

**THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE**

Spécialité

Acoustique, traitement du signal et informatique appliqués à la musique

Présentée par

M. Damien Tardieu

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Modèles d'instruments pour l'aide à l'orchestration

soutenue le 15/12/2008

devant le jury composé de :

M. Xavier Rodet	Directeur de thèse
M. Philippe Leray	Rapporteur
M. François Pachet	Rapporteur
M. Gaël Richard	Examineur
M. Stephen McAdams	Examineur
M. Jean Dominique Polack	Examineur
M. Geoffroy Peeters	Examineur

Résumé

Modèles d'instruments pour l'aide à l'orchestration

Cette thèse traite de la conception d'une nouvelle méthode d'orchestration assistée par ordinateur. L'orchestration est considérée ici comme l'art de manipuler le timbre d'un orchestre par l'assemblage des timbres des différents instruments. Nous proposons la formulation suivante du problème de l'orchestration assistée par ordinateur : Il consiste à trouver les combinaisons de sons instrumentaux dont le timbre se rapproche le plus possible d'un timbre cible fourni par le compositeur. Cette cible est spécifiée sous forme de son enregistré. Les connaissances du système sur le timbre des instruments doivent être extraites de bases d'échantillons. La première étape est donc de définir un ensemble de descripteurs du signal sonore qui permette de décrire le timbre des instruments de l'orchestre. Ces descripteurs doivent être liés à la perception du timbre. La seconde étape consiste à définir une méthode d'apprentissage de ces descripteurs pour les différents instruments. Dans ce but, nous proposons une taxinomie des modes de jeu instrumentaux. L'apprentissage consiste alors à estimer la distribution des descripteurs pour les différentes valeurs des variables de la taxinomie. Ensuite une méthode d'agrégation des distributions permet de déduire des modèles pour un vaste ensemble de sons. Cette méthode d'agrégation, issue du domaine de la décision collective, permet un haut niveau de généralisation et donc la déduction de modèles de sons absents de la base d'apprentissage. À partir des modèles d'instruments, nous pouvons déduire le modèle d'une combinaison de sons instrumentaux et donc évaluer la probabilité que cette combinaison soit perceptivement proche de la cible.

Mots-clefs : orchestration, modèle probabiliste d'instrument, identification d'instrument, timbre, descripteur audio, logarithmic opinion pool, modes de jeu instrumentaux.

Abstract

Instrument models for computer aided orchestration

This thesis deals with the design of a new method for computer aided orchestration . Orchestration is seen as the art of manipulating the timbre of an orchestra by the combination of different instruments. The orchestrating task we focus on is the following one : To find combinations of instrumental sounds whose timbre is close to a target sound provided by the composer. This target is specified by a recorded sound. The knowledge of the system on the instrument timbre is extracted from sample databases. First, a set of features that describe the timbre of the instruments is presented. Then we propose a probabilistic model of these features. This model is based on a taxonomy of the various instruments playing techniques and on the approximation of the feature distributions by a product of Gaussians. The model parameters are learned on the sample databases. This model and the proposed learning technique allow the inference of sounds that are missing in the databases. From the instrument models, we find the model of a combination of sounds to assess the similarity between this combination and the target sound. Proposed methods are tested on solo and polyphonic sound classification tasks.

Keywords : orchestration, probabilistic instrument models, instrument classification, timbre, audio features, logarithmic opinion pool, instrument playing technique.

Table des matières

Introduction	15
1 Différents points de vue sur l'orchestration	17
1.1 Introduction	17
1.2 les traités	17
1.3 Les formalisations	18
1.3.1 Koechlin	18
1.3.2 La psychologie	20
1.4 Orchestration assistée par ordinateur	22
2 Formulation et questions soulevées	25
2.1 Approche choisie	25
2.2 Décrire les sons	26
2.2.1 Approche psychoacoustique	26
2.2.2 Approche computationnelle	27
2.3 Comparer des sons	28
2.3.1 Qu'est ce que la similarité	28
2.3.2 Variabilité de la similarité	29
2.3.3 Proposition de mesure de similarité	30
2.4 Généralisation du savoir instrumental	30
2.5 Combiner des descripteurs et les modèles	31
2.6 La combinatoire	31
2.7 Description du système d'orchestration	32
2.7.1 Introduction	32
2.7.2 Un problème multicritères	32
2.7.3 Un problème probabiliste	33
2.7.4 Définition formelle du problème de l'orchestration	34
3 Instruments et modes de jeu considérés	35
3.1 Introduction	35
3.2 Description des instruments et des modes de jeu	35
3.2.1 Vue d'ensemble	35
3.2.2 Les modes de jeu des cordes	36

3.2.3	Les modes de jeu des cuivres	38
3.2.4	Les modes de jeu des bois	38
3.2.5	Les modes de jeu des anches doubles	39
3.2.6	Combinaisons de modes de jeu	39
3.3	Bases d'échantillons utilisées et répartition	39
3.4	Organisations des modes de jeu	40
3.4.1	Similarité de techniques entre instruments	40
3.4.2	Taxinomie proposée	41
3.4.3	Utilisation de la taxinomie pour la simulation des orchestrations	45
4	Description du son et des combinaisons	49
4.1	Problématique du choix des descripteurs	49
4.2	Les descripteurs utilisés	49
4.2.1	La sonie et l'énergie	49
4.2.2	Structure harmonique	52
4.2.3	L'énergie par bandes de Mel	56
4.2.4	L'enveloppe spectrale sur les bandes de Mel	56
4.2.5	Le centroïde et l'étendue spectrale	56
4.2.6	L'enveloppe du bruit	57
4.2.7	Moyenne des descripteurs spectraux et normalisation	57
4.2.8	Les modulations	58
4.2.9	Temps d'attaque	60
4.3	Liste complète des descripteurs et abréviations utilisées	61
5	L'addition des descripteurs et des modèles	63
5.1	Introduction	63
5.2	Addition de signaux	64
5.2.1	Énergie d'une somme de signaux	64
5.2.2	Interaction entre sons sinusoïdaux	64
5.2.3	L'effet de salle	68
5.2.4	Conclusion	69
5.3	Expériences perceptives	72
5.3.1	Introduction	72
5.3.2	Expérience 1 : Jugement de fusion	72
5.3.3	Expérience 2 : similarité entre mélanges	76
5.3.4	Conclusion	78
5.4	Formules d'additions	79
5.4.1	Introduction	79
5.4.2	Les descripteurs du spectre	79
5.4.3	Le centroïde spectral	80
5.4.4	La noisiness	80
5.4.5	L'ambitus de la modulation d'énergie	80
5.4.6	Le logarithme du temps d'attaque	82

5.5	Distribution des descripteurs des combinaison	82
5.5.1	Cas d'une distribution normale	82
5.5.2	Cas d'une distribution log-normale	82
6	Modèle d'instrument	85
6.1	Introduction	85
6.2	Distribution modélisée	85
6.3	Problème posé par la formulation.	86
6.4	Grouper hauteurs et nuances	86
6.4.1	Principe	86
6.4.2	Relation entre le niveau de sonie et l'énergie	87
6.4.3	Calcul de la distribution conditionnelle	88
6.5	Mode de Jeu	89
6.5.1	Motivations	89
6.5.2	Factorisation de la distribution	90
6.5.3	Utilisation de la taxinomie des modes de jeu	91
6.5.4	Apprentissage des modèles des sous-problèmes	91
6.5.5	Calcul des modèles finaux : résumé	91
6.6	Agrégation des distributions finales	92
6.6.1	Introduction	92
6.6.2	Les méthodes d'agrégation	93
6.6.3	Cas gaussien pour le logarithmic opinion pool	94
6.6.4	Choix des poids	95
6.7	Évaluation d'une cible avec les modèles	95
7	Évaluation sur des échantillons solo et choix des paramètres	97
7.1	Introduction	97
7.2	Exemple de modèle appris	97
7.3	Critères d'évaluations des tâches de classification	98
7.4	Évaluation des modèles initiaux	101
7.4.1	Évaluation de chaque descripteur	101
7.4.2	Classification avec un sous-ensemble de descripteurs	102
7.4.3	Résultats	102
7.5	Évaluation des modèles finaux	106
7.5.1	Procédure	106
7.5.2	Résultats	107
7.6	Évaluation des méthodes d'agrégation	109
7.6.1	Procédure et classifieur utilisés	109
7.6.2	Résultats	109
7.7	Conclusion	111

8	Évaluation sur des mélanges synthétiques	115
8.1	Introduction	115
8.2	Procédure	115
8.3	La base de test	116
8.3.1	Génération des combinaisons	116
8.3.2	Sous problèmes considérés et descripteurs utilisés	116
8.4	Critères d'évaluation	117
8.4.1	Rang de la solution théorique	117
8.4.2	Classement de la solution théorique connaissant les poids	117
8.4.3	Similarité symbolique entre la solution théorique et la solution du front . . .	118
8.5	Résultats	118
8.5.1	Mélanges de 2 sons	118
8.5.2	Mélanges de 3 sons	120
8.6	Conclusion	121
	Conclusion	125
A	Descripteurs utilisés dans les expériences perceptives	129
B	Taxinomie complète des modes de jeu	131
C	Liste des instruments et des abréviations	133

Table des figures

1.1	Extraits des instruments et combinaisons de remplacement du traité d'orchestration de Alfred Blatter [Bla97]	19
1.2	Les échelles d'intensité et de volume dans le traité de l'orchestration de Koechlin [Koe54]	21
2.1	Schéma général du système	26
2.2	Exemple de front de Pareto	33
3.1	La famille des cuivres	36
3.2	La famille des cordes	36
3.3	La famille des anches doubles	36
3.4	La famille des bois	37
3.5	Répartitions des instruments dans les différentes bases	40
3.6	Répartition des modes de jeu dans les différentes bases	41
3.7	Répartitions des nuances pour les modes de jeu de chaque famille	42
3.8	Arbre de décision pour les modes de jeu des cordes	44
3.9	Arbre de décision pour la simulation des résultats	46
4.1	Sonie globale estimée par la méthode N_6 en fonction de la nuance pour des sons entretenus	51
4.2	Sonie globale estimée par la méthode N_6 en fonction de la nuance pour des sons de cordes pizzicato	51
4.3	Sonie globale estimée par la méthode N_6 en fonction de l'octave pour des sons de cordes entretenus	52
4.4	Partiels extraits d'un son de tuba la#1 mf et partiels sélectionnés par la classification .	54
4.5	Assignement des partiels d'une cible à deux fréquences fondamentales	55
4.6	Logarithme du spectre de Mel d'un son de trompette jouant un do6 <i>mf</i>	56
4.7	Exemple d'extraction de l'enveloppe du bruit (extrait de [YR06])	57
4.8	Log noisiness des sons de clarinette et de flûte pour les modes de jeu ordinario, aeolien+ordinario et aeolien	58
4.9	Centroïde spectrale de l'enveloppe du bruit pour les sons aéolien de flûte, clarinette, flûte basse et clarinette basse	58

4.10	Ambitus de la modulation d'énergie pour les sons <i>flatterzunge</i> de Studio en Ligne en linéaire et en logarithme	59
4.11	Vitesse et logarithme de l'ambitus de la modulation d'énergie en fonction du mode de jeu	60
5.1	Déviation de la fréquence fondamentale du son mesurée sur les sons ordinaires de Studio Online par rapport à la fréquence fondamentale théorique	67
5.2	Écart interquartile de la fréquence fondamentale mesurée	67
5.3	Pression mesurée le long d'une ligne à fréquence constante et mesurée en variant la fréquence à une position fixée (extrait de [Kut00])	68
5.4	Amplitude des partiels d'un alto jouant un <i>do4 mf</i> sans effet de salle (en bleu) et avec effet de salle (en rouge)	70
5.5	Amplitude des partiels d'un cor jouant un <i>do4 ff</i> sans effet de salle (en bleu) et avec effet de salle (en rouge)	71
5.6	Moyenne des jugements de fusion pour les instruments percussifs	75
5.7	Moyenne des jugements de fusion pour les instruments entretenus	75
5.8	Dissemblance perçue dans l'expérience 2 lorsque les sons percussifs appartiennent à la même famille ou non	77
5.9	Dissemblance perçue dans l'expérience 2 lorsque les sons entretenus appartiennent à la même famille ou non	78
5.10	Prédiction des jugements de similarité par une combinaison linéaire du spectre de Mel en log et du rapport des logarithmes des centroïdes temporels des deux sons	79
5.11	Prédiction du logarithme de l'ambitus de modulation pour de combinaisons de 2 sons	81
5.12	Prédiction du logarithme de l'ambitus de modulation pour de combinaisons de 3 sons	81
6.1	Dépendances entre les variables d'état et un descripteur pour le modèle de l'équation 6.1	86
6.2	Dépendance entre les variables pour le modèle permettant de regrouper les hauteurs et les nuances dans les ensembles d'apprentissage	87
6.3	Niveau de sonie et log énergie pour des sons de contrebasse à l'octave 2 à gauche et de trompette à l'octave 4 à droite	88
6.4	Exemple de regroupement des modes de jeu pour l'apprentissage	90
6.5	Dépendance entre les variables pour la variable d'état <i>instrument</i> . Le réseau est le même pour les différents <i>modes de jeu</i> de la taxinomie	90
6.6	Exemple de produit d'expert gaussien [Hin99]	93
7.1	Moyenne et écart type des distributions de l'amplitude des harmoniques pour le modèle de la clarinette en Sib pour une note (<i>la4</i>) à différentes nuances (à gauche), et pour une nuance (<i>mf</i>) à différentes hauteurs	98
7.2	Moyenne et écart type des distributions de l'amplitude des harmoniques pour le modèle de trombone pour une note (<i>la3</i>) à différentes nuances (à gauche), et pour une nuance (<i>mf</i>) à différentes hauteurs	99

7.3	Moyenne et écart type des distributions de l'ambitus (gauche) et de la vitesse (droite) de la modulation d'énergie pour le modèle de tremolo et non tremolo	99
7.4	Résultats par descripteur pour la base de test Studio Online	103
7.5	Résultats par descripteur pour la base de test Studio Online	104
7.6	Résultats globaux pour toutes les bases (rappel moyen précision moyenne (nombre de classes))	105
7.7	Effet des approximations pour le sous-problème <i>Instrument</i>	108
7.8	Effet des approximations pour le sous-problème <i>SourdineString</i>	108
7.9	Résultats de classification par sous-problème pour les vents	110
7.10	Résultats de classification par sous-problème pour les cordes selon la méthode d'agrégation utilisée. Les traits tillés représente le résultat d'un classifieur aléatoire	112
7.11	Résultats de classification par sous-problème pour les cordes avec $Z = 1$ selon la méthode de calcul des poids utilisée. Les traits tillés représente le résultat d'un classifieur aléatoire	113
8.1	Histogramme cumulé du rang de la solution théorique pour la note sol3	119
8.2	Classement de la solution théorique selon la norme qu'elle induit pour les combinaisons de deux sons en sol3	119
8.3	Histogramme de la distance de Hamming entre la solution théorique et la combinaison de rang 1 la plus proche (à gauche) et entre toutes les combinaisons (à droite) pour les combinaisons de deux sons	120
8.4	Histogramme cumulé du rang de la solution théorique pour la note sol2 à gauche et sol4 à droite	121
8.5	Histogramme cumulé du rang de la solution pour les combinaisons de 3 sons pour la recherche des modes de jeu	121
8.6	Histogramme de la distance de Hamming entre la solution théorique et la combinaison de rang 1 la plus proche (en bas) et entre toutes les combinaisons (en haut) pour les combinaisons de trois sons	122
8.7	Classement de la solution théorique selon la norme qu'elle induit pour les combinaisons de trois sons	123

Liste des tableaux

3.1	Exemple de taxinomie	43
3.2	Ensemble des variables et des valeurs possibles	43
4.1	Niveau de sonie des différentes nuances	52
4.2	Liste des descripteurs utilisés	61
5.1	Instruments utilisés pour les sons entretenus	73
5.2	Instruments utilisés pour les sons percussifs	73
5.3	Mélanges sélectionnés pour la seconde expérience	76
7.1	ensemble des sous-problèmes considérés et des valeurs possibles	111
8.1	ensemble des sous-problèmes considérés et des valeurs possibles	117

Introduction

L'orchestration tient une place particulière dans la composition. Cet art de la manipulation du timbre de l'orchestre, du passage de la notation musicale à la réalisation acoustique, repose, bien plus que les autres aspects de la composition, sur l'expérience du compositeur, sur sa connaissance du timbre des instruments et sur sa capacité à prédire comment sonneront ces instruments mis ensemble. Comme l'explique Berlioz, l'orchestration a peu été formalisée, ou peut-être ne peut pas l'être, et reste par conséquent difficile à transmettre : « *Considéré sous son aspect poétique, cet art s'enseigne aussi peu que celui de trouver de beaux chants, de belles successions d'accords et des formes rythmiques originales et puissantes. On apprend ce qui convient aux divers instruments, ce qui pour eux est praticable ou non, aisé ou difficile, sourd ou sonore ; [...] quant à leur association par groupes, par petits orchestres et par grandes masses, quant à l'art de les unir, de les mêler, de façon à modifier le son des uns par celui des autres, en faisant résulter de l'ensemble un son particulier que ne produirait aucun d'eux isolément, [...] on ne peut que signaler les résultats obtenus par les maîtres en indiquant leurs procédés ; résultats qui sans doute, seront encore modifiés de mille manières en bien ou en mal par les compositeurs qui voudront les reproduire.* » [Ber70]. Tous les compositeurs se retrouvent face à la même difficulté quand il s'agit d'expliquer l'orchestration. Ainsi pour Rimsky-Korsakov « *L'orchestration est création et on apprend pas à créer* »¹. Plus d'un siècle plus tard, même si l'orchestration a évolué dans ses pratiques, elle reste toujours une discipline éminemment empirique. C'est sans doute la raison pour laquelle très peu de travaux se sont intéressés à une approche informatique de ce domaine, contrairement aux autres aspects de la composition. Il n'est pas possible de définir un ensemble de règles à coder dans un programme. Mais la situation a changée avec l'avènement de l'analyse du son et de l'apprentissage automatique. Une approche informatique empirique est maintenant possible. L'idée de travailler sur l'aide à l'orchestration est née d'un groupe de discussion à l'Ircam. Ce groupe, composé de plusieurs compositeurs et de chercheurs, avait pour but de se demander comment la science et l'informatique pouvait être utilisés comme aide dans certaines tâches de l'orchestration. Le souhait le plus souvent exprimé par les compositeurs était celui d'avoir un système capable de faire des propositions d'orchestrations auxquelles ils n'auraient pas pensé, des propositions inhabituelles ou surprenantes, en permettant une exploration des possibilités sonores de l'orchestre. En effet, l'orchestration étant principalement basée sur l'expérience, il est parfois difficile de sortir des chemins connus. Un autre souhait était la possibilité d'automatiser en partie certaines tâches. A ce titre l'exemple du compositeur Yan Maresz est particulièrement significatif. Ce dernier écrit en 1995 une oeuvre pour Trompette et électronique (Métallics [Mar01a]). En 2001, L'Ensemble intercontemporain lui commande une version orchestrale de cette oeuvre. Il passe donc plusieurs mois à "traduire" la partie

¹Cité par Koechlin dans [Koe54]

électronique de *Metallics* en partition pour orchestre et crée *Metal Extension* [Mar01b]. Il est à ce moment persuadé qu'un programme informatique lui aurait permis de passer beaucoup moins de temps à cette "traduction". Des discussions de ce groupe émerge une approche de l'orchestration assistée par ordinateur : le compositeur spécifie une cible, par exemple sous forme de son enregistré, que le système essaie d'approcher ou d'imiter en combinant plusieurs sons instrumentaux. Les sons instrumentaux sont extraits des bases d'échantillons aujourd'hui largement disponibles. Cette approche a plusieurs avantages. D'abord elle ne suppose aucune connaissance théorique préalable, mais uniquement des connaissances sur le timbre des instruments qui peuvent être extraites automatiquement des bases d'échantillons grâce aux techniques de description du signal musical. Cette absence de connaissance musicale permet précisément à l'ordinateur de proposer des solutions n'ayant rien à voir avec la tradition musicale ou avec les habitudes d'un compositeur. Ensuite en basant la connaissance sur des bases d'échantillons, les possibilités de chaque instrument comme leur registre peuvent être implicitement prises en compte. Enfin une telle architecture remplit les besoins d'automatisation suscités par des problèmes tels que l'écriture de *Metal Extension*.

La démarche ainsi formulée soulève de nombreuses questions sur le plan scientifique. La plus importante est peut-être celle de la description du timbre des instruments et des combinaisons d'instruments. Le choix de cette description détermine en effet les types de sons que le système est capable de traiter. Ensuite il faut déterminer une méthode de comparaison entre sons qui reflète la perception de la similarité. La troisième question relève de la généralisation des connaissances extraites des bases d'échantillons. La dernière concerne l'exploration de l'espace des combinaisons, qui relève de la combinatoire et ne sera pas traitée dans cette thèse. Après une description, au premier chapitre, des différentes approches de l'orchestration qui ont été proposées jusqu'à aujourd'hui, ces questions sont discutées dans le second chapitre. Les réponses apportées nous permettent de proposer une définition formelle du problème de l'orchestration qui servira de base à la suite de la thèse. Les chapitres suivants sont consacrés à la description des différents blocs du système d'orchestration. Le chapitre 3 détaille la base d'échantillons utilisées pour l'extraction des connaissances sur le timbre des instruments. Cette description nous permet de proposer une taxinomie des modes de jeu instrumentaux qui sera utilisée pour structurer l'espace des connaissances et permettre un apprentissage plus simple, plus robuste et plus général de ces modes de jeu et donc des différents timbres de chaque instrument. Dans le quatrième chapitre, nous proposons un ensemble de descripteurs du timbre adapté au problème de l'orchestration. Le chapitre suivant aborde la question de l'addition de ces descripteurs ou comment, à partir des descripteurs d'un ensemble de son, nous pouvons estimer les descripteurs de la combinaison de ces sons. Ce problème fait intervenir des considérations d'ordre acoustique, mathématique mais aussi perceptive. Nous présentons à ce propos un travail sur la perception de combinaisons de sons instrumentaux réalisé à l'université de McGill en collaboration avec le Professeur Stephen McAdams. Enfin le chapitre six présente le modèle d'instrument utilisé pour décrire les sons instrumentaux. Ce modèle probabiliste est basé sur l'apprentissage de la distribution des descripteurs pour les différentes valeurs des variables de la taxinomie proposée au chapitre 3. Nous proposons ensuite une méthode d'agrégation des distributions issue du domaine de la prise de décision collective. Les deux derniers chapitres sont consacrés à l'évaluation de la méthode proposée, d'abord par des tâches de classification d'échantillons solo au chapitre 7 et d'échantillons polyphoniques au chapitre 8.

Chapitre 1

Différents points de vue sur l'orchestration

1.1 Introduction

Le but de ce chapitre est de présenter les différentes manières dont l'orchestration a été abordée jusqu'à aujourd'hui et ainsi de comprendre comment est née l'idée de l'orchestration assistée par ordinateur. Nous décrivons d'abord les traités d'orchestration et la manière empirique qu'ils ont de présenter le problème, puis les premières formalisations. Enfin nous décrivons les différents systèmes d'orchestration assistée par ordinateur qui ont inspiré le travail de cette thèse.

1.2 les traités

La description des techniques d'orchestration se fait principalement sous forme de traités. Citons les principaux : le Grand Traité d'instrumentation et d'orchestration modernes (1855) d'Hector Berlioz [Ber70], les Principes d'orchestration de Rimsky-Korsakov (1908) [RK64], le Traité de l'orchestration (4 vol., 1935-1943) de Charles Koechlin [Koe54] et la Technique de l'orchestre contemporain (1958) de Casella et Mortari [CM58]. Les traités les plus récents viennent des États-Unis. Mais, généralement, ceux-ci ne portent guère d'attention (en tout cas, pas de manière approfondie et satisfaisante) aux connaissances et techniques acquises dans la musique écrite depuis les années cinquante, principalement en Europe. Tous ces ouvrages donnent une image assez fidèle des connaissances de leur époque et ont l'avantage d'en fixer et répertorier les règles.

Rappelons ici la différence entre l'instrumentation qui étudie le fonctionnement et les possibilités de chaque instrument et l'orchestration qui concerne l'assemblage des timbres des instruments. Le traité de Berlioz est plutôt un traité d'instrumentation dans la mesure où il décrit ce qui est techniquement faisable avec les différents instruments de l'orchestre, comment ils sonnent ou dans quel contexte il faut les utiliser. La description du timbre des instruments est très empirique parfois même poétique : « *Le basson est à l'orchestre d'une grand utilité dans une foule d'occasion. Sa sonorité n'est pas très forte, et son timbre, absolument dépourvu d'éclat et de noblesse, a une propension au grotesque, dont il faut toujours tenir compte quand on le met en évidence* », ou encore concernant l'alto « *Il est aussi*

agile que le violon, le son de ses cordes graves a un mordant particulier, ses notes aigües brillent par leur accent tristement passionné, et son timbre en général, d'une mélancolie profonde, diffère de celui des autres instruments à archet ». Ces descriptions reflètent généralement la sensation que lui associe Berlioz. Il est donc difficile d'en extraire quelques généralités. Concernant la combinaison des sons, aucun chapitre ne lui est dédié, il faut chercher au fil des pages pour trouver des conseils tirés de l'expérience du compositeur : « *En général, les maîtres modernes écrivent les flûtes trop constamment dans le haut ; ils semblent toujours craindre qu'elles ne se distinguent pas assez au dessus de la masse de l'orchestre. Il en résulte qu'elles prédominent, au lieu de se fondre dans l'ensemble, et que l'instrumentation devient perçante et dure plutôt que sonore et harmonieuse.* » ou encore : « *Les sons graves du hautbois, disgracieux lors qu'ils sont à découvert, peuvent convenir dans certaines harmonies étranges et lamentables, unis aux notes graves des clarinettes, et au Ré, Mi, Fa et Sol bas des flûtes et des cors anglais* ». Arrive ensuite le traité de Rimsky-Korsakov, toujours très empirique, la description des combinaisons d'instruments devient plus systématique. Des chapitres sont consacrés exclusivement à la description de combinaison d'instruments de la même famille et de familles différentes, ainsi il détaille l'ensemble des combinaisons de un, deux, trois ou quatre vents, en décrit le timbre et explique pourquoi elles peuvent être utilisées : « *Bassoon + Clarinet. Very full quality. The gloomy character of the clarinet prevails in the lower register, the sickly quality of the bassoon in the higher.* »

Les traités plus récents, mis à part celui de Koechlin que nous traitons dans le paragraphe suivant, sont assez similaires à celui de Korsakov. Ils mettent généralement à jour les possibilités des différents instruments qui sont de plus en plus grandes, mais n'apportent pas de nouvelle approche dans la présentation de l'orchestration et décrivent ainsi au cas par cas les différentes combinaisons instrumentales sans donner de méthode systématique. Citons cependant le traité de Alfred Blatter [Bla97] qui propose des combinaisons de remplacement pour tous les instruments. C'est à dire que pour chaque instrument, il propose selon le registre un instrument ou une combinaison d'instruments qui ont un timbre proche (voir figure 1.1).

Ces traités mettent en évidence le fait qu'il est nécessaire de très bien connaître l'instrumentation avant d'étudier l'orchestration. Un système prétendant aider à l'orchestration devra donc nécessairement connaître l'instrumentation, aussi bien concernant le timbre des instruments que leurs possibilités techniques (registre, variation de nuance ...).

1.3 Les formalisations

1.3.1 Koechlin

Le traité de Koechlin se distingue des autres en ce qu'il expose clairement des buts et propose des méthodes pour les atteindre [Chi07]. Il propose ainsi deux échelles de description des instruments : le volume et l'intensité (figure 1.2). Alors que l'intensité est souvent étudiée dans les autres traités, le volume est une qualité spécifique au traité de Koechlin. Aucune définition précise de cette notion n'est donnée, mais le classement des instruments proposés permet de s'en faire une idée intuitive. Cette échelle est utilisée pour obtenir différents plans orchestraux. Si des instruments jouant simultanément ont des volumes et des intensités similaires, ils seront perçus comme appartenant à un même plan.

		Orchestration
386		
Bass clarinet (throat)	flute breath tones; stopped horn; string bass harmonics	
Bass clarinet (low)	cup-muted trombone; cello	bassoon with muted horn
Bassoon (high)		muted horn with oboe; clarinet with soprano saxophone
Bassoon (low)	horn (muted)	bass clarinet with cello
Contra bassoon (high)	viola or cello <i>ponticello</i>	
Contra bassoon (low)	muted tuba	trombone with string bass
Soprano saxophone (high)	clarinet; English horn; bassoon	
Soprano saxophone (low)	cup-muted trumpet; flugelhorn into hat	
Alto saxophone (high)	violin	violin with oboe; clarinet with flute
Alto saxophone (low)		horn with cello (or viola); clarinet with horn
Tenor saxophone (high)		muted horn with bassoon; cello with clarinet
Tenor saxophone (low)		tuba with cello; bassoon with horn (or trombone)
Baritone saxophone (high)		oboe with clarinet; viola with horn
Baritone saxophone (low)		tuba with string bass; trombone with cello
Bass saxophone (high)	string bass; cello (<i>ponticello</i>)	
Bass saxophone (low)	electric bass	tuba with string bass
Horn (high)	flugelhorn; trombone; trumpet with harmon mute (without stem); flute; alto flute	
Horn (low)	bassoon	cello with bassoon; (bass) clarinet with trombone
Stopped horn (high)		clarinet or flute with oboe and violin (<i>ponticello</i>)
Stopped horn (low)		bassoon with bass clarinet (or clarinet) and viola (or cello) <i>ponticello</i>
Muted horn (high)	bassoon; clarinet (throat tones); or flute (breath tones)	
Muted horn (low)	string bass (<i>sul tasto</i>)	cello with clarinet
Trumpet (high)		oboe with clarinet and flute
Trumpet (low)		horn with clarinet and flute
Cornet (high)	trumpet (into stand); flugelhorn	clarinet with oboe and flute
Cornet (low)	trumpet (into stand); flugelhorn	flute(s) with muted horn
Straight-muted trumpet (high)	oboe	soprano saxophone with flute
Straight-muted trumpet (low)		stopped horn with clarinet; stopped horn with flute and oboe

FIG. 1.1 – Extraits des instruments et combinaisons de remplacement du traité d'orchestration de Alfred Blatter [Bla97]

Koechlin parle alors d'équilibre. L'équilibre des sonorités n'est pas essentiel en soi dans l'écriture d'une oeuvre selon Koechlin, mais il est important de savoir écrire des orchestrations équilibrées, donc de comprendre la notion d'équilibre des intensités et des volumes, pour pouvoir ensuite réaliser des orchestrations inégales à dessein. L'intérêt réside donc dans le fait de proposer des dimensions permettant de classer les instruments et de prévoir l'effet de leur combinaison. [Chi07] propose une comparaison de la notion de volume avec les espaces de timbre (voir paragraphe 2.2.1). Il ressort de cette analyse que le volume serait lié au centroïde ou à l'étendue spectrale des sons instrumentaux et que ces derniers attributs seraient importants pour analyser une orchestration.

1.3.2 La psychologie

Certains chercheurs en psychologie cognitive ont appliqué cette science à la compréhension des règles de l'orchestration. L'analyse de scènes auditives (ASA) [Bre90] est un domaine particulièrement utile à cette compréhension. L'ASA a pour but de comprendre comment le système auditif organise le mélange de sons qu'il reçoit de l'environnement en un ensemble cohérent de sources, c'est à dire comment le système auditif regroupe les composantes sonores issues d'une même source pour former une *image mentale* de cette source. Par exemple comment pouvons-nous, à l'écoute d'un quatuor à cordes, regrouper l'ensemble des harmoniques du violon en un seul son que nous associons à ce violon. Deux types de regroupements sont mis en évidence par [Bre90]. Le regroupement séquentiel et le regroupement simultané. Le premier explique comment une séquence de son est associée à une même source, le second explique comment les composantes spectrales sont regroupées. Les facteurs de regroupement séquentiels sont, entre autres, la proximité de fréquence ou de note, la proximité temporelle, la régularité de tempo et la similitude de timbres. Les facteurs de regroupement simultané sont l'harmonicité (une série de partiels entretenant des rapports harmoniques aura plus de chance d'être associée à une même source), le synchronisme d'attaque, des fluctuations d'amplitude cohérente et la cohérence de l'enveloppe spectrale (mcadams). [Bre91] explique comment les facteurs de regroupement des flux peuvent être utilisés pour analyser certaines règles de l'orchestration ou du contrepoint. En créant deux flux séquentiels distincts, on peut se permettre de créer des dissonances acoustiques entre certaines notes simultanées des deux flux sans pour autant créer de dissonance perceptive, le groupement horizontal, dans ce cas l'emporte sur le groupement vertical. Ces études sont les premières à montrer que certaines règles d'orchestration peuvent être expliquées par les sciences cognitives et inversement, à supposer qu'il est possible d'utiliser la science pour définir des règles de l'orchestration.

Ainsi [KC91], tente de déterminer comment sont perçus des sons de duo instrumentaux. Dans cette expérience, il est demandé aux sujets de juger le degré de similarité entre toutes les combinaisons de cinq instruments : le hautbois, la clarinette, la flûte, le saxophone alto et la trompette, dans six contextes différents. Une analyse MDS (Multidimensional Scaling analysis) est ensuite appliquée aux résultats, puis des expériences de verbalisation sont conduites pour caractériser les axes retournés par cette analyse. Les trois axes trouvés sont : nasal/non nasal, brillant/riche, simple/complexe. Ils démontrent ainsi que comme des sons simples, des paires peuvent être placés dans un espace perceptif multidimensionnel. Dans un second article [KC93], les auteurs essaient de relier ces dimensions au degré de fusion et d'identification des paires d'instruments. Le degré de fusion est assez bien prédit par la distance entre les deux instruments dans un espace de timbre. De plus la position des paires

288

4^e ÉCHELLE DES VOLUMES, ÉCHELLE DES INTENSITÉS : Pour les *Volumes* comme pour les *Intensités*, on admet en général équivalence entre : 3 pupitres de *Contrebasses* – 4 pup. de *Violoncelles* – 5 pup. d'*Altos* – 6 pup. de 2^{es} *Violons* ou de 1^{er} *Violons* (mais le plus souvent, afin de mettre davantage les 1^{ers} V. en dehors, on en prend 7 pupitres pour 6 de 2^{es} V.)

S'il s'agit de *Bois* ou de *Cuivres*, il faut classer séparément les Volumes et les Intensités, car il n'y a pas correspondance exacte.

Dans l'ensemble, et toutes réserves faites sur des cas particuliers (car il est difficile de préciser), on donnerait les indications suivantes, en commençant par les volumes ou les intensités les plus considérables :

<i>Volumes :</i>					
Tubas Contrebasses en cuivre	Cors Saxhorns	Trombones	Trompettes Flûtes dans le grave Bassons Saxophones	Clarinettes Hautbois dans le grave Flûtes dans le medium	Hautbois (medium et aigu) Flûtes (aigu) Clarinettes (suraigu)
<i>Intensités</i> ⁽¹⁾					
Ondes Martenot Timbales (et autres instruments de Percussion)	Orgue (<i>jeux d'anches</i>) Trombones Cordes (<i>tutti</i>) Petite flûte (<i>à l'aigu</i>) Trompettes (<i>à l'aigu</i>)		Trompettes V. I + V. II <i>unis (et ff)</i>	Tubas Cors Bassons <i>dans le grave</i> Clarinettes <i>à l'aigu</i> Flûtes <i>à l'extrême aigu</i>	
Hautbois <i>dans le grave</i> Saxophones Flûtes <i>à l'aigu</i>	Hautbois (<i>medium et aigu</i>) Clarinettes (<i>dans le grave</i>) Cor anglais (<i>grave et medium</i>) Bassons (<i>dans le medium</i>)		Basson <i>à l'aigu</i> Cor anglais <i>à l'aigu</i> Clarinettes <i>dans le medium</i>	Flûtes (medium et grave)	
Notons également, pour la <i>facilité plus ou moins grande à jouer pp</i> :					
le plus difficilement pp basson dans le grave cor <i>à l'aigu</i> trompette <i>à l'aigu</i> trombone <i>à l'aigu</i> petite flûte et flûte, <i>à l'extrême aigu</i> clarinette <i>à l'extrême aigu</i> hautbois dans le grave (<i>sib, si, do, do#</i>)	p à la rigueur possibles ⁽²⁾ cor dans le grave cor anglais dans le grave hautbois dans le bas du medium flûte <i>à l'aigu</i> clarinette <i>à l'aigu</i> saxophone dans le grave basson dans le medium basson <i>à l'extrême aigu</i>		p sans difficulté cor, medium hautbois) 2 ^d registre C. A. trompette, medium basson dans le haut du medium clarinette basse <i>à l'aigu</i> saxophone <i>à l'aigu</i>		
facilement p (et même pp) trombone, medium cor avec sourdine trompette avec sourdine clar. basse, grave ou medium saxophone, medium basson (<i>do à sol</i>)	le plus facilement pp cordes cors sons bouchés clar. medium ou grave flûte medium ou grave		facilement ppp cordes avec sourdines clar. medium, flûte medium cors sons bouchés, dans le medium.		

⁽¹⁾ Ces deux tableaux ne peuvent donner que des *points de repère*; il ne faut point les tenir pour rigoureux. La *réalité* est en somme *beau coup plus complexe*. Il était impossible, notamment, d'indiquer les variations de *Volume* du Basson et du Saxophone, selon les tessitures de l'instrument. Il aurait fallu tenir compte, aussi, du Saxophone Soprano, beaucoup moins gros que le Ténor ou le Baryton, et déjà plus mince que l'Alto. En outre, les volumes ne sont point exactement les mêmes chez tous les instrumentistes; certains hautboïstes et clarinettes possèdent un aigu assez ample; il y a des bassonistes dont l'ampleur de son se manifeste au moins jusqu'aux *lab, la, sib* (au-dessus des lignes de la *clef d'ut*); chez certains saxophonistes c'est presque le volume du Cor, etc. On a, dans le tableau précédent, pris une sorte de moyenne, et l'on reconnaît que des exceptions (peut être assez nombreuses) peuvent aller à l'encontre de ce que nous avons indiqué.

De même pour les *Intensités*. Le fait, par exemple, d'avoir classé: Trombones - Trompettes - Tubas, peut être discuté; mais il nous semble répondre à peu près à la généralité des cas (sauf pour les trompettistes très vigoureux, dont l'aigu sort triomphal). Et si nous avons placé le Tuba dans la même rangée que le Cor, il n'est pas douteux qu'au grave le Tuba est plus puissant; mais l'on doit tenir compte aussi de la force du Cor *à l'aigu*. En outre, si solide que soit le Tuba, son timbre plus mou que celui des Trombones le fera toujours paraître moins puissant. - D'ailleurs, en matière d'intensité, la tessiture joue son rôle; et nous ne pouvions faire autrement que de reconnaître au *suraigu* de la *petite flûte*, une *intensité considérable*, mais tout en accordant que cette intensité n'est pas de même nature que celle du Trombone, ou des *Jeux d'anches* de l'Orgue. En somme, pour classer ces diverses intensités, on a cherché ce qui, dans l'orchestre, ressort le plus vigoureusement. - Et, là encore, c'est question d'espèce puisque la force sonore varie avec les instrumentistes.

Comme toutes les classifications, celles-ci sont contestables, sans doute. On souhaite seulement qu'elles se trouvent vérifiées par la majorité des cas, et qu'ainsi elles servent à *guider* l'élève vers la réalisation, toujours difficile, de l'équilibre orchestral.

⁽²⁾ Sax. dans le grave, fl. et cl. *à l'aigu*, moins facilement **p**, que les autres instruments de cette série, cor dans le grave, plus **p**.

FIG. 1.2 – Les échelles d'intensité et de volume dans le traité de l'orchestration de Koechlin [Koe54]

dans l'espace précédent donne des indications sur la fusion : des combinaisons qualifiées de nasales fusionnent moins bien que des combinaisons brillantes ou riches. Les auteurs notent que le degré de fusion est probablement corrélé au ratio entre le premier partiel et les autres. En effet, ils avaient précédemment montré que ce ratio (qu'ils appellent distorsion harmonique totale) est corrélé au degré de nasalité et donc au degré de fusion.

Dans une expérience similaire [San95] essaie de déterminer les paramètres sous-jacents à la perception de la fusion entre instruments. Les sons utilisés sont ceux de Grey [Gre77]. Il est demandé à des auditeurs musiciens de juger du degré de fusion (blend) entre les instruments présentés par paires. Les résultats sont ensuite corrélés à 9 descripteurs acoustiques :

- Le centroïde composite ou centroïde spectral du mélange,
- La différence de centroïde spectral moyen entre les deux sons,
- La corrélation entre les deux centroïdes spectraux au cours du temps,
- La corrélation entre les deux courbes de sonie,
- le contraste d'attaque : moyenne de la différence entre les 2 courbes de sonie entre 0 et 80 ms,
- La synchronie totale : corrélation entre les canaux d'une décomposition du son de la paire sur une base d'ERB.
- La synchronie d'attaque,
- La synchronie de crête,
- La synchronie d'arrêt des sons.

Lorsque les deux sons de la paire sont à l'unisson, le centroïde composite est le descripteur qui explique le plus de variance suivi du contraste d'attaque et de la corrélation de sonie. Lorsque les sons sont présentés en intervalle de tierce mineur, la variance est expliquée par (en ordre décroissant) : la différence de centroïde, le contraste d'attaque, le centroïde composite, et la synchronie d'arrêt.

Ces expériences montrent qu'il est possible de formaliser certains principes d'orchestration et comme le dit Bregman « *qu'il devrait être possible de rédiger un traité d'orchestration énonçant des principes assez abstraits pour être appliqués à tous les styles de musique. Ce ne serait pas un recueil de préceptes accumulant impératifs et interdits, mais une sorte de guide indiquant la marche à suivre pour obtenir tel ou tel son, et laissant toute liberté au compositeur de décider du résultat souhaité.* »

1.4 Orchestration assistée par ordinateur

Cette idée formulée par Bregman semble partagée par plusieurs compositeurs. Avec le développement de l'informatique musicale et du traitement du signal audio, l'idée de créer un programme permettant de trouver automatiquement des orchestrations commence à émerger. Généralement issus de la collaboration entre compositeurs et scientifiques, ces programmes sont tous basés sur le même principe. Le but est d'aider le compositeur à trouver une combinaison de sons instrumentaux dont le timbre s'approche le plus possible d'un timbre cible, c'est à dire de l'idée du compositeur. Ce but est spécifié par un son enregistré qui peut être de n'importe quel type (instrumental ou non), le programme cherche alors une combinaison de sons d'une base d'échantillons d'instruments qui s'approche le plus possible de la cible. Les programmes proposés utilisent différentes heuristiques pour trouver cette combinaison. Ils diffèrent aussi par le type d'analyse du signal utilisé pour décrire le timbre.

[Pse03] utilise un algorithme itératif pour trouver la combinaison de sons qui s'approche le plus de la cible. Chaque instrument est décrit par sa tessiture, un intervalle de nuances et une collection de pics spectraux extraits du spectre à court terme de différentes notes et différentes nuances. Le programme extrait les pics de la cible, puis cherche le son dont les pics sont les plus proches en fréquence et en amplitude. Le classement se fait par une distance euclidienne entre les amplitudes des pics de fréquence proche. Les pics du meilleur son sont ensuite soustraits à ceux de la cible et le programme itère. Il s'arrête lorsque l'amplitude résiduelle des pics de la cible est inférieur à un seuil donné. L'utilisation des pics spectraux permet un bon résumé de la structure harmonique des sons. [Hum05] utilise à peu près le même principe, mais décrit les sons par leur enveloppe spectrale. Ce choix de description rend le système plus efficace pour les sons bruités, comme la voix chuchotée, mais produit des résultats dont la structure harmonique peut être très différente de celle de la cible. Dans les deux cas, l'utilisation d'algorithmes itératifs permet de trouver rapidement une solution, mais ne facilite pas l'exploration des possibilités timbrales de l'orchestre. En effet, le résultat final dépend beaucoup des premiers sons choisis, le système favorise donc des solutions où le premier son est très proche de la cible au détriment de solutions où chaque son est peut être plus loin de la cible mais dont la combinaison sera plus satisfaisante. D'autre part ces deux études font apparaître un aspect intéressant du problème qui est le choix de la description du son. Ce choix détermine en grande partie les solutions accessibles, ainsi [Pse03] trouve des solutions intéressantes pour des sons de structure harmonique complexe tandis que la description utilisée dans [Hum05] est plus efficace pour des cibles bruitées.

Enfin, [RH05] utilise une approche assez différente. La description, toujours exclusivement spectrale, est le spectre moyen sur 4096 points. Il utilise une méthode basée sur la décomposition en valeurs singulières (SVD) pour trouver une somme pondérée de ces spectres qui approche la cible. C'est à dire si F est la matrice contenant les spectres de tous les sons de la base d'échantillons et f le spectre du son cible, alors le problème de l'orchestration peut être formulé de la manière suivante :

$$\hat{a} = \arg \min_a \|aF - f\|_2 \quad (1.1)$$

Où $a = (a_i)$ est un vecteur de pondération qui représente finalement l'amplitude de chaque son dans le mélange. Ce problème de moindre carré peut se résoudre par SVD, mais les a_i trouvés peuvent être négatifs ce qui ne correspond pas une réalité physique. Les auteurs ne donnent pas le détail de leur méthode. La solution est donc l'ensemble des sons qui ont les poids les plus élevés, mais le choix final est laissé à l'utilisateur. Cette méthode est très rapide pour un ensemble réduit de sons, mais devient impossible lorsque la taille de cet ensemble augmente. De plus il semble difficile de spécifier des contraintes comme l'incapacité d'un instrument donné de jouer deux sons en même temps.

Chapitre 2

Formulation et questions soulevées

2.1 Approche choisie

L'approche générale de l'orchestration assistée par ordinateur choisie dans cette thèse ne diffère pas beaucoup de celle des méthodes précédemment citées, mais nous allons voir que cette approche soulève des questions qui doivent être étudiées pour permettre la réalisation d'un programme réellement utilisable par des compositeurs.

Le principe de base est de trouver des combinaisons de sons instrumentaux dont le timbre se rapproche le plus possible du timbre voulu par un compositeur. Ce timbre est spécifié par un son enregistré. Le recours à un son comme représentant de la volonté du compositeur est la solution la plus simple et surtout celle qui permet de limiter au maximum les ambiguïtés. En effet, une autre solution aurait pu être de passer par une description verbale du timbre cible, mais les études sur la verbalisation du timbre ont montré qu'il était très difficile d'obtenir un ensemble partagé de mots décrivant le timbre [Fau00]. Nous aurions pu aussi chercher des méthodes permettant de spécifier la cible par un ensemble de valeurs de descripteurs précis comme la fréquence, l'amplitude des partiels ou la modulation de fréquence, mais cela revient finalement à créer un synthétiseur de son à partir de ces descripteurs, puis à utiliser le son synthétisé comme cible. Nous laissons donc libre choix au compositeur quant à la source de la cible.

Le système d'orchestration comprend trois entités principales (figure 2.1) :

- une base de connaissances composée des échantillons et éventuellement d'un ensemble de connaissances symboliques structurées concernant les instruments de musique (notes jouables, modes de jeu possibles ...),
- une cible spécifiant le son à atteindre,
- un algorithme de recherche qui explore l'espace des solutions pour trouver les meilleurs.

Cette formulation pose au moins cinq questions que nous détaillons dans la suite du chapitre : Comment décrire les sons ? Comment les comparer ? Comment prendre en compte la variabilité du son instrumental dans la description ? Comment combiner les descriptions des sons pour obtenir la description d'un mélange de sons ? Comment résoudre le problème combinatoire ?

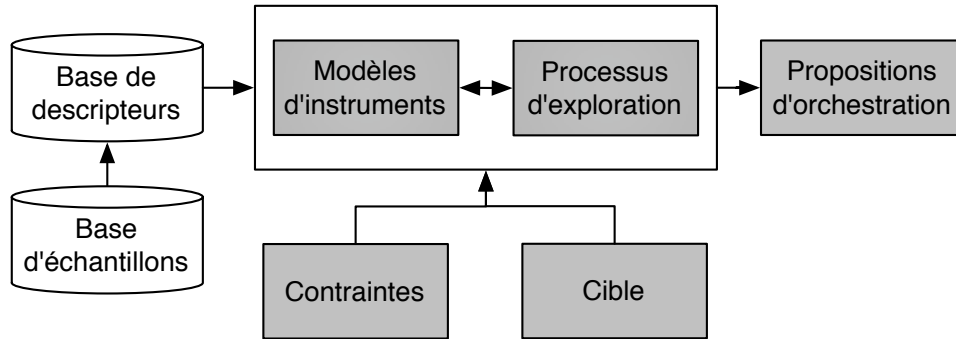


FIG. 2.1 – Schéma général du système

2.2 Décrire les sons

L'orchestration peut-être considérée comme l'art de la manipulation du timbre de l'orchestre. Cette manipulation n'est possible que grâce à une connaissance approfondie du timbre des différents instruments de l'orchestre. Une approche scientifique et informatique de ce problème suppose de donner cette connaissance à l'ordinateur et donc, en premier lieu de disposer d'une description quantitative du timbre. Cette description a fait l'objet d'un grand nombre de recherches dans les domaines de la perception, de la synthèse et de l'analyse des sons.

2.2.1 Approche psychoacoustique

La première étude formelle du timbre est due à Helmholtz [Hel98] qui a montré que le timbre est lié à la composition spectrale du son, c'est à dire à l'importance relative des différents harmoniques du son dans le cas des sons périodiques. Il a aussi suggéré que certaines particularités temporelles du son comme la manière dont il commence ou fini peuvent avoir une importance sans identifier précisément ces caractéristiques. Il faut ensuite attendre l'avènement de l'analyse et de la synthèse informatique des sons et de nouvelles méthodes expérimentales pour préciser et dépasser ces idées [RM69].

D'un autre coté, des méthodes de psychologie cognitive sont appliquées à l'analyse de la perception du timbre [Gre77]. Le but est de déterminer quels sont les indices utilisés par un ensemble de sujets pour juger de la similarité entre deux sons. Dans ces expériences, les participants notent le degré de similarité qu'ils attribuent à différentes paires de sons, puis une analyse MDS (Multi-dimensional Scaling) est appliquée aux jugements. Ce type d'analyse permet d'identifier les éventuelles dimensions sous-jacentes aux jugements de similarités. Les différentes études basées sur des sons instrumentaux simples trouvent en général des espaces à trois dimensions. Il est ensuite possible de chercher à relier les dimensions perceptives trouvées à des descriptions acoustiques du son. Cette recherche se fait d'abord de manière intuitive [Gre77] puis de manière plus systématique en cherchant parmi un ensemble de descripteurs du son ceux qui sont le mieux corrélés aux dimensions [Kri93], [MWD⁺95], [PMH00]. Ces études ont montré que le logarithme du temps d'attaque et le centre de gravité du spectre sont très corrélés aux deux premières dimensions et donc sans doute utilisés dans les jugements de similarité. La troisième dimension à d'abord été associée à la varia-

tion temporelle du spectre, mais des études plus récentes ont montré que l'étendue du spectre était une meilleure explication [PMH00]. Ce genre de méthode s'est donc révélée efficace pour découvrir les attributs utilisés dans la comparaison de sons instrumentaux. Cependant, elles ne permettent pas d'aller au-delà de ces attributs de base. Il est en effet évident que, même si ces attributs ont une grande importance dans la description du timbre, ils n'en décrivent pas la totalité : des sons différents peuvent avoir le même centroïde spectral et le même temps d'attaque. Pour répondre à ces questions [GM77], [Cha81] et [MBM99] utilisent une autre méthode basée sur l'analyse et la synthèse des sons. Un ensemble de sons instrumentaux est analysé, les paramètres issus de l'analyse sont ensuite simplifiés de plusieurs manières et ces simplifications sont utilisées pour la synthèse. L'ensemble de ces sons et de leurs simplifications est enfin utilisé dans des tâches de discrimination. Les simplifications proposées sont par exemple la diminution ou l'annulation des petites variations de l'amplitude des harmoniques au cours du temps ou le lissage de l'enveloppe spectrale. Ce type d'expérience permet à [MBM99] de montrer que l'annulation des petites variations temporelles de la fréquence des harmoniques est peu perçue pour les sons étudiés alors que les variations temporelles de l'amplitude le sont beaucoup plus. Elle permet aussi de déduire une formule pour la distance entre un son et ses diverses simplifications. Cette formule est basée sur la somme pondérée du log de la différence des centroïdes spectraux et de la différence RMS entre les amplitudes des harmoniques.

Toutes ces recherches permettent de mettre en lumière un ensemble de descripteurs utiles pour la comparaison du timbre de deux sons quasi-stationnaires et harmoniques. En résumé, ces descripteurs sont le logarithme du temps d'attaque, le centroïde et l'étendue spectrale, l'amplitude des harmoniques et la variation temporelle de cette amplitude. Cet ensemble, quoique déjà assez vaste, ne permet pas de décrire la totalité des sons instrumentaux. La complexité des protocoles expérimentaux rend la recherche de description plus complète très difficile. L'émergence des problématiques de reconnaissance automatique des sons et d'extraction d'information de la musique (MIR) va donner un autre point de vue sur le problème.

2.2.2 Approche computationnelle

La classification automatique des sons d'instruments de musique émerge à la faveur du développement des grandes bases d'échantillons et découle de la volonté de pouvoir annoter automatiquement ces enregistrements. Elle est aussi une première étape vers l'extraction automatique d'informations musicales comme la transcription ou la séparation de sources. Le but est donc de créer des algorithmes capables de reconnaître automatiquement les instruments contenus dans un enregistrement. Pour cela, deux ingrédients sont nécessaires : d'abord une description adéquate du son, puis l'algorithme de classification à proprement parlé qui utilise la description d'un son pour prendre une décision. Le problème n'est donc plus ici de savoir quels attributs du son sont perçus par un auditeur, mais quels attributs permettront à une machine de distinguer deux instruments. La question de savoir si ces attributs ont une importance perceptive est secondaire. Ainsi, les différentes études ([Mar99], [Ero01a], [BHM01]) utilisent des descripteurs issus de la reconnaissance automatique de locuteurs comme les coefficients cepstraux, de la psychoacoustique ou des descripteurs ad hoc comme l'amplitude et la fréquence des modulations d'énergie et de fréquence fondamentale. Pour une liste détaillée, on pourra se référer à [HBKD06] et à [Pee04]. L'intérêt de ce genre de méthodes réside donc dans le fait qu'elles permettent l'utilisation, le développement et la vérification d'un grand nombre de

descripteurs. Même si il n'a pas été formellement démontré que ces descripteurs constituaient des dimensions de la perception, ils mesurent généralement des caractéristiques audibles. De plus certains d'entre eux d'abord utilisés pour la classification, se sont ensuite avérés très bien corrélés à des dimensions perceptives.

2.3 Comparer des sons

2.3.1 Qu'est ce que la similarité

Une fois la description du son choisie il faut pouvoir comparer les sons à l'aide d'une mesure de similarité. Malgré son apparente simplicité, le concept de similarité est assez complexe à appréhender. Il en existe plusieurs modèles. Nous présentons les deux principaux (pour plus de détails voir [GS05]).

Le premier est le *modèle géométrique* dans lequel la similarité entre deux objets est inversement proportionnelle à la distance entre leur représentation dans un espace multidimensionnel. La distance entre deux objets i et j est définie par :

$$d(i, j) = \left(\sum_{k=1}^K |X_{ik} - X_{jk}|^r \right)^{1/r} \quad (2.1)$$

avec K la dimension de l'espace, X_{ik} la k -ième coordonnée de i , r généralement égal à deux, l'indice de la norme. Ainsi un objet peut être représenté dans un espace multidimensionnel où chaque dimension correspond à un attribut mesurable de l'objet. Une dimension est généralement définie comme une variable à laquelle peut être assignée une valeur et comme un attribut psychologique qui peut être utilisé pour ordonner des stimuli d'une manière cohérente [GT82]. Ce modèle est utilisé pour décrire la similarité entre deux sons (voir paragraphe 2.2.1). Dans notre cas, les dimensions sont les différents descripteurs et la similarité est déduite de la distance dans cet espace de descripteurs.

Le second modèle est le *contrast model* introduit par Amos Tversky en 1977 [Tve77]. Il remet sérieusement en cause le modèle géométrique. Tversky explique que les jugements de similarité ne suivent pas toujours les propriétés de base du modèle géométrique et qu'il est possible de trouver des expériences qui violent chacune de ses propriétés. De plus, ce modèle ne prédit pas qu'en ajoutant une qualité commune à tous les objets, on augmente la similarité perçue. En effet, ajouter une dimension dans l'espace de représentation sur laquelle tous les objets ont la même coordonnée ne modifie pas la distance entre les objets. Considérant tous ces problèmes, Tversky propose le *contrast model* dans lequel les objets sont représentés par un ensemble de qualités. La similarité entre deux objets A et B est exprimée par une fonction de trois arguments : les qualités communes de A et B , les qualités qui appartiennent à A et pas à B et enfin celles qui appartiennent à B et pas à A . La fonction peut par exemple être une application linéaire dans laquelle les poids permettent de décider de l'importance relative des qualités communes et non communes :

$$S(a, b) = \theta f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A), \quad \theta, \alpha, \beta \geq 0 \quad (2.2)$$

Il n'y a pas de restriction sur ce que peut être une qualité. Cela peut-être n'importe quelle propriété de l'objet. Bien que des méthodes aient été proposées pour représenter des dimensions continues, le

contrast model est plus adapté à la représentation de qualités discrètes. Dans notre cas cela peut-être “contient un la4” ou “est modulé en amplitude”. L’intérêt d’une telle approche réside dans le fait qu’elle permet d’accorder une importance explicite aux caractéristiques communes et différentes et de spécifier les poids. Cela donne un point de vue très différent sur la similarité entre la cible et une orchestration.

Dans l’approche géométrique, il faut calculer la position d’une orchestration dans l’espace puis comparer cette position à celle de la cible, ce qui impose en théorie de calculer la position de toutes les combinaisons de sons possibles. Dans une approche de type *contrast model*, s’il est possible de prédire que certaines qualités d’un son se propagent à l’orchestration qui le contient, alors nous pouvons éliminer certains sons de la base et donc diminuer considérablement le nombre d’orchestration à évaluer. Ainsi, si nous considérons la qualité “contient un la4”, nous pouvons supposer que cette qualité se propage à une orchestration, c’est à dire que si une orchestration contient un son de la4, alors cette orchestration contiendra un la4. Si notre cible ne contient pas de la4 nous pouvons enlever tous les sons contenant un la4 dans la base. Cela se traduit par l’affectation d’un poids très élevé (infini) aux qualités distinctes des sons dans l’équation 2.2.

2.3.2 Variabilité de la similarité

Un problème important dans la compréhension de la similarité est sa variabilité. Les jugements de similarité entre deux objets peuvent varier selon l’âge, l’expertise, l’environnement ou la méthode de présentation [GS05]. Ainsi, [Tve77] montre que la Suède est plus souvent choisie que la Hongrie comme étant similaire à l’Autriche dans l’ensemble {Suède, Autriche, Hongrie} alors que c’est le contraire dans l’ensemble {Suède, Norvège, Hongrie}. [Tve77] l’explique par l’effet de diagnosticité. Selon cette théorie, l’importance d’une qualité dans la similarité dépend de son efficacité pour la catégorisation. Dans le premier cas de l’exemple précité, la qualité qui ressort est “type de gouvernement” alors que dans le second cas c’est la qualité “scandinave”. La similarité change car elle est évaluée sur des critères différents. Cet exemple montre qu’il est difficile d’établir une relation de similarité unique, il montre aussi que des qualités efficaces pour la catégorisation acquièrent une importance particulière dans l’évaluation de la similarité.

Les études sur la perception du timbre montrent aussi que les poids affectés aux différentes dimensions dans une tâche de jugement de similarité varie selon les individus ou les classes d’individus. Ainsi [MWD⁺95] utilise des classes latentes dans la MDS et trouve plusieurs classes d’utilisateurs, dont une équi-pondère toutes les dimensions tandis que les autres affectent des poids élevés à seulement une ou deux dimensions. Cela montre ainsi que certains participants n’avaient pas utilisé toutes les dimensions dans leurs jugements. De même [DMW94] montre que dans certaines conditions, les jugements de similarité varient selon le contexte dans lequel les sons sont présentés. Des sons d’instruments à vent sont jugés moins similaires entre eux quand ils sont présentés seuls qu’en présence de sons plus percussifs. Elle montre aussi que le fait de diminuer l’intervalle de variation le long d’une dimension diminue l’importance de cette dimension. Toutes ces expériences mettent en évidence la variation de la perception de similarité selon les personnes mais aussi pour une même personne selon le contexte.

2.3.3 Proposition de mesure de similarité

Il semble donc difficile, sinon impossible, de proposer une distance de type euclidienne à pondérations fixes. Nous proposons donc de ne pas agréger les différents descripteurs dans une seule distance et d'explorer l'espace des solutions selon différentes pondérations. Cette méthode est liée à l'optimisation multicritères. Dans notre cas nous parlerons de préférences plutôt que de contexte.

2.4 Généralisation du savoir instrumental

Une autre question est soulevée par le problème de l'orchestration tel que nous l'avons formulé, celle de la variabilité inhérente aux sons instrumentaux et donc de la généralisation des connaissances extraites des bases d'échantillons. En effet, le son d'un instrument varie selon le modèle considéré (facteur, luthier, marque) et l'instrumentiste qui le joue. Dans le cas d'enregistrement, les conditions (la salle) et les méthodes d'enregistrement (position des micros, matériel utilisé) ont aussi une influence. Tous ces facteurs font qu'il ne peut exister d'enregistrement de référence d'un instrument. Ainsi dans la classification automatique des instruments de musique [LR03] et [Pee03] constatent que l'utilisation d'une seule base pour apprendre les caractéristiques des instruments donne de mauvais résultats pour la reconnaissance de sons venant d'une autre base. Il montre qu'en utilisant plusieurs bases différentes pour l'apprentissage, les résultats sont nettement améliorés, ils passent ainsi de 30% à 57% de bonne reconnaissance pour [Pee03]. Outre les variations inhérentes précédemment décrites, le son émis par un instrument varie selon le geste réalisé par l'instrumentiste. Nous ne parlons pas ici des différents modes de jeu d'un instrument mais de la manière dont l'instrumentiste réalise ces modes de jeu. Ainsi pour un vibrato, l'instrumentiste décide de la fréquence et de l'ambitus, le son d'un ponticello de violon sera différent selon la position de l'archet sur la corde. Il semble pertinent d'essayer d'éviter au maximum d'apprendre les caractéristiques précises du son tel qu'il a été réalisé au moment de l'enregistrement en généralisant la connaissance disponible sur les instruments et les modes de jeu. Une solution à ce problème est donc de diversifier l'ensemble des sons servant à l'apprentissage. Mais ceci oblige à passer par une phase explicite d'apprentissage qui permet d'organiser la connaissance contenue dans la base. Nous voyons ici l'hypothèse implicite contenue dans les systèmes d'orchestration précédent : un échantillon représente le son de l'instrument. A l'opposé nous supposons qu'il est nécessaire de donner à un système d'orchestration une connaissance plus générale sur les instruments. Pour cela nous utilisons des algorithmes d'apprentissage capable d'organiser la connaissance extraite de bases contenant plusieurs instances d'un même son venant de sources différentes. On entend par instance du même son, deux sons produits par le même instrument avec le même mode de jeu, la même hauteur et la même nuance.

De telles méthodes sont utilisées dans le traitement automatique de la parole, la reconnaissance automatique d'instruments de musique, la transcription ou la séparation de sources. Les méthodes utilisées pour ces tâches sont généralement des modèles probabilistes génératifs. D'autres méthodes telles que les K plus proches voisins (KNN) ou les machines à vecteurs de support (SVM) sont aussi utilisées mais ne sont pas directement liées à notre propos, nous n'en parlerons donc pas. Pour une description détaillée de ces méthodes dans le cadre de la classification des instruments de musique, voir [HBKD06]. [Pee03], [Ero01b], [BHM01] et [Bro99] utilisent des modèles gaussiens ou des mélanges de gaussiennes pour approximer la distribution jointe des descripteurs des échantillons

de chaque instrument. L'utilisation de mélanges de gaussiennes permet d'approximer des distributions arbitrairement complexes, mais au prix d'une augmentation du nombre de paramètres et d'une difficulté accrue dans le calcul des modèles de combinaison de sons [BBG06]. En revanche, les modèles gaussiens simplifient à la fois l'apprentissage et le calcul des modèles des combinaisons. Cependant, ils rendent nécessaire le prétraitement des données pour rendre les distributions plus proches de la loi normale.

2.5 Combiner des descripteurs et les modèles

Pour comparer une combinaison avec la cible, il faut connaître les descripteurs de cette combinaison. Une première approche consiste à calculer les descripteurs directement sur des échantillons contenant ces combinaisons [ERD06], comme des enregistrements de musique orchestrale, précédemment annotés. Outre le problème de l'annotation, il semble difficile d'obtenir des exemples pour un nombre élevé de combinaisons, d'analyser ces exemples et de conserver les analyses pour les comparer à la cible. Une autre méthode consiste à générer des mélanges qui satisfont les contraintes (nombre et type d'instruments disponibles, notes possibles ...) et de calculer les descripteurs à la volée [RV07a] [RV07b]. Il en résulte un temps de calcul dépassant de beaucoup la patience d'un utilisateur. Finalement, la démarche permettant d'obtenir, en un temps réduit, la description d'un maximum de combinaisons avec une base d'échantillons de taille raisonnable repose sur l'estimation des descripteurs d'une combinaison à partir des descripteurs de ses composants. Le premier problème dans ce cas est de savoir "additionner" les descripteurs. Pour les descripteurs spectraux, on suppose généralement une addition des spectres de puissance [Vin06], [BBG06], [Gal98]. Pour les autres, des heuristiques devront être trouvées. A partir de ces formules d'addition de descripteurs, et si les descripteurs sont modélisés par des distributions de probabilité, il faut déterminer comment se combinent les distributions. Ainsi, dans le cas où les spectres de puissance sont modélisés par des gaussiennes, la combinaison est décrite par une gaussienne dont les paramètres sont trouvés par addition des moyennes et des variances [BBG06]. En revanche si on modélise le spectre de log-puissance il faut utiliser des formules approximatives permettant de trouver les paramètres du modèle de la combinaison [Gal98].

2.6 La combinatoire

Enfin la dernière question vient de l'aspect intrinsèquement combinatoire du problème. En simplifiant, prenons un petit ensemble de douze instruments. Chaque instrument peut jouer 12 notes sur 3 octave, c'est à dire 36 notes, dans 3 nuances différentes. De plus mettons que chacun de ces instruments puisse utiliser 3 modes de jeu différents. Chaque instrument peut donc jouer 324 sons différents. Le problème du choix d'une combinaison est finalement de choisir un son parmi 325 (324 plus la possibilité que l'instrument ne joue pas) pour les 12 instruments. Cela donne un ensemble de possibilités dont le cardinal est de l'ordre de 10^{30} ce qui est évidemment intraitable de manière exhaustive. Il faut donc trouver des heuristiques de résolution. Ce problème est traité dans la thèse de Grégoire Carpentier. Une description de l'approche choisie est aussi donnée dans [CTA⁺07].

2.7 Description du système d'orchestration

2.7.1 Introduction

Les paragraphes précédents nous permettent maintenant de préciser la forme de notre système d'orchestration. Il est composé d'une base d'échantillons de sons instrumentaux desquels sont extraits un ensemble de descripteurs du timbre. Des modèles probabilistes, entraînés sur ces descripteurs, décrivent les instruments et leurs modes de jeu. Une cible est donnée par l'utilisateur sous forme de son enregistré. Les descripteurs sont extraits de cette cible. Un processus d'exploration est chargé de parcourir l'ensemble des combinaisons pour trouver celles qui se rapprochent le plus de la cible. L'évaluation de la distance entre la cible et une combinaison se fait de manière multicritères à l'aide des modèles d'instruments.

2.7.2 Un problème multicritères

La variabilité de la similarité nous incite à traiter le problème de l'orchestration comme un problème d'optimisation multicritères.

2.7.2.1 Optimisation multicritères

Dans le contexte multicritères l'évaluation d'une solution se fait en regardant séparément chaque critère (ou descripteur), c'est à dire qu'aucune fonction d'agrégation, comme une somme pondérée par exemple, n'est définie a priori. Ainsi, un problème d'optimisation multicritères est défini comme :

$$\min_{c \in \mathcal{K}} (f_1(c), \dots, f_J(c)), \quad (2.3)$$

Où $\mathcal{K}$ est l'ensemble des solutions faisables et f est une fonction objectif vectorielle. Dans notre cas, c représente une orchestration possible, c'est à dire une combinaison de sons instrumentaux définie par les instruments, les hauteurs, les notes et les modes de jeu utilisés. $\mathcal{K}$ représente l'ensemble de toutes les orchestrations faisables. f_j représente l'évaluation de la combinaison c selon le descripteur j . Une solution c est jugée optimale si elle est non-dominée dans l'espace des critères, c'est à dire qu'il n'existe pas de solution c' au moins aussi bonne sur tous les critères et strictement meilleure sur au moins un critère. L'ensemble des solutions non dominées est appelé *front de Pareto* et les solutions du front sont dites de rang 1. Si on retire les solutions appartenant au front de Pareto, les nouvelles solutions non dominées sont dites de rang 2. La figure 2.2 montre le front de Pareto et les solutions de rang 2 pour un problème à deux critères. Toutes les solutions du front sont également optimum pour le problème considéré. Évidemment, pour un problème combinatoire le front de Pareto théorique est impossible à extraire puisqu'il faudrait évaluer toutes les solutions possibles. Le but de l'algorithme d'exploration est donc de trouver des solutions appartenant à ce front ou en étant le plus proche possible et couvrant le plus possible le front.

Le front de Pareto constitue l'ensemble des orchestrations "proches" de la cible dans la mesure où nous ne connaissons pas les préférences de l'utilisateur concernant le poids des descripteurs. Par exemple s'il attache plus d'importance aux critères spectraux qu'aux critères temporels.

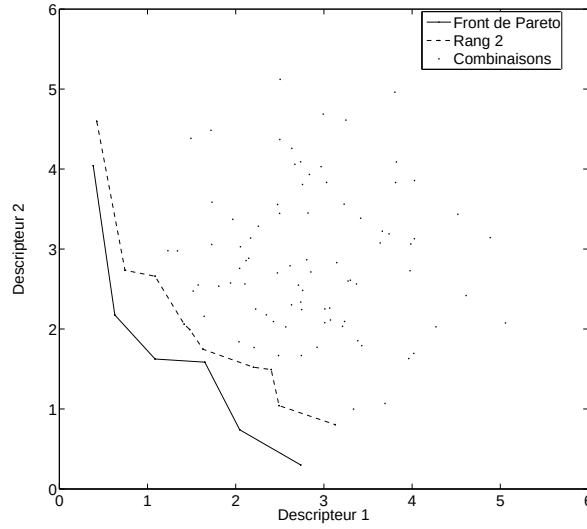


FIG. 2.2 – Exemple de front de Pareto

2.7.2.2 Possibilité d'interaction

Une fois le front approximatif obtenu par le processus d'exploration, il est possible d'entrer dans une phase d'interaction avec l'utilisateur. Par exemple en lui présentant quelques solutions bien réparties sur le front et en lui demandant d'en choisir une. Le choix d'une solution préférée donne indirectement le jeu de poids préféré de l'utilisateur. Il est ainsi possible de relancer une exploration monocritère en agrégeant les critères par une somme pondérée par les poids trouvés ou une exploration multicritères sur un "cône" de poids plus réduits et donc d'obtenir un ensemble de solutions basé sur les préférences de l'utilisateur.

2.7.3 Un problème probabiliste

Compte tenu des problèmes de généralisation évoqués au paragraphe 2.4, nous utilisons une méthode de modélisation probabiliste pour décrire le timbre. Le but est d'obtenir la distribution des descripteurs sachant l'instrument et la manière de le jouer. Comme nous nous plaçons dans un contexte multicritères, nous obtenons une distribution par descripteur :

$$f(d_j|x_m) \quad (2.4)$$

où d_j est un descripteur et x_m représente l'instrument. A partir de ces distributions nous calculons la distribution du descripteur pour une combinaison donnée :

$$f(d_j|x_1, \dots, x_M) = G_j(f(d_j|x_1), \dots, f(d_j|x_M)) \quad (2.5)$$

où G_j est la fonction qui permet de combiner les modèles du descripteur d_j de chaque son pour obtenir celui de la combinaison.

2.7.4 Définition formelle du problème de l'orchestration

Finalement, nous pouvons définir formellement le problème de l'orchestration comme suit :

$$\arg \max_{c \in \mathcal{K}} (f(d_j = d_j^t | c = \{x_1, \dots, x_M\}))_{j=1..J} \quad (2.6)$$

$$\text{avec } f(d_j | c) = G_j(f(d_j | x_1), \dots, f(d_j | x_M)) \quad (2.7)$$

où J est le nombre de descripteurs, c est une combinaison composée des sons $x_1 \dots x_M$, (d_j^t) représente les descripteurs de la cible et G_j est la fonction d'addition de modèle du descripteur j . $f(d_j = d_j^t | c = \{x_1, \dots, x_M\})$ représente la probabilité d'obtenir la valeur d_j^t pour le descripteur j avec la combinaison c . Cette probabilité est évaluée sur la densité $f(d_j | c)$ qui est elle-même obtenue grâce à la fonction G_j à partir des modèles de chaque son.

Dans cette thèse, nous traitons de l'évaluation des combinaisons d'instruments par rapport à la cible. Nous ne parlons pas de méthodes d'exploration de l'espace des solutions.

Chapitre 3

Instruments et modes de jeu considérés

3.1 Introduction

Les instruments et leurs divers modes de jeu tiennent une place centrale dans ce travail. Le développement de méthodes d'orchestration assistée par ordinateur suppose avant tout de donner à l'ordinateur une connaissance approfondie du timbre des instruments. Chaque instrument peut produire une grande variété de timbres selon la volonté de l'interprète et l'espace de ces variations et généralement continu. Pour les besoins de l'écriture cet espace a été divisé en catégorie. Ce sont ces différentes catégories que nous appellerons modes de jeu ou techniques de jeu. Ainsi, la position de l'archet sur la corde d'un violon peut varier continûment mais ce continuum a été divisé en trois positions : proche du chevalet (sul ponticello), proche de la touche (sul tasto) et au milieu (ordinario). La catégorisation en modes de jeu est basée soit sur le geste effectué pour obtenir le son, soit sur le son lui-même. La position de l'archet relève du premier cas, alors que le vibrato relèverait plutôt du second : un effet sonore similaire (une variation périodique de la fréquence et de l'amplitude du son) est obtenu par des gestes différents selon les instruments. Nous nous attachons à expliciter le lien entre le timbre et le mode de jeu correspondant. En terme informatique, cela revient à faire de la classification automatique. Nous utiliserons beaucoup les noms des différents modes de jeu, nous les décrivons donc dans ce chapitre. Plus de détails sont disponibles dans les traités d'instrumentation ou sur le site [Hug]. Nous présentons aussi les bases d'échantillons qui seront utilisées comme ensemble d'apprentissage et de test pour les algorithmes proposés. La grande variété des modes de jeu et la distribution inégale de ces techniques dans les bases d'échantillons nous obligerons à proposer des méthodes d'apprentissage basées sur une taxinomie des modes de jeu que nous décrivons à la fin du chapitre.

3.2 Description des instruments et des modes de jeu

3.2.1 Vue d'ensemble

Nous étudions les instruments pouvant émettre des sons entretenus. Ils se divisent en quatre famille : les cuivres, les cordes, les anches doubles et les vents. Les figures 3.1, 3.2, 3.3 et 3.4 montrent les instruments de ces familles que nous étudierons. Les cases grisées représentent les familles réduites introduites pour faciliter la lecture mais que nous n'utiliserons pas. Le concept de famille est

important car souvent un mode de jeu est réalisable par tous les membres d'une même famille. La liste des abréviations utilisées pour les instruments est donnée en annexe C.

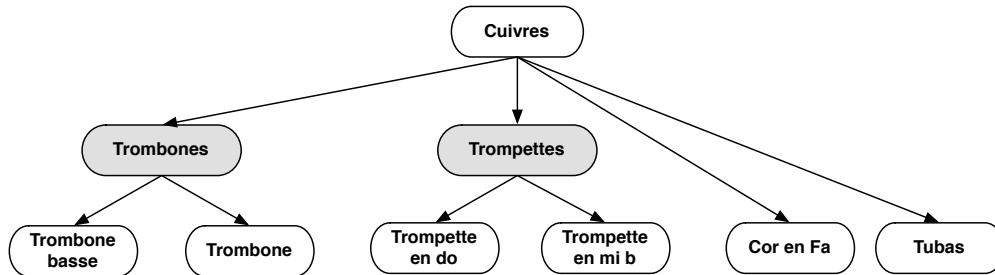


FIG. 3.1 – La famille des cuivres

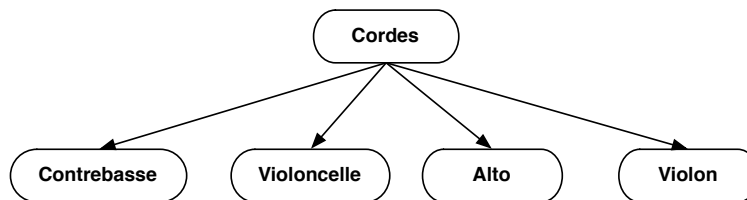


FIG. 3.2 – La famille des cordes

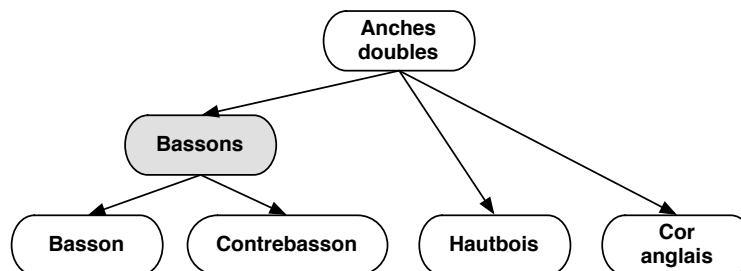


FIG. 3.3 – La famille des anches doubles

3.2.2 Les modes de jeu des cordes

Les sourdines

Les sourdines atténuent le son et changent le timbre de l'instrument. Nous en considérons deux types : la sourdine en caoutchouc et la sourdine de plomb.

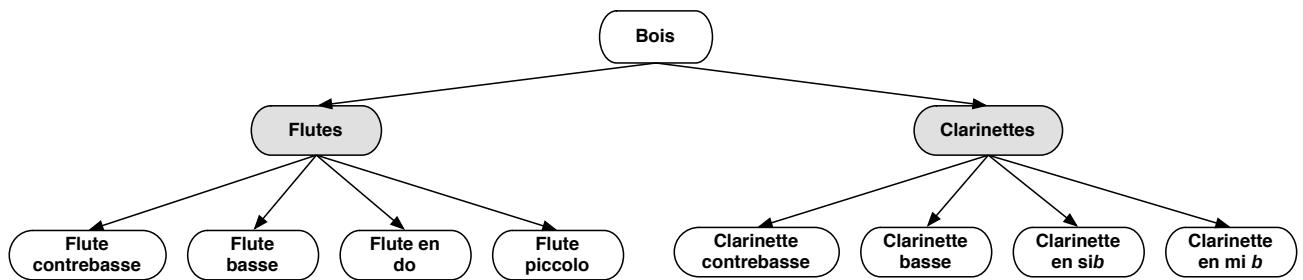


FIG. 3.4 – La famille des bois

Le vibrato

Le vibrato est réalisé par un mouvement du doigt et du poignet de la main gauche. L'effet produit est une modulation périodique de la hauteur et de l'amplitude du son. L'amplitude et la fréquence du vibrato des cordes frottées a été étudiée par [DHAT00], [FS67] et [MW00]. La fréquence observée est généralement de l'ordre de 6 Hz et peut aller jusqu'à 12 Hz. L'amplitude observée par [MW00] est de l'ordre de 15 cents.

Le tremolo

Le tremolo¹ est réalisé par des allers-retours rapides de l'archet. L'effet produit est une modulation périodique de l'amplitude du son. La fréquence de modulation mesurée par [Moe04] est du même ordre que celle du vibrato, quoique légèrement plus rapide dans les tempi élevés.

Le pizzicato

Le pizzicato² consiste à jouer les cordes avec les doigts de la main droite au lieu d'utiliser l'archet. Le son n'est donc plus entretenu mais percussif.

La position de l'archet

La position de l'archet sur la corde donne lieu à trois modes de jeu différents. La position la plus proche du chevalet est appelée sul ponticello³, le son obtenu est bruité avec des harmoniques qui peuvent ressortir de manière d'autant plus incontrôlée que l'on se rapproche du chevalet. La position proche de (ou sur) la touche est appelée sul tasto⁴ et produit un son plus sourd que le mode normal.

La technique d'archet

L'archet peut être utilisé avec la crin (normal) ou avec le bois. Le cas où la corde est frottée avec le bois est appelé col legno tratto. Le son obtenu est bruité et irrégulier. Il est généralement moins fort

¹www.mti.dmu.ac.uk/~ahugill/manual/audio/Violin/3007.mp3

²www.mti.dmu.ac.uk/~ahugill/manual/violin/movies/plucking.mov

³www.mti.dmu.ac.uk/~ahugill/manual/audio/Violin/1833.mp3

⁴www.mti.dmu.ac.uk/~ahugill/manual/audio/Violin/1880.mp3

que le son obtenu avec le crin⁵.

Les harmoniques artificiels

Les harmoniques artificiels sont produits en posant le doigt normalement sur la corde puis en effleurant du petit doigt la corde au point correspondant à la quarte⁶ ou à la quinte⁷ supérieur. La note produite dans le premier cas est deux octaves au dessus de la note correspondant au premier doigt et dans le second cas une octave et une quinte au-dessus. D'autres harmoniques peuvent être obtenus mais sont très rarement utilisés.

3.2.3 Les modes de jeu des cuivres

Le vibrato

Le vibrato⁸ des cuivres est réalisé avec les lèvres ou le conduit vocal, il peut aussi être obtenu avec la coulisse dans le cas du trombone.

Le flatterzunge

Le flatterzunge⁹ est obtenu en faisant rouler la langue sur le palais, le son résultant est une modulation rapide de l'amplitude. La fréquence de cette modulation peut aller jusque 50 Hz [SdS04].

Les sourdines

Il existe beaucoup de sourdines différentes pour les trompettes et les trombones¹⁰. Nous en étudions quatre : les sourdines sèche (straight), wawa, bol (cup), et harmon. Une seule est en revanche utilisée pour les tubas.

3.2.4 Les modes de jeu des bois

Le vibrato

Voir les cuivres.

Le flatterzunge

Voir les cuivres

⁵www.mti.dmu.ac.uk/~ahugill/manual/violin/movies/collegno.mov

⁶www.mti.dmu.ac.uk/~ahugill/manual/violin/movies/artharm.mov

⁷www.mti.dmu.ac.uk/~ahugill/manual/violin/movies/artharm2.mov

⁸www.mti.dmu.ac.uk/~ahugill/manual/trombone/movies/vibrato.mov

⁹www.mti.dmu.ac.uk/~ahugill/manual/trombone/movies/flz.mov

¹⁰www.mti.dmu.ac.uk/~ahugill/manual/trumpet/mutes.html

Aeolien

Cette technique consiste à varier la manière de souffler dans l'instrument (l'angle du jet d'air dans la flûte par exemple) pour obtenir un son avec plus ou moins de souffle. Dans les bases d'échantillons, cela donne lieu à deux modes de jeu : *aeolian* qui est très soufflée et *aeolian+ordinario* qui l'est moins.

3.2.5 Les modes de jeu des anches doubles**Le vibrato**

Voir les cuivre

Le flatterzunge

Voir les cuivres

3.2.6 Combinaisons de modes de jeu

Les modes de jeu décrits précédemment peuvent être combinés entre eux à certaines conditions, généralement d'ordre gestuel. Si les gestes réalisés pour obtenir deux modes de jeu sont réalisables en même temps, alors, les modes de jeu peuvent être combinés. Ainsi, un tremolo peut être réalisé avec différentes positions de l'archet, un flatterzunge peut être aussi aéolien.

3.3 Bases d'échantillons utilisées et répartition

La base d'échantillons est un des éléments principaux du système d'orchestration proposé, puisque c'est elle qui détermine l'ensemble des sons disponibles. Elle revêt d'autant plus d'importance que nous nous plaçons dans le cadre de l'apprentissage et de la classification automatique. La base doit ainsi être assez grande pour permettre un bon apprentissage et doit contenir des sons venant de plusieurs sources pour permettre l'évaluation des capacités de généralisation des algorithmes proposés. Par source, nous entendons musicien différent avec un instrument différent dans une salle différente. Ainsi, pour chaque classe à apprendre, par exemple un instrument donné, il faut disposer de beaucoup d'échantillons de cet instrument provenant de différentes sources. La mise en place d'une telle base d'échantillons est, en soi, un problème difficile et la difficulté augmente quand on augmente la précision, et donc le nombre, des classes à apprendre. Pour l'orchestration, nous souhaitons apprendre les différents modes de jeu des instruments, nous créons une base spécifique par la réunion de plusieurs bases.

Nous utilisons cinq bases d'échantillons : Studio Online [BB99], [Irc], Vienna Symphonic Library [Vie], Iowa [Uni], RWC [Got04] et Virtual Orchestra. La figure 3.5 montre la répartition des instruments dans les différentes bases. Nous disposons au total de 21 instruments, mais la répartition est très déséquilibrée. D'une part, les cordes sont mieux représentées car il y a plus de modes de jeu pour les cordes. D'autre part certains instruments ne sont contenus que dans une ou deux bases. Ainsi la flûte contre basse n'est présente que dans Virtual Orchestra, le cor anglais n'est présent que dans

Iowa et Vienna Symphonic Library. Le problème est le même pour les modes de jeu (voir figure 3.6) avec encore plus de déséquilibre. Beaucoup de modes de jeu ne sont contenus que dans Vienna Symphonique Library ou Studio Online. En résumé, chaque base apporte des caractéristiques différentes à l'ensemble. SOL, Vienna et Virtual Orchestra contiennent beaucoup de modes de jeu différents et rares alors que Iowa, Virtual Orchestra et Vienna contiennent beaucoup d'instruments. Un autre problème vient de la répartition des nuances et des hauteurs. Comme le montre la figure 3.7 certains modes de jeu ne sont enregistrés que pour une dynamique, de plus il est fréquent de ne disposer que d'un nombre limité de hauteurs qui couvrent l'ensemble de la tessiture. Le but de notre système d'apprentissage va donc être d'apprendre le mieux possible les différents timbres des instruments à partir de ces bases très déséquilibrées. Pour cela nous proposons d'abord une organisation ou taxinomie des instruments et modes de jeu, nous montrerons ensuite comment tirer parti de cette taxinomie pour apprendre efficacement les timbres des instruments.

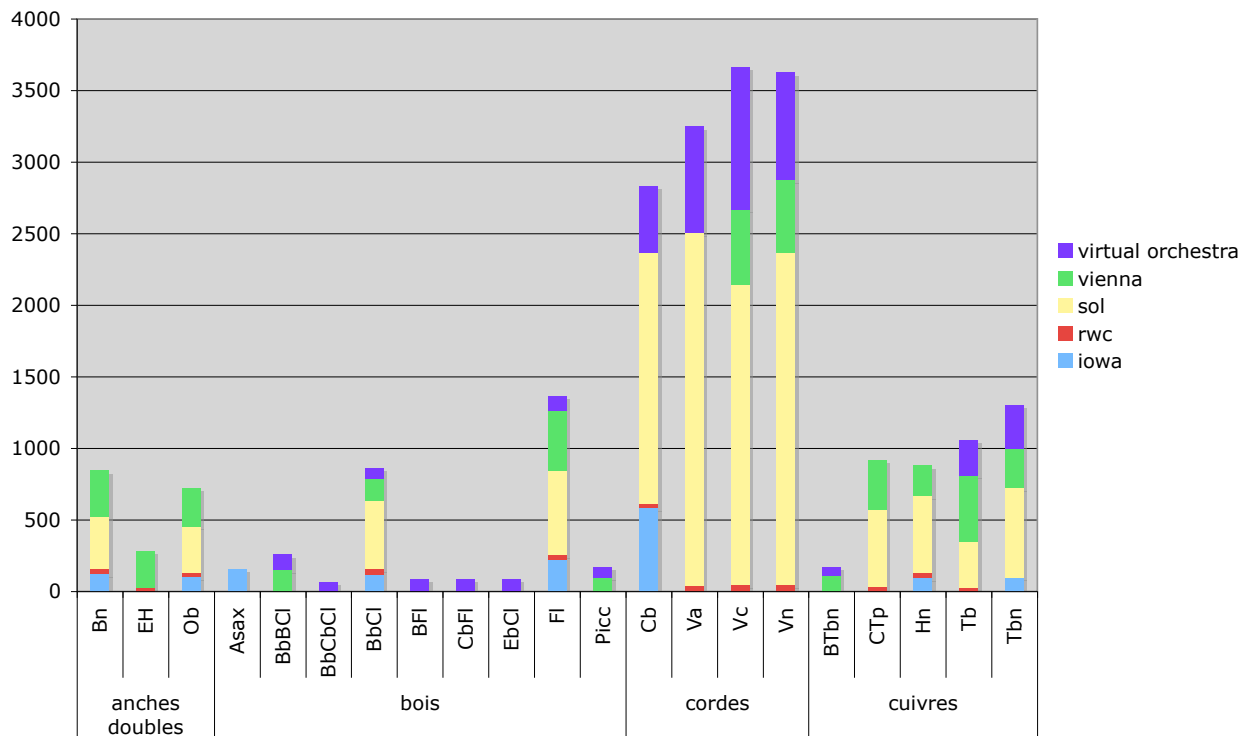


FIG. 3.5 – Répartitions des instruments dans les différentes bases

3.4 Organisations des modes de jeu

3.4.1 Similarité de techniques entre instruments

En regardant les différents modes de jeu décrits précédemment trois constats s'imposent. Premièrement, les mêmes techniques sont partagées par différents instruments aussi bien au niveau des familles qu'à un niveau supérieur. Ainsi, le vibrato peut être réalisé par tous les instruments. Les bois,

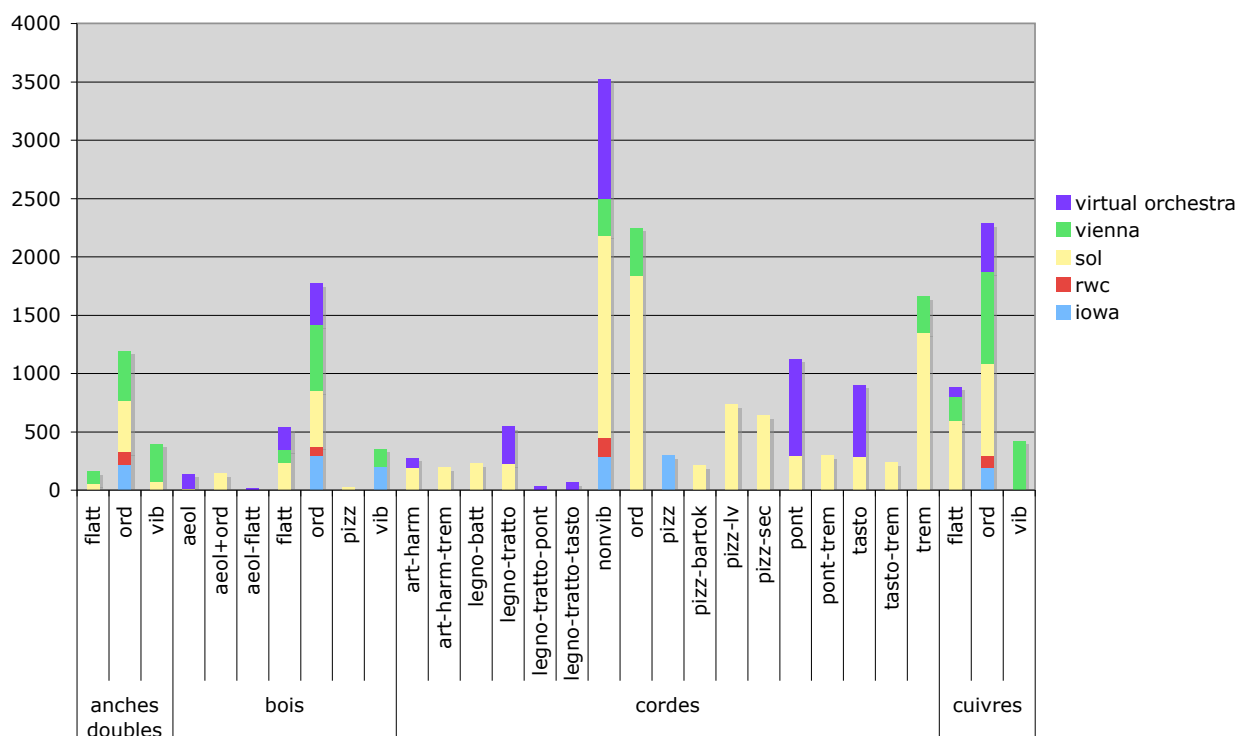


FIG. 3.6 – Répartition des modes de jeu dans les différentes bases

les anches doubles et les cuivres peuvent joués flatterzunge, ou toutes les cordes peuvent faire des tremolos. Deuxièmement, certaines de ces techniques peuvent être combinées entre elles, le violoniste peut utiliser une sourdine et jouer un tremolo sul tasto, un flûtiste peut jouer flatterzunge et aeolien. Enfin, les différentes techniques et leurs combinaisons ne sont pas toutes disponibles dans les bases d'échantillons, ou si elle le sont, seules quelques nuances ou hauteurs existent. Nous pouvons ici soit nous contenter de ce que contient la base d'échantillons et réaliser, quand c'est possible, les enregistrements manquants, soit trouver des méthodes permettant d'extraire un maximum d'informations de la base et de généraliser le plus possible ces informations. Par exemple des méthodes permettant d'appliquer à un instrument les caractéristiques d'un mode de jeu appris sur un autre instrument en sachant que ce mode de jeu est réalisable, ou bien d'utiliser le comportement d'un mode de jeu lors des changements de hauteurs et de nuances pour déduire les hauteurs et nuances manquantes pour un autre mode de jeu ou même un autre instrument. C'est cette seconde approche qui est choisie dans cette thèse et que nous décrirons au chapitre 6.

3.4.2 Taxinomie proposée

3.4.2.1 Description

Pour extraire et généraliser correctement les connaissances des bases d'échantillons nous proposons une taxinomie des modes de jeu. Cette taxinomie rend compte des différentes possibilités des

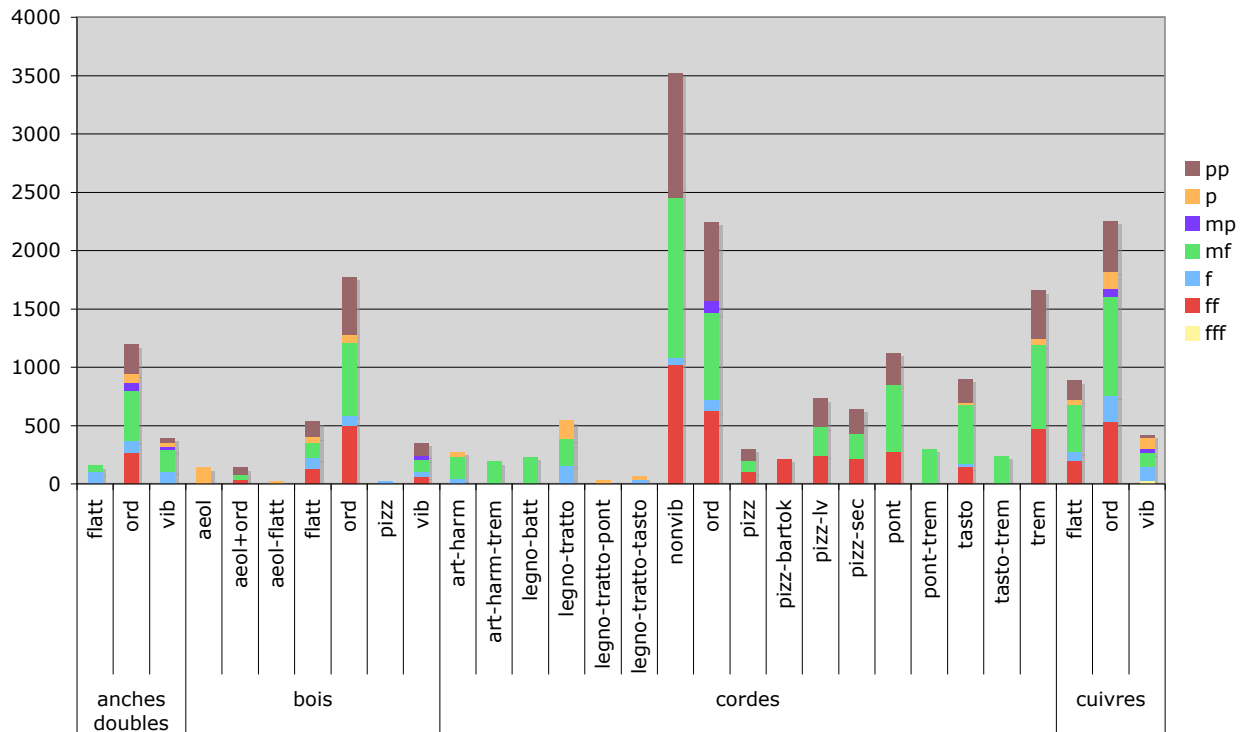


FIG. 3.7 – Répartitions des nuances pour les modes de jeu de chaque famille

instruments, en précisant les modes de jeu réalisables par une famille d'instruments, et des contraintes entre modes de jeu, c'est à dire lesquels sont réalisables en même temps et lesquels ne le sont pas. Cette taxinomie est décrite par un ensemble de variables et un ensemble de dépendances. Chaque variable représente un mode de jeu. L'ensemble des valeurs que peut prendre cette variable est limité et est précisé dans la taxinomie. Ainsi la variable *vib* peut prendre les valeurs *yes* ou *no* selon que le mode de jeu est vibrato ou non, de même, la variable *maindroite*, qui ne s'applique qu'aux cordes frottées, peut prendre la valeur *archet* si l'instrument est joué à l'archet ou *pizz* s'il est joué pizzicato. Chaque variable est associée à des dépendances permettant de préciser dans quels cas le mode de jeu est réalisable, comme dans l'exemple précédent pour exprimer : « ne s'applique qu'aux cordes frottées ». Ces dépendances sont décrites par des variables et les valeurs que doivent prendre ces variables pour que le mode de jeu soit réalisable. La table 3.1 montre un exemple de taxinomie. La variable *val_s* est une structure ayant pour champs l'ensemble des variables. Chaque champ a pour valeur un tableau contenant les valeurs possibles de la variable et donc du mode de jeu. La variable *dep_s* est une structure contenant les mêmes champs. Chaque champ a pour valeur un tableau décrivant les dépendances. Le premier élément de chaque ligne est une variable et les éléments suivants sont les valeurs que peut prendre cette variable pour que le mode de jeu décrit soit réalisable. Ainsi, pour que le mode de jeu *archet* soit réalisable, il faut que la famille soit celle des cordes (strings) et que la variable *maindroite* vaille *archet*. L'ensemble des variables utilisées est donné en annexe sous cette même forme est résumé dans le tableau 3.2.

```

val_s.maindroite = { 'archet', 'pizz' };
dep_s.maindroite = { 'famille', 'strings' };

val_s.archet = { 'ord', 'legno-tratt' };
dep_s.archet = { 'famille', 'strings',
                 'maindroite', 'archet' };

val_s.posArchet = { 'ord', 'pont', 'tasto' };
dep_s.posArchet = { 'famille', 'strings',
                    'maindroite', 'archet' };

```

TAB. 3.1 – Exemple de taxinomie

Mode de jeu	Abréviation	Valeurs possibles
Vibrato	vib	'no' 'yes'
Main droite pour les cordes	maindroite	'archet' 'pizz'
Harmonique artificiel	artHarm	'no' 'yes'
Technique d'archet	archet	'ord' 'legno-tratt' 'legno-batt'
Position de l'archet	posArchet	'ord' 'pont' 'tasto'
Sourdine pour les cordes	sourdineString	'N' 'S' 'SP'
Tremolo	trem	'no' 'yes'
Flatterzunge	flatt	'no' 'yes'
Sourdine pour les cuivres	sourdineBrass	'N' 'C' 'H' 'P' 'S' 'W'

TAB. 3.2 – Ensemble des variables et des valeurs possibles

3.4.2.2 Représentation sous forme d'arbre de décision

Cette taxinomie peut aussi être représentée sous forme d'arbre de décision où chaque noeud est une variable est chaque branche est une valeur de cette variable. L'ordre des noeuds permet de coder les dépendances entre variables. La figure 3.8 montre l'arbre de décision correspondant à la famille des cordes. Les sous arbres symbolisés par les flèches en pointillés sont identiques aux sous arbres qui partent des noeuds correspondant à la même variable.

3.4.2.3 Représentation symbolique des sons

Cette taxinomie permet d'obtenir une représentation vectorielle symbolique des sons. Chaque valeur d'un champs est associée à un entier correspondant à sa position dans la taxinomie. Ainsi un son de violon à l'archet, col legno tratto, sul tasto avec une sourdine de plomb aura pour représentation (1,1,1,2,3,3,1,0,0). Lorsqu'un mode de jeu n'est pas applicable, la valeur du champ est 0. Ensuite en associant une valeur à chaque hauteur (basée sur le protocole MIDI), à chaque nuance et à chaque ins-

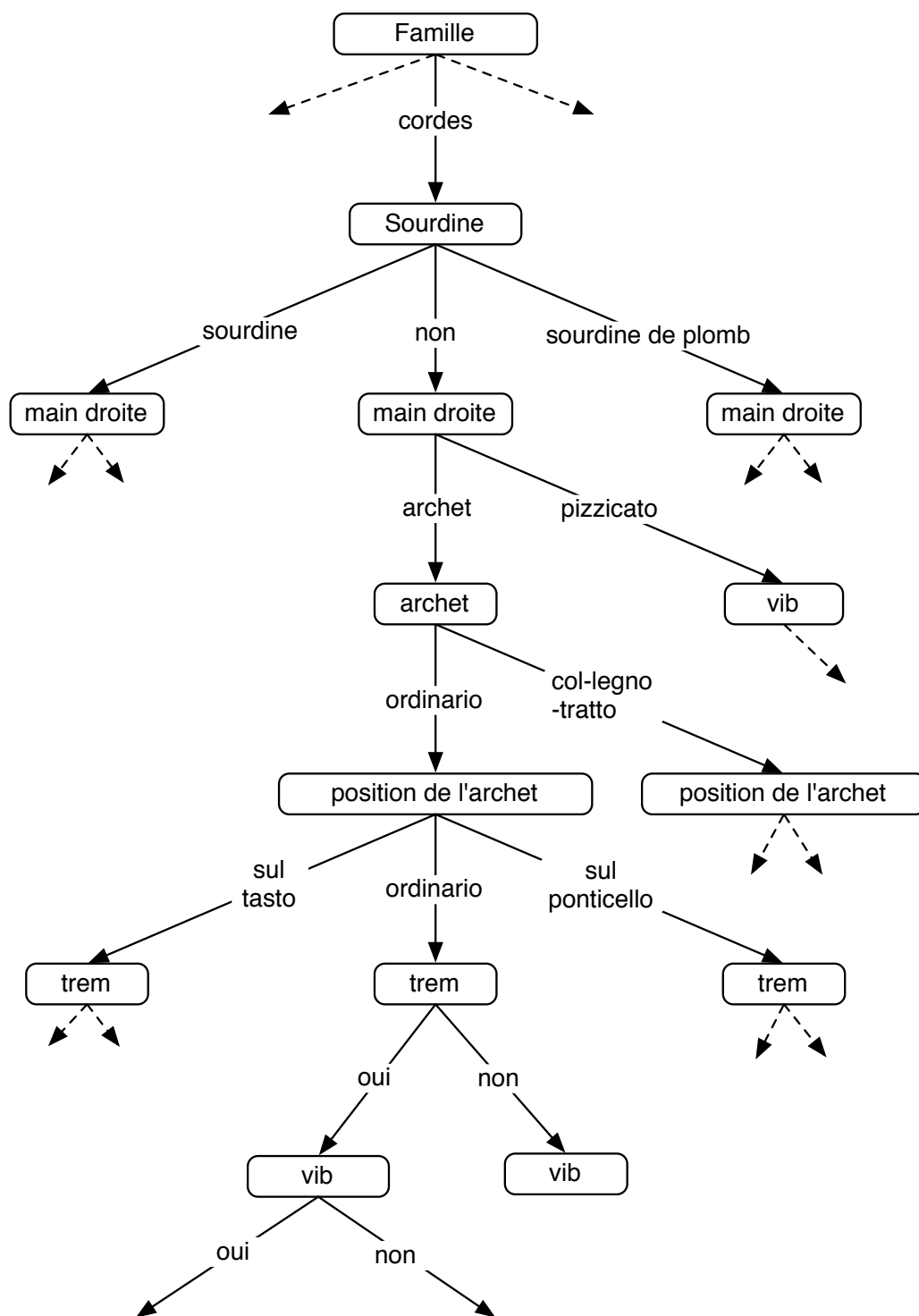


FIG. 3.8 – Arbre de décision pour les modes de jeu des cordes

trument, nous obtenons une représentation symbolique complète de chaque son. Cette représentation sera utilisée dans les chapitres suivant sous le nom *représentation symbolique*.

3.4.3 Utilisation de la taxinomie pour la simulation des orchestrations

3.4.3.1 Position du problème

Le résultat attendu du système d'orchestration est une partition. Cependant pour permettre l'interaction et pour que l'utilisateur puisse se faire une idée du son de l'orchestration proposée il est utile de pouvoir la synthétiser. Dans le cas d'un système basé directement sur une base d'échantillons, cela ne pose pas de problème particulier, il suffit de prendre les sons choisis et de les additionner en ajoutant éventuellement de la reverb. Dans notre cas des modèles sont appris sur plusieurs bases et surtout certains modèles seront extrapolés, comme des nuances et des hauteurs absentes mais aussi des combinaisons de modes de jeu. Ainsi le système peut proposer des orchestrations contenant des sons dont nous n'avons pas d'échantillons. Pour réaliser ces solutions, nous proposons de construire un arbre de décision à partir de la taxinomie.

3.4.3.2 Solution par arbre de décision

Chaque échantillon est décrit par un ensemble de variables symboliques constituée des variables de la taxinomie auxquelles s'ajoutent l'instrument, la famille, la hauteur et la dynamique. Chacune de ces informations constitue la réponse à une question posée à un sommet de l'arbre. Le but de la représentation par arbre de décision est de pouvoir d'une part trouver rapidement un son à partir des résultats de la procédure d'orchestration, d'autre part, si ce son n'existe pas, de pouvoir proposer rapidement une alternative dont le timbre ne soit pas trop éloigné. En partant du constat que deux sons ayant presque les mêmes représentations symboliques ont en général des timbres très proches nous utilisons ce type de méthode plutôt qu'une recherche par similarité sur les descripteurs de timbre.

L'arbre de décision construit ici n'est pas le même que celui de la figure 3.8. Ce dernier n'est qu'une représentation de la taxinomie et l'ordre des sommets n'a pas d'importance du moment qu'il est compatible avec la taxinomie. Au contraire, dans l'arbre construit pour la simulation des solutions, l'ordre des sommets est très important car il représente l'importance relative des informations symboliques. En effet, comme nous parcourons l'arbre en profondeur d'abord, c'est de cet ordre que découlera le choix du son de remplacement. Plus une variable est haute dans l'arbre, plus elle est supposée importante pour le timbre. Par exemple, il est pertinent de mettre la dynamique tout en bas de l'arbre, ainsi, si une orchestration contient un son *mf* dont nous ne possédons pas d'échantillons mais dont nous possédons une version *ff*, une requête dans l'arbre retournera ce dernier échantillon et il ne restera plus qu'à diminuer la sonie de l'échantillon retourné pour obtenir une approximation correcte de la requête initiale. La figure 3.9 donne un exemple d'arbre de décision construit pour la famille des cordes. Dans cet exemple, la note est tout en haut pour que cette caractéristique ne soit jamais changée. Les modulations d'énergie sont aussi placées dans le haut de l'arbre, ainsi une requête concernant un son de violon avec sourdine tremolo retournera un son de violon sans sourdine avec tremolo plutôt qu'un son avec sourdine sans tremolo. Comme pour le problème des poids dans une recherche multicritères, le problème de l'ordre des variables symboliques ne possède pas de solution unique. Pour résoudre ce problème, il serait possible de construire un arbre différent pour chaque orchestration à

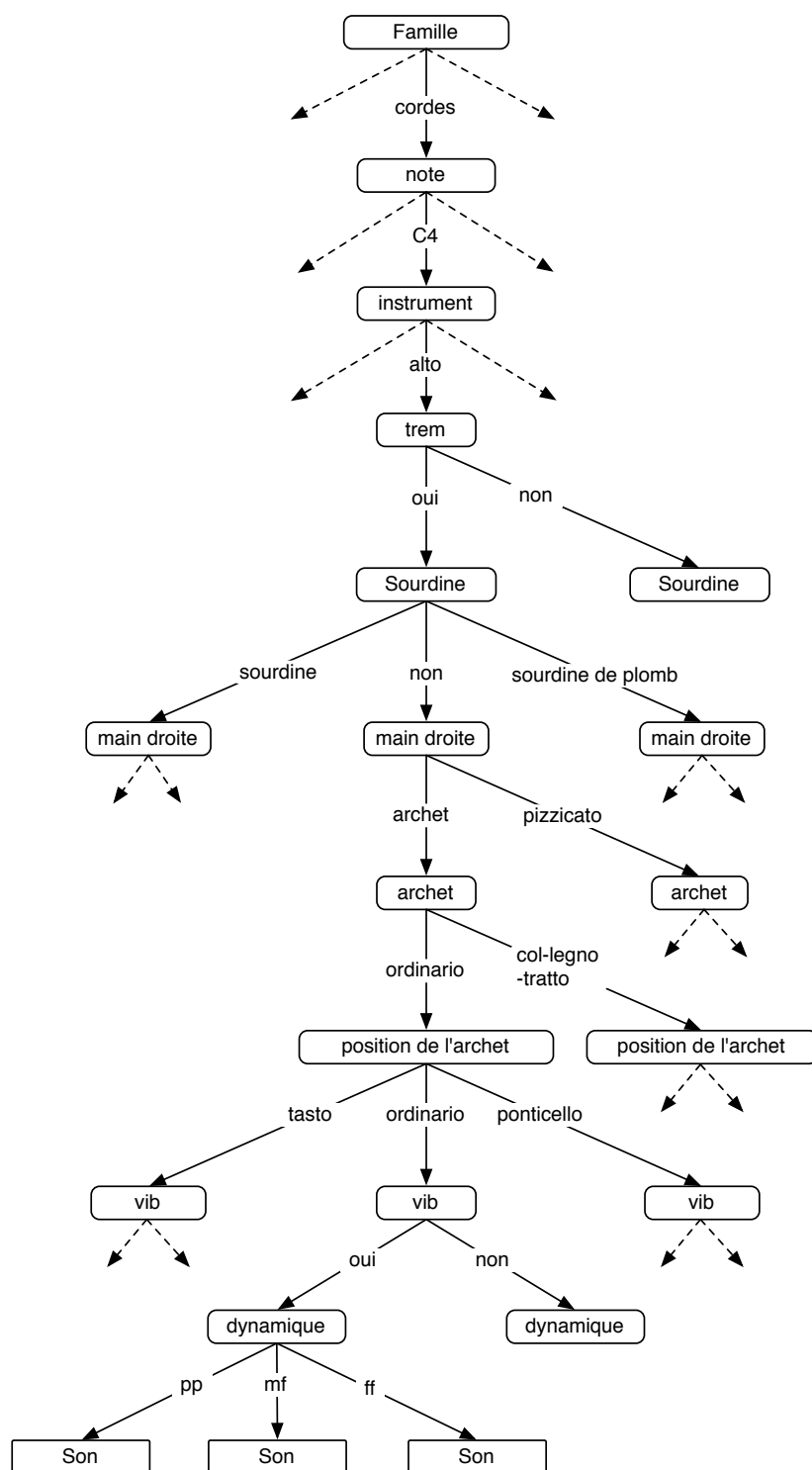


FIG. 3.9 – Arbre de décision pour la simulation des résultats

simuler. En utilisant le jeu de poids associé à cette solution, il est possible de décider d'un ordre pour les variables symboliques en utilisant des méthodes comme les arbres de régression et de classification (CART) [L⁺84]. Cela ne pose pas de problème particulier mais risque d'être trop lent pour une véritable interaction. Notons aussi que le chemin parcouru dans l'arbre donne des informations intéressantes pour éventuellement transformer le son trouvé pour le rapprocher du son recherché. Ainsi si nous cherchons un son avec vibrato et que le système renvoie un son sans vibrato, nous pourrions ajouter artificiellement du vibrato au son trouvé. Si des transformations sont possibles pour une variable, il est plus pertinent de la mettre dans le bas de l'arbre. Ainsi la variable *note* pourrait facilement être descendue tout en bas de l'arbre en ajoutant une procédure d'ajustement de hauteur.

Chapitre 4

Description du son et des combinaisons

4.1 Problématique du choix des descripteurs

Le choix des descripteurs pour l'orchestration est soumis à plusieurs contraintes. Tout d'abord, ils doivent permettre une bonne description des sons instrumentaux et une distinction entre les différents modes de jeu. Ensuite, ils doivent pouvoir être extraits de sons monodiques et polyphoniques puisque la cible à atteindre peut être polyphonique. La procédure d'extraction n'est pas toujours la même dans les deux cas, nous les distinguons quand c'est nécessaire. Les descripteurs doivent aussi décrire des caractéristiques de suffisamment haut niveau pour permettre une interaction avec l'utilisateur. C'est à dire des caractéristiques proche de dimensions perceptives. Enfin il doit être possible de définir une procédure "d'addition" des descripteurs, c'est à dire une procédure permettant de trouver le descripteur d'une combinaison à partir des descripteurs des sons qui la composent. Nous proposons certains descripteurs pour lesquels nous n'avons pas de procédure d'addition, ils seront utilisés dans la modélisation des instruments et des modes de jeu mais pas pour l'application finale d'orchestration.

4.2 Les descripteurs utilisés

4.2.1 La sonie et l'énergie

4.2.1.1 Introduction

Le signal résultant de la combinaison de sons dépend de l'énergie relative de ces sons. Dans le cas de sons instrumentaux ce problème prend une importance particulière puisque l'énergie est corrélée avec le spectre du signal. Il est donc important de connaître l'énergie des sons instrumentaux considérés pour pouvoir les sommer de manière réaliste. Dans notre contexte, toutes les informations sur les sons instrumentaux sont contenues dans les bases d'échantillons. Or, dans de telles bases, l'énergie du signal n'est pas une donnée fiable, en effet les sons peuvent être normalisés en amplitude ou les niveaux d'enregistrement peuvent varier. Il faut extraire cette information d'une autre source. Tous les échantillons contiennent une information de nuance (*pp*, *mf*, *ff*, ...) indiquant comment l'instrument a été joué. Nous utilisons cette information pour déduire la sonie du son puis son énergie. Pour cela nous sélectionnons un ensemble de sons dont l'énergie relative est correcte et évaluons la sonie moyenne pour chaque nuance. Ces moyennes sont ensuite utilisées pour tous les sons. Le lien entre nuance et

sonie est complexe et dépend généralement de l'instrument et du mode de jeu. Ainsi, un fortissimo de trompette sera généralement plus fort qu'un fortissimo de flûte. De même un son col-legno-tratto de violon sera moins fort qu'un son ordinaire à la même nuance. Nous adoptons cependant ici l'hypothèse que les sons de même nuance ont la même sonie. Cette hypothèse n'a pas d'influence sur la suite car la méthode proposée permet de changer les valeurs de sonie de référence selon l'instrument ou le mode de jeu dans le cas où on dispose de mesures plus précises.

4.2.1.2 Estimation de la Sonie

Nous utilisons une méthode simple d'estimation de la sonie. La sonie instantanée est calculée à partir du spectre de Mel. L'échelle de Mel est définie de la manière suivante :

$$M = \begin{cases} f & \text{si } f < 1000 \text{ Hz} \\ 1000 \left(1 + \log_{10} \left(\frac{f}{1000} \right) \right) & \text{sinon} \end{cases} \quad (4.1)$$

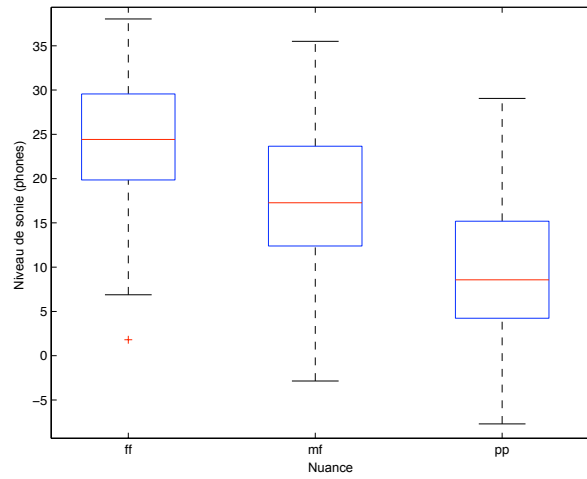
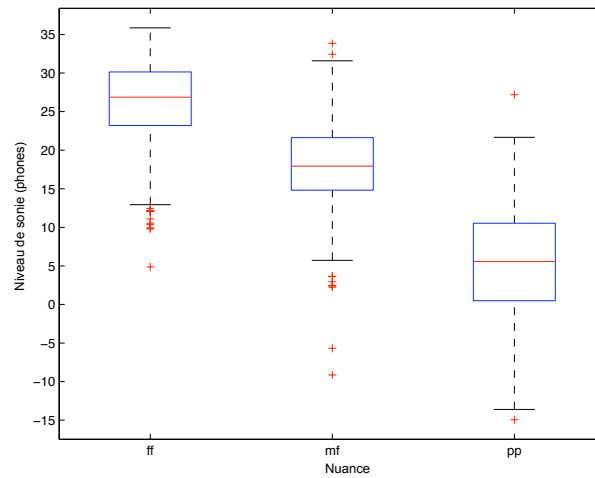
Le spectre de Mel $S(m, t)$ de la bande de Mel m et de la trame t est ensuite calculé de la manière suivante : l'axe des fréquences est divisé en 70 bandes égales. Après pondération par une fenêtre triangulaire pour chaque bande, L'énergie des points de la transformée de Fourier contenus dans chaque bande est sommée. Enfin la sonie instantanée $l(t)$ de la trame t peut être calculer :

$$l(t) = \sum_{m=1}^M S(m, t)^{0.6} \quad (4.2)$$

A partir de la sonie instantanée, nous estimons la sonie globale. Il existe plusieurs méthodes d'estimation de cette grandeur [Sus99]. Nous utilisons une méthode de type N_m . Cette méthode consiste d'abord à sélectionner les portions de signal dont l'énergie est supérieure à la valeur crête diminuée de 30 dB. Puis à extraire le N_m , c'est à dire le niveau de sonie dépassé pendant $m\%$ du temps. La valeur de m qui permet d'obtenir des estimations similaires pour les sons entretenus et les pizzicati et $m = 6$.

4.2.1.3 Estimation de la sonie moyenne sur un sous ensemble de Studio Online

Pour choisir les valeurs standard de sonie de chaque nuance nous étudions la sonie d'un sous ensemble de Studio Online dont les sonies relatives sont correctes. La banque de sons Studio Online a été enregistrée en conservant les positions des micros et les niveaux d'enregistrement. Sur l'ensemble des sons ordinaires, nous disposons de trois nuances (*pp*, *mf*, *ff*). Nous utilisons l'ensemble des sons ordinaires, ainsi que l'ensemble des pizzicati de cordes pour lesquelles nous disposons aussi de trois nuances. L'utilisation des pizzicati nous permet de déterminer la valeur de m dans le calcul de la sonie qui permet une estimation "correcte" de la sonie des sons percussifs. Nous avons donc choisi $m = 6$ pour lequel la sonie des sons entretenus est à peu près identique à celle des sons percussifs. Les figures 4.1 et 4.2 représente le niveau de sonie ($10 \cdot \log_2 l$ en phones) respectivement des sons tenus et pizzicati pour trois nuances. La figure 4.3 représente le niveau de sonie en fonction de la hauteur (octave) pour des sons entretenus. Cela permet de vérifier que la mesure de sonie ne dépend pas de la hauteur du son. Les valeurs de niveau de sonie choisies sont données dans le tableau 4.1. Les valeurs proposées

FIG. 4.1 – Sonie globale estimée par la méthode N_6 en fonction de la nuance pour des sons entretenusFIG. 4.2 – Sonie globale estimée par la méthode N_6 en fonction de la nuance pour des sons de cordes pizzicato

pour les nuances pp , mf et ff correspondent à la médiane du niveau de sonie des sons étudiés arrondie à l'entier le plus proche. Les valeurs pour les autres nuances sont des interpolations approximatives.

4.2.1.4 Adaptation de la sonie des échantillons

Les valeurs de référence obtenues sont utilisées pour adapter la sonie des échantillons des bases avant l'extraction des autres descripteurs. Ainsi l'énergie des signaux est cohérente avec la sonie de référence.

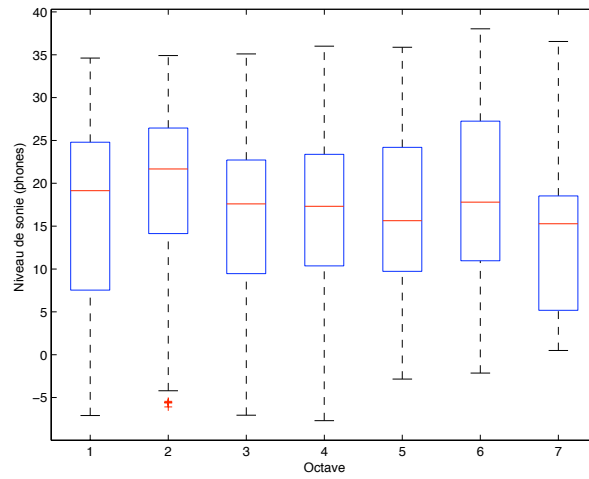


FIG. 4.3 – Sonie globale estimée par la méthode N_6 en fonction de l’octave pour des sons de cordes entretenus

Nuances	Phones
fff	33
ff	28
f	23
mf	18
mp	15
p	13
pp	8
ppp	3

TAB. 4.1 – Niveau de sonie des différentes nuances

4.2.2 Structure harmonique

4.2.2.1 Introduction

La structure harmonique d’un son, c’est à dire les composantes sinusoïdales qu’il contient, a une très grande influence sur la manière dont il est perçu. Nous appelons les composantes sinusoïdales des *partiels*. Tout d’abord, la fréquence des partiels définie en partie les hauteurs perçues dans le sons. Si les fréquences de plusieurs partiels ont une relation harmonique alors ces partiels auront tendance à être groupés dans un seul flux auditif et à créer une hauteur. Ensuite l’amplitude des partiels définie en partie le timbre du son. Dans notre contexte, les sons de la base sont supposés monodiques, c’est-à-dire qu’ils ne contiennent qu’une seule hauteur. La cible, en revanche, peut être polyphonique et donc contenir plusieurs hauteurs. Dans ce cas, trouver les hauteurs induites par le son est un problème complexe, et certains sons créent des sensations de hauteurs différentes selon

les personnes. Nous ne proposerons donc pas de méthode automatique d'estimation des hauteurs contenues dans la cible. Dans la suite de la thèse ces hauteurs seront supposées connues, c'est-à-dire fournies par le compositeur.

Nous décrivons la structure harmonique d'un son par un ensemble de hauteurs et un ensemble de partiels. Pour les sons de la base une seule hauteur est suffisante, tandis que plusieurs peuvent être nécessaires pour décrire la cible.

4.2.2.2 Les sons de la base

Le fréquence fondamentale Pour calculer la fréquence fondamentale (f_0) des sons de la base, nous utilisons la méthode du maximum de vraisemblance [DR93]. Pour éviter toute erreur d'octave dans le calcul, la zone de recherche de la fréquence fondamentale est réduite à un intervalle de deux tons autour de la hauteur théorique du son. Cette hauteur théorique est fournie par le nom de l'échantillon, puisque dans toutes les bases que nous utilisons, le nom des échantillons contient la hauteur.

Les partiels A partir de la fréquence fondamentale, nous pouvons estimer les paramètres des partiels. Les paramètres que nous calculons sont l'énergie et l'index des harmoniques. L'index est égal au multiple de la f_0 auquel correspond le partiel. Ainsi le partiel correspondant à la f_0 aura un index de 1, le premier harmonique aura un index de 2 et ainsi de suite. Nous supposons donc que les partiels sont parfaitement harmoniques ou, autrement dit, nous négligeons l'effet sur le timbre d'une petite déviation de la fréquence des partiels par rapport à la fréquence théorique d'un spectre harmonique. Cette hypothèse nous permettra de simplifier grandement le calcul de l'amplitude des partiels des combinaisons. L'index et l'énergie des partiels sont calculés par la méthode de [Rod97] qui utilise la fréquence fondamentale pour déterminer les pics spectraux les plus proches des fréquences harmoniques de la f_0 et ensuite estimer précisément leur fréquence, leur énergie et leur phase.

4.2.2.3 La cible

Pour les sons de la base, la fréquence fondamentale est calculée en premier puis est utilisée pour estimer les paramètres des partiels. Pour la cible nous procédons dans l'autre sens. Les partiels sont d'abord extraits puis assignés à une ou plusieurs f_0 . Les hauteurs contenues dans la cible sont supposées connues.

Les partiels L'amplitude et la fréquence des partiels de la cible sont calculées à l'aide du programme *Pm2* de l'Ircam. Pour chaque trame, l'algorithme utilisé calcule les paramètres des pics spectraux qui dépassent un seuil d'amplitude relatif au maximum de l'amplitude du spectre. Ces pics sont ensuite reliés au cours du temps pour former des partiels. Les partiels proches en fréquence et en amplitude sont connectés entre eux. Enfin les partiels trop court sont éliminés. Nous calculons la moyenne et l'écart type de l'amplitude et de la fréquence des partiels obtenus (voir paragraphe 4.2.7). Cette méthode permet d'extraire du spectre un grand nombre de partiel.

Pour obtenir une représentation compacte de la structure harmonique de la cible, nous proposons une méthode permettant de sélectionner les partiels les plus importants et de ne conserver que ceux-ci.

Par important nous entendons d'énergie élevée et résolu par le système auditif. Deux partiels sont résolus par le système auditif si leurs fréquences sont suffisamment éloignées (voir paragraphe 5.2.2.1). Dans ce cas la différence de fréquence entre deux partiels se mesure sur une échelle liée au bandes critique pour rendre compte de la variation de la sélectivité fréquentielle de l'oreille. La fréquence des partiels est donc convertie en ERB (Equivalent Rectangular Bandwidth) [GM90]. Si f est la fréquence en Hz, alors

$$ERB_N = 21.4 \log_{10}(0,00437f + 1); \quad (4.3)$$

Les partiels sont ensuite regroupés par un algorithme de classification ascendante hiérarchique à partir des distances entre leurs fréquences sur cette échelle. La distance entre les classes est calculée par la méthode du centroïde. Le nombre de classe est fixé et égal au nombre de partiels désirés (en général 20). Si un partiel forme une classe à lui seul, il est loin de tous les autres et probablement résolu par le système auditif. Il a donc une importance perceptive particulière et nous le conservons dans la description de la cible. Dans les classes qui contiennent plusieurs partiels, nous sélectionnons celui dont l'amplitude est la plus élevée. La figure 4.4 montre les partiels extraits d'un son de Tuba jouant un la#1 et les partiels sélectionnés par la classification. Les premiers partiels sont tous sélectionnés alors que dans les fréquences plus élevées, seuls les partiels les plus forts le sont.

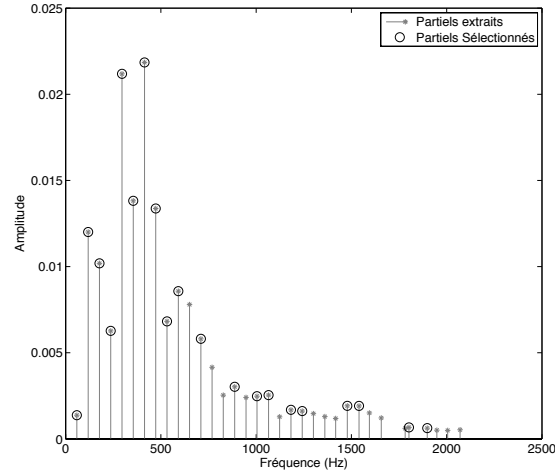


FIG. 4.4 – Partiels extraits d'un son de tuba la#1 mf et partiels sélectionnés par la classification

Les fréquences fondamentales Dans le cadre de cette thèse nous supposons que les hauteurs contenues dans la cible sont connues. Une méthode semi automatique permettant de suggérer un ensemble de hauteurs au compositeur a été proposée par Grégoire Carpentier dans [TCA⁺06].

Assignation des partiels de la cible à une f0 Chaque partiel est assigné à une ou plusieurs f0 précédemment trouvées dans la cible. Un partiel p de fréquence f est assigné à une fréquence fondamentale

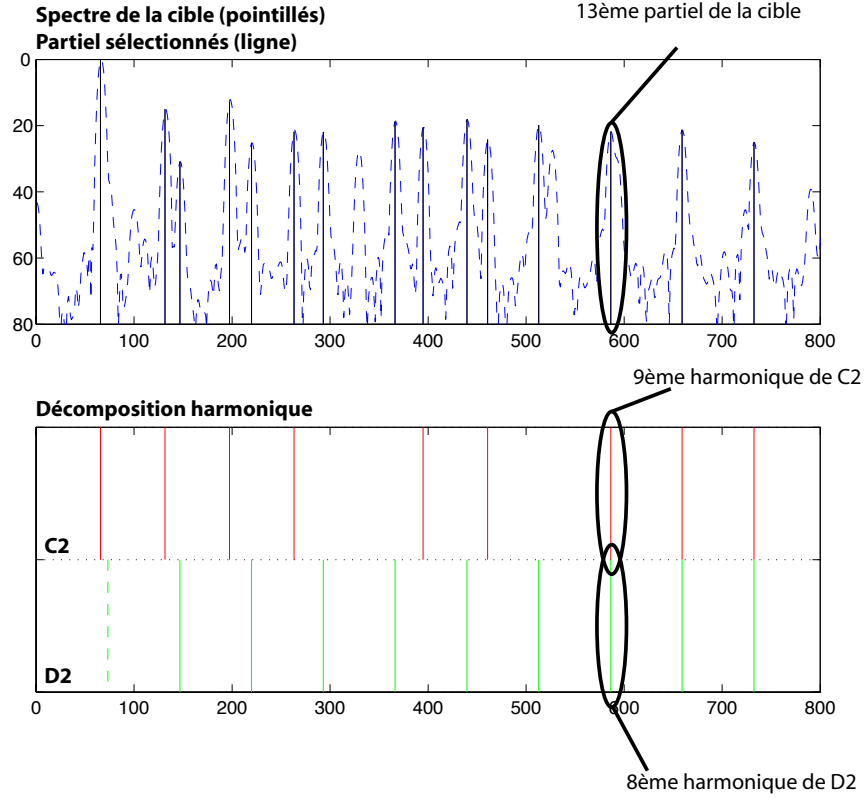


FIG. 4.5 – Assignment des partiels d’une cible à deux fréquences fondamentales

f_0 si il satisfait la condition suivante :

$$\exists N \in \mathbb{N}, \frac{N}{1 + \delta} \leq \frac{f}{f_0} \leq N(1 + \delta) \quad (4.4)$$

Où δ est un facteur de tolérance d’inharmonicité calculé comme suit :

Soit (p_i) un ensemble de partiel de fréquence (f_i) . Soit $\hat{f}_i$ la fréquence de la note, au quart de ton près, la plus proche en fréquence de p_i . Alors

$$\delta = \max_i \frac{f_i - \hat{f}_i}{\hat{f}_i} \quad (4.5)$$

Ainsi la cible est décrite par un ensemble de f_0 et un ensemble de partiels correspondant à chaque f_0 . La figure 4.5 montre un exemple de décomposition. La cible, dont les partiels sont représentés en haut, est décomposée sur un do2 et un ré2. Chacune des notes explique une partie des partiels de la cible. Certains partiels de la cible sont contenus dans les deux notes. Ainsi, le 13ème partiel de la cible correspond au 9ème harmonique de do2 et au 8ème de ré2. Cette représentation nous permettra de simplifier grandement l’addition des spectres de partiels.

4.2.3 L'énergie par bandes de Mel

Pour décrire le spectre nous utilisons aussi le spectre de Mel décrit au paragraphe 4.2.1.2. La figure 4.6 le spectre de Mel moyen en log d'une trompette jouant un *do6 mf*.

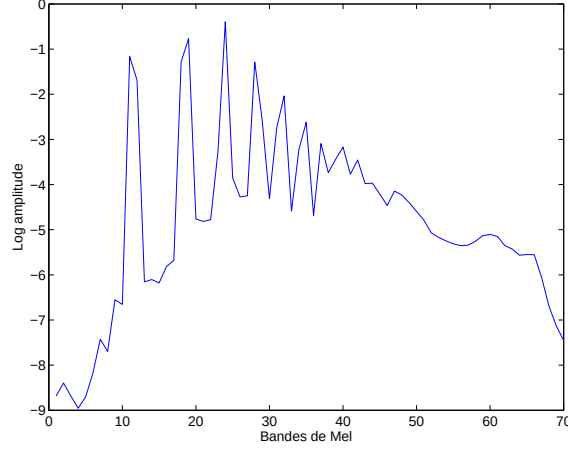


FIG. 4.6 – Logarithme du spectre de Mel d'un son de trompette jouant un *do6 mf*

4.2.4 L'enveloppe spectrale sur les bandes de Mel

Le calcul de l'enveloppe se fait par le calcul des MFCC (Mel Frequency Cepstrum Coefficient), c'est à dire en prenant la DCT du log du spectre de Mel. Puis on conserve seulement les 15 premiers coefficients qu'on utilise dans la transformation inverse.

4.2.5 Le centroïde et l'étendue spectrale

Nous calculons le centroïde et l'étendue spectral sur la spectre de puissance du son. Soit $p(f)$ le spectre de puissance d'une trame, où f est la fréquence. Alors le centroïde spectral s_c de la trame est :

$$s_c = \frac{\int_{f=0}^F f p(f) df}{\int_{f=0}^F p(f) df} \quad (4.6)$$

Où F est la fréquence de Nyquist.

L'étendue spectrale est définie comme l'écart type du spectre.

$$s_s = \sqrt{\frac{\int_{f=1}^F (f - s_c)^2 p(f) df}{\int_{f=1}^F p(f) df}} \quad (4.7)$$

Ces deux descripteurs permettent une représentation très synthétique du spectre liée à la perception des sons instrumentaux.

4.2.6 L'enveloppe du bruit

Pour décrire la partie non-sinusoïdale, nous extrayons dans un premier temps l'enveloppe de cette partie par la méthode de [YR06]. Cette méthode suppose que l'amplitude des pics spectraux dans une bande étroite suit une distribution de Rayleigh. Après une première classification des pics par la méthode de [RZR04], la majorité des pics harmoniques est retirée du spectre. Ensuite en utilisant l'hypothèse sur la distribution des pics du bruit, un processus itératif permet d'estimer une enveloppe lisse du bruit. Un exemple du résultat de cette méthode sur un signal polyphonique est présenté dans la figure 4.7. Cette méthode permet d'extraire une enveloppe lisse de tout type de signaux, monodiques ou polyphoniques et peut donc être appliquée aux sons de la base comme à la cible.

Deux descripteurs sont extraits à partir de cette enveloppe :

- Le centroïde de la partie non-sinusoïdale,
- Le rapport de l'énergie de la partie non-sinusoïdale sur l'énergie totale (*la noisiness*).

Ces descripteurs sont utiles pour des modes de jeu comme les aeolien des vents où le son est composé en grande partie par le bruit du souffle de l'instrumentiste dans l'embouchure, ou pour les trémolos des cordes. Ainsi la figure 4.8 montre la valeur du logarithme de la noisiness pour des sons de flûte et de clarinette pour les modes de jeu ordinario, aeolien+ordinario et aeolien. Plus le son contient de souffle, plus la valeur est élevée. De même la figure 4.9 présentes les valeurs du centroïde spectrale de son aeolien de flûte, clarinette, flûte basse et clarinette basse. Ce descripteur permet bien de distinguer les deux familles d'instruments.

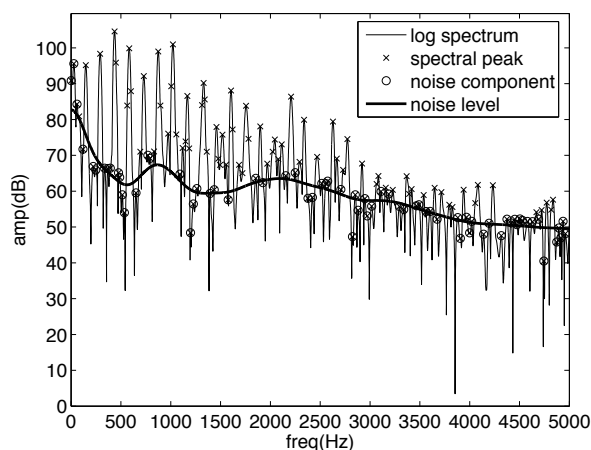


FIG. 4.7 – Exemple d'extraction de l'enveloppe du bruit (extrait de [YR06])

4.2.7 Moyenne des descripteurs spectraux et normalisation

Tous les descripteurs du spectre sont ensuite moyennés dans le temps. La moyenne est pondérée par la sonie locale du sons. L'écart type pondéré du descripteur au cours du temps est aussi utilisé comme descripteur.

Tous les descripteurs multidimensionnels du spectre sont normalisés par l'énergie. Nous prenons ensuite le logarithme des partiels et du spectre de Mel.

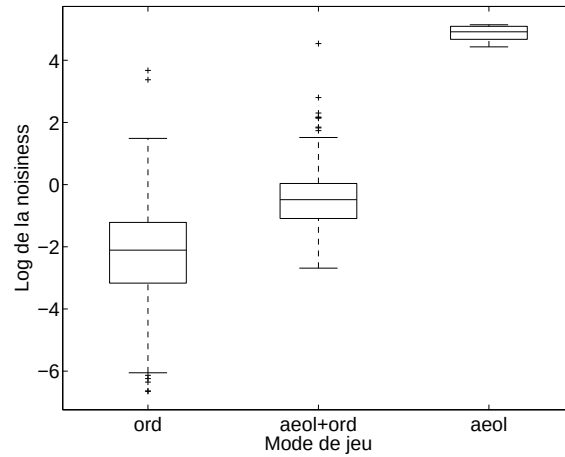


FIG. 4.8 – Log noisiness des sons de clarinette et de flûte pour les modes de jeu ordinario, aeolien+ordinario et aeolien

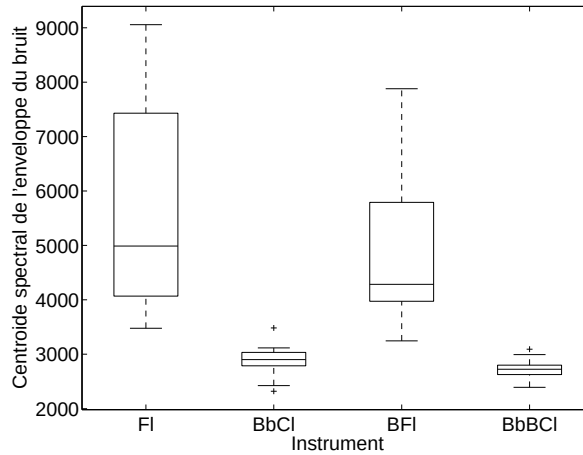


FIG. 4.9 – Centroïde spectrale de l'enveloppe du bruit pour les sons aéolien de flûte, clarinette, flûte basse et clarinette basse

4.2.8 Les modulations

Nous calculons la fréquence et l'amplitude des modulations de fréquence et d'énergie de la manière suivante :

Modèle Le modèle utilisé pour l'énergie et la fréquence fondamentale est le même. Appelons $x(t)$ la grandeur modélisée. Après un découpage en trame, nous décomposons cette grandeur en une moyenne, une variation linéaire et une composante cosinusoidale, ainsi pour la trame k :

$$x_k(t) = E(x_k(t)) + m_k t + p_k + a_k E(x_k(t)) \cos(2\pi f_k t + \phi_k) \quad t \in [t_k, t_{k+1}] \quad (4.8)$$

avec $E(m_k t + p_k) = 0$

où $E(x_k(t))$ représente la moyenne de $x_k(t)$ sur la trame k , a_k représente l'amplitude, f_k la fréquence et ϕ_k la phase de la modulation.

Estimation des paramètres du modèle Pour chaque trame $x_k(t)$, la moyenne et la tendance linéaire sont d'abord retirées. Puis les paramètres a_k , f_k et ϕ_k sont estimés sur la DFT du signal résultant. Le calcul se fait en détectant le plus grand pic de la DFT dans la bande de fréquence comprise entre 0 Hz et 60 Hz ou 0 Hz et 10 Hz, et en estimant son amplitude et sa fréquence. Nous utilisons ces deux bandes de fréquence pour être capable de détecter les modulations des sons vibrato, trémolo et flatterzunge.

Descripteurs Les descripteurs utilisés sont :

- moyenne de la vitesse et de la modulation d'énergie,
- logarithme de la moyenne de l'ambitus de la modulation d'énergie. Le logarithme permet de rendre la distribution plus gaussienne (voir figure 4.10),
- moyenne de l'ambitus de la modulation de fréquence en cents,
- moyenne de la vitesse de la modulation de fréquence.

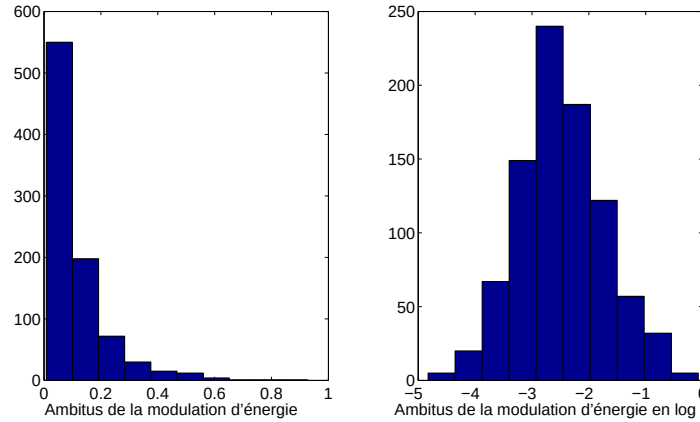


FIG. 4.10 – Ambitus de la modulation d'énergie pour les sons flatterzunge de Studio en Ligne en linéaire et en logarithme

La figure 4.11 montre la distribution de la vitesse de la modulation d'amplitude en fonction du mode de jeu, nous voyons que les valeurs obtenues correspondent bien à la littérature avec des valeurs comprises entre 10 et 15 Hertz pour les trémolo, entre 20 et 40 Hertz pour les flatterzunge et des valeurs autour de 5 Hz pour le vibrato. Les valeurs obtenues pour les sons ordinario ne sont pas significatives puisqu'elles correspondent à de faibles ambitus de modulation comme le montre la figure

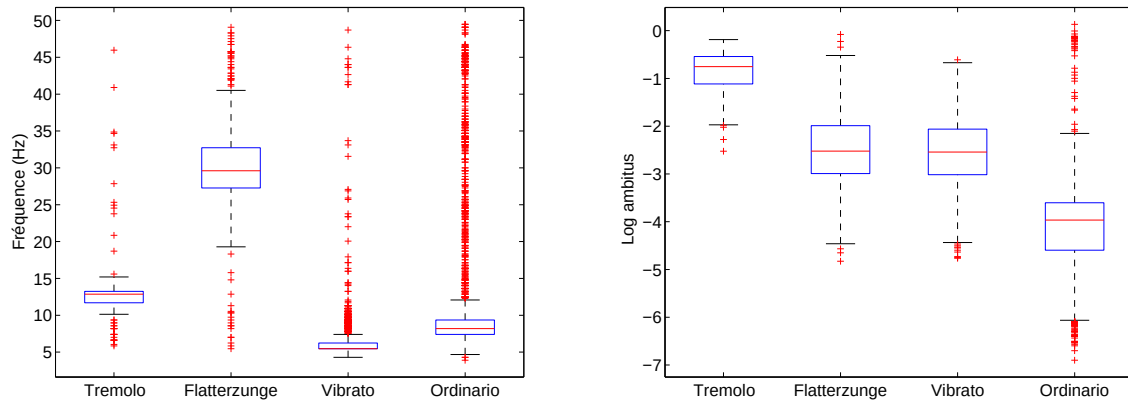


FIG. 4.11 – Vitesse et logarithme de l’ambitus de la modulation d’énergie en fonction du mode de jeu

4.11. Cette figure permet aussi de constater que le flatterzunge et le vibrato créent des modulations d’énergie dont les ambitus sont similaires, tandis que les modulations provoquées par le trémolo sur les cordes sons beaucoup plus grandes.

4.2.9 Temps d’attaque

L’estimation du temps d’attaque se fait par la méthode de seuil adaptatif de [Pee04]. D’abord, l’enveloppe d’énergie du signal est extraite et normalisée par sa valeur maximale. L’intervalle des énergies est ensuite divisé en parties égales (10). Les limites de ces parties sont des seuils d’énergie. Le temps mis pour aller d’un seuil au suivant est appelé effort. Un effort élevé correspond à un temps élevé donc à une pente d’énergie faible. La moyenne M de l’effort est calculée. Le début de l’attaque est défini comme le dernier seuil pour lequel l’effort est inférieur à $K * M$ où K est une constante. De même la fin de l’attaque est définie comme le premier seuil (qui suit le début de l’attaque) pour lequel l’effort est supérieur à $K * M$. K est fixé à 3. Le temps d’attaque est défini comme la différence entre la fin et le début de l’attaque.

4.3 Liste complète des descripteurs et abréviations utilisées

Abréviation	Description	Paragraphe
f0	Fréquence fondamentale	§4.2.2
partP	Puissance des partiels	§4.2.2
partPstd	Déviati��n standard de la puissance des partiels au cours du temps	§4.2.2
partPstdM	Moyenne de la déviati��n standard de la puissance des partiels au cours du temps	§4.2.2
sc	Centro��ide spectral	§4.2.5
ss	Etendue spectrale	§4.2.5
melP	Puissance par bandes de Mel (spectre de Mel)	§4.2.3
melPstd	Déviati��n standard du spectre de Mel au cours du temps	§4.2.3
melPstdM	Moyenne de la déviati��n standard du spectre de Mel au cours du temps	§4.2.3
melPenv	Enveloppe du spectre de Mel	§4.2.4
scEnvn	Centro��ide spectral de l'enveloppe de la partie non sinuso��dale	§4.2.6
ssEnvn	Etendue spectrale de l'enveloppe de la partie non sinuso��dale	§4.2.6
noisiness	Rapport de l'��nergie de la partie non sinuso��dale sur l'��nergie totale	§4.2.6
lat	Logarithme du temps d'attaque	§4.2.9
F0ma	Ambitus de la modulation de fr��quence	§4.2.8
F0mf	Vitesse de la modulation de fr��quence	§4.2.8
ema	Ambitus de la modulation d'��nergie (fc=60Hz)	§4.2.8
emf	Vitesse de la modulation d'��nergie (fc=60Hz)	§4.2.8
ema1	Ambitus de la modulation d'��nergie (fc=10Hz)	§4.2.8
emf1	Vitesse de la modulation d'��nergie (fc=10Hz)	§4.2.8

TAB. 4.2 – Liste des descripteurs utilis  s

Chapitre 5

L'addition des descripteurs et des modèles

5.1 Introduction

Lorsque deux ou plusieurs sons sont émis simultanément, ils interagissent entre eux. Dans le contexte de l'aide à l'orchestration, nous devons modéliser cette interaction pour obtenir des opérateurs simples nous permettant de combiner des sons, ce qui signifie d'être capable, à partir de la connaissance des propriétés des sons émis, de déterminer les propriétés de la combinaison de ces sons. Cette interaction dépend de plusieurs phénomènes liés aux propriétés des émetteurs (les instruments), du transmetteur (la salle) et du récepteur (l'auditeur). Sa compréhension fait donc intervenir des notions d'acoustique et de perception. La perception détermine comment est traitée l'information par l'auditeur tandis que l'acoustique permet de comprendre comment le signal émis par les instruments est transformé par la salle. Il faut aussi prendre en compte les variations du signal émis par les instruments, conséquences des propriétés acoustiques des instruments mais aussi des gestes des instrumentistes. Nous présentons dans ce chapitre un ensemble de propriétés et de résultats expérimentaux concernant ces phénomènes. Ces résultats nous permettent de proposer une approximation de l'interaction des sons instrumentaux dans un orchestre.

Dans la partie 5.2, nous étudions comment les signaux instrumentaux s'additionnent lorsqu'ils sont émis simultanément. Nous en déduisons une méthode d'addition du spectre des sons entretenus. Mais l'approche mathématique et acoustique ne permet pas de tout expliquer. Ainsi si un son entretenu est émis en même temps qu'un son percussif (une clarinette et un vibraphone par exemple) que peut-on dire des caractéristiques du son qui en résulte ? comment "s'additionnent perceptivement" ces deux sons ? Pour répondre à cette question nous avons mené une expérience perceptive en collaboration avec Stephen McAdams à l'université de McGill. Cette expérience est décrite dans la partie 5.3. Le but de cette expérience est double. D'abord nous cherchons à obtenir une description concise du timbre de combinaison contenant un son percussif et un son entretenu. Ensuite, en partant du constat que le recours à ce genre d'expérience est nécessaire pour élaborer un système d'orchestration assistée par ordinateur, nous essayons de proposer un protocole expérimental et des méthodes d'analyse des résultats pouvant être utilisés pour traiter d'autres problèmes liés à l'orchestration. Ainsi le protocole proposé pourrait par exemple être appliqué à des sommes de sons modulés en amplitude pour obtenir

une description du timbre du signal résultant.

Enfin dans la partie 5.4 nous donnons les formules utilisées pour additionner les descripteurs proposés au chapitre précédent. Ces formules sont déduites des deux premiers paragraphes, ou quand cela n'est pas possible, constituent des approximations basées sur l'observation des signaux.

5.2 Addition de signaux

5.2.1 Énergie d'une somme de signaux

Dans le cas idéal de signaux émis sans aucun effet de salle regardons comment se comporte l'énergie de la combinaison de signaux.

Soit s_1 et s_2 deux signaux quelconques.

$$Energie(s_1(t) + s_2(t)) = E\left(\int_{T_1}^{T_2} (s_1(t) + s_2(t))^2 dt\right) \quad (5.1a)$$

$$= E\left(\int_{T_1}^{T_2} s_1(t)^2 dt\right) + E\left(\int_{T_1}^{T_2} s_2(t)^2 dt\right) + 2E\left(\int_{T_1}^{T_2} (s_1(t)s_2(t))^2 dt\right) \quad (5.1b)$$

Si s_1 et s_2 sont décorrélés sur $[T_1, T_2]$ alors $E(\int_{T_1}^{T_2} (s_1(t)s_2(t))^2 dt) = 0$ donc $Energie(s_1(t) + s_2(t)) = Energie(s_1(t)) + Energie(s_2(t))$.

Ainsi pour du bruit, l'énergie de deux signaux émis simultanément est égale à la somme des énergies des signaux. Dans le cas des instruments, ceci s'applique à la partie non-sinusoïdale du signal (paragraphe 4.2.6). Pour la partie sinusoïdale nous ne pouvons pas supposer la décorrélation, nous devons donc étudier dans quel cas ces signaux interagissent, et dans le cas d'une interaction, comment se comporte le troisième terme de l'équation 5.1b

5.2.2 Interaction entre sons sinusoïdaux

5.2.2.1 La perception de sons sinusoïdaux simultanés

Les sons purs L'émission simultanée de deux sons purs résulte, en terme acoustique, en la production de battements. La perception de ce phénomène a été largement étudiée (voir [Pre98]). Helmholtz [Hel98] distingue trois zones différentes selon l'écart fréquentiel entre les sons. Si l'écart est faible, les battements sont entendus clairement, lorsque l'écart augmente, leur perception s'estompe, il devient par exemple impossible de les compter, et apparaît une sensation qu'il appelle rugosité. Enfin au-delà d'un certain écart les deux sons peuvent être perçus séparément.

L'étude de ce phénomène a depuis donné lieu à plusieurs interprétations en particulier les interprétations dites spectrale et temporelle. Nous nous intéressons ici uniquement à l'interprétation spectrale. Celle-ci postule que la rugosité peut être prédite par la composition spectrale des sons. Pour deux sons purs, la rugosité disparaît lorsqu'ils sont résolus par l'oreille interne, c'est-à-dire lorsque l'écart fréquentiel Δf est supérieur à la largeur de la bande critique à la fréquence moyenne considérée. De plus

le maximum de rugosité est atteint pour un écart fréquentiel approximativement égal à une fraction constante de la largeur de la bande critique à la fréquence moyenne considérée.

Nous pouvons ainsi isoler trois zones et préciser leurs limites :

- $\Delta f < 3 - 4Hz$, les battements sont perçus.
- $3 - 4Hz < \Delta f < 110\%$ bande critique, perception de rugosité, avec un maximum de rugosité à 45% de la bande critique.
- $\Delta f > 110\%$ de la bande critique, les deux sons sont entendus séparément.

Ainsi, il faut étudier l'interaction entre deux sons sinusoïdaux uniquement dans les deux premiers cas. Nous supposons dans les paragraphes suivants que l'écart fréquentiel entre les sons est inférieur à 110 % de la largeur d'une bande critique.

Les sons complexes Les sons instrumentaux ne sont pas des sons purs, mais contiennent un ensemble de sinusoïdes à différentes fréquences. Il est donc intéressant d'étudier comment les résultats précédents s'étendent à des sons complexes. Pour ce type de sons, les théories et les résultats divergent beaucoup plus, sans doute du fait de la difficulté de mise en place d'expériences et d'analyse des résultats. Cependant, les résultats s'accordent généralement sur le fait que si deux sons purs modulés en amplitude sont émis simultanément à des fréquences différentes, une incohérence entre les deux modulations, même dans des canaux auditifs différents, provoque une diminution de la détection des modulations. Donc, lorsque deux sons contenant des partiels de fréquences proches sont émis simultanément, les partiels proches vont interagir en créant des modulations d'amplitude. Mais si ces modulations sont incohérentes alors leur détection sera plus difficile. Ce phénomène est étudié dans la littérature sur l'interférence de détection des modulations ([HG91], [OD01]) ou sur la rugosité. Ainsi [PM99] a montré que la combinaison de deux sons modulés en amplitude de manière à produire de la rugosité est moins rugueuse lorsque les deux modulations sont incohérentes.

5.2.2.2 Addition de sons sinusoïdaux

Considérons qu'un ensemble de sons purs de fréquences proches au sens des bandes critiques sont émis simultanément. Ces sons ne sont pas résolus par l'oreille et interagissent donc.

Posons $s_k(t) = A_k(t)\cos(\phi_k(t))$ ($k \in 1, \dots, N$). où $s_k(t)$ représente le k ème son émis.

$\hat{s}(t) = \sum_{k=1}^N s_k(t)$ est le signal résultant. Nous cherchons à déterminer l'amplitude et la phase de $\hat{s}(t)$ c'est-à-dire que nous cherchons $\hat{A}(t)$ et $\hat{\phi}(t)$ tels que $\hat{s}(t) = \hat{A}(t)\cos(\hat{\phi}(t))$. Ainsi,

$$\begin{aligned}
 \hat{s}(t) &= \sum_{k=1}^N s_k(t) \\
 &= \operatorname{Re}\left\{\sum_{k=1}^N A_k(t)e^{j\phi_k(t)}\right\} \\
 &= \operatorname{Re}\{R(t) + jI(t)\} \\
 &= \hat{A}(t)\cos(\hat{\phi}(t))
 \end{aligned}$$

Donc,

$$\hat{A}(t) = \sqrt{R(t)^2 + I(t)^2} \quad (5.2)$$

$$\hat{\phi}(t) = \text{sign}(I(t)) \arccos(R(t)/\hat{A}(t)) \quad (5.3)$$

Ainsi pour l'amplitude nous obtenons,

$$\hat{A}(t)^2 = \left(\sum_k A_k(t) \cos(\phi_k(t)) \right)^2 + \left(\sum_k A_k(t) \sin(\phi_k(t)) \right)^2 \quad (5.4)$$

ou en réarrangeant les termes :

$$\hat{A}(t)^2 = 2 \sum_k \sum_{n>k} A_k(t) A_n(t) \cos(\phi_k(t) - \phi_n(t)) + \sum_k A_k^2(t) \quad (5.5)$$

Le deuxième terme de l'équation 5.5 est constant et correspond à la somme en puissance (ou en énergie) des signaux, tandis que le premier terme est susceptible de varier. Dans le cas où les fréquences sont égales ou proches, ce terme est constant et dépend de la différence de phase entre les signaux. Par proche, nous entendons que l'inverse de la différence entre les fréquences (qui est donc la période de modulation) est grand devant la durée des signaux considérés. Dans le cas où les fréquences sont plus éloignées, ce premier terme oscille et crée des battements de forme d'autant plus complexe que le nombre de composantes croît.

Le premier terme dépend de la différence de phase au cours du temps entre les deux signaux et donc de la différence de fréquence. Sur l'ensemble des réalisations, nous pouvons supposer que la phase est une variable uniforme sur $[0, 2\pi]$, et donc que la différence de phase suit la même distribution. De ce fait, le premier terme de l'équation 5.5 s'annule en moyenne, donc l'énergie résultante est en moyenne égale à la somme des énergies des signaux. Le problème est de savoir ce qui se passe pour chaque réalisation. La figure 5.1 présente la déviation de la fréquence des sons mesurée sur les sons ordinaires de Studio Online par rapport à la fréquence théorique. Cette mesure correspond à la différence entre la médiane de la fréquence fondamentale mesurée et la fréquence fondamentale théorique. Nous voyons qu'il est fréquent que la hauteur du son joué diffère d'un dixième de ton. Cette déviation est constante en ton par rapport à la hauteur du son. Donc plus le son est aigu plus la différence de fréquence est grande. De plus la différence entre les partiels augmente avec l'index du partiel. Si nous supposons la fréquence fondamentale fixe au cours du temps, l'addition de deux sons instrumentaux provoque généralement des battements lents en basse fréquence et plus rapides en haute fréquence. Mais la fréquence fondamentale n'est pas constante dans le temps. La figure 5.2 montre la dispersion de la fréquence fondamentale pour chaque son sous la forme de l'écart interquartile en tons. La dispersion se situe autour de 0,04 ton et n'est donc pas négligeable. Cette variation de la fréquence fondamentale ajoute de l'irrégularité aux battements.

La déviation de la fréquence fondamentale par rapport à la fréquence théorique et sa dispersion permettent de supposer que dans le cas anéchoïque l'émission simultanée de deux sons instrumentaux de même hauteur ou avec des partiels communs résulte en l'addition de l'énergie des signaux en moyenne si la période de temps considéré est suffisamment grande devant l'écart fréquentiel entre les partiels qui interagissent. L'interaction crée des battements d'autant plus rapides et de forme d'autant plus complexe que la fréquence des partiels augmente. D'après ce que nous avons vu au paragraphe

5.2.2.1, l'incohérence des battements les rend difficilement détectables. Nous allons voir que l'effet de salle ajoute encore à l'incohérence des battements et permet ainsi de diminuer l'intervalle temporel permettant de considérer que les énergies s'ajoutent en moyenne.

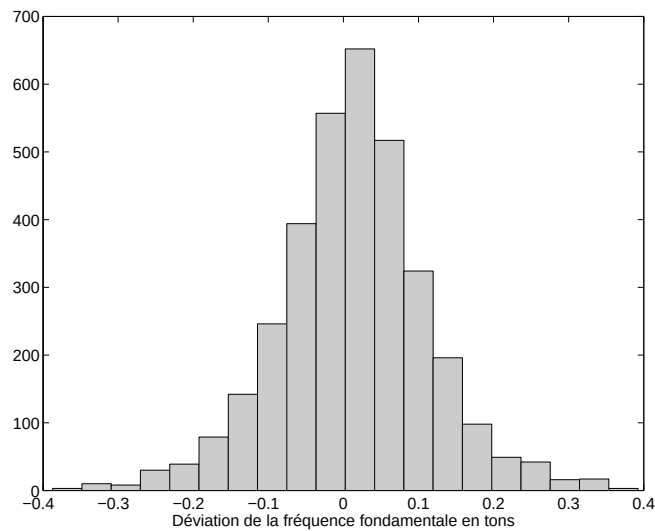


FIG. 5.1 – Déviation de la fréquence fondamentale du son mesurée sur les sons ordinaires de Studio Online par rapport à la fréquence fondamentale théorique

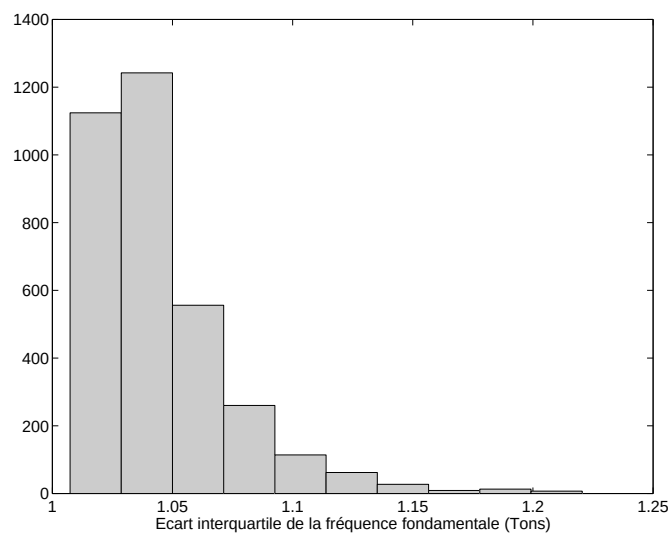


FIG. 5.2 – Écart interquartile de la fréquence fondamentale mesurée

5.2.3 L'effet de salle

Rappelons pour commencer quelques éléments d'acoustique des salles. Nous pouvons étudier la propagation du son dans une salle de deux manières différentes, soit du point de vue "modal" soit du point de vue "géométrique" en utilisant les rayons sonores. Nous utilisons ici le point de vue modal. Les modes propres dépendent de la géométrie de la salle, cependant, lorsque la fréquence croît, la densité fréquentielle des modes propres augmente. Ainsi si un son pur est émis, la réponse de la salle en un point est constituée de la contribution d'un grand nombre de modes de phases aléatoires, la répartition de l'énergie dans la salle est alors donnée par une variable aléatoire indépendante des caractéristiques de la salle. La fréquence à partir de laquelle on suppose ce modèle est appelée fréquence de Schroeder : $f_s \approx 2000\sqrt{T/V}$. Où T représente le temps de réverbération de la salle et V son volume en mètres cubes. Par la suite nous nous placerons au-dessus de cette fréquence ¹. Dans ce cas, le carré de la valeur absolue de la pression p mesurée en un point divisé par sa moyenne fréquentielle ou spatiale, qui est proportionnel à l'énergie, est distribué comme une loi exponentielle. La figure 5.3 représente la pression mesurée le long d'une ligne à fréquence constante et mesurée en variant la fréquence à une position fixée. L'écart entre les maxima spatiaux est égal à $.79 \lambda$ (2.7 cm à 10 kHz, 2.7 cm à 100 Hz), où λ correspond à la longueur d'onde du son. L'écart entre les maxima en fréquence peut être approximé par $\Delta f = 4/T$ où T représente le temps de réverbération, ce qui correspond à des valeurs entre 1 Hz et 4 Hz pour une salle comme l'Espace de Projection de l'Ircam.

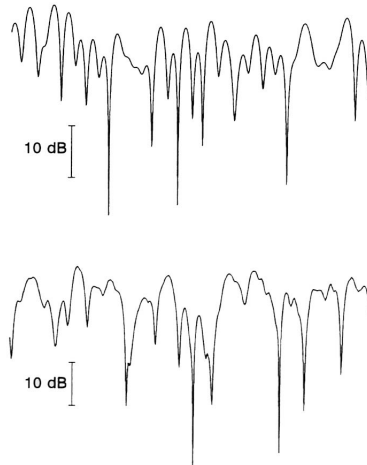


FIG. 5.3 – Pression mesurée le long d'une ligne à fréquence constante et mesurée en variant la fréquence à une position fixée (extrait de [Kut00])

Ainsi lorsque la position de la source ou du récepteur, ou la fréquence de l'émetteur change légèrement, l'énergie subit de grandes variations. La salle a donc pour effet d'ajouter des variations aléatoires aux amplitudes et aux fréquences des partiels des sons. Nous pouvons observer cet effet sur les figures 5.4 et 5.5 qui représentent l'amplitude des partiels 1 à 9 d'un son d'alto et d'un son de cor

¹La fréquence de Schroeder est de l'ordre de 40 Hz dans une salle de concert comme l'Espace de Projection de l'Ircam

sans effet de salle, c'est-à-dire que l'enregistrement est réalisé en champs proche, et des mêmes sons auxquels sont ajoutés une réverbération par convolution d'une réponse impulsionnelle mesurée dans l'Espace de Projection de l'Ircam. Les amplitudes sont normalisées en divisant par la moyenne.

5.2.4 Conclusion

L'effet de salle, en ajoutant de l'aléatoire, crée des modulations différentes pour chaque son et chaque partiel des sons. Lorsque ses sons sont émis simultanément dans une salle, le signal reçu par un auditeur est donc très complexe. Les variations d'amplitude dans chaque canal auditif sont grandes, rapides et incohérentes. Nous pouvons donc supposer que les signaux s'ajoutent en moyenne en énergie, c'est-à-dire que le premier terme de l'équation 5.5 à une moyenne nulle sur un court intervalle de temps.

Nous n'avons pas parlé jusqu'ici de deux autres effets de la salle et de la transmission : le filtrage et l'atténuation du son. Dans cette thèse, nous supposerons que les instruments sont tous à la même distance de l'auditeur et que le filtrage effectué par la salle est le même pour tous les instruments.

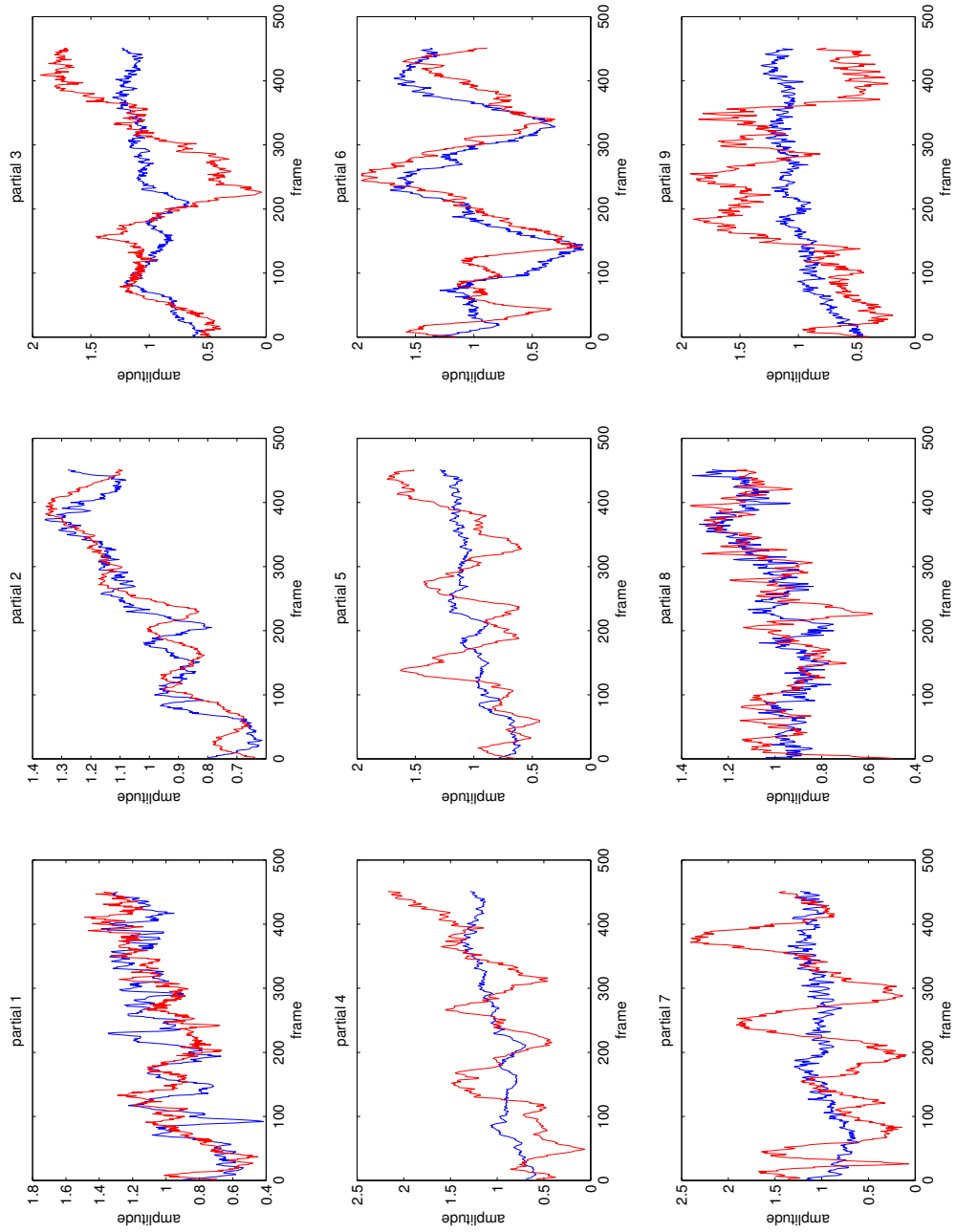


FIG. 5.4 – Amplitude des partiels d'un alto jouant un do_4 *mf* sans effet de salle (en bleu) et avec effet de salle (en rouge)

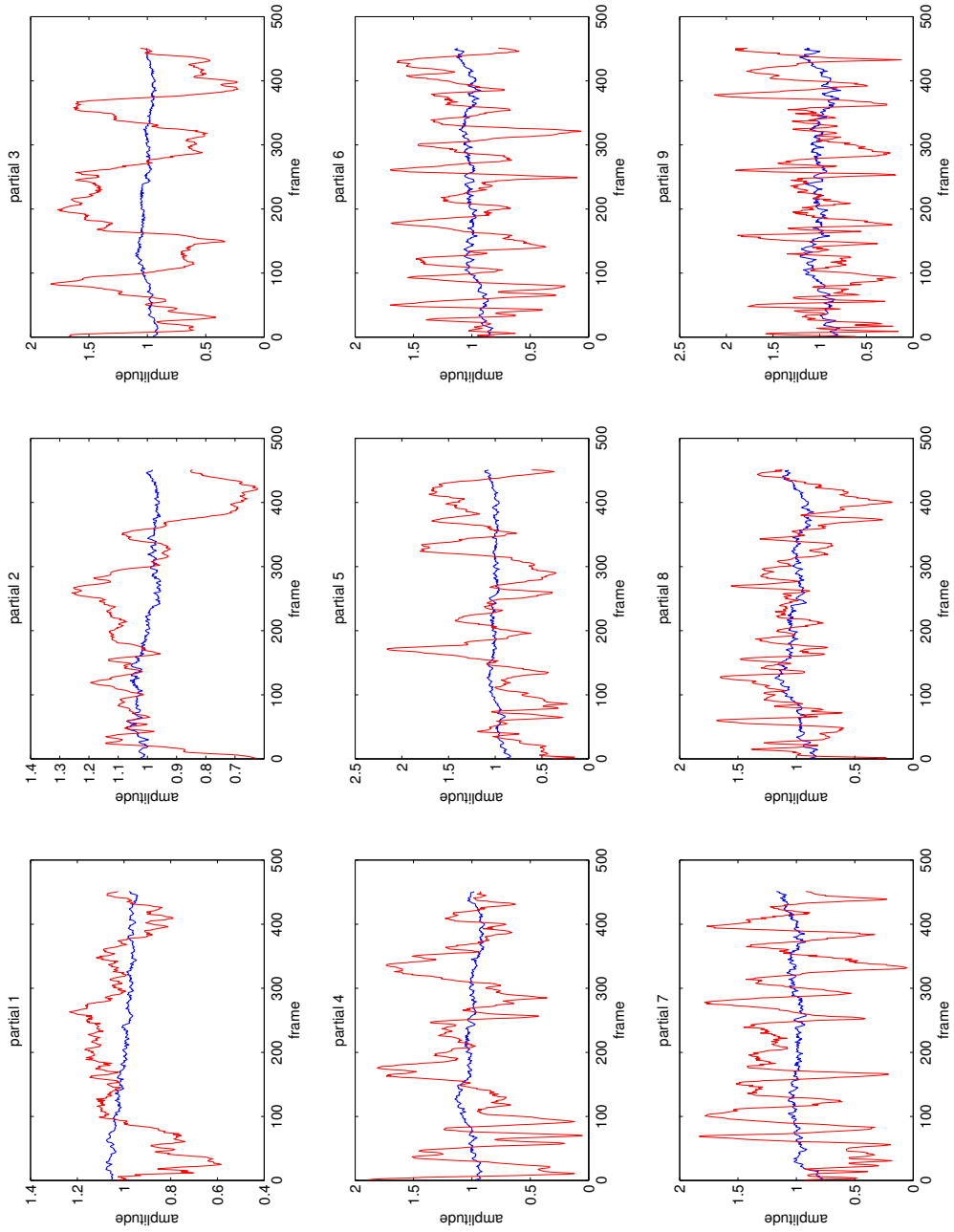


FIG. 5.5 – Amplitude des partiels d'un cor jouant un do4 *ff* sans effet de salle (en bleu) et avec effet de salle (en rouge)

5.3 Expériences perceptives

5.3.1 Introduction

Pour des sons entretenus, nous supposons que les spectres de puissance moyens s'additionnent. Mais cette hypothèse ne s'étend pas aux sons percussifs. Lorsqu'un son percussif et un son entretenu sont émis simultanément, ils créent, dans certaines conditions un objet sonore unique : ils fusionnent. Dans ce cas comment cet objet est-il perçu ? Pouvons-nous prédire la similarité entre ces objets à partir d'une description des sons qui les composent ? La réponse à ces questions ne se trouve pas dans les signaux eux-mêmes mais dans leur perception. Nous proposons donc deux expériences sur la perception de ces sons. Ces expériences ont été réalisées au laboratoire MPCL du département de Music Technology de l'université de McGill avec la collaboration du professeur Stephen McAdams. Dans la première expérience, nous étudions les conditions qui favorisent la fusion entre un son percussif et un son entretenu. Plus précisément nous cherchons les caractéristiques des sons du mélange qui expliquent et permettent de prédire le degré de fusion. Dans la seconde expérience, nous utilisons les combinaisons qui présentent un degré de fusion élevé dans une tâche de dissemblance par paires. Les résultats nous permettent de proposer un ensemble de descripteurs permettant de prédire la similarité perçue entre ces sons.

5.3.2 Expérience 1 : Jugement de fusion

5.3.2.1 Description

Le but de cette première expérience est de tester dans quelle mesure la combinaison d'un son entretenu et d'un son percussif peut fusionner perceptivement.

5.3.2.2 Stimuli

Onze sons entretenus et onze sons percussifs ont été utilisés. Les sons entretenus ont été extraits de la base Studio Online, les sons percussifs de la base Vienna Symphonic Library. Les instruments pour les sons entretenus sont le basson, la clarinette en Sib, le hautbois, la flûte, la trompette en ut, le cor, le trombone, le violon, le violoncelle et la contrebasse. Les instruments pour les sons percussifs sont la flûte pizzicato, le woodblock, le marimba forte et piano, le vibraphone forte et piano, deux pizzicati de violon différents et deux pizzicati de harpe, l'un est normal, l'autre est un harmonique. La liste des instruments, de leur famille et des abréviations utilisées est donnée dans les tables 5.1 et 5.2. Tous les sons sont des $\text{do}\sharp$ au-dessus du do_4 (ie. 277.2 Hz). Les sons ont été coupés pour obtenir une durée maximum de 2.5 secondes et une rampe descendante a été appliquée. La décroissance du marimba et du vibraphone a donc été raccourcie, mais la pente de décroissance n'a pas été modifiée. Les sons ont été normalisés en sonie par une tâche d'égalisation réalisée par 6 participants. Tous les mélanges possibles contenant un son percussif et un son entretenu ont été construits par simple addition pour donner un total de 121 stimuli.

5.3.2.3 Procédure

A chaque essai, les participants entendent un mélange. Ils notent dans quelle mesure ce mélange fusionne (“blend”) sur une échelle continue : la gauche de l’échelle porte la mention “twoness” pour indiquer une absence de fusion, la droite de l’échelle porte la mention “oneness” pour indiquer une fusion en un seul timbre. Les réponses sont données en ajustant un curseur avec la souris. Chaque participant note quatre fois les 121 stimuli en quatre sessions. Dans chaque session, les stimuli sont diffusés dans un ordre aléatoire. Les quatre sessions sont enchaînées sans pause. Pour que les participants se fassent une idée de la tâche, l’ensemble des mélanges est diffusé une fois avant de commencer la tâche. La durée de l’expérience est d’environ une heure.

Les sons sont stockés sur l’ordinateur en mono 16 bits à un taux d’échantillonnage de 44100 Hz et diffusés sur des haut-parleurs à un niveau d’écoute confortable.

Instrument	Famille	Abréviation
Clarinete en Sib	Bois	BbCl
Basson	Bois	Bn
Flûte en do	Bois	Fl
Hautbois	Bois	Ob
Contrebasse	Corde	Cb
Contrebasse, harmonique artificiel	Corde	Cb-ah
Violoncelle	Corde	Vc
Violon	Corde	Vn
Trompette en ut	Cuivre	CTp
Cor	Cuivre	Hn
Trombone	Cuivre	Tbn

TAB. 5.1 – Instruments utilisés pour les sons entretenus

Instrument	Famille	Abréviation
Woodblock	Block	Hbl-GaHa
Woodblock	Block	Hbl-Ho
Flûte pizzicato	Bois	Fl-pz
Violon, pizzicato Bartok	Corde	Vn-pz-bk
Violon, pizzicato ordinaire	Corde	Vn-pz-lv
Marimba forte	Lame	Ma-f
Marimba mezzo forte	Lame	Ma-mf
Vibraphone forte	Lame	Vib-f
Vibraphone piano	Lame	Vib-p
Harpe, pizzicato, harmonique artificiel	String	Hp-harm
Harpe, pizzicato	String	Hp-pz

TAB. 5.2 – Instruments utilisés pour les sons percussifs

5.3.2.4 Participants

Les participants, au nombre de 23, étaient tous des étudiants en musique de l'université de McGill. Leur audition était préalablement testée avec un audiogramme. Trois participants ont été rejetés pour cause de mauvais résultats au test d'audition et ont été payés 5 dollars canadiens. Les autres ont été payés 10 dollars canadiens.

5.3.2.5 Résultats

Deux participants ont été retirés parce que leurs résultats différaient trop de ceux des autres. Sur une échelle allant de 0 à 1, la moyenne des résultats va de 0.2 pour le violon pizzicato Bartok mélangé à la trompette à 0.85 pour le vibraphone piano mélangé au cor. L'écart type des résultats est à peu près similaire pour tous les mélanges, de l'ordre de 0.23. De tels mélanges peuvent donc être perçus comme fusionnant très bien et donner la sensation d'un seul timbre ou être perçu comme fusionnant très mal.

ANOVA Nous avons effectué deux analyses de variance sur les jugements de fusion avec respectivement l'instrument entretenu et l'instrument percussif comme variable (figures 5.6 et 5.7). Les deux variables ont un effet significatif sur la fusion perçue. L'effet de l'instrument percussif est cependant plus fort ($F(1,784)=25.4$ $p<0.0001$) que celui de l'instrument entretenu ($F(2,482)=9.9$ $p<0.006$). Cela pourrait indiquer que le choix de l'instrument percussif est plus important que le choix de l'instrument entretenu dans le contrôle de la sensation de fusion pour de tels mélanges. Nous observons aussi un effet de l'interaction des deux variables ($F(11,56)=4.5$ $p<0.0001$), montrant ainsi que le choix des instruments ne peut être fait indépendamment. Un point intéressant est que les valeurs moyennes dans la figure 5.6 sont très fortement corrélées négativement au centroïde spectral du son correspondant ($r^2 = .85$). Ainsi plus le centroïde du son percussif est faible (moins le son est brillant) plus ce son va fusionner facilement. Ceci est en accord avec les résultats de [San95].

Corrélation avec des descripteurs acoustiques Nous étudions ensuite la corrélation entre les jugements et un ensemble de descripteurs acoustiques dont la liste est donnée en annexe. Chacun de ces descripteurs est calculé sur le son percussif de la combinaison, sur le son entretenu et sur le mélange des deux. Nous calculons aussi la différence entre le descripteur du son entretenu et celui du son percussif.

La meilleure corrélation est obtenue pour le centroïde spectral du son percussif calculé sur une échelle de fréquence divisée en bande de Bark ($r^2 = .68$), ce qui confirme ce que nous avons constaté au paragraphe précédent. Si nous ne prenons pas en compte les descripteurs des sons seuls, le plus corrélé est la différence entre les centroïdes temporels ($r^2 = .62$). Mais ce descripteur est très corrélé au centroïde temporel du son percussif parce que les centroïdes temporels des sons entretenus varient peu. Il est également très corrélé au centroïde spectral du son percussif ($r^2 = .63$). Il est donc difficile de savoir si la perception de fusion pour ce type de mélanges est plus influencée par des critères spectraux ou temporels. De plus, l'intervalle de variation des descripteurs est beaucoup plus petit pour les sons entretenus que pour les sons percussifs, l'influence de ces derniers est donc beaucoup plus visible. Finalement, en terme pratique, nous pouvons dire que pour qu'un mélange de ce type fusionne, il

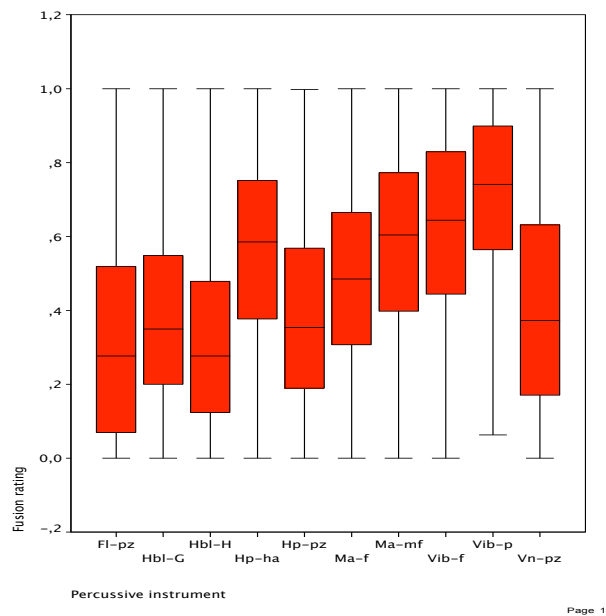


FIG. 5.6 – Moyenne des jugements de fusion pour les instruments percussifs

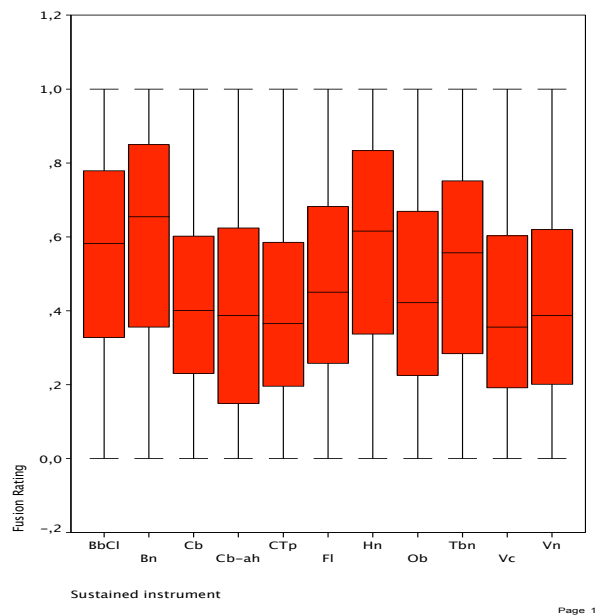


FIG. 5.7 – Moyenne des jugements de fusion pour les instruments entretenus

faut que le son percussif ait un centroïde spectral bas (qu'il soit peu brillant), ou qu'il ait une attaque douce, ce qui pour les sons considérés ici revient quasiment au même.

5.3.3 Expérience 2 : similarité entre mélanges

5.3.3.1 Description

Le but de cette expérience est d'étudier la similarité perçue entre des mélanges de sons contenant un son percussif et un son entretenu. Nous souhaitons aussi trouver une description acoustique du signal des mélanges permettant d'expliquer la similarité entre mélange.

5.3.3.2 Stimuli

Nous utilisons 16 sons de l'expérience précédente. Les sons ont été sélectionnés manuellement selon deux critères. Les paires sélectionnées doivent obtenir un jugement de fusion élevé dans l'expérience précédente et leurs timbres doivent être variés. L'ensemble des mélanges sélectionnés est donné dans la table 5.3.

Son percussif	Son entretenu
HBL-GaHa	Bn
Hp-harm	BbCl
Hp-harm	Bn
MA-f	Bn
MA-mf	Fl
MA-mf	Hn
MA-mf	Tbn
Vib-f	Cb
Vib-f	Vc
Vib-p	CTp
Vib-p	Hn
Vib-p	Ob
Vib-p	Vn
Vn-pz-lv	BbCl
Vn-pz-lv	Bn
Vn-pz-lv	Tbn

TAB. 5.3 – Mélanges sélectionnés pour la seconde expérience

5.3.3.3 Procédure

La tâche consiste à estimer la dissemblance entre chaque paire possible des 16 mélanges sur une échelle continue allant de “très proche” (very close) à “très distant” (very distant). Chaque paire est présentée une fois, l'ordre de présentation est aléatoire. Les participants peuvent répéter les sons d'une

paire autant de fois qu'ils le souhaitent. Le nombre total de comparaisons à effectuer pour chaque participant est de 120.

5.3.3.4 Participants

Les participants, au nombre de 25, étaient tous des étudiants en musique de l'université de McGill. Leur audition était préalablement testée avec un audiogramme. Quatorze d'entre eux avaient participé à l'expérience précédente. Aucun participant n'a raté l'audiogramme. Ils étaient payés 10 dollars canadiens.

5.3.3.5 Résultats

Trois participants ont été retirés car ils avaient des résultats très différents des autres.

ANOVA Nous testons l'influence des variables “les deux instruments percussifs appartiennent à la même famille” et “les deux instruments entretenus appartiennent à la même famille” sur les jugements moyens de dissemblance. Les deux variables ont une influence sur la similarité perçue ($F(1)=15$ $p<.005$ et $F(1)=26$ $p<.005$), mais nous n'observons pas d'effet de synergie. Dans les deux cas, si les instruments appartiennent à la même famille, la distance est plus faible (voir figure 5.8 et 5.9). Cependant l'appartenance à une même famille pour les sons entretenus a plus d'influence sur la similarité globale.

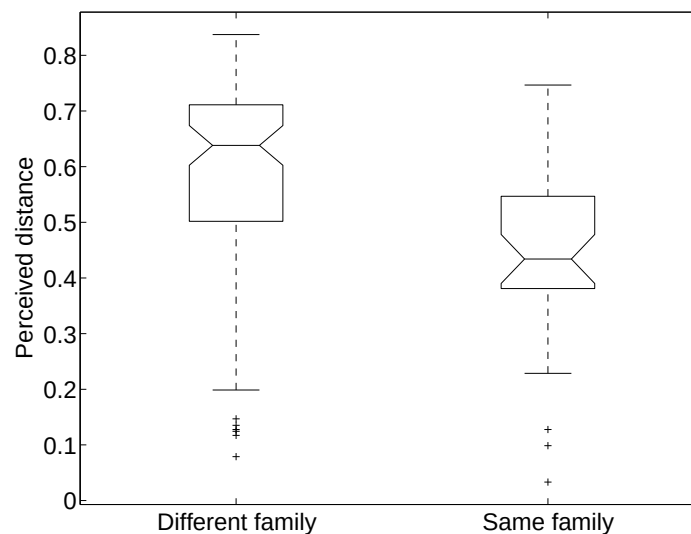


FIG. 5.8 – Dissemblance perçue dans l'expérience 2 lorsque les sons percussifs appartiennent à la même famille ou non

Corrélation Nous voulons maintenant prédire les jugements de dissemblance à partir de descripteurs acoustiques. Idéalement, nous voudrions que cette prédiction ne se base que sur les descripteurs

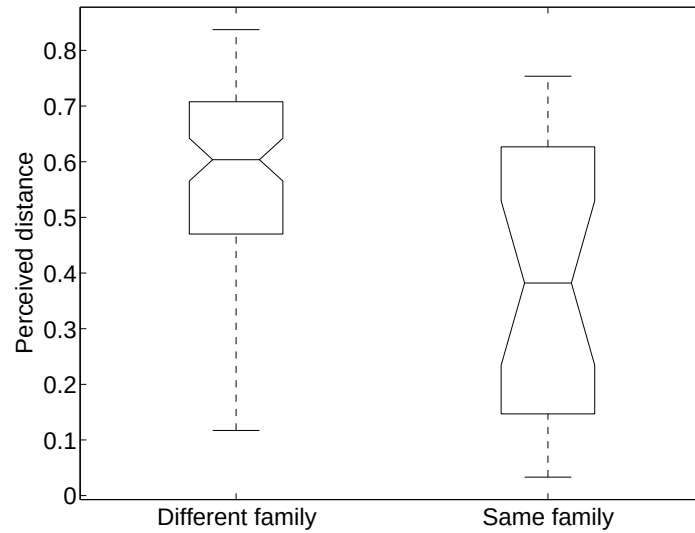


FIG. 5.9 – Dissemblance perçue dans l'expérience 2 lorsque les sons entretenus appartiennent à la même famille ou non

des sons qui composent le duo, et non sur le signal du duo lui-même, pour que le résultat soit applicable au problème de l'orchestration tel que nous l'avons formulé. Pour déterminer le potentiel prédictif d'un descripteur nous calculons la corrélation entre les distances entre sons induites par ce descripteur et les distances moyennes données par les participants de l'expérience. Nous testons les mêmes descripteurs que dans l'expérience précédente. Les distances sont assez bien expliquées par la distance euclidienne entre les spectres de log-amplitude sur une échelle de Mel des sons entretenus ($r^2 = .61$). La corrélation obtenue pour le même descripteur sur les sons percussifs est faible. En revanche si nous calculons ce descripteur sur le signal du duo nous obtenons une corrélation plus élevée ($r^2 = .7$). Si nous estimons le spectre de Mel du duo en supposant que les spectres moyens s'additionnent en énergie, la corrélation obtenue avec cette estimation est de .81 ($r^2 = .65$). Cette approximation n'a aucun fondement acoustique, mais elle est pratique puisque c'est ce que nous supposons pour les sons entretenus. Ces résultats montrent que le spectre du son entretenu a plus d'influence sur les jugements de similarité que celui du son percussif, ce qui est en accord avec les résultats des ANOVA. Utiliser la somme des spectres de puissance donne un descripteur intéressant pour la prédiction de la similarité.

Le descripteur qui améliore le mieux les résultats ensuite est un descripteur ad-hoc. Il s'agit du logarithme du centroïde temporel du son entretenu divisé par celui du son percussif. En utilisant ce descripteur en combinaison avec le spectre de Mel du mélange, nous obtenons une corrélation de .89 ($r^2 = .8$). La figure 5.10 montre les valeurs de prédiction obtenues.

5.3.4 Conclusion

Par ces expériences, nous avons montré qu'un mélange constitué d'un son percussif et d'un son entretenu pouvait fusionner et ainsi donner la sensation d'un timbre unique. Nous avons proposé un critère basé sur le centroïde temporel du son percussif ou sur son centroïde spectrale permettant de

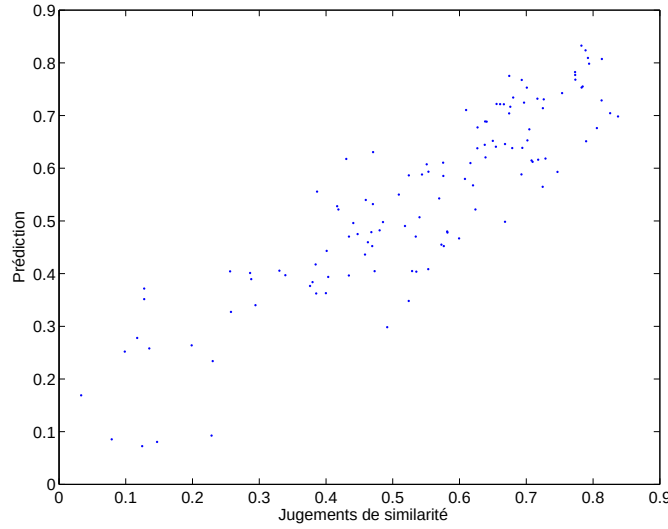


FIG. 5.10 – Prédiction des jugements de similarité par une combinaison linéaire du spectre de Mel en log et du rapport des logarithmes des centroïdes temporels des deux sons

prédire la fusion entre les deux sons. Étant donné la corrélation entre les critères spectraux et temporels sur les sons considérés, il n'a pas été possible de distinguer leurs effets respectifs. Dans la seconde expérience nous avons proposé une description des mélanges permettant une bonne prédiction de la similarité perçue. Cette description est basée sur le logarithme du spectre de Mel et le rapport entre les logarithmes des centroïdes temporels des sons du mélange. Nous avons aussi montré qu'une bonne estimation pouvait être obtenue par l'addition des spectres de Mel ce qui nous sera très utile par la suite.

5.4 Formules d'additions

5.4.1 Introduction

Dans cette partie, nous utilisons les résultats des parties précédentes pour proposer des formule d'estimation des descripteurs des combinaisons. Notre but est de proposer des formules permettant d'estimer les descripteurs d'une combinaison à partir des descripteurs des sons qui la composent. Nous noterons $c = x_1, \dots, x_M$ une combinaison composée des M sons $x_1, \dots, x_M$. Dans le chapitre 6 nous modéliserons la distribution des descripteurs par une loi normale. Nous donnons donc aussi dans ce chapitre, les formules permettant "d'additionner" les distributions des descripteurs.

5.4.2 Les descripteurs du spectre

Nous supposons que les spectres de puissance à court terme s'additionnent. Cette hypothèse est souvent utilisée pour la séparation de sources (voir paragraphe 2.5) et a été justifiée dans les parties précédentes de ce chapitre. Nous supposons de plus que les spectres moyens de puissance s'additionnent pour les sons considérés. Ainsi pour un ensemble de sons de spectre de puissance moyen et normalisé

en énergie d_m et d'énergie moyenne e_m , le spectre moyen de la combinaison de ces sons est donné par la formule suivante :

$$e_c d_c = \sum_{m=1}^M e_m d_m \quad (5.6)$$

5.4.3 Le centroïde spectral

Sachant que les spectres de puissance s'additionnent et que le centroïde spectral est calculé sur le spectre de puissance, le centroïde d'une combinaison est égal à la moyenne des centroïdes des composants pondérés par l'énergie :

$$s_{cc} = \frac{\sum_{m=1}^M e_m s_{cm}}{\sum e_m} \quad (5.7)$$

Cette formule est valable pour le centroïde du spectre total et du spectre de la partie non-sinusoidale.

5.4.4 La noisiness

La noisiness d'une combinaison est égale à la somme des noisiness des composants pondérés par l'énergie. En effet, la noisiness du son x_m est :

$$N_m = \frac{e_{n_m}}{e_m} \quad (5.8)$$

Où e_{n_m} est l'énergie de la partie non-sinusoidale et e_m est l'énergie totale du son x_m .

Comme les énergie s'additionnent, l'énergie totale de la combinaison est $e_{tc} = \sum_{m=1}^M e_m$ et l'énergie de la partie non-sinusoidale est $e_{nc} = \sum_{m=1}^M e_{n_m}$. Ainsi la noisiness de la combinaison est égale à :

$$N_c = \frac{e_{nc}}{e_c} \quad (5.9)$$

$$= \frac{\sum_{m=1}^M e_{n_m}}{\sum_{m=1}^M e_m} \quad (5.10)$$

$$= \frac{\sum_{m=1}^M e_m N_m}{\sum_{m=1}^M e_m} \quad (5.11)$$

5.4.5 L'ambitus de la modulation d'énergie

Pour l'ambitus de la modulation d'énergie, il n'est pas possible de trouver une formule analytique théorique. En effet, les modulations considérées sont à des fréquences différentes, l'addition de deux signaux complexes modulés en énergie donne un signal dont il est très difficile de prévoir les modulations. Nous utilisons une formule heuristique, qui consiste à supposer que l'ambitus de la modulation résultante est la somme des ambitus de modulation de chaque son pondérée par l'énergie de chaque son.

$$a_c = \frac{\sum_{m=1}^M e_m a_m}{\sum_{m=1}^M e_m} \quad (5.12)$$

Il semble pertinent de supposer que plus un son est fort dans une mixture plus c'est sa propre modulation qui va prendre le dessus perceptivement. L'écoute de combinaisons synthétiques de sons instrumentaux confirme cette hypothèse.

Nous avons testé cette hypothèse en comparant la modulation mesurée sur les mixtures synthétiques du chapitre 8 avec l'estimation proposée. Les corrélations obtenues sont de l'ordre de 0,86 pour des mixtures de deux sons et de 0,77 pour des mixtures de trois sons. Les figures 5.11 et 5.12 montrent les résultats obtenus respectivement pour les mixtures de deux et trois sons. La droite affichée correspond au résultat désiré, c'est-à-dire une droite de pente 1 et d'ordonnée à l'origine 0. Nous constatons que, outre une bonne corrélation, nous obtenons des valeurs correctes. Cette vérification n'a, bien entendu aucune signification perceptive.

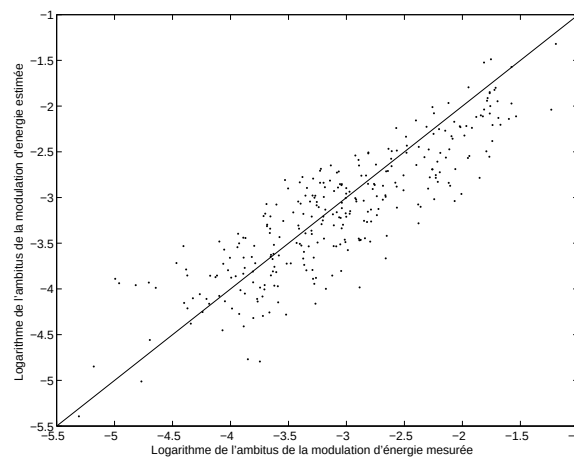


FIG. 5.11 – Prédiction du logarithme de l'ambitus de modulation pour de combinaisons de 2 sons

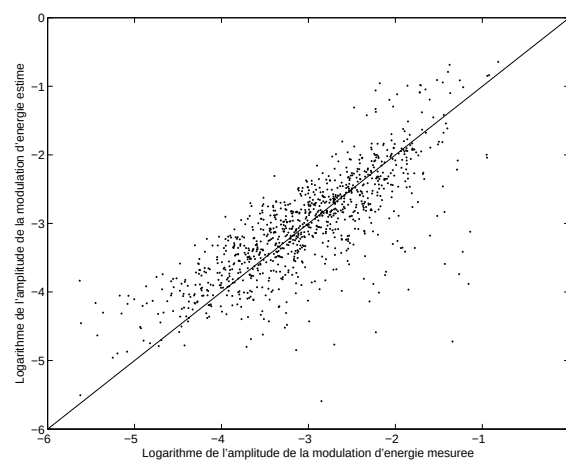


FIG. 5.12 – Prédiction du logarithme de l'ambitus de modulation pour de combinaisons de 3 sons

5.4.6 Le logarithme du temps d'attaque

Comme pour la modulation d'énergie nous utilisons une formule heuristique pour le logarithme du temps d'attaque :

$$lat_c = \frac{\sum_{m=1}^M l_m lat_m}{\sum_{m=1}^M l_m} \quad (5.13)$$

Où l_m est la sonie du son m . Nous avons testé cette hypothèse sur la même base que précédemment, les corrélations obtenues sont de l'ordre de 0,7.

5.5 Distribution des descripteurs des combinaison

Dans le chapitre 6 nous proposons une modélisation de la distribution des descripteurs des sons par des loi normales. Nous avons donc besoin d'une méthode permettant de calculer la distribution des descripteurs d'une combinaison de sons à partir de la distribution des descripteurs des sons qui la composent. Ces méthodes se basent sur les formules précédentes et sur le type de distributions utilisées. Les formules sont toutes des sommes de descripteurs pondérées par l'énergie. Deux cas se présentent ensuite. Soit la variable de la somme est le descripteur lui-même et nous avons affaire à la somme pondérée de variables aléatoires normales, soit la variable de la somme et le logarithme du descripteur, comme c'est le cas pour la noisiness, et nous sommes confronté à une somme pondérée de variables aléatoires de distribution log-normale². Nous présentons les deux cas pour un descripteur d et une combinaison $c = x_1, \dots, x_M$. Chaque son x_m a pour descripteur d_m et pour énergie e_m . Dans tous les cas le descripteur d_c de la combinaison est égale à :

$$d_c = \frac{\sum_{m=1}^M e_m d_m}{\sum_{m=1}^M e_m} \quad (5.14)$$

5.5.1 Cas d'une distribution normale

Supposons $f(d_m) = \mathcal{N}(\mu_m, \sigma_m^2)$. Alors, $f(d_c) = \mathcal{N}(\mu_c, \sigma_c^2)$. avec :

$$\mu_c = \frac{\sum_{m=1}^M e_m \mu_m}{\sum_{m=1}^M e_m} \quad (5.15)$$

et,

$$\sigma_c^2 = \frac{\sum_{m=1}^M e_m^2 \sigma_m^2}{\sum_{m=1}^M e_m^2} \quad (5.16)$$

Ces formules s'étendent sans problème au cas multidimensionnel avec des matrices de covariance pleine.

5.5.2 Cas d'une distribution log-normale

Supposons $f(d_m) = \mathcal{LN}(\mu_m, \sigma_m^2)$. Cette notation est équivalente à $f(\log(d_m)) = \mathcal{N}(\mu_m, \sigma_m^2)$. En effet, les paramètres usuelles de la loi log-normale ne sont pas ses moments mais ceux de la loi

²Si le logarithme d'une variable aléatoire a une distribution normale, alors la variable a une distribution log-normale

normale associée.

Il n'existe pas d'expression analytique de la distribution d'une somme de variables aléatoires log-normales. Plusieurs méthodes d'approximation ont été proposées [Szy07]. Nous utilisons la méthode de Fenton-Wilkinson [Fen60]. Cette méthode consiste à supposer que la somme de distributions log-normale est elle-même log-normale et à estimer ses deux premiers moments.

Pour alléger l'écriture nous donnons les formules pour une somme simple. Les formules pour une somme pondérée se déduisent simplement en rappelant que $f(\log(e_m d_m)) = f(\log(e_m) + \log(d_m)) = \mathcal{N}(\mu_m + \log(e_m), \sigma_m^2)$.

Posons $d_c = \sum_{m=1}^M d_m$. Alors en supposant que $f(d_c) = \mathcal{LN}(\mu_c, \sigma_c^2)$ et que les d_m ne sont pas corrélés, la méthode de Fenton-Wilkinson approche les deux premiers moments de la distribution de d'_c par :

$$E(d_c) = \sum_{m=1}^M e^{\mu_m + \frac{1}{2}\sigma_m^2} \quad (5.17)$$

$$E(d_c^2) = \sum_{m=1}^M e^{2\mu_m + 2\sigma_m^2} + 2 \sum_{m=1}^M \sum_{j>m}^M e^{\mu_m + \mu_j + \frac{1}{2}(\sigma_m^2 + \sigma_j^2)} \quad (5.18)$$

A partir des moments, on retrouve les paramètres de la loi de d_c :

$$\mu_c = 2\ln E(d_c) - \frac{1}{2}\ln E(d_c^2) \quad (5.19)$$

$$\sigma_c^2 = -2\ln E(d_c) + \ln E(d_c^2) \quad (5.20)$$

Cette méthode ne s'étend pas à des distributions multidimensionnelles à matrices de covariance pleines. Par la suite nous supposerons que les matrices de covariance sont diagonales lorsque nous utiliserons cette méthode.

Chapitre 6

Modèle d'instrument

6.1 Introduction

Dans ce chapitre, nous présentons le modèle utilisé pour décrire le timbre des instruments. Nous proposons un modèle probabiliste génératif des descripteurs du timbre. Le modèle est donc constitué de la distribution des descripteurs sachant l'instrument et le mode de jeu utilisés ainsi que la hauteur et la nuance de la note jouée. La méthode d'apprentissage proposée est basée sur deux transformations de la distribution. La première transformation, décrite dans la partie 6.4, consiste à apprendre la distribution jointe d'un descripteur, de la fréquence fondamentale et de l'énergie et de déduire de celle-ci la distribution recherchée. Dans la seconde transformation, décrite dans la partie 6.5, nous considérons la distribution recherchée comme le résultat d'un processus de décision collective basé sur la taxinomie de mode de jeu proposée au chapitre 3. Chaque variable de la taxinomie est associée à un ensemble d'experts qui fournissent une distribution pour le descripteur. L'agrégation de ces distributions, décrite dans la partie 6.6, permet d'obtenir la distribution recherchée.

6.2 Distribution modélisée

Nous souhaitons connaître la valeur des descripteurs d'un instrument pour une hauteur, une nuance et un mode de jeu donnés. Nous décrivons cette valeur par la densité de probabilité $f((\mathbf{d}_j)|i, h, n, m)$ de l'ensemble des descripteurs $(\mathbf{d}_j)_{j \in \{1, \dots, J\}}$ sachant les variables d'état, c'est-à-dire l'instrument i , la hauteur h , la nuance n et le mode de jeu m . $(\mathbf{d}_j)_{j \in \{1, \dots, J\}}$ représente l'ensemble des descripteurs du chapitre 4, ainsi un $\mathbf{d}_j$ peut être multidimensionnel, comme dans le cas des partiels, ou unidimensionnel, comme le centroïde spectral. Nous le considérons toujours comme un vecteur, éventuellement de dimension 1. Pour une utilisation dans un contexte multicritères, nous avons besoin de la distribution de chaque descripteur séparément. Nous supposons donc l'indépendance conditionnelle des descripteurs sachant les variables d'état, ainsi

$$f((\mathbf{d}_j)|i, h, n, m) = \prod_j f(\mathbf{d}_j|i, h, n, m) \quad (6.1)$$

Supposer l'indépendance conditionnelle revient à négliger la corrélation entre les descripteurs pour des variables d'état données.

La figure 6.1 représente les dépendances entre variables sous forme de réseau bayésien pour cette formulation du modèle.

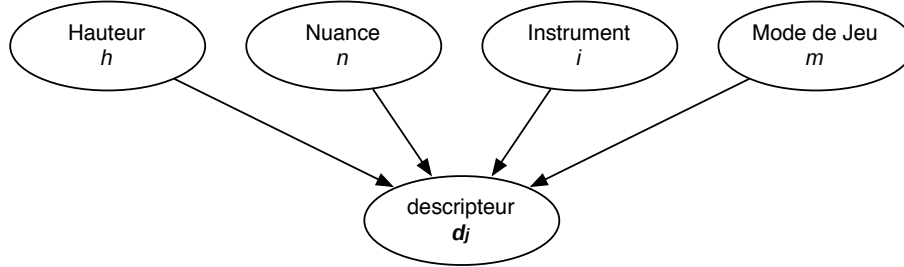


FIG. 6.1 – Dépendances entre les variables d'état et un descripteur pour le modèle de l'équation 6.1

6.3 Problème posé par la formulation.

L'équation 6.1 suppose l'apprentissage d'une distribution par descripteur et par combinaison de variables d'état, il faudrait donc disposer de plusieurs représentants de chaque hauteur, nuance et mode de jeu pour obtenir des modèles robustes. Construire une telle base d'apprentissage semble difficile sinon impossible, de plus il serait souhaitable que pour ajouter de nouveaux sons (un nouvel instrument par exemple), il suffise de donner seulement quelques représentants. Enfin les distributions de l'équation 6.1 permettent peu d'extrapolation ou du moins ne facilitent pas l'extrapolation de sons absents de la base d'apprentissage. Par exemple si l'ensemble d'apprentissage contient une trompette jouant un sol3 et un la3, déterminer la distribution de la trompette jouant un sol#3 suppose d'interpoler entre les distributions du sol3 et du la3. Nous proposons deux modifications de la distribution permettant un apprentissage plus robuste et l'extrapolation simple de modèles absents de l'ensemble d'apprentissage.

6.4 Grouper hauteurs et nuances

6.4.1 Principe

La première modification consiste à grouper les différentes hauteurs et nuances pour apprendre la distribution suivante :

$$f(\mathbf{d}_j, f_0, \log(e)|p) \quad (6.2)$$

Où $p = (i, m)$ est une variable d'état représentant conjointement l'instrument et le mode de jeu. Ici, la hauteur et la nuance ne sont plus vues comme des états mais comme des variables continues, respectivement la fréquence fondamentale f_0 et le logarithme de l'énergie e . Cette méthode est utilisée en classification des instruments de musique [Pee03] avec des gaussiennes avec la différence qu'il apprend la distribution jointe de tous les descripteurs. Elle est aussi utilisée pour la prédiction de la fréquence fondamentale de la voix à partir de l'enveloppe spectrale [ENRC03] ou pour la prédiction

de l'enveloppe spectrale à partir de la fréquence fondamentale [KM98] avec des mélanges de gaussiennes.

Le graphique de la figure 6.2 montre le réseau de dépendance obtenu pour chaque descripteur. Cette modification de la distribution à apprendre permet d'avoir beaucoup plus d'instances dans l'ensemble d'apprentissage et de déduire simplement la distribution pour n'importe quelle hauteur ou nuance par une interpolation implicite, mais suppose une dépendance linéaire entre la fréquence fondamentale et l'énergie d'une part et le descripteur d'autre part.

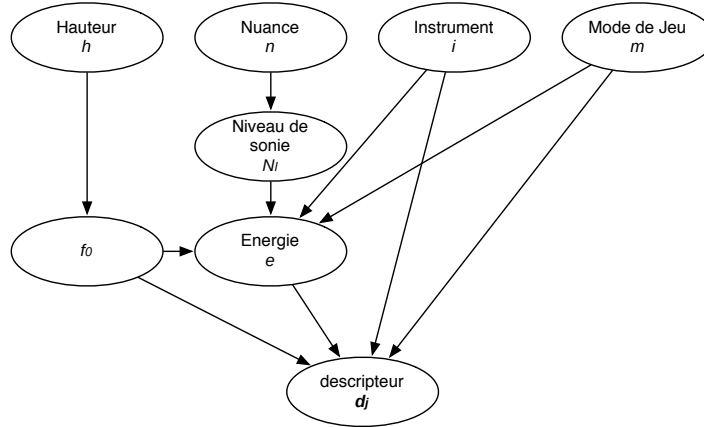


FIG. 6.2 – Dépendance entre les variables pour le modèle permettant de regrouper les hauteurs et les nuances dans les ensembles d'apprentissage

Le calcul de la distribution $f(d_j|p, h, n)$ à partir de la distribution 6.2 se fait en deux temps. D'abord nous calculons l'énergie correspondant à la nuance n et la f_0 correspondant à la hauteur h . Puis nous calculons la distribution $f(d_j|p, \log(e), f_0)$ en utilisant les valeurs de f_0 et d'énergie précédemment obtenues.

6.4.2 Relation entre le niveau de sonie et l'énergie

L'utilisation de la distribution 6.2 suppose la connaissance de la relation entre la hauteur et la fréquence fondamentale d'une part et entre la nuance et l'énergie d'autre part. La première relation est évidente, mais la seconde est plus complexe car elle dépend des autres variables d'état. En d'autres termes la connaissance de la distribution 6.2 permet d'obtenir la distribution d'un descripteur en fonction de l'énergie du signal, mais nous ne connaissons pas a priori la valeur de cette énergie pour différentes nuances de jeu. Il faut donc l'apprendre. Nous connaissons la relation entre la nuance de jeu et le niveau de sonie (voir paragraphe 4.2.1). Nous introduisons le niveau de sonie N_l comme une variable intermédiaire entre la nuance et l'énergie et apprenons la distribution jointe de l'énergie, de la fréquence fondamentale et du niveau de sonie sachant la variable d'état :

$$f(N_l, f_0, \log(e)|p) \quad (6.3)$$

De cette distribution, nous pouvons déduire la distribution de l'énergie pour différentes hauteurs et nuances et différentes valeurs des variables d'états en utilisant la même méthode que dans le para-

graphe 6.4.3.

$$f(\log(e)|N_l, f_0, p) \quad (6.4)$$

Nous utilisons la moyenne de cette dernière distribution comme valeur d'énergie. Cette méthode permet de décrire de manière simple le lien entre la nuance de jeu d'un instrument et l'énergie qui lui est associée. Il faut cependant que le niveau de sonie et la variable décrivant l'énergie suivent une relation linéaire, c'est pourquoi nous utilisons le logarithme de l'énergie et non l'énergie directement. Le log de l'énergie et le niveau de sonie sont reliés par une relation linéaire comme le montre à titre d'exemple la figure 6.3 où sont représentés le niveau de sonie et le log de l'énergie de sons de contre-basse et de trompette, ainsi que les droites de régression entre les deux descripteurs. Dans les deux figures, on peut voir les trois groupes formés par les trois nuances *pp*, *mf* et *ff*.

Notons que la distribution 6.3 est de la même forme que celle des descripteurs (équation 6.2), le niveau de sonie est donc considéré comme un descripteur dans la procédure d'apprentissage.

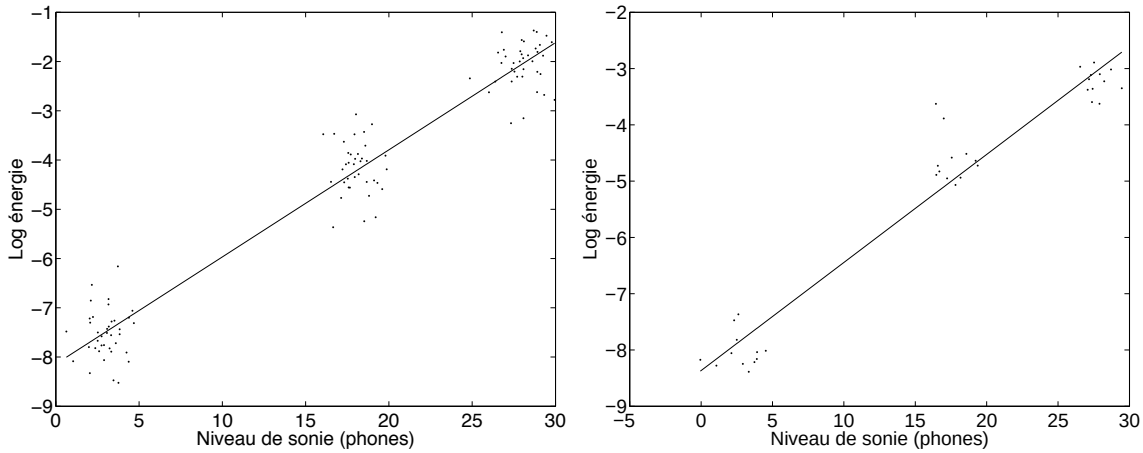


FIG. 6.3 – Niveau de sonie et log énergie pour des sons de contrebasse à l'octave 2 à gauche et de trompette à l'octave 4 à droite

6.4.3 Calcul de la distribution conditionnelle

A partir de la distribution 6.2 et des valeurs d'énergie et de fréquence fondamentale obtenues précédemment nous calculons la distribution $f(\mathbf{d}_j|p, h, n)$.

Pour simplifier l'écriture, posons $\mathbf{v} = (f_0, \log(e))$. Dans le cas gaussien, c'est-à-dire

$$f(\mathbf{d}_j, f_0, \log(e)|p) = f(\mathbf{d}_j, \mathbf{v}|p) = \mathcal{N}(\mathbf{d}_j, \mathbf{v}; \boldsymbol{\mu}_{j,p}, \boldsymbol{\Sigma}_{j,p}) \quad (6.5)$$

les paramètres de la distribution $f(\mathbf{d}_j|p, \log(e), f_0) = \mathcal{N}(\mathbf{d}_j; \boldsymbol{\mu}'_{j,p}, \boldsymbol{\Sigma}'_{j,p})$ sont déduits de 6.5 de la manière suivante :

1. Nous décomposons $\mu_{j,p}$ en leurs éléments correspondant à $\mathbf{d}_j$ et $\mathbf{v}$.

$$\mu_{j,p} = \begin{pmatrix} \mu_{j,p}^{\mathbf{v}} \\ \mu_{j,p}^{\mathbf{d}_j} \end{pmatrix} \quad \text{et} \quad \Sigma_{j,p} = \begin{pmatrix} \Sigma_{j,p}^{\mathbf{v}} & \Sigma_{j,p}^{\mathbf{v}\mathbf{d}_j} \\ \Sigma_{j,p}^{\mathbf{d}_j\mathbf{v}} & \Sigma_{j,p}^{\mathbf{d}_j} \end{pmatrix} \quad (6.6)$$

2. Nous obtenons :

$$\mu'_{j,p} = \mu_{j,p}^{\mathbf{d}_j} + \Sigma_{j,p}^{\mathbf{d}_j\mathbf{v}} \Sigma_{j,p}^{\mathbf{v}\mathbf{v}}^{-1} (\mathbf{v} - \mu_{j,p}^{\mathbf{v}}) \quad (6.7a)$$

$$\Sigma'_{j,p} = \Sigma_{j,p}^{\mathbf{d}_j} + \Sigma_{j,p}^{\mathbf{d}_j\mathbf{v}} \Sigma_{j,p}^{\mathbf{v}\mathbf{v}}^{-1} \Sigma_{j,p}^{\mathbf{v}\mathbf{d}_j} \quad (6.7b)$$

6.5 Mode de Jeu

6.5.1 Motivations

La seconde modification consiste à supposer que les différents modes de jeu ont des effets similaires sur les instruments. Par exemple l'effet *tremolo* se caractérise par une modulation périodique d'énergie et un rapport harmonique sur bruit faible quel que soit l'instrument à cordes frottées utilisé. Ainsi il est possible d'apprendre d'une part les caractéristiques liées à l'instrument et d'autre part les caractéristiques liées au mode de jeu et finalement d'agréger les descriptions. Cette approche présente plusieurs avantages. Premièrement, en regroupant les différents échantillons d'un même mode de jeu pour l'apprentissage, les modèles appris sont plus robustes, grâce à un ensemble d'apprentissage plus grand, et plus généraux. Prenons l'exemple du vibrato et des cordes frottées : l'instrumentiste qui a enregistré les sons vibrato d'alto a, pour des raisons d'habitude ou par hasard, joué le vibrato avec une certaine fréquence et un certain ambitus. Un modèle simple prendrait ces valeurs pour représentatives de l'alto, ce qui n'est évidemment pas le cas, et pourrait éliminer l'alto dans une orchestration sous prétexte qu'il ne peut pas jouer, par exemple, de vibrato plus rapide. Grouper tous les vibratos permet d'apprendre un modèle plus général de vibrato en supposant que les différents instruments ont les mêmes possibilités. Deuxièmement, les modèles de modes de jeu absents de la base d'apprentissage peuvent être déduits par combinaison. La figure 6.4 montre un exemple dans lequel, en appliquant le modèle de vibrato appris sur les cordes nous déduisons le modèle d'une flûte vibrato. La figure représente le nombre d'échantillons disponibles pour chaque type de son. En bas et à droite se trouve le nombre d'échantillons obtenus si on regroupe par instrument et par présence ou absence de vibrato. Dans ce cas, il est possible d'apprendre des modèles génériques de violon, violoncelle et flûte d'une part, vibrato et non vibrato d'autre part et enfin de combiner les modèles obtenus. Nous pouvons donc extrapoler un modèle approximatif de flûte vibrato sans en avoir d'échantillon. Une fois de plus, cela suppose que les modes de jeu ont les mêmes effets sur tous les instruments, ce qui dans le présent exemple n'est pas tout à fait juste puisque le vibrato de flûte présente beaucoup moins de modulation de fréquence que le vibrato de corde, donc l'application du modèle de vibrato appris sur des cordes à une flûte constitue une approximation. Nous supposons cette approximation valide dans notre cadre et choisissons de favoriser une plus grande généralité des modèles à une trop grande précision.

	Violin	Cello	Flute	
Vibrato	100	100	0	Vibrato model 200
Non vibrato	30	30	100	Non vibrato model 160
	Violin model 130	Cello model 130	Flute model 100	

FIG. 6.4 – Exemple de regroupement des modes de jeu pour l'apprentissage

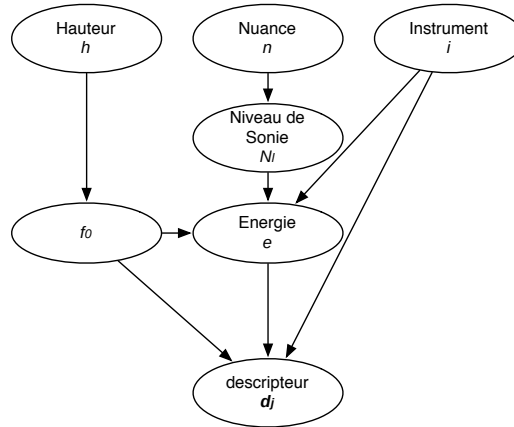
6.5.2 Factorisation de la distribution

Ces hypothèses se traduisent par une transformation de la distribution 6.2 en une fonction de distributions élémentaires. Soit $p = (p^k)_{k \in 1, K}$ un ensemble de variables d'état décrivant l'instrument, et les modes de jeu, soit $f_{j,k} = f(\mathbf{d}_j, f_0, \log(e)|p^k)$ la distribution jointe de $\mathbf{d}_j$, f_0 et $\log(e)$ connaissant l'état p^k . Alors nous postulons que :

$$f(\mathbf{d}_j, f_0, \log(e)|p^k) = T(f_{j,1}, \dots, f_{j,K}) \quad (6.8)$$

$$\text{avec } f_{j,k} = \mathcal{N}(\mathbf{d}_j, f_0, \log(e); \boldsymbol{\mu}_{j,k}, \boldsymbol{\Sigma}_{j,k}) \quad (6.9)$$

Où T est une fonction d'agrégation des distributions. Cette formulation transforme le problème de l'apprentissage de la distribution 6.2 en l'apprentissage d'une distribution $f_{j,k}$ pour chaque mode de jeu élémentaire et chaque descripteur et un problème de prise de décision collective par un ensemble d'*experts* [GZ86]. L'agrégation des distributions élémentaires est traitée au paragraphe 6.6.

FIG. 6.5 – Dépendance entre les variables pour la variable d'état *instrument*. Le réseau est le même pour les différents *modes de jeu* de la taxinomie

Nous choisissons la loi normale pour des raisons de simplicité, il existe une abondante littérature sur l'utilisation de cette loi dans la modélisation et la classification de données. Deux raisons pour-

raient aller à l'encontre de ce choix. La première serait la polymodalité des distributions de descripteurs. Dans notre cas, ces distributions sont unimodales. La deuxième raison est la non-gaussiannité des données. Nous avons donc essayé de fournir des descripteurs dont la distribution soit la plus proche possible d'une distribution gaussienne.

6.5.3 Utilisation de la taxinomie des modes de jeu

Cette formulation du problème oblige à apporter au système des connaissances relatives aux modes de jeu. En effet, rien n'empêche l'agrégation de n'importe quels modes de jeu. Il faut donc spécifier quels modes de jeu sont possibles avec quels instruments. C'est ici que nous utilisons la taxinomie de modes de jeu définie au chapitre 3. Chaque variable de la taxinomie devient une variable d'état p^k avec, en plus, l'instrument. Chaque variable définit un problème d'apprentissage que nous appellerons *sous-problème* par la suite. Un sous-problème, aussi appelé expert dans la terminologie de la prise de décision collective, consiste en la distribution des descripteurs selon les valeurs de la variable correspondante. Ces sous-problèmes peuvent être testés par des tâches de classification automatique. Ainsi la variable *tremolo* qui peut valoir *yes* ou *no* définit un problème de classification consistant à identifier automatiquement les sons trémolo.

6.5.4 Apprentissage des modèles des sous-problèmes

L'apprentissage consiste maintenant à estimer les moyennes et les matrices de covariance des distributions $f_{j,k}$ de l'équation 6.8 sur les descripteurs des échantillons des bases décrites au chapitre 3. Nous appelons ces modèles *modèles initiaux*. Les modèles utilisés pour l'orchestration sont appelés *modèles finaux* et seront dérivés des modèles initiaux. Nous utilisons ces derniers pour tester les descripteurs et les sous-problèmes au chapitre 7.

6.5.5 Calcul des modèles finaux : résumé

Dans les modèles précédents, pour chaque mode de jeu et pour chaque descripteur, il n'y a qu'une seule distribution pour toutes les hauteurs et toutes les nuances. Ainsi ces modèles sont très utiles pour tester rapidement la qualité des descripteurs et la pertinence de la taxinomie des modes de jeu puisque avec une seule distribution nous pouvons évaluer l'ensemble des sons pour un descripteur et une classe. Mais, dans le contexte de l'orchestration, nous devons disposer d'une distribution par nuance et par hauteur pour pouvoir "additionner" facilement les modèles. Le passage des modèles initiaux aux modèles finaux se fait en plusieurs étapes dictées par le réseau de dépendance de la figure 6.5 et en utilisant quelques approximations.

Tout d'abord, il faut choisir l'ensemble des hauteurs désirées et l'ensemble des nuances. Pour les instruments, l'ensemble des hauteurs est donné par leur tessiture théorique propre. Pour les modes de jeu, toutes les hauteurs possibles de tous les instruments doivent être générées. Une fois choisis ces deux ensembles, nous pouvons générer le modèle de chaque hauteur et chaque nuance.

Soit h et n la hauteur et la nuance dont on souhaite obtenir le modèle.

Initialement, pour un état p^k nous disposons de la distribution suivante pour le descripteur $\mathbf{d}_j$:

$$f(\mathbf{d}_j, f_0, \log(e)|p^k) \quad (6.10)$$

et de la distribution du niveau de sonie :

$$f(N_l, f_0, \log(e)|p^k) \quad (6.11)$$

Cette dernière équation nous permet de déterminer la distribution de l'énergie e en fonction du niveau de sonie N_l et de la f_0 en utilisant les mêmes équations que 6.6 à 6.7b. La sonie N_l dépend directement de la nuance et la f_0 dépend directement de la hauteur.

$$f(e|N_l, f_0, p^k) = f(\log(e)|n, h, p^k) = \mathcal{N}(\log(e); \mu_e, \sigma_e) \quad (6.12)$$

Ensuite nous calculons la distribution conditionnelle en utilisant une nouvelle fois les équations 6.6 à 6.7b et la moyenne de la distribution 6.12 comme énergie.

$$f(\mathbf{d}_j|h, n, p^k) = f(\mathbf{d}_j|f_0, \log(e) = \mu_e, p^k) \quad (6.13)$$

Cas des descripteurs multidimensionnels Si $\mathbf{d}_j$ est multidimensionnel, nous ne conservons que la diagonale de la matrice de covariance. Nous utilisons cette dernière approximation pour d'une part pouvoir utiliser la méthode de Fenton-Wilkinson pour l'addition de loi log-normale et d'autre part pour diminuer l'espace mémoire occupé par les modèles. Nous verrons l'impact de cette simplification au chapitre 7.

6.6 Agrégation des distributions finales

6.6.1 Introduction

Le problème de l'agrégation de distributions est lié à la prise de décision collective, c'est-à-dire comment, à partir des opinions de plusieurs experts sur un problème donné (où des préférences de plusieurs individus concernant un choix), obtenir une décision consensuelle. Ce problème a été largement étudié. Notre cas correspond à l'agrégation d'opinions d'experts exprimées sous forme de densités de probabilité [GZ86]. Chaque sous-problème étant un expert donnant un avis sur la distribution des descripteurs ou sur la vraisemblance de la cible sachant l'état. Le problème principal est de savoir comment agréger les différents avis donnés par les experts. [GZ86] propose un état de l'art des différentes méthodes, nous décrivons les deux principales au paragraphe 6.6.2. Un exemple intéressant est le produit d'experts [Hin99], C'est une méthode très efficace pour modéliser des données de grande dimension qui satisfont simultanément plusieurs contraintes de dimension faible. Chaque expert se spécialise pour donner une probabilité élevée aux données qui satisfont seulement une contrainte. Les données qui satisfont cette contrainte mais violent les autres seront éliminées par les autres experts qui donneront une probabilité faible. La figure 6.6 montre un exemple de mixture d'experts appris automatiquement. On voit que certaines gaussiennes se sont spécialisées sur une dimension. [Hin99]

se focalise sur l'apprentissage conjoint des experts. Dans notre cas les contraintes sont explicites et les experts peuvent donc être entraînés séparément. Dans la même optique [MS06] étudie les produits d'experts modélisés par des mélanges de gaussiennes appliqués à la reconnaissance de parole. Ces méthodes peuvent être étendues en utilisant des Modèles de Markov Cachés (HMM) pour donner des HMM à plusieurs flux pour la reconnaissance de parole audiovisuelle [NPL⁺00], c'est à dire utilisant simultanément l'information audio et l'image pour la reconnaissance. Le problème de l'agrégation est aussi central dans la combinaison de classifieurs [JDM00] où on utilise plusieurs classifieurs différents, c'est-à-dire plusieurs méthodes différentes ou des classifieurs entraînés sur des données différentes, pour une même tâche de classification. Chaque classifieur donne une décision et un *combineur* est chargé de prendre la décision finale.

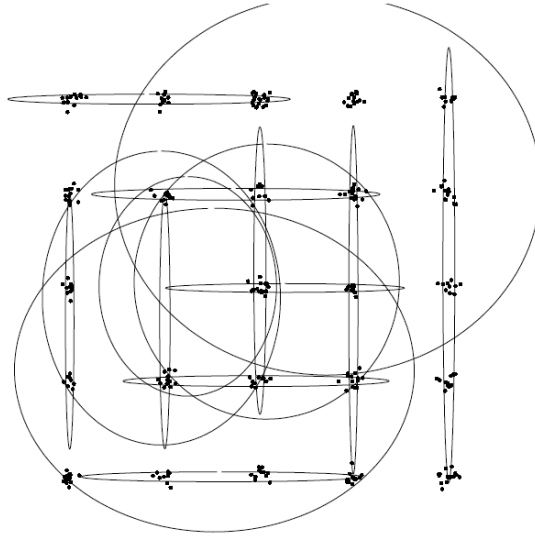


FIG. 6.6 – Exemple de produit d'expert gaussien [Hin99]

6.6.2 Les méthodes d'agrégation

Nous présentons ici les deux méthodes d'agrégation les plus couramment utilisées. Il s'agit du *linear opinion pool* [Sto61] et du *logarithmic opinion pool* [GWZ84]. Dans le premier, la fonction d'agrégation est la somme pondérée des densités. Soit $f_1 \dots f_K$ un ensemble de distribution à agréger. Alors la fonction d'agrégation est définie par :

$$T(f_1, \dots, f_K) = \sum_{k=1}^K w_k f_k \quad \text{avec} \quad \sum_{k=1}^K w_k = 1 \quad (6.14)$$

Elle est typiquement multimodale et dispersée et peut amener à une dictature, c'est-à-dire à une décision imposée par un seul expert. En effet si un expert donne une probabilité très élevée et que tout les autres donne une probabilité nulle, la probabilité résultante sera tout de même très élevée. Ce

comportement n'est pas adapté à notre problème puisque nous avons besoin du consensus entre les différents experts. Le *logarithmic opinion pool* est plus à même de produire un consensus. Il est égal au produit pondéré des distributions élémentaires :

$$T(f_1, \dots, f_K) = \frac{1}{Z} \prod_{k=1}^K f_k^{w_k} \quad (6.15)$$

Z est une constante de normalisation utilisée pour que la distribution résultante somme à un. Cette méthode donne généralement des distributions unimodale et moins dispersée que le linear opinion pool. Elle permet le veto d'un expert : si tous les experts donnent une probabilité élevée sauf un qui donne une probabilité nulle, la probabilité résultante sera nulle. Ce comportement est bien adapté à notre problème. Notons que cette fonction d'agrégation correspond à la moyenne pondérée des log vraisemblance, d'où son nom. Si les poids sont égaux à 1 cela donne un produit d'expert [Hin99]. Ces deux méthodes ont beaucoup été comparées [TDvB97], [AK99]. Le choix d'une méthode d'agrégation reste cependant l'un des problèmes les plus difficiles rencontrés lors de la conception d'un système à plusieurs classifieurs. En effet, les travaux qui traitent ce problème testent généralement un ensemble de méthodes de combinaison pour retenir la plus adéquate en fonction de certains critères de performance. Cependant, les résultats obtenus restent étroitement dépendants des applications traitées et par conséquent, difficiles à généraliser en dehors d'un contexte applicatif donné.

6.6.3 Cas gaussien pour le logarithmic opinion pool

[MS06] a étudié en détail le cas où les experts sont des gaussiennes ou des mélanges de gaussiennes. En particulier il fournit une forme analytique de la distribution résultant du produit de gaussienne. Dans le cas pondéré, la distribution est encore une loi normale. Soient $f_k = \mathcal{N}(\mathbf{d}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. La distribution issue de l'agrégation des f_k avec les poids w_k est :

$$f(\mathbf{d}_j | f_1 \dots f_K) = \mathcal{N}(\mathbf{d}_j; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \quad (6.16)$$

avec

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\Sigma} \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \right) \quad (6.17)$$

$$\hat{\boldsymbol{\Sigma}} = \left(\sum_{k=1}^K w_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1} \quad (6.18)$$

$$Z = \frac{(2\pi)^{d/2} |\boldsymbol{\Sigma}|}{\prod_s (2\pi)^{\frac{dw_s}{2}} |\boldsymbol{\Sigma}_s|^{\frac{w_s}{2}}} \exp \left[\frac{1}{2} \left(\boldsymbol{\mu}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \sum_s w_s \boldsymbol{\mu}_s' \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\mu}_s \right) \right] \quad (6.19)$$

Pour optimiser la rapidité des calculs, il serait intéressant de pouvoir supposer que Z est constante égale à 1. Nous comparons donc les résultats de classification obtenus en utilisant cette constante ou en la supposant égale à 1.

6.6.4 Choix des poids

Il n'existe pas de méthode unique pour choisir les poids dans les fonctions d'agrégation. Il est généralement admis qu'ils sont supposés refléter la qualité de chaque expert concernant le problème. [Win68] souligne que la qualité d'un expert doit être vue comme un jugement subjectif émis par une tierce personne, et que cette personne doit prendre en compte toutes les informations disponibles. Il propose alors plusieurs méthodes pour déterminer les poids dont les trois suivantes :

- Poids égaux, en général $w_k = \frac{1}{K}$ pour K experts pour que les poids somment à 1,
- Poids proportionnels à un classement. En supposant que les experts peuvent être classés, le poids de chaque expert est une fonction de son classement r_k , par exemple $w_k = \frac{r_k}{\sum_k r_k}$,
- Poids proportionnels à une auto-évaluation.

Nous comparons deux méthodes de calcul des poids. La première consiste à utiliser des poids égaux. La seconde permet de calculer les poids en fonction des résultats de classification de chaque expert pour son sous-problème. Par exemple pour le sous-problème *tremolo*, l'amplitude de modulation donne un bon résultat de classification, il sera donc affecté un poids élevé à cet expert pour ce descripteur. Le poids est une fonction linéaire du rappel moyen R dans la tâche de classification (voir paragraphe 7.3) :

$$w = \begin{cases} \frac{1}{R_{max}-R_h} R - \frac{R_h}{R_{max}-R_h} & \text{si } R > R_{max} \\ 0 & \text{sinon} \end{cases} \quad (6.20)$$

Où R_{max} représente le rappel obtenu par le meilleur descripteur pour le sous-problème considéré et R_h représente le rappel moyen d'une classification au hasard. Ainsi, cette équation associe le poids 0 à un expert qui ne fait pas mieux que le hasard et 1 au meilleur expert. Ensuite les poids sont normalisés à 1.

6.7 Évaluation d'une cible avec les modèles

La probabilité que le descripteur $\mathbf{d}_j$ de la combinaison c soit égale à celui de la cible t est évalué de la manière suivante :

$$f(\mathbf{d}_j = \mathbf{d}_j^t | c = \{x_1, \dots, x_M\})$$

$$\text{avec } \begin{cases} f(\mathbf{d}_j | c) &= G_j(f(\mathbf{d}_j | x_1), \dots, f(\mathbf{d}_j | x_M)) \\ f(\mathbf{d}_j | x_m) &= T(f(\mathbf{d}_j | p_m^1, n_m, h_m), \dots, f(\mathbf{d}_j | p_m^K, n_m, h_m)) \end{cases} \quad (6.21)$$

où c est une combinaison composée des sons $x_1 \dots x_M$, $\mathbf{d}_j^t$ est le descripteur de la cible, G_j est la fonction d'addition du descripteur j et T est la fonction d'agrégation. Les p^k sont les différents sous-problèmes.

Pour les descripteurs multidimensionnels, c'est-à-dire les descripteurs du spectre, il faut prendre en compte le fait que nous ne connaissons pas l'énergie de la cible. En effet, l'échantillon fourni pour spécifier la cible ne dit rien de la sonie réelle du son et donc sur son énergie. Pour palier à cette incertitude, nous cherchons l'énergie de la cible qui maximise la probabilité ci-dessus.

Si $f(\mathbf{d}_j|c)$ est une loi normale de moyenne $\boldsymbol{\mu}$ et de covariance diagonale $\boldsymbol{\Sigma}$, alors :

- Si le descripteur est en log :

$$\hat{e}^t = \arg \max_{e^t} f(\mathbf{d}_j = \log(e^t) + \mathbf{d}_j^t | c) \quad (6.22)$$

$$e^t = \frac{1}{\text{Tr}(\boldsymbol{\Sigma}^{-1})} ((\mathbf{d}_j^t + \boldsymbol{\mu})' * \boldsymbol{\sigma}^{-1}) \quad (6.23)$$

avec $\boldsymbol{\sigma}$ un vecteur contenant les éléments diagonaux de $\boldsymbol{\Sigma}$ et $\boldsymbol{\sigma}^{-1}$ un vecteur contenant les éléments diagonaux de $\boldsymbol{\Sigma}^{-1}$.

- Si le descripteur n'est pas en log :

$$\hat{e}^t = \arg \max_{e^t} f(\mathbf{d}_j = e^t \mathbf{d}_j^t | c) \quad (6.24)$$

$$e^t = \frac{\mathbf{d}_j^t \boldsymbol{\Sigma} \boldsymbol{\mu}'}{\mathbf{d}_j^t \boldsymbol{\Sigma} \mathbf{d}_j^{t'}} \quad (6.25)$$

Chapitre 7

Évaluation sur des échantillons solo et choix des paramètres

7.1 Introduction

Dans ce chapitre, nous présentons plusieurs évaluations des modèles d'instruments décrits au chapitre 6. Tout d'abord, nous donnons quelques exemples de modèles appris pour vérifier que les distributions des descripteurs correspondent à la réalité acoustique. Ensuite nous évaluons les modèles initiaux, les modèles finaux et les méthodes d'agrégation dans des tâches de classification. Préalablement, nous présentons, dans la partie 7.3, les critères d'évaluation utilisés. Dans la partie 7.4, les modèles initiaux sont évalués de deux manières différentes. D'abord descripteur par descripteur, puis en sélectionnant un sous-ensemble de descripteurs efficaces pour chaque sous-problème, nous donnons des résultats de classification globaux. La première approche permet de vérifier les capacités de discrimination de chaque descripteur pour chaque sous-problème et donc d'identifier les descripteurs efficaces pour chaque sous-problème. La seconde approche utilise une méthode de sélection de descripteurs basée sur les résultats de la première approche pour tester les capacités globale de discrimination des modèles. Dans la partie 7.5, nous testons les modèles finaux et les différentes approximations et transformations décrites au paragraphe 6.5.5. Finalement, nous évaluons les différentes méthodes d'agrégation de modèles dans la partie 7.6. L'ensemble de ces tests nous permet de choisir les méthodes et les paramètres utilisés dans le système d'orchestration.

7.2 Exemple de modèle appris

Nous décrivons trois exemples de modèles appris. La figure 7.1 montre la moyenne des distributions de la log puissance des harmoniques de la clarinette en Sib pour la hauteur $1a_4$ (440 Hz) et différentes nuances (à gauche) et pour la dynamique mf à différentes hauteurs (à droite). On observe bien la différence d'amplitude entre les harmoniques pairs et impairs typique de la clarinette. De plus on vérifie que cette différence décroît lorsque la dynamique diminue et lorsque la hauteur augmente. Enfin la figure de droite montre que la décroissance du spectre augmente quand la hauteur augmente. Sur la figure 7.2 qui représente la log puissance des partiels pour un trombone jouant un $1a_3$ à diffé-

rentes nuances. Nous observons aussi l'augmentation de la pente du spectre lorsque la nuance passe de *ff* à *pp*.

La figure 7.3 montre la vitesse et l'ambitus de la modulation d'énergie pour les sons tremolo et non tremolo. Concernant les sons tremolo, l'ambitus est cohérent avec les observations de [Moe04]. Ce dernier a constaté une augmentation de la fréquence avec l'augmentation de la hauteur, ce que nous retrouvons dans le modèle. Concernant le modèle des non tremolo, nous observons une déviation standard de la vitesse élevée qui n'est pas significative et montre seulement que la modulation est inexistante ou très faible et ainsi que sa vitesse n'a pas de sens. La forte augmentation de la vitesse de modulation pour les notes graves est provoquée par le fait que la modulation observée est en fait due à la fréquence fondamentale et non à un tremolo.

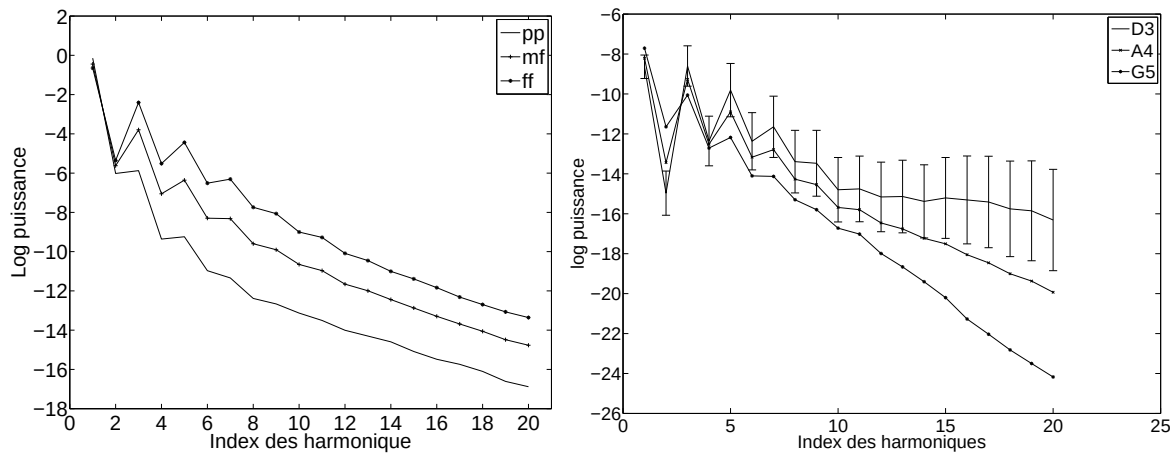


FIG. 7.1 – Moyenne et écart type des distributions de l'amplitude des harmoniques pour le modèle de la clarinette en Sib pour une note (la4) à différentes nuances (à gauche), et pour une nuance (*mf*) à différentes hauteurs

Ces trois exemples montrent bien les possibilités de modélisation de la méthode proposée et comment avec très peu de paramètres, il est possible de décrire la distribution de descripteurs unidimensionnel ou multidimensionnel, ainsi que l'influence de la hauteur et de la nuance sur ces distributions. Dans le paragraphe suivant, nous testons la méthode de modélisation proposée par des tâches de classification de sons.

7.3 Critères d'évaluations des tâches de classification

Nous utilisons trois critères pour évaluer les résultats de classification : le rappel, la précision et la F-mesure [JHD99]. Ces critères sont basés sur la matrice de confusion. Chacun de ces critères est calculé pour chaque classe puis moyenné pour obtenir une mesure global. Le rappel correspond à la proportion d'éléments d'une classe qui ont été correctement classés, la précision correspond à la proportion d'éléments affectés à une classe qui appartiennent effectivement à cette classe, enfin la F-mesure correspond à la moyenne harmonique de la précision et du rappel.

Soit un problème de classification à I classes $\{p_1 \dots p_I\}$. Chaque classe contient N_i éléments. Soit c_i

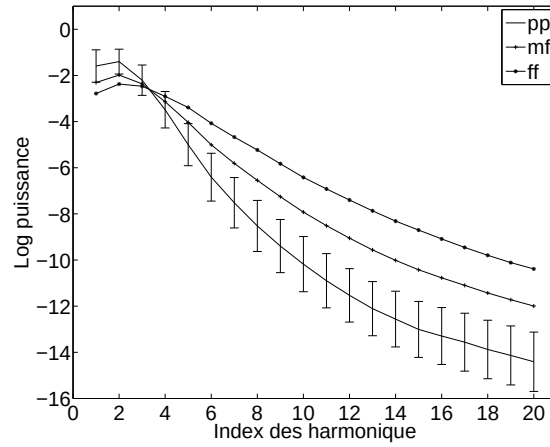


FIG. 7.2 – Moyenne et écart type des distributions de l'amplitude des harmoniques pour le modèle de trombone pour une note (la3) à différentes nuances (à gauche), et pour une nuance (*mf*) à différentes hauteurs

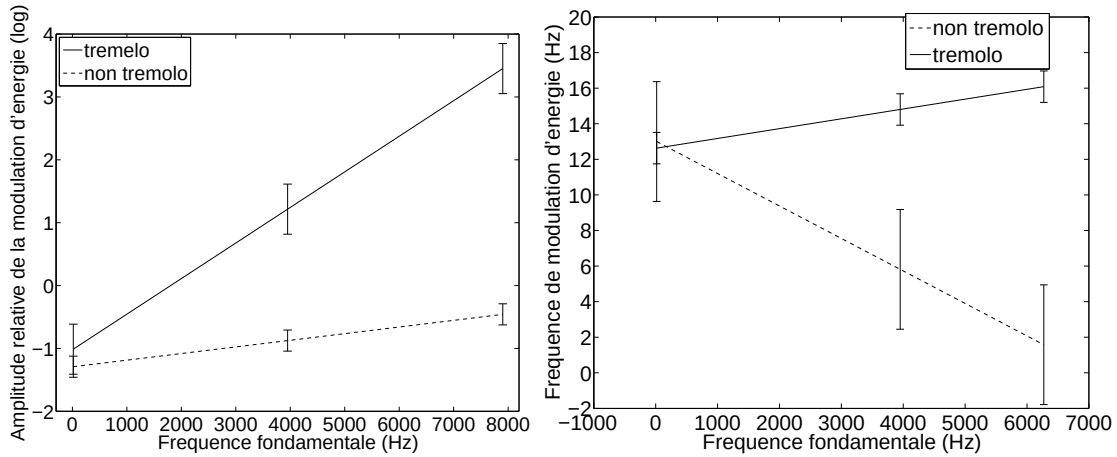


FIG. 7.3 – Moyenne et écart type des distributions de l'ambitus (gauche) et de la vitesse (droite) de la modulation d'énergie pour le modèle de tremolo et non tremolo

le nombre d'éléments de la classe i correctement classés. Soit M_i le nombre total d'éléments assignés (correctement ou non) à la classe i . Le rappel, la précision et la F-mesure valent respectivement :

$$R_i = \frac{c_i}{N_i} \quad (7.1)$$

$$P_i = \frac{c_i}{M_i} \quad (7.2)$$

$$F_i = \frac{2P_i R_i}{P_i + R_i} \quad (7.3)$$

Nous utiliserons généralement les valeurs moyennes de ces mesures sur toutes les classes. Pour ce faire une idée du comportement de ces mesures, nous étudions leurs valeurs pour deux classifieurs particuliers.

Le premier est le classifieur aléatoire. Un classifieur aléatoire est un classifieur qui affecte aléatoirement une classe à un élément. Dans ce cas :

$$R_i = \frac{1}{I} \quad (7.4)$$

$$P_i = \frac{N_i}{\sum_{k=1}^I N_k} \quad (7.5)$$

$$F_i = \frac{2N_i}{2N_i + \sum_{k=1}^I N_k} \quad (7.6)$$

Nous constatons ici que le rappel d'une classe avec un classifieur aléatoire est toujours égal à 1 sur le nombre de classes, alors que ce n'est pas le cas pour la précision et la F-mesure qui dépendent du nombre d'éléments de chaque classe. Concernant les valeurs moyennes, nous obtenons :

$$\bar{R} = \frac{1}{I} \quad (7.7)$$

$$\bar{P} = \frac{1}{I} \quad (7.8)$$

En moyenne, le rappel et la précision sont tous deux indépendants du nombre d'éléments de chaque classe. La F-mesure moyenne reste dépendante de ce paramètre.

Le second classifieur est celui qui affecte toujours la même classe quel que soit l'élément. Ce cas est intéressant, car certains des problèmes que nous étudions ont des classes de cardinalité très différentes. Il faut donc être sûr qu'un classifieur qui affecterait la classe qui contient le plus d'élément à toutes les données ne se verrait pas surévalué. Supposons que la classe 1 récupère toutes les données. Pour ce classifieur, cette classe obtient un rappel de 1, mais toutes les autres un rappel de 0 :

$$R_i = \begin{cases} 1 & \text{si } i = 1 \\ 0 & \text{sinon} \end{cases} \quad (7.9)$$

La moyenne du rappel reste donc de 1 sur le nombre de classes :

$$\bar{P} = \frac{1}{I} \quad (7.10)$$

La précision de chaque classe vaut

$$P_i = \begin{cases} \frac{1}{1 + \frac{\sum_{k \neq i} N_k}{N_i}} & \text{si } i = 1 \\ 0 & \text{sinon} \end{cases} \quad (7.11)$$

Et la précision moyenne est égale à

$$\bar{P} = \frac{1}{I \left(1 + \frac{\sum_{k \neq 1} N_k}{N_1} \right)} \quad (7.12)$$

La F-mesure de chaque classe et la F-mesure moyenne valent respectivement :

$$F_i = \begin{cases} \frac{2}{1 + \frac{\sum_{k=1}^I N_k}{N_i}} & \text{si } i = 1 \\ 0 & \text{sinon} \end{cases} \quad (7.13)$$

$$\bar{F} = \frac{2}{I \left(1 + \frac{\sum_{k=1}^I N_k}{N_1} \right)} \quad (7.14)$$

Ainsi la précision et la F-mesure moyennes dépendent de la répartition des classes. Lorsque le nombre d'élément de la classe qui "récupère" tous les éléments devient très grand devant le nombre total d'élément, ces deux mesures tendent vers 1 sur le nombre de classes. Un classifieur de ce type n'aura jamais une mesure de performance qui dépasse 1 sur le nombre de classes.

La F-mesure moyenne donne un bon résumé des performances mais dépend de la répartition des classes pour les deux classifieurs étudiés. Ce dernier point rend cette mesure délicate à interpréter dans le cas où on ne dispose pas de référence, par exemple un autre algorithme utilisé sur les mêmes données. De plus la dépendance de la F-mesure d'un classifieur à la répartition des classes empêche l'utilisation intuitive de cette mesure dans la comparaison d'une méthode avec un classifieur aléatoire. Nous utiliserons donc la F-mesure pour comparer des algorithmes entre eux ou pour comparer différentes valeurs de paramètres. Dans ce cas, puisque les différentes méthodes sont utilisées sur une même base de test, la F-mesure donne un bon classement. En revanche, pour évaluer un algorithme lorsque nous n'avons pas de référence, nous utiliserons le rappel et la précision. En effet, en l'absence de référence, nous comparons les résultats à ceux d'un classifieur aléatoire, pour que les résultats soient faciles à interpréter l'utilisation du rappel et de la précision, qui valent toujours 1 sur le nombre de classes en moyenne pour un classifieur aléatoire, est plus adaptée.

7.4 Évaluation des modèles initiaux

7.4.1 Évaluation de chaque descripteur

7.4.1.1 Procédure

Les modèles obtenus pour chaque sous-problème sont testés par des tâches de classification. Nous utilisons cinq bases d'échantillons dont le contenu est décrit au chapitre 3. Pour chaque sous-problème, les modèles sont appris sur quatre bases et testés sur la cinquième. Cette procédure, parfois appelée *Leave One Database Out* permet d'assurer un maximum d'indépendance entre les données d'apprentissage et les données de test et ainsi d'éviter le surapprentissage. Les différentes bases ne contenant pas les mêmes instruments ou modes de jeu, il n'est pas toujours possible de tester tous les modèles appris ou d'apprendre des modèles pour tous les échantillons de la base de test. Nous présentons les résultats obtenus en utilisant Studio Online comme base de test car elle contient beaucoup de modes de jeu.

7.4.1.2 Critère de décision

Le critère de décision pour le sous-problème p^k et le descripteur $\mathbf{d}_j$ est le suivant :

$$\hat{p}_j^k = \arg \max_{p^k} f(\mathbf{d}_j, \log(e), f_0|p^k) \quad (7.15)$$

Où $\hat{p}_j^k$ est la classe obtenue dans le sous-problème k avec le descripteur j .

7.4.2 Classification avec un sous-ensemble de descripteurs

7.4.2.1 Procédure

La procédure est la même que précédemment, mais nous présentons les résultats pour toutes les bases de test.

7.4.2.2 Critère de décision

Il s'agit ici de se ramener à un critère unique pour choisir la classe des échantillons testés. Soit le sous-problème défini par la variable d'état $p^k = \{p_1^k \dots p_I^k\}$. Soit $\{\mathbf{d}_1 \dots \mathbf{d}_J\}$ les descripteurs de l'élément à classer.

$$\hat{p}^k = \arg \max_{p^k} f((\mathbf{d}_j), \log(e), f_0|p^k) \quad (7.16)$$

$$= \arg \max_{p^k} \prod_{j=1}^J f(\mathbf{d}_j, \log(e), f_0|p^k) \quad (7.17)$$

7.4.2.3 Sélection de descripteurs

Ce dernier critère suppose d'utiliser tous les descripteurs lors de la classification, or certains d'entre eux ne sont pas pertinents pour certains sous-problèmes. Nous devons donc sélectionner un sous-ensemble de descripteurs utiles pour la tâche de classification. Nous utilisons une méthode de type *wrapper* [GE03] réalisée par un algorithme glouton. Les descripteurs sont ajoutés un par un en commençant par celui qui donne le meilleur résultat de classification mesuré par la F-mesure. L'algorithme s'arrête quand plus aucun descripteur n'améliore les résultats de classification. Par la suite nous appelons D^k , l'ensemble des descripteurs sélectionnés pour le sous-problème k . Le critère de décision est finalement le suivant :

$$\hat{p}^k = \arg \max_{p^k} \prod_{j \in D^k} f(\mathbf{d}_j, \log(e), f_0|p^k) \quad (7.18)$$

7.4.3 Résultats

Les figures 7.4 et 7.5 montrent les résultats de classification par sous-problème et par descripteur lorsque la base de test est Studio Online (critère 7.15). La figure 7.6 montre les résultats globaux obtenus avec la sélection automatique de descripteurs pour chaque base de test (critère 7.16).

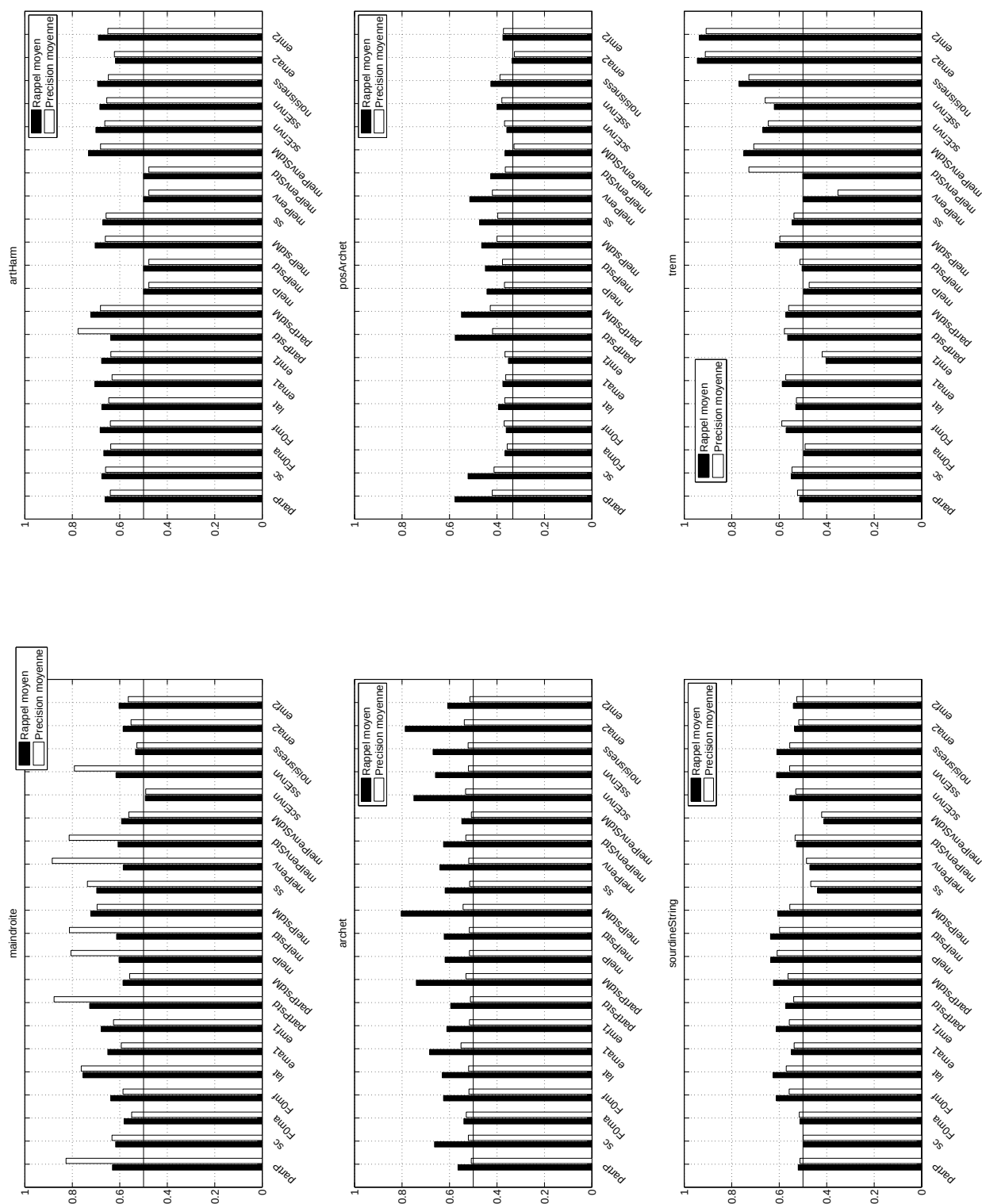


FIG. 7.4 – Résultats par descripteur pour la base de test Studio Online

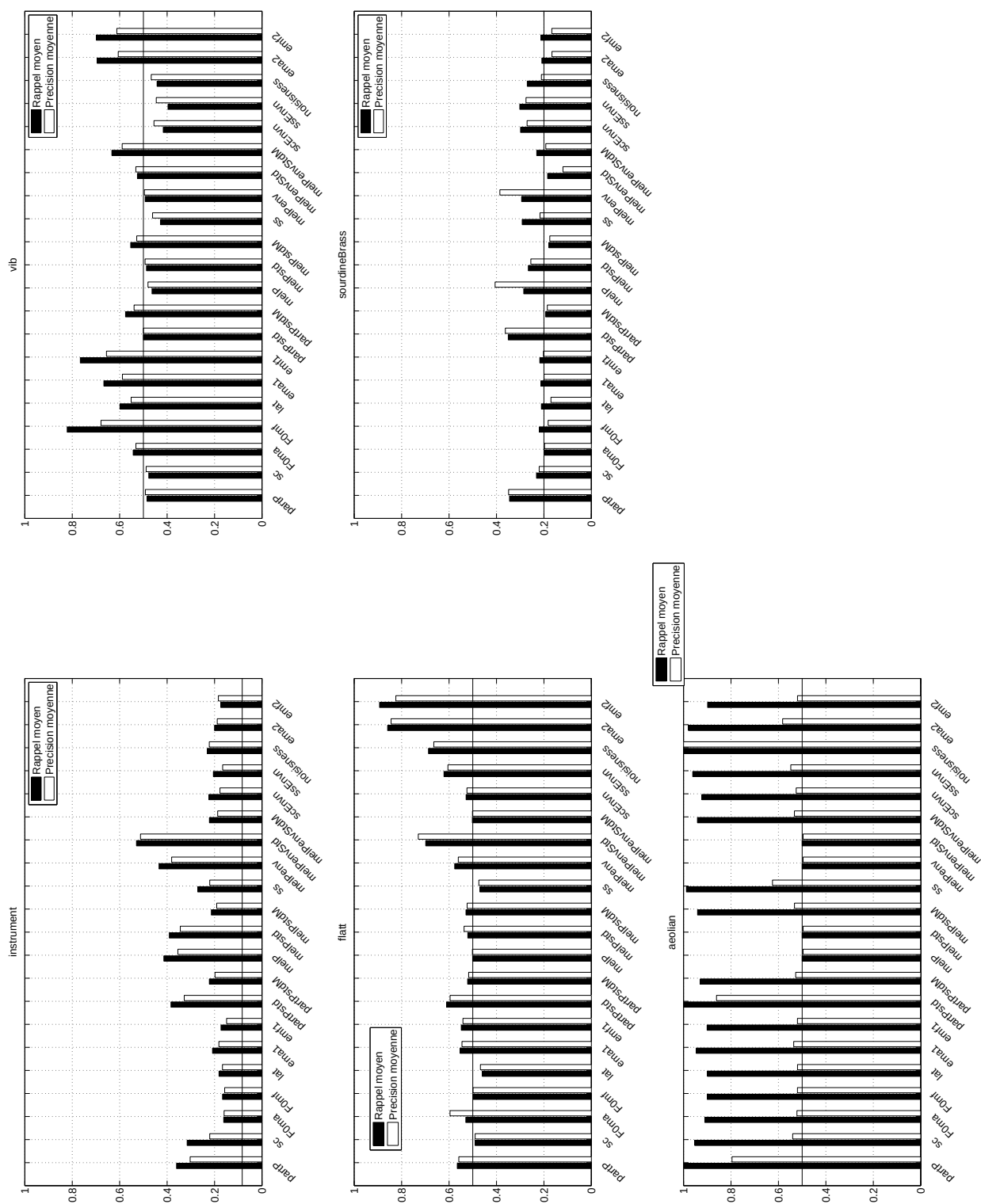


FIG. 7.5 – Résultats par descripteur pour la base de test Studio Online

	Virtual Orchestra	Vienna Symphonic Library	Studio Online	RWC	Iowa
instrument	0,37 0,39 (11)	0,38 0,43 (14)	0,58 0,56 (12)	0,68 0,64 (12)	0,73 0,75 (7)
maindroite			0,88 0,84 (2)		0,97 0,98 (2)
artHarm	0,72 0,78 (2)		0,71 0,73 (2)		
archet	0,57 0,59 (2)		0,83 0,61 (2)		
posArchet	0,42 0,54 (3)		0,54 0,47 (3)		
sourdineString	0,66 0,67 (2)		0,73 0,68 (2)		
trem		0,95 0,93 (2)	0,97 0,98 (2)		
vib		0,92 0,90 (2)	0,87 0,76 (2)		0,80 0,77 (2)
flatt	0,78 0,73 (2)	0,88 0,83 (2)	0,92 0,88 (2)		
sourdineBrass	0,33 0,32 (5)		0,40 0,36 (5)		
aeolian	0,98 0,97 (2)		1,00 1,00 (2)		

FIG. 7.6 – Résultats globaux pour toutes les bases (rappel moyen | précision moyenne (nombre de classes))

Classification des instruments Les résultats de classification des instruments sont difficiles à comparer avec l'état de l'art car ils dépendent fortement du nombre d'instruments et de la procédure de test adoptée. De plus pour chaque instrument, il y a beaucoup de modes de jeu différents, et donc beaucoup de sons différents, ce qui n'est pas le cas dans les études précédentes qui n'étudient que le mode de jeu "ordinaire". Concernant ce dernier point, la comparaison est plus pertinente lorsque les bases de test sont Iowa et RWC puisque elles ne contiennent que le mode de jeu "ordinaire". Pour ces deux tests, les résultats sont comparables à l'état de l'art avec respectivement 68% et 73% de rappel moyen pour 7 et 12 instruments. Sur Studio Online, tous les descripteurs donnent un résultat supérieur au hasard, indiquant qu'ils sont tous au moins un peu discriminants. Les descripteurs principaux sont les descripteurs de spectre comme la puissance des partiels, le centroïde spectral ou les descripteurs d'enveloppe. Il est intéressant de constater que les descripteurs qui mesure la déviation standard du spectre en fonction du temps donnent de bons résultats. La déviation standard de l'énergie par bande de Mel donne ainsi un rappel de 0,4.

Classification des modes de jeu Les résultats sont assez disparates selon le mode de jeu considéré. Le résultat dépend de quatre facteurs : la différence de timbre entre les classes considérées, la capacité des descripteurs à rendre compte de cette différence, la validité du modèle proposé et enfin, la pertinence des classes testées. Ainsi des classes dont les timbres sont bien différents et qui repose sur des descripteurs robustes donnent de très bons résultats. Les sous-problèmes *vibrato*, *tremolo* et *flutter-zunge* qui reposent principalement sur les modulations de fréquence et d'énergie, donnent des taux de rappel compris entre 0,78 et 0,98 et des taux de précision du même ordre. Le tremolo est le mieux reconnu quelle que soit la base de test (autour de 0,95 de rappel). Le vibrato est légèrement moins bien

reconnu. Les *flutterzunge* de Virtual orchestra obtiennent un rappel de 0,78 et un précision de 0,73 alors qu’avec les autres bases les résultats sont autour de 0,85. Cela indique peut-être une différence d’enregistrement entre Virtual Orchestra d’une part et Vienna Symphonic Library et Studio Online d’autre part. La reconnaissance du vibrato repose principalement sur la fréquence de modulation de la fréquence fondamentale et dans une moindre mesure sur l’ambitus de cette modulation et sur l’ambitus et la vitesse de la modulation d’énergie. Le tremolo et le *flutterzunge* sont principalement reconnus grâce à la modulation d’énergie et la noisiness qui est plus élevée pour les sons modulés. Le mode de jeu aeolien est parfaitement reconnu lorsque la base de test est Studio Online et presque parfaitement lorsque la base est Virtual Orchestra. Le principal descripteur est la noisiness. La distinction entre cordes jouées à l’archet et *pizzicato* est aussi très bonne en utilisant le temps d’attaque mais aussi les descripteurs du spectre.

Les autres sous-problèmes sont plus complexes et les résultats sont généralement moins bons mais parfois surprenants. Par exemple la distinction entre les sons de cordes avec et sans sourdine donne de bons résultats en se basant sur le spectre de Mel, la déviation standard de la puissance des harmoniques ou encore le spectre de Mel du bruit résiduel, montrant ainsi qu’il est possible de généraliser l’effet de la sourdine au différentes cordes frottées. Le sous-problème *position de l’archet* donne aussi des résultats intéressants quoique plus faibles. Les rappels et précisions sont de l’ordre de 0,5, mais lorsqu’on regarde le détail des résultats, on constate que les confusions les plus fréquentes se produisent entre *ordinario* et *tasto* qui à l’oreille sont souvent difficile à distinguer. Les sons *ponticello* sont en revanche bien discriminés, grâce à la déviation standard de l’énergie des partiels qui est généralement plus élevée pour les *ponticello* du fait de l’instabilité du son.

En général, les résultats de classification sont moins bon lorsque la base de test est Virtual Orchestra, ce qui indique une différence d’enregistrement avec toutes les autres bases utilisées. Cette différence se confirme à l’écoute des sons. Globalement les classes envisagées et la méthode proposée donnent donc des résultats satisfaisant, considérant que notre but n’est pas la classification. Ainsi, si deux modes de jeu se ressemblent beaucoup, il n’est pas vraiment problématique de les confondre dans une tâche de classification.

7.5 Évaluation des modèles finaux

Les modèles initiaux permettent de classer rapidement un grand nombre de sons, et sont donc utilisés pour vérifier la pertinence des sous problèmes et l’efficacité des descripteurs. Mais il faut vérifier que les différentes approximations menant aux modèles finaux sont pertinentes et ne dégradent pas trop la qualité des distributions. Pour cela nous présentons quelques tests de classification dans lesquels nous évaluons l’impact de chacune des transformations et simplifications décrites dans le paragraphe 6.5.5. Il y a donc deux étapes à comparer au modèle initial : L’utilisation de matrices de covariance diagonales, et l’introduction du niveau de sonie. Nous testons aussi la méthode d’estimation de l’énergie de la cible décrite au paragraphe 6.7.

7.5.1 Procédure

Nous utilisons la base Studio Online comme base de test et les modèles sont appris sur les quatre autres bases. Les tests sont réalisés sur le sous ensemble de notes {sol1, fa1, ré2, sol2, mi4, la4, fa5,

si5, do6, fa6, ré7, sol7}.

La règle de décision pour les 3 premiers tests est la suivante :

$$\hat{p}_j^k = \arg \max_{p^k} f(\mathbf{d}_j | p^k, \log(e) = \log(e^t), f_0 = f_0^t) \quad (7.19)$$

Où e^t et f_0^t représente l'énergie et la fréquence fondamentale du son à classer. La matrice de covariance est pleine dans le premier test, ce qui est équivalent au modèle initial, elle est diagonale dans le second. L'énergie e^t de la cible est estimée comme décrit au paragraphe 6.7 dans le troisième test. La dernière règle utilise les modèles dans lesquels l'énergie e est prédite à partir de la nuance :

$$\hat{p}_j^k = \arg \max_{p^k} f(\mathbf{d}_j | p^k, n, h) \quad (7.20)$$

7.5.2 Résultats

7.5.2.1 Instruments

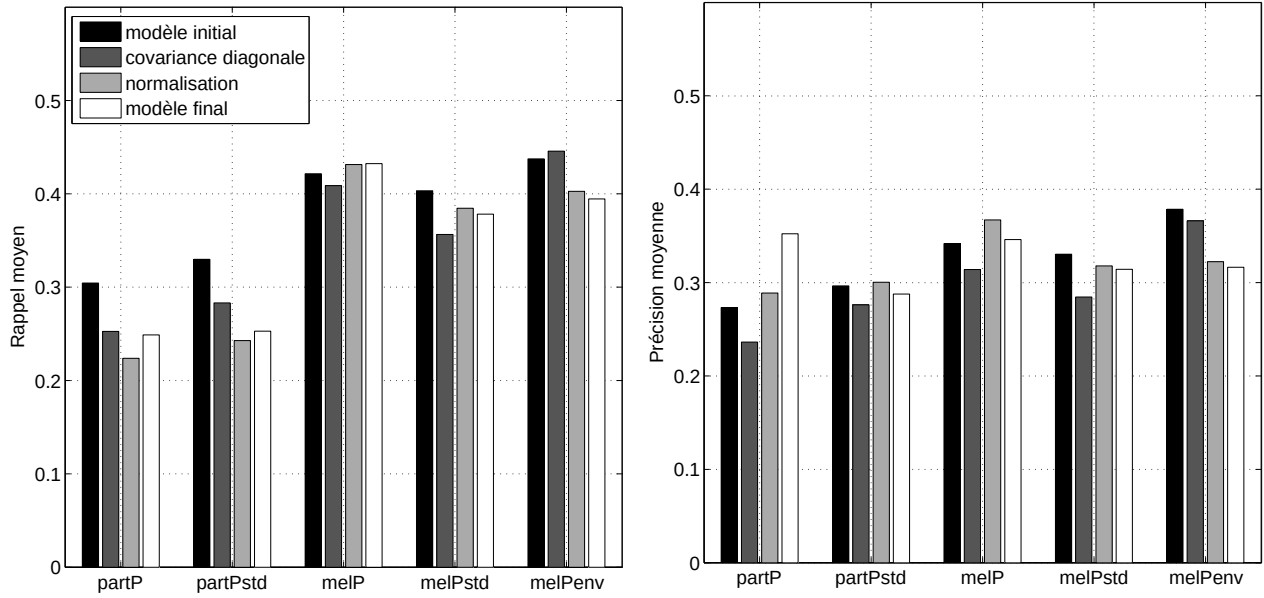
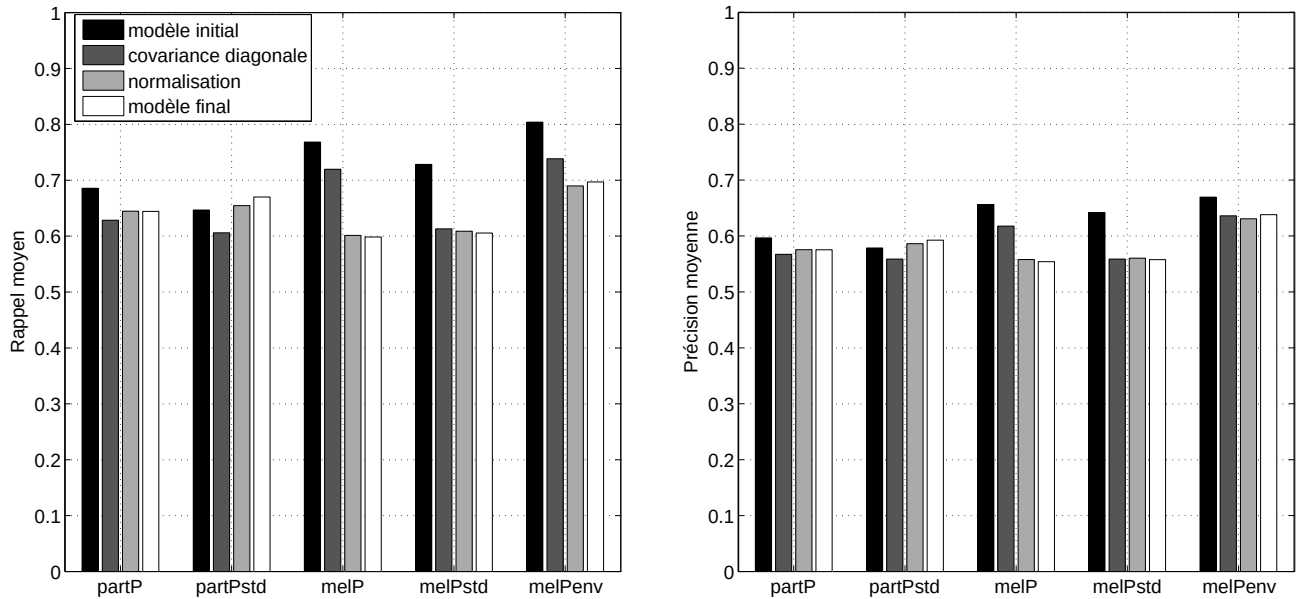
La figure 7.7 montre les résultats pour le sous-problème *instrument* pour les descripteurs multidimensionnels qui sont les seuls sur lesquels l'utilisation de matrices de covariance diagonales et l'estimation de l'énergie de la cible ont un effet. L'effet principal est constaté pour la matrice de covariance diagonale, surtout pour l'énergie des partiels avec une perte de 0,05 point de rappel. Cette approximation fait toujours perdre sur le rappel et sur la précision sauf pour l'enveloppe de Mel. La perte est généralement inférieure à 0,05 point. L'effet des autres approximations est assez variable, elles peuvent parfois faire augmenter les résultats. Ainsi l'estimation de l'énergie de la cible fait gagner en précision (de l'ordre de 0,02 point) pour l'énergie des partiels. Globalement, l'énergie des partiels et l'enveloppe de Mel sont plus sensibles aux modifications que les autres descripteurs. Pour les descripteurs unidimensionnels, seule la dernière approximation a un effet. L'effet constaté est généralement très faible avec une perte moyenne de F-mesure de 0,005, un écart type de 0,008, une perte maximum de 0,014 et un gain maximum de 0,019.

7.5.2.2 Sourdine des cordes

La figure 7.8 donne le rappel et la précision moyenne pour le sous-problème *sourdineString*. Comme précédemment, c'est l'utilisation d'une matrice de covariance diagonale qui a le plus d'effet. Les autres approximations ont un impact assez faible, mis à part l'estimation de l'énergie de la cible sur le spectre de Mel. A l'inverse des instruments, ce sont l'énergie des partiels et l'enveloppe de Mel qui subissent le moins les effets des approximations. Concernant les descripteurs unidimensionnels, les écarts sont encore une fois très faibles avec une moyenne de perte de F-mesure de 0,0032, un écart type de 0,006, une perte et un gain maximum respectivement de 0,003 et 0,013.

7.5.2.3 Tremolo

Pour le sous-problème *tremolo*, nous n'étudions que les descripteurs unidimensionnel qui sont ceux qui ont une influence. Les effets sont du même ordre que précédemment. La moyenne est de 0,0024, l'écart type de 0,008, la perte et le gain maximum sont de 0,015 et 0,016.

FIG. 7.7 – Effet des approximations pour le sous-problème *Instrument*FIG. 7.8 – Effet des approximations pour le sous-problème *SouridineString*

7.5.2.4 conclusion

Les effets étudiés ne sont pas négligeables, mais n'entraînent pas non plus de trop grandes pertes de performance. Le constat intéressant est que la prédiction de l'énergie à partir de la sonie est correcte et que la méthode d'estimation de l'énergie de la cible qui permet de ne pas prendre en compte la

nuance de la cible donne de bons résultats.

7.6 Évaluation des méthodes d'agrégation

7.6.1 Procédure et classifieur utilisés

Les tâches de classification de ce paragraphe sont destinées à comparer les différentes méthodes d'agrégation, les méthodes de calcul des poids et l'effet de la constante de normalisation dans le logarithmic opinion pool. Elles ont aussi pour but de montrer que l'apprentissage par sous-problème suivi par l'agrégation des distributions est une méthode d'apprentissage efficace. La base de test est Studio Online et les modèles sont appris sur les quatre autres bases. Deux tests sont présentés, le premier est fait sur les bois et le second sur les cordes.

Soit $P = \{p^1, \dots, p^K\}$ l'ensemble des sous-problèmes considérés. Chaque sous-problème p^k peut prendre I_k valeurs différentes dans l'ensemble $\{p_1^k, \dots, p_{I_k}^k\}$. Dans ce cas, nous obtenons un problème de classification à $\prod_{k=1}^K I_k$ classes différentes. Pour chacune de ces classes un modèle est obtenu par agrégation pour chaque descripteur :

$$f(\mathbf{d}_j | n, h, p^1, \dots, p^K) = T(f_{j,1}, \dots, f_{j,K}) \quad (7.21)$$

$$\text{avec } f_{j,k} = f(\mathbf{d}_j | p^k, n, h) \quad (7.22)$$

La règle de décision pour un descripteur est alors :

$$(\hat{p}^1, \dots, \hat{p}^K) = \arg \max_{p^1 \dots p^K} f(\mathbf{d}_j | p^1, \dots, p^K, n, h) \quad (7.23)$$

Les descripteurs utilisés sont :

- la puissance des harmoniques,
- le centroïde spectral,
- le logarithme du temps d'attaque,
- le spectre de Mel,
- l'enveloppe du spectre de Mel,
- le centroïde de l'enveloppe de la partie non-sinusoïdale,
- la noisiness
- l'ambitus de la modulation d'énergie.

7.6.2 Résultats

Nous ne présentons pas de résultats de classification globaux, mais des résultats pour chaque sous-problème. Un descripteur n'est généralement pas efficace pour tous les sous-problèmes, les résultats globaux ne sont donc pas très significatifs. Nous cherchons en revanche à vérifier que le modèle issu de l'agrégation conserve les performances des modèles agrégés. Pour vérifier cela il est plus intéressant d'étudier les résultats pour chaque sous-problème. Cela signifie que nous extrayons de la classe retournée par le classifieur, les classes correspondant à chaque sous-problème, c'est-à-dire chacun des $\hat{p}^k$ dans l'équation 7.23.

7.6.2.1 Un cas simple

Nous effectuons d'abord un test sur un cas simple en classant les bois selon l'instrument et le mode de jeu *flutterzunge*, ce qui donne un problème à quatre classes : flûte normale, flûte *flutterzunge*, clarinette normal, clarinette *flutterzunge*. Dans ce test, les différences entre les méthodes d'agrégation sont très faibles sinon inexistantes. En effet, les deux sous-problèmes concernent des descripteurs différents, les descripteurs spectraux pour *instrument* et les descripteurs de modulation pour *flutterzunge*, les distributions agrégées sont donc presque totalement déterminées par la distribution d'un des deux sous-problème, en l'occurrence celle qui a la plus petite variance. Les différences entre les méthodes sont de l'ordre de 4% c'est-à-dire 0.03 de rappel moyen. Nous présentons uniquement les résultats pour la méthode du logarithmic opinion pool avec $Z=1$ (figure 7.9). Ces résultats montrent que pour des problèmes simples la technique consistant à agréger des distributions fonctionne. Les résultats trouvés pour chaque sous-problème sont proches des résultats obtenus lorsque les sous-problèmes sont testés séparément, avec par exemple un rappel moyen supérieur à 0.9 pour le *flutterzunge* avec l'amplitude de la modulation d'énergie ou encore 0.8 pour l'*instrument* avec l'énergie par bandes de Mel.

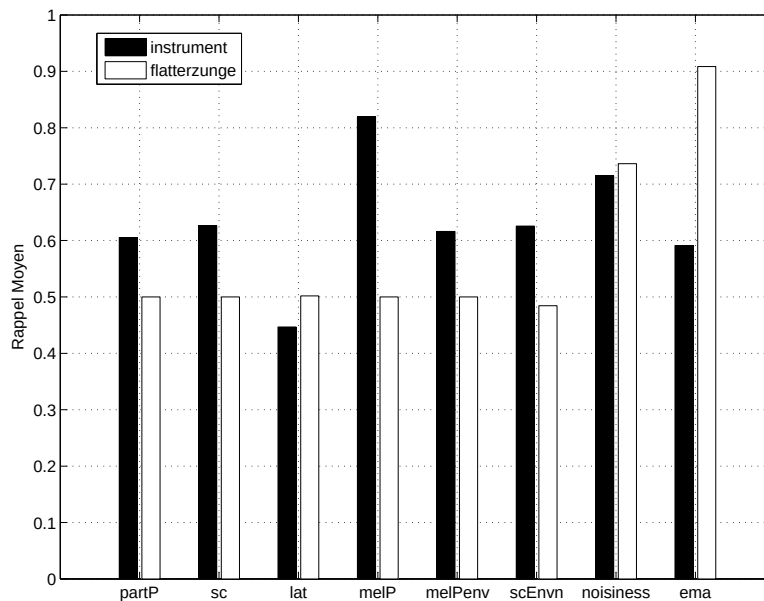


FIG. 7.9 – Résultats de classification par sous-problème pour les vents

7.6.2.2 Un cas complexe

Nous étudions maintenant un cas complexe. L'ensemble des sous-problèmes considérés et des valeurs possibles est présenté dans le tableau 7.1.

Comme précédemment nous présentons les résultats pour chaque sous-problème. La figure 7.10 montre les résultats moyen obtenus pour les trois méthodes d'agrégation : le logarithmic opinion pool (appelé produit), le logarithmic opinion pool avec $Z=1$ et le linear opinion pool (appelé somme). En

sous-problème	Valeurs
Instrument	Violon, Alto, Violoncelle, Contrebasse
Main droite	Archet, Pizzicato
Harmonique artificielle	Oui, Non
Archet	Ordinario, Legno-tratto, NA
Sourdine	Non, Sourdine normale
Tremolo	Non, Oui, NA

TAB. 7.1 – ensemble des sous-problèmes considérés et des valeurs possibles

premier lieu, il est important de noter que les sous-problèmes *archet* et *trem* ne sont plus, comme le montre le tableau 7.1, des problèmes à 2 mais 3 classes, puisque le sous-problème peut ne pas être valable dans certain cas. Par exemple le tremolo n'est pas possible en pizzicato. Le résultat d'un classifieur aléatoire passe donc de $1/2$ à $1/3$. Le premier constat est que le linear opinion pool fournit de mauvais résultats à quelques rares exceptions près, ce qui correspond aux résultats attendus. Ensuite une observation plus intéressante est que l'approximation du facteur de normalisation Z par 1 ne dégrade pas les résultats et les améliore même parfois. Globalement les résultats du logarithmic opinion pool avec et sans approximation sont à peu près équivalents. Enfin, nous voyons que même dans un cas complexe avec beaucoup de sous-problèmes les résultats de classification restent bon.

La figure 7.11 présente les résultats en fonction de la méthode de calcul des poids utilisée. Les différences sont assez faibles. Pour le sous-problème *instrument*, les résultats sont identiques. Pour les autres sous-problèmes des différences existent mais ne sont pas systématiquement en faveur de l'une ou l'autre des méthodes. Pour le sous-problème *archet*, la méthode des poids égaux est ainsi meilleure pour le spectre de Mel et pour la noisiness mais moins bonne pour les partiels et pour le centroïde spectral.

On constate, comme pour la tâche précédente, que l'agrégation des modèles conserve les résultats des modèles utilisés seul pour la plupart des descripteurs. Ainsi pour le sous-problème *maindroite*, tous les descripteurs conservent leur pouvoir de discrimination. Les résultats obtenus avec la noisiness sont même supérieurs lorsque les modèles sont agrégés. Pour le sous-problème *artHarm*, trois descripteurs subissent une dégradation (la noisiness, le centroïde de l'enveloppe de la partie non-sinusoidale et l'ambitus de la modulation d'énergie), mais les autres restent bien discriminants. Pour les sous-problèmes *trem* et *archet* l'analyse est moins évidente puisque ce sont maintenant des problèmes à trois classes. Nous pouvons constater que l'allure général des résultats reste la même. Pour le tremolo, l'ambitus de la modulation d'énergie amène toujours à un résultats à peu près deux fois supérieur au hasard, la noisiness aussi.

7.7 Conclusion

Dans ce chapitre nous avons présenté plusieurs évaluations des modèles d'instruments par des tâches de classification. En utilisant les modèles initiaux, nous obtenons de très bon résultats de classification pour les instruments et les modes de jeu. Nous montrons ainsi que chaque mode de jeu a

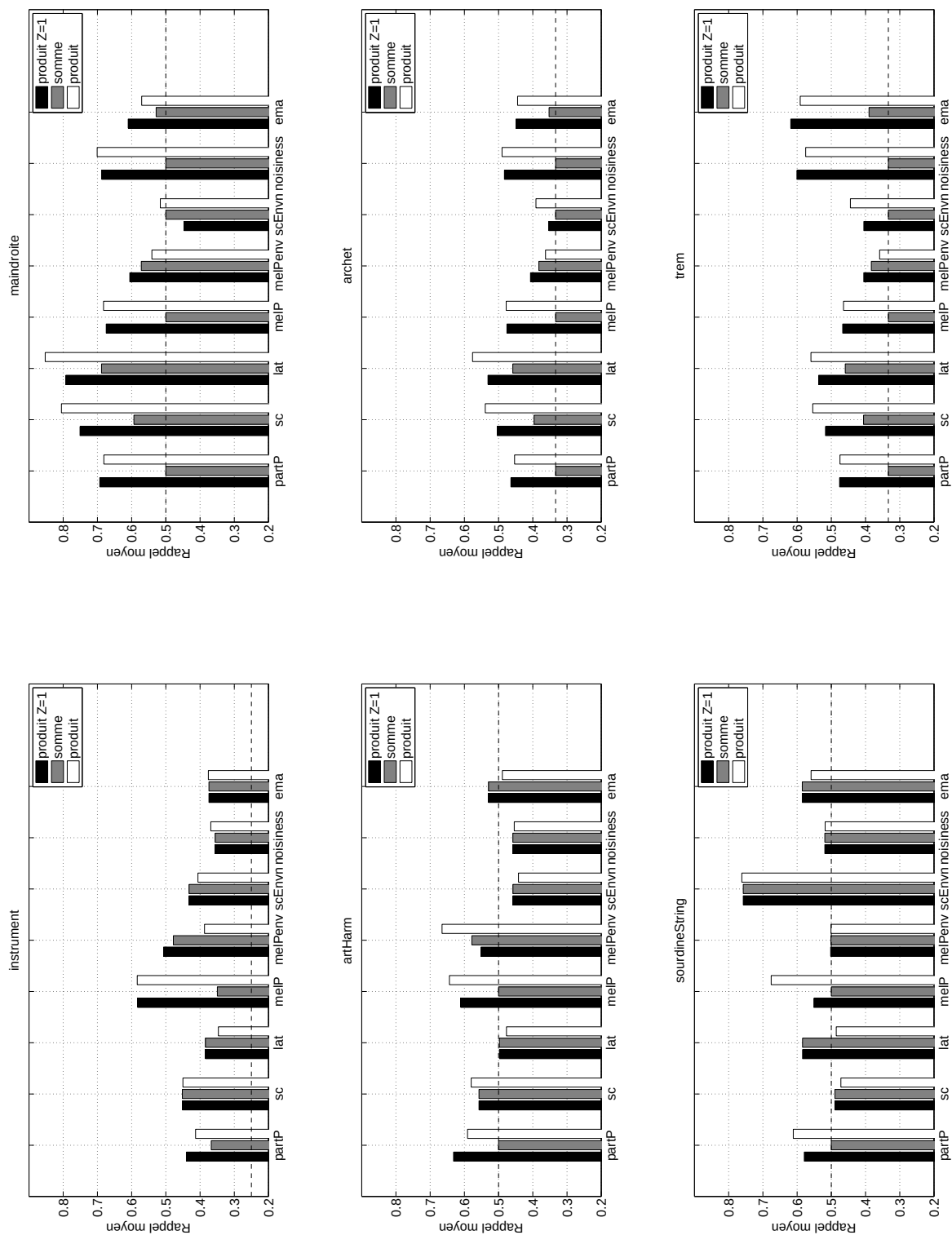


FIG. 7.10 – Résultats de classification par sous-problème pour les cordes selon la méthode d'agrégation utilisée. Les traits tillés représente le résultat d'un classifieur aléatoire

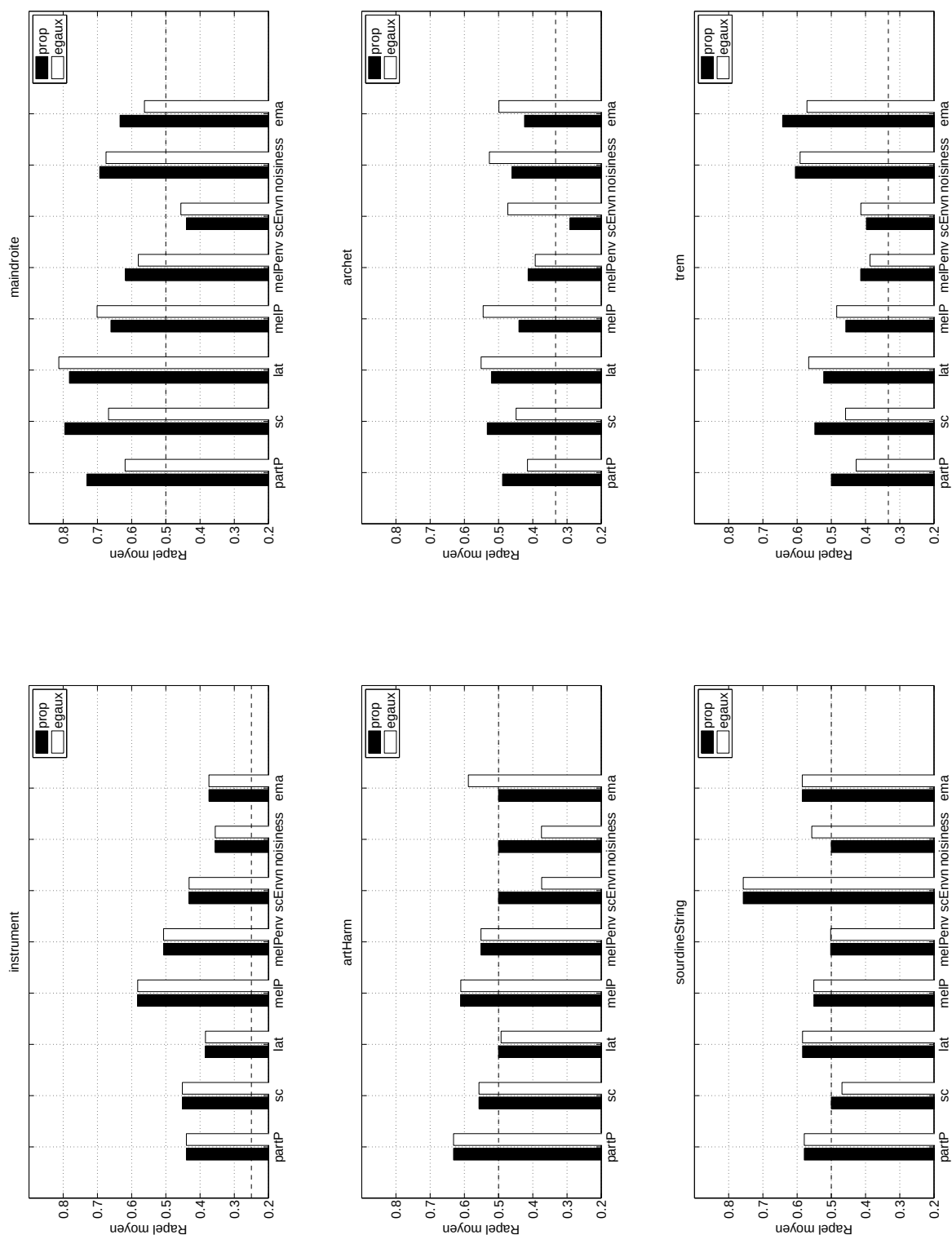


FIG. 7.11 – Résultats de classification par sous-problème pour les cordes avec $Z = 1$ selon la méthode de calcul des poids utilisée. Les traitsillés représente le résultat d'un classifieur aléatoire

des caractéristiques particulières généralisable à tous les instruments. Nous avons ensuite testé les méthodes d'agrégation de modèles. Les résultats obtenus montrent que l'approche fonctionne et que le modèle issu de l'agrégation conserve la capacité de discrimination des modèles agrégés.

Concernant le choix des paramètres, nous utiliserons par la suite le logarithmic opinion pool comme méthode d'agrégation avec $Z=1$ et la méthode des poids proportionnels.

Chapitre 8

Évaluation sur des mélanges synthétiques

8.1 Introduction

Pour évaluer l'ensemble du système nous proposons plusieurs tâches de reconnaissance des sons (instrument, mode de jeu, dynamique) qui composent un mélange synthétique. Ces tests consistent à créer des mélanges artificiels d'échantillons et à étudier dans quelle mesure le système proposé permet de retrouver les sons qui composent le mélange. Le problème de l'orchestration est plus un problème de similarité que de classification, cependant la reconnaissance automatique de mélange donne une mesure objective des performances du système, mesure qu'il est beaucoup plus difficile d'obtenir pour un problème de similarité.

8.2 Procédure

Un ensemble de mélanges synthétiques est d'abord créé selon certaines contraintes (voir paragraphe 8.3). Les descripteurs sont ensuite extraits de ces mélanges. Pour chaque mélange cible, l'ensemble des combinaisons correspondant aux contraintes est généré. Les notes contenues dans la cible sont supposées connues, mais pas le nombre de fois où elles apparaissent. Ainsi, les combinaisons contenant plusieurs fois la même note sont testées, autrement dit, toutes les notes fournies en entrée ne sont pas forcément représentées dans toutes les combinaisons testées. Le modèle des combinaisons est créé, et la cible t est évaluée par rapport à une combinaison c de la manière suivante :

$$y_j^t = (1/f(\mathbf{d}_j = \mathbf{d}_j^t | c = \{x_1, \dots, x_M\}))_{j=1..J}$$
$$\text{avec } \begin{cases} f(\mathbf{d}_j | c) &= G_j(f(\mathbf{d}_j | x_1), \dots, f(\mathbf{d}_j | x_M)) \\ f(\mathbf{d}_j | x_m) &= T(f(\mathbf{d}_j | p_m^1, n_m, h_m), \dots, f(\mathbf{d}_j | p_m^K, n_m, h_m)) \end{cases} \quad (8.1)$$

Où J est le nombre de descripteurs, c est une combinaison composée des sons $x_1 \dots x_M$, $(\mathbf{d}_j^t)$ sont les descripteurs de la cible, G_j est la fonction d'addition du descripteur j et T est le logarithmic opinion pool avec $Z = 1$. Les p^k sont les différents sous-problèmes. Les paramètres qui varient pour chaque combinaison sont les hauteurs h_m , les nuances n_m et les valeurs des variables associées aux sous-problèmes p_m^k .

Cette fonction d'évaluation est à valeur dans $\mathbb{R}_n^+$ donc l'analyse des résultats doit tenir compte de l'aspect multicritères du problème. Notons que dans le contexte de l'optimisation multicritères, le problème doit être un problème de minimisation avec pour minimum 0 c'est pourquoi nous prenons l'inverse de la fonction de densité. Ainsi la cible se trouve à l'origine du repère. La combinaison dont les paramètres sont les mêmes que ceux de la cible (notes, nuances, modes de jeu, instruments) ne se trouve évidemment pas à l'origine du repère. Notre but est qu'elle s'en rapproche le plus possible. Par la suite cette combinaison est appelée solution théorique et notée t_s .

8.3 La base de test

8.3.1 Génération des combinaisons

Les sons utilisés pour créer les combinaisons sont ceux de la base Studio Online. Il faut générer des solutions variées pour tester le plus de cas possibles. Pour cela nous choisissons de maximiser la variété des combinaisons en termes de critères symboliques, c'est-à-dire d'instruments et de modes de jeu. Dans le cas des tests sur des combinaisons de deux sons nous choisissons une seule note et une seule nuance. Nous générons ainsi trois bases de test dont les hauteurs sont *sol2*, *sol3* et *sol4* et la nuance est *mf*. Pour les combinaisons de trois sons nous prenons les *do* et *sol* de toutes les octaves disponibles pour que les harmoniques des sons se recouvrent ce qui correspond au cas le plus difficile et laissons le choix parmi les nuances *pp*, *mf* et *ff*. Les combinaisons sont générées de la manière suivante : Nous cherchons d'abord tous les sons de la base de test (Studio Online) qui correspondent aux critères de hauteur et de nuance désirés. Chaque son est décrit par sa représentation symbolique. Ensuite, nous créons les représentations symboliques de toutes les combinaisons possibles de ces sons. Si le nombre de ces combinaisons est inférieur au nombre de combinaisons désiré alors nous les conservons toutes. Sinon les combinaisons sont regroupées par mode de jeu sans tenir compte de l'instrument, puis un représentant de chaque groupe est choisi au hasard. Finalement, si le nombre de groupe est supérieur au nombre de sons souhaité, les représentants sont tirés au sort et certains groupes ne seront pas représentés, sinon, tous les représentants sont conservés et d'autres combinaisons sont choisies au hasard pour compléter l'ensemble.

Les sons des combinaisons sont ensuite obtenus par simple addition en tenant compte de la sonie théorique (voir chapitre 4).

8.3.2 Sous problèmes considérés et descripteurs utilisés

Les descripteurs utilisés sont :

- La log-puissance des partiels,
- Le centroïde spectral,
- Le logarithme du temps d'attaque,
- L'ambitus de la modulation d'énergie ($f_c = 60$ Hz),
- Le spectre de Mel en log-puissance,
- Le centroïde de l'enveloppe de la partie non-sinusoïdale
- La noisiness.

La table 8.1 récapitule l'ensemble des variables prises en compte et les valeurs possibles.

Sous problème	Valeurs
Instrument	Vn, Va, Vc, Cb, CTp, Hn, Tb, Tbn, BbCl, Bn, Fl
Main droite	Archet, Pizzicato
Harmonique artificiel	Non, Oui
Sourdine	Non, Sourdine normale
Tremolo	Non, Oui
Flatterzunge	Non, Oui

TAB. 8.1 – ensemble des sous-problèmes considérés et des valeurs possibles

8.4 Critères d'évaluation

Ces tests sont assez difficiles à évaluer, le fait d'être dans un contexte multicritères empêche de fournir un seul chiffre, de bonne classification par exemple. Nous proposons trois mesures de performance.

8.4.1 Rang de la solution théorique

Nous utilisons le rang de la solution théorique comme mesure de performance (voir paragraphe 2.7). Le but est que la solution théorique obtienne le rang le plus petit possible, idéalement 1. Le rang est calculé de la manière suivante : Le front de Pareto est extrait avec l'algorithme [Cao], les points du front se voient affecter le rang 1 et sont retirés de l'ensemble, puis le front de Pareto du nouvel ensemble est extrait, les points de ce front sont de rang 2, et le programme itère jusqu'à ce qu'il n'y ait plus de point.

Nous présentons les histogrammes cumulés des rangs obtenus que nous comparons à chaque fois à l'histogramme cumulé qu'obtiendrait un algorithme qui donnerait sa décision aléatoirement.

8.4.2 Classement de la solution théorique connaissant les poids

Lorsque la solution théorique n'est pas de rang 1, nous voulons évaluer si elle est difficile à atteindre en partant du front. Rappelons que la procédure d'orchestration consiste à trouver des combinaisons du front de Pareto (ou proches de ce front) qui soient, aussi, bien répartis sur le front. Ensuite le compositeur choisit une combinaison selon ses préférences. Le choix du compositeur permet de retrouver les poids induits par cette combinaison pour chaque descripteur. Ensuite une procédure explore l'espace localement en prenant en compte les poids trouvés. Nous cherchons donc à savoir si une fois les poids connus, la solution théorique est bien classée dans l'ordre induit par les poids et donc facilement atteignable. il faut d'abord déterminer le jeu de poids correspondant à la solution théorique. Il est donné par l'équation :

$$\lambda_j^{t_s} = \frac{\prod_{i \neq j} y_i^{t_s}}{\sum_{j=1}^J \prod_{i \neq j} y_i^{t_s}}, j = 1, \dots, J \quad (8.2)$$

Où $\lambda_j^{t_s}$ est le poids associé à la j -ième dimension, c'est-à-dire au j -ième descripteur, et $y_j^{t_s}$ est l'évaluation de la cible par rapport à la solution théorique. Ces poids donnent une norme de Chebyshev pondérée :

$$D_{\text{Chebyshev}} = \max_j (\lambda_j^{t_s} |y_j|) \quad (8.3)$$

La norme de Chebyshev est la distance utilisée pour déterminer le front de Pareto. On classe donc l'ensemble des combinaisons selon cette norme. Plus la norme d'une combinaison est faible plus cette combinaison est optimale au sens de la norme induite par les poids. Le classement de la solution théorique selon cette norme donne une bonne indication de la possibilité d'atteindre la solution théorique dans une procédure d'orchestration.

8.4.3 Similarité symbolique entre la solution théorique et la solution du front

Une autre mesure intéressante lorsque la solution théorique n'est pas dans le front est la similarité entre la solution théorique et la solution la plus proche qui appartient au front. La solution de rang 1 la plus proche de la solution théorique est la première dans le classement issu de la norme précédemment trouvée. Pour quantifier la distance entre ces deux combinaisons nous utilisons une distance de Hamming qui est une distance symbolique entre deux combinaisons. Chaque combinaison est décrite par sa représentation symbolique obtenue par concaténation des représentations symboliques des sons qui la composent (voir chapitre 3). La distance de Hamming correspond au nombre de valeurs qui diffèrent entre deux représentations a et b :

$$d(a, b) = \text{Card}\{i : a_i \neq b_i\} \quad (8.4)$$

Par exemple un violon tremolo et un violon non tremolo auront une distance de 1, de même qu'un violon et un violoncelle tremolo. Généralement, la distance de Hamming entre deux sons donne une assez bonne indication de la similarité sonore.

8.5 Résultats

8.5.1 Mélanges de 2 sons

La figure 8.1 montre l'histogramme cumulé du rang pour la note sol3. Nous constatons que la méthode proposée permet d'avoir un peu plus de 60% des solutions recherchées dans le front de Pareto contre un peu moins de 25% pour un algorithme aléatoire. Notons que ce dernier chiffre correspond à la proportion totale de solutions qui appartiennent au front. Il est très élevé car nous travaillons dans un espace de dimension 7, et plus le nombre de dimensions est élevé plus le front de Pareto est grand [KC07]. De plus, dans ce cas, toutes les combinaisons contiennent des sons de même hauteur et

nuance, et sont donc très proches perceptivement. La taille du front de Pareto met aussi en évidence la limite d'une approche multicritères lorsque le nombre de critères augmente. Cela montre la nécessité de mettre en place des techniques d'interaction avec le compositeur pour découvrir ses préférences, c'est-à-dire les poids associés à chaque descripteur, ou de méthode permettant de ne pas prendre en compte certains descripteurs selon le type de cible.

La figure 8.2 présente le classement de la solution théorique par la norme induite lorsque la solution

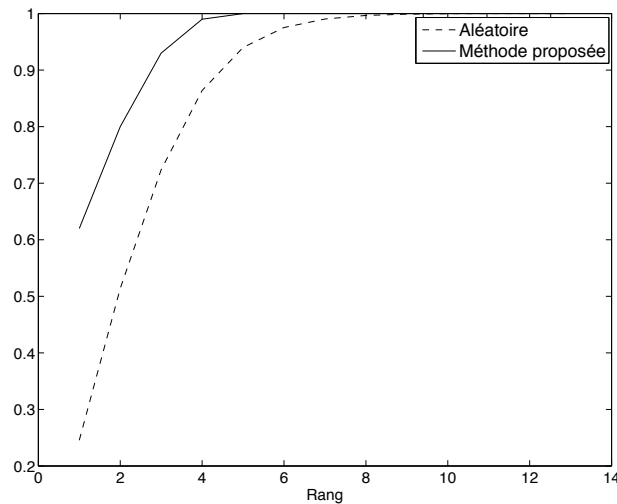


FIG. 8.1 – Histogramme cumulé du rang de la solution théorique pour la note sol3

théorique n'est pas de rang 1. Dans tous les cas la position de la solution théorique est inférieure à 30 et très souvent inférieure à 10 pour un nombre total de combinaisons de l'ordre de 5000, elle est donc facilement atteignable depuis le front en supposant que l'on obtienne les bons poids. Un autre

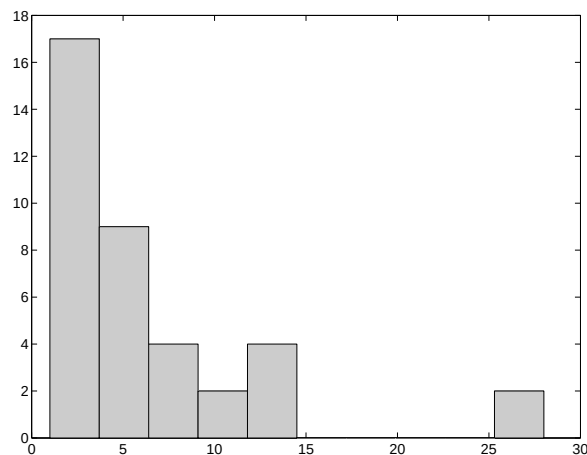


FIG. 8.2 – Classement de la solution théorique selon la norme qu'elle induit pour les combinaisons de deux sons en sol3

point de vue sur les résultats est donné par l’histogramme des distances de Hamming entre la solution théorique et la solution la plus proche qui appartient au front. L’histogramme présenté dans la figure 8.3 ne montre que les cas où il y a des erreurs ce qui explique l’absence de distance nulle. La majorité des erreurs donnent une distance inférieure ou égale à 4. La comparaison de cette distribution avec l’histogramme des distances de Hamming entre toutes les combinaisons possibles (figure 8.3) montre bien l’apport de la méthode proposée dans le sens où elle rapproche du front les solutions similaires à la cible.

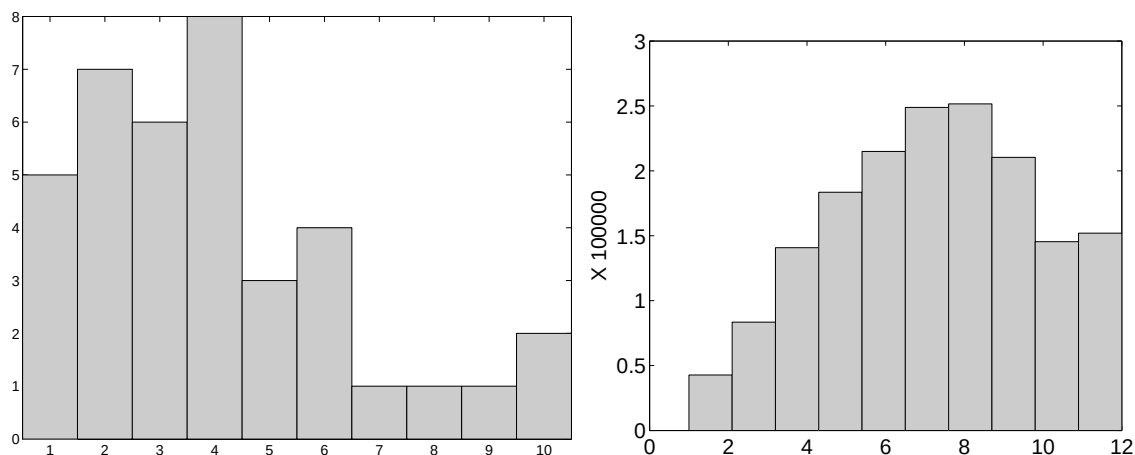


FIG. 8.3 – Histogramme de la distance de Hamming entre la solution théorique et la combinaison de rang 1 la plus proche (à gauche) et entre toutes les combinaisons (à droite) pour les combinaisons de deux sons

Les résultats pour les notes sol2 et sol4 sont très proches de ceux de sol3, nous ne les détaillons donc pas. Nous donnons pour information dans la figure 8.4 les histogrammes cumulés des rangs obtenus pour ces hauteurs.

8.5.2 Mélanges de 3 sons

Dans les tests sur des mélanges de trois sons, les combinaisons peuvent contenir plusieurs notes et plusieurs dynamiques différentes. Le nombre de combinaisons possibles est donc très grand. Il est impossible d’évaluer la totalité des combinaisons, cela demande trop d’espace mémoire. Pour diminuer le nombre de combinaisons, nous fournissons la liste des instruments contenus dans la cible. La recherche se fait donc sur un sous-ensemble réduit d’instrument, sur les nuances, les hauteurs et les modes de jeu. Le nombre de combinaisons possibles est alors de l’ordre de 10^5 . Le nombre de cibles testées est de 464.

Les résultats sont assez similaires à ceux des mixtures de deux sons. La figure 8.5 présente l’histogramme cumulé du rang obtenu. Dans 54% des cas, la solution théorique est dans le front contre 29% dans le cas aléatoire. Dans le cas où la solution théorique n’est pas dans le front, elle est très bien classée par la norme induite (figure 8.6 en bas). Dans la grande majorité des cas elle se classe à une position inférieure à 10. De plus il existe une solution du front qui présente beaucoup de similarité symbolique avec la cible. En effet, la distance de Hamming entre la solution théorique et la solution

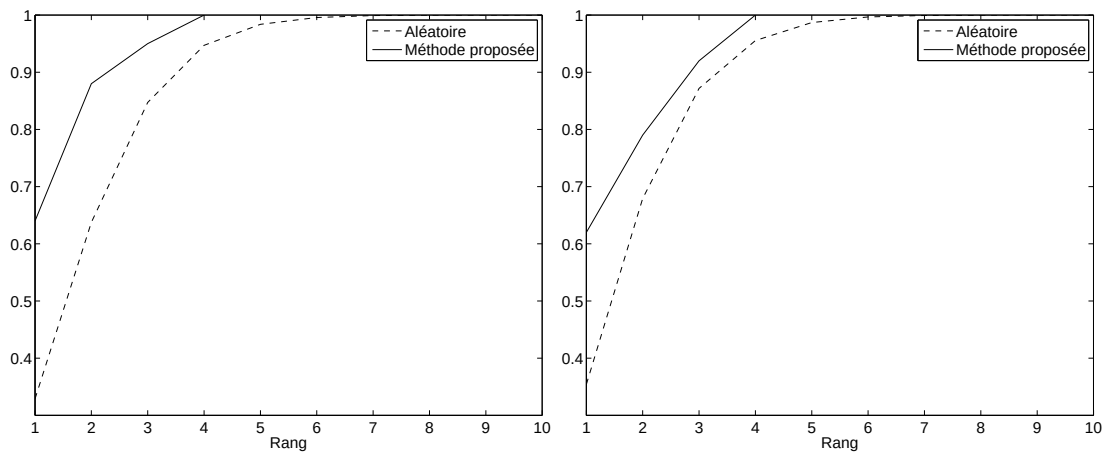


FIG. 8.4 – Histogramme cumulé du rang de la solution théorique pour la note sol2 à gauche et sol4 à droite

du front la plus proche est souvent inférieure à 3 et jamais supérieur à 10 pour un maximum possible de 21.

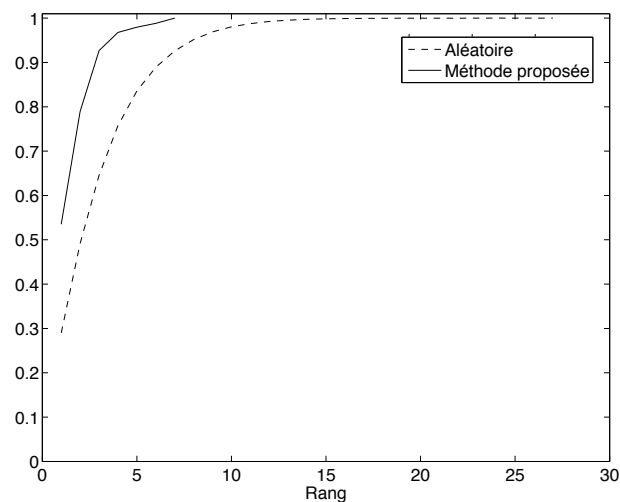


FIG. 8.5 – Histogramme cumulé du rang de la solution pour les combinaisons de 3 sons pour la recherche des modes de jeu

8.6 Conclusion

Dans ce chapitre, nous avons proposé une méthode d'évaluation objective du système proposé. Cette méthode consiste à retrouver les éléments d'une combinaison artificielle de sons. Nous avons, en outre, proposé trois mesures permettant d'analyser les résultats des tests. Les résultats montrent

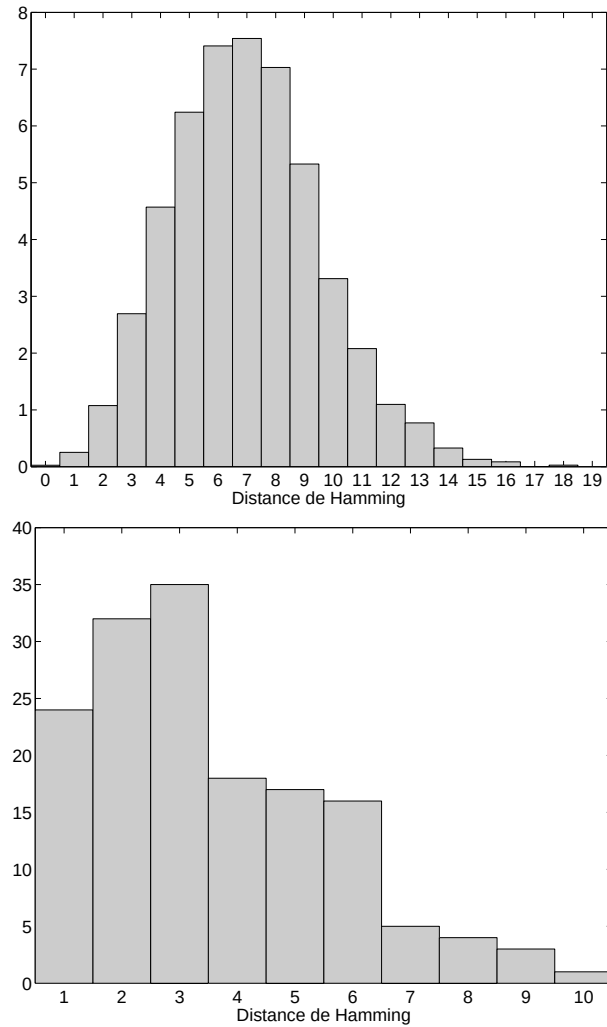


FIG. 8.6 – Histogramme de la distance de Hamming entre la solution théorique et la combinaison de rang 1 la plus proche (en bas) et entre toutes les combinaisons (en haut) pour les combinaisons de trois sons

que la méthode proposée organise l'espace des combinaisons en rapprochant la solution théorique du front de Pareto. Ils montrent aussi les limites d'une approche multicritères lorsque le nombre de critères augmentent et la nécessité de trouver des méthodes de présélection des descripteurs ou des sons selon le type de cible.

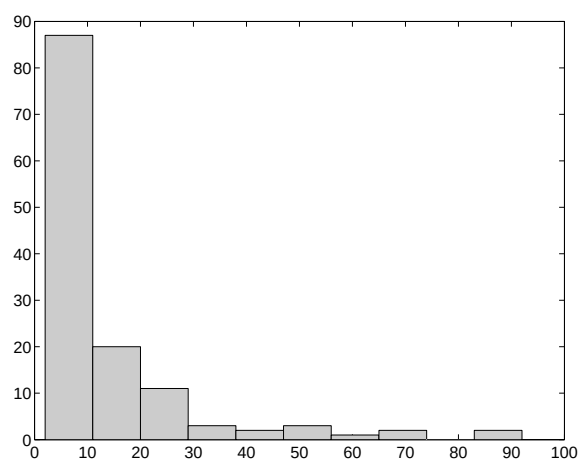


FIG. 8.7 – Classement de la solution théorique selon la norme qu'elle induit pour les combinaisons de trois sons

Conclusion

Au cours de ce travail nous avons proposé une méthode d'orchestration assistée par ordinateur. Dans le système proposé, le compositeur peut fournir une cible sous forme de son enregistré. Le programme recherche alors les combinaisons de sons instrumentaux qui s'approchent le plus de la cible. L'élaboration d'un tel système soulève des questions liées à la perception des sons musicaux, à l'acoustique, à la description et à la modélisation du timbre des instruments. Les réponses à ces questions nous ont permis de proposer une définition formelle du problème de l'aide à l'orchestration. Il est ainsi vu comme un problème d'optimisation multicritères probabiliste. Dans cette thèse, nous avons abordé uniquement la description des sons et des combinaisons de sons ainsi que la comparaison de la cible avec les combinaisons.

Nous avons proposé un ensemble de descripteurs du signal permettant de décrire plusieurs dimensions du timbre instrumental. Pour un sous-ensemble de ces descripteurs, nous avons fourni les méthodes permettant d'estimer les descripteurs du timbre d'une combinaison à partir des descripteurs des sons qui la compose. Pour le cas des combinaisons contenant un son percussif et un son entretenu, nous avons réalisé deux expériences perceptives. Ces deux expériences nous ont permis de trouver un critère acoustique expliquant la fusion entre les sons et une description acoustique du timbre de ces combinaisons. Ensuite, nous avons proposé une méthode de modélisation des instruments et de leurs modes de jeu. Le modèle est basé sur une taxinomie des modes de jeu instrumentaux et sur l'apprentissage de la distribution des descripteurs pour chaque valeur des variables de la taxinomie. Les modèles des différents modes de jeu sont ensuite agrégés par une méthode issue du domaine de la décision collective, le logarithmic opinion pool. Nous avons montré par des tâches de classification que les différents modes de jeu pouvaient être appris séparément et que la méthode proposée permettait d'apprendre des modèles robustes à partir de peu d'échantillons. Les résultats de ces tests révèlent une bonne capacité de discrimination des modes de jeu. Nous avons aussi montré que ces modèles pouvaient être agrégés tout en conservant leur capacité de discrimination. Grâce à cette méthode, nous obtenons les modèles d'un grand nombre de sons instrumentaux éventuellement absents de la base d'apprentissage. Enfin nous avons évalué la méthode proposée par des tâches de classification d'échantillons polyphoniques. Ces tests ont permis de constater que la méthode proposée rapproche les solutions intéressantes du front de Pareto. Nous proposons maintenant plusieurs perspectives de recherche.

La première concerne la description des modulations d'énergie et de fréquence. Le descripteur que nous utilisons, qui calcule la modulation sur la totalité du signal, fonctionne très bien pour les sons monodiques mais ne rend pas compte de toute l'information disponible dans les cas polyphoniques et donc en particulier pour la cible. Il serait intéressant de calculer les modulations sur plusieurs bandes de fréquence du spectrogramme et ainsi d'obtenir des vecteurs d'ambitus et de vitesse de modulation.

Cela permettrait d'avoir une description plus fine des caractéristiques temporelles du son.

La description des combinaisons pourrait aussi être améliorée. Nous n'avons pas fourni de méthode d'addition pour tous les descripteurs proposés, en particulier pour les descripteurs de variation du spectre. Estimer ces descripteurs pour une combinaison de sons suppose de préciser un modèle temporel de l'addition de signaux dans une salle. Comme nous l'avons vu, lorsque deux sons sont émis simultanément dans une salle, le son résultant contient des variations complexes et différentes dans chaque bande de fréquence. Nous en avons déduit un comportement moyen du spectre de la combinaison, mais n'avons pas abordé la question de la variation temporelle. Il faudrait donc étudier le comportement dans une salle d'un son contenant initialement des variations : comment ces variations sont transmises par la salle, comment elles s'ajoutent à celles d'un autre son émis simultanément et enfin ce que perçoit un auditeur. Ce dernier point souligne la nécessité de continuer les recherches sur la perception des combinaisons de sons. Nous avons proposé deux exemples d'expériences permettant de traiter cette question. Ces méthodes pourraient être utilisées pour étudier les combinaisons de sons variant au cours du temps. Il serait aussi intéressant d'étudier des combinaisons contenant beaucoup d'instruments. La question de la fusion, par exemple, se poserait alors différemment. De telles combinaisons fusionnent souvent bien et la question est, en général, plutôt de savoir si un instrument émerge de l'ensemble. Le protocole expérimental devrait être différent de celui que nous avons proposé puisqu'une approche exhaustive comme la nôtre n'est plus possible. Plutôt que la notion de fusion, il s'agirait de traiter celle de l'émergence. Cette question est très proche de celles de l'analyse de scène auditive. De telles expériences permettraient d'inclure les notions de fusion et d'émergence dans le système. Les résultats dont nous disposons aujourd'hui ne portent que sur deux sons et ne sont donc pas directement applicables.

Concernant l'apprentissage, il serait sans doute profitable de trouver automatiquement si certaines variables de la taxinomie méritent d'être apprises conjointement. Dans le travail présenté, chaque variable est apprise séparément, mais il est probable que l'apprentissage conjoint de certaines variables améliore les résultats. Il s'agit ici de trouver un compromis entre la taille de l'ensemble d'apprentissage, l'interaction entre les effets des variables sur les descripteurs, le qualité de modélisation et les possibilités de généralisation. Pour cela, il serait possible d'étudier directement les interactions entre variables pour chaque descripteur avec des méthodes de type analyse de variance ou de tester les résultats de classification pour les différentes combinaisons de variables.

Sur un plan plus prospectif, une meilleure prise en compte de l'aspect temporel de l'orchestration pourrait être utile. Mis à part certains descripteurs de l'enveloppe temporelle du son, le temps est vu dans notre travail comme une succession de segments. Pour orchestrer un son variant dans le temps, il faut le découper en segments et orchestrer chaque segment séparément en ajoutant éventuellement des contraintes entre les différents segments pour éviter des sauts permanents de hauteur ou des changements trop fréquents d'instruments. Cette approche fonctionne pour beaucoup de sons en particulier dans le cas des sons entretenus. En revanche elle montre ses limites pour des sons de structure spectro-temporelle très complexe qui seraient mieux reproduits, par exemple, par des ensembles de percussions. Le problème est ici d'une part de ne pas trop augmenter la combinatoire et d'autre part conserver un temps d'évaluation raisonnable pour les combinaisons. Ainsi des approches de type HMM utilisées en séparation de sources semblent d'emblée exclues. Il pourrait être intéressant de passer à une vision spectro-temporelle du problème. Nous utilisons des spectres moyens et tous les

sons sont supposés commencer en même temps. Il faudrait que les sons puissent commencer n'importe quand et pour cela que la description ne soit plus le spectre moyen mais le spectrogramme. Il serait alors possible de s'inspirer des méthodes de type matching pursuit pour résoudre le problème. Dans le cadre d'un système contenant une description explicite du temps il serait utile d'ajouter des connaissances sur l'instrumentation. De même que nous avons intégré des connaissances sur les modes de jeu possibles ou sur la tessiture des instruments, les possibilités d'enchaînements entre les différents doigtés ou les différents modes de jeu pourraient être introduites. Ceci présente deux avantages. D'abord cela oblige le système à proposer des solutions faisables ou dont le degré de facilité peut être mesuré. Ensuite une telle base de connaissance pourrait être très utile aux compositeurs. Son élaboration nécessite d'abord de définir une structure de données capable de représenter ce genre d'information. La taxinomie proposée dans ce travail peut servir de point de départ à l'élaboration d'une telle structure de donnée.

Une autre perspective concerne la simulation des résultats. Nous créons les modèles correspondant à des sons qui ne sont pas contenus dans la base d'apprentissage, ainsi au moment de la simulation des résultats il se peut qu'il manque des sons. Nous avons proposé une solution par arbre de décision en montrant que l'arbre pouvait être utilisé pour connaître les transformations à effectuer sur un son proche du son manquant. Les transformations les plus évidentes sont celles de l'amplitude et de la hauteur du signal. La connaissance des modèles des modes de jeu nous donne beaucoup plus d'information et pourrait permettre de transformer les sons de manière beaucoup plus radicale, par exemple spectralement. En approfondissant cette idée, nous pourrions imaginer de synthétiser tous les modes de jeu à partir d'une base ne contenant que les sons ordinaires et ainsi gagner beaucoup d'espace mémoire.

Pour finir, la perspective la plus importante est l'intégration des modèles proposés dans un algorithme d'exploration de l'espace des combinaisons tel que celui de [CTA⁺07]. Les premières expériences effectuées permettent de trouver des propositions d'orchestration intéressantes, mais montrent que l'évaluation des combinaisons par les modèles peut être biaisée pour certains descripteurs en faveur des combinaisons contenant peu d'instruments. Cela est sans doute dû aux méthodes d'addition de modèles qui donnent une variance élevée aux modèles des combinaisons contenant beaucoup d'instruments. Mis à part ce biais, l'utilisation des modèles d'instruments dans la recherche d'orchestration offre des perspectives intéressantes. Ces perspectives reposent principalement sur l'exploitation du lien, révélé par les modèles, entre description symbolique et description acoustique. Une première possibilité est d'utiliser cette caractéristique dans une procédure de recherche locale. Si une solution est bien évaluée sur tous les descripteurs sauf un, il est possible de trouver la variable de la taxinomie qui a le plus d'influence sur ce dernier descripteur et de ne changer que la valeur de cette variable dans les sons de la combinaison. D'autre part, la connaissance du lien entre variables symboliques et variables acoustiques peut être utilisée pour sélectionner un sous-ensemble de sons et de descripteurs avant de commencer la recherche. En classant la cible avec les modèles de modes de jeu nous pourrions décider que certains modes de jeu sont inutiles et que les descripteurs qui les expliquent le mieux peuvent être omis. Par exemple en classant la cible avec le modèle de *tremolo*, nous pourrions constater que la cible ne contient pas de modulation d'amplitude et donc que les sons de la base contenant des modulations d'amplitude sont inutiles, de même que les descripteurs de modulations.

Annexe A

Descripteurs utilisés dans les expériences perceptives

Nous détaillons ici l'ensemble des descripteurs utilisés dans le chapitre 5, partie 5.3. Tous ces descripteurs sont calculés sur le sons percussif, sur le son entretenu et sur le son du mélange. Les trois versions sont utilisées comme descripteurs du mélange. Nous utilisons aussi la différence entre le descripteur du son entretenu et celui du son percussif.

Le centroïde et l'étendue spectral

Soit $p(f)$ le spectre d'une trame, où f est la fréquence. Alors le centroïde spectral s_c de la trame est :

$$s_c = \frac{\int_{f=0}^F f p(f) df}{\int_{f=0}^F p(f) df} \quad (\text{A.1})$$

Où F est la fréquence de Nyquist.

L'étendue spectrale est définie comme l'écart type du spectre.

$$s_s = \sqrt{\frac{\int_{f=1}^F (f - s_c)^2 p(f) df}{\int_{f=1}^F p(f) df}} \quad (\text{A.2})$$

Ces deux descripteurs sont calculés sur deux échelles fréquentielles : linéaire et logarithmique, et sur trois échelles d'amplitude : linéaire, carré (homogène à l'énergie) et logarithmique. Il sont aussi calculé sur une échelle fréquentielle de Bark en amplitude carré.

Le spectre de Mel

Le spectre de Mel est calculé comme dans le chapitre 4.

Le logarithme du temps d'attaque

Le logarithme du temps d'attaque est calculé comme dans le chapitre 4.

La croissance temporelle

La croissance est définie comme la pente moyenne de l'énergie pendant l'attaque [Pee04].

La décroissance temporelle

La décroissance temporelle est une mesure de la décroissance de l'énergie du signal. L'énergie du signal est modélisée par :

$$\hat{e}(t) = A \exp(-\alpha(t - t_{max})) \quad t > t_{max} \quad (\text{A.3})$$

où t_{max} est le temps correspondant au maximum de l'énergie. Les paramètres A et α sont estimés par régression linéaire sur le logarithme de l'enveloppe d'énergie.

Le centroïde temporel

Le centroïde temporel est la moyenne temporelle de l'enveloppe d'énergie :

$$t_c = \frac{\sum_t t e(t)}{\sum_t e(t)} \quad (\text{A.4})$$

Moyenne temporelle des descripteurs spectraux

Les descripteurs spectraux sont moyennés en pondérant par la sonie.

Annexe B

Taxinomie complète des modes de jeu

```
val_s.vib    = { 'no', 'yes' };  
dep_s.vib    = { 'famille', 'strings', 'woodwinds', 'brass', 'doublereed' };
```

```
val_s.maindroite = { 'archet', 'pizz' };  
dep_s.maindroite = { 'famille', 'strings' };
```

```
val_s.artHarm    = { 'no', 'yes' };  
dep_s.artHarm    = { 'famille', 'strings' };
```

```
val_s.archet     = { 'ord', 'legno-tratt', 'legno-batt' };  
dep_s.archet     = { 'famille', 'strings',  
                    'maindroite', 'archet' };
```

```
val_s.posArchet  = { 'ord', 'pont', 'tasto' };  
dep_s.posArchet  = { 'famille', 'strings',  
                    'maindroite', 'archet' };
```

```
val_s.sourdineString = { 'N', 'S', 'SP' };  
dep_s.sourdineString = { 'famille', 'strings' };
```

```
val_s.trem       = { 'no', 'yes' };  
dep_s.trem       = { 'famille', 'strings',  
                    'maindroite', 'archet' };
```

```
val_s.flatt      = { 'no', 'yes' };  
dep_s.flatt      = { 'famille', 'woodwinds', 'brass', 'doublereed' };
```

```
val_s.aeolian    = { 'ord', 'aeol+ord', 'aeol' };  
dep_s.aeolian    = { 'famille', 'woodwinds' };
```

```
val_s.sourdineBrass = {'N', 'C', 'H', 'P', 'S', 'W'};  
dep_s.sourdineBrass = {'famille', 'brass'};
```

Annexe C

Liste des instruments et des abréviations

Saxophone alto	Asax
Clarinete basse	BbBCl
Clarinete contrebasse	BbCbCl
Clarinete en sib	BbCl
Clarinete en mib	EbCl
Hautbois	Ob
Cor anglais	EH
Flûte piccolo	Picc
Flûte	Fl
Flûte basse	BFl
Flûte contrebasse	CbFl
Basson	Bn
Trombone	Tbn
Trombone basse	BTbn
Trompette en ut	CTp
Cor	Hn
Tuba	Tb
Violon	Vn
Alto	Va
Violoncelle	Vc
Contrebasse	Cb

Bibliographie

- [AK99] F. M. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20(11-13) :1361–1369, 1999.
- [BB99] G. Ballet and R. Borghesi. Studio online 3.0 : An internet "killer application" for remote access to Ircam sounds and processing tools. In *Journées d'Informatique Musicale*, 1999.
- [BBG06] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1) :191 – 199, Jan 2006.
- [Ber70] H. Berlioz. *Traité d'instrumentation et d'orchestration*. Gregg International Publishers, réimpression de l'éd. de paris, 1855 edition, 1970.
- [BHM01] J. C. Brown, O. Houix, and S. McAdams. Feature dependence in the automatic identification of musical woodwind instruments. *The Journal of the Acoustical Society of America*, 109(3) :1064–1072, 2001.
- [Bla97] A. Blatter. *Instrumentation and Orchestration*. Cengage Learning, 1997.
- [Bre90] A. S. Bregman. *Auditory Scene Analysis : The Perceptual Organization of Sound*. MIT Press, 1990.
- [Bre91] A. S. Bregman. Timbre, orchestration, dissonance et organisation auditive. In *Le timbre, métaphore pour la composition*. Christian Bourgeois, 1991.
- [Bro99] J. C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *The Journal of the Acoustical Society of America*, 105(3) :1933–1941, 1999.
- [Cao] Yi Cao. Pareto front. <http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17251&objectType=File>.
- [Cha81] Gerard R. Charbonneau. Timbre and the perceptual effects of three types of data reduction. *Computer Music Journal*, 5(2) :10–19, 1981.
- [Chi07] Frederic Chiasson. L'orchestration selon koechlin : recherche d'un point de rencontre entre la pratique musicale et les sciences cognitives. In *Proceedings of Composer au XXIe Siècle - Processus et Philosophie*, Montréal, Canada, mars 2007.
- [CM58] A. Casella and V. Mortari. *La Technique de l'orchestre contemporain*. Ricordi, 1958.

- [CTA⁺07] Grégoire Carpentier, Damien Tardieu, Gérard Assayag, Xavier Rodet, and Emmanuel Saint-James. An Evolutionary Approach to Computer-Aided Orchestration. In Mario Giacobini et al., editor, *Applications of Evolutionary Computing. EvoWorkshops 2007 : EvoCOMNET, EvoFIN, EvoIASP, EvoINTERACTION, EvoMUSART, EvoSTOC and EvoTRANSLOG*, pages 488–497, Valencia, Spain, April 2007. Springer. Lecture Notes in Computer Science Vol. 4448.
- [DHAT00] P.W.M. Desain, H.J. Honing, M.L.P. Aarts, and R. Timmers. Rhythmic aspects of vibrato. In P.W.M Desain and W.L. Windsor, editors, *Rhythm Perception and Production*, pages 203–216. Routledge, 2000.
- [DMW94] Sophie Donnadieu, Stephen McAdams, and Suzanne Winsberg. Context effects in timbre space. In *Conference of the Society for Music Perception and Cognition*, pages 311–312, Liège, Belgique, Juillet 1994.
- [DR93] B. Doval and Xavier Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *Proc. ICASSP*, pages 221–224, Minneapolis, USA, 1993.
- [ENRC03] Taoufik En-Najjary, Olivier Rosec, and Thierry Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *EUROSPeech*, pages 1753–1756, 2003.
- [ERD06] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1) :68–80, Jan 2006.
- [Ero01a] A Eronen. Comparison of features for musical instrument recognition. In *WASPAA (IEEE Workshop on Applications of Signal Processing to Audio and Acoustics)*, New York, USA, 2001.
- [Ero01b] Antti Eronen. Automatic musical instrument recognition. Master’s thesis, Tampere university of technology, 2001.
- [Fau00] Anne Faure. *Des sons aux mots : Comment parle-t-on du timbre musical ?* Thèse de doctorat, Paris, Ecoles des Hautes Etudes en Sciences Sociales, Paris, 2000.
- [Fen60] L. Fenton. The sum of log-normal probability distributions in scatter transmission systems. *IEEE Transactions on Communications*, 8(1) :57– 67, Mar 1960.
- [FS67] Harvey Fletcher and Larry C. Sanders. Quality of violin vibrato tones. *The Journal of the Acoustical Society of America*, 41(6) :1534–1544, 1967.
- [Gal98] M. J. F. Gales. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25(1-3) :49–74, 1998.
- [GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3 :1157–1182, 2003.
- [GM77] John M. Grey and James A. Moorer. Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62(2) :454–462, 1977.

- [GM90] B.R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear Res*, 47(1-2) :103–138, Aug 1990.
- [Got04] Masataka Goto. Development of the rwc music database. In *18th International Congress on Acoustics (ICA 2004)*, volume I, pages 553–556, April 2004.
- [Gre77] John M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5) :1270–1277, 1977.
- [GS05] R.L. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison (Eds.), editors, *Cambridge Handbook of Thinking and Reasoning*, pages 13–36. Cambridge University Press, 2005.
- [GT82] I Gati and A Tversky. Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology : Human Perception and Performance*, (8) :325–340, 1982.
- [GWZ84] C. Genest, S. Weerahandi, and J. V. Zidek. Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17(1) :61–70, 1984.
- [GZ86] Christian Genest and James V. Zidek. Combining probability distributions : A critique and an annotated bibliography. *Statistical Science*, 1(1) :114–135, feb 1986.
- [HBKD06] Perfecto Herrera-Boyer, Anssi Klapuri, and Manuel Davy. Automatic classification of pitched musical instrument sounds. *Signal Processing Methods for Music Transcription*, pages 163–200, 2006.
- [Hel98] H.L.F. Helmholtz. *Sensations of tone*. Bristol : Thoemmes, 1998.
- [HG91] Joseph W. Hall and John H. Grose. Some effects of auditory grouping factors on modulation detection interference (mdi). *The Journal of the Acoustical Society of America*, 90(6) :3028–3035, 1991.
- [Hin99] G. Hinton. Products of experts. In *Proc. ICANN*, pages 1–6, Edinburgh, Scotland, 1999.
- [Hug] Andrew Hugill. The Orchestra : A User’s Manual. <http://www.mti.dmu.ac.uk/~ahugill/manual/intro.html>.
- [Hum05] Thomas A. Hummel. Simulation of human voice timbre by orchestration of acoustic music instruments. In *Proceedings of International Computer Music Conference 2005*, 2005.
- [Irc] Ircam Solo Instruments. www.ultimatesoundbank.com/uviscp1.html.
- [JDM00] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :4–37, 2000.
- [JHD99] M. Junker, R. Hoch, and A. Dengel. On the evaluation of document analysis components by recall, precision, and accuracy. In IEEE, editor, *Proceedings of the Fifth International Conference on Volume*, pages 713 – 716, 1999.
- [KC91] R.A. Kendall and E.C. Carterette. Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8(4) :369–404, 1991.

- [KC93] R.A. Kendall and E.C. Carterette. Identification and blend of timbre as a basis for orchestration. *Contemporary Music Review*, 9 :51–67, 1993.
- [KC07] Joshua Knowles and David Corne. Quantifying the effects of objective space dimension in evolutionary multiobjective optimization. *Evolutionary Multi-Criterion Optimization*, pages 757–771, 2007.
- [KM98] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 285–288, 1998.
- [Koe54] Charles. Koechlin. *Traité de l'orchestration*. Max Eschig, 1954.
- [Kri93] Jochen Krimphoff. *Analyse acoustique et perception du timbre*. Thèse de doctorat, Le Mans, Université du Maine, 1993.
- [Kut00] Heinrich Kuttruff. *Room acoustics*. Spon Press, 4. ed. edition, 2000.
- [L⁺84] L.Breiman et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.
- [LR03] Arie Livshin and Xavier Rodex. The importance of cross database evaluation in sound classification. In *ISMIR 2003*, Baltimore, Etats-Unis, Octobre 2003.
- [Mar99] K. Martin. *Sound Source Recognition : A Theory and Computational Model*. PhD thesis, MIT, Medialab, 1999.
- [Mar01a] Y. Maresz. *Metallics*. Durand, Paris, 1995-2001.
- [Mar01b] Y. Maresz. *Metal extensions*. Durand, Paris, 2001.
- [MBM99] Stephen McAdams, James W. Beauchamp, and Suzanna Meneguzzi. Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *The Journal of the Acoustical Society of America*, 105(2) :882–897, 1999.
- [Moe04] Dirk Moelants. Temporal aspects of instrumentalists performance of tremolo, trills and vibrato. In *Proceedings of the International Symposium on Musical Acoustics (ISMA)*, pages 281–284, Nara, Japan, April 2004.
- [MS06] M. J. F. Gales and S. S. Airey. Product of gaussians for speech recognition. *Computer Speech and Language*, 20(1) :22–40, january 2006.
- [MW00] Maureen Mellody and Gregory H. Wakefield. The time-frequency characteristics of violin vibrato : Modal distribution analysis and synthesis. *The Journal of the Acoustical Society of America*, 107(1) :598–611, 2000.
- [MWD⁺95] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58 :177–192, 1995.
- [NPL⁺00] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Technical Report WS00AVSR, Johns Hopkins University, CLSP, 2000.
- [OD01] Andrew J. Oxenham and Torsten Dau. Modulation detection interference : Effects of concurrent and sequential streaming. *The Journal of the Acoustical Society of America*, 110(1) :402–408, 2001.

- [Pee03] Geoffroy Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *AES 115th Convention*, New-York, USA, Octobre 2003.
- [Pee04] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification). in the CUIDADO project. Paris, IRCAM, 2004.
- [PM99] D. Pressnitzer and S. McAdams. An effect of the coherence between envelopes across frequency regions on the perception of roughness. In T. Dau, V. Hohmann, and B. Kollmeier, editors, *Psychophysics, Physiology and Models of Hearing*, pages 105–108. World Scientific, London, 1999.
- [PMH00] G. Peeters, S. McAdams, and P. Herrera. Instrument description in the context of mpeg-7. In *Proceedings of International Computer Music Conference 2000*, Berlin, Germany, 2000.
- [Pre98] Daniel Pressnitzer. *Perception de rugosité psychoacoustique : D'un attribut élémentaire de l'audition à l'écoute musicale*. Thèse de doctorat, Université Paris VI/IRCAM, Paris, 1998.
- [Pse03] David Psenicka. Sporch : An algorithm for orchestration based on spectral analyses of recorded sounds. In *Proceedings of International Computer Music Conference 2003*, 2003.
- [RH05] François Rose and James Hetrick. Spectral analysis as a resource for contemporary orchestration technique. In *Proceedings of Conference on Interdisciplinary Musicology 2005*, 2005.
- [RK64] Nikolai Andreïevitch Rimski-Korsakov. *Principles of orchestration*. Dover Publications Inc., 1913 edition, 1964.
- [RM69] J.C Risset and M. V. Matthews. Analysis of musical instrument tones. *Physics Today*, 22(2), 1969.
- [Rod97] X. Rodet. Musical sound signal analysis/synthesis : Sinusoidal + residual and elementary waveform models. In *IEEE Time-Frequency and Time-Scale Workshop 97, Coventry, Grande Bretagne*, 1997.
- [RV07a] Gustavo Reis and Francisco Vega. A novel approach to automatic music transcription using electronic synthesis and genetic algorithms. In *GECCO '07 : Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation*, pages 2915–2922, New York, NY, USA, 2007. ACM.
- [RV07b] Gustavo Reis and Francisco Fernandez Vega. Electronic synthesis using genetic algorithms for automatic music transcription. In *GECCO '07 : Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1959–1966, New York, NY, USA, 2007. ACM.
- [RZR04] Axel Roebel, Miroslav Zivanovic, and Xavier Rodet. Signal decomposition by means of classification of spectral peaks. In *International Computer Music Conference (ICMC)*, pages 446–449, Miami, USA, Novembre 2004.

- [San95] G.J. Sandell. Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. *Music Perception*, 13(2) :209–246, 1995.
- [SdS04] Gary Scavone and Andrey da Silva. Frequency content of breath pressure and implications for use in control. In *NIME '05 : Proceedings of the 2005 conference on New interfaces for musical expression*, pages 93–96, Singapore, Singapore, 2004. National University of Singapore.
- [Sto61] M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32(4) :1339–1342, dec 1961.
- [Sus99] Patrick Susini. *Perception évolutive et globale de sons non stationnaires*. Thèse de doctorat, Ircam (Paris), Paris, 1999.
- [Szy07] S.S. Szyszkowicz. *Interference in Cellular Networks : Sum of Lognormals Modeling*. PhD thesis, Carleton University, 2007.
- [TCA⁺06] Damien Tardieu, Gregoire Carpentier, Gérard Assayag, Xavier Rodet, and Emmanuel Saint-James. Imitative and generative orchestrations using pre-analysed sounds databases. In *SMC*, Marseille, France, 2006.
- [TDvB97] D. Tax, R. Duin, and M. van Breukelen. Comparison between product and mean classifier combination rules. In *Workshop on Statistical Pattern Recognition.*, 1997.
- [Tve77] A. Tversky. Features of similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.
- [Uni] University of Iowa Musical Instrument Samples. theremin.music.uiowa.edu/MIS.html.
- [Vie] Vienna Symphonic Library. www.vsl.co.at.
- [Vin06] Emmanuel Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1) :91–98, Jan 2006.
- [Win68] Robert L. Winkler. The consensus of subjective probability distributions. *Management Science*, 15(2) :B61–B75, oct 1968.
- [YR06] Chunghsin Yeh and Axel Roebel. Adaptive noise level estimation. In *9th Int. Conf. on Digital Audio Effects (DAFx'06)*, Septembre 2006.