Mise en oeuvre d'un synthétiseur 3D de Langage Parlé Complété

Guillaume Gibert[1]*, Gérard Bailly*[1]*, Frédéric Eliséi*[1]*, Denis Beautemps*[1]*, Rémi Brun*[2]

(1) Institut de la Communication Parlée UMR CNRS 5009, INPG/U3,
46, av. Félix Viallet - 38031 Grenoble France
Tél. : +33 (0)4 76 57 45 34 - Fax : +33 (0)4 76 57 47 10
Mél : gibert@icp.inpg.fr
(2) Attitude Studio SA, 100 Avenue du Général Leclerc, 93692 Pantin France

ABSTRACT

We present here our efforts for characterizing the 3D movements of the right hand and the face of a French female during the production of manual cued speech. We analyzed the 3D trajectories of 50 hand and 63 facial fleshpoints during the production of 238 utterances carefully designed for covering all possible diphones of the French language. Linear and non linear statistical models of the hand and face deformations and postures have been developed using separate and joint corpora. We implement a concatenative audiovisual text-to-cued speech synthesis system.

1. INTRODUCTION

Si les mouvements de la mâchoire, des lèvres et des joues sont immédiatement visibles, les mouvements des organes sous-jacents tels que le larynx, le velum ou la langue ne le sont pas : les mouvements de la langue sont faiblement corrélés avec les mouvements visibles du visage ($R \sim 0.7$) [17, 10] et cette corrélation est insuffisante pour retrouver les signes phonétiques importants comme le lieu d'articulation linguale par exemple [7, 2]. La lecture labiale seule est insuffisante dû à un manque d'information sur le point d'articulation de la langue, des modes d'articulation (nasalité, voisement) et à la similarité de certaines formes de lèvres pour certains phonèmes (aussi appelés sosies labiaux tels que [u] vs. [y]). Dans tous les cas, même le meilleur décodeur ne peut pas identifier plus de 50% de phonèmes dans des syllabes sans sens [14] ou dans des mots ou des phrases [4]. Le Langage Parlé Complété a été construit pour compléter la lecture labiale. Développé par Cornett [5] et adapté à plus de 50 langues [6], ce système est basé sur l'association articulation faciale/clés (formées par la main). En même temps qu'il parle, le locuteur utilise sa main pour indiquer une position sur le visage (déterminant un sous-ensemble de voyelles) et une forme de main (déterminant un sous-ensemble de consonnes cf. figure 1) (voir `http://retore.chez.tiscali.fr/LPC`). De nombreuses études ont montré l'accroissement de l'intelligibilité par ce codage comparé à la lecture labiale seule [13, 16] et l'apport en terme de facilité d'apprentissage de la langue[11]. De nombreux travaux sont consacrés à l'étude de la perception du L.P.C. mais peu de travaux s'attachent à la production. Nous décrivons ici une série d'expériences pour rassembler des données et caractériser les mouvements de la main et du visage de la codeuse L.P.C. en vue d'implémenter un synthétiseur de L.P.C. à partir du texte.

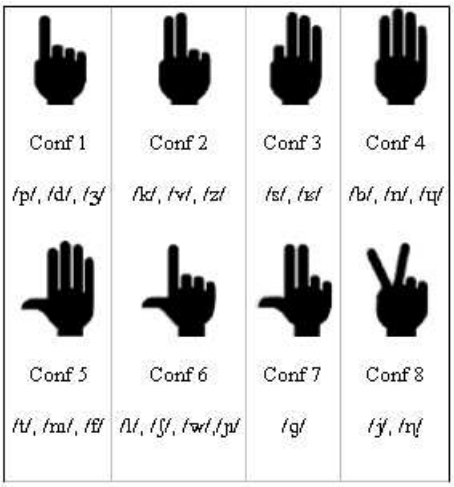


FIG. 1: Système de codage L.P.C. pour les consonnes.

2. DONNÉES ISSUES DE CAPTURE DE MOUVEMENT

Nous avons enregistré les positions 3D de 113 marqueurs collés sur la main et le visage du sujet grâce à un système Vicon© de capture de mouvements utilisant 12 caméras. Ce système délivre les positions 3D des marqueurs (cf. figure 4(a)) à 120 images/s. Deux formats de montage des caméras nous ont permis d'enregistrer 3 corpora distincts :

- un corpus de transitions de formes de main dans un espace libre : la codeuse produit toutes les transitions possibles entre les 8 différentes formes de main.
- un corpus de visèmes sans forme de main associée. Ce corpus est similaire à celui habituellement utilisé pour construire les clones à l'I.C.P. [1].
- un corpus de 238 phrases prononcées et codées.

Les corpora 1 et 2 ont été utilisés pour construire les modèles statistiques des mouvements de la main et du visage séparément. Ces modèles sont ensuite utilisés pour reconstruire les données manquantes dans le corpus 3 : quand le sujet code le L.P.C. le visage cache des parties de la main et vice versa.

3. MODÈLES ARTICULATOIRES DU VISAGE ET DE LA MAIN

La motivation scientifique pour construire des modèles statistiques à partir de données issues de capture de mouvements brutes concerne l'étude du L.P.C. : si les positions des marqueurs sont ainsi toujours accessibles et fiables, la cinématique des articulations, du bout des doigts et des

constrictions doigts/visage offrent un moyen parfait pour étudier la production du L.P.C. et les lois de coordinations entre l'acoustique, les mouvements du visage et de la main.

3.1. Le visage

La méthodologie utilisée à l'I.C.P. pour construire des clones articulatoires consiste en une série d'analyses en composantes principales de différents sous-ensembles de points de peau [1, 15] : la contribution de la rotation de la mâchoire, du geste d'arrondissement des lèvres, du mouvement vertical propre de la lèvre supérieure et inférieure, celui des coins de lèvres et le mouvement de la gorge sont soustraits itérativement aux données originales issues de la capture de mouvements. Cette méthodologie est normalement appliquée à des têtes quasi-statiques. Or le mouvement de la tête est libre dans les corpora 2 et 3, donc nous devons résoudre le problème de la répartition de la variance des positions des 18 marqueurs placés sur la gorge entre les mouvements de tête et ceux du visage. Ce problème est résolu en 3 étapes :

- Estimation d'un mouvement de tête utilisant l'hypothèse d'un mouvement rigide des marqueurs placés sur les oreilles, le nez et le front. Une analyse en composantes principales sur les 6 paramètres de roto-translation extraits du corpus 3 est calculée et les nmF premières composantes sont retenues comme paramètres de contrôle de la tête.
- Le clonage des mouvements articulatoires du visage est effectué en inversant le mouvement rigide sur toutes les données. naF composantes sont retenues comme paramètres de contrôle des mouvements articulatoires du visage.
- Les mouvements de la gorge sont considérés comme égaux aux mouvements de tête pondérés par des facteurs inférieurs à 1. Une optimisation des poids et des déformations du visage est ensuite calculée en gardant la même valeur pour les prédicteurs nmF et naF.

Toutes ces opérations sont faites sur les mouvements du visage des corpus 2 et 3 où tous les marqueurs sont visibles. Une simple quantification vectorielle nous assurant d'un minimum de distance 3D entre les trames sélectionnées (égal ici à 2mm) est mis en oeuvre avant la modélisation.

3.2. La main

Construire un modèle statistique des déformations de la main est plus complexe. Si on considère l'avant-bras comme étant le support de la main (les 50 marqueurs suivent un mouvement rigide qui peut etre considéré comme le mouvement de l'avant-bras), les mouvements du poignet, de la paume et des phalanges ont une influence non-linéaire certaine sur les positions 3D des marqueurs. Ces positions reflètent faiblement les rotations des articulations sous-jacentes : la déformation de la peau engendrée par les tissus musculaires et la peau produit d'importantes variations de distances entre les marqueurs collés sur une même phalange (ex. : variation de 3mm sur une distance de 1.6cm, pour les points situés sur la deuxième phalange du majeur). Le modèle de déformation de la main est construit en 4 étapes :

- Estimation des mouvements de la main en utilisant l'hypothèse d'un mouvement rigide des marqueurs placés sur l'avant-bras. Une analyse en composantes principales est ensuite calculée sur les 6 paramètres de mouvement de la main et on conserve les nmH premières composantes comme paramètres de contrôle des mouvements de la main.
- Tous les angles entre les différents segments composant la main et l'avant-bras ainsi qu'entre les phalanges successives sont calculés (rotation, écartement, torsion) soit 23 angles.
- Une analyse en composantes principales est ensuite calculée sur tous ces angles et les naH premières composantes sont retenues comme paramètres de contrôle de la forme de la main.
- Nous calculons ensuite les sinus et cosinus de toutes ces valeurs prédites et faisons une régression linéaire entre les 2*naH+1 valeurs et les coordonnées 3D des marqueurs collés sur la main.

L'étape 4 fait l'hypothèse que le déplacement induit par une rotation pure au niveau d'une articulation produit un mouvement elliptique de la surface de la peau.

3.3. Résultats de la modélisation

Dans le corpus 1, les données d'apprentissage pour les formes de main comportent 8446 trames. Dans le corpus 2 et 3, les données d'apprentissage pour les déformations faciales comportent 4938 trames. Nous avons retenu $naH = 12$ paramètres de contrôle pour la forme de la main et $naF = 7$ paramètres de contrôle articulatoire du visage. Quant aux mouvements de roto-translation, nous avons gardé $nmF = 5$ et $nmH = 5$ paramètres de contrôle de mouvement de la tête et de la main. L'erreur absolue de modélisation pour la position d'un marqueur visible est de 2mm pour la main et 1mm pour le visage.

4. ANALYSES COMPLÉMENTAIRES DES DONNÉES

Des analyses complémentaires ont été effectuées afin de vérifier si la codeuse avait effectivement réalisé les bonnes transitions de forme et de position de main en fonction de la chaine phonétique de chaque phrase. Par la suite, l'ensemble des données sont considérées. Les mouvements et déformations de la main et du visage sont régularisés et reconstruits en utilisant les modèles décrits plus haut. Globalement, le codage LPC consiste en un modèle de constriction : avec une certaine forme de main une constriction est effectuée (la plupart du temps un contact) soit une occlusion entre la main et le visage. La place de la constriction détermine la voyelle (ou plutôt un sous-ensemble de voyelles) et la forme détermine un sous-ensemble de consonnes.

4.1. Reconnaissance de la forme de la main et des consonnes

Nous avons segmenté manuellement les 238 phrases aux instants de constriction maximale en utilisant notre système d'animation MOTHER OPENGL© [15] et étiqueté la valeur appropriée de la clé, c'est-à-dire un chiffre entre 0 et 8 : 0 correspondant à la position de repos choisie par la codeuse (poing fermé à l'écart du visage). 4114 formes de main ont été identifiées et segmentées. Les 7 paramètres caractéristiques suivant ont été déterminés pour chaque instant cible :

- Pour chaque doigt (hormis le pouce), la distance entre le marqueur près de la paume et celui du bout du doigt est

calculée : une valeur maximale correspond à une extention du doigt alors qu'une valeur minimale correspond à une rétraction.
– La distance entre les marqueurs placés sur les bouts des doigts index et majeur est déterminée pour éviter toute confusion entre les formes 2 et 8.
– La distance entre le bout du pouce et la paume est déterminée pour différencier les formes 1 et 6, 2 et 7.

Ces 7 paramètres associés aux formes de main correspondantes permettent d'estimer des modèles Gaussiens pour chaque forme de main. La probabilité a posteriori de chaque nouvelle trame d'appartenir à une des 8 formes de main peut etre calculée. Un exemple de ces probabilités au cours du temps sur la première phrase du corpus est représentée sur la figure 2 avec le signal acoustique. Le taux de reconnaissance est assez élevé (98.78%). Les erreurs sont en général dues à des problèmes de réductions consonantiques voire à des omissions (notamment des "glides" dans des séquences complexes CCCV).

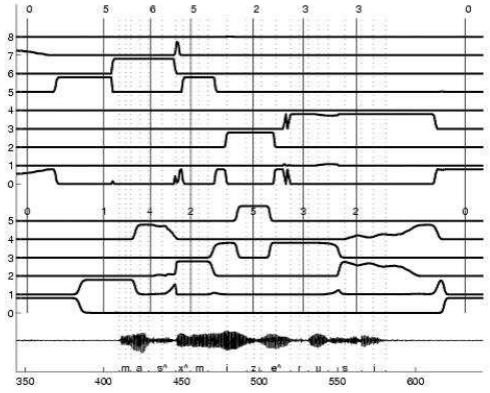


FIG. 2: Variation des probabilités issues des modèles gaussiens pour la forme (haut) et la position (bas) de la main pour la phrase "ma chemise est roussie".

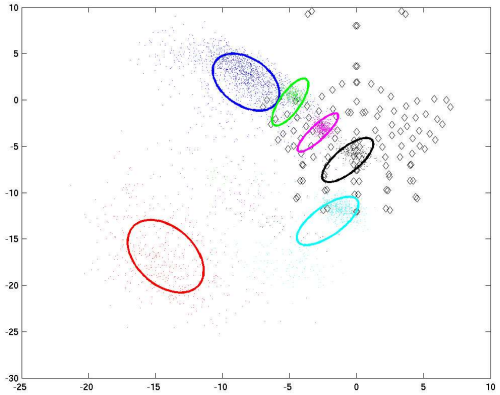


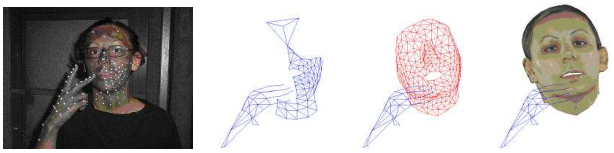
FIG. 3: Données et ellipses de dispertion de la position du bout du doigt le plus long pour chaque atteinte de cible.

4.2. Reconnaissance de la position de la main et des voyelles

Nous avons ajouté à l'étiquetage précédent 6 valeurs pour la position de la main : la position 0 correspondant à la position de repos. Nous avons caractérisé la position de la main pour chaque cible dans un référentiel 3D rattaché à la tête : la position 3D du doigt le plus long (l'index pour les configurations 1 et 6 et le majeur dans les autres cas) (cf. figure 3) a été enregistrée et des modèles Gaussiens ont été estimés comme précédemment. Sur les 4114 trames, 96.76% ont été identifiées pour un total de 133 erreurs de reconnaissance. Il y a trois sources d'erreurs essentielles :
– la plus importante source d'erreur vient de la position 1 (coté). Cette position est aussi utilisée pour coder des consonnes précédées d'une consonne et pour des schwas : la codeuse pointe la position coté mais ne l'atteint pas.
– La position de repos 0 a une grande variance et les positions 1 et 4 réalisées trop loin du visage sont parfois capturées par le modèle gaussien de la position 0.
– des confusions de codage des voyelles intermédiaires (/e/ vs. /ϵ/ par exemple).

5. VERS UN SYSTÈME DE SYNTHÈSE DU L.P.C. À PARTIR DU TEXTE

Ce corpus fournit un ensemble important de mouvements du L.P.C. et nous avons crée un premier synthétiseur L.P.C. à partir du texte utilisant la concaténation de segments de parole multi-modale. Si la synthèse par concaténation utilisant un grand vocabulaire et des unités multi-représentées est largement utilisée en synthèse acoustique [9] et plus récemment pour l'animation faciale [12], ce système est à notre connaissance le premier système générant des mouvements de main et de visage avec le son qui utilise la concaténation d'unités acoustiques et gestuelles. Deux types d'unités seront considérés par la suite : les diphones pour la génération acoustique et des mouvements du visage et les di-clés pour la génération des mouvements de la tête et de la main.



(a) Positions des marqueurs sur la codeuse

(b) Données issues de la capture de mouvements

(c) Modèle de forme haute définition

(d) Modèle d'apparence

FIG. 4: Passage des données issues de la capture de mouvement au modèle d'apparence pour un rendu vidéo-réaliste.

Ce corpus était initialement construit pour faire de la synthèse par concaténation de diphones acoustiques. La couverture des polysons est quasi-optimale : nous avons un minimum de deux exemples de chaque polyson avec un nombre de phrases minimum.

Bien que non totalement indépendantes, les positions et les formes de main sont presque orthogonales. La couverture du corpus en terme de succession de formes de main et de positions de main est satisfaisante : toutes les transitions de formes de main et toutes les transitions de positions de

main sont présentes. Un premier système de synthèse a été développé et il fonctionne en 2 étapes :

– le son et les mouvements faciaux sont traités par un premier système de synthèse par concaténation utilisant des polysons (des diphones si nécessaire) comme unités de base.
– les mouvements de tête, les mouvements de la main (forme et position) sont traités par un second système de synthèse par concaténation utilisant les di-clés comme unités de base.

Une procédure de lissage anticipatoire [3] est implémentée dans les 2 étapes. Elle permet d'éliminer dans la deuxième étape tout problème dû à l'absence d'une di-clé en la remplaçant par une di-clé de même forme et de position différente. En effet, cette interpolation linéaire à l'intérieur de la di-clé se charge d'adoucir les sauts trop brutaux entre 2 clés. Cette procédure à 2 étapes génère une synthèse L.P.C. acceptable. On considère en fait que le mouvement de la tête contribue à la réalisation de la constriction main/visage (20% du geste de constriction est dû à la tête) et on utilise une approximation brute de coordination geste/son (déduite d'une analyse des données) [8] : la cible est atteinte au milieu de la consonne dans le cas d'une séquence CV et en début de phone dans le cas de C ou V isolée. La figure 4 met en évidence le passage des données brutes à la réalisation d'un clone vidéo-réaliste au niveau du visage ; le passage à un modèle haute définition et à un texturage de la main est en cours.

6. CONCLUSIONS ET PERSPECTIVES

L'observation des codeurs en action est un pré-requis pour le développement de technologies de communication pour les malentendants (apprentissage, transcodage, synthèse, etc.). Cette analyse des données nous a permis de construire un premier système de synthèse de Langage Parlé Complété. L'analyse approfondie des données enregistrées sur notre codeuse L.P.C. nous permettra de mieux comprendre les coordinations temporelles entre le son et les mouvements. Tout ces informations permettront d'améliorer notre système de synthèse qui remplacera à la demande les sous-titrages télétextes dans le cadre du projet ARTUS. Une série de tests perceptifs est en cours d'élaboration.

7. REMERCIEMENTS

Nous tenions à remercier Yasmine Badsi, notre codeuse L.P.C. pour avoir accepté les contraintes de l'enregistrement. Nous remercions aussi Virginie Attina pour son expertise dans le domaine du LPC. Ce travail a été financé par le projet RNRT ARTUS.

RÉFÉRENCES

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

[2] G. Bailly and P. Badin. Seeing tongue movements from outside. In *International Conference on Speech and Language Processing*, pages 1913–1916, Boulder, Colorado, 2002.

[3] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, 2002.

[4] L. E. Bernstein, M. E. Demorest, and P. E. Tucker. Speech perception without hearing. *Perception and Psychophysics*, 62 :233–252, 2000.

[5] R. O Cornett. Cued speech. *American Annals of the Deaf*, 112 :3–13, 1967.

[6] R. O Cornett. Cued speech, manual complement to lipreading, for visual reception of spoken language. principles, practice and prospects for automation. *Acta Oto-Rhino-Laryngologica Belgica*, 42(3) :375–384, 1988.

[7] O. Engwall and J. Beskow. Resynthesis of 3d tongue movements from facial data. In *EuroSpeech*, Geneva, 2003.

[8] G. Gibert, G. Bailly, D. Beautemps, Eliséi F., and R. Brun. Analysis and synthesis of the 3d movements of the head, face and hands of a speech cuer. *Journal of the Acoustical Society of America*, submitted for publication.

[9] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, pages 373–376, Atlanta, GA, 1996.

[10] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer. On the correlation between facial movements, tongue movements and speech acoustics. In *Proceedings of International Conference on Speech and Language Processing*, pages 42–45, Beijing, China, 2000.

[11] J. Leybaert. The role of cued speech in language processing by deaf children : an overview. In *Auditory-Visual Speech Processing*, pages 179–186, St Jorioz, France, 2003.

[12] S. Minnis and A. P. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, pages 759–762, Beijing, China, 1998.

[13] G. Nicholls and D. Ling. Cued speech and the reception of spoken language. *Journal of Speech and Hearing Research*, 25 :262–269, 1982.

[14] E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28 :381–393, 1985.

[15] L. Revéret, G. Bailly, and P. Badin. Mother : a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *International Conference on Speech and Language Processing*, pages 755–758, Beijing, China, 2000.

[16] R. Uchanski, L. Delhorne, A. Dix, L. Braida, C. Reed, and N. Durlach. Automatic speech recognition to aid the hearing impaired : Prospects for the automatic generation of cued speech. *Journal of Rehabilitation Research and Development*, 31 :20–41, 1994.

[17] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26 :23–43, 1998.