CARACTERISATION DES MECANISMES DE PRODUCTION DE LA PAROLE: UNE APPROCHE BIOMETRIQUE ET MODELISATRICE MONO-LOCUTEUR ET MULTI-DISPOSITIFS

CHARACTERISATION OF SPEECH PRODUCTION MECHANISMS: A SINGLE-SPEAKER AND MULTI-SETUP BIOMETRIC AND MODELLING APPROACH

Hommage à notre collègue et ami Bernard Teston

BADIN PIERRE[1], SAVARIAUX CHRISTOPHE[2], BAILLY GERARD[1], ELISEI FREDERIC[1], BOË LOUIS-JEAN[1]

INTRODUCTION

Depuis plus de trente ans, Bernard Teston, soutenu et épaulé par ses complices aixois, a contribué de manière considérable au développement et à l'exploitation de systèmes d'acquisition de données articulatoires et aéroacoustique pour l'étude de la production de la parole, qu'elle soit normale ou pathologique. Depuis 1967, Bernard Teston n'a eu de cesse de développer ou de mettre en œuvre des dispositifs pour l'acquisition simultanée et synchrone de divers types de signaux sur des locuteurs : ElectroMyoGraphie (EMG, [1]), ElectroPalatoGraphie (EPG, [2]), aéroacoustique (station EVA : [3], [4], [5]), Articulographie ElectroMagnétique (EMA, [6]), fibroscopie ([7]), cinéradiographie ([7]). Un triple sceau marque cette carrière au service de la recherche en parole : (1) le soin et la rigueur apportés au développement des dispositifs et le souci constant de leur évaluation ; (2) la volonté de faire cohabiter des dispositifs complémentaires pour augmenter la richesse et la complétude des informations obtenues (acoustique + EVA + EMA, acoustique + EVA + EPG, acoustique + cinéradiographie + fibroscopie) ; et (3) l'ambition d'acquérir des bases de données comprenant de très nombreux locuteurs permettant de construire des outils d'évaluation des dysfonctionnements de la production de la parole,et donc de diagnostic clinique. Notons en particulier l'étude de [8] portant sur 449 locuteurs (incluant 391 patients dysphoniques) citée par [9] : cette étude montre qu'une classification de la qualité de la voix établie automatiquement à partir d'une combinaison de six paramètres physiques (étendue vocale, coefficient de Lyapunov, pression sous-glottique estimée, temps maximal de phonation, débit d'air oral et rapport signal/bruit) mesurés instrumentalement à l'aide de la station EVA aboutit à des résultats identiques à ceux établis de manière perceptive par des experts dans 82% des cas

En écho à cette expérimentation massive sur une large population de locuteurs, le propos du présent article est de décrire l'approche complémentaire menée depuis plus de vingt ans à l'Institut de la Communication Parlée (ICP), devenu en 2007 le Département Parole et Cognition du laboratoire GIPSA, à l'aide de dispositifs souvent similaires à ceux utilisés par Bernard Teston et ses collaborateurs, pour étudier un nombre beaucoup plus limité de locuteurs – souvent un seul – mais avec l'ambition de caractériser et modéliser très finement les phénomènes sous-jacents aux mécanismes de production de parole. Cet article décrit donc les différents dispositifs employés, leurs avantages et leurs inconvénients, les combinaisons que nous avons utilisées, et les modèles qui ont été élaborés à partir des données acquises. La dernière section de l'article décrit les aboutissements les plus marquants obtenus grâce à cette approche : synthèse articulatoire des fricatives, clones orofaciaux 3D basés sur la modélisation articulatoire, et plus récemment inversion de l'acoustique vers l'articulatoire à l'aide de modèles statistiques basés données.

La philosophie sous-jacente à notre approche est globalement la suivante. Nous analysons, caractérisons et modélisons les locuteurs individuellement afin d'échapper aux problèmes de normalisation et d'éviter les risques de brouiller les phénomènes importants, en particulier au niveau des stratégies de contrôle articulatoire. En effet, chaque locuteur choisit et adapte ses stratégies en fonction de sa propre morphologie, qui peut être significativement différente de celle des autres locuteurs (ne serait-ce qu'entre locuteurs hommes, femmes et enfants), afin de produire un signal de parole compréhensible pour son interlocuteur.

[1] Pierre.Badin@gipsa-lab.grenoble-inp.fr, *Auteur correspondant* ; Gerard.Bailly@gipsa-lab.grenoble-inp.fr ; Frederic.Elisei@gipsa-lab.grenoble-inp.fr ; Louis-Jean.Boe@gipsa-lab.grenoble-inp.fr
GIPSA-lab / DPC, UMR 5216, CNRS – Grenoble INP – Université Stendhal, France, 11 rue des Mathématiques, BP 46 - 38402 Saint Martin d'Hères cedex, France ;
[2] Christophe.Savariaux@gipsa-lab.grenoble-inp.fr
GIPSA-lab / DPC, UMR 5216, CNRS – Grenoble INP – Université Stendhal, France, BP 25 - 38040 Grenoble cedex 9;
Article reçu le 10.05.2012, accepté le 12.09.2012

Comme souligné ci-dessus, l'étude fine nécessite la mesure de phénomènes caractérisant la production de parole dans les différents domaines tels que la cinématique des articulateurs, la géométrie du conduit vocal, ou encore les pressions et débits aérodynamiques et acoustiques. Ces mesures simultanées sont souvent difficiles ou impossibles à mettre en œuvre pour de multiples raisons : résolutions temps-fréquence et spatiales différentes (par ex. IRM anatomique vs. EMA), ou dispositifs difficilement compatibles à cause de leur encombrement (EMA + EPG) qui peut rendre la tâche du locuteur trop complexe.

Ainsi, en fonction des possibilités, nous collectons les signaux associés à ces domaines soit en mettant en œuvre simultanément de manière synchrone plusieurs dispositifs complémentaires, soit en demandant aux locuteurs de répéter les mêmes corpus avec une élocution aussi semblable que possible dans différents dispositifs dont la mise en œuvre expérimentale est incompatible. Nous utilisons ensuite la modélisation pour fusionner les signaux de manière cohérente.

Finalement, il est important de mentionner que notre approche s'appuie sur un triple principe : pour une question scientifique donnée, (1) nous choisissons le locuteur et les dispositifs adéquats ; (2) nous concevons le corpus de parole susceptible de répondre à la question ; (3) nous construisons un modèle à partir de ces données. Ce modèle et sa validation constituent la réponse à la question, ou suggèrent des pistes qui permettent d'affiner la question.

Dans la mesure où nous nous intéressons en premier lieu aux aspects articulatoires et (aéro-) acoustiques de la production de parole, nous avons développé ou mis en œuvre des dispositifs destinés à mesurer des paramètres liés à l'un ou l'autre de ces domaines. Nous présenterons les dispositifs qui permettent de mesurer la fonction de transfert acoustique du conduit vocal, les débits d'écoulement d'air ou la pression intra-orale. Pour le domaine articulatoire ou géométrique, nous décrirons les diverses méthodes d'imagerie médicales utilisées, ainsi que les traitements nécessaires pour en tirer des données utiles pour la parole.

Le Tableau 1 récapitule les méthodes mises en œuvre et leurs principales caractéristiques. Globalement, ces techniques diffèrent par leur résolution spatiale, leur résolution temporelle, leur caractère plus ou moins invasif ainsi que leur niveau de risque pour la santé du locuteur.

Notons que le locuteur désigné dans la suite par « PB » a été sujet de toutes les expériences qui vont être décrites. Comme nous l'avons indiqué plus haut, les données permettent d'alimenter des modèles, et l'utilisation d'un seul et même locuteur maximise la cohérence de ces modèles entre eux.

TABLEAU 1. — METHODES DE MESURES EN PRODUCTION DE PAROLE.

	Méthode	Résolution temporelle	Résolution spatiale	Commentaires
Aérodynamique / Acoustique	Microphone acoustique			Mesure du signal acoustique.
	Mesure de fonctions de transfert acoustique			Banc de mesure par excitation transcutanée au niveau du larynx et mesure de pression aux lèvres.
	Capteur de pression			Mesure de pression basse fréquence.
	Pneumotachographe de Rothenberg			Mesure du débit acoustique grâce à un masque facial percé de trous recouverts d'une grille métallique. Perturbe légèrement l'articulation.
	Station EVA 2			Mesure du débit acoustique : flux oral grâce à un masque facial en latex qui guide le débit vers une grille métallique; flux nasal grâce à des tuyaux en latex connectés aux narines par des olives souples et guidant le total du flux vers une autre grille métallique
Géométrique / Articulatoire	IRM	(~5 - 40 sec. / articulation)	plans 2D ~1mm / pixel	25-50 coupes 2D. Reconstruction 3D possible. Durée d'acquisition encore longue. Structures osseuses non visibles. La position couché sur le dos perturbe légèrement l'articulation.
	Tomodensitométrie	(~10-20 sec. / articulation)	plans 2D ~0 5 mm / pixel	~150 coupes 2D. Reconstruction 3D précise. Structures osseuses visibles. La position couché sur le dos perturbe légèrement l'articulation. Danger lié aux rayons X.
	Téléradiographie	(~2 sec. / articulation)	continu 2D	Projection sur un plan de l'ensemble des structures de la tête. Complet, mais difficile à tracer. Dangereux pour la santé.
	Labiométrie vidéo	50 / 400 Hz	continu 2D ½ / ~400 points 3D	Enregistrement vidéo mono/multi caméra du visage avec marqueurs (maquillage lèvres, billes collées sur le visage). Reconstruction 3D.
	Cinéradiographie	50 Hz	continu 2D	*cf.* téléradiographie, mais en mouvement. Films très longs à dépouiller manuellement.
	Articulographie Electromagnétique (EMA)	500 Hz	10 - 15 points 2D ou 3D	Bobines électromagnétiques collées sur les lèvres, la langue, les incisives, le velum. Perturbe plus ou moins l'articulation.
	ElectroPalatoGraphie (EPG)	50 Hz	~64 points 3D	Détecte les contacts entre la langue et un palais artificiel porté par le locuteur. Perturbe légèrement l'articulation.
	Imagerie par échographie ultrasonique de la langue	50-70 Hz	continu 2D	Principe de l'échographie. Seule la partie centrale de la langue est observable. Calage par rapport aux structures osseuses non maîtrisé.
	Vidéo rapide des cordes vocales	4000 Hz	continu 2D	Images des cordes vocales vues de dessus par l'intermédiaire d'une fibre optique nasale.

FONCTIONS DE TRANSFERT ACOUSTIQUE

Pour caractériser et modéliser la relation articulatori-acoustique, il est essentiel de pouvoir mesurer la fonction de transfert acoustique du conduit vocal sur des locuteurs humains. En effet, pour valider les modèles acoustiques de conduit vocal (cf. [10]), il est nécessaire de pouvoir mesurer les caractéristiques de résonance du conduit vocal sans qu'elles soient perturbées par le signal d'excitation glottique ou consonantique. Nous avons donc développé un dispositif permettant une telle mesure ([11], [12]). Ce dispositif est basé sur l'excitation transcutanée du conduit vocal au niveau du larynx par un pot vibrant alimenté par un bruit blanc, et sur l'enregistrement par un microphone du son rayonné aux lèvres (cf. FIGURE 1). Il a été ensuite amélioré par le remplacement du bruit blanc par une séquence pseudo-aléatoire connue qui autorise une phonation d'amplitude modérée pendant la mesure ([13]). Il a ainsi été possible de caractériser et modéliser les sources de bruit des consonnes fricatives pour le locuteur PB ([14]), ce qui a constitué, en conjonction avec la modélisation articulatoire du même locuteur, une contribution importante à la synthèse des fricatives par modélisation articulatoire, aérodynamique et acoustique ([15]). Notons par ailleurs qu'une autre étude acoustique, sur des moulages de lèvres réalisés sur le même locuteur, a permis d'établir expérimentalement une équivalence acoustique du pavillon labial utile pour la modélisation acoustique du conduit vocal ([16]).

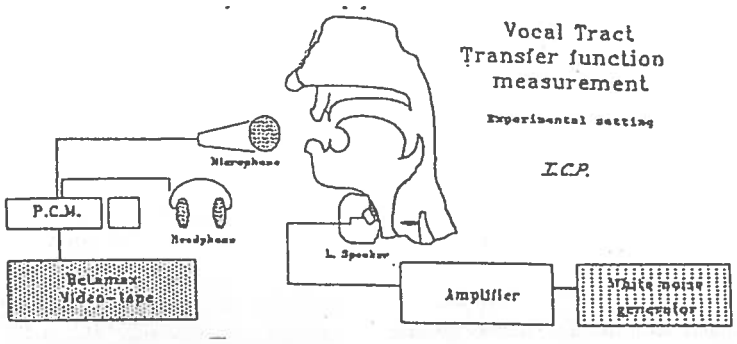

FIGURE 1. — *Schéma du dispositif de mesure de fonction de transfert acoustique du conduit vocal.*

AERODYNAMIQUE

Les sources d'excitation acoustique du conduit vocal, que ce soit la vibration des cordes vocales ou les bruits de friction générés au niveau des constrictions pour les consonnes fricatives, ont des caractéristiques étroitement liées à l'état de l'écoulement aérodynamique dans le conduit vocal. Une première étude a permis d'établir des relations quantitatives entre le niveau global et les caractéristiques spectrales du son rayonné aux lèvres en fonction de la chute de pression à travers la constriction orale et l'aire transversale de cette constriction à partir d'enregistrements réalisés sur le locuteur PB ([17], [18]). Ces mesures de débit aux lèvres et de pression ont été réalisées à l'aide d'un masque pneumotachographique (cf. FIGURE 2) dont nous avons par ailleurs évalué les caractéristiques ([19]). Notons au passage que ces mesures ont également permis de déterminer l'aire transversale aérodynamiquement équivalente de la constriction grâce à l'*équation de l'orifice* qui la relie à la chute de pression et au débit ; ceci a par ailleurs constitué un excellent complément à la détermination des fonctions d'aire des fricatives qui doivent être particulièrement précises dans la région de la constriction, et pour laquelle les techniques d'imagerie médicale disponibles n'offrent pas la résolution suffisante.

Des mesures de fonction de transfert acoustique et des mesures aérodynamiques sur le même locuteur PB ont ensuite permis de construire un modèle fonctionnel des variations du niveau global et des caractéristiques spectrales de la source de bruit de différentes consonnes fricatives en fonction des variables aérodynamiques ([20], [21]).

D'un autre côté, les paramètres d'un modèle à deux masses des cordes vocales ont été ajustés afin de l'adapter au mieux au comportement mesuré sur le même locuteur ([22]). En conjonction avec un modèle simplifié d'écoulement aérodynamique (voir dans [23]), ces données et modèles ont été au cœur de la synthèse articulatoire des consonnes fricatives basée sur le même locuteur PB ([15]).

FIGURE 2. — *Masque pneumotachographique de Rothenberg : la mesure de débit se fait par la mesure de la différence de pression de part et d'autre de la grille calibrée à travers de laquelle s'écoule le flux d'air lorsque le locuteur applique le masque sur le visage. On distingue également le tuyau qui est inséré dans la bouche du locuteur pour mesurer la pression buccale.*

TELE- ET CINE-RADIOGRAPHIE

La radiographie permet de fournir une image 2D des articulateurs 3D en procédant par une illumination latérale par rayons X du complexe orofacial. Elle a constitué un dispositif de référence en parole jusqu'à ce qu'elle soit aujourd'hui quasiment exclue des recherches sur des sujets sains à cause des risques dus au rayonnement ionisant.

La téléradiographie fournit des images d'une grande finesse, mais qui sont statiques. [24] l'a utilisée pour guider la détermination des fonctions d'aires des consonnes fricatives pour le même locuteur, en se basant sur les contours sagittaux extraits de ces images, sur les spectres de signal acoustique rayonné, sur la simulation de la fonction de transfert acoustique à partir de la fonction d'aire, et sur le calcul des fonctions de sensibilité acoustique.

La cinéradiographie offre l'avantage de pouvoir suivre le mouvement, avec une résolution temporelle de 50 images par seconde ou plus, avec une image de qualité moins bonne, mais cependant exploitable, comme on peut le voir à la FIGURE 3. [25] ont en outre couplé à ce dispositif un système d'enregistrement vidéo synchrone qui permet de capturer ainsi les informations articulatoires au niveau labial (cf. FIGURE 3). Les données acquises pour le locuteur PB ont permis le développement d'un modèle articulatoire ([26]) qui a servi de base à la synthèse articulatoire des fricatives ([15]). La FIGURE 3 illustre ces mesures, ainsi que les mouvements reproduits par le modèle articulatoire.

La même approche a été utilisée pour d'autres locuteurs pour étudier les compensations articulatoires ([27]), ou encore pour analyser les différentes stratégies de coordination articulatoire entre langue et mâchoire ([28]).

Notons enfin pour mémoire que [7] avaient également développé un système de ce type, produisant des images vidéos synchrones du voile du palais (par fibroscopie), des lèvres et d'une vue sagittale du conduit vocal (par radiographie).

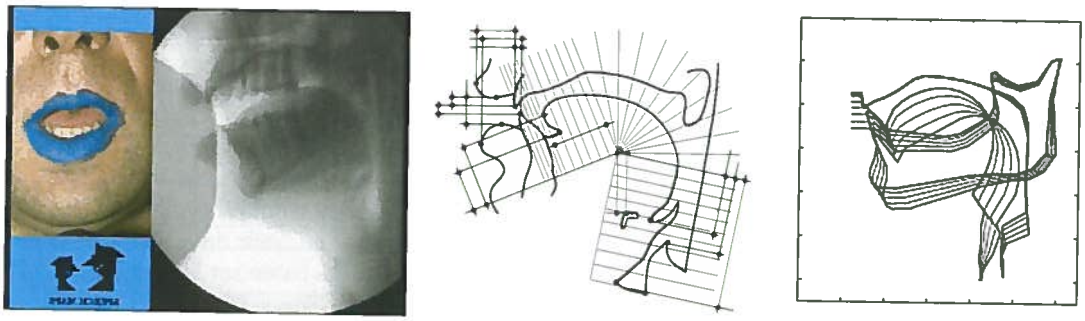

FIGURE 3. — *Exemple d'images cinématographique et vidéo de face synchrones (gauche) ; exemple de tracé manuel et des diverses mesures articulatoires réalisées automatiquement, en particulier grâce à l'utilisation d'une grille semi-polaire extensible dont les extrémités sont attachées à la pointe et à la racine de la langue (milieu) ; illustration du mouvement de la langue lié à un mouvement d'ouverture / fermeture de la mâchoire pour un modèle articulatoire médiosagittal.*

IMAGERIE PAR RESONANCE MAGNETIQUE ET TOMODENSITOMETRIE

À cause de sa nocivité attestée, la radiographie n'est aujourd'hui plus utilisée pour l'étude des sujets sains, au profit de l'Imagerie par Résonance Magnétique (IRM), qui semble être la technique la mieux adaptée à la mesure tridimensionnelle des articulateurs internes. L'IRM volumique peut fournir des séries d'images sagittales parallèles couvrant l'ensemble de la région des articulateurs. Ce type d'acquisition est relativement long, puisque le locuteur doit maintenir de manière artificielle chaque articulation pendant approximativement vingt à vingt-cinq secondes. Le protocole que nous utilisons actuellement produit 25 images sagittales acquises sur des volumes de 4 mm d'épaisseur sans recouvrement avec une résolution d'image d'environ de 1 pixel / mm ([29], [30], [31]) (FIGURE 4).

Notons pour mémoire le développement de plus en plus important de l'IRM dynamique, qui permet d'obtenir aujourd'hui des images à des taux d'échantillonnage d'images reconstruites qui peuvent aller jusqu'à une vingtaine de Hertz (voir par exemple [32] ou plus récemment [33]).

Dans la mesure où l'IRM ne permet pas la visualisation des structures osseuses, un scanner tomodensitométrique complet de la tête du sujet a été obtenu afin d'obtenir de bonnes images de référence de ces différentes structures : mâchoire et os hyoïde pour les structures typiquement mobiles ; palais dur, sinus maxillaires et sphénoïdal, pour les structures fixes de référence qui servent également à recaler dans un même repère commun les images acquises pour les différentes articulations. La FIGURE 4 montre l'image médio sagittale extraite de la pile d'images tomodensitométriques. Elle illustre également le comportement de la composante « Dos de Langue » pour le modèle articulatoire tridimensionnel de la langue.

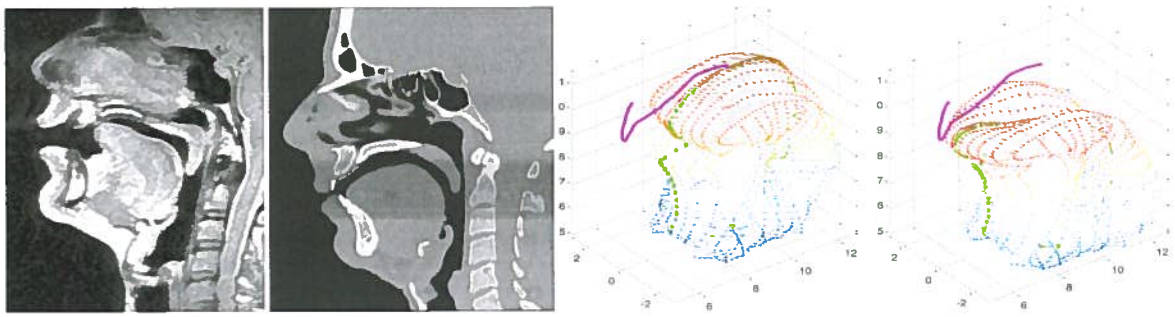

FIGURE 4. — *Exemple d'images médiosagittales obtenues par IRM et par TomoDensitométrie (images de gauche); illustration de la composante « Dos de Langue » pour le modèle articulatoire tridimensionnel de langue (valeurs du paramètre de contrôle TB -2 et +2)(droite).*

ARTICULOGRAPHIE ELECTROMAGNETIQUE

La résolution temporelle des méthodes d'imagerie (cinéradiographie, ou IRM dynamique) n'est pas suffisante pour suivre dans le détail la dynamique des articulateurs. L'articulographe électromagnétique (en anglais ElectroMagnetic Articulograph, EMA) constitue donc un dispositif complémentaire très intéressant qui permet de suivre les coordonnées d'une dizaine de petites bobines électromagnétiques réceptrices collées sur les organes du locuteur (pour une description détaillée du principe, voir par exemple [34]). Dans le cas de l'articulographe médiosagittal que nous utilisons, les bobines peuvent être fixées sur la mâchoire, la langue, le voile du palais, ou encore les lèvres (FIGURE 5). Les avantages de cette méthode sont d'une part sa bonne résolution temporelle (typiquement plusieurs centaines de Hertz), et d'autre part sa capacité à suivre des points de chair et non pas des contours. L'inconvénient majeur réside dans sa mauvaise résolution spatiale, qui ne permet qu'une vue très partielle des articulateurs ; nous verrons cependant dans la suite que cet inconvénient peut être dépassé si l'on fait appel à des modèles des organes que l'on désire suivre. Le caractère partiellement invasif, dû à la présence des bobines et des fils à l'intérieur de la bouche du locuteur constitue par contre un inconvénient qui n'est malheureusement pas négligeable et limite malgré tout la gamme des locuteurs que l'on peut enregistrer. L'articulographe peut être utilisé simultanément et synchronisé avec un dispositif d'enregistrement vidéo (FIGURE 6).

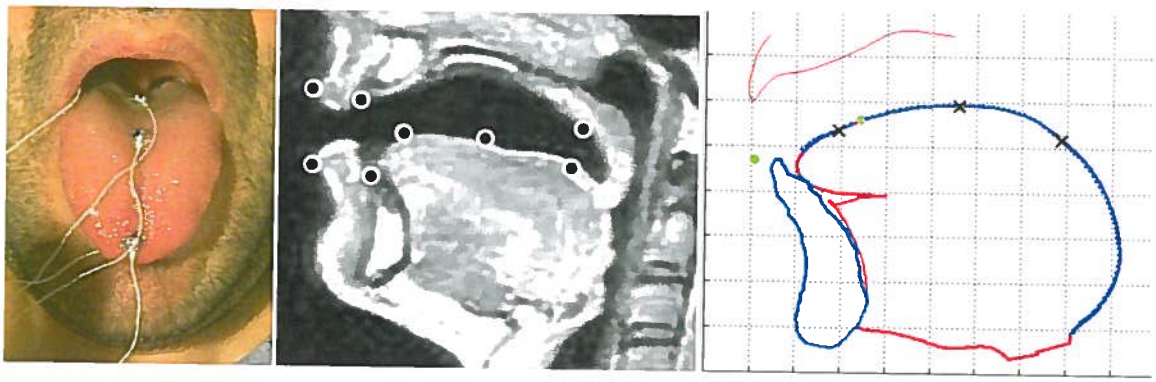

FIGURE 5. — *Photo de bobines réceptrices d'un articulographe électromagnétique fixées sur la langue d'un locuteur (gauche); illustration du positionnement de huit bobines dans le plan médiosagittal (milieu) ; exemple de reconstruction du contour médiosagittal de la langue à partir des coordonnées de la bobine de mâchoire et des trois bobines de langue (droite ; les points gris [verts] correspondent aux bobines, les croix noires à l'approximation par le modèle).*

FIGURE 6. — *Exemple d'une image issue de la vidéo enregistrée simultanément avec l'articulographe. On distingue les éléments du casque qui supporte les bobines électromagnétiques émettrices, les fils qui relient les bobines réceptrices au dispositif d'acquisition, un certain nombre de billes colorées collées sur le visage du sujet, et des traits tracés pour permettre de mieux suivre les mouvements.*

Le locuteur PB a fait l'objet d'au moins trois enregistrements EMA significatifs. [35] ont réalisé un enregistrement avec une bobine sur le velum, et deux sur l'avant de la langue, afin de caractériser le mouvement du voile du palais pour la nasalité (voir

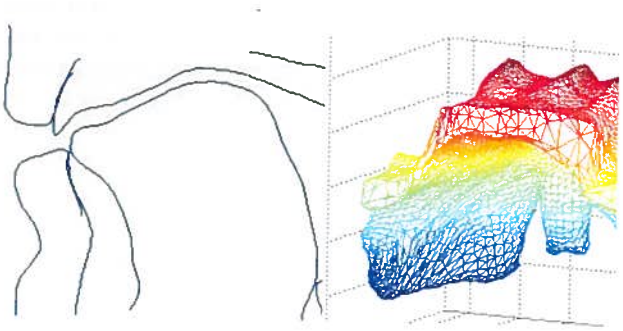

FIGURE 7). [36] ont enregistré le locuteur PB avec trois bobines sur la langue, mais sans bobine sur le velum, en synchronie avec deux caméras vidéo (voir FIGURE 6), pour l'étude des capacités de lecture linguale de sujets humains ; les données ont également été utilisées par [37] pour mettre au point une méthode de détermination de l'articulation à partir du son. Par ailleurs, [38] ont enregistré le locuteur PB sur un double corpus de parole et de mastication / déglutition dans le cadre de la recherche sur les origines du langage parlé.

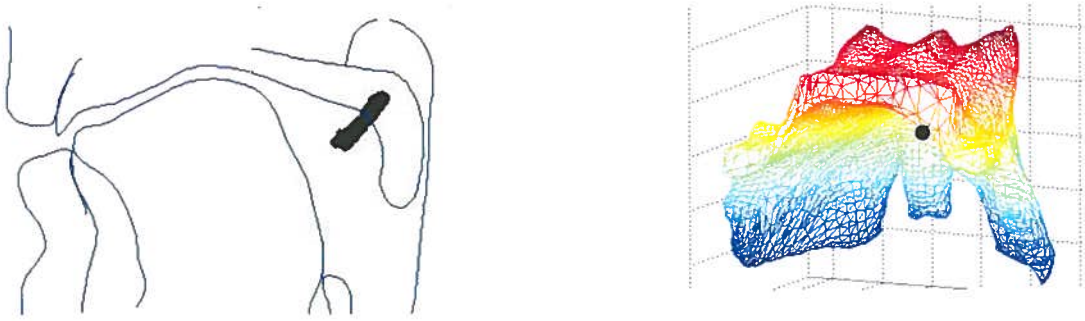

FIGURE 7. — *Espace couvert dans la parole par un point de chair du velum obtenu à l'aide d'une bobine EMA (gauche) ; point du maillage tridimensionnel du velum associé à ce point de chair (droite, gros cercle noir).*

VIDEO

Pour les organes externes, facilement visibles, l'utilisation de caméras vidéo permet un enregistrement non invasif. De longs corpus, avec un bon compromis entre résolution spatiale (de l'ordre de 2 pixels / mm) et résolution temporelle (typiquement 50 Hz), peuvent être enregistrés et stockés facilement au laboratoire.

[39] a développé le premier dispositif de la biométrie basé sur la vidéo à l'ICP. Ce système, basé sur une détection automatique des contours labiaux sur des images vidéos de face et de profil du visage du locuteur dont les lèvres ont été maquillées en bleu, a permis d'obtenir plusieurs ensembles de données labiales très importantes pour la modélisation articulatoire ([25], [26]), et pour l'inversion de paramètres articulatoires à partir du signal audio-visuel ([15]). Ce système fournit un nombre limité de paramètres labiaux (aperture, étirement, protrusion), mais n'est pas en mesure de proposer une reconstruction 3D complète et fidèle de la géométrie du pavillon labial.

Afin de pallier ce problème, nous avons développé des méthodes plus générales pour la mesure de la géométrie du visage et des lèvres à partir d'enregistrements vidéo. Environ 250 marqueurs colorés sont fixés sur le visage du locuteur ([40], [29], [41]). L'utilisation de caméras multiples, synchrones et calibrées, permet de déterminer les coordonnées de ces marqueurs avec une précision meilleure que le millimètre. La surface du visage de la tête parlante est représentée par un maillage d'environ 450 triangles dont les sommets s'appuient sur ces points, comme illustré à la FIGURE 8. Par ailleurs, les lèvres sont définies à l'aide d'un modèle 3D générique à 30 points de contrôle. Celui-ci, indépendant du locuteur, est déformé puis ajusté aux images par un expert, en utilisant les vues multiples pour capturer le contour externe des lèvres et modéliser leur surface visible. Ce dispositif peut également prendre en compte les yeux et en particulier les paupières pour la mesure et la modélisation du regard ([42]) ainsi que dans le cas de la recherche sur la parole expressive ([43]). Ces modèles spécifiques au locuteur peuvent être ensuite utilisés pour réaliser l'inversion acoustico-articulatoire. Par exemple, [44] ont utilisé une approche qui consiste à effectuer un suivi de type « analyse par la synthèse » à l'aide d'un modèle de forme et d'apparence du locuteur dont les paramètres de contrôle articulatoire sont optimisés de façon à reconstruire au mieux les images vidéo. La FIGURE 9 illustre le comportement de la composante « mouvement d'ouverture / fermeture de la mandibule » pour le modèle articulatoire tridimensionnel de lèvres et de visage.

Notons enfin qu'une éclisse mandibulaire (FIGURE 10) peut également être utilisée ([29]) : c'est un dispositif mécanique qui se fixe de manière solidaire sur l'arc dentaire de la mâchoire, et permet de reporter de manière visible à l'extérieur du conduit vocal les mouvements de cette dernière et donc d'en déterminer la position exacte.

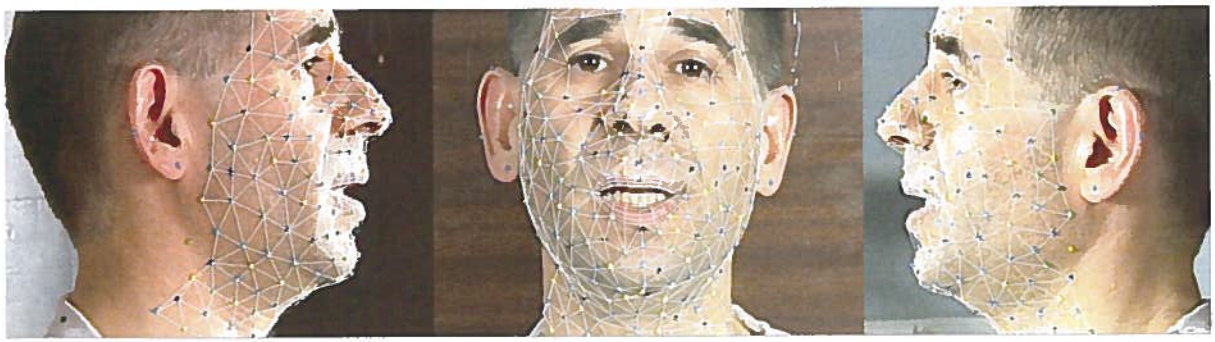

FIGURE 8. — *Exemple de vues de la tête du locuteur produites par trois caméras synchrones et calibrées, avec superposition des maillages de visage et de lèvres.*

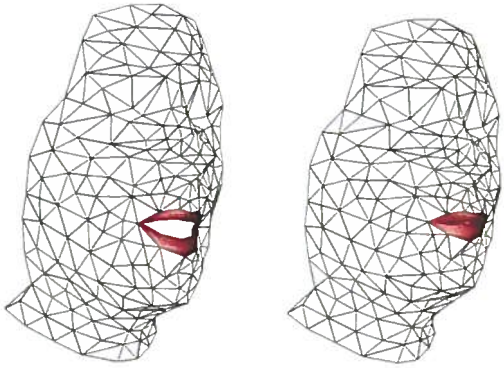

FIGURE 9. — *Illustration de la composante « mouvement d'ouverture / fermeture de la mandibule » pour le modèle articulatoire tridimensionnel de lèvres et de visage (valeurs du paramètre de contrôle JH -2 et +2).*

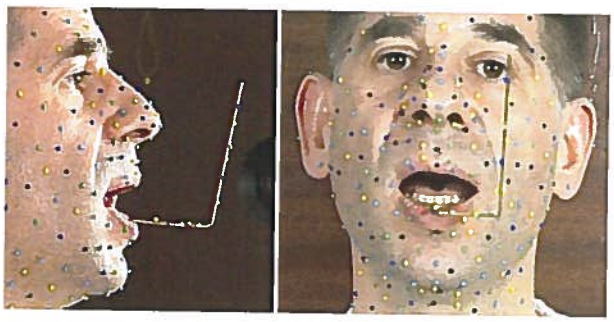

FIGURE 10. — *Exemple de vues du locuteur équipé avec une éclisse mandibulaire.*

COMPARAISON DES DIFFERENTES METHODES

Le Tableau 1 résume les avantages et inconvénients des différentes méthodes qui ont été décrites ci-dessus. Il a été complété pour mémoire par quelques informations sur l'ElectroPalatographie (EPG), l'imagerie par échographie ultrasonique, et la vidéo rapide pour les cordes vocales.

DIFFERENTES REALISATIONS

Dans cette section, nous décrivons les résultats majeurs obtenus grâce à notre approche mono-locuteur multi-dispositifs.

Synthèse articulatoire des fricatives par copie

Le synthétiseur articulatoire que nous avons développé (voir [15] pour plus de détails) est constitué d'un certain nombre de modèles interconnectés, tous basés de manière plus ou moins directe sur des données acquises sur le locuteur PB.

Le modèle articulatoire 2D est un modèle médiosagittal basé sur un corpus cinéradiographique acquis sur le même locuteur PB ([26]) ; le modèle de passage qui permet de déterminer la fonction d'aire à partir de la coupe sagittale a été adapté par optimisation aux données articulatoires et acoustiques du locuteur ([45], [26]). Le modèle aérodynamique est un modèle simplifié de chute de pression qui prend en compte les deux constrictions localisées dans le conduit vocal, la glotte et la constriction orale ([23]). Le modèle de source de bruit de friction est un modèle fonctionnel basé sur des données issues du locuteur PB ([45], [21]). La source de voisement est un modèle à deux masses de cordes vocales dont les paramètres ont été ajustés pour représenter au mieux le comportement phonatoire du locuteur PB en fonction des paramètres aérodynamiques ([22]). Finalement, le modèle de propagation et de rayonnement acoustiques est une ligne analogue à réflexion du conduit vocal ([46]).

Ce synthétiseur articulatoire, adapté à tous les niveaux possibles aux données mesurées sur le locuteur PB, est globalement contrôlé par deux jeux de paramètres articulatoires : les paramètres supra-laryngés qui pilotent le modèle articulatoire, et les paramètres qui contrôlent les cordes vocales (pression sous-glottique, longueur des cordes vocales et hauteur de la glotte au repos), qui doivent être soigneusement coordonnés pour générer de la parole de haute qualité. Nous avons montré ([15]) qu'il est

possible d'imiter de manière précise, à l'aide de ce synthétiseur articulatoire, des séquences Voyelle-Fricative-Voyelle produites par le locuteur PB. Notre approche, basée sur la synthèse articulatoire par copie, consistait à construire des trajectoires articulatoires aussi proches que possible de celles du locuteur. Un test d'évaluation perceptive a a montré l'excellente qualité des résultats de resynthèse, avec des scores d'identification des stimuli synthétiques de 98.6 %, très proches de ceux des signaux originaux. Ce travail de synthèse articulatoire par copie a finalement permis de valider le synthétiseur, l'intégration de tous ses modules, et la possibilité de déterminer des commandes qui produisent une synthèse articulatoire de parole d'excellente qualité. Il valide également notre choix d'une approche anthropomorphique basée sur les données et modèles cohérents issus d'un même locuteur.

Clones orofaciaux 3D et leur contrôle

Les données nous ont permis de développer des modèles tridimensionnels de surface des principaux articulateurs de la parole pour le locuteur PB : mâchoire, lèvres et visage à partir de la vidéo pour la parole neutre ([29]), et plus récemment, incluant des expressions ([41], [43]) ; mâchoire et langue ([29], [30]), ou encore velum, à partir des données issues de l'IRM ([31]). Ces différents modèles sont intégrés dans un clone orofacial du locuteur, qui peut prendre divers aspects, en fonction des besoins, comme illustré à la FIGURE *11*.

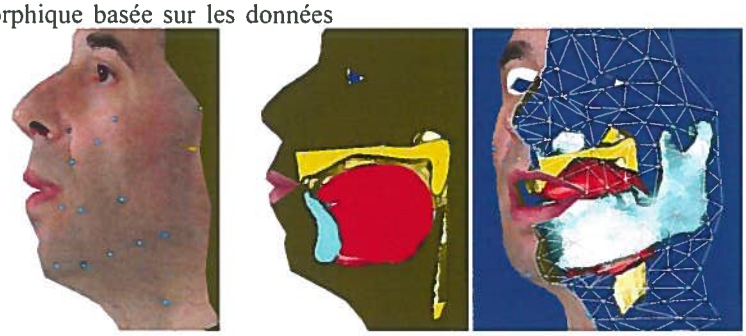

FIGURE 11. — *exemples de différents modes de présentation du clone orofacial du locuteur PB (gauche : tête complète vue de profil avec texture de peau ; milieu : écorché de profil ; droite : vue de trois quarts à moitié texturée).*

Ces clones peuvent être ensuite contrôlés par trois types de méthodes : (1) *capture de mouvement*, (voir 8) ; (2) *inversion à partir du signal audiovisuel* enregistré par le sujet; (3) *synthèse à partir du texte*, qui demande une mise en œuvre complexe (voir par exemple [47], pour un autre locuteur).

Nous avons indiqué plus haut que l'articulographe électromagnétique médiosagittal 2D permet de suivre les coordonnées médiosagittales de petites bobines fixées en différents points de chair sur les articulateurs de la parole et donc d'en capter le mouvement : si le nombre de ces points est suffisant, il est alors possible de déterminer par optimisation les paramètres de commande des modèles articulatoires de manière à faire coïncider les points correspondants des modèles avec les points mesurés. Ainsi, [48] montrent (1) qu'une bobine collée sur les incisives inférieures permet de déterminer l'ouverture de mâchoire, (2) que deux bobines collées sur les lèvres permettent de contrôler le modèle de lèvres et de visage – du moins pour de la parole neutre –, et (3) que les trois bobines collées sur la langue, en complément de celle de la mâchoire, permettent de déterminer les six paramètres de contrôle du modèle de langue. Pour compléter, [31] montrent qu'une bobine collée sur le voile du palais permet de contrôler les deux paramètres du modèle correspondant. Cette approche par capture de mouvement permet de piloter les divers articulateurs du clone orofacial de manière très naturelle. Les stimuli ainsi générés ont été utilisés dans un test de *lecture linguale* ([48]).

CONCLUSIONS ET PERSPECTIVES

Nous avons présenté succinctement un ensemble important de travaux d'acquisition de données acoustiques, aérodynamiques et articulatoires employant des dispositifs divers associés entre eux pour un même locuteur prononçant des corpus similaires. Les données recueillies ont permis de développer des modèles de production de la parole cohérents entre eux, puisque basés sur le même locuteur et les mêmes corpus : cette approche présente l'avantage crucial de pouvoir tirer parti au mieux des caractéristiques des différents dispositifs. Nous avons ainsi développé un clone orofacial possédant une bonne résolution géométrique que nous pouvons ensuite contrôler à partir de données possédant une haute résolution temporelle. Parmi les nombreux résultats, nous pouvons citer la synthèse articulatoire des consonnes fricatives, les modèles tridimensionnels d'articulateurs de la parole. Notre ambition est maintenant d'étendre cette approche à d'autres locuteurs afin de déterminer les caractéristiques qui sont universelles et celles qui relèvent de stratégies idiosyncratiques. Nous nous attaquerons ainsi dans ce cadre aux problèmes de normalisation et d'adaptation entre locuteurs (cf. [49]). Nous pourrons nous référer à un travail préliminaire au cours duquel nous avons comparé les stratégies de synergie mâchoire / langue pour trois locuteurs, dont le locuteur PB ([28]).

REMERCIEMENTS

Nous tenons à remercier ici toute une série de personnes qui ont été associées à ces travaux à diverses périodes, et à divers titres, et sans lesquels ce travail de longue haleine serait impossible. Merci à nos doctorants : Atef Ben Youssef, Oxana Govokhina, Tahar Lallouache, Khaled Mawass, Mathias Odisio, Yen Pham Thi Ngoc, Antoine Serrurier, Yuliya Tarabalka, Christophe Vescovi. Merci à nos anciens doctorants devenus collègues : Denis Beautemps, Eric Castelli, Lionel Revéret, Solange Rossato, Anne Vilain. Merci à nos collègues Amar Djéradi, Bernard Guérin, Kunitoshi Motoki, Pascal Perrier. Merci à Monica Baciu, Gilbert Brock et Christoph Segebarth qui nous ont donné accès à la cinéradiographie et à l'IRM.

Et enfin un grand merci à Bernard Teston qui, même s'il n'est jamais intervenu dans nos travaux de manière explicite, a toujours été pour nous un modèle, un inspirateur et un conseiller précieux.

BIBLIOGRAPHIE

[1] TESTON B and LIPCEY A (1974) — Etude et réalisation d'une unité de traitement des signaux électromyographiques Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA) 2 55-85

[2] AUTESSERRE D and TESTON B (1979) — Current issues of the Phonetic Sciences, ed H Hollien and P Hollien (Amsterdam: John Benjamins) pp 407-22

[3] TESTON B and AUTESSERRE D (1975) — Réalisation d'une unité d'analyse polyphonométrique; sa contribution à l'étude de la nasalisation vocalique et de la nasalité consonantique en français parlé à Marseille Cahiers de Linguistique d'Orientalisme et de Slavistique 5-6 415-37

[4]] TESTON B and GALINDO B (1995) — A diagnostic and rehabilitation aid workstation for speech and voice pathologies. In: 4th EuroSpeech Conference, ed J M Pardo, et al. (Madrid, Spain: Gráficas Brens) pp 1883-6

[5] TESTON B (2007) — Les Dysarthries, ed P Auzou, et al. (Marseille: SOLAL) pp 115-7

[6] AUTESSERRE D and TESTON B (1979) — Etude électropalato-graphique et aérodynamique simultanée des occlusives du français Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA) 6 15-22

[7] TESTON B and AUTESSERRE D (1986) — Description d'un dispositif d'enregistrement simultané des mouvements des organes articulatoires. In: XV^{èmes} Journées d'Etude sur la Parole (JEP), (Aix-en-Provence, France: Groupement des acousticiens de langue française (GALF)) pp 65-8

[8] YU P, GARREL R, NICOLLAS R, OUAKNINE M and GIOVANNI A (2007) — Objective voice analysis in dysphonic patients: new data including nonlinear measurements Folia Phoniatrica et Logopaedica 59 20-30

[9] GHIO A, GIOVANNI A, TESTON B, REVIS J, YU P, OUAKNINE M, ROBERT D and LEGOU T (2008) — Bilan et perspectives de quinze ans d'évaluation vocale par méthodes instrumentales et perceptives. In: Journées d'Etude sur la Parole (Avignon, France: LIA) pp 309-12

[10] BADIN P and FANT G (1984) — Notes on vocal tract computation Speech Transmission Laboratory - Quarterly Progress Status Report - Stockholm 2-3/1984 53-108

[11] CASTELLI E and BADIN P (1989) — Nasopharyngeal tract transfer functions measurements with white noise excitation. In: 13th International Conference on Acoustics, pp 511-4

[12] CASTELLI E and BADIN P (1988) — Vocal tract transfer functions measurements with white noise excitation. Application to the naso-pharyngeal tract. In: 7th FASE Symposium, (Edinburgh, UK pp 415-22

[13] DJÉRADI A, GUÉRIN B, BADIN P and PERRIER P (1991) — Measurement of the acoustic transfer function of the vocal tract: a fast and accurate method Journal of Phonetics 19 387-95

[14] PHAM THI NGOC Y and BADIN P (1994) — Vocal tract acoustic transfer function measurements: further developments and applications. In: Journal de Physique IV, Colloque C5, Supplément au Journal de Physique III. 3rd French Congress of Acoustics, (Toulouse, France pp 549-52

[15] MAWASS K, BADIN P and BAILLY G (2000) — Synthesis of French fricatives by audio-video to articulatory inversion Acta Acustica 86 136-46

[16] BADIN P, MOTOKI K, MIKI N, RITTERHAUS D and LALLOUACHE T M (1994) — Some geometric and acoustic properties of the lip horn Journal of the Acoustical Society of Japan (English) 15 243-53

[17] BADIN P (1989) — Acoustics of voiceless fricatives: production theory and data Speech Transmission Laboratory - Quarterly Progress Status Report - Stockholm 3/1989 33-55

[18] BADIN P and FANT G (1989) — Fricative production modelling: aerodynamic and acoustic data. In: 1st EuroSpeech Conference, (Paris, France pp 23-6

[19] BADIN P, HERTEGÅRD S and KARLSSON I (1990) — Notes on the Rothenberg mask Speech Transmission Laboratory - Quarterly Progress Status Report - Stockholm 1/1990 1-7

[20] BADIN P, SHADLE C H, PHAM THI NGOC Y, CARTER J N, CHIU W, SCULLY C and STROMBERG K (1994) — Frication and aspiration noise sources: contribution of experimental data to articulatory synthesis. In: 3rd International Conference on Spoken Language Processing, (Yokohama, Japan pp 163-6

[21] MAWASS K, BADIN P, VESCOVI C and BEAUTEMPS D (1996) — Evaluation d'un modèle de source de friction pour la synthèse articulatoire des consonnes fricatives. In: XXI^{ès} Journées d'Etude sur la Parole (JEP), (Avignon, France pp 367-70

[22] VESCOVI C, CASTELLI E and PELORSON X (1995) — Adaptation of a two-mass model of the vocal cords to a particular speaker. In: 4th EuroSpeech Conference, ed J M Pardo, et al. (Madrid, Spain: Gráficas Brens) pp 1933-6

[23] ABRY C, BADIN P, MAWASS K AND PELORSON X (1998) — The Equilibrium Point Hypothesis and control space for relaxation movements or "When is movement actually needed to control movement ?", Commentary on target paper: P. Perrier, D.J. Ostry & R. Laboissière (1996), The Equilibrium Point Hypothesis and its application to speech motor control (JSHR, 39, 365-378) Les Cahiers de l'ICP, Bulletin de la Communication Parlée 4 27-33

[24] BADIN P (1991) — Fricative consonants: acoustic and X-ray measurements Journal of Phonetics 19 397-408

[25] BADIN P, GABIOUD B, BEAUTEMPS D, LALLOUACHE T M, BAILLY G, MAEDA S, ZERLING J P and BROCK G (1995) — Cineradiography of VCV sequences: articulatory-acoustic

data for a speech production model. In: 15th International Conference on Acoustics, (Trondheim, Norway pp 349-52

[26] Beautemps D, Badin P and Bailly G (2001) — Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling Journal of the Acoustical Society of America 109 2165-80

[27] Maeda S (1990) — Speech Production and Modelling, ed W J Hardcastle and A Marchal (Kluwer: Academic Publishers) pp 131-49

[28] Bailly G, Badin P and Vilain A (1998) — Synergy between jaw and lips/tongue movements: Consequences in articulatory modelling. In: 5th International Conference on Spoken Language Processing, ed R H Mannell and J Robert-Ribes (Sydney, Australia pp 1859-62

[29] Badin P, Bailly G, Revéret L, Baciu M, Segebarth C and Savariaux C (2002) — Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images Journal of Phonetics 30 533-53

[30] Badin P and Serrurier A (2006) — Three-dimensional linear modeling of tongue: Articulatory data and models. In: 7th International Seminar on Speech Production, ISSP7, ed H C Yehia, et al. (Ubatuba, SP, Brazil: UFMG, Belo Horizonte, Brazil) pp 395-402

[31] Serrurier A and Badin P (2008) — A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data Journal of the Acoustical Society of America 123 2335-55

[32] Demolin D, Metens T and Soquet A (2000) — Real time MRI and articulatory coordinations in vowels. In: 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, (Kloster Seeon, Germany pp 93-6

[33] Lee S, Bresh E, Adams J, Kazemzadeh A and Narayanan S S (2006) — A study of emotional speech articulation using a fast magnetic resonance imaging technique. In: Interspeech 2006 - ICSLP, (Pittsburgh, Pennsylvania, USA pp 2234-7

[34] Perkell J S, Cohen M M, Svirsky M A, Matthies M L, Garabieta I and Jackson M T T (1992) — Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements Journal of the Acoustical Society of America 92 3078-96

[35] Rossato S, Badin P and Bouaouni F (2003) — Velar movements in French: an articulatory and acoustical analysis of coarticulation. In: 15th International Congress of Phonetic Sciences, ed M-J Solé, et al. (Barcelona, Spain pp 3141-4

[36] Tarabalka Y, Badin P, Elisei F and Bailly G (2007) — Can you "read tongue movements"? Evaluation of the contribution of tongue display to speech understanding. In: 1ère Conférence internationale sur l'accessibilité et les systèmes de suppléance aux personnes en situation de handicaps (ASSISTH'2007), ed N Vigouroux, et al. (Toulouse, France: Editions Cépaduès, Toulouse) pp 187-93

[37] Ben Youssef A, Badin P, Bailly G and Heracleous P (2009) — Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models. In: Interspeech 2009, (Brighton, UK pp 2255-8

[38] Serrurier A, Barney A, Badin P, Boë L-J and Savariaux C (2008) — Comparative articulatory modelling of the tongue in speech and feeding. In: 8th International Seminar on Speech Production, ISSP8, ed R Sock, et al. (Strasbourg, France pp 325-8

[39] Lallouache M T (1990) — Un poste 'Visage-Parole'. Acquisition et traitement de contours labiaux. In: 18èmes Journées d'Etude sur la Parole, (Montreal, Quebec, Canada pp 282-6

[40] Odisio M, Elisei F, Bailly G and Badin P (2001) — Clones parlants 3D vidéo-réalistes: application à l'analyse de messages audio-visuels. In: 7èmes Journées d'Études et d'Échange "Compression et représentation des signaux audivisels" (CORESA'2001), (Dijon, France pp 141-4

[41] Bailly G, Elisei F, Badin P and Savariaux C (2006) — Degrees of freedom of facial movements in face-to-face conversational speech. In: International Workshop on Multimodal Corpora, (Genoa, Italy pp 33-6

[42] Elisei F, Bailly G and Casari A (2007) — Towards eyegaze-aware analysis and synthesis of audiovisual speech. In: AVSP 2007, AudioVisual Speech Processing Conference, (Hilvarenbeek, The Netherlands pp 120-5

[43] Bailly G, Bégault A, Elisei F and Badin P (2008) — Speaking with smile or disgust: data and models. In: Auditory-Visual Speech Processing Workshop, AVSP 2008, (Moreton Island, Australia pp 111-4

[44] Odisio M, Bailly G and Elisei F (2004) —Tracking talking faces with shape and appearance models Speech Communication (Special Issue on Audio Visual speech processing) 44 (1-4) 63-82

[45] Beautemps D, Badin P and Laboissière R (1995) — Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data Speech Communication 16 27-47

[46] Kelly J L and Lochbaum C C (1962) — Speech synthesis. In: 4th International Conference on Acoustics, p G42

[47] Govokhina O, Bailly G and Breton G (2007) — Learning optimal audiovisual phasing for a HMM-based control model for facial animation. In: 6th ISCA Workshop on Speech Synthesis, (Bonn, Germany

[48] Badin P, Tarabalka Y, Elisei F and Bailly G (2010) — Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding Speech Communication 52 493-503

[49] Ananthakrishnan G, Badin P, Valdés Vargas J A and Engwall O (2010) — Predicting unseen articulations from multi-speaker articulatory models. In: Interspeech 2010 (11th Annual Conference of the International Speech Communication Association), ed T Kobayashi, et al. (Makuhari, Japan pp 1588-91

2012 • Tome 30 • N°1-2

ISSN 1279 • 7863

BIOMÉTRIE HUMAINE ET ANTHROPOLOGIE

Revue de la Société Internationale de Biométrie Humaine fondée en 1932

TOME 30 janvier - juin 2012 N° 1 - 2

ISSN 1279-7863

Biométrie humaine et Anthropologie

SOMMAIRE / *CONTENTS*

Internet http : //www.biometrie-humaine.org
E-mail : sbh@evolhum.cnrs.fr

Dupli Imprimerie 03 23 53 02 41