Rapport de stage

Modification expressive de voix chantée

Stagiaire :
Maxime Dickerson
Encadré par :
Axel Roebel
Luc Ardaillon
Organisme d'accueil :
IRCAM - Équipe Analyse/Synthèse

Master ATIAM

Année universitaire 2015/2016

Résumé

Le travail présenté dans ce document concerne l'effort vocal dans la synthèse de voix chantée, son réalisme et son contrôle. Son but est de mettre au point une règle de modification de la source de la voix et de préciser l'estimation des formants, les résonances du conduit vocal, dans le cadre du chant pour permettre leur manipulation. Pour cela, une règle liant les paramètres R_d et E_e du modèle de source glottique de Liljencrants-Fant [1] est utilisée et paramétrée puis testée perceptivement par vingt-trois participants. Une méthode de descente de gradient est introduite pour estimer les paramètres des formants d'un modèle du filtre du conduit vocal à quatre filtres d'ordre quatre en parallèle. Une variante de cette méthode est évaluée et comparée au logiciel de traitement de parole Praat sur des enregistrements de chant. Les résultats du test perceptif ne permettent pas de valider ou d'invalider la règle de modification et la comparaison de la méthode d'estimation des formants montre des résultats proches du logiciel Praat pour les enregistrements étudiés. Un second test perceptif pourra tenter de mieux évaluer la règle de modification de la source et la méthode de descente de gradient pourra être évaluée plutôt que sa variante, plus rapide mais moins efficace.

Abstract

This document deals with voice effort control in a source-filter singing synthesis engine. It aims at designing modification rules for the voice source and estimating formants, the vocal tract resonances, in order to manipulate them. A rule based on the R_d shape parameter and E_e energy of the Liljencrants-Fant glottic source model is used, enhanced and has been perceptually tested by 23 people. A gradient descent based method is proposed to estimate formant parameters, according to four fourth order filters in a parallel structure. A derived and quicker method is evaluated and compared with the speech processing software, Praat, on singing recordings. The perception test results do not allow to reject or retain the source modification rule and the formant estimation method shows promising similarities with the Praat software on the studied recordings. Another perception test could attempt at correcting the first one in evaluating the source modification rule and the estimation formant method could be evaluated, instead of its less efficient yet quicker derived version.

Remerciements

Je tiens d'abord à remercier mon tuteur de stage, Axel Roebel, pour m'avoir confié ce sujet passionnant et très riche lié à la synthèse et la modification de la voix chantée, et pour tous les conseils scientifiques et méthodiques que j'ai pu recevoir. Ses nombreuses explications sur la production de la parole et sur les coûts de calcul des différentes méthodes utilisées ont été d'une aide précieuse.

Je me dois de remercier Luc Ardaillon sans qui ce travail n'aurait pas pu avoir lieu. Son aide et ses conseils tout au long du stage ont été particulièrement utiles tout comme les connaissances en matière de synthèse de voix chantée qu'il a partagées.

Enfin je tiens à remercier l'équipe Analyse-Synthèse pour son accueil chaleureux et enrichissant, ainsi que tous les stagiaires ATIAM (entre autres) et les thésards pour tous les bons moments que j'ai passés durant ce stage et en dehors.

Table des matières

Notations & abréviations

LF	Modèle de source glottique de la voix de Liljencrants-Fant
SVLN	Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise
PaReSy	Deterministic plus Stochastic Model-based Parametric Re-Synthesis

f_0	Fréquence fondamentale
f_{VU}	Fréquence de séparation voisé/non voisé
R_d	Paramètre de forme des impulsions glottiques du modèle LF
E_e	Norme des impulsions glottiques du modèle LF
VTF	Filtre du conduit vocal
f_i	Fréquence centrale du i-ième formant
BW_i	Bande passante du i-ième formant

Introduction

L'étude du chant et de la parole appellent à l'analyse de la production de la voix, de ses modifications et de sa modélisation dans le but de la synthétiser. Les travaux de recherche effectuées jusqu'à présent permettent de proposer un nombre d'outils de création et d'analyse conséquent. Des modèles de haute qualité ont été mis au point pour permettre notamment le traitement automatique de la parole pour en extraire le contenu du discours ou la génération de segments chantés. Ces derniers ont pu trouver une place dans l'industrie musicale japonaise à travers le logiciel Vocaloïd [2].

Bien que la voix soit à l'origine de la parole et du chant, les mécanismes en jeu dans ces phénomènes diffèrent, tout comme les aspects qui sont étudiés. Dans le domaine de la parole, tout ce qui contient le sens du discours, les mots, est le centre d'intérêt principal des méthodes d'analyse. Dans le domaine de la voix chantée, une plus importante partie des études effectuées concernent tout ce qui n'est pas sémantique, comme le timbre, la qualité vocale. Le chant démontre plus de variations d'intensité, de modulations et d'effets tels que le *vibrato* ou plus radicalement distincts de la parole des effets tels que la voix rauque, cassée ou presque inhumaine que l'on retrouve dans les musiques du genre *death metal*.

Les intérêts de la synthèse de chant sont variés. Elle représente un défi pour la communauté scientifique tant le chant possède de facettes différentes dans des styles musicaux différents, comme les prestations de professionnels de l'opéra, les chanteuses de musique populaire aux voix puissantes d'une manière encore différente appelée le *belting*, les chants diphoniques mongoles ou encore les différents types de cris et de voix rauques de rock extrèmes. À cette multitude de potentiel de recherche s'ajoute la possibilité de création. Il s'agit de créer des instruments aussi précis et flexibles que possible permettant la synthèse de chant et la modification de chant, aussi fidèle que souhaité.

L'objectif de ce stage est d'apporter de nouvelles règles offrant plus d'expressivité dans la synthèse de la voix chantée, notamment à travers la reconstruction et la modification des variations d'intensité, d'effort, dont la voix chantée est capable.

Cadre

Le projet ANR ChaNTeR a débuté en janvier 2014, pour une durée de 42 mois. Les partenaires de ce projet sont le LIMSI (coordinateur), ACAPELA, DUALO et l'IRCAM. Son objectif est de réaliser un système de synthèse concaténative de voix chantée de haute qualité, pour la langue française. Ce système doit être capable de synthétiser du chant à partir d'un texte et d'une mélodie (mode *chant à partir du texte*), et pourra aussi être utilisé comme un instrument, à travers un mode appelé *chanteur virtuel*. Dans ce second mode, l'utilisateur contrôle le chant en temps réel au moyen d'interfaces de captures de mouvement. Différents styles sont concernés : lyrique, classique ou encore populaire. Le système, devenu instrument dans la main de l'utilisateur, pourra offrir une nouvelle approche du chant aux compositeurs, artistes ainsi qu'à un plus large public.

Ce stage a été effectué à l'IRCAM, du 2 février au 30 juin 2016, sous la direction d'Axel Röbel qui dirige l'équipe Analyse-Synthèse de l'institut. Il s'inscrit dans la construction du mode *chant à partir du texte* et vise à améliorer le réalisme de la synthèse pour l'application de variations

d'intensité (qui peuvent être déterminées par l'utilisateur ou bien générées automatiquement).

Objectifs

Les objectifs sont donc d'analyser la base de données de segments chantées par des professionnels pour en capturer la nature des variations d'effort et d'intensité, en s'inspirant des études de la littérature. Puis il s'agit de les reproduire en laissant l'utilisateur du programme maître de cet effet. Ce travail a fait appel à la mise en place d'outils de traitement du signal usant d'une modélisation source-filtre du chant, en divisant le développement des règles de modifications pour la source puis pour le filtre du conduit vocal.

L'état de l'art de la synthèse du chant, de son expressivité et de la détection de formants, considérés comme les résonances du conduit vocal, est exposé dans une première partie, puis la méthode de modification de la source du modèle est étudiée et évaluée dans une deuxième partie. Dans une troisième partie, le processus d'estimation des formants mis au point dans le cadre de ce travail est présenté et évalué.

Chapitre 1

Etat de l'art

La synthèse de voix chantée a été la source d'un nombre croissant de recherches depuis les années 60. Dans ce chapitre, nous présenterons différentes méthodes qui ont été mises au point, en particulier celle qui est à la base de l'algorithme de synthèse du laboratoire. Puis nous introduirons les notions d'expressivité dans lesquelles s'introduit ce travail et nous terminerons par une brève présentation de l'estimation des formants, problématique qui s'est avérée indispensable à notre étude.

1.1 Systèmes de synthèse de voix chantée

La majorité des méthodes de synthèse de chant sont issues des méthodes de synthèses de paroles. Le modèle source-filtre y est largement répandu. Il est détaillé dans [3] et consite à modéliser la production de la voix en plusieurs éléments spectraux linéairement associés : la source glottique, le filtre du conduit vocal et potentiellement le rayonnement depuis la bouche jusqu'à l'appareil de captation. Le spectre de la parole peut donc être formulé de la manière suivante :

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot L(\omega) \tag{1.1}$$

où $S(\omega)$ est le spectre du signal, $G(\omega)$ le spectre de la source glottique, $C(\omega)$ le filtre du conduit vocal et $L(\omega)$ la fonction de transfert du rayonnement au niveau des lèvres. La source modélise les variations d'onde de débit glottique permises par les fermetures et ouvertures des cordes vocales. Elle peut être un train d'impulsions ou plus fidèlement un signal paramétré tel que le modèle de Liljencrants-Fant [1]. Le filtre du conduit vocal contient les formants, ses résonances.

1.1.1 Méthodes de synthèse

Les différentes approches de la synthèse de chant sont présentées et comparées dans les articles de Rodet [4] et d'Umbert [5]. Dans ce dernier, les différentes approches pour le contrôle de l'expressivité sont également décrites.

Un premier type de méthodes est la synthèse par modèle physique. Ce procédé se base sur la représentation physiologique du phénomène de la production du chant. Les paramètres utilisés pour cette synthèse sont les caractéristiques du conduit vocal : la longueur, la position de la langue, ainsi que des paramètres concernant les cordes vocales comme sa tension et tente de reproduire avec le moins de modélisation possible le chant. Le désavantage est le fait que beaucoup d'approximations doivent être effectuées pour obtenir des équations manipulables, rendant le signal peu naturel. Aussi, le contrôle des modèles physiques nécessite la specification de paramètres assez complexes (position de la langue, par exemple) qui sont actuellement difficiles à produire.

Les méthodes de synthèse formantique sont basées sur la représentation des résonances du conduit vocal. Physiquement, chacun de ces formants peut être relié à une propriété physique du conduit vocal, comme la position de la langue. Une telle méthode de synthèse présente des filtres de formants et une source et doit contenir un grand nombre de règles contrôlant leur variation en fréquence, en amplitude et en bande passante. Chaque voyelle introduite dans le système doit avoir été analysée pour être resynthétisée. Les limites de ce système se trouvent au niveau des consonnes et des transitions entre consonnes et voyelles et nécessitent un nombre de règles pouvant être considérable.

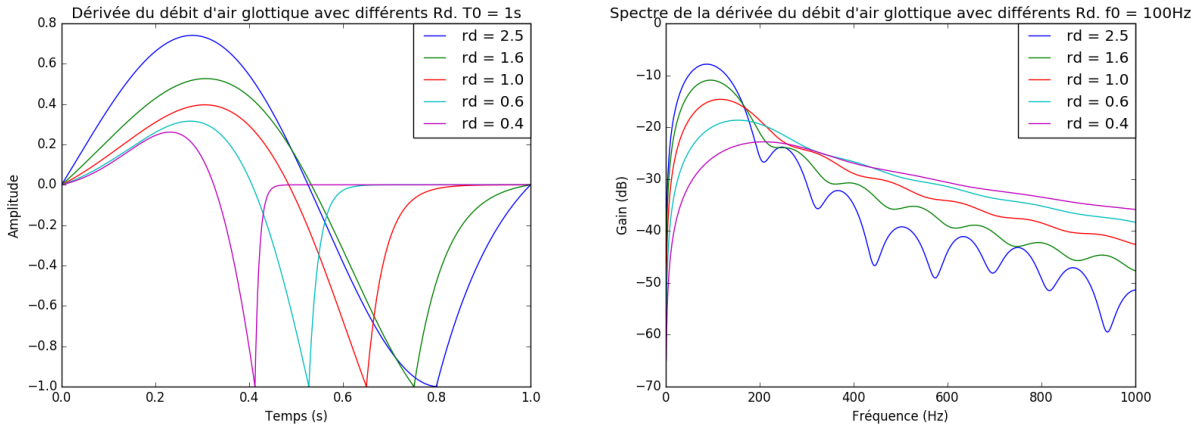
Les méthodes de synthèse pas modèle statistique se basent essentiellement sur des Modèles de Markov Cachés (synthèse par HMM). A l'aide d'une base de données de segments chantées et analysées, les variations de paramètres tels que la fréquence fondamentale, l'intensité ou l'enveloppe spectrale sont appris en association avec des styles et des contextes différents, c'est-à-dire les variations de notes, les mots ou encore les phrases. Cette apprentissage permet ensuite, à partir d'une mélodie et de paroles, de générer chacun des paramètres nécessaires à la synthèse. Ces méthodes peuvent utiliser un grand nombre d'aspects et de paramètres différents simultanément mais souffrent d'un lissage de l'information plus la base de données est grande. Il est donc plus difficile de capturer un style particulier à l'aide de cette méthode et d'atteindre un haut niveau de naturel.

Les méthodes dont les résultats sont les plus proches de la réalité sont, pour l'instant, les systèmes de synthèse concaténative comme celui sur lequel est basé ce travail. La source de ce succès, qui rend ces méthodes commercialisables depuis plusieurs années auprès d'artistes, de compositeurs, est la base de données utilisée pour synthétiser le chant. Elle est constituée d'enregistrements de chanteurs professionnels, un ou plusieurs, dont des segments, appelés unités, sont transformés, lissés et joints pour obtenir le signal de voix chantée. La création de la base de données, des règles de sélection des unités et le choix des méthodes de concaténation sont des étapes indispensables à l'élaboration de ces méthodes. En effet, la langue va définir la nature des phonèmes et donc des unités nécessaires. La sélection d'unités doit permettre un moindre coût des transformations et les méthodes de transformation comme les transpositions doivent être réalistes en altérant correctement la source et le filtre du conduit vocal. La problématique à laquelle fait face ce genre d'approches est la transformation de voix, puisque le domaine de variation de la voix est immense et que la base de données ne peut pas capturer toutes les combinaisons de hauteur, d'intensité et de timbre.

1.1.2 Transformation du signal de chant

La méthode de transformation du signal qui est utilisée dans le système de chant - et par conséquent aussi lors des travaux ménés dans ce stage - est basé sur des approches récentes d'analyse/synthèse de la parole [3, 6, 7]. Ces méthodes représentent le signal de parole avec une approche de type source-filtre avec comme modèle de la source le celui de l'impulsion glottique proposé par Liljencrants et Fant [1], qu'il convient d'introduire.

Modèle de source glottique de Liljencrants-Fant (LF) Dans le modèle LF, le signal provenant de la source glottique peut être caractérisé par un unique paramètre de forme, R_d. Ce modèle de source est facilement manipulable. La figure 1.1 présente la dérivée du débit d'air glottique et son spectre : c'est le signal source de la voix. Deux paramètres entrent également en jeu lors de la synthèse de l'impulsion : sa fréquence fondamentale f_0 et le minimum de cette dérivée, E_e, également l'amplitude de l'excitation.



(a) Signal temporel

(b) Spectre

FIGURE 1.1 – Modèle de source pour 5 valeurs de R_d

Ce modèle peut être préféré à un train d'impulsions car il contient des caractéristiques perceptives spécifiques de la voix. En effet, il regroupe les trois paramètres de forme (R_g, R_k, R_a) proposés par Fant [8] et qui avaient pu être rattachés à des paramètres acoustiques et physiologiques [9]. Cependant, ce modèle souffre de la nécessité d'estimer R_d et de gérer convenablement son intervalle de variation. Pour cela, la référence [3] couvre différentes méthodes tirant parti des instants de fermetures glottiques. Généralement, R_d est compris entre 0.3 et 2.7 à l'intérieur de voyelles et peut prendre des valeurs hors de cet intervalle lors de transitions, de fins de phrase [9]. Il est naturellement plus élevé pour les femmes (0.8-2.5) que pour les hommes (0.5-1.5) dans le contexte de la parole.

Plusieurs méthodes d'analyse/synthèse du signal de parole basées sur une representation de la source glottique avec le modèle ont été développées dans l'équipe Analsyse-Synthèse. Il seront presentés par la suite.

Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise (SVLN)

Le système SVLN est détaillé en [3] [6] et il utilise le modèle de source de Liljencrants-Fant [1] pour représenter la partie déterministe de la source, les impulsions glottiques. Il est complété par l'utilisation d'un bruit gaussien pour représenter le bruit de la voix. Il est en effet nécessaire de synthétiser le bruit de souffle lors de la prononciation des voyelles ou encore tous les bruits au sein des consonnes. Il existe une fréquence f_{VU} à partir de laquelle le spectre de la source se fond dans celui du bruit (VU pour *Voiced/Unvoiced*) et qui permet d'estimer le niveau du bruit σ_g. La connaissance de f_{VU}, σ_g et des paramètres de la source f_0, R_d, E_e permet d'estimer le filtre du conduit vocal. En-dessous de f_{VU}, la division du spectre du signal par le spectre de la source LF permet de l'estimer. Au-delà de cette fréquence, il est donné par l'enveloppe du signal.

Les paramètres précédemment décrits sont utilisés pour synthétiser des segments qui sont ensuite joints. Chaque segment possède une durée égale à la période fondamentale pour les parties voisées et de durée fixe pour les parties non voisées, où la source se réduit au bruit gaussien. Le bruit subit un filtrage passe-haut et une modulation synchrone aux impulsions glottiques pour des questions de réalisme. Il est superposé aux impulsions glottiques par une méthode de recouvrement-addition (OLA). La source glottique et le bruit sont ensuite filtrés par le conduit vocal et la fonction de transfert du rayonnement.

Deterministic plus Stochastic Model-based Parametric Re-Synthesis (PaReSy)

La méthode PaReSy [7] se prête à la conversion et à la transformation de parole. Elle étend la méthode précédente en modifiant notamment la représentation de la partie bruitée de la voix. Il ne s'agit plus d'un bruit gaussien dont on aurait estimé l'énergie mais d'un résiduel d'un modèle sinusoïdal. Ce résiduel est ensuite ajusté pour correspondre au niveau de bruit au-delà de la fréquence f_{VU}, ce qui prévient les effets d'erreurs d'estimation du modèle sinusoïdal lors de transitoires rapides. Cette partie non voisée n'est également plus sujette à la méthode de recouvrement-addition. L'estimation du filtre du conduit vocal diffère de la méthode précédente car elle se base sur la division de l'enveloppe du signal par l'enveloppe du spectre de la source glottique LF. Cela permet de limiter les irrégularités induites sur le spectre par des valeurs de Rd trop grandes (voir figure 1.1). Aussi, la fonction de transfert du rayonnement est implicitement incluse dans le spectre de la source.

Ces améliorations ont permis à cette méthode de recevoir une meilleure évaluation face à la méthode SVLN, comme cela est montré dans [7].

Système actuel et base de données

Le système utilisé dans ce travail se base sur les deux modèles précédents avec des modifications qui n'ont pas encore été publiées et qui ne peuvent donc pas être décrites plus en détail ici. Pour les travaux menés dans ce stage, les différences par rapport au modèle PaReSy n'ont que très peu d'incidence et des résultats similaires auraient pu être produits avec le système PaReSy, qui n'est cependant pas intégré dans le système de chant et ne pouvait donc pas être utilisé dans le contexte de ce stage.

Pour la synthèse de chant, tous les systèmes de synthèse décrits précédemment nécessitent une base de données. Cette base contient environ un millier d'enregistrements. Chaque enregistrement est un mot chanté par une chanteuse professionnelle ou un chanteur professionnel, avec le moins de nuances possibles. Les analyses suivantes sont alors réalisées : estimation de la fréquence fondamentale f_0, de la fréquence f_{VU}, de l'enveloppe spectrale et du paramètre R_d. Les parties bruitées sont également extraites.

Les paramètres en entrée de la synthèse sont une courbe de la fréquence fondamentale en fonction du temps ainsi que la liste des phonèmes (les paroles traduites en alphabet SAMPA) et leur durée. Il est possible d'avoir seulement les notes au lieu d'une courbe de fréquence fondamentale et le programme génère alors une telle courbe en fonction du temps. Cette dernière possède des attaques, du vibrato et des transitions qui n'ont donc pas besoin d'être fournis par l'utilisateur. Il est également possible de donner à l'entrée du programme une courbe de l'intensité sonore souhaitée.

1.2 Expressivité de la voix chantée

Pour améliorer le réalisme des synthèses, il est possible d'agir sur plusieurs aspects de la voix : le contour de la fréquence fondamentale, le contour de l'intensité, l'alignement temporel entre la partition et l'interprétation, ou encore le timbre. La référence [5] présente ces différents éléments. Ici, c'est le timbre qui est concerné, notamment pour rendre compte du relâchement ou de l'effort vocal. Dans cette optique, la source glottique et les formants sont au coeur des transformations.

1.2.1 Variations de la source

Un grand nombre de recherches ont pu montrer ou s'appuyer sur le lien entre l'intensité de la voix et sa pente spectrale. Plus il y a de l'effort dans la production de la voix, plus cette pente spectrale diminue, donnant une plus grande amplitude aux fréquences plus hautes, ceci dans la parole comme dans le chant [10] [11]. Le spectre de la source de la voix est généralement approximée par une pente spectrale d'une valeur de -12dB/octave. Comme la figure 1.1 l'illustre, la valeur de R_d influe sur cette grandeur. Elle augmente avec R_d [9]. Il peut ainsi convenir d'associer à des valeurs élevées (supérieures à 2) de R_d une voix relachée et à l'inverse à des valeurs faibles (inférieures à 1) une voix plus tendue comme cela a été modélisé pour le système PaReSy. D'autres systèmes n'utilisant pas le modèle LF modifient tout de même la pente spectrale de la source. C'est le cas du modèle Excitation plus Resonance (EpR) [12], ou de la modélisation proposée dans [13].

Fant [9] donne des indications quant à la modification de son modèle pour différentes qualités de la voix. Il est possible de faire varier l'effort perçu dans la voix en faisant augmenter simultanément $1/R_d$ et E_e, le premier par pas de 1dB et le second par pas de 2dB. Cette variation conserve une cohérence dans l'évolution de la forme de l'impulsion et son énergie. Les relations entre forme et amplitude sont les mêmes pour les voyelles que pour les consonnes voisées. Il ajoute que dans la partie basse et moyenne des fréquences d'un registre d'une personne, il est possible de considérer que E_e est proportionnel à f_0^p où p est compris entre 1.5 et 2. Cela peut s'observer en-dessous de la fréquence critique de l'interlocuteur (110-160Hz pour les hommes et 200-300Hz pour les femmes). Au-dessus de cette fréquence critique, soit E_e chute, soit sa valeur reste constante, selon l'individu. Il souligne aussi que R_d n'est pas affecté par la fréquence fondamentale seule. Le travail présenté ici s'inspire de la règle donnée par Fant concernant l'effort.

Les effets de voix rauque et d'intensité plus extrêmes (cri, ou voix de rock extrême) font partie des possibilités qu'offrent les variations de la source glottique. En prenant en compte les rapports entre les macro-impulsions et les micro-impulsions glottiques, il est possible de générer des sous-harmoniques particulières, à la source de ces styles. Ces cas particuliers de voix intense aux timbres plus complexes sont décrits dans [14] ainsi que des façons de les mettre en place.

1.2.2 Variations des formants

Les formants représentent les résonances du conduit vocal. Leurs caractéristiques sont leur fréquence, leur bande passante et leur amplitude. Les deux premiers formants (dont les fréquences sont les plus basses) permettent essentiellement de former les voyelles. Le premier est lié à l'ouverture de la mâchoire, alors que le second dépend plutôt de la position de la langue. Le troisième formant est lié à la région sous la langue et le quatrième à la longueur du conduit vocal et aux dimensions du larynx [11].

Des corrélations ont pu être observées entre l'effort dans la parole et les positions des formants [15], notamment du premier, dont la fréquence augmente avec l'effort de la voix. Ce résultat a été observé en rapportant la définition de l'effort à la distance de communication estimée par une personne entendant la voix. Différents niveaux de paroles ont été captés et mesurés de cette manière. D'autres études sont parvenues à des résultats similaires [16], mettant aussi en valeur les variations d'amplitudes des formants, qui reflètent la diminution de la pente spectrale avec l'augmentation de l'effort. Les deuxième et troisième formants n'ont pas eu de variation significative observée lors de ces travaux.

Le filtre du conduit vocal est sujet à des variations lors du chant. Un comportement redondant a été mis en valeur par Sundberg [11] d'après l'étude des comportements particuliers des

chanteurs et chanteuses d'opéra : le "formant du chanteur". Dans le but d'être mieux entendu par le public malgré un grand orchestre jouant simultanément, les chanteurs positionnent leur conduit vocal de manière à renforcer les fréquences au niveau de leur troisième, quatrième et cinquième formant. Ce groupe rapproché de formants résulte en une forte amplitude spectrale autour des 2kHz et 3kHz, permettant l'amplification. Certains modèles utilisent un filtre passe-bande pour mettre en place ces variations dans le spectre [17].

D'autres travaux [13] se basent directement sur l'extraction des formants et leur manipulation selon un modèle dérivé du synthétiseur de Klatt [18]. Le programme de traitement de parole Praat est utilisé pour détecter les formants et leur étude se base sur l'analyse de voyelles chantées à différentes intensités pour capturer les variations des formants en fonction de l'intensité. Les résultats de leurs travaux peuvent être consultés sur leur page internet.

Un certain nombre d'études ont mis en relief des stratégies mises en place par des professionnels pour accorder leurs formants avec les harmoniques de la note chantée. Cela dépasse le cadre de cette étude car, pour prendre en compte ce phénomène, il faut donc faire varier les formants. Les travaux suivants [19] [20] [21] confirment l'existence de ces comportements pour les voix de soprano à des notes élevées et listent des stratégies différentes pour les voix plus graves. Les possibilités qu'ouvrent ce travail dans ce domaine sont énoncées à la fin de ce document.

1.3 Estimation des formants

L'estimation des fréquences de résonance du conduit vocal est un procédé difficile à évaluer sur un signal de parole. En effet, la source glottique est périodique et le signal capté est du à l'échantillonnage du filtre du conduit vocal au niveau de ces harmoniques de la note. Cela signifie que l'information contenue dans le signal peut être pauvre et insuffisante. Par exemple, dans le cas du chant, il est courant que les notes soient suffisamment hautes pour dépasser le premier formant. Cela explique en partie la difficulté à apprécier les voyelles lors de chants très aigus.

1.3.1 Méthodes matérielles durant la production de la voix

Une méthode très efficace pour estimer les formants utilise ainsi plus que le son capté. Elle consiste à exciter, par un bruit ou un signal connu, le conduit vocal d'un chanteur lorsqu'il chante, en mesurant la réponse en fréquence à ce signal d'excitation et la voix simultanément. Elle nécessite toutefois la présence du chanteur et le matériel adéquat. Plusieurs études sur les stratégies des chanteurs tirent profit de ce type de procédé [20] [22].

1.3.2 Méthodes de codage prédictif linéaire (LPC)

Le codage prédictif linéaire est une des méthodes les plus répandues pour l'extraction des formants. Elle utilise la prédiction linéaire d'un modèle tous pôles du filtre de la voix et repose sur l'hypothèse que chaque échantillon du signal peut être approché par une combinaison linéaire des échantillons précédents. L'enveloppe spectrale estimée contient le filtre de la source glottique, le filtre du conduit vocal et le filtre du rayonnement au niveau des lèvres. Cette estimation est permise par l'application de l'algorithme de Levinson à la fonction d'autocorrélation ou par une décomposition de la matrice de covariance[23]. Un algorithme de résolution de racines peut être appliquée pour déterminer des fréquences et les bandes passantes des formants du modèle à partir de ces coefficients. Cependant, cette approche comporte des limitations : la distincton des formants lorsqu'ils sont très proches est difficile et l'ordre du filtre tous-pôles affecte directement le nombre de formants estimés [24]. Plusieurs méthodes, comme ARX [25], dérivent de la méthode LPC avec une source plus complexe voire un modèle du filtre du conduit vocal incluant des pôles.

1.3.3 Méthodes basées sur le cepstre

En supposant que la source de la voix est à phase maximale et que le filtre du conduit vocal est à phase minimal, le calcul du cepstre complexe permet de différencier ces éléments dans des bandes de quéfrences différentes. Les premiers coefficients du cepstre peuvent être associés au filtre du conduit vocal en les transformant vers le domaine spectral et les coefficients supérieurs fournit des informations sur la source en les transformant également vers le domaine spectral. Ce type d'approche peut être utilisé pour estimer les formants [26] et sert de base à une estimation par itérations de l'enveloppe spectrale *true enveloppe* [27] qui passe par les pics des harmoniques.

Chapitre 2

Modification de la source de la voix

La première étape dans notre élaboration de règles liées à l'intensité concerne la modification de la source utilisée pour la synthèse, comme elle a été présentée au chapitre précédent. La règle sélectionnée pour cette étude est basée sur des indications de Fant [9]. En effet, Fant propose une règle permettant de faire varier l'effort de la voix. L'ajout d'un incrément d'intensité d'unité suffisamment petit mais significatif selon lui peut être défini par une rédution d'un décibel du paramètre de forme R_d et d'une augmentation de deux décibels du paramètre E_e. Après avoir présenté des modifications du système utiles à cette étude, nous verrons quels comportements peuvent être observés dans les enregistrements à notre disposition et nous détaillerons la méthode de paramétrisation de la règle en fonction de la base de données des segments chantés.

2.1 Modification des données de l'algorithme de synthèse

Afin de pouvoir modifier la source après avoir estimé le filtre du conduit vocal, la synthèse a été modifiée pour utiliser ce filtre au lieu de l'enveloppe. La synthèse ôtait alors la source dans l'enveloppe originale avant de modifier la source.

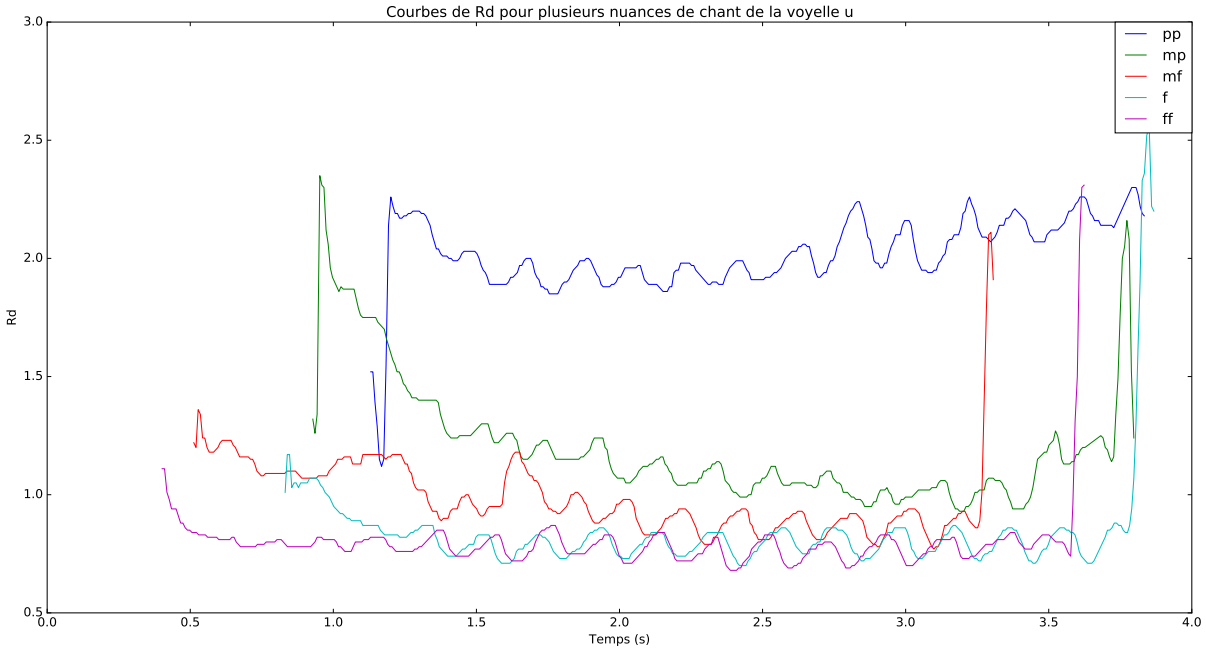
En effet, l'étude a commencé par l'écriture de fonctions permettant de déterminer le filtre du conduit vocal en fonction du temps pour chaque enregistrement de la base de données. Pour cela, R_d et f_0 étaient calculés, ainsi que la fréquence f_{VU} et l'enveloppe spectrale en fonction du temps. Pour un enregistrement, en ôtant la source glottique LF à l'enveloppe spectrale à chaque instant, nous obtenons le filtre du conduit vocal. Cette analyse est utile puisque ces filtres peuvent être modifiées lors de la mise en place des règles portant sur les formants. Ces données sont stockées dans des fichiers au format SDIF, consacré à plusieurs types de descripteurs à l'IRCAM.

Pour tester les différentes règles concernant l'intensité et le paramètre de forme R_d, une fonction de synthèse a été mise en place à partir de ce paramètre et du filtre du conduit vocal. Elle est aussi utilisée lors des tests d'estimation de formants. Avec cette formulation, la synthèse sans modification ni de R_d ni du filtre du conduit vocal est identique à celle utilisée précédemment dans l'algorithme, à partir de l'enveloppe.

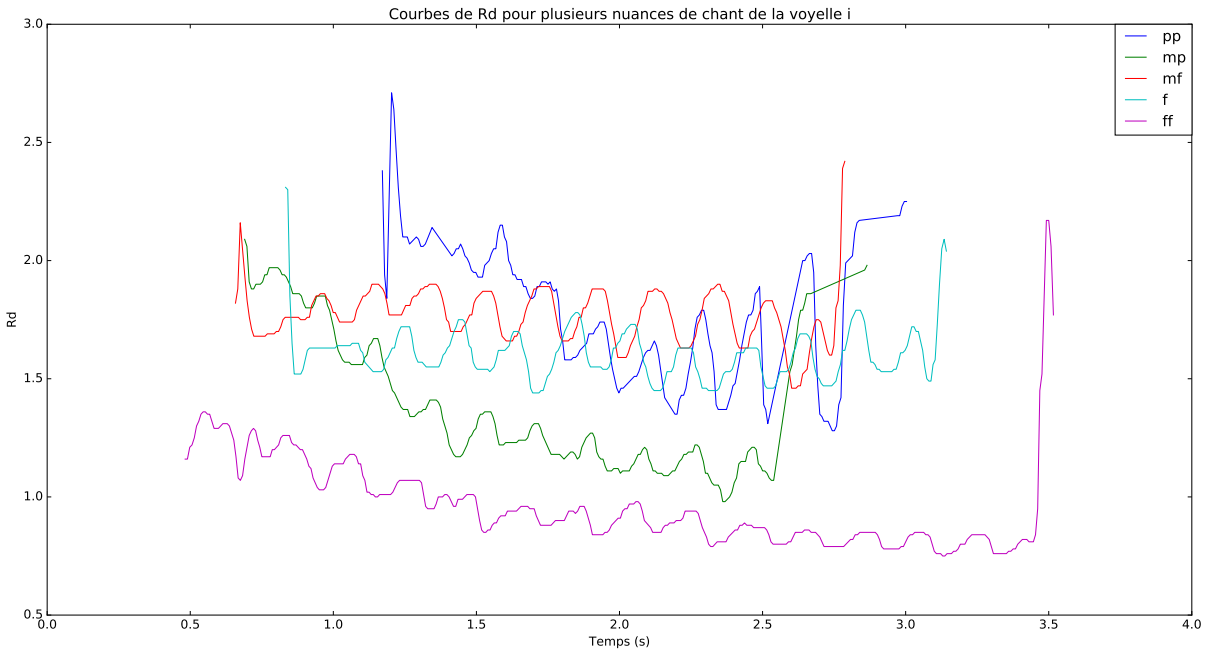
2.2 Analyse de voyelles chantées

La base de données du chanteur Raphaël contient des enregistrements spécifiques concernant l'intensité. Il s'agit de 15 voyelles de la langue française chantées chacune à 5 nuances différentes : *pianissimo*, *mezzopiano*, *mezzoforte*, *forte*, *fortissimo*. Le paramètre R_d a été estimé sur chacun de ces fichiers dans le but d'obtenir une loi et d'observer la pertinence de celle de Fant. Néanmoins, il est très courant que les valeurs de ce paramètre ne respectent pas la logique décrite précédemment le liant à l'intensité de la voix. La figure ci-dessous présente deux voyelles : pour /u/, plus la nuance est intense plus R_d est petit, mais pour /i/, on observe moins bien ce comportement. Une

raison pour cela peut être que l'estimation du paramètre R_d reste une tâche difficile [3] dont la fiabilité peut varier suivant les enregistrements et notamment les différentes voyelles.



(a) Voyelle u



(b) Voyelle i

FIGURE 2.1 – Courbe de Rd pour cinq nuances de chant des voyelles i et u

De même, en observant les valeurs de R_d en fonction de l'intensité sonore des enregistrements, aucun comportement générique n'a pu être observé pour toutes les voyelles.

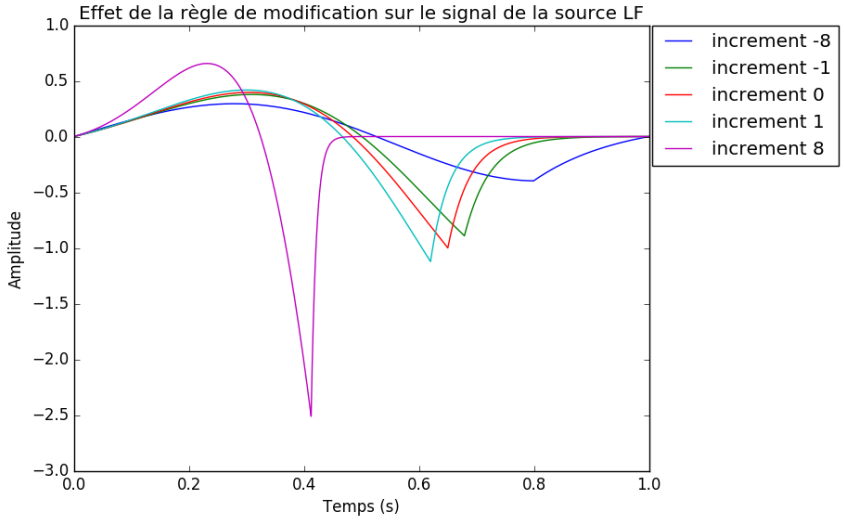
Cette analyse a permis de tirer une première conclusion sur la règle que nous souhaitions apporter : il est difficile de se baser sur des enregistrements pour tenter de mettre au point une règle de modification de la source. La règle de Fant semble réaliste dans plusieurs cas mais les courbes de R_d observés ne nous permettent pas d'en savoir plus.

Cette règle permet une grande variété de sons et les chanteurs peuvent varier R_d et E_e

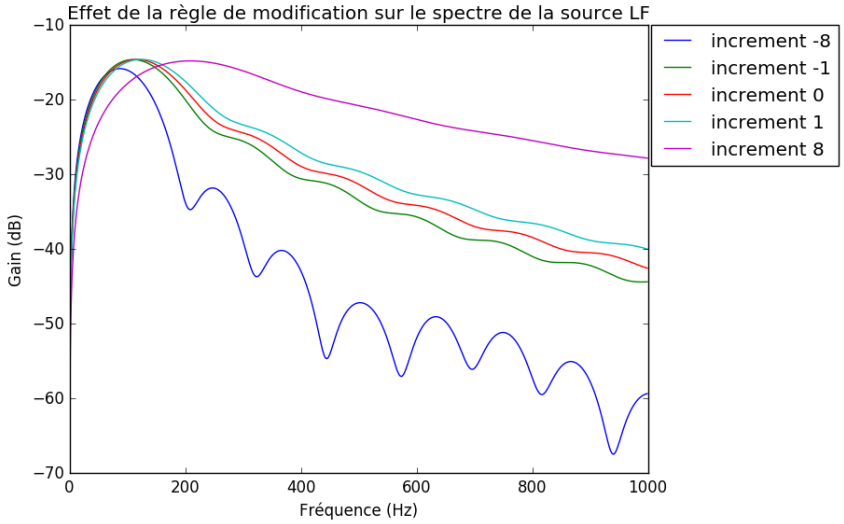
pour adapter le timbre en fonction du contexte. Nous allons utiliser une règle modifiant ces paramètres et développer un réglage permettant à l'utilisateur de faire varier les relations entre ces paramètres selon ses préférences.

2.3 Test informel de la règle de Fant

La figure suivante montre les formes des impulsions glottiques et du spectre de la source pour plusieurs incréments de la règle de modification. La diminution de R_d permet d'augmenter la pente spectrale et l'augmentation de E_e augmente le niveau de la source.



(a) Signal temporel



(b) Spectre

FIGURE 2.2 – Effet de la règle de Fant pour 4 modifications : -8, -1, +1 et +8 incréments d'intensité

À l'aide de la nouvelle organisation de l'algorithme basée sur les analyses du filtre du conduit vocal et des valeurs de R_d, la règle a pu être testée de manière informelle sur des synthèses de chant longues (environ 30 secondes).

Les tests ont été effectués en respectant les limites de R_d : $[0.3; 2.7]$, intervalle à l'extérieur duquel le signal source n'est plus représentatif de la voix humaine. Pour cela, une courbe d'intensité en fonction du temps a été ajoutée comme paramètre de la synthèse de chant. Cette courbe,

dont les valeurs peuvent être négatives, fournit directement le nombre d'incréments d'intensité à apporter au chant. Pour chaque valeur de R_d et de E_e à un instant donné, la courbe d'intensité fournit un nombre d'incréments (qui peut ne pas être entier). La règle a été testée de différentes manières :

- Courbe d'intensité constante : le même incrément est appliqué à tous les instants. Les valeurs utilisés qui ont été jugées significatives sont -4, -2, +2, +4 ;
- Courbe d'intensité croissante : le nombre d'incréments d'intensité en fonction du temps est une droite croissante. Les valeurs initiale et finale étant de -4 et +4 ;
- Courbe d'intensité décroissante : les valeurs initiale et finale étant de +4 et -4 ;

Plusieurs questions ont du être traitées pour juger de l'efficacité de cette paramétrisation. Il a fallu être certain que ce n'est pas la différence entre niveau de la partie harmonique de la source et le bruit du chant qui jouait sur l'impression d'effort, sur l'intensité. Cela a pu être vérifié en ne modifiant que la forme des impulsions et pas leur paramètre E_e. Une écoute informelle a permis de distinguer un changement de timbre résultant de la variation de R_d

Ensuite, les parties bruitées ont du subir une modification pour conserver une cohérence de niveau sonore. En effet, l'analyse d'enregistrements de voyelles chantées avec différentes nuances a montré que le niveau du bruit fluctue avec l'intensité, ce qui est d'ailleurs soutenu dans [13]. Ce sont les résiduels non voisés des enregistrements qui ont été analysés et qui ont révélé qu'une variation du niveau de bruit était nécessaire. La variation de l'énergie E_e de l'impulsion glottique a donc aussi été appliquée comme un gain aux parties bruitées, les consonnes et le bruit de souffle présents dans la prononciation des voyelles.

Cette règle a permis d'ajouter du contrôle dans le timbre de la voix en permettant du relâchement et de la tension, comme la littérature sur les effets de modification de pente spectrale le laissaient attendre. La problématique à soulever est ensuite de lier cette règle aux variations d'intensité, qui constitue une entrée de l'algorithme de synthèse, et pas seulement à un nombre d'incréments abstraits. Cela permettra notamment de copier un signal de chant en fournissant les descripteurs de fréquence fondamentale, d'intensité, les informations sur les paroles, en modifiant les valeurs de R_d associées aux unités de la base de données. Cela pourra modifier la sensation d'effort selon les fluctuations de la courbe d'intensité et appuyer la cohérence du timbre.

2.4 Méthode proposée de détermination d'une règle de modification

L'utilisateur du système souhaite synthétiser un chant respectant une certaine courbe d'intensité sonore. Cette mesure perceptive doit être liée à l'effort de la voix de sorte à ce que les segments de forte intensité donnent l'impression d'un effort vocal plus important, et inversement. Pour cela, compte tenu des éléments discutés précédemment, la méthode proposée pour paramétrer la règle est la suivante.

1. Définir une valeur de R_{d0} et de E_{e0} initiales ;
2. Générer une grille de valeurs de R_{di} et de E_{ei} en appliquant le nombre d'incréments i (qui peut être négatif) selon la règle de Fant aux valeurs initiales précédentes ;
3. Définir un ensemble de valeurs de la fréquence fondamentale f_0 associé à l'ambitus de la voix ;
4. Pour chaque voyelle, sélectionner un filtre du conduit vocal VTF disponible dans la base de données ;
5. Pour chaque voyelle, synthétiser un segment pour chaque f_0 et chaque couple (R_{di}, E_{ei}) en filtrant la source par le VTF associé ;

6. Mesurer l'intensité sonore de chacun de ces signaux ;
7. Utiliser le lien entre i et les variations d'intensité pour préciser la règle de modification de la source ;

Ces données d'intensité fournissent une information importante : le comportement de la règle de Fant par rapport à l'intensité sonore, selon les mesures du conduit vocal du chanteur et la voyelle chantée. Cette méthode permet de tester en quelque sorte la synthèse sur des éléments minimaux qui sont semblables à ceux qu'elle utilise, puisqu'elle resynthétise la source par le modèle LF et la filtre via des informations extraites sur les enregistrements de la base de données. Elle conjecture le comportement de la loi de Fant dans le but de mieux la maîtriser, mais comme elle ne peut pas tenir compte de toutes les possibilités du conduit vocal du chanteur, elle est approximative. Elle doit donc être complétée par une étape d'ajustement où un gain est appliqué au signal synthétisé pour le rapprocher de l'intensité attendue.

2.4.1 Méthode de mesure de l'intensité

La mesure d'intensité utilisée est celle de l'intensité perçue mise au point par l'équipe et qui se base sur l'analyse harmonique de la parole. C'est une mesure de l'intensité sonore simplifiée au sens où elle utilise seulement les harmoniques des parties voisées, et somme leurs valeurs de sonie pour obtenir la sonie globale. Elle ne prend pas en compte les effets de masquages fréquentiels et temporels qui font partie des modèles de l'état de l'art. Cette mesure est celle utilisée pour générer des courbes d'intensité en entrée de la synthèse et à l'intérieur de l'algorithme pour évaluer l'ajustement à appliquer.

2.4.2 Intervalles de valeurs des paramètres et leur influence

Cette étude a été réalisée à l'aide de filtres du conduit vocal analysés sur la base de données de Raphaël. Il y a 15 voyelles françaises qui ont été étudiées. Les notes pouvant être chantées par les hommes se trouvent entre 80 et 590 Hz, ce qui a été l'intervalle de variation de la fréquence fondamentale. Les valeurs choisies correspondent aux notes de musique présentes dans cet intervalle.

R_{d0} et E_{e0} ont été fixés à 1. Pour enrichir les informations, l'ensemble des incréments appliqués I est constitué de multiples de 0.5 et de -0.5. Avec ces données, 38 modifications ont été appliquées pour générer des couples de valeurs (R_{di}, E_{ei}) où R_{di} est toujours compris dans l'intervalle $[0.3; 2.7]$.

2.4.3 Surfaces liant l'intensité et les incréments de la règle de modification

La figure 2.3 présente le type de surface que fournit cette étude.

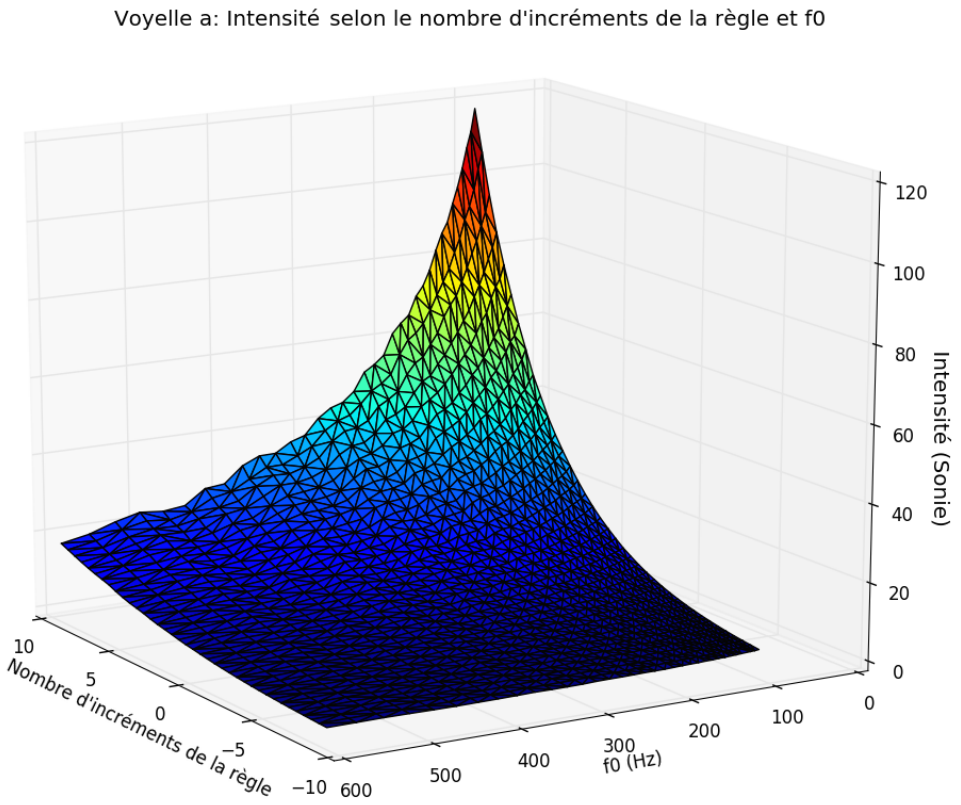


FIGURE 2.3 – Surface obtenue pour la voyelle a

Il est possible d'observer que l'intensité augmente bien avec le nombre d'incréments de la règle de modification de la source. Néanmoins cette surface est difficile à intégrer dans la règle pour préciser le contrôle.

Pour obtenir l'information qui nous intéresse en particulier, il faut considérer le logarithme de l'intensité sonore divisée par l'intensité sonore de référence, c'est-à-dire calculée aux points initiaux, pour la source (R_{d0}, E_{e0}, f_0). Il s'agit du gain d'intensité en dB donné par l'application des incréments. Les surfaces représentant le nombre d'incréments en fonction de ce gain et la fréquence fondamentale se présentent comme à la figure 2.4.

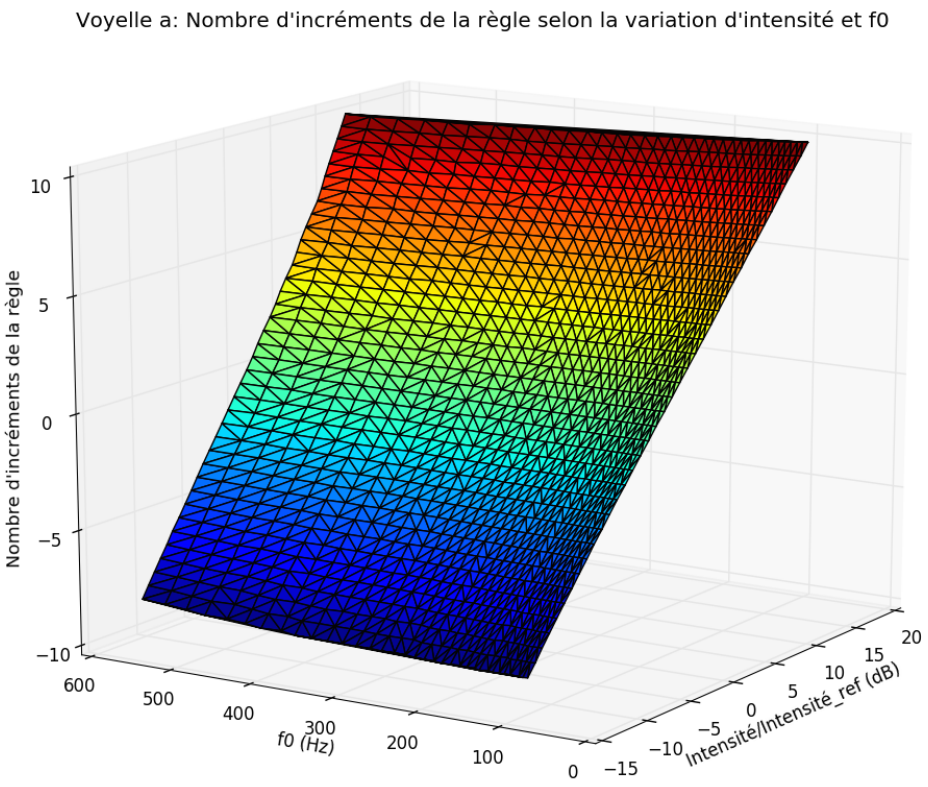


FIGURE 2.4 – Surface avec le logarithme de la variation d'intensité

Un aspect remarquable de ces surfaces est la possibilité d'utiliser des régressions linéaires pour différentes valeurs de f_0 liant le nombre d'incréments de la règle appliqués et l'évolution de l'intensité sonore.

Au sein d'une même voyelle, les variations entre les pentes de ces régressions sont dues à R_d qui déforme le spectre de la source et à la fréquence fondamentale qui échantillonne le filtre du conduit vocal, rendant les variations d'intensité plus complexes.

2.4.4 Détermination de coefficients pour la règle par régressions linéaires

Pour chaque f_0 et chaque voyelle un coefficient peut être déterminé via une ligne des surfaces précédentes : la ligne comportant les variations du nombre d'incréments. Cette ligne est approchée par régression linéaire et fournit un coefficient multiplicatif qui, associé à la variation d'intensité souhaitée, donne le nombre d'incréments de la règle de modification à appliquer. Les coefficients de corrélation de ces régressions linéaires se sont avérés très proche de 1, autour de 0.999.

La figure 2.5 présente les valeurs de ce coefficient pour quelques voyelles en fonction de la fréquence fondamentale de la source.

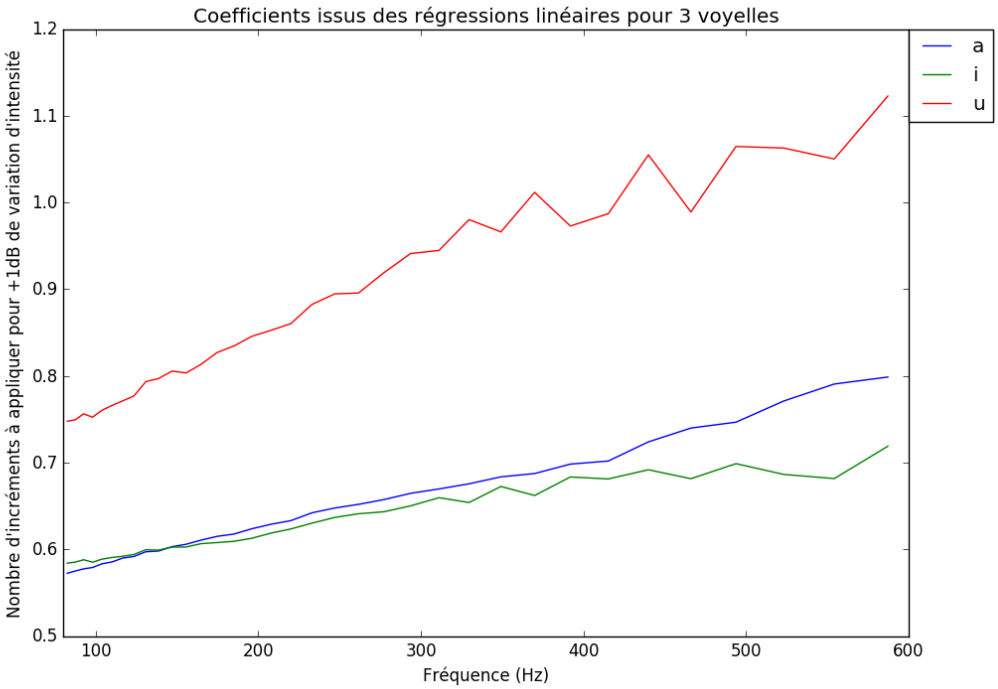


FIGURE 2.5 – Résultat de la méthode pour 3 voyelles

En choisissant une courbe associée à une voyelle et une valeur de la fréquence fondamentale, il suffit de lire l'ordonnée du point associé sur la courbe pour savoir combien d'incréments d'intensité doivent être appliqués pour augmenter l'intensité de 1 dB. Les données ont été établies pour les 15 voyelles de notre étude mais ne figurent pas toutes simultanément pour des questions de lisibilité. Elles suivent la même tendance : plus la fréquence augmente plus le nombre d'incréments doit être grand pour faire varier l'intensité. Ce résultat peut sembler logique car avec R_d petit, l'amplitude du signal glottique diminue, ce qui doit être compensé par une variation plus importante de l'énergie E_e.

Les irrégularités des courbes au-delà de 400Hz proviennent du fait que l'intensité n'est pas augmentée ou diminuée de la même façon dans les hautes notes que dans les basses notes. En effet, le filtre du conduit vocal est mieux échantillonné entre 100 et 200Hz, alors que vers 400-600Hz, une variation de f_0 engendre un changement radical des amplitudes du filtre qui sont échantillonnées, donc un comportement différent par rapport à la règle de Fant. La solution a donc été de lisser ces variations car elles sont l'effet d'un échantillonnage et ne sont pas systématiques.

2.4.5 Influence des paramètres

Cette méthode a été testée avec différents filtres du conduit vocal pour chaque voyelle et s'est avéré ne pas subir de variations importantes en fonction de ce paramètre. Un point important rendant cette étude utilisable dans l'algorithme est que les filtres du conduit vocal estimés proviennent d'enregistrements où le chanteur avait une intensité constante (*mezzoforte*) et un ton neutre. Toute la base de données ayant été enregistrée dans les mêmes conditions, la variabilité du conduit vocal est suffisamment petite. Par exemple, il n'y a pas de changement de registre.

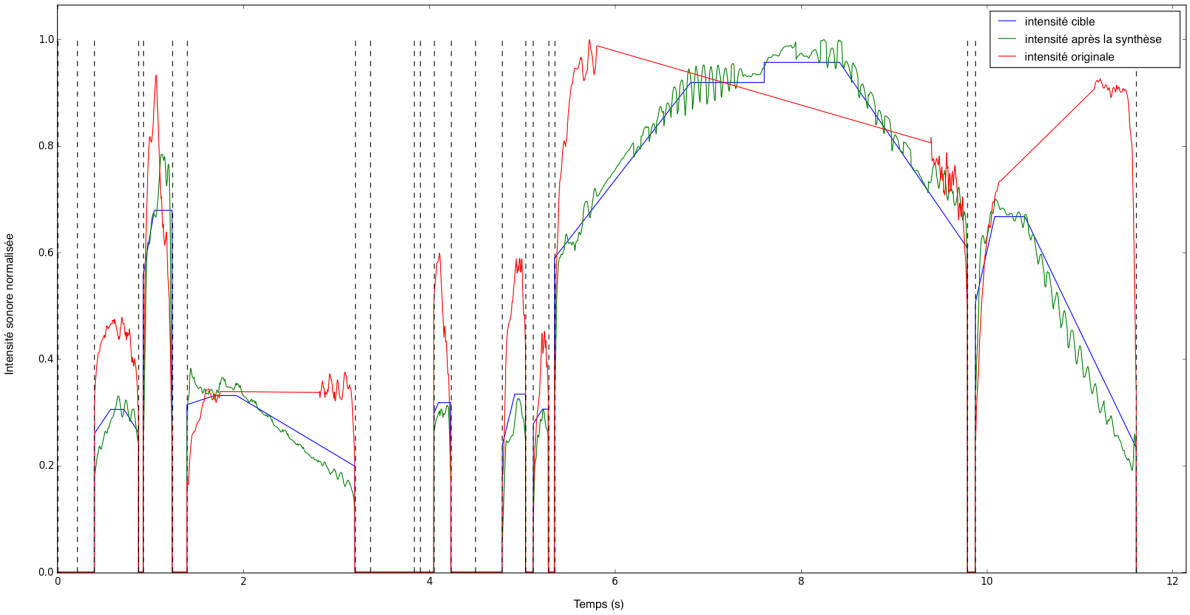


FIGURE 2.6 – Intensité cible de la synthèse et intensité permise par la règle de modification

2.5 Intégration à l'algorithme de synthèse

La règle de modification a été ajoutée à l'algorithme de synthèse en utilisant les coefficients des régressions présentées à la figure précédente. La figure 2.6 illustre le type de résultat que cette méthode permet d'obtenir. La courbe d'intensité cible est obtenue en sortie d'un module de contrôle qui peut utiliser des paramètres spécifiés directement par l'utilisateur ou bien appris sur un corpus. La courbe d'intensité originale provient de l'analyse de la base de chants enregistrés. Ces deux courbes sont utilisées avec la fréquence fondamentale, la voyelle et la règle de modification pour déterminer à chaque instant le nombre d'incréments d'intensité à appliquer. Sur cette figure, nous pouvons observer la courbe du signal synthétisé à l'aide de la méthode.

Une correction est ensuite appliquée pour que la courbe se rapproche encore plus de l'intensité cible. Mais la règle de modification fournit déjà des incréments de modification qui sont cohérents avec les mesures d'intensité.

2.6 Évaluation par test d'écoute

La règle issue de Fant [9] et adaptée d'après la méthode présentée a été évaluée dans le cadre d'un test d'écoute. En effet, la littérature concernant les modifications de la qualité de la voix lors de synthèses de chant se base très fréquemment sur ce type de tests pour évaluer le réalisme et la pertinence des résultats. Comme le soulignent Umbert et al. [5], il n'existe pas encore de cadre d'évaluation commun dans ce domaine. Devant la difficulté de le mettre en place, l'alternative consiste en la réalisation de tests perceptifs suivant des consignes précises [28] que nous avons utilisées comme notre ligne directrice et qui sont détaillées ci-après. Le test a été publié en ligne sur le site `http://recherche.ircam.fr/anasyn/ardaillon/testIntensityMaxime2016/`. Une version de démonstration avec l'identification des signaux est disponible à l'adresse `http://recherche.ircam.fr/anasyn/ardaillon/testIntensityMaxime2016/demo.php`.

2.6.1 Participants

Les réponses de 5 femmes et de 18 hommes ont pu être recueillies. Parmi eux, il est possible de compter : 14 professionnels de l'audio, 15 personnes familières avec les tests d'écoute, 7 personnes

familières de la synthèse de voix, 21 personnes affirmant jouer d'un instrument de musique ou chanter et 6 personnes qui jugent leur de niveau de chant au-dessus de 3 sur 5. Ces informations faisaient l'objet d'un formulaire en fin de test. Ce sont les réponses des 22 personnes ayant utilisé un casque audio qui ont été conservées, la dernière personne ayant utilisé des haut-parleurs.

2.6.2 Description du test

Après avoir étudié et écouté les possibilités de la règle de modification qui a été ajoutée à l'algorithme de la synthèse, nous avons choisi pour cette évaluation de présenter des sons comportant des variations d'intensité, nous semblant plus à même de rendre compte des modifications du timbre apportées par la règle. À forte intensité, la pente spectrale est faible et le spectre comporte plus de hautes fréquences, les basses fréquences étant atténuées. À faible intensité, cette pente est forte et la source possède de l'énergie essentiellement dans les basses fréquences. Pour évoquer au mieux cette effet, les signaux de tests sélectionnés sont des crescendos chantés avec les voyelles /a/, /i/ et /u/. De plus, des enregistrements expérimentaux correspondant à ces crescendos étaient disponibles au laboratoire depuis la création de la base de données du projet. Ces originaux ont servi de vérité terrain, ils sont dus au même chanteur qui à contribué à la base de données. La possibilité de soumettre aux participants des signaux à intensité constante a été écartée car la stationnarité de certains paramètres trahissent davantage l'aspect artificiel de la synthèse. Lors de variations telles que des crescendos, ce risque peut être évité.

Le test est de type CMOS, ou Comparaison par Note d'Opinion Moyenne (*Comparison Mean Opinion Score*), avec une évaluation par paire de sons inspirée des recommendations de [28]. Ainsi, le participant écoute deux sons et évalue le meilleur des deux en lui attribuant 1, 2, ou 3 points et 0 s'il juge les deux sons équivalents. Des indications ont été données pour orienter l'écoute vers le timbre et son évolution. Pour chaque voyelle, trois signaux sont comparés deux à deux : l'original, la synthèse sans la règle de modification de la source et la synthèse avec la modification de la source. Sans modification de la source, c'est un gain qui est appliqué au signal synthétisé pour qu'il satisfasse une certaine courbe d'intensité en entrée. Les signaux sont ordonnés aléatoirement au sein de chaque voyelle à chaque actualisation de la page internet. Ces consignes suivent les indications données par Umbert et al. [5] : la façon de voter et les aspects sur lesquels le participant doit porter son intérêt sont explicités. Selon cet article, les tests par comparaison sont plus efficaces qu'une Note d'Opinion Moyenne. Ce format implique que neuf classements de préférence devaient être réalisés par chacun des participants.

2.6.3 Analyse des résultats

Les résultats du test sont présentés figure 2.7. En confondant les résultats des trois voyelles, notre système ne semble pas apporter d'amélioration en particulier. Les intervalles de confiance à 95% ne se distinguent pas suffisamment des résultats de la synthèse sans la règle de modification. Ce résultat est surprenant étant donné qu'un pré-test sur quelques personnes semblait confirmer que l'effet était suffisamment remarquable et que l'observation des différences spectrales soutenait la présence d'un changement de timbre. Les résultats sont plus significatifs s'ils sont traités voyelle par voyelle. La synthèse du /i/ est meilleure avec la règle à tel point que son intervalle de confiance se démarque de celui de la synthèse sans la règle pour se rapprocher de manière conséquente de celui de l'enregistrement réel. Pour la voyelle /a/, il ne semble pas y avoir eu de distinction de l'effet de la règle par les participants. Notre écoute portait à croire que l'efficacité des modifications y serait les plus probantes, mais cela doit venir du fait que nous sommes entraînés et préparés lors de l'écoute. En effet, le début du crescendo ressemble bien plus à un piano qu'à une simple diminution du niveau du son, comme c'est le cas dans la synthèse originale. La distinction entre un niveau plus bas dans pour cette dernière et la voix plus relâchée de notre méthode est particulièrement discernable dans ce début de crescendo. Le crescendo de la voyelle /u/ présente d'importantes variations de timbre qui montrent que la

règle peut altérer de manière trop importante la voix jusqu'à la rendre moins naturelle. Cela peut provenir du fait que le paramètre R_d ait été surestimé lors des mesures utilisées pour la synthèse. Pour résoudre ce problème il faudrait diminuer ce paramètre manuellement avant l'application de notre règle de modification ou vérifier l'estimation du R_d sur les unités utilisées et leurs analyses.

Le nombre de cas traités est restreint et bien que la règle sur les trois voyelles étudiées confondues ne semble globalement pas améliorer le naturel du son perçu par les auditeurs, elle a été efficace pour un des trois cas. Cela ne suffit pas à la rendre valide, mais cela montre qu'elle peut représenter un apport pour l'expressivité de la synthèse. À l'inverse, elle ne peut pas être écartée dans l'immédiat. Notre choix de signal a été ambitieux et les participants ont pu prendre en compte un nombre trop important de paramètres : les variations lors du crescendo, les parties finales plutôt qu'initiales, où notre travail semblait le plus prometteur. Cette étude montre que le test perceptif doit être révisé par des sons plus basiques ou moins de paramètres peuvent attirer l'attention du participant. De tels signaux pourraient être synthétisés avec une intensité constante, à différents nombres d'incréments d'intensité, ce qui permettrait d'éviter la considération de la variation du crescendo. Un prochain test couvrira plus de cas, plus de voyelles, de manière plus précise et plus restreinte pour mieux cibler l'attention et mieux permettre la comparaison. En concentrant le participant sur le timbre par rapport à l'impression de distance, nous espérons montrer tout l'intérêt que ces modifications peuvent apporter sur la sensation d'effort.

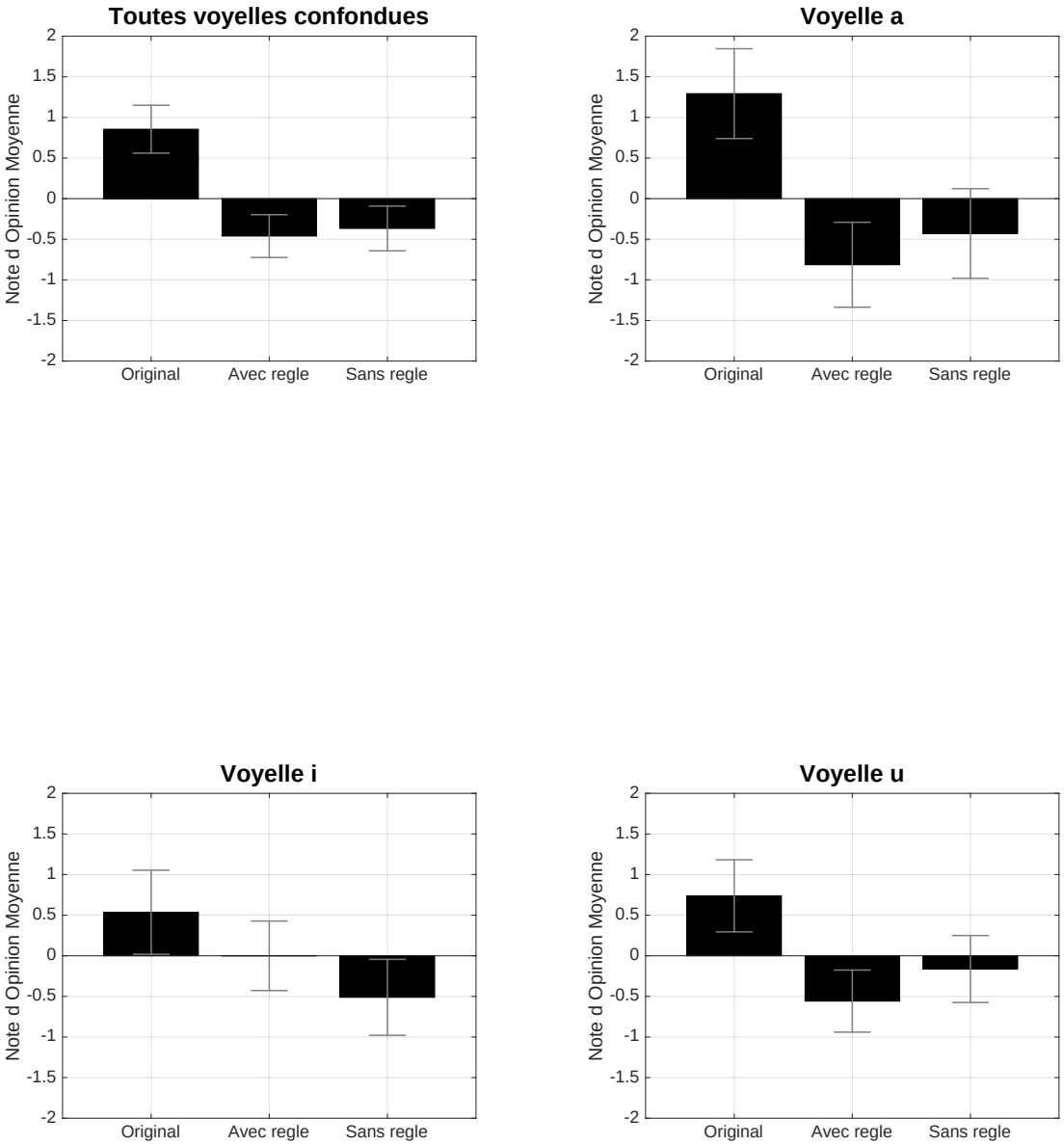


FIGURE 2.7 – Résultats du test CMOS pour les voyelles /a/, /i/, /u/ et ces trois voyelles confondues

Chapitre 3

Estimation des formants

Comme évoqué précédemment, la modification de la source permet d'induire une variation du timbre cohérente avec l'intensité de la voix, mais ne suffit pas à elle seule à reproduire toutes les variations inhérentes à un chant naturel. Il est aussi nécessaire de modifier l'enveloppe spectrale. Cependant, cette modification, via l'utilisation de règles, nécessite une réprésentation paramétrique de l'enveloppe, avec la connaissance des paramètres des formants. Ainsi, ces derniers doivent tout d'abord être estimés.

Dans cette partie, nous abordons la méthode élaborée pour estimer les formants et leurs caractéristiques. Après avoir décrit l'approche choisie, nous détaillerons la descente du gradient qui est effectuée et les améliorations qui ont été ajoutées pour rendre cette opération plus rapide et éviter les minima locaux. Enfin cette méthode est évalué sur des filtres synthétiques et sur des enregistrements en comparaison avec le logiciel Praat.

3.1 Approche choisie

Différemment des méthodes présentées au chapitre 1, l'approche se base sur les paramètres f_0, R_d et VTF qui ont été estimés sur la base de données. En effet, la méthode LPC fonctionne sur le signal temporel pour estimer les pôles du filtre du conduit vocal alors qu'ici, la connaissance de la forme de la source a permis l'estimation de ce filtre et c'est sur ce dernier que notre étude fonctionne. Ainsi, la méthode LPC ne permet pas de différencier filtre de la source et filtre du conduit vocal. Pour permettre un maximum de contrôle sur les formants, un modèle paramétrique doit être établi pour remplacer l'estimation du filtre du conduit vocal. L'objectif final est de pouvoir remplacer dans la synthèse actuelle ce filtre par un modèle de formants. En soustrayant au filtre le modèle des formants nous pouvons obtenir un résiduel de l'enveloppe que l'on peut conserver et ré-ajouter après la modification des formants. La synthèse utiliserait alors les formants modifiés et ce résiduel pour conserver plus d'information sur le filtre du conduit vocal, tout en modifiant celle qui nous intéresse en particulier. Aussi, cette estimation permet d'analyser les enregistrements de voyelles chantées avec des nuances différentes. La mesure des variations qui pourraient exister peut être réutilisée pour former de nouvelles règles de modification des formants, comme proposé dans [13].

L'approche choisie consiste à approcher le module du filtre du conduit vocal par une méthode de descente de gradient, où il sera modélisé par une somme de filtres passe-bas d'ordre 4. Ces filtres sont inspirés de [13], où l'estimation des formants est effectuée par le logiciel de traitement de parole Praat pour servir d'entrée au modèle et analyser et resynthétiser différentes nuances de chant. Praat est basé sur une méthode LPC et suppose que la pente spectrale est constante, égale à -6dB/octave, pour réaliser ses analyses. La synthèse utilisé dans ce travail appelle naturellement à considérer un autre type de méthode. Il s'agit d'insérer de l'information à priori dans l'extraction des formants. En effet, les méthodes d'estimation de l'enveloppe spectrale telles que la méthode

LPC, le cepstre ou la méthode True-Enveloppe, ont des limites lorsque l'enveloppe en sous-échantillonnée, par exemple lorsqu'un formant se situe entre 2 harmoniques et devient difficile à détecter. La méthode par descente de gradient a pour vocation de tenter d'améliorer ces cas en trouvant les meilleurs paramètres qui peuvent expliquer les valeurs des harmoniques observées, même lorsque le formant est très peu échantillonné par ces harmoniques.

Pour un formant i, de fréquence f_i et de facteur de qualité q_i et d'amplitude g_i, le filtre associé est le suivant :

$$|R_i(f)| = \frac{g_i}{(1 - \frac{f^2}{f_i^2})^2 + (\frac{f}{q_i f_i})^2} \tag{3.1}$$

Notre modèle de filtre du conduit vocal comporte 4 formants en parallèle, dans le but de pouvoir comparer nos résultats avec les sons synthétisés par [13] et de tester leurs régressions concernant l'intensité. Un modèle en parallèle a de plus l'avantage de permettre des modifications formant par formant en particulier en matière d'amplitude. L'amplification liée au "formant du chanteur" est plus facilement réalisable si l'on peut manipuler les amplitudes des formants une à une, contrairement au modèle en cascade ou un seul gain régit le filtre. D'après Klatt [18], la structure en parallèle peut atteindre une qualité moindre que celle en cascade pour plusieurs phonèmes. Cependant, l'utilisation du spectre résultant abordé précédemment doit permettre de conserver suffisamment d'information pour préserver la qualité de la synthèse.

Le filtre du conduit vocal est estimé par le modèle suivant :

$$R(f) = \sum_{i=1}^{4} |R_i(f)| \tag{3.2}$$

3.2 Descente de gradient

La fonction effectuant la descente de gradient n'a pas été développée dans le cadre de ce travail. Elle se base sur [29].

3.2.1 Fonction de coût

Étant donné que le filtre du conduit vocal est excité par une source harmonique de fréquence f_0, ce sont les valeurs H_n du module de son spectre aux harmoniques $n.f_0$ qui sont les seules réelles valeurs nous fournissant de l'information sur le filtre. La fonction de coût $C(\theta)$ est ainsi basée sur l'erreur entre les valeurs du filtre du conduit vocal et de notre modèle au niveau des harmoniques :

$$C(\theta) = \sum_{j=1}^{N} (H_j - R(j.f_0))^2 \tag{3.3}$$

Où l'ensemble θ des paramètres contient les quatre amplitudes, fréquences et bandes passantes des filtres des formants. Dans cette formule précédente, le carré est une opération classique permettant de rendre cette fonction positive. Aussi, le nombre d'harmoniques N influe directement sur la complexité du calcul de cette fonction et par extension de son gradient. Dans le but d'estimer quatre formants, les harmoniques qui doivent être prises en compte doivent atteindre les 4000Hz, la fréquence maximum du quatrième formant dans la plupart des cas [30]. Le nombre d'harmoniques à considérer dépend donc de la fréquence de la note chantée : plus celle-ci est élevée, plus le nombre d'harmoniques diminue et le calcul est rapide. Les voix d'hommes sont donc plus longues à analyser avec cette méthode.

3.2.2 Formulation

Le gradient est calculé en utilisant le fait que les filtres des formants à paramétrer sont évalués au niveau des harmoniques de la fréquence fondamentale.

Ainsi, nous adoptons la formulation suivante :
— g_i l'amplitude du i-ième formant
— $p_i = \frac{f_0}{f_i}$ le rapport de la fréquence fondamentale et de la fréquence du formant i
— $q_i = \frac{f_0}{BW_i}$ le facteur de qualité du formant i
— H_j le module du filtre du conduit vocal à la j-ième harmonique de la fréquence fondamentale
— $K_j = j^2$ l'index du partiel de la j-ième harmonique de la fréquence fondamentale au carré
Alors θ est l'ensemble des paramètres g_i, p_i, et q_i.

Nous pouvons réécrire le dénominateur de nos filtres passe-bas d'ordre quatre en fonction du i-ième formant et de la j-ième harmonique de la façon suivante :

$$P_{ij} \;=\; \left(1 \;-\; K_j\, {p_i}^2\right)^2 \;+\; K_j \left(\frac{p_i}{q_i}\right)^2 \tag{3.4}$$

Avec ces formulations, la valeur d'un filtre d'un formant i à une harmonique j est donnée par :

$$R_i(j f_0) \;=\; \frac{g_i}{P_{ij}} \tag{3.5}$$

Considérons également C_j et E_j tel que :

$$C_j = \left(H_j \;-\; \sum_{i=1}^{4} |R_i(j\, f_0)|\right)^2 \;=\; {E_j}^2 \tag{3.6}$$

La fonction de coût devient :

$$C = \sum_{j=1}^{N} C_j \tag{3.7}$$

3.2.3 Gradient

Le gradient est fourni grâce à l'écriture suivante :

$$\frac{\partial C}{\partial g_i} = \sum_{j=1}^{N} \frac{\partial C_j}{\partial g_i}, \qquad \frac{\partial C}{\partial p_i} = \sum_{j=1}^{N} \frac{\partial C_j}{\partial p_i}, \qquad \frac{\partial C}{\partial q_i} = \sum_{j=1}^{N} \frac{\partial C_j}{\partial q_i} \tag{3.8}$$

Les formules des dérivées nécessaires sont les suivantes :

$$\frac{\partial C_j}{\partial g_i} \;=\; -\,2\,\frac{E_j}{P_{ij}} \tag{3.9}$$

$$\frac{\partial C_j}{\partial p_i} \;=\; -\,4\,(2\,{p_i}^2\,{q_i}^2\,K_j \;-\; 2\,{q_i}^2 \;+\; 1)\frac{g_i\,p_i\,K_j\,E_j}{{q_i}^2\,{P_{ij}}^2} \tag{3.10}$$

$$\frac{\partial C_j}{\partial q_i} \;=\; -\,4\,\frac{g_i\,{p_i}^2\,K_j\,E_j}{{q_i}^{1.5}\,{P_{ij}}^2} \tag{3.11}$$

3.2.4 Initialisations

L'initialisation du paramètre $\theta = \left(g_i, p_i = \frac{f_0}{f_i}, q_i = \frac{f_i}{BW_i}\right)_{1 \le i \le 4}$ n'est pas efficace si elle est aléatoire. Pour cause, la fonction de coût comporte un grand nombre de minima locaux qui peuvent mettre fin à la descente de gradient rapidement. Une étape supplémentaire s'est avérée

indispensable pour parer à cela. Il s'agit de mettre en place plusieurs initialisations $\{\theta_n\} = \Theta$ pour mieux parcourir l'espace des combinaisons. Chaque combinaison $\theta \in \Theta$ fait l'objet d'une descente de gradient dont nous stockons le résultat $\theta' \in \Theta'$. La taille de Θ peut être grande donc le nombre d'itérations maximal de ces descentes de gradient est fixé à 50. Le paramètre $\theta'_{opt} \in \Theta'$ minimisant le mieux la fonction de coût est conservé et sert d'initialisation pour une dernière descente de gradient, dont le nombre d'itérations maximal est plus grand, égal à mille. La façon de construire l'ensemble Θ est décrite ci-dessous. Nous décrivons également une façon d'initialiser les amplitudes améliorant le procédé.

Fréquences f_i

La fréquence fondamentale étant connue, l'initialisation des valeurs de p_i peut s'exprimer par les fréquences f_i. Chaque fréquence f_i est initialisée dans un intervalle particulier, lié au formant à estimer. Ces intervalles correspondent à ceux associés aux voyelles françaises [30] :

- f_1 est initialisé entre 250 et 860 Hz pour correspondre au premier formant
- f_2 entre 850 et 2250 Hz pour le deuxième formant
- f_3 entre 2200 et 3000 Hz pour le troisième formant
- f_4 entre 2600 et 3900 Hz pour le quatrième formant

L'ensemble Θ est construit en considérant plusieurs combinaisons possibles pour le vecteur $(p_{1,init},\ p_{2,init},\ p_{3,init},\ p_{4,init})$. Pour quadriller l'ensemble des possibilités, les variables f_i sont initialisées sur une harmonique ou entre deux harmoniques de f_0. Par exemple, si la fréquence fondamentale est de 200Hz, f_1 sera initialisée successivement à 300, 400, 500, 600, 700 et 800Hz, soit toutes les fréquences appartenant à l'intervalle [250 - 860Hz] et situées sur une harmonique ou entre deux harmoniques. De la même manière, f_2, f_3 et f_4 sont initialisées successivement sur plusieurs valeurs. L'ensemble Θ est constitué de toutes les combinaisons possibles.

Facteurs de qualité q_i

Pour chaque combinaison $(p_{1,init}, p_{2,init}, p_{3,init}, p_{4,init})$, des facteurs de qualité initiaux ensuite sont calculés. L'initialisation des valeurs de $q_i = \frac{f_i}{BW_i}$ peut s'exprimer par les bandes passantes BW_i. Les valeurs initiales des bandes passantes ont été fixées : $BW_{1,init} = 60$Hz, $BW_{2,init} = 110$Hz, $BW_{3,init} = 170$Hz, $BW_{4,init} = 250$Hz. Ces valeurs sont dues à [31], elles correspondent aux valeurs maximales fournies par la référence et nous ont fourni de meilleurs résultats que les valeurs minimales (30, 35, 46 et 50Hz) et que les valeurs centrales (60, 72.5, 108 et 150Hz). La qualité de la méthode pourrait être améliorée en approfondissant cette initialisation.

Amplitudes g_i

Pour chaque combinaison $(p_{1,init},\ p_{2,init},\ p_{3,init},\ p_{4,init},\ q_{1,init},\ q_{2,init},\ q_{3,init},\ q_{4,init})$, des gains initiaux sont ensuite déterminés. Il s'agit de la dernière étape de la construction de l'ensemble Θ.

Ce calcul a pour objectif de sélectionner des amplitudes qui rapprochent le filtre issu de notre paramètre initial θ du filtre du conduit vocal à analyser. Pour cela, le module du spectre harmonique $(H_j)_{1 \leq j \leq N}$ à approcher est utilisé. Il y a deux cas possibles pour déterminer $g_{i,init}$:

- Si $f_{i,init}$ est égale à une harmonique j de la fréquence fondamentale.
  Alors $g_{i,init}$ est calculé pour satisfaire l'équation suivante :

$$|R_i(jf_0)| = H_j \tag{3.12}$$

— Si $f_{i,init}$ est située entre deux harmoniques j et $j+1$ de la fréquence fondamentale.
Alors $g_{i,init}$ est calculé pour satisfaire l'équation suivante :

$$|R_i(jf_0)| + |R_i((j+1)f_0)| = H_j + H_{j+1} \tag{3.13}$$

Les filtres R_i sont des filtres passe-bas. En les sommant, le filtre R ne sera pas égal au spectre cible au niveau des harmoniques malgré des équations 3.12 ou 3.13 satisfaites.

Les conditions ci-dessous permettent d'initialiser les amplitudes de manière un peu plus précise. Nous tirons profit de l'aspect passe-bas des filtres : nous ne considérons l'effet de chaque filtre R_i sur les valeurs de $R(f)$ uniquement si $f \leq f_i$. En commençant par déterminer R_4 comme précédemment (équation 3.12 ou 3.13), il y a deux cas possibles pour déterminer $g_{i<4,init}$:

— Si $f_{i,init}$ est égale à une harmonique j de la fréquence fondamentale
Alors $g_{i,init}$ est calculé pour satisfaire l'équation suivante :

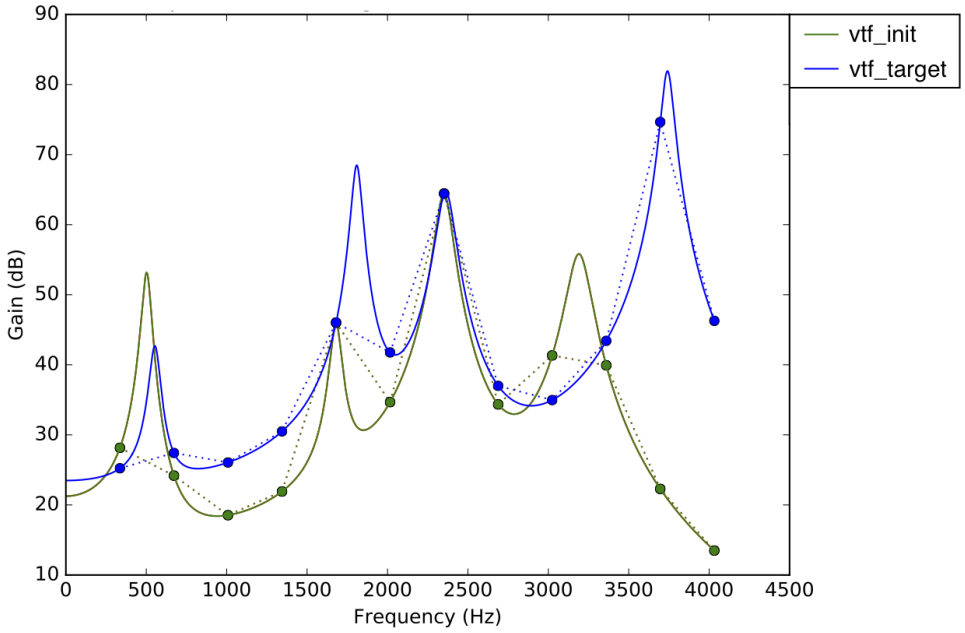
$$|R_i(jf_0)| = H_j - \sum_{k>i}^{4} |R_k(jf_0)| \tag{3.14}$$

— Si $f_{i,init}$ est située entre deux harmoniques j et $j+1$ de la fréquence fondamentale
Alors $g_{i,init}$ est calculé pour satisfaire l'équation suivante :

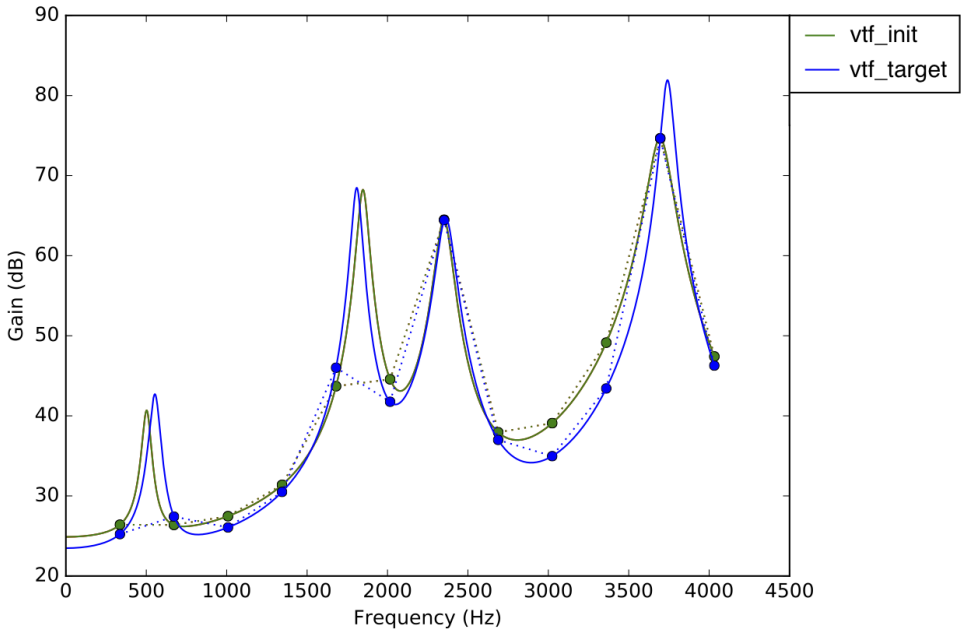
$$|R_i(jf_0)| + |R_i((j+1)f_0)| = H_j + H_{j+1} - \left(\sum_{k>i}^{4} |R_k(jf_0)| + |R_k((j+1)f_0)| \right) \tag{3.15}$$

Ces deux dernières équations permettent de mieux approcher les valeurs H_j avec les valeurs initiales $g_{i,init}$. Une limite est qu'en utilisant les équations 3.14 et 3.15, ces amplitudes peuvent être négatives. L'algorithme est écrit de façon à utiliser 3.12 et 3.13 si cela se produit.

Deux initialisations θ_1 et θ_2 sont présentées à la figure 3.1, elles illustrent les possibilités de la méthode d'initialisation présenté ci-dessus.



(a) Initialisation θ_1



(b) Initialisation θ_2

FIGURE 3.1 – Filtre synthétisé cible (bleu) et deux initialisations de l'ensemble Θ (vert). Les valeurs des spectres aux harmoniques sont représentés par des ronds, les courbes pleines représentent le spectre pour toutes les valeurs intermédiaires, $f_0 = 300$Hz

3.3 Limitations et variante

Le temps de calcul nécessaire à notre méthode l'a rendue difficile à évaluer. Pour des fréquences fondamentales basses (autour de 100Hz), le nombre de partiels à prendre en compte dans la fonction de coût est de plusieurs dizaines. Or, la dimension de l'ensemble Θ, sans

considérer de contraintes au niveau des intervalles de fréquence des formants, est décrite par $dim(\Theta) = \binom{2N}{4}$, avec 4 formants et N harmoniques. Le nombre de descentes de gradients à effectuer pour chaque combinaison de cette ensemble devient donc très grand avec N. Cela vaut aussi en considérant les contraintes des intervalles d'initialisation des fréquences des formants. L'utilisation de notre méthode ne nécessite toutefois pas qu'elle soit rapide : une analyse de la base de données peut avoir lieu une seule fois. Néanmoins, dans le cadre de ce travail, une solution a du être proposée pour réduire les temps de calcul et permettre les tests.

Pour cela, nous avons comparé les temps de calcul de la méthode décrite précédemment et d'une variante moins coûteuse en temps mais moins précise. Il s'agit de n'effectuer aucune descente de gradient lors du parcours de l'ensemble Θ. Chaque élément de cet ensemble est évalué au moyen de la fonction de coût, et la meilleure combinaison de paramètres initiaux est sélectionnée. Cela précède une unique descente de gradient dont l'initialisation utilise cette combinaison. Pour un filtre du conduit vocal à estimer avec une quarantaine d'harmoniques, le temps de calcul est diminué d'un facteur 100. Aussi, la descente de gradient pour chaque initialisation devient moins nécessaire si la f_0 est basse, puisqu'il y a déjà plus d'informations à l'origine (plus d'harmoniques) et qu'il est plus aisé pour ces f_0 d'être plus proche du résultat optimal sans faire les descentes de gradients, puisque l'espace des solutions possibles est mieux échantilloné. La précision et les résultats de la méthode originale sont cependant meilleurs et c'est bien cette approche qui doit être utilisée mais cette variante a été nécessaire pour réaliser des tests.

3.4 Évaluation

Faute de connaître la vérité terrain sur les formants pour les enregistrements des bases de données à notre disposition, il a fallu tester la descente de gradient présentée précédemment sur des filtres du conduit vocal artificiels dont nous générons les formants.

Deux types d'évaluations ont pu être mises en place. La première étudie les résultats de notre estimation avec des signaux artificiels dont les formants sont connus. La seconde compare les résultats de notre approche et ceux de Praat sur des enregistrements de la base de données, où la vérité terrain est inconnue, dans le but de mesurer les variations de chacune des méthodes et de comparer notre méthode à une référence.

Notre objectif étant de modifier les formants avec un modèle de filtre du quatrième ordre, nous avons donc souhaité vérifier ici que ce modèle permettait de reconstituer fidèlement le filtre du conduit vocal original et de conserver l'information de la voyelle et du timbre. Ce travail servirait aussi de comparaison à [13] qui utilise les données estimées par Praat pour des filtres de formants du quatrième ordre et fournit des règles de variation des formants selon l'intensité de la voix.

3.4.1 Praat

Praat applique d'abord un filtrage passe-haut au signal pour réduire la pente spectrale et réduire l'effet du filtre glottique. La pente spectrale à réduire est fixée à -6dB/octave. Après ce filtrage, une méthode de codage prédictif linéaire est appliquée dont l'utilisateur gère certains paramètres : le pas temporel d'analyse, le nombre maximum de formants, le formant maximal en Hertz, la longueur de la fenêtre en secondes, la fréquence de coupure du filtre passe-haut compensant la pente spectrale. Les paramètres sélectionnés pour notre étude suivent les indications de la documentation en ligne `http://www.fon.hum.uva.nl/praat/manual/Source-filter_synthesis_4__Using_existing_sounds.html`. Contrairement à notre méthode où les filtres des formants sont modélisés par un filtre d'ordre quatre, ce logiciel estime chaque formant par deux pôles. Les valeurs estimées de bande passante sont donc incompatibles entre ces deux méthodes.

3.4.2 Tests sur des filtres artificiels

Les filtres du conduit vocal synthétisés pour cette étude s'inspirent du synthétiseur avec des filtres résonateurs en cascade décrit par Klatt [18]. Ce modèle est très commun en traitement de la parole d'où son utilisation ici. Voici la formule de ce modèle de filtre résonateur, avec BW sa bande passante, f_i sa fréquence centrale et T_s la période d'échantillonnage :

$$T_i(f) = \frac{A}{1 - Bz^{-1} - Cz^{-2}} \tag{3.16}$$

où :

$$\begin{aligned} z &= exp(j\,2\,\pi\,f\,T_s) \\ C &= -e^{-\,2\,\pi\,BW\,T_s} \\ B &= e^{-\,\pi\,BW\,T_s} cos(2\,\pi\,f_i\,T_s) \\ A &= 1 - B - C \end{aligned}$$

Paramètres

Les paramètres de la synthèse des signaux de ce test sont générés aléatoirement. Seuls leurs intervalles de variations ont été fixés (valeurs issues de [30] [31]) :

Paramètre	Valeur minimale	Valeur maximale
f_0	100 Hz	600 Hz
f_1	250 Hz	860 Hz
BW_1	30 Hz	90 Hz
f_2	850 Hz	2250 Hz
BW_2	35 Hz	110 Hz
f_3	2200 Hz	3000 Hz
BW_3	46 Hz	170 Hz
f_4	2600 Hz	3900 Hz
BW_4	50 Hz	250 Hz

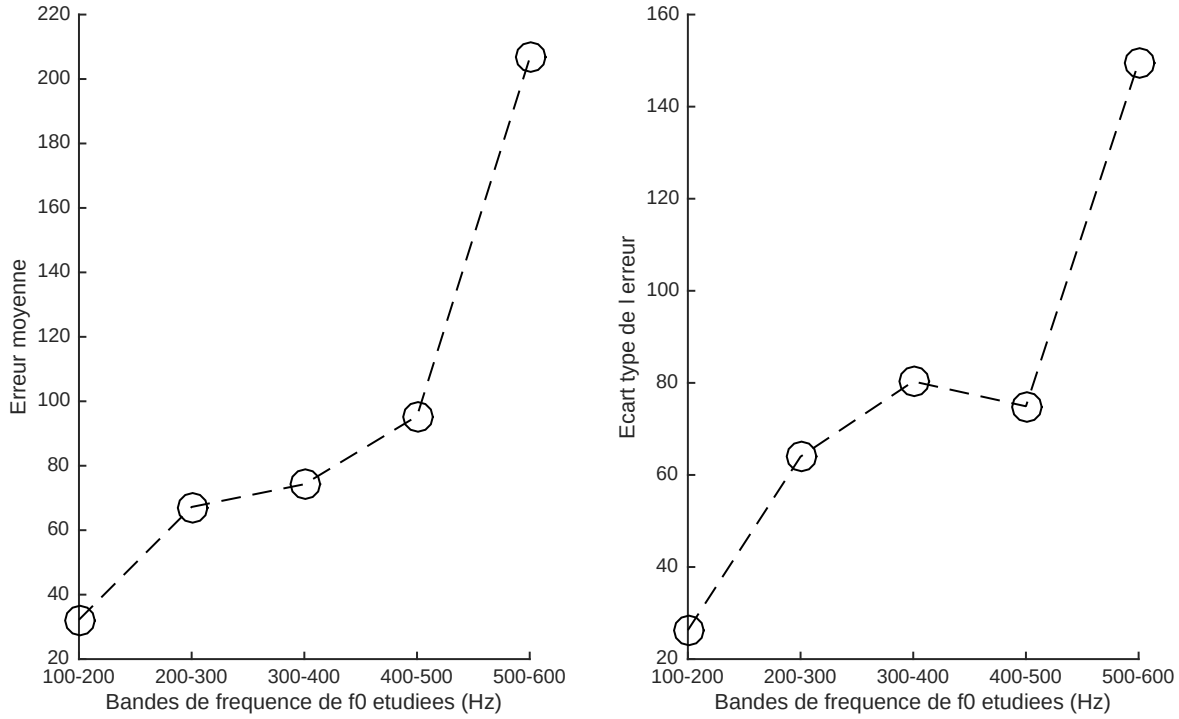
TABLE 3.1 – Intervalles de variation des paramètres des signaux de test

Nous avons synthétisé 30 filtres pour 5 bandes de fréquence fondamentale différentes : $100-200$Hz, $200-300$Hz, $300-400$Hz, $400-500$Hz, $500-600$Hz. En effet, la fréquence fondamentale qui définit le nombre d'harmoniques considérés et donc la quantité d'information sur le filtre du conduit vocal est le paramètre déterminant de ce type d'analyse.

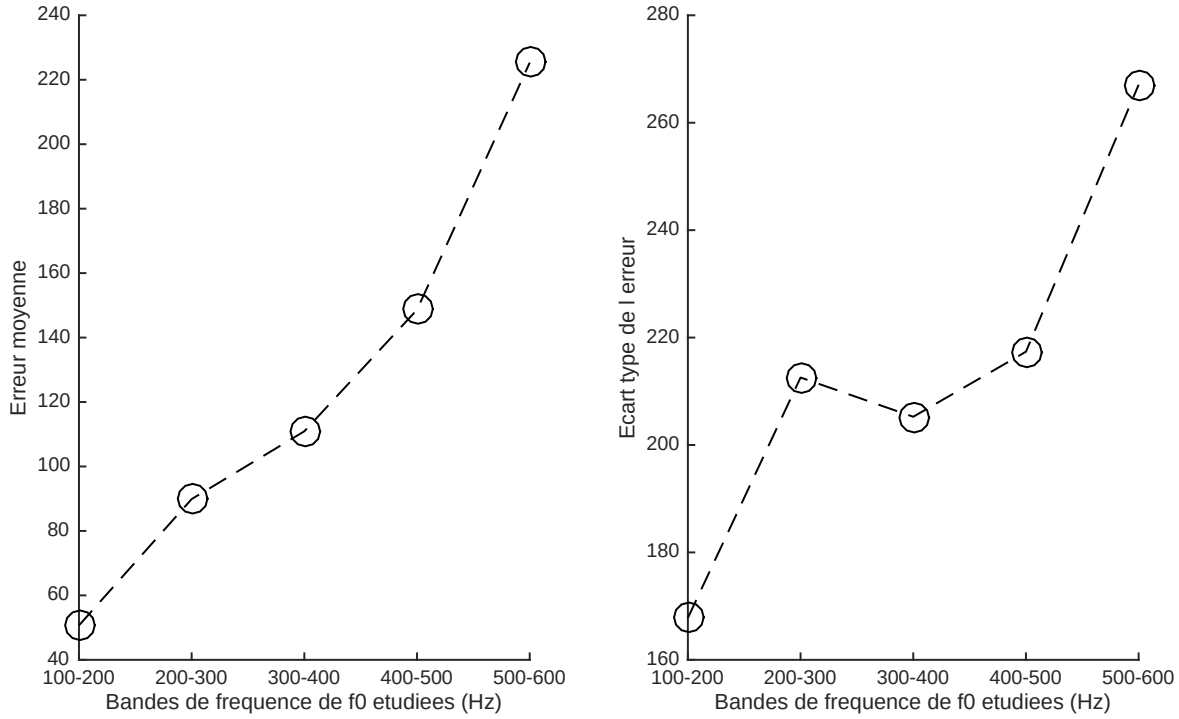
Résultats

Pour chaque bande de fréquence fondamentale, la moyenne de l'erreur entre paramètres utilisés pour la synthèse et paramètres estimés a été calculée ainsi que son écart-type pour relater de la variabilité des estimations.

La figure 3.2 illustre ces résultats pour l'estimation de la fréquence du premier et du second formant. Les résultats pour les troisième et quatrième formants figurent en annexe 3.4.3.



(a) 1er formant f_1



(b) 2ème formant f_2

FIGURE 3.2 – Moyenne et écart-type de l'erreur d'estimation de la fréquence des deux premiers formants. 30 estimations sur 5 bandes de fréquence de f_0 avec des spectres différents sont à l'origine de ces données

La fréquence fondamentale apparaît bien déterminante dans l'estimation des formants avec

notre système. L'erreur est de plus en plus grande quand f_0 grandit. L'estimation de la fréquence du premier formant montre une moyenne et un écart-type de l'erreur plutôt bas pour f_0 entre 100 et 200Hz : autour de 20 Hz pour ces deux grandeurs. La fréquence du second formant, qui varie ici entre 850 et 2250 Hz pourraît sembler correctement estimée mais l'étude de l'écart-type de son erreur d'estimation est très grand. De manière générale, l'écart-type de l'erreur de l'estimation des paramètres est très grande dans ces résultats. Les résultats montrent que les estimations du formant 3 et du formant 4 sont peu robustes et ne permettent plus l'observation de la dépendance à la fréquence fondamentale à laquelle nous nous attendions, comme sur la figure 3.2.

L'estimation des bandes passantes possède aussi de fortes variations. Avec une erreur moyenne d'au moins 20 Hz sur toutes les bandes de fréquence de f_0 et un écart-type de l'erreur d'au moins 20 Hz, notre méthode semble mal estimer ce paramètre.

L'analyse de ces résultats un par un a révélé que la méthode pouvait engendrer de fortes erreurs selon l'initialisation sélectionnée pour la descente de gradient. Si cette initialisation est loin de la réalité malgré le fait d'avoir l'erreur minimum sur notre ensemble de combinaisons possibles, alors les paramètres estimés après la descente de gradient seront également très erronnés. Les fortes valeurs de l'erreur moyennes et de l'écart-type de l'erreur sont principalement le résultat de quelques importantes erreurs d'estimation qui ont perturbé l'évaluation de la qualité de notre méthode. La variante présentée au 3.3 et testée ici a ainsi montré une limite. Sur quelques-unes de ces estimations de filtres présentant une forte erreur, la méthode avec une descente de gradient sur chaque combinaison d'initialisation possible a été testée et a montré de biens meilleurs résultats. Mais le temps que prend cette méthode originale est considérable et il n'a pas été possible de la traiter dans celui qui était imparti. Un test prolongé de cette méthode pourra préciser le comportement en fonction du paramètre f_0 et à mettre au point des résultats quantitatifs plus significatifs.

3.4.3 Tests sur des enregistrements

Parallèlement à l'étude sur des filtres synthétiques, l'autre étude réalisée sur notre système concerne les enregistrements à notre disposition sur la base de données. Pour traiter ces signaux, le filtre du conduit vocal a été estimé toutes les 70 ms, ce qui a permis d'observer les résultats de la méthode proposée sur un plus grand nombre d'enregistrements. Aussi, elle n'est appliquée qu'aux instants ou R_d est estimé, c'est-à-dire dans les parties voisées.

Il s'agit de comparer les estimations de la variante de notre méthode avec les estimations de Praat puis de resynthétiser ces enregistrements sous une forme simple : la source estimée avec le modèle LF filtrée par des filtres du conduit vocal estimés. Les filtres utilisés pour resynthétiser les enregistrements d'après les estimations de Praat sont les filtres du second ordre présentés en 3.4.2, dans une structure en cascade.

Enregistrements utilisés

Plusieurs bases de données ont été sujets à cette étude :
— 15 voyelles chantées à 5 nuances différentes par Raphaël (f_0 est d'environ 115Hz)
— 30 premiers mots de la base de données de chant de Raphaël (f_0 est d'environ 115Hz)
— 30 premiers mots de la base de données de chant de Marlène (f_0 est d'environ 315Hz)

Résultats

Pour ces enregistrements de Raphaël, notre méthode fournit des résultats très proches de ceux du logiciel Praat. Les formants sont estimés avec une variabilité faible, comme l'illustre la

figure suivante :

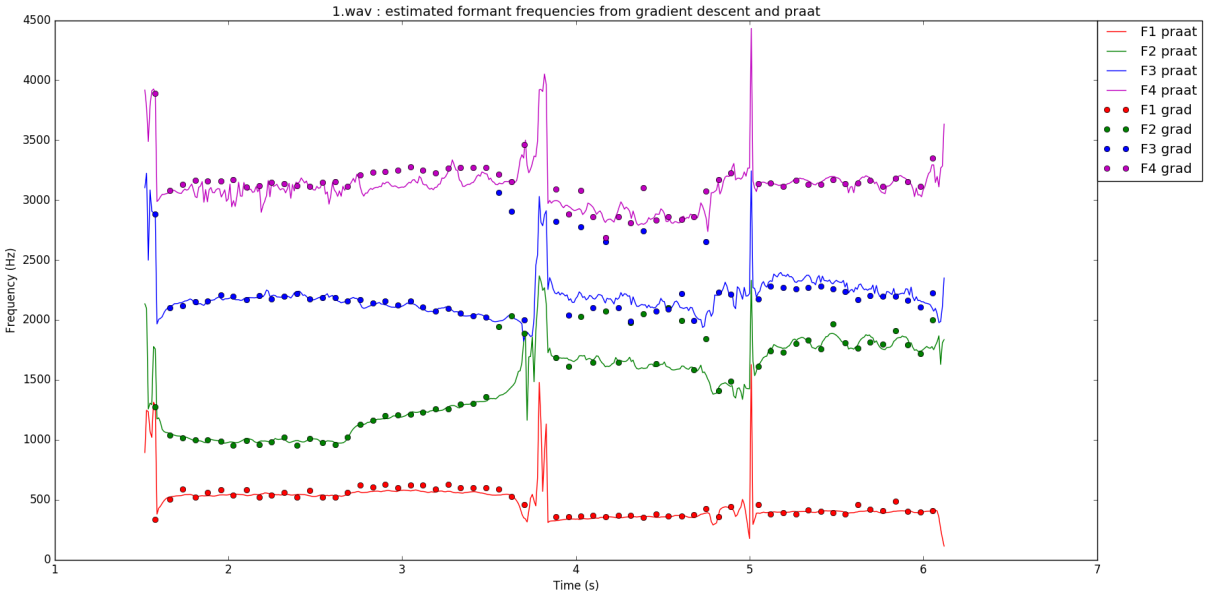


FIGURE 3.3 – Comparaison de l'estimation des fréquences des formants pour le mot "co-inculpé" chanté par Raphaël

L'erreur entre les fréquences estimés par les deux méthodes est faible mais Praat a des résultats plus lisses et plus constants, comme le montre la figure. Il arrive en effet qu'un formant soit estimé par deux filtres simultanément. Ce défaut induit des sauts de formants importants : autour de 4.3 secondes dans la figure 3.3, le deuxième et le troisième formant peuvent être estimés autour de l'estimation du troisième formant selon Praat. Pour d'autres trames d'analyse autour de cet instant, la distinction est correctement effectuée.

Les méthodes ne se comportent de la même façon pour la voyelle nasale /o $\tilde{/}$ ("b**on**"). D'après [32], cette nasalisation peut provoquer l'atténuation voire la disparition du deuxième formant pour les hommes sujets lors de cette étude, ou encore le déplacement de ce formant vers de plus hautes fréquences. Un couple de formants proches est estimé par chacune des méthodes pour le mot "onzième" chanté par Raphaël, les deux formants étant plus éloignés pour notre méthode que pour celle du logiciel Praat. Ces observations sont présentées à la figure 3.4.

Les synthèses de ce mot à partir de ces deux analyses ne nous ont pas permis de déterminer la meilleure estimation de façon perceptive. Une analyse perceptive plus approfondie permettrait d'évaluer la qualité de notre modèle de façon perceptive. Les résultats ont pu montrer que la variante de notre méthode est suffisamment efficace pour s'approcher nettement des estimations de Praat.

Les analyses des enregistrements de Marlène ont montré les limites de chacune des méthodes. Les estimations possèdent de fortes variations d'une trame à une autre. Cela provient du fait qu'un formant peut ne pas être détecté, ce qui décale en fréquence plusieurs formants estimés. La figure 3.5 présente ainsi les estimations pour le mot "co-inculpé".

Cette analyse montre bien qu'il aurait été difficile d'utiliser les données de Praat pour l'estimation des formants sans une certaine correction manuelle des formants. Notre méthode présente le même défaut et n'est, sous cette forme en particulier, pas suffisante pour identifier de manière robuste les formants et permettre leur modification si la fréquence fondamentale est trop élevée.

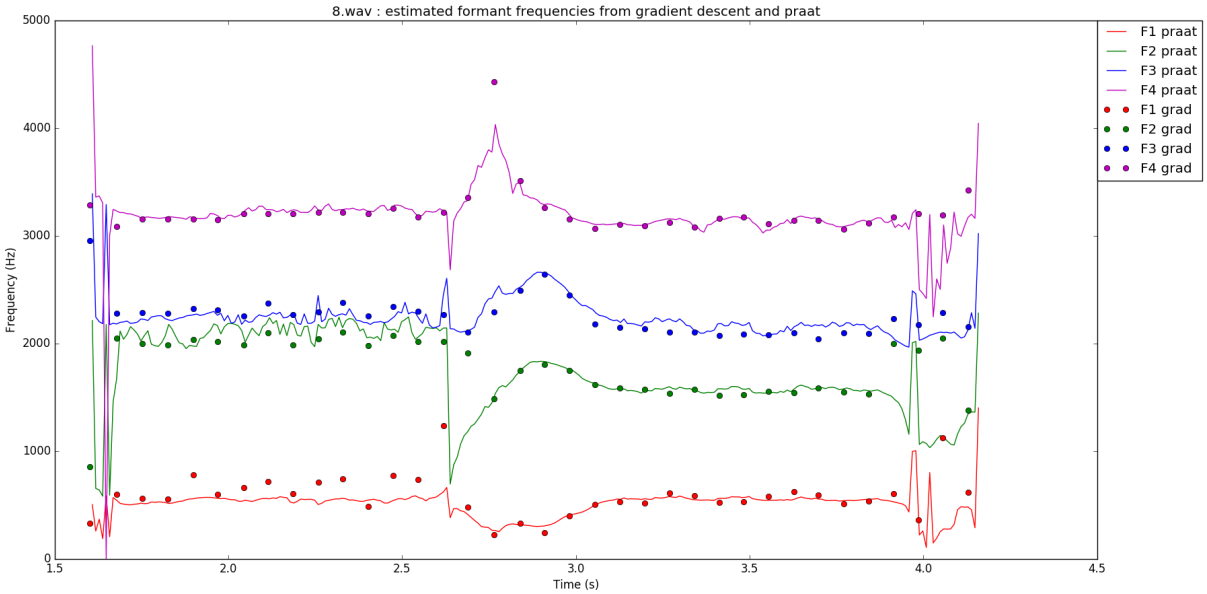


FIGURE 3.4 – Comparaison de l'estimation des fréquences des formants pour le mot "onzième" chanté par Raphaël

Malgré une évaluation sur des filtres synthétiques qui indiquait une faible robustesse de notre méthode, la comparaison au logiciel Praat a montré son efficacité. Les segments chantés par Raphaël ont donné de meilleurs résultats, démontrant la dépendance à la fréquence fondamentale de notre méthode mais aussi celle utilisée par Praat. Les deux méthodes ne se comportent pas de la même façon lorsque deux formants sont proches et plus de tests sont nécessaires pour départager la meilleure possibilité. Notre méthode plus coûteuse en calcul que la variante qui a été testée ici permettrait d'améliorer l'estimation des formants et pourrait faire l'objet d'une étude approfondie. La méthode présentée ici amène à penser à une règle de modification générale qui peut provenir de l'analyse des voyelles chantées à différentes nuances. Une autre façon de modéliser des variations des formants selon l'intensité serait d'interpoler les différents paramètres des formants entre différentes intensités et de synthétiser les filtres du conduit vocal correspondant. Ce procédé diffère de la déformation de l'enveloppe car elle permet de juxtaposer des formants et pourrait permettre de modéliser le formant du chanteur.

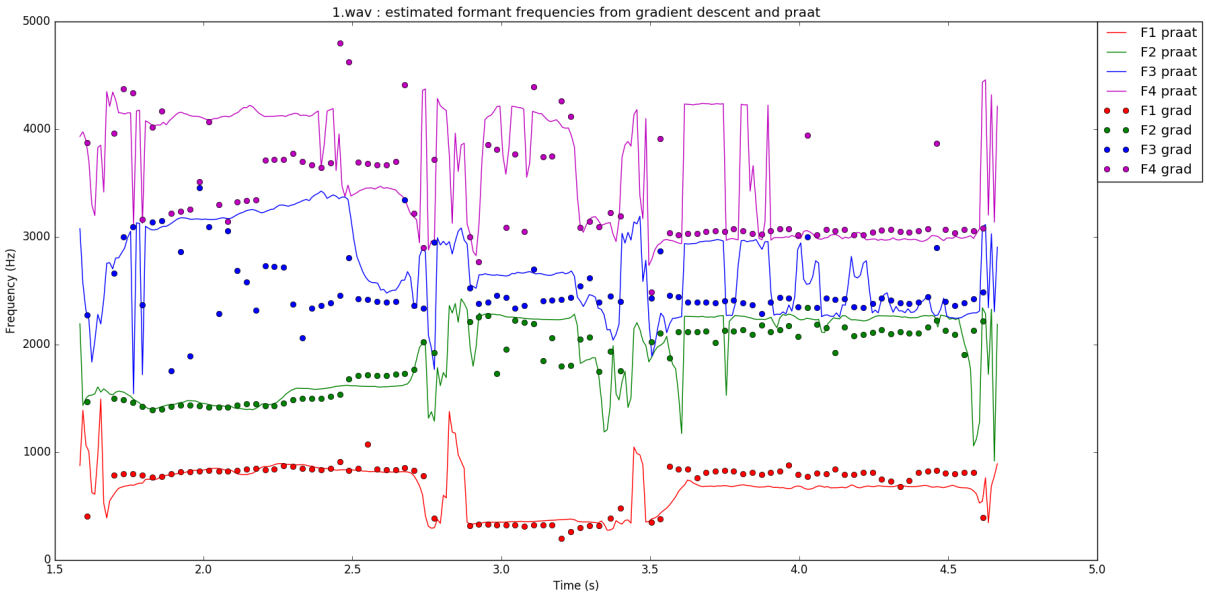


FIGURE 3.5 – Comparaison de l'estimation des fréquences des formants pour le mot "co-inculpé" chanté par Marlène

Conclusion

Le travail effectué durant ce stage s'est intéressé à la modification de l'effort dans la voix et a abouti sur la mise au point d'une méthode permettant de lier intensité sonore perçue et effort vocal. L'analyse d'enregistrements à différentes nuances a montré la complexité de relier l'effort au paramètre de forme R_d de la source du modèle LF. Elles ne laissaient paraître aucune règle de modification de ce paramètre en fonction de la nuance. Le lien entre ce paramètre, l'intensité perçue et la voyelle chantée était également si complexe qu'une règle propre à chaque voyelle s'est avérée nécessaire. Un postulat a donc été émis en étudiant une règle de modification, due à Fant, de R_d et de l'énergie de la source E_e en fonction de l'effort. La problématique liant cette règle avec l'intensité perçue a été soulevée par une étude basée sur les filtres du conduit vocal estimé sur des enregistrements de la base de données. La règle a été paramétrée par cette étude pour faire varier R_d et E_e avec différents coefficients selon la note chantée, la voyelle prononcée et le gain d'intensité sonore souhaité. Un test perceptif a prouvé son efficacité dans une seule des trois voyelles proposées aux participants, récoltant des scores moindres dans les autres cas.

La nécessité d'estimer les formants pour compléter un modèle de règle de modification de la source et du filtre de la voix a mené à considérer quatre filtres d'ordre quatre en parallèle. Ce modèle a été estimé par une descente de gradient dont la fonction de coût n'utilise que les valeurs du filtre du conduit vocal au niveau des harmoniques de la note chantée. Un grand ensemble de combinaisons initiales a été introduit pour parer le nombre important de minima locaux et une initialisation complexe des gains des filtres a été proposée. Les temps de calculs trop importants nous ont empêchés de tester cette méthode et nous avons donc testé à la place une variante simplifiée. Des tests sur signaux artificiels ont montré que d'importantes erreurs pouvaient être produites mais n'ont pas invalidé la méthode. D'autres tests sur des enregistrements ont pu la comparer au logiciel Praat et montrer son intérêt, ses similitudes avec ce logiciel ainsi que les limitations des deux méthodes avec une fréquence fondamentale plus importante. La synthèse des extraits chantés à partir des estimations des deux méthodes et de celle de la source a appuyé ces résultats et la possibilité de considérer une telle méthode pour l'estimation des formants dans notre modèle.

En prolongement de ce travail, un second test perceptif pourra être effectué pour corriger les lacunes du premier, notamment sa trop grande complexité. Cela devrait permettre de mettre en valeur l'atout de la règle de modification de la source de la voix proposée pour une voix chantée plus relâchée. La méthode de descente de gradient pourraît être testée sur plus de signaux et d'enregistrements et sous sa première forme au lieu de la variante qui s'est montrée nécessaire ici. Des résultats plus significatifs pourraient ainsi être obtenus pour montrer son efficacité. Egalement, de la même manière que cette approche profite de connaissance à priori sur le filtre du conduit vocal, plusieurs améliorations sont possibles : utiliser l'information de la voyelle chantée pour contrôler l'initialisation ou le résultat de la descente de gradient, tirer profit de la continuité du conduit vocal au sein de voyelle pour lier l'initialisation de la descente de gradient sur chaque trame au résultat de la trame précédente, ajouter de l'apprentissage pour optimiser les estimations et leur continuité.

Annexe A : Scripts Praat pour l'analyse de formants

Script formant_analysis.praat

```
form Test command line calls
sentence Directory sounds_folder
sentence Directory2 analysis_folder
endform

strings = Create Strings as file list: "list", directory$ + "*.wav"
numberOfFiles = Get number of strings
for ifile to numberOfFiles
    selectObject: strings
    fileName$ = Get string: ifile
    Read from file... 'directory$''fileName$'
    name$ = selected$("Sound")
    # Read from file... 'directory$''name$'.wav
    # For a male voice : max_freq=5000, for a female voice : max_freq=5500
    To Formant (burg)... 0.01 5 5000 0.025 50
    select Formant 'name$'
    Save as text file... 'directory2$''name$'.txt
endfor
```

Appel Praat en ligne de commande

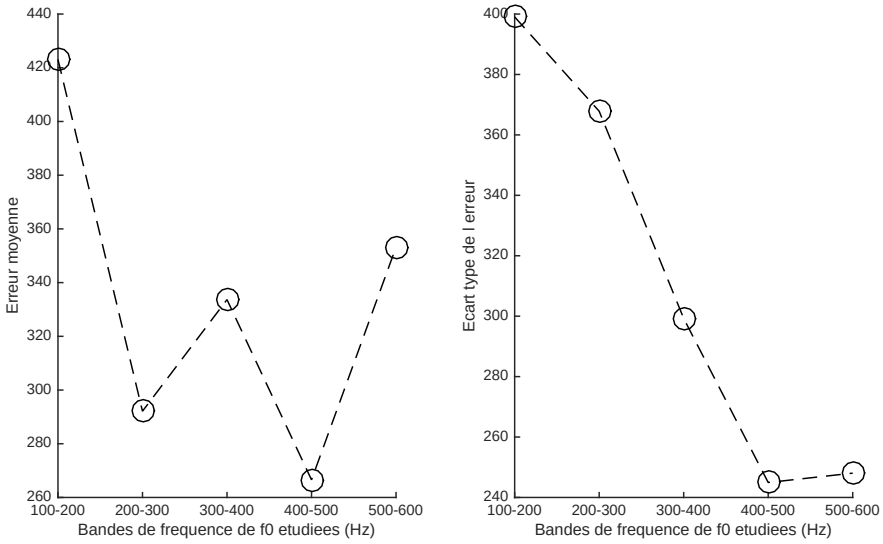
```
> Praat --run formant_analysis.praat sounds_folder analysis_folder
```

Annexe B : Aperçu d'un fichier de résultat de Praat

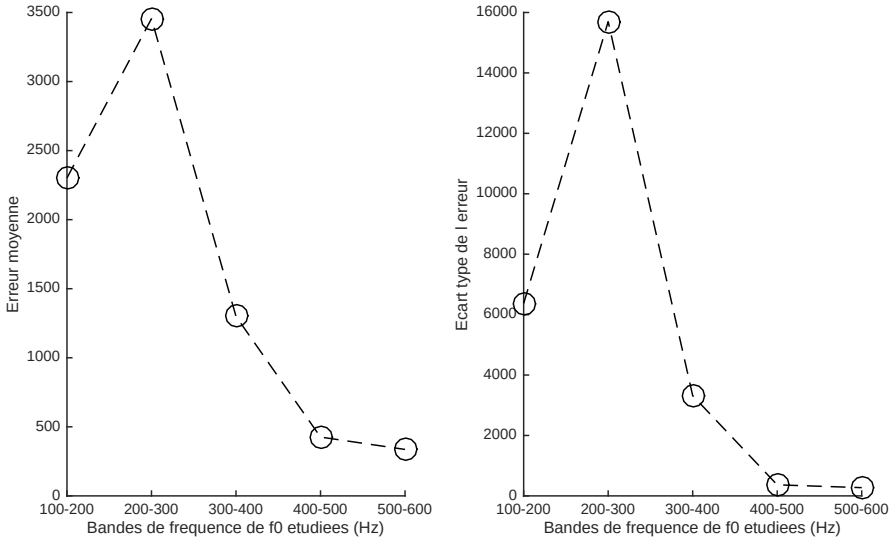
```
File type = "ooTextFile"
Object class = "Formant 2"

xmin = 0
xmax = 1
nx = 96
dx = 0.01
x1 = 0.024999999999999967
maxnFormants = 5
frames []:
    frames [1]:
        intensity = 0.024377370981765505
        nFormants = 5
        formant []:
            formant [1]:
                frequency = 250.74490073219218
                bandwidth = 555.5070451269141
            formant [2]:
                frequency = 1214.0877686247263
                bandwidth = 1112.144361630217
            formant [3]:
                frequency = 2205.549409291365
                bandwidth = 1295.1985206795252
            formant [4]:
                frequency = 3205.9923436292247
                bandwidth = 1376.9290726002223
            formant [5]:
                frequency = 4234.392607740555
                bandwidth = 1364.5929573161209
    frames [2]:
        intensity = 0.024134047916236034
 . . .
```

Annexe C : Estimation des troisième et quatrième formants



(a) 3ème formant f_3



(b) 4ème formant f_4

FIGURE 6 – Moyenne et écart-type de l'erreur d'estimation de la fréquence des troisième et quatrième formants

Bibliographie

[1] G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2 :3, 1995.

[2] Hideki Kenmochi and Hayato Ohshita. Vocaloid - commercial singing synthesizer based on sample concatenation. In *Proc. INTERSPEECH*, pages 4009–4010, 2007.

[3] Gilles Degottex. *Glottal source and vocal tract separation.* PhD thesis, UPMC-IRCAM-UMR9912-STMS, Paris, France, 2010.

[4] Xavier Rodet. Synthesis and processing of the singing voice. In *In Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002*, 2002.

[5] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg. Expression control in singing voice synthesis : Features, approaches, evaluation, and challenges. *IEEE Signal Processing Magazine*, 32(6) :55–73, Nov 2015.

[6] Gilles Degottex, Pierre Lanchantin, Axel Roebel, and Xavier Rodet. Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Commun.*, 55(2) :278–294, February 2013.

[7] Stefan Huber and Axel Roebel. On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system. In *16th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, Dresden, Germany, September 2015. ISCA.

[8] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QSPR*, 26(4) :1–13, 1982.

[9] Gunnar Fant. The voice source in connected speech. *Speech Communication*, 22(2) :125–139, 1997.

[10] Ingo R. Titze and Johan Sundberg. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5) :2936–2946, 1992.

[11] Johan Sundberg. *The Science of the Singing Voice.* Northern Illinois University Press, 1987.

[12] J. Bonada, Ò. Celma, A. Loscos, J. Ortolà, and X. Serra. Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models. In *International Computer Music Conference*, Havana, Cuba, 15/09/2001 2001. Singing Voice.

[13] Emilio Molina, Isabel Barbancho, Ana M. Barbancho, and Lorenzo J. Tardón. Parametric model of spectral envelope to synthesize realistic intensity variations in singing voice. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 634–638. IEEE, 2014.

[14] Oriol Nieto. Voice transformations for extreme vocal effects. Master's thesis, Pompeu Fabra University, 2008.

[15] H. Traunmüller and A. Eriksson. Acoustic effects of variation in vocal effort by men, women, and children. *Acoustical Society of America Journal*, 107 :3438–3451, June 2000.

[16] Jean S. Liénard and Maria G. Di Benedetto. Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America*, 106(1) :411–422, July 1999.

[17] T. Saitou, M. Goto, M. Unoki, and M. Akagi. Speech-to-singing synthesis : Converting speaking voices to singing voices by controlling acoustic features unique to singing voices. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 215–218, Oct 2007.

[18] Dennis H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3) :971–995, 1980.

[19] Nathalie Henrich, John Smith, and Joe Wolfe. Vocal tract resonances in singing : Strategies used by sopranos, altos, tenors, and baritones. *Journal of the Acoustical Society of America*, 129(2) :1024–1035, 2011.

[20] Nathalie Henrich Bernardoni, John Smith, and Joe Wolfe. Vocal tract resonances in singing : variation with laryngeal mechanism for male operatic singers in chest and falsetto registers. *Journal of the Acoustical Society of America*, 135(1) :491–501, 2014.

[21] Johan Sundberg, Filipa M. B. Lã, and Brian P. Gill. Formant tuning strategies in professional male opera singers. *Journal of Voice*, 27(3) :278–288, 2012.

[22] Martine E. Bestebreurtje and Harm K. Schutte. Resonance strategies for the belting style : Results of a single female subject study. *Journal of Voice*, 14(2) :194–204, 2000.

[23] John E. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[24] Gautam K. Vallabha and Betty Tuller. Systematic errors in the formant analysis of steady-state vowels. *Speech Commun.*, 38(1) :141–160, September 2002.

[25] D. Vincent, O. Rosec, and T. Chonavel. Estimation of LF glottal source parameters based on ARX model. In *Interspeech-Eurospeech, Lisboa, sept. 05*, 2005.

[26] Med Ali Kammoun, Dorra Gargouri, Mondher Frikha, and A Ben Hamida. Cepstral method evaluation in speech formant frequencies estimation. In *Industrial Technology, 2004. IEEE ICIT'04. 2004 IEEE International Conference on*, volume 3, pages 1612–1616. IEEE, 2004.

[27] Satoshi Imai and Yoshiharu Abe. Spectral envelope extraction by improved cepstral method. *Journal of IEICE*, 62 :217–223, 1979.

[28] ITURBS Recommendation. 1284-1 : General methods for the subjective assessment of sound quality. *International Telecommunications Union, Geneva*, 2003.

[29] Martin F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *NEURAL NETWORKS*, 6(4) :525–533, 1993.

[30] Laurianne Georgeton, Nikola Paillereau, Simon Landron, Jiayin Gao, and Takeki Kamiyama. Analyse formantique des voyelles orales du français en contexte isolé : à la recherche d'une référence pour les apprenants de FLE. In *Conférence conjointe JEP-TALN-RECITAL 2012*, pages 145–152, Grenoble, France, June 2012.

[31] G Fant. Formant bandwidth data. 1962.

[32] Véronique Delvaux, Thierry Metens, and Alain Soquet. Propriétés acoustiques et articulatoires des voyelles nasales du français. *XXIVèmes Journées d'étude sur la parole, Nancy*, 1 :348–352, 2002.

Table des figures

Liste des tableaux