Estimation de la longueur du conduit vocal pour l'inversion acoustique-articulatoire

B. Elie et Y. Laprie
LORIA, UMR 7503, BP 239, 54506 Vandoeuvre-Les-Nancy, France
benjamin.elie@inria.fr

La géométrie complexe du conduit vocal rend le problème d'inversion acoustique-articulatoire difficile, notamment de par son caractère fortement mal-posé. La régularisation passe par l'ajout de contraintes, soit articulatoires (modèle articulatoire, nécessitant peu de paramètres, mais nécessitant d'être adapté à chaque locuteur), soit sur les valeurs des fonctions d'aires. Dans ce cas, la longueur du conduit vocal est généralement fixée à une certaine valeur arbitraire, ne permettant pas d'analyser des éventuelles protrusions ou des élongations/raccourcissements du pharynx. L'étude présentée ici propose une approche permettant d'estimer la longueur du conduit vocal de tout locuteur à partir de l'enregistrement du signal de parole. La méthode utilisée est une méthode analyse par synthèse consistant à retrouver la fonction d'aire générant les formants estimés du signal de parole du locuteur. Elle est effectuée à partir d'une fonction d'aire initiale que l'on modifie itérativement selon la méthode des fonctions de sensibilités, d'après la théorie développée par Fant et Pauli sur les perturbations de sections à l'intérieur du conduit vocal. Les travaux présent dans la littérature utilisant cette méthode imposent cependant une longueur fixe des fonctions d'aire, et par conséquent une longueur du conduit vocal fixe. Notre approche permet de régler ce problème en prenant en compte aussi les perturbations de longueur du conduit vocal. Une étude numérique et expérimentale permet de valider la technique dans le cas de voyelles orales du français.

1 Introduction

Le problème de l'inversion acoustique-articulatoire, à savoir l'estimation de la géométrie du conduit vocal à partir du signal acoustique de la parole est étudié depuis les années 60. Les méthodes communément admises sont classifiées en 2 principales catégories : les méthodes par apprentissage [1–3], et les méthodes analyse par synthèse [4–9], qui utilisent un modèle acoustique du conduit vocal dans le but d'estimer les paramètres d'entrée du modèle générant un vecteur acoustique de sortie correspondant à celui, observé, du locuteur. La méthode ici présenté appartient à cette seconde catégorie.

Dans ce papier, le susmentionné paramètre d'entrée du modèle acoustique est la fonction d'aire du conduit vocal, qui est modélisé par une concaténation de N tubes cylindriques. Ce paramètre est défini par un vecteur contenant les N aires et N longueurs de ces tubes. Le vecteur acoustique de sortie est composé des fréquences des 3 ou 4 premiers formants, tel qu'il est défini dans nombre d'études présentes dans la littérature [4, 5, 8, 10, 11]. Cette limite des 3 ou 4 premiers formants correspond grossièrement à la limite de validité de l'hypothèse onde plane, à savoir en-dessous de 4 kHz.

La méthode utilise les fonctions de sensibilité dérivées de Fant [12], et ultérieurement reprises par Story [8, 13] et Carré [5] pour des problèmes d'inversion. C'est un méthode itérative : une nouvelle fonction d'aire est générée à chaque itération de manière à réduire la distance entre les fréquences des formants générés et celles des formants observés. La nouvelle fonction d'aire est déformée par rapport à celle de l'itération précédente selon les fonctions de sensibilité calculées précédemment. Les itérations sont répétées jusqu'à ce que la distance soit inférieure à un seuil arbitraire. Cela requiert de générer une fonction initiale, à partir de laquelle l'algorithme commence ses itérations. Ce problème étant connu comme étant mal posé [14, 15], il est recommandé d'inclure des connaissances *a priori* sur les configurations du conduit vocal humain. Pour cela, cet article propose une inversion dite multimodale, par addition de la connaissance de l'ouverture aux lèvres, au moyen d'un logiciel de capture de mouvements faciaux. Grâce à cela, nous pouvons estimer trois paramètres articulatoires : l'ouverture aux lèvres, l'ouverture de la mâchoire et la protrusion des lèvres.

Une technique proche de celle présentée ici a précédemment été proposée par Bunton *et al.* [8]. Dans cet article, les auteurs fixent la longueur totale du conduit vocal à une valeur arbitraire. Cette contrainte peut devenir problématique lors de l'estimation de trajectoires articulatoires. en effet, la longueur totale du conduit vocale est très susceptible de varier au cours du temps (hauteur du pharynx, protrusion des lèvres, ouverture de la mâchoire) pour un même locuteur, et de manière générale, elle varie d'un locuteur à un autre. L'article présenté ici propose premièrement d'améliorer la méthode en estimant à la fois la forme du conduit vocal et sa longueur totale. La longueur du conduit vocal est estimée suivant la même méthode que pour l'aire des tubes. Deuxièmement, des contraintes sur les trajectoires articulatoires sont ajoutées à l'estimation. Ces contraintes permettent une meilleure régularisation du problème [7, 11], notamment lorsque que l'on traite l'inversion dynamique.

L'article est organisé de manière à décrire les principaux aspects de notre approche. Le paragraphe 2 détaille le modèle de conduit vocal utilisé, ainsi que le calcul des matrices de sensibilité. Ces dernières sont calculées à partir de la théorie des perturbations d'aire et de longueur de Fant [12]. Le paragraphe 3 détaille l'algorithme permettant l'estimation simultanée des aires et des longueurs des tubes. Les différentes contraintes utilisées pour régulariser le problème sont également décrites dans ce paragraphe. La méthode est validée au paragraphe 4 à l'aide de signaux de synthèse générés à partir de fonctions d'aire connues. Finalement, le paragraphe 5 présente des résultats numériques et expérimentaux pour des configurations statiques et dynamiques.

2 Principes théoriques

Ce paragraphe présente la théorie sous-jacente relative à l'algorithme détaillé ultérieurement dans l'article.

2.1 Description du conduit vocal

Le conduit vocal forme les frontières d'un volume d'air (ou colonne d'air). La complexité théorique de traiter de la propagation acoustique à l'intérieur d'un volume quelconque peut être outrepassée en considérant les ondes acoustiques comme des ondes planes. Pour des dimensions similaires à celles du conduit vocal humain, cette hypothèse est communément admise dans la littérature pour des fréquences inférieures à 4 kHz. Ainsi, seule l'aire à l'intérieur du conduit vocal est nécessaire pour décrire ses caractéristiques acoustiques. Le conduit vocal est alors caractérisé par la fonction d'aire $a(x, t)$, correspondant à l'aire du conduit vocal à une distance x de l'origine (prise généralement comme étant la glotte), à un instant donné t. Ceci est valide si la coupe du conduit vocal est considérée comme étant uniformément circulaire.

Cette fonction d'aire, continue, peut être discrétisée de manière à ce que le conduit vocal soit modélisé comme un ensemble de tubes acoustiques connectés en série. À noter que l'échantillonnage de $a(x, t)$ n'est pas nécessairement régulier le long de l'axe x. Pour des raisons de clarté dans les notations, la dépendance temporelle t est volontairement omise dans la plupart des équations de l'article. Dans ce cas précis, la grandeur définie est celle à un instant donné t. Le conduit vocal ainsi échantillonné est alors décrit par deux vecteurs :

$$\begin{cases} \mathbf{a} = [a_1, a_2, \ldots, a_n, \cdots, a_N]^T \\ \mathbf{l} = [l_1, l_2, \ldots, l_n, \cdots, l_N]^T \end{cases}, \tag{1}$$

Pour le reste de l'article, a_1 correspond à l'aire du conduit vocal au niveau de la glotte, et a_N est l'aire aux lèvres, c'est-à-dire l'ouverture aux lèvres.

2.2 Matrices de sensibilité

Ce paragraphe traite du calcul des matrices des dérivées partielles des fréquences des formants par rapport aux fonctions d'aire et de longueur.

2.2.1 Dérivée des fréquences des formants par rapport à l'aire et la longueur des tubes

La théorie de la perturbation de Fant [12] établit la relation existante entre une faible variation d'aire et/ou de longueur le long du conduit vocal et la variation de fréquence des formants. Les Eq. (2) et (3) donnent la variation relative de la fréquence du $m^{\text{ième}}$ formant pour une variation relative de **a** et **l** respectivement :

$$\left[\frac{\Delta F_m}{F_m}\right]_{\mathbf{a}} = \frac{\sum_{n=1}^{N} \left[\mathcal{T}_n(F_m) - \mathcal{V}_n(F_m)\right] \frac{\Delta a_n}{a_n}}{\sum_{n=1}^{N} \left[\mathcal{T}_n(F_m) + \mathcal{V}_n(F_m)\right]}, \tag{2}$$

et

$$\left[\frac{\Delta F_m}{F_m}\right]_{\mathbf{l}} = \frac{\sum_{n=1}^{N} \Delta\lambda_n \left[\mathcal{T}_n(F_m) + \mathcal{V}_n(F_m)\right]}{\sum_{n=1}^{N} \left[\mathcal{T}_n(F_m) + \mathcal{V}_n(F_m)\right]}, \tag{3}$$

où

$$\Delta\lambda_n = -\frac{\Delta l_n}{l_n + \Delta l_n}, \tag{4}$$

et où $\mathcal{T}_n(F_m)$ et $\mathcal{V}_n(F_m)$ sont les énergies cinétiques et potentielles à l'intérieure du $n^{\text{ième}}$ tube, à la fréquence F_m. Les énergies potentielles et cinétiques, pour une pression P et un débit U sont données par

$$\mathcal{T}_n(F_m) = \frac{1}{2}\frac{\rho l_n}{a_n}\left|U_n(F_m)\right|^2 \tag{5}$$

$$\mathcal{V}_n(F_m) = \frac{1}{2}\frac{\rho c_s^2}{a_n l_n}\left|P_n(F_m)\right|^2. \tag{6}$$

Nous pouvons alors définir deux fonctions de sensibilité, $S_n^a(F_m)$ et $S_n^l(F_m)$ pour respectivement les perturbations d'aire et de longueur.

$$S_n^a(F_m) = \frac{\mathcal{T}_n(F_m) - \mathcal{V}_n(F_m)}{\mathcal{H}(F_m)}, \tag{7}$$

et

$$S_n^l(F_m) = \frac{\mathcal{T}_n(F_m) + \mathcal{V}_n(F_m)}{\mathcal{H}(F_m)}, \tag{8}$$

où l'indice m indique la fréquence du $m^{\text{ième}}$ formant, et $\mathcal{H}(F_m)$ est l'énergie totale à l'intérieur du conduit vocal à la fréquence du $m^{\text{ième}}$ formant, d'où

$$\mathcal{H}(F_m) = \sum_{n=1}^{N} \left[\mathcal{T}_n(F_m) + \mathcal{V}_n(F_m)\right]. \tag{9}$$

Nous pouvons alors définir les matrices de sensibilité suivantes

$$\mathbf{J_a} = \begin{bmatrix} S_1^a(F_1) & S_2^a(F_1) & \cdots & S_N^a(F_1) \\ S_1^a(F_2) & S_2^a(F_2) & \cdots & S_N^a(F_2) \\ \vdots & \ddots & \ddots & \vdots \\ S_1^a(F_M) & S_2^a(F_M) & \cdots & S_N^a(F_M) \end{bmatrix}, \tag{10}$$

et

$$\mathbf{J_l} = \begin{bmatrix} S_1^l(F_1) & S_2^l(F_1) & \cdots & S_N^l(F_1) \\ S_1^l(F_2) & S_2^l(F_2) & \cdots & S_N^l(F_2) \\ \vdots & \ddots & \ddots & \vdots \\ S_1^l(F_M) & S_2^l(F_M) & \cdots & S_N^l(F_M) \end{bmatrix}, \tag{11}$$

qui quantifient les variations relatives des fréquences des formants par rapport aux variations relatives des aires et des longueurs le long du conduit vocal.

Pour cette étude, les matrices de sensibilité sont calculées à l'aide du paradigme des matrices en chaîne de Sondhi et Schroeter [16]. Les valeurs des constantes de propagation acoustiques sont identiques à celles de la référence [16].

3 Algorithme récursif pour estimer l'aire et la longueur des tubes le long du conduit vocal

Ce paragraphe décrit l'algorithme utilisé pour l'estimation des aires et des longueurs des tubes le long du conduit vocal, à partir des mesures des fréquences des formants, à l'aide de la technique des matrices de sensibilité.

3.1 Principe général de l'algorithme

La technique utilise un calcul itératif de la fonction d'aire, visant à faire correspondre les fréquences des formants mesurées à celles générées par le modèle. L'algorithme requiert une fonction initiale. Nous utilisons dans ce papier la technique du Jacobien inverse [17]. Étant donné un vecteur initial $\mathbf{a}_0$ ou $\mathbf{l}_0$, indiqué par $\mathbf{x}_0$, générant un vecteur de fréquences de formants initial $\mathbf{f}_0$, le problème est écrit

$$\Delta\mathbf{f} = \mathbf{J}_x \Delta\mathbf{x}, \tag{12}$$

où $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ et $\Delta\mathbf{f}$ indiquent les différences entre les fréquences de formants cibles ($\mathbf{f}$) et initiales ($\mathbf{f}_0$). $\Delta\mathbf{x}$ est alors un vecteur qu'on additionne à $\mathbf{x}_0$ pour atteindre la cible. Ce vecteur est alors écrit

$$\Delta\mathbf{x} = \psi\hat{\mathbf{J}}\Delta\mathbf{f}, \tag{13}$$

où $\hat{\mathbf{J}}$ est soit le pseudo-inverse de Moore-Penrose de $\mathbf{J}$, soit la transposée, et ψ est un coefficient scalaire servant à pondérer le vecteur des différences afin qu'il soit suffisamment petit pour rester dans le domaine linéaire. Cette opération est appliquée itérativement jusqu'à ce que $\Delta\mathbf{x}$ s'annule. Pour cette étude, nous avons choisi empiriquement de prendre la transposée de $\mathbf{J}$ car elle s'est avérée être la technique la plus efficace et rapide.

Les itérations pour les corrections de longueur et d'aires sont appliquées simultanément. Cela revient à prendre en considération un vecteur $\Delta\mathbf{x}$ constitué de la concaténation verticale de $\Delta\mathbf{a}$ et $\Delta\mathbf{l}$. $\mathbf{J}_x$ est alors la concaténation horizontale de $\mathbf{J}_a$ et $\mathbf{J}_l$. Par conséquent, les corrections sur l'aire et sur la longueur ne sont pas concurrentes et il n'existe aucun inconvénient à les appliquer simultanément.

3.2 Estimation de la fonction d'aire

A partir d'une certaine itération k, la fonction d'aire suivante, à l'itération $k+1$ est calculée suivant

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \psi_{\mathbf{a}}\mathbf{A}_k\mathbf{J}_{\mathbf{a}}^T\delta\mathbf{f}_k, \tag{14}$$

où $\mathbf{A}_k \in \mathbb{R}_+^{*N\times N}$ est une matrice diagonale dont les éléments diagonaux sont ceux du vecteur de fonction d'aire $\mathbf{a}_k$, à savoir $\mathbf{A}_k = \text{diag}\,(a_1, a_2, \ldots, a_N)$. $\mathbf{J_a} \in \mathbb{R}^{M\times N}$ est la matrice définie par l'Eq. (10) et $\delta\mathbf{f}_k \in \mathbb{R}^M$ est la fonction des différences entre les fréquences relatives des formants mesurées et générées à l'itération k :

$$\delta\mathbf{f}_k = \left[\delta f_{1_k}, \delta f_{2_k}, \ldots, \delta f_{m_k}, \ldots, \delta f_{M_k}\right]^T, \tag{15}$$

où

$$\delta f_{m_k} = \left[\frac{F_m - F'_{m_k}}{F'_{m_k}}\right]. \tag{16}$$

Dans l'Eq. (16), F_m est la fréquence du $m^{\text{ième}}$ formant mesuré et F'_{m_k} la fréquence du $m^{\text{ième}}$ formant généré par la fonction d'aire de la $k^{\text{ième}}$ itération. Le symbole $\psi_{\mathbf{a}}$ de l'Eq. (14) indique un facteur de vitesse, supérieur à 1, permettant d'accélérer l'algorithme. Une valeur de l'ordre de la dizaine est généralement acceptable [8]. Cependant, il est possible d'optimiser cette valeur en la modifiant à chaque itération selon les valeurs de $\delta\mathbf{f}_k$ et $\mathbf{J}$, selon [17]

$$\psi_{\mathbf{a}} = \frac{\delta\mathbf{f}^T\mathbf{J}_{\mathbf{a}}\mathbf{J}_{\mathbf{a}}^T\delta\mathbf{f}}{(\mathbf{J}_{\mathbf{a}}\mathbf{J}_{\mathbf{a}}^T\delta\mathbf{f})^T(\mathbf{J}_{\mathbf{a}}\mathbf{J}_{\mathbf{a}}^T\delta\mathbf{f})} \tag{17}$$

3.3 Estimation de la fonction de longueur

L'estimation de la longueur du conduit vocal est basée sur la même technique. Premièrement la variation de λ est calculée suivant

$$\delta\lambda = \psi_{\mathbf{l}}\mathbf{J}_{\mathbf{l}}^T\delta\mathbf{f}_k, \tag{18}$$

où, d'après l'Eq. (4),

$$\delta\lambda = \left[\frac{l_{1_k}}{l_{1_{k+1}}} - 1, \frac{l_{2_k}}{l_{2_{k+1}}} - 1, \ldots, \frac{l_{N_k}}{l_{N_{k+1}}} - 1\right]^T. \tag{19}$$

Le nouveau vecteur $\mathbf{l}_{k+1}$ est alors

$$\mathbf{l}_{k+1} = \mathbf{\Lambda}_k\mathbf{l}_k, \tag{20}$$

où $\mathbf{\Lambda}_k \in \mathbb{R}^{N\times N}$ est une matrice diagonale dont les éléments diagonaux sont tels que $\mathbf{\Lambda}_k = \text{diag}\left(\frac{1}{1+\delta\lambda_1}, \frac{1}{1+\delta\lambda_2}, \ldots, \frac{1}{1+\delta\lambda_N}\right)$. La valeur de $\psi_{\mathbf{l}}$ est calculée de manière similaire à l'Eq. (17), en substituant les indices **l** à **a**.

Le processus itératif continue jusqu'à ce que la norme L1 de $\delta\mathbf{f}$ soit inférieure à un certain seuil arbitraire ϵ, c'est-à-dire $\|\delta\mathbf{f}_k\|_1 < \epsilon$.

Étant donné la faible précision de l'estimation des fréquences des formants, il n'est pas nécessaire de choisir un seuil extrêmement bas. Fixer $\epsilon = 1\%$ s'est avéré très suffisant : une plus petite valeur peut augmenter considérablement le temps de calcul sans améliorer l'estimation de manière significative.

3.4 Définition de la fonction initiale

La fonction initiale est fixée à une position neutre, qui est la fonction d'aire correspondant à la configuration articulatoire requérant le moindre effort de la part du locuteur (position au repos), au regard de l'ouverture aux lèvres mesurée à l'aide du logiciel de suivi de mouvements faciaux décrit au paragraphe 5.1. Pour déterminer cette position au repos, nous utilisons le modèle articulatoire de Maeda [18], qui utilise un vecteur de 7 coefficients correspondant aux positions des composantes articulatoires (position de la mâchoire, position du dos de la langue, arrondi de la langue, position de l'apex, hauteur des lèvres, protrusion des lèvres, et hauteur du larynx). Parmi ces 7 paramètres, seuls trois modes de déformation (hauteur de la mâchoire, hauteur des lèvres et protrusion des lèvres) contrôlent l'ouverture aux lèvres. Ces paramètres peuvent être estimés à partir de la connaissance de l'aire d'ouverture aux lèvres par une quelconque technique inverse (telle que la technique du Jacobien inverse, par exemple). La fonction initiale est alors celle calculée en considérant ces 3 paramètres estimés d'après l'ouverture aux lèvres, et en fixant les 4 autres à une valeur nulle.

3.5 Contraintes

Le problème étant mal posé, c'est-à-dire qu'une infinité de configurations de conduit vocal peuvent générer les mêmes fréquences de formants, le problème doit être régulariser au mieux. En plus de la contrainte sur l'ouverture aux lèvres, expliquée au paragraphe 5.1, la solution est soumise à des contraintes supplémentaires : une reliée à l'énergie potentielle articulatoire, l'autre reliée à l'énergie cinétique articulatoire. La seconde n'est importante que lorsque l'on estime une trajectoire articulatoire en configuration dynamique.

3.5.1 Contrainte sur l'énergie potentielle articulatoire

Cette contrainte permet d'éviter des configurations de conduit vocal non réaliste morphologiquement en contraignant la fonction d'aire à ne pas dévier trop loin de sa configuration au repos, *i.e.* la fonction d'aire initiale. Soit $\mathcal{V}_{art}$ l'énergie potentielle articulatoire (qui ne doit pas être confondue avec l'énergie potentielle acoustique définie dans l'Eq. (6)), définie par

$$\mathcal{V}_{art} = \|\mathbf{a} - \mathbf{a}_0\|_2^2, \tag{21}$$

où $\mathbf{a}_0$ est la fonction d'aire au repos, donnée par la fonction d'aire initiale. A noter que l'expression de l'énergie potentielle articulatoire pour la longueur est obtenue par simple substitution de **a** par **l** dans l'Eq. (21).

Le terme de contrainte $C_{\mathcal{V}}$ est alors

$$C_{\mathcal{V}} = \frac{\partial\mathcal{V}_{art}}{\partial\mathbf{a}}\mathcal{V}_{art}, \tag{22}$$

où

$$\frac{\partial\mathcal{V}_{art}}{\partial\mathbf{a}} = 2\left[\mathbf{a} - \mathbf{a}_0\right]. \tag{23}$$

3.5.2 Contrainte sur l'énergie cinétique articulatoire

Cette contrainte n'est utilisée qu'en cas d'inversion dynamique, à savoir lors ce que l'on estime une trajectoire articulatoire. Elle permet de réduire la différence entre une fonction d'aire à un instant t et la fonction d'aire suivante à l'instant $t+1$. Soit $\mathcal{T}_{art}$ l'énergie cinétique articulatoire (qui ne doit pas être confondue avec l'énergie cinétique acoustique définie dans l'Eq. (5)), définie par

$$\mathcal{T}_{art}(t) = \|\Delta\mathbf{a}(t)\|_2^2, \tag{24}$$

où $\Delta\mathbf{a}(t) = \mathbf{a}(t+1) - \mathbf{a}(t)$ est la différence entre deux fonctions d'aire successive dans le temps aux instants t et $t+1$. De manière similaire, l'expression de l'énergie cinétique articulatoire pour la fonction de longueur est obtenue par substitution de **a** par **l** dans l'Eq. (24).

Le terme de contrainte $C_{\mathcal{T}}$ est alors

$$C_{\mathcal{T}}(t) = \frac{\partial\mathcal{T}_{art}(t)}{\partial\mathbf{a}(t)}\mathcal{T}_{art}(t), \tag{25}$$

où

$$\frac{\partial\mathcal{T}_{art}}{\partial\mathbf{a}}(t) = \begin{cases} 2\Delta\mathbf{a}(t), & t = 1 \\ 2\left[\Delta\mathbf{a}(t) - \Delta\mathbf{a}(t-1)\right], & 2 \le t \le t_{max} - 1 \\ 2\Delta\mathbf{a}(t-1), & t = t_{max} \end{cases} \tag{26}$$

3.6 Configuration dynamique

Lorsque l'on inverse un segment acoustique, l'algorithme présenté au paragraphe 3 est légèrement modifié. En effet, les trames temporelles sont inversées simultanément. L'équation (14) devient alors :

$$\tilde{\mathbf{a}}_{k+1} = \tilde{\mathbf{a}}_k + \tilde{\mathbf{A}}_k\left[(1 - c_{kin} - c_{pot})\tilde{\mathbf{J}}_{\mathbf{a}}^T\delta\tilde{\mathbf{f}}_k + c_{kin}\tilde{C}_{\mathcal{T}} + c_{pot}\tilde{C}_{\mathcal{V}}\right], \tag{27}$$

où

$$\begin{aligned}
\tilde{\mathbf{a}} &= [\mathbf{a}(0), \mathbf{a}(1), \ldots, \mathbf{a}(t), \ldots, \mathbf{a}(t_{max})]^T \\
\tilde{\mathbf{A}} &= \text{diag}(\tilde{\mathbf{a}}) \\
\delta\tilde{\mathbf{f}} &= [\delta\mathbf{f}(0), \delta\mathbf{f}(1), \ldots, \delta\mathbf{f}(t), \ldots, \delta\mathbf{f}(t_{max})]^T \\
\tilde{\mathbf{J}}_{\mathbf{a}} &= \text{diag}\left(\mathbf{J}_{\mathbf{a}}(0), \mathbf{J}_{\mathbf{a}}(1), \ldots, \mathbf{J}_{\mathbf{a}}(t), \ldots, \mathbf{J}_{\mathbf{a}}(t_{max})\right) \\
\tilde{C}_{\mathcal{V}} &= [C_{\mathcal{V}}(0), C_{\mathcal{V}}(1), \ldots, C_{\mathcal{V}}(t), \ldots, C_{\mathcal{V}}(t_{max})]^T \\
\tilde{C}_{\mathcal{T}} &= [C_{\mathcal{T}}(0), C_{\mathcal{T}}(1), \ldots, C_{\mathcal{T}}(t), \ldots, C_{\mathcal{T}}(t_{max})]^T,
\end{aligned}$$

c_{kin} et c_{pot} sont des coefficients pondérateurs plus petits que 1 appliqués aux contraintes énergétiques cinétiques et potentielles, respectivement.

Le nouveau seuil d'arrêt du processus itératif $\tilde{\epsilon}$ doit être proportionnel au nombre total de trames temporelles inversées T, à savoir $\tilde{\epsilon} = T\epsilon$.

4 Validation numérique

4.1 Synthèse acoustique

La méthode est dans un premier temps validé à l'aide de simulations numériques. Nous disposons d'une large base de données comprenant des images aux rayons X d'une locutrice prononçant des bouts de phrases en français, et également des enregistrements audios de ces phrases. Les fonctions d'aire sont calculées à partir des coupes sagittales obtenues aux rayons X à l'aide du modèle proposé par Heinz et Stevens [19], qui utilise une loi de puissance reliant l'aire de la section $A(x)$ à la distance sagittale $d(x)$

$$A(x) = \alpha(x) d(x)^{\beta(x)}, \tag{28}$$

où α et β sont des coefficients *ad hoc* qui varient le long du conduit vocal, en s'adaptant à la géométrie de la section.

Les fonctions d'aires ainsi calculées sont mises en entrée d'un synthétiseur vocal. Le synthétiseur employé est celui de Maeda [20, 21].

4.2 Résultats

Les valeurs des différents paramètres et contraintes sont indiquées dans le tableau 1.

Tableau 1 – Paramètres constants choisis pour l'étude présentée

Paramètre	Valeur	Paramètre	Valeur
Nombre de tubes N	40	ψ	15
Nombre de formants M	3	ϵ_1 (%)	1
c_s (m.s^{-1})	343	c_{kin}	0.9
ρ (kg.m^{-3})	1.204	c_{pot}	10^{-3}

Les formants du signal acoustique enregistré sont estimés par une stratégie de courbes concurrentes utilisant une LPC (*Linear Predictive Coding*) [22]. L'inversion est également valide en utilisant d'autres techniques d'estimation de formants, telles que les techniques basées sur les coefficients cepstraux, par exemple [9].

4.2.1 Voyelles statiques

L'ouverture aux lèvres est contrainte en imposant une valeur fixe à a_N, correspondant à la valeur cible obtenue à partir des images aux rayons X. La figure 1 montre la fonction d'aire cible, la fonction d'aire initiale, et celle estimée. Les voyelles correspondantes sont un /a/ et un /i/.

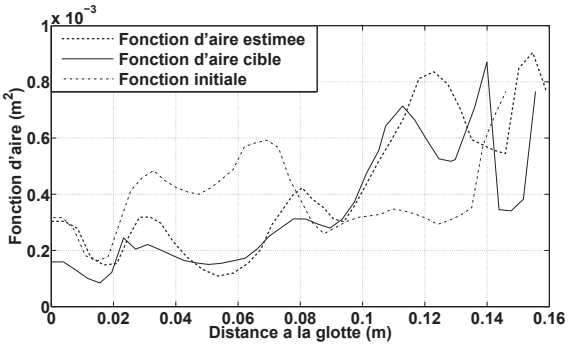
Les fonctions d'aire estimées sont très proches de la fonction cible. Nous observons que la qualité de l'inversion concerne non seulement la forme générale du conduit vocal (les aires des sections), mais également la longueur totale du conduit. En effet, les longueurs estimées sont de 15.91 et 15.97 cm, pour respectivement le /a/ et le /i/, ce qui est très proche des valeurs cibles de 15.56 cm et 16.08 cm, soit des erreurs de respectivement 3.5 mm et 1.1 mm.

4.2.2 Configuration dynamique

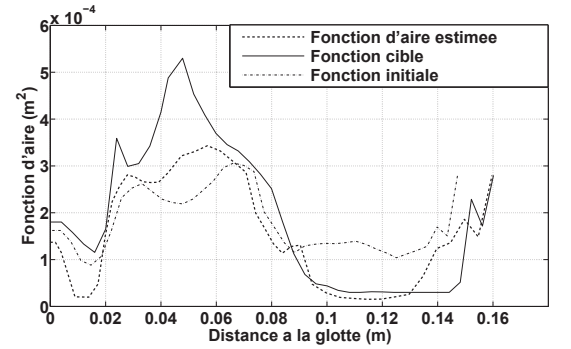
Le test concerne l'inversion d'un ensemble de voyelles prononcées à la suite. La transition est /aøeiou/. La figure 2 montre les résultats de la simulation.

Les résultats de l'estimation montrent que cette dernière est précise : l'erreur ne dépasse que très rarement les 5 mm. L'inversion réalisée sur l'ensemble du segment acoustique permet de bien retrouver la trajectoire de la longueur du conduit vocal. En effet la tendance à l'augmentation de cette longueur se retrouve sur la figure 2.

Les résultats obtenus à partir de voyelles de synthèse, à la fois sur des voyelles statiques et sur des transitions dynamiques nous permet d'effectuer des estimations de géométrie de conduit vocal



(a) /a/



(b) /i/

Figure 1 – Résultats de l'inversion du signal de synthèse : fonction d'aire cible (ligne pleine), fonction initial (ligne traitillée/pointillée) et fonction d'aire estimée (ligne traitillée). Deux voyelles sont inversées, un /a/ et un /i/.

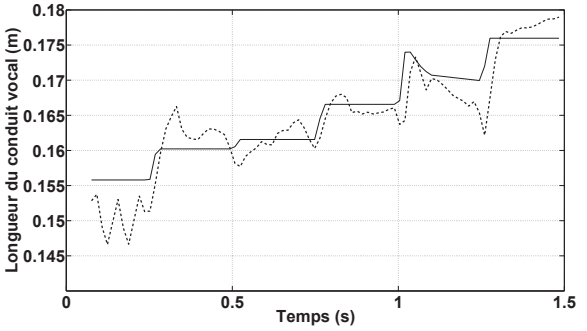


Figure 2 – Longueur estimée (ligne traitillée) du conduit vocal lors de la phrase de synthèse /aøeiou/. Pour comparaison, la longueur mesurée par rayon X du conduit vocal est indiquée en ligne pleine.

avec confiance à partir de signaux réels enregistrés sur des locuteurs. Ceci constitue le sujet de la partie suivante.

5 Expériences

Ce paragraphe détaille le protocole expérimental pour acquérir les données d'entrée de l'algorithme, ainsi que des résultats d'inversion.

5.1 Mesure de l'ouverture aux lèvres

L'ouverture aux lèvres est estimée à l'aide du logiciel de suivi de mouvements faciaux *Faceshift* [1], qui ne nécessite pas de positionner physiquement des marqueurs sur le locuteur. Un détecteur de profondeur permet de visualiser et d'obtenir en temps-réel le contour des lèvres (*cf.* figure 3), ainsi que la position des points souhaités dans l'espace. La vitesse de la caméra est de

1. http://www.faceshift.com/

30 images par secondes. La connaissance des coordonnées dans l'espace à 3 dimensions des points formant le contour des lèvres permet d'obtenir l'aire de l'ouverture aux lèvres.

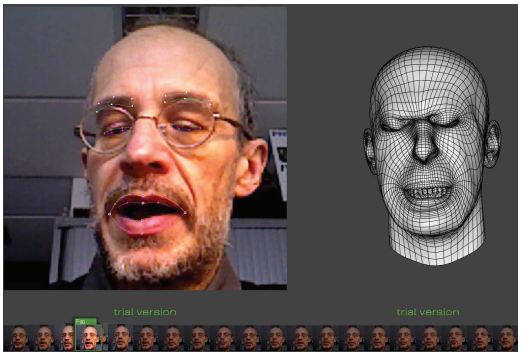


FIGURE 3 – Capture d'écran du logiciel de suivi de mouvements faciaux FaceShift. La profondeur et la position des points (inscrits de manière digitale) sont relevées à chaque trame temporelle (30 images par secondes).

5.2 Résultats

Ce paragraphe montre des exemples d'inversion de voyelles françaises produites par un locuteur dont le français est la langue maternelle. Le paragraphe 5.2.1 présente des résultats en configuration statique, le paragraphe 5.2.2 présente des estimations de trajectoires, d'après une inversion en configuration dynamique (transition entre deux voyelles).

5.2.1 Voyelles statiques

La figure 4 montre des résultats obtenus pour des voyelles statiques. Les fonctions d'aire estimées (ligne pleine) sont en bon accord avec les formes de conduit vocal attendues pour ces voyelles. /i/ et /e/ présentent une large cavité postérieure et une cavité antérieure étroite, alors que /a/ présente une cavité postérieure étroite et tend à s'ouvrir dans la cavité antérieure, débouchant sur une ouverture aux lèvres très large. À noter que /a/ présente le conduit vocal le plus petit, ce qui est également en accord avec ce qui est attendu : le conduit vocal générant un /a/ est raccourci, du fait de l'ouverture très large aux lèvres et de l'absence de protrusion. Le conduit vocal estimé pour /u/ présente deux cavités bien distinctes, séparées par une constriction, situées aux environs du milieu du conduit vocal. Comme attendu, d'après les études antérieures, la longueur du conduit vocal estimée pour /u/ est la plus grande, du fait d'une faible ouverture aux lèvres et d'une forte protrusion.

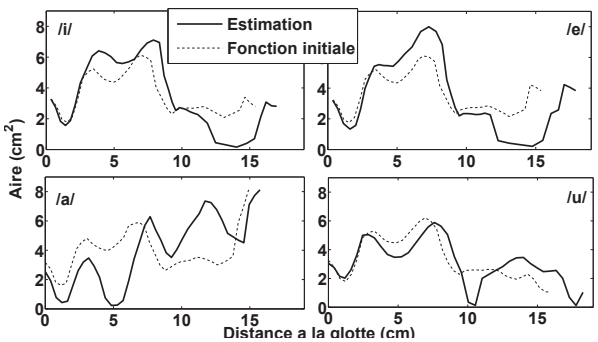


FIGURE 4 – Fonctions d'aire estimées (ligne pleine) de 4 voyelles statiques du français. La fonction d'aire initiale est représentée en lignes traitillées.

Le temps de calcul pour les différentes estimations vont de 0.95 s pour le /u/ à 4.73 s pour le /a/. Ces très bas temps de calcul constituent une avancée optimiste vers une utilisation de l'inversion en temps-réel. Cela pourrait être éventuellement réalisé après optimisation algorithmique.

5.2.2 Transition dynamique

La figure 5 montre des trajectoires de fonctions d'aire allant d'un /a/ à un /u/, ainsi que la variation de la longueur du conduit vocal estimé lors de la même transition. Comme attendu, la longueur du conduit vocal tend à augmenter durant la transition de /a/ à /u/. La transition entre les deux sons peut être clairement identifiée autour du milieu de cette dernière : il existe une augmentation soudaine de la longueur du conduit vocal, de la cavité buccale, et un soudain rétrécissement de la constriction autour d'une position située à 9 cm de la glotte. Nous noterons que l'évolution de la longueur du conduit vocal n'est cependant pas monotone, elle tend à diminuer au début du /a/ pour atteindre rapidement un minimum, puis augmente jusqu'à un maximum au début du /u/. Elle tend à se stabiliser après une légère diminution au début du /u/. Néanmoins, les ordres de grandeurs de ces variations étant dans les marges d'erreurs observées lors des simulations du paragraphe 4.2.2, il n'est pas possible de tirer de réelles conclusions sur de possibles ajustements articulatoires du locuteur pendant la transition.

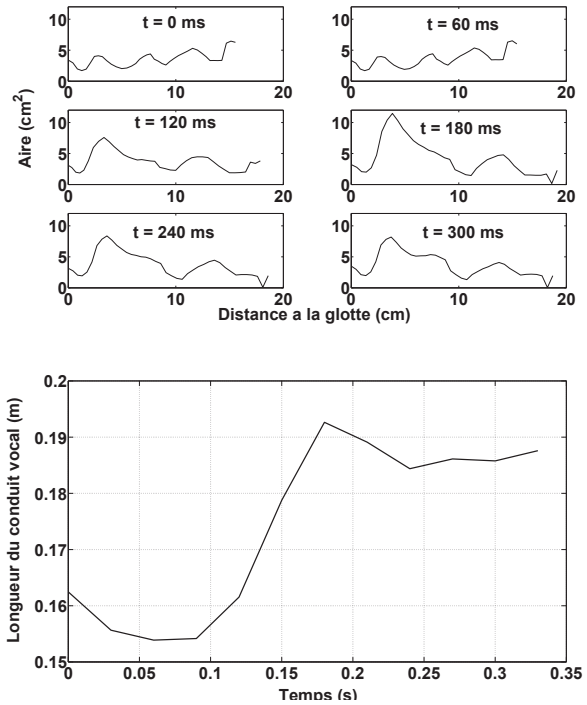


FIGURE 5 – a) Évolution de la fonction d'aire estimée de /a/ à /u/. L'intervalle de temps entre chaque figure est de 60 ms. La figure supérieure gauche correspond à t = 0 ms, la figure en bas à droite correspond à t = 300 ms. Le temps de calcul total est de 126 s. b) Évolution de la longueur du conduit vocal estimé de /a/ à /u/.

6 Conclusions et perspectives

La méthode proposée dans cet article permet une bonne estimation des fonctions d'aire et de longueur du conduit vocal à partir d'une acquisition simultanée du signal acoustique et de l'aire d'ouverture aux lèvres. En comparaison avec les méthodes existantes, la méthode présentée ici possède l'avantage de s'exécuter sans le besoin, ni d'un apprentissage *a priori* d'une base de données, ni d'une recherche dans un dictionnaire pré-établi, qui sont des méthodes très coûteuses en termes de temps de calcul. La technique s'adapte également aisément à n'importe quel locuteur,

et ne nécessite pas de fixer une longueur de conduit vocal à une valeur arbitraire. Ce point permet donc d'étudier les variations de longueur de conduit vocal d'un locuteur ou les variations inter-locuteurs. Le temps de calcul étant bas, moins de 5 secondes pour une voyelle, en configuration statique, la méthode pourrait être implémentée dans un logiciel d'estimation en temps-réel. Cela constitue un point important pour les applications potentielles, telles que la réhabilitation vocale, ou l'apprentissage de langues, par exemple. Cependant, la technique ne fonctionne que pour des voyelles orales pour l'instant. Pour les autres phonèmes, tels que les voyelles nasales ou les fricatives, une solution pratique consisterait à construire les matrices de sensibilité de manière numérique, et non plus de manière analytique.

L'algorithme requiert l'introduction de plusieurs contraintes pertinentes dans le but de régulariser le problème. Pour cela, nous proposons de contraindre :

- l'aire de l'ouverture aux lèvres, à l'aide d'un logiciel d'acquisition de mouvements faciaux.
- la fonction d'aire estimée à être le plus proche possible de sa fonction d'aire initiale, correspondant à celle requérant le moindre effort de la part du locuteur (position au repos). Cette contrainte permet de prévenir des éventuelles configurations irréalistes d'un point de vue morphologique.
- la différence entre deux fonctions d'aires successives, lors d'inversion dynamiques, à être la plus petite possible, afin d'éviter des mouvements articulatoires trop brusques.

Le principal défi de la technique est d'ajuster au mieux les différentes contraintes. En effet, leurs modifications peuvent changer l'estimation de la fonction d'aire. Par conséquent, elles doivent être choisies de manière appropriée. Une définition robuste de ces contraintes est un défi important à prendre en compte pour la suite de ces travaux.

Références

[1] A. Soquet, M. Saerens, and P. Jospa, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model," in *The ESCA Workshop on Speech Synthesis*, 1991, pp. 1–5.

[2] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using a HMM-based speech production model," *IEEE Trans. Speech Audio Proc.*, vol. 12(2), pp. 175–185, 2004.

[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, no. 3, pp. 215–227, 2008.

[4] Z. Yu, "A method to determine the area function of speech based on perturbation theory," *STL-QPSR*, vol. 34(4), pp. 77–96, 1993.

[5] R. Carré, "From an acoustic tube to speech production," *Speech communication*, vol. 42, pp. 227–240, 2004.

[6] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118(1), pp. 444–460, 2005.

[7] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2144–2162, 2011.

[8] K. Bunton, B. H. Story, and I. R. Titze, "Estimation of vocal tract area functions in children based on measurement of lip termination area and inverse acoustic mapping," in *Proceedings of meetings on acoustics*, 2013, vol. 19, pp. 1–8.

[9] J. Busset and Y. Laprie, "Acoustic-to-articulatory inversion by analysis-by-synthesis using cepstral coefficients," in *Proceeding of meetings on Acoustics*, 2013, vol. 19.

[10] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41(5), pp. 1283–1294, 1967.

[11] B. Potard and Y. Laprie, "A robust variational method for the acoustic-to-articulatory problem," in *Interspeech, Brighton 2009*, 2009.

[12] G. Fant, "Vocal-tract area and length perturbations," *Roy. Swedish Academy of Music*, vol. 16(4), pp. 1–14, 1975.

[13] B. H. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *J. Acoust. Soc. Am.*, vol. 119(2), pp. 715–718, 2006.

[14] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, 1960.

[15] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, vol. 41(4), pp. 1002–1010, 1967.

[16] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 35(7), pp. 955–967, 1987.

[17] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation*, vol. 17, pp. 1–19, 2004.

[18] S. Maeda, "Un modele articulatoire de la langue avec des composantes linéaires," 1979, pp. 152–162.

[19] John M Heinz and Kenneth N Stevens, "On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech," in *Proc. 5th Int. Congress of Acoustics*, 1965, vol. 44.

[20] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech communication*, vol. 1, pp. 199–229, 1982.

[21] S. Maeda, "Phonemes as concatenable units : VCV synthesis using a vocal-tract synthesizer," *Phonetica*, pp. 127–232, 1996.

[22] Y. Laprie, "A concurrent curve strategy for formant tracking," Jegu, Korea, Oct. 2004.