Evaluation d'un système de synthèse 3D de Langue française Parlée Complétée

G. Gibert, G. Bailly, F. Elisei

Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex, France
{gibert, bailly, elisei}@icp.inpg.fr

ABSTRACT

This paper presents the virtual speech cuer built in the context of the ARTUS project aiming at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then incrusted - on demand and at the reception - in the original broadcast as an alternative to subtitling. The paper presents the multimodal text-to-speech synthesis system and the first evaluation performed by deaf users.

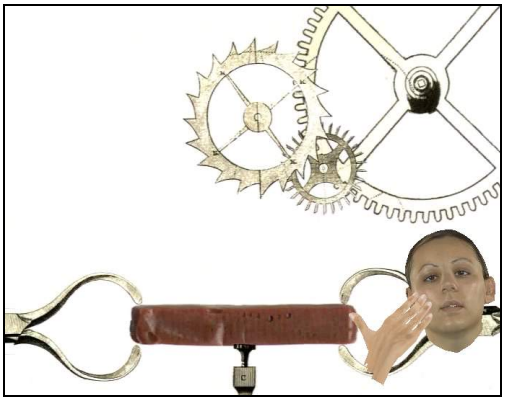

Figure 1: Incrustation du clone ARTUS dans un documentaire produit par la chaîne ARTE.

1. INTRODUCTION

Les personnes sourdes ou malentendantes dépendent grandement de la lecture labiale qui est basée sur l'information visuelle délivrée par les lèvres et le visage. Cependant, la lecture labiale seule est insuffisante dû à un manque d'information sur le point de l'articulation de la langue, des modes d'articulation (nasalité, voisement) et à la similarité de certaines formes labiales pour certains phonèmes (aussi appelés sosies labiaux tels que [u] *vs* [y]). Dans tous les cas, même le meilleur décodeur ne peut identifier plus de 50% de phonèmes dans des syllabes sans sens [16] ou dans des mots ou des phrases [5]. Le système de codage de la Langue française Parlée Complétée a été construit pour compléter la lecture labiale. Développé par Cornett [7, 9] et adapté depuis à plus de 50 langues [8], ce système est basé sur l'association de l'articulation faciale avec des clés formées par la main. Une clé est caractérisée par une position sur le visage (déterminant un sous-ensemble de voyelles) et une forme de main (déterminant un sous-ensemble de consonnes). Le code LPC pour les consonnes est représenté sur la Figure 2. De nombreuses études ont montré le gain d'intelligibilité apporté par ce codage comparé à la lecture labiale seule [15, 19] et son efficacité dans l'apprentissage de la langue écrite et orale [13, 14].

Des travaux sont consacrés à l'étude de la perception du code LPC et à sa production [1] mais peu de travaux s'attachent à la synthèse de celui-ci [voir les systèmes basés sur des règles dans 2, 10]. Nous allons décrire le système de synthèse multimodal produisant du code LPC à partir du texte et la première campagne d'évaluation auprès de personnes sourdes et malentendantes.

2. LE SYSTEME DE SYNTHESE MULTIMODAL

Le système de synthèse 3D de code LPC développé dans le cadre du projet ARTUS convertit une série de sous-titrage télétexte en un flux de paramètres d'animation pour les mouvements de la tête, du visage, du bras et de la main et produit également un signal acoustique. Les modèles de contrôle, de forme et d'apparence ont été déterminés à l'aide de plusieurs enregistrements multimodaux d'une locutrice oralisant et codant LPC.

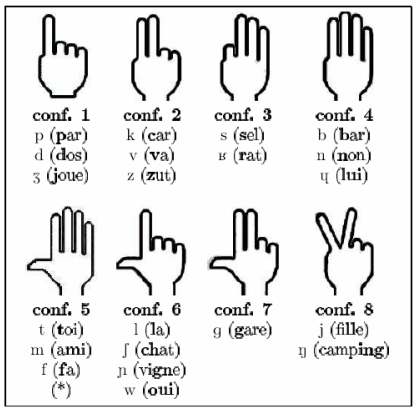

Figure 2 : Système de codage LPC pour les consonnes.

2.1. Expérimentation et modélisation

Les différentes configurations expérimentales pour enregistrer notre codeuse et capturer ses mouvements sont décrites dans [12]. Les configurations incluent (a) un système de capture de mouvements avec une bonne résolution temporelle (120Hz) et une bonne précision spatiale (0.1mm) de quelques dizaines de marqueurs rétro-réfléchissants collés sur le visage et la main de la codeuse (voir Figure 3), lors de la production de code LPC de 238 phrases; (b) un système de capture vidéo de plusieurs centaines de billes colorées collées sur le visage de notre locutrice (voir Figure 4), lors du codage de simples syllabes et

(c) des textures cylindriques de la tête, des moulages de sa main et de ses dents. A l'aide de toutes ces données, des modèles précis de forme et d'apparence de la tête et du visage ainsi que de la main de la codeuse ont été développés [11]. Ces deux types de modèles (forme et apparence) du visage et de la main sont pilotés par des paramètres quasi-articulatoires issus d'une analyse en composantes principales guidée par des connaissances articulatoires a priori.

2.2 Système de synthèse 3D de code LPC

Le système de synthèse de parole à partir du texte COMPOST [3] a été paramétré et des modules ont été ajoutés afin de générer la Langue française Parlée Complétée. Les ajouts sont les suivants:

Traitements linguistiques. Les sous-titrages ayant une ponctuation lâche voire absente, un module spécifique considère le début de chaque morceau de texte comme le début potentiel d'une phrase et abandonne les hypothèses peu probables. Pour respecter la synchronisation avec le contenu visuel, issue du télétexte originel, des marqueurs temporels sont insérés au début des phrases détectées. Ensuite, le module rythmique adapte la durée des pauses inter-phrases pour attendre ces rendez-vous.

Prosodie. Même si les codeurs LPC sont capables de minimiser l'impact de l'ajout d'un geste modal sur le débit de parole, le codage d'une consonne isolée (dans un cluster de consonnes ou dans une coda) impose un débit de parole plus faible et une hyperarticulation des syllabes complexes. L'intonation est également affectée. Le modèle prosodique SFC [4] a donc par conséquent été entraîné sur les données expérimentales. Avec le modèle ainsi appris, trois émissions télévisuelles ont été interprétées : seules quatre phrases n'ont pas été prononcées dans le temps imparti (retard moyen de 120 ms). Pour note, les règles gouvernant la création des sous-titres considèrent uniquement le nombre de lettres d'un groupe et non le temps de lecture.

Synthèse par concaténation d'unités multimodales. Le son et les mouvements de synthèse sont produits par sélection, lissage et concaténation d'unités multimodales multi-représentées. Deux types de segments sont considérés et synchronisés suivant des repères temporels acoustiques et gestuels déduits de règles spécifiques [12]: les "polysons" capturant le signal acoustique et les mouvements faciaux entre deux cibles acoustiques stables (les sons tels que les glides sont inclus dans des unités plus grandes) et les "diclés" qui contiennent les mouvements du bras, de la main et de la tête entre deux clés successives. Les mouvements de la tête contribuent, dans le cas de notre codeuse, significativement à la constriction main/visage : même si la main contribue pour la plus grande partie du mouvement menant la main à une position par rapport au visage, la tête effectue en moyenne à 16,43% de la distance à parcourir. Notons qu'une telle contribution des gestes posturaux à la structuration du discours a déjà été rapportée pour les signeurs natifs [6].

Les segments multi-représentés sont sélectionnés par un algorithme classique de programmation dynamique qui utilise un coût de sélection et un coût de concaténation. Le coût de concaténation prend en compte la contribution relative de chaque paramètre d'animation par rapport à la variance du mouvement total expliqué.

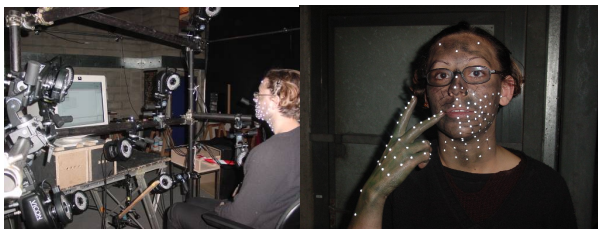

Figure 3 : Capture de mouvement avec un système Vicon® (12 caméras, 120Hz, 50 marqueurs sur la main et 63 sur le visage)

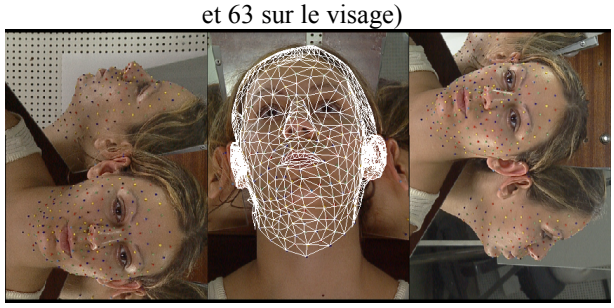

Figure 4 : Capture vidéo avec 247 billes collées sur le visage de la codeuse.

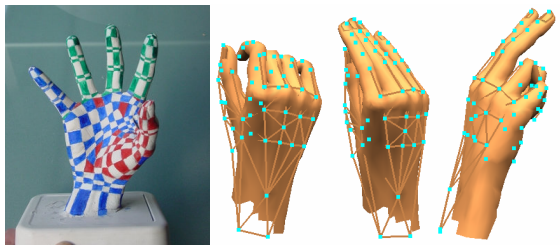

Figure 5 : A gauche : un moulage de la main. A droite : le modèle de main déduit est contrôlé par les données issues de la capture de mouvements.

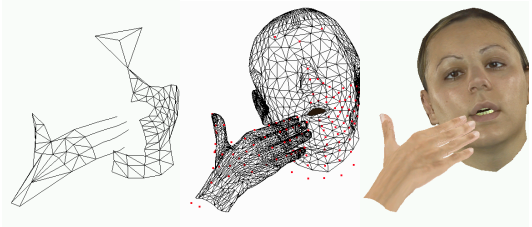

Figure 6 : Le clone du projet ARTUS. De la capture du mouvement à une animation vidéo-réaliste.

2.3 Animation vidéoréaliste

Un modèle d'apparence vidéo-réaliste a été développé pour la main de notre codeuse en utilisant des moulages de la main et la technique de *skinning* [20] de l'infographie (voir Figure 5). Des modèles génériques de lèvres, du crâne, de dents et d'yeux ont été adaptés à la morphologie du sujet. Le résultat de

ces procédures est la création d'un modèle de forme haute définition texturé par un mélange de textures cylindriques et contrôlé par un ensemble de paramètres quasi-articulatoires. Le clone vidéoréaliste résultant est représenté sur la Figure 6.

3. EVALUATION

Une première série d'expériences a été conduite afin d'évaluer l'intelligibilité de notre codeur face à des personnes malentendantes ou sourdes utilisant le code LPC. La première campagne d'évaluation est dédiée à l'intelligibilité segmentale tandis que la seconde est consacrée à la compréhension.

3.1 Intelligibilité segmentale

Paires minimales. Le test développé pour l'occasion est une adaptation du test de rhyme adapté pour le français par Peckels et Rossi [17]: les paires minimales ne testent pas, dans notre cas, les traits acoustiques mais les traits gestuels. Une liste de mots appariés de type CVC a été construite pour tester systématiquement les paires de consonnes en position initiale qui ne différent que dans la forme de main associée. Nous avons choisi toutes les paires dans chacun des 8 sous-ensembles codant les consonnes en LPC et qui étaient visuellement très proches [18]. Les voyelles centrales ont été choisies de telle sorte que toutes les positions par rapport au visage soient présentes et les consonnes finales ont été choisies afin de tester la capacité du système à gérer correctement la coarticulation. Comme toutes les paires minimales n'ont pu être générées dans tous les contextes vocaliques, nous obtenons une liste finale de 196 mots.

Conditions. Les stimuli par paires minimales sont présentés aléatoirement dans les deux ordres. La modalité lecture labiale seule est testée en premier. La modalité incluant le code LPC est présentée dans un deuxième temps afin de réserver les ressources cognitives pour la tâche la plus difficile i.e. la première tâche.

Stimuli. Pour la modalité de présentation « lecture labiale », afin d'éviter la présentation d'une tête complètement statique qui pourrait sembler non naturelle, nous avons divisé par 10 les mouvements de tête fournis par le système de synthèse. Aucune modification des mouvements segmentaux ou suprasegmentaux n'a été effectuée de manière à hyper-articuler.

Sujets. Les sujets sont au nombre de huit, ils sont sourds ou malentendants ayant appris la Langue française Parlée Complétée dès l'âge de 3 ans.

Résultats. Le taux de reconnaissance moyen pour la modalité « lecture labiale » est de 52.36%. Ce taux signifie que les paires proposées ne sont pas distinguables. Il revient donc au même de répondre au hasard. Il s'agit d'un premier résultat qui confirme que nos formes labiales entre sosies labiaux sont assez proches pour être confondues.

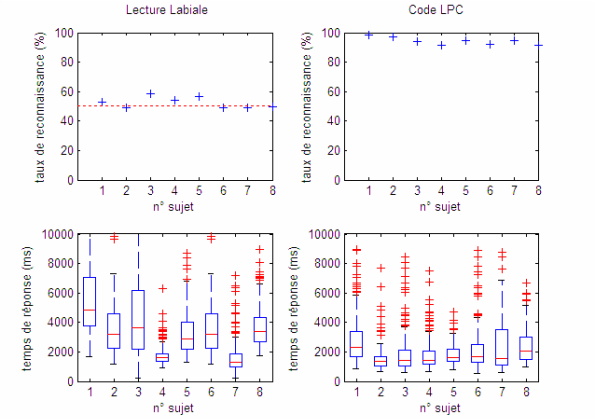

Figure 7 : Taux de reconnaissance et temps de réponse des 8 sujets pour les deux modalités (lecture labiale et lecture labiale + code LPC).

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	l	j
d	48	18	-	18	20	-	-	-	-	-	-	-	-	-	-	-	-
t	29	71	-	-	23	-	-	-	-	-	16	-	-	-	-	19	2
n	-	-	69	10	17	-	-	-	-	-	7	-	-	-	-	11	6
z	11	-	4	51	14	-	-	-	-	-	-	-	-	-	-	-	-
s	22	15	16	22	77	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	15	25	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	15	41	24	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	24	16	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	57	17	9	-	-	-	-	-	5
ʃ	-	-	-	-	-	-	-	-	17	61	12	14	-	-	-	-	-
g	-	14	7	-	-	-	-	-	15	19	115	9	-	-	17	-	4
k	-	-	-	-	-	-	-	-	-	13	11	45	-	-	-	-	3
v	-	-	-	-	-	-	-	-	-	-	-	-	17	15	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	18	14	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	17	-	-	-	51	17	3
l	-	22	27	-	-	-	-	-	-	-	-	-	-	-	24	47	-
j	-	4	3	-	-	-	-	-	13	-	1	2	-	-	7	-	26

Figure 8 : Matrice de confusion de la consonne initiale pour la modalité « lecture labiale ».

	d	t	n	z	s	p	b	m	ʒ	ʃ	g	k	v	f	ʁ	l	j
d	102	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
t	2	155	-	-	-	-	-	-	-	-	1	-	-	-	-	2	-
n	-	-	99	-	17	-	-	-	-	-	1	-	-	-	-	3	-
z	2	-	-	74	4	-	-	-	-	-	-	-	-	-	-	-	-
s	-	-	1	-	151	-	-	-	-	-	-	-	-	-	-	-	-
p	-	-	-	-	-	40	-	-	-	-	-	-	-	-	-	-	-
b	-	-	-	-	-	1	74	5	-	-	-	-	-	-	-	-	-
m	-	-	-	-	-	-	5	35	-	-	-	-	-	-	-	-	-
ʒ	-	-	-	-	-	-	-	-	85	2	1	-	-	-	-	-	-
ʃ	-	-	-	-	-	-	-	-	3	100	1	-	-	-	-	-	-
g	-	3	1	-	-	-	-	-	3	6	180	2	-	-	5	-	-
k	-	-	-	-	-	-	-	-	-	4	6	62	-	-	-	-	-
v	-	-	-	-	-	-	-	-	-	-	-	-	32	-	-	-	-
f	-	-	-	-	-	-	-	-	-	-	-	-	3	29	-	-	-
ʁ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-
l	-	1	1	-	-	-	-	-	-	-	-	-	-	-	1	117	-
j	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55

Figure 9 : Matrice de confusion de la consonne initiale pour la modalité « lecture labiale + code LPC ».

Le taux de reconnaissance moyen pour la modalité « lecture labiale + code LPC » est de 94.26%. La différence de taux de reconnaissance entre les deux modalités montre que notre codeur LPC apporte une information significative en terme de mouvements de main. On peut voir sur les matrices de confusion (Figure 8 et Figure 9), les erreurs faites par les sujets. Pour le groupe des consonnes bilabiales (encadré rouge), la consonne /p/ n'est pas reconnue dans la modalité « lecture labiale » (25 fois sur 40 elle est reconnue comme un /b/) alors qu'elle est toujours reconnue dans la modalité « lecture labiale + code LPC ».

Les temps de réponse qui sont un indice de la charge cognitive imposée aux sujets pour effectuer la tâche de discrimination entre les paires sont significativement différents (ANOVA à un facteur à mesures répétées $F(1,3134)=7.5$, $p<0.01$). Il est ainsi plus aisé pour les sujets de répondre à la tâche incluant le code LPC qu'à la tâche « lecture labiale ». Le gain est donc double, en termes de reconnaissance et en termes de charge cognitive nécessaire.

3.2 Compréhension

Afin d'évaluer la compréhension globale de notre système, nous avons demandé au sujets de l'étude précédente de visualiser un reportage de l'émission *Karambolage* de la chaîne ARTE dans lequel le clone LPC est incrusté (voir Figure 1). A la fin de la séance, nous avons demandé aux sujets de répondre à un questionnaire. Ce questionnaire se compose de 10 questions portant tant sur les informations apportées par la vidéo originale que par le clone ARTUS.

Le nombre moyen de réponses correctes est de 3 sur 10. Les sujets rapportent comprendre des mots isolés mais pas l'ensemble du discours. Cette observation est également constatée si l'on présente la vidéo de la codeuse incrustée. Une explication de ces résultats se trouve dans la complexité de la tâche proposée (rythme élevé, manque de marqueurs « prosodiques », etc.).

CONCLUSIONS ET PERSPECTIVES

L'observation et l'enregistrement d'une codeuse en action nous a permis de développer un système complet de synthèse LPC 3D à partir du texte. Les résultats préliminaires des tests perceptifs appliqués à ce système soulignent l'énorme gain en intelligibilité apporté par notre système. Cette série de tests doit se poursuivre sur un plus grand nombre de sujets pour pouvoir généraliser les résultats et quantifier plus finement la charge cognitive imposée aux sujets. Elle est actuellement complétée par des tests qui comparent l'usage du clone par rapport au télétexte à l'aide d'un dispositif oculométrique.

REMERCIEMENTS

Nous tenions à remercier Yasmine Badsi, notre codeuse LPC. Nous remercions également Martine Marthouret, Marie-Agnès Cathiard, Denis Beautemps et Virginie Attina pour leur aide dans l'élaboration des tests perceptifs. Nous n'oublions pas les sujets qui ont bien voulu participer.

BIBLIOGRAPHIE

[1] Attina, V. (2006) *La Langue française Parlée Complétée (LPC) : Production et Perception*. PhD Thesis. Institut National Polytechnique: Grenoble - France.

[2] Attina, V., Beautemps, D., Cathiard, M.-A., and Odisio, M. (2004) *A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer.* Speech Communication, **44**: p.197-214.

[3] Bailly, G. and Alissali, M. (1992) *COMPOST: a server for multilingual text-to-speech system.* Traitement du Signal, **9**(4): p.359-366.

[4] Bailly, G. and Holm, B. (2005) *SFC: a trainable prosodic model.* Speech Communication, **46**(3-4): p.348-364.

[5] Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (2000) *Speech perception without hearing.* Perception & Psychophysics, **62**: p.233-252.

[6] Brentari, D. (1999) *A prosodic model of sign language phonology*. Boston, MA: MIT Press.

[7] Cornett, R.O. (1967) *Cued Speech.* American Annals of the Deaf, **112**: p.3-13.

[8] Cornett, R.O. (1988) *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): p.375-384.

[9] Cornett, R.O. (1982) *Le Cued Speech*, in *Aides manuelles à la lecture labiale et perspectives d'aides automatiques*, F. Destombes, Editor. Centre scientifique IBM-France: Paris.

[10] Duchnowski, P., Lum, D.S., Krause, J.C., Sexton, M.G., Bratakos, M.S., and Braida, L.D. (2000) *Development of speechreading supplements based on automatic speech recognition.* IEEE Transactions on Biomedical Engineering, **47**(4): p.487-496.

[11] Elisei, F., Bailly, G., Gibert, G., and Brun, R. (2005) *Capturing data and realistic 3D models for cued speech analysis and audiovisual synthesis*. in *Auditory-Visual Speech Processing Workshop*. Vancouver, Canada

[12] Gibert, G., Bailly, G., Beautemps, D., Elisei, F., and Brun, R. (2005) *Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech.* Journal of Acoustical Society of America, **118**(2): p.1144-1153.

[13] Leybaert, J. (2000) *Phonology acquired through the eyes and spelling in deaf children.* Journal of Experimental Child Psychology, **75**: p.291-318.

[14] Leybaert, J. (2003) *The role of Cued Speech in language processing by deaf children: an overview*. in *Auditory-Visual Speech Processing*. St Jorioz - France. p.179-186.

[15] Nicholls, G. and Ling, D. (1982) *Cued Speech and the reception of spoken language.* Journal of Speech and Hearing Research, **25**: p.262-269.

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

[17] Peckels, J.P. and Rossi, M. (1973) *Le test de diagnostic par paires minimales. Adaptation au francais du 'Diagnostic Rhyme Test' de W.D. Voiers.* Revue d'Acoustique, **27**: p.245-262.

[18] Summerfield, Q. (1991) *Visual perception of phonetic gestures*, in *Modularity and the motor theory of speech perception*, I.G. Mattingly and M. Studdert-Kennedy, Editors. Lawrence Erlbaum Associates: Hillsdale, NJ. p. 117-138.

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

[20] Woodward, C.D. (1988) *Skinning techniques for interactives B-spline surface interpolation.* Computer-Aided Design, **20**(8): p.441- 451.