Adjonction de contraintes visuelles pour l'inversion acoustique-articulatoire

Blaise Potard, Yves Laprie

LORIA / Équipe PAROLE
Campus Scientifique - BP 239
54506 VANDŒUVRE-lès-NANCY CEDEX, France
Mél : {Blaise.Potard, Yves.Laprie}@loria.fr
http ://www.loria.fr/equipes/parole/

ABSTRACT

The goal of this work is to investigate audiovisual-to-articulatory inversion. It is well established that acoustic-to-articulatory inversion is an under-determined problem. On the other hand, there is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues : the view of visible articulators, i.e. jaw and lips, improves speech intelligibility. It is thus interesting to add constraints provided by the direct visual observation of the speaker's face. Visible data were obtained by stereo-vision and enable the 3D recovery of jaw and lip movements. These data were processed to fit the nature of parameters of Maeda's articulatory model. Inversion experiments show that constraints on visible articulatory parameters enable relevant articulatory trajectories to be recovered and substantially reduce time required to explore the articulatory codebook.

1. INTRODUCTION

La principale difficulté de l'inversion acoustico-articulatoire est le fait qu'il n'existe pas de relation directe du domaine acoustique vers l'articulatoire : un grand nombre de formes différentes de conduit vocal peuvent produire le même spectre de parole. C'est un problème sous-déterminé, puisqu'il y a plus d'inconnues que de données en entrée. Un des enjeux principaux réside dans l'étude de contraintes qui soient suffisamment restrictives et pertinentes d'un point de vue phonétique, de façon à éliminer des solutions manifestement peu réalistes.

La parole est un signal bimodal, comportant une composante acoustique, et une composante visuelle : la vue du locuteur. Ces deux modalités sont fortement corrélées et redondantes. Il a été observé à de nombreuses reprises que les locuteurs et auditeurs humains exploitent la nature multimodale de la parole, et plus particulièrement les indices articulatoires : l'intelligibilité de la parole augmente dans des conditions difficiles (déficiences auditives, environnement bruyant...) lorsque l'auditeur voit le locuteur[10, 1, 9, 4].

L'objectif du présent travail est d'adjoindre aux données acoustiques classiquement utilisées dans l'inversion acoustique-articulatoire, des données issues de l'observation des articulateurs visibles (lèvres et mâchoire), obtenues par stéréovision.

2. MÉTHODE D'INVERSION

Notre méthode d'inversion est fondée, comme beaucoup d'autres, sur l'analyse par synthèse, et le processus d'inversion comporte trois étapes.

Une étape préalable au processus d'inversion est la construction d'une table articulatoire, ou codebook, qui associe des vecteurs articulatoires (à 7 dimensions, correspondant aux 7 paramètres du modèle de Maeda) à leurs correspondants acoustiques, dans notre cas, le triplet des fréquences des 3 premiers formants. La force de notre méthode d'inversion réside dans la résolution acoustique quasi uniforme du codebook. Cette propriété est garantie par la façon dont est construite la table : on explore l'espace articulatoire récursivement en évaluant à chaque étape la linéarité locale de la relation articulatoire-acoustique [7]. Cette table est organisée de manière à retrouver facilement tous les vecteurs articulatoires qui permettent de générer un triplet de formants donné et est propre à chaque locuteur : sa construction nécessite une adaptation préalable du modèle articulatoire.

La première étape du processus d'inversion proprement dit consiste à générer un grand nombre de solutions potentielles à partir du codebook. Comme il existe *a priori* une infinité de vecteurs articulatoires permettant d'obtenir un vecteur acoustique il est nécessaire d'échantillonner l'espace des solutions de façon suffisamment concise mais précise pour trouver des solutions proches de la solution réelle.

La deuxième étape de notre méthode consiste à reconstruire une trajectoire articulatoire qui soit suffisamment régulière au cours du temps. Nous utilisons pour cela un algorithme de programmation dynamique qui minimise une fonction de coût représentant la « distance » couverte par les articulateurs.

La dernière étape consiste à améliorer la fidélité acoustique et la régularité articulatoire de la solution obtenue à l'étape précédente en utilisant un algorithme de régularisation variationnelle.

2.1. Exploration de l'espace nul de la relation articulatoire acoustique

Pour chaque vecteur acoustique représentée par les trois premières fréquences formantiques, le processus d'inversion consiste en la recherche de tous les hypercubes qui peuvent générer le triplet de formants observé. Il faut ensuite trouver un ensemble de solutions dans chacun de

ces cubes. Comme l'inversion consiste à trouver 7 paramètres à partir de 3, l'espace des solutions a *a priori* 4 degrés de liberté. La relation articulatoire acoustique (notée R) est supposée être localement linéaire au niveau du centre P_0 de l'hypercube (c'est-à-dire que l'application $P - P_0 \mapsto R(P) - R(P_0)$ est supposée être une application linéaire). Trouver l'ensemble des solutions n'est pas un problème trivial car il s'agit de trouver l'intersection d'un espace à 4 dimensions (l'espace nul de la relation précédente, c'est-à-dire l'ensemble des antécédents de 0 pour l'application linéaire) et d'un hypercube à 7 dimensions, ce que l'on ne sait pas faire de manière formelle. Une première approximation de l'intersection est obtenue par programmation linéaire. Puis l'espace nul est échantillonné, et l'appartenance à l'intersection de chacun des points est testée[8].

3. CONTRAINTES VISUELLES

3.1. Acquisition des données

Pour acquérir les données sur les articulateurs visibles nous avons utilisé un système d'acquisition et de suivi de données tridimensionnelles. Ce système a été réalisé par l'équipe de vision par ordinateur de notre laboratoire [11]. L'un de ses intérêts est d'être peu cher et facilement utilisable. Par ailleurs, il est plus flexible que les systèmes de *motion-capture* qui utilisent généralement des caméras infrarouges et des marqueurs collés sur la peau.

Il utilise simplement deux caméras, un PC, et des marqueurs peints qui ne perturbent pas l'articulation ; il permet une acquisition suffisamment rapide pour reconstituer de façon précise les trajectoires des points 3D.

Pour faire une reconstitution des mouvements des articulateurs en stéréovision, il est nécessaire d'être capable de suivre les même points physiques au cours du temps. Comme la peau naturelle n'est pas assez contrastée, nous avons choisi de peindre des marqueurs sur le visage du locuteur. Cette méthode permet de contrôler la taille, la densité et la position des points intéressants. Par exemple, nous avons peints 210 marqueurs sur le visage (46 sur les lèvres) pour permettre d'obtenir une information précise sur la déformation de la forme des lèvres (fig. 1), dans l'optique de construire une tête parlante de bonne qualité.

Dans le cas du corpus utilisé dans cette étude, élaboré dans l'optique d'étudier la variabilité interlocuteurs de la coarticulation labiale, nous avons choisi de peindre seulement 15 marqueurs sur le visage du locuteur (seulement 4 marqueurs sur les lèvres, fig. 2) de façon à conserver un temps de préparation raisonnable pour les sujets de l'étude. En plus des marqueurs utilisés pour étudier les mouvements des lèvres, nous avons placés 6 marqueurs dans la partie supérieure du visage de façon à compenser le mouvement global de la tête. Deux caméras monochromes sont utilisées, car leur vitesse d'acquisition (environ 120 images par seconde) étant plus rapide que celle des caméras couleurs, elles permettent de suivre des mouvements très rapides des articulateurs, par exemple les occlusives.

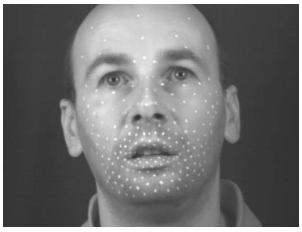

FIG. 1: 210 marqueurs blancs sont peints sur le visage du locuteur.

FIG. 2: Images en stéreovision de deux locuteurs, 15 marqueurs sont peints sur le visage de chaque locuteur.

3.2. Intégrer les données visuelles au modèle de Maeda

Le modèle articulatoire de Maeda[5] a été établi à partir d'images radiographiques de coupes sagittales du conduit vocal en y appliquant une analyse factorielle permettant le choix explicite des composantes linéaires. Les mouvements de la mâchoire, en particulier, peuvent être facilement déterminés en mesurant la position des incisives qui apparaissent très clairement sur les images. L'ouverture latérale des lèvres ne peut par contre pas être évaluée à partir des radiographies, et ce n'est donc pas un des paramètres du modèle.

Les données 3D du visage du locuteur permettent de mesurer directement l'étirement et l'ouverture des lèvres à partir de la position des marqueurs sur les lèvres (voir Fig. 2). La protrusion peut aussi être estimée à partir de ces points, mais comme il s'agit d'un mouvement complexe qui implique un « dépliement » des lèvres, les mouvements de marqueurs peints sur les lèvres dans le plan sagittal ne peuvent rendre compte que partiellement de ce mouvement complexe. Par conséquent, la protrusion est probablement légèrement sous-estimée.

Contrairement aux images radiographiques, les données 3D du visage du locuteur ne permettent pas de mesure précise des mouvements de la mâchoire. En effet, le mouvement des marqueurs peints sur le menton (que nous utilisons pour évaluer les mouvements de la mâchoire) est lié à la mâchoire, mais aussi à celui de la lèvre inférieure qui déplace ces marqueurs quand elle bouge. Par conséquent, le mouvement de la mâchoire n'est pas non plus connu avec précision.

3.3. *Ajustement des données visuelles*

À partir des données visuelles acquises, nous calculons 4 paramètres : l'ouverture de la bouche, l'étirement des lèvres et les mouvements de la mâchoire, ces paramètres se calculant facilement à partir de la position des marqueurs, et la protrusion des lèvres, dont l'évaluation est plus complexe.

- L'ouverture de la bouche est donnée par la distance entre les deux points des lèvres situés dans le plan sagittal.
- L'étirement des lèvres est la distance entre les deux points situés aux commissures des lèvres.
- Le mouvement de la mâchoire est la distance entre les points du menton et un point fixe. Nous prenons la moyenne des positions des 4 points du menton. En faisant cela, nous négligeons de façon implicite l'influence des mouvements des lèvres sur la position de ces points.
- La protrusion des lèvres est plus complexe à calculer. Le paramètre est déterminé en utilisant la distance des points des lèvres inférieure et supérieure à un plan de référence défini par la position moyenne des 4 points des lèvres.

Normalisation Comme notre objectif est d'utiliser les données 3D obtenues avec le système de stéréovision comme des contraintes sur les paramètres régissant les articulateurs visibles du modèle de Maeda, nous devons établir une correspondance entre les paramètres articulatoires du modèle et les paramètres observés que nous venons de définir. Les données géométriques mesurées par Maeda étaient centrées et normalisées avant d'être traitées par analyse factorielle. Chacun des 7 paramètres du modèle articulatoire de Maeda peut ainsi varier dans un intervalle de $\pm 3\sigma$ (où σ est l'écart-type du paramètre). Nous appliquons la même transformation aux paramètres issus des données tridimensionnelles : chacun des paramètres est centré autour de sa position moyenne et réduit.

Décorrélation L'étape de normalisation précédente permet d'obtenir des paramètres observés qui ont les mêmes dimensions que les paramètres articulatoires. Cependant, l'analyse factorielle de Maeda permettait en plus de cela de retirer l'effet de la mâchoire des autres paramètres de façon à obtenir des paramètres indépendants. Nous devons donc retirer l'effet des mouvements de la mâchoire des autres paramètres. De la même façon que Maeda nous calculons la corrélation entre la mâchoire et chacun des deux autres paramètres (l'ouverture et la protrusion des lèvres puisque nous n'utilisons pas l'étirement qui ne peut pas être utilisé dans le modèle) et soustrayons la corrélation des mesures normalisées.

Le principal problème de cette méthode est que, contrairement aux radiographies où le mouvement de la mâchoire est mesurable avec précision, le mouvement de la mâchoire inférieure n'est ici connu que de manière approchée. Cette étape de décorrélation ne permet d'obtenir par conséquent que des approximations de chacun des paramètres.

3.4. *Intégration au processus d'inversion*

Nous obtenons ainsi trois paramètres compatibles avec le modèle de Maeda, mais malheureusement imprécis. Nous devons donc compenser cette imprécision, ce que nous faisons en permettant aux paramètres visuels des solutions de l'inversion de varier dans un domaine assez important.

Pour cela, nous n'utilisons les paramètres observés que lors de la sélection des hypercubes : nous n'échantillonnons que les hypercubes dont les centres ont des paramètres visuels proches des paramètres observées. De cette manière l'imprécision des données visuelles ne se répercute pas directement sur les résultats de l'inversion et, comme nous le verrons plus tard, cette utilisation très simple permet d'accélérer considérablement le processus d'inversion et d'améliorer le réalisme des solutions.

Les paramètres observés ne sont donc pour l'instant utilisés que dans la première étape du processus d'inversion : la génération d'un grand nombre de solutions possibles. Dans la deuxième étape, on construit une trajectoire initiale optimale parmi cet ensemble de solutions en minimisant un critère biomécanique ou de régularité sur les paramètres articulatoires. Dans cette étude, on minimise un critère portant sur la « vitesse globale » des articulateurs à l'aide de l'algorithme de Ney[6]. Dans la troisième étape, on lisse cette trajectoire en utilisant un algorithme de régularisation variationnelle[3] pour améliorer la fidélité acoustique et la régularité de la solution.

4. EXPÉRIENCES

4.1. *Corpus*

La phrase que nous présentons dans cette étude est extraite d'un corpus de données 3D enregistré par 10 locuteurs français natifs (5 hommes et 5 femmes), chacun s'exprimant pendant environ 120 secondes. Ce corpus était principalement destiné à étudier la variabilité interlocuteurs de la coarticulation labiale, et comporte essentiellement des logatomes. Il comporte aussi une phrase « Le joaillier a broyé les cailloux de la voyageuse.» construite spécialement pour faciliter l'inversion, puisque la plupart des sons la composant sont des voyelles, des semi-voyelles ou d'autres sons voisés.

Nous présentons ici les résultats obtenus pour l'un des locuteurs. Le modèle articulatoire a été adapté au locuteur en utilisant la méthode de Galvan-Rodrigez[2]. Bien que nous ayons choisi une précision acoustique assez faible pour la construction du codebook (1 Bark), la précision acoustique moyenne reste très bonne : l'erreur RMS moyenne est d'environ 15 Hz pour F3.

4.2. *Inversion de séquences*

Nous présentons ici les résultats détaillés pour deux parties voisées de la phrase précédente : « joaillier » (/ʒɔɑje/) et « cailloux » (/kaju/) qui présentent des mouvements articulatoires intéressants. Nous ne présentons ici que les résultats non lissés, qui sont plus significatifs des forces et des faiblesses du système actuel.

La figure 3 présente les résultats de l'inversion pour la séquence « joaillier ». Nous affichons les trajectoires trouvées pour les 4 principaux paramètres (mâchoire, ouverture des lèvres, protrusion, position de la langue). Nous affichons aussi en trait pointillé les paramètres observés donnés comme contrainte. Comme on peut le voir sur le graphique de la mâchoire, l'inversion rencontre des difficultés au milieu de la séquence (instant 21400), pour la

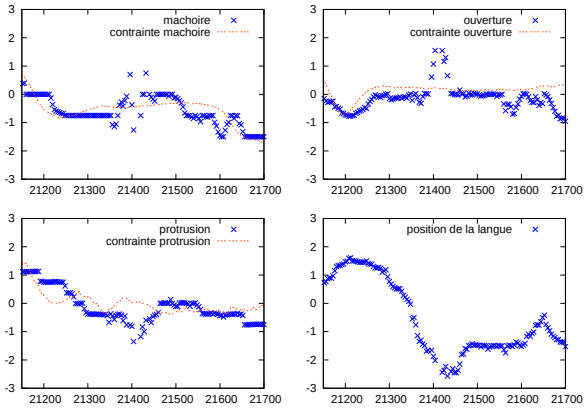

FIG. 3: Résultat de l'inversion pour « joaillier ». Pour chaque graphique, la courbe en trait fin pointillé correspond au paramètre issu des données visuelles, la courbe discontinue à la trajectoire du paramètre correspondant trouvée par le processus d'inversion. L'abscisse représente le temps (en ms), l'ordonnée la valeur du paramètre qui peut varier entre -3 et 3.

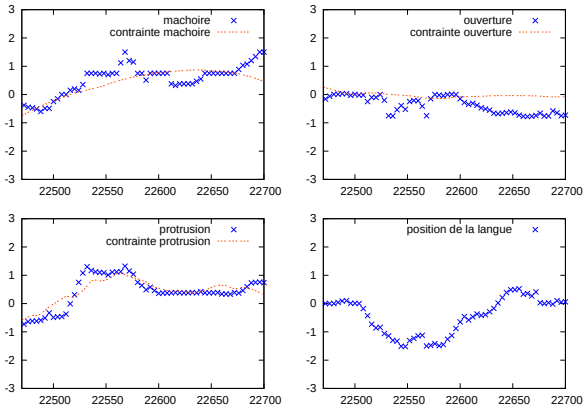

FIG. 4: Résultats de l'inversion pour « cailloux ».

transition /ɑj/. On retrouve la même irrégularité, moins marquée, pour les deux autres paramètres. Nous pouvons également observer que la position de la langue a une trajectoire cohérente avec le mouvement attendu : postérieure pour prononcer /ɔɑ/, elle avance beaucoup pour le /j/, puis recule légèrement pour /e/ (plus la valeur de ce paramètre est élevée, plus la langue est en arrière). Faute de place, nous n'affichons pas les trajectoires des autres articulateurs, d'autant que les évolutions de ces paramètres sont moins facilement interprétables.

La séquence « cailloux » (/aju/ en fait, notre système ne pouvant pas inverser le /k/) a été elle aussi inversée avec succès, et cela avec moins de difficulté pour satisfaire les contraintes visuelles. Comme cela apparaît sur la figure 4 les trois paramètres des articulateurs visibles ont des trajectoires très proches de leur contrainte. Le quatrième paramètre est la position de la langue, qui a, là aussi, une trajectoire phonétiquement réaliste bien que l'amplitude soit beaucoup plus faible (légèrement postérieure pour le /a/, antérieure pour /j/, postérieure pour /u/).

5. CONCLUSION ET PERSPECTIVES

Nous avons effectué le même travail sur les différents locuteurs de ce corpus avec des résultats très similaires. Cette approche du couplage entre paramètres visuels et acoustiques pour l'inversion audiovisuelle articulatoire est donc prometteuse. En effet, le modèle parvient à trouver des solutions satisfaisant les contraintes visuelles tout en prenant en compte les données acoustiques alors que les paramètres articulatoires visuels ne peuvent pas être récupérés avec une très grande précision. Par ailleurs, et il s'agit là d'un point essentiel, les trajectoires articulatoires récupérées sont réalistes d'un point de vue phonétique.

La poursuite de ce travail s'effectue suivant plusieurs axes. Nous travaillons à présent sur un autre corpus beaucoup plus précis géométriquement et plus long car il ne porte que sur une locutrice. Nous étudions en particulier une méthode pour établir une correspondance directe (plutôt que statistique) entre les données visuelles et les paramètres correspondants du modèle de Maeda. En effet, les paramètres obtenus actuellement ne sont pas tout à fait équivalents à ceux du modèle. Ensuite, nous souhaitons évaluer l'influence de l'adéquation géométrique du modèle articulatoire d'analyse au locuteur sur les trajectoires articulatoires inversées. Il est en effet probable que l'adjonction des contraintes conduise à des effets de compensation articulatoire artificiels.

RÉFÉRENCES

[1] C. Benoît, T. Mohamadi, and S. Kandel. Effect of phonetic context on audio-visual intelligibility of french. *Journal of Speech,Language and Hearing Research*, 37 :1195–1203, October 1994.

[2] A. Galván-Rdz. *Études dans le cadre de l'inversion acoustico-articulatoire : Amélioration d'un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des occlusives*. Thèse de l'Institut National Polytechnique de Grenoble, 1997.

[3] Y. Laprie and B. Mathieu. A variational approach for estimating vocal tract shapes from the speech signal. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 929–932, Seattle, USA, May 1998.

[4] B. Le Goff. Automatic modeling of coarticulation in text-to-visual speech synthesis. In *Eurospeech'97 Proceedings*, volume 3, pages 1667–1670, Rhodes, Greece, 1997. European Speech Communication Association.

[5] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

[6] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2) :163–173, March 1983.

[7] Slim Ouni and Yves Laprie. Improving acoustic-to-articulatory inversion by using hypercube codebooks. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, China*, volume II, pages 178–181, October 2000.

[8] Slim Ouni and Yves Laprie. Studying articulatory effects through hypercube sampling of the articulatory space. In *17th International Congress on Acoustics , Rome, Italy*, volume 4, September 2001.

[9] J. Robert-Ribes, J-L. Schwartz, and P. Escudier. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, 9 :323–346, 1994.

[10] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *JASA*, 26(2) :212–215, 1954.

[11] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low cost stereovision based system for acquisition of visible articulatory data. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, pages 145–150, Vancouver, 2005.