ c_{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.4}$$

Trouver une constriction à la $i^{ème}$ ligne de la grille à partir d'une configuration quelconque équivaut à trouver, à partir des valeurs des variables articulatoires données, un ensemble de valeurs satisfaisant l'équation 2.4 ci-dessus. Représentons la configuration de départ par un point dans l'espace $\mathbb{R}^5$ nous chercherons le point (toujours en $\mathbb{R}^5$) le plus proche de ce point de départ.

Nous utiliserons une approche générale pour un espace $\mathbb{R}^n$ contraint à un ensemble de m équations linéaires.

Soient :

- $A = [A'|A'']$ une matrice constante de taille $m \times n$, où A' est une matrice de taille $m \times m$ et A'' est une matrice de taille $m \times (n-m)$.
- $\mathcal{A} = [A']^{-1} \times A''$ avec a^i la $i^{ème}$ ligne de $\mathcal{A}$.
- B une matrice constante de taille $m \times 1$.
- $\mathcal{B} = [b_1, b_2 \dots b_m]^T = [A']^{-1} \times B$
- $X = [x_1, x_2 \dots x_n]^T$ une variable en $\mathbb{R}^n$.
- $P = [p_1, p_2 \dots p_n]^T$, $Q = [q_1, q_2 \dots q_n]^T$ deux points en $\mathbb{R}^n$.
- I la matrice identité de taille $m \times m$.

Le point Q le plus proche d'un point P donné et contraint à l'équation (multidimensionelle linéaire) $AX + B = 0$ est obtenu d'après les équations suivantes :

$$\begin{aligned} \begin{bmatrix} q_{m+1} \\ q_{m+2} \\ \vdots \\ q_n \end{bmatrix} &= \left\{ I + \sum_{j=1}^{m} [a^j]^T \times a^j \right\}^{-1} \times \left\{ \begin{bmatrix} p_{m+1} \\ p_{m+2} \\ \vdots \\ p_n \end{bmatrix} + \sum_{j=1}^{m} (p_j - b_j)[a^i]^T \right\} \\ \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix} &= \mathcal{A} \times \begin{bmatrix} q_{m+1} \\ q_{m+2} \\ \vdots \\ q_n \end{bmatrix} + \mathcal{B} \end{aligned} \tag{2.5}$$

Soit P (ci-dessus) une configuration articulatoire quelconque, nous pouvons maintenant trouver la configuration articulatoire Q faisant une constriction à la ligne i de la grille de référence.

Avec les équations ci-dessus et l'équation 2.2, nous pouvons contraindre la pointe de la langue à se positionner à un endroit bien précis. Si nous la positionnons à l'intersection de la $i^{ème}$ ligne de la grille et du palais, il est possible de réaliser des occlusives rétroflexes. La figure 2.8 montre le résultat d'une constriction à la deuxième ligne de la grille soit avec le dos, soit avec la pointe de la langue [6].

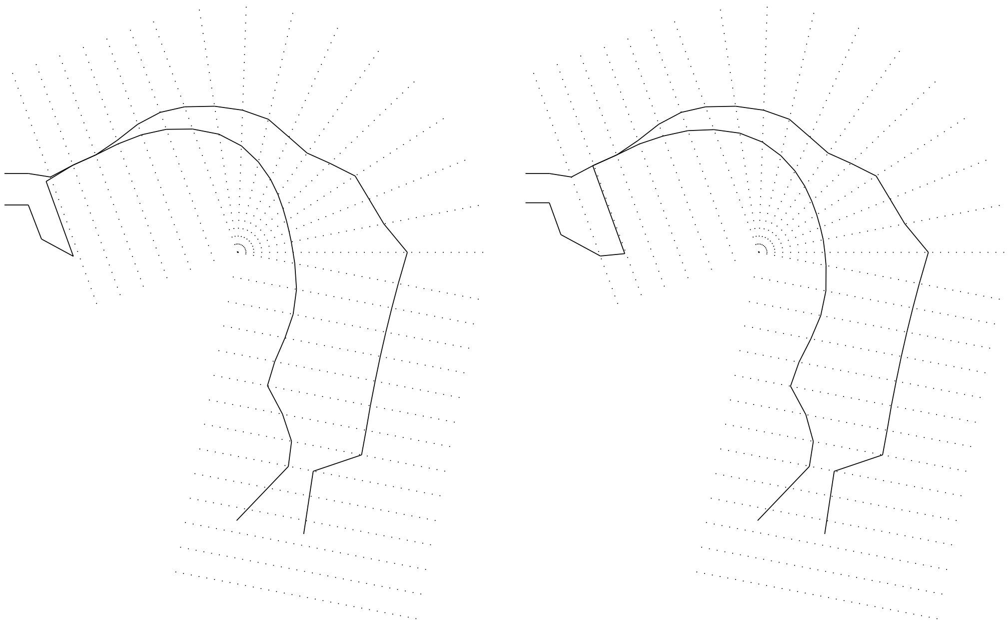

FIG. 2.8: Constriction à la deuxième ligne de la grille avec le dos (gauche) et avec la pointe de la langue (droite).

Nous avons à présent une façon systématique de produire des occlusions à des endroits bien spécifiques du conduit vocal à partir de configurations articulatoires quelconques. Il nous reste à trouver une façon (systématique aussi) de produire une constriction aux lèvres. Nous savons que l'aire transversale aux lèvres est fonction de la position de la mâchoire, ainsi que de l'ouverture et de la protrusion des lèvres (Abry & Boë 1986). Afin de faire une constriction labiale à partir d'une configuration articulatoire quelconque nous supposerons que le paramètre protrusion des lèvres ne changera pas ; nous n'aurons à spécifier que des variations sur les paramètres mâchoire et ouverture des lèvres. La participation de chacun de ces deux paramètres à la position des lèvres ayant été définies (cf. section précédente),

[6] Étant donné que l'équation 2.2 ne prend en compte que la position de la pointe de la lange, nous avons contraint l'intersection entre la langue et la $i+3^{ème}$ ligne de la grille à être séparée du palais. Ceci afin d'éviter des constrictions derrière la pointe de la langue.

nous forcerons l'aire aux lèvres à être nulle, tout en respectant la proportion de participation de chacun de ces deux paramètres.

Nous utiliserons dorénavant la variable X_c pour définir une constriction à un endroit donné. Cette variable prendra les valeurs suivantes :

- X_c= t3 : constriction à la troisième ligne de la grille effectuée avec la pointe de la langue (rétroflexe).
- X_c= t2 : constriction à la deuxième ligne de la grille effectuée avec la pointe de la langue (rétroflexe).
- X_c= t1 : constriction à la première ligne de la grille effectuée avec la pointe de la langue.
- X_c= ll : constriction aux lèvres.
- X_c= d2 : constriction à la deuxième ligne de la grille effectuée avec le dos de la langue.[7]
- X_c= d3 ... d9 : constriction à la troisième ... neuvième ligne de la grille effectuée avec le dos de la langue.

2.5.2 Vérification de la Séparabilité Acoustique

Notre objectif étant d'inférer des commandes articulatoires d'occlusives à partir de cibles spécifiées dans l'espace acoustique, nous devons nous assurer que les différents lieux d'articulation passibles d'être obtenus par le robot parlant se différencient acoustiquement entre eux.

Ce travail a déjà été entrepris lors d'une étude précédente (Laboissère & Galván 1995). À partir des configurations articulatoires de la base de données originale, nous avons produit une occlusion à chaque lieu de constriction possible. À chaqu'un de ces lieux, nous avons associé une mixture de Gaussiennes dans l'espace F2–F4, le but étant de trouver une description de la dispersion acoustique pour chaque endroit de constriction.[8]

Appelons $\mathcal{M}_j$ la mixture correspondant à la constriction j. Un point y quelconque dans l'espace F2–F4 aura un probabilité $P(y|\mathcal{M}_j)$ d'appartenir à cette mixture, donnée par l'équation suivante :[9]

$$P(y|\mathcal{M}_j) = \sum_i g_{ji}|C_{ji}|^{-1/2}e^{-(y-y_{ji})^T C_{ji}^{-1}(y-y_{ji})} \tag{2.6}$$

[7] X_c ne peut prendre la valeur d1 étant donnée qu'une constriction à la première ligne de la grille implique toujours la pointe de la langue.

[8] Le premier formant de chacune de ces configurations étant systématiquement proche de 190Hz (Fant 1969, Calliope 1989, chapitre 3), il n'a pas été pris en compte.

[9] les valeurs des g_{ij} doivent être tels que $\int dy P(y|\mathcal{M}_j) = 1$.

Où : g_{ij} – probabilités à priori.

y_{ij} – vecteurs des moyennes.

C_{ij} – matrices de covariance.

Soit y_{X_c} un point dans l'espace formantique F2–F4 produit par le modèle articulatoire en faisant une constriction à X_c, nous avons généré pour chaque lieu de constriction X_c 406 points correspondant aux positions articulatoires de la base de données originale ne présentant pas de constriction. Le tableau 2.1 montre pour un lieu de constriction X_c, le nombre de points (parmi les 406) qui a été reconnu comme appartenant à la mixture des Gaussiennes $\mathcal{M}_j$.

	t3	t2	t1	ll	d2	d3	d4	d5	d6	d7	d8	d9
$\mathcal{M}_{t3}$	292	18	1	13	34	66	9	5	6	13	16	1
$\mathcal{M}_{t2}$	32	294	37	5	60	14	3	1	0	8	16	4
$\mathcal{M}_{t1}$	17	41	349	2	20	7	4	1	0	0	3	3
$\mathcal{M}_{ll}$	32	21	6	361	12	16	10	8	7	8	1	1
$\mathcal{M}_{c2}$	2	14	2	0	248	29	2	0	3	2	3	6
$\mathcal{M}_{c3}$	4	0	0	2	19	252	4	1	1	4	2	4
$\mathcal{M}_{c4}$	8	0	4	2	0	7	249	27	26	17	1	2
$\mathcal{M}_{c5}$	1	2	1	0	3	2	41	191	59	32	16	11
$\mathcal{M}_{c6}$	0	3	1	5	0	1	13	57	192	25	3	4
$\mathcal{M}_{c7}$	3	4	2	8	0	3	30	44	50	225	29	23
$\mathcal{M}_{c8}$	11	4	2	1	0	7	23	34	28	41	259	36
$\mathcal{M}_{c9}$	4	5	1	7	10	2	18	37	34	31	57	311

TAB. 2.1: Matrice de confusion pour les produits acoustiques de chaque lieux de constriction du modèle articulatoire.

Remarquons que les fortes valeurs de probabilité se concentrent sur la diagonale principale. Ceci équivaut à dire que les constrictions que notre modèle est capable de produire peuvent être différenciées par leurs résultats acoustiques.

2.6 Conclusions

Nous avons décrit dans ce chapitre le modèle articulatoire que nous utiliserons pour essayer d'inférer des commandes articulatoires correspondant à des séquences voyelle-consonne. Nous avons étudié les problèmes qui ont été rencontrés avec des modèles articulatoires semblables à celui que nous avons utilisé.

Nous avons mis en œuvre une méthode simple et efficace pour produire des constrictions et nous avons observé que les différents lieux de constriction susceptibles d'être produits par le modèle se différencient dans l'espace formantique. Il reste à vérifier si les résultats acoustiques des séquences que le modèle est capable de produire sont compatibles avec les mêmes séquences prononcées par des locuteurs quelconques.

Notre modèle ayant été obtenu à partir d'un sujet précis, il est clair qu'une modification du modèle s'impose si l'on veut lui faire suivre des séquences acoustique provenant de sujets quelconques. Le chapitre suivant est consacré à résoudre ce point.

Chapitre 3

Normalisation de l'Espace Acoustique

La normalisation de l'espace acoustique d'un ensemble de locuteurs est depuis longtemps un problème d'une grande importance. Il est intéressant pour plusieurs raisons d'extraire les différences entre les locuteurs pour arriver à une seule représentation de chaque phonème dans l'espace acoustique. Ceci a une importance capitale dans la présente étude, puisque nous avons pour but de faire suivre au robot parlant une trajectoire présentée par n'importe quel locuteur. La représentation de la trajectoire doit être telle que le robot soit capable de la suivre et doit en plus avoir une certaine cohérence avec ce que le locuteur présente comme entrée au système.

Ce chapitre est consacré à l'étude qui a été effectuée en vue de l'obtention de paramètres pertinents qui permettront au modèle articulatoire de suivre les trajectoires présentées au système par un locuteur quelconque dans l'espace distal (i.e. l'espace acoustique). Nous nous intéressons notamment à la simulation du conduit vocal d'un locuteur en utilisant celui d'un autre locuteur.

3.1 Introduction

Les formants produits par le robot étant liés aux longueurs des cavités du modèle articulatoire, il est évident qu'une modification de celui-ci s'impose si l'on veut qu'il y ait une correspondance entre les réalisations acoustiques d'un même phonème par le locuteur d'une part et par le robot parlant d'autre part (sauf, bien entendu, dans le cas où le locuteur est celui à partir duquel le modèle a été construit). Ceci nous amène à faire une normalisation du conduit vocal du modèle de façon à ce que les valeurs des formants se correspondent le mieux possible.

L'intérêt central de notre travail est de modifier le conduit vocal du robot de façon à ce qu'il soit capable de suivre les trajectoires formantiques présentées à l'entrée du système et plus précisément, de répondre aux deux questions suivantes :

- De quelle façon faut-il modifier le conduit vocal du robot parlant pour qu'il puisse produire les formants présentés par un locuteur quelconque ?
- La configuration articulatoire utilisée pour produire les formants en question correspond-elle au phonème présenté par le locuteur ?

Avant de continuer, nous soulignons que la totalité de mesures formantiques des sujets ainsi que celles du modèle articulatoire est faite dans l'échelle des Barks, utilisant les expressions analytiques proposées par Zwicker & Terhardt (1980). De cette manière, les distances entre configurations formantiques qui seront calculées respecteront la sensitivité fréquentielle de l'appareil auditif humain, ce qui n'aurait pas été le cas si nous avions utilisé l'échelle en Hz.

3.2 Quel genre de paramètres choisir, et combien ?

L'objectif d'une normalisation est de trouver une fonction permettant de représenter l'espace acoustique d'un locuteur quelconque dans un espace acoustique de référence.

Il existe dans la littérature de nombreuses études ayant pour but la normalisation vocalique d'un ensemble de locuteurs. Parmi ces études il y a celles qui ne prennent en compte que les caractéristiques acoustiques des voyelles à normaliser. Il y a aussi celles qui cherchent à faire une correspondance entre les traits du signal acoustique et l'anatomie du conduit vocal du sujet qui l'a produit. Puisque nous travaillons avec un modèle articulatoire obtenu à partir de données anatomiques réelles d'un locuteur, nous nous sommes penchés surtout sur les dernières. Nous citons néanmoins des études traitant du premier cas car elles sont également importantes dans le domaine de la normalisations des espaces acoustiques.

Parmi le premier type de normalisation, nous trouvons notamment Matsumoto & Wakita (1986), Lobanov (1971), Harshman & Papcun (1976) et Zahorian & Jagharghi (1991).

Zahorian & Jagharghi (1991) utilisent deux méthodes purement mathématiques pour arriver à une normalisation relativement bonne. Le point intéressant dans leur approche est le fait qu'ils s'occupent des caractéristiques statiques **et** dynamiques des voyelles. Ceci est important pour notre étude car nous voulons que le robot parlant suive des **trajectoires** et pas seulement des entrées stationnaires. Ils ont obtenu des résultats pour les traits dynamiques qui sont en fait meilleurs que ceux des traits statiques. Nous remarquons qu'il existe des techniques qui peuvent être utilisées pour normaliser à la fois des traits acoustiques statiques et dynamiques.

En ce qui concerne le deuxième type d'approche, nous trouvons les études de Nordström (1977), Mattingly (1966) et Wakita (1977).

Mattingly (1966) compare la distribution des valeurs des formants pour chacun des groupes (homme, femme, enfant) des données formantiques de Peterson & Barney (1952). D'après l'hypothèse selon laquelle les différences entre les valeurs des formants parmi des individus du même groupe sont dues aux dimensions du conduit vocal, on s'attendrait à ce qu'il y ait une haute corrélation entre les distributions de chaque groupe. Il s'avère cependant que les corrélations sont plutôt basses. On conclut que la taille du conduit vocal seule ne suffit pas à expliquer les différences entre les groupes de sujets.

Nordström (1977) modifie des fonctions d'aire obtenues pour un sujet masculin de façon à simuler les conduits vocaux des femmes et des enfants. Cette étude montre qu'une distorsion non-linéaire peut rendre compte en partie des différences formantiques des voyelles produites par des sujets de ces trois groupes. Même si les auteurs reconnaissent que comprendre la normalisation chez l'auditeur nécessite une meilleure compréhension des traitements auditifs des différences formantiques, leurs résultats nous confortent dans notre approche, à être décrite ci-dessous.

Parmi les exemples de normalisation liés aux caractéristiques anatomiques des individus nous citerons celle proposée par Wakita (1977). La méthode consiste, tout d'abord, en l'estimation de la longueur du conduit vocal du sujet ayant produit les voyelles pour calculer par la suite un coefficient de correction des formants. En multipliant les formants des voyelles étudiées par le coefficient correspondant, on obtient une bonne réduction des ellipses de dispersion.

Nous trouvons donc qu'il n'y a pas que les dimensions longitudinales globales du conduit vocal d'un individu qui soient déterminantes de la structure formantique des phonèmes qu'il produit. Payan & Perrier (1993),avec une approche semblable à celle que nous présenterons, se servent d'un modèle articulatoire et des cavités acoustiques (Bailly 1995*a*) et réussissent à ramener les formants d'un locuteur à l'espace formantique du locuteur de base.[1] Bien que leur méthode ait donné des résultats satisfaisants, elle ne peut être utilisée que si l'on connaît les affiliations des formants aux cavités acoustiques. Toute trajectoire formantique pour laquelle on ne connaît pas les affiliations de façon continue reste donc non normalisable par leur méthode.

Nous remarquons néanmoins que, d'après l'étude de Wakita (1977), les dimensions du conduit vocal ont un rôle non négligeable dans l'effet acoustique d'un phonème. Nous allons donc chercher d'abord à inférer les dimensions longitudinales du conduit vocal du sujet pour ensuite essayer de trouver une configuration articulatoire telle que le conduit vocal modifié produise les formants présentés par le locuteur. Cette démarche doit en principe compenser à la fois les différences anatomiques entre les locuteurs (grâce à la distorsion longitudinal) et les différentes stratégies articulatoires qu'ils utilisent lorsqu'ils produisent le même phonème (grâce à la recherche des nouvelles configurations articulatoires).

[1]cf. aussi Payan (1992) pour une description plus complète de la méthode et Payan, Apostol & Perrier (1995) pour une évaluation de cette méthode

Des données anthropométriques (Scheideman et al. 1980) ont été étudiées pour décider du nombre de paramètres à utiliser. La figure 3.1 présente des données moyennes pour l'homme et pour la femme obtenues à partir des clichés radiographiques. Nous avons marqué des points entre lesquels nous mesurons les distances. Pour de raisons de visualisation, la figure ne montre que les points sur les données de la femme.. Le tableau 3.1 présente les distances moyennes obtenues pour l'homme et pour la femme.

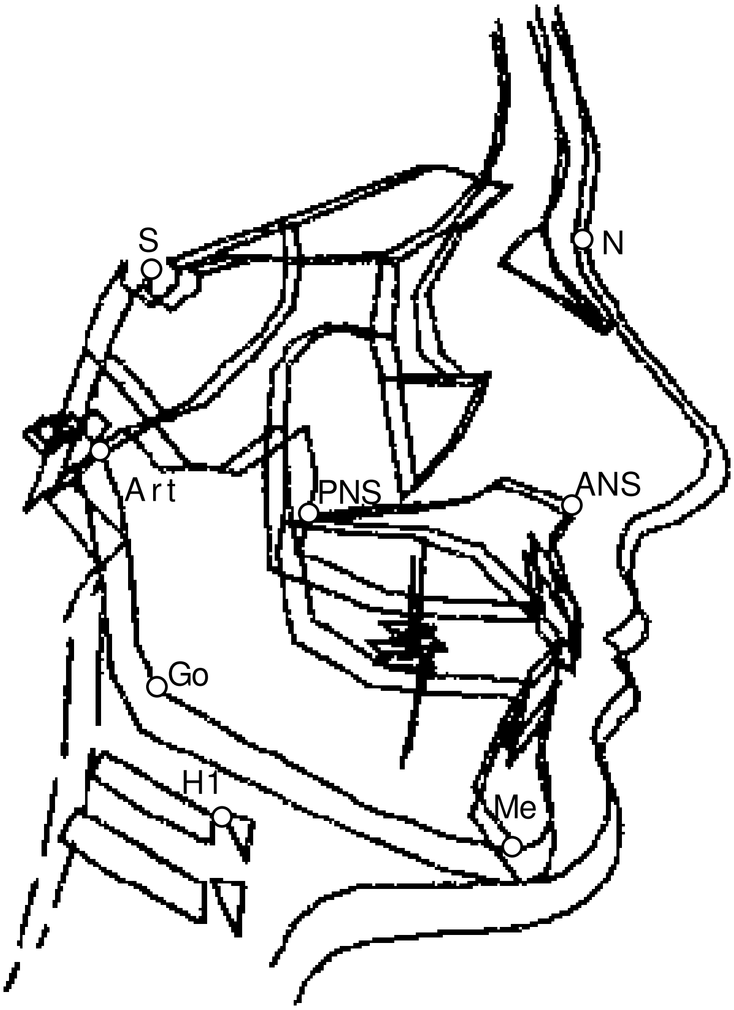

FIG. 3.1: Données anthropométriques moyennes pour l'homme et pour la femme (d'après (Scheideman et al. 1980)).

On observe que les distances moyennes des hommes sont systématiquement plus grandes que celles des femmes d'une valeur entre 4 et 6%. Ceci est vrai pour toutes les distances sauf celles marquées « S–H1 » et « Art–Go ». Ces deux distances correspondent à des mesures prises le long de la partie arrière du conduit vocal. Nous choisissons donc d'utiliser deux paramètres de modification de la longueur du conduit vocal correspondant à l'élongation des sections orale et pharyngale. La figure 3.2 indique l'endroit où nous avons placé cette

Mesure	Homme	Femme	Rapport
Go–Me	80.3	76.7	1.047
PNS–ANS	56.4	53.1	1.063
N–S	76.3	72.6	1.051
Art–N	103.0	96.6	1.066
N–ANS	56.4	54.1	1.043
S–Me	136.8	128.5	1.065
S–H1	115.6	101.7	1.137
Art–Go	52.6	47.1	1.117

TAB. 3.1: Mesures anatomiques d'hommes et de femmes avec leur rapport.

division.[2]

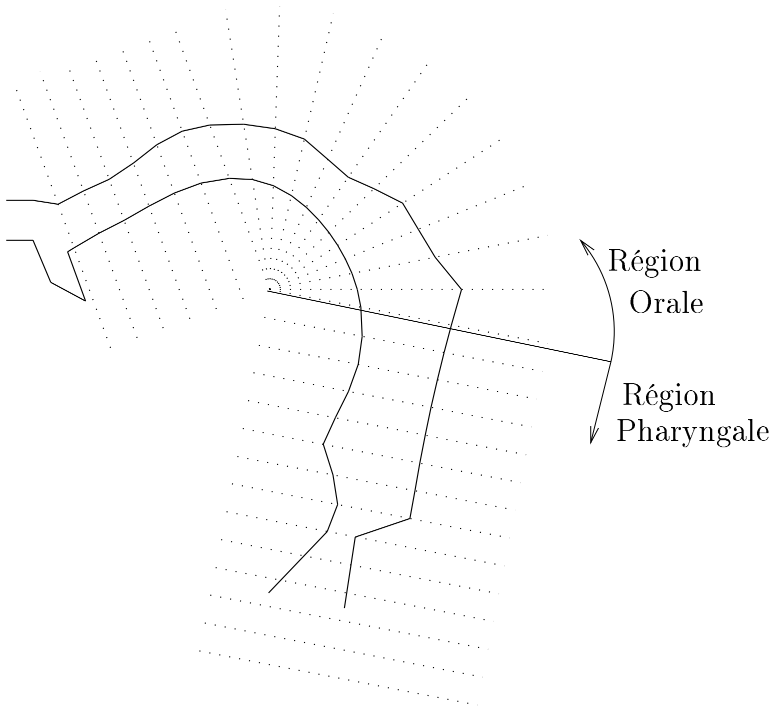

FIG. 3.2: Séparation du conduit vocal en partie orale et pharyngale.

Nous allons donc utiliser deux paramètres pour représenter les dimensions du conduit vocal d'un locuteur quelconque. Ces deux paramètres représenteront la longueur du conduit vocal du sujet par rapport à celle du modèle articulatoire utilisé. Ainsi, une valeur de $d_o = 1.1$ correspond à une longueur de la cavité orale 10% plus grande que celle du robot. Cette élongation (ou éventuel rétrécissement) se fera au niveau de la fonction d'aire et nous l'appellerons dorénavant « distorsion ».

[2]Nos conclusions ne font que confirmer ce qui est présenté classiquement dans la littérature (voir, par exemple, Fant 1966)

3.3 Sélection des paramètres d_o et d_p

Il nous reste à définir la façon de choisir les valeurs des paramètres appropriées à chaque locuteur. Les valeurs des formants étant directement liées aux cavités du conduit vocal, il a souvent été dit que les formants des voyelles extrêmes représentent un bon moyen d'estimer les dimensions du conduit vocal (Verbrugge, Shankweiler & Edman 1976). Il est à noter que nous ne cherchons pas à trouver une meilleure méthode de classification de voyelles avec cette déformation, Verbrugge et al. (1976) ayant démontré que cette information n'aide pas à cette tâche de façon significative. Nous voulons, dans un premier temps, que le modèle articulatoire adapté soit capable de produire les mêmes formants que le sujet présente à l'entrée du système.

Nous commençons par obtenir les voyelles extrêmes du robot de la base de données à partir de laquelle le modèle a été obtenu. La figure 3.3 présente l'emplacement de ces voyelles dans l'espace acoustique du robot parlant. On peut remarquer que ces voyelles ne sont pas aux limites de l'espace acoustique. Ceci est dû au fait qu'elles ont été extraites des phrases et sont donc l'objet d'une réduction vocalique[3].

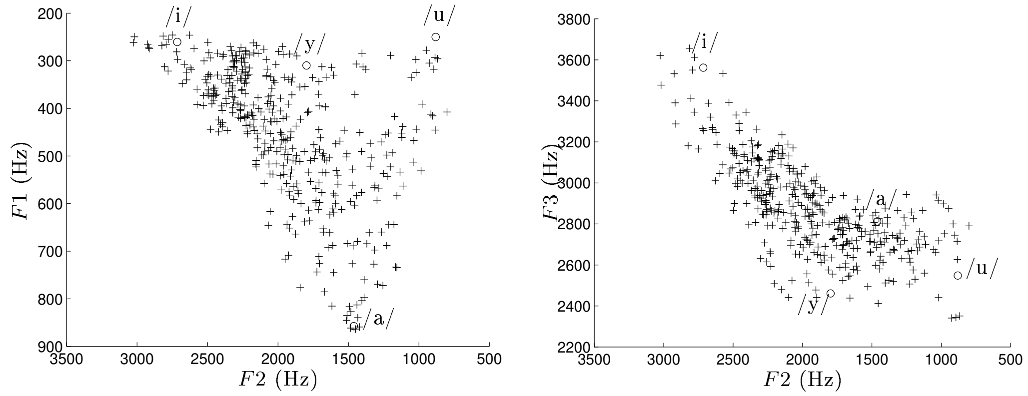

FIG. 3.3: Formants des voyelles cardinales dans la base de données.

Afin d'estimer les dimensions du conduit vocal du locuteur, nous comparerons les formants des voyelles extrêmes du locuteur par rapport à ceux du robot parlant. Nous avons vu que la relation entre les valeurs des formants d'une part et les dimensions du conduit vocal d'autre part n'est pas évidente. Nous sommes donc obligés de suivre une démarche telle qu'il y ait une relation aussi directe que possible entre les dimensions du conduit vocal du robot et les formants des voyelles extrêmes correspondants. Nous ferons un échantillonnage hexagonal de l'espace d_o–d_p et obtiendrons les formants des voyelles extrêmes pour chacune des paires (d_o, d_p). Nous prenons en compte les trois premiers formants de chaque voyelle, ce qui nous donne pour chaque point (d_o, d_p) dans l'espace $\mathbb{R}^2$, un point correspondant dans l'espace $\mathbb{R}^{12}$ (quatre voyelles avec trois formants pour chacune). Nous créons ainsi une

[3]cf. Loevenbruck & Perrier (1993). Pour une étude plus approfondie, voir Fowler (1980), Whalen (1990).

« grille » de points dans l'espace d_o, d_p et une « grille » correspondante dans l'espace $\mathbb{R}^{12}$ (cf. figure 3.4).

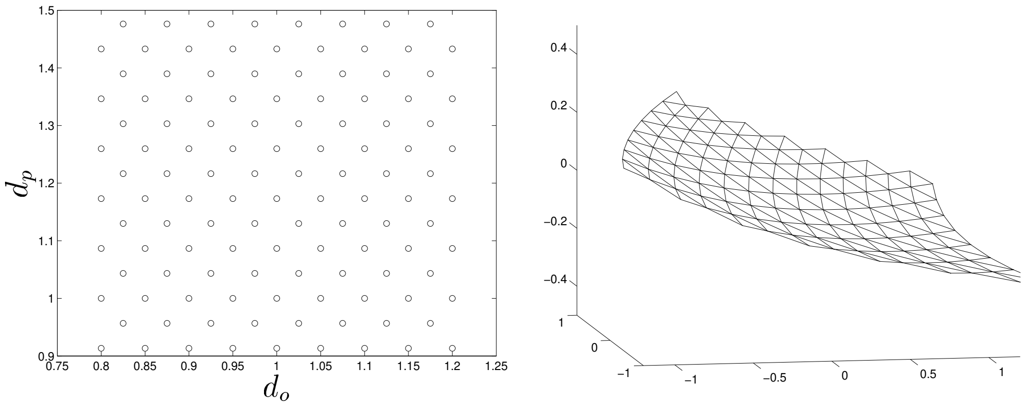

FIG. 3.4: Échantillonnage des paramètres d_o, d_p (droite) et la grille correspondant en $\mathbb{R}^{12}$ (gauche). Seuls les trois composantes principales de l'espace $\mathbb{R}^{12}$ sont représentés

Pour trouver les valeurs appropriées de $(d_o,\ d_p)$ d'un locuteur quelconque nous suivons les étapes suivantes :[4]

1. Nous prenons les trois formants de chacune des voyelles extrêmes du locuteur pour créer un point P_l dans l'espace $\mathbb{R}^{12}$.
2. Nous projetons le point P_l sur le grille en $\mathbb{R}^{12}$.
3. La valeur appropriée des paramètres (d_o, d_P) est le point correspondant sur le grille en $\mathbb{R}^2$.

Une représentation de notre méthode en termes de schéma blocs est montré à la figure 3.5

Nous testons d'abord cette méthode avec la base de données obtenue par Neagu (1994). Cette base de données est composée de dix voyelles prononcées par douze locuteurs (six hommes, six femmes). Les valeurs des formants de chacune des voyelles sont extraites du milieu d'une phrase ; nous nous attendons donc à ce qu'il y ait une réduction vocalique, comme c'est le cas pour le robot parlant. La figure 3.6 présente les valeurs obtenues des paramètres d_o et d_p à partir des projections des locuteurs sur la grille en $\mathbb{R}^{12}$.

On observe qu'il y a une bonne distinction entre les valeurs des hommes et des femmes et qu'en plus ces valeurs sont en concordance avec ce que l'on aurait pu espérer : les valeurs des paramètres des femmes sont plus petites que celles des hommes.

[4]pour une description plus détaillée, voir l'annexe A.

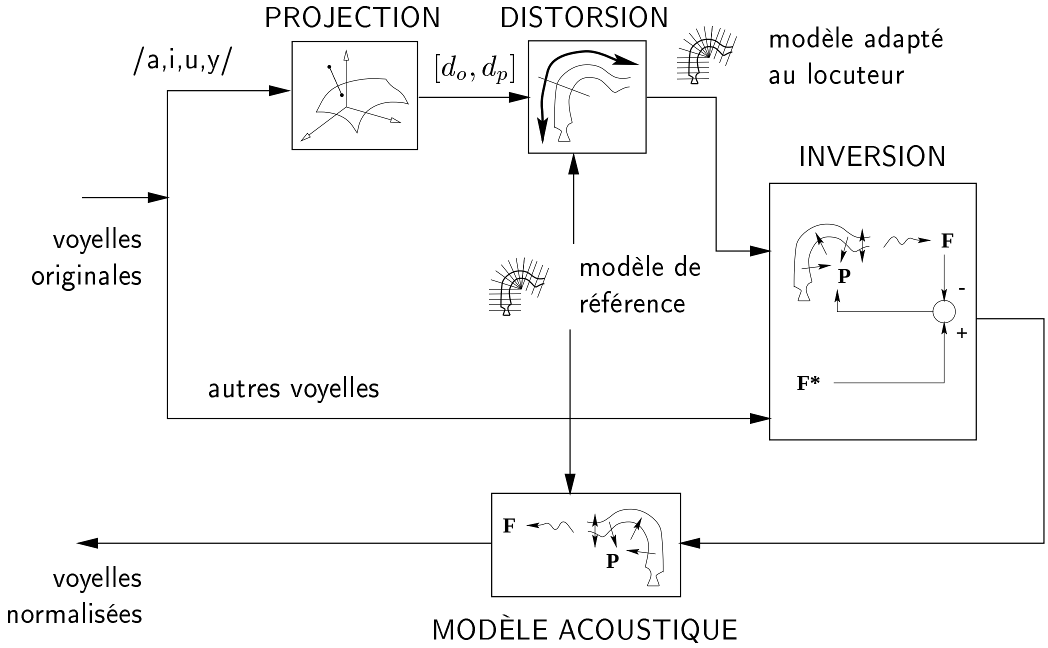

FIG. 3.5: Schéma synoptique de la méthode de normalisation du locuteur fondée sur une distorsion longitudinal du modèle articulatoire.

3.4 Approximation aux formants du locuteur

Dans la section précédente nous avons décrit la façon d'obtenir les valeurs optimales des paramètres d_o et d_p pour un locuteur quelconque. Nous savons que, bien que ceci soit important lors de la normalisation (Wakita 1977), il ne suffit pas de modifier la longueur du conduit vocal pour obtenir les formants d'un locuteur. Nous chercherons alors à trouver une configuration articulatoire telle que le robot parlant, dont le conduit vocal a été l'objet d'une distorsion, produise les formants présentés à l'entrée du système.

Avant de continuer, nous définirons une « fibre articulatoire » comme l'ensemble de toutes les configurations articulatoires qui correspondent à un seul point dans l'espace $F1$–$F3$. Une question importante se pose : une fibre articulatoire en reste-t-elle une si l'on fait subir une distorsion au conduit vocal qui la produit ?

Pour répondre à cette question, nous prendrons plusieurs points d'une fibre articulatoire et nous modifierons ensuite la longueur du conduit vocal qui la produit pour observer la distribution des points correspondant en $F1$–$F3$. Nous utiliserons des fibres articulatoires des voyelles /i/, /a/, /y/ et /u/ et nous les testerons sur des conduits vocaux modifiés par les paramètres trouvés dans la section précédente.[5]

[5]Chaque fibre a été obtenue en trouvant, à partir de plusieurs points de départ, plusieurs configurations articulatoires produisant les formants $F1$–$F3$ de la voyelle désirée

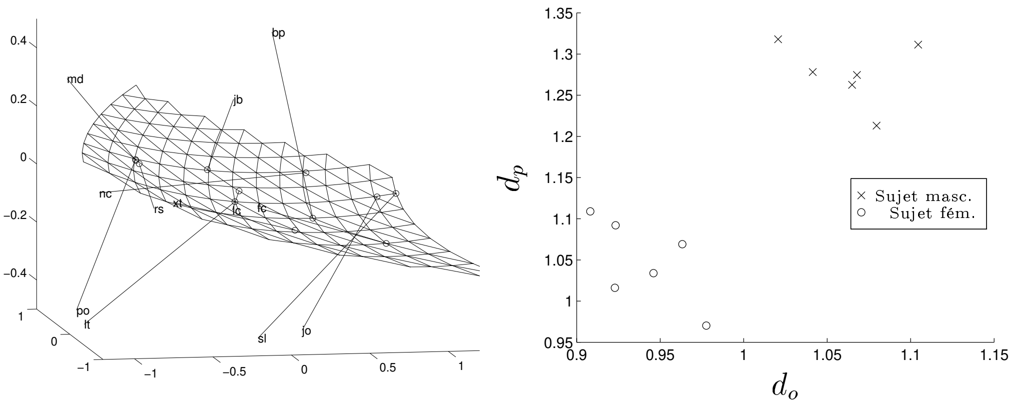

FIG. 3.6: Exemples de projections des données pour chaque locuteur sur la grille correspondante au modèle articulatoire (gauche) et valeurs des paramètres d_o et d_p pour la base de données 2 à partir des formants des locuteurs (droite).

Le tableau 3.2 présente les configurations articulatoires utilisées. On remarque que les paramètres articulatoires présentent une variance non négligeable et que les nuages de points correspondants dans l'espace $F1$–$F3$ (voir figure 3.7) peuvent être représentés par un seul point. Les coupes sagittales qui en résultent sont présentées sur la figure 3.8.

Nous observons d'après la figure 3.9 que les fibres articulatoires des voyelles testées se rassemblent, après la distorsion, en un nuage assez serré dans l'espace formantique correspondant. Sur la base de ces observations, nous ferons la généralisation suivante : la fibre articulatoire d'un phonème quelconque reste une fibre articulatoire lorsque le conduit vocal qui la produit subit une distorsion.

À présent, pour arriver à répondre à la première question du début du chapitre (page 22) nous essayerons de trouver une configuration articulatoire telle que, le conduit vocal ayant subi la distorsion appropriée, le point correspondant en $F1$–$F3$ soit égal au point présenté par le locuteur. Pour ce faire nous utiliserons des algorithmes d'optimisation. Nous trouverons une configuration articulatoire pour chacune des voyelles de chacun des locuteurs de la base de données obtenue de Neagu (1994). La figure 3.10 présente les résultats de cette procédure. Pour chaque locuteur, chaque voyelle est représentée par un cercle et les croix représentent les points atteints par le robot parlant, dont le conduit vocal a été l'objet d'une distorsion.

En observant les points atteints par le robot parlant, nous sommes en mesure de dire que nous avons trouvé une réponse à la première question (page 22). Mais nous devons encore répondre à la deuxième.

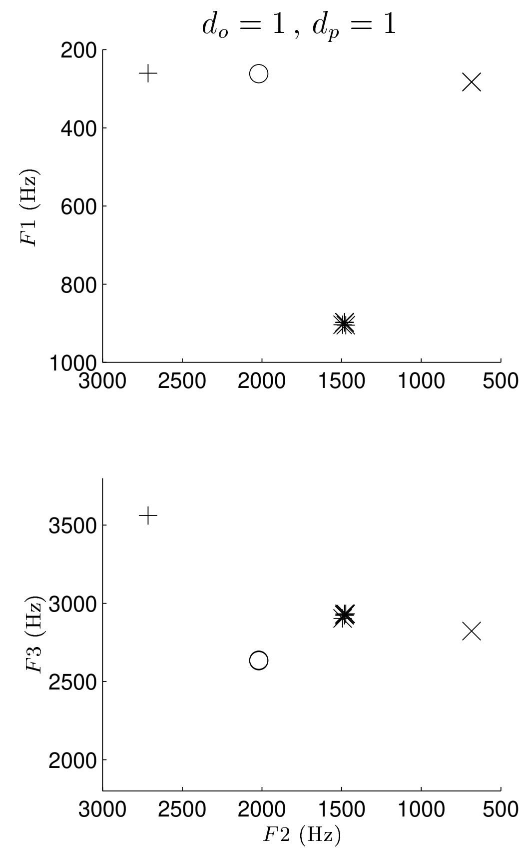

FIG. 3.7: Formants des fibres articulatoires pour chacune des voyelles extrêmes.

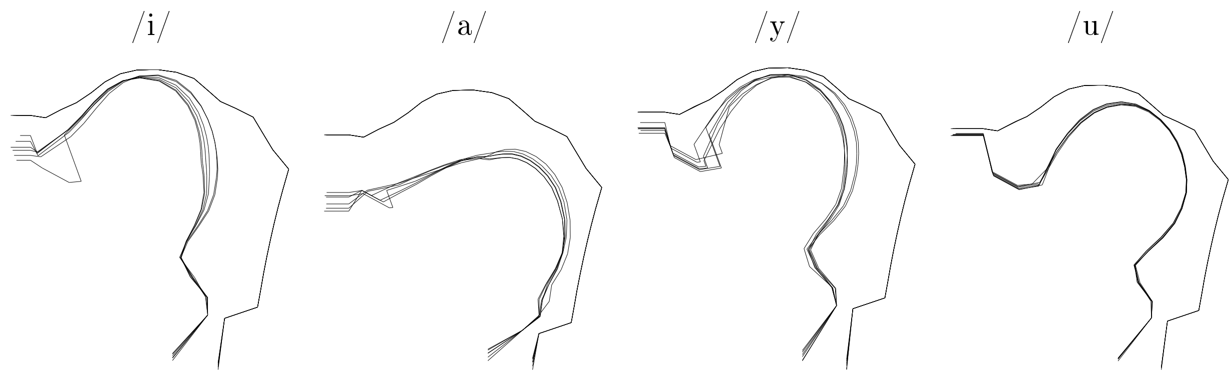

FIG. 3.8: Coupes sagittales pour les configurations au long des fibres articulatoires de chaque voyelle extrême .

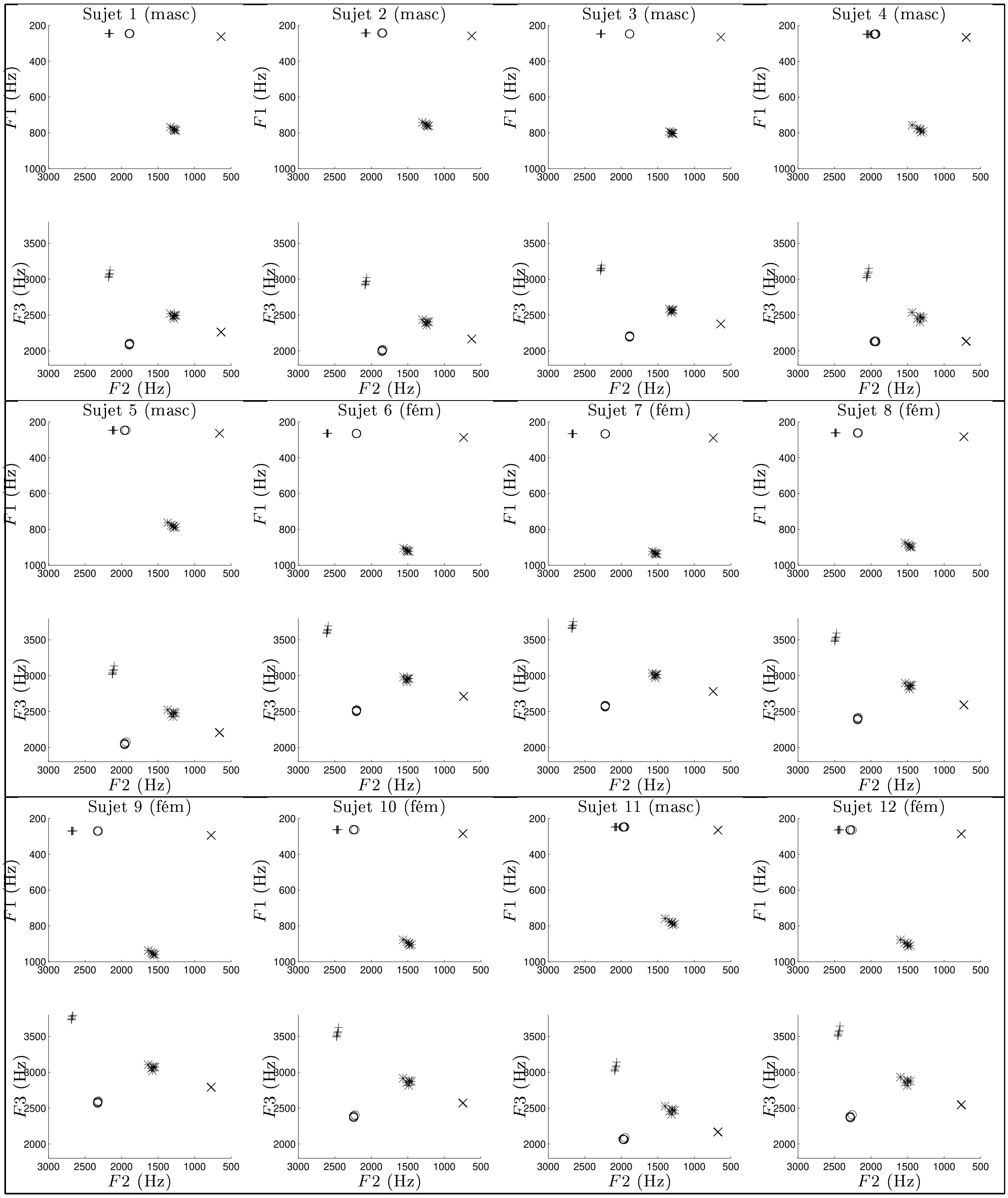

FIG. 3.9: Formants résultants de la fibre articulatoire après application des distorsions au modèle articulatoire.

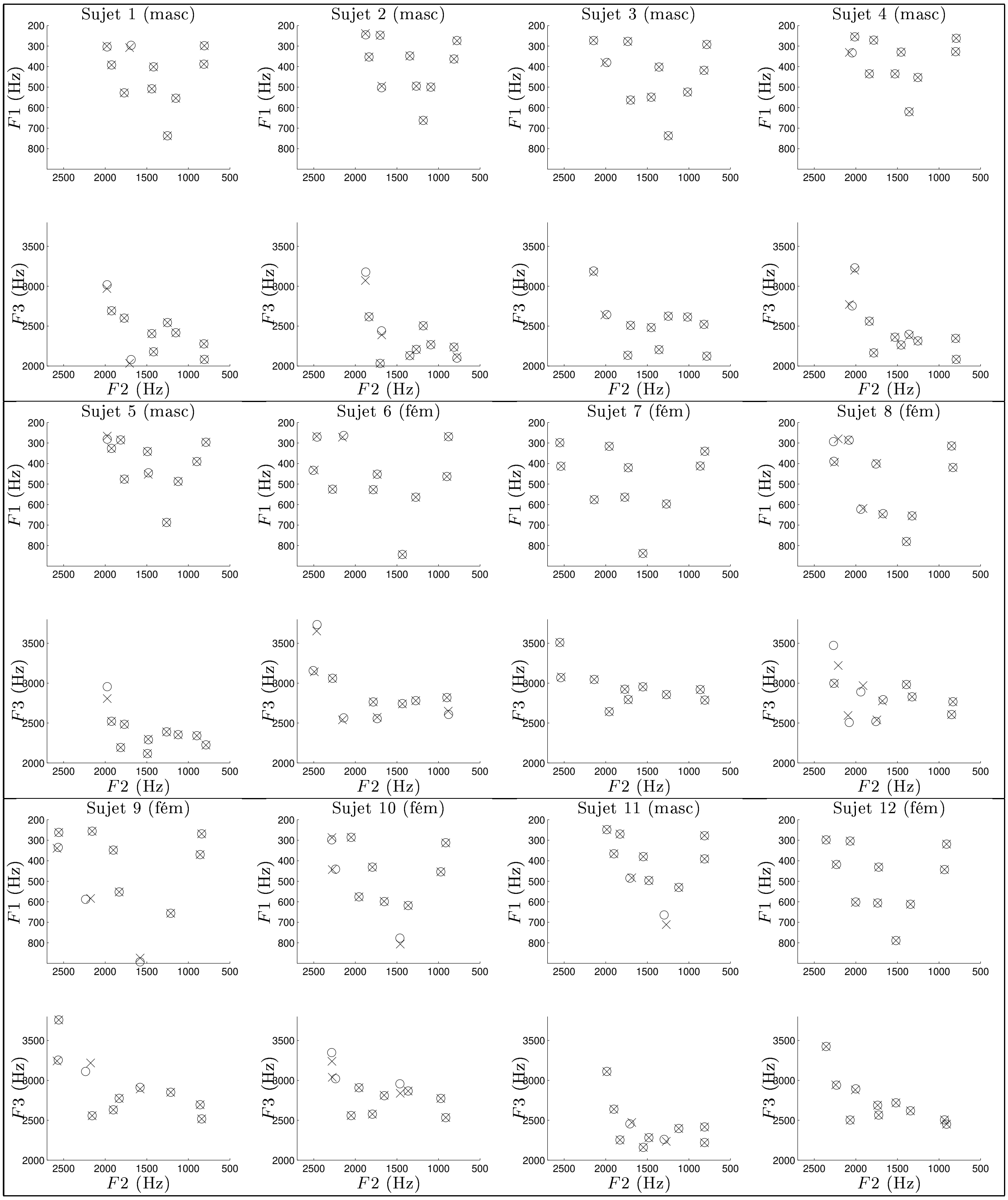

FIG. 3.10: Base de données 2 (cercles) et points atteints par le robot parlant (croix).

phonème	j	b	d	t_x	t_y	l_x	l_y	p
/i/	-0.893	-2.103	0.806	-0.775	-0.479	0.275	0.485	0.368
	-1.000	-2.308	0.454	-0.435	-0.663	0.590	0.292	0.207
	-1.091	-2.378	0.357	-0.671	-0.749	-0.975	-0.394	0.164
	-1.246	-1.657	2.127	-1.228	0.144	0.022	0.724	0.855
	-0.623	-1.616	1.526	1.290	1.228	-0.369	1.095	0.556
	-0.754	-1.872	1.236	-1.242	-0.606	-0.236	-0.009	0.617
/a/	-0.681	2.334	-0.732	-0.746	-1.215	1.123	2.566	-1.752
	-1.435	1.350	-1.585	0.632	-1.384	1.490	2.868	-0.940
	-0.873	1.627	-1.688	0.126	-2.208	1.185	2.784	-0.578
	-1.475	1.486	-1.175	1.230	-1.664	1.083	1.769	-1.333
	-2.108	1.342	-0.781	1.837	-0.130	1.404	1.864	-1.655
	-1.441	1.511	-1.163	1.540	-1.745	1.174	2.044	-1.352
/y/	2.972	0.412	1.719	2.038	-1.617	1.788	0.329	2.005
	1.463	-0.099	2.279	0.619	-1.829	1.472	-0.064	1.956
	1.088	-1.063	1.252	1.555	-2.054	1.878	-0.245	1.370
	0.511	-1.190	1.324	2.237	0.181	1.946	-0.308	1.121
	0.902	-1.166	0.825	2.080	0.307	1.983	-0.323	0.942
	2.429	-0.322	0.982	0.245	-1.831	1.539	-0.385	1.580
/u/	0.449	1.040	2.504	1.927	-2.461	3.000	-0.884	0.636
	0.805	1.238	2.620	1.041	-2.669	2.207	-0.929	0.844
	0.572	1.069	2.506	2.003	-2.184	2.657	-0.914	0.653
	0.195	0.859	2.591	2.016	-1.306	2.605	-0.996	0.557
	0.011	0.752	2.731	1.975	-1.402	2.743	-0.982	0.620
	-0.086	0.653	2.697	2.423	-1.072	2.111	-1.095	0.546

TAB. 3.2: Configurations articulatoires donnant un seul point dans l'espace formantique

3.5 Vérification de la normalisation

Cette section est consacrée à la réponse de la deuxième question de la page 22. Nous démarrons avec les hypothèses suivantes :

1. À chaque phonème correspond une fibre articulatoire, et la réalisation de chaque phonème par un locuteur quelconque appartient à la fibre correspondante produite par un conduit vocal dont on ignore les dimensions.
2. Toutes les réalisations d'un même phonème par un locuteur appartiennent à la même fibre articulatoire.

Dans la section précédente nous avons trouvé un moyen d'estimer la longueur des conduits vocaux. Pour ce faire nous avons eu besoin de connaître les formants des quatre voyelles extrêmes du locuteur en question. Nous avons vu de plus qu'une fibre articulatoire reste relativement invariante lorsque le conduit vocal qui la produit est l'objet d'une distorsion. Nous allons ensuite vérifier les hypothèses ci-dessus avec la base de données de Neagu (1994).

Nous avons déjà trouvé pour chaque locuteur des valeurs appropriées des paramètres d_o, d_p et nous avons en outre trouvé des configurations articulatoires qui produisent chacun des phonèmes de chacun des locuteurs de cette base de données. Si les hypothèses ci-dessus sont vraies, toutes les réalisations d'un seul phonème par le robot parlant (ayant subi une distorsion) font partie de la même fibre articulatoire. Nous procéderons donc à la réalisation des configurations articulatoires utilisées pour obtenir les formants de la figure 3.10 sur le conduit vocal du robot parlant.

Nous présentons sur la figure 3.11 les ellipses de dispersion (un écart type) de chacune des voyelles de la base de données de Neagu (1994). La figure 3.12 présente les ellipses des mêmes voyelles produites par le robot parlant en utilisant les configurations articulatoires trouvées pour la figure 3.10, avec $d_o = 1, d_p = 1$.

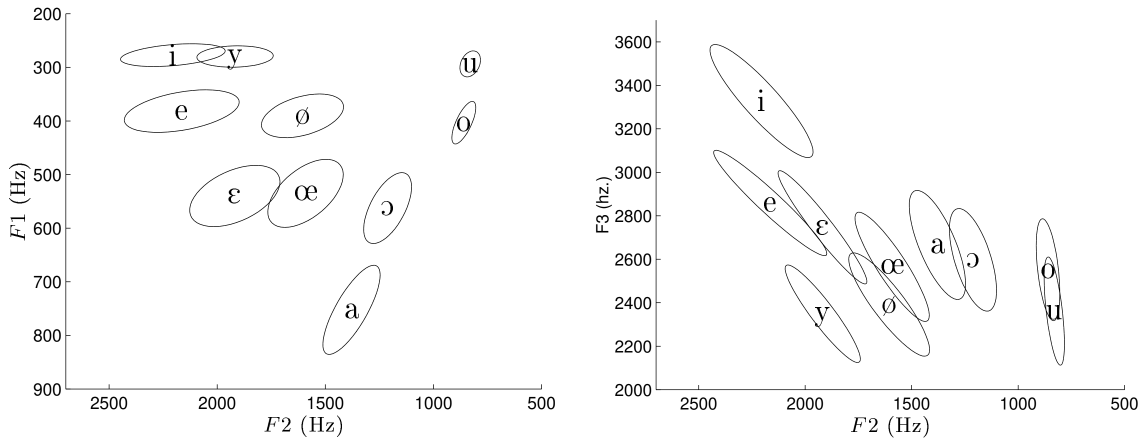

FIG. 3.11: Ellipses de dispersion des voyelles de la base de données 2.

Nous procéderons ensuite à l'évaluation des résultats obtenus. Si nous supposons que les voyelles produites par le robot appartiennent chacune à une fibre articulatoire, alors les volumes des ellipsoïdes de dispersion[6] des locuteurs de la base de données et du robot parlant doivent être de grandeur comparable.

Nous savons que le volume d'un ellipsoïde (figure 3.13) décrit par l'équation :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \tag{3.1}$$

est donné par l'expression $V = \frac{4}{3}\pi abc$, les valeurs des variables a^2, b^2, c^2 correspondant aux valeurs propres de la matrice de covariance des données représentées par l'ellipsoïde.

[6]Puisque nous travaillons dans l'espace $F1 - -F3$ nous utiliserons comme mesure les volumes des ellipsoïdes de dispersion et non pas les aires des ellipses.

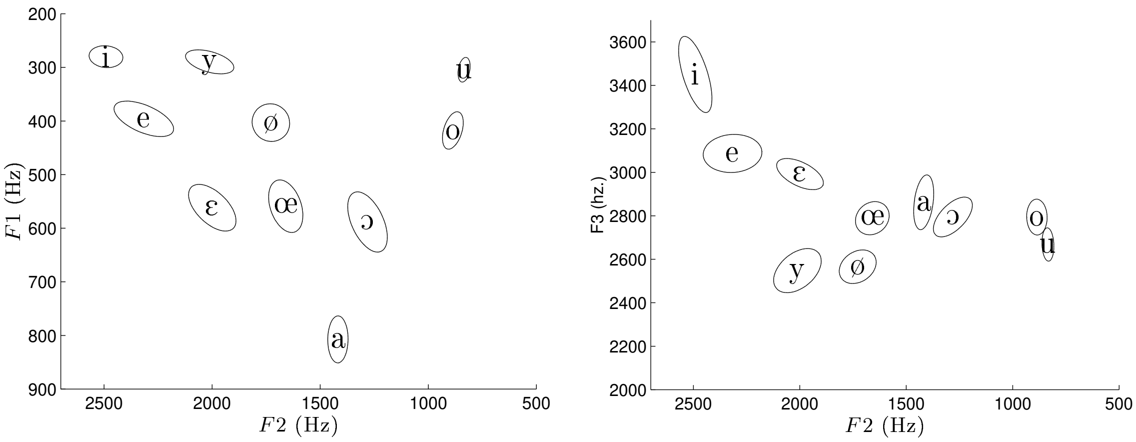

Fig. 3.12: Ellipses de dispersion du robot parlant.

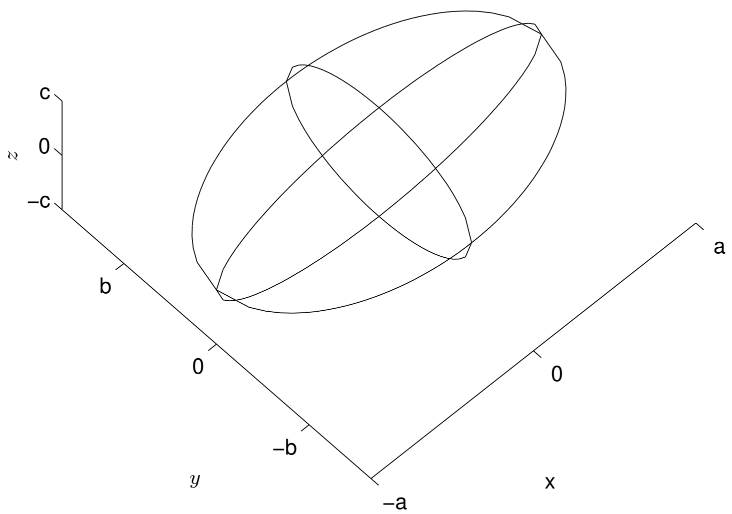

Fig. 3.13: Exemple d'un ellipsoïde.

Le tableau 3.3 présente ces volumes pour chaque voyelle de chaque locuteur. Nous présentons également les volumes des ellipsoïdes produits par le robot parlant. Nous observons que tous les volumes des ellipsoïdes du robot sont bien dans le rang des volumes des locuteurs. Nous sommes alors en mesure de dire que l'hypothèse no. 1 est vraie.

locuteur	œ	ø	a	ɛ	ɔ	e	i	o	u	y	somme
1(m)	.021	.010	.048	.017	.083	.007	.015	.035	.068	.018	.321
2(m)	.047	.017	.067	.017	.045	.005	.005	.008	.019	.005	.235
3(m)	.054	.035	.078	.092	.042	.015	.012	.064	.034	.009	.433
4(m)	.061	.037	.114	.016	.163	.006	.014	.029	.101	.018	.561
5(m)	.101	.030	.227	.083	.112	.011	.022	.118	.034	.028	.766
6(f)	.018	.007	.186	.018	.083	.005	.003	.014	.100	.011	.444
7(f)	.023	.035	.514	.042	.094	.009	.016	.046	.123	.012	.914
8(f)	.046	.008	.039	.061	.022	.010	.018	.020	.195	.021	.441
9(f)	.019	.107	.293	.067	.291	.015	.015	.169	.036	.055	1.067
10(f)	.022	.013	.080	.050	.108	.024	.031	.043	.094	.034	.499
11(m)	.009	.009	.020	.013	.051	.004	.005	.035	.059	.010	.214
12(f)	.010	.023	.194	.013	.170	.006	.008	.042	.043	.017	.527
robot	.055	.026	.038	.047	.058	.096	.058	.044	.020	.063	.505

TAB. 3.3: Volumes des ellipses de chaque locuteur et du robot parlant.

En regardant la figure 3.12 nous observons que les ellipses de dispersion sont placées aux endroits correspondants aux voyelles de la base de données. Nous sommes alors en mesure d'affirmer que nous avons répondu à la deuxième question de la page 22.

3.6 Formalisation et évaluation la méthode de normalisation

Dans les sections précédentes nous avons répondu aux deux questions que nous nous sommes posées au début du chapitre, mais nous devons encore être sûrs que la méthode utilisée est appropriée. C'est à ce sujet que cette section est consacrée.

Nous commençons par formaliser la méthode que nous avons utilisée dans la section précédente :

1. Pour chaque locuteur de la base de données, estimer les paramètres de normalisation du conduit vocal à l'aide des formants des voyelles extrêmes (voir la description dans la page 27).
2. Trouver pour chaque phonème de chaque locuteur une configuration articulatoire telle que appliquée au robot parlant (dont le conduit vocal a été l'objet d'une distorsion avec les paramètres appropriés) produise les mêmes formants.

3. Les formants produits par les configurations articulatoires trouvées dans l'étape précédente (sans distorsion) seront les formants des phonèmes normalisés.

Analysons ensuite les points qu'il restent à considérer :

1. Pourrions-nous obtenir les paramètres d_o, d_p d'une autre façon ? Et cette façon-là aurait-elle de l'influence sur le degré de normalisation obtenu ?
2. La méthode utilisée fonctionnerait-elle sur une base de données dont les voyelles n'ont pas subi de réduction vocalique ?
3. Fonctionnerait-elle sur une base de données d'une autre langue ?

Pour répondre à ces questions, nous avons effectué le travail de normalisation sur trois bases de données. Pour chacune, plusieurs méthodes d'estimation de la longueur du conduit vocal ont été utilisées. Les sections suivantes présentent les résultats obtenus.

Nous utiliserons les critères d'évaluation introduits par Ferrari-Disner (1980), selon lesquels le critère pour le degré de succès d'une méthode de normalisation est associé à la réduction de la variance dans chaque groupe de voyelles censées appartenir à la même cible lors qu'elle est produite par des locuteurs différents, tout en préservant la séparation entre les groupes censés représenter des cibles différentes.

La méthode proposée par Ferrari-Disner (1980) utilise la relation des aires des ellipses normalisées par rapport aux ellipses de départ comme mesure d'évaluation. Nous évaluerons la relation des volumes des ellipsoïdes de dispersion des points obtenus.

3.6.1 Base de donnés 1

La première base de donnés utilisée est celle recueillie par Bailly (1995a), elle est composée par des valeurs de formants de voyelles dans la phrase *c'est VV ça ?* Nous avons utilisé neuf locuteurs et trois locutrices de cette base de données. Remarquons qu'il n'y a pas de réduction vocalique dans les formants de cette base de données. Des résultats erronés sont donc à prévoir si l'on se sert des configurations articulatoires utilisées pour la base de données précédente. La figure 3.14 présente les ellipses de dispersion dans les plans $F1/F2$ et $F2/F3$.

Étant donné le fait que la base de données en question ne présente pas de réduction vocalique, un changement de méthodologie s'impose afin de trouver des valeurs appropriées pour les variables d_o et d_p. Nous chercherons d'abord à simuler les stratégies utilisées par les locuteurs pour produire les voyelles de la base de données.

Nous trouvons dans la littérature plusieurs études visant à expliquer les choix des voyelles dans une langue. Schwartz, Boë & Vallée (1995) utilisent la théorie dite de « disper-

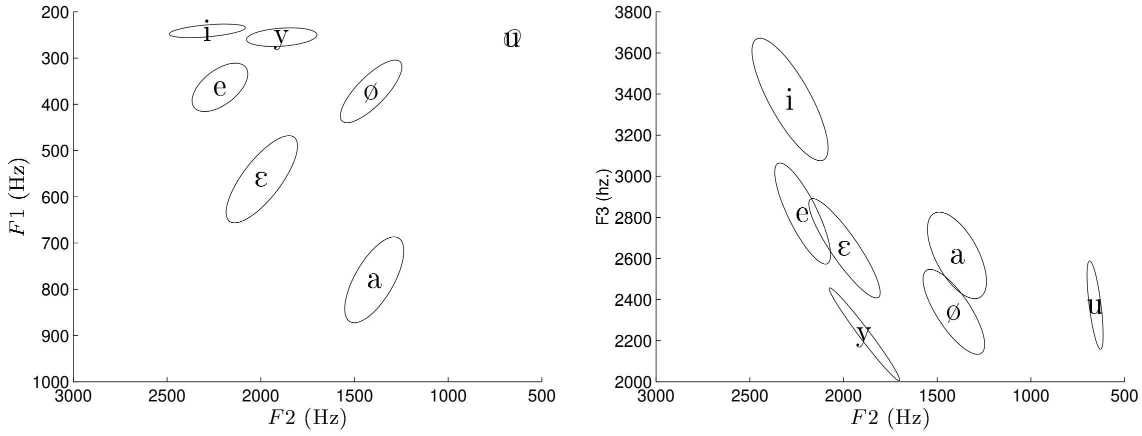

FIG. 3.14: Base de données 1.

sion/focalisation » pour prédire des systèmes de voyelles dans une langue. Boë, Schwartz, Laboissière & Vallée (1996) vont plus loin et y intègrent un modèle articulatoire. En utilisant la notion de « vie artificielle », Glotin (1995) arrive aussi à prédire des systèmes de voyelles. Ce dernier comporte un intérêt particulier dans le cadre de cette thèse, puisque le modèle articulatoire utilisé est celui que nous avons crée (chapitre 2).

Dans la ligne des travaux cités ci-dessus, nous élargissons le plus possible l'espace vocalique du robot parlant. En prenant comme point de départ les voyelles de la figure 3.3, nous utilisons les critères décrits dans le tableau 3.4 pour obtenir des voyelles dites « optimisées ». Nous utilisons encore un algorithme d'optimisation pour trouver des configurations articulatoires qui correspondent aux voyelles optimisées. La figure 3.15 présente l'emplacement de ces voyelles dans l'espace $F1 - F2$ et $F2 - F3$.

Voyelle	Critère Utilisé
/i/	min($F1$) et min($F4 - F3$)
/a/	max($F1$) et min($F2 - F1$)
/y/	min($F1 - F2$) et min($F3$)
/u/	min($F2 - F3$) et max($F3$)

TAB. 3.4: Critères d'optimisation des voyelles extrêmes.

Nous avons trouvé les stratégies que le robot utiliserait pour produire les quatre voyelles en question ; ceci étant dit, rien ne nous assure la validité de ces stratégies si le conduit vocal du robot est l'objet d'une distorsion. Nous chercherons donc des stratégies pour chacun des points extrêmes de la grille (figure 3.4) et nous ferons ensuite une interpolation pour trouver les configurations articulatoires appropriées pour chacun des points du grille(Abramowitz & Stegun 1972).

En d'autres termes, les points dans la grille en $\mathbb{R}^{12}$ ne seront pas construits en utilisant

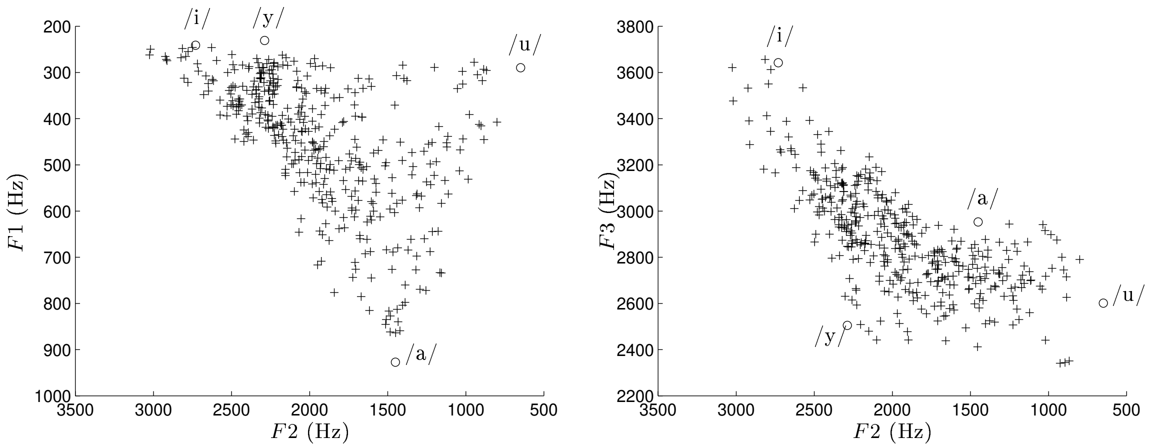

FIG. 3.15: Emplacement des voyelles optimisées.

tous la même configuration articulatoire. La configuration articulatoire utilisée pour chaque voyelle sera le résultat de l'interpolation des valeurs des articulateurs obtenues aux extrêmes de la grille.

3.6.2 Base de données 2

Cette base de données est celle que nous avons utilisée dans les sections précédentes. Comme nous l'avons déjà indiqué, les formants des voyelles sont l'objet d'une réduction vocalique. Nous avons obtenu des résultats satisfaisants du point de vue qualitatif. Cette section évaluera ces résultats quantitativement.

C'est avec cette base de données que nous avons répondu aux questions du début du chapitre et vérifié des hypothèses que nous considérons cruciales pour poursuivre l'étude qui nous occupe. La figure 3.11 montre les ellipses de dispersion de cette base de données.

3.6.3 Base de données 3

La troisième base de données que nous utiliserons dans notre étude est celle de Peterson & Barney (1952). Nous avons utilisé les 61 locuteurs adultes (33 masculins, 28 féminins) de cette base.

Les valeurs des formants de cette base de données ont été extraites du noyau vocalique des mots du type /hVd/ ; bien que minimale, une réduction vocalique est à prévoir.

La figure 3.16 montre les ellipses de dispersion des voyelles dans le plan $F1/F2$ et

$F2/F3$. Nous observons que la région correspondant à la voyelle /a/ dans les deux autres bases de données n'est pas occupée. Nous utiliserons la moyenne des voyelles /ɑ/ et /æ/ de chaque locuteur pour estimer la longueur du conduit vocal correspondant.

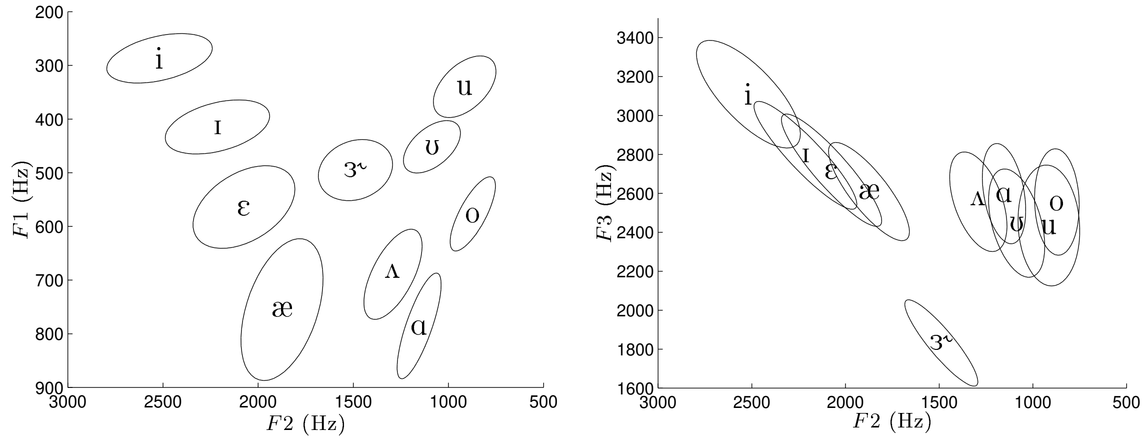

FIG. 3.16: Base de données 3.

3.6.4 Procédure d'évaluation à utiliser

Sur chaque base de données, nous avons effectué une normalisation à partir de chacun des points suivants :

1. Estimation des paramètres d_o, d_p avec tous les formants des voyelles extrêmes[7].
2. Estimation des paramètres d_o, d_p avec les formants des voyelles /i/, /a/ et /u/ qui présentent le plus de variance entre hommes et femmes.
3. Idem à ci-dessus, en utilisant en plus une relation linéaire entre les paramètres d_o, d_p. Ceci revient à utiliser un seul paramètre d'estimation de la longueur du conduit vocal.

Les normalisations qui en résultent ont été évaluées selon les critères cités dans la section 3.6 (page 36).

3.7 Résultats de l'évaluation

Cette section présente les résultats obtenus des normalisations effectuées sur les trois bases de données. Pour l'obtention des paramètres d_o, d_p, nous avons suivi les étapes décrites sur la page 27.

[7] /i/, /a/, /y/ et /u/ pour les bases de données 1 et 2 ; /i/, /a/ et /u/ pour la 3.

3.7.1 Estimation des paramètres avec tous les formants des voyelles extrêmes

Les paramètres d_o, d_p correspondant à la base de données 2 ont été présentés dans la section 3.3 (figure 3.6). Les figures 3.17 et 3.18 présentent les valeurs des paramètres correspondant aux bases de données 1 et 3 respectivement.

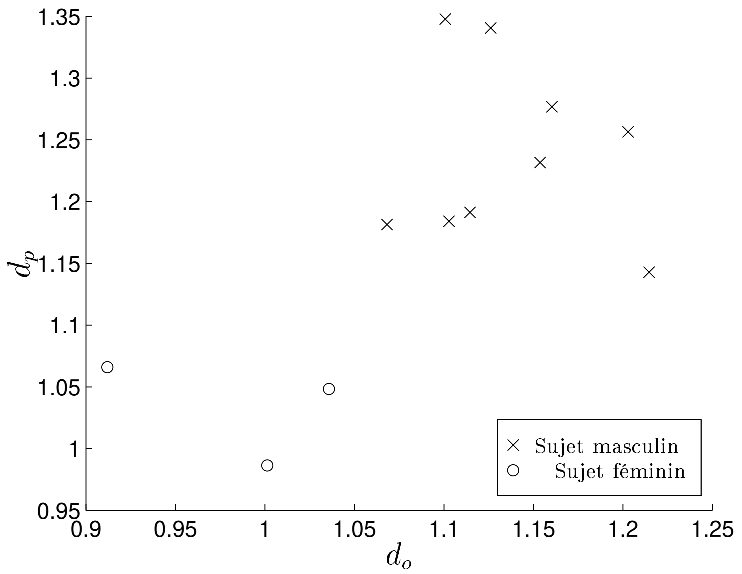

FIG. 3.17: Valeurs des paramètres d_o, d_p de la base de données 1.

Nous observons que le groupe masculin est bien séparé du féminin et en plus que les valeurs des paramètres des femmes correspondent à un conduit vocal plus court que ceux qui correspondent aux conduits des hommes, et cela pour les deux bases de données. Ceci est en concordance avec ce que nous avons trouvé dans la section 3.3 pour la base de données 2.

Les ellipses de dispersion pour la base de données 2 ont été présentées dans la section 3.5 (figure 3.12). Les figures 3.19 et 3.20 présentent les ellipses des bases de données 1 et 3 respectivement.

Nous observons déjà, en comparant les ellipses de dispersion normalisées (figures 3.19, 3.12 et 3.20) avec celles des bases de données originales (figures 3.14, 3.11 et 3.16) que la méthode atteint un assez bon degré de normalisation.

Nous procédons maintenant à l'évaluation du degré de réduction obtenu (en concordance avec les critères de Ferrari-Disner (1980)). Les tableaux 3.5, 3.6 et 3.7 présentent, pour chaque voyelle de chaque base de données, le volume de l'ellipsoïde (en Bark[3]) correspondant avant et après normalisation, et la somme des volumes des ellipsoïdes. La

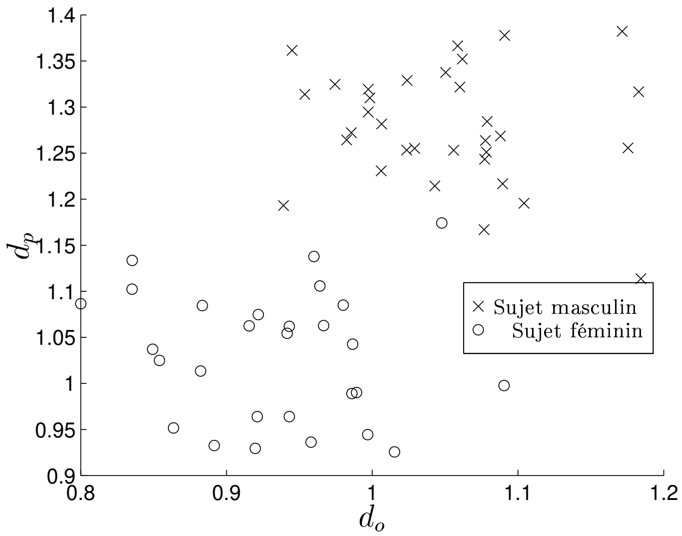

FIG. 3.18: Valeurs des paramètres d_o, d_p de la base de données 3.

troisième colonne présente le rapport (en pourcentage) du volume de l'ellipsoïde normalisé par rapport a celui du non normalisé.

phonème	volume non normalisé	volume normalisé	rapport
i	0.0552	0.0032	5.86 %
a	0.2787	0.0683	24.51 %
y	0.0472	0.0255	54.12 %
u	0.0856	0.0271	31.70 %
e	0.1608	0.0709	44.09 %
ε	0.1808	0.0845	46.75 %
ø	0.2585	0.0971	37.56 %
somme	1.0668	0.3767	35.31 %

TAB. 3.5: Volumes des ellipsoïdes avec leur rapport, base de données 1.

3.7.2 Normalisation avec des paramètres estimés à partir de formants sélectionnés

Nous allons à présent faire la normalisation en estimant d'abord les paramètres d_o, d_p en ne prenant en compte que les formants des voyelles /i/, /a/, et /u/ qui présentent le plus de variance entre les hommes et les femmes.

Les figures 3.21, 3.22 et 3.23 présentent les valeurs des formants pour chacune des voyelles à considérer. Pour faciliter la visualisation, les locuteurs ont été ordonnés par

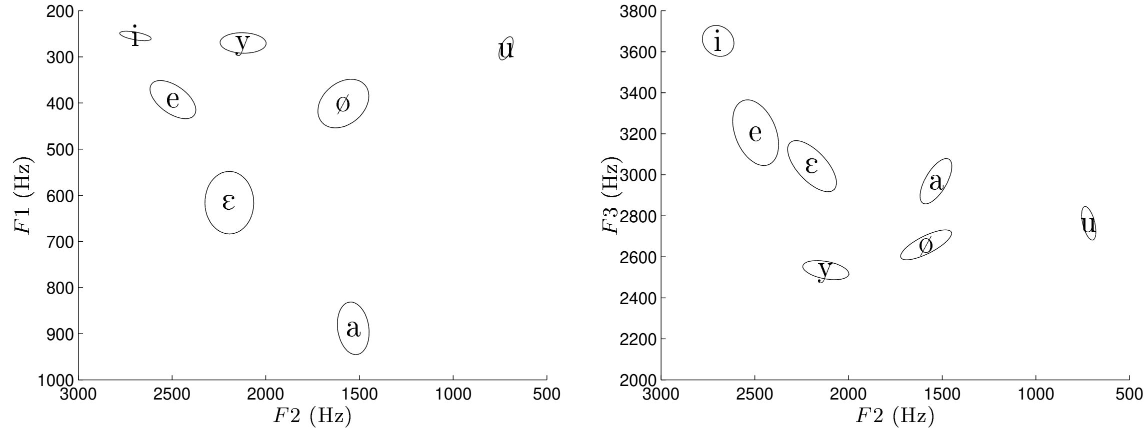

FIG. 3.19: Base de données 1 normalisée.

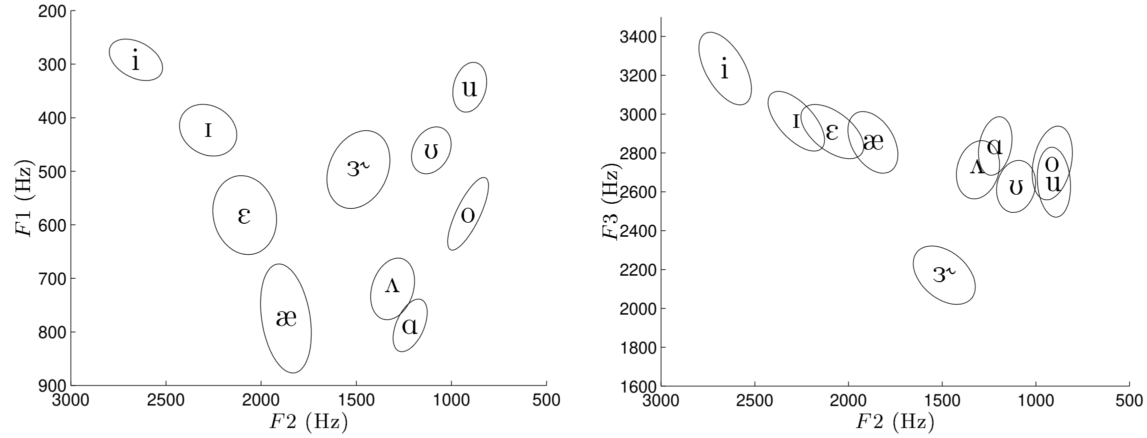

FIG. 3.20: Base de données 3 normalisée.

phonème	volume non normalisé	volume normalisé	rapport
o	0.1508	0.0554	36.73 %
u	0.1208	0.0258	21.36 %
ɛ	0.0686	0.0380	55.42 %
y	0.0561	0.0473	84.27 %
a	0.2762	0.0581	21.02 %
ɔ	0.2720	0.0956	35.12 %
ø	0.2473	0.0584	23.62 %
e	0.1100	0.0436	39.61 %
i	0.0559	0.0201	35.87 %
œ	0.2047	0.0633	30.90 %
somme	1.5625	0.5055	32.35 %

TAB. 3.6: Volumes des ellipsoïdes avec leur rapport base de données 2.

phonème	volume non normalisé	volume normalisé	rapport
i	0.2479	0.0801	32.31 %
I	0.2599	0.1224	47.10 %
i	0.4328	0.2206	50.97 %
æ	0.6750	0.2662	39.43 %
ʌ	0.7265	0.2740	37.71 %
ɑ	0.5365	0.1478	27.55 %
o	0.7661	0.3171	41.39 %
ʊ	0.7141	0.2489	34.85 %
u	1.1193	0.3360	30.02 %
ɝ	0.5190	0.5628	108.44 %
somme	5.9971	2.5759	42.95 %

TAB. 3.7: Volumes des ellipsoïdes avec leur rapport base de données 3.

ordre croissant du troisième formant de chaque voyelle présentée. Les trois lignes continues correspondent à chacun des formants des locuteurs masculins. Les lignes en pointillé correspondent à leur tour aux trois formants des locuteurs féminins.

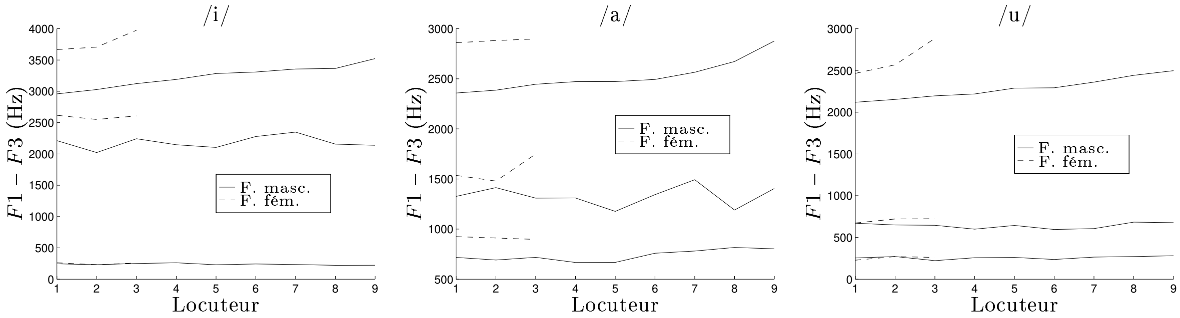

FIG. 3.21: Formants des voyelles, base de données 1.

Les figures montrent une différence significative entre les deux groupes pour le deuxième et troisième formant du /i/, les trois formants du /a/ et juste le troisième formant du /u/. Nous n'avons donc utilisé que ces formants pour estimer la longueur du conduit vocal de chacun des locuteurs[8]. Les figures 3.24, 3.25 et 3.26 montrent les valeurs des paramètres d_o, d_p obtenues pour les trois bases de donnés, en n'utilisant que ces formants.

Les figures montrent que, là encore, les valeurs des paramètres d_o, d_p correspondant aux hommes sont bien séparés de celles correspondant aux femmes, et cela pour les trois bases de données. Nous procédons maintenant à la normalisation des formants en utilisant la méthode décrite dans la page 36. Les figures 3.27, 3.28 et 3.29 montrent les résultats obtenus sur les trois bases de données.

Les figures montrent des réductions de volume comparables à celles obtenues avec tous

[8]Le grille en $\mathbb{R}^{12}$ est maintenant un grille en $\mathbb{R}^5$

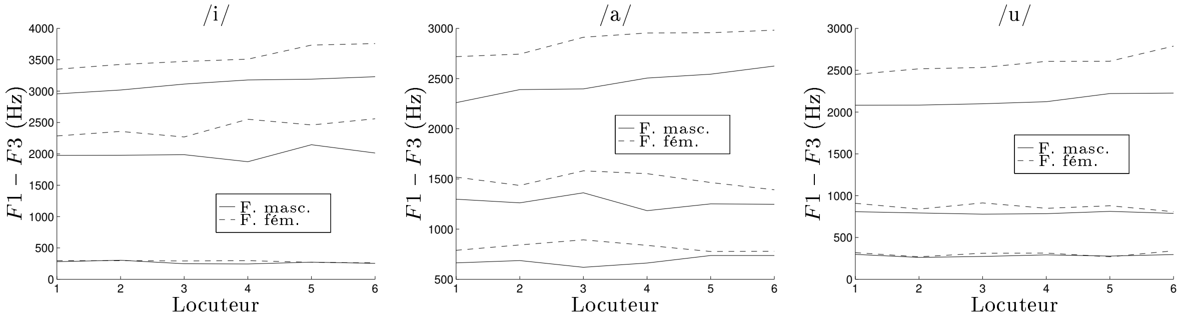

FIG. 3.22: Formants des voyelles, base de données 2.

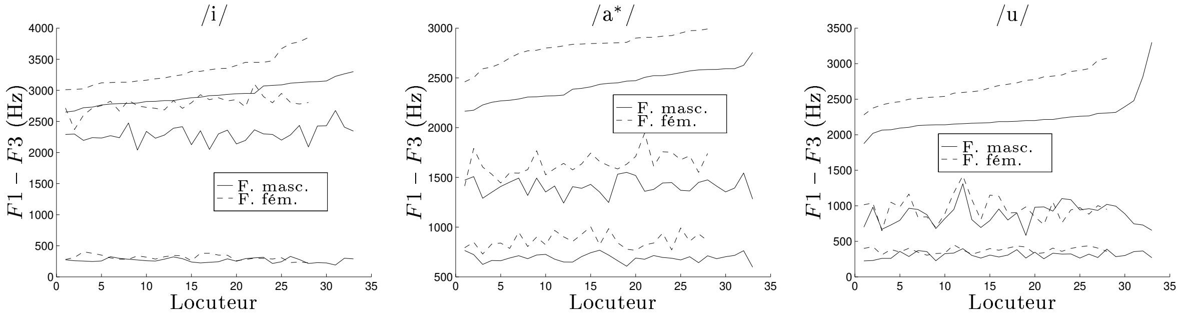

FIG. 3.23: Formants des voyelles, base de données 3.

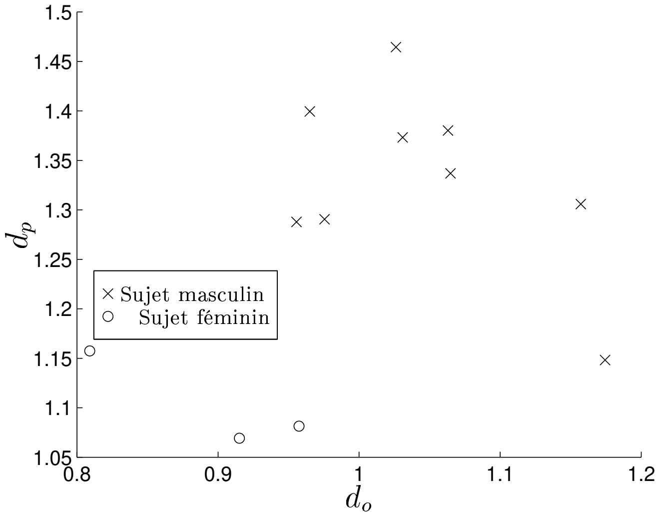

FIG. 3.24: Valeurs des paramètres d_o, d_p de la base de données 1, formants sélectionnés.

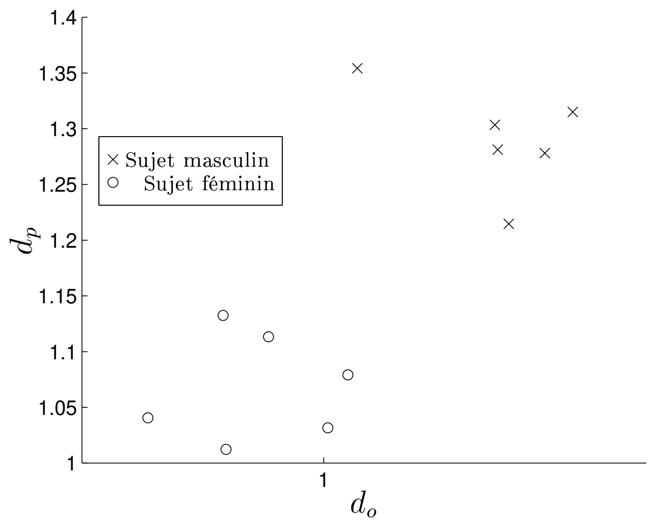

FIG. 3.25: Valeurs des paramètres d_o, d_p de la base de données 2, formants sélectionnés.

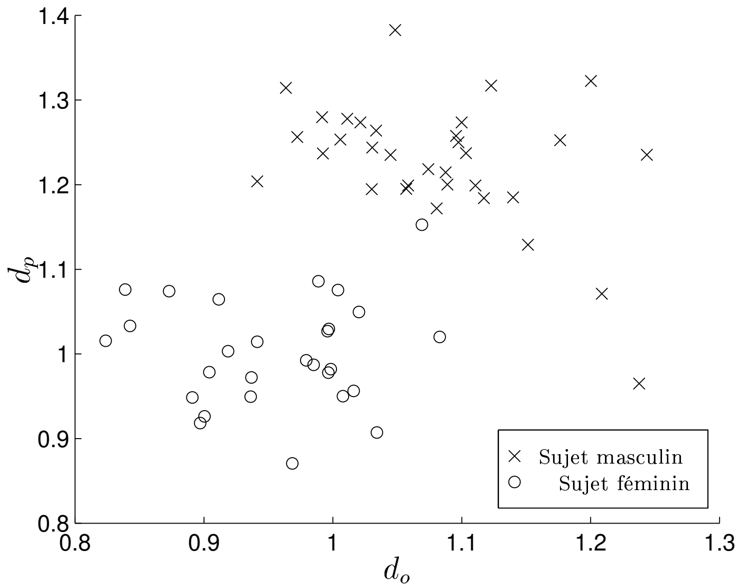

FIG. 3.26: Valeurs des paramètres d_o, d_p de la base de données 3, formants sélectionnés.

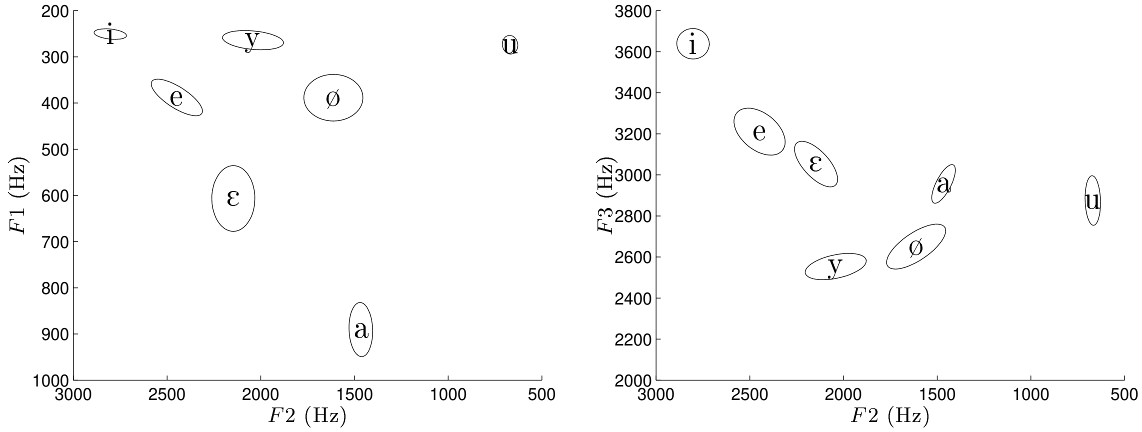

FIG. 3.27: Base de données 1 normalisée, formants sélectionnés.

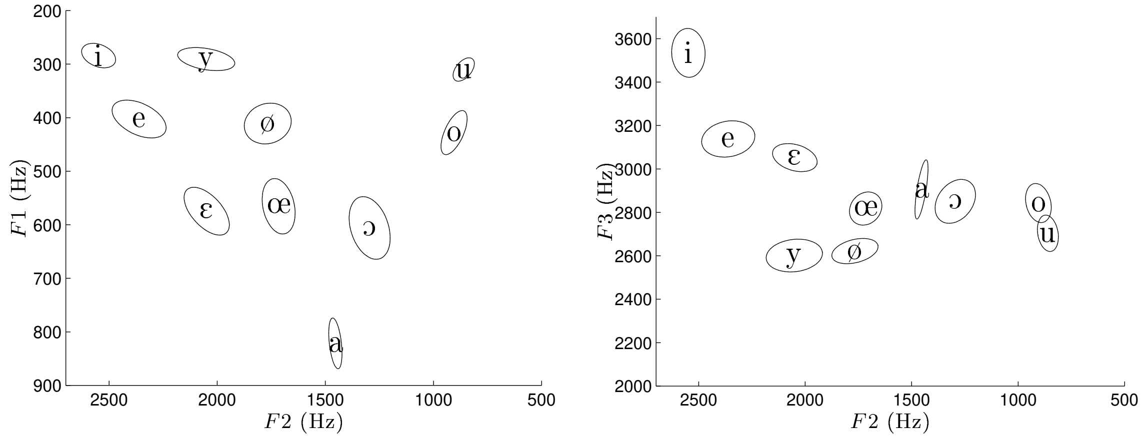

FIG. 3.28: Base de données 2 normalisée, formants sélectionnés.

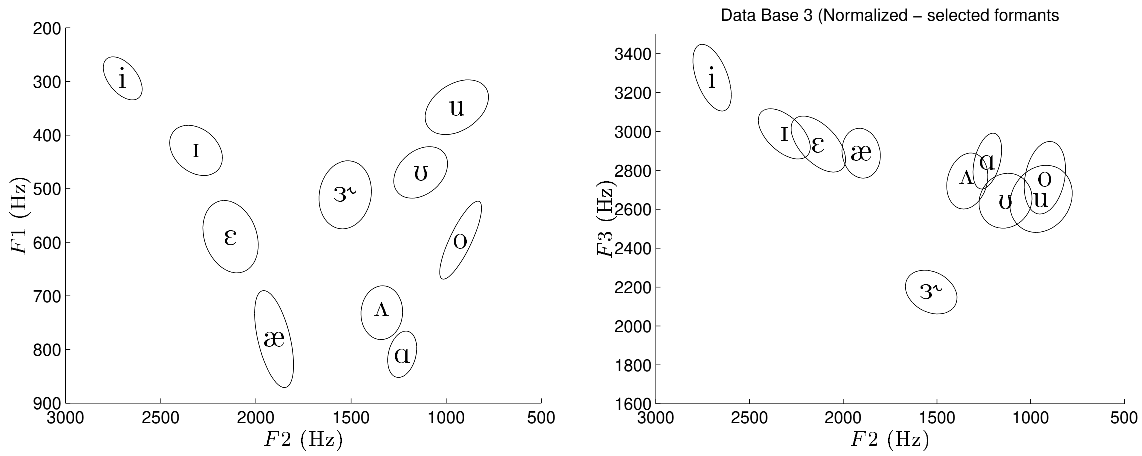

FIG. 3.29: Base de données 3 normalisée, formants sélectionnés.

les formants des voyelles extrêmes. Ce fait est illustré également sur les tableaux 3.8, 3.9 et 3.10 pour les bases de données 1, 2 et 3 respectivement.

phonème	volume non normalisé	volume normalisé	rapport
i	0.0552	0.0040	7.30 %
a	0.2787	0.0470	16.85 %
y	0.0472	0.0351	74.46 %
u	0.0856	0.0454	53.05 %
e	0.1608	0.0424	26.39 %
ɛ	0.1808	0.0899	49.74 %
ø	0.2585	0.1956	75.66 %
somme	1.0668	0.4594	43.07 %

TAB. 3.8: Volumes des ellipsoïdes de la base de données 1 avec leur rapport, formants sélectionnés.

phonème	volume non normalisé	volume normalisé	rapport
o	0.1508	0.0689	45.66 %
u	0.1208	0.0402	33.29 %
ɛ	0.0686	0.0353	51.53 %
y	0.0561	0.0438	78.05 %
a	0.2762	0.0266	9.62 %
ɔ	0.2720	0.1293	47.53 %
ø	0.2473	0.0527	21.32 %
e	0.1100	0.0403	36.61 %
i	0.0559	0.0097	17.43 %
œ	0.2047	0.0676	32.99 %
somme	1.5625	0.5144	32.92 %

TAB. 3.9: Volumes des ellipsoïdes de la base de données 2 avec leur rapport, formants sélectionnés.

3.7.3 Normalisation avec des paramètres d_o, d_p linéairement dépendants

Nous allons estimer à présent les paramètres d_o, d_p de chaque locuteur en utilisant la relation $d_p = a \times d_o + b$. Les variables a, b sont obtenues à partir d'une régression linéaire sur les valeurs d_o, d_p de la section précédente (voir les figures 3.24, 3.25, et 3.26). Ceci équivaut à utiliser un seul paramètre pour l'estimation de la longueur du conduit vocal. La grille en $\mathbb{R}^5$ devient alors une ligne de points (toujours en $\mathbb{R}^5$) sur laquelle nous projetons, pour chaque locuteur, les points composés des valeurs de formants sélectionnés correspondantes. Les figures 3.30, 3.31, et 3.32 montrent les valeurs des paramètres obtenues.

Nous montrons ensuite les ellipses de dispersion obtenues en utilisant les valeurs des paramètres ci-dessus pour les trois bases de données. Nous observons là encore (figures 3.33,

phonème	volume non normalisé	volume normalisé	rapport
i	0.2479	0.0516	20.83 %
ɪ	0.2599	0.0966	37.18 %
e	0.4328	0.1725	39.86 %
æ	0.6750	0.1391	20.61 %
ʌ	0.7265	0.2191	30.17 %
ɑ	0.5365	0.1030	19.20 %
o	0.7661	0.3103	40.51 %
ʊ	0.7141	0.3643	51.02 %
u	1.1193	0.5941	53.08 %
ɝ	0.5190	0.3236	62.34 %
somme	5.9971	2.3744	39.59 %

TAB. 3.10: Volumes des ellipsoïdes de la base de données 3 avec leur rapport, formants sélectionnés.

3.34 et 3.35) des réductions de volume comparables à celles obtenues avec les deux autres méthodes d'estimation des paramètres d_o, d_p. Les tableaux 3.11, 3.12 et 3.13 reflètent ce fait.

phonème	volume non normalisé	volume normalisé	rapport
i	0.0552	0.0149	26.97 %
a	0.2787	0.0682	24.48 %
y	0.0472	0.0982	208.09 %
u	0.0856	0.0461	53.89 %
e	0.1608	0.0649	40.39 %
ɛ	0.1808	0.0829	45.88 %
ø	0.2585	0.2304	89.12 %
somme	1.0668	0.6057	56.78 %

TAB. 3.11: Volumes des ellipsoïdes de la base de données 1 avec leur rapport, réduction à un seul paramètre.

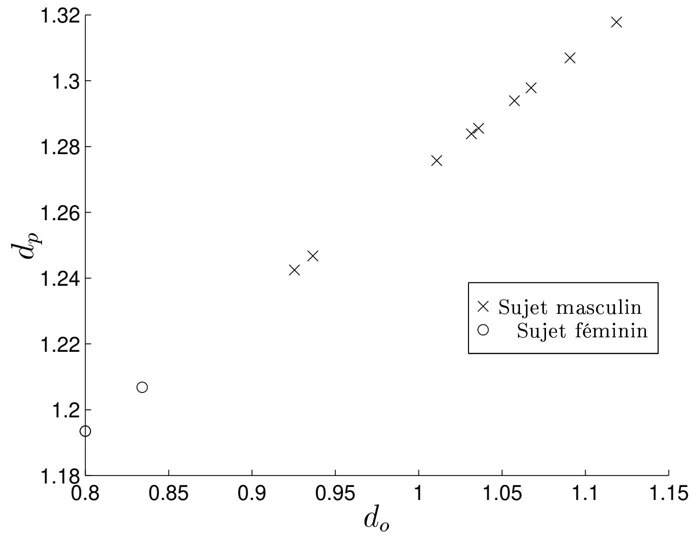

FIG. 3.30: Valeurs des paramètres d_o, d_p de la base de données 1, réduction à un seul paramètre.

3.8 Conclusions

Nous avons étudié, dans ce chapitre, les relations existant entre la géométrie du conduit vocal d'une part et les formants des voyelles d'autre part. Avec le concept de distorsion longitudinal, nous avons trouvé une façon de modifier le conduit vocal du robot pour faire en sorte que le robot parlant puisse suivre les formants présentés à l'entrée du système par un sujet quelconque.

En exploitant le concept de fibre articulatoire, nous avons démontré que les différentes réalisations d'un même phonème par plusieurs locuteurs peuvent être ramenées à un volume de grandeur comparable à celui des locuteurs d'origine.

La méthode utilisée pour normaliser l'espace vocalique d'une base de données s'est avérée robuste du point de vue de la méthode d'estimation de la longueur du conduit vocal. Certaines objections à l'utilisation des voyelles extrêmes pour la normalisation peuvent néanmoins être faites, comme celles suggérées par Verbrugge et al. (1976). Dans cette étude, les auteurs vérifient comment le taux d'identification des voyelles dans les contextes /h_d/ et /p_p/ varie selon les cas suivants : (1) les auditeurs n'ont pas de connaissance du locuteur qui prononce la voyelle et (2) les auditeurs entendent préalablement les voyelles /i/, /a/ et /u/ du locuteur. Aucune différence significative n'a été trouvée entre les taux de reconnaissance dans les deux cas.

D'après les résultats de cette expérience, on est forcé de conclure qu'aucune information concernant les points extrêmes de l'espace acoustique du locuteur n'est nécessaire pour la classification correcte de ses voyelles. Nous estimons cependant que ce genre d'expérience

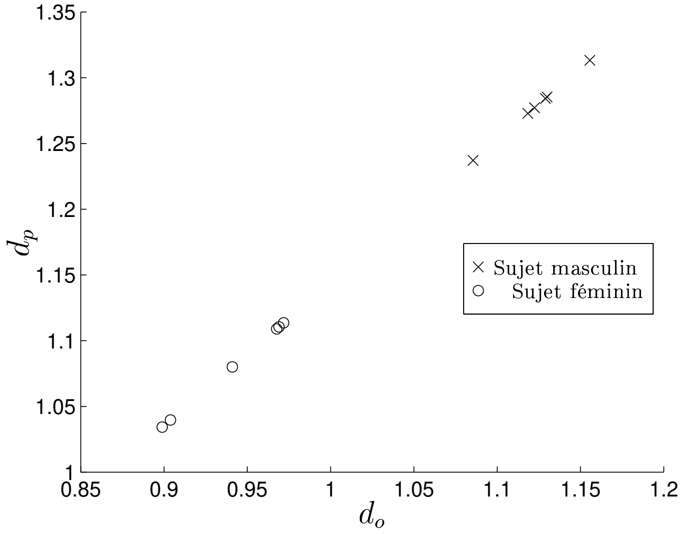

FIG. 3.31: Valeurs des paramètres d_o, d_p de la base de données 2, réduction à un seul paramètre.

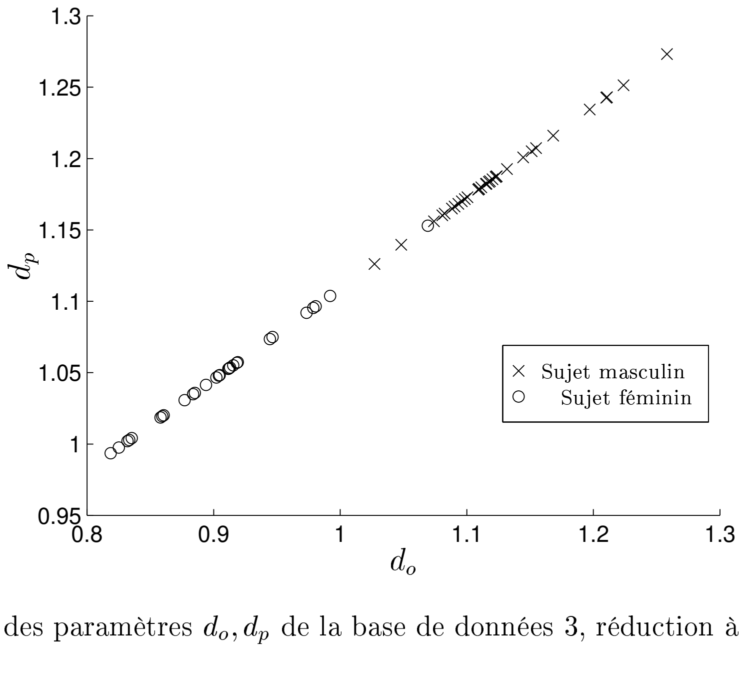

FIG. 3.32: Valeurs des paramètres d_o, d_p de la base de données 3, réduction à un seul paramètre.

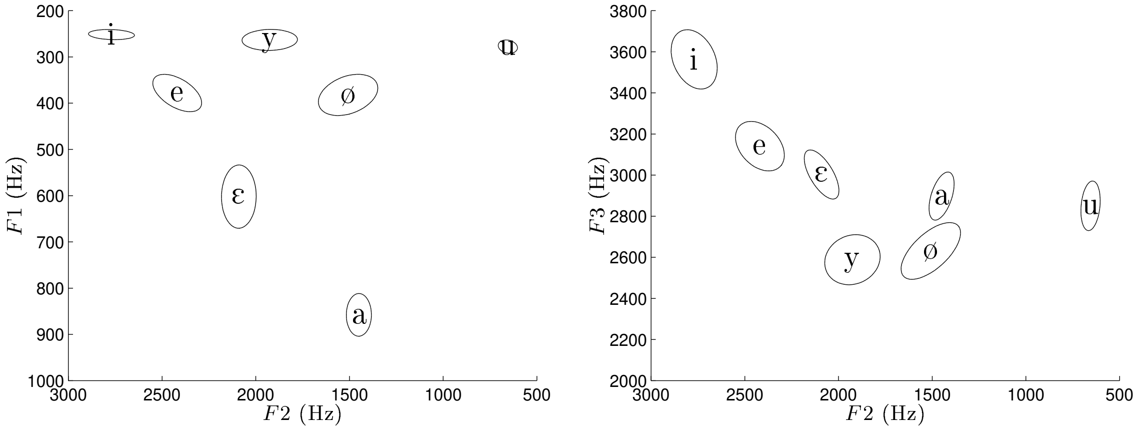

FIG. 3.33: Base de données 1 normalisée, un seul paramètre.

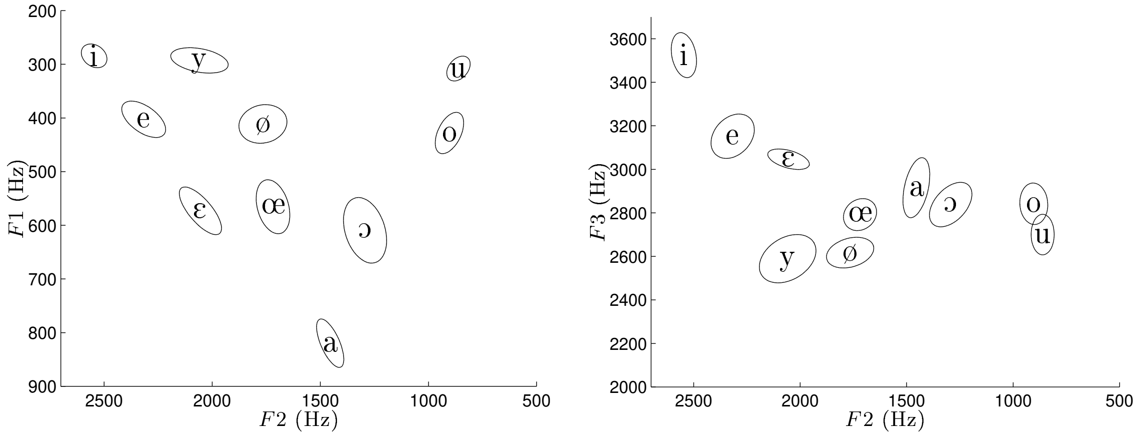

FIG. 3.34: Base de données 2 normalisée, un seul paramètre.

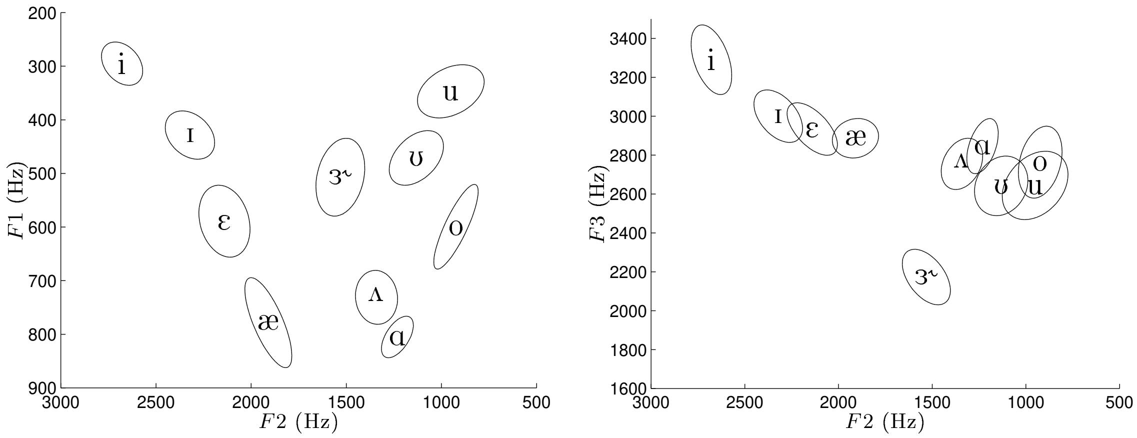

FIG. 3.35: Base de données 3 normalisée, un seul paramètre.

phonème	volume non normalisé	volume normalisé	rapport
o	0.1508	0.0834	55.30 %
u	0.1208	0.0557	46.14 %
ɛ	0.0686	0.0177	25.84 %
y	0.0561	0.0683	121.65 %
a	0.2762	0.0478	17.30 %
ɔ	0.2720	0.1237	45.49 %
ø	0.2473	0.0640	25.89 %
e	0.1100	0.0363	33.03 %
i	0.0559	0.0106	18.88 %
œ	0.2047	0.0657	32.11 %
somme	1.5625	0.5733	36.69 %

TAB. 3.12: Volumes des ellipsoïdes de la base de données 2 avec leur rapport, réduction à un seul paramètre.

perceptive n'invalide pas l'approche décrite dans ce chapitre parce que, d'une part, le fait l'utilité de la connaissance préalable des voyelle extrêmes pour la normalisation du locuteur chez l'auditeur n'est pas décelée n'implique pas que ces voyelles extrêmes ne sont pas porteuses d'informations d'ordre anatomiques. Verbrugge et al. (1976) suggèrent même que le fait de présenter les voyelle dans un contexte fixe peut biaiser les résultats. Les informations portées par les consonnes peuvent donner des informations suffisantes sur les caractéristiques particulières du locuteur. D'autre part, nous ne prétendons aucunement que la technique décrite dans ce chapitre soit celle que les êtres humains utilisent en pratique pour effectuer la normalisation des locuteurs. Notre seul but était de trouver une méthode pratique d'adaptation d'un modèle articulatoire à un locuteur donné sans avoir recours à des mesures anatomiques de celui-ci.

En conclusion, nous avons à présent une façon de représenter un phonème (ou, plus précisément, une configuration formantique) quelconque réalisé par un locuteur quelconque dans l'espace des phonèmes (ou l'espace formantique) du robot parlant. Ceci s'avérera crucial lors de la mise en œuvre du contrôleur.

phonème	volume non normalisé	volume normalisé	rapport
i	0.2479	0.0656	26.45 %
I	0.2599	0.0947	36.44 %
e	0.4328	0.1436	33.19 %
æ	0.6750	0.1119	16.58 %
ʌ	0.7265	0.1972	27.15 %
ɑ	0.5365	0.0654	12.20 %
o	0.7661	0.3201	41.78 %
ʊ	0.7141	0.4077	57.10 %
u	1.1193	0.5949	53.14 %
ɝ	0.5190	0.4281	82.49 %
somme	5.9971	2.4294	40.51 %

TAB. 3.13: Volumes des ellipsoïdes de la base de données 3 avec leur rapport, réduction à un seul paramètre.

Chapitre 4

Inversion de Séquences Voyelle-Plosive

Le but de chacun des travaux entrepris jusqu'à présent a été de construire les bases d'un modèle articulatoire capable de suivre des trajectoires formantiques présentées au modèle de contrôle par des locuteurs quelconques. Ce chapitre est consacré à l'étude qui a été entreprise en vue de l'obtention des paramètres articulatoires qui permettront au modèle de suivre ces trajectoires. Les résultats du chapitre précédant nous permettent d'ajuster la longueur du conduit vocal de façon à ce que le robot soit capable de suivre ces trajectoire d'une manière cohérente.

4.1 Introduction

Le but ultime de notre travail consiste à mettre en œuvre le contrôleur décrit par la figure 4.1 (Laboissière 1992) de façon à le rendre capable de suivre des trajectoires voyelle-consonne-voyelle. Dans une étude similaire, Bailly (1996) se consacra à ce même problème en utilisant un modèle articulatoire similaire au nôtre (Badin, Gabioud, Beautemps, Lallouache, Bailly, Maeda, Zerling & Brock 1995) et un espace sensori-moteur séparé en trois sous-espaces (articulatoire, géométrique et acoustique). Ce modèle, bien qu'il donne des résultats satisfaisants, n'a pas été testé pour produire des sorties correspondant à des locuteurs différents de celui à partir duquel le modèle a été obtenu.

Lors de notre travail, nous nous sommes intéressés à la transition voyelle-consonne de ces séquences. Nous centrerons notre étude sur des indices acoustiques, et plus précisément formantiques, dont le contrôleur pourrait se servir afin d'inférer des commandes articulatoires compatibles avec de tels indices. Nous chercherons notamment à retrouver le lieu d'articulation des occlusives à partir des transitions formantiques d'un signal de parole, en nous basant sur les transitions que notre modèle articulatoire est capable de produire.

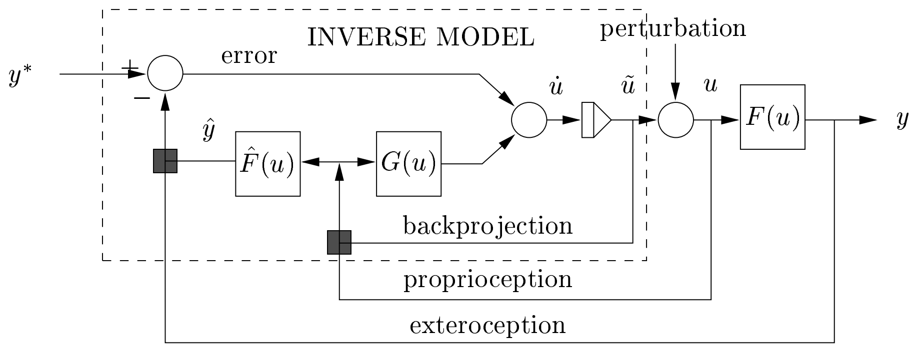

FIG. 4.1: Architecture du contrôleur articulatoire (d'après laboissiere-galvan/icph95).

Plusieurs études ont été menées en vue de trouver des invariants articulatoires, et plus précisément le lieu de constriction des plosives, à partir d'un signal acoustique. Une approche souvent utilisée est celle des équations du locus (« locus equations » en anglais). Öhman (1966) a étudié les transitions formantiques produites par des locuteurs prononçant des séquences voyelle-plosive-voyelle. Il n'a pas réussi à en déduire de locus invariant de F2 ou F3 et à partir de ses interprétations conclut qu'une réévaluation de la théorie du locus était nécessaire.

Kewley-Port (1982) a réalisé des analyses spectrales pour essayer de trouver des invariants acoustiques pour le lieu d'articulation à partir de séquences du type plosive-voyelle. Le signal acoustique a été analysé par prédiction linéaire (LPC) pour obtenir des mesures temporelles et spectrales détaillées. Pour un seul locuteur prononçant les consonnes /b,d,g/ suivies par huit voyelles, les valeurs de F2 au moment du relâchement de l'occlusion se sont avérées utiles pour séparer les trois consonnes lorsqu'elles étaient suivies par toutes les voyelles sauf /i/ et /a/. Les mesures de F3 n'ont été utiles que lorsque la consonne était suivie par la voyelle /æ/. On y conclut que les valeurs de F2 ne sont pas suffisantes pour trouver le lieu de constrictions dans les séquences étudiées.

Plus récemment, Sussman, McCaffrey & Mattews (1991) ont réétudié les équations du locus avec une base de données de 20 locuteurs (10 masculins et 10 féminins). Avec une analyse discriminante utilisant les valeurs du F2, des taux acceptables d'identification du lieu d'articulation ont été obtenus (84%, 81% et 69% pour les /b/, /d/ et /g/ des locuteurs et 82%, 78% et 67% pour ceux des locutrices). La relation (fortement linéaire) entre les valeurs du F2 au moment de l'ouverture et celle correspondant à la voyelle qui suit a encore été mise en évidence par Sussman, Hoemeke & Ahmed (1993) dans une étude incluant plusieurs langues.

Une autre approche souvent utilisée est celle de l'étude du spectre du bruit d'explosion (« burst » en anglais) qui suit l'ouverture d'une occlusion. Dans l'étude par Stevens & Blumstein (1978), on présente une interprétation théorique des caractéristiques acoustiques des occlusives qui correspond à des données réelles et qui réussit relativement bien à caractériser le lieu d'articulation à partir du spectre du bruit d'explosion. Seul le spectre du bruit d'explosion y est utilisé pour caractériser le lieu d'articulation. Ces résultats sont en accord avec la théorie selon laquelle les caractéristiques spectrales des premières 10 à 20 ms après l'ouverture d'une constriction suffiraient à caractériser le lieu d'articulation. Cette théorie est encore renforcée par les résultats des études de Blumstein & Stevens (1979) et Blumstein & Stevens (1980).

Des résultats provenant des études ultérieures par Blumstein, Isaacs & Mertus (1982) suggèrent, néanmoins, que la forme générale du spectre du bruit d'explosion pourrait ne pas être à l'origine des attributs caractérisant le lieu d'articulation. D'ailleurs, la métrique utilisant cette forme générale n'étant pas utile pour classifier des plosives des langues telles que le malayam et le français (Lahiri & Blumstein 1981), elle a dû être modifiée afin de la généraliser aux langues étudiées. Lahiri, Gewirth & Blumstein (1984) ont mis en œuvre une méthode pour séparer les différents lieux d'articulation, méthode basée sur le changement relatif de la distribution d'énergie entre le bruit d'explosion et le début de la voyelle (« voicing onset »). Eek & Meister (1995) étudient les équations du locus et, dans une approche similaire à celle de Lahiri et al. (1984), se servent du plus fort sommet du spectre au moment de l'ouverture et du « centre de gravité » (Schwartz, Boë, Vallée & Abry 1997) de la voyelle qui suit pour obtenir une séparation satisfaisante des différents lieux d'articulation.

4.2 La base de données

La base de données que nous avons utilisée est composée de trois réalisations de chacune des combinaisons du type $V_1CV_2CV_1ZE$ possibles avec C= /b/, /d/, /g/ et $V_{1,2}$= /a/,/i/,/y/,/u/ (e.g., ABABAZE-1, ABABAZE-2, ABAZE-3, ABIBAZE-1, ABIBAZE-2, etc.). Étant donné le temps de calcul que l'analyse de chaque séquence implique (cf. sections suivantes) nous n'utiliserons qu'une réalisation de chaque type voyelle-consonne possible, soit quatre voyelles $\times$ trois consonnes = 12 séquences.

Piquemal, Schwartz, Berthommier, Lallouache & Escudier (1996) ont utilisé cette base de données lors d'une étude visant à inférer des indices de détection du bruit d'explosion généré lors de la production d'occlusives. Après observation des signaux, nous avons décidé d'utiliser le signal à partir de 120 ms avant le premier bruit d'explosion ayant été marqué par cette étude (i.e., jusqu'au début de la deuxième voyelle).

Nous avons utilisé la première paire voyelle-consonne de la première séquence symétrique de chaque type. Une autre séquence symétrique était utilisée lorsque l'étiquette n'était pas considérée comme bien placée ou bien si les formants ne pouvaient pas être obtenus de façon non ambiguë. Aucune séquence symétrique correspondant à « AG » et satisfaisant ces critères n'ayant pu être trouvée, nous avons alors utilisé « AGIGAZE-1 ». Voici les séquences que nous avons finalement utilisées :

ABABAZE-1	OUBOUBOUZE-3
ADADAZE-1	OUDOUDOUZE-3
AGIGAZE-1	OUGOUGOUZE-3
IBIBIZE-1	UBUBUZE-2
IDIDIZE-1	UDUDUZE-2
IGIGIZE-1	UGUGUZE-1

Le signal a d'abord été préaccentué par un filtre de fonction de transfert $H(z) = 1 - 0.9z^{-1}$ (Calliope 1989). Quatre formants ont ensuite été obtenus à l'aide d'un traitement LPC avec 24 pôles. Une fenêtre de 256 échantillons (16 ms, la fréquence d'échantillonnage étant de 16 kHz) décalée de seize échantillons (1 ms) à chaque itération a été utilisée. Nous avons alors une séquence de formants composée de trames séparées d'une milliseconde.

La figure 4.2 présente les formants que nous avons utilisés (les lignes verticales correspondent aux étiquettes désignant les différentes parties des signaux, cf. section 4.4, page 60 pour une description détaillée).

4.3 Application au Modèle de Contrôle

Notre modèle articulatoire ayant déjà été testé pour produire des voyelles de plusieurs locuteurs de façon satisfaisante (chapitre précédent) et le modèle de contrôle de la figure 4.1 pour des trajectoires voyelle-voyelle (Laboissière 1992) nous utiliserons ce dernier pour la partie vocalique des séquences voyelle-consonne.

En ce qui concerne l'occlusion, nous étudierons les résultats acoustiques de chaque constriction que le modèle est capable de produire (cf. chapitre 2) afin d'obtenir une mesure probabiliste reliée à chacune de ces constrictions.

4.4 Inversion des séquences

Bien que la plupart des études dans la littérature utilisent des indices à partir de la consonne et jusqu'à la voyelle, les trajectoires formantiques à partir de la voyelle et jusqu'à

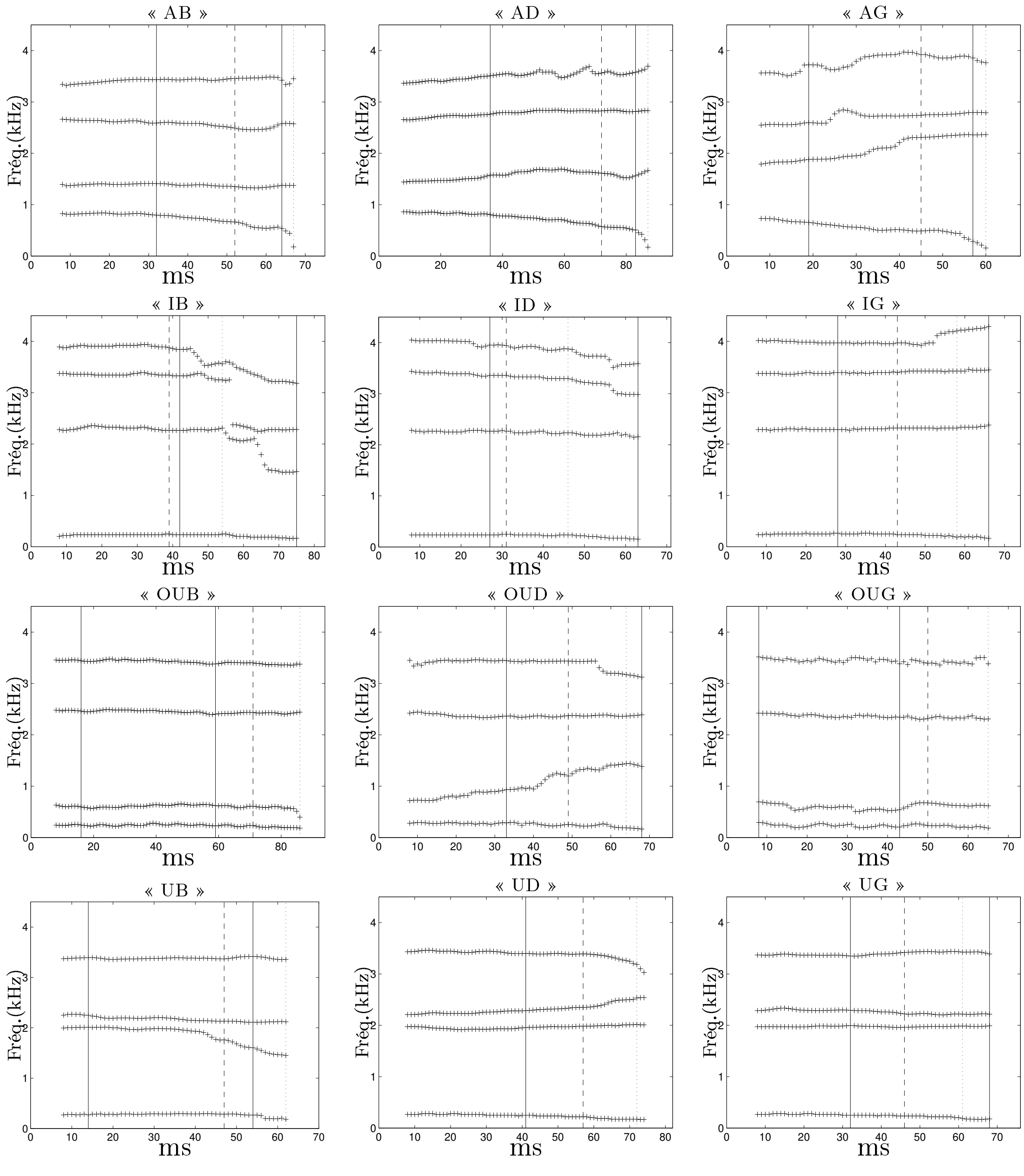

Fig. 4.2: Formants de la base de données.

la constriction doivent leur être semblables (Stevens & Blumstein 1978)[1]. Nous avons vérifié ce fait avec les séquences symétriques ($V_1 = V_2$) de la base de données. La figure 4.3 montre les formants obtenus pour ABA, ADA et AGA. Observons que les formants avant et après la fermeture de la constriction sont assez symétriques, les différences observées étant certainement dûes en partie au phénomène de réduction vocalique (après la deuxième voyelle, il reste une deuxième constriction à réaliser).

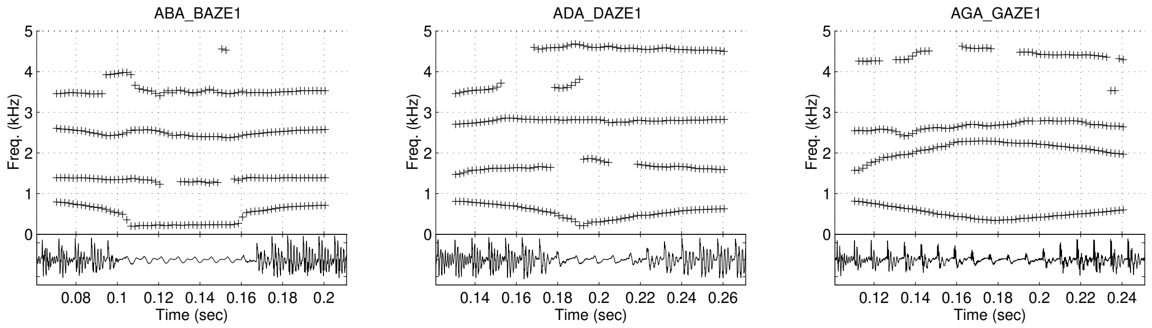

FIG. 4.3: Formants des séquences symétriques.

Afin de réaliser un étiquetage de chacun des signaux que nous nous proposons d'inverser, une analyse de l'énergie à chaque trame de formants a été effectuée. Les figures 4.4 et 4.5 montrent le résultat de cette analyse, l'énergie du signal (en haut) synchronisée au signal lui même. Afin d'automatiser l'étiquetage des signaux, la courbe d'énergie a été approximée par une courbe polynomiale de troisième ordre. À l'aide de cette information et celle concernant les formants nous avons étiqueté le signal aux endroits suivants :

- E_m : Maximum de la courbe d'énergie, ligne solide gauche sur les figures.
- E_x : Point d'inflexion de la courbe d'énergie, ligne solide droite sur les figures.
- $F1_x$: Endroit où $F1$ descend au dessous de 190 Hz, en pointillé sur les figures[2].
- $F1_m$: $F1_x{-}15$ ms, en traits sur les figures.

4.4.1 Inversion de la partie vocalique

L'inversion de séquences vocaliques a déjà été mise en œuvre à l'aide d'un modèle direct (« modele forward » en anglais) (Jordan & Rumelhart 1992) par Laboissière (1992). Une procédure de ce type impliquerait dans notre cas l'implémentation d'un tel modèle pour chaque ensemble de paramètres de distortion (cf. chapitre 3). Nous avons donc choisi de

[1]Le modèle de contrôle étant en mesure de nous fournir une configuaration articulatoire correspondant à une voyelle, nous chercherons à partir de cette configuration-là l'occlusion produisant les formants désirés.

[2]Lors d'une occlusion dans le conduit vocal, la valeur du premier formant descend environ à 190 Hz mais jamais au-dessous de cette valeur, dû aux phénomènes de vibration des parois (Fant 1969, Calliope 1989, chapitre 3).

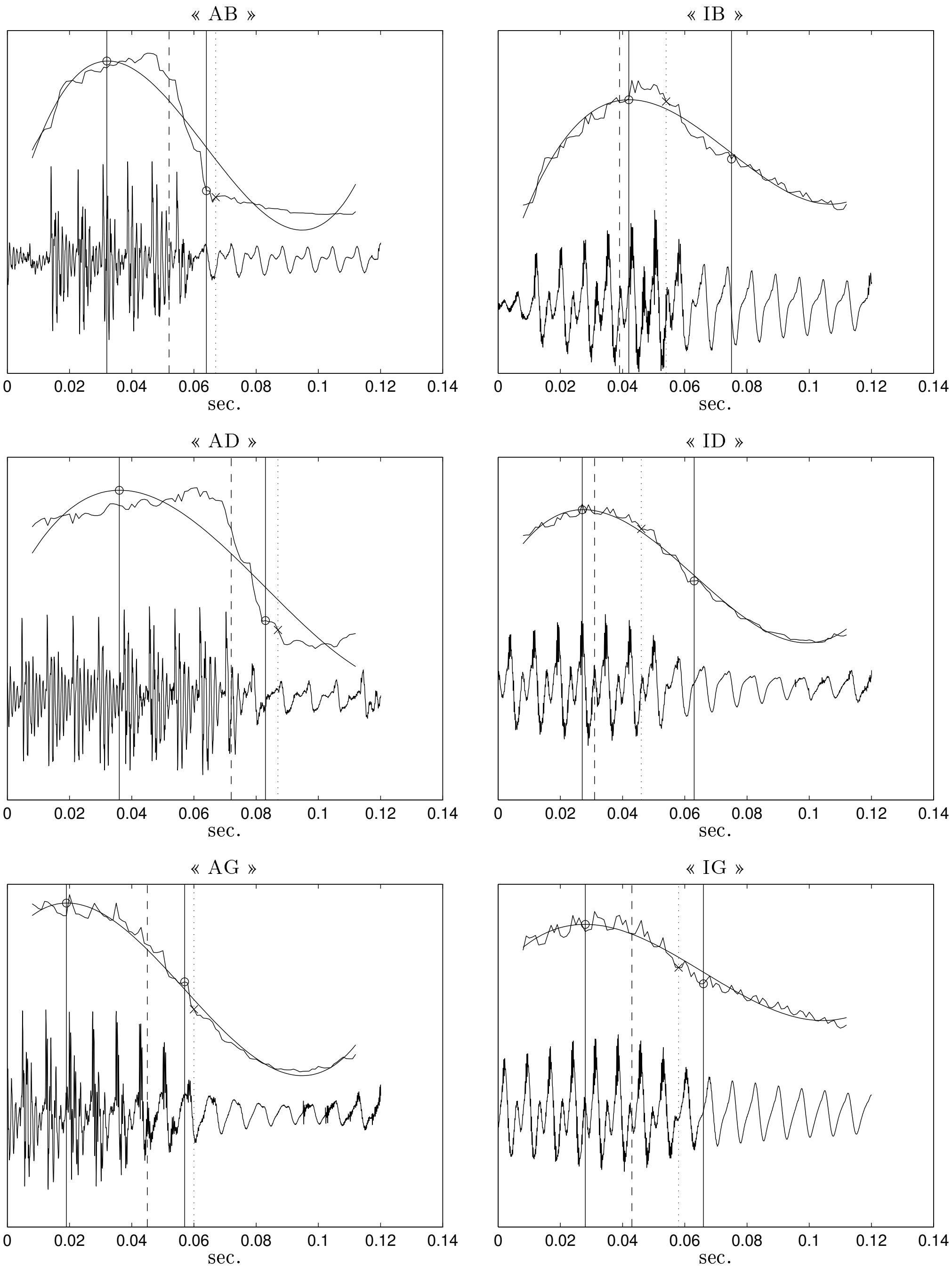

FIG. 4.4: Étiquetage des signaux : /a/ (gauche) et /i/ (droite).

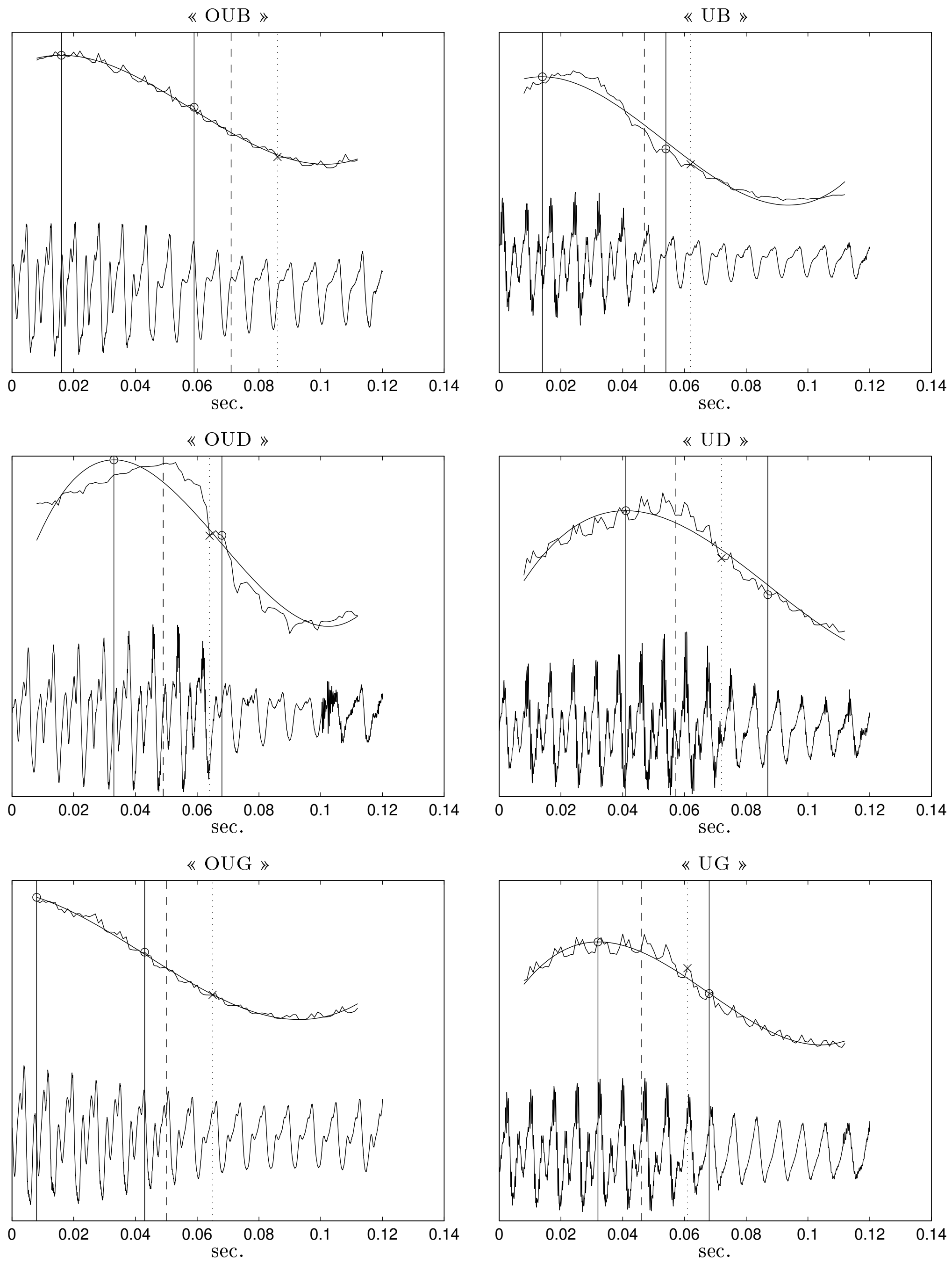

FIG. 4.5: Étiquetage des signaux : /u/ (gauche) et /y/ (droite).

mettre en œuvre la partie vocalique de l'inversion à l'aide des algorithmes d'optimisation , en commençant par la configuration articulatoire de la base de données originale dont les formants, après la distorsion correspondante, s'approchent le plus des formants désirés. Nous faisons ensuite pour chaque trame une optimisation à partir de la configuration obtenue pour la trame précédente.

Notre modèle acoustique (Badin & Fant 1984) n'etant pas capable de produire des formants au dessous de 190 Hz, nous arrêterons l'inversion de la partie vocalique 15 ms avant le passage du premier formant au dessous de 190 Hz. La figure 4.6 montre les résultats de l'application des configurations articulatoires obtenues de l'inversion (en cercle, à gauche), superposées aux formants de la base de données (en '+') pour les séquences AB, AD et AG. Les coupes sagittales des configurations articulatoires correspondantes sont montrées à droite. Les résultats obtenus pour les autres séquences sont quasiment identiques.

4.4.2 Inversion de la partie consonantique : Vérification des trajectoires

Avant d'essayer d'inverser la partie consonantique des séquences, nous regarderons de près les formants des trajectoires de constriction générées par le modèle. Nous avons pris des prototypes vocaliques pour chacune des voyelles dans la base de données (i.e., /i/, /a/, /y/ et /u/) et nous avons forcé le modèle a réaliser une constriction correspondant à chacun des lieux X_c possibles (cf. chapitre 2, section 2.5). En faisant une interpolation linéaire nous avons obtenu trente configurations articulatoires correspondant à la trajectoire formantique du modèle. Les figures 4.7 à 4.10 montrent les résultats acoustiques de ces trajectoires.

Étudions cas par cas[3] :

- /a/ : Pour le lieu de constriction X_C = ll les formants restent invariants jusqu'à quelques échantillons avant la fermeture, surtout le premier, qui présente une brusque chute en finale. Le premier formant descend de façon plus ou moins linéaire dans le reste des cas. Les trois autres formants prennent des valeurs d'autant plus hautes que la constriction est arrière. Les troisième et quatrième formants redescendent à partir de X_c = d7.
- /i/ : La constriction labiale produit un très brusque changement dans les trois derniers formants juste avant la fermeture. Le quatrième formant prend la place du troisième et celui-ci celle du deuxième. Le deuxième formant descend très brusquement de 2400 Hz jusqu'à 1000 Hz. Les constrictions en avant (tt1, d2, d3) ne présentent de différence significative qu'au niveau du troisième formant qui descend un peu pour X_C = tt1, un peu moins pour X_C = d2 et presque pas pour X_C = d3. Les quatre

[3]Étant donné que la base de données ne comprend aucune occlusive rétroflexe, les lieux de constriction X_c = tt3 ,tt2 ne sont pas étudiés.

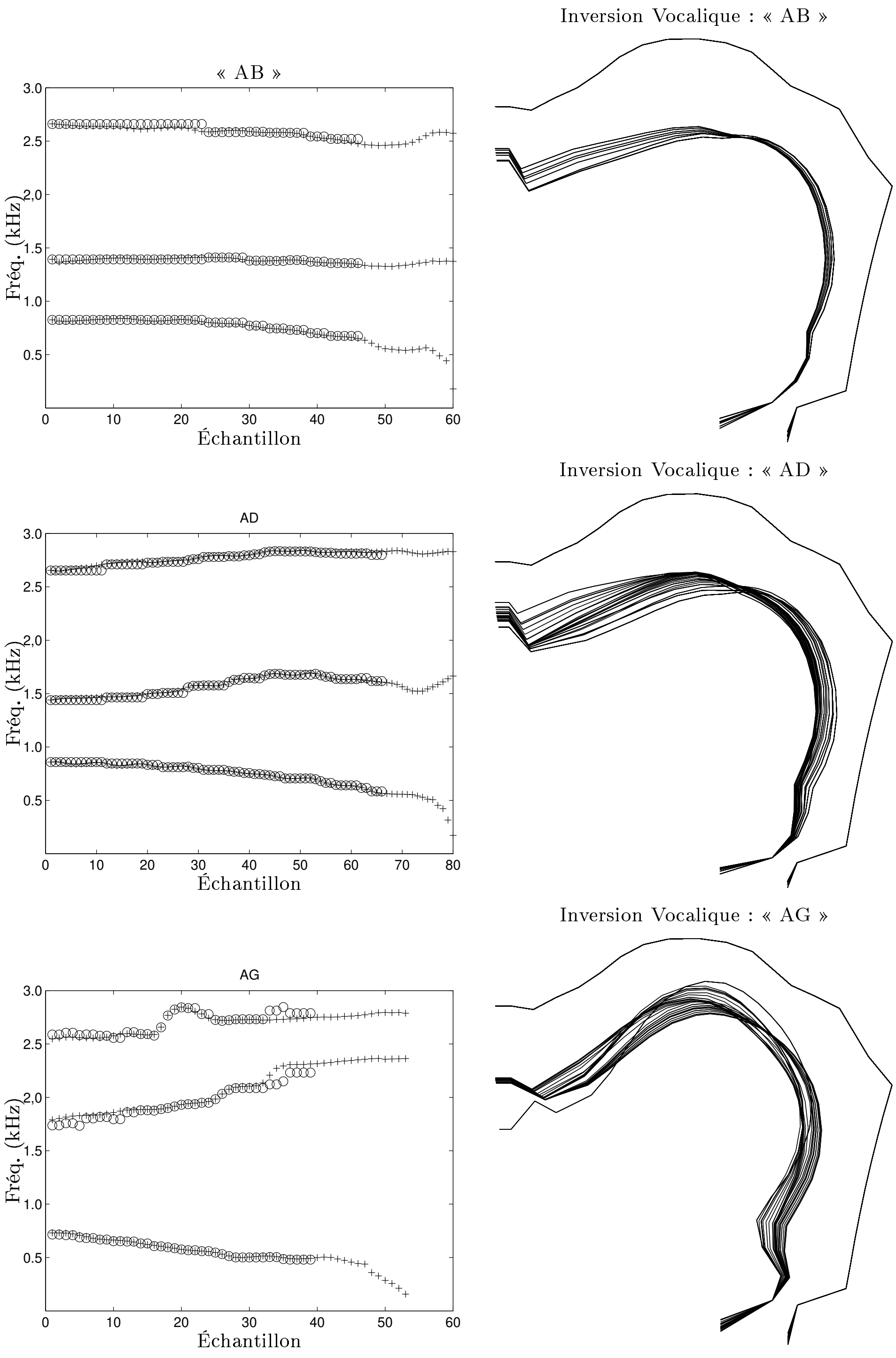

FIG. 4.6: Résultats de la partie vocalique de l'inversion.

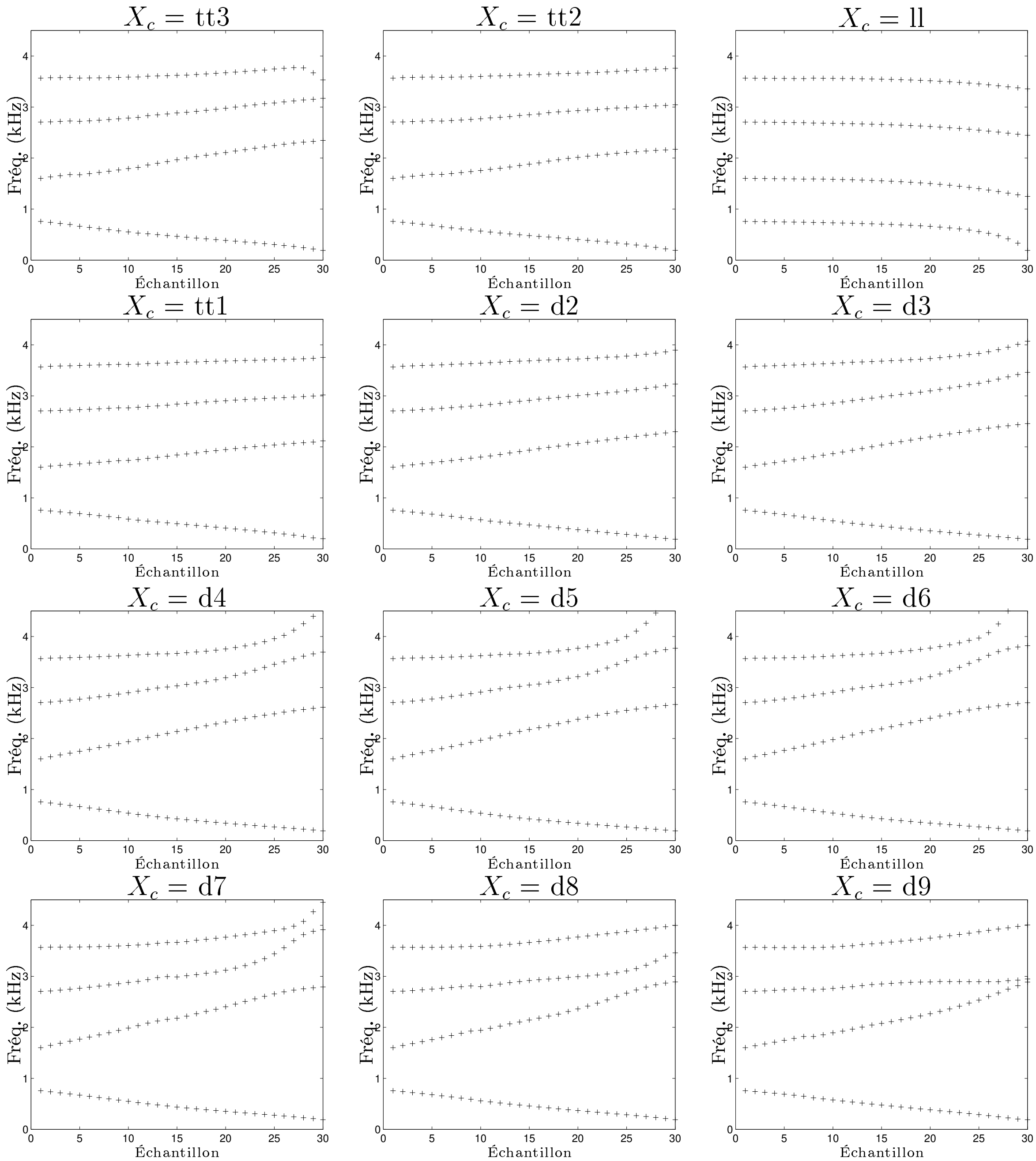

FIG. 4.7: Formants produits par les différents lieux de constriction à partir de /a/.

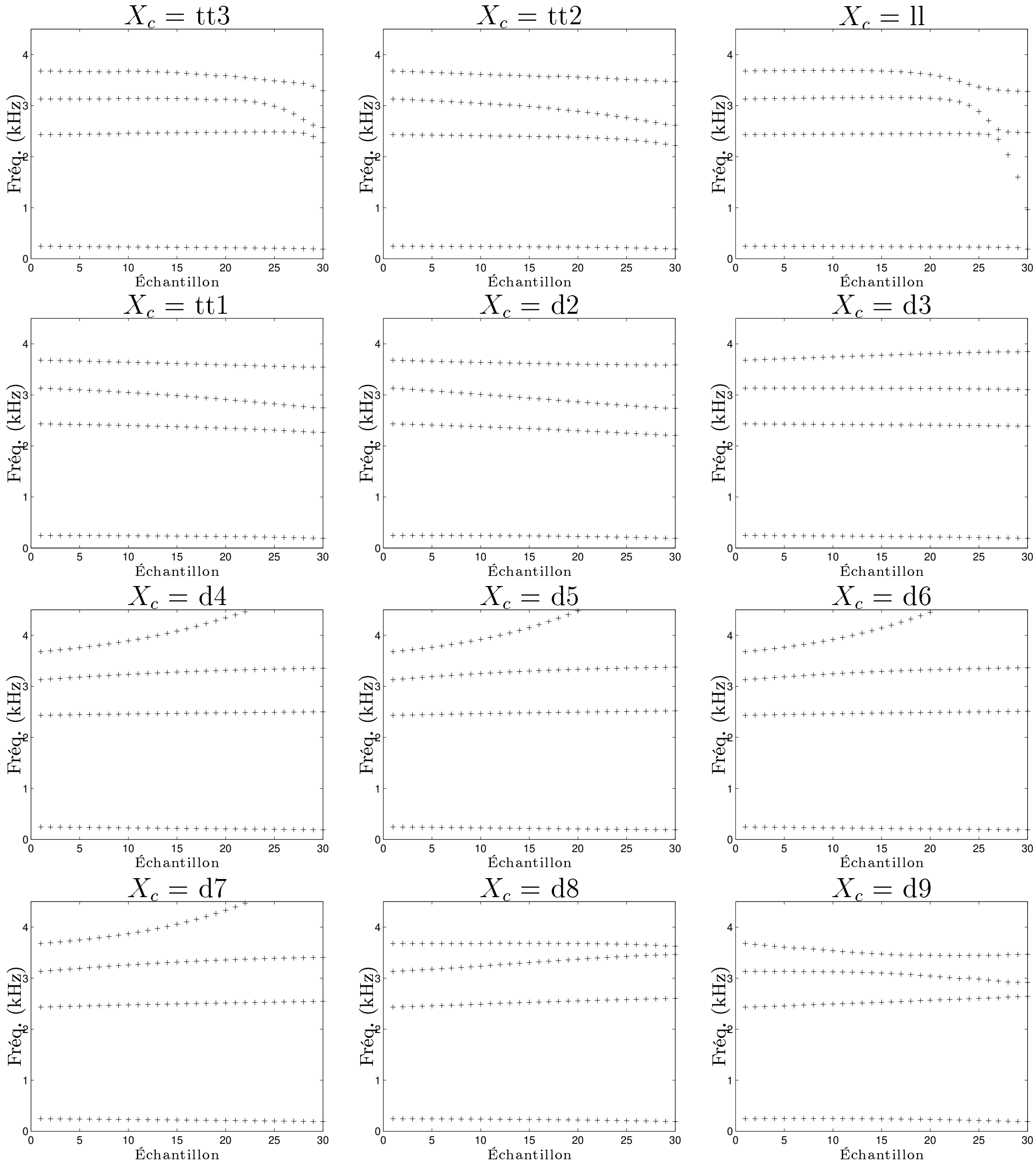

FIG. 4.8: Formants produits par les différents lieux de constriction à partir de /i/.

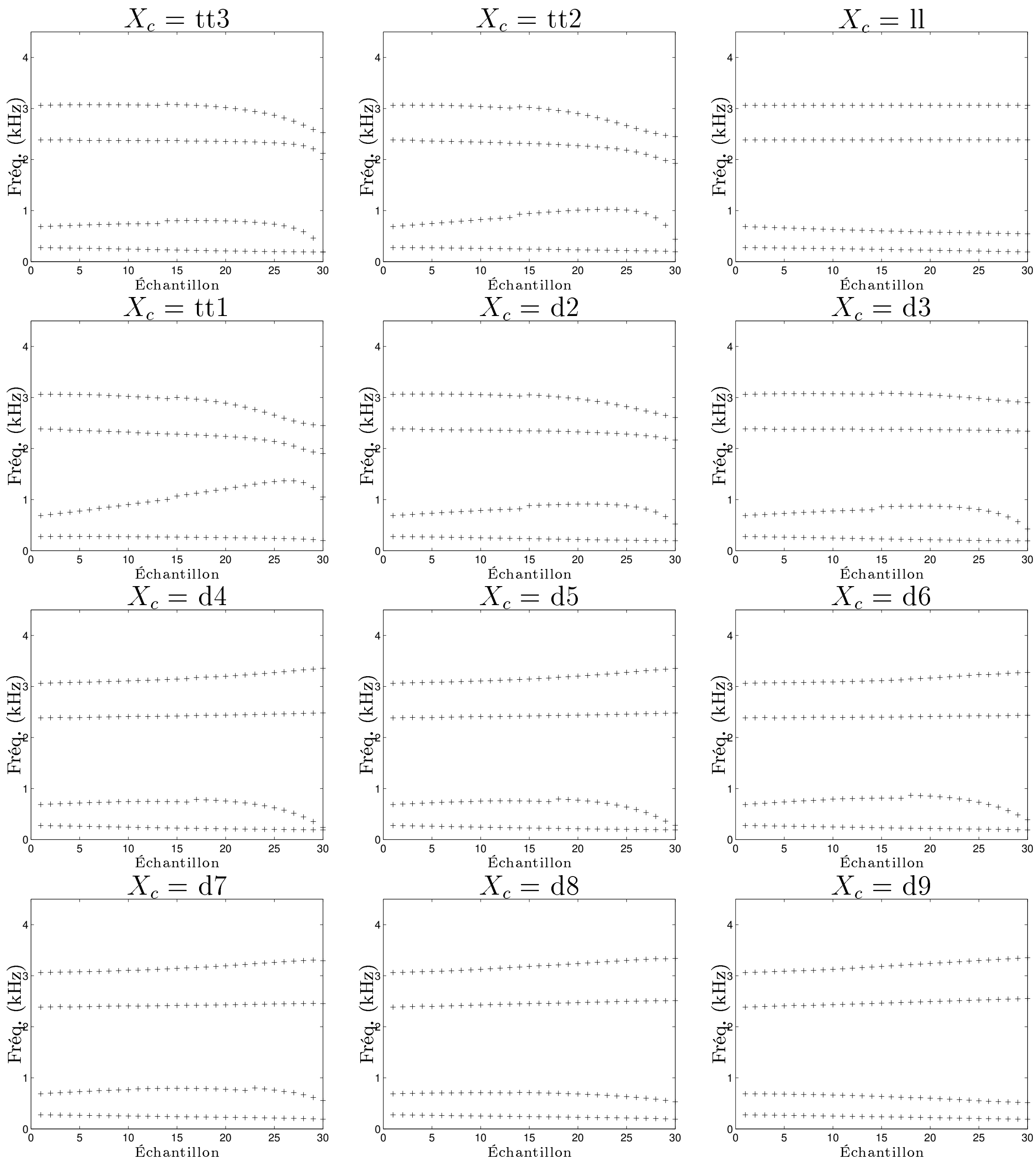

FIG. 4.9: Formants produits par les différents lieux de constriction à partir de /u/.

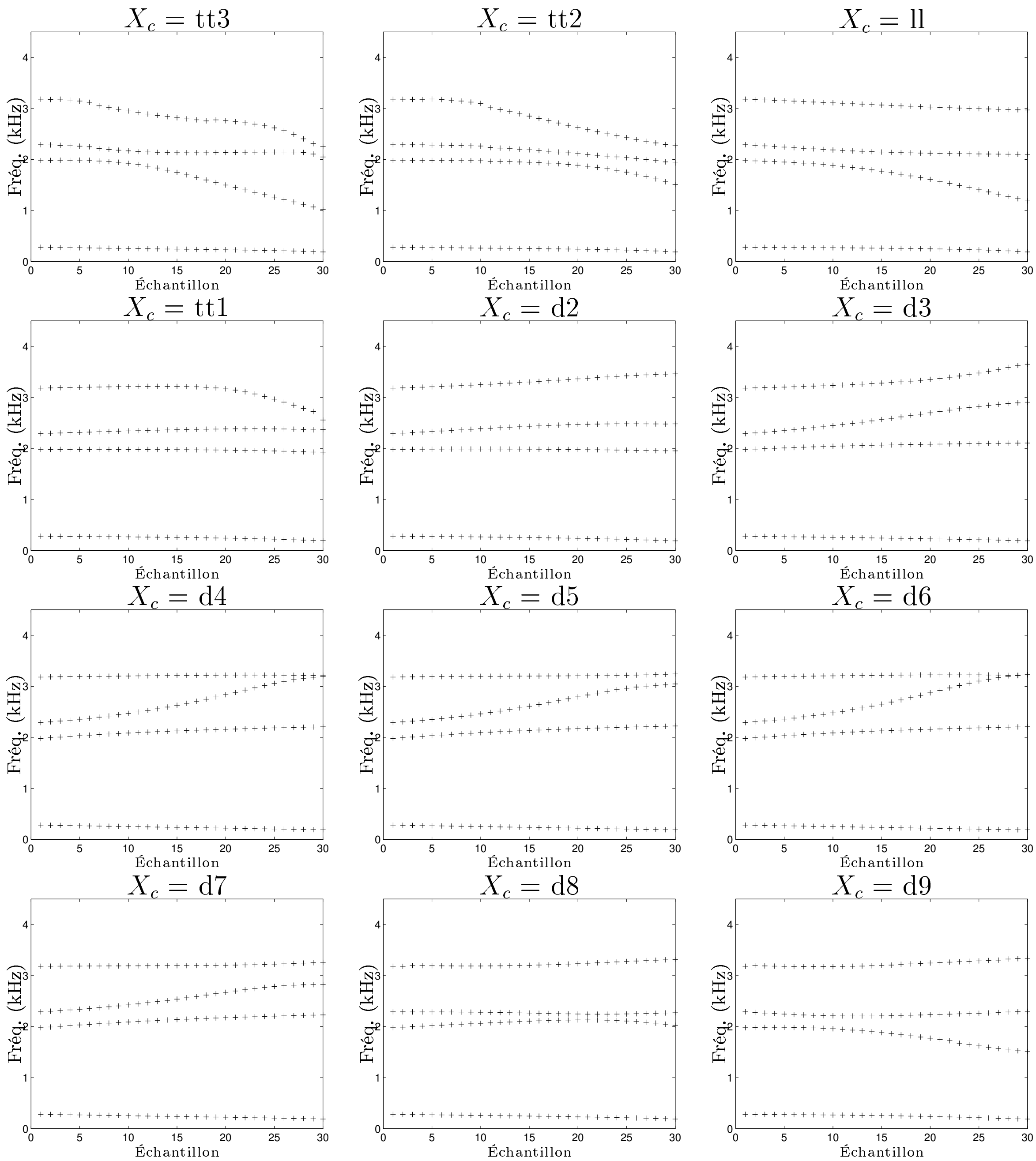

FIG. 4.10: Formants produits par les différents lieux de constriction à partir de /y/.

suivants (X_C = d4 à d7) présentent toutes une très grande montée du quatrième formant. X_C = d8 présente une montée du troisième formant tandis que X_C = d9 présente une descente de ce même formant, les deuxième et quatrième formant restant relativement stables.

- /u/ : La constriction aux lèvres ne fait changer les valeurs des formants que de très peu. Une constriction très en avant (X_C = tt1) produit une forte montée du deuxième formant avec une légère redescente à la fin et une descente des troisième et quatrième formants. Le deuxième formant se comporte de façon très semblable, bien qu'avec une beaucoup plus faible montée et redescente pour les constrictions correspondant à X_C = d2 à d7. Les troisième et quatrième formants changent d'une légère descente vers une légère montée pour ces constrictions. Les deux dernières constrictions n'ont presque pas d'effet sur les formants. Nous remarquons la forte ressemblance entre les formants correspondant aux constrictions X_C = ll, d8 et d9.
- /y/ : Une constriction aux lèvres produit une descente du deuxième formant et laisse les autres invariants. La constriction correspondant à X_C = tt1 présente une descente du quatrième formant de façon à presqu'atteindre le niveau du troisième. Les deux constrictions suivantes produisent une montée des troisième et quatrième formants. X_C = d4 à d7 produisent une montée du troisième formant qui rejoint le quatrième. X_C = d8 fait monter légèrement le deuxième formant avant de le faire redescendre. Finalement X_C = d9 fait descendre le deuxième formant et monter le troisième, ce dernier de très peu. Remarquons la ressemblence entre X_C = ll et d9.

Nous remarquons que les tendances observées correspondent de façon satisfaisante à la base de données que nous essayons d'inverser.

4.4.3 Inversion de la partie consonantique : Méthode à utiliser

Pour la partie du signal acoustique de la voyelle et jusqu'à l'occlusion, nous utiliserons une approche probabiliste basée sur les capacités de production de notre modèle, afin d'inférer les configurations articulatoires produisant les formants désirés. Nous définirons une « séquence d'occlusion » comme la transition formantique à partir de n ms avant l'etiquette indiquant l'occlusion du conduit vocal. Nous appellerons le temps n ms avant la fermeture t_0 et celui au moment de la fermeture t_n. La procédure décrite dans la section précédente nous fournissant la configuration articulatoire a_0, nous cherchons l'ensemble de configurations $[a_1, a_2 \dots a_n]$ capables de produire les formants $F = [f_1, f_2 \dots f_n]$ donnés.

À partir d'une configuration articulatoire quelconque nous avons une méthode pour arriver à une constriction donnée (chapitre 2). Cette méthode nous donne une configuration articulatoire à partir d'une configuration initiale et un lieu de constriction.

En prenant la dernière configuration articulatoire générée par l'inversion vocalique, nous sommes capables de produire une configuration finale pour chaque lieu de constriction possible. En faisant une interpolation linéaire entre ces deux points dans l'espace articulatoire, nous aurons à partir d'une configuration initiale a_0 et un lieu de constriction X_c un ensemble d'articulateurs $A_c = \mathcal{F}(a_0, X_c) = [a_{1,c}\ , \ a_{2,c} \dots\ a_{n,c}]$ (cf. ci-dessus). L'ensemble des formants correspondant à ces articulateurs $F_c = [f_{1,c}\ , \ f_{2,c} \dots\ f_{n,c}]$ pourra alors être comparé aux formants désirés F.

Toutefois, nous n'avons qu'une configuration de départ, et rien ne nous assure que celle-ci soit représentative de l'ensemble des configurations capables de produire les formants désirés. Une approche probabiliste exige que toutes les configurations capables de produire le signal soient considérées pour ensuite choisir l'ensemble correspondant au lieu d'articulation le plus probable. Nous devons alors considérer la fonction de probabilité appropriée sur l'ensemble des configurations produisant les formants f_0.

Soit $\mathcal{A}$ la fibre articulatoire contenant toutes les configurations a_0 capables de produire les formants f_0, nous cherchons pour chaque lieu de constriction X_c, la probabilité que les formants $F' = [f_0, F]$ aient été produits par la trajectoire correspondante A_c, et cela sur toute la fibre $\mathcal{A}$. En d'autres termes, nous cherchons, étant donné F' une évaluation de la constriction X_c à partir de toutes les $a_0 \in \mathcal{A}$ (équation 4.1).

$$P(X_c|F') = \int_{\mathcal{A}} P(X_c, a_0|F')\ da_0 \tag{4.1}$$

L'équation 4.1 vise à estimer la probabilité $P(X_c|F')$ à partir des formants F' donnés. Il est important d'inclure dans le modèle probabiliste, un terme correspondant aux autres connaissances que nous pourrions avoir (e.g., position des lèvres d'après données visuelles, connaissances contextuelles d'après les phonèmes déjà identifiés). Dans la tradition des méthodes probabilistes Bayesiennes, ces connaissances sont dénommées « a priori » (Jaynes 1997). Soit I l'ensemble de toutes ces connaissances, nous cherchons alors une évaluation de la constriction X_c étant donné les formants F' et les connaissances I (cf. équation 4.2).

$$P(X_c|F', I) = \int_{\mathcal{A}} P(X_c, a_0|F', I)\ da_0 \tag{4.2}$$

Après quelques manipulations des fonctions de probabilité, nous arrivons à l'équation :

$$P(X_c|F', I) = \frac{1}{P(F'|I)} \int_{\mathcal{A}} P(F'|a_0, X_c, I)\ P(X_c|a_0, I)\ P(a_0|I)\ da_o \tag{4.3}$$

Si en plus nous considérons que F' peut être separée en f_0 et $F = [f_1, f_2, \dots f_{15}]$ alors nous obtenons $P(F'|a_0, X_c, I) = P(F|f_0, a_o, X_c, I)\ P(f_0|a_0, X_c, I)$. Ceci nous donne l'équation :

$$P(X_c|F', I) = \frac{1}{P(F')} \int_{\mathcal{A}} P(F|f_0, a_0, X_c, I)\ P(X_c|a_0, I)\ P(f_0|a_0, X_c, I)\ P(a_0, I)\ da_0 \tag{4.4}$$

Où : $P(F')$ – terme de normalisation de la fonction de probabilité.

$P(F|f_0, a_0, X_c, I)$ – probabilité que les formants F aient été produits par la trajectoire générée à partir de la configuration a_0 et la constriction X_c. La trajectoire ne dépendant pas de f_0, ce terme devient alors $P(F|A_c, I)$ (A_c étant une fonction de a_0 et S_c).

$P(f_0|a_0, X_c, I)$ – probabilité d'avoir les formants f_0 à partir de la configuration articulatoire a_0, i.e. probabilité que a_0 appartienne à la fibre articulatoire $\mathcal{A}$. Les formants f_0 ne dépendant que des articulateurs a_0, ce terme devient $P(f_0|a_0)$.

$P(X_c|a_0, I)$ – probabilité d'avoir une constriction à X_c à partir de a_0 : anticipation articulatoire.

$P(a_0, I)$ – terme correspondant aux contraintes articulatoires; Nous admettrons ici que les valeurs des paramètres de a_0 suivent une distribution de probabilité normale dont la variance est déterminée par les données radiographiques ayant servi a établir le modèle. Nous avons inclure des contraintes géométriques correspondant aux valeurs limites : $\mid a_0 \mid \leq 3$. De plus, a_0 est tel que l'aire la plus petite dans la fonction d'aire correspondant à a_0 est supérieure à 0.1 cm^2.

En prenant en compte les considérations ci-dessus, nous arrivons à l'équation 4.5.

$$P(X_c|F', I) = \frac{1}{P(F')} \int_{\mathcal{A}} P(F|A_c, I)\ P(X_c|a_0, I)\ P(f_0|a_0, I)\ P(a_0, I)\ da_0 \tag{4.5}$$

4.4.4 Inversion de la partie consonantique : Solution

La solution de l'équation 4.5 nécessite l'obtention de la totalité de la fibre articulatoire associée aux formants f_0. La relation entre les articulateurs du modèle et les résultats acoustiques d'une configuration quelconque étant fort compliquée, nous essayerons de trouver une estimation aussi proche que possible de cette fibre. Pour ce faire, nous considérerons que les formants $\tilde{f}_0$ correspondant à une configuration articulatoire a_0 quelconque ont une probabilité $P(\tilde{f}_0 = f_0)$ d'appartenir à une fibre articulatoire $\mathcal{A}$ correspondant aux formants f_0. Soit C la matrice de covariance des formants associée à f_0[4], $\tilde{f}_0$ les formants calculés pour la configuration a_0 et n le nombre de dimensions du point f_0 (nombre de formants considérés), nous considérerons que cette probabilité suit une distribution gaussienne (équation 4.6).

$$P(\tilde{f}_0 = f_0) = \frac{1}{\sqrt{(2\pi)^n [C]}}\ \exp\Big\{-\frac{1}{2}(\tilde{f}_0 - f_0)^T C^{-1} (\tilde{f}_0 - f_0)\Big\} \tag{4.6}$$

[4]Cette matrice de covariance a été calculée à partir des mesures d'incertitude liées à l'analyse LPC.

Nous utiliserons l'algorithme de *Monte-Carlo* pour obtenir une estimation de la solution à l'équation 4.5. Il n'entre pas dans le cadre de cette thèse de faire une étude approfondie de cet algorithme, nous l'utiliserons comme un outil pour résoudre le problème qui nous concerne. Pour une description détaillée, voir Neal (1993).

L'algorithme *Monte-Carlo* est fondé sur l'approximation décrite par l'équation 4.7, où l'on estime la somme sur un ensemble de variables $\tilde{x}_1, \ldots \tilde{x}_n$ par la somme sur un échantillon $x_1^{(t)} \ldots x_n^{(t)}$ de taille N.

$$\sum_{\tilde{x}_1} \cdots \sum_{\tilde{x}_n} g(\tilde{x}_1, \ldots \tilde{x}_n)\ P(X_1 = \tilde{x}_1, \ldots X_n = \tilde{x}_n)\ \approx\ \frac{1}{N} \sum_{t=0}^{N-1} g(x_1^{(t)} \ldots x_n^{(t)}) \tag{4.7}$$

Un point crucial dans l'algorithme est le choix de l'échantillon à considérer. Il doit être représentatif de la distribution de probabilité $P(X_1 = \tilde{x}_1, \ldots X_n = \tilde{x}_n)$. Dans notre cas, étant donné que nous cherchons à trouver dans tout l'espace articulatoire le sous-espace qui produit les formants f_0, les valeurs des variables j, b, d etc. doivent être représentatives de la fibre articulatoire.

Choisir l'échantillon approprié dans l'ensemble d'articulateurs possibles équivaut à choisir les points a_0 qui représentent la fibre articulatoire $\mathcal{A}$ recherchée. Il est à remarquer que la fibre articulatoire est contrainte aux valeurs des articulateurs dans le rang approprié ($[-3\ ,\ +3]$) ainsi qu'aux valeurs respectant les conditions d'aire minimale. Nous incluons alors dans la distribution de probabilité (dont les articulateurs choisis doivent être représentatifs) le terme dans l'équation 4.5 correspondant à cette contrainte ainsi qu'à la probabilité que a_0 appartienne à la fibre. Nous aurons alors la solution représentée par l'équation 4.8.

$$\begin{aligned} P(X_c|F', I) &= \frac{1}{P(F')} \int_{\mathcal{A}} P(F|A_c, I)\ P(f_0|a_0)\ P(X_c|a_0, I)\ P(a_0, I)\ da_0 \\ &\approx \frac{1}{P(F')} \frac{1}{N} \sum_{i=0}^{N-1} P(F|A_c(i), I)\ P(X_c|a_0(i), I) \end{aligned} \tag{4.8}$$

Où : $a_0(i)$ suit la distribution de probabilité $P_f(a_0) = P(f_0|a_0, I)\ P(a_0)$.

4.4.5 Estimation de la fibre articulatoire

Afin d'obtenir un échantillon représentatif de la fibre articulatoire, nous avons utilisé l'algorithme *Metropolis* (cf. Neal (1993)) avec la configuration a_0 comme configuration articulatoire initiale, que nous appellerons dorénavant « état initial ». Soit a_0 l'état initial, nous cherchons un ensemble $a_0(1)\dots a_0(N)$ tel que $P(f_0|a_0(i), I)$ corresponde à la distribution de probabilité normale (cf. équation 4.6) et contrainte aux valeurs limites des articulateurs et d'aire minimale. L'algorithme *Metropolis* est décrit par les pas suivants :

1. Choisir un état candidat a_0^* tel que tous les composants, sauf le k^{ème} soient les mêmes que l'état précédant a_0. Le k-ième composant est choisi selon une distribution $S_k(a_0, a_0^*)$ qui peut éventuellement dépendre de a_0.
2. Accepter cet état candidat avec probabilité $A(a_0, a_0^*)$; sinon, le rejeter et garder l'état actuel a_0. Cette étape est accomplie en sélectionnant au hasard un numéro r de la distribution uniforme en [0, 1) pour ensuite choisir l'état suivant a'_0 de la façon suivante :

$$a'_0 = \begin{cases} a_0^* & \text{si } r < A(a_0, a_0^*) \\ a_0 & \text{sinon} \end{cases} \tag{4.9}$$

Nous utiliserons la fonction d'acceptation $A(a_0, a_0^*)$ décrite par l'équation 4.10, avec $P_f(a_0)$ – Probabilité que l'état a_0 appartienne à la fibre articulatoire $\mathcal{A}$ tout en respectant les valeurs des articulateurs ainsi que les critères d'aire minimale (cf. équation 4.8).

$$A(a_0, a_0^*) = \min(1, P_f(a_0^*)/P_f(a_0)) \tag{4.10}$$

Étant donné que nous pouvons à tout moment obtenir une estimation du jacobien des formants par rapport aux articulateurs par simple perturbation de l'état actuel, nous allons utiliser une variation de l'algorithme décrit ci-dessus afin d'accélérer la recherche et couvrir davantage l'espace articulatoire. Soit $J(a_0)$ le jacobien des formants par rapport aux articulateurs autour de l'état actuel a_0 avec :

$$J(a_0) = \begin{bmatrix} \frac{\partial F1}{\partial j_0} & \frac{\partial F1}{\partial b_0} & \cdots & \frac{\partial F1}{\partial p_0} \\ \frac{\partial F2}{\partial j_0} & \frac{\partial F2}{\partial b_0} & \cdots & \frac{\partial F2}{\partial p_0} \\ \frac{\partial F3}{\partial j_0} & \frac{\partial F3}{\partial b_0} & \cdots & \frac{\partial F3}{\partial p_0} \\ \frac{\partial F4}{\partial j_0} & \frac{\partial F4}{\partial b_0} & \cdots & \frac{\partial F4}{\partial p_0} \end{bmatrix} \tag{4.11}$$

Soient $[u_1, u_2, u_3, u_4, v_1, v_2, v_3, v_4]$ les vecteurs propres du jacobien $J(a_0)$, tels que les vecteurs v_i décrivent un espace dans le noyau $\mathcal{N}$ de $J(a_0)$ perpendiculaire à l'espace décrit par les vecteurs u_i, nous pouvons ajouter une perturbation Δa à l'état actuel a_0 telle que les formants qui en résultent soient égaux aux formants de a_0 (cf. équation 4.12).

$$\Delta a = \sum_{i=1}^{4} c_i\, v_i \tag{4.12}$$

Cette approche correspond à l'un des algorithmes décrits par Neal (1993, chapitre 5). Pour rester fidèles à cet algorithme, nous devons ajouter un terme hors du noyau $\mathcal{N}$ à la perturbation Δa. Nous utiliserons une combinaison des vecteurs u_i, décrite par l'équation 4.13 pour calculer la perturbation Δa.

$$\Delta a = \sum_{i=1}^{4} c_i\, v_i \;+\; \sum_{i=1}^{4} d_i\, u_i \tag{4.13}$$

Les coefficients c_i, d_i seront tirés d'après une distribution de probabilité normale avec moyenne zéro et écart types σ_c et σ_d respectivement. Les coefficients c_i étant liés au noyau du jacobien, ils peuvent prendre des valeurs aussi grandes que le jacobien est linéaire autour de l'état actuel, et ceci sans perturber les valeurs des formants correspondants. Les valeurs des coefficients d_i par contre, ne doivent pas être trop grandes car cela signifierait que l'on n'est plus dans la fibre articulatoire et l'état proposé serait rejeté par la fonction d'acceptation. Nous devons donc utiliser des écart types $\sigma_c \gg \sigma_d$. En pratique nous avons utilisé $\sigma_c = 0.1$ et $\sigma_d = 0.005$.

Nous allons donc utiliser $a_o^* = a_0 + \Delta a$ comme état candidat. La distribution de probabilité $S_k(a_0, a_0^*)$ de l'algorithme correspond alors à une combinaison linéaire des vecteurs propres de $J(a_0)$. Un point très important dans l'algorithme *Metropolis* est l'ergodicité : la probabilité d'aller d'un état a_0 à un état a_0^* doit être strictement égale à la probabilité d'aller de l'état a_0^* à l'état a_0. Ceci resterait vrai si $J(a_0) = J(a_0^*) \quad \forall\, a_0, a_0^*$. Comme cette égalité ne peut pas être assurée, un terme de compensation doit être inclus dans la fonction d'acceptation 4.10, qui devient alors :

$$A(a_0, a_0^*) = \min\Big(1, \frac{S(a_0^*, a_0)}{S(a_0, a_0^*)} P_f(a_0')/P_f(a_0)\Big) \tag{4.14}$$

Soient $[u_1' \ldots u_4', v_1' \ldots v_4']$ les vecteurs propres correspondant au jacobien $J(a_0^*)$, la probabilité de revenir de a_0^* vers a_0 dépendra des valeurs des coefficients c_i', d_i' tels que $a_0 = a_0^* + \Delta a'$, où $\Delta a'$ est donné par l'équation 4.15.

$$\Delta a' = \sum_{i=1}^{4} c_i'\, v_i' \;+\; \sum_{i=1}^{4} d_i'\, u_i' \tag{4.15}$$

La figure 4.11 illustre la façon d'aller de a_0 vers a_0^* et vice versa. Observons que les coefficients c_i', d_i' peuvent être calculés de la façon suivante ("$\langle$" et "$\rangle$" indiquent le produit scalaire) :

$$
\begin{aligned}
c'_i &= \langle a_0 - a_0^* \, , \, v'_i \rangle \\
d'_i &= \langle a_0 - a_0^* \, , \, u'_i \rangle
\end{aligned}
\tag{4.16}
$$

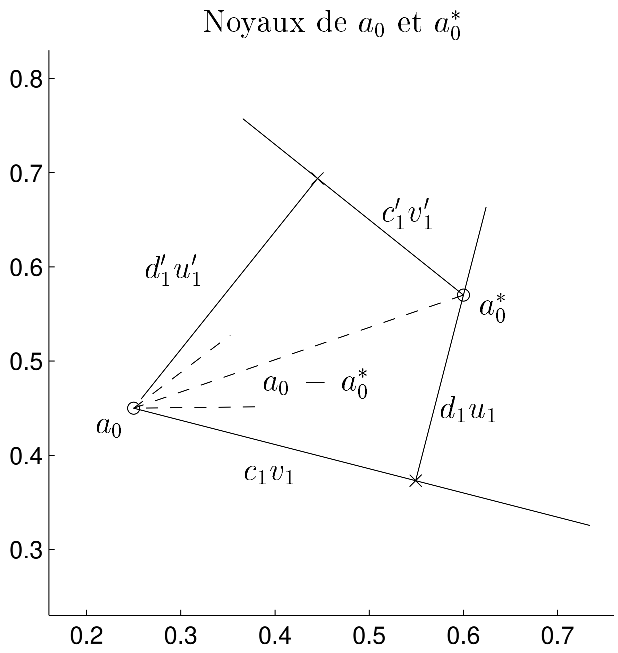

FIG. 4.11: Terme de compensation en fonction des noyaux.

Nous pouvons à présent calculer le rapport des probabilités entre $S(a_0^*, a_0)$ et $S(a_0, a_0^*)$. Ce rapport est décrit par l'équation 4.17.

$$
\frac{S(a_0^*, a_0)}{S(a_0, a_0^*)} = \frac{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{i=1}^{4}\Big(\frac{c'_i}{\sigma_c}\Big)^2 - \dfrac{1}{2}\displaystyle\sum_{i=1}^{4}\Big(\frac{d'_i}{\sigma_d}\Big)^2 \Big\}}{\exp\Big\{ -\dfrac{1}{2}\displaystyle\sum_{i=1}^{4}\Big(\frac{c_i}{\sigma_c}\Big)^2 - \dfrac{1}{2}\displaystyle\sum_{i=1}^{4}\Big(\frac{d_i}{\sigma_d}\Big)^2 \Big\}}
\tag{4.17}
$$

Après simplification, nous obtenons l'équation 4.18.

$$
\frac{S(a_0^*, a_0)}{S(a_0, a_0^*)} = \exp\Big\{ \frac{1}{2\sigma_c^2}\sum_{i=1}^{4}(c_i^2 - c_i'^2) \; + \; \frac{1}{2\sigma_d^2}\sum_{i=1}^{4}(d_i^2 - d_i'^2) \Big\}
\tag{4.18}
$$

Où : c'_i, d'_i – sont calculés d'après l'équation 4.16.

Nous avons obtenu de cette façon une estimation de la fibre articulatoire correspondant aux formants n ms avant l'occlusion pour chacune des séquences voyelle-consonne étudiées. Les formants correspondant à l'ensemble des configurations articulatoires obtenues de cette façon doivent représenter un seul point dans l'espace des quatre premiers formants tout en couvrant le plus possible l'espace articulatoire. Les figures 4.12 et 4.13 montrent les formants correspondant à la fibre articulatoire (en cercle superposés aux formants de la séquence AB) et un histogramme pour chacune des variables[5]. Observons que la fibre présente une variabilité importante dans l'espace articulatoire et une variance presque nulle dans l'espace formantique.

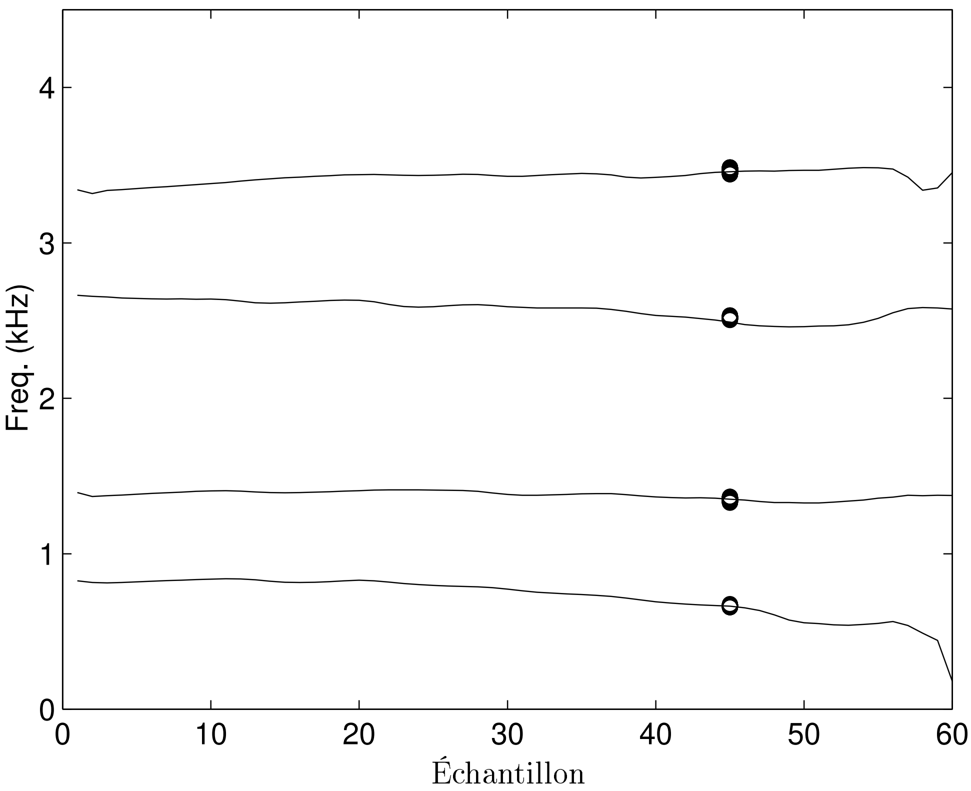

FIG. 4.12: Formants de la fibre articulatoire (séquence AB).

4.4.6 Implémentation

Étant donné l'approche que nous avons utilisée, notamment en ce qui concerne la séparation de l'inversion en partie vocalique et consonantique, il est clair que nous devons choisir l'instant où ces deux parties se séparent (ou éventuellement se croisent). De nombreuses

[5]La totalité des fibres articulatoires obtenues est présentée dans les annexes B et C.

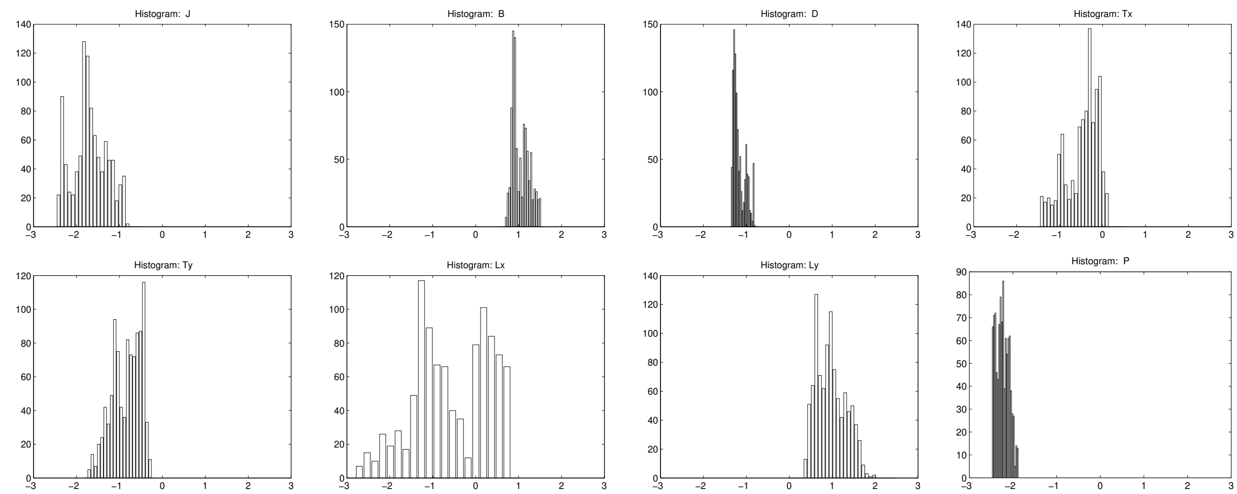

FIG. 4.13: Histogramme de chacune des variables de la fibre articulatoire.

études ont été menées en vue de décrire cette interaction de commandes, dite « coarticulation » (Öhman 1966, Fowler 1980, Whalen 1990). L'évaluation de ce phénomène sortant du cadre de notre étude, nous avons choisi de porter notre attention sur les résultats de notre méthode d'inversion à partir de deux instants possibles où les commandes correspondant à l'occlusion pourraient apparaître.

Deux fibres articulatoires ont donc été obtenues pour chacune des séquences étudiées, une correspondant au point E_m, qui sera comparée aux formants à partir de E_m et jusqu'à E_x ; et une autre au point $F1_m$, qui sera comparée aux formants à partir de $F1_m$ et jusqu'à $F1_x$ (cf. section 4.4, page 60). Chacune des fibres est composée de 2000 configurations articulatoires.

Soit $A_{c,\mathcal{A}}$ l'ensemble des trajectoires allant de chacune des configurations de la fibre articulatoire $\mathcal{A}$ jusqu'au lieu de constriction X_c, chaque élement de $A_{c,\mathcal{A}}$ aura vingt éléments correspondant à l'interpolation linéaire entre la configuration articulatoire $a_0(i) \in \mathcal{A}$ et la configuration faisant une constriction à X_c.

Soient $\mathcal{A}_F$ et $\mathcal{A}_E$ les fibres articulatoires estimées aux points $F1_m$ et E_m respectivement, nous calculerons pour chacune des phrases à inverser et pour chaque fibre, douze probabilités $P(X_c|F')$ (équation 4.8), une pour chacun des lieux de constriction X_c possibles.

Ceci nous donnera douze valeurs proportionnelles à la probabilité que les formants F' aient été produits par la constriction X_c à partir de toutes les configurations articulatoires produisant les formants f_0.

L'équation 4.8 nécessite l'estimation du terme $P(X_c|a_0(i), I)$; cependant, nous ne disposons pas des moyens permettant d'estimer ce terme. Nous limiterons alors notre étude au premier terme de l'équation 4.8.

Le but de notre approche étant de trouver le lieu de constriction le plus probable parmi ceux que le robot est capable de produire, nous nous concentrerons sur le rapport entre les probabilités associées à chacun de ces lieux de constriction. Étant donné cette approche, le terme de normalisation $\frac{1}{P(F')}$ de l'équation 4.8 ne sera pas pris en compte.

Nous essayerons maintenant d'obtenir une estimation de la probabilité qu'une séquence de formants F a d'être produite par une trajectoire articulatoire A_c avec les formants $[f_{1,c}\ \ldots\ f_{n,c}]$ correspondants.

En représentant chacune des séquences formantiques F par un ensemble de formants $[f_1 \ldots f_n]$ avec une mesure d'incertitude des valeurs des formants (mesure liée par exemple à la méthode d'obtention des formants) nous serions capables d'évaluer une trajectoire donnée A_c avec formants $[f_{1,c} \ldots f_{n,c}]$. Analysons le cas où les formants seraient obtenus avec une certitude suivant une distribution normale. F sera alors représenté comme un ensemble de distributions normales avec moyennes $[f_1 \ldots f_n]$ et matrices de covariance $[C_1 \ldots C_n]$.

Nous aurons pour chaque trajectoire A_c avec les formants correspondants $[f_{1,c} \ldots f_{n,c}]$ la fonction de probabilité suivante[6] :

$$
\begin{aligned}
P(F|a_0, X_c) &= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^4 [C_i]}} P(f_{i,c} = f_i) \\
&= \Big(\frac{1}{\sqrt{(2\pi)^4 [C_i]}} \Big)^n \exp \Big\{ -\frac{1}{2} \sum_{i=1}^{n} (f_{i,c} - f_i)^T C_i^{-1} (f_{i,c} - f_i) \Big\}
\end{aligned} \tag{4.19}
$$

Soient :

$$
\mathbb{F} = [f_{1,c} \ldots f_{n,c}]^T - [f_1 \ldots f_n]^T \quad , \quad \mathbb{C} = \begin{bmatrix} C_1 & 0 & 0 & \ldots & 0 \\ 0 & C_2 & 0 & \ldots & 0 \\ 0 & 0 & C_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & C_n \end{bmatrix}
$$

L'équation 4.19 peut être représentée par une seule distribution normale (en $\mathbb{R}^{4n}$) :

$$
P(F|a_0, X_c) = \frac{1}{\sqrt{(2\pi)^{4n} [\mathbb{C}]}} \exp \Big\{ -\frac{1}{2} \mathbb{F}^T \mathbb{C}^{-1} \mathbb{F} \Big\} \tag{4.20}
$$

[6]Cette approche suppose que le nombre d'élements de la trajectoire A_c est égal au nombre de formants F avec lesquels nous devons les comparer.

Comparer les formants produits aux formants désirés à chaque instant de la séquence, ce qu'implique la fonction de probabilité ci-dessus, équivaut à supposer une stricte synchronisation au niveau des commandes à l'origine de ces formants, la moindre déviation de formants étant exponentiellement « pénalisée ». Un modèle permettant de synchroniser les commandes articulatoires de façon précise s'impose si une comparaison du type impliqué par l'équation 4.20 est utilisée.

Cependant, l'état actuel de notre modèle ne lui permet que de fournir des configurations articulatoires à partir d'une configuration initiale et un lieu de constriction. Nous sommes donc contraints à utiliser une méthode basée sur la forme des signaux à comparer sans considérer l'échelle du temps.

L'obtention d'un paramètre lié à la similitude de deux séquences est un ancien problème dans la reconnaissance automatique de la parole (Itakura 1975, Sakoe & Chiba 1978, Rabiner, Rosenberg & Levinson 1978). Des interpolations linéaires ne donnant pas de résultats satisfaisants, nous utiliserons une méthode incluant un paramètre de distortion temporelle.

Sakoe & Chiba (1978) proposent une méthode simple basée sur le « Dynamic Time Warping ». Cette méthode donne une estimation de la distance entre deux séquences n'ayant pas la même echelle de temps. La méthode inclut en plus une matrice de pondération que nous supposerons égale à une constante multipliée par la matrice identité. Nous utiliserons cette distance comme variable aléatoire de la fonction de probabilité $P(F|A_c(i), I)$.

Soit $D(F, A_c)$ la distance entre les formants F donnés et les formants produits par l'ensemble de configurations articulatoires A_c, nous définirons la probabilité d'avoir les formants F à partir de la séquence A_c comme :

$$P(F|A_c(i), I) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}D^2(F, A_c)\right\} \tag{4.21}$$

Observons que l'équation ci-dessus correspond à une distribution de probabilité normale avec moyenne zéro et écart-type 1. Les paramètres correspondant à un écart-type quelconque sont pris en compte dans la fonction $D(F, A_c)$ (Sakoe & Chiba 1978).

4.5 Résultats

Les figures 4.14 à 4.17 montrent les résultats de l'évaluation de l'équation 4.8 pour chaque phrase à partir de la fibre articulatoire $\mathcal{A}_F$ avec $P(F|A_c(i), I)$ donné par l'équation 4.21 et $P(X_c|a_0(i), I) = \frac{1}{12} \quad \forall X_c$.

Chaque figure correspond à une phrase et chaque bloc représente la probabilité $P(X_c|F')$ (normalisée de façon à ce que la somme des probabilités soit égale à un). Le bloc de gauche représente la constriction aux lèvres $X_c = \text{ll}$ et celui de droite $X_c = \text{d9}$. Les blocs inférieurs correspondent aux constrictions rétroflexes $X_c = \text{t2}, \text{t3}$.

Les figures 4.18 à 4.21 montrent les résultats de l'évaluation de l'équation 4.8 pour chaque phrase à partir de la fibre articulatoire $\mathcal{A}_E$ avec $P(F|A_c(i), I)$ donné par l'équation 4.21 et $P(X_c|a_0(i), I) = \frac{1}{12} \quad \forall X_c$.

Nous observons que de meilleurs résultats sont obtenus avec la fibre articulatoire $\mathcal{A}_E$. Un cas très particulier est celui du « OUD », nous observons que seule la constriction $X_c = \text{t1}$ produit les formants appropriés (figures 4.2 et 4.9) ; cependant ce fait n'est pas reconnu lorsque l'on utilise la fibre articulatoire $\mathcal{A}_F$. En regardant de près nous observons que cette fibre articulatoire produit des formants dont le deuxième a déjà une valeur élevée (plus de 1000 Hz, en traits sur la figure 4.2). Il se trouve qu'à partir de ces formants d'autres constrictions produisent des formants qui restent proches aux formants désirés.

Nous observons le même cas de figure pour « AG ». Une constriction labiale produit des formants stables jusqu'à la constriction. Si la fibre articulatoire est calculée à un endroit où les formants sont déjà près de leurs valeurs finales, c'est une occlusion labiale que le système reconnaît.

Les probabilités obtenues avec la fibre articulatoire $\mathcal{A}_E$ reflètent mieux les similitudes que nous avons observées entre les signaux de la base de données et ceux que notre modèle produit à partir des voyelles (figures 4.2 et 4.7 à 4.10), notamment une grande confusion entre « OUB » et « OUD » (constrictions $X_c = \text{ll}$ et $X_c = \text{d9}$ dans « OUB » et « OUG »). Aucune confusion n'est observée pour le « OUD ».

Un cas assez particulier est celui du « IG » (cf. figure 4.20). Le système confond systématiquement les constrictions arrières à partir de $X_c = \text{d3}$. Cette confusion entre « IB » et « IG » a souvent été observée (Stevens & Blumstein 1978, Blumstein & Stevens 1980).

4.6 Conclusions

Nous avons étudié dans ce chapitre les relations entre les formants produits par un locuteur et ceux produits par notre modèle articulatoire. Nous observons que le modèle prédit de façon satisfaisante les lieux d'articulation à partir de la représentation formantique du signal d'un locuteur.

Bien que nous n'ayons utilisé que douze séquences, les probabilités que le système produit sont conformes à la totalité. Le pas suivant consiste à tester notre méthode primo avec la totalité de la base de données utilisée, secundo avec d'autres bases de données.

Nous sommes en mesure d'affirmer que la méthode que nous proposons, à savoir : nor-

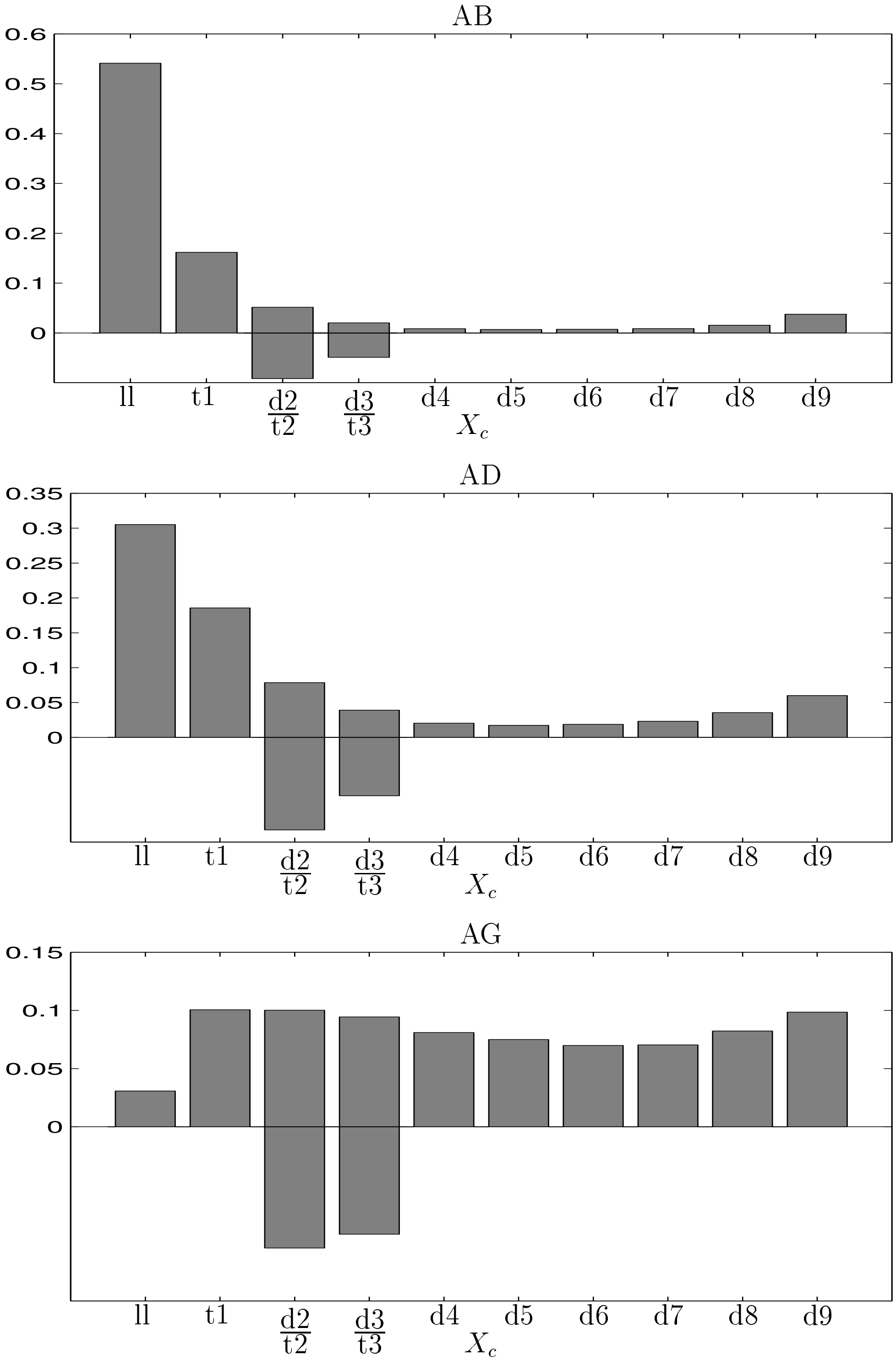

FIG. 4.14: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /a/, fibre articulatoire à $F1_m$.

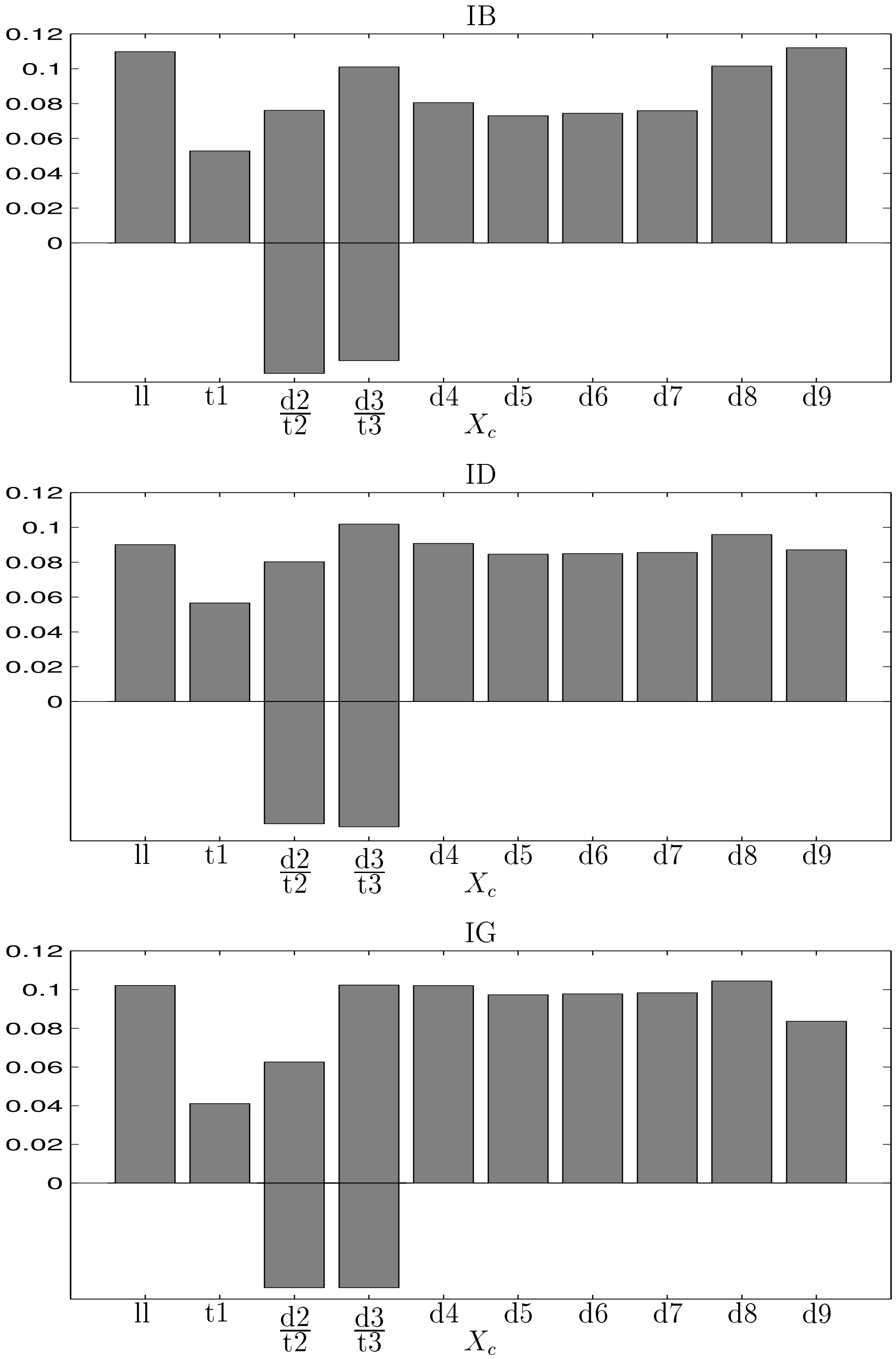

FIG. 4.15: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /i/, fibre articulatoire à $F1_m$.

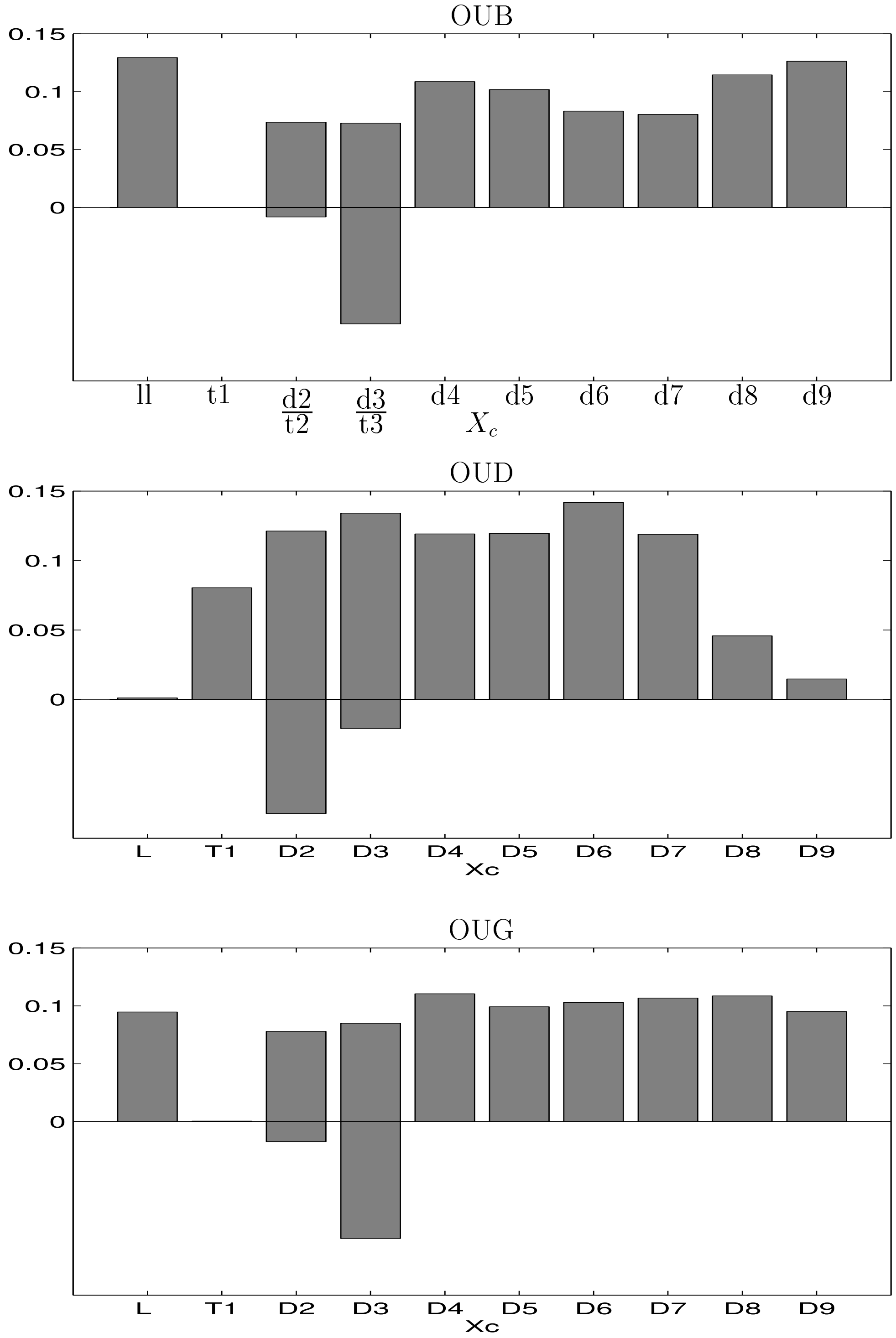

FIG. 4.16: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /u/, fibre articulatoire à $F1_m$.

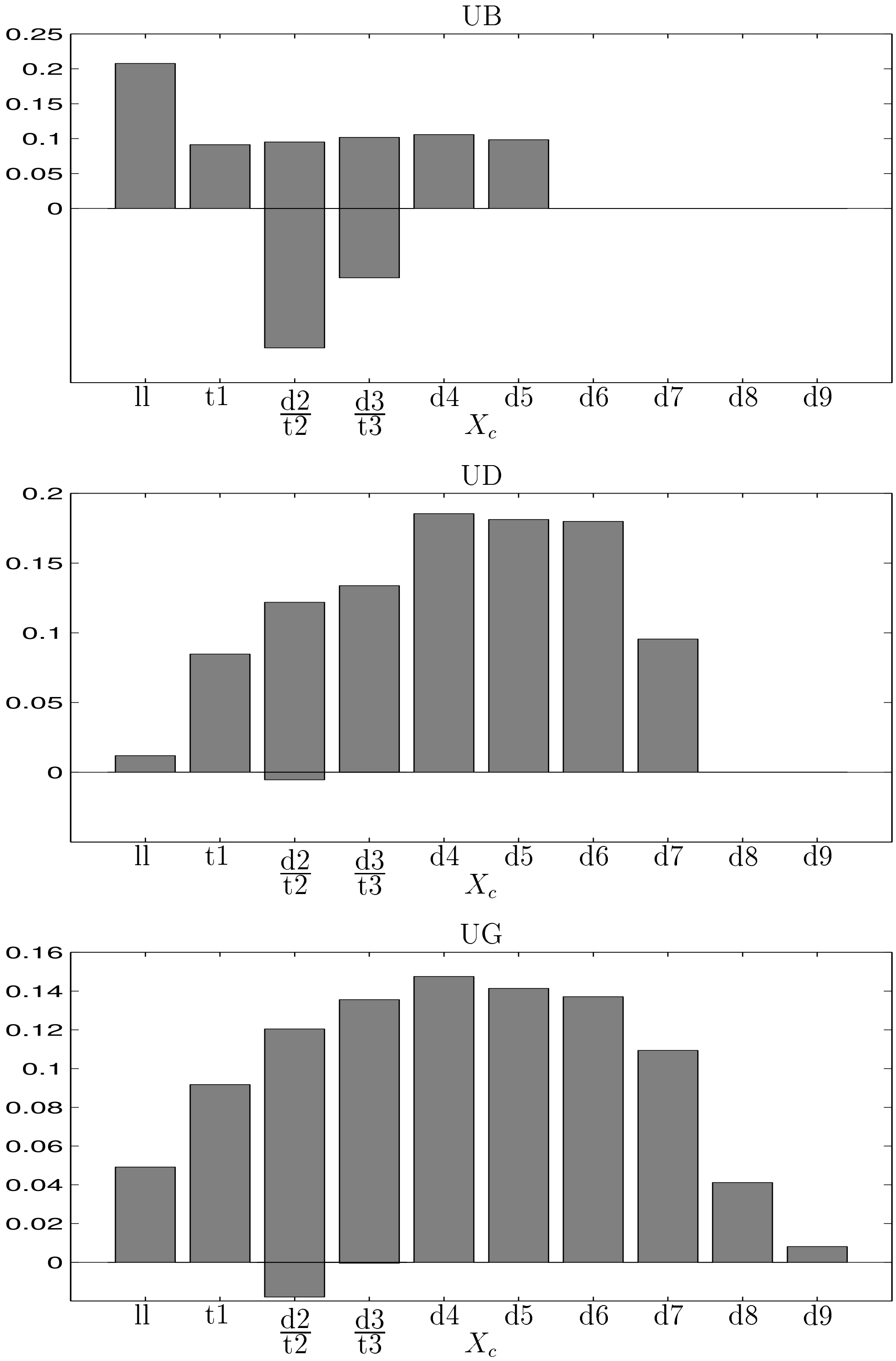

FIG. 4.17: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /y/, fibre articulatoire à $F1_m$.

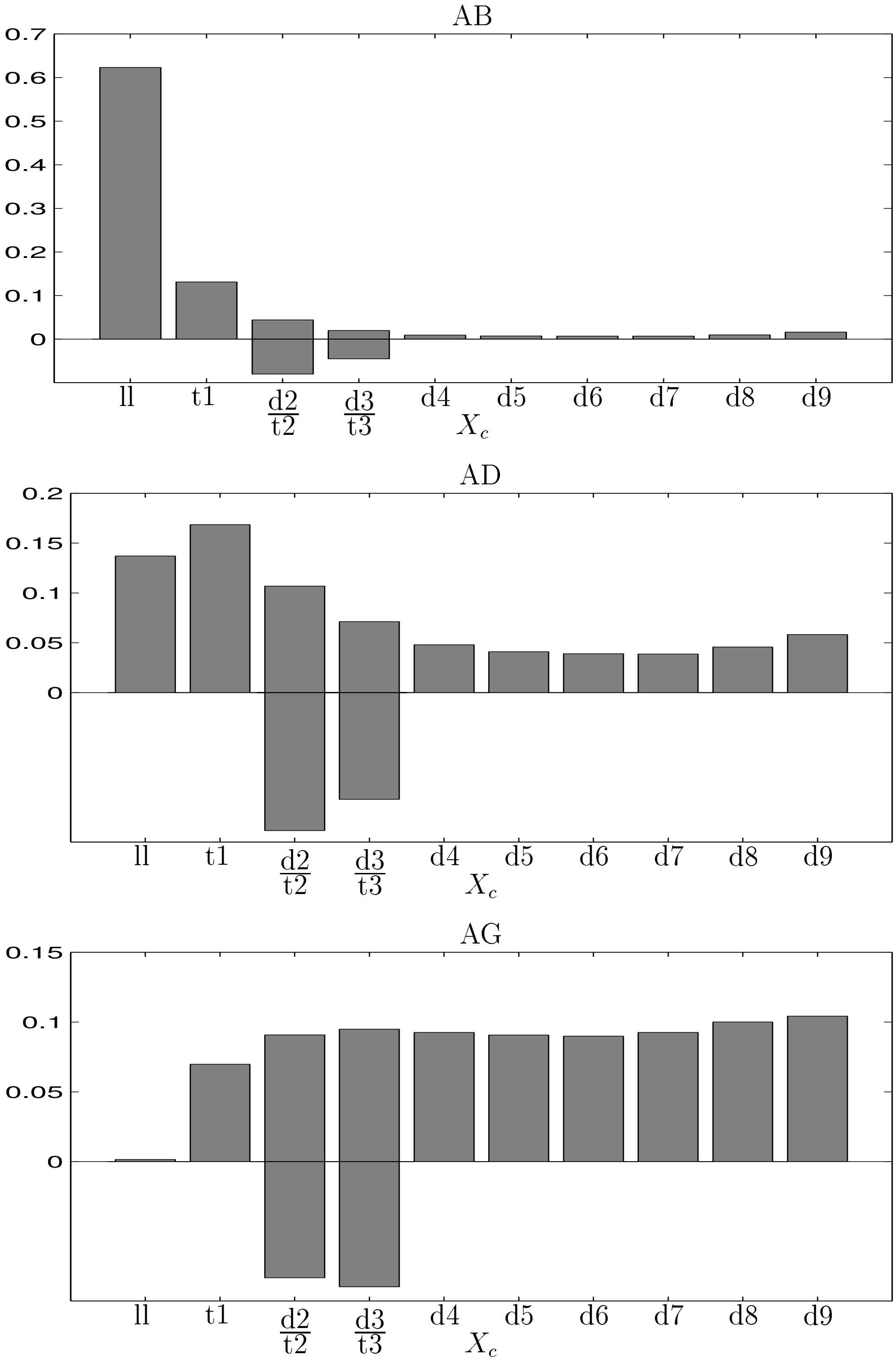

FIG. 4.18: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /a/, fibre articulatoire à E_m.

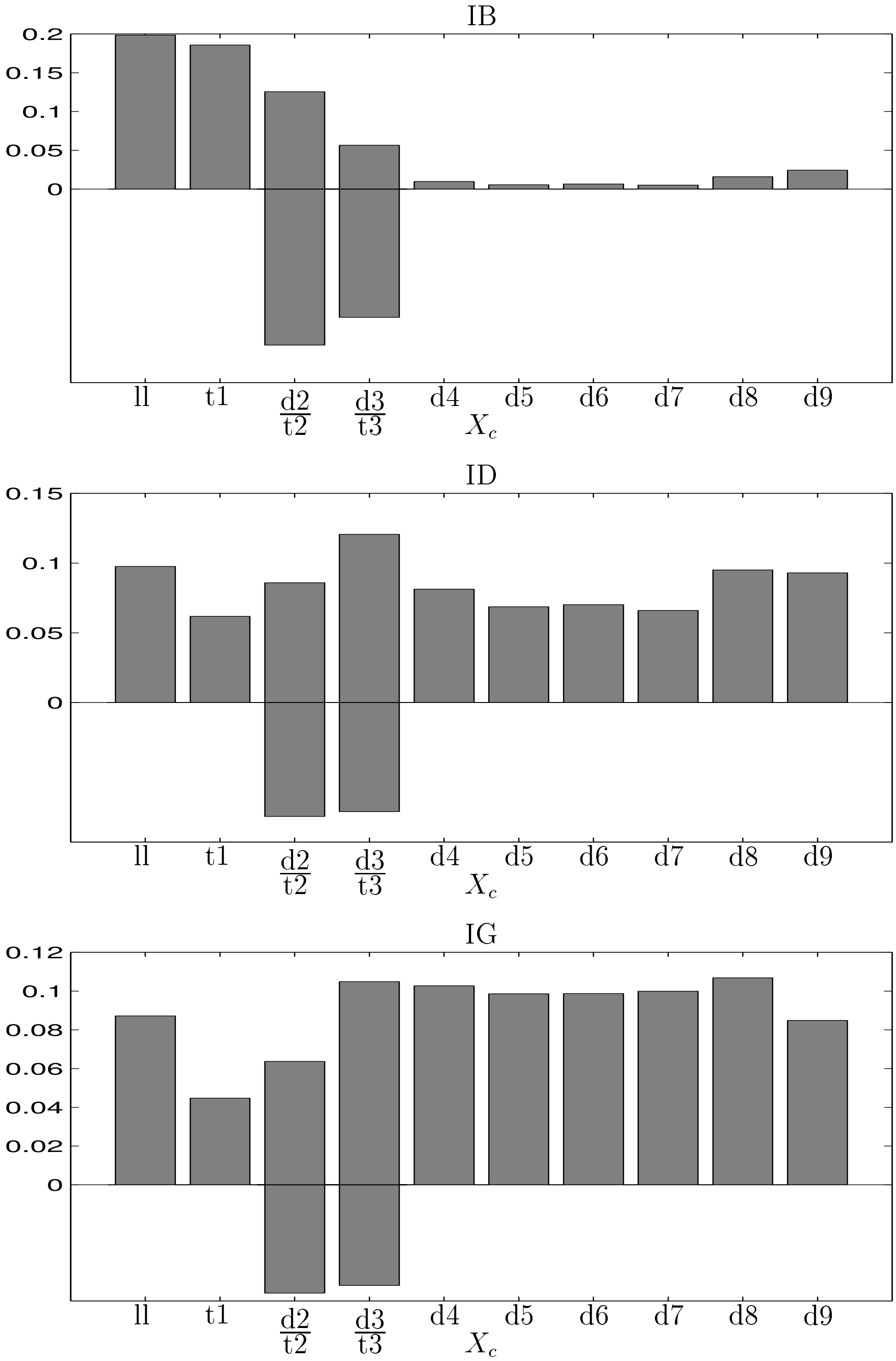

FIG. 4.19: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /i/, fibre articulatoire à E_m.

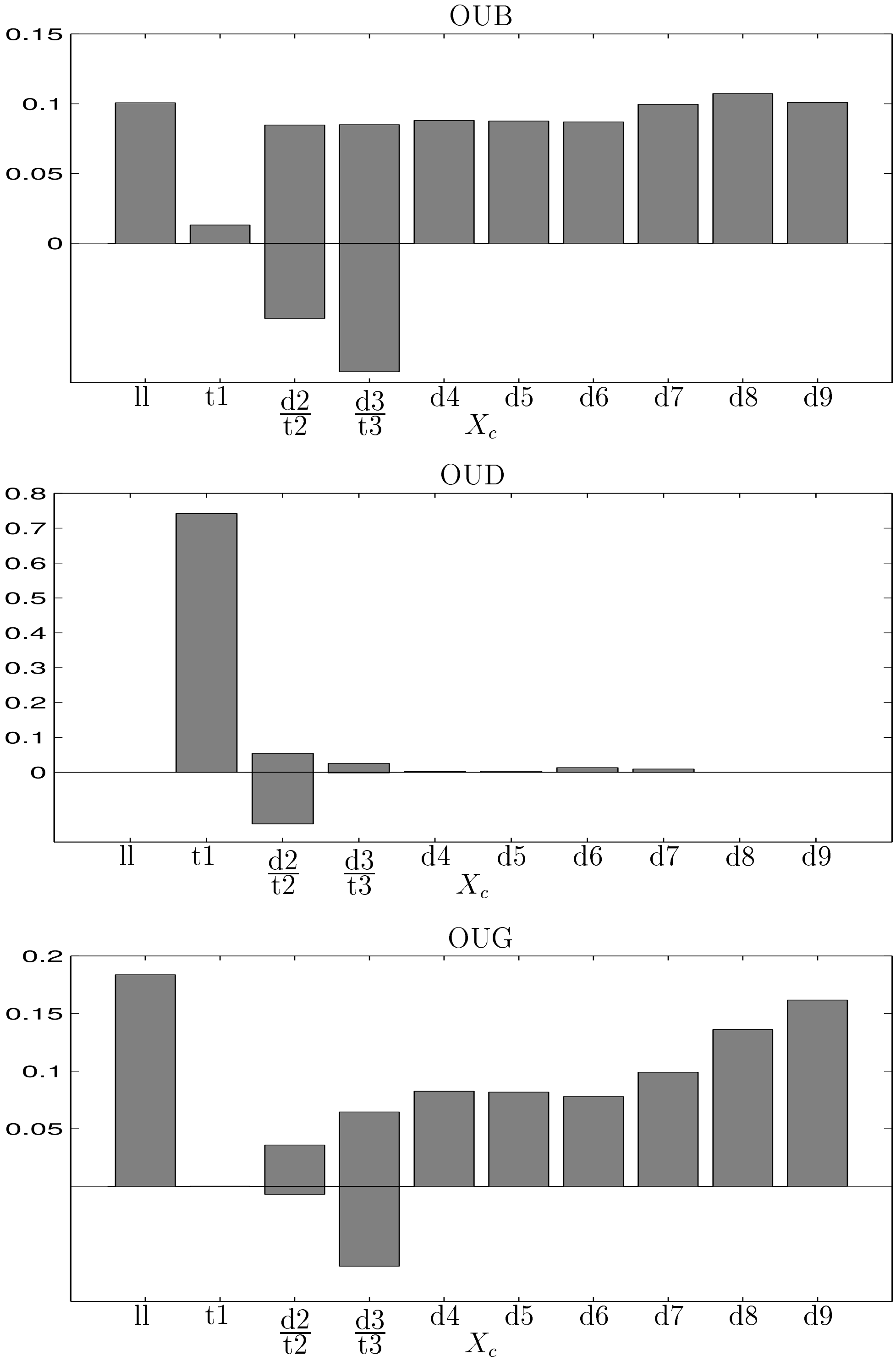

FIG. 4.20: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /u/, fibre articulatoire à E_m.

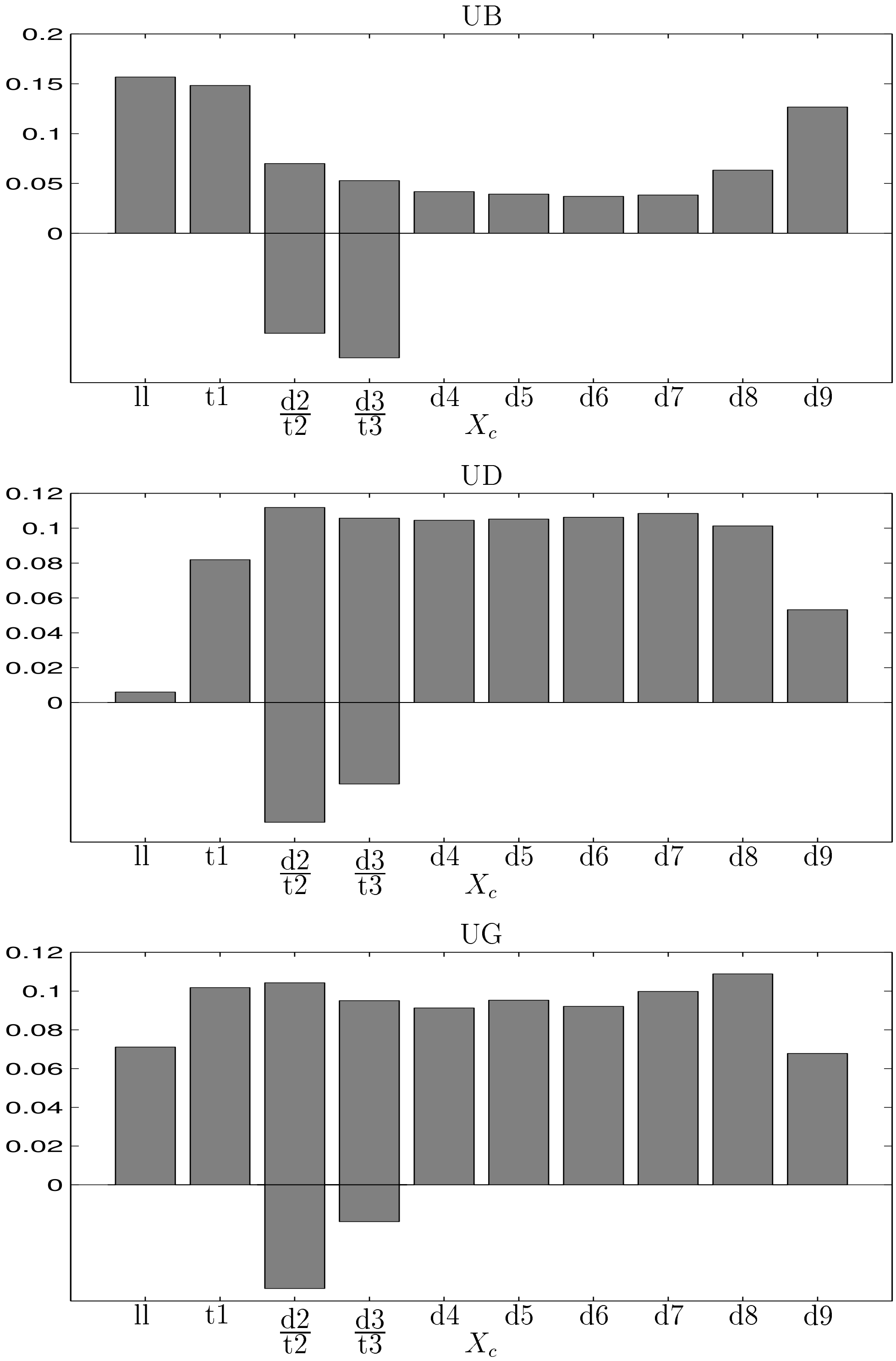

FIG. 4.21: Probabilité « acoustique » pour chaque valeur de X_c, à partir de /y/, fibre articulatoire à E_m.

malisation par distorsion, estimation de la fibre articulatoire et intégration par probabilité, a fait ses preuves du moins vis-à-vis des étapes qui la composent.

Une amélioration qui nous semble tout à fait réalisable à présent consisterait à inclure une partie visuelle correspondant à l'aire des lèvres. Ceci peut aisément être intégré dans le terme $P(X_c|a_0, I)$ de l'équation 4.8, la confusion entre « OUB » et « OUD » (cf. ci-dessus) serait tout de suite éliminée.

Chapitre 5

Conclusion

Nous avons présenté au long du présent document les travaux qui ont été entrepris en vue de faire en sorte qu'un modèle du conduit vocal arrive à inférer des commandes correspondant à des séquences voyelle-consonne présentées par de locuteurs quelconques. Nous présentons dans ce chapitre une sorte de résumé des techniques et méthodologies utilisées. Tout au long de cette thèse nous avons essayé de rester le plus élémentaire possible, bien que cela puisse sembler incompatible avec la nature même de la parole.

5.1 Le modèle articulatoire

Notre travail concernant le modèle articulatoire a consisté essentiellement à étudier les modèles déjà existant et à en choisir un surtout d'après les études dont ils avaient été l'objet.

Nous avons implémenté une version améliorée d'un modèle articulatoire ayant fait ses preuves dans plusieurs domaines (Payan 1992, Laboissière 1992, entre autres). Bien que ce modèle nous ait été très utile afin d'arriver au but que nous nous étions fixé, il a encore des limitations, et un nouveau modèle capable de décrire davantage de types de phonèmes (e.g., fricatives, rétroflexes, nasales, etc.) serait très utile dans des études similaires à la nôtre. Des modèles en trois dimensions (Wilhelms-Tricarico 1996) ainsi que d'autres incluant une modélisation des muscles (Payan 1996) ont été récemment implémentés et leur utilité nous semble plus qu'évidente.

Des techniques capables d'estimer l'aire transversale du conduit vocal avec plus de précision nous semblent aussi nécessaires, les études visant à estimer cette aire ne donnant pas de résultat en concordance les uns avec les autres (Perrier et al. 1992, Baer, Gore, Gracco & Nye 1988, Johansson, Sundberg, Wilbrand & Ytterbergh 1983).

Un point qui est resté assez obscure est celui concernant le volume de la cavité avant. Bien que le modèle ait été capable de retrouver le lieu d'articulation approprié aux constrictions où ce volume a un rôle important, notamment celles correspondant au « D » (Bailly 1995*b*), il nous semble important de réaliser une étude systématique de l'effet acoustique de cette cavité.

5.2 Normalisation

La méthode que nous avons implémentée est composée des étapes suivantes :

- Estimation des paramètres de distorsion à l'aide des formants des voyelles dites extrêmes.
- Obtention des paramètres articulatoires correspondant aux phonèmes, en utilisant le modèle articulatoire modifié selon les paramètres appropriés à chaque locuteur.
- Application des paramètres articulatoires au modèle de base afin d'obtenir les formants normalisés.

Nous considérons que cette méthodologie possède l'atout de s'adresser à la fois au problème de la longueur du conduit vocal ainsi qu'à celui de la stratégie articulatoire utilisée par chaque locuteur lors de la production des phonèmes. Résoudre ces deux problèmes était crucial étant donné le travail que nous voulions réaliser.

La méthode présente aussi une flexibilité liée au fait qu'elle est séparée en trois étapes. Si de nouvelles techniques d'estimation de la longueur du conduit vocal sont proposées, elle peuvent aisément y être intégrées.

La procédure de vérification qui nous semble logique à présent consiste à comparer les longueurs estimées avec les longueur réelles des sujets dans la base de données. Cette procédure nécessitera des techniques d'imagerie déjà utilisées lors de plusieurs études (Baer, Gore, Gracco & Nye 1991, Hassan 1988). Le cas échéant, une révision de la technique d'estimation des paramètres de distorsion s'imposera.

Un modèle forward (Jordan & Rumelhart 1992) incluant les paramètres de distorsion nous semble nécessaire afin d'accélérer l'obtention des paramètres articulatoires à partir des formants des locuteurs.

En regardant les résultats des normalisations, nous observons que la réduction des volumes des ellipsoïdes de dispersion est systématiquement moindre pour la base de données 1 que pour la base de données 2. Il se peut que ceci soit lié au fait que la première base de données présente des voyelles non-réduites alors que notre modèle a été obtenu à partir d'un sujet prononçant des phrases en continu, donc avec des voyelles réduites. Nous concluons alors qu'un modèle incluant des voyelles isolées doit être utilisé pour ce type de bases de données, et le cas écheant que la technique d'obtention des voyelles extrêmes devra être révisée.

Nous remarquons aussi que la base de données en anglais présente des réductions des volumes des ellipsoïdes plus petites que la base de données 2, bien que du même ordre que celles de la base de données 1. Le fait que cette base de données n'ait pas le même /a/ que le français est peut-être à l'origine d'une mauvaise estimation des paramètres de distortion. Notre méthode reste à être évaluée avec des bases de données d'autres langues.

5.3 Inversion des séquences

Nous avons choisi une méthodologie semblable à celle que nous avons utilisée lors de la normalisation, cela du point de vue de la séparation en parties, chacune pouvant être modifiée. Ces étapes sont les suivantes :

- Inversion de la partie vocalique afin d'obtenir une configuration articulatoire correspondant à chaque trame de formants.
- Choix des trames initiale et finale de la « séquence d'occlusion » (cf. chapitre 4, section 4.4.3).
- Obtention de la fibre articulatoire correspondant à la première trame de la séquence d'occlusion.
- Obtenir, à partir de chaque élément de la fibre, une séquence A_c pour chaque occlusion X_c possible et évaluer dans l'équation 4.8.

L'objectif de la méthode que nous avons utilisée a été avant tout de formaliser la procédure d'inversion en la séparant en différentes parties pour ensuite analyser chacune de ces parties. Nous avons implémenté chacune de ces parties d'une façon qui nous a semblé à la fois simple et logique, des techniques plus sophistiquées pouvant eventuellement y être incluses.

L'aproche probabiliste que nous avons utilisée a servi avant tout à formaliser notre méthode. Nous avons démontré en utilisant une formalisation probabiliste qu'il existe une très forte relation entre les trajectoires formantiques liées à une séquence d'occlusion et les commandes articulatoires pouvant les produire, et cela indépendamment de la configuration utilisée, pourvu que celle-ci produise les formants appropriés du début de la séquence.

Notre implémentation contient, certes, des approximations, notamment en ce qui concerne l'estimation de la distance entre ce que le robot produit et ce que le locuteur présente comme transition formantique. Nous considérons néanmoins que ces approximations sont judicieuses. Cela étant dit, l'implémentation d'une autre fonction de probabilité $P(F|A_c, I)$ quelconque peut directement remplacer celle que nous avons utilisée et les résultats immédiatement évalués.

Comme nous l'avons mentionné dans le chapitre 4, une approche multimodale (Robert-Ribes 1995) peut également être incluse dans le calcul des fonctions de probabilité.

Un des points qui reste à approfondir est celui concernant la détermination du moment de l'apparition de la séquence d'occlusion. Ce point implique une révision des theories existant sur la coarticulation (Fowler 1980). Étant donné nos résultats (la fibre articulatoire $\mathcal{A}_E$ donnant de meilleurs résultats que la fibre $\mathcal{A}_F$) nous nous penchons sur la théorie selon laquelle la coarticulation ferait partie intégrale des séquences voyelle-consonne-voyelle (Whalen 1990).

Nous estimons que le point ci-dessus nécessite d'abord une étude visant à mieux comprendre les commandes à la base des mouvements des articulateurs de la parole (e.g., des études comme celles présentées par Ostry & Keller (1983), ou celle de Saltzman & Munhall (1989)). Une intégration de ce type de commandes implique obligatoirement un modèle articulatoire capable de prendre en compte ces nouvelles commandes.

5.4 Conclusions générales

Le but que nous nous étions posé au début de ce travail de thèse est à présent considéré comme atteint. Certes, beaucoup reste à faire dans le vaste domaine de la parole avant d'être en mesure de comprendre tous les mécanismes intervenant dans le processus que nous avons appelé « répéter » en introduction du présent document, en commençant par mieux comprendre le processus d'apprentissage de la parole (Markey 1994) en passant par les méthodes visant à expliquer les relations et interactions entre les commandes sensori-motrices, la prise en compte (avec modélisation) de l'ensemble de structures (anatomiques et autres) participant à la production et perception entre autres.

Bibliographie

Abramowitz, M. et Stegun, I. A. (1972), *Handbook of Mathematical Functions*, Dover, New York.

Abry, C., Badin, P. et Scully, C. (1994), Sound-to-gesture inversion in speech : the speech maps approach, *in* K. Varghese, S. Pfleger et J. Lefèvre, eds, 'Advanced speech applications', Springer Verlag, Berlin, pp. 182–196.

Abry, C. et Boë, L. (1986), 'Laws for lips', *Speech Communication* **5**, 97–104.

Atal, B. S., Chang, J. J., Mathews, M. V. et Tukey, J. W. (1978), '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.

Badin, P. et Fant, G. (1984), 'Notes on vocal tract computation', *STL-QPSR* **2/3**, 53–108.

Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J. et Brock, G. (1995), Cineradiography of vcv sequences : Articulatory-acoustic data for a speech production model, *in* 'Proceedings of the 15^{th} International Congress on Acoustics', Vol. 4, pp. 349–352.

Baer, T., Gore, J., Gracco, L. et Nye, P. (1988), 'Vocal tract dimensions obtained from magnetic resonance images', *J. Acous. Soc. Am.* **84**(Suppl. 1), S125.

Baer, T., Gore, J., Gracco, L. et Nye, W. (1991), 'Analysis of vocal tract shape and dimensions using magnetic resonance imaging : Vowels', *J. Acous. Soc. Am* **90**(2), 799–828.

Bailly, G. (1995*a*), Characterising formant trajectories by tracking vocal tract resonances, *in* C. Sorin, J. Mariani, H. Meloni et J. Schoentgen, eds, 'Levels in Speech Communication : Relations and Interactions', Elsevier Science B.V., Amsterdam, The Netherlands, pp. 91–102.

Bailly, G. (1995*b*), Recovering place of articulation for occlusives in vcv's, *in* 'Proceedings of the XIII^{th} ICPhS', Vol. 2, pp. 230–233.

Bailly, G. (1996), Sensory-motor control of speech movements, *in* 'Proccedings : 1^{st} ETRW on Speech Production Modeling / 4^{th} Speech Production Seminar', pp. 145–154.

Beautemps, D., Badin, P. et Laboissière, R. (1995), 'Deriving vocal-tract area functions from midsaggital profiles and formant frequencies : A new model for vowels and fricative consonants base on experimental data', *Speech Communication* **16**, 27–47.

Blumstein, S., Isaacs, E. et Mertus, J. (1982), 'The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants', *J. Acous. Soc. Am.* **72**(1), 43–50.

Blumstein, S. et Stevens, K. (1979), 'Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants', *J. Acous. Soc. Am.* **66**(4), 1001–1017.

Blumstein, S. et Stevens, K. (1980), 'Perceptual invariance and onset spectra for stop consonants in different vowel environments', *J. Acous. Soc. Am.* **67**(2), 648–662.

Boë, L.-J., Perrier, P. et Bailly, G. (1992*a*), 'The geometric vocal tract variables controlled for vowel production : Proposals for constraining acoustic-to-articulatory inversion', *Journal of Phonetics* **20**(1), 27–38.

Boë, L., Perrier, P. et Bailly, G. (1992*b*), 'The geometric vocal tract variables controlled for vowel production : Proposal for constraining acoustic-to-articulatory inversion', *Journal of Phonetics* **20**, 27–38.

Boë, L., Perrier, P., Guérin, B. et Schwartz, J. (1989), Maximal vowel space, *in* 'Proceedings of the 1^{st} European Conference on Speech Communication and Technology', Vol. 2, pp. 281–284.

Boë, L., Schwartz, J., Laboissière, R. et Vallée, N. (1996), Integrating articulatory-acoustic constraints in the prediction of sound structures, *in* 'Proccedings : 1^{st} ETRW on Speech Production Modeling / 4^{th} Speech Production Seminar', pp. 163–166.

Botherel, A., Simon, P., Wioland, F. et Zerling, J. (1986), *Cinéradiographie des voyelles et des consonnes du français. Recueil de documents synchronisés pour quatre sujets : Vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques*, Institut de Phonétique de Strasbourg, Strasbourg, France.

Bouabana, S. (1995), Modélisation des mouvements articulatoires de la langue par la méthode de la LPC multi-impulsionnelle, PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France.

Calliope (1989), *La parole et son traitement automatique*, Masson, Paris, France. ouvrage collectif édité par J.-P. Tubach.

Coker, C. et Fujimora, O. (1966), 'Model for specification of the vocal tract area function', *J. Acouys. Soc. Am.* **40**(5), 1271.

Eek, A. et Meister, E. (1995), The perception of stop consonants : Locus equations and spectral integration, *in* 'Proceedings of the XIII^{th} ICPhS', Vol. 1, pp. 18–21.

Fant, G. (1969), 'Stops in cv-syllables', *STL-QPSR* **4**, 1–25.

Fant, G. (1966), 'A note on vocal tract size factors and non-uniform F-pattern scalings', *STL-QPSR* **4**, 22–30.

Feng, G., Abry, C. et Guérin, B. (1986), The nasopharyngeal tract : A target for nasality. acoustic simulations vs. sweep tone measurements, *in* 'Proceedings of the 12^{th} International Congress on Acoustics', pp. A3–8.

Ferrari-Disner, S. (1980), 'Evaluation of vowel normalization procedures', *J. Acous. Soc. Am.* **67**(1), 253–261.

Flanagan, J. L. (1972), *Speech Analysis Synthesis and Perception*, second edn, Springer-Verlag, Berlin.

Fowler, C. (1980), 'Coarticulation and theories of extrinsic timing', *Journal of Phonetics* **8**, 113–133.

Fujimora, O. et Lindqvist, J. (1971), 'Sweep-tone measurements of vocal tract characteristics', *J. Acous. Soc. Am.* **49**(2), 541–558.

Gabioud, B. (1994), Articulatory models in speech synthesis, *in* 'Fundamentals of Speech Synthesis and Speech Recognition', John Wiley & Sons, Chichester, England, pp. 215–230.

Glotin, H. (1995), 'La vie artificielle d'une société de robots parlants : Émergence et changements du code phonétique', Rapport de Stage du D.E.A de l'I.N.P.G.

Harshman, R. et Papcun, G. (1976), 'Vowel normalization by linear transformation of each speaker's acoustic space', *J. Acous. Soc. Am.* **59**(S1), S71.

Hassan, O. (1988), Quelques aspects de l'utilisation des données radiographiques : Extraction automatique des fonctions d'aire, macro-sensibilités acoustiques et transitions voyelle-voyelle, PhD thesis, INPG.

Heinz, J. et Stevens, K. (1965), On the relations between lateral cineradiographs, area functions and acoustic spectra of speech, *in* 'Actes du $5^{\text{ème}}$ Congrès International d'Acoustique', pp. 1–4.

Itakura, F. (1975), 'Minimum prediction residual principle applied to speech recognition', *IEEE Transactions on Acoustics Speech and Signal Processing* **ASSP-23**(1), 67–72.

Jaynes, E. (1997), 'Probability theory : The logic of science', ouvrage non-publié, disponible en version éléctronique à l'URL : `http://omega.albany.edu:8008/JaynesBook.html`.

Johansson, C., Sundberg, C., Wilbrand, H. et Ytterbergh, C. (1983), From sagittal distance to area, *in* 'STL-QPSR', Vol. 4, pp. 39–49.

Jordan, M. et Rumelhart, D. (1992), 'Forward models :supervised learning with a distal teacher', *Cognitive Science* **16**, 307–354.

Kewley-Port, D. (1982), 'Measurement of formant transitions in naturally produced stop consonant-vowel syllables', *J. Acous. Soc. Am.* **72**(2), 379–389.

Kiritani, S., Itioh, K. et Fujimura, O. (1975), 'Tongue-pellet tracking by a computer controlled x-ray microbeam system', *J. Acous. Soc. Am.* **57**(6), 1516–1520.

Laboissère, R. et Galván, A. (1995), Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences, *in* 'Proceedings of the XIII^{th} ICPhS', Vol. 1, pp. 358–361.

Laboissière, R. (1992), Préliminaires pour une robotique de la communication parlée : Inversion et contrôle d'un modèle articulatoire du conduit vocal, PhD thesis, INPG.

Lahiri, A. et Blumstein, S. (1981), 'A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants : Evidence from a cross-language study', *J. Acous. Soc. Am.* **70**(S1), S39.

Lahiri, A., Gewirth, L. et Blumstein, S. (1984), 'A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants : Evidence from a cross-language study', *J. Acous. Soc. Am.* **76**(2), 391–404.

Lin, Q. et Fant, G. (1989), 'Vocal tract area function parameters from formant frequencies', *Proceedings of the European Conference on Speech Communication and Technology* **2**, 673–676.

Lindblom, B. et Sundberg, J. (1971), 'Acoustical consequences of lip, tongue, jaw, and larynx movement', *J. Acous. Soc. Am.* **50**(2), 1166–1179.

Lobanov, B. (1971), 'Classification of russian vowels spoken by different speakers', *J. Acous. Soc. Am.* **49**(2), 606–608.

Loevenbruck, H. et Perrier, P. (1993), Vocalic reduction : Prediction of acoustic and articulatory variabilities with invariant motor commands, *in* 'Proceedings of the 3^{rd} European Conference on Speech Communication and Technology', Vol. 1, pp. 85–88.

Maeda, S. (1979), Un modèle articulatoire de la langue avec des composantes linéaires, *in* 'Actes des $X^{èmes}$ Journées d'Étude sur la Parole', pp. 152–162.

Maeda, S. (1984), Une paire de pics spectraux comme correlat acoustique de la nasalisation des voyelles, *in* 'Actes des $XIII^{èmes}$ Journées d'Étude sur la Parole', pp. 223–224.

Maeda, S. (1988), 'Improved articulatory model', *J. Acous. Soc. Am.* **81**(S1), S146.

Majid, R. (1986), Modélisation articulatoire du conduit vocal : Exploration et exploitation, fonctions de macro-sensibilité paramétriques et voyelles du français, PhD thesis, INPG.

Majid, R., Abry, C., Boë, L. et Perrier, P. (1987), Contribution à la classification articulatori-acoustique des voyelles : Étude des macro-sensibilités à l'aide d'un modèle articulatoire, *in* 'Proceedings of the XI^{th} ICPhS', Vol. 2, pp. 348–351.

Markey, K. (1994), The sensorimotor foundations of phonology : A computational model of early childhood articulatory and phonetic development, PhD thesis, University of Colorado at Boulder, Departement of Computer Science.

Matsumoto, H. et Wakita, H. (1986), 'Vowel normalization by frequency warped spectral matching', *Speech Communication* **5**(2), 239–251.

Mattingly, I. (1966), 'Speaker variation and vocal tract size', *J. Acous. Soc. Am.* **39**(6), 1219.

Neagu, A. (1994), 'Mise en œuvre des traitement distribués pour sidoc. application à la caractérisation des espaces vocaliques en termes de résonances du conduit vocal', Rapport de Stage du D.E.A. de l'I.N.P.G.

Neal, R. (1993), Probabilistic inference using markov chain monte carlo methods, Technical Report CRG-TR-93-1, Departement of Computer Science, University of Toronto.

Nordström, P. (1977), 'Female and infant vocal tracts simulated from male area functions', *Journal of Phonetics* **5**, 81–92.

Öhman, S. (1966), 'Coarticulation in vcv utterances : Spectrographic measurements', *J. Acous. Soc. Am.* **39**(1), 151–168.

Öhman, S. (1967), 'Numerical model of coarticulation', *J. Acous. Soc. Am.* **41**(2), 310–320.

Ostry, D. et Keller, E. (1983), 'Similarities in the control of the speech articulators and the limbs : Kinematics of tongue dorsum movement in speech', *Journal of Experimental Psychology : Human Perception and Performance* **9**(4), 622–636.

Overall, J. (1962), 'Orthogonal factors and uncorrelated factor scores', *Psychological Report* **10**, 651–662.

Payan, Y. (1992), '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 Stage du D.E.A de l'I.N.P.G.

Payan, Y. (1996), Modèle biomécanique et contrôle de la langue en parole, PhD thesis, INPG.

Payan, Y., Apostol, L. et Perrier, P. (1995), Evaluation of a vowel normalisation procedure based on speech production knowledge, *in* 'Proceedings of the 4^{th} European Conference of Speech Communication and Technology', Vol. 3, pp. 1925–1928.

Payan, Y. et Perrier, P. (1993), Vowel normalization by articulatory normalization : First attempt for vowel transitions, *in* 'Proceedings of the 3^{rd} European Conference on Speech Communication and Technology', Vol. 1, pp. 417–420.

Perkell, J. et Nelson, W. (1985), 'Variability in production of the vowels /i/ and /a/', *J. Acous. Soc. Am.* **77**(5), 1889–1895.

Perrier, P., Boë, L. et Sock, R. (1992), 'Vocal tract function estimation from midsagittal dimensions with ct scans and a vocal tract cast : Modeling the transition with two sets of coefficients', *Journal of Speech and Hearing Research* **35**, 53–67.

Peterson, G. et Barney, H. (1952), 'Control methods used in a study of vowels', *J. Acous. Soc. Am.* **20**, 528–535.

Piquemal, M., Schwartz, J., Berthommier, F., Lallouache, T. et Escudier, P. (1996), Détection et localisation auditive et visuelle d'explosions consonantiques dans des séquences vcv bruitées, *in* 'Actes des XXI^{èmes} Journées d'Étude sur la Parole', pp. 355–358.

Rabiner, L., Rosenberg, A. et Levinson, S. (1978), 'Considerations in dynamic time warping algorithms for discrete word recognition', *IEEE Transactions on Acoustics, Speech and Signal Proccessing* **ASSP-26**(6), 575–582.

Robert-Ribes, J. (1995), Modèles d'intégration audiovisuelle de signaux linguistiques, PhD thesis, INPG.

Saerens, M., Soquet, A. et Jospa, P. (1991), 'Trying to determine place of articulation of plosives with a vocal tract model', *XIIth International Congress of Phonetic Sciences* **2**, 66–69.

Sakoe, H. et Chiba, S. (1978), ‘Dynamic programming algorithm optimization for spoken word recognition’, *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-26**(1), 43–49.

Saltzman, E. et Munhall, K. (1989), ‘A dynamical approach to gestural patterning in speech production’, *Ecological Psychology* **1**(4), 333–382.

Sanchez, H. et Boë, L. (1984), De la coupe sagittale à la fonction d’aire du conduit vocal, *in* ‘Actes des XIII^{èmes} Journées d’Étude sur la Parole’, pp. 23–25.

Scheideman, G., Bell, W., Leagan, H., Finn, R. et Reisch, J. (1980), ‘Cephalometric analysis of dentofacial normals’, *Am. J. Orthod.* **78**(4), 404–420.

Schoentgen, J. et Ciocea, S. (1995), Cinematic acoustic-to-geometric mapping, *in* ‘Proceedings of the XIII$^{\mathrm{th}}$ ICPhS’, Vol. 2, Stockholm, Sweden, pp. 194–197.

Schroeter, J. et Sondhi, M. M. (1994), ‘Techniques for estimating vocal tract shapes from the speech signal’, *IEEE Transactions on Speech and Audio Processing* **2**(1), 133–150.

Schwartz, J., Boë, L. et Vallée, N. (1995), Testing the dispersion-focalization theory : Phase spaces for vowel systems, *in* ‘Proceedings of the XIIIth ICPhS’, Vol. 1, pp. 412–415.

Schwartz, J., Boë, L., Vallée, N. et Abry, C. (1997), ‘The dispersion - focalisation theory of vowel systems’, *Journal of Phonetics* .

Sorokin, V. N. (1992), ‘Determination of vocal tract shape for vowels’, *Speech Communication* **11**, 71–85.

Stevens, K. (1989), ‘On the quantal nature of speech’, *Journal of Phonetics* **17**, 3–45.

Stevens, K. et Blumstein, S. (1978), ‘Invariant cues for place of articulation’, *J. Acous. Soc. Am.* **64**(5), 1358–1368.

Sundberg, J. et Lindblom, B. (1990), ‘Acoustic estimations of the front cavity in apical stops’, *J. Acous. Soc. Am.* **88**(3), 1313–1317.

Sussman, H., Hoemeke, K. et Ahmed, F. (1993), ‘A cross- linguistic investigation of locus equations as a phonetic descriptor for place of articulation’, *J. Acous. Soc. Am.* **94**(3), 1256–1268.

Sussman, H., McCaffrey, H. et Mattews, S. (1991), ‘An investigation of locus equations as a source of relational invariance for stop place categorization’, *J. Acous. Soc. Am.* **90**(3), 1309–1325.

Verbrugge, R.R.and Strange, W., Shankweiler, D. et Edman, T. (1976), ‘What information enables a listener to map a talkers vowel space ?’, *J. Acous. Soc. Am.* **60**(1), 198–212.

Wakita, H. (1977), ‘Normalization of vowels by vocal-tract length and its application to vowel identification’, *IEEE Transactions on Acoustics, Speech and Signal Processing* **25**(2), 183–192.

Whalen, D. (1990), ‘Coarticulation is largely planned’, *Journal of Phonetics* **18**, 3–35.

Wilhelms-Tricarico, R. (1996), ‘A biomechanical and physiologically-base vocal tract model and its control’, *Journal of Phonetics* **24**, 23–38.

Zahorian, S. et Jagharghi, A. (1991), 'Speaker normalization of static and dynamic vowel spectral features', *J. Acous. Soc. Am.* **90**(1), 67–75.

Zwicker, E. et Terhardt, E. (1980), 'Analytical expressions for critical-band rate and critical-band width as a function of frequency', *J. Acous. Soc. Am.* **68**(5), 1523–1525.

Annexe A

Obtention des Paramètres de Distorsion

Cette annexe décrit la façon dont nous avons obtenu les paramètres de distorsion longitudinale d_o, d_p à partir des formants des voyelles /i/, /a/, /y/, et /u/ produits par un locuteur quelconque.

A.1 Présentation

Soient :

- A_i, A_a, A_y, A_u – les configurations articulatoires correspondant aux voyelles /i/, /a/, /y/, /u/ respectivement[1].
- $F_{i(o,p)}$, $F_{a(o,p)}$, $F_{y(o,p)}$, $F_{u(o,p)}$ – les trois premiers formants produits à partir des configurations articulatoires A_i, A_a, A_y, A_u, respectivement, par le robot parlant ayant subi une distorsion avec $d_o = o$, $d_p = p$ (cf. chapitre 3).
- S_i, S_a, S_y, S_u – les trois premiers formants des voyelles /i/, /a/, /y/, /u/ respectivement, produits par un locuteur S quelconque.

Nous pouvons associer à chaque point $P = [o\,,\, p]$ dans l'espace des paramètres d_o , d_p, un point $\mathcal{P} = [F_{i(o,p)}\,,\, F_{a(o,p)}\,,\, F_{y(o,p)}\,,\, F_{u(o,p)}]$ correspondant aux formants des quatre voyelles produites par le robot parlant ayant subi la distorsion correspondante. Nous appellerons dorénavant l'espace des paramètres d_o , d_p « $\mathbb{R}^2$ » et l'espace des trois premiers formants de chaque voyelle « $\mathbb{R}^n$ »[2].

L'ensemble des points $\mathcal{P} \in \mathbb{R}^n$ que le robot est capable de produire décrit une surface (dimension 2) en $\mathbb{R}^n$. Chaque point de cette surface correspond à une paire de paramètres d_o , d_p. Nous cherchons à trouver pour chaque locuteur S le point $P_S \in \mathbb{R}^2$ tel que le

[1] Ces configurations ont été obtenues de la base de données originale.

[2] Pour le cas que nous sommes en train d'analyser, $n = 12$ (4 voyelles $\times$ 3 formants chacune) ; néanmoins, ce choix n'étant pas unique (cf. chapitre 3) nous laisserons la dimension n des points $\mathcal{P}$ variable.

point correspondant $\mathcal{P}_S \in \mathbb{R}^n$ faisant partie de cette surface s'approche le plus possible des formants $\mathcal{S} = [S_i\ ,\ S_a\ ,\ S_y\ ,\ S_u]$ du locuteur en question.

La surface en $\mathbb{R}^n$ étant liée aux points correspondants en $\mathbb{R}^2$ de façon non-linéaire, nous procéderons à l'échantillonnage de l'espace $\mathbb{R}^2$ de façon hexagonale (cf. figure A.1, gauche) afin d'associer à chaque ensemble de trois points $P_i\ ,\ Q_i\ ,\ R_i\ \in \mathbb{R}^2$, un plan (dimension 2) $\mathcal{T}_i \in \mathbb{R}^n$ passant par le triangle $\mathcal{PQR}_i$ délimité par les points correspondants $\mathcal{P}_i, \mathcal{Q}_i, \mathcal{R}_i\ \in \mathbb{R}^n$. Ce plan décrira la partie (localement linéaire) de la surface délimitée par le triangle $\mathcal{PQR}_i$ (cf. figure A.1, droite).

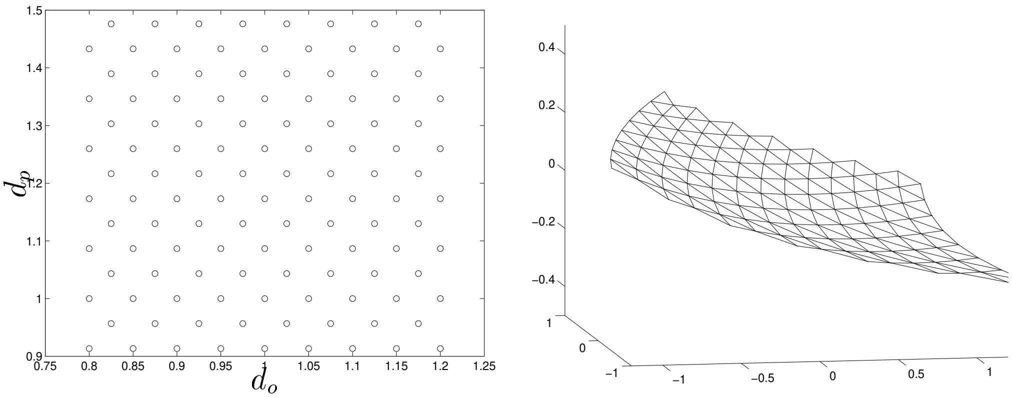

FIG. A.1: Projection des points $\in \mathbb{R}^2$ en leur point correspondant $\in \mathbb{R}^n$, $(n = 3)$.

Pour chaque plan $\mathcal{T}_i$ correspondant au triangle $\mathcal{PQR}_i$ nous associons un point $\mathcal{S}_i$ de la façon suivante :

- Si la projection $\mathcal{S}'$ du point $\mathcal{S}$ sur le plan $\mathcal{T}_i$ est à l'intérieur du triangle $\mathcal{PQR}_i$, alors $\mathcal{S}_i = \mathcal{S}'$ (cf. figure A.2).
- Si la projection $\mathcal{S}'$ sur le plan est en dehors du triangle (cf. figure A.3), nous obtiendrons la projection du point $\mathcal{S}$ sur chacune des arêtes du triangle. Si une ou plusieurs de ces projections sont dans le triangle (sur la figure A.3 une de ces projections est en dehors et les deux autres sont dedans) $\mathcal{S}_i$ sera la projection la plus proche du point $\mathcal{S}$.
- Si aucune des projections sur les arêtes du triangle n'est dans ce dernier, alors $\mathcal{S}_i$ sera le point le plus proche de $\mathcal{S}$ parmi les points $\mathcal{P}_i\ ,\ \mathcal{Q}_i\ ,\ \mathcal{R}_i$, délimitant le triangle.

Le point $\mathcal{P}_S$ recherché sera alors le point $\mathcal{S}_i$ le plus proche du point $\mathcal{S}$. Trois aspects restent à résoudre :

1. Trouver la projection d'un point $\mathcal{S}$ sur le plan passant par trois points $\mathcal{P}\ ,\ \mathcal{Q}\ ,\ \mathcal{R}$ quelconques en $\mathbb{R}^n$ et en outre déterminer si cette projection est dans le triangle défini par ces trois points.

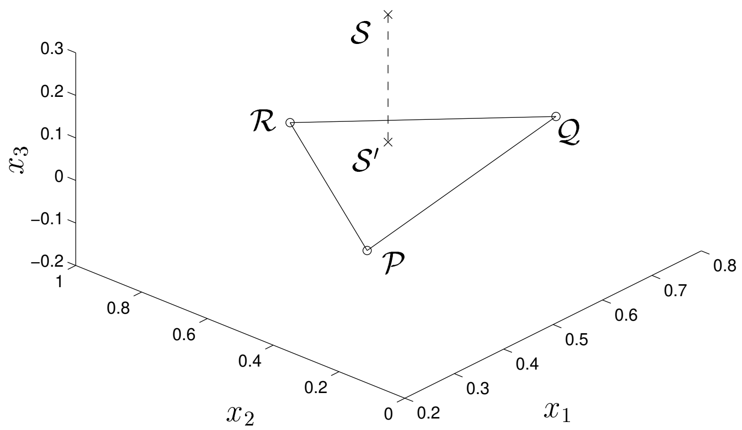

FIG. A.2: Projection du point $\mathcal{S}$ sur un plan.

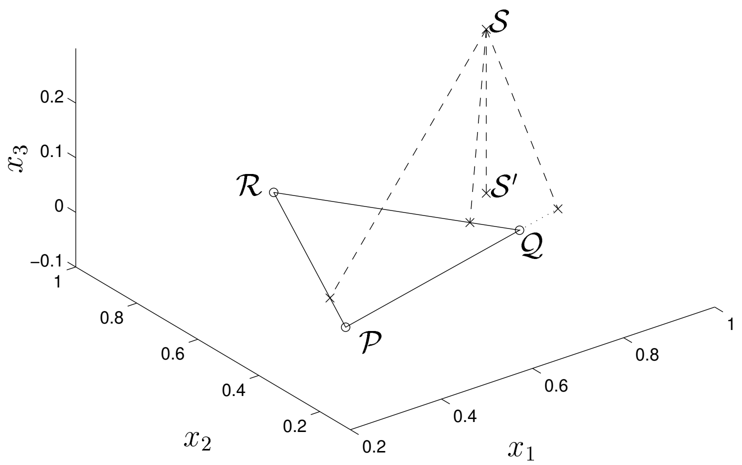

FIG. A.3: Projection du point $\mathcal{S}$ sur les arêtes d'un triangle.

2. Trouver la projection d'un point $\mathcal{S}$ sur une ligne passant par deux points quelconques en $\mathbb{R}^n$ et déterminer si cette projection est dans le segment de ligne délimitée par ces deux points.
3. À partir d'un point en $\mathbb{R}^n$ trouver le point correspondant en $\mathbb{R}^2$ alors que les correspondances entre les points en $\mathbb{R}^2$ et ceux en $\mathbb{R}^n$ ne sont définies que pour les points d'une maille hexagonale.

A.2 Projection d'un point sur un plan en $\mathbb{R}^n$

Nous cherchons à trouver une façon appropriée de représenter un plan en $\mathbb{R}^n$ à partir de trois points quelconques pour ensuite trouver la projection d'un point $\mathcal{S}$ sur ce plan. Soient $[x_1 \, \dots \, x_n]$ les coordonnées d'un point quelconque en $\mathbb{R}^n$, nous représenterons un plan en $\mathbb{R}^n$ de la façon décrite par l'équation A.1, où chacune des dernières $n-2$ dimensions est une fonction linéaire des deux premières.

$$\begin{bmatrix} x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} A_3 & B_3 \\ A_4 & B_4 \\ \vdots & \vdots \\ A_n & B_n \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} C_3 \\ C_4 \\ \vdots \\ C_n \end{bmatrix} \tag{A.1}$$

Soient $\mathcal{P} = [p_1 \dots p_n]$, $\mathcal{Q} = [q_1 \dots q_n]$ et $\mathcal{R} = [r_1 \dots r_n]$ les points à partir desquels nous cherchons à représenter un plan en $\mathbb{R}^n$. Les coefficients $A_3 \dots A_n$, $B_3 \dots B_n$, $C_3 \dots C_n$ doivent être tels que chacun des plans représentés par les lignes de l'équation ci-dessus passe par ces trois points. Nous trouvons donc les coefficients $A_i \, , B_i \, , C_i \quad \forall \; i \in [3 \dots n]$ en résolvant l'équation A.2.

$$\begin{bmatrix} p_i \\ q_i \\ r_i \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & 1 \\ q_1 & q_2 & 1 \\ r_1 & r_2 & 1 \end{bmatrix} \times \begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix} \tag{A.2}$$

Soit $\mathcal{S} = [s_1 \dots s_n]$ un point en $\mathbb{R}^n$, la projection $\mathcal{S}' = [s'_1 \dots s'_n]$ de $\mathcal{S}$ sur le plan décrit par l'équation A.1, est telle que la distance entre $\mathcal{S}$ et $\mathcal{S}'$ est minimale. Nous trouverons les coordonnées de S' en minimisant le carré $\mathbb{D}$ de la distance entre le point $\mathcal{S}$ et le plan décrit par l'équation A.1 (cf. équation A.3).

$$\mathbb{D} = (x_1 - s_1)^2 + (x_2 - s_2)^2 + \sum_{i=3}^{n} (x_n - s_n)^2 \tag{A.3}$$

Où : $x_i = A_i x_1 + B_i x_2 + C_i \quad \forall\, i \in [3 \ldots n]$

Le minimum de l'équation A.3 se trouve dans le point où toutes les dérivés partielles $\frac{\partial \mathbb{D}}{\partial x_i}$ sont égales à zéro $\forall\, i, \in [1 \ldots n]$. Puisque seules les variables x_1 et x_2 sont indépendantes, nous n'aurons à résoudre que deux équations (cf. équation A.4).

$$\begin{bmatrix} \frac{\partial \mathbb{D}}{\partial x_1} \\ \frac{\partial \mathbb{D}}{\partial x_2} \end{bmatrix} = \begin{bmatrix} (x_1 - s_1) + \sum_{i=3}^{n} (A_i x_1 + B_i x_2 + C_i - s_i) A_i \\ (x_2 - s_2) + \sum_{i=3}^{n} (A_i x_1 + B_i x_2 + C_i - s_i) B_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{A.4}$$

La solution de l'équation ci-dessus est donnée par l'équation A.5.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 + \sum_{i=3}^{n} A_i^2 & \sum_{i=3}^{n} A_i B_i \\ \sum_{i=3}^{n} A_i B_i & 1 + \sum_{i=3}^{n} B_i^2 \end{bmatrix}^{-1} \times \begin{bmatrix} s_1 + \sum_{i=3}^{n} (s_i - C_i) A_i \\ s_2 + \sum_{i=3}^{n} (s_i - C_i) B_i \end{bmatrix} \tag{A.5}$$

En remplaçant x_1, x_2 dans l'équation A.1, nous trouvons les coordonnées $[x_1, x_2 \ldots x_n]$ du point $\mathcal{S}'$.

Nous avons trouvé la projection $\mathcal{S}'$ d'un point $\mathcal{S}$ quelconque sur un plan passant par trois points $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$. Nous décrivons à présent une façon pour déterminer si le point $\mathcal{S}'$ se trouve dans le triangle $\mathcal{PQR}$.

Étant donné que le triangle $\mathcal{PQR}$ et le point $\mathcal{S}'$ sont tous les deux dans le même plan, nous pouvons affirmer que la projection de l'ensemble triangle $\mathcal{PQR}$ – point $\mathcal{S}'$ sur un plan quelconque gardera la même topologie ; c'est-à-dire, la projection du point sera dans la projection du triangle si et seulement si le point est dans le triangle. Afin d'éviter des calculs, nous délimiterons l'espace de travail à $[x_1\,,\, x_2\,,\, x_3]$ et nous choisirons le plan $x_3 = 0$. Les points sur ce nouveau plan seront alors $\mathcal{P}_p = [p_1, p_2, 0]$, $\mathcal{Q}_p = [q_1, q_2, 0]$, $\mathcal{R}_p = [r_1, r_2, 0]$ et $\mathcal{S}'_p = [s'_1, s'_2, 0]$.

Soient les vecteurs : $\mathcal{V}_{QP} = \mathcal{Q}_p - \mathcal{P}_p$, $\mathcal{V}_{RQ} = \mathcal{R}_p - \mathcal{Q}_p$, $\mathcal{V}_{PR} = \mathcal{P}_p - \mathcal{R}_p$, les trois vecteurs « faisant le tour » du triangle et soient les vecteurs $\mathcal{V}_{SP} = \mathcal{S}'_p - \mathcal{P}_p$, $\mathcal{V}_{SQ} = \mathcal{S}'_p - \mathcal{Q}_p$, $\mathcal{V}_{SR} = \mathcal{S}'_p - \mathcal{R}_p$, les trois vecteurs allant de chacun des points délimitant le triangle au point $\mathcal{S}'_p$ (cf. figure A.4). Le point $\mathcal{S}'_p$ est dans le triangle si et seulement si les composants x_3 de tous les produits vectoriels $\mathcal{V}_{QP} \times \mathcal{V}_{SQ}$, $\mathcal{V}_{RQ} \times \mathcal{V}_{SR}$ et $\mathcal{V}_{PR} \times \mathcal{V}_{SP}$ sont du même signe.

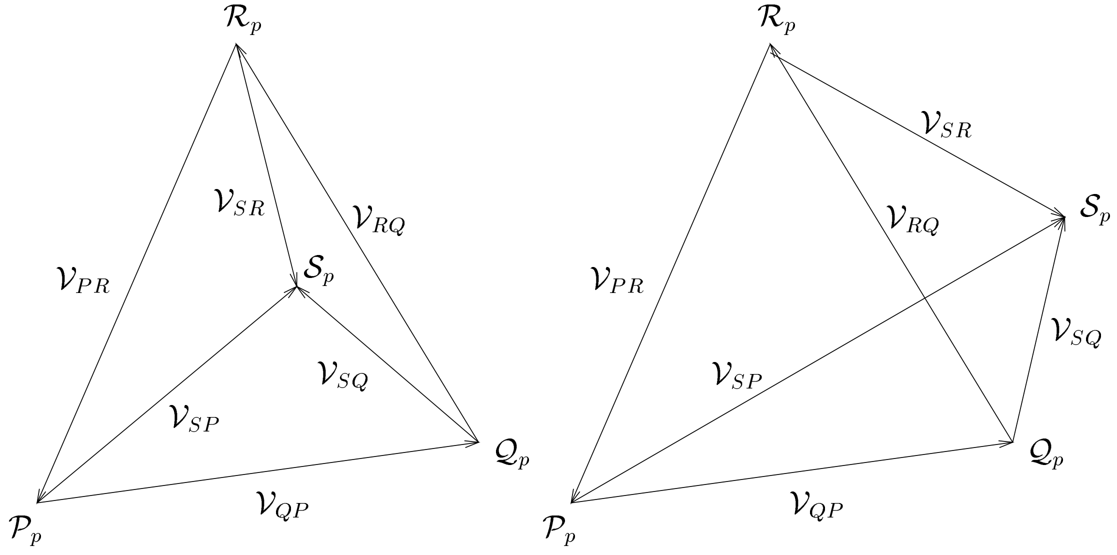

FIG. A.4: Point $\mathcal{S}_p$ à l'intérieur (gauche) et à l'extérieur (droite) du triangle.

A.3 Projection d'un point sur une ligne en $\mathbb{R}^n$

Nous avons trouvé dans la section précédente la façon d'obtenir la projection d'un point sur un plan en $\mathbb{R}^n$. Si le point projeté n'est pas dans le triangle formé par les trois points définissant le plan, il nous faut trouver la projection du même point sur une ligne pour ensuite vérifier si le point projeté est dans la partie ce cette ligne comprise entre les deux points donnés.

Nous représenterons une ligne en $\mathbb{R}^n$ d'une façon semblable à celle que nous avons utilisée dans la section précédente. L'équation A.6 montre les $n-1$ dernières dimensions comme étant une fonction linéaire de la première :

$$\begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} A_2 \\ A_3 \\ \vdots \\ A_n \end{bmatrix} \times \begin{bmatrix} x_1 \end{bmatrix} + \begin{bmatrix} C_2 \\ C_3 \\ \vdots \\ C_n \end{bmatrix} \tag{A.6}$$

Soient $\mathcal{P} = [p_1 \ldots p_n]$ et $\mathcal{Q} = [q_1 \ldots q_n]$ les points à partir desquels nous cherchons à représenter une ligne en $\mathbb{R}^n$. Les coefficients $A_2 \ldots A_n$, $C_2 \ldots C_n$ doivent être tels que chacune des lignes représentées par les lignes de l'équation ci-dessus passe par ces deux points. Nous trouvons donc les coefficients $A_i, C_i \quad \forall\ i \in [2 \ldots n]$ en résolvant l'équation A.7.

$$\begin{bmatrix} p_i \\ q_i \end{bmatrix} = \begin{bmatrix} p_1 & 1 \\ q_1 & 1 \end{bmatrix} \times \begin{bmatrix} A_i \\ C_i \end{bmatrix} \tag{A.7}$$

De la même façon, soit $\mathcal{S} = [s_1 \dots s_n]$ un point en $\mathbb{R}^n$, la projection $\mathcal{S}' = [s'_1 \dots s'_n]$ de $\mathcal{S}$ sur la plan décrit par l'équation A.6, est tel que la distance entre $\mathcal{S}$ et $\mathcal{S}'$ est minimale. Nous trouverons les coordonnées de S' en minimisant le carré $\mathbb{D}$ de la distance entre le point $\mathcal{S}$ et la ligne décrite par l'équation A.6 (cf. équation A.8).

$$\mathbb{D} = (x_1 - s_1)^2 + \sum_{i=2}^{n} (x_n - s_n)^2 \tag{A.8}$$

Où : $x_i = A_i x_1 + B_i x_2 + C_i \quad \forall\, i \in [2 \dots n]$

Le minimum de l'équation A.8 se trouve dans les points où toutes les dérivés partielles $\frac{\partial \mathbb{D}}{\partial x_i}$ sont égales à zero $\forall\, i \in [1 \dots n]$. Puisque seule la variable x_1 est indépendente, nous n'aurons à résoudre qu'une équation (cf. équation A.9).

$$\frac{\partial \mathbb{D}}{\partial x_1}(x_1 - s_1) + \sum_{i=2}^{n} (A_i x_1 + B_i x_2 + C_i - s_i) A_i = 0 \tag{A.9}$$

La solution de l'équation ci-dessus est donnée par l'équation A.10.

$$x_1 = \frac{s_1 + \displaystyle\sum_{i=2}^{n} (s_i - C_i) A_i}{1 + \displaystyle\sum_{i=2}^{n} A_i^2} \tag{A.10}$$

En remplaçant x_1 dans l'équation A.6, nous trouvons les coordonnées $[x_1, x_2 \dots x_n]$ du point $\mathcal{S}'$.

Soient $\mathcal{D}_{SP}$, $\mathcal{D}_{SQ}$ les distances entre le point projeté $\mathcal{S}'$ et les points $\mathcal{P}$, $\mathcal{Q}$ respectivement, et soit $\mathcal{D}_{PQ}$ la distance entre les deux points $\mathcal{P}$, $\mathcal{Q}$, le point $\mathcal{S}'$ sera dans la partie de la ligne délimitée par les points $\mathcal{P}$, $\mathcal{Q}$ si et seulement si $\mathcal{D}_{PQ} = \mathcal{D}_{SP} + \mathcal{D}_{SQ}$.

A.4 Obtention du point correspondant en $\mathbb{R}^2$ à partir du point en $\mathbb{R}^n$

Nous avons à présent une façon d'obtenir parmi les points en $\mathbb{R}^n$ que le robot parlant peut produire, celui qui s'approche le plus d'un point $\mathcal{S}$ produit par un locuteur quelconque. Le pas final consiste à déduire le point S en $\mathbb{R}^2$ correspondant au point $\mathcal{S}'$.

Nous savons que le point $\mathcal{S}'$ calculé est lié au triangle $\mathcal{PQR}$. Soient : PQR le triangle correspondant en $\mathbb{R}^2$; PQ le point en $\mathbb{R}^2$ à mi-chemin entre les points P et Q ; et $\mathcal{PQ}$, le point correspondant en $\mathbb{R}^n$ (cf. figure A.5).

Les vecteurs $\vec{V}_1$, $\vec{V}_2$ (unitaires et orthogonaux) sur la partie gauche de la figure A.5, correspondent aux vecteurs $\vec{\mathcal{V}}_1$, $\vec{\mathcal{V}}_2$ sur la partie droite de la même figure. Nous décrirons le point $\mathcal{S}'$ par rapport aux vecteurs $\vec{\mathcal{V}}_1$, $\vec{\mathcal{V}}_2$ afin de trouver ensuite le point S recherché par rapport aux vecteurs $\vec{V}_1$, $\vec{V}_2$ correspondants.

Soient A'_1 , A'_2 les normes des projections du point $\mathcal{S}'$ sur les vecteurs $\vec{\mathcal{V}}_1$ et $\vec{\mathcal{V}}_2$ respectivement, le point $\mathcal{S}'$ peut être décrit de la façon suivante :

$$\mathcal{S}' = \mathcal{PQ} + A'_1\vec{\mathcal{V}}_1 + A'_2\vec{\mathcal{V}}_2 \tag{A.11}$$

Étant donné que les vecteurs $\vec{\mathcal{V}}_1$, $\vec{\mathcal{V}}_2$ ne sont pas orthogonaux, A'_1 et A'_2 sont calculés d'après l'équation A.12 (où $\langle\ ,\ \rangle$ indiquent le produit scalaire).

$$\begin{bmatrix} A'_1 \\ A'_2 \end{bmatrix} = \begin{bmatrix} 1 & \langle \vec{\mathcal{V}}_1 , \vec{\mathcal{V}}_2 \rangle \\ \langle \vec{\mathcal{V}}_2 , \vec{\mathcal{V}}_1 \rangle & 1 \end{bmatrix}^{-1} \times \begin{bmatrix} \langle (\mathcal{S}' - \mathcal{PQ}) , \vec{\mathcal{V}}_1 \rangle \\ \langle (\mathcal{S}' - \mathcal{PQ}) , \vec{\mathcal{V}}_2 \rangle \end{bmatrix} \tag{A.12}$$

Finalement, soient A_1 , A_2 les projections du point S recherché sur les vecteurs $\vec{V}_1$, $\vec{V}_2$ respectivement, S est donné par l'équation suivante :

$$S = PQ + A_1\vec{V}_1 + A_2\vec{V}_2 \tag{A.13}$$

Les valeurs de A_1 , A_2 seront obtenus de l'équation A.14 (où $||X||$ indique « la norme du vecteur X »).

$$\begin{bmatrix} \dfrac{A_1}{||(Q - PQ)||} \\ \dfrac{A_2}{||(R - PQ)||} \end{bmatrix} = \begin{bmatrix} \dfrac{A'_1}{||(\mathcal{Q} - \mathcal{PQ})||} \\ \dfrac{A'_2}{||(\mathcal{R} - \mathcal{PQ})||} \end{bmatrix} \tag{A.14}$$

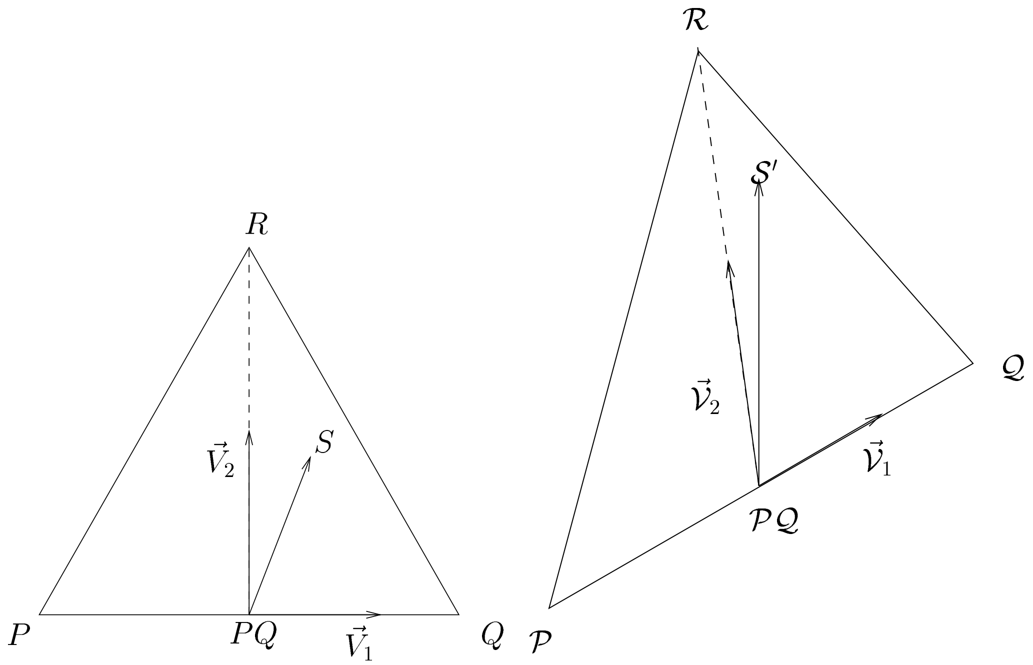

FIG. A.5: Triangle $PQR \in \mathbb{R}^2$ et le correspondant $\mathcal{PQR} \in \mathbb{R}^n$.

Annexe B

Fibres Articulatoires $\mathcal{A}_F$

Nous présentons la totalité des fibres articulatoires obtenues d'après la mèthode décrite dans le chapitre 4. La matrice de covariance utilisée (cf. équation 4.8) a été obtenu à partir de la mesure d'incertidude des formants par l'analyse LPC. Chaque figure présente en haut les formants que la fibre articulatoire produit (en cercle) superposés aux formants (en ligne continue) que nous cherchons à inverser.

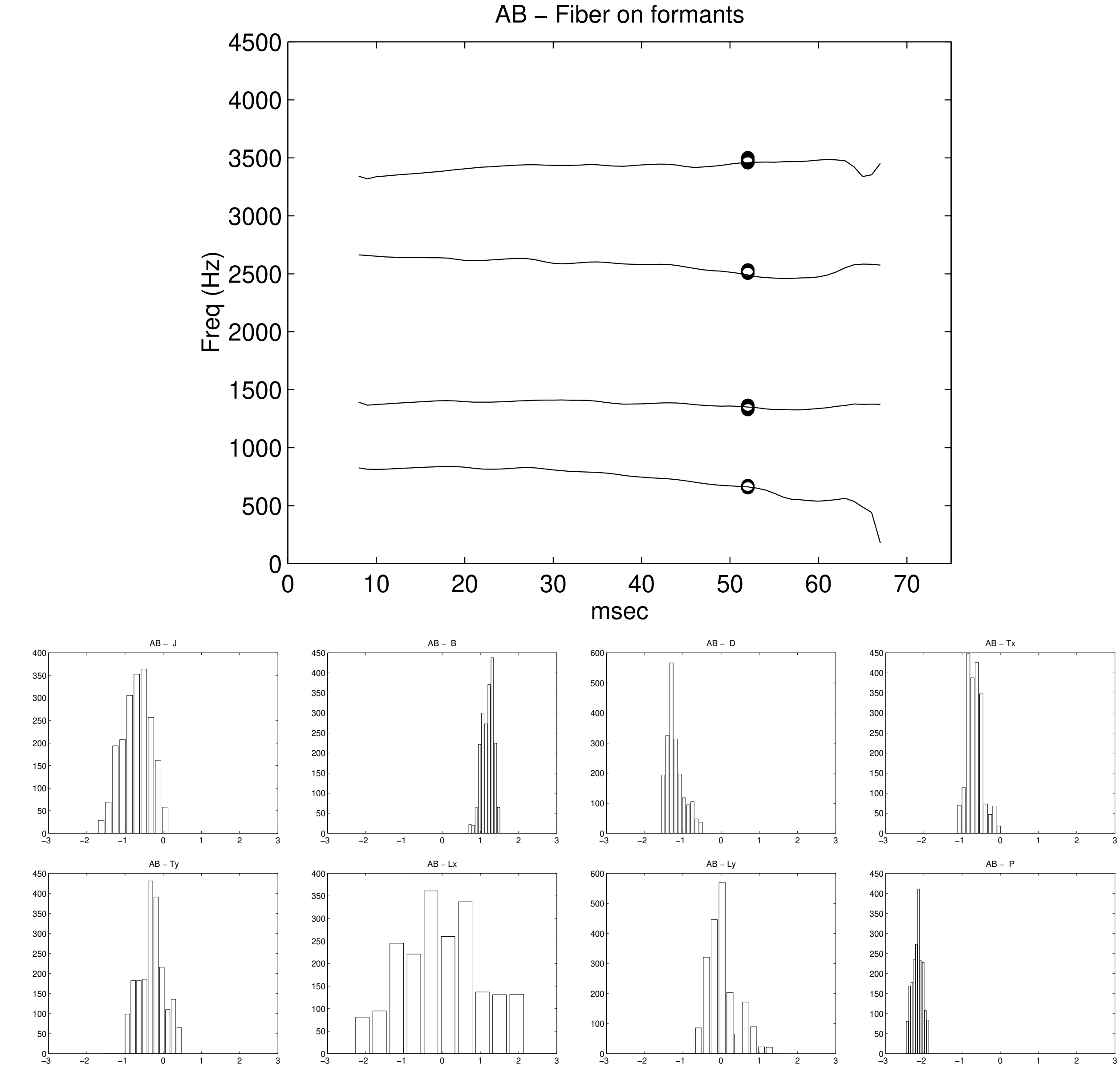

FIG. B.1: Fibre articulatoire du « AB », fibre à $\mathcal{A}_F$.

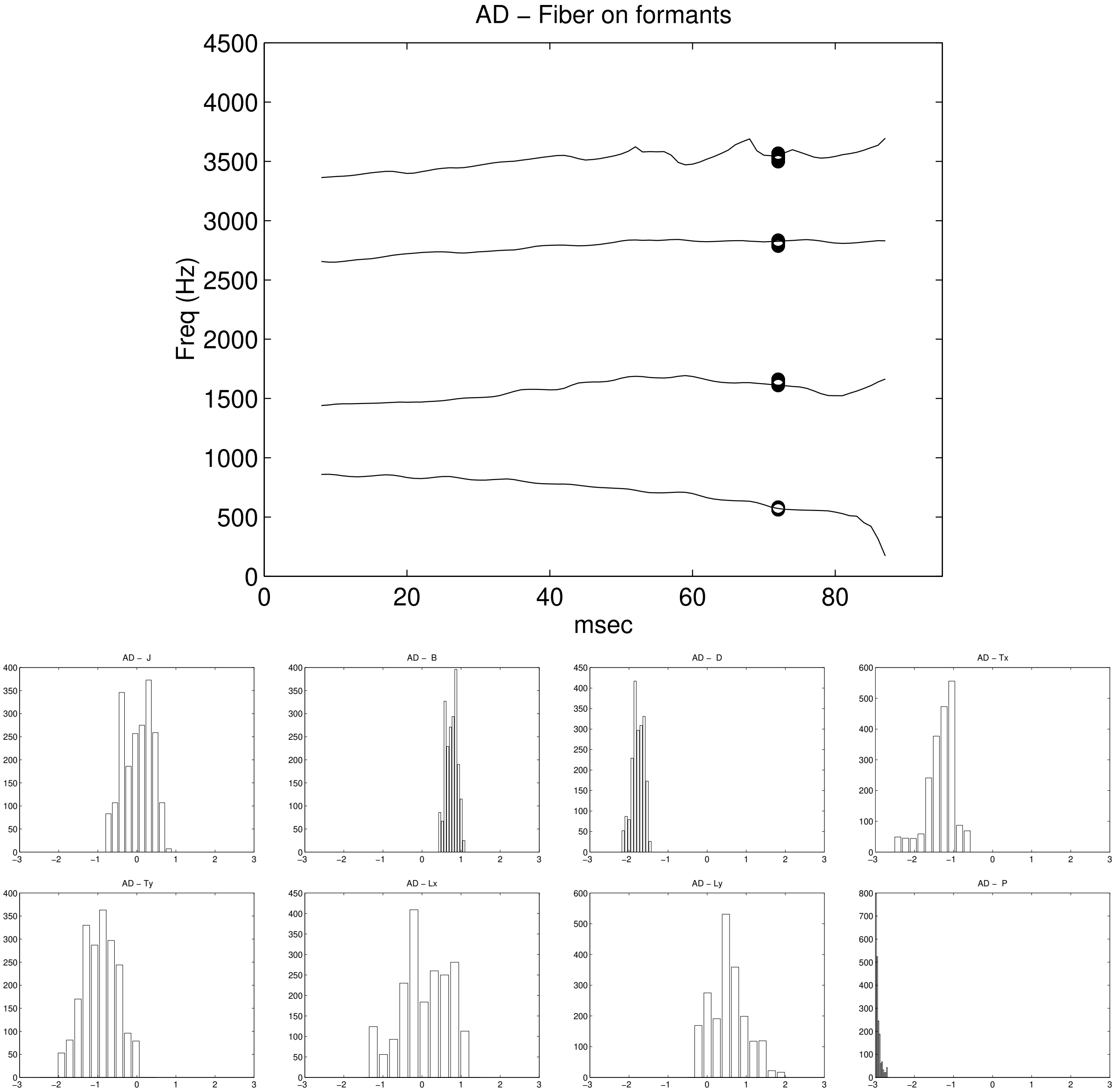

FIG. B.2: Fibre articulatoire du « AD », fibre à $\mathcal{A}_F$.

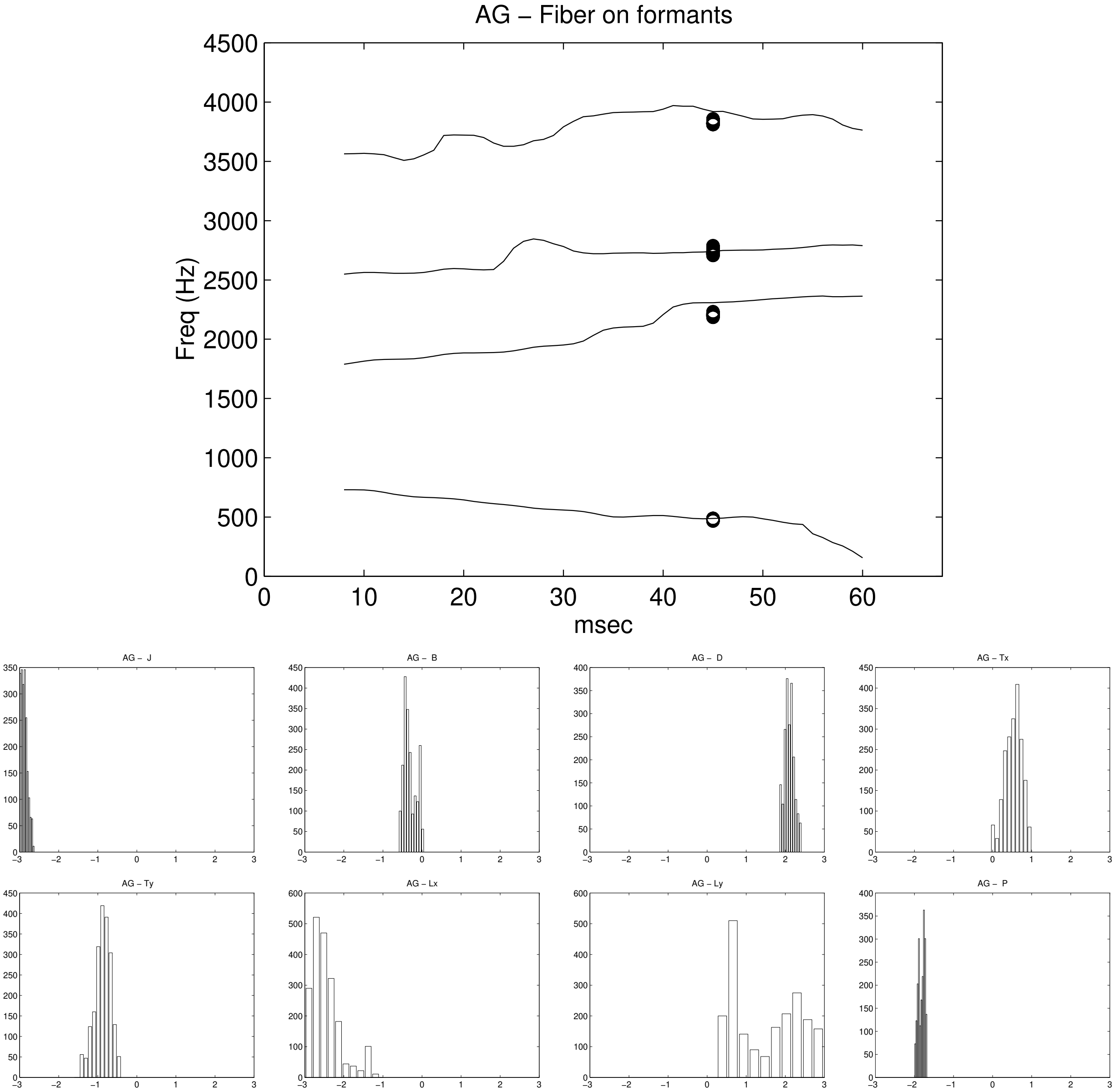

FIG. B.3: Fibre articulatoire du « AG », fibre à $\mathcal{A}_F$.

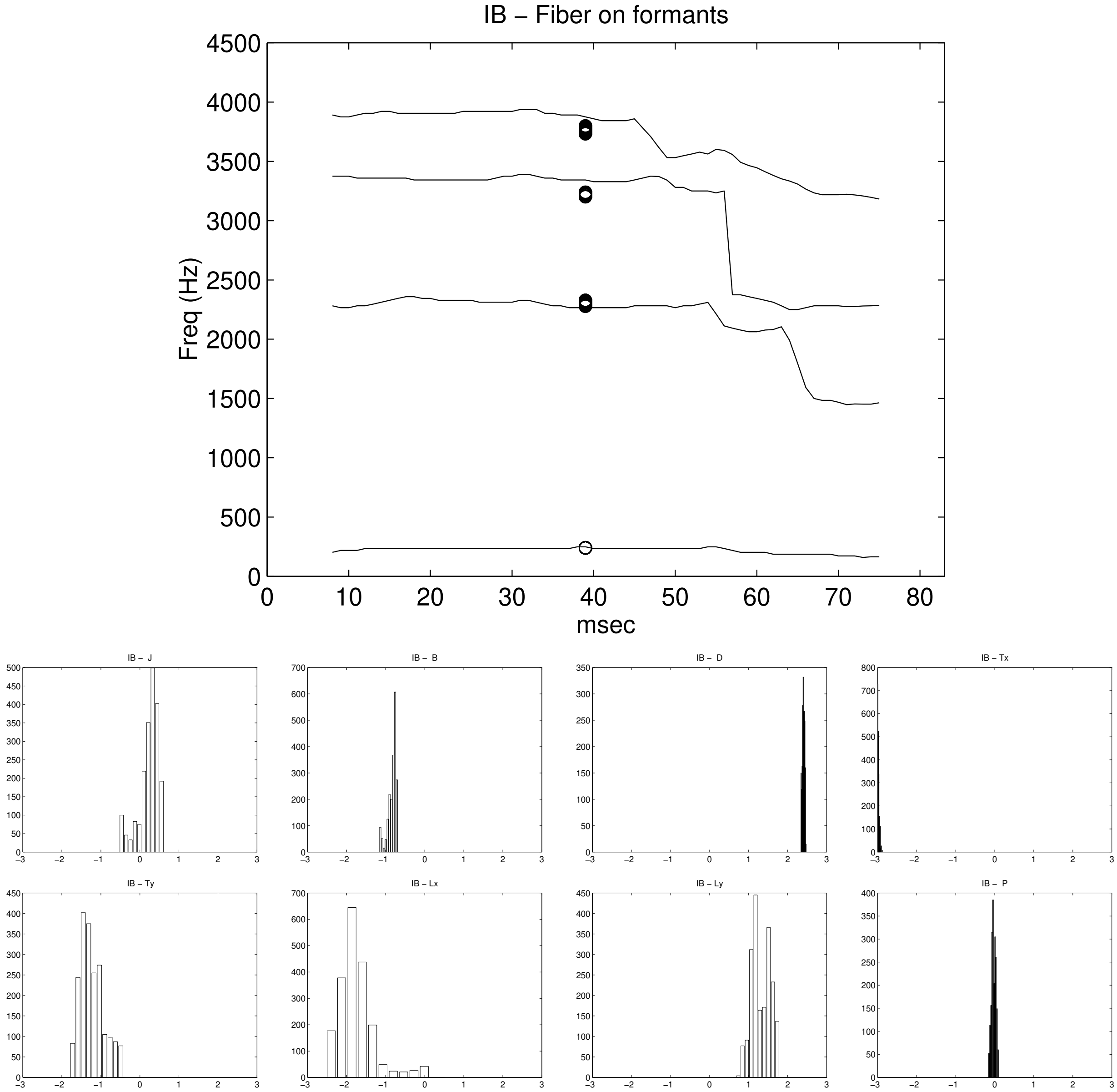

FIG. B.4: Fibre articulatoire du « IB », fibre à $\mathcal{A}_F$.

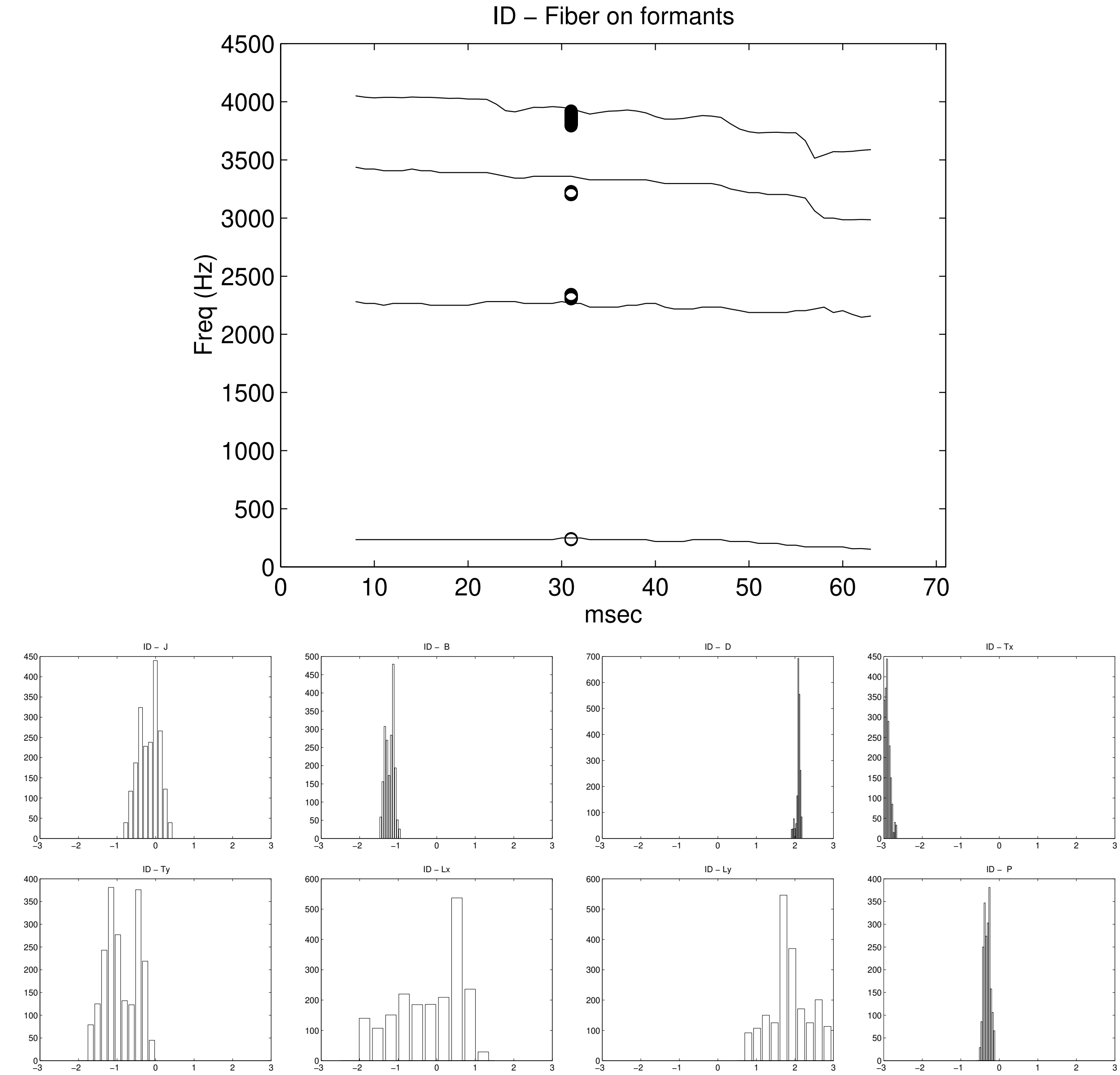

Fig. B.5: Fibre articulatoire du « ID », fibre à $\mathcal{A}_F$.

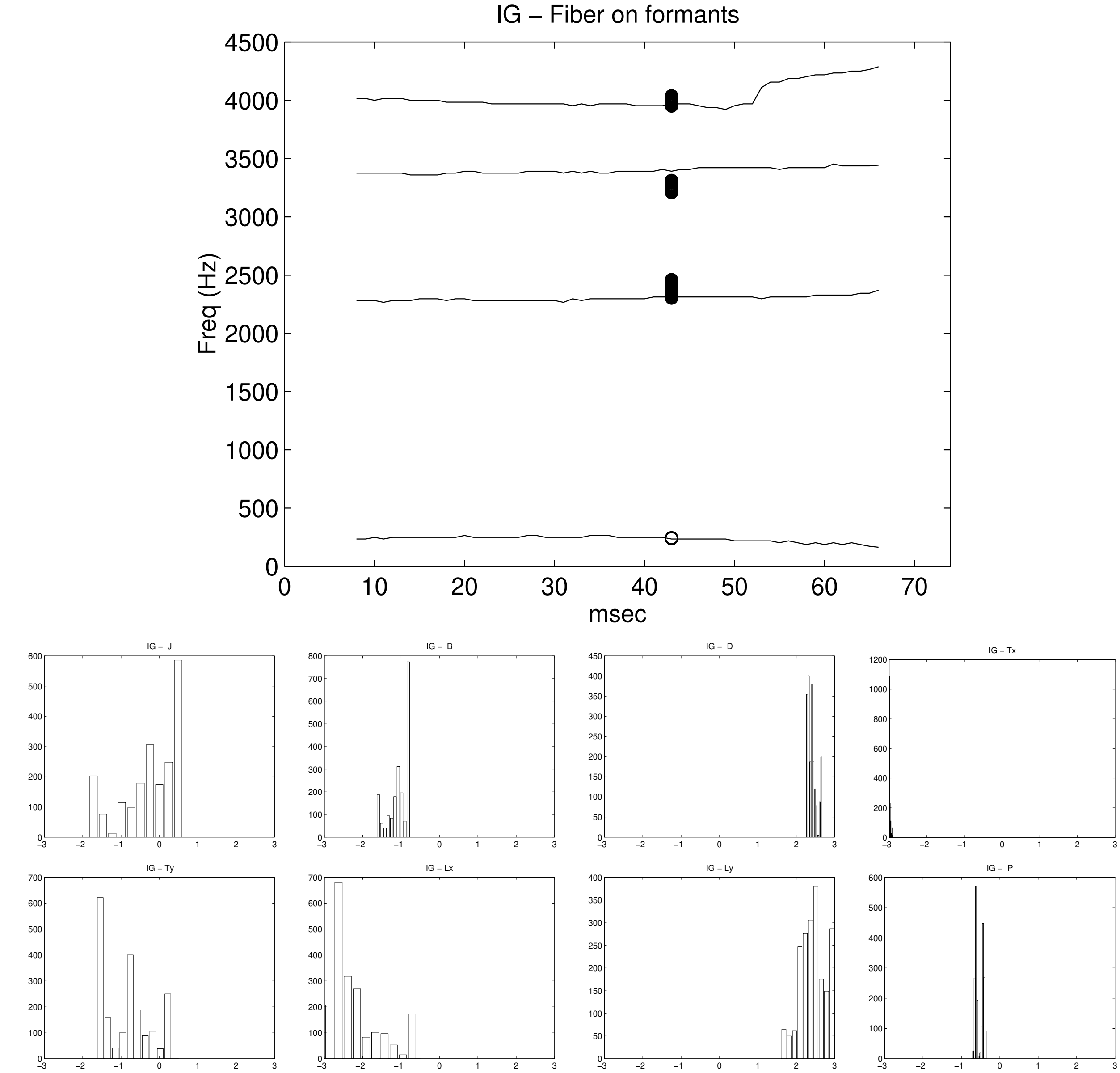

Fig. B.6: Fibre articulatoire du « IG », fibre à $\mathcal{A}_F$.

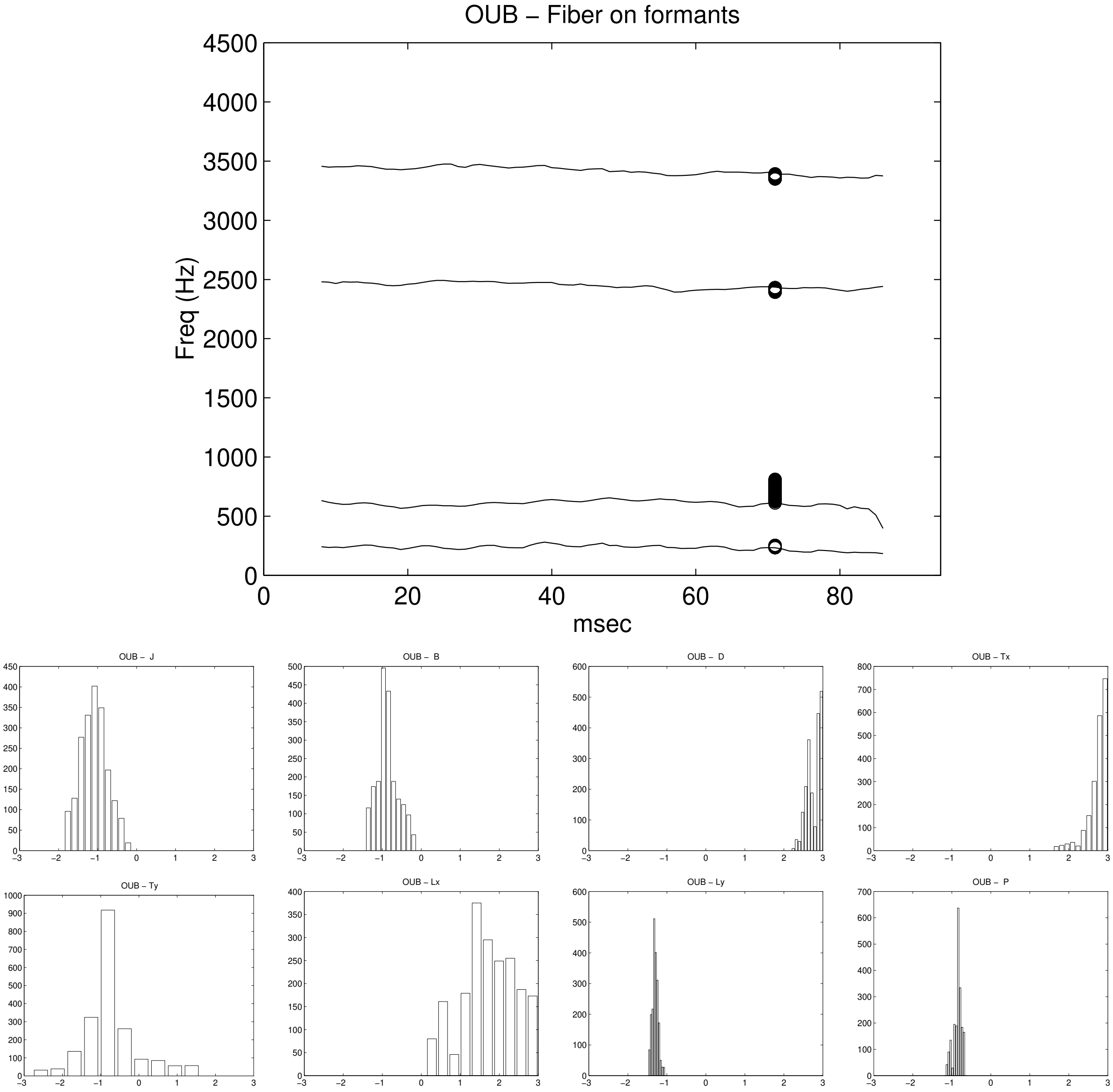

FIG. B.7: Fibre articulatoire du « OUB », fibre à $\mathcal{A}_F$.

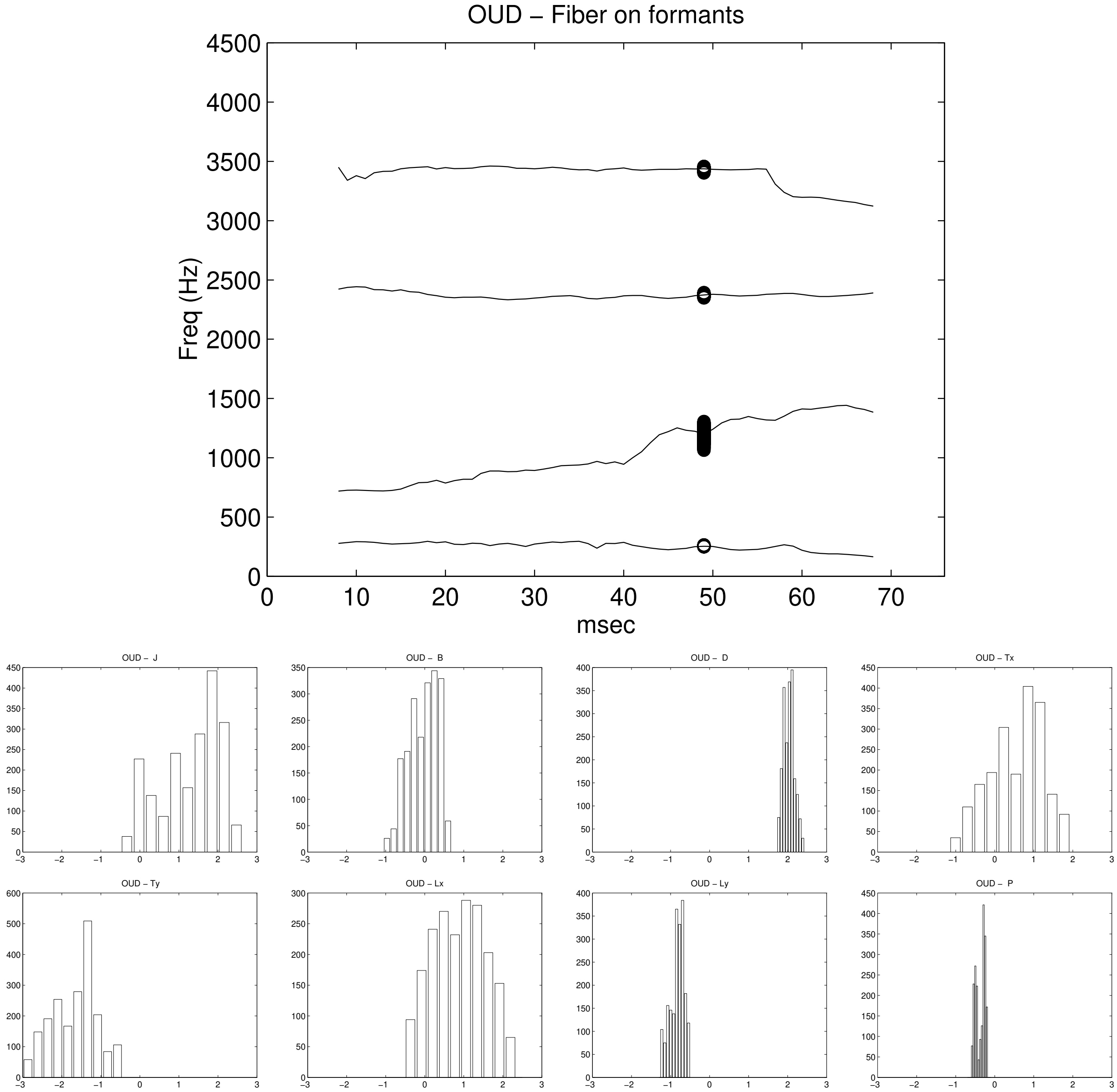

FIG. B.8: Fibre articulatoire du « OUD », fibre à $\mathcal{A}_F$.

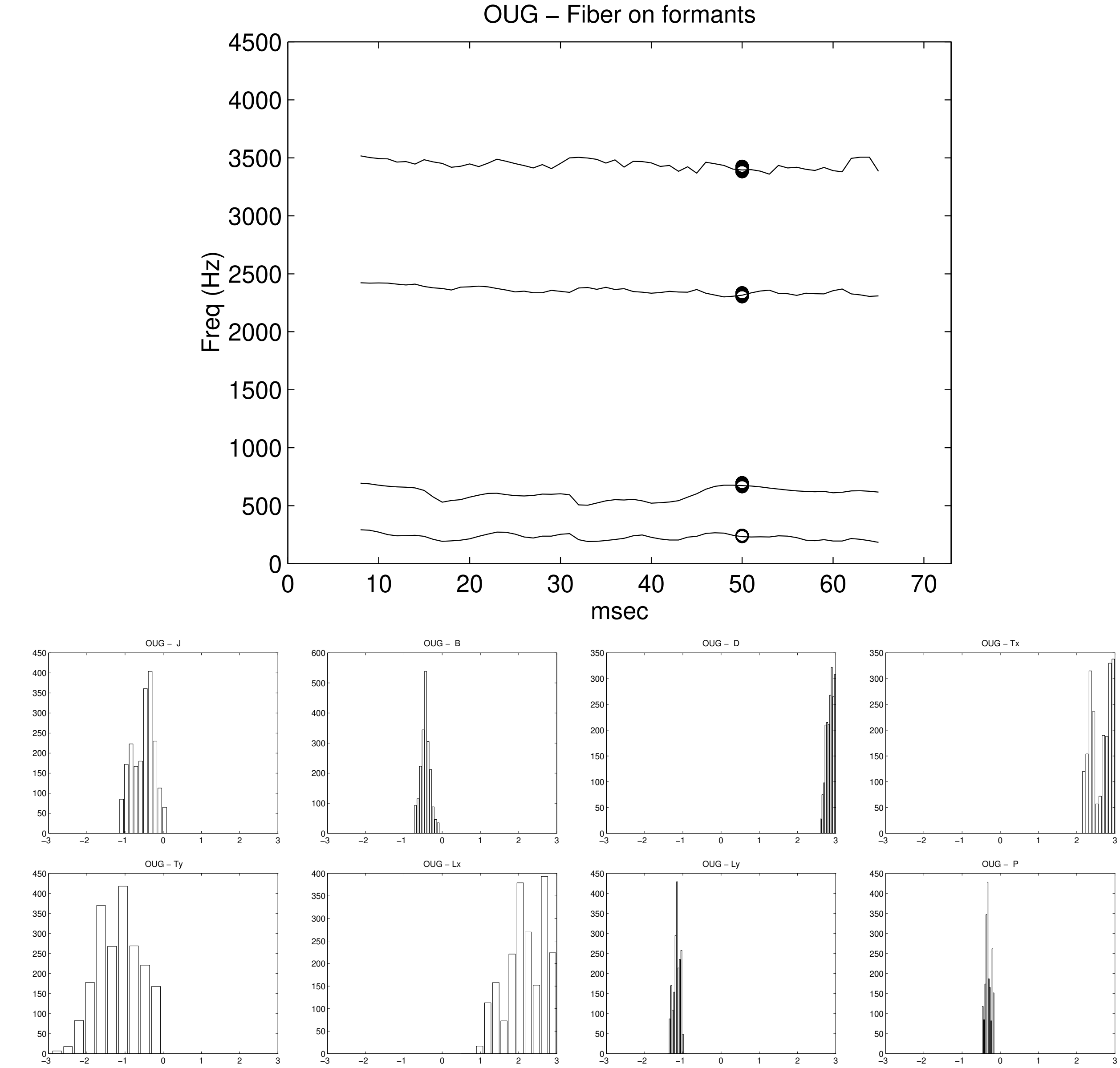

FIG. B.9: Fibre articulatoire du « OUG », fibre à $\mathcal{A}_F$.

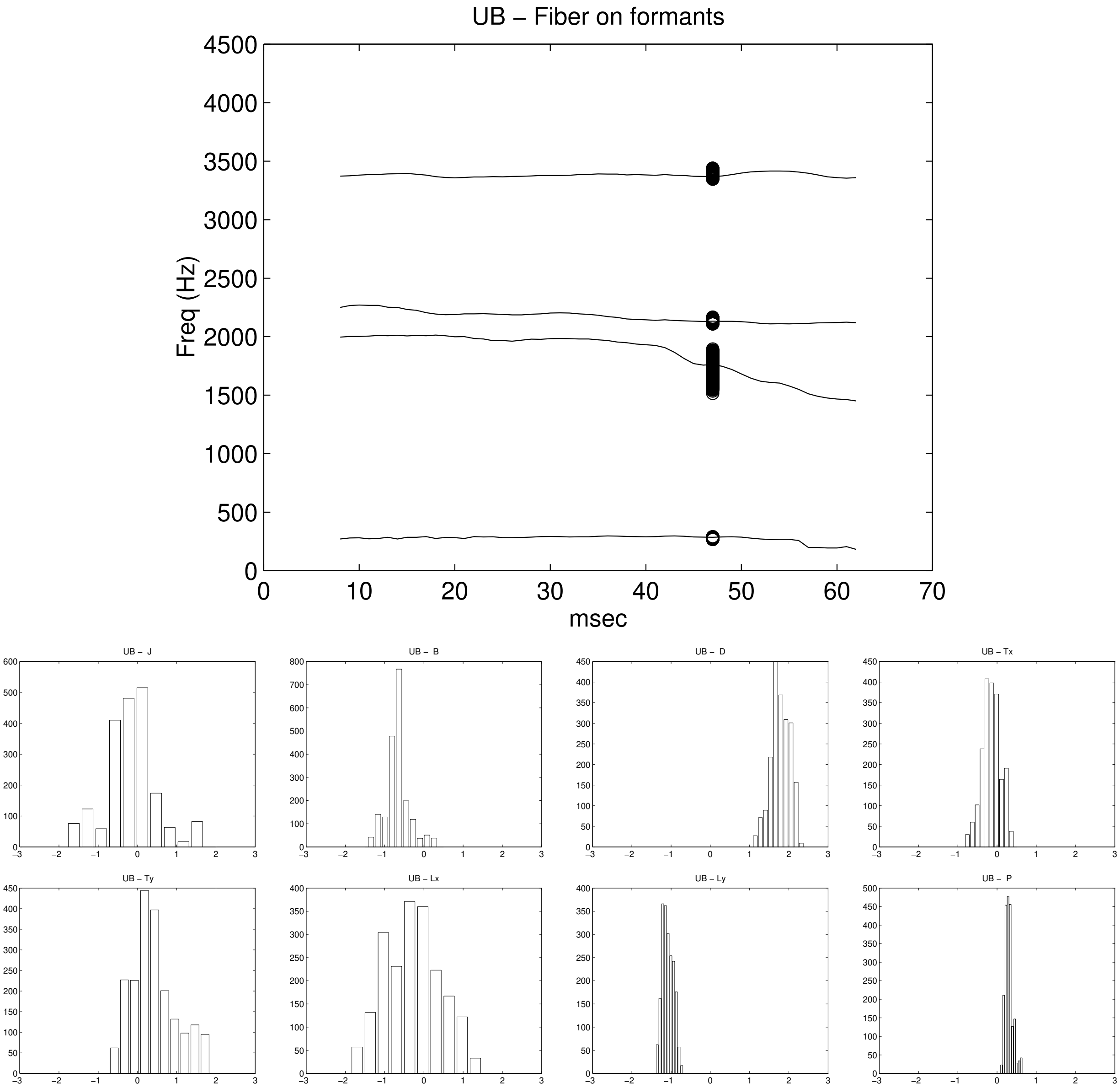

FIG. B.10: Fibre articulatoire du « UB », fibre à $\mathcal{A}_F$.

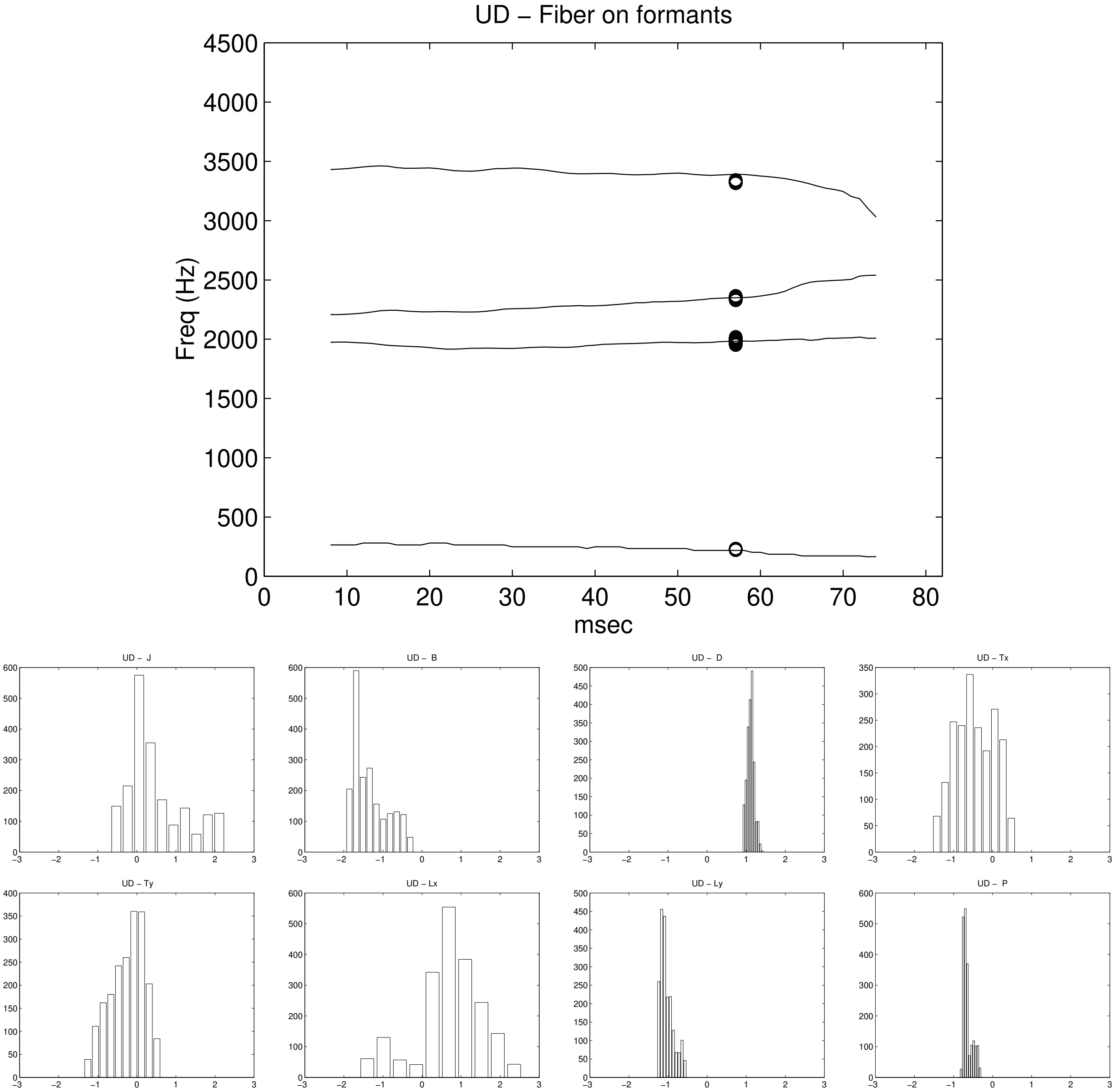

FIG. B.11: Fibre articulatoire du « UD », fibre à $\mathcal{A}_F$.

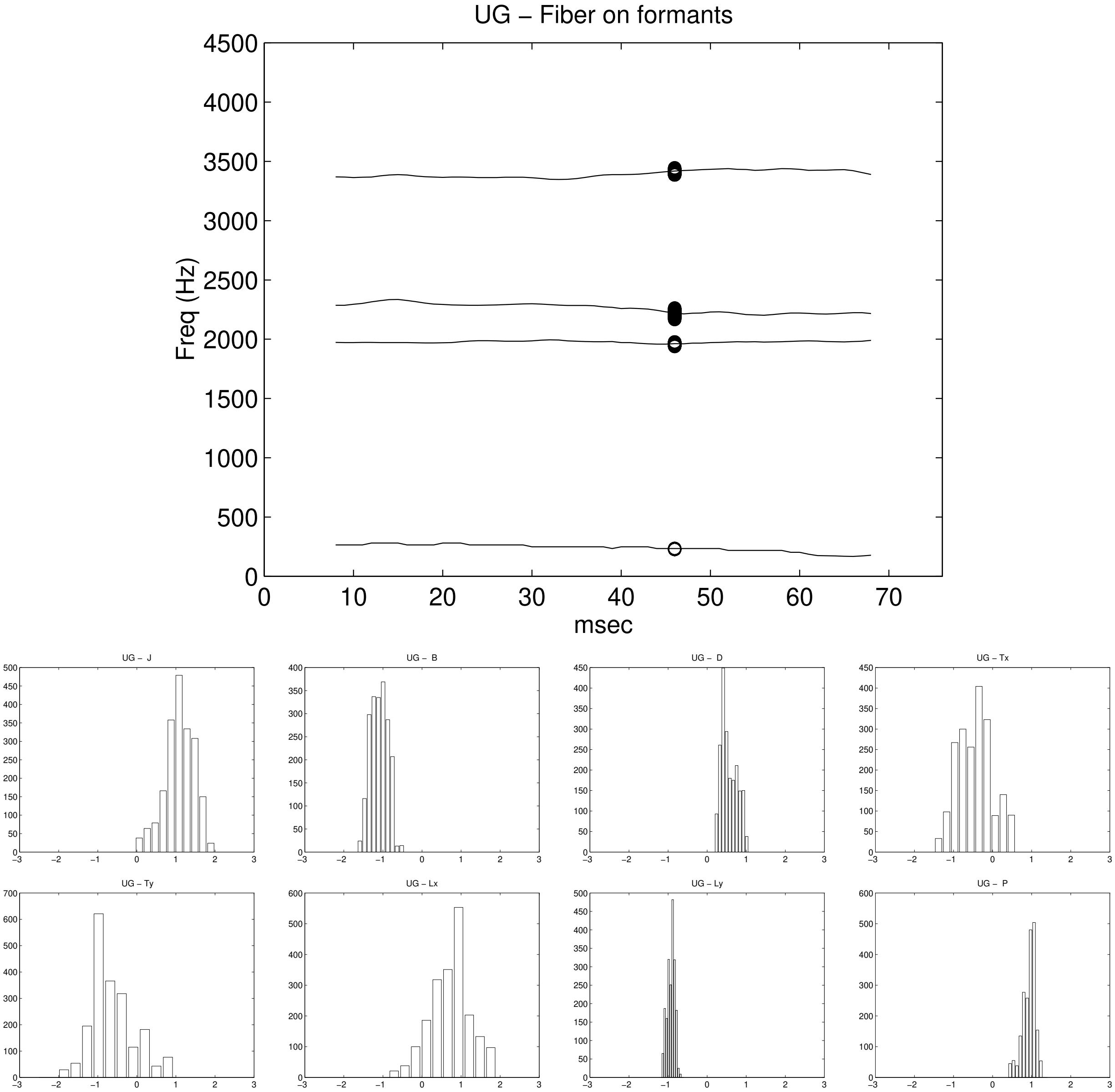

FIG. B.12: Fibre articulatoire du « UG », fibre à $\mathcal{A}_F$.

Annexe C

Fibres Articulatoires $\mathcal{A}_E$

Nous présentons la totalité des fibres articulatoires obtenues d'après la mèthode décrite dans le chapitre 4. La matrice de covariance utilisée (cf. équation 4.8) a été obtenu à partir de la mesure d'incertidude des formants par l'analyse LPC. Chaque figure présente en haut les formants que la fibre articulatoire produit (en cercle) superposés aux formants (en ligne continue) que nous cherchons à inverser.

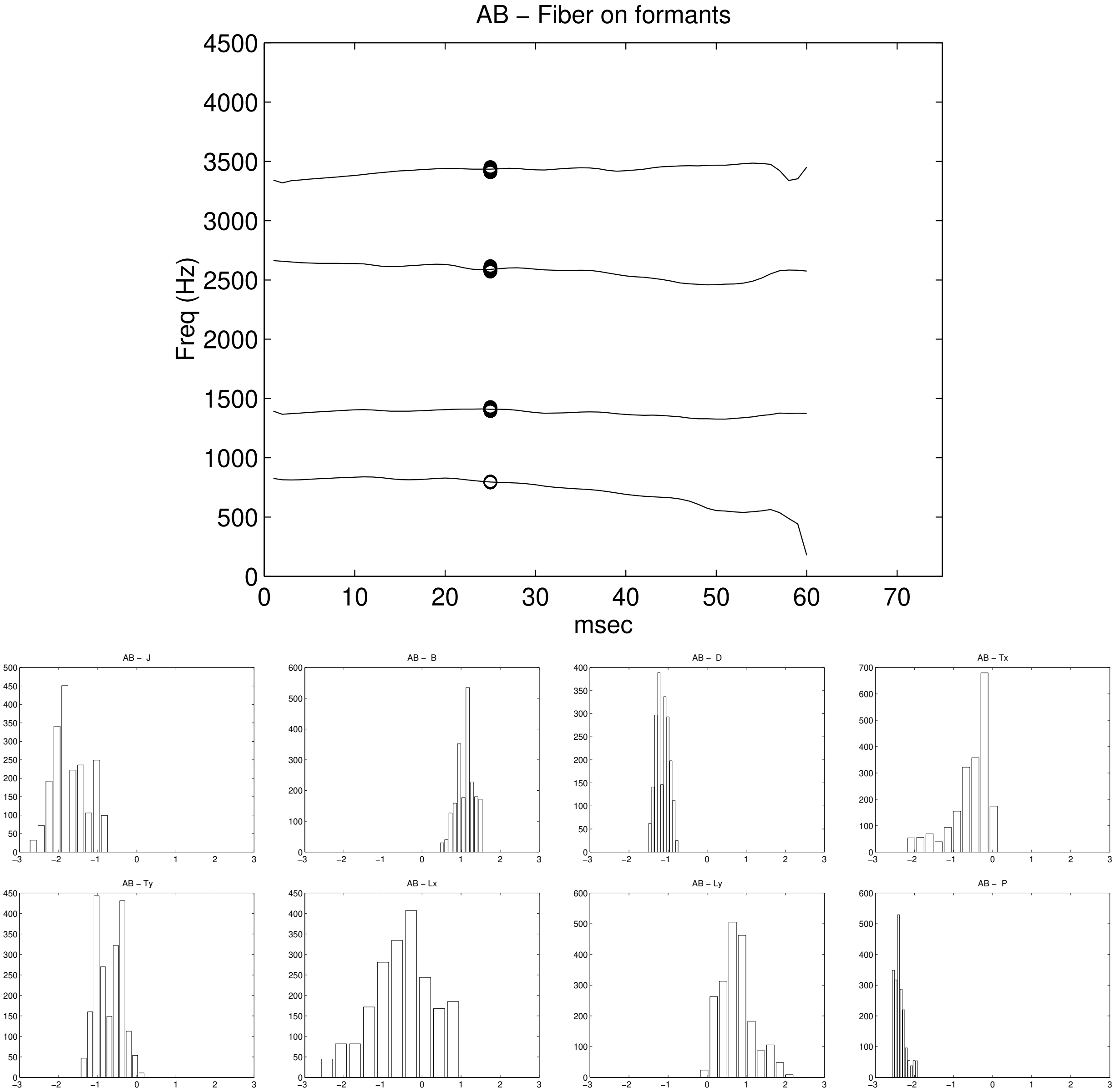

FIG. C.1: Fibre articulatoire du « AB », fibre à $\mathcal{A}_E$.

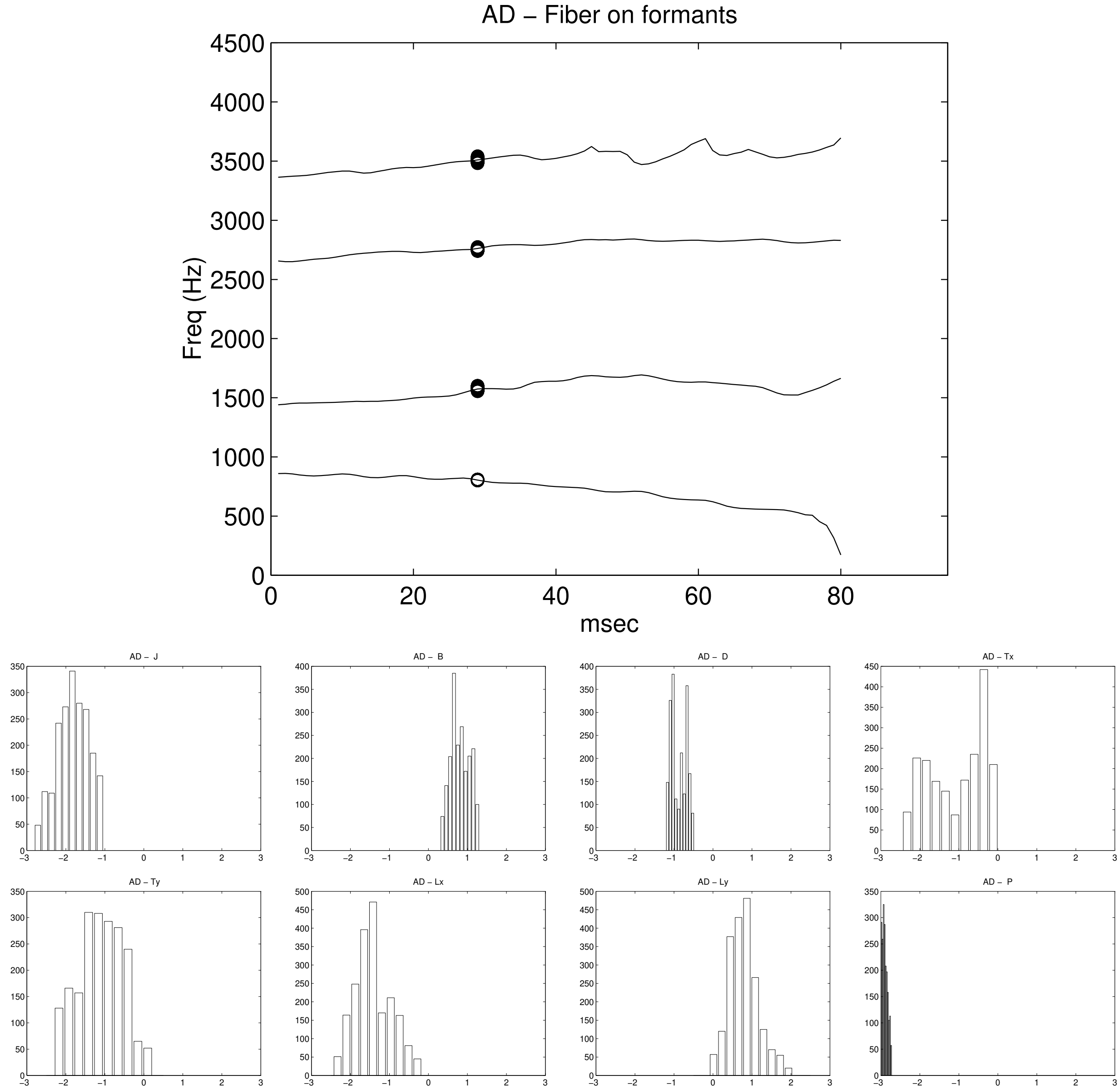

FIG. C.2: Fibre articulatoire du « AD », fibre à $\mathcal{A}_E$.

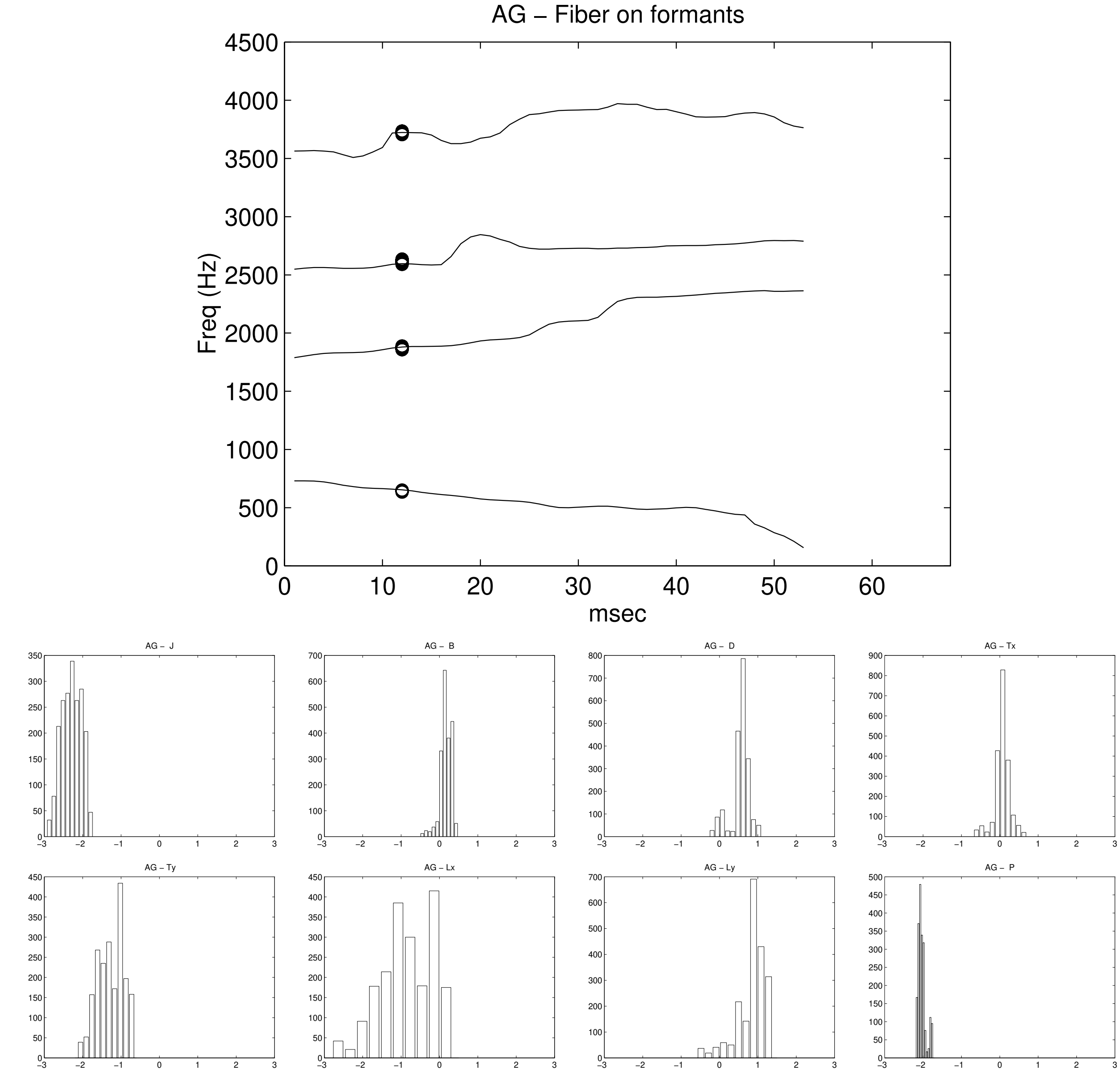

FIG. C.3: Fibre articulatoire du « AG », fibre à $\mathcal{A}_E$.

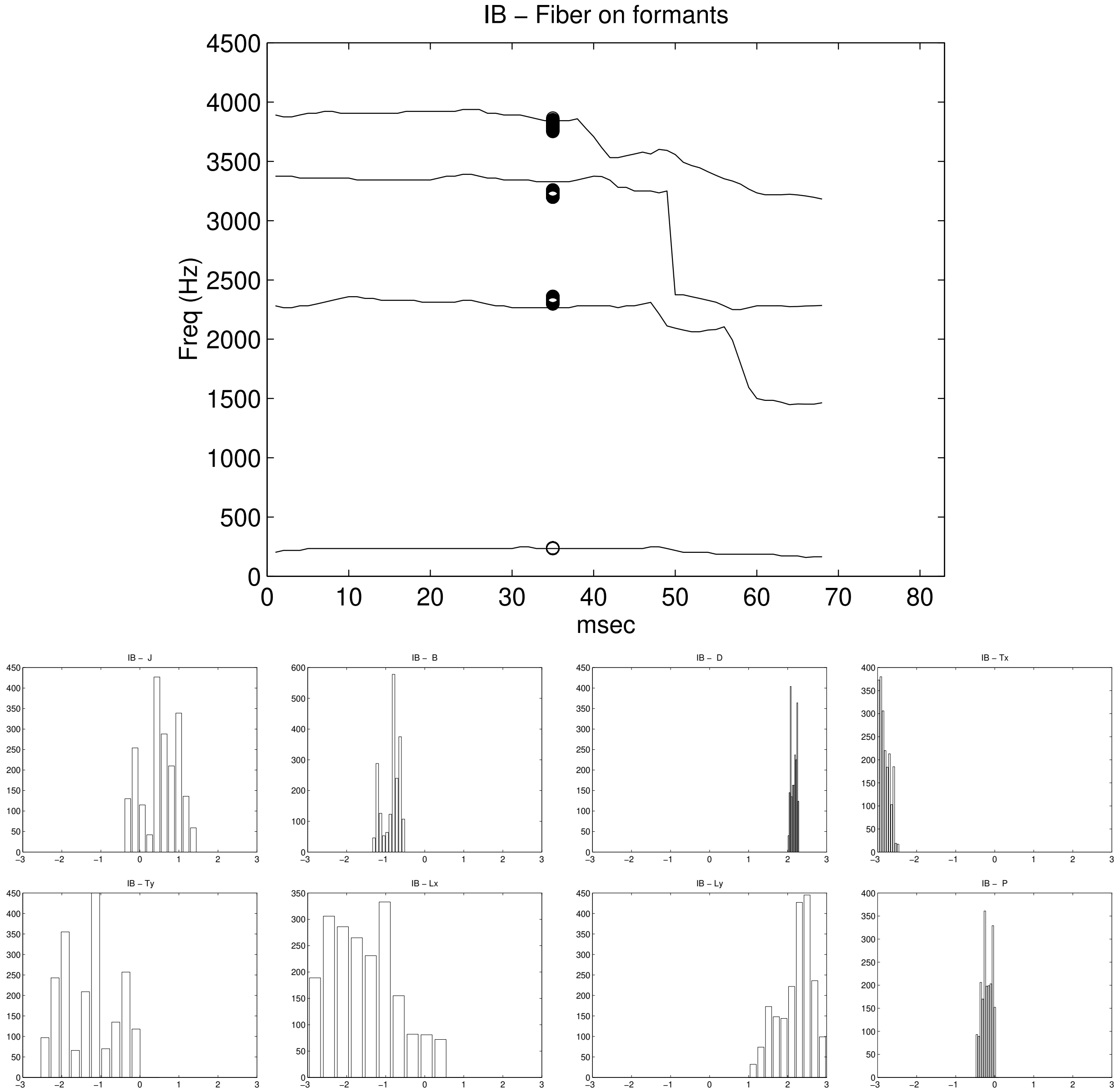

FIG. C.4: Fibre articulatoire du « IB », fibre à $\mathcal{A}_E$.

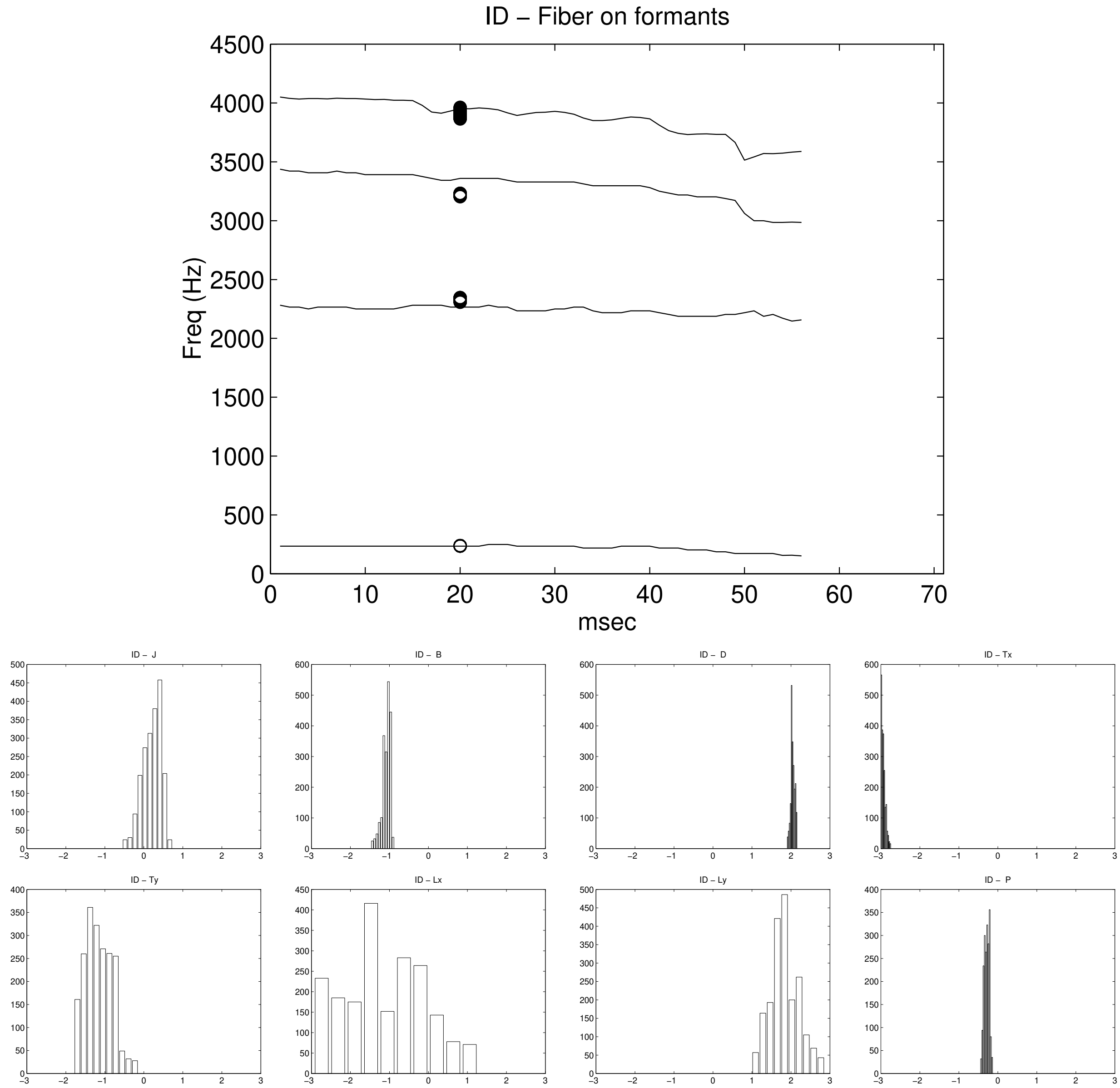

FIG. C.5: Fibre articulatoire du « ID », fibre à $\mathcal{A}_E$.

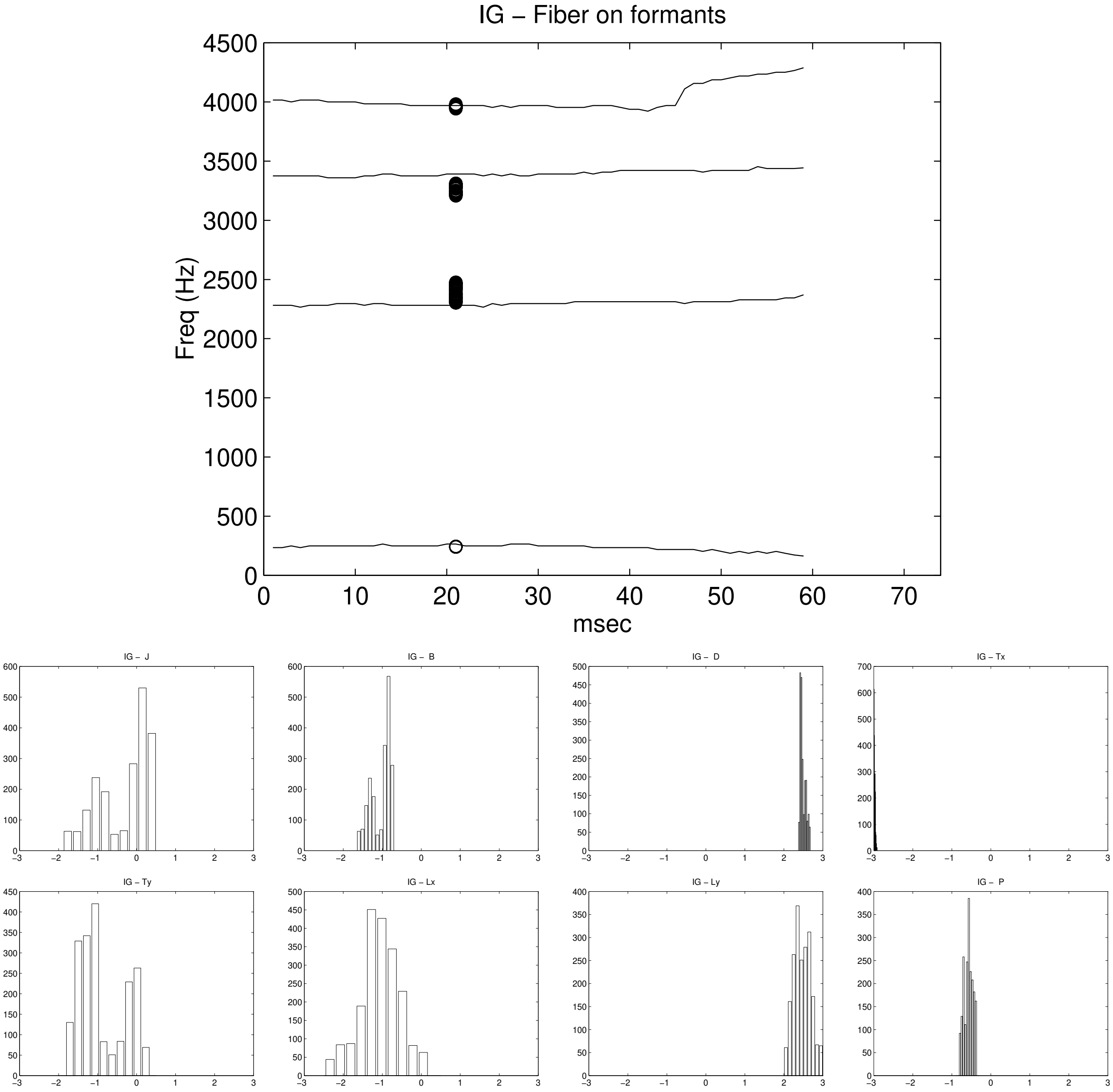

FIG. C.6: Fibre articulatoire du « IG », fibre à $\mathcal{A}_E$.

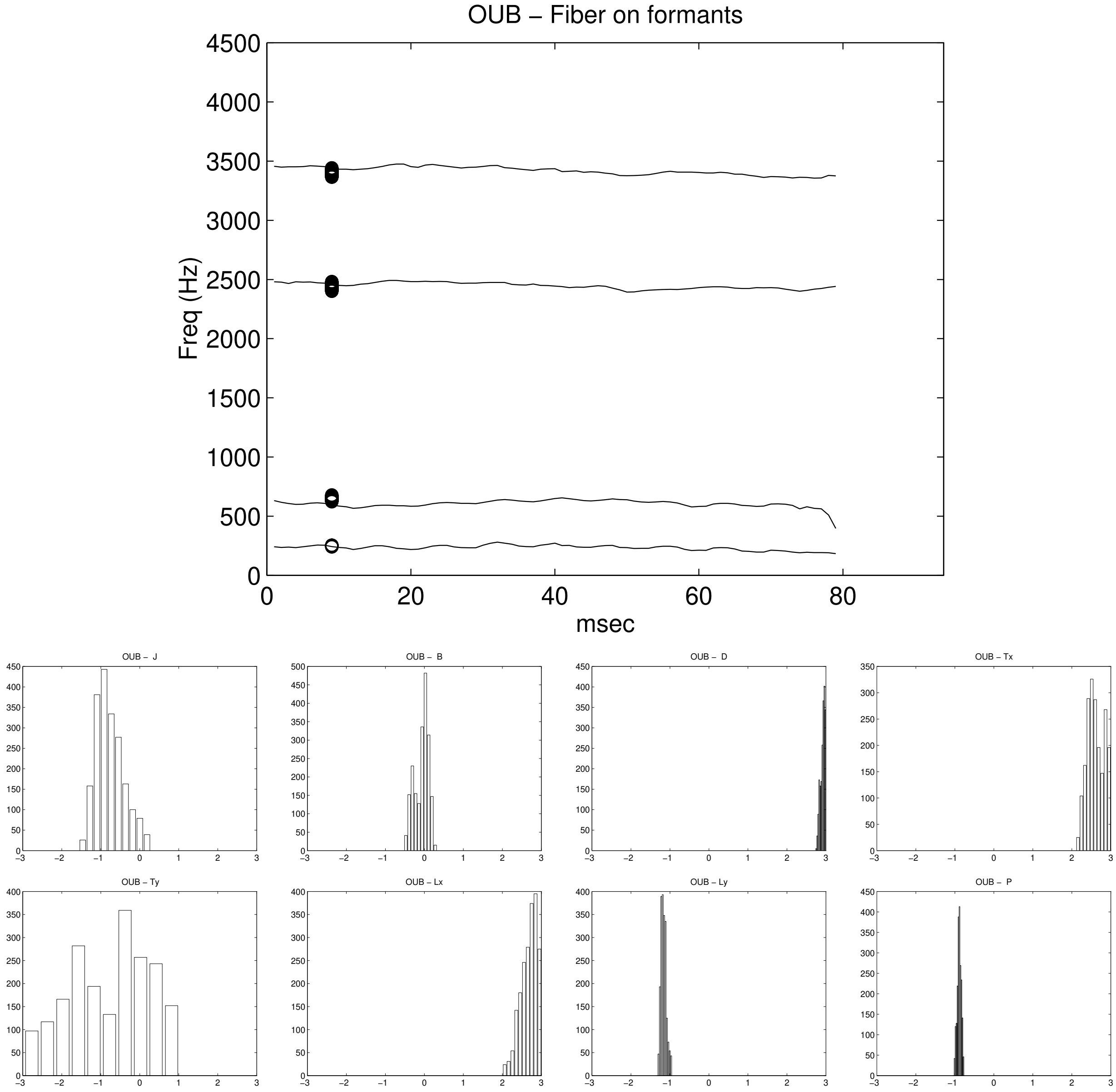

FIG. C.7: Fibre articulatoire du « OUB », fibre à $\mathcal{A}_E$.

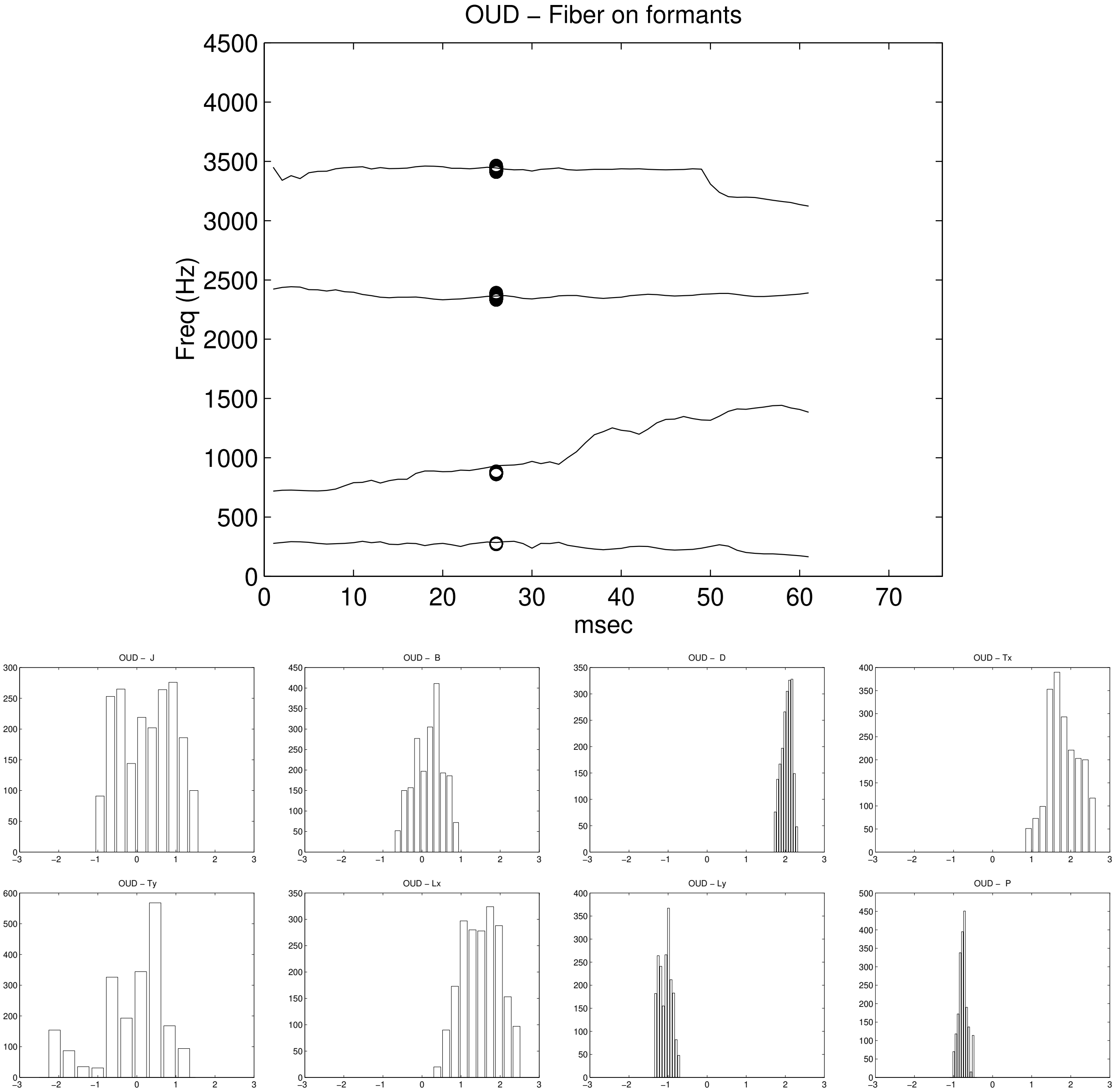

FIG. C.8: Fibre articulatoire du « OUD », fibre à $\mathcal{A}_E$.

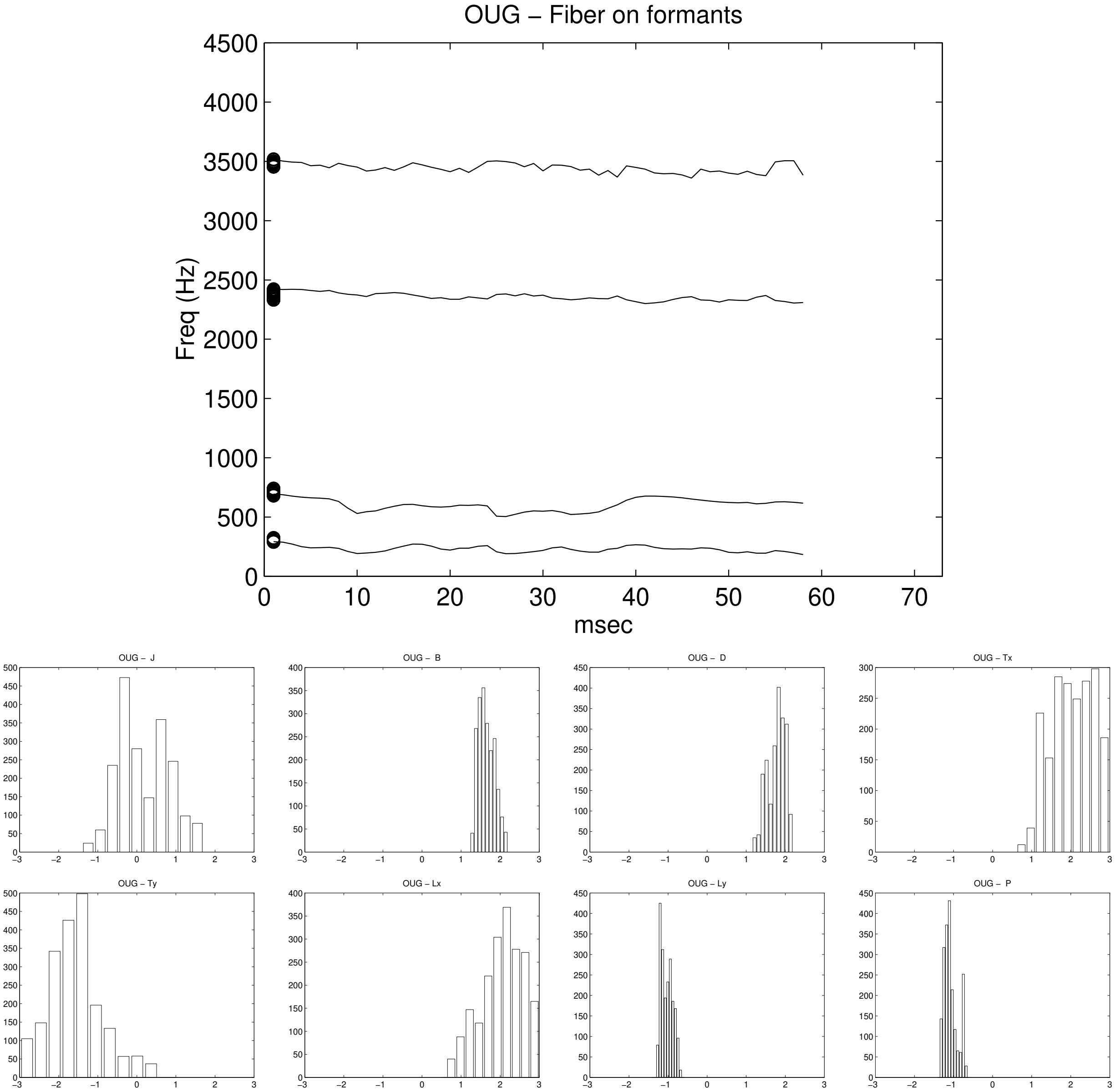

FIG. C.9: Fibre articulatoire du « OUG », fibre à $\mathcal{A}_E$.

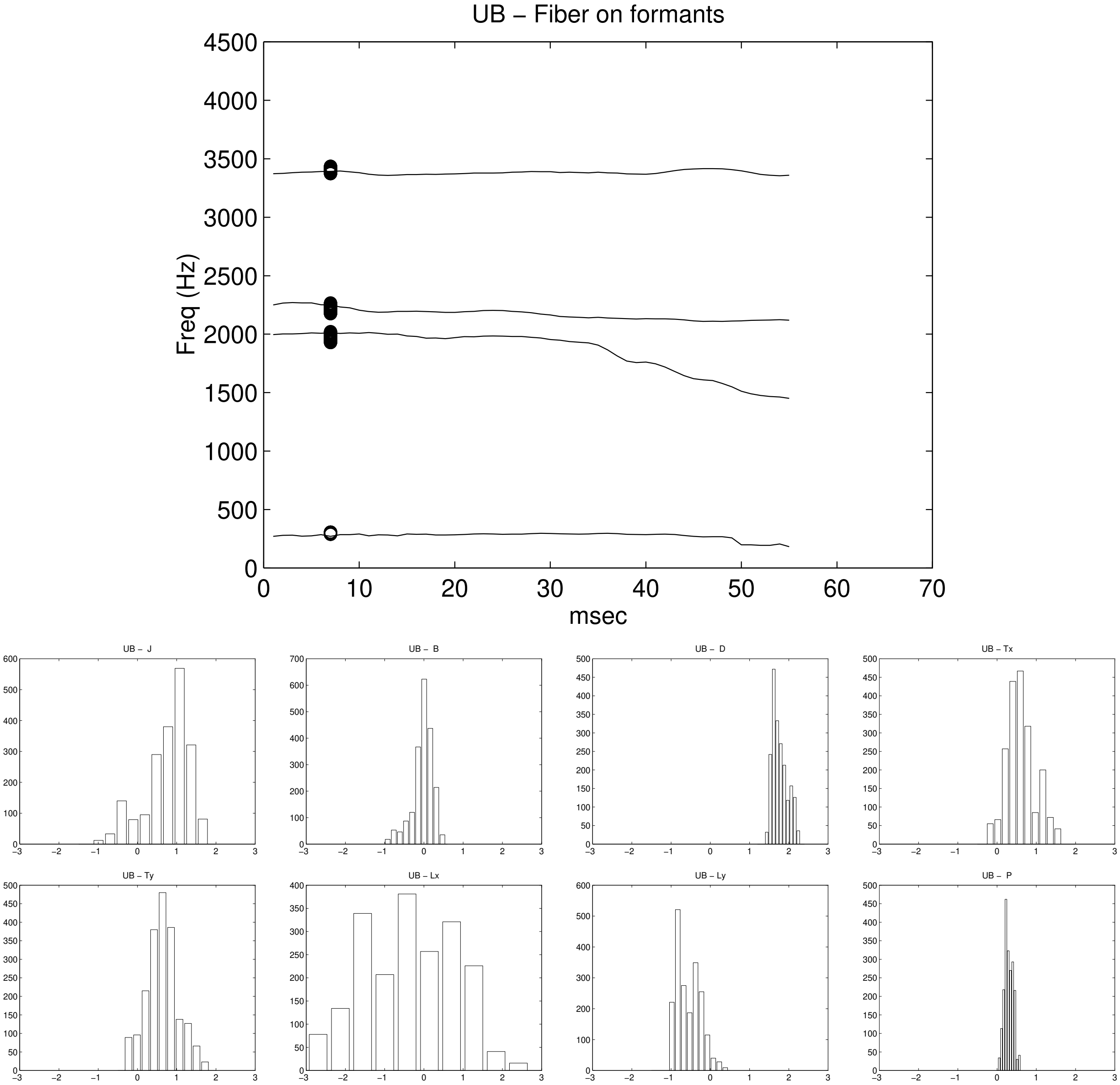

FIG. C.10: Fibre articulatoire du « UB », fibre à $\mathcal{A}_E$.

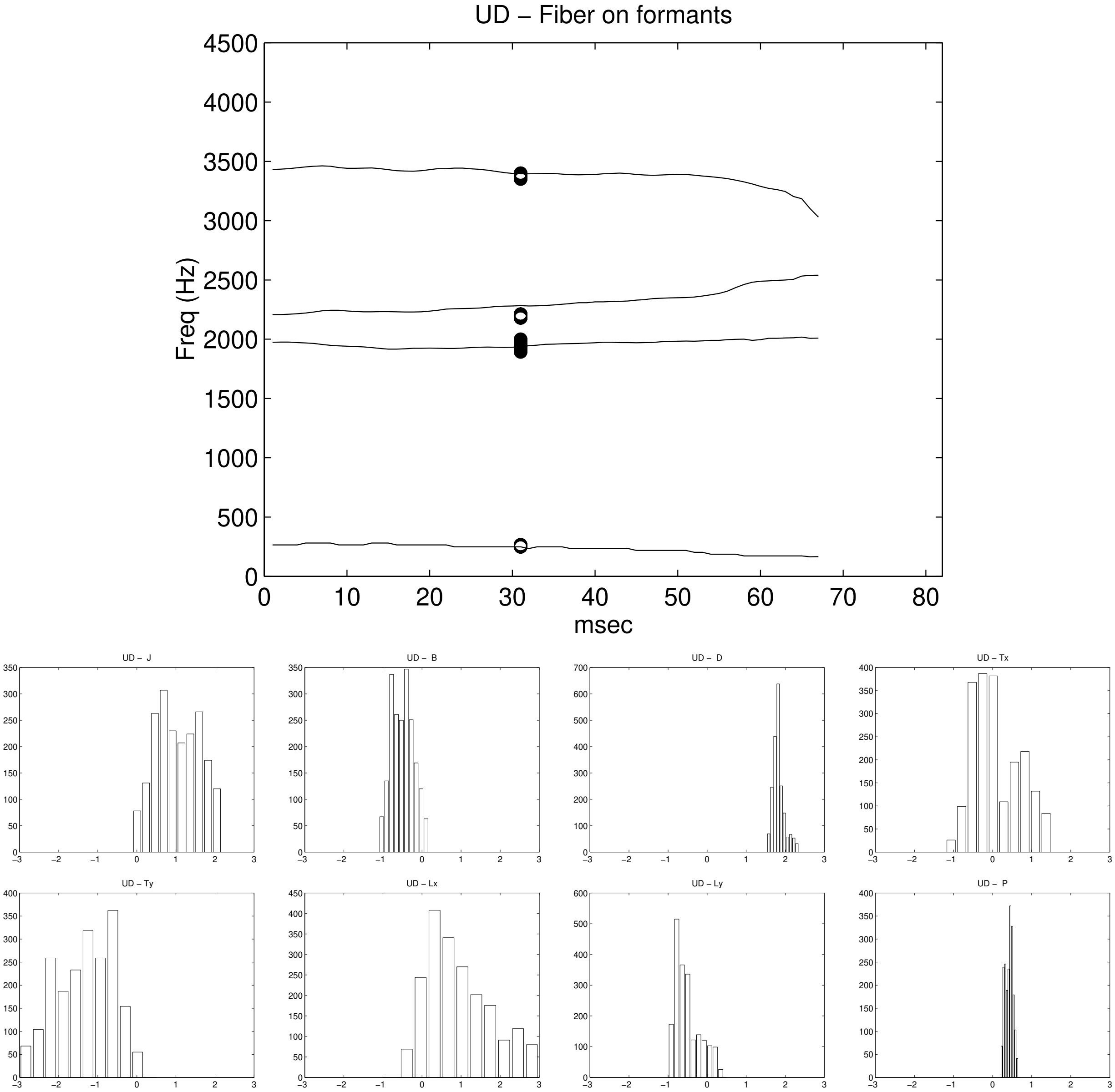

FIG. C.11: Fibre articulatoire du « UD », fibre à $\mathcal{A}_E$.

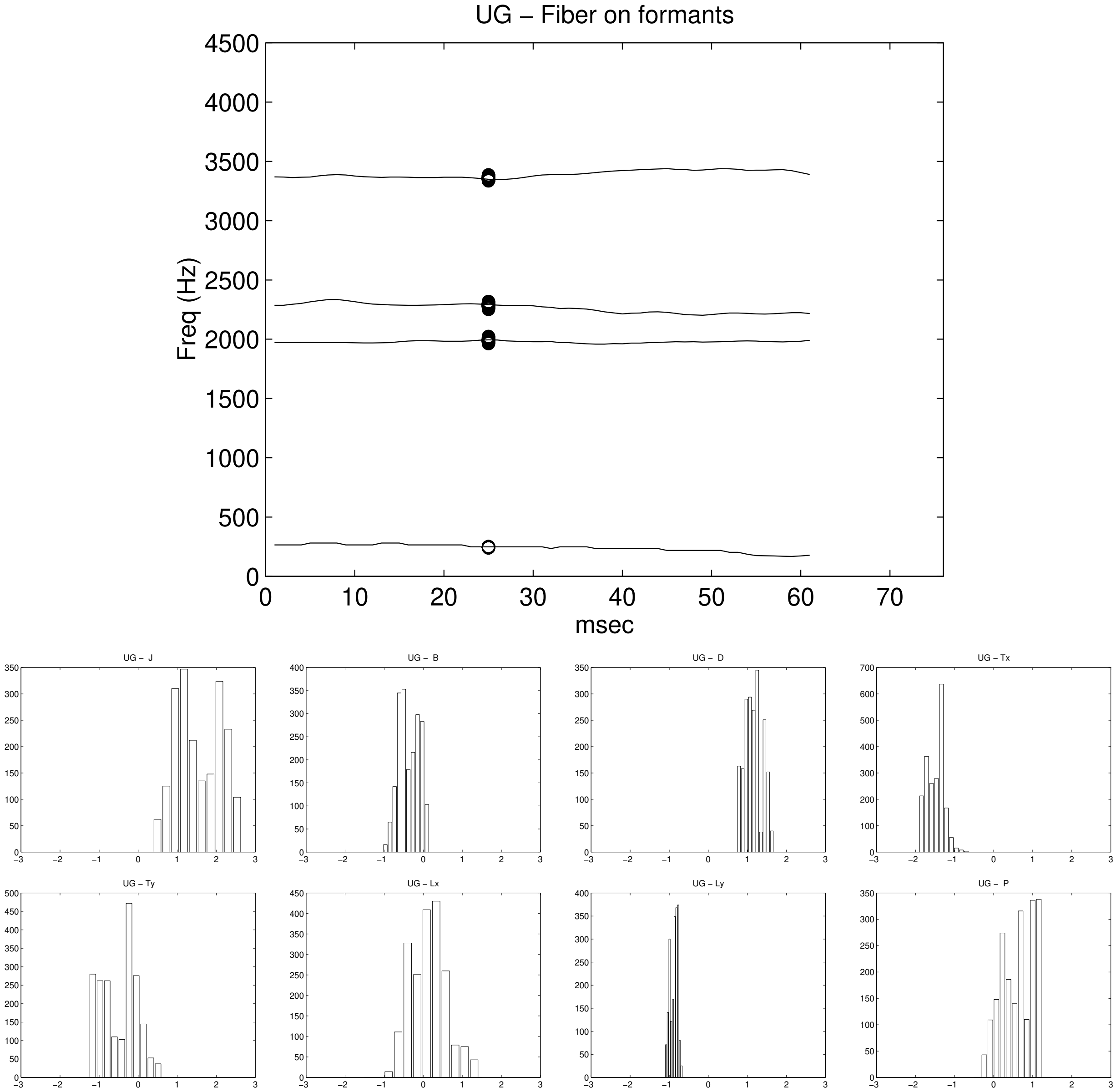

FIG. C.12: Fibre articulatoire du « UG », fibre à $\mathcal{A}_E$.