Rapport de stage

Transformation statistique de la voix
Application à la modification de l'âge du locuteur

Auteur
Xavier FAVORY

Encadrant
Nicolas Obin

30 juillet 2014

Remerciements

Je tiens à remercier dans un premier temps, toute l'équipe pédagogique de l'ENSEA et les intervenants professionnels responsables de la formation pour avoir assuré la partie théorique de celle-ci.

Je remercie également Sylvain Reynal pour l'aide et les conseils qu'il m'a apporté lors des différents suivis concernant les missions évoquées dans ce rapport.

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance aux personnes suivantes, pour l'expérience enrichissante et pleine d'intérêt qu'elles m'ont fait vivre durant ces trois mois au sein de l'IRCAM :

Axel Roebel, responsable du département Analyse et synthèse des sons, pour son accueil et la confiance qu'il m'a accordée dès mon arrivée dans l'Institut.

Mon encadrant Nicolas Obin, pour m'avoir intégré rapidement au sein de l'équipe et m'avoir accordé toute sa confiance, pour le temps qu'il m'a consacré tout au long de cette période, sachant répondre à toutes mes interrogations, sans oublier son aide au cheminement de ce rapport

Tous les chercheurs du département Analyse et synthèse des sons, pour leur aide et leur conseil ; ainsi que tous le personnel de l'IRCAM, pour leur accueil sympathique et leur coopération professionnelle.

Et enfin ma famille, pour le soutien et la patience qu'elle m'a témoignée.

Table des matières

Introduction

La transformation de la voix est un sujet de recherche très présent depuis de nombreuses années à l'IRCAM. De nombreux outils ont été développés, que ce soit pour la voix parlée ou pour la voix chantée. L'étude que j'ai développé pendant mon stage porte sur la voix parlée. Plus concrètement l'objectif de ma recherche est d'essayer de la transformer pour modifier des caractéristiques perceptives de la voix, comme l'âge ou le sexe.

Actuellement un logiciel de transformation de la voix est disponible sur le marché, sous le nom de IRCAM Trax Transformer. Ce logiciel s'insère parfaitement dans une chaine de traitement numérique. Basé essentiellement sur SuperVP, il permet de manipuler des caractéristiques de la voix comme par exemple le sexe, l'âge, l'expression, ...
Ces algorithmes de transformation ont déjà fait leur preuve dans des projets cinématographique (Farinelli, Vatel, Tirésia, Les amours d'Astrée and Céladon).

En outre, des travaux basés sur des connaissances physiologiques ont conduit à des résultats pour l'application à la transformation de la voix. Dans ce sens le projet Vivos a pour but de développer des techniques de synthèse et de traitement spécifiques au caractère expressif des voix. En passant par la transformation de l'identité d'une voix, la transformation du type et de la nature de la voix, la synthèse de voix expressive ou la synthèse à partir de texte. Par exemple, pour le passage d'une voix d'homme à une voix de femme, on comprend aisément que l'homme ayant les cordes vocales et un conduit vocal plus grands que ceux de la femme aura un pitch plus faible et une coloration spectrale plus grave.

D'autres recherches ont été réalisées pour étudier l'évolution des paramètres acoustiques de la voix en fonction de l'âge.

En parallèle, des méthodes de transformations statistiques de la voix ont été développées pour modifier l'âge perçu dans la voix. Ces méthodes sont basées principalement sur des techniques de régression multiple des paramètres acoustiques de la voix en fonction de l'âge.

Cependant, on remarque que les transformations développées sont les mêmes pour toutes les voix, et ne s'adaptent donc pas aux spécificités de la voix de chaque individu. Par exemple, pour deux individus la voix peut vieillir de manière différente. Ces transformations ne se basent que sur la modification de la longeur des cordes vocales (hauteur tonale), ou sur la taille du conduit vocal (enveloppe spectrale, position des formants).

L'objectif de mon étude est d'essayer d'apporter de nouvelles techniques dans la transformation statistique de la voix, avec application à la modification de l'âge du locuteur. Pour ce faire il est important d'explorer une représentation acoustique

"avancée" de la voix, notamment l'utilisation de la source glottique (qualité vocale). Une autre idée importante est de réaliser une représentation statistique permettant d'adapter la transformation au spécificités de la voix d'un individu, pour prendre en compte qu'il existe de multiples manières de vieillir une voix.

1 Représentation acoustique de la voix

L'équipe Analyse et synthèse des sons a développé un nombre important d'outils d'analyse pour la voix. SuperVP (Super vocodeur de phase) est une technologie pour le traitement des sons en temps différé et en temps réel. Il s'agit d'une sorte de vocodeur de phase amélioré qui m'a permis d'analyser les voix de locuteurs, et de les transformer. Par ailleurs, Le travail de recherche de Gilles Degotex sur la séparation source filtre (SVLN), m'a également été très utile. Avant de pouvoir réaliser les analyses, il a fallu avoir à disposition un corpus d'échantillons de voix. Ainsi j'avais à ma disposition un corpus de phrases en français lues par 80 locuteurs (BREF80). Mais j'ai également du créer mon propre corpus adapté pour l'étude des paramètres influents sur l'âge perçu dans la voix des locuteurs.

1.1 Constitution d'un corpus de locuteurs

BREF-80 est le résultat d'efforts de chercheurs du LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur) , laboratoire propre du CNRS. Ce corpus a été conçu pour disposer d'un échantillon suffisamment ample de parole pour développer des systèmes de reconnaissance de parole, mais aussi pour fournir un large corpus de parole pour l'acquisition de connaissances acousto-phonétiques sur le français. Cependant BREF-80 ne m'a pas permis d'obtenir de résultats pour ma mission : les échantillons de voix des locuteurs ne sont disponibles qu'à une seule prériode de leur vie. Il est donc difficile d'observer l'évolution des paramètres acoustiques de la voix suivant l'âge. De plus, la répartition des âges des locuteurs n'était pas assez large.

Ainsi, il s'est avéré nécessaire de constituer un corpus à partir de fichiers récupérés sur internet. Avec l'aide de Nicolas Obin, nous avons constitué un corpus d'échantillons de voix de personnalités. Nous avons décidé de travailler avec la langue française uniquement, pour que les particularités de chaques langues n'influent pas sur nos résulats. Nous avons pu obtenir des paroles de mêmes locuteurs à différents âges. Ceci nous a permis d'observer l'évolution des paramètres acoustiques des voix de différents locuteurs, et ainsi pouvoir extraire différentes trajectoires d'évolution des paramètres.

Les fichiers audio ont été extraits de fichiers video mp4, et ensuite transformés en wave. Résolution de 16 bits, sous-échantilloné à 12kHz pour que la représentation spectrale ne soit pas influé par le haut du spectre (< 6 kHz). L'inconvénient majeur de ce corpus est que les extraits ont été enregistrés dans des conditions à chaque fois différentes, et pour certains dans de mauvaises conditions. Ceci a eu un impact sur la précision des analyses.
Voici un descriptif de Age Corpus contenant au total 88 fichiers de 13 locuteurs français :

Locuteurs	Nombre fichiers	Répartition age
Yannick Noah	7	14 - 50
Isabelle Huppert	7	23 - 57
Jeanne Moreau	10	27 - 79
Simone Veil	6	46 - 77
Bernard Pivot	6	32 - 72
Marguerite Duras	8	27 - 79
Annie Girardot	6	27 - 72
Brigitte Bardot	4	25 - 71
Catherine Deneuve	5	29 - 58
Jacques Chirac	5	38 - 74
JeanPaul Belmondo	7	27 - 77
Johnny Hallyday	8	17 - 68
Michel Drucker	9	22 - 70

1.2 Paramètres acoustiques de la voix

L'instrument

La voix est un instrument complexe permettant de produire des sons, de communiquer des informations ou même des émotions. La synthèse de la voix se fait à partir d'une source et d'un corps sonore. L'air provenant des poumons fait vibrer les cordes vocales (la source). C'est au niveau du conduit vocal que le son est amplifié et filtré (le corps sonore). Le conduit vocal est constitué du pharynx, de la cavité buccale, et des articulations (voir figure 1). La voix est basée sur un modèle dit source/filtre. La source détermine la hauteur tonale du son, et la qualité du son. Le conduit vocale donne au son une couleur en fonction des mouvements d'articulations du locuteur.

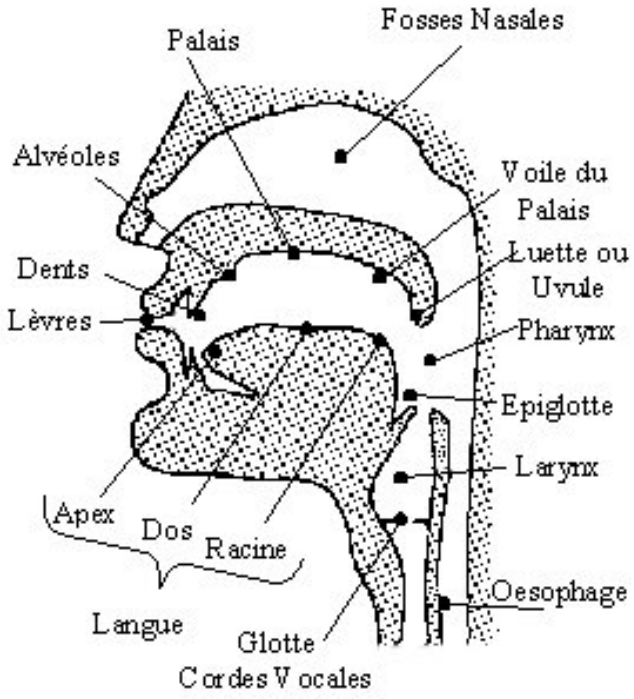


FIG. 1 – Représentation de l'appareil phonatoire

Le modèle source filtre utilisé dans les travaux actuels sont souvent basés sur un modèle "standard". L'utilisation d'un modèle "avancé" nous permet de prendre en compte des paramètres importants pour pouvoir modifier la qualité de la voix.

Modèle source/filtre standard

physique	signal	perception
source glotique	F0 (+bruit)	hauteur
conduit vocal	enveloppe spectrale	timbre

Modèle source/filtre "avancé"

physique	signal	perception
source glotique	F0 (+bruit) / LF-Rd / GCI	hauteur / qualité vocale
conduit vocal	enveloppe spectrale	timbre

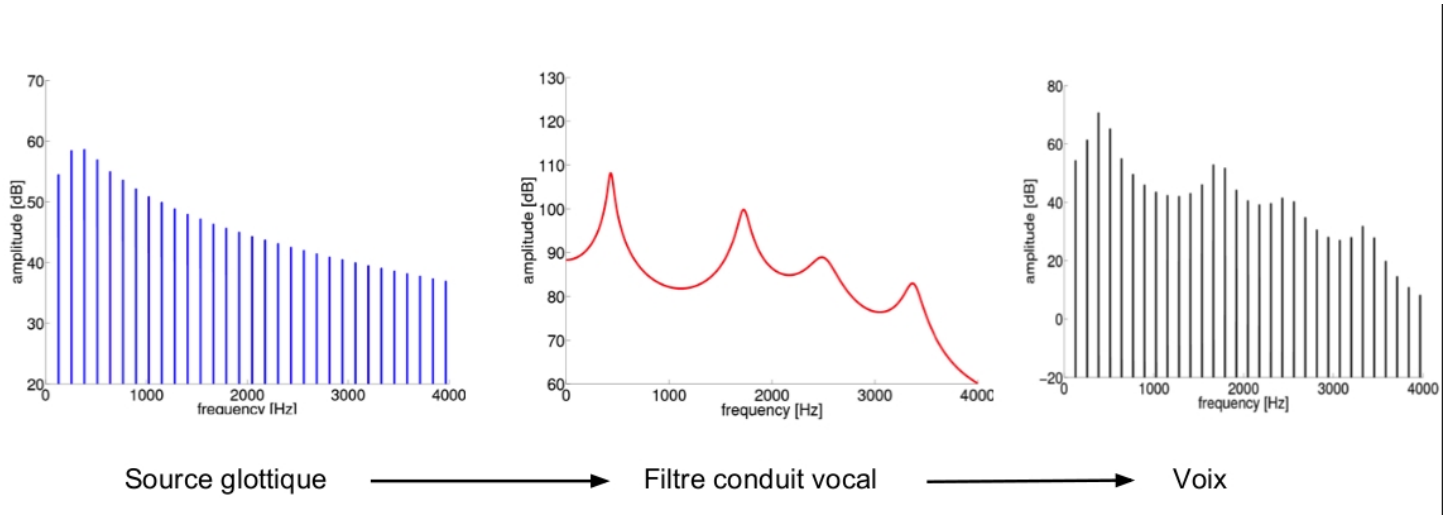


FIG. 2 – Représentation du modèle de synthèse de la voix

La qualité vocale

L'étude de la qualité vocale est traitée dans de nombreux sujets de recherche de traitement de la parole. Pour la transformation de la voix, la source glottique est très importante. Un modèle plus avancé que le modèle source/filtre standard nous permet de modéliser certaines caractéristiques de la voix, comme les voix soufflées, craquées, rauques, ... [14]Le coefficient de relaxation Rd nous permet de résumer un ensemble de paramètres modélisants la forme du pulse glottique.

Enveloppe spectrale

L'enveloppe spectrale est un aspect très important de la voix. Elle est fixée par le conduit vocal qui évolue au cours de notre vie. En effet le conduit vocal change de forme et donc les fréquences de résonances changent. C'est donc un paramètre important pour notre travail.

L'estimation de l'enveloppe spectrale est un sujet récurrent en traitement de la parole. La pratique courante consiste à effectuer cette estimation trame par trame en considérant chacune de ces trames séparément. Pour ce faire, des méthodes à base de prédiction linéaire ou de cepstre sont couramment employées. La difficulté arrive lorsqu'on est dans des applications nécessitant des transformations de l'enveloppe variant dans le temps. Dans mon cas, je veux faire une sorte de conversion de voix, entre un locuteur source, et un locuteur cible (locuteur source à un autre age). L'un des problèmes cruciaux est de préserver la cohérence temporelle du signal converti. SuperVP dispose d'un outils d'analyse et de synthèse de l'enveloppe, c'est la True Enveloppe [22]. Elle donne une analyse cepstral du signal de parole.

Cepstre

Le cepstre d'un signal est une transformation de ce signal dans un domaine analogue au domaine temporel. Il s'exprime comme étant la transformé inverse du logarithme de la transformé de Fourier du signal.

Cepstre = $F^{-1}(\ln(F(x(t))))$; où x(t) est le signal temporel.

C'est une représentation du signal bien adapté pour la voix. En effet, si le signal est issue d'un signal source x(t) filtré par un filtre de réponse impulsionnel h(t).
Le signal s'exprime $y(t) = x(t) \otimes y(t)$; où $\otimes$ est l'opérateur de convolution.

De part les propriétés de la transformé de Fourier, on peut alors exprimer la transformé du signal y(t) :

F(y(t)) $= F(x(t)) \times F(h(t))$

$\ln(\mathrm{F}(\mathrm{y}(\mathrm{t}))) = \ln(x^*) + \ln(h^*)$

et finalement : $Cesptre_y = Cesptre_x + Cesptre_h$

Si x(t) correspond à un signal variant rapidement, et h plutôt à la réponse d'un filtre "lent", on peut alors aisément séparer les deux.

1.3 Paramètres liés à l'âge

L'étude de l'évolution des paramètres acoustiques de la voix liés à l'âge ont fait l'objet de beaucoup de recherche, pour des systèmes de transformation de la voix, ou pour par exemple évaluer l'impact du vieillissement du locuteur sur un système de reconnaissance [23]. Les différents paramètres sont basés sur un modèle source/filtre "standard". [17] [1]

F0
La fréquence fondamentale de vibration des cordes vocales F0 représente la hauteur tonale de la voix. Son évolution en fonction de l'âge du locuteur a été étudiée dans [16]. Pour les hommes il a été observé que cette fréquence F0 diminuait entre l'adolescence et l'âge adulte, pour finalement augmenter légérement au delà de 60 ans.

Intensité et niveau de bruit
Les variations de l'intensité du signal de parole ont également été étudiées [12]. D'autres travaux se sont intéressés au niveau de bruit de la source glottique [10].

Les Formants
Les formants correspondent à des résonances dans le spectre. Les positions des résonances changent selon la voyelle prononcée. En se limitant aux deux premiers formants, on peut déjà réaliser une classification vocalique (voir figure 2). Des travaux sur l'évolution de la fréquence fondamentale (f0) et des fréquences des premiers formants (F1, F2, F3) en fonction de l'âge nous apprennent qu'il y a des changements importants. [11] Ils ont pu observer grâce à des mesures sur des personnes prises entre leur 29 et 50 ans : Une diminution de F0 et F1 ; une légère diminution de F2 ; et aucun changement, voire une augmentation de F3. Ces résultats sont assez difficile à obtenir en pratique. Une telle étude des fréquences des formants selon l'âge et les voyelles demande d'avoir des enregistrements réalisés sous les mêmes conditions, en faisant prononcer aux locuteurs les mêmes voyelles, ou phrases.

En vue des difficultés à obtenir de bonnes analyses pour modéliser l'espace acoustique phonétique et ses évolutions en fonction de l'âge du locuteur, on s'attend à encore plus de difficultés pour transformer l'espace acoustique d'un locuteur pour modifier l'âge perçue. Mon but a donc été de trouver une manière simple de transformer le spectre de la voix, en choisissant de garder la représentation cepstrale.

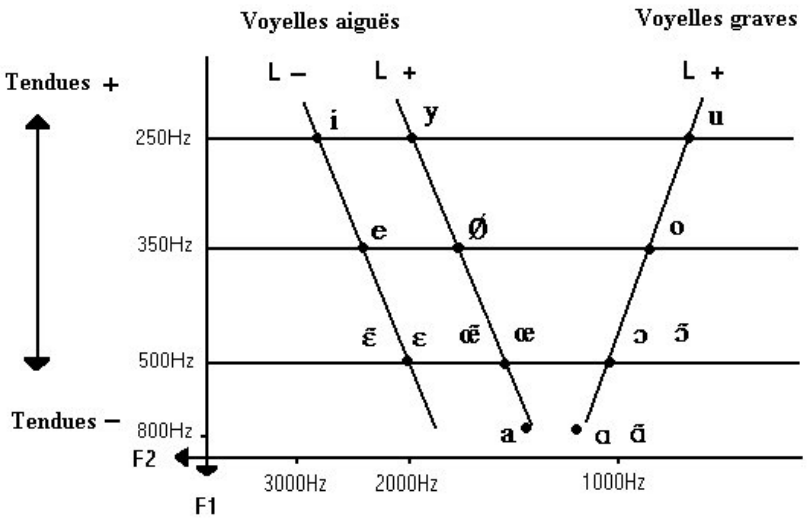


FIG. 3 – Représentation des voyelles en français dans le plan (F1,F2)

Paramètres glottiques

La glotte qui est située dans le larynx comprend les deux cordes vocales. La synthèse de la source pour la voix humaine est du à des alternances d' ouvertures et de fermetures de la glotte. Ceci a pour effet de créer ce qu'on appelle un pulse glottique, qui contient un fondamental, et des harmoniques. On note également l'ajout d'une source de bruit à la voix, qui peut être du à une fuite d'air. Il faut noter que lorsque la glotte émet un son, on obtiendra un son voisée, c'est à dire harmonique, avec une hauteur tonale.

Model source filtre LF

Le modèle de Liljencrantz-Fant donne une représentation de la forme du pulse glottique. Des travaux ont permis de résumer un ensemble de paramètres glottiques en un seul [6]. C'est le coefficient Rd de relaxation, qui modélise le caractère "tendu" de la voix. On peut voir l'effet d'un changement du coefficient Rd sur le spectre du pulse (figure 5).

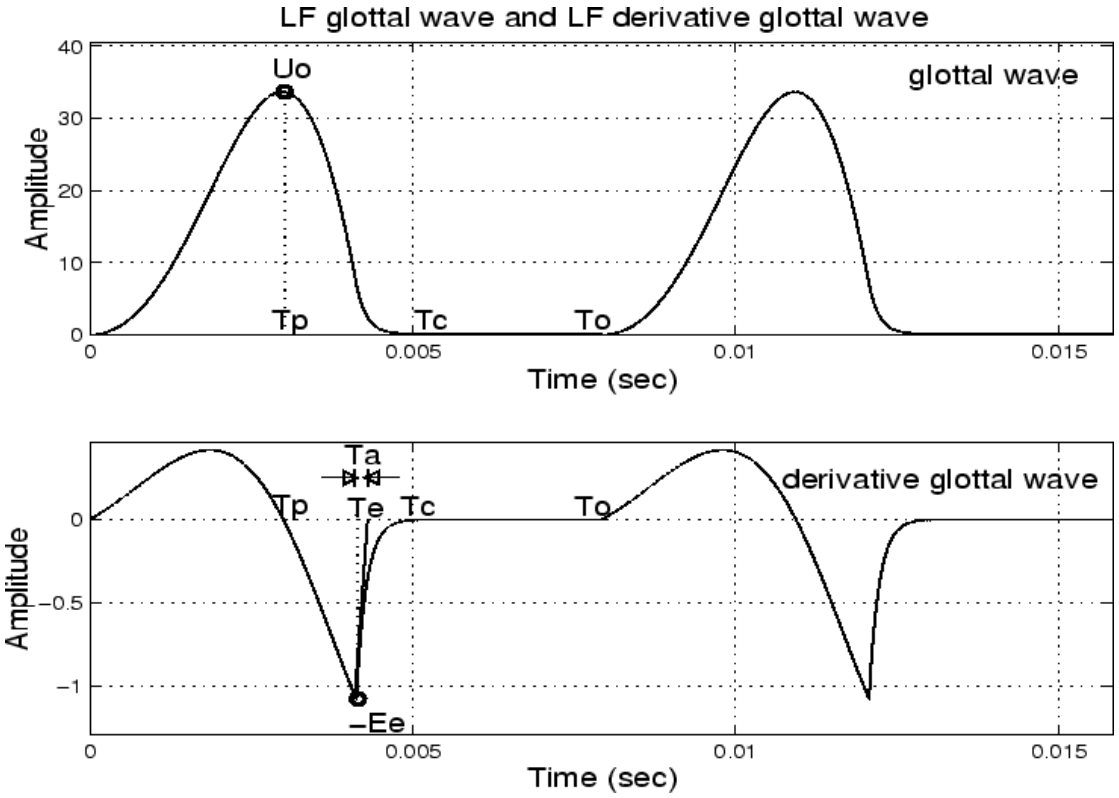


FIG. 4 – Modèle LF

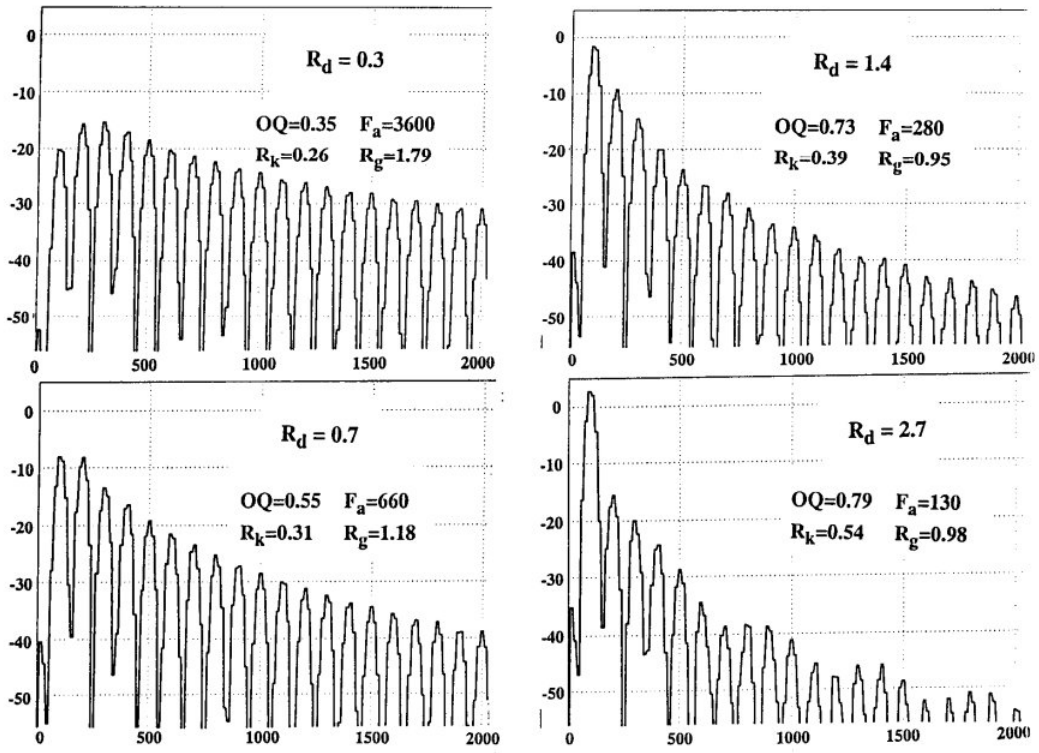


Fig. 5A. Glottal flow derivative spectra at F_0=100 Hz and varying R_d.

FIG. 5 – Représentation des spectres pour différents Rd

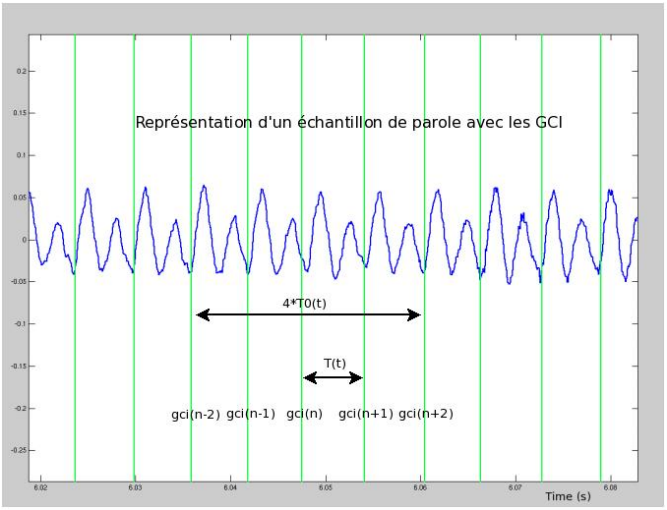
Relative Noise Ratio Dans le modèle source filtre peut se rajouter une source de bruit en plus du signal harmonique. Cela permet d'obtenir des voix soufflées, ou souvent qualifiées de breathy. La mesure nous donne un niveau en dB, représentant la différence entre l'énergie du pulse glottique E et le niveau de bruit.

Glottal pulse aperiodicity

Un phénomène important survenant sur certains locuteurs quand ils deviennent âgés, est le craquement. Il serait intéressant de pouvoir générer des voix craquées, mais cela pourrait donner lieu à un sujet de stage entier. Ce craquement vient d'une irrégularité des pulses glottiques. Une apériodicité dans les instants de fermeture de la glotte, et une fluctuation de l'énergie des pulses provoquent un effet de voix "rauque".

Le modèle de séparation de source conduit vocale (SVLN) réalisé par Gilles [4], m'a permis de travailler sur cet aspect. J'ai pu ajouté une irrégularité dans les instants de fermeture glottique pour obtenir une voix craquée. Il a fallu trouver une caractérisation de ce craquement, et donc de l'apériodicité des GCI (Glottal Closure Instant). J'ai implémenté une fonction de calcul du jitter pour pouvoir observer sur mon corpus, l'évolution du taux de craquement en fonction de l'âge des locuteurs.

Une analyse acoustique des paramètres glottiques réalisée avec SuperVP m'a permis d'obtenir les GCI. Le calcul du jitter que j'ai implémenté se base sur le travail de Mireia Farrús [8].



$$\text{jitter} = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}|T0_{n+1}-T0_n|}{\frac{1}{N}\sum_{n=1}^{N}(T0_n)} \times 100$$

$$\text{jitter} = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}|gci_{n+1}-2gci_n+gci_{n-1}|}{\frac{1}{N}(gci_N-gci_0)} \times 100$$

Pour éviter d'obtenir des valeurs importantes de jitter au moment du saut entre deux segments voisés, je ne prends pas en compte les valeurs qui sont calculées pour des changements de T0 trop grands.

J'ai utilisé cette méthode sur les fichiers de mon corpus de voix, malheureusement je n'obtiens pas de résultats très cohérents.

Le problème sur les fichiers de mon corpus vient de l'analyse glottique. En effet, cette analyse nécessite des fichiers audio très propres pour donner des résultats cohérents. Pour tester l'efficacité de ma formule du jitter. J'ai, à partir d'un fichier propre, créer de l'apériodicité dans les GCI. La modélisation SVLN m'a permis de déplacer les pulses glottiques. J'ai donc ajouté un bruit gaussien aux positions des pulses, de moyenne nulle, et de variance plus ou moins forte. De cette façon j'ai obtenu des voix très craquées pour des variances fortes.

Le fichier non transformé obtient une valeur moyenne de jitter de 0.21
Pour les valeurs de variances suivantes j'ai obtenue les résultats suivants :
variance = 10% de la période : 30 %
variance = 30% de la période : 34 %
variance = 60% de la période : 39 %

Le résultats du jitter augmente bien avec la variance du bruit, ce qui est rassurant. Néanmoins, on a une assez faible augmentation du jitter pour une variance pourtant bien élevée. En vérifiant les positions des pulses glottiques, on s'aperçoit que l'analyse n'est pas très performante et n'arrive pas à détecter tous les pulses.

Paramètres retenus pour l'étude

Pour mon étude, je me suis basé sur le modèle source/filtre "avancé". L' analyse de la fréquence fondamentale des corde vocales F0 est réalisée par l'analyseur SWIPE [2].
L'enveloppe spectrale représenté par les coefficients cepstraux, est estimé par la méthode de True-Enveloppe [22].
Les paramètres glottiques sont basés sur le modèle LF-Rd [6]. Sont compris l'énergie du pulse glottique E, le coefficient de relaxation Rd, et le niveau de bruit dans la source RNL.
A ceci s'ajoute la position des pulses glotiques, et l'analyse du jitter pour qualifier une voix de craquée ou non.

Les paramètes de la source glottique peuvent avoir un impact très important sur la qualité vocale, cependant l'estimation de ces paramètres n'est pas aisée. En effet, l'analyseur n'est pas très robuste et le fait que mes fichiers soient bruités a limité les résultats.

2 Représentations statistiques pour la transformation de la voix

2.1 Modèles de mélange de gaussiennes

2.1.1 Généralités

Après avoir bien assimilé les caractéristiques importantes et les outils d'analyse correspondant, il a fallu s'intéresser au modèle statistique.

Modèle Gaussien

Unidimensionnelle
La loi normale est une des lois de probabilité les plus adaptées pour modéliser des événements aléatoires. Elle est définie par deux paramètres : Sa moyenne μ et sa variance σ^2. Sa densité de probabilité est donnée par :
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}\right)$

Multidimensionnelle
Lorsqu'on considère plusieurs paramètres, on doit généraliser la loi normale. La loi multi-normale est paramétrée par un vecteur $\boldsymbol{\mu}$ représentant le centre, et une matrice semi-définie positive $\boldsymbol{\Sigma}$, qu'on appelle matrice de covariance. La fonction de densité de probabilité, par analogie, est donnée par :

$f(x) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, N étant la dimension.

Un aspect important pour mes travaux est que si les paramètres sont indépendants, on a alors la matrice de covariance $\boldsymbol{\Sigma}$ qui est diagonale. Par exemple les coefficients cesptraux peuvent être considérés comme indépendants et nous n'avons donc pas à modéliser leur inter-corrélation.

Modèle de mélange de gaussiennes
Un modèle multi-gaussien (appelé couramment GMM, Gaussian Mixture Model), est un modèle mélange. On modélise une distribution par une somme de gaussiennes.
Pour un modèle de mélange de gaussiennes comprennant g composantes :

$g(x) = \sum_{k=1}^{g} \alpha_k f_k(x)$

Où f_k est la densité de probabilité de la k^{ime} gaussienne.

2.1.2 Application à l'acoustique

Rappelons les paramètres acoustiques qui nous intéressent :
La fréquence fondamentale de vibration des cordes vocales : F0
La représentation spectrale de la voix : le cepstre (61 coefficients)
Les paramètres glottiques : Rd, E
Le rapport signal bruit : RNL
L'apériodicité de la glotte : Jitter

Il est important pour bien estimer les paramètres de nos modèles de traiter les données obtenues lors des analyses. Ainsi on procède à une analyse du volume, avec l'aide de la fonction Floudness1770. On peut donc éliminer les résultats de mesures lors des silences.

Il faut aussi considérer uniquement les segments voisés pour la F0, les paramètres glottiques, le RNL et le Jitter. Ceci est fait à l'aide du résultat f0_confidence donné par l'analyse SWIPE. On fixe un seuil de confiance à partir duquel on considère les segments comme voisés et donc délivrant des résultats pertinents pour ces paramètres.
Pour la représentation spectrale, il n'est pas nécessaire de se contenter des segments voisés. Certes le spectre des sons non voisés est plus bruité et ne comprend pas d' harmoniques, mais le conduit vocal agit sur tout le signal de parole.

La difficulté majeure a été de représenter correctement le cepstre. L'état de l'art nous apprend comment représenter l'espace acoustique [21].
J'ai premièrement essayé d'appliquer un modèle GMM au cesptre. Pour bien représenter l'espace acoustique entier, il faut normalement prendre un nombre de gaussiennes important. Seulement, la taille de mes fichiers étant bien réduite, j'ai décidé de travailler avec beaucoup moins de gaussiennes, de l'ordre de 4, 8 ou 16. La difficulté avec le spectre est que les résonances qu'on appellent les formants, dépendent du phonème.
J'ai voulu réaliser un clustering de groupes de phonèmes pour pouvoir relativement bien représenter mon espace acoustique. Ma base de données n'étant pas complète en diversité de phonèmes et comportant des fichiers enregistrés dans des conditions à chaque fois différentes fait qu'il est assez difficile d'avoir une très bonne représentation. Mon but était de trouver une représentation assez simple.

Algorithme EM

L'algorithme EM (Expectation Maximisation) nous permet d'évaluer les paramètres des modèles par une méthode itérative [5]. Ce principe comporte deux étapes :
- Une étape d'évaluation de l'espérance de la vraisemblance des données sur le modèle.
- Une étape de maximisation, où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape précédente.

On utilise ensuite les paramètres trouvés lors de la deuxième étape comme point de départ d'une phase d'expectation. On réitère ainsi les étapes jusqu'à convergence de l'espérance.

2.1.3 Adaptation de modèles

Modèles Gaussiens

Lorsqu'on veut faire une conversion de voix, c'est à dire transformer la voix d'un locuteur source en celle d'un locuteur cible, on adapte les paramètres acoustiques grâce aux modèles statistiques. Pour les modèles gaussiens, il faut donc changer la moyennes et la variance des données. On applique donc la formule suivante :

$$f(x) = \nu + \frac{\gamma}{\sigma} \times (x - \mu)$$

Avec μ et ν respectivement les moyennes des modèles source et target. σ et γ les variances des modèles source et target.

GMM

Pour les modèles multi-gaussiens, on utilise une formule analogue au cas précédent :

$$f(x) = \sum_{n=1}^{N} p_n(x)(\nu_n + \frac{\gamma_n}{\sigma_n} \times (x - \mu_n))$$

Avec N le nombre de gaussiennes.
μ_n, ν_n, σ_n, γ_n les paramètres des n^{ime} gaussiennes des modèles source et target.
$p_n(x)$ la probabilité à posteriori de x sur le modèle n.

On s'aperçoit rapidement de la difficulté de l'application de cette méthode. Il faut que les gaussiennes des modèles source et target soient associées deux à deux. Or je dispose de fichiers non alignés, c'est à dire que les locuteurs ne prononcent pas les mêmes choses. Des méthodes d'alignement existent mais paraissent compliquées à appliquer pour un moteur de synthèse devant à terme fonctionner en temps réel. [15]

Pour associer les gaussiennes des modèles, j'ai essayé différents critères : Minimiser la distance de Bhattacharyya entre les gaussiennes. Minimiser la KL-divergence. [13] Et une méthode consistant à maximiser le clustering commun des données sur un fichier audio.

Ces méthodes donnent des résultats, et dans certains cas on arrive à obtenir des résultats correctes. Néanmoins on obtient souvent des artefacts dégradant le signal une fois transformé.

J'ai également essayé une technique de transformation originale mais sans succès : Faire une adaptation des gaussiennes sur toute les gaussiennes, et non pas associées deux à deux et pondérer chaque transformation par un facteur :

$b(cc, n, p) = \frac{p_{xn}(cc) \times p_{yp}(cc)}{Bhattacharyya_{np}}$

qu'on normalise ensuite pour avoir $\sum_{n,p} b(cc, n, p) = 1$

Où :

- cc est un vecteur de coefficients cesptraux, n et p sont respectivement les indices des gaussiennes des modèles source et target.
- $p_{xn}(cc)$ la probabilité à posteriori de cc sur la gaussienne n du modèle source.
- $p_{yp}(cc)$ la probabilité à posteriori de cc sur la gaussienne p du modèle target.
- $Bhattacharyya_{np}$ la distance de Bhattacharyya entre la gaussienne n du modèle source et la gaussienne p du modèle target.

Pour obtenir les transformations, je me suis concentré sur une méthode fondamentalement plus simple, des régressions polynomiales des paramètres des gaussiennes. Mais avec un principe de clustering de trajectoires.

2.2 Régression multiple de GMM

Les recherches actuelles utilisent des modèles de mélange de gaussiennes pour représenter l'espace acoustique. A partir du modèle cible et target joint préalablement établi, ils réalisent une adaptation des données [20].
Ce genre de méthode est utilisé en conversion de la voix, c'est à dire quand on veut changer l'identité d'un locuteur en celle d'un autre locuteur.

Les résultats obtenus ne sont pas très performants. D'une part la conversion de l'identité n'est pas toujours efficace. [19].

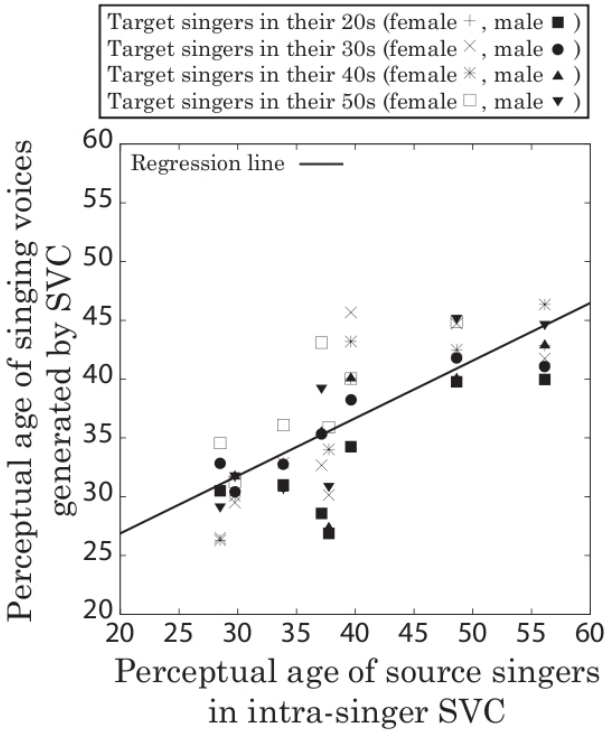


FIG. 6 – Performance d'une méthode de conversion de voix pour l'application du changement de l'âge

Mais il y a également des artefacts dû à la transformation qui viennent dégrader le signal de parole. En effet les modèlisations GMM ne donnent pas de toujours de transformations stable dans l'espace acoustique. Les problèmes viennent de la représentation spectrale du signal, où il est diffcile de segmenter l'espace en phonème. La mauvaise estimation des formants est également un problème pour l'élaboration d'un moteur de synthèse statistique basé sur des modèles GMM.

2.3 Mélange de régressions

Un des points nouveaux de mon travail a été d'utiliser une méthode de régressions multiples. En effet le fait de pouvoir trouver différentes trajectoires d'évolution des paramètres séduit car cela pourrait permettre d'extraire différentes transformations pour faire vieillir la voix. On remarque aisément que les voix ne vieillissent pas toutes de la même façon selon les locuteurs. On peut ainsi faire des distinctions entre genre, ou type de vieillissement comme par exemple l'apparition de craquement ou de souffle dans la voix.

Pour réaliser ces régressions multiples avec clustering de trajectoires, j'ai utiliser un principe basé sur une estimation de paramètres de gaussiennes basé sur l'algorithme EM. [7] [3] [9]

Le modèle est donné comme suit :

$$y_i = \begin{cases} x_i^T\beta_1 + \epsilon_{i1} & ; \quad \text{avec une probabilité } \pi_1 \\ x_i^T\beta_2 + \epsilon_{i2} & ; \quad \text{avec une probabilité } \pi_2 \\ \quad \vdots & \\ x_i^T\beta_J + \epsilon_{iJ} & ; \quad \text{avec une probabilité } \pi_J \end{cases}$$

où y_i est la variable image de la $i^{\text{ème}}$ observation ; x_i^T (i=1,...,n) correspond à la transposée du vecteur de variables aléatoires indépendantes de la $i^{\text{ème}}$ observation, de dimension p+1 ; β_j (j=1,...,J) correspond au vecteur de régression pour la composante j ; π_j sont les probabilités des mixtures. Et ϵ_{ij} sont les erreurs aléatoires. Sous l'hypothèse de loi normale, on a $\epsilon_{ij} \sim N(0, \sigma_j^2)$

On cherche donc à estimer les paramètres $\theta = (\pi_1, ..., \pi_J, \beta_1, ..., \beta_J, \sigma_1^2, ..., \sigma_J^2)$. L'algorithme EM est parfait pour estimer ce genre de paramètres dans un cas gaussien. Un implémentation de l'algorithme a été faite en matlab. Mais j'ai par la suite trouvé un toolbox permettant de faire des regressions multiples, et avec des méthodes d'alignement.

Exemple d'utilisation sur des données générées à la main

A titre d'exemple pour mettre en avant les avantages d'une telle méthode de régression, j'ai généré des valeurs suivant trois différentes droites, mais avec un bruit ajouté. J'utilise maintenant le principe mélange de régression à l'ordre un, et avec trois clusters.

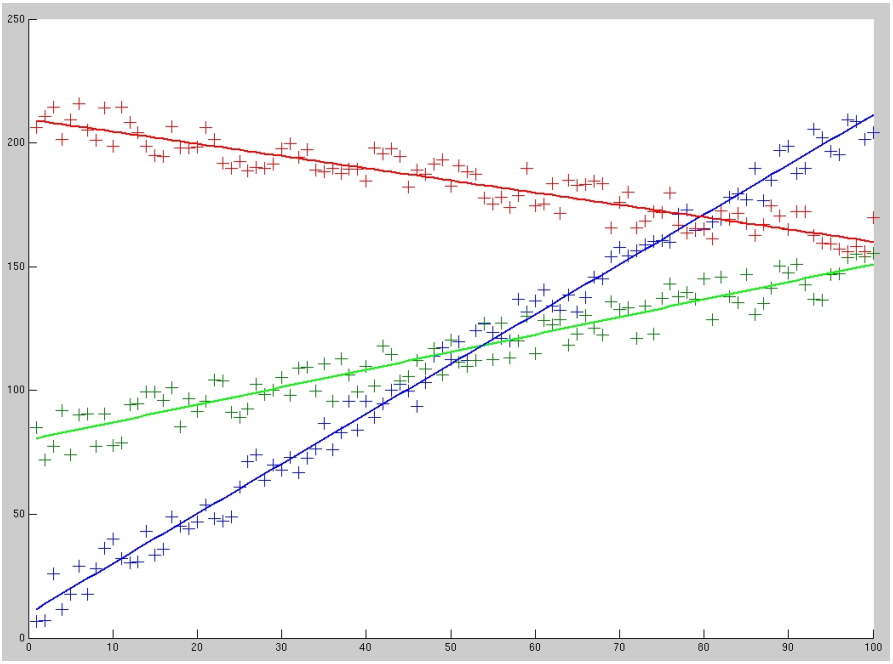


FIG. 7 – Régression multiple sur un exemple de données générées à la main

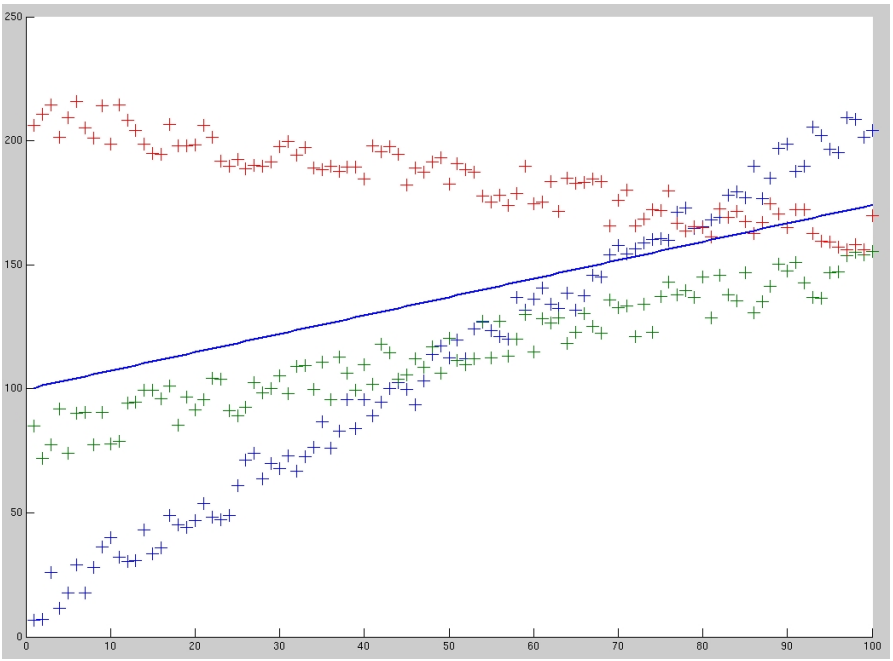


FIG. 8 – Régression simple sur un exemple de données générées à la main

Exemple de résultats obtenus sur les paramètres acoustiques

Pour représenter l'évolution de la F0 moyenne des locuteurs, j'ai utilisé une régression polynomiale d'ordre 4, avec 2 clusters sur le logarithme de la fréquence. On observe que le clustering est efficace puisque les deux groupes de regression correspondent pour l'un aux femmes et pour l'autres aux hommes. On a alors deux trajectoires différentes à utiliser selon le sexe du locuteur. On remarque cependant quelques défauts des régressions polynomiales quand on veut représenter la F0 pour des âges en dehors de la tranche 20-70 ans. Les polynomes tendent vers des valeurs élevées en valeur absolue.

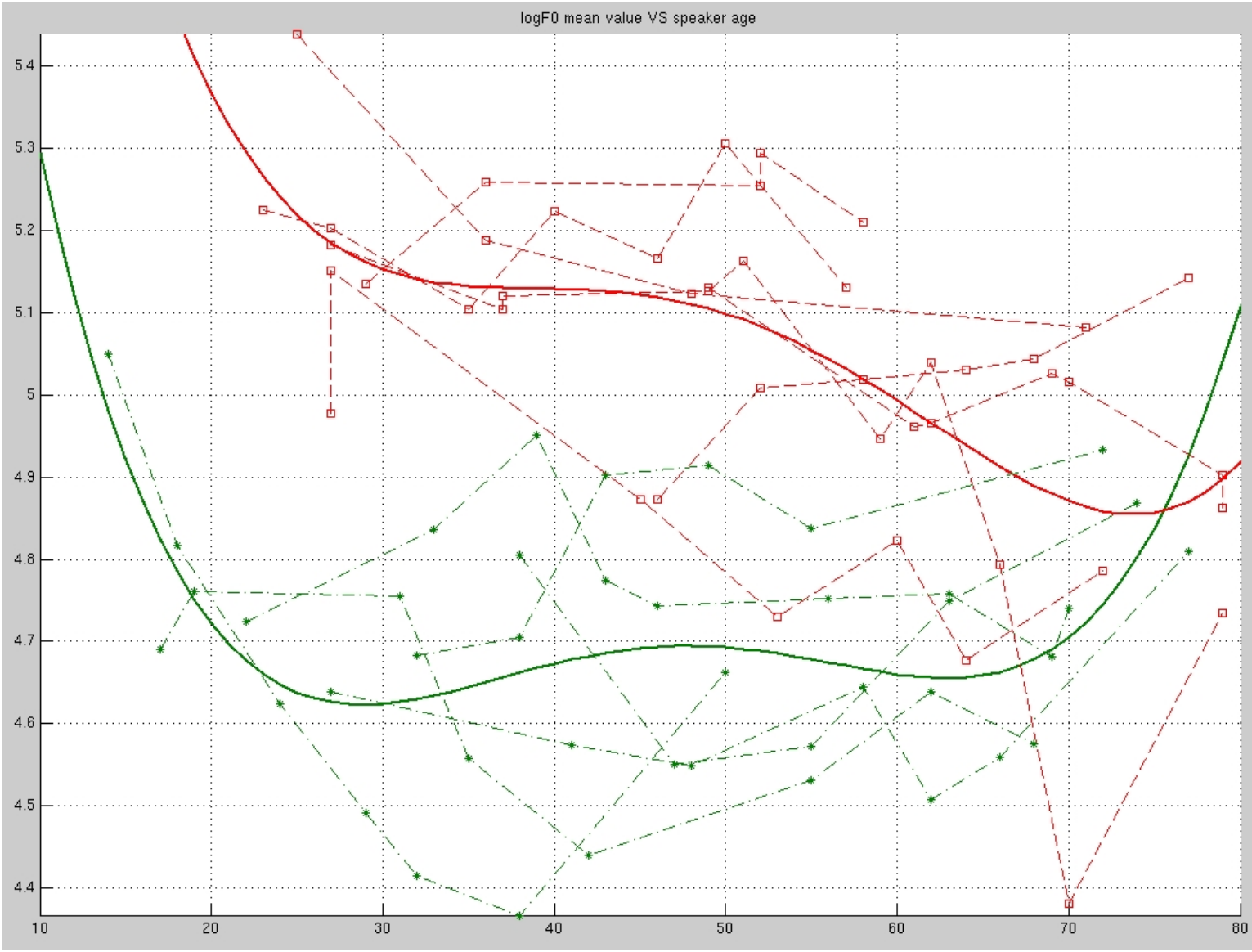


FIG. 9 – Régression multiple des trajectoires de F0

Ainsi pour tous les paramètres retenus pour l'étude, j'ai pu observer les tendances d'évolution suivantes :

	20	40	60	80
F0 (homme)	++	=	=	+
F0 (femme)	++	+	=	-
F0 var	=	+	++	+
SNR	+	=	=	+
Rd	+	=	=	+
E	+	=	=	+
GCI 1	=	+	=	=
GCI 2	=	=	+	++

Comme annoncé par l'état de l'art, la fréquence fondamentale F0 diminue au cours de notre vie. Pour les hommes, on constate une légère augmentation de la valeur moyenne lorsqu'on dépasse 60 ans.

Il est intéressant de noter qu'on a obtenu deux trajectoires d'évolution du paramettre de jitter (l'apériodicité des instants de fermeture glottique). La première tendance observée correspond à celle d'un individu sans particularité : la voix devient un peu craqué vers les 30 ans, et ensuite elle redevient claire. La deuxième tendance met en avant le caractère craquée apparaissant lorsque le locuteur devient âgé. Comme exemple pris du corpus, Marguerite Duras suit typiquement cette évolution.

3 Moteur de synthèse SVLN

J'ai implémenté un moteur de synthèse sur Matlab, à partir des différents analyseurs, du modèle source/filtre avancé, et des résultats obtenus. Le moteur de synthèse a nécessité l'élaboration d'une étape d'entrainement réalisée grâce au corpus. L'étape de transformation se sert ensuite des résultats obtenus pour pouvoir établir le modèle cible en fonction du modèle source et de l'âge qu'on veut donner au locuteur.

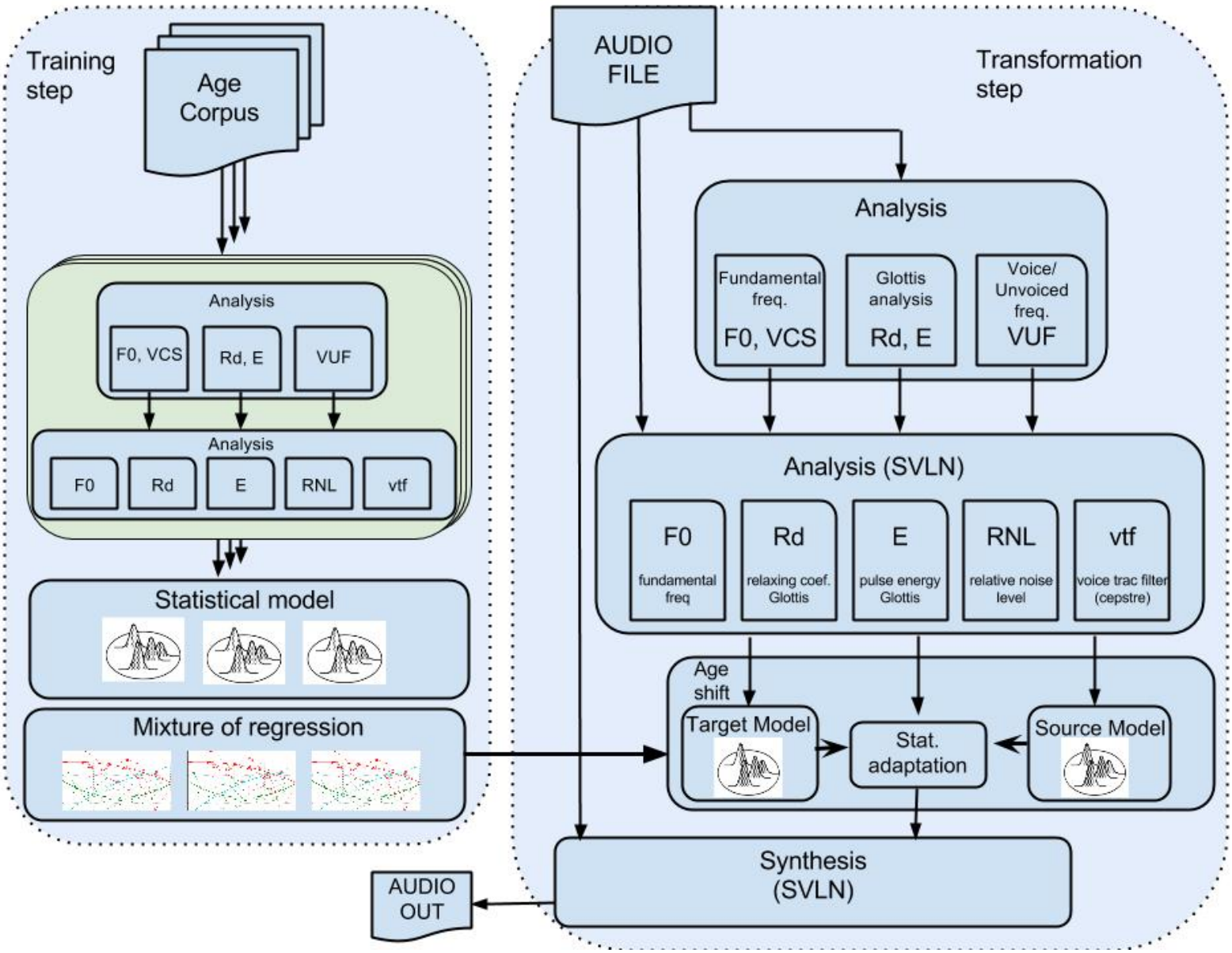


FIG. 10 – Principe de Transformation

3.1 Transformation

Pour réaliser les transformations des paramètres donnés par SVLN, j'ai réalisé une adaptation des paramètres du modèle source du locuteur, sur un modèle target obtenu grâce aux regressions réalisées dans la phase d'apprentissage. Pour trouver les paramètres du modèle target, je leur ai fait suivre l'évolution donnée par les polynomes obtenues précédemment. Ainsi pour une regression au premier ordre on a :

$$\begin{aligned} F(x) = &\quad a \times x + b \\ F(T+X) = &\quad a \times T + a \times X + b = F(T) + A \times X \end{aligned}$$

Au second ordre :

$$\begin{aligned} F(x) = &\quad a \times x^2 + b \times x + c \\ F(T+X) = &\quad a \times (T+X)^2 + b \times (T+X) + c \\ F(T+X) = &\quad F(T) + a \times X \times (X + 2T) + b \times X \end{aligned}$$

Au troisième ordre :

$$\begin{aligned} F(x) = &\quad a \times x^3 + b \times x^2 + c \times x + d \\ F(T+X) = &\quad F(T) + \\ &\quad a \times X \times ((T+X)^2 + (T+X) \times T + T^2) + \\ &\quad b \times X \times (X + 2T) + \\ &\quad c \times X \end{aligned}$$

Au quatrième ordre :

$$\begin{aligned} F(x) = &\quad a \times x^4 + b \times x^3 + c \times x^2 + d \times x + e \\ F(T+X) = &\quad F(T) + \\ &\quad a \times X \times ((T+X)^2 + T^2) \times (X + 2T) + \\ &\quad b \times X \times ((T+X)^2 + (T+X) \times T + T^2) + \\ &\quad c \times X \times (X + 2T) + \\ &\quad d \times X \end{aligned}$$

Avec F le polynôme obtenu par régression sur un des paramètres vocaux, T l'âge du locteur et X le changement d'âge que l'on veut réaliser.
Cette méthode permet de ne pas prendre le terme constant des polynomes, et ainsi ne considérer que la variations des paramètres entre l'âge supposé du locuteur et l'âge cible.

3.2 Synthèse de craquement dans la voix

Comme on l'a vu dans la partie 3, le craquement vient d'une apériodicité dans la position des pulses glottiques (GCI). Ce craquement a été difficile à mesurer et quantifier. Néanmoins, il semble important dans l'évolution des voix des locuteurs. J'ai donc implémenté dans le moteur de synthèse SVLN, une fonction pour ajouter du craquement. J'ai commencé d'abord par ajouter un bruit aux positions des pulses. En faisant de cette façon, on obtient une voix craquée qui parait assez irréelle. En effet, le craquement s'applique alors à tout le fichier audio. En me basant sur des études faites sur la prosodie, j'ai segmenté la voix par registre du pitch. [18]

Le craquement semble apparaître souvent sur les parties graves des paroles. Sur les parties où la F0 est faible. J'ai utilisé un modèle séparant la F0 en 3 domaines : High Medium Low. Le domaine High commence à partir de 2 semi-tons au dessus de la valeur moyenne du pitch sur le fichier. Le domaine Low commence à partir de 2 semi-tons en dessous de la valeur moyenne.

J'ai ensuite appliqué la méthode du craquement lorsqu'on se situait dans le registre Low. Pour rendre l'apparition du craquement moins brutale, j'ai modulé la valeur de variance du bruit par un coefficient. J'ai pris la forme du sigmoïde pour faire passer la variance de 0 à sa valeur maximale en fonction de la différence de pitch avec le pitch moyen.

On définit d'abord le pitch en semi-tons : $F0_{st} = 12 \times \log_2 \frac{F0}{F0mean}$

Et le coefficient de modulation : $alpha = 1 - \frac{1}{1+\exp(3\times(-F0_{st}-1))}$

Les coefficients dans l'exponentielle sont déterminés de façon à avoir une valeur alpha très proche de 1 lorsqu'on arrive à $F0_{st} = -2st$, et proche de zéro à $F0_{st} = 0st$.

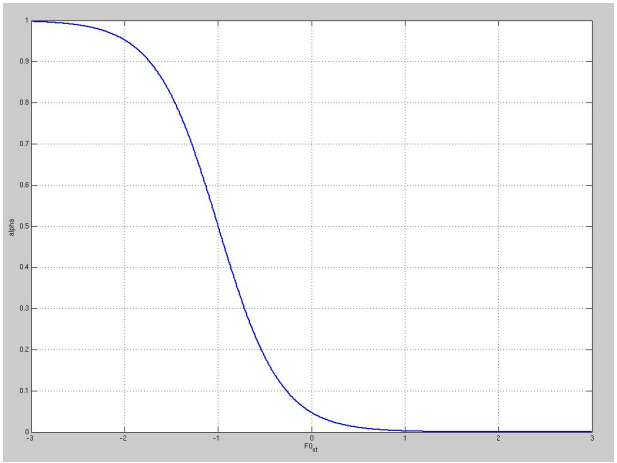


FIG. 11 – Coefficient de modulation en fonction de l'écart au pitch moyen

4 Résultats

4.1 Test perceptif

Pour évaluer la méthode mis en place lors de ce travail, il a fallu réaliser un test perceptif. A l'aide de mon moteur de synthèse, j'ai transformé la voix de locuteurs pour changer l'âge perçu dans leur voix. Le but étant d'évaluer l'intérêt de l'utilisation des paramètres glottiques, j'ai transformé les voix en utilisant différents jeux de paramètres :

- (F0, enveloppe spectrale)
- (F0, enveloppe spectrale, Rd)
- (F0, enveloppe spectrale, SNR)
- (F0, enveloppe spectrale, GCI)
- (F0, enveloppe spectrale, Rd, SNR, GCI)

Ne connaissant pas l'âge du locuteur, je l'ai supposé égale à 35 ans. J'ai transformé les voix également selon plusieurs âges cibles : 15 ans, 25 ans, 45 ans, 55 ans, 65 ans, 75 ans. Cela nous fait déjà $5 \times 6 = 30$ écoutes, auxquelles j'ai rajouté le fichier original et le fichier analysé puis synthétisé sans transformation. Donc au total les participants à l'expérience devront écouter 32 fichiers.

Pour chaque fichier, il faut répondre à deux questions :

- Quel est l'âge perçu dans la voix du locuteur ? Avec comme réponses possibles les tranches 10-20 ans, 20-30 ans, 30-40 ans, 40-50 ans, 50-60 ans, 60-70 ans, 70-80 ans.
- A quel point le fichier audio sonne-t-il naturel ou synthétique ? Avec comme réponses possibles
  - Excellent pour une synthèse imperceptible.
  - Good pour une synthèse perceptible mais non génante.
  - Fair pour une synthèse légèrement génante.
  - Poor pour une synthèse génante.
  - Bad pour une synthèse très génante.

Nicolas Obin a implémenté le test en ligne, et pour ce rapport, j'ai pu avoir 8 participants. Un autre test perceptif est envisagé, pour montrer un autre résultat : l'intérêt des mélanges de régressions par rapport aux régressions simples.

4.2 Résultats

Les résultats du test sont encourageants. Ils montrent comme attendu l'intérêt de l'utilisation d'un modèle de la source glottique. Voici les résultats de corrélation entre l'âge cible et l'âge perçu du locuteur, par jeux de paramètres, ainsi que les scores de qualité de synthèse en moyenne.

jeux de paramètres	Coef de corrélation
(F0, env spec)	0.69
(F0, env spec, GCI)	0.70
(F0, env spec, SNR)	0.55
(F0, env spec, Rd)	0.44
(F0, env spec, Rd, SNR, GCI)	0.80

Fichiers	Score de qualité
Original	3.38
Synthèse	3.21

Voici les graphiques représentant l'âge perçu du locuteur sur les fichiers audio synthétisés en fonction de l'âge cible.

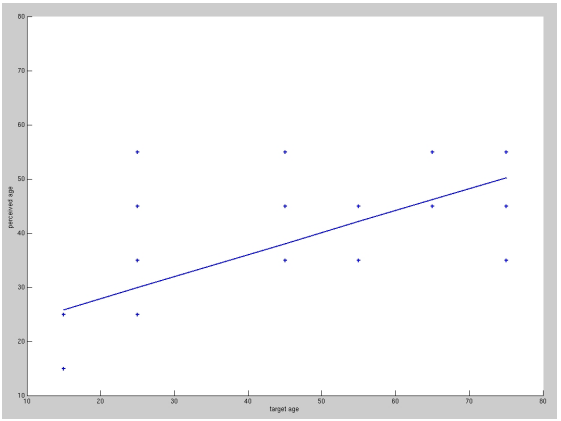


FIG. 12 – Performance de la méthode de transformation pour le jeux de paramètres (F0, enveloppe spectrale)

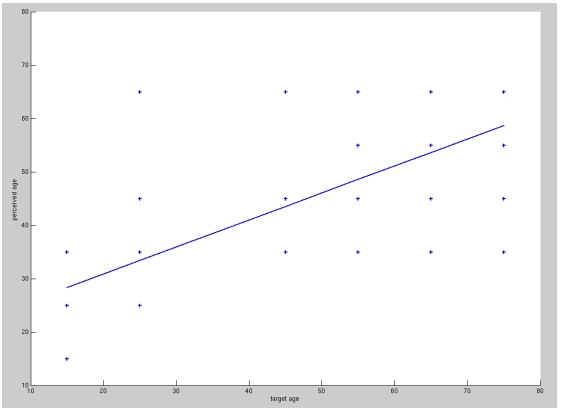


FIG. 13 – Performance de la méthode de transformation pour le jeux de paramètres (F0, enveloppe spectrale, GCI)

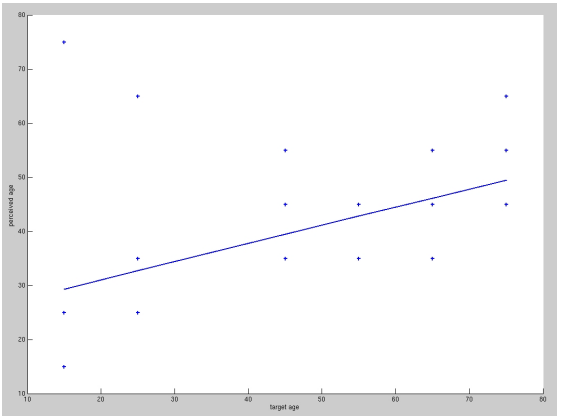


FIG. 14 – Performance de la méthode de transformation pour le jeux de paramètres (F0, enveloppe spectrale, SNR)

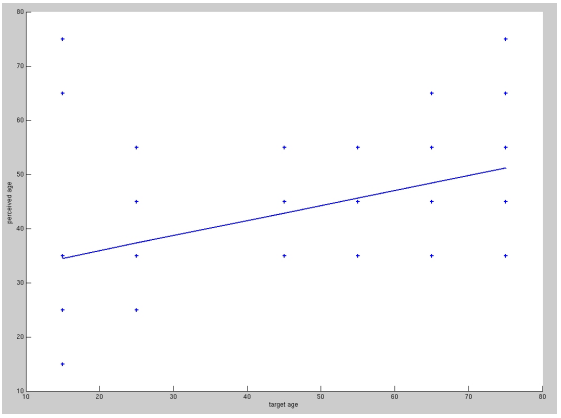


FIG. 15 – Performance de la méthode de transformation pour le jeux de paramètres (F0, enveloppe spectrale, Rd)

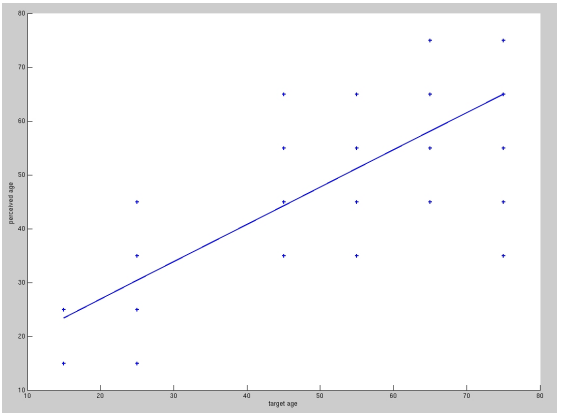


FIG. 16 – Performance de la méthode de transformation pour le jeux de paramètres (F0, enveloppe spectrale, Rd, SNR, GCI)

Ces résultats montrent plusieurs choses. Tout d'abord, les résultats de corrélation montrent qu'il y a une forte corrélation entre l'âge perçu du locuteur et l'âge cible. Ensuite, le résultat de corrélation pour le jeux de paramètre complet est le meilleurs. Cela montre que l'utilisation des paramètres glottique apporte réellement quelquechose sur l'âge perçu dans la voix. Par ailleurs, on s'apperçoit que pour le jeu de paramètres (F0, enveloppe spectrale, Rd), le résultat de corrélation est le plus faible. Cela montre l'importance de l'utilisation d'une représentation complète du modèle de synthèse vocale. Enfin, les résultats sur la qualité de la transformation montrent que le son n'est pas très dégradé par le moteur de synthèse, contrairement aux méthodes de conversion utilisant les représentations GMM, pour lesquelles il est difficile d'avoir une synthèse propre.

Conclusion

Ce travail sur la transformation statistique de la voix, et l'application au changement de l'âge perçu du locuteur a permis d'amener des réponses. L'exploration d'une représentation acoustique de la voix pour prendre en compte les paramètres glottiques liés à la qualité vocale a fait ses preuves lors d'un premier test perceptif. Grâce aux travaux de recherche actuels sur la glotte, il sera certainement possible de réaliser de meilleurs analyses des paramètres, et ainsi rendre encore plus exploitable mon travail. De plus, nous avons pu mettre en place un principe pour extraire les différentes façons de transformer les paramètres vocaux, et pouvoir ainsi se rendre compte des différentes façons de vieillir la voix.

Ce stage a été très intéressant et m'a beaucoup intéressé, j'ai pu découvrir le département d'Analyse et synthèse des sons de l'IRCAM, ce qui m'a permis de comprendre un peu plus le fonctionnement de la recherche. J'ai rencontré des personnes compétentes et généreuses, qui m'ont donné envie de poursuivre dans cette voie. J'ai aussi découvert un monde où des scientifiques et des artistes se réunissent autour de projets communs, ce qui m'encourage à poursuivre dans cette voix. Je compte d'ailleurs suivre la formation du Master ATIAM l'an prochain.

Références

[1] Markus Bruckl and Walter Sendlmeier. Aging female voices : An acoustic and perceptive analysis. In *ISCA Tutorial and Research Workshop on Voice Quality : Functions, Analysis and Synthesis*, 2003. 10

[2] Arturo Camacho. *SWIPE : A sawtooth waveform inspired pitch estimator for speech and music*. PhD thesis, University of Florida, 2007. 15

[3] Gerben de Vries and Maarten van Someren. Clustering vessel trajectories with alignment kernels under trajectory compression. In *Machine Learning and Knowledge Discovery in Databases*, pages 296–311. Springer, 2010. 22

[4] Gilles Degottex. Glottal source and vocal-tract separation. *IRCAM, Paris*, 2010. 13

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 18

[6] Gunnar Fant. The lf-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3) :40, 1995. 11, 15

[7] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2) :201–225, 2010. 22

[8] Mireia Farrus, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *INTERSPEECH*, pages 778–781, 2007. 13

[9] Scott J Gaffney and Padhraic Smyth. Joint probabilistic curve clustering and alignment. In *Advances in neural information processing systems*, pages 473–480, 2004. 22

[10] James D Harnsberger, William S Brown Jr, Rahul Shrivastav, and Howard Rothman. Noise and tremor in the perception of vocal aging in males. *Journal of Voice*, 24(5) :523–530, 2010. 10

[11] Jonathan Harrington, Sallyanne Palethorpe, and Catherine I Watson. Age-related changes in fundamental frequency and formants : a longitudinal study of four speakers. In *INTERSPEECH*, pages 2753–2756, 2007. 10

[12] F Sean Hodge, Raymond H Colton, and Richard T Kelley. Vocal intensity characteristics innormal and elderly speakers. *Journal of Voice*, 15(4) :503–511, 2001. 10

[13] Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1) :52–60, 1967. 19

[14] J Laver. The analysis of vocal quality : from the classical period to the twentieth century. *Towards a history of phonetics*, pages 79–99, 1981. 8

[15] Larbi Mesbahi. *Transformation automatique de la parole-Etude des transformations acoustiques*. PhD thesis, Université Rennes 1, 2010. 19

[16] S Mwangi, W Spiegl, F Honig, T Haderlein, A Maier, and E Noth. Effects of vocal aging on fundamental frequency and formants. In *Proceedings of the International Conference on Acoustics NAG/DAGA*, pages 1761–1764, 2009. 10

[17] Binh Phu Nguyen and Masato Akagi. Spectral modification for voice gender conversion using temporal decomposition. *Journal of Signal Processing*, 2007. 10

[18] Nicolas Obin, Julie Beliao, Christophe Veaux, Anne Lacheret, et al. Slam : Automatic stylization and labelling of speech melody. *Speech Prosody 7*, pages 246–250, 2014. 28

[19] Kumi Ohta, Tomoki Toda, Yamato Ohtani, Hiroshi Saruwatari, and Kiyohiro Shikano. Adaptive voice-quality control based on one-to-many eigenvoice conversion. 2010. 21

[20] Tomoki Toda. Eigenvoice-based approach to voice conversion and voice quality control. 2009. 21

[21] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8) :2222–2235, 2007. 17

[22] Fernando Villavicencio, Axel Robel, and Xavier Rodet. Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006. 9, 15

[23] Ravichander Vipperla, Steve Renals, and Joe Frankel. Ageing voices : The effect of changes in voice parameters on asr performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010 :5, 2010. 10