POLYPUBLIE

Techniques d'apprentissage machine appliquées à la modélisation du timbre d'un individu pour la synthèse de voix réaliste

Pierre-Alexandre Fournier

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2005
Type	Mémoire ou thèse / Dissertation or Thesis Mémoire de maîtrise/Master's thesis
URL de PolyPublie PolyPublie URL	https://publications.polymtl.ca/7756/
Conditions d'utilisation Terms of use	© Tous droits réservés / All rights reserved
Référence Citation	Fournier, P.-A. (2005). Techniques d'apprentissage machine appliquées à la modélisation du timbre d'un individu pour la synthèse de voix réaliste [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7756/

Détails de la thèse
Thesis details

Directeurs de recherche Advisors	Jean-Jules Brault
Programme Program	Apprentissage machine

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal
https://publications.polymtl.ca

UNIVERSITÉ DE MONTRÉAL

TECHNIQUES D'APPRENTISSAGE MACHINE APPLIQUÉES À LA MODÉLISATION DU TIMBRE D'UN INDIVIDU POUR LA SYNTHÈSE DE VOIX RÉALISTE

PIERRE-ALEXANDRE FOURNIER
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(APPRENTISSAGE MACHINE)
JUILLET 2005

© Pierre-Alexandre Fournier, 2005.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

TECHNIQUES D'APPRENTISSAGE MACHINE APPLIQUÉES À LA MODÉLISATION DU TIMBRE D'UN INDIVIDU POUR LA SYNTHÈSE DE VOIX RÉALISTE

présenté par: FOURNIER Pierre-Alexandre
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de:

M BOYER François-Raymond, PhD, président
M BRAULT Jean-Jules, PhD, membre et directeur de recherche
M LABIB Richard, PhD, membre

À mes parents.

REMERCIEMENTS

Je dois d'abord remercier mon directeur Jean-Jules Brault pour m'avoir fait découvrir une tonne de sujets de recherche intéressants. Des réseaux de neurones aux modèles graphiques probabilistes en passant par l'informatique quantique et la génétique, nos discussions m'ont ouvert les yeux sur les limites et les possibilités de domaines dont je n'avais pas conscience auparavant. Je voudrais le remercier aussi pour m'avoir laissé assez de liberté et de temps pour faire mes expériences dans tous ces domaines, pour avoir pris du temps pour discuter avec moi de sujets beaucoup plus larges que ma recherche, et d'avoir insisté pour que j'aie du plaisir à chercher à comprendre l'univers.

Merci au Fonds québécois de la recherche sur la nature et les technologies pour avoir financé une partie de mes recherches. L'aide financière est un aspect très important de la vie des étudiants gradués, je suis reconnaissant d'avoir eu la chance d'être soutenu par le FQRNT pendant les vingt mois que duraient ma bourse.

Je veux aussi remercier mes collègues et amis pour leur support, leurs idées et leurs conseils techniques et personnels. Je pense entre autres à Guillaume Germain, Jean-François Roy, Ghislain Provost, Olivier Duval et Jeffrey Dungen, pour n'en nommer que quelques-uns. Un merci spécial à François Lanctôt pour ses nombreuses idées et suggestions pour le traitement de la voix.

Merci à mes parents pour m'avoir soutenu et encouragé à étudier depuis 1978. J'ai parfois pris des libertés et des détours dans la vie mais vous m'avez toujours fait confiance.

Enfin merci à mon épouse Sahar pour son amour, sa patience et sa foi.

RÉSUMÉ

La synthèse de voix par ordinateur permet maintenant de produire des phrases quelconques avec des voix réalistes pour les interfaces humain-machine et la création artistique. La construction de chaque voix nécessite cependant des mois de travail à une équipe de linguistes, informaticiens et ingénieurs. Chaque phonème ou diphone doit être isolé et caractérisé pour pouvoir être synthétisé par la suite dans un mot ou une phrase.

Nous voudrions arriver un jour au point où un programme aurait une assez bonne connaissance des propriétés acoustiques de la voix humaine pour pouvoir apprendre automatiquement une nouvelle voix à partir de quelques exemples de phonèmes. Notre objectif est d'entamer la recherche menant à ce but. Jusqu'ici, les recherches publiées sur la variabilité du timbre entre les locuteurs avaient principalement pour objectif d'améliorer la reconnaissance vocale. À notre connaissance, nos travaux sont les premiers à étudier le concept de variabilité du timbre dans un contexte de synthèse vocale.

Notre recherche porte sur la prédiction du timbre de voyelles. Nous avons posé l'hypothèse que lorsque la distance spectrale entre deux voyelles de deux locuteurs est petite, elle l'est en général pour toutes les voyelles. La prédiction des caractéristiques acoustiques d'une voyelle est une étape importante dans la reproduction du timbre vocal d'un individu, parce que les voyelles contiennent beaucoup d'information sur l'identité du locuteur. Cette recherche étudie donc un sous-problème de l'approximation du timbre de locuteurs dans le but de mieux en comprendre les difficultés et les possibilités.

Ce mémoire compare quelques méthodes d'apprentissage machine pour compléter le corpus de voyelles d'un locuteur à partir des voyelles disponibles de celui-ci et des corpus d'autres locuteurs. Il n'y a pas de modèle de relations acoustiques simple et général pour décrire les variations phonétiques entre les individus. C'est

pourquoi l'approche par apprentissage machine, qui nous permet de calculer ces relations à partir de données, nous a semblé raisonnable. Nous avons choisi des modèles à noyaux, l'analyse par composantes principales et l'analyse par composantes indépendantes comme méthodes d'estimation (de prédiction) de voyelles. La faible quantité de données (5 voyelles provenant de 31 locuteurs) et le grand nombre de dimensions de celles-ci (200 par voyelle) nous a contraint à utiliser ces méthodes, les réseaux de neurones et les modèles graphiques probabilistes (pour ne nommer que ces méthodes) ayant besoin de plus de données pour éviter le surentraînement. Nous avons utilisé le modèle appelé harmonique et résiduel (*harmonic and noise model*) pour représenter la voix[1]. Ce modèle nous a semblé le meilleur pour reproduire une voix d'assez bonne qualité pour retrouver les caractéristiques acoustiques d'un locuteur. La qualité de la représentation de la voix est importante dans notre cas, parce que nous voulons que la synthèse soit réaliste et non seulement intelligible. Nous avons comme objectif ultime que des voyelles prédites d'un locuteur puissent être identifiées comme ayant le timbre de ce locuteur lors de la synthèse d'une phrase.

La première partie de ce document introduit les concepts de production et de perception de la parole ainsi que quelques modèles mathématiques utilisés pour l'analyse et la synthèse de voix, soit la prédiction linéaire, le modèle sinusoïdal harmonique et le modèle harmonique et résiduel (HNM).

La seconde partie expose les différents problèmes de l'apprentissage machine et décrit les différentes méthodes de prédiction explorées dans le cadre de cette recherche, soit les noyaux et les algorithmes génétiques.

La troisième partie décrit la méthodologie utilisée dans nos expériences pour la saisie de données, la normalisation des voyelles des locuteurs et l'estimation (la

[1]Attention : ne pas confondre les modèles de relations acoustiques issus des méthodes d'apprentissage machine (noyaux, PCA, ICA) avec le modèle de représentation de la voix (HNM) pour la synthèse vocale.

prédiction) des caractéristiques acoustiques des voyelles manquantes.

La dernière partie compare les résultats expérimentaux obtenus avec les différentes méthodes d'apprentissage machine étudiées. Nous montrons qu'il est possible de réduire de 36,6% le niveau de distorsion harmonique moyen d'une voyelle inconnue d'un locuteur à partir de la connaissance de quatre autres voyelles. La réduction de distorsion harmonique est calculée en comparant la prédiction avec une voyelle moyenne de tous les locuteurs disponibles. La fonction d'erreur utilisée est l'erreur quadratique du spectre de puissance harmonique. Le meilleur résultat est obtenu avec un noyau gaussien asymétrique utilisant une fonction de distance optimisée par un algorithme génétique.

Nos résultats permettent de penser qu'il est possible d'apprendre un modèle général du timbre permettant de tirer profit de l'information acoustique disponible d'un individu afin d'en compléter le corpus de phonèmes.

ABSTRACT

Computer-based voice synthesis can now produce arbitrary utterances with realistic voices for human-machine interfaces and artistic creation. To achieve this, we have to first go through the demanding and expensive process of building a voice. This process needs linguists, computer scientists and engineers to work for months to build a single voice. Each phoneme needs to be isolated and characterized in order to be synthesized in a word.

We would like to write a computer program that would know enough about acoustical properties of the human voice so that it could learn automatically a new voice with only a few examples. Our objective is to make the first steps towards that goal. The research on acoustical speaker variability has before been concentrated on improving speech recognition. As far as we know, our work is the first to study speaker variability in a speech synthesis context.

This research is about predicting the timbre of vowels. Our hypothesis is that when speakers have a similar timbre for a vowel, they generally have a similar timbre for other vowels. Predicting the acoustical characteristics of a vowel is an important step for reproducing the timbre of an individual, because vowels carry a lot of information on the identity of a speaker. This research puts the emphasis on this subset of the timbre prediction problem to better understand its difficulties and possibilities.

This thesis compares a few machine learning techniques to complete the vowels' corpus of a speaker with the information of other vowels from that same speakers and other speakers' corpora. There is no simple and general model describing the phonetical variations between individuals. This is why a machine learning approach seemed reasonable to us. We chose non-parametric kernel methods, principal components analysis and independent components analysis as prediction techniques to achieve our goal. The low number of data available (5 vowels from 31 speakers) and

the great number of dimensions (200 per vowel) forced us to use these methods, because neural networks and graphical probabilistic models (just to name a few) need more data to avoid overtraining.

We used the harmonic and noise model (HNM) as a representation of voice for speech synthesis[2]. This model seems to be the best available to reproduce a speaker's voice with a quality high enough to keep his acoustical characteristics. The speech synthesis quality is an important factor because we want the resulting voice to be realistic, and not only intelligible. Our ultimate goal is to predict missing vowels for a speaker so they can be easily identified as his vowels.

The first part of this document introduces concepts of speech production and perception and some important mathematical models used in speech analysis and synthesis : linear prediction, sinusoidal models and the harmonic and noise model (HNM).

The second part exposes the fundamental problems of machine-learning and describes different methods used in this research : non-parametric kernels and genetic algorithms.

The third part describes the methodology used in our experiments to normalize and estimate (predict) the acoustical characteristics of missing vowels.

The last part compares the methods studied here and presentsour experimental results. We show that it is possible to reduce the harmonic distortion rate of a predicted vowel by 36.6% using four other known vowels. The harmonic distortion is obtained by comparing the predicted vowel with a mean vowel from all the available speakers in our database. This distortion is defined as the quadratic error of the harmonic spectrum. The best result was obtained with an asymmetric Gaussian kernel using a distance function optimized with a genetic algorithm.

Our results allow us to think it is possible to learn a general model of the timbre

[2]Note : the model representing the acoustical relations between vowels (kernels, PCA, ICA) should not be confused with the model representing the speech signal (HNM).

so that we can complete an individual corpus using its available acoustical information.

TABLE DES MATIÈRES

DÉDICACE iv

REMERCIEMENTS v

RÉSUMÉ vi

ABSTRACT ix

TABLE DES MATIÈRES xii

LISTE DES FIGURES xvii

LISTE DES TABLEAUX xx

LISTE DES NOTATIONS ET DES SYMBOLES xxi

LISTE DES ANNEXES xxiii

INTRODUCTION 1

CHAPITRE 1 CARACTÉRISATION DU SIGNAL VOCAL 4

1.1 Production du signal vocal 4

1.1.1 Conduit vocal 4

1.1.1.1 Sons voisés 5

1.1.1.2 Sons non-voisés 6

1.1.2 Analyse en fréquence des voyelles 8

1.1.2.1 Spectrogramme 8

1.1.2.2 Sonogramme 9

1.1.2.3 Formants et triangle vocalique 10

1.1.3 L'oreille humaine 12

1.2 Définition de quelques propriétés timbrales 14
1.2.1 La hauteur ou fréquence harmonique fondamentale 14
1.2.2 Autres propriétés timbrales 15
1.2.2.1 Puissance moyenne 15
1.2.2.2 Clarté (brightness) 16
1.2.2.3 Ratio d'harmonicité 16
1.2.2.4 Vacillement de la fréquence fondamentale (*F0 jitter*) et vibrato 16
1.2.2.5 Étalement spectral 17
1.2.2.6 Blancheur spectrale 17
1.3 Modélisation du signal vocal 18
1.3.1 Prédiction linéaire (*LPC* et *CELP*) 19
1.3.2 Modèles sinusoïdaux 21
1.3.2.1 Modélisation des sons voisés 23
1.3.2.2 Modélisation des sons non-voisés 23
1.3.2.3 Synthèse avec un modèle sinusoïdal 24
1.3.2.4 Notes sur le modèle sinusoïdal 24
1.3.3 Modèle harmonique et résiduel 25
1.3.3.1 Estimation de la fréquence fondamentale 26
1.3.3.2 Enveloppe de puissance et de phase des harmoniques 34
1.3.3.3 Enveloppe de puissance du signal résiduel 36
1.3.3.4 Algorithme de synthèse 37
1.3.3.5 Exemples de transformations possibles de la voix avec le modèle harmonique et résiduel 38
1.4 Adaptation de la reconnaissance pour différents locuteurs 39
1.5 Notre contribution aux techniques de traitement de la parole . . . 40

CHAPITRE 2 THÉORIE DES SYSTÈMES ADAPTATIFS UTILISÉS . . 43

2.1 Types de problèmes de l'apprentissage machine 43
2.1.1 Apprentissage supervisé et non-supervisé 44
2.1.2 Classification 44
2.1.3 Régression 45
2.1.4 Estimation de densité de probabilité 45
2.1.5 Notes sur les bases de données servant à l'entraînement . . . 46
2.2 Noyaux 46
2.2.1 Mesures de distance 46
2.2.1.1 Échelle de Mel 47
2.2.1.2 Courbes d'isosonie 48
2.2.1.3 Norme dB(A) 48
2.2.2 Types de noyaux 49
2.2.2.1 Les k plus proches voisins 50
2.2.2.2 Moyenne pondérée par un noyau Nadaraya-Watson 51
2.2.2.3 Malédiction de la dimensionnalité 52
2.3 Algorithmes génétiques 52
2.3.1 Théorie des algorithmes génétiques 53
2.4 Analyse en composantes principales 54
2.4.1 Algorithme pour trouver les l composantes principales . . . 55
2.4.2 Décomposition en vecteurs propres 56
2.5 Analyse en composantes indépendantes 57
2.5.1 Maximum d'entropie 58
2.5.2 Minimiser la divergence de Kullback-Leibler 59
2.6 Conclusion du chapitre 60

CHAPITRE 3 MÉTHODOLOGIE 61
3.1 Description des données utilisées 61
3.2 Normalisation des phonèmes des locuteurs 63

3.2.1 Extraction des paramètres 63
3.2.2 Élimination des enveloppes atypiques 64
3.2.3 Calcul de l'enveloppe typique moyenne 66
3.2.3.1 La fréquence fondamentale et le problème d'estimation des formants 67
3.2.3.2 Lissage de l'enveloppe typique avec un noyau . . . 67
3.3 Estimation des phonèmes manquants 69
3.3.1 Estimer une voyelle inconnue d'un locuteur 69
3.3.1.1 Estimation avec les noyaux 69
3.3.1.2 Estimation avec les composantes principales 71
3.3.1.3 Estimation avec les composantes indépendantes . . 72
3.3.2 Validation croisée 72
3.3.3 Mesure d'erreur 73
3.3.3.1 Notes sur la distorsion harmonique et le rapport signal sur bruit 73
3.4 Optimisation de la fonction de distance avec un algorithme génétique 74

CHAPITRE 4 RÉSULTATS 76
4.1 Noyaux 76
4.1.1 Effet de la taille du voisinage 77
4.1.2 Notes sur l'information mutuelle entre les voyelles 79
4.1.3 Distribution des résultats 83
4.2 Analyse en composantes principales 86
4.3 Analyse en composantes indépendantes 87
4.4 Importance de ces résultats pour la synthèse de voix 88

CONCLUSION 90

RÉFÉRENCES 92

ANNEXES 97

LISTE DES FIGURES

FIG. 1.1	Voies aériennes supérieures.	4
FIG. 1.2	Diagramme temporel du signal vocal du mot *chat* en français, contenant un son non-voisé et un son voisé.	5
FIG. 1.3	Sonogramme du signal vocal du mot *chat* en français.	6
FIG. 1.4	Détail de la partie non-voisée du signal vocal du mot *chat* en français.	7
FIG. 1.5	Détail de la partie voisée du signal vocal du mot *chat* en français.	8
FIG. 1.6	Spectrogramme et enveloppe harmonique d'une voyelle /a/.	9
FIG. 1.7	Sonogramme d'une voyelle /a/.	10
FIG. 1.8	Position approximative des centres des deux premiers formants d'un système de 5 voyelles et leur relation avec l'ouverture de la mâchoire et la position de la langue.	11
FIG. 1.9	Anatomie de l'oreille humaine.	12
FIG. 1.10	Réponse en fréquence des cellules ciliées selon leur position dans la cochlée.	13
FIG. 1.11	Graphique temporel de l'excitation périodique des cordes vocales.	20
FIG. 1.12	Schéma du système de modélisation harmonique et résiduelle du signal vocal.	26
FIG. 1.13	Spectrogramme d'une voyelle /a/.	28
FIG. 1.14	Spectrogramme d'une voyelle illustrant la cause du *frequency halfing*.	29
FIG. 1.15	Diagramme temporel d'une voyelle /a/.	30
FIG. 1.16	Une fenêtre d'analyse de la voyelle /a/.	31
FIG. 1.17	La fenêtre de pondération de Hanning utilisée.	32

Fig. 1.18	Une fenêtre d'analyse pondérée de la voyelle /a/.	32
Fig. 1.19	Coefficients temporels d'un filtre à délai à réponse finie (FIR).	33
Fig. 1.20	Réponse fréquentielle du filtre à délai de la figure 1.19. . . .	34
Fig. 1.21	Voyelle /o/ (signal original $s(t)$).	35
Fig. 1.22	Voyelle /o/ (signal harmonique $h(t)$).	36
Fig. 1.23	Voyelle /o/ (signal résiduel $r(t)$).	37
Fig. 2.1	Échelle perceptuelle de Mel.	48
Fig. 2.2	Courbes d'isosonie pour l'oreille humaine.	49
Fig. 2.3	Perception de la puissance du son selon la norme dB(A). . .	50
Fig. 2.4	Le problème de séparation aveugle de sources.	58
Fig. 3.1	Spectrogramme montrant trois fenêtres consécutives d'une voyelle /i/.	64
Fig. 3.2	Enveloppe spectrale normalisée des harmoniques d'une fenêtre par une approximation linéaire à chaque 20 Hz.	65
Fig. 3.3	Détail des basses fréquences ($f < 2$ kHz) de l'enveloppe typique d'une voyelle /a/ d'un locuteur construite à l'aide des points des fenêtres valides avec une approximation par noyau gaussien à chaque 20 Hz.	68
Fig. 3.4	Vecteur s_i contenant les enveloppes de puissance harmoniques des voyelles connues /a/, /e/, /o/ et /u/ d'un locuteur. . .	70
Fig. 3.5	Vecteur y contenant l'enveloppe de puissance harmonique d'une voyelle /i/ inconnue.	71
Fig. 3.6	Composition de la population d'une nouvelle génération de solutions.	75
Fig. 4.1	Enveloppes de puissance harmonique réelle et estimée pour une voyelle /o/.	78

FIG. 4.2 Distorsion de l'enveloppe harmonique moyenne pour chaque voyelle pour l'algorithme des k plus proches voisins (KNN) en fonction de k. 79
FIG. 4.3 Distorsion de l'enveloppe harmonique moyenne pour chaque voyelle pour différentes largeurs λ de noyau gaussien. 80
FIG. 4.4 Relation linéaire entre la somme des paramètres et la distorsion moyenne de la prédiction pour les cinq voyelles étudiées. 82
FIG. 4.5 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.04$ pour la voyelle /a/. 83
FIG. 4.6 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.10$ pour la voyelle /e/. 83
FIG. 4.7 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.09$ pour la voyelle /i/. 84
FIG. 4.8 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.07$ pour la voyelle /o/. 84
FIG. 4.9 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.09$ pour la voyelle /u/. 85
FIG. 4.10 Cinq premières composantes principales de notre base de données pour la voyelle /a/. 87
FIG. 4.11 Cinq composantes indépendantes de notre base de données pour la voyelle /a/. 88

LISTE DES TABLEAUX

TAB. 3.1 Statistiques sur le nombre de fenêtres d'analyse d'un locuteur pour chaque voyelle. 62

TAB. 4.1 Distorsion harmonique moyenne avec chaque méthode . . . 77

TAB. 4.2 Coefficients λ_v des noyaux gaussiens asymétriques pour chaque voyelle à prédire (vecteur Λ). 81

LISTE DES NOTATIONS ET DES SYMBOLES

LISTE DES ACRONYMES

ADN	Acide DésoxyriboNucléique
CELP	Code Excited Linear Prediction (Prédiction linéaire avec excitation codée)
dB	déciBels
FFT	Fast Fourier Transform (Transformée rapide de Fourier)
FIR	Finite Impulse Response (filter) (Filtre à réponse finie)
GMM	Gaussian Mixture Model (Mélange de gaussiennes)
GSM	Global System for Mobile Communications (Système Global pour les Communications Mobiles)
HNM	Harmonic and Noise Model (Modèle harmonique et résiduel)
ICA	Independant Component Analysis (Analyse en Composantes Indépendantes)
IP	Internet Protocol (Protocole Internet)
KL	Kullback-Liebler (Divergence de)
KNN	K-Nearest Neighbors (K Plus Proches Voisins)
LPC	Linear Predictive Coding (Encodage par Prédiction Linéaire)
PCA	Principal Component Analysis (Analyse en Composantes Principales)
SEEVOC	Spectral Envelope Estimation Vocoder (Vocoder à Estimation d'Enveloppe Spectrale)
SNR	Signal to Noise Ratio (rapport signal sur bruit)

LISTE DES SYMBOLES

h_k	Amplitude du k^e sinusoïde d'un système harmonique.
p_i	Composante principale i de l'ensemble des spectres de puissance harmonique des voyelles des locuteurs dont le corpus est complet.
s_i	Vecteur formé des spectres de puissance harmonique des voyelles connues du locuteur i.
y_i	Vecteur formé du spectre de puissance harmonique d'une voyelle cible connue du locuteur i.
$\hat{y}_i$	Vecteur formé du spectre de puissance harmonique d'une voyelle cible prédite du locuteur i.
μ	Moyenne des vecteurs s_i.
K_λ	Noyau gaussien de paramètre λ.
K_Λ	Noyau gaussien de paramètres λ_v pour chacune des voyelles v.

LISTE DES ANNEXES

ANNEXE I CÉDÉROM CONTENANT DES EXEMPLES SONORES 97

INTRODUCTION

Le désir de construire une machine permettant de reproduire la voix humaine n'est pas nouveau. La légende veut qu'un peu avant l'an 1000, Gerbert d'Aurillac [3] aurait inventé une tête de bronze pouvant répondre par "oui" ou "non" à des questions. Des têtes similaires auraient aussi été inventées au 13e siècle par Albertus Magnus et Roger Bacon. Des essais mieux documentés ont été tentés à la fin du 18e siècle (entre autres par Wolfgang von Kempelen en Autriche), mais ce n'est qu'au milieu du 20e siècle, avec l'apparition des ordinateurs électroniques, que ces tentatives sont devenues capables de reproduire un discours intelligible.

Il existe maintenant plusieurs technologies pour synthétiser des voix réalistes[4], mais la construction d'une voix complète nécessite des mois de travail à une équipe de linguistes, informaticiens et ingénieurs. Le locuteur qu'on veut reproduire doit avoir été enregistré pendant plusieurs dizaines d'heures. Dans le cas d'un système de synthèse par concaténation de diphones en particulier, les enregistrements doivent contenir toutes les possibilités de diphones[5], soit entre 1000 et 2000 différents pour la plupart des langues. Idéalement, il devrait y avoir quelques exemples de chaque diphone, permettant de sélectionner le meilleur selon le contexte et de minimiser les extrapolations lors de la synthèse. Des tests sur la qualité de la synthèse permettent ensuite aux ingénieurs de détecter les erreurs inévitables de la saisie de données. Dans le meilleur des cas, le locuteur est disponible pour enregistrer sa voix s'il manque des exemples de diphones.

[3] Qui fut aussi pape catholique de 999 à 1003 sous le nom de Sylvestre II.

[4] On entend par voix réaliste une voix non seulement intelligible, mais de qualité assez bonne pour faire croire à celle d'un individu particulier.

[5] Un diphone est la transition entre la fin d'un phonème et le début du phonème suivant d'un mot.

Les systèmes de synthèse de voix de haute qualité présentement disponibles n'offrent pour cette raison que quelques voix différentes. Par exemple, Le système d'AT&T[6] offre quatre voix anglaises, une espagnole, deux allemandes et une française, le système Vocaloid[7] de Yamaha pour le chant n'offre que trois voix anglophones et une japonaise, et le système gratuit MBROLA[8], développé entre autres par la Faculté Polytechnique de Mons (Belgique), offre une soixantaine de locuteurs pour un total d'une trentaine de langues. Le système MBROLA est le seul qui permet à l'usager d'ajouter une voix. La qualité de la synthèse est cependant généralement moins bonne que les systèmes de Yamaha et de AT&T, surtout à cause des données fournies, parfois incomplètes ou de mauvaise qualité. Cette constatation nous fait souhaiter qu'un programme puisse un jour avoir une assez bonne connaissance des propriétés acoustiques de la voix humaine pour pouvoir apprendre automatiquement une nouvelle voix à partir d'exemples.

La modélisation automatique du timbre d'un individu pourrait avoir de nombreuses applications dans des domaines variés. Elle permettrait la synthèse d'une voix à partir d'un corpus incomplet, la conservation du timbre pour la traduction, la reconstitution d'enregistrements endommagés et l'apprentissage progressif du timbre d'un individu par une machine.

L'objectif de cette recherche est d'explorer différentes techniques d'apprentissage machine pour compléter le corpus de voyelles d'un locuteur à partir d'un sous-ensemble d'exemples de ce locuteur. La prédiction de voyelles et des sons voisés en général est une étape importante dans la reproduction de l'identité timbrale d'un individu.

Ce document débute avec une description acoustique de la voix et de ses propriétés.

[6]http ://www.naturalvoices.att.com/demos/
[7]http ://www.vocaloid.com
[8]http ://festvox.org/mbrola/

Nous considérons ensuite le modèle harmonique et résiduel (*harmonic and noise model*) comme système de base pour l'analyse-synthèse de voix. Le premier chapitre se termine avec une revue de la littérature sur les variances acoustiques entre les locuteurs.

Le deuxième chapitre couvre les différentes méthodes d'apprentissage machine explorées dans le cadre de cette recherche, soit les noyaux, l'analyse par composantes principales et l'analyse par composantes indépendantes. Ces méthodes permettent de prédire les caractéristiques acoustiques des voyelles pour lesquelles nous n'avons aucun exemple provenant d'un locuteur.

Le troisième chapitre décrit la méthodologie utilisée pour nos expériences pour la normalisation des voyelles des locuteurs et l'estimation des caractéristiques acoustiques des voyelles manquantes.

Le quatrième chapitre compare les différentes méthodes utilisées et montre nos résultats expérimentaux.

CHAPITRE 1

CARACTÉRISATION DU SIGNAL VOCAL

1.1 Production du signal vocal

1.1.1 Conduit vocal

La voix humaine est produite par le passage de l'air sortant des poumons dans le conduit vocal. Toutes les parties du conduit vocal participent à la sonorité finale, mais le timbre produit est particulièrement influencé par les cordes vocales, la langue, les lèvres et les dents.

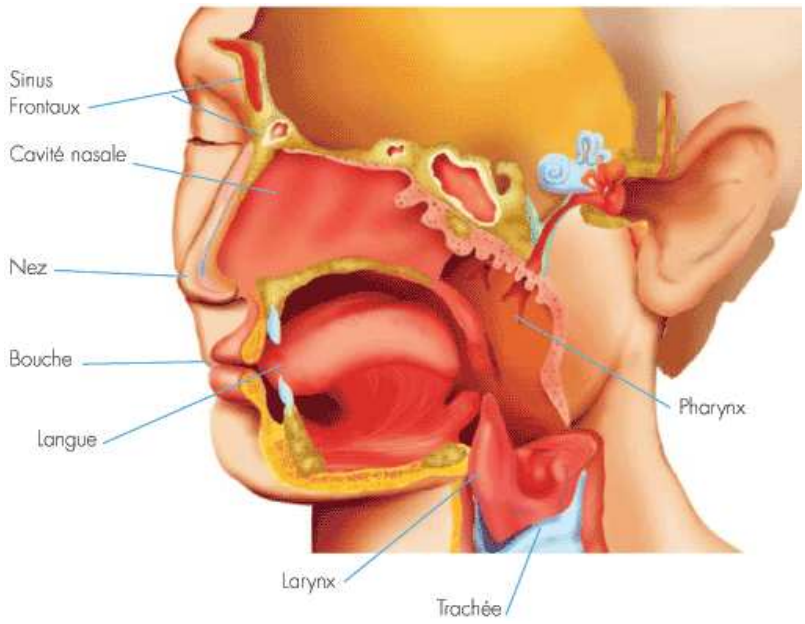


FIG. 1.1 Voies aériennes supérieures.
Source : http ://www.astrazeneca.fr

La position et la forme de chacune des parties du conduit vocal permet à l'individu

de faire varier la réponse fréquentielle de sa voix. De cette manière, il peut produire une grande variété de sons. Les sons produits par le conduit vocal sont appelés *phonèmes*.

Du point de vue du traitement de signal, il est utile de séparer les phonèmes en deux catégories : les *sons voisés* et les *sons non-voisés*.

1.1.1.1 Sons voisés

Les sons voisés sont produits lorsque les cordes vocales sont suffisamment contractées pour vibrer avec le passage de l'air. Ces sons ont tous une composante périodique harmonique. Cette catégorie comprend, en plus des voyelles, d'autres groupes de phonèmes comme par exemple les consonnes nasales /m/ et /n/, les fricatives /v/ et /Z/ et les plosives /b/ et /d/.

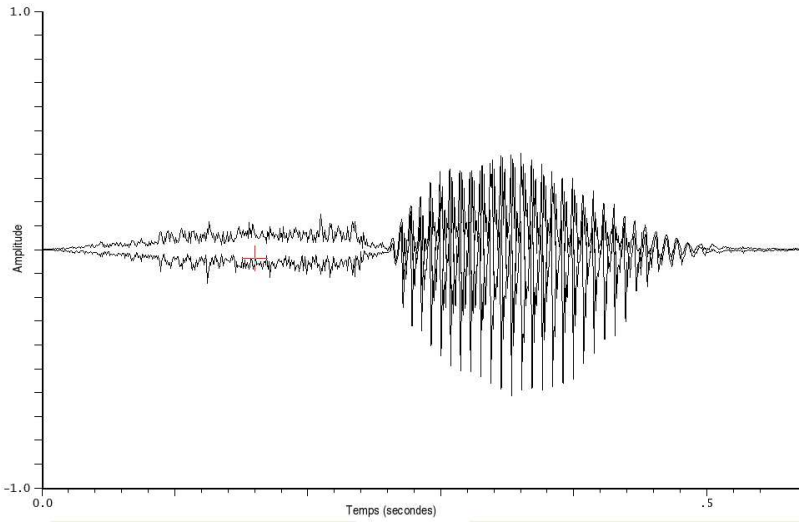


FIG. 1.2 Diagramme temporel du signal vocal du mot *chat* en français, contenant un son non-voisé et un son voisé.

1.1.1.2 Sons non-voisés

Les sons non-voisés comprennent tous les sons qui ne sont pas voisés, par exemple les plosives /p/ et /t/ et les fricatives /s/ et /S/. Ces sons sont des signaux transitoires (comme la plosive /p/ dans le mot "patate") ou des signaux qui n'ont pas de composante périodique (comme la fricative /S/ dans le mot "chat"). Ils sont produits dans le premier cas par les lèvres et dans le deuxième cas par les turbulences de l'air passant par le conduit vocal. Dans ce dernier cas, la position des parties du conduit vocal modifie le spectre du signal non périodique. Une position particulière de la langue permet d'avoir le son /s/ (comme dans le mot "signal") ou le son /S/.

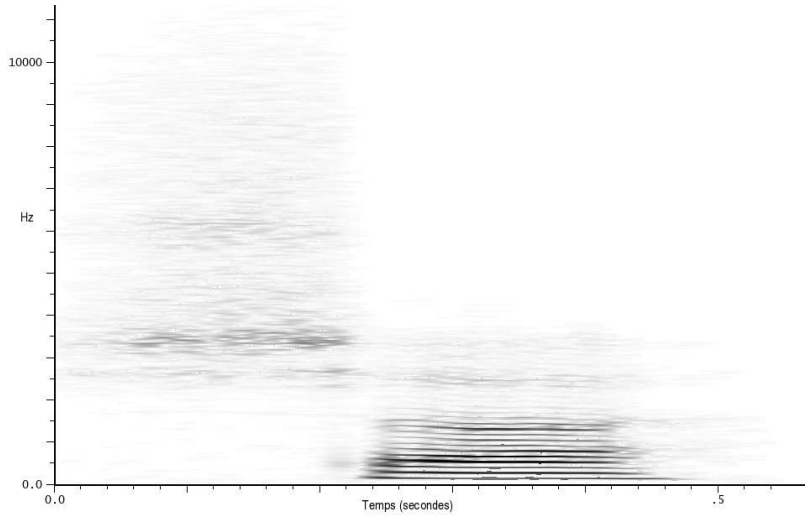


FIG. 1.3 Sonogramme du signal vocal du mot *chat* en français.

Les figures 1.2 à 1.4 illustrent bien l'antagonisme entre les deux classes de sons (ou phonèmes). Le mot *chat* en français est composé d'un phonème non-voisé /S/ et d'une voyelle /a/. Sur le graphique du signal temporel de la figure 1.2, on distingue clairement les deux phonèmes, le premier (entre 0 et 0.25 seconde) étant non-périodique et le deuxième (entre 0.25 et 0.5 seconde) étant clairement périodique. Sur le sonogramme (figure 1.3), nous pouvons voir que le phonème /S/ est caractérisé par une énergie répartie entre 2 et 11 kHz, avec des sommets entre 2500 à

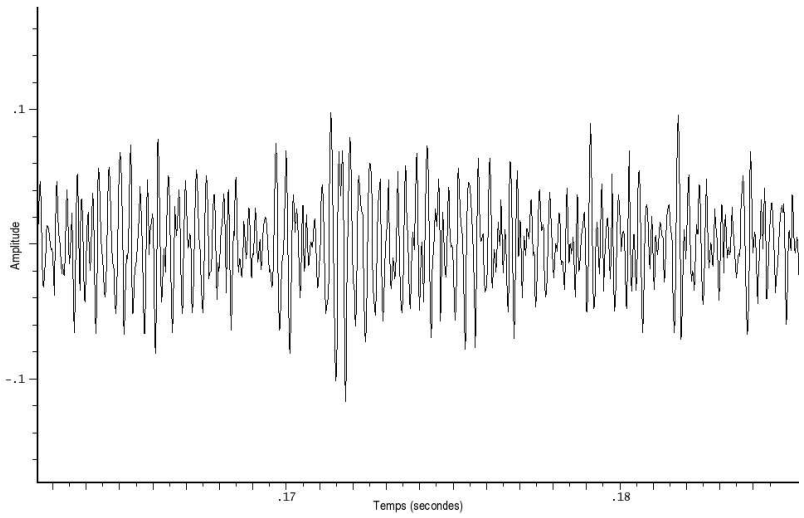


FIG. 1.4 Détail de la partie non-voisée du signal vocal du mot *chat* en français.

4000Hz et autour de 6000 Hz. Sur ce graphique, une couleur plus foncée indique une énergie plus élevée pour une bande de fréquence à un moment donné. Le phonème /a/ est quant à lui caractérisé par une énergie concentrée dans les composantes harmoniques périodiques sous 4 kHz. Les lignes noires également espacées sur le graphique soulignent cette caractéristique harmonique du signal.

Les figures 1.4 et 1.5 montrent respectivement le détail des signaux temporels non-voisés et voisés. L'apparence de périodicité sur la figure 1.4 est due aux sommets d'énergie du son non-voisé autour de 3500 Hz. Sur la figure 1.5, nous pouvons observer la périodicité du signal construit par une somme de sinusoïdes harmoniques. La période du signal de la figure 1.5 correspond à la période fondamentale du système harmonique, soit dans ce cas 7.63 ms ou 131 Hz. Les lignes noires sous 4 kHz de la figure 1.3 situées entre 0.25 et 0.5 seconde sont donc espacées d'environ 131 Hz.

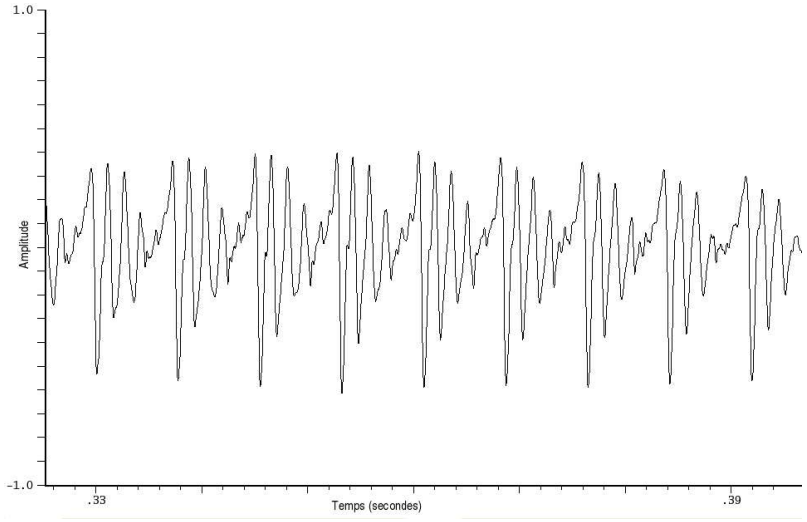


FIG. 1.5 Détail de la partie voisée du signal vocal du mot *chat* en français.

1.1.2 Analyse en fréquence des voyelles

Notre recherche se concentre sur l'analyse d'un sous-ensemble de phonèmes importants pour la reconnaissance du locuteur, les voyelles /a/, /e/, /i/, /o/ et /u/. Nous introduisons donc ici quelques techniques pour analyser et visualiser ces phonèmes.

1.1.2.1 Spectrogramme

La figure 1.6 représente un exemple de spectrogramme montrant l'enveloppe harmonique d'une fenêtre d'analyse d'une voyelle /a/ du locuteur 11 de notre ensemble de données. Une des principales caractéristiques des sons voisés est de contenir la puissance des harmoniques dans une enveloppe de puissance caractéristique. Cette enveloppe de puissance est reliée au filtre créé par la forme du conduit vocal lors de la vibration des cordes vocales qui produisent les harmoniques. Les harmoniques d'une voyelle resteront à l'intérieur de cette enveloppe peu importe la fréquence

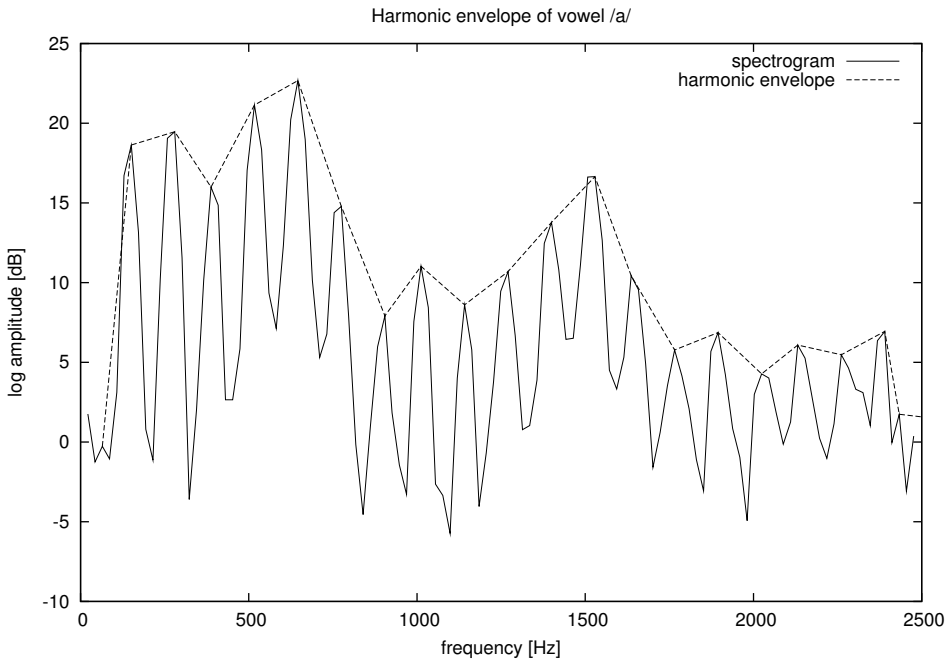


FIG. 1.6 Spectrogramme et enveloppe harmonique d'une voyelle /a/.

fondamentale de la voix[1].

Sur le spectrogramme de la figure 1.6, les pics d'énergie également espacés sont appelés harmoniques. La ligne pointillée représente approximativement l'enveloppe de puissance harmonique dans laquelle les harmoniques doivent se situer pour que la voyelle corresponde au son /a/.

1.1.2.2 Sonogramme

La figure 1.7 montre le sonogramme d'une voyelle /a/. Un sonogramme permet de voir la puissance d'un spectre de puissance varier dans le temps. Sur la figure, un

[1]L'enveloppe dont nous parlons est une représentation idéale que nous jugeons suffisante pour un système de synthèse vocale réaliste. Elle néglige toutes les légères variations qui se retrouvent dans une voix normale, mais celles-ci peuvent être ajoutées par la suite dans un système de synthèse de voix, pour correspondre au contexte et à l'émotion désirée dans la voix.

point plus foncé correspond à une fréquence dont la puissance est plus élevée à un instant donné. Sur ce sonogramme nous pouvons voir que même lorsque les harmoniques changent de fréquence, leur puissance respecte l'enveloppe de puissance harmonique correspondant au /a/.

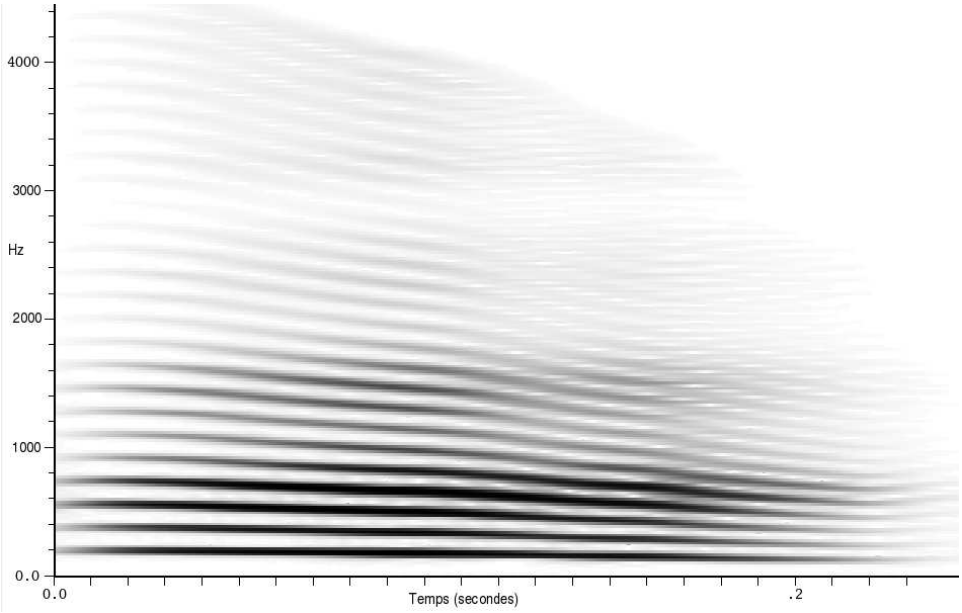


FIG. 1.7 Sonogramme d'une voyelle /a/.

1.1.2.3 Formants et triangle vocalique

Les sommets de l'enveloppe de puissance harmonique sur la figure 1.6 sont appelés *formants*. On associe normalement la position de ces formants avec la perception psychoacoustique des différentes voyelles. Il est d'ailleurs possible classer les voyelles avec moins de 20% d'erreur (pour un ensemble de 5 voyelles) simplement à partir de la fréquence centrale des trois premiers formants(Pal et Dutta Majumder, 1977).

Le premier formant, celui dont la fréquence est la plus basse, est associé à l'ouverture de la mâchoire. Plus la mâchoire est ouverte, plus le formant est situé à une fréquence élevée. Le premier formant est normalement situé entre 300 et 700Hz.

Le deuxième formant, dont la fréquence centrale est plus élevée que le premier formant, est associé à la position de la langue. Plus la langue est située à l'avant de la bouche, plus le deuxième formant est situé à une fréquence élevée. Le deuxième formant est normalement situé entre 800 et 1900Hz.

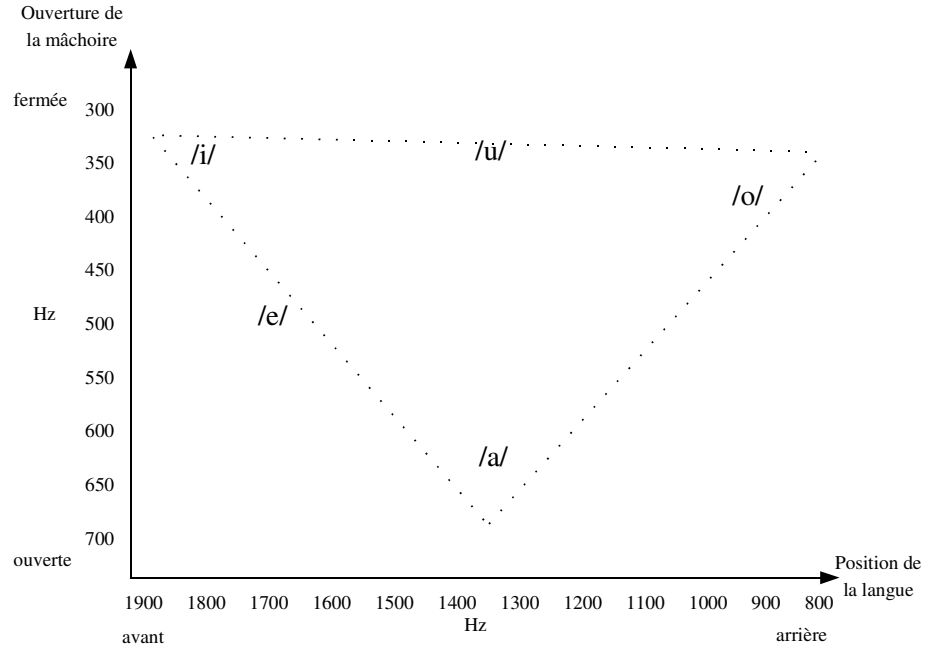


FIG. 1.8 Position approximative des centres des deux premiers formants d'un système de 5 voyelles et leur relation avec l'ouverture de la mâchoire et la position de la langue.

La figure 1.8 montre le triangle vocalique[2] utilisé pour classer les voyelles à partir de leurs deux premiers formants. L'axe vertical représente la fréquence centrale du premier formant, relié à l'ouverture de la mâchoire. L'axe horizontal représente la fréquence centrale du deuxième formant, relié à position de la langue. Les valeurs décroissantes des fréquences sur les axes est une convention en phonétique pour cette représentation graphique du triangle vocalique.

[2]Le triangle vocalique est parfois présenté sous la forme d'un trapèze, dépendant de la langue étudiée et de l'auteur.

1.1.3 L'oreille humaine

L'oreille est l'organe perceptuel du signal sonore. Nous décrivons habituellement cet organe en trois parties : l'oreille externe, moyenne et interne.

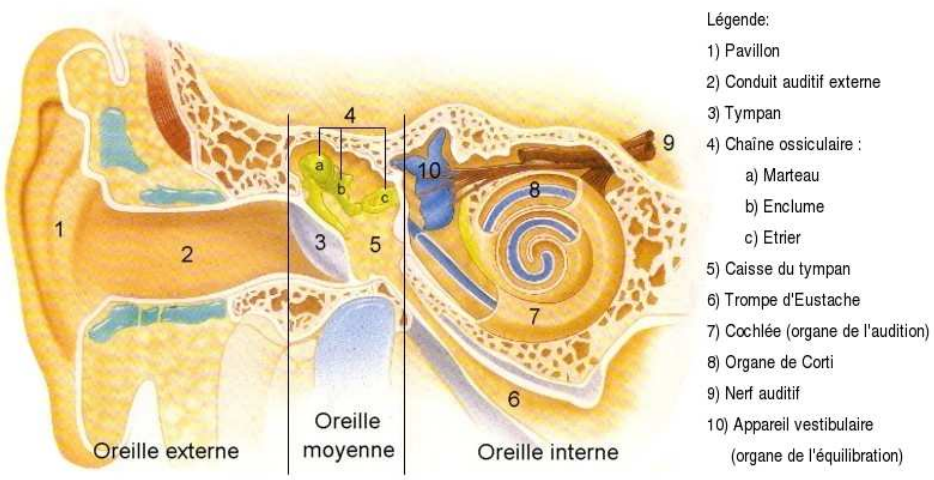


FIG. 1.9 Anatomie de l'oreille humaine.
Source : http ://www.audition-grenier.com

L'oreille externe est constituée du *pavillon* et du conduit auditif externe. Contrairement à d'autres mammifères, l'humain n'a pas la capacité d'orienter son pavillon pour mieux percevoir les sons. La géométrie du pavillon et du conduit auditif externe amplifie les sons entre 1 et 5 kHz.

L'oreille moyenne est composée du *tympan* et des osselets, trois petits os. Les osselets s'appellent le *marteau*, *l'enclume*, et *l'étrier*. Les *trompes d'Eustache*, reliées au conduit respiratoire, permettent d'équilibrer la pression entre les deux côtés du tympan lors d'un changement d'altitude ou lorsque l'individu plonge sous l'eau. Si cet équilibre n'est pas maintenu, la fragile membrane du tympan peut se rompre. Les osselets permettent de transmettre les vibrations sonores perçues par le tympan à l'oreille interne.

L'oreille interne contient la *cochlée* (ou *limaçon*), mais aussi le *vestibule*, qui est l'organe de l'équilibration. La cochlée est l'organe de l'ouïe. Elle a la forme d'une spirale creuse dont l'aire des deux cavités se réduit progressivement. Les cavités, séparées par la *membrane basilaire*, sont remplies d'un liquide appelé *endolymphe* et sont tapissées par les *cellules ciliées*. Ces cellules et les membranes de la cochlée sont appelées l'*organe de Corti*.

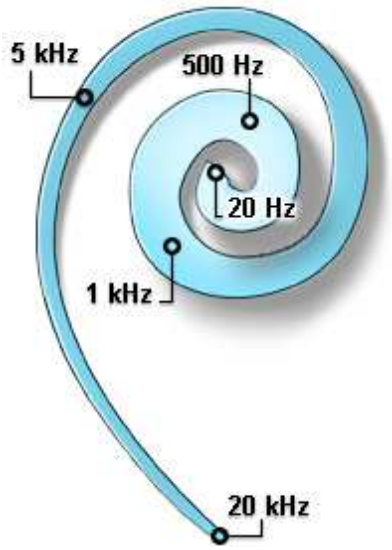


FIG. 1.10 Réponse en fréquence des cellules ciliées selon leur position dans la cochlée.
Source : http ://www.mpsycho.uqam.ca/percept/Bloc3/bloc3-audition11.html

Lorsque les cellules ciliées sont agitées par les vibrations sonores reçues par la cochlée, elle transmettent un signal au nerf auditif. La traduction des ondes sonores mécaniques en signal nerveux électrique s'appelle la *transduction mécanoélectrique*. Les cellules les plus près de l'entrée de la cochlée sont plus sensibles aux hautes fréquences et celles les plus éloignées répondent aux basses fréquences, ce qui permet de distinguer la hauteur d'un son (voir figure 1.10). La réponse fréquentielle des cellules ciliées suit une loi logarithmique : la résolution en fréquence est plus fine en basse fréquence.

L'anatomie de l'oreille est déterminante pour la perception psychoacoustique humaine. Elle permet de comprendre quelles caractéristiques sonores transportent le plus d'information pour la reconnaissance des individus. Les détails mathématiques de la perception de la puissance et de la fréquence des sons seront couverts dans la section 2.2.1.

1.2 Définition de quelques propriétés timbrales

1.2.1 La hauteur ou fréquence harmonique fondamentale

Lorsque les cordes vocales vibrent, elles produisent un signal pouvant être approximé comme une somme de sinusoïdes appelés *harmoniques* dont la fréquence est un multiple d'une *fréquence harmonique fondamentale*. Cette fréquence fondamentale du signal est aussi appelée *hauteur*. Le sonogramme de la figure 1.7 et le spectrogramme de la figure 1.6 montrent bien que l'énergie d'un son voisé est concentrée dans une série de sinusoïdes également espacés en fréquence. Les harmoniques peuvent être décrits de cette manière :

$$h(t) = \sum_k a_k(t) sin(2\pi kt \frac{f}{f_e} + \phi_0)$$

où f_e est la fréquence d'échantillonnage du signal numérique, $a_k(t)$ est l'amplitude de l'harmonique k au temps t et f est la fréquence fondamentale du signal harmonique. ϕ_0 est la phase de l'harmonique au temps $t = 0$.

La hauteur est sans doute une des notions les plus fondamentales pour l'analyse du signal vocal. Déterminer la hauteur d'une voix peut sembler simple à première vue, mais lorsque le mathématicien ou l'ingénieur sont confrontés à des signaux vocaux réels, plusieurs problèmes apparaissent. La présence de bruits ou d'autres signaux

périodiques et les voix pathologiques rendent parfois très difficile l'estimation du signal vocal harmonique. Une règle simple et infaillible pour la détermination de la hauteur semble impossible : le problème est un sujet de recherche en soi. Heureusement, certaines méthodes fonctionnent dans la plupart des cas qui nous intéressent. Quelques-unes de ces méthodes seront décrites dans la section 1.3.3.1.

1.2.2 Autres propriétés timbrales

Nous définissons dans cette section quelques propriétés timbrales de la voix de manière mathématique. Ces définitions nous permettront de mieux comprendre le travail de traitement de la voix présenté dans les prochains chapitres.

1.2.2.1 Puissance moyenne

La puissance se définit comme la quantité d'énergie contenue dans le signal vocal dans une période de temps donnée. Dans le cas d'un signal numérique, nous sommes intéressés par la puissance moyenne du signal $s(t)$ dans une fenêtre de longueur T :

$$P_{moyenne} = \frac{1}{T} \sum_{t=1}^{T} s(t)^2$$

La mesure de puissance est utile pour positionner les fenêtres d'analyse pour les algorithmes de synthèse du modèle harmonique et bruit dans la section 1.3.3. La puissance est aussi une caractéristique prise en compte lorsque nous voulons définir si une voix est présente dans le signal analysé.

1.2.2.2 Clarté (brightness)

La clarté d'un signal audio est définie comme la position du centroïde spectral. Le centroïde spectral est la fréquence médiane de l'énergie E d'un signal (Jehan, 2001).

$$E(0, f_{centroide}) = E(f_{centroide}, \infty)$$

Nous pouvons modifier la clarté d'une voix simplement en appliquant un filtre sur le signal vocal qui réduit ou augmente la puissance des basses ou hautes fréquences.

1.2.2.3 Ratio d'harmonicité

Le ratio d'harmonicité est le rapport entre l'énergie pouvant être expliquée par un signal périodique harmonique E_h et l'énergie du signal total E_s. Ce ratio peut être mesuré pour le signal complet ou pour une bande de fréquence en particulier (Quatieri, 2001).

$$R_{harm} = \frac{E_h}{E_s}$$

Ce ratio peut être modifié lors de la synthèse avec le modèle harmonique et bruit pour obtenir les effets décrits dans la section 1.3.3.5.

1.2.2.4 Vacillement de la fréquence fondamentale (*F0 jitter*) et vibrato

Le vacillement de la fréquence fondamentale est sa variation moyenne sur une période de temps donnée. On peut le représenter simplement comme un écart-type quand il n'est pas périodique. Dans le cas d'une voix chantée, le léger vacillement périodique de la fréquence fondamentale est appelé vibrato. La période du vibrato varie généralement entre 100 et 400 ms.

Il est possible de reproduire cet effet lors de la synthèse. Le vacillement périodique d'une voix chantée peut être recréé avec une simple fonction sinusoïdale ou un vacillement provenant de l'analyse d'un signal vocal réel.

1.2.2.5 Étalement spectral

L'étalement spectral est une caractéristique timbrale dont la mesure est prévue pour l'analyse de signaux audio dans la future norme MPEG-7 sur la description de contenu multimédia (Lee, 2002)[3]. L'étalement spectral est défini comme la portion du spectre effectivement utilisée, par exemple la portion se situant au-dessus d'un certain seuil de puissance. Sa définition mathématique peut varier, voici une forme qu'elle peut prendre :

$$ES = \frac{N_{bandes\ audessus\ du\ seuil}}{N_{total\ de\ bandes}}$$

Les bandes sont définies comme des sections du spectre du signal. Elles sont arbitrairement définies selon l'application, mais les points du spectre donné par la transformée de Fourier rapide (FFT) sont souvent utilisés.

1.2.2.6 Blancheur spectrale

La blancheur spectrale est une autre caractéristique timbrale dont la mesure est prévue pour l'analyse de signaux audio dans la norme MPEG-7 (Lee, 2002). Un spectre blanc a une densité de puissance constante dans toutes les bandes[4]. La blancheur spectrale est une fonction qui définit la distance entre un spectre blanc

[3]La norme MPEG-7 contient aussi les autres caractéristiques timbrales décrites dans ce document, mais l'étalement spectral n'est pas utilisé par les autres travaux étudiés dans le cadre de cette recherche, d'où la référence particulière à la norme MPEG-7 faite ici.

[4]Un bruit blanc est un bruit dont l'espérance de la densité puissance est une constante pour toutes les bandes.

et le spectre du signal mesuré. On peut mesurer la blancheur spectrale ainsi :

$$BS = 1 - \frac{VAR(E_i)}{E_s}$$

Comme la variance des bandes i de puissance E_i du signal est toujours inférieure ou égale à l'énergie totale E_s du signal, cette fonction est comprise entre 0 et 1, où la valeur 0 indique un signal sinusoïde pur (de l'énergie pour une seule fréquence), et la valeur 1 indique un spectre blanc. La blancheur spectrale est parfois utilisée comme critère pour définir si un signal est harmonique ou non.

1.3 Modélisation du signal vocal

Les algorithmes de synthèse de la parole construisent en général un modèle de la voix à partir de vrais segments d'une voix. Ils utilisent un modèle intermédiaire qui permet parfois de resynthétiser la voix avec un rythme et une hauteur (fréquence) arbitraires. Une fois tous les segments modélisés, il est possible de les synthétiser et de les concaténer pour créer l'expression désirée.

Le domaine des télécommunications appelle ces modèles *codecs* (coder-decoders) lorsqu'ils sont utilisés pour la compression de voix. Ils ne sont pas alors construits pour une seule voix comme dans le cas de la synthèse, et sont souvent organisés en dictionnaires (*codebooks*), ce qui permet d'envoyer des codes minimaux durant la transmission pour une reconstitution intelligible.

Dans cette section, nous décrivons d'abord deux modèles fréquemment utilisés depuis la fin des années 1970 pour la synthèse de la voix, soit la prédiction linéaire et les modèles sinusoïdaux (ou banques d'oscillateurs). Nous décrivons ensuite le modèle harmonique et résiduel utilisé dans cette recherche. Ce dernier modèle per-

met une reproduction beaucoup plus fidèle du timbre d'un individu.

1.3.1 Prédiction linéaire (*LPC* et *CELP*)

Le codage par prédiction linéaire (Linear Predictive Coding ou simplement *LPC*) est une représentation qui permet de reconstruire le signal vocal en supposant un modèle source/filtre du conduit vocal. La source dans ce cas est le mouvement des cordes vocales, et le filtre est un filtre FIR représentant le conduit vocal (Rabiner et Schafer, 1978). Les p paramètres a_i du filtre sont recalculés à chaque fois que le signal vocal transgresse un critère de stationnarité. La voix est resynthétisée en filtrant une excitation d'une période et puissance variable avec les filtres correspondant aux phonèmes cibles. L'algorithme de resynthèse s'appelle CELP (*Code Excited Linear Prediction*) (Huang, Acero et Hon, 2001).

La figure 1.11 montre un exemple d'excitation des cordes vocales utilisée comme source d'un algorithme CELP. Ce mouvement peut être enregistré par un laryngographe[5] (Boite et al, 2000; Quatieri, 2001).

La prédiction linéaire estime le prochain échantillon $\hat{x}(t)$ à partir des p derniers échantillons :

$$\hat{x}(t) = -\sum_{i=1}^{p} a_i x(t-i)$$

L'erreur sur la prédiction est calculée ainsi :

$$e(t) = x(t) - \hat{x}(t)$$

Les coefficients a_i sont trouvés en minimisant l'erreur quadratique $E[e^2(t)]$. La

[5]Un laryngographe (ou électrolaryngographe) est un appareil permettant de mesurer l'impédance du larynx durant la phonation.

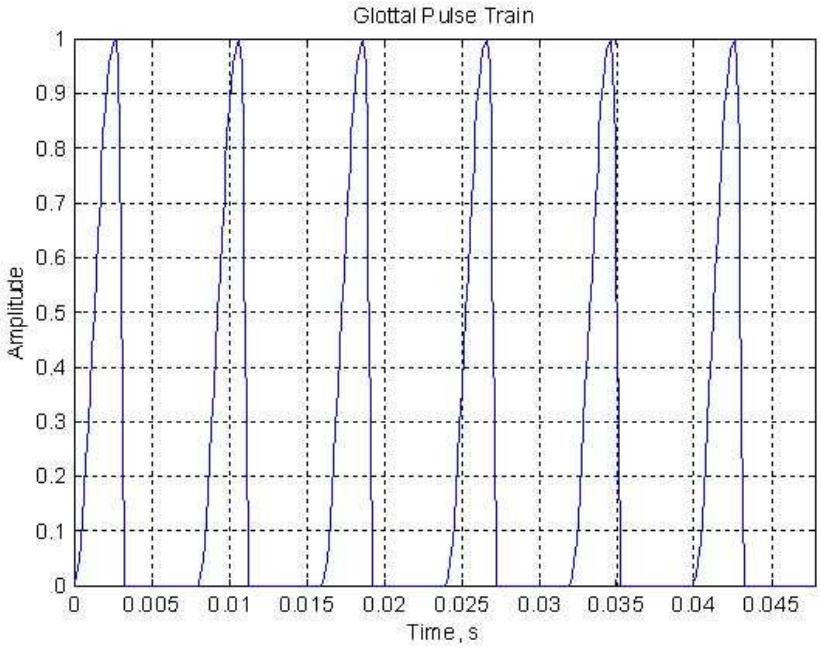


FIG. 1.11 Graphique temporel de l'excitation périodique des cordes vocales. Source : http ://umsis.miami.edu/˜balberta/een540project2/malee.htm

solution est trouvée en solutionnant l'équation de Yule-Walker (Kleijn et Paliwal, 1995) :

$$\mathbf{a^T R} = -[\mathbf{R}(1)\mathbf{R}(2)\mathbf{R}(3)\ldots\mathbf{R}(p)]$$

où $\mathbf{R}$ est la matrice d'autocorrélation composée des éléments $\mathbf{R}(i-j)$:

$$\mathbf{R}(i-j) = E[x(i)x(j)]$$

et $E[\cdot]$ représente l'espérance, ici celle du produit $x(i)x(j)$, normalement calculée sur une durée d'au moins 15 ms. La matrice d'autocorrélation $\mathbf{R}$ (symétrique, Toeplitz), est composée d'éléments $\mathbf{R}_{i,j} = \mathbf{R}(i-j)$, et le vecteur $\mathbf{a}$ contient les coefficients du filtre FIR. Les algorithmes de Levinson (Trench, 1964) ou de Schur (Kleijn et Paliwal, 1995) sont normalement utilisés pour calculer récursivement cette solution de manière efficace (complexité $\Theta(n^2)$).

Il est courant d'utiliser 10 à 13 coefficients pour le filtre FIR. Cela permet d'avoir une approximation raisonnable des formants et donne une voix suffisamment intelligible pour être utilisée dans des applications commerciales.

En plus d'être utilisé couramment dans les systèmes de synthèse de voix (comme le système libre et gratuit *festival*[6], développé à l'université d'Edinburgh), l'encodage LPC est utilisé dans plusieurs systèmes de compression de voix comme le standard *GSM* pour la téléphonie cellulaire.

1.3.2 Modèles sinusoïdaux

Les modèles sinusoïdaux (Kleijn et Paliwal, 1995)(Quatieri, 2001) permettent de synthétiser ou de compresser des voix avec une performance généralement supérieure aux algorithmes basés sur la prédiction linéaire. Cependant, la quantité de calcul nécessaire pour analyser et resynthétiser la voix a retardé leur application dans les systèmes de téléphonie[7]. Le contexte qui nous intéresse, la synthèse de voix réaliste, est moins sensible au temps de calcul et au délai. Les modèles sinusoïdaux sont donc intéressants à étudier dans le cadre de notre recherche.

Les modèles sinusoïdaux se basent sur le principe de la transformée de Fourier qui dit que tout signal peut se représenter comme étant une somme de sinus. La modélisation de la voix peut donc être faite en trouvant un ensemble de fonctions sinusoïdales représentant l'essentiel de l'énergie du signal vocal. Il suffit ensuite de reproduire ces sinusoïdes principales pour resynthétiser la voix. Il existe une

[6]Site web de festival : http ://www.cstr.ed.ac.uk/projects/festival/

[7]L'encodage des modèles sinusoïdaux entraîne un délai d'entre 10 et 25 ms dans les communications. Ce délai gêne la conversation, particulièrement dans le cas où la ligne téléphonique est sujette à un écho important (un écho dont le délai est de plus de 20 ms déplaît aux usagers des systèmes téléphoniques). Par ailleurs, l'encodage de la voix n'est pas la seule étape responsable des délais de transmission, c'est pourquoi les codeurs de style LPC ou ADPCM sont généralement préférés.

variété d'algorithmes basés sur ce principe, nous nous contenterons ici d'en exposer un. La principale caractéristique du modèle sinusoïdal que nous présentons est de contraindre les sinus à avoir une relation harmonique, c'est à dire que leur fréquence est multiple d'une fréquence fondamentale. Même si cela ne semble pas toujours être le cas des sons voisés (Kleijn et Paliwal, 1995), l'approximation apparaît raisonnable à nos yeux et à nos oreilles. Le modèle prend donc la forme mathématique suivante :

$$s(t) = \sum_k a_k(t) sin(\phi_k(t))$$

Où $a_k(t)$ et $\phi_k(t)$ sont respectivement l'amplitude et la phase de l'harmonique k au temps t et $s(t)$ est la valeur du signal harmonique au temps t.

D'autres modèles n'assument pas que les sinusoïdes ont une relation harmonique. Certains algorithmes choisissent les k sinusoïdes les plus puissantes de chaque fenêtre d'analyse de la FFT (Quatieri, 2001). D'autres suivent l'apparition et la disparition de sinusoïdes en définissant certains critères d'existence (puissance soutenue, continuité de la fréquence du sinusoïde) (Serra, 1997). Il est aussi possible d'exprimer les puissances des sinusoïdes de fréquence f_k avec un filtre d'ordre 10 à 22 (Kleijn et Paliwal, 1995). Nous présentons ici le modèle sinusoïdal harmonique pour sa simplicité, nous n'aborderons pas le sujet de la compression des coefficients pour la transmission sur un lien de communication à bas-débit[8]. Le modèle sinusoïdal harmonique est également la base du modèle harmonique et résiduel présenté dans la section 1.3.3 ; son étude est donc une bonne introduction à ce dernier modèle.

[8]La transmission des coefficients de resynthèse sur des liens de communications à bas-débit (< 10 kbits/s) est abordée dans (Kleijn et Paliwal, 1995).

1.3.2.1 Modélisation des sons voisés

La modélisation des sons voisés se fait en trouvant la fréquence fondamentale (voir section 1.3.3.1) et en calculant la puissance et la phase de chaque harmonique. L'analyse est faite généralement sur des fenêtres de signal contenant de deux à quatre périodes fondamentales pondérées par des fonctions qui atténuent les effets de bords (p.e. une fonction triangulaire (Kleijn et Paliwal, 1995) ou de Hanning comme sur la figure 1.17). Nous ne nous attarderons pas ici aux détails techniques de l'analyse. La problématique ainsi que différentes méthodes sont décrites dans la section harmonique du modèle harmonique et résiduel présenté dans la section 1.3.3.

1.3.2.2 Modélisation des sons non-voisés

Les sons non-voisés (transitoires ou bruités) ne sont pas sinusoïdaux à la manière des sons voisés, mais peuvent tout de même être modélisés comme une somme de sinus ayant une relation harmonique. Comme dans le cas de la modélisation des sons voisés, nous chercherons une série de fonctions sinusoïdales expliquant la majeure partie de l'énergie du signal vocal. Nous voulons également que la fréquence fondamentale utilisée soit inférieure à 150 Hz, pour éviter que les sons bruités aient la sonorité résonnante caractéristique des signaux périodiques. Utiliser une basse fréquence fondamentale dans le cas de sons bruités permet d'approcher l'approximation sinusoïdale de la transformée de Fourier. La transformée de Fourier a une résolution de $\frac{f_e}{N}$ pour une fenêtre d'analyse de N points et un signal échantillonné à f_e Hz (donc environ 43 Hz pour une fenêtre de 1024 points à 44,1 kHz). Lorsque la fréquence fondamentale de l'approximation harmonique du bruit est près de cette valeur (p.e. 43 Hz), le modèle s'approche de la représentation complète du signal de la transformée de Fourier.

1.3.2.3 Synthèse avec un modèle sinusoïdal

Pour resynthétiser le signal vocal, il suffit de faire la somme des k sinusoïdes de fréquence instantanée $kf_0(t)$ et de puissance instantanée $a_k(t)$. La phase $\phi_k(t)$ du sinusoïde k est calculée en additionnant une différence de phase $2\pi\frac{f_0(t)}{f_e}$ à la phase précédente $\phi_k(t-1)$. Les fréquences et puissances instantanées peuvent être interpolées linéairement entre les centres des fenêtres d'analyse (voir 1.16). Dans (Moulines et Charpentier, 1990), on suggère d'utiliser la méthode de recouvrement et d'addition (*overlap-and-add*), qui consiste à synthétiser une fenêtre à la fois pour éviter une accumulation de l'erreur sur la phase due à l'analyse et l'interpolation. Les fenêtres se recouvrent aux extrémités et sont pondérées de manière à reproduire le signal original sans modification de puissance. La figure 1.18 montrée plus loin illustre une fenêtre d'analyse pondérée par une fenêtre de Hanning, pouvant être utilisée par la méthode de recouvrement et d'addition.

1.3.2.4 Notes sur le modèle sinusoïdal

Le modèle sinusoïdal que nous venons de décrire comporte quelques lacunes dues à sa simplicité. Premièrement, le modèle néglige toujours une partie du signal (sauf dans le cas purement périodique et harmonique), ayant comme conséquence une perte de qualité inévitable du signal vocal. De plus, alors qu'il permet des modifications du signal comme le changement de hauteur ou la dilatation temporelle, il ne permet pas de jouer sur le ratio d'harmonicité de la voix (section 1.2.2.3), par exemple pour rendre une voix plus suave ou pour transformer une voix en murmure. Alors que ces limites sont sans conséquence grave pour les systèmes téléphoniques de basse qualité (p.e. les téléphones cellulaires ou portables), elles deviennent dramatiques dans le cas de la synthèse de voix réaliste, où le locuteur doit non seulement être intelligible mais reconnaissable et naturel, dans le sens

où nous ne voulons pas que l'auditeur ait l'impression qu'une modification ait été appliquée au signal sonore. Le thème de la synthèse de voix réaliste est repris plus en détail à la fin de ce chapitre.

1.3.3 Modèle harmonique et résiduel

Le but de ce modèle est de séparer la voix en deux signaux : un signal périodique harmonique et un signal résiduel[9]. La partie harmonique (comparable au modèle sinusoïdal) permet de modéliser certaines caractéristiques acoustiques dues aux cordes vocales (comme la hauteur du son) et du conduit vocal. La partie résiduelle permet de modéliser le reste du signal, principalement dû aux turbulences de l'écoulement de l'air des poumons dans le conduit vocal. Parmi les trois modèles étudiés dans cette recherche, celui-ci est celui qui offre la plus grande qualité de voix. Il permet aussi d'appliquer des modifications de manière générale à la voix pour créer une grande variété d'intonations et de transformations sur le timbre. C'est donc celui sur lequel nous insisterons tout au long de ce document.

Un système de modélisation harmonique et résiduel peut devenir rapidement complexe lorsqu'on y ajoute des parties servant à traiter des cas spéciaux. Dans notre cas, nous avons bâti un système relativement simple qui nous permet d'arriver à de bons résultats. Le système est principalement composé de trois parties :

- Estimation de la fréquence fondamentale
- Mesure de la partie harmonique
- Mesure de la partie résiduelle

[9]Le signal résiduel est parfois appelé partie *bruitée* ou *stochastique*, mais nous préférons nous en tenir au terme "résiduel". Même si cette partie est resynthétisée avec un générateur de nombres pseudo-aléatoires, le signal résiduel peut aussi contenir des signaux transitoires et peut avoir un spectre fréquentiel contenant peu d'information (donc peu "aléatoire").

La schéma de la figure 1.12 résume les étapes de notre système de décomposition et modélisation harmonique et résiduel.

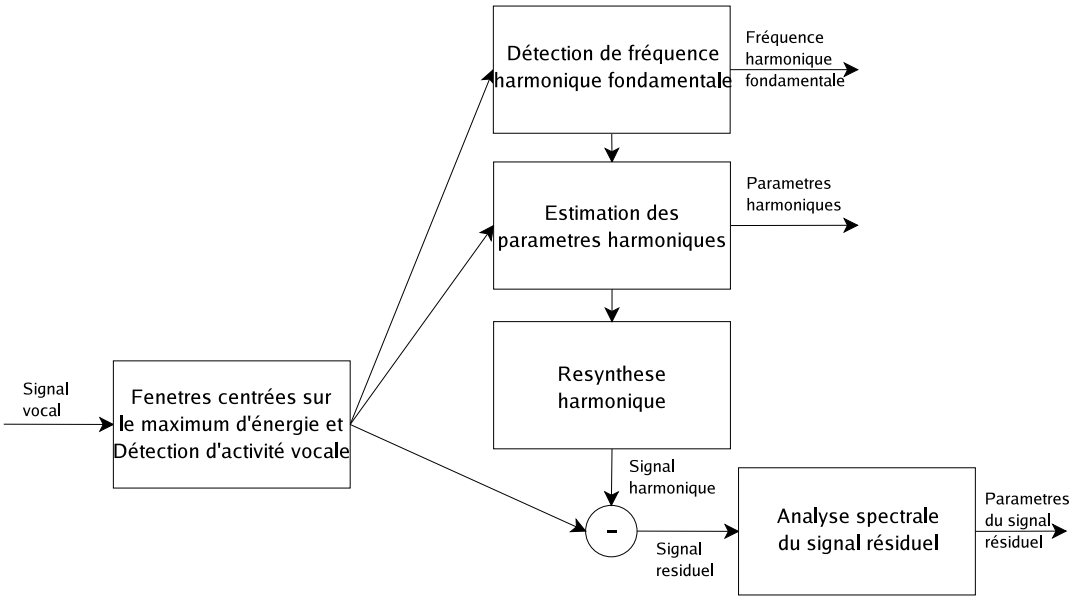


FIG. 1.12 Schéma du système de modélisation harmonique et résiduelle du signal vocal.

1.3.3.1 Estimation de la fréquence fondamentale

Il existe une littérature abondante sur le problème d'estimation de la fréquence fondamentale d'un signal harmonique. L'estimation de la fondamentale est un problème clé dans plusieurs algorithmes de compression et de traitement de la voix. Pour cette raison, presque tous les ouvrages concernant le traitement de la parole contiennent une section réservée à ce sujet. Nous couvrirons ici quelques méthodes que nous jugeons intéressantes, soit l'analyse des sommets du spectre de puissance de la transformée de Fourier rapide (FFT), la méthode d'autocorrélation, le filtre à délai ainsi que quelques modifications permettant d'améliorer les résultats de ces méthodes.

Dans (Serra, 1997), l'auteur décrit une méthode pour trouver la fréquence fondamentale d'un signal harmonique à l'aide des **sommets du spectre de puissance de la FFT**. Dans un signal vocal harmonique, les harmoniques se distinguent par des sommets prononcés dans le spectre de puissance, particulièrement dans les basses fréquences (voir la figure 1.13). La méthode de Serra est donc de trouver une fréquence fondamentale pour laquelle ces sommets contiennent la plus grande partie de la puissance. L'algorithme parcourt le domaine des fréquences fondamentales f_0 possibles (fixé par l'usager), et calcule la puissance expliquée par les harmoniques situés aux fréquences $k \times f_0$. La fréquence fondamentale f_0 pour laquelle le spectre contient le plus d'énergie harmonique est choisie comme étant la fréquence fondamentale du signal. Cette méthode a comme avantage d'être simple et rapide, mais elle dépend de la précision de l'interpolation du spectre entre les points calculés par la FFT. Par exemple, une FFT a 1024 points d'un signal à 44,1 kHz a une résolution de 21.53 Hz. Pour les fréquences entre ces points, une approximation doit être faite pour estimer la puissance. Une approximation quadratique avec les trois points les plus élevés autour de cette fréquence peut être utilisée pour estimer un sommet harmonique.

Nous voulons aussi souligner quelques autres difficultés qui s'appliquent à cette méthode ainsi qu'à toutes les autres décrites dans cette section. Premièrement, lorsque l'on calcule l'énergie d'un signal harmonique dont la fondamentale est située à une fraction de la fréquence fondamentale réelle (par exemple $\frac{1}{2}$ ou $\frac{1}{3}$) nous mesurons toujours une énergie égale ou supérieure à celle mesurée pour la vraie fréquence fondamentale. Ceci est vrai parce qu'en plus d'additionner les énergies de tous les harmoniques, nous ajoutons de l'énergie située entre les sommets harmoniques du spectre de puissance. Ce phénomène est couramment appelé *frequency halfing*, parce qu'il survient le plus fréquemment à la moitié de la fréquence fondamentale réelle, tel qu'illustré sur la figure 1.14. Sur cette figure, les vrais sommets harmoniques sont

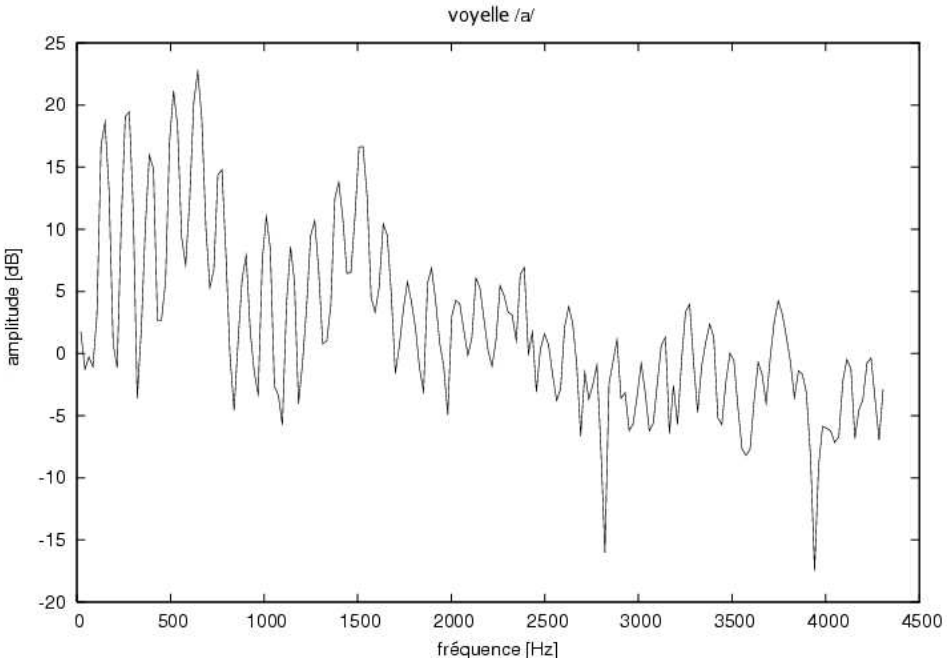


FIG. 1.13 Spectrogramme d'une voyelle /a/.

encerclés par des ronds noirs, et les harmoniques trouvés dans un cas de *frequency halfing* sont encerclés par des ronds rouges pointillés.

Pour résoudre ce problème, il est possible de fixer un seuil d'énergie pour le premier harmonique. Cependant, sur certains vieux enregistrements, les basses fréquences (inférieures à 100 Hz) ont été filtrées par de mauvais supports audio, et le seuil n'est pas rencontré pour les voix d'hommes très basses[10]. À cause de cela, il est préférable de fixer un seuil pour chacun des premiers harmoniques, et vérifier que la majorité de ces harmoniques atteigne le seuil.

Nous devons aussi souligner que la fréquence fondamentale d'un signal vocal change constamment. Elle change faiblement dûes à des microvariations de la tension des cordes vocales et de manière plus importante à cause de l'intonation et de l'accen-

[10]Par exemple, la voix de l'acteur Humphrey Bogart sur le DVD du film *Casablanca* a parfois une hauteur de 75 Hz, et l'enregistrement rend l'harmonique fondamental presque absent.

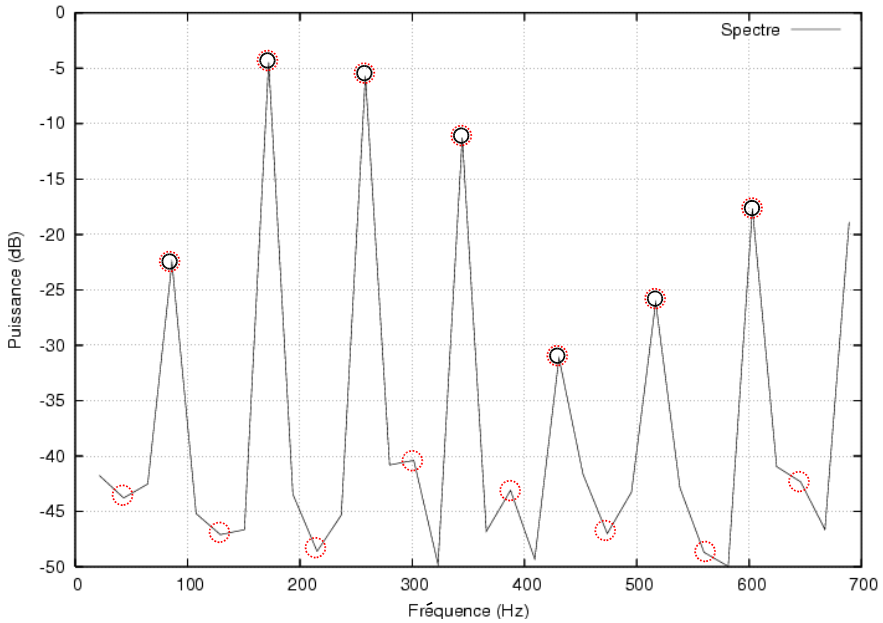


FIG. 1.14 Spectrogramme d'une voyelle illustrant la cause du *frequency halfing*.

tuation. Même les individus dont la voix semble très monotone accentuent normalement certaines syllables[11]. Le problème est que les méthodes d'analyse numérique supposent une certaine stabilité du spectre fréquentiel du signal à l'intérieur d'une fenêtre d'analyse, condition qui n'est pas toujours satisfaite. Toutes les techniques décrites dans ce mémoire supposent que les propriétés du signal restent constantes sur une durée d'au moins deux (pour les locuteurs avec une voix basse) à quatre (pour les locuteurs avec une voix haute) périodes fondamentales[12], ce qui veut dire concrètement une stabilité sur une période de 15 à 25 ms[13]. Les variations brusques

[11]L'accentuation de certaines syllables est très importante pour une bonne compréhension de certaines langues européennes comme l'anglais, l'espagnol ou l'italien, et elle est essentielle pour des langues comme le chinois. L'absence totale d'intonation est anormale, voire pathologique pour un locuteur dans la majorité des langues.

[12]D'autres méthodes d'analyse-synthèse qui ne sont pas couvertes par ce document (par exemple la méthode PSOLA et ses variantes) traitent les périodes du signal harmonique une à la fois et n'ont pas ce problème. Malgré leur très grande popularité récemment, nous avons cependant décidé de ne pas travailler avec ces méthodes après avoir été déçu par certains logiciels l'utilisant.

[13]Cette hypothèse permet de travailler avec des fenêtres d'une longueur de 1024 points pour

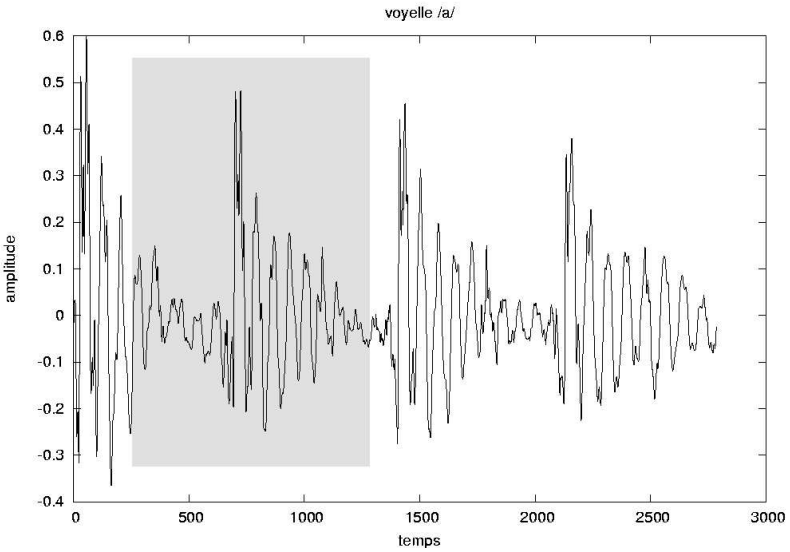


FIG. 1.15 Diagramme temporel d'une voyelle /a/.

de la hauteur peuvent donc rendre les sommets du spectre de puissance "flous". C'est le problème de l'incertitude temps-fréquence vu par l'ingénieur de traitement de signal[14]. Ce problème d'incertitude est en partie corrigé par une pondération de la fenêtre d'analyse qui atténue les extrémités (par exemple une pondération avec une fenêtre de Hanning ou de Blackman), ce qui permet de concentrer l'analyse du signal sur ce qui se passe au milieu de la fenêtre. Le processus de fenêtrage et d'atténuation des bords est illustré dans les figures 1.15, 1.16, 1.17 et 1.18.

La deuxième technique que nous voulons présenter est la **méthode de l'autocorrélation** (Rabiner et Juang, 1993) (Boite et al, 2000). Le principe ici est encore assez simple. La périodicité est définie par un événement qui se reproduit à intervalle régulier. Si nous calculons l'autocorrélation du signal avec lui-même à différents intervalles, le résultat maximal sera atteint pour un intervalle j égal à un multiple de la période fondamentale. La formule suivante représente l'autocorrélation d'un signal de moyenne nulle avec un retard j sur un intervalle T.

des signaux dont la fréquence d'échantillonnage est de 44,1 kHz, donc de 23,2 ms.

[14]Par analogie au problème de l'incertitude position-vitesse en physique quantique.

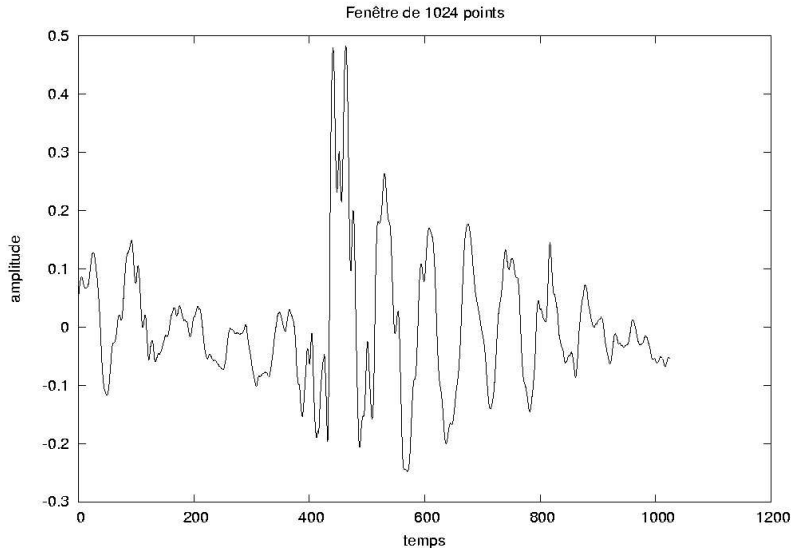


FIG. 1.16 Une fenêtre d'analyse de la voyelle /a/.

$$r(j) = \sum_t x_t x_{t-j}$$

La valeur de **r** pour différentes valeurs de j peut être calculée simplement avec une convolution, qui est équivalente pour des signaux réels à moyenne nulle :

$$r(j) = (x(t) * x(j-t))$$

Dans le domaine fréquentiel, cette opération revient à multiplier le spectre avec lui-même pour des intervalles de fréquence $\frac{1}{j}$ et de calculer l'énergie résultante. Lorsque les sommets fréquentiels sont alignés, nous obtenons l'énergie maximale.

Une autre méthode pour trouver la fréquence fondamentale expliquant le plus de contenu harmonique est la **méthode du filtre à délai** (*tap-delay filter*). Avec

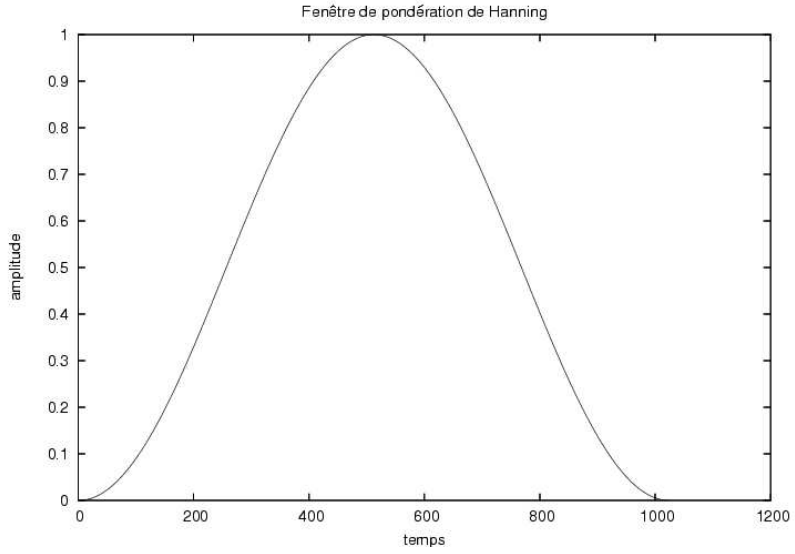


FIG. 1.17 La fenêtre de pondération de Hanning utilisée.

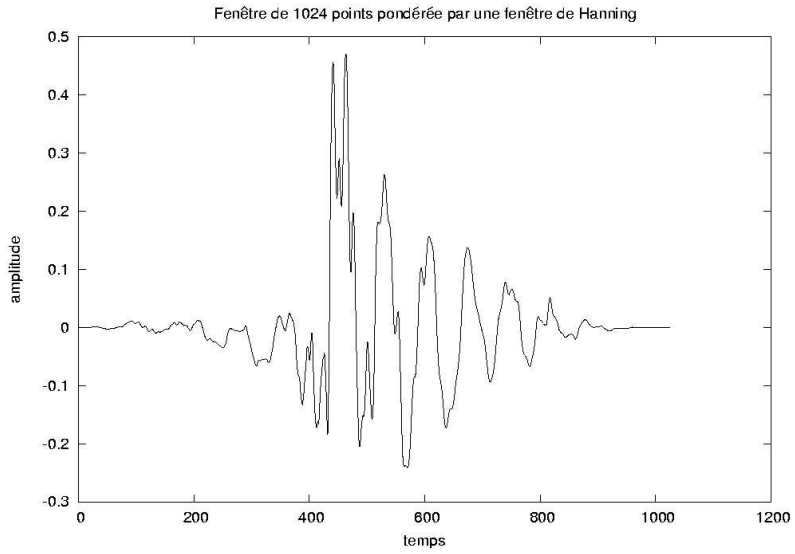


FIG. 1.18 Une fenêtre d'analyse pondérée de la voyelle /a/.

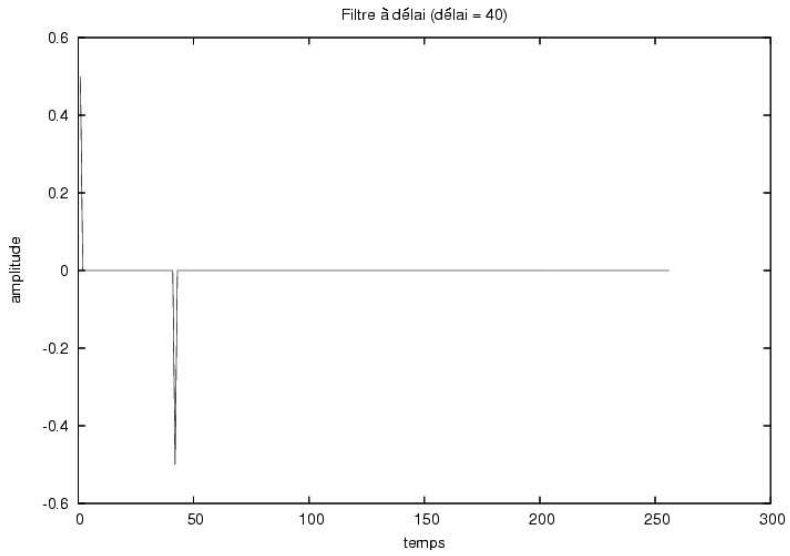


FIG. 1.19 Coefficients temporels d'un filtre à délai à réponse finie (FIR).

cette technique, chaque échantillon est soustrait par un échantillon j points plus tôt. Une fois le signal filtré, on mesure l'énergie résultante. Le résultat minimal s'obtient alors pour la période j correspondant à la fréquence fondamentale.

$$Y(j) = \sum_t (0.5x_t - 0.5x_{t-j})^2$$

Dans le domaine fréquentiel, cette opération est équivalente à un filtre comportant des zéros à toutes les fréquences $\frac{k}{j}$ (voir la figure 1.20). Pour cette raison, le filtre est parfois appelé *comb filter*. Évidemment, un long délai atténue plus de fréquences, et augmente le risque de *frequency halfing*. Nous pouvons alors appliquer les tests de seuil sur les harmoniques comme expliqué plus haut.

Dans (Quatieri, 2001), l'auteur décrit une **méthode de régularisation** qui tient compte de l'énergie captée par des filtres avec différents délais pour éviter de faire des tests sur la présence de certains harmoniques. Les filtres à long délai (donc

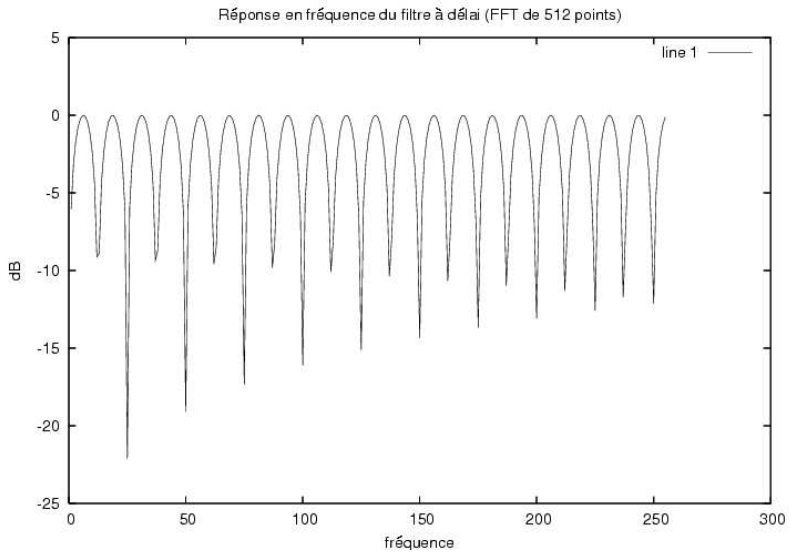


FIG. 1.20 Réponse fréquentielle du filtre à délai de la figure 1.19.

basse fréquence fondamentale) sont pénalisés par rapport aux filtres à court délai, ce qui permet de discriminer les minima d'énergie situés à des fractions de la fréquence fondamentale réelle. Grâce à cette méthode, nous pouvons calculer les résultats de différents filtres à délai pour tout le signal, et ensuite appliquer une fenêtre coulissante pondérée (par exemple une pondération gaussienne) sur chaque résultat pour prendre le minimum d'énergie à chaque point. Cette technique s'avère très précise (en temps et en fréquence), et c'est celle que nous avons choisie pour notre système.

1.3.3.2 Enveloppe de puissance et de phase des harmoniques

Une fois la fréquence fondamentale trouvée pour une fenêtre d'analyse donnée, la séparation des harmoniques du reste du signal est une opération relativement triviale. Notre méthode consiste à soustraire successivement des sinusoïdes de fréquence kf_0 de la fenêtre d'analyse. L'amplitude et la phase de ces sinusoïdes sont trouvées

avec une transformée discrète de Fourier centrée sur la fréquence d'intérêt.

$$h(t) = \sum_k sin(2\pi \frac{kf_0}{f_e} t + \phi_k) = \sum_k a_k sin(2\pi \frac{kf_0}{f_e} t) + b_k cos(2\pi \frac{kf_0}{f_e} t)$$

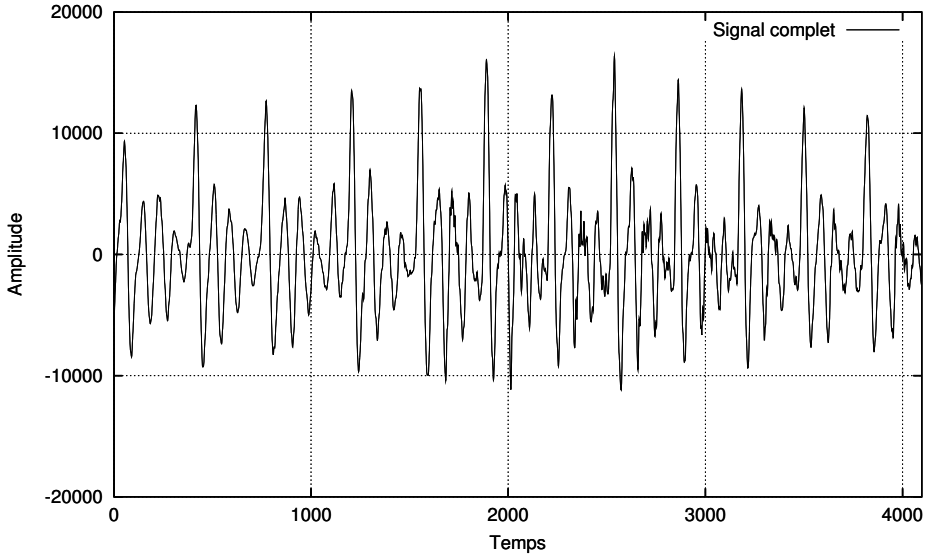


FIG. 1.21 Voyelle /o/ (signal original $s(t)$).

La fréquence kf_0 est considérée constante sur la fenêtre d'analyse, une hypothèse raisonnable étant donné que la fenêtre contient généralement moins de quatre périodes fondamentales du signal. De plus, 92% de l'énergie de la fenêtre de Hanning est contenu dans 50% des points situés au centre de la fenêtre, donc sur moins de deux périodes fondamentales (tel qu'illustré par la figure 1.18).

Les figures 1.22 et 1.23 représentent respectivement les parties harmoniques et résiduelles de la voyelle /o/ de la figure 1.21.

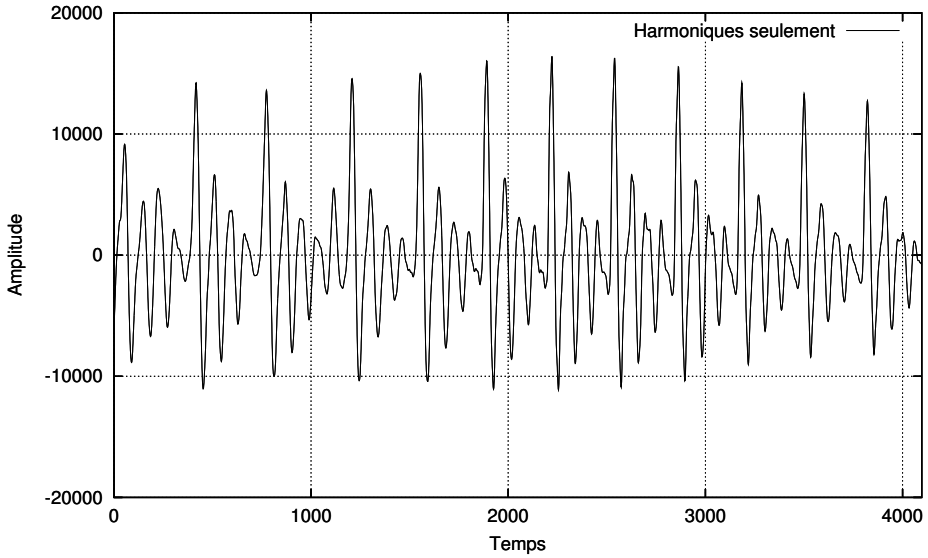


FIG. 1.22 Voyelle /o/ (signal harmonique $h(t)$).

1.3.3.3 Enveloppe de puissance du signal résiduel

Ce qui reste après la soustraction des harmoniques est considéré comme la partie résiduelle. Afin de conserver les propriétés timbrales dans le cas d'une dilatation dans le temps du signal, nous modélisons ce signal par l'enveloppe de puissance de son spectre fréquentiel. Nous lissons également cette enveloppe pour éviter les artefacts dus à la soutraction d'harmoniques. En effet, une partie du bruit est soustraite avec les harmoniques, ce qui produit parfois un effet de voix fantôme lorsqu'une transformation de hauteur est faite sur les harmoniques et que le bruit est gardé tel quel. L'équation est donc aussi simple que :

$$r(t) = s(t) - h(t)$$

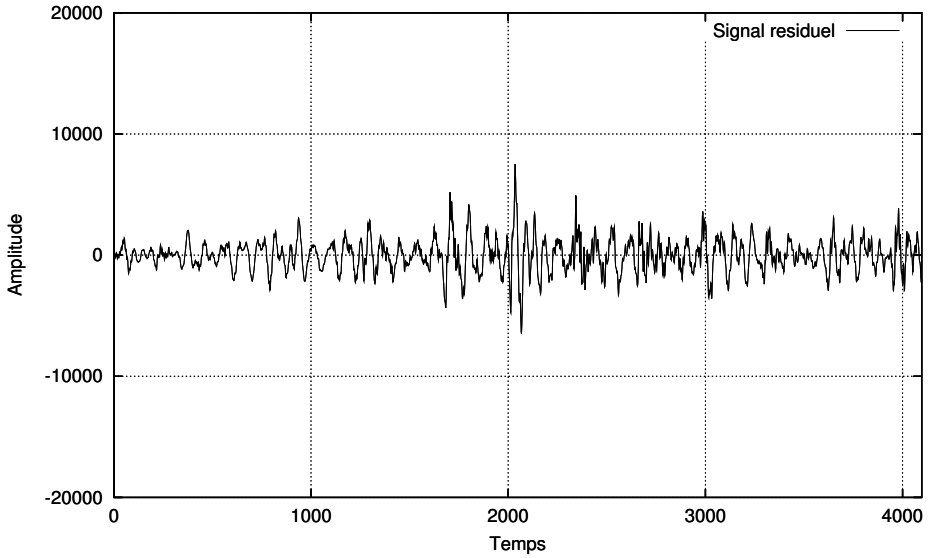


FIG. 1.23 Voyelle /o/ (signal résiduel $r(t)$).

Où $s(t)$ est le signal complet, $h(t)$ est la partie harmonique (pondérée par une fenêtre de Hanning comme le signal complet $s(t)$ de la fenêtre d'analyse), et $r(t)$ est la partie résiduelle. Le lissage (ou régularisation) de la partie résiduelle est fait avec un noyau gaussien d'environ 100Hz d'écart-type afin de bien couvrir l'espace entre deux harmoniques.

1.3.3.4 Algorithme de synthèse

Une fois le modèle harmonique et résiduel construit, la synthèse se fait relativement simplement. Les harmoniques sont synthétisés point par point en calculant l'amplitude et la phase de chaque harmonique k au temps t. L'amplitude instantanée est calculée par une interpolation linéaire entre les points d'analyse. La phase est calculée en estimant la fréquence instantanée linéairement entre le milieu des fenêtres d'analyse originales. L'interpolation linéaire s'est révélée suffisante pour toutes

nos expériences. Une interpolation par splines cubiques a aussi été essayée dans plusieurs situations, sans apporter d'amélioration acoustique intéressante. L'interpolation linéaire a donc été choisie pour l'approximation des amplitudes et des fréquences instantanées pour sa simplicité et son faible coût computationnel. La phase originale est choisie au milieu d'un segment d'analyse et non pas au début d'une syllabe, où le calcul de phase est moins précis.

Le résidu est synthétisé en filtrant un bruit blanc pour lui donner le spectre mesuré et régularisé de la partie résiduelle originale. Nous avons décidé de le synthétiser fenêtre par fenêtre, et de juxtaposer les segments en appliquant la même fenêtre de Hanning utilisée pour la décomposition harmonique et résiduelle. Cette technique nous permet de retrouver la partie résiduelle originale.

1.3.3.5 Exemples de transformations possibles de la voix avec le modèle harmonique et résiduel

À partir du modèle harmonique et résiduel, certaines transformations peuvent être appliquées systématiquement sur le modèle afin d'obtenir une voix intelligible et caractéristique d'un locuteur. Nous décrivons brièvement dans cette section comment il est possible de changer la hauteur, le rythme et le ratio d'harmonicité de la voix.

Le changement de hauteur par un coefficient η se fait en synthétisant des harmoniques de fréquence $\eta k f_0$. L'amplitude de chaque harmonique doit suivre l'enveloppe de puissance harmonique lorsqu'elle est disponible, sinon l'amplitude est interpolée linéairement entre les mesures des amplitudes des harmoniques originaux[15]. La partie résiduelle est synthétisée sans modification. Dans (Stylianou,

[15]La transposition d'un facteur η se fait généralement sans altérer le timbre significativement

1998), l'auteur décrit comment il applique cette technique pour la synthèse de voix par concaténation de diphone avec de très bons résultats.

Le changement de rythme se fait en rapprochant ou en espaçant les centres des fenêtres d'analyse lors de la resynthèse. La partie harmonique est synthétisée normalement, en interpolant les fréquences et amplitudes entre les points d'analyse. La partie résiduelle est synthétisée normalement, en ajustant la largeur des fenêtres de Hanning pour qu'elles correspondent à l'espacement des fenêtres.

Les variations du taux d'harmonicité permettent de créer des effets intéressants. Pour obtenir un chuchottement, une voix "jazzée" ou plus sensuelle, on peut augmenter la puissance du résidu par rapport à la partie harmonique. On peut aussi réduire cette caractéristique de la voix en augmentant la puissance relative des harmoniques.

D'autres transformations de la voix qui n'entrent pas dans le cadre de notre recherche sur les voix réalistes sont possibles, par exemple les voix de robots, cristallines, etc. Il serait pertinent d'écrire un ouvrage s'intéressant spécialement à ces voix, qui sont souvent des créations artisanales et mal documentées (du moins à notre connaissance). Nous n'avons cependant pas l'intention de les décrire ici, alors revenons à notre sujet.

1.4 Adaptation de la reconnaissance pour différents locuteurs

Les publications concernant la variabilité du timbre selon le locuteur se sont jusqu'ici surtout concentrées sur le problème de la reconnaissance vocale. Encore ici,

lorsque η est entre 0,6 et 1,7. Lorsque ce coefficient est plus grand que 2, l'altération est souvent perceptible, car l'enveloppe de puissance harmonique diffère selont la hauteur. Pour cette raison, il est préférable d'avoir des échantillons originaux situés à plusieurs endroits différents du registre lorsque nous voulons faire des transpositions réalistes pour la voix chantée.

le sujet est rarement abordé. Une cause possible de cette rareté serait que les systèmes de reconnaissance pouvant s'adapter au locuteur sont souvent propriété d'entreprises voulant protéger leurs secrets industriels. Une autre explication serait que l'adaptation au locuteur est peut-être difficile ou inutile étant donné la courte durée moyenne des conversations téléphoniques humain-machine.

Il est mentionné dans (Huang, Acero et Hon, 2001) que les systèmes de reconnaissance vocale conçus pour s'adapter au locuteur nécessitent des exemples provenant d'environ 500 locuteurs. Le système mentionné utilise des mélanges de gaussiennes (GMM) pour caractériser certains timbres typiques. Ces GMM sont entraînées sur les coefficients cepstraux[16] des phonèmes, couramment utilisés pour la détection de caractéristiques spectrales en reconnaissance.

1.5 Notre contribution aux techniques de traitement de la parole

Nous voudrions dans cette section souligner notre contribution aux techniques de traitement de la parole en divisant le domaine en trois catégories : la compression, la reconnaissance de caractéristiques et l'analyse-synthèse.

La compression de la voix est un sujet de recherche pour lequel il y a une grande littérature. D'abord étudié pour la transmission de la voix sur les premiers systèmes de télécommunication numériques, le sujet s'est enrichi par les besoins issus de la démocratisation du téléphone cellulaire et la perspective du transport de voix sur des réseaux utilisant le protocole Internet (IP). L'objectif de la compression est d'encoder la voix avec le moins d'information possible tout en la gardant intelligible. Une bonne compression permet d'utiliser plus efficacement les liens de

[16] Les coefficients cepstraux sont le résultat de la transformée de Fourier du logarithme de la transformée de Fourier d'un signal ($TF[log(TF[s(t)])]$). Ils permettent d'extraire certaines caractéristiques du spectre du signal vocal.

télécommunication.

La reconnaissance de caractéristiques a aussi été motivée en partie par la téléphonie. La reconnaissance vocale est utilisée dans divers systèmes de réponse automatique (comme l'annuaire téléphonique ou la confirmation de vols d'avion) de même que dans les logiciels de dictée automatique. Elle a pour objectif d'identifier les phrases et les mots prononcés par un locuteur. La reconnaissance d'émotions (Polzin et Waibel, 1998)(Batliner et al, 2000) essaie de distinguer certains états émotifs du locuteur (par exemple patient/impatient, colérique/doux, triste/joyeux). Dans (Levit et al, 2001), on s'intéresse au niveau d'alcoolémie du locuteur, dans le but de savoir s'il peut conduire un véhicule.

L'analyse-synthèse est l'analyse du signal vocal dans le but de faire des modifications, de la segmentation ou de la concaténation pour la synthèse de voix. Nous incluons dans cette catégorie la synthèse de la parole en général (à partir d'un texte[17] ou non). Les voix de synthèse sont créées en représentant une voix réelle par un modèle qui se compare parfois à ceux utilisés pour la compression. Cependant, le modèle pour la synthèse ne vise pas nécessairement à réduire l'information nécéssaire à restituer la voix, il veut surtout pouvoir la restituer dans plusieurs contextes en y appliquant des modifications. Pour cette raison nous en faisons une catégorie à part.

Dans cette recherche, nous ne traitons pas le langage (comme les systèmes de reconnaissance de la voix ou la synthèse à partir d'un texte), seulement la voix comme une suite de phonèmes. Nous ne tentons pas non plus de compresser la voix. Notre projet est de *compléter le corpus d'un individu* lorsque celui-ci est incomplet pour la synthèse d'un phonème en particulier. Il n'y a pas à notre connaissance de littérature sur la prédiction de phonèmes pour la synthèse vocale de locuteurs

[17]La synthèse à partir d'un texte est appelée communément *text-to-speech*.

particuliers. Nous partirons donc de la théorie expliquée dans ce chapitre et le suivant pour construire notre expérience.

CHAPITRE 2

THÉORIE DES SYSTÈMES ADAPTATIFS UTILISÉS

Dans le chapitre précédant, nous avons décrit le signal vocal, ses propriétés ainsi que la problématique entourant sa modélisation. Dans ce chapitre, nous allons nous pencher sur l'apprentissage machine.

Nous commencerons par décrire quelques concepts généraux de l'apprentissage machine. Par la suite, nous introduirons les méthodes utilisées dans le cadre de cette recherche, soit l'approximation par noyaux, l'analyse en composantes principales et l'analyse en composantes indépendantes. La lecture de ce chapitre permettra de mieux comprendre la démarche détaillée dans le prochain chapitre.

2.1 Types de problèmes de l'apprentissage machine

Avant de décrire les différentes méthodes d'apprentissage utilisées dans cette recherche, nous nous devons de définir quelques concepts généraux de l'apprentissage machine. Dans cette section nous expliquons d'abord la différence entre l'apprentissage supervisé et non-supervisé, puis nous séparons les problèmes de l'apprentissage machine en trois groupes, soit la classification, la régression et l'estimation de densité de probabilité. Nous finissons par quelques notes sur les bases de données utilisées pour entraîner les algorithmes d'apprentissage.

2.1.1 Apprentissage supervisé et non-supervisé

L'apprentissage supervisé est possible quand les exemples disponibles sont sous forme de paires entrées-sorties. Par exemple, des exemples de différentes classes déjà étiquetés ou les entrées et sorties d'une fonction quelconque. On souhaite alors qu'une machine apprenne à prédire la sortie d'une nouvelle entrée qui lui est présentée. Le terme supervisé vient du fait que les exemples sont déjà "interprétés" par un "professeur" qui leur a assigné une sortie. Les algorithmes tels les *noyaux*, les *réseaux bayesiens*, la *régression* et les *réseaux de neurones* font partie de cette catégorie.

Dans l'apprentissage non-supervisé, on ne dispose que d'un ensemble d'exemples d'entrées. Nous désirons alors qu'une machine regroupe ces entrées ou apprenne à leur assigner une probabilité. Les algorithmes de partitionnement comme les *k-moyennes* et d'estimation de densité comme les *mélanges de gaussiennes* font partie de cette catégorie. Les algorithmes de réduction de dimensionnalité (comme le PCA ou l'ICA) font aussi partie de cette catégorie.

2.1.2 Classification

La tâche de classification consiste à apprendre à assigner une classe à de nouveaux exemples à partir d'une source d'exemples déjà étiquetés. L'apprentissage est donc toujours supervisé. Des problèmes classiques dans la littérature sont, par exemple, la reconnaissance de caractères manuscrits, la classification de champignons selon certaines mesures faites sur les spécimens, la présence ou non de cancer chez des patients à partir de données médicales ou les types de terrain (champ, route, lac) présents sur une photo satellite.

2.1.3 Régression

Les algorithmes cherchant à approximer la sortie d'une fonction (linéaire ou non) accomplissent la tâche de régression. Par exemple, on cherche à approximer le prix d'une résidence en fonction de différentes caractéristiques (accès à la mer, nombre de pièces, grandeur du terrain) ou construire un modèle de l'épaisseur de l'aorte suivant l'âge et le sexe d'une personne (Hastie, Tibshirani et Friedman, 2001) à partir d'exemples.

2.1.4 Estimation de densité de probabilité

L'estimation de densité consiste à approximer la distribution des données. Le modèle doit donner un estimé de la densité de probabilité d'un échantillon provenant de la même distribution que les données d'entraînement.

Estimer la densité de probabilité de plusieurs classes permet aussi de résoudre le problème de classification. Cependant, l'estimation de densité de probabilité est normalement un problème beaucoup plus difficile que les deux premiers étant donné le faible nombre d'échantillons d'entraînement habituellement disponibles (voir la section 2.1.5 plus bas).

Certains modèles ainsi construits permettent de générer des échantillons probables selon la distribution estimée. On appelle ces modèles *génératifs* (Ackley, Hinton et Sejnowski, 1985) (Frey, 1998) (Haykin, 1999).

2.1.5 Notes sur les bases de données servant à l'entraînement

Les données servant à l'entraînement sont habituellement limitées, contrainte qui rend souvent beaucoup d'algorithmes complètement inefficaces devant la tâche à accomplir. C'est le cas de cette recherche, où il ne nous était pas possible de recueillir les caractéristiques acoustiques de milliers de locuteurs. Les modèles génératifs auraient été particulièrement intéressants dans notre cas, mais ils n'étaient pas pertinents à cause de notre base de données restreinte.

2.2 Noyaux

Les noyaux sont des algorithmes de prédiction basés sur une mesure de distance entre différents exemples appelés échantillons. On dit aussi que ce sont des modèles basés sur des distances à des prototypes (Vincent, 2003). Ils utilisent une fonction locale d'approximation pour prédire la valeur d'une fonction à l'aide de la valeur de cette fonction dans un certain voisinage. Ce voisinage est défini par une mesure de distance et un noyau. Nous définissons ces concepts dans cette section.

2.2.1 Mesures de distance

La définition commune et aussi la plus intuitive[1] de mesure de distance est la distance euclidienne :

$$d_2(s_i, s_j) = \sqrt{||s_i - s_j||^2}$$

[1]Nous disons intuitive car elle utilise le théorème de Pythagore.

où s_i et s_j sont deux vecteurs. Cette mesure de distance peut être généralisée par la mesure de *p-distance*[2] :

$$d_p(s_i, s_j) = \sqrt[p]{||s_i - s_j||^p}$$

La distance euclidienne est alors vue comme la 2-distance. La 1-distance est aussi appellée distance de Manhattan.

Lorsque les différentes dimensions des vecteurs s_i sont de nature différente, il est courant de les normaliser. Par exemple, il est d'usage en apprentissage machine de normaliser toutes les variances des dimensions d'un groupe d'échantillons à 1.

Dans les sous-sections suivantes, nous décrivons quelques échelles permettant de pondérer le spectre de puissance sonore en tenant compte de propriétés psychoacoustiques dues à l'anatomie de l'oreille interne.

2.2.1.1 Échelle de Mel

Dans le cas du spectre de puissance vocal, l'échelle de Mel sert à reproduire l'importance que l'oreille donne à chaque bande de fréquence. L'oreille a une plus grande définition en basse fréquence qu'en haute fréquence. L'échelle de Mel approxime cette définition en donnant plus d'importance aux basses fréquences dans l'évaluation psychoacoustique du spectre de puissance. La figure 2.1 décrit la résolution de l'oreille en fonction de la fréquence.

[2]La p-distance est aussi appellée métrique de Minkowski (Duda, Hart et Stork, 2001)

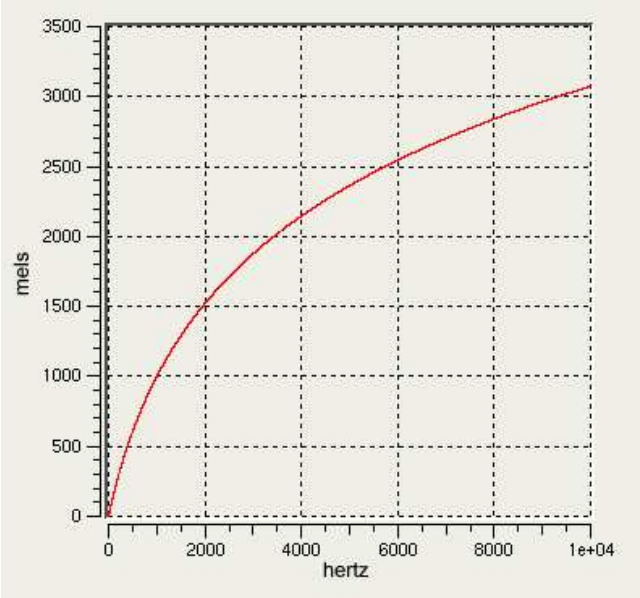


FIG. 2.1 Échelle perceptuelle de Mel.
Source : http ://en.wikipedia.org/wiki/Image :Mel-Hz-plot.png

2.2.1.2 Courbes d'isosonie

La sensibilité de l'oreille varie selon la fréquence du signal sonore. Les courbes d'isosonie représentent cette variation de sensibilité, et peuvent être utilisées pour pondérer le spectre sonore afin de pouvoir en comparer les parties suivant leur importance psychoacoustique. La figure 2.2 montre que la sensibilité maximale se situe entre 3 et 4 kHz.

2.2.1.3 Norme dB(A)

La norme dB(A) [3] définit une courbe de pondération pour le spectre de puissance acoustique selon l'importance psychoacoustique de chaque fréquence. Cette courbe est plus simple que la courbe d'égale intensité sonore et permet d'être appliquée fa-

[3]Cette norme est décrite dans le standard militaire *MIL-STD-188* et le standard fédéral *1037C* du gouvernement américain.

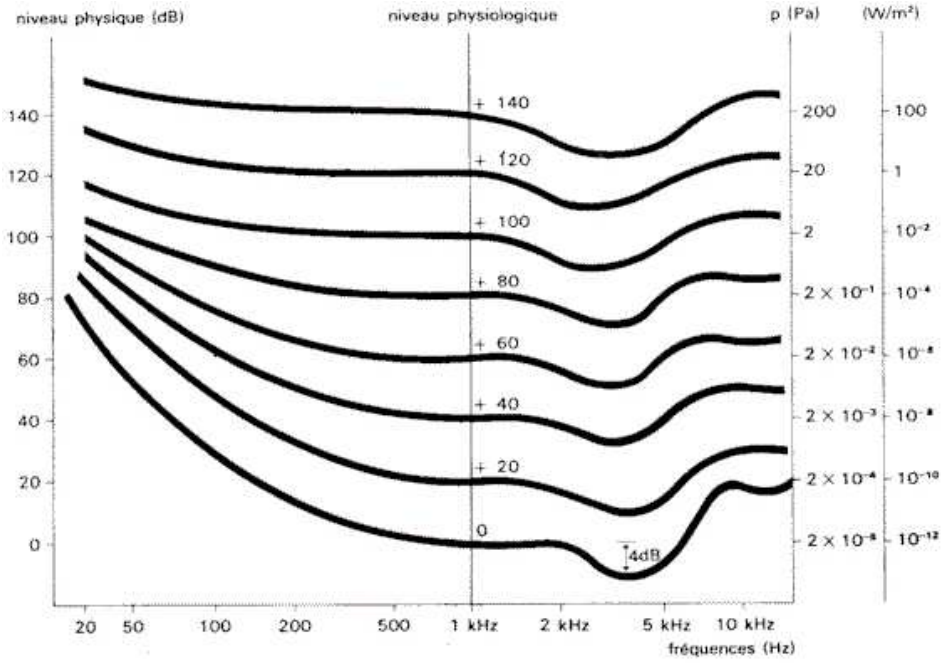


FIG. 2.2 Courbes d'isosonie pour l'oreille humaine.
Source :
http ://www.epsic.ch/branches/electronique/techn99/acous/aqvoitxt.html

cilement sur des signaux mesurés. La norme dB(A) situe le maximum de sensibilité autour de 3 kHz (voir figure 2.3).

2.2.2 Types de noyaux

Dans la sous-section précédente, nous avons défini quelques mesures de distance utilisées pour comparer les spectres de puissance sonore. Une fois calculée la distance entre les échantillons, nous devons décider comment nous allons utiliser les échantillons du voisinage. Dans cette section nous décrirons quelques noyaux pouvant être utilisés pour l'approximation de fonction. Ces noyaux font tous l'hypothèse que la fonction à approximer varie lentement localement. Évidemment, pour approximer une fonction en un point à l'aide de noyaux, nous devons avoir

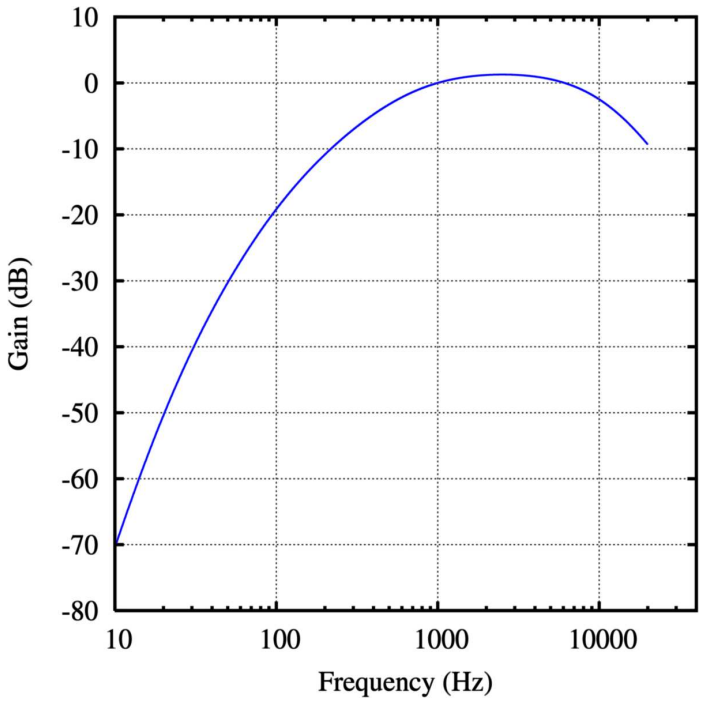


FIG. 2.3 Perception de la puissance du son selon la norme dB(A).
Source : http ://en.wikipedia.org/wiki/Image :Dba_plot.png

des échantillons semblables autour du point d'observation. La taille du voisinage utilisée et la pondération accordée aux voisins définira la qualité de la prédiction pour un problème en particulier.

2.2.2.1 Les k plus proches voisins

L'algorithme des k plus proches voisins (KNN) calcule une prédiction en faisant la moyenne des k plus proches voisins d'un vecteur d'entrée. Dans notre cas, ce vecteur d'entrée est composé des spectres des voyelles connues d'un nouveau locuteur s_j, et la prédiction doit permettre d'approximer la voyelle manquante y_j de ce nouveau locuteur. Cette approximation est résumée par la formule suivante :

$$\hat{y}_j(s) = Ave(y_i|s_i \in N_k(s))$$

où $Ave()$ est la fonction moyenne, s est l'ensemble de tous les locuteurs (sauf le nouveau locuteur s_j), et $N_k(s)$ est l'ensemble des k plus proches voisins de s_j dont les spectres y_i sont utilisés dans la prédiction.

2.2.2.2 Moyenne pondérée par un noyau Nadaraya-Watson

Le noyau Nadaraya-Watson est une forme générale pour calculer la prédiction $\hat{y}_j(s_j)$ d'un noyau défini par un voisinage $K_\lambda(s_j, s_i)$ (Hastie, Tibshirani et Friedman, 2001). La fonction K_λ peut être n'importe quelle fonction bornée ou non dont la valeur est finie sur le domaine d'intérêt. En général la fonction K_λ est monotone décroissante lorsque l'on s'éloigne du point à estimer.

Le noyau de Nadaraya-Watson calcule une moyenne pondérée de chaque échantillon du voisinage selon la valeur de la fonction de voisinage K_λ :

$$\hat{y}_j(s_j) = \frac{\sum_i K_\lambda(s_j, s_i) y_i}{\sum_i K_\lambda(s_j, s_i)}$$

La fonction de voisinage gaussienne est souvent utilisée parce qu'elle est continue, dérivable et définie sur tout le domaine des distances $[0, \infty+]$:

$$K_\lambda(s_j, s_i) = \frac{1}{\lambda} exp[\frac{-||s_i - s_j||^2}{2\lambda}]$$

Le paramètre λ définit la largeur du voisinage. Un λ plus grand signifie un voisinage plus grand. Ce noyau s'adapte lui-même relativement bien à une densité d'échantillonnage irrégulière.

2.2.2.3 Malédiction de la dimensionnalité

Il est facile de voir que le nombre d'échantillons nécessaires pour avoir une certaine densité dans un espace croît exponentiellement avec le nombre de dimensions. Lorsque le nombre de dimensions d'entrée de la fonction approximée est grand, la densité devient rapidement faible et la distance entre les échantillons tend à être égale dans toutes les dimensions. Ce phénomène est communément appelé la *malédiction de la dimensionnalité*. La perte de résolution exponentielle de l'échantillonnage en fonction du nombre de dimensions est aussi appellée *myopie* par analogie avec le problème optique.

La fonction de distance des noyaux doit être bien choisie pour éviter ce problème, en tenant compte des traits les plus significatifs des échantillons. Le choix de la fonction de distance dans l'objectif d'une bonne généralisation est un problème difficile, pour lequel il n'existe pas à notre connaissance d'algorithme efficace.

Cependant, nous assumons que dans plusieurs problèmes (comme le nôtre), l'ensemble des échantillons se trouvent sur une surface de relativement petite dimension. Ceci permet tout de même d'utiliser des algorithmes à noyaux malgré la très grande dimensionnalité de l'espace des modèles de locuteurs (environ 1000 dimensions).

2.3 Algorithmes génétiques

Les noyaux que nous venons d'exposer comportent une difficulté : la qualité de leur prédiction est déterminée par la largeur du noyau et la fonction de distance utilisée. Il n'existe pas de méthode permettant d'entraîner ces paramètres à l'aide d'un gradient. Nous devons donc explorer l'espace de solution en minimisant la

durée du calcul pour le faire dans un temps raisonnable. Pour optimiser nos noyaux gaussiens, nous avons choisi d'utiliser un algorithme génétique.

2.3.1 Théorie des algorithmes génétiques

Dans un algorithme génétique, les caractéristiques des solutions (appelées *individus*) sont appelés *chromosomes*, par analogie avec l'ADN des êtres vivants. Une fonction d'erreur[4] est nécessaire pour déterminer quelles sont les meilleures solutions pour notre ensemble de données. Une série de *générations* de solutions permettront de faire évoluer notre *population* afin de trouver les meilleurs *individus*.

L'algorithme commence en générant un ensemble de solutions (la *population initiale*) au hasard. La distribution de probabilité initiale doit être choisie judicieusement, car d'elle dépend la convergence de l'algorithme. Les solutions sont évaluées par notre fonction d'erreur, puis les meilleures sont conservées (par exemple, les 50% meilleures). À partir de ces solutions *survivantes*, nous allons engendrer de nouvelles solutions selon deux principes : la *mutation* et l'*accouplement*. La génération de nouveaux individus permet à la population de revenir à sa taille initiale.

Les mutations permettent de faire varier sensiblement les paramètres de nos solutions pour explorer leur voisinage. Nous allons donc créer de nouvelles solutions à partir de légères variations sur les solutions survivantes.

L'accouplement permet de mélanger certaines caractéristiques de deux solutions survivantes pour en créer une nouvelle. Une certaine proportion (par exemple 50-50% ou 40-60%) des caractéristiques des solutions A et B sont prises pour former la nouvelle solution C.

[4]La fonction d'erreur doit normalement être monotone.

Une fois la population renouvellée grâce aux mutations et aux accouplements, l'opération d'élimination recommence. Après plusieurs générations, l'ensemble converge vers un meilleur ensemble de solutions (Duda, Hart et Stork, 2001). Lorsque la meilleure solution nous satisfait, nous pouvons arrêter l'algorithme. Le pseudo-code suivant résume l'algorithme.

0- **Initialisation et sélection** : générer un ensemble de solutions au hasard et conserver les 50% meilleures.

1- **Mutation** : Générer de nouvelles solutions à partir de légères modifications des individus survivants.

2- **Accouplement** : Générer de nouvelles solutions à partir de combinaisons des paramètres des individus survivants.

3- **Sélection** : Conserver seulement les 50% meilleures solutions.

4- **Condition de fin** : Si la meilleure solution est satisfaisante, arrêter l'algorithme. Sinon, continuer en (1).

2.4 Analyse en composantes principales

L'analyse en composantes principales[5] (Hastie, Tibshirani et Friedman, 2001) (Haykin, 1999) permet, grâce à la décomposition en vecteurs propres, de projeter les données sur une nouvelle base d'une dimensionnalité plus petite. Contrairement à d'autres transformées permettant de réduire la dimensionnalité, comme la transformée en cosinus ou la transformée en ondelettes, l'analyse en composantes principales n'a pas de vecteurs de base fixes : ces vecteurs sont calculés à partir des

[5] Aussi appelée transformation de Karhunen-Loève.

données. La nouvelle base est choisie afin de **maximiser la variance des l premières dimensions**, $0 < l < m$. Le changement de base correspond à une rotation de l'espace de m dimensions $\Re^m$ pour remplir cette condition. La rotation optimale est trouvée en faisant la décomposition en vecteurs propres de la matrice d'autocorrélation des échantillons, comme expliqué dans la prochaine section.

2.4.1 Algorithme pour trouver les l composantes principales

Voici le pseudo-code de l'algorithme permettant de trouver les l composantes principales d'une matrice de données $\mathbf{D}$ comportant n vecteurs de m dimensions.

1. Calculer le vecteur μ : les moyennes des m dimensions des données.

2. Soustraire le vecteur μ de chaque vecteur de données pour obtenir la matrice $\mathbf{S}$.

$$\mathbf{S} = \mathbf{D} - \mu$$

3. Construire la matrice de covariance $\mathbf{C}$ des données (matrice carrée de dimension m) :

$$\mathbf{C} = \mathbf{S}\mathbf{S}^T.$$

4. Calculer les vecteurs propres (voir section 2.4.2) de la matrice $\mathbf{C}$ et les classer en ordre décroissant de valeur propre.

5. Retenir les l premiers vecteurs propres dans la matrice de projection $\mathbf{P}$. La matrice $\mathbf{P}$ est de dimension $m \times l$.

2.4.2 Décomposition en vecteurs propres

La décomposition en vecteurs propres a pour but de trouver les vecteurs propres $\mathbf{x}$ et les valeurs propres λ qui satisfont l'équation (Kreyszig, 1993) :

$$\mathbf{Dx} = \lambda\mathbf{x}$$

Cette équation peut également s'écrire sous la forme :

$$(\mathbf{D} - \lambda\mathbf{I})\mathbf{x} = 0$$

Les valeurs propres sont les racines de l'équation caractéristique de la matrice D :

$$det(\mathbf{D} - \lambda\mathbf{I}) = 0$$

Une fois les valeurs propres trouvées, on résout l'équation :

$$(\mathbf{D} - \lambda\mathbf{I})\mathbf{x} = 0$$

pour obtenir les vecteurs propres. Ceux-ci ont une norme égale à 1, ce qui permet d'obtenir une base orthonormale. Il existe plusieurs algorithmes d'algèbre linéaire permettant de résoudre ce système d'équation, et la plupart des bibliothèques de fonctions pour le calcul numérique (comme Matlab, GNU Scientific Library ou LAPACK) fournissent des méthodes pour solutionner un système linéaire de ce genre.

Si tout ce qui nous intéresse est de trouver les composantes, il est plus pratique de prendre une librairie qui calcule les valeurs et vecteurs propres directement à partir de la matrice $\mathbf{C}$, par exemple avec la décomposition en valeurs singulières (SVD).

2.5 Analyse en composantes indépendantes

L'analyse en composantes indépendantes (*ICA*) permet de projeter des données sur une base de dimensionnalité réduite dont les vecteurs de projection sont aussi indépendants que possible au niveau de leur apparition dans les données. Cela permet d'isoler des caractéristiques qui ne sont pas normalement présentes dans un même échantillon. Cette technique est utilisée dans divers domaines pour débruiter des signaux (p.e. pour enlever le bruit atmosphérique des images en astronomie) ou pour séparer n sources indépendantes mélangées dans n signaux mesurés.

Un bon exemple d'application des techniques itératives pour trouver les composantes indépendantes d'un signal est la séparation aveugle de sources (*blind source separation*), pour le problème du *cocktail party*, dans lequel plusieurs sources indépendantes (des voix) sont mesurées par un seul (Jang et Lee, 2002) ou plusieurs (Bell et Sejnowski, 1995) senseurs (des oreilles), et doivent être séparées. Ce problème est omniprésent dans les systèmes de téléconférence.

Dans le cas de plusieurs senseurs, le problème dans sa forme matricielle peut être exprimé comme un mélange linéaire de sources (Haykin, 1999) :

$$\mathbf{Y} = \mathbf{WX} = \mathbf{WAU}$$

où $\mathbf{U}$ est la matrice des vecteurs sources, $\mathbf{A}$ est la matrice de mixage, $\mathbf{X}$ est la matrice des signaux mélangés et $\mathbf{W}$ est la matrice de démixage (dont le but est de ressembler à $\mathbf{A}^{-1}$). Le but est de trouver une matrice de signaux $\mathbf{Y}$ égale à la matrice des vecteurs sources inconnus $\mathbf{U}$.

La figure 2.4 illustre le problème de la séparation aveugle de sources. Il existe plusieurs méthodes itératives pour trouver la matrice $\mathbf{W}$. Nous en présentons

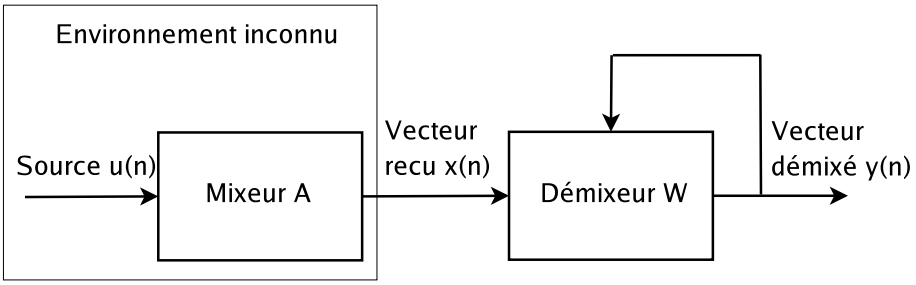


FIG. 2.4 Le problème de séparation aveugle de sources.

deux dans les pages qui suivent, sans cependant nous attarder aux détails de la démonstration qui mène à leur règle d'apprentissage.

2.5.1 Maximum d'entropie

Dans l'article (Bell et Sejnowski, 1995), les auteurs décrivent une méthode de séparation de sources basée sur le principe Infomax (Haykin, 1999). L'objectif est de minimiser la redondance d'information entre les sorties $\mathbf{Y}$ d'un système de démixage. Minimiser la redondance revient à maximiser l'entropie $h(\mathbf{Y})$ des sorties $\mathbf{Y}$. L'entropie $h(X)$ d'une variable aléatoire discrète X est définie comme suit :

$$h(\mathbf{X}) = -\sum_{k=-K}^{K} p_k log(p_k)$$

où p_k est la probabilité que la variable aléatoire X soit égale à la k^{ieme} valeur discrète possible.

La méthode décrite dans l'article de Bell (Bell et Sejnowski, 1995) utilise une non-linéarité pour représenter le problème de la séparation de source :

$$\mathbf{y} = g(\mathbf{W}\mathbf{x} + \mathbf{w}_0)$$

La fonction non-linéaire $g(u)$ est la fonction sigmoïde :

$$g(u) = \frac{1}{1 + e^{-u}}$$

La règle d'apprentissage pour arriver à trouver la matrice de déconvolution $\mathbf{W}$ est donnée par ces équations :

$$\mathbf{W} = \mathbf{W} + \Delta\mathbf{W}$$

où

$$\Delta\mathbf{W} = \eta([\mathbf{W}^T]^{-1} + (1 - 2\mathbf{y})\mathbf{x}^T)$$

$$\Delta\mathbf{w}_0 = \eta(1 - 2\mathbf{y})$$

Ces règles d'apprentissage poussent les vecteurs de $\mathbf{W}$ à apprendre des caractéristiques différentes des vecteurs d'entrée $\mathbf{x}$. L'algorithme converge tant que le déterminant de la matrice $\mathbf{W}$ est plus grand que 0. Un déterminant près de 0, même positif, peut toutefois faire diverger l'algorithme.

2.5.2 Minimiser la divergence de Kullback-Leibler

Dans (Haykin, 1999), on suggère une autre méthode pour calculer la matrice des poids $\mathbf{W}$ en minimisant la divergence de Kullback-Liebler $D_{f||\tilde{f}}(\mathbf{W})$ entre la densité de probabilité de $\mathbf{Y}$ (notée $f_{\mathbf{Y}}(\mathbf{y}, \mathbf{W})$) et la distribution factorielle de $\mathbf{Y}_i$ (notée $\tilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{W})$).

$$D_{f||\tilde{f}}(\mathbf{W}) = -h(\mathbf{Y}) + \sum_{i=1}^{m} \tilde{h}(\mathbf{Y}_i)$$

Minimiser la divergence KL est similaire à maximiser l'entropie $h(\mathbf{Y})$ de la sortie

du système. La règle d'apprentissage de $\mathbf{W}$ est donnée par :

$$\mathbf{W} = \mathbf{W} + \Delta\mathbf{W}$$

où

$$\Delta\mathbf{W} \ = \eta[I - \psi(\mathbf{y})\mathbf{y}^T](\mathbf{W}^T)^{-1}$$

Contrairement à Bell, Haykin n'utilise pas de fonction non-linéaire dans son modèle. Il utilise cependant une fonction d'activation non-linéaire et non-monotone $\psi(y)$ dans l'équation d'apprentissage présentée ci-haut. Cette fonction est décrite dans (Madhuranath et Haykin, 1998) comme étant :

$$\psi(y) = \frac{1}{2}y^5 + \frac{2}{3}y^7 + \frac{15}{2}y^9 + \frac{2}{15}y^{11} - \frac{112}{3}y^{13} + 128y^{15} - \frac{512}{3}y^{17}$$

2.6 Conclusion du chapitre

Ce chapitre nous a permis d'exposer les bases de l'apprentissage machine ainsi que quelques algorithmes permettant d'estimer certaines fonctions, i.e. qui accomplissent la tâche de régression. Ces algorithmes, bien que jouissant d'une justification mathématique solide, ne s'appliquent pas directement en général. Certains ajustements sont nécessaires dépendant du type de données et de la quantité d'échantillons disponibles. Lorsque confrontées à des cas réels, ces techniques ont leurs limites : c'est à ce moment que la théorie fait place à l'ingénierie[6].

[6]On parle parfois du "no free-lunch theorem" : il n'y a pas de "recette magique" ou d'algorithme unique qui permet de résoudre automatiquement tous les problèmes d'apprentissage que nous soumettons à nos machines.

CHAPITRE 3

MÉTHODOLOGIE

Dans les deux premiers chapitres, nous avons d'abord décrit les mécanismes d'audition et de phonation, puis nous avons présenté différents modèles mathématiques permettant d'analyser et de reconstruire le signal vocal. Dans ce chapitre, nous allons détailler notre méthode de modélisation et d'estimation acoustique de locuteurs. Nous décrirons d'abord les données utilisées et notre méthode pour extraire les paramètres d'un locuteur à partir d'exemples de phonèmes, puis nous expliquerons de quelle manière les méthodes du chapitre 2 sont appliquées pour estimer les caractéristiques acoustiques des phonèmes manquants.

3.1 Description des données utilisées

Nous n'avons trouvé aucune base de données publique de voix dans laquelle les voix étaient étiquetées par locuteur. Les bases de données utilisées pour la reconnaissance vocale contiennent des dizaines de locuteurs, mais sont généralement étiquetées par phonème et il est impossible de retrouver les locuteurs originaux à partir des fichiers publiés. Nous avons donc entrepris de construire nous-mêmes une base de données de locuteur pour cinq voyelles en particulier : /a/, /e/, /i/, /o/ et /u/[1]. Ces voyelles ont été utilisées parce qu'elles sont répandues dans un grand nombre de langues européennes.

Les locuteurs, tous masculins, étaient des volontaires du département de génie

[1]En français, ces voyelles sont prononcées “a”, “é”, “i”, “o” et “ou”.

électrique de l'École Polytechnique de Montréal. Nous leur avons demandé de lire un texte de manière intelligible, lentement et en prononçant bien les syllables. Puis, nous leur avons demandé de dire les cinq voyelles étudiées sur trois tons, bas, moyen et haut.

Nous avons ensuite découpé des segments des voyelles qui nous intéressaient dans les extraits sonores des volontaires. Nous avons pris soin de prendre des exemples représentant bien le phonème et d'éviter de choisir les erreurs de prononciation des locuteurs[2]. Nous avions entre 5 et 15 exemples d'une longueur de 50 à 500 ms pour chaque voyelle pour chacun des 31 locuteurs[3]. Ces exemples étaient divisés en unités appelées *fenêtres* pour l'analyse spectrale. Comme l'illustre le tableau 3.1, le nombre de fenêtres pouvait varier beaucoup d'un locuteur à l'autre. Ces fenêtres sont la base de la sélection et de la normalisation des phonèmes des locuteurs expliquées dans la section suivante.

TAB. 3.1 Statistiques sur le nombre de fenêtres d'analyse d'un locuteur pour chaque voyelle.

Voyelle	/a/	/e/	/i/	/o/	/u/
Moyenne de fenêtres d'analyse	47	55	61	64	62
Minimum de fenêtres d'analyse	6	19	17	22	23
Maximum de fenêtres d'analyse	92	134	121	138	161

[2]Par exemple, certains québécois prononcent la voyelle "é" comme la voyelle "è" dans certaines situations. Le mot *québécois* est alors prononcé "quèbècois". Nous avons rejeté ce "è" lorsqu'il était présent dans l'extrait d'un locuteur.

[3]En comparaison, la base de données commerciale *Tactical Speaker Identification Speech Corpus* (TSID) disponible sur le site http ://www.ldc.upenn.edu contient 35 locuteurs, soit 4 femmes et 31 hommes.

3.2 Normalisation des phonèmes des locuteurs

Pour appliquer les algorithmes d'apprentissage machine présentés au dernier chapitre, les données audio des différents locuteurs doivent être normalisées afin de localiser les caractéristiques acoustiques de chaque individu dans un espace mathématique commun. Dans cette section, nous expliquons la méthode que nous avons appliquée pour obtenir les paramètres caractéristiques de chaque locuteur.

Nous commençons en décrivant brièvement l'extraction des paramètres pour chaque fenêtre d'analyse des fichiers audio. Toutes ces fenêtres ne sont pas nécessairement caractéristiques d'un locuteur, c'est pourquoi nous présentons ensuite une méthode d'élimination des enveloppes atypiques. Les enveloppes restantes serviront à calculer l'enveloppe de puissance caractéristique d'un locuteur pour une voyelle.

3.2.1 Extraction des paramètres

La première étape consiste à trouver des fenêtres valides d'analyse dans les extraits des locuteurs. Une fenêtre dont la puissance est trop faible ou contenant un signal transitoire superposé à la voyelle n'est pas souhaitable.

Les fenêtres ont donc été centrées sur des maxima d'énergie du signal temporel tel que vu sur les figures 1.15 et 1.16. Les fenêtres moins puissantes que la moyenne de l'extrait analysé n'ont pas été conservées. Nous avons aussi rejeté les fenêtres dont la ratio d'harmonicité ne dépassait pas 95%. Cette première élimination nous a garanti que les fenêtres restantes représentaient un moment où la voyelle était clairement définie.

Les paramètres extraits de ces fenêtres d'analyse sont les paramètres du modèle harmonique et résiduel tels que décrits dans la section 1.3.3. Seuls les sommets

du spectre harmonique correspondant aux fréquences multiples de la fréquence fondamentale sont mesurés.

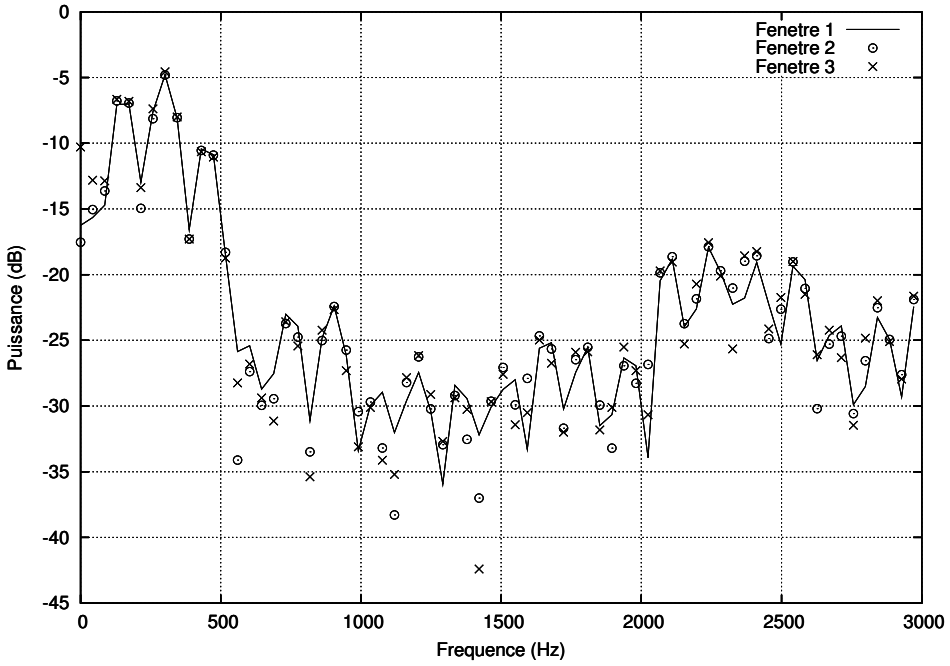


FIG. 3.1 Spectrogramme montrant trois fenêtres consécutives d'une voyelle /i/.

La figure 3.1 montre que les enveloppes varient d'une fenêtre d'analyse à l'autre. La séquence d'élimination et de normalisation expliquée dans les sections suivantes définit l'élimination des fenêtres atypiques et l'utilisation des variations entre les différentes fenêtres pour construire un modèle moyen.

3.2.2 Élimination des enveloppes atypiques

Après la première élimination, il reste des fenêtres dont l'enveloppe de puissance harmonique est loin de la voyelle caractéristique d'un locuteur. Certains formants peuvent apparaître ou être effacés selon le contexte où la voyelle est prononcée ou simplement à cause de phénomènes physiques altérant des périodes de signal de manière imprévisible. Comme ce qui nous intéresse est une enveloppe de puissance

typique d'une voyelle d'un locuteur, nous avons procédé à une seconde élimination en utilisant un algorithme dérivé de l'*algorithme des k-moyennes* (Haykin, 1999) (Hastie, Tibshirani et Friedman, 2001).

L'idée est de classer les fenêtres des cinq voyelles d'un locuteur à l'aide des moyennes de chaque classe. Certaines fenêtres sont plus près de la moyenne d'une autre classe que de la leur. Nous éliminons progressivement les fenêtres mal classées, pour terminer avec cinq ensembles de fenêtres distincts dans l'espace des caractéristiques spectrales.

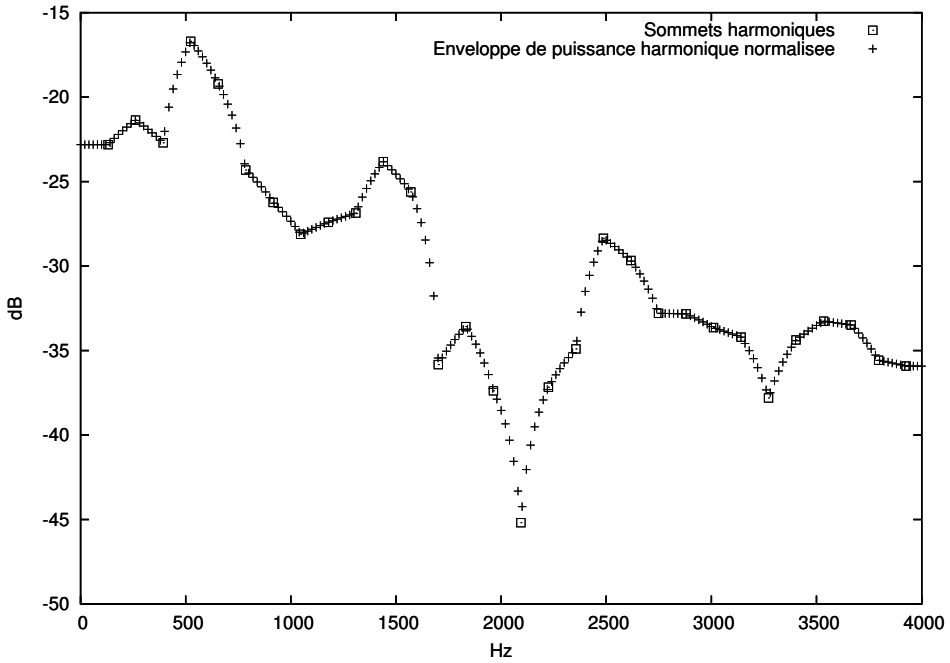


FIG. 3.2 Enveloppe spectrale normalisée des harmoniques d'une fenêtre par une approximation linéaire à chaque 20 Hz.

Mais les fenêtres n'ont pas toutes la même fréquence fondamentale. Pour comparer leurs caractéristiques spectrales, nous devons d'abord normaliser leurs données. Nous avons donc traduit les paramètres harmoniques en 200 points à intervalle de 20 Hz (20 à 4000Hz) avec une interpolation linéaire, tel que suggéré dans (Ramsay et Silverman, 1997). Un exemple de cette normalisation est illustré dans la

figure 3.2. Sous cette forme, il est possible de comparer les spectres des enveloppes harmoniques en utilisant la distance euclidienne.

Voici le pseudo-code de notre algorithme d'élimination des enveloppes atypiques :

1- Calculer les moyennes de chacun des 5 groupes.

2- Calculer les distances entre le point et sa moyenne.

3- S'il reste au moins un échantillon mal classé, éliminer le plus loin de sa moyenne et recommencer à 1.

4- Si toutes les fenêtres sont bien classées, FIN.

Les paramètres harmoniques des enveloppes qui passent ce test sont utilisés dans la suite de la normalisation des voyelles des locuteurs.

3.2.3 Calcul de l'enveloppe typique moyenne

L'enveloppe typique moyenne pour une voyelle d'un locuteur est calculée en utilisant les fenêtres d'analyse qui n'ont pas été éliminées par l'étape précédente. Les données utilisées sont les données brutes de l'analyse harmonique et résiduelle, et non pas l'enveloppe approximée par intervalle de 20 Hz. Nous allons calculer l'enveloppe typique moyenne en faisant une moyenne pondérée du voisinage de chaque fréquence, mais avant nous devons expliquer les raisons de ce choix au lecteur.

3.2.3.1 La fréquence fondamentale et le problème d'estimation des formants

Le calcul de la puissance de chaque harmonique h_k décrit dans la section 1.3.3 nous permet d'avoir une estimation de l'enveloppe harmonique de puissance aux fréquences $k * f0$ des harmoniques h_k. Cependant, la décomposition harmonique et résiduelle ne nous permet pas d'avoir de mesure de l'enveloppe de puissance harmonique entre deux harmoniques h_k et h_{k+1} avec une seule fenêtre d'analyse. Cela peut poser un problème lorsque la fréquence fondamentale est élevée et que l'essentiel de l'information d'un formant se trouve entre deux harmoniques (par exemple lorsqu'ils sont séparés par plus de 150 Hz).

Pour éviter de faire une interpolation inappropriée entre deux points d'estimation de l'enveloppe de puissance harmonique, nous utilisons plusieurs fenêtres d'analyse pour construire un spectre typique pour un locuteur. Cela nous permet de tenir compte des sommets harmoniques de toutes les fenêtres en même temps et d'éviter d'attribuer une valeur trop faible ou trop élevée à une fréquence pour laquelle une fenêtre n'a pas d'information. Cette méthode nous a été inspirée par (Ramsay et Silverman, 1997) et (Shiga et King, 2003a).

3.2.3.2 Lissage de l'enveloppe typique avec un noyau

On obtient une courbe lisse en faisant une moyenne pondérée des points mesurés (f_i, y_i) autour d'une fréquence f_j avec une fonction de distance gausienne K_λ.

$$\hat{y}_j(f_j) = \frac{\sum_i K_\lambda(f_j, f_i) y_i}{\sum_i K_\lambda(f_j, f_i)}$$

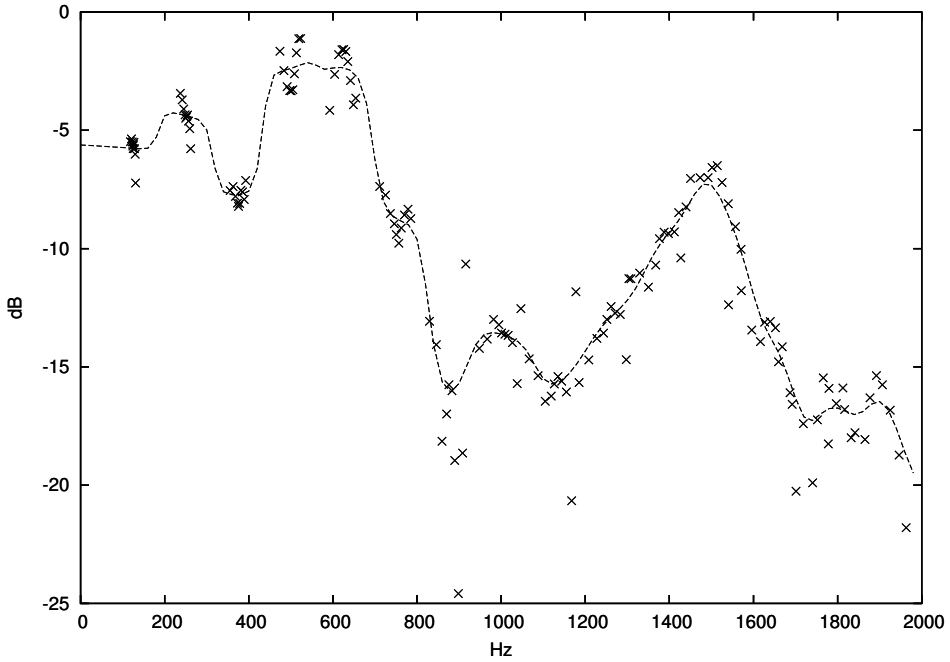


FIG. 3.3 Détail des basses fréquences ($f < 2$ kHz) de l'enveloppe typique d'une voyelle /a/ d'un locuteur construite à l'aide des points des fenêtres valides avec une approximation par noyau gaussien à chaque 20 Hz.

La largeur du noyau λ choisi est de 900, ce qui nous permet d'avoir une estimation très locale, de l'ordre de l'espace entre deux harmoniques d'une voyelle normalement prononcée par un locuteur mâle (écart-type de 30 Hz). Dans (Kleijn et Paliwal, 1995), on appelle l'enveloppe résultante l'enveloppe SEEVOC. Cette enveloppe est une bonne estimation de l'enveloppe typique moyenne du locuteur pour une voyelle donnée.

La figure 3.3 montre un exemple d'un tel lissage par noyau gaussien. Les points représentent les sommets harmoniques calculés dans les fenêtres qui ont survécu à l'étape de l'élimination des enveloppes atypiques. La courbe représente l'approximation qui servira comme donnée dans notre expérience de prédiction d'enveloppe de puissance harmonique de voyelles inconnues. En comparant les figures 3.2 et 3.3, qui proviennent d'une voyelle /a/ d'un même locuteur, nous pouvons observer

que l'enveloppe obtenue à partir de plusieurs fenêtres de la figure 3.3 permet de modéliser une caractéristique autour de 1000 Hz que l'enveloppe de la figure 3.2 ne permet pas de voir.

3.3 Estimation des phonèmes manquants

3.3.1 Estimer une voyelle inconnue d'un locuteur

L'objectif de notre recherche est d'estimer l'enveloppe de puissance harmonique (telle que définie dans la section 3.2) d'une voyelle inconnue d'un locuteur. À l'aide de quatre voyelles connues de l'ensemble /a/, /e/, /i/, /o/, /u/ d'une personne et des données d'autres locuteurs, nous voulons déterminer les caractéristiques acoustiques inconnues de la cinquième voyelle. Dans cette section, nous décrivons la méthode utilisée pour estimer l'enveloppe de la voyelle inconnue avec trois méthodes : les noyaux, les composantes principales et les composantes indépendantes. Nous décrirons ensuite notre méthode d'évaluation des performances de ces différents algorithmes.

3.3.1.1 Estimation avec les noyaux

Nous avons utilisé deux types de noyaux pour l'expérience de prédiction de voyelles : les k-plus proches voisins (KNN) et la moyenne pondérée de Nadaraya-Watson utilisant un noyau gaussien. La fonction de distance utilisée est la distance euclidienne des spectres de puissance. Les résultats de (Stylianou et Syrdal, 2001) sur l'évaluation de la mesure de distance psycho-acoustique faite par un groupe d'auditeurs montrent que la distance euclidienne sur le spectre de puissance donne de bons résultats pour estimer la perception. Seules la distance de Kullback-Liebler

sur ce même spectre et la distance euclidienne sur les coefficients cepstraux donnaient des résultats semblables. Nous avons donc préféré nous en tenir à cette fonction intuitive, qui représente par ailleurs la somme des différences d'énergie de deux enveloppes. Cette mesure de distance étant fixe, notre travail consistait donc à trouver une largeur de noyau idéale pour notre ensemble de données.

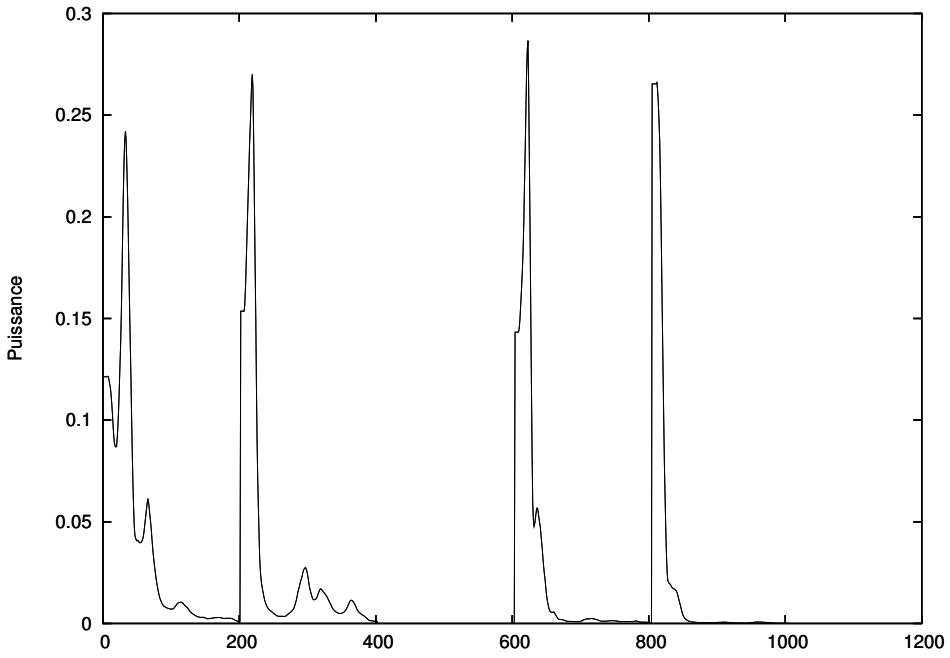


FIG. 3.4 Vecteur s_i contenant les enveloppes de puissance harmoniques des voyelles connues /a/, /e/, /o/ et /u/ d'un locuteur.

– Pour les KNN, nous avons fait varier le nombre k de voisins pour trouver le voisinage optimal pour la prédiction.
– Pour les noyaux gaussiens, nous avons fait varier la largeur λ du noyau K_λ.

$$K_\lambda(s_j, s_i) = \frac{1}{\lambda} exp[\frac{-||s_i - s_j||^2}{2\lambda}]$$

Dans cette équation, les vecteurs s_i contiennent les valeurs des spectres des voyelles connues dont la puissance est normalisée pour être égale à 1 (voir figure 3.4).

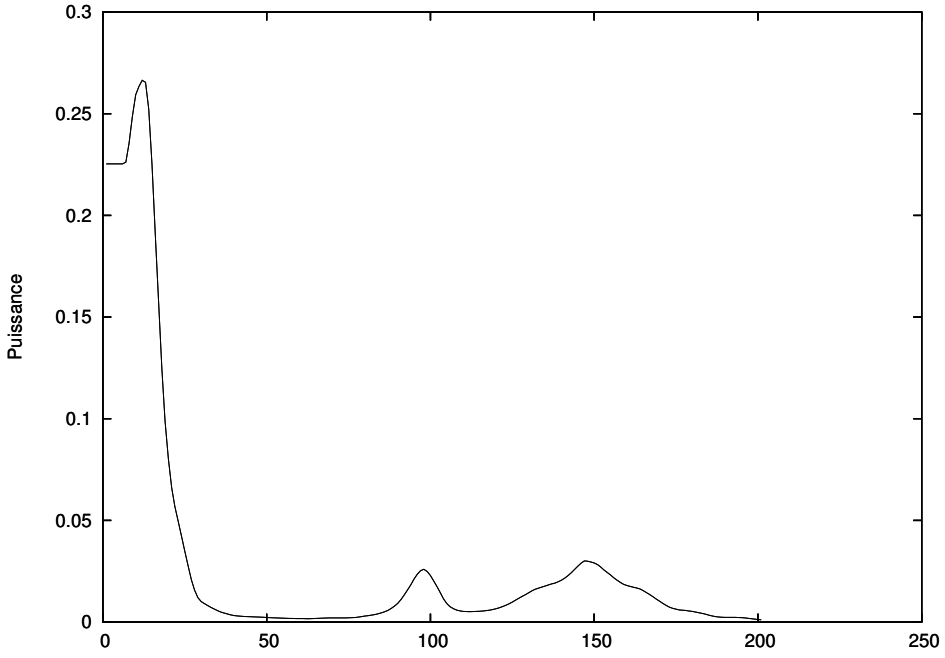


FIG. 3.5 Vecteur y contenant l'enveloppe de puissance harmonique d'une voyelle /i/ inconnue.

Nous l'avons d'abord essayé pour un seul paramètre λ, et nous avons ensuite essayé des λ_i différents pour chaque voyelle connue afin de tenir compte des relations particulières entre les caractéristiques acoustiques de chaque voyelle. La matrice Λ des $\lambda_{i,j}$, que nous avons optimisée avec un algorithme génétique (voir section 3.4), était donc de taille 5×5, les éléments de la diagonale étant nuls (voir tableau 4.2).

3.3.1.2 Estimation avec les composantes principales

Pour estimer la voyelle inconnue d'un locuteur avec les composantes principales, il suffit de projeter les voyelles connues du nouveau locuteur sur les composantes principales de notre base de données. Nous obtenons la voyelle estimée en faisant une moyenne pondérée de cette voyelle pour les composantes principales. L'estimation

se fait donc de cette manière :

$$\hat{y}_j(s_j) = \frac{\sum_i (p_i^T (s_j - \mu)) y_i}{\sum_i (p_i^T (s_j - \mu))}$$

Cette équation permet de normaliser la projection du vecteur des spectres harmoniques des voyelles connues s_i sur les composantes principales p_i. y_i représente ici la portion de la composante principale p_i contenant le spectre harmonique de la voyelle cible et $\hat{y}_j$ représente la voyelle manquante estimée (voir figure 3.5).

3.3.1.3 Estimation avec les composantes indépendantes

L'estimation avec les composantes indépendantes se fait de la même manière que l'estimation avec les composantes principales. Nous avons cependant fait l'hypothèse que les composantes indépendantes nous permettraient de mieux capturer les caractéristiques des voyelles, car elles représentent des vecteurs ayant une information aussi indépendante que possible les uns des autres au niveau de leur apparition dans les données (l'ICA minimise l'information mutuelle).

Dans notre recherche, nous avons utilisé la méthode de (Haykin, 1999) basée sur la divergence de Kullback-Liebler pour trouver les composantes indépendantes de la matrice W, tel que décrit dans la section 2.5.2.

3.3.2 Validation croisée

Étant donné la faible taille de notre base de données (31 locuteurs), nous avons utilisé la validation croisée afin de mesurer la capacité de généralisation des différents estimateurs étudiés dans cette recherche. Nous avons calculé la capacité d'un estimateur à prédire l'enveloppe spectrale en retirant un locuteur de la base de données

et en se servant des autres pour mesurer l'erreur. Cette méthode est communément appelée *leave one out cross-validation*.

3.3.3 Mesure d'erreur

La mesure d'erreur que nous avons utilisée est la distance euclidienne (*2-distance*) entre les enveloppes mesurées et prédites. Cette distance correspond à celle utilisée pour les prédicteurs à noyaux, ce qui permet de garder les concepts de proximité et de mesure d'erreur équivalents pour la comparaison d'échantillons. La formule suivante rappelle la définition de la distorsion harmonique entre les points k des enveloppes mesurée y_j et prédite $\hat{y}_j(s_j)$:

$$Distorsion = \sum_k (y_{jk} - \hat{y}_{jk}(s_j))^2$$

3.3.3.1 Notes sur la distorsion harmonique et le rapport signal sur bruit

La distorsion harmonique, laquelle est notre mesure de distance et notre mesure d'erreur, est en fait l'inverse d'un rapport signal sur bruit, où le signal serait le spectre de la voyelle prédite et le bruit la différence d'énergie entre le spectre prédit et le spectre mesuré. En effet, ce rapport signal sur bruit (SNR), peut être exprimé ainsi (en dB) :

$$SNR = 10log_{10}(\frac{\sum_k y_{jk}^2}{\sum_k (y_{jk} - \hat{y}_{jk}(s_j))^2})$$

où $\hat{y}_{jk}(s_j)$ sont les points du spectre de la voyelle prédite et les y_{jk} ceux de la voyelle mesurée. La mesure que nous appelons distorsion harmonique est simplement le

même rapport inversé :

$$Distorsion = 10log_{10}(\frac{\sum_k (y_{jk} - \hat{y}_{jk}(s_j))^2}{\sum_k y_{jk}^2})$$

Car $\sum_k y_{jk}^2$ est dans notre cas normalisé à 1. Minimiser la distorsion harmonique équivaut donc à maximiser le rapport signal sur bruit de la voyelle inconnue, i.e. maximiser notre information sur cette voyelle manquante.

3.4 Optimisation de la fonction de distance avec un algorithme génétique

L'algorithme génétique conçu pour optimiser la matrice de relations entre les voyelles utilisées par les prédicteurs à noyaux gaussiens est décrit dans la section 2.3. Les paramètres de l'algorithme étaient les suivants :

- La population, composée de 20 individus, était constante d'une génération à l'autre.
- Les invididus initiaux étaient composés de paramètres suivant des lois gaussiennes indépendantes de moyenne 0,05 et d'écart-type 0,05[4]. Les paramètres étaient contraints par une borne inférieure de 0,0001.
- Le taux de survie était de 50 % à chaque génération.
- Pour les nouveaux individus, le taux d'individus mutants était de 50 %. L'autre 50 % était composé d'individus hybrides.
- Les mutations suivaient une loi gaussienne d'écart-type égale à 0,01, et les mutations des différents paramètres étaient indépendantes les unes des autres. Les paramètres étaient également contraints par une borne inférieure de 0,0001.
- Les individus hybrides étaient composés de 10 paramètres d'un individu survivant et de 10 autres paramètres d'un autre individu survivant pour avoir un ensemble

[4]Cette distribution a été choisie à la lumière des résultats des estimateurs à noyaux gaussiens symétriques pour toutes les voyelles.

complet de 20 paramètres. Les combinaisons de paramètres venant d'un individu ou d'un autre variaient d'un individu hybride à un autre.

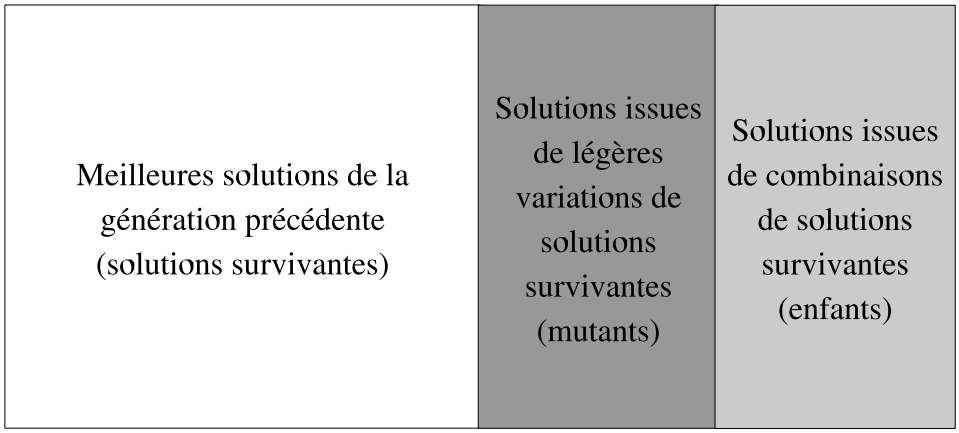


FIG. 3.6 Composition de la population d'une nouvelle génération de solutions.

CHAPITRE 4

RÉSULTATS

Dans le chapitre précédent, nous avons détaillé les méthodes menant à la prédiction du spectre de puissance harmonique des voyelles manquantes du corpus d'un locuteur. Nous avons appliqué ces techniques à notre base de données de locuteurs mâles. Ce chapitre présente et analyse les résultats de la prédiction faites avec les différents noyaux, les composantes principales et les composantes indépendantes.

4.1 Noyaux

Le tableau 4.1 résume les résultats obtenus avec les différentes méthodes à noyau présentées dans le chapitre 3. Nous y comparons les distorsions moyennes de l'enveloppe de puissance harmonique prédites par rapport à l'enveloppe mesurée. Les méthodes sont présentées en ordre croissant de performance, de la voyelle moyenne jusqu'aux noyaux dont la largeur a été optimisée par un algorithme génétique. L'écart-type est exprimé en pourcentage d'énergie mal placée, donc il suit le même calcul que le pourcentage de distorsion. Le gain est défini comme la réduction de la distorsion harmonique par rapport à la méthode de la voyelle moyenne. Ces résultats ont été publiés précédemment dans (Fournier et Brault, 2005).

La figure 4.1 montre une voyelle prédite avec une distorsion harmonique de -26 dB. Nous pouvons voir que les formants de l'enveloppe de puissance harmonique prédite sont correctement positionnés, et que leur puissance est bien estimée.

Il est important de noter qu'une erreur de 2 dB à -20 dB est beaucoup moins

TAB. 4.1 Distorsion harmonique moyenne avec chaque méthode

Noyau	Distorsion	Distorsion	Écart-type	Gain
Voyelle moyenne	-9.58 dB	11.02 %	7.81 %	N/A
3-NN	-10.74 dB	8.43 %	7.37 %	23.5 %
Gaussien ($\lambda = 0.1$)	-10.91 dB	8.11 %	7.20 %	26.4 %
Gaussien ($\lambda = 0.06$)	-11.00 dB	7.95 %	7.69 %	27.9 %
Gaussien (meilleur λ par voyelle)	-11.41 dB	7.72 %	7.47 %	34.4 %
Gaussien (Algorithme génétique)	-11.56 dB	6.98 %	6.77 %	36.6 %

importante qu'une erreur de 2 dB à -10 dB. De plus, la distorsion de quantification de l'encodage μ-law dans les systèmes téléphoniques entraîne une distorsion de l'ordre de -38 dB[1]. La voyelle présentée ici est donc pratiquement indistingable de la vraie voyelle du locuteur, relativement à la distorsion normale lors d'une conversation téléphonique.

La colonne *écart-type* du tableau 4.1 montre que les résultats de prédiction varient énormément d'un locuteur à l'autre. Nous analysons ce phénomène dans la section 4.1.1. Ces résultats varient aussi d'une voyelle à l'autre. Nous discutons de ces écarts dans la section 4.1.3.

4.1.1 Effet de la taille du voisinage

Nos expériences montrent que l'effet de la taille du voisinage λ sur la qualité de la voyelle à prédire varie significativement selon la voyelle. Sur les figures 4.2 et 4.3, nous observons que le voisinage optimal pour prédire la voyelle /o/ (2 voisins ou $\lambda = 0.07$) est plus petit que le voisinage optimal pour prédire la voyelle /i/ (3 voisins

[1]La distorsion de quantification n'inclut pas les bruits causés par le combiné téléphonique ou la ligne elle-même. La distorsion est donc plus élevée que -38 dB en pratique.

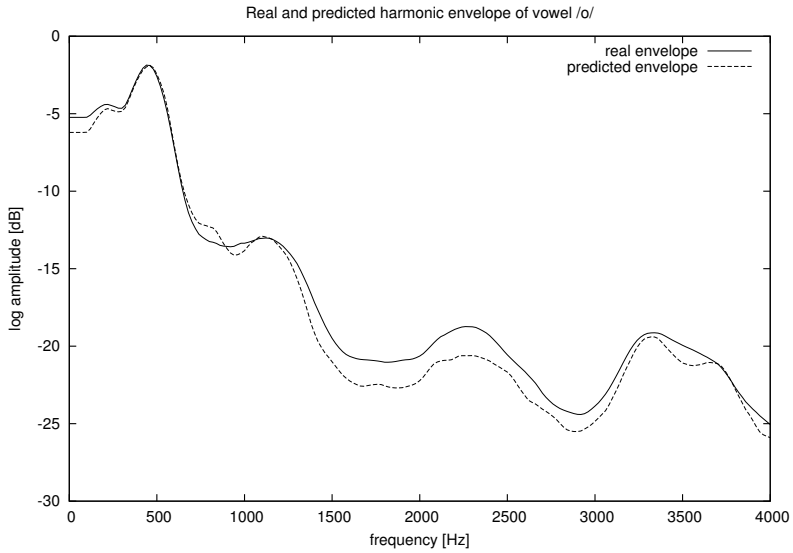


FIG. 4.1 Enveloppes de puissance harmonique réelle et estimée pour une voyelle /o/.

ou $\lambda = 0.09$). La voyelle /a/, avec ses voisinages optimaux $k = 3$ pour l'algorithme KNN et $\lambda = 0.04$ pour le noyau gaussien contredit cependant l'hypothèse que les voisinages optimaux suivent la même relation d'un algorithme à l'autre pour toutes les voyelles. Une voyelle dont le voisinage optimal avec l'algorithme KNN est plus grand que pour une autre voyelle n'aura pas nécessairement un voisinage optimal plus grand avec un noyau gaussien. Nous pouvons donc seulement dire que les voyelles peuvent profiter d'un voisinage sur mesure pour chacune d'elle, dépendamment de l'algorithme utilisé.

Les figures 4.2 et 4.3 soulignent aussi que l'estimateur KNN n'a pas une fonction d'erreur convexe en fonction du nombre de voisins. Contrairement aux noyaux gaussiens, la fonction d'erreur de l'estimateur KNN a plusieurs minima locaux. Cette observation souligne que l'algorithme KNN est un estimateur plus bruité que le noyau gaussien. Cela est dû au fait que l'estimateur KNN a une surface formée de discontinuitées aux points où un échantillon devient soudainement plus près qu'un autre qui était inclus dans le voisinage. Le noyau gaussien pondère ces échantillons

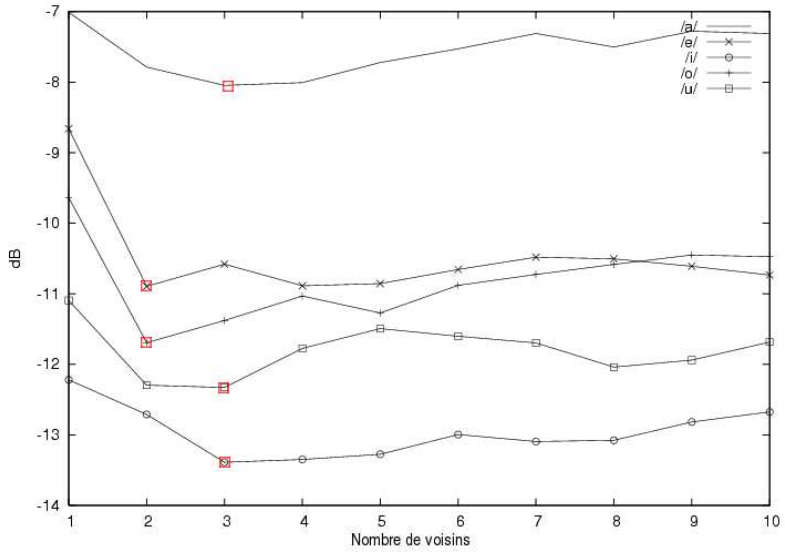


FIG. 4.2 Distorsion de l'enveloppe harmonique moyenne pour chaque voyelle pour l'algorithme des k plus proches voisins (KNN) en fonction de k.

selon leur distance, ce qui donne une surface lisse, et en fait un estimateur plus robuste.

Outre la comparaison des voisinages optimaux, nous pouvons également observer sur les figures 4.2 et 4.3 que certaines voyelles sont plus faciles à prédire que d'autres en moyenne. Il y a environ 5 dB d'écart entre la prédiction moyenne des voyelles /a/ et /i/, donc 70 % moins d'erreur dans le cas de la voyelle /i/. Il semble que les formants des /a/ soient plus difficiles à prédire à l'aide des autres voyelles. Nous discuterons de ce problème dans la section 4.1.2 sur l'information mutuelle entre les voyelles ci-bas.

4.1.2 Notes sur l'information mutuelle entre les voyelles

Comme on pourrait s'y attendre, certaines voyelles sont plus utiles que d'autres pour prédire une voyelle manquante. Nos résultats sur l'optimisation de la largeur de noyaux asymétriques avec un algorithme génétique montrent que les voyelles

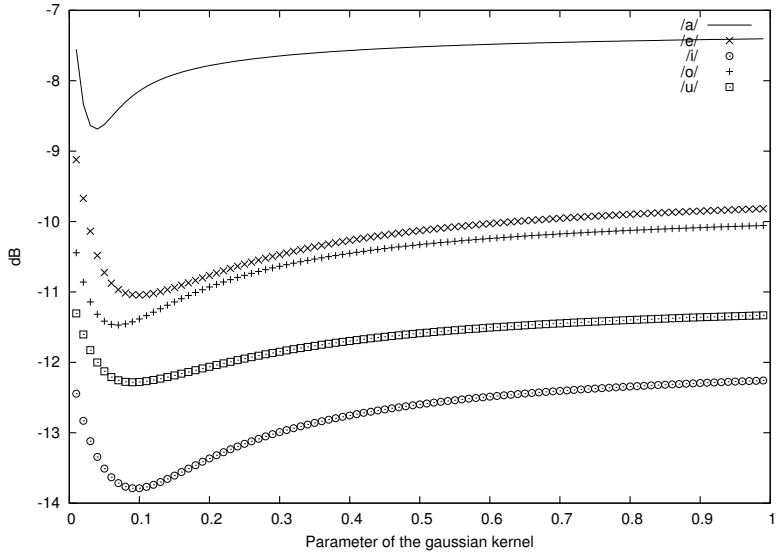


FIG. 4.3 Distorsion de l'enveloppe harmonique moyenne pour chaque voyelle pour différentes largeurs λ de noyau gaussien.

connues donnent parfois très peu d'information pour prédire d'autres voyelles.

Le but de l'optimisation était de trouver 4 paramètres pour la largeur d'un noyau gaussien, soit un par voyelle disponible. Ces paramètres permettent de pondérer l'importance relative des voyelles connues dans la tâche de prédiction. La formule de Nadaraya-Watson est donc réécrite avec un noyau défini par un vecteur de paramètres Λ :

$$\hat{y}_j(s_j) = \frac{\sum_i K_\Lambda(s_j, s_i) y_i}{\sum_i K_\Lambda(s_j, s_i)}$$

où le noyau K_Λ est défini comme :

$$K_\Lambda(s_j, s_i) = \sum_v \frac{1}{\lambda_v} exp[\frac{-||s_{iv} - s_{jv}||^2}{2\lambda_v}]$$

Les spectres des voyelles v des locuteurs s_i et s_j sont notées s_{iv} et s_{jv}.

Le tableau 4.2 présente les résultats de l'optimisation du vecteur de paramètres Λ. Les valeurs de la colonne v représentent les valeurs de λ_v pour chaque voyelle à

prédire de la colonne de gauche. Plus les valeurs sont élevées, plus le noyau est large pour la voyelle v. Un paramètre relativement plus étroit signifie une information moins grande contenue dans la voyelle v pour prédire la voyelle de la colonne de gauche.

TAB. 4.2 Coefficients λ_v des noyaux gaussiens asymétriques pour chaque voyelle à prédire (vecteur Λ).

Voyelle à prédire	λ_a	λ_e	λ_i	λ_o	λ_u
/a/	N/A	0.0122	0.0001	0.0001	0.0001
/e/	0.0256	N/A	0.0001	0.0087	0.0001
/i/	0.0340	0.0130	N/A	0.0018	0.0010
/o/	0.0169	0.0218	0.0001	N/A	0.0010
/u/	0.0001	0.0307	0.0001	0.0241	N/A

Ce tableau montre les disparités de l'information mutuelle entre les voyelles de notre base de données. Par exemple, seul le /e/ semble donner de l'information pour prédire la voyelle /a/. Les autres valeurs λ_v de la première rangée ont été limitées par la borne inférieure fixée par notre algorithme génétique (0.0001). Le noyau asymétrique optimal pour prédire le /a/ considère donc tous les locuteurs très éloignés pour les voyelles /i/, /o/ et /u/, et ne considère que la voyelle /e/ comme mesure de similarité pour prédire la voyelle /a/.

Nous semblons retrouver dans les valeurs du tableau 4.2 l'ordre des erreurs sur la prédiction des voyelles du graphique 4.3. Dans ce dernier graphique, les voyelles les plus faciles à prédire étaient dans l'ordre le /i/, le /u/, le /o/ et le /e/ ; la voyelle /a/ étant le plus sujet à une grande erreur. Dans le tableau 4.2, les voyelles bénéficiant de plus d'information pour la prédiction semblent être dans l'ordre le /i/, le /u/, le /o/, le /e/ et finalement le /a/. En effet, seul le /e/ paraît être utile pour prédire le /a/, alors que toutes les voyelles jouent un rôle pour prédire

le /i/. Les voyelles /u/, /o/ et /e/ ont progressivement des valeurs de λ_v moins importantes, avec deux valeurs touchant la borne inférieure de 0.0001 pour le /e/.

La figure 4.4 présente cette intuition de manière quantitative. Nous pouvons y voir que l'erreur de prédiction suit une loi[2] à peu près linéaire en fonction de la somme des éléments λ_v du vecteur Λ. Les valeurs relatives des vecteurs Λ semblent donc nous donner une mesure de l'information qu'il est possible d'obtenir sur la voyelle à prédire à partir des voyelles connues des locuteurs.

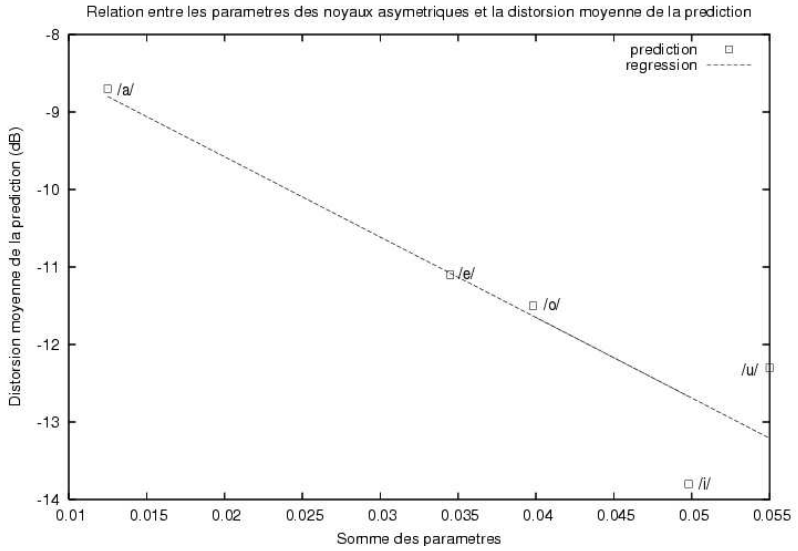


FIG. 4.4 Relation linéaire entre la somme des paramètres et la distorsion moyenne de la prédiction pour les cinq voyelles étudiées.

Il serait intéressant de valider cette théorie sur un ensemble de données avec plus de locuteurs et plus de phonèmes afin de vérifier si la linéarité de cette relation tient toujours. Cinq points ne nous permettent pas avec certitude d'établir une règle, mais nous pensons que la piste est intéressante.

[2]Nous avons comparé différentes relations, soit l'erreur de prédiction en fonction de la somme des carrés des paramètres, l'erreur de prédiction en fonction de la variance ou de l'écart-type des paramètres et l'erreur de prédiction en fonction de la somme des logarithmes des paramètres. L'erreur de prédiction en fonction de la somme des paramètres, en plus d'être une relation très simple, est celle qui minimise l'erreur quadratique dans notre cas.

4.1.3 Distribution des résultats

Pour un même fonction de voisinage K_λ et une même voyelle, les résultats de la prédiction varient énormément d'un locuteur à l'autre. Les histogrammes des figures 4.5 à 4.9 nous montrent les distributions des erreurs de prédiction pour des noyaux gaussiens symétriques optimaux.

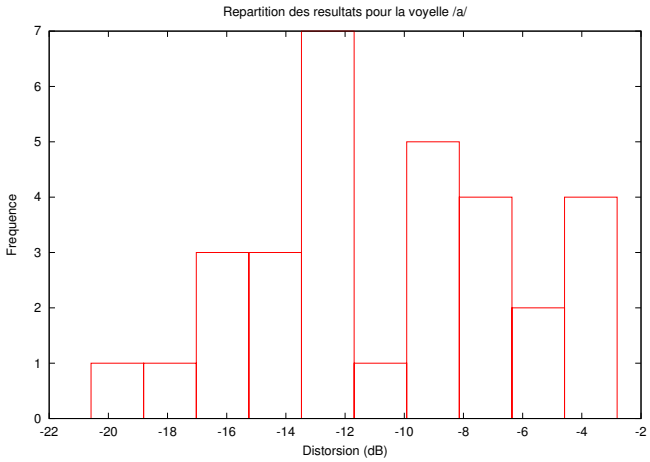


FIG. 4.5 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.04$ pour la voyelle /a/.

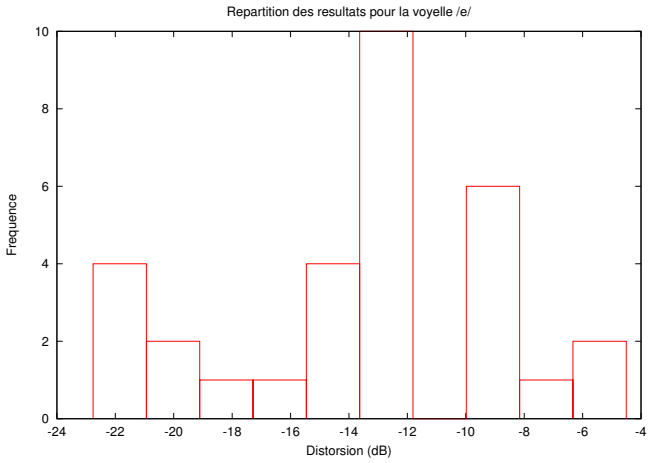


FIG. 4.6 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.10$ pour la voyelle /e/.

L'écart le plus grand entre la plus petite et la plus grande erreur de prédiction est

de -19 dB pour la voyelle /u/. L'écart le plus petit est de -7 dB pour la voyelle /i/. Il ne semble pas y avoir de lien entre la largeur de la distribution de l'erreur et sa moyenne.

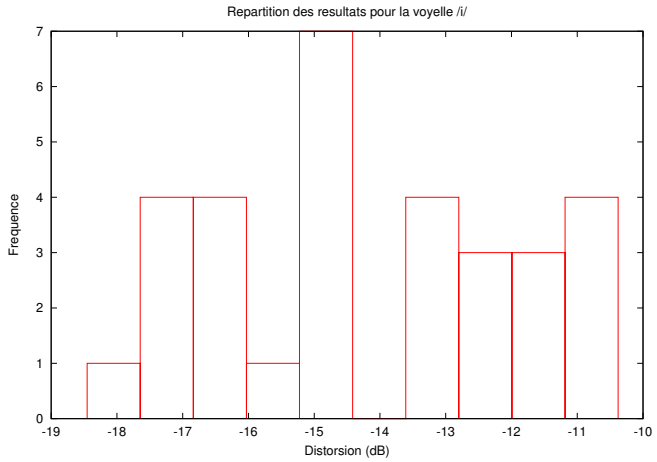


FIG. 4.7 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.09$ pour la voyelle /i/.

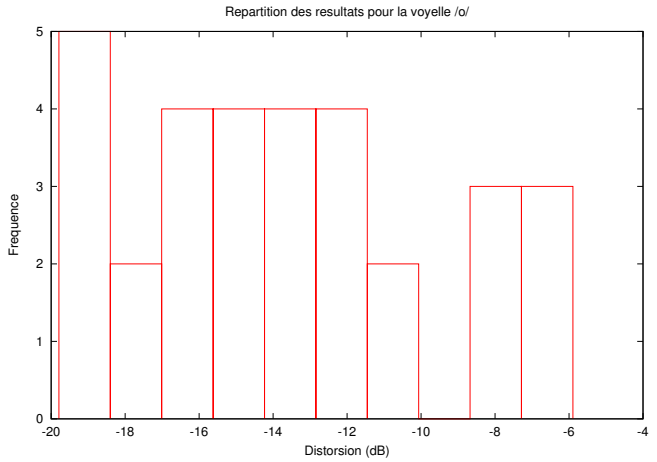


FIG. 4.8 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.07$ pour la voyelle /o/.

La forme même des distributions varie significativement. Sur la figure 4.8, on observe que la distribution est à peu près uniforme entre -6 et -20 dB pour la voyelle /o/, tandis que la distribution des résultats pour les voyelles /a/ et /e/ est concentrée autour de -9 à -13 dB (figures 4.5 et 4.6). Toutes les voyelles sauf /o/

ont un mode situé environ au milieu de leur distribution.

Sur les figures 4.6 et 4.7, on observe que la distribution des résultats pour les voyelles /e/ et /i/ est irrégulière, avec des modes autour de -13 dB pour /e/ et -15 dB pour /i/. Cet étrange résultat est probablement dû à notre ensemble de données. Nous anticipions une distribution plus lisse pour un ensemble de locuteurs plus grand.

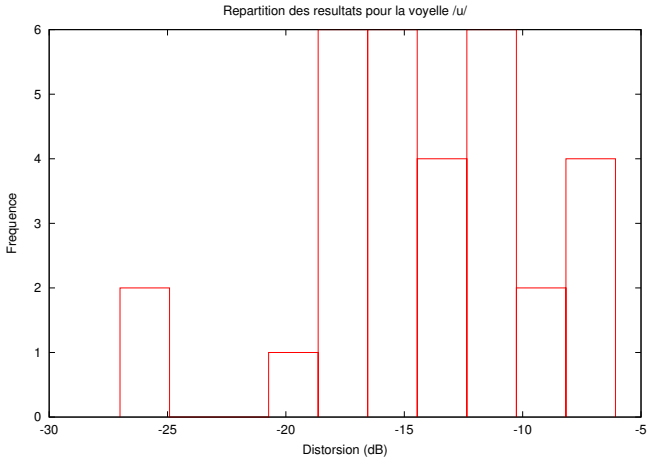


FIG. 4.9 Distribution des résultats de prédiction pour un noyau gaussien de paramètre $\lambda = 0.09$ pour la voyelle /u/.

Les distorsions plus élevées que -7 dB sont rares, sauf pour la voyelle /a/ (30 % des cas, voir figure 4.5). Pour la voyelle /i/ (figure 4.7), la pire erreur est d'enrivon -11 dB. Une erreur plus petite que -20 dB est exceptionnelle, et est habituellement due à la présence de deux locuteurs dont les enveloppes de puissance harmonique sont très similaires pour certaines voyelles. Nous appelons ces locuteurs similaires des *frères acoustiques*[3]. Une base de données idéale devrait contenir assez de locuteurs pour pouvoir fournir des frères acoustiques à la plupart des locuteurs d'une population.

Il serait intéressant de développer une méthode pour estimer l'erreur de la prédiction afin de signaler au système qu'une voyelle prédite n'est pas de bonne qualité. Cela

[3]Nous pourrions peut-être suggérer le terme *acoustical siblings* en anglais.

permettrait à l'usager d'un système de synthèse de voix automatique de prendre les mesures nécessaires pour ajuster manuellement, s'il le faut, les paramètres du phonème. Cela pourrait aussi servir à évaluer le genre de locuteur à rechercher pour compléter l'ensemble de voix d'une base de données afin de couvrir le plus grand nombre de cas timbraux.

4.2 Analyse en composantes principales

Le but de l'analyse en composantes principales (PCA) dans notre problème est de trouver les vecteurs propres de la base de données de locuteurs. Ces vecteurs propres servent de prototypes de locuteurs (*eigenspeakers*).

Nous pensions que la projection d'un locuteur incomplet sur la base de protoypes de locuteurs aurait permis de compléter les données manquantes. Malheureusement, cette méthode ne nous a pas permis de contraindre le spectre de la voyelle prédite à des valeurs positives. Alors que nous ne pensions pas que ce serait un problème, cette limite s'est révélée navrante pour nos résultats.

Les 5 composantes principales de la voyelle /a/ sur la figure 4.10 illustrent ce problème. La première composante principale a des valeurs négatives pour $f > 1600$ Hz. Cette composante étant celle ayant la valeur de projection moyenne la plus élevée, les voyelles prédites résultantes ont presque toujours une puissance négative à un endroit de cette partie du spectre.

Même en contraignant ces valeurs à 0 après la prédiction, le résultat est presque toujours moins bon que la méthode de la voyelle moyenne. Nous décidons donc d'abandonner cette méthode, du moins sous cette forme. Il serait peut-être possible d'adapter la méthode pour contraindre les vecteurs propres à des valeurs positives, mais nous choisissons de ne pas adopter cette direction dans notre recherche et de

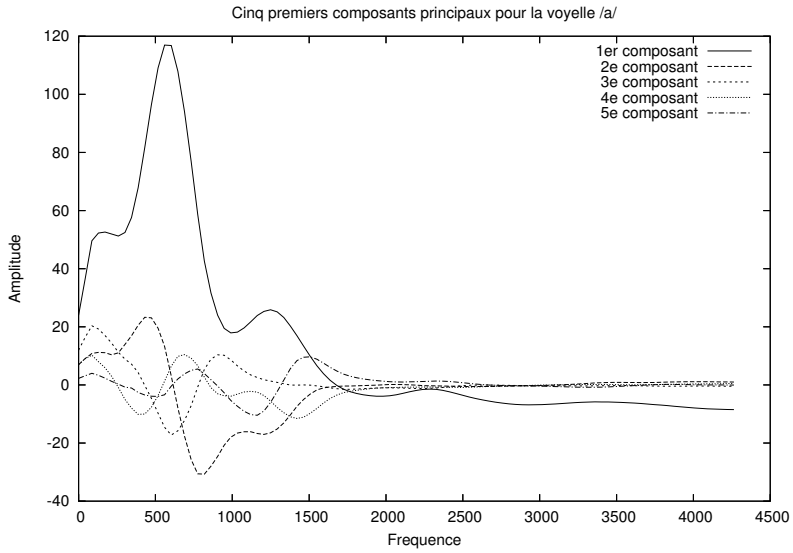


FIG. 4.10 Cinq premières composantes principales de notre base de données pour la voyelle /a/.

nous en tenir à ces résultats.

4.3 Analyse en composantes indépendantes

L'analyse en composantes indépendantes (ICA) devait nous permettre de trouver les caractéristiques timbrales les plus distinctes possibles. Comme pour la PCA, ces caractéristiques nous ont donné une base de prototypes à partir de laquelle nous pouvons compléter les locuteurs. Cependant, nous nous sommes heurtés au même genre de problème que pour la PCA.

La figure 4.11 montre cinq composantes indépendantes prises au hasard pour une voyelle /a/. Ces composantes ont été obtenues après s'être assuré d'une bonne convergence de l'algorithme ICA décrit dans la section 2.5.2. Nous pouvons voir que toutes ces composantes contiennent au moins une partie négative de leur spectre, caractéristique qui nous empêche d'utiliser cette technique pour l'application qui

nous intéresse.

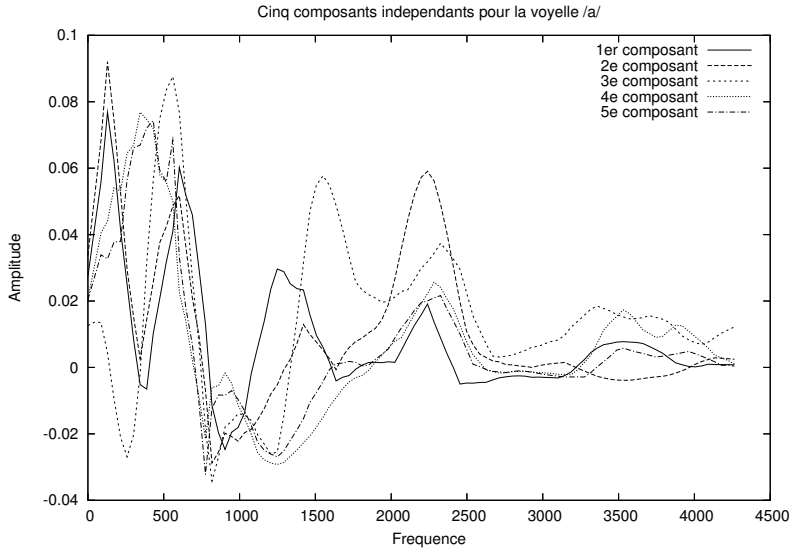


FIG. 4.11 Cinq composantes indépendantes de notre base de données pour la voyelle /a/.

Comme pour la PCA, même en contraignant ces valeurs à 0 après la prédiction, le résultat était généralement beaucoup moins bon que la voyelle moyenne. Il aurait été peut-être possible ici aussi d'adapter la méthode pour contraindre les vecteurs propres à des valeurs positives, mais nous avons décidé de laisser cette idée pour une autre recherche.

4.4 Importance de ces résultats pour la synthèse de voix

Avec une base de données réduite de seulement 31 locuteurs et 5 voyelles, nous sommes arrivés à améliorer sensiblement (de 36,6%) la prédiction du timbre d'une voyelle inconnue d'un locuteur. À notre connaissance, la prédiction de phonèmes n'a pas été étudiée en détail auparavant. Qu'on puisse arriver à de bons résultats avec un estimateur aussi simple que les noyaux, et avec aussi peu de données, nous laisse penser qu'il sera possible de construire des modèles complets et précis des

traits acoustiques qui caractérisent le timbre d'une personne.

De plus, le tableau 4.2 montre que certains phonèmes ne sont pas vraiment utiles à la prédiction d'autres phonèmes. Nous pensons qu'une matrice de poids λ_v contenant plus de phonèmes sera possiblement creuse, ce qui permettra de réduire sa taille et de faire plusieurs économies de calcul lors de la prédiction. Pour d'autres méthodes d'apprentissage machine comme les réseaux bayesiens, cela voudra dire qu'il est possible d'élaguer certaines parties du réseau pour en réduire sa taille.

Cette perspective de réduction de la complexité des algorithmes de prédiction de phonème est une bonne nouvelle, mais elle comporte un côté négatif. Un phonème qui ne peut tirer d'information que d'un seul autre phonème, comme le /a/ dans notre expérience, ne peut pas être prédit avec beaucoup de précision si cet autre phonème est manquant. Bien sûr nous pourrions prédire cet autre phonème en premier (par exemple le /e/), puis s'en servir pour prédire le phonème voulu (par exemple le /a/). Nous reconnaissons cependant que le système a ses limites, et qu'avec peu d'information la tâche de prédiction est beaucoup plus difficile à accomplir.

CONCLUSION

Notre recherche avait comme objectif de vérifier la faisabilité d'un système de complétion de corpus pour la synthèse de voix réaliste. Nous avons développé un modèle harmonique et résiduel (HNM) pour l'analyse et la synthèse de voix réaliste, et nous avons appliqué des techniques d'apprentissage machine pour prédire les phonèmes manquants de locuteurs mâles. Notre recherche s'est restreint à la capacité de prédire une voyelle manquante à l'aide de quatre autres parmi cinq voyelles, soit les sons /a/, /e/, /i/, /o/ et /u/. Les techniques de prédiction étudiées sont les k-plus proches voisins, les noyaux gaussiens, l'analyse en composantes principales et l'analyse en composantes indépendantes.

Les résultats des noyaux gaussiens entraînés par un algorithme génétique présentés dans le dernier chapitre ont permis de réduire de 36,6% la distorsion harmonique d'une voyelle prédite par rapport à la voyelle moyenne correspondante de la base de données. Ce résultat permet d'affirmer qu'il y a un avantage à utiliser l'information disponible d'un locuteur pour compléter son corpus de phonèmes. Par ailleurs, les paramètres des noyaux gaussiens asymétriques nous permettent de voir quelle voyelle nous donne le plus d'information sur la voyelle à prédire, et semble nous indiquer quelle sera l'erreur sur la voyelle prédite. Nous ne connaissons pas d'autres études qui ont abordé le sujet de la complétion de corpus dans le but de faire de la synthèse de voix réaliste. Nous considérons donc nos résultats comme étant une contribution intéressante et originale au domaine de la synthèse de voix.

Nous avons cependant conscience que notre recherche n'a fait qu'effleurer le vaste sujet qu'est la synthèse de voix. Notre étude s'est limitée aux relations entre les spectres de puissance harmonique de cinq voyelles. Une prochaine investigation pourrait étendre ces résultats au spectre de phase harmonique et au spectre de puis-

sance de la partie résiduelle du modèle HNM. Dans ce contexte, il serait nécessaire de construire une base de données contenant quelques centaines de locuteurs de chaque sexe pour couvrir davantage de cas de timbres différents. Cette base de données contiendrait idéalement des phonèmes et des diphones formant la carte phonétique complète d'au moins une langue.

Avec une base de données de cette taille, il serait possible d'utiliser des techniques d'apprentissage machine plus sophistiquées, tels les réseaux bayesiens et les modèles graphiques probabilistes. Ces derniers permettraient aussi d'étudier les relations entre les différents phonèmes en interprétant les poids optimisés par un algorithme d'apprentissage, de manière similaire à ce qui a été accompli avec les noyaux gaussiens dans cette recherche. Il serait également intéressant de développer une méthode plus formelle pour la prédiction de l'erreur de prédiction. La théorie des noyaux est déjà bien définie de ce côté, nous n'avons cependant pas pris le temps de le faire au cours de cette recherche.

Nous entrevoyons plusieurs applications possibles à un système de prédiction des phonèmes d'un locuteur. Outre la complétion simple d'une carte phonétique d'un individu, il serait possible de traduire la voix d'un locuteur dans une langue qu'il ne connaît pas, de créer à partir d'enregistrements un système de synthèse vocale pour des personnes dont les cordes vocales ont été endommagées, de donner la voix d'un joueur à son personnage de jeu vidéo ou d'apprendre progressivement la voix d'une personne durant une conversation avec un ordinateur.

Bien sûr, beaucoup d'autres aspects de la synthèse vocale autres que la modélisation des phonèmes restent à parfaire, comme la synthèse de la prosodie à partir d'un texte (rythme et intonation) et la génération de voix émotives. Nous espérons cependant que notre recherche viendra appuyer et inspirer les travaux en cours dans ces domaines également.

RÉFÉRENCES

ACKLEY, D. H., HINTON, G. E. et SEJNOWSKI T. J., 1985, “A learning algorithm for Boltzmann machnes.”, Cognitive Science 9, 147–169.

BATLINER, A., FISCHER, K., HUBER, R., SPILKER, J., NÖTH, E., 2000, “Desperately Seeking Emotions or : Actors, Wizards and Human Beings”, Proceedings of ISCA Workshop (ITRW) on Speech and Emotion 2000, Belfast, North Ireland.

BELL, A. J. et SEJNOWSKI, T. J., 1995, “An Information-Maximization Approach to Blind Separation and Blind Deconvolution”, Neural Computation, vol. 7, no. 6.

BOITE, R., BOURLARD, H., DUTOIT, T., HANCQ, J. et LEICH, H., 2000, “Traitement de la parole”, Presses polytechniques et universitaires romandes.

BURROWS, Tina-Louise., 1996, “Speech Processing with Linear and Neural Networks Models”, Ph. D. Thesis, Cambridge University Engineering Department.

CAHN, J.E., 1990, “Generating Expression in Synthesized Speech”, Technical Report, M.I.T. Media Laboratory.

DEVERGE, M., 2002, “Ecoulement dans la glotte : Comparaison de modèles théoriques simplifiés aux résultats de mesures sur modèle in-vitro rigide”, Rapport de stage de DEA : Acoustique Appliquée, Université du Maine, Laboratoire d’Acoustique, Technische Universiteit Eindhoven, Fluid Dynamics Laboratory (NT/GDY).

DUTOIT, T., PAGEL, V., PIERRET, N., BATAILLE, F., VAN DER VREKEN, O., 1996, ”The MBROLA Project : Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes”, Proceedings of ICSLP’96, Philadelphia, vol. 3, pp. 1393-1396.

DUDA, R.O., HART, P.E., STORK, D.G., 2001, "Pattern Classification", deuxième édition, John Wiley & Sons.

FOURNIER, P.-A. et BRAULT, J.-J., 2005, "Harmonic Envelope Prediction for Realistic Speech Synthesis Using Kernel Interpolation", Proceedings of the International Joint Conference on Neural Networks 2005 (IJCNN05), Montréal.

FREY, B. J., 1998, "Graphical Models for Machine Learning and Digital Communication", MIT Press, Cambridge, MA, USA.

HASTIE, T., TIBSHIRANI R. and FRIEDMAN J., 2001, "The Elements of Statistical Learning", Springer, New York.

HAYKIN, S., 1999, "Neural Networks, A Comprehensive Foundation, Second Edition", Prentice Hall, New Jersey.

HUANG, X., ACERO, A. et HON, H.-W., 2001, "Spoken Language Processing : A Guide to Theory, Algorithm and System Development", Pearson Education ; 1st edition.

JANG, G., LEE, T., 2002, "A probabilistic Approach to Single Channel Blind Signal Separation", NIPS Conference Proceedings 2002, Vancouver.

JEHAN, T., 2001, "Perceptual Synthesis Engine : An Audio-Driven Timbre Generator", Thèse de maîtrise, Massachusetts Institute of Technology.

JURAFSKY, D. et MARTIN, J.H., 2000, "Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", Prentice Hall.

KLEIJN, W.B. et PALIWAL, K.K., 1995, "Speech Coding and Synthesis", Elsevier, Amsterdam.

KREYSZIG, E., 1993, "Advanced Engineering Mathematics, 7th edition", John Wiley & Sons, New York.

LEE, S. Y., 2002, "MPEG7-Overview : Multimedia Content Description Interface", *http ://www.csie.nctu.edu.tw/˜dychen/MP7_overview1.pdf*

LEVIT, M., HUBER, R., BATLINER, A., NOETH, E., 2001, “use of prosodic speech characteristics for automated detection of alcohol intoxication”, Proceedings of Prosody in Speech Recognition and Understanding 2001, ISCA Tutorial and Research Workshop (ITRW), Red Bank, NJ, USA.

MADHURANATH, H. et HAYKIN, S., 1998, “Improved Activation Function for Blind Separation : Details of Algebraic Derivations”, CRL Internal Report, Communications Research Laboratory, McMaster University, Hamilton, Ontario.

MOULINES, E. et CHARPENTIER, F., 1990, ”Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones,” Speech Communication Vol. 9., numéro 5, pp 453-467.

PAL, S.K. et Dutta Majumder, D., 1977, “Fuzzy sets and decision making approaches in vowel and speaker recognition”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 7, pp. 625-629.

POLZIN, T.H. et WAIBEL, A. H., 1998, “Detecting Emotion in Speech”, Detecting Emotions In Speech, Proceedings of the CMC 1998.

QUATIERI T., 2001, “Discrete-Time Speech Signal Processing : Principles and Practice”, Prentice Hall.

RABINER, L.R. et SCHAFER, W., 1978, “Digital Processing of Speech Signals”, Pearson Education.

RABINER, L.R. et JUANG, B.-H., 1993, “Fundamentals of Speech Recognition”, Prentice Hall.

RAMSAY, J.O. et SILVERMAN, B.W., 1997, “Functional Data Analysis”, Springer, New York.

SERRA, X. et SMITH, J., 1990, “Spectral modeling synthesis : A sound analysis/synthesis system based on a deterministic plus stochastic decomposition.”, Computer Music Journal, 14(4) :12-24, Winter 1990.

SERRA, X., 1997, “Musical Sound Modeling with Sinusoids plus Noise.”, published in “Musical Signal Processing”, C. Roads, S. Pope, A. Picialli, G. De Poli Editors, Swets & Zeitlinger Publishers.

SHIGA, Y. et KING, S., 2003, “Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis”, Eurospeech 2003, Geneva.

SHIGA, Y. et KING, S., 2003, “Estimation of Voice Source and Vocal Tract Characteristics Based on Multi-frame Analysis”, Eurospeech 2003, Geneva.

STYLIANOU, Y., 1996, “Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.” Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris.

STYLIANOU, Y., 1996, “Decomposition of Speech Signals into a Deterministic and a Stochastic Part”, Proc. ICSLP, Philadelphia.

STYLIANOU, Y., 1998, “Concatenative Speech Synthesis using a Harmonic plus Noise Model”, The third ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia.

STYLIANOU, Y., SYRDAL, K.S., 2001, “Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis.”, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001, Salt Lake City.

TRENCH W. F., 1964, “An algorithm for the inversion of finite Toeplitz matrices.” Journal Soc. Indust. Appl. Math., v. 12, pp. 515-522.

THIBAULT, F. et DEPALLE, P., 2004, “Adaptive Processing of Singing Voice Timbre” In Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2004), The Niagara Falls, Ontario, Canada.

VINCENT, P., 2003, “Modèles à noyaux à structure locale” Thèse de doctorat, Département d’informatique et de recherche opérationnelle, Université de Montréal, Montréal.

WOUTERS, J. et MACON, M.W., 2000, “Spectral Modification for Concatenative Speech Synthesis” ICASSP 2000, Istanbul, Turkey.

ANNEXE I

CÉDÉROM CONTENANT DES EXEMPLES SONORES