Apprentissage non-supervisé de Modèles de Markov Cachés Application à l'inversion acoustico-articulatoire

Lionel KOENIG, Hélène LACHAMBRE, Régine ANDRÉ-OBRECHT

IRIT – Université de Toulouse
118 route de Narbonne, 31062 Toulouse Cedex 9, France
{koenig, lachambre, obrecht}@irit.fr

Résumé – Nous présentons une méthode d'inversion acoustico-articulatoire basée sur des Modèles de Markov Cachés (HMM) non supervisés. Un modèle global est tout d'abord appris, à l'aide de l'ensemble des données (acoustiques et articulatoires). Nous en déduisons deux sous-modèles, représentant la partie acoustique et la partie articulatoire des données. Le processus de génération des données articulatoire se fait en deux étapes. Le signal acoustique est tout d'abord décodé à l'aide du HMM acoustique. La suite d'états ainsi déterminée est ensuite transposée dans le modèle articulatoire. Pour cette deuxième étape, deux approches sont étudiées. Notre méthode est testée sur deux corpus : ARTIS et MOCHA-TIMIT. L'erreur RMS est de 2.25 mm pour ARTIS, et 2.45 mm et 2.22 mm pour MOCHA-TIMIT.

Abstract – We present an acoustic-to-articulatory inversion method, based on unsupervised Hidden Markov Models. A global HMM is first trained with all the data (acoustic and articulatory). From this global model, we deduce two submodels, representing the acoustic part and the articulatory part of the data. The generation process is done in two steps. First the acoustic signal is decoded with the acoustic model. The sequence found is then transposed in the articulatory model. For this step, two approches are studied. We test our method on two corpus: ARTIS and MOCHA-TIMIT. The RMS error is 2.25 mm on ARTIS, and 2.45 mm and 2.22 mm on MOCHA-TIMIT.

1 Introduction

L'inversion acoustico-articulatoire a pour but de déterminer la forme de la cavité bucale d'une personne en fonction des sons qu'elle prononce. Utile pour l'étude des processus de production de la parole, elle peut également avoir des applications plus courantes : la parole augmentée (une aide pour les personnes mal-entendantes), ou encore l'apprentissage des langues étrangères (en montrant à un appremnant quelle a été sa prononciation, et celle qu'il aurait dû produire).

Pour modéliser le lien entre l'acoustique et l'articulatoire, les deux approches prédominantes sont l'approche par Modèles de Mélange de Gaussiennes (GMM) [1, 2], et celle par Modèles de Markov Cachés (HMM) [3, 4, 5].

L'approche GMM modélise la distribution conjointe des vecteurs acoustiques et articulatoires par un Mélange de Modèles de Gaussiennes. L'inversion est réalisée par mappage, selon divers critères : le critère MMSE (Minimum Mean Square Error) [1] ou le critère du maximum de vraisemblance [1, 2].

L'approche HMM a pour but de prendre en compte le caractère temporel de la parole, tant dans le domaine acoustique que dans le domaine articulatoire. La partie acoustique est modélisée par un HMM, classiquement appris sur un corpus étiqueté en phonèmes. Plusieurs approches ont été proposées pour la modélisation de la partie articulatoire : [3] propose, pour chaque état du HMM, une régression linéaire entre l'acoustique et l'articulatoire. Dans [4, 5], le modèle de l'articulatoire est appris conjointement à celui de l'acoustique par le biais d'un "HMM multistream". Ce HMM est ensuite décomposé en un HMM acoustique et un HMM articulatoire. Dans tous les cas [3, 4, 5], l'étape d'inversion commence par le décodage du signal acoustique par le HMM acoustique. La séquence d'états ainsi déterminée est alors convertie en paramètres articulatoires soit par régression linéaire [3], soit à l'aide du HMM articulatoire [4, 5]. Dans ce dernier cas [4, 5], la génération des paramètres articulatoires est faite à l'aide des modèles de trajectoire proposés par l'outil HTS [6].

Notre approche se veut à la jonction des deux précédentes. Nous utilisons une modélisation par HMM, afin de tenir compte de l'aspect temporel de la parole. Cependant, l'apprentissage de ce HMM se fait de façon non supervisée, ce qui nous permet de nous abstraire de l'étiquettage (coûteux) des données.

La suite de l'article est organisée comme suit : dans la partie 2, nous présentons les données sur lesquelles nous avons travaillé. Puis, nous proposons notre approche dans la partie 3, et les expériences que nous avons menées dans la partie 4.

2 Corpus

La collecte de corpora assez conséquents et synchronisés entre l'acoustique et l'articulatoire est une difficulté majeure. En tant que partenaires du projet ANR ARTIS [7], nous avons accès à la base de données développée par le Gipsa-Lab. Cette base de données a déjà servi dans de nombreuses expériences [2, 8, 9], une description complète peut en être trouvée dans ces

mêmes articles. Dans ce travail, nous avons également travaillé sur le corpus MOCHA-TIMIT[1], libre d'accès. Dans la suite de cet article, ces deux corpora seront nommés : "corpus ARTIS" et "corpus MOCHA-TIMIT".

Prononcé en français par un locuteur masculin, le corpus ARTIS contient des données acoutiques et articulatoires synchones. Le corpus MOCHA-TIMIT est proposé par le CSTR de l'université Queen Margaret à Edingbourg. Il est prononcé en anglais par plusieurs locuteurs (hommes et femmes), en anglais. Les données acoustiques et articulatoires sont également synchrones. Dans notre étude, nous avons utilisé les données de 2 locuteurs (un homme et une femme).

Les données articulatoires sont enregistrées au moyen d'un dispositif EMA (ElectoMagnetic Articulograph) et représentent la position de capteurs, connus chacun par deux coordonnées. Dans le corpus ARTIS, les capteurs sont au nombre de six : deux sont positionnés sur les lèvres, trois sur la langue et un sur la machoire. Le vecteur articulatoire est de taille 12. Dans le corpus MOCHA-TIMIT, il y a neuf capteurs : deux sont placés sur les lèvres, deux sur les incisives, quatre sur la langue et un sur le palais. Le vecteur articulatoire est de taille 18.

Les données acoustiques sont, pour les deux corpus, classiquement paramétrées : 12 MFCC, l'énergie, accompagnés de leurs dérivées.

Le vecteur global sera noté $\boldsymbol{O} = [\boldsymbol{O}^{ac\,T}\boldsymbol{O}^{art\,T}]^T$ avec $\boldsymbol{O}^{ac}$ le vecteur acoustique (de taille 26) et $\boldsymbol{O}^{art}$ le vecteur articulatoire (de taille 24 ou 36).

3 Méthode

Notre méthode est résumée sur la figure 1.

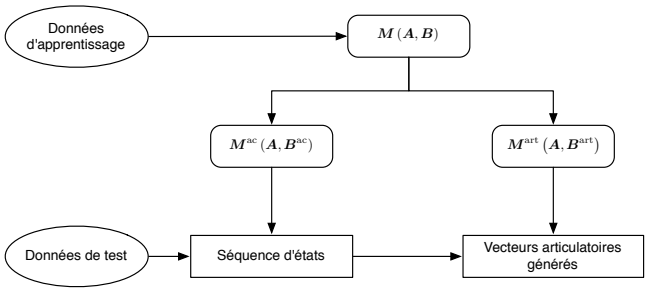

FIG. 1 – Schéma global de notre méthode.

Nous proposons d'apprendre de manière non supervisée un modèle global $\boldsymbol{M}(A, B)$ (noté $\boldsymbol{M}$). Ce modèle est ensuite séparé en deux «sous-modèles» $\boldsymbol{M}_{ac}(A, B_{ac})$ et $\boldsymbol{M}_{art}(A, B_{art})$ (notés $\boldsymbol{M}_{ac}$ et $\boldsymbol{M}_{art}$), représentant les parties acoustique et articulatoire du modèle.

Lors de la génération des vecteurs articulatoires, le signal acoustique est reconnu par le modèle $\boldsymbol{M}_{ac}$. Cette première étape nous donne une suite d'états, suite qui est alors transposée dans le modèle articulatoire.

[1] http ://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

3.1 Apprentissage non supervisé des modèles

3.1.1 Modèle global

Nous proposons l'apprentissage du modèle HMM global $\boldsymbol{M}$ en trois étapes :

- Clustering non supervisé des vecteurs d'apprentissages en Q classes (Q est fixé *a priori*), par exemple à l'aide de l'algorithme des K-means. Un cluster étant assimilé à un état du HMM, chaque vecteur d'apprentissage obtient *a posteriori* un label.
- La probabilité d'émission b_i de l'état i est modélisée par une gaussienne $\mathcal{N}(\mu_i, \Sigma_i)$. Les paramètres de cette loi sont estimés avec les vecteurs d'apprentissages assignés à cet état.
- La matrice de transition A est classiquement estimée en comptant le nombre de transitions entre états parmi les vecteurs d'apprentissage.

3.1.2 Sous-modèles acoustique et articulatoire

Les deux sous-modèles $\boldsymbol{M}_{ac}$ et $\boldsymbol{M}_{art}$ sont déduits de $\boldsymbol{M}$ de la façon suivante :

- Le nombre d'état Q est le même que pour $\boldsymbol{M}$. Chaque vecteur d'apprentissage $\boldsymbol{O}$, assigné à l'état i dans $\boldsymbol{M}$ est séparé en $\boldsymbol{O}^{ac}$ et $\boldsymbol{O}^{art}$. Ces deux sous-vecteurs sont assignés au même état i dans leurs modèles respectifs.
- La matrice de transition A reste exactement la même pour $\boldsymbol{M}_{ac}$ et $\boldsymbol{M}_{art}$ que pour $\boldsymbol{M}$.
- Les lois d'émissions b_i^{ac} et b_i^{art} des états restent des gaussiennes ($\mathcal{N}(\mu_i^{ac}, \Sigma_i^{ac})$ et $\mathcal{N}(\mu_i^{art}, \Sigma_i^{art})$), et sont estimées avec les sous-vecteurs d'apprentissage assignés à chaque état.

Notons que les lois b_i^{ac} et b_i^{art} peuvent être déduites immédiatement de la loi b_i de l'état i du modèle global. En notant $b_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, nous avons :

$$\mu_i = [\mu_i^{ac\,T}, \mu_i^{art\,T}]^T$$

et

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{ac} & \Sigma_i^{ac,art} \\ \Sigma_i^{art,ac} & \Sigma_i^{art} \end{bmatrix}$$

3.2 Génération des vecteurs articulatoires

Deux approches (extrèmement simples) sont proposées pour la génération des vecteurs articulatoires. Le signal acoustique est tout d'abord paramétré, ce qui nous donne une séquence de K vecteurs $\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_K^{ac}$.

3.2.1 Génération par GMM

La première méthode s'inspire de l'approche proposée par Rødbro *et al.* [10]. La loi d'observation des vecteurs générés $\boldsymbol{O}^{art}$ est supposée être un Mélange de Lois Gaussiennes (GMM) (avec b_i^{art} définit ci-avant) :

$$\hat{\boldsymbol{O}}_t^{art^{GMM}} \sim \sum_{i=1}^{Q} P(s_t^{ac} = i|\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_K^{ac}) b_i^{art} \quad (1)$$

Ou encore, avec les notations classiques [11] :

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

$$\begin{aligned} \gamma_t^{ac}(i) &= \frac{\alpha_t^{ac}(i)\beta_t^{ac}(i)}{\sum_{i=1}^{Q} \alpha_t^{ac}(i)\beta_t^{ac}(i)} \\ \alpha_t^{ac}(i) &= P(\boldsymbol{O}_1^{ac}, ..., \boldsymbol{O}_t^{ac}, s_t^{ac} = i) \\ \beta_t^{ac}(i) &= P(\boldsymbol{O}_{t+1}^{ac}, ..., \boldsymbol{O}_K^{ac}|s_t^{ac} = i) \end{aligned} \quad (3)$$

3.2.2 Génération par le meilleur état (BS)

Nous proposons alternativement une simplification de l'équation 2 en ne prenant que le terme prépondérant dans la somme. Nous ne considérons donc que l'état le plus probable :

$$\begin{aligned} \hat{\boldsymbol{O}}_t^{art^{BS}} &= \mu_{\hat{s}_t} \\ \hat{s}_t &= argmax_{i=1,...,Q} \gamma_t(i) \end{aligned} \quad (4)$$

4 Expériences

Pour nos expériences, les données de chacun des locuteurs de chacun des corpus ont été séparées en deux : 2/3 des données pour l'apprentissage des modèles du locuteur, 1/3 pour les tests.

4.1 Configurations testées

Pour l'étape de clustering dans l'apprentissage des modèles, nous testons l'algorithme des K-means (32, 64 et 128 classes), et une approche GMM (128 composantes). Pour l'étape de génération, nous avons testé le meilleur état (BS) et l'approche GMM.

Les configurations testées sont les suivantes :

- $\boldsymbol{M}_{32}^{K_{means}}$-BS : Clustering avec un K-means à 32 classes, génération par le meilleur état.
- $\boldsymbol{M}_{128}^{K_{means}}$-BS : Clustering avec un K-means à 128 classes, génération par le meilleur état.
- $\boldsymbol{M}_{128}^{K_{means}}$-GMM : Clustering à l'aide d'un K-means à 128 classes, génération par l'approche GMM.
- $\boldsymbol{M}_{128}^{EM}$-GMM : Clustering avec un GMM à 128 composantes, génération par l'approche GMM.
- $\boldsymbol{M}_{64}^{EM}$-GMM : Clustering avec un GMM à 64 composantes, génération par l'approche GMM.

Pour la recherche de la meilleure configuration, les expériences ont été menées sur le corpus ARTIS. La meilleure configuration a ensuite été testée sur le corpus MOCHA-TIMIT.

Les résultats sont classiquement donnés en terme de Root Mean Square Error (RMSE) et Pearson Product-Moment Correlation Coefficient (PMCC).

4.2 Structure du modèle non supervisé

Lors de l'apprentissage supervisé d'un HMM, la topologie du modèle est très contrainte, de même que la matrice de transition : de nombreuses transitions entre états sont impossibles.

Nous avons observé que les modèles appris de façon non supervisés sont similaires, dans le sens où de nombreuses transitions entre états sont non significatives. Ainsi, sur la figure 2, est présentée la structure d'un HMM à 32 états appris sur le corpus ARTIS. Seules les transitions de probabilité supérieure à 10^{-2} (relativement peu nombreuses) sont représentées.

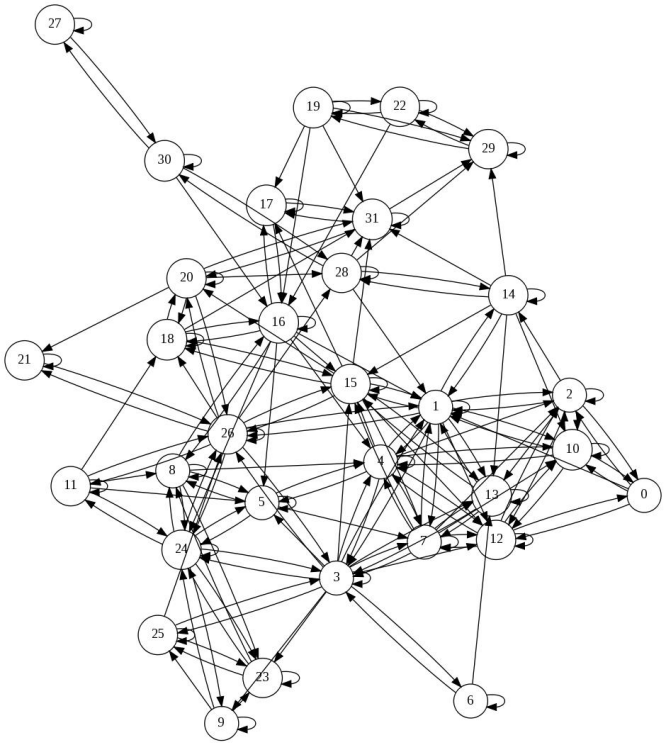

FIG. 2 – Visualisation de la topologie d'un modèle HMM à 32 états appris de façon non supervisée sur le coprus ARTIS. Seules les transitions dont la probabilité est supérieure à 10^{-2} sont représentées.

4.3 Données ARTIS

Les résultats obtenus sur le corpus ARTIS avec chacune des configurations proposées sont présentés dans le tableau 1.

TAB. 1 – RMSE (en mm) et PMCC sur le corpus ARTIS, pour chaque configuration.

	RMSE	PMCC
$\boldsymbol{M}_{32}^{K_{means}}$-BS	2.78	0.50
$\boldsymbol{M}_{128}^{K_{means}}$-BS	2.48	0.55
$\boldsymbol{M}_{128}^{K_{means}}$-GMM	2.47	0.56
$\boldsymbol{M}_{128}^{EM}$-GMM	2.25	0.59

Nous notons une nette amélioration en passant de 32 classes à 128 classes lors du clustering. Des tests ont montré qu'en appliquant l'algorithme du K-means avec 256 classes, certaines classes sont vides. Nous avons donc gardé ce nombre de 128 classes pour les autres expériences. Remarquons que lors de l'apprentissage supervisé d'un HMM, les 35 phonèmes du français sont chacun modélisé par 3 états, soit au total 35*5=105 états, nombre proche des 128 états que nous utilisons.

La réestimation des états donnés par l'algorithme du K-means à l'aide d'un GMM permet également d'améliorer très sensiblement les résultats.

Ces résultats nous semblent très encourageants, puisqu'aucun effort n'a pour l'instant été mis sur la phase de génération.

4.4 Données MOCHA-TIMIT

Les expériences menées sur le corpus ARTIS nous ont permis de déterminer la meilleure configuration : clustering par un GMM, inversion par l'approche GMM.

Le corpus MOCHA-TIMIT ne contenant pas assez de données pour peupler 128 états, nous avons réduit le nombre d'états du HMM à 64. Les résultats ainsi obtenus sont présentés dans le tableau 2. À titre de comparaison, la même configuration est proposée pour le corpus ARTIS.

TAB. 2 – RMSE (en mm) et PMCC sur le corpus MOCHA-TIMIT, pour la meilleure configuration : M_{64}^{EM}-GMM.

	RMSE	PMCC
Homme	2.22	0.46
Femme	2.45	0.48
ARTIS	2.46	0.54

Les performances obtenues sur le corpus MOCHA-TIMIT sont comparables, à modèle équivalent, à celles obtenues sur le corpus ARTIS. Ainsi, notre méthode peut être étendue à d'autres locuteurs, et à d'autres langues.

5 Conclusion et perspectives

Nous avons présenté une nouvelle approche pour l'inversion acoustico-articulatoire, basée sur des Modèles de Markov Cachés. L'originalité de cette méthode repose sur l'apprentissage des modèles, qui se fait de façon non supervisée, sans structure *a priori*, et sans données annotées.

Les résultats nous semblent satisfaisants, puisque indépendants du corpus et proches de ceux de la littérature, compte tenu du fait qu'aucun effort n'a pour l'instant été mis sur la phase d'inversion.

Plusieurs pistes sont à explorer pour améliorer cette méthode. Tout d'abord, une réestimation des modèles par l'algorithme de Baum-Welch permettrait de mieux prendre en compte le contexte. Ensuite, la phase d'inversion doit être améliorée, notamment en ajoutant un modèle de trajectoire.

6 Remerciements

Les auteurs remercient le Gipsa-Lab à Grenoble, pour le partage du corpus ARTIS. Ce travail a été réalisé dans le cadre du projet ANR ARTIS, sous le numéro ANR-08-EMER-001-02.

Références

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

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

[3] S. Hiroya and M. Honda. 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, 2004.

[4] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous. 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, 2009.

[5] L. Zhang and S. Renals. Acoustic-articulatory modeling with the trajectory HMM. *IEEE Signal Processing Letters*, 15 :245–258, 2008.

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

[7] French ANR project. *ARTIS : Articulatory inversion from audio-visual speech for augmented speech presentation*, 2008-2012.

[8] A. Ben Youssef, P. Badin, and G. Bailly. Can tongue be recovered from face ? The answer of data-driven statistical models. In *Interspeech - European Conference on Speech Communication and Technology*, pages 2002–2005, 2010.

[9] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly. Can you "read tongue movements" ? In *Interspeech - European Conference on Speech Communication and Technology*, pages 2635–2638, 2008.

[10] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen. Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5) :1609–1623, 2006.

[11] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA, 1993.