

Fusion de capteurs électromagnétiques et d'échographies pour le suivi de la langue

Coupling electromagnetic sensors and ultrasound images for tongue tracking

M. Aron

E. Kerrien

M.O. Berger

Y. Laprie

INRIA Lorraine - CNRS UMR7503 - Nancy Université

615, rue du Jardin Botanique, 54602 Villers-lès-Nancy, France
{aron,kerrien,berger,laprie}@loria.fr

Résumé

Cet article présente une méthode pour la fusion d'images échographiques avec des données 3D de capteurs électromagnétiques, afin de permettre un suivi complet de la langue durant la production de la parole. Les données électromagnétiques sont superposées aux échographies après un calibrage spatial et un recalage temporel. Après une courte étude préalable sur la validité d'acquisition des données électromagnétiques, des résultats de cette fusion dans des conditions expérimentales sont présentés sur plusieurs sons.

Mots Clef

Fusion, recalage multimodal, capteurs électromagnétiques, échographie, suivi de la langue.

Abstract

This paper describes a new method for coupling ultrasound images with three-dimensional electromagnetic data, to recover larger parts of the tongue during speech production. The electromagnetic data is superimposed on ultrasound images after spatial and temporal calibration. Successful fusion results are presented on various speech sequences. A complete setup for evaluation of the electromagnetic system is further presented.

Keywords

Fusion, electromagnetic sensors, ultrasound images, tongue tracking.

1 Introduction

L'inversion articulatoire en parole consiste à retrouver l'évolution temporelle et la position des articulateurs du conduit vocal (Fig. 1) à partir du signal acoustique. L'inversion s'oppose donc directement à la synthèse de la parole. On cherche donc à connaître précisément la forme et la position des articulateurs (surface du conduit vocal, de la langue, des lèvres...) et leur évolution dans le temps

pour pouvoir évaluer ces méthodes d'inversion. Les domaines d'application sont nombreux : la possibilité d'avoir un rendu visuel des positions et évolutions des articulateurs permettrait de faciliter l'apprentissage des langues étrangères ou encore de mettre en place des modèles réalistes de têtes parlantes.

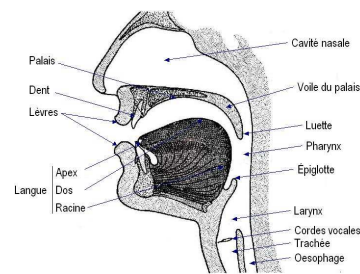


FIG. 1 – Coupe medio-sagittale du conduit vocal.

L'obtention de vastes corpus pour l'inversion de la parole nécessite de disposer d'outils automatisés (ou fortement automatisés) pour l'extraction dynamique des articulateurs. Le système d'acquisition des données articulatoires est donc un point crucial qui doit être aisé à mettre en oeuvre pour faciliter l'acquisition sur des locuteurs multiples, et ne pas nécessiter de matériels d'acquisition trop spécifiques. Il n'existe actuellement encore aucun système répondant à ces critères, l'idée étant de combiner plusieurs types de modalités d'acquisition et de fusionner les données que chacun apporte ([3]) : l'Imagerie à Résonance Magnétique (IRM) effectue des acquisitions de grande qualité du conduit vocal en 3D mais nécessite plusieurs dizaines de secondes pour des acquisitions précises. Les images échographiques permettent un suivi 2D de la langue en temps réel mais occultent certaines parties, notamment l'apex (le bout de la langue) car l'os de la mâchoire et l'air empêchent la propagation des ultrasons ([10]). Les données électromagnétiques permettent à partir de capteurs d'obtenir des informations 3D en temps réel,

mais sur un point de l'espace seulement.

1.1 Objectif

Notre objectif à long terme est de fusionner toutes ces données (IRM, échographie, son, et données capteurs) pour la construction d'un modèle articulaire 3D dynamique ([1]). Cet article traite plus précisément du recalage multimodal temporel entre les images échographiques et les données capteurs. Cette première étape est fondamentale pour la mise en place des méthodes d'inversion, la langue étant l'un des articulateurs les plus difficiles à suivre lors de la production de la parole.

1.2 Dispositifs d'acquisition

[4] a déjà utilisé des données électromagnétiques avec une autre modalité d'acquisition pour le suivi de la langue : dans son travail, la langue est modélisée à partir d'images IRM, et les données électromagnétiques provenant d'un articulographe permettent seulement d'ajouter les aspects dynamiques a-posteriori. Nous avons adopté une stratégie différente : les données dynamiques sont acquises en 2D par l'échographe et des capteurs collés sur la langue permettent de compléter l'information pour l'apex. Les images échographiques sont acquises en maintenant la sonde sous le menton : les capteurs sont aussi utilisés pour suivre la position de la sonde échographique en 3D, ce qui permet au locuteur de pouvoir bouger la tête librement. Ce procédé évite d'utiliser un système de contention pouvant influencer sur la production de la parole, comme le Head Transducer Support System (HATS) de [11]. Pour nos acquisitions électromagnétiques, nous avons utilisé le système Aurora (Northern Digital Inc, Waterloo, Canada).

1.3 Contributions

Les contributions de ce travail sont les suivantes : (i) le système Aurora a été évalué afin de vérifier qu'il répondait bien aux contraintes de précision inhérentes au suivi de la langue. (ii) Une méthode d'acquisition des données électromagnétique et d'images échographiques est présentée. (iii) Le problème de la fusion de ces données, tant sur le plan spatial (calibrage) que sur le plan temporel (recalage), est traité. (iv) Des résultats en conditions réelles sont présentés pour valider notre travail.

2 Les capteurs électromagnétiques

2.1 Présentation du matériel

Le système est composé d'un générateur de champ électromagnétique (GEM), d'une interface de contrôle, et de capteurs miniatures (0.8 mm x 8 mm). Un champ électromagnétique est émis par le GEM, et les capteurs placés dans ce champ fournissent la transformation du repère local au capteur au repère fixe attaché au GEM. Cette transformation à 5 degrés de liberté (DL), que nous appellerons T_{em}

par la suite, est la translation (3 DL) et la rotation (2 DL) entre les deux repères (l'information manquante étant la rotation du capteur autour de son axe z). Pour des raisons de précision, Northern Digital Inc. conseille de restreindre le volume d'acquisition à un cube d'environ 50 cm x 50 cm x 50 cm.

Le constructeur spécifie une précision de 1 – 2 mm pour les translations et 0.6° pour les rotations à l'intérieur du volume d'acquisition. Le système supporte 8 capteurs à 5DL, la fréquence d'acquisition des données étant de 40 Hz si moins de 6 capteurs sont utilisés, et de 20 Hz si 6 capteurs ou plus sont branchés.

Enfin, deux capteurs à 5DL peuvent être utilisés pour fabriquer un capteur à 6DL, permettant ainsi de récupérer toutes les transformations dans l'espace. C'est le cas du stylet MagTrax ou du capteur 'MagTrax 6DOF', que nous avons utilisés dans nos expérimentations.

2.2 Mesures de répétabilité

Nous avons fixé un capteur à 5DL sur une table micrométrique pour évaluer la répétabilité des mesures fournies par le système Aurora. Cette table a une précision de 0.013° en rotation et 0.48 mm en translation. Les tests ont été effectués pour trois positions du capteur sur la table : près du GEM (position 1 à 5 cm), à distance moyenne (position 2 à 30 cm) et enfin la plus éloignée possible (position 3 à 50 cm). Il est important de noter que la position 2 est celle qui correspond à la configuration la plus fréquemment utilisée, notamment dans le cadre de nos futures expérimentations. Chaque mesure a été répétée 100 fois pour chaque position, et comparée à celle donnée par la table micrométrique.

Nous avons répété le même protocole expérimental en fixant cette fois le capteur à 5DL sur une sonde échographique, afin d'évaluer l'influence des perturbations de cette dernière sur la précision des données électromagnétiques. Les résultats de ces diverses expérimentations sont consignés dans la figure 1.

La moyenne de l'erreur en translation est inférieure à 1 mm et en rotation à 0.5° pour les positions 1 et 2. Par contre, les erreurs augmentent de manière significative pour la position 3 qui correspond à une position peu utilisée en pratique. Nos résultats sont similaires à ceux spécifiés dans [8] et [6]. On remarque aussi que les erreurs en translation augmentent lorsque les capteurs sont fixés sur la sonde échographique : ce résultat prévisible est dû à la présence de matériaux ferromagnétiques dans la sonde, altérant le champ électromagnétique et faussant ainsi les mesures du système. On note cependant que les erreurs restent inférieures à 1 mm pour les positions 1 et 2.

2.3 Comparatif avec l'articulographe

Il est intéressant de comparer ce système d'acquisition de données électromagnétiques avec l'articulographe, couramment utilisé par les cliniciens pour l'étude de la parole. L'articulographe (EMA pour ElectroMagnetic Articulograph, Carstens, Lengler, Allemagne) est le seul sys-

	capteur 5DL		capteur 6DL sur la sonde échographique	
	Moyenne des erreurs en translation (en mm)	Moyenne des erreurs en rotation (en degrés)	Moyenne des erreurs en translation (en mm)	Moyenne des erreurs en rotation (en degrés)
Position 1	0.31	0.39	0.87	0.25
Position 2	0.53	0.50	0.76	0.20
Position 3	3.58	0.84	3.39	0.30

FIG. 2 – Précision des capteurs électromagnétiques

tème commercial existant à base de capteurs électromagnétiques pour l'acquisition de données dynamiques sur la langue. Le modèle le plus récent, l'AG500, est composé de six générateurs de champs électromagnétiques placés dans un cube en plexiglas. Le patient fixe des capteurs sur sa langue, son visage, ses lèvres..., positionne sa tête dans le cube et les données peuvent être enregistrées. En comparaison, les capteurs Aurora sont plus longs (+2mm) mais plus fins que les capteurs EMA. En terme de précision, les deux systèmes semblent être équivalents. Un point fort de l'EMA est son taux d'acquisition des données avec une fréquence de 200 Hz. Le système Aurora est pour le moment à 40 Hz avec en prévision un taux avoisinant les 65 Hz. Cependant, le système Aurora permet d'acquérir et de visualiser les données en temps réel alors que EMA nécessite plusieurs heures de calcul pour quelques minutes d'acquisition. De plus, les capteurs Aurora peuvent être intégrés et utilisés dans des outils personnalisés où le développeur a un contrôle complet de ce qu'il fabrique. Ces derniers avantages nous ont conduit à opter pour le système Aurora, aussi substantiellement moins coûteux que l'EMA.

3 Le couplage échographie - capteurs électromagnétiques

3.1 L'échographe

Matériel. Nous avons à disposition une machine échographique Logiq5 (GE Healthcare, the Chalfont St. Giles, Royaume-Uni) avec une sonde micro-convexe 8C produisant des ultrasons à des fréquences comprises entre 5 MHz et 9 MHz.

Utilisation des échographies pour le suivi de la langue.

En positionnant la sonde sous le menton (Fig.3), on acquiert des images de la surface de la langue dans le plan medio-sagittal de la tête. Ces coupes 2D permettent d'obtenir des informations sur la dynamique de la langue en temps réel. Les réglages utilisés pour l'acquisition de ces images se doivent d'être des compromis entre la fréquence de la sonde, la profondeur de pénétration, la largeur de la zone visée, et la fréquence d'acquisitions des images. Par exemple, plus la zone visée est large, plus la fréquence d'acquisition des images sera faible. La surface de la langue, lors de la parole, est située à une distance de 3 cm à 7 cm du menton, ce qui impose de fixer une profondeur de pénétration suffisamment importante. Nous sommes par-

venus à acquérir des séquences vidéo à une fréquence de 50 Hz pour de larges zones (8 cm) et à une fréquence de 150 Hz pour des zones beaucoup plus étroites (3 cm) (voir Fig.4).

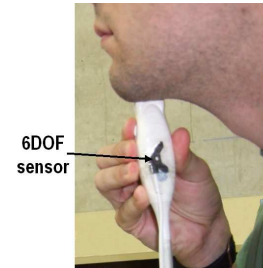


FIG. 3 – Sonde échographique sous le menton avec un capteur à 6DL.

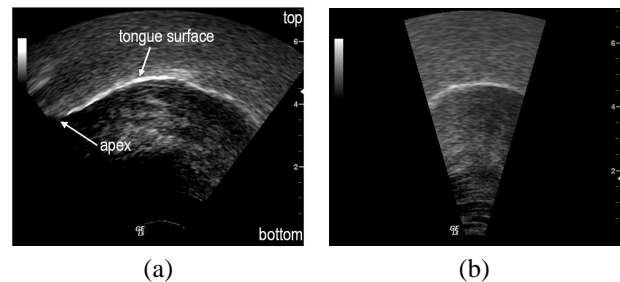


FIG. 4 – La langue durant un /a/. (a) : image échographique d'une séquence à 66 Hz, largeur de 6.8 cm. (b) : image échographique d'une séquence à 152 Hz, largeur de 3.1 cm.

Les réglages que nous avons choisis pour le suivi de la langue nous permettent d'obtenir les caractéristiques suivantes pour les séquences acquises :

- une fréquence d'acquisition de 66 Hz
- une taille d'images de 532x434 pixels
- une résolution de 0.017 cm/pixel
- une profondeur de pénétration maximale de 8 cm

3.2 La fusion des données EM et échographiques

Les images échographiques permettent donc d'obtenir un suivi en temps réel de la langue. Malheureusement, l'extrémité de la langue (l'apex) n'est pas visible sur les images car l'air et l'os de la mâchoire stoppent les ondes ultrasonores et empêchent l'acquisition. L'idée est donc de pla-

cer un capteur électromagnétique à 5DL sur l'apex afin de compléter l'information. On utilise aussi un second capteur à 5DL que l'on place sur le dos de la langue, sur le plan medio-sagittal afin de vérifier la cohérence entre les données électromagnétiques et échographiques.

Enfin, un troisième capteur à 6DL est placé sur la sonde échographique afin de repositionner les données des capteurs sur la langue dans un repère lié à la sonde. Cette étape nécessite un calibrage spatial et un recalage temporel, explicités dans la partie suivante.

4 Calibrage spatial et recalage temporel des données

4.1 Calibrage spatial

Le capteur à 6DL positionné sur la sonde échographique permet d'obtenir la transformation complète (translation et rotation) entre le repère local lié au capteur et le repère global lié au GEM. L'étape de calibrage spatial permet de localiser précisément les images échographiques dans l'espace 3D, c'est-à-dire calculer la transformation rigide T_c entre le plan image et le capteur à 6DL. Une fois cette transformation connue, les positions des capteurs sur la langues peuvent être exprimées dans le plan image lié à la sonde, et les deux modalités sont alors exprimées dans le même repère.

Soit \mathcal{R}_{us} le repère 3D lié à la sonde tel que chaque pixel $p = (u, v)$ dans l'image échographique corresponde à un point 3D $P_{us} = (u, v, 0)$. Ce point 3D s'exprime alors, dans le repère global lié au GEM (voir Fig. 5) :

$$P_{em} = T_{em} \cdot T_c \cdot P_{us}$$

où T_{em} est la transformation du capteur à 6DL et T_c est la transformation rigide cherchée (3 paramètres pour la translation et 3 paramètres pour la rotation).

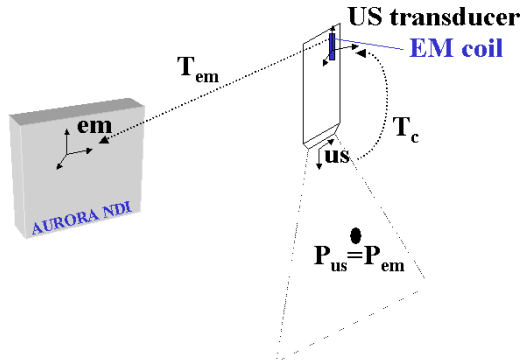


FIG. 5 – Repères et transformations utilisés

Le fait de positionner un capteur sur la sonde échographique apporte une souplesse d'utilisation non négligeable pour les expérimentations puisque la tête de l'utilisateur peut bouger librement, comme dans le système Haskins Optically Corrected Ultrasound System (HOCUS[12]).

4.2 Protocole expérimental pour le calibrage spatial

Afin d'estimer les six paramètres, un fantôme de calibrage, dont les propriétés géométriques 3D sont connues, est utilisé. Il permet une mise en correspondance de points visibles à la fois dans les images échographiques et dans le système électromagnétique. La matrice T_c peut alors être estimée puisque P_{us} (le point détecté dans les images échographiques), T_{em} (donné directement par le système électromagnétique), et P_{em} (le point détecté par le système électromagnétique) sont connus. On trouve dans la littérature de nombreux types de fantômes permettant l'estimation des paramètres de calibrage, chacun apportant sa contribution au niveau de la précision, de la facilité de mise en oeuvre... Parmi les plus utilisés, on trouve les fantômes de type point d'intersection ('cross-wire' en anglais), d'ensemble de points, les fantômes plans ou avec pointeur de localisation (on reporte le lecteur à [9] pour un état de l'art plus précis sur les différents types de fantômes existants). Il est important de noter qu'il n'existe pas aujourd'hui de technique de calibrage surpassant les autres, certaines étant très difficiles à mettre en oeuvre mais légèrement plus précises, d'autres moins efficaces mais beaucoup plus simples à utiliser...

Le fantôme que nous avons choisi d'utiliser est celui de [7] : deux capteurs à 5DL sont fixés aux extrémités (P_0 et P_1) d'une tige en bois de longueur 25 cm et de diamètre environ 3 mm. Cette tige forme alors une ligne dont la position est connue dans l'espace 3D grâce aux capteurs électromagnétiques. Notre choix s'est porté sur un tel fantôme car l'intersection du plan échographique avec la tige est bien plus facile à localiser qu'un point. De plus, sa fabrication et sa mise en oeuvre est facile et rapide. Ensuite, le fantôme est plongé dans de l'eau à température ambiante, et on acquiert des images pour différentes positions et orientations de la sonde (une trentaine d'images). Pour chaque image, la tige apparaît sous la forme d'une ellipse, très souvent bruitée (souvent entre 20 et 25 pixels de diamètre), mais dont le centre de gravité peut être manuellement pointé par un utilisateur (Fig.7). Même si ce pointage reste relativement peu précis, le fait d'utiliser un grand nombre d'images, et donc de positions de la sonde, compense ce problème. Les expérimentations ayant été effectuées dans de l'eau à température ambiante ($20^\circ C$), [2] a montré qu'il fallait corriger les distances observées sur les images échographiques par un facteur correspondant à la vitesse des ultrasons dans les tissus humains ($\approx 1540m/s$) divisé par la vitesse des ultrasons dans le milieu observé (soit $1485m/s$ pour l'eau à $20^\circ C$). Ainsi, un objet apparaissant à 10 cm sur les images échographiques est physiquement situé à $10 * 1540/1485 = 10.4cm$ du point d'origine des émissions ultra-sonores.

Une fois les positions du fantôme dans les images échographiques connues, le problème du calibrage revient à mini-

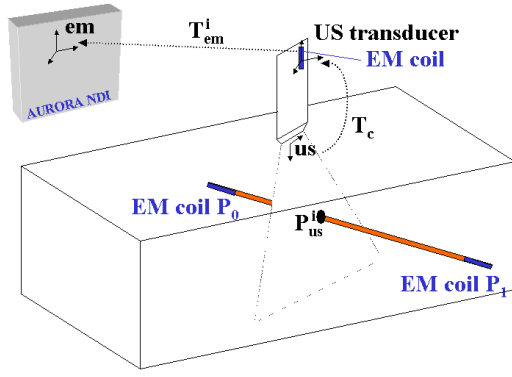


FIG. 6 – Protocole expérimental avec le fantôme pour le calibrage image échographique / capteur électromagnétique.

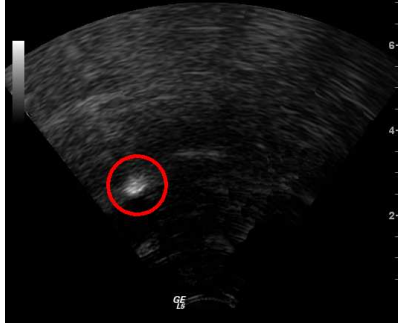


FIG. 7 – Image échographique du fantôme.

miniser le système d'équations suivant ([7]) :

$$[\tilde{T}_c] = \underset{[T_c]}{\operatorname{argmin}} \sum_i \|(P_1 - P_0) \times (T_{em}^i \cdot T_c \cdot P_{us}^i - P_0)\|^2$$

où \times représente le produit scalaire, et P_0, P_1 les extrémités de la tige exprimées dans le repère électromagnétique. Pour nos expériences, nous avons utilisé une minimisation de Powell ([5]) pour retrouver les six paramètres de la transformation rigide.

4.3 Recalage temporel

Une fois le calibrage spatial effectué, les données des capteurs à 5DL fixés sur la langue peuvent être exprimées dans le repère image. L'étape suivante consiste à estimer le délai entre le début des acquisitions des images et des données électromagnétiques pour recalibrer temporellement les deux modalités.

Pour cela, les données capteurs (à 40 Hz) sont sur-échantillonnées par interpolation linéaire à la même fréquence que les images échographiques (à 66 Hz) et sont ensuite projetées dans les images échographiques. Notons t_0 le début de l'acquisition capteur, f la fréquence rééchantillonnée commune ($f = 1/66$), et $c_{t,i}$ la position à l'instant t sur l'image i du capteur c . Si d est le délai d'acquisition alors la donnée image i et la donnée capteur acquise à $t_0 + df + if$ coïncident et le point $c_{t_0+df+if,i}$ appartient

donc au contour de la langue détecté sur l'image i . Calculer d à partir de plusieurs images revient donc à minimiser

$$d = \min_d \sum_i \operatorname{dist}(c_{t_0+df+if,i}, l_i)$$

où l_i représente le contour de la langue détournée par l'utilisateur sur l'image i .

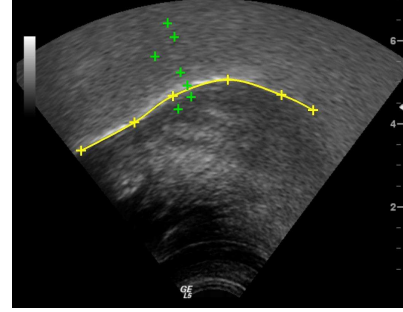


FIG. 8 – Langue détournée et positions du capteur sur le dos de la langue projetées sur l'image échographique.

5 Expérimentations sur un locuteur

5.1 Protocole expérimental

Nous avons ensuite testé notre système sur un locuteur. Pour cela, les capteurs destinés à être fixés dans la bouche ont été protégés par un fin film plastique créé par Northern Digital Inc. Les capteurs sont ensuite collés dans le plan medio-sagittal de la langue à l'aide d'*Histoacryl*¹, une colle chirurgicale utilisée pour les cicatrises (similaire à la *Super - Glue*). Ils restent fixés sur la langue durant environ 30 minutes.

Quatre séquences ont été testées : "/au/, /atu/, /aku/, /ao/, /ako/, /ae/, /ake/, /ate/" (3 fois) et la phrase complète "la bise et le soleil se disputaient, chacun assurant qu'il était le plus fort" (une fois). La fréquence d'acquisition des images échographiques était de 66 Hz.

5.2 Résultats

Il est difficile de valider quantitativement les résultats. Cependant, la représentation visuelle permet de tester leur cohérence, aussi bien dans le domaine spatial que dans le domaine temporel. Les quatre tests sur les séquences ont été concluants, les deux capteurs apparaissant correctement sur la surface de la langue dans les images échographiques. De plus, ils suivent temporellement la surface de la langue et ce malgré la fréquence d'acquisition des capteurs (40 Hz) plus basse que celle des images (66 Hz). Les résultats sont d'autant plus convaincants sur la séquence de la phrase où le mouvement de la langue est parfois très rapide.

Les images suivantes (Fig.9) montrent la fusion des données électromagnétiques avec les images échographiques,

¹<http://www.bbbaun.com>

où les deux capteurs sont représentés par des croix. La Fig.9.a montre les deux capteurs sur la surface de la langue, et la Fig.9.b montre l'utilité du capteur positionné sur l'apex : en effet, dans cette image, l'apex passe au dessus de l'os de la mâchoire et n'est donc plus visible dans les images. Le capteur permet alors de retrouver sa position. Les séquences vidéos complètes peuvent être visualisées sur notre site Internet <http://magrit.loria.fr/Confs/Orasis07>

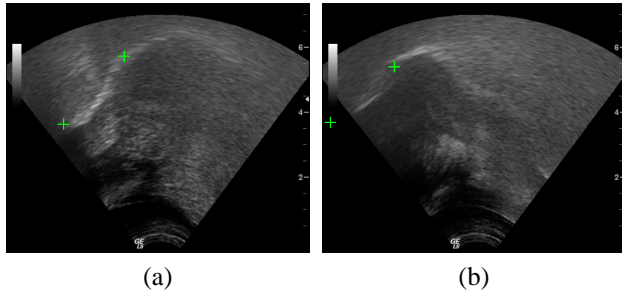


FIG. 9 – Les capteurs (croix) sur la langue : (a) /u/ de /aku/. (b) /k/ de /ake/.

6 Conclusion

Nous avons présenté un protocole complet d'acquisition de données à partir d'images échographiques et de données capteurs en vue du suivi de la langue. Ces acquisitions incluent l'apex rarement visible dans les images échographiques seules. Après avoir montré que le système Aurora était suffisamment précis pour notre application, nous avons détaillé les procédures de calibrage spatial et de recalage temporel permettant de fusionner les informations apportées par les deux modalités. Des expérimentations ont enfin été présentées, validant visuellement la fusion. Ce système représente une alternative à l'articulographe, permettant d'obtenir des données dynamiques de la langue avec seulement deux capteurs électromagnétiques collés sur la langue. La validation de notre travail est pour le moment uniquement visuelle. Outre une validation plus quantitative, nos efforts vont désormais se porter sur la mise en place de méthodes robustes et automatique pour le suivi de la langue.

Remerciements

Ce travail s'inscrit dans le cadre du projet Européen ASPI (IST-2005-021324). Nous remercions Sigmar Froehlich et les personnes de Nothern Digital Inc. en Allemagne pour leur aide et leur support.

Références

- [1] Badin, P., Bailly, G., Revéret, L., Baciú, M., Segebarth, C., and Savariaux, C. Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3) :533-553.
- [2] Bilaniuk, N. and Wong, G. Speed of sound in pure water as function of temperature. *Journal of the Acoustical Society of America*, 93 : 1609–1612, 1993.
- [3] Engwall, O. Are static MRI measurements representative of dynamic speech ? In *ICSLP-2000*, pages 17–20, 2000.
- [4] Engwall, O. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, 410 (2-3) : 303–329, 2003.
- [5] Flannery, B., Teukolsky, S., and Vetterling, W. *Numerical Recipes, 2nd Edition*, Cambridge University Press, 1993.
- [6] Hummel, J., Figl, M., Kollmann, C., and Bergmann, H. Evaluation of a miniature electromagnetic position tracker. *Med. Phys.*, 290 (10) : 2205–2212, 2002.
- [7] Khamene, A. and Sauer, F. A novel phantom-less spatial and temporal ultrasound calibration method. In *MICCAI 2005*, pages 65–72, 2005.
- [8] Kirsch, S. Accuracy assessment of the electromagnetic tracking system aurora. Technical report, NDI Europe GmbH, 2005.
- [9] Mercier, L., Lango, T., Lindseth, F., and Collins, D. A review of calibration techniques for freehand 3-D ultrasound systems. *Ultrasound in Med. and Biol.*, 310 (4) : 449–471, 2005.
- [10] Stone, M. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 190 (6-7) : 455–502, Sept-Nov 2005.
- [11] Stone, M. and Davis, E. A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of the Acoustical Society of America*, 980(6) : 3107–3112, 1995.
- [12] Whalen, D., Iskarous, K., Tiede, M., Ostry, D., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., and Hailey, D. The haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research*, 480(3) : 543–553, 2005.