

# Détermination de la position du voile du palais à partir du signal de parole pour les nasales du français.

*Solange Rossato, Pierre Badin, Gang Feng*

Institut de la Communication Parlée  
UMR 5009, CNRS, Univ. Stendhal, INPG – Grenoble, France  
Tél.: ++33 (0)476 82 41 20 - Fax: ++33 (0)476 82 38 45  
Mél: rossato@icp.inpg.fr

## ABSTRACT

This paper deals with the recovery of the velar position from speech signals for the French nasals for one subject. Articulatory movements were recorded using electromagnetic articulography (EMA) synchronized with the acoustic signals. The maximum of likelihood method was used to estimate the velar position from speech signals. The results show that the velum position is well recovered with a root mean square error of 0.12 cm, when the learning is performed over the complete corpus. Higher precision is achieved when nasal vowels and nasal consonants are considered separately.

## 1. INTRODUCTION

L'inversion de la parole consiste à retrouver les mouvements articulatoires à partir du signal de parole. Ce thème de recherche a donné lieu à de nombreux travaux ces dernières années. Les enjeux sont grands autant pour le codage à bas débit et la synthèse articulatoire que pour l'aide à la rééducation. Retrouver les gestes articulatoires à partir du son est un problème mal défini dans le sens où plusieurs configurations articulatoires peuvent donner des signaux acoustiques très proches ([Abr94]). De nombreuses approches ont besoin de connaissances préalables issues de données, par exemple pour construire un dictionnaire de liens entre paramètres acoustiques et paramètres articulatoires, pour entraîner un réseau de neurones.

Les relations entre espace acoustique et espace articulatoire sont complexes, et l'effet d'un articulateur sur le signal acoustique n'est pas toujours bien connu. Dans le cas particulier de la nasalité, un deuxième conduit, le conduit nasal, entre en jeu et ajoute une complexité supplémentaire. Ainsi, si les voyelles sont bien caractérisées par leurs formants, la nasalisation d'une voyelle introduit des pôles et des zéros qui ont des effets différents suivant la voyelle et le degré de couplage entre conduit oral et conduit nasal ([Che95], [Fen96]). Le principal effet de ces paires de pôles / zéros est un aplatissement du spectre en basse fréquence, un élargissement des bandes passantes, et une atténuation des formants qui deviennent ainsi plus difficiles à détecter.

La grande variabilité du signal de parole et la difficulté d'observations des mouvements du voile du palais nous a conduit à réaliser une première étude ([Ros98]) utilisant

une modélisation des voyelles nasales. L'avantage de la modélisation réside dans le contrôle tous les paramètres ainsi que dans une grande facilité pour générer des données. L'utilisation du maximum de vraisemblance a montré des résultats encourageants avec un taux d'estimation correcte d'environ 90%. Le travail présenté dans cet article vise maintenant à tester la méthode du maximum de vraisemblance non plus sur des données issues d'un modèle mais sur des données enregistrées sur un locuteur homme avec un articulographe électromagnétique (EMA).

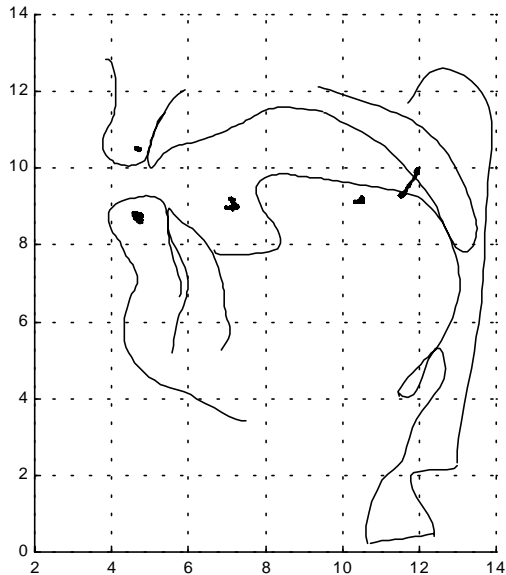
## 2. LES DONNEES

Deux corpus ont été enregistrés pour un seul locuteur homme, l'un pour les voyelles nasales, l'autre pour les consonnes nasales. L'objectif est d'obtenir des articulations ne se différenciant que pour le voile du palais. Le locuteur prononce plusieurs répétitions des quatre voyelles nasales du français /ã/, /õ/, /ẽ/ et /œ/ /, suivies de leur contrepartie orale avec pour consigne de ne pas bouger les articulateurs autres que le voile du palais. Dans le cas des consonnes, des séquences /pVNVpVCV/ ont été enregistrées, où N représente la consonne nasale et C la consonne occlusive correspondante, en contexte /a i u/. Les paires de consonnes nasale/orale ont été choisies sur le critère d'une articulation proche où seule la position du velum diffère : paires n/d et m/b. Le corpus des voyelles est constitué de 36 réalisations et celui des consonnes en contient 102.

### 2.1 Description du dispositif expérimental

Deux capteurs EMA ont été fixés au niveau des incisives supérieures (référence) et des incisives inférieures (mâchoire). Deux autres capteurs ont été positionnés sur la langue (pointe, partie laminaire). Un cinquième capteur a été collé sur le velum (voir figure 1).

Les coordonnées des capteurs sont échantillonnées à 1 kHz. Les mouvements des articulateurs sont relativement lents ce qui permet d'appliquer un filtre passe-bas sur toutes les données afin d'éliminer une partie du bruit de mesure. Un micro permet d'enregistrer le signal de parole sur DAT avec un excellent rapport signal à bruit. Ce signal est ensuite échantillonné à 16 kHz et les trames sont synchronisées à 100 Hz avec les positions des capteurs.



**Figure 1.** Trajectoires des cinq capteurs EMA pour la transition /āa/ superposées avec le contour obtenu par IRM pour le même sujet et pour l’articulation /ā/.

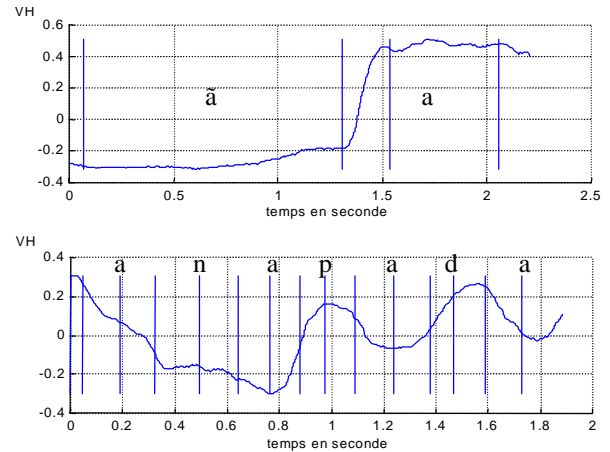
Les trajectoires des capteurs permettent de vérifier les mouvements des articulateurs du locuteur. Concernant le corpus des voyelles nasales et de leur contrepartie orale, les capteurs indiquent un très léger mouvement de la langue lors de la transition (de l’ordre de 1 mm) et aucun mouvement de la mâchoire. L’observation des capteurs pour le corpus des consonnes montre que la durée moyenne de l’occlusion dentale est du même ordre de grandeur pour le /n/ que pour le /d/ avec des évolutions semblables pour les articulateurs autres que le velum. Le critère de départ qui consiste à obtenir des articulations ne se différenciant que pour la position du velum semble donc respecté.

## 2.2 Traitement des signaux

### Paramètre indicateur de la position du voile du palais

Le but de ce prétraitement est de déterminer un paramètre unique qui soit représentatif des mouvements du velum au lieu des deux coordonnées X et Y du capteur EMA du velum. Le mouvement de ce capteur est relativement rectiligne avec une pente qui varie peu d’un item à l’autre. Le choix du paramètre « hauteur » du velum se porte donc vers la projection du point (X,Y) sur une droite. La droite choisie est la moyenne des droites de régression obtenues pour chaque item.

Cette « hauteur » du velum (VH) a un mouvement d’amplitude 0,8 à 1 cm environ pour les voyelles nasales / $\tilde{e}$ / et / $\sim\alpha$ / ; plutôt entre 0,6 et 0,8 cm pour les voyelles plus ouvertes / $\tilde{a}$ / et / $\tilde{o}$ /. Pour les séquences /pVNVpVVCV/, les mouvements sont en général d’amplitude plus faible. La figure 2 présente deux exemples de courbes obtenues pour le paramètre VH.



**Figure 2:** Paramètre VH « hauteur » du velum en cm. Pour l’item /āa/, l’amplitude du mouvement est de l’ordre de 0,8 cm. Pour l’item /panapada/, on remarque trois niveaux : relevé pour /p/ et /d/, abaissé pour /n/, avec une position intermédiaire pour les /a/. La voyelle /a/ qui suit /n/ garde la position très basse du velum atteinte pendant la consonne /n/. Ce phénomène est également observé pour les voyelles /i/ et /u/ suivant la consonne /n/.

**Signaux acoustiques** Dans un premier temps, le signal acoustique est étiqueté pour faciliter l’extraction d’une portion du signal. Le signal acoustique est découpé en trames de longueur 32 ms avec un recouvrement de 22 ms. Pour chaque trame, le signal est d’abord filtré passe-haut puis une analyse par banc de filtre fournit 16 valeurs à partir desquelles on obtient les 16 coefficients melcepstres (MFCC).

## 3. MAXIMUM DE VRAISEMBLANCE

### 3.1 La Méthode

Le signal acoustique, représenté par un vecteur S de coefficients, dépend de façon complexe du paramètre VH. Il s’agit de décider, à partir d’un vecteur S, quelle est la valeur la plus probable de VH, notée  $VH_{est}$ . La méthode du maximum de vraisemblance peut s’appliquer à ce problème : la valeur de  $VH_{est}$  est celle qui maximise la vraisemblance du vecteur S pour le paramètre VH,  $p(S/VH)$ . Puisque VH est un paramètre continu, la probabilité que le paramètre estimé soit une valeur précise  $VH_0$  est très faible : une certaine incertitude existe autour de cette valeur. L’ensemble des valeurs que peut prendre VH est donc découpé en N intervalles. Si les intervalles sont trop petits, les probabilités sont plus faibles et les résultats moins fiables. Si les intervalles sont trop grands, la précision est plus faible. Il faut donc trouver un compromis entre fiabilité et précision dans l’estimation.

Connaissant le vecteur S, on calcule, pour chaque intervalle  $V_i$  de centre  $v_i$ , la probabilité conditionnelle  $p(S/v_i)$  aussi appelée vraisemblance du vecteur S sachant

**Tableau 1** : Le tableau 1 présente les erreurs RMS (en cm) obtenues pour les voyelles nasales ( $\tilde{V}$ ), les transitions (tr) et leur contrepartie orale (V) ainsi que pour les consonnes nasales (N) et orales (C) et les voyelles adjacentes : VNV et VCV, pour différentes bases d'apprentissage.

Apprentissage	$\tilde{V}$	tr	V	V	N	V	V	C	V
Les 2 corpus	0.08	0.22	0.10	0.10	0.11	0.11	0.08	0.08	0.10
Corpus voyelles	0.09	0.14	0.09	-	-	-	-	-	-
Parties vocaliques	0.07	0.20	0.12	0.07	0.23	0.12	0.07	0.08	0.08
Consonnes	-	-	-	-	0.08	-	-	0.07	-

que le paramètre VH a pour valeur  $v_i$ . Pour un vecteur S donné, on obtient une probabilité pour chaque intervalle de centre  $v_i$ . L'estimation du paramètre VH est la valeur  $v_i$  qui a la plus grande probabilité. On peut également calculer l'espérance en associant à chaque centre  $v_i$ , la probabilité  $p(S/v_i)$  normalisée. Le paramètre estimé  $VH_{est}$  est alors un paramètre continu comme l'est le paramètre VH, « hauteur » du velum.

### 3.2 Détermination des probabilités conditionnelles

Pour appliquer la méthode du maximum de vraisemblance, les probabilités conditionnelles  $p(S/v_i)$  doivent être connues au préalable. Ces probabilités sont considérées comme des lois normales estimées à partir de statistiques effectuées sur les données articulatoires et acoustiques extraites des corpus enregistrés : ensemble de vecteurs de coefficients MFCC et valeur du paramètre VH correspondant. La plage de variation de VH est divisée en intervalles de largeur 0,1 cm. Cette taille permet d'avoir assez de données pour chaque intervalle pour déterminer la probabilité conditionnelle  $p(S/v_i)$  : entre 300 et 5000 vecteurs S sur lesquels sont calculées moyenne et matrice de covariance. Les lois normales sont donc entièrement définies. Pour chaque corpus, entre 9 et 12 itérations de chaque item sont prononcées par le locuteur. Les 5 premières sont utilisées pour l'apprentissage, c'est-à-dire pour déterminer les probabilités conditionnelles, les autres sont réservées au test.

## 4. ANALYSE DES RESULTATS

La détermination de la position du voile du palais est effectuée tout d'abord de façon globale en prenant en compte sans distinction le corpus de voyelles et le corpus de consonnes. Ensuite, une étude restreinte au corpus des voyelles est présentée avant d'étendre l'apprentissage à toutes les parties vocaliques des corpus. Enfin, nous nous intéressons à la détermination de la position du velum lors de la production de consonnes.

### 4.1 Analyse globale

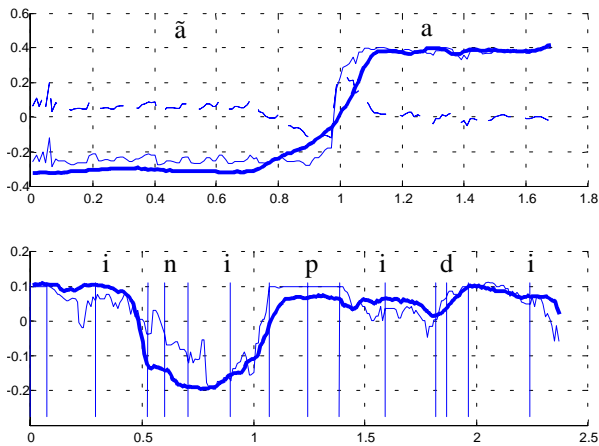
L'apprentissage est effectué « en aveugle » sur 100 000 trames provenant des deux corpus. La variation maximale de la « hauteur » du velum est de 1,17 cm. 11 intervalles égaux de largeur 0.1 cm sont utilisés avec plus de 1000

trames par intervalle. L'erreur quadratique moyenne (RMS) entre la « hauteur » déterminée par la méthode du maximum de vraisemblance  $VH_{est}$  et la « hauteur » mesurée VH est de 0.12 cm pour l'ensemble du corpus de test. Cette valeur, si elle donne une idée globale de l'estimation, ne permet pas de d'établir un lien entre qualité de la détermination de la position du velum et phonème. Les erreurs RMS entre la détermination de  $VH_{est}$  et VH sont présentées de façon détaillée dans le tableau 1. Dans l'ensemble, le paramètre VH est déterminé avec une précision correcte. Cependant, le paramètre  $VH_{est}$  est déterminé avec une erreur RMS importante (0,22 cm) pour les transitions des voyelles nasales vers leur contrepartie orale.

### 4.2 Voyelles nasales et parties vocaliques

Un premier apprentissage a été réalisé sur un corpus limité contenant les quatre voyelles nasales, leurs contreparties orales et les transitions, soit environ 38 000 trames de signal acoustique. L'erreur RMS est de 0.08 cm pour les voyelles nasales et leur contrepartie orale. Pour les transitions, l'erreur RMS de détermination de la position du velum est de 0,14 cm. Cependant, le coefficient de corrélation moyen entre  $VH_{est}$  et VH pour l'ensemble des transitions est de 0,9. L'évolution est donc relativement bien respectée.

Les voyelles /a/ ,/i/ et /u/ en contexte nasal et non nasal (30 000 trames) ont ensuite été intégrées dans la base d'apprentissage. Les erreurs d'estimation sont présentées dans le tableau 1. On observe une erreur importante (0,20 cm) pour les transitions et une corrélation plus faible (0,84). Dans l'ensemble, la détermination du paramètre  $VH_{est}$  est moins précise lorsque les parties vocaliques du corpus des consonnes sont introduites dans la base d'apprentissage. En contrepartie, il est possible de déterminer le paramètre  $VH_{est}$  pour les parties vocaliques du corpus des consonnes avec une faible erreur RMS (de l'ordre de 0,07 cm). Seule la voyelle suivant la consonne nasale se démarque avec une erreur de 0,12 cm. D'après les données articulatoires, le velum reste en position basse durant la voyelle qui suit la consonne nasale (voir figures 1 et 3). L'apprentissage pour ces positions de velum se fait principalement grâce aux transitions des voyelles nasales vers leur contrepartie orale. On retrouve donc des erreurs supérieures, plus proches de celle obtenues pour les transitions.



**Figure 3.** Exemples de  $VH_{est}$  et de  $VH$  pour l'item /ãa/ (en haut) et l'item /pinipidi/ (en bas). Le trait fin représente  $VH_{est}$  tandis que le trait épais correspond à  $VH$ . La différence a été tracée (pointillé) pour l'item /ãa/ : les erreurs sont importantes pour la transition.

### 4.3 Consonnes nasales

Un apprentissage spécifique aux consonnes nasales et orales a été réalisé sur 14 000 trames. La « hauteur » du velum a une plage de variation plus faible (0,8 cm) sur la réalisation des consonnes que sur celle des voyelles (1,1 cm). L'erreur RMS commise lors de la détermination de la « hauteur » du velum est de 0,08 cm pour les consonnes nasales /n/ et /m/ et de 0,07 cm pour les consonnes orales /d/ et /b/.

## 5. DISCUSSION ET PERSPECTIVES

En conclusion, la méthode du maximum de vraisemblance permet de déterminer le paramètre  $VH_{est}$  à partir du signal acoustique avec une erreur de 0,12 cm avec un apprentissage global. Un apprentissage plus ciblé permet d'obtenir des erreurs légèrement inférieures. Les modifications du signal acoustique introduites par l'abaissement du voile du palais semblent permettre de « séparer » les vecteurs acoustiques indépendamment de la qualité de la voyelle.

Ces résultats peuvent être comparé à ceux de [Ric99] qui utilise un réseau de neurones pour estimer la position du velum. L'apprentissage se fait sur un corpus de 400 phrases, et Richmond trouve des erreurs comprises entre 0,14 et 0,20 en unité arbitraire (proches de nos centimètres) pour les voyelles et les consonnes nasales étudiées. Les erreurs obtenues dans notre étude sont légèrement inférieures pour un apprentissage effectué sur des corpus ciblés sur les voyelles et les consonnes nasales. On pourra envisager dans la suite une pré-détermination de la nature du segment (voyelle / consonne) de manière à pouvoir ensuite utiliser la méthode du maximum de vraisemblance avec la base d'apprentissage appropriée à

chaque segment. Cela devrait permettre d'améliorer encore la précision des résultats.

Un point critique de la méthode réside dans le découpage en intervalles. Ce découpage induit une discrétisation arbitraire du paramètre  $VH$ . Le nombre d'intervalles est donc un choix important. L'influence de ce découpage sur la détermination de la « hauteur » du velum peut être étudié avec le modèle articuloire et avec les données mesurées.

La détermination de la position du voile du palais à partir du signal de parole a été appliquée pour un seul locuteur homme il s'agit maintenant d'étendre cette étude à plusieurs locuteurs.

## 6. REMERCIEMENTS

Ce travail n'aurait pu se faire sans Pascal Perrier et Christophe Savariaux que nous remercions. Leur aide et leur expérience ont été d'un grand secours pour les mesures articuloires faites avec l'articulographe

## BIBLIOGRAPHIE

- [Abr94] Abry C., Badin P. et Scully C. (1994) "Sound-to-gesture inversion in speech : The Speech Maps approach", *Advanced speech applications* pp. 182-196.
- [Che95] Chen M.Y. (1995) "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers", *J. Acoust. Soc. Am.* 98 (5), pp. 2443- 2453.
- [Fen96] Feng G. et Castelli E. (1996) "Some acoustic features of nasal and nasalised vowels : A target for vowel nasalisation", *J. acoust. Soc. Am.* 99 (6), pp. 3694-3706.
- [Hog96] Hogden J., Löfqvist A., Gracco V., Zlokarnik I. Rubin P. et Saltzman (1996) "Accurate recovery of articulator positions from acoustics : New conclusions based on human data", *J. Acoust. Soc. Am.* 100 (3), pp. 1819-1834.
- [Wre99] Wrench A.A. (1999) "An investigation of sagittal velar movement and its correlation with lip, tongue and jaw movement", *Proc. Int. Conf. on Phon. Sc. San Francisco*, pp.435-438.
- [Ric99] Richmond K. (1999) "Estimating velum height from acoustics during continuous speech", *Eurospeech Budapest*, pp.149-152.
- [Ros98] Rossato S., Feng G. et Laboissière R. (1998) "Recovering gesture from speech signals: a preliminary study for nasal vowels", *Proc. ICSLP, Sydney*.