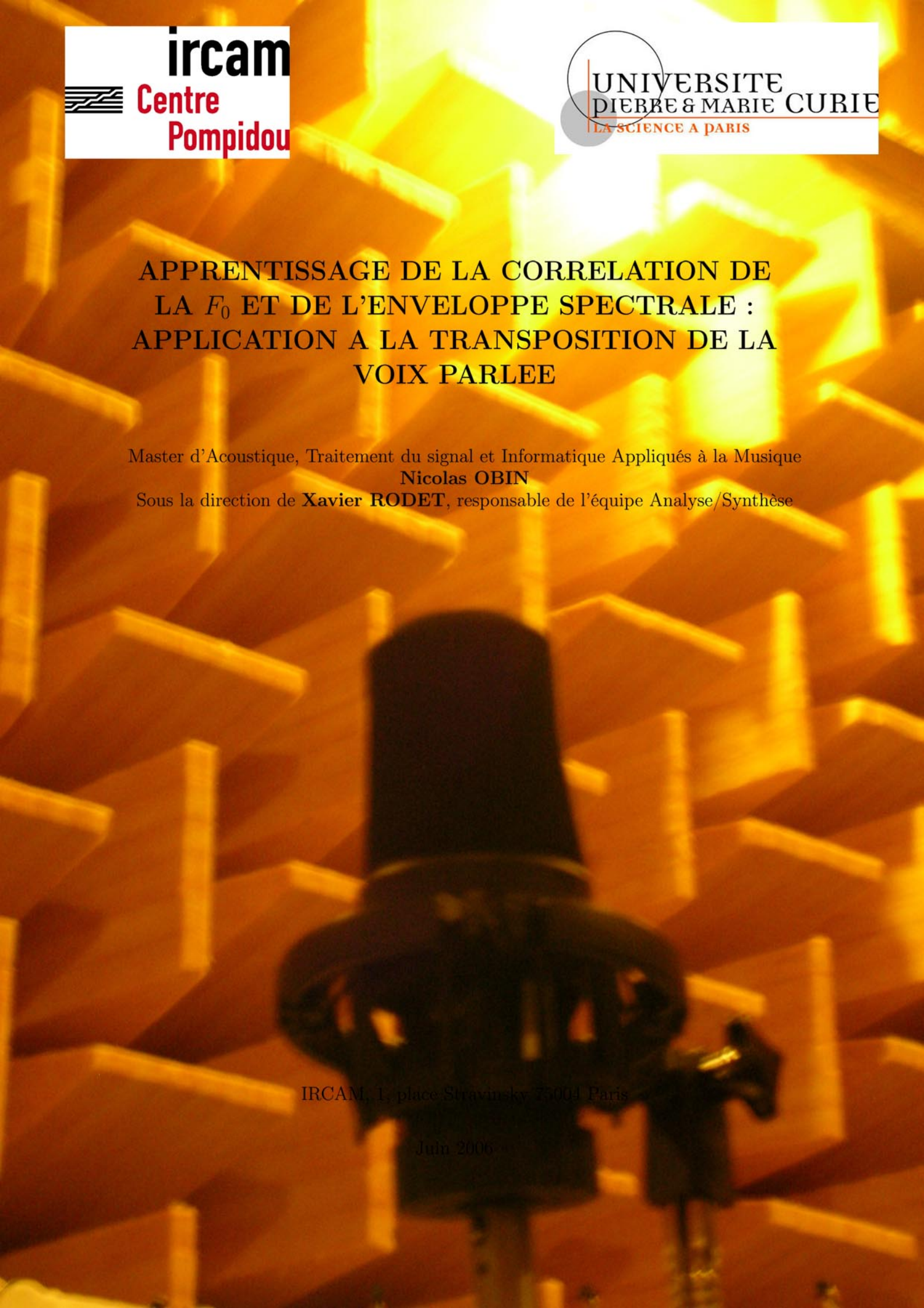


APPRENTISSAGE DE LA CORRELATION DE LA F_0 ET DE L'ENVELOPPE SPECTRALE : APPLICATION A LA TRANSPOSITION DE LA VOIX PARLEE

Master d'Acoustique, Traitement du signal et Informatique Appliqués à la Musique

Nicolas OBIN

Sous la direction de **Xavier RODET**, responsable de l'équipe Analyse/Synthèse

IRCAM, 1, place Stravinsky 75004 Paris

Juin 2006

Etudiant

Nicolas OBIN

Lieu

IRCAM - Equipe Analyse-Synthèse

Responsable

Xavier RODET

1

Table des matières

Table des figures

Résumé

Nous présentons dans ce rapport un travail visant à améliorer les résultats de transposition de la voix parée actuels. Partant de la constatation que les méthodes de transposition actuelles, comme le vocodeur de phase et PSOLA, ne sont pas encore optimales pour des transpositions plus grande approximativement que le doublement de fréquence, notre travail consiste à mettre en évidence et à modéliser la corrélation entre la fréquence fondamentale et l'enveloppe spectrale par un apprentissage sur une large base de données. Nous mettons en évidence l'importance de l'enveloppe spectrale lors de la transposition à partir de quelques simples exemples, pour lesquels nous avons constitué une base de données spécifique. Forts de ce résultat, et prolongeant les recherches qui ont été menées en la matière [Sty95], [Sty96], [Sty98], [Kai00, [Kai01]], nous réalisons un apprentissage supervisé de cette corrélation par un modèle de mélange de gaussiennes sur les classes phonétiques voisées de la langue française. Nous montrons enfin que le résultat est meilleur que celui obtenu par les méthodes classiques.

Introduction

La transposition est une méthode de modification du signal qui consiste à modifier la hauteur d'un son par modification de sa fréquence fondamentale. Appliquée à la voix parlée, la transposition est un enjeu majeur dans les systèmes de synthèse de parole à partir du texte, dit *text to speech*, dans la synthèse de voix expressive ou dans les méthodes de conversion de voix [Kai01], [Sty95], [Sty96], [Sty98]. Ce type de systèmes de synthèse repose généralement sur un corpus réduit d'occurences de paroles que l'on doit modifier de manière à obtenir l'intonation souhaitée. Cette modification se fait soit par des méthodes d'analyse, de modification, et de synthèse du signal comme le vocodeur de phase [Dol86], [Roe05a], [Roe05b], ou bien les méthodes d'addition et de superposition synchrone à la période fondamentale [Pee01]. Ces méthodes procurent généralement des résultats suffisament naturels pour être couramment utilisées. Néanmoins celui-ci se dégrade au fur et à mesure que l'on augmente le facteur de transposition, perdant le naturel et l'identité de la voix du locuteur originel. Ceci semble principalement dû au fait que lorsque nous parlons à des hauteurs différentes, les résonnateurs et les articulateurs sont modifiés, provoquant ainsi une modification de l'enveloppe spectrale ; modification qui n'est pas prise en compte par les méthodes de transposition qui supposent l'enveloppe spectrale invariante avec la fréquence d'excitation, en cohérence avec l'approximation du modèle source-filtre. Des travaux ont déjà montré la corrélation de l'enveloppe spectrale et de la fréquence fondamentale dans le cas de la voix parlée [EnN03], [Kai00], [Sha], [Mil]. Cette corrélation n'est aujourd'hui pas prise en compte par les méthodes de transposition. Notre démarche s'appuie sur des travaux récents portant sur la conversion de la voix qui ont tenté de modéliser la corrélation.

Ce rapport s'articule de la manière suivante : nous présentons dans une première partie des caractéristiques de la production vocale nécessaire è la compréhension des enjeux de la transposition de la voix parlée. Dans une seconde partie, nous définissons l'enveloppe spectrale, présentons et comparons les modèles et les méthodes permettant son estimation. Dans une troisième partie, nous présentons les méthodes actuelles permettant la transposition de la voix parlée : le vocodeur de phase, et la méthode d'addition et superposition synchrone à la période fondamentale PSOLA, et explicitons leur limites, qui provient principalement de l'approximation du filtre comme invariant avec la fréquence d'excitation, en cohérence avec le modèle source-filtre. Dans une cinquième partie, nous mettons en évidence à travers une expérience l'importance de l'enveloppe spectrale sur le résultat de la transposition. Dans une sixième partie, nous proposons une méthode permettant l'apprentissage du modèle de l'évolution de l'enveloppe spectrale en fonction de la fréquence fondamentale reposant sur l'apprentissage par mélange de gaussiennes. Enfin, nous utilisons ce modèle pour mettre en évidence cette corrélation et permettre sa prédiction, et montrons que sa prédiction améliore le résultat de la transposition par rapport au modèle de conservation de l'enveloppe spectrale.

Chapitre 1

Quelques caractéristiques de la voix parlée

Nous nous proposons dans ce chapitre de présenter quelques caractéristiques de la voix parlée utiles dans le cadre de notre démarche. Ce chapitre s'articule de la manière suivante : nous présentons la production vocale du point de vue de sa modélisation physique, puis le modèle prosodique de la production vocale, avant de finir par une présentation d'un descripteur particulier de la production vocale, à savoir le noyau vocalique.

1.1 La modélisation source-filtre

D'un point de vue physique, l'appareil phonatoire, instrument permettant la production de la parole, est consitué d'un générateur sonore - production de trains d'impulsions ou de bruit dû à un écoulement d'air turbulent -, d'un ensemble de cavités réglables - les résonateurs buccal, labial et nasal - et d'articulateurs permettant la modification de ces cavités. L'ensemble est modelisé en première approximation comme la réponse d'un filtre à un train d'impulsions de la source glottique, soit un modèle source-filtre. Lorsque la source d'excitation est une vibration presque périodique, la parole est dite « voisée » , ce qui est le cas des voyelles et de certaines consonnes dans la langue française.

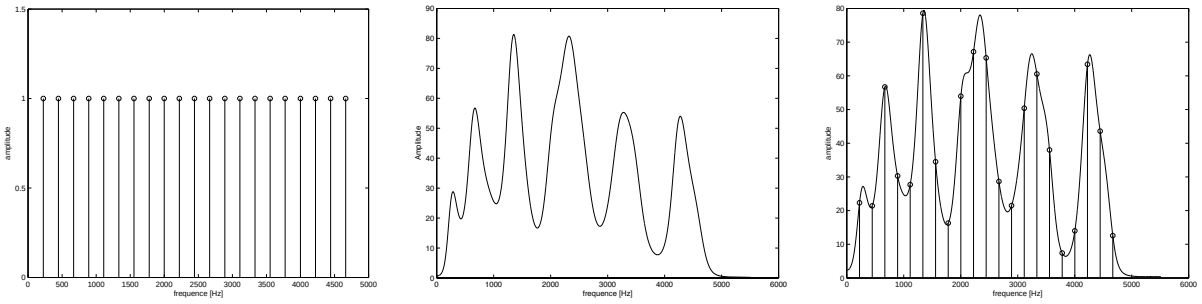
Si e est le signal source, et h le filtre, la réponse impulsionnelle s'écrit par définition :

$$s(t)\widehat{=}(e \otimes h)(t) \tag{1.1}$$

Et la réponse fréquentielle :

$$S(\omega)=E(\omega).H(\omega) \tag{1.2}$$

Si e est presque périodique d'enveloppe spectrale plate de valeur 1 [1.6(a)], alors la réponse fréquentielle du filtre [1.6(b)] est aussi l'enveloppe spectrale de s [1.6(c)].



(a) (b) (c)

FIG. *1.4:* Le modèle source-filtre : spectre du signal source (a) ; réponse frequentielle du filtre (b) ; spectre résultant (c)

1.2 Le formant

Les résonances de la cavité bucale entraînent de fortes concentrations énergetiques aux fréquences de résonnance dans le spectre : ces résonnances sont appelés formants . La définition du formant varie selon les auteurs : pour certains, il s'agit uniquement des résonnances de la cavité buccale [Fan60], alors que d'autres font la distinction entre formants vocaliques - les résonnances de la cavité -, et le formant glottique - provenant de la glotte [Dov03]. Les formants sont caractéristiques de la production vocale humaine : il sont discriminants vis-à-vis de la voyelle prononcée, et jouent un rôle essentiel pour la compréhension (le premier formant) et pour le « timbre » de la voix (le deuxième formant).

1.3 La production de la parole

Le modèle source-filtre que nous venons de voir, s'il donne une idée du mécanisme acoustico-physique en jeu lors de la production de la parole, est cependant insuffisant pour la décrire précisement. L'être humain est avant tout un « être communiquant » et la production vocale un instrument de communication. Les sons que l'homme produit ne sont pas n'importe quel type de sons : la parole s'articule autour d'unités de bases, les phonèmes, pour donner naissance à la langue. Ainsi, la production de la parole est le lieu d'un double codage, acoustique et syntaxico-sémantique [Fon71].

1.3.1 Notion de phonétique

Le phonème est « la plus petite unité distinctive du langage oral ». Les phonèmes du français sont intrinsèquement regroupés par classes d'après des caractéristiques communes, comme par exemple les voyelles et les consonnes.
Nous pouvons ainsi pour la langue française déterminer :
• les voyelles : a/e/i/o/u, etc., qui sont toujours voisées
• les consonnes plosives, caracterisées par un bruit bref, dit d'« explosion ». Elles peuvent être sourdes, c'est-à-dire non voisées : [p],[t],[k] ou sonores, c'est-à-dire voisées : [b],[d],[g].
• les consonnes chuintantes, qui sont caracterisées par l'émission d'un bruit continu. Elles sont

sourdes : [f],[s], ou voisées : [v],[z].
• les liquides : [l], vibrantes [r] et les nasales : [m],[n] sont également voisées.

Les phonèmes sont caracterisés en groupe ou individuellement par des caractéristiques temporelle et fréquentielle qui touchent exactement à notre sujet. Des études ont par exemple montre la corrélation des voyelles et de la position relatives des deux premiers formants [Fan70], ce que l'on a coutume de nommer le « triangle vocalique » : ainsi, chaque voyelle se caractérise par une enveloppe caractéristique. Types de phonèmes et enveloppe spectrale sont donc corrélés : nous devrons prendre en compte cette corrélation, et traiter individuellement chaque phonème lors de la transposition.

1.3.2 La dynamique de la production vocale

Cependant le phonème n'est qu'une abstraction de la production vocale. Celle-ci est un phénomène dynamique qui se caractérise par une succession continue de phonèmes. La production du phonème dépend de son contexte segmental. Pour chaque phonème, les phonèmes le précédent et le suivant directement jouent un rôle essentiel dans la production de ce phonème, et influent sur lui. Les variations formantiques de la voyelle qui suivent et qui précèdent l'articulation d'une consonne sont caractéristiques de l'association de cette consonne et de cette voyelle.

Les interactions dynamiques qui se manifestent durant les transitions entre les phonémes, et qui relèvent du mécanisme de co-articulation, agit à un niveau local de la production vocale, dit segmental. L'évolution à un niveau temporel plus grand - dit supra-segmental - de la production vocale est la prosodie. La prosodie se caractérise par quatre paramètres majeurs : l'intonation, c'est-à-dire la «courbe » de la fréquence fondamentale, l'énérgie, le débit et la « qualité vocale ». La qualité vocale est ce que l'on nomme communément le timbre, et qui dépend notamment de l'enveloppe spectrale.

Nous présentons sur la figure [1.6] les multiples paramétres de la production de la parole qui viennent influencer le comportement de l'enveloppe spectrale.

1.3.3 Le noyau vocalique

Le noyau vocalique d'un phonème est par définition la partie des segments sur lesquels le phonème se réalise, c'est-à-dire que l'on se trouve sur une partie stable du phonèmes, laquelle ne dépend alors pas du contexte phonétique. Cette zone de stabilité se caractérise par trois critères majeurs : relatives stabilités de la fréquence fondamentale et de l'enveloppe spectrale - essentiellement de la position des formants - et présence d'un maximum d'énergie. Nous avons utilisé l'algorithme de détection des noyaux vocaliques par la mesure de l'« *l'Acoustic Center of reliability* » qui a été proposé par P. Mokhtari [Mok02], et que nous présentons dans la figure [1.5].

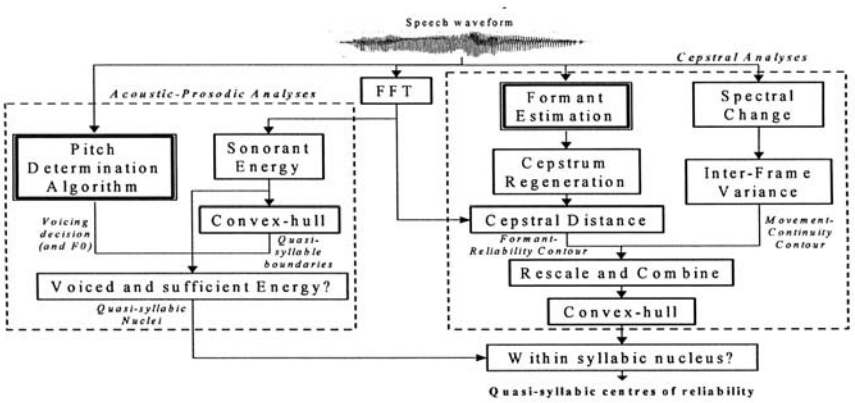


FIG. *1.5:* Algorithme d'estimation de l'« Acoustic center of reliability », d'après [Mok02]

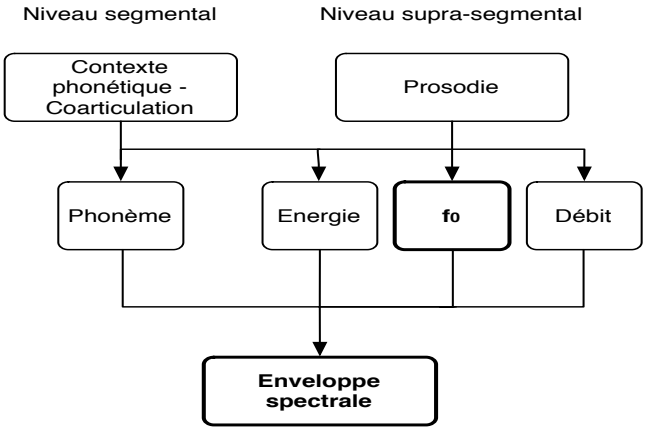


FIG. *1.6:* Représentation des paramètres de la production vocale et de leur influence sur l'enveloppe spectrale

Chapitre 2

Modèles et estimation de l'enveloppe spectrale

2.1 Définition

Il n'y a pas à ce jour de consensus sur une définition de l'enveloppe spectrale. Nous utiliserons la défintion proposée par A. Roebel et F. Villavicencio [Vil06], [Roe05a] : l'enveloppe spectrale est une fonction régulière - *smooth* - qui passe par les sommets des partiels et suffisament lisse pour ne pas modéliser les partiels. L'hypothèse de régularité provient du fait que l'enveloppe doit être une fonction continue et de dérivée continue - ce qui n'est par exemple pas le cas d'une interpolation linéaire entre les sommets des partiels. D'autres modèles de l'enveloppe sont utilisés, que nous allons maintenant présenter dans leurs grandes lignes. Chaque modèle suppose implicitement une certaine définition de l'enveloppe spectrale.

2.2 Une enveloppe spectrale et des représentations

Dans le modèle source-filtre, l'enveloppe spectrale correspond, à un facteur près, au module de la réponse en fréquence du filtre. La *modélisation* auto-régressive, par exemple le modèle du codage par prédiction linéaire - *Linear Predictive Coding*, est fondée sur l'hypothèse que ce filtre ne contient que des pôles - *all-pole model*. La *modélisation* par les coefficients cepstraux est elle fondée sur l'hypothèse d'un filtre à réponse impulsionnelle finie de type « cepstre ». Pour chacun de ces *modèles*, plusieurs méthodes d'estimation des paramètres ont été développées : ainsi les coefficients auto-régressifs peuvent être estimés par la méthode dite d'« auto-correlation » [], par la méthode du filtre « tout-pôles » discrets [ElJ91]. Et l'estimation des coefficients cepstraux par les méthodes du cepstre [Opp68], cepstre discret [Gal90], cepstre discret avec régularisation [Cap96] et par la méthode dite de la *True Envelope* [Ima79] [Roe05a].

2.3 Modélisation auto-régressive

On modélise le signal s par un estimateur $\hat{s}$ suivant un modèle auto-régressif d'ordre p, de la manière suivante :

$$\hat{s}(n)=\sum_{k=1}^{p} a_k s(n-k) \tag{2.1}$$

Celà revient à faire l'hypothèse que dans le modèle source-filtre, le filtre H est un filtre ne contenant que des pôles. Ce filtre s'écrit, en transformée en z :

$$H(z)=\frac{1}{\sum_{i=k}^{p} a_k z^{-k}} \tag{2.2}$$

Les paramètres de ce modèle sont les coefficients $\{a_1, ..., a_p\}$ ainsi que l'énergie du résiduel, définie comme l'erreur quadratique entre le signal et son estimation :

$$\sigma = (s(n) - \hat{s}(n))^2$$

2.3.1 Estimation par la méthode d' « autocorrelation »

C'est l'une des méthodes les plus connues et les plus courantes pour estimer les paramètres $\{\{a_k\}_{k=\{1,...,p\}}, \sigma\}$. Pour celà, on détermine la matrice d'auto-covariance, qui amène directement à l'équation dite de Yule-Walker :

$$\gamma_{XX}(i)=\begin{cases}\sum_{k=1}^{p} a_k \gamma_{XX}(k) + \sigma_B, & \text{si i} = 0\\ \sum_{k=1}^{p} a_k \gamma_{XX}(k-i) & \text{sinon.}\end{cases} \tag{2.3}$$

Soit en notation matricielle :

$$\begin{bmatrix}\gamma_{XX}(1)\\ ...\\ ...\\ \gamma_{XX}(p)\end{bmatrix}=\begin{bmatrix}\gamma_{XX}(0) & ... & ... & \gamma_{XX}(p-1)\\ ... & ... & ... & ...\\ ... & ... & ... & ...\\ \gamma_{XX}(p-1) & ... & ... & \gamma_{XX}(0)\end{bmatrix}\begin{bmatrix}a(1)\\ ...\\ ...\\ a(p)\end{bmatrix} \tag{2.4}$$

Les coefficients $\{a_1, ..., a_p\}$ sont alors obtenus par estimation de l'auto-covariance et inversion de l'équation (2.4) .

L'estimation des paramètres du modèle par la méthode d'auto-correlation est la meilleure estimation linéaire, pour un ordre donné, et dans le cas d'un signal fenêtré par une fenêtre rectangulaire, au sens de l'erreur quadratique moyenne entre le signal et sa prédiction [Kay]. L'inconvénient de cette méthode est qu'elle ne permet pas une estimation de l'enveloppe spectrale d'un signal harmonique au sens où nous l'entendons, c'est-à-dire qu'elle ne passe pas bien par les sommets des partiels. Ceci s'explique par le fait que la mesure de l'erreur comme moyenne de l'erreur quadratique n'est pas adaptée au modèle harmonique, pour lequel seul un certain nombre de points fréquentiels comptent réellement, à savoir le sommet des partiels.

La méthode DAP [ElJ91] a été proposée afin de pallier à cet inconvénient. Son principal avantage réside dans la prise en compte du modèle harmonique du signal. Cette méthode consiste à déterminer itérativement les paramètres du modèle qui minimisent la distance d'Itakura-Saito. Cependant, les résultats sont très dépendant de l'estimation préalable des paramètres des partiels.

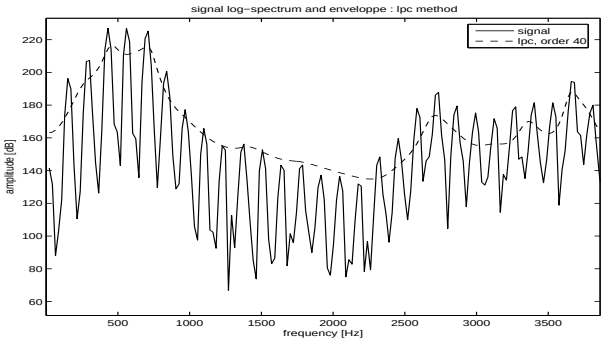


FIG. *2.1:* Estimation de l'enveloppe spectrale par analyse LPC d'ordre 40 pour un signal echantillonné à 44.1kHz

2.3.2 Une autre représentation des coefficients auto-régressifs : Linear Spectral Frequencies

D'autres représentations des coefficients auto-régressifs ont été utilisées. Les LSF, introduites par Itakura [Ita75], constituent une représentation différente des coefficients auto-régressifs.

En reprenant l'équation auto-régressive d'ordre M, soit :

$$\frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{M} a_k z^{-k}} \tag{2.5}$$

On définit les polynômes P et Q de la manière suivante :

$$\begin{cases} P(z) = A(z) + z^{-(M+1)} A(z^{-1}) \\ Q(z) = A(z) - z^{-(M+1)} A(z^{-1}) \end{cases} \tag{2.6}$$

On montre que les racines $z_{P,j}$ et $z_{Q,j}$, respectivement des polynômes P et Q se trouvent sur le cercle unité. Ceci étant, on peut les écrire :

$$\begin{cases} z_{P,j} = e^{j\alpha_{P,j}} \\ z_{Q,j} = e^{j\beta_{Q,j}} \end{cases} \tag{2.7}$$

Les angles $\alpha_{P,j}$ et $\beta_{Q,j}$ sont appelés les *Linear Spectral Frequencies*

Les lignes spectrales présentent l'avantage de donner une représentation locale en fréquence du spectre, car elles modélisent par paires les maximum d'énergie du spectre en fréquences et en amplitudes - notamment les formants. Elle possédent également de bonne propriétés d'interpolation qui nous serons utilies dans la suite, et que nous évoquerons au chapitre 6. Cependant, les lignes spectrales ont l'incovénient de parfois modéliser des résonnances trop fortes, lorsque deux lignes sont très proches l'une de l'autre.

2.4 Les coefficients cepstraux

La représentation de l'envelope spectrale par des coefficients cepstraux est un modèle de représentation dans le domaine fréquentiel. Elle repose sur une modélisation du filtre du modèle source-filtre comme d'un filtre à réponse impulsionnelle finie de type « cepstre ».

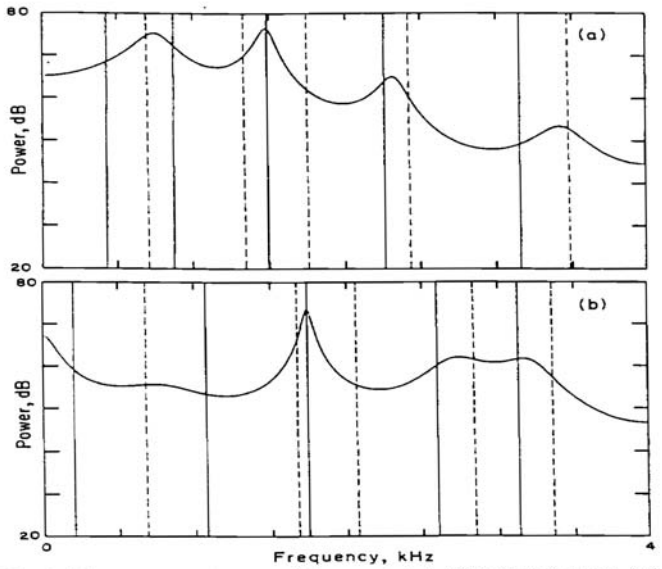


Fig. 1. LP power spectrum and the associated LSFs for (a) vowel /a/ and (b) fricative /s/.

FIG. *2.2:* Lignes fréquentielles

2.4.1 La méthode du « cepstre » :

On considère le modèle source filtre tel que presenté dans la partie précédente :

$$x(t)=(s*e)(t) \tag{2.8}$$

Soit dans le domaine fréquentiel :

$$X(f) = S(f).E(f) \tag{2.9}$$

En ne considérant que le module de la transformée de Fourier, et en passant le logarithme :

$$log(|X(f)|) = log(|S(f)|) + log(|E(f)|) \tag{2.10}$$

En prenant la représentation en logarithme, nous sommes passés d'une multiplication des composantes source et filtre à une simple addition ; ainsi, en supposant que les deux composantes interviennent à des fréquences différentes, nous pouvons donc simplement séparer les deux contributions.

Par transformée en cosinus, on obtient :

$$c(n)=DCT(log(|X(f)|)) \tag{2.11}$$

où DCT est la transformée en cosinus En utilisant les propriétés de parité et de périodicité de $log(|X(f)|)$, on en déduit :

$$log(|X(f)|) \quad = \quad c(0) + 2\sum_{k=1}^{K} c(k)cos(2\pi fk) \tag{2.12}$$

où K est le nombre de points de la DCT.

Le k-ième coefficients cepstral représente donc la constribution de la cosinusoïde de fréquence $2\pi fk$ au spectre d'amplitude logarithmique.

En revenenant sur la séparation des contributions de la source et du filtre, nous pouvons trouver l'ordre p optimal - c'est-à-dire l'ordre du cepstre maximal permettant de ne pas modéliser les partiels - de l'estimation de l'enveloppe spectrale.

Cet ordre est donné par la relation :

$$p \leqslant \frac{f_e}{2f_0} \tag{2.13}$$

où $\begin{cases} f_e \text{ est la fréquence d'échantillonage du signal} \\ f_0 \text{ est la fréquence fondamentale du signal considéré comme harmonique ou quasi-harmonique} \end{cases}$

Les principaux inconvénients de cette méthode d'estimation des coefficients cepstraux sont qu'elle modélise une enveloppe qui ne passe pas par les sommets des partiels du spectre, qu'elle a d'ailleurs tendance à sous-estimer.

2.4.2 Échelle logarithmique

La transformation de l'échelle des fréquences sur une échelle logarithmique et la décomposition du signal sur un banc de filtres modélise le fonctionnement de l'audition humaine au niveau de la cochlée dans le système auditif périphérique. Celà revient à faire une intégration de l'énergie contenue dans des bandes de fréquences [Rab93].

La fonction de transformation des fréquences de l'échelle linéaire à une échelle de Mel est définie de la manière suivante :

$$f_{mel} : f \mapsto 2595 log(1 + \frac{f}{700})$$

Le signal est alors intégré fréquentiellement sur un banc de M filtres régulierement espacés en échelle Mel. Les coefficients cepstraux sont alors estimés sur les points résultants. La représentation du spectre par les MFCC est très utiles et très utilisés dans les problèmes de reconnaissance et de classification, car ils présentent le double avantage de réduire le nombre de coefficients utilisés, et de mettre en avant des caractéristiques saillantes, c'est-à-dire perceptives, de l'information contenue dans le spectre.

2.4.3 La méthode de la « *True envelope* »

Le problème principal de l'estimation des coefficients cepstraux est le même que celui de l'estimation des coefficients auto-régressifs : seul un nombre réduit de points - les sommets des partiels - est réellement intéressant, et minimiser l'erreur quadratique sur l'ensemble du spectre ne donnera en général pas la meilleure estimation sur ces points. La méthode du cepstre dicret [Gal] prend en compte le modèle harmonique et détermine les paramètres en minimisant l'erreur quadratique entre

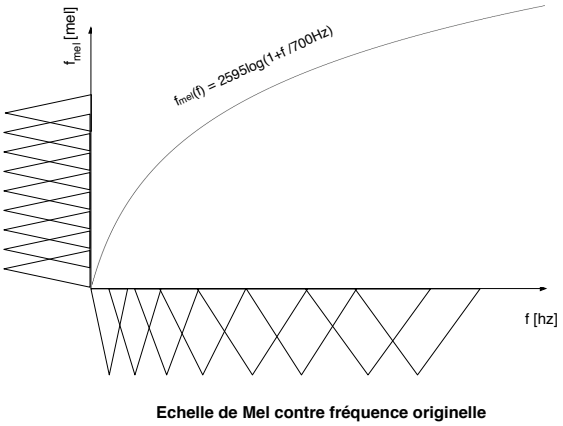


FIG. *2.3:* Relation entre fréquences en échelle linéaire et fréquences en échelle Mel

l'amplitude de l'enveloppe à un certain nombre de points fréquentiels $\{f_1, ..., f_n\}$ et l'amplitude du spectre à ces mêmes points. Enfin la méthode du cepstre dicret régularisé [Cap] ajoute dans la mesure de l'erreur une fonction favorisant une enveloppe lisse. Le principal inconvénient de ces méthodes est que les résultats sont très dépendants de la qualité de l'estimation en fréquence, amplitude et nombre des partiels.

Une autre méthode d'estimation des coefficients cepstraux, dite *true enveloppe*, a été proposée par S. Imai et Y. Abe [Ima79] pour résoudre ce problème sans nécessiter d'estimation préalable des paramètres des partiels. Cette méthode a récemment été reprise et optimisée par A. Roebel [Roe05a], [Roe05b].

La True Envelope est une méthode d'estimation itérative des coefficients cepstraux, qui permet de prendre en compte implicitement l'harmonicité du signal, c'est-à-dire qu'elle ne nécessite pas l'estimation préalable des paramétres précis des partiels.

Soit $X(k)$ le spectre discret de K points fréquentiels d'une trame temporelle, et soit $V_i(k)$ la représentation spectrale donnée par les coefficients cepstraux à l'itération i, c'est-à-dire la transformée de Fourier des p premiers coefficients cepstraux :

$$V_i(k) = c(0) + \sum_{k=1}^{p} c(k) cos(2\pi f k)$$

La méthode d'estimation fonctionne itérativement de la manière suivante :

0. On pose $A_0(k)$=$log(|X(k)|)$ et $V_0(k) = -\infty \quad , \quad \forall k \in [1, ..., K]$

1. L'amplitude du spectre « cible », à l'itération i, est :

$$A_i(k)=max(A_{i-1}(k), V_{i-1}(k)) \quad , \quad \forall k \in [1, ..., K]$$

2. Les coefficients cepstraux du spectre $A_i(k)$, et par conséquence la nouvelle représentation spectrale de l'enveloppe $V_i(k)$ est calculée.

Les étapes 1 et 2 sont répétées jusqu'à ce que qu'un critère d'arrêt soit vérifié. Le critère d'arrêt est par exemple la distance θ entre l'enveloppe estimée V_i et les points fréquentiels du spectre dont

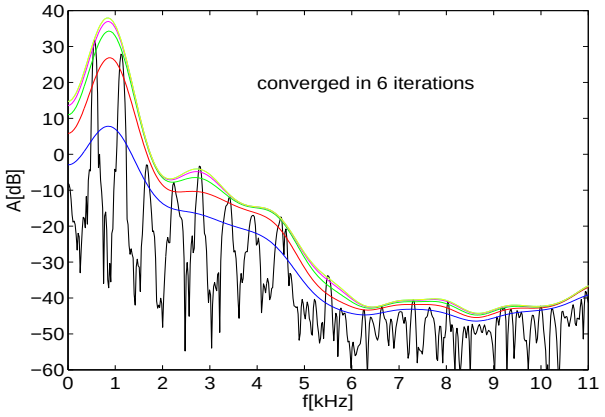


FIG. *2.4:* Estimation de l'enveloppe par la méthode *True Enveloppe* pour 6 itérations, d'après [Roe05al]

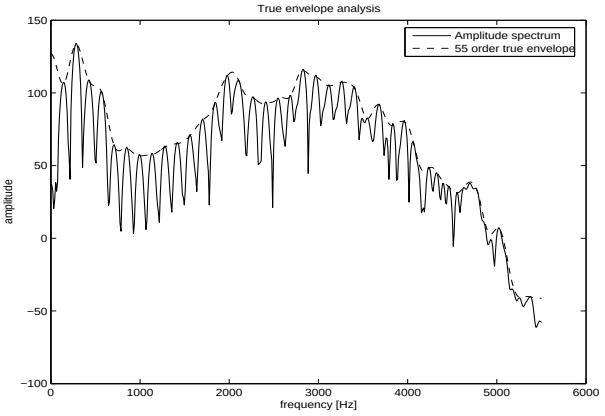


FIG. *2.5:* Estimation de l'enveloppe spectrale par analyse *True Enveloppe* d'ordre 55

l'amplitude est supérieure à cette enveloppe :

$$A_i(k) - V_i(k) \leqslant \theta \quad , \quad \forall k \in [1, ..., K] \tag{2.14}$$

Une exemple typique de valeur de θ est celle qui correspond à 2 dB.

Dans la suite de notre travail, nous avons choisi d'utiliser l'estimation de l'enveloppe spectrale par la méthode *true envelope*, car elle est celle qui donne l'estimation de l'enveloppe spectrale qui correspond le mieux à notre définition. Elle passe par les sommets des partiels, sans pour autant nécessiter l'estimation des paramètres de ces partiels. Nous utiliserons également la représentation en LSF dont l'estimation sera faite à partir des coefficients cepstraux estimés par la méthode true envelope, suivant la méthode proposée dans [Vil06].

Chapitre 3

Les méthodes de transposition actuelles

La transposition est une transformation homothétique des fréquences des partiels dans le domaine fréquentiel. Actuellement, cette transformation est realisée par les méthodes d'analyse, de modification, et de synthèse du signal comme le *vocodeur de phase* et les méthodes fondées sur l'addition et la superposition de formes d'ondes élémentaires, comme par exemple la méthode dite *Pitch Synchronized Overlap-Add* ou *PSOLA* [Pee01]. Nous présentons dans ce chapitre ces deux méthodes et mettons en évidence leurs limites.

3.1 Le vocodeur de phase

3.1.1 Le vocodeur de phase « classique »

Initialement développée au sein des laboratoires Bells par J. Flannagan [Fla66], la méthode du vocodeur de phase a connu depuis de nombreux développements à des fins musicales. Nous donnerons comme exemples D. Moore à San Diego qui a créé le logiciel *p-voc* dans le cadre d'un sysème de musique assistée par ordinateur ; M. Dolson [Dol86] pour le logiciel *Carl*, et qui vint collaborer à l'IRCAM pour développer un système d'analyse-synthèse basé sur le vocodeur de phase. Travail repris et prolongé à l'IRCAM par P. Depalle [Dep], pour finalement aboutir au logiciel *SuperVP*, aujourd'hui développé par A. Roebel [Roe05a], [Roe05b]. De nombreux autres développements de cette méthode ont été réalisés à travers le monde, avec notamment, au MIT, B. Vercoe pour le logiciel *C-SOUND*, J. Laroche [Lar99], et d'autres encore.

Le vocodeur de phase repose sur la décomposition temporelle du signal en trames successives, elles-mêmes décomposées fréquentiellement sur un banc de filtres. Le signal est donc d'abord décomposé en trames temporelles, dont on calcule pour chacune la transformée de Fourier. L'ensemble est réalisé au moyen de la transformée de Fourier à court-terme ou TFCT.

La TFCT à l'instant discret n, et à la fréquence discrète k d'un signal x, s'écrit :

$$X_k(n) = \sum_{r=-\infty}^{\infty} x(r)w(n-r)e^{-j\frac{2\pi rk}{N}} \tag{3.1}$$

Pour opérer une transposition, la méthode du vocodeur de phase consiste en un double rééchantillonage du signal en deux étapes :

- un changement de durée avec conservation de la hauteur et synchronisation des phases pour éviter des phénomènes de distorsion de phase. L'hypothèse de l'unicité de la composante sinusoïdale dans chaque banc de filtre est fondamentale dans cette étape, car c'est à cette condition que l'on peut rigoureusement synchroniser les phases.

Pour modifier le signal sur l'échelle temporelle, il faut idéelement décrire sa TFCT à une vitesse différente :

$$X(rI,k) \mapsto X(rI^{'},k) \tag{3.2}$$

La fréquence instantannée $\mathcal{F}(n,k)$ à l'instant discret n et dans la bande de fréquence k s'écrit :

$$\mathcal{F}(n,k) \widehat{=} \frac{1}{2\pi}\frac{\partial \Phi}{\partial t} \tag{3.3}$$

En supposant que chacun des filtres ne contient qu'une seule composante sinusoïdale, on montre que la fréquence contenue dans chacune des bandes de fréquences k s'écrit :

$$f_0 = \mathcal{F}(n,k) + \frac{k}{N} \tag{3.4}$$

En faisant l'hypothèse de l'unicité de la composante sinusoïdale dans chaque bande de fréquence, la relation précédente nous donne :

$$\begin{cases} \Phi(rI,k) = (f_0 - \frac{k}{M})rI \\ \Phi(r^{'}I,k) = (f_0 - \frac{k}{M})rI^{'} \end{cases} \tag{3.5}$$

Pour que les contributions demeurent en phases, il faut donc ajouter au signal obtenu la phase :

$$\phi = 2\pi(f_0 - \frac{k}{M})(I^{'} - I) \tag{3.6}$$

La modification finalement operée est :

$$X(rI,k) \mapsto X(rI^{'},k)exp(-2\pi(f_0 - \frac{k}{M})(I^{'} - I)) \tag{3.7}$$

- une modification de la hauteur par rééchantillonage temporel, c'est-à-dire une dilatation / compression fréquentielle.

A travers ce modèle de transposition, il apparaît que l'enveloppe spectrale de la trame se trouve dilatée ou compressée en même temps que les fréquences, car la transposition ne modifie pas l'amplitude des partiels. Une pareille modification s'accompagne d'une modification sensible du timbre du son.[1].

3.1.2 Conservation de l'enveloppe spectrale

J. Dolson a proposé une méthode qui permet la préservation de l'enveloppe spectrale au cours de la transposition [Dol86]. Elle est réalisée au moyen d'une dilatation ou compression de l'enveloppe dans le domaine fréquentiel avant la transposition par le vocodeur de phase.

1. Nous savons en effet que le « timbre » - notion au demeurant éminemment complexe et aujourd'hui encore incomplètement définie - dépend entre autres de la position fréquentielle des partiels ainsi que de leurs amplitudes respectives et de leur rapport de fréquence et d'amplitude. De petites variations de ces paramètres entraînent une modification sensible du timbre.

Pour un facteur de transposition f donné, la nouvelle amplitude P au point fréquentiel k est déterminée à partir de l'amplitude initiale A au même point de la manière suivante :

$$P(k)=\frac{A(k.f)}{A(k)} \tag{3.8}$$

Cette méthode nécessite d'estimer l'enveloppe spectrale, et ensuite de modifier l'amplitude des partiels par la formule (3.8).

Cette dilatation ou compression consiste, comme le montre la formule précédente, à normaliser dans un premier temps le spectre et à lui appliquer dans un second temps l'amplitude correspondante de l'enveloppe à la fréquence transposée. Ainsi la transposition revient à une translation des partiels à l'intérieur de l'enveloppe spectrale. Cependant cette réaffectation des amplitudes pose des problèmes lorsque l'on transpose vers le bas : en effet, l'enveloppe n'est pas bien définie pour les fréquences qui se trouvent en-dessous du premier partiel, et l'amplitude de celui-ci peut alors être modifiée de manière non désirable. Pour remédier à celà, A. Roebel [Roe05a] a proposé une nouvelle fonction d'affectation des amplitudes :

$$P(k)=\frac{A(k(D(k)+(1-D(k))f))}{A(k)}, \tag{3.9}$$

avec

$$D(k)=\frac{1}{1+exp(\frac{k-k_1}{T_k})}$$

où $\begin{cases} k_1 \text{ est le point fréquentiel correspondant à la fréquence du premier partiel} \\ T_k \text{ représente la largeur de bande de la zone de transition entre la conservation} \\ \text{de l'enveloppe originelle A et l'enveloppe transformée P.} \end{cases}$

Cette nouvelle fonction permet de traiter l'enveloppe en deux parties fréquentiellement distinctes : une partie basse, autour de la fréquence du premier partiel, dans laquelle les partiels transposés conservent leur amplitude initiale - comme dans le vocodeur de phase « classique » -, et une partie haute dans laquelle l'enveloppe est conservée. Ainsi le probléme soulevé est résolu. Par ailleurs, il doit être noté que lorsqu'un formant se trouve dans une région fréquentielle suffisamment proche de celle du premier partiel, il est déplacé en même temps que celle-ci, conformément aux observations faites sur la corrélation de la position des formants en fonction de la fréquence fondamentale.

3.2 Méthode par superposition et addition de formes d'ondes élémentaires - *Pitch synchronized Overlap-Add* - PSOLA

De nombreuses méthodes de modification du signal reposant sur le principe d'addition et de superposition de formes d'ondes temporelle ont été développées ces 30 dernières années. Nous citerons la méthode SOLA (Synchronized Overlap-Add), WSOLA (Waveform Similarity Overlap-Add) [Ver] et PSOLA (Pitch Synchronized Overl-Add) [Pee01]. La méthode de superposition et d'addition synchrone à la période fondamentale est fondée sur une décomposition du signal en une série de formes d'ondes élémentaires. Ces formes d'ondes élémentaires sont obtenues par un fenêtrage exactement centré sur les périodes fondamentales du signal. La synchronisation à la période fondamentale est obtenue à partir de marques temporelles basées sur des caractéristiques locales du signal. Le signal de synthèse est alors reconstitué par superposition et addition de ces formes d'ondes élémentaires, après d'éventuelles modifications de celles-ci. Dans le cas d'un signal stationnaire et périodique et

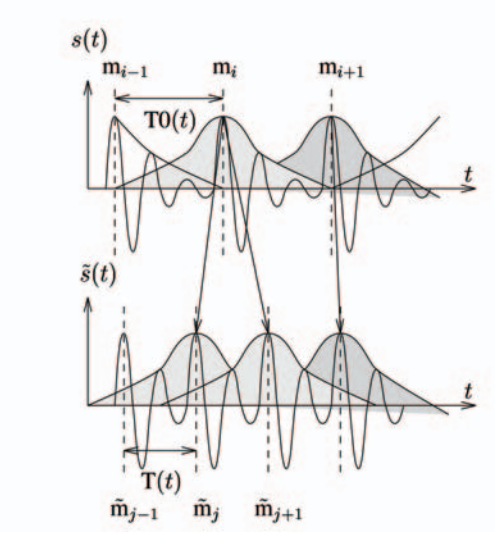


FIG. *3.1:* Transposition par la méthode *PSOLA*, d'après [Pee01]

en utilisant la méthode PSOLA à large bande, *PSOLA-WB*, - c'est-à-dire en prenant une forme d'onde élémentaire constituée de deux périodes du signal - les composantes sinusoïdales du signal ne sont pas résolues fréquentiellement, et donc le spectre de la forme d'onde élémentaire correspond à une approximation de l'enveloppe spectrale.

En admettant par ailleurs la modélisation source-filtre du signal, c'est-à-dire la convolution d'un train d'impulsions à la période fondamentale T_0 avec la réponse impulsionelle d'un filtre :

$$s(t)=\sum_{k=1}^{\infty}(\delta_{(k,T_0)} \otimes h)(t) \tag{3.10}$$

avec

$$\delta_{(k,T_0)}(t) = \delta(t - kT_0)$$

Soit dans le domaine fréquentiel :

$$S(\omega)=\sum_{k=1}^{\infty}\delta(\omega - k\omega_0).H(\omega) \tag{3.11}$$

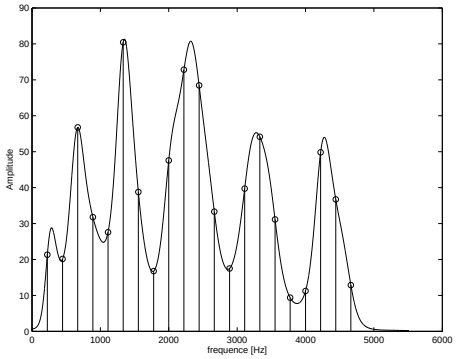
On obtient par transformée inverse :

$$x(t)=\sum_{k=1}^{\infty}\frac{1}{T_0}X_{T_0}(k\omega_0)exp(jk\omega_0 t) \tag{3.12}$$

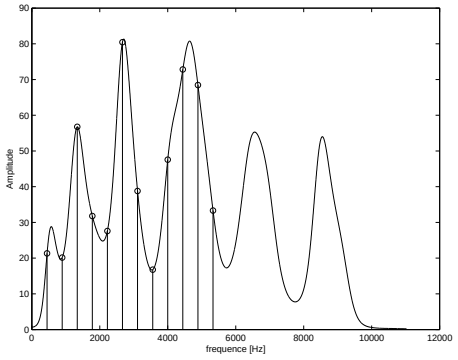
Dans la méthode PSOLA-WB, $S(\omega)$ est une enveloppe spectrale du signal. La transposition du signal consiste simplement en un ré-échantillonage du spectre de la forme d'onde élémentaire X_{T_0} à de nouvelles fréquences $k.f_0$, où k est le facteur de transposition. La méthode PSOLA-WB est donc implicitement une méthode de transposition avec conservation de l'enveloppe spectrale.

3.3 Limites des méthodes de transposition

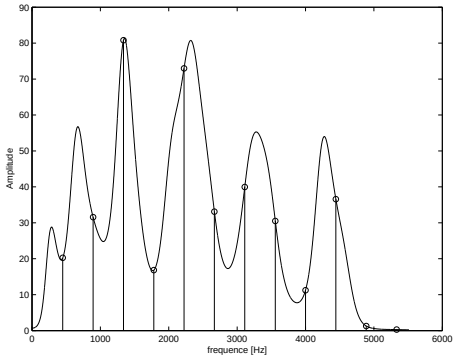
La transposition de parole réalisée par les méthodes présentées n'est pas sans défaut : tant que la transposition reste dans des proportions raisonnables, c'est-à-dire inférieure à une octave environ, le résultat est vaguement correct. Mais dépassé cette limite, le résultat ne semble plus du tout naturel : la voix du locuteur perd peu à peu de son « naturel », voir de son « identité ».



(a)



(b)



(c)

FIG. *3.5:* Modèles de transposition : spectre initial (a), transposition avec dilatation de l'enveloppe comme dans le vocodeur de phase « classique » (b), avec conservation de l'enveloppe comme dans SuperVP et PSOLA-WB (c)

Chapitre 4

Rôle de l'enveloppe spectrale pour la transposition

Nous avons vu dans les méthodes de transposition que la modification de la fréquence fondamentale est sans modification de l'enveloppe spectrale, en cohérence avec le modèle source-filtre, dans lequel le filtre est supposé invariant en fréquence. Cependant, ce modèle est une approximation du comportement réel de l'enveloppe spectrale. La mise en évidence des limites de ces méthodes laisse supposer qu'il n'est plus valable pour des modifications importantes de la fréquence fondamentale. L'examen de la position du premier formant en fonction de la fréquence fondamentale [Syr85] montre que filtre varie en fonction de la fréquence fondamentale. Ce chapitre est consacré à la mise en place d'une expérience visant à mettre en évidence cette assertion. La question s'énonce de la manière suivante : si nous sommes capables de savoir quelle transformation appliquer à l'enveloppe spectrale, celà améliore-t-il le résultat de la transposition ?

4.1 Constitution d'un corpus

Nous avons constitué un corpus reposant sur deux idées principales : constituer une base de test pour juger de l'influence de l'enveloppe spectrale sur la transposition et une base spécifiquement adaptée à l'étude de la modification de l'enveloppe spectrale lors d'un changement de hauteur. Le corpus a été constitué de la manière suivante : un locuteur unique, de sexe masculin, dont la tessiture se trouve approximativement entre 80Hz et 220Hz. L'enregistrement a été réalisé dans une chambre anéchoïque. La diction sera qualifiée de « neutre » et de « naturelle », par opposition aux travaux menés sur l'expressivité de la voix[1].
Pour le test, nous avons choisi une phrase répétée 9 fois, dont le locuteur a varié chaque fois l'intonation.
Ce premier corpus constitue un sous ensemble d'un corpus plus large concu spécifiquement pour l'étude de la corrélation de l'enveloppe spectrale et de la fréquence fondamentale. Dans le premier chapitre, nous avons montré que l'enveloppe spectrale dependait d'une multitude de facteurs qui rendent impossible une étude exhaustive de leurs influences respectives. Pour montrer l'influence

1. Les notions de « neutralité » et de « naturel » peuvent paraître obscures. Pour fixer les choses, nous dirons que nous entendons par neutre une diction qui n'est pas expressive, c'est-à-dire ne véhiculant pas d'expression particulière. Et par « naturelle » une diction qui n'est pas forcée, c'est-à-dire avec un débit standard et une intonation qui reste dans le registre normal du locuteur, c'est-à-dire que le locuteur n'est pas amené à forcer sa voix.

de la seule fréquence fondamentale sur l'enveloppe spectrale, il faut tenter de l'isoler, en « neutralisant » au mieux les autres paramètres pouvant entrer en jeu. Nous avons construit un ensemble de 4 phrases contenant l'ensemble des phonèmes voisés du français. Ces phrases ont été choisies courtes pour permettrent au locuteur de contrôler facilement son intonation d'une phrase à l'autre. A partir de ces quatres phrases originelles, nous avons déduits par permutation des mots à l'intérieur de celles-ci et pour chacune d'elles, un ensemble de quatre sous-phrases en faisant varier la place des phonèmes à l'intérieur de la phrase. Ceci est fait afin d'essayer de neutraliser l'influence supra-segmentale de la position du phonème ainsi que celui du contexte phonétique. Enfin, nous avons défini un ensemble de 9 courbes d'intonations qui permettent de varier les hauteurs pour chacun des phonèmes. Nous avons donc finalement un corpus de 16 phrases, répétées 9 fois chacune par le locuteur, soit un corpus final de 144 phrases, chacune d'une durée approximative de 5 secondes.

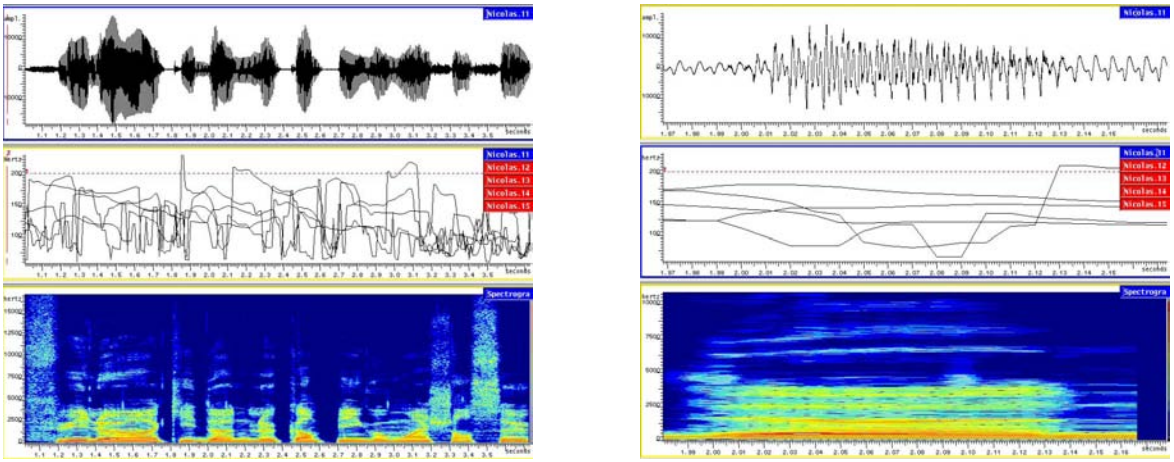


FIG. *4.1:* Constitution d'un corpus spécifique : courbes d'intonations pour une même phrase (a) et fréquences fondamentales sur un même phonème (b)

4.2 Validation de la démarche : importance de l'enveloppe spectrale pour la transposition ?

La première idée était de valider au préalable notre démarche scientifique en constatant si la modification de l'enveloppe spectrale améliorait la qualité de la transposition. Notre démarche est la suivante : transposer une phrase du corpus sur la courbe d'intonation d'une autre version de cette phrase, d'abord sans modification de l'enveloppe, puis avec application de l'enveloppe de la phrase cible. La transposition est réalisée par les méthodes présentées dans le chapitre précédent.

On considère deux phrases, l'une dite source qui est la phrase que l'on veut transposer, et une autre phrase, dite cible, vers laquelle on va transposer. On associe à ces deux phrases respectivement une séquence temporelle source formée du doublet (f_0, env), et une séquence temporelle cible formée du doublet $(f_0^{'}, env^{'})$, où f_0 est la fréquence fondamentale et env une représentation de l'enveloppe spectrale.

Notre démarche consiste à :

(1) Transposer la phrase source vers la phrase cible en conservant l'enveloppe de la phrase source.
(2) Transposer la phrase source vers la phrase cible en appliquant l'enveloppe de la phrase cible sur la transposée de la phrase source.
(3) Comparer les versions (1), (2) ainsi que l'original cible.

L'expérience est réalisée de la manière suivante :

1. Alignement des deux phrases par la méthode classique de l'affectation dynamique des temps - *Dynamic Time Warping* - utilisant la distance euclidienne des 13 premiers coefficients MFCC.
2. Estimation de la fréquence fondamentale et de l'aperiodicité des phrases source et cible par l'algorihtme YIN [Che02].
3, Calcul du facteur de transposition lissé et corrigé par application d'un filtre médian d'ordre 15 sur un signal échantillonné à 44100 Hz. Les parties non-voisées ne sont pas transposées.
4. L'estimation de l'enveloppe spectrale est réalisée par la méthode *True Envelope* avec l'ordre optimal déduit de la fréquence fondamentale.

Le résultat est ensuite comparé aux différentes méthodes de transposition actuelles . Les méthodes utilisées sont :
- le vocodeur de phase « classique »
- le vocodeur de phase avec conservation de l'enveloppe source
- le vocodeur de phase utilisant la réaffectation des amplitudes en deux parties suivant la relation (3.9) du chapitre précédent
- le vocodeur de phase avec application de l'enveloppe de la cible
- PSOLA-WB

Les résultats obtenus, disponibles sur http ://recherche.ircam.fr/equipes/analyse-synthese/vivos/, nous permettent d'aboutir à la conclusion suivante : l'application de l'enveloppe cible améliore sensiblement la qualité de la transposition, et ceci d'autant plus que le facteur de transposition est élevé. La séparation du spectre en deux parties, la basse transposée, la haute conservée, montre de bons résultats. Celà s'explique par le fait que cette réaffectation des amplitudes est supposée rendre compte, de manière évidemment limitée, de la modification de l'enveloppe en fonction de la fréquence fondamentale. Cette simple expérience montre l'importance de l'enveloppe spectrale lors de la transposition. Notre tâche va maintenant consister à tenter de déterminer de manière optimale l'enveloppe spectrale qui sera appliquée lors de la transposition.

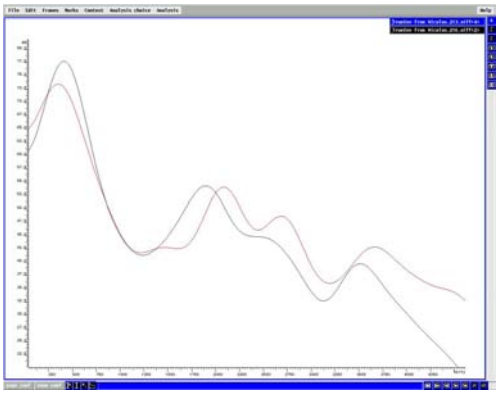


FIG. *4.2:* Exemple de différence d'enveloppe spectrale lors de la prononciation d'un même phonème, dans le même contexte phonétique, mais à des hauteurs différentes. Phonème /on/ dans la phrase : "Ils n'ont pas l'air de croire à leur bonheur"

Chapitre 5

Classification et apprentissage

5.1 Introduction

De récents travaux ont mis en évidence la corrélation entre la fréquence fondamentale et l'enveloppe spectrale, que ce soit dans le cadre de la prédiction de la fréquence fondamentale à partir de l'enveloppe spectrale [EnN03], [EnN05], [Sha], [Mil] ou pour la conversion de voix [Sty96], [Sty98], [Kai00], [Kai01]. Notre démarche consiste à tirer partie de cette corrélation pour permettre la prédiction de l'enveloppe spectrale en fonction de la fréquence fondamentale c'est-à-dire déterminer une fonction : $f : (f_0; env) \mapsto (f_0^{'}; env^{'})$. Les travaux précédemment cités reposent tous sur un apprentissage de cette corrélation. L'explication en est simple : la modélisation du comportement de l'enveloppe spectrale en fonction de la fréquence fondamentale fait intervenir une multitude de paramètres difficilement quantifiables, comme nous l'avons montré dans le premier chapitre. Cette compléxité entraîne une grande variabilité des données. La modélisation statistique et l'*apprentissage* de ces modèles permet de déterminer, à travers la diversité des manifestations des données, une caractéristique commune, un *prototype* qui régit ces données à travers leur variabilité [1].

5.2 Le mélange de Gaussiennes - GMM

Le « mélange de gaussiennes » est un modèle utilisé couramment à la fois pour des problèmes d'apprentissage et de classification de données. L'apprentissage d'un mélange de gaussiennes consiste à déterminer les paramètres du mélange qui expliquent le mieux la distribution des données. L'apprentissage peut être supervisé ou non. Dans le cas supervisé, l'apprentissage est realisé indépendemment sur un certain nombre de classes jugées pertinentes. La classification dite *a posteriori* permet alors, une fois déterminés les paramètres de ces mélanges, de déterminer, pour un vecteur quelconque, les probabilités conditionnelles d'appartenance de ce vecteur à chacune des classes.

Notre objectif est d'estimer et de maximiser la probabilité qu'étant donnée une fréquence fondamentale f_0, nous ayons une enveloppe spectrale env, autrement dit d'estimer la probabilité

1. Cette démarche par *classification* et *apprentissage* s'apparente à la manière dont l'être humain fait lui-même l'apprentissage des données du monde qui l'entoure. Ainsi, un « son de piano » n'existe pas *en soi*, mais est appris et reconnu en tant que tel à travers un processus complexe d'*apprentissage* et de *catégorisation*.

conditionnelle :

$$P(env|f_0) \tag{5.1}$$

Y. Stylianou a montré dans une étude comparative des modèles d'apprentissage [Styl96] que le choix des mélanges de gaussiennes était le modélisation qui donnait les meilleurs résultats dans le cadre de la conversion de la voix.

Nous noterons par la suite :
x = f_0, de dimension 1
y =env de dimension p, où p est l'ordre de l'estimation de l'enveloppe
et z = [x ;y], le vecteur conjoint de x et de y, de dimension d = p+1.

On supposera, nous placant dans le cadre d'une modélisation par mélange de gaussiennes, que la distribution de chacun de ces vecteurs suit une loi gaussienne.

Dans le cas d'une distribution multi-dimensionnelle de dimension d, la densité de probabilité du vecteur z s'écrit :

$$\mathcal{N}(\mathbf{z}|(\mu;\mathbf{\Sigma}))=\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\mathbf{\Sigma})}}exp(-\frac{(\mathbf{z}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{z}-\mu)}{2}) \tag{5.2}$$

$$\text{où} \quad \begin{cases} \mu \text{ est l'espérance de la gausienne, de dimensions } (1 \times d), \\ \mathbf{\Sigma} \text{ est la matrice de variance de la gaussienne, de dimensions } (d \times d) \end{cases}$$

La distribution de probabilité est une une combinaison linéaire de gaussiennes :

$$P(z)=P(x,y) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(z|(\mu_k;\sigma_k)) \quad \text{avec} \quad \sum_{k=1}^{K} \alpha_k=1 \tag{5.3}$$

où K est le nombre supposé de gaussiennes du mélange.

Pour déterminer les paramètres du mélange de gaussiennes, c'est-à-dire les triplets :

$$(\mu_k;\sigma_k;\alpha_k)_{k\in\{1,\ldots,K\}} \tag{5.4}$$

Nous devons minimiser une distance entre les données et le modèle :

$$(\mu;\sigma;\alpha)_{opt}=min(distance(\text{données},\text{modèle})) \tag{5.5}$$

Nous minimisons la distance dite de Manahalobis :

$$d^2(x)=(x-\mu)^T\mathbf{\Sigma}(x-\mu) \tag{5.6}$$

5.2.1 L'algorithme estimation et maximisation - *Expectation Maximization*

Pour déterminer les paramètres $(\alpha_k, \mu_k, \Sigma_k)$ d'un mélange de gaussiennes, nous employons l'estimation dite d'estimation et de maximisation EM, qui est une méthode couramment utilisée dans ce cas. Cette méthode, introduite par A. Dempster [Dem77], est une méthode itérative de convergence vers un minimum local de l'écart entre les données et les gaussiennes au sens de la log-vraisemblance.

Il s'agit d'estimer les paramètres qui maximisent la quantité :

$$\mathcal{L}_k = \sum_{p=1}^{K} log(p_p(x_i)) \tag{5.7}$$

Ce qui est fait itérativement en deux étapes : une étape dite d'estimation, et une étape de maximisation.

L'algorithme fonctionne de la manière suivante :

0. Initialisation des paramètres $(\alpha_k, \mu_k, \Sigma_k)$
1. Etape d'estimation : calcul des probabilités d'appartenance de chacune des données à la classe j

$$P(j|x_p) = \alpha_j \mathcal{N}(x_i, (\mu_p; \Sigma_p)) \tag{5.8}$$

2. Etape de maximisation : calcul des paramètres du modèle en fonction des probabilités précédentes

$$\begin{cases} \alpha_j = \frac{1}{n} \sum_{p=1}^{N} P(j|x_p) \\ \mu_j = \frac{\sum_{p=1}^{N} P(j|x_p) x_p}{\sum_{p=1}^{N} P(j|x_p)} \\ \Sigma_j = \frac{\sum_{p=1}^{N} P(j|x_p)(x_p - \mu_j)(x_p - \mu_j)^T}{\sum_{p=1}^{N} P(j|x_p)} \end{cases} \tag{5.9}$$

3. On réitère les étapes 2 et 3 jusqu'à parvenir à un point fixe de la log-vraisembalnce

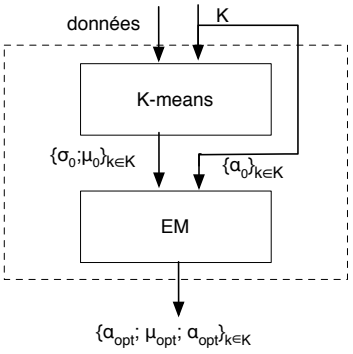


FIG. *5.1:* L'algorithme EM avec initialisation K-means dans le cadre d'une modélisation GMM

Le choix de l'initialisation est crucial dans l'estimation des paramètres, puisque EM va converger vers un minimum, mais pas nécessairement vers le minimum global.

5.2.2 K-means

L'initialisation de l'estimation et maximisation peut être menée entre autres par la méthode des k-moyennes, *k-means*. La méthode des k-moyennes, introduite par J.McQueen [McQ71], est utilisée dans des problèmes de classification, et consiste en le partitionnement d'un ensemble de données en un certains nombre de sous-ensembles disjoints. Chaque sous-ensemble est déterminé par minimisation de l'écart quadratique au barycentre de l'ensemble.

La minimisation se fait par convergence vers un point fixe en terme de barycentre.

On considère un ensemble de données $\{x_1, ..., x_N\}$, avec $x_i \in R^d$, où d est la dimension de l'espace des données. Soit k le nombre de classes utilisées.

La méthode des k-moyennes fonctionne itérativement de la manière suivante :

0. On initialise en distribuant de manière aléatoire les barycentres des classes, ce qui revient à affecter chacune des données aléatoirement à chacune des classes.
1. On détermine ensuite l'appartenance de chacun des vecteurs à une classe en appliquant une fonction :

$$f : x_i \in R^d \mapsto j \tag{5.10}$$

La fonction f est généralement une fonction de mimisation de la distance euclidienne par rapport au barycentre de chacune des classes :

$$f(x_i) = min_j ||x_i - \mu_j||^2 \tag{5.11}$$

2. Les barycentres sont alors recalculés à partir des données appartenant à la classe :

$$\mu_i = \frac{1}{card(x \in C_i)} \sum_{x \in C_i} x \tag{5.12}$$

On réitère les étapes 1 et 2 jusqu'à parvenir à la convergence.

La convergence de cette méthode est montrée, mais elle est locale, et dépend de l'initialisation. D'où la méthode - coûteuse - qui consiste à réitérer la méthode un certain nombre de fois et de choisir ensuite la solution qui minimise la distance sur l'ensemble des itérations.

Les k-moyennes permettent donc de déterminer des paramètres initiaux $(\mu_0; \sigma_0)$ de la méthode d'estimation et de maximisation. Les amplitudes des gausiennes sont choisies initialement égales, initialisées à $\alpha_0 = \frac{1}{K}$.

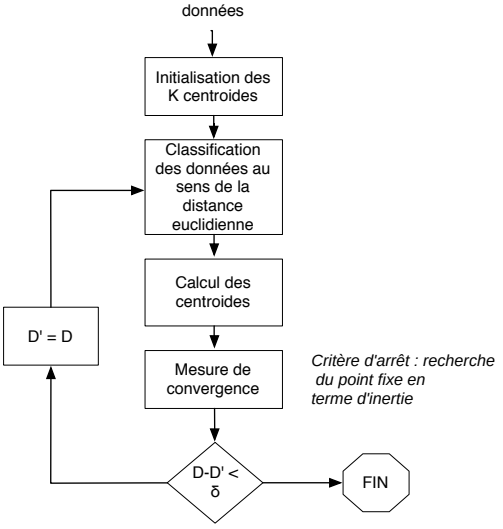


FIG. *5.2:* Diagramme du calcul des K-moyennes

5.3 Modèle de prédiction

Il nous faut donc, dans une étape postérieure à celle de l'apprentissage, déterminer une fonction qui pour une certaine fréquence fondamentale, renvoie une envellope d'après le mod`le estimé.

On montre [Kay93] que dans le cas uni-gaussien, le vecteur cible y qui minimise l'erreur quadratique moyenne entre la prédiction et le modèle est donné par la régression linéaire :

$$E[y|x] = \mu^y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu^x) \tag{5.13}$$

où μ^x et μ^y sont respectivement l'espérance des vecteur x et y :

$$\mu_x \widehat{=} E[x]$$

$$\mu_y \widehat{=} E[y]$$

et Σ_{yx} est la matrice d'auto-covariance du vecteur conjoint z :

$$\Sigma_{yx} \widehat{=} E[(y - \mu_y)(x - \mu_x)^T]$$

déterminés par l'estimation des paramètres du modèle de la probabilité conjointe de z.

Y. Stylianou a proposé [Sty98] une fonction de prédiction pour le cas général du mélange de gaussiennes. Il s'agit simplement de la somme des fonctions de prédiction locale pondérées par la probabilité *a posteriori* que le vecteur source x appartienne à la classe c_i, soit :

$$F(x)=E[y|x] = \sum_{k=1}^{K} h_k(x)[\mu_k^y + \Sigma_k^{yx}(\Sigma_k^{xx})^{-1}(x - \mu_k^x)] \tag{5.14}$$

$$\text{avec} \quad \Sigma_k \quad = \quad \begin{bmatrix} \Sigma_k^{yy} & \Sigma_k^{yx} \\ \Sigma_k^{xy} & \Sigma_k^{xx} \end{bmatrix} \quad \text{et} \quad \mu_k \quad = \begin{bmatrix} \mu_k^y \\ \mu_k^x \end{bmatrix}$$

et

$$h_k(x) = \frac{\alpha_k \mathcal{N}(x|\{\mu_k^x, \Sigma_k^{xx}\})}{\sum_{i=1}^{Q} \alpha_i \mathcal{N}(x|\{\mu_i^x, \Sigma_i^{xx}\})}$$

Chapitre 6

Mise en évidence et apprentissage de la corrélation entre la f_0 et l'enveloppe spectrale

La démarche adoptée dans ce chapitre est de mettre en évidence la corrélation de la fréquence fondamentale et de l'enveloppe spectrale, de montrer que l'on peut modéliser par apprentissage cette corrélation de manière satisfaisante, et que la prise en compte de ce modèle permet d'améliorer le modèle de transposition classique par conservation de l'enveloppe spectrale. Nous présentons dans ce chapitre le cadre général du protocole d'apprentissage et de prédicion que nous avons utilisé, et mettons en évidence des résultats obtenus sur un ensemble restreint d'exemples, démarche que nous généraliserons par la suite.

6.1 Dispositif expérimental

6.1.1 Le corpus de voix

Nous avons utilisé pour notre étude un corpus de vingt minutes environ de voix d'homme enregistrée dans une chambre anéchoïque. La transcription phonétique de ce corpus a été verifiée « à la main ». La base de données que nous avons spécialement réalisée n'est pour le moment pas encore utilisée pour l'apprentissage. Le registre du locuteur se trouve approximativement entre 80Hz et 250Hz.

6.1.2 La base de données *talkapillar*

Nous avons utilisé le système d'analyse et de synthèse *talkapillar* [Bel05], qui est une base de données développée à l'IRCAM initialement pour la synthèse par concaténation d'unités. Ce système est de plus en plus utilisé aujourd'hui pour l'analyse et l'apprentissage de données. Il contient des informations à deux niveaux du contenu des phrases : une représentation symbolique par semi-phone, phone et diphone, ainsi qu'une représentation acoustique qui peut prendre en compte n'importe

quel type de descripteurs acoustique du signal, que nous pouvons définir et ajouter à la base de données.

6.1.3 Les descripteurs utilisés

Phonétisation

La phonétisation est réalisée par le logiciel Euler.

Pré-traitement

Afin de réduire l'ordre des coefficients d'estimation de l'enveloppe spectrale, nous avons sous échantillonné à 11025 Hz, avec filtrage passe-bas préalable pour éviter d'éventuels phénomènes de repliement spectral.

Estimation de l'enveloppe spectrale

L'enveloppe spectrale est estimée sur l'ensemble de la base de données de la manière suivante :

- trames de longueurs fixée à 30ms
- pas d'avancement *pitch-synchroned*, avec des marques calculée par la méthode PSOLA.
- estimation à partir des coefficients cepstraux, d'ordre constant 55, par la méthode true envelope.
- estimation des LSF d'ordre 20 à partir des coefficients cepstraux précédents par la méthode dite *LPC-True Enveloppe* présentée dans [Vil06].

Estimation de la fréquence fondamentale

La fréquence fondamentale est estimée par l'algorithme YIN. L'algorithme YIN consiste en une estimation de la fréquence fondamentale par autocorrelation normalisée.

L'estimation de la fréquence fondamentale se fait de la manière suivante :

- trames de longueur fixée à 30ms
- pas d'avancement *pitch-synchroned*, avec des marques calculée par la méthode PSOLA.
- estimation de la fréquence fondamentale par YIN
- décision voisée/non-voisée à partir de la mesure d'aperiodicité et d'énergie réalisée par YIN.
- filtrage médian d'ordre 15 pour lisser l'estimation, et notamment supprimer d'éventuelles grosses erreurs ponctuelles de l'estimation.

Estimation du noyau vocalique

Le noyau vocalique est déterminé à partir de l'« Acoustic Center of Reliability », tel que défini dans le premier chapitre. La probabilité de se trouver dans un noyau est transformé en une décision binaire par seuillage sur chacun des paramètres.

6.1.4 Mesure de performance

Le critère d'erreur retenu pour mesurer la qualité de prédiction entre l'enveloppe prédite et l'enveloppe connue sont les mesures de distorsions spectrales, en échelle linéaire et en échelle de mel [Rab93]. La première donne une mesure objective de la différence des deux enveloppes, comme l'aire en db de la différence au carré du logarithme des amplitudes des enveloppes connues et prédites. C'est donc une mesure de l'erreur quadratique moyenne entre les deux enveloppes.

La mesure de distorsion spectrale en échelle linéaire s'énonce de la manière suivante :

$$d^2 = \int_{-\pi}^{\pi} |logS(\omega) - logS'(\omega)|^2 \frac{d\omega}{2\pi} \tag{6.1}$$

Ce qui peut autrement s'écrire :

$$d^2 = \sum_{-\infty}^{\infty} (c_n - \widehat{c}_n)^2 \tag{6.2}$$

La deuxième est une mesure perceptive rendant compte de la perception de cette différence. Elle consiste en un préfiltrage par le filtre de l'oreille moyenne des deux enveloppes, puis à la même mesure que précédemment faite sur les énergies intégrées sur des bandes de Mel. Nous avons choisi pour notre procédure une banc de 30 bandes de Mel.

Soit $\tilde{S}(\omega)$ l'énergie du spectre à la sortie de chacune des bandes de fréquences.

On note :

$$\tilde{c}_n = \sum_{k=1}^{K} log\tilde{S}_k.cos[n(k - \frac{1}{2})\frac{\pi}{K}] \quad , \quad n = 1, .., .L \tag{6.3}$$

De manière similaire à (6.2) , la distorsion spectrale en échelle de Mel s'écrit :

$$d^2_{\tilde{c}} = \int_{-\pi}^{\pi} |log\tilde{S}(\omega) - log\tilde{S}'(\omega)|^2 \frac{d\omega}{2\pi} \tag{6.4}$$

6.1.5 Présentation du protocole expérimental général

A partir du corpus utilisé, nous faisons l'extraction des descripteurs phonétique, de la f_0, de l'enveloppe et du noyau vocalique. Cette base de données est alors divisée en une base d'apprentissage et une base de test, selon la démarche courante dans les méthodes d'apprentissage. L'apprentissage est fait avec un nombre de composantes du mélange variable. Une fois le modèle estimé pour chacune de classes phonétiques par apprentissage sur la corpus d'entraînement, nous utilisons les données contenues dans celui-ci pour estimer la qualité de l'apprentissage, et les données contenues dans le corpus de test pour estimer la capacité de généralisation de notre modèle. La prédiction de l'enveloppe, à partir d'un vecteur quelconque (f_{0test}, env_{test}), est réalisée de la manière suivante :

(1) Estimation des probabilités de (f_{0test}, env_{test}) d'appartenance à chacune des classes phonétiques, soit

$$P(c_j|(f_{0test}, env_{test})) \quad , \quad \forall j \in [1, ..., 24]$$

(2) Prédiction de l'enveloppe à partir de f_{0test} pour chacune des classes phonétiques, suivant la fonction de prédiction.

L'enveloppe prédite est alors définie comme la somme des différentes enveloppes prédites pour chacune des classes, pondérées par la probabilité d'appartenance à chacune de ces classes. L'enveloppe

prédite est un barycentre des enveloppes prédites pour chaque classe. La nécessité que nous avons d'interpoler les différentes enveloppe justifie le choix de la modélisation par les LSF de l'enveloppe, car celles-ci présentent de bonnes propriétés d'interpolation [Isl00], [Pal95]. Nous synthétisons l'ensemble de notre protocole d'apprentissage, de prédiction et d'estimation dans les figures [6.1] et [6.2].

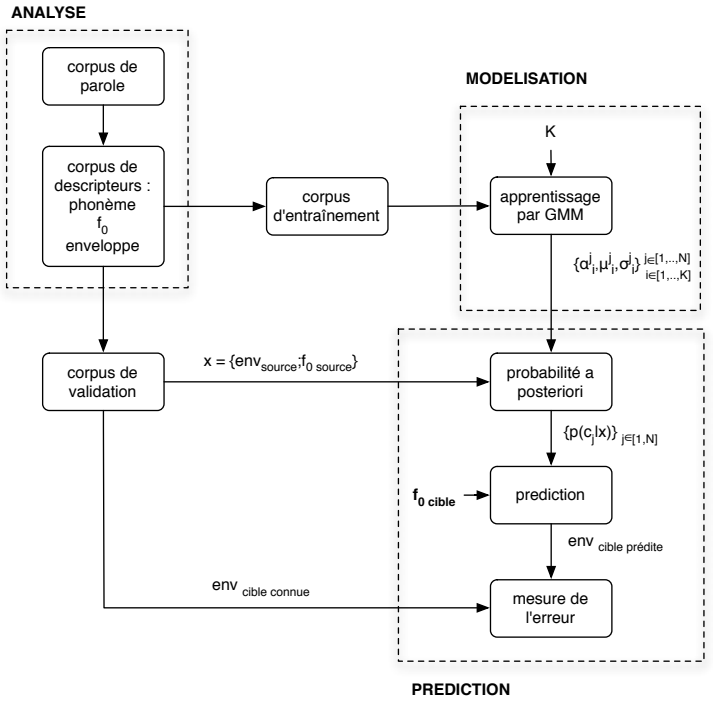


FIG. *6.1:* Diagramme de modélisation par apprentissage, prédiction de l'enveloppe spectrale et mesure de la qualité de la prédiction

6.2 Expérience : apprentissage supervisé par classes phonétiques sur l'ensemble des noyaux vocaliques

6.2.1 Procédure

Pour mettre en évidence et apprendre la modélisation de la corrélation de l'enveloppe spectrale et de la fréquence fondamentale, nous avons réalisé un apprentissage conjoint de ces deux paramètres. L'apprentissage est supervisé sur les phonèmes voisés de la langue française, soit 24 classes, et réalisé sur les noyaux vocaliques afin d'isoler au maximum l'influence de la seule fréquence fondamentale, en supprimant l'influence du contexte phonétique. Le nombre de composantes du mélange est variable. Notre démarche expérimentale est axée en trois étapes : mettre en évidence la corrélation de la fréquence fondamentale et de l'enveloppe spectrale, modéliser le comportement de l'enveloppe spectrale en fonction de la f_0 obtenue par l'apprentissage et comparer le modèle avec le comportement réels des données, et mesurer l'amélioration obtenue par la prédiction de l'enveloppe spectrale par rapport au modèle classique de conservation de l'enveloppe.

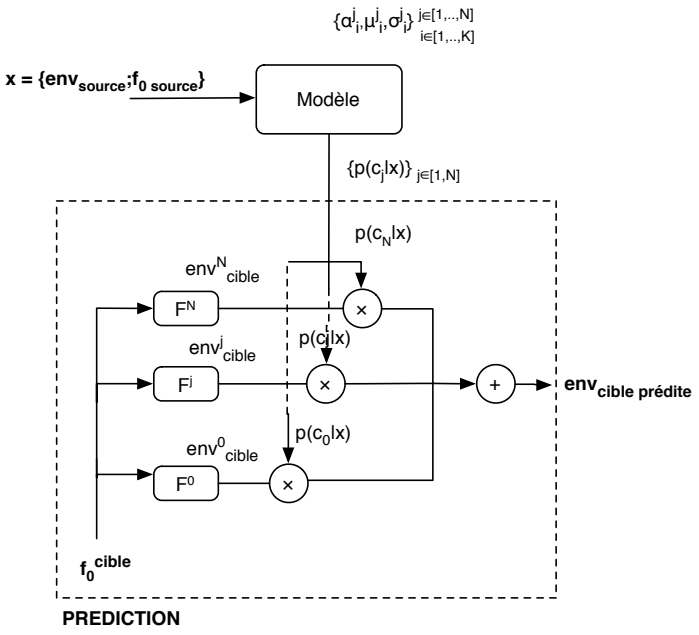


FIG. *6.2:* Diagramme de prédiction de l'enveloppe spectrale

Nous présentons dans le tableau ci-après le nombre de données que contient chacune des classes phonétiques que nous avons utilisé :

Phonème	**nombre de réalisations**	**nombre de trames**
[XSAMPA]	[unités]	[trames]
a	522	83340
e	410	66436
u	199	31982
i	433	65132
o	127	26367
a	252	51081
o	170	32865
9	128	28280
2	96	20637
e	121	25169
@	337	41218
9	67	14430
E	362	59754
O	196	30522
b	164	21317
d	276	30736
g	102	12369
v	177	20937
z	165	23813
m	236	31741
n	258	30985
j	151	19587
l	536	63719
y	190	28130

6.2.2 Résultats

Nous présentons ici les résultats obtenus à partir de tests réalisés à l'intérieur de quelques classes phonétiques, pour mettre en valeur la qualité de l'apprentissage et de la prédiction au niveau de la classe. Les résultats présentés ont valeur d'exemple. La prédiction globale, telle que nous l'avons presentée dans la section précédente , sera menée ultérieurement. Actuellement, la classe phonétique est supposée connue.

Mise en évidence de la corrélation entre la f_0 et l'enveloppe spectrale

Nous calculons pour chaque classe phonétique la distribution de f_0 à l'intérieur de chacune des composantes du mélange de gaussiennes. L'appartenance de la f_0 à la composante j est déterminée comme la f_0 associée à l'enveloppe spectrale qui maximise la probabilité conditionelle *a posteriori* $P(j|env)$. Le résultat obtenu met en évidence l'existence d'une forte correlation entre la f_0 et l'enveloppe spectrale. Ce qui confirme l'étude réalisée par [EnN03], [EnN05] ; et est d'ailleurs plus efficace puisque celui-ci réalisait un apprentissage non-supervisé sur l'ensemble des trames voisées de sa base de parole, ce qui bien évidemment rend moins évidente la corrélation, puisqu'elle ajoute une variabilité relevant de la classe phonétique. La variance des f_0 augmente aux extrêmes valeurs de la f_0, ce qui s'explique bien par le fait que les données sont moins nombreuses dans ces registres « extrêmes », ayant pour conséquence d'augmenter la variance de la gaussienne le long de l'axe des

f_0 comme le montre la figure [6.3]. Nous représentons dans la figure [6.4] la distribution des f_0 à l'intérieur des classes.

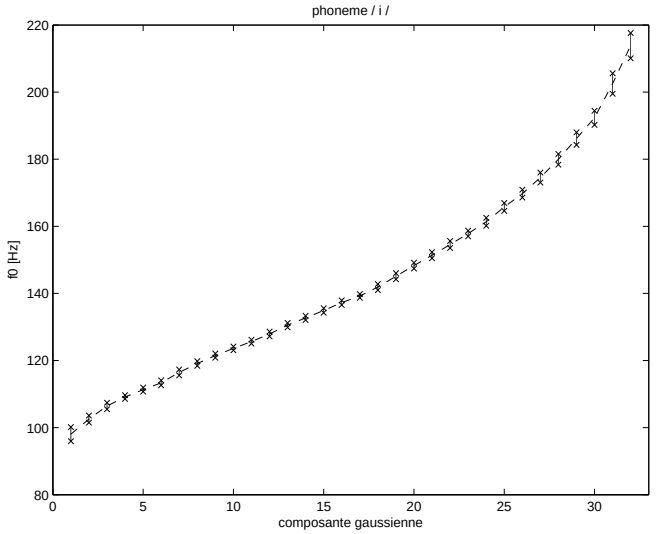


FIG. *6.3:* Représentation de la f_0 contenue dans chacune des classes après apprentissage avec 32 gaussiennes sur le phonème /i/. Les composantes sont ordonnées par ordre croissant de f_0 moyenne. Les barres verticales représentent la déviation standard de la f_0 dans chacune des composantes

Prédiction de l'évolution de l'enveloppe spectrale en fonction de la f_0

Nous avons estimé la qualité de la prédiction sur la base d'entraînement, performance qui donne une idée de la qualité du modèle résultant de l'apprentissage. L'estimation de cette même prédiction sur la base de test donne la capacité de généralisation de ce modèle. Nous n'avons seulement estimé la qualité de la modélisation. Pour ce faire, nous avons réalisé un apprentissage sur le vecteur conjoint de la fréquence fondamental et de 20 LSF. La base d'entraînement contient 80% des données, et la base de test, 20%. L'entraînement est fait de 4 à 64 composantes, par ordre croissant de puissance de 2.

En prenant l'exemple de la prédicion résultante dans la cas du phonème /i/ pour une fréquence fondamentale allant de 100Hz à 210 Hz, nous montrons clairement l'évolution de l'enveloppe spectrale. Les brusques variations de position des lignes spectrales que l'on constate sur la figure [6.8] en dessous et au dessus de ces valeurs sont dues à un manque de données dans ces régions fréquentielles, comme le montre l'histogramme de la figure [6.4]. On peux noter le déplacement important de la 5ème ligne spectrale autour de 1000Hz sur la figure [6.5] et sa conséquence sur l'enveloppe spectrale sur la figure [6.6]. Le déplacement du premier formant avec la fréquence fondamentale est montré sur les figure [6.7] et [6.9][1]. Le deuxième formant est lui relativement stable autour de 3250Hz, mais sa largeur de bande et son énergie augmente avec la fréquence fondamentale [6.10]. On constate la dégradation de la prédiction avec l'augmentation du nombre de composantes du mélange : les figures [6.5] et [6.8] montrent la prédiction résultante d'un mélange respectivement de 16 et de 64 gaussiennes : la variance de la prédiction augmente dont la cause est une sur-modélisation des données.

1. Rectification : l'échelle des fréquences sur l'axe de la fréquences fondamentale est erronnée, il s'agit d'une prédiction entre 100Hz et 220Hz, comme sur la figure [6.10].

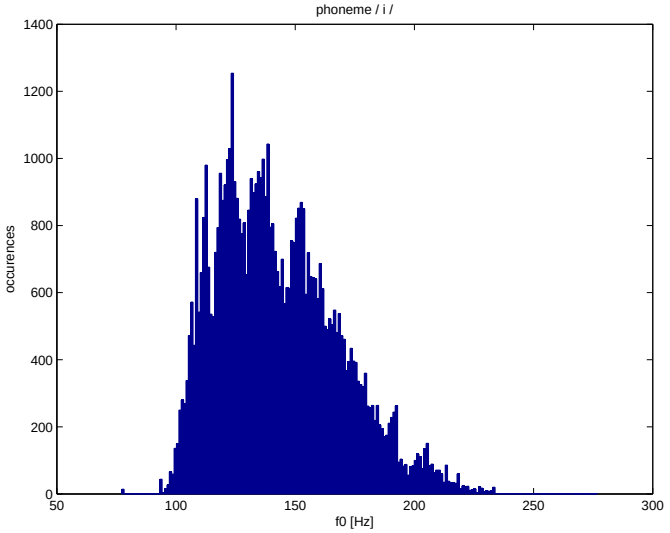


FIG. *6.4:* Distribution des réalisations du phonème /i/ en fonction de la fréquence fondamentale

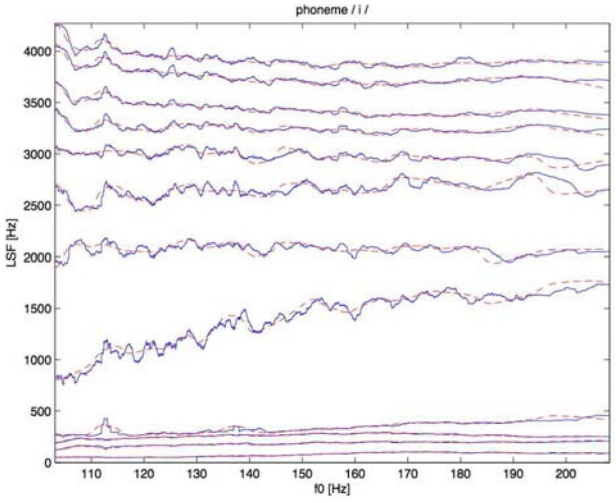


FIG. *6.5:* Prédiction des LSF du phonème /i/ pour un mélange de 16 gaussiennes pour un f_0 variant de 100Hz à 210Hz. La prédiction est en trait pointillé, et les données réelles, filtrées par une moyenne courante, sont en trait plein

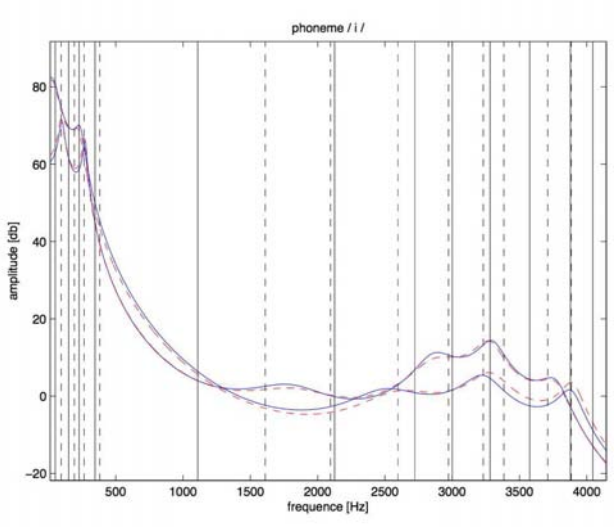


FIG. *6.6:* Prédiction de l'enveloppe spectrale du phonème /i/ pour un mélange de 16 gaussiennes sur l'exemple de f_0 prises à 120Hz et 210Hz. La prédiction est en trait pointillé, et les données réelles, filtrées par une moyenne courante, sont en trait plein

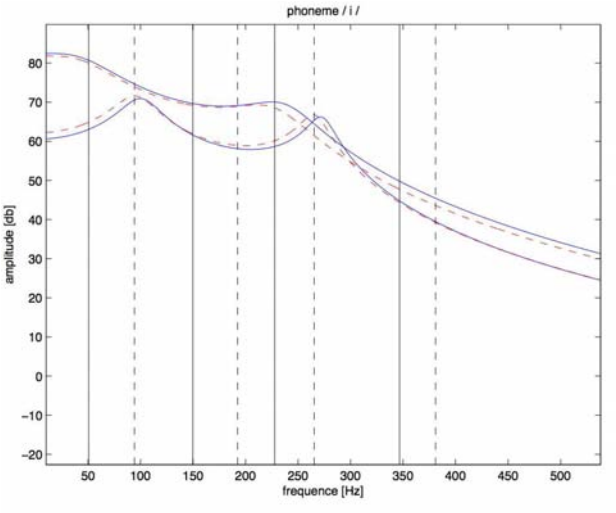


FIG. *6.7:* Prédiction de l'enveloppe spectrale du phonème /i/ pour un mélange de 16 gaussiennes sur l'exemple de f_0 prises à 120Hz et 210Hz. « Gros plan » sur la partie basse du spectre. La prédiction est en trait pointillé, et les données réelles, filtrées par une moyenne courante, sont en trait plein

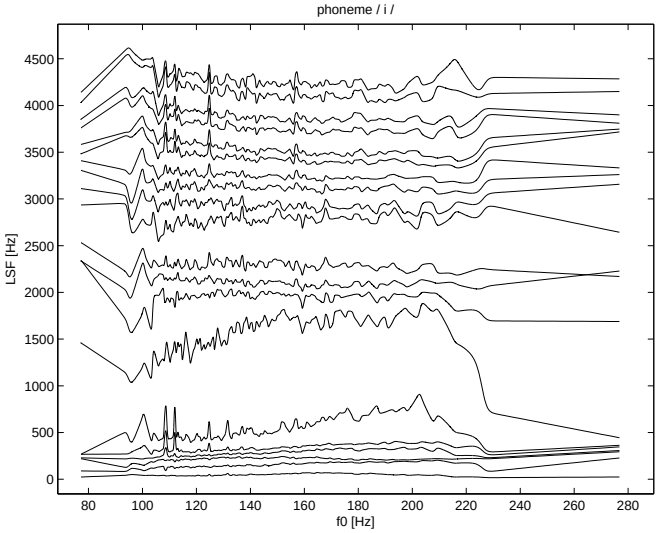


FIG. *6.8:* Prédiction des LSF du phonème /i/ pour un mélange de 64 gaussiennes

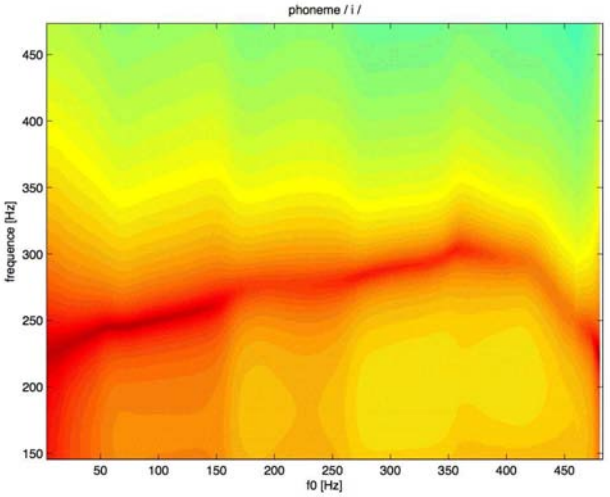


FIG. *6.9:* Déplacement du premier formant du phonème /i/ en fonction de la fréquence fondamentale

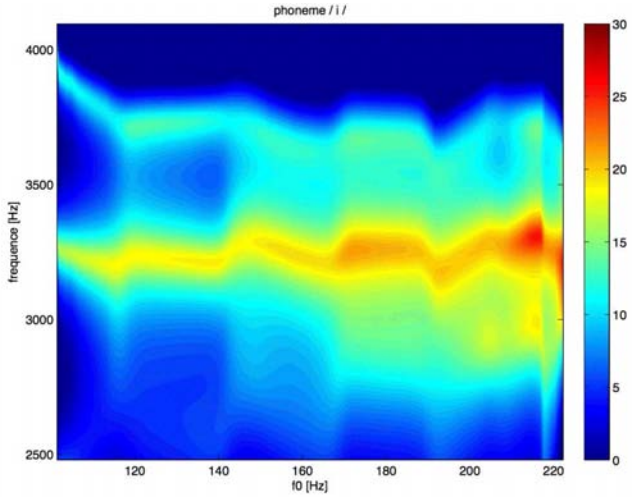


FIG. *6.10:* Prédiction de la position du deuxième formant du phonème /i/ en fonction de la fréquence fondamentale

Comparaison de l'enveloppe determinée la prédiction avec le modèle de conservation de l'enveloppe

Afin de mesurer la performance de notre prédiction, nous avons réalisé l'expérience suivante : pour des valeurs de f_0 source allant de la valeur la plus basse à la valeur la plus haute lors de l'apprentissage, par pas de 1Hz, nous prédisons l'enveloppe obtenue pour chacune des f_0 cibles décrivant le même rang de valeurs. Pour chacun des couples $(f_0, f_0^{'})$, nous connaissons l'enveloppe source, l'enveloppe cible et l'enveloppe prédite.

Nous proposons une mesure de la performance de notre prédiction par rapport au modèle de conservation de l'enveloppe source de la manière suivante :

On pose : $d^2_{conservation}(f_0, f_0^{'})$ la distorsion spectrale entre l'enveloppe conservée, c'est-à-dire l'enveloppe de la trame source, et l'enveloppe de la trame cible.
$d^2_{prediction}(f_0, f_0^{'})$ la distorsion spectrale entre l'enveloppe prédite et l'enveloppe de la trame cible.
Et on note de manière similaire $d^2_{\tilde{c},conservation}$ et $d^2_{\tilde{c},prediction}$, les mêmes mesures faites sur des bandes de fréquences en échelle de Mel après filtrage par l'oreille moyenne.

Le gain obtenu par la prédiction est estimé de la manière suivante :

$$G(f_0, f_0^{'}) = log(\frac{d^2_{conservation}(f_0, f_0^{'})}{d^2_{prediction}(f_0, f_0^{'}))} \tag{6.5}$$

Et le gain perceptif estime sur les bandes de Mel, après filtrage par l'oreille moyenne :

$$G_{subj}(f_0, f_0^{'}) = log(\frac{d^2_{\tilde{c},conservation}(f_0, f_0^{'})}{d^2_{\tilde{c},prediction}(f_0, f_0^{'}))} \tag{6.6}$$

Posée de cette manière, la performance de la prédiction est meilleure que la conservation de l'enveloppe lorsque la valeur du gain est supérieure à 0, moins bonne sinon.

Nous montrons un exemple de résultats dans la cas du phonème /o/ réalisé successivement sur la base d'entraînement et la base de test sur les figures [6.11],[6.12] et [6.13] et [6.14].

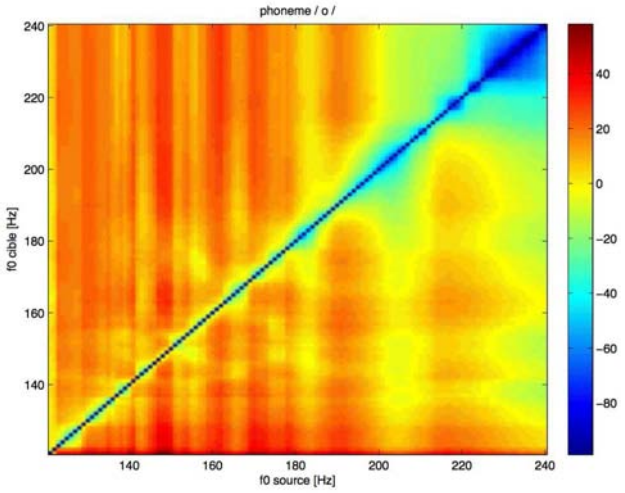


FIG. *6.11:* Gain obtenu par la prédiction de l'enveloppe par rapport à la conservation de l'enveloppe, mesuré sur la base d'entraînement. Exemple du phonème /o/.

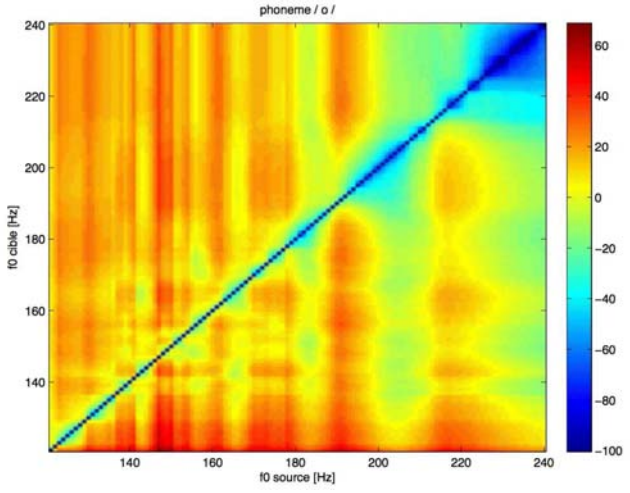


FIG. *6.12:* Gain perceptif obtenu par la prediction de l'enveloppe par rapport a la conservation de l'enveloppe, estime sur la base d'entraînement. Exemple du phonème /o/.

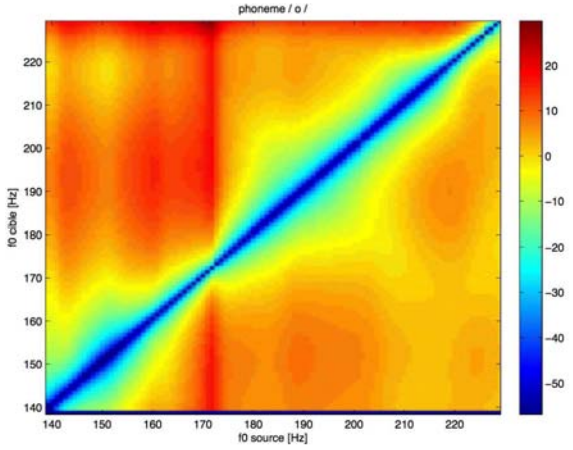


FIG. *6.13:* Gain obtenu par la prédiction de l'enveloppe par rapport à la conservation de l'enveloppe, mesuré sur la base de test. Exemple du phonème /o/.

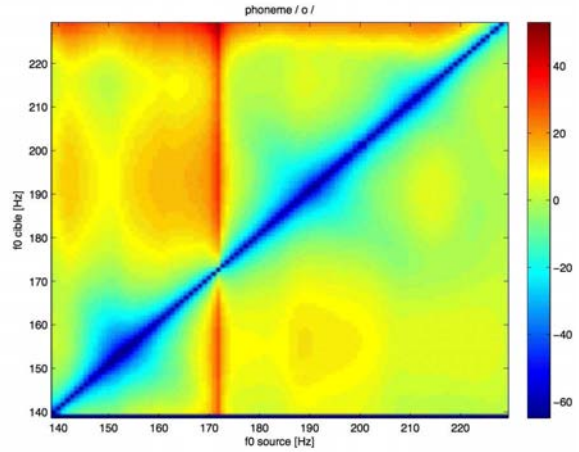


FIG. *6.14:* Gain perceptif obtenu par la prédiction de l'enveloppe par rapport à la conservation de l'enveloppe, mesuré sur la base de test. Exemple du phonème /o/.

Sur la base d'entraînement [6.8] [6.9], la prédiction est meilleure de manière quasi-homogène que la conservation de l'enveloppe, sur la mesure objective comme perceptive. Les « pics et creux de performance » pour certaines fréquences sources correspond simplement à la distribution des f_0 de la base d'entraînement. Sur la base de test [6.10] [6.11], l'étude du résultat conduit aux conclusions suivantes : lorsque la transposition est faible la conservation est meilleure que l'enveloppe prédite. Mais rapidement la prédiction devient meilleure que la conservation, d'autant plus que l'écart de fréquence entre la source et la cible est grande. Le « pic de performance » vers 170 Hz s'explique de la même manière que précédemment. Notons enfin que la matrice de gain n'est pas symétrique : la performance est meilleure lorsque l'on transpose vers le haut que vers le bas.

Conclusions et perspectives

Nous avons mis en évidence le rôle de l'enveloppe spectrale dans la transposition de la voix parlée ainsi que la corrélation de la fréquence fondamentale et de l'enveloppe spectrale. Par un apprentissage de cette corrélation sur les classes phonétiques, nous avons modélisé de manière satisfaisante l'évolution de l'enveloppe spectrale avec la fréquence fondamentale sur l'exemple de quelques phonèmes, dans le cas où le phonème est supposé déjà connu. La prédiction réalisée de l'enveloppe à partir de notre modèle montre de meilleures performances que la conservation de l'enveloppe actuellement utilisée dans les méthodes de transposition de la voix parlée. Il s'agit pour nous maintenant de généraliser les résultats obtenus à l'ensemble des classes phonétiques et de tester la performance de la prédiction dans le cas ou le phonème est supposé inconnu, c'est-à-dire le cas tel qu'il se présente véritablement dans le problème de la transposition. Nous n'avons pris en compte dans notre apprentissage que l'enveloppe spectrale par rapport à la fréquence fondamentale. D'autres descripteurs pourraient avantageusement étre utilisés : par exemple pour étudier la corrélation de l'energie et de la fréquence fondamentale. Mais également les dérivées premières et secondes de l'enveloppe spectrale, ce qui nous permettrait sans doute de modéliser plus finement l'évolution de l'enveloppe spectrale, notamment dans les transitions entre phonèmes, qui est pour le moment modélisée en premiére approximation comme une interpolation pondérée entre les enveloppes des phonèmes. Il serait également intéressant de comparer les résultats obtenus par différentes modélisation de l'enveloppe, ainsi qu'avec ou sans la prise en compte du noyau vocalique. Nous pourrions également utiliser les méthodes de détermination du nombre optimal de composantes du mélange de gaussiennes, comme les algorithmes « gloutons» - EM Greedy Algorithm -, ou d'après le critère BIC. Celà nous affranchirait du choix du nombre de composantes. Enfin, l'augmentation de la taille de notre corpus d'entraînement et l'utilisation du corpus que nous avons enregistré permettrait d'améliorer sensiblement la modélisation de la corrélation, et conséquemment de la prédiction. Toutes choses que nous proposons de réaliser dans la seconde moitié de notre stage.

Phrases du corpus

Phrase d'exemple :
Ils n'ont pas l'air de croire à leur bonheur,
Et leur chanson se mêle au clair de lune.[1]
_ i l n o ~ p a l E R d @ k r w a R a l 9 R b O n 9 R _
e l 9 R S a ~ s o ~ s @ m E l o k l E R d @ l y n _

Phrase 1 :
Fuis le Hun qui donne le camping au moins chanceux.
_ f H i l @ 9 ~ k i d O n l @ k a ~ p i N o m w e ~ S a ~ s 2 _
Donne le moins camping qui fuis au Hun le chanceux.
_ d O n l @ k a ~ p i N k i f H i o 9 ~ l @ S a ~ s 2 _
Le chanceux fuis au moins le camping qui donne Hun.
_ l @ S a ~ s 2 f H i o m w e ~ l @ k a ~ p i N k i d O n 9 ~ _
Au moins donne le Hun au chanceux qui le fuis camping.
_ o m w e ~ d O n l @ 9 ~ o S a ~ s 2 k i l @ f H i k a ~ p i N _
Le camping chanceux fuit moins qui au Hun le donne.
_ l @ k a ~ p i N S a ~ s 2 f H i m w e ~ k i o 9 ~ l @ d O n _

Phrase 2 :
Bonjour Xavier, fais-en une pâte au beurre et aux oignons.
_ b o ~ Z u R g z a v j e _ f E z a ~ y n p A t o b 9 R e o z o J o ~ _
Au Beurre fais-en une Xavier, aux oignons et Bonjour pâte.
_ o b 9 R f E z a ~ y n g z a v j e _ o z o J o ~ e b o ~ Z u R p A t _
Oignons une pâte au beurre et aux Xavier, en fais bonjour.
_ o J o ~ y n p A t o b 9 R e o g z a v j e _ a ~ f E b o ~ Z u R _
Xavier, et une pâte aux oignons bonjour en fais au beurre.
_ g z a v j e _ e y n p A t o z o J o ~ b o ~ Z u R a ~ f E o b 9 R _
Pâte au Xavier, oignons fais-aux bonjour en beurre une.
_ p A t o g z a v j e _ o J o ~ f E z o b o ~ Z u R a ~ b 9 R y n _

1. Paul Verlaine, *Clair de lune* in Romance sans parole

Phrase 3 :
Pourquoi bonté divine faire un gosse aux yeux jeunes.
_ p u R k w a b o ~ t e d i v i n f E R 9 ~ g O s o z j 2 Z 9 n _
Gosse divine faire un bonté pourquoi aux jeunes yeux.
_ g O s d i v i n f E R 9 ~ b o ~ t e p u R k w a O Z 9 n j 2 _
Faire yeux bonté aux gosses pourquoi un jeune divine.
_ f E R Z 9 b o ~ t e O g O s p u R k w a 9 ~ Z 9 n d i v i n _
Jeunes gosse pourquoi faire un divine yeux aux bontés.
_ Z 9 n g O s p u R k w a f E R 9 ~ d i v i n Z 9 O b o ~ t e _
Aux jeunes un faire bonté gosse yeux divine pourquoi.
_ O Z 9 n 9 ~ f E R b o ~ t e g O s z j 2 d i v i n p u R k w a _

Phrase 4 :
Une huile hâtivement cherche à gagner le ring.
_ y n H i l A t i v @ m a ~ S E R S a g a J e l @ R i N _
Le gagner une ring hâtivement à l'huile cherche.
_ l @ g a J e y n R i N A t i v @ m a ~ a l H i l S E R S _
Hâtivement cherche le ring une huile à gagner.
_ A t i v @ m a ~ S E R S E l @ R i N y n H i l a g a J e _
Cherche à une huile gagner le ring hâtivement.
_ S E R S a y n H i l g a J e l @ R i N A t i v @ m a ~ _
A une ring gagner hâtivement cherche le huile.
_ a y n R i N g a J e A t i v @ m a ~ S E R S l @ H i l _

Liste XSAMPA des phonèmes du français moderne

Le code XSAMPA est une extension de la norme SAMPA de transcription en code ASCII de l'alphabet standard API. Il est utilisé comme outil de transcription pour l'analyse de corpus oraux.[1]

1. cf. le site de la phonologie du français contemporain http ://infolang.u-paris10.fr/pfc/

Symbole ascii	Exemple	Transcription
p	pont	po ~
b	bon	bo ~
t	temps	ta ~
d	dans	da ~
k	quand	ka ~
g	gant	ga ~
f	femme	fam
v	vent	va ~
s	sans	sa ~
z	zone	zon
S	champ	Sa ~
Z	gens	Za ~
m	mont	mo ~
n	nom	no ~
J	oignon	oJo ~
N	camping	ka ~piN
l	long	lo ~
R	rond	Ro ~
w	coin	kwe ~
H	juin	ZHe ~
j	pierre	pjER
i	si	si
e	ses	se
E	seize	sEz
a	patte	pat
A	pâte	pAt
O	comme	kOm
o	gros	gRo
u	doux	du
y	du	dy
2	deux	d2
9	neuf	n9f
@	justement	Zyst@ma ~
e ~	vin	ve ~
a ~	vent	va ~
o ~	bon	bo ~
9 ~	brun	bR9 ~

Bibliographie

[Bel03] Beller, G., Hueber, T., Schwarz, D., Rodet, X., « An Overview of Talkapillar »,

[Bel05] Beller, G., Marty, A., « TALKAPILLAR : outil d'analyse de corpus oraux »,

[Cap]Cappé, O., Laroche, J., Moulines, E., « Regularized estimation of cepstrum envelope from discrete frequency points », *IEEE Signal Processing Letters*, vol.3 no4, p.100-102, 1996.

[Che02] De Cheveigné, A., Kawahara, H., « YIN, a fundamental frequency estimator for speech and music », *Journal of Acoustic Society of America*, 2002.

[Dem77] Dempster, A., Laird, N., Rubin, D., « Maximum likelihood from incomplete data via the EM algorithm », *Journal of the Royal Statistical Society*, Series B, 39(1) :p.1-38, 1977.

[DHa] D'Haes W., Rodet, X., « Discrete Cepstrum Coefficients as Perceptual Features »,

[Dol86] Dolson, M.B., « The phase vocoder : a tutorial », *Computer Music Journal*, p.14-27, 1986.

[Dov03] Doval, B., d'Alessandro, C., Henrich, N., « The voice source as causal/anticausal linear filter », *Proc. ISCA ITRW VOQUAL 2003*, Geneva, Switzerland, 2003, p.15-19

[ElJ91] El-Jarorudy, A., Makhoul, J., « Discrete all pole modelling », *IEEE Trans. Acoust., Speech, Signal Processing*, 39(2), 1991, p.411-423.

[EnN03] En-Najjary, T., Rossec, O., Chonavel, T., « A new method for pitch prediction from spectral envelope and its application in voice conversion », EUROSPEECH, 2003.

[EnN05] En-Najjary, T., Conversion de voix pour la synthèse de la parole, Thèse de l'Université de Rennes I, 2005.

[Fan60] Fant, G., *Acoustic theory of speech production*, Edition Mouton, La Hague, 1960.

[Fan70] Fant, G., « On the predictability of formant levels and spectrum envelopes from formant frequencies », *For Roman Jakobson*, The Hague, 1970, p.109-120.

[Fla66] Flanagan, J.L., Golden, R.M., « Phase vocoder », *Bell Systems Technical Journal*, p.1493-1509, 1966.

[Fon71] Fonagy, I., « Double coding in speech », Semiotica, 3, p.189-222, 1971.

[Gal90] Galas, T., Rodet, X., « An improved cepstral method for deconvolution of source filter systems with discrete spectra : Application to musical sound signals », *Proceedings of the International Computer Music Conference*, p.82-84, 1990.

[Gra74] Gray, A., Markel, J.D., « A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech signals », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.ASSP-22, no.3, p.207-217, 1974.

[Ima79] Imai, S.,Abe, Y., « Spectral envelope extraction by improved cepstral method », *Electron. and Commun. in Japan*, vol.62-A, no 4, p10-17, 1979.

[Isl00] Islam, T., Interpolation of Linear Prediction Coefficients for Speech Coding, Master Degree of the McGill University, Canada, 2000.

[Ita75] Itakura, F., « Line spectrum representation of linear predictive coefficients of speech signals », *J. Acoust. Soc. Am.*, vol. 57, no, 535(A), 1975.

[Kai00] Kain, A., Stylianou, Y., « Stochastic modeling of spectral adjustment for high quality pitch modification », *IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, 2000.

[Kai01] Kain, A., High Resolution Voice Transformation, PhD Thesis, Faculty of the Oregon Graduate Insitute of Science and Technology, 111p. 2001.

[Kam96] Kambhatla, N., *Local Models and Gaussian Mixture Models for Statistical Data Processing*, Thèse de l'Oregon Graduate Institute de Portland, 1996.

[Kay] Kay, S., Modern Spectral Estimation : Theory and Application, Prentice-Hall Signal Processing Series,

[Kay93] Kay, S., *Fundamentals of Statistical Signal Processing : Estimation Theory*, Englewood Cliffs, NJ, Prentice Hall, 1993.

[Lar99] Laroche, J., Dolson, M., « New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing, and other exotic audio modifications », *Journal of the Audio Engineering Society*, p.928-936, 1999.

[McQ71] McQueen, J., « Some methods for classification and analysis of multivariate observations », *Proc. Fifth Berkeley Sym. on Mathematicals Statistics and Probability*, p.281-297, 1967.

[Mil] Milner, B., Shao, X., Darch, J., « Fundamental Frequency and Voicing Prediction from MFCC's for Speech Reconstruction from Unconstrained Speech »,

[Mok02] Mokhtari, P., Campbell, N., « Automatic Measurement of Pressed/Breathy Phonation at Acoustic Centres of Reliability in Continuous Speech », *Special Issue on Speech Information Processing*, 2002.

[Mou95] Moulines, E., Laroche, J., « Techniques for pitch-scale and time-scale transformation of speech : part I, Non parametric methods », *Speech Communications*, vol.16, 1995.

[Opp68] Oppenheim, A., « Speech Analysis-Synthesis System Based on Homorphic Filtering », 1968.

[Pal95] Paliwal, K., « Interpolation properties of linear prediction parametric representations », *Proceedings of EUROSPEECH*, 1995.

[Pee01] Peeters, G., Modèles et modification du signal sonore adaptés à des caractéristiques locales, Thèse de doctorat de l'université Paris 6, 2001.

[Rab75] Rabiner, L., « Applications of a non-linear smoothing algorithm to speech processing », IEEE, 1975.

[Rab93] Rabiner, L., Juang, B.H., Fundamentals of speech recognition, Prentice-Hall International, London, 1993, 507p.

[Roe05a] Roebel, A, Rodet, X., « Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation » , *Conference on Digital Audio Effects*, 2005

[Roe05b] Roebel, A., Rodet, X., « Real time transposition with envelope preservation in the phase vocoder », *ICMC Barcelona*, p.672-675, 2005.

[Sch98] Schwarz, D., Spectral envelopes in sound analysis and synthesis, IRCAM / Universität Stuttgart Fakultät Informatik, rapport de stage, 1998.

[Sha] Shao, X., Milner, B., « MAP Prediction of Pitch from MFCC Vectors for Speech Reconstruction »,

[Sin] Sinclair, S., « Frequency Shifting with the Phase Vocoder »,

[Sty95] Stylianou, Y., Laroche, J., Moulines, E., « High-Quality Speech Modification based on Harmonic + Noise Model », EUROSPEECH, 1995.

[Sty96] Stylianou, Y., Baudoin, G., « On the transformation of the speech spectrum for voice conversion », *ICSLP*, 1996.

[Sty98] Stylianou, Y., Cappé, O., Moulines, E., « Continous Probabilistic Transform for Voice Conversion », *IEEE Transactions on speech and audio processing*, vol.6, no.2, 1998.

[Syr85] Syrdal, A., Steele, A., « Vowel F1 as a Function of Speaker Fundamental Frequency », 110th Meeting of JASA, vol.78, Fall 1985.

[Ver] Verhelst, W., Roelands, M., « An Overlap-Add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech »

[Vil06] Villavicencio, F., Roebel, A., Rodet, X., « Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation »,