

Méthodes basées sur les HMMs et les GMMs pour l'inversion acoustico-articulatoire en parole

Atef Ben Youssef, Viet-Anh Tran, Pierre Badin, Gérard Bailly

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Université de Grenoble
961 rue de la Houille Blanche, BP 46, F-38402 Saint Martin d'Hères cedex, France
{Atef.Ben-Youssef, Viet-Anh.Tran, Pierre.Badin, Gerard.Bailly}@gipsa-lab.grenoble-inp.fr

ABSTRACT

Two speech inversion methods are implemented and compared. In the first, multistream Hidden Markov Models (HMMs) of phonemes are jointly trained from synchronous streams of articulatory data acquired by EMA and speech spectral parameters; an acoustic recognition system uses the acoustic part of the HMMs to deliver a phoneme chain and the states durations; this information is then used by a trajectory formation procedure based on the articulatory part of the HMMs to resynthesise the articulatory data. In the second, Gaussian Mixture Models (GMMs) are trained on these streams to associate directly articulatory frames with acoustic frames in context. Over a corpus of 17 minutes uttered by a French speaker, the RMS error was 1,66 mm with the HMMs and 2,25 mm with the GMMs.

Keywords: speech inversion, EMA, GMM, HMM

1. INTRODUCTION

L'inversion en parole a été longtemps basée sur l'analyse par la synthèse. Mais depuis une décade, des techniques d'apprentissage plus sophistiquées sont apparues, grâce à la disponibilité de corpus importants de données articulatoires et acoustiques produites par des dispositifs tels que l'Articulographe Electro-Magnétique (EMA) ou des dispositifs de suivi de marqueurs basés sur la vidéo classique ou infrarouge.

Au moins deux classes de modèles statistiques de production de parole se trouvent dans la littérature récente: les modèles de Markov cachés (HMMs) [1], [2], et les modèles de mélanges de Gaussiennes (GMMs) [3]. En plus de la différence structurelle entre HMMs et GMMs, on peut noter que les HMMs utilisent explicitement les informations phonétiques et l'organisation temporelle, tandis que les GMMs agrègent simplement le comportement multimodal de segments de parole similaires.

Hiroya & Honda [1] ont développé une méthode qui estime les mouvements articulatoires à partir du son à l'aide d'un modèle de production de parole basé sur les HMMs. Pour chaque phone, le modèle comprend un HMM des paramètres articulatoires dépendant du contexte et un associauteur linéaire qui transforme les paramètres articulatoires en spectre de parole pour chacun des états du HMM. Les modèles sont construits à partir d'observations acoustiques et articulatoires simultanées

acquises par EMA. En utilisant le modèle de production, la séquence des états HMM est déterminée en cherchant le maximum de vraisemblance de la séquence de spectres de parole. Les paramètres articulatoires sont ensuite déterminés en cherchant le maximum de l'estimation a posteriori des paramètres articulatoires pour un spectre de parole donné et la séquence des états HMM. L'erreur RMS moyenne obtenue est de 1,73 mm.

Toda et coll. [3] ont décrit une approche statistique à la fois pour le mapping articulatoire vers acoustique et le mapping inverse acoustique vers articulatoire sans information phonétique. Ils modélisent la densité de probabilité conjointe des trames acoustiques et articulatoires en contexte par un modèle GMM entraîné sur une base de données parallèles acoustiques et articulatoires. Ils utilisent deux techniques différentes pour établir le mapping GMM. Avec un critère d'erreur quadratique moyenne minimum (MMSE) sur une fenêtre acoustique de 11 trames et 32 composantes pour le GMM, ils obtiennent des erreurs RMS d'inversion de 1,61 mm pour une locutrice, et de 1,53 mm pour un locuteur. L'utilisation d'une méthode de maximum de vraisemblance (MLE) avec 64 composantes gaussiennes, réduit les erreurs à 1,45 mm pour la locutrice, et à 1,36 mm pour le locuteur.

Les études décrites ci-dessus ne permettent pas de déterminer la méthode d'inversion optimale, puisque les données, les locuteurs et les langues ne sont pas comparables. En outre, les corpus ainsi que les conditions d'apprentissage et de test ne sont pas non plus comparables. Ainsi, l'objectif du présent travail est de comparer, *ceteris paribus*, la méthode HMM utilisée dans [2] avec une méthode GMM similaire à celle de [3] en utilisant les critères MMSE et MLE pour la méthode GMM.

2. DONNEES

Pour cette étude préliminaire, nous avons utilisé un corpus déjà enregistré par un unique locuteur français [4], composé de deux répétitions de 224 séquences VCV, deux répétitions de 109 paires de mots de structure CVC différant par un seul trait, 68 phrases courtes et 20 phrases longues. Les phones sont d'abord étiquetés à partir du signal audio et de la transcription phonétique associée, à l'aide d'une procédure d'alignement forcé basée sur des HMMs. Les étiquettes et les frontières de phones sont

ensuite corrigées manuellement. Les 36 phonèmes sont : [a ε e i y u o ø ɔ œ ã ã̃ ã̄ ã̅ p t k f s ʃ b d g v z ʒ m n ʁ l w ɥ j ə _ _], où _ et __ sont respectivement les pauses internes courtes et les pauses longues en début et fin de phrase. Au total, le corpus, dont les longues pauses ont été exclues, contient approximativement 100.000 trames (~17 mn) correspondant à 5132 allophones.

Les données articulatoires ont été acquises à l'aide d'un Articulographe Electro-Magnétique (EMA) qui permet de suivre dans le plan médiosagittal des points cutanés des articulateurs à l'aide de petites bobines électromagnétiques. Pour cette étude, six bobines ont été utilisées: l'une attachée aux incisives inférieures, trois autres attachées à la pointe, au milieu, et à l'arrière de la langue, et les deux dernières attachées à la limite entre la peau et le vermillon des lèvres supérieure et inférieure. Le signal de parole a été enregistré à 22050 Hz, de manière synchrone avec les coordonnées des bobines EMA enregistrées à 500 Hz, et filtrées passe-bas à 20 Hz afin de réduire le bruit.

3. MODELES HMM

Nous rappelons les expériences menées en [2]. Pour l'apprentissage des HMMs, les vecteurs de traits acoustiques sont composés de 12 coefficients cepstraux en échelle Mel (MFCC) et du logarithme de l'énergie, estimés à partir du signal sous échantillonné à 16 kHz sur des fenêtres de 25 ms à une fréquence de trame de 100 Hz ; ces vecteurs sont complétés par les dérivées premières temporelles. Les vecteurs de traits articulatoires sont composés des 12 coordonnées x et y des six bobines actives, ainsi que leurs dérivées premières. Les trajectoires EMA sont sous-échantillonnées à 100 Hz pour être synchrones avec les vecteurs acoustiques.

Différentes variantes ont été testées: phonèmes sans contexte (*no-ctx*), avec contexte gauche (*L-ctx*) ou droit (*ctx-R*), et avec contextes gauche et droite (*L-ctx-R*). Une méthode de regroupement hiérarchique basée sur la matrice des distances de Manhattan entre les coordonnées des bobines pour chaque paire de phonèmes, a permis de définir six classes cohérentes pour les contextes vocaliques ([a ε ã̃ | ø œ ã̃̄ | e i | y | u | o ɔ ã̃̅]) et dix classes pour les contextes consonantiques ([p b m | f v | ʁ | ʃ ʒ | l | t d s z n | j | ɥ | k g | w]). Le schwa, et les pauses courtes et longues ([ə _ _]) ne sont pas pris en compte comme contextes.

Nous avons utilisé des modèles HMM gauche-droite à trois états, avec une gaussienne par état et une matrice de covariance diagonale. Les procédures d'apprentissage et de test ont été réalisées avec la boîte à outils HTK [5]. Pour l'apprentissage, le critère de maximum de vraisemblance (ML) était implémenté sous forme de maximisation de l'espérance (EM). Les vecteurs de traits acoustiques et articulatoires sont considérés comme deux flux dans la procédure multi-flux de HTK. Les modèles

HMMs obtenus sont ensuite séparés en *HMMs articulatoires* et *HMMs acoustiques*.

Un modèle de langage bigramme considérant les séquences de phones en contexte est appris sur l'ensemble du corpus. L'inversion est réalisée en deux étapes: la première effectue une reconnaissance phonémique basée sur les HMMs acoustiques, et fournit la séquence des allophones reconnus, ainsi que la durée de chaque état. Une procédure d'héritage permet de remplacer un HMM en contexte manquant dans le corpus d'apprentissage par le HMM le plus proche [2]. La seconde étape resynthétise les trajectoires articulatoires à partir de ces informations à l'aide de la procédure de formation de trajectoire proposée par Zen *et al.* [6].

La méthode est ensuite évaluée par la procédure du *jack-knife*: les données sont séparées en 5 parties approximativement homogènes du point de vue de la répartition des phones ; chaque partie est tour à tour utilisée pour évaluer les performances des modèles HMM appris sur le restant des données. Les performances sont évaluées sur l'ensemble des 5 résultats par (1) la racine carrée des erreurs quadratiques moyennes (RMSE), (2) les coefficients de corrélation (r) entre données mesurées et données estimées, et (3) les taux de reconnaissance et de précision agrégés sur l'ensemble du corpus.

Les taux de reconnaissance / précision obtenus varient entre 88,90 / 68,99 % en l'absence de contexte et la meilleure performance de 93,66 / 80,9 % obtenue pour des phones en contexte droit. La procédure d'héritage de HMMs manquant permet de gagner entre 1 et 5 % sur les performances de reconnaissance. Le modèle de langage pour la reconnaissance permet de passer de taux de reconnaissance / précision de 72,29 / 34,22 % à 93,66 / 80,90 %. Cette amélioration spectaculaire a cependant une faible influence sur les performances puisque l'on passe seulement, en contexte droit, de 1,83 à 1,66 mm pour la RMSE et de 0,90 à 0,92 pour la corrélation. On voit sur la Table 1 que l'utilisation de contextes augmente très sensiblement les performances (sauf pour le contexte droit et gauche pour lequel la reconnaissance est nettement moins bonne, vraisemblablement dû à la taille trop petite du corpus).

Afin d'estimer la contribution du processus de formation de trajectoire à l'erreur RMSE de l'inversion complète, nous avons aussi synthétisé les trajectoires en utilisant un alignement forcé des états basés sur étiquettes originales, émulant ainsi un étage de reconnaissance parfaite (voir

Table 1 : RMSE (mm) et coefficient de corrélation r pour la méthode HMM. (1): avec étape de reconnaissance parfaite ; (2) inversion complète.

	no-ctx		L-ctx		ctx-R		L-ctx-R	
	RMSE	r	RMSE	r	RMSE	r	RMSE	r
(1)	1,91	0,90	1,55	0,93	1,55	0,93	1,40	0,94
(2)	2,07	0,87	1,72	0,91	1,66	0,92	1,91	0,89

Table 1). Le niveau relativement élevé de ces erreurs montre que la majeure partie de l'erreur globale (entre 70 et 90 %) est due à l'étape de formation de trajectoire qui lisse en excès les mouvements prédits et ne capture pas de manière appropriée les patrons de coarticulation.

4. MODELES GMM MULTIMODAUX

Nous avons mis en œuvre une mise en correspondance basée sur les GMMs en utilisant le critère de minimum de l'erreur quadratique moyenne (MMSE), souvent utilisé pour la conversion de voix. En outre, afin d'améliorer la précision de l'inversion, nous avons ajouté une étape d'optimisation basée sur l'estimation du maximum de vraisemblance (MLE) [3]. Les trajectoires de paramètres cibles ayant les propriétés statiques et dynamiques adéquates sont déterminées en combinant les estimations locales de la moyenne et de la variance pour chaque trame $p(t)$ et ses dérivées $\Delta p(t)$ par la relation explicite entre les paramètres statiques et dynamiques (*p. ex.* $\Delta p(t) = p(t) - p(t-1)$). Pour chaque trame, le vecteur de traits est la concaténation d'un vecteur articulatoire de 24 composantes (12 coordonnées EMA et leurs dérivées), et d'un vecteur acoustique de 24 composantes. Afin de prendre en compte le contexte acoustique [3], [7], de 9 à 17 vecteurs acoustiques (12 MFCC, énergie log) sont prélevés de manière équirépartie dans une zone temporelle contextuelle de taille variable, et réduits à 24 composantes par Analyse en Composantes Principales. Nous avons fait varier le nombre de composantes gaussiennes de 8 à 64 et la zone contextuelle d'une taille phonémique (~90 ms) à une taille syllabique (~170 ms). Chaque gaussienne est représentée par une matrice de covariance pleine (48×48), un vecteur de moyennes (48) et son coefficient de pondération.

La Table 2 montre les performances pour les différentes expériences déterminées par la même méthode *jack-knife* sur les mêmes parties. L'erreur quadratique moyenne (RMSE) diminue lorsque le nombre de composantes augmente, et atteint un optimum pour une fenêtre contextuelle de 110 ms. L'explication la plus plausible est qu'une fenêtre de taille de diphone contient de manière optimale les traits phonétiques locaux nécessaires à l'inversion. La meilleure précision d'inversion est

finale obtenue pour une fenêtre de 110 ms et 64 composantes qui semblent constituer la meilleure représentation des 36 phonèmes. Nous avons noté par ailleurs que l'étape supplémentaire d'optimisation par MLE augmente les performances de l'ordre de 5 %.

Table 2 : RMSE (mm) et coefficient de corrélation r pour la méthode GMM en fonction du nombre de Gaussiennes (# mix) et de la taille du contexte ctw (ms).

#mix	8		16		32		64	
ctw	RMSE	r	RMSE	r	RMSE	r	RMSE	r
90	2,68	0,78	2,61	0,80	2,38	0,83	2,32	0,84
110	2,68	0,78	2,54	0,80	2,37	0,83	2,25	0,85
130	2,66	0,78	2,51	0,81	2,36	0,83	2,27	0,85
150	2,66	0,78	2,50	0,81	2,44	0,82	2,32	0,84
170	2,65	0,78	2,44	0,82	2,41	0,82	2,29	0,84

5. COMPARAISONS ET COMMENTAIRES

La Figure 1 compare les trajectoires originales et estimées des ordonnées des bobines EMA pour les systèmes étudiés.

Au vu de la littérature, il est surprenant que nos résultats d'inversion basés sur les HMMs soit significativement plus précis ($p < 10^{-6}$) que ceux basés sur les GMMs (1,66 mm vs. 2,25 mm) : dans les deux expériences les plus abouties, Hiroya & Honda [1] trouvent 1,73 mm avec des HMMs (ce qui est proche de nos résultats) comparé aux 1,36 – 1,45 mm trouvés par Toda et coll. [3] avec des GMMs. Même en prenant en compte le fait que ces deux expériences sont basées sur des sujets et des langues différentes, la différence est telle que nous ne nous attendions pas à de tels résultats. Nous n'avons pour l'instant pas d'explication à cette divergence.

Nos deux systèmes peuvent cependant être améliorés. L'inversion à base de HMMs pourrait inclure un traitement plus sophistiqué de l'asynchronie articulatoire / acoustique en introduisant des modèles de retard qui se sont révélés efficaces pour la synthèse multimodale par HMMs [8]. Le système basé sur les GMMs pourrait être amélioré en considérant d'autres techniques de réduction de la dimensionnalité telles que l'Analyse Discriminante Linéaire (LDA) qui sont assez efficaces pour l'inversion

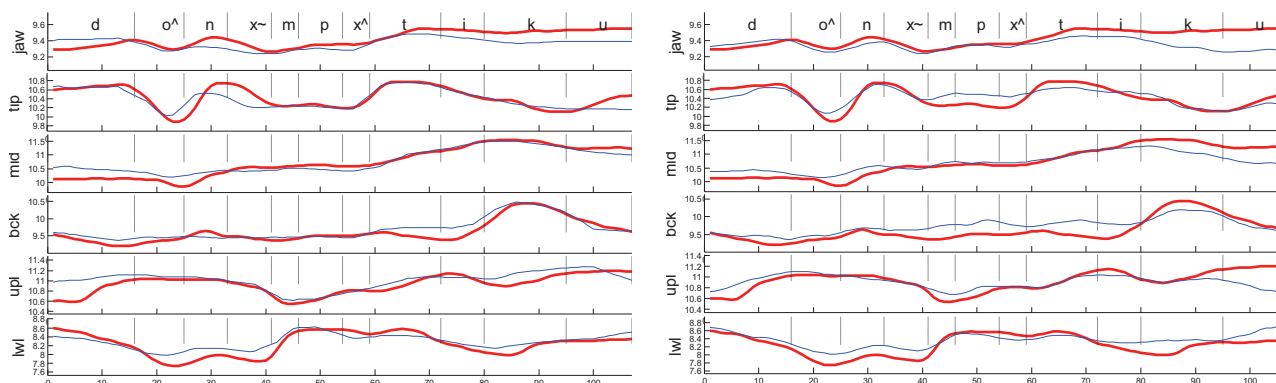


Figure 1 : Comparaison des trajectoires originales (traits épais) et prédites (traits fins) des ordonnées des 6 points de chair EMA. A gauche: inversion et synthèse basée-HMM avec des modèles en contexte droit. A droite: synthèse par GMM avec une fenêtre contextuelle de 110 ms centrée sur la trame courante et un mélange de 64 Gaussiennes.

basée sur les HMMs [7]. Les deux systèmes pourraient aussi gagner à incorporer de l'information visuelle en entrée et à inclure de manière plus intime cette information additionnelle dans le processus d'optimisation qui va considérer la cohérence multimodale entre les paramètres d'entrée et de sortie: les lèvres sont clairement visibles, et la position de la mâchoire est accessible de manière indirecte à partir des mouvements faciaux.

6. CONCLUSIONS ET PERSPECTIVES

Nous avons mis en œuvre et comparé deux techniques d'inversion acoustico-articulatoire en parole qui diffèrent par la façon dont elles capturent et exploitent la cohérence multimodale a priori entre son et articulation. Plusieurs réserves peuvent cependant être faites à propos de ces premières expériences.

Le système basé sur les HMMs bénéficie de la phonotactique du langage cible. Notons que le Français possède un inventaire syllabique riche : nous pouvons ainsi imaginer que les résultats obtenus avec des langues présentant des structures phonologiques très différentes telles que le Japonais, le Polonais ou l'Espagnol présentant des complexités syllabiques diverses pourraient conduire à des résultats différents.

Les mesures objectives globales pourraient ne pas refléter entièrement le comportement spécifique des phones qui peut avoir un impact majeur sur une évaluation subjective de l'articulation générée. La précision de la reconstruction est bien naturellement de la plus haute importance pour l'évaluation, mais d'autres éléments tels que la précision de la récupération d'éléments cruciaux comme les constrictions du conduit vocal sont également très importants.

Nous avons montré [4] que les sujets ont des performances très diverses en *lecture linguale*, et que cette performance augmente avec l'entraînement. Notons ainsi que le réalisme du mouvement pourrait compenser le manque de précision des détails de forme: la cinématique des trajectoires calculées pourrait être plus importante pour la perception que la précision des trajectoires elles-mêmes.

Finalement, les résultats de cette étude vont nous permettre de développer un système de tuteur pour la correction phonétique [9], dans lequel les mouvements articulatoires reconstruits seront utilisés pour piloter une tête parlante virtuelle 3D avec tous les degrés de liberté possibles [10].

7. REMERCIEMENTS

Nous remercions Christophe Savariaux et Coriandre Vilain pour les enregistrements EMA, et Tomoki Toda (NAIST, Japon) pour la mise à disposition du logiciel de GMM. Ce travail a été partiellement financé par le projet ANR-08-EMER-001-02 *ARTIS* et le projet Franco-japonais PHC SAKURA *CASSIS*.

BIBLIOGRAPHIE

- [1] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 175-185, March 2004.
- [2] 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 2009*, Brighton, UK, pp. 2255-2258, 2009.
- [3] T. Toda, A. W. Black, and K. Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, vol. 50, pp. 215-227, 2008/3 2008.
- [4] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly. Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, vol. 52, pp. 493-503, 2010.
- [5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book. Revised for HTK Version 3.4 December 2006, 2006.
- [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 (SSW5)*, Pittsburgh, PA, USA, pp. 191-196, 2004.
- [7] V.-A. Tran, G. Bailly, H. Loevenbruck, and C. Jutten. Improvement to a NAM captured whisper-to-speech system, in *Interspeech*, Brisbane, Australia, pp. 1465-1468, 2008.
- [8] O. Govokhina, G. Bailly, and G. Breton. Learning optimal audiovisual phasing for a HMM-based control model for facial animation, in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.
- [9] P. Badin, G. Bailly, and L.-J. Boë. Towards the use of a Virtual Talking Head and of Speech Mapping tools for pronunciation training, in *Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning*, Stockholm, Sweden, pp. 167-170, 1998.
- [10] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka. An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data, in *Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098)*, F. J. Perales and R. B. Fisher, Eds. Berlin, Heidelberg, Germany: Springer Verlag, pp. 132-143, 2008.