Nancy-Université
Université Henri Poincaré

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Acquisition et modélisation de données articulatoires dans un contexte multimodal

THÈSE

présentée et soutenue publiquement le 12 Novembre 2009

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Michaël Aron

Composition du jury

Rapporteurs :	Phil Hoole	Chercheur LMU, Munich
	Yohan Payan	Directeur de Recherche CNRS, Grenoble
Examinateurs :	Marie-Odile Berger	Chargée de Recherche INRIA, Nancy
	Saida Bouakaz	Professeur des Universités, Lyon I
	Nacer Boudjlida	Professeur des Universités, UHP Nancy
	Erwan Kerrien	Chargé de Recherche INRIA, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria

Mis en page avec la classe thloria.

Remerciements

Je remercie mes rapporteurs, Yohan Payan et Phil Hoole pour l'attention et la curiosité manifestes avec lesquelles ils ont jugé ce manuscrit ; Nacer Boudjlida et Saida Bouakaz d'avoir accepté de faire partie du jury. Enfin, un immense merci à Marie-Odile Berger et à Erwan Kerrien pour m'avoir soutenu et aidé dans ce travail durant ces quatres années.

Je remercie aussi particulièrement Yves Laprie pour son aide et ses précieux conseils en parole et pour avoir toujours accepté de se préter à des expériences occultes avec les capteurs électromagnétiques. Un grand merci aussi à Fabrice Hirsch pour sa disponibilité, sa grande patience et la gentillesse dont il a fait preuve pour les longues séances d'acquisitions de données.

Merci au Professeur René Anxionnat et à Sandrine Lefort du service neuroradiologie du CHU de Nancy, pour leur disponibilité et avoir su être à notre écoute pour mettre en place un protocole IRM comme dans mes rêves les plus fous.

C'est un plaisir et une chance de travailler dans un cadre comme celui du LORIA. Merci à tous ceux de Magrit, ex ou toujours en place : Brigitte, les deux Frédéric, Gilles, les trois Nicolas, Blaise, Cédric, Flavio, Ting, Shrikrishna, Evren, Diego, Sébastien. Merci à Isabelle pour sa gentillesse et ses réserves de stylos, ainsi qu'aux autres doctorants, Farid, Julien, Adrien, Nizar, Zainab. Une pensée aussi aux fans de Demis Roussos et de Nana Mouskouri, Asterios, Nassos et Tassos.

Et puis, toutes celles et ceux avec qui j'aime partager autre chose que des données articulatoires. Merci aux californiens Aude et Adrien ; à Gilles, Céline et Zoé et leur bétonnière avec qui j'ai vécu de grandes et belles choses ; à Manue et Séb, Claire et Momo, Julie, Linda et Fred, Magalie et Bob. Une pensée pour Ang-lem et son unique email par an, aux lyonnais Cécile et Fab, Aurélie et Nicolas et tout leurs enfants. Mille mercis à Audrey. Merci aussi aux indéboulonables Marco, JC, Mimil pour ces week-ends culturels dans les pubs londoniens et ces exquises fondues savoyardes. Merci à Lise, plein de courage pour la fin à toi aussi.

Enfin, un dernier gigantesque merci à mes parents, à ma soeur, à Philippe, Sylvie, ma crapouille, et à toute ma petite famille pour leur infaillaible soutient depuis que je suis tout petit. Je suis fier de vous avoir !

Table des matières

Table des figures

Liste des tableaux

Abbréviations

ACP	Analyse en Composantes Principales
DDL	Degré De Liberté
EM	ÉlectroMagnétique
ICP	Iterative Closest Point
IRM	Imagerie par Résonnance Magnétique
US	UltraSon
VV	Voyelle Voyelle, utilisé pour indiquer la production de deux Voyelles consécutives
VCV	Voyelle Consonne Voyelle, utilisé pour indiquer la production d'une Voyelle, puis d'une consonne, puis d'une voyelle. On trouve également **CVC** pour Consonne Voyelle Consonne, **VCCV** pour Voyelle Vonsonne Consonne Voyelle...

Introduction

L'acquisition et le traitement de données des mouvements des articulateurs (lèvres, langue, parois pharyngales, larynx) du conduit vocal constituent un enjeu crucial pour l'étude de la parole. En effet, l'amélioration de la qualité de ces données acquises à partir de capteurs d'image ou de position a permis de grandes avancées pour les modélisations acoustiques et articulatoires de la parole.

Il n'existe pas encore à l'heure actuelle de technique d'acquisition unique permettant de connaître dans l'espace et dans le temps la position de chacun des articulateurs du conduit vocal. En effet, les techniques modernes présentent toutes une faiblesse. Soit, elles ne capturent que partiellement les articulateurs (caméra vidéo pour suivre les lèvres et la mâchoire, échographie pour imager une partie de la surface de la langue, capteur électromagnétique à des positions précises à l'intérieur du conduit...). Soit, elles n'ont pas une résolution temporelle suffisante (par exemple, l'IRM qui nécessite plusieurs dizaines de secondes sans bouger pour acquérir des images tridimensionnelles du conduit). Soit, elles sont dangereuses, comme la cinéradiographie qui expose le sujet à des radiations nocives pour sa santé. Soit enfin, elles sont très invasives et modifient considérablement l'articulation (l'électropalatographie qui consiste à utiliser un palais artificiel pour chercher les points de contact avec la langue).

Les premières études réalisées avec la volonté d'associer plusieurs modalités d'acquisition de données articulatoires datent du début des années 1940, et ont été effectuées par Chiba et Kajiyama [CK41]. Leurs travaux associaient plusieurs systèmes d'imagerie médicale (photographie aux rayons X, palatographie et observation laryngoscopique du pharynx) afin d'étudier le comportement acoustique du conduit vocal. La fusion de ces informations visait à améliorer l'analyse et la compréhension des différents articulateurs, et ainsi contribuer à la compréhension et à l'interprétation des mouvements des articulateurs.

Parmi les champs d'application des études des mouvements des articulateurs, on peut citer entre autres :

- **la synthèse acoustique** : en parole, on sait générer un signal acoustique à partir des positions des articulateurs. À partir de ces positions observées sur des images rayons X, Fant [Fan60] fut l'un des premiers à mettre en place un modèle de production de parole à partir de la description de conduit vocal sous la forme d'une suite de quatre tubes, de section variable. Cette modélisation, bien que grossière, permet de générer le signal acoustique résultant des positions des articulateurs à un instant donné ;
- **l'inversion acoustique articulatoire** : le problème inverse de la synthèse articulatoire, c'est-à-dire la possibilité de récupérer automatiquement l'évolution temporelle de la forme du conduit vocal à partir du signal acoustique constitue aujourd'hui un problème majeur en traitement automatique de la parole [PLO04]. Dans l'idéal, l'inversion cherche à partir du seul signal acoustique à retrouver la position des articulateurs. Mais ce problème est mal posé et les efforts portent sur l'introduction, par exemple, de contraintes sur les formes et positions possibles du conduit ;

- **l'apprentissage de langues étrangères** : Engwall [Eng08] a montré qu'un retour visuel réaliste sur le positionnement de la langue permet à un individu d'améliorer sa prononciation. Son étude a en effet mis en exergue, dans le cas de locuteurs français, qu'un retour visuel sur les positions de langue à adopter pour parler suédois permettait effectivement d'améliorer leur prononciation. On peut alors songer à des têtes parlantes virtuelles et réalistes représentant les positions et mouvements des articulateurs du conduit vocal. Des applications directes dans l'apprentissage des langues sont possibles, ou encore en orthophonie en guidant les personnes malentendantes dans leur articulation ;
- **la médecine** : dans le cadre de glossectomies par exemple (ablation partielle de la langue), il est encore aujourd'hui très difficile pour les patients de réapprendre à parler. Une meilleure connaissance des corrélations entre la position de la langue et le son résultant permettrait de les aider [BTB+05]. On peut aussi penser à plus long terme à l'étude de prothèses adaptées aux mouvements de langue possibles.

Ces applications nécessitent avant tout une base de données articulatoires (ou corpus). Puisque chaque méthode d'acquisition existante n'apporte qu'une information partielle (spatiale ou temporelle), les recherches se concentrent sur l'aspect multimodal en combinant plusieurs données provenant de systèmes différents.

L'objectif de ce travail de thèse est de proposer un ensemble de méthodes permettant d'acquérir, de traiter et de fusionner automatiquement des données articulatoires multimodales. En effet, les rares systèmes multimodaux existants comportent de nombreux défauts. L'alignement temporel ou spatial des données, et/ou l'extraction des formes des articulateurs sont souvent effectués manuellement et sont conséquemment sujets à de nombreuses imprécisions. Ils utilisent parfois des informations qui n'ont pas été préalablement vérifiées, et reposent ainsi sur des données contructeurs pouvant être biaisées. Ces systèmes considèrent aussi parfois des hypothèses non validées comme, par exemple, la langue supposée avoir une forme identique qu'il y ait phonation ou pas. Ces nombreuses approximations et imprécisions sont susceptibles de fausser, voire d'empêcher, l'utilisation des données pour les applications de modélisation acoustique ou articulatoire. En d'autres termes, les données disponibles actuellement ne sont pas *bien fondées*.

Notre travail cherche à pallier ce manque en proposant des méthodes automatiques pour obtenir des données articulatoires multimodales, tout en mesurant la confiance à apporter à chaque donnée acquise et traitée. Nous vérifierons que l'ensemble des données peut être utilisé pour des applications de modélisation acoustique ou articulatoire. Pour cela, nous mettrons l'accent sur les méthodes à utiliser afin d'obtenir automatiquement des corpus les plus complets possible. Nous prendrons un soin particulier à évaluer indépendamment la précision spatiale et temporelle de chacune des modalités, ainsi que la précision globale des données fusionnées. Nous démontrerons par une première exploitation du corpus acquis la validité des données en les évaluant sur le modèle articulatoire de Maeda [Mae79], et nous proposerons un nouveau modèle de langue. L'ensemble de ces méthodes appliquées à nos données en feront un ensemble de données articulatoires *bien fondées*.

Nous utiliserons pour ces travaux des techniques d'imagerie (échographie, IRM, stéréovision) ainsi que des données électromagnétiques, et définirons des protocoles d'acquisition dédiés permettant d'obtenir à la fois des données statiques et dynamiques tridimensionnelles sur le conduit vocal. Des techniques utilisées en vision par ordinateur seront adaptées afin de recaler spatialement ces données (calibrage et recalage 3D/3D), d'en extraire automatiquement les contours pour les images échographiques (suivi), et de les synchroniser.

Ce travail de thèse s'inscrit dans le cadre du projet européen ASPI[1], débuté en décembre 2005 et terminé en janvier 2009. Ce projet incluait plusieurs partenaires universitaires européens : l'équipe *Multimodal speech synthesis group* du *Kungliga tekniska högskolan (KTH)* à Stockholm en Suède, l'équipe *Computer Vision, Speech Communication and Signal Processing Group* de la *National Technical University of Athens (NTUA)* à Athènes en Grèce, des membres de *l'Université Libre de Bruxelles (ULB)* à Bruxelles en Belgique, des membres du département *Traitement du Signal et de l'Image* de *Télécom ParisTech (ENST)* de Paris en France, ainsi que les équipes *Parole* et *Magrit* du *LORIA* de Nancy. L'objectif de ce projet consistait en l'étude de méthodes d'inversion acoustique articulatoire. Intégrés dans ce projet global, nos travaux ont consisté à mettre en place un système d'acquisition et des méthodes associées afin d'obtenir des données articulatoires *bien fondées* pour l'évaluation des méthodes d'inversion.

Ce mémoire se compose de six chapitres : le premier présente le système articulatoire humain et établit un état de l'art sur les différents procédés utilisés pour l'acquisition de données, ainsi que des modèles articulatoires existants. Le second présente les objectifs de notre travail en détaillant le cahier des charges nécessaire à l'obtention de données articulatoires bien fondées. Il propose aussi une critique des bases de données articulatoires multimodales existantes. Le troisième chapitre traite de notre stratégie d'acquisition de données dynamiques, et le quatrième de leurs traitements spécifiques afin d'être exploitable en tant que données articulatoires. Le cinquième chapitre définit ensuite un protocole original d'acquisition de données statiques à partir d'IRM et présente la façon dont elles ont été recalées avec les données statiques. L'incertitude du recalage faisant intervenir toutes les modalités du système y est ensuite quantifiée. Et enfin, dans le sixième et dernier chapitre, nous détaillons les données acquises avec notre système, présentons une méthode d'évaluation des données recalées dans un contexte de parole, et concluons enfin par des perspectives de recherche.

[1] AudioviSual to articulatory SPeech Inversion, financé par le Programme IST de la Commission de la Communauté Européenne avec le numéro IST-2005-021324, `http://aspi.loria.fr`

Chapitre 1

Données articulatoires et modélisation du conduit vocal

L'objectif de ce chapitre est de présenter succinctement au lecteur le vocabulaire anatomique employé dans le manuscrit, les différentes techniques possibles pour l'acquisition de données articulatoires humaines, et leur utilisation pour la modélisation du conduit vocal.

1.1 Le conduit vocal

Le conduit vocal peut être découpé entre quatre zones d'intérêt (cf figure 1.1) : la cavité buccale, les fosses nasales, le pharynx et le larynx. Il permet d'assurer les fonctions de déglutition, de respiration, et aussi celle de phonation (i.e. la production d'un son), sur laquelle nous nous concentrerons dans ce manuscrit.

1.1.1 Description anatomique

Les descriptions ne se veulent pas exhaustives, mais présentent les bases anatomiques importantes utiles à la compréhension de ce manuscrit. Il est possible de se référer à un atlas anatomique [KHP78] ou [SSS+06] pour plus de détails.

1.1.1.1 Cavité buccale

La cavité buccale est délimitée par les lèvres à l'avant, le plancher buccal en bas, le palais en haut, les joues sur les côtés, et communique avec le pharynx à l'arrière par l'isthme du gosier et l'oropharynx. Elle comprend entre autres des structures rigides (dents, palais) et un organe déformable (la langue), qui seront plus détaillés dans la partie 1.1.2. La cavité buccale est protégée par une structure osseuse, la mâchoire, dont la partie inférieure est appelée la mandibule (ou maxillaire inférieur).

1.1.1.2 Fosses nasales

Les deux fosses nasales, situées au-dessus de la cavité buccale, jouent un rôle très important dans la respiration et la phonation de voyelles ou de consonnes nasalisées (en français, /$\tilde{o}$/ de « bon », /$\tilde{a}$/ de « sans » , /$\tilde{\varepsilon}$/ de « brun »par exemple). Elles sont limitées à l'avant par le nez et communiquent à l'arrière avec le nasopharynx, par l'intermédiaire du voile du palais et de l'uvule palatine (ou luette). L'abaissement ou le relèvement de ces derniers contrôle l'écoulement d'un

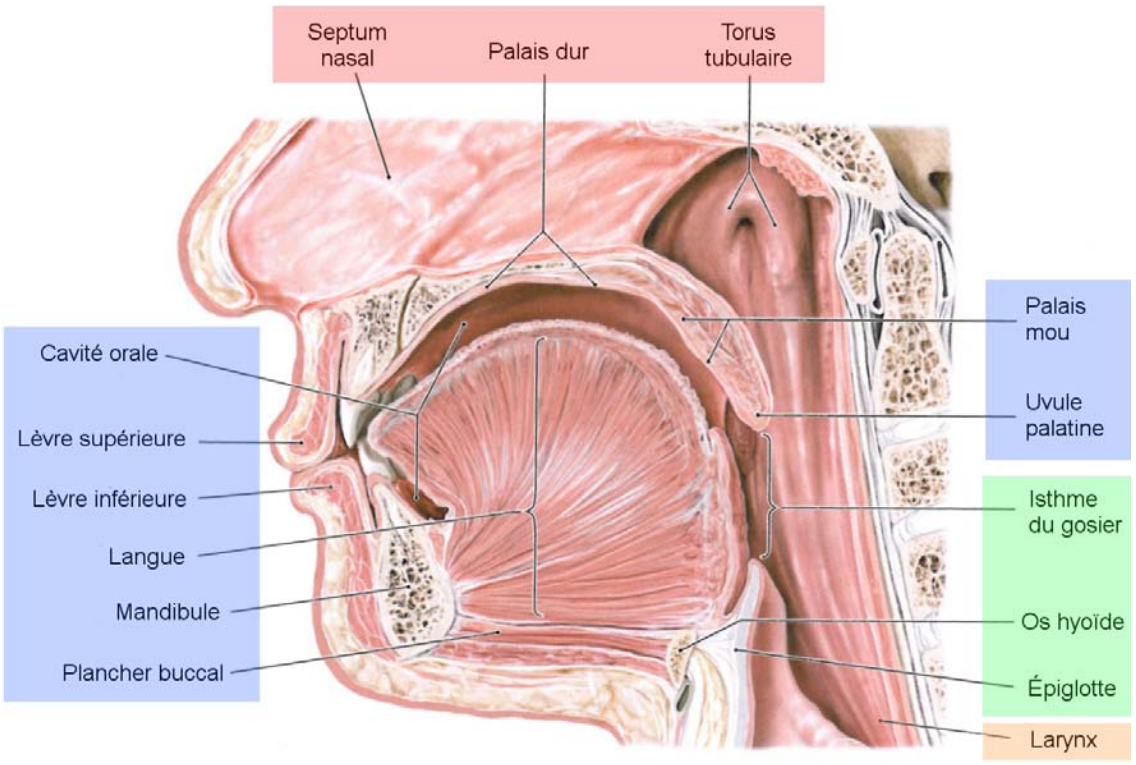


FIG. 1.1 – Schéma du conduit vocal. En rose les fosses nasales, en bleu la cavité buccale, en vert le pharynx et en orange le larynx. D'après [SSS+06].

flux d'air dans les fosses nasales. Pour une étude détaillée des fosses nasales et de leur rôle dans la production de la parole, nous reportons le lecteur à [Ser06].

1.1.1.3 Pharynx

Le pharynx est un conduit musculo-membraneux déformable reliant à son sommet la cavité buccale et les fosses nasales et à son pied le larynx. C'est donc le point de jonction des zones présentées sur la figure 1.1. Il est découpé en trois parties :

- le nasopharynx, en arrière de la cavité nasale,
- l'oropharynx constituant la zone d'interface entre le nasopharynx et le voile du palais en haut, en avant avec l'isthme du gosier et la cavité buccale et en bas avec le laryngopharynx,
- la laryngopharynx constituant la partie inférieure avec notamment l'épiglotte.

1.1.1.4 Larynx

Le larynx est un conduit cartilagineux faisant suite au pharynx et qui assure le lien avec la trachée. Il abrite notamment les cordes vocales, replis de membranes muqueuses dont la vibration est utilisée dans la phonation. L'os hyoïde (ou os lingual) se trouve juste au-dessus du larynx, au-dessous de la base de la langue. Cet os, qui est le seul du corps humain à ne pas être attaché au squelette, est relié à un grand nombre de muscles du pharynx, du larynx et de la langue. Nous verrons par la suite que la présence de cet os à une influence importante sur la formation des images échographiques.

1.1.2 Les principaux articulateurs dans la parole

1.1.2.1 La production de la parole

Nous présentons ici le principe général de la production de la parole. Il est possible de se référer aux ouvrages de [Fla72, Ken97, Ste99] pour plus de détails.

Les sons de parole sont produits par un déplacement du flux d'air dans le conduit vocal. Ce flux d'air, en provenance des poumons, entre dans le conduit du larynx puis entre dans les cordes vocales, qui pour les voyelles et les autres sons voisés, se mettent à vibrer de manière pseudo-périodique (à une fréquence appelée la fréquence fondamentale). L'espace entre les deux cordes vocales par lequel passe l'air est la glotte.
Le flux d'air sortant, l'air laryngé, passe alors dans les cavités supraglottiques que forment le pharynx, la cavité buccale et les fosses nasales. Ces cavités constituent un ensemble de résonateurs acoustiques, dont les formes et les volumes varient au cours du temps grâce aux articulateurs qui leurs sont associés. Ces articulateurs peuvent alors totalement bloquer l'air dans le conduit buccal (cas des occlusives), le laisser passer dans une région très étroite (cas des fricatives) ou plus large (cas des voyelles). Les variations de position de ces articulateurs permettent de produire les différents sons de la parole. Dans la suite de ce manuscrit, on appellera constriction la zone du conduit vocal la plus étroite durant la production d'un son.

En nombre, peu d'articulateurs sont impliqués dans la production de la parole. En revanche, ces articulateurs adoptent de très nombreuses formes et positions dans le conduit vocal. Ces nombreuses combinaisons possibles font toute la force et la richesse de la parole. La figure 1.2 répertorie les différents points d'articulation (l'endroit où s'effectue l'obstruction au passage de l'air) lors de la phonation de consonnes, avec les différents termes qui leur sont associés. Elle permet d'avoir un aperçu de toutes les zones du conduit vocal utilisées, et de se rendre compte du nombre important de lieux d'articulation pouvant être sollicités par les articulateurs pour la production de la parole.

En phonologie, un phonème est un contraste phonétique qui devient porteur de sens dans un langue. Ainsi leur fonction dans une langue est d'établir des oppositions entre les mots de son lexique [Vai06]. Il est aussi susceptible d'être prononcé de façon différente selon les locuteurs ou selon sa position et son environnement au sein du mot (phénomène de coarticulation). Dans ce manuscrit, nous utiliserons la notation de l'alphabet phonétique international[2] pour représenter un phonème.

Ce travail de thèse consiste précisément à utiliser des techniques d'imagerie et de capteurs pour l'étude des formes des articulateurs et de leurs mouvements. Nous nous sommes concentrés sur les articulateurs les plus mobiles et les plus déformables de la cavité buccale : la langue, le palais et les lèvres. Nous les présentons plus en détail dans les sections suivantes.

1.1.2.2 La langue

La langue est un organe musculo-membraneux, de forme ovoïde et reposant sur le plancher de la cavité buccale. Elle est composée de muscles internes, de muscles externes attachés à des os ou des organes voisins, et d'un revêtement muqueux sur lequel reposent les papilles. Elle est liée à l'os hyoïde, à la mandibule, au palais et au plancher buccal par les muscles, et au pharynx par sa muqueuse.

La muqueuse linguale est composée de trois parties :

[2] `http://www.langsci.ucl.ac.uk/ipa`

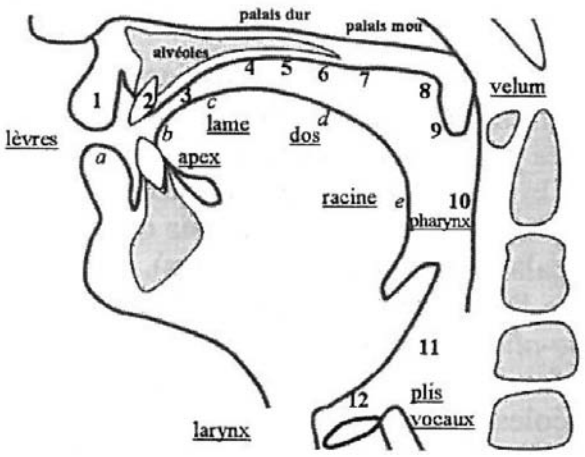


FIG. 1.2 – Dénomination du lieu de l'articulation des consonnes, d'après [Vai06]. Lieu d'articulation : 1. Labiale ; 2. Dentale ; 3. Alvéolaire ; 4. Prépalatale ; 5. Médiopalatale ; 6. Postpalatale ; 7. Prévélaire ; 8. Vélaire ; 9. Uvulaire ; 10. Pharyngale ; 11. Laryngale ; 12. Glottale.
a. Apicale ; b. Laminale ; c.Prédorsale ; d. Médiodorsale ; e. Postdorsale/radicale.

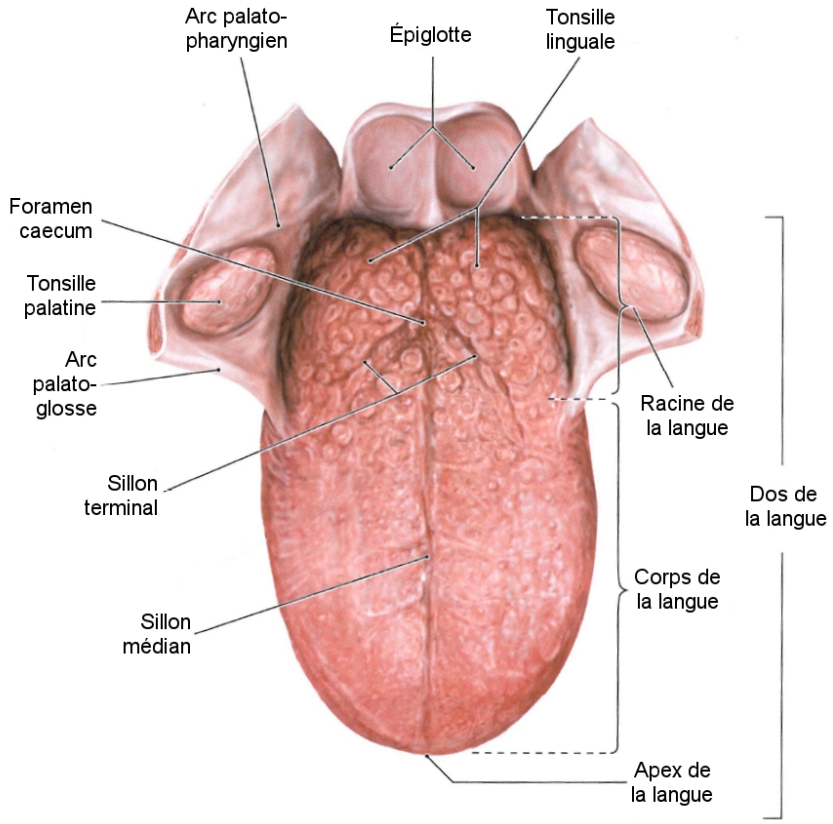


FIG. 1.3 – Schéma anatomique de la langue. D'après [SSS+06].

- la face inférieure de la langue, reliée au plancher buccal. La muqueuse y est fine, et présente un repli muqueux médian, appelé le frein de la langue,

- le dos de la langue composée d'une muqueuse épaisse et adhérente dans la partie buccale, et moins adhérente dans la partie pharyngée. Il est composé de nombreuses papilles de formes différentes. Un sillon peu profond, médian, subdivise le dos de la langue en moitiés droite et gauche, et sépare le dos de la langue de sa racine,
- la racine (ou base) de la langue.

Par la suite, nous utiliserons les termes suivants pour caractériser le dos de la langue (cf figure 1.3) :

- la pointe de la langue, ou l'apex, située au-dessus du plancher buccal,
- le dos de la langue situé en-dessous du palais et dans le pharynx,
- la racine de la langue située à côté de l'épiglotte.

Pour une description plus détaillée du rôle de chacun des muscles de la langue, on réfère le lecteur à [Buc07] et à [Tod09]. La langue constitue l'articulateur le plus mobile du système supraglottique de l'appareil vocal, et participe très activement à la production de la parole [Lad01].

1.1.2.3 Le palais

Le toit de la cavité buccale est formé dans ses deux tiers antérieurs par le palais dur, dans son tiers postérieur par le voile du palais (le velum).

Seul le voile du palais est mobile durant la production de la parole, et représente un articulateur du conduit vocal. Il est constitué d'une membrane et de muscles, et la luette y est attaché. Il permet l'isolement des fosses nasales de la cavité buccale lors de l'articulation de certains phonèmes en s'ouvrant ou se refermant. Il permet de laisser ainsi passer ou pas un flux d'air dans les fosses nasales.

Le palais dur, appelé aussi voûte palatine, est statique par rapport au conduit vocal. La langue a souvent des points de contact avec cette zone du palais, afin de contrôler le flux d'air entrant dans la cavité buccale.

1.1.2.4 Les lèvres

Les lèvres sont les articulateurs externes du conduit vocal. Ce sont deux replis musculo-membraneux très mobiles : les lèvres inférieures et supérieures. Elles sont connectées en leurs extrémités pour former les commissures labiales. Recouvertes de peau, elles sont attachées à des muscles constricteurs et dilatateurs, permettant leurs mouvements.

Lors de la phonation, ce sont les derniers articulateurs que rencontre l'air avant d'être expiré de la cavité buccale. La position des lèvres, plus ou moins fermées, étirées ou pincées finit de moduler le son émis.

1.2 Les modèles articulatoires

Afin d'étudier la production de la parole, de nombreux modèles du conduit vocal ont été proposés. Ces modélisations, plus ou moins élaborées, ont toutes le même objectif : mettre en place un modèle de conduit vocal, souvent commandé par un nombre réduit de paramètres, et ayant un comportement réaliste pour pouvoir simuler la position et le mouvement des articulateurs. Une application directe est la synthèse articulatoire qui, à partir des formes du conduit, consiste à générer le son résultant. Mais on peut aussi penser à des applications en inversion acoustique articulatoire, en médecine ou pour le rendu réaliste de têtes parlantes.

Le plus souvent, les modèles articulatoires représentent la coupe médiosagittale de la tête. En effet, c'est pour cette représentation dans le plan médian du corps humain que l'on dispose

de données de bonne qualité, en assez grande quantité, et qui fournissent suffisamment d'informations pour pouvoir effectuer une modélisation acoustique. Ces modèles 2D représentent aujourd'hui l'essentiel des modèles articulatoires développés dans la littérature, même si, depuis quelques années, on voit arriver de plus en plus de modèles basés sur des données en trois dimensions reposant sur de nouvelles méthodes d'acquisition plus évoluées.

Cette partie ne se veut pas une description exhaustive des différents modèles existants, mais une revue décrivant les différentes classes de modèles, en présentant brièvement l'un des plus représentatifs de chaque classe.

1.2.1 Les modèles à fonction d'aire

La modélisation la plus simple du conduit vocal, mais qui reste encore très utilisée pour la synthèse acoustique, consiste à calculer sa fonction d'aire, c'est-à-dire la donnée de l'aire à la section transversale du conduit vocal le long de la courbe médiane du conduit, de la glotte à l'ouverture des lèvres. Avec cette approche, on considère que le conduit vocal est assimilable à un conduit droit de section variable. Fant [Fan60] est l'un des pionniers dans le calcul des fonctions d'aire, en modélisant le conduit vocal par quatre tubes de rayons variables, chaque tube représentant une région du conduit vocal : un tube pour l'air du larynx et du pharynx, un tube pour la zone de constriction, un tube pour la cavité buccale et un tube pour les lèvres. Chaque son correspond à une configuration avec des tubes de rayons et de longueurs différentes. Malgré la simplicité de ce modèle, il permet de représenter schématiquement les configurations articulatoires des sons de la parole et de les synthétiser.

Cependant, ce type de modèle ne cherche pas à représenter fidèlement le conduit vocal au sens anatomique, mais à simuler le comportement du passage de l'air dans le conduit pour en synthétiser le son.

1.2.2 Les modèles géométriques

Les modèles géométriques représentent chacun des articulateurs du conduit vocal par une forme géométrique simple. Par exemple, les travaux de Coker [CF66] puis de Mermelstein [Mer73] modélisent le corps de la langue par un cercle. Chacun des articulateurs est alors piloté par un nombre variable de paramètres qui agissent comme des commandes sur ces formes : translations, rotations, déformations... Le modèle de Mermelstein a été testé à de nombreuses reprises en synthèse acoustique pour décrire les configurations du conduit vocal sur un petit nombre de voyelles et de consonnes.

Le principal défaut de ces modèles est qu'ils ont été élaborés de façon ad hoc, parfois à partir d'images réelles (rayons X), mais également à partir d'expertise humaine et d'intuition. Ils sont de plus limités à des formes géométriques simples qui ne rendent pas compte de la complexité anatomique du conduit vocal. Par ailleurs, ils nécessitent de nombreux paramètres pour modéliser le conduit vocal.

1.2.3 Les modèles statistiques

Une alternative à l'approche géométrique est l'élaboration de modèles à l'aide d'une analyse factorielle basée sur des données articulatoires réelles. On s'est aperçu dès les premières observations d'images réelles (souvent des images cinéradiographiques) qu'il existe beaucoup de redondance dans les formes possibles du conduit vocal. En appliquant une analyse factorielle à ces données, on peut donc décrire avec un nombre réduit de modes orthogonaux et/ou non corrélés la majeure partie de leur variabilité.

Les modèles statistiques sont construits uniquement à partir de données réelles, contrairement aux modèles géométriques. Par conséquent, ils sont censés caractériser de façon bien plus réaliste les formes et mouvements des articulateurs. Cependant, cet avantage est aussi leur défaut : ils sont entièrement dépendants des données. Si celles-ci comportent de fausses informations (mauvais traitements, mauvais détourage...) ou ne décrivent que de façon partielle l'ensemble des formes adoptables par les articulateurs, alors le modèle résultant souffrira de ces faiblesses.

L'un des plus connus est le modèle de Maeda [Mae79] qui décrit les formes de conduit vocaux à partir de contours dessinés manuellement sur des images rayons X. Nous détaillons ce modèle, car il sera utilisé dans le chapitre 6 de ce manuscrit. Dans le plan médiosagittal, le conduit vocal est décomposé en trois parties indépendantes (les lèvres, la langue et le larynx) qui sont seulement influencées par la position de la mâchoire inférieure, la mandibule. L'analyse factorielle proposée par Maeda pour traiter les données tient compte de cette influence pour soustraire le mouvement de la mâchoire aux autres articulateurs : il s'agit d'une analyse en composantes orthogonales arbitraires (proposée par Overall [Ove62], aussi appelée analyse en composantes principales guidée). Une fois ce mouvement de mâchoire soustrait, chaque zone du conduit vocal (lèvres, langue, larynx) peut alors être étudiée indépendamment. Pour chacune des zones, des paramètres de contrôle sont obtenus par une analyse en composantes principales (ACP) sur les données décorrélées de l'influence de la mâchoire, en retenant suffisamment de composantes pour expliquer l'essentiel de la variance. Le nombre de paramètres est variable suivant la zone ; pour les lèvres, deux paramètres sont nécessaires : ouverture verticale, et protrusion (distance lèvres-mâchoire) ; pour le larynx, un seul paramètre suffit ; pour la langue, trois paramètres sont nécessaires pour décrire 96% de la variance des données. En ajoutant un paramètre pour la position de la mâchoire, un total de 7 paramètres permet de décrire l'ensemble des déformations du conduit vocal (cf figure 1.4). Notons enfin que ce modèle a été établi à partir de données provenant d'un seul locuteur : Meada propose dans [Mae92] une technique pour adapter ce modèle à différentes morphologies de locuteurs, en étirant ou rétrécissant la taille globale du conduit.

La très grande majorité des études effectuées ont concerné le plan médiosagittal, car beaucoup de méthodes d'acquisition de données permettent d'obtenir des informations seulement sur ce plan. Avec les récents progrès sur les systèmes d'acquisition en trois dimensions, il commence cependant à apparaître des modèles statistiques 3D, basés sur le même principe que celui de Maeda : Badin [BBR$^+$02] propose un modèle articulatoire basé sur une ACP en trois dimensions sur des données acquises à partir d'images vidéos des lèvres et d'images IRM du conduit vocal : cinq paramètres sont nécessaires pour décrire 72.2 % de la variance des données sur la langue, cinq paramètres pour les lèvres (96.6 % de la variance totale), le larynx n'ayant pas été étudié. Nous reviendrons dans le chapitre 4 sur les données utilisées pour la mise en place de ce modèle.

1.2.4 Les modèles biomécaniques

La forme la plus complexe, mais aussi la plus complète, de modèles de conduits vocaux repose sur l'intégration d'un maximum de propriétés physiologiques des articulateurs et l'étude de leurs interactions avec des éléments externes (os, muscles...). Dans ce type de modèle, les articulateurs sont décomposés en un sous-ensemble d'éléments, chacun pouvant avoir des propriétés différentes.

Le modèle biomécanique le plus répandu est le modèle masse-ressort où chaque élément possède une masse et est lié à un autre élément par un ressort ayant une masse et une constante élastique propres. Chaque élément a alors un mouvement caractérisé par des forces dépendant des masses et des constantes élastiques des éléments qui lui sont liés.

Le premier modèle physiologique de la langue a été développé par Perkell [Per74] pour ses

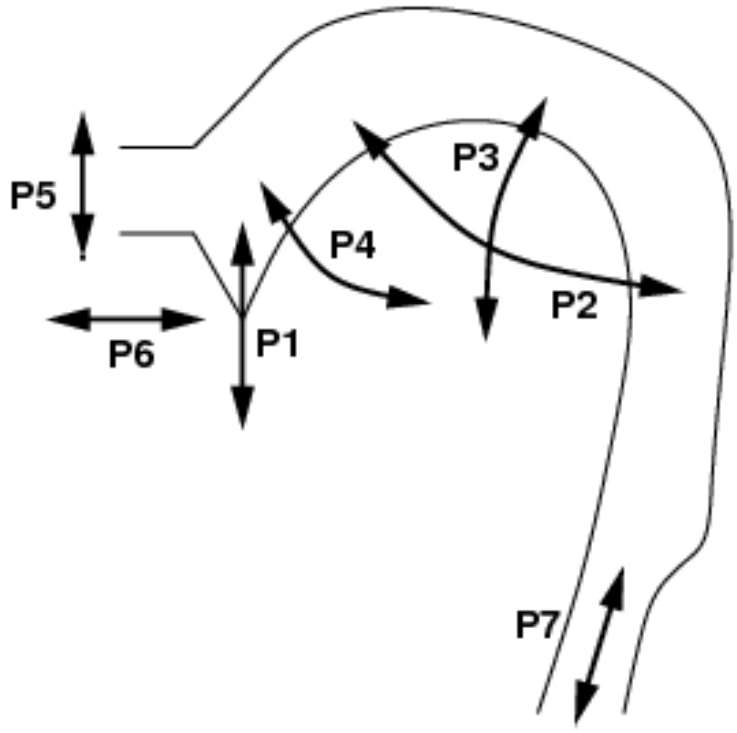


FIG. 1.4 – Les sept paramètres du modèle de Maeda : la mâchoire P1, l'ouverture P5 et la protrusion des lèvres P6, la position du corps de la langue P2, la forme de la langue P3, un terme contrôlant la pointe de la langue P4, et enfin la hauteur du larynx P7.

travaux de thèse en 1974. Son but était d'élaborer un modèle biomécanique et dynamique de la langue pour comprendre les relations existant entre les propriétés phonétiques et les formes physiquement réalisables des articulateurs. Le modèle, établi en 2D dans le plan médiosagittal, est composé de 16 éléments masse-ressort disséminés dans la langue (8 à la surface, et 8 en son milieu), accrochés à des éléments générateurs de tension. Ces derniers se décomposent en deux sous-catégories : 28 éléments actifs, correspondant aux tissus musculaires capables de développer des forces en réponse à une stimulation et 47 éléments passifs, représentant les tissus conjonctifs et les structures molles et rigides du conduit vocal. Pour répondre aux exigences d'incompressibilité de la langue, les éléments actifs sont reliés entre eux en formant des quadrilatères dont l'aire ne peut pas dépasser un seuil minimal. Chacun des éléments est aussi caractérisé par des forces de frottement, de friction et de glissement.

Le modèle de Perkell a influencé tous les autres modèles physiologiques de conduits vocaux mis en place par la suite. Parmi les travaux récents, on peut citer ceux de Gérard et Wilhelms-Tricario [GWTPP03] qui se basent sur la théorie des éléments finis pour une modélisation tridimensionnelle de la langue. Leur approche prend en compte les propriétés physiques non-linéaires (basées sur les lois physiques de l'élasticité non linéaire) pour ajouter au réalisme de leur modèle. D'autres travaux récents, comme ceux [VLB$^+$08] ou [GOL$^+$04] s'attachent à modéliser tout le conduit vocal en utilisant un modèle biomécanique. L'inconvénient majeur des modèles biomécaniques est que beaucoup de paramètres doivent être déterminés. Ainsi, dans l'idéal, il faut imager le conduit pour fixer la géométrie du locuteur. Il faut aussi disposer des outils nécessaires pour mesurer les potentiels électriques musculaires afin de connaître et/ou vérifier les commandes. Bien qu'étant la solution de modélisation la plus réaliste, cette approche demeure encore aujour-

d'hui très coûteuse en ressources humaines et matérielles. Il apparaît de plus très fréquemment des instabilités numériques liées à la complexité des modèles.

Depuis quelques années, des ressources (University of British Columbia à Vancouver au Canada, GIPSA-lab à Grenoble en France, TIMC à Grenoble en France) sont mises en commun sous l'initiative de Fels [FSH+09] pour la mise en place d'un modèle biomécanique le plus complet possible *ArtiSynth*. Le site `http://www.magic.ubc.ca/artisynth` permet de télécharger et tester librement le modèle, permettant de suivre ses dernières évolutions.

1.2.5 Conclusion

Tous les modèles présentés, qu'ils soient géométriques, statistiques, biomécaniques, ou qu'ils simulent le conduit vocal par des fonctions d'aires, dépendent tous de données articulatoires soit pour les construire, soit pour les évaluer, soit les deux. La même problématique relie donc tous ces modèles, à savoir qu'ils nécessitent des données articulatoires pour être comparés, validés, améliorés ou modifiés. Nous nous intéressons dans la section suivante aux différentes méthodes d'acquisition de données articulatoires existantes.

1.3 Les méthodes d'acquisition

Dans le but d'établir des modèles articulatoires réalistes et cohérents, de très nombreuses méthodes ont été testées sur des sujets (ou locuteurs) pour acquérir des données articulatoires du conduit vocal. S'il existe des méthodes d'acquisition physiologiques (l'électromyographie par exemple qui mesure des courants électriques à partir d'électrodes collées sur le visage) ou encore aérocoustiques (mesures de flux d'air...), il est nécessaire de disposer de méthodes permettant d'obtenir des informations anatomiques de position et de mouvement pour les confronter à des modèles articulatoires. Parmi elles, on dénote des techniques d'imagerie (échographies, IRM, cinéradiographie...) et des techniques permettant de récupérer la position de points (articulographie, palatographie...).
Dans l'idéal, la technique d'obtention de données articulatoires devrait :

- couvrir la totalité du conduit vocal et des articulateurs en trois dimensions,
- avoir une fréquence temporelle d'acquisition suffisante pour capturer tous leurs mouvements. On estime que 60 Hz est le seuil inférieur nécessaire pour une observation précise des mouvements articulatoires [MBE+06],
- ne pas être nocive pour le sujet,
- ne pas perturber l'articulation,
- capturer un signal acoustique de bonne qualité.

Malheureusement, aucun système actuel ne satisfait entièrement ces conditions. Toutes les techniques ont des contraintes spécifiques qui ont forcément une influence sur les mesures acquises. Nous présentons dans la section suivante différentes techniques d'imagerie et de capture de position pour récupérer la forme des articulateurs du conduit vocal.

1.3.1 Cinéradiographies, rayons X

Les données cinéradiographiques - ou rayons X - ont longtemps été les plus utilisées pour l'observation d'images du conduit vocal [Dar87]. Les rayons X, découverts par Röntgen à la fin du XIX^{e}siècle, sont des rayonnements électromagnétiques d'énergie suffisamment grande pour qu'une partie du rayonnement traverse les objets tandis que le reste interagit avec le milieu traversé. Le rayonnement subit donc des atténuations avant d'être mesuré par un capteur image.

Les images résultantes offrent un très bon compromis entre résolution spatiale (0.3 mm/pixel) et fréquence d'acquisition (50 images par seconde). Cependant, un pixel de l'image représente l'intégration en un seul point de toutes les différentes atténuations subies par le rayon incident en traversant les tissus. Par conséquent, il est parfois très difficile de distinguer les structures en raison de leur superposition dans l'image.
Par exemple sur la figure 1.5, la langue donne lieu à plusieurs contours à cause d'une concavité longeant le sillon de la langue. Cela explique qu'il soit difficile de détecter ces contours automatiquement dans les images rayons X, et que les tracés utilisés aujourd'hui sont encore effectués manuellement. Soumis à une appréciation subjective, ils restent souvent entachés d'imprécisions, voire parfois d'erreurs.

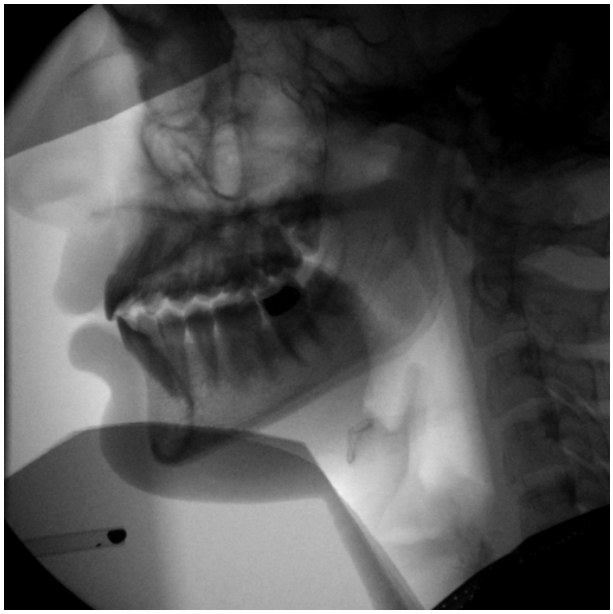


FIG. 1.5 – Image rayons X : les contours sont souvent très difficiles à distinguer, car plusieurs structures se superposent pour un même point de l'image.

Mais le principal inconvénient de cette méthode est qu'elle est nocive pour le sujet : lorsque le rayonnement heurte un atome de matière, l'énergie du choc permet d'éjecter un électron et de modifier ainsi la matière traversée. Les effets sur le patient vont de brûlures localisées (érythèmes) aux cancers. Pour des raisons éthiques et sanitaires évidentes, la cinéradiographie a donc dû être abandonnée au début des années quatre-vingt dans un cadre de recherche.

1.3.2 Micro-faisceaux de rayons X

En 1975, [KIF75] a utilisé les micro-faisceaux de rayons X pour suivre de petits marqueurs (2-3 mm de diamètre) collés sur les articulateurs. Le principe physique reste le même que pour les rayons X, mais la dose de rayons reçue par le sujet est beaucoup plus faible. Il s'agit dans ce cas précis, non plus d'obtenir des images complètes du conduit vocal, mais de suivre quelques points définis au préalable. Ce suivi s'effectue seulement dans un plan, le plus souvent dans le plan médiosagittal du visage.
Cette technique, outre le fait qu'elle est onéreuse, expose tout de même le sujet à des rayons X et son utilisation dans un contexte de recherche est interdite à cause des radiations auxquelles est soumis le sujet. De plus, elle a rapidement été supplantée par une autre méthode d'acquisition

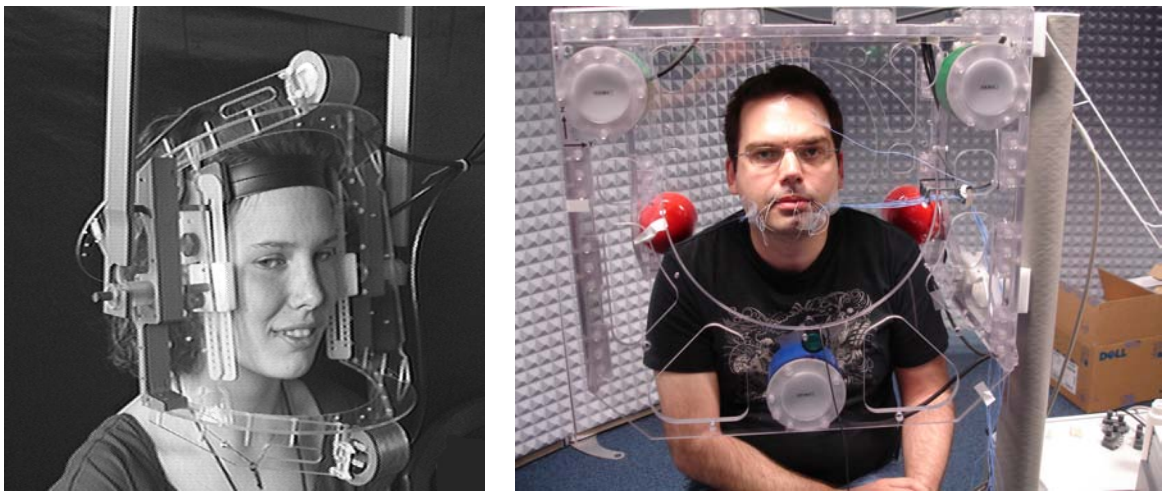
basée elle aussi sur le suivi de marqueurs et qui n'est pas nocive : les acquisitions électromagnétiques.

1.3.3 Données électromagnétiques

L'articulographe (aussi appelé EMA, pour ElectroMagnetic Articulograph) est fréquemment utilisé depuis une dizaine d'années pour suivre des points physiques du conduit vocal grâce à des capteurs de mouvement. Le principe est de coller des capteurs miniatures formés de bobines électriques sur les tissus dont on veut suivre l'évolution dans l'espace et dans le temps. Ces capteurs sont reliés par un fil au système pour pouvoir enregistrer la puissance du champ magnétique reçu par la bobine. Ensuite, on place les capteurs dans un champ magnétique créé par des émetteurs. La puissance du champ enregistrée par chaque bobine est alors inversement proportionnelle à sa distance à l'émetteur, permettant ainsi de retrouver leur position dans l'espace.

Les principaux avantages de tels systèmes sont qu'ils permettent l'acquisition de données à de hautes fréquences - près de 200 Hz pour certains systèmes actuels -, et pour différents articulateurs. On obtient donc une série de données dynamiques, représentant l'évolution spatiale d'un point au cours du temps. De plus, il n'est pas nécessaire de disposer d'une ligne de vue comme c'est le cas avec un système optique pour lequel les capteurs doivent être constamment visibles par une caméra.

Différents articulographes existent : celui du MIT [PCS+92], le Botronic Movetrack [Bra85], et les plus utilisés par la communauté parole, les systèmes AG100, AG200 et AG500 (cf figure 1.6) de Carstens (http ://www.articulograph.de). Les articulographes AG100, AG200, du MIT et de Botronic sont des modèles historiques, pour lesquels le sujet devait porter un casque stabilisé sur sa tête, ce qui pouvait être gênant pour le locuteur. Cette contrainte a disparu avec le système AG500, où la tête du sujet est placée dans une cage, le laissant plus libre de ses mouvements. De plus, seul le système AG500 permet de récupérer des informations en trois dimensions et non plus seulement dans le plan médiosagittal, comme avec les modèles AG100 et AG200.



(a) (b)

FIG. 1.6 – Articulographes : (a) AG200 (sans les capteurs collés sur le locuteur), extrait de la documentation officielle. (b) AG500 utilisé au LORIA.

De nombreuses études ([Hoo93, ZHFE07]...) ont été réalisées à partir de données EMA. L'ar-

rivée de l'AG500 étant relativement récente (aux alentours de 2005), il existe encore très peu de recherches portant véritablement sur les positions 3D des articulateurs, la plupart se concentrant toujours sur des capteurs placés dans le plan médiosagittal. Une configuration typique utilise six capteurs fixés sur les articulateurs : deux pour les lèvres inférieures et supérieures, trois pour la langue (en moyenne à 8 mm, 20 mm et 52 mm de l'apex, ces valeurs dépendant du locuteur), et un sur la mâchoire inférieure. De plus, deux capteurs supplémentaires sont fixés sur le visage (arête du nez) pour compenser les mouvements de tête.

Concernant la précision de ce système, Kroos a montré dans une récente étude [Kro08] qu'elle variait beaucoup selon les mouvements des capteurs. Il a comparé les mesures fournies par les capteurs en les fixant sur un système de suivi optique (Optotrack, NDI) dont la précision des mesures est de l'ordre de 0.15 mm. Les résultats montrent que si on peut s'attendre à une erreur moyenne inférieure à 1 mm pour des mesures statiques, les mesures dynamiques des capteurs en mouvement donnent des erreurs de l'ordre de 4 mm. Il n'a cependant pas pu vérifier s'il y avait une corrélation entre la vitesse du mouvement et les imprécisions résultantes sur les mesures.

Par ailleurs, il n'y a pas à notre connaissance d'études portant sur la gêne que procurent les capteurs sur les articulateurs et les fils sortant de la bouche pour les capteurs collés dans la cavité buccale. Même si leur taille est petite (inférieure à 5 mm dans les trois dimensions pour la majorité des modèles), il serait intéressant d'étudier les modifications qu'ils apportent dans le processus de phonation. La grande majorité des personnes utilisant ce type de système considèrent par défaut que leurs effets sont négligeables.

1.3.4 Échographie

Un échographe est constitué d'une sonde et d'un système informatique affichant l'image traitée. Les échographes sont couramment utilisés dans l'imagerie médicale pour l'acquisition d'images dans un plan 2D.

Le principe de fonctionnement est le suivant : un cristal de céramique (piézo), situé dans la sonde, est soumis à des impulsions électriques, vibre et émet des ultrasons (US). Ces derniers se propagent alors dans les tissus humains et les échos qu'ils renvoient sont captés par la même sonde.

Chaque milieu traversé par l'ultrason possède une impédance acoustique (qui caractérise la vitesse de propagation de l'ultrason dans le milieu concerné) propre. Ce sont ces différences d'impédance acoustique entre deux milieux qui sont à l'origine des différences d'amplitudes observées lorsque la sonde reçoit le signal écho. Ces échos renvoyés sont donc des signatures des obstacles rencontrés par le signal, et caractérisent les interfaces traversées par le faisceau ultrasonore.

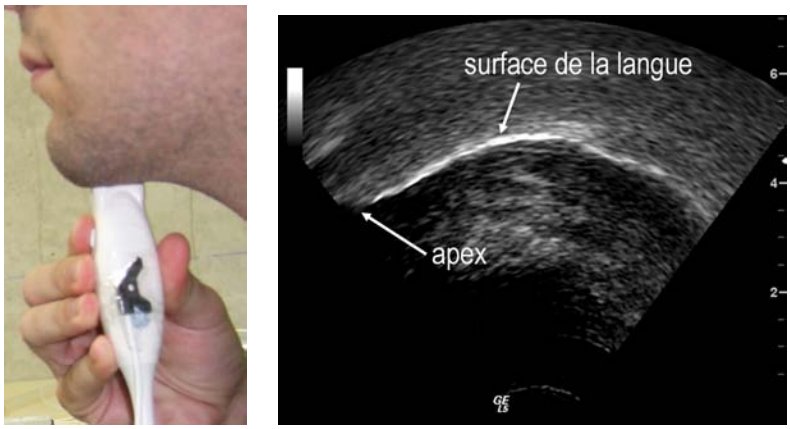
Les échos reçus sont amplifiés et traités par le système échographique, qui les convertit en un signal vidéo en niveau de gris. Le noir représente un écho dont l'amplitude est minimale et le blanc représente l'écho d'amplitude maximale. Il se dégage donc des images échographiques des zones blanches qui caractérisent des changements abrupts de milieu, ce qui correspond, en général, à la surface des organes.

Aucun tissu humain traversé n'est homogène : l'écho qui traverse le tissu est donc constamment perturbé par ces inhomogénéités, ce qui résulte dans l'image échographique par la formation de bruit, appelé speckle[3].

Maureen Stone, pionnière de l'utilisation de l'échographie pour l'étude des mouvements de la langue depuis les années quatre-vingt, détaille un guide d'utilisation de l'échographie pour l'acquisition de données sur la langue [Sto05]. L'utilisation la plus courante consiste à positionner

[3]On gardera la formulation anglaise, car il n'existe pas de consensus sur une traduction française : on parle parfois de scintillement, de granularité ou même de chatoiement pour désigner le speckle

la sonde échographique sous le menton (cf figure 1.7.a) pour obtenir des images de la surface de la langue dans le plan médiosagittal (cf figure 1.7.b).



(a) (b)

FIG. 1.7 – Utilisation de l'échographie pour imager la surface de la langue. (a) Sonde ultrason sous le menton. (b) Image résultante : la zone blanche présente la zone d'interface entre la surface de la langue et l'air. Dans la suite de ce manuscrit, on gardera la même orientation pour le images US : apex à gauche et arrière de la langue à droite de l'image.

Les avantages sont nombreux : tout d'abord, il n'y a pas de risque pour le sujet. Ensuite, les fréquences d'acquisition sont élevées, pouvant atteindre les 200 Hz pour les systèmes les plus récents. Les acquisitions sont non invasives, et la sonde placée sous le menton ne perturbe que très modérément la phonation du locuteur. Son coût est aussi relativement raisonnable. De plus, une fois réglé, un échographe peut être utilisé très rapidement et permet d'obtenir des images en temps réel.

En revanche, cette modalité comporte un certain nombre de limitations : de par la nature même de la formation de l'image échographique, les zones d'interface sont très souvent couplées à du speckle et sont représentées par une zone d'épaisseur variable suivant leurs propriétés acoustiques et leur orientation. De plus, la qualité des images dépend de l'échogénicité des sujets. Les ultrasons ne peuvent ni traverser l'air ni les os. Pour des acquisitions de la surface de la langue par exemple, l'apex est très souvent invisible à cause de l'air présent entre le plancher buccal et la langue, mais également à cause de l'occultation du faisceau par la mandibule. L'os hyoïde forme aussi un cône d'ombre à la racine de la langue. Enfin, la qualité dépend aussi de l'orientation de l'interface imagée par rapport à l'orientation du faisceau ultrasonore : une interface orthogonale à la direction du faisceau sera visible alors qu'elle disparaît presque complètement si elle est alignée selon cette direction.

Pour utiliser un tel système, il s'agit de trouver un bon compromis entre zone de couverture, profondeur, qualité de l'image et fréquence d'acquisition. Les réglages utilisés dans le cadre de notre système sont détaillés dans le chapitre 3. Dans la suite de manuscrit et par abus de langage, on parlera indifféremment d'images échographiques, d'images ultrasons ou encore d'images US.

1.3.5 IRM

1.3.5.1 IRM statiques

L'Imagerie à Résonnance Magnétique nucléaire (IRM), mise au point au cours des années 70 par Paul Lauterbur et Peter Mansfield, est, à ce jour, la plus couramment utilisée pour l'étude de la forme du conduit vocal dans une position statique [RHI$^+$86, BBR$^+$02, Eng00] : elle permet en effet d'obtenir des informations en 3D, sur la totalité du conduit vocal, et avec une bonne résolution spatiale (entre 0.5 et 1 mm/pixel dans un plan et entre 1 mm et 5 mm dans la troisième dimension).

Le principe de fonctionnement est le suivant : la tête du sujet est soumise un champ magnétique élevé. Ce dernier oriente le moment magnétique de l'unique proton du noyau des atomes d'hydrogène constituant la matière et perturbe cette orientation forcée par un gradient d'impulsions de champ magnétique dans un plan de coupe donné. Le signal d'énergie généré par le retour des moments magnétiques à leur état d'équilibre forcé est le signal de résonance magnétique nucléaire. Il mesure donc indirectement la densité des protons des tissus dans la coupe considérée. L'image de niveaux de gris résultante (cf figure 1.8) forme l'image l'IRM, et permet ainsi de différencier des tissus ou structures n'ayant pas la même densité en atomes d'hydrogène. Des volumes 3D sont reconstruits en empilant des images successives. Il existe également des acquisitions nativement 3D, mais ces dernières requièrent un temps d'acquisition encore trop long pour être envisagées dans notre contexte.

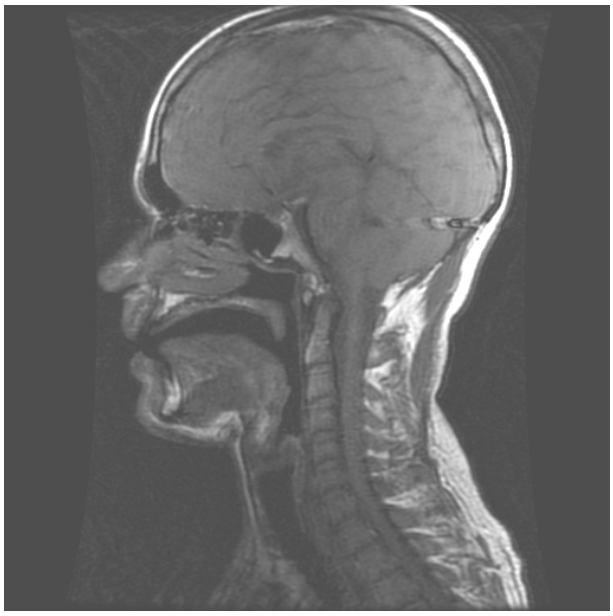


Fig. 1.8 – Image IRM : coupe médiosagittale d'un */a/*.

Contrairement aux images échographiques, l'imagerie IRM permet de couvrir la globalité du conduit vocal et en 3D dimensions. Elle connaît cependant quelques sérieux désavantages : le protocole d'enregistrement est relativement long, souvent plusieurs dizaines de secondes pour une vingtaine d'images espacées de 3 mm par exemple, permettant donc d'imager seulement des positions statiques pouvant être tenues dans le temps. Ensuite, les structures comme les dents ou les os, qui sont des structures cristallines qui ne résonnent pas, n'apparaissent pas sur les images en se confondant avec l'air. De plus, le sujet est en position couchée durant les acquisitions ce

qui peut affecter la phonation. Enfin, son utilisation reste réservée à des spécialistes dans des structures spécialisées, et il n'est parfois pas évident d'y avoir accès dans un cadre de recherche.

Nous reviendrons plus en détail dans le chapitre 5 sur les caractéristiques des protocoles IRM (sens des coupes, temps d'acquisition...) et des images obtenues (taille, nombre d'images, résolution...).

1.3.5.2 IRM dynamiques

De nombreux progrès ont été réalisés ces dernières années dans les temps d'acquisition des images IRM, et on commence à voir apparaître des prototypes de systèmes permettant d'acquérir plusieurs images par seconde. La principale caractéristique des IRM dynamiques est que la reconstruction des images IRM (passage du domaine de Fourier au domaine spatial) est effectuée a posteriori, ce qui diminue considérablement les temps d'acquisition. Les premières utilisations dans le cadre de la parole datent de 1999 avec [SMJC99] : 25 images du plan médiosagittal espacées de 21 ms ont été acquises pendant la prononciation de */pasi/*. Ces images avaient une taille de 128×128 pixels pour une résolution de 1.875 mm/pixel. Deux autres séries, à 11 mm à droite et à 11 mm à gauche du plan médiosagittal, ont été effectuées de la même manière en répétant le son, ce qui fait dire aux auteurs qu'ils effectuent une acquisition 3D, alors que ce n'est vraiment le cas.

Dans de récents travaux [KSN09], les auteurs proposent un système IRM permettant, pour une même taille d'image, d'obtenir une résolution de 1.33 mm/pixel pour des acquisitions d'une dizaine de secondes. Les résultats sont prometteurs, mais pas encore exploitables en tant que tel dans le cadre d'études sur la position précise des articulateurs à un instant donné. De plus, seule une coupe peut être acquise à la fois. Enfin, le matériel utilisé est encore à l'état de prototype et très onéreux.

1.3.6 Récapitulatif

Pour résumer, les principales caractéristiques des méthodes d'acquisition sont présentées dans le tableau 1.1. Il apparaît très clairement qu'aucune de ces modalités ne permet des acquisitions idéales du conduit vocal. Les données EMA permettent d'avoir des points 3D à de très bonnes fréquences, mais sur un nombre très limité de points. Les données IRM couvrent la globalité du conduit vocal en trois dimensions, mais ne peuvent être acquises que pour des sons pouvant être tenus dans le temps, aucun aspect dynamique ne peut être pris en compte avec cette modalité. En revanche, l'échographe permet d'avoir une information 2D dynamique et sur une partie continue de la langue, sans pour autant pouvoir l'imager dans son ensemble. Les données à base de cinéradiographie sont dangereuses pour le sujet et ne peuvent plus être obtenues dans un contexte de recherche.

Aucune n'étant satisfaisante seule, nous allons fusionner plusieurs de ces modalités afin de bénéficier d'informations complémentaires sur les formes et les mouvements des articulateurs. Ces informations constitueront une base de données articulatoires permettant la modélisation (et son évaluation) du comportement du conduit vocal lors de la phonation.

	EMA	**IRM**	**Échographe**	**Rayons X**	**Micro-faisceaux**
Fréquence d'acquisition[1]	200 Hz	-	30-200 Hz	50 Hz	40-160 Hz
Type de l'information	point 3D	volume 3D	coupe 2D	2D	point 2D
Conduit vocal complet	non	oui[2]	non	oui	non
Racine de la langue	non	oui	non	oui	non
Dos de la langue	point	totalité	surface	totalité	capteurs
Apex	point	oui	non	oui	non
Vélum	oui	oui	non	oui	oui
Nocif	non	non	non	oui	oui

TAB. 1.1 – Comparaison entre les différentes techniques d'acquisition. Notes : [1] 60 Hz est le seuil inférieur pour observer les mouvements articulatoires. 1000 Hz semblent nécessaires pour observer tous les mouvements [MBE+06]. [2]La position allongée pendant les acquisitions IRM est susceptible d'avoir un effet sur l'articulation.

Chapitre 2

Acquisition de données articulatoires multimodales : état de l'art et objectifs de la thèse

Que ce soit pour explorer les stratégies d'articulation, pour construire ou évaluer un modèle articulatoire, les acquisitions et les traitements des données constituent le fondement des bases de données de formes des articulateurs du conduit vocal dans l'espace et dans le temps. Ce chapitre présente les différents systèmes d'acquisition multimodaux existants, propose ensuite leur analyse critique pour enfin décrire les objectifs de ce travail de thèse.

2.1 Étude de l'existant

2.1.1 Introduction

Nous avons vu dans le chapitre précédent qu'une seule modalité ne suffisait pas pour obtenir la totalité des formes possibles des articulateurs. L'idée est donc de regrouper plusieurs méthodes d'acquisition pour constituer une base multimodale de données articulatoires. Dans le but de mettre en place notre propre système d'acquisition, nous détaillerons les différentes bases de données existantes telles qu'elles sont décrites dans la littérature par leurs auteurs. Nous nous intéresserons aux systèmes acquérant des données multimodales dynamiques, c'est-à-dire à ceux qui permettent à un instant t d'obtenir des données provenant de plusieurs modalités. Pour cela, nous détaillerons les points particuliers suivants :

Objectifs Nous évoquerons les principales motivations pointées par les auteurs pour leurs acquisitions.

Volume et type des données Nous verrons le volume des données enregistrées par chacun des systèmes, les types de corpus utilisés (phrases, VCV, phonèmes simples...), et le nombre de locuteurs ayant été testés.

Modalités Nous détaillerons ensuite les différentes modalités choisies et donnerons leurs principales caractéristiques.

Synchronisation Au cours d'une séance d'acquisition, lorsque des données dynamiques sont obtenues par différents systèmes, il est nécessaire de les mettre temporellement en correspondance. On utilisera pour cette opération dans la suite ce manuscrit le terme de *synchronisation* des données. La synchronisation implique de définir une référence de temps commune aux modalités. Elle permet ensuite d'étiqueter la donnée acquise par chaque modalité dans cette référence de temps pour un instant t, dit temps d'acquisition. Elle fait apparaître deux notions sous-jacentes : celle de délai, qui correspond à la différence de temps entre t_0, l'origine du temps commun, et le début de l'acquisition de la modalité (cf figure 2.1), et celle de fréquence, qui mesure en Hertz (Hz) le nombre d'acquisitions faites par la modalité en une seconde. Pour une modalité, la durée séparant deux temps d'acquisition successifs, c'est-à-dire le pas d'acquisition, étant en général constante, la fréquence d'acquisition est l'inverse de ce pas. Synchroniser les données revient à maîtriser les valeurs de délai entre les modalités, ainsi que chacune de leurs fréquences d'acquisition. Nous verrons notamment dans la partie 2.1.3 de ce chapitre les conséquences que peuvent avoir des mesures erronées de délai et de fréquence. Nous détaillerons donc pour chaque système d'acquisition la stratégie de synchronisation choisie par les auteurs.

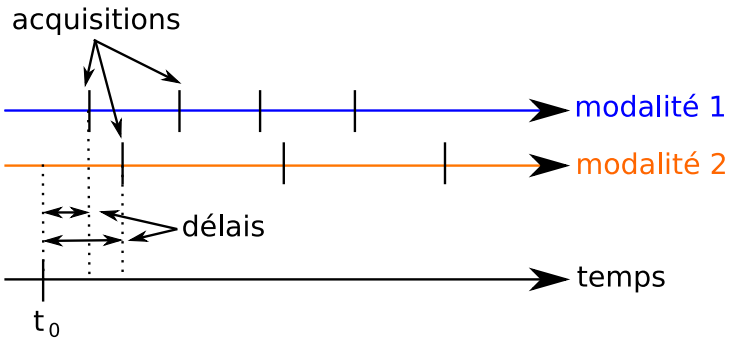


FIG. 2.1 – Principe de la synchronisation : le délai correspond à la différence entre l'origine des temps t_0 et le début de l'acquisition de la modalité. La fréquence d'une modalité est le nombre d'acquisitions effectuées en une seconde.

Recalage De la même façon que la synchronisation met temporellement en correspondance des données multimodales, le *recalage* les lie spatialement, afin de disposer d'un repère spatial commun à toutes les données. Chaque modalité étant acquise dans son propre repère spatial, le recalage consiste à calculer la transformation entre les repères de chacune des modalités. Nous évoquerons donc, si cela a été mentionné, comment le recalage a été effectué pour les systèmes étudiés.

2.1.2 Bases de données de la littérature

2.1.2.1 MOCHA, Edinburgh University

MOCHA (Multi-CHannel Articulatory) [WH00] est une base de données publique recueillie à l'Université Queen Margaret College d'Edimbourgh, en Écosse, développée pour la reconnaissance automatique de la parole et l'étude de la coarticulation. Elle contient des enregistrements de deux locuteurs, un homme et une femme parlant l'anglais britannique, pour 460 phrases d'une dizaine de mots, ce qui représente environ 40 minutes de parole.

Les données enregistrées comprennent :
- de l'audio (16 kHz),
- des données EMA provenant d'un articulographe AG200 (500 Hz). Huit capteurs sont collés dans le plan médiosagittal : 2 pour les lèvres, 2 sur la mâchoire, 3 sur la langue et 1 sur le vélum,
- des données provenant d'un laryngographe : il s'agit d'un système fournissant la valeur de la fréquence fondammentale des cordes vocales (à 16 kHz),
- des données provenant d'un électropalatographe (EPG), palais artificiel moulé donnant l'information binaire (à 200 Hz) de contact entre l'un de ses 62 points et la langue,
- des données d'une caméra vidéo (30 Hz) filmant de face les lèvres des locuteurs et enregistrées sur cassette SVHS.

Selon les auteurs, les données sont synchronisées lors des acquisitions grâce à un trigger. La méthode n'est pas détaillée, et aucune mesure sur les délais de synchronisation et des fréquences d'acquisition n'est spécifiée. À l'usage, il apparaît d'après les utilisateurs de la communauté parole, qu'il y a des décalages temporels entre les acquisitions des différentes modalités. Par exemple, Qin dans [QCP07] teste empiriquement plusieurs valeurs de délai entre les données électromagnétiques et les données audio, pour finalement conclure qu'un écart de 15 ms entre les deux modalités donne une erreur moindre sur sa méthode d'inversion acoustique articulatoire.

Il n'y a pas de recalage de données proposé pour ce système, elles sont géométriquement utilisées de manière indépendante par la communauté parole.

2.1.2.2 Qualisys-Movetrack, KTH Stockholm

La base de données privée Qualisys-Movetrack [BEG03] (KTH, Suède) a été enregistrée sur une locutrice suédoise ayant prononcé 270 phrases de quatre ou cinq mots, 138 VCV et VCCV, et 41 CVC asymétriques. Les auteurs précisent qu'il s'agit d'un système pouvant être utilisé pour la mise en place de têtes parlantes animées, sans pour autant détailler davantage les objectifs de leurs travaux.

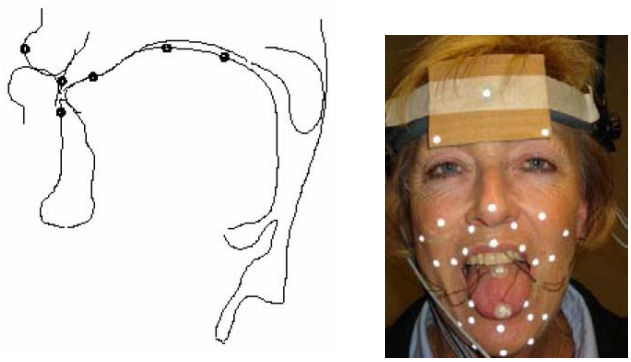
Elle contient des enregistrements simultanés :
- d'audio (16 kHz),
- des données EMA provenant d'un articulographe Movetrack [Bra85] fonctionnant à 200 Hz. Six capteurs électromagnétiques ont été utilisés (cf figure 2.2.a) dans le plan médiosagittal : 3 pour la langue, 2 pour la mâchoire inférieure et supérieure, et 1 capteur sur la lèvre supérieure,
- des données vidéo (60 Hz) provenant de quatre caméras Qualisys[4]. Grâce à 28 marqueurs (points blancs de la figure 2.2.b) dessinés sur le visage de la locutrice, ce système optique permet la reconstruction tridimensionnelle de ces marqueurs.

Le capteur sur la lèvre supérieure, visible par les caméras vidéo, est utilisé pour la synchronisation EMA/vidéo. Elle est effectuée a posteriori, en utilisant l'information redondante visible dans les deux modalités : les mouvements du capteur sur la lèvre de la locutrice visible à la fois dans les données EMA et vidéo sont manuellement mis en correspondance. Au préalable, les données des capteurs EMA et vidéo sont sous-échantillonnées à la fréquence d'acquisition la plus

[4] `http://www.qualisys.se`

faible (60 Hz) pour avoir le même nombre de données dans chaque modalité.

Pour le recalage des données EMA et des données vidéo, les auteurs estiment la position du plan médiosagittal à partir des trois marqueurs vidéo dessinés sur le front (cf figure 2.2.b) et calculent la transformation de ce plan avec celui défini par les capteurs EMA.



(a) (b)

FIG. 2.2 – (a) Placement des capteurs sur les articulateurs avec Movetrack. (b) Marqueurs peints sur le visage pour Qualisys. Extraits de [BEG03].

2.1.2.3 Vocal Tract Visualization (VTV) Laboratory, Maryland University

Il s'agit d'un système d'acquisition proposé par Maureen Stone (Université du Maryland, États-Unis) dans [Sto05]. Les données acquises n'ont pas été rendues publiques, et il n'y a pas d'objectif spécifique présenté dans l'article. Le système est utilisé pour l'étude des formes de la langue lors de l'articulation comme dans [SSB$^+$07] où les différences de position de la surface de la langue entre la position couchée et la position assise sont étudiées dans les images échographiques. Hueber [HCD$^+$07] utilise aussi ce système pour le projet Ouisper qui cherche à synthétiser un phonème à partir d'une forme de la langue et d'une position des lèvres. Pour ces recherches, un corpus de 720 phrases, prononcé par un locuteur masculin parlant l'anglais américain, a été acquis, ce qui représente 43 minutes de parole.

Le système est composé du système de contention HATS [SD95] permettant l'acquisition d'images échographiques de la surface de la langue dans le plan médiosagittal : la tête du locuteur est immobilisée dans un casque et la sonde est fixée sur un support à ressort (cf figure 2.3), de façon à obtenir des images dans le plan médiosagittal tout en ayant une légère latitude dans les mouvements de la sonde grâce au ressort. Ce système de contention de la tête est couplé à :

- un EPG (électropalatographe) pour enregistrer les points de contact de la langue et du palais (à 200 Hz),
- un microphone pour capturer le signal audio (à 16 kHz),
- une caméra vidéo filmant le visage du locuteur de profil,
- une table de mixage vidéo, qui mixe à la fois le signal sonore et les signaux vidéo provenant de l'échographe et des caméras,
- un magnétoscope enregistrant le signal audiovidéo issu de la table de mixage (à 30 Hz).

Le signal vidéo de l'échographe provient de sa sortie analogique, délivrant un flux vidéo au format NTSC cadencé à 30 Hz. À la sortie de la table de mixage, le signal vidéo enregistré sur le magnétoscope regroupe donc le signal de l'échographe, de la caméra vidéo et du signal audio, supposant que les trois signaux sont synchronisés. Cependant, Stone [Sto05] avoue avoir constaté sur les enregistrements des décalages d'une seconde entre les signaux audio et vidéo, soient 30 images vidéo. Elle incrimine la table de mixage qui ne semble pas être suffisamment précise pour effectuer une meilleure synchronisation.

Il n'y a pas de recalage des données. Un des points précisés est de tenter de reprendre approximativement les mêmes positions pour le locuteur entre des sessions d'acquisition différentes, afin de disposer de données capturées dans des conditions similaires. De ce point de vue, Hueber [HCDS08] propose d'afficher le visage du sujet tel qu'il apparaît dans les sessions précédentes et de lui demander de superposer au mieux l'image de son visage sur cette image de référence. Même si cette méthode permet d'obtenir des données à peu près cohérentes entre elles, elle ne permet pas de calculer la transformation spatiale entre les données échographiques et les données vidéo.

2.1.2.4 Haskins Optically Corrected Ultrasound System (HOCUS), Yale University

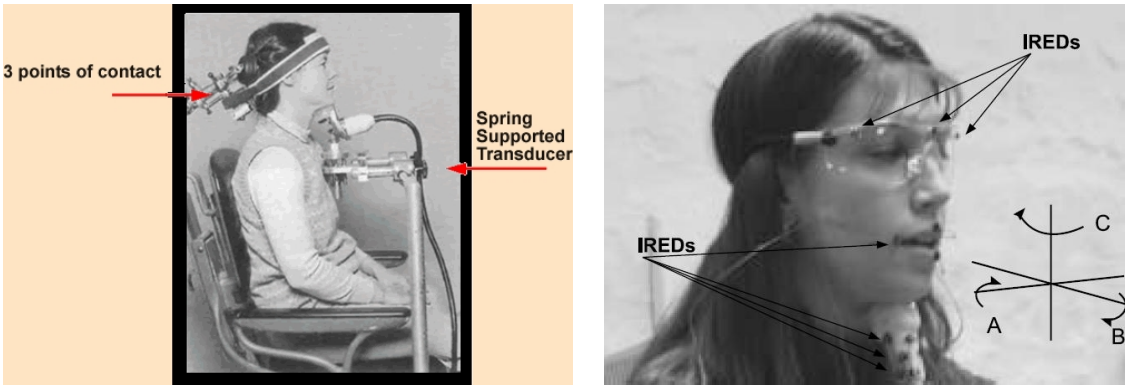
HOCUS est un système d'acquisition développé par le laboratoire Haskins de l'Université de Yale aux États-Unis. Les données n'ont pas été rendues publiques, mais le système est détaillé dans [WTO+05]. Il est présenté dans le cadre d'étude de formes de 11 voyelles anglaises prononcées par une locutrice.

Il s'agit de la combinaison :
- d'un échographe (57 Hz),
- de marqueurs optiques (système Optotrack, NDI fonctionnant à 200 Hz) fixés sur la sonde échographique et sur la tête du locuteur,
- et d'un microphone pour enregistrer le son (16 kHz).

Les marqueurs optiques sont fixés sur deux supports : sur la sonde échographique et sur la tête du locuteur. La position dans l'espace de ces deux éléments est donc connue grâce au suivi optique. Cela permet de les laisser libres lors des acquisitions, sans qu'il soit nécessaire d'utiliser un système de contention comme avec HATS.

Comme pour les données acquises avec le VTV, les images provenant de l'échographe sont enregistrées sur un magnétoscope, en utilisant la sortie vidéo analogique (30 Hz) de l'échographe. Il est spécifié que le signal audio a été manuellement synchronisé avec les images US, sans plus détailler ce point.

Les auteurs sélectionnent les images ultrasons pour lesquelles les mouvements de la sonde et de la tête du locuteur capturés par le système optique sont inférieurs à un seuil (seuillages sur l'angle et la translation) par rapport à la première acquisition, considérée comme la position de référence. Le système optique est donc utilisé comme une modalité permettant de discriminer les images US pour lesquelles les déplacements de la sonde et de la tête sont jugés trop importants par rapport à la position de référence. Ensuite, pour les images considérées comme valables, les auteurs extraient manuellement la surface de la langue de ces images pour leurs études.



(a) (b)

FIG. 2.3 – (a) Système HATS, extrait de `http://speech.umaryland.edu/ahats.html`. (b) Système HOCUS, extrait de [WTO+05].

2.1.2.5 GIPSA-lab (ex ICP), Grenoble

Le laboratoire grenoblois de l'Image, de la Parole, du Signal et de l'Automatique (GIPSA-lab) s'est lui aussi intéressé à l'acquisition de données articulatoires multimodales. Leur stratégie est différente des systèmes présentés précédemment puisqu'est utilisée l'imagerie IRM qui ne peut être acquise qu'indépendamment de toute autre modalité. Dans [BEB+07], Badin décrit des acquisitions de données pour la construction de modèles articulatoires intégrés dans une tête parlante virtuelle. Les données, non publiques, ont toutes été acquises sur un locuteur masculin : le corpus est composé de 46 phonèmes français pouvant tous être artificiellement tenus durant la phonation, les 14 phonèmes voyelles, et les consonnes dans les contextes de coarticulation */a/, /i/, /u/*.

Le système d'acquisition comprend des données (les auteurs ne précisent pas les fréquences d'acquisitions des matériels utilisés) :

- IRM acquises en 53 coupes sagittales sur la totalité du conduit vocal,
- vidéo acquises à partir de caméras de stéréovision pour filmer le visage du locuteur et permettre une reconstruction tridimensionnelle du visage sur lequel ont été peints des marqueurs (notamment les lèvres),
- audio,
- EMA venant d'un articulographe, avec six capteurs utilisés dans le plan médiosagittal : 1 sur la mâchoire, 2 sur les lèvres et 3 sur la langue.

Un premier corpus de données statiques (sons maintenus) a été acquis séparément avec chacune des modalités. Ensuite, un second corpus (non précisé) est acquis avec seulement l'articulographe pour des données dynamiques (sons parlés). Il ne s'agit ici donc pas d'un système d'acquisition multimodal tel que nous l'avons défini dans la partie 2.1.1 car une seule modalité est acquise à un instant t. Il est cependant intéressant de le décrire, car il est le seul proposant d'utiliser à la fois des données dynamiques et des données statiques. Puisque chaque modalité est utilisée indépendamment, il n'y a pas besoin de synchronisation.

Chaque information articulatoire du premier corpus est manuellement extraite des modalités

(positions de la langue et du palais dans les IRM par exemple), puis trois modèles articulatoires sont construits à partir des données IRM et vidéos grâce à des analyses en composantes principales guidées : un modèle articulatoire mâchoire-lèvres-visage, un modèle articulatoire mâchoire-langue, et un modèle de voile du palais.
Pour être utilisés dans la tête parlante, les auteurs proposent de contrôler ces trois modèles articulatoires par un nombre réduit de paramètres. Pour cela, les données dynamiques EMA du second corpus sont utilisées : les points correspondant aux positions des capteurs EMA dans les modèles articulatoires sont manuellement désignés, de façon à attacher à chaque modèle un ou plusieurs points EMA. Ensuite, les déformations de chaque modèle sont calculées par optimisation pour que les modèles collent au mieux à la position des points EMA attachés au modèle. L'une des difficultés d'une telle méthode est de choisir les bons points EMA, c'est-à-dire ceux qui caractérisent au mieux les déformations du modèle articulatoire. Selon les auteurs, les six capteurs électromagnétiques suffisent à contrôler les trois modèles.
Les auteurs ne précisent pas leur méthode pour spatialement recaler les trois modèles articulatoires, laissant supposer que ceci est effectué manuellement.

Il s'agit donc ici principalement d'acquisitions de données en position statique, et pouvant être acquises indépendamment l'une de l'autre. Les données dynamiques ne proviennent que d'une seule modalité, l'articulographe, fournissant des paramètres de contrôle des modèles articulatoires. Ce travail est l'un des seuls à notre connaissance qui s'intéresse au contrôle de modèles articulatoires à partir de quelques données dynamiques.

2.1.3 Discussion

De cette étude des méthodes d'acquisition multimodale existantes se dégagent plusieurs constatations :

- afin de mesurer et caractériser le plus grand nombre de formes du conduit vocal possible, tous les protocoles expérimentaux proposés permettent des acquisitions de données articulatoires sur plusieurs minutes d'acquisition. Certains comme MOCHA ont été testés pour plusieurs locuteurs ;
- afin de récupérer une information sur la forme d'un articulateur à un instant donné, tous les systèmes proposent l'acquisition de données dynamiques, que ce soit avec les images échographiques, des données électromagnétiques, et/ou d'images vidéo des lèvres ;
- pour les acquisitions électromagnétiques ou échographiques, les données sont recueillies dans le plan médiosagittal qui correspond au plan dans lequel les modèles articulatoires 2D sont établis ;
- certains systèmes comme HOCUS ou celui du KTH laissent libre la tête du locuteur au cours des acquisitions dynamiques, lui permettant d'effectuer des mouvements articulatoires naturels. De plus, le gain substantiel de confort pour le locuteur permet d'effectuer théoriquement des sessions d'enregistrement plus longues qu'avec des systèmes de contention comme HATS, très souvent inconfortables ;
- le GIPSA-lab utilise des acquisitions statiques IRM, seule modalité permettant de couvrir l'ensemble des articulateurs du conduit vocal en trois dimensions.

Cependant, cette analyse de l'existant révèle de nombreuses lacunes. Beaucoup d'hypothèses a priori ont été posées dans la littérature, sans avoir été vérifiées ou remises en cause. Parmi ces points peu ou mal traités, on dénombre les aspects décrits dans les sections suivantes.

La synchronisation La synchronisation des données est souvent rapidement décrite. C'est pourtant un problème crucial lorsque plusieurs modalités dynamiques sont acquises lors d'une même session. Pour se donner une idée de l'influence de la valeur du délai, si on prend l'exemple du système HOCUS avec des données EMA acquises à 500 Hz et un délai de 15 ms, cela engendrera une erreur de huit acquisitions EMA. De la même façon, une erreur d'estimation de 1 Hz dans la fréquence engendrera une erreur de dix acquisitions EMA après dix secondes d'acquisition, soient soixante acquisitions EMA après seulement une minute d'acquisition.
Les auteurs du système du KTH proposent une synchronisation a posteriori des données vidéo et électromagnétiques, car ils bénéficient d'information redondante (capteur électromagnétique visible) entre les deux modalités pour les synchroniser. Avec un événement facilement identifiable dans les deux modalités (mouvement brusque par exemple), cette information redondante permet de synchroniser les modalités lors de cet événement. Cependant, les auteurs ne spécifient pas comment l'information redondante a été exploitée et s'ils ont rencontré des difficultés pour la traiter. Nous n'avons aucune idée de la précision de la synchronisation de cette méthode.
De plus, une information redondante n'est pas toujours présente pour d'autres systèmes utilisant des modalités différentes : par exemple, entre des caméras vidéo filmant les lèvres et des images ultrasons de la surface langue, il n'y a pas d'information commune visible sur ces deux modalités.
Les systèmes du VTV et de HOCUS utilisent donc un magnétoscope pour enregistrer en temps réel les différents signaux audio et vidéo, et supposent qu'ils sont synchronisés. Cependant, malgré la confiance que les auteurs ont sur le matériel utilisé, ils constatent tout de même des délais de quelques secondes. Ce délai peut provenir du temps nécessaire pour capturer le signal analogique à la sortie de l'échographe, du temps nécessaire entre l'acquisition US et son affichage... Il est donc absolument nécessaire de contrôler les délais et les fréquences d'acquisition, afin de s'assurer de la bonne synchronisation des données.
Enfin, l'utilisation d'un magnétoscope pour l'acquisition des images ultrasons a pour conséquence de générer un signal vidéo à la fréquence de 30 Hz, car la sortie vidéo analogique de l'échographe est utilisée. Il est dommage pour un tel système d'acquisition, dont l'un des principaux avantages est de pouvoir obtenir des images à des fréquences importantes (souvent plus de 50 Hz), de ne finalement l'utiliser que pour obtenir un signal vidéo sous-échantillonné à 30 Hz.

Le recalage On remarque aussi que les modalités ne sont pas spatialement recalées entre elles. Les informations spatiales sont utilisées de façon différente suivant les systèmes :

- pour HOCUS, l'une des modalités (les capteurs optiques) est utilisée pour fixer un seuil d'amplitude de déplacement au-delà duquel l'autre modalité (les images ultrasons) ne sera pas utilisée. La possibilité de savoir à chaque instant où se situe le plan ultrason en trois dimensions par rapport aux marqueurs optiques n'est pas exploitée, et seuls les déplacements de la tête et de la sonde US sont considérés ;
- le système du VTV et les travaux de Hueber [HCDS08] proposent une méthode pour garder une cohérence spatiale des données entre plusieurs sessions d'enregistrement, mais les données des lèvres ne sont pas recalées avec les données échographiques ;
- le système du KTH est le seul proposant une solution pour le recalage spatial des données. Cependant, peu de détails sont donnés à son propos, et on est en droit de se demander la précision de la méthode. En effet, les plans médiosagittaux sont estimés manuellement, que ce soit en peignant les marqueurs sur le front pour les données vidéo ou en collant les capteurs sur la langue pour les données EMA. Rien n'assure que les capteurs soient tous collés dans un même plan. De plus, les auteurs ne détaillent pas le procédé de calcul utilisé.

Les traitements manuels Il apparaît de cette étude de l'existant que beaucoup d'étapes sont effectuées manuellement. Les auteurs ne détaillent que très sommairement les étapes de synchronisation et de recalage, et les interventions d'un opérateur humain sont nombreuses. De même, certaines opérations, comme l'extraction de la surface de la langue dans les images échographiques du système HOCUS, ou l'attache du point EMA aux modèles articulatoires du GIPSA-lab, sont totalement manuelles. Outre le fait qu'elles font appel à la subjectivité humaine, qui peut être source d'erreurs, ou tout du moins de résultats différents selon la personne qui les effectue, elles ont pour principales conséquences d'empêcher l'exploitation de corpus volumineux en taille, et limitent les études multilocuteurs.

La répétabilité et la variabilité Dans le cadre d'acquisitions statiques comme celles effectuées au GIPSA-lab, les auteurs supposent que les positions des articulateurs ne varient pas pour un même son. Stone dans [ES05] fait la même hypothèse pour reconstruire un modèle de langue tridimensionnel à partir d'images échographiques. Cependant, on ne sait pas si cette hypothèse est vraie. Y a t-il des différences de position des articulateurs entre deux acquisitions d'un même son ? Si oui, quelle est cette influence dans les applications proposées ? Ces questions sont souvent passées sous silence dans la littérature.

L'évaluation Enfin, il manque cruellement d'évaluation des méthodes présentées. Il est, certes, très difficile de proposer une évaluation des données articulatoires, car on ne dispose pas de réalité terrain avec laquelle comparer les données acquises.
Pour chaque système d'acquisition utilisé, les auteurs font naturellement confiance aux mesures de précision données par les constructeurs. Par exemple, une des seules études existantes sur la précision dynamique des données de l'articulographe AG500 n'a été réalisée que fin 2008 par Kroos [Kro08]. Il a comparé des mesures électromagnétiques à des vitesses différentes en se référant à un système de suivi optique plus précis. Les résultats obtenus ont montré que si la précision des capteurs en statique est inférieure à 1 mm, elle peut être supérieure à 4 mm en dynamique. Pourtant, l'AG500 est donné par les constructeurs avec une précision de 0.5 mm, et la grande majorité des études effectuées avec cet articulographe font référence à cette valeur de précision, visiblement fausse ! Il est donc difficile d'avoir une idée de l'ordre de grandeur de la précision attendue sur des données articulatoires acquises.

De plus dans le cadre d'un système multimodal, l'évaluation de la précision des informations fusionnées manque également. Par exemple, pour le système du KTH qui est l'un des seuls à effectuer un recalage, nous ne sommes pas en mesure de dire si le recalage est précis ou pas, et s'il nécessite d'être amélioré ou pas.

Pour des systèmes comme HOCUS où la surface de la langue a été extraite des images US, ou pour celui du GIPSA-lab où les contours du palais ont été dessinés sur les images, nous n'avons pas non plus idée de la précision de cette extraction. Dans le cas où elle a été manuelle, elle dépend de l'opérateur, mais aussi de la résolution de l'image. Dans ce cas, quelle précision est souhaitée ?

Le GIPSA-lab a utilisé dans [BEB+07] les données articulatoires acquises pour la mise en place d'une tête parlante, pour ensuite évaluer l'apport de la vision de la langue à l'intelligibilité de la parole. Badin présente ce travail comme une évaluation préliminaire du modèle articulatoire mis en place, puisqu'il est utilisé dans un contexte applicatif avec succès. L'objectif de ce

travail étant une application purement visuelle, les auteurs ne s'attardent pas sur la précision de leur système. Il n'y pas de résultat quantitatif présenté, et la conclusion « l'analyse des résultats montre un certain effet d'apprentissage implicite de la lecture linguale » prouve qu'il reste encore de nombreux tests à effectuer pour évaluer leurs travaux.

Il n'y a donc pas d'évaluation objective de la précision de chacune des modalités utilisées, ni de la précision globale des systèmes d'acquisition. De plus, de nombreux travaux utilisent les données acquises, sans pour autant vérifier leur validité dans le cadre de l'application visée. Le problème de l'évaluation des données articulatoires reste donc ouvert.

Tableau récapitulatif Les différents points mis en avant précédemment sont synthétisés dans le tableau 2.1.

	MOCHA	**KTH**	**VTV**	**HOCUS**	**GIPSA-lab**
Modalités dynamiques	audio vidéo (face) EMA EPG laryngographe	audio vidéo (stéréo) EMA	audio vidéo (profil) US EPG	audio optotrak US	audio EMA
Modalités statiques	-	-	-	-	IRM vidéo (stéréo)
Synchronisation	trigger	manuelle (a posteriori)	mixage audio-vidéo	manuelle	-
Recalage	non	oui	non	non	manuel
Évaluation	non	non	non	non	visuelle

TAB. 2.1 – Résumé des principaux systèmes d'acquisition multimodaux pour les données articulatoires

Suite à cette étude de l'existant, nous présentons dans la section suivante les objectifs de notre travail.

2.2 Objectifs de la thèse

2.2.1 Corpus et multilocuteurs

Les corpus à acquérir sont fixés par la communauté parole et font émerger des données les mouvements représentatifs de l'espace articulatoire. Le système doit donc permettre d'acquérir des données sur les articulateurs en mouvement, comme des VCV pour étudier les transitions et le phénomène de coarticulation, des phrases pour étudier la dynamique et la vitesse des articulateurs afin de disposer du plus large éventail possible de formes du conduit. La première conséquence engendrée par ces conditions est que le système doit être capable d'acquérir un volume important de données (plusieurs heures d'enregistrement) pour avoir une base significative.

De plus, le système doit être utilisable sur plusieurs locuteurs, afin de pouvoir étudier les différentes stratégies articulatoires interlocuteurs. Par manque de données disponibles, le modèle

de Maeda [Mae79] (cf chapitre 1), est basé sur des données acquises (images rayons X) sur une seule locutrice : posséder des données sur plusieurs sujets permettrait donc d'établir de nouveaux modèles prenant en compte les différences interlocuteurs.

2.2.2 Données multimodales statiques et dynamiques

En regard de l'existant et des caractéristiques de chacune des modalités d'acquisition présentées dans le chapitre 1, nous avons choisi d'utiliser les modalités suivantes pour constituer notre base de données articulatoires :

- **données dynamiques** :
  - les échographies, pour visualiser la surface de la langue dans le plan médiosagittal ;
  - un système électromagnétique pour fixer un capteur sur l'apex pour visualiser sa position et ses mouvements, très souvent invisibles à l'échographie à cause de l'air entre la langue et le plancher de la cavité buccale et l'os du maxillaire inférieur ;
  - un système de stéréovision pour avoir la position et le mouvement des lèvres en trois dimensions ;
  - et enfin un système permettant l'enregistrement du signal audio ;
- **données statiques** : l'IRM sera utilisée pour obtenir des images représentant le conduit en trois dimensions, et ce, pour les phonèmes pour lesquels la phonation peut être maintenue dans le temps.

Cet ensemble forme un système d'acquisition de données statiques et dynamiques des articulateurs du conduit vocal. Ces données sont tridimensionnelles, sauf les échographies. En effet, technologiquement, il n'est actuellement pas possible d'obtenir des données dynamiques tridimensionnelles de toute la surface de la langue. Les IRM dynamiques sont encore à l'état de prototypes et fournissent des images bidimensionnelles de faible résolution (près de 2 mm/pixel pour des images de 128×128 pixels), les échographies tridimensionnelles n'ont pas une fréquence d'acquisition suffisante (une trentaine d'images par seconde pour des images de langue). Pour ces raisons, nous utilisons un échographe dans le plan médiosagittal pour obtenir des images de la dynamique de la surface de la langue.

2.2.3 Analyse des besoins

Avec un tel système pour l'acquisition de données multimodales statiques et dynamiques, et compte tenu de l'analyse de l'existant décrite dans le premier paragraphe, ce travail de thèse a mis l'accent sur les aspects décrits dans les sections suivantes.

2.2.3.1 Automatisation des acquisitions et des traitements

Un aspect fondamental d'un système multimodal est le caractère automatique des acquisitions et des traitements. En effet, puisqu'un des objectifs est d'acquérir des corpus de plusieurs dizaines de minutes, toutes les acquisitions et tous les traitements effectués sur les données se doivent d'être les plus automatiques possible, afin d'alléger la charge de travail manuel, et d'éviter les problèmes de traitements différents suivant l'opérateur qui les effectue.
Par exemple, pour une acquisition de 30 minutes de parole avec des images échographiques acquises à 50 Hz, on obtient près de 90000 images US ! Si l'on inclut aussi des images de stéréovision et des données EM, le volume des données acquises devient rapidement extrêmement conséquent et impossible à traiter si l'on ne dispose pas de méthodes automatiques.

De plus, pour être utilisées pour une application comme l'inversion acoustique articulatoire, certaines données doivent être extraites. C'est le cas des positions de la surface de la langue dans les images échographiques. Un traitement spécifique doit donc être envisagé pour extraire des images US les positions de la langue, afin de pouvoir traiter rapidement un important volume de données.

2.2.3.2 Traitements multimodaux

Le caractère multimodal d'un système d'acquisition de données articulatoires nécessite de prendre en compte les deux aspects fondamentaux que sont la synchronisation et le recalage, souvent négligés dans la littérature. Chaque modalité apportant une information différente, leur fusion dans un même repère spatial et temporel permet de savoir où chacune est située dans l'espace et dans le temps.

Ne disposant pas d'information redondante avec les systèmes d'acquisition choisis, la synchronisation doit être prise en compte lors des acquisitions. Recalées deux à deux et par transitivité, toutes les modalités pourront donc être synchronisées dans un repère temporel commun. Suite aux problèmes rencontrés par le VTV, on prendra soin de mettre en place des méthodes permettant de mesurer les délais d'acquisition entre chaque modalité, et leurs fréquences de fonctionnement. Le système doit aussi permettre d'obtenir des images échographiques qui soient échantillonnées à plus de 30 Hz afin de tirer avantage des importantes fréquences d'acquisition d'un échographe.

Les systèmes HOCUS et du KTH laissent libre la tête du locuteur lors des acquisitions dynamiques, sans qu'il y ait de moyen de contention. Cela nous semble important pour acquérir des données articulatoires les plus naturelles possible. Avec notre système, un capteur EM sera fixé sur la sonde US pour pouvoir repérer la position de la sonde dans l'espace EM. Des capteurs seront aussi fixés sur la tête du locuteur pour que ce dernier puisse bouger au cours des acquisitions. Ces capteurs nous permettront de connaître spatialement la position de la sonde échographique et de la tête dans le repère EM. Ils permettront donc le recalage de ces différentes modalités dans un repère intrinsèque lié à la tête du locuteur. Des études de variabilité et de répétabilité de la position de la surface de la langue lors de la phonation seront également réalisées. On mettra aussi en place une méthode permettant de recaler les données dynamiques US, EM et vidéos avec les données statiques IRM.

Le système permettra donc de fusionner une information de position d'un articulateur statique avec une information dynamique d'un autre articulateur ou plusieurs informations dynamiques ensemble. Une application potentielle, est d'afficher sur une image IRM, où la position du palais est connue, les contours de langue extraits des images US et les capteurs EM collés sur la langue.

2.2.3.3 Précision et évaluation

Les derniers points sur lesquels nous mettrons l'accent concernent la précision et l'évaluation des données.

Nous nous attacherons à fournir pour les modalités d'imagerie, les résolutions spatiales de chacune. Pour les données électromagnétiques, nous préciserons la fiabilité de chacune des mesures, à savoir la mesure de confiance que nous pouvons apporter aux données acquises. Comme Kroos [Kro08], nous analyserons les problèmes que peuvent engendrer des mouvements rapides

sur l'acquisition de données EM et les imprécisions qui en découlent.

Ensuite, nous évaluerons la fusion temporelle et spatiale des données, à savoir la synchronisation et le recalage. Ces deux étapes, qui constituent les deux briques d'un système multimodal, seront évaluées de la façon suivante :

- pour la synchronisation, des méthodes pratiques seront proposées pour estimer les délais d'acquisition entre chaque modalité, ainsi que leurs fréquences d'acquisition ;
- pour le recalage, nous évaluerons la précision de la chaîne permettant de passer d'une modalité et l'autre (cf figure 2.4). On cherchera à évaluer l'impact d'une erreur de quelques pixels dans l'image US recalée dans une image IRM. Ces deux modalités fournissant des images de résolutions différentes, la précision du recalage dépend de cette différence de résolution et du calcul de chacune des transformations impliquées dans la chaîne de recalage.

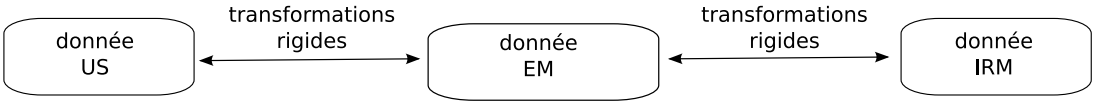


FIG. 2.4 – Principe de la chaîne de recalage : plusieurs transformations nécessitent d'être calculées pour le recalage entre une donnée échographique et une donnée IRM.

Ces études rigoureuses sur les précisions à la fois de chacune des modalités utilisées, et aussi sur la précision globale du système dans l'espace et dans le temps, nous permettront de quantifier les erreurs de notre système. Cela nous semble fondamental pour être en mesure de pouvoir l'améliorer dans le futur. Nous préciserons aussi les mesures d'erreurs et d'incertitudes lors de l'extraction de données articulatoires, comme pour la surface de la langue dans les images échographiques, ou encore de la position du palais dans les images IRM.

Enfin, nous évaluerons les données acquises et traitées. Nous utiliserons les données dans le contexte applicatif du projet européen ASPI dans lequel s'est effectué ce travail de thèse. Avant de pouvoir étudier les méthodes d'inversion acoustique articulatoire dont le but est de retrouver les formes des articulateurs à partir, idéalement, du signal acoustique seul (en pratique, on utilise aussi des contraintes a priori afin de limiter le nombre de formes du conduit vocal), on s'intéresse d'abord à la synthèse articulatoire qui consiste à effectuer l'opération inverse. Elle permet de vérifier que les données articulatoires disponibles sont cohérentes pour générer le signal acoustique correspondant, et ainsi disposer d'une base de formes et de sons pour l'étude des méthodes d'inversion. Le modèle articulatoire statistique le plus utilisé par la communauté parole en synthèse acoustique est celui de Maeda [Mae79] : nous utiliserons donc nos données pour l'estimation des paramètres de ce modèle articulatoire, afin de se rendre compte si les formes que nous avons génèrent des résultats cohérents avec les résultats de Maeda. Cela nous permettra d'évaluer si nos données sont utilisables par la communauté parole pour leurs études sur l'inversion.
Dans le but de savoir si un modèle de déformations de la langue construit à partir de données statiques (IRM) comme au GIPSA-lab suffit à décrire toutes les formes dynamiques de la langue, nous utiliserons un tel modèle pour le suivi sur des images US. Ce travail permettra d'évaluer si un modèle construit à partir de données statiques peut décrire des formes dynamiques.

2.3 Organisation du mémoire

Dans le **chapitre 3**, nous décrivons l'architecture de notre système d'acquisition de données dynamiques. Nous détaillons chacune des modalités utilisées, et leurs précisions temporelles et spatiales. Nous nous intéressons dans la seconde partie de ce chapitre à la synchronisation des données, avec notamment la description des protocoles expérimentaux permettant de mesurer les différentes valeurs de délai et de fréquence des modalités.

Le **chapitre 4** évoque le recalage des données dans le système dynamique, et décrit une technique pour recaler les données échographiques avec les données électromagnétiques. Cette étape permet de disposer de données dynamiques recalées, et notamment de pouvoir connaître la position des capteurs EM par rapport aux images US. La seconde partie de ce chapitre présente notre technique d'extraction des contours de langue dans les images échographiques. Elle se base sur une méthode de suivi bien connue en traitement d'images, les contours actifs, adaptée à notre application en utilisant la position connue des capteurs EM comme une aide au suivi.

Le **chapitre 5** utilise le système d'acquisition présenté précédemment pour étudier la variabilité et la répétabilité des phonèmes statiques. À la suite de cette étude, nous présentons un protocole original d'acquisition IRM pour des phonèmes statiques. Nous décrivons alors la méthode de recalage permettant d'exprimer les données dynamiques dans le même repère que les données statiques de l'IRM. Enfin, nous évaluons l'incertitude globale des données recalées de notre système.

Le **chapitre 6** s'attache à évaluer les données articulatoires acquises. Nous présentons des résultats dans le cadre d'une approche par synthèse articulatoire où les données recalées sont utilisées pour l'estimation des paramètres articulatoires du modèle de Maeda. Nous présentons ensuite un modèle de déformations de la langue construit à partir des données IRM et utilisé pour le suivi dans des séquences US dynamiques. Cela permettra d'évaluer si un modèle construit à partir de données statiques peut suffire à décrire des formes dynamiques.

Chapitre 3

Système d'acquisition de données dynamiques

Nous présentons d'abord l'architecture globale du système d'acquisition des données articulatoires dynamiques. Ensuite, chaque modalité utilisée est détaillée, dont leurs principales caractéristiques techniques ainsi que leurs performances en termes de résolution spatiale. Nous décrivons enfin la procédure de synchronisation des données et caractérisons sa précision temporelle.

3.1 Le système d'acquisition

3.1.1 Architecture globale

Nous avons énuméré dans le chapitre 2 différentes modalités utiles à un système d'acquisition de données dynamiques, à savoir un système échographique, électromagnétique, de stéréovision et audio. Nous avons donc utilisé ces quatre modalités pour mettre en place un système d'acquisition. Son architecture globale est schématisée sur la figure 3.1, et photographiée sur la figure 3.2.

Chaque modalité se présente comme un matériel d'acquisition de données qui nécessite pour fonctionner un système enregistrant ces données, rôle assumé par le PC de contrôle. Dans ce sens, chacune des modalités peut être vue comme un périphérique au PC de contrôle. Ce PC a donc pour rôle de recevoir et d'enregistrer les données de chaque modalité.

Le système échographique a une place à part dans notre système, car il comporte à lui seul un système d'acquisition et un système d'enregistrement (un PC déjà intégré à l'échographe). L'enregistrement est normalement commandé par l'utilisateur. Afin d'éviter cette étape manuelle et automatiser l'enregistrement, le système échographique est relié au PC de contrôle. Ce dernier envoie aussi un signal de déclenchement de l'enregistrement des données US lors de l'acquisition multimodale. Nous verrons plus en détail la façon dont la synchronisation est réalisée dans la section 3.2

En pratique, les différentes étapes d'une session d'acquisition avec notre système sont les suivantes :

- installer le locuteur assis sur la chaise, visible sur la figure 3.2 (chaise de droite). Pour des acquisitions nécessitant un corpus important, nous plaçons un écran face au locuteur de façon à ce que s'affichent les sons et phrases qu'il a à prononcer.
- Fixer les capteurs EM sur sa langue et sur sa tête (cf section 3.1.3).
- Placer la sonde US sous le menton. Cette sonde peut être tenue soit par le locuteur, soit

par un manipulateur assis à ses côtés (chaise à gauche sur la figure 3.2). Nous avons en effet constaté qu'il était plus confortable que le locuteur ait uniquement à se concentrer sur le corpus à prononcer, sans se soucier de la position de la sonde US. Le manipulateur dispose d'écrans de contrôle (cf figure 3.2) lui permettant de voir l'image échographique et la position de la sonde par rapport aux capteurs EM (cf chapitre 4). Ces écrans de contrôle permettent des acquisitions de meilleure qualité.

- Lancer une acquisition. Les données sont alors automatiquement enregistrées par le PC de contrôle et l'échographe. Nous décrivons plus en détail dans la section 3.2 notre méthode de synchronisation automatique des modalités.

Notre système d'acquisition ne nécessite donc que deux manipulateurs : un pour le PC de contrôle, et un pour tenir la sonde US. Il a été utilisé pour acquérir avec succès l'ensemble des données dynamiques présentées dans le chapitre 6 de ce manuscrit.

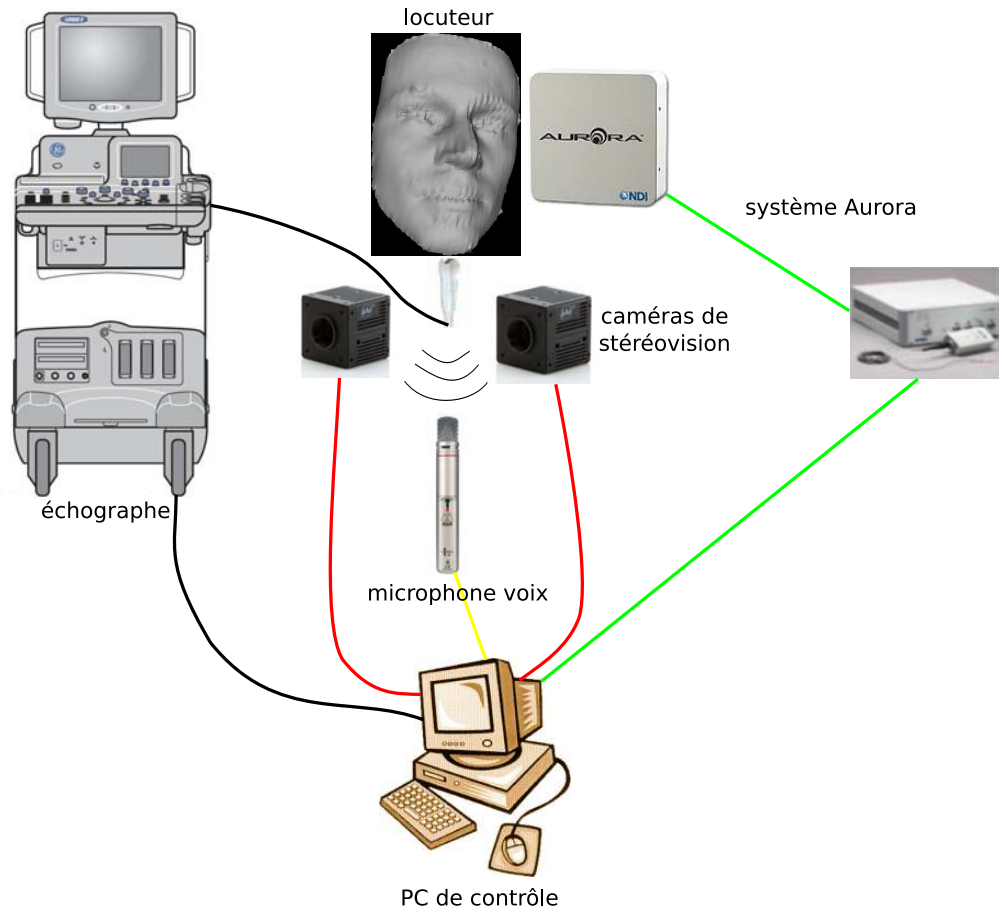


FIG. 3.1 – Architecture globale du système d'acquisition de données dynamiques.

FIG. 3.2 – Photographie du système d'acquisition de données dynamiques.

3.1.2 Les données ultrasons

3.1.2.1 Matériel

Un échographe Logiq5 (GE Healthcare[5], Chalfont St Giles, Royaume-Uni) a été choisi pour les acquisitions US. Il a été acheté par notre laboratoire en fin d'année 2005, après que nous ayons testé différents matériels aux Journées Françaises de Radiologie[6]. Ce système a été retenu, car il fournissait des images de bonne qualité comparativement aux systèmes portables, tout en restant à un coût acceptable. Ce choix pourrait sans doute être remis en question aujourd'hui avec l'arrivée sur le marché de systèmes portables plus performants, comme celui utilisé par Hueber [HCD+07] pour le projet Ouisper. Nous avons aussi choisi une sonde de type microconvexe (sonde 8C) produisant des signaux entre 5 MHz et 9 MHz. Cette sonde a été sélectionnée, car elle nous est apparue comme légère, étroite et confortable lorsqu'elle est positionnée sous le menton du locuteur, par rapport à une sonde plate plus large et pouvant être plus gênante durant la phonation. Ce type de sonde microconvexe offre une largeur et une profondeur d'acquisition suffisamment importante pour pouvoir imager une zone comme la surface de la langue.

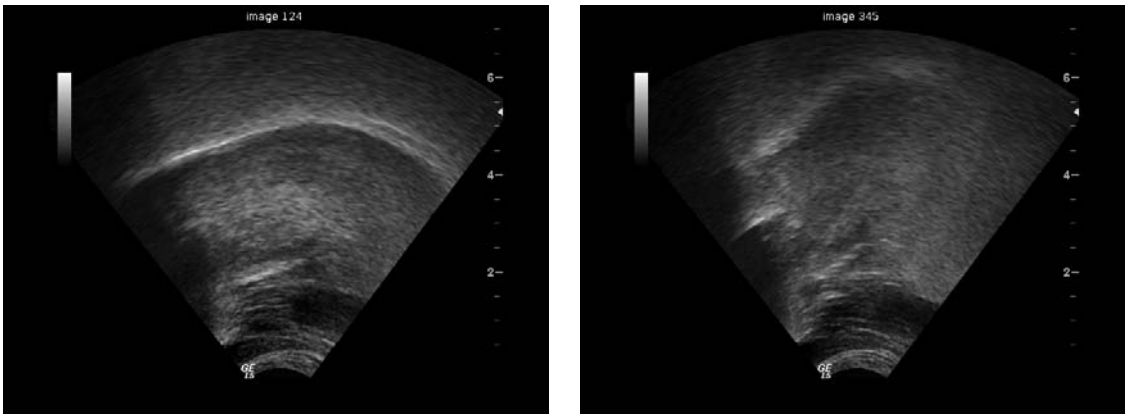
3.1.2.2 Données acquises

Les données échographiques sont enregistrées directement sur le disque dur du système Logiq5, et doivent être transférées avant de pouvoir être utilisées. Ce transfert s'effectue sous la forme de fichiers DICOM[7]. Comme avec le système échographique de Stone [Sto05], ces données peuvent aussi être récupérées en utilisant la sortie vidéo analogique de l'échographe, mais ceci a l'inconvénient de ré-échantillonner les séquences vidéo à une fréquence de 30 Hz. Nous avons donc choisi d'utiliser le transfert DICOM pour préserver la fréquence originale d'acquisition, afin de pouvoir restituer la dynamique de la langue lors de mouvements rapides.
Les images obtenues ont toutes une taille de 534×432 pixels.

[5] `http://www.gehealthcare.com`
[6] JFR : `http://www.sfrnet.org`
[7] **DICOM** (Digital Imaging and COmmunications in Medicine) est le format de fichier standard faisant référence dans le domaine de l'imagerie médicale, `http://medical.nema.org`

3.1.2.3 Réglages

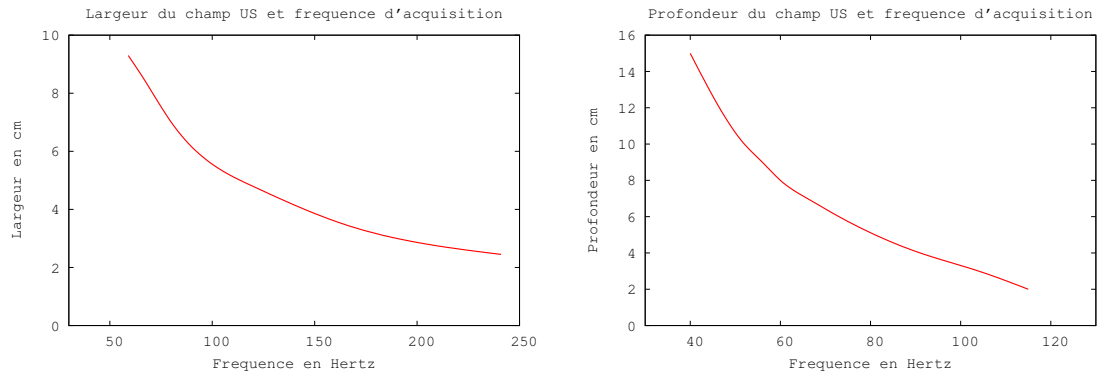
Avec la sonde US placée sous le menton, la meilleure qualité d'images US acquises est atteinte pour les sons où la langue est proche de l'horizontale, comme le /a/ de la figure 3.3.a. Les sons générant des formes plus complexes, où des portions sont proches de la verticale, sont plus difficiles à imager comme le /k/ de la figure 3.3.b [Sto05]. Ce phénomène s'explique de par la nature même de la formation de l'image échographique : les échos réfléchis par la zone de contact langue/air lorsque la langue est horizontale sont plus facilement captés par la sonde que les échos renvoyés par des zones verticales.



(a) (b)

FIG. 3.3 – Images US de la langue : (a) /a/. (b) /k/

Pour toute acquisition échographique, un compromis doit être choisi entre tous les réglages possibles de la machine. Pour des acquisitions sur la langue, ce compromis fait principalement intervenir la fréquence d'acquisition des images, la profondeur sur laquelle on désire imager la zone et la largeur du champ échographique. Plus la profondeur ou la largeur est importante, plus la fréquence d'acquisition des images est faible (cf figures 3.4.a et 3.4.b). La résolution des images US résultantes dépend de la profondeur choisie.



(a) (b)

FIG. 3.4 – (a) Largeur du champ US (cm) en fonction de la fréquence d'acquisition (Hz). (b) Profondeur du champ US (cm) en fonction de la fréquence d'acquisition (Hz).

L'échographie offrant la possibilité d'obtenir une largeur de champ importante pour des acquisitions sur la langue, nous avons choisi cette option, tout en couvrant la profondeur adéquate pour avoir la surface de la langue pour tous ses mouvements. Un réglage typique utilisé est une largeur de 8.6 cm et une profondeur de 7 cm, pour obtenir une fréquence d'acquisition des images de 66 Hz. Cela donne pour les images une résolution annoncée par l'échographe de 0.1753 mm/pixel. Bien que ces réglages conviennent à une majorité de locuteurs, il peut s'avérer nécessaire de les adapter à leur morphologie. Les résultats exposés dans la suite de ce manuscrit sont obtenus pour toutes les images échographiques présentées avec un réglage à 66 Hz.

Stone [Sto05] précise que l'échogénicité est très variable suivant le sujet. Nous avons effectivement constaté ce phénomène : un locuteur féminin, maigre et jeune semble générer des images de meilleure qualité qu'un locuteur masculin, plus âgé, et plus adipeux. Ce phénomène a aussi été constaté sur les enfants qui engendrent des images où les contours de langue apparaissent avec un fort contraste. Ces impressions subjectives ne constituent en aucun cas une règle empirique, mais ce phénomène a été constaté sur divers sujets testés dans notre laboratoire. Il semblerait aussi que des sujets soient plus échogènes que d'autres, sans que nous soyons capables de fournir une explication. La seule solution est de tester chaque locuteur.

3.1.2.4 Particularités de la sauvegarde des données

L'échographe est un système fermé qu'il n'est pas possible de modifier. Il effectue à la fois les acquisitions et les enregistrements, et nous ne pouvons intervenir sur le système pour récupérer par exemple le flux d'images en temps réel. Nous sommes donc contraints d'utiliser la méthode proposée par la machine pour enregistrer les données. Elle a deux particularités : la durée d'enregistrement est limitée à quinze secondes, quelle que soit la fréquence d'acquisition des images ; et l'enregistrement d'une séquence s'effectue a posteriori en pressant un bouton qui a pour effet de sauvegarder les quinze dernières secondes acquises. Cette sauvegarde s'effectue sur le disque dur de l'échographe et prend une trentaine de secondes. Nous verrons que l'enregistrement a posteriori est un point important et délicat pour la synchronisation des données.

3.1.2.5 Résolution spatiale

L'échographe indique une valeur de résolution spatiale pour chaque valeur de profondeur de champ échographique choisie. Prager [PRGB98] pour le calibrage d'un système échographique de type mains libres préconise de calculer la résolution, car il semblerait que les valeurs annoncées par les constructeurs soient différentes des valeurs effectives.

Pour effectuer ce calcul, un fantôme 3D est nécessaire. Ce fantôme possède des caractéristiques géométriques 3D connues, et qui sont visibles dans les images US. Ainsi, les rapports des mesures 3D et imagées sont effectués pour en déduire une valeur de résolution.

Nous avons donc fait l'acquisition d'un fantôme (modèle 055A, cf figure 3.5.a) fabriqué par la société CIRS Inc.[8] (Norfolk, Virginie, États-Unis). Il est constitué de filaments parallèles entre eux, formant un « A » renversé. Les distances relatives entre les filaments sont connues à 0.2 mm près selon le fabricant. En détectant la position de ces filaments dans l'image US (cf figure 3.5.b), et en les rapportant à leurs distances relatives connues, on est en mesure de calculer la résolution de chaque image.

Nous avons effectué des acquisitions US avec le même réglage qu'utilisé pour un locuteur, soit une résolution indiquée par l'échographe de 0.1753 mm/pixel.

[8] `http://www.cirsinc.com`

(a) (b)

FIG. 3.5 – Images du fantôme US CIRS Inc. (a) Vue d'extérieur (d'après la documentation). (b) Image US.

Sur trois séquences US de 975 images, les points du fantôme sont automatiquement extraits sur chaque image. Pour cela, un seuillage par hystérésis est effectué sur l'image pour obtenir un ensemble de taches correspondant aux positions des filaments formant le motif en « A ». Le centre de gravité de chaque tache est alors calculé pour estimer leur position dans l'image. Les distances relatives entre les filaments sur les axes horizontaux et verticaux sont calculées pour en déduire des valeurs moyennes de résolution sur chaque axe pour chaque image. La moyenne des résolutions trouvées pour la séquence est indiquée dans le tableau 3.1.

	Résolution X (mm/pixel)	**Résolution Y (mm/pixel)**
Moyenne	0.17293	0.18134
Écart-type	0.007	0.007

TAB. 3.1 – Calcul de la résolution des images échographiques avec un fantôme dédié.

Les valeurs moyennes obtenues sont différentes suivant les axes horizontaux et verticaux : les pixels US sont donc anisotropiques. Il apparaît aussi très clairement que cette résolution calculée est différente de la résolution donnée par l'échographe (0.1753mm/pixel). De plus, l'écart-type calculé est relativement élevé. Nous étudierons dans le chapitre 5 l'influence de cette incertitude de résolution sur le recalage global des modalités de notre système.

Dans la suite de ce manuscrit, nous utiliserons ces deux valeurs moyennes de résolution.

3.1.3 Les données électromagnétiques

3.1.3.1 Matériel

Pour notre application où nous cherchons un système de localisation pouvant être utilisé à la fois dans la cavité buccale, sur la sonde échographique et sur la tête du locuteur, nous avons choisi un système électromagnétique Aurora de la société Northern Digital Inc.[9] (Waterloo, Ontario, Canada). D'autres systèmes EM existent comme le système de Carstens présenté dans le chapitre 2, mais ce dernier a les désavantages suivants : il n'était pas utilisable en temps réel (il l'est devenu depuis 2007), le locuteur doit positionner sa tête dans un cube plexiglas, rendant difficile l'utilisation avec des caméras de stéréovision, et il est très coûteux comparativement à d'autres systèmes EM. Le système EM « Microbird » de la société Ascension Technology[10] (Burlington, États-Unis) a aussi été envisagé, mais la taille des capteurs proposés ($1.8 \times 8.4mm$) était trop importante pour envisager de les utiliser sur la langue.

Nous décrivons dans les paragraphes suivants les caractéristiques techniques du système Aurora. Ce système a l'avantage de posséder une API (Application Programmer Interface) nous permettant de développer notre propre application, et donc de contrôler l'acquisition à partir du PC de contrôle (cf figure 3.1).

Description des capteurs Le système comporte un générateur de champ magnétique, une unité de contrôle, quatre unités d'interface, et différents types de capteurs (cf figure 3.6.a). Tous les éléments sont reliés à l'unité de contrôle, qui envoie les données des capteurs au PC de contrôle sous forme de données texte via le port série.
Le système repose sur des capteurs à 5 degrés de liberté (DDL en français, ou DOF en anglais pour Degrees Of Freedom) dont 3 DDL donnent la position et 2 l'orientation (la donnée manquante est celle de rotation du capteur autour de son axe Z). Ces capteurs miniatures sont des bobines cylindriques de taille de 0.5 mm × 8 mm et reliées par un fil aux unités d'interface, elles-mêmes reliées à l'unité de contrôle. Deux capteurs 5 DDL peuvent être reliés à une unité d'interface, ce qui autorise l'utilisation d'au maximum huit capteurs à 5 DDL. Deux capteurs 5 DDL peuvent être fixés rigidement l'un par rapport à l'autre pour former un capteur à 6 DDL : NDI propose un outil de calibrage pour déterminer la transformation rigide entre les deux capteurs 5 DDL. Nous avons acquis avec notre système, en plus des capteurs à 5 DDL fournis par NDI, deux capteurs à 6 DDL manufacturés par la société Traxtal[11] (Texas, États-Unis) : un stylet et un capteur, tous deux directement utilisables avec le système Aurora.

Volume utile Les données de position et d'orientation des capteurs sont établies dans un repère dont le centre est situé dans le générateur de champ électromagnétique. L'axe X est horizontal, Y vertical, et l'axe Z caractérise la profondeur, formant ce qu'on appellera par la suite le repère EM. NDI préconise l'utilisation des capteurs dans un volume utile de 50 cm × 50 cm × 50 cm (cf figure 3.6.b). En effet, hors des limites de ce volume, l'intensité du champ magnétique diminue conséquemment et les mesures données sont susceptibles d'être faussées ou tout simplement manquantes.

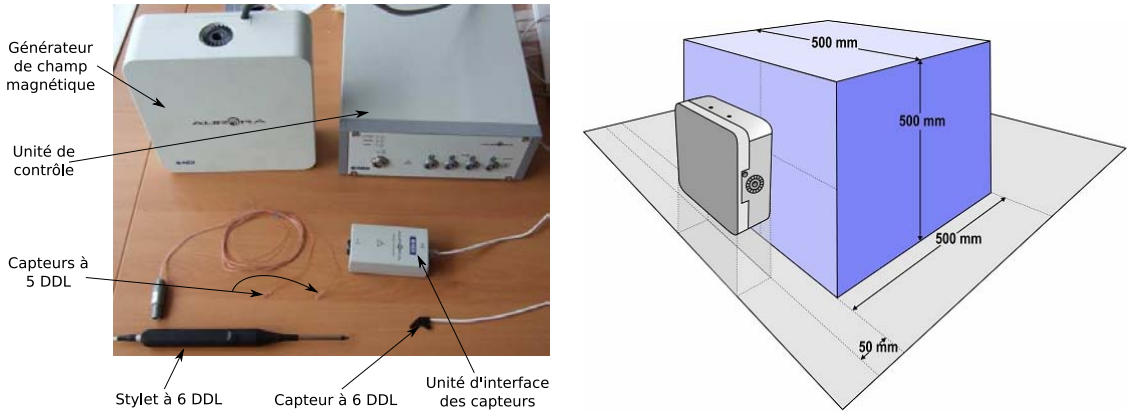
[9] `http://www.ndigital.com`
[10] `http://www.ascension-tech.com`
[11] `http://www.traxtaltech.com`

Fréquences d'acquisition Le système fournit des mesures à une fréquence de 40 Hz (selon le constructeur) si au plus six capteurs sont connectés. Si sept ou huit capteurs à 5 DDL sont utilisés, la fréquence passe à 20 Hz. Puisque nous désirons utiliser au moins un capteur sur l'apex qui a une dynamique importante, nous préférons bénéficier de la fréquence d'acquisition la plus importante possible, et sommes donc limités à utiliser au maximum six capteurs à 5 DDL.

Configuration d'utilisation Dues aux contraintes inhérentes au système d'acquisition, nous avons donc choisi la configuration suivante pour le placement des capteurs EM lors des acquisitions (cf figure 3.7.a) :

- deux capteurs 5 DDL sont utilisés sur la langue : un sur l'apex pour compléter l'information US, et un sur le dos de la langue pour la corroborer. Nous verrons aussi dans le chapitre 4 que ces deux capteurs sont utilisés pour aider le suivi du contour de la langue dans les images échographiques,
- un capteur 6 DDL est fixé sur la sonde US, afin de pouvoir situer dans le repère EM la position de la sonde. Nous verrons dans le chapitre 4 une méthode de calibrage pour pouvoir spatialement lier les repères EM et US grâce à ce capteur,
- deux capteurs 5 DDL sont utilisés pour former un repère tête à 6 DDL. Pour cela, un capteur est fixé derrière chaque oreille du locuteur durant les acquisitions, en prenant soin que leur deux axes ne soient pas parallèles. Le repère tête permet d'exprimer toutes les mesures EM dans un repère indépendant des mouvements de tête. Nous verrons aussi dans le chapitre 5 que ce repère intrinsèque à la tête est utilisé pour le recalage des données US et EM avec des données IRM.

Cette configuration a aussi été choisie suite à l'étude de précision sur les capteurs effectuée en section 3.1.3.3.



(a) (b)

FIG. 3.6 – (a) Description du système Aurora. (b) Volume utile de mesure par rapport au générateur de champ magnétique (d'après la documentation).

3.1.3.2 Utilisation de capteurs sur la langue

Les capteurs EM utilisés sur la langue (cf figure 3.7.b) nécessitent une préparation préliminaire afin de faciliter leur collage sur la langue du locuteur. Nous nous sommes inspirés de la technique utilisée pour le collage des capteurs des articulographes AG100, AG200 et

AG500 et décrite sur la page du laboratoire de phonétique de l'UCLA (University of California Los Angeles, États-Unis) : `http://www.humnet.ucla.edu/humnet/linguistics/faciliti/facilities/physiology/ema.html#Placing_coil`
La préparation consiste à :

- plonger le capteur dans du latex liquide afin de lui fournir une protection fine ;
- coller le capteur sur un morceau de tissu fin (soie...) à l'aide d'une colle de type cyanoacrylate (communément connue sous le nom de Superglue©).

Même si les capteurs fournis disposent d'une protection, la première étape renforce cette protection, et elle permet aussi de faciliter le collage du capteur sur le tissu. La seconde étape permet de faciliter le collage du capteur sur la langue en utilisant un intermédiaire comme la soie. Il suffit de mettre de la colle sur cet intermédiaire et d'appuyer légèrement sur le capteur pour faire tenir le tout sur la langue.
Une telle préparation permet de faire tenir le capteur de 10 à 30 minutes sur la langue. Les temps de collage dépendent de la quantité de colle utilisée, du séchage éventuel de la langue avant le collage, et de la composition salivaire du sujet. Enfin, nous avons noté que cette préparation permet de prolonger la durée de vie des capteurs.

(a) (b)

FIG. 3.7 – (a) Disposition des capteurs EM sur le locuteur (vue de dessus). (b) Deux capteurs EM collés sur la langue.

3.1.3.3 Précision des capteurs en position statique

Les spécifications données par le constructeur indiquent une précision géométrique de 0.9 mm et une précision angulaire de 0.3 ° à l'intérieur du volume utile [Kir05]. Il est à noter que ces valeurs ont évolué au cours de la thèse : lorsque le système a été acquis en 2005, le constructeur annonçait alors une précision de 0.43 mm en translation et de 0.39 ° en rotation. Ces valeurs peuvent également être sensibles à l'environnement magnétique. Nous avons donc évalué la précision des capteurs dans ce contexte, et en particulier, nous avons recherché une influence potentielle de la sonde US sur les données du capteur EM utilisé pour son suivi.

Protocole expérimental Nous avons une table micrométrique (cf figure 3.8) disposant d'un bras articulé pouvant effectuer des translations sur deux axes horizontaux et des rotations sur deux angles pan et tilt. La précision de la table dans son positionnement est de 0.48 mm en translation et de 0.013 ° en rotation. Cette table possède d'après le constructeur une grande répétabilité (non quantifiée). Nous avons fixé un capteur EM à 5 DDL sur le bras articulé. Le générateur de champ EM est placé en bout de table de telle façon que son axe Y soit orthogonal

au plan dans lequel le capteur effectue ses translations. À partir de quelques positions capteurs et de table, nous avons manuellement aligné les axes de translation de la table avec les axes X et Z du repère EM. Ces positions nous ont également permis de calibrer la translation entre les deux repères, de façon que le changement de repère est totalement connu. Dans la suite, nous ne nous référerons qu'au repère EM.

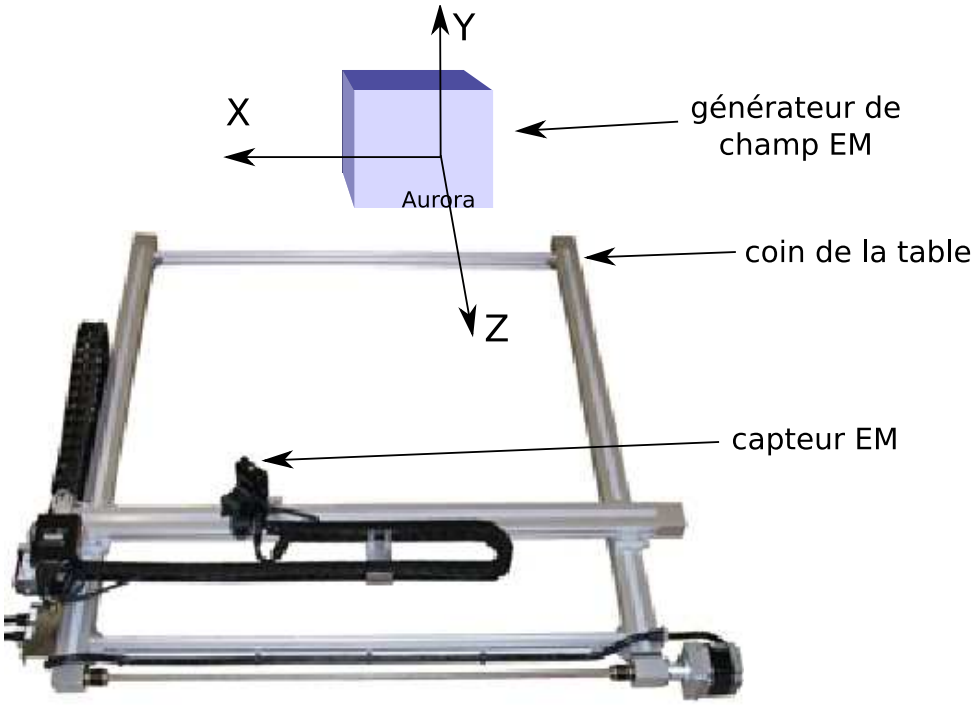


FIG. 3.8 – Table micrométrique utilisée pour mesurer la précision des capteurs EM.

Dans les expérimentations suivantes, des mesures de position du capteur sont effectuées sur les deux axes. Elles sont échantillonnées tous les 10 millimètres sur l'axe X (de -150 mm à +180 mm) et tous les 50 millimètres sur l'axe Z (de 100 mm à 350 mm), ce qui donne 210 positions.

Étude d'erreur Tout d'abord, une première étude a été effectuée en comparant la mesure du position du capteur EM et la mesure de position donnée par la table, considérée comme étant la mesure de référence. La distance entre les deux positions a été calculé. Les moyennes de ces distances pour les 210 positions de la table sont présentées sur la figure 3.9.

On observe un pic d'erreur en position $X = -120$ et $Z = 140$. Il est dû au coin de la table micrométrique désigné sur la figure 3.8. De nombreux fils électriques aboutissent dans ce coin et perturbent les mesures EM. Le système apparaît donc comme très sensible aux perturbations électromagnétiques.
Nous n'observons pas de fortes variations dans ces erreurs moyennes suivant la position du capteur. Même s'il semble y avoir une légère augmentation de l'erreur en $Z = 350$ mm, cela correspond plus aux positions proches du bord de la table ($X = -150$ mm), visiblement plus aptes à perturber les mesures de position.

Cette étude pourrait être améliorée. En effet, nous nous basons pour ces mesures sur la précision de la table qui est de 0.48 mm en translation, soit du même ordre de grandeur que

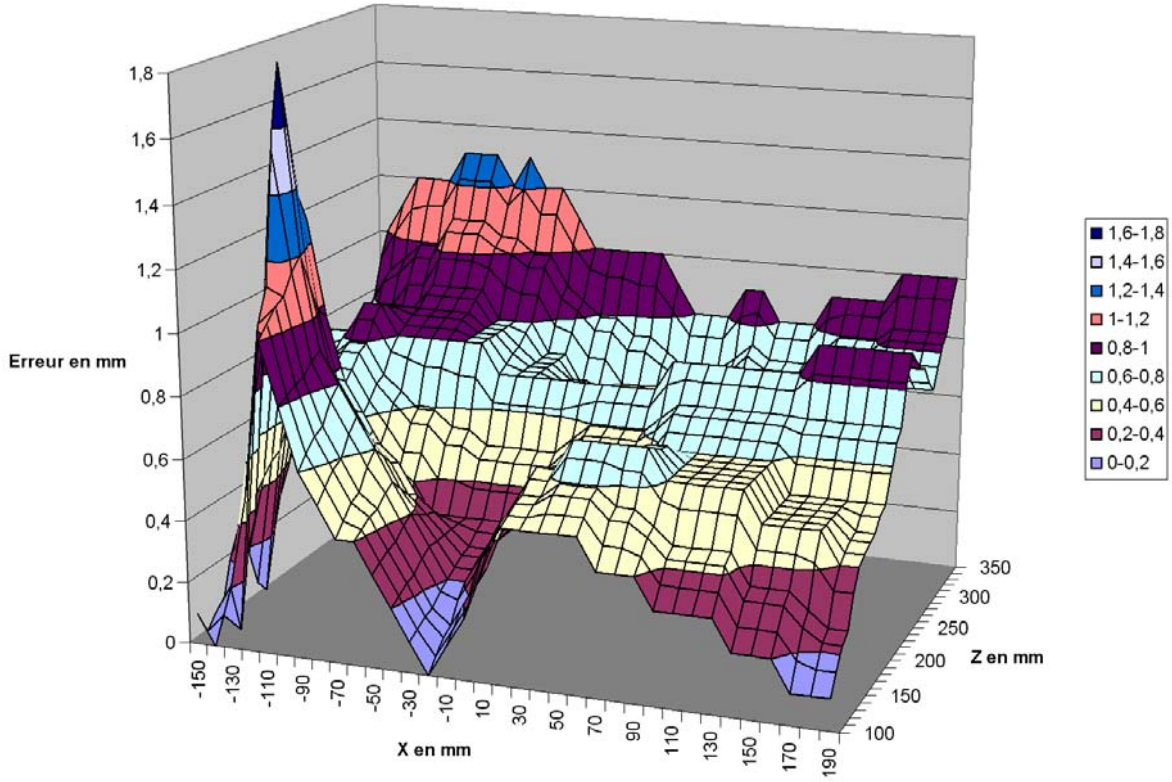


FIG. 3.9 – Erreur moyenne entre les mesures capteur et la valeur de position de la table micrométrique.

la précision trouvée sur le capteur. Nous aurions besoin d'un système de référence plus précis, comme un système optique utilisé par Kroos [Kro08]. Les deux erreurs, celles de la table et du capteur EM, sont donc probablement mutuellement présentes dans les résultats de la figure 3.9.

Puisque l'erreur de positionnement du capteur EM sur la table est sensiblement la même que celle de la table seule, on peut considérer que la précision du capteur EM peut s'estimer par une mesure de répétabilité. Nous étudions dans la section suivante cette mesure.

Étude de la répétabilité Cette étude consiste à observer les différentes valeurs de position données par le capteur pour une même position de la table. 100 mesures capteurs sont effectuées à la même position de table puis la position suivante est acquise. Nous calculons une position moyenne et l'erreur est définie par la racine carrée de la distance quadratique moyenne à cette position moyenne. Les résultats obtenus sont présentés sur la figure 3.10.

La perturbation magnétique due au coin de la table est ici aussi présente. On observe que l'erreur augmente au fur et à mesure de l'éloignement (axe Z) du capteur par rapport à l'origine du repère EM. En revanche, il reste stable sur l'axe X.

Les mesures capteurs sont donc de moins en moins répétables au fur et à mesure de l'éloignement du générateur EM. Les erreurs restent inférieures à 1 mm si le capteur EM est placé à moins de 35 cm du générateur de champ EM.
L'expérience précédente consiste à effectuer 100 mesures à une position p_1, puis 100 mesures à une position p_2... La table ne bouge donc pas pour une même mesure, et seule l'erreur de répétabilité du capteur est mesurée.

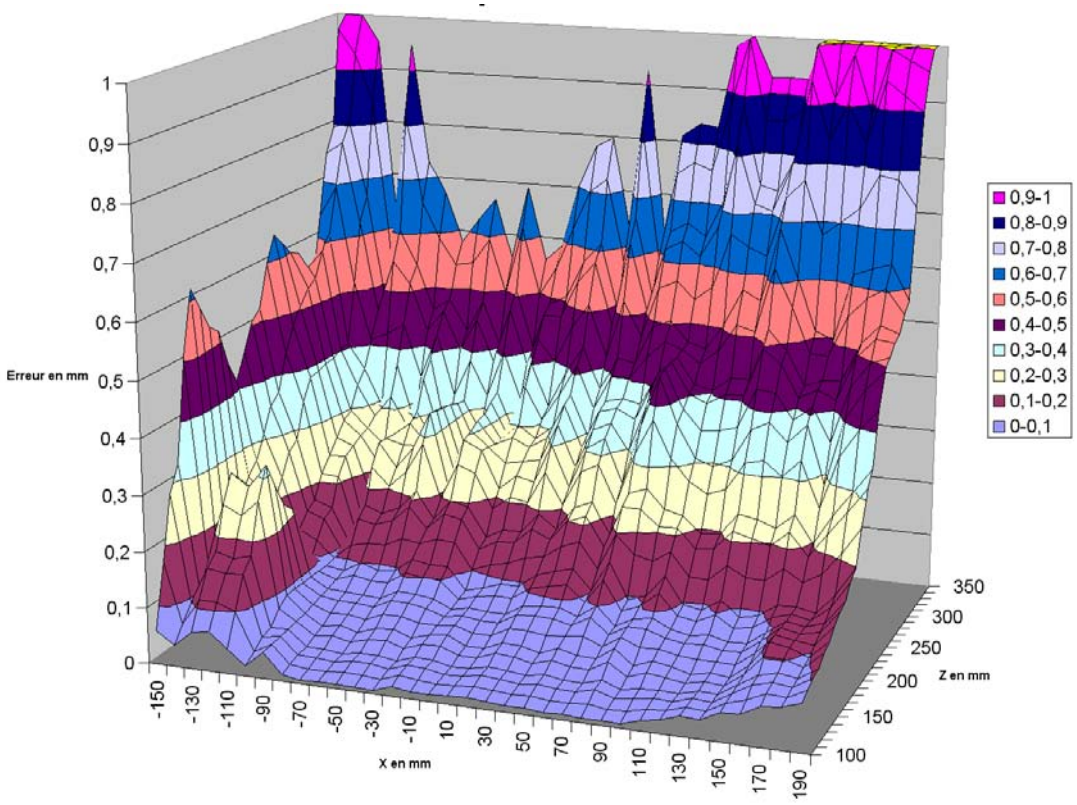


FIG. 3.10 – Erreur de répétabilité des 100 mesures de position du capteur sur la table micrométrique.

Nous avons aussi testé les configurations suivantes :

- une mesure en position p_1 puis une mesure en position p_2, le tout répété 100 fois,
- une mesure en position p_2 puis une mesure en position p_1, le tout répété 100 fois,
- une mesure en position p_1 puis une mesure en position p_2 en passant par une position p_3, le tout répété 100 fois,

Les résultats trouvés ont été similaires aux résultats de la figure 3.10. Cela confirme les résultats de répétabilité. L'imprécision d'un capteur n'est pas lié à sa trajectoire (sens dans lequel il atteint sa position) et ne varie pas dans le temps.

Le tableau 3.2 présente les erreurs de répétabilité trouvées pour différentes valeurs de distance au générateur (distances sur Z moyennées en intégrant les mesures pour des positions différentes selon X). Nous avons aussi effectué une mesure à l'extrême limite du volume utile (à 500 mm de l'origine du repère EM), pour avoir une idée de la précision en ce point.

Distance (mm)	150	300	500
Erreur (mm)	0.31	0.53	3.58

TAB. 3.2 – Erreurs de répétabilité sur les mesures capteur 5 DDL suivant la distance au générateur de champ EM.

Ces mesures corroborent les résultats présentés précédemment. Si le générateur de champ

EM est placé à la limite du volume utile des capteurs, l'erreur augmente drastiquement pour atteindre près de 3.5 mm. En deçà de 30 cm, l'erreur est inférieure au millimètre.

Pour ces trois positions, nous avons aussi testé les rotations grâce à la tourelle de la table robotique qui peut effectuer des rotations de type « pan-tilt », dans un intervalle de [-159 ° ; 159 °] pour le pan et [-47 ° ; 31 °] pour le tilt. L'erreur angulaire pour une position est définie comme la moyenne des différences angulaires entre les données capteurs deux à deux, pour cette position. Le tableau 3.3 présente les erreurs sur l'angle de rotation (en degré) pour un capteur EM à 5 DDL.

Distance (mm)	150	300	500
Erreur (degré)	0.39	0.50	0.84

TAB. 3.3 – Erreur de répétabilité en rotation sur un capteur 5 DDL en fonction de la distance au générateur de champ EM.

Les mesures capteurs semblent plus répétables pour la rotation que pour la translation. Elles sont semblables à celles données par NDI, en restant inférieures à un degré dans le volume utile, voire même à un demi-degré si l'on reste proche du générateur de champ EM.

Résultats capteur EM sur la sonde Les études précédentes montrent que les capteurs EM sont très sensibles aux perturbations électromagnétiques. Dans notre système d'acquisition où un capteur EM est fixé sur la sonde US, il convient d'étudier le comportement du capteur dans cette configuration.
Un capteur EM a été fixé sur la sonde US en fonctionnement, elle-même fixée sur la table robotique. L'étude de répétabilité effectuée ci-dessus a été réitérée dans cette nouvelle configuration. Les résultats sont présentés dans le tableau 3.4.

Distance (mm)	150	300	500
Erreur (mm)	0.87	0.76	3.39

TAB. 3.4 – Erreur de répétabilité sur un capteur 5 DDL fixé sur la sonde US suivant la distance au générateur de champ EM.

Nous observons peu de différence avec les précédents résultats. La sonde semble ne perturber que très modérément les mesures de position des capteurs EM. Il faut toutefois bien prendre soin de ne pas éloigner la sonde US au-delà de 30 cm du générateur. Les mesures sur la rotation sont similaires à celles indiquées sur le tableau 3.3.

Conclusion L'erreur moyenne des capteurs du système EM est inférieure à 1 mm en translation et à 0.5 ° en rotation, et est similaire aux résultats annoncés par NDI. Pour des mesures fiables, il est nécessaire de rester proche du générateur du champ EM. Une distance de 30 cm est acceptable à la fois pour ce qui est de la précision et d'un point de vue pratique pour englober la tête du locuteur.

Pour nos expérimentations où des capteurs sont fixés sur la tête du locuteur et sur sa langue, la meilleure configuration possible est donc de positionner sa tête la plus proche possible du générateur de champ EM. Nous avons décidé de placer ce dernier au dessus de l'épaule du

locuteur lors des acquisitions afin de garder la surface de son visage visible par les caméras de stéréovision (cf figure 3.7.a).

La sonde échographique ne perturbe que modérément les mesures EM, et son utilisation dans le champ EM est donc envisageable. Cependant, les mesures capteurs sont très sensibles aux perturbations électromagnétiques. Il est donc nécessaire de bien prendre soin d'ôter tout matériau ferromagnétique (lunettes, montre...) lors de l'utilisation du système EM.

3.1.3.4 Précision des capteurs en dynamique

Dans notre configuration (cf figure 3.7.a), les capteurs fixés sur la tête et sur la sonde ont des mouvements lents. En revanche, les deux capteurs fixés sur la langue sont soumis à des mouvements rapides de langue, et leur précision doit aussi être évaluée dans cette configuration.

Kroos dans son étude [Kro08] a comparé la précision à des vitesses différentes des capteurs d'un système électromagnétique AG500, en se référant à un système de suivi optique plus précis. Il précise qu'il a effectué des mouvements rapides sans pour autant les quantifier. Ses résultats montrent qu'il y a une importante disparité dans la précision entre des mesures statiques et dynamiques : il obtient moins d'un millimètre de précision en statique contre plus de quatre millimètres en dynamique.
Nous avons effectué des mesures sur nos capteurs en les fixant rigidement l'un par rapport à l'autre et en plaçant le dispositif dans la bouche d'un locuteur pour être dans nos conditions expérimentales. Nous avons ensuite calculé la distance entre ces deux capteurs en position statique pendant dix secondes. Puis nous avons fait effectuer des mouvements rapides au dispositif en le secouant le plus rapidement possible par le fil pendant dix secondes. Les résultats obtenus sont présentés dans le tableau 3.5. La vitesse indiquée est donnée seulement à titre indicatif : elle correspond à la moyenne des vitesses calculées entre deux positions consécutives. Ces mesures de position étant pour la plupart imprécises, cette mesure ne doit être considérée que comme donnant un ordre d'idée de la vitesse.

Expérience	**Vitesse des mouvements**	**Durée**	**Distance moyenne**	**Distance minimale**	**Distance maximale**	**Écart type**
statique	0.0 mm/s	10 sec	14.54 mm	14.40 mm	14.69 mm	0.07 mm
dynamique	19.4 mm/s	10 sec	15.74 mm	5.26 mm	87.28 mm	4.92 mm

Tab. 3.5 – Comparaison entre les positions de deux capteurs fixés rigidement l'un par rapport à l'autre lors d'acquisitions statiques et dynamiques.

Comme Kroos, nous observons de grandes disparités entre les mesures statiques et dynamiques. Certaines mesures des capteurs EM à une vitesse importante peuvent être aberrantes, et les mesures des capteurs EM placés sur la langue dans notre système d'acquisition doivent donc être considérées avec précaution.

Remarque Nous avons d'abord pensé que les capteurs EM pouvaient être perturbés par les plombages des locuteurs. Nous avons récupéré chez un prothésiste dentaire, différents alliages utilisés pour les couronnes et les bridges :

- alliage à base de cobalt (64% de cobalt, 28% de chrome, 5.1% de manganèse)
- alliage pour couronne et bridge (42% de fer, 28% de nickel, 22% de chrome, 4% de silicium, 3% de molybdène)

- alliage pour céramique (62 % de nickel, 26.2% de chrome, 9.4% de molybdène, 2.4% de silicium)

En approchant et éloignant ces alliages d'un capteur, on peut observer la variabilité des mesures et déduire ainsi si la mesure capteur est perturbée par l'alliage. Aucun ne la perturbe de façon significative, et ne peut expliquer les grandes disparités observées dans le tableau 3.5. Les variabilités trouvées sont similaires à celles obtenues sans alliage.
La même expérience réalisée sur un support rigide, mais en dehors de la bouche donne des variations similaires à celle à l'intérieur de la bouche. Ces variations sont donc inhérentes aux capteurs EM, et même à la technologie du calcul du positionnement d'après des mesures EM puisque le phénomène est similaire pour les systèmes Aurora et AG500.

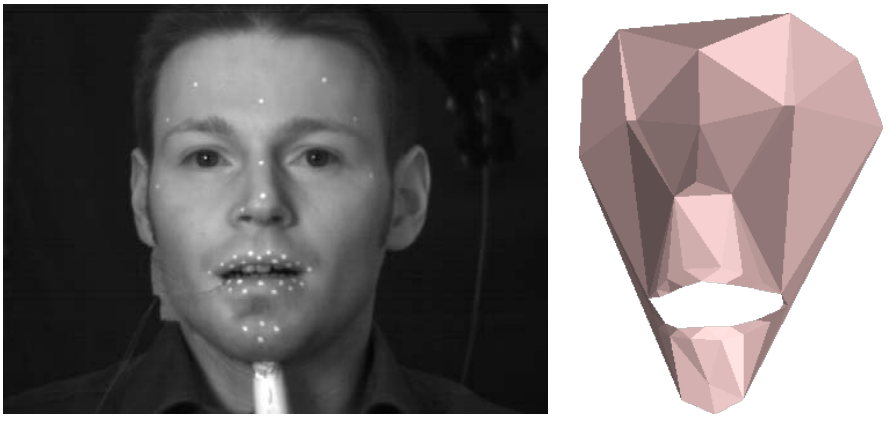
3.1.4 Les données de stéréovision

3.1.4.1 Matériel et utilisation

Nous avons utilisé un système existant, préalablement développé et utilisé dans notre laboratoire. Nous décrivons ici les principales caractéristiques de ce système, et nous invitons le lecteur à se reporter à [WDBP+05] pour plus de détails.

Deux caméras de stéréovision (JAI A33, Stemmer Imaging[12], Puchheim Allemagne) sont utilisées pour filmer le visage du locuteur, notamment la position de ses lèvres durant les acquisitions. Ces deux caméras fournissent des images noir et blanc (au format PGM, « Portable Gray Map ») de taille 640×480 pixels. La fréquence d'acquisition annoncée par le constructeur est de 120 Hz.

Ces deux caméras sont calibrées en début de chaque acquisition dynamique à l'aide d'une mire de calibrage à la précision micrométrique. Des marqueurs sont peints sur le visage du locuteur (cf figure 3.11.a) pour permettre la reconstruction tridimensionnelle de la surface de son visage (cf figure 3.11.b). Deux projecteurs à intensité d'éclairage réglable permettent d'éclairer la surface du visage du locuteur durant les acquisitions.



(a) (b)

FIG. 3.11 – (a) Visage du locuteur avec des marqueurs peints sur son visage. (b) Reconstruction tridimensionnelle de la surface de son visage à partir d'une paire d'images de stéréovision.

[12] `http://www.stemmer-imaging.de`

3.1.4.2 Précision

Des mesures d'incertitude ont été effectuées sur la reconstruction tridimensionnelle effectuée à partir des données de stéréovision d'un visage. Cette incertitude dépend de la détection des marqueurs dans les images (1 pixel dans notre cas). Elle a été estimée à 1.1 mm sur le plan fronto-parallèle est à 2.4 mm en profondeur.

Remarque Nous n'avons pas travaillé directement avec le système de stéréovision pour cette thèse, mais nous l'avons utilisé pour compléter le système d'acquisition de données articulatoires. Pour cette raison, nous n'avons pas effectué d'étude exhaustive sur la précision de ces données. En particulier, l'incertitude sur la détection des marqueurs mériterait d'être précisée, car l'estimation grossière que nous avons employée mène à des erreurs de reconstruction supérieures à celles rencontrées en pratique.

3.1.5 Récapitulatif

Les principales caractéristiques de chaque modalité de notre système sont résumées dans le tableau 3.6.

	EM	**US**	**Stéréovision**	**Audio**
Fréquence	40 Hz	66 Hz	120 Hz	44100 Hz
Temps d'enregistrement	illimité	15 secondes	illimité	illimité
Format des données	fichiers texte	DICOM	images PGM	fichiers WAV
Type d'enregistrement	temps réel	a posteriori	temps réel	temps réel

Tab. 3.6 – Résumé des principales caractéristiques des modalités du système d'acquisition de données dynamiques.

La modalité limitant la durée des enregistrements est l'US qui sauvegarde quinze secondes de données. À cause de cela, plusieurs acquisitions de quinze secondes sont nécessaires pour acquérir un corpus de données. La fréquence d'acquisition des différentes modalités est satisfaisante pour des acquisitions sur le conduit vocal. Les données EM, avec une fréquence de 40 Hz restent en-deça du minimum requis. Cependant, cette limitation devrait être levée dans un futur proche dans les évolutions du système Aurora (cf chapitre 6).

Toutes les modalités sont reliées par un PC de contrôle (cf figure 3.1) qui contrôle l'enregistrement des données de chacune. Nous nous intéressons dans la partie suivante à la synchronisation de ces modalités.

3.2 Synchronisation des données

3.2.1 Principe

Nous avons vu dans le chapitre 2 que la synchronisation des données consiste à mettre temporellement en correspondance toutes les modalités utilisées dans notre système ([HGK04], [SBW07]). En effet, chaque modalité acquiert ses données indépendamment et rien n'assure qu'elles soient synchronisées.

Pour cela, le principe adopté pour la synchronisation consiste à étiqueter chaque donnée provenant d'une modalité par son temps de réception sur le PC de contrôle. Techniquement,

ce temps de réception correspond au nombre de cycles et à la fréquence du microprocesseur du PC de contrôle. Sa précision est de l'ordre de la nanoseconde, ce qui est suffisant pour notre application où la modalité ayant la fréquence d'échantillonnage la plus élevée, hormis l'audio, est la stéréovision (120 Hz). Ce processus a l'avantage de pouvoir être effectué automatiquement par le PC de contrôle.

Il reste cependant à mesurer le temps écoulé entre l'acquisition de la donnée sur une modalité et sa réception sur le PC de contrôle. Ce temps, appelé délai, peut par exemple correspondre à un temps de traitement interne à la modalité d'acquisition. Il doit être déterminé pour toutes les modalités utilisées. Nous traiterons le système échographique différemment des autres modalités, car c'est un système fermé, qui effectue à la fois les acquisitions et les enregistrements. Nous ne pouvons donc pas étiqueter les données de cette modalité.

3.2.2 Estimation des délais entre les modalités

3.2.2.1 Principe

Pour être en mesure d'estimer les délais d'acquisition entre les modalités, il est nécessaire de pouvoir mettre en correspondance un événement identifiable dans deux modalités. Cet événement est repéré dans chaque modalité, et les instants d'étiquetage par le PC de contrôle sont comparés. Leur différence fournit une mesure du délai recherché.

Pour notre système d'acquisition, nous avons choisi la modalité audio comme modalité de référence. En effet, nous proposons des dispositifs expérimentaux simples à mettre œuvre avec l'audio, où l'événement est identifiable à la fois dans les données audio et la modalité concernée.

Lors de ces expériences, le microphone est placé proche du lieu de l'événement. Nous avons donc considéré que le temps de propagation du son de l'événement jusqu'au microphone était négligeable. Nous détaillons dans les sections suivantes les divers protocoles expérimentaux utilisés permettant de trouver les délais entre l'audio et les autres modalités du système d'acquisition.

3.2.2.2 Délai entre les données audio et les données EM

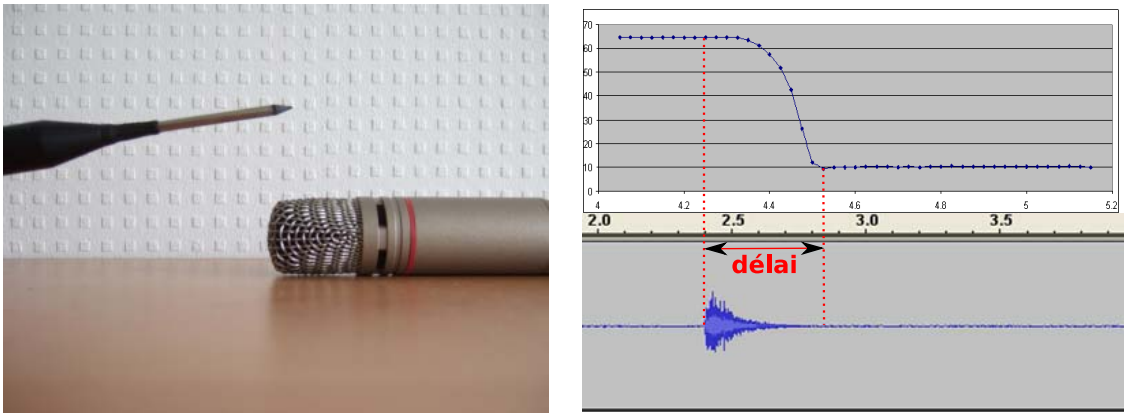
Le dispositif expérimental permettant d'avoir un événement identifiable dans les données audio et sur les données EM est décrit sur la figure 3.12. Le stylet EM vient frapper le microphone plusieurs fois et les données audio et EM sont enregistrées par le PC de contrôle. En mettant en correspondance les deux événements reçus par le PC de contrôle, on est en mesure d'estimer le délai entre les deux modalités (cf figure 3.12).

Pour 20 mesures effectuées, le délai constaté est de 67.8 ms avec un écart-type de 8.9 ms. L'écart-type est donc inférieur à la fréquence d'échantillonnage du système EM (40 Hz soit 25 ms). Cela signifie que le délai entre les deux modalités peut être considéré comme constant.

3.2.2.3 Délai entre les données audio et les données de stéréovision

Le dispositif expérimental permettant d'avoir un événement identifiable dans les données audio et sur les données de stéréovision est décrit sur la figure 3.13. Une balle de golf vient heurter le fond d'un réceptacle en plastique transparent. L'événement est enregistré par le microphone et par les caméras de stéréovision. En mettant en correspondance les deux événements reçus par le PC de contrôle, on est en mesure d'estimer le délai entre les deux modalités.

Pour 20 mesures effectuées, le délai constaté est de 9.78 ms avec un écart-type de 0.23 ms. Comme pour les capteurs EM, l'écart-type est largement inférieur à la fréquence d'échantillonnage



(a) (b)

FIG. 3.12 – (a) Dispositif de synchronisation des données EM avec l'audio. La sonde EM vient taper sur le microphone. (b) En haut, position en Y du stylet EM. En bas, données du fichier audio reçues par le PC de contrôle. Le délai entre les deux modalités est mesuré.

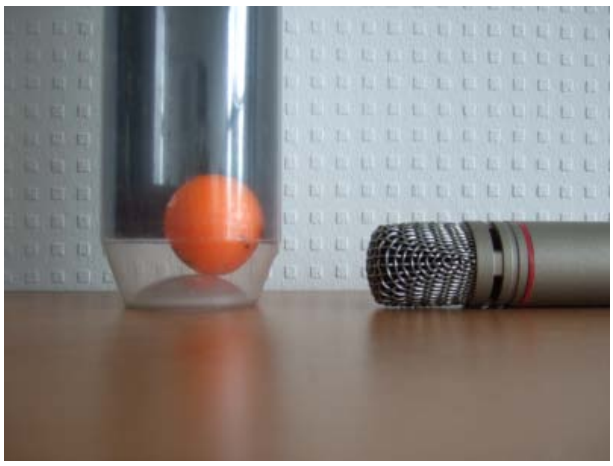


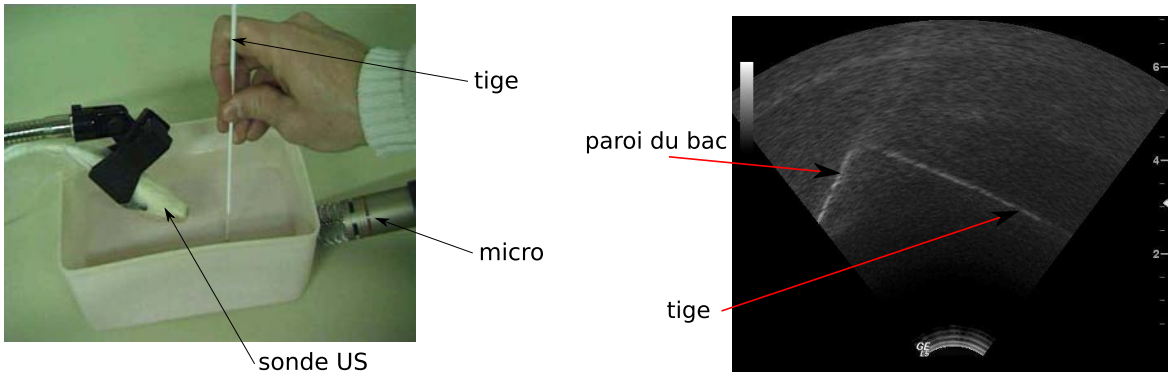
FIG. 3.13 – Dispositif de synchronisation des données de stéréovision avec l'audio. Lorsque la balle de golf heurte le fond, le son est enregistré par le microphone. L'événement est aussi filmé par les caméras de stéréovision.

des caméras (120 Hz soit 8.3 ms). Le délai entre les deux modalités peut donc être considéré comme constant.

3.2.2.4 Synchronisation avec l'échographe

Nous avons vu dans la présentation du système échographique que le processus d'enregistrement des données est effectué a posteriori en pressant un bouton dédié. Nous avons relié l'échographe au PC de contrôle via le port série, et simulé l'appui sur ce bouton par l'envoi d'un signal sur ce port. Cette opération permet de déclencher l'enregistrement sur l'échographe à partir du PC de contrôle.
Il reste à mesurer le délai entre l'envoi du signal du PC de contrôle et l'arrêt effectif de l'acquisition des images sur l'échographe. Pour cela, nous avons utilisé le protocole expérimental suivant :

une tige immergée dans un bac en plastique rempli d'eau à 50 ° C (cf chapitre 3) a été utilisée pour frapper la paroi du bac (cf figure 3.14). Ce choc est visible sur les images échographiques et peut être corrélé avec le son émis lorsque la baguette heurte la paroi. Une vidéo de l'expérience est accessible sur `http://www.loria.fr/~aron/these.html`.



(a) (b)

FIG. 3.14 – (a) Dispositif de synchronisation de l'US avec l'audio : la sonde US est immergée dans un bac d'eau chaude et la tige frappe le fond du bac. Le son est enregistré par le microphone. (b) La tige dans les images US.

Cette expérience a été répétée 20 fois. La moyenne de ce délai est de 14.9 ms avec un écart-type de 8.2 ms. Cette variabilité de 8.2 ms représente en terme d'images US deux images. Nous sommes donc en mesure d'assurer une synchronisation des données US avec les autres modalités à plus ou moins une image US près.

De plus, cette expérience nous a aussi permis de vérifier la fréquence d'acquisition des images échographiques en calculant d'après le fichier audio le temps écoulé entre deux images où le choc se produit. Pour une fréquence annoncée de 66 Hz, nous avons mesuré une fréquence expérimentale de 65.92 Hz en moyenne sur les 20 mesures avec un écart-type de 0.02 Hz. Les fréquences d'acquisition de l'échographe sont donc très stables d'après l'écart-type trouvé. Cette fréquence mesurée permet de corriger celle indiquée par l'échographe. Nous mettons ici en avant une fréquence calculée différente de celle indiquée par le constructeur et un problème pointé dans le chapitre 2 : au bout de 15 secondes d'acquisition US, cette différence dans la fréquence s'exprime par plus d'une image US. À supposer que l'on dispose d'un système permettant d'enregistrer des séquences plus longues, cette différence correspond à près de cinq images au bout d'une minute. Il est donc nécessaire de calculer cette valeur de fréquence pour synchroniser les données. Cette valeur calculée est utilisée dans la suite de ce manuscrit comme la valeur réelle de fréquence d'acquisition du système US.

Puisque le délai entre l'acquisition et l'enregistrement de la dernière image de la séquence US, et la valeur de la fréquence, ont été mesurés, on est alors en mesure d'étiqueter chaque image US par son temps d'acquisition, modulo l'incertitude de 8.2 ms mesurée.

3.2.3 Remarques sur la synchronisation

Lors des acquisitions, afin de ne pas perturber les données EM, le microphone enregistrant le locuteur n'est pas placé directement à proximité de sa bouche, mais en environ 50 cm. Si l'on considère que la vitesse du son dans l'air est de 343 m/s, cela revient à un temps de propagation

de la bouche au microphone de près de 1.5 ms. Pour une distance de un mètre, cela fait presque 3 ms, et deux mètres correspondent à 6 ms. Au delà d'un mètre, le délai ne devient donc plus négligeable par rapport à la fréquence d'acquisition des données de stéréovision et il faut donc bien veiller à disposer le microphone à moins d'un mètre du locuteur lors des acquisitions.

Nous avons proposé dans [AFK+07] une stratégie de synchronisation différente de celle décrite ici. Le système d'acquisition comportait alors deux PC, l'un pour enregistrer des données et l'autre pour commander les modalités. Les deux PC étaient synchronisés en leur faisant émettre des bips audio, permettant de mettre en correspondance les données temporelles des deux machines. Cette stratégie s'est simplifiée au cours de la thèse avec l'acquisition d'un matériel plus puissant permettant de centraliser commande et enregistrement, pour ainsi éliminer les bips audio.

3.3 Conclusion

Nous avons présenté l'architecture de notre système d'acquisition de données dynamiques. Les valeurs de résolution de chaque modalité ont été détaillées. Nous avons ensuite proposé une méthode pour synchroniser automatiquement toutes les données acquises à partir d'un PC de contrôle. Les délais d'acquisition de chaque modalité ont été calibrés pour être automatiquement corrigés dans les données acquises.

Les méthodes présentées sont simples à mettre en œuvre. Le système d'acquisition nécessite un seul manipulateur pour l'utiliser (pour le PC de contrôle). En pratique, un autre manipulateur est souvent nécessaire pour tenir la sonde US sous le menton du locuteur. En plus de le soulager de cette tâche, le locuteur peut se concentrer uniquement sur le corpus à prononcer. Le manipulateur lui tenant la sonde peut aussi se concentrer pour viser le plan médiosagittal avec la sonde US.

Contrairement aux systèmes d'acquisition de données articulatoires présentés dans le chapitre 2, nous avons pris soin de mesurer pour chaque modalité sa valeur de précision et/ou d'incertitude sans nous conformer aux mesures constructeurs. Nous avons également proposé des méthodes pour mesurer les délais d'acquisition entre les modalités, et synchroniser automatiquement les acquisitions. Les mesures trouvées prouvent que cette étape est absolument nécessaire lorsque l'on souhaite synchroniser des données provenant de systèmes d'acquisition différents. Cette remarque est d'autant plus vraie lorsque les fréquences d'acquisition nécessaires sont élevées. Notre méthode a d'ailleurs été reprise par Hueber [HCDS08] et le système Ouisper pour mesurer le délai entre des données échographiques et des données audio.

Chapitre 4

Traitement des données dynamiques

Ce chapitre présente dans un premier temps le recalage des données dynamiques US et EM par le calibrage de ces deux modalités. Après avoir décrit le principe théorique du calibrage, nous évoquons les principales méthodes existant à travers la littérature. Nous proposons ensuite un dispositif expérimental pour notre système.
Dans la seconde partie de ce chapitre, nous abordons le problème du suivi de la surface de la langue dans les images échographiques. Après un état de l'art sur les techniques de suivi utilisées dans des images échographiques, nous présentons notre méthode adaptée à la spécificité du suivi de la langue dans les images US.

4.1 Calibrage des données échographiques et électromagnétiques

4.1.1 Principe

L'objectif de ce travail est d'exprimer les données US et EM dans un même repère. On pourra ainsi connaître la position des capteurs EM fixés sur la langue par rapport aux images US, ou inversement retrouver la position de l'image US dans le repère EM. Nous le verrons dans ce manuscrit, cette étape est cruciale, car le calibrage EM/US sera utilisé pour le suivi de la langue dans les images échographiques (cf section 2), et aussi pour le recalage du système de données dynamiques avec les données statiques (cf chapitre 5).

Grâce à la synchronisation des données EM et des données US, nous sommes en mesure de temporellement lier les données des deux modalités. Le calibrage permet de les lier spatialement en calculant la transformation entre leurs deux repères.

Les images échographiques sont exprimées dans le repère $\mathcal{R}_{us}$ de la sonde US. Les données EM sont quant à elles exprimées dans le repère $\mathcal{R}_{em}$ du générateur de champ EM. Pour lier spatialement ces deux repères, il est nécessaire de connaître la position de la sonde US dans $\mathcal{R}_{em}$. Pour cela, il suffit de fixer un capteur EM à 6 DDL sur la sonde, donnant la transformation T_{em} entre le capteur et $\mathcal{R}_{em}$. Ce capteur définit un nouveau repère EM solidaire de la sonde. Il reste alors à calculer la transformation rigide (translation et rotation) T_c entre $\mathcal{R}_{us}$ et le repère du capteur EM fixé sur la sonde. Ce principe, le calibrage EM/US, est présenté sur la figure 4.1.

4.1.2 Formulation

Pour calibrer la transformation T_c, on utilise un fantôme de calibrage, qui possède des caractéristiques géométriques 3D connues dans son repère $\mathcal{R}_{fant}$. Ce fantôme permet d'imager des points 3D visibles à la fois dans $\mathcal{R}_{fant}$ et $\mathcal{R}_{us}$. Le calibrage consiste à identifier le point 3D dans les

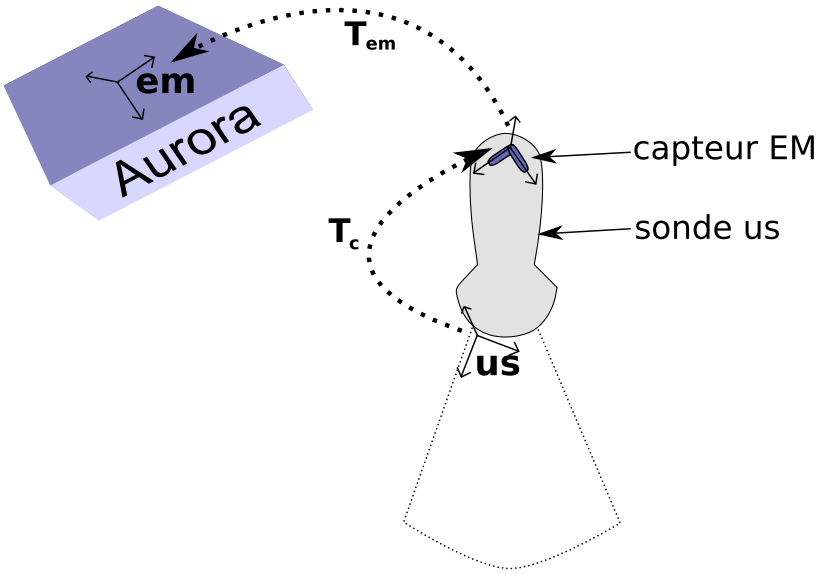


FIG. 4.1 – Principe du calibrage EM/US estimer la transformation T_c.

deux modalités, à apparier ses coordonnées dans les deux repères, pour estimer la transformation T_c. Géométriquement, il faut au moins trois appariements pour calculer cette transformation.

Le calibrage d'un système échographique avec un système de localisation est un problème abondamment traité dans la littérature [PRGB98, Rou03, MLLC05, PR05]. De nombreuses méthodes existent, toutes ayant pour objectif de mettre en place un fantôme de calibrage facilitant la détection et l'appariement de points 3D visibles à la fois dans le repère du fantôme et dans le repère US.

Prager [PRGB98] formule le problème du calibrage de la façon suivante[13] : soit en coordonnées homogènes un point 3D $P_{us} = (s_x.u, s_y.v, 0, 1)^T$ correspondant à un point 2D $p = (u, v, 1)^T$ de l'image US dans $\mathcal{R}_{us}$. Les termes s_x et s_y correspondent aux facteurs de résolution (en mm/pixel) de l'image US. Ce point s'exprime dans le repère $\mathcal{R}_{fant}$ du fantôme de calibrage par :

$$P_{fant} = T_{fant}.T_{em}.T_c.P_{us} \tag{4.1}$$

avec T_{em} la transformation rigide du capteur de localisation dans $\mathcal{R}_{em}$, T_{fant} la transformation rigide de $\mathcal{R}_{em}$ à $\mathcal{R}_{fant}$, et T_c la transformation rigide entre $\mathcal{R}_{us}$ et le capteur EM de la sonde US (cf figure 4.2).

T_{em} est connue, car elle est donnée par le capteur EM fixé à la sonde US. T_c, ainsi que les deux paramètres d'échelle s_x et s_y sont à estimer. T_{fant} peut être soit mesurée avec le système de localisation, soit être estimée.
En coordonnées homogènes, une transformation rigide s'écrit :

$$T = \begin{bmatrix} R(\alpha, \beta, \gamma) & t(t_x, t_y, t_z) \\ 0 & 1 \end{bmatrix} \tag{4.2}$$

où $t(t_x, t_y, t_z)$ est un vecteur de translation, α, β, γ les angles de rotation autour des axes du repère $\mathcal{R}_{us}$, et $R(\alpha, \beta, \gamma)$ la matrice de rotation associée. Une transformation rigide est déterminée par

[13] Pour nous rattacher à notre travail, nous prenons un système EM comme système de localisation, mais ce peut être aussi un système optique ou acoustique

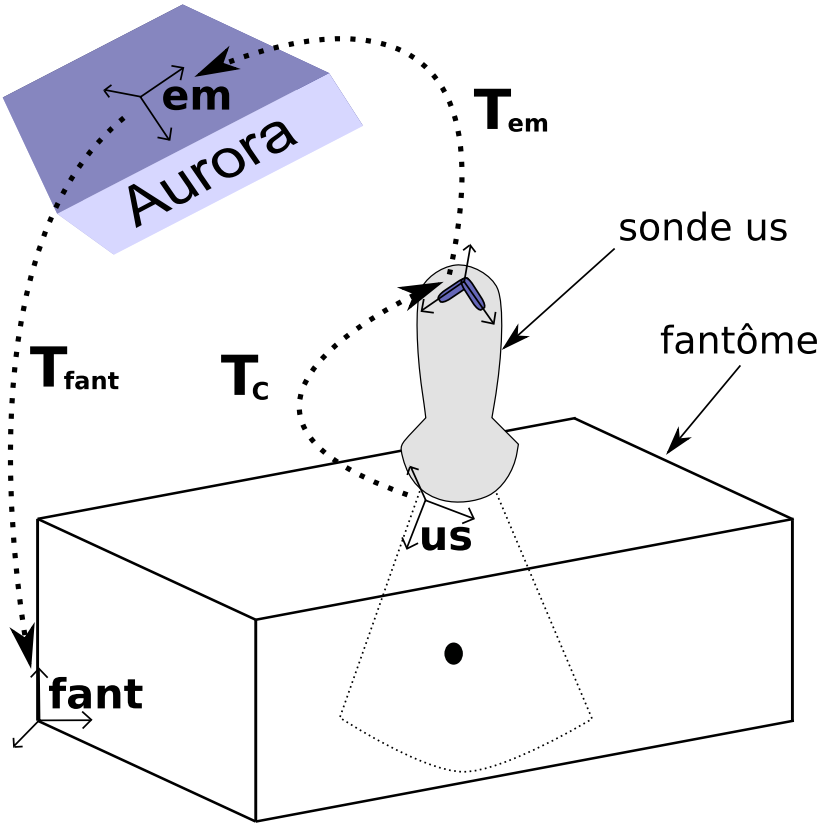


FIG. 4.2 – Principe du calibrage EM/US avec un fantôme.

six paramètres, trois pour la translation et trois pour la rotation. Le calibrage consiste consiste à retrouver les six paramètres de T_c, les deux paramètres d'échelle s_x et s_y, ainsi qu'éventuellement les six paramètres de la transformation T_{fant}, soit un total de quatorze paramètres.

4.1.3 Méthodes existantes

Nous présentons ici les fantômes couramment utilisés, en détaillant leurs principales caractéristiques. Pour plus de détails, nous invitons le lecteur à consulter les états de l'art sur les méthodes de calibrage des systèmes échographiques mains libres, comme celui de Mercier [MLLC05]. On trouve :

- fantôme de type point d'intersection (« cross-wire » en anglais),
- fantôme filaire,
- fantôme de type ensemble de points,
- fantôme plan,
- fantôme multimodal

4.1.3.1 Fantôme de type « point d'intersection »

Detmer [DBH+94] fut l'un des premiers à proposer un fantôme de calibrage pour un système échographique mains libres. Il utilise un système EM pour localiser la sonde US. Le fantôme est composé de deux fils qui se croisent en un point 3D, et qui sont plongés dans un bac d'eau. Ce point d'intersection est fixe dans l'espace 3D, et considéré comme l'origine du repère $\mathcal{R}_{fant}$ du

fantôme. Pour m positions et orientations de la sonde, le point est imagé. L'équation 4.1 s'écrit :

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = T_{fant}.T_{em}.T_c.P_{us} \tag{4.3}$$

Les trois premières lignes de l'équation 4.3 donnent trois équations impliquant les mesures T_{em} et P_{us} et les inconnues T_{fant}, T_c et s_x et s_y. Le repère $\mathcal{R}_{fant}$ a pour origine le point d'intersection, mais son orientation n'est pas fixée : elle peut être quelconque. Géométriquement, cela signifie que les trois paramètres de rotation de la transformation T_{fant} ne sont pas identifiables. Il sont fixés arbitrairement à zéros pour la résolution du système d'équation [PRGB98], ce qui laisse onze paramètres à identifier. Avec m mesures, on forme un système d'équations de taille $3m$, minimisé en utilisant une méthode de Powell ou de Levenberg-Marquadt [FTV93].

Pour étudier la précision de ce système, Prager [PRGB98] propose d'observer la variabilité de la position des points 3D reconstruits autour de la position moyenne 3D de cet ensemble de points. Il estime cette erreur à 1.65 mm.

Ce type de fantôme a largement été utilisé depuis ([AKJ+01]...) et cette méthode est devenue une méthode de référence. Cependant, les auteurs notent la difficulté de localiser le point d'intersection dans l'image US. En effet, il est difficile d'imager un point précis tout en faisant varier les positions et orientations de sonde US. On se retrouve fréquemment à imager le point avec des positions de sonde spatialement très proches.

4.1.3.2 Fantôme filaire

Carr [Car96] propose un fantôme constitué de trois fils orthogonaux. En scannant plusieurs fois les trois fils avec la sonde US, les paramètres de calibrage sont estimés en utilisant la contrainte d'orthogonalité des trois fils. En considérant le fil orienté suivant l'axe x, un pixel appartenant à ce fil vérifie l'équation :

$$\begin{pmatrix} x \\ 0 \\ 0 \\ 1 \end{pmatrix} = T_{fant}.T_{em}.T_c.P_{us} \tag{4.4}$$

L'opération est répétée le long des deux autres fils. De la même façon que pour un fantôme de type point d'intersection, en effectuant m mesures de T_{em} et P_{us}, les deux composantes nulles de l'équation 4.4 donnent un système de taille $2m$. Une minimisation aux moindres carrés permet de retrouver tous les paramètres de T_c et T_{fant} [PRGB98].

L'argument mis en avant par ses auteurs pour ce fantôme est qu'il est beaucoup plus facile, avec la sonde échographique, d'imager un fil qu'un point comme avec un fantôme de type point d'intersection. Cependant, la précision de la méthode repose sur la qualité de fabrication du fantôme et il n'est pas évident de disposer de 3 fils parfaitement orthogonaux.
Prager [PRGB98] estime l'erreur de localisation obtenue égale à 2.67 mm en utilisant le même procédé et le même système EM que celui décrit dans la section 4.1.3.1.

4.1.3.3 Ensemble de points

Les fantômes de type ensemble de points consistent en une structure complexe, dont le modèle 3D est connu. Le fantôme est localisé par le système de localisation dans un espace 3D. Les paramètres de calibrage sont estimés en mettant en correspondance les points extraits de l'image US avec les structures 3D du fantôme dont les positions sont connues. Comeau [CFP98] a été l'un des tout premiers a proposer un fantôme avec un motif avec des fils parallèles entre eux, formant un « Z » sur l'image échographique. L'avantage de ce type de fantôme est que les étapes de détection et d'appariements des points 3D peuvent se faire automatiquement, car on recherche dans l'image US un motif connu. Cependant, ces fantômes ont l'inconvénient majeur d'être très difficile à fabriquer.

Mercier [MLLC05] dans son état de l'art indique des erreurs moyennes de 1.52 mm pour ce type de fantôme. Cependant, ce résultat ne doit pas être comparé aux précédents annoncés pour les autres types de fantômes, car il a été obtenu avec un système de localisation optique, plus précis qu'un système électromagnétique. Il en donne cependant une idée.

4.1.3.4 Fantôme plan

Prager [PRGB98] a remarqué que le fond des fantômes de type point d'intersection ou filaire - le plus souvent des bacs remplis d'eau - formait dans l'image échographique une ligne bien visible. Il a utilisé cette ligne pour mettre en place un fantôme plan. Le repère du fantôme est choisi pour que l'axe z soit orthogonal au plan. Chaque point de l'image US vérifie une équation de type :

$$\begin{pmatrix} x \\ y \\ 0 \\ 1 \end{pmatrix} = T_{fant}.T_{em}.T_c.P_{us} \tag{4.5}$$

L'idée est séduisante, car, contrairement aux autres méthodes présentées, la fabrication du fantôme est très facile. La composante nulle de l'équation 4.5 donne une équation par mesure. On peut désigner sur chaque image US, deux points appartenant au plan (ce dernier apparaît sous la forme d'une ligne dans l'image US), ce qui donne deux équations par mesure.
Avec une telle méthode, les paramètres de translation x et y de T_{fant} et un de ses paramètres de rotation (autour de de l'axe x) ne sont pas identifiables [PRGB98]. Cela laisse onze paramètres à identifier pour T_{fant} et T_{em}. Comme avec les méthodes de point d'intersection ou filaire, ces paramètres sont retrouvés en minimisant un système de $2m$ équations.

Prager estime l'erreur de localisation obtenue égale à 3.43 mm en utilisant un système de localisation EM.

Il a ensuite mis au point un dispositif complexe pour améliorer sa méthode en fixant la sonde US sur un système à deux roues [PRGB98]. Ce dispositif a pour rôle de faciliter la détection des lignes dans les images US tout en faisant effectuer de nombreuses rotations et translations à la sonde US. Il obtient une précision de 2.17 mm. Ce fantôme a l'inconvénient d'être très complexe à construire.

4.1.3.5 Fantôme multimodal

Blackall [BRCM$^+$00] propose d'utiliser un modèle 3D préalablement scanné dans une autre modalité (IRM, CT. . . ...). L'estimation des paramètres de calibrage est alors effectuée par une technique de recalage entre les images US et le modèle 3D. Comme pour le fantôme avec ensemble

de points, cette technique facilite la détection et l'appariement des points des images US avec les points 3D du fantôme, car on recherche un motif connu. Blackall obtient une précision similaire avec un fantôme de type point d'intersection.

4.1.3.6 Récapitulatif

Les techniques de calibrage présentées reposent sur le principe suivant :

- un fantôme dont les propriétés géométriques 3D sont connues est imagé avec la sonde US,
- les paramètres de calibrage sont calculés en retrouvant cette géométrie dans les images US.

Il n'existe pas encore de consensus sur le fantôme idéal, chaque technique possède ses avantages et ses inconvénients. Les précisions obtenues avec les fantômes varient entre 1.52 mm pour le fantôme par ensemble de points et 3.43 mm pour le fantôme plan. Dans l'évaluation des méthodes de calibrage, Rousseau [Rou03] prend en compte d'autres critères comme le coût nécessaire à la mise en œuvre de la méthode de calibrage, la simplicité de construction du fantôme, son utilisation et le temps employé à l'utiliser. Sa conclusion est qu'il n'y a aucune méthode surpassant vraiment les autres.

Les auteurs de ces méthodes précisent aussi qu'il n'est pas évident d'imager un point avec un système échographique. En effet, en pratique, il est difficile d'obtenir des positions et orientations T_{em} différentes de la sonde US permettant d'obtenir un grand nombre m d'images où le point 3D est visible. Pour les méthodes utilisant un fil, on retrouve aussi cette difficulté à varier les positions et orientations de sondes différentes. Cela a pour effet de mal conditionner le système de $2m$ ou $3m$ équations du type de l'équation 4.1, et rend sensible l'estimation des quatorze paramètres inconnus (T_{fant}, T_c et s_x et s_y) [PRGB98].

Dans le problème du calibrage tel qu'il a été présenté ci-dessus, la transformation T_{fant} et les paramètres s_x et s_y sont estimés alors qu'ils ne concernent pas directement la transformation T_c recherchée. Une solution pour réduire le nombre de paramètres à estimer est de calculer a-priori T_{fant} et s_x et s_y afin d'effectuer le calcul de minimisation du calibrage seulement sur les paramètres de T_c. Pour cela, Khamene [KS05] propose une méthode permettant d'estimer d'abord la transformation T_{fant} indépendamment, et ensuite calculer les paramètres de T_c.

4.1.3.7 Méthode de Khamene

Khamene [KS05] propose de reprendre la méthode du fantôme filaire décrite en section 4.1.3.2, mais en fixant sur une tige un capteur EM. La position des deux extrémités de la tige dans le repère EM est d'abord pré-calibrée en utilisant une méthode de type « pivot ». Ensuite cette tige est utilisée comme fantôme pour le calibrage EM/US où elle est imagée pour plusieurs positions et orientations de sonde US. La position des deux extrémités de la tige étant connue dans le repère EM, la transformation T_{fant} de l'équation 4.1 ne fait plus partie des paramètres à estimer. L'équation de calibrage devient :

$$P_{em} = T_{em}.T_c.P_{us} \tag{4.6}$$

Il n'y a donc plus que huit paramètres inconnus : la résolution s_x et s_y ainsi que les six paramètres de la transformation T_c.
Les auteurs ne spécifient pas la précision obtenue sur leur système. On peut cependant penser que la précision est semblable est celle d'un fantôme filaire.

4.1.3.8 Conclusion

La méthode de Khamene [KS05] présente l'avantage d'être facile à mettre en œuvre, et de réduire le nombre de paramètres à estimer en ne cherchant que ceux qui concernent la transformation T_c. Dans notre système, nous avons fixé dans le chapitre 3 la résolution des images US grâce au fantôme CIRS Inc. permettant ce calcul. En utilisant une méthode reposant sur le principe de celle de Khamene, nous réduisons à six le nombre de paramètres à estimer dans l'équation 4.6.

Nous avons donc choisi cette méthode représentant un bon compromis entre facilité de fabrication du fantôme et précision attendue. Nous étudierons plus en détail la précision du système dans le chapitre 5 de ce manuscrit. Nous verrons aussi dans les perspectives du chapitre 6 que l'expertise acquise avec ce fantôme nous a permis d'effectuer le design d'un fantôme de type ensemble de points qui devrait améliorer la précision du calibrage EM/US.

Remarque La littérature fait état de vitesses de propagation du son différentes dans l'eau et dans les tissus humains (en moyenne 1540 m/s). Il apparaît en effet que le son se propage à des vitesses différentes suivant la température de l'eau. Bilaniuk [BW93] et Marczak [Mar97] ont établi la relation entre la température de l'eau et la vitesse de propagation du son (cf figure 4.3).

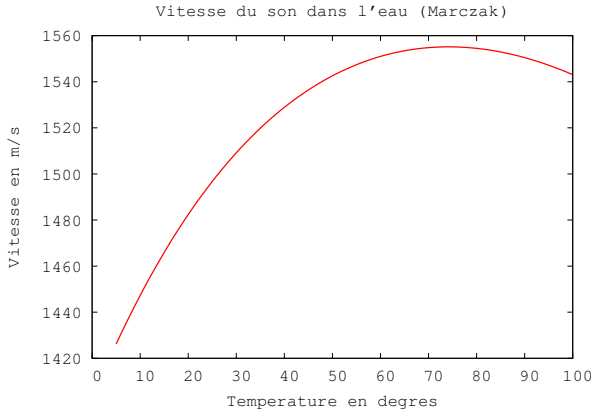


FIG. 4.3 – Vitesse de propagation des US dans l'eau en fonction de la température (d'après Marczak [Mar97]).

Ainsi dans de l'eau à 20 °C, le son se propage à une vitesse moyenne de 1482 m/s. La différence entre la vitesse de propagation du son dans l'eau et dans les tissus humains a pour effet d'introduire des distorsions dans l'image US [AMT00]. La solution la plus simple et que nous avons adoptée consiste à utiliser de l'eau chaude, à 50 °C [BJC+03], pour que le son se propage à 1540 m/s. Rousseau [Rou03] propose une solution à base d'eau et d'éthanol à 20 °C pour éviter d'avoir recours à de l'eau chaude.

4.1.4 Protocole expérimental

Pour fabriquer un fantôme de calibrage, nous avons repris le principe utilisé par Khamene [KS05]. Mais au lieu de devoir pré-calibrer la position de deux extrémités de la tige, nous avons directement utilisé deux capteurs EM fixés aux extrémités d'une tige en bois de 25 cm de longueur. Ainsi, l'équation de la droite définie par le fil dans le repère du système EM est connue, et elle peut facilement être imagée par la sonde US. Ce principe est détaillé sur la figure 4.4. Nous avons

utilisé une baguette en bois (3 mm de diamètre) pour ne pas perturber les mesures des capteurs EM (cf figure 4.5).

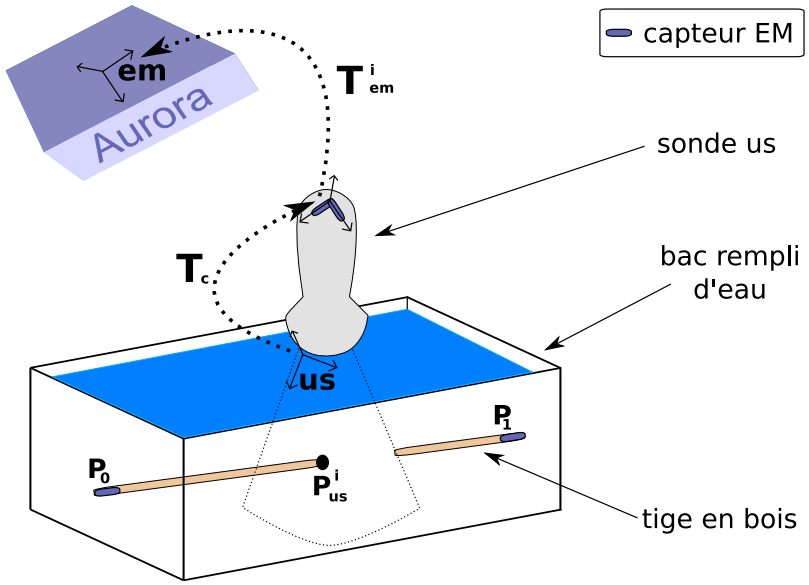


FIG. 4.4 – Protocole expérimental utilisé pour le calibrage EM/US : la sonde US image l'intersection du plan US avec une baguette en bois dont la position est connue dans le repère EM grâce à 2 capteurs EM fixés aux extrémités de la baguette. Plusieurs images sont acquises pour diverses positions de la sonde.

FIG. 4.5 – Photographie du fantôme utilisé pour le calibrage EM/US.

En notant P_0^i et P_1^i les coordonnées 3D dans le repère EM des deux points situé aux extrémités de la baguette pour une donnée EM i, le calibrage peut être défini comme la minimisation de la fonction de coût suivante :

$$\widetilde{T_c} = \underset{[T_c]}{\arg\min} \sum_i || \frac{(P_1^i - P_0^i)}{||P_1^i - P_0^i||} \times (T_{em}^i . T_c . P_{us}^i - P_0^i) ||^2 \tag{4.7}$$

où $\times$ représente le produit vectoriel. Ce dernier diminue lorsque la distance du point détecté dans l'image US à la droite (P_0P_1) diminue. Une méthode de minimisation de type Powell [FTV93] est utilisée pour résoudre l'équation 4.7, et retrouver les six paramètres de calibrage.

La minimisation de l'équation 4.7 fait intervenir un calcul de distance au sens des moindres carrés. Notre calibrage est soumis à des données aberrantes : mauvaise détection du point US dans l'image, donnée capteur EM aberrante... Pour rendre le calcul de la minimisation plus robuste, nous avons utilisé le M-estimateur développé par Huber [Hub81]. La technique consiste à remplacer le calcul du résidu avec les moindres carrés par un calcul de résidu faisant intervenir une fonction d'influence ρ. Pour un réel c, cette fonction d'influence (cf équation 4.8) s'écrit :

$$\rho(x) = \begin{cases} x^2/2 & \text{si} |x| \leq c \\ c(|x| - c/2) & \text{sinon} \end{cases} \tag{4.8}$$

La constante c représente le seuil à partir duquel l'influence de x est bornée. Cette valeur, fixée empiriquement, permet de seuiller l'influence de données aberrantes pour le calcul du résidu.

En pratique, 30 images US de la baguette ont été acquises, pour des orientations et des positions de sonde différentes. La figure 4.6 présente des images US obtenues avec notre fantôme. Sur ces deux images, la baguette est visible seulement partiellement (partie inférieure), et il appartient à un expérimentateur d'estimer la position du centre de la baguette. L'image 4.6.a montre la baguette proche (3 cm environ) de la sonde. Dans cette configuration, son centre peut être facilement estimé. En revanche, plus la sonde s'éloigne de la baguette, plus l'estimation est délicate. L'image 4.6.b présente ce cas de situation : à 6 cm de la sonde, le point est assimilable à une ellipse dont la largeur dépasse dix pixels sur l'image. Il est donc difficile d'automatiser cette étape de détection, qui ne peut être soumise qu'à une expertise humaine. Pour cette raison, les points d'intersection entre le plan US et la baguette dans les images US ont été manuellement détectés par deux expérimentateurs.

Notons que nous avons testé le fantôme avec un fil plus fin que la baguette en bois : fil à coudre, fils nylon de diamètres différents..., sans pour autant obtenir de différence dans l'image US résultante.

4.1.5 Résultats

4.1.5.1 Résultats numériques

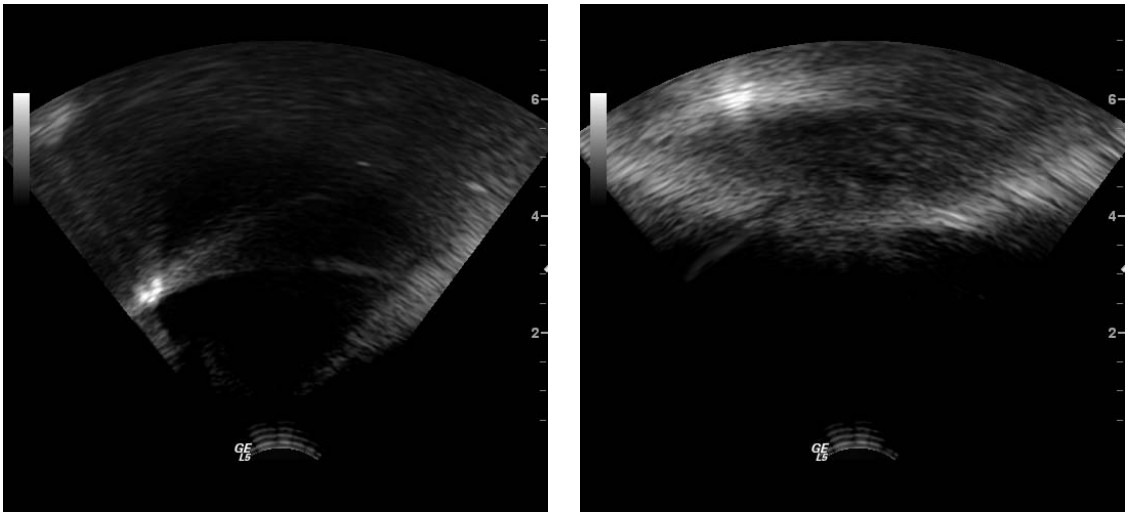
Le calibrage EM/US a été testé pour un réglage typique d'utilisation de l'échographe pour nos acquisitions de données articulatoires, c'est-à-dire avec une profondeur de champ de 7 cm, et une fréquence d'acquisition à 66 Hz.

Nous pouvons tout d'abord calculer l'erreur de pointage effectuée par les deux expérimentateurs pour la désignation du point d'intersection du plan US avec la baguette dans les images US. Sur les trente images, la distance moyenne entre les deux positions manuellement désignées est de 2.09 pixels, soit 0.37 mm. Cette distance est donc relativement faible compte tenu de la largeur possible du point dans l'image.

Dans l'équation 4.7, pour chaque donnée EM i $(1 \leq i \leq m)$ de T^i_{em}, de P^i_0 et P^i_1, nous faisons correspondre un point P^i_{us} manuellement désigné dans l'image US par un expérimentateur. Cette étape est répétée pour le second expérimentateur. Nous fabriquons ainsi un système de $2m$ équations qui est minimisé par la méthode de Powell [FTV93].

Le calcul normalisé du résidu de calibrage utilisant le M-estimateur de Huber (cf équation 4.8, c est fixé à 1 mm) donne :

$$E_{res} = 0.59\ mm \tag{4.9}$$



(a) (b)

FIG. 4.6 – Images US obtenues avec notre fantôme pour le calibrage EM/US. (a) Le centre de la baguette est facilement repérable dans l'image. (b) La baguette apparaît sur une largeur de 15 pixels et le point est très bruité : il est difficilement repérable dans l'image.

Cette valeur de résidu indique la distance moyenne des droites estimées ($\widetilde{P}P_0$) en utilisant la position $\widetilde{P}$ du point US calculé d'après la matrice de calibrage à la droite (P_0P_1) donnée par les capteurs EM (cf équation 4.7). Ce résidu est satisfaisant compte tenu des incertitudes présentes à la fois sur les données capteurs EM et sur le pointage du centre de la baguette dans les images US.

Remarque Nous pouvons aussi facilement dans la minimisation de l'équation 4.7 intégrer les paramètres de facteur d'échelle s_x et s_y. On obtient alors un résidu de calibrage égal à 0.45 mm et des résolutions pixelliques de 0.1687 mm/pixel sur X et 0.1774 mm/pixel sur Y. On retrouve bien le caractère anisotropique sur X et Y de la résolution US. Les valeurs calculées sont légèrement plus faibles que celles trouvées dans le chapitre 3. Nous gardons cependant les valeurs de résolution fixées avec le fantôme CIRS Inc. qui est spécialement dédié aux études de résolution.

4.1.5.2 Aide aux acquisitions dynamiques

Le calibrage EM/US permet de connaître la position du plan US dans le repère EM. Cette connaissance est utilisée pour faciliter les acquisitions dynamiques en permettant au manipulateur ayant la sonde US en main (locuteur lui-même ou une tierce personne) de visualiser en temps réel la position du plan US par rapport aux capteurs EM (langue, tête et sonde). Nous avons pour cela développé une application de visualisation dont le résultat est visible sur la figure 4.7. Une alarme (cercle rouge) s'affiche dans la fenêtre de visualisation si la distance d'un capteur au plan US franchit un seuil, fixé empiriquement à un centimètre. Cette application s'est révélée utile pour éviter au plan US de trop s'éloigner de la position des capteurs langue durant les acquisitions dynamiques, et de préserver l'alignement du plan US dans le plan médiosagittal.

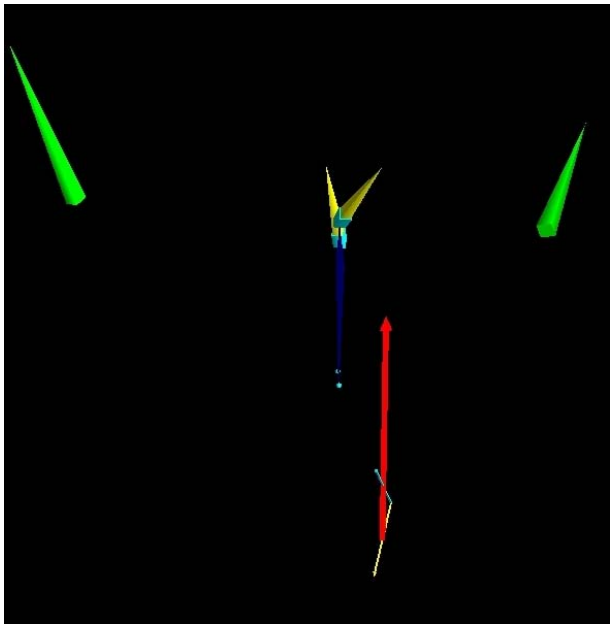
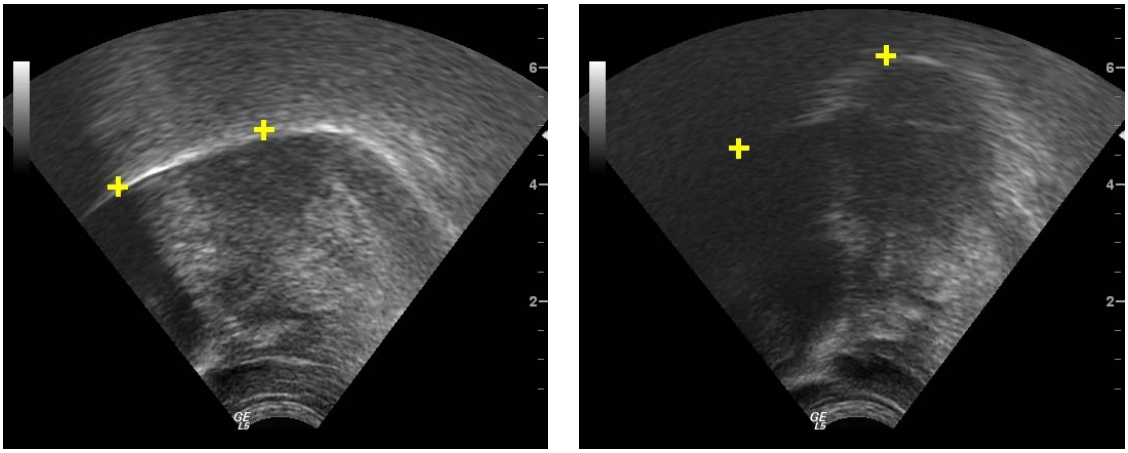


FIG. 4.7 – Interface de visualisation en temps réel de la position du plan US par rapport aux positions des capteurs EM. En vert, capteurs EM tête, en jaune, capteurs EM langue. En rouge, capteur EM sonde. En bleu : plan US.

4.1.5.3 Capteurs EM sur la langue

Les deux capteurs EM de la langue sont collés de façon à être les plus proches possible du plan médiosagittal, théoriquement visé par la sonde US. Grâce au calibrage EM/US et à la synchronisation des deux modalités US et EM (cf chapitre 3), on peut aussi projeter la position des capteurs EM de la langue (apex et dos) dans les images US. La figure 4.8 présente des résultats de cette projection.



(a) (b)

FIG. 4.8 – Images US de la langue avec les positions des capteurs EM. (a) /a/. (b) /ʃ/.

Nous avons vu dans le chapitre 3 que les fréquences d'acquisition entre les modalités EM et

US sont différentes. Pour un réglage typique d'utilisation (EM à 40 Hz et US à 66 Hz), il y a donc une donnée EM pour deux images US sur trois.

Cette étape permet aussi de vérifier que la synchronisation des deux modalités est correcte, en visualisant que les deux capteurs bougent temporellement en cohérence avec la surface de la langue.

Enfin, la figure 4.8.b permet de se rendre compte de l'utilité du capteur sur l'apex : un effet, cette zone de la langue n'est pas imagée à cause de l'air entre la langue et le plancher de la cavité buccale. Sa position peut être maintenant retrouvée grâce au capteur EM.

4.1.6 Conclusions

Nous avons présenté une méthode permettant de calculer la transformation entre les repères des modalités US et EM. Couplé à leur synchronisation, le calibrage de ces modalités permet de les fusionner temporellement et spatialement. On est ainsi en mesure d'afficher les positions des capteurs collés sur la langue dans les images échographiques pour notamment retrouver la position de l'apex.

La fréquence d'acquisition du système EM étant plus faible que celle du système US, nous disposons de données EM seulement pour deux images US sur trois. Nous verrons dans le chapitre 5 traitant du recalage multimodal qu'il est nécessaire de connaître la position et l'orientation de chaque image US acquise. L'amplitude des mouvements de la tête du locuteur et de la sonde US lors des acquisitions étant faible, la solution adoptée consiste à interpoler les données des capteurs EM de la tête et de la sonde pour avoir la position de chaque image US dans le repère EM. Le traitement effectué sur ces données EM consiste en une interpolation linéaire sur la translation, et une interpolation par la méthode « Spherical Linear intERPolation (SLERP) » [Sho85] pour la rotation. Ce traitement n'est pas effectué sur les données des capteurs de la langue qui par définition ont des mouvements brusques et rapides. On les affiche donc sur l'image US seulement lorsque l'information EM est présente.

Puisque nous utilisons le résultat du calibrage pour l'application de visualisation décrite en section 4.1.5.2, le calibrage EM/US doit être effectué en amont d'une séance d'acquisition de données dynamiques. Ensuite, tant que le capteur reste fixé sur la sonde US, il n'est pas nécessaire de répéter cette étape de calibrage.

Nous allons maintenant nous intéresser à autre facette du traitement des données dynamiques, à savoir l'extraction du contour de la langue dans les images échographiques.

4.2 Suivi du contour de la langue dans les séquences US

4.2.1 Spécificités du problème

Nous cherchons dans la seconde partie de ce chapitre à mettre en place une méthode pour extraire automatiquement le contour de la langue dans les images échographiques.

Avec la sonde placée sous le menton du locuteur, le contour de la langue est représenté dans l'image échographique par une bande de largeur variable (de 2 à plus de 20 pixels) selon son orientation, caractérisant la zone d'interface entre la langue et l'air (cf chapitre 1). Stone précise dans son guide d'utilisation de l'imagerie US pour la langue [Sto05] que la surface physique de langue correspond au contour inférieur de la bande visible dans les images US[14]. Nous cherchons

[14] À noter que nous avons cherché à vérifier cette hypothèse en recherchant un matériel échogène que nous aurions fixé sur la langue, et qui soit suffisamment visible à l'image US. Malgré nos recherches, nous n'avons pas trouvé un tel type de matériel.

donc à retrouver la position de ce contour pour toutes les images d'une séquence US, et donc à appliquer une méthode de suivi de ce contour à travers la séquence.

Certaines propriétés évidentes peuvent être énoncées sur ce contour : il est ouvert (la langue ne peut pas se replier sur elle-même et l'apex ne peut pas venir toucher l'arrière), libre à ses extrémités, continue, et lisse (il n'y a pas de point de rebroussement). Une autre caractéristique de la langue, et qui est déterminante pour le choix de la méthode de suivi, est qu'elle n'est pas rigide et qu'elle subit des déformations élastiques.
La langue a aussi des mouvements rapides d'une image à l'autre. Selon Perkell [Per69], elle peut atteindre la vitesse de 80 cm/s. Avec notre système US à 66 Hz et à une résolution de 0.17 mm/pixel, on obtient un mouvement de langue de 1.2 cm entre deux images, soit environ 70 pixels.
Une autre difficulté est la nature de l'image sur laquelle la langue est représentée. L'image US est en effet fortement sujette au bruit, le speckle (cf chapitre 1). De plus, les contours peuvent disparaître lorsque sa surface est proche de la verticale, comme pour un */u/* par exemple.

Cet ensemble de contraintes rend très spécifique la méthode de suivi à utiliser. Elle doit à la fois être capable de s'adapter à la langue et à ses déformations rapides tout en étant suffisamment robuste au bruit des images US.
Nous présentons une revue de techniques utilisées dans la littérature sur le suivi de courbes, majoritairement pour des applications différentes de la nôtre, car les travaux sur la langue sont rares. De nombreux travaux de suivi dans des images échographiques existent cependant notamment en cardiographie. Même si les objectifs sont différents (courbe fermée, mouvements relativement réguliers...), beaucoup de similarités existent avec notre travail.

4.2.2 Segmentation de courbes dans les images échographiques : le rôle des contours actifs

Le bruit présent dans les images échographiques est une difficulté majeure pour l'extraction et le suivi de contours. De nombreux travaux cherchent à le réduire en utilisant des filtres appropriés [YA02, Tau05] avant de procéder à la phase d'extraction, ce qui conduit à une amélioration relativement modeste de la qualité des images. La majorité des travaux utilisent donc des méthodes de type contours actifs [KWT88] pour extraire les contours d'intérêt dans les images. Un contour actif (ou snake) est une courbe $\mathcal{C}$ qui évolue sous l'influence d'une fonction d'énergie caractérisant les propriété de la courbe recherchée. La fonctionnelle à minimiser est caractérisée par la somme de deux termes d'énergie :

$$\phi(\mathcal{C}) = E_{img}(\mathcal{C}) + \lambda E_{int}(\mathcal{C}), \lambda \in \mathbb{R}^+ \tag{4.10}$$

E_{img} caractérise photométriquement la structure à mettre en évidence, un terme de type attraction par les gradients $E_{img} = -||\nabla I||$ étant fréquemment utilisé. Le terme E_{int} introduit des contraintes de lissage et d'élasticité sur les courbes à extraire via un terme le plus souvent du type $\alpha||v'||^2 + \beta||v''||^2$. v est une paramétrisation de la courbe $\mathcal{C}$. Le terme $|v'|^2$ influence la longueur de la courbe (on l'appelle la tension ou la rigidité de la courbe) et le terme $|v''|^2$ influence sa courbure (on l'appelle l'élasticité). Ces termes sont pondérés par les coefficients réels α et β choisis par l'utilisateur, et sont appelés les termes de régularisation du contour actif.

Les contours actifs, en introduisant un certain degré de connaissances a priori sur les formes, permettent d'éviter que la segmentation des contours soit trop perturbée par le bruit présent dans les images et sont d'usage très fréquent dans le traitement des données échographiques. La convergence du résultat dépend cependant de la proximité de la courbe initiale avec le contour à détecter. Pour éviter ces problèmes, des modèles paramétriques de contours actifs ont été

développés. Ils permettent de faire évoluer le contour actif à l'intérieur d'un espace de formes prédéfinies ce qui permet d'éviter d'obtenir des formes quelconques incompatibles avec l'application. Cette idée, déjà présente dans la technique de motifs déformables de Yuille [YCH92] a été étendue en particulier par Bascle et Deriche [Bas94] qui ont été parmi les pionniers de cette technique en définissant des snakes paramétriques. Ce concept a largement été utilisé depuis, une catégorie particulièrement intéressante de contours actifs étant les espaces de formes construits par le biais d'une analyse en composantes principales [CTCG95, JNB98, Lev00], lorsque de nombreux exemples des formes souhaitées sont disponibles.

Il est cependant souvent difficile de construire de tels espaces de formes, faute de disposer d'exemples en nombre suffisant. Par ailleurs, dans le domaine médical en particulier, certaines applications nécessitent de construire un modèle par patient pour atteindre la robustesse souhaitée, alors que d'autres applications ont besoin d'un modèle moins spécifique. Enfin, même si on dispose d'exemples et que l'on peut raisonnablement envisager une phase d'apprentissage, il est parfois difficile de mettre en correspondance ces exemples, ce qui est nécessaire à la construction du modèle, le modèle étant sensible à la présence d'erreurs dans la mise en correspondance [DCT01].

Les contours actifs qu'ils soient guidés par un modèle ou pas, ne peuvent cependant pas forcément être utilisés directement en suivi : le contour détecté dans une image fournit le plus souvent une initialisation inappropriée dans l'image suivante, car trop éloignée, dès que le mouvement apparent entre images est conséquent. La plupart des méthodes de suivi utilisent donc une phase de prédiction modélisant la dynamique des contours qui est suivie ou couplée à une phase de détection, qu'elle soit ou non basée sur les contours actifs.

4.2.3 Utilisation de la dynamique

Soit X^t le vecteur d'état représentant la courbe à suivre à l'instant t. Spécifier la dynamique revient à définir l'évolution de X_t au cours du temps par une fonction $X^{t+1} = f^t(X^t) + W^t$ où W^t est le bruit de prédiction à l'instant t, et dont la covariance est supposée connue.

Cette prédiction f peut être définie a priori en utilisant des fonctions classiques semblant raisonnables : des modèles de déplacement rigides, affines ou homographiques [Bas94, BWL99] ou des modèles autorégressifs [PHVG02] sont ainsi couramment utilisés. Dans le cas où les images sont particulièrement bruitées ou lorsque les mouvements sont très particuliers, il est utile de mieux préciser ce modèle. Le modèle de prédiction de la dynamique est alors acquis à la suite d'un apprentissage. Les travaux de Blake sont parmi les plus aboutis du domaine [CBZ92, IB96]. Cette phase d'apprentissage permet de contraindre les déformations possibles et permet ainsi d'accroître notablement la robustesse du processus. Des applications de ce principe existent en particulier dans le domaine du suivi cardiaque dans des images échographique [JNB98].

Historiquement, on a d'abord utilisé des méthodes en deux passes pour lesquelles la prédiction de la dynamique est faite via un mouvement paramétrique [Bas94, BWL99], cette phase étant suivie par une phase de convergence. Les méthodes de filtrage particulaire [IB96, PHVG02, AMP09] ont ensuite permis de considérer explicitement des fonctions f complexes, en particulier non linéaires, avec la possibilité de prendre en compte des déviations par rapport à ce modèle, c'est à dire un bruit de prédiction. Ces modèles de suivi probabilistes sont actuellement un champ de recherche très actif.

4.2.4 Travaux sur le suivi de la langue

Il existe peu de travaux sur le suivi de langue, les travaux les plus connus étant ceux de M. Stone. Dans [LKS03], elle propose d'utiliser un modèle de contours actifs pour le suivi de la langue dans les images échographiques. Des contraintes spécifiques sont utilisées dans les termes d'énergie. Pour l'énergie interne E_{int}, un terme classique de lissage est utilisé, ainsi qu'un terme de similarité entre le contour courant et le contour initial pour contraindre les déformations du contour. L'énergie externe utilise le terme classique de gradient pondéré par un terme de pénalité contraignant le contour à détecter la base inférieure de la zone blanche définissant le contour de la langue. Des contraintes de ce type directionnel ont déjà été utilisées par exemple dans [MDM99] dans le cadre de l'extraction du ventricule pour éviter la convergence vers un contour inadéquat. Une méthode de programmation dynamique [AWJ90] est appliquée pour minimiser l'énergie du contour sur trois images, la précédente, la courante, et la suivante et permettre une cohérence temporelle dans le suivi. Cette approche a été implémentée dans le logiciel public Edgetrak [LKS03]. Elle permet une bonne détection des contours de la langue dans les images US sous réserve d'une initialisation appropriée. Cependant, elle n'est pas bien adaptée à des mouvements rapides entre deux images, car aucune estimation du déplacement n'est effectuée. De plus, à l'utilisation, il semble que des conditions aux limites inadéquates soient utilisées, car le contour à tendance à « s'aplatir » dans l'image, avec les extrémités qui sont attirées par les bords de l'image.

Des améliorations de la méthode ont été proposées dans [LKS06] en proposant une approche de type *level set*. Cette méthode est inspirée de [Lev00] et utilise un modèle statistique de formes apprises sur un certain nombre d'images de la séquence à analyser. Bien que les idées développées dans ce papier soient très intéressantes, il est dommage que l'évaluation fournie dans l'article soit très incomplète : seul un petit nombre d'exemples sont présentés. Par ailleurs, la construction du modèle n'est pas explicitée. C'est pourtant un point délicat de la méthode, car la mise en correspondance des structures extraites des images échographiques en vue de construire un modèle statistique est difficile puisque la langue n'est pas vue dans son intégralité.

Fontecave [FB05] propose une méthode d'extraction des mouvements de la langue dans une base d'images rayons X par une méthode d'apprentissage. Pour cela, un ensemble d'images clés sont aléatoirement sélectionnées et manuellement détourées. Chaque image de la base est indexée par une image clé en calculant la similarité entre l'image courante et les images clés. La distance utilisée est basée sur les coefficients DCT (Discrete Cosine Transform) des images clés. En pratique, la similarité des images n'impliquant pas la similarité des courbes, plusieurs images clés sont retenues. La position des contours des images de la base est alors calculée en interpolant les contours des images clés. Cette méthode, testée sur des images cinéradiographiques, nécessite de devoir effectuer un détourage manuel de nombreuses images clés (une centaine) et l'erreur de suivi reste importante (environ douze pixels quand trois images clés sont utilisées pour l'indexation). Cette méthode ne semble pas avoir été testée sur des images échographiques.

4.2.5 Nos choix pour le suivi

Nous avons succinctement présenté dans cette section différentes méthodes de suivi de courbes utilisées notamment dans des séquences d'images échographiques. Cette étude a montré qu'un suivi robuste peut être atteint si l'on dispose de suffisamment de contraintes sur les formes admissibles et/ou sur la dynamique des courbes.

Nous avons choisi d'utiliser un modèle de type prédiction suivi par une phase de raffinement utilisant les contours actifs. Cependant, ne disposant pas a priori d'un modèle de langue pour

le locuteur considéré, nous avons choisi de développer une méthode sans modèle de déformation spécifique. Il repose sur un modèle affine du mouvement qui s'est avéré efficace pour de nombreuses séquences. L'estimation de la dynamique utilise le flot optique calculé entre les images et intègre également les contraintes délivrées par les capteurs EM. Ce suivi est présenté en détail dans la section suivante. Nous reviendrons sur ce problème du suivi dans le chapitre 6 et y discuterons de l'impact de l'introduction d'un modèle de formes.

4.2.6 Principe : suivi avec contraintes

4.2.6.1 Principe du suivi

Nous sommes partis d'une méthode préalablement développée dans notre laboratoire par Berger [BWL99] pour le suivi du ventricule gauche dans les séquences échocardiographiques. Cette méthode a ensuite a été adaptée à nos besoins. Comme dans les travaux d'Isard [IB96] et de Blake [BCZ93], l'idée générale est de restreindre le mouvement de la forme à suivre entre deux images à un ensemble possible de mouvements paramétriques (mouvement rigide, similitude, affine...) décrivant au mieux l'évolution de la forme. Restreindre la nature du mouvement permet d'être plus robuste vis-à-vis du bruit présent dans les images US. Ensuite une approche par contours actifs est utilisée pour affiner la position du contour dans l'image.

Le contour à suivre est manuellement défini dans la première image de la séquence. Ensuite le suivi effectue automatiquement les deux étapes suivantes :

- estimation du déplacement de la forme à suivre entre deux images en utilisant la contrainte de mouvement,
- raffinement de la position de la forme dans l'image en utilisant la méthode des contours actifs.

La seconde étape permet de prendre en compte les déformations élastiques de la langue et ainsi d'adapter le contour à la forme dans la nouvelle image. Nous détaillons dans les paragraphes suivants ces deux étapes utilisées pour le suivi, et résumées sur la figure 4.9.

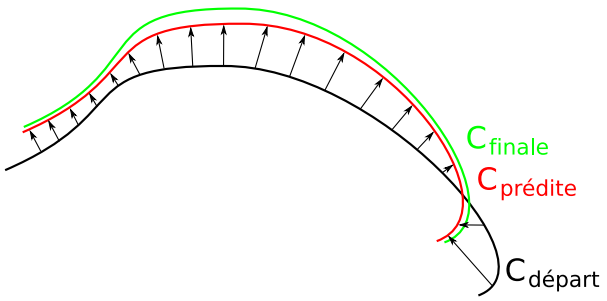


FIG. 4.9 – Principe du suivi : à partir d'une courbe initiale C_{depart} (noir), sa position est prédite en une courbe courbe $C_{predite}$ (rouge) d'après le calcul de l'estimation de mouvement (flèches noires). La courbe C_{finale} (vert) est la position de la courbe cherchée et est retrouvée grâce à la seconde étape du calcul.

Estimation itérative du déplacement L'estimation du déplacement 2D entre deux images est basée sur un calcul itératif de la composante normale du flot optique [HS80], qui est la seule composante pouvant être estimée de manière fiable (problème « d'ouverture »), contraint à un mouvement paramétrique D. Soit $\mathcal{C}_d$ le contour initial, l'estimation du déplacement est effectuée

seulement sur les points de ce contour à suivre, et de telle façon que les intensités sur les points de $\mathcal{C}_d$ et sur les points de $D(\mathcal{C}_d)$ soient similaires.

Le mouvement paramétrique est contraint ici à une transformation affine. Cette estimation, même si elle est grossière, s'est avérée suffisante puisque l'étape suivante permet d'affiner la forme de la courbe au contour recherché dans l'image.

Soit $(M_i)_{0\leq i\leq N}$ les points du contour $\mathcal{C}_d$, $f_0^{\perp}(M_i)$ la composante normale du flot optique sur les points (M_i) et n_i le vecteur unitaire, normal à la courbe $\mathcal{C}_d$ en M_i. Soit $\mathcal{C}_p$ la courbe finale à atteindre. Le déplacement 2D D_0 minimisant

$$\sum_{0\leq i\leq N} |(M_i D_0(M_i).n_i)n_i - f_0^{\perp}(M_i)|^2 \tag{4.11}$$

est une première estimation grossière du déplacement des points M du contour $\mathcal{C}_d$. La courbe $D_0(\mathcal{C}_d)$ est donc plus proche de la courbe $\mathcal{C}_p$ et l'estimation du déplacement est affinée en calculant la composante normale du flot optique $f_1^{\perp}$ sur $D_0(\mathcal{C}_d)$. En effet, le calcul du flot optique repose sur une hypothèse différentielle de petit mouvement. Plus l'écart entre la courbe estimée et la courbe à atteindre est faible, plus l'estimation du flot sera fiable. Ceci explique la convergence de cette estimation itérative. La composition des déplacements infinitésimaux D_0,...,D_j,... permettent de calculer la courbe $\mathcal{C}_j = D_j \circ \ldots \circ D_0(\mathcal{C})$ qui converge vers la courbe $\mathcal{C}_p$.

Raffinement en utilisant les contours actifs Une fois le déplacement estimé, la méthode classique des contours actifs, décrite en section 4.2.2, est utilisée pour attirer la courbe vers la zone de fort gradient, c'est-à-dire la zone d'interface entre la surface de la langue et l'air dans les images US.

4.2.6.2 Ajout de contraintes

Nous avons adapté le suivi à notre application spécifique. Grâce au calibrage EM/US et à la synchronisation de ces deux modalités, nous disposons des données de position du plan US dans le repère EM et de la position de capteurs EM projetés dans les images US. Ces informations peuvent être facilement intégrées au suivi sous la forme de contraintes afin d'en améliorer la qualité.

Correction des mouvements de la sonde US Nous avons remarqué lors de l'acquisition d'une séquence d'images US que la sonde bouge en fonction de l'articulation. La position de la sonde dans la première image dévie et se retrouve à imager une zone plus en arrière ou plus en avant de la langue. Afin de corriger ces mouvements de sonde, nous appliquons le mouvement de la sonde US à la courbe du suivi lors de l'initialisation du contour à chaque image de la séquence (cf figure 4.10).

Contraintes aux limites Dans les méthodes de type contours actifs, il est important de définir des contraintes aux limites appropriées surtout dans le cas de courbes ouvertes. En effet, les contours actifs ont naturellement tendance à se rétracter. Pour éviter cela, nous avons défini des contraintes aux frontières de l'image.

Nous avons tout d'abord contraint les deux extrémités du contour à rester sur les deux segments de droite définis sur la figure 4.11. Ces deux segments partent de l'origine de la sonde et passent par les extrémités du premier contour manuellement défini pour le suivi. Nous appellerons par la suite cette contrainte « apex et arrière frontières ».

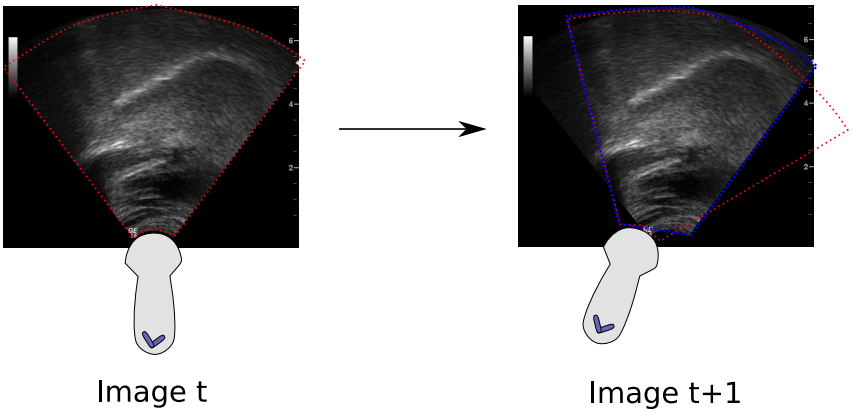


FIG. 4.10 – Correction des mouvements de la sonde US pour le suivi : en rouge, la position de la 1ère image. En bleu, la zone de recherche dans l'image courante.

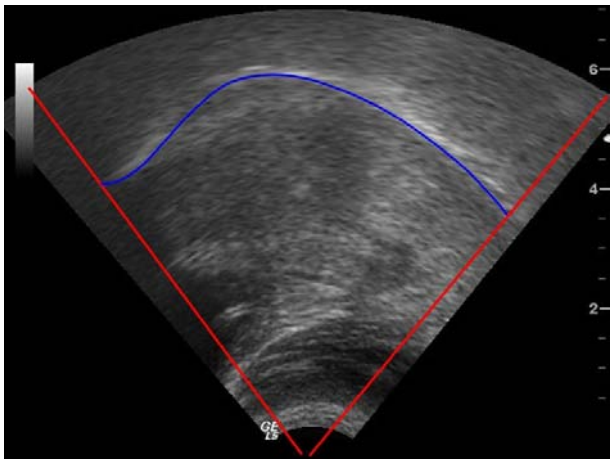


FIG. 4.11 – Frontières (en rouge) du suivi (en bleu).

Mais cette contrainte n'est pas physiologiquement valide si elle appliquée telle quelle. En effet, pour une image comme celle de la figure 4.11 l'arrière de la langue coupe systématiquement la frontière droite de l'image. En revanche, la langue n'a pas systématiquement une intersection avec la frontière gauche : lors d'un /o/ par exemple, l'apex se retrouve vers le centre haut de l'image US. Nous avons donc décidé d'utiliser la position du capteur EM situé sur l'apex pour redéfinir la position de cette frontière à chaque fois qu'une donnée EM est disponible pour l'image US. L'apex est contraint à se déplacer sur le segment de droite défini par l'origine de la sonde et la position du capteur EM de l'apex. Lorsque l'information capteur n'est pas présente, nous contraignons le contour à rester sur cette même ligne. Puisqu'une information capteur est disponible environ 2 images sur 3, nous avons considéré que cette information était suffisante. Nous appellerons par la suite cette contrainte « apex EM et arrière frontière ».

Techniquement, il est numériquement possible de calculer un contour actif à extrémités fermées, libres ou fixes [Ber91]. Il est facile de contraindre les extrémités à rester sur la verticale (« $x =$ **constant** ») ou sur l'horizontale (« $y =$ **constant** »), mais beaucoup plus difficile de numériquement contraindre les extrémiter du contour actif à rester sur une droite de type $y = ax+b$, comme c'est le cas avec les frontières de l'image US. Nous avons donc choisi de « déplier » l'image

US en l'exprimant dans un système de coordonnées polaires, dont le centre est l'intersection des frontières de l'image US et le rayon est donné par la distance du centre au sommet de l'image (cf figure 4.12). Une fois l'image US dépliée, nous avons ajouté la contrainte pour que les extrémités du contour actif soient libres sur l'axe vertical et fixes sur l'axe horizontal (« $x =$ **constant** et $y =$ **libre** »). Nous calculons ensuite le contour actif sur l'image gradient de cette image dépliée et exprimons enfin les points du contour obtenu dans le système de coordonnées original.

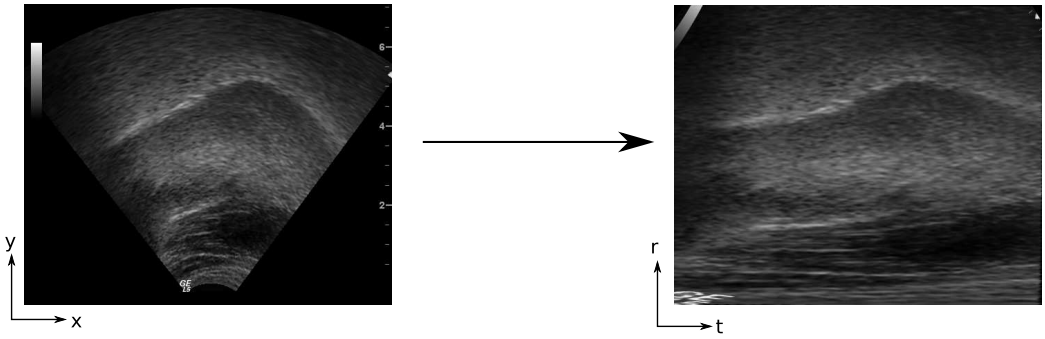


FIG. 4.12 – Dépliement de l'image US.

Utilisation des capteurs EM L'étape d'estimation du mouvement affine par le calcul du flot optique dans le suivi est dépendante du bruit dans les images échographiques. Nous l'avons vu dans le chapitre 3, il arrive pour certaines positions de langue qu'une partie de la surface de la langue ne soit plus visible. Ce manque d'information dans les images US a pour effet de ne pas engendrer une bonne estimation du mouvement. Afin d'aider le suivi à retrouver le contour correct, les projections des capteurs EM dans les images US sont utilisées avant le calcul par contours actifs. La courbe est initialisée en passant par la position des capteurs fixés sur la langue (cf figure 4.13). Cela permet d'aider le contour retrouver le contour de langue, tout en le laissant libre de se déplacer lors du calcul par les contours actifs (cf figure 4.14).
Nous aurions aussi pu choisir de contraindre le contour actif à passer par les positions capteurs. Mais il arrive que les capteurs donnent des mesures aberrantes et ne soient pas projetés là où ils devraient l'être. Notre choix s'est donc porté sur la solution intermédiaire entre laisser le contour totalement libre et le contraindre à passer par des positions pouvant être fausses.

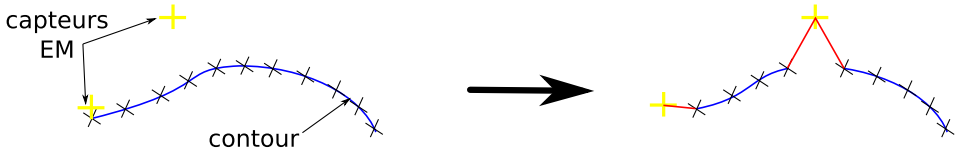


FIG. 4.13 – Ré-initialisation de la position du contour avant le calcul du raffinement de sa position par les contours actifs.

4.2.7 Résultats

Afin d'évaluer notre méthode de suivi, nous avons choisi une séquence composée de quatre groupes de phonème VV : */ae/, /ai/, /ao,/ /ay/*. Elle comporte 390 images (environ 6 secondes)

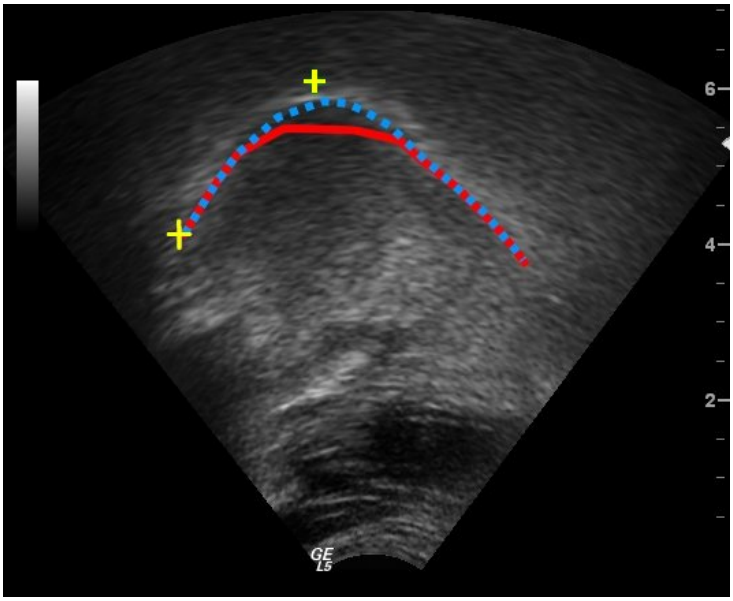


FIG. 4.14 – Suivi en utilisant les positions des capteurs EM : les deux courbes représentent un résultat de suivi après les deux étapes d'estimation et de raffinement. La courbe rouge est le contour obtenu sans utiliser les positions des capteurs EM et la courbe bleue en pointillées est le contour obtenu en utilisant les positions des capteurs EM.

et des données EM acquises avec notre système.
Sur les deux premiers groupes de phonèmes (*/ae/, /ai/*, 200 images), les contours de langue sont bien visibles dans les images US. Sur le second groupe (*/ao/, /ay/*, 190 images), les contours sont moins perceptibles sur les images, car le mouvement est rapide et des parties de la surface de la langue sont proches de la verticale.

La figure 4.15 présente le résultat du suivi appliqué avec succès sur cette séquence, pour 50 itérations dans le calcul de l'estimation du déplacement (cf équation 4.11) et des valeurs de α et β fixés à 1 pour le calcul par contours actifs. La vidéo complète est consultable sur `http://www.loria.fr/~aron/these.html`.

4.2.7.1 Étude comparative

Nous avons testé différentes méthodes de suivi : une méthode existante (EdgeTrak [LKS03]), et la méthode de suivi avec les diverses contraintes énoncées dans la section 4.2.6.2. Cela donne pour les méthodes testées :

- méthode 1 : EdgeTrak
- méthode 2 : contrainte « apex et arrière frontières » ;
- méthode 3 : contrainte « apex EM et arrière frontière ».

La contrainte de correction du mouvement de la sonde est utile seulement lorsqu'il y a un large mouvement de la sonde US durant les acquisitions. Elle n'apporte pas d'améliorations sur la séquence étudiée.

Ces trois méthodes ont été comparées à un suivi effectué manuellement image par image et considéré comme le contour de référence. L'erreur est calculée en effectuant la somme des distances (en millimètres) des points de la courbe à comparer à la courbe de référence pour

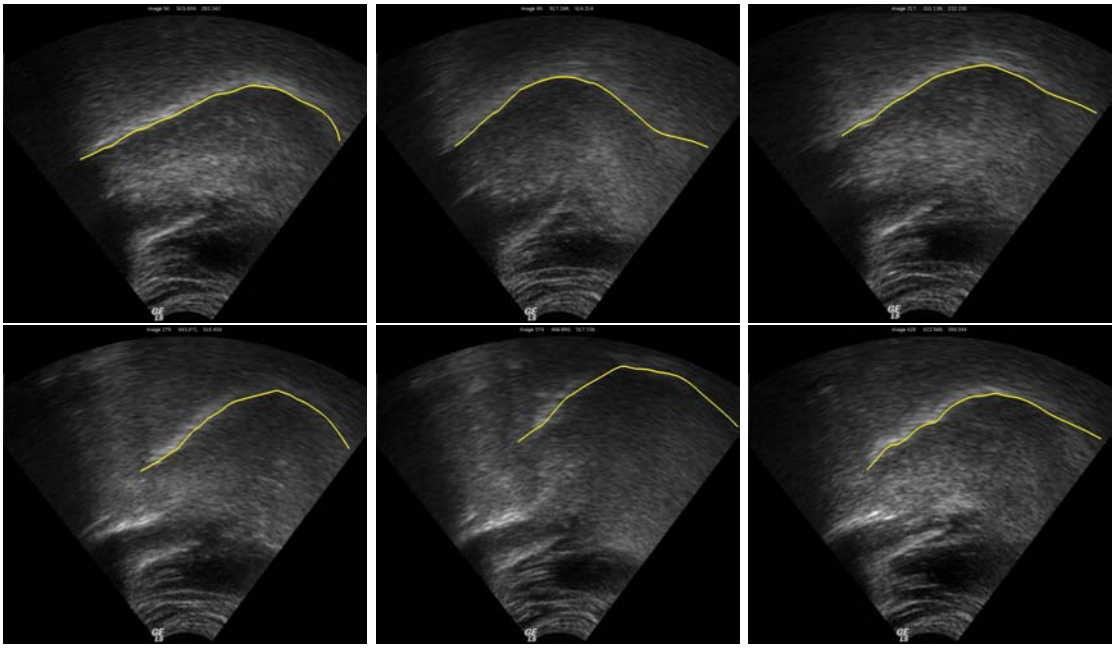


FIG. 4.15 – Résultat du suivi pour six images de la séquence */ae/ /ai/ /ao/ /ay/*. Son prononcé pour les images présentées (de gauche à droite, de haut en bas) : */a/* de */ae/*, */e/*, */a/* de */ao/*, */o/*, */y/* (début), */y/* (fin).

chacune des images. Ensuite la moyenne de ces distances est calculée pour connaître l'erreur moyenne d'une méthode de suivi par rapport au tracé manuel. Nous indiquons aussi le pourcentage d'images de la séquence pour lesquelles l'erreur est inférieure à 2 mm. Nous avons en effet constaté qu'au-delà de cette valeur, le suivi semblait visuellement faux. Les résultats sont présentés dans le tableau 4.1.

Méthode #	1	2	3
Erreur moyenne (mm)	1.36	1.18	1.34
Écart-type (mm)	0.58	0.51	0.65
% d'images avec une erreur > 2 mm	17	9.5	14.514

TAB. 4.1 – Résultats des différentes contraintes appliquées au suivi et testées sur le groupe de phonèmes */ae/* et */ai/* (200 images - 3 sec).

Toutes les méthodes ont une erreur similaire, comprise entre 1.18 mm et 1.36 mm. Cette séquence ne présente pas de difficultés majeures pour le suivi car les contours sont bien visibles dans les images.

Les méthodes ont ensuite été testées sur le groupe de phonèmes */ao/ /ay/*, pour lesquelles la langue a une dynamique plus importante.

Pour ces deux groupes de phonèmes où les mouvements de langue sont plus rapides, on peut observer sur le tableau 4.2 les apports de notre méthode de suivi et des contraintes sur la qualité du suivi. Alors qu'avec EdgeTrak, l'erreur moyenne dépasse les 5 mm sur cette séquence, elle

Méthode #	1	2	3
Erreur moyenne (mm)	5.68	1.83	1.79
Écart-type (mm)	2.57	0.51	0.56
% d'images avec une erreur > 2 mm	93.2	35.3	34.2

TAB. 4.2 – Résultats des différentes contraintes appliquée au suivi et testées sur le groupe de phonèmes */aɔ/* et */aj/* (190 images - 2.9 sec).

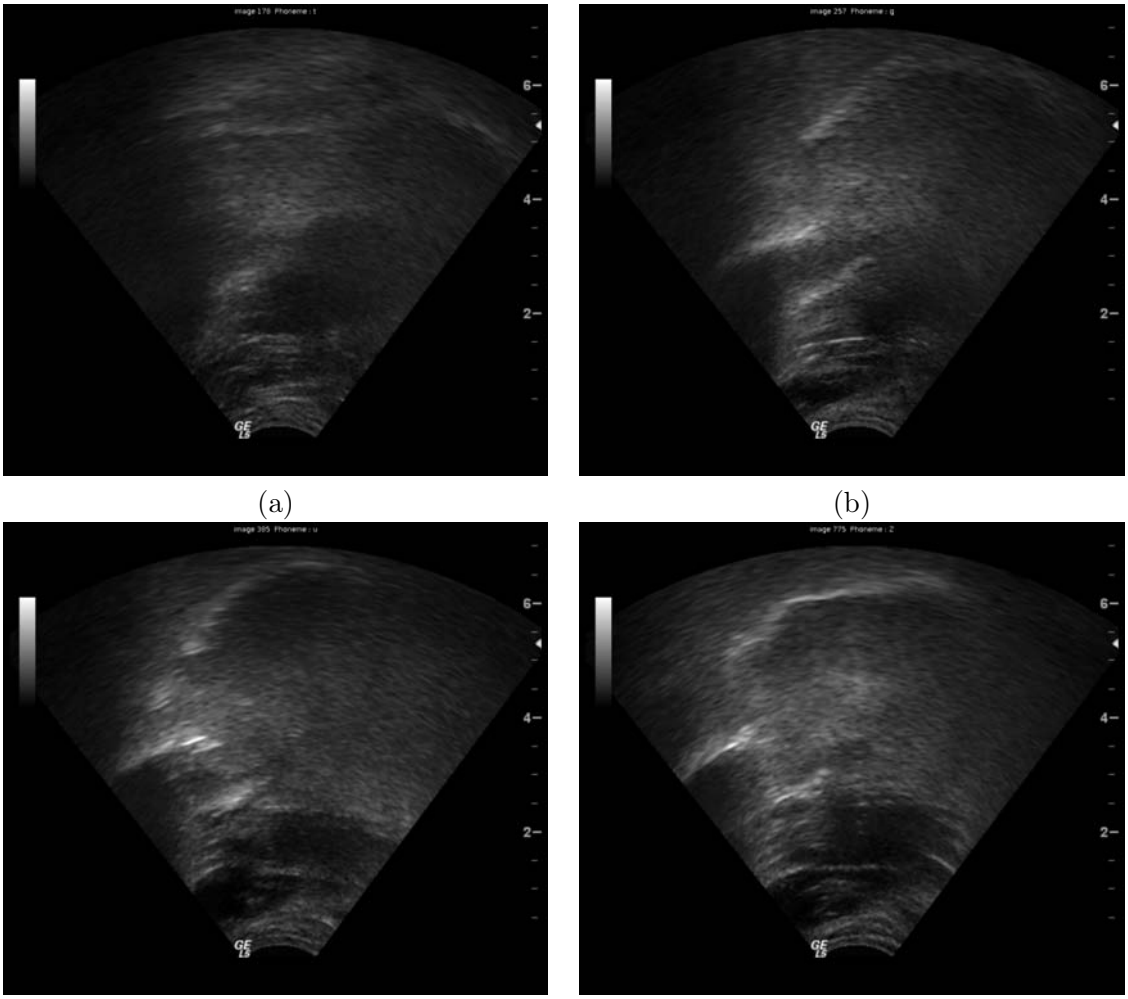
reste inférieure à 2 mm avec notre méthode. L'utilisation de la contrainte « apex EM et arrière frontière »permet de légèrement affiner la précision du suivi sur cette séquence. Malgré la prédiction du mouvement dans notre suivi et l'utilisation de contraintes liées aux capteurs EM, plus de 30% des images ont une erreur supérieure à 2 mm. La figure 4.16 présente des exemples typiques d'image US où l'arrière de la langue n'est pas visible. Ces images ne contiennent pas suffisamment d'information a priori pour inférer un contour correct. Suivant les termes de régularisation utilisés dans le calcul du snake, ou suivant le lissage effectué sur l'image, on obtient des contours différents à cause de ce manque d'information. Il est aussi très difficile pour un utilisateur de visuellement évaluer la position de ce contour dans l'image. La visualisation de la séquence en dynamique permet souvent de manuellement estimer la position du contour, même si elle reste très grossière et soumise à la subjectivité d'un utilisateur.

Lors de nos acquisitions de données dynamiques, nous sommes souvent confrontés à ce genre de situation. Il est très difficile de quantifier le nombre d'images US présentant cette configuration, car de nombreux paramètres interviennent : échogénicité du sujet, position de la sonde lors de l'acquisition, son prononcé... Idéalement un capteur EM dans cette zone permettrait d'introduire une nouvelle contrainte pour retrouver la position de la langue. Mais il n'est pas envisageable de coller un capteur à cet endroit, qui de plus gênerait davantage le locuteur pour la phonation. Une solution à envisager serait d'utiliser un modèle de déformations de la langue dans le suivi pour contraindre davantage la forme du contour. Nous discuterons cette possibilité dans la partie perspective du chapitre 6.

Afin de pouvoir rapidement traiter l'important volume de séquences acquises avec notre système, et de manuellement corriger le suivi lorsque ce dernier échoue, nous avons développé une interface de contrôle permettant à l'utilisateur de contrôler et éventuellement corriger le suivi. Nous présentons dans la partie suivante cette interface.

4.2.8 Interface de suivi

Une interface de visualisation et de correction du suivi a été développée et est présentée sur la figure 4.17. Elle est constituée d'une fenêtre de commande et d'une fenêtre de visualisation.
La fenêtre de commande permet à l'utilisateur de fixer les paramètres comme le nombre d'itérations utilisées pour le calcul du flot optique, le type de déplacement choisi (rigide, similitude, affine), les constantes de rigidité d'élasticité du contour actif, et les contraintes à utiliser pour le calcul (utilisation des capteurs EM, comportement des extrémités du contour...).
La fenêtre de visualisation affiche l'image et le résultat du calcul en temps réel. Il est possible d'interrompre le calcul en cours pour manuellement éditer la courbe et relancer le calcul.



(a) (b)

(c) (d)

FIG. 4.16 – Exemples typiques d'image US pour lesquelles le suivi échoue. (a) /t/. (b) /g/. (c) /u/. (d) /ʒ/.

4.3 Conclusion

Nous avons présenté deux méthodes pour le traitement automatique des données US et EM. La première consiste à calibrer la position du capteur EM sur la sonde US afin de connaître la transformation rigide liant les repères électromagnétiques et échographiques. Après avoir brièvement présenté les diverses méthodes existantes et les difficultés à les utiliser, nous avons opté pour l'une d'entre elles qui s'avère simple et rapide à mettre en œuvre. Nous avons mis en exergue et quantifié les imprécisions de ce calibrage, notamment la résolution US. Nous verrons dans le chapitre 5 que cette imprécision à une forte influence sur l'incertitude dans le recalage global de toutes les modalités utilisées dans le système d'acquisition de données articulatoires.

Nous avons ensuite présenté une méthode de suivi des contours de langue dans les images échographiques. Nous avons contraint cette méthode en utilisant notamment les capteurs EM qui représentent une aide substantielle au suivi. L'erreur du suivi, quantifiée sur une séquence, est estimée à 1.5 mm en moyenne. Une interface de contrôle de suivi a été développée afin de

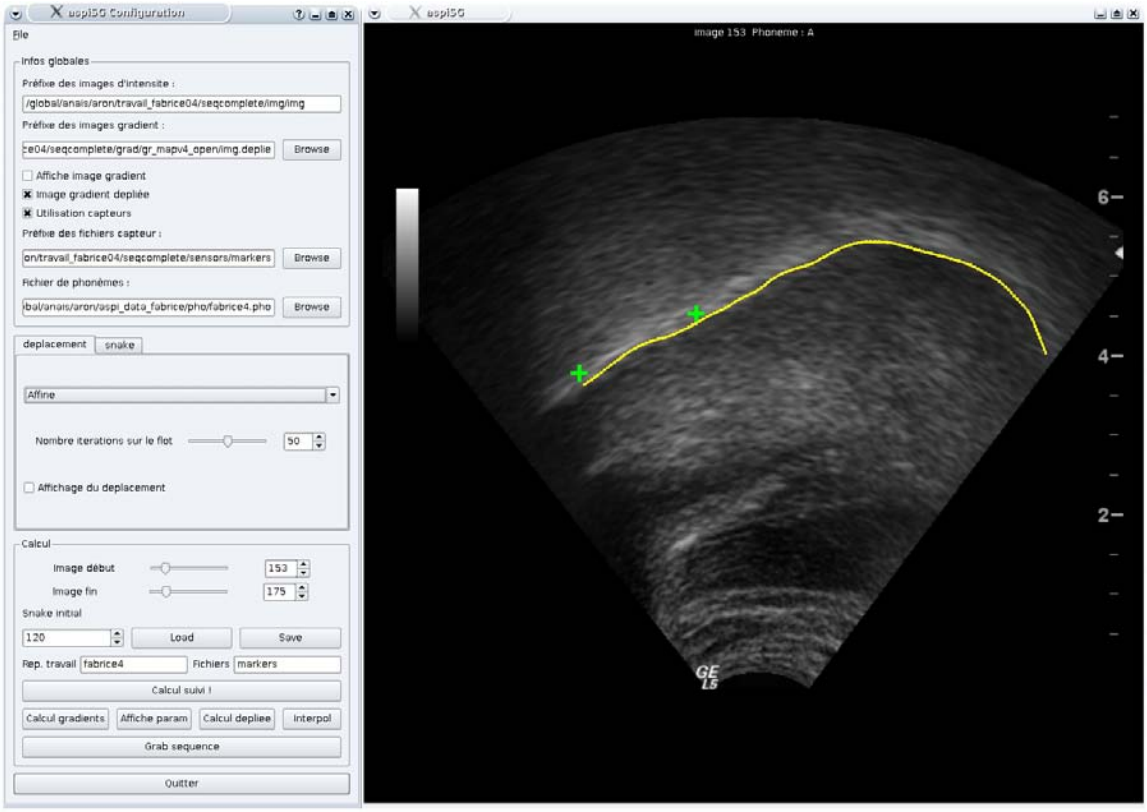


FIG. 4.17 – Interface pour le suivi.

manuellement corriger les images pour lesquelles le contour obtenu avec le suivi ne semble pas correct, et traiter ainsi facilement un important volume de données.

Nous disposons donc à ce stade d'un système d'acquisition de données dynamiques, ainsi que de méthodes pour traiter et extraire des données des informations articulatoires. Nous allons nous intéresser par la suite aux données statiques (IRM) et au recalage de toutes les données dynamiques et statiques dans un repère spatial commun.

Chapitre 5

Données statiques IRM : acquisition et recalage avec les données dynamiques

Ce chapitre présente le protocole d'acquisition des données statiques IRM avec les spécificités à prendre en considération lors de l'acquisition de données articulatoires. Une étude de répétabilité et de variabilité est notamment effectuée à partir du système dynamique pour mettre en place un protocole d'acquisition IRM adapté.
Dans un second temps, le recalage de ces données statiques avec les données dynamiques est détaillé. Une évaluation de l'incertitude du recalage est enfin proposée afin de quantifier la précision globale de notre système.

5.1 Données statiques : IRM

5.1.1 Introduction

Il n'est pas question de prétendre détailler ici le principe de fonctionnement d'une IRM, la formation de l'image et les différentes possibilités et termes de réglages inhérents à cette modalité, qui dépassent largement le cadre de cette thèse. Nous ne présentons que les notions d'IRM utiles pour la mise en place d'un protocole d'acquisition. Pour plus de détails, nous invitons le lecteur à se reporter à un ouvrage spécialisé, comme celui de Kastler [KVPG06].

Une acquisition IRM consiste à obtenir une série d'images (appelées aussi coupes) selon une orientation : sagittale (de gauche à droite du sujet), coronale (de l'avant vers l'arrière du sujet) ou axiale (de haut en bas du sujet). Chaque coupe possède une résolution définie par une taille pixel, et une épaisseur sur laquelle sont imagées les structures. La distance entre chacune des coupes est appelée l'espacement, et est utilisée pour définir un voxel (taille pixel multipliée par l'espacement entre les coupes). Lorsque l'espacement entre les coupes est égal à leur épaisseur, toute l'information tridimensionnelle de la structure imagée est contenue dans les coupes. Elles sont alors dites jointives. En général, les voxels sont anisotropes : la distance entre les coupes est différente de la taille pixel.

L'IRM est habituellement utilisée à des fins médicales et constitue une imagerie anatomique très riche. Cependant, le temps d'acquisition est très long (plus de 4 minutes pour une acquisition du crâne en entier à une résolution de 0.625 mm/pixel et des images de 512×512 pixels). Par conséquent, elle est particulièrement adaptée à l'imagerie des organes statiques et non dé-

formables.
Ce n'est bien évidemment pas le cas dans le conduit vocal pendant la phonation. La contrainte la plus forte qui a guidé la définition de notre protocole d'acquisition a donc été de réduire au maximum le temps d'acquisition afin d'éviter les mouvements des articulateurs, tout en conservant une qualité d'image satisfaisante.

Il existe deux principaux modes d'acquisition à l'IRM : le mode « spin écho » et le mode « écho de gradient ». Le premier permet des acquisitions de bonne qualité, mais souvent avec un temps d'acquisition long. Le second a des temps plus rapides aux dépens d'une résolution spatiale un peu moins bonne. En outre, le mode écho de gradient engendre de nombreux artefacts (les « artefacts de susceptibilité magnétique ») lorsqu'un grand nombre de zones d'interfaces entre l'air et les tissus est présent. C'est le cas dans le conduit vocal où ce mode d'acquisition n'est généralement pas utilisé.

Le temps d'acquisition dépend du nombre de coupes, de leur épaisseur, de l'espacement et de beaucoup d'autres réglages propres à l'IRM (temps de répétition (TR), temps d'écho (TE), taille de la matrice d'acquisition...). Il influe sur la qualité des images résultantes. Un compromis entre qualité d'image et temps d'acquisition doit donc être trouvé pour l'acquisition de données articulatoires. De plus, nous devons nous assurer lors des acquisitions IRM que le conduit vocal reste statique.

5.1.2 État de l'art : protocoles IRM pour l'acquisition de données articulatoires

Chaque machine IRM possède ses propres caractéristiques d'acquisition. C'est une des raisons pour lesquelles on trouve dans la littérature d'acquisition de données articulatoires avec l'IRM un protocole différent à chaque fois. Ces protocoles sont d'autant plus difficiles à reproduire que des informations capitales sont parfois omises, comme la résolution de l'image obtenue par exemple. La littérature existante fait cependant ressurgir deux stratégies possibles, liées au temps d'acquisition.

5.1.2.1 Protocoles d'acquisition longs

La première stratégie, et la plus couramment utilisée par la communauté parole, consiste à effectuer une seule acquisition *longue* en demandant au sujet de maintenir l'articulation une fois qu'il ne peut plus produire de parole. C'est ce que proposent Yang [YK94], Badin [BBR+02], et Engwall [Eng04] dans leurs protocoles d'acquisition. Badin et Engwall utilisent le mode « spin écho » pour obtenir 53 images sagittales de 256×256 pixels avec une résolution de 1 mm/pixel pour des coupes de 3.6 mm d'épaisseur, espacées tous les 4 mm. L'acquisition d'un phonème dure 43 secondes : le locuteur commence donc par prononcer le son, et suivant sa capacité pulmonaire arrête la phonation au bout de quelques secondes (en moyenne une vingtaine d'après nos tests) et tente de maintenir ses articulateurs en position tout en respirant doucement jusqu'à la fin de l'acquisition.

5.1.2.2 Protocoles avec pauses

Baer [BGGN91] puis Story [STH96] proposent une autre stratégie : découper l'acquisition d'un phonème en *sous-acquisitions* de quelques secondes entrecoupées de pauses, afin de per-

mettre au sujet de reprendre sa respiration. Dans ce cas, une acquisition complète dure plusieurs minutes, mais la phonation est assurée pendant les sous-acquisitions. Story obtient avec un tel protocole en mode « spin écho » 26 coupes sagittales de 256×256 pixels avec une résolution de 0.94 mm/pixel pour des coupes de 5 mm d'épaisseur (l'espacement entre les coupes n'est pas précisé). L'acquisition complète d'un phonème dure 10 min, avec des sous-acquisitions de 8 secondes.

5.1.2.3 Discussion

On remarque tout d'abord que les protocoles proposés utilisent le mode spin écho avec des coupes sagittales : en effet, il s'agit du sens dans lequel le moins de coupes sont nécessaires pour couvrir tout le conduit - des lèvres aux cordes vocales. En moyenne, le conduit vocal couvre plus de 9 cm en axial, plus de 7 cm en coronal, et 4-5 cm en sagittal. Même en prenant quelques coupes supplémentaires de part et d'autre (environ 1 cm) pour être certain d'avoir tout le conduit (par exemple, jusqu'aux joues du locuteur en sagittal), la dimension sagittale reste la plus étroite et par conséquent celle où le temps nécessaire à l'acquisition sera le plus court.

Dans le premier cas des acquisitions longues, la difficulté est de conserver la même articulation qu'il y ait ou non phonation, qu'il y ait ou non respiration. Autrement dit, la question est de savoir si la forme de conduit vocal imagée correspond effectivement à la phonation. À notre connaissance, cette étude n'a jamais été réalisée. La littérature considère que la forme du conduit est identique avec ou sans phonation, sans pour autant avoir vérifié ce postulat.
Pour les acquisitions avec pauses, la difficulté est de savoir s'il est possible de répéter la même articulation aussi précisément que possible un grand nombre de fois. Nous avons évalué ces deux stratégies à l'aide de notre système d'acquisition dynamique, car ce dernier permet de visualiser facilement l'évolution des formes de langue au cours du temps.

5.1.3 Faisabilité des protocoles IRM

Nous avons examiné la variabilité articulatoire de la voyelle */i/* qui, d'après les spécialistes en parole, présente une variabilité moyenne, moins forte que celle de */a/* mais plus importante que celle de */u/*. Les formes de langue extraites des images US sont comparées. Grâce à notre système d'acquisition dynamique, ces comparaisons sont effectuées dans le repère intrinsèque à la tête du locuteur, c'est-à-dire dans un repère où les mouvements de la sonde et de la tête ont été retirés.

5.1.3.1 Importance de la phonation

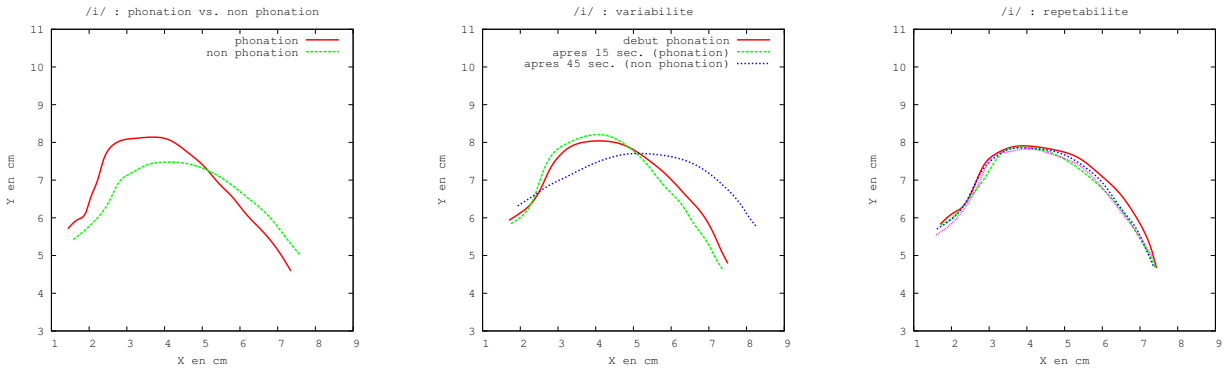
Nous avons d'abord étudié l'influence de la phonation sur la forme du conduit vocal en demandant au sujet de produire le son de la voyelle durant une à deux secondes, et ensuite d'arrêter la phonation tout en conservant la même position articulatoire, simulant ainsi l'approche de Badin [BBR+02] et Engwall [Eng04]. Une première image US est capturée lors de la phonation et une seconde dès qu'elle s'arrête. La figure 5.1.a montre que la distance entre les deux contours est très sensible (de l'ordre de 10 mm) à l'avant de la langue, c'est-à-dire la partie de la langue utilisée pour réaliser la constriction de */i/*. Sans phonation, la langue a tendance à retourner à la position neutre sans que le locuteur en ait conscience.

5.1.3.2 Variabilité pendant la phonation

Nous avons ensuite comparé la variabilité de la langue pendant la phonation à celle provoquée par un temps d'arrêt de 30 secondes de la phonation. Pour cela, le locuteur a reçu la consigne de maintenir la phonation pendant 15 secondes (durée pour laquelle tout locuteur est capable d'émettre un son avant de ne plus avoir d'air), puis de l'arrêter tout en conservant la même position articulatoire pendant 30 secondes. Nous avons capturé trois images : la première au début de la phonation ($t = 0\ s$), la deuxième à la fin de la phonation ($t = 15\ s$) et la troisième 30 secondes après l'arrêt de la phonation ($t = 45\ s$). La figure 5.1.b montre que la variabilité articulatoire induite par l'arrêt de la phonation est beaucoup plus forte que celle observée pendant la phonation : environ 10 mm par rapport à 2 mm. De plus, après 30 secondes de respiration, la langue revient à sa position neutre.

5.1.3.3 Répétabilité

Ces deux premières expériences montrent que la langue n'a pas la même position avec et sans phonation. De plus, il n'est pas possible de maintenir la même position sans phonation. Il semble donc que la meilleure stratégie d'acquisition IRM consiste à répéter plusieurs fois la même articulation, à condition bien sûr de vérifier que le sujet est capable de répéter la même articulation. C'est l'objet de cette dernière expérience pour laquelle la forme de langue a été capturée pour plusieurs occurrences de la même voyelle. La figure 5.1.c confirme que la variabilité articulatoire liée à la répétition du */i/* est faible (inférieure à 2 mm).



(a) (b) (c)

FIG. 5.1 – Comparaisons de la variabilité et de la répétabilité articulatoire. (a) phonation vs. non phonation. (b) variabilité pendant la phonation vs. variabilité entre phonation et non phonation. (c) variabilité due à la répétition.

5.1.3.4 Discussion

Cette étude indique clairement que les formes de langue, et a fortiori du conduit vocal, sont différentes s'il n'y a pas phonation. La meilleure stratégie d'acquisition d'images IRM semble donc de répéter plusieurs fois la même articulation en maintenant la phonation.

Nous n'avons présenté qu'un seul phonème acquis pour un seul locuteur. Sur trois locuteurs, nous avons observé la même tendance pour la langue à retourner à sa position neutre lorsqu'il n'y a plus phonation. C'est d'autant plus vrai que le temps écoulé depuis l'arrêt de la phonation est important. En revanche, il semblerait qu'il y ait d'importantes disparités pour la variabilité de certains phonèmes pour des locuteurs.
Pour deux locuteurs de notre laboratoire, nous avons extrait les formes de langue acquises pendant quinze secondes d'enregistrement US pendant lesquels la phonation est effectuée. Ces deux locuteurs ont l'habitude d'effectuer des acquisitions de données articulatoires. La figure 5.2 présente les résultats pour le /*a*/ et le /*u*/. Le locuteur 1 a une variabilité plus importante que le locuteur 2. En effet, pour le /*a*/, le locuteur 1 montre des variations de positions de la langue proches de 7 mm sur toute sa surface, alors que le second a des variations proches de 2 mm. Pour le /*u*/, les 2 locuteurs ont une variabilité moins importante au niveau de la constriction qu'au niveau de l'apex et du bas de la langue. Si les variations sur l'apex sont moins importantes pour le locuteur 2 (environ 3 mm) que pour le locuteur 1 (près de 6 mm), elles sont semblables au niveau de la racine de la langue (près de 10 mm). Ces résultats montrent que même si la forme générale de la langue reste semblable au cours de la phonation, elle peut malgré tout être soumise selon le locuteur et le phonème à une variabilité importante.

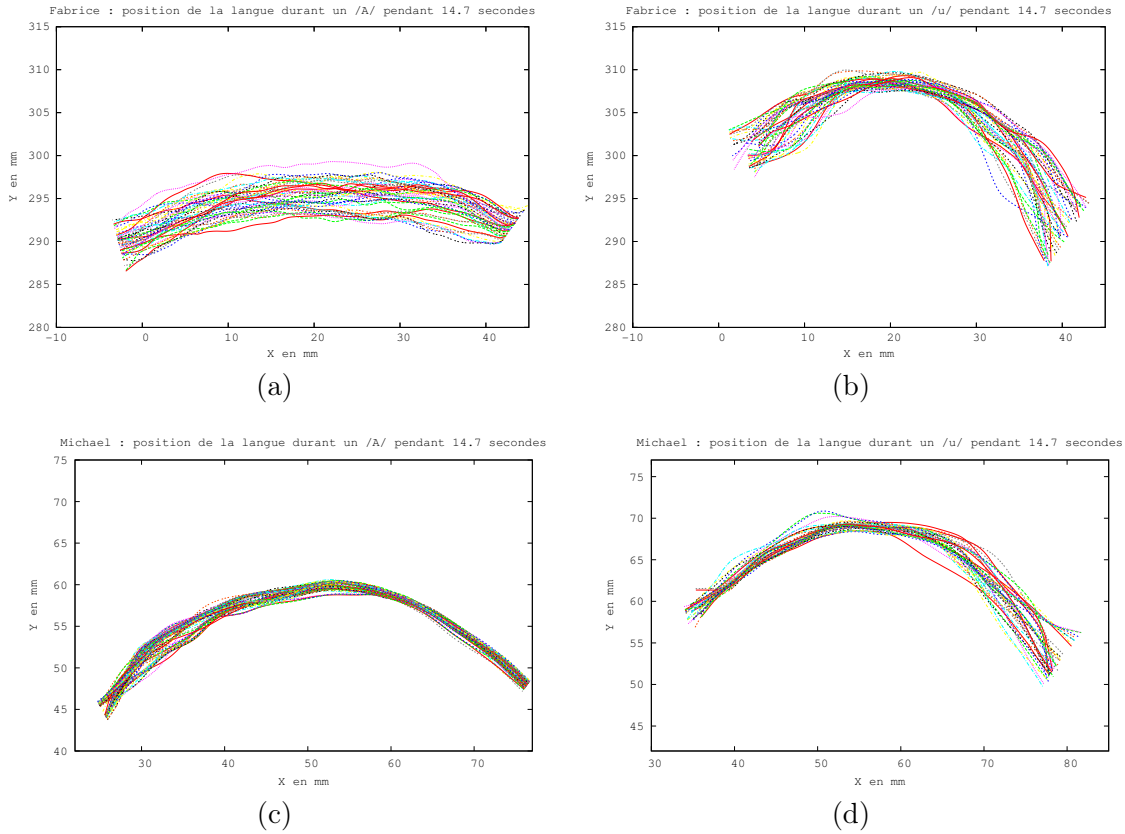


FIG. 5.2 – Variabilité pendant 15 secondes de phonation entre 2 phonèmes de 2 locuteurs différents. L'apex est à gauche et la racine de la langue à droite. (a) /*a*/ du locuteur 1. (b) /*u*/ du locuteur 1. (c) /*a*/ du locuteur 2. (d) /*u*/ du locuteur 2.

5.1.4 Protocole d'acquisition

5.1.4.1 Protocole pour l'acquisition d'un phonème

Un protocole d'acquisition basé sur le protocole de Story [STH96] a été mis en place pour pouvoir obtenir des coupes IRM. Ces acquisitions ont été effectuées au service de neuroradiologie du CHU de Nancy. L'acquisition d'un phonème est découpée en sous-acquisitions pendant lesquelles le locuteur effectue la phonation entrecoupée de pauses pour reprendre sa respiration. Dans la suite de ce manuscrit, nous nommerons **IRM phonème** les images ainsi acquises.

Pour cela, nous sommes partis d'un protocole d'acquisition IRM cholangiographiques, utilisées pour la recherche de pathologies sur les voies biliaires, et que nous avons adapté pour le conduit vocal. Ce protocole propose des acquisitions en mode spin écho avec pauses.

Pour des coupes jointives de 512×512 pixels, de 3 mm d'épaisseur espacées tous les 3 mm (résolution de 0.625 mm/pixel), la figure 5.3.a présente les temps d'acquisition nécessaires. Le plus petit temps d'acquisition atteignable est de 16 secondes pour obtenir indifféremment de 1 à 4 coupes. Pour des acquisitions sagittales sur le conduit vocal qui nécessitent une largeur d'environ 7 cm-8 cm en incluant les joues, il est donc nécessaire d'avoir de 25 à 28 coupes, soit 1 minute et 53 secondes d'acquisition. Le temps minimal d'acquisition étant de 16 secondes, cela donne un total de 7 sous-acquisitions pour obtenir toutes les coupes nécessaires. Si la morphologie du sujet le permet, on peut réduire à 6 sous-acquisitions pour obtenir de 21 à 25 coupes.
En comptant en moyenne une dizaine de secondes de pause entre chaque sous-acquisition, le temps total nécessaire pour l'acquisition d'une IRM phonème est de :

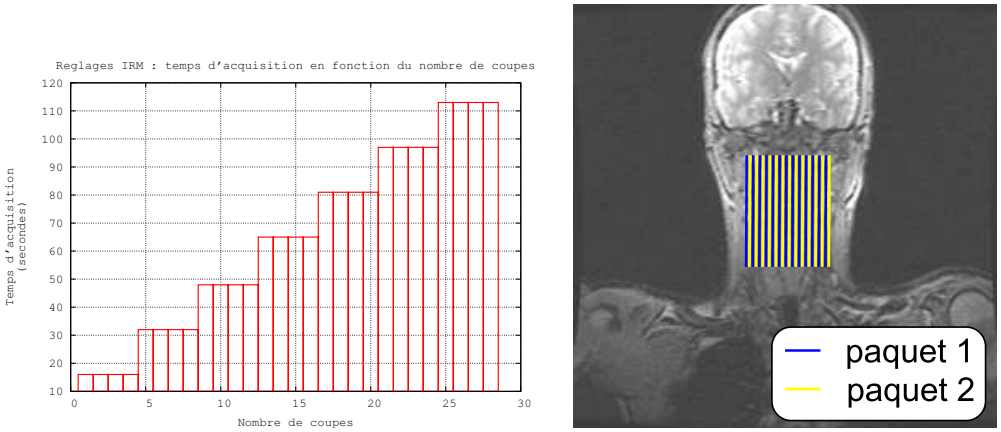
$$(16 \text{ secondes} + 10 \text{ secondes}) \times 7 = 3 \text{ minutes } 2 \text{ secondes}$$

Le temps de pause peut être adapté au locuteur, car c'est l'opérateur qui, manuellement, décide de relancer une sous-acquisition en prévenant le sujet au moyen du microphone intégré à l'IRM. Ce microphone permet en outre de s'assurer que la phonation est effectuée avant de lancer la prochaine sous-acquisition.

En utilisant le mode spin-écho, si l'espacement entre les coupes n'est pas supérieur de plus de 10 % à l'épaisseur de ces coupes, les impulsions magnétiques peuvent exciter partiellement les coupes adjacentes (« phénomène d'excitation croisée »), et créer de nombreux artefacts sur les coupes résultantes. La solution consiste alors à entrelacer (cf figure 5.3.b) deux paquets de coupes ayant des espacements importants pour ne pas imager sur la même excitation les coupes adjacentes. Cela permet d'obtenir des coupes jointives sans perturber l'acquisition des coupes adjacentes.

Les différents réglages présentés dans le tableau 5.1 représentent le meilleur compromis trouvé pour avoir une qualité d'image satisfaisante tout en ayant des temps de sous-acquisition permettant au sujet d'effectuer la phonation.

Des images IRM dans le plan médiosagittal sont présentées sur la figure 5.4. Par rapport aux acquisitions de Badin [BBR$^+$02], nous avons un protocole permettant d'obtenir des formes de conduit qui correspondent réellement à la phonation (pas de maintien artificiel de la position des articulateurs), tout en ayant une meilleure résolution d'image (images de 512×512 pixels à 0.625 mm/pixel au lieu d'images de 256×256 pixels à 1 mm/pixel) et des coupes jointives apportant une information sur la totalité du conduit vocal.



(a) (b)

FIG. 5.3 – (a) Temps d'acquisition en fonction du nombre de coupes avec les réglages de la table 5.1. (b) Principe d'entrelacement des coupes IRM en mode spin écho.

Machine	GE Medical Signa HDx 1.5T
Mode	Fast Spin Echo
Nombre de coupes[1]	26 (2 paquets de 13)
Épaisseur de coupe	3 mm
Espacement entre les coupes[2]	2.6 mm et 3.2 mm
TE (echo delay time)	20.712 ms
TR (repetition time)	400 ms
Image	512x512 pixels
Résolution d'image	0.625 mm/pixels

TAB. 5.1 – Résumé des principaux paramètres d'acquisition IRM pour un phonème. Notes : [1] ce chiffre peut varier suivant la morphologie du sujet. [2] le chiffre est donné à titre indicatif, car l'espacement dépend de la position des deux paquets l'un par rapport à l'autre.

5.1.4.2 Protocole pour un locuteur

Une **session d'acquisition** correspond en un groupement de plusieurs acquisitions IRM phonèmes. En effet, les acquisitions IRM ayant lieu au service de neuroradiologie du CHU de Nancy, nous disposions de la machine IRM pour seulement 1 ou 2 heures. Nous avons donc été contraints d'effectuer des acquisitions en plusieurs fois pour un même sujet.

Pour chaque session d'acquisition, le protocole d'acquisition IRM consiste à :

- positionner le sujet dans la machine IRM en contraignant les mouvements de tête avec des cales en mousse pour éviter des mouvements trop amples durant les acquisitions ;
- effectuer un repérage de la région à imager et définir manuellement la position des grilles entrelacées des coupes IRM de la figure 5.3.a suivant la morphologie du sujet ;
- effectuer une première acquisition où le sujet est au repos. Les paramètres d'acquisition sont détaillés dans le tableau 5.2. Cette première acquisition ne comporte pas de pauses et le temps d'acquisition est supérieur à 4 minutes, car les réglages utilisés permettent

FIG. 5.4 – Exemples d'images IRM acquises (coupes médiosagittales) : (a) */i/* (b) */l/* au début de l'articulation du */la/* (c) */â/* (d) IRM référence, position au repos.

d'obtenir une bonne qualité d'image. Elle sera utilisée par la suite pour le recalage des données IRM entre elles. On nommera cette acquisition **l'IRM de référence**. Elle permet aussi au sujet de s'habituer à l'environnement d'une acquisition IRM ;

- effectuer ensuite les acquisitions d'IRM phonèmes décrites dans la partie 5.1.4.1.

5.1.4.3 Difficultés et solutions

L'IRM est très sensible à de nombreux paramètres. Tout d'abord, certains sujets, de par leur morphologie et leur constitution physique, engendrent des images de moins bonne qualité que d'autres sujets (dans le jargon médical, on parle de sujets qui « résonnent bien » et d'autres qui « résonnent mal »). On ne sait pas prévoir si un sujet résonnera bien ou mal, la seule solution est de tester une acquisition sur lui.

Dans les images présentées sur la figure 5.4, on observe un léger « repliement spectral » :

Machine	GE Medical Signa HDx 1.5T
Mode	Fast Spin Echo
Nombre de coupes	34
Épaisseur de coupe	3 mm
Espacement entre les coupes	3 mm
TE (echo delay time)	21.624 ms
TR (repetition time)	680 ms
Image	512x512 pixels
Résolution d'image	0.625 mm/pixels

TAB. 5.2 – Résumé des principaux paramètres d'acquisition des IRM de référence.

l'arrière du crâne se retrouve sur la gauche de l'image IRM, devant son nez. Ce phénomène peut être évité en augmentant la taille de la zone d'acquisition, mais dans ce cas, le temps des sous-acquisitions augmente lui aussi.
L'image est très sensible aux mouvements du sujet. Il se peut que la tête bouge légèrement durant les acquisitions. Cela entraîne des effets de flou - sur la figure 5.5, on peut observer que les contours de la langue au niveau de l'apex ne sont pas visibles - et des artefacts - sur cette même figure, des artefacts sont visibles au niveau de la mandibule et des cordes vocales, et aussi sur le cerveau du sujet.

Notre protocole permet de se rendre aisément compte du bougé éventuel des articulateurs du locuteur pendant l'acquisition d'une IRM phonème. De par le principe de l'entrelacement, deux coupes adjacentes sont acquises à des instants différents. On peut ainsi observer si un articulateur à bougé durant ce temps (voile du palais...). Sur une trentaine de phonèmes acquis pour un sujet, nous avons constaté ce phénomène sur trois séquences pour lesquelles les positions des articulateurs étaient différentes dans les deux paquets. Cela peut être dû soit à un arrêt trop précoce de la phonation, ou soit à la déglutition. Ces séquences mal acquises ont été refaites.
Pour habituer le locuteur à garder une position de langue stable pendant et entre les sous-acquisitions, on pourrait imaginer l'entraîner avec le système échographique en lui faisant visualiser sa langue. Nous n'avons pas pris le temps d'effectuer un tel entraînement, mais c'est toutefois une solution à envisager pour minimiser les risques de bougé lors des acquisitions IRM. Enfin, on pourrait aussi songer à l'entraîner à prononcer un son dans un environnement très bruité comme l'est l'IRM. En effet certains locuteurs peuvent s'avérer déstabilisés devant le bruit important imposé par une acquisition IRM.

5.1.5 Traitement des images IRM

5.1.5.1 Approche

Une fois les images IRM acquises, des traitements sont nécessaires pour en extraire des informations comme la surface du palais par exemple, ou celle du visage, qui servira au recalage des IRM avec les données dynamiques. La segmentation d'images IRM est un domaine de recherche à part entière. Par manque de temps au cours de ce travail de thèse, nous ne nous sommes pas focalisés sur des techniques de traitement d'image avancées pour l'exploitation de ces images, mais nous nous sommes plutôt concentrés sur les méthodes pour fusionner ces données avec d'autres modalités.

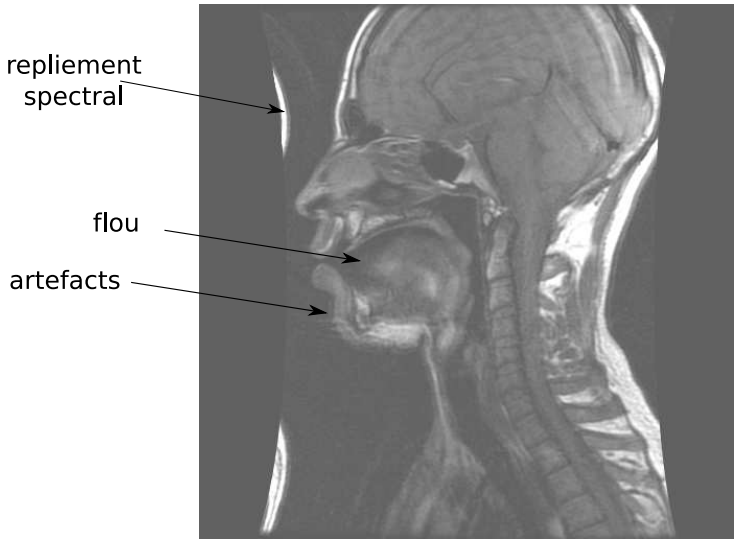


FIG. 5.5 – Exemple d'une mauvaise acquisition IRM : le sujet a bougé.

Pour reconstruire en 3D des surfaces du conduit vocal, nous avons utilisé la méthode des marching cubes [LC87]. Cette méthode permet d'extraire une isosurface sous forme de maillage polygonal, à partir des données IRM. Pour cela, à partir de la configuration (i.e. à l'intérieur ou à l'extérieur de la surface à reconstruire) de 8 voxels voisins dans le volume IRM, l'algorithme des marching cubes établit une configuration de 15 formes polygonales de base, qui appliquées à tous les voxels forment une surface maillée.

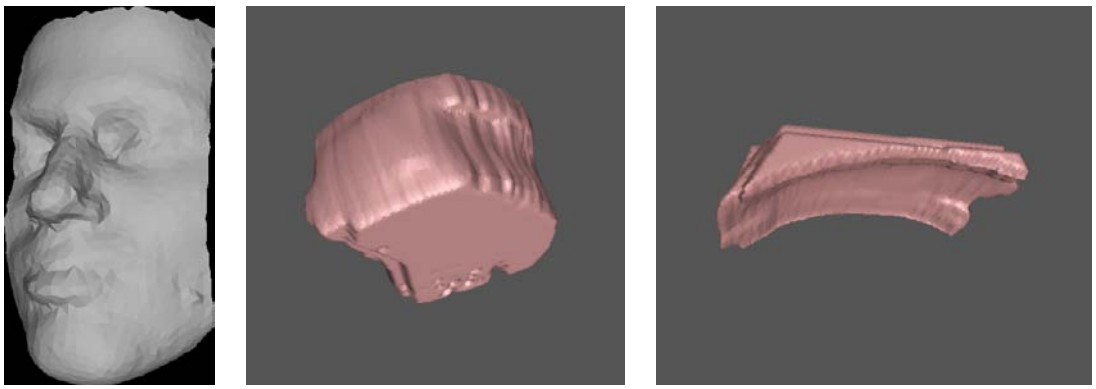
Pour segmenter seulement les structures désirées dans les IRM, et pour séparer notamment le conduit vocal des fosses nasales et de l'air ambiant, une zone d'intérêt autour de chaque élément à extraire est manuellement délimitée sur chaque image. Ensuite, un seuillage est manuellement choisi et appliqué aux images pour segmenter la structure à reconstruire. L'algorithme des marching cubes est enfin appliqué pour reconstruire la surface considérée.

5.1.5.2 Résultats

La figure 5.6 présente une reconstruction de la surface du visage, de la langue lors de la phonation d'un /i/, et du palais.

La délimitation de la zone d'intérêt dans les images IRM a été réalisée manuellement. Cette approche a été choisie, au moins dans un premier temps, car le nombre d'acquisitions est réduit pour un même locuteur (une trentaine d'IRM phonèmes). De plus, la surface du visage et le palais, qui sont des structures statiques par rapport à la tête, ne sont reconstruits que pour l'IRM de référence.

Pour disposer de toutes les structures du conduit vocal, nous avons aussi tenté de faire apparaître les dents du locuteur dans l'IRM. Takemoto [TKNH04] propose de faire boire au locuteur du jus de groseilles juste avant l'acquisition, car ce jus possède des pigments qui se collent aux dents et qui seraient visibles sur les images. Malheureusement, nous ne sommes pas parvenus à obtenir de résultats satisfaisants avec cette technique. Une solution comme dans les travaux de Serrurier [Ser06] consistant à effectuer au préalable un moulage dentaire du locuteur et de le recaler avec les données IRM ensuite est envisagée actuellement dans notre laboratoire.



(a) (b) (c)

FIG. 5.6 – Exemples de surfaces IRM extraites par la méthode des marching cubes. (a) Surface du visage au repos (b) Langue lors d'un /i/. (c) Palais.

5.1.6 Recalage des IRM

5.1.6.1 Introduction

Pour chaque session d'acquisition IRM, la position de la tête du locuteur est différente dans la machine. De plus, nous nous sommes rendu compte qu'au sein même d'une session de plusieurs heures, sa tête bouge entre les acquisitions. Afin de pouvoir disposer d'un même repère spatial de référence dans lequel exprimer toutes ces différentes acquisitions, intra et inter-session, il est nécessaire de les recaler.

Les IRM de référence effectuées en début de chaque session d'acquisition ont l'avantage d'avoir été effectuées au repos pour le locuteur. L'une de ces IRM de référence est choisie comme référence absolue. Toutes les autres données IRM de toutes les sessions d'acquisition seront exprimées dans son repère par un recalage basé image.

5.1.6.2 Technique

Le recalage de deux séquences d'images IRM est un problème de recalage 3D/3D. De nombreuses méthodes ont été présentées dans la littérature [MV98]. Les méthodes se basant sur des recalages géométriques, comme sur la correspondance de contours ou de gradient entre les images, se révèlent précises, mais peu robustes aux erreurs locales [MHH$^+$05], contrairement aux méthodes iconiques qui s'appuient sur des mesures de similarité dans la globalité des images. Au milieu des années 90, le critère de *l'information mutuelle* [Col95, Vio95] a été proposé, et est devenu depuis très populaire pour le recalage d'images médicales. Ce critère a l'avantage d'être robuste aux différences locales entre deux séquences d'images, ainsi qu'aux variations d'intensité. Dans notre application, nous disposons d'IRM phonèmes et d'une IRM de référence sur laquelle les recaler. Les IRM phonèmes caractérisent des positions différentes des articulateurs : toute la zone basse du visage est donc susceptible de bouger. En revanche, la partie haute du visage contenant le cerveau est très stable, quelle que soit l'acquisition. L'information mutuelle est donc parfaitement adaptée à notre problème, ayant la capacité de recaler le haut du crâne de manière robuste face aux données aberrantes, au regard d'un recalage rigide, que constitue la mâchoire inférieure.

L'information mutuelle est basée sur un calcul d'entropie des images. On rappelle que l'entropie de Shannon H d'une image A est définie par

$$H(A) = -\sum_a p_a.log(p_a) \tag{5.1}$$

avec p_a la probabilité d'apparition du niveau de gris a dans l'image A. L'entropie conjointe H de deux images A et B est définie par :

$$H(A,B) = -\sum_a \sum_b p_{a,b}.log(p_{a,b}) \tag{5.2}$$

avec $p_{a,b}$ la probabilité qu'un pixel ait la valeur a dans l'image A et b dans l'image B. On définit alors l'information mutuelle $I(A,B)$ comme la quantité d'information contenue à la fois dans les images A et B. On l'exprime de la façon suivante :

$$I(A,B) = H(A) + H(B) - H(A,B) \tag{5.3}$$

où H représente l'entropie marginale et $H(A,B)$ l'entropie conjointe. Lorsque l'information mutuelle entre deux images est maximale, elles sont recalées. Pour plus de détails, nous invitons le lecteur à lire le document de synthèse de Pluim [PMV03] sur l'information mutuelle.

Cette méthode est facilement utilisable, car elle est disponible dans la bibliothèque gratuite *itk* (Insight Segmentation and Registration Toolkit[15]), qui implémente la méthode de Mattes [MHV+03]. Elle a donc été utilisée pour recaler toutes les IRM phonèmes en trois dimensions quelque soit la session par rapport à une IRM de référence pour un même locuteur. Le principe est résumé sur la figure 5.7.

5.2 Recalage multimodal

Nous avons vu dans la partie précédente le processus d'obtention de données IRM, et une méthode pour que cet ensemble de données statiques soit exprimé dans un repère spatial intrinsèque à la tête. Les données avec les données dynamiques présentées dans les chapitres 3 et 4 de ce manuscrit ont, de leur côté, été elles aussi exprimées dans un repère commun intrinsèque à la tête. Le second volet de ce chapitre expose notre méthode pour estimer la transformation qui lie ces deux repères réalisant par ce biais le recalage des données statiques et dynamiques.

5.2.1 Introduction

Nous disposons d'une part de données US et EM (recalées grâce au calibrage EM/US présenté dans le chapitre 4) et d'autre part de données IRM. Pour un même locuteur, le calcul de la transformation rigide entre les repères US et IRM permettra de superposer à des positions d'articulateurs statiques extraits de l'IRM des informations dynamiques de position de la langue extraites des données US, et de disposer ainsi d'un ensemble de données toutes exprimées dans un même repère.

Un recalage, qu'il soit géométrique ou iconique, se base sur des informations communes entre les modalités à recaler. Maintz [MV98] précise dans son état de l'art sur les différentes

[15] `http://www.itk.org`

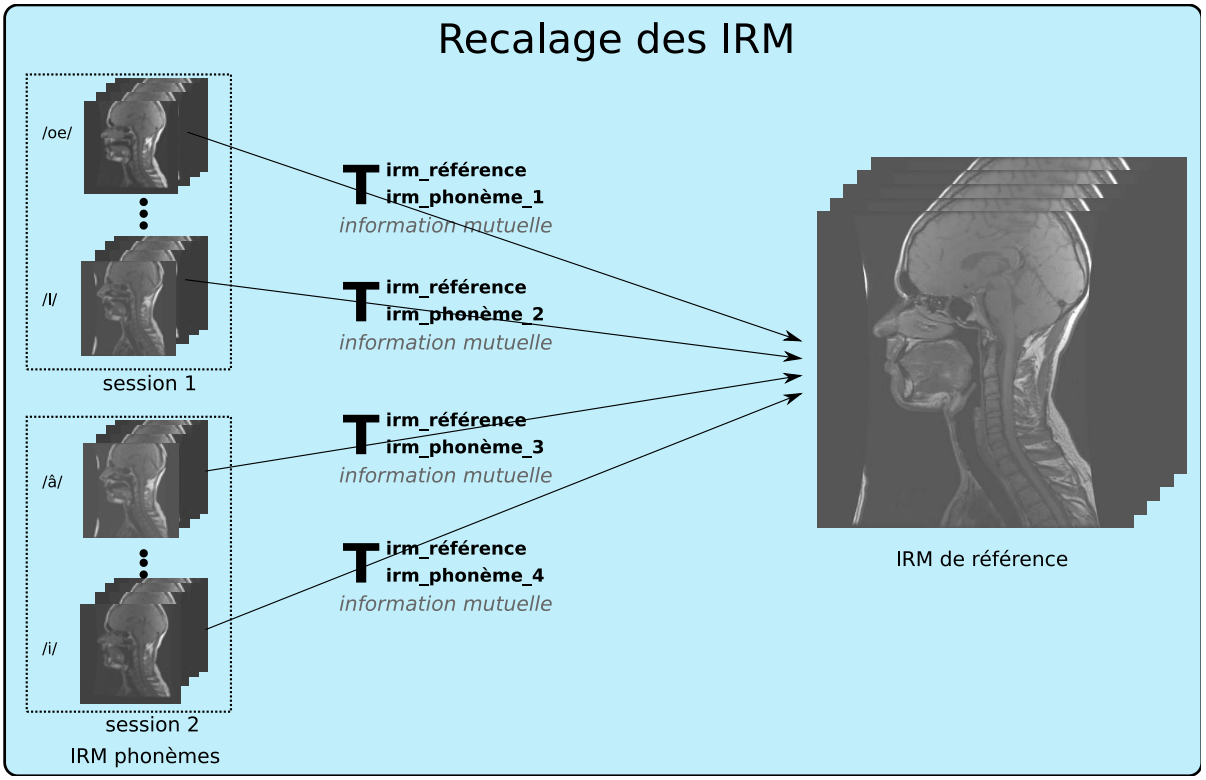


FIG. 5.7 – Schéma récapitulatif du recalage utilisant l'information mutuelle entre les séquences IRM : chaque IRM phonème de chaque session d'acquisition est recalée sur l'IRM de référence.

méthodes de recalage que très peu de travaux ont été effectués sur le recalage US/IRM. Cela vient probablement du fait de la trop grande disparité entre les deux modalités, de la pauvreté de l'information des images US et de leur aspect extrêmement bruité. Certains travaux se sont cependant penchés sur le recalage US/IRM.

Les travaux de Roche [RPMA01] proposent de recaler images US et IRM du cerveau en utilisant une mesure de similarité robuste aux changements d'intensité, par le calcul d'un rapport de corrélation basé sur la dépendance fonctionnelle d'une image par rapport à l'autre. Il dispose d'images US et IRM où des structures caractéristiques du cerveau (circonvolutions par exemple) sont visibles dans les deux modalités.

Pagoulatos [PHK00] propose d'initialiser le recalage entre l'image US et l'IRM en utilisant un système électromagnétique : des marqueurs sont collés sur le sujet, et sont visibles à l'IRM. Ils sont ensuite détectés par le système EM pour initialiser le recalage entre les repères EM et IRM. Comme avec notre système, un capteur EM est placé sur la sonde US, et moyennant un calibrage, les deux repères sont recalés. Le système est ensuite utilisé sur un fantôme spécialement manufacturé pour leur application, où des structures sont visibles à la fois dans les images US et IRM. Le critère de l'information mutuelle est enfin appliqué pour affiner le recalage entre les deux modalités.

Notre système est proche de celui de Pagoulatos [PHK00] avec la modalité EM. Mais contrairement à ces travaux ou à ceux de Roche, nous n'avons aucune structure commune visible dans les images IRM et dans les images US et/ou dans les données EM pour effectuer un recalage iconique. Il est donc nécessaire de mettre en place une procédure spécifique pour obtenir des informations communes entre les données dynamiques et statiques.

5.2.2 Méthode

5.2.2.1 Principe

Nous avons vu dans la première partie de ce chapitre comment extraire des données des acquisitions IRM. La partie 5.1.5.2 présente notamment la surface du visage du locuteur extraite à partir de l'IRM de référence. La surface du visage peut aussi être numérisée avec les capteurs EM, en la balayant avec un stylet EM lors de chaque session d'acquisition de données dynamiques. Elle représente une information commune entre les deux modalités, utilisable pour leur recalage. Les données US étant préalablement recalées avec les données EM par le calibrage EM/US, elles sont corollairement recalées avec les données IRM en composant les transformations rigides.

Le palais a aussi été testé comme structure commune : sa surface, balayée par le stylet EM, peut aussi être recalée avec celle extraite de l'IRM. Cependant, le palais ne couvre qu'une très faible surface, sans structure saillante sur laquelle le recalage pourrait s'ancrer, le rendant ainsi très incertain. De plus, les acquisitions avec le stylet EM sur le palais sont beaucoup moins pratiques à réaliser, et très inconfortables pour le locuteur, par rapport au balayage de la surface de son visage.

5.2.2.2 Technique

La méthode la plus couramment utilisée pour le recalage de deux surfaces 3D est l'*Iterative Closest Point* (ICP), proposée par Besl et McKay [BM92]. Elle consiste à chercher itérativement la transformation rigide T qui minimise la distance entre les deux surfaces. Pour le recalage d'une surface S_1 sur une surface S_2, la surface S_1 est exprimée selon en ensemble de N points 3D $(P_i)_{0\leq i\leq N}$, et le recalage par ICP consiste à chercher la transformation $\widetilde{T}$ suivante :

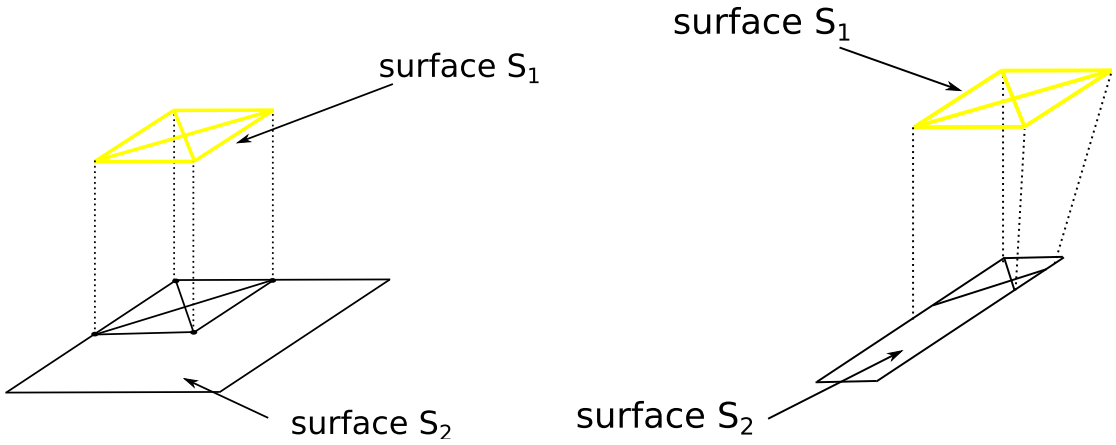
$$\widetilde{T} = \arg\min_T \sum_{P_i \in S_1} dist(T(P_i), S_2) \tag{5.4}$$

Il est important lors de l'utilisation directe de cette méthode que la surface S_1 soit totalement recouverte par la surface S_2. Dans le cas contraire (cf figure 5.8), l'algorithme cherche à placer les points de la surface S_1 sur une région de la surface S_2 qui ne lui correspond pas, et la minimisation n'a plus de sens. Il existe des méthodes de recalage par ICP plus robustes [RL01] et qui en pondérant les points de la surface S_1 autorisent un recouvrement partiel des deux surfaces. Nous verrons par la suite que leur utilisation n'est pas nécessaire dans notre contexte, car nous avons choisi une surface de référence S_2 toujours plus grande et plus précise que les surfaces S_1.

Cette référence est la surface obtenue à l'aide d'un numériseur 3D (3D mega capturor, Inspeck) permettant de numériser la surface du visage rapidement en la modélisant par un maillage 3D complet et dense (plus de 10000 sommets). Les modalités IRM, EM et stéréo sont recalées sur cette surface de référence par ICP.

5.2.2.3 Recalage des données dynamiques

En début de chaque session d'acquisition de données dynamiques, la surface du visage du locuteur est balayée par le stylet EM. Elle est exprimée dans le repère tête et recalée, par ICP sur la surface du visage numérisée. En notant T_a^b la transformation rigide permettant de calculer le passage de la modalité a vers la modalité b, on note $T_{em\_tete}^{numer}$ cette transformation.
Les données de stéréovision sont aussi recalées en choisissant un couple d'images pour lesquelles la tête du locuteur est au repos. La surface du visage correspondant à ce couple d'images est



(a) (b)

FIG. 5.8 – Recouvrement entre deux surfaces : (a) la surface S_1 en jaune, est inclus dans la surface S_2, et est bien recalée par ICP. (b) la surface S_1 est mal recalée sur S_2 par ICP car les correspondances entre les points sont fausses.

reconstruite, et constitue la surface du visage de stéréovision de référence par rapport à laquelle toutes les autres données de stéréovision sont exprimées. Elle est recalée par ICP sur la surface du visage numérisée (T_{stereo}^{numer}).
Le principe de recalage des données dynamiques est résumé sur la figure 5.9.

5.2.2.4 Recalage des données dynamiques et statiques

La surface du visage extraite de l'IRM de référence présentée en section 5.1.5.2 est recalée par ICP sur la surface du visage numérisée (T_{irm}^{numer}).

À ce stade, toutes les données statiques et dynamiques peuvent être recalées entre elles, en composant les transformations rigides obtenues.
Les différents repères utilisés dans le recalage sont :

- *us* : le repère US,
- *em_sonde* : le repère lié au capteur EM attaché à la sonde US,
- *em_tete* : le repère intrinsèque à la tête du locuteur et défini par les deux capteurs EM fixés derrière les oreilles du locuteur,
- *stereo* : le repère de stéréovision,
- *numer* : le repère de la surface du visage numérisé,
- *irm_reference* : le repère de l'IRM de référence,
- *irm_phonemes* : les repères de chaque IRM phonème.

Les différentes transformations utilisées dans le recalage sont :

- $T_{us}^{em\_sonde}$: obtenue par le calibrage EM/US présenté au chapitre 4,
- $T_{em\_sonde}^{em\_tete}$: donnée par le système EM,
- $T_{em\_tete}^{numer}$: calculée par ICP entre la surface du visage numérisé et les points 3D obtenus par le stylet EM,
- T_{numer}^{stereo} : calculée par ICP entre le repère des données de stéréovision et de la surface du

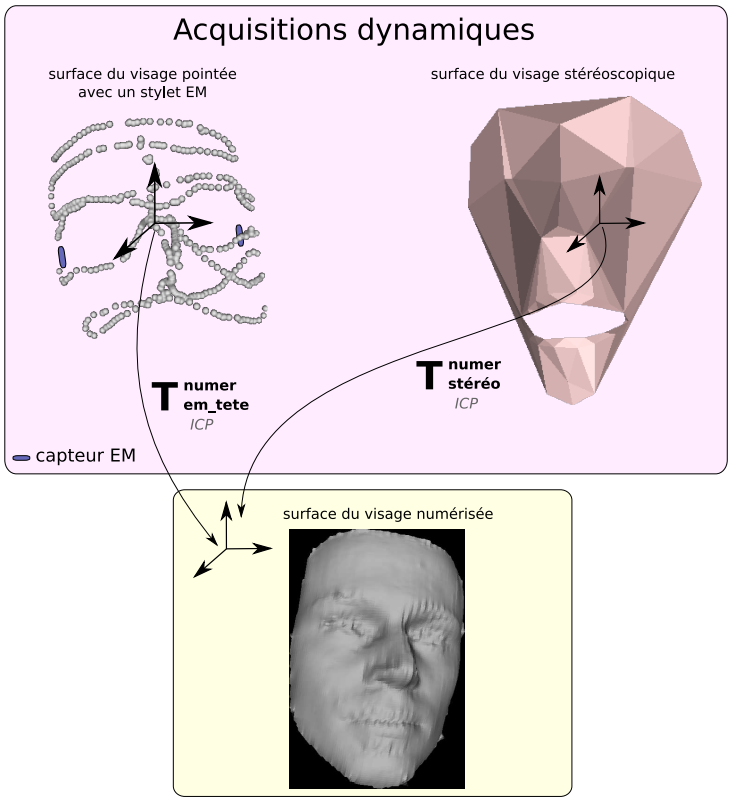


FIG. 5.9 – Schéma récapitulatif du recalage des données US, EM et de stéréovision dans un repère commun aux trois modalités.

visage numérisée.

- $T_{numer}^{irm\_reference}$: calculée par ICP entre la surface du visage extraite de l'IRM et celle obtenue avec le numériseur 3D
- $T_{irm\_reference}^{irm\_phoneme\_i}$: calculées par information mutuelle.

La figure 5.10 récapitule tous les recalages utilisés dans notre système d'acquisition. La transformation permettant de passer d'une donnée IRM phonème à une donnée US est donc donnée par :

$$T_{us}^{irm\_phoneme\_i} = T_{irm\_reference}^{irm\_phoneme\_i} \cdot T_{numer}^{irm\_reference} \cdot T_{em\_tete}^{numer} \cdot T_{em\_sonde}^{em\_tete} \cdot T_{us}^{em\_sonde} \quad (5.5)$$

De la même façon, le recalage de données de stéréovision avec une donnée IRM phonème est donné par :

$$T_{stereo}^{irm\_phoneme\_i} = T_{irm\_reference}^{irm\_phoneme\_i} \cdot T_{numer}^{irm\_reference} \cdot T_{stereo}^{numer} \quad (5.6)$$

5.2.2.5 Discussion

Un repère intermédiaire, celui correspondant à la surface du visage numérisé, a été choisi comme repère de référence sur lequel recaler toutes les données dynamiques (EM, US et stéréovision). On aurait pu essayer de recaler directement le balayage EM exprimé dans le repère tête avec la surface du visage IRM. Cependant, nous nous sommes rendu compte que seuls 50% des

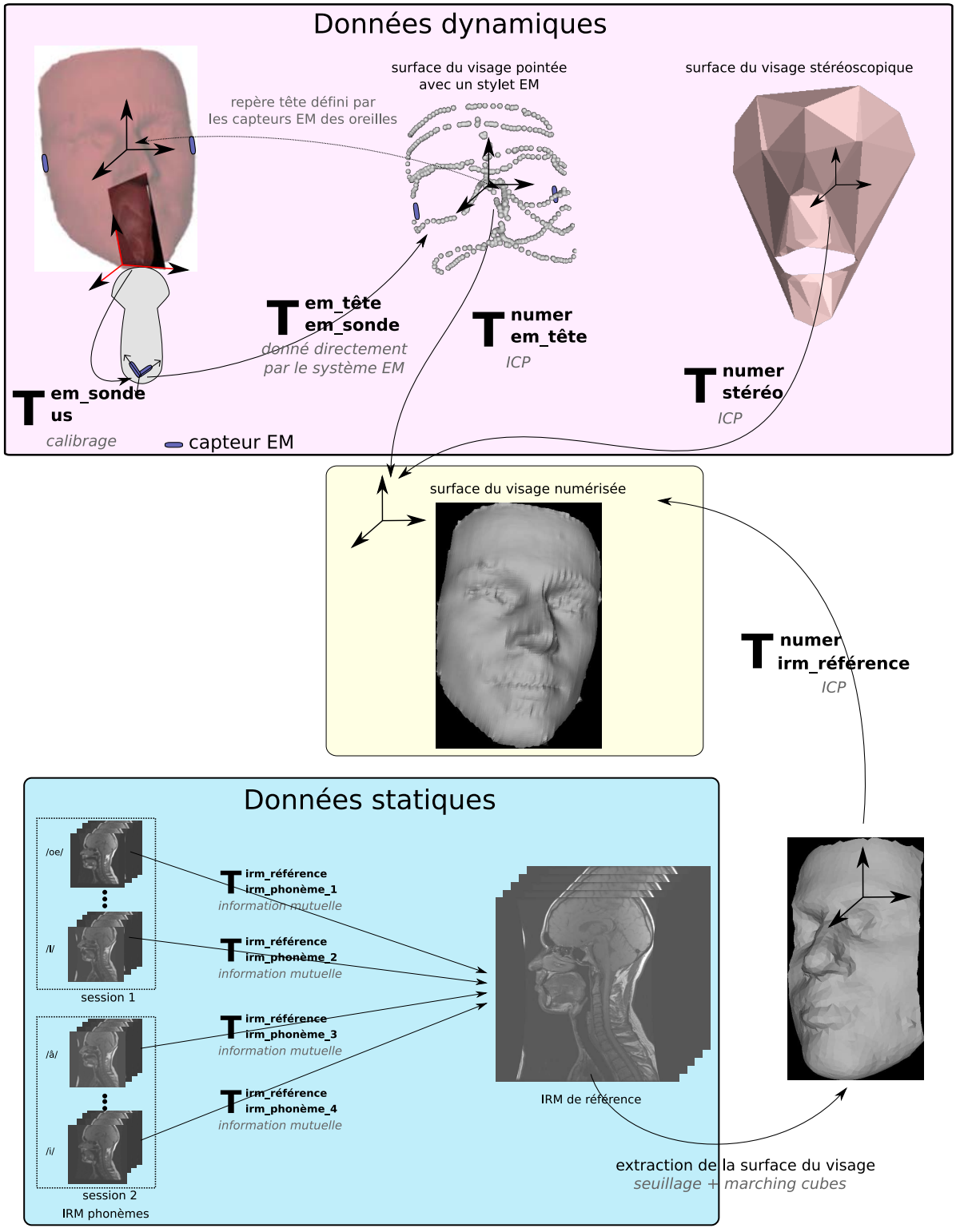


FIG. 5.10 – Schéma récapitulatif présentant les différents recalages utilisés entre les modalités. La méthode utilisée pour le calcul de chaque transformation T est inscrite en grisée.

points EM recouvraient la surface du visage extraite de l'IRM. En effet, cette dernière représente le visage sur une largeur de 10 cm environ autour du plan médiosagittal (en reprenant les paramètres du tableau 5.2, 34 coupes de 3 mm donnent une largeur de 10.2 cm), alors que les points EM balayent tout le visage, de l'oreille droite à l'oreille gauche. Le recouvrement entre les deux modalités était donc très partiel.
La solution aurait consisté à effectuer une acquisition complète de la tête du locuteur à l'IRM afin de disposer d'un recouvrement total entre les données EM et IRM. Ne disposant pas de cette donnée au moment où le recalage a été mis en place, nous n'avons pas pu tester cette solution. Cependant, elle sera testée pour les futures acquisitions, et elle peut aussi être envisagée par les équipes ne disposant pas de numériseur 3D.

De plus, cette surface intermédiaire a l'avantage d'être représentée par un maillage dense (plus de 10000 sommets) alors que les autres modalités ont des maillages plus épars (1000 points pour le balayage EM, 1500 points pour le visage IRM, 50 points pour la stéréovision). Le recalage par ICP est plus précis si le maillage utilisé comme référence est dense.

Un recalage direct des données de stéréovision avec les données EM a aussi été envisagé : le stylet pointeur a été utilisé pour détecter les points de la mire nécessaire au calibrage des caméras de stéréovision. Malheureusement, la mire contient des matériaux ferromagnétiques qui empêchent toute mesure EM lorsque le stylet est proche d'elle.

5.3 Résultats et évaluations

Grâce au recalage précédemment présenté, nous sommes désormais en mesure d'exprimer toutes les modalités dans un même repère spatial pour obtenir une image 3D fusionnée. La figure 5.11 présente un exemple de visualisation 3D de données US, EM, et IRM recalées dans le repère IRM de référence. Les séquences dont les images de ce paragraphe sont extraites sont disponibles à l'adresse `http://www.loria.fr/~aron/these.html`.

Ce recalage permet aussi de se rendre compte d'erreurs ponctuelles dans les séquences dynamiques. En effet, malgré l'utilisation du système de visualisation 3D de la position du plan US par rapport aux capteurs EM durant les acquisitions dynamiques, il n'est parfois pas facile dans l'action de se rendre compte que le plan US s'éloigne du plan médiosagittal, comme sur la figure 5.12. Ces séquences peuvent être facilement repérées grâce au recalage.

Le recalage effectué dans ces exemples met en œuvre l'équation 5.5. Celle-ci est la composition de transformations rigides qui ont chacune une précision liée à la façon dont elles sont calculées. Nous allons nous intéresser dans cette partie à la précision globale de notre recalage, en étudiant plus précisément les incertitudes qui le composent afin d'évaluer la qualité de nos images 3D fusionnées de conduit vocal.

5.3.1 Évaluation perceptive

Avant toute évaluation quantitative des incertitudes, la première chose à faire est de visuellement se rendre compte de la qualité des données acquises.

Le palais d'un locuteur a été extrait des données IRM, puis son intersection avec le plan US a été calculée produisant une courbe que nous avons superposée à l'image US pour toutes les séquences dynamiques acquises avec le même locuteur. L'impression laissée par la visualisation

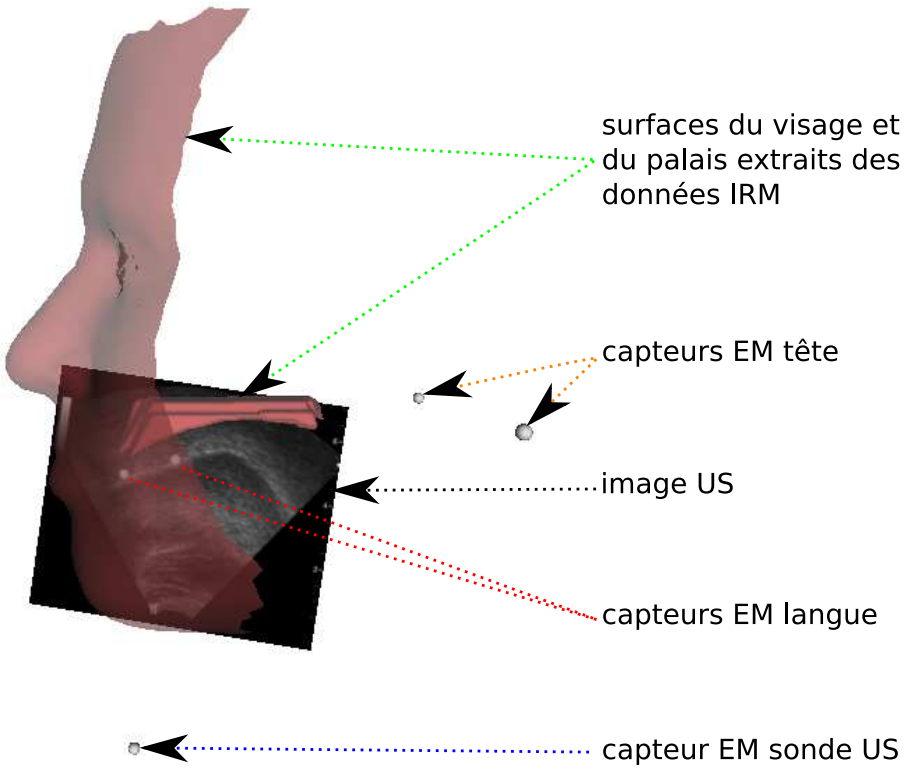


FIG. 5.11 – Visualisation 3D du recalage entre les modalités US, EM et IRM.

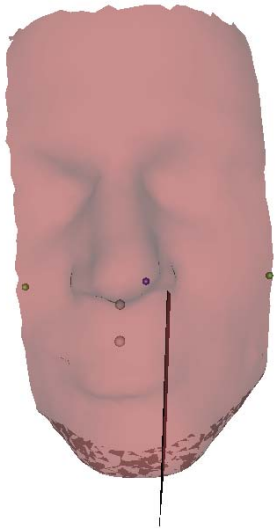
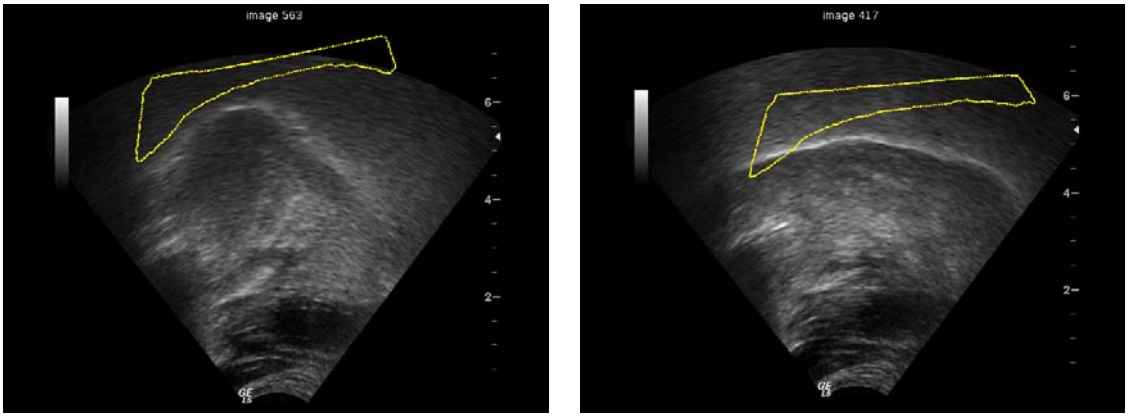


FIG. 5.12 – Visualisation 3D du recalage entre les modalités US, EM et IRM. Le plan US s'est éloigné du plan médiosagittal.

des séquences est cohérente, comme pour la figure 5.13.a qui présente le phonème /*u*/ de /*au*/ pour lequel la langue s'approche du palais sans jamais le toucher.

Cependant, la figure 5.13.b présente une acquisition effectuée quelques minutes plus tard où le locuteur prononce la transition entre le */g/* et le */e/* du mot « nager » de la phrase « ce petit canard apprend à nager ». La langue vient traverser le palais d'une dizaine de pixels (soit 1.7 *mm*) sur l'image US au niveau de l'apex, ce qui prouve que des erreurs sont bien présentes dans notre recalage.

Il est difficile de fournir une mesure quantitative sur le volume des données présentant un recalage problématique par rapport aux volumes des données acquises, les résultats étant très variables selon les séquences. Sur une séquence dynamique de 975 images pour des acquisitions VV, aucune ne comporte une telle superposition incohérente. Sur une séquence prise quelques minutes plus tard avec les mêmes recalages entre les modalités, une séquence pour laquelle une phrase est prononcée comporte plus de 370 images (plus de 38%) où la langue semble traverser le palais.



(a) (b)

FIG. 5.13 – Exemples de contours de palais reconstruit à partir de données IRM recalées dans les images US : (a) */u/* de */au/*, (b) (a) */e/* de « nager ».

Résolutions Les impressions laissées par cette première étude perceptive doivent être modérées par le problème de la visualisation de données extraites d'images ayant des résolutions différentes. En effet, dans les exemples présentés, les données IRM sont extraites d'images ayant une résolution de 0.625 mm/pixel et sont affichées dans une image US qui a pour résolution 0.17 mm/pixel. Il y a donc un rapport d'environ de 1 à 4 entre ces deux résolutions. Cela signifie qu'en omettant toute erreur intermédiaire, un pixel IRM appartenant par exemple à un contour du palais peut être représenté par 16 pixels US. En reprenant de la figure 5.13.b, on peut effectuer l'opération inverse, c'est-à-dire superposer le contour de langue extrait de l'US sur l'image IRM (cf figure 5.14). L'impression visuelle laissée par la superposition des données des deux modalités recalées est cette fois bien meilleure.

5.3.2 Mesures d'incertitudes

5.3.2.1 Introduction

Pour notre application où plusieurs modalités sont fusionnées, une étude de la précision globale incluant les mesures de précision de chaque modalité est nécessaire. Nous aimerions être en

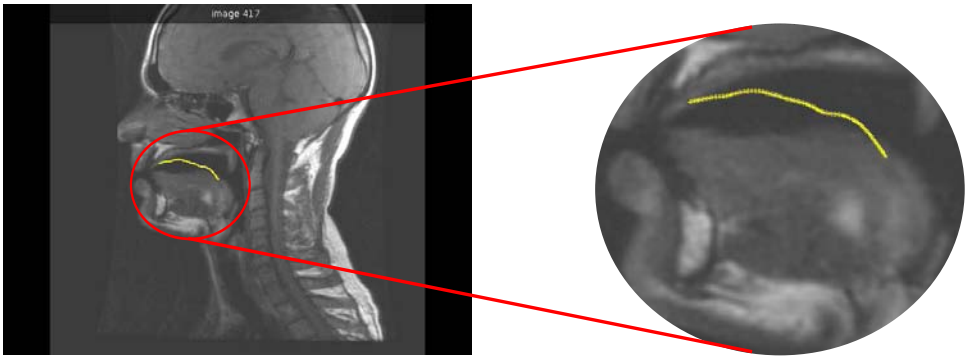


FIG. 5.14 – Visualisation du contour de langue extrait de l'US dans une image IRM.

mesure de quantifier l'incertitude présente sur les données à l'issu du recalage multimodal. Évaluer quantitativement cette incertitude est un problème difficile. En effet, nous ne possédons pas de vérité terrain avec laquelle comparer nos données, et fournir ainsi des mesures fiables. De plus, nous sommes limités par les moyens d'acquisition que nous utilisons. À cause de la limitation du système EM à six capteurs par exemple, nous ne pouvons pas ajouter un troisième capteur sur la tête pour avoir un calcul du repère tête plus robuste, ou tout simplement être en mesure de vérifier qu'aucun de ces capteurs ne bouge l'un par rapport à l'autre pendant les acquisitions. Cette absence de redondance dans les données est une contrainte forte de notre système. Pourtant, le système EM fournit des données avec des erreurs de positionnement, et il nous faut composer avec.

Nicolau [Nic04] propose, dans un cadre de recalage d'une aiguille 3D sur une image 2D, de quantifier l'incertitude de son système en effectuant une analyse des perturbations sur les transformations utilisées dans le recalage. Pour cela il étudie analytiquement la propagation linéaire des covariances de chaque transformation en utilisant les matrices jacobiennes associées.
Dans notre cas, les transformations utilisées dans le recalage font intervenir des méthodes de minimisation non linéaires (le calcul du calibrage EM/US par exemple), rendant difficile l'expression analytique des covariances associées. Cependant, l'incertitude de notre système peut être étudiée en adoptant une approche statistique de type Monte Carlo [HZ00] : elle permet d'estimer la covariance globale d'un système en étudiant l'influence d'un bruit appliqué aux données utilisées pour les calculs des transformations intermédiaires.

Nous effectuons cette étude sur la plus longue chaîne de recalage utilisée dans notre système, c'est-à-dire le recalage entre les données US et les données IRM, afin de quantifier l'incertitude globale de notre système.

5.3.2.2 Principe

Une approche de type Monte Carlo consiste à utiliser une méthode empirique pour étudier l'influence d'un bruit appliqué à un ensemble de données. Cette approche découle de la loi forte des grands nombres qui stipule qu'en appliquant un grand nombre de fois une même expérience aléatoire à un calcul numérique, la moyenne des résultats obtenus tend à se rapprocher de l'espérance mathématique de l'expérience.

Partant de la position d'un point P de l'image d'origine, on cherche à étudier l'incertitude de la position de ce point dans l'image fusionnée, après qu'il ait subi les transformations utilisées

dans le recalage $\mathcal{T}$ entre l'image d'origine et l'image fusionnée. Pour cela, la méthode consiste à bruiter les différentes données $\mathcal{D}$ utilisées pour le calcul de $\mathcal{T}$ suivant un bruit défini par leur covariance. On note f le procédé permettant le calcul de $\mathcal{T}$, d'après les données $\mathcal{D}$:

$$\mathcal{T} = f(\mathcal{D}) \tag{5.7}$$

On étudie donc l'influence sur $\mathcal{T}$ d'un bruit appliqué sur $\mathcal{D}$. Cette approche permettant d'obtenir l'estimation de la covariance globale de la transformation est résumée dans le tableau 5.3.

Cette méthode peut être appliquée à plusieurs transformations $\mathcal{T}$, qu'il suffit de composer pour avoir la transformation globale. En faisant varier le bruit appliqué aux données permettant le calcul des $\mathcal{T}$, on peut en outre cibler les étapes ayant une influence importante dans le recalage pour les améliorer.

Algorithme Monte Carlo (Point P, Données $\mathcal{D}$)

(i) Pour k allant de 1 à K tirages
  Génération aléatoire de données $\mathcal{D}_k$ en ajoutant un bruit sur les données $\mathcal{D}$
  Calcul de la transformation $\mathcal{T}_k$, $\mathcal{T}_k = f(\mathcal{D}_k)$
  Calcul de P'_k, $P'_k = \mathcal{T}_k(P)$
(ii) Calcul de l'erreur RMS entre les positions des points P'_k

TAB. 5.3 – Principe de la méthode de Monte Carlo appliquée à notre système.

L'erreur RMS (Root Mean Square) est définie par l'équation 5.8.

$$err_{rms} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} |P'_k - \overline{P'}|^2} \tag{5.8}$$

avec $\overline{P'}$ représentant la moyenne des points P'_k.

Les données $\mathcal{D}$ correspondent aux données de position de chacune des modalités utilisées dans notre système. Chacune de ces données intervenant dans le calcul du recalage de l'équation 5.5 est bruitée suivant ce principe. Elles sont répertoriées dans le tableau 5.4.

Transformation	**Données**
$T_{us}^{em\_sonde}$	positions des capteurs EM du fantôme (calibrage EM/US) positions du capteur EM sur la sonde US (calibrage EM/US) pointage manuel du point US (calibrage EM/US) résolution US
$T_{em\_sonde}^{em\_tete}$	positions du capteur EM sur la sonde US positions des deux capteurs EM tête
$T_{em\_tete}^{numer}$	pointage avec le stylet EM pour le recalage par ICP
$T_{numer}^{irm\_reference}$	extraction de la surface IRM pour le recalage par ICP

TAB. 5.4 – Modalités intervenant dans chaque étape du calcul du recalage US/IRM.

Nous avons vu dans le chapitre 3 que les données des capteurs EM avaient une incertitude de 0.53 mm sur les données de translation pour les capteurs à 5 DDL, et de 0.76 mm pour le capteur fixé sur la sonde. Les données de rotations sont quant à elles soumises à une incertitude de 0.5 °. Les capteurs EM pouvant avoir des mesures erronées, on peut assimiler le bruit sur ces données de position à un bruit B gaussien, de moyenne nulle et de matrice de covariance Σ. Soit X la vraie position et Y la position bruitée : on modélise très simplement la perturbation sur les données de position par :

$$Y = X + B \tag{5.9}$$

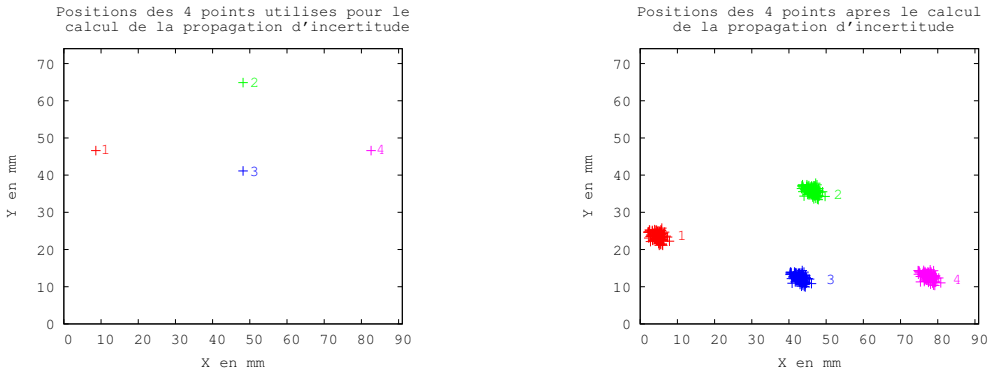
où le bruit B suit une loi normale $\mathcal{N}(0, \Sigma)$. Nous faisons ici l'hypothèse d'un bruit gaussien sur ces données, ce qui n'a pas été rigoureusement prouvé. Nous pouvons cependant considérer que cette hypothèse est la plus vraisemblable et même si elle ne permet d'obtenir une valeur exacte de l'incertitude, elle en donnera une approximation valable.
Le chapitre 3 a aussi montré que la valeur de résolution dans les images US est entachée d'une incertitude. Les différentes valeurs testées font état d'une résolution ayant un écart-type de 0.007 mm/pixel. On peut donc raisonnablement appliquer un bruit uniforme de 0.007 mm/pixel à cette valeur.
Dans le chapitre 4, nous avons vu que l'incertitude du pointage manuel des points d'intersection entre le plan US et le fantôme est de 2.09 pixels. Nous pouvons assimiler ce bruit à un bruit gaussien comme pour le capteur EM.
Enfin l'extraction des surfaces à partir des données IRM comporte elle aussi une incertitude. Nous n'avons pas de moyen de la calculer, mais nous pouvons au moins l'assimiler à la valeur de résolution des images IRM (0.625 mm) sur les axes X et Y, et à l'épaisseur de coupe (3 mm) suivant l'axe Z. Elles correspondent à un bruit uniforme sur les données IRM.
Ces valeurs sont résumées dans le tableau 5.5. Nous supposons le bruit sur les données est soit uniforme lorsque nous estimons que l'erreur ne peut pas être supérieure à l'incertitude, soit gaussien lorsque des mesures aberrantes peuvent survenir (typiquement les capteurs EM). Ces hypothèses devraient être rigoureusement vérifiées, mais elles nous apparaissent comme les plus vraisemblables dans notre configuration.

Méthode bruitée	**Type de bruit**	**Écart-type**
capteurs EM	gaussien	0.50 mm et 0.76 mm sur la sonde US 0.5 ° en rotation
pointage US	gaussien	2.09 pixels
résolution US	uniforme	0.007 mm/pixel
IRM	uniforme	0.625 mm sur X et Y, 3 mm sur Z

Tab. 5.5 – Incertitudes sur chacune des données des modalités utilisées dans le calcul du recalage.

5.3.2.3 Résultats

Principe expérimental Les positions de quatre points 2D (cf figure 5.15.a), en millimètres, correspondant à des positions de la langue et du palais dans les images US sont choisies comme points de mesure. Ces 4 points correspondent respectivement à la position de l'apex à la gauche de l'image (point 1), la position du palais au centre haut de l'image (point 2), la position du dos de la langue au centre de l'image (point 3), et la position de l'arrière de la langue à la droite de l'image (point 4).



(a) (b)

FIG. 5.15 – (a) Positions des 4 points utilisés pour le calcul des incertitudes : le point 1 correspond à une position de l'apex, le 2 à une du palais, le 3 au milieu de la langue et le 4 à l'arrière de la langue. (b) Calcul de la propagation de l'erreur en utilisant toutes les incertitudes sur le recalage (capteurs EM, pointage US, résolution US, et ICP IRM). Les points sont projettés dans le plan sagittal.

La méthode de Monte Carlo décrite précédemment est alors appliquée à ces points suivant l'équation 5.5, les points intervenant dans le calcul de l'erreur RMS (cf équation 5.8) sont exprimés dans le repère 3D de l'IRM. Plusieurs études sont réalisées pour évaluer l'influence du bruit sur chacune des données impliquées dans le calcul du recalage. Les résultats de ces expériences sont présentés dans le tableau 5.6 en indiquant l'erreur RMS obtenue sur les points bruités. En pratique, 1000 tirages ont été effectués pour les calculs de Monte Carlo.

Incertitude liée au recalage Les résultats montrent que les incertitudes se situent entre 0.5 mm et 2.5 mm, ce qui correspond à 3 à 15 pixels dans les images US, et 1 à 4 pixels dans les images IRM. La dernière ligne du tableau 5.6 présente toute l'incertitude du recalage global, située entre 2.2 mm et 2.5 mm. Ce niveau d'incertitude permet d'expliquer les comportements problématiques observés sur la figure 5.13.b. Sur la séquence citée en section 5.3.1, sur plus de 38% des images US, la langue semblait traverser le palais. Si on tient compte de l'incertitude calculée sur la position du palais, il ne reste que 4% des images US où la langue semble le traverser (cf figure 5.16).

Le tableau 5.6 permet aussi de se rendre compte de l'influence de chaque incertitude prise indépendamment dans le calcul. Ainsi, en appliquant seulement une incertitude sur les données EM, nous arrivons déjà à plus de 2 mm d'erreur RMS. Les capteurs EM interviennent dans de très nombreuses étapes de notre système (calibrage, sonde US, repère tête, balayage tête) ce qui explique cette importante influence. Cependant, nous tenons à souligner que les capteurs impliqués dans ces recalages ne sont pas soumis à des vitesses importantes. Le niveau de bruit considéré ici en ce qui les concerne, et qui correspond à des mesures statiques, est donc pleinement justifié.

Nous avons aussi évalué l'incertitude de la résolution US. En effet, bien que nous disposons d'un fantôme pour déterminer cette résolution (cf chapitre 3), nous avons étudié l'influence de l'incertitude sur le recalage complet. Les résultats montrent qu'elle est située entre 0.3 mm et 1 mm selon la position du point dans l'image. Cette incertitude n'est pas négligeable dans notre système, et une attention particulière doit donc être portée sur cet effet.

Bruit				Incertitude	
Capteurs EM	**Pointage US**	**Résolution US**	**IRM**	**Point**	**Erreur RMS (mm)**
×				1	2.06
				2	2.14
				3	2.10
				4	2.28
		×		1	0.32
				2	0.57
				3	0.67
				4	1.02
×	×	×		1	2.50
				2	2.56
				3	2.34
				4	2.55
×	×	×	×	1	2.41
				2	2.32
				3	2.26
				4	2.53

TAB. 5.6 – Incertitudes de recalage en ajoutant aux données un bruit blanc gaussien de moyenne nulle et d'écart-type correspondant à la précision de la méthode concernée. Les 4 indices de la colonne Point correspondent à une expérience particulière où seules certaines données d'entrée sont bruitées. Ces données bruitées sont identifiées par une croix dans leur colonne.

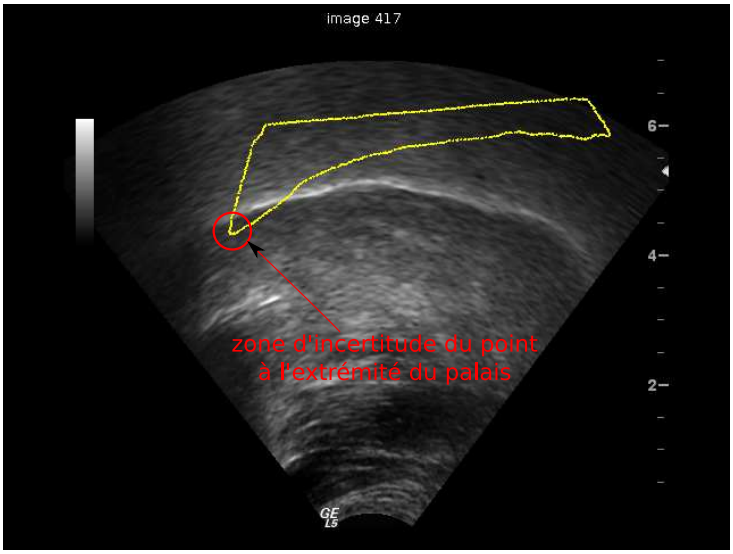


FIG. 5.16 – Incertitude de 20 pixels du point à l'extrémité avant du palais.

La troisième ligne du tableau 5.6 prend en compte l'incertitude du calibrage EM/US et des capteurs EM. Elle est similaire à l'incertitude du recalage lorsque toutes les modalités sont considérées. L'incertitude sur les données IRM a une influence négligeable dans notre système,

car le recalage par ICP permet de compenser le bruit sur les données IRM.

Les données EM et le calibrage EM/US constituent donc un point sensible de notre système, et leur incertitude influe considérablement sur l'incertitude globale du recalage.

Incertitude intégrant les traitements Les incertitudes calculées concernent le recalage. Pour avoir une idée de l'incertitude globale de notre système, nous pouvons composer avec les incertitudes des traitements appliqués, à savoir l'extraction des contours de langue dans les images US. Pour cela, nous avons effectué le même calcul, mais en composant avec une incertitude de 1.5 mm sur les positions des points P de l'image d'origine. Elle correspond à l'incertitude de l'extraction des contours de langue dans les images US. Les résultats sont présentés dans le tableau 5.7.

Bruit		**Incertitude**	
Recalage	**Extraction US**	**Point**	**Erreur RMS (mm)**
×	×	1	3.35
		2	3.20
		3	3.41
		4	3.49

TAB. 5.7 – Incertitude globale du système : incertitude de recalage et incertitude d'extraction des données des images US. Les 4 indices de la colonne Point correspondent à 4 positions de points de la langue et du palais dans les images.

On peut aussi calculer la propagation de cette erreur suivant chaque axe tridimensionnel, pour connaître la dispersion de l'incertitude du calcul sur chacun de ces axes. Les résultats sont présentés sur la figure 5.17.

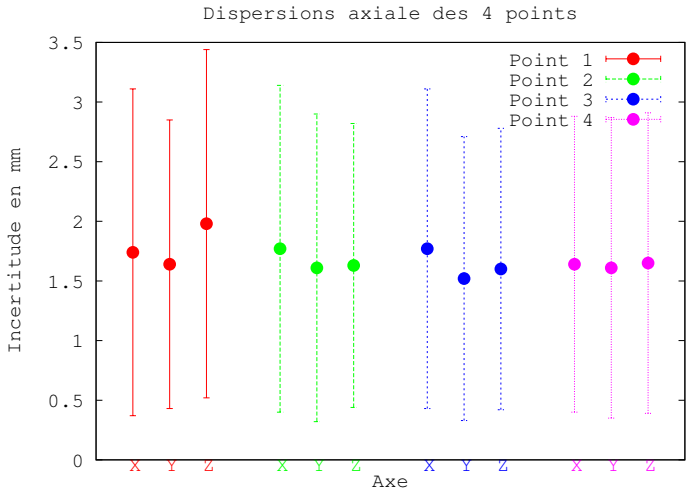


FIG. 5.17 – Incertitudes du chacun des axes de chacun des quatre points. La valeur moyenne de l'incertitude est représentée par un point, et son écart-type par la barre d'erreur associée à chaque point.

Il apparaît que l'incertitude globale de notre système, pour superposer une courbe US extraite automatiquement dans les images US à un palais extrait de l'image IRM, est comprise entre

3.2 mm et 3.5 mm. Eu égard au nombre important de modalités, de traitements et de transformations impliquées dans ce calcul, et des incertitudes associées à chaque étape, ces chiffres ne sont pas surprenants. Sur la figure 5.17, on observe que l'incertitude de chaque point est équirépartie sur chaque axe.

5.3.2.4 Conclusion

Nous avons présenté une méthode permettant d'évaluer les incertitudes de recalage et de notre système complet, basée sur une analyse statistique d'un bruit appliqué aux données.

Idéalement, un fantôme dynamique visible à la fois à l'IRM, dans les images US et utilisable avec des capteurs EM et des caméras de stéréovision sur lequel tous les mouvements et positions pourraient être contrôlés et quantifiés permettrait d'évaluer précisément l'erreur moyenne de notre système. Pagoulatos [PHK00] qui utilise un fantôme spécialement manufacturé pour son application de recalage US/IRM obtient une erreur moyenne de 1.88 mm pour des acquisitions statiques. Ce fantôme laisse apparaître des structures visibles à la fois à l'US et à l'IRM afin de pouvoir utiliser une méthode de recalage basé image. Cette méthode est supposée bien plus précise qu'un recalage utilisant une tierce modalité, comme un système EM, qui possède en outre des imprécisions de mesure de l'ordre du millimètre.

Ne disposant pas d'un tel fantôme pour comparer nos mesures à une réalité terrain, nous ne pouvons dans notre cas que donner une valeur d'incertitude moyenne sur le recalage. Si on ne considère que les incertitudes liées aux différents recalages, un point identifié dans l'image US peut être localisé en 3D dans l'IRM à 3 mm près en moyenne.

En ajoutant les traitements en amont et en aval de la chaîne de recalage, une incertitude globale de 3.5 mm caractérise notre système d'acquisition de données. Ces résultats constituent en tout état de cause les premières mesures globales d'incertitude annoncées pour un système d'acquisition de données multimodales.

Nous avons mis en avant l'influence des capteurs EM et du calibrage EM/US, notamment du calcul de la résolution des images US, dans l'incertitude finale du recalage. Ces deux aspects constituent les points clés pour le recalage des différentes modalités. Les efforts à faire pour améliorer notre système doivent donc d'abord se concentrer sur ces deux aspects.

Il faut enfin noter que ces résultats proviennent de données acquises sur un seul locuteur. Il serait approprié d'effectuer les mêmes tests sur de nouveaux locuteurs pour étudier les éventuelles variations de ces résultats. Nous pensons en effet qu'ils peuvent varier suivant les acquisitions et tous les paramètres qui y sont attachés : capacité du sujet à rester immobile durant le balayage EM de la surface du visage lors des acquisitions dynamiques, positions de sa tête vis-à-vis du générateur de champ EM, échogénicité...

5.4 Conclusion

Ce chapitre présente la dernière étape nécessaire à l'élaboration d'un système d'acquisition de données articulatoires statiques et dynamiques. Nous avons d'abord montré qu'il devait y avoir phonation pour effectuer l'acquisition statique d'un articulateur. Un protocole d'acquisition IRM original a été mis en place pour satisfaire cette contrainte. Ensuite toutes les données IRM ont été recalées dans un même repère. Nous avons alors proposé une solution de recalage de ces données statiques avec les données dynamiques en utilisant la surface du visage du locuteur comme surface intermédiaire. Les incertitudes du système ont enfin été quantifiées.

Il n'est pas possible de comparer ce travail à d'autres recherches similaires, puisque c'est le seul proposant à la fois des données articulatoires statiques et dynamiques. Nous nous sommes cependant attachés à évaluer les incertitudes à tous les niveaux afin de pouvoir améliorer notre système et le comparer à de futurs travaux similaires.

Nous verrons dans le chapitre 6 que malgré la valeur de précision de 3.5 mm trouvée sur l'incertitude globale du recalage, des résultats pertinents ont été produits par nos partenaires de parole à partir de ces données.

Le caractère automatique des méthodes présentées permet d'obtenir rapidement un grand volume de données articulatoires sur plusieurs locuteurs, afin de pouvoir mettre en place des études intra et inter-locuteurs. L'ensemble de ces données traitées représente une base de données articulatoires, où les données sont temporellement synchronisées et spatialement recalées.

Le dernier chapitre de ce manuscrit présente l'ensemble des données acquises avec notre système sur plusieurs locuteurs. Il s'intéresse ensuite à l'utilisation de ces données dans un contexte de parole. Cette ultime étape constitue une évaluation des données utilisées dans leur contexte applicatif et nous permettra de conclure sur de futures perspectives pouvant faire suite à ce travail.

Chapitre 6

Base de données articulatoires, évaluation et perspectives

Ce travail de thèse s'inclut dans un projet européen (ASPI) dont l'un des objectifs était d'acquérir un important volume de données sur plusieurs locuteurs. Ces données doivent permettre l'étude de méthodes sur l'inversion acoustique articulatoire. Pour cela, ces méthodes s'attachent à générer des formes de conduit vocal dans le plan médiosagittal à partir du signal acoustique. La base de données acquise doit alors permettre de valider les formes obtenues.
Ce dernier chapitre présente l'ensemble des données statiques et dynamiques acquises avec notre système, pour ensuite évaluer ces données traitées dans un cadre applicatif en parole. Nous concluons par des perspectives de recherches pouvant faire suite à ce travail.

6.1 Base de données articulatoires

6.1.1 Données dynamiques

Les données dynamiques ont été acquises sur trois locuteurs : un Français (cf tableau 6.1) et deux Suédois (cf tableau 6.2). Pour le locuteur français, deux sessions d'acquisition ont eu lieu.

Certaines de ces acquisitions ne comportent pas de données de stéréovision. Ceci est dû au temps d'enregistrement de ces données : en effet, après chaque acquisition dynamique (15 secondes au maximum à cause de l'échographe), le temps de sauvegarde des images de stéréovision était supérieur à 3 minutes. Nous avons donc pris la décision d'acquérir, pendant ces temps de sauvegarde, de nouvelles données US, EM et audio en excluant celles de stéréovision. Depuis peu, cette contrainte de temps de sauvegarde a été levée avec l'achat d'un nouveau matériel permettant la sauvegarde en temps réel des données de stéréovision.

Locuteur	Session	Corpus	Temps	# images US	# images stéréovision
Fabrice	1	112 phrases	5min45sec	22425	41400 × 2
		46 phrases	2min30sec	9750	-
		VCV et VV	1min30sec	5850	10800 × 2
	2	phonèmes	5min	4875	36000 × 2
		phonèmes	4min30sec	17550	-

Tab. 6.1 – Données dynamiques enregistrées sur le locuteur français.

Locuteur	Session	Corpus	Temps	# images US	# images stéréovision
Anne-Marie	3	152 phrases	5min45sec	22425	41400×2
		VCV	5min	19500	36000×2
Olov	3	109 phrases	5min15sec	20475	37800×2
		VCV	5min	19500	36000×2

TAB. 6.2 – Données dynamiques enregistrées sur les locuteurs suédois.

Les tableaux 6.1 et 6.2 montrent que notre base est constituée de 142350 images US (36 minutes et 30 secondes), et de 239400×2 images de stéréovision (33 minutes et 15 secondes). Elle comprend aussi des données EM pour la langue (environ 2 images US sur 3 à cause de la différence de fréquence pour ces deux modalités, 40 Hz vs. 66 Hz), et pour la position sonde et tête (interpolées pour chaque image US, cf chapitre 4). Enfin, chaque séquence est accompagnée de sa donnée audio. Toutes ces données ont été automatiquement synchronisées (cf chapitre 3), recalées (cf chapitre 4 pour les données EM/US et chapitre 5 pour le recalage des données stéréo et EM), et traitées (cf chapitre 4 pour l'extraction de la surface de la langue dans les images US).

On se rend bien compte que le temps de traitement manuel de ce volume très important de données est inimaginable et que l'automatisation des acquisitions et des traitements trouve ici toute sa justification.

Pour chaque locuteur ayant participé à une acquisition de données dynamiques, la surface de son visage a été numérisée pour être en mesure de recaler les données dynamiques sur cette surface de référence (cf chapitre 5).
Enfin, lors de chaque session d'acquisition dynamique, le visage du locuteur est balayé avec le stylet EM afin de pouvoir recaler le repère EM tête avec le repère défini par le numériseur 3D (cf chapitre 5).

6.1.2 Données statiques

Les données IRM ont été acquises au service de neuroradiologie du CHU de Nancy suivant le protocole décrit au chapitre 5. Les mêmes locuteurs acquis avec le système dynamique ont aussi été acquis à l'IRM (un Français, cf tableau 6.3 et deux Suédois, cf tableau 6.4). Deux locuteurs français supplémentaires, un homme et une femme, ont aussi été acquis à l'IRM (cf tableau 6.3), et le seront prochainement avec le système dynamique.

N'ayant jamais accès à la machine IRM pour plusieurs heures consécutives, plusieurs sessions d'acquisition ont été organisées pour acquérir tous les phonèmes souhaités. Elles sont numérotées de 1 à 5 dans les tableaux présentés. Les phonèmes acquis sont également spécifiés en utilisant la notation de l'alphabet phonétique international[16]. Pour la coarticulation, les consonnes sont tenues en maintenant l'occlusion tout en pensant à la voyelle suivante : pour un /ʃ/ en contexte /a/ par exemple, le locuteur commence à prononcer le /ʃ/ en pensant au /a/ suivant, comme dans le mot « chat ».

Pour un même locuteur, les données IRM sont recalées sur l'IRM de référence en utilisant le critère de l'information mutuelle (cf chapitre 5). On extrait enfin la surface du palais à partir de ces données recalées (cf chapitre 5).

[16] `http://www.langsci.ucl.ac.uk/ipa`

Locuteur	Numéro session	Phonèmes
Fabrice	1	IRM référence voyelles : /*i*/ /*u*/ /*e*/ /*ɛ*/ /*o*/ /*œ*/ /*y*/ /*o* / /*ɔ*/ coarticulation : /*ʃ*/ et /*s*/ en [i u] /*k*/ en [a i u] consonnes : /*l*/
	2	IRM référence voyelles : /*i*/ /*ɛ*/ /*a*/ /*ɔ*/ coarticulation /*p*/ et /*t*/ en [a i] /*k*/ en [i a]
	4	IRM référence coarticulation : /*l*/ et /*r*/ en [a i u] /*ʃ*/ et /*s*/ en [a] /*f*/ en [a u] nasales : / *a*/ / *e*/
Yves	2	IRM référence voyelles : /*i*/ /*e*/ /*ɛ*/ /*a*/ /*œ*/ /*o*/ /*u*/ /*y*/ / *o*/ /*ɔ*/
	4	IRM référence co-articulation : /*p*/ /*t*/ /*k*/ /*s*/ /*ʃ*/ /*l*/ en [a i u]
Amélie	5	IRM référence voyelles : /*e*/ /*o*/ /*y*/ /*ɔ*/ /*a*/ /*ɛ*/ /*O*/ /*u*/ / *o*/

TAB. 6.3 – Données statiques enregistrées sur les locuteurs français.

Locuteur	Numéro session	Phonèmes
Olov	3	IRM référence voyelles : /*a*/ /*i*/ /*u*/ /*o*/ /*y*/ /*ü*/ consonnes : /*sj*/ /*rs*/ /*tj*/ /*s*/ /*t*/ /*k*/ /*l*/
Anne-Marie	3	IRM référence voyelles : /*a*/ /*i*/ /*u*/

TAB. 6.4 – Données statiques enregistrées sur les locuteurs suédois.

6.1.3 Bilan des acquisitions

Les données statiques ont été recalées avec les données dynamiques suivant la méthode exposée dans le chapitre 5. Ainsi, pour chaque locuteur, nous disposons d'un ensemble de données traitées, synchronisées et recalées, caractérisant les formes et mouvements des articulateurs du conduit vocal.

Le système dynamique mis au point au cours de ce travail de thèse nous a permis de nous forger une expérience dans sa mise en place et son utilisation. En effet, des événements inattendus sont parfois intervenus sur le système d'acquisition de données dynamiques au cours de ce travail. Parmi eux, nous pouvons citer :

- des images US de mauvaise qualité dues à un locuteur ayant une mauvaise échogénicitée, ou tout simplement comme nous l'avons vu dans le chapitre 3, des sons qui génèrent des contours invisibles à l'image US ;
- le plan US visé s'éloignant du plan médiosagittal du locuteur. Une attention particulière

doit en effet être portée sur la position de ce plan durant les acquisitions, car il est facile à cause du gel et des mouvements du locuteur de dévier de la position médiosagittale ;
- un capteur EM qui se décolle lors d'une acquisition sans que l'on s'en soit rendu compte. Toutes les données EM et US acquises après cet événement ne peuvent plus être alors recalées dans le repère tête, rendant impossible leur exploitation ;
- le disque dur de l'échographe saturé par un volume trop important de données enregistrées. Cela nécessite de reporter la fin de l'acquisition de quelques heures, car le transfert des 40 Go de données DICOM peut facilement prendre plus de deux heures ;
- la présence sur le système d'exploitation du PC de contrôle d'outils de mise à jour que nous avions oublié de désactiver, et qui se mettent en route lors des acquisitions, faussant les délais de synchronisation.

Ces événements nécessitent une attention particulière, et nous ont permis de progressivement rendre le système plus robuste à ces aléas à chaque session d'acquisition de données dynamiques. Pour les acquisitions statiques, il peut aussi survenir des impondérables liés au comportement des locuteurs dans la machine : certains d'entre eux ont naturellement tendance à beaucoup bouger lors des acquisitions, rendant difficile, voire impossible, l'exploitation des images résultantes. D'autres se sentent vite très mal à l'aise dans la machine et préfèrent en sortir avant qu'un corpus significatif ait été acquis (cas de la locutrice suédoise du tableau 6.4).

Ces aléas rendent ardue la mise en place d'un système d'acquisition de données, très souvent soumise à des imprévus dont certains ont pour conséquence de faire échouer toute la session ou de rendre impossible l'exploitation des données.
Des solutions ont été proposées au cours de ce travail, comme la mise en place d'un outil de visualisation en temps réel du plan US par rapport aux positions capteurs (cf chapitre 4). Cet outil, utilisé pendant les acquisitions dynamiques, permet d'aider le manipulateur à rester proche du plan médiosagittal, et ainsi d'améliorer les acquisitions US.
D'autres éléments, comme un décollement des capteurs tête, peuvent être vérifiés directement ou a posteriori en calculant la distance entre les deux capteurs. Si cela est fait lors des acquisitions, le manipulateur sur le PC de contrôle est averti lorsque cette distance varie au cours de l'acquisition. Nous songeons aussi à un système EM acceptant plus de capteurs que celui utilisé pour en fixer davantage sur la tête du locuteur, afin d'introduire de la redondance dans les mesures données. Une telle redondance permettrait de rendre les mesures du système EM plus robustes. Nous sommes actuellement en train de tester un tel système.
Des vérifications a posteriori ont été effectuées sur les données acquises, par exemple le calcul de la position du plan US par rapport au plan médiosagittal lors du recalage. Lorsque cette distance était trop importante, les données correspondantes ont été retirées. Il peut s'agir soit d'une partie de la séquence soit de la séquence complète.

Les données statiques et dynamiques présentées dans ce chapitre sont celles fournies à nos partenaires du projet ASPI, qui vont désormais les utiliser pour tester leur méthode d'inversion acoustique articulatoire. Ce travail a donc permis de constituer avec succès une base de données statiques et dynamiques sur plusieurs locuteurs. Cette base sera par la suite enrichie en la complétant par de nouvelles données sur les mêmes locuteurs et sur des nouveaux.

Bien qu'ayant calculé dans le chapitre 5 les incertitudes géométriques de ces données recalées, nous devons maintenant savoir si ces incertitudes sont acceptables pour que nos données puissent être utilisées dans un cadre applicatif en parole. Pour cela, poursuivant la démarche exposée au chapitre 2, nous proposons de les évaluer en vérifiant qu'elles sont cohérentes avec les formes

du modèle de Maeda [Mae79]. Nous nous concentrons sur les données articulatoires statiques et dynamiques acquises avec le locuteur français. Le corpus acquis avec ce locuteur est décrit en annexe A.

6.2 Évaluation des données recalées sur le modèle de Maeda

Nous avons présenté dans le chapitre 1 le modèle de Maeda [Mae79], un modèle articulatoire statistique largement utilisé dans la communauté parole. Il décrit bien les formes de conduit vocal dans le plan médiosagittal pour les voyelles, et permet, couplé à une simulation acoustique, de générer le son correspondant à une forme [Mae82].
Nous nous sommes donc naturellement tournés vers ce modèle pour savoir si les formes de notre base de données étaient cohérentes avec celles du modèle. De plus, avec nos données, nous ne disposons que d'une partie de la langue dans les images US. Nous avons aussi la position du palais grâce au recalage US/IRM et les positions des lèvres avec les données de stéréovision. En vérifiant que le modèle de Maeda peut s'adapter à nos données, nous vérifierons aussi que ces données couvrant partiellement le conduit vocal sont suffisantes pour retrouver une forme du conduit dans le plan médiosagittal. L'idée sous-jacente est de vérifier que nos données permettent de contrôler les déformations de conduit vocal [BEB+07].
Ce travail a été réalisé en collaboration avec l'équipe Parole du LORIA.

6.2.1 Méthode

Le modèle de Maeda [Mae79] a été construit à partir de contours détourés manuellement dans des images rayons X d'un locuteur féminin. Une grille semi-polaire, dessinée sur la figure 6.1, a été utilisée pour obtenir les points d'intersection de cette grille avec les contours médiosagittaux du conduit vocal. Une analyse en composantes principales sur les points d'intersection a permis d'obtenir les sept composantes linéaires principales guidant les plus importantes déformations du conduit vocal (cf chapitre 1).
L'objectif de ce travail est de retrouver une combinaison linéaire de ces sept composantes en ajustant ce modèle linéaire à nos données articulatoires. Nous devons aussi prendre en compte l'aspect temporel afin que les déformations du modèle aient un comportement temporel réaliste. Nous détaillons ici toutes les étapes nécessaires à ce travail et explicitées dans notre article [ATB+09].

6.2.1.1 Adaptation du modèle au locuteur

Le modèle de Maeda a été établi sur le conduit vocal d'un locuteur féminin dont la morphologie est sensiblement différente de celui d'un locuteur masculin (pharynx moins long...). Il est donc nécessaire d'adapter le modèle à un locuteur masculin sur lequel sera testée la méthode. Pour cela, la grille semi-polaire sur laquelle apparaît la paroi externe du conduit vocal (articulateur fixe dans le modèle de Maeda) est manuellement superposée à la coupe IRM médiosagittale de l'IRM référence de notre locuteur. Ce placement manuel est réalisé de telle façon que la grille et la paroi externe du conduit vocal se superposent au mieux à l'image en jouant sur la translation, la rotation, et les paramètres d'échelle de la grille.

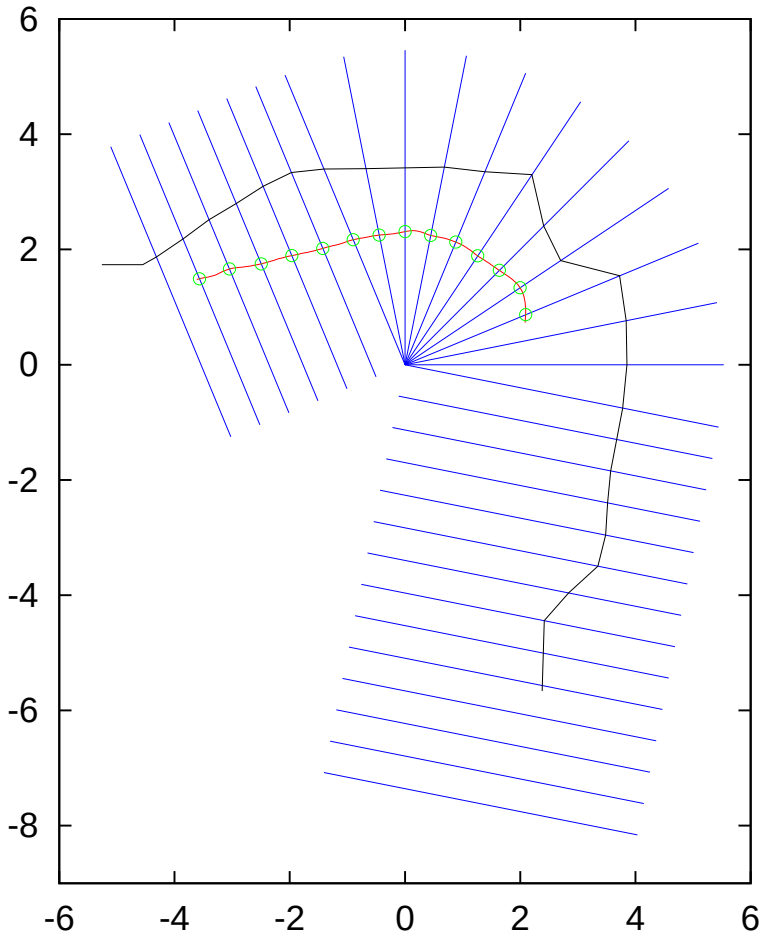


FIG. 6.1 – Grille semi-polaire de Maeda, paroi externe du conduit vocal, et contours US de langue lors d'un /a/.

6.2.1.2 Des contours US aux données de la langue

Pour obtenir les données de la langue nécessaires au modèle de Maeda, les contours de langue extraits des images US, recalés dans le repère IRM et projetés dans le plan médiosagittal de la grille sont ensuite utilisés. Ces intersections sont représentées par des cercles verts sur la figure 6.1.

6.2.1.3 Des données de stéréovision aux données des lèvres

Les données de stéréovision sont aussi utilisées pour récupérer la position de la mâchoire, et les valeurs de l'ouverture, de l'écartement et la protrusion des lèvres. Ces données proviennent directement des marqueurs dessinés sur le visage du locuteur lors des acquisitions.

6.2.1.4 Des données aux paramètres articulatoires

Dans le modèle de Maeda, les données géométriques et les paramètres articulatoires sont reliés linéairement :

$$\mathbf{v} = \mathbf{Ap} \tag{6.1}$$

où $\mathbf{v}$ est un vecteur de dimension 29 correspondant aux éléments de la grille de Maeda ; $\mathbf{A}$ est la matrice des composantes linéaires de l'ACP de dimension 29×6 ; et $\mathbf{p}$ est un vecteur de dimension 6 décrivant les paramètres articulatoires (position de la mâchoire, position du dos de la langue, forme du dos de la langue, position de l'apex, ouverture des lèvres et protrusion). Le paramètre du larynx est exclu, car nous ne disposons pas d'information à son propos avec nos données. Toutes les données utilisées dans le vecteur $\mathbf{v}$ sont centrées et normalisées.
Étant donné l'ensemble C des données $\mathbf{v}$ à un instant donné, nous pouvons approcher les paramètres articulatoires leur correspondant en minimisant la quantité suivante :

$$I_s(p) = \sum_{i \in C} \left(v_i - \sum_{j=1}^{6} a_{i,j} p_j \right)^2 \tag{6.2}$$

où $a_{i,j}$ sont les éléments de la matrice $\mathbf{A}$ et p_j les éléments du vecteur $\mathbf{p}$. La contrainte :

$$p_j \in [-3, 3], j = 1, \ldots, 6 \tag{6.3}$$

est ajoutée pour que les coefficients appliqués au modèle engendrent des formes plausibles. La minimisation de l'équation 6.2 sous la contrainte de l'équation 6.3 est un problème de programmation quadratique avec contraintes linéaires résolu dans [Fle87]. Elle permet d'obtenir un ensemble de paramètres articulatoires correspondants à nos données, chacun étant calculé pour un instant donné.
Afin d'assurer une cohérence temporelle des trajectoires articulatoires, une technique de régularisation [Bon93] est utilisée sur ces paramètres articulatoires. Elle consiste à minimiser un fonction de coût, faisant apparaître la quantité de l'équation 6.2 tout en contrôlant les variations temporelles du vecteur de paramètres $\mathbf{p}$ (vitesse et accélération). Cette régularisation temporelle consiste donc, sur l'intervalle de temps $[t_s, t_f]$ considéré, à minimiser la fonction de coût suivante :

$$I_d(p) = \int_{t_s}^{t_f} \sum_{i \in C} \left(v_i(t) - \sum_{j=1}^{6} a_{i,j} p_j(t) \right)^2 dt + \lambda \int_{t_s}^{t_f} \sum_{j=1}^{6} p'_j(t)^2 dt + \beta \int_{t_s}^{t_f} \sum_{j=1}^{6} p_j(t)^2 dt \tag{6.4}$$

Les trois intégrales de cette équation caractérisent respectivement : la distance entre les variables observées et générées par le modèle ; la vitesse de changements des paramètres articulatoires ; et l'effort articulatoire. Les constantes λ et β sont choisies empiriquement. Pour plus de détails sur la technique de régularisation temporelle des paramètres articulatoires, nous invitons le lecteur à se reporter aux travaux de Laprie [LM98].

6.2.2 Résultats

Cette méthode a été testée sur une séquence VV de notre locuteur Fabrice. La figure 6.2 présente la transition entre un */a/* et un */e/* provenant de la séquence VV */ae/*. Les courbes de langue issues des images US sont superposées aux formes de conduit du modèle de Maeda.

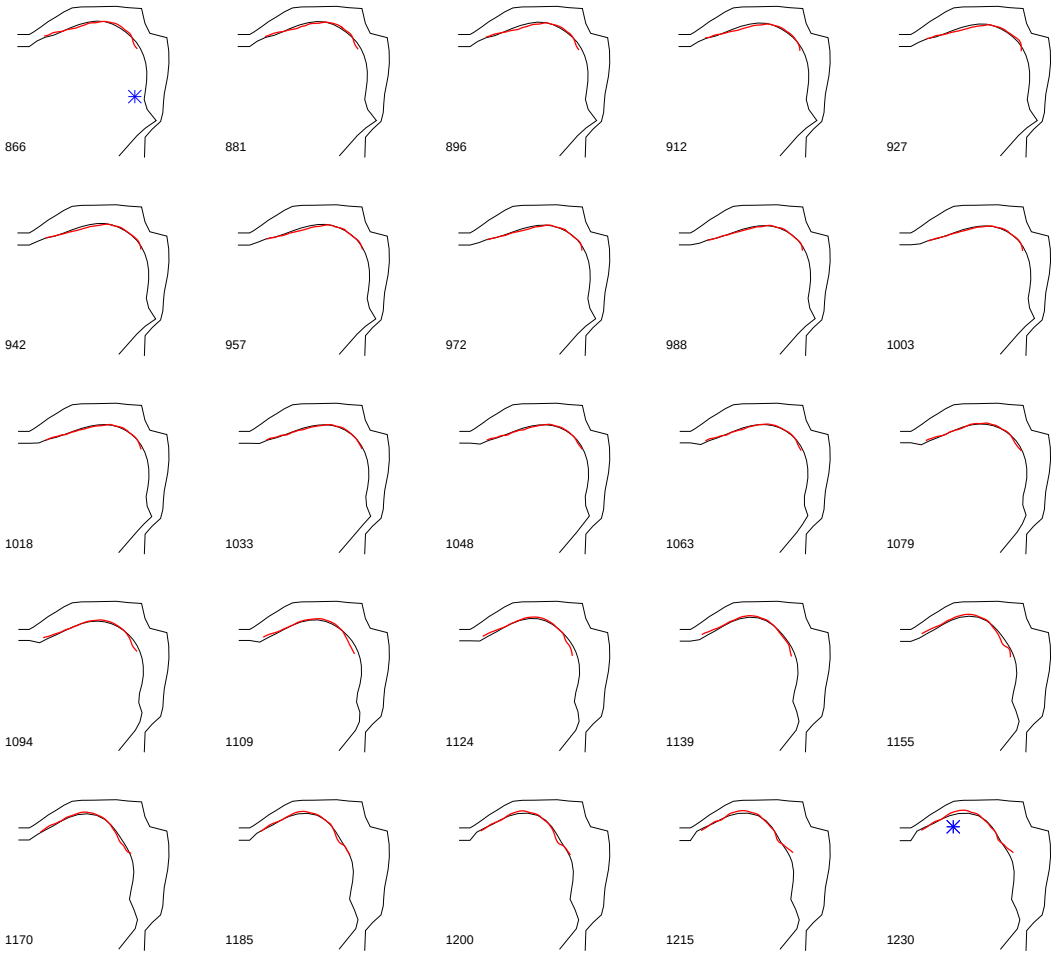


FIG. 6.2 – Contours US superposés au modèle de Maeda ajusté à ces données pour la séquence VV */ae/*. Le temps en millisecondes est indiqué sur chaque forme. Les astérisques indiquent le lieu de constriction.

On observe, pour cette transition VV, que les formes de la langue US et du modèle se correspondent. Puisque ces données utilisent le recalage proposé dans le chapitre 5 (palais IRM et contour de langue US), cela nous permet de vérifier que la précision de ce recalage semble suffisante.
De plus, ces données permettent de générer un ensemble de formes cohérentes sur le plan médiosagittal du conduit vocal : ce dernier a l'apparence physique d'un conduit pour un */ae/*. Les informations partielles sur les articulateurs utilisées pour cette étude semblent suffisantes pour contrôler les paramètres d'un modèle articulatoire existant.

La séquence de la figure 6.3 présente une transition VV */ay/*. Sur cette séquence, on observe que le modèle a plus de difficultés à s'ajuster à nos données, notamment pour la dernière forme de langue. Il ne s'agit pas ici d'un problème de recalage qui se traduirait par un décalage constant entre la forme US et le modèle, mais la forme n'est pas atteignable par le modèle. Cela ouvre le champ à deux possibilités : soit l'extraction de la surface de la langue dans les images US est

incorrecte, soit le modèle de Maeda ne permet pas de décrire cette forme.
En visualisant la séquence US, il n'y a pas d'ambiguïté visuelle : la forme de langue correspond bien à la forme extraite. Il n'est donc pas possible de générer cette forme par une combinaison linéaire des paramètres articulatoires du modèle de Maeda. Cela nous amène donc naturellement à reconsidérer la construction de ce modèle.

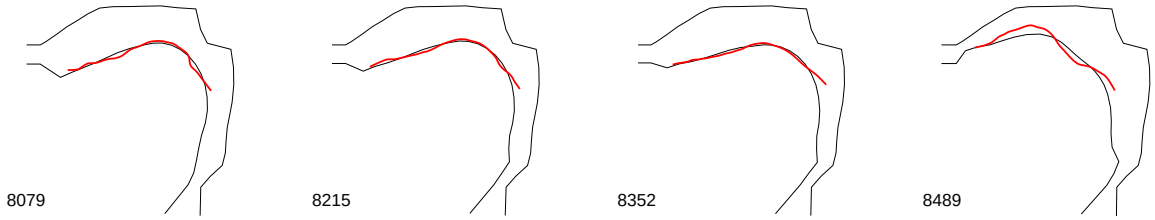


FIG. 6.3 – Contours US superposés au modèle de Maeda ajusté à ces données pour la séquence VV */ay/*. Le temps en millisecondes est indiqué sur chaque forme.

6.2.3 Le modèle articulatoire de Maeda : critiques

Les données ayant été utilisées pour la construction du modèle de Maeda proviennent de 1000 croquis décrivant les positions des articulateurs dans le plan médiosagittal du conduit vocal, réalisés dans les années quatre-vingt par l'Institut Phonétique de Strasbourg [WZBS86]. Ces croquis ont été dessinés à partir de données cinéradiographiques acquises sur une seule locutrice, et les formes de conduit ont été segmentées manuellement. Le corpus était constitué de phrases phonétiquement équilibrées.
Nous avons vu dans le chapitre 1 que les images rayons X ne présentent pas une vue du plan médiosagittal, mais la superposition de coupes sagittales de la tête. En raison de ces superpositions, il est difficile de déterminer sur ces images le contour médiosagittal de la langue(cf figure 6.4). Les détourages effectués par des phonéticiens sur les croquis sont donc basés à la fois sur une estimation visuelle des contours médiosagittaux du conduit vocal, et sur leurs connaissances a priori. Mais l'absence de concavités dans ces croquis montre que ces détourages ne sont pas corrects. La présence de concavités dans le plan médiosagittal, à la fois sur les données IRM et US de notre base de données, corrobore cette constatation. Le modèle de formes de Maeda ne peut donc pas s'ajuster à des formes concaves de langue, comme avec le */y/* par exemple.

Par ailleurs, le support utilisé par le modèle pour la mise en correspondance des points du conduit est la grille semi-polaire présentée sur la figure 6.1. Lorsqu'un point de contour n'a pas d'intersection avec la grille, un point sur le plancher buccal est choisi. La figure 6.5 présente une telle situation, où la langue n'a pas d'intersection avec le premier élément de la grille.

Il apparaît ici une autre faiblesse de ce modèle : les points utilisés pour l'ACP ne ne correspondent pas physiquement. Il suffit que la langue subisse une translation horizontale pour que les points de la grille soient physiquement différents. Ainsi, le modèle suppose que les déformations du conduit s'effectuent dans la direction de chaque élément de la grille, ce qui est physiquement faux.
La mise en correspondance de points physiques est un problème difficile, et particulièrement sur

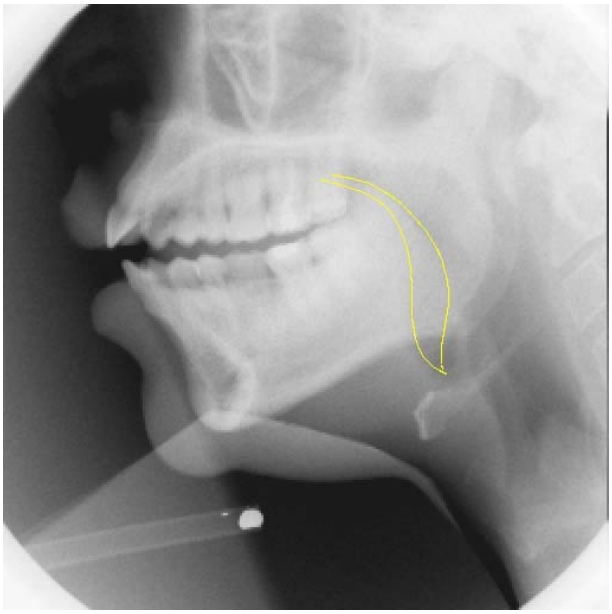


FIG. 6.4 – Exemple d'image rayons X où deux contours de langue (délimités en jaune) sont visibles.

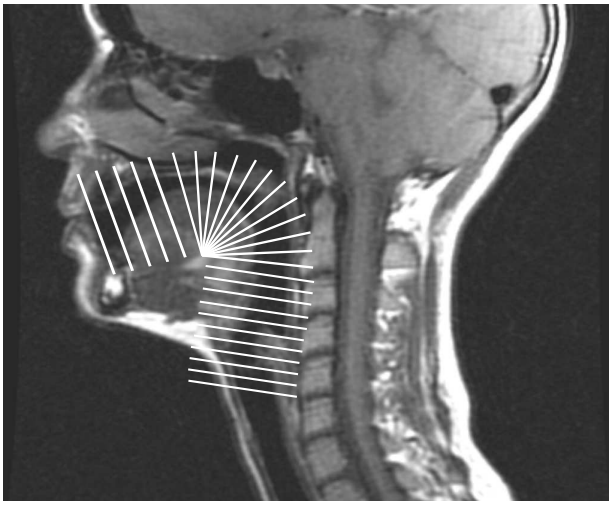


FIG. 6.5 – Exemple d'image où tous les éléments de la grille à l'avant du conduit n'ont pas de point d'intersection avec la langue. Le son prononcé est un /o/.

la langue de par sa nature déformable et élastique. Nous proposons dans la dernière partie de ce travail de thèse une autre approche que celle de Maeda pour générer un modèle de langue physiquement plus cohérent, en nous basant sur les coupes IRM médiosagittales.

6.3 Vers un nouveau modèle de déformations de langue

Nous présentons un travail actuellement en cours dans notre laboratoire. Les résultats obtenus avec le modèle de Maeda nous ont naturellement amenés à chercher à mettre en place un nouveau modèle de déformations de la langue. Des résultats préliminaires sont exposés.

6.3.1 Principe

Nous nous sommes basés sur les données statiques IRM pour établir un nouveau modèle de formes de la langue, car les coupes médiosagittales IRM offrent la possibilité de voir sans ambiguïté les cavités de la langue. Afin de prendre en compte l'élasticité et les déformations de la langue, nous avons cherché une méthode pour que le support permettant d'effectuer l'ACP sur les points de contour soit adapté à chaque position de la langue.

Pour cela, les contours de la langue dans les coupes IRM médiosagittales sont manuellement détourés, du point de contact entre la langue et le plancher de la cavité buccale à l'avant, à sa base au niveau de l'épiglotte. L'abscisse curviligne est alors calculée sur ce contour pour le discrétiser en n points (40 en pratique) régulièrement espacés. Ensuite, une ACP est effectuée sur ces points de contours.

Cette méthode a pour avantage de prendre en compte les déformations élastiques de la langue dans toutes les dimensions, qui ne sont plus contraintes dans une seule direction comme avec la grille semi-polaire de Maeda. On peut donc espérer que la correspondance physique des points soit mieux préservée, même si elle reste physiquement inexacte.

6.3.2 Résultats

Les 36 coupes médiosagittales du locuteur Fabrice ont été traitées. Les résultats de l'analyse en composantes principales sont détaillés dans le tableau 6.5.

Mode	1	2	3	4	5	6
Variance expliquée en %	50.4	28.1	10.4	4.7	3	1.3
Variance cumulée en %	50.4	78.5	88.9	93.6	96.6	97.9

Tab. 6.5 – Variance et variance cumulée pour les 6 premiers modes de l'ACP sur les contours de langue.

Les quatre premiers modes de l'ACP expliquent plus de 90% de la variance totale de nos données. Comme dans le modèle de Maeda, chaque mode semble avoir une interprétation physique des déformations de la langue. Cependant, il est plus difficile de distinguer clairement le rôle de chaque composante. La position de l'apex est, par exemple, influencée par chacun des modes. On peut cependant les décrire de la façon suivante : le premier mode décrit prioritairement les variations de position du corps de la langue, le second celles de la forme du dos de la langue, le troisième celles de position de la racine de la langue et le quatrième concerne essentiellement les variations de position de l'apex. L'influence de ces principaux modes peut être observée sur la vidéo présentée sur `http://www.loria.fr/~aron/these.html`.
Conformément à ce qui était attendu avec l'utilisation de données IRM, notre modèle laisse apparaître des formes avec des concavités sur la langue. On distingue notamment la présence d'une concavité au niveau de l'apex (cf figure 6.6), plus ou moins prononcée suivant la combinaison linéaire des vecteurs propres de l'ACP, et qui est inexistante dans le modèle de Maeda.

6.3.3 Utilisation du modèle de langue sur le suivi

Ce modèle de langue, établi seulement à partir de données statiques, est ensuite utilisé sur les données dynamiques. L'objectif est d'estimer si un tel modèle suffit à décrire toutes les formes dynamiques adoptées par la langue, ou si le modèle de déformations doit être complété.

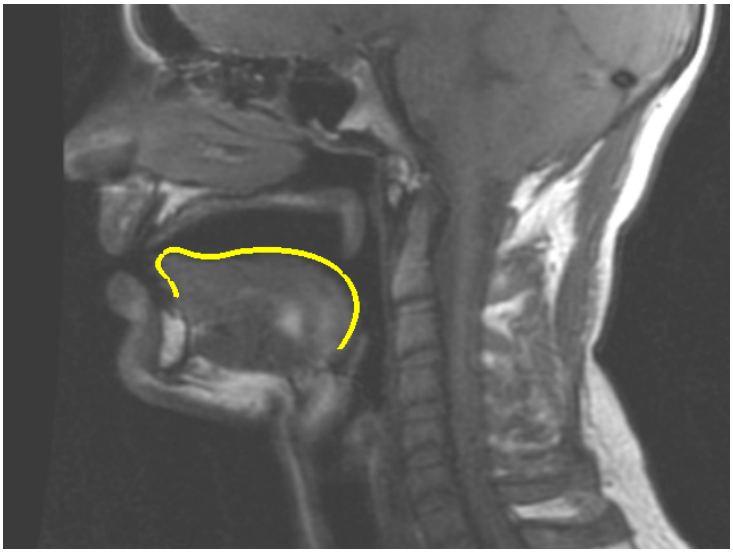
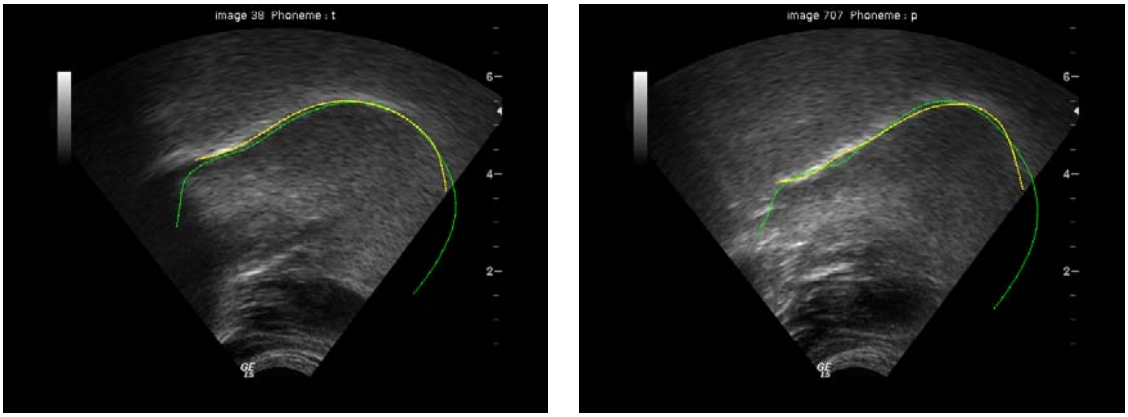


FIG. 6.6 – Exemple de forme obtenue avec notre modèle de langue, avec notamment une concavité au niveau de l'apex. Le modèle est affiché sur une image IRM pour une meilleure visualisation.

Pour cela, nous pouvons reprendre le principe de l'ajustement du modèle aux données US vu précédemment. Cet ajustement est effectué sur le modèle de langue présenté ci-dessus. Des résultats préliminaires sont visibles sur la figure 6.7 : aucune régularisation temporelle des paramètres n'a été effectuée afin d'observer le comportement purement géométrique de notre modèle. Le suivi complet peut être vu sur la vidéo disponible à l'adresse `http://www.loria.fr/~aron/these.html`.

Ces résultats, obtenus pour des phrases où la dynamique est importante, sont prometteurs : il semblerait que le modèle de déformations de la langue construit à partir de données statiques ait un intérêt dans la prédiction de la forme de la langue lorsqu'elle a des mouvements rapides. Ils sont toutefois à confirmer en effectuant notamment une étude quantitative et qualitative, car certaines formes sont parfaitement décrites par le modèle (cf figure 6.7.a) et d'autres semblent plus difficiles à atteindre (cf figure 6.7.b).

Ce modèle sera prochainement intégré dans le processus de suivi. Nous pensons que l'intégration d'un modèle de déformations de la langue dans la contrainte de mouvement par le calcul du flot optique peut améliorer la phase de prédiction. En effet, nous utilisons pour le moment la contrainte d'un mouvement paramétrique affine, ce qui est un a priori moins adapté qu'un modèle de déformations appris sur des formes de langue. L'intérêt d'une telle étude est double : elle permet d'évaluer quantitativement l'apport d'un modèle de déformations établi à partir de données statiques dans un contexte dynamique. Elle permet aussi en testant le suivi sur plusieurs locuteurs de déterminer si le modèle de déformations construit à partir d'un seul locuteur peut être utilisé sur plusieurs locuteurs.



(a) (b)

FIG. 6.7 – Exemples d'ajustement du modèle (courbe verte) à la courbe de suivi (courbe jaune). (a) Premier /t/ du mot « autorisation ». (b) Second /p/ du mot « propre ».

6.4 Perspectives

6.4.1 À court terme

6.4.1.1 Calibrage EM/US

L'étape de calibrage EM/US (cf chapitre 4) a une influence conséquente sur l'incertitude globale du recalage (cf chapitre 5). Nous avons utilisé pour le calibrage EM/US un fantôme manuellement fabriqué composé d'une baguette en bois avec deux capteurs collés à ses extrémités, le tout plongé dans un bac d'eau. Ce dispositif expérimental comporte de nombreuses imprécisions, incluant l'incertitude des capteurs EM, l'incertitude de pointage du point dans les images US, et l'incertitude de résolution. Elles peuvent toutes être diminuées. Pour cela, grâce au savoir-faire acquis sur ce fantôme, nous avons fait le design d'un nouveau fantôme de type ensemble de points. Nous l'avons commandé auprès de la société CIRS Inc., et nous devrions le recevoir d'ici la fin de l'année 2009. Constitué d'un gel spécifiquement étudié pour simuler la vitesse de propagation des ultrasons dans les tissus humains, de plusieurs filaments s'entrecroisant visibles sur l'image US et pouvant être repérés avec le système EM, ce fantôme devrait améliorer la qualité de l'étape de calibrage EM/US, et par conséquent réduire les incertitudes de recalage.

6.4.1.2 Système EM à huit capteurs et à 100 Hz

Un nouveau système EM est proposé depuis fin août 2009 par NDI. Nous avons récemment fait l'acquisition de ce système qui sera prochainement intégré et testé avec notre système d'acquisition dynamique. Il possède l'avantage d'avoir une fréquence d'acquisition de 100 Hz, permettant ainsi de disposer de données EM à des fréquences plus élevées pour les capteurs langue. De plus, il offre la possibilité d'utiliser simultanément huit capteurs EM. Une fois sa précision évaluée, nous pourrons fixer un ou deux capteurs EM supplémentaires sur le visage du locuteur, afin d'ajouter de la redondance dans les données EM pour le calcul du repère tête. Cela devrait permettre de diminuer l'incertitude sur ce calcul, et contribuer ainsi à améliorer la qualité du recalage entre les données dynamiques et statiques.

6.4.1.3 Améliorations du système d'acquisition dynamique

Nous améliorons actuellement le système d'acquisition dynamique en essayant de rendre plus silencieux chacun des matériels utilisés. En effet, il a été constaté que le bruit généré par les diverses ventilations utilisées (des modalités US, EM, du PC d'acquisition, des projecteurs éclairant le visage du locuteur durant les acquisitions de stéréovision...) était audible dans les enregistrements sonores et en perturbait la qualité.

Nous pensons aussi effectuer un recalage entre les données EM et US et la surface du visage numérisée en début même de chaque session d'acquisition dynamique (et non a posteriori). Cette étape améliorerait l'interface de visualisation présentée dans le chapitre 4 où, pour le moment, seules les données EM et US sont visibles. L'ajout de la surface du visage numérisée permettrait d'améliorer le confort du manipulateur gérant la sonde US en visualisant directement la position du plan US et des données EM par rapport à la surface du visage numérisée. Par ailleurs, nous pourrions immédiatement évaluer visuellement le recalage des données dynamiques entre elles.

6.4.1.4 Traitements des données statiques

Pour le traitement des données IRM (cf chapitre 5), le faible volume des données acquises au début de ce travail a été manuellement traité (délimitation de la région d'intérêt) et les surfaces ont été extraites en utilisant des méthodes de seuillage et de reconstruction tridimensionnelles simples (marching cubes). Nous commençons à acquérir de plus en plus de données IRM et des méthodes de segmentation automatique de ces données IRM doivent être envisagées. Pour cela, des segmentations basées modèles [Lev00] du conduit vocal sont actuellement explorées dans notre laboratoire.

6.4.2 À long terme

6.4.2.1 Évaluations

Les données acquises sont aussi actuellement utilisées par l'équipe Parole de notre laboratoire pour évaluer leurs méthodes d'inversion acoustique articulatoire. Leurs travaux consistent à étudier les formes de conduit vocaux obtenues à partir du signal acoustique avec celles données par notre système d'acquisition.

L'évaluation des données de notre base en utilisant la méthode décrite dans la section 2 de ce chapitre n'a été faite que sur un petit ensemble de données sur un seul locuteur. Cela peut paraître faible compte tenu de l'importance du volume de données présentées dans la section 1, mais, hormis le fait que ces données n'ont été acquises que récemment, la visée de notre étude était préliminaire. Il nous semble désormais important d'acquérir des données sur plusieurs nouveaux locuteurs pour la consolider et étudier le comportement du recalage et de l'incertitude selon les locuteurs.

6.4.2.2 Modélisation articulatoire

Notre système rend possible l'obtention automatique de données articulatoires sur un grand nombre de locuteurs différents. Cet aspect peut être exploité pour étudier les aspects multilocuteurs de la production de la parole. Nous sommes actuellement en train de tester le comportement d'un modèle de déformations 2D de la langue pour le suivi. Dans ce cadre, il est seulement utilisé pour la prédiction de la forme de la langue. Mais on peut songer à aller plus loin en étudiant

la possibilité d'avoir un modèle de déformations générique capable de s'adapter à plusieurs locuteurs. Nous aimerions donc savoir si la mise en place d'un tel modèle est possible, et si oui, comment adapter un modèle d'un locuteur à l'autre ?

Une question que l'on se pose à l'issue de ce travail de thèse est de savoir si un modèle de déformations 2D construit à partir de données statiques permet de décrire toute la dynamique du conduit vocal. Étant donnée la richesse des mouvements possibles des articulateurs, il paraît en effet difficile d'imaginer que c'est effectivement le cas. S'il s'avérait que cette hypothèse est vraie, la question à se poser ensuite est de savoir comment compléter un modèle statique à partir de données dynamiques ? Notre système permet d'envisager une telle étude.

Nous nous sommes focalisés dans ce travail pour les acquisitions dynamiques sur le plan médiosagittal du locuteur. Mais notre système est parfaitement adaptable à d'autres types d'acquisitions, et plus particulièrement en 3D : on peut ainsi songer à acquérir la surface de la langue avec des coupes US coronales, ou même faire des acquisitions US où l'on balaye le conduit vocal pour le reconstruire en trois dimensions [SEKL05]. En prenant soin de bien effectuer au préalable une étude de répétabilité et de variabilité de la langue en 3D, nous pourrions mettre en place des modèles 3D dynamiques de la langue avec notre système d'acquisition.

Les mêmes études que celles suggérées en 2D sont alors possibles : rechercher l'existence de modèles permettant à la fois de décrire toute la dynamique des articulateurs, et des modèles 3D capables d'être utilisés pour plusieurs locuteurs.

L'objectif final serait de savoir s'il est possible de mettre en place un modèle articulatoire tridimensionnel dynamique et suffisamment générique pour être adapté à tous les locuteurs possibles. L'application visée pourrait être une tête parlante 3D, incluant les déformations du conduit vocal, pilotée seulement à partir des données articulatoires dynamiques de notre système d'acquisition. Avant d'en arriver à une telle application, le chemin à parcourir est encore long.

Ce travail, l'un des tous premiers à fusionner données statiques et dynamiques sur le conduit vocal, permet en tout cas d'ouvrir de telles perspectives de recherche.

6.5 Conclusion

Ce chapitre présente dans un premier temps l'ensemble des données acquises et traitées avec succès avec notre système d'acquisition, en prenant soin de souligner les nombreuses difficultés inhérentes à la mise en place de tout système d'acquisition de données articulatoires.

Nous avons ensuite montré comment ces données pouvaient être utilisées pour retrouver les paramètres articulatoires du modèle de Maeda. Cette approche permet d'évaluer la pertinence des données dans un contexte applicatif en parole. En effet, il ne suffit pas d'être en mesure d'acquérir un grand volume de données pour justifier d'un système d'acquisition fiable. Il faut qualifier ces données en leur attribuant un niveau de confiance exprimé par l'incertitude que nous avons calculée dans le chapitre 5. Il faut également vérifier que l'ensemble de ces données acquises et traitées peut être utilisé pour des applications en parole, par exemple en vérifiant leur cohérence avec des modèles articulatoires établis. Ces deux étapes sont indispensables avant de pouvoir dire que ces données sont *bien fondées*.

Enfin, nous avons ouvert notre travail sur de nombreuses perspectives de recherche, pour certaines actuellement explorées dans notre laboratoire. À court terme, nous continuerons à améliorer le système d'acquisition, à tester un modèle de langue sur le suivi et en étudiant plus de données de différents locuteurs. À plus long terme, ces données peuvent être utilisées pour de nombreuses applications, que ce soit pour la mise en place de modèles de déformations et pour

des études multilocuteurs en 2D et en 3D. Nous pensons avoir ici montré tout l'intérêt du travail réalisé en pointant ces diverses pistes de recherche à étudier.

Conclusion

L'objectif de ce travail était la mise en place d'un ensemble de méthodes pour l'acquisition, la fusion et la validation d'un ensemble de données articulatoires multimodales. Il s'inscrit dans la lignée des systèmes d'acquisition HOCUS [WTO+05], MOCHA [WH00], du VTV [Sto05] et du GIPSA-lab [BEB+07] qui utilisent des données articulatoires issues de plusieurs modalités. Cependant, contrairement à ces systèmes, nous avons proposé des méthodes pour fusionner (recaler et synchroniser) automatiquement des volumes importants de données multimodales. Par ailleurs, nous avons avons attaché aux données articulatoires issues de notre système une précision spatiale et temporelle, et nous avons validé les données en les utilisant avec succès pour des applications en parole. Ces étapes font des données issues de nos travaux des données *bien fondées*, contribuant ainsi à améliorer la qualité des données articulatoires disponibles. Les perspectives principales que nous envisageons à ce travail ayant été exposées dans le chapitre précédent, nous reprenons dans cette conclusion les apports essentiels de notre travail de thèse avant d'en dresser un court bilan.

Les acquisitions Nous avons décrit un ensemble de protocoles et de méthodes permettant d'acquérir des données statiques et dynamiques sur le conduit vocal. Nous avons détaillé l'architecture d'un système multimodal dynamique. Ce système basé sur les échographies, des données électromagnétiques, de stéréovision et acoustiques constitue un sous-système pouvant être reproduit dans un laboratoire et être éventuellement complété avec d'autres modalités. Nous avons pris soin de proposer un protocole d'acquisition IRM adapté aux acquisitions de phonèmes.

Ces méthodes permettent d'acquérir automatiquement un important corpus sur plusieurs locuteurs.

Le traitement et la fusion Nous avons aussi présenté des techniques de traitement des données statiques et dynamiques. Ces traitements concernent aussi bien l'aspect temporel pour synchroniser les modalités entre elles, et l'aspect spatial en les recalant dans un même repère. Nous avons présenté une méthode d'extraction des contours de la langue dans les images échographiques en adaptant une méthode classique de suivi utilisée en traitement d'image aux spécificités de notre application (mouvements rapides et élastiques). Nous utilisons notamment les capteurs électromagnétiques pour contraindre le suivi.

Ces méthodes permettent de disposer d'un corpus de données articulatoires pour lequel les formes des articulateurs sont extraites et sont toutes représentées dans un même repère spatial et temporel.

L'évaluation Enfin, la précision de chaque modalité a été étudiée. Nous avons proposé des méthodes pour mesurer les délais d'acquisition. Nous avons quantifié l'incertitude de chacune des modalités, des sous-systèmes statiques et dynamiques et enfin du recalage global. C'est à

notre connaissance les seuls travaux faisant état de telles mesures. Nous avons enfin évalué des données acquises et traitées en les confrontant à un modèle articulatoire existant.

Ces méthodes permettent de qualifier les données articulatoires du corpus, et de leur associer, ainsi qu'au système complet, une mesure de confiance.

Bilan Nous avons acquis pour le projet européen ASPI un volumineux corpus sur plusieurs locuteurs. Ces acquisitions ont été effectuées récemment, et nous n'avons encore que partiellement dépouillé et traité toutes ces données. Elles sont actuellement utilisées et testées par les membres du projet pour leurs recherches [KRM+08, TMAB08, ATB+09]. Nous nous sommes souvent posé la question durant ces quatre années de la précision nécessaire pour évaluer des méthodes d'inversion acoustique articulatoire. La réponse communément formulée par la communauté parole est que les données doivent être les plus précises possible. Ce travail permet d'apporter une réponse quantitative, en qualifiant chaque donnée par une mesure d'incertitude. Nous pourrons ainsi grâce aux retours sur les travaux d'inversion de nos partenaires connaître avec plus d'exactitude la précision nécessaire sur des données articulatoires. Une fois ce travail effectué, le projet ASPI prévoit pour la fin de l'année 2010 de rendre une partie des données publiques.

Au cours de ce travail, nous avons couvert un grand nombre de problématiques, allant d'aspects matériels et pratiques à des aspects utilisant des techniques de traitement d'image avancées. Il a fallu faire preuve de pragmatisme pour répondre aux nombreuses contraintes inhérentes au matériel d'acquisition (nous aurions aimé par exemple avoir accès au PC de l'échographe pour faciliter l'enregistrement et la synchronisation des données), et proposer des solutions les plus automatiques et les plus fiables possibles. Ce large spectre de techniques utilisées est à la fois l'un des points forts et l'un des points faibles de ce travail : il est pluridisciplinaire mais ne se focalise pas sur une thématique précise liée à l'acquisition de données articulatoires. Partant de zéro au début du projet, le résultat de ce choix initial dans l'orientation de la thèse est que nous bénéficions aujourd'hui d'une base de travail solide qui peut dorénavant être complétée et améliorée par des travaux qui pourraient se focaliser et approfondir une thématique particulière. Elle ouvre de nombreuses perspectives présentées dans la dernière section du sixième chapitre.

D'un point de vue global, nous avons appliqué à un domaine attaché à celui de la parole des méthodes utilisées dans l'imagerie. Nous avons appliqué des techniques de vision par ordinateur pour fusionner les différentes données et vérifier leur validité en les testant à des applications de parole. Nous pensons ainsi avoir contribué à améliorer les acquisitions et les traitements de données articulatoires.

Annexe A

Corpus

Nous présentons ici le corpus français acquis avec le système dynamique. Ce corpus a été mis en place par l'équipe Parole du LORIA. Il comprend 4 parties :

- une partie où des fricatives sont prononcées dans un contexte de VCV ;
- une partie VV ;
- une partie VCV ;
- une partie où des phrases sont prononcées. Le corpus est définit par Combescure dans [Com81].

Fricatives

- ʃ : aʃa aʃɛ ɛʃɛ iʃi iʃe eʃe uʃu uʃo oʃo aʃy iʃy aʃø iʃø iʃø
- s : asa asɛ ɛsɛ isi ise ese usu uso oso asy isy asø isø isø
- f : afa afɛ ɛfɛ ifi ife efe ufu ufo ofo afy ify afø ifø

VV

- ɛkɛ aka utu oto iti ity ito ito
- aɛ ae ai aɔ ao au ay aø aœ aœ
- ie iɛ ia iy iø iœ iɔ io iu iu
- yi ye yɛ ya yø yœ yɔ yo yu yu
- ui ue uɛ ua uy uø uœ uɔ uo uo

VCV

- aka aki aku ika iki iku uka uki uku uku
- ata ati atu ita iti itu uta uti utu utu
- apa api apu ipa ipi ipu upa upi upu upu
- aʃa aʃi aʃu iʃa iʃi iʃu uʃa uʃi uʃu uʃu
- asa asi asu isa isi isu usa usi usu usu
- afa afi afu ifa ifi ifu ufa ufi ufu ufu

Phrases Voici l'ensemble phrases utilisées lors des acquisitions dynamiques :

Il se garantira du froid avec ce bon capuchon. (0)
Annie s'ennuie loin de mes parents. (1)
Les deux camions se sont heurtés de face. (2)
Un loup s'est jeté immédiatement sur la petite chèvre. (3)

Dès que le tambour bat, les gens accourent. (4)
Mon père m'a donné l'autorisation. (5)
Vous poussez des cris de colère ? (6)
Ce petit canard apprend à nager. (7)
La voiture s'est arrêtée au feu rouge. (8)
La vaisselle propre est mise sur l'évier. (9)
Leur chienne a hurlé toute la nuit. (10)
Pour se protéger, il s'est couché près de ma porte. (11)
Sa voisine est inimitable. (12)
Le renard se hâte vers son gîte. (13)
Le bouillon fume dans les assiettes. (14)
Le caractère de cette femme est moins calme. (15)
Le camp d'été s'est passé au bord du fleuve. (16)
Un train entre déjà en gare. (17)
Souvent, je m'accoude au muret de ce pont. (18)
A l'Ouest, mes pommiers donnent peu. (19)
Lentement des canes se dirigent vers la mare. (20)
Une goélette déploie ses voiles. (21)
Le facteur va porter le courrier. (22)
Bien sûr, je connais son nom. (23)
Maman prend un verre et une assiette. (24)
Désormais, je me tournerai quand il partira. (25)
Les avions tournent au-dessus de la place. (26)
Mettez la faux, ici sous ma tente. (27)
Je suis resté sourd à ses cris. (28)
Le chameau est loin de son abri. (29)
Il pense être de retour ici, avant la nuit. (30)
Des chiens nous montraient leurs crocs pointus. (31)
La jeune fille se peigne devant sa glace. (32)
Il a été condamné pour un vol de voiture. (33)
Je ne veux pas que vous le changiez pour le moment. (34)
Nous avons pris froid en jouant au tennis. (35)
Il est désormais accablé par son travail. (36)
Ce bonbon contenait trop de sucre. (37)
A la hâte, le métayer ensilait ses récoltes avant l'hiver. (38)
Une brume épaisse s'est formée sur la mer. (39)
Le menuisier a scié une planche et l'a rabotée. (40)
Maman a préparé une galette pour jeudi. (41)
Le football, voilà ce qui l'intéresse. (42)
C'est un charmant spectacle, je t'assure. (43)
Ils m'ont apporté des friandises à mon anniversaire. (44)
Ces élèves prendront l'autocar tout à l'heure. (45)
Parfois, mon épicière vend à crédit. (46)
Personne n'a applaudi ce beau discours. (47)
Je me demande pourquoi on court sans cesse. (48)
Il se repend de ce qu'il vient de faire. (49)
Des gens se sont levés dans les tribunes. (50)
Vous éplucherez les légumes du pot-au-feu. (51)

Ce chasseur projette encore de partir d'ici, ce matin. (52)
La poire est un fruit à pépins. (53)
Plus nous le connaissons, plus nous le respectons. (54)
Là-haut, monte la voix du pâtre qui ramène ses moutons. (55)
Le courrier arrive en retard en ce moment. (56)
Cette cage contient mon oiseau. (57)
Des lièvres jouent à l'orée du bois. (58)
Je te dis que ma bouteille s'abîme à la cave. (59)
II s'est réfugié dans ma chambre. (60)
Le troupeau s'abreuvait au ruisseau. (61)
Le client s'attend à ce que vous fassiez une réduction. (62)
Chaque fois que je me lève, ma plaie me tire. (63)
Une rançon est exigée par les ravisseurs. (64)
Ainsi, cette comédie est en un acte. (65)
Papa aime mon vin quand il est bon. (66)
Le ciel est tout noir, il va tomber des cordes. (67)
On dit que l'essor de ce village est important. (68)
Ce soir, nous ne nous coucherons pas tard. (69)
Vous avez du plaisir à jouer avec ceux qui ont un bon caractère. (70)
Le chevrier a corné pour rassembler ses troupeaux. (71)
Mon cordonnier a ressemelé tes souliers. (72)
L'oie est dans sa main, son cœur bat et saute. (73)
Une rivière dessinait des méandres dans sa prairie. (74)
L'alpiniste continuait à grimper le long d'une roche. (75)
Effrayé par l'insecte, je rentre précipitamment. (76)
Je me suis entretenu avec l'institutrice de ma jeune fille. (77)
Quand le soleil se lève, je saute de mon lit. (78)
Le fermier est parti pour la foire. (79)
L'été, tout le monde se mettait aux fenêtres. (80)
Le cocher a fouetté sa jument. (81)
Je rends souvent visite à mon oncle. (82)
Ma soirée se passera sans incident. (83)
La police veut les papiers du chauffeur. (84)
Jean, quant à lui, est très grand pour son âge (85)
Le microscope, qui est sur pied, est le mien. (86)
Le jardin entoure un petit lac. (87)
Il a broyé du noir depuis la perte de son ami. (88)
Le forçat s'est évadé du bagne. (89)
Un fort crédit est consenti par une banque. (90)
Le passereau lance une roulade et s'enfuit. (91)
Des hannetons voletaient autour de ce prunier. (92)
Ces légendes me rappellent les temps anciens. (93)
Qu'est ce que vous regardez comme oiseau ? (94)
Ce sont mes meilleurs chevaux dont voici les noms. (95)
Je parcours les rues des villages avec sa mère. (96)
Ma partition est sous ce pupitre. (97)
Il arrive demain d'Italie par la route. (98)
Le tapis était élimé sur le bord. (99)

Ma mère et moi faisons de courtes promenades. (100)
La poupée fait la joie de cette très jeune fille. (101)
Mais le temps lui a manqué. (102)
Il aura été retardé par quelqu'importun. (103)
Une grenouille verte saute sur les nénuphars. (104)
Des violettes emplissent l'air de subtiles senteurs. (105)
Au bois, j'ai ramassé de si bons champignons. (106)
Fais ce que je veux dès ce midi ! (107)
Papa coupe l'herbe dans le jardin. (108)
Vous porterez ces caisses dans vos voitures. (109)
A midi les collégiens vont au réfectoire. (110)
Des pommes mûres se détachent de l'arbre. (111)
Il tombe lourdement sur un sol plat. (112)
Nous partons avant demain vers Paris. (113)
Il a été arrêté par des policiers. (114)
Cette voyageuse a loué une voiture sans chauffeur. (115)
Jean semblait calme tout à coup. (116)
Je n'irai sûrement pas danser à son mariage. (117)
Elle le lui redit sans cesse. (118)
Une guerre nucléaire ferait de nombreux morts. (119)
La lune se lève maintenant au-dessus des arbres. (120)
Des rires montent de la cour de récréation. (121)
Ça et là, la prairie se piquait de fleurs. (122)
Une grosse poutre maintient la misérable charpente. (123)
Tout s'est animé, dès que le soleil s'est levé. (124)
Ma voiture est en panne devant ce pont. (125)
Ces femmes portent encore une coiffe. (126)
Je lui rapporte des fruits très rouges. (127)
Je vois ma table en bois vert. (128)
Dans le taillis est cache un nid de fauvettes. (129)
On entend les gazouillis d'un oiseau dans le jardin. (130)
La barque du pêcheur a été emportée par une tempête. (131)
Ce livre provient de la bibliothèque. (132)
J'en conclus qu'il n'y a personne à voir. (133)
Le mal s'envenime, faute de soins. (134)
Je suis sûr que vous connaissez ces noms. (135)
Il s'arrêtait tout l'été, ici. (136)
Voilà toujours deux choux pour le repas de midi. (137)
Les manches de son manteau sont décousues. (138)
Ce vaisseau parcourt les mers à travers le monde. (139)
Vous achèterez mes moules minuscules ! (140)
Une jolie bague scintille au doigt de ta fille aînée. (141)
A six heures, un voyageur attendait le train. (142)
Ses locataires sont rentrés très tard. (143)
Ce que j'ai prévu se produira. (144)
Le capitaine regarde par le hublot de sa cabine. (145)
Virginie a mis le couvert pour sa fête. (146)
Votre portrait est expose au salon. (147)

Maman se demande ce qu'il va dire. (148)
Des moineaux se sont querelles dans mon champ. (149)
Une société de musique va bientôt défiler. (150)
Le juge veut prolonger l'interrogatoire. (151)
Ici, ma mère a acheté des coupons de tissu. (152)
Pierre cogne par derrière comme un sourd. (153)
La pluie ne fait pas le beau temps. (154)
Sans fleurs, la maison est triste. (155)
Elle a vraiment toujours des doigts menus. (156)
Ce boucher n'a encore plus de lard à l'étalage. (157)
Confie-moi à quoi tu penses. (158)
Ce dont nous discutons vous laisse rêveur. (159)
Elle habite à proximité du champ de foire. (160)
Ma concierge veillait sur mon appartement durant les vacances. (161)
Ils sont allés travailler bien qu'ils fussent fatigués. (162)
Un colonel commandait le régiment. (163)
Je vous dis de recoudre ce bouton. (164)
Nos parents sont nos tuteurs naturels. (165)
Les mésanges y pondaient des œufs tachetés. (166)
Vous lui défendez de jouer sur cette route le soir. (167)
Ma goélette noire est rentrée au port. (168)
Ce passeport n'avait pas de visa. (169)
La pieuvre saisit sa proie avec ses tentacules. (170)
La neige couvre la cime des montagnes. (171)
Un mouflon se cache dans les anfractuosités. (172)
J'ai entendu ce que vous tachez de jouer au piano. (173)
Ce moyeu de roue grince continuellement. (174)
Nous voulons tous nous promener à bicyclette. (175)
Je ne peux atteindre les bocaux de confiture. (176)
Dans cette crémerie, on vend du fromage fort. (177)
La pie se précipité vers ce qui brille. (178)
Un petit lièvre est terré dans le buisson. (179)
Je ménage une surprise à mon ami. (180)
Les boulangers façonnent des pains. (181)
Vos livres devront être couverts. (182)
Mangeras-tu de cette tarte aux prunes. (183)
Le chapeau de Monique est sur la table. (184)
Il s'est glissé loin des spectateurs. (185)
II s'empresse de réclamer ce qu'on lui a promis. (186)
Vous tremblez parce que vous avez froid. (187)
C'est le soir qu'il travaille le mieux. (188)
Un serpent noir fuit sous une pierre. (189)
Je me souviens des beaux jours que j'ai vécus. (190)
Mon cousin a été très vexé par ce qu'elle avait dit. (191)
Grand-père, sois donc un peu raisonnable ! (192)
En ce moment, les soirées à l'opéra sont données. (193)
Tu as beaucoup changé depuis que tu es parti. (194)
Il a souffert pendant des semaines. (195)

Vous voyez tout le temps cette femme triste. (196)
Ce sentier mène à la route du village. (197)
Le docteur a ordonné un médicament. (198)
Il faut aussi arriver à temps. (199)

Bibliographie

[AFK+07] M. Aron, N. Ferveur, E. Kerrien, M.O. Berger, and Y. Laprie. Acquisition and synchronization of multimodal articulatory data. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1398–1401, Anvers, Belgique, 2007.

[AKJ+01] D. Amin, T. Kanade, B. Jaramaz, A.M. Di Gioia, C. Nikou, R. LaBarca, and J.E. Moody. Calibration method for determining the physical location of the ultrasound image plane. In *Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001)*, Octobre 2001.

[AMP09] C. Avenel, E. Memin, and P. Perez. Tracking closed curves with non-linear stochastic filters. In *Conference on Scale Space and Variational Methods (SSVM'09)*, Voss, Norvège, Juin 2009.

[AMT00] M.E. Anderson, M.S. McKeag, and G.E. Trahey. The impact of sound speed errors on medical ultrasound imaging. *Journal of the Acoustical Society of America (JASA)*, 107(6) :3540–3548, Juin 2000.

[ATB+09] M. Aron, A. Toutios, M.O. Berger, E. Kerrien, B. Wrobel Dautcourt, and Y. Laprie. Registration of Multimodal Data for Estimating the Parameters of an Articulatory Model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[AWJ90] A.A Amini, T.E. Weymouth, and R.C. Jain. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9) :855–867, Septembre 1990.

[Bas94] B. Bascle. *Contributions et applications des modéles déformables en vision par ordinateur*. PhD thesis, Université de Nice-Sophia Antipolis, Juin 1994.

[BBR+02] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional articulatory modeling of the tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3) :533–553, 2002.

[BCZ93] A. Blake, R. Curwen, and A. Zisserman. A framework for spatiotemporal control in the tracking of visual contours. *International Journal of Computer Vision*, 11(2) :127–145, Octobre 1993.

[BEB+07] P. Badin, F. Elisei, G. Bailly, C. Savariaux, A. Serrurier, and Y. Tarabalka. Têtes parlantes audiovisuelles virtuelles : données et modèles articulatoires - applications. *Rev. Laryngol. Otol. Rhinol.*, 128(5) :289–295, 2007.

[BEG03] J. Beskow, O. Engwall, and B. Granström. Resynthesis of Facial and Intraoral Motion from Simultaneous Measurements. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pages 431–434, Barcelone, Espagne, 2003.

[Ber91] M.O. Berger. *Les contours actifs : modélisation, comportement et convergence*. PhD thesis, Institut National Polytechnique de Lorraine (INPL), Nancy, France, Février 1991.

[BGGN91] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging : Vowels. *Journal of the Acoustical Society of America (JASA)*, 90(2) :799–828, 1991.

[BJC+03] E.M. Boctor, A. Jain, M.A. Choti, R.H. Taylor, and G. Fichtinger. Rapid calibration method for registration and 3D tracking of ultrasound images using spatial localizer. In *Medical Imaging (SPIE)*, volume 5035, pages 521–532, 2003.

[BM92] P.J. Besl and N.D. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2) :239–256, 1992.

[Bon93] M. Bonvalet. *Les principes variationnels*. Masson, 1993.

[Bra85] P. Branderud. Movetrack - a movement tracking system. In *the French-Swedish Symposium on Speech*, pages 113–122, Grenoble, France, 1985.

[BRCM+00] J.M. Blackall, D. Rueckert, Jr. C.R Maurer, G.P. Penney, D.L.G. Hill, and D.J. Hawkes. An Image Registration Approach to Automated Calibration for Freehand 3D Ultrasound. In *MICCAI '00 : Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 462–471, Londres, Royaume-Uni, 2000. Springer-Verlag.

[BTB+05] T. Bressmann, P. Thind, C.M. Bollig, C. Uy, R.W. Gilbert, and J.C Irish. Quantitative three-dimensional ultrasound analysis of tongue protrusion, grooving and symmetry : Data from twelve normal speakers and a partial glossectomee. *Clinical Linguistics and Phonetics*, 19 :573–588, 2005.

[Buc07] S. Buchaillard. *Activations musculaires et mouvements linguaux : modélisation en parole naturelle et pathologique*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, Décembre 2007.

[BW93] N. Bilaniuk and G. Wong. Speed of sound in pure water as function of temperature. *Journal of the Acoustical Society of America (JASA)*, 93 :1609–1612, 1993.

[BWL99] M.O. Berger, G. Winterfeldt, and J.P. Lethor. Contour tracking in echocardiographic sequences without learning stage : Application to the 3d reconstruction of the beating left ventricule. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 508–515, Cambridge, Royaume-Uni, 1999.

[Car96] J. Carr. *Surface Reconstruction in 3D Medical Imaging*. PhD thesis, University of Canterbury, Christchurch, Nouvelle Zélande, 1996.

[CBZ92] R. Curwen, A. Blake, and A. Zisserman. Real time Visual Tracking for surveillance and Path Planning. In *Proceedings of 7th European Conference on Computer Vision (ECCV)*, pages 879–883, Copenhague, Danemark, Juin 1992.

[CF66] C. Coker and O. Fujimura. Model for the specification of the vocal tract area function. *Journal of the Acoustical Society of America (JASA)*, 40 :1271, 1966.

[CFP98] R.M. Comeau, A. Fenster, and T.M. Peters. Integrated MR and ultrasound imaging for improved image guidance in neurosurgery. In *Medical Imaging (SPIE)*, volume 3338, pages 747–754, 1998.

[CK41] T. Chiba and M. Kajiyama. *The Vowel : Its Nature and Structure*. Tokyo-Kaseikan, 1941.

[Col95] A.E.A. Collignon. Automated Multi-Modality Image Registration Based on Information Theory. In *Proceedings of the 14th Conference on Information Processing in Medical Imaging*, volume 3, pages 263–274, 1995.

[Com81] P. Combescure. Vingt listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 14(56), 1981.

[CTCG95] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - Their training and application. *Computer Vision and Image Understanding*, 61(1) :38–59, 1995.

[Dar87] S. Dart. A bibliography of X-ray studies of speech. *UCLA Phonetics Laboratory Group*, 66, 1987.

[DBH+94] P.R. Detmer, G. Bashein, T. Hodges, K.W. Beach, E.P. Filer, D.H. Burns, and D.E Jr Strandness. 3D ultrasonic image feature localization based on magnetic scanhead tracking : in vitro calibration and validation. *Ultrasound in Medicine and Biology*, 20(9) :923–936, 1994.

[DCT01] R.H. Davies, T.F. Cootes, and C.J. Taylor. *A Minimum Description Length Approach to Statistical Shape Modelling*, volume 2082. Information Processing in Medical Imaging, 2001.

[Eng00] O. Engwall. A 3D tongue model based on MRI data. In *Proceedings of the International Conference on Language and Signal Language Processing (Interspeech)*, Beijing, Chine, Octobre 2000.

[Eng04] O. Engwall. From real-time MRI to 3D tongue movements. In Soon Hyob Kim and Dae Hee Youn, editors, *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, pages 1109–1112, Jeju Island, Corée du Sud, Octobre 2004.

[Eng08] O. Engwall. Can audio-visual instructions help learners improve their articulation ? - an ultrasound study of short term changes. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2631–2634, Brisbane, Australie, 2008.

[ES05] M.A. Epstein and M. Stone. The tongue stops here : Ultrasound imaging of the palate. *Journal of the Acoustical Society of America (JASA)*, 2005.

[Fan60] G. Fant. *Acoustic Theory of Speech Production*. The Hague : Mouton & Co., 1960.

[FB05] J. Fontecave and F. Berthommier. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database. In *Annual Conference of the International Speech Communication Association (Interspeech)*, Lisbonne, Portugal, Septembre 2005.

[Fla72] J.L. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, 2nd edition, New York, 1972.

[Fle87] R. Fletcher. *Practical methods of optimization*. Wiley-Interscience New York, NY, USA, 1987.

[FSH+09] S. Fels, I. Stavness, A.G. Hannam, J.E. Lloyd, P. Anderson, C. Batty, H. Chen, C. Combe, T. Pang, T. Mandal, B. Teixeira, S. Green, R. Bridson, A. Lowe, F. Almeida, J. Fleetham, and R. Abugharbieh. Advanced tools for biomechanical modeling of the oral, pharyngeal, and laryngeal complex. In *International Symposium on Biomechanics Healthcare and Information Science*, Février 2009.

[FTV93] B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes, 2nd Edition*. Cambridge University Press, 1993.

[GOL+04] J.M. Gérard, J. Ohayon, V. Luboz, P. Perrier, and Y. Payan. Indentation for estimating the human tongue soft tissues constitutive law : application to a 3d biomechanical model to study speech motor control and pathologies of the upper airways. *Lecture Notes in Computer Science*, 3078 :77–83, 2004.

[GWTPP03] J.M. Gérard, R. Wilhelms-Tricarico, P. Perrier, and Y. Payan. A 3D Dynamical Biomechanical Tongue Model to Study Speech Motor Control. *Research Developments in Biomechanics*, pages 49–64, 2003.

[HCD+07] T. Hueber, G. Chollet, B. Denby, M. Stone, and L. Zouari. Ouisper : Corpus Based Synthesis Driven by Articulatory Data. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, pages 2193–2196, Saarbrücken, Allemagne, 2007.

[HCDS08] T. Hueber, G. Chollet, B. Denby, and M. Stone. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, pages 365–369, Strasbourg, France, 2008.

[HGK04] K. Huang, S. Graham, and P.R. Kumar. Temporal alignment of distributed sensors with an application to characterization of plant delay. In *IEEE 43rd International Conference on Decision and Control*, Paradise Island, Bahamas, Décembre 2004.

[Hoo93] P. Hoole. Methodological considerations in the use of electromagnetic articulography in phonetic research. Technical Report Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation, Universität München, 1993.

[HS80] B. Horn and B. Schunk. Determining Optical Flow. Ai-memo 572, Massachusetts Institute of Technology (MIT), Cambridge, MAS, États-Unis, 1980.

[Hub81] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[HZ00] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN : 0521623049, 2000.

[IB96] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proceedings of 4th European Conference on Computer Vision*, volume 1064, pages 343–356, Cambridge, Royaume-Uni, 1996.

[JNB98] G. Jacob, A. Noble, and A. Blake. Robust Contour Tracking in Echographic Sequences. In *Proceedings of 6th International Conference on Computer Vision*, pages 408–413, Bombay, Inde, Janvier 1998.

[Ken97] R.D. Kent. *Speech Sciences*. Singular, 1997.

[KHP78] W. Kahle, H.Leonhardt, and W. Plater. *Anatomie - Tome 2 : Viscères*. Flammarion, 1978.

[KIF75] S. Kiritani, K. Itoh, and O. Fujimura. Tongue-pellet tracking by a computer controlled X-ray microbeam system. *Journal of the Acoustical Society of America (JASA)*, 48 :1516–1520, 1975.

[Kir05] S. Kirsch. Accuracy assessment of the electromagnetic tracking system aurora. Technical report, NDI Europe GmbH, 2005.

[KRM+08] A. Katsamanis, A. Roussos, P. Maragos, M. Aron, and M.O. Berger. Inversion from Audiovisual Speech to Articulatory Information by Exploiting Multimodal

Data. In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, Strasbourg, France, 2008.

[Kro08] C. Kroos. Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, Strasbourg, France, 2008.

[KS05] A. Khamene and F. Sauer. *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2005)*, volume 3750, chapter A Novel Phantom-Less Spatial and Temporal Ultrasound Calibration Method, pages 65–72. 2005.

[KSN09] Y.C Kim, S. Shhrikanth, and K.S. Nayak. Accelerated 3D MRI of vocal tract shaping using compressed sensinf and parallel imaging. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[KVPG06] B. Kastler, D. Vetter, Z. Patay, and P. Germain. *Comprendre l'IRM : Manuel d'auto-apprentissage*. Masson, 2006.

[KWT88] M. Kass, A. Witkin, and D. Terzopoulos. Snakes : Active Contour Models. *International Journal of Computer Vision*, 1 :321–331, 1988.

[Lad01] P. Ladefoged. *A Course in Phonetics, 4th edition*. Heinle, 2001.

[LC87] W. Lorensen and H.E. Cline. Marching Cubes : A High Resolution 3D Surface Construction Algorithm. In *Proceedings of SIGGRAPH*, volume 2, pages 163–170, Juillet 1987.

[Lev00] M.E. Leventon. *Statistical Models for Medical Image Analysis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, États-Unis, 2000.

[LKS03] M. Li, C. Kambhamettu, and M. Stone. Snake for band edge extraction and its applications. In *Computer Graphics and Imaging*, pages 261–266, 2003.

[LKS06] M. Li, C. Kambhamettu, and M. Stone. A level set approach for shape recovery of open contours. In *7th Asian Conference on Computer Vision (ACCV)*, volume 3851, pages 601–611, Hyderabad, Inde, Septembre 2006. Springer.

[LM98] Y. Laprie and B. Mathieu. A variational approach for estimating vocal tract shapes from the speech signal. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 929–932, Seattle, WA, États-Unis, Mai 1998.

[Mae79] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, France, Mai 1979.

[Mae82] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1 :199–229, 1982.

[Mae92] S. Maeda. Modélisation articulatoire du conduit vocal. *Journal de physique*, 2 :191–198, 1992.

[Mar97] W. Marczak. Water as standard in the measurements of speed of sound in liquids. *Journal of the Acoustical Society of America (JASA)*, 102(5) :2776–2779, 1997.

[MBE+06] S. Maeda, M.O. Berger, O. Engwall, Y. Laprie, P. Maragos, B. Potard, and J. Schoentgen. Technology inventory of audiovisaul-to-articulatory inversion. Technical report, FET ASPI Report 2.0, Novembre 2006.

[MDM99] J. Montagnat, H. Delingette, and G. Malandain. Cylindrical echocardiographic images segmentation based on 3D deformable models. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'99)*, volume 1679 of *Lectures Notes in Computer Science*, pages 168–175, Cambridge, Royaume-Uni, Septembre 1999. Springer.

[Mer73] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America (JASA)*, 53 :1070–1082, 1973.

[MHH+05] R.A. McLaughlin, J. Hipwell, D.J. Hawkes, J.A. Noble, J.V. Byrne, and T.C. Cox. A comparison of a similarity-based and a feature-based 2D-3D registration method for neurointerventional use. *IEEE Transactions On Medical Imaging*, 24(8) :1058–1066, 2005.

[MHV+03] D. Mattes, D.R. Haynor., H. Vesselle, T.K. Lewellen, and W. Eubank. PET-CT image registration in the chest using free-form deformations. *IEEE Transactions On Medical Imaging*, 22(1) :120–128, Janvier 2003.

[MLLC05] L. Mercier, T. Lango, F. Lindseth, and D.L. Collins. A reviwe of calibration techniques for freehand 3-d ultrasound systems. *Ultrasound in Medicine and Biology*, 31(4) :449–471, Novembre 2005.

[MV98] J.B.A. Maintz and M.A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1) :1–16, 1998.

[Nic04] S. Nicolau. *Un système de réalité augmentée pour guider les opérations du foie en radiologie interventionnelle.* PhD thesis, Université de Nice-Sophia Antipolis, Nice, France, Novembre 2004.

[Ove62] J.E. Overall. Orthogonal Factors and Uncorrelated Factor Scores. *Psychological Reports*, 10 :651–662, 1962.

[PCS+92] J.S. Perkell, M.H. Cohen, M.A. Svirsky, M.L. Matthies, I. Garabieta, and M.T.T. Jackson. Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America (JASA)*, 92(6) :3078–3096, 1992.

[Per69] J.S. Perkell. *Physiology of speech production : results and implications of a quantitative cineradiographic study.* MIT Press, Cambridge, MA, États-Unis, 1969.

[Per74] J.S. Perkell. *A physiologically-oriented model of tongue activity in speech production.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, États-Unis, 1974.

[PHK00] N. Pagoulatos, D.R. Haynor, and Y. Kim. Image-based registration of ultrasound and magnetic resonance images : a preliminary study. In *SPIE Medical Imaging*, volume 3976, pages 156–164, 2000.

[PHVG02] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proceedings of 7th European Conference on Computer Vision (ECCV)*, pages 661–675, Copenhague, Danemark, Juin 2002.

[PLO04] B. Potard, Y. Laprie, and S. Ouni. Expériences d'inversion basées sur un modèle articulatoire. In *Actes des Journées d'Etudes sur la Parole (JEP)*, 2004.

[PMV03] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images : a survey. *IEEE Transactions On Medical Imaging*, 22(8) :986–1004, Août 2003.

[PR05] T.C. Poon and R.N Rohling. Comparison of calibration methods for spatial tracking of a 3D ultrasound probe. *Ultrasound in Medicine and Biology*, 31(8) :1095–1108, Avril 2005.

[PRGB98] R.W. Prager, R.N. Rohling, A.H. Gee, and L. Berman. Rapid calibration for 3-D freehand ultrasound. *Ultrasound in Medicine and Biology*, 24(6) :855–869, Mars 1998.

[QCP07] C. Qin and M.Á. Carreira-Perpiñán. A comparison of acoustic features for articulatory inversion. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2469–2472, Anvers, Belgique, 2007.

[RHI+86] M. Rokkaku, K. Hashimoto, S. Imaizumi, S. Nimi, and S. Kirtani. Measurements of the Three-Dimensional Shape of the Vocal Tract Based on the Magnetic Resonance Imaging Technique. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, 20 :47–54, 1986.

[RL01] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. In *Proceedings of the Third International Conference on 3D Digital Imaging and Modeling*, pages 145–152, 2001.

[Rou03] F. Rousseau. *Méthodes d'analyse d'images et de calibration pour l'échographie 3D en mode main-libre*. PhD thesis, Université de Rennes I, Rennes, France, Décembre 2003.

[RPMA01] A. Roche, X. Pennec, G. Malandain, and N. Ayache. Rigid registration of 3D ultrasound with MR images : a new approach combining intensity and gradient information. *IEEE Transactions on Medical Imaging*, 20(10) :1038–1049, Octobre 2001.

[SBW07] P. Shrstha, M. Barbieri, and H. Weda. Synchronization of multi-camera video recordings based on audio. In *MULTIMEDIA '07 : Proceedings of the 15th international conference on Multimedia*, pages 545–548, New York, NY, États-Unis, 2007. ACM.

[SD95] M. Stone and E. Davis. A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement. *Journal of the Acoustical Society of America (JASA)*, 98(6) :3107–3112, 1995.

[SEKL05] M. Stone, M.A. Epstein, C. Kambhamettu, and M. Li. *Predicting 3D tongue shapes from midsagittal contours.*, chapter 18, pages 315–330. Speech Production : Models, Phonetic Processes, and Techniques, J.Harrington and M. Tabain edition, 2005.

[Ser06] A. Serrurier. *Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production de nasales*. PhD thesis, Institut National Polytechnique de Grenoble, 2006.

[Sho85] K. Shoemake. Animating rotation with quaternion curves. In *SIGGRAPH '85 : Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, New York, NY, USA, 1985. ACM.

[SMJC99] C. Shadle, M. Mohammad, P. Jackson, and J. Carter. Multi-planar dynamic magnetic resonance imaging : New tools for speech research. In *Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS)*, pages 623–626, 1999.

[SSB+07] M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, V. Parthasarathy, J. Prince, M. Li, and C. Kambhamettu. Comparison of speech production in upright and

supine position. *Journal of the Acoustical Society of America (JASA)*, 122(1) :532–541, 2007.

[SSS+06] M. Schünke, E. Schulte, U. Schumacher, M. Voll, and K. Wesker. *Atlas d'anatomie prométhée : tête et neuro-anatomie*, volume 3. Pollina s.a. France, 2006.

[Ste99] K.N. Stevens. *Acoustic Phonetics*. MIT Press, 1999.

[STH96] B.H. Story, I.T. Titze, and E.A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America (JASA)*, 100(1) :537–554, 1996.

[Sto05] M. Stone. A guide to analyzing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 19(6-7) :455–502, Septembre-Novembre 2005.

[Tau05] C. Tauber. *Filtrage anisotrope robuste et segmentation par B-spline snake : application aux images échographiques*. PhD thesis, Institut National Polytechnique de Toulouse, Toulouse, France, Février 2005.

[TKNH04] H. Takemoto, T. Kitamura, H. Nishimoto, and K. Honda. A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions. *Acoustical Science and Technology*, 25(6) :468–474, 2004.

[TMAB08] M. Toda, S. Maeda, M. Aron, and M.O. Berger. Modeling Subject-Specific Formant Transition Patterns in /aSa/ Sequences. In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, pages 357–360, Strasbourg, France, 2008.

[Tod09] M. Toda. *Étude articulatoire et acoustique des fricatives sibilantes*. PhD thesis, Université Paris III, Paris, France, 2009.

[Vai06] J. Vaissière. *La Phonétique*. 2006.

[Vio95] P.A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachussets Institute Of Technology, Cambridge, MA, États-Unis, 1995.

[VLB+08] F. Vogt, J.E. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S.S. Fels. An Efficient Biomechanical Tongue model for Speech Research. In *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, Strasbourg, France, 2008.

[WDBP+05] B. Wrobel-Dautcourt, M.O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low cost stereovision based system for acquisition of visible articulatory data. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, pages 145–150, Vancouver, Canada, 2005.

[WH00] A. Wrench and W.J. Hardcastle. A multichannel articulatory speech database and its application for automatic speech recognition. In *Proceedings of the 5th International Seminar on Speech Production (ISSP)*, pages 305–308, Kloster Seeon, Allemagne, 2000.

[WTO+05] D.H. Whalen, M.K. Tiede, D.J. Ostry, H. Lehnert-LeHouillier, E. Vatikiotis-Bateson, and D.S. Hailey. The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language and Hearing Research*, 48 :543–553, Juin 2005.

[WZBS86] F. Wioland, J.P. Zerling, A. Bothorel, and P. Simon. Cinéradiographies des voyelles et consonnes du Français. *Travaux de l'Institut de Phonétique de Strasbourg (IPS)*, 1986.

[YA02] Y. Yu and S.T. Acton. Speckle reducing anisotrpic diffusion. *IEEE Transactions on Image Processing*, 11(11) :1260–1270, Novembre 2002.

[YCH92] A. Yuille, D. Cohen, and P. Hallinan. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2) :99–111, Août 1992.

[YK94] C.S. Yang and H. Kasuya. Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects. In *Proceedings of the International Conference on Language and Signal Language Processing (Interspeech)*, volume 2, pages 623–626, Yokohama, Japon, Septembre 1994.

[ZHFE07] C. Zeroual, P. Hoole, S. Fuchs, and J. Esling. EMA Study of the Coronal Emphatic and Non-emphatic Plosive Consonants of Moroccan Arabic. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, pages 397–400, Saarbrücken, Allemagne, 2007.

Résumé

La connaissance des positions et des mouvements des articulateurs (lèvres, palais, langue...) du conduit vocal lors de la phonation est un enjeu crucial pour l'étude de la parole. Puisqu'il n'existe pas encore de système permettant l'acquisition de ces positions et de ces mouvements, ce travail de thèse s'intéresse à la fusion de plusieurs modalités d'imagerie et de capteurs de localisation pour l'acquisition des positions des articulateurs dans l'espace et dans le temps. Nous décrivons un ensemble de protocoles et de méthodes pour obtenir et fusionner automatiquement un important volume de données échographiques (imageant en 2D la dynamique de la langue), stéréoscopiques (imageant en 3D la dynamique des lèvres), de capteurs électromagnétiques (capturant des points 3D de la langue et du visage), et d'Imagerie par Résonance Magnétique (IRM) pour acquérir en 3D l'ensemble des articulateurs en position statique. Nos contributions concernent plus particulièrement la synchronisation temporelle, le recalage spatial des données et l'extraction automatique des formes à partir des données (suivi de la langue dans les images échographiques). Nous évaluons la précision sur chaque donnée extraite, ainsi que sur l'ensemble des données fusionnées. Nous les validons enfin sur un modèle articulatoire existant. Ces travaux permettent l'obtention de données bien fondées pour la mise en place et l'étude de modèles articulatoires pour des applications en parole.

Mots-clés: vision par ordinateur, imagerie quadridimensionnelle, traitement d'images, fusion multimodale d'imagerie, données articulatoires

Abstract

There is no single technique that will allow all relevant behaviour of the speech articulators (lips, tongue, palate...) to be spatially ant temporally acquired. Thus, this thesis investigates the fusion of multimodal articulatory data. A framework is described in order to acquire and fuse automatically an important database of articulatory data. This includes : 2D Ultrasound (US) data to recover the dynamic of the tongue, stereovision data to recover the 3D dynamic of the lips, electromagnetic sensors that provide 3D position of points on the face and the tongue, and 3D Magnetic Resonance Imaging (MRI) that depict the vocal tract for various sustained articulations. We investigate the problems of the temporal synchronization and the spatial registration between all these modalities, and also the extraction of the shape articulators from the data (tongue tracking in US images). We evaluate the uncertainty of our system by quantifying the spatial and temporal inacuracies of the components of the system, both individually and in combination. Finally, the fused data are evaluated on an existing articulatory model to assess their quality for an application in speech production.

Keywords: computer vision, dynamic 3D images, image treatment, multimodal fusion of images, articulatory data