Vers une inversion acoustico-articulatoire d'un locuteur étranger

Hélène Lachambre, Régine André-Obrecht
IRIT - Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9
lachambre@irit.fr, obrecht@irit.fr

RÉSUMÉ

Nous présentons une extension de notre méthode d'inversion acoustico-articulatoire basée sur des Modèles de Markov Cachés non supervisés. La génération des vecteurs articulatoires est inspirée par l'approche "GMM". Dans le cadre de l'aide à l'apprentissage des langues étrangères, nous étudions le comportement de cette approche dans le cas de données (phonèmes) manquants.

ABSTRACT

Toward an acoustic to articulatory inversion of a foreign speaker

We present an extension of our acoustic-to-articulatory inversion method, based on unsupervised Hidden Markov Models. The articulatory vectors' generation is based on the "GMM" approach. Considering the application of our method to the teaching of foreing languages, we study the performances of this approach in the case of missing data.

MOTS-CLÉS : Inversion acoustico-articulatoire, HMM non supervisé, données manquantes.

KEYWORDS: Acoustic-to-articulatory inversion, unsupervised HMM, missing data.

1 Introduction

L'inversion acoustico-articulatoire consiste à déterminer la forme du conduit bucal à partir d'un enregistrement audio de parole. Il s'agit plus précisément de reconstruire la trajectoire de divers points situés sur la langue, les lèvres et la machoire (et éventuellement le palais) à partir du signal acoustique. Intéressante en tant que telle pour l'étude des processus de production de la parole, l'inversion acoustico-articulatoire a également des applications plus "grand public" : par exemple, la parole augmentée (pour l'aide à la compréhension des mal-entendants) ou encore l'aide à l'apprentissage des langues étrangères (montrer à un apprenant comment il a prononcé un son, et comment il devrait le prononcer).

Deux principales approches sont utilisées dans la littérature, pour l'inversion acoustico-articulatoire : l'approche GMM (Toda *et al.*, 2008; Ben Youssef *et al.*, 2010) (Modèles de Mélanges de Gaussiennes) et l'approche HMM (Modèles de Markov Cachés) (Hiroya et Honda, 2004; Ben Youssef *et al.*, 2009; Zhang et Renals, 2008; Zen *et al.*, 2010). L'approche GMM consiste à modéliser la distribution conjointe des vecteurs acoustiques et articulatoires par un modèle GMM. L'inversion est considérée comme une recherche de données manquantes et est réalisée par mappage, selon divers critères : MMSE (Minimum Mean Square Error) (Toda *et al.*, 2008) ou Maximum de vraisemblance (Toda *et al.*, 2008; Ben Youssef *et al.*, 2010). L'approche HMM

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 169–176,
Grenoble, 4 au 8 juin 2012.

vise à prendre en compte le caractère temporel de la parole, et les conséquences en termes de contraintes tant au niveau acoustique qu'articulatoire. La partie acoustique est alors modélisée par un HMM. (Hiroya et Honda, 2004) propose une régression linéaire entre l'acoustique et l'articulatoire pour modéliser cette dernière. Dans (Ben Youssef *et al.*, 2009; Zhang et Renals, 2008; Zen *et al.*, 2010), la partie articulatoire est modélisée par un HMM appris conjointement à celui de l'acoustique. La phase d'inversion commence toujours par un décodage du signal audio par le HMM acoustique. La séquence d'états (phonèmes, biphones ou triphones) ainsi déterminée est alors convertie en paramètres articulatoires soit par régression linéaire (Hiroya et Honda, 2004), soit à l'aide du HMM articulatoire (Ben Youssef *et al.*, 2009; Zhang et Renals, 2008; Zen *et al.*, 2010). Dans ce dernier cas, l'inversion inclut des modèles de trajectoire (HTS (Zen *et al.*, 2004)), qui prennent en compte la dynamique des vecteurs articulatoires. Selon les travaux, l'apprentissage des modèles est fait en tenant compte des trajectoires (Zen *et al.*, 2010) ou non (Ben Youssef *et al.*, 2009; Zhang et Renals, 2008).

La modélisation par HMM considère l'aspect temporel de la parole, mais nécessite un étiquetage phonétique coûteux. L'approche GMM considère chaque instant indépendamment des autres, mais l'apprentissage se fait de manière non supervisée. L'approche que nous avons déjà proposée (Lachambre *et al.*, 2011) se place à un niveau intermédiaire : la modélisation se fait par des HMMs, afin de tenir compte de l'aspect temporel de la parole. Cependant, l'apprentissage se fait de manière non supervisée. Pour la phase d'inversion, nous avions précédemment proposé deux approches simples, basées sur des combinaisons linéaires d'états. Nous proposons une nouvelle approche, basée sur le Maximum de vraisemblance.

Dans le cadre particulier de l'apprentissage des langues étrangères, le processus complet consiste à imager la parole de l'apprenant dans l'espace articulatoire d'une personne connue parlant la langue cible (il n'est pas envisageable, pour des questions de coût et de confort de l'apprenant, d'acquérir des données articulatoires de l'apprenant.). Deux problèmes principaux se posent alors. Le premier, qui a été abordé récemment (Ben Youssef *et al.*, 2011), est lié au passage de l'acoustique de l'apprenant à l'articulatoire de la cible, alors que seul le modèle acoustico-articulatoire de la cible est connu. Le second réside dans le fait que l'apprenant est susceptible de prononcer des sons inconnus dans la langue cible, sons qu'il faudra malgré tout imager. Nous nous proposons ici d'étudier la capacité de généralisation de notre modèle confronté, pendant la phase d'inversion, à des sons inconnus lors de l'apprentissage.

Après une présentation du corpus utilisé dans la partie 2, nous rappelons l'apprentissage du modèle dans la partie 3.1. Dans la partie 3.2, nous décrivons l'inversion par Maximum de vraisemblance. Enfin, nous étudions ce modèle en contexte de données manquantes dans la partie 4.

2 Corpora

En tant que partenaire du projet ANR ARTIS [1], nous avons accès à la base de donnée développée par le Gipsa-Lab à Grenoble. Ce corpus a déjà été utilisé dans de nombreuses publications sur l'inversion acoustico-articulatoire (Ben Youssef *et al.*, 2010; Lachambre *et al.*, 2011).

Sont présents des prononciations des 34 phonèmes du français : [i y e ɛ ɛ̃ œ œ̃ ə a ɑ̃ u ø o ɔ ɔ̃ p

1. ARTIS : Articulatory inversion from audio-visual speech for augmented speech presentation, ANR-08-EMER-001-02

b m t d s z n f v ʁ l ʃ ʒ k g j ɥ w]. Le corpus est composé de deux répétitions de 224 séquences VCV (Voyelle-Consonne-Voyelle), deux répétitions de 109 mots courts (CVC) français réels, 68 phrases courtes et 20 phrases longues.

Les données articulatoires sont acquises à l'aide d'un ElectroMagnetic Articulographe (EMA), et sont constituées des coordonnées (X,Y) de six capteurs placés dans un plan sagital. Deux capteurs sont positionnés sur les lèvres (inférieure et supérieure), un sur la machoire, et trois sur la langue (devant, au milieu et au fond). Les données audio sont acquises au format WAV. Elles sont représentées par 12 MFCC, l'énergie, et leurs dérivées. Tous ces paramètres sont calculés toutes les 10 ms, il en résulte des données acoustiques et articulatoires synchrones. Le vecteur global sera noté $\boldsymbol{O} = [\boldsymbol{O}^{ac\,T}\boldsymbol{O}^{art\,T}]^T$ avec $\boldsymbol{O}^{ac}$ et $\boldsymbol{O}^{art}$ les vecteurs acoustique et articulatoire.

3 Approche markovienne non supervisée

Notre approche repose sur un modèle de Markov Caché global $\boldsymbol{M}(A, B)$. Ce modèle induit deux sous-modèles $\boldsymbol{M}_{ac}(A, B_{ac})$ et $\boldsymbol{M}_{art}(A, B_{art})$, représentant respectivement les parties acoustique et articulatoire du signal.

Lors de l'étape d'inversion, le signal acoustique est classiquement décodé à l'aide du modèle acoustique $\boldsymbol{M}_{ac}$, résultant en une suite d'états. Cette suite d'état est ensuite transposée dans le modèle articulatoire pour générer les signaux articulatoires (figure 1).

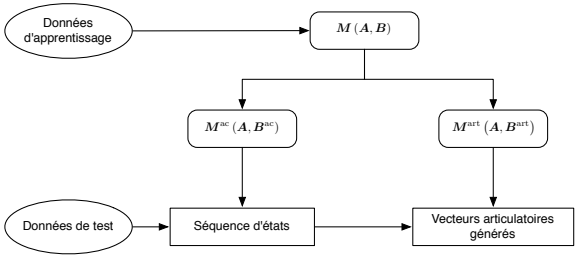


FIGURE 1 – Schéma global de notre méthode.

3.1 Apprentissage

L'apprentissage du modèle global $\boldsymbol{M}$ est réalisé en mode non supervisé (Lachambre *et al.*, 2011). Il se fait en trois étapes :

1. Un clustering non supervisé à l'aide d'un GMM est appliqué à l'ensemble d'apprentissage. Le nombre Q de composantes est fixé *a priori*. Chaque vecteur d'apprentissage est affecté *a posteriori* à une gaussienne et associé par conséquent à un label.
2. Le modèle de Markov global induit se compose d'autant d'états que de clusters. La probabilité d'émission associé à l'état i est modélisée par une loi gaussienne $\mathcal{N}(\mu_i, \Sigma_i)$. Les

paramètres de cette loi sont estimés à l'aide des vecteurs portant le label i.

3. La matrice de transition A est classiquement estimée en comptant le nombre de transitions sur les séquences de labels, associées aux séquences des vecteurs d'apprentissage.

Du modèle global $\boldsymbol{M}$ sont déduits les modèles acoustique et articulatoire $\boldsymbol{M}_{ac}$ et $\boldsymbol{M}_{art}$:
– Le nombre d'états des deux modèles est le même que pour $\boldsymbol{M}$. Chaque vecteur d'apprentissage $\boldsymbol{O}$, portant le label i dans $\boldsymbol{M}$, est séparé en sa partie acoustique $\boldsymbol{O}^{ac}$ et sa partie articulatoire $\boldsymbol{O}^{art}$, chacun assigné à l'état i du modèle correspondant.
– Les matrices de transitions A sont inchangées par rapport à celle de $\boldsymbol{M}$.
– Les probabilités d'émission pour chaque état i de chaque modèle sont des gaussiennes $\mathcal{N}(\mu_i^{ac}, \Sigma_i^{ac})$ et $\mathcal{N}(\mu_i^{art}, \Sigma_i^{art})$, dont les paramètres sont estimés avec les vecteurs portant le label i.

Notons que nous avons les relations suivantes :

$$\mu_i = [\mu_i^{ac\,T}, \mu_i^{art\,T}]^T \qquad \Sigma_i = \begin{bmatrix} \Sigma_i^{ac} & \Sigma_i^{ac,art} \\ \Sigma_i^{art,ac} & \Sigma_i^{art} \end{bmatrix}$$

3.2 Procédure d'inversion

Deux approches peuvent être envisagées résultant d'une résolution de type soit moindres carrés (MMSE) soit maximum de vraisemblance (ML). Compte tenu des modèles de Markov sous jacents, les deux approches prennent en compte partiellement la dimension temporelle ; elles diffèrent par la prise en compte à chaque instant de la correlation entre acoustique et articulatoire.

3.2.1 Inversion MMSE

Lors de la phase d'inversion, le signal acoustique est paramétré en une séquence de K vecteurs $\boldsymbol{O}_1^{ac} ... \boldsymbol{O}_K^{ac}$. Dans une précédente étude (Lachambre *et al.*, 2011), l'approche MMSE a été utilisée pour générer les vecteurs articulatoires correspondants de la manière suivante (avec les notations classiques (Rabiner et Juang, 1993)) :

$$\hat{\boldsymbol{O}}_t^{art} = \sum_{i=1}^{Q} \gamma_t^{ac}(i) \mu_i^{art} \qquad \begin{array}{c} \gamma_t^{ac}(i) = \frac{\alpha_t^{ac}(i)\beta_t^{ac}(i)}{\sum_{i=1}^{Q} \alpha_t^{ac}(i)\beta_t^{ac}(i)} \\ \alpha_t^{ac}(i) = P(\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_t^{ac}, s_t^{ac} = i) \\ \beta_t^{ac}(i) = P(\boldsymbol{O}_{t+1}^{ac}, ..., \boldsymbol{O}_K^{ac} | s_t^{ac} = i) \end{array} \quad (1)$$

3.2.2 Inversion ML

L'approche ML conduit à prendre en compte la correlation instantanée entre les données articulatoires et acoustiques. La comparaison avec les approches classiques basées GMM (cf introduction) montre une différence au niveau de la prise en compte de la dimension temporelle.

$$\hat{\boldsymbol{O}}_t^{art} = \sum_{i=1}^{Q} \gamma_t^{ac}(i) (\mu_i^{art} + \Sigma_i^{ac,art} \Sigma_i^{ac^{-1}} (\boldsymbol{O}_t^{ac} - \mu_i^{ac})) \quad (2)$$

Des expériences précédentes (Lachambre *et al.*, 2011) sur l'approche MMSE ayant montré que ne considérer que le terme prépondérant (l'état le plus probable) dans l'équation 1 donne des résultats équivalents, nous simplifions de la même façon l'approche ML :

$$\hat{\boldsymbol{O}}_t^{art} = \mu_{\hat{s}_t}^{art} + \Sigma_i^{ac,art}\Sigma_i^{ac^{-1}}(\boldsymbol{O}_t^{ac} - \mu_{\hat{s}_t}^{ac}), \quad \hat{s}_t = argmax_{i=1,\ldots,Q}\gamma_t(i) \tag{3}$$

4 Evaluation

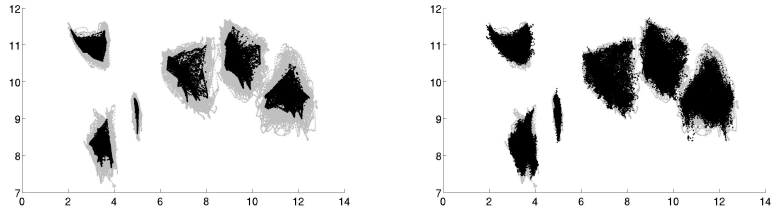
4.1 Evaluation du modèle proposé - Comparaison des méthodes d'inversion

Nous avons montré (Lachambre *et al.*, 2011) qu'un clustering à 128 états est performant pour l'approche MMSE sur ce corpus. Nous avons repris cette valeur pour comparer les performances de l'approche MMSE à l'approche ML. Les résultats quantitatifs (Root Mean Square Error (RMSE) et Pearson Product-Moment Correlation Coefficient) sont présentés dans le tableau 1.

TABLE 1 – Comparaison des approches "MMSE" et "ML" pour l'inversion

Méthode	RMSE	PMCC
MMSE	2.25 mm	0.59
ML	1,83 mm	0.64

Il est clair que l'approche ML est plus perfomante que l'approche MMSE. Une visualisation de la couverture de l'espace articulatoire atteinte lors de l'inversion, pour chacune des deux méthodes, est visible sur la figure 2. La méthode "Maximum de vraisemblance" permet d'atteindre des points beaucoup plus proches de la frontière de l'espace articulatoire, ce qui explique les meilleures performances de l'approche proposée ici.



(a) MMSE (b) ML

FIGURE 2 – Dispersion des vecteurs articulatoires pour les deux méthodes d'inversion (gris : référence, noir : estimation).

4.2 Inversion pour l'apprentissage d'une langue étrangère

Dans le cadre d'inversion acoustico-articulatoire pour l'aide à l'apprentissage des langues, il faut tenir compte du fait que l'apprenant ne connait pas tous les sons de la langue cible, et qu'il va éventuellement ajouter des sons inexistants. Nous étudions ici le comportement de notre modèle, dans le cas où le système doit inverser des sons inconnus lors de l'apprentissage.

Le protocole que nous avons suivi est le suivant :
– L'apprentissage du modèle global et des modèles acoustique et articulatoire sont effectués en enlevant la totalité des prononciations de certains phonèmes de l'ensemble d'apprentissage.
– L'inversion est effectuée sur l'ensemble de test complet : des prononciations de phonèmes inconnus des modèles sont présentes.

Nous avons évalué plusieurs configurations ; pour chacune d'elles, nous préciserons :
– Les phonèmes manquants lors de l'apprentissage,
– La proportion de phonèmes enlevés et restants en terme de durée,
– Le RMSE pour les phonèmes inconnus, les phonèmes connus, et l'ensemble des données.

4.2.1 Les voyelles

En nous basant sur des études phonétiques des langues (Vallée, 1994; Pellegrino, 1998), elles-mêmes basées sur la base de données UPSID [2], il s'avère que 90% des langues utilisent le sous système vocalique [a i u] qui correspondent aux trois configurations extrêmes d'un point de vue articulatoire ; d'autre part, le système vocalique [a e i o u] est le plus représenté dans UPSID. Il apparait donc pertinent d'étudier le comportement de notre système en ne gardant que ces trois ou cinq voyelles dont les résultats apparaissent respectivement dans les tableau 2 et tableau 3.

TABLE 2 – Performances en l'absence de toutes les voyelles centrales - Phonèmes exclus de l'apprentissage : [y e ɛ ɛ̃ œ œ̃ ə ɑ̃ ø o ɔ ɔ̃]

	RMSE	% du temps
Phonèmes manquants	2,51 mm	40 %
Phonèmes connus	1,99 mm	60 %
Tous phonèmes	2,22 mm	100 %

TABLE 3 – Performances en l'absence des voyelles centrales suivantes : [y ɛ ɛ̃ œ œ̃ ə ɑ̃ ø ɔ ɔ̃]

	RMSE	% du temps
Phonèmes manquants	2,49 mm	34 %
Phonèmes connus	1,95 mm	66 %
Tous phonèmes	2,15 mm	100 %

Il est à noter qu'en enlevant 40 % du corpus d'apprentissage qui correspondent à 1/3 des phonèmes, les performances de notre système restent tout à fait honorables avec un RMSE

2. UCLA Phonological Segment Inventory Database

d'environ 2,22 mm sur l'ensemble du corpus de test. Cependant de fortes différences sont à observer selon les phonèmes : parmi les phonèmes inconnus, le mieux reconstruit est le [e] avec un RMSE de 1,88 mm, et le moins bien reconstruit est le [o] avec un RMSE de 3,03 mm. Lors de la comparaison des eux expériences, nous avons noté que les phonèmes bien ou mal reconstruits le sont dans les deux cas.

Nous pensons que l'approche, non supervisée lors de la phase d'apprentissage, permet effectivement une assez bonne capacité de généralisation : le noyau dont dérive chaque loi gaussienne, n'est pas pur en terme de classes phonétiques, puisque, en moyenne, il contient 70 % d'un seul phonème, son identité n'est donc pas une identite phonétique, mais sans doute plus proche d'une configuration articulatoire.

4.2.2 Vers l'inversion du français avec un modèle anglais

Afin de préfigurer l'apprentissage de l'anglais par un français, nous proposons l'expérience suivante : afin d'apprendre un "'modèle d'inversion proche d'un modèle d'inversion pour l'anglais", les phonèmes connus du français, et manquant en anglais sont retirés. Il est évident que dans la réalité, il manquerait des consonnes ou semi consonnes propres à l'anglais. Les résultats sont présentés dans le tableau 4.

TABLE 4 – Performances en l'absence des phonèmes inconnus d'un anglais : [y e ø ɛ̃ œ ɑ̃ o ɔ̃ ʁ ɥ]

	RMSE	% du temps
Phonèmes manquants	2,63 mm	31,5 %
Phonèmes connus	1,94 mm	68,5 %
Tous phonèmes	2,18 mm	100 %

Dans cette expérience, les résultats sont cohérents avec les résultats précédents, en terme de voyelles plus ou moins bien reconstruites. Les deux consonne/semi-consonne que nous avons retirées de l'apprentissage sont parmi les zones les moins bien reconstruits (RMSE de 3,45 mm pour le son [ɥ]).

5 Conclusion et perspectives

Dans cet article, nous avons proposé une nouvelle méthode pour l'inversion acoustico-articulatoire, à mi chemin entre les deux principales approches couramment développées : l'utilisation d'un modèle HMM mais dont l'apprentissage est non supervisé, et la définition de la fonction d'inversion exploitant l'approche ML des GMM. Cette nouvelle proposition améliore les résultats de notre approche.

Par ailleurs, en nous plaçant dans le cadre de l'apprentissage des langues étrangères, nous avons proposé une première étude du cas de phonèmes manquants lors de la phase d'apprentissage. Les premiers résultats sont à la fois encourageants et conformes à nos prévisions : plus le nombre de phonèmes enlevé est important, plus la tâche d'inversion est difficile ; les consonnes inconnues sont plus difficiles à reconstruire que les voyelles inconnues.

Pour la suite, nous allons étudier plus avant les performances de l'inversion pour chacun des phonèmes et nous étudierons à titre de comparaison les performances des approches classiques (HMM, GMM) en l'absence de données manquantes.

Remerciements

Les auteurs remercient le Gipsa-Lab à Grenoble, pour le partage du corpus ARTIS et les nombreux échanges scientifiques sur ce sujet.

Références

Ben Youssef, A., Badin, P. et Bailly, G. (2010). Acoustic-to-articulatory inversion in speech based on statistical models. *In 9th International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 160–165.

Ben Youssef, A., Badin, P., Bailly, G. et Heracleous, P. (2009). Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme Hidden Markov Models. *In Interspeech - European Conference on Speech Communication and Technology*, pages 2255–2258.

Ben Youssef, A., Hueber, T., Badin, P. et Bailly, G. (2011). Toward a multi-speaker visual articulatory feedback system. *In 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 589–592.

Hiroya, S. et Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Audio, Speech, and Language Processing*, 12(2):175–185.

Lachambre, H., Koenig, L. et André-Obrecht, R. (2011). Articulatory parameter generation using unsupervised hidden markov models. *In European Signal Processing Conference (EUSIPCO)*, pages 456–459.

Pellegrino, F. (1998). *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques*. Thèse de doctorat, Université Paul Sabatier, Toulouse.

Rabiner, L. et Juang, B.-H. (1993). *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA.

Toda, T., Black, A. W. et Tokuda, K. (2008). Statistical Mapping between Articulatory Movements and Acoustic Spectrum Using a Gaussian Mixture Model. *Speech Communication*, 50:215–227.

Vallée, N. (1994). *Systèmes vocaliques : de la typologie aux prédictions*. Thèse de doctorat, Université Stendhal, Grenoble.

Zen, H., Nankaku, Y. et Tokuda, K. (2010). Continuous stochastic feature mapping based on trajectory hmms. *IEEE Transaction on Audio, Speech, and Language Processing*, 19(2):417–430.

Zen, H., Tokuda, K. et Kitamura, T. (2004). An introduction of trajectory model into HMM-based speech synthesis. *In Fifth ISCA ITRW on Speech Synthesis*.

Zhang, L. et Renals, S. (2008). Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, 15:245–258.