# Modélisation sinusoïdale et applications à l'indexation audio

Michaël Betser

10/04/2008

*A mes parents,*
*A mes grands-parents*

# Remerciements

Je tiens à remercier tout d'abord Patrice Collen et Jean Bernard-Rault pour m'avoir donné l'opportunité de faire cette thèse au sein l'équipe Tech/Iris à France Télécom R&D, Rennes. Ils m'ont apporté un soutient précieux tout au long de mon travail parmis eux.

Je remercie également mes directeurs de thèse : Gaël Richard, pour sa disponibilité et son accueil lors de mes nombreuses visites à Télécom Paris, et Bertrand David, pour ses conseils avisés et mathématiquement très pointus.

Je remercie les membres du jury de m'avoir fait l'honneur de s'intéresser à mon travail.

Merci enfin à tous ceux qui m'ont aidés et soutenus pendant ces trois longues années : mes collègues de France Telecom, mes amis de Rennes et de Paris, ma famille et ma chère Sophie.

# Table des matières

**Table des matières** **v**

**Table des figures** **xi**

**Liste des tableaux** **xv**

**Acronymes** **xvii**

**Notations** **xix**

**1 Introduction générale** **1**

**I Analyse sinusoïdale** **7**

**2 Modélisation sinusoïdale** **9**
2.1 Les sinusoïdes généralisées . . . . . 10
2.2 Le modèle sinusoïdes+bruit . . . . . 11
2.3 Modèle sinusoïdal local . . . . . 11
2.3.1 Erreur de modélisation . . . . . 13

**3 Méthodes classiques d'estimation** **17**
3.1 Maximum de vraisemblance . . . . . 17
3.1.1 Principe général . . . . . 18
3.1.2 Estimateur ML pour un modèle polynomial en amplitude et en phase . . . . . 19
3.1.3 Estimateur ML pour un modèle log-polynomial en amplitude et polynomial en phase . . . . . 20
3.1.4 Estimateur ML pour une sinusoïde . . . . . 21
3.1.4.1 Modèles **M01** . . . . . 21
3.1.4.2 Modèle **MKQ** . . . . . 21
3.2 Méthodes à "haute résolution" . . . . . 22
3.3 Méthodes reposant sur la transformée de Fourier . . . . . 23
3.3.1 Vue d'ensemble d'un système d'analyse sinusoïdale basé sur la Transformée de Fourier à Court Terme (TFCT) . . . . . 23

3.3.2 Signal bien résolu par la transformée de Fourier . . . . . 24
3.3.3 Comparaison de la résolution de deux fenêtres de Fourier . . . 25
3.3.4 Ordre de complexité . . . . . 26
3.3.5 Transformée discrète/transformée continue . . . . . 27
3.3.6 Quelques rappels sur la transformée de Fourier . . . . . 27
3.3.6.1 Interpolation du spectre par zéro-padding . . . . . 27
3.3.6.2 Transformée de Fourier zéro-phase . . . . . 27
3.4 Protocole pour la comparaison expérimentale des estimateurs . . . . . 29
3.4.1 Mode opératoire . . . . . 29
3.4.2 Bornes de Cramer-Rao . . . . . 30

**4 Etat de l'art des méthodes d'analyse basées sur la TFCT 33**
4.1 Estimation pour le modèle **M01** . . . . . 33
4.1.1 Estimation de phase et d'amplitude . . . . . 34
4.1.2 Estimation de fréquence . . . . . 35
4.1.2.1 Le vocodeur de phase . . . . . 35
4.1.2.2 La Méthode de la dérivée . . . . . 37
4.1.2.3 Interpolation du spectre de Fourier discret . . . . . 39
4.1.2.4 Interpolation utilisant l'information de phase . . . . . 41
4.2 Estimation pour le modèle **M02** . . . . . 43
4.2.1 Estimation de fréquence . . . . . 43
4.2.1.1 Méthode du réassignement spectral . . . . . 43
4.2.2 Estimation de la variation de fréquence . . . . . 46
4.2.2.1 Estimateur de Röbel . . . . . 46
4.2.2.2 Estimateurs de Master . . . . . 47
4.2.2.3 Estimateur de Liu . . . . . 52
4.2.2.4 Estimateur de Marques et Almeida, Estimateur de Masri . . . . . 53
4.3 Estimation pour le modèle **M12** . . . . . 53
4.3.1 Méthode d'interpolation quadratique de la TF (QIFFT) . . . . 54
4.4 Comparaison expérimentale des estimateurs de l'état-de-l'art . . . . . 56
4.4.1 Estimateurs de fréquence . . . . . 57
4.4.2 Estimateurs d'amplitude et de phase . . . . . 60
4.4.3 Estimateurs de la variation de fréquence . . . . . 61
4.4.4 Estimateurs de la variation d'amplitude . . . . . 62
4.5 Conclusion . . . . . 63

**5 Etude critique des méthodes d'estimation reposant sur la TFCT 65**
5.1 Analyse des estimateurs de fréquence basés sur la TFCT . . . . . 65
5.1.1 Estimation bien résolue par la transformée de Fourier . . . . . 66
5.1.2 Principes généraux . . . . . 66
5.1.2.1 Estimation des paramètres d'ordre supérieur . . . . . 68
5.2 Déterminer un estimateur à partir d'une combinaison particulière de bins . . . . . 68
5.2.1 En utilisant des approximations de Taylor . . . . . 68

5.2.2 En utilisant des modélisations . . . . . 69
5.3 Le temps d'estimation . . . . . 71
5.3.1 Temps d'estimation et bornes de Cramer-Rao . . . . . 71
5.3.2 Le temps d'estimation du réassignement . . . . . 73
5.4 Quelques éléments de réflexion sur les propriétés statistiques des estimateurs . . . . . 74
5.5 Conclusion . . . . . 76

**6 Développement de nouveaux estimateurs 77**
6.1 Estimation de fréquence pour le modèle **M01** . . . . . 78
6.1.1 Approximations de Taylor . . . . . 78
6.1.1.1 Exemple d'estimateur à deux bins . . . . . 80
6.1.1.2 Propriétés statistiques . . . . . 80
6.1.1.3 Performances . . . . . 83
6.1.2 Algorithme Général pour les estimateurs modélisés . . . . . 84
6.1.2.1 Evaluation . . . . . 85
6.2 Estimation de fréquence pour les modèles **M02** et **M12** basée sur la méthode du vocodeur de phase . . . . . 87
6.2.1 Suivi du maximum (MB) . . . . . 89
6.2.2 Estimation du déroulement de phase . . . . . 91
6.2.3 Le vocodeur de phase par suivi des maximum (MB) . . . . . 92
6.2.4 Le Vocodeur à phase corrigées (PCV) . . . . . 92
6.2.5 Le vocodeur réassigné (RV) . . . . . 95
6.2.6 Propriétés statistiques des estimateurs . . . . . 97
6.2.7 Evaluation . . . . . 99
6.3 Estimation des paramètres d'ordre supérieur pour le modèle **M12** . . . 102
6.3.1 Estimation basée sur le réassignement . . . . . 104
6.3.2 Estimation basée sur le vocodeur de phase . . . . . 106
6.3.3 Evaluation . . . . . 107
6.4 Estimation d'amplitude et de phase . . . . . 108
6.4.1 Biais et variance . . . . . 110
6.4.2 Evaluation . . . . . 111
6.5 Conclusion . . . . . 112

**II Applications de la modélisation sinusoïdale à l'indexation audio 113**

**7 Estimation de fréquence fondamentale 115**
7.1 Introduction . . . . . 115
7.2 Etat de l'art . . . . . 116
7.2.1 Méthodes basées sur l'auto-similarité . . . . . 117
7.2.1.1 RAPT . . . . . 118
7.2.1.2 YIN . . . . . 119
7.2.2 Méthodes de type reconnaissance de forme . . . . . 120

7.2.2.1 La méthode de Doval et Rodet . . . 122
7.2.2.2 La méthode de Maher . . . 124
7.3 Algorithme de suivi de pitch basé sur la modélisation sinusoïdale (SPS) 126
7.3.1 Génération d'hypothèses . . . 126
7.3.2 Identification harmonique . . . 127
7.3.3 Mesure d'harmonicité . . . 128
7.3.3.1 Modification de la mesure de Maher et Beauchamp . . 128
7.3.3.2 Mesure probabiliste simplifiée . . . 130
7.3.4 Suivi de fréquence fondamentale . . . 132
7.4 Comparaison avec des méthodes existantes . . . 135
7.4.1 Protocole d'évaluation . . . 136
7.4.2 Choix de la méthode d'estimation sinusoïdale . . . 138
7.4.3 Résultats et commentaires . . . 138
7.5 Conclusion . . . 141

**8 Identification audio 143**
8.1 Introduction . . . 143
8.2 Contraintes de l'audio ID . . . 144
8.2.1 Les dégradations du signal . . . 145
8.2.2 Les contraintes concernant la mise en oeuvre . . . 145
8.3 Etat de l'art . . . 146
8.3.1 Méthode de Pinquier . . . 148
8.3.1.1 Formation de l'empreinte . . . 148
8.3.1.2 Comparaison . . . 148
8.3.1.3 Décision . . . 149
8.3.2 Méthode de Philips . . . 150
8.3.2.1 Formation de l'empreinte . . . 150
8.3.2.2 Comparaison et décision . . . 151
8.3.2.3 Réduction de l'espace de recherche . . . 151
8.3.3 Méthode de Shazam . . . 153
8.3.3.1 Formation de l'empreinte . . . 153
8.3.3.2 Comparaison et décision . . . 155
8.3.4 Discussion . . . 155
8.4 Description d'un système d'empreinte basé sur la modélisation sinusoïdale . . . 156
8.4.1 Création de l'empreinte lors de la phase d'entraînement . . . 156
8.4.1.1 Présélection . . . 157
8.4.1.2 Analyse sinusoïdale . . . 157
8.4.1.3 Sélection des composantes pertinentes et création de la signature . . . 157
8.4.1.4 Compaction de la signature . . . 159
8.4.2 Création de l'empreinte lors de la classification . . . 159
8.4.3 Module de comparaison des empreintes . . . 160
8.4.4 Accélération de la comparaison . . . 160
8.4.4.1 Comparaison par blocs de trames et d'objets . . . 161

8.4.4.2 Table d'index des trames . . . 164
8.4.5 Aspects calculatoires . . . 165
8.4.5.1 Mémoire . . . 165
8.4.5.2 Complexité . . . 166
8.5 Evaluation expérimentale . . . 167
8.5.1 Identification de jingles . . . 167
8.5.2 Identification de morceaux de musique . . . 169
8.6 Conclusion . . . 170

**9 Conclusion générale et perspectives 173**

**Appendices 177**

**A Propriétés de la transformée de Fourier 179**

**B Propriétés des fenêtres d'analyse 181**
B.1 Relation entre les fenêtres rectangulaire, triangulaire et les fenêtres de Blackman et de Hann . . . 181
B.2 Deux propriétés de la fenêtre Gaussienne . . . 182
B.3 Propriétés asymptotiques des fenêtres . . . 183

**C Bornes sur l'erreur de modélisation 187**
C.1 Modèle polynomial en amplitude . . . 187
C.2 Modèle log-polynomial en amplitude . . . 188
C.3 Erreur sur la TF du signal . . . 189

**D Dérivation de la variance de l'estimateur** (6.1.7) **191**

**E Seuil adaptatif pour la sélection de pic sinusoïdaux 193**

**F Comparaison des complexités des représentations temporelle et fréquentielle 195**

**G Deux articles de l'auteur 199**

**Bibliographie de l'auteur 229**

**Bibliographie 231**

# Table des figures


2.1 Variation maximale tolérée en fonction de l'ordre de modélisation . . . . . 14
2.2 Modélisation d'une modulation lente d'amplitude ou de fréquence . . . . . 15

3.1 Méthode d'analyse basée sur la FFT . . . . . 23
3.2 Réponse fréquentielle d'un chirp linéaire fortement modulé . . . . . 25
3.3 Résolution fréquentielle à -3dB . . . . . 26
3.4 Spectre d'amplitude et de phase d'une transformée de Fourier à phase nulle 28

4.1 Erreur des estimateurs arccos, arcsin et arctan en fonction de la fréquence. 38
4.2 Spectre de la fenêtre à réponse triangulaire . . . . . 40
4.3 Erreur des estimateurs d'interpolation en fonction la fréquence. . . . . 41
4.4 Erreur du réassignement en fonction de la variation de fréquence. . . . . 45
4.5 Erreur de l'estimateur de Röbel en fonction de la variation de fréquence. Num et 1/Den sont respectivement le numérateur et l'inverse du dénominateur de l'estimateur (4.2.9) . . . . . 46
4.6 Erreur des estimateurs de Master dérivés des intégrales de Fresnel en fonction de la variation de fréquence . . . . . 48
4.7 Erreur de l'estimateur de Master dérivé de l'approximation de Taylor en fonction de la variation de fréquence . . . . . 50
4.8 Erreur de l'estimateur de Liu en fonction de la variation de fréquence . . . 52
4.9 Spectre de la fenêtre Gaussienne pour de fortes modulations . . . . . 55
4.10 Comparaison de la fenêtre de Hann et de la fenêtre Gaussienne . . . . . 56
4.11 Evaluation des estimateurs classiques de fréquence . . . . . 58
4.12 Evaluation des estimateurs classiques d'amplitude et de phase . . . . . 60
4.13 Evaluation des estimateurs classiques de variation de fréquence. 'QIS' correspond à la QIFFT courte et 'QIL' à la QIFFT longue. . . . . 61
4.14 Evaluation des estimateurs classiques de variation d'amplitude . . . . . 62

5.1 Résumé de la méthode de modélisation des estimateurs . . . . . 70
5.2 Evolution des CRBs pour les paramètres du modèle **M12** en fonction du temps d'estimation . . . . . 72
5.3 Le temps du réassignement varie suivant la fréquence de la TFCT . . . . . 73

6.1 Choix des bins dans la méthode d'estimation de fréquence basée sur le modèle **M01**. Ici $X_i = X(t, \omega_i; h)$ et $\omega_i$ est la fréquence du bin $k_i$. . . . . 78

6.2 Variance théorique et expérimentale de l'estimateur (6.1.8) . . . . . 83
6.3 Performances de l'estimateur (6.1.8) . . . . . 84
6.4 Comparaison de l'estimateur (6.1.8) et de l'estimateur modélisé . . . . . 86
6.5 Performances de l'estimateur (6.1.20) sans ajout de bruit . . . . . 87
6.6 Deux exemples de fonction à modéliser . . . . . 87
6.7 Transformation en phase quadratique discrète . . . . . 89
6.8 Schéma de fonctionnement du Vocodeur à Phases Corrigées (PCV) . . . . 93
6.9 Schéma de fonctionnement du Vocodeur Réassigné (RV) . . . . . 96
6.10 Comparaison de la variance théorique et expérimentale pour la méthode RV . . . . . 98
6.11 Choix de la résolution fréquentielle . . . . . 99
6.12 Comparaison des méthodes PCV, RV, réassignement et QIFFT pour le modèle **M02** . . . . . 102
6.13 Comparaison des méthodes RV, réassignement et QIFFT pour le modèle **M12** . . . . . 103
6.14 Bins utilisés par les méthodes complètes d'estimation des paramètres du modèle **M12** . . . . . 105
6.15 Evaluation des méthodes d'estimation des paramètres d'ordre supérieur pour le modèle **M12** . . . . . 107
6.16 Evaluation de la méthode d'estimation de l'amplitude et de la phase pour le modèle **M12** . . . . . 111

7.1 Principe de l'auto-corrélation . . . . . 117
7.2 Schéma de fonctionnement des méthodes basées sur l'auto-similarité . . . 118
7.3 Schéma de fonctionnement de l'algorithme RAPT . . . . . 119
7.4 Schéma de fonctionnement des méthodes basées sur la reconnaissance de forme . . . . . 120
7.5 Schéma de fonctionnement de la méthode d'estimation de pitch sinusoïdale 121
7.6 Identification des fréquences dans la méthode de Doval et Rodet . . . . . 122
7.7 Identification des fréquences dans la méthode de Maher et Beauchamp . . 125
7.8 Fonctionnement de l'estimation du meilleur pitch dans la méthode sinusoïdale . . . . . 129
7.9 Exemple de modélisation pour la mesure d'harmonicité probabiliste . . . . 131
7.10 Fonctionnement de l'algorithme de suivi de pitch . . . . . 134
7.11 Exemple de spectre de laryngographe . . . . . 137
7.12 Performance des mesures d'harmonicité en fonction du seuil sur la vraisemblance . . . . . 139

8.1 Identification audio basé sur l'extraction d'empreintes . . . . . 146
8.2 Principe de l'analyse en sous-bandes . . . . . 148
8.3 Comparaison bloc par bloc, trame à trame . . . . . 149
8.4 Décision par double seuil . . . . . 150
8.5 Extraction de l'empreinte de Philips . . . . . 151
8.6 Exemple d'empreinte de Philips . . . . . 152
8.7 Réduction de l'espace de recherche par une table de look-up . . . . . 153

8.8 Sélection des points d'intérêt dans la méthode de Shazam . . . . . 154
8.9 Extraction de l'empreinte sinusoïdale . . . . . 156
8.10 Extrait d'une empreinte sinusoïdale . . . . . 158
8.11 Comparaison des empreintes sinusoïdales . . . . . 159
8.12 Représentations temporelles et fréquentielles . . . . . 161
8.13 Représentation fréquentielle de l'empreinte de référence et de l'empreinte à identifier . . . . . 162
8.14 Table utilisée pour l'accumulation des vraisemblances . . . . . 163
8.15 Création des clés sinusoïdales, pour la réduction de l'espace de recherche . 164
8.16 Identification des morceaux de musique en temps réel . . . . . 170

# Liste des tableaux

3.1 CRBs pour les modèles **M01**, **M11**, **M02** et **M12** . . . . . 32

6.1 Exemples de combinaison de bins donnant des fenêtres symétriques ou antisymétriques . . . . . 79
6.2 Valeurs des bornes pour différentes fenêtres d'analyse. . . . . 81
6.3 Valeur maximale des pas d'avancement pour le modèle **M02** . . . . . 92

7.1 Comparaison des deux mesures d'harmonicité sans seuillage . . . . . 139
7.2 Comparaison des algorithmes d'estimation de fréquence fondamentale . . 140

8.1 Description du corpus de test utilisé pour l'identification de jingles . . . . 167
8.2 Comparaison des paramètres . . . . . 168
8.3 Rappel en terme d'occurrence . . . . . 169
8.4 Rappel en terme de durée . . . . . 169

A.1 Propriétés de la transformée de Fourier continue . . . . . 179

# Acronymes

| | |
|---|---|
| **TF** | Transformée de Fourier |
| **TFC** | Transformée de Fourier Continue |
| **FFT** | Transformée de Fourier Rapide |
| **TFCT** | Transformée de Fourier à Court Terme |
| **ML** | Maximum de Vraisemblance |
| **ESPRIT** | Estimation of Signal Parameters via Rotational Invariance Techniques |
| **MUSIC** | MUltiple SIgnal Classification |
| **HR** | Haute Résolution |
| **AM** | Modulation d'Amplitude |
| **FM** | Modulation de Fréquence |
| **ESM** | Exponential Sinusoidal Model |
| **QIFFT** | Quadratically Interpolated FFT |
| **RV** | Vocodeur Réassigné |
| **PCV** | Vocodeur à Phases Corrigées |
| **MSE** | Mean Squared Error |
| **SNR** | Signal to Noise Ratio |
| **CRB** | Cramer-Rao Bound |
| **dB** | décibels |
| **NCCF** | Normalized Cross-Correlation Function |
| **YIN** | algorithme de suivi de pitch de [de Cheveigné and Kawahara, 2002] |
| **SPS** | algorithme de Suivi de Pitch Sinusoidal |
| **GER** | Gross Error Rate |
| **BER** | Bit Error Rate |
| **RH** | Rappel Harmonique |
| **PH** | Précision Harmonique |

# Notations

**Opérateurs**

| | |
|---|---|
| $\bar{f}$ | Conjugué complexe de $f$ |
| $\hat{f}$ | Estimateur de $f$ |
| $f^{(k)}$ | Dérivée d'ordre $k$ de $f$ |
| $f^H$ | Transposée hermitienne |
| $\Im()$ | Opérateur partie imaginaire |
| $E()$ | Espérance mathématique |
| $O()$ | Ordre de complexité |
| $P()$ | Opérateur de probabilité |
| $\Re()$ | Opérateur partie réelle |
| arg() | Argument principal |
| Card() | Cardinal |
| mod() | Modulo $2\pi$ |
| sgn() | Signe |
| var() | Variance |

**Modélisation**

| | |
|---|---|
| $A(t)$ | Fonction d'amplitude |
| $A$ | Amplitude initiale |
| $A_i(t)$ | Fonction d'amplitude de la sinusoïde numéro $i$ |
| $A_{i,k}$ | Coefficient du développement de Taylor d'ordre $k$ de $A_i(t)$ |
| $L(t)$ | Fonction d'amplitude logarithmique, $L(t) = \log(A(t))$ |
| $\Omega(t)$ | Fonction de fréquence, $\Omega(t) = \frac{\partial \Phi}{\partial t}(t)$ |

| | |
|---|---|
| $\Phi(t)$ | Fonction de phase |
| $\Phi$ | Phase initiale |
| $\Phi_i(t)$ | Fonction de phase de la sinusoïde numéro $i$ |
| $\Phi_{i,k}$ | Coefficient du développement de Taylor d'ordre $k$ de $\Phi_i(t)$ |
| $\alpha$ | Paramètre de phase initial |
| $\alpha_i$ | Phase de la sinusoïde à instant $t_i$ : $\alpha_i = \Phi(t_i)$ |
| $\beta$ | Paramètre de fréquence initiale |
| $\beta_i$ | fréquence de la sinusoïde à l'instant $t_i$ : $\beta_i = \Phi^{(1)}(t_i)$ |
| $\gamma$ | Paramètre de variation de fréquence. Si $t$ est le temps d'analyse, $\gamma = \Phi^{(2)}(t)$ |
| $\lambda$ | Paramètre de log-amplitude initiale |
| $\lambda_i$ | Log-amplitude à l'instant $t_i$ : $\lambda_i = L(t_i)$ |
| $\mu$ | Paramètre de variation de log-amplitude. Si $t$ est le temps d'analyse, $\mu = L^{(1)}(t)$ |
| $\epsilon_\Phi$ | Erreur de modélisation sur la fonction $\Phi(t)$ |
| $\epsilon_A$ | Erreur de modélisation sur la fonction $A(t)$ |
| $\epsilon_L$ | Erreur de modélisation sur la fonction $L(t)$ |

**Estimation Sinusoïdale**

| | |
|---|---|
| $F$ | Fréquence d'échantillonnage |
| $K$ | Ordre de modélisation de la fonction de log-amplitude $L$ |
| $M$ | Index du milieu de la fenêtre d'analyse, $M = (N-1)/2$ |
| $N$ | Taille de la fenêtre d'analyse de Fourier |
| $P$ | Taille de la transformée de Fourier, $P \geq N$ |
| $Q$ | Ordre de modélisation de la fonction de phase $\Phi$ |
| $T$ | pas d'avancement en secondes entre deux fenêtres d'analyse consécutives |
| $W$ | Nombre total d'échantillons utilisés pour l'estimation, $W \geq N$ |
| $\omega$ | Pulsation ou, par abus de langage, fréquence |
| $\omega_{k_i}$ | Fréquence du bin $k_i$ : $\omega_k = 2\pi k_i F/P$. Lorsqu'il n'y a pas d'ambiguité le $k$ est omis : $\omega_i = \omega_{k_i}$ |
| $t$ | Temps de l'analyse, aussi appelé temps global, ou temps de la transformée de Fourier |

$t_{m_i}$ Temps correspondant à l'index $m_i$ : $t_{m_i} = m_i/F$. Lorsqu'il n'y a pas d'ambiguité le $m$ est omis : $t_i = t_{m_i}$

$\tau$ Temps dans la fenêtre d'analyse, aussi appelé temps local

$\tau_n$ Temps local correspondant à l'index $n$ : $\tau_n = n/F$

$\Delta\beta_i$ Différence entre la fréquence de la sinusoïde au temps $t_{m_i}$ et la fréquence $\omega_{k_i}$

$h(\tau)$ Fenêtre d'analyse

$\dot{h}(\tau)$ Dérivée de $h(\tau)$ par rapport au temps $t$

$x(t)$ Signal analysé

$n(t)$ Bruit aléatoire blanc Gaussien centré

$s(t)$ Signal bruité, $s(t) = x(t) + n(t)$

$H(\omega)$ Transformée de Fourier de la fenêtre $h$, pour la fréquence $\omega$

$\Gamma(\mu, \beta, \gamma; h)$ Transformation polynomiale de $h$, voir équation (6.2.3), page 86

$\Gamma(\beta; h)$ Transformée de Fourier de $h$, $\Gamma(\beta; h) = \Gamma(0, \beta, 0; h) = H(\omega)$

$\Gamma(h)$ Somme des éléments de $h$, $\Gamma(h) = \Gamma(0, 0, 0; h) = H(0)$

$X(\omega)$ Si $t$ et $h$ ne sont pas informatifs, on note $X(\omega) = X(t, \omega; h)$

$X(t, \omega; h)$ Transformée de Fourier à court terme du signal $x(t)$

$X_c(t, \omega; h)$ Transformée de Fourier à court terme continue de $x(t)$

$X_i$ Raccourci pour $X(t_i, \omega_i; h_i)$

$\Delta X$ Différence de phase entre deux TFCT $X_1$ et $X_2$ : $\Delta X = \arg(X_2 \bar{X}_1)$

$\mathcal{H}$ Combinaison de bins

$\delta$ Différence de fréquence. Dans les cas ambigus, on distingue $\delta_t$, une différence temporelle et $\delta_\omega$ une différence fréquentielle

$\epsilon_D$ Erreur déterministe

$\epsilon_N$ Erreur stochastique

$\eta$ Rapport signal à bruit

$\sigma^2$ Variance du bruit $n$

$\theta$ Vecteur de paramètres à estimer

### Suivi de Pitch

$p$ Fréquence fondamentale

$A_i$ Amplitude de l'harmonique numéro $i$

$f_i$ Fréquence de l'harmonique numéro $i$. $f_i = i.p$

| | |
|---|---|
| $A_m$ | Amplitude maximale parmi les harmoniques hypothèses |
| $F_{max}$ | Fréquence supérieure de l'intervalle de recherche du pitch |
| $F_{min}$ | Fréquence inférieure de l'intervalle de recherche du pitch |
| $\tilde{K}$ | Nombre de pics estimés |
| $K$ | Nombre d'harmoniques |
| $K_{max}$ | Nombre d'harmoniques maximum. $K \leq K_{max}$ |
| $M$ | Nombre d'harmoniques manquantes parmi les $K$ harmoniques du pitch |
| $P$ | Nombre d'harmoniques présentes parmi les $K$ harmoniques du pitch |
| $Q$ | Ensemble des pics du peigne harmonique hypothèse |
| $q_i$ | Harmonique numéro $i$ |
| $\sigma$ | Variance de la la densité de probabilité de fréquence |
| $\tilde{A}_b$ | Amplitude du bruit estimée |
| $\tilde{A}_i$ | Amplitude du pic estimé numéro $i$ |
| $\tilde{f}_i$ | Fréquence du pic estimé numéro $i$ |
| $\tilde{A}_m$ | Amplitude maximale parmi les pics estimés |
| $T_f$ | Tolérance fréquentielle pour l'identification des pics |
| $\tilde{Q}$ | Ensemble des pics sinusoïdaux estimés |
| $\tilde{q}_i$ | Pic estimé numéro $i$ |

**Identification Audio**

| | |
|---|---|
| $B$ | Nombre de bits utilisés pour coder les fréquences |
| $B_{j,n}$ | Sous-bloc de l'objet de référence $j$, de taille $N_t^{'}$, commençant à l'index $n$ |
| $B^{'}$ | Bloc à identifier |
| $D_p$ | Densité de pics dans l'objet de référence, en pics par secondes |
| $D_p^{'}$ | Densité de pics dans le bloc à identifier, en pics par secondes |
| $F_c$ | Fréquence de coupure |
| $B_s$ | Nombre de bits utilisés pour coder une fréquence dans une clé. $B_s \leq B$ |
| $K_p$ | Nombre de pics présents dans le bloc de taille $T_s$ |
| $K_s$ | Nombre de fréquences utilisées pour former une clé |
| $M$ | Nombre de pics retenus dans le bloc de référence de taille $T_b$ |
| $N_b$ | Nombre de sous-bandes |

| | |
|---|---|
| $N_t$ | Nombre de trames de l'objet de référence |
| $N_t^{'}$ | Nombre de trames dans le bloc à identifier |
| $T_b$ | Taille de l'objet de référence, en secondes |
| $T_d$ | Taille en seconde de l'intervalle d'analyse pour la décision |
| $T_h$ | Décalage entre deux trames, en secondes |
| $T_s$ | Taille du bloc utilisé pour former une clé |
| $T_b^{'}$ | Taille du bloc à identifier, en secondes |

CHAPITRE 1

# Introduction générale

Dans le domaine du traitement du son, l'analyse sinusoïdale est dès l'origine utilisée pour la transformation et la génération des sons. C'est une application du théorème de Fourier qui montre que tout signal périodique peut être modélisé par une somme de sinusoïdes avec différentes fréquences et amplitudes. Les premiers systèmes exploitant ce principe, comme le vocodeur de phase, ont été développés vers la fin des années soixante [Flanagan and Golden, 1966] pour modéliser les signaux harmoniques. Dans les années soixante-dix, l'apparition de l'analyse numérique à permis le développement d'algorithmes rapides pour ces systèmes, notamment ceux reposant sur l'algorithme de la Transformée de Fourier Rapide (FFT). Toujours basées sur la modélisation sinusoïdale, de nouvelles méthodes de synthèse additive [McAulay and Quatieri, 1986] sont apparues, ainsi que des schéma complets d'analyse, de transformation et de synthèse des signaux audio [Serra, 1989]. Les applications visées dans cette thèse concernant l'indexation audio, seule la partie analyse sinusoïdale va ici nous intéresser. Toutes ces méthodes sont basées sur la transformée de Fourier et son implantation rapide, la FFT. Elles sont aujourd'hui toujours utilisées grâce à leur facilité de mise en oeuvre et à leur souplesse d'utilisation dans des domaines aussi divers que le codage, l'analyse musicale, l'analyse de parole etc.

## Le modèle sinusoïdal

Le modèle sinusoïdal est très adapté à la modélisation des signaux harmoniques, c'est à dire de signaux dont les fréquences sont des multiples de la fréquence fondamentale, et pour les superpositions de tels signaux dans le cas général. En effet dans ce cas particulier, un faible nombre de sinusoïdes est requis pour représenter ces signaux. Par contre dans le cas de bruits aléatoires par exemple, cette représentation, bien que toujours applicable, devient beaucoup moins adaptée, car un grand nombre de composantes sinusoïdales est alors requis. C'est pourquoi la modélisation a été perfectionnée, conduisant au modèle sinusoïdes+bruit [Serra, 1989]. D'autres modèles ont également été proposés afin de prendre en compte la non-stationnarité

des signaux, en particulier les signaux fortement variables ou transitoires, comme le modèle sinusoïdes+transitoires+bruit [Verma and Meng, 1998; Levine and Smith, 1998]. Ce modèle reste cependant un cas particulier du modèle sinusoïde+bruit, car les transitoires sont également des sinusoïdes, mais dont les paramètres varient très rapidement. Le modèle sinusoïdes+bruit reste ainsi le modèle à la fois le plus simple et le plus général pour représenter un signal.

Le nombre total de paramètres à estimer dépend du modèle utilisé pour représenter les sinusoïdes. Le modèle le plus simple et le plus utilisé consiste à considérer des sinusoïdes dont la fréquence et l'amplitude sont localement constantes. D'autre modèles plus précis, mais avec plus de paramètres peuvent être utilisés, pouvant améliorer la représentation des signaux fortement variables, comme les transitoires [Hermus et al., 2002], [Boyer and Abed-Meraim, 2004].

## Estimation des paramètres sinusoïdaux

L'estimation des paramètres sinusoïdaux est l'un des thèmes les plus étudiés dans le domaine du traitement du signal et de nombreuses approches ont été proposées pour résoudre ce problème. Un bon nombre d'entre elles s'attachent à améliorer l'analyse basée sur la Transformée de Fourier (TF), comme les estimateurs de fréquence basés sur le vocodeur de phase [Portnoff, 1981], [Rife and Boorstyn, 1976], sur des techniques d'interpolation [Abe and Smith III, 2004], [Quinn, 1994], [Macleod, 1998], [Betser et al., 2006a], ou le réassignement spectral [Auger and Flandrin, 1995].

D'autres méthodes existent basées sur l'analyse des moindres carrés non-linéaire [Stoica and Nehorai, 1988], [Choi, 1997] ou sur les méthodes dites de sous-espace, comme ESPRIT [Roy et al., 1986] et MUSIC [Schmidt, 1986]. Le principal problème lié aux méthodes basées sur la TF est la limitation de leur résolution, ce que les autres méthodes s'efforcent de contourner. Nous verrons cependant que la limitation de la résolution n'est plus un défaut, lorsque l'on utilise des modèles plus complexes, plus proches des signaux que l'on est amené à rencontrer.

## Application de la modélisation sinusoïdale à l'indexation audio

Après l'étude du modèle sinusoïdal nous nous attacherons à mettre en oeuvre des algorithmes dédiés à l'indexation audio et strictement basés sur les paramètres sinusoïdaux. Deux tâches ont été abordées au cours de cette thèse, le suivi de fréquence fondamentale dans le cas monophonique et l'identification audio.

### Suivi de fréquence fondamentale dans le cas monophonique

Le suivi de fondamentale, ou suivi de pitch a été choisi car c'est un domaine très riche, qui a déjà une longue histoire. Le pitch est par définition la fréquence la plus basse d'un signal harmonique pur. L'importance de l'estimation du pitch vient du fait que beaucoup de sons de notre environnement sont harmoniques ou presque

harmoniques et que la donnée du pitch permet de déterminer presque complètement ce type de signaux. De nombreuses méthodes d'estimation ont été développées, basées notamment sur la fonction d'auto-corrélation [Rabiner and Juang, 1993], ou sur les paramètres sinusoïdaux [Doval and Rodet, 1991]. Le suivi de pitch nous a en fait permis une première approche de la problématique de l'indexation audio basée sur les paramètres sinusoïdaux.

### Identification audio

La deuxième tâche abordée au cours de cette thèse est l'indentification audio ou audio ID. Elle consiste à essayer de retrouver un document audio, ou des informations concernant ce document audio, à partir d'un fragment sonore non identifié (requête audio). Des déformations ont pu altérer ce document lors de la requête, donc l'audio ID consiste en fait à retrouver un document identique, tout en étant robuste à aux différentes altérations possibles. L'audio ID est une tâche dont l'intérêt industriel et les applications sont récentes et en plein essor, à cause de la multiplication des média et de leurs contenus. Un certain nombre de méthodes dédiée à l'identification audio à grande échelle existent déjà. Ces méthodes ainsi que celle que l'on a développé sont basées sur l'extraction d'une "empreinte" audio. L'utilisation des paramètres sinusoïdaux pour former cette empreinte est par contre presque inexistante, le plus proche exemple étant celui de l'algorithme de Shazam [Wang, 2003]. Le domaine est donc relativement propice au développement de nouvelles méthodes.

## Principaux résultats apportés dans le cadre de la thèse

Le premier apport de la thèse est une discussion théorique sur les modèles utilisés pour l'estimation de paramètres sinusoïdaux. Nous avons distingué un modèle global et un modèle local, qui est une approximation du modèle global. Nous avons introduit une représentation du modèle local par développement de Taylor en amplitude et en phase du modèle global. Nous avons montré qu'un développement d'ordre faible, qui est le modèle employé le plus souvent, ne permet pas de décrire convenablement les signaux audio.

Parmi les méthodes d'estimation existantes, notre travail s'est concentré sur une famille d'entre elle, les méthodes basées sur la transformée de Fourier. La plupart des méthodes de cette famille sont basées sur un modèle sinusoïdal simple. Nous avons présenté sous un même formalisme un très grand nombre de ces méthodes, ce qui a permis de mettre en évidence des liens très forts entre certaines d'entre elles, principalement pour les méthodes d'estimation de fréquence basées sur la phase. Nous avons également implanté et réalisé une comparaison complète de ce type de méthodes. Nous avons notamment évalué la méthode d'interpolation quadratique de la FFT (Quadratically Interpolated FFT (QIFFT) en anglais) et les méthodes d'estimation de variation de fréquence, méthodes n'ayant jamais été comparées aux autres estimateurs.

Nous avons ensuite mis en évidence des mécanismes communs à tous les estimateurs basés sur la transformée de Fourier. A partir de ces constatations, nous

avons développé un certain nombre d'algorithmes pour améliorer les performances des estimateurs, notamment les estimateurs basés sur des modèles plus complexes, permettant d'être plus robuste à des modulations du signal.

Dans le cas des estimateurs de fréquence basés sur un modèle sinusoïdal simple, sans modulation, nous avons mis au point une méthode pour réduire de façon arbitraire le biais d'une méthode existante. Nous avons également développé une méthode pour trouver des estimateurs du type "interpolateur de spectre utilisant la phase" [Macleod, 1998], utilisable avec n'importe quel type de fenêtre. Ces estimateurs étaient auparavant limités à la fenêtre rectangulaire.

Nous nous sommes intéressés ensuite à des méthodes robustes à des modulations à la fois d'amplitude et de fréquence. Nous avons développé deux méthodes alternatives au réassignement pour l'estimation de fréquence, moins complexes et donnant des performances similaires dans le cas d'une modulation d'ordre 2 en phase et d'une amplitude constante. Dans le cas d'une modulation d'amplitude d'ordre 1 et d'une modulation de phase d'ordre 2, nous avons proposés deux schémas de calcul complets des paramètres, à la fois moins complexes et plus performants que la QIFFT [Abe and Smith III, 2004], qui est la seule méthode équivalente déjà existante.

Nous avons également esquissé un algorithme général d'inversion des fonctions, qui éviterait d'avoir recours à une formule analytique approchée. Cette méthode constituerait une alternative à l'optimisation multidimensionnelle des méthodes du type maximum de vraisemblance, pour l'estimation des paramètres des modèles. Dans le cas général cet algorithme semble cependant délicat à mettre en oeuvre. Enfin nous avons présenté une méthode alternative au développement asymptotique pour dériver les variances des estimateurs. Cette méthode est particulièrement utile pour le cas des modèles qui varient en amplitude et qui ne sont pas bornés asymptotiquement.

Dans la deuxième partie, nous nous sommes attachés à développer des algorithmes d'indexation audio basés uniquement sur ces paramètres sinusoïdaux. Deux tâches ont été abordées, le suivi de pitch et l'identification audio.

L'estimation de pitch basé sur l'analyse sinusoïdale s'inspire des travaux de [Maher and Beauchamp, 1994] et [Doval and Rodet, 1993]. Un nouvel algorithme de suivi de pitch a également été développé. L'algorithme complet donne de bons résultats par rapport à l'état de l'art en termes d'erreur sur le pitch et de détection des zones harmoniques. La méthode repose sur deux principes qui peuvent être réutilisés pour d'autres tâches d'indexation audio : une identification des pics sinusoïdaux présents, qui cherche à associer par couple les pics du signal à identifier et les pics du peigne harmonique hypothèse, et une mesure de similarité sur ces couples.

Le dernier apport de cette thèse est un algorithme d'identification audio basé sur une empreinte sinusoïdale et utilisant ces deux principes. Ici l'identification des pics se fait entre le signal à identifier et un signal de référence. L'empreinte sinusoïdale permet de reconnaître de façon fiable des évènements sonores très courts de l'ordre de 1 s et d'être plus robuste que la plupart des empreintes utilisées couramment en audio ID. Enfin une méthode de réduction de l'espace de recherche par table de look-up a été présentée pour réduire la complexité de l'algorithme.

## Organisation du document et pistes de lectures

Le document est divisé en deux parties presque indépendantes. Les personnes que n'intéressent que les applications de la modélisation sinusoïdale, pourront se contenter d'un tour d'horizon des méthodes d'estimation sinusoïdale au chapitre 3, avant d'aborder la deuxième partie.

Pour les lecteurs intéressés uniquement par les méthodes d'estimation sinusoïdales basées sur la FFT, nous suggérons d'entamer la lecture par la section 3.3 avant d'attaquer l'état de l'art au Chapitre 4 et les développements réalisés au cours de cette thèse au Chapitre 6.

### Première partie : *Analyse sinusoïdale*

Nous avons rapidement évoqué la possibilité de faire intervenir des modèles plus précis pour modéliser les signaux audio. Le type de modélisation utilisé va en grande partie déterminer les limitations de performances des algorithmes utilisés. Ces limitations vont bien entendu dépendre des signaux étudiés. La modélisation étant en quelque sorte la base de l'estimation, le chapitre 2 lui sera consacré. En particulier nous allons expliquer pourquoi nous nous sommes intéressés à des modèles faisant intervenir des variations de fréquence et d'amplitude. Nous parlerons également des avantages et des inconvénients des modèles étudiés. Dans le chapitre 3, nous verrons les trois grands types de méthodes permettant d'estimer les paramètres des sinusoïdes, à savoir la méthode du maximum de vraisemblance, les méthodes de sous-espaces, aussi appelées méthodes Haute Résolution (HR), et les méthodes basées sur la transformée de Fourier. Nous évoquerons notamment les avantages et les inconvénients qui sont liés à chacune de ces méthodes. Le chapitre 4 est un état de l'art sur les estimateurs basés sur la transformée de Fourier, pour les différents modèles étudiés dans cette thèse. Nous réunirons sous un même formalisme des méthodes assez disparates, et nous présenterons des comparaisons détaillées de ces estimateurs, pour des signaux modulés en fréquence et en amplitude. Le chapitre 5 présentera une étude critique de ces estimateurs. Nous soulignerons les points communs qui unissent ces méthodes d'estimation et nous présenterons les méthodes utilisées dans la thèse pour dériver de nouveaux estimateurs, ainsi que leur propriétés statistiques. Enfin le chapitre 6 présentera les estimateurs développés dans le cadre de la thèse. Nous détaillerons notamment deux nouvelles méthodes complètes pour l'estimation des paramètres d'un modèle avec variation d'amplitude et de fréquence.

### Deuxième partie : *Applications de la modélisation sinusoïdale à l'indexation audio*

La deuxième partie est divisée en deux chapitres, chacun dédié à une tâche d'indexation : le suivi de fondamentale au Chapitre 7 et l'identification audio au Chapitre 8. Chaque chapitre est organisé de la même manière et peut se lire indépendamment. Une introduction replace le contexte spécifique à chaque tâche. Puis un état de l'art fait un rapide tour d'horizon des méthodes existantes avec une description plus

approfondie des méthodes qui serviront à la comparaison, et des méthodes proches, basées sur une analyse sinusoïdale. On détaille ensuite dans les deux cas une méthode originale basée uniquement sur les paramètres sinusoïdaux, que nous évaluons ensuite par rapport aux méthodes présentées dans l'état de l'art.

# Première partie

# Analyse sinusoïdale

CHAPITRE 2

# Modélisation sinusoïdale

Le son est un phénomène vibratoire, la vibration étant par définition un phénomène oscillant autour d'une position d'équilibre. Trois paramètres caractérisent une vibration élémentaire : l'écart maximal par rapport à la position d'équilibre, l'amplitude, une mesure de la rapidité de vibration, la pulsation, et enfin l'état initial de la vibration, la phase initiale. Le phénomène oscillatoire le plus simple est celui dont l'amplitude $A$ et la pulsation $\omega$ sont constants dans le temps et peut être complètement décrit par une fonction sinusoïdale :

$$x(t) = A\cos(\omega t + \Phi) \tag{2.0.1}$$

où $\Phi$ est la phase en $t = 0$.

Les vibrations les plus aisément observables dans la nature sont toutes réelles, comme l'est le son. Néanmoins même ces phénomènes peuvent être décrits de façon élémentaire comme une décomposition de sinusoïdes complexes. La plupart des systèmes d'analyse, et en particulier la transformée de Fourier, décomposent le signal en sinusoïdes complexes plutôt que réelles. C'est pourquoi dans la suite de cet ouvrage cette représentation sera préférée.

La représentation complexe associée à un signal réel est appelé représentation analytique. Un signal analytique est par définition un signal dont la transformée de Fourier est nulle pour les fréquences négatives. Pour passer du signal réel (modélisation réelle) au signal analytique (modélisation complexe), il y a plusieurs solutions. La plus simple consiste à utiliser la propriété de symétrie hermitienne de spectres des signaux réels. On met toutes les fréquences négatives à zéro, et on multiplie par deux les valeurs des fréquences positives. Une autre solution est d'utiliser la transformée de Hilbert[1].

Dans la nature, les phénomènes vibratoires ne sont jamais parfaitement stationnaires, les paramètres vont évoluer au cours du temps. Ils peuvent aussi subir des perturbations dont il faut tenir compte pour représenter le signal. Le modèle sinusoïdal simple est donc généralement insuffisant, c'est pourquoi on introduit le modèle

[1] Voir par exemple la référence [Kunt, 1999] pour plus de détails.

sinusoïdal généralisé à la section 2.1, dont les paramètres peuvent varier dans le temps, et le modèle sinusoïdes+bruit qui tient compte de la partie non-déterministe à la section 2.2. Le modèle sinusoïdal généralisé ne donnant pas une représentation directement manipulable, on l'approche localement par un modèle paramétrique plus simple décrit à la section 2.3.

## 2.1 Les sinusoïdes généralisées

Pour décrire des phénomènes vibratoires non stationnaires, on a recours à une forme généralisée de la fonction sinusoïdale, où l'amplitude et la pulsation peuvent varier avec le temps. Une sinusoïde de ce type est souvent appelée "oscillateur" ou "partiel".

$$x(t) = A(t)\exp(j\Phi(t)) \tag{2.1.1}$$

où $A$ et $\Phi$ sont maintenant des fonctions $C^\infty$ par rapport au temps. La fonction d'amplitude doit être positive pour être consistante avec la définition d'une sinusoïde réelle. Soit $f_e$ la fréquence minimale du support fréquentiel du spectre de $\exp(j\Phi(t))$. Pour que $x$ soit analytique, le support fréquentiel du spectre de $A(t)$ doit être compris entre $[-f_e, f_e]$, c'est à dire que $A(t)$ doit être une fonction basse fréquence et $\exp(j\Phi(t))$ une fonction haute fréquence. $x$ est alors le signal analytique de $A(t)\cos(\Phi(t))$ [Picinbono, 1997]. Dans beaucoup de cas de figure étudiés par la suite, $x$ ne sera qu'approximativement analytique, car il y aura souvent un chevauchement des supports fréquentiels de $A(t)$ et $\exp(j\Phi(t))$, mais on supposera cette erreur négligeable. Enfin on suppose que les fonctions d'amplitude et de phase sont lentement variables dans le temps, c'est à dire que sur un intervalle de temps suffisamment petit, ces fonctions vont pouvoir être représentées par des modèles plus simples, paramétriques, que l'on va appeler modèles locaux par la suite.

Certains auteurs préfèrent utiliser l'amplitude logarithmique $L(t) = \log(A(t))$ :

$$x(t) = \exp(L(t) + j\Phi(t)) \tag{2.1.2}$$

Cette représentation est particulièrement utilisée pour représenter des systèmes amortis libres ou des attaques de parole ou d'instruments.

Les sons naturels étant rarement des oscillateurs purs, ils vont plutôt être décrits comme un mélange d'oscillateurs. En effet, une grande partie des signaux musicaux, et de parole sont caractérisés par des vibrations, qui peuvent être considérées comme stationnaires à court terme. Cet aspect vibrant des signaux peut être efficacement modélisé par une somme de sinusoïdes dont les amplitudes et les phases peuvent évoluer dans le temps.

$$x(t) = \sum_{i=1}^{M} A_i(t)\exp(j\Phi_i(t)) \tag{2.1.3}$$

où $x$ est un mélange de $M$ partiels, les $\Phi_i$ sont les fonctions de phase et les $A_i$ les fonctions d'amplitude du partiel $i$.

Dans certains cas de figure cette représentation peut s'avérer ambigüe, car il peut y avoir plusieurs formulations acceptables du même phénomène. L'exemple classique

de cette ambiguïté est la modulation sinusoïdale d'amplitude [Zwicker and Feldtkeller, 1981] :

$$x_m(t) = A(1 + \cos(\omega_m t)) \exp(j\omega t) \tag{2.1.4}$$

où $A(t) = A(1 + \cos(\omega_m t))$ et $\Phi(t) = \omega t$. Il s'agit d'une sinusoïde de fréquence $\omega$ et d'amplitude variable. Elle peut se réécrire, grâce aux formules trigonométriques usuelles :

$$x_m(t) = A \exp(j\omega t) + \frac{A}{2} \exp(j(\omega_m + \omega)t) + \frac{A}{2} \exp(j(\omega - \omega_m)t) \tag{2.1.5}$$

Ici le signal est représenté par la somme de trois sinusoïdes d'amplitude et de fréquence constantes.

Dans un système réel d'analyse du son, cette ambiguïté est levée dans la plupart des cas grâce à la résolution du système, c'est à dire sa capacité à distinguer deux sinusoïdes proches. Si $|\omega - \omega_m|$ est plus petit que la résolution du système, la première formulation (eq. (2.1.4)) est préférée, tandis la seconde formulation (eq. (2.1.5)) sera préférée dans le cas contraire. En particulier l'oreille présente, comme tout système réel, une limite à sa résolution [Moore, 1997]. Si $|\omega - \omega_m|$ est suffisamment grand, l'oreille va distinguer plusieurs sinusoïdes d'amplitude constante, dans le cas contraire, celles-ci seront fusionnées en un seul son d'amplitude variable.

## 2.2 Le modèle sinusoïdes+bruit

Parmis les signaux réels, on trouve d'autres types de signaux purement stochastiques pour lesquels la représentation sinusoïdale ne sera pas adaptée. De plus, même si la partie vibrante d'un signal est très énergétique, il y aura toujours un résidu non déterministe du signal pour les signaux réels. Le modèle sinusoïde+bruit [Serra, 1989] a été introduit pour pouvoir représenter cette grande variété de signaux :

$$s(t) = \sum_{i=1}^{M} A_i(t) \exp(j\Phi_i(t)) + n(t) \tag{2.2.1}$$

$n$ est le résidu non déterministe du signal.

Dans cette partie, nous ne nous intéresserons qu'à l'analyse du signal déterministe, mais les méthodes décrites seront également utilisables dans des schémas d'analyse dédiés au modèle sinusoïdes+bruit.

## 2.3 Modèle sinusoïdal local

Le modèle sinusoïdal local est une approximation à court terme du modèle sinusoïdal généralisé, dont les fonctions d'amplitude et de phase peuvent être quelconques.

Ces fonctions sont généralement analytiques[2] et peuvent donc être approchées localement par un modèle polynomial. Ce modèle polynomial local est équivalent à un développement de Taylor dans le voisinage d'un temps $t$. Dans le cas de la modélisation log-polynomiale, on utilisera le modèle suivant :

$$L_K(t+\tau) \triangleq \sum_{k=0}^{K} l^{(k)}(t)\frac{\tau^k}{k!}, \qquad \Phi_Q(t+\tau) \triangleq \sum_{k=0}^{Q} \phi^{(k)}(t)\frac{\tau^k}{k!} \tag{2.3.1}$$

L'avantage d'une telle représentation est celui des approximations de Taylor, à savoir que l'erreur de modélisation décroît très vite lorsque le nombre de termes augmente.

La notation utilisée par la suite pour désigner les différents modèles fait référence aux ordres des polynômes du modèle : **MKQ** va être le modèle d'ordre $K$ en amplitude et d'ordre $Q$ en phase. Les modèles les plus étudiés dans la littérature sont les suivants :

- **M01.** $K = 0$, $Q = 1$. Le modèle le plus simple avec une amplitude et une fréquence constantes. C'est également le plus utilisé pour la modélisation sinusoïdale ;
- **M11.** $K = 1$, $Q = 1$. Le modèle de modulation d'amplitude (AM) de premier ordre, aussi connu sous le nom Modèle Sinusoïdal Exponentiel (Exponential Sinusoidal Model (ESM) en anglais), qui considère des amplitudes modulées exponentiellement ;
- **M02.** $K = 0$, $Q = 2$. Le modèle de modulation de phase de second ordre, aussi appelé chirp, qui considère une fréquence modulée (FM) linéairement ;
- **M12.** $K = 1$, $Q = 2$. Un modèle plus général de premier ordre à la fois en amplitude et en fréquence.

Le modèle le plus utilisé est le modèle sinusoïdal classique avec une amplitude et une fréquence constante, mais un tel modèle va conduire à des difficultés dans la modélisation des signaux modulés rapides, comme les vibratos ou les transitoires. Pour palier à ce problème, des modèles hybrides ont été proposés comme le modèle sinusoïdes+transitoires+bruit. Nous préférons ici plutôt remettre en cause l'utilisation d'un modèle local trop simple.

Les trois derniers modèles sont de plus en plus étudiés. Le modèle AM a été récemment utilisé pour décrire des transitoires [Hermus et al., 2002],[Boyer and Abed-Meraim, 2004], comme des sons percussifs en musique, des attaques de parole, et des systèmes librement amortis [Jensen et al., 2004], comme les cordes pincées. Le modèle FM a également été beaucoup utilisé en analyse de parole et de musique, dans des schémas d'analyse-transformation-synthèse [Peeters and Rodet, 1999], et pour décrire des glissandi ou des transitions entre des phonèmes. D'autres applications dans le domaine des radars et des sonars [Peleg and Porat, 1991], [Djuric and Kay, 1990], et en sismologie [Zhou et al., 1996], existent. Ces deux modèles ont aussi servi en codage audio [Vafin et al., 2001b], [Badeau et al., 2002]. Le dernier modèle est

[2]On rappelle qu'une *fonction* analytique est une fonction complexe continue et dérivable par rapport à ses paramètres complexes. A ne pas confondre avec un *signal* analytique, dont on a parlé précédemment, qui par définition est un signal dont la transformée de Fourier est nulle pour les valeurs de fréquence négatives.

cependant moins utilisé pour l'instant à cause des difficultés liées à l'estimation de ses paramètres.

Enfin, on peut évoquer le modèle **M13**, très utilisé en traitement de parole, mais uniquement dans l'étape de synthèse [McAulay and Quatieri, 1986] et non dans l'étape d'analyse qui nous intéresse ici.

Bien que les modèles polynomiaux en amplitude soient également répandus, nous ne nous intéresserons ici qu'aux modèles log-polynomiaux car ils permettent d'obtenir des équations plus simples à manipuler[3].

Certains travaux traitent le cas plus général d'un ordre de modélisation quelconque [Friedlander and Francos, 1993]. Des méthodes basées sur le critère du maximum de vraisemblance, qui seront décrites plus en détails dans le chapitre suivant, permettent en effet de traiter le cas d'une somme de sinusoïdes avec une décomposition polynomiale d'ordre quelconque à la fois pour l'amplitude et la phase. Un ordre de modélisation très élevé n'est souvent pas nécessaire, les modèles simples donnent en effet de bons résultats dans une grande majorité de situations. La question qu'on peut légitimement se poser est donc le type de modélisation à utiliser. On se propose de discuter brièvement de ce problème dans le paragraphe suivant.

### 2.3.1 Erreur de modélisation

Dans cette section, nous allons étudier l'erreur de modélisation dûe à l'approximation d'une sinusoïde généralisée par un développement de Taylor. Un modèle est adapté à un type de signal donné si son erreur de modélisation est faible dans la majorité des cas de figure du signal étudié. Une borne sur l'erreur de modélisation peut donc donner de précieuses indications sur le domaine de validité du modèle [Mallat, 2000]. Le détail des démonstrations est donné dans l'annexe C.

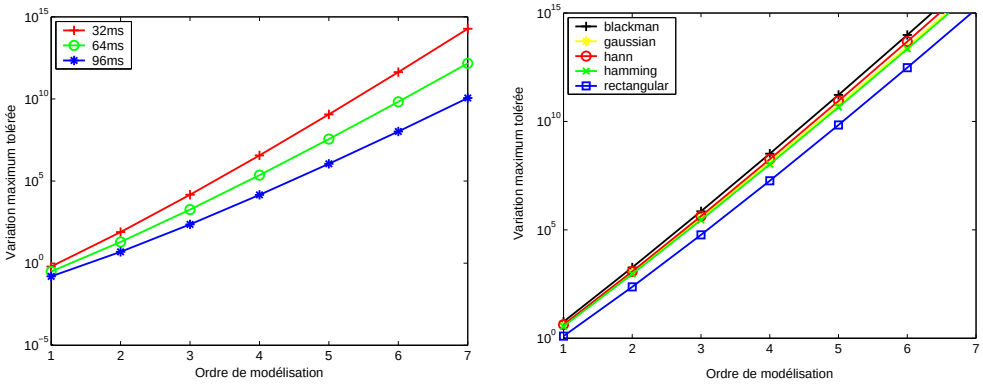
L'erreur de modélisation est la différence entre la vraie fonction $x$, et le modèle utilisé $x_M$. Ici $x$ est une sinusoïde généralisée et $x_M$ est l'approximation de Taylor d'ordre $q-1$ en phase et $k-1$ en amplitude. Cette approximation est faite localement, c'est à dire sur un intervalle d'analyse de longueur $N$. En supposant le signal $x_M$ non nul, nous utilisons l'erreur normalisée $\epsilon$, afin de faciliter l'interprétation de l'erreur, qui doit alors être comparée à l'unité. On peut tout d'abord montrer que l'erreur de modélisation $\epsilon$ est formée de deux termes, un dû à l'approximation sur la phase $\epsilon_\Phi$ et l'autre à l'approximation sur l'amplitude $\epsilon_A$ :

$$\epsilon = \frac{x(t) - x_M(t)}{x_M(t)} = \epsilon_\Phi + \epsilon_A \tag{2.3.2}$$

Ces deux termes sont explicités dans l'annexe C. Dans le cas du modèle log-polynomial en amplitude il y a un terme croisé supplémentaire mais qui est négligeable par rapport aux deux autres.

Les bornes sur $\epsilon_A$ vont être différentes si le modèle est polynomial en amplitude ou log-polynomial. Pour un ordre de modélisation $q-1$ en phase et $k-1$ en amplitude,

[3] Voir également la discussion à la fin de ce chapitre, à la section 2.3.1.




(a) Erreur sur le signal

(b) Erreur sur la transformée de Fourier du signal

FIG. 2.1: Valeur maximum tolérée de variation ($M_{\Phi^q}$, $M_{L^k}$, $M_{A^k}$) en fonction de l'ordre de modélisation ($q-1$ ou $k-1$), pour une erreur de 1%

les bornes du cas polynomial sont :

$$|\epsilon_\Phi| \leq \frac{{\tau_N}^q}{2^q q!} M_{\Phi^q} \tag{2.3.3}$$

$$|\epsilon_A| \leq \frac{{\tau_N}^k}{2^k k!} \frac{M_{A^k}}{m_A} \tag{2.3.4}$$

où $N$ est la taille la fenêtre d'analyse. $M_{\Phi^q}$ est le maximum de variation tolérée sur le paramètre d'ordre $q-1$, ou en d'autres termes, la valeur maximum de la dérivée d'ordre $q$ de la fonction de phase. $M_{A^k}$ est l'équivalent pour la fonction d'amplitude. $m_A$ est la valeur minimum de l'amplitude du signal sur l'intervalle considéré. $\frac{M_{A^k}}{m_A}$ représente donc une variation normalisée d'amplitude.

Dans le cas log-polynomial ces bornes deviennent :

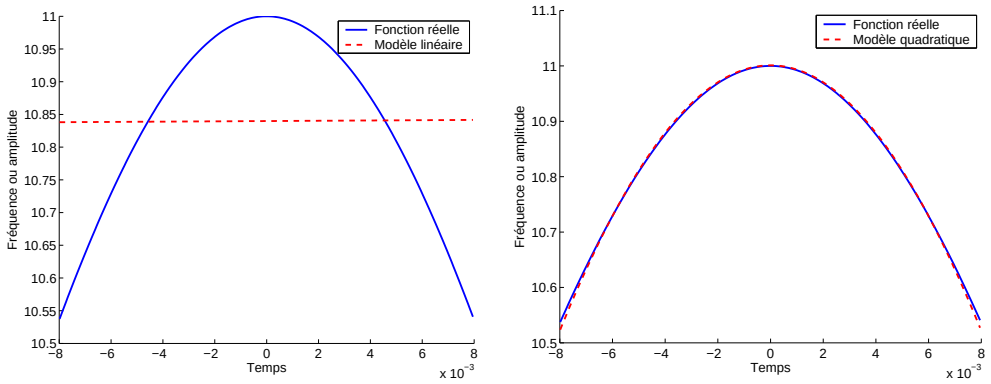
$$|\epsilon_\Phi| \leq \frac{{\tau_N}^q}{2^q q!} M_{\Phi^q} \tag{2.3.5}$$

$$|\epsilon_L| \leq \frac{{\tau_N}^k}{2^k k!} M_{L^k} \tag{2.3.6}$$

La borne sur l'amplitude est valable uniquement si le terme croisé reste négligeable, concrètement si l'erreur reste inférieure à 20%[4]. Comme le modèle est exponentiel, ici $M_{L^k}$ représente directement une variation normalisée d'amplitude.

Toutes les bornes présentent la même forme de fonction. La figure 2.1(a) représente la valeur limite des paramètres $M_{\Phi^q}$, $M_{L^k}$ et $M_{A^k}$ pour que ces bornes restent en deçà de 1% d'erreur, c'est-à-dire la fonction $0.01\frac{2^k k!}{{\tau_N}^k}$. Ces courbes sont fonction de l'ordre de modélisation $k$. Comme l'on pouvait s'y attendre, la valeur maximale acceptable pour $M$ augmente très vite avec l'ordre de modélisation. Pour les signaux

[4]Voir Annexe C pour plus de détails.




(a) Avec un modèle linéaire (b) Avec un modèle quadratique

FIG. 2.2: Modélisation d'une modulation lente d'amplitude ou de fréquence, sur un intervalle de 32ms.

sonores réels on s'attend à ce que ces paramètres soient bornés et proches de zéro. Augmenter l'ordre de modélisation va donc très rapidement faire décroître l'erreur de modélisation. La courbe a été tracée pour trois tailles de fenêtres différentes. On voit que la limite est plus faible lorsque la taille de fenêtre augmente, ce qui est cohérent avec la notion de modèle local du signal : plus l'intervalle d'analyse est grand plus le signal est difficile à modéliser.

Les estimateurs étudiés dans cette partie sont tous basés sur la transformée de Fourier. L'erreur induite par la modélisation sur la transformée de Fourier $X$ de $x$ va donc apporter une information sur les performances que l'on peut attendre de la modélisation choisie. La TF étant linéaire, l'erreur sera également composée de deux termes, comme dans l'équation (2.3.2), et les bornes auront des formes très similaires aux précédentes (cf Annexe C). La limite sur les paramètres $M$ a été tracée sur la Figure 2.1(b) pour différents types de fenêtres utilisées pour calculer la TF de $x$ sur l'intervalle du modèle local. La taille de la fenêtre est fixée à 32ms et l'erreur tolérée est toujours de 1%. Une erreur relative inférieure à 1% sur la TF assure que l'erreur de modélisation n'aura qu'un impact limité sur une estimation des paramètres sinusoïdaux basée sur la TF. Pour un modèle à fréquence constante par exemple (ordre 1 en phase), la pente de fréquence maximum constatée ($q = 2$ sur la figure) sur le signal réel ne doit pas dépasser 1202 $rad.Hz.s^{-1}$, pour une fenêtre de Hann de 32ms, soit 191 $Hz.s^{-1}$, valeur très faible et souvent dépassée dans les signaux réels, et ce particulièrement sur les attaques. Cela plaide en faveur de l'augmentation de l'ordre de modélisation. En effet, il est important de modéliser convenablement le signal étudié, car toute erreur de modélisation va se répercuter directement sur l'estimation.

Maintenant la question est de savoir à quel ordre s'arrêter, afin d'avoir le meilleur compromis pour le type de signal étudié. Pour le traitement du son, un modèle d'ordre 1 en amplitude et d'ordre 1 en phase permet déjà une meilleure modélisation des attaques des sons, par rapport au modèle classique **M01**, comme l'ont déjà montré

plusieurs études [Vafin et al., 2001a], [Nieuwenhuijse et al., 1998]. De fortes variations d'amplitude et de fréquence sont constatées simultanément lors des transitions entre phonèmes ou pour certaines attaques de parole par exemple. Une modélisation avec variation de fréquence, de type **M12**, devrait donc également permettre de mieux représenter ce type signal. Un autre cas de figure typique des signaux est la modulation sinusoïdale lente d'amplitude[5] ou de fréquence. Un modèle linéaire en amplitude ou en fréquence ne sera pas suffisant pour avoir une modélisation convenable de tels phénomènes sur une durée de l'ordre d'une demi-période de la modulation, comme le montre pour un cas critique la figure 2.2(a). En revanche, un modèle d'ordre 2 en amplitude et d'ordre 3 en phase permettrait une représentation très correcte de ces signaux à cette échelle (figure 2.2(b)). Avec un tel modèle, tous les cas de figures typiques d'un signal devraient pouvoir être bien représentés, et donc idéalement ce serait un modèle comme celui-ci qui devrait être utilisé. Les estimateurs étudiés s'arrêteront cependant au modèle **M12**, qui est déjà une nette avancée sur le modèle **M01**, le plus répandu dans la littérature.

On pourrait objecter qu'au lieu de chercher à augmenter l'ordre de modélisation, il serait plus aisé de diminuer la taille de la fenêtre d'analyse du signal. Cette façon de faire présente néanmoins deux inconvénients majeurs : une fenêtre d'analyse plus courte signifie moins d'échantillons utilisés, et donc une moins bonne estimation des paramètres du modèle, ensuite une analyse basée sur la transformée de Fourier réclame une résolution suffisamment importante pour pourvoir distinguer les composantes sinusoïdales.

Enfin, jusqu'ici on a parlé de deux modèles possibles pour l'amplitude, l'amplitude linéaire et l'amplitude exponentielle[6], lequel des deux préférer ? Leur capacité de modélisation est très similaire. Le modèle exponentiel en amplitude a été utilisé pour les méthodes de sous-espace et les méthodes basées sur la transformée de Fourier essentiellement parce qu'il était adapté à ces méthodes. On peut dire la même chose du modèle linéaire pour la méthode du maximum de vraisemblance. Le seul argument en faveur du modèle d'amplitude exponentielle est que la représentation habituelle de l'amplitude est en décibels, et que dans ce domaine ce modèle est linéaire, donc plus simple.

[5] Il s'agit d'un battement cf paragraphe 2.1.

[6] Voir le paragraphe 2.1 sur les sinusoïdes généralisées.

CHAPITRE 3

# Méthodes classiques d'estimation

Dans ce chapitre nous allons parler des grandes catégories de méthodes existantes pour l'estimation de paramètres des sinusoïdes, à savoir l'estimation au maximum de vraisemblance, (Maximum likelihood (ML) en anglais), les méthodes à "haute résolution", et les méthodes basées sur la transformée de Fourier. Les deux derniers types d'approches peuvent être considérés comme des méthodes approchées de la méthode du maximum de vraisemblance, permettant de palier à certains de ses inconvénients. Dans la section 3.1, nous nous attarderons donc un peu sur la méthode du maximum de vraisemblance. Nous passerons ensuite rapidement sur les méthodes à haute résolution dans la section 3.2, avant de parler des méthodes basées sur la transformée de Fourier dans la section 3.3, dont l'étude sera l'objet principal de cette partie. Nous finirons par quelques rappels rapides sur la transformée de Fourier, qui nous servirons dans toute la suite.

## 3.1 Maximum de vraisemblance

La méthode du maximum de vraisemblance est une méthode très générale d'estimation de paramètres. Elle a été beaucoup utilisée pour estimer les paramètres sinusoïdaux car les estimateurs obtenus par cette méthode sont asymptotiquement optimaux et sans biais.

Néanmoins, dans le cas de modèles complexes, il n'existe pas nécessairement de solutions analytiques. Il faut alors faire appel à des méthodes d'optimisation non-linéaires telle la méthode de Newton. L'inconvénient d'une telle méthode d'optimisation est qu'elle est itérative, qu'elle peut tendre vers un minimum local si la méthode a été mal initialisée, et enfin qu'elle est très coûteuse en temps de calcul.

Beaucoup de travaux se sont attachés à trouver les estimateurs ML pour les différents types de modèles sinusoïdaux. Parmi les modèles traités nous pouvons citer les suivants :

– [Djuric and Kay, 1990] : amplitude constante, fréquence linéaire (modèle **M02**)

– [Wolcin, 1980] : amplitude constante, fréquence linéaire par morceaux[1].
– [Friedlander and Francos, 1993] : amplitude et phase représentée par un modèle paramétrique linéaire[2].
– Sinusoïde amortie exponentiellement, fréquence constante : par exemple la thèse de Badeau [Badeau, 2005] (modèle **M11**).

La méthode présentée dans ces articles, et rappelée dans la section 3.1.1, est toujours identique. L'estimateur ML dans le cas général d'une somme de sinusoïdes avec amplitude log-polynomiale et phase polynomiale n'a jamais été abordé dans la littérature à notre connaissance. Etant donné que certaines méthodes décrites par la suite dans cette partie font intervenir ce type de modèle, et l'importance de l'estimateur de ML pour comprendre les estimateurs de Fourier, nous avons décidé de détailler la dérivation des estimateurs ML pour ce modèle dans la section 3.1.3. La dérivation est similaire à celle utilisée dans l'article [Friedlander and Francos, 1993] pour une somme de sinusoïdes polynomiales en amplitude, qui est rappelée brièvement dans la section 3.1.2. Nous terminerons en donnant dans la section 3.1.4 les estimateurs dans le cas particulier d'une seule sinusoïde. Ce cas correspond aux méthodes basées sur la transformée de Fourier lorsque les sinusoïdes sont bien séparées dans le domaine fréquentiel.

### 3.1.1 Principe général

Soit $s$ un signal perturbé par un bruit $n$ complexe que l'on suppose blanc, Gaussien, circulaire et de variance $\sigma^2$ :

$$s(\tau) = x(\tau) + n(\tau)$$

$x$ est une fonction déterminée par un vecteur de paramètre, $\theta$, que l'on cherche à estimer. La densité de probabilité de l'observation est alors une Gaussienne de variance $\sigma^2$. On va alors tenter de maximiser la log-vraisemblance[3] des observations par rapport au vecteur de paramètres recherchés, afin de trouver les estimateurs de $\theta$ et de $\sigma$.

L'intervalle temporel d'analyse est échantillonné, on aura donc un vecteur colonne de $N$ valeurs, que l'on note $\tau = [0, ..., N-1]/F$. $F$ est la fréquence d'échantillonage. De façon générale, pour une fonction scalaire $f(\tau)$, on va noter le vecteur colonne contenant les valeurs de $f(\tau)$ par $\mathbf{f}$. On montre tout d'abord que la variance du bruit peut se calculer comme la variance du signal $s$ auquel on a soustrait la partie déterministe $x$ :

$$\sigma^2 = \frac{1}{N}\|\mathbf{x} - \mathbf{s}\|^2 \tag{3.1.1}$$

Ensuite on déduit les estimateurs des paramètres sinusoïdaux qui sont linéaires par rapport au modèle, comme les paramètres d'amplitude et de phase d'ordre zéro pour

[1] Ce modèle ne correspond pas à la définition des modèles **MKQ**, car plusieurs morceaux linéaires sont considérés pour la même fenêtre de signal.

[2] Idem, mais cette fois parce que le modèle n'est pas log-linéaire.

[3] Par commodité on va plutôt maximiser la log-vraisemblance plutôt que la vraisemblance, ce qui est tout à fait équivalent étant donné que la fonction log est monotone croissante.

le modèle log-polynomial en amplitude, ou tous les coefficients d'amplitude dans le cas polynomial en amplitude. On réinjecte ces estimateurs pour obtenir une nouvelle fonction de vraisemblance à maximiser ne dépendant que des paramètres non-linéaires du modèle. Ces derniers paramètres seront estimés en utilisant une méthode itérative d'optimisation multidimensionnelle, comme la méthode de Newton [Wolcin, 1980], [Saha and Kay, 2002], [Abotzoglou, 1986].

Lors de l'estimation, la méthode du maximum de vraisemblance se présente donc en trois étapes[4] :

1. Estimation des paramètres non-linéaires avec une méthode itérative comme la méthode de Newton
2. Estimation des paramètres linéaires en injectant les estimations des paramètres non-linéaires dans les formules d'estimation ML de ces paramètres.
3. Estimation de la variance du bruit en soustrayant du signal étudié les sinusoïdes reconstruites avec les paramètres estimés.

Les méthodes d'optimisation non-linéaires demandent généralement une bonne initialisation des paramètres, pour éviter de tomber dans un maximum local. Une méthode de type Fourier peut par exemple être utilisée. Elles demandent également une estimation au préalable du nombre de composantes sinusoïdales. Enfin, l'estimation des paramètres linéaires nécessite l'estimation préalable des paramètres non-linéaires. On va donc avoir inévitablement une propagation des erreurs faites sur les paramètres non-linéaires vers les paramètres linéaires.

Les méthodes qui seront présentées par la suite, à savoir les méthodes de type Fourier et les méthodes de sous-espace, sont proches de la méthode ML, dont elles reprennent les mêmes étapes. La première étape est remplacée par une technique de calcul analytique (non itérative), les deux autres restant identiques.

Nous allons maintenant donner les formules ML pour les deux modèles présentés dans la section 2.3.

### 3.1.2 Estimateur ML pour un modèle polynomial en amplitude et en phase

Le signal étudié est constitué de $M$ composantes sinusoïdales, que l'on a approché par des développement de Taylor et perturbé par un bruit $n$ :

$$s(\tau) = \sum_{i=1}^{M} \Big( \sum_{k=0}^{K} \frac{A_{i,k}}{k!} \tau^k \Big) \exp\Big( j \sum_{q=0}^{Q} \frac{\Phi_{i,q}}{q!} \tau^q \Big) + n(\tau)$$

On cherche à estimer les paramètres $A_{i,k}$ et $\Phi_{i,q}$. Les paramètres linéaires sont ici les coefficients d'amplitude $A_{i,k}$, et les paramètres non-linéaire les coefficients de phase $\Phi_{i,q}$. La dérivation des formules ML pour les paramètres linéaires est un cas particulier de la méthode présentée dans [Friedlander and Francos, 1993]. Pour simplifier la

[4] Les trois étapes de l'estimation sont inversées par rapport aux trois étapes pour trouver les estimateurs, décrites au paragraphe précédent.

notation, on écrit $Z_i(\tau) = \exp(j \sum_{q=1}^{Q} \frac{\Phi_{i,q}}{q!} \tau^q)$. $\tau^{\mathbf{K}}\mathbf{Z_i}$ est le vecteur colonne contenant les valeurs de la fonction $\tau^K Z_i(\tau)$ pour le vecteur des temps échantillonnés $\tau$.

$$\begin{aligned}
\mathbf{\Phi}_i &= [\mathbf{Z_i}, ..., \tau^{\mathbf{K}}\mathbf{Z_i}] \\
\mathbf{\Phi} &= [\mathbf{\Phi}_1, ..., \mathbf{\Phi}_M] \\
a_i &= [A_{i,0} \exp(j\Phi_{i,0}), ..., A_{i,K} \exp(j\Phi_{i,0})] \\
a &= [a_1, ..., a_M]^T
\end{aligned}$$

Avec cette notation l'équation précédente devient :

$$\mathbf{s} = \mathbf{\Phi} a + \mathbf{n} \tag{3.1.2}$$

La densité de probabilité de $\mathbf{s} - \mathbf{\Phi} a$ est une Gaussienne de variance $\sigma^2$. En maximisant cette densité de probabilité, on montre alors que les paramètres d'amplitude s'estiment ainsi :

$$a = (\mathbf{\Phi}^H \mathbf{\Phi})^{-1} \mathbf{\Phi}^H \mathbf{s} \tag{3.1.3}$$

Les paramètres de phase quand à eux vont être estimés en maximisant la fonction :

$$J(\theta) = \mathbf{s}^H \mathbf{\Phi} (\mathbf{\Phi}^H \mathbf{\Phi})^{-1} \mathbf{\Phi}^H \mathbf{s} \tag{3.1.4}$$

où $\theta$ est le vecteur de paramètres à estimer, ici $\theta = \{\Phi_{i,q}\}_{i \in [1..M], q \in [1..Q]}$.

### 3.1.3 Estimateur ML pour un modèle log-polynomial en amplitude et polynomial en phase

Le modèle log-polynomial pour $M$ composantes sinusoïdales est noté :

$$s(\tau) = \sum_{i=1}^{M} \exp\Big( \sum_{k=0}^{K} \frac{L_{i,k}}{k!} \tau^k + j \sum_{q=0}^{Q} \frac{\Phi_{i,q}}{q!} \tau^q \Big) + n(\tau)$$

Les paramètres linéaires sont les phases et les amplitudes constantes $L_{i,0}$ et $\Phi_{i,0}$. La dérivation des formules ML pour les paramètres linéaires est similaire à la méthode présentée dans [Friedlander and Francos, 1993]. Comme précédemment, on va simplifier la notation mais en posant cette fois $Z_i(\tau) = \exp(\sum_{k=1}^{K} \frac{L_{i,k}}{k!} \tau^k + j \sum_{q=1}^{Q} \frac{\Phi_{i,q}}{q!} \tau^q)$. $\mathbf{Z_i}$ désignera le vecteur colonne des valeurs de la fonction $Z_i(\tau)$ pour le vecteur de temps échantillonné $\tau$.

$$\begin{aligned}
\mathbf{\Phi} &= [\mathbf{Z_1}, ..., \mathbf{Z_M}] \\
a &= [\exp(L_{1,0} + j\Phi_{1,0}), ..., \exp(L_{M,0} + j\Phi_{M,0})]^T
\end{aligned}$$

Le modèle $s$ devient comme dans le paragraphe précédent :

$$\mathbf{s} = \mathbf{\Phi} a + \mathbf{n} \tag{3.1.5}$$

Les paramètres d'amplitude et de phase d'ordre 1 s'estiment comme précédemment :

$$a = (\mathbf{\Phi}^H \mathbf{\Phi})^{-1} \mathbf{\Phi}^H \mathbf{s} \tag{3.1.6}$$

Les paramètres d'amplitude et de phase d'ordre supérieur vont être estimés en maximisant la fonction :

$$J(\theta) = \mathbf{s}^H \mathbf{\Phi} (\mathbf{\Phi}^H \mathbf{\Phi})^{-1} \mathbf{\Phi}^H \mathbf{s} \tag{3.1.7}$$

Dans ce cas le nombre de paramètres non-linéaires augmente très vite avec l'ordre de modélisation. La fonction $J(\theta)$ sera donc plus difficile à optimiser et le temps de calcul va être encore plus important.

### 3.1.4 Estimateur ML pour une sinusoïde

#### 3.1.4.1 Modèles M01

C'est un cas particulier du cas précédent où l'on ne considère qu'une seule sinusoïde avec une amplitude constante et une fréquence constante. En notant $\omega_1 = \Phi_{1,1}$ et $Z_1(\tau) = \exp(j\omega_1\tau)$, les vecteurs $a$ et $\mathbf{\Phi}$ se simplifient en :

$$\begin{aligned} \mathbf{\Phi} &= \mathbf{Z_1} \\ a &= \exp(A_{1,0} + j\Phi_{1,0}) \end{aligned}$$

$a$ est maintenant un scalaire et $\mathbf{\Phi}$ est un vecteur colonne. On en déduit alors $\mathbf{\Phi}^H \mathbf{\Phi} = N$ et la fonction à maximiser :

$$\begin{aligned} J(\theta) &= \frac{1}{N} \mathbf{s}^H \mathbf{\Phi} \mathbf{\Phi}^H \mathbf{s} \\ &= \frac{1}{N} |\sum_{n=0}^{N-1} s(\tau_n) \exp(j\omega_1 \tau_n)|^2 \\ &= \frac{1}{N} |S(\omega_1)|^2 \end{aligned}$$

où $\tau_n = n/F$ est le temps échantillonné et $S(\omega_1)$ est la transformée de Fourier de $s$ pour la fréquence $\omega_1$. La fréquence estimée de la sinusoïde $\hat{\omega}_1$ est donc donnée par le maximum du spectre d'amplitude de $s$.

Enfin, l'amplitude complexe $a$ est la transformée de Fourier $S$ de $s$ à cette fréquence :

$$a = \frac{1}{N} S(\hat{\omega}_1) \tag{3.1.8}$$

On peut montrer que le cas de $N$ sinusoïdes peut se résoudre par la transformée de Fourier à condition que les écarts entre les fréquences soient supérieurs à la résolution de la transformée de Fourier[5]. Il suffit alors de trouver les $N$ maxima de la transformée de Fourier, et d'appliquer la méthode d'estimation à chacun des maxima.

#### 3.1.4.2 Modèle MKQ

Comme pour le modèle **M01**, $a$ est un scalaire et $\mathbf{\Phi}$ est un vecteur colonne. En notant $Z_1(\tau) = \exp(\sum_{k=1}^{K} \frac{L_{1,k}}{k!}\tau^k + j\sum_{q=1}^{Q} \frac{\Phi_{1,q}}{q!}\tau^q)$, les vecteurs $a$ et $\mathbf{\Phi}$ se simplifient

[5] Voir par exemple l'article Stoica et al. [2000] sur l'estimation d'amplitude.

en :

$$\begin{aligned}\mathbf{\Phi} &= \mathbf{Z_1} \\ a &= \exp(L_{1,0} + j\Phi_{1,0})\end{aligned}$$

La fonction $J(\theta)$ à maximiser pour trouver les paramètres d'ordre supérieur peut s'écrire

$$J(\theta) = \frac{1}{\sum_{n=0}^{N-1} \exp(2\sum_{k=1}^{K} \frac{L_{1,k}}{k!}\tau_n^k)} \left| \sum_{n=0}^{N-1} s(\tau_n) \exp\Big(\sum_{k=1}^{K} \frac{L_{1,k}}{k!}\tau_n^k - j\sum_{q=1}^{Q} \frac{\Phi_{1,q}}{q!}\tau_n^q\Big) \right|^2 \tag{3.1.9}$$

Il s'agit en quelque sorte d'une généralisation du périodogramme et de la transformée de Fourier, que l'on appellera transformation polynomiale. Cette transformation est normalisée par l'énergie de la transformation $\|\mathbf{\Phi}\|^2 = \sum_{n=0}^{N-1} \exp(2\sum_{k=1}^{K} \frac{L_{1,k}}{k!}\tau_n^k)$, dans le cas de la transformée de Fourier cette énergie était simplement égale à $N$.

Une fois que l'on a maximisé la fonction $J(\theta)$, on déduit les paramètres d'ordre 0 avec la formule :

$$a = \frac{1}{\sum_{n=0}^{N-1} \exp(2\sum_{k=1}^{K} L_{1,k}\tau_n^k)} \sum_{n=0}^{N-1} s(\tau_n) \exp\Big(\sum_{k=1}^{K} L_{1,k}\tau_n^k - j\sum_{q=1}^{Q} \Phi_{1,q}\tau_n^q\Big) \tag{3.1.10}$$

Une fois les paramètres d'amplitude et de phase d'ordre supérieur à 1 estimés, cette formule nous permet d'obtenir les amplitudes et phases d'ordre 0. Elle pourra donc être utilisée conjointement avec les méthodes alternatives d'estimation des paramètres d'ordre supérieur, comme les méthodes reposant sur la transformée de Fourier.

## 3.2 Méthodes à "haute résolution"

Les méthodes à haute résolution sont des méthodes paramétriques du signal, comme les méthodes issues du maximum de vraisemblance. Elles ont été développées pour palier au problème de l'optimisation multidimensionnelle inhérente à ces dernières. Elles ont également une meilleure résolution que les méthodes basées sur la transformée de Fourier, particulièrement pour des fenêtres très courtes.

A l'origine, les méthodes HR reposent sur des techniques de prédiction linéaire pour l'estimation de sommes d'exponentielles [Riche de Prony, 1795] ou de sinusoïdes [Pisarenko, 1973]. Ces méthodes s'étant révélées peu robustes en présence de bruit, de nouvelles approches ont été développées reposant sur les propriétés particulières de la matrice de covariance du signal. Le point commun de ces nouvelles méthodes est de séparer cette matrice en deux sous-espaces, un correspondant aux sinusoïdes, l'espace signal, l'autre à l'espace bruit. Parmi les plus étudiées, on peut citer la méthode MUSIC (MUltiple SIgnal Classification) [Schmidt, 1986], la méthode ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) [Roy et al., 1986] et les méthodes de matrix pencil [Hua and Sarkar, 1990].

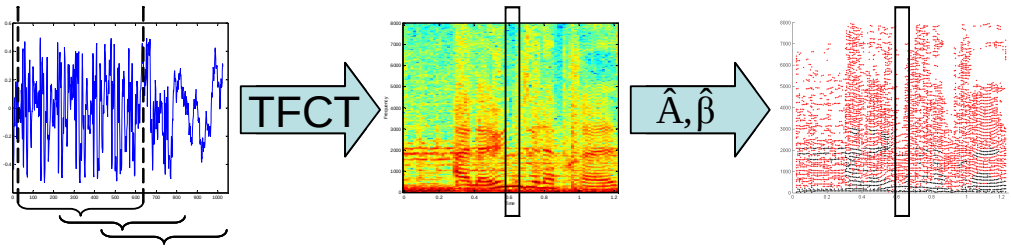



FIG. 3.1: Méthode d'analyse basée sur la FFT

Ces méthodes présentent également un certain nombre d'inconvénients majeurs. Le premier est de reposer sur une décomposition en valeurs singulières du signal, décomposition assez coûteuse en temps de calcul, par rapport à des méthodes basées sur la transformée de Fourier. En particulier lorsque le nombre de sinusoïdes est élevé, c'est à dire supérieur à 25 sinusoïdes [Badeau, 2005], comme c'est le cas pour beaucoup de signaux polyphoniques ou de signaux harmoniques de basse fréquence, ces méthodes deviennent inapplicables. Un deuxième inconvénient est de reposer fortement sur un modèle à fréquence constante. Un tel modèle est peu adapté à la description d'une sinusoïde avec une fréquence variable, qui va demander alors un grand nombre de sinusoïdes à fréquence constante. Une solution est de réduire la taille des fenêtres, mais on réduit d'autant les performances d'estimation attendues. Un dernier désavantage, par rapport à des méthodes basée sur la TF est de demander une estimation préalable du nombre de sinusoïdes présentes dans le signal.

## 3.3 Méthodes reposant sur la transformée de Fourier

Les méthodes basées sur la transformée de Fourier sont toujours les plus utilisées en analyse du signal. Leur principal avantage réside dans l'existence d'un algorithme de faible complexité pour le calcul de la transformée de Fourier discrète, appelé Fast Fourier Transform (FFT), et leur facilité de mise en oeuvre.

### 3.3.1 Vue d'ensemble d'un système d'analyse sinusoïdale basé sur la Transformée de Fourier à Court Terme (TFCT)

La TFCT est une transformée de Fourier discrète, à support temporel fini et fenêtrée. Dans le reste du document nous utiliserons essentiellement la version centrée de la TFCT, définie ainsi :

$$X(t_m, \omega_k; h) = \sum_{n=-(N-1)/2}^{(N-1)/2} x(\tau_n + t_m)\, h(\tau_n) \exp(-j\, \tau_n \omega_k) \qquad (3.3.1)$$

où $N$ est la taille en échantillons du support temporel de la fenêtre $h$, $m$ est le numéro d'échantillon dans le flux audio, $k$ est le numéro de bin fréquentiel, $\tau_n = n/F$ et $t_m = m/F$ sont respectivement le temps local et le temps global en seconde. $F$ est

la fréquence d'échantillonnage du signal et enfin $\omega_k = \frac{2\pi kF}{N}$ est la pulsation du bin $k$. La fenêtre $h$ utilisée est généralement paire et réelle. On notera $M = (N-1)/2$ par la suite, pour alléger la notation[6]. Enfin $P$ sera la taille de la transformée de Fourier. Un signal de taille $N \leq P$ sera donc complété par des zéros, avant de faire la TF[7].

La méthode d'analyse basée sur la TFCT consiste à découper le signal en trames avec recouvrement, et à effectuer une analyse de Fourier discrète sur chacune de ces trames. On obtient alors une représentation temps-fréquence discrète du signal (première étape sur la Figure 3.1) où les axes temporel et fréquentiel sont échantillonnés de façon régulière. Les méthodes d'estimation des paramètres sinusoïdaux que nous allons étudier dans cette partie opèrent directement dans le domaine de la TFCT et permettent une visualisation plus nette des trajectoires sinusoïdales qu'avec la TFCT seule (deuxième étape sur la Figure 3.1).

On suppose que la transformée de Fourier permet de séparer les sinusoïdes dans le domaine fréquentiel. Les pics du spectre de plus forte amplitude vont donc correspondre à des sinusoïdes. Ces bins maximums sont généralement déterminés en utilisant un critère très simple. Le bin $k$ est un maximum à l'instant $t_m$ s'il vérifie :

$$|X(t_m, \omega_{k-1}; h)| < |X(t_m, \omega_k; h)|$$
$$|X(t_m, \omega_{k+1}; h)| < |X(t_m, \omega_k; h)|$$

D'autres définitions des bins maximums sont possibles, certaines faisant intervenir l'adéquation des bins avec le modèle sinusoïdal étudié [McAulay and Quatieri, 1986], [Peeters and Rodet, 1999]. Pour toutes les expériences que nous présenterons, nous nous contenterons de cette définition.

Dans la section 3.1 nous avons précisé que le maximum de la transformée de Fourier était un estimateur de la fréquence, pour une sinusoïde à fréquence constante. Cet estimateur, comme tous ceux du maximum de vraisemblance, est asymptotiquement optimal. Cependant lorsque l'estimation est réalisée avec un nombre d'échantillons fini, on peut trouver des méthodes beaucoup plus précises, combinant les bins de la transformée de Fourier proches du maximum. Ce sont ces estimateurs que nous allons étudier dans les prochains chapitres.

### 3.3.2 Signal bien résolu par la transformée de Fourier

Le principal inconvénient de la transformée de Fourier est la limitation de sa résolution. Comme nous l'avons vu dans le paragraphe 3.1, les méthodes basées sur la transformée de Fourier sont équivalentes au maximum de vraisemblance, à condition que les sinusoïdes puissent être séparées dans le domaine de Fourier. Augmenter la taille de la fenêtre d'analyse, permet de contourner ce problème pour des sinusoïdes à fréquence constante, mais pour des signaux réels on ne peut augmenter indéfiniment la taille de la fenêtre d'analyse car le modèle de signal ne sera alors plus valide.

En particulier pour des signaux à fréquence variant linéairement (modèle **M02**), l'augmentation de la résolution peut conduire à la formation de plusieurs pics dans

[6] $N$ peut prendre des valeurs paires ou impaires. Si $N$ est paire, $n$ prend des valeurs demi-entières, mais la notation reste valable.

[7] Voir la section 3.3.6.1 sur le zéro-padding.

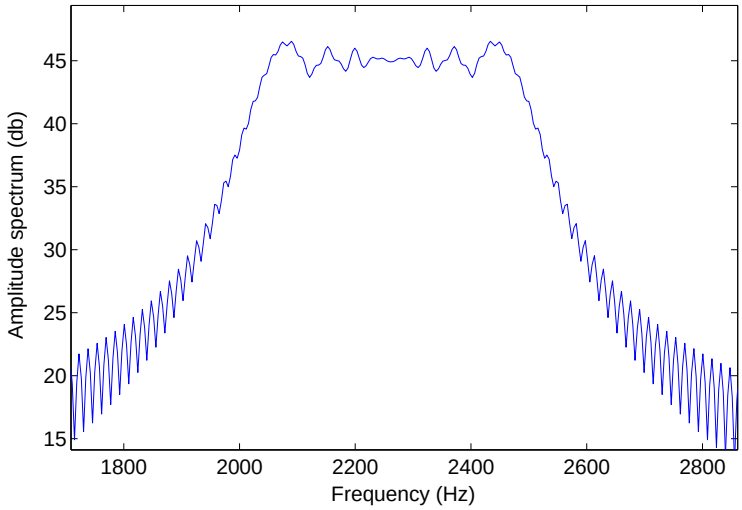



FIG. 3.2: Réponse fréquentielle d'un chirp linéaire fortement modulé

le domaine de Fourier au lieu d'un seul. C'est le même problème que l'on rencontre dans les méthodes à "haute résolution" : la résolution devient trop fine pour identifier correctement le signal, comme étant une seule entité (voir Figure 3.2).

Dans le cas de la transformée de Fourier, la résolution fréquentielle devra être optimisée en fonction du type de signal étudié. Toutes les méthodes présentées par la suite ne seront valables que pour des sinusoïdes compatibles avec la résolution choisie. Elle ne seront valables également, que si le bruit ne masque pas complètement le pic principal de la TF. Par sinusoïdes "bien résolues" nous entendrons donc deux choses : que la résolution de la TF est adaptée au signal étudié, et que le bruit est suffisamment faible, par rapport à l'énergie du signal. Une définition plus précise mathématiquement sera donnée dans la section 5.1.1.

### 3.3.3 Comparaison de la résolution de deux fenêtres de Fourier

Dans les expérimentations, nous serons amené à comparer des méthodes utilisant des fenêtres différentes. Comme les performances des méthodes dépendent de la résolution de la fenêtre utilisée, il sera souhaitable si possible de prendre des fenêtres avec la même résolution.

La mesure choisie pour comparer et régler la résolution fréquentielle des fenêtres est la largeur de bande passante de ces fenêtres, c'est à dire la largeur du lobe principal à -3 dB du maximum. Les -3 dB correspondent à l'amplitude efficace de la sinusoïde. Pour un signal dont l'amplitude est normalisée à 1, on aura donc $A_{\text{Eff}} = \frac{1}{\sqrt{2}}$. Si cette amplitude est mesurée en dB, on retrouve bien -3 dB : $20\log_{10}(A_{\text{Eff}}) \approx -3.01\text{dB}$.

Certains auteurs préfèrent définir la résolution comme étant la largeur fréquentielle à la base du pic. Cependant la transformée de Fourier pourra toujours distinguer deux sinusoïdes même si la séparation entre celles-ci est inférieure à la base du lobe

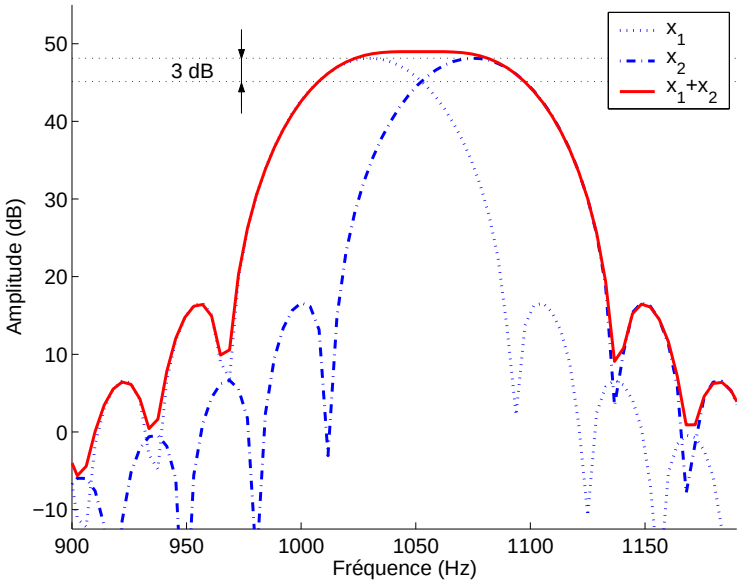



FIG. 3.3: Cas limite pour lequel on ne peut plus distinguer deux sinusoïdes non modulées, de même amplitude et en quadrature de phase. L'écart minimal entre les deux sinusoïdes doit être égal à la largeur du pic à -3dB.

principal . La résolution est donc mesurée de façon plus précise en étant plus près du maximum du pic.

### 3.3.4 Ordre de complexité

La plupart des méthodes d'estimation ne réclament qu'une à trois transformées de Fourier. En terme de complexité, cela veut dire que toutes les méthodes qui seront présentées par la suite seront du même ordre. En effet, la transformée de Fourier rapide est d'ordre $O(N \log_2(N))$, ce qui signifie concrètement que la complexité $C(N)$ est bornée par :

$$C(N) \leq C_1 N \log_2(N) + C_2 N + C_3 \tag{3.3.2}$$

où $C_1$, $C_2$ and $C_3$ sont trois constantes. $N$ est ici la taille de la TF. On peut voir qu'une méthode faisant intervenir un nombre fini et constant (indépendant de $N$) de transformées de Fourier est également d'ordre $O(N \log_2(N))$.

Cependant toutes les méthodes n'auront pas les mêmes performances. La différence va se faire bien évidemment sur le nombre de TF utilisées mais aussi sur la taille de transformée utilisée, certaines méthodes nécessitant l'utilisation de zéro-padding[8].

[8]Le zéro-padding est une technique d'interpolation du spectre discret définie à la section 3.3.6.1.

### 3.3.5 Transformée discrète/transformée continue

Pour les besoins de certaines démonstrations nous utiliserons également la définition continue de la Transformée de Fourier Continue (TFC) :

$$X_c(t,\omega;h) \triangleq \int_{-\infty}^{+\infty} h(\tau)x(t+\tau)\,\mathrm{e}^{-j\omega\tau}\,d\tau \tag{3.3.3}$$

Si le nombre d'échantillons utilisés pour effectuer la transformée de Fourier discrète est suffisamment important, l'approximation faite en passant de la TF continue à la TF discrète sera négligeable. Toutefois certaines propriétés des transformées continues ne sont plus valides en discret. C'est le cas de la propriété de dérivation temporelle de la transformée de Fourier. Dans l'annexe A, nous avons récapitulé les propriétés respectives de ces deux transformations.

### 3.3.6 Quelques rappels sur la transformée de Fourier

Cette section a pour but de rappeler quelques méthodes liées à la transformée de Fourier qui seront fortement utilisées par la suite, la technique du zéro-padding, qui permet d'interpoler le spectre discret, et la méthode de zéro-phasing qui permet de passer d'une transformée de Fourier à réponse de phase linéaire à une transformée zéro-phase.

#### 3.3.6.1 Interpolation du spectre par zéro-padding

Le zéro padding, ou bourrage de zéros en français est une opération qui consiste à ajouter des zéros à la fin ou au début d'un signal ou d'un spectre avant de réaliser la transformée de Fourier discrète, ou la transformée inverse, de ce signal. Si cette opération est réalisée sur le signal temporel, cela va correspondre à une interpolation dans le domaine fréquentiel, car on utilise alors plus de bins fréquentiels pour représenter le spectre, pour la même taille de signal et pour la même fréquence d'échantillonnage. Réciproquement, si des zéros sont ajoutés au début ou à la fin du spectre, cela va correspondre à une interpolation dans le domaine temporel.

Bien entendu, l'interpolation ne conduit pas à une augmentation de la résolution, cela signifie simplement que le spectre est représenté avec plus de points.

#### 3.3.6.2 Transformée de Fourier zéro-phase

Souvent les implémentations de la transformée de Fourier ne sont pas centrées, c'est à dire que la variable temporelle varie entre 0 et $(N-1)/F$ au lieu de varier entre $-(N-1)/(2F)$ et $(N-1)/(2F)$. Pour mettre en oeuvre la plupart des méthodes décrites par la suite, il faudra alors passer d'une transformation non centrée à une transformation centrée.

Pour les deux transformations, la réponse d'amplitude est identique, seule la réponse de phase sera différente. Pour un signal symétrique réel, la réponse de phase pour la transformée de Fourier non centrée est linéaire par rapport à la fréquence, alors qu'elle est nulle pour une transformation de Fourier centrée. On dit alors que

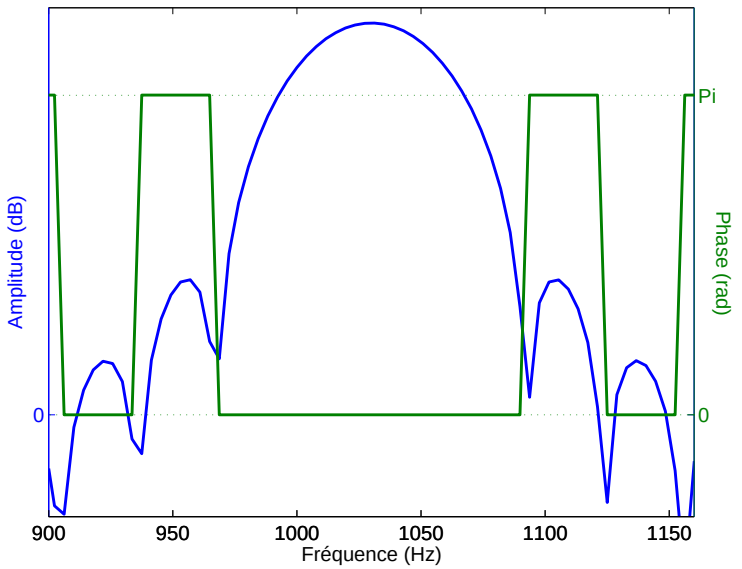



FIG. 3.4: Spectre d'amplitude et de phase d'une transformée de Fourier zéro-phasée. Le signal est une sinusoïde non modulée.

la transformée de Fourier est zéro-phase. En effet, si $h$ est paire, $h(\tau) = h(-\tau)$, la transformée de Fourier de la fenêtre $h$ devient :

$$H(\omega) = \sum_{n=-M}^{n=M} h(\tau_n)\cos(\omega\tau_n) + j\sum_{n=-M}^{n=+M} h(\tau_n)\sin(\omega\tau_n) \tag{3.3.4}$$

Comme $h$ est paire et le sinus impaire, la partie imaginaire va être nulle et donc la phase ne dépend que du signe de la partie réelle.

$$\arg(H(\omega)) = \begin{cases} 0, & \text{si } \Re(H) \geq 0 \\ \pi, & \text{si } \Re(H) < 0 \end{cases} \tag{3.3.5}$$

Les sauts de phase de $\pi$ correspondent aux annulations du spectre d'amplitude de $h$ (Figure 3.4).

Il existe deux méthodes pour obtenir une représentation à phase nulle à partir de la transformation de Fourier linéaire. La première méthode consiste à effectuer une permutation circulaire du signal fenêtré, de façon à placer le milieu de la fenêtre exactement sur le premier échantillon temporel [Marchand, 2000]. Dans ce cas on doit utiliser une fenêtre de longueur impaire, de façon à conserver l'échantillon correspondant au milieu de la fenêtre. Lorsque l'on utilise des implantations rapides de la transformée de Fourier, les tailles des TF sont paires, donc dans ce cas une opération de zéro-padding est nécessaire avant la permutation circulaire.

Si $x$ est le signal original, le nouveau signal permuté peut s'écrire ainsi :

$$y(\tau_n) = \begin{cases} h(\tau_{n+M})x(\tau_{n+M}), & \text{si } n \in [0, M] \\ 0, & \text{si } n \in [M+1, P-M+1] \\ h(\tau_{n-P+M})x(\tau_{n-P+M}), & \text{si } n \in [P-M, P-1] \end{cases} \tag{3.3.6}$$

Le signal a été en quelque sorte retardé de $\tau_M = (N-1)/(2F)$ secondes. Le nouveau temps de référence $\tau^{'}$ est lié à l'ancien temps de référence par la relation : $\tau = \tau^{'} + \tau_M$.

La deuxième méthode pour rendre la TF zéro-phase est plus simple, et convient à toutes les tailles de fenêtres, paires ou impaires. Elle consiste à compenser la distorsion linéaire de phase par le facteur complexe $\exp(j\omega\tau_M)$. La TF zéro-phase $Y$ est donc obtenue à partir de la TF linéaire en posant $Y(\omega) = X(\omega)\,\mathrm{e}^{j\omega\tau_M}$.

## 3.4 Protocole pour la comparaison expérimentale des estimateurs

Dans cette section nous allons présenter rapidement le protocole opératoire utilisé pour comparer les méthodes d'estimation décrites. De façon classique, les méthodes seront aussi comparées aux bornes d'estimation théoriques, les bornes de Cramer-Rao.

### 3.4.1 Mode opératoire

On va comparer l'erreur quadratique moyenne des estimateurs (Mean Squared Error (MSE) en anglais). Les MSE seront tracées en fonction du rapport signal à bruit (Signal to Noise Ratio (SNR) en anglais), mesuré en décibel (dB) que l'on va faire varier généralement entre -10 et 100 dB. La MSE va être calculée pour un grand nombre d'expériences indépendantes, 10000 sauf indication contraire. Chaque expérience consiste d'abord à tirer aléatoirement les paramètres du modèle dans un certain intervalle de définition, qui sera précisé à chaque expérience. Ensuite on ajoute le bruit avec le SNR souhaité et on réalise une estimation des paramètres. Enfin on met à jour la MSE.

Pour toutes les méthodes d'estimation basées sur la transformée de Fourier, une détection de pic est nécessaire au préalable. La plupart des évaluations expérimentales de ces méthodes sélectionnent le bin d'amplitude maximum comme le pic correct. Si cette façon de faire est très satisfaisante pour un fort rapport signal à bruit, elle peut conduire à d'importantes erreurs pour des SNRs faibles. Comme nous ne souhaitons pas évaluer les performances de la méthode de sélection de bin, nous supposerons que le bin maximum est connu, *i.e.* le bin le plus proche de la fréquence moyenne de la sinusoïde. Lorsque les méthodes requièrent l'utilisation de plusieurs trames consécutives, le bin correct sera supposé connu seulement pour la trame du milieu, afin d'avoir une comparaison équitable entre les différentes méthodes.

Deux intervalles de variation d'amplitude et de fréquence seront généralement utilisés dans les expériences : un intervalle pour des modulations faibles, $[0, 1000]$

pour la variation de fréquence $\gamma$ et $[0, 10]$ pour la variation d'amplitude $\mu$, et un intervalle pour des modulations fortes, $[0, 8000]$ et $[0, 100]$ respectivement. Une modulation d'amplitude d'environ 100 correspond à une augmentation d'amplitude de 870 dB par seconde, ce qui ramené à la taille des fenêtres usuelle est une variation forte mais pas irréaliste. Par exemple, une attaque rapide de trompette peut présenter une augmentation d'amplitude de 30 dB en moins de 30 ms, conduisant à une variation d'amplitude d'environ 1000 dB par seconde. En ce qui concerne la variation de fréquence, on a constaté que dans des cas critiques, comme les transitions rapides entre phonèmes dans le cas de la parole, la fréquence fondamentale pouvait croître avec une vitesse de 5000 Hz/s, sans parler des harmoniques de cette fréquence fondamentale qui peuvent possèder bien sûr des variations beaucoup plus importantes. Ces deux cas critiques ont motivé nos choix pour les intervalles de variation de fréquence et d'amplitude.

### 3.4.2 Bornes de Cramer-Rao

La théorie de l'estimation permet de trouver des bornes correspondant à la meilleure estimation possible lorsque le signal est perturbé par du bruit additif. La meilleure estimation correspond à l'estimateur non biaisé de variance minimale et ses performances sont données par la borne de Cramer-Rao (Cramer-Rao Bound (CRB) en anglais). Ces bornes sont très importantes lorsque l'on analyse la performance d'un estimateur, car elles servent de référence.

Parfois les estimateurs à comparer peuvent être légèrement biaisés. La comparaison à la CRB est toujours pertinente à condition dans ce cas de comparer les erreurs quadratiques moyennes (MSE an anglais). En effet un biais va accroître la MSE par rapport à l'estimateur non biaisé, donc la CRB reste la borne d'erreur minimale.

La dérivation des bornes de Cramer-Rao est un problème qui a été étudié pour un grand nombre de modèles. Soit $x$ la variable aléatoire observée et $\theta$ l'ensemble des paramètres du modèle. Soit $P(x|\theta)$ la vraisemblance du signal $x$ pour l'ensemble de paramètres $\theta$, et $l = \ln(P(x|\theta))$ la log-vraisemblance du modèle. Si le modèle statistique est régulier, c'est à dire que $P(x|\theta)$ est dérivable par rapport aux paramètres $\theta$, l'inégalité de Cramer-Rao établit que la variance minimale de l'estimateur non biaisé est donnée par les termes diagonaux de la matrice de Fisher inverse [Kay, 1993] :

$$\mathrm{var}(\theta_i) \geq J_{ii}^{-1} \tag{3.4.1}$$

$$J_{ij}(\theta) = E\left(\frac{\partial l(x,\theta)}{\partial \theta_i}\frac{\partial l(x,\theta)}{\partial \theta_j}\right) \tag{3.4.2}$$

où $J$ est la matrice d'information de Fisher.

Dans le cas où le bruit est blanc et Gaussien de variance $\sigma$ et si le signal est une sinusoïde bruitée du type $x(t) = \mathrm{e}^{L(t)+j\Phi(t)} + n(t)$, on peut montrer que les paramètres $\theta_L$ de l'amplitude et les paramètres $\theta_\Phi$ de la phase sont découplés [Zhou et al., 1996]. On a donc deux matrices de Fisher indépendantes pour les paramètres d'amplitude

et de phase, ce qui facilite l'inversion. Les termes restants s'expriment ainsi :

$$J_{L,ij} = \frac{2}{\sigma^2} \sum_{n=-(W-1)/2}^{(W-1)/2} \mathrm{e}^{2L(\tau_n)} \frac{\partial L(\tau_n)}{\partial \theta_{L,i}} \frac{\partial L(\tau_n)}{\partial \theta_{L,j}} \quad \text{Pour les paramètres d'amplitude} \tag{3.4.3}$$

$$J_{\Phi,ij} = \frac{2}{\sigma^2} \sum_{n=-(W-1)/2}^{(W-1)/2} \mathrm{e}^{2L(\tau_n)} \frac{\partial \Phi(\tau_n)}{\partial \theta_{\Phi,i}} \frac{\partial \Phi(\tau_n)}{\partial \theta_{\Phi,j}} \quad \text{Pour les paramètres de phase} \tag{3.4.4}$$

$W$ est le nombre d'échantillons total utilisés pour l'estimation, on le distingue de $N$ le nombre d'échantillons utilisés dans la TFCT, car pour des méthodes faisant intervenir plusieurs trames successives, $N$ sera différent de $W$. En particulier si la méthode utilise deux trames successives séparées par $H$ échantillons, on aura $W = N + H$. Dans le cas d'un modèle **MKQ** avec $K \leq 2$ et $Q \leq 2$ ces deux matrices sont facilement inversibles.

Les formules des CRBs utilisées dans les parties expérimentales sont regroupées dans le tableau 3.1. Dans ces formules $\epsilon_q$ est défini par :

$$\epsilon_q = \sum_{n=-(W-1)/2}^{(W-1)/2} \tau_n^q \, \mathrm{e}^{2\mu\tau_n} \tag{3.4.5}$$

Le modèle local le plus général utilisé par la suite est le modèle log-polynomial **M12** :

$$x_{12}(\tau) = \mathrm{e}^{\lambda+\mu\tau+j(\alpha+\beta\tau+\frac{\gamma}{2}\tau^2)} \tag{3.4.6}$$

Les modèles **M01**, **M02** et **M11** sont des cas particuliers de ce modèle. Pour avoir la CRB du modèle **M01** par exemple, il suffit de faire tendre $\gamma$ et $\mu$ vers zéro. La CRB pour le modèle **M01** peut être trouvée dans n'importe quel livre traitant de l'estimation sinusoïdale, par exemple dans [Kay, 1993]. La CRB pour le modèle **M02** est donnée par exemple dans l'article [Djuric and Kay, 1990]. Pour les modèles **M11** et **M12**, on s'est inspiré de l'article [Zhou et al., 1996]. A la différence de [Zhou et al., 1996], l'amplitude n'est pas normalisée ici, ce qui conduit à des formules légèrement différentes. On a choisi l'origine des temps au milieu de la fenêtre, pour être cohérent avec notre définition de la transformée de Fourier. Cependant le choix de l'origine des temps est un problème délicat car les performances attendues vont en dépendre fortement[9].

Parfois on préfère exprimer les CRBs en fonction du rapport signal à bruit $\eta$. L'énergie de la sinusoïde dépend de sa fonction d'amplitude, donc l'expression du SNR sera différent suivant l'ordre du polynôme d'amplitude. En supposant toujours que le bruit est blanc Gaussien, le SNR pour une fonction d'amplitude constante et

[9] On discute de ce problème dans la section 5.3.1.

TAB. 3.1: CRBs pour les modèles **M01**, **M11**, **M02** et **M12**

| | Amplitude ordre 0 | Amplitude ordre 1 |
|---|---|---|
| Phase ordre 1 | $CRB_\lambda = \frac{F\sigma^2}{2W\,\mathrm{e}^{2\lambda}}$<br>$CRB_\alpha = \frac{\sigma^2}{\mathrm{e}^{2\lambda}}\frac{(2W-1)F}{W(W+1)}$<br>$CRB_\beta = \frac{\sigma^2}{\mathrm{e}^{2\lambda}}\frac{6F^3}{W(W^2-1)}$ | $CRB_\lambda = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_2}{D_1}$<br>$CRB_\mu = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_0}{D_1}$<br>$CRB_\alpha = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_2}{D_1}$<br>$CRB_\beta = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_0}{D_1}$ |
| Phase ordre 2 | $CRB_\lambda = \frac{F\sigma^2}{2W\,\mathrm{e}^{2\lambda}}$<br>$CRB_\alpha = \frac{\sigma^2}{\mathrm{e}^{2\lambda}}\frac{(9W^2-21)F}{8W(W^2-4)}$<br>$CRB_\beta = \frac{\sigma^2}{\mathrm{e}^{2\lambda}}\frac{6F^3}{W(W^2-1)}$<br>$CRB_\gamma = \frac{\sigma^2}{\mathrm{e}^{2\lambda}}\frac{90F^5}{W(W^2-1)(W^2-4)}$ | $CRB_\lambda = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_2}{D_1}$<br>$CRB_\mu = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_0}{D_1}$<br>$CRB_\alpha = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_2\epsilon_4-\epsilon_3^2}{D_2}$<br>$CRB_\beta = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_0\epsilon_4-\epsilon_2^2}{D_2}$<br>$CRB_\gamma = \frac{\sigma^2}{2\,\mathrm{e}^{2\lambda}}\frac{\epsilon_0\epsilon_2-\epsilon_1^2}{D_2}$ |
| | | Avec $D_1 = \epsilon_0\epsilon_2 - \epsilon_1^2$<br>Avec $D_2 = \epsilon_0\epsilon_2\epsilon_4 - \epsilon_1^2\epsilon_4 - \epsilon_0\epsilon_2^2 + 2\epsilon_1\epsilon_2\epsilon_3 - \epsilon_2^3$ |

pour une modulation linéaire d'amplitude s'écrivent ainsi :

$$\eta = \frac{\mathrm{e}^{2\lambda}}{\sigma^2} \qquad \text{Pour une amplitude constante}$$

$$\eta = \frac{\mathrm{e}^{2\lambda}}{\sigma^2}\frac{\sinh(\mu W)}{\mu W} \qquad \text{Pour une amplitude log-linéaire}$$

$\mathrm{e}^{2\lambda}\frac{\sinh(\mu W)}{\mu W}$ est la puissance de la sinusoïde modulée. Si on fait tendre $\mu$ vers zéro dans la deuxième équation, on retrouve bien la première.

On peut remarquer que le SNR pour une amplitude log-linéaire tend vers l'infini lorsque $W$ tend vers l'infini. En fait pour un tel signal, il n'y aura pas d'analyse asymptotique possible. Dans la réalité, ce type de signal n'existe de toute façon qu'à court terme, comme sur une attaque de parole par exemple. On peut remarquer également que la CRB dépend d'un paramètre que l'on doit estimer, $\mu$. Pour évaluer les méthodes selon le protocole décrit dans la section 3.4.1, nous serons amenés à faire des expériences avec des valeurs de $\mu$ différentes, et donc des CRBs différentes. Pour ces expériences, la CRB représentée sera la CRB moyenne.

CHAPITRE 4

# Etat de l'art des méthodes d'analyse basées sur la TFCT

Dans ce chapitre nous allons présenter plus en détail les méthodes d'estimation basée sur la TFCT. Nous allons aborder chacun des modèles **M01**, **M02** et **M12** l'un après l'autre. Pour les méthodes basées sur la TF, le modèle **M11** n'est pas abordé dans la littérature. Les méthodes seront placées dans la section correspondant au modèle sous-jacent le plus général qui leur est associé dans la littérature. Bien évidemment, si les méthodes peuvent estimer les paramètres d'un modèle **MKQ**, elles pourront également estimer les paramètres d'un modèle moins général d'ordre $K^{'} \leq K$ et $Q^{'} \leq Q$. Nous verrons par la suite que certaines des méthodes présentées sont valables dans des cas plus généraux que ceux de la littérature.

Le principe de chaque méthode sera décrit de façon complète, et avec un formalisme unifié, qui permet de mieux saisir les ressemblances et différences entre ces méthodes. Nous terminerons par une comparaison expérimentale de ces estimateurs à la section 4.4.

## 4.1 Estimation pour le modèle M01

Le modèle **M01** est le modèle le plus étudié dans la littérature. On rappelle qu'il fait l'hypothèse d'une fréquence et d'une amplitude constantes :

$$x(\tau) \triangleq A\,\mathrm{e}^{j(\alpha_M+\beta\tau)} \tag{4.1.1}$$

où $\alpha_M = \Phi(t_M)$ est la phase initiale de la sinusoïde. Les paramètres $A$ et $\beta$ dépendent aussi implicitement de l'indice $M$, mais il a été supprimé pour souligner que $A$ et $\beta$ ont des valeurs constantes sur l'intervalle d'analyse. Un grand nombre d'estimateurs de fréquence ont été développés pour ce modèle, que l'on peut regrouper en deux familles, les estimateurs de type vocodeur de phase, qui utilisent la dérivée discrète de la phase, et les interpolateurs de spectre. Pour l'estimation des amplitudes et des

phases, il existe plusieurs techniques d'estimation, dont une basée sur la TFCT. C'est cette méthode qui est utilisée dans la plupart des systèmes d'analyse.

### 4.1.1 Estimation de phase et d'amplitude

Dans la section 3.1, on a vu que la méthode du maximum de vraisemblance nous donnait des estimateurs pour les phases et amplitudes de mélanges de sinusoïdes. Dans les méthodes basées sur la TFCT, on utilise cependant les estimateurs correspondant à une seule sinusoïde dans du bruit, car dans ce cas la phase et l'amplitude peuvent être obtenues sans calculs supplémentaires de TFCT, et ne dépendent que de la fréquence de la sinusoïde, ce qui rend la méthode particulièrement simple et rapide.

Dans notre cas, où le bruit est supposé blanc Gaussien, il y a équivalence entre le maximum de vraisemblance et le critère des moindres carrés. Dans le cas où le bruit est non blanc, d'autres méthodes existent comme les moindres carrés pondérés, ou les méthodes basées sur un banc de filtres adaptés (matched filter bank). Toutes ces méthodes étant optimales asymptotiquement, elles vont donner des résultats comparables si $N$ est suffisamment grand [Stoica et al., 2000]. Nous nous contenterons donc ici de décrire l'estimateur de Fourier, le plus utilisé.

#### L'estimateur de Fourier

L'estimateur de Fourier est donné par la formule (3.1.8) :

$$a = \frac{1}{N} X(t, \beta; h_{rec}) \tag{4.1.2}$$

où $h_{rec}$ est la fenêtre rectangulaire. En pratique, lorsque $N$ est fini, on préférera utiliser des transformées de Fourier fenêtrées pour réduire les perturbations causées par des sinusoïdes proches. En notant $H(\omega)$ la transformée de Fourier de la fenêtre $h$ pour la fréquence $\omega$, cet estimateur devient :

$$a = \frac{1}{H(0)} X(t, \beta; h) \tag{4.1.3}$$

$a$ désigne l'amplitude complexe :$a = A\exp(j\alpha_M)$. Pour éviter un calcul de transformée de Fourier pour la fréquence $\beta$, on utilise généralement le bin $k$ de la TFCT le plus proche de $\beta$ :

$$a = \frac{1}{H(\omega_k - \beta)} X(t, \omega_k; h) \tag{4.1.4}$$

Si on utilise une transformée de Fourier à phase nulle (cf. section 3.3.6.2), $\arg(H(\omega - \beta)) = 0$ pour toute fréquence $\omega$ au voisinage de $\beta$, en particulier pour $\omega_k$. On obtient donc comme estimateurs pour la phase et l'amplitude :

$$\hat{A} = \frac{1}{|H(\omega_k - \beta)|} |X(t, \omega_k; h)| \tag{4.1.5}$$

$$\hat{\alpha}_M = \arg(X(t, \omega_k; h)) \tag{4.1.6}$$

La fonction $|H(\omega)|$ est une fonction connue, ne dépendant que de $h$. Elle peut donc être facilement prétabulée. On note $R = \pi F/N$ la demi précision de la transformée de Fourier. L'intervalle utile pour la tabulation est $\omega \in [0, R]$, car la fonction est symétrique et l'écart maximum entre $\beta$ et $\omega_k$ est égal à $R$.

### 4.1.2 Estimation de fréquence

Parmi les méthodes d'estimation de la fréquence, deux catégories émergent, les interpolateurs de spectre et les méthodes basées sur la dérivée discrète de la phase. Dans la première catégorie, on trouve deux types de méthodes présentant des différences importantes : les méthodes d'interpolation du spectre d'amplitude, et les méthodes appelées "interpolateurs de spectre utilisant la phase" [Macleod, 1998]. Ce dernier nom est plutôt mal choisi car contrairement aux premières elles n'utilisent pas de fonction d'interpolation prédéfinie. Le seul point commun entre ces méthodes est qu'elles utilisent des bins de la TFCT qui diffèrent en fréquence, alors que les méthodes basées sur la dérivée de la phase, utilisent des bins qui diffèrent en temps. Nous utilisons ici une typologie différente, caractérisée par le paramètre temps-fréquence de la TFCT sur lequel ces méthodes travaillent : la coordonnée temporelle pour les méthodes de dérivation de la phase (méthodes d'estimation à fréquence constante), la coordonnée fréquentielle pour les méthodes de type interpolation (méthodes d'estimation à temps constant).

Dans cette section, nous allons présenter une revue des principaux estimateurs utilisés dans la littérature. Parmi les méthodes à temps variable, deux méthodes sont détaillées, la méthode utilisée dans le vocodeur de phase et la méthode dite de la dérivée, ainsi que deux variantes très proches. Parmi les méthodes à fréquence variable il y aura deux méthodes d'interpolation du spectre d'amplitude, l'interpolation parabolique et l'interpolation triangulaire, ainsi que deux méthodes utilisant la phase, la méthode de Quinn et la méthode de MacLeod.

#### 4.1.2.1 Le vocodeur de phase

Le terme "vocodeur", dérivé de "voice coder" en anglais, fait référence à un système d'analyse [Puckette and Brown, 1998] et synthèse [Griffin and Lim, 1984] de parole complet. Le vocodeur de phase est un codeur de voix particulier qui utilise une représentation basée sur la TFCT [Flanagan and Golden, 1966]. Cette méthode est devenue célèbre pour ses possibilités de manipulation du son [Laroche and Dolson, 1999], [Puckette, 1995], [Moulines and Laroche, 1995] et son utilisation comme instrument de musique électronique [Grey and Moorer, 1977]. En particulier, Portnoff a proposé une implantation très efficace, basée sur la transformée de Fourier rapide, qui a permis un grand nombre d'applications temps réel [Portnoff, 1981].

Historiquement, le premier estimateur de fréquence basé sur la TFCT est celui du vocodeur de phase. C'est également la première méthode cherchant à extraire les paramètres sinusoïdaux dans une étape d'analyse. La fréquence instantanée est calculée grâce à une dérivation discrète de la phase, comme une approximation de la

définition continue de la fréquence[1]. Si $\Phi(t)$ est la fonction de phase de la sinusoïde, la fréquence correspondante est définie par :

$$\Omega(t) = \frac{\partial \Phi}{\partial t}(t) \tag{4.1.7}$$

Le modèle **M01** a deux propriétés particulières qui le rendent particulièrement facile à manipuler. Tout d'abord, lorsque la fréquence est constante, la dérivation discrète de la phase est un équivalent exacte de la dérivation continue :

$$\Omega(t) = \beta = \frac{\alpha_2 - \alpha_1}{T} \tag{4.1.8}$$

où $\alpha_i$ est la phase de la sinusoïde à un instant $t_i$ et $T = t_2 - t_1$ est le pas d'avancement entre les deux instants d'estimation. $\beta$, la fréquence, est constante pour tout $t$. La deuxième propriété du modèle **M01** est que la phase de la sinusoïde coïncide exactement avec la phase de la transformée de Fourier du signal. Si on suppose que la sinusoïde est bien résolue[2] fréquentiellement autour du bin $k$, on a donc $\hat{\alpha}_i = \arg(X(t_i, \omega_k; h))$ et la fréquence $\beta$ peut s'estimer ainsi :

$$\hat{\beta} = \frac{\arg(X(t_2, \omega_k; h)) - \arg(X(t_1, \omega_k; h))}{T} \tag{4.1.9}$$

Cette méthode ne fait aucune approximation et ne va présenter aucun biais pour le modèle **M01**.

Le réglage de l'intervalle $T$ est un point délicat de la méthode. A l'origine, le vocodeur de phase décrit dans [Flanagan and Golden, 1966] utilise un intervalle égal à un échantillon, $T = 1/F$. Dans ce cas, l'estimation de fréquence demande deux TFCT adjacentes, sauf dans le cas de fenêtres très particulières comme les fenêtres rectangulaire et de Hann, où la deuxième TFCT peut être déduite de façon récursive, à partir de la première [Brown and Puckette, 1993]. Quand la différence de phase est effectuée sur des échantillons consécutifs, on constate également que les estimations de fréquence présentent une variance importante. La variance peut être réduite en utilisant des intervalles plus grands [Puckette and Brown, 1998], et on parle alors de vocodeur de phase à long terme. Un problème d'indétermination de phase peut alors apparaître lorsque l'incrément de phase $\beta T$ entre deux TFCT consécutives est plus grand que $2\pi$. L'estimation de fréquence s'exprime alors ainsi

$$\hat{\beta} = \frac{\Delta X + 2\pi n}{T} \tag{4.1.10}$$

où $n$ est un entier calculé en pratique par déroulement de phase. Le déroulement de phase consiste à "suivre" la sinusoïde avec une autre sinusoïde de fréquence connue

[1] Un autre usage répandu du terme vocodeur de phase fait référence à cet estimateur de fréquence particulier.

[2] Le terme "bien résolu" est défini à la section 3.3.2.

$\Omega$ et telle que la différence d'accroissement de phase entre les deux sinusoïdes est inférieur à une période. Cette condition peut s'écrire ainsi :

$$|\Omega - \beta| < \frac{\pi}{T} \tag{4.1.11}$$

lorsque cette condition est respectée, on pourra donc estimer $n$ avec cette formule [McAulay and Quatieri, 1986] :

$$\hat{n} = \text{round}\big((\Omega T - \Delta X)/(2\pi)\big) \tag{4.1.12}$$

En général on va choisir $\Omega$ comme étant la fréquence du bin $\omega_k$ le plus proche de $\beta$. Dans ce cas on sait que $|\omega_k - \beta| \leq \frac{\pi F}{N}$, et on peut en déduire la condition suivante sur $T$ :

$$T < \tau_N \tag{4.1.13}$$

Cette condition est différente de celle que l'on peut trouver dans les articles [Puckette, 1995], [Laroche and Dolson, 1999]. En effet, à l'origine, le vocodeur de phase était appliqué sur tous les bins et en particulier on exigeait que la condition (4.1.11) soit respectée pour tous les bins du lobe principal de spectre d'amplitude. En notant $\omega_h$ la fréquence de coupure de la fenêtre d'analyse, on aura donc $|\Omega - \beta| < \omega_h$ et on peut en déduire la condition suivante sur $T$ :

$$T < \frac{\pi}{\omega_h} \tag{4.1.14}$$

Pour les fenêtres d'analyse usuelles (*e.g.* Hann, Hamming), la fréquence de coupure est quasiment proportionnelle à la longueur de la fenêtre d'analyse $\omega_h = 4\pi/\tau_N$ [Puckette, 1995]. La condition (4.1.14) se simplifie alors en : $T < \tau_N/4$. Dans ce cas de figure, le minimum de recouvrement des fenêtres est de 75%, ce qui est bien plus contraignant que dans le cas précédent.

#### 4.1.2.2 La Méthode de la dérivée

Cette méthode a été développée récemment par Sylvain Marchand [Desainte-Catherine and Marchand, 2000], [Hofbauer, 2004]. C'est une méthode basée sur le temps, qui utilise deux transformées de Fourier successives, comme la méthode du vocodeur de phase. La démonstration d'origine de cette méthode fait intervenir la dérivation discrète du signal. Une autre démonstration est préférée ici, qui montre la forte relation qu'il y a avec le vocodeur de phase.

On note $R$ le rapport :

$$R = \frac{X(t+T, \omega; h)}{X(t, \omega; h)} \tag{4.1.15}$$

On rappelle que l'estimateur de fréquence utilisé par le vocodeur de phase peut s'exprimer ainsi :

$$\hat{\beta} = \frac{\arg(R)}{T} \tag{4.1.16}$$

$\arg(R)$ est une phase déroulée et $\beta \geq 0$, donc $\arg(R) \geq 0$.

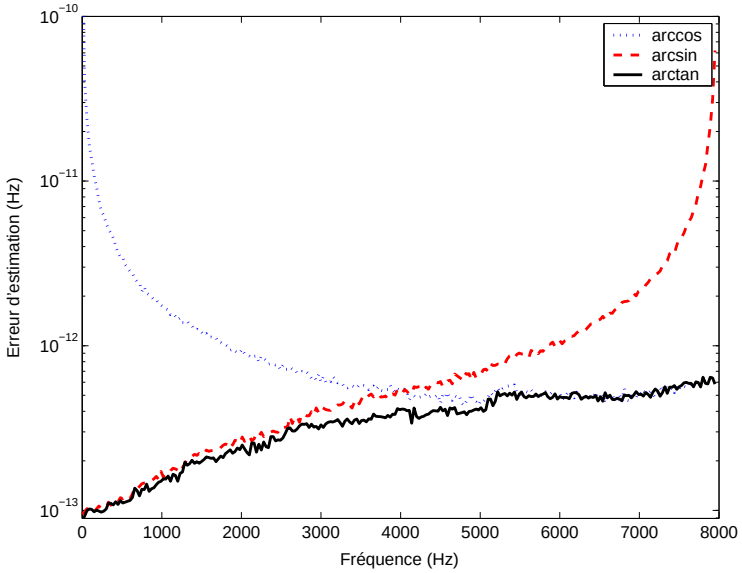



FIG. 4.1: Erreur des estimateurs arccos, arcsin et arctan en fonction de la fréquence.

On considère maintenant le rapport suivant, qui est celui de la méthode de la dérivée :

$$R_{\sin} = \frac{X(t+T,\omega;h) - X(t,\omega;h)}{X(t,\omega;h)} \tag{4.1.17}$$

$$R_{\sin} = R - 1 \tag{4.1.18}$$

Comme l'amplitude et la fréquence sont constants, on sait que $|R| = 1$. En utilisant les propriétés trigonométriques, on peut montrer que le vocodeur de phase et la méthode de la dérivée estiment le même angle :

$$R_{\sin} = \cos(\arg(R)) - 1 + j\sin(\arg(R)) \tag{4.1.19}$$

$$|R_{\sin}|^2 = 2(1 - \cos(\arg(R))) \tag{4.1.20}$$

$$|R_{\sin}|^2 = 4\sin^2\Big(\frac{\arg(R)}{2}\Big) \tag{4.1.21}$$

$$\arg(R) = 2\arcsin\Big(\frac{|R_{\sin}|}{2}\Big) \tag{4.1.22}$$

On en déduit donc l'estimateur de fréquence suivant :

$$\hat{\beta} = \frac{2}{T}\arcsin\big(\frac{|R_{\sin}|}{2}\big) \tag{4.1.23}$$

De la même manière, on aurait pu considérer les rapports :

$$R_{\cos} = R + 1 \tag{4.1.24}$$

$$R_{\tan} = \frac{X(t+T,\omega_k;h) - X(t,\omega_k;h)}{X(t,\omega_k;h) + X(t+T,\omega_k;h)} \tag{4.1.25}$$

Ces rapports permettent également d'estimer le même angle $\arg(R)$ et nous conduisent respectivement aux estimateurs suivants :

$$\hat{\beta} = \frac{2}{T}\arccos\left(\frac{|R_{\cos}|}{2}\right) \tag{4.1.26}$$

$$\hat{\beta} = \frac{1}{T}\arctan\left(\frac{|R_{\tan}|}{2}\right) \tag{4.1.27}$$

Ces nouveaux estimateurs correspondent aux trois façons possibles de calculer un angle, en utilisant soit le cosinus, le sinus ou la tangente. Les propriétés statistiques de ces différents estimateurs vont cependant être assez différentes, dépendant essentiellement de la sensibilité des fonctions arccos, arcsin et arctan, la dernière fonction étant plus stable que les deux autres [Betser et al., 2006a]. C'est ce que l'on peut voir sur la Figure 4.1 : les trois méthodes sont non biaisées, et atteignent presque la précision machine, cependant on peut voir que pour le sinus et le cosinus l'erreur devient beaucoup plus importante pour les hautes fréquences pour le premier et les basses fréquences pour le second. Enfin il faut souligner que le vocodeur de phase est aussi un estimateur qui utilise la tangente et va présenter des résultats quasiment identiques au dernier estimateur.

#### 4.1.2.3 Interpolation du spectre de Fourier discret

Cette méthode est une méthode à temps constant, faisant varier le paramètre de fréquence uniquement et ne nécessite donc qu'une seule transformée de Fourier discrète. Le principe de cette méthode est d'interpoler le spectre d'amplitude discret pour trouver le maximum. Ces méthodes utilisent une fonction d'interpolation, comme le triangle ou la parabole, qu'elles vont ajuster au spectre d'amplitude. La réponse fréquentielle d'une sinusoïde fenêtrée doit bien sûr être cohérente avec la fonction choisie.

L'interpolation triangulaire a été développée par Keiler et Zölzer [Keiler and Zölzer, 2001]. Ils proposent d'utiliser une fenêtre ayant une réponse fréquentielle presque triangulaire, l'interpolation utilisant deux segments de droite. Les paramètres des droites sont estimés en utilisant un critère des moindres carrés et la fréquence du maximum est calculée comme l'intersection de ces deux droites. Pour une description complète de l'algorithme, on pourra consulter [Keiler and Zölzer, 2001]. Un filtre triangulaire dans le domaine fréquentiel ne possède pas de support temporel fini. La fenêtre va donc être tronquée dans le domaine temporel, conduisant à une réponse fréquentielle approximativement triangulaire (cf. Figure 4.2). La fenêtre triangulaire utilisée par les auteurs est en fait un cas particulier des fenêtres de Blackman et pour une estimation basée sur deux bins on retombe sur la fenêtre de Hann[3].

L'interpolation parabolique [Smith III and Serra, 1987], [Abe and Smith III, 2004] consiste à interpoler le spectre d'amplitude avec une parabole. Plusieurs types de fenêtres peuvent être utilisées avec cette méthode, comme la fenêtre de Hann ou la fenêtre Gaussienne tronquée. En échelle logarithmique, ces fenêtres sont paraboliques

[3] Voir l'annexe B pour une explication plus détaillée.

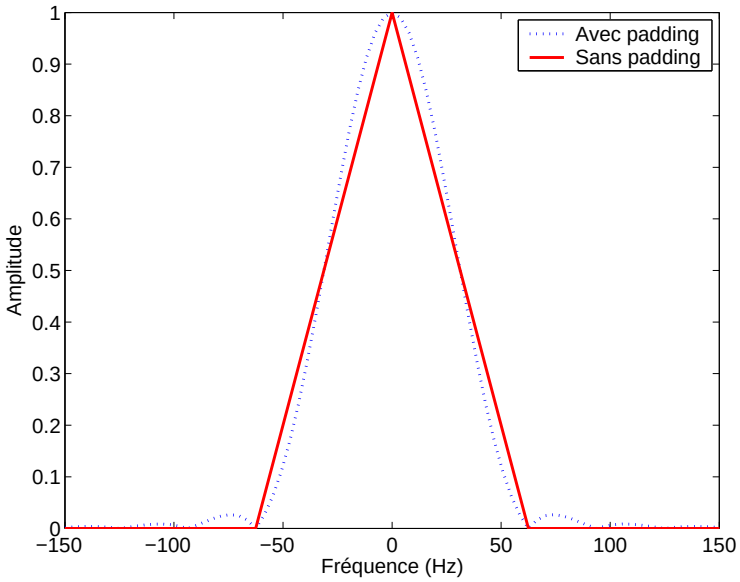



FIG. 4.2: La fenêtre à réponse triangulaire tronquée pour une interpolation à 2 échantillons (fenêtre de Hann)

au voisinage du maximum. Pour avoir une meilleure approximation parabolique, le zéro-padding[4] est souvent utilisé avec ces méthodes. Si $x$ est un sinus complexe avec une fréquence $\beta$ et si $p$ est l'index du maximum du spectre d'amplitude discret,

$$\hat{\beta} = 2\pi F/N(p + \Delta) \tag{4.1.28}$$

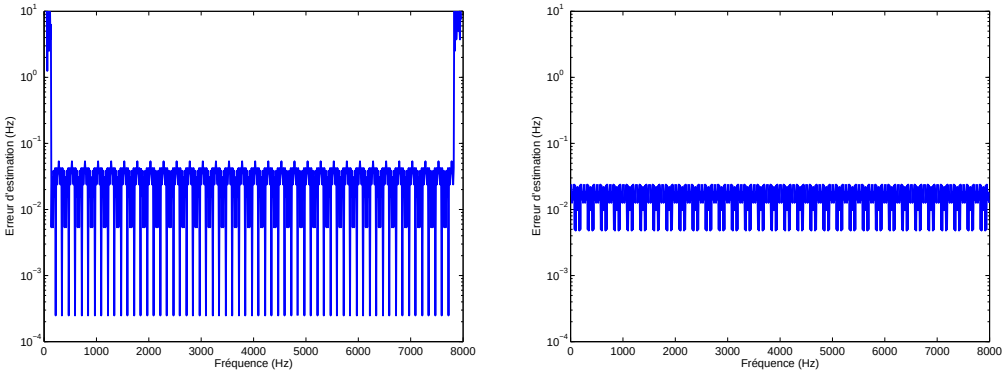
et $\Delta$ est donné par l'interpolation parabolique :

$$\hat{\Delta} = 0.5\frac{X_{db}(p-1) - X_{db}(p+1)}{X_{db}(p-1) - 2X_{db}(p) + X_{db}(p+1)} \tag{4.1.29}$$

où $X_{db}(k) = 20\log_{10}(|X(k)|)$. Les performances de cette méthode dépendent fortement du type de fenêtre utilisé. Certaines fenêtres ont été spécialement conçues pour cette méthode [Keiler and Marchand, 2002]. Une étude complète de l'influence des différents paramètres, type de fenêtre, longueur de fenêtre, facteur de zéro-padding, peut être trouvée dans [Abe and Smith III, 2004].

La Figure 4.3 propose une comparaison de ces estimateurs sur toute la plage de fréquence étudiée. La méthode de l'interpolation parabolique nécessite un fort coefficient de padding, ici 4. Pour la méthode triangulaire, les performances se dégradent si on ajoute du zéro-padding, car alors les segments de droites présentent des perturbations dues à la troncature de la fenêtre. L'erreur de cette méthode (Figure 4.3(a)) augmente fortement près des fréquences zéro et $F/2$ car on a moins de points pour estimer les droites. Pour les deux méthodes, la précision de l'estimation est meilleure

[4]Ce terme est défini à la section 3.3.6.1.




(a) Interpolation triangulaire (b) Interpolation quadratique

FIG. 4.3: Erreur des estimateurs d'interpolation en fonction la fréquence.

lorsque le bin maximum tombe sur un bin FFT et décroit lorsqu'il tombe entre deux bins, ce qui explique les variations périodiques sur les figures.

#### 4.1.2.4 Interpolation utilisant l'information de phase

Ces méthodes utilisent la transformée de Fourier complexe au lieu du spectre d'amplitude. Contrairement aux méthodes précédentes, on ne cherche pas ici à interpoler une fonction, mais on utilise des propriétés de la transformée de Fourier pour obtenir un estimateur de fréquence. Ces estimateurs ont été introduits par Quinn [Quinn, 1994], et ont été généralisés par MacLeod [Macleod, 1998]. Une comparaison avec les autres estimateurs de la littérature peut être trouvée dans [Hainsworth and Macleod, 2003a].

Le développement de ces méthodes repose sur l'existence d'une forme analytique simple de la transformée de Fourier d'un signal sinusoïdal **M01** pondéré par une fenêtre rectangulaire.

$$X(t,\omega_i;1) = A\exp(j\alpha)\frac{1-\mathrm{e}^{j(\beta-\omega_i)\frac{N}{F}}}{1-\mathrm{e}^{j(\beta-\omega_i)\frac{1}{F}}} \tag{4.1.30}$$

$\omega_i$ est un bin FFT, donc $\mathrm{e}^{j\omega_i N/F} = 1$. Si on considère que $(\beta-\omega_i)\frac{1}{F} << 1$, on peut utiliser l'approximation pour les petits angles de l'exponentielle, et on en déduit l'approximation suivante [Aboutanios and Mulgrew, 2005] :

$$X(t,\omega_i;1) \approx \frac{b}{\beta-\omega_i} \tag{4.1.31}$$

où $b$ est une constante complexe indépendante de la fréquence $\omega_i$.

La première étape consiste à choisir une fréquence de référence $\omega_k$, généralement la fréquence la plus proche du bin maximum du spectre d'amplitude. On note $R(m) = X(\omega_{k+m})$. Ensuite on élimine la constante $b$ en appliquant l'approximation (4.1.31)

sur plusieurs bins, dans le voisinage de $k$, afin d'obtenir une fonction ne dépendant que de $\beta$.

Quinn [Quinn, 1994] propose d'utiliser les estimateurs suivants[5] :

$$\hat{\beta}_+ = \omega_k - \Re\Big(\frac{R(1)}{R(0)-R(1)}\Big) \tag{4.1.32}$$

$$\hat{\beta}_- = \omega_k + \Re\Big(\frac{R(-1)}{R(0)-R(-1)}\Big) \tag{4.1.33}$$

MacLeod [Macleod, 1998] suggère de combiner ces deux estimateurs en utilisant ce test :

$$\hat{\beta} = \begin{cases} \hat{\beta}_+ & \text{si } |R(-1)| > |R(1)|, \\ \hat{\beta}_- & \text{sinon.} \end{cases} \tag{4.1.34}$$

Dans le même article, Macleod propose également des estimateurs utilisant trois et cinq échantillons :

$$\gamma_3 = \Re\Big(\frac{R(-1)-R(1)}{2R(0)+R(-1)+R(1)}\Big) \tag{4.1.35}$$

$$\gamma_5 = \Re\Big(\frac{4(R(-1)-R(1))+2(R(-2)-R(2))}{12R(0)+8(R(-1)+R(1))+R(-2)+R(2)}\Big) \tag{4.1.36}$$

$$\hat{\beta}_3 = \omega_k + \frac{(\sqrt{1+8\gamma_3^2}-1)}{4\gamma_3} \tag{4.1.37}$$

$$\hat{\beta}_5 \approx \omega_k + 0.4041 \arctan(2.93\gamma_5) \tag{4.1.38}$$

Certains auteurs, au lieu de combiner directement les bins FFTs, définissent $R$ par[6] :

$$R(m) = \Re(X(\omega_m).\bar{X}(\omega_k)) \tag{4.1.39}$$

$$R(m) \approx \frac{|b|^2}{(\beta-\omega_{k+m})(\beta-\omega_k)} \tag{4.1.40}$$

$\frac{|b|^2}{(\beta-\omega_k)}$ sera éliminé par le rapport de transformées de Fourier, et on retrouve les même estimateurs que précédemment.

MacLeod compare ces estimateurs de façon exhaustive dans [Macleod, 1998]. L'estimateur à cinq échantillons n'apportant que peu d'améliorations par rapport à l'estimateur à trois échantillons, seul ce dernier sera retenu pour les comparaisons.

[5] D'après la formule (4.1.31), le rapport $\frac{R(1)}{R(0)-R(1)}$ devrait être réel, mais comme cette formule n'est qu'une approximation, il est préférable de considérer la partie réelle du rapport. Idem pour le rapport $\frac{R(-1)}{R(0)-R(-1)}$.

[6] Comme précédemment $X(\omega_m).\bar{X}(\omega_k)$ devrait être réel, mais comme cette formule n'est qu'une approximation, il est préférable de considérer la partie réelle.

## 4.2 Estimation pour le modèle M02

Dans cette section, les méthodes d'estimation pour le modèle de phase quadratique, le modèle **M02**, sont passées en revue. Ce modèle est très souvent utilisé dans des problèmes d'estimation de fréquence. Il suppose une fréquence localement linéaire (phase quadratique) et une amplitude constante :

$$x(\tau) \triangleq A\,\mathrm{e}^{j(\alpha_M+\beta_M\tau+\gamma\tau^2/2)} \tag{4.2.1}$$

où $\alpha_M$, $\beta_M$, et $\gamma$ sont respectivement la phase, la fréquence et le taux de variation de la fréquence, pour l'instant $t_M$. Afin de souligner que l'amplitude et le taux de variation sont constants à l'intérieur de l'intervalle d'analyse, l'indice $M$ a été supprimé pour ces deux paramètres.

### 4.2.1 Estimation de fréquence

#### 4.2.1.1 Méthode du réassignement spectral

Le réassignement spectral est une méthode initiée par Kodera, Gendrin et Villedary pour améliorer la lisibilité des spectrogrammes [Kodera et al., 1978]. Le but est d'obtenir une représentation très bien localisée à la fois en temps et en fréquence pour des signaux usuels, à partir d'un spectre échantillonné régulièrement en fréquence. Les signaux couverts par cette représentation sont les sinusoïdes, les chirps linéaires. La méthode propose de dessiner le spectrogramme en déplaçant les échantillons discrets du spectre au centre de gravité de leur contributions d'énergie, c'est à dire au point :

$$\begin{aligned}\hat{\beta}(t,\omega) &= \frac{\partial}{\partial t}\arg(X(t,\omega;h))\\ \hat{t}(t,\omega) &= t-\frac{\partial}{\partial \omega}\arg(X(t,\omega;h))\end{aligned}$$

La méthode originelle propose de calculer ces différences continues par des approximations discrètes :

$$\begin{aligned}\hat{\beta}(t,\omega) &\approx \frac{1}{2\delta_t}\arg(X(t+\delta_t,\omega;h)\bar{X}(t-\delta_t,\omega;h))\\ \hat{t}(t,\omega) &\approx t-\frac{1}{2\delta_\omega}\arg(X(t,\omega+\delta_\omega;h)\bar{X}(t,\omega-\delta_\omega;h))\end{aligned}$$

En général pour le calcul de la fréquence, on utilisera deux transformées de Fourier séparées d'un échantillon. On peut voir ici que la méthode des différences finies, utilisée dans le vocodeur de phase est une approximation de l'estimateur de fréquence du réassignement.

Auger et Flandrin [Auger and Flandrin, 1995] ont reformulé cette méthode et l'ont généralisée pour tous les types de transformations bilinéaires. Ils ont notamment prouvé que cette transformation permet de localiser parfaitement les chirps. Ils ont

également proposé un autre algorithme basé sur la TFCT, en utilisant la fenêtre dérivée, $\dot{h}$, et la fenêtre multipliée par une rampe temporelle, $\tau h$ :

$$\begin{aligned} \hat{\beta}(t,\omega) &= t + \Re\Big(\frac{X(t,\omega;\dot{h})}{X(t,\omega;h)}\Big) \\ \hat{t}(t,\omega) &= \omega - \Im\Big(\frac{X(t,\omega;\tau h)}{X(t,\omega;h)}\Big) \end{aligned} \tag{4.2.2}$$

Une démonstration originale de cet estimateur pour le chirp linéaire est donnée au paragraphe suivant.

Les signes devant les opérateurs partie réelle et partie imaginaire peuvent varier suivant les auteurs, dépendant de la définition donnée à la TFCT. Si le signal analysé possède une fréquence constante, le temps d'estimation n'a plus d'importance, car la fréquence est la même sur toute la fenêtre d'analyse. Si le signal possède une fréquence variable, le temps d'estimation ne peut être omis, mais pour une fréquence qui varie suffisamment lentement, ce temps est à peu près égal à $t_M$, le temps du milieu de la fenêtre.

La méthode d'Auger et Flandrin nécessite trois TFCT différentes pour calculer la fréquence, plus que la méthode des différences finies qui n'en demande que deux. Cependant, la méthode des différences finies requière un coefficient de zéro padding plus élevé que la méthode d'Auger et Flandrin pour une précision équivalente. Pour la fenêtre de Hann, il existe une approximation ne requérant qu'une TFCT [Hainsworth, 2004], mais cette approximation reste assez grossière. Pour la fenêtre Gaussienne, les fenêtres $\dot{h}$ et $\tau h$ sont identiques, à une constante multiplicative près, donc la méthode d'Auger et Flandrin ne nécessite que deux TFCTs pour cette fenêtre[7].

La Figure 4.4 montre l'erreur du réassignement en fonction de la variation de fréquence, pour les méthodes d'Auger et Flandrin et pour la méthode des différences finies. Le biais de la première méthode est très faible, inférieur à $10^{-3}$. L'erreur croit légèrement en fonction de la variation de fréquence. Celui de la méthode des différences finies est beaucoup plus important, il est dû essentiellement à l'approximation de la dérivation fréquentielle, c'est à dire à la limitation de précision de la transformée de Fourier. Sur cette figure, les transformées de Fourier pour la méthode des différences finies ont été calculées avec un coefficient de padding de 4, contre un coefficient de 1 pour l'autre méthode. La méthode d'Auger et Flandrin est donc à la fois plus précise et plus rapide en temps de calcul car il est moins long de calculer trois transformées de longueur N que deux transformées de longueur 4N. Seule la méthode d'Auger et Flandrin sera comparée par la suite.

Nous proposons maintenant une démonstration originale de la validité de l'estimateur du réassignement dans le cas d'un chirp. L'intérêt de cette démonstration est de montrer la relation qu'il y a entre temps et fréquence réassignés, qui, du point de vue de l'estimation, sont indissociables. En effet lorsque la fréquence varie, on va estimer une fréquence (la fréquence réassignée) pour un temps donné (le temps réassigné).

[7] Voir l'annexe B pour plus de détails.

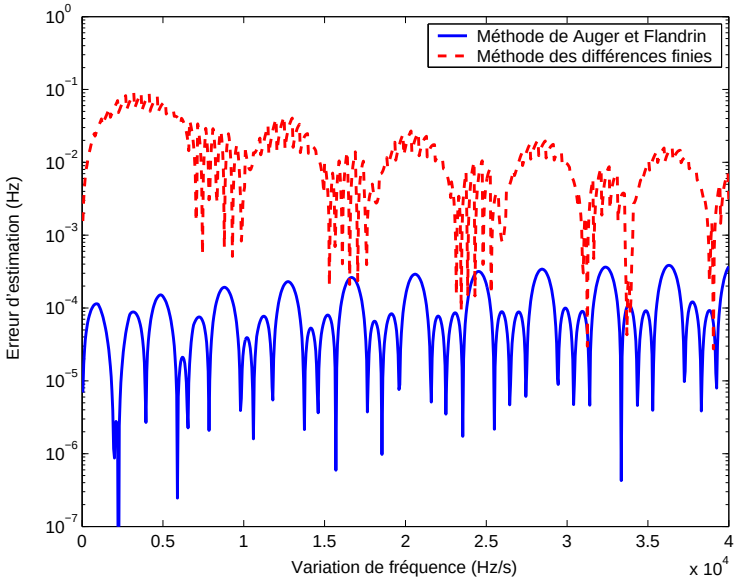



FIG. 4.4: Erreur du réassignement en fonction de la variation de fréquence.

### Une démonstration pour le réassignement d'un chirp avec la TFCT

Cette démonstration a été publiée dans [Betser et al., 2006a]. On utilise une formulation fréquentielle par souci de concision. La transformée de Fourier $X_c$ de $x$ est ici continue et définie par (3.3.3). Soit $f(\tau) = x(\tau).h(\tau)$. Le signal $x$ suit toujours le modèle **M02** donné par l'équation (4.2.1) et $h$ est une fenêtre continue, dérivable, et à support temporel fini.

$$\frac{df}{d\tau}(\tau) = \frac{dx}{d\tau}(\tau)h(\tau) + \dot{h}(\tau)x(\tau) \tag{4.2.3}$$

$$\frac{df}{d\tau}(\tau) = j(\beta + \gamma\tau)x(\tau)h(\tau) + \dot{h}(\tau)x(\tau) \tag{4.2.4}$$

La transformée de Fourier continue est alors appliquée à cette relation pour une fréquence $\omega$ telle que la sinusoïde ait une énergie non nulle pour cette fréquence particulière ($X_c(t,\omega;h) \neq 0$).

En utilisant la propriété de dérivation[8], on obtient :

$$j\omega X_c(t,\omega;h) = j\beta X_c(t,\omega;h) + j\gamma X_c(t,\omega;\tau h) + X_c(t,\omega;\dot{h}) \tag{4.2.5}$$

$$\beta + \gamma\frac{X_c(t,\omega;\tau h)}{X_c(t,\omega;h)} = \omega + j\frac{X_c(t,\omega;\dot{h})}{X_c(t,\omega;h)} \tag{4.2.6}$$

C'est la partie réelle de cette dernière équation qui nous intéresse. Le terme de gauche est la fréquence de la sinusoïde pour le temps

$$\hat{t} = t + \Re\Big(\frac{X_c(t,\omega;\tau h)}{X_c(t,\omega;h)}\Big), \tag{4.2.7}$$

[8] cf. Annexe A pour un rappel des propriétés de la TF.

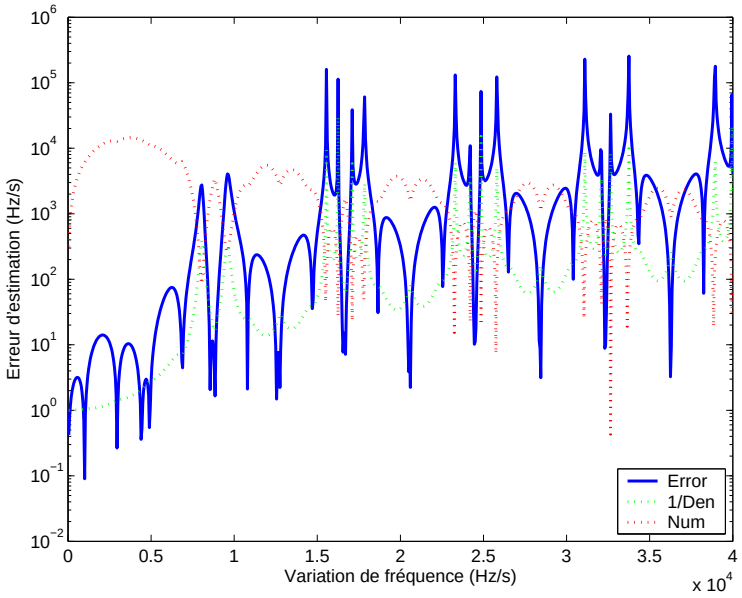



FIG. 4.5: Erreur de l'estimateur de Röbel en fonction de la variation de fréquence. Num et 1/Den sont respectivement le numérateur et l'inverse du dénominateur de l'estimateur (4.2.9)

c'est à dire le temps réassigné. La seconde partie de l'équation correspond à l'opérateur de fréquence réassigné :

$$\hat{\beta} = \omega - \Im\Big(\frac{X_c(t,\omega;\dot{h})}{X_c(t,\omega;h)}\Big). \tag{4.2.8}$$

On voit donc clairement que temps réassigné et fréquence réassignée sont simultanés, on estime une fréquence à un temps donné.

La version discrète de la méthode du réassignement a été rappelée à l'équation (4.2.2). La discrétisation introduit un léger biais dans l'estimation [Hainsworth and Macleod, 2003b], [Betser et al., 2006a]. En fait la démonstration précédente n'est pas rigoureusement valide pour une TF discrète car il n'existe pas d'équivalent discret de la propriété de dérivation temporelle de la TF continue.

### 4.2.2 Estimation de la variation de fréquence

#### 4.2.2.1 Estimateur de Röbel

L'estimateur de Röbel est basé sur le réassignement fréquentiel. Pour un bin fréquentiel donné, $\omega$, les équations du réassignement (4.2.8) et (4.2.7) sont des fonctions

du temps :

$$\hat{t}(t) = t + \Re\left(\frac{X_c(t,\omega;h.t)}{X_c(t,\omega;h)}\right)$$
$$\hat{\beta}(t) = \omega - \Im\left(\frac{X_c(t,\omega;\dot{h})}{X_c(t,\omega;h)}\right)$$

Or on sait que $\hat{\beta}$ est la fréquence au temps $\hat{t}$, donc par définition la variation de fréquence est la dérivée de la fréquence réassignée par rapport au temps réassigné :

$$\hat{\gamma} = \frac{d\hat{\beta}}{d\hat{t}} = \frac{\frac{d\hat{\beta}}{dt}}{\frac{d\hat{t}}{dt}} \tag{4.2.9}$$

En utilisant la propriété de dérivation temporelle de la transformée de Fourier[9], puis en approchant les transformées de Fourier continues par des transformées discrètes, on montre en outre que :

$$\frac{d\hat{\beta}}{dt} \approx -\Re\Big(\frac{X(t,\omega;\dot{h}.t)}{X(t,\omega;h)}\Big) + \Re\Big(\frac{X(t,\omega;\dot{h})X(t,\omega;h.t)}{X(t,\omega;h)^2}\Big)$$
$$\frac{d\hat{t}}{dt} \approx \Im\Big(\frac{X(t,\omega;\ddot{h})}{X(t,\omega;h)}\Big) - \Im\Big(\big(\frac{X(t,\omega;\dot{h})}{X(t,\omega;h)}\big)^2\Big)$$

Dans cette méthode, on doit calculer cinq transformées de Fourier, pour les fenêtres $h$, $\tau h$, $\dot{h}$, $t.\dot{h}$ et $\ddot{h}$. La méthode est donc beaucoup plus complexe que toutes les autres méthodes d'estimation de $\gamma$ présentées dans cet ouvrage. D'autre part, elle utilise plusieurs fois la propriété de dérivation temporelle, qui n'existe pas dans le cas discret[10]. L'erreur résultante se révèle très importante pour certaines valeurs de variation de fréquence, comme l'illustre la Figure 4.5. Pour certaines valeurs de variation de fréquence, on voit que $\frac{d\hat{\beta}}{dt}$ et $\frac{d\hat{t}}{dt}$ se rapprochent tous deux de zéro. L'erreur dû aux approximations devient alors très importante.

#### 4.2.2.2 Estimateurs de Master

Dans deux articles [Master, 2002], [Master and Liu, 2003a], Master décrit un certain nombre d'estimateurs tous dédiés à l'estimation de variation de fréquence pour le modèle **M02**. Ces estimateurs se rangent en deux catégories : ceux découlant de l'approximation des intégrales de Fresnel, et ceux découlant d'une approximation de Taylor. Ces deux méthodes sont maintenant présentées rapidement.

[9] Voir la table A.1 dans l'annexe A.
[10] Voir annexe A.

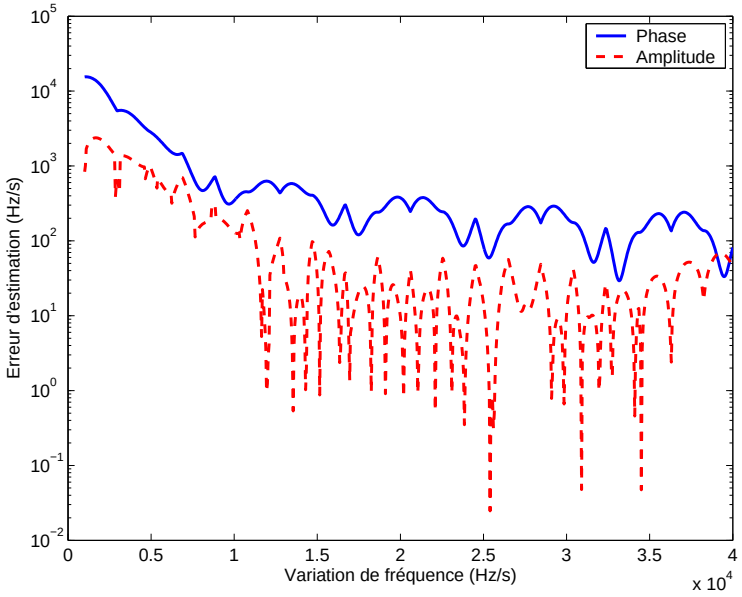



FIG. 4.6: Erreur des estimateurs de Master dérivés des intégrales de Fresnel, en fonction de la variation de fréquence. 'Phase' correspond à l'estimateur (4.2.13), 'amplitude' à l'estimateur (4.2.14).

### Approximation des intégrales de Fresnel

Les intégrales de Fresnel sont définies par :

$$C(u) \triangleq \int_0^u \cos(\frac{\pi}{2}x^2)dx$$
$$S(u) \triangleq \int_0^u \sin(\frac{\pi}{2}x^2)dx$$

Une bonne approximation asymptotique ($u \gg 1$) est donnée par l'expression :

$$C(u) \approx \frac{1}{2} + \frac{1}{\pi u}\cos(\frac{\pi}{2}u^2)$$
$$S(u) \approx \frac{1}{2} + \frac{1}{\pi u}\sin(\frac{\pi}{2}u^2) \qquad (4.2.10)$$

Si l'on considère la transformée de Fourier discrète du signal **M02**, pour une fenêtre rectangulaire, nous avons :

$$X(t,\omega;h_{\text{rec}}) = A\,\mathrm{e}^{j\alpha}\sum_{-(N-1)/2}^{(N-1)/2}\exp(j(\beta-\omega)t + j\frac{\gamma}{2}t^2)$$

En approchant tout d'abord $X(t,\omega;h_{\text{rec}})$ par son équivalent continu, puis en effectuant un changement de variable, on peut exprimer $X$ comme une somme d'intégrales

de Fresnel. Chaque intégrale sera remplacée par l'approximation (4.2.10). L'expression résultante est longue et inutilisable telle quelle, c'est pourquoi elle n'est pas reproduite ici. Master propose alors de considérer la fenêtre de Hann. En effet, la transformée de Fourier d'une fenêtre de Hann peut s'exprimer comme une somme de trois transformées de fenêtres rectangulaire[11], et cette combinaison va conduire à une approximation beaucoup plus simple :

$$\arg(X(t,\omega;h_{\text{han}})) \approx \alpha + \text{sgn}(\gamma)\frac{\pi}{4} - \frac{(\omega-\beta)^2}{2\gamma} \tag{4.2.11}$$

$$|X(t,\omega;h_{\text{han}})| \approx \frac{A\sqrt{\pi}}{2\sqrt{|\gamma|}}\Big(1+\cos\big((\omega-\beta)\frac{2\pi F}{N\gamma}\big)\Big) \tag{4.2.12}$$

Dans notre cas, les approximations des intégrales de Fresnel seront valides seulement si $N$ ou $\gamma$ sont suffisamment grands. Lorsque $\omega$ est au voisinage de $\beta$, cette condition peut se résumer ainsi :

$$\gamma\tau_N^2 \gg 4\pi$$

Par exemple, si l'on choisit $\tau_N = 32$ms, alors on ne pourra estimer que des valeurs de $\gamma$ telles que $\gamma \gg 2000$Hz/s, *i.e.* à partir de 10000 Hz/s environ. Cette contrainte réduit beaucoup l'intérêt de la méthode pour l'étude des signaux de parole et de musique, dont les variations sont souvent plus faibles.

Cette approximation nous conduit à deux types d'estimateurs possibles, basés sur la phase ou sur l'amplitude. Pour l'estimation basée sur la phase on peut éliminer $\alpha$ et $\beta$ en dérivant deux fois l'équation (4.2.11) par rapport à $\omega$. On approche ensuite la dérivation continue par une dérivation discrète et on obtient l'estimateur :

$$\hat{\gamma} = \left(\frac{\arg(X(\omega_0)^2\bar{X}(\omega_0+\delta_\omega)\bar{X}(\omega_0-\delta_\omega))}{\delta_\omega^2}\right)^{-1} \tag{4.2.13}$$

Par exemple, on peut choisir pour l'estimation le bin maximum pour $\omega_0$ et ses deux bins adjacents. Dans ce cas, $\delta_\omega = 2\pi F/P$ est la précision de Fourier.

Pour l'estimation basée sur l'amplitude, il faut remarquer que si $\omega_{h/2}$ est la fréquence correspondante exactement à la moitié de la hauteur du pic d'amplitude, on doit avoir $|X(\omega_{h/2},t;h_{\text{han}})| = \frac{A\sqrt{\pi}}{2\sqrt{|\gamma|}}$, ce qui implique que le terme en cosinus de l'équation (4.2.12) doit s'annuler pour cette valeur. On en déduit l'estimateur suivant :

$$\hat{\gamma} = \text{sgn}(\gamma)|\omega_{h/2}-\beta|\frac{4F}{N} \tag{4.2.14}$$

D'après cette formule, si $\gamma$ est suffisamment grand, alors l'étalement du lobe principal de la TF est directement proportionnel à $\gamma$, ce qui est un résultat assez remarquable. La première étape de cette méthode est donc de trouver la fréquence correspondant à la mi-hauteur du pic. Afin d'avoir une fréquence aussi précise que possible, on va interpoler linéairement la valeur entre les deux bins les plus proches de la mi-hauteur.

[11] Voir l'annexe B pour plus de détails.

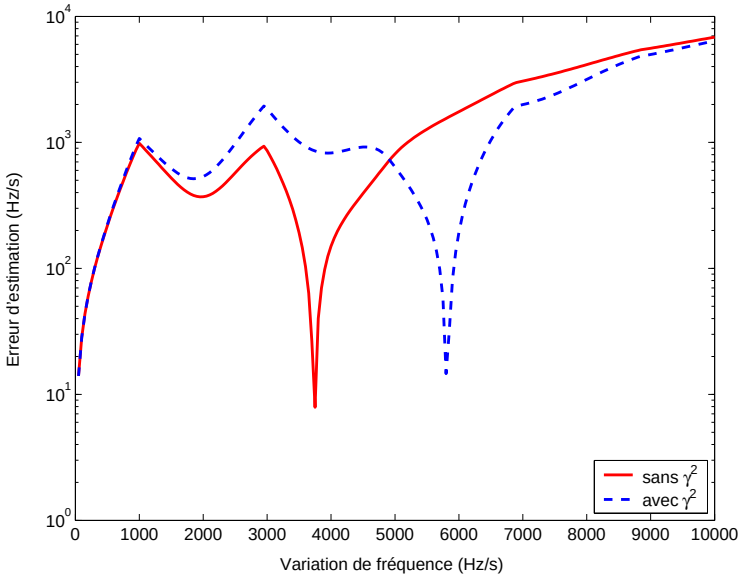



FIG. 4.7: Erreur de l'estimateur de Master dérivé de l'approximation de Taylor en fonction de la variation de fréquence. Il est décliné en deux versions : avec et sans le terme d'ordre 2 (équation (4.2.16)).

La deuxième étape consiste à déterminer le signe de $\gamma$ : il suffit de regarder la courbure de la phase au voisinage du maximum, *i.e.* le signe de $\arg(X(\omega_0)^2\bar{X}(\omega_0+\delta_\omega)\bar{X}(\omega_0-\delta_\omega))$.

Le fait de devoir déterminer la fréquence de mi-hauteur va rendre l'estimateur d'amplitude plus sensible au bruit. Il est en fait possible d'estimer $\gamma$ en utilisant n'importe quelle fraction de la hauteur du pic. Cependant plus on se rapproche du maximum, plus l'argument du cosinus va être petit (tend vers zéro à la limite), rendant l'estimation instable. Master préconise l'emploi de la mi-hauteur comme compromis.

La Figure 4.6 montre l'erreur de ces deux estimateurs en fonction de la variation de fréquence, sans bruit additif. Ici la taille de fenêtre utilisée est de 32 ms et on constate bien que les estimateurs n'ont une précision satisfaisante qu'à partir de 10000Hz/s (erreur inférieure à 100 Hz/s). Ainsi que l'avait constaté Master [Master, 2002], l'estimateur d'amplitude est plus précis que l'estimateur basé sur la phase. Etant donné que pour l'étude des signaux de musique et de parole, on va s'intéresser plus particulièrement à des variations de fréquence inférieures à 10000 Hz/s, ces estimateurs ne seront pas retenus lors des comparaisons.

### Approximation de Taylor

Ici au lieu de chercher un développement asymptotique de la transformée de Fourier, on va effectuer un développement limité d'ordre 1 de $\exp(j\gamma\tau/2)$ au voisinage

de 0 :

$$e^{j\frac{\gamma}{2}\tau} \approx 1 + j\frac{\gamma}{2}\tau$$

Pour que cette approximation soit valable, on doit vérifier la condition suivante :

$$\gamma\tau_N^2 \ll 4\pi$$

Si on reprend le même exemple que précédemment, pour $\tau_N = 32$ms, on a alors $\gamma \ll 2000$ Hz/s. La méthode ne sera valable que pour des valeurs très petites de $\gamma$, environ de zéro à 200Hz/s.

On trouve l'approximation suivante de la transformée de Fourier :

$$X(t,\omega;h) \approx A\,e^{j\alpha}(H(\omega-\beta) + j\frac{\gamma}{2}HT(\omega-\beta)) \tag{4.2.15}$$

où $H$ est la TFD de $h$ et $HT$ celle de $h\tau^2$. On utilisera également par la suite $HTT$ la TFD de $h\tau^4$. Cette approximation est valable pour n'importe quel type de fenêtre $h$.

Master propose de dériver deux fois l'argument de la formule précédente au point $\omega = \beta$, pour trouver un estimateur. Comme la fenêtre $h$ est symétrique et zéro-phase, la TF de $\tau^i h$ sera purement imaginaire si $i$ est impaire, et sera purement réelle si $i$ est paire. De plus pour la fréquence $\omega = \beta$, la TF de $\tau^i h$ sera nulle si $i$ est impaire. On en déduit la formule suivante :

$$\left.\frac{d^2 \arg(X(t,\omega;h))}{d\omega^2}\right|_{\omega=\beta} \approx \frac{\gamma}{2}\frac{HT(0)^2 - H(0)HTT(0)}{H(0)^2 + \frac{\gamma^2}{4}HT(0)^2}$$

On fait alors deux approximations supplémentaires. Tout d'abord, on considère que la taille de la TF est suffisamment grande pour que le bin maximum $\omega_0$ soit presque égal à $\beta$ (*i.e.* $\omega_0 - \beta \approx 0$). Ensuite on remplace la dérivée continue par un filtre dérivateur d'ordre 2. Contrairement au cas de l'estimateur précédent (4.2.13), ici l'emploi de la dérivation est bien une approximation supplémentaire, car il n'y a pas identité entre les dérivations discrètes et continues.

Ensuite Master propose d'estimer $\gamma$ en résolvant le polynôme d'ordre 2 en $\gamma$. Cependant, comme l'estimateur ne fonctionne que pour de faibles valeurs de $\gamma$, le terme d'ordre 2 va être négligeable pour les fenêtres usuelles. On retrouve donc comme Masri[12] une dépendance linéaire entre $\gamma$ et le filtre dérivateur de phase d'ordre 2, mais cette fois avec une formule explicite :

$$\hat{\gamma} = 2\frac{H(0)^2}{HT(0)^2 - H(0)HTT(0)}\left(\frac{\arg(X^2(\omega_0)\bar{X}(\omega_0+\delta_\omega)\bar{X}(\omega_0-\delta_\omega))}{\delta_\omega^2}\right) \tag{4.2.16}$$

où $\delta_\omega$ est l'écart fréquentiel entre les bins utilisés. En général on va choisir le bin maximum et ses deux bins adjacent, c'est à dire $\delta_\omega = 2\pi F/P$. L'erreur de cet estimateur en fonction de la variation de fréquence a été tracé sur la Figure 4.7 avec et sans le terme d'ordre 2. La taille de fenêtre utilisée est de 32 ms, et on constate bien

[12] L'estimateur de Masri est décrit à la section 4.2.2.4.

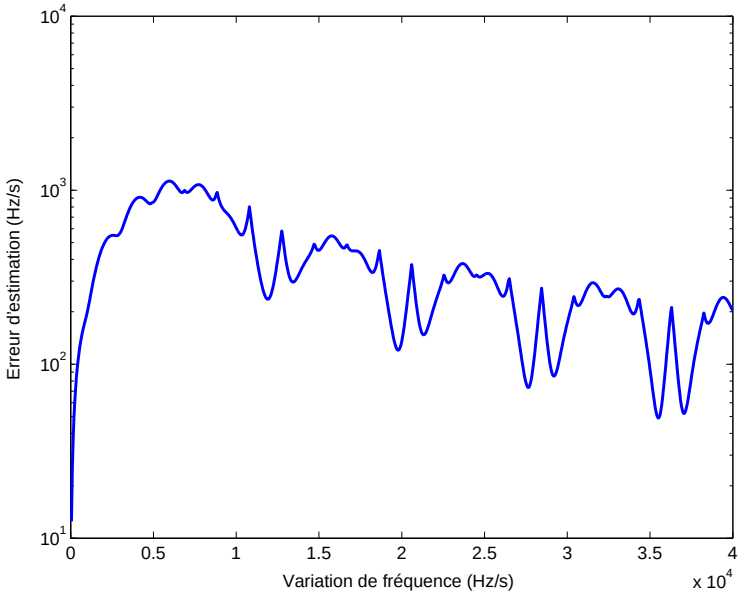



FIG. 4.8: Erreur de l'estimateur de Liu en fonction de la variation de fréquence

que dans ce cas, l'erreur est faible en dessous de 200 Hz/s. Lorsque l'on néglige le terme d'ordre 2, l'estimation est meilleure jusqu'à 5000 Hz/s environ.

On peut remarquer que beaucoup d'autres estimateurs auraient pu être formés à partir de l'équation (4.2.15), basés sur des rapports de transformées de Fourier. Mais toutes ces méthodes vont présenter les même limites que l'estimateur de Master, elles ne seront valides que pour des valeurs faibles de $\gamma$.

#### 4.2.2.3 Estimateur de Liu

La méthode de Liu [Master and Liu, 2003b] repose sur une propriété particulière de la transformée de Fourier continue d'un chirp avec une fenêtre rectangulaire. En utilisant la propriété de dérivation fréquentielle de la TF continue[13] et en faisant une intégration par partie, la dérivée fréquentielle de cette fonction prend cette forme :

$$\frac{dX_c(t,\omega;h_{\text{rec}})}{d\omega} = \frac{-jA\,\mathrm{e}^{j\alpha}}{\gamma}\left(-2\,\mathrm{e}^{j\frac{\gamma}{2}\tau_M^2}\sin((\omega-\beta)\tau_M)+(\omega-\beta)X_c(t,\omega;h_{\text{rec}})\right) \quad (4.2.17)$$

On rappelle que $\tau_M = \frac{N-1}{2F}$ représente la moitié de la durée de la fenêtre d'analyse.

Pour la fenêtre de Hann la dérivée fréquentielle a une expression encore plus simple. Sachant que la fenêtre de Hann peut se décomposer en une somme de trois fenêtres rectangulaires[14], on peut montrer que :

$$\frac{dX_c(t,\omega;h_{\text{han}})}{d\omega} = \frac{-j(\omega-\beta)}{\gamma}X_c(t,\omega;h_{\text{han}}) \quad (4.2.18)$$

[13] Voir l'annexe A.

[14] Pour plus détails, voir l'annexe B.

et en dérivant une nouvelle fois :

$$\frac{d^2 X_c(t,\omega;h_{\text{han}})}{d\omega^2} = \frac{-j}{\gamma} X_c(t,\omega;h_{\text{han}}) \tag{4.2.19}$$

Si $N$ est suffisamment grand pour que $X_c(t,\omega;h_{\text{han}}) \approx X(t,\omega;h_{\text{han}})$ et en approchant la dérivée continue de la TF par une dérivée discrète, on trouve l'estimateur suivant :

$$\hat{\gamma} = -jX(\omega)\left(\frac{X(\omega+\delta_\omega)+X(\omega-\delta_\omega)-2X(\omega)}{\delta_\omega^2}\right)^{-1} \tag{4.2.20}$$

où $X(\omega) = X(t,\omega;h_{han})$. On rappelle que cette formule n'est valable que pour la fenêtre de Hann. En général on va choisir d'utiliser le bin maximum et les deux bins adjacents. Dans ce cas $\omega$ est la fréquence du bin maximum et $\delta_\omega$ est la précision de Fourier : $\delta_\omega = 2\pi F/P$.

La Figure 4.8 représente l'erreur de cet estimateur en fonction de la variation de fréquence. Contrairement aux estimateurs de Master il n'y a pas de limite sur l'intervalle de définition de $\gamma$, même si expérimentalement, on constate un pic d'erreur assez important entre 5000 et 10000 Hz/s, pour une fenêtre de 32 ms.

#### 4.2.2.4 Estimateur de Marques et Almeida, Estimateur de Masri

L'estimateur de Marques et Almeida [Marques and Almeida, 1986] utilise une propriété particulière de la fenêtre Gaussienne, à savoir que la TF d'une Gaussienne est également une Gaussienne en $\omega$ (cf. Annexe B). Par interpolation quadratique on peut donc déterminer les coefficients de la parabole et en déduire à la fois la fréquence et la variation de fréquence. C'est la méthode qui est à l'origine de QIFFT (quadratically interpolated FFT), décrite par Abe et Smith [Abe and Smith III, 2005], qui permet de trouver tous les paramètres du modèle **M12**. Cette méthode est décrite en détail dans le paragraphe 4.3.1. Comme c'est une généralisation de la méthode Marques et Almeida, celle-ci n'a pas été retenue dans la comparaison.

De son côté, l'estimateur de Masri [Masri, 1996] provient de la constatation expérimentale que le filtre dérivateur d'ordre 2 est proportionnel à $\gamma$, pour de faible valeurs de $\gamma$. C'est exactement le résultat que trouve Master avec son approximation de Taylor (cf. section 4.2.2.2), mais en proposant une formule analytique, contrairement à Masri. C'est pourquoi on ne détaille pas cette référence ici.

## 4.3 Estimation pour le modèle M12

Dans cette section, on décrit la seule méthode qui traite du modèle **M12** avec modulation d'amplitude et de fréquence, la méthode de la QIFFT (Quadratically Interpolated FFT). Le modèle **M12** est plus réaliste que le précédent car dans de nombreux signaux réels les variations de fréquence sont souvent combinées avec des variations d'amplitude. Ce modèle peut s'écrire ainsi :

$$x(\tau) \triangleq \mathrm{e}^{\lambda_M+\mu\tau}\,\mathrm{e}^{j(\alpha_M+\beta_M\tau+\gamma\tau^2/2)} \tag{4.3.1}$$

où $\gamma$ est le taux de variation de fréquence, $\lambda_M$ est la log-amplitude instantanée et $\mu$ est le taux de variation de log-amplitude. Comme pour les autres modèles locaux, ces paramètres correspondent au temps $t_M$, et l'indice $M$ a été supprimé pour $\gamma$ et $\mu$ pour souligner les variations d'amplitude et de fréquence sont supposées constantes sur l'intervalle d'analyse.

### 4.3.1 Méthode d'interpolation quadratique de la TF (QIFFT)

La méthode QIFFT est basée sur une propriété des fenêtres Gaussiennes, à savoir que la transformée de Fourier d'une fonction Gaussienne est une autre fonction Gaussienne. Marques et Almeida [Marques and Almeida, 1986] avaient déjà proposé une méthode analytique utilisant cette propriété pour estimer la fréquence de modèles à phase quadratique (modèle **M02**). Masri [Masri, 1996] a étudié les distorsions causées par des modulations d'amplitude et de fréquence sur le spectre de la fenêtre Gaussienne. A partir des mesures de déformation, il a déduit une série d'abaques permettant d'estimer les taux de variation de fréquence et d'amplitude. Peeters [Peeters and Rodet, 1999] a rassemblé ces deux travaux pour donner une méthode complète d'estimation de tous les paramètres du modèle **M12**. Enfin Abe et Smith [Abe and Smith III, 2005] ont repris ces travaux pour leur donner un cadre théorique cohérent, et ont également proposé un algorithme rapide.

Ils ont notamment montré [Abe and Smith III, 2005] que la log-amplitude et la phase de la transformée de Fourier d'un signal de type **M12**, avec une fenêtre Gaussienne, sont toutes deux des fonctions quadratiques du paramètre de fréquence $\omega$ de la transformée de Fourier. Une interpolation quadratique de ces deux fonctions permet de calculer tous les paramètres du modèle.

La démonstration se fait dans le cas d'une fenêtre continue et infinie, mais dans la pratique, le signal est discrétisé et la fenêtre Gaussienne est tronquée. On a pu constater [Betser et al., 2008] que l'utilisation d'une fenêtre Gaussienne tronquée peut avoir de graves conséquences sur les performances des estimations, lorsqu'on est confronté à des taux de variation élevés, car la réponse fréquentielle n'est alors plus parabolique (voir Figure 4.9). Augmenter la taille de la fenêtre va atténuer ce problème. L'autre défaut de cette méthode est la nécessité d'utiliser de forts coefficients de zéro-padding [15] afin de réaliser une interpolation parabolique convenable, ce qui réduit d'autant plus la rapidité de l'algorithme.

En reprenant les notations de [Abe and Smith III, 2005], on note $p$ l'inverse de la variance de la fenêtre Gaussienne : $p = 1/(2\sigma^2)$. La fenêtre Gaussienne est définie par :

$$w(t) = \sqrt{\frac{p}{\pi}}\, \mathrm{e}^{-pt^2}$$

La log-amplitude et la phase sont des fonctions quadratiques de la fréquence $\omega$, c'est à dire qu'elles sont respectivement décrites par les fonctions $a_0\omega^2 + a_1\omega + a_2$ et $b_0\omega^2 + b_1\omega + b_2$. Les coefficients $a_i$ et $b_i$ sont calculés grâce à une interpolation parabolique sur les trois bins les plus proches du maximum du spectre d'amplitude (cf.

[15] Voir l'article [Betser et al., 2007], reproduit dans l'annexe G.

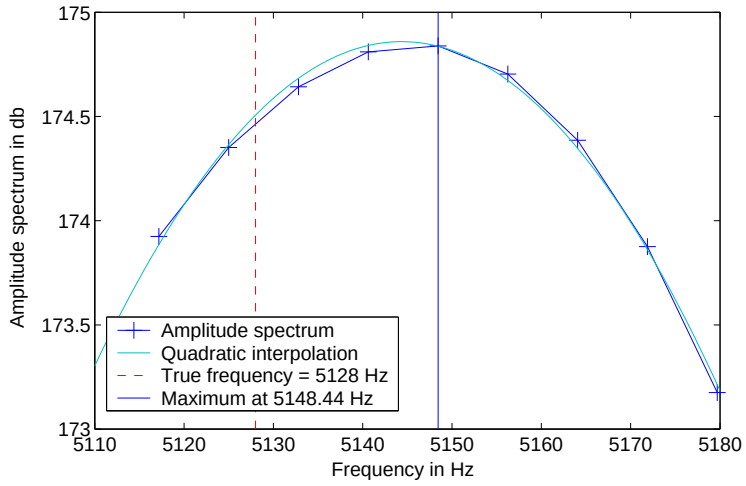



FIG. 4.9: La réponse de la fenêtre Gaussienne n'est plus parabolique pour de fortes modulations ($\mu = 50$, $\gamma = 8000$)

[Abe and Smith III, 2005] pour plus de détails). On montre alors que les paramètres du modèle **M12** peuvent s'estimer ainsi :

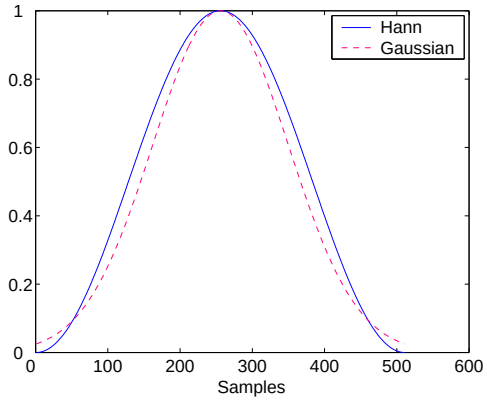
$$\hat{\omega}_0 = -\frac{\hat{a}_1}{2\hat{a}_0}, \qquad \hat{\mu} = -2p(2\hat{b}_0\hat{\omega}_0 + \hat{b}_1),$$

$$\hat{\gamma} = p\hat{b}_0/\hat{a}_0, \qquad \hat{\beta}_M = \hat{\omega}_0 + \hat{\mu}\hat{\gamma}/p,$$

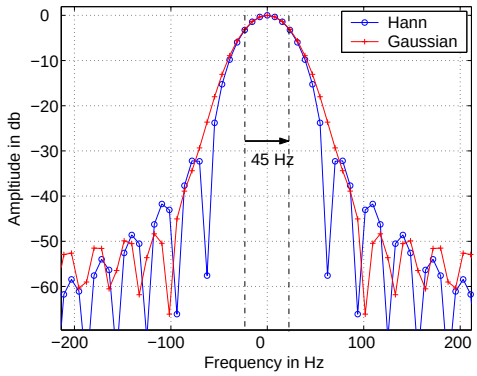
$$\hat{\alpha}_M = \hat{b}_0\hat{\omega}_0^2 + \hat{b}_1\hat{\omega}_0 + \hat{b}_2 + \frac{\hat{\mu}^2\hat{\gamma}}{4p^2} - \frac{1}{2}\arctan\left(\frac{\hat{\gamma}}{p}\right),$$

$$\hat{\lambda}_M = \hat{a}_0\hat{\omega}_0^2 + \hat{a}_1\hat{\omega}_0 + \hat{a}_2 - \frac{\hat{\mu}^2}{4p} + \frac{1}{4}\log\left(1 + \left(\frac{\hat{\gamma}}{p}\right)^2\right)$$

Ce cadre théorique ne s'applique qu'à la fenêtre Gaussienne, mais dans le cas des autres fenêtres, une forme empirique peut être dérivée des formules Gaussiennes en utilisant des coefficients d'adaptation [Abe and Smith III, 2005]. Ces coefficients d'adaptation ne donnent qu'une approximation de la réponse fréquentielle de la fenêtre Gaussienne de même résolution et la méthode souffre toujours des problèmes décrits précédemment. C'est pourquoi dans ce chapitre et dans nos expériences par la suite nous ne discuterons que du cas Gaussien pour cette méthode. Le réglage de la résolution de la fenêtre Gaussienne est un choix important pour faire une comparaison équitable entre les estimateurs. On va imposer la même résolution que la fenêtre de Hann de longueur 32 ms, c'est à dire 45 Hz. La fenêtre résultante est comparée à la fenêtre de Hann sur la Figure 4.10. Le choix de la résolution de 45 Hz est discuté dans la section 6.2.7.




(a) Fenêtre de Hann et fenêtre Gaussienne normalisée




(b) Réponses fréquentielles

FIG. 4.10: Comparaison de la fenêtre de Hann et de la fenêtre Gaussienne pour une résolution de 45 Hz, et une longueur de 32 ms.

## 4.4 Comparaison expérimentale des estimateurs de l'état-de-l'art

Un certain nombre de travaux se sont attachés à comparer les estimateurs existants basés sur Fourier, dans le cas du modèle **M01** [Hofbauer, 2004], [Hainsworth and Macleod, 2003a]. Cependant peu d'articles ont étudié les performances de ces estimateurs face aux modulations d'amplitude et de fréquence.

Dans cette section nous allons donc comparer les estimateurs classiques pour tous les paramètres du modèle **M12**. Pour chaque méthode les quatre modèles étudiés, **M01**, **M02**, **M11** et **M12**, seront testés. Les méthodes développées pour des ordres peu élevés, comme par exemple toutes les techniques de type vocodeur de phase, vont bien sûr présenter des biais importants pour les modèles d'ordre supérieur. Un des buts de cette section est de mettre en évidence les limitations des différentes méthodes existantes en fonction du type de modulation utilisé.

La méthode QIFFT, plus récente que les autres, n'a pas été comparée de façon systématique aux autres estimateurs dans la littérature. Cette section a également pour but de remédier à ce manque. Par ailleurs la QIFFT étant une méthode conçue pour le modèle **M12**, elle va nous donner une idée du gain de performance que l'on peut atteindre en améliorant les méthodes existantes basées sur un modèle d'ordre faible.

Le protocole d'évaluation a été déjà complètement décrit dans la section 3.4. Dans les expériences, les paramètres des sinusoïdes sont tirés au hasard, avec une distribution uniforme. L'erreur déterministe présentée sur les courbes, ainsi que l'erreur stochastique, sont en fait des moyennes par rapport à la distribution des parmètres. Par la suite, on parlera simplement d'erreur déterministe et d'erreur stochastique.

### 4.4.1 Estimateurs de fréquence

Nous étudions tout d'abord l'estimation de fréquence, pour les estimateurs suivants : vocodeur à long terme (section 4.1.2.1), QIFFT (section 4.3.1), réassignement (section 4.2.1.1), interpolation triangulaire et parabolique (section 4.1.2.3), estimation par la méthode des sinus, des cosinus et des tangentes (section 4.1.2.2), estimateur de Macleod à 3 bins (section 4.1.2.4). Pour les méthodes des sinus, cosinus et tangentes, le pas d'avancement utilisé est de 1 échantillon[16]. Pour le vocodeur à long terme, le pas d'avancement est de 8 ms[17]. La méthode de la QIFFT est déclinée en deux versions, avec une fenêtre courte, 32 ms, et avec une fenêtre longue, 48 ms[18]. Dans les deux cas, la résolution de la fenêtre Gaussienne est la même. La méthode de Macleod utilise une fenêtre rectangulaire de 32 ms[19] et toutes les autres méthodes des fenêtres de Hann de 32 ms[20]. Toutes les méthodes utilisent un facteur de zéro-padding de 4, excepté l'interpolation triangulaire. La CRB représentée est la borne pour $W = 732$ échantillons (48 ms pour $F = 16000$).

Les courbes vont toutes avoir la même allure typique. Pour les hauts SNRs, elles vont présenter une partie horizontale qui correspond à l'erreur déterministe de l'estimateur. Pour des SNRs faibles, l'erreur déterministe, causée par les approximations des différentes méthodes, devient plus faible que l'erreur stochastique, dûe au bruit additif. Dans cette zone, l'erreur est parallèle à la CRB. Pour des SNRs très faibles, certaines hypothèses sur lesquelles reposent les méthodes peuvent ne plus être valable, engendrant une très forte erreur[21].

Dans le cas du modèle classique **M01**, Figure 4.11(a), tous les estimateurs sont proches de la borne théorique, la CRB. Excepté les estimateurs sinus et cosinus, toutes les méthodes ont une erreur stochastique très proche. Seuls les erreurs déterministes présentent des différences vraiment significatives. Toutes les méthodes basées sur l'estimation de la variation de phase, *i.e.* les estimateurs cosinus, sinus, tangente et vocodeur à long terme, sont non biaisées, avec une erreur déterministe nulle. Le réassignement et la QIFFT longue présentent une erreur déterministe très faible. Enfin pour les autres méthodes cette erreur est un peu pluse élevé, mais reste assez faible. On voit que les méthodes tangente et vocodeur ont des résultats quasiment identiques, ce qui est cohérent avec le fait qu'elles estiment le même angle et toutes deux avec la fonction tangente (voir section 4.1.2.2). L'estimateur de Macleod a une erreur stochastique légèrement plus faible que les autres, à cause de la fenêtre rectangulaire qui a une résolution plus fine que les autres.

La Figure 4.11(b) montre le comportement des estimateurs face à une forte modulation d'amplitude ($\mu_m = 100$). L'erreur déterministe de toutes les méthodes aug-

[16] C'est la valeur utilisée dans la littérature.

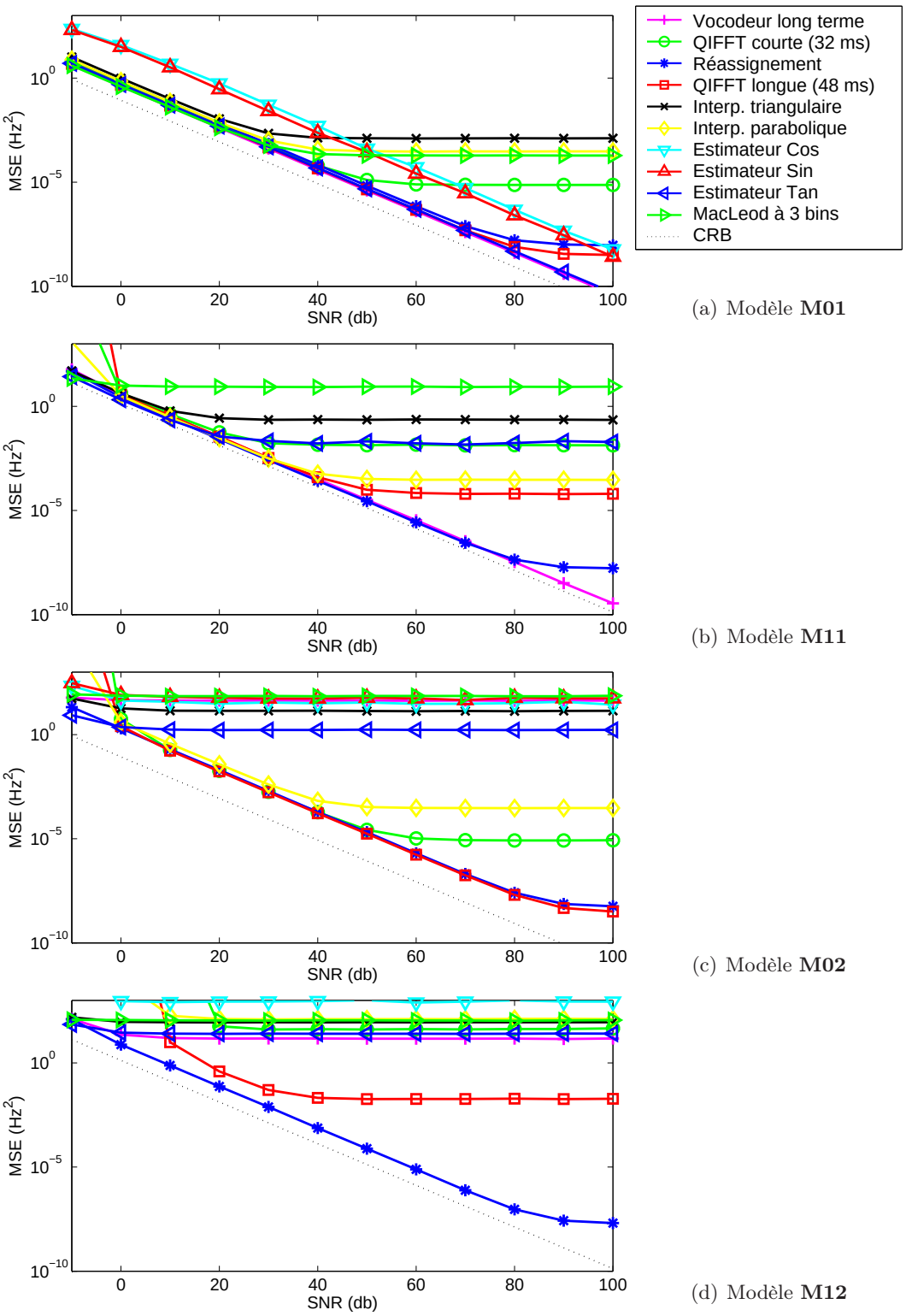
[17] Cela correspond au pas d'avancement entre deux trames pour un système classique d'analyse audio. Cette méthode va donc utiliser plus d'échantillons que les autres méthodes.

[18] Les performances de la QIFFT sont évaluées avec deux fenêtres différentes pour être cohérent avec les expériences du chapitre 6, même si la méthode de la QIFFT longue utilise en fait plus d'échantillons que les autres méthodes.

[19] La méthode de Macleod ne fonctionne que pour la fenêtre rectangulaire.

[20] Même l'interpolation triangulaire. Voir la description de la méthode à la section 4.1.2.3.

[21] Pour plus d'explications sur l'allure des courbes on se référera à la section 5.4.




(a) Modèle **M01**

(b) Modèle **M11**

(c) Modèle **M02**

(d) Modèle **M12**

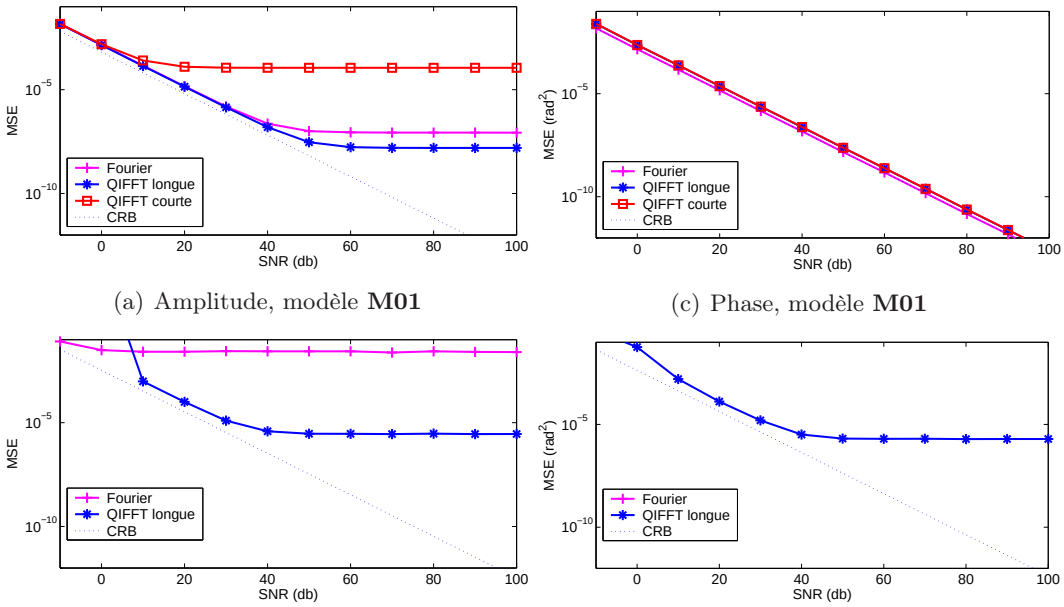
FIG. 4.11: Evaluation des estimateurs classiques de fréquence

mente excepté pour la méthode du vocodeur de phase, qui est toujours non biaisée[22], et pour la méthode du réassignement. Pour ces deux méthodes, les performances sont maintenant très proches de la borne théorique. Les erreurs déterministes des méthodes sinus et cosinus sont très importantes, supérieurs à $10^2$ et les performances de ces méthodes tombent en dehors de l'intervalle d'erreur de la figure. Pour la QIFFT, on voit l'importance de l'utilisation d'une fenêtre suffisamment longue pour bien prendre en compte les déformations causées par la modulation d'amplitude. Ces déformations ont déjà été mentionnées dans la section 4.3.1, et causent une erreur déterministe importante sur les estimations. Pour des SNRs bas, l'algorithme QIFFT devient instable. Cette erreur est due à la propagation des erreurs à partir des paramètres d'ordre supérieur (ici la variation d'amplitude) vers les paramètres d'ordre inférieur.

La Figure 4.11(c) présente les résultats pour une modulation de fréquence forte ($\gamma_m = 8000$). Pour toutes les méthodes d'estimation basées sur le modèle **M01**, on a assigné comme temps d'estimation par défaut le temps du milieu de la fenêtre. Toutes les méthodes basées sur le modèle **M01** présentent des erreurs déterministes très importantes. Seules les méthodes du réassignement et de la QIFFT, continuent d'exhiber de bons résultats, même si leur erreur stochastique augmente sensiblement. On peut noter également la bonne performance de l'interpolation parabolique. Lors d'une modulation de fréquence seule, le maximum du spectre de Fourier correspond exactement à la fréquence $\beta_M$ que l'on cherche à estimer, et d'autre part, le spectre d'une fenêtre de Hann reste quasiment parabolique au voisinage du maximum, ce qui explique ce résultat.

Enfin la dernière Figure 4.11(d), nous montre les performances des estimateurs pour de fortes modulations à la fois d'amplitude et de fréquence. La méthode du réassignement donne les même résultats que dans le cas précédent. La QIFFT avec fenêtre longue présente toujours une estimation correcte de la fréquence, mais avec des erreurs déterministes et stochastiques plus importantse. Lors d'une modulation simultanée de l'amplitude et de la fréquence, le maximum de Fourier ne correspond plus à la fréquence $\beta_M$ que l'on cherche à estimer[23], le principe de l'interpolation parabolique ne peut plus s'appliquer tel quel et les performances de la méthode chuttent (la courbe de performance est sous la courbe correspondant à l'estimateur de Macleod).

Pour conclure, on peut dire que la méthode de la littérature la plus robuste pour estimer la fréquence est le réassignment. Elle continue à donner des performances très proches du CRB, même pour de fortes modulations.




(a) Amplitude, modèle **M01**

(c) Phase, modèle **M01**

(b) Amplitude, modèle **M12**

(d) Phase, modèle **M12**

FIG. 4.12: Evaluation des estimateurs classiques d'amplitude et de phase

### 4.4.2 Estimateurs d'amplitude et de phase

Nous comparons maintenant les deux méthodes d'estimation des paramètres d'ordre 0, celle basée sur Fourier et dédiée au modèle **M01** et celle basée sur la QIFFT, qui elle est dédiée au modèle **M12**. La QIFFT est toujours déclinée en deux versions, une avec fenêtre longue et une avec fenêtre courte. Pour la méthode basée sur Fourier, le choix s'est porté sur le vocodeur de phase à long terme pour l'estimation de fréquence[24]. Pour l'estimation d'amplitude, nous avons échantillonné 20 valeurs de la fonction $H(\omega_k - \beta)$ sur l'intervalle $[-R, R]$ (voir équation (4.1.5)). Les valeurs intermédiaires sont calculées par interpolation linéaire. L'estimateur de phase ne nécessite pas de fonction particulière, car la transformée de Fourier est zéro-phase[25].

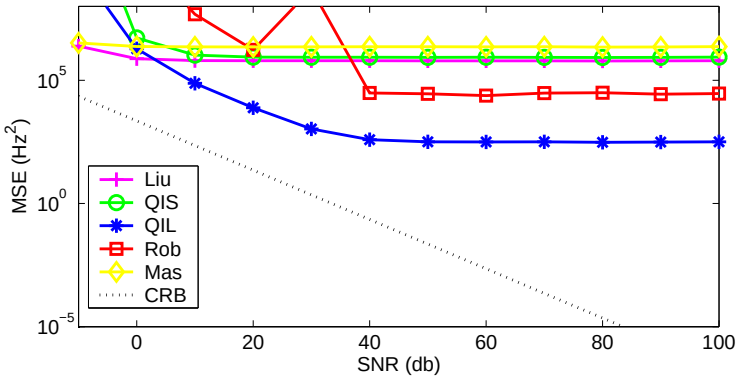
Les courbes des Figures 4.12(a) et 4.12(c) sont une comparaison des deux méthodes Fourier et QIFFT pour des sinusoïdes non modulées. On voit que dans ce cas, les deux méthodes sont très proches du CRB et présentent des erreurs déterministes faibles. Sur la Figure 4.12(a), l'erreur déterministe de la méthode basée sur Fourier est dûe à l'échantillonnage de la fonction $H$ et peut être réduit en augmentant le

[22]En effet, pour une modulation d'amplitude seule, il y a toujours identité entre la phase de Fourier sur le lobe principal et la phase initiale de la sinusoïde. La formule d'estimation utilisée dans le vocodeur de phase est donc toujours une formule exacte, sans approximations. On rappelle que pour les modèles **M01** et **M11** la phase de Fourier est constante sur tout le lobe principal car la transformée de Fourier est centrée.

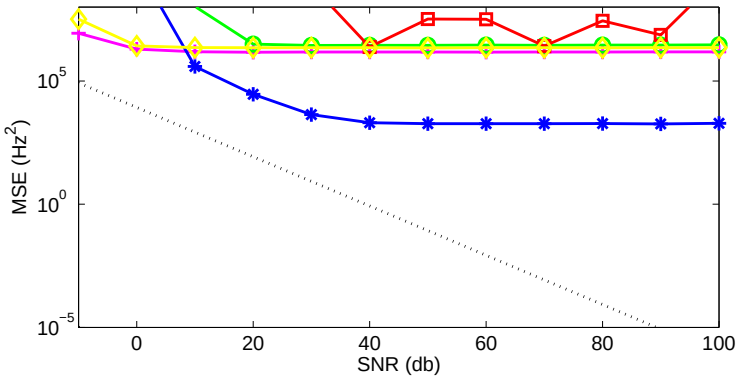
[23]Voir la section 4.3.1 pour plus de détails.

[24]La méthode demande en effet une estimation au préalable de la fréquence. Voir la section 4.1.1.

[25]Voir la section 3.3.6.2.




(a) Modèle **M02**



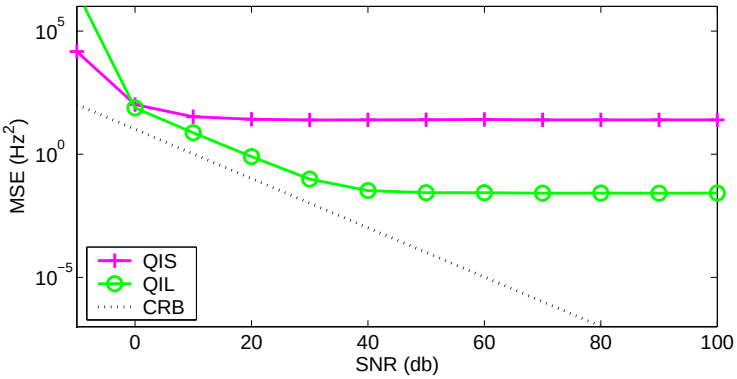

(b) Modèle **M12**

FIG. 4.13: Evaluation des estimateurs classiques de variation de fréquence. 'QIS' correspond à la QIFFT courte et 'QIL' à la QIFFT longue.

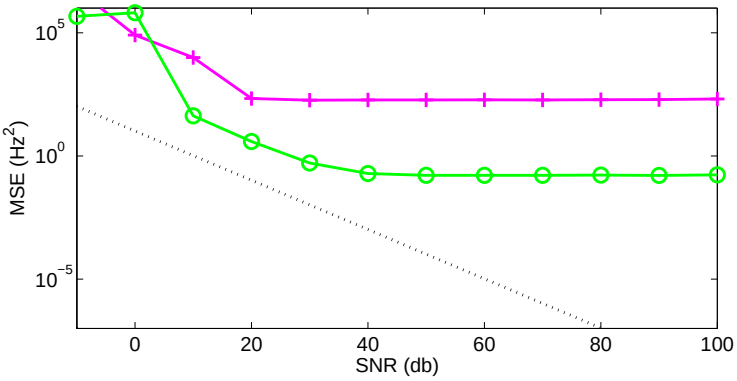
nombre d'échantillons. L'erreur déterministe de la méthode de la QIFFT est toujours dûe à l'utilisation d'une fenêtre Gaussienne tronquée. Lorsqu'on ajoute des modulations fortes, comme pour les courbes des Figures 4.12(b) et 4.12(d), l'estimation directement basée sur la transformée de Fourier ne fonctionne plus. Seule la QIFFT permet d'estimer correctement les amplitudes et les phases dans ce cas.

### 4.4.3 Estimateurs de la variation de fréquence

Comme nous l'avions constaté dans la section traitant des méthodes d'estimation de fréquence pour le modèle **M02**, toutes les méthodes présentent des erreurs déterministes trop importantes pour être vraiment applicables sur un large intervalle de variation de fréquence, $\gamma \in [0, 8000]$. La méthode de Röbel est légèrement plus performante que les autres. On remarque que la méthode QIFFT donne aussi une très mauvaise estimation lorsque la fenêtre est courte. Seule la QIFFT avec une fenêtre longue permet d'estimer la variation de fréquence avec une précision raisonnable, avec une erreur déterministe d'environ 20 Hz.s$^{-1}$.




(a) Modèle **M11**




(b) Modèle **M12**

FIG. 4.14: Evaluation des estimateurs classiques de variation d'amplitude

Dans le cas du modèle **M12**, représenté sur la Figure 4.13(b), la méthode de Röbel devient trop biaisée pour donner une estimation correcte. On remarque aussi que l'erreur déterministe de la méthode QIFFT devient plus forte, et que la méthode décroche pour des SNRs inférieurs à 10 dB, contre 0 dB dans le cas précédent.

### 4.4.4 Estimateurs de la variation d'amplitude

Pour l'estimation de la variation d'amplitude, on ne dispose que de la méthode de la QIFFT, toujours déclinée en deux versions, avec une fenêtre longue et une fenêtre courte. On constate toujours le même problème avec la fenêtre courte, la réponse fréquentielle n'étant qu'approximativement parabolique, l'erreur déterministe est trop forte pour donner une estimation correcte. Dans le cas du modèle **M11** (cf. Figure 4.14(a)), l'estimation avec la fenêtre longue présente une erreur déterministe assez faible, de l'ordre de $10^{-1}$, et l'estimation est assez proche de la borne théorique. Dans le cas du modèle **M12**, l'erreur déterministe est quasiment la même, mais l'erreur stochastique augmente légèrement. Comme pour l'estimation de variation de fréquence, la méthode décroche aux alentours de 10 dB, contre 0 dB dans le

cas du modèle **M11**. Sachant que l'estimation de la fréquence de la QIFFT utilise les valeurs estimées des variations d'amplitude et de fréquence, on comprend pourquoi la méthode décroche également vers 10 dB pour l'estimation de fréquence (voir Figure 4.11(d)), contrairement au réassignement.

## 4.5 Conclusion

Dans cette section nous avons présenté sous un même formalisme un grand nombre de méthodes d'estimation basées sur la transformée de Fourier. Nous avons ensuite effectué une comparaison rigoureuse de ces méthodes. Un certain nombre d'entre elles, comme la QIFFT et les méthodes d'estimation de variation de fréquence de la section 4.2.2, n'avait à notre connaissance jamais été évalué de façon poussée. Nous avons également étudié en détail les performances des estimateurs classiques pour des modulations d'amplitude et de fréquence, et nous avons mis en évidence la faible robustesse aux modulations d'amplitude et de fréquence de la plupart des méthodes.

CHAPITRE 5

# Etude critique des méthodes d'estimation reposant sur la TFCT

Dans ce chapitre nous allons décortiquer les méthodes d'estimation basées sur la TFCT. Nous verrons dans la section 5.1, que tous ces estimateurs sont basés sur les mêmes principes, puis dans la section 5.2, nous esquisserons deux méthodes pour dériver un estimateur à partir d'une combinaison particulière de bins de la TFCT. Nous nous intéresserons ensuite au temps d'estimation de ces méthodes dans la section 5.3. Pour des paramètres qui varient au cours du temps, les performances attendues pour un estimateur vont varier en fonction du temps d'estimation choisi. La question de savoir à quel instant on aura la meilleure estimation est donc cruciale. Nous tenterons d'apporter des éléments de réponse en étudiant les performances de la borne théorique (CRB) en fonction du temps d'estimation. Enfin nous verrons comment dériver les propriétés statistiques des estimateurs, biais et variance dans la section 5.4. Dans la littérature, la variance est généralement dérivée en étudiant le comportement asymptotique de l'estimateur, lorsque le nombre d'échantillons $N$ tend vers l'infini. Pour certains des modèles étudiés ici, *i.e.* les modèles avec variation d'amplitude, ce type d'étude ne sera pas possible, l'énergie de ces modèles n'étant pas bornée lorsque $N$ tend vers l'infini. Nous proposons donc une méthode alternative, basée sur l'hypothèse de transformation de Fourier bien résolue évoquée dans la section 3.3.2 et qui sera précisée dans la section 5.1.1.

## 5.1 Analyse des estimateurs de fréquence basés sur la TFCT

Le problème de l'estimation de paramètres basée sur la transformée de Fourier, peut être résumé de la façon suivante. En entrée de notre estimateur nous avons un ensemble de points temps-fréquences complexes, calculés à l'aide d'une ou plusieurs TFCT. L'estimateur devra combiner ces points pour calculer le paramètre recherché.

Toutes les méthodes décrites dans la section bibliographique rentre dans ce cas de figure.

### 5.1.1 Estimation bien résolue par la transformée de Fourier

On considère un signal $s = x + n$, où $n$ est le bruit et $s$ le signal perturbé. On note $X_i = X(t_i, \omega_i; h_i)$ et $N_i = N(t_i, \omega_i; h_i)$ les transformées de Fourier de $x$ et $n$ pour le point temps-fréquence $(t_i, \omega_i)$ et la fenêtre $h_i$.

La première étape pour dériver un estimateur consiste à sélectionner les points temps-fréquences (ou bins) que l'on va utiliser, et de choisir un point de référence. Pour que l'estimation soit valable, il faut que la sinusoïde soit "bien résolue" par la transformée de Fourier pour les bins en question. Pour cela deux conditions doivent être vérifiées.

Il faut tout d'abord que le bruit soit négligeable pour les bins choisis, c'est à dire que $N_i \ll X_i$[1]. Il faudra également s'assurer que le modèle utilisé est résolvable par la TF dans le voisinage considéré. Dans le cas d'une modulation de fréquence par exemple, la transformée de Fourier ne résoudra plus la sinusoïde lorsque le nombre d'échantillon $N \to \infty$[2]. Une condition pour s'assurer que le modèle est bien résolu par la TF consiste à supposer que la largeur du lobe principal de spectre du signal doit être supérieur à la variation de fréquence maximale à l'intérieur de la fenêtre d'analyse. Dans le cas contraire, des pics multiples vont apparaître sur le lobe principal du spectre. Par exemple, si $\gamma_m$ est la plus grande variation de fréquence tolérée pour le signal considéré, et si le lobe principal de la sinusoïde possède une largeur de $K$ bins (indépandemment de la taille de la fenêtre $N$), cette condition peut s'exprimer ainsi : $\gamma_m \frac{N}{F} < 2\pi \frac{KF}{N} \Leftrightarrow \frac{N}{F} < \sqrt{\frac{2\pi K}{\gamma_m}}$. Pour des chirps avec de très fortes variations ($\gamma_m \gg 1\pi K$), cette condition devient $\tau_N = N/F \ll 1s$.

On dira donc que l'estimation sera bien résolue par la transformée de Fourier, si les deux conditions précédentes sont vérifiées pour tous les bins utilisés par l'estimateur. Nous verrons par la suite comment utiliser ces conditions pour dériver les propriétés statistiques des estimateurs.

### 5.1.2 Principes généraux

Pour pouvoir estimer un paramètre, la combinaison des points temps-fréquences choisis doit permettre d'éliminer les autres paramètres. Pour trouver une bonne combinaison, on doit d'abord exprimer les points temps-fréquences en fonction du point temps-fréquence de référence $(t_0, \omega_0)$. On va se restreindre au modèle **M12**, mais on précisera lorsque certains résultats sont valables pour des modèles plus généraux. Pour un point temps-fréquence quelconque $(t, \omega)$, on définit $\delta_t = t - t_0$ et $\delta_\omega = \omega - \omega_0$.

[1] Soit $a$ et $b \in C$. Par $a \ll b$, on entend que $a+b$ est dans un voisinage très proche de $b$, c'est à dire que $a + b$ est inclus dans un disque autour de $b$ de rayon $\epsilon$ avec $\epsilon \ll 1$. Voir par exemple [Dieudonné, 1980] pour un rappel sur l'étude des fonctions complexes.

[2] En effet pour un signal modulé en fréquence, quand $N$ tend vers l'infini, le chirp couvre toute la plage de fréquence avec la même énergie. La sinusoïde n'est donc plus résolue fréquentiellement.

Pour le modèle **M12**, la transformée de Fourier peut s'exprimer ainsi :

$$X(t,\omega;h) = \mathrm{e}^{\mu\delta_t + j(\beta\delta_t + \gamma\frac{\delta_t^2}{2})}\, X(t_0, \omega_0 + \delta_\omega - \delta_t\gamma; h) \tag{5.1.1}$$

L'étape suivante est l'estimation des paramètres d'ordre supérieur[3]. Pour cela on doit d'abord éliminer l'amplitude complexe [4].

**Elimination de l'amplitude complexe**

Quel que soit le modèle utilisé, on sait que l'amplitude complexe $\tilde{A} = \exp(\lambda + j\alpha)$ va se mettre en facteur pour tous les points temps-fréquences calculés : $X(t,\omega;h) = \tilde{A}f(\mu,\beta,\gamma)$. La phase initiale peut être éliminée de deux façons différentes : par division des transformées en deux points temps-fréquences (5.1.2), ou par multiplication de la transformée en un point temps-fréquence par le conjugué de la transformée en un autre (5.1.3).

$$X_2/X_1 = f_2(\mu,\beta,\gamma)/f_1(\mu,\beta,\gamma) \tag{5.1.2}$$
$$X_2.\bar{X}_1 = \mathrm{e}^{2\lambda}.f_2(\mu,\beta,\gamma)\bar{f}_1(\mu,\beta,\gamma) \tag{5.1.3}$$

De façon similaire, on va pouvoir éliminer $\lambda$ de deux façons différentes, par division des transformées en deux points temps-fréquences (5.1.4) ou en prenant l'argument de la transformée en un point temps-fréquence (5.1.5) :

$$X_2/X_1 = f_2(\mu,\beta,\gamma)/f_1(\mu,\beta,\gamma) \tag{5.1.4}$$
$$\arg(X_1) = \alpha + \arg(f_1(\mu,\beta,\gamma)) \tag{5.1.5}$$

Toutes les méthodes d'estimation basées sur la TFCT vont combiner ces possibilités de façon à éliminer à la fois l'amplitude et la phase initiale. On remarque que l'emploi d'un rapport de transformées permet directement d'éliminer les deux paramètres. On remarque aussi que si l'on considère une combinaison linéaire de bins, alors l'élimination se fera exactement de la même façon.

Un grand nombre de méthodes, notamment certains interpolateurs de spectre, utilise des combinaisons linéaires de transformées en différents points temps-fréquences. Les combinaisons utilisées par ces méthodes peuvent toutes se mettre sous la forme :

$$\mathcal{H} = \frac{\sum_i \kappa_i X_i}{\sum_i \nu_i X_i} \tag{5.1.6}$$

où $\kappa_i$ et $\nu_i$ sont des facteurs complexes qui pondèrent les transformées. Les estimateurs de Macleod et les méthodes d'interpolation de spectre d'amplitude, comme l'interpolation parabolique, éliminent d'abord la phase, en utilisant un bin conjugué, avant d'éliminer l'amplitude. Les premiers utilisent la fonction $R_2 = \Re(X_1\bar{X}_2)$, et les autres le module, $|X| = \sqrt{X\bar{X}}$. On voit que ces fonctions préservent le facteur d'amplitude, qui sera par la suite éliminé grâce au rapport.

[3] C'est à dire d'ordre supérieur strict à 0. Dans le cas du modèle **M12**, il s'agit des paramètres $\mu$, $\beta$ et $\gamma$.

[4] L'amplitude complexe est incluse dans le terme $X(t_0, \omega_0 + \delta_\omega - \delta_t\gamma; h)$ dans l'équation 5.1.1.

#### 5.1.2.1 Estimation des paramètres d'ordre supérieur

Pour le modèle **M01**, l'amplitude et la phase étant éliminées, la combinaison $\mathcal{H}$ de bins choisie est une fonction $l$ qui ne dépend que de la fréquence $\beta$. Si cette fonction est inversible, une estimation de $\beta$ peut alors être réalisée grâce à $l^{-1} : \hat{\beta} = l^{-1}(\mathcal{H})$. Tous les estimateurs présentés dans la section 4, sont obtenus en utilisant des propriétés de la TF et des approximations, afin d'avoir une fonction inversible simple. Dans le cas général, les fonctions obtenues par combinaison de bins sont rarement inversibles sur tout l'intervalle de définition des fréquences. Nous verrons dans la section 5.2 des méthodes génériques pour développer un estimateur à partir d'une combinaison particulière de bins.

Pour les modèles d'ordre supérieur, le problème est plus délicat. Une fois l'amplitude et la phase initiale éliminées, la combinaison de bins est alors une fonction des paramètres d'ordre supérieur. On suppose que cette fonction est réelle et qu'il y a $U$ paramètres d'ordre supérieur. Pour estimer ces paramètres de façon non ambiguë, une fonction $l$ de $\mathbb{R}^U$ dans $R$ ne sera pas suffisante. On peut généraliser la remarque du cas **M01** en supposant que $l$ est une fonction de $\mathbb{R}^U$ dans $\mathbb{R}^V$ injective, où $V \geq U$. $l$ sera donc formée par $V$ fonctions indépendantes de $\mathbb{R}^U$ dans $\mathbb{R}$. En d'autres termes, $l$ est formé de $V$ combinaisons $\mathcal{H}$ indépendantes. La propriété de l'injectivité nous assure qu'à tout élément de l'ensemble de définition correspond un élément dans le nouvel espace, et que tout élément de l'espace d'arrivée ne possède au plus qu'un antécédent dans l'ensemble de définition des paramètres. Donc si on analyse un signal dont les paramètres $x$ sont dans l'ensemble de définition, alors $l(x)$ n'aura qu'un seul antécédent, qui sera $x$. Dans le cas où $U = V$, $l$ doit être bijective. Elle correspond alors à un changement de variable tel que, dans le nouvel espace, tous les paramètres soient directement calculables avec les TFCT. Comme dans le cas du modèle **M01**, il est cependant difficile de trouver une fonction injective sur tout l'intervalle de définition.

## 5.2 Déterminer un estimateur à partir d'une combinaison particulière de bins

Nous venons de voir qu'il est assez aisé de définir des combinaisons de bins qui ne dépendent que des paramètres d'ordre supérieur ou égal à 1. Il s'agit maintenant de pourvoir estimer ces paramètres grâce à ces combinaisons de bins. Nous allons exposer les principes des deux méthodes que nous utiliserons dans le chapitre suivant pour dériver des estimateurs. La première cherche à linéariser la fonction par rapport à ces paramètres via des approximations de Taylor. La deuxième ne cherche pas de formule explicite pour l'estimateur mais plutôt à modéliser la fonction inverse.

### 5.2.1 En utilisant des approximations de Taylor

L'approximation de Taylor a déjà été utilisée dans certaines méthodes comme les interpolateurs de spectre utilisant la phase décrit à la section 4.1.2.4, ou l'estimateur de variation de fréquence de Master décrit à la section 4.2.2.2. Nous allons tout

d'abord généraliser son emploi, en ne faisant que des hypothèses générales sur la fenêtre d'analyse utilisée, à la façon de Master. On rappelle que pour les méthodes de la section 4.1.2.4 par exemple, on utilise uniquement la fenêtre rectangulaire car cela permet d'avoir une formule explicite de la TF, facilitant les calculs. Nous utiliserons la formule suivante pour trouver nos développements de Taylor :

$$X(t,\omega;h) = \sum_{i=0}^{Q} \frac{(-j)^i(\omega-\omega_0)^i}{i!} X(t,\omega_0;\tau^i h) \tag{5.2.1}$$

Cette formule découle directement de la propriété de dérivation de la transformée de Fourier[5].

Il s'agit d'un développement de Taylor par rapport au paramètre de fréquence. En général, le maximum de la TFCT nous donne déjà une bonne approximation de la fréquence de la sinusoïde. Donc si le développement de Taylor est effectué aux alentours de cette fréquence, l'erreur sera d'autant plus petite et aura une borne indépendante de la valeur maximum du paramètre, comme nous le verrons dans le chapitre suivant. Ce n'est pas le cas par exemple pour un développement de Taylor par rapport au paramètre $\gamma$. Ce type de développement est moins judicieux, car sa zone de validité sera beaucoup plus restreinte, comme on a pu le voir pour l'estimateur de Master 4.2.2.2.

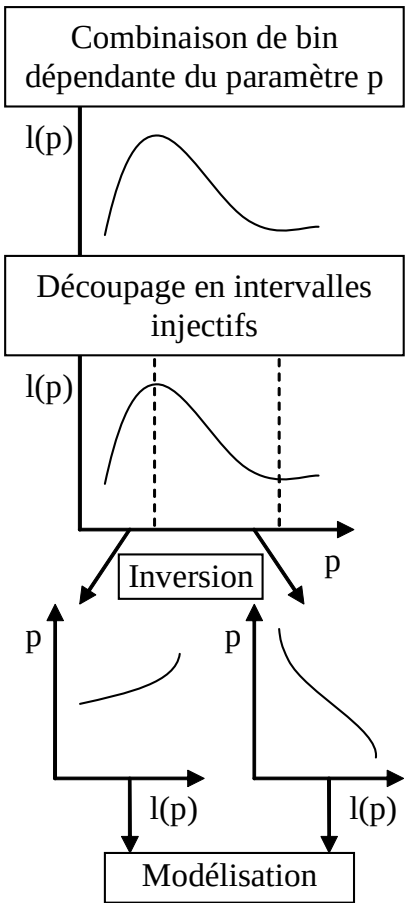
Le choix de la fréquence de référence va dépendre de la méthode utilisée. Il faut bien évidemment prendre celle qui minimise l'erreur déterministe de l'estimateur. Pour le modèle **M01**, si la combinaison de bin utilisée est symétrique, il faut utiliser comme référence le centre de symétrie des bins, car la symétrie permet alors d'éliminer des termes d'erreur (cf. section 6.1.1).

### 5.2.2 En utilisant des modélisations

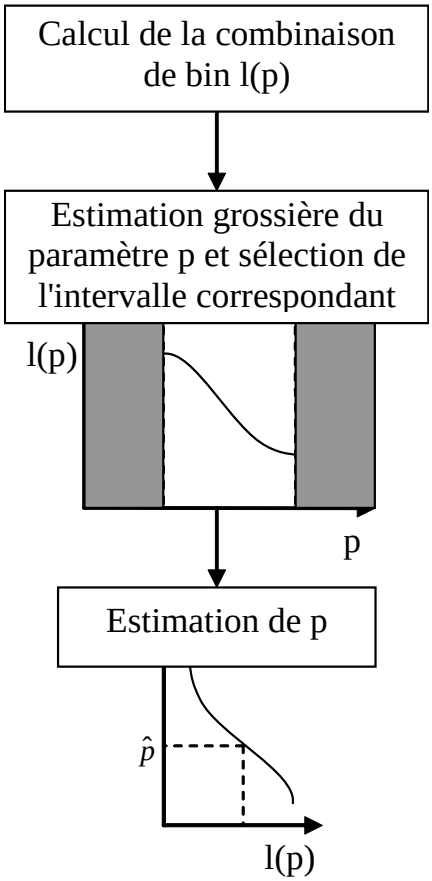
Plutôt que de chercher une formule explicite pour l'estimateur, on peut choisir de modéliser cette fonction. On peut par exemple l'approcher par une famille de fonctions, comme la famille des polynômes, on peut également décider de ne faire intervenir aucun modèle et de pré-tabuler les valeurs de la fonction. Les valeurs intermédiaires sont alors trouvées par interpolation. Le principal avantage de la modélisation est que l'on peut réduire de façon arbitraire l'erreur déterministe en augmentant l'ordre de modélisation, ou le nombre d'éléments prétabulés.

Dans la section 5.1.2, nous avons déjà souligné que lorsque l'on dérive un estimateur à partir d'une combinaison de bins $l$, il fallait que $l$ soit une fonction injective de $\mathbb{R}^U$ dans $\mathbb{R}^V$ ($V \geq U$), pour que l'estimateur existe. Or ce n'est pas le cas en général. La seule façon de rendre la fonction injective, c'est de restreindre l'intervalle d'analyse. En découpant l'intervalle de définition des paramètres judicieusement, chaque fragment de la courbe sera injectif et on pourra alors modéliser l'inverse de chaque fragment. Notre estimateur sera caractérisé par l'ensemble de ces modèles.

[5]Les propriétés de la TF sont rappelées dans l'annexe A.




(a) Dérivation de l'estimateur




(b) Algorithme d'estimation

FIG. 5.1: Résumé de la méthode de modélisation des estimateurs

La dernière étape est de déterminer pour une estimation donnée le bon intervalle de paramètres et d'utiliser le modèle correspondant. La TFCT découpe déjà le paramètre des fréquences de façon assez fine, et le choix du bon intervalle est réalisé par la détermination du maximum d'amplitude. Dans le cas du modèle **M01**, la méthode de la modélisation est donc facilement mise en oeuvre, et sera détaillée dans le section 6.1.2.

Dans le cas des modèles d'ordre supérieur, l'augmentation du nombre de paramètres rend la modélisation plus difficile, car elle devient multidimensionnelle. De plus le découpage en fréquence de la TFCT n'assure l'injectivité de $l$ que si l'intervalle de définition des paramètres non fréquentiels est un voisinage de zéro, même si certaines fonctions particulières peuvent être injectives partout, comme la plupart des estimateurs présentés dans la section 4. Pour des valeurs suffisamment grandes de ces paramètres, l'injectivité peut ne plus être vérifiée. Il faudra alors découper également leur intervalle de définition et mettre en place un mécanisme pour sélectionner le bon

intervalle. On peut noter qu'il existe des méthodes pour vérifier numériquement l'injectivité d'une fonction sur un intervalle, et également pour découper la fonction en intervalles injectifs [Lagrange et al., 2007]. Enfin, la sélection du bon intervalle pourra se faire en faisant une première estimation grossière des paramètres. La Figure 5.1 résume la méthode dans le cas général.

En conclusion bien qu'il doit être possible d'adapter cette méthode pour des modèles d'ordre supérieur quelconque, sa mise en place semble difficile. Nous nous contenterons donc de la mettre en place pour le modèle **M01** et d'en avoir esquissé les grandes lignes pour les autres modèles.

## 5.3 Le temps d'estimation

Pour une sinusoïde dont les paramètres varient avec le temps, la question du temps d'estimation se pose. En particulier on aimerait savoir quel est le meilleur temps d'estimation pour les modèles d'ordre supérieur à **M01**. Nous allons discuter de cette question dans cette section, d'abord en étudiant le comportement des bornes de Cramer-Rao, qui sont censées décrire le comportement de l'estimateur "optimal". Ensuite nous parlerons du temps donné par le réassignement.

### 5.3.1 Temps d'estimation et bornes de Cramer-Rao

Les bornes de Cramer-Rao pour les différents modèles étudiés dans cette thèse sont données dans la section 3.4.2. Au lieu d'estimer les paramètres du modèle au temps correspondant au milieu de la fenêtre d'analyse, on peut décaler ce temps d'estimation d'un temps $t_0$. Pour le modèle **M12**, le signal étudié est alors :

$$x_{12}(t) = \mathrm{e}^{\lambda_0+\mu(t-t_0)}\,\mathrm{e}^{j(\alpha_0+\beta_0(t-t_0)+\frac{\gamma}{2}(t-t_0)^2)} \tag{5.3.1}$$

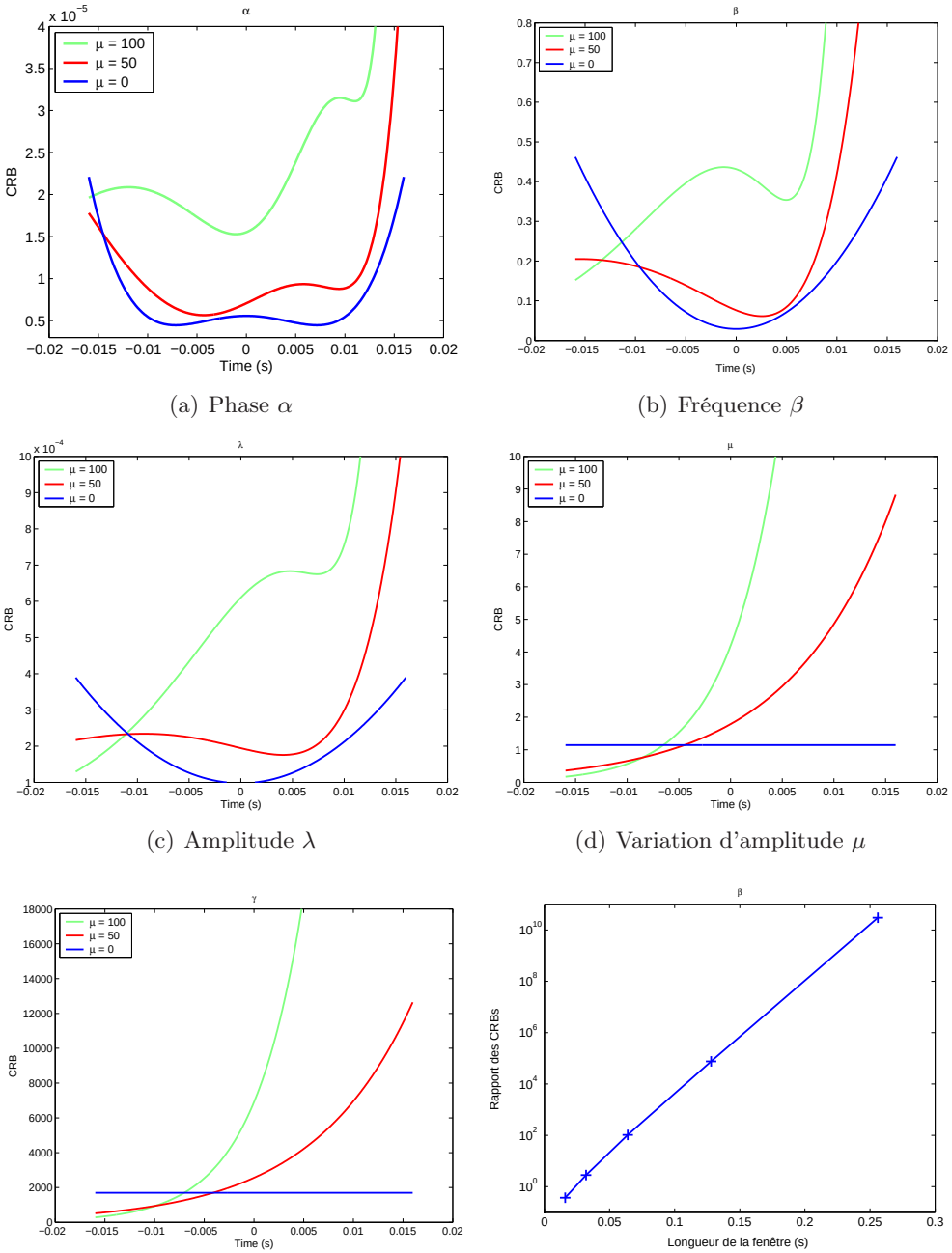
Les bornes de Cramer-Rao correspondant à ce modèle seront alors des fonctions du temps d'estimation $t_0$. En faisant varier le temps d'estimation sur toute la fenêtre d'analyse, on va trouver une valeur minimale pour la CRB qui nous donnera le meilleur temps d'estimation possible.

Pour exprimer les CRBs en fonction du temps d'estimation $t_0$, il suffit de redéfinir $\epsilon_q$[6] par :

$$\epsilon_q = \sum_{n=-M}^{M} (\tau_n - t_0)^q\,\mathrm{e}^{2\mu(\tau_n-t_0)} \tag{5.3.2}$$

Les courbes pour les paramètres du modèle **M12** sont tracées sur la figure 5.2. On voit que les minima des CRBs dépendent fortement de la valeur de $\mu$. De façon surprenante, on voit que pour de fortes valeurs de $\mu$, tous les paramètres, sauf la phase, présente une CRB minimum sur le bord de la fenêtre d'analyse le moins énergétique. Ce résultat est assez contre intuitif, car on aurait pu penser que les paramètres seraient plus facilement estimable là où l'énergie est la plus forte.

[6] voir section 3.4.2, équation (3.4.5).




(a) Phase $\alpha$

(b) Fréquence $\beta$

(c) Amplitude $\lambda$

(d) Variation d'amplitude $\mu$

(e) Variation de fréquence $\gamma$

(f) Rapport $\frac{CRB(t=0)}{CRB(t=-t_M)}$ pour $\beta$, en fonction de la taille de fenêtre et avec $\mu = 100$

FIG. 5.2: Evolution des CRBs pour les paramètres du modèle **M12** en fonction du temps d'estimation, pour une fenêtre de 32 ms et un SNR de 10.

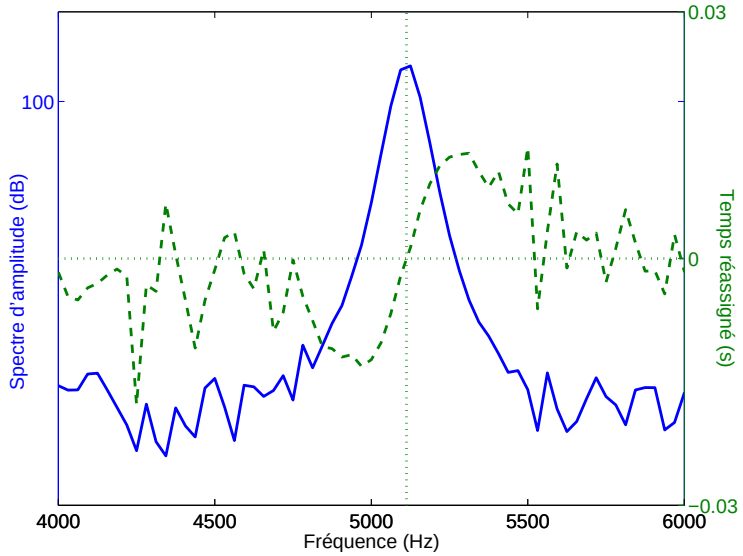



FIG. 5.3: Le temps du réassignement varie suivant la fréquence de la TFCT

En résumé, l'estimation de la phase a toujours une CRB proche du minimum pour une origine des temps centrée. Pour tous les autres paramètres, si aucune information n'est disponible sur le paramètre $\mu$, il vaut mieux choisir d'estimer tous les paramètres avec une origine centrée. Si l'on connaît le signe de $\mu$, alors les bornes de Cramer-Rao nous suggèrent qu'il est préférable de choisir une origine des temps excentrée pour l'estimation des paramètres $\mu$ et $\gamma$. Si $\mu$ est faible, il vaut mieux estimer les paramètres $\lambda$ et $\beta$ au centre de la fenêtre. Si $\mu$ est fort, alors il faut estimer $\lambda$ et $\beta$ avec une origine excentrée, c'est à dire sur le bord le moins énergétique de la fenêtre d'analyse.

Pour un estimateur optimal le gain d'estimation en terme de MSE, par rapport à une estimation centrée, devient important seulement dans le cas d'une très forte variation d'amplitude $\mu \geq 50$. Dans le cas où $\mu = 100$, le gain est d'un facteur 10 pour tous les paramètres sauf $\alpha$. Plus la fenêtre sera longue et plus le gain sera important, comme le montre la Figure 5.2(f) pour la fréquence.

### 5.3.2 Le temps d'estimation du réassignement

Dans la méthode du réassignement temps-fréquence, le temps d'estimation varie suivant le signal analysé. On peut donc se demander si le temps du réassignement n'est pas lié au temps optimal d'estimation.

La localisation de l'estimation est déplacée au centre de gravité des contributions énergétiques de la transformation temps-fréquence, ici le spectrogramme. Ce centre de gravité n'est cependant par lié de façon simple au centre de gravité de la sinusoïde. De plus, si pour un modèle **M02**, on calcule le réassignement pour différents bins d'une même TFCT, *i.e.* on fait varier la fréquence uniquement, le temps d'estimation ne va pas être constant, comme on peut le voir sur la Figure 5.3 pour une fenêtre de

Hann de 32 ms. Le temps du réassignement ne nous donne donc pas d'indice sur le meilleur instant d'estimation de la sinusoïde.

## 5.4 Quelques éléments de réflexion sur les propriétés statistiques des estimateurs

Tous les estimateurs qui seront présentés dans le prochain chapitre seront basés sur des combinaisons de bins. Soit $\hat{p}$ un estimateur du paramètre $p$, utilisant $Q$ bins FFT. Il aura la forme suivante[7] :

$$\hat{p} = g \circ f(X_1, ..., X_Q) \tag{5.4.1}$$

où $f$ est une fonction analytique sur un intervalle $I \subset \mathbb{C}^Q$ vers $\mathbb{C}$ et $g$ est une fonction de $\mathbb{C}$ vers $\mathbb{R}$ du type $\Re()$, $\Im()$, arg() ou $|.|$. Comme précédemment, les $X_i$ correspondent aux valeurs des bins, $X_i = X(t_i, \omega_i; h_i)$. L'estimateur possède généralement une erreur déterministe $\epsilon_D$, ce qui nous donne la relation suivante par rapport à la vraie valeur $p$ :

$$\hat{p} = p + \epsilon_D \tag{5.4.2}$$

Maintenant si l'on ajoute du bruit, que l'on suppose ici blanc Gaussien et centré, chaque paramètre $X_i$ sera perturbé par un bruit $N_i$ de moyenne nulle :

$$\hat{p} = g \circ f(X_1 + N_1, ..., X_Q + N_Q) \tag{5.4.3}$$

Pour trouver une expression de la variance de l'estimateur, la méthode la plus courante est de considérer un développement asymptotique de cette expression, lorsque le nombre d'échantillons $N$ tend vers l'infini. Dans le cas du modèle **M01**, cela implique que l'on aura[8] $N_i \ll X_i$ pour les bins $i$ proches du maximum du spectre d'amplitude. C'est une des deux conditions d'une transformée de Fourier bien résolue[9] : l'énergie des perturbations est négligeable par rapport à l'énergie de chaque bin FFT utilisé. L'hypothèse d'une transformée bien résolue est indispensable pour que l'estimation soit valide. Pour certains modèles, qui ne peuvent exister asymptotiquement, comme le modèle **M12**, cette hypothèse remplacera la condition asymptotique et conduira à des propriétés équivalentes.

Comme $N_i \ll X_i$, un développement de Taylor au premier ordre de $f$ nous donnera une très bonne approximation de l'erreur statistique :

$$f(X_1 + N_1, ..., X_Q + N_Q) = f(X_1, ..., X_Q) + \sum_{i=1}^{Q} N_i \frac{\partial f}{\partial X_i}(X_1, ..., X_i, ..., X_Q) \tag{5.4.4}$$

[7] $\circ$ est l'opérateur de composition de fonctions.

[8] Soit $h$ une fenêtre symétrique, positive réelle. D'une part $E(|N_i|^2) = \sigma^2 \sum_i h_i^2 \sim N$ d'après la relation B.3.1, en remarquant que $h^2$ est également une fenêtre symétrique, et d'autre part $|X_i|^2 \sim N^2$ d'après la relation B.3.2, pour les bins proches du maximum du spectre d'amplitude. On aura donc $|X_i|^2 \gg |N_i|^2$ pour $N$ suffisement grand.

[9] Voir la définition à la section 5.1.1.

Si $g$ est la fonction $\Re()$ ou $\Im()$, alors l'équation (5.4.3) se mettra sous la forme :

$$\begin{aligned} \hat{p} &= p + \epsilon_D + \epsilon_N \\ \epsilon_N &\triangleq g\big(\sum_{i=1}^{Q} N_i \frac{\partial f}{\partial X_i}(X_1, ..., X_i, ..., X_Q)\big) \end{aligned} \tag{5.4.5}$$

où $\epsilon_N$ est l'erreur stochastique. On voit immédiatement que son espérance est nulle.

Si $g$ est l'argument ou le module, d'après l'hypothèse de résolution on aura également :

$$f(X_1, ..., X_Q) \gg \sum_{i=1}^{Q} N_i \frac{\partial f}{\partial X_i}(X_1, ..., X_i, ..., X_Q) \tag{5.4.6}$$

et l'équation (5.4.3) aura une forme similaire au cas précédent :

$$\begin{aligned} \hat{p} &= p + \epsilon_D + \epsilon_N \\ \epsilon_N &\triangleq h\big(\frac{1}{f(X_1, ..., X_Q)} \sum_{i=1}^{Q} N_i \frac{\partial f}{\partial X_i}(X_1, ..., X_i, ..., X_Q)\big) \end{aligned} \tag{5.4.7}$$

$h$ sera la partie imaginaire $\Im()$ dans le cas où $g$ est la fonction argument ($g(x) = \arg(x)$) et la partie réelle $\Re()$ dans le cas $g$ est la fonction module ($g(x) = |x|$).

L'espérance de l'erreur quadratique, la MSE, pour tous les estimateurs étudiés aura donc la forme suivante :

$$E((p - \hat{p})^2) = \epsilon_D^2 + \epsilon_N^2 \tag{5.4.8}$$

Leur biais et leur variance seront donnés par :

$$E(p - \hat{p}) = \epsilon_D \tag{5.4.9}$$

$$E((p - \hat{p} - E(p - \hat{p}))^2) = \epsilon_N^2 \tag{5.4.10}$$

Il faudra donc trouver l'erreur déterministe pour trouver le biais de l'estimateur et ensuite calculer $\epsilon_N$ pour trouver la variance de l'estimateur. Dans les expériences, on est amené à faire une moyenne de ces erreurs par rapport à $p$. Or les erreurs $\epsilon_D$ et $\epsilon_N$ dépendent de $p$ dans le cas général, donc les courbes des expériences vont en fait faire intervenir $E_p(\epsilon_D^2)$ et $E_p(\epsilon_N^2)$. Par abus de langage, on appelera simplement "biais" ou "erreur déterministe" $E_p(\epsilon_D^2)$, et "variance" ou "erreur stochastique" $E_p(\epsilon_N^2)$, lors de l'analyse des courbes.

On a supposé dans toute cette section que la perturbation $N_i$ du signal $X_i$ était stochastique. Dans le cas d'une erreur $D_i$ non plus stochastique, mais déterministe, mais vérifiant toujours $D_i \ll X_i$, on peut utiliser exactement le même raisonnement et les formules exprimant l'erreur d'estimation (5.4.5) et (5.4.7) seront les mêmes.

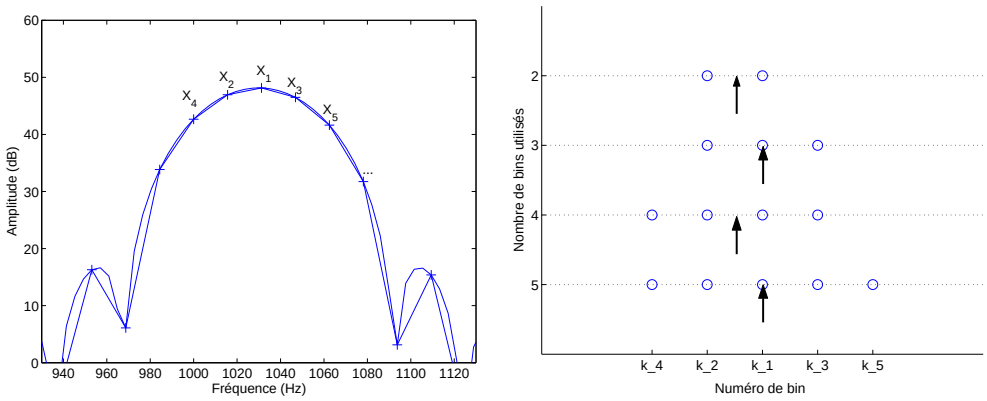
## 5.5 Conclusion

Dans cette section nous avons mis en évidence un certain nombre de mécanismes communs à tous les estimateurs basés sur la transformée de Fourier. Nous avons tout d'abord montré que ces estimateurs, de façon similaire à la méthode du maximum de vraisemblance, cherchaient tout d'abord une expression indépendante des paramètres linéaires, ici la phase et l'amplitude initiale. Ensuite, nous avons remarqué que les méthodes cherchaient à inverser cette fonction des paramètres d'ordre supérieur. Afin de rendre cette inversion plus simple, la plupart des méthodes proposent de linéariser cette fonction en utilisant des approximations, comme des approximations de Taylor. Nous avons également esquissé un principe général d'inversion des fonctions, qui éviterait d'avoir recours à une formule analytique approchée. Cette méthode constituerait une alternative à l'optimisation multidimensionnelle des méthodes du type maximum de vraisemblance. Dans le cas général, ce principe semble cependant difficile à mettre en oeuvre.

CHAPITRE 6

# Développement de nouveaux estimateurs

Dans ce chapitre nous nous penchons sur les estimateurs développés dans le cadre de la thèse. Ils sont le résultat des idées présentées dans le chapitre précédent (chapitre 5) et ont fait l'objet de plusieurs publications. Cependant leur présentation dans ce chapitre sera parfois différente, car nous avons tenté de généraliser les résultats lorsque cela s'est avéré possible. Nous commencerons par décrire dans la section 6.1 une généralisation des méthodes d'estimation de fréquence présentées dans la section 4.1.2.4 pour le modèle **M01** . Cette généralisation avait déjà été esquissée dans l'article [Betser et al., 2006c]. Dans la section 6.2, nous décrirons des méthodes d'estimation de fréquence pour les modèles **M02** et **M12**, basée sur le principe du vocodeur de phase. Ces méthodes ont été publiées dans les IEEE Transactions on Signal Processing [Betser et al., 2008]. Puis dans la section 6.3 deux nouvelles méthodes d'estimation permettant de trouver tous les paramètres d'ordre supérieurs du modèle **M12** seront présentée. Nous terminerons enfin ce chapitre par la section 6.4 dans laquelle nous développerons de nouveaux estimateurs d'amplitude et de phase, généralisant le travail présenté dans [Betser et al., 2006b] à un modèle d'ordre quelconque. En particulier nous appliquerons la méthode à un modèle d'ordre **M12**. En combinant les estimateurs des sections 6.4 et 6.3 on dispose donc d'une nouvelle méthode complète pour estimer les paramètres du modèle **M12**. Pour chaque méthode nous étudierons les propriétés statistiques des estimateurs, et nous ferons une comparaison avec les méthodes classiques d'estimation présentées dans la section 4.




(a) Numérotation des bins par ordre d'amplitude.

(b) Centres de symétrie des bins utilisés.

FIG. 6.1: Choix des bins dans la méthode d'estimation de fréquence basée sur le modèle **M01**. Ici $X_i = X(t, \omega_i; h)$ et $\omega_i$ est la fréquence du bin $k_i$.

## 6.1 Estimation de fréquence pour le modèle M01

### 6.1.1 Approximations de Taylor

La première étape pour trouver un estimateur basé sur la TFCT est de choisir une combinaison de bins qui permette d'éliminer l'amplitude et la phase initiale[1]. Dans cette section, nous considérerons des rapports de bins à l'intérieur d'une même trame, qui permettent d'éliminer les deux paramètres en même temps. Il s'agit de la forme de combinaison utilisée dans la littérature par les interpolateurs de spectre utilisant la phase[2]. D'une façon générale, ce type de rapport peut être mis sous la forme :

$$\mathcal{H} = \frac{\sum_q \kappa_q X(t, \omega_q; h)}{\sum_l \nu_l X(t, \omega_l; h)} \tag{6.1.1}$$

où $\kappa_q$ et $\nu_l$ sont des coefficients complexes. $h$ est une fenêtre symétrique réelle.

La deuxième étape consiste à utiliser le développement de Taylor (5.2.1) en $\omega_0 = \beta$, pour linéariser l'équation par rapport à la fréquence. Il n'est cependant pas judicieux d'appliquer directement cette approximation à une combinaison complètement quelconque de transformées de Fourier. En effet, dans le cas général, $X(t, \beta; \tau h) = 0$ à cause de la symétrie de $h$ et il faudrait alors faire un développement à l'ordre 2 pour obtenir une fonction dépendant de $\beta$. La fonction serait donc en $\beta^2$. Il est préférable de choisir une combinaison de bins symétrique qui permet de se ramener à un rapport de deux transformées de Fourier uniquement, du type :

$$\mathcal{H} = \frac{\xi_{as} X(t, \omega_{as}; h_{as})}{\xi_{sy} X(t, \omega_{sy}; h_{sy})} \tag{6.1.2}$$

[1] Les différentes étapes pour former un estimateur ont été discutées à la section 5.1.2.

[2] Ces estimateurs ont été décrits à la section 4.1.2.4.

TAB. 6.1: Exemples de combinaison de bins $X_q$ conduisant à des fenêtres $h^{'}$ symétriques (S) ou antisymétriques (AS) : $\sum_q \kappa_q X_q = \xi^{'} X(t, \omega^{'}; h^{'})$. $c$ est une constante réelle.

| Combinaison | $h^{'}$ | $\omega^{'}$ | $\xi^{'}$ | S/AS |
|---|---|---|---|---|
| $X_1$ | $h(\tau)$ | $\omega_1$ | 1 | S |
| $X_2 - X_1$ | $\sin(\delta\omega_1\tau)h(\tau)$ | $\omega_b$ | 2j | AS |
| $X_3 - X_2$ | $\sin(\delta\omega_2\tau)h(\tau)$ | $\omega_1$ | 2j | AS |
| $X_1 + X_2$ | $\cos(\delta\omega_1\tau)h(\tau)$ | $\omega_b$ | 2 | S |
| $X_2 + X_3 + c \cdot X_1$ | $(\cos(\delta\omega_2\tau) + .5c)h(\tau)$ | $\omega_1$ | 2 | S |
| $c \cdot (X_1 + X_2) + X_3 + X_4$ | $(c\cos(\delta\omega_1\tau) + \cos(\delta\omega_3\tau))h(\tau)$ | $\omega_b$ | 2 | S |
| $c \cdot (X_2 - X_1) + X_4 - X_3$ | $(c\sin(\delta\omega_1\tau) + \sin(\delta\omega_3\tau))h(\tau)$ | $\omega_b$ | 2j | AS |
| Avec : | $\delta\omega_1 = .5(\omega_2 - \omega_1)$ | $\delta\omega_2 = .5(\omega_3 - \omega_2)$ | | |
| | $\delta\omega_3 = .5(\omega_4 - \omega_3)$ | $\omega_b = .5(\omega_2 + \omega_1)$ | | |

où $\xi_{as}$ et $\xi_{sy}$ sont deux constantes complexes, $\omega_{as}$ et $\omega_{sy}$ sont les centres de symétrie des bins utilisés (cf. Figure 6.1(b)). Enfin $h_{as}$ et $h_{sy}$ sont deux nouvelles fenêtres, telles que $h_{as}$ soit antisymétrique et $h_{sy}$ soit symétrique. On applique l'approximation de Taylor (5.2.1) en $\omega_0 = \beta$ à l'ordre 1, au numérateur et au dénominateur :

$$X(t, \omega; h_{as}) \approx X(t, \beta; h_{as}) - j(\omega - \beta)X(t, \beta; \tau h_{as})$$
$$X(t, \omega; h_{sy}) \approx X(t, \beta; h_{sy}) - j(\omega - \beta)X(t, \beta; \tau h_{sy})$$

Comme $h_{as}$ est antisymétrique et $h_{sy}$ symétrique, on aura $X(t, \beta; h_{as}) = X(t, \beta; \tau h_{sy}) = 0$. En posant $\Gamma(h) = \sum_{i=-(N-1)/2}^{(N-1)/2} h(\tau_i)$, on aura également $X(t, \beta; h) = A\,\mathrm{e}^{\alpha}\,\Gamma(h)$ si $h$ est symétrique. On en déduit pour le rapport $\mathcal{H}$ :

$$\mathcal{H} \approx j(\beta - \omega_{as})\frac{\xi_{as}\Gamma(\tau h_{as})}{\xi_{sy}\Gamma(h_{sy})} \tag{6.1.3}$$

où sont des constantes réelles ne dépendant que des fenêtres utilisées. On trouve alors un estimateur de $\beta$ en inversant cette équation :

$$\hat{\beta} = \omega_{as} + \frac{\Gamma(h_{sy})}{\Gamma(\tau h_{as})}\Re\left(\frac{\mathcal{H}\xi_{sy}}{j\xi_{as}}\right) \tag{6.1.4}$$

Enfin pour que l'approximation soit la plus précise possible, il vaut mieux choisir $\omega_{sy}$ et $\omega_{as}$ le plus près possible de $\beta$.

Le tableau 6.1 donne quelques exemples de combinaison pouvant conduire à des fenêtres symétriques ($h_{sy}$) ou antisymétriques ($h_{as}$). La numérotation des bins $X_i = X(t, \omega_i; h)$ est croissante avec l'éloignement du maximum du spectre d'amplitude, comme indiqué sur la Figure 6.1(a). Bien d'autres combinaisons sont certainement possibles, en faisant par exemple intervenir plus de bins comme $k_5$, $k_6$ etc. Notons que les équivalences présentées dans ce tableau sont vraies quelles que soit le signal analysé. Par contre la méthode de dérivation ne sera plus valide pour un modèle plus complexe que **M01**. En effet la symétrie de la fenêtre ne sera plus suffisante pour que certains termes du développement de Taylor s'annulent.

Pour former un estimateur, il suffit de prendre une combinaison antisymétrique au numérateur et symétrique au dénominateur. Certains rapports symétrique/antisymétrique ont déjà été utilisés dans [Macleod, 1998], [Aboutanios and Mulgrew, 2005] et [Quinn, 1994][3], mais pour des fenêtres rectangulaires uniquement. Nous allons maintenant détailler un exemple d'estimateur en utilisant cette méthode, et nous donnerons ensuite les propriétés statistiques, *i.e.* biais et variance, de l'estimateur.

#### 6.1.1.1 Exemple d'estimateur à deux bins

Nous allons prendre comme exemple l'estimateur développé dans une de nos publications [Betser et al., 2006c]. La combinaison de bins utilisée est la suivante :

$$\mathcal{H} = \frac{X(t,\omega_1;h) - X(t,\omega_2;h)}{X(t,\omega_1;h) + X(t,\omega_2;h)} \tag{6.1.5}$$

$$= j\,\frac{X(t,\omega_b;h_s)}{X(t,\omega_b;h_c)} \tag{6.1.6}$$

où $\omega_b$, $h_s$ et $h_c$ nous sont donnés par le Tableau 6.1, $h_s(\tau) = \sin(\delta\omega_1\tau)h(\tau)$ et $h_c(\tau) = \cos(\delta\omega_1\tau)h(\tau)$.

On applique les équations (6.1.3) et (6.1.4) avec $h_{sy} = h_c$, $h_{as} = h_s$, $\xi_{as} = 2j$, $\xi_{sy} = 2$ et $\omega_{as} = \omega_b$. Pour un modèle **M01**, le développement de Taylor à l'ordre 1, en $\beta$, du dénominateur et du numérateur nous donne :

$$\mathcal{H} \approx (\omega_b - \beta)\frac{\Gamma(\tau h_s)}{\Gamma(h_c)} \tag{6.1.7}$$

Et on obtient finalement l'estimateur suivant :

$$\hat{\beta} = \omega_b - \frac{\Gamma(h_c)}{\Gamma(\tau h_s)}\Re(\mathcal{H}) \tag{6.1.8}$$

La méthode complète est la suivante :

1. Initialisation : calculer $\Gamma(\tau h_s)$ et $\Gamma(h_c)$ pour la fenêtre d'analyse $h$ de la FFT.
2. Calculer la FFT zéro-phase pour le temps $t$
3. Sélectionner le bin maximum $k = \arg\max_i |X(t,\omega_i;h)|$ et le deuxième bin maximum $k' = \arg\max_{i\in\{k+1,k-1\}} |X(t,\omega_i;h)|$
4. Calculer le rapport $\mathcal{H} = \frac{X(t,\omega_k;h) - X(t,\omega_{k'};h)}{X(t,\omega_k;h) + X(t,\omega_{k'};h)}$
5. Calculer la fréquence estimée :
   $\hat{\beta} = \omega_b - \Re(\mathcal{H})\frac{\Gamma(h_c)}{\Gamma(\tau h_s)}$, où $\omega_b = (\omega_k + \omega_{k'})/2$

#### 6.1.1.2 Propriétés statistiques

Nous allons maintenant énoncer quelques propriétés concernant le biais et la variance de l'estimateur. La plupart des démonstrations seront omises. Pour plus de détails on se référera à l'article [Betser et al., 2006c], qui a été placé dans l'annexe G.

[3] Ces estimateurs sont décrits à la section 4.1.2.4.

| | Han | Ham | Rec | Bla | Gau |
|---|---|---|---|---|---|
| $b_1$ | 1.5e-4 | 2.1e-2 | 3.9e-6 | 5.4e-3 | 2.4e-2 |
| $b_2$ | 1.3e-1 | 1.4e-1 | 2.5e-1 | 1.0e-1 | 1.3e-1 |
| Borne sur l'erreur(Hz) | 2.6e-3 | 3.8e-1 | 8.3e-5 | 9.4e-2 | 4.3e-1 |

TAB. 6.2: Valeurs des bornes pour différentes fenêtres d'analyse.

**Biais**

Pour trouver une expression du biais, on va utiliser l'expression du développement de Taylor avec reste de Lagrange. On rappelle que $\Gamma(\omega; h)$ est la transformée de Fourier zéro-phase de $h$ pour la fréquence $\omega$ et que $\Gamma(h) = \Gamma(0; h)$ et on note $\delta = \omega_b - \beta$. Pour le modèle **M01**, on aura $\Gamma(\tau^i h_s) = 0$ si $i$ est pair, et $\Gamma(\tau^i h_c) = 0$ si $i$ est impair, ce qui simplifie le développement en :

$$\mathcal{H} = \frac{\Gamma(\tau h_s)\delta - \Gamma(c_1; \tau^3 h_s)\frac{\delta^3}{6}}{\Gamma(h_c) - \Gamma(c_2; \tau^2 h_c)\frac{\delta^2}{2}} \tag{6.1.9}$$

$$= \delta \frac{\Gamma(\tau h_s)}{\Gamma(h_c)} \frac{\big(1 - P\big)}{\big(1 - Q\big)} \tag{6.1.10}$$

où $c_1$ et $c_2$ sont deux constantes dans $[0, \delta]$. $P$ et $Q$ sont les restes de Lagrange,

$$P \triangleq \frac{\Gamma(c_1; \tau^3 h_s)}{\Gamma(\tau h_s)} \frac{\delta^2}{6}, \qquad Q \triangleq \frac{\Gamma(c_2; \tau^2 h_c)}{\Gamma(h_c)} \frac{\delta^2}{2} \tag{6.1.11}$$

Avec ces définitions, l'erreur de l'estimateur (6.1.8) peut s'écrire ainsi :

$$\beta - \hat{\beta} = \frac{(Q - P)}{(1 - Q)} \delta \tag{6.1.12}$$

Comme $\beta$ est dans l'intervalle $[\omega_1, \omega_2]$, $|\delta|$ est donc inférieur à la demi-précision de la TFCT : $R = \pi F/N$.

On suppose que la fenêtre $h$ est symétrique, réelle, positive, normalisée ($h(t) \leq 1$) et que son énergie est concentrée au milieu de la fenêtre, c'est à dire qu'elle vérifie la propriété suivante :

$$\sum_{|n| \leq N/2} h(n) \geq 2 \sum_{N/4 \leq |n| \leq N/2} h(n) \tag{6.1.13}$$

Toutes les fenêtres usuelles vérifient ces propriétés. On peut alors montrer que l'erreur $\beta - \hat{\beta}$ est bornée et qu'elle est $O(N^{-1})$. Pour certaines fenêtres, $P$ et $Q$ seront du même ordre, et l'erreur pourra avoir un ordre encore plus faible. Il a été démontré que l'estimateur est $O(N^{-2})$ pour la fenêtre rectangulaire [Aboutanios and Mulgrew, 2005].

On note maintenant $b_1$ et $b_2$ les bornes sur $|P-Q|$ et $|Q|$ respectivement.

$$b_1 \triangleq \frac{\Gamma(\tau^2 h_c)}{\Gamma(h_c)}\frac{R^2}{2} \tag{6.1.14}$$

$$b_2 \triangleq \sum_{i=1}^{\infty} \frac{R^{2i}}{(2i+1)!}\left|\frac{\Gamma(\tau^{2i+1}h_s)}{\Gamma(\tau h_s)} - (2i+1)\frac{\Gamma(\tau^{2i}h_c)}{\Gamma(h_c)}\right| \tag{6.1.15}$$

La borne $b_2$ s'obtient en développant le reste de Lagrange $P$ en série, et en utilisant l'inégalité triangulaire. $b_2$ est une série infinie, mais la règle d'Alembert prouve que la série converge, et comme $R$ est généralement petit, elle va converger très vite. Les quelques premiers termes vont donc constituer une très bonne approximation de cette borne. Le tableau 6.2 donne les valeurs des bornes pour les fenêtres usuelles avec $N = 512$ et $F = 16000$. Comme $b_2$ est $O(1)$, on voit que si $N$ est suffisamment grand, on aura $b_2 < 1$. On en déduit la borne suivante sur l'erreur :

$$|\beta - \hat{\beta}| \leq \frac{b_1}{1-b_2}|\delta| \leq \frac{b_1}{1-b_2}R \tag{6.1.16}$$

**Variance**

Dans l'article [Betser et al., 2006c], la variance théorique de l'estimateur a été trouvée en utilisant un développement asymptotique de l'estimateur, en s'inspirant de la méthode décrite par Quinn [Quinn, 1994], [Quinn and Hannan, 2001]. Par manque de place, les calculs intermédiaires ont été omis dans [Betser et al., 2006c]. En Annexe D, on donne une démonstration complète en utilisant cette fois la méthode décrite dans la Section 5.4. On arrive à un résultat identique[4] :

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2\Gamma(h_c)^2}{2A^2\Gamma(\tau h_s)^2\Gamma(\delta;h_c)^2}\left[\Gamma(h_s^2) + \frac{\Gamma(\delta;h_s)}{\Gamma(\delta;h_c)}\Gamma(h_c^2)\right] \tag{6.1.17}$$

Dans le cas le plus défavorable $\delta = R$, on a la borne suivante sur la variance de l'estimateur :

$$\mathrm{var}(\hat{\beta}) \leq \frac{\sigma^2\Gamma(h_c)^2}{2A^2\Gamma(\tau h_s)^2\Gamma(R;h_c)^2}\left[\Gamma(h_s^2) + \frac{\Gamma(R;h_s)}{\Gamma(R;h_c)}\Gamma(h_c^2)\right] \tag{6.1.18}$$

La Figure 6.2 donne un exemple d'utilisation de ces formules, pour la fenêtre de Hann et un SNR de 30 dB.

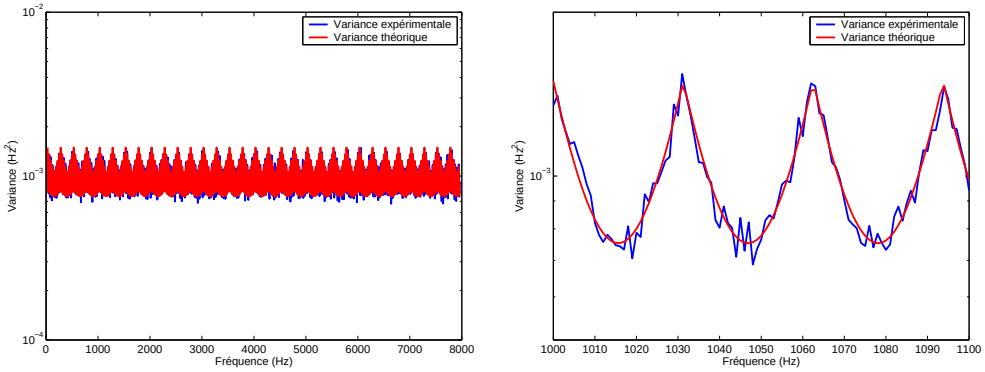
En utilisant les formules asymptotiques des fenêtres données en annexe B, on peut montrer que si $N$ est suffisamment grand, la variance peut s'exprimer ainsi :

$$\mathrm{var}(\hat{\beta}) = C{t_N}^{-3}\eta^{-1} \tag{6.1.19}$$

où $C$ est une constante, $t_N = N/F$ est la taille de la fenêtre d'analyse et $\eta$ est le SNR. On constate que la variance de l'estimateur est en $N^{-3}$, comme la borne théorique[5].

[4] Il y a une erreur typographique dans l'article [Betser et al., 2006c] sur cette équation. En effet à l'équation 4.4, il faut remplacer $2\sigma^2$ par $\sigma^2/2$.

[5] Voir le tableau 3.1.




(a) Variance (b) Variance (zoom)

FIG. 6.2: Variance théorique et expérimentale de l'estimateur pour une fenêtre de Hann de 32 ms et un SNR de 30 dB.

#### 6.1.1.3 Performances

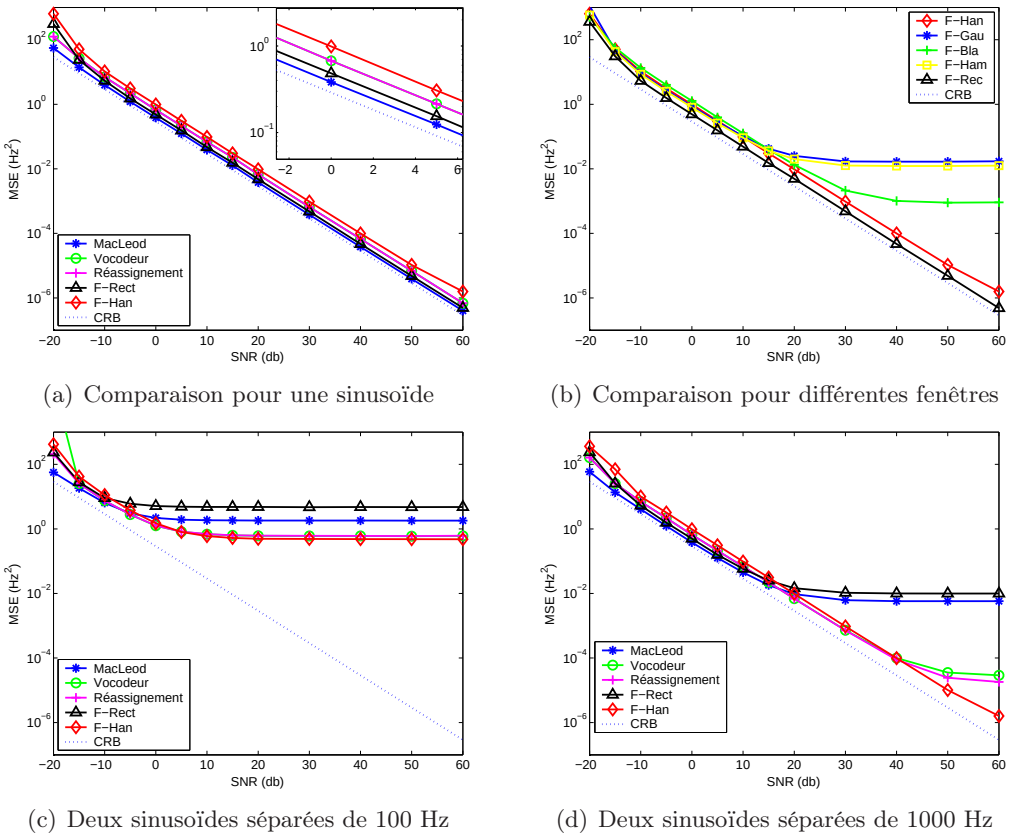
Le nouvel estimateur décrit dans les paragraphes précédents sera appelé 'F', suivi des trois premières lettres de la fenêtre utilisée. La Figure 6.3(b) montre les performances de cet estimateur, pour une sinusoïde et pour différentes fenêtres d'analyse. Lorsque le SNR croit, le bruit devient négligeable, et le biais dû à l'approximation de Taylor apparaît.

L'estimateur est ensuite comparé toujours pour une sinusoïde, avec deux autres méthodes, la méthode de Macleod (Section 4.1.2.4) et la méthode du vocodeur de phase (Section 4.1.2.1). Les méthodes utilisant la fenêtre rectangulaire sont toujours légèrement meilleures dans le cas d'une sinusoïde. Avec la fenêtre rectangulaire notre estimateur donne des résultats très proches de la méthode de Macleod. Les deux méthodes sont en fait très similaires, le léger avantage de l'estimateur de Macleod provient de l'utilisation d'un développement de Taylor à l'ordre 2. Si on regarde la formule de l'estimateur (4.1.37), on s'aperçoit qu'elle correspond à une racine de polynôme d'ordre 2. Le gain de performance obtenu avec un développement supérieur reste négligeable.

Pour des SNRs bas (-20dB), les performances de F-Rec et F-Han chutent plus vite que pour les autres méthodes. Cela est dû à l'asymétrie de la combinaison de bins utilisée. On a besoin de trouver le maximum et le second bin le plus énergétique. Pour ces niveaux de bruit, le second maximum peut s'avérer difficile à trouver[6].

Les Figures 6.3(c) et 6.3(d) illustrent l'ajout d'une perturbation d'une deuxième sinusoïde de même amplitude avec 100 Hz et 1000 Hz d'écart respectivement par rapport à la première. On voit évidemment l'intérêt de pouvoir utiliser une fenêtre de Hann. La bonne atténuation de ses lobes secondaires permet un gain significatif de performance. La méthode F-Han apparaît donc comme un bon compromis avec

[6]L'estimateur à 3 bins de Macleod ne présente pas ce problème car on besoin de déterminer le bin maximum uniquement.




(a) Comparaison pour une sinusoïde

(b) Comparaison pour différentes fenêtres

(c) Deux sinusoïdes séparées de 100 Hz

(d) Deux sinusoïdes séparées de 1000 Hz

FIG. 6.3: Performances de l'estimateur

des performances comparables aux méthodes les plus utilisées, vocodeur de phase et réassignement, pour le modèle **M01**.

### 6.1.2 Algorithme Général pour les estimateurs modélisés

Comme précédemment, la première étape est toujours de choisir une combinaison $\mathcal{H}$ de bins de la TFCT, qui ne dépende que du paramètre de fréquence.

La méthode pour dériver un estimateur modélisé à partir d'une combinaison a déjà été esquissée dans la section 5.2.2. Nous allons maintenant détailler le cas du modèle **M01**. L'étape suivante consiste à estimer les paramètres de la fonction de modélisation. Pour une combinaison $\mathcal{H}$ quelconque de points TFCT qui élimine $\alpha$ et $A$, c'est à dire que l'on a $\mathcal{H} = l(\beta)$, la fonction $l$ va dépendre du bin de référence $k$ choisi pour l'estimation. On va donc devoir estimer un modèle différent pour chaque bin TFCT $k$. Le temps de référence est préférablement choisi comme étant le temps moyen des bins utilisés, la fréquence de référence sera préférablement un bin maximum. La fréquence que l'on cherche à estimer sera donc dans l'intervalle $[-R + \omega_k, R + \omega_k]$ où $R$ est la demi-précision de Fourier. Chaque fonction inverse $l_k^{-1}$ n'a donc besoin d'être modélisée que pour cet intervalle. Le modèle utilisé peut

être par exemple une interpolation polynomiale, dont les paramètres sont estimés en utilisant un critère des moindres carrés. Pour certaines combinaisons de bins, $l_k^{-1}$ sera identique pour tous les bins $k$.

L'algorithme pour modéliser l'estimateur est le suivant :

1. Pour chaque bin $k \in [0, P/2[$
   a) Initialiser la table des couples $(\mathcal{H}, \beta)$ à zéro.
   b) Soit $\Omega = \{\omega_k + \frac{2\pi i F}{P.Q} \,/\, i \in [0, Q]\}$ un ensemble contenant les valeurs échantillonnées de l'intervalle $[\omega_k, \omega_{k+1}]$.
   c) Pour chaque fréquence $\beta$ dans $\Omega$
      i. Génération de $x(n) = \mathrm{e}^{j\beta\tau_n}$ pour $n \in [0, N + H[$.
      ii. Calcul des FFTs fenêtrées de $x$ pour les temps utilisés dans le rapport $\mathcal{H}$.
      iii. Calcul de $\mathcal{H}$.
      iv. Mise à jour de la table des couples $(\mathcal{H}, \beta)$.
   d) En utilisant la table, estimation des paramètres du modèle pour le bin $k$ : $l_k^{-1}(\mathcal{H}) = \beta$.

L'algorithme d'estimation est le suivant :

1. Calculer les FFTs pour les temps requis
2. Pour chaque bin maximum $k$ correspondant au temps de référence $t_r$ :
3. Calculer $\mathcal{H}$
4. Calculer $\hat{\beta} = l_k^{-1}(\mathcal{H})$

#### 6.1.2.1 Evaluation

On va évaluer l'algorithme pour deux estimateurs différents. Le premier est l'estimateur décrit dans la section 6.1.1.1. Le deuxième est un nouvel estimateur, qui va illustrer les possibilités et les inconvénients de la méthode.

**Application à un estimateur existant biaisé**

On reprend l'estimateur décrit par l'équation (6.1.8). Cet estimateur est biaisé, mais l'utilisation de l'algorithme de modélisation sur le rapport (6.1.5) va nous permettre de réduire ce biais. Pour cet estimateur, la fonction inverse sera identique pour tous les bins $k$.

La Figure 6.4 montre une comparaison de l'estimateur modélisé par rapport à l'estimateur analytique approché de la section précédente. La Figure est tracée pour une modélisation polynomiale d'ordre 5. On voit qu'un ordre de modélisation peu élevé permet déjà de réduire de façon significative le biais, sans changer la variance. En augmentant l'ordre de modélisation, on peurra réduire arbitrairement le biais.

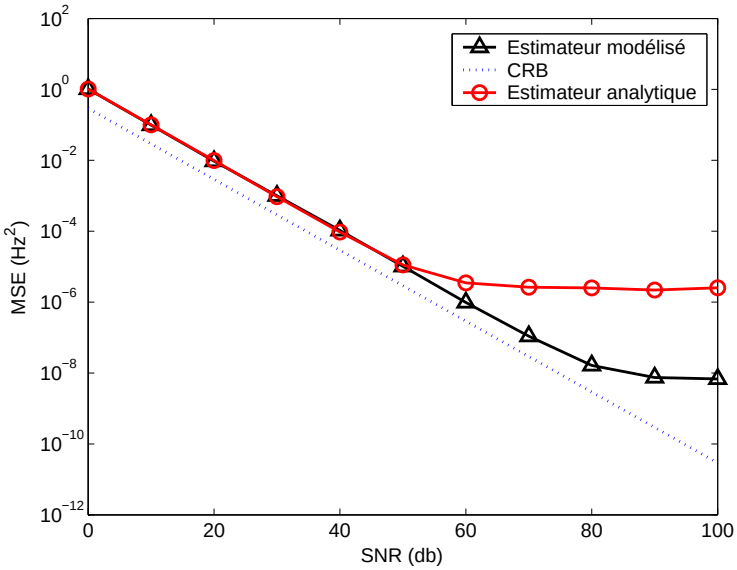



FIG. 6.4: Comparaison de l'estimateur dérivé par approximation de Taylor et de l'estimateur modélisé

### Application à un nouvel estimateur

On note ici $X(m,k) = X(t_m, \omega_k; h)$. Afin d'illustrer les possibilités de dérivation de nouveaux estimateurs, nous allons choisir une combinaison de bin des quatre points temps-fréquence $X(m,k)$, $X(m,k+1)$, $X(m+1,k)$ et $X(m+1,k+1)$ :

$$\mathcal{H}_{new} = \arg\left(\frac{X(m,k).\bar{X}(m,k+1) - X(m,k+1)\bar{X}(m+1,k)}{X(m,k)\bar{X}(m+1,k+1)}\right) \tag{6.1.20}$$

Dériver analytiquement un estimateur de cette combinaison est plus difficile que pour l'estimateur précédent, car il n'y a pas de symétrie entre les bins utilisés. L'utilisation du bin $X(m+1,k)$ au numérateur va également introduire un déphasage qui va varier en fonction du bin $k$ choisi. Donc contrairement à l'exemple précédent, on aura une fonction inverse différente par bin $k$ choisi.

La Figure 6.5 représente l'erreur de l'estimateur en fonction de la fréquence, sans ajout de bruit. La modélisation est réalisée avec un polynôme d'ordre 5 comme précédemment. On peut voir que l'erreur est globalement assez faible sur l'ensemble du spectre, excepté un pic aux alentours de 100 Hz. Ce pic est dû à une discontinuité de la combinaison (6.1.20) par rapport à la fréquence, illustré sur la Figure 6.6(a). A titre de comparaison la Figure 6.6(b) donne l'allure de la fonction pour un autre bin $k$, inversible cette fois. La discontinuité sur la Figure 6.6(a) est due à un problème de déroulement de phase causé par la fonction arg(), qui projette les valeurs sur $[0, 2\pi]$. L'erreur pourrait sûrement être évitée, soit en effectuant un déroulement de phase correct, soit en modifiant légèrement la combinaison de bins utilisée.

Cet exemple montre que l'algorithme permet facilement de dériver un estimateur à partir d'une combinaison, même si on voit clairement qu'une combinaison prise au

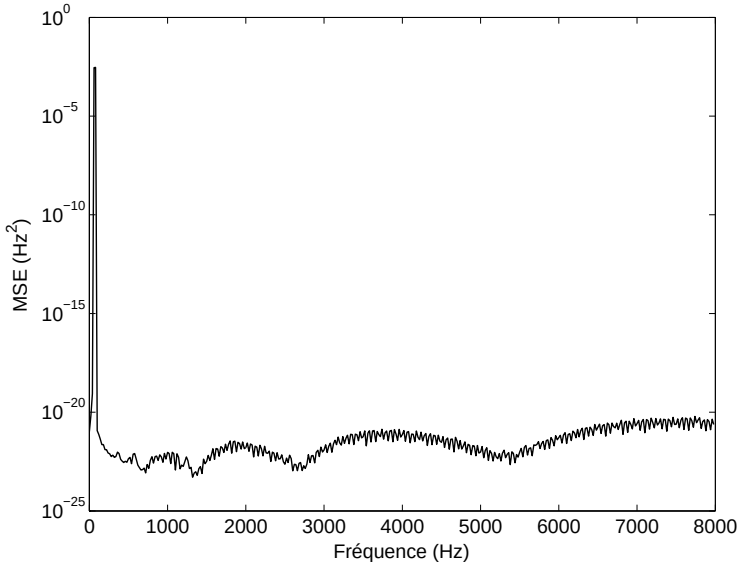
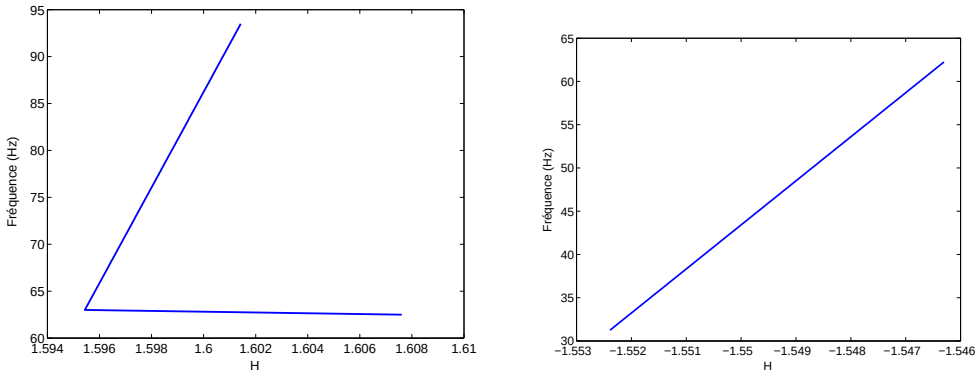



FIG. 6.5: Performances de l'estimateur (6.1.20) sans ajout de bruit




(a) Fonction non inversible (b) Fonction inversible

FIG. 6.6: Deux exemples de fonction à modéliser

hasard peut présenter des parties non inversibles, particulièrement avec la fonction arg().

## 6.2 Estimation de fréquence pour les modèles M02 et M12 basée sur la méthode du vocodeur de phase

L'estimateur de fréquence utilisé dans le vocodeur de phase, présenté pour le modèle **M01** dans la section 4.1.2.1, repose sur une dérivation discrète de la phase. On avait remarqué que la dérivation discrète donnait un estimateur exact pour les

modèles **M01** et **M11**. En fait, ce principe peut être étendu à un modèle avec phase quadratique et un ordre quelconque en amplitude (modèle **MK2**).

Considérons deux trames centrées respectivement en $t_{m_1}$ et $t_{m_2}$ telles que $t_M - t_{m_1} = t_{m_2} - t_M = \frac{T}{2}$. Sur l'intervalle d'analyse $W$, le paramètre d'ordre 2 en phase est considéré comme constant. Donc pour un temps $t_{m_i}$ sur cet intervalle, la phase correspondante est $\alpha_i = \alpha_M + \beta_M \tau_{m_i} + \gamma \frac{\tau_{m_i}^2}{2}$, et la fréquence locale est égale à $\beta_i = \beta_M + \gamma \tau_{m_i}$. On en déduit donc la relation suivante :

$$\beta_M = \frac{\alpha_2 - \alpha_1}{T} \tag{6.2.1}$$

La dérivation discrète de la phase nous donne un estimateur de fréquence pour l'instant $t_M$.

La principale différence avec le modèle **M01**, c'est que la transformée de Fourier n'est plus un estimateur direct de la phase dans le cas d'un modèle d'ordre 2 en phase. En fait, la variation de fréquence va introduire un terme d'erreur $\Gamma$, qui s'exprime de la façon suivante[7] pour le modèle **M12** :

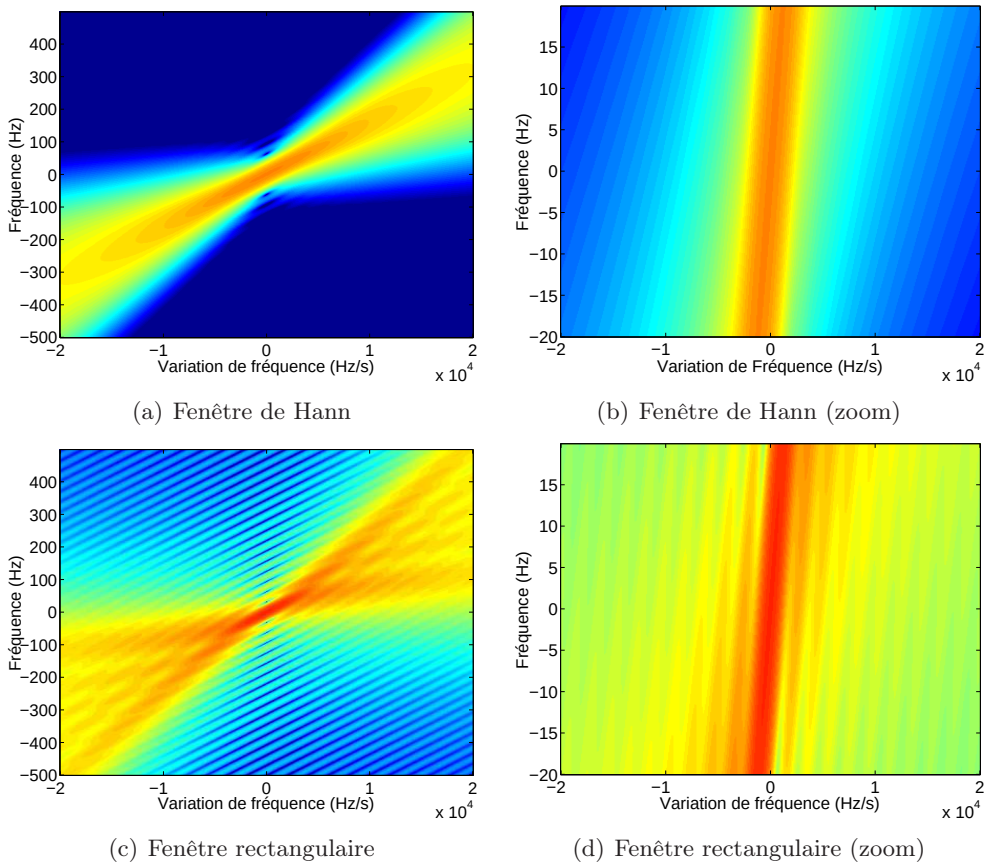
$$\alpha_i = \arg(X(t_{m_i}, \omega_{k_i}; h)) - \arg(\Gamma(\mu, \Delta\beta_i, \gamma; h)) \ [2\pi] \tag{6.2.2}$$

$$\Gamma(\mu, \Delta\beta_i, \gamma; h) \triangleq \sum_{n=-(N-1)/2}^{(N-1)/2} h(\tau_n)\, \mathrm{e}^{\mu\tau_n}\, \mathrm{e}^{j(\Delta\beta_i \tau_n + \frac{\gamma}{2}{\tau_n}^2)} \tag{6.2.3}$$

où $\Delta\beta_i$ est la différence entre la fréquence de la sinusoïde à l'instant $t_{m_i}$ et la fréquence du bin $\omega_{k_i}$. Le terme $\Gamma$ a la forme d'une transformation polynomiale. Lorsque $\mu = 0$, on retrouve la transformation en phase quadratique, utilisée par exemple dans [Ikram et al., 1996], [Xia, 2000]. De même que pour le modèle **M01**, si $T$ est suffisamment grand, un facteur de déroulement de phase $n$ sera nécessaire, quand on considérera une différence de phase $\alpha_2 - {\alpha_1}$[8].

La relation (6.2.2) permet d'estimer la phase de façon non biaisée, à condition de connaître les paramètres $\beta$ et $\gamma$ [Betser et al., 2006b]. On appelle cette estimation de la phase, phase de Fourier corrigée ou phase corrigée. A la section 6.4, nous reprendrons plus en détail cette idée, en la généralisant à un modèle d'ordre quelconque.

Dans cette section nous allons d'abord présenter deux briques utilisées par les algorithmes d'estimation développés par la suite, le suivi du maximum, à la section 6.2.1, et le déroulement de phase, à la section 6.2.2. Ensuite, nous présenterons trois méthodes d'estimation de fréquence basées sur le vocodeur de phase et dédiées aux modèles **M02** et **M12** : l'estimation par suivi de maximum, à la section 6.2.3, le vocodeur à phases corrigées, à la section 6.2.4 et le vocodeur réassigné 6.2.5. Enfin nous finissons par un développement théorique de la variance des estimateurs, à la section 6.2.6, avant une comparaison expérimentale avec le réassignement fréquentiel et la QIFFT, à la section 6.2.7.




(a) Fenêtre de Hann

(b) Fenêtre de Hann (zoom)

(c) Fenêtre rectangulaire

(d) Fenêtre rectangulaire (zoom)

FIG. 6.7: Transformation en phase quadratique discrète pour deux fenêtres d'analyse

### 6.2.1 Suivi du maximum (MB)

La Figure 6.7 donne un exemple de transformation en phase quadratique pour un signal de type **M02**. On voit sur les Figures 6.7(b) et 6.7(d) que lorsque l'on est proche du maximum, c'est à dire dans l'intervalle $[-R, R]$ où $R$ est toujours la demi-précision de Fourier, l'amplitude de la transformée est quasiment constante pour une variation de fréquence donnée. Le terme d'erreur de l'équation (6.2.2) est donc dominé par l'influence de la variation de fréquence lorsque l'on est proche du maximum. Donc si l'on suppose que cette variation est identique sur deux trames consécutives, et si on considère la différence de phase entre deux bins proches du maximum de leur trame respective, les termes d'erreur devraient se compenser.

Une façon plus mathématique de justifier l'importance du suivi de bin lorsque la fréquence varie est de revenir à l'équation (5.1.1) :

$$X(t, \omega; h) = \mathrm{e}^{\mu\delta_t + j(\beta\delta_t + \gamma\frac{\delta_t^2}{2})}\, X(t_0, \omega_0 + \delta_\omega - \delta_t\gamma; h)$$

[7] Voir la section 6.4 pour plus de détails.

[8] Voir la section 4.1.2.1 sur le vocodeur de phase à long terme.

On rappelle que $\delta_\omega = \omega - \omega_0$ et $\delta_t = t - t_0$. On peut choisir indépendamment le point de référence $(t_0, \omega_0)$ et $\delta_\omega$. En particulier on peut ajuster $\delta_\omega$ de façon à compenser exactement $\delta_t\gamma$. Chaque TFCT s'exprime alors en fonction de la TFCT au point de référence.

Par exemple, considérons deux points temps-fréquence $(t_1, \omega_1)$ et $(t_2, \omega_2)$. On choisit comme point de référence le milieu de ces points $(\frac{t_2+t_1}{2}, \frac{\omega_1+\omega_2}{2})$, de façon à avoir $\delta_{\omega_1} = -\delta_{\omega_2}$ et $\delta_{t_1} = -\delta_{t_2}$. Si les fréquences $\omega_1$ et $\omega_2$ sont telles que $\delta_\omega$ compense $\delta_t\gamma$, alors la différence discrète des transformées de Fourier nous donne une estimation directe de la fréquence en $t_M$ :

$$\beta_M = \frac{\arg(X(t_2, \omega_2; h)) - \arg(X(t_1, \omega_1; h))[2\pi]}{2\delta_t} \tag{6.2.4}$$

Pour cela il faut connaître $\delta_t\gamma$, or le suivi des maximums du spectre nous donne justement une estimation de $\delta_t\gamma$, pour un modèle d'ordre **M12**. En effet, si on suppose que $\omega_i = \text{argmax}_\omega |X_i|$, alors on aura :

$$\omega_1 = \text{argmax}_\omega |X(t_0, \omega - \gamma\delta_t; h)| \tag{6.2.5}$$
$$\omega_2 = \text{argmax}_\omega |X(t_0, \omega + \gamma\delta_t; h)| \tag{6.2.6}$$

On a donc la relation suivante entre $\omega_1$ et $\omega_2$ : $\omega_2 = \omega_1 + 2\gamma\delta_t$, et on trouve bien que :

$$\gamma\delta_t = \frac{\omega_2 - \omega_1}{2} = \delta_\omega \tag{6.2.7}$$

Même si le maximum du spectre ne donne plus directement une estimation de la fréquence de la sinusoïde, le suivi des maximums donne une estimation de $\gamma\delta_t$. Si la précision fréquentielle de la TFCT est suffisamment élevée pour avoir une bonne approximation de $\gamma\delta_t$, et grâce à la différence des phases de Fourier de l'équation (6.2.4), on en déduit la fréquence en $t_M$. Le dernier élément manquant pour pouvoir estimer $\beta_M$ est l'estimation du déroulement de phase, car l'argument est défini modulo $2\pi$. Nous discuterons de ce problème dans la section suivante.

### Algorithme de suivi de maximum

Une méthode simple de suivi de bin peut être adaptée de [McAulay and Quatieri, 1986]. Le suivi est réalisé localement sur trois trames consécutives, et la variation de fréquence maximum tolérée dans [McAulay and Quatieri, 1986], correspond ici à une borne sur la variation de bin, $\Delta k = \frac{NT\gamma_m}{4\pi F}$, pour un intervalle temporel de $T/2$ entre deux trames consécutives. La procédure peut se résumer ainsi :

1. Calculer les FFTs zéro-phase $X_1$, $X_2$, $X_M$ pour les temps $t_{m_1}$,$t_{m_2} = t_{m_1} + T$ et $t_M = t_{m_1} + T/2$
2. Calculer le bin maximum $\hat{k}_M$ pour le temps $t_M$
3. Calculer $\hat{k}_1 = \arg\max_{k\in\{k_M \pm \Delta k\}} |X(t_{m_1}, \omega_k; h)|$
4. Calculer $\hat{k}_2 = \arg\max_{k\in\{k_M \pm \Delta k\}} |X(t_{m_2}, \omega_k; h)|$

### 6.2.2 Estimation du déroulement de phase

Nous avons vu que pour le modèle **M01**, le déroulement de phase était estimé grâce à l'équation (4.1.12) que nous rappelons ici :

$$\hat{n} = \text{round}\big((\Omega T - \Delta X)/(2\pi)\big)$$

où $\Omega$ est une fréquence de référence suffisamment proche de la vraie fréquence pour que le déroulement de phase soit le même[9]. Dans le cas d'une fréquence constante, on va choisir comme référence le bin maximum $\omega_k$. Dans le cas où la fréquence varie le bin maximum ne sera plus le même d'une trame à la suivante et l'on va choisir cette fois la moyenne des bins maximum $\omega_M$ comme référence. C'est cette référence qui est utilisée dans [McAulay and Quatieri, 1986] dans le cadre de la synthèse de phase cubique.

Comme dans la section 4.1.2.1, ce choix impose une limite théorique sur l'écart temporel maximum toléré entre les trames utilisées, que nous allons maintenant étudier. Pour deux trames successives séparées par un temps $T$, on note $\Delta X_m$ la différence de phase des bins maximum $k_1$ et $k_2$ :

$$\Delta X_m = \arg(X(t_2, \omega_{k_2}; h)) - \arg(X(t_1, \omega_{k_1}; h)) \tag{6.2.8}$$

En choisissant comme fréquence de référence $\omega_M$, le facteur de déroulement de phase $n$ va donc être estimé ainsi :

$$\hat{n} = \text{round}\big((\omega_M T - \Delta X_m)/(2\pi)\big) \tag{6.2.9}$$

Nous avons vu au début de la section que, pour le modèle **M12**, la phase de la sinusoïde à un instant $t_i$ pouvait s'exprimer comme la somme de la phase du bin maximum de la TF et d'un terme correctif (équation (6.2.2)). On en déduit la relation suivante :

$$T\beta_M = \Delta X_m + \arg(\Gamma_1 \bar{\Gamma}_2) + 2\pi n \tag{6.2.10}$$

où $\Gamma_i = \Gamma(\mu, \Delta\beta_i, \gamma; h)$ et $n$ est le facteur de déroulement de phase $n \in \mathcal{Z}$. Dans le cas où les bins maximum correspondent exactement au maximum de la TF, on aura $\Gamma_1 = \Gamma_2$ et on retombe sur l'équation (6.2.4). En notant $\Delta\beta_M = \beta_M - \omega_M$, on va maintenant faire apparaître la fréquence de référence $\omega_M$ :

$$n = \frac{1}{2\pi}\big[\omega_M T - \Delta X_m + \Delta\beta_M T - \arg(\Gamma_1 \bar{\Gamma}_2)\big] \tag{6.2.11}$$

En comparant cette dernière équation avec l'estimateur (6.2.9), on a montré dans l'article [Betser et al., 2007], reproduit dans l'annexe G, qu'une condition suffisante pour qu'il y ait identité entre la vraie valeur de $n$ et la valeur estimée $\hat{n}$ est la suivante :

$$H \leq N\left(1 - \frac{\Gamma_m(R, \mu_m, \gamma_m; h)}{\pi}\right) \tag{6.2.12}$$

[9] Voir l'estimation du déroulement de phase dans le cas du vocodeur de phase à long terme, à la section 4.1.2.1.

TAB. 6.3: Valeur maximale des pas d'avancement, correspondant à l'équation (6.2.12), mesurée en nombre d'échantillons pour le modèle **M02** ($N = 512$, $F = 16000$)

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $\gamma_m = 1000$ | 508 | 505 | 509 | 506 |
| $\gamma_m = 8000$ | 495 | 492 | 500 | 494 |

où $\Gamma_m$ est la valeur maximale du terme correctif pour les intervalles de paramètres considérés :

$$\Gamma_m(\Delta_m, \mu_m, \gamma_m; h) = \max_{|\Delta\beta_i| \leq \Delta_m, |\mu| \leq \mu_m, |\gamma| \leq \gamma_m} |\arg(\Gamma_1 \bar{\Gamma}_2)| \tag{6.2.13}$$

Dans le cas des modèles **M01** et **M11**, c'est à dire lorsque la fréquence reste constante, on a $\arg(\Gamma_1\bar{\Gamma}_2) = 0$, et on retrouve la condition de déroulement de phase classique $H \leq N$. Dans le cas des modèles **M02** et **M12** on a également la condition $H \leq N$ si les bins maximums choisis sont les maximums exacts de la TF. En effet, on a fait remarquer au paragraphe précédent que dans ce cas $\Gamma_1 = \Gamma_2$, ce qui signifie que l'on aura également $\arg(\Gamma_1\bar{\Gamma}_2) = 0$.

Si les bins maximums s'écartent du maximum exact de la TF, alors l'écart temporel maximum toléré diminue selon l'équation (6.2.12). La valeur $\Gamma_m$ est difficile à calculer analytiquement dans le cas général, mais pour un jeu de paramètres donné, une évaluation numérique peut facilement être réalisée. Le tableau 6.3 donne les valeurs maximales tolérées dans le cas du modèle **M02**, pour différent types de fenêtres utilisées. On peut voir que l'écart maximum théorique entre les deux trames utilisées pour l'estimation diminue lentement lorsque la variation de fréquence augmente. Pour les applications usuelles qui utilisent des écarts beaucoup plus petits que ces limites, on peut donc en conclure que la variation de fréquence n'aura pas d'impact sur l'estimation du facteur de déroulement de phase $n$.

### 6.2.3 Le vocodeur de phase par suivi des maximum (MB)

En se basant sur le suivi du maximum décrit à la section 6.2.1 et le déroulement de phase décrit à la section 6.2.2, on peut directement déduire une méthode basée sur la formule (6.2.4) pour calculer la fréquence. La procédure, notée MB dans les expériences, peut se résumer ainsi :

1. Suivi des bins maximums : les FFTs, $\hat{k}_M$, $\hat{k}_1$ et $\hat{k}_2$ sont calculés
2. Estimation du déroulement de phase $\hat{n}$
3. Estimation de la fréquence pour le temps $\hat{t} = t_M$ :
   $\hat{\beta} = \frac{\arg(X(t_2,\omega_{k_2};h)\bar{X}(t_1,\omega_{k_1};h)) + 2\pi\hat{n}}{T}$

### 6.2.4 Le Vocodeur à phase corrigées (PCV)

Nous verrons dans la section expérimentale 6.2.7 que pour avoir une estimation suffisamment précise de la variation de fréquence, la méthode des bins maximum nécessite un très fort coefficient de zéro-padding. Cela rend la méthode assez complexe,

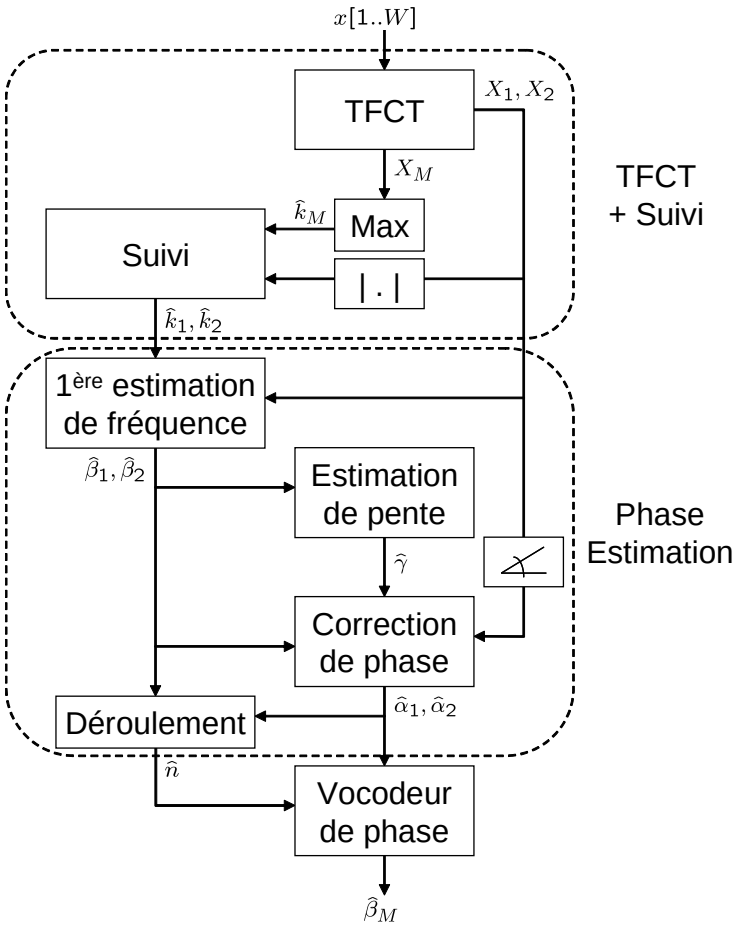



FIG. 6.8: Schéma de fonctionnement du Vocodeur à Phases Corrigées (PCV)

et il va être donc intéressant de considérer des méthodes approchées, nécessitant peu ou pas de padding. C'est ce nous allons étudier dans ce paragraphe et le suivant.

Comme nous l'avons mentionné précédemment, la TF n'est plus directement un estimateur de la phase pour les signaux dont la fréquence varie. On a vu au début de la section qu'une estimation non biaisée de la phase était possible à condition de posséder une estimation des paramètres d'ordre supérieur. Une première idée pour améliorer l'estimation consiste à corriger les phases de Fourier en utilisant la formule (6.2.2), comme dans l'article [Betser et al., 2006b]. La méthode d'estimation de fréquence utilisant la correction de phase a été développée en détail dans les articles [Betser et al., 2008] et [Betser et al., 2007].

Le modèle considéré ici est le modèle **M02**. Le schéma d'analyse du vocodeur à phase corrigées est rappelé sur la Figure 6.8. La méthode est en deux étapes :

1. Estimation des phases corrigées (modulo $2\pi$) $\hat{\alpha}_1$ et $\hat{\alpha}_2$, et du facteur de déroulement de phase $\hat{n}$.

2. Estimation de $\beta_M$ en utilisant la formule du vocodeur de phase [10]

$$\hat{\beta}_M = \frac{\text{mod}(\hat{\alpha}_2) - \text{mod}(\hat{\alpha}_1) + 2\pi\hat{n}}{T} \tag{6.2.14}$$

La première étape commence par une estimation des fréquences $\beta_1$ et $\beta_2$ réalisée grâce à une méthode de type interpolation de spectre (cf. Figure 6.8). Ce type de méthode a été développée pour un modèle **M01** et sera donc biaisée pour le modèle **M02**. Une erreur sera donc introduite dans l'équation (6.2.2) pour l'estimation des phases corrigées. Malgré cela l'article [Betser et al., 2006b] a montré que ce schéma permettait d'améliorer grandement la précision d'estimation des phases, par rapport à l'utilisation directe de la méthode des bins maximum. Les deux autres paramètres nécessaires à la première étape, la variation de fréquence $\gamma$ et le déroulement de phase $n$ peuvent être déduits des fréquences $\beta_1$ et $\beta_2$ grâce à ces formules :

$$\hat{\gamma} = \frac{\hat{\beta}_2 - \hat{\beta}_1}{T}$$
$$\hat{n} = \text{round}\left(\frac{1}{2\pi}\left(\text{mod}(\hat{\alpha}_1) - \text{mod}(\hat{\alpha}_2) + \frac{\hat{\beta}_1 + \hat{\beta}_2}{2}T\right)\right)$$

Une estimation de $\gamma$ aurait pu être réalisée en utilisant les méthodes décrites dans la section 4.2.2. Il a cependant été constaté, que l'estimateur ci-dessus donnait de meilleurs résultats [Betser et al., 2006b].

En résumé la procédure d'estimation est donc la suivante :

1. Suivi des bins maximum : les FFTs et $\hat{k}_M$, $\hat{k}_1$ et $\hat{k}_2$ sont calculés (cf. section 6.2.1)
2. Estimation des phases corrigées :
   - Calculer une estimation de $\hat{\beta}_1$ (resp. $\hat{\beta}_2$) de la fréquence en $t_{m_1}$ (resp. $t_{m_2}$).
   - Calculer une estimation de la variation de fréquence : $\hat{\gamma} = (\hat{\beta}_2 - \hat{\beta}_1)/T$
   - Calculer les phases corrigées $\hat{\alpha}_1$, $\hat{\alpha}_2$ : $\hat{\alpha}_i = \arg(X(t_{m_i}, \omega_{k_i}; h)) - \arg(\Gamma(\omega_{k_i} - \hat{\beta}_i, \hat{\gamma}; h)$.
   - Calculer le facteur de déroulement de phase : $\hat{n} = \text{round}\big(\frac{\text{mod}(\hat{\alpha}_1) - \text{mod}(\hat{\alpha}_2) + .5(\hat{\beta}_1 + \hat{\beta}_2)T}{2\pi}\big)$
3. Estimation de la fréquence en $\hat{t} = t_M$ : $\hat{\beta} = \frac{\text{mod}(\hat{\alpha}_2) - \text{mod}(\hat{\alpha}_1) + 2\pi\hat{n}}{T}$

Pour conclure, nous remarquerons que bien qu'une généralisation de la méthode soit possible pour un modèle **M12**, nous ne disposons pas d'estimateur de variation d'amplitude $\mu$ qui nous permettrait d'utiliser la formule d'estimation de phase du modèle **M12**. Nous nous en tiendrons donc dans les expériences à la méthode appliquée au modèle **M02**, comme dans les articles [Betser et al., 2008] et [Betser et al., 2007]. Notons également que ce schéma d'analyse peut être utilisé pour améliorer la précision des fréquences après un suivi sinusoïdal comme ceux présentés dans [McAulay and Quatieri, 1986; Serra, 1997].

[10] mod() est la fonction modulo $2\pi$.

### 6.2.5 Le vocodeur réassigné (RV)

Dans cette section nous décrivons une approche différente pour améliorer l'estimation de la méthode des bins maximum. Elle a été décrite, comme la méthode précédente, dans les articles [Betser et al., 2008] et [Betser et al., 2007]. Elle est basée sur une approximation de Taylor des termes d'erreur dans l'équation (6.2.2) pour le modèle **M12**.

Le point de départ de la méthode est l'équation des différences de phase (6.2.10), que l'on rappelle ici :

$$T\beta_M = \Delta X_m + \arg(\Gamma_1 \bar{\Gamma}_2) + 2\pi n$$

où $\Gamma_i = \Gamma(\mu, \Delta\beta_i, \gamma; h)$. La première étape est d'exprimer $\arg(\Gamma_1\bar{\Gamma}_2)$ en fonction de la fréquence à estimer, $\beta_M$. Pour cela, nous allons décomposer $\Delta\beta_1$ et $\Delta\beta_2$ en deux termes bornés $B = \beta_M - \omega_M$ et $G = \delta_\omega - \gamma\frac{T}{2}$. Nous allons ensuite effectuer un développement de Taylor en $G = 0$.

On rappelle que $\omega_M$ est la fréquence moyenne des bins, $\delta_\omega$ est la demi-variation des bins (6.2.7). On peut donc en déduire que les $\Delta\beta_i$ peuvent s'écrire ainsi[11] :

$$\Delta\beta_1 = B + G, \qquad \Delta\beta_2 = B - G \tag{6.2.15}$$

Dans le cas du modèle **M02**, $\omega_{k_1}$ et $\omega_{k_2}$ sont les bins les plus proches de $\beta_1$ et de $\beta_2$, et donc dans ce cas on aura[12] : $|B| < R$. Dans le cas du modèle **M12** cette relation ne sera plus vraie car la variation d'amplitude va éloigner le bin maximum $\omega_{k_i}$ de la fréquence $\beta_i$. Par contre pour les deux modèles, la relation $|G| < R$ sera vérifiée. Le fait que $G$ soit une quantité bornée, nous assure d'avoir une borne sur l'erreur commise en faisant une approximation de Taylor en $G = 0$.

Nous rappelons ici le développement de Taylor d'ordre 1 en $G = 0$ de l'équation (6.2.10), sans donner les calculs intermédiaires. On se référera à [Betser et al., 2007], reproduit dans l'annexe G, pour plus de détails :

$$\begin{aligned}\hat{\beta} &= \frac{\Delta X_m + 2\pi n}{T} + 2\frac{\delta_\omega}{T}\Re\Big(\frac{X(t_M, \omega_M; \tau h)}{X(t_M, \omega_M; h)}\Big) \\ \hat{t} &= t_M + \Re\Big(\frac{X(t_M, \omega_M; \tau h)}{X(t_M, \omega_M; h)}\Big)\end{aligned} \tag{6.2.16}$$

On remarque que $\hat{t}$ est le temps réassigné. $\hat{\beta}$ est composé de deux termes, un terme principal lié à la méthode d'estimation classique du phase vocodeur, et un terme correctif lié au temps réassigné. Notons que l'équation (6.2.16) indique que le temps d'estimation de l'estimateur classique du phase vocodeur (*i. e.* avec $\delta_\omega = 0$) est en fait le temps réassigné d'une trame centrée en $t_M$ et pour la fréquence $\omega_M = \omega_{k_1}$. Cet estimateur peut être également vu comme une forme approchée du réassignement. L'avantage est qu'ici on utilise seulement deux fenêtres différentes $h$ et $\tau h$ au lieu de trois.

[11] Voir l'article [Betser et al., 2007] reproduit dans l'annexe G pour plus de détails.

[12] On rappelle que $R$ est la demi précision de Fourier.

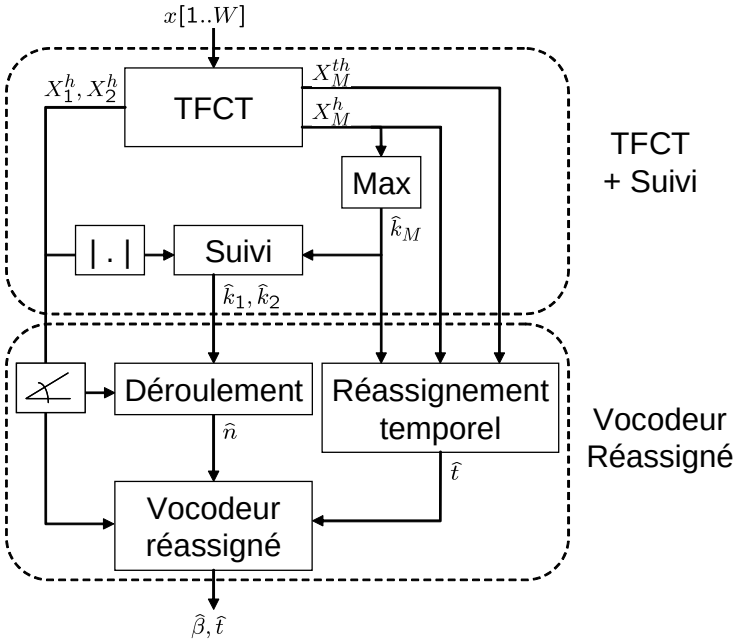



FIG. 6.9: Schéma de fonctionnement du Vocodeur Réassigné (RV). $X_i^h$ désigne la FFT calculée avec la fenêtre $h$ au temps $t_{m_i}$.

L'estimation du facteur de déroulement de phase a été discuté dans la section 6.2.2. La méthode du vocodeur réassigné a été implantée en utilisant un schéma similaire à celui du vocodeur à phases corrigées, avec trois trames successives localisées aux temps $t_{m_1}$, $t_M$ et $t_{m_2}$ (voir Figure 6.9).

Lorsque les bins maximum $k_1$ et $k_2$ sont différents, la fréquence $\omega_M = (k_1 + k_2)F/(2N)$ pour la trame centrée en $t_M$ peut ne pas correspondre à un bin de la TFCT. Afin de pouvoir calculer le temps réassigné en utilisant les TFCTs (cf. équation (6.2.16)), $k_1$ ou $k_2$ doit être décalé sur un bin adjacent. Le nouveau bin choisi est celui qui a l'amplitude la plus haute. On résume ci-dessous la méthode complète :

1. Calcul des TFCTs et suivi de bins maximums : calcul des FFTs pour le temps $t_M$, en utilisant les fenêtres $h$ et $\tau h$. calcul des FFTs pour les temps $t_{m_1} = t_M - T/2$ et $t_{m_2} = t_M + T/2$, avec la fenêtre $h$. Calcul de $k_M$, $k_1$ et $k_2$.

2. Calcul du facteur de déroulement de phase en deux étapes :
   $\hat{n}_1 = \text{round}\big(\frac{\arg(X_1) - \arg(X_M) + (\omega_{k_1} + \omega_M)T/4}{2\pi}\big)$
   $\hat{n}_2 = \text{round}\big(\frac{\arg(X_M) - \arg(X_2) + (\omega_M + \omega_{k_2})T/4}{2\pi}\big)$
   $\hat{n} = \hat{n}_1 + \hat{n}_2$

3. Calcul de la fréquence estimée :
   $\hat{\beta} = \frac{\Delta X_m + 2\pi\hat{n}}{T} + 2\frac{\delta_\omega}{T}\Re\big(\frac{X(t_M, \omega_M; \tau h)}{X(t_M, \omega_M; h)}\big)$
   L'estimation est réalisée pour le temps :
   $\hat{t} = t_M + \Re\big(\frac{X(t_M, \omega_M; \tau h)}{X(t_M, \omega_M; h)}\big)$

Afin de réduire les possibilités d'erreur sur le déroulement de phase, on a choisi de calculer le facteur de déroulement $n$ en deux étapes (cf. section 6.2.2).

### 6.2.6 Propriétés statistiques des estimateurs

Dans cette section nous allons étudier la variance des trois estimateurs qui viennent d'être présentés, la méthode des bins maximum (MB), le vocodeur à phases corrigées (PCV) et le vocodeur réassigné (RV). Toutes ces méthodes vont présenter des comportements similaires. La méthode pour dériver la variance est celle présentée dans la section 5.4.

On suppose que le signal $x$ est perturbé par un bruit blanc Gaussien centré $s = x + n$ et on note $S_i = X_i + N_i$ la TF au point temps-fréquence $(t_{m_i}, \omega_{k_i})$. D'après l'équation (5.4.7), si on choisit $f(S_1, S_2) = S_1\bar{S}_2$ et $g(X) = \arg(X)$, on aura :

$$\arg(S_2\bar{S}_1) \approx \arg(X_2\bar{X}_1) + \Im(Z)$$

où $\Im(Z) = \Im(N_2/X_2) - \Im(N_1/X_1)$.

En remplaçant $\arg(X_2\bar{X}_1)$ par l'équation (6.2.10), on en déduit que :

$$\arg(S_2\bar{S}_1) \approx T\beta_M + \arg(\Gamma_2\bar{\Gamma}_1) + 2\pi n + \Im(Z)$$

Toutes les méthodes vont calculer des estimations du déroulement de phase $n$, mais en considérant que les sinusoïdes sont suffisamment résolues, aucune erreur ne sera faite dans l'estimation de $n$. Les méthodes PCV et RV, vont utiliser des estimations de $\arg(\Gamma_2\bar{\Gamma}_1)$, mais on peut montrer, en utilisant toujours l'hypothèse de sinusoïdes bien résolues par la TF, que l'erreur stochastique provenant de l'estimation de $\arg(\Gamma_2\bar{\Gamma}_1)$ est négligeable par rapport à $\Im(Z)$[13]. Dans le cas de la méthode des bins maximum, on va considérer que la précision de la TFCT est suffisamment élevée pour ne pas avoir à estimer $\arg(\Gamma_2\bar{\Gamma}_1)$.

Pour les trois méthodes, on peut donc écrire :

$$\hat{\beta} \approx \beta + \epsilon + \frac{\Im(Z)}{T} \tag{6.2.17}$$

où $\beta$ est la fréquence à estimée, qui est $\beta_M$ pour les méthodes PCV et MB et $\beta_M + \gamma\Re(\frac{X(t_M,\omega_M;\tau h)}{X(t_M,\omega_M;h)})$ pour la méthode RV. $\epsilon$ est le biais déterministe, différent pour chaque méthode. La variance est donc la même pour les trois méthodes et sera égale à :

$$\mathrm{var}(\hat{\beta}) = \frac{E(\Im(Z)^2)}{T^2} \tag{6.2.18}$$

La démonstration complète de la dérivation de la variance est donnée dans l'article [Betser et al., 2007], reproduit dans l'annexe G. On ne rappelle ici que le résultat final :

$$\mathrm{var}(\hat{\beta}) = \frac{\sinh(\mu\tau_W)}{\mu\tau_W} \frac{\Big[\cosh(\Delta\lambda)H_0 - \cos(\Delta\Phi)H_1\Big]}{\eta T^2|\Gamma_2\Gamma_1|} \tag{6.2.19}$$

[13] Voir l'article [Betser et al., 2007], reproduit dans l'annexe G.

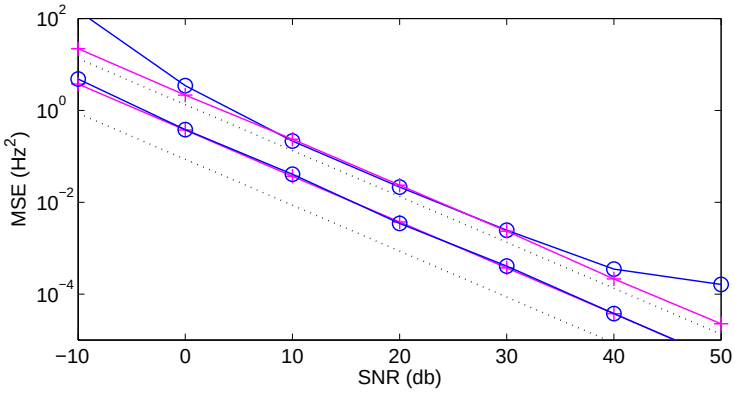



FIG. 6.10: Comparaison des variances théoriques (marqueur '+') à la CRB (en pointillés) et à la MSE de la méthode RV (marqueur 'o'). Les courbes du haut correspondent au modèle **M12** avec $\mu \in [0, 100]$ et $\gamma \in [0, 8000]$, et les courbes du bas au modèle **M02**.

où $\eta$ est le rapport signal à bruit, $\Delta\lambda$ et $\Delta\Phi$ sont respectivement la différence de log-amplitude et de phase entre $X_2$ et $X_1$. Enfin $H_0$ et $H_1$ sont deux facteurs qui ne dépendent que de la fenêtre $h$.

$$
\begin{aligned}
\eta &\triangleq \frac{e^{\lambda_1+\lambda_2}}{\sigma^2}\frac{\sinh(\mu\tau_W)}{\mu\tau_W} \\
\Delta\lambda &\triangleq T\mu + \log(|\Gamma_2|) - \log(|\Gamma_1|) \\
\Delta\Phi &\triangleq T(\beta_M - \omega_M) + \arg(\Gamma_2\bar{\Gamma}_1) \\
H_0 &\triangleq \sum_{i=-N/2}^{N/2} h_i^2 \\
H_1 &\triangleq \sum_{i=-(N-H)/2}^{(N-H)/2} h_{i+\frac{H}{2}}\, h_{i-\frac{H}{2}} \cos(\tau_i(\omega_{k_1} - \omega_{k_2}))
\end{aligned}
$$

De ces équations, la variance pour les modèles **M11** et **M02** peut être déduite directement.

Dans le cas du modèle classique **M01**, $\mu = 0$, $\gamma = 0$, $\Delta\lambda = 0$ et $\omega_{k_1} = \omega_{k_2} = \omega$, $\beta_1 = \beta_2 = \beta$. L'équation (6.2.19) se simplifie en :

$$
\mathrm{var}(\hat{\beta}) = \frac{\Big[H_0 - \cos(\Delta\Phi)H_1\Big]}{\eta T^2 |\Gamma|^2} \tag{6.2.20}
$$

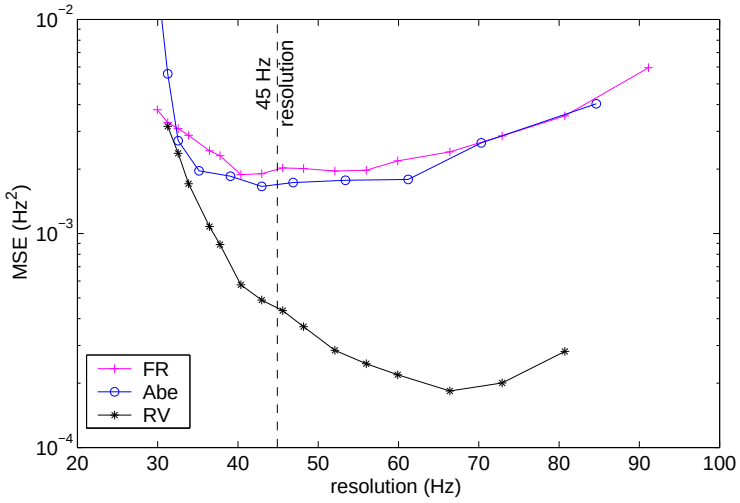



FIG. 6.11: Performances des algorithmes pour un modèle **M02** en fonction de la résolution fréquentielle, pour un SNR=30 dB et W=48 ms.

où

$$\eta = \frac{\mathrm{e}^{2\lambda}}{\sigma^2}, \qquad H_1 = \sum_{i=-(N-H)/2}^{(N-H)/2} h_{i+\frac{H}{2}}\, h_{i-\frac{H}{2}},$$

$$\Delta\Phi = T(\beta_M - \omega_M), \qquad \Gamma = \sum_{i=-N/2}^{N/2} h_i$$

Cette dernière équation est la même que celle trouvée dans [Abeysekera and Padhi, 2006].

Deux exemples sont donnés sur la Figure 6.10, un pour le modèle **M02** (courbe du bas) et un pour le modèle **M12** (courbe du haut). Dans les zones où l'erreur stochastique domine, la variance théorique correspond exactement à l'erreur expérimentale (MSE). Pour le modèle **M12** (courbe du haut), des biais apparaissent pour les hauts SNRs et les bas SNRs. Dans le premier cas, le biais est dû à l'erreur déterministe de l'estimateur, et dans le deuxième cas, il est dû au schéma de suivi de maximum.

### 6.2.7 Evaluation

Le protocole expérimental a été décrit dans la section 3.4. Les algorithmes des bins maximum, du vocodeur à phases corrigées et du vocodeur réassigné vont être comparés à l'algorithme de la QIFFT et au réassignement classique. Dans la section décrivant cette dernière méthode, nous avions dit que la résolution de la fenêtre Gaussienne pouvait être ajustée, et qu'elle serait fixée à 45 Hz pour les expériences[14]. Nous allons expliquer ici pourquoi.

[14]La section 3.3.3 décrit le critère utilisé pour mesurer la résolution.

En fait, la performance de tous les algorithmes dépend de la résolution de la TFCT, comme le montre la Figure 6.11. Pour qu'une comparaison soit valable, il faut donc que les méthodes comparées aient approximativement la même résolution. Or on voit sur la Figure 6.11 que la résolution de 45 Hz correspond à la zone où toutes les méthodes à comparer donnent de bons résultats. 45 Hz correspond également à la résolution de la fenêtre de Hann de 32 ms, qui est un compromis classique entre la résolution temporelle et fréquentielle pour des signaux variables comme la parole.

Pour une comparaison plus juste avec les méthodes MB, PCV et RV basées sur le vocodeur de phase , le réassignement et la QIFFT vont également être appliquées sur trois trames successives (RF et QIFFT3). Les bins maximums sont choisis en faisant un suivi de maximum local comme dans la section 6.2.1. L'estimation de fréquence finale sera la moyenne des estimations sur les trois fenêtres. Comme la QIFFT est instable lorsque l'on utilise des fenêtres courtes pour analyser des signaux avec des fortes variations d'amplitude et de fréquence (section 4.3.1), une méthode d'estimation utilisant seulement une longue fenêtre est aussi présentée (QIFFT1).

L'augmentation du facteur de zéro-padding réduit le biais des méthodes et sera tout particulièrement bénéfique aux méthodes PCV et QIFFT3, qui sont fortement biaisée dans le cas **M12**. Quand ce n'est pas précisé, toutes les méthodes sont utilisées avec un facteur de padding de 3. Pour les modèles **M01** et **M11**, les méthodes PCV, RV et MB sont toutes équivalentes à la méthode du vocodeur de phase à long terme (LV) décrite dans la section 4.1.2.1. La comparaison pour ces modèles a donc déjà été faite dans la section 4.4.

### Modèle M02

Cette section compare les estimateurs pour le modèle **M02**. Les expériences sont réalisées pour une fenêtre d'analyse avec $W = 767$ et pour une variation de fréquence comprise dans l'intervalle $[0, 8000]$ (Figure 6.12). La fréquence d'échantillonnage est toujours $F = 16000$.

La Figure 6.12(a) montre les améliorations successives réalisées par rapport au vocodeur de phase à long terme (LV). Dans cette expérience, le facteur de padding est de 1 pour toutes les méthodes. La courbe du haut correspond au vocodeur de phase à long terme standard, et on retrouve un biais très significatif pour les fortes variations de fréquence. Une première amélioration est obtenue lorsque le temps réassigné est utilisé pour l'estimation de la méthode LV. Cela correspond à une approximation du premier ordre avec $\Delta\omega = 0$, comme on l'a vu dans la section 6.2.5, équation (6.2.16). La méthode des bins maximums, décrite dans la section 6.2.1, améliore encore l'estimation. Elle est asymptotiquement non biaisée lorsque la précision de la TF tend vers zéro, mais pour un zéro padding peu élevé les performances restent faibles. De meilleurs performances sont obtenues avec la méthode PCV, particulièrement pour les hauts SNRs. Le biais restant est principalement dû au biais sur les estimations de variation de fréquence. Enfin le vocodeur réassigné qui combine l'utilisation des bins maximum et du temps réassigné améliore fortement les performances pour les hauts SNRs. L'estimateur est en fait quasiment sans biais pour les intervalles de variation de fréquence considérés.

La Figure 6.12(b) est une comparaison des méthodes RV et PCV, avec les méthodes de l'état-de-l'art. Le réassignement fréquentiel et la QIFFT donnent des résultats comparables pour des SNRs au delà de 0 dB. Pour la QIFFT, et des SNRs très faibles, il y a un phénomène important de propagation d'erreur des paramètres d'ordre supérieur (ici la variation de fréquence) vers les paramètres d'ordre inférieur [15], qui explique les mauvaises performances de la méthode dans cette région. Pour la région de SNR comprise entre $[0, 40]$, toutes les méthodes donnent des performances très proches avec un léger avantage pour la QIFFT. Enfin pour les hauts SNRs, on voit apparaître les biais des différentes méthodes. On rappelle que les biais sont dûs aux approximations respectives utilisées dans ces méthodes : l'approximation de Taylor pour la méthode RV, la propagation des erreurs de la première estimation pour la méthode PCV, l'utilisation d'une fenêtre Gaussienne tronquée pour la QIFFT et l'approximation d'une intégrale par une somme dans le cas du réassignement.

### Modèle M12

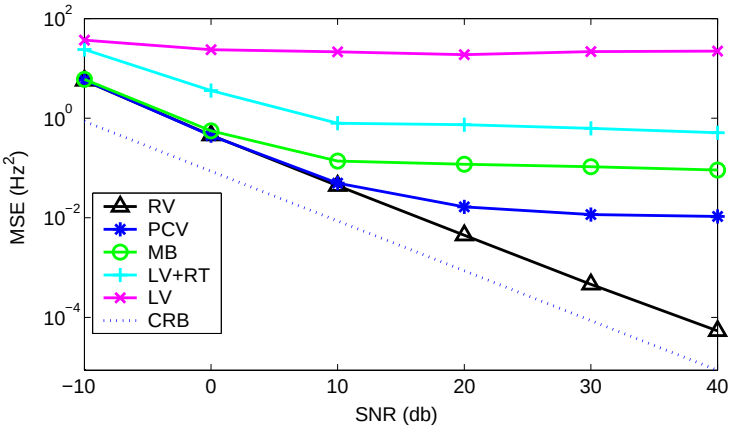
Dans cette section, les méthodes sont comparées pour le modèle **M12**. Sur la Figure 6.13(d), la variation de fréquence est dans l'intervalle $[0, 1000]$ et la variation d'amplitude dans l'intervalle $[0, 10]$, ce qui correspond à des sinusoïdes moyennement modulées. Pour toutes les autres figures, la variation de fréquence est comprise dans $[0, 8000]$ et la variation d'amplitude dans $[0, 100]$.

Les Figures 6.13(a) et 6.13(c) illustrent les performances obtenues pour différentes tailles de fenêtre d'analyse. Pour un faible recouvrement temporel (Figure 6.13(c)), le vocodeur réassigné utilise des bins identiques ($k_1 = k_2 = k_M$), car la variation de fréquence est très faible dans ce cas. Son biais disparaît et ses performances sont alors presque identiques à celles du réassignement : pour les deux méthodes le biais apparaît autour de 80 dB.

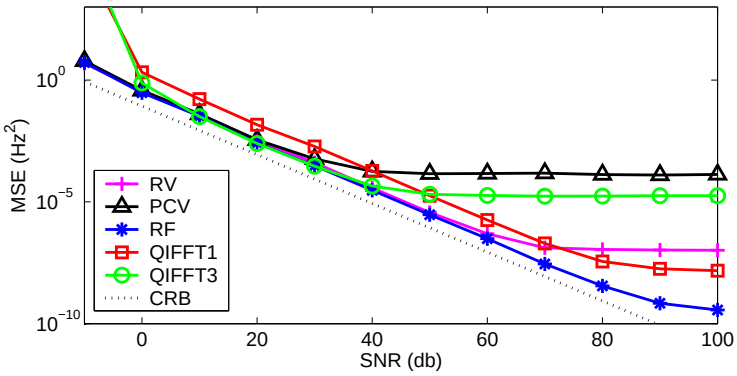
Pour de large pas d'avancement et de faibles SNRs, le schéma d'estimation avec trois trames successives devient légèrement instable. En fait, la difficulté de la tâche augmente lorsque l'on considère de large pas d'avancement, car la modulation d'amplitude décale la distribution d'énergie de la sinusoïde sur un des bords de la fenêtre $W$. L'autre bord a une énergie plus faible, surpassée par le bruit lorsque le SNR est suffisamment bas, ce qui pose des problèmes au schéma de suivi. Cet effet apparaît légèrement sur la Figure 6.13(a) pour les courbes RV et RF. Un pas d'avancement de 16 ms est un bon compromis, permettant de garder de bonnes performances avec une longueur raisonnable entre les trames. Le biais du réassignement est le même que dans les expériences précédentes, tandis que le biais de la méthode RV augmente légèrement. Pour la région de SNR comprise entre $[0, 40]$, les méthodes RV et RF sont toutes deux très proches du CRB avec un léger avantage à la méthode RV (Figure 6.13(b)).

Comme nous l'avions déjà mentionné dans la section 4.3.1, une fenêtre Gaussienne trop petite n'a plus une réponse parabolique, ce qui cause des problèmes d'estimation. Lorsqu'on augmente la taille de la fenêtre, les résultats de la QIFFT s'améliorent,

[15] Voir la section 4.3.1 qui décrit la méthode QIFFT.




(a) Comparaison des méthodes basées sur le vocodeur de phase : Vocodeur Long (LV), LV utilisant les bins maximums (MaxBin), LV utilisant le Temps Réassigné (LV+RT), RV et PCV



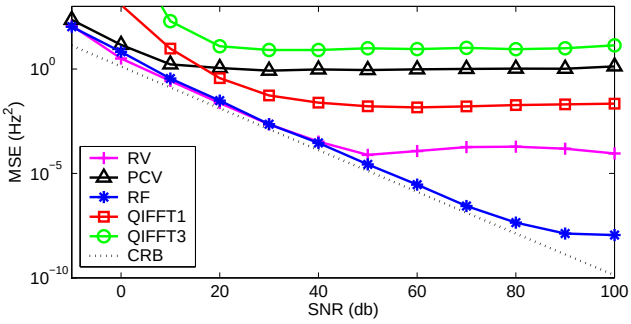

(b) Comparaison du vocodeur à phases corrigées (PCV), du Réassignement Fréquentiel (RF), de la QIFFT, et du Vocodeur Réassigné (RV)

FIG. 6.12: Comparaison des méthodes PCV, RV, réassignement et QIFFT pour le modèle **M02** ($\gamma \in [0, 8000]$).

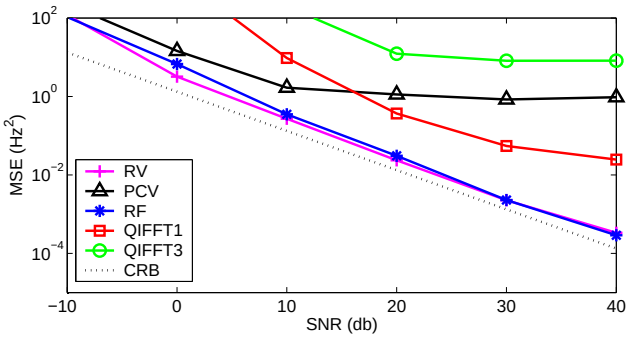
mais elle reste instable pour de faibles SRNs et moins performante que les autres estimateurs. La méthode d'estimation PCV n'est plus valide dans ce cas de figure car elle ne prend pas en compte les variations d'amplitude 6.2.4. La Figure 6.13(d) montre que toutes les méthodes se comportent bien pour des sinusoïdes faiblement modulées avec un léger avantage à l'algorithme de la QIFFT, dans la région SNR$\in [0, 40]$.

## 6.3 Estimation des paramètres d'ordre supérieur pour le modèle M12

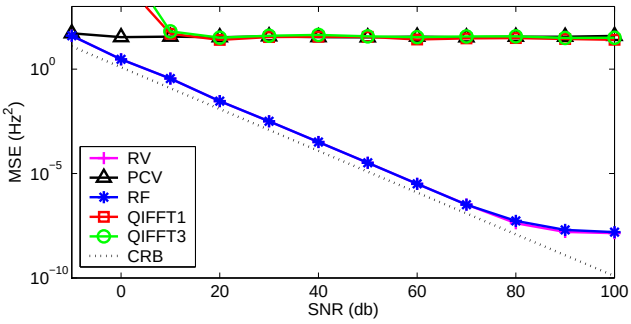
Dans cette section nous allons utiliser les résultats précédents pour estimer tous les paramètres du modèle **M12**. Deux méthodes sont présentées, une reposant sur le



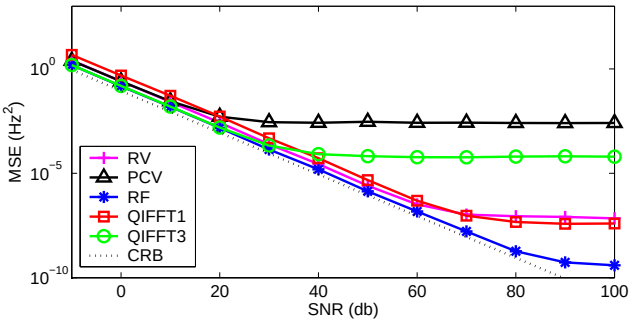

(a) $\gamma \in (0, 8000)$, $\mu \in (0, 100)$, $W = 767$, $T = 16$ms




(b) $\gamma \in (0, 8000)$, $\mu \in (0, 100)$, $W = 767$, $T = 16$ms (zoom)




(c) $\gamma \in (0, 8000)$, $\mu \in (0, 100)$, $W = 513$, $T = 0.1$ms




(d) $\gamma \in (0, 1000)$, $\mu \in (0, 10)$, $W = 767$, $T = 16$ms

FIG. 6.13: Comparaison des méthodes RV, réassignement et QIFFT pour le modèle **M12**

réassignement décrit dans la section 4.2.1.1, et une autre reposant sur le vocodeur réassigné décrit dans la section 6.2.5. Les deux méthodes vont présenter des similarités très fortes.

### 6.3.1 Estimation basée sur le réassignement

Dans la section 4.2.1.1, nous avons proposé une nouvelle démonstration de la méthode du réassignement pour la transformée de Fourier et pour le modèle **M02**. En utilisant la même méthode appliquée au modèle **M12**, nous allons montrer qu'il est possible de dériver un couple d'équations linéaires par rapport aux trois paramètres $\beta_M$, $\gamma$ et $\mu$. Nous verrons ensuite comment utiliser ce couple d'équations pour estimer facilement ces paramètres.

En utilisant la même méthode que dans la section 4.2.1.1, on montre que l'équation (4.2.6) devient pour le modèle **M12** :

$$\beta - j\mu + \gamma \frac{X_c(t,\omega;h.t)}{X_c(t,\omega;h)} \quad = \quad \omega + j\frac{X_c(t,\omega;\dot{h})}{X_c(t,\omega;h)} \tag{6.3.1}$$

On rappelle que $X_c$ désigne la TF continue.

En séparant la partie réelle et imaginaire et en remplaçant les TF continues par des TF discrètes, on obtient les équations suivantes :

$$\begin{aligned} \beta + \gamma t_r &= \omega - d_i \\ \mu - \gamma t_i &= -d_r \end{aligned} \tag{6.3.2}$$

où $t_r$, $d_r$, $t_i$ et $d_i$ sont définis par :

$$t_r = \Re\left(\frac{X(t,\omega;\tau h)}{X(t,\omega;h)}\right) \qquad t_i = \Im\left(\frac{X(t,\omega;\tau h)}{X(t,\omega;h)}\right)$$
$$d_r = \Re\left(\frac{X(t,\omega;\dot{h})}{X(t,\omega;h)}\right) \qquad d_i = \Im\left(\frac{X(t,\omega;\dot{h})}{X(t,\omega;h)}\right)$$

$t_r$, $d_r$, $t_i$ et $d_i$ sont directement calculables à partir des TFCTs pour les trois fenêtres $h$, $\dot{h}$ et $\tau h$. Pour estimer les trois paramètres $\beta_M$, $\gamma$ et $\mu$, il suffit d'appliquer le couple d'équations (6.3.2) à $L$ bins différents. On obtient alors $2L$ équations distinctes pour trois inconnues. $\beta$ va être différent d'une trame à l'autre, mais on peut facilement exprimer toutes les équations en fonction de la fréquence de la trame du milieu $\beta_M$, $\beta = \beta_M + \gamma\delta_t$ où $\delta_t$ est le temps séparant la trame considérée de la trame centrale :

$$\begin{aligned} \beta_M + \gamma(t_r + \delta_t) &= \omega - d_i \\ \mu - \gamma t_i &= -d_r \end{aligned} \tag{6.3.3}$$

En appliquant ce couple d'équation à $L$ bins, une simple inversion matricielle de dimension 3 nous permet donc de trouver une estimation de nos paramètres. On note $\mathbf{t_i}$, $\mathbf{t_r} + \delta_\mathbf{t}$, $\omega + \mathbf{d_i}$ et $\mathbf{d_r}$ les vecteurs colonnes des coefficients des équations, c'est à dire que $\mathbf{t_i}$ contient les $t_i$ de plusieurs bins. On aura alors :

$$\mathbf{Ap} = \mathbf{b} \tag{6.3.4}$$

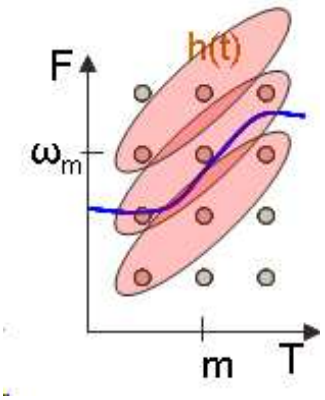



FIG. 6.14: Bins utilisés (en rose) par les méthodes du Réassignement et du vocodeur de phase. En bleu est dessiné l'évolution réelle de la fréquence de la sinusoïde. Dans le premier cas le couple d'équations est appliqué à chaque bin (18 équations), et dans le deuxième cas aux triplets représentés par des ellipses (6 équations).

où $\mathbf{A}$, $\mathbf{b}$ et $\mathbf{p}$ sont définis par :

$$\mathbf{A} = \begin{bmatrix} \mathbf{1} & \delta_\mathbf{t} + \mathbf{t_r} & \mathbf{0} \\ \mathbf{0} & -\mathbf{t_i} & \mathbf{1} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \omega - \mathbf{d_i} \\ -\mathbf{d_r} \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \beta_M \\ \gamma \\ \mu \end{bmatrix} \tag{6.3.5}$$

$\mathbf{1}$ et $\mathbf{0}$ sont des vecteurs colonnes avec uniquement des 1 et des zéros. On en déduit une estimation de $\mathbf{p}$ :

$$\hat{\mathbf{p}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{6.3.6}$$

$\mathbf{A}^T\mathbf{A}$ est une matrice symétrique réelle de dimension 3, donc facilement inversible.

La procédure d'estimation retenue utilise trois trames consécutives, comme dans le cas des méthodes présentées dans la section 6.2. On effectue un suivi de bins maximums et on applique les équations (6.3.2) au bin maximum et à ses deux bins adjacents, et ceci pour les trois trames consécutives, c'est à dire pour 9 bins différents en tout (voir Figure 6.14). La procédure d'estimation complète peut se résumer ainsi :

1. Calcul des TFCTs et suivi de bin maximum : calcul des FFTs pour le temps $t_M$, $t_{m_1} = t_M - T/2$ et $t_{m_2} = t_M + T/2$, en utilisant les fenêtres $h$, $\tau h$ et $\dot{h}$. Calcul de $k_M$, $k_1$ et $k_2$.

2. On applique le couple d'équations (6.3.3) aux bins $k_M$, $k_1$, $k_2$, $k_M \pm 1$, $k_1 \pm 1$, $k_2 \pm 1$, et on remplit les matrices $\mathbf{A}$ et $\mathbf{b}$.

3. Estimation des paramètres :$\hat{\mathbf{p}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$

### 6.3.2 Estimation basée sur le vocodeur de phase

De façon analogue au réassignement, l'équation du vocodeur de phase (6.2.10) va avoir un équivalent en amplitude, nous conduisant au couple d'équations suivant :

$$\begin{aligned} \beta_M + \gamma(t_r + \delta_t) &= \frac{\arg(Y)}{T} + \frac{\delta_\omega}{T} t_r \ [2\pi] \\ \mu - \gamma t_i &= \frac{\log(|Y|)}{T} - \frac{\delta_\omega}{T} t_i \end{aligned} \tag{6.3.7}$$

où $Y = X(t_2, \omega_2; h)/X(t_1, \omega_1; h)$. La première équation est l'équation du vocodeur de phase réassigné (6.2.16) et la deuxième son pendant en amplitude. Cette dernière équation est dérivée de façon analogue à la première, par développement de Taylor d'ordre 1.

On va ensuite utiliser une méthode similaire à la méthode décrite dans la section précédente pour estimer tous les paramètres. Il y a cependant quelques différences. Tout d'abord, on applique ici les équations sur des triplets de bins. Les triplets en question doivent suivre la variation de fréquence pour s'assurer que le biais des équations reste borné. Cela limite donc fortement le nombre de triplets de bins sur lequel on peut appliquer ces équations.

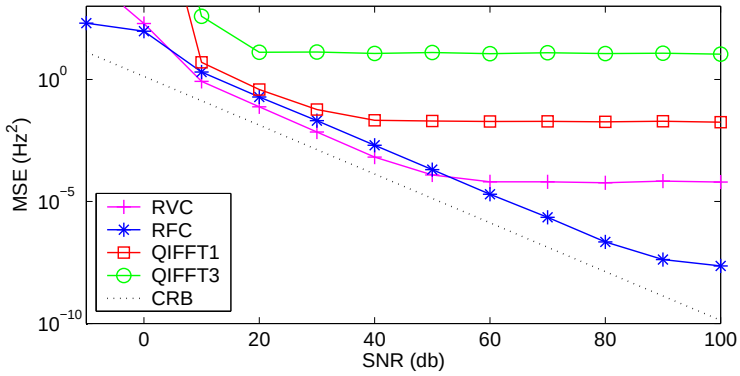
Comme pour le réassignement, on va considérer trois trames successives, et on applique les équations sur les trois bins de la trajectoire du maximum et sur les deux trajectoires adjacentes. Ces trajectoires sont représentées par des ellipses sur la Figure 6.14. Le nombre d'équations total est cette fois de 6 contre 18 dans le cas du réassignement. Un autre problème apparaît, lié au faible nombre d'équations. En effet, on avait fait remarquer déjà que le temps réassigné $t_r$ tend vers zéro lorsque $\gamma$ tend vers zéro. Donc lorsque $\gamma$ va tendre vers zéro, la première équation du couple (6.3.7) sera quasiment identique pour les trois triplets de bins de la Figure 6.14 et l'estimation sera instable. Il est donc préférable d'utiliser uniquement la deuxième équation pour estimer $\gamma$, et ensuite de déduire $\beta_M$ et $\mu$. Si l'on reprend les notations de la section précédente, on va donc estimer les paramètres ainsi

$$\hat{\mathbf{p_2}} = (\mathbf{A_2}^T \mathbf{A_2})^{-1} \mathbf{A_2}^T \mathbf{b_2} \tag{6.3.8}$$

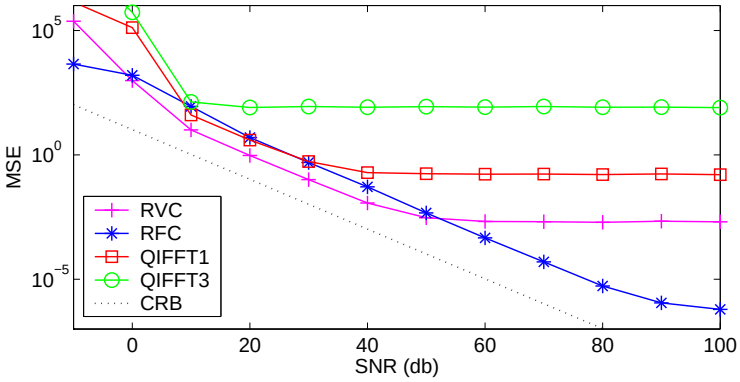
avec $\mathbf{A_2} = \begin{bmatrix} \mathbf{1} & -\mathbf{t_i} \end{bmatrix}$, $\mathbf{b_2} = -\mathbf{d_r}$ et $\mathbf{p_2} = \begin{bmatrix} \mu \\ \gamma \end{bmatrix}^T$. On en déduit ensuite $\beta_M$ en remplaçant $\gamma$ par son estimée dans la première équation.

L'algorithme proposé dans ce cas est donc le suivant :

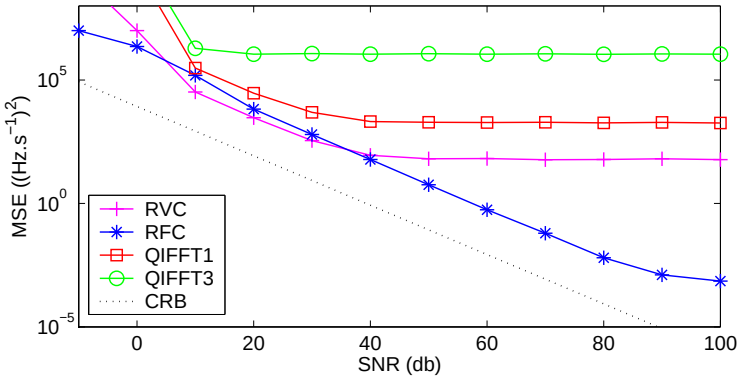
1. Calcul des TFCTs et suivi de bin maximum : calcul des FFTs pour le temps $t_M$, $t_{m_1} = t_M - T/2$ et $t_{m_2} = t_M + T/2$, en utilisant les fenêtres $h$, $\tau h$ et $\dot{h}$. Calcul de $k_M$, $k_1$ et $k_2$.
2. On applique le couple d'équations (6.3.7) aux triplets de bins $(k_1, k_M, k_2)$, $(k_1 - 1, k_M - 1, k_2 - 1)$ et $(k_1 + 1, k_M + 1, k_2 + 1)$. On remplit les matrices $\mathbf{A_2}$ et $\mathbf{b_2}$.
3. Estimation des paramètres $\mu$ et $\gamma$ :$\hat{\mathbf{p_2}} = (\mathbf{A_2}^T \mathbf{A_2})^{-1} \mathbf{A_2}^T \mathbf{b_2}$
4. Estimation de $\beta_M$




(a) $\beta_M$




(b) $\mu$




(c) $\gamma$

FIG. 6.15: Evaluation des méthodes d'estimation des paramètres d'ordre supérieur pour le modèle **M12** ($\gamma_m = 8000$,$\mu_m = 100$)

### 6.3.3 Evaluation

Dans cette section on va évaluer les nouvelles méthodes d'estimation des paramètres d'ordre supérieur du modèle **M12**. On rappelle que ces paramètres sont la

fréquence à l'instant $t_M$, et les variations d'amplitude et de fréquence. La méthode notée RFC est la méthode décrite à la section 6.3.1, et la méthode RVC celle décrite à la section 6.3.2. Ces deux méthodes d'estimation des paramètres supérieurs du modèle **M12** sont originales. On les compare à la seule méthode équivalente déjà existante, la QIFFT. Ici pour avoir des méthodes de complexité à peu près équivalente, on a choisi un facteur de padding de 4 pour la QIFFT, de 2 pour la méthode RV et de 1 pour la méthode du réassignement. On rappelle que le nombre de TF requis par chaque méthode est respectivement de 1 pour la QIFFT, de 2 pour la méthode RV, et de 3 pour le réassignement. On peut signaler que la méthode de la QIFFT est dans ce cas de figure, la plus complexe des trois, particulièrement la version avec fenêtre longue.

Dans la section précédente, on a vu que les méthodes RV et réassignement fournissent également une estimation de fréquence pour le temps réassigné. Si on compare les performances d'estimation de $\beta$ pour le temps réassigné (Figure 6.13(a)) et pour le temps $t_M$ (Figure 6.15(a)), on s'aperçoit que les performances sont meilleures dans le premier cas. Cependant, il est intéressant de pouvoir estimer la fréquence à un temps bien défini si l'on veut estimer les paramètres $\alpha$ et $\lambda$. En effet, on verra dans la section 6.4 que les fonctions de correction utilisées pour estimer $\alpha$ et $\lambda$ sont définies à un temps particulier, ici $t_M$. Dans un schéma complet d'estimation, la fréquence estimée au temps $t_M$ sera incontournable.

Les deux méthodes proposées sont plus performantes que la QIFFT, présentant toutes les deux un biais et une variance plus faibles, pour chacun des paramètres. La méthode RVC présente un biais plus important que la méthode basée sur le réassignement (RFC) dû à l'approximation (6.2.16) et au faible coefficient de padding utilisé. Sa variance est également légèrement plus faible que pour le réassignement, ce qui permet à cette méthode d'être la plus proche du CRB pour l'estimation des trois paramètres $\beta_M$, $\gamma$ et $\mu$. Enfin on voit que la méthode décroche maintenant, comme la QIFFT, lorsque le SNR devient trop faible. Ceci est dû à l'utilisation des bins adjacents pour l'estimation des paramètres, qui sont moins robustes au bruit que dans le cas du bin maximum. Dans la méthode basée sur le réassignement (RFC) cet effet est très atténué grâce au nombre important d'équations différentes utilisées pour l'estimation.

## 6.4 Estimation d'amplitude et de phase

Dans la section 3.1, nous avons vu que l'estimation au sens du maximum de vraisemblance de l'amplitude et de la phase constante se faisait en deux temps : d'abord estimation des paramètres d'ordre supérieur puis estimation des amplitudes (équation (3.1.10)). Dans le cas du modèle **M01**, nous avons rappelé le résultat classique établissant que l'estimateur d'amplitude de Fourier est équivalent à l'estimateur ML pour une sinusoïde. Nous allons maintenant voir que dans le cas général du modèle **MKQ**, une méthode similaire en deux étapes et basée sur la transformée de Fourier peut être dérivée. Pour cela on revient à la transformée de Fourier d'une sinusoïde

pour ce modèle :

$$X(t,\omega;h) = \mathrm{e}^{A_0+j\Phi_0} \sum_{i=-N/2}^{N/2} h(\tau_i)\, \mathrm{e}^{\sum_{k=1}^{K} A_k \tau_i^k + j(\Phi_1-\omega)\tau_i + j\sum_{q=2}^{Q}\Phi_q\tau_i^q}$$

$$X(t,\omega;h) = \mathrm{e}^{A_0+j\Phi_0}\, \Gamma(A_{1..K}, \Phi_1-\omega, \Phi_{2..Q}; h)$$

On en déduit que $A_0$ et $\Phi_0$ vérifient les formules suivantes :

$$\Phi_0 = \arg(X(t,\omega;h)) - \arg(\Gamma(A_{1..K}, \Phi_1-\omega, \Phi_{2..Q}; h))[2\pi] \tag{6.4.1}$$

$$A_0 = \log(|X(t,\omega;h)|) - \log(|\Gamma(A_{1..K}, \Phi_1-\omega, \Phi_{2..Q}; h)|) \tag{6.4.2}$$

$\Gamma$ est une fonction connue des paramètres $\{A_i\}_{1..K}, \{\Phi_i\}_{1..Q}$. Si l'on dispose d'une estimation de ces paramètres alors les équations précédentes nous donnent une estimation de l'amplitude et de la phase. Ces formules sont similaires à celles de l'estimation ML, car elles font toutes deux intervenir une transformation polynomiale, ici la fonction $\Gamma$. La première différence est qu'il y a une pondération par une fenêtre d'analyse. On peut remarquer qu'une estimation ML pondérée donnerait lieu à des formules équivalentes. La deuxième différence est que l'on se sert de la transformée de Fourier pour réduire l'intervalle des valeurs possibles pour la fréquence, c'est à dire que l'on a remplacé $\Phi_1$ par $\Phi_1 - \omega$ dans la transformation polynomiale.

La fonction $\Gamma$, comme l'estimateur ML sont des fonctions assez coûteuses en temps de calcul. Il est donc intéressant de prétabuler ces fonctions pour effectuer un calcul rapide. L'avantage de la deuxième formulation est alors évident, car il permet de réduire considérablement le nombre de valeur de l'intervalle de fréquence à tabuler. En effet, si les valeurs des paramètres des autres paramètres $\{A_i\}_{1..K}, \{\Phi_i\}_{2..Q}$ sont dans un voisinage de zéro, $\Phi_1 - \omega$ sera dans un intervalle borné et proche de $[-R, R]$, où $R$ est la précision de la TF. De plus si les valeurs des paramètres $\{A_i\}_{1..K}, \{\Phi_i\}_{2..Q}$ sont dans un voisinage de zéro, la fonction $\Gamma$ dépendra de façon quasi linéaire de ces paramètres, diminuant encore le nombre de valeurs nécessaire à prétabuler. Une fois la fonction prétabulée, une valeur particulière est déduite de la table par interpolation linéaire.

Dans l'article [Betser et al., 2006b] nous avons utilisé cette méthode dans le cas de l'estimation de phase pour un modèle **M02**. Nous avons constaté qu'en fait la dépendance de $\Gamma$ en fonction de la variation de fréquence était linéaire sur un large intervalle de fréquence autour de zéro. Nous avons également présenté une dérivation analytique du coefficient linéaire, via un développement de Taylor. Ce chapitre est donc une généralisation pour un modèle **MKQ** de la méthode présentée dans cet article.

Dans les expériences par la suite, nous ne nous intéresserons qu'au modèle **M12**. Nous résumons donc ici la méthode pour ce modèle uniquement :

1. Initialisation : Prétabulation de la fonction $\Gamma(\mu, \beta, \gamma; h)$ pour $|\beta| \leq R$, $|\mu| \leq \mu_m$ et $|\gamma| \leq \gamma_m$.
2. Calcul de la FFT zéro-phase pour le temps $t$
3. Sélection du bin maximum $k = \arg\max_i |X(t, \omega_i; h)|$

4. On suppose que l'on dispose d'une estimation de $\mu$, $\beta$ et $\gamma$, et on en déduit :
   a) L'amplitude : $\lambda = \log(|X(t,\omega_k;h)|) - \log(|\Gamma(\hat{\mu},\hat{\beta}-\omega_k,\hat{\gamma};h)|)$
   b) La phase : $\alpha = \arg(X(t,\omega_k;h)) - \arg(\Gamma(\hat{\mu},\hat{\beta}-\omega_k,\hat{\gamma};h))[2\pi]$

### 6.4.1 Biais et variance

Le bais et la variance des estimateurs d'amplitude et de phase dépendent directement des biais et variances des estimations des paramètres d'ordre supérieur, $\{\hat{A}_i\}_{1..K}$ et $\{\hat{\Phi}_i\}_{1..Q}$. On va faire les trois suppositions suivantes sur ces estimateurs :

- Tous les estimateurs sont basés sur Fourier et sont formés par des combinaisons de bins ;
- Pour tous ces estimateurs, le signal étudié est bien résolu par la transformée de Fourier ;
- Le biais déterministe de ces estimateurs est faible par rapport à la valeur à estimer.

D'après la section 5.4, si $\hat{p}$ est un estimateur de $p$, les deux premières hypothèses nous permettent de dire que toutes les estimations auront la forme : $\hat{p} = p + \epsilon_D + \epsilon_N$. La deuxième hypothèse nous assure également que $\tau_N \ll 1$. La dernière hypothèse nous permet de dire que $\epsilon_D + \epsilon_N << p$. On peut donc appliquer directement la méthode de la section 5.4, et en déduire que les estimations d'amplitude et de phase auront aussi la forme $\hat{p} = p + \epsilon_D + \epsilon_N$. Afin d'alléger la notation, on appelle $\mathbf{p}$ l'ensemble des paramètres de la fonction $\Gamma$, $\mathbf{p} = (A_{1..K}, \Phi_1 - \omega, \Phi_{2..Q})$. Dans le cas de la phase, on aura :

$$\epsilon_D = -\sum_{k=1}^{K} \epsilon_{D_k} \Im\left(\frac{\Gamma(\mathbf{p};h\tau^k)}{\Gamma(\mathbf{p};h)}\right) - \sum_{q=1}^{Q} \epsilon_{D_q} \Re\left(\frac{\Gamma(\mathbf{p};h\tau^q)}{\Gamma(\mathbf{p};h)}\right) \tag{6.4.3}$$

$$\epsilon_N = \Im\left(\frac{N}{X}\right) - \sum_{k=1}^{K} \epsilon_{D_k} \Im\left(\frac{\Gamma(\mathbf{p};h\tau^k)}{\Gamma(\mathbf{p};h)}\right) - \sum_{q=1}^{Q} \epsilon_{D_q} \Re\left(\frac{\Gamma(\mathbf{p};h\tau^q)}{\Gamma(\mathbf{p};h)}\right) \tag{6.4.4}$$

où $\epsilon_{D_k}$ et $\epsilon_{N_k}$ (resp. $\epsilon_{D_q}$ et $\epsilon_{N_q}$) sont les erreurs déterministes et stochastiques de l'estimateur de $A_k$ (resp. de $\Phi_q$). Dans le cas de l'amplitude, on obtient des fonctions quasiment identiques, il suffit de remplacer les opérateurs de partie imaginaire par des opérateurs partie réelle, et vice versa.

Maintenant on ajoute cette hypothèse aux trois précédentes :

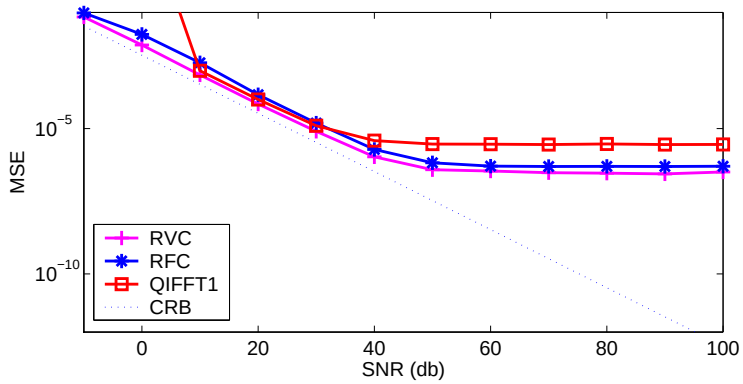
- $\epsilon_{N_k}$ est du même ordre de grandeur ou inférieur à $\Im(N/X)$, c'est à dire que $\mathrm{var}(\epsilon_{N_k}) \leq \mathrm{var}(\Im(N/X))$.

Cette hypothèse sera vérifiée par la plupart des estimateurs basés sur Fourier. Le fait que $\tau_N \ll 1$ nous permet de dire que :

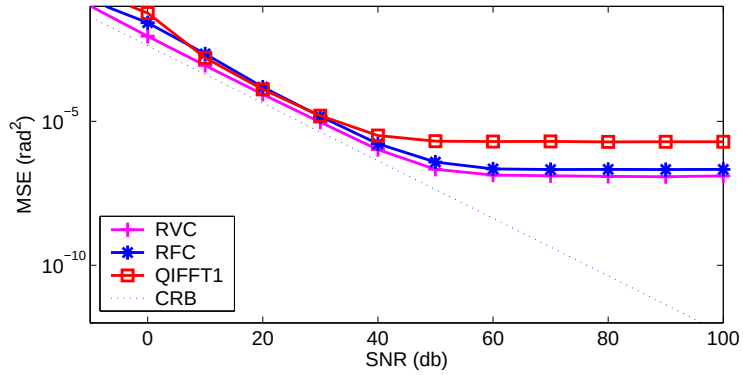
$$\Im\big(\frac{\Gamma(\mathbf{p};h\tau^i)}{\Gamma(\mathbf{p};h)}\big) \ll 1, \qquad \Re\big(\frac{\Gamma(\mathbf{p};h\tau^i)}{\Gamma(\mathbf{p};h)}\big) \ll 1$$

pour tout $i \geq 1$. On en déduit que l'erreur stochastique sur l'estimation de phase se simplifie en :

$$\epsilon_N = \Im\big(\frac{N}{X}\big) \tag{6.4.5}$$




(a) $\lambda$




(b) $\alpha$

FIG. 6.16: Evaluation de la méthode d'estimation de l'amplitude et de la phase pour le modèle **M12** ($\gamma_m = 8000$,$\mu_m = 100$)

Pour l'estimateur d'amplitude, on déduira la même chose en remplaçant $\Im()$ par $\Re()$.

Ces quatre hypothèses, formulées légèrement différemment, ont déjà été utilisées avec succès dans l'article [Betser et al., 2007] pour trouver la variance dans le cas de la méthode d'estimation de fréquence utilisant des phases estimées par la fonction (6.4.1) (méthode PCV). Cette méthode a été décrite dans la section 6.2.4.

### 6.4.2 Evaluation

Dans le cas du modèle **M01**, on retombe sur l'estimateur de Fourier classique décrit dans la section 4.1.1. Cet estimateur a déjà été testé dans la section 4.4.

La méthode d'estimation des phases et des amplitudes n'est testée ici que dans le cas du modèle **M12**. Etant donné que cette méthode nécessite des estimations de paramètres d'ordre supérieur, elle a été couplée avec les méthodes d'estimation décrites dans la section 6.3, dédiées à l'estimation de ces paramètres. Deux versions de l'algorithme d'estimation des phases et des amplitudes sont donc présentées : la première, notée RVA, est couplée avec le réassignement (méthode RFC) pour l'esti-

mation des paramètres d'ordre supérieur, la deuxième est couplée avec le vocodeur réassigné (méthode RVC). Ces deux versions sont comparées sur la Figure 6.16 à la QIFFT qui possède ses propres estimateurs pour la phase et l'amplitude.

Comme dans la section précédente, les méthodes utilisent respectivement des coefficients de padding de un pour RFA, deux pour RVA et quatre pour la QIFFT. Ici la QIFFT est employée uniquement dans la version avec une longue fenêtre, de 48 ms. Les deux premières méthodes sont donc d'une complexité à peu près équivalente. La QIFFT utilisant des transformées de Fourier beaucoup plus longues est plus complexe que les deux premières, comme nous l'avons déjà expliqué dans la section 6.3.3. Les fonctions d'amplitude et de phase ont été prétabulées avec 8000 points répartis sur les intervalles étudiés.

La Figure 6.16, nous montre que les estimateurs ont des résultats similaires sur l'estimation des amplitudes et sur l'estimation des phases. Dans les deux cas, leur variance est très proche de la borne théorique, avec un léger avantage, au vocodeur réassigné. Le biais des méthodes, visible pour les hauts SNRs, est légèrement plus faible pour les méthodes RV et RF. Le biais de la méthode RV est cohérent avec les biais des estimations des paramètres d'ordre supérieur (voir Figure 6.15), car il apparaît pour la même valeur de SNR. Pour la méthode RF le biais apparaît ici pour des SNRs plus faibles. Ce phénomène est dû à l'échantillonnage des fonctions d'amplitude et de phase. Pour diminuer le biais, il faudrait augmenter le nombre de points utilisés pour représenter ces fonctions.

## 6.5 Conclusion

Dans cette section nous avons présenté les estimateurs développés dans le cadre de la thèse. Un effort de généralisation important a été réalisé. Dans le cas des estimateurs de fréquence basés sur le modèle **M01**, cela nous a permis de mettre au point une méthode pour réduire de façon arbitraire le biais d'une méthode existante, et également d'adapter les estimateurs du type interpolateur de spectre utilisant la phase, comme la méthode de Macleod, à n'importe quel type de fenêtre. Cependant le gain de performance attendu pour ces nouveaux estimateurs est assez faible, car les méthodes existantes présentent des résultats déjà très proches de la borne théorique.

Nous avions également constaté dans la comparaison expérimentale des estimateurs de la littérature, dans la section 4.4, que très peu de ces estimateurs sont réellement robustes à ces modulations : les plus robustes sont la méthode du réassignement pour l'estimation de fréquence et la méthode de la QIFFT, qui est dédiée à l'étude du modèle **M12**. Nous nous sommes donc intéressés ensuite à des méthodes robustes à des modulations à la fois d'amplitude et de fréquence. Nous avons développé trois méthodes alternatives au réassignement pour l'estimation de fréquence, dans le cas du modèle **M02** donnant des performances similaires, et requérant moins de TF à calculer. Dans le cas du modèle **M12**, nous avons proposé deux schémas de calcul complets des paramètres, à la fois moins complexes et plus performants que la QIFFT, qui est la seule méthode équivalente déjà existante.

Deuxième partie

# Applications de la modélisation sinusoïdale à l'indexation audio

CHAPITRE 7

# Estimation de fréquence fondamentale

## 7.1 Introduction

La plupart des signaux de parole ou de musique sont caractérisés par l'harmonicité. Dans le cas de la parole, on parle généralement de voisement. Physiquement, un son harmonique est composé de plusieurs partiels, dont les fréquences sont réparties de façon régulière le long de l'axe fréquentiel. Plus précisément, la fréquence de chaque partiel est un multiple de la fréquence du partiel le plus grave, appelé fondamentale :

$$x_h(t) = \sum_{i=1}^{Q} A_i(t)\, \mathrm{e}^{j(\Phi_i + i\int_0^t \Omega_0(t)dt)} \tag{7.1.1}$$

$\Omega_0(t)$ dénote ici la fréquence instantanée de la fondamentale. $A_i(t)$ est l'amplitude instantanée du partiel $i$ et $\Phi_i$ est sa phase initiale. Pour tout instant $t$, on a bien $\Omega_i(t) = i\Omega_0(t)$. On voit immédiatement que le déroulement de phase sera également un multiple du déroulement du partiel fondamental. Dans le cas des sons harmoniques, la fréquence perçue par l'oreille humaine, appelée pitch, correspond à la fréquence fondamentale. Avec l'amplitude des partiels, et les phases initiales, la fréquence fondamentale permet de complètement déterminer le signal harmonique. Il existe des cas où l'oreille perçoit un pitch sans que le signal ne possède de fréquence fondamentale, par exemple dans le cas des bruits blancs filtrés, donc rigoureusement il n'y a pas équivalence entre le pitch et la fréquence fondamentale. Cependant nous ne nous intéresserons ici qu'aux parties harmoniques des signaux, et nous utiliserons donc l'un où l'autre terme indifféremment.

Un certain nombre des estimateurs de la littérature, dont celui de Cheveigné [de Cheveigné and Kawahara, 2002], sont suffisamment performants pour que la tâche d'estimation de pitch d'un signal monophonique soit considérée comme presque résolue dans le cas où ce signal est enregistré dans de bonnes conditions. L'estimation

de fondamentale dans un ensemble polyphonique reste quand à elle une tâche très difficile, encore sujette à de nombreuses recherches [Klapuri, 2004].

Dans ce chapitre, nous ne nous intéresserons qu'à l'estimation de pitch pour les signaux monophoniques. Le but que nous nous sommes fixé est d'utiliser la modélisation sinusoïdale comme paramètre d'entrée de systèmes d'indexation. La tâche d'estimation de pitch monophonique est une tâche d'indexation qui nous a paru être une première étape intéressante avant de s'attaquer à des problèmes plus complexes, comme l'identification audio. La première raison est que certains algorithmes dédiés à cette tâche sont très proches de l'algorithme envisagé. La deuxième raison est que l'estimation de pitch est une tâche plus simple que l'identification audio, dans le sens où l'on possède un modèle clairement identifié (équation (7.1.1)) et un nombre de paramètres restreints.

Nous présenterons d'abord un bref état de l'art dans la section 7.2, non exhaustif mais couvrant les principales techniques utilisées. Deux méthodes de référence de l'état de l'art seront présentées plus en détail, celle de Talkin [Talkin, 1995] et celle de Cheveigné [de Cheveigné and Kawahara, 2002]. Dans la section 7.2 nous verrons que des méthodes d'estimation de pitch basées sur les pics sinusoïdaux existent déjà. Nous allons ensuite décrire une méthode simple inspirée de ces dernières méthodes à la section 7.3. Une comparaison avec les algorithmes de Cheveigné et de Talkin sur un très large corpus de parole sera présenté dans la section 7.4.

## 7.2 Etat de l'art

Il existe de très nombreuses méthodes pour estimer la fréquence fondamentale d'un signal monophonique. Ces algorithmes sont particulièrement utilisés en analyse et en codage de parole, car il existe de très nombreux documents sonores où la parole est présente seule, ou très prédominante (émission radiophonique ou télévisuelle, téléphonie etc.). Le terme d'estimation de pitch prédominant est donc souvent employé pour ces algorithmes. Le but ici n'est pas de détailler toutes les techniques existantes, mais de faire un tour des grandes familles de méthodes. Des revues et des comparaisons plus complètes peuvent être trouvées dans [Rabiner and Juang, 1993], [Hess, 1991], [de Cheveigné and Kawahara, 2001] et [Huang et al., 2001].

Il n'y a pas de découpage idéal pour définir ces familles de méthodes. Le découpage le plus évident est celui du domaine temporel ou fréquentiel dans lequel opèrent ces méthodes. Cependant la plupart des méthodes présentent des similitudes indépendantes du domaine de calcul. Nous préférons les regrouper ici par similitude conceptuelle, permettant une présentation plus succincte. Nous parlerons donc d'abord des méthodes basées sur une mesure de similarité du signal avec lui-même dans la section 7.2.1, auxquelles appartiennent les méthodes de Talkin et de Cheveigné, choisies comme références. Ces deux méthodes seront décrites plus en détail. Ensuite nous verrons dans la section 7.2.2 les méthodes basées sur la similarité du signal avec un modèle prédéfini, qui sont des méthodes du type reconnaissance de forme. Les méthodes basées sur les fréquences instantanées, comme la méthode que nous avons développée, appartiennent à cette famille.

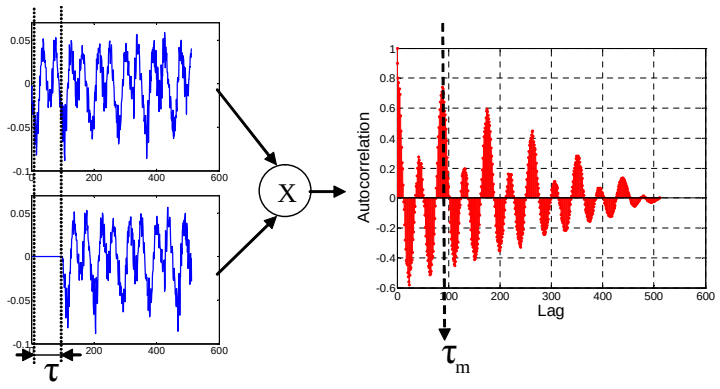



FIG. 7.1: Principe de l'auto-corrélation. Le signal est multiplié avec lui même décalé d'un temps $\tau$. Le pitch est donné par $1/\tau_m$, où $\tau_m$ est le décalage du pic maximal.

Les algorithmes vont souvent présenter une étape de pré-traitement, ou de post-traitement spécifique. Typiquement le pré-traitement peut inclure un filtrage passe-bas, une amplification des composantes de haute fréquence ou d'autres traitements spécifiques. Le post-traitement quand à lui est une étape de lissage du pitch, essentiellement pour éliminer les erreurs de double pitch et de demi pitch. Ces deux étapes varient souvent d'un auteur à l'autre et ne seront pas détaillées ici.

### 7.2.1 Méthodes basées sur l'auto-similarité

L'idée des méthodes basées sur l'auto-similarité, consiste à évaluer la similitude d'un signal avec lui-même, mais décalé d'un certain intervalle prédéfini. La Figure 7.1 illustre ce mécanisme dans le cas de l'auto-corrélation. Le signal d'entrée peut être temporel, fréquentiel, voire toute représentation susceptible de présenter une périodicité liée au pitch. La fonction de similitude peut-être la fonction d'auto-corrélation (ACF) bien sûr, et toutes ses variantes : la corrélation croisée (CCF), la corrélation croisée normalisée (NCCF) [Talkin, 1995], la corrélation croisée entre deux segments adjacents (SRPD) [Medan et al., 1991], [Bagshaw et al., 1993], le cepstre [de Cheveigné and Kawahara, 2002]. On trouve également la fonction des différences d'amplitude moyenne (AMDF) [Umapathy et al., 1984] [Hermes, 1988] [Boersma, 1993], la fonction des différences d'amplitude moyenne cumulatives et normalisées [de Cheveigné and Kawahara, 2002] etc.

Souvent la mesure de similarité du signal présente elle aussi une périodicité liée au pitch, donc de nouvelles mesures de similarité peuvent être formées par des applications successives de mesures de similarité, comme l'a souligné Sylvain Marchand [Marchand, 2001] dans le cas de la transformée de Fourier. Des schémas faisant intervenir un calcul d'auto-similarité par sous-bandes est également possible, comme dans le cas du modèle perceptuel unitaire [Medan et al., 1991] [Klapuri, 2004].

Toutes les méthodes basées sur ce principe vont suivre le même schéma de fonctionnement qui est donné sur la Figure 7.2. Une fois la fonction d'auto-similarité

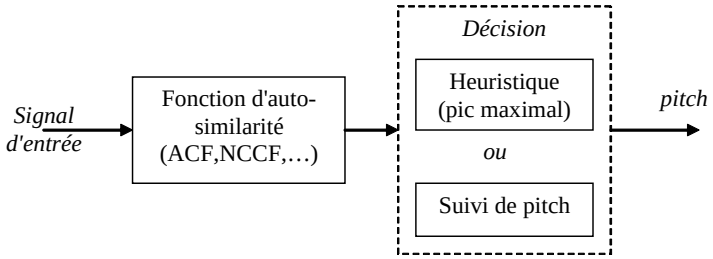



FIG. 7.2: Schéma de fonctionnement des méthodes basées sur l'auto-similarité

calculée, la période fondamentale sera donnée soit simplement par le maximum de la fonction d'auto-similarité, soit par une heuristique particulière. Cette fonction est souvent obtenue en deux étapes pour restreindre la complexité des algorithmes. Elle est d'abord évaluée grossièrement, c'est à dire que les intervalles entre deux mesures vont être assez importants. Une fois que les pics ont été identifiés, un calcul plus précis du maximum est effectué à leur voisinage [de Cheveigné and Kawahara, 2002].

Nous allons maintenant décrire plus en détail les deux algorithmes retenus pour la comparaison. Ces deux algorithmes sont disponibles gratuitement sur internet et sont généralement utilisés comme références pour l'évaluation des algorithmes d'estimation de pitch.

#### 7.2.1.1 RAPT

L'algorithme robuste pour le suivi de pitch (RAPT) est décrit dans [Talkin, 1995]. Le RAPT est un algorithme rapide, basé sur la corrélation croisée normalisée (NCCF[1]) du signal temporel :

$$\phi_t(\tau) = \frac{\sum_{j=t+1}^{t+W} x_j x_{j+\tau}}{\sqrt{e_t e_{t+\tau}}} \tag{7.2.1}$$

$\phi_t(\tau)$ est la NCCF de l'index temporel $t$ pour une latence de $\tau$ échantillons, et pour une fenêtre d'intégration de taille $W$. $x_t$ est l'échantillon $t$ du signal et $e_t$ est l'énergie du signal entre les échantillons $t$ et $t+W$ :

$$e_t = \sum_{j=t+1}^{t+W} x_j^2 \tag{7.2.2}$$

Dans le cas de la parole, [Talkin, 1995] a montré que si l'on exclue le pic en $\tau = 0$, le maximum de la NCCF correspond généralement au pitch et que sa valeur est très proche de 1. Lorsque plusieurs maximums sont proches de 1, le pitch est généralement celui correspondant au plus petit décalage $\tau$. Enfin on constate que dans les zones de parole non-voisées (i.e. non pitchées), le maximum de la NCCF est généralement très faible. Toutes ces propriétés rendent la NCCF particulièrement intéressante pour analyser les signaux de parole.

[1]Normalized Cross Correlation Function.

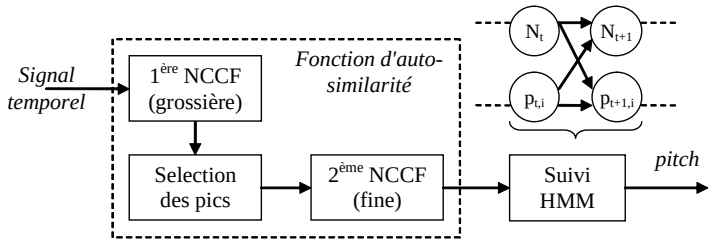



FIG. 7.3: Schéma de fonctionnement de l'algorithme RAPT

Le schéma de l'algorithme est présenté sur la Figure 7.3. Pour réduire la complexité de l'algorithme la NCCF est calculée en deux passes. On calcule d'abord une première NCCF du signal fortement sous-échantillonné, ce qui donne une grossière approximation de la NCCF. On sélectionne les maximums de cette NCCF, puis pour chaque maximum, on calcule une NCCF plus précise à ses alentours. Chaque maximum est alors affiné et gardé comme hypothèse de pitch, noté $P_{t,i}$ sur la Figure 7.3.

En ajoutant en plus l'hypothèse de non voisement, noté $N_t$, un algorithme de programmation dynamique (post-traitement) est utilisé pour trouver la meilleure succession d'états sur l'ensemble des trames du signal. Les transitions entre les états et la probabilité de l'état non-voisé sont des formules empiriques faisant intervenir beaucoup de constantes de réglage, et ne seront pas rappelées ici. On pourra se référer à [Raghuram, 2002] par exemple pour plus de détails. Le meilleur chemin est ensuite calculé en utilisant un algorithme de type Viterbi [Rabiner and Juang, 1993]. L'inconvénient de ce type de traitement dynamique, c'est que le fichier doit être traité d'un bloc, rendant la méthode inapplicable pour les applications en temps réel. Pour de telles applications, le post-traitement doit être désactivé et le pitch estimé est alors tout simplement la fréquence correspondant au plus grand pic de la NCCF. Mais les performances sont dans ce cas beaucoup moins bonnes.

#### 7.2.1.2 YIN

L'algorithme YIN a été développé par Kawahara et de Cheveigné [de Cheveigné and Kawahara, 2002]. Basé sur l'auto-corrélation, il fait intervenir un certain nombre de nouveautés qui rendent l'algorithme particulièrement robuste et précis.

La fonction d'auto-similarité utilisée est une version améliorée de l'auto-corrélation. L'auto-corrélation est notée $r_t(\tau)$ :

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{7.2.3}$$

La fonction des différences $d_t(\tau)$ est définie par :

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \tag{7.2.4}$$

Cette fonction est moins sensible à des variations d'amplitude que l'auto-corrélation simple grâce aux deux premiers termes qui ajustent le niveau d'énergie en fonction

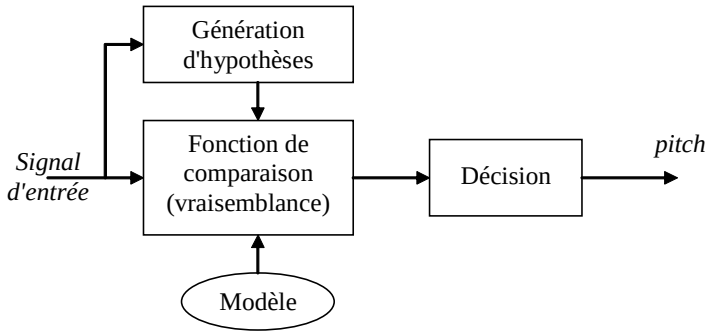



FIG. 7.4: Schéma de fonctionnement des méthodes basées sur la reconnaissance de forme. Elle font intervenir une mesure de comparaison avec un modèle.

du décalage (lag) courant $\tau$ [de Cheveigné and Kawahara, 2002]. On cherche ici à minimiser la fonction $d_t(\tau)$[2]. La version normalisée de cette fonction permet de sélectionner plus facilement le minimum correspondant au pitch :

$$d_t'(\tau) = \begin{cases} 1 & \text{si } \tau = 0, \\ d_t(\tau) / \Big[ (1/\tau) \sum_{j=1}^{\tau} d_t(j) \Big] & \text{sinon.} \end{cases} \tag{7.2.5}$$

Le minimum est alors sélectionné selon l'heuristique suivante :

- S'il n'y a aucun minimum inférieur à un seuil absolu fixé, le minimum sélectionné est le minimum global
- Sinon on sélectionne le plus petit $\tau$ donnant un minimum inférieur à ce seuil.

Cette heuristique permet de diminuer fortement les erreurs d'octave. Enfin la valeur est affinée par interpolation parabolique, puis par un lissage local des résultats.

L'algorithme YIN ne possède pas de détection de voisement spécifique. Une heuristique simple sera appliquée si cette information est requise, en définissant un seuil absolu sur la fonction d'auto-similarité.

### 7.2.2 Méthodes de type reconnaissance de forme

Contrairement aux méthodes précédentes les méthodes basées sur la reconnaissance de forme impliquent une modélisation du signal recherché : ici un peigne de sinusoïdes décrit par l'équation (7.1.1). On va également devoir définir une distance entre le signal et le modèle choisi, ou encore une mesure de vraisemblance du signal par rapport au modèle défini. Le pitch est estimé comme étant un extremum de cette fonction de comparaison [Griffin and Lim, 1988], [Ferrari et al., 1992], [Chan and So, 2004], [Goto, 2001], [Tabrikian et al., 2004]. Le schéma d'un tel système est résumé sur la Figure 7.4. Le modèle en question peut également être un modèle plus général qui aura été appris sur un corpus d'entraînement.

[2]Dans le YIN, la mesure de similarité est une distance. En effet, la fonction des différences a un signe opposé à la NCCF.

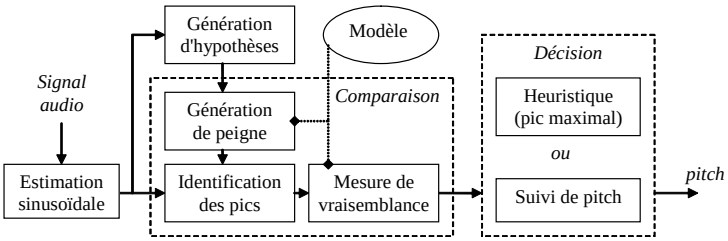



FIG. 7.5: Schéma de fonctionnement de la méthode envisagée. Il s'agit d'une méthode de reconnaissance de forme basée sur les pics sinusoïdaux.

### Méthodes basées sur les fréquences instantanées

Notre approche se situe dans une sous catégorie particulière, qui utilise une représentation intermédiaire, les fréquences instantanées, avant de calculer l'adéquation du signal au modèle. La spécificité de ces méthodes est d'être en deux étapes, la première étant obligatoirement le calcul des fréquences instantanées. Ces fréquences peuvent être calculées dans le domaine temporel, en utilisant des bancs de filtres comme dans l'algorithme TEMPO [Kawahara et al., 1999], et dans d'autres approches [Yang et al., 1994], [Abe et al., 1995] ou plus récemment [Klapuri, 2003]. Les algorithmes les plus proches de notre méthode, ceux qui vont nous intéresser ici, vont calculer les fréquences instantanées dans le domaine fréquentiel, c'est à dire qu'ils vont faire intervenir une étape d'estimation sinusoïdale dans le domaine de Fourier, selon une des méthodes présentées dans la première partie de la thèse, aux chapitres 4 et 6. Leur schéma général de fonctionnement est décrit par la Figure 7.5.

Depuis les travaux précurseurs de MacAulay et Quatieri [McAulay and Quatieri, 1990], un certain nombre de méthodes d'analyse basées sur les pics sinusoïdaux a été proposé. Brown [Brown, 1991], [Brown, 1992] utilise une version modifiée de la transformée de Fourier, la transformation à Q constant, avant de réaliser la reconnaissance de forme. Nous terminons notre tour d'horizon par deux méthodes très proches de la méthode que l'on veut mettre en place, recherchant une correspondance pic à pic entre les pics extraits du spectre et les pics du modèle. La première développé par Maher [Maher and Beauchamp, 1994] utilise une mesure de correspondance empirique et une association des pics dans les deux sens : spectre vers modèle et modèle vers spectre. La deuxième, décrite par Doval et Rodet [Doval and Rodet, 1991], teste toutes les combinaisons de pics et assigne une vraisemblance à chacune de ces combinaisons. Etant donné la proximité de ces méthodes avec celle que l'on veut développer, il est intéressant de les examiner plus en détail.

On note $\tilde{Q} = \{\tilde{q}_i\}_{i\in[1..\tilde{K}]}$ l'ensemble des pics présents dans le signal. $\tilde{K}$ est le nombre de pics trouvés. Chaque pic est caractérisé par l'amplitude et la fréquence : $\tilde{q}_i = (\tilde{A}_i, \tilde{f}_i)$. La phase est considérée comme non informative. Le modèle de signal est aussi un ensemble de pics $Q = \{q_i(p)\}_{i\in[1..K]}$, caractérisé uniquement par la fréquence fondamentale $p$ et les amplitude $A_i$ : $q_i(p) = (A_i, ip)$.

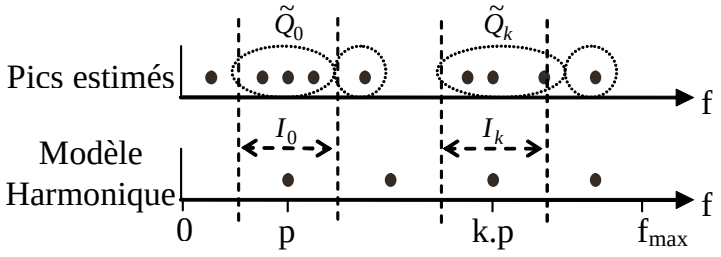



FIG. 7.6: Identification des fréquences dans la méthode de Doval et Rodet

#### 7.2.2.1 La méthode de Doval et Rodet

La méthode de Doval et Rodet [Doval and Rodet, 1991] est une méthode probabiliste qui évalue la vraisemblance d'une hypothèse de pitch connaissant les pics présents dans le signal. Etant donné une hypothèse de pitch $p$, le spectre est partitionné suivant $K$ intervalles : $I_k = [(k-0.5)p, (k+0.5)p]$. Sur chacun de ces intervalles, il y a un ensemble $\tilde{Q}_k$ de pics présents (voir Figure 7.6). $\tilde{Q}_k$ représente donc une partition de l'ensemble des pics présents : $\tilde{Q} = \bigcup_k \tilde{Q}_k$. On va alors chercher à évaluer la vraisemblance de l'ensemble des pics observés $\tilde{Q}$ conditionnellement à la fréquence fondamentale hypothèse $p$ :

$$L(p) = P(\tilde{Q}|p) = \prod_{k=1}^{K} P(\tilde{Q}_k|p) \tag{7.2.6}$$

Pour simplifier l'expression de cette vraisemblance les auteurs font une simplification importante, que Maher utilise implicitement et que nous utiliserons également par la suite. Au lieu de calculer la probabilité $P(\tilde{Q}_k|p)$ pour tous les pics présents dans $\tilde{Q}_k$, on ne va retenir que le pic le plus probable. En d'autre terme, avant de calculer la vraisemblance, on va associer chaque harmonique $kp$ du modèle avec le pic le plus vraisemblable. Le critère d'association des harmoniques utilisé par les auteurs est de retenir le pic $\tilde{q}_i$ ayant la fréquence la plus proche de $k.f$ et ayant une amplitude proche de l'enveloppe du spectre. Cette dernière condition sert à éliminer les pics pouvant être causés par du bruit. Les auteurs supposent également, mais sans le dire explicitement, que si pour une harmonique $k$, le pic correspondant le plus vraisemblable a une probabilité plus faible que la probabilité d'être absent, alors l'harmonique est comptée comme manquante.

Cette simplification se justifie par le fait que la loi de probabilité, pour être efficace, va décroître très vite lorsque la fréquence du pic s'écarte de la fréquence hypothèse $k.p$. Etant donné que la résolution de la transformée de Fourier dans les systèmes de traitement de son usuels est de quelques dizaines de Hertz, on est sûr que si l'harmonique n'est pas manquante, alors le pic le plus proche correspond à cette harmonique.

On note $\tilde{q}_k = (\tilde{A}_k, \tilde{f}_k) \in \tilde{Q}$ le pic correspondant à l'harmonique $k$. $\{K\}$ est l'ensemble des indices des harmoniques présentes et $\{M\}$ celui des harmoniques

manquantes. En utilisant les notations définies dans la section précédente la log-vraisemblance s'exprime ainsi :

$$\begin{aligned}\log(L(p)) = &\sum_{k\in\{K\}} \log(P(E_k)) + \log(P(\tilde{Q}_k - \{\tilde{q}_k\})) + \log(g(\frac{\tilde{f}_k}{p} - k)) + \log(h(\tilde{A}_k)) \\ &+ \sum_{k\in\{M\}} \log(P(\bar{E}_k)) + \log(P(\tilde{Q}_k))\end{aligned} \tag{7.2.7}$$

$P(E_k)$ (resp. $P(\bar{E}_k)$) est la probabilité de présence (resp. d'absence) de l'harmonique numéro $k$. $P(\tilde{Q}_k - \{\tilde{q}_k\})$ est la probabilité d'observer des pics en plus de l'harmonique $k$, $P(\tilde{Q}_k)$ représente la même chose dans le cas où l'harmonique est manquante. Les auteurs définissent cette probabilité comme étant tout simplement le nombre de pics supplémentaires, c'est à dire $\mathrm{Card}(\tilde{Q}_k - \{\tilde{q}_k\})$ dans le cas où l'harmonique est présente et $\mathrm{Card}(\tilde{Q}_k)$ quand elle est manquante. Enfin $g$ est la distribution de probabilité de l'harmonique $k$ autour de la fréquence $k.p$ et $h$ est la distribution de probabilité de l'amplitude. Les auteurs suggèrent de choisir une densité de probabilité Gaussienne pour $g$. Pour évaluer la probabilité de l'amplitude $h$, ils proposent de considérer la distribution de l'amplitude autour de l'enveloppe du spectre, avec encore une fois une densité de probabilité Gaussienne [Doval and Rodet, 1993].

En supposant en plus que $P(E_k) = P(E)$ est identique pour tout $k$, la mesure de probabilité concrètement utilisée est donc :

$$\begin{aligned}\log(L(p)) = &\sum_{k\in\{K\}} \log(\mathrm{Card}(\tilde{Q}_k) - 1) - \frac{(\tilde{f}_k - k.p)^2}{2\sigma_f^2 p^2} - \frac{(\tilde{A}_k - A_e(k.p))^2}{2\sigma_a^2} \\ &+ \sum_{k\in\{M\}} \log(\mathrm{Card}(\tilde{Q}_k)) \\ &+ \mathrm{Card}(\{K\})C_1 + \mathrm{Card}(\{\bar{K}\})C_2\end{aligned} \tag{7.2.8}$$

où $\sigma_a$ et $\sigma_f$ sont les variances de $h$ et $g$. $\sigma_a$ est supposé indépendante de la fréquence. $A_e(f)$ est la valeur de l'enveloppe du spectre pour la fréquence $f$. $C_1$ et $C_2$ sont deux constantes :

$$\begin{aligned}C_1 &= \log(P(E)) - \log(\sigma_a) - \log(\sigma_f) - \log(2\pi) \\ C_2 &= \log(1 - P(E))\end{aligned}$$

L'équation possède donc en tout trois paramètres à régler : $P(E)$, $\sigma_a$ et $\sigma_f$.

La valeur de pitch retenue est celle qui maximise l'équation (7.2.8), sur l'intervalle de pitch possible $[Fmin, Fmax]$. Le calcul est réalisé en deux étapes. Tout d'abord la fonction est évaluée grossièrement pour des valeurs régulièrement espacées, ce qui nous donne un histogramme de la fonction $L$. On sélectionne le maximum et on l'affine en recalculant la fonction dans son voisinage.

Dans [Doval and Rodet, 1993], les auteurs proposent un algorithme de programmation dynamique similaire à celui utilisé dans la méthode RAPT décrite au paragraphe 7.2.1.1. Les états et les probabilités correspondantes sont donnés par l'histogramme de $L$. L'algorithme de Viterbi [Rabiner and Juang, 1993] est utilisé pour

trouver le meilleur chemin parmi tous ces états. Les probabilités de transition entre les états sont modélisées par une densité Gaussienne sur l'écart fréquentiel entre deux états, ce qui rajoute un paramètre à l'algorithme (la variance). Contrairement au RAPT, les auteurs ne font pas intervenir d'état non pitché, ce qui laisse des interrogations sur le comportement de l'algorithme dans ces zones.

**Discussion**

Dans leur article, Doval et Rodet soulignent l'importance de l'association pic à pic pour avoir une mesure de vraisemblance qui soit rapide tout en restant robuste. Nous avons d'ailleurs insisté sur la justification en pratique de leur méthode d'association. La méthode aurait donc pu être présentée plus simplement après la phase d'identification des pics, c'est à dire en considérant que les pics hypothèses sont déjà associés aux pics estimés.

A priori le nombre de pics additifs peut être complètement arbitraire, sauf si on fait l'hypothèse qu'il n'y a aucun pic de bruit supplémentaire, ce qui n'est pas le cas. La probabilité des pics supplémentaires devrait donc être uniforme et n'apporte pas d'information supplémentaire à la mesure de vraisemblance.

Ensuite, les auteurs ont souligné que dans le cas de la parole la probabilité de présence des pics $P(E_k)$ était très proche de 1 pour tout $k$. Ils ajoutent que cette probabilité peut servir dans certains cas particuliers d'instruments de musique, qui présentent des spectres bien particulier, avec des harmoniques systématiquement manquantes. Cependant, dans le cas général, cette probabilité n'est pas très informative. De plus si le signal est bruité ou filtré, certaines harmoniques peuvent disparaître de façon arbitraire, rendant encore moins intéressante l'introduction de cette probabilité. Une information similaire, mais plus générale et plus robuste, serait le nombre d'harmoniques présentes parmi les $K$ harmoniques de l'hypothèse. Plus le nombre d'harmoniques présentes est important, plus le peigne hypothèse sera vraisemblable.

Enfin, l'algorithme de suivi de pitch ne fait pas intervenir l'état de non-voisement (ou non-harmonicité) comme dans l'algorithme RAPT. L'algorithme ne permet donc pas de définir les zones non-harmoniques et on peut légitimement se demander si les pics de bruit des zones non-harmoniques ne peuvent pas perturber l'optimalité du chemin trouvé par l'algorithme de Viterbi.

#### 7.2.2.2 La méthode de Maher

Le méthode de Maher et Beauchamp [Maher and Beauchamp, 1994] se base sur le même principe que la méthode précédente, décrite sur la Figure 7.5. Les différences se font sur la façon dont les pics sont associés et sur la mesure de correspondance entre le modèle et les pics estimés. Les auteurs ne proposent pas de post-traitement, mais l'algorithme de programmation dynamique décrit par Doval et Rodet [Doval and Rodet, 1993] peut par exemple être utilisé.

L'identification des pics est réalisé ici en deux étapes. On calcule pour chaque harmonique du pitch hypothèse le pic estimé le plus proche, comme dans la méthode

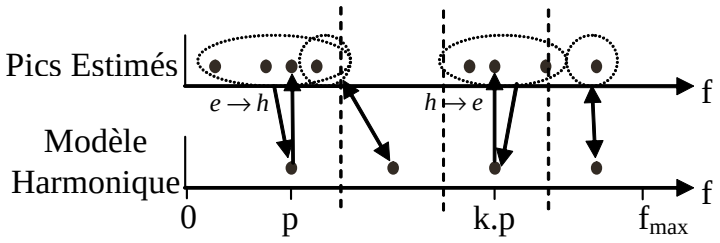



FIG. 7.7: Identification des fréquences dans la méthode de Maher et Beauchamp

de Doval et Rodet. On note cette première association[3] $h \rightarrow e$. Ensuite on cherche l'information symétrique, notée $e \rightarrow h$ : pour chaque pic estimé, on associe le pic hypothèse le plus proche.

La première association nous donne un ensemble de couple $C_{h\rightarrow e} = \{(q_k, \tilde{q}_k)\}_{1..K}$. $\tilde{q}_k = (\tilde{A}_k, \tilde{f}_k)$ appartient à l'ensemble des pics estimés $\tilde{Q}$ et $q_k = (A_k, f_k)$ à l'ensemble des pics hypothèses. Pour l'harmonique $q_k$, l'amplitude hypothèse choisie est celle du pic estimé associé, $A_k = \tilde{A}_k$, et la fréquence est un multiple de la fréquence fondamentale $p$ : $f_k = k.p$. La deuxième association nous donne l'ensemble : $C_{e\rightarrow h} = \{(\tilde{q}_k, q_k)\}_{1..N}$. Dans le cas général, ces deux ensembles de couples de pics sont différents (voir Figure 7.7). En appliquant la fonction d'erreur sur les deux ensembles, on obtient deux mesures complémentaires que l'on va combiner pour avoir une erreur totale. La mesure d'erreur est donnée par :

$$Err(C_{h\rightarrow e}) = \sum_{k=1}^{\text{Card}\, C_{h\rightarrow e}} |\tilde{f}_k - f_k|(f_k)^{-p_1} + \frac{A_k}{A_m}.\left(p_2|\tilde{f}_k - f_k|(f_k)^{-p_1} - p_3\right) \quad (7.2.9)$$

où $p_1$, $p_2$, $p_3$ sont trois paramètres à régler et $A_m$ est l'amplitude maximale parmi l'ensemble des pics estimés $\tilde{Q}$. Pour obtenir $Err(C_{e\rightarrow h})$, il suffit d'intervertir $\tilde{f}_k$ et $f_k$, $\tilde{A}_k$ et $A_k$. Enfin on combine ces deux mesures ainsi :

$$Err(f) = Err(C_{h\rightarrow e}) + p_4 Err(C_{e\rightarrow h}) \quad (7.2.10)$$

où $p_4$ est un paramètre de plus à régler.

Enfin, comme dans la méthode précédente, le minimum de cette fonction est calculé en deux étapes : un histogramme grossier d'abord, puis un calcul plus précis autour du minimum de l'histogramme.

### Discussion

L'association à double sens semble être une implantation plus précise de la vraisemblance totale $P(\tilde{Q}_k)$ de l'équation (7.2.7), car au lieu de retenir le meilleur pic, elle va dans certains cas associer plusieurs pics à une harmonique hypothèse (cf Figure 7.7). Ici les auteurs ne font pas intervenir de partition des fréquences comme

[3] $h \rightarrow e$ signifie pic du peigne Hypothèse vers pic Estimé.

Doval et Rodet. Il en résulte que parfois une harmonique hypothèse va être associée à un pic qui sera très proche d'une autre harmonique (Figure 7.7), ce qui d'une part n'est pas logique et d'autre part va causer des problèmes de normalisation pour le cas où il y a des harmoniques manquantes. En fait cette association à double sens est une façon de palier à l'absence de concept d'harmonique manquante. Mais au lieu d'une pénalisation fixe pour une harmonique manquante, on aura en fait une pénalité aléatoire qui dépend du pitch hypothèse et du pic estimé le plus proche.

D'un point de vue calculatoire, la méthode de calcul en deux étapes présente des avantages et des inconvénients. Le fait de ne pas faire intervenir l'enveloppe de l'amplitude pour associer les pics, comme dans la méthode Doval et Rodet, va accélérer le processus pour un pic, mais comme l'association est faite dans les deux sens, il y aura plus du double de pics à associer. Donc globalement il n'y a pas non plus d'amélioration sur la vitesse qui justifierait un tel procédé.

## 7.3 Algorithme de suivi de pitch basé sur la modélisation sinusoïdale (SPS)

Le schéma de l'algorithme est le même que pour les méthodes du paragraphe 7.2.2, et visible sur la Figure 7.5. Nous allons chercher à améliorer chaque partie spécifique, c'est à dire le générateur d'hypothèses, l'identification des pics, la mesure de vraisemblance et la méthode de suivi, afin d'avoir une méthode au moins aussi robuste que les méthodes existantes, tout en étant plus simple et plus rapide.

### 7.3.1 Génération d'hypothèses

La première étape d'un système d'estimation de pitch basé sur les pics sinusoïdaux est de générer un certain nombre d'hypothèses de pitch possible à partir des pics sinusoïdaux estimés. Pour la plupart des algorithmes de la littérature, et en particulier les algorithmes décrits dans les sections précédentes, cette étape consiste à définir un intervalle de recherche $[F_{min}, F_{max}]$, et à l'échantillonner de façon régulière.

L'article de P. Cano [Cano, 1998] propose une solution alternative intéressante. Il s'agit de considérer comme hypothèses possibles $p$ les fréquences des pics sinusoïdaux $\tilde{f}$ divisées par des valeurs entières simples :

$$p = \tilde{f}/m, \qquad m \in \mathbb{N} \tag{7.3.1}$$

Les facteurs entiers sont limités à $m \leq 4$, pour que le nombre d'hypothèses reste faible.

Le pitch, par définition, est un sous multiple des fréquences des pics des harmoniques. A condition que le facteur entier soit suffisamment élevé, le pitch sera donc forcément contenu dans ces hypothèses. Si on limite les facteurs possibles à $m \leq 4$, une bonne hypothèse de pitch sera formulée à condition qu'une des quatre premières harmoniques soit présente, ce qui est quasiment toujours vérifié, en particulier dans le cas de la parole.

Dans le cas des signaux parfaitement harmoniques, utiliser une harmonique supérieure pour faire une hypothèse de pitch, permet de réduire l'erreur sur le pitch. En effet la variance du bruit de l'estimation sinusoïdale ne dépend pas de la fréquence, mais seulement du SNR. Prenons l'exemple d'une harmonique $f_k$, estimée avec une erreur $e_k$ : $\tilde{f}_k = f_k + e_k$. Si $m = k$, l'hypothèse de pitch $\tilde{p}$ s'exprime alors ainsi :

$$\tilde{p} = \frac{\tilde{f}_k}{k} = \frac{k.p + e_k}{k} = p + \frac{e_k}{k} \tag{7.3.2}$$

$p$ est la vraie valeur du pitch. Dans le cas où les harmoniques présentent des perturbations $e_k$ de magnitude équivalente pour tout $k$, on voit clairement que plus le numéro de l'harmonique est élevé, plus l'influence de l'erreur sera réduite.

L'autre avantage de cette méthode est de réduire fortement le nombre d'hypothèses possibles, ce qui permet d'éviter une estimation en deux étapes, une étape grossière et une étape fine, comme dans le cas de la plupart de algorithmes de la littérature.

La proposition de l'article [Cano, 1998] peut néanmoins être améliorée en terme de complexité. En effet en considérant les quatre sous harmoniques de chaque pic, on génère une certaine redondance dans les hypothèses générées, et des hypothèses superflues dues à des pics de bruit vont se glisser. On rappelle que le but de l'algorithme est d'estimer le pitch prédominant. A condition que le signal ne soit pas noyé dans le bruit, la composante sinusoïdale la plus forte devrait être une harmonique du pitch prédominant. Donc la bonne hypothèse de pitch sera contenue dans l'ensemble des sous-harmoniques du pic de plus grande amplitude, ou des $M$ composantes de plus forte amplitude. L'algorithme proposé repose donc sur les deux étapes suivantes :

1. Parmi l'ensemble des pics estimés $\{\tilde{f}\}$, sélectionner les $M$ composantes de plus forte amplitude
2. Calculer toutes les sous-harmoniques de ces composantes contenues dans l'intervalle d'analyse $[F_{min}, F_{max}]$.

### 7.3.2 Identification harmonique

Une fois les hypothèses de pitch générées, il faut calculer la probabilité du peigne harmonique correspondant. Pour cela, nous allons, comme pour les méthodes de Maher et de Doval, identifier les harmoniques, c'est-à-dire faire correspondre à chaque harmonique $kp$ du peigne hypothèse un des pics estimés $\tilde{f}$. La méthode choisie est proche de l'algorithme de Doval et Rodet (Figure 7.6). On va associer à chaque harmonique le pic le plus proche fréquentiellement.

Pour tenir compte de la possibilité que l'harmonique soit manquante, on va comme dans le cas de Doval limiter l'intervalle de fréquence autour de la fréquence de l'harmonique hypothèse, avec une certaine tolérance $T_f$. Cette tolérance dépend de la qualité de l'estimation des pics sinusoïdaux. On a vu dans la première partie que la qualité d'estimation dépendait du rapport signal à bruit de la composante sinusoïdale considérée et de la taille de la transformée de Fourier. Comme le SNR de chaque pic n'est pas forcément connu, on va considérer une tolérance suffisamment large pour prendre en compte la plupart des cas de figure attendus.

Pour résumer, la partie "identification" de l'algorithme proposé est la suivante :

- Pour chaque hypothèse de pitch $p$
  - Pour chaque harmonique $f_k = k.p$, $k \in [1, K_{max}]$
  - Identifier le pic estimé $\{\tilde{A}, \tilde{f}\}$ le plus proche fréquentiellement
  - Si $|\tilde{f} - k.p| \leq T_f$ alors l'harmonique est présente et ses paramètres sont $\tilde{A}$ et $\tilde{f}$.
  - Sinon l'harmonique est manquante

### 7.3.3 Mesure d'harmonicité

Une fois que les pics estimés ont été identifiés, on va calculer la vraisemblance de cet ensemble de pics par rapport à l'hypothèse de pitch considérée $p$, que nous appelleront mesure d'harmonicité. Cet ensemble est caractérisé par le nombre total d'harmoniques $K$, $K \leq K_{max}$, par le nombre d'harmoniques manquantes $M$ parmi ces $K$ harmoniques, et par les paramètres des harmoniques présentes $\{\tilde{f}_k, \tilde{A}_k\}_{k \in \{P\}}$.

La mesure d'harmonicité doit permettre d'éviter les doublements et les demi-pitchs. Elle doit être robuste au bruit et doit pouvoir indiquer le voisement ou la présence d'instruments. Pour que ces propriétés soient vérifiées, il faut que la mesure respecte les règles suivantes :

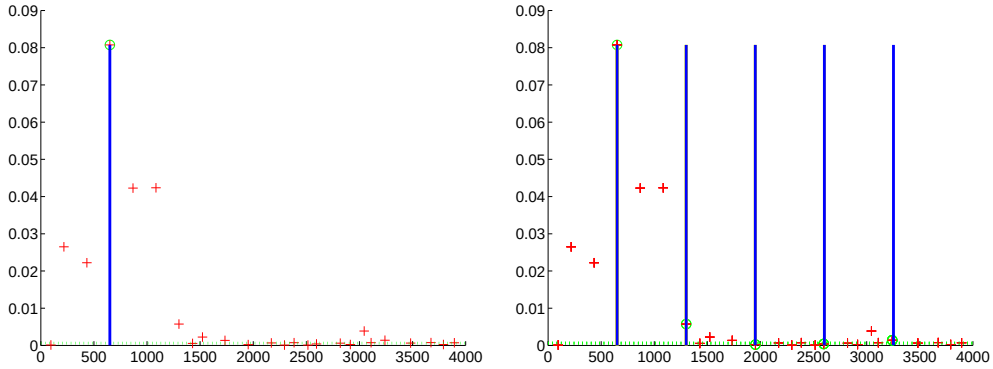
- Donner plus de poids aux harmoniques de plus grande amplitude car leur estimation de fréquence est plus fiable ($R_1$)
- Donner plus de poids lorsque le nombre d'harmoniques $K$ est élevé ($R_2$)
- Pénaliser les erreurs d'estimation de fréquence et les in-harmonicités ($R_3$)
- Pénaliser les harmoniques manquantes ($R_4$)

Contrairement à Doval et Rodet, nous allons considérer que les probabilités de présence des harmoniques et des pics supplémentaires sont non informatives (cf section 7.2.2.1). Ces propriétés vont nous guider pour définir les densités de probabilités utilisées pour modéliser le problème.

Deux mesures harmoniques compatibles avec l'identification harmonique définie dans la section précédente ont été comparées. La première est une adaptation de la mesure de Maher et Beauchamp qui tient compte des harmoniques manquantes. La deuxième mesure est une simplification de la mesure de Doval et Rodet, implantant les règles données dans le paragraphe précédent. La Figure 7.8 donne un aperçu du fonctionnement complet du mécanisme combinant la génération d'hypothèses, l'identification de pics et la mesure de vraisemblance.

#### 7.3.3.1 Modification de la mesure de Maher et Beauchamp

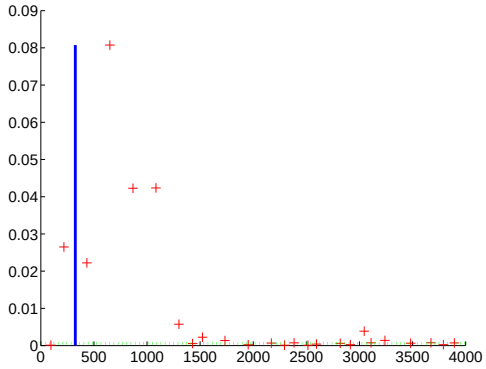
Comme nous l'avons suggéré dans la section 7.2.2.2, la procédure d'identification à double sens est une façon de tenir compte des harmoniques manquantes, mais qui va causer des problèmes de normalisation dans la mesure de similarité. Nous proposons donc d'utiliser la mesure de Maher dans le cadre de l'identification harmonique de la



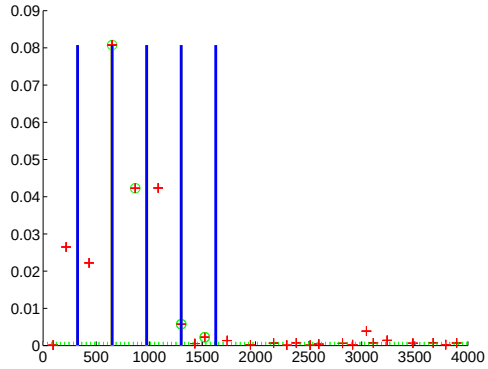
(a) $p = \tilde{f}_m$, $K = 1$, $l = 0.99$



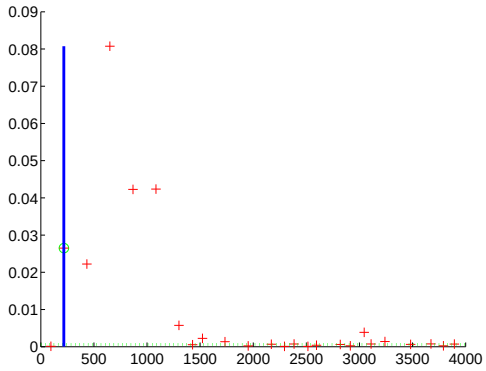
(b) $p = \tilde{f}_m$, $K = 5$, $l = 0.86$



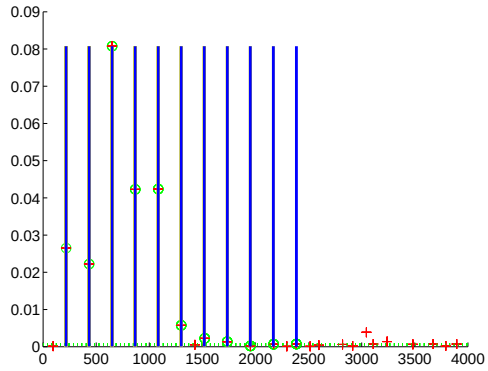
(c) $p = \tilde{f}_m/2$, $K = 1$, $l = -0.5$



(d) $p = \tilde{f}_m/2$, $K = 5$, $l = -0.44$



(e) $p = \tilde{f}_m/3$, $K = 1$, $l = 0.32$



(f) $p = \tilde{f}_m/3$, $K = 11$, $l = 2.67$

FIG. 7.8: Fonctionnement de l'estimation du meilleur pitch. Les croix indiquent les positions fréquentielles des pics, les barres verticales les harmoniques hypothèses et les ronds les pics qui ont été identifiés. Le niveau de bruit estimé est indiqué par des pointillés. On teste toutes les sous-harmoniques du pic maximal $\tilde{f}_m$ pour tous les nombres d'harmoniques $K$ possibles. La vraisemblance $l$ maximale est l'hypothèse $\tilde{f}_m/3$ avec 11 harmoniques.

section 7.3.2[4] en lui ajoutant un terme de pénalisation des harmoniques manquantes :

$$L(p) = \sum_{k\in\{P\}} -|\tilde{f}_k - f_k|(f_k)^{-p_1} - \frac{A_k}{A_m}\Big(p_2|\tilde{f}_k - f_k|(f_k)^{-p_1} - p_3\Big) - T_f \sum_{k\in\{M\}} (f_k)^{-p_1} \tag{7.3.3}$$

où $\{P\}$ est l'ensemble des indices des harmoniques présentes et $\{M\}$ l'ensemble des indices des harmoniques manquantes. $T_f$ est toujours la tolérance fréquentielle lors de l'identification des pics. En fait l'erreur des pics absents correspond à la plus forte erreur possible pour un pic présent c'est à dire pour un pic d'amplitude nulle et d'erreur fréquentielle égale à $T_f$. Ainsi l'erreur d'un pic absent est bornée et reste la même d'une trame à l'autre. Par rapport à l'équation (7.2.9), le signe a été également inversé afin d'avoir une mesure de qualité comme dans le cas de Doval et Rodet et non plus une mesure d'erreur.

On peut constater que pour cette mesure les règles $R_{1-4}$ énoncées dans la section précédente sont respectées.

#### 7.3.3.2 Mesure probabiliste simplifiée

On va chercher à maximiser la probabilité à posteriori des paramètres du système $M$, $K$, $\{\tilde{f}_k, \tilde{A}_k\}_{k\in\{P\}}$ par rapport à la valeur du pitch hypothèse $p$. La méthode est similaire à celle de Doval et Rodet, la vraisemblance du système étant ici évaluée après la phase d'identification des pics. En considérant la log vraisemblance, on ne change pas le problème, ce qui nous donne comme mesure :

$$L(p) = \log\big(P(K, M, \{\tilde{f}_k, \tilde{A}_k\}_{k\in\{P\}}|p)\big)$$

En considérant la probabilité à posteriori et non à priori, on fait l'hypothèse implicite que les probabilités des différentes valeurs de pitch et des perturbations d'amplitude suivent une loi uniforme.

On fait ensuite les hypothèses simplificatrices suivantes, déjà évoquées à la section 7.2.2.1 :

- la valeur de chacune des amplitudes est indépendante des autres paramètres.
- la probabilité de chacune des fréquences ne dépend que de $p$.
- le nombre d'harmoniques total $K$ et le nombre d'harmoniques manquantes $N$ sont indépendants de tous les autres paramètres.

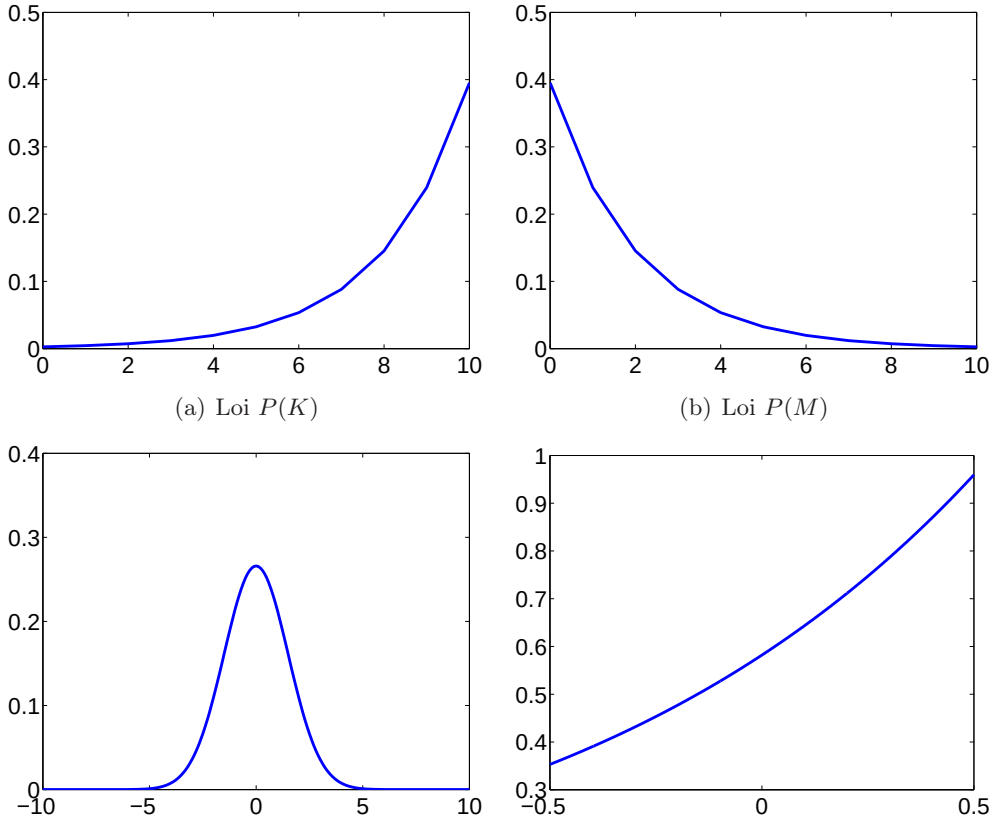
La mesure de vraisemblance s'écrit alors :

$$L(p) = \sum_{k\in\{P\}} \log(P(\tilde{A}_k)) + \log(P(\tilde{f}_k)|p) + \log(P(K)) + \log(P(M))$$

Pour respecter les règles $R_{1-4}$ énoncées dans le paragraphe 7.3.3, on choisit la modélisation suivante :

- $P(\tilde{A}_k) = \frac{1}{C_1}\,\mathrm{e}^{\frac{\tilde{A}_k - \tilde{A}_b}{\tilde{A}_m}}$ où $\tilde{A}_m$ est l'amplitude du pic maximal et $\tilde{A}_b$ est l'amplitude du bruit. L'amplitude normalisée $A' = \frac{\tilde{A}_k}{\tilde{A}_m}$ suit une loi uniforme sur $[0, 1]$. Enfin $C_1$ est une constante de normalisation : $C_1 = \mathrm{e}^{-\tilde{A}_b/\tilde{A}_m}(e-1)$.

[4] C'est à dire juste l'erreur harmonique vers pic, noté $Err_(C_{h\to p})$ dans la section 7.2.2.2.




(a) Loi $P(K)$ (b) Loi $P(M)$

(c) Loi $P(\tilde{f}_k|p)$ (d) Loi $P(\tilde{A}_k)$

FIG. 7.9: Exemple de modélisation typique pour les différents paramètres. Pour $P(\tilde{A}_k)$ la loi dessinée correspondant à $\tilde{A}_b = 0.5$ et $\tilde{A}_m = 1$. Ici $K_{max} = 10$ donc $P, K \in [0..10]$

- $P(\tilde{f}_k|f) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(\tilde{f}_k - k.p)^2}{2\sigma^2}}$ loi Gaussienne dont la variance $\sigma$ est directement liée à la variance de l'estimateur de fréquence.
- $P(M) = \frac{1}{C_2} \mathrm{e}^{-\lambda_1.M}$, loi exponentielle discrète avec $\lambda_1 > 0$, $C_2$ constante de normalisation. $M$ est limité à $[0, K_{max}]$ pour la normalisation.
- $P(K) = \frac{1}{C_3} \mathrm{e}^{\lambda_2.K}$ loi exponentielle discrète avec $\lambda_2 > 0$. $K$ est également limité à $[0, K_{max}]$.

Les densités de probabilités des différentes lois sont dessinées sur la Figure 7.9. Comme dans le cas de la mesure de Maher, l'amplitude est normalisée par l'amplitude du pic maximum $\tilde{A}_m$ pour la trame considérée. Afin d'être plus robuste au bruit, seuls les pics dont l'amplitude est supérieur au niveau de bruit doivent augmenter la vraisemblance : l'amplitude du bruit $\tilde{A}_b$, considéré ici blanc Gaussien est soustrait à l'amplitude des pics. Nous avons choisi une méthode simple pour évaluer l'amplitude du bruit en prenant l'amplitude médiane des pics. En effet lors du calcul des pics, si tous les pics du spectre sont calculés sans filtrage, une majorité d'entre eux seront dus au bruit. Ce mécanisme est illustré sur la Figure 7.8.

La mesure de vraisemblance s'écrit alors ainsi :

$$L(p) = \Big[ \sum_{k\in\{P\}} \frac{\tilde{A}_k - \tilde{A}_b}{\tilde{A}_m} - \frac{(\tilde{f}_k - k.p)^2}{2\sigma^2} \Big] - (K-M).C_4 - \lambda_1.M + \lambda_2 K + C_5$$

$$\text{où } C_4 = -\log(C_1) - \log(\sqrt{2\pi}\sigma)$$
$$C_5 = -\log(C_2) - \log(C_3)$$

On peut alors remarquer que $C_5$ est une constante indépendante du pitch hypothèse $p$ qui peut être supprimée sans changer le problème. Pour réduire le nombre de paramètres de la mesure, on peut choisir $\lambda_2 = C_4$. En effet comme $e-1 > C_1 > [1-\frac{1}{e}]$ et qu'en pratique $\sigma > 1$, $C_4$ sera une constante positive comme $\lambda_2$. Enfin on pose $\alpha_1 = \frac{1}{2\sigma^2}$ et $\alpha_2 = \lambda_1 - C_4$. La mesure de vraisemblance simplifiée s'écrit alors :

$$L(p) = \Big[ \sum_{k\in\{P\}} \frac{\tilde{A}_k - \tilde{A}_b}{\tilde{A}_m} - \alpha_1(\tilde{f}_k - k.p)^2 \Big] - \alpha_2.M$$

Etant donné que les lois de probabilité ont été choisies de façon à vérifier les règles $R_{1-4}$ énoncées dans le paragraphe 7.3.3, cette mesure va bien entendu les vérifier. On peut noter qu'elle est plus simple que la mesure de Maher modifiée, et qu'elle ne fait intervenir que deux paramètres contre trois pour la mesure de Maher. Elle sera donc plus facile à régler.

On a fait remarquer que la variance $\sigma$ sur l'erreur de fréquence est directement liée à la variance de l'estimateur de fréquence. On a vu dans la première partie de la thèse que cette variance dépend de la résolution de la TF, c'est à dire de la taille de la fenêtre d'analyse $t_N = N/F$ et du SNR[5]. Une fois que la longueur $t_N$ de la fenêtre d'analyse est fixée, et en réglant le système pour le cas du SNR attendu le plus défavorable, la constante $\alpha_1$ reste identique si on fait varier $N$ ou $F$.

Une harmonique sera considérée comme absente si l'erreur fréquentiel est inférieure au poids de pénalisation des harmoniques manquantes $\alpha_2$. Cela nous donne la tolérance fréquentielle effective $T_f$ suivante :

$$T_f = \sqrt{\frac{\alpha_2}{\alpha_1}} \tag{7.3.4}$$

Lorsqu'une harmonique est absente on a fait ici l'hypothèse que son amplitude est égale à celle du bruit $A_b$.

### 7.3.4 Suivi de fréquence fondamentale

La dernière partie de l'algorithme, le suivi de fréquence fondamentale, est facultative. Dans le cas où il n'est pas utilisé, l'hypothèse de pitch retenue est celle ayant la mesure de vraisemblance la plus élevée. L'intérêt du suivi de pitch est de pouvoir

[5]En effet le tableau 3.1 montre que le CRB pour la fréquence s'écrit $C.{t_N}^{-3}.\eta^{-1}$. Dans le cas du modèle **M01**, la variance de la plupart des estimateurs a une forme similaire, comme on l'a vu par exemple pour l'estimateur de la section 6.1.1.1.

éliminer les derniers cas de demi-pitch, car le demi-pitch peut parfois avoir une vraisemblance plus forte avec le type de mesures décrites dans le paragraphe précédant. Il permet également de délimiter plus facilement les zones pitchées, et offre une plus grande robustesse aux bruits.

Dans la littérature, on trouve essentiellement deux types de suivi de partiel. Le premier utilise la programmation dynamique, comme par exemple l'algorithme de Viterbi. Deux méthodes de ce type ont déjà été décrites, dans le cas de l'algorithme de Doval et Rodet [Doval and Rodet, 1993], et de l'algorithme RAPT [Talkin, 1995]. Le principal désavantage de ces deux algorithmes est de demander une analyse globale sur le signal analysé, ou tout du moins une analyse sur une fenêtre temporelle large, rendant difficile les applications temps-réel. Pour définir les zones voisées, comme dans RAPT, ces algorithmes doivent faire intervenir un état supplémentaire difficile à modéliser pour décrire l'hypothèse de non-harmonicité. Le deuxième type de suivi est un "guide d'onde". Cet algorithme est souvent utilisé pour suivre des sinusoïdes seules, mais il peut également être utilisé pour suivre des peignes de sinusoïdes (i.e. le pitch). Son principe est de relier les pics sinusoïdaux consécutifs les plus proches fréquentiellement, à l'intérieur d'une certaine tolérance fréquentielle [McAulay and Quatieri, 1986], [Serra, 1989]. Il fait intervenir les concepts (ou états) de naissance, vie et mort de la sinusoïde. L'avantage de cet algorithme est sa simplicité, et la possibilité de délimiter les zones harmoniques.

Nous proposons ici un algorithme qui combine les avantages des deux méthodes. Il s'agit d'un algorithme de programmation dynamique, qui va chercher le meilleur chemin à travers les hypothèses de pitch, en utilisant à la fois une accumulation des valeurs des vraisemblances de chaque hypothèse mais également le principe du guide d'onde, c'est à dire en liant les pics les plus proches fréquentiellement et en utilisant les états naissance, vie et mort. Pour permettre des applications en temps réel, un mécanisme pour forcer une décision au bout d'un temps fixe a également été mis en place. Cet algorithme peut être vu comme un algorithme de type Viterbi à faisceau (on limite les chemins possibles à un certain intervalle fréquentiel), et fenêtré (la décision est forcée).

Le fonctionnement de l'algorithme est décrit sur la Figure 7.10. L'élément de base de l'algorithme est un "noeud" (représenté par un point sur la Figure 7.10). Un noeud est caractérisé par :

- une hypothèse de pitch.
- un lien avec un et un seul noeud de la trame précédente.
- un état naissance (B), vie (L), mort (D) ou solitaire (S).
- une vraisemblance "locale" $l$ qui correspondant à la mesure d'harmonicité de l'hypothèse de pitch.
- une vraisemblance "cumulée" $l_n$ depuis l'origine du chemin. Cette origine peut être trouvé en remontant de proche en proche les liens avec les trames précédentes.

La trame entrante, contient des noeuds non connectés qui sont tous mis à l'état S (solitaire). Le buffer d'hypothèses de pitch est une table contenant des noeuds organisés en trames. Ce buffer contient un nombre fixe de trames ($k-1$ sur la Figure 7.10), et un nombre variable de noeuds par trame. Grâce au lien qui relie un

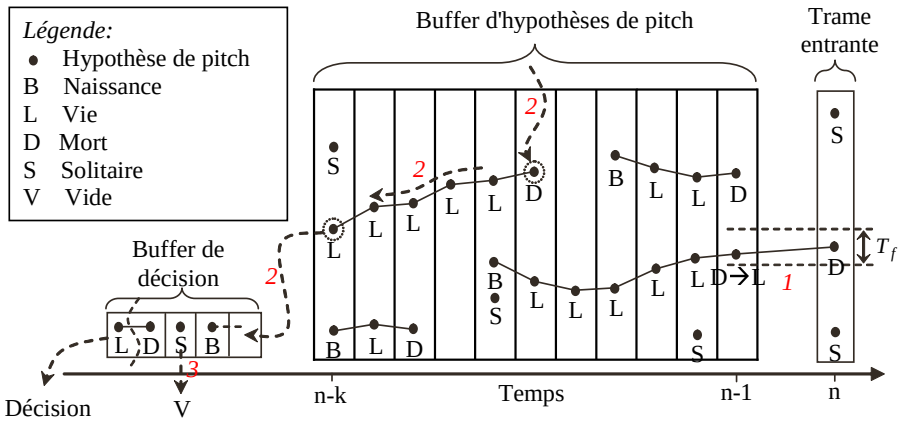



FIG. 7.10: Fonctionnement de l'algorithme de suivi de pitch

noeud à un autre noeud de la trame précédente, les noeuds sont également organisés en "segments". Un segment est soit un noeud de type S (segment de longueur 1), soit une succession de noeud commençant par un état B et finissant par un état D (segment de longueur $> 1$). Entre le noeud B et D, il peut y avoir un nombre quelconque de noeuds L. Enfin, le buffer de décision contient le meilleur chemin trouvé par l'algorithme. Il contient un nombre $E$ de noeuds, certains noeuds pouvant être vides si par exemple une trame du buffer d'hypothèse ne contenait aucun noeud. Ces noeuds sont organisés en segments de type naissance-vie-mort, comme dans le buffer d'hypothèses de pitch. Le buffer de décision sert à éliminer les segments trop courts, de longueur $< E$. Sur l'exemple de la Figure 7.10, on est dans le cas où $E = 4$ et on élimine tous les segments de longueur inférieure ou égale à 3.

Les étapes 1 à 3 de l'algorithme s'effectuent de la façon suivante :

1. connexion de la trame entrante (indice $n$)
   - un point de $n$ est lié si sa vraisemblance $l$ est supérieure à un seuil $s_v$ et si la tolérance $T_f$ est respectée.
   - s'il y a plusieurs candidats possibles en $n-1$, on connecte avec celui dont la vraisemblance $l_{n-1}$ est la plus grande.
   - on change l'état du point connecté en $n-1$ en L (vie), et celui du point en $n$ en D (mort).
   - on accumule la vraisemblance : $l_n = l_{n-1} + l$.
2. sélection du meilleur candidat en $n-k$
   - sélection du noeud le plus vraisemblable parmi tous les noeuds D et S (noeuds finaux) qui sont connectés à la trame $n-k$.
   - à partir du noeud final le plus vraisemblable, on remonte le chemin correspondant jusqu'à $n-k$ (début du buffer d'hypothèse).
   - le noeud trouvé en $n-k$ est l'hypothèse de pitch la plus vraisemblable pour la trame $n-k$. Ce noeud est ajouté au buffer de décision et la trame $n-k$ est supprimée.

– si le noeud trouvé en $n-k$ est connecté avec le noeud précédent du buffer de décision, *i.e.* le noeud trouvé en $n-k-1$, le noeud est ajouté tel quel (cas de la Figure 7.10). Sinon le noeud $n-k$ est changé en B. Le noeud en $n-k-1$ est changé en D si sa valeur était L et S si sa valeur était B.

3. suppression des segments trop courts et décision finale pour la trame $n-k-E$ :
   – on retourne la décision finale pour la trame $n-k-E$, c'est-à-dire le premier noeud du buffer de décision, et on supprime ce noeud dans le buffer de décision.
   – si un segment complet (naissance-vie-mort) de longueur inférieure stricte à $E$ est contenu dans le buffer de décision, il est supprimé et remplacé par des noeuds vides.

Contrairement à l'algorithme de MacAulay, ici on ne lie pas des pics sinusoïdaux, mais des hypothèses de pitch. L'algorithme fait également intervenir un état supplémentaire par rapport à celui de MacAulay, l'état solitaire (S). Il s'agit simplement du cas particulier ou l'hypothèse de pitch n'est pas liée, c'est à dire qu'il s'agit à la fois d'une mort et d'une naissance.

En éliminant les pics ayant une mesure d'harmonicité faible, et en filtrant les partiels de courte durée (par exemple inférieur à 30 ms), l'algorithme va également définir avec précision les zones harmoniques et les zones non-harmoniques car les chemins dus au bruit sont généralement très courts.

L'algorithme fait intervenir un certain nombre de paramètres : la taille du buffer d'hypothèse de pitch, la taille du buffer de décision, la valeur de la variation de fréquence $T_f$ maximale tolérée et le seuil $s_v$ sur la vraisemblance des hypothèses de pitch. Le choix de $T_f$ a déjà été discuté dans la partie sur l'estimation sinusoïdale dans la section 3.4. On sait que pour la plupart des signaux de parole et de musique, une valeur de variation de pitch maximale de 8000 Hz par seconde semble être un bon compromis. Les valeurs des autres paramètres ne sont pas critiques et peuvent être réglées grossièrement.

## 7.4 Comparaison avec des méthodes existantes

Dans cette partie nous allons évaluer la méthode d'estimation de pitch basée sur les pics sinusoïdaux, faisant intervenir les mesures d'harmonicité simplifiées décrites dans les paragraphes précédents. Pour cela nous allons appliquer la même démarche que Cheveigné dans l'article [de Cheveigné and Kawahara, 2002], et que nous redécrivons brièvement à la section 7.4.1. Dans la section 7.4.2, nous discuterons de la méthode d'estimation sinusoïdale à utiliser, car comme nous l'avons vu dans la première partie de la thèse, des améliorations sur l'estimation des paramètres des pics ont été développées depuis les travaux de Maher et Doval. La méthode sera ensuite comparée à la section 7.4.3 aux deux algorithmes état-de-l'art YIN et RAPT décrits précédemment. Les deux mesures d'harmonicité y seront également comparées et leurs avantages respectifs discutés.

### 7.4.1 Protocole d'évaluation

Le protocole de test est celui utilisé couramment pour comparer les algorithmes d'estimation de pitch prédominant. Il est très bien décrit dans un certain nombre d'ouvrages comme [Rabiner and Juang, 1993] ou [de Cheveigné and Kawahara, 2001]. Nous allons ici en rappeler les grandes lignes.

#### Corpus

Le principal problème pour évaluer les algorithmes d'estimation de pitch est d'avoir une vérité terrain réaliste et fiable. L'option retenue dans la plupart des articles récents sur le sujet est d'utiliser des bases de données de parole fournie avec un enregistrement de laryngographe. Les vérités terrains sont extraites de cet enregistrement, plus facile à traiter. Le principe du laryngographe et la méthode pour extraire la vérité terrain sont décrites brièvement dans le paragraphe suivant.

Il existe maintenant un grand nombre de bases de données avec laryngographe facilement disponibles sur internet. Quatre bases ont été retenues, similaires à celles utilisées dans [de Cheveigné and Kawahara, 2002] :

- timit mocha (460 phrases énoncées par 1 femme et 2 hommes)
- timit kdt (460 phrases énoncées par 1 homme)
- fda (50 phrases énoncées par 1 femme et 1 homme)[6]
- keele ( 30 secondes de parole pour 5 femmes et 5 hommes)[7]

Le temps total cumulé des enregistrements est de 2h30, complètement indexés grâce aux enregistrements de laryngographe. Dans les deux bases timit ce sont les mêmes phrases qui sont prononcées mais par des personnes différentes et avec des accents différents, ce qui fait qu'elles ne sont pas redondantes.

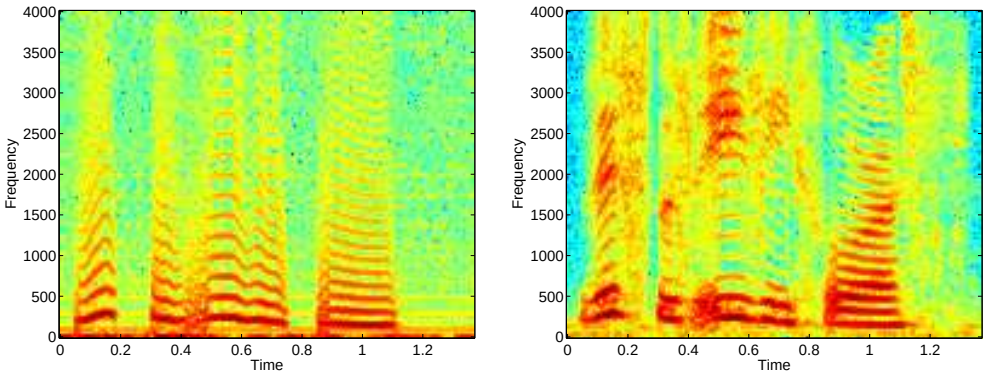
#### Laryngographe

Le laryngographe permet d'enregistrer le pitch directement à la sortie du larynx. Le principe est le suivant. Un faible courant est induit entre deux électrodes des deux cotés du larynx. L'ouverture et la fermeture de la glotte fait varier la conductivité du Larynx ce qui crée une modulation d'amplitude. Le signal est ensuite démodulé, amplifié et filtré passe-haut pour supprimer l'influence du déplacement vertical lent du larynx. Le spectre de ce signal n'a pas été modifié par les filtres successifs que sont le conduit vocal et la bouche. Il est donc généralement très simple, le pitch étant donné par l'harmonique qui est à la fois la plus basse et la plus forte (Voir l'exemple sur la Figure 7.11). Le bruit est généralement très faible si le laryngographe a été bien utilisé. Toutes les conditions sont réunies pour pouvoir très facilement estimer le pitch.

Les références de pitch sont générées à partir de ces enregistrements en utilisant une version modifiée d'un algorithme d'estimation de pitch, se contentant de suivre

[6] Créée par [Bagshaw et al., 1993], téléchargeable à l'adresse http ://www.cstr.ed.ac.uk/pcb/fda/eval.tar.gz

[7] Peut être téléchargée à l'adresse ftp ://ftp.cs.keele.ac.uk/pub/pitch/Speech. On doit écrire à l'auteur pour avoir le mot de passe.




(a) Laryngographe (b) Signal audio correspondant

FIG. 7.11: Exemple de spectre de laryngographe

l'harmonique la plus basse. L'algorithme utilisé est ici une version légèrement simplifiée de l'algorithme d'estimation de pitch sinusoïdal, avec le suivi de pitch pour délimiter les zones harmoniques. Les références ont ensuite été vérifiées manuellement et corrigées le cas échéant.

### Les mesures

Les mesures de performances utilisées sont les mêmes que celles décrites dans l'étude de Rabiner [Rabiner and Juang, 1993], et visent à évaluer quatre différents types d'erreur :

- l'erreur de pitch grossière (GER), c'est à dire les doublements de pitch et les demi pitch par exemple.
- l'erreur fine de pitch, c'est à dire les imprécisions sur le pitch mesuré (FER)
- le taux de rappel de la détection de l'harmonicité (RH)
- le taux de précision de l'harmonicité (PH)

L'erreur la plus déterminante est le GER, car ce sont les erreurs les plus graves. Une erreur est considérée comme grossière lorsque l'erreur relative dépasse un certain seuil. Le nombre d'erreurs grossières est normalisé par le nombre total d'estimations réalisées, qui est donné par le nombre de trames de référence classées comme harmoniques. Le seuil généralement utilisé est de 20%, ce qui permet de s'assurer que tous les doublements et demi-pitchs seront comptés dans cette mesure. Les erreurs de faux rejets de la mesure d'harmonicité sont aussi considérées comme des erreurs grossières. Les deux dernières mesures sont également importantes car elles vont compléter le GER. Il s'agit des erreurs de classement classiquement utilisées en théorie de la décision :

$$RH = \frac{N_t - N_{fr}}{N_t} \tag{7.4.1}$$

$$PH = \frac{N_t - N_{fr}}{N_t - N_{fr} + N_{fa}} \tag{7.4.2}$$

où $N_t$ est le nombre total de trames, $N_{fr}$ est le nombre de faux rejets, c'est-à-dire les trames classées non-harmoniques, alors qu'elles étaient harmoniques dans la référence, et $N_{fa}$ est le nombre de fausses alarmes, c'est-à-dire les trames classées harmoniques, alors qu'elles étaient non-harmoniques dans la référence. Le FER est quand à lui moins déterminant, surtout en considérant que la précision du laryngographe est inconnue. Il sera donc pas donné dans les expériences. Toutes ces mesures d'erreur sont comprises entre 0 et 1, 0 étant la meilleur performance possible en terme de GER et 1 en terme de RH et PH .

### 7.4.2 Choix de la méthode d'estimation sinusoïdale

Nous avons vu dans la première partie un grand nombre de méthodes d'estimation des paramètres sinusoïdaux, certaines étant beaucoup plus performantes que les méthodes utilisées dans les articles [Doval and Rodet, 1993] et [Maher and Beauchamp, 1994]. Cependant, à cause du protocole de test et des mesures de performance utilisées, la précision de l'estimation d'amplitude et de fréquence ne sera pas critique ici. En effet, la seule mesure qui pourrait être fortement influencée par la précision de l'analyse sinusoïdale est le FER, mais celle-ci n'a pas été retenue ici, car la précision fréquentielle de la référence du corpus de test n'est malheureusement pas connue.

Le choix s'est donc porté sur la méthode décrite à la section 6.1.1.1, qui est l'estimateur à 2 bins présenté dans [Betser et al., 2006c]. Il s'agit d'une des méthodes d'estimation parmi les plus rapides[8], avec de bonnes performances dans le cas d'une sinusoïde constante, et présentant un bon compromis entre précision et résolution, grâce à l'utilisation de la fenêtre de Hann. La taille de fenêtre utilisée est de 32 ms et le pas de 8 ms, valeurs classiques en analyse de parole.

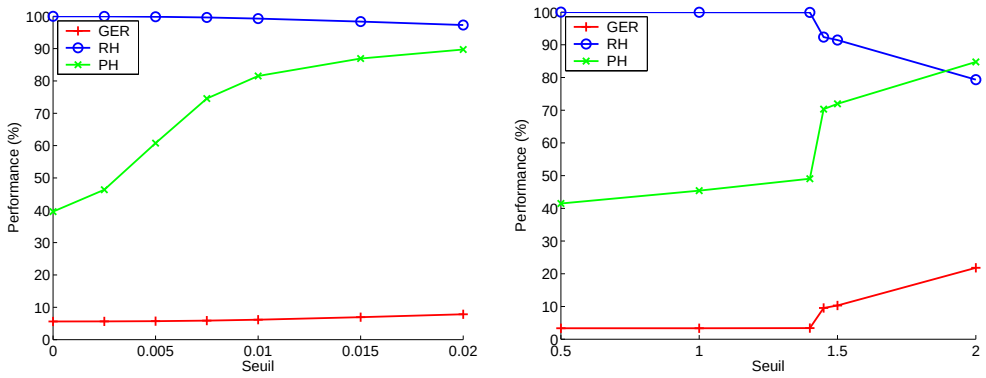
### 7.4.3 Résultats et commentaires

Deux expériences ont été réalisées pour évaluer l'algorithme. La première concerne les mesures d'harmonicité, afin de déterminer laquelle, de la mesure de Maher modifiée décrite dans la section 7.3.3.1 ou de la mesure probabiliste de la section 7.3.3.2, est la plus performante. La deuxième expérience concerne la comparaison de l'algorithme complet avec les méthodes état-de-l'art.

Les paramètres de notre algorithme concernant l'estimation sinusoïdale ont été choisis conformément à la section 7.4.2. Les paramètres des deux mesures d'harmonicité ont été réglés sur un petit corpus de développement constitué de quelques fichiers de parole fortement bruités. Les paramètres optimaux trouvés pour la mesure de Maher sont les mêmes que ceux donnés dans l'article [Maher and Beauchamp, 1994], ce qui confirme que ces valeurs sont indépendantes du corpus utilisé[9]. Ces valeurs sont : $p_1 = .5$, $p_2 = .5$ et $p_3 = 1.4$. La tolérance utilisée dans le cas de la mesure

[8] Cette méthode est plus rapide que l'interpolation parabolique simple par exemple, car celle-ci requière un coefficient de zéro-padding élevé ($\geq 4$) [Abe and Smith III, 2004] pour donner des performances similaires. Pour la définition du zéro padding voir la section 3.3.6.1.

[9] A noter que les valeurs des paramètres $q$ et $r$ de l'article [Maher and Beauchamp, 1994], page 2257, ont été inversées.




(a) Mesure probabiliste simplifiée (b) Mesure de Maher modifiée

FIG. 7.12: Performance des mesures d'harmonicité sans suivi de pitch, en fonction du seuil sur la vraisemblance

de Maher modifiée est la tolérance maximale possible, *i.e.* $p/2$ pour une hypothèse de pitch $p$. C'est la tolérance qui donne le meilleur résultat avec cette mesure. Les paramètres de la mesure probabiliste sont quand à eux : $\alpha_1 = .0011$ et $\alpha_2 = .3$. Pour la mesure probabiliste, la tolérance utilisée pour l'identification de pics est d'après l'équation (7.3.4) $T_f = 16$ Hz. Enfin le suivi de pitch utilise un buffer d'hypothèses de 20 trames et, lorsque ce n'est pas précisé, un filtrage des chemins plus courts que 4 trames. Le seuil minimum sur la vraisemblance pour lier un pic est de 0.005. Ces deux derniers paramètres vont déterminer la qualité du suivi de pitch.

TAB. 7.1: Comparaison des deux mesures d'harmonicité sans seuillage

| | GER (%) |
|---|---|
| Mesure probabiliste simplifiée | 5,6 |
| Mesure de Maher modifiée | 3,3 |

Pour la comparaison des mesures d'harmonicité, le suivi de pitch a été désactivé. Seule l'hypothèse de pitch la plus vraisemblable a été gardée[10]. Le tableau 7.1 donne les performances des deux mesures, sans seuillage. Le rappel et la précision sont superflus ici car une estimation est donnée pour toutes les trames qu'elles soient harmoniques ou non. On voit que la mesure de Maher modifiée est meilleure que la mesure probabiliste dans ce cas de figure, avec un gain de performance non négligeable. La plupart des erreurs de la mesure probabiliste correspond à des demi-pitchs, des erreurs qui seront facilement corrigées par le suivi de pitch comme nous le verrons par la suite.

Les performances des deux mesures ont ensuite été comparées en utilisant un seuil sur la vraisemblance des hypothèses de pitch, pour tester à la fois le GER et la détection d'harmonicité. Les courbes correspondantes sont données sur la Figure 7.12. Ici apparaît un gros défaut de la mesure de Maher et Beauchamp simplifiée. Dans le cas

[10] voir section 7.3.4 pour plus de détails.

TAB. 7.2: Comparaison des algorithmes d'estimation de fréquences fondamentales. Les mesures sont en pourcentage.

| Algo | GER | Précision | Rappel |
|---|---|---|---|
| RAPT | 5,7 | 77,6 | 97,9 |
| YIN | 3,6 | 39,4 | 100 |
| SPS0 | 1,9 | 39,4 | 100 |
| SPS1 | 1,96 | 68,3 | 99,9 |
| SPS2 | 1,97 | 70,9 | 99,9 |
| SPS3 | 1,97 | 72,7 | 99,9 |
| SPS4 | 2,0 | 74,1 | 99,9 |

où l'harmonique fondamentale est aussi l'harmonique la plus forte, l'hypothèse la plus probable correspond à un peigne avec une seule harmonique ($K = 1$), et la valeur de vraisemblance obtenue est alors $L(p) = p_3$ (voir l'équation 7.3.3.1). Un signal avec $K$ harmoniques et dont la première harmonique est la plus forte aura la même vraisemblance qu'un signal avec une seule harmonique : la mesure de Maher et Beauchamp simplifiée n'ordonne pas correctement les hypothèses de pitch. En conséquence, même si cette mesure donne de bons résultats avec un seuil faible, elle va présenter un gros palier d'erreur pour un seuil de $L(p) = p_3 = 1.4$ (Figure 7.12(b)), contrairement à la mesure probabiliste, dont l'erreur progresse linéairement (Figure 7.12(a)). Or l'algorithme de suivi de pitch a besoin d'une mesure qui ordonne convenablement les hypothèses pour fonctionner correctement. Ce désavantage rend donc la mesure de Maher modifiée moins intéressante que la mesure probabiliste si le suivi de pitch est utilisé. C'est cette dernière mesure qui sera finalement gardée pour la comparaison avec les algorithmes état-de-l'art.

L'algorithme de suivi de pitch sinusoïdal (SPS) est ensuite comparé au YIN et au RAPT et les résultats sont donnés dans le tableau 7.2. La mesure d'harmonicité utilisée est la mesure probabiliste. L'algorithme SPS est décliné en plusieurs versions. Le chiffre après SPS indique la taille minimale pour qu'un chemin soit gardé dans le buffer de décision (voir Figure 7.10). 0 signifie qu'il n'y a aucun post-traitement, et 4 signifie que tous les chemins d'une longueur inférieure à 4 trames ont été filtrés (4 trames=32 ms car on rappelle que le pas est de 8 ms). Ce filtrage permet d'éliminer les chemins causés par le bruit et de délimiter les zones harmoniques, comme le montre le tableau 7.2. En effet les performances de la détection des zones harmoniques augmentent très rapidement tout en gardant un GER très faible. Pour un filtrage supérieur à 4, les performances commencent à chuter assez vite, car on élimine alors des zones harmoniques importantes.

Dans leur version disponible librement, l'algorithme YIN est employé sans suivi de pitch, il doit donc être comparé à la ligne SPS0 de notre algorithme, contrairement à l'algorithme RAPT qui en utilise un et qui doit donc être comparé à la ligne SPS4. On voit que l'algorithme SPS est bien meilleur que le RAPT, le gain de performance étant de presque 4% en terme de GER, pour un rappel/précision à peu près égal. Il est également légèrement meilleur que le YIN, même si la différence est cette fois

moins significative. On peut constater également la qualité de la détection des zones harmoniques, car les performances restent meilleures que le YIN sans détection des zones harmoniques. Enfin si on se réfère à l'article [de Cheveigné and Kawahara, 2002] qui compare YIN avec un grand nombre d'algorithmes de la littérature, dont RAPT, et sur des bases de données très similaires, l'algorithme SPS avec la mesure probabiliste simplifiée devrait se placer parmi les plus performant.

A titre indicatif, le temps requis pour traiter les données est de l'ordre d'un dixième du temps réel, sur un pentium 1700 MHz. L'algorithme a été implanté en C/C++, sans optimisation de code particulière. Le coût principal de calcul est en fait la mesure d'harmonicité, qui bien que très simple peut prendre du temps si le nombre d'hypothèses de pitch générées est important. La solution est bien entendu de faire un tri parmi ces hypothèses avant le calcul de la mesure d'harmonicité, mais il ne nous a cependant pas semblé nécessaire pour l'usage que nous avions de l'algorithme.

## 7.5 Conclusion

Nous avons présenté une méthode simple et rapide pour estimer le pitch à partir des pics sinusoïdaux. Cette méthode s'inspire des travaux de [Maher and Beauchamp, 1994] et [Doval and Rodet, 1993]. Deux mesures d'harmonicité basées sur les pics sinusoïdaux ont été développées, toutes deux rapides et simples à mettre en oeuvre : la mesure de Maher modifiée et une mesure probabiliste simplifiée. La première donne de meilleurs résultats si le suivi de pitch est désactivé, mais moins bons lorsqu'il est activé. Le suivi de pitch est une méthode originale combinant programmation dynamique et guide d'onde. Il est également fenêtré, ce qui permet de l'utiliser pour des applications temps réel.

L'algorithme complet avec la mesure probabiliste simplifiée donne de bons résultats par rapport à l'état de l'art en terme de GER et de détection des zones harmoniques. Sur le type de base utilisée, c'est à dire de la parole monophonique très peu bruitée, beaucoup d'algorithmes ont des performances très bonnes, mais notre algorithme permet tout de même d'améliorer légèrement les performances. Les tests seraient certainement plus significatifs en modifiant le protocole de test de ces méthodes, en choisissant par exemple d'ajouter divers bruits typiques, facilement reproductibles, au corpus de parole. C'est une première perspective de ce travail.

Les autres perspectives concernent d'abord l'extension de l'algorithme à l'analyse multirésolution. En effet, il est connu que la résolution optimale varie en fonction du locuteur, en particulier s'il s'agit d'un homme ou d'une femme. Notre algorithme peut facilement prendre en compte des pics sinusoïdaux provenant d'analyses à des résolutions différentes. Les seules modifications dans ce cas concernent la partie analyse sinusoïdale, en amont de l'algorithme présenté dans cette partie.

Une dernière perspective intéressante est l'analyse multipitch, auquel se prête bien cet algorithme. Des algorithmes d'estimation multipitch basés sur les pics sinusoïdaux existent déjà, certains donnant de très bons résultats [Klapuri, 2004].

L'estimation de pitch basé sur l'analyse sinusoïdale s'est avérée performante, tout en restant simple et rapide. Des mécanismes comme l'identification de pics sinusoï-

daux et la mesure d'harmonicité peuvent être réutilisés pour d'autres tâches d'indexation, comme nous allons le montrer dans le chapitre suivant dans le cas de l'identification audio.

CHAPITRE 8

# Identification audio

## 8.1 Introduction

L'identification de documents audio, ou audio ID, se place dans le cadre de la recherche par requête audio. A partir d'un exemplaire audio on cherche à retrouver de l'information textuelle, sonore ou audio-visuelle, comme par exemple des métadonnées, des contenus audio identiques ou similaires etc. On peut distinguer plusieurs catégories d'applications pour la recherche par requête audio, qui dépendent du degré de similitude recherché vis-à-vis du document original :

- Recherche de documents rigoureusement identiques. Ex : Vérification de l'intégrité des données pour le codage audio, la cryptographie.
- Recherche de documents perceptuellement identiques mais qui ont pu souffrir de détériorations ou de modifications mineures. C'est dans cette catégorie que se situe l'audio ID.
- Recherche de documents pouvant être très différents mais présentant des ressemblances suivant un certain critère. Ex : classification de morceaux de musiques, de programmes TV, recherche par mélodie chantée, ou pianotée, recherche par rythme, aide à la découverte de nouveaux artistes (recherche par "goût" musical).

L'audio ID se place dans une catégorie à mi-chemin entre les extrêmes que constituent la détection exacte et le classement par genre. A l'intérieur des applications potentielles de l'audio ID, le degré de similarité recherché peut également varier. C'est pour cette raison qu'il est important de préciser le degré de similarité recherché pour chaque application envisagée et d'essayer de cerner sur quelles caractéristiques de l'objet sonore va se porter la recherche de similarités. Par objet sonore, nous entendrons tout type de document sonore ou de fragment de document sonore, dont les reproductions ont la propriété de rester perceptuellement similaires.

Il existe un grand nombre d'applications potentielles de l'audio ID, la plupart ayant émergé très récemment. Nous pouvons les distinguer en trois catégories suivant l'usage auxquelles elles sont destinées :

- La recherche d'information sur un document : identification des pistes d'un CD audio par rapport à une base de CD de référence, ou plus généralement retrouver les métadonnées d'un document audio inconnu, effacer les dupliquas d'une base de donnée audio, identification d'une chanson "en direct" (entendue à la radio par exemple) par l'intermédiaire d'un téléphone portable, ou de tout autre appareil d'enregistrement, etc.
- La recherche d'occurrences à des fins de structuration : détection de jingle comme première étape pour l'analyse d'émissions radiophoniques ou télévisuelles (recherche d'information, fabrication automatique de résumés), détection des jingles de publicité en vue de leur suppression etc.
- La recherche d'occurrences à des fins de contrôle de la diffusion : confirmation de la diffusion de publicités pour les sponsors, filtrage des documents audio légalement distribués pour des services en-ligne comme Napster, détection d'utilisation illégale de contenus multimédia, analyses statistiques ("charts analysis"), systèmes visant la rétribution des propriétaires des droits des documents diffusés, etc.

A toutes ces applications, on peut ajouter leur équivalent multimédia, c'est à dire les tâches d'identification conjointes audio et vidéo. Les systèmes audio utilisés pour le multimédia restent très similaires à leur version uni-media, car les traitements audio et vidéo se font généralement en parallèle et la décision conjointe est réalisée par fusion des résultats des différents systèmes. La liste d'applications présentée ici n'est certainement pas exhaustive et il est très probable que d'autres applications vont émerger avec l'évolution des technologies et des besoins.

Dans ce chapitre nous verrons tout d'abord, à la section 8.2, les différentes contraintes auxquelles sont soumis les algorithmes d'audio ID, et en particulier nous parlerons des détériorations de l'objet sonore. Dans la section 8.3, nous parlerons des méthodes existantes, et nous nous intéresserons plus particulièrement aux méthodes basées sur la création d'empreintes. Nous décrirons trois méthodes caractéristiques de l'état-de-l'art en détail. Elles nous serviront de point de comparaison dans les expériences. Ensuite nous aborderons à la section 8.4 le problème de l'audio ID basé sur l'estimation sinusoïdale. En particulier nous essaierons de montrer les avantages et les inconvénients spécifiques d'une telle méthode. Enfin, nous comparerons à la section 8.5 la méthode développée avec l'état de l'art sur deux tâches caractéristiques de l'audio ID, la détection de jingles radiophoniques et l'identification de morceaux de musique.

## 8.2 Contraintes de l'audio ID

Les systèmes d'identification audio sont confrontés à des contraintes, que l'on peut séparer en deux catégories. D'une part il y a des contraintes physiques, c'est à dire des contraintes que l'on ne maîtrise pas, essentiellement les déformations que l'on peut attendre entre un document original et une copie de ce document. D'autre part il y a des contraintes techniques, qui dépendent de la méthode utilisée : la vitesse de calcul, la taille de stockage etc. Sans faire une liste exhaustive, nous allons détailler les

principales difficultés généralement rencontrées, et auxquelles nous serons confrontés dans les deux applications envisagées.

### 8.2.1 Les dégradations du signal

Beaucoup de travaux présentent des revues de ces modifications, qui dépendent en fait du type d'application envisagée [Allamanche et al., 2001], [Haitsma and Kalker, 2003], [Cano et al., 2005]. Parmi ces dégradations, on trouve :

- l'ajout de "bruit" au document : du bruit dû à la mauvaise qualité de la transmission (bruit aléatoire), mais aussi des sons superposés au document original, comme de la parole d'un animateur radio sur un morceau de musique.
- les distorsions et les interférences causées par la transmission du signal : par exemple les transmissions Hertziennes, la conversion analogique/digitale etc.
- le changement de vitesse du document ('pitching') très fréquent sur les radios.
- les changements d'amplitude qui peuvent être globaux (volume sonore) ou fréquence par fréquence ('equalization'). A l'extrême on peut avoir une diminution de la bande de fréquence du document audio, comme lorsque l'on passe de la radio FM à la radio AM.
- les effets dus à la compression audio avec perte d'information (codage perceptuel), comme le mp3.
- la granularité du système : sa capacité à reconnaître un document à partir d'un extrait du document, comme le refrain d'une chanson par exemple.
- le décalage temporel entre l'original et la copie.

L'ajout de bruits autres que des bruits aléatoires, comme de la parole, est un problème très fréquemment rencontré dans le cas de l'analyse de flux radiophoniques ou télévisuels, mais qui n'est pas évoqué à notre connaissance dans la littérature. Ce type de modification peut altérer de façon significative le document audio, mais doit rester suffisamment faible pour que ce dernier reste reconnaissable pour des humains.

### 8.2.2 Les contraintes concernant la mise en oeuvre

Ce sont des contraintes d'ordre technique et applicatif. On peut citer [Cano et al., 2005] :

- les performances souhaitées : est-ce que l'on va plutôt tolérer des erreurs de fausse détection (on détecte quelque chose alors qu'il n'y a rien), ou de détection manquée (on rate un document)...
- la rapidité requise par l'application, dépendant évidement de la quantité de données à traiter, c'est à dire le nombre de documents à détecter dans un flux continu ou la taille de la base de données où l'on cherche notre document. Dans le premier cas, la contrainte est souvent celle du traitement en temps réel.
- la taille de stockage nécessaire dans le cas des bases de données.

Ces contraintes sont parfois contradictoires, et résultent souvent en un compromis, par exemple entre les performances souhaitées et la rapidité. L'enjeu principal des méthodes d'empreinte consiste à extraire l'empreinte qui offre le meilleur compromis.

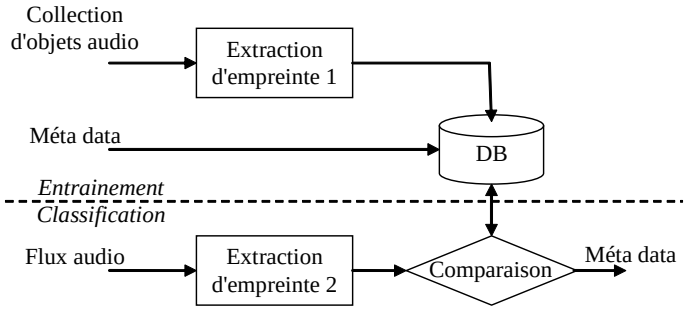



FIG. 8.1: Identification audio basé sur l'extraction d'empreintes

## 8.3 Etat de l'art

Parmi les techniques d'audio ID, et plus généralement d'identification multimédia, on trouve deux principaux types de méthodes : les méthodes dites de tatouage ou de "watermarking" [Cox et al., 1996], [Boney et al., 1996], et les méthodes utilisant des empreintes, dites de "fingerprinting" [Cano et al., 2005]. Elles reposent sur deux principes différents : la première se propose de cacher les informations essentielles à l'identification dans le document audio, tandis que la seconde extrait une référence, l'empreinte, directement calculée sur le document audio et va chercher les informations utiles dans une base de données de références. Nous nous intéressons ici aux méthodes dites de "fingerprinting", qui sont avantagées par leur robustesse et leur caractère non-intrusif.

Ces méthodes présentent toutes le même schéma d'analyse, très bien décrit dans [Cano et al., 2005] pour le cas de l'identification de morceaux de musique. Ce schéma est représenté sur la Figure 8.1 dans un cadre plus général que celui de [Cano et al., 2005]. Une empreinte audio peut être vue comme un résumé court d'un objet audio [Haitsma and Kalker, 2002]. Ce résumé doit être unique pour chaque objet, afin de pouvoir les distinguer, et doit donc se baser sur le contenu de l'objet. Il est généralement extrait à partir des composants de l'objet sonore les plus pertinents perceptuellement [Cano et al., 2002]. On rappelle que par objet sonore on entend tout type de documents sonores ou de fragments de documents sonores. Les méta-données désignent littéralement les "données sur les données", c'est-à-dire toutes les informations sémantiques additionnelles qui peuvent se rapporter à l'objet sonore. On peut citer par exemple, le nom du document, le nom de l'auteur, la date de création etc. Enfin un flux audio désigne tout type de media, allant des fichiers audio stockés, à la diffusion radio ou télévisuelle.

La Figure 8.1 montre qu'un système d'identification audio par empreinte est constitué de trois parties :

- Un module d'extraction d'empreintes
- Un module de stockage ('DB' pour database)
- Un module de comparaison d'empreintes

Le module de stockage peut être facultatif pour certaines tâches, les empreintes de référence pouvant être calculées à l'initialisation du système au lieu d'être stockées. Le principal travail dans le développement d'un algorithme d'audio ID portera cependant essentiellement sur les deux autres modules et plus particulièrement l'extraction de l'empreinte.

Ainsi qu'il a été mentionné ci-dessus, les reproductions des objets sonores que l'on cherche à retrouver doivent rester perceptuellement similaires. Nous avons déjà décrit les types de dégradations auxquels les objets sonores vont être soumis à la section 8.2. La clé dans le développement d'une empreinte robuste à ces dégradations est d'identifier quelles vont être les caractéristiques du signal les plus robustes à ces modifications.

Beaucoup de travaux décrivent des techniques d'extraction d'empreinte suivant le cadre défini ci-dessus :

- [Laroche, 2001] décrit une empreinte formée à partir du spectre de modulation du flux énergétique de quelques bandes de fréquence.
- [Pinquier and André-Obrecht, 2004] décrit une empreinte formée directement à partir des magnitudes spectrales d'une analyse en sous-bandes
- [Burges et al., 2001] propose une empreinte calculée par application successive de plusieurs analyses en composantes principales.
- [Gomes et al., 2003] utilise un descripteur MFCC et une modélisation HMM comme empreinte.
- [Fragoulis et al., 2001] et [Papaodysseus et al., 2001] proposent une empreinte basée sur un état binaire (activé, non-activé) des sous-bandes spectrales.
- [Haitsma and Kalker, 2002, 2003] propose d'utiliser le signe des différences, à la fois temporelles et fréquentielles, d'énergies des sous-bandes. Cette méthode est intégrée dans la solution de Philips dédiée à l'identification de morceaux de musique (cf. Gracenote).
- [Herre, 2004] décrit une empreinte basée sur la mesure de platitude spectrale.
- [Wang, 2003] propose un descripteur basé sur les attaques sinusoïdales. Cette méthode est intégrée dans la solution de Shazam dédiée à l'identification de morceaux de musique.

D'autres méthodes existent mais les travaux décrits ci-dessus sont représentatifs de l'état de l'art. Toutes ces méthodes, hormis la dernière, sont basées sur le principe d'un découpage du spectre en sous-bandes. Parmi ces méthodes nous retiendrons la méthode de Pinquier, qui est la méthode la plus simple parmi celles citées ci-dessus, et qui nous servira d'introduction à la section 8.3.1. Nous retiendrons également la méthode de Philips qui est parmi les plus citées et qui a donné lieu à un certain nombre de variantes. Elle sera décrite à la section 8.3.2. Enfin nous parlerons à la section 8.3.3 de la méthode de Shazam qui est, parmi les approches citées, la plus proche de celle que l'on propose.

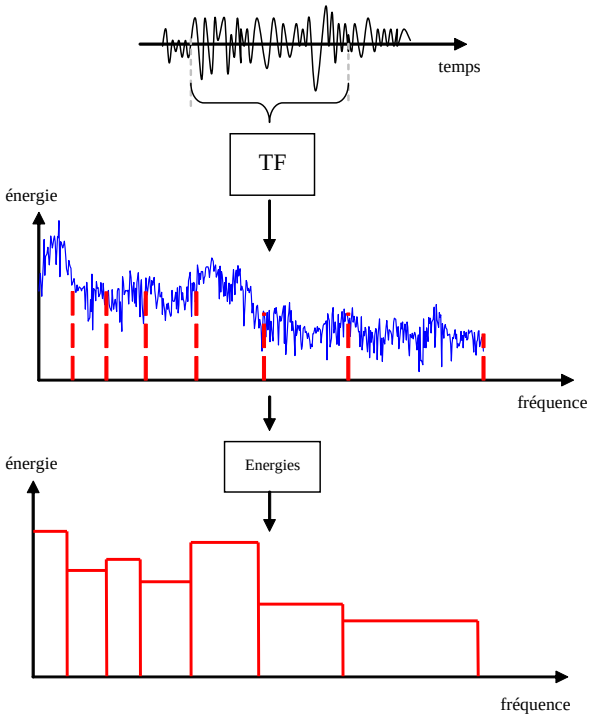



FIG. 8.2: Principe de l'analyse en sous-bandes

### 8.3.1 Méthode de Pinquier

#### 8.3.1.1 Formation de l'empreinte

Le principe général est décrit à la Figure 8.2. Le signal audio temporel est tout d'abord fenêtré et découpé en trames de quelques dizaines de millisecondes avec un recouvrement typiquement de 50% ; on réalise ensuite une transformée temps/fréquence, ici une transformée de Fourier, sur chacune de ces trames. Le spectre ainsi obtenu est ensuite découpé en $N_b$ sous-bandes, selon une échelle Bark afin de prendre en compte la sensibilité de l'oreille humaine. On calcule ensuite l'énergie de chacune des sous-bandes, formant ainsi un vecteur de taille $N_b$.

#### 8.3.1.2 Comparaison

Le principe de la méthode est illustré sur la Figure 8.3. La comparaison va se faire bloc par bloc puis trame à trame. Considérons que le signal à identifier est un bloc, noté $B^{'}$, composé de $N_t^{'}$ trames. On rappelle que chaque trame est un vecteur composé des énergies des sous-bandes et de taille $N_b$. On va ici comparer $B^{'}$ à tous les blocs possibles $B_{j,n}$ de longueur $N_t^{'}$ parmi tous les objets de référence, où $j$ est le numéro de l'objet de référence et $n$ l'index de trame initiale dans l'objet de référence

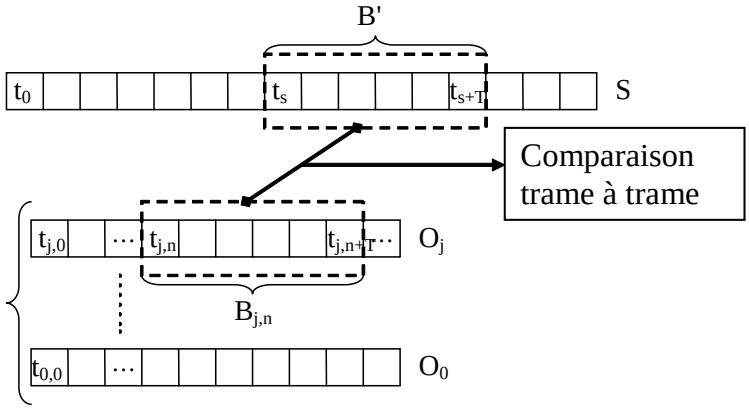



FIG. 8.3: Comparaison bloc par bloc, trame à trame. $S$ est le signal à analyser, et $O_j$ sont les objets de référence.

$j$. La mesure de similarité entre un bloc $B_{j,n}$ et le bloc de référence $B^{'}$ est ici la distance euclidienne entre ces deux blocs.

En fait dans la méthode de Pinquier [Pinquier and André-Obrecht, 2004], la comparaison se fait directement entre l'objet de référence complet et un bloc du signal à analyser $B^{'}$ de même taille que l'objet en question. Les objets pouvant être de tailles très variable, cela peut poser des soucis de normalisation et de mise en oeuvre. On a donc préféré une approche "bufferisée", utilisée dans quasiment tous les autres systèmes.

#### 8.3.1.3 Décision

La décision est prise en utilisant un système de seuillage adaptatif. Faisons l'hypothèse que le bloc $B^{'}$ correspond exactement au bloc $B_{j,n}$. Alors d'une part la mesure de similarité sera à priori élevée aux "alentours" du bloc $B_{j,n}$, ce qui justifie l'usage d'un premier seuil absolu. D'autre part la mesure de similarité présentera un maximum marqué au temps $t_{j,n}$. Ce pic dans la mesure de vraisemblance sera d'autant plus marqué que le bloc à comparer sera long et varié. Une étude de la finesse de ce pic peut donner une indication sur la qualité de la détection. Si les blocs sont courts (par exemple d'une seconde), elle peut servir notamment à filtrer les blocs peu informatifs, comme des notes simples tenues. Nous verrons dans l'algorithme d'analyse sinusoïdale qu'un tel seuil permet d'éviter les fausses alarmes, causées par ce cas de figure. La décision se fait donc par un double seuillage, illustré sur la Figure 8.4 :

- un premier seuil absolu $S_1$ pour la hauteur du pic de distance
- un deuxième seuil relatif $S_2$ pour la finesse du pic de distance.

Pour le deuxième seuil, on trouve d'abord les minima à gauche et à droite du pic sur l'intervalle d'analyse de taille fixée $T_d$. On calcule $H_1$ et $H_2$ les hauteurs relatives du pic par rapport aux minima et enfin, on teste si $H_1$ et $H_2$ sont supérieurs au seuil $S_2$.

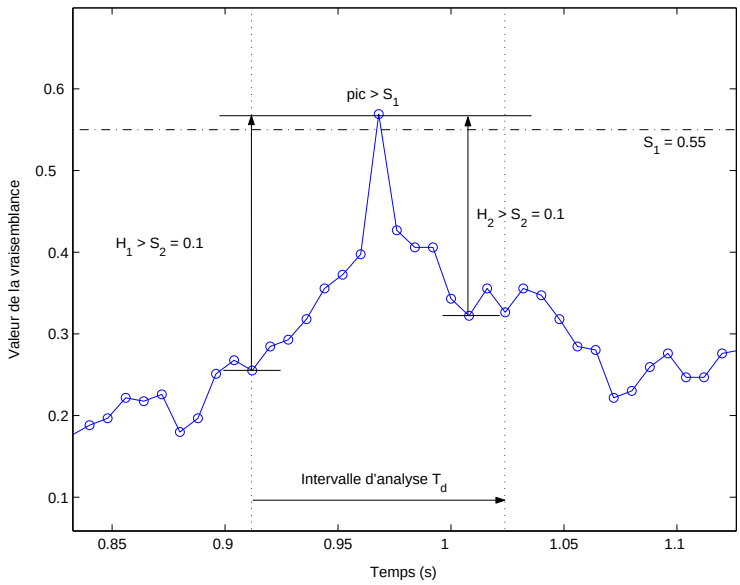



FIG. 8.4: Décision par double seuil

La méthode est légèrement différente de celle de [Pinquier and André-Obrecht, 2004] qui ne prend pas les deux minima à gauche et à droite du pic mais les valeurs de la mesure de similarité respectivement à $-T_d/2$ et à $+T_d/2$ du maximum. On a constaté expérimentalement que considérer les minima était plus efficace.

### 8.3.2 Méthode de Philips

#### 8.3.2.1 Formation de l'empreinte

L'empreinte de la méthode de Philips [Haitsma and Kalker, 2002] utilise, comme celle de Pinquier, une analyse en banc de filtres. Pour rendre l'empreinte plus robuste, les auteurs proposent de calculer le signe des différences d'énergie, simultanément le long de l'axe temporel et fréquentiel (Figure 8.5). L'empreinte pour une trame est alors un vecteur binaire $F_t$. En notant $E_{t,i}$ l'énergie de la sous bande numéro $i$ au temps $t$, le bit numéro $i$ de $F_t$ est calculé ainsi :

$$F_{t,i} = \begin{cases} 1 \text{ si } (E_{t,i} - E_{t,i+1}) - (E_{t-1,i} - E_{t-1,i+1}) > 0 \\ 0 \text{ si } (E_{t,i} - E_{t,i+1}) - (E_{t-1,i} - E_{t-1,i+1}) \leq 0 \end{cases} \tag{8.3.1}$$

Diverses variations de cette empreinte ont été développées, certaines par les auteurs eux-mêmes [Haitsma and Kalker, 2003], dans le but de rendre l'empreinte plus robuste à des déformations comme le changement de vitesse de lecture. Les gains ne sont cependant pas très importants, et seule l'approche de base de Philips a été implantée et testée.

Un vecteur est calculé toutes les 12 ms, ce qui donne environ 86 trames par seconde. Le nombre d'éléments à stocker pour une base de données musicale de 10000 morceaux de 5 minutes est environ de 260 millions.

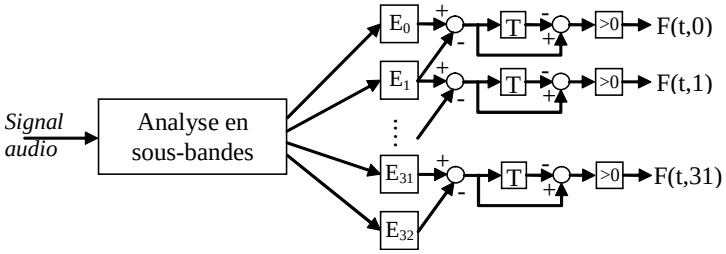



FIG. 8.5: Extraction de l'empreinte de Philips. $E_j$ est l'énergie de la sous-bande numéro $j$. $T$ est ici l'opérateur de retardement d'une trame.

#### 8.3.2.2 Comparaison et décision

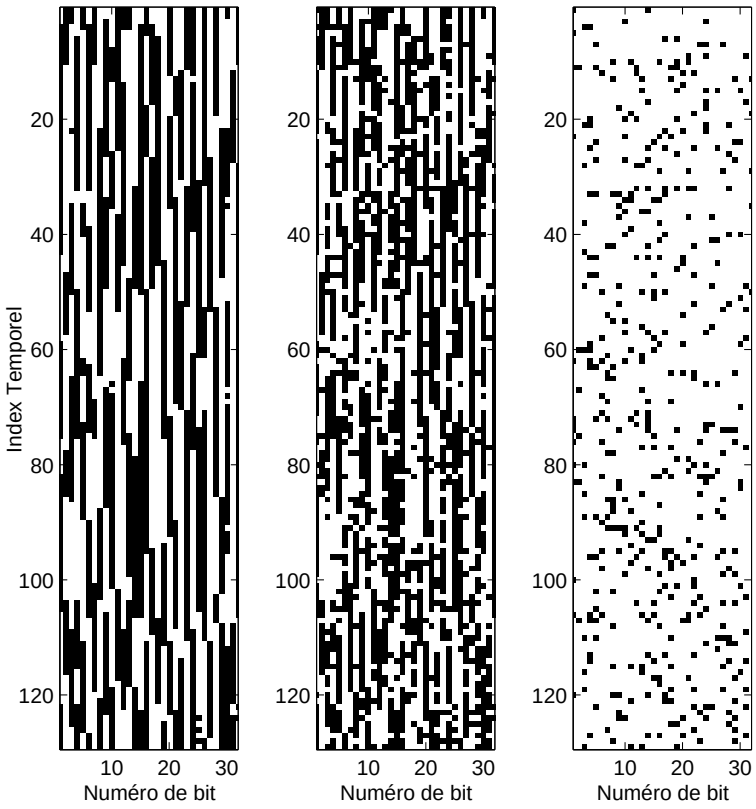
La comparaison entre deux empreintes est une comparaison bit à bit de deux blocs de bits, effectuée à l'aide d'une opération OU-exclusif. La Figure 8.6 donne un exemple d'une telle comparaison. La mesure de similarité utilisée est le taux d'erreur du résiduel (BER), c'est à dire le nombre de bits faux divisé par le nombre total de bits. La Figure 8.6(c) donne un exemple de résiduel et le BER correspondant.

L'empreinte inconnue est considérée comme identifiée si le BER est inférieur à un certain seuil. Les auteurs [Haitsma and Kalker, 2002] ont montré que pour une valeur de 0.35, la probabilité d'une fausse alarme est extrêmement faible, de l'ordre de $10^{-20}$.

#### 8.3.2.3 Réduction de l'espace de recherche

Dans le cas de grandes bases de données, typiquement une base de donnée de 10000 chansons de 5 minutes de moyenne, le temps de comparaison selon une méthode exhaustive, comme la méthode décrite dans la section 8.3.1.2, serait trop long, de l'ordre de plusieurs dizaines de minutes, alors que l'identification d'un morceau doit se faire de façon quasi-instantanée. Il est donc indispensable de réduire l'espace de recherche. La méthode adoptée par Philips est d'indexer toutes les trames des objets de référence dans une table. Si le nombre de sous-bandes utilisé est $N_b$, alors chaque trame sera un vecteur de $N_b$ bits et la table de look-up sera une table à $2^{N_b}$ entrées. Chaque entrée, appelée clé, va pointer vers tous les objets qui possèdent exactement cette entrée, au temps correspondant (Figure 8.7). Le nombre d'entrées de la table est tellement énorme, que si les vecteurs étaient répartis de façon uniforme dans les objets de référence, le nombre d'éléments par clé serait très faible, de l'ordre de $10^{-1}$ pour une base de données de 10000 morceaux de musique [Haitsma and Kalker, 2002]. Le temps de comparaison selon un tel procédé serait alors négligeable.

Cependant ce mécanisme est confronté à deux difficultés. La première est la limitation de la mémoire disponible. La table de look-up est en effet trop volumineuse pour être chargée en mémoire. La solution est d'utiliser une table de hash à la place d'une table de look-up simple. Dans une table de hachage l'indice du tableau n'est plus directement la clé, mais une valeur calculée à partir de la clé grâce à une fonction,




(a) Empreinte originale (b) Empreinte bruitée (c) Différence des empreintes

FIG. 8.6: Exemple d'empreinte de Philips, pour un extrait de jingle radio. Pour les Figures (a) et (b), le pixel est blanc si le bit vaut 1. Pour la Figure (c) les erreurs sont représentées en noir et correspondent à un BER de 0.1

la fonction de hash[1]. Lorsque l'ensemble des clés réellement utilisées est beaucoup plus petit que le nombre total d'entrée de la table, la table de hachage requière moins de place de stockage [Cormen et al., 2001].

La deuxième difficulté concerne les déformations que le signal à identifier va subir. En effet pour que la comparaison fonctionne, il faut que les clés de table de look-up soit reproduites de façon exacte dans le signal à analyser. Les auteurs de la méthode [Haitsma and Kalker, 2002] ont remarqué que si les blocs utilisés pour la comparaison sont suffisamment grands, alors il y aura quasiment toujours des trames qui présenterons une clé exactement identique. Pour augmenter les chances de trouver des clés identiques, les auteurs proposent de considérer pour chaque clé du signal à analyser toutes les clés présentant une distance inférieure à $k$, c'est à dire toutes les clés ayant moins de $k$ bits différents. Avec $k = 2$ par exemple, il y aura en tout $C^0_{N_b}+C^1_{N_b}+C^2_{N_b} = 529$ clés à tester. Pour réduire le nombre de clés à tester les auteurs

[1]Le lecteur se référera au chapitre 11 de [Cormen et al., 2001] pour plus de détails.

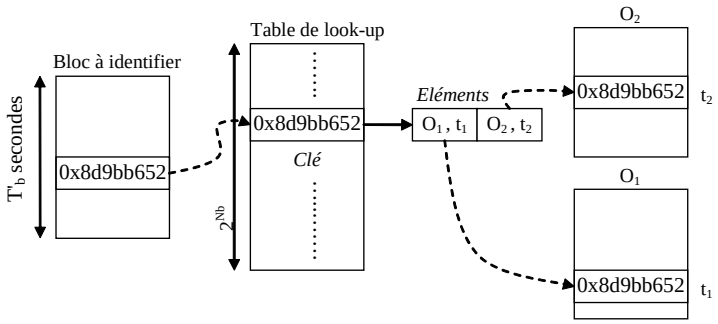



FIG. 8.7: Réduction de l'espace de recherche par une table de look-up

proposent également de sélectionner celles dont les bits sont plus "fiables" en utilisant un critère de confiance. Les auteurs soulignent que leur mesure de confiance n'est pas toujours parfaite et conduit à une amélioration limitée. Ce dernier mécanisme n'a donc pas été testé.

### 8.3.3 Méthode de Shazam

L'approche de Shazam [Wang, 2003], [Wang, 2006] n'a pas été implantée ni testée, mais elle présente des caractéristiques proche d'un algorithme basé sur les pics sinusoïdaux. C'est pourquoi nous la décrivons en détail dans cette section.

#### 8.3.3.1 Formation de l'empreinte

L'empreinte est basée sur les bins de la transformée de Fourier présentant les variations d'amplitude les plus fortes, qui seront appelés "points d'intérêts". La plupart de ces points d'intérêts vont correspondre à des attaques, ou des relâchements sinusoïdaux (voir Figure 8.8). Par un seuillage, on va sélectionner un nombre restreint de points d'intérêt, de l'ordre d'une dizaine par seconde. Pour avoir une couverture à peu près uniforme du spectre de fréquence, on va utiliser un seuillage différent par bande de fréquence [Ogle and Ellis, 2007].

Une fois les points d'intérêts déterminés, on va les combiner pour augmenter la quantité d'information apportée, car une fréquence seule n'est pas suffisamment discriminante. On va former des paires de points d'intérêt, chaque paire donnant naissance à une clé. Pour deux points d'intérêt $(f_1, t_1)$ et $(f_2, t_2)$, la clé est formée par le triplet[2] $(f_1, f_2, t_2 - t_1)$. On considère maintenant un point d'intérêt $(f, t)$. Pour éviter une explosion combinatoire, les paires associées à ce point sont choisies de la façon suivante :

– on limite les candidats à l'appariement à un horizon temporel $[t+1, t+\Delta_t]$ et à un horizon fréquentiel $[f-\Delta f, f+\Delta f]$. On évite les temps inférieurs ou

[2]Le temps $t_1$ n'est pas retenu dans la clé, car on souhaite s'affranchir du temps absolu. On ne garde donc que le temps relatif $t_2 - t_1$.

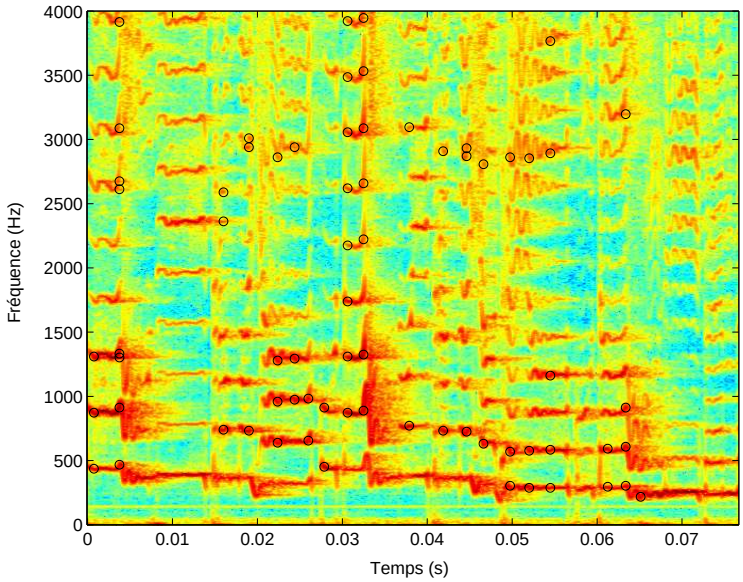



FIG. 8.8: Sélection des points d'intérêt dans la méthode de Shazam

égaux à $t$ pour empêcher les doublons et les combinaisons avec une différence temporelle nulle.
– parmi les candidats restant on choisit les $K$ plus énergétiques, typiquement $K \leq 10$.

On peut remarquer que, comme l'horizon fréquentiel est limité, une fréquence $f_2$ telle que $|f_2 - f_1| > \Delta f$ ne sera jamais retenue comme candidat à l'appariement, donc la clé correspondante $(f_1, f_2, t_2 - t_1)$ ne sera jamais atteinte. Le gain d'information apporté par l'appariement de deux points d'intérêt serait probablement mieux représenté par une différence de fréquence, $(f_1, f_2 - f_1, t_2 - t_1)$, pour éviter les valeurs de clé impossibles à atteindre.

Chacune de ces clés va représenter une entrée de la table de look-up. Comme dans la méthode de Philips, chaque entrée de la table va contenir une liste de références aux objets possédant cette clé. Ici ce seront des couples $(j, t_1)$ où $j$ est le numéro de l'objet, et $t_1$ le temps à l'intérieur de l'objet.

Si on reprend le cas d'une base de données musicale de 10000 morceaux de musique d'une durée moyenne de 5 minutes, et en considérant que le nombre de clés gardées par seconde est en moyenne de 100, on trouve que le nombre d'éléments total est 300 millions, soit légèrement plus que pour la méthode de Philips. Dans les deux cas, les auteurs proposent de représenter les éléments sur 32 bits, ce qui signifie que les tailles de stockage sont à peu près équivalentes. Enfin cela signifie aussi qu'une table de hachage est également souhaitable pour la méthode de Shazam.

#### 8.3.3.2 Comparaison et décision

Pour un flux à identifier, on calcule de la même façon que précédemment les triplets formant les empreintes. On considère que l'on cherche à identifier un buffer de taille fixe. L'identification est réalisée de la façon suivante :

- Pour chaque clé $(f_1^{'}, f_2^{'}, t_2^{'} - t_1^{'})$ contenue dans le buffer
  - Sélection dans la table de look-up des éléments correspondant à cette entrée.
  - Pour chaque éléments $(j, t_1)$ on calcule le décalage temporel $d = t_1^{'} - t_1$. L'objet $j$ est ajouté comme candidat.
- Pour chaque objet $j$ candidat
  - Calcul de l'histogramme des décalages $d$.
  - Si l'histogramme présente un pic supérieur à un certain seuil, l'objet $j$ est considéré comme identifié.

Contrairement à la méthode de Philips, le calcul de la distance se fait directement sur les clés de la table de look-up, la clé ne sert pas seulement à restreindre l'espace de recherche.

### 8.3.4 Discussion

Perceptuellement, ce qui permet de reconnaître un objet sonore fortement altéré, réside essentiellement dans l'information portée par les principales composantes sinusoïdales de cet objet. Bien qu'aucune étude psychoacoustique complète n'existe sur ce sujet, des expériences semble le confirmer [Papaodysseus et al., 2001], [Fragoulis et al., 2001] et [Wang, 2003].

La plupart des méthodes de l'état-de-l'art, basées sur les bancs de filtres, ne tiennent que partiellement compte de l'information sur ces composantes. Un autre défaut récurrent est de tenir compte des parties moins informatives du signal dans leurs empreintes, ce qui a comme effet de les rendre plus fragiles. Concrètement, face au problème de l'ajout de bruit non aléatoire et fortement énergétique, comme la parole par exemple, la plupart de ces méthodes vont échouer à des degrés divers. Prenons le cas d'un évènement audio constitué d'une seule fréquence à 1000 Hz, stationnaire pendant une seconde. Même si dans le flux audio à traiter, viennent se superposer d'autre signaux plus complexe (signature noyée dans du bruit ou dans de la parole par exemple), la méthode d'extraction de composantes sinusoïdales sera toujours capable de retrouver au moins cette fréquence particulière et on pourra espérer identifier la signature.

La méthode de Shazam tient compte de cette information, car elle est grosso modo basée sur les attaques sinusoïdales. Cette méthode est cependant destinée à identifier des objets audio longs. En effet, un objet court, comme un jingle de radio, ne présenterait pas suffisamment d'attaques pour que la mesure utilisée dans ce cas soit suffisamment fiable.

Les avantages d'une méthode basée sur l'analyse sinusoïdale sont donc les suivants :

- Une meilleure modélisation du signal à reconnaître. On se focalise en effet uniquement sur les parties informatives du signal (typiquement les composantes

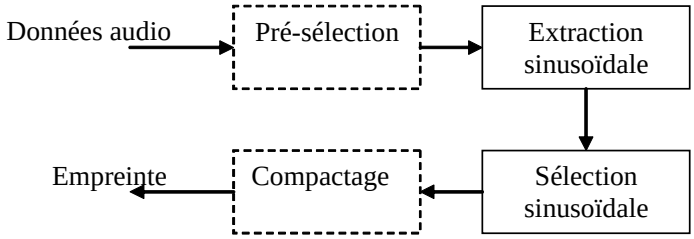



FIG. 8.9: Extraction de l'empreinte sinusoïdale

sinusoïdales de forte énergie) en laissant de coté les parties moins informatives du spectre (typiquement le bruit).
- Une robustesse accrue face à l'ajout de bruit et à la compression. Les composantes sinusoïdales de forte énergie sont en effet peu altérées par ce type de dégradation.
- Une robustesse accrue face à la superposition d'autres signaux.
- Une robustesse accrue face aux phénomènes de filtrage en sous-bandes (procédé d'équalization qui vise à modifier l'enveloppe spectrale des signaux), si comme dans la méthode de Shazam le descripteur utilisé est purement fréquentiels.
- Une meilleure granularité. L'idée est de ne pas se limiter aux seules attaques sinusoïdales comme dans le cas de Shazam.

La méthode de Shazam partage ces avantages, sauf le dernier.

## 8.4 Description d'un système d'empreinte basé sur la modélisation sinusoïdale

Dans cette section, nous allons présenter un système dédié à l'audio ID basé sur une empreinte directement extraite à partir de l'analyse sinusoïdale. On s'est attaché à développer une méthode qui soit suffisamment flexible pour s'adapter à toute sorte d'objets sonores de nature et de tailles différentes, et à toutes sortes de tâches liées à l'identification audio. Le schéma de fonctionnement général est celui des techniques basées sur l'extraction d'empreinte. La seule différence notable est que la création des empreintes va se faire de façon différente pour la phase d'entraînement et la phase d'identification (Figure 8.1). Cela va permettre d'obtenir de meilleures signatures lors de la phase d'entraînement, tout en gardant une faible complexité pour la phase d'identification.

### 8.4.1 Création de l'empreinte lors de la phase d'entraînement

Le calcul de l'empreinte, représenté sur la Figure 8.9, est un module d'analyse sinusoïdal constitué de deux principaux éléments et de deux éléments optionnels.

D'abord un module d'extraction sinusoïdal estime les paramètres sinusoïdaux d'un objet audio, puis un module sélectionne les composantes sinusoïdales les plus pertinentes. Ces deux étapes peuvent également être réalisées simultanément. On a déjà mentionné la possibilité de faire un calcul d'empreinte différent dans la phase d'entraînement et dans la phase de classification. Une version allégée du calcul d'empreinte peut être préférable lors de la phase de classification pour des raisons de rapidité, par exemple en changeant ou en supprimant le module de sélection des composantes.

Les deux autres modules sont optionnels : le premier est un pré-traitement permettant d'éliminer des parties inutiles de l'objet audio pour la reconnaissance. Le dernier module est un post-traitement qui sert à réduire la taille de l'empreinte.

Nous allons maintenant décrire bloc par bloc le système qui a été implanté.

#### 8.4.1.1 Présélection

La présélection est une étape de sélection grossière, se déroulant avant l'analyse sinusoïdale. Le seul critère retenu est un filtrage passe-bas, afin d'être robuste à une éventuelle limitation de bande passante. On ne va sélectionner ici que les composantes sinusoïdales comprises entre 0Hz et 4kHz. En pratique on a choisit de sous-échantillonner le signal à 8kHz, car cela permet d'avoir un gain de complexité pour le calcul de la FFT. On note $F_c = 4$ kHz la fréquence de coupure du signal filtré.

#### 8.4.1.2 Analyse sinusoïdale

L'analyse sinusoïdale est réalisée selon le principe décrit à la section 3.3. La méthode d'estimation de fréquence utilisée est, comme dans le cas de l'estimation de pitch, la méthode à deux bins de la section 6.1.1.1. Ici, plus que pour le pitch, une estimation précise des composantes est superflue, à cause des déformations potentiellement importantes que le signal peut subir. On a donc encore une fois privilégié un algorithme rapide, même si moins précis. On rappelle que la première étape de l'analyse sinusoïdale fait intervenir le calcul d'une TCFT de taille $P$, avec une fenêtre $h$ de taille $N \leq P$.

#### 8.4.1.3 Sélection des composantes pertinentes et création de la signature

Parmi toutes les composantes sinusoïdales trouvées, on sélectionne celles qui vont constituer la signature de l'objet sonore à modéliser. Ce module joue un rôle important dans les performances du système d'audio ID complet puisque les composantes sinusoïdales retenues doivent à la fois être les plus représentatives de l'évènement à modéliser, tout en étant robustes aux dégradations éventuelles du signal. Le nombre de composantes retenues doit également être suffisement faible pour permettre une comparaison rapide.

Pour simplifier la création de la signature, on a choisi de découper les objets en bloc de taille $N_t$ trames, au plus. Ce découpage en bloc est arbitraire et, à condition que $N_t$ soit suffisamment grand, il n'affecte que très peu la création de la signature.

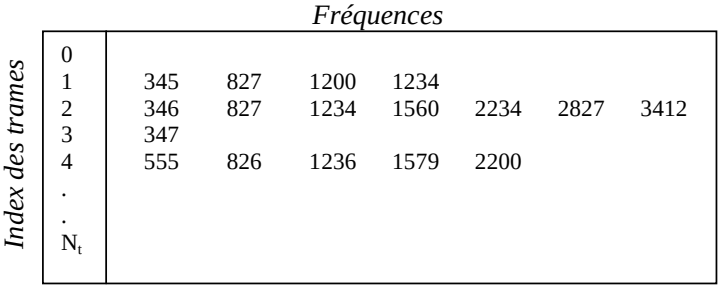



FIG. 8.10: Extrait d'une empreinte sinusoïdale

Ces blocs peuvent être considérés comme des objets indépendants, possédant des identifiants identiques s'ils appartiennent au même objet.

Afin d'être robuste à l'ajout de bruit, on réalise un premier filtrage basé sur l'énergie des composantes sinusoïdales extraites. Pour ce faire, on simule l'ajout d'un bruit blanc par un seuil $A_l$ et on ne garde que les sinusoïdes dont l'amplitude est supérieure à ce seuil[3] :

$$A_l = \alpha P_s 10^{\frac{Z}{10}} \tag{8.4.1}$$

où $P_s$ est la puissance du signal, $\alpha = \frac{2\sum_{i=0}^{N} h(i)^2}{|H(F_c/P)|}$ est une constante qui ne dépend que de la fenêtre $h$. $H(F_c/P)$ est la TF de $h$ pour la fréquence $F_c/P$. Enfin $Z$ est une constante qui indique le niveau de fiabilité requis pour les pics. Par exemple si on veut garder les sinusoïdes qui ressortent d'au moins 10 dB d'un bruit correspondant à un SNR (signal-to-noise ratio) de 20dB on va choisir $Z = 10 - 20 = -10$dB. Le seuil va en fait s'adapter automatiquement au SNR mesuré sur un horizon de quelques secondes.

L'analyse des fréquences va se faire avec un pas de $T_h = 16$ ms. Il pourra donc y avoir un décalage temporel fixe entre l'analyse faite à l'entraînement et l'analyse réalisée lors de l'identification. Afin d'être robuste à ce décalage, on réalise deux autres analyses sinusoïdales avec un décalage temporel égal respectivement à $+.5T_h$ et $-.5T_h$ et on ne garde que les composantes qui restent à l'intérieur d'une tolérence fréquentielle $T_f$.

Afin d'être robuste à l'ajout de bruit et afin de limiter le nombre de composantes présentes dans la signature, on réalise une seconde sélection de pics basée sur un critère énergétique. Pour un bloc donné, on ne gardera que les $M$ pics les plus énergétiques, parmi les pics restant. On rappelle que les objets sont découpés en blocs de taille $N_t$ fixe. Or dans la plupart des cas la longueur de ces objets n'est pas un multiple de $N_t$, donc leur dernier bloc sera de taille inférieur à $N_t$. Comme tous les blocs n'ont pas la même taille, on préférera calculer $M$ à partir d'une densité de pics par seconde $D_p$ fixée : $M = D_p T_b$, où $T_b$ est la taille en secondes du bloc en question. $T_h$ est le décalage temporel entre deux trames, donc on aura $T_b = T_h N_t$.

[3]Voir l'annexe E pour une démonstration.

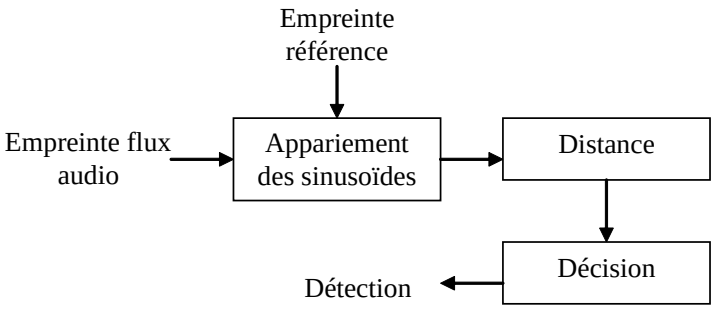



FIG. 8.11: Comparaison des empreintes sinusoïdales

La signature sera ensuite formée des seules fréquences des composantes sinusoïdales sélectionnées. Ne garder que l'information de fréquence nous permettra, comme dans le cas de l'algorithme de Shazam[4], d'être robuste aux phénomènes de filtrage en sous-bandes (equalization) et aux bruits additifs. La signature est alors constituée d'un vecteur de réels.

#### 8.4.1.4 Compaction de la signature

Dans le but de réduire l'espace de stockage et afin d'accélérer le traitement lors de la phase de comparaison, la signature va être compactée. Les fréquences sont quantifiées et codées sur $B = 12$ bits. Pour une fréquence de coupure de $4kHz$, un codage sur 12 bits permet d'avoir une précision fréquentielle de l'ordre de 1Hz. La signature finale est ainsi constituée d'un vecteur de $M$ mots de $B$ bits. Un exemple de signature décodée est présenté sur la Figure 8.10. Cette signature est composé de $N_t$ trames et est constituée à la trame 4 des fréquences 555Hz, 826Hz, 1236Hz, 1579Hz et 2200 Hz.

### 8.4.2 Création de l'empreinte lors de la classification

Le procédé d'extraction des composantes sinusoïdales est le même que celui décrit au paragraphe précédent, à l'exception du module de sélection. Pour un bloc de longueur $T_b^{'} = T_h.N_t^{'}$, le module de sélection de la Figure 8.9 se réduit ici à garder les $D_p^{'}T_b^{'}$ pics sinusoïdaux de plus grande amplitude, parmi les composantes extraites. $D_p^{'}$ est la densité de pics par seconde du signal à identifier. On va choisir $D_p^{'} > D_p$, afin de permettre la reconnaissance même dans le cas d'une perturbation sinusoïdal additive très intense.

### 8.4.3 Module de comparaison des empreintes

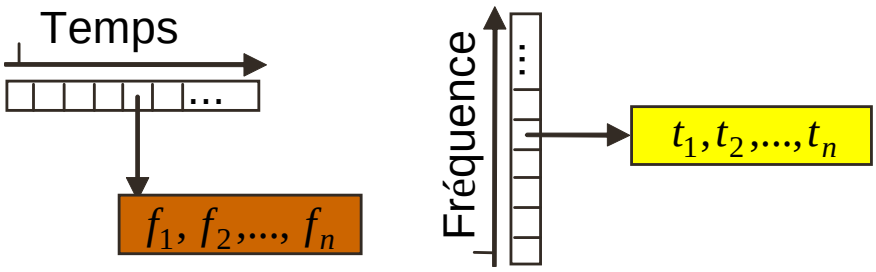
Le principe du module de comparaison est représenté sur la Figure 8.11. Il reprend les principaux éléments du module de comparaison de l'algorithme d'estimation de pitch, à savoir l'appariement des sinusoïdes et l'utilisation d'une mesure de similarité[5]. L'empreinte de référence a été calculée selon la méthode de la section 8.4.1 et l'empreinte à identifier selon la section 8.4.2. Comme dans la méthode de Pinquier, on va comparer le bloc à identifier de taille $T_b^{'}$ à tous les sous-blocs de taille $T_b^{'}$ du bloc de référence. $T_b^{'}$, la taille des blocs à identifier, sera généralement très inférieure à $T_b$, la taille des blocs des objets de référence[6]. Typiquement on peut prendre $T_b^{'} = 1$ s et $T_b = 30$ s.

Le module de comparaison se compose de trois sous-modules :

- Le premier sous-module d'appariement sélectionne chaque composante sinusoïdale de l'empreinte de référence, et cherche la plus proche fréquentiellement parmi les composantes de l'empreinte à identifier. Elle utilise une tolérance fixe $T_f = 3$ Hz sur l'écart de fréquence.
- La mesure de similarité est une mesure très simplifiée pour permettre un calcul rapide. Il s'agit du nombre de composantes sinusoïdales trouvées $K$, normalisé par le nombre de pic moyen dans l'objet de référence, $D_j.T_b^{'} : L = \frac{K}{D_j.T_b^{'}}$. Cette normalisation permet de favoriser les parties de la référence qui ont un nombre de pics supérieur à la moyenne et qui sont donc plus fiables.
- La mesure de similarité, calculée à des instants successifs va présenter un pic comme dans le cas de la mesure de similarité de Pinquier. La décision sera prise de la même manière, en utilisant le double seuillage décrit à la section 8.3.1.

### 8.4.4 Accélération de la comparaison

Nous allons nous pencher plus en détail sur la mesure de similarité. En effet si la comparaison était réalisée bloc par bloc et trame par trame comme dans la méthode de Pinquier (Figure 8.3), le temps de calcul serait prohibitif, sauf pour des bases de données contenant très peu d'objets. Deux améliorations pour la comparaison ont été envisagées. La première amélioration, décrite à la section 8.4.4.1 est une représentation différente des objets, strictement équivalente, c'est à dire requérant la même taille de stockage et permettant un gain de temps considérable dans le cas d'un calcul exhaustif. La deuxième amélioration, présentée à la section 8.4.4.2, est l'utilisation d'une table de look-up pour restreindre l'espace de recherche, de façon similaire à l'algorithme de Philips[7]. Cette méthode n'est pas exhaustive, mais le temps de calcul va rester presque constant en fonction de la taille de la base de données.




(a) Indexation par les fréquences (b) Indexation par les temps

FIG. 8.12: Représentations temporelles et fréquentielles

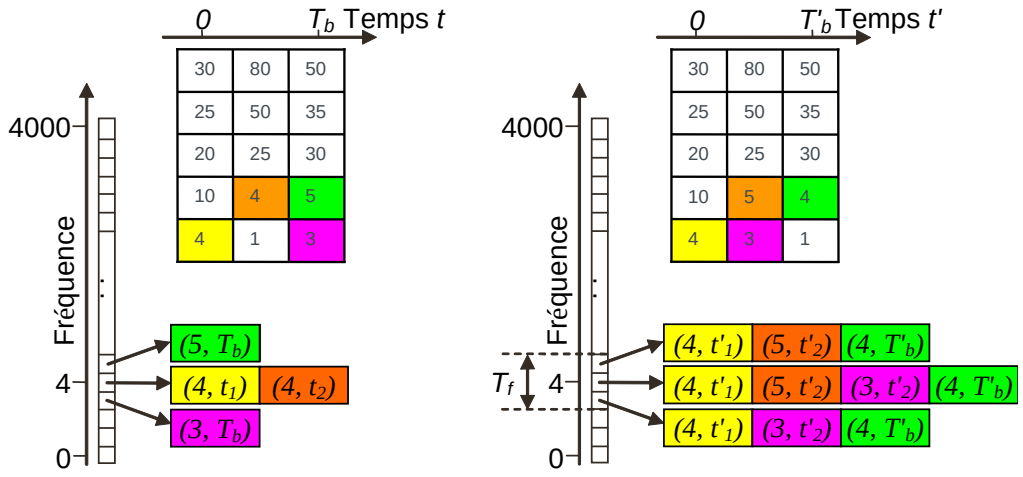
#### 8.4.4.1 Comparaison par blocs de trames et d'objets

Dans la méthode de comparaison bloc par bloc, trame à trame de la Figure 8.3, les pics sinusoïdaux sont indexés d'abord temporellement puis fréquentiellement, comme montré sur la Figure 8.12(a). C'est la représentation utilisée dans l'algorithme de Pinquier, mais aussi de Philips. Pour les pics sinusoïdaux, qui sont caractérisés par un couple de valeurs (temps,fréquence), une autre représentation tout aussi naturelle est d'indexer les pics d'abord par fréquence puis temporellement, comme sur la Figure 8.12(b). La représentation est strictement équivalente, et la taille de stockage quasiment identique. Comme la précision sur les fréquences est limitée par le nombre de bits $B$ utilisés pour représenter les fréquences, le nombre d'entrées de la table des fréquences est de $2^B$. Dans le cas où $B = 12$ bits la table possède 4096 entrée, c'est à dire que les fréquences auront une précision d'à peu près 1Hz pour une bande passante de 4kHz.

Pour tenir compte de la tolérance $T_f$ sur les fréquences, les pics du bloc à identifier comme celui de la Figure 8.13(b), vont être dupliqués dans la table des fréquences correspondante : un élément $(t, f)$ est ajouté à toutes les entrées $f^{'}$ de la table des fréquences telles que $|f - f^{'}| \leq T_f$. Si on considère, comme dans l'exemple de la Figure 8.13(b), que les entrées de la table des fréquences sont les fréquences entières de 0 à 4000 Hz, et que la tolérence $T_f$ est de 3 Hz, alors un élément dont la fréquence est égale à 4 Hz sera ajouté aux entrées 3, 4 et 5 de la table des fréquences. On pourrait choisir de dupliquer les pics des objets de référence, mais ce choix ne serait pas judicieux, car les blocs à identifier sont de taille fixe et beaucoup plus petits que l'ensemble des objets de référence ! Les pics de la table de fréquence de l'objet de référence numéro $r$ ne sont donc pas dupliqués (Figure 8.13(a)).

Une fois la duplication réalisée, la comparaison entre les deux tables de fréquences se fait de la manière suivante :

[4] Décrit à la section 8.3.3.
[5] La section 7.3 décrit le module de comparaison de l'algorithme de suivi de pitch.
[6] Voir le paragraphe sur la sélection de composantes, section 8.4.1.3.
[7] Cet algorithme est décrit à la section 8.3.2.




(a) Empreinte de référence (b) Empreinte à identifier

FIG. 8.13: Représentation fréquentielle de l'empreinte de référence et de l'empreinte à identifier. Dans la table des fréquences du bloc à identifier, les couples $(f^{'}, t^{'})$ sont dupliqués pour chaque fréquence dans l'intervalle de tolérance $T_f$ (les couples identiques ont la même couleur)

- Pour chaque pic de la table du bloc à identifier (Figure 8.13(b)), on consulte la liste de fréquences correspondante dans la table de l'objet $r$ (Figure 8.13(a)).
- Pour un pic à identifier $(f, t^{'})$ et un pic de référence de même fréquence $(f, t)$, on ajoute $+1$ à la case $t - t^{'}$ de la table de vraisemblance (Figure 8.14).

La table de vraisemblance sert à accumuler les pics trouvés au temps initial du bloc à identifier dans l'objet de référence $r$. Cette table à un rôle similaire à l'histogramme de la méthode de Shazam[8].

La représentation fréquentielle (Figure 8.12(b)) est très importante car elle permet une comparaison beaucoup plus rapide entre les empreintes, par rapport à la représentation temporelle (Figure 8.12(a)). On se propose de comparer maintenant les complexités respectives du calcul de vraisemblance dans les deux méthodes.

### Comparaison des complexités des méthodes fréquentielle et temporelle

Le calcul de ces complexités est détaillé dans l'annexe F, on ne donne ici que le résultat final :

$$\begin{aligned} C_t &= (T_b - T_b^{'})T_b^{'}(9D_p + 4D_p^{'})/T_h \\ C_f &= 9T_bT_b^{'}N_qD_pD_p^{'}2^{-B} \end{aligned}$$

où $C_t$ est la complexité de la méthode d'indexation temporelle et $C_f$ celle de la méthode d'indexation fréquentielle. On rappelle que $T_b$ et $T_b^{'}$ sont les longueurs des

[8]Voir le fonctionnement de la comparaison dans la méthode de Shazam, section 8.3.3.2.

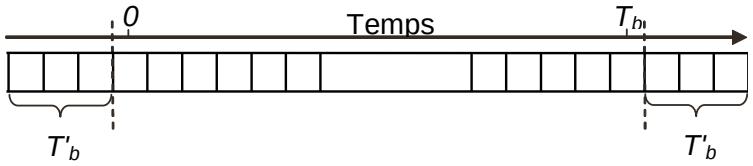



FIG. 8.14: Table utilisée pour l'accumulation des vraisemblances. Elle est utilisée en entrée du module de décision.

blocs de trames en secondes, $D_p$ et $D_p{}'$ sont les densités de pics par secondes, $T_h$ est la longueur d'une trame en secondes et $2^B$ est le nombre de valeurs quantifiées pour les fréquences. Enfin on a introduit $N_q = (2T_f 2^B/F_c + 1)$ l'intervalle de tolérance fréquentielle mesurée en nombre de valeurs quantifiées. $N_q$ correspond au nombre de duplications de fréquence dans la table de l'objet à identifier. Ces complexités sont données en nombre d'additions entières, dans le cas où les fréquences sont réparties uniformément sur l'ensemble des $2^B$ valeurs possibles. La méthode fréquentielle est en $2^{-B}$, valeur généralement très petite, ce qui explique son avantage sur l'autre méthode.

### Cas d'un calcul exhaustif

Dans le cas d'un calcul exhaustif, on teste l'ensemble les objets de référence, on a donc $T_b \gg T_b^{'}$. Le rapport des complexités s'exprime donc ainsi :

$$\frac{C_t}{C_f} = \frac{(D_p + \frac{4}{9}D_p^{'})2^B}{N_q D_p D_p^{'} T_h} \tag{8.4.2}$$

Dans un cas typique, on prendra $B = 12$, $D_p = 300$ pics/s, $D_p^{'} = 2D_p$, $T_h = 0.016$ s et $N_q = 2 \cdot 3 \cdot 2^{12}/4000 + 1 = 7$. Le rapport de complexité est alors un facteur 1000.

### Cas d'un calcul non exhaustif

Dans la section suivante nous présenterons une méthode qui permet de réduire l'espace de recherche. Dans cette méthode, la comparaison se fait localement dans l'objet à identifier et non plus sur l'ensemble de l'objet. Si on suppose que le module de décision nécessite $V$ calculs de vraisemblance, la comparaison va se faire entre un bloc de $T_b^{'}$ secondes et un bloc de $T_b = T_b^{'} + VT_h$ secondes. Le rapport de complexité devient alors :

$$\frac{C_t}{C_f} = \frac{VT_h}{T_b^{'} + VT_h} \frac{(D_p + \frac{4}{9}D_p^{'})2^B}{N_q D_p D_p^{'} T_h} \tag{8.4.3}$$

Comme $VT_h$ est petit par rapport à $T_b^{'}$, l'avantage de la méthode sera moins marqué. Pour l'exemple précédent et en prenant $V = 15$ et $T_b^{'} = 1$s, on trouve un rapport de complexité égal à 200. Dans les deux cas de figure envisagés la représentation fréquentielle est donc préférable.

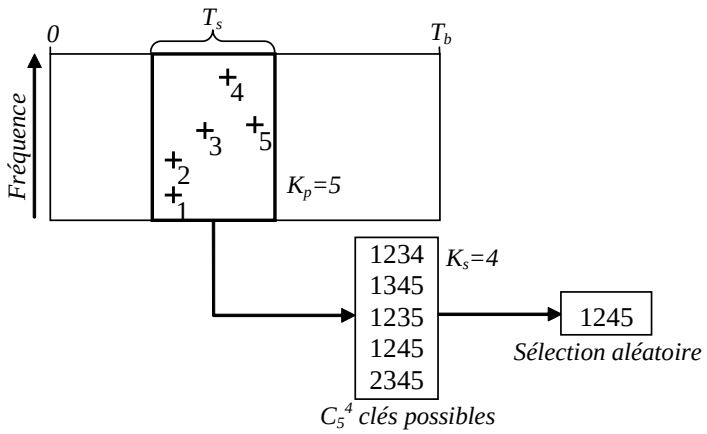



FIG. 8.15: Création des clés sinusoïdales, pour la réduction de l'espace de recherche

**Passage de la représentation temporelle à la représentation fréquentielle**

Même si la représentation fréquentielle est plus efficace pour le calcul, il peut s'avérer nécessaire de stocker les empreintes en utilisant la représentation temporelle. En effet, la méthode présentée dans la section suivante permet de réduire l'espace de recherche dans le domaine temporel uniquement. Il va donc falloir passer d'une représentation temporelle à une représentation fréquentielle. La complexité requise pour passer de l'une à l'autre est de $8T_bD_p$ additions entières pour un bloc de taille $T_b$ secondes et une densité de $D_p$ pics par seconde[9]. Ce coût est donc négligeable par rapport au coût de la comparaison.

### 8.4.4.2 Table d'index des trames

Dans les deux méthodes de comparaison décrites dans la section précédente 8.4.4.1 et dans la section 8.3.1.2, l'algorithme était exact. Dans le cas de larges bases de données, il est préférable de faire une présélection afin de réduire le nombre de candidats potentiels. L'algorithme ne sera alors plus exact, mais réduira considérablement le temps de calcul.

Comme dans le cas de la méthode de Philips, nous allons utiliser une table de look-up. Chaque clé va être calculée à partir des composantes sinusoïdales. Une fréquence est trop peu informative pour constituer une clé, on va donc associer les fréquences de plusieurs pics, un peu comme dans la méthode de Shazam. Le mécanisme de création des clés est illustré à la Figure 8.15. Appelons le nombre de fréquences constituant l'empreinte $K_s$ et considérons que lors de la création des signatures le nombre de pics par seconde est de $D_p$ pics en moyenne. On accumule alors les pics sur un bloc de trames de taille $T_s$ tel que le nombre de pics moyen sur ce bloc soit supérieur à $K_s$, c'est à dire $K_s < T_sD_p$.

[9]Voir l'annexe F pour plus de détails.

Les pics sont numérotés de façon croissante suivant leur temps dans le bloc, puis selon leur fréquence le cas échéant, comme indiqué sur la Figure 8.15. On forme alors toutes les combinaisons possibles de $K_s$ pics, ordonnés par ordre croissant. Comme $T_s D_p$ est une valeur moyenne, pour certains blocs le nombre effectif $K_p$ de pics présents sera inférieur à $K_s$, pour d'autres il sera supérieur. On a donc deux cas de figure :

- $K_p < K_s$ : aucune combinaison n'est possible. Ce bloc ne possède pas de clé.
- $K_p \geq K_s$ : il y a $C_{K_p}^{K_s}$ combinaisons possibles. On retient une de ces combinaisons prise au hasard.

Ne retenir qu'une seule combinaison permet de ne pas surcharger la table de look-up et la prendre au hasard permet de mieux répartir les clés présentes dans la table.

La clé alors formée est composée des $K_s$ fréquences correspondantes codées chacune sur $B_s$ bits, $B_s$ pouvant être inférieur à $B$ le nombre de bits utilisé pour coder les fréquences dans les empreintes. Si on suppose que $B_s$ et $B$ peuvent être différents, cela veut simplement dire que l'on s'offre la possibilité de sous-échantillonner les fréquences à la volée, de façon à garder une taille de clé raisonnable.

Lors de la phase d'identification, le processus est similaire. On considére également des blocs de taille $T_s$ secondes, et on garde les $D_p^{'} T_s$ meilleurs pics. Les pics sont encore ordonnés de façon croissante suivant leurs temps dans le bloc, puis selon leur fréquence. Enfin on forme toutes les suites croissantes possibles de $K_s$ pics. La seule différence est que chacune des clés ainsi formées sera gardée et testée. Chaque clé va pointer sur la liste des blocs qui la possèdent. Une fois qu'une clé est trouvée dans le signal à identifier, on va faire un calcul exhaustif selon la méthode de la section 8.4.4.1 pour tous les blocs correspondant. Ici les clés ne servent qu'à réduire l'espace de recherche, comme dans la méthode de Philips.

Dans la pratique on va utiliser des clés contenant $K_s = 4$ fréquences et codées sur 32 bits, c'est à dire avec $B_s = 32/4 = 8$. Pour de gros volumes de clés, nous utiliserons encore une fois une table de hash. L'algorithme utilisé dans la section expérimentale s'attache à détecter des évènements très courts, de l'ordre d'une seconde. Pour que l'identification soit efficace, le nombre de pics gardés par trame sera très élevé, de l'ordre de 5 pics par trame, donc dans ce cas, les clés seront formées simplement par 4 fréquences prises au hasard sur la trame, ordonnées par ordre croissant.

### 8.4.5 Aspects calculatoires

Dans cette section, nous allons comparer la complexité et la taille en mémoire requises par notre algorithme, avec les deux autres algorithmes présentés dans la section état-de-l'art, les méthodes de Philips et de Shazam.

#### 8.4.5.1 Mémoire

Nous allons d'abord comparer la taille requise en mémoire pour notre méthode par rapport aux deux autres méthodes. Le nombre moyen de pics par seconde est toujours noté $D_p$, et le nombre de bits utilisés pour coder une fréquence, $B$. Donc la taille requise pour un bloc de 1 seconde est de $BD_p = 12 \times 300 \approx 3600$ bps (bits par

seconde). La méthode de Philips utilise des trames codées sur 32 bits, avec un pas d'avancement de 11.6 ms [Haitsma and Kalker, 2002], c'est à dire que le flux est égal à $32/0.0116 = 2800$ bps. La méthode de Shazam quand à elle réclame 32 bits pour un couple de fréquence. Nous avons vu à la section 8.3.3 que le nombre moyen de couples par seconde est de de l'ordre de 100, c'est à dire un flux de $100 \times 32 = 3200$ bps. On considère que chaque méthode utilise une table de look-up identique, avec une clé de taille 32 bits. La méthode de Shazam ne requiert aucune mémoire complémentaire, car chaque élément est stocké directement par des mots de 32 bits dans la table. Notre méthode et celle de Philips, vont utiliser des pointeurs (32 bits) vers respectivement le bloc correspondant à la clé, et la trame correspondant à la clé. Notre méthode ainsi que celle de Philips réclameront donc environ deux fois plus de mémoire que la méthode de Shazam dans ce cas de Figure.

La densité de pics dans les objets de référence utilisée par notre algorithme est de $D_p = 300$ pics par seconde, et correspond à la détection d'évènements de longueur $T_b^{'} = 1$ seconde. Cependant dans la plupart des cas, comme l'identification de morceaux de musique, les évènements à identifier sont plus longs. Or la qualité de l'estimation réclame un nombre constant de pics présents quel que soit la taille de l'évènement $T_b^{'}$ à détecter, de l'ordre de quelque centaines de pics. Par exemple pour un évènement d'une durée de $T_b^{'} = 3$ secondes, la densité pourra être réduite à $D_p/T_b^{'} = 100$ et la taille du flux sera de $12 \times 100 \approx 1200$ bps. Dans ce cas de figure, on voit que notre méthode requière moins d'espace que les méthodes de Philips et de Shazam. Elle en demandera d'autant moins que la taille de l'évènement à comparer $T_b^{'}$ est longue.

#### 8.4.5.2 Complexité

Pour la méthode de Philips, la comparaison réclame $C_p = O(T_b^{'}/T_{h_p})$ opérations, où $T_{h_p}$ est la taille d'une trame pour la méthode de Philips, et pour la méthode de Shazam $C_s = O(T_b^{'} D_s)$ opérations où $D_s$ est le nombre de clés retenues par seconde, typiquement 100. On rappelle que la complexité de notre algorithme est[10] $C_f = O(T_b^{'} T_b D_p D_p^{'} N_q 2^{-B})$.

La comparaison n'est pas évidente, car elle va dépendre des réglages choisis. Pour donner un ordre idée, on se propose de les comparer dans leur configuration la plus typique, pour un bloc $T_b^{'} = 1$s de signal. En posant $T_{h_p} = 0.0116$, $D_s = 100$, $T_b = T_b^{'} + VT_h = 1.24$, $D_p = 300$, $D_p^{'} = 2D_p$, $N_q = 7$, $B = 12$, on obtient les complexités suivantes : $C_f \propto 381$, $C_p \propto 86$, $C_s \propto 100$. La complexité de l'algorithme sinusoïdal est donc légèrement plus importante dans ce cas de figure.

On rappelle que la complexité globale dépend de la qualité de la réduction de l'espace de recherche par la table de look-up qui dépend elle-même du taux de remplissage de cette table. Si les clés sont suffisamment bien réparties, le taux de remplissage maximum d'une entrée de la table sera très faible, inférieur à 1 [Haitsma and Kalker, 2002]. La différence de complexité d'une comparaison de bloc ne sera donc pas critique.

[10] Voir la section 8.4.4.1 et l'annexe F pour plus de détails.

| Nom du corpus | Description | Durée totale | Nombre d'occurrences |
|---|---|---|---|
| FI_AM | France Info en AM | 18h | 243 |
| FI_AM+MP3 | corpus AM et compression | 18h | 243 |
| FI_AM+PAR | corpus AM et ajout de parole | 18h | 243 |
| RS | Skyrock et RFM | 48h | 0 |

TAB. 8.1: Description du corpus de test utilisé pour l'identification de jingles

## 8.5 Evaluation expérimentale

Nous allons essayer de mettre en évidence la robustesse des descripteurs sinusoïdaux, déjà partiellement constatée dans la méthode de Shazam. La tâche principale sur laquelle notre algorithme a été évalué est une tâche d'identification de jingles radiophoniques, présentée à la section 8.5.1. Cette tâche est particulièrement utile pour la structuration de documents sonores. La méthode sera comparée avec celle de Philips.

La capacité de charge de la méthode a été également évaluée indépendamment sur une tâche d'identification musicale à la section 8.5.2.

### 8.5.1 Identification de jingles

#### Tâche et corpus

Le système est évalué sur une tâche de détection de jingles radiophoniques. La tâche consiste à détecter 30 extraits de jingles de la radio France Info. Les jingles à détecter ont une longueur qui varie de 3 à 10 secondes. Leur longueur n'est pas déterminante pour la reconnaissance car les algorithmes ont été réglés pour détecter des évènements d'une longueur d'une seconde ($T_b^{'} = 1$ s). Le corpus d'entraînement pour la création des empreintes est composé d'un exemple de chaque jingle enregistré en qualité FM. Le corpus de développement utilisé pour le réglage des paramètres est composé de 15 extraits d'une minute de France Info enregistrés séparément, chacun contenant un jingle.

Le corpus de test est composé de 18 heures de France Info enregistré en qualité AM[11] (FI_AM). Un total de 243 occurrences de jingles sont présents dans le corpus. Parmi eux, 33 correspondent à des versions courtes des jingles. Dans les programmes radio, ces jingles courts sont utilisés pour annoncer des nouveaux sujets par exemple. Ce sont généralement des fragments d'une seconde environ de leur version plus longue. Pour tester la robustesse des algorithmes aux bruits, ce corpus a été altéré en utilisant deux types de distorsions supplémentaires : compression à 16 kb/s (FI_AM+MP3) et de l'addition de parole avec un SNR de 0 dB (FI_AM+PAR). Nous avons également ajouté 48 heures provenant de deux autres radios musicales, RFM et skyrock, pour

[11] Le corpus d'apprentissage est en FM. Le passage du FM au AM s'accompagne d'une réduction de bande ainsi que d'un fort bruit blanc.

| | Sinusoidal | Philips |
|---|---|---|
| Fréquence d'échantillonnage | 8000 | 8000 |
| Taille de trame | 512 | 4096 |
| Décalage entre 2 trames | 128 | 96 |

TAB. 8.2: Comparaison des paramètres

tester les fausses alarmes (RS). Les caractériques de ces différents corpus sont résumés dans la table 8.1.

On souhaite ici évaluer la qualité des empreintes respectives des deux algorithmes. Le calcul est ici fait de manière exhaustive, c'est à dire que la réduction de l'espace de recherche par table de look-up est désactivé.

### Paramètres

Dans la table 8.2, on résume les paramètres utilisés par les deux algorithmes. "Sinusoidal" se réfère à l'algorithme présenté dans la section 8.4 et "Philips" à l'algorithme de Philips décrit à la section 8.3.2.

Les deux méthodes utilisent des paramètres FFT très différents : la méthode de Philips utilise des transformées de Fourier longues avec un fort recouvrement, tandis que la méthode sinusoïdale utilise des transformées de Fourier courtes et un recouvrement plus faible. La comparaison va se faire sur des blocs d'une seconde, soit $8000/128 = 62$ trames pour l'algorithme sinusoïdal et $8000/96 = 83$ trames pour l'algorithme de Philips.

Les autres paramètres ont été optimisés sur le corpus de développement. La tolérance $T_f$ a été réglée à 3 Hz, comme suggéré à la section 8.4.3. L'algorithme de Philips utilise 33 bandes de fréquences réparties uniformément sur une échelle Bark, et le seuil sur le BER a été fixé à 0.25, ce qui est la valeur suggérée dans [Haitsma and Kalker, 2003].

### Résultats

Les algorithmes sont comparés en terme de rappel/précision. Néanmoins, comme les deux algorithmes n'ont présenté aucune fausse alarme, la précision a été omise, étant dans tous les cas égale à 100%. Deux mesures de rappel différentes sont utilisées. Le rappel en terme d'occurrences est le nombre de jingles correctement détectés divisé par le nombre de jingles total présent dans le corpus (table 8.3). Le rappel en terme de durée est la longueur totale des blocs correctement détectés, divisés par la longueur totale des jingles dans le corpus de test (table 8.4).

Une différence significative apparaît entre le rappel d'occurrence et le rappel de durée. Comme chaque jingle est d'une longueur d'au moins quelques secondes, il y a toujours une forte probabilité de trouver au moins un des blocs correspondant à une occurrence de jingle.

Pour une perturbation AM seule, les deux algorithmes ont des rappels élevés, avec un avantage pour l'algorithme sinusoïdal. Les erreurs restantes pour l'algorithme

| | FI_AM | FI_AM+MP3 | FI_AM+PAR |
|---|---|---|---|
| Sinusoidal | 97 | 95 | 83 |
| Philips | 89 | 85 | 67 |

Tab. 8.3: Rappel en terme d'occurrence (%)

| | FI_AM | FI_AM+MP3 | FI_AM+PAR |
|---|---|---|---|
| Sinusoidal | 79 | 68 | 53 |
| Philips | 60 | 57 | 34 |

Tab. 8.4: Rappel en terme de durée (%)

sinusoïdal, et dans le cas de la mesure d'occurrence, proviennent des versions courtes des jingles. Certains d'entre eux sont trop courts pour permettre une détection fiable (inférieurs à 1 s). Les deux algorithmes sont robustes à une forte compression mp3, en terme d'occurrence. Comme nous l'attendions[12], l'algorithme sinusoïdal offre de meilleurs performances pour l'ajout de parole par rapport à l'algorithme de Philips. Le rappel de durée décroît de façon très significative pour cet algorithme.

### 8.5.2 Identification de morceaux de musique

Un test de charge sur l'algorithme complet, avec réduction de l'espace de recherche, a été effectué sur une base de données musicales contenant 586 morceaux de musique populaire anglaise d'une durée moyenne de 5 minutes. L'expérience est schématisée sur la Figure 8.16. Des extraits de morceaux contenus dans la base de données sont enregistrés et analysé en temps réel. La durée totale de signal à analyser est de 200 secondes. On va chercher ici à identifier 200 blocs d'une seconde chacun.

Le signal, qui passe à travers un micro de qualité médiocre, est fortement bruité. On constate notamment un fort rehaussement des hautes fréquences, un filtrage des basses fréquences, plus de bruit blanc et des composantes sinusoïdales moins nettes.

La taille en mémoire requise par l'algorithme est de 550 Mo. Sur un ordinateur avec une mémoire vive totale de 1.5 Go, le nombre maximum de morceaux de musique que l'on peut traiter est donc d'environ 1800, en supposant que les morceaux ont une taille moyenne identique. Comme on l'a déjà remarqué à la section 8.4.5.1, ce faible nombre est dû à la brièveté des évènements à détecter (1 s). En augmentant la taille des évènements à détecter, on peut réduire la densité de pics et donc la taille en mémoire de la base de données.

Sur un pentium cadencé à 1700 MHz, le temps de calcul effectif est de 6 secondes au total pour 200 secondes de signal à analyser. On a constaté que la majorité du temps de calcul était dû à des clés trop fréquentes dans la base de données. On envisage deux solutions à ce problème :

[12] Voir la discussion à la section 8.3.4.

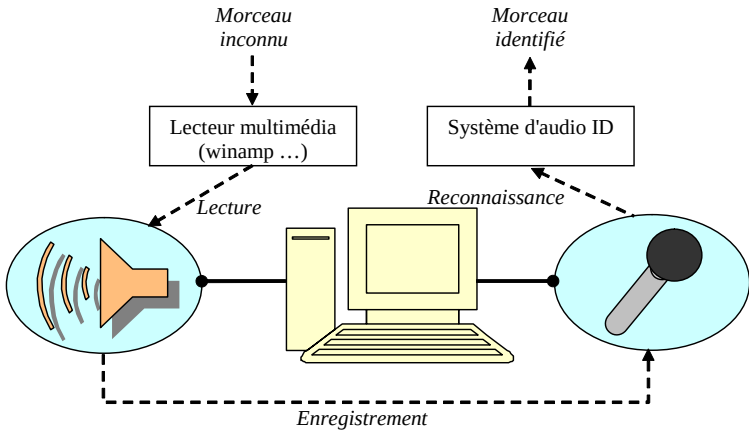



FIG. 8.16: Identification des morceaux de musique en temps réel

– une première consiste à mieux répartir les fréquences retenues sur le plan temps-fréquence, en utilisant par exemple un seuil différent par bande de fréquences comme pour l'algorithme de Shazam [13].
– la deuxième solution est d'augmenter la taille des blocs à identifier. En effet la brièveté des blocs dans l'expérience (1 s) engendre une forte redondance dans les fréquences et donc dans les clés calculées.

Le rappel en terme de durée sur ces 200 secondes est de 44 %. La déformation la plus proche étudiée dans la section précédente est celle de l'expérience AM+MP3, avec un score de 68%. Le score diminue fortement dans cette expérience. Tout d'abord les déformations sont plus marquées que dans l'expérience AM+MP3. De plus certaines introductions de morceaux de musique présentes durant ces 200 secondes à identifier sont apparues être purement rythmiques, donc sans présence de composantes sinusoïdales. Si on retire ces zones du calcul, on trouve un rappel de 52%, un rappel plus proche des résultats précédents. La plupart des erreurs supplémentaires sont dues au procédé de réduction de l'espace de recherche, ce qui est un phénomène déjà observé dans le cas l'algorithme de Philips [Haitsma and Kalker, 2002].

## 8.6 Conclusion

Dans cette partie nous avons présenté un algorithme dédié à l'identification audio et basé sur l'estimation sinusoïdale. L'algorithme d'identification est basé sur une méthode similaire à celle de l'algorithme d'estimation de pitch, mettant en jeu une identification des pics sinusoïdaux et une mesure de vraisemblance simplifiée.

Les avantages d'une empreinte basée sur les pics sinusoïdaux sont de deux ordres par rapport aux empreintes basées sur les sous-bandes, les plus utilisées dans la littérature. Tout d'abord une meilleure modélisation du signal à analyser permet de

[13] Voir la section 8.3.3 décrivant l'algorithme de Shazam.

reconnaître de façon fiable des évènements sonores très courts de l'ordre de 1 s. Deuxièmement, l'empreinte sinusoïdale, basée uniquement sur les fréquences des sinusoïdes, est plus robuste aux distortions, comme la compression et surtout l'ajout de bruit non-aléatoire comme de la parole.

Le principal inconvénient de ce type d'empreinte est son inefficacité pour traiter des données non-sinusoïdales, comme on a pu le constater pour le cas des morceaux rythmiques.

Nous avons vu que la réduction de l'espace de recherche par table de look-up permet d'avoir des performances similaires en terme de taille en mémoire et de complexité par rapport aux algorithmes existants. Des améliorations peuvent être apportées à la clé utilisée pour réduire l'espace de recherche. Une meilleure répartition dans l'espace temps-fréquence des pics utilisés pour créer cette clé est notamment souhaitable.

Parmi les perspectives également envisageables, on pourrait par exemple chercher à adapter l'algorithme d'identification pour tenir compte des étirements temporels possibles.

CHAPITRE 9

# Conclusion générale et perspectives

## Bilan de l'étude

### Analyse sinusoïdale

Le principal apport de cette thèse concerne l'estimation des paramètres sinusoïdaux. Les méthodes étudiées sont basées sur la Transformée de Fourier (TF). Comme les méthodes de la littérature étaient basées sur des modèles différents, la première question concernait la modélisation utilisée. Nous avons montré que pour les signaux usuels, la modélisation couramment utilisée, avec fréquence constante et amplitude constante, était insuffisante.

Nous avons présenté sous un formalisme unifié un grand nombre des méthodes existantes, ce qui a permis de mettre en évidence des liens très forts entre certaines méthodes, principalement pour les méthodes d'estimation de fréquence basées sur la phase. Une comparaison complète de toutes ces méthodes a été réalisée. Nous avons notamment évalué la méthode d'interpolation de la TF, appelée QIFFT, et les méthodes d'estimation de variation de fréquence, méthodes n'ayant jamais été comparées aux autres estimateurs. Enfin, nous avons étudié en détail les performances des estimateurs classiques pour des modulations d'amplitude et de fréquence, et nous avons mis en évidence la faible robustesse de la plupart des méthodes.

Un certain nombre de mécanismes communs à tous les estimateurs basés sur la transformée de Fourier ont été constatés. Nous avons tout d'abord montré que ces estimateurs, de façon similaire à la méthode du maximum de vraisemblance, cherchaient tout d'abord une expression indépendante des paramètres linéaires, ici la phase et l'amplitude initiale. Ensuite, nous avons remarqué que les méthodes cherchaient à inverser cette fonction des paramètres d'ordre supérieur. Afin de rendre cette inversion plus simple, la plupart des méthodes proposent de linéariser cette fonction en utilisant des approximations, comme des approximations de Taylor. Nous

avons également esquissé un algorithme général d'inversion des fonctions, qui éviterait d'avoir recours à une formule analytique approchée. Cette méthode constituerait une alternative à l'optimisation multidimensionnelle des méthodes du type maximum de vraisemblance. Dans le cas général cet algorithme semble cependant malaisé à mettre en oeuvre. Enfin nous avons présenté une méthode alternative au développement asymptotique pour dériver les variances des estimateurs. Cette méthode est particulièrement utile pour le cas des modèles qui varient en amplitude et qui ne sont pas bornés asymptotiquement.

Grâce à cette étude plusieurs nouveaux estimateurs ont pu être développés. Dans le cas des estimateurs de fréquence basés sur un modèle à fréquence et amplitude constantes, nous avons mis au point une méthode pour réduire de façon arbitraire le biais d'une méthode existante, et une autre méthode permettant d'adapter les estimateurs du type interpolateur de spectre utilisant la phase à n'importe quel type de fenêtres. Cependant le gain de performance attendu pour ces nouveaux estimateurs est assez faible, car les méthodes existantes présentent des résultats déjà très proches de la borne théorique. Nous avions également constaté dans la comparaison expérimentale des estimateurs de la littérature que très peu de ces estimateurs sont réellement robustes à des modulations, à l'exception de la méthode du réassignement pour l'estimation de fréquence et la méthode de la QIFFT, qui est dédiée à l'étude du modèle avec modulation linéaire de fréquence et d'amplitude. Nous nous sommes donc intéressés ensuite à des méthodes robustes à des modulations à la fois d'amplitude et de fréquence. Nous avons développé deux méthodes alternatives au réassignement pour l'estimation de fréquence dans le cas d'une modulation d'amplitude seule, donnant des performances similaires et requérant moins de TFs à calculer. Dans le cas du modèle avec modulation linéaire d'amplitude et de fréquence, nous avons proposés deux schémas de calcul complets des paramètres, à la fois moins complexes et plus performants que la QIFFT qui est la seule méthode équivalente déjà existante.

### Indexation Audio

Dans une deuxième partie, nous nous sommes attaché à développer des algorithmes d'indexation audio basé uniquement sur ces paramètres sinusoïdaux. Deux tâches ont été abordées, le suivi de pitch et l'identification d'audio.

Concernant l'estimation de pitch, nous avons présenté une méthode inspirée des travaux de [Maher and Beauchamp, 1994] et [Doval and Rodet, 1993]. Deux mesures d'harmonicité basées sur les pics sinusoïdaux ont été introduites, toutes deux rapides et simples à mettre en oeuvre : la mesure de Maher modifiée et une mesure probabiliste simplifiée. La première donne des résultats meilleurs si le suivi de pitch est désactivé, mais moins bons lorsqu'il est activé. Le suivi de pitch est une méthode originale combinant programmation dynamique et guide d'onde. Sur le type de base utilisé, c'est à dire de la parole monophonique très peu bruitée, beaucoup d'algorithmes ont des performances très bonnes en terme de GER. L'algorithme complet avec la mesure probabiliste simplifiée permet d'améliorer légèrement les performances de l'état de l'art en terme de GER et de détection des zones harmoniques.

L'estimation de pitch basé sur l'analyse sinusoïdale s'est avérée remarquablement performante, tout en restant simple et rapide. Des mécanismes comme l'identification de pics sinusoïdaux et la mesure d'harmonicité présentent des avantages intéressants qui peuvent être réutilisés pour d'autres tâches d'indexation, comme nous l'avons montré ensuite dans le cas de l'identification audio.

Le dernier apport de cette thèse est un algorithme d'identification audio basé sur les pics sinusoïdaux. L'algorithme est basé sur une méthode similaire à celle de l'algorithme de suivi de pitch : elle met en jeu une identification des pics sinusoïdaux et une mesure de vraisemblance simplifiée.

Les avantages d'une empreinte basée sur les pics sinusoïdaux sont de deux ordres par rapport aux empreintes basées sur les sous-bandes, les plus utilisées dans la littérature. Tout d'abord une meilleure modélisation du signal à analyser permet de reconnaître de façon fiable des évènements sonores très courts de l'ordre de 1 s. Deuxièmement, l'empreinte sinusoïdale, basée uniquement sur les fréquences des sinusoïdes, est plus robuste aux distortions, comme la compression et surtout l'ajout de bruit non-aléatoire comme de la parole. Le principal inconvénient de ce type d'empreinte est son inefficacité pour traiter des données non-sinusoïdales, comme on a pu le constater pour le cas des morceaux rythmiques.

Enfin, nous avons constaté que la réduction de l'espace de recherche par table de look-up permet d'avoir des performances similaires en terme de taille en mémoire et de complexité par rapport à deux algorithmes existants, les méthodes de Philips et de Shazam.

## Perspectives

Les algorithmes d'estimation des paramètres sinusoïdaux développés sont à la fois rapides et performants. Ils trouverons des applications notamment en analyse-synthèse et en codage. Leur utilisation dépasse le cadre du traitement des signaux audio, et pourrons éventuellement être utilisés dans d'autres domaines tels que les radars et la sismologie.

Des généralisations des méthodes employées et des améliorations peuvent être envisagées. Les méthodes développées pourraient être adaptées à d'autres transformations temps-fréquences similaires à la transformée de Fourier, comme par exemple la transformée en ondelettes. Des généralisations à certaines classes de transformées sont probablement possibles. La principale amélioration concerne le développement de méthodes similaires pour des modèles plus complexes. Nous avons notamment mis en avant le modèle de modulation quadratique en amplitude et cubique en phase, déjà utilisé avec succès en synthèse sinusoïdale. Ce type de modèle devrait être à même de représenter efficacement les vibratos et les trémolos par exemple.

Concernant l'estimation de pitch nous avons constaté que les algorithmes existant étaient déjà très performants sur le protocole de test communément utilisé. Cela ne veut pas dire pour autant que tous les algorithmes se valent, car les bases de test utilisées sont relativement propres. Les tests seraient certainement plus significatifs en modifiant le protocole de test de ces méthodes, en choisissant par exemple d'ajou-

ter divers bruits typiques, facilement reproductibles, au corpus de parole. Les autres perspectives concernent d'abord l'extension de l'algorithme de suivi de pitch à l'analyse multirésolution. En effet, il est connu que la résolution optimale varie en fonction du locuteur, en particulier s'il s'agit d'un homme ou d'une femme. Notre algorithme peut facilement prendre en compte des pics sinusoïdaux provenant d'analyse à des résolutions différentes. Les seules modifications dans ce cas concernent la partie analyse sinusoïdale, en amont de l'algorithme de suivi de pitch. Une dernière perspective intéressante est l'analyse multipitch, auquel se prête bien cet algorithme. Des algorithmes d'estimation multipitch basés sur les pics sinusoïdaux existent déjà, certains donnant de très bons résultats [Klapuri, 2004]. Une adaptation du suivi de pitch présenté dans la section 7.3.4 est envisageable et pourrait donner de bons résultats.

Concernant l'identification audio basée sur l'analyse sinusoïdale, nous avons constaté des faiblesses au niveau du type de clé utilisé pour réduire l'espace de recherche. Des améliorations peuvent être facilement apportées. Une meilleure répartition dans l'espace temps-fréquence des pics utilisés pour créer cette clé est notamment souhaitable. Des améliorations sur la sélection des composantes sinusoïdales sont envisageables, en ajoutant d'autres critères de sélection, comme la continuité, en utilisant par exemple un suivi sinusoïdal. Enfin, concernant la comparaison, on pourrait chercher à adapter l'algorithme d'identification pour tenir compte des étirements temporels possibles. Cela entraînerait bien évidement un coût de calcul supplémentaire mais qui peut être acceptable pour certaines tâches d'indexation.

Pour conclure sur l'application des paramètres sinusoïdaux à l'indexation audio, nous pouvons dire que ces paramètres sont particulièrement bien adaptés dans le cas où l'identification est exacte, comme on l'a vu dans le cas du suivi de pitch, de l'identification de jingles ou de morceaux de musique. Mais dans le cas où il y a une forte variabilité, comme par exemple pour la reconnaissance de parole ou l'identification de "bruits" (bruits de pas etc.) la modélisation sinusoïdale ne semble pas adaptée. Différentes expériences effectuées durant la thèse mais également des expériences réalisées par d'autres chercheurs vont dans ce sens. Notamment [Ogle and Ellis, 2007] a tenté d'appliquer la méthode de Shazam à des sons fortement variables comme des bruits de pas et des fermetures de portes. Les performances se sont révélées très mauvaises. Un travail non publié [Mandel, 2005] a tenté d'appliquer la méthode Shazam pour trouver des similarités entre des signaux audio différents, sans succès. Dans le cadre de notre thèse nous avons pu constater que pour des signaux rythmiques notamment, l'identification sinusoïdale échouait, alors qu'ils restaient clairement identifiable pour un auditeur. Enfin des expériences non reproduites ici concernant l'application de l'algorithme d'identification audio au cas de la reconnaissance de phonèmes ont également été très peu concluantes.

La paramétrisation sinusoïdale seule ne pourra donc pas suffire pour indexer la totalité des signaux audio. Il faudra soit trouver une représentation plus souple, soit, comme dans le cas de la modélisation sinusoïdes+bruits, lui adjoindre d'autres descripteurs plus efficaces pour les signaux fortement variables.

# Appendices

ANNEXE A

# Propriétés de la transformée de Fourier

Dans cette annexe on rappelle les propriétés des transformation de Fourier continue et discrète. On appelle $f$ une fonction quelconque du temps et $F$ sa transformée de Fourier. Les propriétés de la transformation continue, définie à l'équation (3.3.3), sont donnée dans le tableau A.1. La transformation discrète, définie à l'équation (3.3.1), présente des propriétés équivalentes sauf pour les dérivées temporelles. La démonstration fait en effet intervenir des particularités des intégrales, notamment des intégrations par partie etc.

TAB. A.1: Propriétés de la transformée de Fourier continue

| **Propriété** | **Fonction** | **Transformée de Fourier** |
|---|---|---|
| Inverse | $F(t)$ | $2\pi f(-\omega)$ |
| Convolution | $f_1 \star f_2(t)$ | $F_1(\omega)F_2(\omega)$ |
| Multiplication | $f_1(t)f_2(t)$ | $\frac{1}{2\pi}F_1 \star F_2(\omega)$ |
| Translation | $f(t-u)$ | $\mathrm{e}^{-ju\omega}\,F(\omega)$ |
| Modulation | $\mathrm{e}^{jut}\,f(t)$ | $F(\omega-u)$ |
| Changement d'échelle | $f(t/u)$ | $\lvert u\rvert F(u\omega)$ |
| Dérivées temporelles | $f^{(p)}(t)$ | $(j\omega)^p F(\omega)$ |
| Dérivées fréquentielles | $(-jt)^p f(t)$ | $F^{(p)}(\omega)$ |
| Complexe conjugué | $\bar{f}(t)$ | $\bar{F}(\omega)$ |
| Symétrie hermitienne | $f(t) \in \mathbb{R}$ | $F(-\omega) = \bar{F}(\omega)$ |

ANNEXE B

# Propriétés des fenêtres d'analyse

Dans cette section nous donnons les définitions de la plupart des fenêtres utilisées dans ce document, ainsi que certaines propriétés de ces fenêtres, qui sont à chaque fois justifiées. Parmi ces propriétés, il y a tout d'abord les relations entre les fenêtres usuelles triangulaire, rectangulaire, Blackman et Hann, à la section B.1. Ensuite à la section B.2, nous parlerons de deux propriétés de la fenêtre Gaussienne concernant la TF et le réassignement. Enfin à la section B.3, nous parlerons des propriétés asymptotiques générales de toutes ces fenêtres.

## B.1 Relation entre les fenêtres rectangulaire, triangulaire et les fenêtres de Blackman et de Hann

La fenêtre de Hann est définie par l'équation[Kunt, 1999] :

$$h_h(t) = \begin{cases} 0.5 + 0.5\cos(2\pi\frac{t}{T}) \text{ Pour } |t| \leq T/2 \\ 0 \text{ partout ailleurs} \end{cases} \tag{B.1.1}$$

La fenêtre de Blackman est définie par [Kunt, 1999] :

$$h_b(t) = \begin{cases} a_0 + 2\sum_{l=1}^{L} a_l \cos(\frac{2\pi tl}{T}) \text{ Pour } |t| \leq T/2 \\ 0 \text{ partout ailleurs} \end{cases} \tag{B.1.2}$$

Les coefficients $a_l$ doivent satisfaire la condition :

$$a_0 + 2\sum_{l=1}^{L} a_l = 1 \tag{B.1.3}$$

La fenêtre triangulaire est définie par[1] :

$$h_t(t) = \begin{cases} \frac{1}{S}\left[1 + 2\sum_{l=1}^{S-1}(1 - \frac{l}{S})\cos(\frac{2\pi tl}{T})\right] \text{ Pour } |t| \leq T/2 \\ 0 \text{ partout ailleurs} \end{cases} \quad \text{(B.1.4)}$$

où $S$ est la largeur en bins de la réponse fréquentielle de la fenêtre triangulaire.

Les propriétés suivantes sur ces fenêtres sont utilisées dans le document :

- La fenêtre de Hann est un cas particulier de la fenêtre de Blackman généralisée
- La fenêtre triangulaire est un cas particulier de la fenêtre de Blackman généralisée
- La fenêtre de Hann est équivalente à une fenêtre triangulaire à deux bins
- La fenêtre de Hann est la somme de trois fenêtres rectangulaires

Pour la première propriété, il suffit de choisir $L = 1$, $a_0 = 0.5$, $a_1 = 1/4$, on retombe sur la fenêtre de Hann.

Pour la deuxième propriété, on voit que la fenêtre triangulaire est un cas particulier de la fenêtre de Blackman, avec $L = S$, $a_0 = 1/S$ et $a_l = \frac{1}{S}(1 - \frac{l}{S})$. En effet, on montre aisément que $a_0 + 2\sum_{l=1}^{S-1}(1 - \frac{l}{S}) = 1$.

Pour la troisième propriété, il suffit de prendre $S = 2$ dans la formule de la fenêtre triangulaire et on retombe sur la fenêtre de Hann.

En appelant $H_r$ la réponse fréquentielle, de la fenêtre rectangulaire, on peut montrer que la fenêtre de Blackman généralisée est en fait un combinaison linéaire $M = 2L + 1$ répliques de $H_r$ [Kunt, 1999]. En particulier pour la fenêtre de Hann :

$$H_h(\omega) = 0.5H_r(\omega) + 1/4H_r(\omega + \delta_\omega) + 1/4H_r(\omega - \delta_\omega) \quad \text{(B.1.5)}$$

où $\delta_\omega = 2\pi/T$ est l'écart fréquentiel entre deux bins FFT.

## B.2 Deux propriétés de la fenêtre Gaussienne

La fenêtre Gaussienne est définie par :

$$h_g(t) = \begin{cases} e^{-\frac{t^2}{2\sigma^2}} \text{ Pour } |t| \leq T/2 \\ 0 \text{ partout ailleurs} \end{cases} \quad \text{(B.2.1)}$$

où $\sigma$ est la variance de la Gaussienne. La fenêtre, comme celle du précédent paragraphe est normalisée par rapport à sa valeur maximale dans le domaine temporel. Ici l'énergie de la Gaussienne n'est donc pas normalisée. La résolution de la fenêtre de la Gaussienne est déterminée par le paramètre $\sigma$ et ne change pas lorsque $N$ augmente.

Deux propriétés importantes de la fenêtre Gaussienne sont utilisées dans ce document.

- La transformée de Fourier de la fenêtre Gaussienne est une autre fonction Gaussienne : $H_g(\omega) = \sqrt{2\pi\sigma^2}e^{-\frac{\sigma^2}{2}\omega^2}$.

[1] Elle est définie dans l'article [Keiler and Zölzer, 2001] à l'équation (7). Ici on remplacé 1/N par 1/S pour que la fenêtre soit normalisée, *i.e.* le maximum temporel soit égal à 1.

– Pour une fenêtre Gaussienne, le calcul du temps et de la fréquence réassignés ne requière que deux transformées de Fourier.

La première propriété n'est valable que dans le cas où la transformée de Fourier est continue et infinie. Elle sera approximativement valable dans le cas discret et fini, à condition que $N$ soit suffisamment grand.

Démonstration de la première propriété :

$$H_g(\omega) = \int_{-\infty}^{+\infty} h_g(t) e^{-j\omega t} dt \tag{B.2.2}$$

$$= \left[ -\frac{1}{j\omega} h_g(t) e^{-j\omega t} \right]_{-\infty}^{+\infty} - \frac{1}{j\omega\sigma^2} \int_{-\infty}^{+\infty} t h_g(t) e^{-j\omega t} dt \tag{B.2.3}$$

$$= -\frac{1}{\omega\sigma^2} H_g^{'}(\omega) \tag{B.2.4}$$

La deuxième ligne est obtenue par intégration par partie. La dernière équation est une équation différentielle dont la solution générale est :

$$H_g(\omega) = C e^{-\frac{\sigma^2}{2}\omega^2} \tag{B.2.5}$$

La constante est déterminée pour $\omega = 0$ :

$$C = H_g(0) = \int_{-\infty}^{+\infty} h_g(t) dt = \sqrt{2\pi\sigma^2} \tag{B.2.6}$$

On en déduit donc bien la première propriété.

Démonstration de la deuxième propriété :

On considère que un signal $x$ et sa transformée de Fourier continue, définie à l'équation (3.3.3). Pour la fenêtre $\dot{h} = \frac{-t}{\sigma^2} h(t)$, on aura donc :

$$X_c(t, \omega; \dot{h}) = \frac{-1}{\sigma^2} X_c(t, \omega; \tau.h) \tag{B.2.7}$$

$X_c(t, \omega; \dot{h})$ peut donc être déduit directement de $X_c(t, \omega; \tau.h)$. Il ne faut donc plus que deux transformées de Fourier pour calculer les temps et fréquences réassignés (cf. section 4.2.1.1).

## B.3 Propriétés asymptotiques des fenêtres

On considère une fenêtre $h(t)$ réelle, positive, symétrique, définie et intégrable sur $[-1/2, 1/2]$. Pour obtenir une fenêtre de taille $t_N \neq 1$, on va dilater temporellement la fenêtre $h$ : $h_T(t) = h(\frac{t}{t_N})$. On note $F$ la fréquence d'échantillonage et $N$ le nombre de points correspondant à une fenêtre de taille $t_N$, c'est à dire $t_N = N/F$. $\Gamma(h)$ et $\Gamma(\delta; h)$ sont définis comme dans la partie 1 par $\Gamma(h) = \sum_{n=-N/2}^{N/2} h(\tau_n)$ et $\Gamma(\delta; h) = \sum_{n=-N/2}^{N/2} h(\tau_n) e^{-j\delta\tau_n}$. On a alors les propriétés asymptotiques suivantes :

$$\Gamma(h_T) \sim C.N \tag{B.3.1}$$

$$\Gamma(\delta; h_T) \sim C.N \text{ si } |\delta| \geq R \tag{B.3.2}$$

$$\Gamma(t^q h_T) \sim C.N^{q+1} \text{ si q est pair} \tag{B.3.3}$$

où $C \leq 0$ est une constante ne dépendant que de la fenêtre. $R$ est la précision de la transformée de Fourier : $R = \frac{\pi F}{N}$.

Démonstration :

Pour la relation (B.3.1) :

$$\Gamma(h_T) = \sum_{n=-N/2}^{N/2} h_T(\tau_n) = \sum_{n=-N/2}^{N/2} h(\frac{\tau_n}{t_N}) \tag{B.3.4}$$

Cette dernière expression est une somme de Riemann, donc asymptotiquement, on a :

$$\frac{1}{F}\Gamma(h_T) \leftrightarrow \int_{-t_N/2}^{t_N/2} h(\frac{t}{t_N} dt \tag{B.3.5}$$

Or d'après les hypothèses, on a :

$$\int_{-t_N/2}^{t_N/2} h(\frac{t}{t_N}) dt = t_N \int_{-.5}^{.5} h(t) dt = C.t_N \tag{B.3.6}$$

$C$ est une constante positive car $h$ est positive. On peut donc conclure que :

$$\Gamma(h_T) \leftrightarrow C.N \tag{B.3.7}$$

Pour la relation (B.3.2) :

$$\Gamma(\delta; h_T) = \sum_{n=-N/2}^{N/2} h_T(\tau_n) e^{-j\delta\tau_n} \tag{B.3.8}$$

$$= \sum_{n=-N/2}^{N/2} h(\frac{\tau_n}{t_N}) \cos(\delta\tau_n) \tag{B.3.9}$$

car la fenêtre $h$ est symétrique. En supposant que $\delta \leq R = \frac{\pi F}{N}$, on a $\delta\tau_n \in [-\frac{\pi}{2}; \frac{\pi}{2}]$ et $1 \geq \cos(\delta\tau_n) \geq 0 \ \forall n \in [-N/2, N/2]$. De plus $\cos(\delta\tau_n)$ est symétrique et monotone décroissante sur $n \in [0, N/2]$, donc $1 \geq \cos(\delta\tau_n) \geq \cos(R\tau_n) \ \forall |\delta| \in R, \forall n \in [-N/2, N/2]$. On en déduit les inégalités suivantes :

$$\Gamma(R; h_T) \leq \Gamma(\delta; h_T) \leq \Gamma(0; h_T) = \Gamma(h_T) \tag{B.3.10}$$

On pose $h_T^{'}(t) = h(\frac{t}{t_N}) \cos(Rt) = h(\frac{t}{t_N}) \cos(\pi \frac{t}{t_N})$ et $h^{'}(t) = h(t) \cos(\pi t)$. Comme $h^{'}$ est symétrique, réelle, positive et intégrable, on peut appliquer la relation (B.3.1). On en déduit que :

$$C^{'}.N \lesssim \Gamma(\delta; h_T) \lesssim C.N \tag{B.3.11}$$

où $C^{'}$ et $C$ sont deux constantes positives.

Pour la relation (B.3.3) :

$$\Gamma(t^q h_T) = {t_N}^q \sum_{n=-N/2}^{N/2} (\frac{\tau_n}{t_N})^q h(\frac{\tau_n}{t_N}) \tag{B.3.12}$$

On pose $h^{'}(t) = \frac{t}{t_N}^q h(\frac{t}{t_N})$. Comme $q$ est pair, $h^{'}$ est symétrique réelle, définie positive et intégrable, donc on peut utiliser la relation (B.3.1). On en déduit que :

$$\Gamma(t^q h_T) \leftrightarrow C{t_N}^{q+1} \tag{B.3.13}$$

ANNEXE C

# Bornes sur l'erreur de modélisation

## C.1 Modèle polynomial en amplitude

Considérons le développement de Taylor d'ordre $q-1$ et $k-1$ respectivement pour les fonctions d'amplitude et de phase :

$$A(t_m+\tau_n) = S_A + \frac{\tau_n^k}{k!}\alpha(t_m,\tau_n)$$
$$\Phi(t_m+\tau_n) = S_\Phi + \frac{\tau_n^q}{q!}\beta(t_m,\tau_n)$$

$t_m = m/F$ est le temps global et $\tau_n = n/F$ est le temps local. $S_A$ et $S_\Phi$ sont les développements de Taylor polynomiaux, que nous préférons garder sous cette forme compacte. Ils dépendent aussi de $t_m$ et de $\tau_n$, mais ceux-ci ont été omis par souci de commodité. $\alpha$ et $\beta$ sont les restes de Lagrange et sont bornés par :

$$|\alpha| \leq \sup_{r\in[t_m,t_m+\tau_n]} |A^k(r)|$$
$$|\beta| \leq \sup_{r\in[t_m,t_m+\tau_n]} |\Phi^q(r)|$$

Le modèle de signal $x_m$ d'ordre $M_{q-1,k-1}$ est donc donné par :

$$x_M = S_A\, \mathrm{e}^{jS_\Phi}$$

En injectant les développements de Taylor dans la définition du signal réel, on obtient :

$$x(t_m+\tau_n) = S_A\, \mathrm{e}^{j(S_\Phi+\frac{\tau_n^q}{q!}\beta(t_m,\tau_n))} + \frac{\tau_n^k}{k!}\alpha(t_m,\tau_n)\, \mathrm{e}^{j(S_\Phi+\frac{\tau_n^q}{q!}\beta(t_m,\tau_n))}$$

On effectue un développement de Taylor d'ordre 1 de la fonction exponentielle du premier terme :

$$x(t_m+\tau_n) = S_A\,\mathrm{e}^{jS_\Phi}(1+\frac{\tau_n^q}{q!}\beta(t_m,\tau_n)\eta(t_m,\tau_n)) + \frac{\tau_n^k}{k!}\alpha(t_m,\tau_n)\,\mathrm{e}^{j(S_\Phi+\frac{\tau_n^q}{q!}\beta(t_m,\tau_n))}$$

$\eta$ est le reste de Lagrange d'ordre 1 de la fonction exponentielle complexe, et il est donc borné par :

$$|\eta| \leq \sup_{r\in[0,\frac{\tau_n^q}{q!}\beta(t_m,\tau_n)]} |\,\mathrm{e}^{jr}\,| \leq 1$$

On peut donc déduire que l'erreur de modélisation s'exprime ainsi :

$$\epsilon = \frac{x-x_m}{x_m} = \epsilon_\Phi + \epsilon_A \tag{C.1.1}$$

$$\epsilon_\Phi = \frac{\tau_n^q}{q!}\beta(t_m,\tau_n)\eta(t_m,\tau_n) \tag{C.1.2}$$

$$\epsilon_A = \frac{\tau_n^k}{k!}\frac{\alpha(t_m,\tau_n)}{S_A}\,\mathrm{e}^{j\frac{\tau_n^q}{q!}\beta(t_m,\tau_n)} \tag{C.1.3}$$

En considérant que l'intervalle d'analyse est de longueur $N$ et qu'il est centré, les bornes sur les erreurs sont donc les suivantes :

$$\begin{aligned}
|\epsilon| &\leq |\epsilon_\Phi| + |\epsilon_A| \\
|\epsilon_\Phi| &\leq |\frac{\tau_n^q}{q!}\beta(t_m,\tau_n)| \leq \frac{\tau_N^q}{2^q q!}M_{\Phi^q} \\
|\epsilon_A| &\leq |\frac{\tau_n^k}{k!}\frac{\alpha(t_m,\tau_n)}{S_A}| \leq \frac{\tau_N^k}{2^k k!}\frac{M_{A^k}}{m_A}
\end{aligned}$$

où $M_{\Phi^q}$ et $M_{A_k}$ sont les valeurs maximales possibles pour, respectivement, la dérivée de phase d'ordre $q$ et la dérivée d'amplitude d'ordre $k$. $m_A$ est la valeur minimale de l'amplitude modélisée, sur l'intervalle d'analyse. Plus l'amplitude du signal va être faible, plus l'erreur relative sera sensible à une variation du paramètre d'ordre $k$.

## C.2 Modèle log-polynomial en amplitude

Pour un modèle exponentiel en amplitude (log-polynomial), on dérive de la même façon les bornes sur l'erreur. $x$ peut s'écrire ainsi :

$$\begin{aligned}
x &= x_M(1+\epsilon_\Phi+\epsilon_L+\epsilon_{L\Phi}) \\
\epsilon_L &= \frac{\tau_n^k}{k!}\alpha\eta_2 \\
\epsilon_{L\Phi} &= \eta\eta_2\beta\alpha\frac{\tau_n^q}{q!}\frac{\tau_n^k}{k!}
\end{aligned}$$

$\epsilon_\Phi$ est donné par l'équation (C.1.2), et $\eta_2$ est le reste de Lagrange de la fonction exponentielle, qui est bornée ainsi :

$$\eta_2 \leq \sup_{r \in [0, \frac{\tau_n^q}{q!}\beta]} |\, \mathrm{e}^r \,| \leq \mathrm{e}^{\frac{\tau_n^q}{q!} M_{L^k}} \tag{C.2.1}$$

Les bornes sur les erreurs sont dans ce cas :

$$\begin{aligned} |\epsilon| &\leq |\epsilon_\Phi| + |\epsilon_L| + \epsilon_{L\Phi} \\ |\epsilon_\Phi| &\leq \frac{\tau_N^q}{2^q q!} M_{\Phi^q} \\ |\epsilon_L| &\leq \frac{\tau_N^k}{2^k k!} M_{L^k}\ \mathrm{e}^{\frac{\tau_N^k}{2^k k!} M_{L^k}} \\ |\epsilon_{L\Phi}| &\leq |\epsilon_L||\epsilon_\Phi| \end{aligned}$$

Si les erreurs $\epsilon_L$ et $\epsilon_\Phi$ sont petites devant 1 (moins de 20% d'erreur), alors $\epsilon_{L\Phi}$ devient négligeable et on aura également $\frac{\tau_N^k}{2^k k!} M_{L^k} \ll 1$. On peut donc utiliser l'approximation $x\,\mathrm{e}^x \approx x$ pour la borne de $\epsilon_L$. Les bornes deviennent alors :

$$\begin{aligned} |\epsilon| &\leq |\epsilon_\Phi| + |\epsilon_L| \\ |\epsilon_\Phi| &\leq \frac{\tau_N^q}{2^q q!} M_{\Phi^q} \\ |\epsilon_L| &\leq \frac{\tau_N^k}{2^k k!} M_{L^k} \end{aligned}$$

## C.3 Erreur sur la TF du signal

On définit l'erreur relative sur la transformée de Fourier par :

$$E = \frac{|X - X_M|}{N M_A} \tag{C.3.1}$$

où $X$ est la TF du signal réel et $X_M$ la TF du signal modélisé. L'erreur est normalisée par l'amplitude maximum $M_A$ du signal modélisé ainsi que par la taille de la fenêtre $N$ de façon à avoir une erreur relative moyenne.

La transformée de Fourier étant linéaire, l'erreur va se décomposer de la même façon que précédemment. En outre, comme $|\sum_i x_i| \leq \sum_i |x_i|$, on peut déduire que les bornes de l'erreur sur la TF du signal, dans le cas d'une amplitude polynomiale, sont :

$$\begin{aligned} |E| &\leq |E_\Phi| + |E_A| \\ |E_\Phi| &\leq \frac{h_q}{N q!} M_{\Phi^q} \\ |E_A| &\leq \frac{h_k}{N k!} \frac{M_{A^k}}{M_A} \end{aligned}$$

où $h_k$ est défini par $h_k = \sum_{i=-N/2}^{N/2} h(\tau_i)|\tau_i|^k$. $h$ est la fenêtre d'analyse de la TF.

Dans le cas d'une amplitude log-polynomiale et en supposant que $M_{A^k}\frac{\tau_N^k}{2^k k!} \ll 1$ , on trouve de la même façon que les bornes sur l'erreur sont :

$$|E| \leq |E_\Phi| + |E_L|$$
$$|E_L| \leq \frac{h_k}{Nk!} M_{A^k}$$

ANNEXE D

# Dérivation de la variance de l'estimateur (6.1.7)

On applique la méthode décrite dans le chapitre 5.4. Ici la fonction $g$ utilisée est la partie réelle, $g(x) = \Re(x)$ et la fonction $f$ combinant les bins est définie par :

$$f(X_1, X_2) = \omega_b - \frac{\Gamma(h_c)}{\Gamma(\tau h_s)} \frac{X_1 - X_2}{X_1 + X_2} \tag{D.0.1}$$

D'après la formule (5.4.5), $\epsilon_N$ nous est donné par :

$$\epsilon_N = \frac{\Gamma(h_c)}{\Gamma(\tau h_s)} \Re\Big(\sum_i N_i \frac{\partial f}{\partial X_i}\Big) \tag{D.0.2}$$

$$= \frac{2\Gamma(h_c)}{\Gamma(\tau h_s)} \Re\Big(\frac{N_1 X_1}{(X_1 + X_2)^2} - \frac{N_2 X_2}{(X_1 + X_2)^2}\Big) \tag{D.0.3}$$

On exprime ensuite $X_1$ et $X_2$ en fonction de $\delta = \omega_b - \beta$ :

$$X_1 = A\,\mathrm{e}^{j\alpha} \sum_i h_i \cos((\omega_b - \delta_\omega + \beta)\tau_i) \tag{D.0.4}$$

$$= A\,\mathrm{e}^{j\alpha}[\Gamma(\delta; h_c) - j\Gamma(\delta; h_s)] \tag{D.0.5}$$

$$X_2 = A\,\mathrm{e}^{j\alpha}[\Gamma(\delta; h_c) + j\Gamma(\delta; h_s)] \tag{D.0.6}$$

$$X_1 + X_2 = 2A\,\mathrm{e}^{j\alpha}\,\Gamma(\delta; h_c) \tag{D.0.7}$$

On rappelle que $\delta_\omega = (\omega_2 - \omega_1)/2$. On passe de l'équation D.0.4 à l'équation D.0.5 en développant le cosinus et en utilisant les définitions de $h_s$ et $h_c$.

On en déduit pour $\epsilon_N$ :

$$\epsilon_N = \frac{\Gamma(h_c)}{2A\,\mathrm{e}^{j\alpha}\,\Gamma(\tau h_s)\Gamma(\delta; h_c)} \big((1 + \mathcal{H})\Re(N_1\,\mathrm{e}^{-j\alpha}) - (1 - \mathcal{H})\Re(N_2\,\mathrm{e}^{-j\alpha})\big) \tag{D.0.8}$$

$$= \frac{\Gamma(h_c)}{2A\,\mathrm{e}^{j\alpha}\,\Gamma(\tau h_s)\Gamma(\delta;h_c)}\big((1+\mathcal{H})(N_1^R\cos(\alpha)+N_1^I\sin(\alpha)) - (1-\mathcal{H})(N_2^R\cos(\alpha)+N_2^I\sin(\alpha))\big) \tag{D.0.9}$$

où $N_i^R = \Re(N_i)$ et $N_i^I = \Im(N_i)$.

Lorsque l'on va élever $\epsilon_N$ au carré, on aura une fonction de termes du type $N_i^x N_j^y$ où $x$ et $y$ désigne soit la partie réelle soit la partie imaginaire et $i,j \in \{1,2\}$. Pour trouver la variance de $\epsilon_N$ il faudra donc connaître l'espérance de tous ces termes :

$$E(N_1^{R2}) = E(N_2^{R2}) = \frac{\sigma^2}{2}\Gamma(h^2) \tag{D.0.10}$$

$$E(N_1^R N_1^I) = E(N_2^R N_2^I) = 0 \tag{D.0.11}$$

$$E(N_1^R N_2^R) = E(N_1^I N_2^I) = \frac{\sigma^2}{2}\big[2\Gamma(h_c^2) - \Gamma(h^2)\big] \tag{D.0.12}$$

$$E(N_1^R N_2^I) = -E(N_2^R N_1^I) = \sigma^2\Gamma(h_s h_c) \tag{D.0.13}$$

En remarquant en plus que $\Gamma(h^2) = \Gamma(h_c^2) + \Gamma(h_s^2)$, on en déduit la variance de l'estimateur :

$$\mathrm{var}(\hat{\beta}) = E(\epsilon_N^2) = \frac{\Gamma(h_c)}{2A\,\mathrm{e}^{j\alpha}\,\Gamma(\tau h_s)\Gamma(\delta;h_c)}\big[\Gamma(h_s^2) + \frac{\Gamma(\delta;h_s)^2}{\Gamma(\delta;h_c)^2}\Gamma(h_c^2)\big] \tag{D.0.14}$$

Comme $\Gamma(\delta;h_s)$ est une fonction monotone croissante sur $[0,R]$ et $\Gamma(\delta;h_c)$ est une fonction monotone décroissante, on en déduit que la variance est maximum pour $\delta = R$. Plus l'écart avec la fréquence de référence $\omega_b$ est importante, plus l'erreur est importante.

ANNEXE E

# Seuil adaptatif pour la sélection de pic sinusoïdaux

L'espérance de l'estimateur d'amplitude (4.1.5) est donnée par la formule :

$$E(\tilde{A}^2) = A^2 + \frac{2\sigma^2 \sum_{i=0}^{N} h_i^2}{|H(f_k - f)|^2}$$

- $A$ est l'amplitude de la sinusoïde,
- $\tilde{A}$ est l'amplitude estimée,
- $\sigma$ est la variance du bruit (supposé blanc, Gaussien),
- $h_i$ est l'échantillon i de la fenêtre h,
- $H$ est la TF de $h$,
- $N$ est la taille de la fenêtre $h$,
- $f$ est la fréquence (exacte) de la sinusoïde,
- $f_k$ est la fréquence du bin k le plus proche de $f$ ,
- E est l'opérateur d'espérance mathématique.

On voit qu'il s'agit d'un terme qui dépend de l'amplitude et d'un terme dû au bruit. On cherche $A_l$ tel que le terme dû au bruit soit négligeable par rapport au terme d'amplitude, avec un rapport de $X$ dB. On exprime également la variance du bruit s'exprime à partir de la puissance du signal et du SNR :

$$X = 10 \log_{10}\Big(\frac{A_l^2}{E(\tilde{A}^2) - A^2}\Big)$$
$$\sigma^2 = P_s.10^{-\frac{SNR}{10}}$$

On déduit de ces trois formules la valeur limite :

$$A_l^2 = \alpha.P_s.10^{\frac{Z}{10}}$$

où $Z = X - SNR$ et $\alpha = \frac{2\sum_i h_i^2}{|H(f_k - f)|^2}$.

On veut que $A_l$ soit un seuil inférieur indépendant de $f$. On impose donc $\alpha = \frac{2\sum_i h_i^2}{|H(R)|^2}$, avec $R = \frac{F_s}{2P}$ et $P$ étant la taille de la TF.

Le seuil ne dépend alors que d'un paramètre $Z$. Par exemple si on veut garder les sinusoïdes qui ressortent d'au moins 10 dB d'un bruit correspondant à un SNR de 20dB on va choisir $Z = 10 - 20 = -10$dB.

ANNEXE F

# Comparaison des complexités des représentations temporelle et fréquentielle

Dans cette annexe, nous allons détailler le calcul des complexités des algorithmes de comparaison, pour la représentation temporelle et pour la représentation fréquentielle. Cette compraison est dominée dans les deux cas par le calcul de la vraisemblance entre un ensemble de pics $T$ appartenant à un objet de référence et un ensemble de pics $T^{'}$ appartenant à l'objet à identifier. Nous terminons la partie par le calcul du coût du passage d'une représentation temporelle à une représentation fréquentielle.

## Indexation Temporelle

Dans la méthode de l'indexation temporelle, les données sont d'abord indexées temporellement avant d'être indexée fréquentiellement, comme indiqué sur la Figure 8.12(a). Soit $T$ l'ensemble des $N$ trames de l'objet de référence et $T^{'}$ celui des $N^{'} < N$ trame de l'objet à identifier. On note $t$ une trame de l'objet $T$ et $t^{'}$ une trame de l'objet $T^{'}$. $t$ contient $n$ éléments et $t'$ contient $n^{'}$ éléments.

L'élément de base de la méthode consiste à identifier les fréquences communes entre une trame $t$ et une trame $t'$ quelconque. Une fréquence $f$ de $t$ est dite trouvée s'il existe une fréquence $f'$ de $t'$ telle que $|f - f^{'}| < T_f$ où $T_f$ est la tolérance fréquentielle. On met alors la vraisemblance de la trame à jour en l'incrémentant de 1. On rappelle que les fréquences sont représentées par des mots de $B$ bits. Elles peuvent donc être considérées comme des entiers pour les calculs.

```
Identifier_Trame(t,t')                    Coût    Fois
1   i <- 0
2   j <- 1
3   Pour i = 1..n                           1      n+1
```

```
4     Faire val <- t[i] - Tf                2       n
5           Tant que t'[j] - val < 0        2       n+n'
6              Faire j <- j + 1             1       n'
7                    Si j - n' > 0          1       n'
8                    Alors Retourne l
9           Si t'[j] - t[i] - Tf < 0        4       n
10          Alors l <- l + 1
11  Retourne l
```

Le test $|f - f^{'}| < T_f$ est décomposé en deux parties en s'appuyant sur le fait que les fréquences de $t$ et $t^{'}$ sont ordonnées de façon croissante. La boucle qui commence à la ligne 5 correspond au test $x_2 - x_1 > -T_f$. La ligne 10 correspond au test $x_2 - x_1 < T_f$. Le coût est donné en nombre d'additions entières et il est calculé en supposant que les fréquences sont réparties de façon uniforme. Un accès à un tableau est compté comme une addition entière et on néglige les comparaisons à 0 et les opérations de conditionnement. Dans le cas où les fréquences sont réparties uniformément et si $n$ et $n^{'}$ sont suffisement grands, la boucle de la ligne 5 sera approximativement appelée $n^{'}$ fois. La ligne 10 n'est pas compté car la probabilité de trouver une fréquence ne s'exprime pas de façon simple. Le coût total de la méthode est donc :

$$C \approx 9.n + 4.n^{'}$$

En considérant que les pics spectraux sont répartis uniformément à la fois fréquentiellement et temporellement, on aura en moyenne $C \approx (9D_p + 4D_p^{'})T_h$, où $D_p$ et $D_p^{'}$ sont les densités de pics par secondes des deux bloc à comparer et $T_h$ est la taille d'une trame en secondes.

La méthode complète pour calculer la vraisemblance va identifier pour tous les blocs de $N'$ trames consécutives parmis les $T$ trames de l'objet de référence, les fréquences qui sont présentes dans l'objet à identifier, lui aussi de taille $N'$.

```
Vraisemblance_Méthode_Temporelle(T,T')                Coût  Fois
1   Pour i = 1..N-N'
2   Faire l[i] <- 0
3         Pour j = 1..N'
4         Faire ll <- Identifier_Trame(T[i+j],T'[j])   C   (N-N')N'
5               l[i] <- l[i] + ll
```

Le coût principal de cette méthode est dû à la ligne 4. Le coût moyen en nombre d'additions du calcul de vraisemblance par la méthode temporelle est donc le suivant :

$$\begin{aligned} C_t &\approx (N - N^{'})N^{'}(9D_p + 4D_p^{'})T_h \\ &\approx (T_b - T_b^{'})T_b^{'}(9D_p + 4D_p^{'})/T_h \end{aligned}$$

où $T_b$ et $T_b^{'}$ sont les longueurs en secondes des blocs de trames.

## Indexation Fréquentielle

Dans la méthode de l'indexation fréquentielle, les données sont d'abord indexées fréquentiellement avant d'être indexée temporellement, comme indiqué sur la Figure 8.12(b). $T$ et $T^{'}$ sont ici des tables possédant $2^B$ entrées et chaque entrée $i$ possède une liste d'éléments de taille $t_i$ et $t_i^{'}$ respectivement.

```
Vraisemblance_Méthode_Fréquentielle(T,T')                    Coût  Fois
1   Pour i = 1..2^B
2     Faire Pour j = 1..ti
3             Faire Pour k = 1..t'i                           1     S
4                     Faire cle <- T[i][j]-T'[i][k]           5     S
5                           l[cle] <- l[cle] + 1              3     S
```

La valeur de la variable *cle* peut être négative, $cle \in [-N', N+N']$. Le coût principal de cette méthode est dû à la boucle la plus intérieure, c'est à dire aux lignes 3 à 4. $S$ est la somme suivante :

$$S = \sum_{i=1}^{2^B} t_i.t_i^{'}$$

On suppose comme pour la représentation temporelle que les fréquences sont réparties de façon uniforme sur l'ensemble des valeurs possibles, pour l'objet de référence et l'objet à identifier. Le nombre d'éléments $t_i$ ou $t_i^{'}$ pour la valeur de fréquence $i$ suit donc une loi binomiale. L'espérance de la binomiale est égale au nombre de tirages, ici correspondant aux $nN$ pics, multiplié par la probabilité d'un tirage, ici égale à $2^{-B}$. On a donc en moyenne $t_i = n.N/2^B$. Dans le cas de l'objet à identifier, on duplique les fréquences pour tenir compte de la tolérance fréquentielle[1]. On a alors $t_i^{'} = n^{'}.N^{'}.N_q/2^B$, où $N_q = 2T_f 2^B/F_c + 1$ est le nombre de fois que l'on insère chaque fréquence dans la table. $T_f$ est toujours la tolérance fréquentielle en hertz et $F_c$ est la fréquence de coupure. Donc le coût total en nombre d'additions de la méthode est le suivant :

$$C_f = 9NN^{'}N_q n n^{'} 2^{-B}$$

On réécrit cette équation en terme des densités de pics par seconde $D_p$ et $D_p^{'}$ et des longueurs en secondes des blocs de trames $T_b$ et $T_b^{'}$ :

$$C_f = 9T_b T_b^{'} N_q D_p D_p^{'} 2^{-B}$$

## Passage d'une représentations à l'autre

On présente ici la méthode pour passer d'une représentation temporelle $T$ à une représentation fréquentielle $F$. $T$ est un tableau de $N$ entrées chaque entrée ayant une dimension $t_i$. $F$ est un tableau possèdant $2^B$ entrées, chaque entrée possédant une dimension $t_i^{'}$. Si on suppose que les pics spectraux sont disposées uniformément

[1] Voir la section 8.4.4.1.

temporellement et fréquentiellement, on a en moyenne $t_i = D_p T_h = n$. On rappelle que $D_p$ est la densité de pics par seconde et $T_h$ la taille d'une trame en secondes.

```
Passage(T,F)                                  Coût  Fois
1   Pour i = 1..2^B
2     Faire m[i] <- 0
3   Pour i = 1..N
4     Faire Pour j = 1..ti                      1    N+1
5             Faire cle <- T[i][j]              2   N(n+1)
6                  F[cle][m[cle]] <- i          2    N.n
7                  m[cle] <- m[cle]+1           3    N.n
```

La complexité totale est donc de $C \approx 8Nn = 8T_b D_p$ additions, où $T_b = NT_h$ est la durée en secondes du bloc à convertir.

ANNEXE G

# Deux articles de l'auteur

## Premier article : *Experimental and theoretical complements to the article 'Estimation of frequency for AM/FM models using the phase vocoder framework'*

Le premier article présenté est un article technique qui détaille les démonstrations concernant les estimateurs RV et PCV. Cet article est un complément à l'article publié dans les Transactions of Signal Processing [Betser et al., 2008].

# Experimental and theoretical complements to the article "Estimation of frequency for AM/FM models using the phase vocoder framework"

Michaël Betser, Patrice Collen, Bertrand David and Gaël Richard

14/02/2007

## Contents

**1 Introduction** **1**

**2 Reassignment and the AM/FM model** **2**

**3 The phase vocoder frequency estimator and the AM/FM model** **2**
3.1 Phase corrected vocoder (PCV) . . . . . 3
3.2 Reassigned vocoder (RV) . . . . . 4
3.3 Study of the unwrapping factor . . . . . 5
3.4 Study of the bias . . . . . 7
3.5 Study of the variance . . . . . 8

**4 Conclusion** **15**

**Abstract**

This technical report is a theoretical and experimental complement to the article [4]. Several demonstrations concerning Fourier-based estimators are presented. The first demonstration concerns the reassignment, and shows that this method is still valid for the AM/FM case. The other demonstrations concern the phase-vocoder-based frequency estimation, in the AM/FM case.

# 1 Introduction

The model under study in this report is the first-order AM/FM model, defined as:

$$x(\tau) \triangleq \mathrm{e}^{\lambda+\mu\tau} \cdot \mathrm{e}^{j(\alpha+\beta\tau+\gamma\tau^2/2)} \tag{1.1}$$

where $\alpha$ is the phase, $\beta$ is the frequency, $\gamma$ is the Frequency Change Rate (FCR), $\lambda$ is the log-amplitude and $\mu$ is the Log-Amplitude Change rate (ACR). $\tau$ is the local time. The time of the $n^{th}$ sample is $\tau_n = n/F$ where $F$ is the sampling frequency.

## 2 Reassignment and the AM/FM model

The frequency reassignment is known to perfectly localize chirp signals [2]. A simple demonstration for the continuous Fourier Transform in the FM case is presented in [3]. Using the same method, it can be shown easily that the time-frequency Reassignment is also perfectly valid for the AM/FM model.

Let's define the continuous Fourier Transform as:

$$FT(x;\omega) \triangleq \int_{-\infty}^{+\infty} x(\tau)\,\mathrm{e}^{-j\omega\tau}\,d\tau \tag{2.1}$$

Let $g$ be $g(\tau) \triangleq x(\tau)h(\tau)$, where $h$ is the Fourier Transform window. Using (1.1), we have:

$$\begin{aligned}\frac{dg}{d\tau}(\tau) =& \frac{dx}{d\tau}(\tau)h(\tau) + \frac{dh}{d\tau}(\tau)x(\tau)\\ =& (j(\gamma\tau+\beta)+\mu)x(\tau)h(\tau) + \frac{dh}{d\tau}(\tau)x(\tau)\end{aligned} \tag{2.2}$$

The FT is applied to the relation (2.2) for a frequency $\omega$ such that the sinusoid $x$ has a non-zero energy for this particular frequency ($FT(g;\omega) \neq 0$), on the definition interval of the window $h$:

$$\begin{aligned}j\omega FT(g;\omega) = j\gamma FT(\tau g;\omega) + (j\beta+\mu)FT(g;\omega)\\ + FT(\frac{dh}{d\tau}x;\omega)\end{aligned}$$

$$\Leftrightarrow \beta - j\mu + \gamma\frac{FT(\tau g;\omega)}{FT(g;\omega)} = \omega + j\frac{FT(\frac{dh}{d\tau}x;\omega)}{FT(g;\omega)} \tag{2.3}$$

$$\Rightarrow \beta + \gamma\Re\Big(\frac{FT(\tau g;\omega)}{FT(g;\omega)}\Big) = \omega - \Im\Big(\frac{FT(\frac{dh}{d\tau}x;\omega)}{FT(g;\omega)}\Big) \tag{2.4}$$

This formula is exactly the same as the one obtained in [3] for the FM model. The influence of the ACR has been removed by taking the real part of equation (2.3). Accordingly to the usual formulation of the reassignment, let's define $\dot{h}(\tau) \triangleq \frac{dh}{d\tau}(\tau)$. The first term of equation (2.4) is the frequency of the partial for the time

$$\hat{t} = t + \Re(\frac{FT(thx;\omega)}{FT(hx;\omega)}) \tag{2.5}$$

which is the time reassignment operator. The second part of the equation corresponds to the frequency reassignment operator

$$\hat{\beta} = \omega - \Im(\frac{FT(\dot{h}x;\omega)}{FT(hx;\omega)}) \tag{2.6}$$

## 3 The phase vocoder frequency estimator and the AM/FM model

In this part the model is supposed valid on an interval $W$, and $\lambda$, $\alpha$, $\beta$ corresponds to the log-amplitude, phase and frequency for the time $t_M$, *i.e.* the time corresponding to the middle of the window $W$).

$$x(\tau) = \mathrm{e}^{\lambda_M + j\alpha_M}\,\mathrm{e}^{\mu\tau + j(\beta_M\tau + \gamma\tau^2/2)} \tag{3.1}$$

Let's define $\Gamma$ as:

$$\Gamma(\beta, \mu, \gamma; h) = \sum_{i=-(N-1)/2}^{(N-1)/2} h(\tau_i)\, \mathrm{e}^{\mu\tau_i}\, \mathrm{e}^{j(\beta\tau_i + \gamma\frac{\tau_i^2}{2})} \tag{3.2}$$

The STFT of $x$ for the time $t_{m_i}$ and the frequency $\omega_{k_i}$ is:

$$X(t_{m_i}, \omega_{k_i}; h) = \mathrm{e}^{\lambda_i + j\alpha_i}\, \Gamma(\beta_i - \omega_{k_i}, \mu, \gamma; h) \tag{3.3}$$

$$X(t_{m_i}, \omega_{k_i}; h) = \mathrm{e}^{\lambda_i + j\alpha_i}\, \Gamma_i \tag{3.4}$$

where $\alpha_i = \alpha_M + \beta_M(t_{m_i} - t_M) + \gamma(t_{m_i} - t_M)^2/2$, $\beta_i = \beta_M + \gamma(t_{m_i} - t_M)$, $\omega_{k_i}$ is the frequency of the closest maximal bin $k_i$ to $\beta_i$ for the time $t_{m_i}$. Finally, $h$ is the window.

If we take the argument of this last equation:

$$\psi_i \triangleq \arg(X_i) = \alpha_i + \arg(\Gamma_i) \tag{3.5}$$

If we consider the phase difference between two time-frequency points, $X_1$ and $X_2$, such that $t_{k_2} - t_M = t_M - t_{k_1} = T/2$:

$$\Delta\psi = \arg(X_2) - \arg(X_1) \tag{3.6}$$

$$= T\beta_M + \arg(\Gamma_2\bar{\Gamma}_1) + 2\pi n \tag{3.7}$$

The first two subsections will recall how work the two methods presented in the article [4], namely the phase corrected vocoder and the reassigned vocoder. Then, in the next three subsections, the property of these estimators, concerning the unwrapping factor $n$, their biases and their variances, will be derived.

## 3.1 Phase corrected vocoder (PCV)

This method is derived for the FM model ($\mu = 0$). The Fourier transform is not a direct estimator of the phase for chirp signals. An improvement to the phase vocoder consists in correcting the Fourier phase estimation, as in [5], using the error function $\Gamma(\Delta\beta, 0, \gamma; h)$. The estimation scheme proposed involves two steps:

1. Estimation of the corrected phases (modulo $2\pi$) $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and the unwrapping factor $\hat{n}$.

2. Estimation of $\beta_M$ using the phase vocoder formula[1]

$$\hat{\beta}_M = \frac{\mathrm{mod}(\hat{\alpha}_2) - \mathrm{mod}(\hat{\alpha}_1) + 2\pi\hat{n}}{T} \tag{3.8}$$

The function $\Gamma$ requires the knowledge of the frequencies corresponding to $t_{m_1}$ and $t_{m_2}$ (namely $\beta_1$ and $\beta_2$). Therefore the first step of the estimation scheme will involve a first frequency estimation for $\beta_1$ and $\beta_2$. As there is no knowledge about the FCR in this step, it is proposed to use one of the frequency estimators based on the classical sinusoidal model. Although these estimators

[1] mod() is the modulo $2\pi$ function

are biased for the FM model, it is shown in [5] that this scheme can greatly improve the precision on the phase estimates.

The parameter $\gamma$, and the unwrapping factor $n$ can be deduced from the frequencies $\beta_1$ and $\beta_2$, using the formulas:

$$\hat{\gamma} = \frac{\hat{\beta}_2 - \hat{\beta}_1}{T}$$
$$\hat{n} = \text{round}\left(\frac{1}{2\pi}\left(\text{mod}(\hat{\alpha}_1) - \text{mod}(\hat{\alpha}_2) + \frac{\hat{\beta}_1 + \hat{\beta}_2}{2}T\right)\right)$$

## 3.2 Reassigned vocoder (RV)

For this method, an accurate approximation of $\arg(\Gamma_2\bar{\Gamma}_1)$ will be derived. The first step is to express $\arg(\Gamma_1\bar{\Gamma}_2)$ as a function of $\beta_M$, the frequency to be estimated. In this process, $\Delta\beta_1$ and $\Delta\beta_2$ can be decomposed into two bounded terms $B$ and $G$, described below. Let's define $\omega_M$ as the mean bin frequency and $\Delta\omega$ as half the frequency variation in bins:

$$\omega_M = \frac{\omega_{k_1} + \omega_{k_2}}{2}, \qquad \Delta\omega = \frac{\omega_{k_2} - \omega_{k_1}}{2} \tag{3.9}$$

In addition, from the definition of the quadratic phase model (3.1), the FCR $\gamma$ follows this relation:

$$\gamma\frac{T}{2} = \frac{\beta_2 - \beta_1}{2} \tag{3.10}$$

Let $B = \beta_M - \omega_M$ and $G = \Delta\omega - \gamma\frac{T}{2}$. From the previous definitions, $\Delta\beta_i$ can be expressed as:

$$\Delta\beta_1 = B + G, \qquad \Delta\beta_2 = B - G \tag{3.11}$$

And since $\omega_{k_1}$ and $\omega_{k_2}$ are respectively the closest maximal bins to $\beta_1$ and $\beta_2$, then $|G| < R$, where $R$ is half the FT precision. In the pure FM case the relation $|B| < R$ is also verified. In the AM case the maximal bin $\omega_i$ are not anymore the closest bin to the $\beta_i$, and $B$ will be bounded by a constant $C$, depending on ACR, and such that $C > R$. The theoretical value of $C$ is difficult to obtain, but numerical analysis shows that $C$ increases very slowly when the ACR increases.

The first-order Taylor expansion in $G = 0$ of $\Gamma_1$ and $\Gamma_2$ is given by:

$$\Gamma_1 = \Gamma(B, \mu, \gamma; h) + G\Gamma'(B, \mu, \gamma; h) + \epsilon_1$$
$$\Gamma_2 = \Gamma(B, \mu, \gamma; h) - G\Gamma'(B, \mu, \gamma; h) + \epsilon_2$$

Where $\epsilon_1$ and $\epsilon_2$ are the Lagrange remainders. The frequency derivation property of the STFT leads to:

$$\Gamma_1 = \Gamma(B, \mu, \gamma; h) + jG\Gamma(B, \mu, \gamma; th) + \epsilon_1$$
$$\Gamma_2 = \Gamma(B, \mu, \gamma; h) - jG\Gamma(B, \mu, \gamma; th) + \epsilon_2$$

For an order 1 Taylor expansion in 0 of the argument function, we obtain:

$$\arg(\Gamma_1\bar{\Gamma}_2) = 2G\Re\left(\frac{\Gamma(B, \mu, \gamma; th)}{\Gamma(B, \mu, \gamma; h)}\right) + \epsilon \tag{3.12}$$

This approximation has proven to be quite accurate for the intervals of parameter considered. Indeed the deterministic bias has a magnitude of $10^{-3}$Hz in average for $\mu \in [0, 100]$ and $\gamma \in [0, 8000]$ (cf. section 3.4).

$\Re(\frac{\Gamma(B,\mu,\gamma;th)}{\Gamma(B,\mu,\gamma;h)})$ is in fact equivalent to the discrete version of the reassigned time. Indeed, the STFT can be rewritten as a function of $\Gamma$:

$$X(t_M, \omega_M; th) = \mathrm{e}^{\lambda_M + j\alpha_M}\, \Gamma(B, \mu, \gamma; th)$$
$$X(t_M, \omega_M; h) = \mathrm{e}^{\lambda_M + j\alpha_M}\, \Gamma(B, \mu, \gamma; h)$$

where $th(\tau) \triangleq \tau h(\tau)$ and $\alpha_M = \alpha + \beta\tau_M + \frac{\gamma}{2}\tau_M^2$. We can therefore conclude that:

$$\Re(\frac{\Gamma(B,\mu,\gamma;th)}{\Gamma(B,\mu,\gamma;h)}) = \Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)})$$

and

$$\arg(\Gamma_1\bar{\Gamma}_2) = 2G\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) + \epsilon$$

Using the previous expression in equation (3.7):

$$T\beta_M = \Delta\psi + 2\pi n + 2G\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) + \epsilon$$

Replacing $G$ by its definition leads to:

$$\begin{aligned}\beta_M + \gamma\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) = \frac{\Delta\psi + 2\pi n}{T} \\ + 2\frac{\Delta\omega}{T}\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) + \frac{\epsilon}{T}\end{aligned} \tag{3.13}$$

The left part of this expression is the frequency for the time: $\hat{t} = t_M + \Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)})$. The right part is the vocoder estimator corrected by a term depending on the reassigned time and on $\frac{2\Delta\omega}{T}$, which can be interpreted as a first FCR estimate using frequency bins.

## 3.3 Study of the unwrapping factor

The last problem to solve is the computation of the unwrapping factor $n$. It will be achieved using the estimator

$$\hat{n} = \mathrm{round}\big((\Omega T - \Delta\psi)/(2\pi)\big) \tag{3.14}$$

where $\Omega$ is a chosen reference frequency $\Omega = \omega_M$ [8]. This choice imposes a theoretical limit on the hop-size length of the phase vocoder, which is now discussed. From (3.7), $n$ verifies this relation:

$$n = \frac{1}{2\pi}\big[\omega_M T - \Delta\psi + \Delta\beta_M T - \arg(\Gamma_1\bar{\Gamma}_2)\big] \tag{3.15}$$

where $\Delta\beta_M = \beta_M - \omega_M$. The chosen estimator of $n$ is:

$$\hat{n} = \mathrm{round}\left(\frac{\omega_M T - \Delta\psi}{2\pi}\right) \tag{3.16}$$

Table 1: Maximal hop-size values in samples for the RV method ($N = 512$, $F = 16000$)

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $\gamma_m = 1000$ | 508 | 505 | 509 | 506 |
| $\gamma_m = 8000$ | 495 | 492 | 500 | 494 |

The condition for identity between $n$ and $\hat{n}$ is:

$$n = \hat{n} \Leftrightarrow |T\Delta\beta_M - \arg(\bar{\Gamma}_1\Gamma_2)| \leq \pi \tag{3.17}$$

But from the triangular inequality, we have:

$$|T\Delta\beta_M - \arg(\bar{\Gamma}_1\Gamma_2)| \leq T\Delta_m + \Gamma_m(\Delta_m, \mu_m, \gamma_m; h) \tag{3.18}$$

where $\Delta_m$ is the largest difference between $\beta_M$ and the maximal bin $\omega_M$. $\Gamma_m$ is the maximal value of the corrective term for the system parameters considered:

$$\Gamma_m(\Delta_m, \mu_m, \gamma_m; h) = \max_{|\Delta\beta_i| \leq \Delta_m, |\mu| \leq \mu_m, |\gamma| \leq \gamma_m} |\arg(\Gamma_1\bar{\Gamma}_2)| \tag{3.19}$$

Therefore, a sufficient condition for (3.17) to be verified is:

$$T\Delta_m \leq \pi - \Gamma_m(\Delta_m, \mu_m, \gamma_m; h) \tag{3.20}$$

As $T = H/F$ ($H$ is the hop-size in samples), we finally get:

$$H \leq \frac{\pi F}{\Delta_m}\left(1 - \frac{\Gamma_m(\Delta_m, \mu_m, \gamma_m; h)}{\pi}\right) \tag{3.21}$$

In the classical and AM cases ($\gamma_m = 0$), $\arg(\Gamma_1\bar{\Gamma}_2) = 0$ and we find the classical unwrapping condition $H \leq N$.

In the FM case, $\Delta_m$ is equal to $R$, half the Fourier precision:

$$H \leq N\left(1 - \frac{\Gamma_m(R, \mu_m, \gamma_m; h)}{\pi}\right) \tag{3.22}$$

Table 1 presents a numerical evaluation of the maximal hop-size values for various system parameters. It can be seen that the maximal theoretical hop-sizes decrease very slowly when the FCR increases. For usual applications, which use much lower hop-sizes than this limit, this means that the FCR will have no impact on the unwrapping estimation. The rectangular window cannot be used with this method (the time reassignment requires smooth functions) and is therefore not present in the table.

When the amplitude varies, the energy attributed to each frequency will be shifted depending on the ACR. The maximum of energy will no longer correspond to $\beta_M$, the sinusoid frequency for the middle of the window. Therefore, $\Delta_m$ will be superior to $R$, and will increase as the ACR increases. In the AM/FM rate model, the maximal theoretical hop-sizes for the unwrapping estimation are more difficult to compute in this case, as $\Delta_m$ is not explicitly known. They should be lower than the values presented in table 1. Nevertheless, this problem can be minimized by using intermediate phases.

Table 2: Evaluation of $\epsilon_{RV}$ for different values of $\mu_m$, and $\gamma_m$: maximal bias and absolute mean in Hz.

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $\mu_m = 10, \gamma_m = 2\pi 1000$ | 2E-2;1.6E-3 | 3E-2;2.6E-3 | 6.6E-3;5.3E-4 | 1.9E-2;1.8E-3 |
| $\mu_m = 100, \gamma_m = 2\pi 8000$ | 0.39;2.8E-3 | 0.49;3.2E-2 | 0.19;9.5E-3 | 0.37;2.3E-2 |

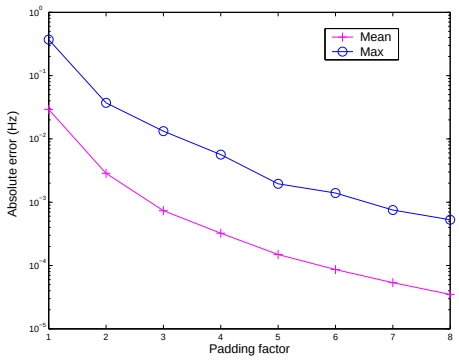



Figure 1: Influence of the padding factor on the RV method ($\mu_m = 100, \gamma_m = 2\pi 8000$, Hann window)

## 3.4 Study of the bias

All experiments are done for a window size of 32ms and a hop-size of 8ms.

### For the RV method

For high FCR, a bias will appear, caused by the approximation (3.12). Although the bias does not have a simple expression, it can be easily evaluated numerically as

$$\epsilon_{RV} = \frac{1}{T}\left|(\arg(\Gamma_1\bar{\Gamma}_2) - 2G\Re(\frac{\Gamma(B,\mu,\gamma;th)}{\Gamma(B,\mu,\gamma;h)})\right| \tag{3.23}$$

Table 2 shows values of this bias for different windows and two different ACR and FCR intervals. In each case, the first figure corresponds to the maximal bias, and the second figure is the mean value. The RV method is applied to the maximal bins of the Fourier Transform and is used without padding. The mean value is an average of 10000 experiments. It can be seen that the biases are kept within 1 Hz even for strong AM/FM modulations. If $\gamma_m = 0$, all the biases disappear. The bias only slightly increases when $\mu_m$ increases and is lower when the window is more concentrated in time, as for the Blackman window. The bias will be reduced even more if a padding factor is used (cf. Figure 1).

### For the PCV method

As for the RV method, the bias in the PCV does not have a simple mathematical expression and is evaluated numerically:

$$\epsilon_{PCV} = \frac{1}{T}\left|(\arg(\Gamma_1\bar{\Gamma}_2) - \arg(\hat{\Gamma}_1\bar{\hat{\Gamma}}_2))\right| \tag{3.24}$$

Table 3: Evaluation of $\epsilon_{PCV}$ and for different values of $\mu_m$, and $\gamma_m$: maximal bias and absolute mean in Hz.

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $\mu_m = 0, \gamma_m = 2\pi 8000$ | 0.5;5.6E-2 | 0.65;9.7E-2 | 0.27;2.3E-2 | 0.56;8.6E-2 |
| $\mu_m = 10, \gamma_m = 2\pi 1000$ | 0.37;8.6E-2 | 0.47;0.1 | 0.26;6.5E-2 | 0.43;9.6E-2 |
| $\mu_m = 10, \gamma_m = 2\pi 8000$ | 1.1;0.14 | 1.2;0.19 | 0.64;9.6E-2 | 1.0;1.7E-2 |

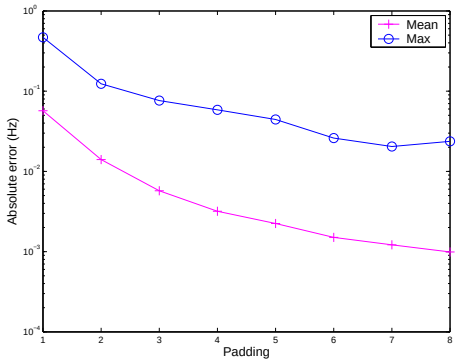



Figure 2: Influence of the padding factor on the PCV method ($\mu = 0, \gamma \in [0, 8000]$, Hann window)

In the first step, the frequency estimation method chosen is the interpolator described in [6]. The method is applied to maximum bins, without padding, and the results are based on an average of 10000 experiments. In the PCV case, Table 3 shows that the bias is within 1Hz for high FCR, if the ACR stays relatively small. For high ACRs, the method no longer works, as the PCV does not take into account this parameter. The use of a padding factor greatly decreases the biases (cf. Figure 2).

### 3.5 Study of the variance

The influence of a white noise on the classical phase vocoder is studied in [9]. A more recent reference [1] uses the same method, but presents a simpler formula applicable to any window. The results presented in [9, 1] are generalized here for the AM/FM model.

It is well known that when the frequency is constant, the Fourier Transform asymptotically resolves the sinusoid. In the FM case, the Fourier Transform will no longer resolve the chirp when $N \to \infty$[2]. Instead of an asymptotic property of the estimator, we will suppose that the sinusoid is well resolved. This has two consequences. First the energy of the noise is negligible compared to the energy of the sinusoids within the chosen bins, $X_i \gg N_i$. Secondly, for a chirp to be resolved, the width of the main lobe window must be superior to the maximum frequency variation of the chirp inside the window. If not, multiple peaks will appear on the main lobe. For example, if $\gamma_m$ is the largest FCR possible for the

[2]Indeed,for a pure FM signal, when $N$ tends to infinity, the chirp covers all the frequency range, with equal energy, and will no longer be resolved by the Fourier Transform.

signal considered, and if the main lobe of the window frequency response has a width of $K$ bins (independently of the window size $N$), this condition can be expressed as: $\gamma_m \frac{N}{F} < 2\pi \frac{KF}{N} \Leftrightarrow \frac{N}{F} < \sqrt{\frac{2\pi K}{\gamma_m}}$. For strong chirps ($\gamma_m \gg 1\pi K$), this condition can be expressed $\tau_N = N/F \ll 1s$.

If $S_i = S(t_{m_i}, \omega_{k_i}; h)$ and $N_i = N(t_{m_i}, \omega_{k_i}; h)$ are the Fourier Transform of $s$ and $n$ respectively, then:

$$S_i = X_i + N_i$$
$$S_i = X_i(1 + N_i^{'})$$

where $N_i^{'} \triangleq N_i/X_i$. The conjugate product $S_2\bar{S}_1$ can be written as:

$$S_2\bar{S}_1 = X_2\bar{X}_1(1 + Z) \tag{3.25}$$

where $Z \triangleq 1 + N_2^{'} + \bar{N}_1^{'} + N_2^{'}\bar{N}_1^{'}$.

As it is assumed that the STFT resolves the sinusoid $x$ from the disturbance $n$, $X_i \gg N_i$ for the bins close to the maximum, and $\arg(1 + Z) \approx \Im(Z) \approx \Im(N_2^{'}) - \Im(N_1^{'})$. $\arg(S_2\bar{S}_1)$ can be written as

$$\arg(S_2\bar{S}_1) \approx \arg(X_2\bar{X}_1) + \Im(Z)$$

From equation (3.7), this relation becomes:

$$\arg(S_2\bar{S}_1) \approx T\beta_M + \arg(\Gamma_2\bar{\Gamma}_1) + 2\pi n + \Im(Z)$$

And the expression of $\beta_M$ is:

$$\beta_M \approx \frac{\arg(S_2\bar{S}_1) - \arg(\Gamma_2\bar{\Gamma}_1) + 2\pi n - \Im(Z)}{T}$$

Let's note $\epsilon_{N,voc} = \Im(Z)$. The PCV and RV methods both use an estimate of $\arg(\Gamma_2\bar{\Gamma}_1)$ and $n$ to compute the frequency. As the sinusoid are supposed resolved from the noise, there will be no error in the estimation of $n$. It will now be proved that the stochastic error resulting from the estimation of $\arg(\Gamma_2\bar{\Gamma}_1)$ is negligible compared to $\epsilon_{N,voc}$.

## PCV case

In the PCV case, $\Gamma_i$ will be replaced by an estimate $\hat{\Gamma}_i = \Gamma(\hat{\beta}_i - \omega_i, 0, \hat{\gamma}; h)$, where $\hat{\beta}_i$ and $\hat{\gamma}$ are estimates computed using other Fourier-based estimators. The PCV estimation scheme has been derived for the FM model, *i.e.* $\mu = 0$. As there is no knowledge on $\gamma$ in a first step, the frequency estimator used will be based on the classical model, and will be biased when the slope is present. It will now be proved that the influence of the stochastic error from this first step is negligible.

Let $\epsilon_{D_i}^{\beta}$ and $\epsilon_{N_i}^{\beta}$ be respectively the deterministic and the stochastic error of the first step estimator for the frequency $\beta_i$. It is supposed that this estimator verifies the following assumptions:

1. $\hat{\beta}_i = \beta_i + \epsilon_{D_i}^{\beta} + \epsilon_{N_i}^{\beta}$
2. $\tau_N \epsilon_{N_i}^{\beta} \ll 1$

3. $E(\epsilon^{\beta}_{N_i}) = 0$

4. $\text{var}(\epsilon^{\beta}_{N_i}) \leq \text{var}(\Im(N_i^{'}))$

In the classical case, many Fourier-based estimators verify these assumptions asymptotically, in particular the discrete Fourier spectrum interpolators using phase, such as the methods described in [6, 10, 7]. If we suppose that the sinusoids are well resolved within the bins used, assumptions 1-4 will remain true.

The FCR estimate is defined as:

$$\hat{\gamma} = \frac{\hat{\beta}_2 - \hat{\beta}_1}{T} \tag{3.26}$$

Therefore, from the first assumption, the stochastic error of $\hat{\gamma}$ will also verify:

$$\hat{\gamma} = \gamma + \epsilon^{\gamma}_{D} + \epsilon^{\gamma}_{N} \tag{3.27}$$

$$\epsilon^{\gamma}_{N} \triangleq \frac{\epsilon^{\beta}_{N_2} - \epsilon^{\beta}_{N_1}}{T} \tag{3.28}$$

Let's define:

$$\Gamma_{k,i} = \Gamma(\beta_i + \epsilon^{\beta}_{D_i}\omega_i, 0, \gamma + \epsilon^{\gamma}_{D}; h\tau^k) \tag{3.29}$$

From the second assumption, the following approximation will hold:

$$\arg(\hat{\Gamma}_i) \approx \arg(\Gamma_{0,i}) + \arg(1 + j(\epsilon^{\beta}_{N}\frac{\Gamma_{1,i}}{\Gamma_{0,i}} + \epsilon^{\gamma}_{N}\frac{\Gamma_{2,i}}{2\Gamma_{0,i}})) \tag{3.30}$$

$$\approx \arg(\Gamma_i) + \arg(\frac{\Gamma_{0,i}}{\Gamma_i}) + \epsilon^{\beta}_{N_i}C_{1,i} + \epsilon^{\gamma}_{N}C_{2,i} \tag{3.31}$$

$$\approx \arg(\Gamma_i) + \epsilon_{D_i} + \epsilon_{N_i} \tag{3.32}$$

where,

$$C_{1,i} = \Im\left(\frac{\Gamma_{1,i}}{\Gamma_{0,i}}\right), \qquad C_{2,i} = \Im\left(\frac{\Gamma_{2,i}}{2\Gamma_{0,i}}\right) \tag{3.33}$$

Using equation (3.28), $\epsilon_{N_1}$ can be written as:

$$\epsilon_{N_1} = (C_{1,i} - \frac{C_{2,i}}{T})\epsilon^{\beta}_{N_1} + \frac{C_{2,i}}{T}\epsilon^{\beta}_{N_2} \tag{3.34}$$

As explained earlier, for a system aimed at analyzing chirps with high FCR, the relation $\tau_N \ll 1$ holds. From this relation, it can be proved that $C_{1,i} \ll 1$ and $C_{2,i}/T \ll 1$. From assumption 4, we can conclude that $\text{var}(\epsilon^{\beta}_{N_1}) \ll \text{var}(\Im(N_1^{'}))$. Similarly, it can be shown that $\text{var}(\epsilon^{\beta}_{N_2}) \ll \text{var}(\Im(N_1^{'}))$. Combining these results, we therefore have $\text{var}(\epsilon^{\beta}_{N_1} + \epsilon^{\beta}_{N_2}) \ll \text{var}(\epsilon_{N,voc})$.

In summary, if the frequency estimator of the first step verifies the assumption 1-4 and if the sinusoid is well resolved, the stochastic error due to the first estimates will be negligible, and the noised PCV equation can be written as:

$$\beta_M \approx \frac{1}{T}\arg(S_2\bar{S}_1) - \frac{1}{T}\arg(\hat{\Gamma}_2\bar{\hat{\Gamma}}_1) + 2\pi n + \epsilon_{PCV} - \frac{\Im(Z)}{T} \tag{3.35}$$

where $\epsilon_{PCV}$ is the deterministic error of the PCV method.

Table 4: Value of $Q$ for usual windows

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $Q(h)$ | 0.02 | 0.02 | 0.01 | 0.02 |

## RV case

In the RV case, $\arg(\Gamma_2\bar{\Gamma}_1)$ is replaced by the approximation (3.12), with a noise perturbation. In keeping with the previous notations, $S_M = S(t_M, \omega_M; h)$ (idem for $X_M$ and $N_M$) and $S_{M,1} = S(t_M, \omega_M; th)$ (idem for $X_{M,1}$ and $N_{M,1}$).

$$\begin{aligned}\widehat{\arg(\Gamma_1\bar{\Gamma}_2)} &= 2G\Re(\frac{S_{M,1}}{S_M}) \\ &= 2G\Re(\frac{X_{M,1}+N_{M,1}}{X_M+N_M})\end{aligned}$$

As in the PCV case, the sinusoid is supposed well resolved from the noise, $X(t_M, \omega_M; h) \gg N(t_M, \omega_M; h)$. The previous equation can be approximated as:

$$\begin{aligned}\widehat{\arg(\Gamma_1\bar{\Gamma}_2)} &\approx 2G\Re(\frac{X_{M,1}}{X_M}) + \epsilon_{N,RV} \\ \epsilon_{N,RV} &\triangleq 2G\Re(\frac{N_M}{X_M}\frac{\bar{X}_{M,1}}{\bar{X}_M} + \frac{N_{M,1}}{X_M})\end{aligned} \tag{3.36}$$

As in the PCV case, it will now be proved that $\epsilon_{N,RV}$ is negligible compared to $\epsilon_{N,voc}$.

We know that $\mathrm{var}(N_1) = \mathrm{var}(N_2) = \mathrm{var}(N_M)$. As $k_1$, $k_2$ and $k_M$ are maximum bins, if $\mu = 0$ then $|X_1| \approx |X_M| \approx |X_2|$. Recall that $N_i^{'} = N_i/X_i$, therefore if $\mu \neq 0$ we have:

$$\begin{aligned}\mathrm{e}^{\mu T}\,|X_1|^2 \approx |X_M|^2 \approx \mathrm{e}^{-\mu T}\,|X_2|^2 \\ \mathrm{e}^{-\mu T}\,\mathrm{var}(N_1^{'}) \approx \mathrm{var}(N_M^{'}) \approx \mathrm{e}^{\mu T}\,\mathrm{var}(N_2^{'}) \\ \mathrm{var}(N_M^{'}) \leq \max\left(\mathrm{var}(N_1^{'}), \mathrm{var}(N_2^{'})\right)\end{aligned}$$

But $\epsilon_{N,voc}$ also verifies:

$$\begin{aligned}\mathrm{var}(\epsilon_{N,voc}) &= \mathrm{var}(\Im(N_1^{'})) + \mathrm{var}(\Im(N_2^{'})) - E(\Im(N_1^{'})\Im(N_2^{'})) \\ \mathrm{var}(\epsilon_{N,voc}) &= \frac{1}{2}(\mathrm{var}(N_1^{'}) + \mathrm{var}(N_2^{'})) - E(\Im(N_1^{'})\Im(N_2^{'})) \\ \mathrm{var}(\epsilon_{N,voc}) &\approx \frac{1}{2}\max\left(\mathrm{var}(N_1^{'}), \mathrm{var}(N_2^{'})\right)\end{aligned}$$

Therefore the following relation is approximately verified:

$$\begin{aligned}\mathrm{var}(\epsilon_{N,voc}) &\geq \frac{1}{2}\,\mathrm{var}(N_M^{'}) \\ \mathrm{var}(\epsilon_{N,voc}) &\geq \mathrm{var}(\Re(N_M^{'}))\end{aligned} \tag{3.37}$$

From the definition of $G$, we have $|G| \leq R$ and $R = \pi\frac{F}{P_c N} = \frac{\pi}{P_c \tau_N}$, where $P_c$ is the padding factor used. As $\mathrm{var}(N_{M,1}) = \sigma^2 \sum_i h_i^2 \tau_i^2$ and $\mathrm{var}(N_M) =$

Table 5: Value of $Q^{'}$ for usual windows, for $N = 512$ and $\mu = 100$

| | Hann | Hamming | Blackman | Gaussian |
|---|---|---|---|---|
| $Q^{'}(h, 512, 100)$ | 0.03 | 0.04 | 0.02 | 0.04 |

$\sigma^2 \sum_i h_i^2$, the variance of $2G\Re(\frac{N_{M,1}}{X_M})$ verifies these inequalities:

$$\begin{aligned}
\operatorname{var}(2G\Re(\frac{N_{M,1}}{X_M})) &\leq \frac{2\pi^2}{P_c^2\tau_N^2}\frac{\sigma^2}{|X_M|^2}\sum_i h_i^2\tau_i^2 \\
\operatorname{var}(2G\Re(\frac{N_{M,1}}{X_M})) &\leq \frac{4\pi^2}{P_c^2\tau_N^2}\frac{\sum_i h_i^2\tau_i^2}{\sum_i h_i^2}\operatorname{var}(\Re(N_M^{'})) \\
\operatorname{var}(2G\Re(\frac{N_{M,1}}{X_M})) &\leq \frac{4\pi^2}{P_c^2}Q(h,N)\operatorname{var}(\Re(N_M^{'})) \\
Q(h,N) &\triangleq \frac{\sum_i h_i^2\tau_i^2}{\tau_N^2\sum_i h_i^2}
\end{aligned}$$

As $\frac{\sum_i h_i^2\tau_i^2}{\sum_i h_i^2}$ is $O(N^2)$, $Q(h,N)$ will have a finite limit $Q(h)$ as $N$ tends to infinity. For a rectangular window, $Q(h)$ is equal to $1/12$. For the other windows, a numerical evaluation of $Q(h)$ has been done in Table 4. From the value of Table 4, we can see that $Q(h) \ll 1$. Given that the padding factor $P_c$ is large enough, we have:

$$\operatorname{var}(2G\Re(\frac{N_{M,1}}{X_M})) \ll \operatorname{var}(\Re(N_M^{'})) \tag{3.38}$$

The variance of $2G\Re(\frac{X_{M,1}}{X_M}N_M^{'}))$ verifies this relation:

$$\operatorname{var}(2G\Re(\frac{X_{M,1}}{X_M}N_M^{'})) \leq \frac{4\pi^2}{P_c^2\tau_N^2}\left|\frac{X_{M,1}}{X_M}\right|^2\operatorname{var}(\Re(N_M^{'}))$$

Parseval's theorem states that for any signal $y$ with a DFT equal to $Y_k$ for the bin $k$:

$$\sum_{i=-(N-1)/2}^{(N-1)/2}|y_i|^2 = \frac{1}{N}\sum_{k=0}^{N-1}|Y_k|^2 \tag{3.39}$$

Applying this formula to $y_i = h(\tau_i)\tau_i x(\tau_i)$, we can conclude that:

$$|X_{M,1}|^2 < \sum_{k=0}^{N-1}|X(t_{m_M};\omega_{k_M};th)|^2 \tag{3.40}$$

$$|X_{M,1}|^2 < N\,\mathrm{e}^{2\lambda_M}\sum_{i=-(N-1)/2}^{(N-1)/2}h_i^2\tau_i^2\,\mathrm{e}^{2\mu\tau_i} \tag{3.41}$$

Consider now $y_i = h(\tau_i)x(\tau_i)$. From the hypothesis that the signal is well resolved, the Fourier Transform of $y_i$, which corresponds to $X(t_{m_M};\omega_{k_M};h)$, has its energy concentrated near the maximum $k_M$. Therefore, for the closest bin

$k_M$ to the maximum, we have: $|X(t_{m_M};\omega_{k_M};h)|^2 \approx \sum_{k=0}^{N-1} |X(t_{m_M};\omega_{k_M};h)|^2$. From Parseval's theorem, we can conclude that:

$$|X_M|^2 \approx N\,\mathrm{e}^{2\lambda_M} \sum_{i=-(N-1)/2}^{(N-1)/2} h_i^2\,\mathrm{e}^{2\mu\tau_i} \tag{3.42}$$

From equations (3.41) and (3.42), the following relation holds:

$$\mathrm{var}(2G\Re(\frac{X_{M,1}}{X_M}N_M^{'})) \leq \frac{4\pi^2}{P_c^2}Q^{'}(h,N,\mu)\,\mathrm{var}(\Re(N_M^{'}))$$
$$Q^{'}(h,N,\mu) \triangleq \frac{\sum_{i=-(N-1)/2}^{(N-1)/2} h_i^2\tau_i^2\,\mathrm{e}^{2\mu\tau_i}}{\tau_N^2\sum_{i=-(N-1)/2}^{(N-1)/2} h_i^2\,\mathrm{e}^{2\mu\tau_i}}$$

$Q^{'}$ is an increasing function of $\mu$, and has a finite limit when $N$ tends to infinity. When $\mu = 0$, we find that $Q^{'}(h,N,0) = Q(h,N)$. In the case that $\mu = 100$ and $N = 512$, Table 5 shows that $Q^{'}(h,N) \ll 1$. Given that the padding factor $P_c$ is large enough, we can say that:

$$\mathrm{var}(2G\Re(\frac{X_{M,1}}{X_M}N_M^{'})) \ll \mathrm{var}(\Re(N_M^{'})) \tag{3.43}$$

From (3.43) and (3.38), it holds that $\mathrm{var}(\epsilon_{N,RV}) \ll \mathrm{var}(\Re(N_M^{'}))$, and, using equation (3.37), that $\mathrm{var}(\epsilon_{N,RV}) \ll \mathrm{var}(\epsilon_{N,voc})$. Therefore, the RV noised estimation formula can be written as:

$$\beta_M + \gamma\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) \approx \frac{1}{T}\arg(S_2\bar{S}_1) + 2\frac{\Delta\omega}{T}\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)}) + 2\pi n + \epsilon_{RV} - \frac{\Im(Z)}{T}$$

where $\epsilon_{RV}$ is the deterministic error of the RV method.

## Expression of the variance for both methods

In summary, if the sinusoids are well resolved, the noised expression of the PCV and RV estimators can both be written as:

$$\hat{\beta} \approx \beta + \epsilon + \frac{\Im(Z)}{T} \tag{3.44}$$

where $\beta$ is the frequency to be estimated, which is $\beta_M$ for the PCV and $\beta_M + \gamma\Re(\frac{X(t_M,\omega_M;th)}{X(t_M,\omega_M;h)})$ for the RV estimator. $\hat{\beta}$ is the estimator for $\beta$ and is given by equation (3.8) for the PCV and equation (3.13) for the RV. $\epsilon$ is the deterministic bias. For both methods, the stochastic error is approximately $\Im(Z)/T$.

The expectation of the estimators is $\beta + \epsilon$, and their variance is given by:

$$\mathrm{var}(\hat{\beta}) = \frac{E(\Im(Z)^2)}{T^2} \tag{3.45}$$
$$= \frac{E((\Im(N_2^{'}) - \Im(N_1^{'}))^2)}{T^2} \tag{3.46}$$
$$= \frac{E\left(\Im(N_1^{'})^2 + \Im(N_2^{'})^2 - 2\Im(N_2^{'})\Im(N_1^{'})\right)}{T^2} \tag{3.47}$$

From the definition of $N_k^{'}$, we have:

$$\Im(N_k^{'}) = \mathrm{e}^{-L_k}\left(\sum_i h_i n^I_{m_k+i}\cos(\omega_k\tau_i+\Theta_k) - \sum_i h_i n^R_{m_k+i}\sin(\omega_k\tau_i+\Theta_k))\right) \tag{3.48}$$

where $L_k = \lambda_k + \log(|\Gamma_k|)$, $\Theta_k = \alpha_k + \arg(\Gamma_k)$. $m_k$ is the sample corresponding to the middle of the STFT number $k$. $n^I_{m_k+i}$ (resp. $n^R_{m_k+i}$) is the imaginary part (resp. real part) of the noise for the sample $m_k + i$. $\Im(N_k^{'})$ is a linear combination of independent, zero-mean random variables $n_i$, with the same variance and with real coefficients $a_i$: $\Im(N_k^{'}) = \sum_i a_i n_i$. Using the property $E(\Im(N_k^{'})^2) = E(n_i^2)\sum_i a_i^2$, we get:

$$E(\Im(N_k^{'})^2) = \frac{\sigma^2}{2}\,\mathrm{e}^{-2L_k}\,H_0 \tag{3.49}$$

where $H_0 = \sum_i h_i^2$.

Let's define the following variables:

$$\begin{aligned}
\Delta\lambda &\triangleq L_2 - L_1 = \tau_H\mu + \log(|\frac{\Gamma_2}{\Gamma_1}|)\\
\Delta\Theta &\triangleq \Theta_2 - \Theta_1 = \tau_H\beta_M + \arg(\Gamma_2\bar{\Gamma}_1)\\
\Delta\Phi &\triangleq \Delta\Theta - \tau_H\omega_M\\
H_1 &\triangleq \sum_{i=-(N-1-H)/2}^{(N-1-H)/2} h_{i+H/2}h_{i-H/2}\cos(\tau_i(\omega_1-\omega_2))
\end{aligned}$$

Now, the expectation of the cross term $\Im(N_2^{'})\Im(N_1^{'})$ will be derived. From the definition of $N_k^{'}$, we have:

$$\begin{aligned}
N_2^{'}\bar{N}_1^{'} &= \mathrm{e}^{-L_1-L_2-j\Delta\Theta}\sum_i h_i n_{m_2+i}\,\mathrm{e}^{-j\omega_{k_2}\tau_i}\sum_i h_i\bar{n}_{m_1+i}\,\mathrm{e}^{j\omega_{k_1}\tau_i}\\
E(N_2^{'}\bar{N}_1^{'}) &= \mathrm{e}^{-L_1-L_2-j\Delta\Theta}\sum_{i=-(N-1-H)/2}^{(N-1-H)/2} h_{i+H/2}h_{i-H/2}E(|n_i|^2)\,\mathrm{e}^{j(\omega_1\tau_{i+H/2}-\omega_2\tau_{i-H/2})}
\end{aligned}$$

In the second equation, we have used the fact that $E(n_k\bar{n}_l) = 0$ if $k \neq l$.

$$\begin{aligned}
E(N_2^{'}\bar{N}_1^{'}) &= \sigma^2\,\mathrm{e}^{-L_1-L_2-j\Delta\Phi}\sum_{i=-(N-1-H)/2}^{(N-1-H)/2} h_{i+H/2}h_{i-H/2}\,\mathrm{e}^{j\tau_i(\omega_1-\omega_2)}\\
E(N_2^{'}\bar{N}_1^{'}) &= \sigma^2\,\mathrm{e}^{-L_1-L_2-j\Delta\Phi}\,H_1
\end{aligned}$$

The last equality comes from the parity hypothesis on $h$. In a similar way, it can be proved that $E(N_2^{'}N_1^{'}) = 0$ because $E(n_i^2) = E({n_i^R}^2) - E({n_i^I}^2) + 2E(n_i^R n_i^I) = 0$. Let's remark that

$$\Im(N_1^{'})\Im(N_2^{'}) = \frac{1}{2}\Re(\bar{N}_1^{'}N_2^{'} - N_1^{'}N_2^{'}) \tag{3.50}$$

$$E(\Im(N_1^{'})\Im(N_2^{'})) = \frac{1}{2}(\Re(E(\bar{N}_1^{'}N_2^{'})) - \Re(E(N_1^{'}N_2^{'}))) \tag{3.51}$$

We can therefore conclude that

$$E(\Im(N_1^{'})\Im(N_2^{'})) = \sigma^2 \, \mathrm{e}^{-\Delta\lambda} \cos(\Delta\Phi) H_1 \tag{3.52}$$

Using equations (3.49) and (3.52), the variance of the estimators is finally obtained:

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\mathrm{e}^{L_1+L_2}} \frac{\Big[\cosh(\Delta\lambda) H_0 - \cos(\Delta\Phi) H_1\Big]}{T^2} \tag{3.53}$$

$$\mathrm{var}(\hat{\beta}) = \frac{\sinh(\mu\tau_W)}{\mu\tau_W} \frac{\Big[\cosh(\Delta\lambda) H_0 - \cos(\Delta\Phi) H_1\Big]}{\eta T^2 |\Gamma_2 \Gamma_1|} \tag{3.54}$$

where $\eta$ is the Signal to Noise Ratio (SNR),

$$\eta \triangleq \frac{\mathrm{e}^{\lambda_1+\lambda_2}}{\sigma^2} \frac{\sinh(\mu\tau_W)}{\mu\tau_W}$$

From equation (3.54), the variance for the AM, FM and classical models can be deduced directly.

For the classical model, $\mu = 0$, $\gamma = 0$, $\Delta\lambda = 0$ and $\omega_{k_1} = \omega_{k_2} = \omega$, $\beta_1 = \beta_2 = \beta$. Equation (3.54) simplifies to:

$$\mathrm{var}(\hat{\beta}) = \frac{\Big[H_0 - \cos(\Delta\Phi) H_1\Big]}{\eta T^2 |\Gamma|^2} \tag{3.55}$$

where,

$$\eta = \frac{\mathrm{e}^{2\lambda}}{\sigma^2}, \qquad H_1 = \sum_{i=-(N-H)/2}^{(N-H)/2} h_{i+\frac{H}{2}} \, h_{i-\frac{H}{2}},$$

$$\Delta\Phi = T(\beta_M - \omega_M), \qquad \Gamma = \sum_{i=-N/2}^{N/2} h_i$$

This last equation is the same as in [1].

Four examples are given on Figure 3(a) and 3(b). On Figure 3(a) the three upper curves correspond to the FM model and three lower one to the AM/FM model. In areas where the stochastic errors dominate, the theoretical variance matches the experimental MSE of the estimators. For the AM/FM model (upper curves), biases appear at high SNRs and low SNRs. In the former case, it is caused by the deterministic error of the estimator, and in the latter case, by the tracking scheme.

On Figure 3(a) the three upper curves corresponds to the FM model and three lower one to the AM/FM model. The theoretical variance matches the experimental curves. For the AM model, in the low SNRs case, the error due to the tracking scheme is slightly visible.

# 4 Conclusion

In this paper, it has been proved that the Fourier-based reassignment method is valid for an AM/FM model, using an original method.

The phase-vocoder frequency estimator has also been studied in the case of an AM/FM model. Two modified phase-vocoder-based schemes have been proposed: the Phase Corrected Vocoder (PCV) which aims at correcting the biased Fourier phases, and the Reassigned Vocoder (RV) which is an accurate estimator involving time reassignment. For both methods, the theoretical variance has been derived for a white-Gaussian-noise perturbation, and an experimental study of the biases has been done.

# References


[1] S. Abeysekera and K. Padhi. An investigation of window effects on the frequency estimation using the phase vocoder. *IEEE Transactions on Audio Speech and Signal Processing*, 14(4):1432–1439, Jul 2006.

[2] François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representation by the reassignment method. *IEEE Transaction on Signal Processing*, 43(5):1068–1088, May 1995.

[3] M. Betser, P. Collen, B. David, and G. Richard. Review and discussion on STFT-based frequency estimation methods. *Audio Engineering Society 120th Convention*, 2006.

[4] M. Betser, P. Collen, B. David, and G. Richard. Estimation of frequency for AM/FM models using the phase vocoder framework. *Journal of Signal Processing*, 2007.

[5] M. Betser, P. Collen, and J.-B. Rault. Accurate FFT-based phase estimation for chirp-like signals. *Audio Engineering Society 120th Convention*, 2006.

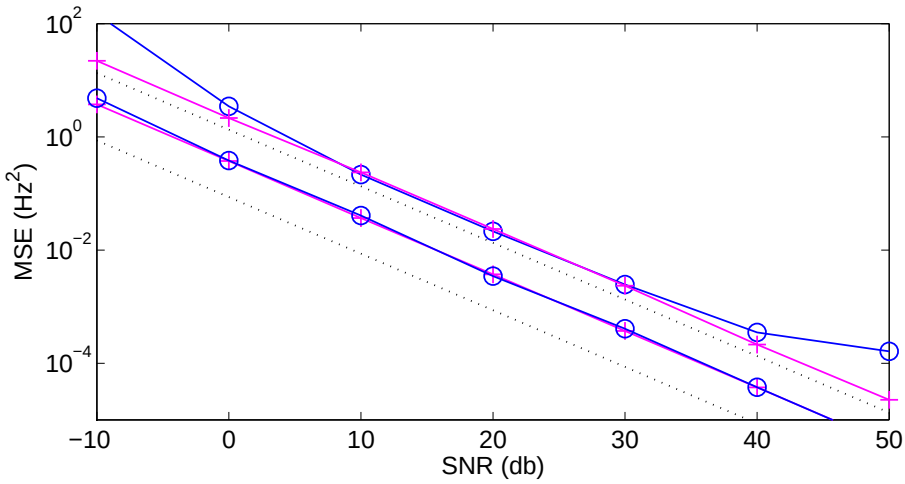
[6] M. Betser, P. Collen, and G. Richard. Frequency estimation based on adjacent DFT bins. *Proc. of EUSIPCO*, 2006.

[7] M. Macleod. Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Transaction on Signal Processing*, 46(1):141–148, Jan 1998.

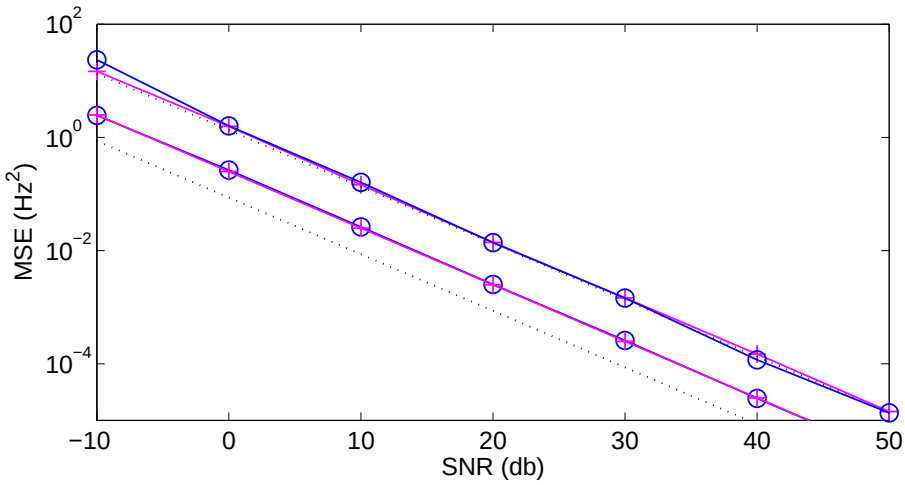
[8] R.J. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE transactions on Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754, Aug 1986.

[9] Miller S. Puckette and Judith C. Brown. Accuracy of frequency estimate using the phase vocoder. *IEEE Transaction on Speech and Audio Processing*, 6(2):166–176, Mar 1998.

[10] Barry G. Quinn. Estimating frequency by interpolation using Fourier coefficients. *IEEE Transactions on Signal Processing*, 42(5):1264–1268, May 1994.




(a) Upper curves correspond to the AM/FM model with $\mu \in [0, 100]$ and $\gamma \in [0, 8000]$, and lower curves to the FM model.




(b) Upper curves correspond to the AM model with $\mu \in [0, 100]$ and lower curves to the classical model.

Figure 3: Comparison of the theoretical vocoder variance ('+' markers) to the CRB (doted lines) and to the MSE of the RV method ('o' markers).

## Deuxième article : *Frequency estimation based on adjacent DFT bins*

Le deuxième article présenté concerne l'estimateur (6.1.8). Cet article, publié à EUSIPCO 2007, donne le détail de certaines démonstrations seulement esquissées à la section 6.1.1.1.

# FREQUENCY ESTIMATION BASED ON ADJACENT DFT BINS

*Michaël Betser, Patrice Collen,*

France Telecom R&D
35512 Cesson-Sévigné, France
first.name@francetelecom.com

*Gaël Richard.*

Telecom Paris
75634 Paris, France
first.name@enst.fr

## ABSTRACT

This paper presents a method to derive efficient frequency estimators from the Discrete Fourier Transform (DFT) of the signal. These estimators are very similar to the phase-based Discrete Fourier Spectrum (DFS) interpolators but have the advantage to allow any type of analysis window (and especially non-rectangular windows). As a consequence, it leads to better estimations in the case of a complex tone (cisoid) perturbed by other cisoids. Overall, our best estimator leads to results similar to those of phase vocoder and reassignment estimators but at a lower complexity, since it is based on a single Fast Fourier Transform (FFT) computation.

## 1. INTRODUCTION

Sinusoidal modeling [1] is a very popular and efficient representation for speech and music signals. It has led to numerous applications such as audio coding, analysis, synthesis and sound transformation. However, to be eligible for such applications, these models require accurate parameter estimations and, in particular, accurate frequency estimation.

Many frequency estimators use the Short Time Fourier Transform (STFT) as a starting point. Such estimators can be classified into three main categories: namely, *time methods* where the time of the STFT varies as in the phase vocoder [2] and the derivative method [1], *window methods* where the window varies as in the spectral reassignment [3], and *frequency methods* where the frequency varies as in the amplitude spectrum interpolation [4] or the phase-based DFS interpolators [5, 6]. A comparison of the state-of-the-art frequency methods can be found in [7, 8]. The latter methods have the advantage to require only one FFT computation since they only use adjacent frequency bins of this single FFT.

This article presents a new frequency estimator that is rather similar to the phase-based DFS interpolator. However, the method used to derive the frequency estimator is original and presents the advantage to allow the use of windowing, which was not the case in [5]. Similarly to previous work, it is supposed in this paper that the studied signal is a complex sinusoid with quasi-constant amplitude and frequency, in the neighborhood of a time $t$. Such a signal can be written as:

$$x(t+\tau) \triangleq \tilde{A}\,\mathrm{e}^{j\beta\tau} + w(t+\tau) \tag{1.1}$$

where $\tilde{A} \triangleq A\,\mathrm{e}^{j\alpha}$ is the complex amplitude of the sinusoid, ($A$ is the real amplitude and $\alpha$ the constant phase), and $\beta$ is the pulsation, both for the time $t$. $\tau$ is the local time in the neighborhood of the time $t$. This sinusoid can be perturbed by a complex noise $w$, which is supposed to be zero-mean white Gaussian. The asymptotic properties of the estimator presented in section 4 also holds under weaker noise properties described in [9, 6].

The analysis is based on the Short Time Fourier Transform (STFT) of the partial, defined as:

$$X(t,\omega_k;h) \triangleq \sum_{n=-N/2}^{N/2} x(\tau_n+t)\,h(\tau_n)\exp(-j\,\tau_n\omega_k) \tag{1.2}$$

where $N$ is the size in samples of the window support $h$, $F$ is the sampling frequency, $k$ is the frequency bin, $\tau_n = n/F$ is the time in seconds of the corresponding sample number ($n$ is an integer). Finally, $\omega_k = \frac{2\pi kF}{N}$ is the pulsation of the bin $k$. Here $N$ is supposed to be odd. This STFT corresponds to the centered form of the FT, which is sometimes called zero-phased FT, because the phase spectrum response of a symmetric window $h$ has a phase equal to zero in the neighborhood of the frequency zero. This definition is preferred here, because it will simplify the developments presented in this article. If $N$ is even, a centered form of the STFT is also possible, but will be slightly different as the sum will not be on integer values anymore. The STFT of the signal (1.1) can be put under the form:

$$X(t,\omega_k;h) = \tilde{A}\Gamma(\omega_k-\beta;h) + W(t,\omega_k;h) \tag{1.3}$$

where $\Gamma(\omega;h)$ is the discrete time FT, using the definition (1.2), of the window $h$ for the pulsation $\omega$, and $W$ is the STFT of the noise.

In practice, the definition of the FT used is often the linear-phased FT (i.e. the sum in the FT is done from 0 to $N-1$), as for many implementations of the FFT for example. A practical way to zero-phase the FFT is to perform a $(N-1)/2$ sample circular permutation of the windowed signal before computing the FFT [1]. In the remainder of this article, all the FT will be zero-phase.

## 2. PROPOSED METHOD

The proposed method combines Fourier Transforms (FT) computed for two different frequencies $\omega_1$ and $\omega_2$. The window $h$ is supposed to be symmetric, real and positive. In this part, the noise is not considered. Let's introduce the following FT ratio:

$$\mathscr{H} \triangleq \frac{X(t,\omega_1;h) - X(t,\omega_2;h)}{X(t,\omega_1;h) + X(t,\omega_2;h)} \tag{2.1}$$

$$= \frac{\Gamma(\omega_1-\beta;h) - \Gamma(\omega_2-\beta;h)}{\Gamma(\omega_1-\beta;h) + \Gamma(\omega_2-\beta;h)} \tag{2.2}$$

Because of its particularity, one can show that this ratio can be understood as a ratio of two FT differing in windows. Let $\Delta\omega \triangleq \frac{\omega_2-\omega_1}{2}$, $\omega_b \triangleq \frac{\omega_1+\omega_2}{2}$ and $\delta \triangleq \omega_b - \beta$, then $\mathscr{H}$ can

be written as:

$$\mathscr{H} = j\frac{\Gamma(\delta;h_s)}{\Gamma(\delta;h_c)} \tag{2.3}$$

where $h_s$ and $h_c$ are new analysis windows defined by:

$$h_s(\tau) \triangleq \sin(\Delta\omega\tau)h(\tau), \qquad h_c(\tau) \triangleq \cos(\Delta\omega\tau)h(\tau)$$

If $h$ is even, then $h_s$ is odd and $h_c$ is even.

Remark that $\mathscr{H}$ is necessarily real. In fact, as $h_c$ is symmetric, $\Gamma(\omega;h_c)$ is purely real, and as $h_s$ is anti-symmetric, $\Gamma(\omega;h_s)$ is purely imaginary. It will now be shown that an estimator can be defined from (2.3) and the parity hypothesis on $h$.

Taylor expansions around the frequency zero will be done. The frequency derivative property of the FT states that:

$$\frac{\partial^i\Gamma}{\partial\omega^i}(\omega;h) = (-j)^i\Gamma(\omega;\tau^i.h) \tag{2.4}$$

Let $\Gamma(h) = \Gamma(0;h)$. As $\Gamma(\tau^i.h_s) = 0$ if $i$ is even, and $\Gamma(\tau^i.h_c) = 0$ if $i$ is odd, the upper part of (2.3) will be expanded to an order 1 and the lower part to an order 0. $c_1$ and $c_2$ in $[0,\delta]$ exist such that:

$$\mathscr{H} = \frac{\Gamma(\tau.h_s)\delta - \Gamma(c_1;\tau^3.h_s)\frac{\delta^3}{6}}{\Gamma(h_c) - \Gamma(c_2;\tau^2h_c)\frac{\delta^2}{2}} \tag{2.5}$$

$$= \delta\frac{\Gamma(\tau.h_s)}{\Gamma(h_c)}\frac{(1-P)}{(1-Q)} \tag{2.6}$$

where $P$ and $Q$ are the Lagrange remainders,

$$P \triangleq \frac{\Gamma(c_1;\tau^3.h_s)}{\Gamma(\tau.h_s)}\frac{\delta^2}{6}, \qquad Q \triangleq \frac{\Gamma(c_2;\tau^2.h_c)}{\Gamma(h_c)}\frac{\delta^2}{2} \tag{2.7}$$

The values of $\Gamma(\tau^i.h_s)$ and $\Gamma(\tau^i.h_c)$ depend only on the analysis window $h$ and are known in advance. So if the corrective terms $P$ and $Q$ are small compared to 1 (cf section 3), an estimation of the frequency can be obtained as:

$$\hat{\beta} = \omega_b - \Re(\mathscr{H})\frac{\Gamma(h_c)}{\Gamma(\tau.h_s)} \tag{2.8}$$

In practice, $\mathscr{H}$ is never exactly real, this is why the real part ($\Re()$) of $\mathscr{H}$ is taken. When using an FFT, formula (2.8) can be applied to the two most energetic bins of the cisoid: the maximum DFS bin $k$ and the highest DFS bin between $k+1$ and $k-1$. In this case, $\Delta\omega$ is the half frequency resolution of the FFT and $\omega_b$ is the middle of the two bins selected.

*Algorithm:*

1. Initialization: compute $\Gamma(\tau.h_s)$ and $\Gamma(h_c)$ for the window $h$ used in the FFT.
2. Compute the zero-phased FFTs for a time $t$
3. Select the maximum bin $k = \arg\max_i |X(t,\omega_i;h)|$ and the second maximum $k' = \arg\max_{i\in\{k+1,k-1\}} |X(t,\omega_i;h)|$
4. Compute the ratio $\mathscr{H} = \frac{X(t,\omega_k;h) - X(t,\omega_{k'};h)}{X(t,\omega_k;h) + X(t,\omega_{k'};h)}$
5. Compute the estimated frequency: $\hat{\beta} = \omega_b - \Re(\mathscr{H})\frac{\Gamma(h_c)}{\Gamma(\tau.h_s)}$, where $\omega_b = (\omega_k + \omega_{k'})/2$

| | Han | Ham | Rec | Bla | Gau |
|---|---|---|---|---|---|
| $b_1$ | 1.5e-4 | 2.1e-2 | 3.9e-6 | 5.4e-3 | 2.4e-2 |
| $b_2$ | 1.3e-1 | 1.4e-1 | 2.5e-1 | 1.0e-1 | 1.3e-1 |
| Error bound(Hz) | 2.6e-3 | 3.8e-1 | 8.3e-5 | 9.4e-2 | 4.3e-1 |

Table 1: Bound values for different analysis windows.

## 3. ERROR BOUND

In this section the performances of the algorithm without noise are studied. Using equation (2.6) and the definition of the estimator (2.8), the error between $\beta$ and the estimation $\hat{\beta}$ can be rewritten as:

$$\beta - \hat{\beta} = \frac{(Q-P)}{(1-Q)}\delta \tag{3.1}$$

Since $\beta$ is inside $[\omega_1,\omega_2]$, $|\delta|$ is lower than the half frequency resolution of the DFT, $R = \pi F/N$.

We will first try to bound $P$ and $Q$. The FT of a real symmetric and positive window reaches its maximum in $\omega = 0$ and is decreasing for $\omega \in [0;R]$. Therefore $P$ and $Q$ are positive, and tight bounds on $P$ and $Q$ are:

$$P \le \frac{\Gamma(\tau^3.h_s)}{\Gamma(\tau.h_s)}\frac{R^2}{6}, \qquad Q \le \frac{\Gamma(\tau^2.h_c)}{\Gamma(h_c)}\frac{R^2}{2} \tag{3.2}$$

If no additional hypotheses on the window $h$ are made, $|P| \le \pi^2/24$ and $|Q| \le \pi^2/8$. It means that $Q$ could be equal to 1. Let's now suppose that $h$ verifies the following property:

$$\sum_{|n|\le N/2} h(n) \ge 2\sum_{N/4\le|n|\le N/2} h(n) \tag{3.3}$$

For all the usual windows, the energy of the center is superior to the energy of the edges. Consequently all the usual windows verify (3.3). With this hypothesis, the bound on $Q$ becomes: $|Q| \le 5\pi^2/64 < 1$, which proves that $1/(1-Q)$ is $O(1)$ [1]. As $P$ is also $O(1)$, and $\delta$ is $O(N^{-1})$, then, from (3.1), the estimate error $\beta - \hat{\beta}$ is $O(N^{-1})$, for all the windows verifying property (3.3). For some windows, the corrective terms $P$ and $Q$ will be of the same order, leading to smaller order of error: it has been shown that for the rectangular window, the estimator is $O(N^{-2})$ [6].

In order to find the bound on $P-Q$, the Lagrange remainders $P$ and $Q$ will be rewritten as :

$$P = \sum_{i=1}^{\infty}(-1)^i\frac{\Gamma(\tau^{2i+1}.h_s)}{\Gamma(\tau.h_s)}\cdot\frac{\delta^{2i}}{(2i+1)!} \tag{3.4}$$

$$Q = \sum_{i=1}^{\infty}(-1)^i\frac{\Gamma(\tau^{2i}.h_c)}{\Gamma(h_c)}\cdot\frac{\delta^{2i}}{(2i)!} \tag{3.5}$$

[1] Let $n$ be an integer variable which tends to infinity, let $g(n)$ be a positive function and $f(n)$ any function. Then $f = O(g)$ means that $|f| \le A.g$ for some constant $A$ and all values of $n$. [10]

$$E\{(\beta-\hat{\beta})^2\} = \delta^2\frac{(Q-P)^2}{(1-Q)^2} + \frac{2\sigma^2}{A^2\Gamma(\delta;h_c)^2}\left[\delta^2\Gamma({h_c}^2)\frac{(1-P)^2}{(1-Q)^2} + \frac{\Gamma(h_c)^2\Gamma({h_s}^2)}{\Gamma(\tau.h_s)^2}\right] + O(N^{-3.5}\ln(N)^{1.5}) \quad (4.4)$$

$$E\{(\beta-\hat{\beta})^2\} \lesssim \frac{R^2{b_1}^2}{(1-b_2)^2} + \frac{2\sigma^2}{A^2\Gamma(R;h_c)^2}\left[\frac{R^2\Gamma({h_c}^2)}{(1-b_1)^2} + \frac{\Gamma(h_c)^2\Gamma({h_s}^2)}{\Gamma(\tau.h_s)^2}\right] \quad (4.5)$$

---

A bound on $P-Q$ is therefore the infinite sum:

$$|P-Q| \leq \sum_{i=1}^{\infty}\frac{R^{2i}}{(2i+1)!}\left|\frac{\Gamma(\tau^{2i+1}.h_s)}{\Gamma(\tau.h_s)} - (2n+1)\frac{\Gamma(\tau^{2i}.h_c)}{\Gamma(h_c)}\right| \quad (3.6)$$

Alembert's rule shows that this bound is a convergent series, and as $R$ is usually small, this series converges fast. The first terms give a good approximation of this bound.

Let's note $b_1$ and $b_2$ the bounds on $P-Q$ (3.6) and $Q$ (3.2) respectively. Values of these bounds for different typical windows are given in table 1. As $b_2 < 1$, one can conclude that a bound on the error is:

$$|\beta-\hat{\beta}| \leq \frac{b_1}{1-b_2}|\delta| \quad (3.7)$$

**Theorem 1** *If the window h is real, symmetric, positive, and verifies the property* (3.3)*, then the error of the estimator* (2.8) *without considering the noise influence is at least* $O(N^{-1})$ *and is bounded by* $\frac{b_1}{1-b_2}R$*, where* $b_1$ *and* $b_2$ *are the bounds in equations* (3.6) *and* (3.2) *respectively, and R is the half Fourier resolution.*

In the last line of table 1, the bounds are given in Hz for $F = 16000$ and $N = 512$, and for various analysis windows. The small values of the bound $b1$ show that the two error terms $P$ and $Q$ compensate each other, especially for the rectangular window. This is why the first order Taylor expansion of equation (2.5), which seems a bit rough at first, can nevertheless give good results, depending on the window used. It can be noted that superior order Taylor expansions of (2.3) can lead to more precise estimators, but at the cost of an increase in complexity. In this case the frequency estimation is now one of the roots of a polynomial which has the same order as the expansion order.

## 4. STATISTICAL PROPERTIES OF THE ESTIMATOR

The noise influence on the performance of the estimator is discussed in this section. The analysis will be very similar to the one given in [6], as they consider almost the same estimator but only in the case of a rectangular window. It also follows the strategy adopted by Quinn in [9, 10].

Let $\mathscr{Z} \triangleq j\frac{\Gamma(\delta;h_s)}{\Gamma(\delta;h_c)}$ and $d \triangleq \tilde{A}\Gamma(\delta;h_c)$. From equation (2.3), if no noise is present, there is identity between $\mathscr{Z}$ and $\mathscr{H}$. From the definition of $\mathscr{H}$ in equation (2.1), and using equation (1.3), the ratio $\mathscr{H}$ has the form:

$$\mathscr{H} = \frac{\mathscr{Z} + \frac{(W_1-W_2)}{d}}{1+\frac{(W_1+W_2)}{d}} \quad (4.1)$$

where $W_1$ and $W_2$ are the STFT of the noise for the frequencies $\omega_1$ and $\omega_2$ respectively.

In [11], it is proved that if $w$ is white Gaussian, then $W(t,\omega;h)$ is $O(\sqrt{N\ln(N)})$ almost surely. This property is still true for more general assumptions on the noise, which are described in [11, 9]. If the function used to construct the discrete window $h$ is continuous, positive, and normalized, i.e. $\leq 1$ on the interval of definition, then $\sum_n h_n$ will be $O(N)$ and $1/(\sum_n h_n)$ will be $O(N^{-1})$. This property will be useful to determine the function orders. As $\Gamma(\delta;h_c)$ is $O(N)$, then $(W_1+W_2)/d$ is $O(N^{-1/2}\ln(N)^{1/2})$.

$$\mathscr{H} = \left(\mathscr{Z} + \frac{(W_1-W_2)}{d}\right)\left(1 - \frac{(W_1+W_2)}{d} + \frac{(W_1+W_2)^2}{d^2} + O(N^{-1.5}\ln(N)^{1.5})\right) \quad (4.2)$$

The expansion has been done to an order 2, because order 1 terms will be canceled when considering the expectation.

What interests us is the expectation of the squared error between the true frequency and the estimated frequency, which corresponds to the Mean Squared Error (MSE) in section 5:

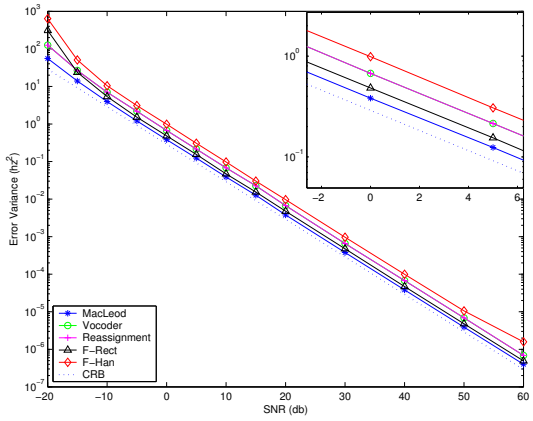
$$E\{(\beta-\hat{\beta})^2\} = E\left\{\left(\frac{\Gamma(h_c)}{\Gamma(\tau.h_s)}\Re(\mathscr{H}) - \delta\right)^2\right\} \quad (4.3)$$

Substituting equation (4.2) inside eq (4.3) leads us to an asymptotic development of the MSE. After some simplifications, one can found that this development is given by equation (4.4). If $N$ is large enough, a variance estimate, or at least a tight bound on the variance, could be computed for each value of $\delta$. We have chosen to present only the worst estimation case bound which is given by equation (4.5). This bound has been computed for different typical windows, represented in figure 1. The bound is composed of two terms: one corresponding to the deterministic error, and another to the error caused by the noise. If we consider this bound as a function of the SNR: $\sigma^2/A^2 = 10^{(-SNR/10)}$, the deterministic error will be constant, and the noise error will be a linear function of the SNR in a log scale. Therefore, when the deterministic error is dominant - i.e. for high SNRs - the estimator error will be constant, and when the noise error becomes dominant, for low SNRs, the error will be linear (in log scale). This explains the shape of the curves, in two parts, of the figure 1.

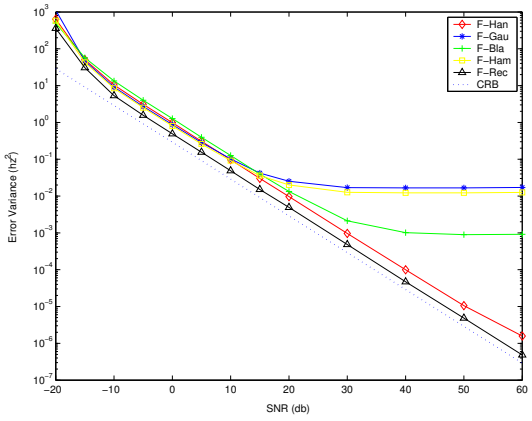
**Theorem 2** *If the function used to construct the window h is real, continuous, positive, normalized, symmetric and verifies the property* (3.3)*, then the estimator defined in* (2.8) *is asymptotically unbiased and, when N is large enough, a worst case bound on the variance is given by equation* (4.5)*.*

## 5. PERFORMANCE COMPARISON

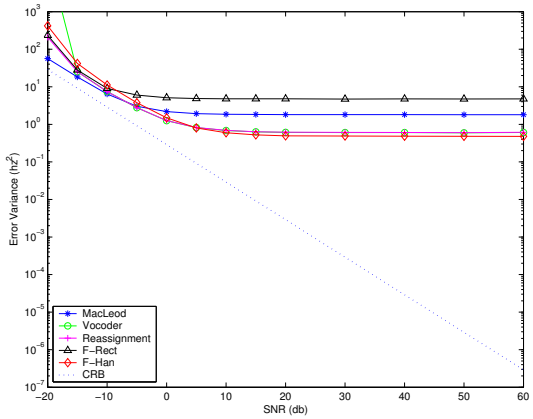
The purpose of this section is to compare the behavior of the new estimators to the classical ones. As studies comparing



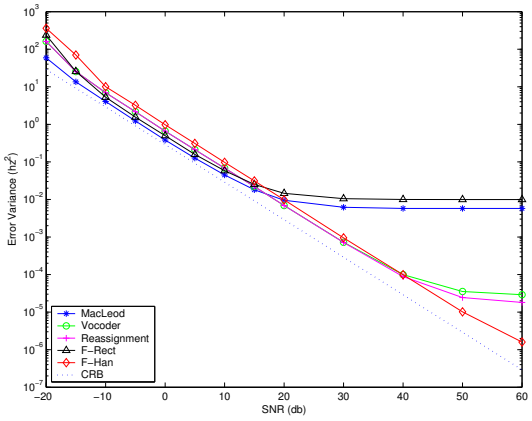

(a) Single cisoid comparison




(b) Window comparison




(c) Two cisoids with 100Hz separation




(d) Two cisoids with 1000Hz separation

Figure 2: Performance comparison

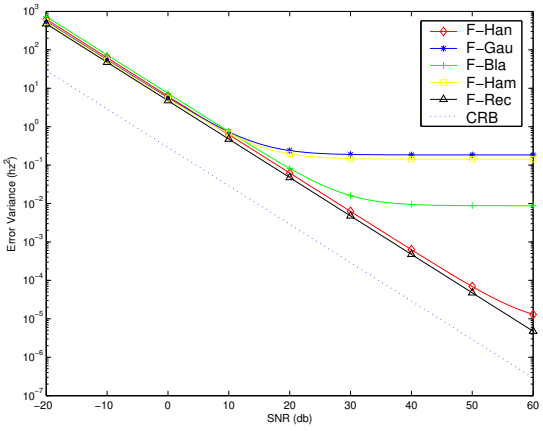



Figure 1: Theoretical window comparison

the classical frequency estimation methods have been done more than once [7], only the algorithms giving the best results will be considered, namely the classical phase vocoder ('Vocoder'), the reassignment method ('Reassignment'), and another interesting method, Macleod's 3 samples interpolator ('Macleod'). All the methods are summarized in table 2. The new method is named '$F$' followed by the first three letters of the window used.

In order to achieve a frequency estimation, peak detection is needed, but as our purpose is to compare the frequency estimators, it will be assumed in all experiments that the correct maximum bins are known. The second maximum bin is still supposed unknown.The classical Cramer-Rao Bound (CRB) framework is used to compare the estimators for different Signal to Noise Ratios (SNR) [5]. The CRB is represented by a dashed line.

The experiments are presented for $F = 16000$ and $N = 512$. The error between the true and estimated values is based on an average of the possible causes of error: on noise, using $K$ independent realizations, on frequency values, using randomly picked frequency in regularly spaced interval (10Hz size) over all the spectrum, and on the initial phase and amplitude, using random values between $[0, 2\pi[$ and $[0.1, 0.9]$ respectively. The noise variance is computed from the current amplitude value. In the experiments 2(c) and 2(d), the second sinusoid has the same amplitude as the first one.

Figure 2(b) shows the raw performances of the $F$ estimator on a single cisoid, for different analysis windows. As the SNR increases, the noise becomes negligible, and the inherent bias due to approximation (2.6) appears, as explained in section 4. The theoretical curves in figure 1 have the same shape and the same performance relations as the experimental curve. They have also approximately the same magnitude order as the experimental MSE. The shift between the theoretical curves and the experimental MSE is explained by the fact that the bound (4.5) corresponds to the worst estimation

| Name | Estimation | Window |
|---|---|---|
| Vocoder | $\hat{\beta} = F.(\angle X(t+1/F,\omega;h) - \angle X(t,\omega;h))$ | Han |
| Reassignment | $\hat{\beta} = \omega + \Im(\frac{X(t,\omega;h')}{X(t,\omega;\tau.h)})$ | Han |
| Macleod | $\hat{\beta} = \omega + \Delta\omega \frac{(\sqrt{1+8\delta^2}-1)}{4\delta}$, where $\delta = \frac{\Re((X(t,\omega-\Delta\omega;h)-X(t,\omega+\Delta\omega;h))X^*(t,\omega;h))}{\Re((2X(t,\omega;h)+X(t,\omega-\Delta\omega;h)+X(t,\omega+\Delta\omega;h))X^*(t,\omega;h))}$ | Rec |

Table 2: Summary of the different methods compared.

case.

The $F$ estimator is compared to the classical methods described in table 2. For clarity, only two different versions of the F estimator have been retained: F-Rec and F-Han. In the case of a single cisoid estimation, all methods give similar values and perform quite well as all the MSE are contained within 1db of the CRB. F-Han, because of its inherent bias, does not perform as well as the other estimators. The best results are obtained with Macleod's estimator, but F-Rec is very close. These two methods use the rectangular window, and the way Macleod derived his in [5] makes them very similar. If one takes a closer look to the formula of Macleod's estimator (table 2), one can see that it is the solution of an order 2 polynomial. As it has been mentioned in the previous section, the error made with approximation (2.6) may be reduced by using higher Taylor expansion orders, which is what is done in Macleod's estimator. The increase in performance is nonetheless quite small. For low SNR (-20db), the performances of F-Rec and F-Han drop faster than for other estimators. This is caused by the asymmetry of the method: the estimation is best done when using the maximum bin and the second highest bin, but for this noise level the second highest bin is hard to find. A solution, as for Quinn's estimator, is to compute the estimation for both bins around the maximum, and to use a test to determine which estimation is best [5]. But this error appears only in a failure area where all estimators perform badly.

Figures 2(c) and 2(d) represent the errors for an estimation perturbed by a second cisoid which is, respectively, at 100Hz and 1000Hz from the first cisoid. The methods using rectangular windows now give worse results than the others, except when the perturbation due to the second sinusoid becomes smaller than the noise perturbation. The other methods perform better because they use a Hann window which has a better side lobe attenuation. The F-Han method appears to be a good compromise between side lobe attenuation and single cisoid precision, comparable to the phase-vocoder and the reassignment.

## 6. CONCLUSION

This paper has presented a new frequency estimator based on frequency variations of the FFT. This estimator is very similar to the estimation called DFS interpolator using the phase, but the method presented allows the use of non-rectangular windows, which was not possible before. The advantage of using non-rectangular windows is to keep good performances even if the estimation is perturbed with close cisoids, which was the main default of the DFS interpolator using the phase. The results using a Hann window are similar to the performances of the classical phase vocoder and the reassignment method. In future work, we believe that using superior order expansion and testing different ratio filters $\mathscr{H}$ will lead to estimators performing better than the classical ones in all the scenarii.

## REFERENCES


[1] Sylvain Marchand, *Sound models for computer music (analysis, transformation, synthesis)*, Ph.D. thesis, University of Bordeaux, 2000.

[2] Eric Moulines and Jean Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, 1995.

[3] François Auger and Patrick Flandrin, "Improving the readability of time-frequency and time-scale representation by the reassignment method," *IEEE Transaction on Signal Processing*, vol. 43, no. 5, May 1995.

[4] Mototsugu Abe and Julius O. Smith III, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of fft magnitude peaks," *Audio Engineering Society 117th Convention*, 2004.

[5] M. Macleod, "Fast nearly ml estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Transaction on Signal Processing*, vol. 46, no. 1, pp. 141–148, Jan 1998.

[6] Elias Aboutanios and Bernard Mulgrew, "Iterative frequency estimation by interpolation on fourier coefficients," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1237–1241, Apr 2005.

[7] Konrad Hofbauer, "Estimating frequency and amplitude of sinusoids in harmonic signals," Tech. Rep., Graz University of Technology, Apr 2004.

[8] Stephen Hainsworth and Malcolm Macleod, "On sinusoidal parameter estimation," *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx)*, 2003.

[9] Barry G. Quinn, "Estimating frequency by interpolation using fourier coefficients," *IEEE Transactions on Signal Processing*, vol. 42, no. 5, pp. 1264–1268, May 1994.

[10] B. G. Quinn and E. J. Hannan, *The estimation and tracking of frequency*, Cambridge Univ. Press, 2001.

[11] Z.-G. Chen H.-Z. An and E. J. Hannan, "The maximum of the periodogram," *J. Multivariate Anal.*, vol. 13, pp. 383–400, 1983.

# Bibliographie de l'auteur

## Articles publiés pendant la thèse


Betser, M., Collen, P., David, B., and Richard, G. (2006a). Review and discussion on STFT-based frequency estimation methods. *Audio Engineering Society 120th Convention*.

Betser, M., Collen, P., David, B., and Richard, G. (2007a). Experimental and theoretical complements to the article 'estimation of frequency for AM/FM models using the phase vocoder framework'. *GET Technical report*.

Betser, M., Collen, P., David, B., and Richard, G. (2008). Estimation of frequency for AM/FM models using the phase vocoder framework. *IEEE transactions on Signal Processing*, 56(2) :505–518.

Betser, M., Collen, P., and Rault, J.-B. (2006b). Accurate FFT-based phase estimation for chirp-like signals. *Audio Engineering Society 120th Convention*.

Betser, M., Collen, P., and Rault, J.-B. (2007b). Audio identification using sinusoidal modelling, and application to jingle detection. *Proc. of ISMIR*.

Betser, M., Collen, P., and Richard, G. (2006c). Frequency estimation based on adjacent DFT bins. *Proc. of EUSIPCO*.


## Articles antérieurs


Ben, M., Betser, M., Gravier, G., and Bimbot, F. (2004). Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. *In proc. ICSLP (International Conference on Speech and Language Processing)*.

Betser, M. and Gravier, G. (2004a). Multiple events tracking in sound tracks. *In proc. ICME (International Conference on Multimedia and Expo)*.

Betser, M. and Gravier, G. (2004b). Recherche d'événements multiples dans les bandes sons. *In proc. CORESA (compression et représentation des signaux audiovisuel)*.

Betser, M., Gravier, G., and Gribonval, R. (2003). Extraction of information from video sound tracks : Can we decribe simultaneous events ? *In proc. CBMI (Contempt Based Multimedia Indexing).*

Coldefy, F., Bouthemy, P., Betser, M., and Gravier, G. (2004). Tennis video abstraction from audio and visual cues. *In Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP).*

## Brevêts déposés pendant la thèse

Betser, M., Collen, P., and Rault, J.-B. (2005). Patent : Procédé d'estimation de phase pour la modélisation sinusoïdale d'un signal numérique. Number : 05 53891-FR.

Betser, M., Collen, P., and Rault, J.-B. (2006a). Patent : Procédé d'estimation de phase pour la modélisation sinusoïdale d'un signal numérique. Number : 06 53507-FR.

Betser, M., Collen, P., and Rault, J.-B. (2006b). Patent : Procédé d'identification d'un objet audionumérique dans un flux audionumérique. Number : 06 55933-FR.

# Bibliographie


Abe, M. and Smith III, J. O. (2004). Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks. *Audio Engineering Society 117th Convention*.

Abe, M. and Smith III, J. O. (2005). AM/FM rate estimation for time-varying sinusoidal modeling. *Proc. of ICASSP*.

Abe, T., Kobayashi, T., and Imai, S. (1995). Harmonics tracking and pitch extraction based on instantaneous frequency. *Proc. of ICASSP*, pages 756–759.

Abeysekera, S. and Padhi, K. (2006). An investigation of window effects on the frequency estimation using the phase vocoder. *IEEE Transactions on Audio Speech and Signal Processing*, 14(4) :1432–1439.

Abotzoglou, T. (1986). Fast maximum likelihood joint estimation of frequency and frequency rate. *Proc. of ICASSP*, pages 1409–1412.

Aboutanios, E. and Mulgrew, B. (2005). Iterative frequency estimation by interpolation on Fourier coefficients. *IEEE Transactions on Signal Processing*, 53(4) :1237–1241.

Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., and Cremer, M. (2001). Audioid : towards content-based identification of audio material. *Audio Engineering Society 110th convention*.

Auger, F. and Flandrin, P. (1995). Improving the readability of time-frequency and time-scale representation by the reassignment method. *IEEE Transaction on Signal Processing*, 43(5) :1068–1088.

Badeau, R. (2005). *Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées.Application aux signaux de musique.* PhD thesis, Telecom Paris.

Badeau, R., Boyer, R., and David, B. (2002). Eds parametric modeling and tracking of audio signals. *Proc. of DAFx-02*, pages 139–144.

Bagshaw, P., Hiller, S., and Jack, M. (1993). Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. *Proc. of Eurospeech.*

Betser, M., Collen, P., David, B., and Richard, G. (2006a). Review and discussion on STFT-based frequency estimation methods. *Audio Engineering Society 120th Convention.*

Betser, M., Collen, P., David, B., and Richard, G. (2007). Experimental and theoretical complements to the article 'estimation of frequency for AM/FM models using the phase vocoder framework'. *GET Technical report.*

Betser, M., Collen, P., David, B., and Richard, G. (2008). Estimation of frequency for AM/FM models using the phase vocoder framework. *IEEE transactions on Signal Processing*, 56(2) :505–518.

Betser, M., Collen, P., and Rault, J.-B. (2006b). Accurate FFT-based phase estimation for chirp-like signals. *Audio Engineering Society 120th Convention.*

Betser, M., Collen, P., and Richard, G. (2006c). Frequency estimation based on adjacent DFT bins. *Proc. of EUSIPCO.*

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Institute of Phonetic Sciences*, 17 :97–110.

Boney, L., Tewfik, A., and Hamdy, K. (1996). Digital watermarks for audio signals. *in Proc. IEEE Multimedia.*

Boyer, R. and Abed-Meraim, K. (2004). Damped and delayed sinusoidal model for transient signals. *IEEE Transactions on Signal Processing*, 53(5).

Brown, J. (1991). Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1) :425–434.

Brown, J. (1992). Musical fundamental frequency tracking using a pattern recognition method. *Journal of the Acoustical Society of America*, 92(3) :1394–1402.

Brown, J. and Puckette, M. (1993). A high resolution fundamental frequency determination based on phase changes of the Fourier transform. *J. Acoust. Soc. Amer.*, 94 :662–667.

Burges, C., Patt, J., and Jana, S. (2001). Distortion discriminant analysis for audio fingerprinting. Technical Report MSR-TR-2001-116, Microsoft.

Cano, P. (1998). Fundamental frequency estimation in the sms analysis. *Proc. of the Digital Audio Effects Workshop (DAFX).*

Cano, P., Batlle, E., Gaomez, E., Gomes, L., and Bonnet, M. (2005). Audio fingerprinting : concepts and applications. *Studies in computational intelligence*, 2 :233–245.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2002). A review of algorithms for audio fingerprinting. *In International Workshop on Multimedia Signal Processing*.

Chan, K. and So, H. (2004). Accurate frequency estimation for real harmonic sinusoids. *IEEE Signal Processing Letters*, 11(7).

Choi, A. (1997). Real-time fundamental frequency estimation by least-square fitting. *IEEE Transaction on Speech and Audio Processing*, 5(2) :201–205.

Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (2001). *Introduction à l'algorithmique*. MIT presse.

Cox, I., Kilian, J., Leighton, T., and Shamoon, T. (1996). A secure, robust watermark for multimedia. *Workshop on information hiding*.

de Cheveigné, A. and Kawahara, H. (2001). Comparative evaluation of f0 estimation algorithms. *Proc. of Eurospeech*.

de Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4).

Desainte-Catherine, M. and Marchand, S. (2000). High precision Fourier analysis of sounds using signal derivatives. *J. Audio Eng. Soc.*, 48(7/8) :654.

Dieudonné, J. (1980). *Calcul infinitésimal*. Hermann, Paris.

Djuric, P. M. and Kay, S. M. (1990). Parameter estimation of chirp signals. *IEEE trans. on Acoustics, Speech and Signal Processing*, 38(12) :2118–2126.

Doval, B. and Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. *Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing*.

Doval, B. and Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmm's. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

Ferrari, A., Alengrin, G., and Theys, C. (1992). Estimation of the fundamental frequency of a noisy sum of cisoids with harmonic related frequencies. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 5 :517–520.

Flanagan, J. and Golden, R. (1966). Phase vocoder. *Bell System Technical Journal*, pages 1493–1509.

Fragoulis, D., Rousopoulos, G., Panagopoulos, T., Alexiou, C., and Papaodysseus, C. (2001). On the automated recognition of seriously distorted musical recordings. *TSP*, 49(9) :898–908.

Friedlander, B. and Francos, J. (1993). Estimation of amplitude and phase of non-stationary signals. *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, pages 848–861.

Gomes, L., Cano, P., Gomez, E., Bonnet, M., and Batlle, E. (2003). Audio watermarking and fingerprinting : for which applications ? *Journal of New Music Research*, 32(1) :65–81.

Goto, M. (2001). A predominant-F0 estimation method for CD recordings : MAP estimation using EM algorithm for adaptive tone models. *Proc. Workshop on Consistent and reliable acoustic cues for sound analysis*, pages 3365–3368.

Grey, J. M. and Moorer, J. A. (1977). Perceptual evaluation of synthesized musical intrument tone. *J. Acoust. Soc. Amer.*, 62 :454–462.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE trans. on Acoustics, Speech and Signal Processing*, ASSP-32 :236–243.

Griffin, D. and Lim, J. (1988). Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(8) :1223–1235.

Hainsworth, S. (2004). Time-frequency reassignment, a review and analysis. Technical Report CUED/F-INFENG/TR.459, Campbridge University Engineering Department.

Hainsworth, S. and Macleod, M. (2003a). On sinusoidal parameter estimation. *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx)*.

Hainsworth, S. and Macleod, M. (2003b). Time-frequency reassignment : measures and uses. *Proc. Cambridge Music Processing Colloquium*, page 36.

Haitsma, J. and Kalker, T. (2002). A highly robust audio fingerprinting system. *Proc. of the international conference on music information retrieval*.

Haitsma, J. and Kalker, T. (2003). Speed-change resistant audio fingerprinting using auto-correlation. *in Proc. ICASSP*.

Hermes, D. (1988). Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1) :257–264.

Hermus, K., Verhelst, W., and Wambacq, P. (2002). Psychoacoustic modeling of audio with exponentially damped sinusoids. *Proc. of ICASSP*, 2 :1821–1824.

Herre, J. (2004). Patent : Method and device for producing a fingerprint and method and for identifying an audio signal. Number : US0172411.

Hess, W. J. (1991). *Advances in speech signal processing*, chapter Pitch and voicing determination. Marcel Dekker, Inc., New York.

Hofbauer, K. (2004). Estimating frequency and amplitude of sinusoids in harmonic signals. Technical report, Graz University of Technology.

Hua, Y. and Sarkar, T. K. (1990). Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *Proc. IEEE Trans. Acoust., Speech, Signal Processing*, 38(5) :814–824.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing*. Prentice Hall, Upper Saddle River, NJ.

Ikram, M. Z., Abed-Meraim, K., and Hua, Y. (1996). Fast discrete quadratic phase transform for estimating the parameters of chirp signals. *Proc. IEEE conf. on Signals, Systems and Computer*, 1 :798–802.

Jensen, J., Heusdens, R., and Jensen, S. H. (2004). A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids. *IEEE Trans. Speech Audio*, 12(2) :121–132.

Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. (1999). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. *Proc. of Eurospeech*, 6 :2781–2784.

Kay, S. M. (1993). *Estimation Theory*. Prentice Hall, Upper Saddle River, NJ 07458.

Keiler, F. and Marchand, S. (2002). Survey on extraction of sinusoids in stationary sounds. *Proc. of the 5th Int. Conference on Digital Audio Effects*.

Keiler, F. and Zölzer, U. (2001). Extracting sinusoids from harmonic signals. *Journal of New Music Research*, 30(3) :243–258.

Klapuri, A. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE trans. on Speech and Audio Processing*, 11(6) :804–816.

Klapuri, A. (2004). *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology.

Kodera, K., Gendrin, R., and de Villedary, C. (1978). Analysis of time-varying signals with small BT values. *IEEE transaction on Acoustic, Specch and Signal Processing*, 26(1) :64–76.

Kunt, M. (1999). *Traitement numérique des signaux*. Presses Polytechniques et Universitaires Romandes.

Lagrange, S., Delanoue, N., and Jaulin, L. (2007). On sufficient conditions of injectivity, developpment of a numerical test via interval analysis. *Reliable Computing*, 13(5) :409–421.

Laroche, J. (2001). Patent : Process for identifying audio content. Number : WO88900.

Laroche, J. and Dolson, M. (1999). New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 91–94.

Levine, S. and Smith, J. (1998). A sines+transients+noise audio representation for data compression and time/pitch scale modifications. *Proc. of the 105th AES convention.*

Macleod, M. (1998). Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Transaction on Signal Processing*, 46(1) :141–148.

Maher, R. C. and Beauchamp, J. W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of Acoustic al Society of America*, pages 2254–2263.

Mallat, S. (2000). *Une exploration des signaux en ondelettes.* Ellipse.

Mandel, M. (2005). Audio fingerprinting for recognition.

Marchand, S. (2000). *Modélisation informatique du son musical (analyse, transformation, synthèse).* PhD thesis, Université de Bordeaux.

Marchand, S. (2001). An efficient pitch-tracking algorithm using a combination of Fourier transforms. *Proc. of Conference on Digital Audio Effects.*

Marques, J. S. and Almeida, L. B. (1986). A background for sinusoid-based representation of voiced speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1233–1236.

Masri, P. (1996). *Computer modeling of sound for transformation and synthesis of musical signals.* PhD thesis, University of Bristol.

Master, A. S. (2002). Non-stationary sinusoidal model frequency parameter estimation via Fresnel integral analysis. Technical Report EE 391, Stanford University.

Master, A. S. and Liu, Y. (2003a). Robust chirp parameter estimation for hann windowed signals. *IEEE Int. Conf. on Multimedia and Expo (ICME).*

Master, A. S. and Liu, Y.-W. (2003b). Non-stationary sinusoidal modeling with efficient estimation of linear frequency chirp parameters. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

McAulay, R. and Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE transactions on Acoustics, Speech and Signal Processing*, ASSP-34(4) :744–754.

McAulay, R. and Quatieri, T. (1990). Pitch estimation and voicing detection based on a sinusoidal speech model. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing.*

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE transactions on Signal Processing*, 39(1) :40–48.

Moore, B. (1997). *Introduction to the psychology of hearing.* Academic press ; 4th edition.

Moulines, E. and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech communication*, pages 174–215.

Nieuwenhuijse, J., Heusdens, R., and Deprettere, E. (1998). Robust exponential modeling of audio signals. *in Proc. ICASSP.*

Ogle, J. and Ellis, D. (2007). Fingerprinting to identify repeated sound events in long-duration personal audio recordings. *in Proc. ICASSP.*

Papaodysseus, C., Roussopoulos, G., Fragoulis, D., Panagopoulos, T., and Alexiou, C. (2001). A new approach to the automatic recognition of musical recordings. *Journal of the audio engineering society*, 49(1) :23–35.

Peeters, G. and Rodet, X. (1999). Sinola : A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum. *Proc. of ICMC.*

Peleg, S. and Porat, B. (1991). Linear FM signal parameter estimation from discrete-time observations. *IEEE Trans. Aerosp. Electron. Syst.*, 27 :607–614.

Picinbono, B. (1997). On instantaneous ampplitude and phase of signals. *IEEE transaction on Signal Processing*, 45(3) :552–560.

Pinquier, J. and André-Obrecht, R. (2004). Jingle detection and identification in audio document. *Proc. of ICASSP.*

Pisarenko, V. F. (1973). The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomy Society*, 33 :347–366.

Portnoff, M. R. (1981). Short-time Fourier analysis of sampled speech. *IEEE transaction on Acoustics, Speech and Signal Processing*, 29(3) :364–374.

Puckette, M. S. (1995). Phase-locked vocoder. *Proc. IEEE ASSP workshop on Applications of Signal Processing to Audio and Acoustics.*

Puckette, M. S. and Brown, J. C. (1998). Accuracy of frequency estimate using the phase vocoder. *IEEE Transaction on Speech and Audio Processing*, 6(2) :166–176.

Quinn, B. G. (1994). Estimating frequency by interpolation using Fourier coefficients. *IEEE Transactions on Signal Processing*, 42(5) :1264–1268.

Quinn, B. G. and Hannan, E. J. (2001). *The estimation and tracking of frequency.* Cambridbge Univ. Press.

Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, New Jersey.

Raghuram, R. (2002). Pitch and voicing determination of speech signals. Master's thesis, Indian Institute of Technology, Madras.

Riche de Prony, G. M. (1795). Essai expérimental et analytique : sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool à différentes températures. *Journal de l'école polytechnique*, 1(22) :24–76.

Rife, D. C. and Boorstyn, R. R. (1976). Multiple-tone parameter estimation from discrete-time observations. *The Bell system technical journal*, 55(9) :1389–1410.

Roy, R., Paulraj, A., and Kailath, T. (1986). Esprit : A subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5) :1340–1342.

Saha, S. and Kay, S. M. (2002). Maximum likelihood parameter estimation of superimposed chirps using Monte-Carlo importance sampling. *IEEE trans. on Signal Processing*, 50(2) :224–230.

Schmidt, R. O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transaction on Antennas and Propagation*, 34(3) :276–280.

Serra, X. (1989). *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, CCRMA, Department of Music, Stanford University.

Serra, X. (1997). *Musical sound modeling with sinusoids plus noise*. Swets&Zeitlinger.

Smith III, J. O. and Serra, X. (1987). Parshl : A program for the analysis/synthesis of inharmonic sounds based on a sinusoidal representation. *ICMC*.

Stoica, P., Li, H., and Li, J. (2000). Amplitude estimation of sinusoidal signals : Survey, new results, and an application. *IEEE Transactions on Signal Processing*, 48(2) :338–351.

Stoica, P. and Nehorai, A. (1988). Statistical analysis of two non-linear least-squares estimators of sine wave parameters in the colored noise case. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 2408–2411.

Tabrikian, J., Dubnav, S., and Dickalov, Y. (2004). Maximum a posteriori probability pitch tracking in noisy envirinments using harmonic model. *IEEE Transaction on Speech and Audio Processing*, 12(1) :76–87.

Talkin, D. (1995). *Speech coding and synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT). Elsevier, New York.

Umapathy, K., Krishnan, S., and Jimaa, S. (1984). A pitch detection algorithm with hypothesis and test strategy by means of fast surface AMDF. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*.

Vafin, R., Heusdens, R., and Kleijn, W. B. (2001a). Improved modeling of audio signals by modifying transient locations. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Vafin, R., Heusdens, R., and Kleijn, W. B. (2001b). Modifying transients for efficient coding of audio. *Proc. of ICASSP*.

Verma, T. and Meng, T. (1998). An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio. *Proc. of ICASSP*.

Wang, A. (2003). An industrial-strength audio search algorithm. *Proc. of the 4th Int. Symposium on Music Information Retrieval (ISMIR)*.

Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM*, 49(8) :44–48.

Wolcin, J. (1980). Maximum a posteriori estimation of narrow-band signal parameters. *Journal of the Acoustical Society of America*, 68(1) :174–178.

Xia, X. (2000). Discrete chirp-Fourier transform and its application to chirp rate estimation. *IEEE transactions on signal processing*, 48(11) :3122–3133.

Yang, H., Qiu, L., and Koh, S. (1994). Application of instantaneous frequency estimation for fundamental frequency detection. *Proc. of ICASSP*, pages 616–619.

Zhou, G., Giannakis, G., and Swami, A. (1996). On polynomial phase signals with time-varying amplitudes. *IEEE transactions on signal processing*, 44(4) :848–861.

Zwicker, E. and Feldtkeller, R. (1981). *Psychoacoustique*. MASSON.