Mapping de l'espace spectral vers l'espace visuel de la parole: Les voyelles du Français en Langue Française Parlée Complétée

Zuheng Ming[1], *Gang Feng*[1], *Denis Beautemps*[1]

(1) Gipsa-lab, 11 rue des Mathématiques, Grenoble Campus, BP 46, F - 38402 SAINT MARTIN D'HERES Cedex

Denis.Beautemps@grenoble-inp.fr

RESUME

Cet article présente les résultats de l'approche statistique GMM pour le mapping des paramètres spectraux du signal acoustique de la parole vers les paramètres visuels de la Langue Parlée Complétée (LPC) au sens des moindres carrés, à un bas niveau d'interfaçage ce qui est innovant par rapport à l'approche classique texte-parole visuelle. A toute fin d'évaluation de l'approche GMM, nous présentons aussi les résultats de l'approche de modélisation multi-linéaire. Les résultats montrent que la méthode GMM améliore très significativement le mapping, tout particulièrement dans le cas de faible niveau de corrélation entre certains paramètres cibles comme ceux du LPC et les prédicteurs constitués des paramètres spectraux du signal acoustique de parole.

ABSTRACT

Mapping of the spectral space to the visual speech space for French vowels cued in Cued Speech

In this paper, we present a statistical method based on GMM modeling to map the acoustic speech spectral features to visual features of Cued Speech in the sense of least square error in a low signal level which is innovative and different with the classic text-to-visual approach. In comparison with the GMM based mapping modeling we first present the results with the use of a multi-linear model also at the low signal level and study the limitation of the approach. The experimental results demonstrate that the GMM based mapping method can significant improve the mapping performance compared with the multi-linear based mapping model especial in the sense of the weak linear correlation between the target and the predictor such as the hand positions of Cued Speech and the acoustic speech spectral features.

MOTS-CLES : LPC, LSP, MFCC, PARAMETRES LABIAUX, CONVERSION, MODELE LINEAIRE, GMMS.
KEYWORDS : Cued Speech, LSP, MFCC, Lips, Linear modeling, GMMs.

1 Introduction

Le cadre de cet article est la communication parlée chez les personnes sourdes. En France, cinq à six millions de personnes sont atteintes de surdité. Le recours à la lecture labiale est dans ce cas primordial pour la perception de la parole. Or l'information fournie par la forme des lèvres est ambigüe au sens où plusieurs sons de paroles peuvent avoir

des formes aux lèvres similaires ([p], [b], [m] par exemple) ce qui de ce fait rend difficile la perception complète de la parole sans information complémentaire (auditive, sémantique,...). Avec des conséquences pour le développement du langage chez l'enfant. La méthode du Cued Speech (Cornett, 1967) a été introduite pour combler ce manque. C'est un code manuel conçu pour désambiguïser la lecture labiale. Le locuteur, tandis qu'il parle, utilise une de ses mains pour pointer des positions particulières sur le visage, le côté du visage ou le cou (pour coder les voyelles) tout en présentant le dos de la main avec des formes particulières (8 clés digitales pour coder les consonnes). La main en position et présentant une clé digitale code ainsi une syllabe Consonne-Voyelle. Avec ce système, les sons similaires aux lèvres sont désambiguïsés par des positions ou clés digitales distinctes. Etendue à plus de 60 langues depuis son invention en 1967, dont la langue Française avec la Langue Parlée Complétée (LPC), cette méthode permet aux enfants sourds congénitaux stimulés par cette méthode depuis leur plus jeune âge d'accéder à un système phonologique complet de la langue parlée et d'avoir un développement du langage similaire à des enfants normo-entendants (Leybaert, 2000). Enfin, cette méthode renvoie vers l'audition comme l'indique son utilisation dans la pratique orthophoniste des enfants implantés cochléaires. L'objectif du travail présenté dans cet article est la conversion automatique du son de parole en paramètres de formes labiales et LPC. Ces paramètres sont les sorties des systèmes de synthèse visuelle incluant la modalité LPC (voir par exemple Attina et al., 2004 ; Gibert et al., 2005). Des travaux antérieurs ont développé de nombreux dispositifs visant à traduire automatiquement le son de parole en clés LPC. Tous s'appuient sur le couplage d'un système de reconnaissance automatique avec un système visuel de génération des clés du LPC (Autocuer, Cornett, 1988) ou un système de synthèse de la main (Duchnovski et al., 2000) ou encore un système de parole audio visuelle (Attina, et al., 2004 ; Gibert et al., 2005 ; Beautemps et al., 2007). Dans ces différents dispositifs le recours au niveau syntaxique est une des clés du système. Ceci a l'inconvénient de perdre la richesse contenue dans la variabilité du signal de parole. L'objectif visé ici est ainsi d'étudier les méthodes de mapping des paramètres acoustiques du son de parole vers les paramètres visuels (labiaux et LPC) en utilisant un bas niveau d'interfaçage de type signal et donc sans le recours à la reconnaissance automatique de la parole. L'introduction de la composante manuelle du LPC dans ce programme constitue une véritable originalité de ce travail avec des retombées claires pour les systèmes de communication utilisant le geste associé à la parole ou non, tels que le LPC mais aussi des gestes de pointage ou la Langue des Signes. Nous abordons ce programme en traitant le cas des voyelles orales du Français. Le mapping consistera ici à déterminer les coefficients d'une combinaison linéaire reliant les paramètres de l'espace acoustique (les prédicteurs) aux paramètres de l'espace visuel (lèvres et LPC) en minimisant l'erreur au sens des moindres carrés entre le résultat de la prédiction et les valeurs des paramètres visuels. Dans la suite, nous présenterons tout d'abord l'expérimentation et les paramètres considérés pour caractériser chacun des espaces, puis nous étudierons les limites de l'approche par prédiction multi-linéaire pour finir par les résultats de la prédiction multi-gaussiennne GMM.

2 Expérimentation, paramètres spectraux, visuels et LPC

2.1. Dispositif expérimental

Les données sont composées d'un enregistrement vidéo d'un locuteur prononçant et codant un corpus de 50 mots isolés. Les mots étaient constitués de 32 nombres (de zéro à 31), des douze mois de l'année et de six mots couramment rencontrés en Français. Chaque mot était présenté sur un moniteur placé en face du locuteur, dans un ordre

aléatoire. Le corpus a ainsi été répété 10 fois. Le locuteur est une femme de langue maternelle française, codeuse et diplômée en LPC. L'enregistrement a été réalisé en chambre sourde à la fréquence vidéo de 25 Hz conformément au banc expérimental vidéo du laboratoire. Le locuteur était assis en face de la caméra et d'un micro pour la bande son numérisée à la fréquence de 44100 Hz. Des pastilles colorées étaient placées sur la peau entre les arcades sourcilières et à l'extrémité des doigts pour permettre l'extraction des coordonnées après numérisation des images. Enfin, un panneau quadrillé placé dans le plan du visage a été enregistré par la caméra pour permettre une conversion pixel/centimètre pour la suite des traitements. L'enregistrement vidéo réalisé en format PAL, a été numérisé comme des images RGB constituées de l'entrelacement des deux ½ trames vidéo (composées des lignes impaires et paires respectivement). Pour chaque image ainsi numérisée, les deux ½ trames ont été reconstituées et pour chacune les lignes manquantes obtenues par interpolation linéaire des autres lignes de façon à obtenir deux images complètes séparées de 20 ms.

2.2. Extraction des paramètres de lèvres, de main et spectraux

Ces trames définissent l'ensemble des images à la cadence de 50 Hz auxquelles il sera fait référence dans la suite. De cet ensemble de données, des images correspondant aux instants t_0 des voyelles en position labiale cible ont été sélectionnées manuellement par un des expérimentateurs ainsi que l'extraction des coordonnées du contour interne des lèvres desquels les paramètres labiaux d'étirement (A), d'aperture (B) et d'aire intéro-labiale (S) ont alors été calculés selon les formules classiques dans le domaine (Lallouache, 1991). De même, les images correspondants aux instants t_1 auxquels la main pointant la position LPC atteint la cible de la voyelle ont été sélectionnées et les coordonnées (x,y) de l'extrémité du doigt « pointeur » ont été extraites en référence au centre de la pastille placée entre les arcades sourcillères. Enfin pour chaque instant t_0, deux extraits du son de parole correspondant (pondérés par une fenêtre de Hamming) d'une durée de 20 et 32 ms centrés sur t_0 sont utilisés pour le calcul des 16 coefficients spectraux LSP et MFCC respectivement. Enfin, 4 formants ont été extraits de l'enveloppe spectrale obtenue à partir des coefficients LSP. L'ensemble de ces traitements a ainsi permis de constituer deux bases de données de 331 et 263 éléments (Table 1) avec pour chacun, les 4 composantes principales (par Analyse en Composantes Principales) de l'ensemble des 4 formants, les 16 composantes principales des coefficients LSP, ceux des coefficients MFCCs, les 32 composantes principales de l'ensemble des coefficients LSP et MFCC, les coordonnées (x,y) liés à la main et les 3 paramètres labiaux (A,B,S).

Taille	[a]	[i]	[u]	[y]	[ø]	[œ]	[e]	[ɛ]	[o]	[ɔ]
331 (apprentissage)	24	67	22	10	28	37	51	68	15	9
263 (test)	35	51	14	14	14	19	50	42	15	9

Table 1- Bases de données des 331 voyelles pour l'apprentissage et 263 pour le test.

3 La modélisation multi-linéaire

3.1. Méthode

L'objectif de cette partie est de prédire les paramètres labiaux (A,B,S) et de main (x,y) à partir des paramètres spectraux (leurs composantes principales F_i) par la méthode de prédiction multi-linéaire. De façon à ordonner ces prédicteurs F_i, le coefficient de corrélation ρ de chacun des F_i avec le paramètre à prédire est calculé. Les prédicteurs sont alors ordonnés dans le sens décroissant des valeurs de leur ρ pour donner un ensemble ordonné $\{F_1,F_2,\ldots F_{32}\}$. Ainsi pour le paramètre B, celui-ci est tout d'abord centré sur sa valeur moyenne puis soumis à une régression linéaire avec F_1 ce qui permet d'obtenir le coefficient k_1 de régression linéaire. L'erreur résiduelle de prédiction est alors soumise à une régression linéaire avec F_2 et ainsi de suite jusqu'à l'ordre p (p entier tel que $1 \leq p \leq 32$). La formule de prédiction à l'ordre p s'écrit alors :

$$\hat{B} = k_1 F_1 + k_2 F_2 + \ldots + k_p F_p + \overline{B}$$

3.2. Résultats sur les paramètres labiaux

Cette formule de prédiction a été appliquée aux paramètres labiaux (A,B,S) à partir de l'analyse de la base de données composée des 331 observations des voyelles (Table 1). La figure 1 présente les résultats de prédiction. Elle montre une décroissance de la variance résiduelle en fonction du nombre de prédicteurs. Ce résultat conduit à trois remarques. Tout d'abord, la variance résiduelle reste élevée avec l'utilisation des formants (de l'ordre de 30% de la variance totale). Ce résultat est vraisemblablement dû au manque de dimension. En effet, l'utilisation de 16 paramètres MFCC et LSP améliore très significativement le résultat (variance résiduelle entre 15 et 20%). Une seconde remarque consiste à constater que les MFCCs permettent une décroissance rapide tandis que les LSPs permettent d'atteindre une plus faible variance résiduelle. La prédiction utilisant les 16 composantes principales de l'ensemble des LSPs et MFCCs permet à la fois une baisse rapide et une valeur finale basse. La troisième remarque est liée au constat que l'erreur reste malgré tout relativement élevée (10%). Ces résultats servent de référence pour la suite de la modélisation, notamment pour le choix des prédicteurs performants.

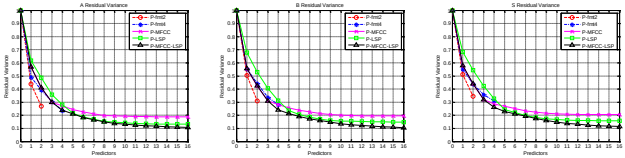

Figure 1 – Variance résiduelle de la prédiction des paramètres labiaux, de gauche à droite respectivement A, B et S, exprimés chacun relativement à leur variance totale, en fonction du nombre de prédicteurs (leurs composantes principales) et pour chaque ensemble de prédicteurs : 2, 4 formants (fmt), MFCC, LSP et l'ensemble de MFCC-LSP.

3.3 Résultats sur les paramètres LPC de main

La même méthode d'analyse a été appliquée pour les coordonnées (x,y) de la main. La valeur finale de la variance résiduelle atteint 40,8 % pour la coordonnée x et 35,5% pour y, même dans le cas des meilleurs prédicteurs constitués de l'ensemble des paramètres LSPs et MFCCs. Cette valeur élevée de la variance résiduelle finale s'explique par les valeurs faibles des corrélations ρ (0,43 et 0, 42 respectivement pour les coordonnées x et y). Cette faible corrélation est vraisemblablement liée à la relation d'ordre entre les coordonnées et les paramètres spectraux, ce qui constitue une limite de la méthode. Afin de vérifier cette hypothèse nous avons redistribué les positions (x, y) du LPC de façon

cohérente avec le triangle vocalique défini par l'espace des deux premiers formants, étant donné la forte corrélation entre les formants et les paramètres spectraux (0,96 et 0,98 pour les 2 premiers formants). Nous avons alors pu observer une grande baisse de la variance résiduelle pour atteindre finalement 7,85% et 7,08% respectivement pour x et y).

4 La modélisation multi-gaussienne GMM

4.1. Méthode

Dans cette partie, l'espace spectral est caractérisé par les 16 premières composantes principales de l'ensemble des coefficients LSP et MFCC, composant les éléments du vecteur x de dimension N (1≤N≤32): En référence à la formulation de Kain (2001) (voir aussi Hueber et al., 2011), l'estimateur (au sens des moindres carrés) du paramètre y est une combinaison de l'observation x pondérées par les m probabilités $P(c_i/x)$ additionnée d'un biais :

$$\hat{y} = F(x) = \sum_{i=1}^{m} (W_i x + b_i) \cdot P(c_i \mid x)$$

$P(c_i/x)$ étant la probabilité conditionnelle a posteriori que l'observation x soit générée par le modèle gaussien c_i (de moyenne μ_i^X et de covariance $\sum_i^{YX}$),

W_i et b_i étant respectivement la matrice de transformation et de biais associés à c_i.

$$b_i = \mu_i^Y - \sum_i^{YX} (\sum_i^{XX})^{-1} \mu_i^X$$

$$W_i = \sum_i^{YX} (\sum_i^{XX})^{-1}$$

$$P(c_i \mid x) = \frac{\alpha_i N(x, \mu_i^X, \sum_i^{XX})}{\sum_{p=1}^{m} \alpha_p N(x, \mu_p^X, \sum_p^{XX})}$$

où :

α_i est le coefficient de pondération du modèle gaussien c_i, la somme de tous les coefficients valant 1 ; $\sum_i^{YX}$ étant la matrice de covariance entre $\boldsymbol{x}$ et le $\boldsymbol{y}$ calculée sur le sous-ensemble de données i et μ_i^Y la moyenne du paramètre $\boldsymbol{y}$ de ce même sous-ensemble. Dans cette expérience, le nombre de gaussiennes a été fixé à $m=3$ pour l'estimation des paramètres labiaux. En effet dans cet espace les voyelles sont traditionnellement réparties en 3 sous-ensembles de voyelles (les visèmes) : [a,i,e,ɛ], [y,o,u,ø] et [œ, ɔ]. Pour les coordonnées de main, le nombre de gaussiennes a été fixé à m=5 pour rendre compte des 5 positions du LPC pour les voyelles du Français ce qui donne dans ce cas 5 sous-ensembles de voyelles : [a, o, œ], [i], [ɛ,u,ɔ], [ø] et [y,e]. Les moyennes et matrices de covariances des modèles gaussiens ont été calculées sur ces sous-ensembles. Enfin, les 16 composantes spectrales sont ordonnées dans l'ordre décroissant de leur explication de la variance du paramètre estimé. Les N composantes du vecteur x sont alors les N premières composantes selon cet ordre.

4.2 Résultats de la modélisation

Nous avons appliqué la modélisation au corpus d'apprentissage des 331 observations des voyelles (Table 1). La Figure 2 présente les résultats de l'analyse de la variance des

données. Elle montre une décroissance de la variance résiduelle en fonction de la dimension de l'espace spectral pour atteindre une valeur très basse en dessous de 5%. En comparaison des meilleurs résultats de la modélisation linéaire, on observe une amélioration significative. La décroissance rapide montre que la méthode converge avec peu de dimensions (5 pour les lèvres et 8 pour la main). Nous avons tenté d'améliorer ce résultat en utilisant l'algorithme k-means et l'algorithme EM (Expected Maximization). Nous n'avons pas obtenu de meilleurs résultats même en augmentant le nombre de gaussiennes de manière significative.

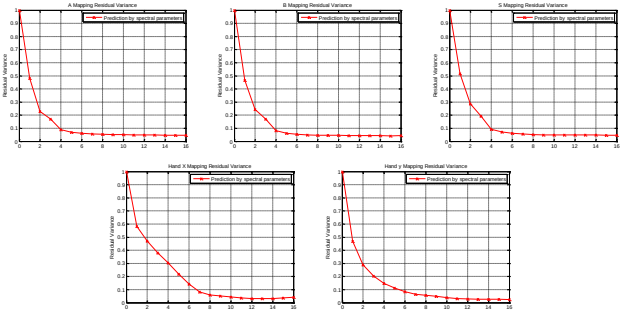

Figure 2 - Variance résiduelle des paramètres labiaux et de main exprimés chacun relativement à leur variance totale, en fonction de la dimension de l'espace spectral.

4.3 Evaluation

L'évaluation consiste à tester l'estimateur en fonction des dimensions de l'espace spectral sur la base des données de test (Table 1). Nous présentons les résultats de l'évaluation tout d'abord en terme de variance expliquée (Figure 3) pour permettre une comparaison avec la phase d'analyse puis en terme d'erreur quadratique moyenne (Figure 4). Ainsi sur la Figure 3 relative au paramètre labial *A*, le point de la courbe à l'abscisse p correspond au résultat du test en utilisant l'estimateur composé de 3 gaussiennes à p dimensions dont les paramètres ont été calculés sur la base d'apprentissage. De façon similaire sur la courbe correspondant à la coordonnée *y*, le point à l'abscisse *p* correspond au résultat du test en utilisant l'estimateur composé de 5 gaussiennes à *p* dimensions dont les paramètres ont été calculés sur la base d'apprentissage. Les valeurs finales de la Figure 3 sont de l'ordre de 10 à 12 %. Ces valeurs sont supérieures à celles obtenues avec les données d'apprentissage tout en restant proches. Avec l'ordre déterminé dans l'apprentissage, on constate de légères fluctuations dans la décroissance. Néanmoins, la décroissance reste rapide. L'ensemble montre une bonne adéquation du modèle. Ainsi la Figure 4 montre que les erreurs quadratiques moyennes suivent la même évolution pour atteindre une précision finale de 0,4 cm, 0,15 cm et 0,6 cm2 pour A, B et S respectivement et 1 cm pour les coordonnées *x* et *y*. Pour ces derniers, la variance de la prédiction à l'intérieur de chaque position LPC est similaire à celle des données et est centrée sur leur valeur moyenne.

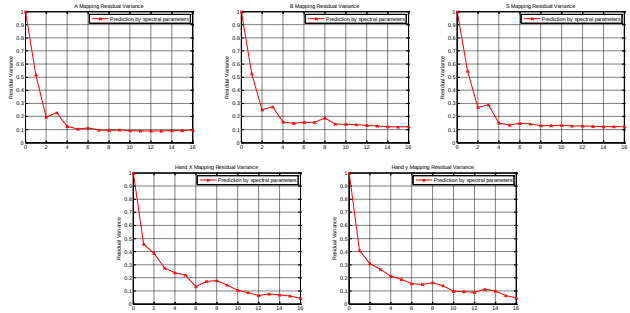

Figure 3 - Variance résiduelle sur les données de test des paramètres labiaux et de main exprimés chacun relativement à leur variance totale, en fonction de la dimension de la dimension de l'espace spectral.

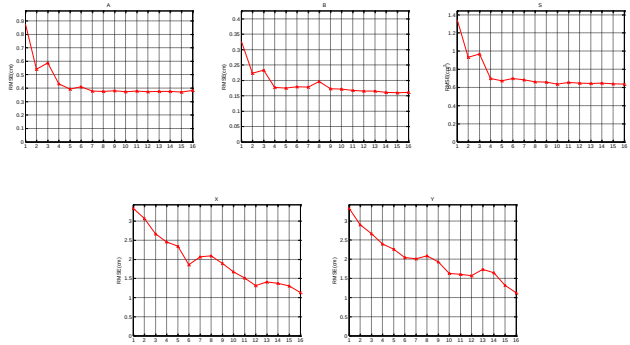

Figure 4 - Erreur quadratique moyenne (RMSE) sur les données de test des paramètres labiaux et de main en fonction de la dimension de l'espace spectral.

5 Conclusion

Cet article étudie les relations entre l'espace des paramètres spectraux de la parole et l'espace visuel de la parole et de la Langue parlée Complétée (LPC) dans l'objectif d'un mapping de l'un vers l'autre. Nous avons abordé ce programme avec le cas des voyelles du Français. Ainsi nous avons tout d'abord exploré la modélisation multi-linéaire pour convertir les paramètres spectraux vers les paramètres labiaux ainsi que les paramètres de la Langue Parlée Complétée. Les résultats montrent que les meilleurs prédicteurs sont 16 paramètres issus d'une analyse en composantes principales de l'ensemble composé de 16 coefficients LSP et 16 coefficients MFCC. L'approche linéaire a montré ses limites pour le cas de la composante manuelle du LPC. Nous avons ensuite testé l'approche GMM avec la modélisation multi-gaussienne de l'espace spectral. Les résultats ont été améliorés aussi bien pour les paramètres de lèvres que pour ceux de la composante LPC avec une explication de 95% de la variance totale sur le corpus d'apprentissage. Ces résultats (les meilleurs obtenus) ont été observés lorsque les gaussiennes étaient

distribuées en fonction de connaissances phonétiques sur les représentations labiales et LPC des voyelles du Français. Dans le cadre de la conversion automatique, ces résultats prometteurs restent à être intégrés dans les systèmes de synthèse visuelle de la parole pour une évaluation perceptive auprès des personnes sourdes utilisatrices du LPC.

Remerciements

Les auteurs souhaitent remercier Myriam Diboui, la codeuse en LPC, pour avoir accepté les contraintes d'enregistrement. Ces travaux sont soutenus par l'Agence Nationale de la Recherche Française au travers des projets TELMA et PLASMODY.

Références

ATTINA, V., BEAUTEMPS, D., CATHIARD, M. A. & ODISIO, M. (2004). "A pilot study of temporal organization in Cued Speech production of French syllables: rules for Cued Speech synthesizer," Speech Communication, vol. 44, pp. 197-214.

BEAUTEMPS, D., GIRIN, L., ABOUTABIT, N., BAILLY, G., BESACIER, L., BRETON, G., BURGER, T., CAPLIER, A., CATHIARD, M.A., CHÊNE, D., CLARKE, J., ELISEI, F., GOVOKHINA, O., LE, V.B., MARTHOURET, M., MANCINI, S., MATHIEU, Y., PERRET, P., RIVET, B., SACHER, P., SAVARIAUX, C., SCHMERBER, S., SÉRIGNAT, J.F., TRIBOUT, M., VIDAL, S. (2007), "TELMA: Telephony for the Hearing-Impaired People, From Models to User Tests," In Proceedings of. ASSISTH 2007, pp. 201–208.

CORNETT, R. O. (1967). "Cued Speech," American Annals of the Deaf ,112, 3-13, 1967.

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): 375-384.

DUCHNOVSKI, P., D. S. LUM, J. C. KRAUSE, M. G. SEXTON, M. S. BRATAKOS AND L. D. BRAIDA (2000). "Development of speechreading supplements based on automatic speech recognition." IEEE Transactions on Biomedical Engineering 47(4): 487-496.

GIBERT, G., BAILLY, G., BEAUTEMPS, D., ELISEI, F., & BRUN, R. (2005). "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using Cued Speech," J. Acoust. Soc. Am., vol. 118(2), pp. 1144–1153.

HUEBER, T., BENAROYA, E.L, DENBY, B., CHOLLET, G. (2011). "Statistical Mapping between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface", Proceedings of Interspeech, pp. 593-596, Firenze, Italia.

KAIN, A. (2001). High-resolution voice transformation (PhD, OGI School of Science & Engineering, Oregon Health & Science University).

LALLOUACHE, M.T. (1991). "UN POSTE VISAGE-PAROLE COULEUR. ACQUISITION ET TRAITEMENT AUTOMATIQUE DES CONTOURS DES LEVRES," PH.D. THESIS, INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, 1991.

LEYBAERT, J., 2000. PHONOLOGY ACQUIRED THROUGH THE EYES AND SPELLING IN DEAF CHILDREN. JOURNAL OF EXPERIMENTAL CHILD PSYCHOLOGY 75, 291-318.