Systèmes de Recherche de Documents Musicaux par Chantonnement

Thèse de Doctorat
Ecole Nationale Supérieure des Télécommunications
Spécialité : Signal et Images

Matthieu Carré
matcarre@yahoo.com

Thèse soutenue le 6 juin 2002 devant le jury composé de :

Xavier Rodet (Président)
Régine André-Obrecht (Rapporteur)
Christian Wellekens (Rapporteur)
Marie-José Caraty (Examinateur)
Frédéric Bimbot (Examinateur)
Pierrick Philippe (Examinateur - Directeur technique)
Nicolas Moreau (Examinateur - Directeur administratif)

Je tiens à remercier toutes les personnes qui m'ont accompagné tout au long de ce travail de thèse.

Je remercie Pierrick Philippe et Nicolas Moreau pour leur encadrement à la fois sérieux et amical. L'expérience humaine que nous avons partagée constitue l'une des principales richesses que m'a apportées ce travail de thèse.

Permanents et thésards de France Télécom R&D Rennes ont contribués au bon déroulement de ce travail, tant par leurs connaissances que par leur bonne humeur. Je remercie en particulier Henri Sanson et Vincent Marcatté pour être des chefs à l'écoute de leurs équipes, Françoise Coudray et Nicole Champenois du centre de documentation pour m'avoir fourni les précieuses références nécessaires à ma recherche, et Christophe Apélian pour m'avoir épaulé dès son arrivée dans l'équipe.

Une thèse de doctorat est un travail bien trop personnel pour pouvoir se passer du soutien de ses proches. Je tiens tout particulièrement à remercier Christine pour son soutien sans faille ; David pour m'avoir accueilli à Rennes et accompagné dans sa découverte ; Guillaume, Jérôme, Corentin, et Nicolas pour nos folles improvisations (musicales, mais pas seulement...) ; Peggy, Stéphanie, Cyril, Christian et Tata pour l'intensité de mes retours parisiens ; Eza, Manu et Laurent pour nos heures bordelaises ; Lætitia pour ses encouragements et les désormais mythiques fêtes lannionaises.

Trois années permettent de nombreuses découvertes qui vont bien au-delà du cadre scientifique. Je remercie Ilham pour son enthousiasme à partager, en particulier cette fabuleuse aventure qu'est la passion du chant. Je remercie également Delphine pour son écoute et sa générosité dans cet étrange voyage qu'est l'existence.

Enfin, je remercie mes parents pour m'avoir donné la vie et le reste.

Résumé

Avec l'explosion des données numériques disponibles (notamment via Internet), la question de l'accès aux documents reçoit depuis quelques années une attention accrue. En effet, l'indexation des documents, traditionnellement fondée sur la description textuelle, atteint rapidement ses limites en particulier lorsque le contenu concerné est musical.

Cette thèse focalise sur la recherche de documents musicaux par chantonnement. Nous présentons un système qui permet de retrouver une musique à partir d'un extrait chanté par l'utilisateur. Sa réalisation a nécessité deux études préalables qui ont comblé quelques lacunes d'un domaine de recherche encore jeune. Nous nous sommes intéressés, d'une part, à la "justesse" des mélodies chantonnées (par l'étude de 500 requêtes), et d'autre part, à certains aspects de la similarité mélodique (par la réalisation de tests subjectifs).

Grâce à ces études, nous proposons un système de recherche original et performant. Refusant une description tempérée de la requête (i.e. comportant une quantification des notes au demi-ton), le système proposé retrouve plus de 90% des documents musicaux attendus, pour une taille de requête moyenne (13 notes). La base de données consultée est constituée de 20.000 fichiers MIDI (40 millions de notes indexées). Le temps d'attente est acceptable puisqu'il ne faut que quelques secondes au système pour fournir sa réponse (i.e. la liste des documents les plus similaires à la requête, ceux-ci étant classés par ordre de similarité).

Cette thèse apporte également une aide dans le processus d'évaluation de la qualité de tels systèmes. En effet, nous proposons une modélisation de l'imprécision des mélodies chantonnées. Celle-ci permet la génération de requêtes artificielles qui peuvent être substituées aux requêtes réelles lors du test de systèmes. Cette alternative permet d'alléger le processus de test tout en conservant une stimulation réaliste.

Mots-clés : audio, indexation, moteur de recherche, musique, mélodie, chantonnement, fredonnement, MPEG-7.

Abstract

With the explosion of the digital data available (via Internet particularly), a growing attention has been paid to the means of access to documents. Document indexing, traditionally based on text, provides imprecise descriptions, especially for musical contents.

This thesis focuses on the retrieval of musical documents by humming. We propose a system that performs a search for music initiated from a short excerpt, hummed by the user. This work includes two preliminary studies that forefill some missing elements of a still young research domain. We consider the accuracy of hummed melodies observing 500 queries. We also reveal some aspects of melodic similarity by carrying out subjective tests.

Thanks to these studies, we propose a both original and efficient retrieval system. Refusing a tempered description of the query (i.e. using a quantization of notes into semi-tones), our system retrieves more than 90% of the expected documents from a mean length query (13 notes). The database is made of 20,000 MIDI files (40 millions indexed notes). The search time is acceptable for common use, as the system takes only a few seconds to return the result (i.e. the list of documents that are the most similar to the query).

This thesis also contributes to the quality evaluation process of such systems. We propose a model of the imprecision of hummed melodies. This allows artificial queries to be constructed and substitued to real queries for system testing. This makes the test process easier while preserving a realistic stimulation of tested systems.

Keywords : audio, indexing, search engine, retrieval, music, melody, humming, MPEG-7.

Table des matières

Chapitre 1

Introduction Générale

A l'heure actuelle, nous disposons d'une quantité d'information audio importante et rapidement grandissante par le biais de bases de données publiques (sites Internet, cédéroms, bibliothèques[1]) ou privées (archives INA, SACEM). Ce formidable potentiel n'est malheureusement pas exploité au mieux. En effet, le volume et la richesse des documents disponibles rendent difficile l'accès rapide à l'information recherchée.

L'indexation, ou la gestion de documents en fonction de leur contenu, tente de remédier à ce problème. Les techniques classiques, fondées sur l'annotation textuelle, limitent fortement l'accès aux documents. Pour un contenu musical, le titre, le nom du compositeur ou de l'interprète n'assurent qu'une description éloignée du contenu concerné. De plus, celle-ci demande une intervention humaine qui peut introduire une subjectivité dans la description. Puisqu'il s'agit de musique, langage universel, dépassant les différences linguistiques, l'indexation devrait plutôt se fonder sur la musique elle-même, via l'extraction automatique de descripteurs.

Depuis quelques années, face à ce constat d'une sous exploitation du potentiel informatif disponible, l'indexation multimédia fait l'objet de recherches intenses. Cette tendance s'illustre par la mise en place d'un nouveau standard, MPEG-7, dont la première version a été publiée en septembre 2001.

MPEG (Moving Picture Coding Experts Group, groupe de travail de l'ISO (International Standards Organization), est chargé du développement de standards internationaux consacrés aux documents audiovisuels. Alors que les versions 1, 2 et 4 de MPEG s'attachaient à la transmission des contenus audiovisuels (représentation, compression[2]), MPEG-7 est consacré à leur description [Mar01, QL01].

Baptisé "Multimedia Content Description Interface," MPEG-7 est destiné à décrire les contenus multimédias à des fins de recherche, de filtrage et de consultation. MPEG-7 définit un jeu de descriptions fondées sur les index traditionnels (titre, auteur, droits), sur le contexte (qui ? quoi ? quand ? où ?) et sur les caractéristiques structurelles du contenu audiovisuel. Concernant l'audio, quelques applications-phares ont motivé la mise en place de descriptions spécifiques. Il s'agit de la reconnaissance de flux, de l'identification d'instruments, de la reconnaissance de parole, de la reconnaissance et de l'indexation de sons, et enfin, de la recherche de musique par chantonnement.

Le but de cette thèse est la réalisation d'un système de recherche de documents musicaux

[1]La bibliothèque nationale de France possède 300.000 heures d'enregistrement audio, dont 90% de musique.

[2]dont le fameux MP3, contraction de MPEG-1 Layer 3

destiné au grand public. Cette thèse s'inscrit donc directement dans le contexte que nous venons de décrire, en s'intéressant d'une part à la description de contenus audio musicaux, et d'autre part à son exploitation par un système de recherche (ce deuxième point sortant du cadre de la normalisation MPEG-7).

Afin de nous adresser à une large population d'utilisateurs, la recherche permise par notre système s'effectuera à partir d'une requête chantonnée. La figure 1.1 présente les mécanismes constituant un système de recherche de documents musicaux par chantonnement.

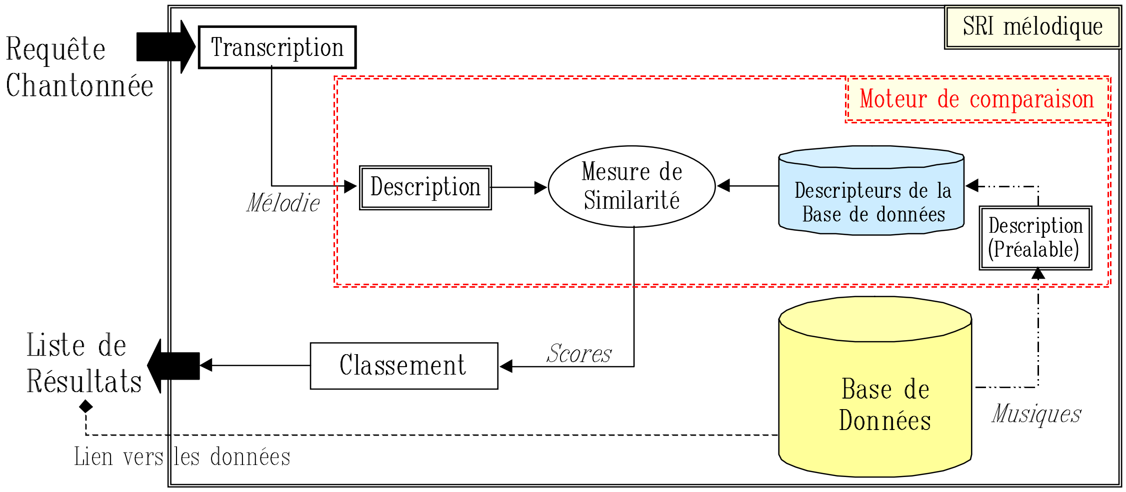


figure 1.1 : *Système de recherche de documents musicaux par chantonnement.*

Lors d'une recherche, la requête soumise par l'utilisateur subit une transcription qui extrait la mélodie chantonnée. Le moteur de comparaison juge alors de la similarité existant entre les musiques de la base et la mélodie recherchée, via leurs descripteurs respectifs. Les scores établis permettent de renvoyer à l'utilisateur une liste de documents ordonnés selon leur ressemblance à la requête. Le moteur de comparaison englobe les descriptions des musiques de la base (descriptions effectuées préalablement au processus de recherche), la description de la requête, et la mesure de similarité.

Ce mémoire de thèse aborde les sujets suivants :

Dans le Chapitre 2, nous traiterons de l'indexation audio musicale en général, afin d'en saisir les enjeux, les principes fondamentaux, et un ordre d'idée des performances actuelles.

Dans le Chapitre 3, nous focaliserons sur le domaine précis de l'indexation mélodique. Un état de l'art détaillé révèlera les lacunes de ce domaine encore jeune.

En réaction à ce constat, le Chapitre 4 viendra améliorer l'état des connaissances sur la requête chantonnée. Celle-ci, particulièrement adaptée aux systèmes de recherche dédiés au grand public, n'a pas encore reçu l'attention méritée.

Le Chapitre 5 décrira le processus détaillé de la conception de quatre moteurs de comparaison mélodique. Composé de descripteurs associés à une mesure de similarité, le moteur de comparaison mélodique constitue le cœur de notre système de recherche.

Dans le Chapitre 6, nous présenterons des tests subjectifs effectués afin d'améliorer les connaissances concernant la similarité mélodique. Ces tests nous permettront de définir des critères objectifs utilisés pour la qualification des moteurs proposés.

Le Chapitre 7 présente l'évaluation de la qualité de système de recherche complets. Les configurations testées illustreront l'intérêt des choix effectués lors de la conception de nos moteurs de comparaison. Ces évaluations permettront également d'élire le meilleur des quatre moteurs proposés, et de le situer par rapport à l'état de l'art.

Enfin, avant d'exposer conclusion et perspectives, nous proposerons, dans le Chapitre 8, un outil facilitant le processus de qualification effectué au Chapitre 7.

Chapitre 2

Indexation Audio Musicale

2.1 Introduction

Par les termes "indexation audio musicale" nous considérons l'indexation des documents audio en excluant ceux présentant des contenus dits "de parole", c'est-à-dire ceux spécifiquement visés par les techniques de reconnaissance automatique de la parole et d'identification du locuteur. Bien avant le récent essor de l'indexation audio musicale, ces techniques faisaient déjà l'objet de recherches intensives. Elles constituent un domaine à part entière. Bien qu'indispensables à un système complet d'indexation audio, nous les écartons du cadre de cet exposé. Par contre, les sons et bruits divers qui habillent fréquemment les enregistrements audio en font partie. D'origines très variées (sons d'ambiances naturelles, bruitages artificiels...), les sons dits "environnementaux" sont difficiles à séparer des contenus de musique. Il serait difficile d'établir une frontière précise séparant les deux domaines. En effet, la musique est un art qui a, depuis plusieurs décennies maintenant, investi des matériaux sonores tels que bruits domestiques, cris d'animaux... Ceux-ci peuvent être considérés comme des contenus relevant de l'indexation audio musicale.

L'indexation de documents consiste en une gestion de ceux-ci selon leur contenu, afin de permettre un accès efficace à l'information qu'ils contiennent. L'accès aux documents audio musicaux peut revêtir différentes formes. Par exemple, on peut envisager la recherche de musique contenant de la trompette, ou bien une mélodie particulière, ou plus généralement, un style musical donné (classique, tango...). La richesse du matériau audio musical disponible engendre de multiples profils de recherche. Dans ce chapitre, nous verrons en quoi consiste l'indexation audio musicale, ainsi que ses différentes déclinaisons selon les applications visées. [Foo99] propose une description technique plus complète de l'indexation audio.

2.2 Principe et intérêt

L'indexation audio musicale a pour but d'améliorer l'accès à l'information sonore musicale. Traditionnellement, l'indexation passe par une description textuelle des documents. Cependant, appliquée aux contenus audio musicaux, l'association de mots clés n'assure qu'une description limitée et généralement éloignée du contenu sonore (titre, compositeur, année...). Une description textuelle du contenu réel est cependant envisageable, par exemple en recensant les instruments utilisés dans un morceau de musique[1], ou bien en qualifiant un son particulier... Malheureusement,

[1] cf. SmarTuner sur http ://www.mzz.com ou version démo sur http ://www.lecielestbleu.com/html/smartuner.htm

compte tenu du volume de données disponibles et à venir, cette description implique un investissement humain dissuasif. Par ailleurs, cette option peut engendrer des descriptions trop subjectives pour être efficaces. Prenons l'exemple d'un son particulier. Le qualifier textuellement est une tâche ardue. En effet, des termes tels que "agressif" ou "chaleureux" pourront être employés, mais ils évoqueront des sonorités différentes selon les personnes. Une description précise sera alors délicate. L'indexation audio musicale a donc besoin d'outils (semi-)automatiques permettant une description objective du contenu sonore, afin d'assurer une gestion efficace des données.

La question de la description se pose en terme d'extraction automatique d'information représentative du contenu sonore. Les descripteurs obtenus permettent de manipuler les documents selon leur contenu, en fonction de l'application visée. Il existe nombre d'applications à l'indexation audio musicale : il y a bien sûr la consultation d'"audiothèques" où, même lorsqu'il est connu, le nom du compositeur ou de l'interprète ne suffit pas toujours à accéder rapidement au document désiré (par exemple, comment savoir lequel des 80 albums de Frank Zappa contient l'air qui nous intéresse ? !). Même si le titre de l'œuvre est connu, retrouver un passage désiré peut être délicat si la durée du document est importante (concert, opéra). A la différence de la vidéo où l'utilisateur peut disposer d'images clés, la recherche d'un passage dans un document audio implique encore une écoute exhaustive de celui-ci. L'indexation audio musicale tend à permettre une consultation *non-linéaire* des documents musicaux, afin de réduire le temps de recherche. D'autres domaines sont concernés : le commerce de la musique sur Internet (recherche de contenu), la vérification de droits d'auteurs (reconnaissance de contenu), l'aide à la post-production de films et à la composition de musique électronique (recherche de sons particuliers). L'indexation du contenu audio permet également d'envisager un écrémage des documents télé et radio diffusés : on peut imaginer un poste radio qui change de station dès qu'il n'y a plus de musique, ou qui recherche un contenu précis comme des concerts ou de la musique techno. Avant de détailler les applications possibles, arrêtons-nous sur les qualités attendues d'un système d'indexation.

Imaginons un système d'indexation idéal. La réponse qu'il fournit à une recherche doit être à la fois *rapide et précise*. Il doit proposer des documents pertinents vis-à-vis de la requête formulée, et ce, quelle qu'elle soit. Cela implique tout d'abord qu'idéalement, tous les types de requête sont permis. L'utilisateur peut donc choisir la manière dont il soumet sa requête au système : texte, son, image, vidéo, bref, n'importe quelle information multimédia. Ensuite, indépendamment de la nature de la requête, l'angle de recherche que peut adopter l'utilisateur est également libre. On peut donc concevoir une requête du type : "Je souhaite trouver un fichier sonore au relief aussi prononcé que cette image". La nature de la requête est multimédia (texte+image), et l'angle de recherche se fonde sur une notion de "relief" pour obtenir un document audio. Si notre système d'indexation idéal renvoie une réponse adaptée à chaque requête, cela implique deux hypothèses exclusives l'une de l'autre :
– soit, tous les angles de recherche possibles ont été anticipés, l'utilisateur s'inscrivant dans l'un d'entre eux ;
– soit, le système de recherche est capable d'interpréter le désir de l'utilisateur. Ainsi, de manière autonome, il procède à une description des contenus disponibles conformément aux attentes de l'utilisateur.

Il est évident que les contraintes imposées par la réalité rendent un tel système impossible. L'état des outils d'analyse n'autorise pas encore tous les types de requêtes, et il est illusoire d'espérer prévoir tous les angles de recherche. Enfin, si la réponse d'un système passait par une nouvelle description des contenus disponibles, le temps de réponse résultant serait rédhibitoire. Dans un système d'indexation, la description est préalable à la recherche. Consulter les descripteurs et non le contenu lui-même permet -entre autres- de réduire le temps de réponse.

Une autre vertu est également attendue du système d'indexation idéal, il s'agit de sa *tolérance*. Un système doit être robuste aux perturbations qui séparent requête et données recherchées. Par exemple, se tromper sur une note ne doit pas condamner la recherche d'une mélodie. Cette attente étant antagoniste avec le désir de précision évoqué plus tôt, on voit se dégager un *compromis précision/tolérance*, sous la contrainte de *rapidité*. L'existence de ce compromis montre que, par essence, il n'existe pas de système d'indexation idéal.

La multiplicité des profils de recherche et les attentes en matière de performances amènent à réduire le domaine d'application d'un système d'indexation. Celui-ci se spécialise : le compromis précision/tolérance sous contrainte de rapidité est géré pour la recherche de certains contenus, et pour certains types de requête seulement.

Nous allons maintenant détailler les différentes facettes qui constituent l'indexation audio musicale :
- l'extraction de descripteurs,
- l'étiquetage de contenus,
- la recherche et la consultation de documents audio musicaux.

Pour chacune d'elles nous verrons les spécificités qu'entraînent les applications possibles.

2.3 Extraction de descripteurs

L'extraction de descripteurs du contenu audio consiste à sélectionner l'information représentative du contenu. Cette action commence par le découpage temporel du signal en portions dont sont extraites des caractéristiques pertinentes. Celles-ci doivent révéler la spécificité des différents contenus en éliminant l'information non discriminante. Le choix des descripteurs est bien sûr lié à l'application visée. Par exemple, une recherche de musique contenant de la trompette n'utilisera pas les mêmes descripteurs qu'une recherche de mélodie particulière. Dans le premier cas, c'est l'instrument qui compte (quelle que soit la mélodie), dans le deuxième, ce sont les notes jouées (quels que soient les instruments).

Le choix des descripteurs peut être guidé par des observations auditives. Par exemple, si l'on souhaite séparer la parole de la musique, on essaiera de révéler les spécificités audibles de chacun. Dans les signaux de parole, on peut percevoir une alternance voisée/non-voisée caractéristique. Les sons voisés ont une hauteur perceptible, ce qui se traduit par un spectre harmonique. Ils correspondent à la prononciation de sons tels que "a", "i", "ou", "en", mais aussi "m", "n", "g", "b". Les sons non-voisés sont plutôt assimilables à du bruit (spectre inharmonique, riche en hautes fréquences). Ils correspondent à la prononciation de "f", "ch", "s", "p", "t"... Un descripteur de la répartition en fréquence du spectre du signal permettra, entre autres, de distinguer les contenus de parole des contenus de musique, dont les propriétés sont différentes.

La figure 2.1 illustre le comportement d'une des caractéristiques couramment utilisées dans la discrimination Parole/Musique. Le taux de passage à zéro de la forme d'onde temporelle (*Zero Crossing Rate* ou *ZCR*) renseigne sur l'étalement fréquentiel d'un spectre. Appliquée à un signal de parole (première partie du signal), cette mesure révèle l'alternance voisée/non-voisée typique des signaux de parole. Cette alternance est traduite par les brusques variations d'un profil temporel par ailleurs plutôt faible. Par contre, la musique (deuxième partie du signal), globalement plus tonale (ou harmonique), possède un taux présentant moins de variations et une moyenne plus élevée.

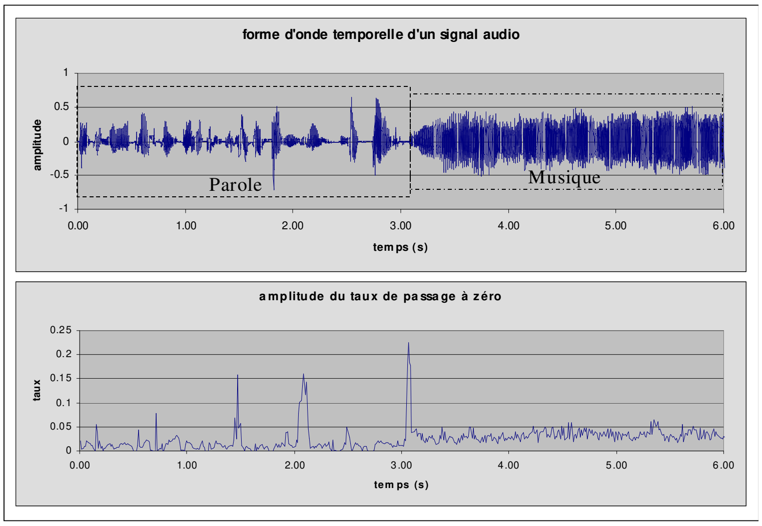


figure 2.1 : forme d'onde temporelle d'un signal audio, amplitude du taux de passage à zéro

Le taux de passage à zéro peut engendrer différents descripteurs représentatifs du contenu spectral : l'amplitude du *ZCR*, la variance de sa dérivée, son moment du 3ème ordre, son dépassement par rapport à un seuil, sa différence de répartition de part et d'autre de la moyenne [Sau96, SS97, ZK98].

Un descripteur efficace ne correspond pas forcément à un trait acoustique. Cela signifie qu'il n'illustre pas une caractéristique précise perceptible à l'oreille. Cependant, ses propriétés discriminantes en font une illustration efficace du contenu. A titre d'exemple, on peut citer les coefficients cepstraux pondérés selon les propriétés perceptives des bandes de fréquences (*Mel Frequency Cepstral Coefficient* ou *MFCC*). Leur utilisation vient du domaine de la parole (reconnaissance de parole, identification de locuteur) où ils ont fait leurs preuves [Foo97].

Pour un exposé plus complet des descripteurs usités ainsi que de leur manipulation, le lecteur pourra consulter l'état de l'art en indexation audio (musicale) présenté par l'auteur de cette thèse dans [CP00].

Nous l'avons évoqué en section 2.2, les descripteurs extraits de contenus sont destinés à rechercher ces derniers. Mais ils peuvent également servir à les classifier. Dans la section suivante, nous allons nous intéresser à l'opération d'étiquetage.

2.4 Etiquetage de contenus

L'étiquetage est l'identification de contenus constituant un document. Un étiquetage consiste par exemple à distinguer les passages de parole des passages de musique, ou encore, à détecter la présence d'un style de musique particulier. Cette application permet d'améliorer la consultation de documents audio en associant aux documents des annotations directement reliées à leur contenu. Dans [SW97] par exemple, une interface graphique permet de naviguer dans un morceau de jazz en visualisant les instruments qui y interviennent. Au delà d'une aide à la consultation des documents audio, l'identification de contenus permet l'application de traitements adaptés à la nature de chacun d'eux. Pour une indexation de flux audio radiophonique où alternent musique et commentaires, l'identification permettra, par exemple, l'utilisation d'un processus de reconnaissance musicale (quelle musique est diffusée ?) d'une part et une reconnaissance automatique de parole (quel discours l'animateur tient-il ?) d'autre part.

L'étiquetage est assuré grâce aux descripteurs extraits du contenu étudié. Ceux-ci sont comparés à des références représentatives des contenus discriminés (parole, musique ou encore jazz, rock, classique). Ces références, qu'elles soient seuils, modèles statistiques ou autres, sont issues d'un entraînement. Celui-ci est assuré grâce à un corpus de données préalablement étiquetées (ce qui nécessite généralement une intervention humaine). Cette opération d'entraînement, appelée apprentissage, est indispensable si l'on veut pouvoir effectuer l'opération "inverse", c'est-à-dire présenter un segment dont le contenu est inconnu et obtenir en retour l'étiquette correspondante (i.e. le type de contenu).

L'étiquetage d'un flux audio s'effectue conjointement à sa segmentation. En effet, la détermination de zones de contenu homogène implique le découpage du flux considéré. Cette double action de découpage/étiquetage se fonde généralement sur l'observation de plusieurs descripteurs. Ceux-ci proposent des instants de segmentation qui correspondent aux variations significatives des caractéristiques qu'ils représentent. La délimitation d'une zone homogène relève d'une décision fondée sur les instants de découpage proposés par les différents descripteurs.

Une segmentation peut cependant être effectuée sans identification. Ainsi, Foote propose une mesure du degré de nouveauté d'un flux audio [Foo00]. Le principe est de comparer le signal à sa modélisation afin de révéler ses changements significatifs. La nature des changements détectés dépend de l'horizon temporel de la modélisation. Le type de segmentation est ajustable à l'application visée. Du plus précis au plus général, on pourrait citer comme exemple la segmentation d'une musique en notes, ou bien en parties (type *couplet/refrain*), ou encore la discrimination *Parole/Musique*... La méthode, qui ne s'appuie pas sur des traits acoustiques, est annoncée comme dispensée d'apprentissage. Cependant, il faut noter que deux choix non triviaux sont à déterminer pour la mise en œuvre d'une application donnée : celui d'un seuil de nouveauté et celui d'une résolution temporelle adéquate.

On observe une partition des recherches concernant l'étiquetage des contenus. Deux secteurs se dessinent en fonction du type d'étiquette attribué. D'un coté, l'étiquetage en catégories générales concerne des contenus tels que parole et musique par exemple. De l'autre, l'étiquetage en catégories fines identifie des contenus plus précis comme, par exemple, un instrument particulier.

2.4.1 Etiquetage en catégories générales

Des systèmes d'étiquetage temps-réel du contenu ont été présentés dans différentes publications. Nous abordons d'abord celles dont les catégories distinguées sont générales : discrimination *Musique/Parole&publicité* [Sau96], *Parole/Musique* [SS97]. Notons que lorsque la musique est opposée à la parole (typiquement, dans la discrimination *Parole/Musique*), il s'agit généralement de musique instrumentale, c'est-à-dire sans intervention vocale. En effet, la voix parlée et la voix chantée présentent des similitudes qui rendent leur dissociation difficile. Or les contenus hybrides tels que des mélanges musique+voix sont fréquents : chansons, commentaires sur fond musical (radio, publicité...), musique rap...

Zhang et Kuo, tout en s'attachant aux différences entre voix parlée et voix chantée, prennent en compte leur contexte d'apparition. Ainsi, ils consacrent des catégories spécifiques aux principaux mélanges recensés [ZK99]. Les types audio discriminés sont les suivants :
– silence
– musique "pure" (i.e. instrumentale)
– chanson
– parole "pure" (i.e. seule)
– parole sur fond musical
– sons environnementaux
– sons environnementaux sur fond musical.

La catégorie des sons environnementaux correspond aux sons et bruits divers qui habillent fréquemment les enregistrements audio. D'origines très variées (sons d'ambiance naturelles, bruitages artificiels...), ils sont -dans le système de Zhang- sous-classifiés selon leurs propriétés acoustiques (harmonicité, périodicité et stabilité). Pratiquement, la catégorie des sons environnementaux recueille au final, les contenus jusqu'alors non identifiés. Les expériences de Zhang et Kuo sont significatives dans ce domaine de l'étiquetage en catégories générales. Le taux d'erreur, correspondant à la proportion des segments mal étiquetés, est annoncé inférieur à 10%. Or, jusqu'ici, les performances diminuaient avec l'augmentation du nombre de catégories discriminées (d'autant plus lorsqu'il s'agissait de catégories hybrides, mélange de catégories de base). A titre d'exemple, Scheirer et al. passèrent de 10 à 35% d'erreur en ajoutant une troisième classe *Parole et Musique simultanés* à leur système de discrimination *Parole/Musique* [SS97]].

Cependant, les résultats de telles expériences sont fortement liés à la qualité des données utilisées pour les phases d'apprentissage et de test. Cette dépendance implique une confiance relative vis-à-vis des systèmes actuels.

Premier maillon d'une chaîne d'analyses successives (par exemple, détection de parole puis transcription du texte prononcé), l'identification de types généraux doit présenter une bonne robustesse aux fluctuations de la qualité des données d'entrée. En effet, la "qualité" de séquences audio quelconques est extrêmement variable, tant au niveau du son (bande passante, niveau de bruit), qu'au niveau du contenu (mixture de différentes classes). Les techniques d'étiquetage en catégories générales sont d'ores et déjà utilisées, notamment comme aide à l'indexation vidéo (utilisation de la bande son) [PHMW98, ZK01]. On peut citer comme exemple le système d'archivage vidéo Virage[2] qui utilise le discriminateur Parole/Musique de MuscleFish[3] [WBKW99].

2.4.2 Etiquetage en catégories fines

Parallèlement à l'amélioration de l'étiquetage de contenus généraux, des travaux sont menés sur un étiquetage plus précis de certains contenus audio. Des systèmes plus spécialisés traitent des signaux dont la catégorie générale (*Musique*, par exemple) est supposée connue. Ces systèmes spécialisés sont complémentaires aux systèmes d'étiquetage généraux. Dans un système complet traitant de contenus quelconques, l'étiquetage en catégories fines succèderait à un étiquetage grossier des contenus. Par exemple, [SS97]] présente un descripteur soulignant le rythme prononcé et régulier de musiques telles que la salsa, la techno, certains types de rock, etc. De tels descripteurs

[2]http ://www.virage.com
[3]une démonstration interactive de ce dernier est disponible sur http ://www.musclefish.com

pourraient permettre une discrimination au sein des contenus préalablement étiquetés *Musique*. Une imbrication de telles techniques permettrait, par exemple, la distribution personnalisée de musique [AT00, CV00] (1er stade : sélection des contenus musicaux ; 2ème stade : sélection des styles musicaux désirés).

Pour le moment, l'amélioration des étages d'un système complet s'effectue de manière indépendante. Concernant les systèmes d'étiquetage fin, les pré-requis imposés aux données d'entrée peuvent aller au delà d'une connaissance a priori du type général (Parole, Musique...). En effet, ces systèmes spécialisés peuvent attendre des contenus présentés qu'ils soient pré-segmentés. Typiquement, les données d'entrée seront des segments sonores isolés. Dans ce cas, la tâche n'est plus de délimiter des zones de contenu homogène à identifier, mais uniquement d'étiqueter la portion présentée.

MuscleFish propose un système d'étiquetage de sons isolés dans lequel 16 variétés sont discriminées [WBKW99]]. On y retrouve des sons de cloches, des cris d'animaux, des bruits de foules, des rires, différents types instruments (cuivres, cordes frottées, percussions...), des bruits d'eau, des voix parlées. D'autres expériences ont été menées sur la discrimination de ces mêmes 16 catégories, avec un taux d'erreur ramené à moins de 10% [Li00]. Plus spécifique (dans l'application), une identification de sons issus d'une trentaine d'instruments de musique est réalisée avec 20% d'échec dans [EK00]. Comme nous l'avons dit plus tôt, lorsque le nombre de catégories discriminées diminue, les performances sont généralement meilleures. Ainsi, Eronen et *al.* obtiennent un taux d'erreur de 5% lorsqu'ils considèrent simplement la famille des instruments (cuivres, corde pincées, cordes frottées...).

Les performances des systèmes d'étiquetage fins et généraux s'améliorent significativement, mais les conditions d'utilisation qu'ils imposent les écartent encore d'un système d'étiquetage de documents audio musicaux quelconques.

Nous allons maintenant aborder une autre action qui tire profit de l'extraction de descripteurs : il s'agit de la recherche de contenus.

2.5 Recherche de contenus

Les systèmes de recherche de contenus constituent une autre facette de l'indexation. La tâche est de fournir à un utilisateur l'information qu'il désire. Cependant, les moyens dont dispose cet utilisateur pour spécifier ses attentes doivent être adaptés au type d'information qu'il recherche. Pour l'audio musical, deux possibilités, non exclusives, se dégagent. La première, traditionnelle, est la requête textuelle, la deuxième, plus spécifique, est la requête dite "par l'exemple".

2.5.1 Types de requête

La requête textuelle

Nous l'avons vu, l'étiquetage automatique de contenus sonores permet d'annoter les documents disponibles. Si les caractéristiques discriminantes correspondent à des traits perceptibles (style musical, instrument...), une description textuelle du contenu sonore recherché est envisageable. Seulement, s'il est le langage habituel des moteurs de recherche, le texte atteint ses limites lorsqu'il s'agit de décrire un son. En effet, des termes tels que "brillant", "chaleureux" ou "agressif" peuvent évoquer des sons différents selon les sujets. Il est donc difficile de s'accorder sur les termes à employer et ce qu'ils désignent précisément. Une adaptation du système au vocabulaire de chaque utilisateur est envisageable, mais celle-ci implique une phase d'apprentissage (à l'instar des systèmes de dictée pour traitement de texte). La requête textuelle est donc utile mais insuffisante pour la recherche de contenus audio musicaux.

La requête "par l'exemple"

Elle désigne une requête de même nature que l'information recherchée, dans notre cas : du son. Cette voie permet d'appliquer à la requête un traitement identique à celui des données indexées. Les mêmes descripteurs peuvent donc être extraits d'une requête facilitant ainsi sa comparaison avec les contenus disponibles. La requête "par l'exemple" offre une alternative de qualité à la requête textuelle. L'utilisateur peut présenter une requête audio issue des documents dont il dispose (extrait de bande-son, par exemple), ou bien produire lui-même l'exemple à soumettre au système de recherche (onomatopées, fredonnement).

La requête est le moyen de spécifier l'objet d'une recherche. Comme nous allons le voir dans la sous-section suivante, la recherche de contenus possède un cas particulier : la reconnaissance.

2.5.2 Reconnaissance de contenus

Les systèmes de reconnaissance sont des systèmes de recherche particuliers. La réponse attendue est unique : le document reconnu. La requête est une portion de contenu, éventuellement dégradé, présent dans la base de données (d'où le terme *re-connaissance*). Il s'agit donc d'une requête par l'exemple.

Une application phare de l'indexation audio musicale est la reconnaissance automatique de musique. Mobiquid propose un service de reconnaissance des programmes musicaux de certaines radios françaises. Le client utilise son téléphone mobile pour transmettre une musique radiodiffusée inconnue, et le système lui envoie un message contenant le titre et l'interprète de la chanson, tout en lui permettant d'acquérir le produit[4]. Le même type d'application est proposé par "Audible Magic Corporation"[5]. Le service appelé "Clango" s'applique à la musique diffusée sur Internet par certaines stations radio américaines[6]. Dans ces deux systèmes de recherche, la requête est une portion d'un matériau musical issu du disque d'un artiste. La reconnaissance doit résister aux différentes dégradations possibles du signal (ajout de bruit, réduction de bande passante...) qu'engendrent sa diffusion (vers l'auditeur) et sa retransmission (vers le système).

[4]http ://www.mobiquid.com
[5]possesseur de MuscleFish depuis octobre 2000
[6]http ://www.audiblemagic.com

figure 2.2 : interface "clango" après la phase d'analyse

A notre connaissance, peu de publications sont disponibles sur le sujet. La plus significative est [PRF$^+$01] dans laquelle plus de 86% des extraits musicaux présentés sont identifiés parmi 350 enregistrements, en un temps moyen d'une minute. Les portions candidates à la reconnaissance ne doivent cependant pas dépasser un taux de distorsion "moyen" par rapport aux extraits originaux.

Sans aller jusqu'à la réalisation d'un système complet, Wold et *al.* avaient proposé un descripteur utilisable dans ce genre d'application [WBKW99]. Pikrakis et *al.* ont également mené des travaux sur la reconnaissance de musique, concernant des motifs de clarinette jouée suivant la tradition musicale grecque [PTK98].

Notons que les descriptions utilisées pour la reconnaissance se veulent simplement suffisantes pour permettre une comparaison efficace des séquences traitées. L'analyse mise en œuvre ne correspond pas à une analyse musicale, telle qu'elle intéresserait les musicologues, par exemple.

Dans les applications de reconnaissance, on cherche à retrouver un élément unique parmi un ensemble important. Ce type d'application n'est pas destiné à associer deux interprétations différentes d'une même chanson. Leur vocation n'est donc pas, par exemple, de retrouver une chanson donnée en fournissant comme requête la version d'un autre artiste. Un extrait de "Mon manège à moi" d'Edith Piaf ne permettrait pas de retrouver la reprise chantée par Etienne Daho.

Les systèmes de reconnaissance sont un cas particulier des systèmes de recherche plus traditionnels qui retournent généralement une liste de documents similaires à la requête. La reconnaissance consiste en une limitation de leurs possibilités :

– seule la requête par l'exemple est permise
– la réponse du système se limite à un seul élément (généralement au premier document de la liste des réponses. Celui-ci étant le plus proche de la requête, il correspond le plus probablement à celui dont est issue la portion à reconnaître).

2.5.3 Moteurs de recherche

Les systèmes de recherche retournent généralement plusieurs documents classés par ordre de similarité à la requête. Comme nous l'avons déjà vu, dans le domaine audio musical, la similarité peut se fonder sur une requête textuelle. Par exemple, la réponse à une recherche de sons de flûte sera une liste de documents comportant l'étiquette "flûte". Cependant, si la démarche consiste à rechercher des sons qui ne proviennent pas d'une flûte mais qui y ressemblent fortement, l'étiquetage ne suffit plus. Dans ce cas, la requête par l'exemple prend le relais de la requête textuelle, et compare les contenus, non plus grâce à leurs étiquettes, mais directement via leurs descripteurs. Par exemple, après avoir trouvé un son de flûte grâce à une requête textuelle, l'utilisateur peut réinjecter l'extrait sonore obtenu pour obtenir des sons similaires. La figure 2.3 illustre cet exemple effectué grâce au moteur de recherche "FindSounds" de Comparisonics[7].

La fenêtre "Search for" a préalablement permis la recherche de sons grâce au mot "flute". Puis l'option "Find Sounds like this one" a fait du son de flûte "F7.B.au" une nouvelle requête. La figure 2.3 présente les premiers éléments de la liste des réponses obtenue. On retrouve le même extrait sonore ("F7.B.au") au début de la liste[8]. En deuxième position, la source est différente puisqu'il s'agit d'un son d'oiseau : une grive. Présentant une bonne similarité avec la requête (confirmée par l'écoute), celui-ci pourra, par exemple, être utilisé dans une composition musicale. Il pourra se substituer au son habituel de la flûte, apportant ainsi une couleur originale à la musique.

[7] http ://www.comparisonics.com

[8] La similarité de 99% indique qu'il est très proche de la requête, sans pour autant être identique. Ce très faible écart s'explique sans doute par la différence de précision de représentation des deux extraits sonores (16bits contre les 8bits de la requête)

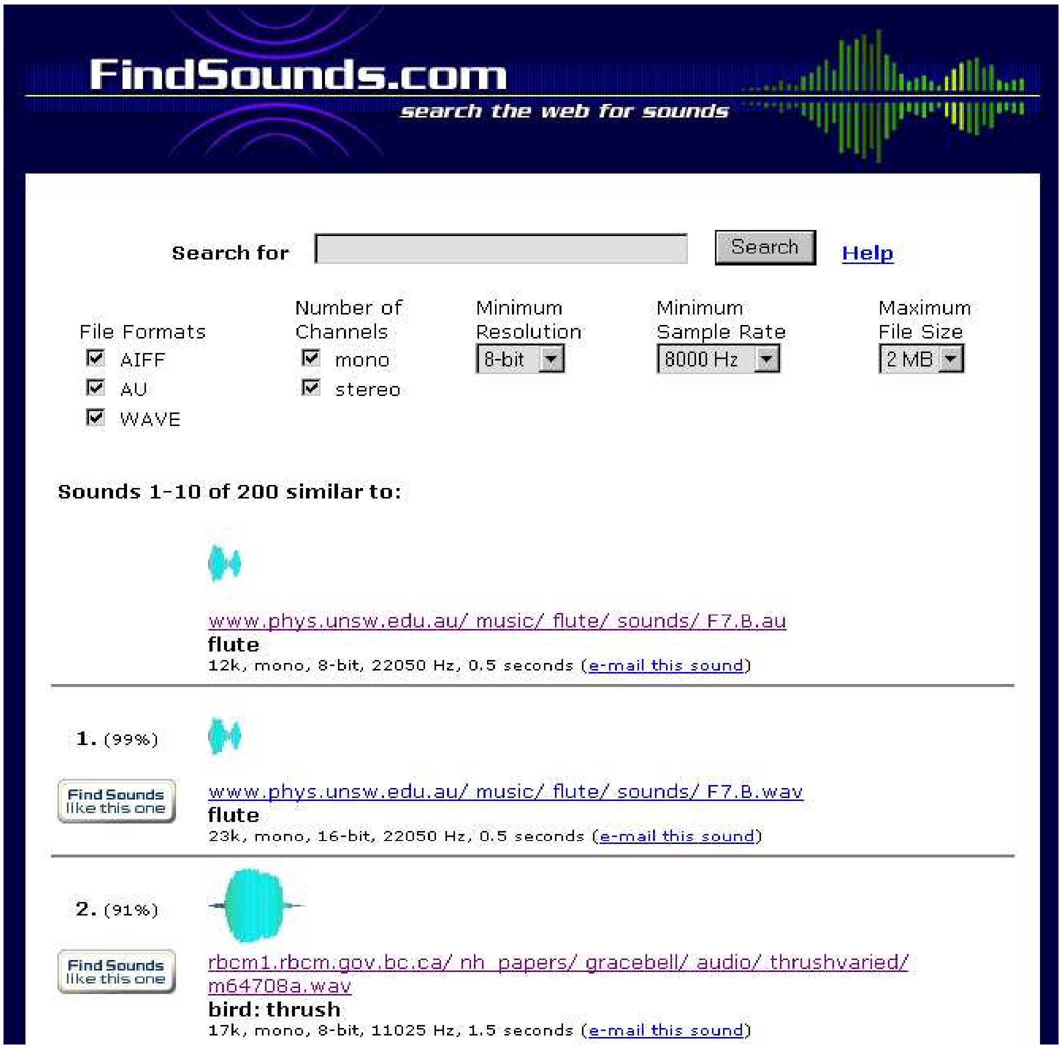


figure 2.3 : interface de FindSounds

D'autres systèmes de recherche sont proposés. Un moteur de recherche de sons par "critère psychoacoustique de similarité" est intégré au "Studio Online" de l'Ircam[9]. 125.000 sons de seize instruments enregistrés par des solistes réputés y sont disponibles. Une démonstration du système MuscleFish de recherche par le contenu est disponible sur leur site[10]. 400 sons appartenant aux 16 variétés citées plus tôt (cf. 2.4.2) peuvent être ordonnés selon leur similarité à l'un d'entre eux. Par ailleurs, MuscleFish développe un système flexible de gestion d'extraits sonores. Nommé "SoundFisher", ce système est dédié à l'organisation, la recherche et la consultation de contenus audio, ainsi qu'à leur transmission à d'autres applications. La Figure 2.4 présente un exemple d'interface du programme.

[9] http ://www.ircam.fr/produits/technologies/sol/
[10] http ://www.musclefish.com/cbrdemo.html

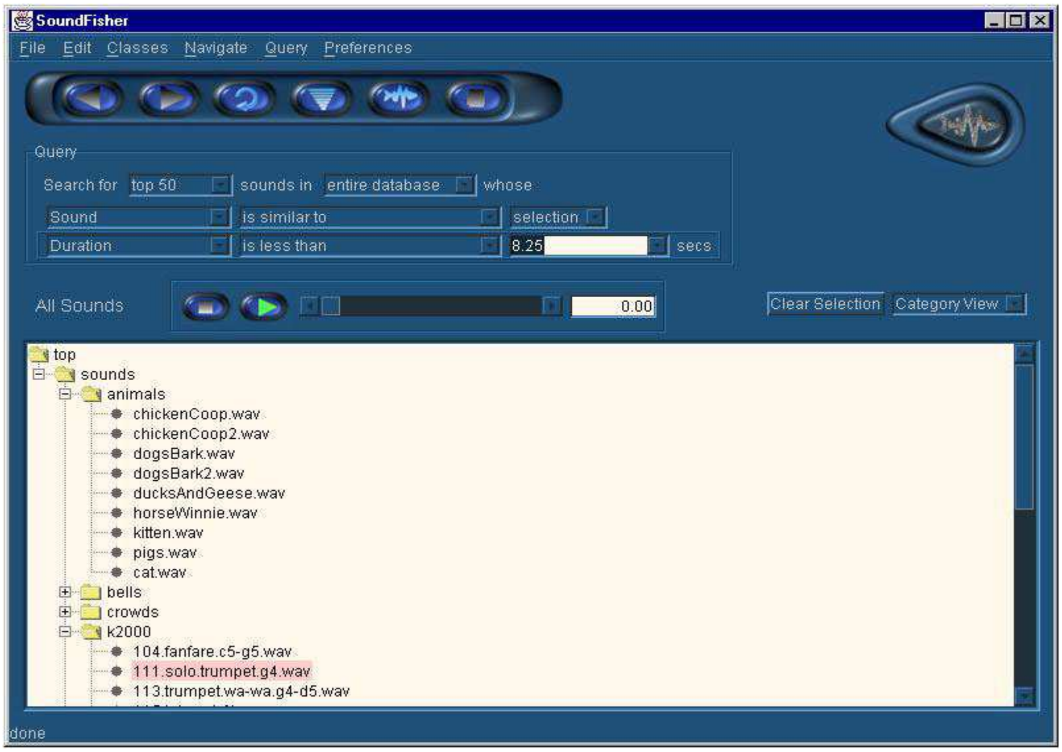


figure 2.4 : interface de SoundFisher

Les descripteurs mis en jeu dans les systèmes de recherche que nous venons de voir tentent souvent de représenter le *timbre* des sons. Ce dernier peut être défini comme la qualité permettant de faire la différence entre deux sons de même hauteur, de même puissance et de durée identique[11] Le timbre nous permet donc de différencier un violon d'une trompette, même si ceux-ci jouent la même musique. Difficile à modéliser, il représente un excellent facteur pour la discrimination des sons. Cependant, cette caractéristique forte passe au second plan lorsqu'un autre aspect du contenu audio musical est considéré. Cet aspect est la *mélodie* qui constitue un élément fondamental du matériau musical.

2.5.4 La mélodie : un contenu "haut-niveau"

Comme nous l'avons vu dans la discrimination *Parole/Musique*, une manière de décrire le matériau audio musical est de considérer la manière dont il a été produit. Par exemple, il est évident que les instruments utilisés dans une musique influencent directement le contenu audio final. Une description de la musique peut donc consister à signaler les instruments qui y participent. Une autre information très importante de la création d'une musique peut être représentée par la notion de partition. Description visuelle d'une musique, la partition spécifie notamment les notes que chaque instrument exécute.

L'intérêt principal de fonder une description sur la partition musicale est la manipulation d'une musique indépendamment de sa réalisation. On s'intéresse ici à un autre aspect du processus de

[11]Les hauteur, puissance et durée évoquées ici sont en réalité des caractéristiques perçues. En effet, deux sons émis avec une même hauteur peuvent ne pas être perçus comme tels. Il en est de même avec la puissance et la durée.

production du contenu musical : au lieu de focaliser sur les conséquences de l'arrangement adopté (en particulier les instruments utilisés), cet angle met l'accent sur les mélodies jouées.

Contrairement à la reconnaissance musicale évoquée plus tôt (cf. 2.5.2), cet angle de description permet de retrouver simultanément les deux versions d'une même chanson. Quelques notes de la mélodie de "Mon manège à moi", par exemple, permettront de retrouver aussi bien la version d'Edith Piaf que celle d'Etienne Daho. D'aspect final différent, les deux documents audio présentent une indiscutable similitude puisque c'est la même chanson qui est interprétée. Cette similitude est mélodique. Nous la qualifions de "haut-niveau" car elle s'appuie sur l'*organisation* dans le temps de certaines caractéristiques de base des sons (typiquement, hauteur et durée des notes qui constituent la mélodie).

L'indexation mélodique s'oppose en quelque sorte à la reconnaissance musicale évoquée plus tôt. Ici, nul besoin d'une portion issue de l'information recherchée pour soumettre sa requête (e.g. l'extrait d'un CD retransmis à la radio). La requête peut revêtir différents aspects tant que l'information mélodique demeure (peu importent l'instrument, le tempo, la tonalité). Deux types de requêtes par l'exemple sont possibles. Dans le sens où la musique dispose d'un langage spécifique pour son écriture, une requête "graphique" (pour ne pas dire "textuelle") constitue une requête par l'exemple. Cependant, la population des utilisateurs connaissant le solfège est réduite (une partie des musiciens, les musicologues...). La musique étant avant tout une information acoustique, la requête par l'exemple *audio* apparaît la plus naturelle. Elle constitue l'alternative qui élargit l'accès à l'information musicale. En effet, elle permet à chacun de rechercher une musique, par exemple, en fredonnant quelques notes. Chanter pour retrouver un air de trompette est possible puisque l'information mélodique est conservée. Comme pour la recherche d'extraits sonores vue précédemment, la requête doit simplement permettre l'extraction des descripteurs qui permettront la comparaison avec les données de la base.

Ici encore, la requête par l'exemple offre un meilleur accès à l'information musicale par une alternative de qualité à la requête textuelle. Le principal problème de ce type d'indexation est l'accès à la partition de documents audio musicaux quelconques. Cette limitation a pour conséquence de contraindre fortement les données concernées par ce type d'indexation.

L'indexation mélodique est le domaine dans lequel s'inscrit cette thèse. Le chapitre suivant l'aborde de manière détaillée. Nous y verrons les nombreux aspects de la mise en œuvre de systèmes d'indexation mélodique.

2.6 Conclusion

L'indexation audio musicale peut revêtir plusieurs formes, toutes fondées sur l'extraction de descripteurs représentatifs du contenu. L'étiquetage (plus ou moins grossier) consiste à l'identification de contenu provenant de catégories telles que "parole", "musique", "rires", "violon", "bruits d'eau"... Les repères que fournit l'étiquetage permettent d'améliorer la consultation des documents audio (éviter l'écoute exhaustive en accédant rapidement à l'information voulue). De plus, l'étiquetage permet l'adaptation de traitements ultérieurs à la nature des matériaux sonores (par exemple, compléter l'indexation en opérant une transcription textuelle lorsqu'un contenu de parole est identifié).

L'indexation englobe également la recherche de contenus (avec un cas particulier : la reconnaissance). Différents angles de recherche se dessinent suivant les différentes approches du matériau audio musical. On peut, par exemple, rechercher des musiques contenant de la trompette (quel que soit l'air joué), ou bien des musiques contenant une mélodie donnée (quel que soit l'instrument qui l'exécute). Dans les moyens disponibles pour la recherche de contenus, la requête "par l'exemple" offre une alternative de qualité à la requête textuelle. Elle désigne une requête de même nature que l'information recherchée, facilitant ainsi la localisation de contenus similaires. L'utilisateur peut, par exemple, présenter un extrait de bande-son pour retrouver le film ou obtenir des sons similaires, ou encore fredonner l'air qu'il recherche.

L'indexation audio musicale est un domaine jeune qui, compte-tenu des enjeux économiques, fait l'objet d'une attention croissante. Comme l'illustre l'émergence de produits commercialisées (Mobiquid, AudibleMagic, Comparisonics), certaines techniques présentent d'ores et déjà une bonne efficacité (notamment grâce à l'héritage de domaines plus anciens tels que la reconnaissance automatique de parole ou l'identification de locuteurs).

Cependant, constitué d'une imbrication de technologies diverses, le système complet d'indexation audio musicale (permettant d'étiqueter un matériau sonore quelconque et/ou d'en rechercher le contenu) est une chaîne limitée par ses maillons les plus faibles. L'incontournable compromis précision/tolérance de tout système d'indexation est encore difficile à gérer de manière satisfaisante.

Au stade actuel, les techniques disponibles ne sont pas assez robustes pour envisager des systèmes décrivant de manière complète et séquentielle des contenus quelconques (description de la plus grossière à la plus fine). Les recherches focalisent donc sur l'amélioration de tel ou tel maillon de la chaîne, souvent extrait pour être appliqué à une tâche spécifique. En contrepartie, la part d'intervention humaine peut demeurer importante. Par exemple, l'étiquetage en catégories fines (autres que *Parole*, *Musique*, *Parole sur fond musical*...) implique un conditionnement strict des contenus candidats (présegmentation, qualités particulières du matériau audio...).

Néanmoins, à l'image des techniques audio utilisées comme aide à la segmentation vidéo [PHMW98, LMB97], l'association d'outils s'adressant à des contenus divers marque les premiers pas d'une réelle indexation multimédia.

Ce travail de thèse n'échappe pas au contexte actuel de la recherche en indexation. Il focalise sur un domaine d'application précis, destiné à terme à s'insérer dans un système d'indexation de contenus (audio) quelconques. Cette thèse s'inscrit dans le domaine de l'indexation mélodique, en proposant un système de recherche de musique dont la requête s'effectue par chantonnement.

Dans le chapitre suivant, nous allons décrire l'état de l'art de ce domaine afin de fournir les informations nécessaires à la compréhension de ce travail de thèse.

Chapitre 3

Etat de l'Art en Indexation Mélodique

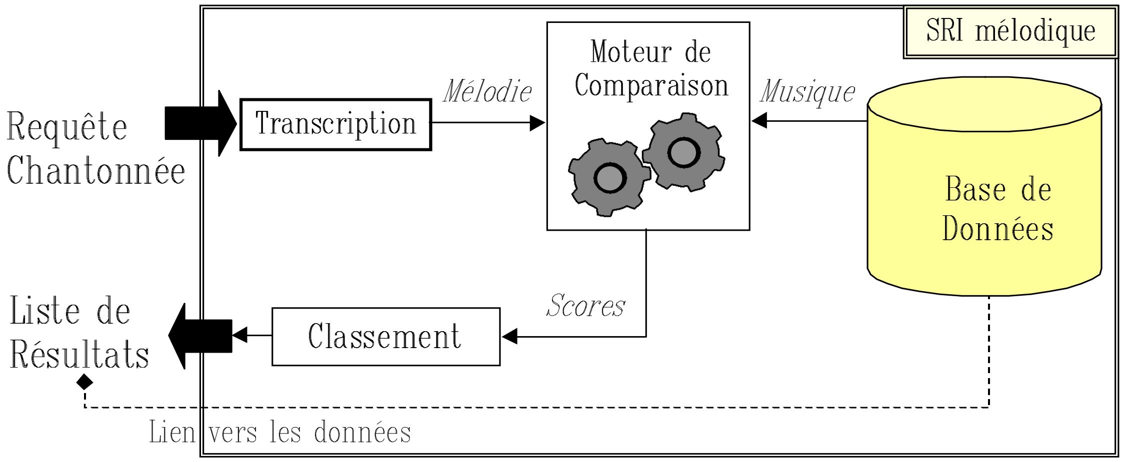


figure 3.1 : *Système de recherche de documents musicaux par chantonnement.*

3.1 Introduction

La mélodie est une information essentielle pour la description d'une séquence audio musicale. Ligne de sons successifs en hauteur et en durée [Gui96], elle possède une capacité à conserver le même contenu sémantique malgré une modification de sa hauteur, et/ou une modification de sa vitesse d'exécution. Ces propriétés permettent une certaine liberté dans la formulation d'une requête et une tolérance dans le choix des réponses retournées. De plus, la mélodie est invariante à l'instrument utilisé. Toutes ces qualités font de la mélodie le dénominateur commun de deux séquences d'une même œuvre, produites par deux instruments différents, dans deux tons différents et à deux tempos différents.

Dans ce chapitre, nous présentons l'état des recherches concernant l'indexation fondée sur la mélodie, c'est-à-dire l'indexation mélodique. En section 3.3, nous nous intéresserons au moteur de comparaison. Cœur du système d'indexation mélodique, sa tâche consiste à définir la similarité des mélodies qui lui sont présentées. Nous traiterons ensuite de la nature des bases de données concernées par les systèmes d'indexation mélodique (section 3.4), ainsi que des moyens offerts à l'utilisateur pour effectuer une recherche (section 3.5). Enfin, avant de présenter quelques systèmes disponibles en ligne (section 3.7), nous aborderons la question de l'évaluation de leur qualité (section 3.6).

Mais commençons dans un premier temps par poser quelques notions musicales fondamentales.

3.2 Notions musicales

Le but de cette section est de fournir au lecteur non musicien une première approche du vocabulaire nécessaire à la compréhension de ce document. Les notions musicales que nous allons présenter peuvent comporter certaines approximations.

Pour une connaissance plus précise des notions considérées, le lecteur pourra se référer à [Gui96].

Intervalle

Ecart séparant deux sons. Exprimé en demi-tons, l'intervalle témoigne du rapport des fréquences fondamentales des sons en cause. Soient deux sons ayant pour fréquences fondamentales f_{0_1} et f_{0_2} (exprimées en hertz). L'intervalle existant entre ces deux sons est :

$$Intervalle = 12 * Ln(\frac{f_{0_2}}{f_{0_1}}) \qquad (3.1)$$

Octave

Intervalle perçu par l'auditeur non exercé comme une similitude totale entre deux notes. Physiquement, il correspond à une fréquence fondamentale double (e.g. $f_{0_2} = 2 * f_{0_1}$).

Les gammes et les modes occidentaux furent conçus pour que les notes se répartissent à l'intérieur de l'octave. Dans ce contexte, l'octave est l'intervalle qui sépare deux notes qui portent le même nom. L'étendue des notes audibles est découpée en octaves numérotées, ce qui permet de définir la hauteur d'une note sans ambiguïté (e.g. mi2, fa6).

Tempérament

Manière de répartir les intervalles de la gamme sur un clavier ou un instrument à sons fixes. A la différence de la voix ou d'un violon qui peuvent emprunter le continuum des hauteurs, et donc adapter la hauteur d'une note à son contexte, les instruments à sons fixes doivent se plier à des approximations, dont aucune ne peut être considérée comme parfaite. De multiples variantes ont été adoptées au fil des âges, la dernière étant le tempérament égal, qui divise l'octave en douze demi-tons égaux.

Un instrument, ou encore, des hauteurs de notes, sont qualifiées de *tempérés* lorsqu'ils appartenant à ce système (tempérament égal).

Demi-ton

De la même manière que l'échelle des hertz est associée aux fréquences, l'échelle des demitons est associée aux hauteurs de notes. Elle témoigne de la perception logarithmique que nous avons des variations de fréquences fondamentales.

Le tempérament égal divisant l'octave en douze demi-tons égaux, le demi-ton correspond à la multiplication d'une fréquence fondamentale par $2^{1/12}$. L'écart en fréquence est donc fonction de la hauteur de la note initiale. Par contre, l'écart auditivement perçu est identique quel que soit le registre investi (i.e. quelle que soit le domaine de fréquence où l'on se place) ce qui est traduit fidèlement par l'unité de demi ton.

Le demi-ton correspond, par exemple, à l'écart existant entre deux touches de piano contiguës ou entre deux frettes successives d'une guitare.

Tonalité

La tonalité se compose des deux éléments distincts qui sont le *ton* et le *mode*. Le mode donne la "couleur" d'une tonalité (e.g. mineur, majeur), et le ton sa "hauteur globale" (e.g. sol4).

Armature ou Armure

Terme désignant la ou les altérations (i.e. $\sharp$ ou $\flat$) constitutives d'une tonalité. Ecrites immédiatement après la clef, ces altérations affectent toutes les notes de même nom, quelle que soit leur octave, et leur effet se prolonge pendant toute la durée du morceau.

Tessiture et Ambitus

La tessiture est la zone de hauteurs dans laquelle une voix, ou plus généralement un instrument, sonne *bien*. Les limites de la tessiture sont donc floues, mais appartiennent à l'*ambitus*, qui se définit par les notes extrêmes (dans le grave comme dans l'aigu) que peut atteindre l'instrument.

Tempo

Le tempo est lié à la rapidité d'exécution d'une musique. La manière dite *métronomique* de l'indiquer consiste à fournir un nombre de battements par minute (*bpm*), tout en assignant un battement à une unité de temps donnée (représentée par une *valeur de note*, e.g. blanche, noire, croche).

Valeur de note

La durée des notes d'une mélodie dépend du tempo auquel elle est jouée. Plus le tempo est rapide, plus les notes sont courtes. Dans la notation musicale classique, chaque note possède une *valeur*. Celle-ci renseigne sur le *rapport* existant entre les durées des notes présentes dans la musique considérée. Le tempo de la musique fixe la vitesse d'exécution de celle-ci.

Par exemple, la valeur *croche* désigne une durée deux fois moins longue que la valeur *noire* (la *noire*, elle, dure deux fois moins longtemps qu'une *blanche*). Jouées à un tempo de "60*bpm* à la noire", la *noire* dure 1s, la *croche* dure 0.5s, et la *blanche* dure 2s. Avec un tempo supérieur, (par exemple 120*bpm*), ces mêmes valeurs correspondront à des durées moins longues (respectivement 0.5, 0.25 et 1s).

Rythme

Le rythme résulte de la manière dont s'articulent entre elles, non seulement les durées, mais aussi et surtout les points d'appui sur lesquels celles-ci se placent.

Mesure et Métrique

La mesure est une unité rythmique. Sur une partition, les mesures sont délimitées par des barres verticales découpant la portée. La structure interne à la mesure est définie par la métrique. Cette dernière intervient dans la localisation des points d'appui du rythme d'une musique.

Temps (unité)

Unité rythmique, subdivision d'une mesure. Selon le type de mesure (simple/composée), le nombre de temps qu'une mesure peut contenir varie.

Les notions musicales indispensables étant posées, nous allons maintenant passer à ce qui constitue le cœur de systèmes d'indexation mélodique : le moteur de comparaison.

3.3 Le moteur de comparaison

Le moteur de comparaison a pour tâche de témoigner de la similarité entre mélodies. Sa conception nécessite la connaissance de liens entre des caractéristiques mesurables de ces mélodies et leur similarité perçue.

Dans cette section, nous allons voir sur quels principes se fondent les moteurs de comparaison, les contraintes auxquelles ils sont soumis, ainsi que les différents types proposés jusqu'à présent.

3.3.1 Fondements du moteur de comparaison

La théorie musicale offre des notions propices à la mise en place de mécanismes de comparaison mélodique. Par exemple, la connaissance du *mode* constitue un atout dans l'estimation de la similarité. En effet, remplacer une note par une autre est moins perturbant si le *mode* de la mélodie initiale est conservé que dans le cas contraire [Dow78]. Seulement, si ce type de notions est adapté à la quantification de différences mélodiques, leur caractère haut-niveau les rend difficilement accessibles (i.e. extractibles automatiquement). C'est pourquoi elles sont souvent considérées comme étant connues a priori (e.g. tempo, métrique, tonalité). Dans ce cas, la comparaison dépend d'informations à fournir en complément des mélodies soumises. La nature de ces informations (donc le type d'utilisateur capable de les fournir) destine le moteur de comparaison à un emploi plus ou moins spécialisé (e.g. outil d'analyse pour musicologue, système de recherche grand public).

La similarité mélodique possède encore de nombreuses facettes et frontières inexplorées. A notre connaissance, les expériences menées dans le domaine de la psychologie ne permettent pas, à ce jour, de concevoir un moteur de comparaison à la fois efficace et convivial. Celui-ci fournirait des jugements conformes à la perception sans que les mélodies soumises soient complétées pas des informations supplémentaires.

Invariance

Le point de départ communément adopté peut être illustré par une définition de la mélodie décrite comme l'entité présentant une invariance à la transposition, au changement de tempo, de timbre (instrument), et d'environnement acoustique.

Conformément à cette définition, un moteur de comparaison mélodique doit présenter les propriétés suivantes, notamment vis-à-vis des mélodies comparées :

– Invariance vis-à-vis des tons auxquels sont jouées les mélodies. Une différence de ton ne constitue pas une dissimilarité. Par ailleurs, pour un système de recherche, il serait trop contraignant de demander à un utilisateur de connaître la hauteur originale de la musique qu'il recherche.

– Invariance vis-à-vis des tempos empruntés. De même que pour le ton d'un morceau, une même séquence de notes jouée à deux vitesses différentes ne doit pas engendrer de dissimilarité. De plus, on ne peut attendre d'un sujet recherchant une musique qu'il connaisse sa vitesse d'exécution exacte.

Sensibilité aux erreurs

Nous venons de voir deux types de différences que le moteur de comparaison devait *ignorer*. Au contraire, d'autres différences doivent être prises en compte par le moteur de comparaison. La qualité d'un moteur de comparaison repose sur sa capacité à tolérer ces différences tout en les pénalisant conformément à la dissimilarité qu'elles engendrent.

Les différences -ou erreurs- les plus fréquemment considérées par les moteurs de comparaison sont les suivantes :
– transposition : hauteur de note erronée,
– insertion/omission : ajout/absence d'une note,
– fragmentation/consolidation : division d'une note en plusieurs/fusion de plusieurs notes en une.

Les connaissances sur la similarité mélodique ne permettent pas d'établir des hiérarchies entre ces différents types d'erreurs. Leur catégorisation est encore trop grossière pour le permettre. Un témoin de l'évolution - trop lente - de cette catégorisation est l'introduction du type d'erreur *fragmentation/consolidation* par Mongeau [MD90]. La définition de ce phénomène provient du fait que celui-ci, jusqu'alors exprimé en terme d'insertion/omission, semblait trop pénalisé par rapport à la dissimilarité perçue. En effet, il apparaît inadapté de pénaliser par le coût de trois insertions le remplacement d'une *blanche* par quatre *croches* (en particulier lorsqu'elles possèdent la même hauteur).

Cet exemple montre qu'il est difficile de disposer de règles permettant de pénaliser justement les différences entre mélodies. Par conséquent, la conception de moteurs de comparaison implique des concessions.

Compromis tolérance/précision

Les invariances et la sensibilité que nous venons d'aborder sont le résultat du compromis entre tolérance et précision déjà évoqué en 2.2. Le manque de tolérance d'un moteur de comparaison se traduit par une pénalisation excessive de documents pertinents. Néanmoins, ce besoin de tolérance s'accompagne d'un désir de précision. En effet, il est nécessaire d'assurer une bonne discrimination dans les réponses du système, surtout lorsque la base de données consultée devient volumineuse. Le manque de précision d'un moteur de comparaison se traduit par une trop grande clémence vis-à-vis de documents non pertinents.

Le moteur de comparaison gère le compromis tolérance/précision. Une grande tolérance autorisera d'importantes erreurs dans la requête. La faible précision induite se traduira par la présence de séquences non désirées dans ses réponses proposées.

Répartition des propriétés

Un moteur de comparaison est composé des deux éléments suivants :

– la *description* des mélodies à comparer : chaque mélodie est représentée par un *descripteur*. Ce dernier ne contient que les caractéristiques qui permettront une comparaison pertinente, c'est-à-dire conforme aux invariants et aux tolérances requis.

– la *mesure de similarité* qui fournit un score en rapport avec la différence témoignée par les deux descripteurs.

La gestion des propriétés du moteur de comparaison peut donc s'opérer à deux niveaux : celui de la représentation de la mélodie, et celui de la mesure de similarité.

Rapidité

La rapidité du processus entre évidemment en ligne de compte. Le temps d'attente de l'utilisateur est directement lié au temps de consultation de la base, donc au calcul de la similarité (et par conséquent, à la compacité des représentations mélodiques).

Dans les sous-sections suivantes, nous allons voir comment divers moteurs de comparaison témoignent de la similarité mélodique. La question de l'information nécessaire à la comparaison des mélodies sera systématiquement abordée. En effet, dans le contexte d'un système de recherche, ce point détermine la complexité de la requête attendue d'un utilisateur.

3.3.2 Prise en compte du contexte musical

Participation experte de l'utilisateur

Les descripteurs mélodiques utilisés par Mongeau et Sankoff s'expriment sur la base du ton et du tempo des mélodies concernées [MD90]. En effet, chaque hauteur de note est représentée en fonction de sa distance[1] au ton de la mélodie , et chaque durée en fonction de la période fixée par le tempo (cf. *valeur de note* p. 29). La mesure de similarité gère la sensibilité aux erreurs grâce à un algorithme de programmation dynamique. Une erreur est considérée comme une transformation mélodique qui possède un coût. La similarité de deux mélodies dépend du nombre et de la nature des transformations nécessaires au passage de l'une à l'autre. Lorsque plusieurs voies sont possibles, la similarité choisie correspond au coût minimum recensé.

Mongeau et Sankoff pénalisent les transpositions en fonction du contexte musical. Cela est rendu possible par la connaissance du *mode* des mélodies. Omissions et insertions de notes sont

[1] exprimée en demi-ton.

prises en compte, ainsi que deux phénomènes jusqu'alors ignorés : la fragmentation et la consolidation. Division d'une note en plusieurs et fusion de plusieurs notes en une, ces phénomènes sont gérés grâce à l'utilisation conjointe des informations fréquentielles (hauteurs) et temporelles (durées) contenues dans les descripteurs mélodiques.

Le moteur de comparaison de Mongeau et Sankoff présente donc les invariances requises et prend en compte (en les étendant) les différences mélodiques recensées. La similarité mélodique fournie témoigne des propriétés de la perception en tenant compte du contexte musical. En contrepartie, le moteur proposé requiert des données haut-niveau concernant les mélodies à comparer. En effet, les informations de ton, de mode, et de tempo doivent être fournies en complément aux mélodies elles-mêmes. En l'absence de processus d'extraction automatique, ces informations doivent être données par les utilisateurs, ce qui réduit considérablement leur nombre.

Processus autonome

Coyle et Shmulevich utilisent également une description comprenant à la fois des informations fréquentielles et temporelles [CS98]. Mais, à la différence de Mongeau et Sankoff, ces informations sont traitées séparément avant d'être combinées pour fournir la mesure de similarité finale.

L'information fréquentielle est exploitée de deux manières différentes. Le premier descripteur fréquentiel recueille les intervalles séparant les hauteurs successives. Cette solution est la plus largement utilisée car elle est représentative de la perception humaine. En effet, d'une séquence de hauteurs successives, nous mémorisons moins les valeurs absolues que les intervalles qui les séparent. Indépendants du ton, ils justifient le fait qu'un changement de ce dernier ne constitue pas une forte dissimilarité mélodique.

Le second descripteur prend en compte le contexte musical en exprimant les hauteurs par rapport au ton et au mode empruntés. Ceux-ci n'ont pas à être fournis par l'utilisateur car Coyle et Shmulevich les obtiennent par extraction automatique. Dans ce descripteur dit *perceptif*, l'indépendance au ton est également assurée.

L'information temporelle est codée sous forme de durées relatives. La durée initiale mesurée correspond au temps écoulé entre le début d'une note et le début de la suivante (*IOI* ou *Inter-Onset Interval*). La durée relative d'une note est égale à sa durée initiale divisée par la durée initiale de la note précédente. A l'instar des intervalles, l'information relative représente la perception humaine en assurant l'indépendance du descripteur au tempo emprunté.

La comparaison de deux mélodies chaque descripteur entraîne donc la comparaison de trois paires de descripteurs. Les deux descripteurs fréquentiels sont comparés à leurs homologues via une mesure de distance (norme L1). Une valeur unique est ensuite obtenue par combinaison des deux scores. La participation de la description *perceptive* dépend de l'indice de confiance qualifiant l'extraction de tonalité. Ainsi, lorsque la tonalité extraite est peu fiable, la similarité mesurée en tient peu compte. Les descripteurs temporels sont comparés via une mesure spécifique à leur nature logarithmique, la norme L1 n'assurant pas une mesure équitable. La combinaison des scores temporel et fréquentiel est laissée à l'appréciation de l'utilisateur qui peut ainsi favoriser l'aspect de sa requête en lequel il a le plus confiance (un mode de répartition est tout de même suggéré).

Le moteur de comparaison de Coyle et Shmulevich se contente des seules mélodies pour établir une similarité. Contrairement à celui de Mongeau et Sankoff, il ne demande donc aucun complément d'information à l'utilisateur. Le moteur proposé présente les invariances requises (ton et tempo) et traite les erreurs de transposition au regard du contexte mélodique. Seulement, la fiabilité de l'extraction de tonalité reste inconnue (nous y reviendrons au Chapitre 5), et la gestion des erreurs de type insertion/omission et fragmentation/consolidation n'est pas prévue.

Nous venons de voir deux moteurs de comparaison qui jugeaient de la dissimilarité de transpositions au regard du contexte musical. Nous avons également vu que cette prise en compte de la perception impliquait soit une participation experte de l'utilisateur, soit la mise en place d'un processus automatique. Or, la première solution réduit considérablement la population des utilisateurs, et la seconde est d'une fiabilité inconnue (cette question sera traitée en 5.2.2).

Nous allons maintenant voir le type de comparaisons qu'assurent les moteurs ne couvrant pas cet aspect de la similarité mélodique.

3.3.3 Contour mélodique et *string-matching*

Les travaux de Dowling ont montré que le *mode* et le *contour mélodique* étaient les deux principaux facteurs fréquentiels de mémorisation de mélodies [Dow78]. Le contour mélodique consiste en une *quantification* brutale des intervalles qui n'en conserve que le signe, soit le sens de variation des hauteurs de notes successives (la description est donc indépendante du ton). Les éléments d'un contour mélodique n'empruntent que trois états :

- hauteur plus haute que la précédente (intervalle positif),
- hauteur moins haute de la précédente (intervalle négatif),
- hauteurs identiques (intervalle nul).

Partant du principe que l'information de *mode* n'était pas disponible, des moteurs de comparaison ont été conçus sur l'exploitation du *contour mélodique*.

Dans [GLCS95], Ghias et *al.* limitent la description mélodique au simple contour mélodique. Aucune information concernant la durée des notes n'est conservée. La description assure donc l'indépendance au ton et au tempo empruntés. Le contour mélodique consistant en une succession d'états, il peut être codé sous la forme d'une chaîne de caractères. Le résultat est fréquemment appelé *représentation UDS*, en rapport à l'alphabet de 3 lettres U, D, et S, caractérisant respectivement les directions ascendante (Up), descendante (Down) et constante (Same) des hauteurs.

Deux mélodies étant représentées par des chaînes de caractères (ou *strings*), leur comparaison peut être effectuée grâce aux techniques d'appariement textuel (ou *string-matching*). La requête pouvant comporter des erreurs, Ghias et *al.* utilisent une méthode de *string-matching* flexible autorisant la substitution, l'insertion et l'omission de caractères [BYP92].

L'indexation textuelle a reçu beaucoup plus d'attention que l'indexation mélodique. La récupération de techniques éprouvées (en particulier le *string-matching*) n'est donc pas surprenante. Notons cependant que l'attrait d'un héritage (par sa facilité de mise en œuvre) ne fait pas de ce dernier la solution idéale à un problème.

Une similarité inéquitable

Dans la chaîne représentant un contour mélodique, un caractère ne correspond pas à une hauteur mais à un *sens de variation* de hauteur. Une erreur de transposition (i.e. hauteur erronée) n'a pas toujours la même influence sur la description, et par conséquent sur la comparaison. Ainsi, une erreur d'1 demi-ton sur un intervalle nul provoque un changement de symbole, alors qu'une erreur de 5 demi-tons sur l'intervalle +7 ne modifie pas le symbole résultant (U).

Pour illustrer plus précisément l'influence de ce phénomène sur la comparaison, nous prenons l'exemple de "Au clair de la lune". La représentation UDS des premières hauteurs ("do,do,do,ré,mi,ré") est SSUUD. Prenons maintenant trois requêtes visant cette mélodie, mais comportant chacune une erreur de transposition :

– requête 1 : "do,do,do,*ré♯*,mi,ré" - contour mélodique : SSUUD
– requête 2 : "do,do,do,*mi*,mi,ré" - contour mélodique : SSU**S**D
– requête 3 : "do,*ré*,do,ré,mi,ré" - contour mélodique : **UD**UUD

Malgré la présence d'une transposition, la première requête possède une description identique à la portion mélodique visée. La transposition de la deuxième entraîne le changement d'un seul symbole (noté en gras) dans la représentation UDS. Pour la troisième requête, deux symboles sont modifiés. Il apparaît donc clair que ces trois requêtes seront traitées différemment alors qu'elles comportent le même type d'erreur.

Il en est de même pour les erreurs d'insertion/omission. Dans la liste suivante, les requêtes 4 et 5 comportent une insertion, et les requêtes 6 et 7 une omission :

– requête 4 : "do,do,do,*do*,ré,mi,ré" - contour mélodique : SS**S**UUD
– requête 5 : "do,do,*ré*,do,ré,mi,ré" - contour mélodique : S**UD**UUD
– requête 6 : "do,do,ré,mi,ré" - contour mélodique : S UUD
– requête 7 : "do,do,do,ré,ré" - contour mélodique : SSU **U**

L'insertion d'une note provoque l'insertion d'un symbole et, selon les cas, le changement d'un symbole contigu. Le constat est analogue pour l'omission d'une note : la description subit la perte d'un symbole (notée " ") et, selon les cas, la modification d'un symbole contigu.

Les phénomènes de fragmentation/consolidation de note ne peuvent être pris en compte par ce moteur de comparaison. En effet, les durées de notes, indispensables à l'identification de tels phénomènes, sont absentes de la description mélodique.

Le moteur de comparaison fondé sur le contour mélodique témoigne donc d'une similarité uniquement fréquentielle, indépendante du ton et du tempo. Malheureusement, ce moteur souffre d'un *manque d'équité* dans son jugement sur la similarité mélodique. En effet, comme nous venons de le voir, la tolérance assurée par la description est arbitraire vis-à-vis de la similarité mélodique. Ainsi, avec ce genre de moteur, une transposition peut passer inaperçue, ou bien provoquer une, voire deux erreurs.

Mesure de la qualité de la similarité fournie

Le fait que le contexte musical ne soit pas pris en compte rend difficile le jugement de la similarité fournie par un moteur. En l'absence de critères objectifs de qualification des moteurs de comparaison, la solution consiste à placer ces derniers dans un contexte d'utilisation. Les réponses d'un système de recherche à qui l'on soumet des requêtes illustrent alors la manière dont le moteur de comparaison gère le compromis tolérance/précision évoqué en 3.3.1.

Notons au passage que les qualifications obtenues de la sorte dépendent de facteurs très variables d'une expérience à l'autre (e.g. critère de qualité, base de données, requêtes utilisées...). Les performances que nous allons évoquer ne seront généralement pas comparables, puisque issues de configurations différentes. La question de l'évaluation de systèmes sera abordée en section 3.6.

Pour revenir au cas du moteur de Ghias et *al.*, le fort potentiel de tolérance assuré par l'utilisation du contour mélodique se paie par une précision fragile. Les 90% de discrimination assurés par des séquences de 11-13 notes sont obtenus sur une base de 183 chansons[2]. Ce niveau de performance est trop faible pour que le système résiste à une augmentation de la base de données.

Ne conserver qu'un seul des deux facteurs révélés par Dowling s'avère donc insuffisant pour assurer une bonne discrimination. Le manque de précision du contour mélodique doit donc être corrigé par une augmentation de l'information conservée dans les descripteurs.

3.3.4 Variations sur la notion de contour

Contour mélodique du 2ème ordre

Une idée évoquée par Lindsay [Lin96], a été développée Blackburn [Bla99]. Il s'agit du contour mélodique du 2ème ordre. A la différence du contour mélodique classique qui compare une note avec la précédente, celui-ci compare une note avec celle qui la précède en 2ème position (par exemple, la 5ème note d'une mélodie avec la 3ème). Prenons pour exemple les deux suites de hauteurs suivantes :

- do,mi,ré,fa,mi,sol,fa,la,sol
- sol,la,fa,sol,mi,fa,ré,mi,do

Leurs contours mélodiques sont identiques (UDUDUDUD) alors que les séquences présentent des sens opposés. Les caractères ascendant de l'une et descendant de l'autre sont clairement révélés par leur contour du deuxième ordre :

- UUUUUUU
- DDDDDDD

Associer ce type de contour au contour mélodique classique permet d'augmenter la discrimination. Notons que cela n'élimine pas pour autant le caractère inéquitable de la similarité issue de contours (cf. p. 35). A notre connaissance, les performances d'un moteur de comparaison associant contours mélodiques du premier et du second ordre n'ont pas été mesurées.

Contour rythmique

Une autre manière d'augmenter la précision est de compléter la description par une information rythmique. Dans [SGM98], Sonoda et *al.* appliquent le principe du contour aux durées des notes. Ainsi, les variations de durées sont réparties en trois états : durée plus longue, égale, et plus courte que la durée précédente. La description adoptée est donc indépendante du tempo emprunté.

[2] Cela signifie qu'en moyenne une séquence de 11-13 notes suffit à juger comme non similaires les 90% des 183 documents de la base de données.

Les variations de hauteurs sont décrites par une représentation *UDS*. Un algorithme de programmation dynamique est utilisé pour déterminer la similarité fréquentielle et la similarité temporelle. Calculées de manière indépendante, ces deux contributions sont ajoutées pour déterminer la similarité finale.

Les expériences de Sonoda et *al.* montrent qu'associer contour mélodique et rythmique augmente sensiblement la précision. En effet, d'après leurs tests, une telle configuration présente une efficacité de 90% alors que les contours utilisés seuls mènent à des taux de réussite bien plus faibles[3] : 47% pour le contour mélodique seul, et 13% pour le contour rythmique seul.

Notons que ces performances sont obtenues sur une base de données modeste (200 chansons cf. tableau 3.1 page 49).

La volonté d'exploiter les algorithmes de *string-matching* disponibles a largement favorisé la description par séquences d'états. Face au manque de précision du contour, plusieurs travaux se sont dirigés vers une extension du concept en augmentant le nombre d'états utilisés.

Extension du contour

Dans un soucis d'amélioration de la précision, on assiste à une augmentation du nombre d'états servant à quantifier les variations de hauteur. Compte-tenu du type de description (valeurs relatives), les moteurs de comparaison proposés sont invariants aux tons employés. L'invariance au tempo et dépend de l'éventuel complément de description rythmique. La sensibilité aux erreurs, en partie gérée par la tolérance de la description, dépend du type d'algorithme de *string-matching* investi.

Dans [KCGV00], Kim et *al.* comparent les performances de représentations à 3, 5, et 7 états. Pour les deux derniers cas, plusieurs jeux de frontières d'états sont testés. Downie et Nelson poussent plus loin le nombre d'états en comparant des représentations à 3, 7, et 15 états (leurs découpages diffèrent de ceux testés par Kim et *al.*) [DN00]. Lemström et Laine structurent l'espace de variations en délimitant 7 zones procédant des recouvrements partiels [LL98]. Cette configuration mène à 11 états distincts dont la gestion (par la mesure de similarité) augmente la tolérance vis-à-vis des transpositions. Sonoda et *al.* proposent un système où le niveau de précision dépend de la discrimination voulue [SGM98] : la recherche débute via une représentation à 3 états, puis, si c'est nécessaire, passe à 9, puis 27 états. Cette augmentation itérative de la précision peut améliorer les résultats, mais au prix d'un temps de calcul accru. Autre facteur d'amélioration des performances, Sonoda et *al.* déterminent dynamiquement les frontières des états proposés en fonction de la répartition des données de la base.

La limite d'extension du contour correspond à une quantification des intervalles au demi-ton près. Nous rappelons que le demi-ton correspond à l'intervalle de base de la musique occidentale (cf. 3.2). Dans ce cas, il y a autant d'états que d'intervalles tempérés possibles. Aucun regroupement de valeurs n'étant appliqué, la précision de description est maximum[4]. La répartition des propriétés entre description et mesure de similarité désigne cette dernière comme seule gérante de la tolérance nécessaire aux transpositions. Ce type d'erreur étant déjà pris en compte par les algorithmes de *string-matching* (en plus des insertions/omissions), la mesure de similarité n'a pas

[3] La réussite est atteinte lorsque le document recherché se retrouve en première position sur la liste des réponses fournies par le système. La base de données utilisées comporte 200 chansons.

[4] Nous verrons que cela n'est pas le cas pour une mélodie chantonnée.

à être modifiée. Ce type de moteur de comparaison évite l'inéquité de traitement des erreurs de transposition, observée lorsqu'une partie de leur gestion est assurée par la description. La charge de calcul s'en trouve augmentée, mais la similarité mélodique fournie est mieux contrôlée.

La représentation par intervalles a largement été envisagée [KMT93, MSWH00, LL98]. Downie a montré que, compte-tenu des intervalles généralement rencontrés, une représentation à 15 états suffisait à assurer une précision équivalente à la précision maximum [DN00].

Concernant les quantifications plus grossières, plus le nombre d'états est important plus la discrimination est grande. Kim et *al.* tirent de leur expérience une sorte de point optimal, compromis entre précision et tolérance, pour une de leurs représentations à 5 états [KCGV00]]. Mais ce choix est lié au reste de la configuration adoptée : absence d'information rythmique et appariement exact, entre autres. Le changement d'un de ces éléments modifie la répartition *précision/tolérance*, et donc la place du "point optimal". Parmi les facteurs intervenant dans ce compromis *précision/tolérance*, l'information rythmique a son importance. Seulement, nous allons voir qu'à l'instar du mode, sa prise en compte nécessite souvent une participation de l'utilisateur.

3.3.5 Place de l'information rythmique

La considération de l'information rythmique nécessite, la plupart du temps, un apport d'information spécifique. Certes, l'invariance au tempo peut être assurée par un contour rythmique ou par des durées relative calculées sur la base de la durée de la première note [LL98][5]. Cependant, une description type *valeur de note* est généralement préférée. Pour cela, l'information pré-requise peut simplement être le tempo [SJ98, CC98], mais ce dernier doit parfois être accompagné d'informations plus précises telles que la métrique [KCGV00], ou les valeurs de note et de silence minimales [MSWH00]]. Quoi qu'il en soit, les descriptions proposées sont loin de témoigner de la réelle notion de rythme. En effet, nous rappelons que ce dernier résulte de la manière dont s'articulent entre elles, non seulement les durées, mais aussi et surtout les points d'appui sur lesquels celles-ci se placent (cf. page 29).

Les expériences de Sonoda et *al.* montrent qu'à représentation identique[6], le pouvoir discriminant de l'information fréquentielle est supérieur à celui de l'information rythmique [SGM98]. La prise en compte de cette dernière permet néanmoins d'améliorer la précision, avec pour conséquence l'ajout d'un acteur supplémentaire dans la répartition *précision/tolérance* du moteur de comparaison.

La contribution de l'information temporelle s'effectue généralement de manière indépendante à celle de l'information fréquentielle. La similarité finale provient alors de deux contributions distinctes. Par exemple, dans [KMT93], le score de chaque note correspond à la somme d'une similarité fréquentielle et d'une similarité rythmique. Une interaction est cependant possible. Toujours dans [KMT93], la contribution du score de la note dans la similarité finale dépend de la durée de la note considérée. L'information temporelle exerce une influence sur la similarité fréquentielle : la durée de la note est prise comme témoin de son importance dans la mélodie considérée.

La prise en compte de l'information temporelle permet d'améliorer la similarité mélodique

[5] Lemström et Laine assurent alors l'invariance au tempo par une mesure de similarité spécifique

[6] Il s'agit en l'occurrence de la représentation relative consistant à décrire l'information fréquentielle par les intervalles entre hauteurs (i.e. le rapport des fréquences fondamentales exprimé en demi-tons) et l'information temporelle par le rapport des durées successives.

mesurée. Cependant, dans les moteurs que nous avons vus, l'indépendance des calculs de similarités fréquentielle et rythmique empêche la prise en compte de phénomènes tels que la fragmentation/consolidation. En effet, ceux-ci nécessitent la considération simultanée des deux types d'information.

D'autres phénomènes peuvent cependant être révélés. Le moteur de Chen et *al.* utilise uniquement l'information temporelle [CC98]. Le tempo étant supposé connu, une mélodie est représentée par une séquence de valeurs de notes et de silences. La mesure de similarité permet de prendre en compte des phénomènes d'extension/contraction, de déplacement, et d'éclatement/fusion de durées[7].

Dans [MSWH00], McNab a comparé quelques-unes des nombreuses configurations possibles. A notre connaissance, McNab est le premier à avoir intégré une base de donnée relativement conséquente. Celle qui a été utilisée pour cette expérience contient plus de 9.000 mélodies monophoniques. Le classement suivant présente les configurations testées par précisions croissantes :

1. Contour mélodique et rythme (appariement flexible)
2. Intervalles et rythme (appariement flexible)
3. Contour mélodique seul
4. Intervalles seuls
5. Contour mélodique et rythme
6. Intervalles et rythme

L'association du rythme et du contour mélodique assure donc une discrimination supérieure à celle fournie par la description par intervalles seuls. A fortiori, les représentations fréquentielles à plus de 3 états associées au rythme surpassent également la description par intervalles seuls.

Par ailleurs, on observe que la flexibilité d'appariement fait baisser la discrimination. Il faut cependant noter que la sensibilité aux erreurs autorisée par cette flexibilité fait également partie des propriétés essentielles d'un moteur de comparaison. Cette remarque permet de souligner l'insuffisance du seul facteur "discrimination" pour la qualification de systèmes de recherche. Nous reviendrons sur le sujet en section 3.6.

3.3.6 Autre type de représentation par états

La représentation par séquences d'états est généralement appliquée de manière séparée aux hauteurs et aux durées. Or, nous avons vu que l'indépendance de traitement qui en découle n'est pas forcément idéale pour témoigner de la similarité mélodique.

Chou et *al.* proposent une description plus globale [CCL96]. Leur approche consiste à regrouper des notes successives afin d'en tirer un unique représentant : un accord [8]. Une mélodie est donc représentée par une séquence d'accords. L'intérêt d'une telle description est qu'elle reste invariante à toutes les erreurs recensées, à la condition que celles-ci respectent la couleur musicale locale (représentée par le nom de l'accord), c'est-à-dire qu'elles n'altèrent pas la mélodie au point de changer la séquence d'accords qui la représente. Une forte tolérance est donc assurée au niveau de la description. Malheureusement, en plus d'importants pré-requis (tempo, métrique et

[7] L'éclatement/fusion diffère de la fragmentation/consolidation par le fait qu'aucune information fréquentielle n'intervient dans son identification.

[8] Initialement, la notion d'accord désigne une combinaison d'au moins trois notes jouées simultanément.

armure sont supposés connus), le moteur proposé est dépendant du ton utilisé. Ce défaut pourrait être contourné en déclinant la requête dans tous les tons possibles, mais le temps de calcul serait largement accru.

3.3.7 Représentations associées à des distances

En marge des nombreux moteurs de comparaison exploitant les techniques de *string-matching* héritées de l'indexation textuelle, quelques moteurs tirent leur mesure de similarité du calcul de distances. Le moteur de Coyle et Shmulevich déjà évoqué en 3.3.2 en fait partie. Nous allons en voir deux autres exemples.

Dans [Bee97], Beeferman propose un descripteur décrivant la structure dynamique[9] des hauteurs successives. Correspondant à une version modifiée de la transformée en ondelettes de Haar, cette description témoigne de la répartition dynamique des notes depuis une vue globale (e.g. portion mélodique ascendante) jusqu'à la vue la plus locale (e.g. paire de hauteurs descendantes). Une distance euclidienne permet de comparer le descripteur de la requête avec ceux des mélodies de la base (des portions de taille fixe en sont préalablement extraites). La description et la mesure de similarité adoptées assurent une sensibilité à toutes les erreurs recensées. Malheureusement, le type de tolérance assuré par la description entraîne, une fois encore[10], une sensibilité arbitraire vis-à-vis des transcriptions et des insertions/omissions.

Dans [CCC$^+$00], Chen et *al.* commencent par extraire un "tracé mélodique" de la séquence de notes considérée (courbe de type "marches d'escalier" dans le plan temps/hauteur). Quatre formes élémentaires paramétrées permettent de décrire ce tracé. La description mélodique consiste en une séquence de formes dont la hauteur et la durée[11] sont spécifiées. La similarité de deux mélodies est calculée si celles-ci possèdent la même séquence de formes (indépendamment de la valeur des paramètres). Les paramètres de hauteur et de durée permettent alors de mesurer deux *distances* qui sont combinées pour donner la mesure de similarité finale.

Cette description est invariante à certaines erreurs, comme par exemple la fragmentation d'une *blanche* en quatre *croches* de même hauteur. Les transpositions sont encore une fois sources d'influences variables. Elles peuvent entraîner une simple modification des paramètres d'une séquence de formes, ou carrément changer la séquence par l'ajout ou l'élimination de formes. Ces différences de traitement n'étant pas justifiées par des différences humainement perceptibles, ce moteur présente lui aussi une mesure de la similarité mélodique inéquitable.

Distance vs. *string-matching*

L'utilisation de techniques de *string-matching* pour la comparaison d'états entraîne une mesure de similarité discrète. Le minimum de similarité témoigné (i.e. la plus grande dissimilarité) dépend du nombre d'erreurs tolérées. Si ce dernier n'est pas limité, la pire similarité mesurée est égale au nombre maximum d'erreur qu'une mélodie peut présenter. Cette quantité est directement liée à la taille de la description.

Au contraire, l'utilisation d'une distance permet une mesure de similarité continue, non bornée, et toujours dépendante de la nature des descripteurs comparés (et non de leur taille).

[9] L'information concernée, appelée *pitch dynamics*, possède une précision située entre le contour mélodique et les intervalles.

[10] cf. p. 35

[11] Il s'agit plutôt de *valeurs* car les durées sont exprimées en fonction du tempo, supposé connu (cf. 3.2).

3.3.8 Conclusion

Il apparaît difficile d'assurer un jugement de qualité sur la similarité mélodique sans fournir au moteur de comparaison des informations concernant les mélodies soumises (e.g. mode, tempo, métrique). Or, cet apport d'information requiert généralement des compétences musicales. C'est pourquoi les moteurs les plus fidèles à la similarité mélodique ne peuvent s'adresser qu'à une faible population d'utilisateurs experts (e.g. musiciens, musicologues).

Les recherches se sont donc tournées vers des moteurs de comparaison plus autonomes, mais moins conformes aux connaissances sur la similarité mélodique. Les premières tentatives de description ont été associées à des mesures de similarité fondées sur des techniques de *string-matching*. Les travaux suivants ont en majorité continué dans cette voie, excepté quelques cas fondant leur mesure de similarité sur des distances ou des algorithmes de programmation dynamique. Ces voies alternatives, un peu vite éclipsées par les techniques héritées de l'indexation textuelle, gagneraient à être d'avantage investies.

Les propriétés de précision/tolérance d'un moteur de comparaison peuvent être réparties entre description et mesure de similarité. Or, en dehors des invariances de base (ton et tempo), la tolérance assurée par la description se traduit généralement par une sensibilité aux erreurs inhomogène. La mesure de similarité qui en découle n'est donc pas équitable.

Une solution semble consister en une description précise des mélodies, la sensibilité aux erreurs étant gérée par la mesure de similarité. Cette constatation s'applique surtout à l'information fréquentielle des mélodies. L'information temporelle, moins étudiée pour l'instant, vient généralement en complément de description et nécessite souvent une participation de l'utilisateur (qui doit fournir le tempo).

La question du coût de calcul est peu abordée. En effet, le domaine de recherche étant récent, la priorité est donnée au développement de moteurs de comparaison témoignant d'un jugement sur la similarité mélodique qui soit satisfaisant.

3.4 La base de données musicales

Idéalement, la base de données devrait être constituée de documents musicaux sous leur forme usuelle, c'est-à-dire des formes d'ondes temporelles brutes (cf. format CD, wav, aiff, pcm...) ou bien compressées (e.g. MP3). Cependant, il existe un décalage entre cette représentation de la musique et les informations nécessaires aux moteurs de comparaison (i.e. la partition). Nous allons voir que le passage de l'un à l'autre (appelé transcription) n'est pas une opération aisée. Cette difficulté motive l'utilisation de données musicales synthétiques, ce qui constitue une concession par rapport à la base de données idéale.

Dans cette section, nous verrons également quelles sont les données musicales qui seront effectivement comparées à la requête via le moteur de comparaison. Nous nous intéresserons à l'indexation proprement dite, c'est-à-dire à la classification des données de la base en fonction de leur contenu. Enfin, nous ferons un tour d'horizon des bases de données utilisées dans les publications dont nous disposons.

3.4.1 L'accès à la partition

Nous l'avons vu plus tôt, les moteurs de comparaison s'appuient sur des éléments correspondants à la partition des musiques à comparer. Les informations requises vont du niveau le plus bas (hauteur et durée des notes) au niveau le plus haut (tonalité, métrique). Seulement, même les informations les plus basiques sont difficilement extractibles des documents musicaux usuels, du moins d'une manière automatique. En effet, la musique qu'ils contiennent est bien souvent *multi-timbrale* et *polyphonique*. Cela signifie d'une part, que plusieurs instruments y interviennent simultanément, et d'autre part, qu'un instrument peut produire plusieurs notes simultanément. La partition structure cette double simultanéité en représentant séparément les parties des différents instruments. Les techniques de transcription automatique actuelles sont malheureusement incapables de fournir un résultat comparable [12] [Kla98, Mar96].

Si l'on exclut la transcription manuelle extrêmement coûteuse en investissement humain, deux solutions se dégagent :

1. Se limiter aux documents musicaux adaptés aux performances des outils d'analyse (semi-) automatique actuels, c'est-à-dire acceptant des contenus audio *monophoniques*[13], ou des configurations simples et connues a priori (e.g. piano solo sans réverbération).
2. Se limiter aux documents musicaux offrant un accès à (une part de) l'information désirée, comme des fichiers de musique synthétique (e.g. fichiers MIDI).

Dans les travaux dont nous avons eu connaissance, la deuxième solution est généralement adoptée. En effet, les fichiers de musique synthétique sont nettement plus nombreux que les documents susceptibles d'être analysées avec de bons résultats. De plus, ils offrent un accès rapide et précis à une partie de l'information désirée, à la différence des transcriptions qui sont à la fois coûteuses en calcul et imprécises.

Dans la sous-section suivante, nous allons introduire les principaux éléments de la norme MIDI adoptée par la majorité des fichiers musicaux synthétiques existants.

3.4.2 La norme MIDI

Le MIDI (Musical Instrument Digital Interface) est une représentation conçue pour la communication entre instruments de musique électroniques. La spécification MIDI définit aussi bien le protocole d'envoi et de réception des messages de commande que les connections physiques existant entre les différents éléments en relation. Un système MIDI se compose de matériel producteur de son (synthétiseurs, échantillonneurs), d'éléments de contrôle (clavier, batterie électronique), et de logiciels, appelés séquenceurs, fonctionnant sur micro-ordinateur.

Les éléments de base de la représentation sont les évènements de début et de fin de note, accompagnés des informations de hauteur, de vélocité (= force d'attaque d'une note). En général, chaque instrument est assigné à une piste, elle-même diffusée dans l'un des 16 canaux MIDI.

La représentation MIDI peut être stockée dans des fichiers au format standard MIDI (extension ".mid") qui sont manipulables par les séquenceurs : changer de ton, de tempo, occulter un instrument, modifier certaines notes sont par exemple des transformations aisées. Cette flexibilité en fait un outil très répandu chez les musiciens et les compositeurs. Les premiers l'utilisent

[12]à partir d'un document sonore quelconque
[13]contenus où n'apparaît qu'une seule voix à la fois.

souvent à des fins d'accompagnement (type *karaoké*), et les seconds pour disposer d'un retour sonore (même si le rendu est primaire par rapport à ce que donnerait, par exemple, un réel orchestre).

Les fichiers MIDI sont très répandus sur Internet. Le fait que ces fichiers ne contiennent que les instructions correspondant au directives de jeu, et non la musique elle-même entraîne deux conséquences importantes.

La première est une taille de fichier extrêmement faible, par exemple 14Ko pour un morceau de piano solo de 4 minutes, contre 40Mo pour une séquence de durée équivalente codée en qualité CD[14], et environ 12 fois moins, soit quand même plus de 3Mo pour une version MP3 de bonne qualité (100kbit/s). Cette compacité fait du MIDI un moyen avantageux pour diffuser de la musique à très bas débit (e.g. habillage musical de sites web).

La deuxième conséquence importante est un résultat sonore dépendant du matériel utilisé pour jouer le fichier MIDI (synthétiseur de carte son par exemple). Quoiqu'il en soit, le rendu sonore est limité par les performances des méthodes de synthèse mises en jeu. Tous les instruments ne sont donc pas équitablement représentés. Parmi les défavorisés, les parties vocales sont généralement assurées par des instruments de type flûte ou orgue.

La qualité d'un fichier MIDI dépend de son mode de création. Nous distinguons deux cas extrêmes. Dans le premier, l'utilisateur du séquenceur spécifie une à une les informations contenues dans son fichier, qu'il s'agisse du tempo, de la métrique, de l'armure (et de leur éventuels changement au cours de la musique) et évidemment des notes. Dans ce cas, l'information contenue dans le fichier MIDI est très proche de celle contenue dans une partition. On peut d'ailleurs soupçonner l'utilisateur de disposer de cette dernière, et de désirer écouter de son rendu sonore.

Dans le second cas de figure, l'utilisateur dispose d'un instrument MIDI[15] qui lui permet de jouer la musique qu'il désire stocker. L'avantage est la prise en compte des nuances qui évitent le côté mécanique d'une partition "parfaitement" réalisée. Par contre, rien n'oblige l'utilisateur à compléter sa performance par les valeurs de tempo, de métrique ou de tonalité correspondantes.

Les seules informations systématiquement présentes dans un fichier MIDI sont les paramètres de notes (hauteur, durée, vélocité) et les instruments qui les jouent. En règle générale, le tempo est spécifié, d'avantage que la métrique, elle-même beaucoup plus fréquente que l'armure. Un fichier MIDI n'est donc pas toujours l'équivalent de la partition. Ce constat pénalise les moteurs de comparaison qui nécessitent d'avantage d'information que les seules notes jouées. Seule l'extraction automatique des informations complémentaires requises (e.g. la tonalité) pourrait permettre l'utilisation de ces moteurs. Nous détaillerons ces questions au chapitre 5, lors de la construction de notre moteur de comparaison.

3.4.3 Sélection des données pertinentes

Comme nous venons de le voir, les fichiers MIDI sont structurés en pistes, chacune d'elle étant généralement associée à un instrument particulier. Il peut être avantageux de tirer profit de cette structure afin de *réduire l'espace de recherche* en sélectionnant les données effectivement décrites.

Par ailleurs, la musique possède des propriétés particulières (e.g. répétition de thèmes) également exploitables pour améliorer les performances des systèmes de recherche.

[14] Qualité CD : stéréo, 16bits, échantillonnée à 44.1kHz

[15] Il peut s'agir d'un clavier, d'une guitare munie d'un capteur, ou encore d'un contrôleur de souffle (type "instrument à vent").

Identification de types de contenus musicaux

Dans [BD00], Blackburn et DeRoure proposent de répartir les pistes MIDI en quatre catégories : *accompagnement*, *basse*, *percussion/rythme*, et *soliste*. Les descripteurs d'une piste témoignent de sa densité en notes et en accords, de la hauteur et de la durée des notes présentes, des intervalles parcourus, du degré de polyphonie et de répétition, du canal MIDI et de l'instrument assigné.

Trois méthodes de classification sont testées (statistique, heuristique, et k-plus-proches-voisins), menant à des performances proches de 70% de classification correcte. Les 90% sont atteints lorsque le nombre de classes est réduit à trois, par la fusion des classes *soliste* et *accompagnement*.

Ce genre de méthode permet de limiter les données à parcourir pour la recherche. Pour éviter de s'adresser à l'ensemble de la base de données, la requête doit comporter l'information permettant d'identifier le domaine de la recherche. Concernant la *navigation* dans les bases de données (originalité de Blackburn et DeRoure), la requête est constituée d'une portion de document consulté. Les portions retournées par le système appartiendront à des pistes de classe identique à celle contenant la portion sélectionnée en requête (e.g. *percussion/rythme*).

Pour appliquer ce type de réduction de l'espace de recherche à un système reposant sur une requête produite par l'utilisateur (e.g. jouée, chantée), il faut contraindre ce dernier à préciser la classe du contenu qu'il recherche (e.g. *basse*).

Dans [TYK00], Tang et *al.* tentent d'identifier automatiquement la piste contenant ce qu'ils appellent la *ligne mélodique* d'un fichier (notion voisine de l'étiquette *soliste* dans le paragraphe précédent). Ainsi, un fichier serait uniquement représenté par le contenu de cette piste. Les auteurs supposent donc que les requêtes ne s'adressent qu'aux lignes mélodiques des musiques de la base. Seulement, sur les cinq descripteurs testés[16], le plus fiable n'est pas toujours extractible. En effet, il s'agit du descripteur fondé sur le *champ textuel* associé à chaque piste qui témoigne de la présence de mots-clés tels que "melody", "vocal", "voice", "solo". Or, selon les auteurs, moins de 50% des fichiers MIDI disponibles possèdent des pistes annotées. Par conséquent, cette méthode n'apparaît pas vraiment efficace.

L'expérience de Tang et *al.* révèle la supériorité d'un étiquetage manuel sur des descripteurs trop simplistes. Bien que plus élaborés, les descripteurs proposés par Blackburn et DeRoure peinent à dissocier les classes *soliste* et *accompagnement*, facilement différentiables à l'écoute. L'identification de contenus musicaux haut-niveau à partir de données MIDI ne semble donc pas encore assez précise pour permettre une sélection pertinente de données musicales.

Par ailleurs, nous allons voir qu'identifier l'information pertinente d'un fichier MIDI est plus complexe que la simple sélection d'une piste.

Réduction de la polyphonie

Les moteurs de comparaison sont chargés de définir la similarité de mélodies. *Ligne de sons successifs en hauteur et durée* [Gui96], la mélodie vient s'opposer à la nature des musiques constituant la base de données. En effet, la première est monophonique alors que les autres sont polyphoniques. Certes, le format MIDI permet de décrire séparément les différents instruments, mais

[16] Il s'agit de la vélocité moyenne des notes de la piste (valeur haute attendue), le rapport entre passages polyphoniques et passages monophoniques (valeur basse attendue), le taux de silence (valeur basse attendue), l'ambitus (valeur moyenne attendue), le nom de la piste (mots-clés significatifs attendus).

la polyphonie intrinsèque à chacun demeure. Nous allons donc voir les solutions proposées pour ramener cette polyphonie à un matériau adapté aux moteurs de comparaison.

Uitdenbogerd et Zobel se sont intéressés à la question d'une mélodie unique représentant une musique polyphonique [UZ98]. Quatre méthodes d'extraction ont été testées sur dix fichiers MIDI. Huit auditeurs ont jugé les mélodies obtenues, au regard de la capacité de celles-ci à représenter la musique polyphonique originale. Cette expérience a montré que la meilleure méthode est celle qui consiste à sélectionner à chaque instant la note la plus haute de la musique considérée.

Dans [UZ99], Uitdenbogerd et Zobel ont à nouveau comparé ces techniques d'extraction, mais cette fois dans le contexte d'un système de recherche (i.e. soumission d'une requête, observation des réponses du système). Les auteurs ont ajouté une cinquième méthode consistant à appliquer la technique élue en [UZ98] à *chacune des pistes* de manière indépendante. Dans ce cas, une musique n'est plus représentée par une mélodie unique, mais par autant de mélodies que son fichier MIDI possède de pistes/instruments. Selon les auteurs, cette variante donne de moins bons résultats. Nous pensons cependant que cette solution est intéressante, et ce, pour deux raisons :

1. La première provient du fait qu'une requête puisse viser la mélodie d'un instrument particulier, sans que celui-ci joue dans le registre aigu de la musique recherchée (typiquement une ligne de basse). Représenter chaque piste par une mélodie autorise ce genre de requêtes.

   En contrepartie, il est évident que le nombre de mélodies candidates à la comparaison est plus important (autant de mélodies que de pistes et non plus autant de mélodies que de fichiers). Cette augmentation des candidats à l'appariement est au cœur de la deuxième raison qui nous pousse à relativiser la suprématie de l'approche "*une mélodie par fichier*".

2. Les performances de l'approche "*une mélodie par piste*" sont jugées inférieures à cause l'augmentation des fausses alarmes[17] due au plus grand nombre de candidats perturbateurs. Or, nous pensons que cette conclusion est à relativiser au regard des références choisies par Uitdenbogerd et Zobel pour le critère de qualité utilisé (i.e. critère précision/rappel, cf. 3.6).

   En effet, les documents jugés *pertinents* par les auteurs (i.e. références attendues en réponse pour une requête donnée) sont les différents *arrangements*[18] de la musique visée. Ceux-ci sont identifiés grâce au *nom* des fichiers qui les contiennent. Cela signifie qu'un document possédant une portion similaire à la requête serait considéré comme indésirable dans la liste des réponses si son nom ne correspondait pas au *titre* de la musique recherchée !

   Prenons l'exemple d'une musique n'appartenant pas aux arrangements identifiés par Uitdenbogerd et Zobel, mais possédant malgré tout une portion mélodique similaire à la requête. Un processus d'extraction révélant cette portion placerait naturellement la musique dont elle est issue dans les premiers documents proposés. Malheureusement, le critère choisi par les auteurs pénaliserait injustement le processus utilisé en jugeant -à tort, puisque indépendamment de la similarité du contenu- le document proposé comme étant indésirable.

Il n'existe pas *une* méthode d'extraction remportant tous les suffrages. Mue par la volonté de ne conserver que des mélodies susceptibles d'être recherchées, la réduction de la polyphonie mériterait d'être approfondie.

[17] Une fausse alarme correspond à la présence d'un document indésirable dans la liste des réponses fournie par un système de recherche.

[18] Une différence d'arrangement correspond à un ton différent, un nombre de parties différent, ou des différences de rythme, de dynamique ou de structure [UZ99].

Au delà de cette réduction nécessaire à la comparaison, la sélection des données pertinentes peut être poursuivie. Pour cela, on utilise une caractéristique forte de la musique qui consiste en la répétition de motifs mélodiques particuliers, appelés *thèmes*.

Détection de thèmes

Les mélodies représentant les documents de la base peuvent comporter des redondances, dues à la répétition du refrain d'une chanson par exemple. La répétition (exacte ou approximative) de motifs est un critère constitutif d'un thème -et d'une manière générale- d'une mélodie facilement mémorisable.

Les méthodes de détection des redondances peuvent servir à éliminer celles-ci de la description de la base de données. En effet, il apparaît superflu de comparer une requête aux différentes répétitions d'une même mélodie au sein de la base de données. Une autre approche peut consister à n'assurer la description que des mélodies considérées comme des thèmes. Néanmoins, ce choix s'accompagne de l'hypothèse forte que seuls les thèmes sont susceptibles d'être recherchés.

La détection de redondances est essentiellement un problème d'identification de contenus similaires. En l'occurrence, il s'agit d'extraire des portions mélodiques et de voir si d'autres portions leur ressemblent. Le cœur du processus est donc bien connu (cf. 3.3) : les mélodies sont représentées par des descripteurs et comparées grâce à une mesure de similarité. Le type de moteur de comparaison détermine la similitude des redondances révélées.

Le nombre de candidats possibles étant d'autant plus important que la base de données est volumineuse, la difficulté de la détection de thèmes réside en une optimisation de la recherche. Le but est de dégager les portions redondantes les plus longues de la manière la plus efficace possible.

Dans le cadre de leur projet *Muse*, dédié au développement d'un système d'indexation musicale, Hsu et *al.* ont successivement testé trois méthodes permettant de découvrir des *répétitions exactes de sous-chaînes de caractères* [HLC01]. La similitude des thèmes détectés dépend donc de ce que représentent ces caractères. Selon la description choisie, un caractère peut représenter une hauteur, un intervalle, un sens de variation, un accord... Toujours dans la domaine de la comparaison textuelle, Tseng a proposé une méthode dérivée d'un algorithme de recherche de mots-clés dans les documents textuels [Tse99].

De son côté, Rolland propose une méthode baptisée *FlExPat* [19] fondée sur la combinaison de correspondances de base [Rol99]. A l'échelle de la note, ces correspondances peuvent prendre en compte des phénomènes tels que (multi-)insertion/ommission, fragmentation/consolidation, ce qui n'était pas le cas des méthodes précédentes. Ici encore, la similitude des portions considérées comme redondantes dépend de la description mélodique employée. *FlExPat* fait partie de l'environnement *Imprology*, qui offre une grande richesse de description, de la plus élémentaire (e.g. contours type UDS) à la plus élaborée (e.g. modes, schémas harmoniques/cadences). Le type de redondances détectées dépend de la description adoptée ainsi que des transformations mélodiques prises en compte dans la comparaison (toutes sont spécifiables par l'utilisateur).

Les méthodes citées ont toutes adopté des descriptions *dépendantes* du ton des mélodies. Rendant l'exposé plus simple, ce choix ne permet pas de révéler les répétitions *transposées* que l'on

[19] *FlExPat* pour *Flexible Extraction of Patterns*

peut rencontrer au sein d'un document (modulations) ou de la base de données entière (musique considérée présente avec des tons différents). En effet, au delà d'une extraction des thèmes d'une musique, l'identification des redondances gagnerait à être appliquée à l'échelle de la base de données.

Après avoir sélectionné les données pertinentes des fichiers MIDI constituant la base de données, il peut être avantageux d'organiser leurs descripteurs dans des structures favorisant le processus de recherche. Cette opération constitue, à proprement parler, une *indexation*.

3.4.4 Indexation

L'indexation désigne le classement de documents en fonction de leur contenu. Cette opération est destinée à retrouver efficacement une information recherchée. Le terme *indexation* est devenu l'appellation générique des différentes opérations liées à la description, à la manipulation et à la recherche des documents. Dans cette sous-section, nous nous intéressons au sens original du terme.

La possibilité de décrire de la musique grâce à des chaînes de caractères a permis l'héritage de travaux initialement destinés à gérer automatiquement les documents textuels. Nous avons vu que les techniques de *string-matching* sont largement exploitées dans les mesures de similarité (cf. 3.3). Seulement, le langage musical et le langage écrit présentent des différences qui nécessitent parfois une adaptation des techniques héritées.

Une spécificité de la musique est qu'elle ne possède pas d'entité équivalente au *mot*, ou si c'est le cas, les limites n'en sont pas aussi clairement définies. Kageyama restreint les points de comparaison aux débuts des musiques et après chaque silence [KMT93], mais cette solution présume illégitimement de la nature des requêtes, empêchant l'utilisateur d'accéder à une grande partie du contenu initialement disponible. Plus généralement, on assiste à une absence de restriction sur le point d'entrée d'une recherche : la portion mélodique visée par une requête est recherchée sur l'ensemble de la description des données.

Premières tentatives

A notre connaissance, Chou et *al.* sont les premiers à utiliser une structure d'index pour éviter le parcours exhaustif des descripteurs de la base de données musicale [CCL96]. L'arborescence choisie, appelée *PAT-tree*, leur permet d'indexer l'ensemble des portions extractibles de la description des données de la base.

Dans [LHU98], Lemström et *al.* démontrent la nette supériorité de la stratégie d'indexation sur l'approche linéaire (représentée par l'algorithme *Boyer-Moore*). En effet, l'approche exhaustive implique un temps de recherche augmentant linéairement avec la taille de la base de données alors que l'approche par index permet un temps de recherche quasiment constant et d'un niveau sensiblement inférieur. La structure d'index utilisée, appelée *suffix-trie*, permet de classifier l'ensemble des suffixes extractibles des descripteurs de la base. Cette structure se révèle cependant inadaptée à la gestion de volumes de données réalistes. En effet, même dans des versions tronquées (i.e. profondeur d'arbre limitée), le *suffix-trie* requiert un espace mémoire trop important.

Pour appréhender des bases de données supérieures aux leurs (i.e. supérieures à 100 chansons), Lemström et *al.* envisageaient d'implémenter une structure plus compacte, appelée *suffix tree*. Comme nous l'avons vu précédemment, cette structure a servi de point de départ aux expériences de Liu et *al.* concernant l'extraction de thèmes [LHC99].

Vers un cahier des charges plus réaliste

Les contributions que nous venons d'évoquer se limitent à un seul descripteur par mélodie. Or, il n'est pas rare qu'à une mélodie correspondent *plusieurs* descripteurs (e.g. une chaîne de caractères pour l'information fréquentielle et une autre pour l'information temporelle). Dans [LC00], Lee et Chen, présentent quatre structures d'index dédiées à la recherche conjointe sur deux descripteurs.

Autre avancée, Lee et Chen considèrent aussi bien l'appariement exact qu'approximatif. En l'absence de processus de recherche permettant les erreurs, la solution consiste à générer les différentes variantes possibles de la requête, puis de lancer autant de recherches d'appariement exact. Cette solution devient vite coûteuse lorsque la tolérance autorisée et la taille des requêtes augmentent.

Bilan

Les structures d'index exploitables sont limitées par le désir de disposer de base de données volumineuses et l'absence de restriction sur les points d'entrée à la recherche. Les expériences d'indexation menées jusqu'à présent concernent des volumes de données éloignés de proportions réalistes : le maximum recensé est une base de données de 1000 chansons [LC00].

La question de l'indexation vient s'ajouter à celle, déjà difficile, du choix du moteur de comparaison mélodique. C'est sans doute la raison pour laquelle les travaux qui focalisent sur le sujet sont encore peu nombreux. Si les recherches futures peinent à trouver des solutions mieux adaptées aux contenus musicaux, la démarche consistant à se rapprocher des contenus textuels pourrait être préférée, afin d'exploiter au maximum les outils d'indexation textuels classiques.

Downie, par exemple, manipule des entités comparables aux mots [DN00], ce qui lui permet d'utiliser une méthode de recherche issue du *SMART Information Retrieval System*. La difficulté de définir les limites de "mots" musicaux est contournée grâce à l'extraction de portions de descripteurs de taille constante (appelées *n-grams*, *n* étant le nombre de caractères définissant la taille choisie). Aucun point d'entrée n'est négligé car les portions extraites présentent un recouvrement maximum (le pas d'avancement est égal à un caractère). Cette solution permet en outre de gérer l'insertion/omission de caractères.

3.4.5 Caractéristiques des bases de données utilisées

L'industrie de la musique produit 10.000 nouveaux albums chaque année, et dans le même temps, 100.000 œuvres originales sont déposées pour protection [UZ99]. Comme nous l'avons vu dans cette section et la précédente, de nombreux tests mettent en jeu des bases de données trop modestes pour être réalistes. Le tableau 3.1 reprend les volumes annoncés dans diverses publications.

Quelques centaines de fichiers MIDI ne suffissent pas à révéler les limites d'un moteur de comparaison quelque peu performant. Depuis peu, on observe une augmentation des volumes de données utilisés mais un manque de précision sur celles-ci empêche la comparaison des résultats d'expériences différentes. Les fichiers sont-ils polyphoniques ? Si c'est le cas, toutes les pistes

Référence	Volume de données
[KMT93]	500 fichiers MIDI
[GLCS95]	183 fichiers MIDI
[SGM98]	200 fichiers MIDI
[LL98]	154 fichiers MIDI
[KCGV00]	50 fichiers MIDI
[UZ99]	10.000 fichiers MIDI
[Tse99]	135 fichiers MIDI
[Bee97]	1.000 fichiers MIDI
[Bla99]	8.000 fichiers MIDI
[TYK00]	16.000 fichiers MIDI
[MSWH00]	9.400 mélodies

TAB. 3.1 : *Volumes de données investis dans l'évaluation de performances.*

sont-elles indexées ? Une requête peut-elle s'adresser à n'importe quelle portion de la base ?...

Le nombre de notes constituant les mélodies pouvant faire l'objet d'une recherche semble constituer un indicateur pertinent pour renseigner sur le volume d'une base de données. Par exemple, Tseng tirent de leurs 135 fichiers MIDI de musique pop chinoise 1052 lignes mélodiques de 746 notes en moyenne. Grand vainqueur des volumes cités, les 100.000 fichiers MIDI de l'université de Waikato [BNMW$^+$99], représentent plus de 500 millions de notes ! A notre connaissance, cette base de données n'a encore été utilisée pour aucun test. Le défaut de l'indicateur "nombre de notes" est qu'il ne renseigne pas sur la durée, ni sur la nature des mélodies, cette dernière pouvant influencer les performances. Une solution serait la création d'un corpus référence. Les 9.400 mélodies que McNab et *al.* ont mis à la disposition de Downie [DN00] en constituent peut être une amorce.

3.4.6 Conclusion

Les bases de données constituées de musiques sous leur forme usuelle (e.g. MP3) ne permettent pas un accès à l'information nécessaire à la comparaison mélodique. En attendant une meilleure robustesse des outils de transcription automatique, la solution généralement adoptée consiste à utiliser des fichiers MIDI. Destinés à être joués par des synthétiseurs, ces fichiers contiennent les caractéristiques principales[20] des notes de chacun des instruments intervenant dans une musique.

L'identification automatique de contenus (de type *mélodie*, *accompagnement*...) permettrait de réduire l'espace de recherche ou d'offrir une navigation thématique. Malheureusement, les méthodes proposées pour s'acquitter de cette tâche sont encore loin de fournir des résultats satisfaisants. La réduction de la polyphonie intrinsèque à un instrument est à un stade plus avancé. Il existe des solutions pour représenter une musique polyphonique par une (ou plusieurs) mélodie(s). Cependant, cette question mériterait d'être approfondie.

L'identification de mélodies redondantes au sein d'une base de données reçoit une attention soutenue. Permettant d'économiser espace mémoire et temps de calcul, ce processus, préalable à la recherche, repose sur la similarité mélodique. Le matériau considéré étant exclusivement MIDI,

[20] Il s'agit de la hauteur, la durée, et l'instant d'apparition

des informations comme le tempo sont disponibles. Autre avantage pour une comparaison mélodique de qualité, les mélodies dont sont extraites des portions candidates au statut de *thème* sont entières. Toute portion mélodique considérée peut donc être décrite au regard des caractéristiques globales de la mélodie dont elle est issue. Cela n’est pas le cas avec une requête soumise à un système de recherche.

Ultime aspect de la gestion des bases de données, l’indexation -ou classement des documents en fonction de leur contenu- est également destinée à accélérer la recherche d’information. Seulement, les structures d’index exploitables sont limitées par le désir de disposer de base de données volumineuses et l’absence de restriction sur les points d’entrée à la recherche. En effet, à la différence des mots d’un texte, séparés par des espaces, les portions mélodiques qu’une requête peut viser peuvent commencer n’importe où dans la base de données. Directement liée à la mesure de similarité employée par le moteur de comparaison, cette question doit elle aussi être approfondie pour aboutir à des solutions exploitables.

Dans les deux dernières sections, nous avons abordé les moteurs de comparaison mélodique ainsi que les données musicales auxquelles il fournit un accès. Nous allons maintenant nous intéresser à un autre élément essentiel d’un système de recherche : la requête.

3.5 La requête

Comme nous l’avons vu Section 2.5.4, la requête d’un système d’indexation mélodique peut prendre plusieurs formes. Le solfège permet l’écriture musicale, cependant on lui préfère généralement la requête *audio*, qui permet d’élargir la population des utilisateurs. En effet, peu de gens ont la capacité d’écrire la musique, par contre, ils sont plus nombreux à savoir jouer d’un instrument, et d’avantage encore à pouvoir chantonner.

Dans le cas d’une requête *audio*, une *transcription* est nécessaire afin de disposer de la mélodie recherchée. Le succès de cette opération n’étant pas évident, des contraintes sont imposées aux utilisateurs afin de faciliter l’analyse automatique de leur requête.

Dans cette section, nous considèrerons uniquement la requête chantonnée car elle constitue le type de requête le plus "ouvert" pour les systèmes d’indexation mélodique. Cela explique que de nombreux système l’aient adoptée [KMT93, GLCS95, SGM98, LL98, KCGV00, MSWH00]. Notons cependant, que certains aspects de la transcription (e.g. segmentation en notes) concernent toute requête acoustique, quel que soit l’instrument utilisé.

3.5.1 Transcription automatique d’une requête chantonnée

A de rares exceptions près, la voix est un instrument monophonique. Un chantonnement est donc susceptible d’être analysé avec succès par les algorithmes de détection de hauteurs actuels. Cependant, cette qualité ne suffit pas à rendre l’extraction de la mélodie-requête possible. En effet, l’information attendue de la transcription est une séquence de notes, chacune d’elle étant qualifiée par sa hauteur, sa durée et son instant d’apparition. Il faut donc définir les paramètres de chaque note pour que l’information mélodique soit complète.

Dans [SGM98] et [MSWH00] par exemple, l’interprétation de la requête acoustique est fondée sur la détection des zones voisées (desquelles on perçoit une hauteur). Comme c’est systématiquement le cas dans les systèmes de transcription utilisés pour la recherche de mélodies par chantonnement, la contrainte suivante est imposée. L'utilisateur doit prononcer une syllabe du type "ti" ou "da" à chaque nouvelle note. La prononciation du "t" ou du "d" entraînant un arrêt de l'émission de hauteur (à la différence d'un "l" ou d'un "m", par exemple), le repérage des limites temporelles des notes successives est facilité[21]. Toute nouvelle note est ainsi précédée d'un passage non voisé.

Dans [MSWH00], la hauteur choisie pour représenter la note correspond à la valeur de hauteur la plus élevée parmi celles (régulièrement) calculées au cours de la zone voisée. Pour la détection des hauteurs, Mc Nab et al. préconisent une méthode temporelle, mais des méthodes spectrales peuplent également la large littérature disponible [Hes83, Nol67, MYC91].

Des post-traitements peuvent améliorer la transcription. Par exemple, Lindsay effectue un lissage du profil de fréquence fondamentale et applique un filtrage par médian pour éliminer les discontinuités souvent dues aux erreurs d'octave commises par la détection de hauteurs.

Le principal défaut des méthodes d'extraction actuellement utilisées est la contrainte qu'elles imposent à l'utilisateur (*ta-ta-ta...*). En effet, celui-ci préfèrerait sans doute pouvoir chanter librement. Il pourrait ainsi s'appuyer sur les paroles de la chanson, imiter le timbre de instrument original, se permettre notes liées, glissandos... Seulement, si ces libertés sont synonymes d'information supplémentaire permettant à l'humain une meilleure identification de la musique recherchée, les moteurs de comparaison actuels ne sont pas prêts à les traiter de manière efficace.

A ce stade, la mélodie qui va être recherchée dans la base de données est définie. Chaque note est caractérisée par sa hauteur, sa durée et son instant d'apparition. La qualité d'une requête acoustique dépend des trois facteurs suivants :

1. qualité de la mémorisation de la mélodie recherchée
2. qualité d'interprétation (justesse, rythme)
3. qualité de la transcription.

Seulement, la mélodie transcrite possède une particularité qui n'a pas été évoquée jusqu'à présent. Il s'agit de l'absence de tempérament de ses hauteurs.

3.5.2 Non tempérament des requêtes chantonnées

La voix emprunte le continuum des fréquences. Ainsi, elle peut emprunter toutes les hauteurs de son ambitus (à la différence d'un piano, par exemple, où la commande par clavier réduit l'espace des fréquences aux seules hauteurs tempérées).

La quantification généralement appliquée aux hauteurs chantonnées ne semble motivée que par la propension à se conformer à la nature des mélodies constituant la base de données. En l'occurrence, cela se traduit par l'adoption de l'échelle tempérée pour la représentation des hauteurs de la requête. Seulement, pour une requête chantonnée, les intervalles tempérés (i.e. quantifiés au demi-ton près) ne correspondent plus au maximum de précision que la description puisse avoir. Par conséquent, ce qui était vrai pour la représentation par contour mélodique (et les contours étendus), l'est aussi pour la description par intervalles tempérés appliquée à la requête chantonnée : la tolérance assurée par une description quantifiée n'est pas équitable (cf. 3.3.3 et 3.3.4).

[21] Le terme *chantonnement* est donc préféré au terme *fredonnement* puisque ce dernier désigne une production vocale bouche *fermée*.

L'absence d'intérêt concernant la quantification des hauteurs de la requête chantonnée est illustrée par le peu de contributions concernant ce point. A notre connaissance, McNab et *al.* sont les seuls à proposer une alternative à la quantification classiquement appliquée, cette dernière consistant en une quantification directe sur les intervalles mesurés (cf. p. 108). Leur méthode repose sur un traitement adaptatif des hauteurs successives [MSWH00]. En effet, à chaque nouvelle note considérée, l'échelle tempérée déterminant la quantification subit un recalage. Ce recalage est fondé sur l'erreur de quantification observée la note précédente. Cette méthode est destinée à suivre un glissement progressif du ton d'une la requête.

Malgré l'influence considérable que la requête peut avoir sur un système de recherche, la requête chantonnée n'a pas fait l'objet d'études nombreuses. Celles-ci sont décrites dans la section suivante.

3.5.3 Caractérisation des mélodies chantonnées

A notre connaissance, seules deux études concernant ce sujet ont été effectuées à ce jour [Lin96, MSWH00]. Les approches de Lindsay et de McNab et *al.* sont différentes, et leurs conclusions concernent des aspects différents des mélodies chantonnées.

Concernant la précision fréquentielle des notes chantées, Lindsay conclue que l'on peut considérer les erreurs commises comme indépendantes de l'intervalle visé. L'imprécision généralement observée sur un intervalle n'augmenterait donc pas avec la taille de ce dernier. De son côté, McNab observe des tendances à la compression des grands intervalles chantés, particulièrement pour des amplitudes de 7 à 9 demi-tons. Par ailleurs, il relève une tendance à l'extension des plus petits intervalles (1 et 2 demi-tons), lorsqu'ils constituent des séquences ascendantes ou descendantes (i.e. intervalles de même signe).

Lindsay note également que les erreurs ne s'accumulent pas : les sujets corrigent leurs erreurs d'une note à l'autre. McNab relève un phénomène contraire, à savoir une dérive progressive du ton de la mélodie. Il recense également une modification brutale et durable du ton de la mélodie. Dans cette thèse, ces trois types d'erreur seront respectivement nommés :
- Erreur Locale, ou EL,
- Glissement de Ton, ou GT,
- Rupture de Ton, ou RT.

Un point fondamental est la différence de stratégies expérimentales adoptées dans les deux publications. Lindsay diffuse des mélodies que ses sujets doivent immédiatement chanter. L'avantage est de pouvoir constituer un corpus de données relativement équilibré. En effet, les mélodies présentées sont telles que, sur l'ensemble du test, les intervalles sont équitablement représentés.

Par contre, les mélodies n'étant pas connues des sujets, Lindsay s'écarte d'un cadre réaliste de production. Or, les différences de précision entre mémoire court terme (qu'il sollicite) et mémoire à long terme des mélodies peuvent influencer significativement les résultats [UZ98]. De plus, les mélodies que présente Lindsay ont un tempo rapide (4 notes par seconde). Les mélodies chantées présentent donc des notes très courtes entraînant généralement une instabilité des fréquences chantées, qui entraîne à son tour une confiance relative dans les hauteurs de notes issues de la transcription. Enfin, les mélodies étant courtes (5 notes), il est compréhensible qu'à la différence de McNab, ni rupture, ni glissement de ton (RT&GT) n'ait été révélé.

La démarche de McNab a l'avantage de rester dans un cadre de production plus réaliste. En effet, il demande à ses sujets de chanter des mélodies populaires dont seul le titre leur est présenté.

Outre les observations concernant les tendances à l'extension et à la compression des intervalles chantés, McNab relève la présence d'insertions/omissions de notes, et également, d'anticipations qui provoquent des différences rythmiques avec la mélodie visée. Le défaut de cette expérience est la faible qualité des données recueillies. Les 91 requêtes du corpus limitent l'expérience dans sa capacité à dégager des observations générales sur les mélodies chantonnées.

Ces deux études n'apportent que peu d'éléments fiables et généraux sur la précision des intervalles chantés. Lindsay s'appuyant sur la mémoire à court terme de ses sujets, ses conclusions sont peut être inadaptées au contexte de la recherche de musique. McNab ne proposant pas de mesures précises, les observations tirées ne sont pas représentatives d'un aspect général des mélodies chantonnées.

L'expérience idéale combinerait les avantages des deux méthodes : d'une part, le cadre d'expérimentation réaliste (les sujets chantent "de mémoire"), et d'autre part, un corpus de données homogène et important. Malheureusement, ces deux qualités sont difficilement compatibles, à moins de trouver des mélodies populaires, dont les intervalles couvriraient de manière homogène, le domaine d'intervalles étudié.

Dans le Chapitre 4, nous proposons une expérience qui tente de s'approcher du cas idéal puisqu'elle reprend la stratégie de McNab, tout en constituant un corpus plus important (500 requêtes).

3.5.4 Conclusion

Parmi les modes de requête possibles, la requête chantonnée est souvent préférée pour l'importante population d'utilisateurs qu'elle autorise. Malheureusement, son traitement par les moteurs de comparaison actuels impose des contraintes lors de sa production (prononciation de *ta-ta-ta...*). Par ailleurs, la nature des mélodies ainsi recueillies n'a encore reçu que peu d'attention. Par conséquents, les moteurs de comparaison manquent d'éléments pour adapter leurs mécanismes aux caractéristiques de la requête attendue. En particulier, l'aspect non tempéré des hauteurs est généralement éliminé par leur quantification au demi-ton près. Or, nous avons vu section 3.3 que ce genre de processus entraînait une mesure de similarité inéquitable.

A notre connaissance, seules deux contributions se sont intéressées à la manière dont des mélodies étaient chantonnées. Les conclusions dégagées motivent la poursuite des recherches afin de d'augmenter les connaissances exploitables dans la conception de moteurs de comparaison.

3.6 Evaluation de systèmes d'indexation mélodique

En section 3.3 (p. 35), nous avons vu qu'il n'était pas simple de qualifier la similarité fournie par des moteurs de comparaison mélodique. Pour juger sa méthode de détection de thèmes, Rolland compare ses résultats à ceux d'une étude musicologique [Rol99]. En effet, la comparaison du jugement d'un moteur avec le jugement humain constitue une excellente voie de qualification d'un système d'indexation. Concernant les systèmes de recherche mélodique, on ne dispose malheureusement pas de liste de mélodies classées par ordre de similarité à une référence particulière. Une solution consiste à juger le comportement du système de recherche complet. Ainsi, les réponses aux requêtes soumises permettent de qualifier les performances du moteur de comparaison.

Dans ce type de démarche, deux éléments sont fondamentaux. Il s'agit du *critère de qualification* choisi et de la *nature des requêtes* choisies pour la stimulation.

Dans [MSWH00], McNab qualifie les réponses des systèmes testés grâce aux critères suivants :

– Nombre de réponses fournies par le système en fonction de la taille de la requête soumise,
– Taille minimum de requête pour qu'une seule mélodie soit retenue, et ce, en fonction de la taille de la base de données.

Ces critères témoignent de la discrimination assurée par le système testé. Or, si ce critère renseigne sur la diversité des mélodies de la base de données[22], nous avons vu qu'il était insuffisant à qualifier les différentes facettes de la similarité (cf. p. 39). De plus, ce critère ne s'applique qu'aux moteurs effectuant une sélection des documents constituant la liste des réponses. D'autres types de moteurs (notamment ceux fondés sur l'utilisation de distances) définissent la liste des réponses par le classement des documents de la base en fonction de leur similarité avec la requête. La liste des réponses fournies n'est donc pas limitée en nombre.

Les critères de précision et de rappel, utilisés depuis longtemps en indexation textuelle, paraissent mieux adaptés au type de performances évaluées. Les formules les définissant sont [Ref [23]] :

$$Précision = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ de\ documents\ retrouvés} \tag{3.2}$$

$$Rappel = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ total\ de\ documents\ pertinents} \tag{3.3}$$

Ces critères sont utilisés dans [UZ99], mais comme nous l'avons déjà vu, la manière dont Uitdenbogerd et Zobel déterminent les documents pertinents nous paraît inadaptée (cf. p. 45).

Downie utilise une version normalisée du critère de précision (NPREC) et détermine les documents pertinents à l'aide de portions mélodiques issues de la base de données. Ces portions sont ensuite perturbées et à nouveau soumises au système testé. Ainsi, ce dernier est stimulé par des requêtes imparfaites visant des documents préalablement identifiés. Le défaut de cette démarche est le manque de réalisme des imperfections générées. Ici encore, le manque de connaissances concernant les requêtes chantonnées se fait ressentir.

Une fois de plus, l'alternative consiste à se rapprocher du contexte réel d'utilisation du système. La collecte de requêtes produites par des sujets permet une stimulation idéale. Kageyama et *al.* constituent un corpus de 100 requêtes [KMT93], et Sonoda et *al.* rassemblent 112 requêtes[24] [SGM98]. Cependant, recherchant uniquement à retrouver le document dont est tiré la portion chantée, ils ne considèrent pas la totalité des contenus similaires à la requête. Dans le cas où un critère de type *précision/rappel* est utilisé, l'identification des documents pertinents demande un lourd investissement. En effet, celle-ci passe par l'identification des mélodies visées par les requêtes collectées (avec caractérisation des erreurs commises), et la création des requêtes parfaites correspondantes.

Face à l'absence de critères objectifs permettant de juger de la qualité de moteurs de comparaison, ce sont les systèmes entiers qui font l'objet d'évaluations. Après quelques tentatives peu

[22] En effet, plus les mélodies sont originales (i.e. différentes les unes des autres), plus le nombre de notes nécessaire à leur identification est faible.
[23] Salton/McGill
[24] respectivement produites par 10 et 12 sujets.

adaptées à la réelle qualification de la similarité fournie, l'évaluation des performances se dirige vers l'utilisation de critères de type *précision/rappel*. Seulement, la stimulation des systèmes pose problème. L'utilisation de requêtes réelles demande un fort investissement humain. Certes, des requêtes artificielles peuvent leur être substituées, mais le manque de réalisme des imperfections simulées biaise les performances mesurées. Ce constat est un élément supplémentaire encourageant l'étude des requêtes chantonnées.

3.7 Systèmes disponibles en ligne

Quelques systèmes de recherche de documents musicaux sont disponibles sur Internet. A notre connaissance, seuls deux autorisent une requête acoustique. Le site *Tuneserver* semble fonder son moteur de comparaison sur le simple contour mélodique [25]. Plus de 10.000 documents sont accessibles.

Le système de recherche de la "New Zealand Digital Library" de l'université de Waikato[26] est plus évolué. En effet, le système MELDEX (MELody inDEX) proposé par McNab et al. permet de choisir le type de comparaison effectué, en combinant les options suivantes :

- transcription adaptative ou non de la requête (cf. 3.5.2),
- représentation par contour mélodique (*UDS*) ou bien par intervalles tempérés,
- représentation avec ou sans prise en compte du rythme (requiert des compétences musicales),
- comparaison exacte ou approximative (transpositions, insertions, omissions admises),
- recherche au début de chaque chanson ou bien sur l'ensemble de la base de données.

Il s'agit du programme le plus complet disponible en ligne. La base de données est constituée de 10.000 thèmes de musique traditionnelle de différentes cultures. La requête s'effectue par un fichier son de musique monophonique. Celui-ci est analysé automatiquement, l'utilisateur pouvant écouter le résultat (et éventuellement recommencer sa requête). La possibilité de soumettre une requête sans avoir à la compléter par des informations fait de MELDEX un système adapté à un large public.

Deux autres sites parmi ceux consacrés à la recherche de mélodie par le contenu sont présentés ici. L'un est une démonstration du *Themefinder* de l'Université de Stanford[27]. Ce système permet la recherche de thèmes de musique classique (les résultats sont retournés sous forme de partitions). La base de données est pour l'instant uniquement constituée de partitions issues des œuvres de Ludwig Van Beethoven, mais nous en ignorons le nombre. La requête s'effectue textuellement par des informations du type mélodie (hauteurs uniquement), métrique, tonalité, intervalles, et différents contours mélodiques. Ce système, conçu pour les musicologues, requiert des connaissances musicales, et n'est donc pas adapté au grand public.

L'autre site, émanant de l'Université de Southampton, propose un service de recherche de fichiers MIDI appelé *Search By Humming* [28]. La base de données est composée de 180 fichiers. Contrairement à ce qu'indique le nom du système, la requête s'effectue directement sous forme des contours (de 1er ordre et de 2d ordre). Un algorithme de *string-matching* flexible permet de

[25]http ://tuneserver.de/
[26]http ://www.nzdl.org/
[27]http ://musedata.stanford.edu
[28]http ://audio.ecs.soton.ac.uk/sbh/

gérer l'insertion, l'omission et la substitution de symboles. Cependant, ce système n'est pas adapté au grand public puisque l'utilisateur doit lui-même produire les contours représentant sa mélodie.

Quelques systèmes de recherche sont donc disponibles en ligne. Assez représentatifs de l'état de l'art, ils témoignent du fort désir d'aboutir rapidement à des systèmes opérationnels. Le manque de maturité des technologies actuelles n'empêche donc pas la mise à disposition du public.

3.8 Conclusion

L'indexation mélodique fait l'objet d'une intensification des recherches récente, motivée par la place importance qu'occupe la musique aujourd'hui. Le moteur de comparaison, témoignant de la similarité mélodique, peine à concilier qualité de jugement et simplicité d'utilisation. Le difficile compromis *précision/tolérance* est géré par la répartition des propriétés entre descripteur et mesure de similarité (couple constituant le moteur de comparaison). La tendance générale est à la quantification des mélodies afin de les représenter comme une succession d'états et ainsi profiter des techniques d'appariement de chaînes de caractères (*string-matching*). Malheureusement, cette voie témoigne inéquitablement de la similarité mélodique.

Les bases de données musicales concernées par les systèmes développés ne peuvent contenir des documents musicaux usuels (e.g. MP3). En effet, les outils de transcription actuels ne permettent pas d'accéder aux notes jouées par les divers instruments intervenant dans une musique polyphonique quelconque. Les bases de données utilisées sont donc constituées de fichiers de musique synthétique (MIDI) qui fournissent l'information désirée ($\approx$ la partition).

Nombre de questions concernant la description de ces données doivent encore être approfondies. Quelle(s) mélodie(s) peu(ven)t représenter fidèlement une musique polyphonique ? Comment identifier des contenus musicaux tels que *mélodie*, *accompagnement* ? Comment organiser les descripteurs pour assurer un temps de recherche minimum ? La détection de thèmes redondants, permettant de réduire l'espace de recherche, semble plus avancée.

Concernant la manière d'effectuer une recherche, la requête chantonnée s'impose grâce à la large population d'utilisateurs qu'elle autorise. La faible quantité d'études la concernant montre que, paradoxalement, elle n'a pas encore reçu l'attention qu'elle mérite compte-tenu de son importance.

Un système d'indexation mélodique est constitué d'éléments dont l'imbrication rend difficile une appréhension globale du problème. De plus, les différents domaines concernés n'ont pas encore la maturité suffisante pour que les solutions proposées soient satisfaisantes en terme d'efficacité et de convivialité.

Dans le Chapitre 4, nous participerons à l'enrichissement des connaissances sur les mélodies chantonnées. Dans le Chapitre 6, nous tenterons de dégager des éléments objectifs sur la similarité mélodique. En effet, ceux-ci manquent cruellement pour le jugement de la qualité de moteurs de comparaison. Ces contributions seront appliquées aux moteurs de comparaison proposés au Chapitre 5. Les différents stades de leur conception seront détaillés, afin de bien poser les alternatives possibles, les conséquences des choix effectués, ainsi que les hypothèses sur lesquelles ils reposent. Les chapitres restants (Chapitres 7 et 8) concerneront l'évaluation de la qualité de systèmes de recherche mélodiques.

Chapitre 4

Etude de Requêtes Chantonnées

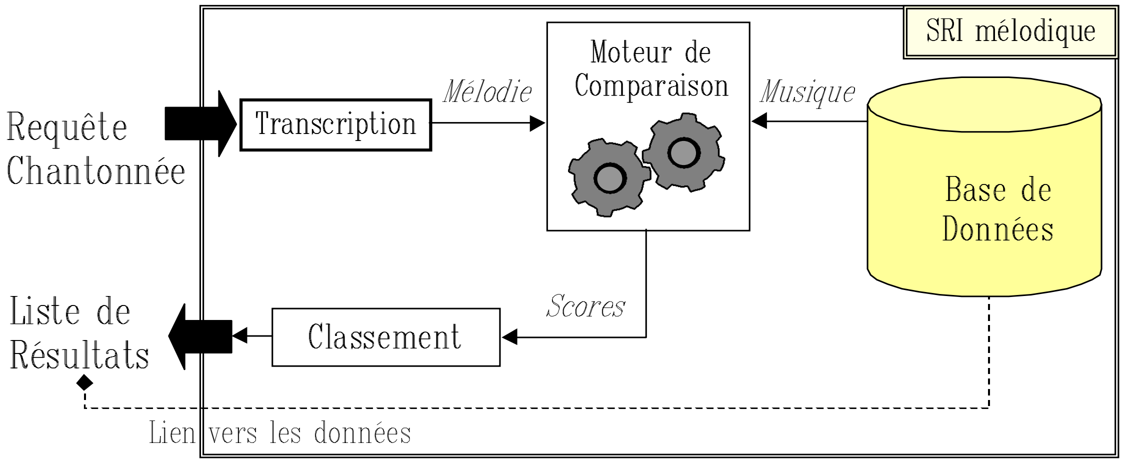


figure 4.1 : *Système de recherche de documents musicaux par chantonnement.*

4.1 Introduction

La requête constitue l'information reçue par le système de recherche. Malgré sa participation naturelle aux systèmes de recherche mélodique dédiés au grand public, la requête chantonnée n'a pas fait l'objet d'études nombreuses. Comme nous l'avons vu en 3.5.3, les contributions sur le sujet sont peu nombreuses. Ce chapitre est consacré à l'amélioration des connaissances sur la requête chantonnée.

Nous commencerons par décrire le processus de transcription que nous utilisons dans notre système de recherche. Extrayant les notes du flux acoustique contenant la requête chantonnée, celui-ci nous permettra d'étudier les mélodies-requêtes fournies au moteur de comparaison. Nous limiterons nos observations aux seules caractéristiques fréquentielles des mélodies, en particulier, aux hauteurs de notes chantées. En effet, nous verrons que le caractère non tempéré de la voix motive cette étude.

4.2 Transcription automatique d'une mélodie chantonnée

La transcription d'une mélodie chantonnée consiste à traiter le signal acoustique capté par le microphone afin d'en extraire les notes chantonnées par l'utilisateur. La détermination des notes implique deux actions. La segmentation en notes définit les instants de début et de fin des notes. L'étiquetage consiste à assigner une hauteur à chacune des notes segmentées.

Nous allons voir que, dans notre système, ces deux actions s'appuient sur un traitement unique : la détection de fréquences fondamentales.

4.2.1 Détection de fréquences fondamentales

Puisque la requête consiste en une suite de notes dont on perçoit la hauteur, il est naturel de s'intéresser à la présence de fréquences fondamentales au sein du signal.

L'algorithme *eSRPD* [1] détecte les passages voisés et non-voisés d'un signal vocal [Bag94]. Les premiers correspondent à la présence d'une fréquence fondamentale, les seconds, non. La méthode utilisée est adaptée au suivi d'intonations de la voix parlée. Elle a été testée par comparaison de ses résultats avec des signaux issus de l'observation directe de l'activité laryngée[2]. Ses performances en font un outil efficace pour la détection de passages voisés, et le suivi des variations de fréquence fondamentale.

Les fréquences fondamentales produites par la voix chantée occupent un domaine de fréquence plus large que celles généralement produites par la voix parlée. Nous avons cependant conservé une zone de fonctionnement basse et relativement restreinte (40-450Hz) afin de ne pas sortir de la configuration dans laquelle les bonnes performances de l'algorithme ont été constatées. Le module de transcription assure un traitement correct des hauteurs inférieures ou égales au *sol♯4*. Ce point devrait être amélioré dans une perspective de système commercial.

Toutes les 5ms [3], l'algorithme *eSRPD* fournit une information sur le voisement du flux acoustique. Deux cas sont distingués :
- cas non-voisé : la valeur nulle est fournie (aucune hauteur n'a été détectée),
- cas voisé : la valeur de la fréquence fondamentale estimée est fournie.

Les valeurs de fréquences fondamentales ne sont pas directement utilisées pour déterminer la mélodie chantée par l'utilisateur. En effet, chaque passage voisé (i.e. séquence de valeurs non nulles consécutives) subit un traitement. Dans un premier temps, une élimination des ambiguïtés d'octave est opérée, puis un lissage est appliqué afin d'éliminer les fines perturbations non significatives. Ainsi, les valeurs de fréquences fondamentales finalement fournies sont exemptes de discontinuités (problème évoqué dans [Lin96]). La figure 4.4 (courbe grisée) illustre le type de résultat obtenu.

[1] *enhanced version of Super Resolution Pitch Determinator*
[2] Le larynx joue un rôle essentiel dans l'émission de la voix.
[3] 5ms correspondent au pas d'avancement de la fenêtre d'analyse (de taille 20ms).

4.2.2 Segmentation en notes

Notre segmentation en notes repose sur l'hypothèse qu'à chaque zone de voisement correspond une note. Cependant, un passage voisé n'est pris en compte que s'il dure 40ms au minimum. Ce seuil, ajusté empiriquement, est le résultat d'un compromis. En effet, nous voulons permettre à un utilisateur de produire des notes courtes, mais nous voulons éviter que des bruits intempestifs viennent ajouter des notes indésirables.

Une contrainte (classique) est, par conséquent, imposée à l'utilisateur : celui-ci doit obligatoirement faire figurer un passage non voisé entre deux notes afin que celles-ci ne soient pas fusionnées. Il peut donc insérer volontairement des silences entre ses notes, ou encore prononcer "ta" -ou toute autre syllabe commençant par un "t"- à chaque nouvelle note. La prononciation du "t", incompatible avec l'émission d'une hauteur, entraîne l'absence de voisement qui facilite la délimitation temporelle des notes.

Les figures 4.2 et 4.3 illustrent un signal de requête et sa segmentation en notes.

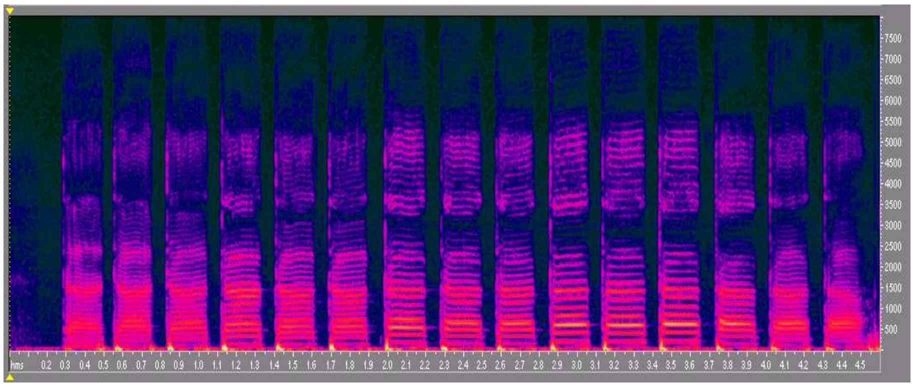


figure 4.2 : *Spectrogramme d'une requête chantonnée.*

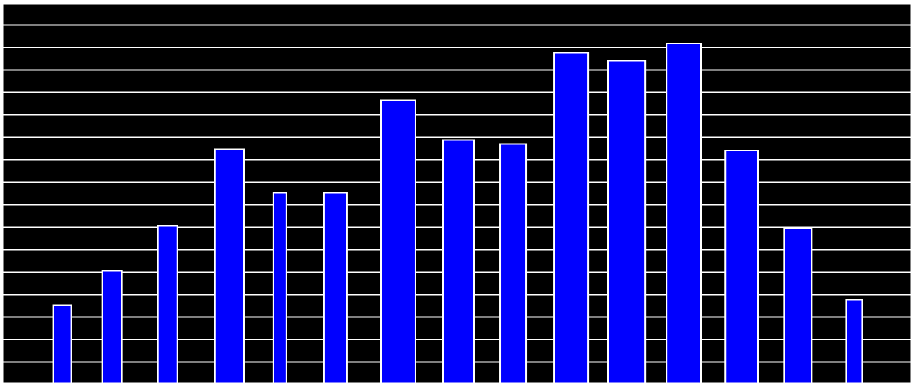


figure 4.3 : *Segmentation en notes de la requête chantonnée dont le spectrogramme est présenté figure 4.2. Chaque note détectée est illustrée par une colonne, dont la largeur est proportionnelle à la durée détectée.*

A ce stade, les limites temporelles des notes sont fixées. Il reste donc à définir la hauteur de chacune d'elles.

4.2.3 Etiquetage des hauteurs

L'information fréquentielle de chaque note est constituée d'une séquence de fréquences fondamentales temporellement espacées de 5ms. La tâche à accomplir consiste à déterminer la fréquence fondamentale *unique* qui va fournir sa hauteur à la note.

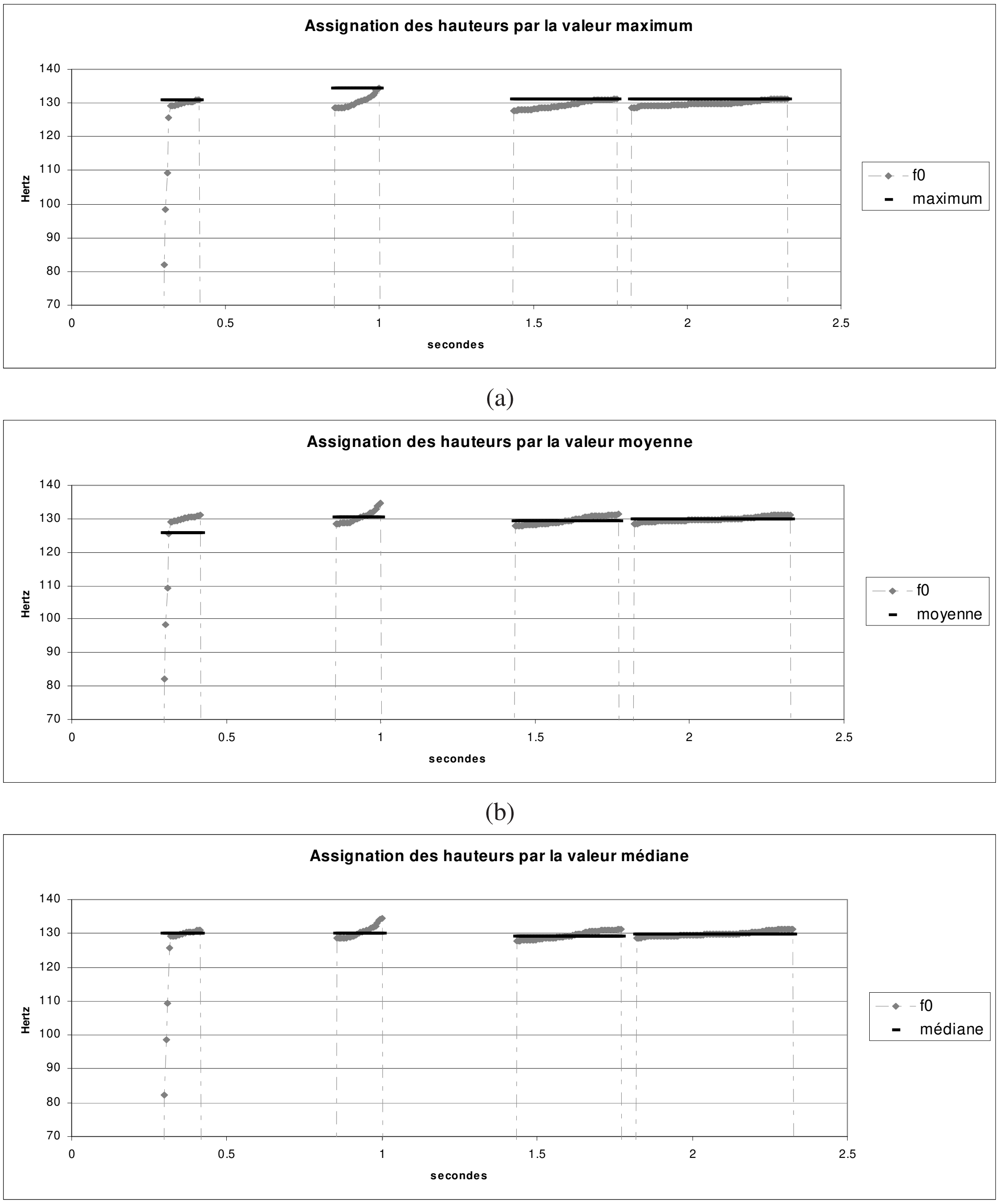
L'observation des séquences de fréquences fondamentales extraites nous a poussé, à l'instar de Lindsay (cf. 3.5.1), à élire le médian de la séquence pour représenter cette dernière. En effet, nous avons observé que les valeurs extraites pouvaient varier sensiblement au cours d'une note (i.e. d'un passage voisé). La Figure 4.4 illustre le résultat de trois méthodes d'étiquetage différentes. Les séquences de fréquences fondamentales extraites (f0) sont superposées aux valeurs représentant la hauteur des notes détectées. Le signal à transcrire contient une mélodie de quatre notes dont les hauteurs sont perçues comme identiques. Sur cet exemple, le médian montre sa capacité à faire abstraction de valeurs perturbatrices.

Cette illustration est complétée par le tableau 4.1. Celui-ci présente les valeurs de fréquence fondamentale fournies par les trois méthodes que nous venons d'illustrer. N1, N2, N3 et N4 désignent les quatre notes composant la mélodie. Les intervalles, témoignant des différences entre hauteurs successives, sont également présentés. Ils sont calculés conformément à la formule 3.1 (p. 28).

Méthode	f0/Intervalles	N1	N2	N3	N4
Maximum	f0	131.0 Hz	134.6 Hz	131.3 Hz	131.3 Hz
	Intervalles	-	+0.47 demi-ton	-0.43 demi-ton	0.00 demi-ton
Moyenne	f0	125.7 Hz	130.4 Hz	129.5 Hz	129.9 Hz
	Intervalles	-	+0.64 demi-ton	-0.12 demi-ton	+0.05 demi-ton
Médian	f0	130.0 Hz	130.1 Hz	129.3 Hz	129.7 Hz
	Intervalles	-	+0.01 demi-ton	-0.11 demi-ton	+0.05 demi-ton

TAB. 4.1 : *Résultats de l'assignation d'une fréquence fondamentale unique à chaque note selon 3 méthodes différentes. La mélodie chantonnée est une succession de notes dont la hauteur est perçue comme* identique.

Comme on peut le voir, les intervalles issus des hauteurs extraites peuvent être, selon la méthode employée, assez différents. Malgré la constance des hauteurs *perçues*, certains intervalles sont proches du quart-de-ton (+0.47 et -0.43 avec la méthode du maximum), voire le dépassent largement (+0.64 avec la méthode de la moyenne).



(a)

(b)

(c)

figure 4.4 : *Assignations de la hauteur des notes selon la valeur maximum (a), moyenne (b), médiane (c).*

Afin de faciliter la comparaison des hauteurs entre elles, les fréquences fondamentales peuvent être exprimées en demi-ton. La numérotation MIDI, qui suit cette échelle, sert de base à l'étiquetage des hauteurs issues de la transcription.

Le tableau 4.2 présente le résultat de la conversion des fréquences fondamentales médianes.

Comme l'illustre cet exemple, l'utilisateur n'ayant pas de repère de hauteur sur lequel appuyer sa requête, les notes chantées sont généralement non tempérées, les valeurs obtenues sont, par conséquent, non entières.

	N1	N2	N3	N4
f0 (hertz)	130.0	130.1	129.3	129.7
Hauteurs (demi-ton)	35.89	35.91	35.80	35.85

TAB. 4.2 : *Conversion des fréquences fondamentales élues en hauteurs (non-tempérées).*

4.2.4 Interface de visualisation du module de transcription

Nous avons implémenté une interface permettant d'observer le résultat de la transcription d'une requête. La figure 4.5 présente la fenêtre disponible après analyse du signal. La mélodie chantée correspond au début de "La Marseillaise".

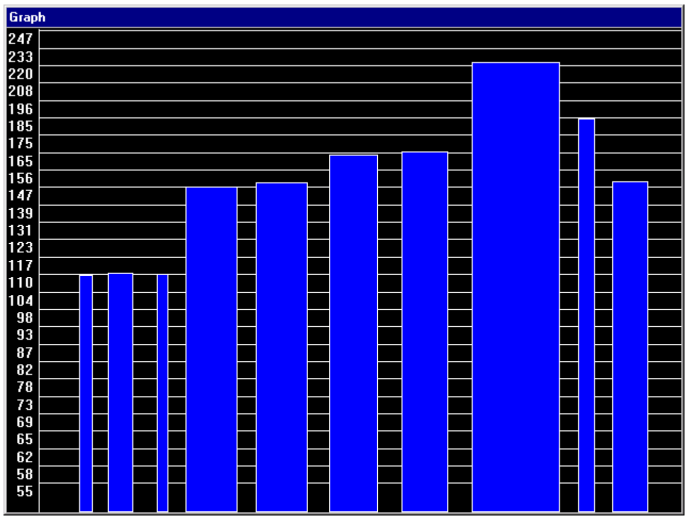


figure 4.5 : *Fenêtre de visualisation de la requête transcrite. Dans ce cas, il s'agit des 10 premières notes de "La Marseillaise" : "A-llons en-fants de la pa-tri-i-e".*

Il s'agit d'une représentation temps/fréquence de la mélodie transcrite. Chaque note détectée est représentée par un pavé. La largeur de ce dernier témoigne de la durée de la note, et sa hauteur témoigne de la fréquence fondamentale élue. L'échelle fréquentielle est logarithmique afin de s'adapter à la perception humaine fondée sur l'octave. Elle est découpée en zones d'un demi-ton, chacun trait étant associé à sa valeur en hertz. Cette double indication permet à la fois de situer la hauteur générale de la requête, et également d'estimer la valeur des intervalles chantés. Dans cet exemple, la requête se situe entre 110 et 230Hz, et la séquence des intervalles chantés est approximativement [0 , 0 , +5 , 0 , +2 , 0 , +5 , -3 , -4].

4.2.5 Conclusion

A la sortie du module de transcription, nous disposons donc d'une séquence de notes, chacune d'entre elles étant définie par :
- ses instants de début et de fin
- sa hauteur

Notre méthode de transcription de la mélodie ne se démarque donc pas sensiblement de l'état de l'art (cf. 3.5.1). En effet, elle contraint elle aussi l'utilisateur dans sa manière de produire sa requête.

Son principal défaut concerne l'insertion et l'omission de notes. En effet, il arrive, avec certains sujets en particulier, que des notes soient scindées en deux. L'interruption prend généralement place en début ou fin de note, ce qui produit l'insertion d'une note de courte durée. En l'état actuel de notre module de transcription, ce phénomène est difficile à éradiquer puisque, par ailleurs, nous désirons permettre la transcription de notes courtes.

Une piste d'amélioration serait l'observation du contexte d'apparition des notes courtes afin d'identifier les scissions indésirables. La prise en compte de la proximité des notes voisines pourrait être un indicateur utile.

Maintenant que nous disposons de l'information mélodique issue de la requête, nous allons pouvoir en étudier certaines caractéristiques.

4.3 Analyse de mélodies chantonnées

Dans cette section, nous allons étudier certaines caractéristiques des mélodies telles qu'elles sont transmises au moteur de comparaison (i.e. telles qu'elles sont fournies par le module de transcription). Les différences présentes entre mélodies transcrites et mélodies visées peuvent avoir des causes diverses puisque, comme nous l'avons déjà vu en 3.5.1, le processus amont est triple : *souvenir d'une musique - production vocale - transcription automatique*.

4.3.1 Références et critère de comparaison

Pour juger de la qualité d'une mélodie chantonnée, il faut pouvoir la comparer à la mélodie visée. La première tâche consiste donc à identifier la partition qui servira de référence. Ce rôle est généralement joué par une portion mélodique issue de la base de données.

Détermination de l'information observée

Les différences de tempo et de ton n'étant pas considérées comme perturbant l'information mélodique, il faut s'affranchir de leur influence. D'avantage intéressés par l'aspect fréquentiel des mélodies chantonnées, nous laissons de côté leur aspect temporel. Nous considèrerons donc les séquences de hauteurs, en dehors du tempo et de toute autre information rythmique.

Les éventuelles insertions/omissions de notes entraîneront une adaptation manuelle de la référence. La priorité donnée à l'imprécision fréquentielle des notes chantées est motivée par le caractère *inévitable* du phénomène. Les insertions/omissions de notes sont certes, non négligeables, mais également non systématiques.

Concernant le ton, nous pourrions envisager d'ajuster les deux mélodies (chantonnée et référence) grâce à leur ton respectif afin d'éliminer l'éventuelle différence. Cependant, comme nous le verrons en détail au Chapitre 5, l'extraction du ton d'une mélodie n'est pas une question évidente. Par conséquent, afin de ne pas ajouter un biais supplémentaire aux données étudiées, nous renonçons à cette voie. Ce refus est possible car, comme nous l'avons déjà vu, l'observation des intervalles d'une mélodie permet de s'affranchir du ton utilisé. Ce choix rejoint celui de McNab et également de Lindsay, bien que l'expérience de ce dernier (i.e. faire répéter des mélodies à ses sujets) aurait peut être pu lui permettre d'éviter le problème.

L'observation de l'imprécision fréquentielle des notes chantées se fondera sur les *intervalles* séparant les hauteurs successives.

Adaptation des références

L'adaptation des références est effectuée "à l'oreille". Une écoute de la requête chantonnée dicte la modification de la référence en fonction des notes ajoutées, omises, mais également transposées. Concernant les transpositions[4], l'ajustement de la référence a pour but d'éviter de prendre pour une forte imprécision, ce qui n'est, en réalité, qu'une hauteur visée différente de celle de la référence initiale.

La figure 4.6 illustre le début de quatre références concernant une même chanson[5]. Les notes dont la hauteur diffère par rapport à la première référence sont surmontées d'un petit carré (leur

[4]Une transposition est un changement de hauteur de note.

[5]Il s'agit de la chanson "Les portes du pénitencier" (J. Hallyday), également connue dans sa version originale anglo-saxonne "The house of the rising sun".

valeur est écrite en gras dans le tableau 4.3).

figure 4.6 : *Exemple de références multiples issues des requêtes visant une chanson identique.*

Référence	Numéro MIDI des hauteurs de notes							
1	67	67	67	70	74	72	67	67
2	67	67	67	70	74	72	67	**70**
3	67	67	**69**	70	74	72	67	67
4	67	67	**69**	70	74	72	67	**70**

TAB. 4.3 : *Hauteurs correspondant aux références multiple présentées figure 4.6.*

L'adaptation des références est un processus *lourd*, mais nécessaire à la collecte d'une grande quantité de requêtes qui ne contraigne pas trop les sujets. Cet effort permet de rester dans un cadre réaliste d'utilisation du genre de systèmes auxquels nous nous intéressons, ce qui n'était pas le cas des expériences de Lindsay.

4.3.2 Constitution du corpus de données

Neuf sujets (dont l'auteur) ont participé à la constitution du corpus de données. Les candidats ont été choisis afin de couvrir un large spectre de pratique musicale (instrumentale et vocale). Celle-ci est détaillée, pour chaque sujet, en Annexe A.

Principe

Pendant son passage, le sujet doit chantonner (avec un "ta" à chaque note) des mélodies appartenant à une liste de titres de musique populaire. Cette liste, composée de 21 chansons et "jingles" publicitaires, est présentée tableau 4.4.

Titre	Nombre de requêtes chantées	Taille de requête moyenne
Prendre un enfant (Y. Duteil)	26	12.6 notes
Amsterdam (J. Brel)	20	15.9 notes
La Marseillaise (C.J. Rouget de Lisle)	27	12.6 notes
Les portes du pénitencier (J. Hallyday)	17	14.6 notes
Frère Jacques (trad.)	27	13.4 notes
Au clair de la lune (trad.)	27	13.0 notes
Petit Papa Noël (trad.)	27	12.9 notes
A la claire fontaine (trad.)	26	12.6 notes
Noir, c'est noir (J. Hallyday)	18	9.0 notes
Les sucettes (F. Gall)	18	12 notes
Yellow submarine (Beatles)	25	16.8 notes
Yesterday (Beatles)	18	12.1 notes
Cadet roussel (trad.)	20	17.2 notes
La mer (C. Trenêt)	24	10.3 notes
La vie en rose (E. Piaf)	24	14.2 notes
Le petit vin blanc (trad.)	27	12.7 notes
La rirette (trad.)	24	15.9 notes
France Télécom (jingle)	27	7 notes
Bouygues Télécom (jingle)	24	5 notes
Lu (jingle)	27	4 notes
Direct Assurance (jingle)	27	5 notes

TAB. 4.4 : *Nombre de requêtes et taille moyenne pour les différents motifs proposés.*

Afin d'augmenter la taille du corpus, chaque sujet doit produire trois réalisations de chaque élément connu de la liste (donc 63 requêtes au maximum par sujet). Sachant que deux requêtes ne sont jamais strictement identiques, cette tactique permet d'augmenter la variété des intervalles collectés. Cependant, cette voie ne renouvelle pas les contextes auxquels ils appartiennent. Ce point est important puisqu'il n'est pas exclu que les notes entourant un intervalle aient une influence sur la manière dont ce dernier est chanté.

Cette procédure d'expérimentation possède l'avantage de conserver un cadre réaliste d'utilisation de systèmes de recherche de musique par chantonnement. Elle fait appel à la mémoire à long-terme des sujet, au contraire de Lindsay, qui sollicitait leur mémoire à court terme.

L'amélioration par rapport à la procédure de McNab, est la constitution d'un corpus de données important (500 requêtes contre 91), permettant une étude plus générale.

Les musiques proposées aux sujets sont différentes de celles mises en jeu dans les expériences précédemment citées. Puisque l'on ne connaît pas l'influence du contexte mélodique sur la précision du chant, cette différence peut entraîner un biais dans les observations effectuées.

Cependant, le fait que les musiques en question soient populaires et proviennent d'une même culture musicale (occidentale) permet de relativiser cet éventuel inconvénient. En effet, ces deux arguments peuvent laisser penser que les musiques proposées possèdent, localement, de fortes ressemblances (gammes parcourues, résolutions...).

Bien que la taille de notre corpus nous permette une étude plus globale que celle de McNab, un inconvénient fondamental demeure : les différents intervalles n'y sont pas représentés équitablement. Certains, plus rarement usités (e.g. $\pm$ 6 demi-tons, cf. figure 5.12 page 115), ne sont même pas représentés. Nos observations seront donc incomplètes et d'une précision variable.

Cette expérience nous permettra néanmoins d'améliorer la connaissance des requêtes chantonnées, en généralisant certaines conclusions des études précédentes, et en en apportant de nouvelles.

Remarques sur la collecte des requêtes

1. Certains sujets ont signifié leur gêne d'avoir à substituer la syllabe "ta" aux paroles des chansons. Cette observation encourage à trouver des solutions autorisant des requêtes plus libres.

2. Par ailleurs, la quasi-totalité des sujets a été incapable de se souvenir de la mélodie des "jingles" publicitaires, présentés par leur marque associée. Les mélodies leur ont donc été diffusées afin qu'ils se rendent compte s'ils les connaissaient ou pas. Si cela était le cas, ils produisaient leurs requêtes. Ainsi, chacune des requêtes constituant notre corpus est bien issue de la mémoire à long terme des sujets.

   Lorsque ce cas s'est présenté, les sujets ont spontanément tenté d'adopter la tonalité de la mélodie diffusée. Ce mimétisme les a généralement entraînés hors de la zone des hauteurs qu'ils pouvaient chanter confortablement.

   Cette remarque illustre le biais que peut introduire la procédure expérimentale qui consiste à faire répéter des mélodies diffusées. Afin de les replacer dans un contexte plus réaliste d'utilisation de systèmes de recherche de musique par chantonnement, les sujets ont été encouragés à trouver un ton qui leur convenait.

3. Enfin, nous avons remarqué que le début des requêtes (en particulier le premier intervalle) pouvait présenter une certaine imprécision. En effet, chanter n'est pas forcément une action facile, la voix peut donc manquer de stabilité au début de l'émission. Nous verrons par la suite si une imprécision supérieure en début de requête confirme cette observation.

Propriétés générales du corpus

Taille des requêtes

La taille moyenne des requêtes collectées est de 11.7 notes (écart type = 5). En excluant les "jingles" dont la taille est fixe, les requêtes de taille dite libre, ont une taille moyenne de 13.4 notes (écart type = 4.2). Le détail des requêtes produites est présenté Annexe A. On y trouvera le nombre de requêtes et taille moyenne pour les différents motifs proposés, ainsi que la taille moyenne des requêtes "libre" pour chacun des sujets.

Répartition des intervalles visés

Les 500 mélodies-requêtes collectées comportent 5845 notes, constituant donc 5345 intervalles. La répartition des intervalles visés est présentée figure 4.7.

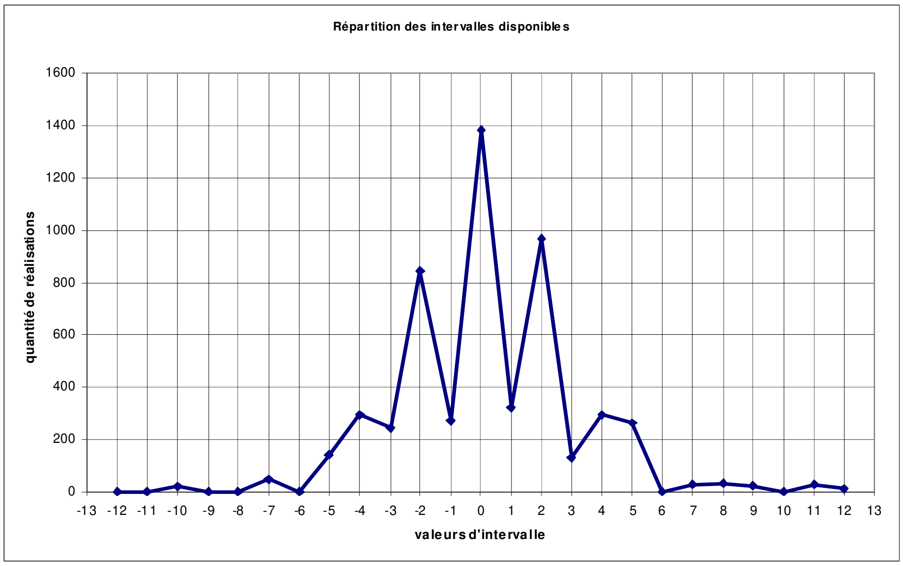


figure 4.7 : *Répartition des intervalles visés par les requêtes collectées.*

Comme on peut le voir, les intervalles d'amplitude supérieure à 5 demi-tons sont faiblement représentés (voire absents, pour certains). Les contextes mélodiques dans lesquels ils apparaissent sont donc certainement moins variés. Par conséquent, nous garderons à l'esprit que les résultats les concernant puissent être l'expression de cas particuliers.

Dans la suite, nous allons procéder à l'analyse de l'imprécision des 5345 intervalles chantonnés constituant notre corpus. Avant de procéder à une étude quantitative des erreurs collectées, nous allons voir si notre corpus témoigne de certains types d'erreurs recensés dans les études passées.

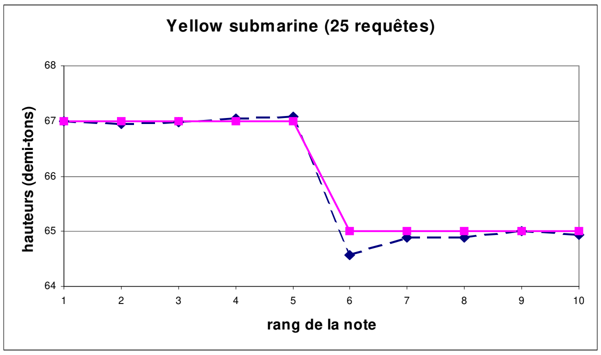
Convention : Les erreurs observées correspondent à la valeur de l'intervalle chanté à laquelle on a retranché la valeur de l'intervalle visé.

4.3.3 Types d'erreurs recensés

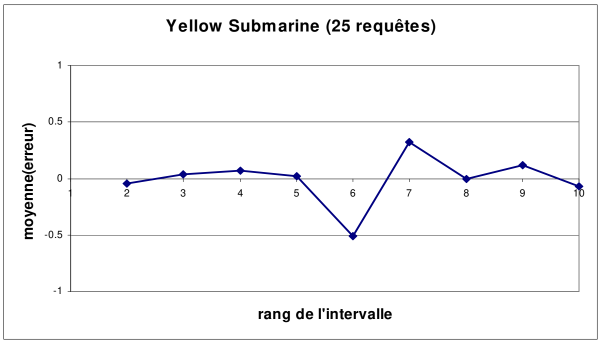
Trois types d'erreurs ont été recensées par McNab et Lindsay. Nous en avons constaté l'existence au sein de notre corpus. Les exemples que nous allons citer seront illustrés à la fois par les hauteurs (ajustement manuel, tonalité non significative) et les erreurs d'intervalles. Les premières permettront d'appréhender simplement les phénomènes traités, les secondes nous permettront d'en observer les conséquences sur les données effectivement étudiées dans ce chapitre.

Erreur Locale - EL

Considérons l'erreur moyenne des 25 requêtes visant une portion du motif "Yellow Submarine". La figure 4.8 présente les erreurs effectuées sous les deux formes annoncées (hauteurs successives (a) et intervalles successifs (b)).



(a)



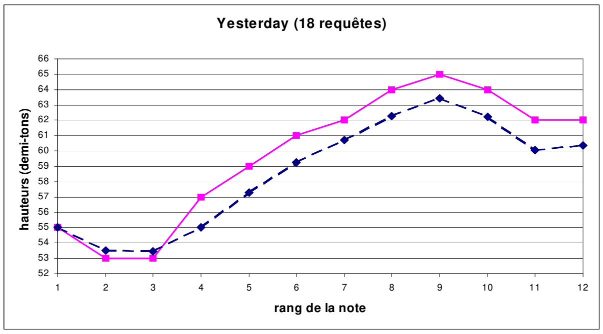
(b)

figure 4.8 : *Erreurs moyennes sur une portion du motif "YellowSubmarine" - Illustration d'une EL sur (a) les hauteurs (référence = trait plein) ; (b) les intervalles (référence = 0).*

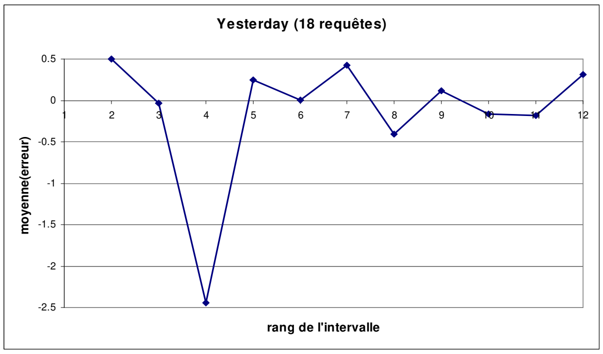
Il apparaît que l'écart observable sur la 6ème hauteur de la figure (a) entraîne deux erreurs consécutives à la fois importantes et opposées sur la figure (b). L'intervalle d'abscisse 6 comporte une erreur moyenne de -0.51. Celle-ci est entraînée par une large majorité (87%) d'erreurs négatives (de moyenne -0.60). Au contraire, l'intervalle suivant comporte une erreur moyenne de 0.32, entraînée par une large majorité (80%) d'erreurs positives (de moyenne 0.46). Cette opposition traduit le phénomène de correction immédiate définissant l'EL. Cette interprétation est appuyée par le fait que les intervalles contigus au phénomène (i.e. intervalles d'abscisse 5 et 8) comportent des erreurs de moyennes quasi-nulles.

Rupture de Ton - RT

Considérons à nouveau l'erreur moyenne en fonction du rang de l'intervalle au sein d'un motif visé. La figure 4.9 présente les erreurs moyennées sur les 18 requêtes visant le motif "Yesterday".



(a)



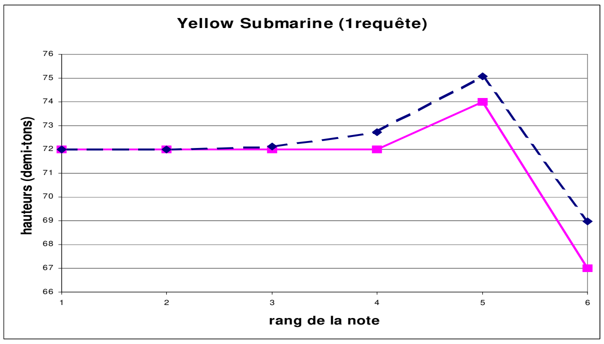
(b)

figure 4.9 : *Erreurs moyennes sur une portion du motif "Yesterday" - Illustration d'une RT sur (a) les hauteurs (référence = trait plein) ; (b) les intervalles (référence = 0).*

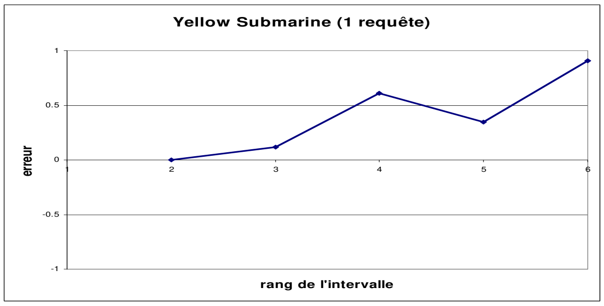
Contrairement au cas précédent, l'écart important observable sur la 4ème hauteur de la figure (a) n'est pas rattrapé sur la 5ème. Cette erreur non compensée entraîne une erreur forte et *isolée* sur la figure (b). En effet, l'intervalle d'abscisse 4 comporte une erreur moyenne de -2.44 demi-tons ! Celle-ci est entraînée par une large majorité (89%) d'erreurs négatives (de moyenne -2.82). L'intervalle visé étant positif, cette valeur traduit une nette tendance à la compression (l'intervalle chanté est plus petit que l'intervalle attendu). En fait, l'amplitude anormalement élevée de cette erreur traduit un brusque changement de ton de la part des sujets. Ce changement n'étant pas rectifié par la suite (absence d'erreur d'amplitude comparable et de signe opposé au rang suivant), il constitue une RT.

Glissement de Ton - GT

Deux contextes mélodiques particuliers (appartenant aux motifs "Yellow Submarine" et "Yesterday") nous ont permis de révéler EL et RT sur les erreurs produites par l'*ensemble* des sujets. Afin d'illustrer le phénomène du GT d'une manière flagrante, nous considèrerons les erreurs d'une requête seulement. Nous verrons par la suite, qu'à l'instar des EL et RT, certains contextes mélodiques favorisent un GT observable sur l'ensemble des sujets. La figure 4.10 présente l'imprécision issue d'une réalisation particulière du motif "Yellow Submarine".



(a)



(b)

figure 4.10 : *Erreurs sur une portion du motif "YellowSubmarine" - Illustration d'un GT sur (a) les hauteurs (référence = trait plein) ; (b) les intervalles (référence = 0).*

Nous voyons sur (a) que plus le temps passe, plus la requête s'écarte de la référence visée. Cet éloignement progressif de la requête vis-à-vis du motif mélodique visé caractérise le GT. Cette dérive se traduit sur (b) par la présence d'erreurs consécutives de signes identiques. Ces erreurs ne sont ni corrigées à la manière des EL, ni isolées à la manière des RT. Au contraire, elles sont aggravées par les erreurs voisines.

Dans la suite, nous allons voir s'il existe un lien entre les erreurs et les intervalles qui les comportent. Nous observerons tout d'abord les éventuelles tendances à la compression et à l'extension de chacun (intervalle chanté plus petit ou plus grand que l'intervalle visé). Ensuite, nous nous intéresserons à l'influence de l'intervalle visé sur l'imprecision en général, que ce soit par son amplitude (e.g. 2 demi-tons), ou encore son rang au sein de la mélodie (e.g. premier intervalle de la mélodie chantée).

4.3.4 Extension/compression des intervalles

Nous l'avons vu Section 3.5.3, McNab observe une tendance à l'extension des petits intervalles (1 et 2 demi-tons) constituant des séquences monotones (montantes ou descendantes). Par ailleurs, il relève une tendance à la compression des grands intervalles (7 à 9 demi-tons). Nous allons voir si les données de notre corpus confirment ces phénomènes.

Signes d'erreurs et phénomènes associés

Nous l'avons déjà évoqué plus tôt, le phénomène d'extension désigne un intervalle chanté plus grand que l'intervalle visé, et celui de compression, un intervalle chanté plus petit que l'intervalle visé. Conformément à la convention adoptée pour le calcul des erreurs (cf. page 68), le phénomène d'extension entraîne, pour les intervalles positifs (i.e. ascendants), une erreur positive, et celui de compression une erreur négative.

Pour les intervalles négatifs (i.e. descendants), c'est donc le contraire. Les liens entre signe d'erreur et type d'imprécision du chantonnement sont rappelés tableau 4.5.

Type d'intervalle	Signe de l'erreur	Phénomène
Intervalle négatif (descendant)	Positif	Compression
	Négatif	Extension
Intervalle positif (ascendant)	Positif	Extension
	Négatif	Compression

TAB. 4.5 : *Compression/extension des intervalles : correspondance entre signe d'erreur et phénomène.*

Nous allons observer la manière dont les phénomènes d'extension et de compression sont représentés dans notre corpus. La répartition des erreurs nous permettra de définir l'occurrence de ces deux phénomènes, et l'amplitude des deux populations constituées témoignera de leur intensité.

Répartition des erreurs et amplitude des phénomènes associés

Selon les correspondances entre signes d'erreur et phénomènes que nous venons de citer, il apparaît que 57% des erreurs correspondent au phénomène d'extension. Globalement, la tendance (légère) est donc à des intervalles chantés plus grand que les intervalles visés.

Le tableau 4.6 présente la répartition des erreurs par valeur d'intervalle. Il contient le nombre d'occurrence des deux types d'erreur (< 0 et ≥ 0), ainsi que la tendance qui en découle, pour les intervalles visés par nos sujets.

Intervalle	Erreurs négatives	Erreurs positives	Tendance au chantonnement
-10	4	14	Compression (78%)
-7	22	24	Compression (52%)
-5	55	86	Compression (61%)
-4	170	124	Extension (58%)
-3	173	72	Extension (71%)
-2	523	324	Extension (62%)
-1	174	97	Extension (64%)
0	563	820	Int. ascendant (59%)
+1	101	220	Extension (69%)
+2	441	525	Extension (54%)
+3	54	77	Extension (59%)
+4	132	164	Extension (55%)
+5	182	82	Compression (69%)
+7	9	19	Extension (68%)
+8	8	24	Extension (75%)
+9	9	15	Extension (63%)
+11	18	9	Compression (67%)
+12	9	2	Compression (82%)

TAB. 4.6 : *Compression/extension des intervalles : phénomène le plus représenté, en terme de population d'erreur, pour chacun des intervalles disponibles.*

Afin de compléter l'observation par l'intensité -ou ampleur- des phénomènes représentés, la figure 4.11 présente l'amplitude moyenne des erreurs de chaque signe. Ces moyennes sont entourées des *intervalles de confiance à 95%* [6], rebaptisées *zones de confiance* afin d'éliminer toute ambiguïté avec le terme *intervalle*, principalement utilisé dans ce travail pour désigner l'écart existant entre deux hauteurs de notes.

De par la constitution de la figure 4.11 (cf. légende), les zones de confiance symétriques par rapport à 0 correspondent aux mêmes *phénomènes*. A titre d'exemple, ceux-ci sont précisés pour les intervalles -4 et +4, les lettres "C" et "E" désignent respectivement les phénomènes de compression et d'extension. Pour l'intervalle -4, la zone de confiance de *gauche* (erreurs négatives) correspond au phénomène d'extension, celle de *droite* (erreurs positives) correspond au phénomène de compression. Pour l'intervalle +4, la zone de confiance de *gauche* (erreurs négatives) correspond au phénomène de compression, celle de *droite* (erreurs positives) correspond au phénomène d'extension.

[6] L'intervalle de confiance à x% est une plage située de part et d'autre une moyenne telle que l'on est sûr à x% que la moyenne en question s'y trouve. Pour le degré de certitude de 95% utilisé ici, l'intervalle de confiance est le suivant : $m \pm 1.96 * (\frac{\sigma}{\sqrt{n}})$. n, m et σ sont respectivement le nombre de données étudiées, la moyenne des valeurs et l'écart type associé.

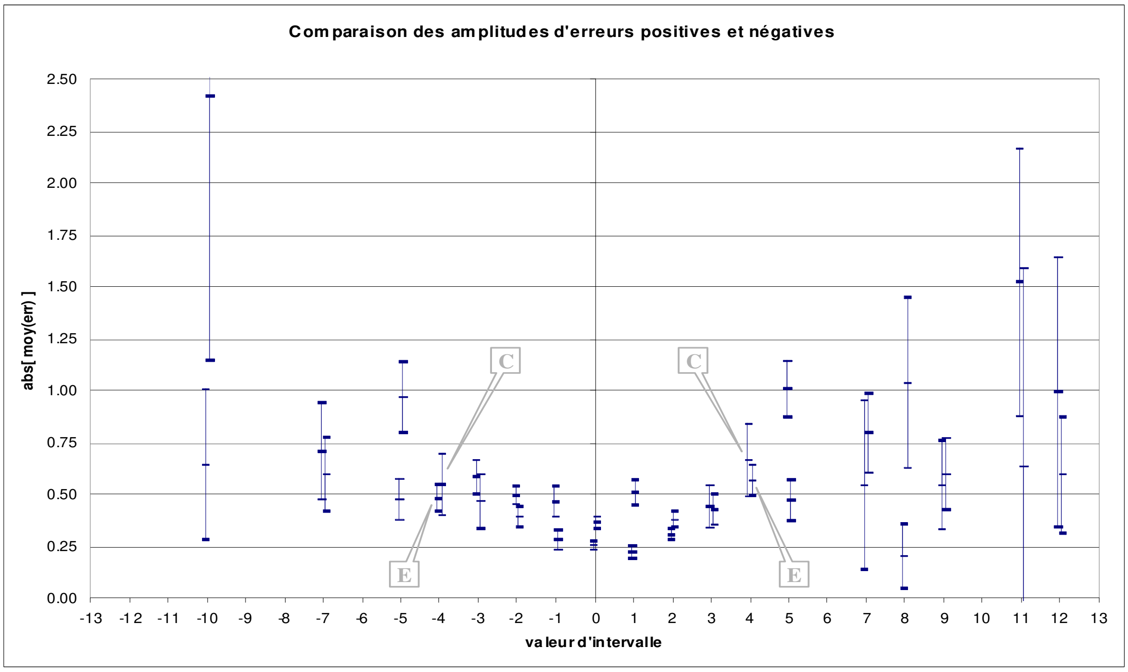


figure 4.11 : *Pour chaque intervalle, de gauche à droite : erreur negative puis erreur positive. Pour permettre la comparaison entre intervalles symétriques, les valeurs représentant les erreurs négatives sont représentées par leur symétrique par rapport à l'axe des abscisses.*

Similitudes sur les intervalles symétriques

On observe, tant au niveau de la répartition des populations d'erreurs (cf. tableau 4.6) qu'au niveau de leurs amplitudes (cf. figure 4.11), que les intervalles symétriques présentent une certaine cohérence.

Concernant la *répartition* des erreurs, les intervalles d'amplitudes identiques présentent des tendances similaires[7].

Cette cohérence se retrouve également dans l'*amplitude* des phénomènes. En effet, 75% des intervalles symétriques présentent une intersection des zones de confiance correspondant aux phénomènes de compression ou d'extension. A titre d'exemple, on peut citer la correspondance entre l'amplitude des erreurs positives de l'intervalle -4 et l'amplitude des erreurs négatives de l'intervalle +4 (cf. figure 4.11). L'ampleur du phénomène de compression est donc comparable pour les intervalles -4 et +4.

Le tableau 4.7 reprend des données illustrées figure 4.11 en présentant, pour chaque paire d'intervalles symétriques, les zones de confiance associées aux deux phénomènes observés. Comme on peut le voir, seuls 3 cas (en gras) sur 12 sont disjoints (ce qui correspond aux 75% de cohérence entre couple d'intervalles symétriques cités à l'instant).

Les phénomènes observés semblent donc dépendre de l'amplitude de l'intervalle visé, indé-

[7]La seule exception est le couple ± 7. Cependant, cette exception est d'autant moins perturbante qu'elle contient le plus faible déséquilibre (52%), et qu'elle correspond au couple d'intervalles le plus faiblement représenté.

Couple	Phénomène	Intervalle	Borne inf	Borne sup
±1	Compression	+1	0.19	0.25
		-1	0.23	0.33
	Extension	+1	0.45	0.57
		-1	0.39	0.54
±2	Compression	+2	**0.28**	**0.34**
		-2	**0.34**	**0.44**
	Extension	+2	**0.35**	**0.42**
		-2	**0.45**	**0.54**
±3	Compression	+3	0.34	0.55
		-3	0.34	0.60
	Extension	+3	**0.35**	**0.50**
		-3	**0.51**	**0.66**
±4	Compression	+4	0.49	0.84
		-4	0.40	0.70
	Extension	+4	0.50	0.64
		-4	0.42	0.55
±5	Compression	+5	0.88	1.15
		-5	0.80	1.14
	Extension	+5	0.38	0.57
		-5	0.38	0.58
±7	Compression	+7	0.14	0.95
		-7	0.42	0.78
	Extension	+7	0.61	0.99
		-7	0.48	0.94

TAB. 4.7 : *Zones de confiances témoignant de l'amplitude des erreurs positives et négatives pour chaque intervalle (elles sont représentées figure 4.11). Les phénomènes correspondants (extension/compression) y sont associés.*

pendamment du signe. Cette cohérence de comportement témoigne d'une similitude des erreurs commises lors du chantonnement d'intervalles symétriques.

Phénomènes prépondérants

A partir des informations de répartition des erreurs, ainsi que l'amplitude des phénomènes associés, nous allons constater l'existence de prépondérances à l'extension ou la compression, pour chacune des amplitude d'intervalle représentée. Nous confronterons nos résultats aux observations de McNab.

Extension des intervalles les plus faibles

Parmi les erreurs d'intervalles faibles, par ailleurs les mieux représentés, la tendance issue de la répartition des erreurs est à l'extension. La prépondérance de ce phénomène est appuyée, pour les intervalles ± 1 et ± 2, par l'amplitude des erreurs (cf. déséquilibre des zones de confiance, figure 4.11). Cela confirme et généralise les observations de McNab qui n'avait relevé cette extension que pour les intervalles constituant des séquences montantes et descendantes.

Concernant l'intervalle nul, le terme d'*extension* n'est pas approprié. La tendance générale dégagée consiste en sa substitution par un intervalle ascendant.

Neutralité des intervalles ± 3 et ± 4

Pour les intervalles d'amplitude 3 et 4, les zones de confiance, témoignant de l'ampleur des phénomènes, sont comparables (cf. figure 4.11). Par ailleurs, les répartitions d'erreurs ne sont pas assez deséquilibrées pour en déduire une tendance nette (cf. tableau 4.6). Les intervalles d'amplitude 3 et 4 ne semblent donc pas présenter de tendance générale à la compression ou l'extension.

Compression des intervalles ± 5

Pour les intervalles d'amplitude 5, la répartition des erreurs ainsi que les amplitudes associées révèlent une nette tendance à la compression.

Ne disposant pas de données concernant les intervalles ± 6, nous passons directement au cas des intervalles de taille supérieure. Nous verrons notamment si la compression des intervalles d'amplitude 7 à 9, observée par McNab est confirmée par notre corpus.

Grands intervalles : tendances variées

A la vue de leurs populations d'erreurs et des amplitudes associées, les grands intervalles représentés dans notre corpus se répartissent dans les trois catégories possibles :
- Neutralité : intervalles ± 7, +9 et +11 ;
- Extension : intervalle +8 ;
- Compression : intervalles -10 et +12.

Ces intervalles étant nettement moins représentés que ceux étudiés jusqu'à présent (cf. figure 4.7), les contextes mélodiques auxquels ils appartiennent sont par conséquent peu variés. Nos observations sont donc spécifiques aux contextes mélodiques recensés, et ceux-ci sont certainement différents de ceux de McNab.

La qualité des sujets influence également les données recueillies, cependant, les ayant sélectionné de manière à couvrir un large domaine de compétence (a l'instar de McNab), nous pensons que l'influence prépondérante provient des contextes mélodiques.

Le phénomène de compression des intervalles d'amplitude 7 à 9 observé par McNab n'est donc pas révélé par notre corpus. Dans les grands intervalles, notre corpus n'est pas assez fourni pour tirer des conclusions générales. Nous pouvons simplement conclure que si, comme McNab l'a observé, certains contextes révèlent la compression de ces intervalles, celle-ci n'est pas systématique.

Faibles populations d'erreurs : illustration de l'influence d'un contexte mélodique donné

Nous allons détailler le cas de l'intervalle +8 qui présente une nette tendance à l'extension, phénomène opposé à celui relevé par McNab. Nous illustrerons ainsi l'influence du contexte sur les résultats obtenus dans le cas de faibles populations d'erreurs.

Les erreurs qui représentent l'intervalle +8 proviennent de deux contextes différents. Le premier, fournissant 75% des erreurs, appartient à "La vie en rose", le second, fournissant les 25% restants, provient de "La Marseillaise". Nous allons observer le comportement des erreurs en fonction du motif visé.

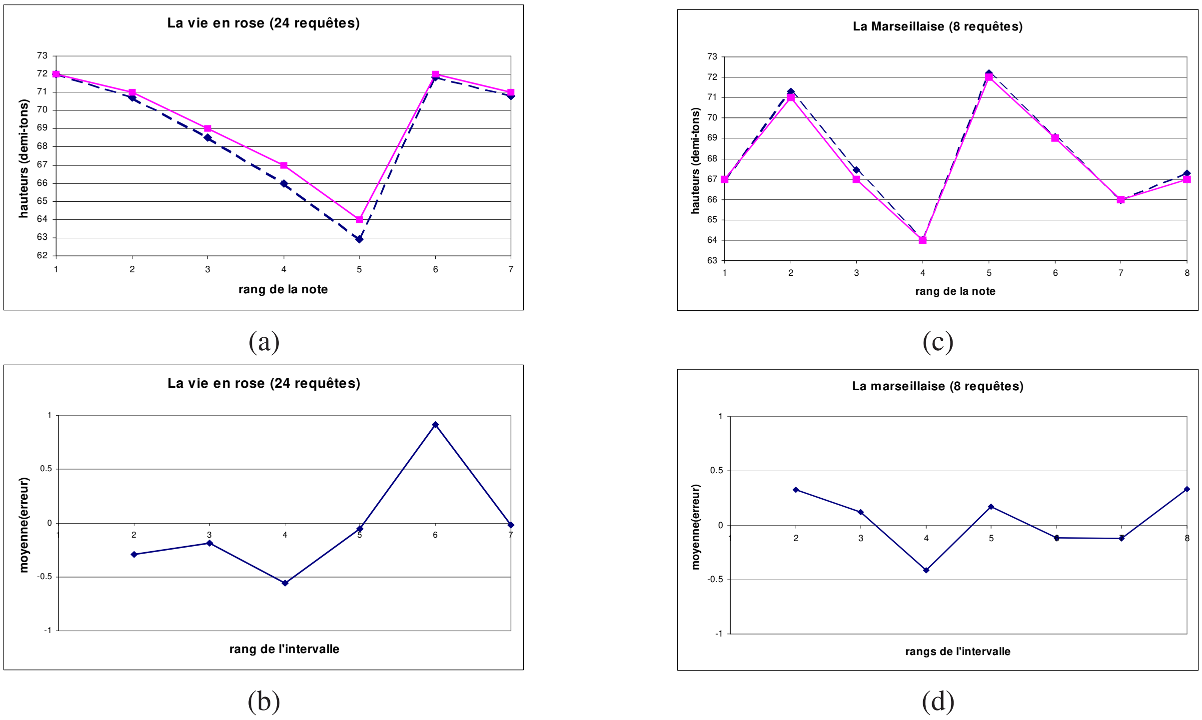


figure 4.12 : *Erreurs moyennes sur deux portions mélodiques contenant un intervalle +8. Figures (a)&(b) : "La vie en rose" ("Quand il me prend dans ses bras") ; Figures (c)&(d) : "La Marseillaise" ("le jour de gloire est a-rri-vé"). - Références : pour les hauteurs la référence est le trait plein (figures (a)&(c)) ; pour les intervalles la référence est 0 (figures (b)&(d)).*

Les figures 4.12 (a) et (b) présentent le début du premier motif considéré : "La vie en rose".

Nous voyons que l'extension de l'intervalle +8 (abscisse 6) vient corriger un GT descendant. Nous pensons que cette correction est favorisée par le fait que la hauteur visée par l'intervalle +8 est la même que la première hauteur chantée. Deux éléments nous y incitent : d'une part, la hauteur ayant été récemment chantée dans la mélodie, on peut penser qu'une nouvelle réalisation s'avère facilitée. D'autre part, le sujet devrait aisément pouvoir chanter cette hauteur puisqu'il l'a spontanément choisie pour commencer sa requête (cela ne serait pas forcément le cas des hauteurs suivantes, guidées par la mélodie visée).

Ayant pris connaissance du contexte majoritaire, intéressons-nous maintenant à la deuxième source d'intervalles +8 : "La Marseillaise".

Comme on peut le voir sur les figures 4.12 (c) et (d), leurs erreurs ne présentent pas les mêmes caractéristiques (cf. abscisse 5). En effet, la moyenne observée est nettement inférieure (0.17 contre 0.91 pour "La vie en rose") et le contexte mélodique est également différent.

Même si, dans les deux contextes, l'intervalle +8 est entouré d'intervalles descendants, leur nombre et leur valeur sont différents. Dans "La Marseillaise", la séquence descendante qui précède contient deux intervalles, de taille plutôt moyenne (-4,-3), alors que dans "La vie en rose", elle contient 4 intervalles de taille plutôt petite (-1, -2, -2, -3).

Nous voyons donc qu'un même intervalle peut comporter des erreurs très différentes lorsqu'il est plongé dans des contextes différents. Le manque de variété de ceux-ci au sein d'un corpus peut donc orienter sensiblement les résultats obtenus.

Cette illustration nous permet également de voir que les types d'erreur recensés en 4.3.3 font partie des plus simples. En effet, dans la cas de "La vie en rose", le GT des premières notes est totalement rectifié par l'extension de l'intervalle +8. Donc, contrairement à l'EL où l'imprécision ne dure que le temps d'une note, la tonalité initialement adoptée est délaissée sur *plusieurs* notes avant d'être rejointe. Cette capacité de rectification à "long-terme" illustre la complexité des erreurs qu'une requête chantonnée peut contenir.

Conclusion

L'observation des phénomènes de compression/extension des intervalles chantés nous a permis de généraliser certaines des tendances révélées par McNab. Grâce à l'important volume de notre corpus, nous avons pu nous intéresser à l'ensemble des intervalles recensés. Nous avons ainsi dégagé des tendances ignorées jusqu'alors.

Nos résultats pourront être mis à profit dans les moteurs de comparaison fondés la quantification des intervalles. Nous avons vu, par exemple, que l'intervalle nul était plutôt chanté comme un intervalle ascendant, et que l'intervalle +1 subissait généralement une extension. Ajuster la valeur du seuil de quantification entre ces deux intervalles (classiquement placé à 0.5 demi-ton) améliorerait la transcription/quantification d'une mélodie chantonnée, en évitant de prendre des intervalles nuls (chantés positifs) pour des intervalles +1 (compressés).

Pour les intervalles les moins représentées, les tendances que nous avons observées ne sont pas conformes à celles de McNab. Nous pensons que ces tendances (tant les nôtres que celles de McNab, dont le corpus est plus modeste) sont représentatives de contextes mélodiques particuliers.

Ce point a permis de souligner l'importance du contexte mélodique d'un intervalle sur sa compression/extension. La prise en compte d'informations complémentaires permettrait sans doute de diviser le cas général que nous avons traité ici en familles de contextes aux tendances spécifiques. Par exemple, la neutralité des intervalles ± 3 et ± 4 pourrait provenir d'un équilibre de deux tendances opposées. Celles-ci pourraient être révélées par la connaissance des intervalles voisins de celui étudié, mais aussi du placement des hauteurs par rapport à la tessiture des sujets.

Par ailleurs, nos résultats témoignent de la forte similitude des erreurs commises lors du chantonnement d'intervalles symétriques. Cette propriété va nous permettre, dans la section suivante, d'effectuer un regroupement de données. Cela avait déjà été effectué dans les travaux de Lindsay, cependant, cette action trouve ici sa justification dans le contexte expérimental qui nous intéresse (les sujets connaissent les mélodies chantées).

4.3.5 Dépendance de l'imprécision à l'amplitude de l'intervalle visé

Dans ce qui suit, nous regroupons les erreurs issues des intervalles de même amplitude. Le tableau 4.8 présente les populations d'erreurs obtenues.

Amplitude	0	1	2	3	4	5	6	7	8	9	10	11	12
Population	1383	592	1813	376	590	405	0	74	32	24	18	27	11

TAB. 4.8 : *Populations représentant les différentes amplitudes d'intervalle, après regroupement des erreurs issues des intervalles symétriques.*

Comparaison avec les travaux de Lindsay

Comme nous l'avons vu au chapitre 3 (p. 52), Lindsay a réalisé une expérience dans laquelle il présente des mélodies que ses sujets doivent immédiatement chanter. Cependant, les mélodies n'étant pas connues des sujets, Lindsay sort d'un cadre réaliste d'utilisation des systèmes qui nous intéressent. En plus du fait qu'il sollicite la mémoire court terme de ses sujets, les mélodies présentées ont un tempo rapide (4 notes par seconde). Les mélodies chantées présentent donc des notes très courtes entraînant généralement une instabilité des fréquences chantées, qui entraîne à son tour une confiance relative dans les hauteurs de notes issues de la transcription.

Il nous paraît intéressant de confronter notre corpus à ses résultats. Nous allons donc observer nos données à la manière dont Lindsay présente les siennes.

Ce dernier a limité ses investigations aux intervalles compris entre -7 et +7. Une sélection de nos données correspondantes nous évitera d'avoir à prendre en compte des intervalles faiblement représentés. De plus, le regroupement des erreurs issues des intervalles symétriques, justifié par la forte similitude de leurs erreurs, augmente les populations considérées. Cela dit, comme le montre le tableau 4.8, les intervalles d'amplitude 7 sont nettement désavantagés par rapport aux précédents (exceptés ceux d'amplitude 6, absents de notre corpus).

La figure 4.13 présente nos résultats et ceux de Lindsay [Lin96]. Pour chaque amplitude d'intervalle visée, la moyenne des erreurs est accompagnée de l'écart-type.

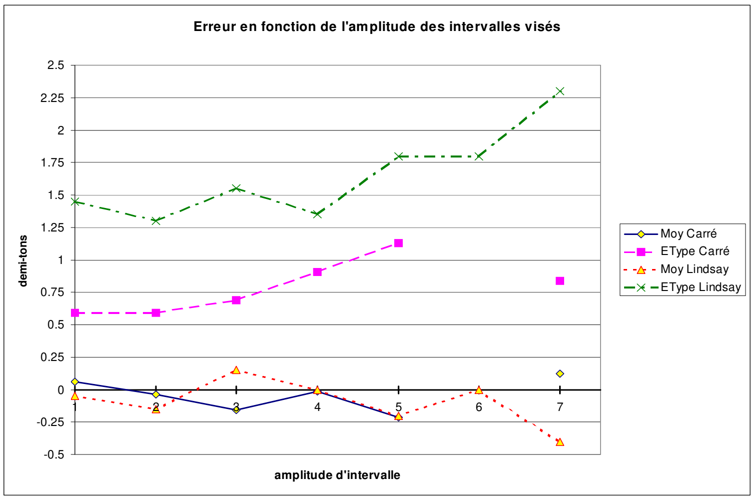


figure 4.13 : *Comparaison des moyennes et écarts-types des erreurs de nos données et de celles de Lindsay.*

Puisqu'il nous manque la valeur de l'imprécision sur l'amplitude 6, il nous est impossible de comparer complètement nos résultats avec ceux de Lindsay. On peut cependant remarquer, sur les données communes, que les moyennes obtenues sont comparables (cf. deux courbes les plus basses).

Concernant les écarts-types (cf. deux courbes les plus hautes), nous restreignons notre observation aux cinq premières valeurs. Ainsi, nous nous fondons sur une courbe continue, sans présumer du comportement de nos données manquantes (amplitude 6), ou faiblement représentées (amplitude 7). Le fait que nos données mènent à des tronçons plus lisses s'explique par la plus grande quantité de données dont nous disposons.

Les tronçons considérés (amplitudes 1 à 5) empruntent des domaines de variation différents. D'étendues comparables (environ 0.5 demi-ton), ils se situent à deux niveaux différents : les données de notre corpus mènent à des écarts-types nettement inférieurs à ceux de Lindsay. Le tableau 4.9 reprend les valeurs d'écart-type de la figure 4.13 ainsi que les tailles de population dont ils sont issus.

Amplitude	Ecart-type		Population	
	Corpus Lindsay	Corpus Carré	Corpus Lindsay	Corpus Carré
1	1.45	0.59	18	592
2	1.3	0.59	18	1813
3	1.55	0.69	18	376
4	1.35	0.91	18	590
5	1.8	1.13	18	405
6	1.8	-	18	0
7	2.3	0.84	18	74

TAB. 4.9 : *Ecarts-types et tailles des populations représentant les erreurs associées aux différentes amplitudes d'intervalle.*

La différence des profils d'écarts-types pourrait être due aux compétences des sujets sélectionnés. Cependant, la répartition des profils rassemblés (d'expert à néophyte) semblant comparable, nous ne privilégierons pas cette interprétation. Les caractéristiques intrinsèques des mélodies visées, ou encore le choix de Lindsay d'utiliser des mélodies rapides pourraient également causer une telle différence de précision. Cependant, nous pensons que ce phénomène pourrait être causé par la différence existant entre mémoire à long-terme et à mémoire à court-terme : il est plus facile de reproduire avec exactitude une mélodie connue qu'une mélodie fraîchement mémorisée [UZ98].

Concernant l'influence de l'amplitude de l'intervalle visé sur la précision, nous avons vu que la *variation* que présente notre corpus est comparable avec celle relevée par Lindsay. Cette influence se retrouve donc également dans le chantonnement de mélodies *connues*.

L'augmentation de l'imprécision avec l'amplitude de l'intervalle visé est-elle négligeable ? Dans ce qui suit, nous allons tenter de répondre à cette question.

Faut-il négliger la dépendance de l'imprécision à l'amplitude de l'intervalle visé ?

Pour disposer d'éléments de réponse, nous allons quitter le mode de visualisation adopté par Lindsay. Afin observer l'influence de l'amplitude de l'intervalle visé sur la précision, nous allons considérer les erreurs, en valeur absolue cette fois-ci. Les phénomènes de compression et d'extension ne seront donc plus distingués. La moyenne des erreurs désignera l'*imprécision* associée à chaque amplitude d'intervalle.

La figure 4.14 représente les moyennes obtenues, ainsi que leurs zones de confiance, en fonction de l'amplitude de l'intervalle visé.

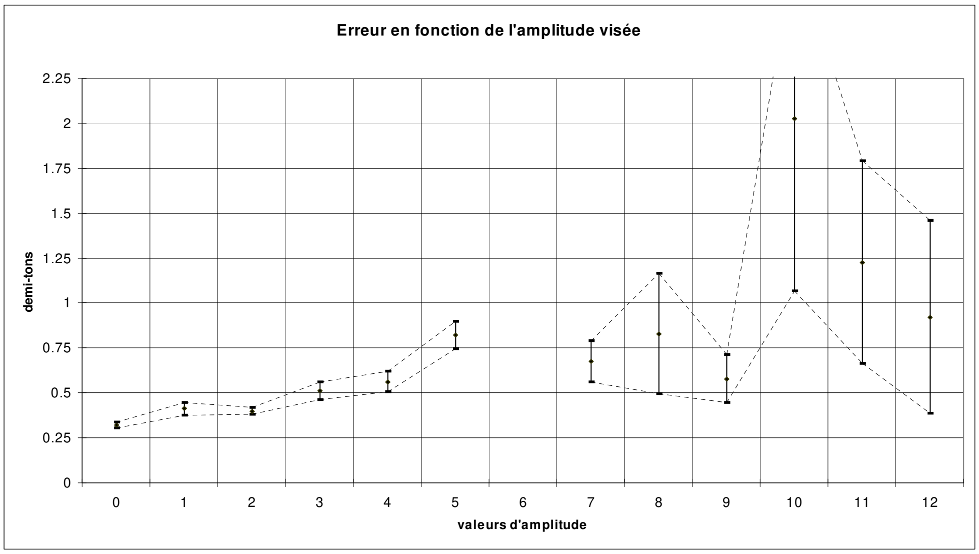


figure 4.14 : *Amplitude moyenne des erreurs en fonction de l'amplitude de l'intervalle visé.*

Sur le premier tronçon de la figure, nous pouvons voir que l'imprécision passe d'une amplitude de 0.3 à plus de 0.75 demi-ton. L'intervalle de base de la musique occidentale étant le demi-ton, cette évolution apparaît comme *non négligeable*. Après une absence de donnée à l'amplitude 6, on observe un deuxième tronçon (amplitudes de 7 à 12) aux amplitudes globalement élevées.

Malheureusement, l'étendue des zones de confiance le constituant ne nous permet pas de définir, de manière précise, la tendance comportementale de l'imprécision.

Par exemple, il est possible que l'intervalle d'octave (12 demi-tons) soit plus précis que d'autres intervalles d'amplitude inférieure. Un tel phénomène pourrait s'expliquer par un lien entre perception et production. En effet, malgré son amplitude importante, *cet intervalle est perçu par l'auditeur non exercé comme une totale similitude entre les deux notes* [Gui96]. Il n'est pas exclu que cette propriété de la perception humaine s'accompagne d'une précision dans la réalisation (chantée). Celle-ci serait meilleure pour cet intervalle que pour d'autres d'amplitude inférieure...

Quoiqu'il en soit, il semble que la dépendance observée ne soit pas négligeable. Cependant, les données dont nous disposons ne nous permettent pas de modéliser précisément l'évolution de l'imprécision en fonction des amplitudes supérieures ou égales à 6. Il serait intéressant, à l'avenir, de constituer un corpus plus fourni dans les fortes amplitudes afin de préciser les observations.

Conséquences de l'imprécision observée

Nous avons vu que les imprécisions observées étaient relativement élevées par rapport à l'intervalle de base de la musique occidentale (le demi-ton). Séparant deux hauteurs tempérées contiguës, ce dernier implique qu'une erreur supérieure à un quart de ton (soit 0.5 demi-ton) entraîne une méprise sur l'intervalle visé par l'utilisateur.

Les systèmes de recherche par chantonnement actuels appliquent une quantification des intervalles de la requête. Pour ceux comportant une erreur inférieure à 0.5, l'intervalle visé sera bien estimé, pour les autres cela ne sera pas le cas.

Prenons le cas de l'intervalle +3 dont l'erreur moyenne se situe aux alentours de 0.5 (cf. figure 4.14). Si l'intervalle chanté vaut 3+0.49 = 3.49, la quantification le ramènera sur la valeur la plus proche, soit +3. L'imprécision de 0.49 sera annulée.

Par contre, si l'imprécision est supérieure de 0.02, soit 0.51, l'intervalle chanté vaudra 3+0.51 = 3.51. La quantification le ramènera vers la valeur la plus proche, soit +4. L'imprécision sera aggravée.

La figure 4.14 montre que si certains intervalles (en particulier l'intervalle nul) ont de bonnes chances de voir leurs erreurs annulées, nombre d'autres subiront le sort contraire. Comme nous le verrons au chapitre 5, cette observation nous amènera à nous démarquer des systèmes appliquant une quantification des intervalles.

Conclusion

Concernant la dépendance de l'imprécision à l'amplitude de l'intervalle visé, nous avons montré que les erreurs constituant notre corpus présentaient une caractéristique commune avec celles étudiées par Lindsay. En effet, l'augmentation de la taille de l'intervalle visé provoque des augmentations de l'imprécision comparables. Par contre, les valeurs représentant cette dernière sont meilleures (i.e. moins élevées) que celles de Lindsay. Parmi les causes possibles de cette meilleure précision, la supériorité de la mémoire à long-terme sur la mémoire à court-terme nous paraît importante (plus grande précision des mélodies chantées de mémoire, par rapport à celles, inconnues, répétées après diffusion).

L'imprécision est donc bien dépendante de l'amplitude de l'intervalle visé. Cependant, contrairement à Lindsay, cette dépendance ne nous paraît pas vraiment négligeable. Malheureusement, les faiblesses de notre corpus d'erreurs ne nous permettent pas d'établir un lien clair entre amplitude de l'imprécision et taille de l'intervalle visé, sur l'*ensemble* du domaine d'intervalle étudié.

Par ailleurs, les imprécisions observées nous montrent que nombre d'entre elles seraient aggravées par l'application d'une quantification, processus unanimement adopté par les systèmes actuels. Ce constat nous poussera, au chapitre 5, à proposer une solution alternative.

4.3.6 Dépendance de l'imprécision au rang de l'intervalle dans la mélodie

Nous avions observé, lors de la constitution du corpus de requêtes chantonnées, que le premier intervalle pouvait être particulièrement imprécis. Le rang des intervalles au sein de la mélodie pourrait donc avoir une influence sur la précision des requêtes.

La figure 4.15 présente l'imprécision des intervalles en fonction de leur rang dans la mélodie. Nous avons limité notre observation aux rangs représentés par un minimum de 100 valeurs, aboutissant aux 14 premiers rangs.

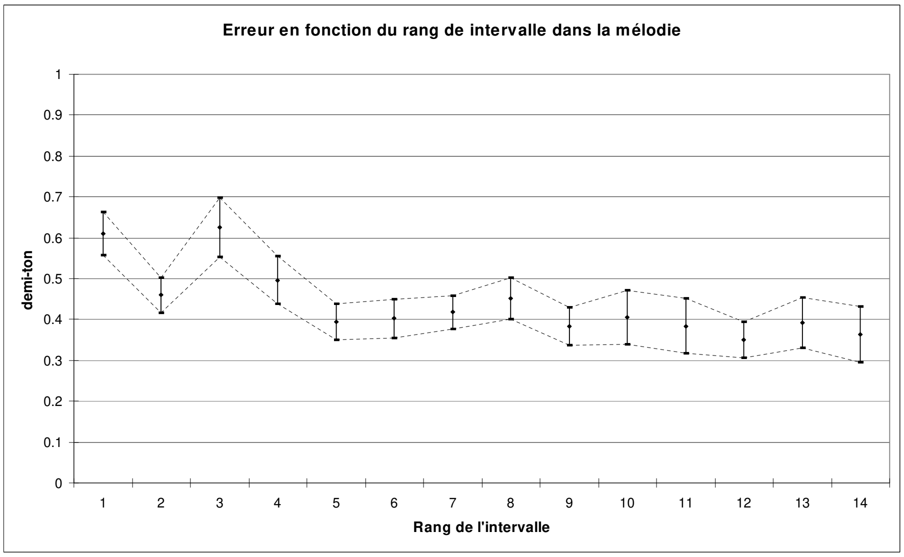


figure 4.15 : *Erreur en fonction du rang de l'intervalle au sein de la requête.*

Il apparaît effectivement que le début de la requête soit le lieu d'imprécisions plus élevées que dans la suite. En effet, la grande majorité des imprécisions se situe aux alentours de 0.4 demi-ton, ce qui n'est le cas ni du rang 1, ni du rang 3, plutôt situés au alentours de 0.6 demi-ton.

Si nous nous attendions à une plus forte imprécision du rang 1, le cas du rang 3 est plus surprenant. Cela dit, comme nous venons de le voir (cf. 4.3.5), tous les intervalles ne sont pas réalisés avec la même précision. Par conséquent, à rang égal, une population constituée d'intervalles d'amplitude 5 devrait présenter une plus grande imprécision qu'une population constituée d'intervalles nuls. Afin de voir si les imprécisions observées sont bien associées aux rangs eux-mêmes, nous allons tenter de nous affranchir de l'influence des amplitudes d'intervalle visées.

Dissocier l'influence du rang de celle des amplitudes d'intervalle visées

Dans la section 4.3.5, nous avons vu que nous ne pouvions pas modéliser avec précision l'influence de *toutes* les amplitudes d'intervalle sur les erreurs commises. La dépendance entre amplitude visée et imprécision est cependant bien réelle. Les différences de précision observées sur les rangs 1 et 3 peuvent être dues à des répartitions particulières des intervalles visés (prépondérance d'intervalles à forte imprécision), et non au rang lui-même.

Amplitudes d'intervalle visées

Afin d'éclaircir cette question, nous commencerons par observer, pour chaque rang, la répartition des intervalles visés. La figure 4.16 présente les 8 premiers. Les 6 autres sont disponibles en annexe A.

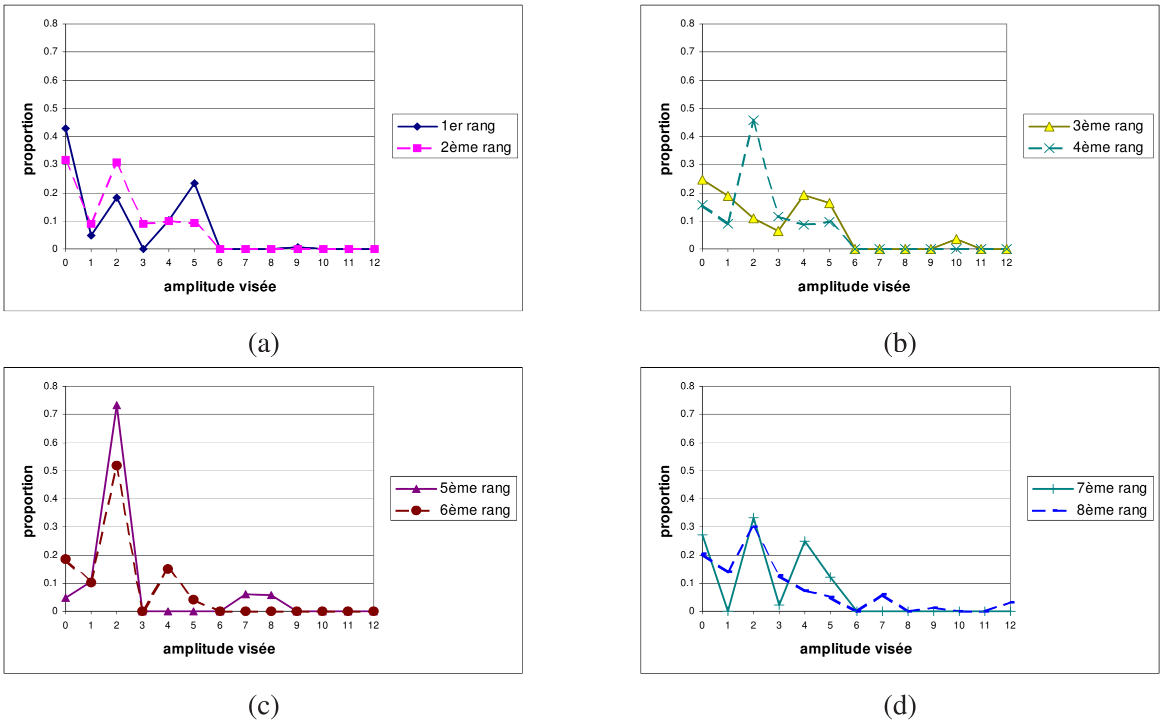


figure 4.16 : *Répartition des intervalles présents pour chacun des 8 premiers rangs. (a) rangs 1 et 2 ; (b) rangs 3 et 4 ; (c) rangs 5 et 6 ; (d) rangs 7 et 8.*

Il apparaît clairement que tous les rangs ne présentent pas des répartitions identiques. Celles-ci pourraient donc être à l'origine des différences d'imprécisions observées figure 4.15.

Imprécisions associées aux amplitudes d'intervalle visées

Si l'imprécision associée à *chaque* amplitude d'intervalle était clairement définie, nous pourrions estimer l'imprécision attendue pour chacun des rangs considérés à partir de leurs répartitions respectives. Puisque cette information n'est connue que pour certaines amplitudes d'intervalle, nous nous préparons à n'en effectuer qu'une estimation *partielle*. Nous verrons par la suite si la différence entre les imprécisions mesurées et estimées traduit une influence du rang.

Le tableau 4.10 présente les 6 imprécisions les plus clairement définies. Comme nous l'avions vu figure 4.14, elles correspondent aux zones de confiance les plus restreintes.

Amplitude	0	1	2	3	4	5
Nom	$impr_0$	$impr_1$	$impr_2$	$impr_3$	$impr_4$	$impr_5$
Valeur	0.32	0.41	0.40	0.51	0.56	0.82

TAB. 4.10 : *Imprécision moyenne des amplitudes d'intervalle les plus représentées.*

Ces 6 moyennes vont nous servir à estimer les contributions à l'imprécision des amplitudes d'intervalle correspondantes, pour chacun des rangs considérés.

A titre d'exemple, considérons le premier rang. La répartition des amplitudes d'intervalles le constituant (illustrée figure 4.16(a)) est présentée tableau 4.11.

p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}
0.43	0.05	0.18	-	0.10	0.23	-	-	-	0.01	-	-	-

TAB. 4.11 : *Proportions définissant la répartition des amplitudes d'intervalles visées, pour le premier rang.*

L'estimation partielle de l'imprécision pour le rang 1 est donnée par la somme des 6 premières proportions, pondérées par les imprécisions moyennes du tableau 4.11, soit

$$\sum_{k=0}^{5} p_k * impr_k \; = \; 0.48$$

Cette valeur, ainsi que les imprécisions partiellement estimées pour les autres rangs concernés, sont présentées dans la quatrième colonne du tableau 4.12. Afin de comparer ces estimations avec imprécision mesurées, nous y avons ajouté, colonnes 2 et 3, les bornes délimitant les zones de confiance précédemment illustrées figure 4.15. La superposition des imprécisions mesurées et (partiellement) estimées est présentée figure 4.17.

Rang	Borne sup.	Borne inf.	Prévision (partielle)	Eléments non pris en compte *[amplitude ; proportion]*
1	0.66	0.56	0.48	[9 ;0.01]
2	0.50	0.42	0.44	
3	0.70	0.55	0.47	[10 ; 0.04]
4	0.55	0.44	0.46	
5	0.44	0.35	0.35	[7 ; 0.06] [8 ; 0.06]
6	0.45	0.36	0.43	
7	0.46	0.38	0.47	
8	0.50	0.40	0.39	[7 ; 0.06] [9 ; 0.01] [12 ; 0.03]
9	0.43	0.34	0.43	
10	0.47	0.34	0.35	[7 ; 0.08]
11	0.45	0.32	0.35	[9 ; 0.05] [11 ; 0.07]
12	0.39	0.31	0.41	
13	0.45	0.33	0.47	
14	0.43	0.30	0.37	[8 ; 0.07]

TAB. 4.12 : *Bornes des zones de confiance issues de la mesure de l'imprécision pour chacun des 14 premiers rangs (illustrées figure 4.15 et 4.17) ; Estimation de la contribution à l'imprécision des 6 premières amplitudes d'intervalle, et caractéristiques des éléments non pris en compte dans l'estimation.*

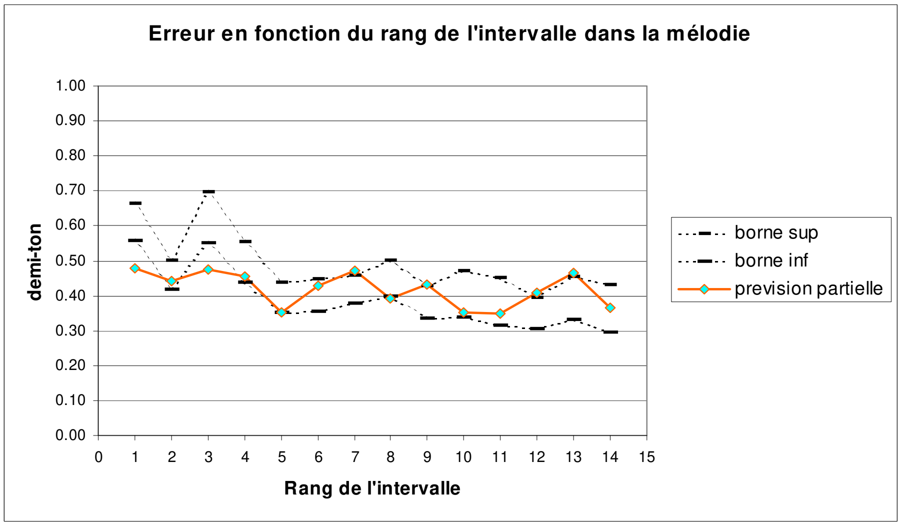


figure 4.17 : *Erreur en fonction du rang de l'intervalle au sein de la requête, et prévisions (partielles).*

Comme nous pouvons le voir dans la dernière colonne du tableau 4.12, la prévision annoncée comme partielle s'avère complète pour la moitié des rangs considérés (2, 4, 6, 7, 9, 12, 13). Or, pour ces derniers, l'imprécision estimée correspond à l'imprécision mesurée (elle peut y être légèrement supérieure, cf. rangs 7, 12 et 13). Nous considèrerons donc que notre estimation est fiable.

Influence du rang dans l'imprécision mesurée

Les rangs 1 et 3 appartiennent aux rangs dont toutes les amplitudes recensées n'ont pas été prises en compte dans l'estimation de l'imprécision. Malgré la faible proportion des amplitudes manquantes (0.01 et 0.04, cf. tableau 4.12), les imprécisions estimées sont nettement inférieures aux imprécisions mesurées.

Concernant le rang 3, la distance séparant l'estimation de l'imprécision mesurée (i.e. borne inférieure de la zone de confiance) est égale à $0.55 - 0.47 = 0.08$ demi-ton. Ainsi, pour que l'estimée complète de l'imprécision corresponde à l'imprécision mesurée, il faudrait que l'imprécision associée à l'amplitude 10 soit égale à $\frac{0.55-0.47}{0.04} = 2$ demi-tons. Aux vues de la figure 4.14, cette valeur n'est pas aberrante. L'imprécision observée au rang 3 semble donc être due au fait que les intervalles qui le peuplent soit plus difficile à chanter.

Par contre, lorsqu'un raisonnement identique est mené pour le rang 1, on en déduit une imprécision associée à l'amplitude 9 égale à $\frac{0.56-0.48}{0.01} = 8$ demi-tons ! Cette valeur étant démesurée, nous en concluons que le premier intervalle est effectivement chanté de manière moins précise que les suivants.

Conclusion

Nous venons de voir que le premier intervalle chanté était globalement moins précis que les suivants. Cependant, cette augmentation de l'imprécision est modeste. Dans le cas général, il paraît donc superflu de diminuer l'influence du premier intervalle dans la comparaison mélodique. Lors de la construction de notre moteur de comparaison (cf. chapitre 5), l'ensemble des intervalles de la requête se verront donc accorder une confiance identique.

Cependant, il faut noter que les grandeurs observées représentent la performance *moyenne* des sujets. Il est donc possible qu'une partie d'entre eux présentent ce phénomène d'une manière plus marquée. Si cette hypothèse était confirmée, deux types de profils d'utilisateurs pourraient être définis (e.g. *précis* et *imprécis*) entraînant une adaptation du moteur de comparaison utilisé dans la recherche. Pour les requêtes produites par des utilisateurs *imprécis*, l'influence du premier intervalle serait diminuée.

4.4 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux requêtes chantonnées, et plus particulièrement à leurs caractéristiques fréquentielles. Nous avons constitué un corpus de 500 mélodies qui nous a permis, dans un premier temps, de confirmer l'existence de types d'erreurs[8] observés dans les deux seules contributions préalables dont nous avions connaissance ([MSWH00, Lin96]).

Par ailleurs, nous nous sommes intéressés aux phénomènes d'extension/compression des intervalles. Nous avons généralisé certains résultats des études précédentes, comme l'extension des petits intervalles. Nous en avons relativisé d'autres, comme la compression des intervalles de grande amplitude, qui semble principalement liée à certains contextes mélodiques. Nous avons complété ces observations par la révélation de tendances concernant les intervalles de taille moyenne et

[8] Erreur Locale, Rupture de Ton et Glissement de Ton (EL, RT et GT).

l'intervalle nul. Les tendances générales révélées pourront être mises à profit dans les moteurs de comparaison fondés sur la quantification des intervalles (e.g. comme guide pour le choix des frontières d'états).

La cohérence des résultats observés sur les intervalles symétriques nous a permis de rassembler les données correspondantes, et ainsi, de nous fonder, pour les conclusions suivantes, uniquement sur l'*amplitude* des intervalles visés par les sujets.

Ainsi, nous avons confirmé la dépendance de l'imprécision à la taille de l'intervalle visé. Au regard des amplitudes d'erreurs obtenues, cette dépendance nous a paru non négligeable (à la différence de Lindsay). Malheureusement, l'hétérogénéité de notre corpus nous a empêché de définir clairement la dépendance de l'imprécision aux intervalles visés, au delà de l'amplitude 5.

Cependant, notre modélisation (partielle) de l'imprécision en fonction de l'amplitude visée nous a tout de même permis de d'estimer l'influence du rang de l'intervalle sur sa précision. Ainsi, nous avons pu montrer qu'il existait une augmentation de l'imprécision sur le premier intervalle chanté. D'une manière générale, ce phénomène est léger ; par conséquent, nous n'en tiendrons pas compte lors de la construction de notre moteur de comparaison. L'influence du premier intervalle ne sera pas donc diminuée par rapport aux suivants. Cependant, nous pensons que ce phénomène pourrait se révéler plus nettement pour une certaine catégorie de sujets. Un système de recherche par chantonnement différenciant les niveaux de précision des utilisateurs pourrait en tirer avantage.

Les imprécisions observées nous ont montré que nombre d'entre elles seraient aggravées par l'application d'une quantification au demi-ton, processus largement utilisé dans les systèmes actuels. Cet inconvénient vient s'ajouter au problème de tolérance non uniforme de la description obtenue (problème souligné en 3.5.2). La quantification des hauteurs de la requête possède donc des conséquences néfastes.

La nature des requêtes chantonnées étant un peu mieux connue, nous allons passer à la conception d'un moteur de comparaison.

Chapitre 5

Conception d'un Moteur de Comparaison

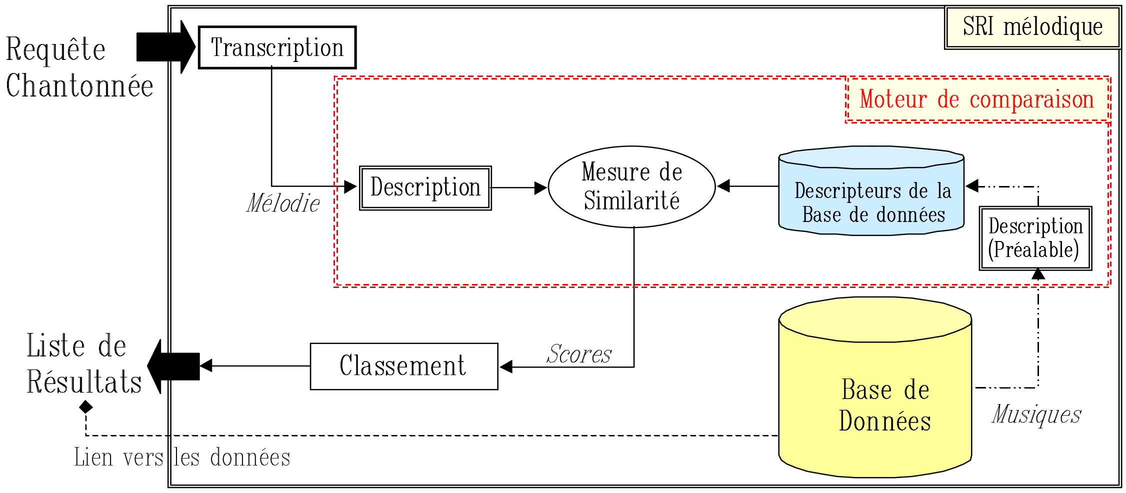


figure 5.1 : *Système de recherche de documents musicaux par chantonnement, avec moteur de comparaison détaillé.*

5.1 Introduction

A l'instar des systèmes rencontrés jusqu'ici, notre système de recherche s'adresse à une base de données constituée de fichiers MIDI. Comme nous l'avons déjà vu en 3.4.2, ceux-ci présentent l'avantage de contenir les principales informations attendues d'une transcription (malheureusement non réalisable automatiquement). Il s'agit des notes jouées par les différents instruments qui interviennent dans une musique.

Concernant la requête, le module de transcription automatique fournit une mélodie, constituée d'une séquence de trios *hauteur - instant de début de note - instant de fin de note* (cf. 4.2).

La tâche du moteur de comparaison est de déterminer la similitude des données de la base vis-à-vis de la mélodie tirée de la requête. Nous l'avons vu en 3.3.1, ce jugement doit être indépendant des notions de tonalité et de tempo. Par contre, il doit tenir compte d'autres types de différences, telles que l'imprécision des hauteurs chantonnées. Tolérance et précision doivent être gérées de manière à sanctionner avec discernement les différences entre requête et portions de la base de données. Dans le cas contraire, on risque d'aboutir à deux cas extrêmes de réponse du système, à savoir, d'une part, une liste de résultats jugés tous similaires à la requête (précision insuffisante), et d'autre part, une liste de résultats jugés tous dissimilaires à la requête (tolérance insuffisante).

Ce compromis *tolérance/précision* est assuré par le moteur de comparaison qui comporte, comme on peut le voir sur la figure 5.1, la *description* des données musicales à comparer, et une *mesure de similarité*.

Dans ce chapitre, nous allons suivre pas à pas le cheminement permettant de construire un moteur de comparaison. Pour chaque problème soulevé, nous tenterons d'envisager toutes les solutions possibles avant de nous prononcer.

Ce chapitre est organisé de la manière suivante. Dans la section 5.2, nous définirons la description de la mélodie issue de la requête. La section 5.3 présentera la description adoptée pour les données de la base. Enfin, la section 5.4 viendra compléter le moteur de comparaison en introduisant la mesure de similarité témoignant des différences significatives entre descripteurs.

5.2 Description de la requête

Au Chapitre 4, nous avons vu que l'information issue de la requête était d'une nature différente de celle des données de la base (notamment des hauteurs non tempérées, souvent négligées par les systèmes actuels). Dans cette section, nous allons voir l'influence de cette nature spécifique sur la description de la requête. Nous commencerons par considérer la manière de représenter l'information temporelle (instants d'apparition et durée des notes), puis nous complèterons la description par la prise en compte de l'information fréquentielle (hauteurs).

5.2.1 Information temporelle

A partir des séquences des *instants de début et de fin de note*, nous allons voir comment représenter l'information temporelle de la requête de manière à ce que la similarité mesurée par notre moteur soit indépendante du tempo, tolère les imperfections rythmiques, tout en gardant la précision nécessaire à l'identification.

S'affranchir du tempo

Si l'on peut s'attendre à ce qu'un utilisateur connaisse approximativement le rythme de la mélodie qu'il recherche, il est difficile de lui demander de chanter sa requête au tempo original. Comme nous l'avons déjà vu au Chapitre 3, la durée des notes dépend du tempo emprunté par l'utilisateur. Désirant s'affranchir de ce dernier, nous aimerions disposer d'une base temporelle à laquelle nous pourrions nous référer pour exprimer la longueur des notes. Si l'on désire être conforme à la notation musicale classique occidentale, cette base temporelle devrait être l'inverse du tempo[1], que nous appèlerons T (exprimée en secondes).

[1] Le tempo est une fréquence. Il s'exprime en battements par minute, ou *bpm*.

Dans un premier temps, nous allons voir différentes manières de fournir au système ce type de référence, puis nous verrons différentes manières de s'en passer. Enfin, nous verrons la solution retenue pour notre moteur de comparaison.

Référence temporelle fournie par l'utilisateur

Nous avons vu qu'il existait des systèmes qui attendent de l'utilisateur qu'il fournisse lui-même le tempo de sa requête. Certains lui font directement spécifier sa valeur, d'autres lui font définir la fréquence d'une pulsation, sur laquelle il s'appuie pour chanter sa requête. Puisque le tempo des mélodies de la base est généralement spécifié dans les fichiers MIDI, la comparaison peut s'effectuer indépendamment des tempos empruntés.

Cependant, ces solutions requièrent une certaine compétence ce qui réduit sensiblement le champs des utilisateurs. Le système développé dans le cadre de cette thèse devant s'adresser au plus grand nombre, ce type de solutions n'a pas été retenu.

Extraction automatique d'une référence temporelle

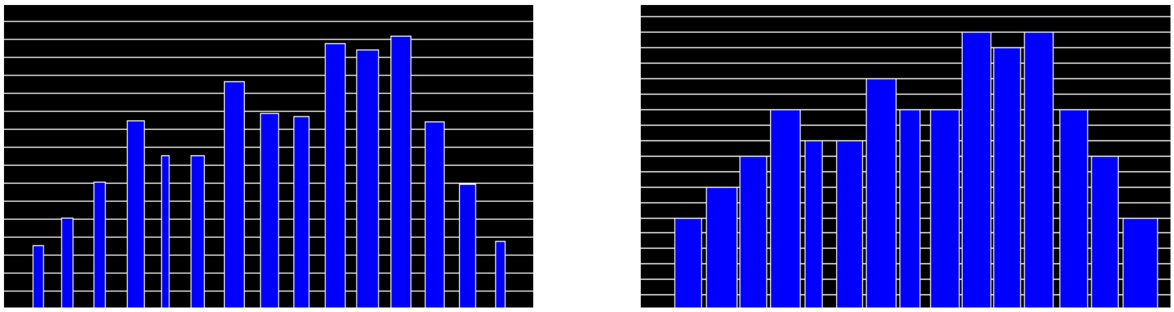
L'extraction automatique d'une base temporelle est envisageable. En effet, l'étude de l'information temporelle des notes chantées permet d'extraire une référence, qui peut servir de base à l'analyse rythmique de la mélodie fournie par l'utilisateur.

Notons qu'il n'est pas forcément nécessaire de disposer de la période T, elle même. La base temporelle extraite peut très bien être un (sous-)multiple de T (typiquement $T/2$, $2T$, $T/3$...). En effet, le but étant de comparer la requête à une portion de la base de données, le même traitement est appliqué aux deux motifs mélodiques. Si les motifs rythmiques comparés sont proches, les bases temporelles extraites témoignent de la différence des tempos employés (rapport des bases temporelles = rapport des tempos). Ainsi, les deux motifs rythmiques pourront être ajustés sans pour autant que les bases temporelles extraites aient pour valeur T.

L'exemple suivant illustre l'ajustement d'une requête et de la portion mélodique qu'elle vise, grâce à l'extraction automatique d'une base temporelle, celle-ci étant différente de T.

Nous commencerons par extraire la base temporelle de la requête (point 1), ainsi que du motif mélodique visé (point 2). Connaissant le tempo de ce dernier, nous verrons que la base temporelle extraite ne correspond pas à T, mais à $\frac{T}{3}$. Les deux valeurs obtenues nous permettront ensuite de normaliser les deux motifs rythmiques (point 3). Enfin, une comparaison des deux mélodies rythmiquement ajustées sera opérée (point 4).

1. La requête : La figure 5.2 (a) présente le profil temps/fréquence issu de la transcription d'une requête visant les premières notes de "Jésus que ma joie demeure" (J.S. Bach).



(a) (b)

figure 5.2 : *Profil temps/fréquence (a) d'une requête réelle visant les premières notes de "Jésus que ma joie demeure", (b) issu de la transcription de la mélodie visée.*

Les durées séparant les instants d'apparition des notes sont présentées tableau 5.1 (pour information, les instants de détection des notes chantées sont présentés en Annexe B, tableau B.1 page 180).

Nous avons mis au point une méthode qui, à partir de ces durées, extrait automatiquement une base temporelle. Elle consiste en un regroupement de celles-ci par tailles voisines ; le groupe le mieux représenté (taille population et cohérence des valeurs) fournit, par sa moyenne, la base temporelle élue. Une présentation détaillée de la méthode est disponible en annexe B.

Notre méthode consiste en une mise en œuvre simple d'une approche dite de *clustering*. Le lecteur intéressé par l'extraction automatique d'information rythmique pourra se référer à [Dix00], qui constitue un bon point d'entrée au domaine.

Sur la base des durées présentées tableau 5.1, nous avons obtenu une base temporelle de 0.286 seconde (cf. bas de la première colonne du même tableau).

2. Motif mélodique visé : la figure 5.3 présente la portion du fichier MIDI contenant la musique visée par la requête. Bien que l'écriture musicale soit nette, les instants d'apparition des notes ne sont pas forcément parfaitement réguliers. Ceux-ci, directement extraits du fichier MIDI, fournissent les durées présentées dans le tableau 5.1. A titre d'illustration, nous présentons figure 5.2 (b), le profil temps/fréquence issu de la transcription de la portion MIDI.

figure 5.3 : *Partition correspondant au motif mélodique visé.*

Notre méthode d'extraction fournit une base temporelle de 0.332 seconde. Celle-ci, présente au bas de la deuxième colonne du tableau 5.1, est différente de T qui est égal à 1 seconde, puisque le tempo est fixé à 60 battements par minutes (cf. figure 5.3). La base temporelle extraite est donc égale à $\frac{T}{3}$.

3. Normalisation des durées : Les durées initiales sont normalisées par leurs bases temporelles respectives. Elles sont présentées dans les colonnes 3 et 4 du tableau 5.1.

Rang	Durées initiales		Durées relatives	
	Requête	Référence	Requête	Référence
1	0.275	0.316	0.960	0.951
2	0.265	0.342	0.925	1.030
3	0.280	0.333	0.978	1.003
4	0.275	0.342	0.960	1.030
5	0.280	0.325	0.978	0.978
6	0.285	0.325	0.995	0.978
7	0.305	0.333	1.065	1.003
8	0.300	0.342	1.047	1.030
9	0.280	0.325	0.978	0.978
10	0.270	0.333	0.943	1.003
11	0.300	0.317	1.047	0.954
12	0.300	0.350	1.047	1.054
13	0.300	0.325	1.047	0.978
14	0.295	0.342	1.030	1.030
Base temporelle extraite des durées initiales	0.286	0.332		

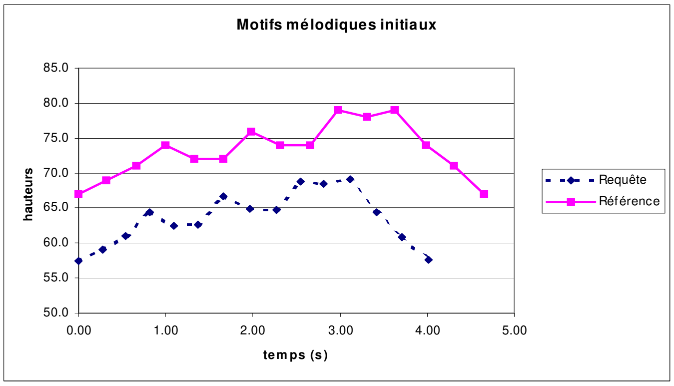
TAB. 5.1 : *Durées inter-débuts de notes pour les deux motifs mélodiques (requête et référence), avant et après application de la base temporelle extraite.*

4. Comparaison des motifs mélodiques rythmiquement ajustés : La figure 5.4 représente deux comparaisons de la requête et du motif mélodique visé. (a) les représente tels qu'ils sont à l'origine, et (b), après application des bases temporelles extraites. Pour information, les instants d'apparitions (déduits des durées normalisées) et les hauteurs correspondants sont présentés tableau B.1 (Annexe B, page 180).

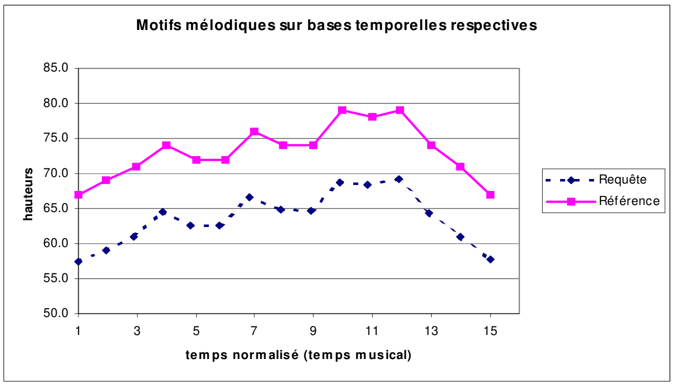
Dans cet exemple, la valeur automatiquement extraite permet l'ajustement temporel des deux mélodies, initialement décalées du fait d'une différence de tempo. L'ajustement temporel rend la similarité des deux mélodies plus flagrante : Les variations de hauteur, proches, sont maintenant quasi-simultanées (une approximation des valeurs temporelles normalisées aux entiers les plus proches mènerait à une parfaite simultanéité).

Ajustement itératif de l'échelle des temps

Nous venons de voir que l'extraction automatique d'une référence pouvait permettre une description temporelle indépendante du tempo emprunté. Une autre solution serait de faire supporter cette invariance à la mesure de similarité. En effet, lors de la comparaison, il serait possible d'itérativement étirer/comprimer l'échelle temporelle de la requête afin de s'adapter au tempo de la portion mélodique recherchée. Le choix du meilleur ajustement serait fondé sur la similarité fréquentielle déduite de chaque itération.



(a)



(b)

figure 5.4 : *Comparaison des motifs mélodiques (a)* avant *application de la base temporelle extraite, (b)* après *application des bases temporelles respectives. Ces dernières sont arbitrairement désignées comme représentant le* temps musical.

Instabilité rythmique des requêtes

Les méthodes d'ajustement temporel vues jusqu'à présent (ajustement par extraction d'une référence, et ajustement itératif) sont fondées sur l'hypothèse d'un tempo stable sur l'ensemble des mélodies comparées. Or, ralentissements, accélérations, troncatures de notes et de silences (phénomènes observés par McNab, cf. Section 3.5) entraînent un *manque de constance* du tempo qui peut les faire échouer. Dans de tels cas, la base temporelle extraite ou l'ajustement effectué ne permettront pas une comparaison efficace. Ces méthodes souffrent d'un manque de robustesse vis-à-vis des variations de tempo.

L'exemple suivant illustre un cas typique d'anticipation de notes. Le motif mélodique visé provient du générique de la série "Mission impossible". Il s'agit d'un motif de flûte dont la partition est présentée figure 5.5.

La figure 5.6 permet de comparer les profils temps/fréquence du motif visé et de la requête.

figure 5.5 : *Partition des 4 mesures correspondants à la mélodie de flûte visée.*

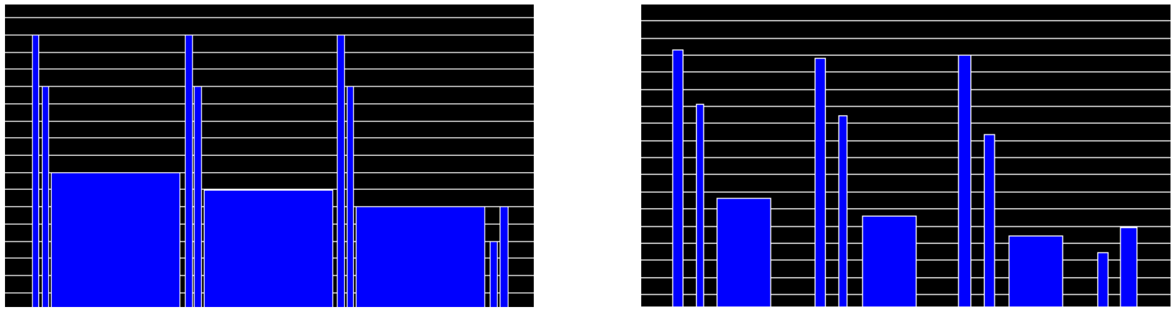


figure 5.6 : *Profil temps/fréquence (a) d'une requête réelle visant les premières notes du motif de flûte dans "Mission impossible", (b) issu de la transcription de la mélodie visée.*

On peut y voir que les couples de notes aiguës sont moins espacés dans (b) que dans (a)[2]. Or, les notes aiguës appartenant à (b) sont plus larges, traduisant un tempo plus lent dans la requête que dans la référence. Cette incohérence traduit le phénomène d'anticipation présent dans la requête : les notes succédant aux notes de rang 3, 6 et 9 sont chantées trop tôt, par rapport à la musique originale visée.

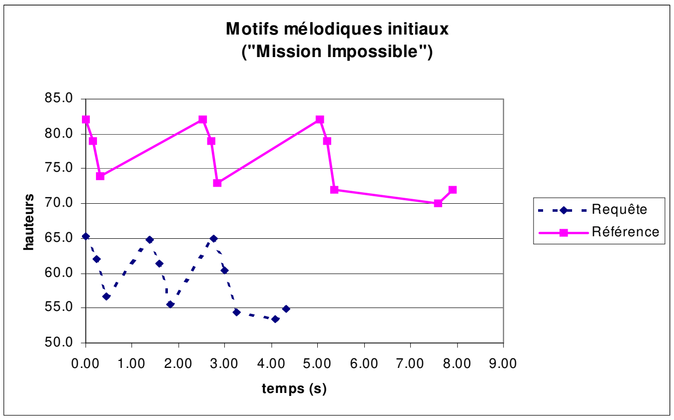
La figure 5.7 représente les comparaisons de la requête et du motif mélodique visé. (a) les représente tels qu'ils sont à l'origine, et (b), après application des bases temporelles extraites. Le détail de l'opération est disponible en Annexe B.

Il apparaît clairement dans cet exemple, que l'extraction d'une base temporelle ne permet pas un ajustement de bonne qualité. En effet, si les notes courtes (qui ont élu aussi bien la base temporelle de la requête que celle de la référence) ont des répartitions comparables sur l'échelle du temps musical[3], les notes longues, elles, révèlent le phénomène d'anticipation des notes présent

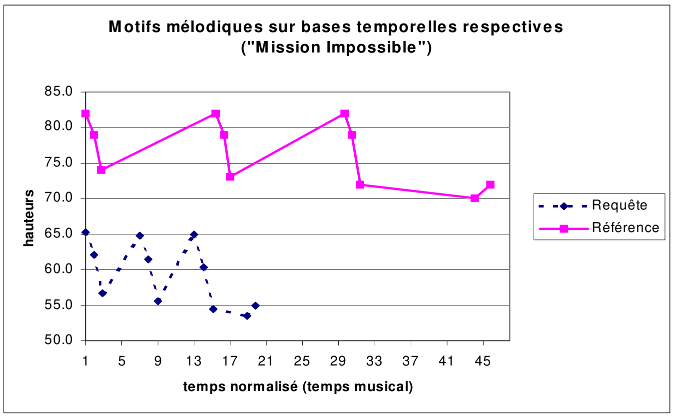
[2]Il s'agit des notes occupant les rangs (1,2), (4,5), et (7,8).

[3]On peut voir que la différence de tempo, particulièrement observable sur les trois premières notes de la figure

dans la requête. Ce dernier, tout comme les autres phénomènes provoquant une inconstance du tempo, met en péril les méthodes d'ajustage temporel.



(a)



(b)

figure 5.7 : *Comparaison des motifs mélodiques (a)* avant *application de la base temporelle extraite, (b)* après *application des bases temporelles respectives. Ces dernières sont arbitrairement désignées comme représentant le* temps musical.

Avec ce genre de différences rythmiques, ni l'extraction d'une base temporelle (appliquée sur cet exemple en Annexe B), ni un ajustement temporel itératif ne permet de témoigner de la similitude (pourtant bien réelle) des deux motifs. Ce type de solutions pourrait néanmoins être exploré plus avant, par exemple, en clarifiant l'importance du phénomène, ou encore, en cherchant non pas *une*, mais *des* bases temporelles, témoignant des éventuelles variations au cours de la requête.

Cependant, nous nous limiterons volontairement sur le sujet afin de favoriser la question de l'aspect *fréquentiel* de la requête. Ce choix rejoint celui de Dowling, dont les travaux ont focalisé sur les principaux traits *fréquentiels* de reproduction et de reconnaissance des mélodies [Dow78]. Nous privilégions ce type d'information dans nos descripteurs mélodiques. Cependant, comme le montre la suite, nous n'abandonnerons pas complètement l'information temporelle contenue dans

5.7(a), a été résorbée par l'application des bases temporelles (cf. trois premières notes de la figure 5.7(b)).

la requête.

Solutions alternatives

Nous venons de le voir, il n'est pas évident de disposer de références temporelles fiables. Sans celles-ci, nous devons renoncer à une représentation temporelle de type *valeur* des notes, qui conserve un maximum d'information tout en s'affranchissant du tempo.

L'alternative consiste à assurer l'indépendance au tempo en se contentant d'une information temporelle *moins riche*. On peut citer les trois possibilités suivantes :

1. Une solution consiste à représenter les notes par le rapport de leur durée et de celle de la note précédente. Dans ce cas, la durée d'une note constitue la base temporelle de la note suivante. Il s'agit donc d'une référence glissante, donc locale. Elle renseigne sur les variations de durée existant d'une note à l'autre. Cette information est indépendante du tempo, lorsqu'il est stable, et peu affectée par ses variations faibles.
   Avec cette solution, la troncation de la durée d'une note (cf. exemple précédent), se traduit par deux différences : la note tronquée est par définition représentée par une valeur moindre, de plus, elle entraîne l'augmentation de la valeur suivante (puisqu'elle en constitue la base de représentation).
   Un brusque changement de tempo n'entraînera qu'une différence. En effet, seule la note suivant le nouveau tempo voit sa durée relative calculée sur la base d'une note appartenant à l'ancien tempo.

   Cette solution peut servir de base à une description plus compacte. En effet, il est possible de conserver uniquement le signe des valeurs obtenues (cf. *contour rythmique* dans 3.3.4). Ce type d'approche ne conserve que le sens de variation des durées successives (plus longue, moins longue, longueur identique). Cette réduction de l'information assure une robustesse aux variations de tempo en perdant de la précision sur le rythme.

2. Une autre solution permettrait d'abandonner le rythme sans pour autant négliger les durées. Ces dernières témoigneraient de l'importance des notes associées. Par exemple, un descripteur de la requête pourrait consister en un histogramme des hauteurs présentes sur la base de leur durée d'apparition au sein du motif traité. L'information conservée serait donc indépendante du tempo (mais sensible à ses variations fortes), et du rythme.

3. Seul l'ordre d'apparition des notes est retenu dans la représentation. Ici, la tolérance est favorisée par rapport à la précision puisque aucune propriété discriminative de l'information temporelle n'est conservée. Cette solution assure l'indépendance au tempo (quelle que soit sa stabilité), au rythme, et aux durées.

Solution adoptée

Dans le système développé au cours de cette thèse, c'est la solution 3 qui a été retenue (ordre d'apparition des notes successives). La description de la requête sera donc indépendante du tempo (même instable), mais également invariante aux imperfections rythmiques. Sa faible précision ne permettra donc pas de traiter ces dernières en fonction de leur importance. Ce choix illustre la volonté de focaliser sur les qualités de l'information fréquentielle des mélodies en s'appuyant sur celle-ci pour établir la similarité mélodique.

Cependant, ce choix pourra, dans les développements ultérieurs à ce travail, être remis en cause. Si la discrimination assurée par l'information fréquentielle conservée s'avérait insuffisante, un enrichissement de l'information temporelle conservée permettrait d'augmenter la sélectivité du système. Ce genre de démarche pourrait être motivée par un volume croissant des données indexées.

5.2.2 Information fréquentielle

A ce stade de l'élaboration de la description de la requête, l'information que nous conservons de la mélodie issue de la transcription est une séquence chronologiquement ordonnée de hauteurs de notes (aucune information concernant la durée des notes n'a été conservée).

Nous allons voir comment représenter l'information fréquentielle de la requête de manière à ce que la similarité mesurée par notre moteur soit indépendante du ton, tolère les imperfections observées au chapitre 4, tout en gardant la précision nécessaire à l'identification.

S'affranchir du ton

Les hauteurs issues de la transcription dépendent du ton emprunté par l'utilisateur. Or, si l'on peut s'attendre à ce qu'un utilisateur connaisse approximativement le motif fréquentiel de la mélodie qu'il recherche, il est difficile de lui demander de chanter sa requête au ton original. Si celui-ci était connu, le descripteur de la requête pourrait en être affranchi (par l'expression des valeurs de hauteur en référence à celui-ci, cf. 3.3.2).

Dans un premier temps, nous allons voir différentes manières de fournir au système ce type de référence, puis nous verrons différentes manières de s'en passer.

Référence fréquentielle fournie par l'utilisateur

Le fait de demander à ce que la requête chantonnée soit complétée par des informations sur celle-ci réduit le nombre des utilisateurs potentiels. Si leur demander le *tempo* des requêtes soumises constitue déjà une limitation sensible, demander le *ton* revient à ne s'adresser qu'à des utilisateurs musiciens. Demander une référence fréquentielle à l'utilisateur est donc tout à fait contraire à la démarche adoptée dans cette thèse. En effet, si la requête chantonnée est intéressante, c'est précisément parce qu'elle permet aux non-musiciens, incapables de formuler textuellement une "requête mélodique", d'effectuer une recherche de musique par l'exemple.

De plus, à la différence du tempo, le ton n'est pas spécifié par le format MIDI[4]. L'information manquerait donc pour les mélodies de la base de données. Cela nous amène directement à envisa-

[4] L'armure peut l'être mais étant optionnelle, elle est loin d'être systématiquement présente.

ger l'extraction automatique d'une référence fréquentielle.

Extraction automatique d'une référence fréquentielle

L'extraction automatique d'une référence fréquentielle pose, comme nous allons le voir, d'avantage de problèmes que l'extraction d'une référence temporelle.

Comme pour l'extraction automatique d'une référence temporelle, la question de la stabilité des mélodies traitées est essentielle. Par analogie à ce qui a été vu concernant l'aspect temporel de la requête, nous allons détailler les deux notions constituant la stabilité de la référence fréquentielle que l'on cherche à extraire.

A l'image d'une musique accélérant ou ralentissant, la première notion concerne la capacité à conserver un référentiel fréquentiel constant (un même tempérament). Lorsque cela n'est pas le cas, on assiste au phénomène de glissement de ton (GT) évoqué par McNab. Cependant, aux vues du corpus étudié au Chapitre 4, ce phénomène nous a paru peu répandu (du moins dans une forme flagrante). Dans ce qui suit, nous considèrerons donc cette stabilité "fréquentielle" comme assurée.

La deuxième notion de stabilité est musicale. Elle est liée à l'aspect "ouvert" de la définition du ton[5] qui, nous le rappelons, est hauteur représentative de la "hauteur globale" d'une musique. Or, lorsque l'on considère une portion mélodique, le ton perçu peut être différent de celui que l'on désignerait pour l'ensemble de la mélodie[6]. Il en était de même dans l'extraction automatique d'une référence temporelle, où nous pouvions nous contenter d'une valeur différente de T ($2T$, $\frac{T}{2}$, $\frac{T}{3}$...), pour peu que nous la retrouvions à la fois dans la requête et dans la portion mélodique visée.

Seulement, l'extraction automatique du ton sur des courtes mélodies semble être beaucoup plus sensible aux changements que ne l'est l'extraction automatique d'une base temporelle. Un problème d'*ambiguïté* est observé au niveau des résultats. Ce problème est causé par une information disponible insuffisante. Comme nous allons le voir, l'extraction automatique d'une référence fréquentielle semble difficilement exploitable pour assurer l'indépendance des descripteurs mélodiques vis-à-vis du ton.

Application d'une extraction automatique du ton d'une mélodie

A titre d'illustration, nous allons comparer les tons automatiquement extraits de trois mélodies. Il s'agit de deux requêtes chantonnées, et de la portion mélodique qu'elles visent. Le motif mélodique visé est le début de la partie du chant de "Amsterdam" (J. Brel)[7]. Nous commencerons par traiter la référence (point 1), puis nous passerons aux deux requêtes (points 2 et 3).

Notre méthode est une version simplifiée de la méthode d'extraction de la tonalité de Krumhansl [Kru90], dont nous avons adapté le principe aux hauteurs non tempérées[8]. Pour trouver la

[5]Idéalement, la référence fréquentielle recherchée est le ton, tout comme la référence temporelle recherchée dans la section précédente était T, soit l'inverse du tempo.

[6]Cela explique l'insuffisance de la connaissance de l'armure, parfois fournie dans un fichier MIDI.

[7]Il s'agit des douze premières notes, correspondant aux paroles : "Dans le port d'Am-ster-dam, y'a des ma-rins qui chantent"

[8]Nos observations nous laissent penser que notre algorithme est représentatif des performances (et des limitations) que l'on peut attendre de l'algorithme original sur des mélodies courtes.

tonalité la mieux représentée, chaque note de la requête est tour à tour prise comme ton potentiel. Deux modes (majeur et mineur) entraîneront au plus deux scores par ton considéré. La tonalité gagnante est celle ayant le plus haut score. Le détail de la méthode est présenté en Annexe C.

1. Référence : portion mélodique visée (les hauteurs sont tempérées). Le Tableau 5.2 présente les résultats du processus d'extraction de tonalité. La première colonne ("Rang") indique la position de chaque note au sein de la mélodie. Les deux suivantes ("Note" et "# MIDI") désignent le nom des notes et le numéro MIDI correspondant. La troisième colonne témoigne du *mode* proposé lorsque la note correspondante est considérée comme étant le ton (notation : "m" désigne le mode mineur et "M" désigne le mode majeur). La quatrième colonne renseigne sur l'importance, au sein de la mélodie, de la tonalité étudiée. La tonalité élue étant celle qui obtient le plus grand score, il y a ici deux premiers *ex aequo* : les tonalités "ré mineur" et "la♯ majeur". Il existe donc une ambiguïté dans le ton automatiquement extrait. Les deux hauteurs candidates au ton, "ré" et "la♯", se retrouvent respectivement aux rangs 1 et 2, et aux rangs 6 et 8.

Le Tableau 5.3 présente les scores obtenus pour chacune des hauteurs successivement considérées comme ton, pour les deux requêtes suivantes.

Rang	Note	# MIDI	Mode	Score
1	ré	62	m	**7**
2	ré	62	m	**7**
3	sol	67	m	6
4	sol	67	m	6
5	la	69	m	3
6	la♯	70	M	**7**
7	do	72	?	3
8	la♯	70	M	**7**
9	la	69	m	3
10	fa	65	M	6
11	fa	65	M	6
12	fa	65	M	6

TAB. 5.2 : *Extraction de tonalité d'*Amsterdam *version tempérée ("m" = mineur et "M" = majeur).*

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Référence-Ton1	0	0	5	5	7	8	10	8	7	3	3	3
Référence-Ton2	-8	-8	-3	-3	-1	0	2	0	-1	-5	-5	-5
Requête1	-2.52	-2.88	1.79	1.83	3.8	4.86	7.1	5.58	4.04	-0.12	0	0.07
Requête2	-7.94	-8.05	-3.41	-2.96	-1.24	0	2.24	0.18	-1.16	-4.59	-5.1	-5.11

TAB. 5.4 : *Descripteurs après centrage par la valeur du ton extrait (La portion mélodique tempérée ayant deux résultats* ex aequo*, deux centrages différents sont proposés : "Référence-Ton1" et "Référence-Ton2").*

Requête 1				Requête 2			
Rang	# MIDI	Mode	Score	Rang	# MIDI	Mode	Score
1	56.11	m	1.86	1	55.89	m	4.38
		M	2.32				
2	55.75	m	4.66	2	55.78	m	5.04
3	60.42	m	3.12	3	60.42	m	1.62
		M	2.84			M	1.62
4	60.46	m	3.28	4	60.87	m	4.52
		M	2.84				
5	62.43	m	1.92	5	62.59	m	1.88
6	63.49	M	3.74	6	63.83	M	**5.18**
7	65.73	?	1.84	7	66.07	?	1.6
8	64.21	M	1.88	8	64.01	M	4.34
9	62.67	m	2.4	9	62.67	m	2.04
10	58.51	M	4.46	10	59.24	M	2.1
11	58.63	M	**4.94**	11	58.73	M	3.9
12	58.70	M	4.82	12	58.72	M	3.92

TAB. 5.3 : *Extraction de tonalité sur deux requêtes visant* Amsterdam *(hauteurs non tempérées).*

2. Requête chantée 1 : Parmi les 12 hauteurs constituant cette requête, notre méthode d'extraction a désigné celle de rang 11 comme étant le ton de la tonalité la mieux représentée (cf. 4ème colonne du tableau 5.3).

   Cette solution ne fait pas partie de celles obtenues avec la portion tempérée visée. Nos deux premiers exemples ne pourront donc pas être ajustés avec succès par notre méthode d'extraction de ton.

3. Requête chantée 2 : Dans ce dernier exemple, la note de rang 6 ayant le score le plus élevé, sa hauteur correspond au ton élu.

   Cette solution est différente de celle de la requête chantonnée précédente (point 2). Si nous avions à ajuster ces deux mélodies (initialement proches), ce serait un échec. Par contre, la solution obtenue est commune avec l'une de celles de la mélodie référence (point 1). L'indépendance au ton pourrait être assurée par notre méthode, à condition de savoir choisir la bonne solution (note de rang 6 ou 8), ou de considérer les deux cas.

La figure 5.8 présente les séquences de hauteurs originales de nos trois exemples. On y perçoit la proximité des deux requêtes chantonnées, et de la portion tempérée (à une différence de ton près). La figure 5.9 présence ces mêmes séquences de hauteurs, une fois centrées par leur ton (automatiquement extrait). Les valeurs correspondantes sont présentées tableau 5.4.

L'extraction de ton appliquée sur ces trois mélodies n'apparaît pas vraiment efficace pour assurer une indépendance aux tons empruntés. Les deux requêtes (exemples 2 et 3) initialement

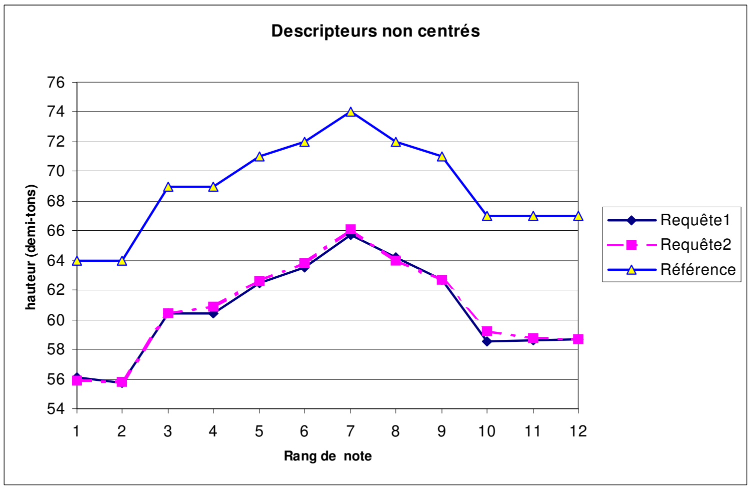


figure 5.8 : *Comparaison des séquences de hauteurs non centrées. Requête1 et Requête2 étant chantées à la même tonalité, la comparaison visuelle de leurs séquences de hauteurs est aisée.*

proches sont, après centrage, séparées de 5 demi-tons ! Quant à la différence de tonalité initialement observée entre requêtes et portion tempérée, elle n'est complètement résorbée que pour un cas ("Référence-Ton2" et "Requête2"). Pour les autres, l'écart reste inacceptable (3, 5, et 8 demi-tons). Si l'on avait à comparer ces mélodies (perceptivement proches), les différences de valeurs observées sur les descripteurs centrés sont telles qu'elles seraient pour la plupart jugées - à tort - dissimilaires.

Ce type d'extraction automatique du ton ne semble pas permettre aux descripteurs de s'affranchir du ton emprunté. Le phénomène d'instabilité mentionné par Krumhansl dans le cas d'un faible nombre de notes (page 80 dans [Kru90]), s'illustre notamment par la présence de premiers *ex aequo* dans l'exemple 1. Le fait d'avoir adapté le principe aux hauteurs non tempérées permet d'éviter les ambiguïtés, mais cette unicité du résultat s'accompagne d'une sensibilité trop importante à la valeur des hauteurs traitées (résultats très différents pour les exemples 2 et 3 pourtant initialement très proches).

Ce manque de fiabilité des résultats d'extraction du ton semble dû à la difficulté de la tâche. En effet, une courte mélodie monophonique constitue bien peu d'information pour extraire la tonalité de la musique dont il est issu.

Il apparaît donc difficile d'assurer l'indépendance au ton emprunté par l'extraction automatique d'une référence fréquentielle. Des solutions existent pour s'affranchir du ton, mais sans pour autant le déterminer. Nous allons voir que la difficulté d'extraire le ton (via la tonalité) présente des conséquences plus difficilement contournables.

Conséquences de l'ignorance de la tonalité des mélodies à comparer

Dowling l'a montré [Dow78], le contexte mélodique est, avec le contour, un élément essentiel dans la perception de la similarité mélodique. En effet, la perturbation entraînée par une erreur n'est pas forcément proportionnelle à son amplitude. Ainsi, une hauteur erronée qui appartiendrait

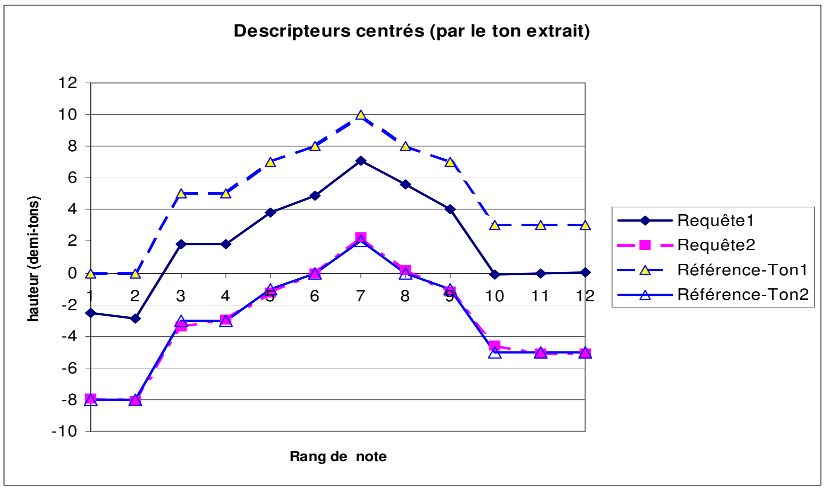


figure 5.9 : *Comparaison des séquences de hauteurs centrés par leur ton (automatiquement extrait) respectif.*

toujours au mode de la mélodie devrait voir son influence amoindrie.

Malheureusement, comme nous l'avons vu, l'ignorance de la tonalité nous empêche de définir un descripteur représentant les hauteurs dans un *référentiel fondé sur le contexte mélodique*. Nous serons donc contraints à nous fier uniquement à l'*amplitude* des différences entre portions mélodiques comparées.

Voyons maintenant comment nous pouvons, malgré l'échec de l'extraction automatique du ton, assurer l'indépendance de la comparaison mélodique à celui-ci.

Solution alternative

Une *description mélodique* qui assure une indépendance au ton peut consister en l'*écart* séparant les hauteurs successives, soit les intervalles. Comme nous l'avons vu au Chapitre 3, cette solution est largement utilisée (cf. 3.3.4).

Dans ce cas, la hauteur de chaque note constitue la référence fréquentielle de la hauteur suivante. Il s'agit donc d'une référence glissante, donc locale. Elle renseigne sur les variations de hauteur existant d'une note à l'autre. Cette description est donc indépendante du ton, lorsqu'il est stable, et peu affectée par ses variations progressives. En effet, les erreurs d'un GT ne s'accumulent pas comme elles le feraient avec une référence fréquentielle unique pour l'ensemble de la mélodie décrite.

Deux solutions envisagées

Afin de considérer deux facettes de l'information fréquentielle (information absolue et relative), nous n'abandonnerons pas la piste de la description par séquences de hauteurs, baptisées *profils de hauteurs*. Puisque cette dernière n'assure pas l'indépendance au ton, nous chargerons la *mesure de similarité* de gérer cette indispensable tolérance. Nous considèrerons également la solution alternative, c'est-à-dire la description par *séquences d'intervalles*. Les descripteurs mélodiques correspondants aux deux voies dégagées sont explicités ci-dessous, pour une mélodie de N notes.

1. **Profil de hauteurs**
   Le descripteur mélodique *Profil de hauteurs* est un vecteur contenant les N hauteurs des notes successives (exprimées en demi-ton). L'information est absolue donc dépendante du ton emprunté. L'élément de base du descripteur est la note.

2. **Séquence d'intervalles**
   Le descripteur mélodique *séquence d'intervalles* est un vecteur contenant les $N-1$ intervalles séparant les N hauteurs successives (exprimées en demi-ton). L'information est relative donc indépendante du ton emprunté. L'élément de base du descripteur est l'intervalle, qui lie deux notes consécutives.

Nous allons maintenant considérer la question de la quantification de l'information fréquentielle. Au Chapitre 3, nous avons vu qu'elle était systématiquement appliquée (cf. 3.5.2). Motivée par un mimétisme (inconscient) vis-à-vis des données (tempérées) de la base, la quantification de l'information fréquentielle de la requête s'accompagne de certains inconvénients.

Faut-il quantifier les données fréquentielles ?

La précision de la transcription est telle que les hauteurs disponibles peuvent être considérées comme appartenant à une échelle continue. Quantifier les hauteurs en se ramenant à l'échelle discrète des demi-tons a l'avantage de ramener au format occidental de représentation musicale. Cependant, représenter des valeurs appartenant à un domaine "continu" par celles d'un domaine discret entraîne leur *approximation*. Les hauteurs d'une requête chantonnée présentant une certaine imperfection, cette approximation engendre, comme nous allons le voir par la suite, deux types de conséquences antagonistes. En effet, la quantification a l'avantage d'*annuler* certaines imprécisions, mais en contrepartie, elle présente l'inconvénient d'en *aggraver* d'autres.

La question de la quantification vient compliquer le choix d'un descripteur, déjà ouvert aux deux possibilités *profil de hauteurs* et *séquence d'intervalles*. Soit h, la séquence des hauteurs issue du module d'analyse, et soient les opérations $Q(.)$ de quantification des valeurs, $I(.)$ d'extraction d'une séquence d'intervalles à partir d'une séquence de hauteurs, et $H(.)$ de construction d'une séquence de hauteurs à partir d'une séquence d'intervalles. La combinaison des possibilités permet d'envisager les six cas présentés dans le Tableau 5.5.

	Profils de hauteurs	Séquences d'intervalles
Non Quantification	h	$I(h)$
Quantification	$Q(h)$	$Q[I(h)]$
	$H[Q[I(h)]]$	$I[Q(h)]$

TAB. 5.5 : *Différentes descripteurs mélodiques possibles selon la quantification de l'information fréquentielle.*

De manière explicite, nous avons :

1. h : le profil de hauteurs non quantifiées ;
2. $Q(h)$: le profil de hauteurs quantifiées ;
3. $I[Q(h)]$: la séquence d'intervalles issus des hauteurs quantifiées ;
4. $I(h)$: la séquence d'intervalles non quantifiées ;
5. $Q[I(h)]$: la séquence des mêmes intervalles, mais quantifiés. En effet, opérer la quantification des données fréquentielles sur les hauteurs *puis* calculer les intervalles ne donne pas forcément le même résultat que calculer les intervalles à partir des hauteurs non quantifiées *puis* appliquer une quantification ;
6. $H[Q[I(h)]]$: le profil de hauteurs reconstruit à partir des intervalles quantifiés précédents.

Nous allons maintenant présenter différentes manières de quantifier l'information fréquentielle de la requête. Comme nous venons de le voir, la quantification peut être appliquée sur une information absolue (hauteurs) ou relative (intervalles).

Application d'une quantification sur les hauteurs

Le point essentiel dans la quantification appliquée aux hauteurs concerne le positionnement des valeurs à quantifier par rapport à l'échelle témoignant des valeurs autorisées. Nous allons voir successivement quatre moyens différents de quantifier les hauteurs issues d'une mélodie non tempérée.

1. La première voie consiste en une quantification brutale des hauteurs, appliquée sur la base de leur comparaison avec l'échelle des demi-tons calée sur le diapason (i.e. son de référence, soit généralement 440Hz). Sur les hauteurs, ce type de quantification n'est pas souhaitable. En effet, selon le ton employé par la mélodie non tempérée, le résultat pourra varier sensiblement. Le pire cas consiste en un ton situé juste au milieu de deux valeurs tempérées. Dans ce cas, la plus faible imprécision pourra faire basculer la hauteur quantifiée d'un demi-ton à son voisin. L'interprétation engendrée par la quantification pourra être catastrophique.

   A titre d'illustration, nous allons effectuer la quantification brutale des hauteurs d'une mélodie non tempérée. Il s'agit de la première requête chantonnée utilisée pour illustrer l'extraction automatique du ton. Les valeurs des hauteurs transcrites sont rappelées tableau 5.6 (deuxième ligne). L'arrondi brutal de ces valeurs est présenté dans la ligne suivante. On peut observer quatre différences (notes de rang 3 à 6) par rapport à la référence. On peut également voir qu'à 0.02 demi-ton près, les hauteurs des rangs 6 et 10 (respectivement 63.49 et

58.51) auraient provoqué deux erreurs supplémentaires.

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Requête	56.11	55.75	60.42	60.46	62.43	63.49	65.73	64.21	62.67	58.51	58.63	58.70
Q directe	56	56	**60**	**60**	**62**	**62**	66	64	63	59	59	59
Référence	56	56	61	61	63	64	66	64	63	59	59	59

TAB. 5.6 : *Quantification brutale des hauteurs d'une requête chantée ("Amsterdam").*

Un *ajustement* préalable des hauteurs de la mélodie chantée par rapport à une échelle des demi-tons est donc souhaitable si l'on désire opérer une quantification quelque peu "maîtrisée".

2. Nous avons vu sous-section 4.3.6 que le premier intervalle d'une requête chantonnée comportait une imprécision supérieure à la moyenne. La précision des suivants étant liée à la mélodie chantée, il serait dangereux de désigner un rang de note donné pour que celle-ci serve systématiquement de référence pour l'ajustement. En effet, l'erreur associée se répercuterait sur le reste de la mélodie, entraînant des résultats potentiellement aussi mauvais que ceux d'une quantification brutale.

   Pour illustrer ce cas, reprenons la requête vue en 1 et choisissons une référence pour l'ajustement parmi les hauteurs disponibles. En l'absence de réel critère de choix (à part éviter les deux premières notes), nous sélectionnons arbitrairement la note de rang 8. Celle-ci ayant pour hauteur 64.21, nous allons ajuster le vecteur descripteur en soustrayant à l'ensemble de ses éléments la valeur 64.21-64=0.21 (séparant la hauteur choisie du demi-ton le plus proche). Le résultat est présenté tableau 5.7.

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Requête	56.11	55.75	60.42	60.46	62.43	63.49	65.73	64.21	62.67	58.51	58.63	58.70
Req. ajustée	55.90	55.54	60.21	60.25	62.22	63.28	65.52	64.00	62.46	58.30	58.42	58.49
Q 8e note	**56**	**56**	60	60	62	63	**66**	**64**	62	58	58	58
Référence bis	55	55	60	60	62	63	65	63	62	58	58	58

TAB. 5.7 : *Quantification d'une requête chantée ("Amsterdam") précédée d'un ajustement fondé sur le choix arbitraire d'une référence parmi les hauteurs.*

La différence de ton ne devant pas constituer un facteur aggravant la dissimilarité, nous avons transposé la référence précédemment utilisée en soustrayant un demi-ton de l'ensemble de ses hauteurs. Nous comptabilisons ainsi le nombre minimum d'erreurs provoquées par la quantification appliquée. Celui-ci s'élève à quatre.

Ce résultat est donc aussi mauvais que celui obtenu avec une quantification brutale. Si nous avions pris la 7ème note comme référence, nous aurions obtenu un résultat parfaitement conforme à la référence. Cela illustre l'aspect aléatoire de cette méthode qui dépend de la qualité (variable et inconnue) de la hauteur choisie en référence pour l'ajustement.

3. Nous l'avons vu au Chapitre 3, McNab propose un ajustement adaptatif de l'échelle des demi-tons, menant à une quantification plus maîtrisée. Cependant, cette démarche, conçue pour palier aux Glissements de Ton, peut avoir un effet indésirable sur d'autres types d'erreur.

   Le tableau 5.8 présente les résultats de cette quantification appliquée à notre requête "Amsterdam".

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Requête	56.11	55.75	60.42	60.46	62.43	63.49	65.73	64.21	62.67	58.51	58.63	58.70
Req. ajustée	56.11	55.64	60.67	61.04	62.97	64.06	66.24	64.48	62.46	57.84	58.12	58.07
Q McNab	56	56	61	61	63	64	66	64	**62**	**58**	**58**	**58**
Référence	56	56	61	61	63	64	66	64	63	59	59	59

TAB. 5.8 : *Quantification d'une requête chantée ("Amsterdam") précédée de l'ajustement adaptatif de McNab.*

   On peut noter que la présence d'une Erreur Locale provoque le décalage des quatre dernières notes quantifiées (entraînant une Rupture de Ton dans la requête quantifiée).

   En effet, la hauteur (de rang) 8 est imprécise puisque qu'elle est à 1.52 demi-ton de la hauteur 7 (contre les 2 demi-tons requis par la référence). Par ailleurs, puisque les hauteurs 7 et 9, séparées de 1.52+1.54=3.06 demi-tons, sont conformes à la référence (3 demi-tons), la hauteur 8 s'avère être une imprécision passagère (une EL donc).

   L'ajustement adaptatif de McNab ne témoigne donc pas de l'EL présente, mais d'un autre type d'erreur (RT). La spécificité de cette méthode, conçue pour suivre les GT, peut donc entraîner une méprise sur la nature de la requête.

   Il faut cependant noter que cette méthode est, à notre connaissance, la seule à prendre en compte le fait que l'hypothèse de stabilité du ton puisse ne pas être respectée.

4. Nous proposons une méthode originale pour l'ajustement préalable à la quantification. Plus précisément, cette méthode permet d'indiquer parmi les hauteurs de la requête, la référence pour l'ajustement. Le but est de palier aux inconvénients des quantifications rudimentaires (cf. 1 et 2) sans pour autant favoriser un type d'erreur particulier (cf. 3).

   Notre méthode est fondée sur l'extraction automatique de la tonalité d'une mélodie utilisée plus tôt (cf. page 99), et décrite en Annexe C. La référence indispensable à l'ajustement du motif à quantifier est la hauteur correspondant au ton extrait.

   L'intérêt de notre méthode réside dans le fait que la quantification ainsi effectuée assure une approximation avantageuse des notes "importantes" de la tonalité élue.

   Parmi les 12 hauteurs constituant cette requête, notre méthode d'extraction a désigné celle de rang 11 comme étant le ton de la tonalité la mieux représentée (cf. Annexe C). C'est donc cette hauteur qui va orienter l'ajustement : toutes les hauteurs de la requête vont être déplacées de 59-58.63=0.37 demi-ton vers le haut (écart séparant la hauteur choisie du demi-ton le plus proche). Le résultat de cet ajustement, ainsi que la quantification qui en découle sont présentés tableau 5.9.

Rang	1	2	3	4	5	6	7	8	9	10	11	12
Requête	56.11	55.75	60.42	60.46	62.43	63.49	65.73	64.21	62.67	58.51	58.63	58.70
Req. ajustée	56.48	56.12	60.79	60.83	62.80	63.86	66.10	64.58	63.04	58.88	59.00	59.07
Q ton	56	56	61	61	63	64	66	**65**	63	59	59	59
Référence	56	56	61	61	63	64	66	64	63	59	59	59

TAB. 5.9 : *Quantification d'une requête chantée ("Amsterdam") précédée d'un ajustement fondé sur le ton automatiquement extrait.*

Sur cet exemple, notre méthode donne donc le meilleur résultat puisqu'une seule hauteur diffère de la référence. Contrairement à la méthode précédente, l'erreur locale que constitue la hauteur 8 est traitée en tant que telle (elle ne provoque qu'une erreur dans la requête quantifiée).

Application d'une quantification sur les intervalles

La quantification sur les intervalles est beaucoup plus aisée puisque, l'information étant relative, l'ajustement n'est plus nécessaire. Par conséquent, elle constitue la méthode la plus utilisée. Le tableau 5.10 présente le résultat de cette quantification appliquée à notre exemple.

Rang	1	2	3	4	5	6	7	8	9	10	11	12
$I(h)_{Req}$	-0.36	4.67	0.04	1.97	1.06	2.24	-1.52	-1.54	-4.16	0.12	0.07	-
Requête	56.11	55.75	60.42	60.46	62.43	63.49	65.73	64.21	62.67	58.51	58.63	58.70
$H[Q[I(h)]]_{Req}$	0	5	0	2	1	2	-2	$-\mathbf{2}$	-4	0	0	-
$H[Q[I(h)]]_{Ref}$	0	5	0	2	1	2	-2	-1	-4	0	0	-

TAB. 5.10 : *Quantification directe des intervalles d'une requête chantée ("Amsterdam").*

On observe donc une seule *erreur d'intervalle*, située sur celui de rang 8. Elle correspond à une RT d'amplitude 1 demi-ton. Ce résultat est identique à celui obtenu avec la méthode 3. La quantification brutale sur les intervalles peut donc également modifier la nature des erreurs présentes sur une requête, mais sans offrir de parade à l'instabilité tonale.

La figure 5.10 illustre les résultats obtenus avec les méthodes présentées.

Conséquences des approximations liées à la quantification

Chacune des quantifications que nous venons d'illustrer présente deux types de conséquences antagonistes. Lorsque la valeur quantifiée rejoint la valeur visée, on assiste à une *annulation de l'imprécision* (e.g. Rang 1, tableau 5.10). Dans le cas contraire (valeur quantifiée différente de la valeur visée), on assiste à une *aggravation de imprécision* (e.g. Rang 8, tableau 5.10).

Or, les observations effectuées au Chapitre 4 nous ont montré qu'une proportion non négligeable des imprécisions ne seraient pas annulées par une quantification au demi-ton (cf. p. 82). En fait, d'après les mélodies chantonnées de notre corpus, entre 25 et 30% des intervalles comportent une imprécision supérieure au quart de ton.

Cela dit, l'aggravation de certaines erreurs n'est pas le seul désavantage de la quantification. En effet, les systèmes de recherche de musique par chantonnement qui appliquent une quantification des hauteurs -ou des intervalles- de la requête chantée, retournent des réponses possédant deux inconvénients :

– D'une part, la quantification de l'information fréquentielle s'accompagne d'une perte de la précision. Celle-ci entraîne une baisse de la discrimination, illustrée par une augmentation d'*ex aequo* dans la liste des réponses.

– D'autre part, la tolérance inéquitable qui accompagne la quantification (cf. 3.5.2) confère un caractère "chaotique" à la liste des réponses d'un système de recherche. En effet, deux requêtes présentant d'infimes variations de hauteur[9] peuvent mener (de par les différences de quantification engendrées) à des réponses très différentes de la part du système de recherche. Comme nous l'avons illustré en 3.3.3 pour le contour mélodique, cette sensibilité n'est pas justifiée par des contextes mélodiques particuliers.

[9] 1 *cent* - soit 1/100ème de demi-ton - peut suffire

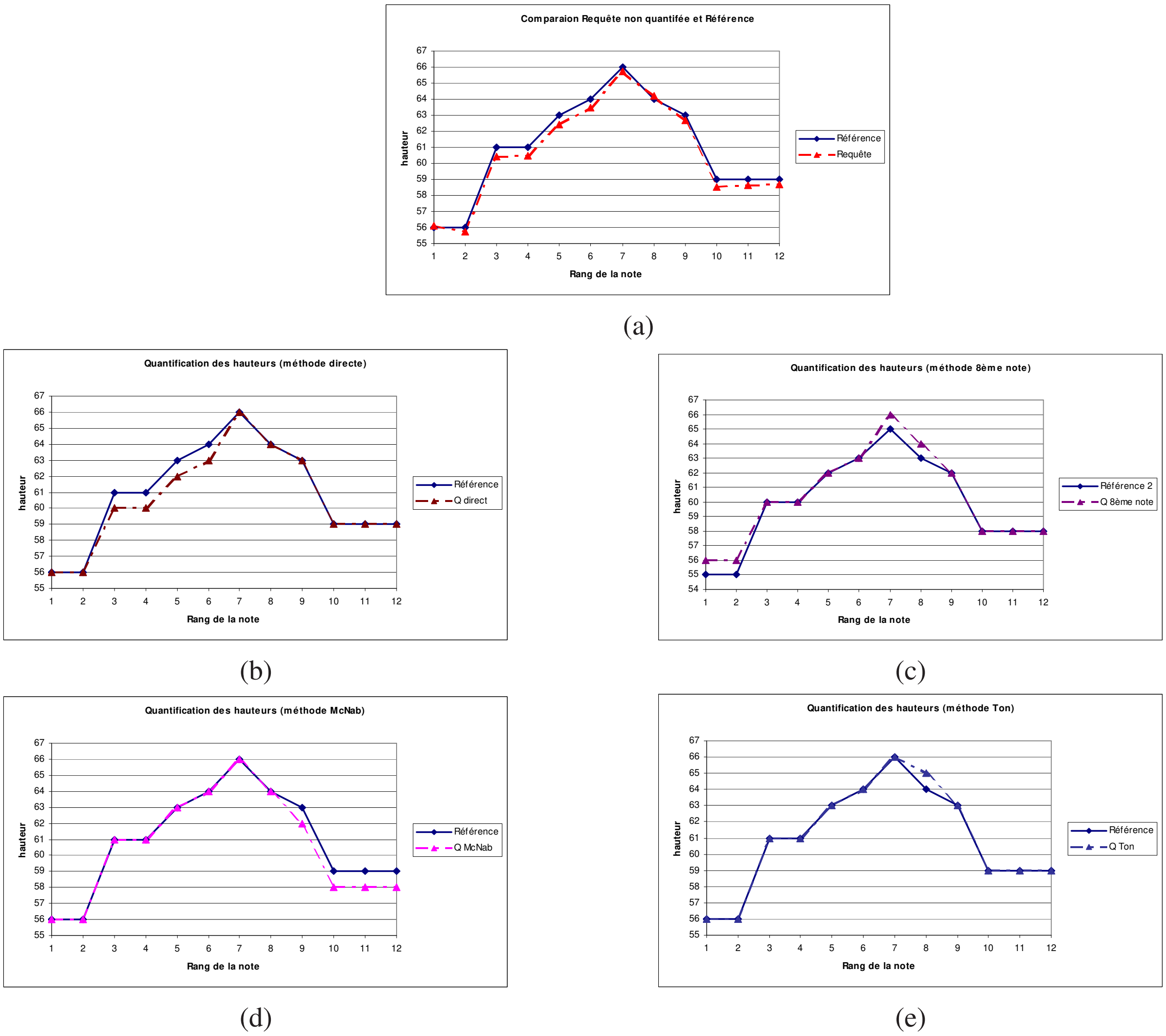


figure 5.10 : *Illustration des différentes méthodes de quantifications présentées. (a) permet de comparer les hauteurs de la requête et celles du motif mélodique visé ("Référence", toujours en trait plein) ; (b) présente le résultat d'une quantification directe ; (c) celui d'une quantification précédée d'un ajustement fondé sur la 8ème hauteur ; (d) celui d'une quantification précédée de l'ajustement adaptatif de McNab (résultat similaire à la quantification directe des intervalles) ; (e) celui d'une quantification précédée de l'ajustement fondé sur la hauteur associée au ton automatiquement extrait.*

Refus de la quantification de l'information fréquentielle

Un choix inédit (à notre connaissance) consiste en la *conservation des valeurs de hauteur exactes* issues de l'analyse de la requête chantée. Ainsi, l'absence de quantification permet d'éviter l'interprétation de la requête au niveau de sa représentation. En recevant les valeurs brutes fournies par le module d'analyse du chantonnement, la mesure de similarité dispose du maximum de précision sur l'information de hauteur. La sensibilité aux erreurs de transposition est ainsi maîtrisée grâce à la gestion du phénomène par la seule mesure de similarité.

Les deux possibilités restantes sur les six citées en début de section 5.2.2 sont donc

1. *h : profil de hauteurs non tempérées*. Le descripteur de la requête est un vecteur de taille N, contenant les valeurs non tempérées des N hauteurs successives de la mélodie chantonnée ;

2. $I(h)$ *: séquence d'intervalles non tempérés*. Le descripteur de la requête est un vecteur de taille $N-1$, contenant les intervalles non tempérés issus des N hauteurs successives de la mélodie chantonnée.

5.2.3 Conclusion

Dans cette partie, nous avons décrit le cheminement effectué pour aboutir à une description de la requête. Concernant l'information temporelle, nous avons choisi de conserver uniquement l'*ordre d'apparition des notes*. Les conséquences sont une indépendance de la description au tempo (même instable), au rythme et aux durées de notes. La description assure donc une grande tolérance vis-à-vis des imprécisions temporelles. Cependant, l'information conservée est trop faible pour pouvoir juger de leur importance. La sensibilité aux imprécisions temporelles sera donc grossière (binaire).

Concernant l'information fréquentielle, nous avons décidé de conserver la précision fournie par le module d'analyse, *nous n'opérons donc pas de quantification* sur les valeurs de hauteur. Cette précision maximum assure une forte discrimination tout en évitant la tolérance inéquitable qui accompagne la quantification. La sensibilité aux erreurs sera donc entièrement déléguée à la mesure de similarité. A la différence de l'aspect temporel, la richesse de l'information qui lui sera fournie lui permettra une finesse dans la sensibilité aux erreurs (malgré l'absence de prise en compte du contexte musical). Notre description mélodique favorise donc nettement l'information fréquentielle des mélodies.

Le choix *profil de hauteurs* vs. *séquence d'intervalles* n'a pas encore été fixé. Nous envisagerons donc deux types de descriptions dans ce qui suit (information fréquentielle absolue et relative). Notons que le premier type étant dépendant du ton emprunté, la mesure de similarité devra gérer les différences de ton, en plus de la sensibilité aux erreurs.

5.3 Description des données de la base

Nous l'avons évoqué au Chapitre 3, la transcription automatique de musiques *polyphoniques* ne fournit pas encore de résultats satisfaisants. A l'instar des systèmes rencontrés jusqu'ici, notre système de recherche de musique par chantonnement s'adresse à une base de données constituée de fichiers MIDI. Ceux-ci nous permettent de disposer de la hauteur et de la durée des notes, ainsi que des instruments qui les produisent.

Les descripteurs des musiques de la base de données et le descripteur de la requête chantonnée sont destinés à être comparés. Par conséquent, les choix précédemment effectués pour la description de la requête vont intervenir dans celle des données de la base.

5.3.1 Sélection des données concernées par la description

Les fichiers MIDI constituant notre base sont en écrasante majorité polyphoniques, ce qui n'est pas le cas de la requête. Cette polyphonie se constate à deux niveaux. D'une part, plusieurs instruments peuvent intervenir simultanément au sein d'une même musique (e.g. un piano accompagnant une flûte). D'autre part, un instrument peut produire plusieurs notes simultanément (e.g. un accord de guitare).

Compte-tenu de la description précédemment adoptée pour la requête, la mise en correspondance de cette dernière, monophonique, avec une musique polyphonique ne peut s'effectuer directement. Heureusement, les instruments jouant *simultanément* peuvent être traités *séparément* grâce au codage MIDI qui assigne une piste à chaque instrument.

Ainsi, il est possible d'éviter de prendre en compte dans la description les pistes de batterie, pour lesquels les numéros MIDI ne représentent pas des hauteurs mais des sons particuliers : grosse caisse, caisse claire, cymbale...

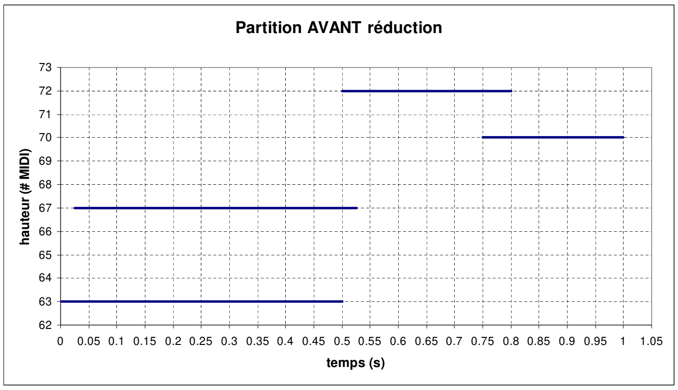
Le reste des pistes sera accessible à l'utilisateur. Ainsi, sa requête pourra viser, par exemple, la mélodie du chant, ou bien une partie piano, ou encore un solo de trompette.

Cependant, le problème de la polyphonie intrinsèque aux pistes demeure. Une solution consiste à opérer une *réduction* des notes d'une piste, afin d'en tirer une mélodie monophonique. La sous-section suivante décrit le processus de réduction appliqué dans notre système de recherche.

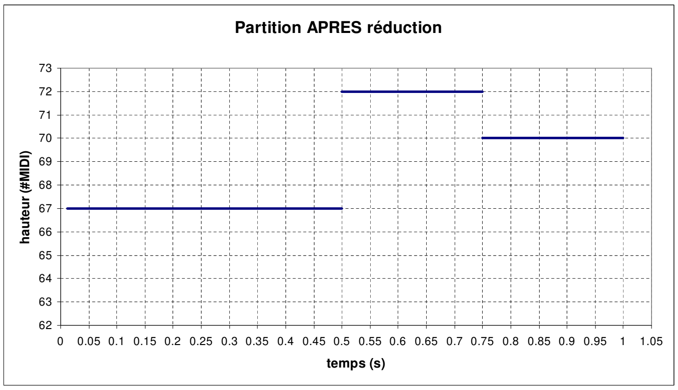
5.3.2 Réduction de la polyphonie intrinsèque à un instrument

La réduction de la polyphonie permet de ramener la partition polyphonique d'un instrument à une mélodie monophonique. La réduction appliquée dépend des instants d'apparition des notes sonnant simultanément.

Comme nous pouvons le voir sur l'exemple illustré figure 5.11, si les instants d'apparition sont assez espacés, l'arrivée d'une nouvelle note met simplement fin à la précédente. Par contre, si ces instants sont proches (i.e. notes quasi-simultanés), seule la note dont la hauteur est la plus élevée est conservée. Comme nous l'avons vu en 3.4.3, ce critère de choix a été jugé comme étant le meilleur parmi les quatre algorithmes de réduction testés dans [UZ98].



(a)



(b)

figure 5.11 : *Musique polyphonique (a) et sa représentation par une mélodie monophonique (b).*

Le jugement de la simultannéité d'apparition de deux notes est fondé sur un critère relatif au tempo. L'horizon de l'intégration temporelle appliquée s'étend sur $\frac{1}{32}$ème de temps, soit $\frac{1}{4}$ de triple croche.

Prenons l'exemple des notes illustrées figure 5.11(a). Leurs caractéristiques sont présentées tableau 5.11(a). La musique dont elles sont tirées ayant un tempo de "60*bpm* à la noire", le seuil décidant de la simultanéité d'apparition correspond à $\frac{60}{tempo} * \frac{1}{32} = 31.25$ms. Les deux premières notes de la musique considérée commencent à moins de 31.25ms d'intervalle. Elles sont donc en concurrence pour figurer dans la description. La note élue est la deuxième (hauteur 67 > 63). Les notes 2 à 4 ont des instants d'apparition assez éloignés (différence > 31.25ms) pour ne pas que l'une n'occulte l'autre. Dans ce cas, on assiste à de simples troncatures de durées.

Note	Instant d'apparition	Instant de disparition	Hauteur (# MIDI)
1	0.000	0.500	63
2	0.025	0.525	67
3	0.500	0.800	72
4	0.750	1.000	70

TAB. 5.11 : *Caractéristiques de notes composant une musique polyphonique.*

La réduction de la polyphonie des pistes sélectionnées permet de se ramener à un matériau musical qui soit comparable à la requête i.e. une mélodie monophonique. Concernant la description de ce matériau, nous allons suivre le même cheminement que pour la requête, en considérant dans un premier temps l'information temporelle, puis l'information fréquentielle.

5.3.3 Information temporelle

Concernant l'information temporelle, notre choix de description des données de la base s'aligne strictement sur celui de la requête : seul l'ordre d'apparition des notes est conservé.

5.3.4 Information fréquentielle

Pour l'aspect fréquentiel, deux cas de figure demeurent selon la description choisie pour la requête : *profil de hauteurs* ou bien *séquence d'intervalles*. Or, les hauteurs des mélodies constituant la base de données sont - par nature - quantifiées au demi-ton près. Par conséquent, si les processus de description de la requête sont appliqués aux données de la base, la seule différence viendra de la précision des valeurs. Pour les données de la base, les descripteurs seront constitués de valeurs *entières*. On aurait donc :

– *profil de hauteurs* : vecteur des hauteurs tempérées consécutives.
– *séquence d'intervalles* : vecteur des intervalles tempérés consécutifs.

Une précision ajustable

Les hauteurs de la base de données étant discrètes, il est possible d'appliquer une nouvelle quantification sans que celle-ci pose les problèmes d'ambiguïté aux frontières d'états rencontrés avec la requête. Cependant, le choix d'un pas de quantification plus grand que le demi-ton entraîne toujours une description à la tolérance inéquitable. Nous allons cependant détailler cette possibilité.

Selon que le matériau quantifié soit *intervalles* ou *hauteurs*, les conséquences sont différentes. Pour les intervalles, nous l'avons vu en 3.3.3 et en 3.3.4 qu'une perte maîtrisée de précision pouvait être désirée. L'étude des propriétés des données de la base permettent de guider le choix d'une quantification supplémentaire. La figure 5.12 montre que certains intervalles sont nettement moins représentés que d'autres. Comptant sur une probabilité de sollicitation faible, une nouvelle quantification (non uniforme) peut donc être appliquée, privilégiant la tolérance, sans trop pénaliser la discrimination.

Ainsi, pour les intervalles, le niveau de précision peut être choisi en fonction de la discrimination désirée. Pour les hauteurs, les conséquences d'une quantification supplémentaire sont moins intéressantes. En effet, puisque toutes les hauteurs sont concernées, c'est l'ensemble des intervalles les séparant -sans distinction- qui est touché. L'intervalle de base de la musique occidentale n'est donc pas préservé. Or, il constitue, après l'intervalle de ton et l'intervalle nul, l'intervalle le plus rencontré dans la base de données (cf. figure 5.12). Le fait qu'il ne puisse pas être représenté avec précision entraîne une perte de précision difficilement acceptable.

Comme nous l'avons vu plus tôt, la quantification prive la description d'une tolérance inéquitable. C'est pourquoi nous conservons le maximum de précision dans notre description fréquentielle des mélodies de la base, et ce, dans les deux cas (séquences d'intervalles, et profils de hauteurs).

Dans la sous-section suivante, nous donnons quelques caractéristiques du matériau musical finalement décrit.

5.3.5 Caractéristiques du matériau musical décrit

Grâce aux nombreux fichiers MIDI disponibles sur le réseau Internet, nous nous sommes constitué une base de données conséquente. Après élimination des redondances flagrantes[10], nous avons sélectionné 19282 fichiers (occupant 649.8Mo, stockables sur un cédérom).

Notre base de données contient de la variété française et internationale, des styles aussi divers que le jazz, le gospel, la disco, le hard-rock, mais aussi de la musique classique, des hymnes nationaux, ainsi que des musiques de film et des génériques TV.

Les caractéristiques présentées tableau 5.12 renseignent sur le volume de données effectivement accessibles. Conformément à ce que nous avons vu en 3.4.5, nous exprimons ce dernier en terme de nombre de notes, afin que d'autres bases puissent facilement s'y comparer.

	minimum	maximum	moyenne
nombre de pistes par fichier	1	40	6.7
nombre de notes par piste	4	15 906	299.4
nombre total de notes	38 679 692		

TAB. 5.12 : *Volume des données mélodiques issues des documents MIDI disponibles.*

La répartition des 40 millions d'intervalles constituant ces mélodies est présentée figure 5.12.

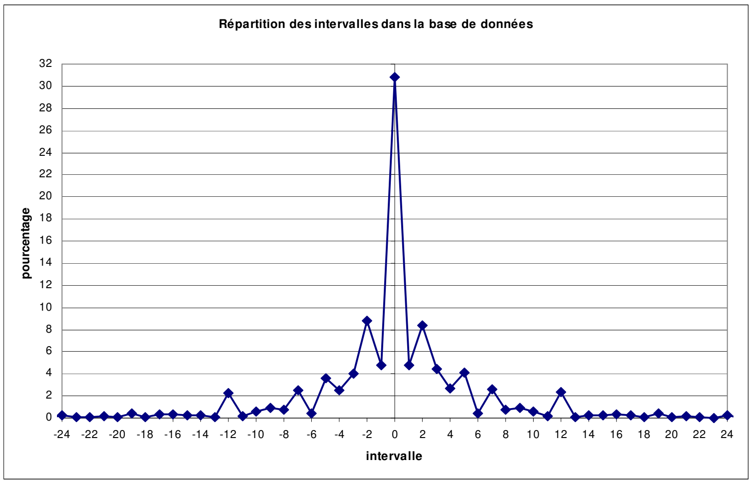


figure 5.12 : *Répartition des intervalles constituant les mélodies accessibles par une requête.*

[10]Même nom de fichier, et taille (quasi-)identique.

Comme nous pouvons le voir, l'intervalle nul est nettement mieux représenté que les autres (près d'un tiers des données). Les intervalles ± 2 se placent en deuxième position, suivis du groupe d'intervalles ± 1, ± 3, ± 5, puis du groupe ± 4, ± 7, ± 12. Aucun des intervalles restants ne dépasse 1% de la population.

Cette hiérarchie correspond (d'une manière moins nette pour les derniers groupes) à celle des intervalles visés par les mélodies chantés étudiées au chapitre 4 (cf. figure 4.7). Les données récoltées étaient donc représentatives des intervalles les plus rencontrés dans notre base de données.

5.3.6 Conclusion

La description des 19.282 fichiers MIDI constituant notre base de données musicale passe donc par une description séparée de leurs pistes. Ainsi, différentes requêtes pourront viser des mélodies jouées par les différents instruments d'une musique. Lorsqu'un instrument suit une partition polyphonique, une réduction de l'information est appliquée afin de se ramener à une mélodie (monophonique) qui permet la description. L'information mélodique accessible représente près de 40 millions de notes.

A l'instar de la description adoptée pour la requête, aucune quantification n'est appliquée. La précision fréquentielle disponible est donc préservée. Pour les données MIDI, tempérées de nature, la représentation est assurée par des valeurs entières. Conformément au choix laissé ouvert pour la requête, nous proposons deux types de descripteurs pour les données de la base : *profils de hauteurs* et *séquences d'intervalles*.

Nous en rappelons rapidement les caractéristiques : ces descripteurs assurent l'indépendance au ton ainsi qu'au rythme des mélodies. Les séquences d'intervalles sont indépendantes du ton des mélodies. Cela n'est pas le cas des profils de hauteurs qui délèguent cette invariance à la mesure de similarité. Celle-ci doit également gérer la sensibilité aux erreurs, et ce, pour les deux types de descripteurs proposés.

5.4 Mesure de similarité

Le refus de la quantification des requêtes entraîne l'impossibilité d'une représentation des mélodies par une succession d'états (en nombre fini). Les techniques de *string matching*, qui jouent habituellement le rôle de mesure de similarité, ne peuvent être utilisées ici. Notre mesure de similarité consistera en une distance entre le vecteur descripteur de la requête et celui d'une portion de la base de données.

5.4.1 Distances

Ce travail de thèse parcourant un vaste domaine, nous n'avons pu envisager un grand nombre de distances. Nous nous sommes limités à envisager les distances L1 et L2, soit

$$d_{L1}(X,Y) = \sum_{j=0}^{N-1} \left|X_j - Y_j\right| \quad et \quad d_{L2}(X,Y) = \sum_{j=0}^{N-1} (X_j - Y_j)^2$$

avec $X = [X_0, \ldots, X_{N-1}]$ et $Y = [Y_0, \ldots, Y_{N-1}]$.

Note : Afin de simplifier les calculs ultérieurs, $d_{L2}(X,Y)$ ne comporte pas de racine carrée. Les distances utilisées étant destinées à établir un classement, cette absence n'a pas d'incidence.

Pondération

L'étude des mélodies chantonnées du Chapitre 4 nous a montré que l'imprécision présente au début des requêtes n'était pas suffisante pour diminuer systématiquement l'influence de certaines notes dans la comparaison mélodique (cf. 4.3.6).

Cependant, si cette imprécision s'avérait spécifique à un type d'utilisateurs dits *débutants*, nous pourrions envisager un système proposant deux profils. Le profil *expert* accorderait une confiance maximum à l'ensemble des notes de la requête, et le profil *débutant* entraînerait une pondération diminuant l'influence du premier intervalle chanté.

Le système que nous concevons dans cette thèse correspondrait donc au profil *expert*.

Normalisation

La distance utilisée sera divisée par la taille N des mélodies mises en jeu afin de se ramener à une mesure de la dissimilarité moyenne par note. Ainsi, notre système de recherche de musique par chantonnement autorisera une comparaison des dissimilarités mesurées sur des requêtes de tailles différentes. Nous utiliserons donc

$$d'_{L1}(X,Y) = \frac{1}{N} \sum_{j=0}^{N-1} \left|X_j - Y_j\right| \quad et \quad d'_{L2}(X,Y) = \frac{1}{N} \sum_{j=0}^{N-1} (X_j - Y_j)^2$$

Nous allons maintenant voir la mise en œuvre de ces deux distances sur nos deux types de descripteurs, *profil de hauteurs* et *séquence d'intervalles*. Nous verrons notamment comment l'invariance au ton est assurée pour le premier d'entre eux.

5.4.2 Cas des profils de hauteurs

Soit $\bar{r} = [r_0, \ldots, r_{N-1}]$ le profil de hauteurs de la requête et soit $\bar{d} = [d_0, \ldots, d_{N-1}]$ celui d'une portion mélodique de la base de données. Les deux mélodies considérées comportent chacune N notes.

Nous savons que cette description des mélodies est dépendante du ton. Dans l'impossibilité d'extraire de manière satisfaisante le ton d'une portion mélodique, nous remplaçons cette notion *musicale* par une valeur issue d'un critère purement *mathématique*. Un changement de ton correspondant à une translation des valeurs sur l'échelle des hauteurs, les deux vecteurs à comparer seront ajustés par un offset λ minimisant la distance mesurée.

La mesure de similarité possède donc l'expression suivante :

$$D_{h\gamma}(\bar{r}, \bar{d}) = \frac{1}{N} \sum_{j=0}^{N-1} |r_j - d_j - \lambda|^{\gamma} \tag{5.1}$$

avec $\gamma = \{1, 2\}$.

Etudions maintenant l'influence de chacune des deux distances envisagées.

Distance L1

Avec une distance L1, l'offset λ_{h1} minimisant la distance est la valeur médiane[11] de la liste constitué des composantes du vecteur $(\bar{r} - \bar{d})$ [AN89] :

$$\lambda_{h1} = median(\bar{r} - \bar{d}) \tag{5.2}$$

Exprimons maintenant les hauteurs de la requête en fonction de celles de la mélodie visée. Si t représente la différence de ton (supposée constante) entre requête et mélodie visée, et b_j l'imprécision ponctuelle sur la hauteur d_j, on a :

$$r_j = d_j + t + b_j \tag{5.3}$$

L'expression 5.1 nous donnant

$$D_{h1}(\bar{r}, \bar{d}) = \frac{1}{N} \sum_{j=0}^{N-1} |t + b_j - \lambda| \tag{5.4}$$

et les équations 5.2 et 5.3

$$\lambda_{h1} = median(t + \bar{b}) = t + median(\bar{b}) \tag{5.5}$$

[11] La valeur médiane d'un ensemble de valeurs est obtenue en ordonnant ces dernières et en choisissant la valeur située au milieu de la liste ordonnée. Si la population est paire, la valeur retenue est la moyenne des deux valeurs centrales.

On peut tirer l'expression de la distance témoignant de la dissimilarité de deux mélodies, représentées par leur profil de hauteurs, et comparées par une distance L1 normalisée :

$$\boxed{D_{h1}(\bar{r},\bar{d}) = \frac{1}{N}\sum_{j=0}^{N-1}|b_j - median(\bar{b})|} \tag{5.6}$$

Distance L2

Une démarche similaire peut être effectuée dans le cas de l'utilisation d'une distance L2. Dans ce cas, l'offset λ_{h2} minimisant la distance est

$$\lambda_{h2} = moyenne(\bar{r} - \bar{d}) \tag{5.7}$$

De même que pour le médian, en utilisant la décomposition de r_j vue en 5.3, on a

$$\lambda_{h2} = moyenne(t + \bar{b}) = t + moyenne(\bar{b}) \tag{5.8}$$

A partir des équations 5.1 et 5.8, on obtient l'expression de la distance témoignant de la dissimilarité de deux mélodies, représentées par leur profil de hauteurs, et comparées par une distance L2 normalisée :

$$\boxed{D_{h2}(\bar{r},\bar{d}) = \frac{1}{N}\sum_{j=0}^{N-1}\left[b_j - moyenne(\bar{b})\right]^2} \tag{5.9}$$

Interprétation

Sous l'hypothèse d'un ton stable sur chacune des deux requêtes ($t = cst$), ces moteurs de comparaison (profils de hauteurs associés à L1 ou L2) permettent de s'affranchir de la différence de ton au prix d'une modification des erreurs observées. Selon la distance choisie, la mesure de similarité témoigne des erreurs sur les hauteurs $\bar{b} = [b_0, \ldots, b_{N-1}]$, mais recentrées, soit par leur médian (D_{h1}), soit par leur moyenne (D_{h2}).

L'étude des requêtes chantonnées du Chapitre 4 portait sur les intervalles et non sur les hauteurs. Les observations effectuées ne nous permettent pas de savoir si les erreurs sont centrées (i.e. si leur moyenne est nulle). Quant bien même nous le saurions, nous ne pourrions affirmer que la solution 5.9 revient à observer les erreurs elles-mêmes, compte-tenu du faible nombre de réalisations qu'une requête contient.

Les éléments dont nous disposons ne nous permettent pas, pour le moment, de nous prononcer en faveur de l'une ou l'autre des solutions (L1 ou L2).

5.4.3 Cas des séquences d'intervalles

Avec toujours $\bar{r} = [r_0, \ldots, r_{N-1}]$ la séquence des hauteurs ordonnées de la requête et $\bar{d} = [d_0, \ldots, d_{N-1}]$ celle d'une portion mélodique de la base de données, la distance de dissimilarité associée aux séquences d'intervalles s'exprime de la manière suivante[12] :

$$D_{i\gamma}(\bar{r}, \bar{d}) = \frac{1}{N} \sum_{j=0}^{N-2} \left| (r_{j+1} - r_j) - (d_{j+1} + d_j) \right|^{\gamma} \tag{5.10}$$

avec $\gamma = \{1, 2\}$.

La mesure de similarité témoigne des écarts existants entre les intervalles comparés deux à deux.

Avec la décomposition de r_j vue en 5.3, l'équation 5.10 devient

$$\boxed{D_{i\gamma}(\bar{r}, \bar{d}) = \frac{1}{N} \sum_{j=0}^{N-2} \left| b_{j+1} - b_j \right|^{\gamma}} \tag{5.11}$$

avec $\gamma = \{1, 2\}$.

Sous l'hypothèse d'un ton constant, et quelle que soit la distance utilisée ($\gamma = 1$ ou 2), ce choix de moteur de comparaison permet donc de s'affranchir du ton, et témoigne des erreurs sur intervalles.

5.4.4 Conclusion

Avec la définition de mesures de similarité, nos différents moteurs de comparaison sont maintenant au complet. La nature non tempérée des descriptions de la requête oriente la mesure de similarité vers une mesure de distance. Deux types sont envisagés (normes L1 et L2). Pour les descripteurs *profil de hauteurs*, la mesure de similarité assure l'indépendance au ton (s'il est stable) par le biais d'un ajustement minimisant la distance mesurée. La mesure de similarité témoigne des erreurs de hauteurs recentrée (par leur médian ou par leur moyenne selon la distance utilisée). Les descripteurs *séquence d'intervalles* présentent déjà une indépendance au ton (s'il est stable également). La mesure de similarité témoigne des erreurs d'intervalles.

[12] Rappel : toutes nos distances sont pondérées par le nombre de notes N constituant les portions comparées, même si, comme c'est le cas ici, la somme porte un nombre d'éléments différent (taille des descripteurs = $N - 1$).

5.5 Conclusion

Nous venons de suivre pas à pas la conception de quatre moteurs de comparaison, chacun étant constitué de descripteurs mélodiques et d'une mesure de similarité. A chaque étape, nous avons envisagé les différentes voies possibles en nous prononçant pour celles offrant la convivialité nécessaire à un système de recherche grand public. Ainsi, nos moteurs de comparaison sont adaptés à la requête chantonnée, sans qu'aucune information complémentaire à la mélodie soumise ne soit requise.

La similarité mesurée est invariante aux tons des mélodies, si ceux-ci sont stables. Le résultat de la comparaison de deux mélodies est également indépendant des rythmes adoptés (i.e. tempos, même instables, et durées des notes). La tolérance assurée au niveau de l'information temporelle des mélodies est donc importante. A l'instar de la majorité des moteurs de comparaison présentés au Chapitre 2, les imprécisions fréquentielles de la requête ne sont pas jugées en fonction du contexte musical. Il s'agit de la conséquence négative de la simplicité d'utilisation offerte. Par contre, nos moteurs de comparaison se démarquent de l'état de l'art en assurant une sensibilité équitable vis-à-vis des transpositions.

Cette qualité de jugement est rendue possible grâce au refus de la quantification de l'information fréquentielle des mélodies. Ce choix est synonyme d'une précision maximum de représentation, ce qui confère aux moteurs proposés leur capacité de discrimination. L'absence de quantification de l'information fréquentielle s'accompagne également de l'impossibilité d'utiliser les techniques de *string-matching* pour la mesure de similarité. Celle-ci repose sur une mesure de distance.

Nos moteurs de comparaison n'assurent pas la gestion des insertion/omission de notes. Cependant, si l'importance du phénomène le justifiait, cette gestion supplémentaire pourrait être assurée, soit par une meilleure exploitation de l'information temporelle (extraction automatique d'une référence), ou encore par un découpage du descripteur de la requête afin de procéder à un appariement par morceaux (cf. 3.4.4).

Les quatre moteurs proposés appartiennent à deux familles distinctes. La première, fondée sur les *profils de hauteurs*, favorise une approche globale de l'information fréquentielle. En effet, l'ajustement permettant l'affranchissement aux tons empruntés dépend de l'ensemble des éléments des deux descripteurs. La seconde famille de moteur correspond à une approche plus locale de la comparaison fréquentielle. Les éléments des descripteurs fondés sur les *séquences d'intervalles* sont comparés deux à deux, indépendamment des autres éléments des descripteurs en jeu.

Par ailleurs, les types de descripteurs contiennent des éléments dont la nature est différente. En effet, les intervalles (assurant l'indépendance au ton) témoignent non pas d'une note, mais d'une relation entre notes contiguës. Ainsi, une Erreur Locale entachera deux éléments du descripteur *séquence d'intervalles* alors que, conformément à la nature de l'erreur, le descripteur *profil de hauteurs* ne verra qu'un seul élément modifié. Inversement, une Rupture de Ton entachera deux éléments du descripteur *profil de hauteurs* alors que le descripteur *séquence d'intervalles* ne verra qu'un seul de ses éléments modifié.

Ces considérations ne suffisent pas à designer le meilleur des moteurs de comparaison proposés. En fait, seule leur conformité avec des caractéristiques *perçues* le pourrait. Dans le chapitre suivant, nous allons donc tenter d'établir une correspondance entre paramètres physiques et perception. Ainsi, nous disposerons d'une mesure objective de la qualité qui, nous l'espérons, nous permettra de juger et de départager les moteurs de comparaison proposés.

Chapitre 6

Mesure Objective de la Qualité de Moteurs de Comparaison

Nous désirons mesurer objectivement la qualité de nos moteurs de comparaison. Ne disposant pas de critère adapté à cette tâche, nous allons tenter d'en définir. Dans ce chapitre, nous procèderons à des tests subjectifs afin de disposer d'éléments concernant la similarité mélodique. Nous utiliserons ensuite leurs résultats pour voir comment les moteurs de comparaison proposés témoignent de cette similarité.

6.1 Introduction

Une des principales difficultés de la conception de moteurs de comparaison est le manque de connaissances sur la similarité mélodique qui soient réellement applicables dans un contexte d'utilisation grand public[1]. En particulier, l'ignorance du contexte tonal oblige à abandonner certains critères musicaux pour ne fonder les mesures de similarité que sur les différences objectives entre mélodies, quelle que soit la perception de ces différences (cf. p. 102). Cette limitation ampute sensiblement le domaine de la similarité mélodique concerné par les moteurs de comparaison actuellement développés.

Dans le domaine restant, nous aimerions disposer de mesures objectives afin de juger de la qualité des moteurs proposés. La définition de critères objectifs témoignant du jugement humain sur la similarité mélodique nécessite des tests subjectifs.

Le but général des tests que nous allons effectuer est la révélation de certaines limites de la similarité mélodique. Plus précisément, il s'agit d'obtenir des règles représentatives de la perception, comme des correspondances entre différents types de perturbation. Par exemple, une des règles attendues pourrait être : "Tel type de dégradation d'une amplitude a et tel autre d'une amplitude b entraînent le même degré de perturbation".

Nous allons tout d'abord décrire les tests subjectifs réalisés (principe, mise en place, résultats), puis nous appliquerons les critères obtenus aux moteurs de comparaison proposés, pour tenter de les départager.

[1] Les seuls critères utilisés sont l'invariance du ton et l'invariance du tempo.

6.2 Tests subjectifs : Principe

L'expérience que nous allons mettre en œuvre consiste à faire entendre à des sujets des mélodies perturbées. Partant du maximum de dégradation, nous diminuons l'amplitude jusqu'à diffuser la mélodie originale (i.e. non perturbée). Ainsi, nous pourrons déterminer le niveau de perturbation auquel la mélodie originale est identifiée.

Il s'agit donc de tests sur la *reconnaissance* de mélodies dégradées. Compte tenu des moteurs de comparaison proposés, les dégradations appliquées entacheront uniquement la hauteur des notes.

A ce stade, deux points doivent être précisés :

1. La *similarité* que nous cherchons à cerner ici, désigne ce qui est *assez similaire pour être reconnu*. Il est légitime de s'appuyer sur cette notion puisque les systèmes de recherche sur lesquels nous travaillons sont, en fait, destinés à reproduire la capacité humaine à *reconnaître* un air donné.

2. Compte-tenu de la nature des moteurs de comparaison développés, cette expérience repose sur l'hypothèse que *les erreurs injectées sont d'autant plus perturbantes que leur amplitude est élevée*. Or, nous l'avons déjà vu, ce genre d'hypothèse n'est pas forcément respecté du point de vue de la perception.
   Par exemple, diminuer la hauteur d'une note d'un demi-ton peut être plus perturbant que la diminuer d'un ton entier. En effet, cela dépend du contexte tonal dans lequel se place la note perturbée. Si cette note correspond à une *quinte* dans la tonalité empruntée, la descendre d'un demi-ton la transformera en *quinte diminuée* (généralement perturbante), alors que la diminuer d'un ton la transformera en *quarte* (généralement mieux acceptée).

   Cette remarque influence notre démarche en deux points :

   (a) Nous nous limitons à des perturbations simples, d'amplitudes relativement faibles. Ainsi, nous conservons une marge vis-à-vis d'amplitudes importantes mais peu perturbantes (l'octave en particulier).
   (b) Par ailleurs, nous évitons de placer les hauteurs modifiées sur l'échelle tempérée. Cette démarche a pour but de contourner le problème du contexte tonal. Les perturbations ne seront ainsi ni amoindries par leur appartenance à la tonalité de la mélodie, ni aggravées par leur exclusion.

6.3 Tests subjectifs : Modalités

Dans cette section, nous allons voir les étapes de la mise en œuvre de ces tests. Nous commencerons par aborder la sélection des mélodies originales, puis nous détaillerons les trois types de dégradations appliquées ; enfin, nous présenterons brièvement les sujets testés.

6.3.1 Sélection des mélodies

Afin que les mélodies puissent être reconnues par le plus grand nombre de sujets, elles doivent être les plus populaires possibles. Certaines mélodies sont reconnaissables avec très peu de notes, alors que d'autres en nécessitent plus. Nous avons donc sélectionné des portions de taille variable (de 6 à 14 notes), mais contenant pour la plupart des phrases musicales *complètes* (c'est-à-dire, ne provocant pas -à l'écoute- de sensation de troncature brutale). Les titres et les tailles des mélodies à identifier sont présentés tableau 6.1.

Titre	Taille (notes)	Rang et signe (EL et RT)	Signe (GT)
C'est à babord (trad.)	10	4+	+
La javanaise (S.Gainsbourg)	13	7-	-
Dès que le vent soufflera (Renaud)	12	7+	+
Noir, c'est noir (J.Hallyday)	9	3-	+
Carmen (Bizet)	8	3-	+
Il était un petit navire (trad.)	9	4+	+
La salsa du démon (G.Orch.du Splendid)	14	6+	-
Pennylane (Beatles)	13	7+	-
San Francisco (M.Leforestier)	6	4+	+
Cadet roussel (trad.)	8	7-	+
Des chiffres et des lettres (TV)	10	4-	+
Prendre un enfant (Y.Duteil)	7	5+	-
La vie en rose (E.Piaf)	7	5+	+
La mer (C.Trenêt)	12	6-	-
Amsterdam (J.Brel)	6	5-	+
Les démons de minuit (Images)	8	3-	-
Les animaux du monde (TV)	7	5-	+
Santiano (H.Aufray)	8	4-	-

TAB. 6.1 : *Motifs mélodiques sélectionnés pour les tests subjectifs de reconnaissance de mélodies dégradées. Les noms sont accompagnés du nombre de notes sélectionnées, ainsi que d'informations sur les dégradations appliquées : emplacement et signe des EL/RT, et signe des GT.*

6.3.2 Dégradations appliquées

Chacune des 18 portions mélodiques sélectionnées est entachée d'une dégradation dont l'amplitude sera diminuée à chaque nouvelle présentation. Nous allons voir les types de perturbations injectées ainsi que les amplitudes empruntées.

Types sélectionnés

Nous utilisons les trois types de dégradation rapportés par McNab et Lindsay, et confirmés par nos observations (cf. Chapitres 3 et 4). Ils constituent des perturbations simples qui pourront facilement être injectées dans les mélodies soumises aux sujets, ainsi que dans l'expression analytique des mesures de similarité des moteurs de comparaison à juger. Ces types d'erreur sont les suivants :

1. Erreur Locale - EL : seule une note voit sa hauteur entâchée d'erreur. Avec le formalisme utilisé dans le Chapitre 5 (cf. expression 5.3, p. 118), la mélodie présentée correspond à

$$\bar{r} = \bar{d} + \bar{b}_{EL(a,\alpha)} \tag{6.1}$$

$\bar{d}$ étant le vecteur descripteur du motif mélodique à reconnaître, et $\bar{b}_{EL(a,\alpha)} = [0,\ldots,0,a,0,\ldots,0]$, l'EL (de valeur a) injectée. $\alpha \in [1\,;N]$ désigne le rang auquel l'erreur apparaît (N = nombre de notes du motif mélodique) ;

2. Rupture de Ton - RT : la hauteur globale de la mélodie est modifiée soudainement et définitivement. La mélodie présentée correspond donc à

$$\bar{r} = \bar{d} + \bar{b}_{RT(a,\alpha)} \tag{6.2}$$

avec $\bar{b}_{RT(a,\alpha)} = [0, \ldots, 0, a, \ldots, a]$, la RT (de valeur a) injectée. Comme précédemment, $\alpha \in [1\,;N]$ désigne le rang d'apparition de l'erreur ;

3. Glissement de Ton - GT : la hauteur globale de la mélodie dérive régulièrement à chaque nouvelle note. La mélodie présentée correspond donc à

$$\bar{r} = \bar{d} + \bar{b}_{GT(a)} \tag{6.3}$$

avec $\bar{b}_{GT(a)} = [0, a, 2a, \ldots, (N-1)a]$, le GT (de valeur a) injecté (Contrairement aux deux cas précédents, a s'exprime en demi-ton/note et non en demi-ton.).

Amplitudes

Le tableau 6.2 présente les amplitudes des paramètres a utilisées.

a pour EL & RT (en demi-ton)	4.5	3.5	2.5	1.5	0.5	0.0
a pour GT (en demi-ton/note)	1.0	0.8	0.6	0.4	0.2	0.0

TAB. 6.2 : *Amplitudes du paramètre a pour les trois types d'erreur utilisés.*

Les airs présentés, la taille des mélodies les représentant, et les amplitudes d'erreur ont été choisies grâce à des tests préliminaires. Les stimuli exploitables sont les mélodies qui ne sont pas reconnues lorsqu'elles sont entachées des erreurs d'amplitude maximum, mais dont l'original est reconnu.

Ainsi, l'amplitude de dégradation à laquelle les mélodies sont reconnues est inférieure à l'amplitude maximum, les données recueillies peuvent être exploitées pour définir les critères objectifs recherchés.

Les autres paramètres d'erreur (emplacement et signe pour EL et RT, signe pour GT) ont été choisis "à l'écoute", avec le même but. Ils sont présentés dans les deux dernières colonnes du tableau 6.1.

Le but de l'expérience étant de comparer différents types de perturbations, les erreurs EL et RT sont, pour une mélodie donnée, systématiquement injectées au *même emplacement*, et avec un *signe identique*. Nous éliminons ainsi un degré de liberté éventuellement perturbant (une erreur donnée peut avoir une influence différente selon le contexte mélodique).

Maintenant que le matériau de test est défini, nous allons voir la manière de le présenter aux sujets.

6.3.3 Mode de présentation des mélodies

Idéalement, nous aimerions que chaque sujet puisse nous renseigner sur les correspondances qu'il perçoit entre deux types de perturbations successivement appliquées sur une même mélodie. Ainsi, nous pourrions en tirer des équivalence entre types de dégradation à contexte mélodique constant. Malheureusement cela paraît difficile puisqu'une fois qu'une mélodie est identifiée, celle-ci ne peut à nouveau être présentée même avec un autre type d'erreur, car elle serait plus facilement reconnaissable.

Des profils différents

Chaque mélodie présentée n'est donc associée qu'à un seul type d'erreur. Cependant, les appariements ne sont pas identiques pour tous les sujets, nous utilisons des profils différents.

Pour un profil de sujet donné, les mélodies dégradées peuvent être séparées en deux groupes. Une moitié est commune à tous les profils (mélodies assignées aux mêmes perturbations), pour l'autre moitié, les associations entre mélodies et perturbations sont spécifiques au profil.

Ainsi, la première moitié permettra d'observer les réactions de l'ensemble des sujets face aux mêmes stimuli. Si cette base révèle une importante cohérence de la part de sujets de profils différents, les résultats de ceux-ci seront exploités afin comparer différents type de dégradation sur une même mélodie (grâce à la deuxième moitié des mélodies).

Par exemple, s'il y a 6 mélodies en tout [m1...m6], m1, m2, et m3 seront respectivement associées au types t1, t2 et t3, et ce, pour tous les sujets. Pour les autres mélodies, soit m4, m5, et m6, les types associés varieront. On aura par exemple :

- Sujet 1 : m1-t1 ; m2-t2 ; m3-t3 ; m4-t1 ; m5-t2 ; m6-t3
- Sujet 2 : m1-t1 ; m2-t2 ; m3-t3 ; *m4-t2* ; *m5-t3* ; *m6-t1*
- Sujet 3 : m1-t1 ; m2-t2 ; m3-t3 ; *m4-t3* ; *m5-t1* ; *m6-t2*

Si, sur les mélodies m1 à m3, les sujets 2 et 3 révélaient un comportement identique, nous pourrions comparer les dégradations appliquées sur les mélodies m4 à m6 :

- m4-t2&t3
- m5-t3&t1
- m6-t1&t2

Eviter l'anticipation des mélodies à venir

Une mélodie originale subit donc différents degrés de perturbation. A la diffusion, il n'est pas souhaitable de faire écouter à la suite des versions de moins en moins perturbées d'une même mélodie. En effet, cela peut permettre au sujet d'anticiper sur l'information à venir. Les motifs mélodiques (présentés tableau 6.1) seront donc mélangés.

Dans notre exemple, si les types d'erreurs t1, t2, t3 étaient utilisés avec les amplitudes $a3 > a2 > a1 > 0$, le Sujet1 se verrait présenter la séquence de stimuli du tableau 6.3 :

Ordre	1	2	3	4	5	6
Mélodie	m1-t1(a3)	m2-t2(a3)	m3-t3(a3)	m4-t1(a3)	m5-t2(a3)	m6-t3(a3)
Ordre	7	8	9	10	11	12
Mélodie	m1-t1(a2)	m2-t2(a2)	m3-t3(a2)	m4-t1(a2)	m5-t2(a2)	m6-t3(a2)
Ordre	13	14	15	16	17	18
Mélodie	m1-t1(a1)	m2-t2(a1)	m3-t3(a1)	m4-t1(a1)	m5-t2(a1)	m6-t3(a1)
Ordre	19	20	21	22	23	24
Mélodie	m1	m2	m2	m4	m5	m6

TAB. 6.3 : *Exemple de séquence de présentation de mélodies.*

Lorsqu'une mélodie est reconnue, ses autres versions sont retirées de la liste des mélodies à venir, donc non présentées au sujet.

6.3.4 Sujets

Huit sujets, âgés de 23 à 32 ans ont participé aux tests de reconnaissance. Leur écoute musicale hebdomadaire est importante (moyenne de 16h, min = 3h, max = 35h), et cinq d'entre eux ont l'expérience d'une pratique musicale (moyenne de 6 ans, min = 4ans, max = 10 ans).

Nous n'avons pas délibérément rassemblé une telle population. Cependant, nous espérons que la relative cohérence des âges, l'écoute et la pratique soutenues de la musique nous assureront un faible taux de musiques non reconnues. Nous redoutons cependant qu'un fort taux de musiques soient reconnues malgré la pire perturbation.

6.4 Tests subjectifs : Résultats

Dans cette section, nous allons établir les critères objectifs témoignant de la similarité mélodique à partir des résultats de nos tests. Nous verrons que ces derniers ne sont pas aussi précis que ce que nous espérions.

6.4.1 Données exploitables

Sur les 8*18 = 144 données recueillies, nous avons déploré un fort taux de déchet. En effet, seules 47% des données initiales sont exploitables. Cela illustre la difficulté de disposer de mélodies qui ne soient pas identifiées à leur première diffusion (i.e. entachés de la pire dégradation) mais reconnues par la suite. Leur répartition est présentée tableau 6.4.

Type d'erreur	Reconnues avec amplitude max		Non reconnues		Données exploitables	
EL	16	33%	10	20%	22	46%
RT	7	15%	26	54%	15	31%
GT	4	8%	13	27%	31	65%
Tous	27	19%	49	34%	68	**47%**

TAB. 6.4 : *Résultats définissant la proportion des données exploitables.*

Nous n'avons pas observé de cohérence assez importante entre sujets de profils différents pour effectuer des regroupements. Notre tactique de répartition des sujets par profils n'a pu être menée à son terme, nous empêchant de comparer les versions dégradées d'une *même* mélodie.

Par conséquent, nos tests vont simplement nous permettre de nous faire une idée des correspondances existantes entre les types de dégradations représentés.

6.4.2 Définition de critères objectifs de qualification

Le tableau 6.5 présente la répartition des données exploitables pour chacun des types d'erreur (indépendamment des mélodies entachées). On y trouve le nombre de stimuli reconnus pour chaque amplitude de perturbation considérée (deuxième ligne de chaque bloc).

EL	Amplitude à la reconnaissance	4.5	3.5	2.5	1.5	0.5	0.0
	Nombre de stimuli	0	2	7	1	1	4
	Pourcentage de reconnaissance	0	13	47	7	7	27
	Pourcentage cumulé	0	13	60	67	73	100
	Amplitude moyenne	1.8					
RT	Amplitude à la reconnaissance	4.5	3.5	2.5	1.5	0.5	0.0
	Nombre de stimuli	0	8	2	4	2	6
	Pourcentage de reconnaissance	0	36	9	18	9	27
	Pourcentage cumulé	0	36	45	64	73	100
	Amplitude moyenne	1.8					
GT	Amplitude d'incrément à la reco.	1.0	0.8	0.6	0.4	0.2	0.0
	Nombre de stimuli	0	4	4	6	10	7
	Pourcentage de reconnaissance	0	13	13	19	32	23
	Pourcentage cumulé	0	13	26	45	77	100
	Amplitude moyenne	0.3					

TAB. 6.5 : *Valeurs des amplitudes pour les trois types d'erreur utilisés. Pour GT, dégradation non locale, il s'agit de l'accentuation de l'erreur à chaque note, exprimée en demi-ton/note. Pour El et RT, l'amplitude d'erreur est exprimée en demi-ton.*

Ces résultats nous permettent, dans un premier temps, de déterminer le taux de reconnaissance en fonction de la dégradation, soit le pourcentage de la population ayant reconnu la mélodie entachée d'une dégradation d'amplitude supérieure ou égale à une amplitude donnée. Ces pourcentages (cumulés) sont présentés dans l'avant dernière ligne de chaque bloc.

La figure 6.1 l'illustre pour chacun des types d'erreur. Nous rappelons que le paramètre a ne correspond pas au même type d'amplitude pour un GT (unité : *demi-ton/note*) que pour une EL ou RT (unité : *demi-ton*). L'axe des abscisses représente donc une double échelle rendant artificiel le placement de la courbe "GT" par rapport aux deux courbes "EL" et "RT".

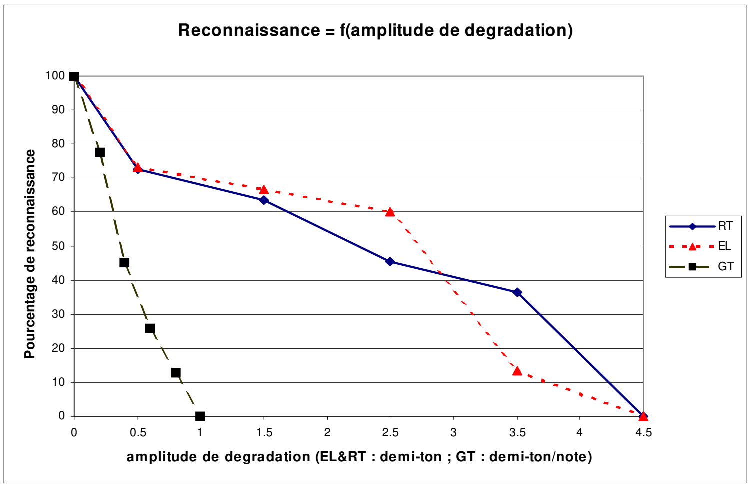


figure 6.1 : *Taux de reconnaissance pour les trois types de dégradation en fonction de l'amplitude de perturbation.*

Nous voyons que les dégradations EL et RT ont des profils relativement similaires, nous pouvons donc en conclure qu'un moteur de comparaison ne devrait pas traiter trop différemment les dissimilarités entraînées par ces deux types d'erreurs.

Cette observation constitue notre *premier critère objectif* de jugement de la qualité des moteurs de comparaison.

Pour comparer les GT aux EL/RT, nous allons considérer l'amplitude moyenne à laquelle l'ensemble de la population a reconnu les mélodies dégradées. Celle-ci est présentée pour chaque type d'erreur tableau 6.5 (dernière ligne de chaque bloc).

Conformément aux premier critère établi, cette moyenne est identique (= 1.8) pour les EL et les RT. Pour les GT, elle vaut 0.3. Nous pourrons donc établir des correspondances entre GT d'amplitude 0.3 demi-ton/note et EL et RT d'amplitude 1.8 demi-tons.

Cette *double correspondance* vient s'ajouter au critère précédemment établi. Nous disposerons donc de *trois critères objectifs* pour estimer la qualité des moteurs de comparaison proposés.

6.4.3 Conclusion

En faisant identifier des musiques à partir d'extraits mélodiques perturbés, nous avons défini des liens entre paramètres physiques et dissimilarité mélodique. En effet, nous avons révélé qu'un moteur de comparaison devait traiter équitablement les EL et les RT d'amplitudes identiques. Nous avons également établi des correspondances entre ces deux types d'erreur et les GT, portant ainsi à trois le nombre de critères objectifs définis.

La principale difficulté de ce genre d'expérience consiste à créer des dissimilarités maîtrisées alors que le but est précisément d'observer l'impact de perturbation sur la similarité. Notre protocole d'expérimentation est original, cependant la difficulté de mise en œuvre limite nos conclusions. En effet, il est difficile de recueillir des mélodies exploitables avec succès, notamment à cause des différences de culture musicale entre sujets (34% de mélodies non reconnues). Les critères objectifs dégagés fournissent néanmoins un ordre d'idée sur le comportement à attendre d'un moteur de comparaison confronté à des dégradations mélodiques simples.

Dans la section suivante, nous allons soumettre nos quatre moteurs de comparaison aux trois critères objectifs dégagés par nos tests. Nous aurons ainsi décrit l'ensemble de notre démarche : de la conception et la mise en œuvre des tests, jusqu'à l'application des critères tirés des résultats.

6.5 Application des critères objectifs aux moteurs de comparaison proposés

Dans cette section, nous allons juger les différents moteurs de comparaison en fonction des trois critères de correspondance tirés de nos tests subjectifs sur la similarité mélodique.

6.5.1 Principe

Pour appliquer nos critères de correspondance, nous nous appuierons sur l'expression analytique des dissimilarités provoquées par l'injection de *chacune* des trois perturbations considérées (EL, RT, et GT), et ce, pour les quatre moteurs de comparaison proposés. Le calcul de ces expressions est disponible en Annexe D.

Par exemple, le moteur de comparaison $h2$ témoigne d'une RT (de paramètres a et α) par la mesure de similarité suivante :

$$D_{h2}(\bar{r} = \bar{d} + \bar{b}_{RT(a,\alpha)}\ ,\ \bar{d}) = \frac{1}{N}\sum_{j=0}^{N-1}(b_{RT(a,\alpha)\ j}\ - moyenne(\bar{b}_{RT(a,\alpha)}))^2 = \frac{(\alpha-1)(N-\alpha+1)a^2}{N^2}$$

Nos critères de jugement consistants en une correspondance entre erreurs, nous estimerons la qualité de nos moteur en observant le comportement du *rapport* des dissimilarités.

Par exemple, pour une correspondance entre une EL (de paramètres a_1 et α_1), et une RT (de paramètres a_2 et α_2), nous observerons le rapport :

$$\frac{D_{h2}(\bar{r} = \bar{d} + \bar{b}_{EL(a_1,\alpha_1)}\ ,\ \bar{d})}{D_{h2}(\bar{r} = \bar{d} + \bar{b}_{RT(a_2,\alpha_2)}\ ,\ \bar{d})}$$

Si ce rapport est égal à 1, nous en conclurons que la correspondance entre EL et RT est parfaitement assurée par le moteur considéré ($h2$ dans notre exemple).

Deux tendances opposées pourrons être comparées grâce à l'expression inverse de l'une des deux. Ainsi, un moteur dont le rapport est égal à 1.33 (favorisant les RT par rapport aux EL), sera jugé aussi équitable qu'un moteur ayant un rapport égal à 0.75 (puisque $\frac{1}{0.75} = 1.33$).

Dans cette section, nous présentons directement les rapports des dissimilarités permettant d'estimer la conformité des moteurs aux critères définis. Les calculs y menant sont disponibles en Annexe D.

Les résultats des tests subjectifs effectués portent sur les trois comparaisons EL/RT, GT/EL et GT/RT. Nous allons présenter et étudier successivement les rapports représentatifs de ces différences de traitement entre types d'erreurs. Nous tenterons de dégager le moteur de comparaison le plus proche des conclusions fournies par les tests sur la similarité mélodique.

6.5.2 Relation entre EL et RT

Afin d'observer la manière dont EL et RT sont traitées l'une par rapport à l'autre, nous étudions le rapport suivant :

$$\frac{D_X(\bar{r} = \bar{d} + \bar{b}_{EL(a_1,\alpha_1)} \ , \ \bar{d})}{D_X(\bar{r} = \bar{d} + \bar{b}_{RT(a_2,\alpha_2)} \ , \ \bar{d})}$$

X désignant le moteur de comparaison concerné, $X = \{h1, h2, i1, i2\}$.

Dans nos tests, pour un motif mélodique donné, EL et RT étaient systématiquement injectées au même endroit afin de conserver le contexte mélodique de l'erreur. Afin d'en témoigner, le rang de l'erreur sera le même pour EL et RT, soit $\alpha_1 = \alpha_2 = \alpha$.

Voici les expressions représentatives des différences de traitement entre EL et RT, pour nos quatre moteurs de comparaison.

Moteur $h1$: Relation entre EL et RT

$$\frac{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{EL(a_1,\alpha_1)} \ , \bar{d})}{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{RT(a_2,\alpha_2)} \ , \bar{d})} = \frac{1}{\alpha-1} * \left|\frac{a_1}{a_2}\right| \qquad \begin{array}{ll} \text{si } \alpha \in [2; \frac{N+1}{2}] & \text{(N impair)} \\ \text{ou} & \\ \text{si } \alpha \in [2; \frac{N}{2}] & \text{(N pair)} \end{array}$$

$$\frac{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{EL(a_1,\alpha_1)} \ , \bar{d})}{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{RT(a_2,\alpha_2)} \ , \bar{d})} = \frac{2}{N} * \left|\frac{a_1}{a_2}\right| \qquad \text{si } \alpha = \frac{N}{2}+1 \quad \text{(N pair)}$$

$$\frac{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{EL(a_1,\alpha_1)} \ , \bar{d})}{D_{h1}(\bar{r}=\bar{d}+\bar{b}_{RT(a_2,\alpha_2)} \ , \bar{d})} = \frac{1}{N-\alpha+1} * \left|\frac{a_1}{a_2}\right| \qquad \begin{array}{ll} \text{si } \alpha \in [\frac{N+3}{2}; N-1] & \text{(N impair)} \\ \text{ou} & \\ \text{si } \alpha \in [\frac{N}{2}+2, N-1] & \text{(N pair)} \end{array}$$

Moteur $h2$: Relation entre EL et RT

$$\frac{D_{h2}(\bar{r} = \bar{d} + \bar{b}_{EL(a_1,\alpha_1)} \ , \ \bar{d})}{D_{h2}(\bar{r} = \bar{d} + \bar{b}_{RT(a_2,\alpha_2)} \ , \ \bar{d})} = \frac{N-1}{(\alpha-1)(N-\alpha+1)} * \left(\frac{a_1}{a_2}\right)^2 \tag{6.4}$$

Moteurs *i* : Relation entre EL et RT

$$\frac{D_{i\gamma}(\bar{r} = \bar{d} + \bar{b}_{EL(a_1,\alpha_1)}\ ,\ \bar{d})}{D_{i\gamma}(\bar{r} = \bar{d} + \bar{b}_{RT(a_2,\alpha_2)}\ ,\ \bar{d})} = 2 * \left|\frac{a_1}{a_2}\right|^{\gamma} \tag{6.5}$$

avec $\gamma = \{1,2\}$.

Pour les moteurs *h*, les expressions dépendent donc de la taille *N* de la requête, ainsi que de l'emplacement α de l'erreur. Par contre, pour les moteurs *i*, nous observons une totale indépendance à ces deux facteurs.

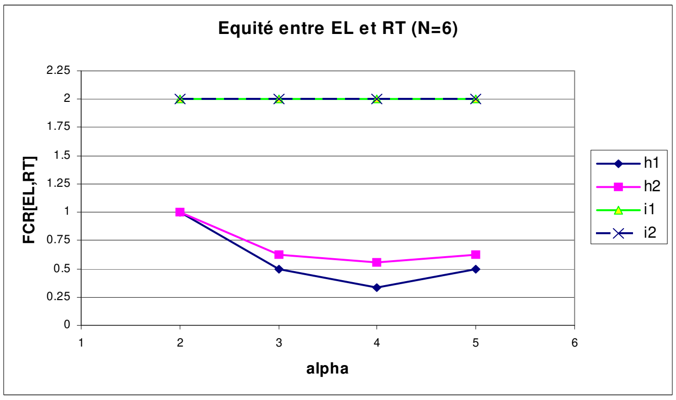
Equité entre EL et RT

Le premier critère dégagé par nos tests est l'équivalence des perturbations provoquées par une EL et une RT de même amplitude. Nous allons donc observer si, lorsque $a_1 = a_2$, les expressions que nous venons de présenter sont proches de 1. La figure 6.2 illustre leur comportement en fonction de α, pour trois valeurs de *N*.

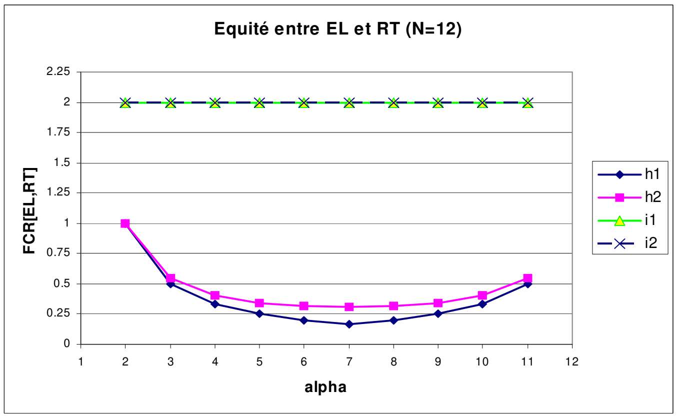
Excepté le cas où les erreurs sont placées sur la deuxième note de la requête, les moteurs ne remplissent pas le critère d'équité. Les moteurs *i* présentent un comportement constant qui consiste à pénaliser deux fois plus les EL que les RT. Les moteurs *h* présentent un comportement dépendant à la fois de *N* et de α.

Sur deux des trois cas de figure illustrés (*N* = 12 et 18), les moteurs *i* s'approchent d'avantage de l'équité désirée. Pour les requêtes courtes (*N* = 6), c'est le moteur *h*2 qui donne de meilleurs résultats, pénalisant cependant les RT par rapport aux EL.

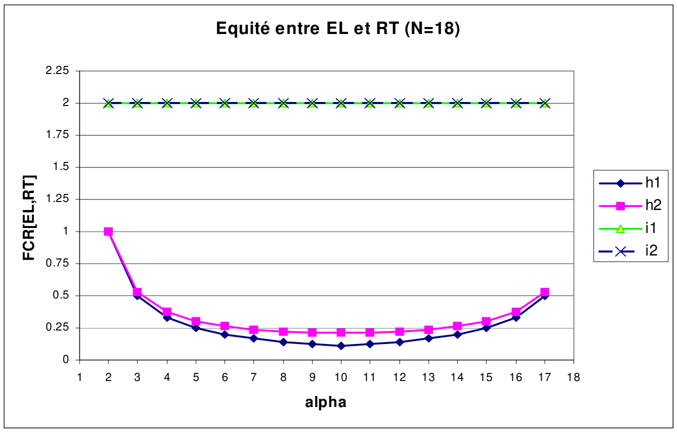
Selon le critère de l'équité entre EL et RT, le moteur *h*2 est donc le meilleur pour les requêtes courtes. Pour les requêtes moyennes et grandes, les moteurs de comparaison *i*, indépendants de *N* et de α, fournissent un traitement plus équitable entre EL et RT.



(a)



(b)



(c)

figure 6.2 : *Influence de l'emplacement des erreurs sur l'équité entre EL et RT, en fonction de α, pour (a) N = 6 ; (b) N = 12 ; (c) N = 18.*

6.5.3 Relation entre GT et EL

Dans cette section, nous allons observer comment les GT sont traités par rapport aux EL. Nous considèrerons donc l'expression suivante :

$$\frac{D_X(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_X(\bar{d}+\bar{b}_{EL(a_2,\alpha)}\ ,\ \bar{d})}$$

X désignant toujours le moteur de comparaison concerné, $X = \{h1, h2, i1, i2\}$.

Voici les expressions représentatives des différences de traitement entre GT et EL, pour nos quatre moteurs de comparaison.

Moteur $h1$: Relation entre GT et EL

$$\frac{D_{h1}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h1}(\bar{d}+\bar{b}_{EL(a_2)}\ ,\ \bar{d})} = \frac{N^2-1}{4} * \left|\frac{a_1}{a_2}\right| \qquad \text{si N impair}$$

$$\frac{D_{h1}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h1}(\bar{d}+\bar{b}_{EL(a_2)}\ ,\ \bar{d})} = \frac{N^2}{4} * \left|\frac{a_1}{a_2}\right| \qquad \text{si N pair}$$

Moteur $h2$: Relation entre GT et EL

$$\frac{D_{h2}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h2}(\bar{d}+\bar{b}_{EL(a_2)}\ ,\ \bar{d})} = \frac{N^2(N+1)}{12} * \left(\frac{a_1}{a_2}\right)^2 \tag{6.6}$$

Moteurs i : Relation entre GT et EL

$$\frac{D_{i\gamma}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{i\gamma}(\bar{d}+\bar{b}_{EL(a_2)}\ ,\ \bar{d})} = \frac{(N-1)}{2} * \left|\frac{a_1}{a_2}\right|^{\gamma} \tag{6.7}$$

avec $\gamma = \{1, 2\}$.

Pour cette relation entre GT et EL, l'ensemble des moteurs présente une indépendance à l'emplacement α de l'EL.

Correspondance entre GT et EL

A la différence de la comparaison précédente (EL avec RT), la correspondance établie porte, cette fois-ci, sur des amplitudes d'erreurs différentes : la perturbation d'une GT d'amplitude 0.3 a été jugée équivalente à celle d'une EL d'amplitude 1.8.

La figure 6.3 présente l'évolution des expressions que nous venons de présenter en fonction de N, pour $a1 = 0.3$ et $a2 = 1.8$.

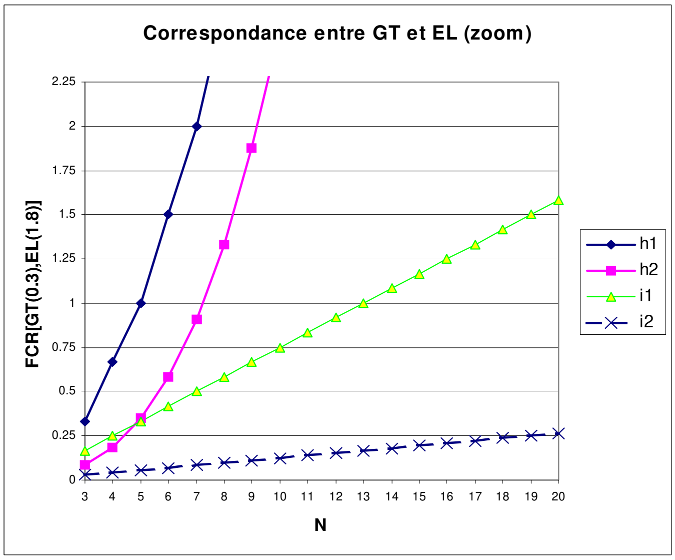


figure 6.3 : *Influence de la taille de la mélodie sur la correspondance entre GT et EL.*

	$\frac{D_X(d+b_{GT(0.3)}\ ,\ d)}{D_X(\bar{d}+\bar{b}_{EL(1.8)}\ ,\ \bar{d})}$			
moteur	$h1$	$h2$	$i1$	$i2$
$N = 3$	**0.33**	0.08	0.17	0.03
$N = 4$	**0.67**	0.19	0.25	0.04
$N = 5$	**1**	0.35	0.33	0.06
$N = 6$	**1.5**	0.58	0.42	0.07
$N = 7$	2	**0.91**	0.5	0.08
$N = 8$	2.67	**1.33**	0.58	0.10
$N = 9$	3.33	1.88	**0.67**	0.11
$N = 10$	4.17	2.55	**0.75**	0.13
$N = 11$	5.00	3.36	**0.83**	0.14
$N = 12$	6.00	4.33	**0.92**	0.15
$N = 13$	7.00	5.48	**1.00**	0.17
$N = 14$	8.17	6.81	**1.08**	0.18
$N = 15$	9.33	8.33	**1.17**	0.19
$N = 16$	10.67	10.07	**1.25**	0.21
$N = 17$	12.00	12.04	**1.33**	0.22
$N = 18$	13.50	14.25	**1.42**	0.24
$N = 19$	15.00	16.71	**1.50**	0.25
$N = 20$	16.67	19.44	**1.58**	0.26

TAB. 6.6 : *Influence de la taille de la mélodie sur la correspondance entre GT et EL.*

Pour ce deuxième critère, les moteurs atteignent d'avantage la valeur unité recherchée. Ils présentent également un changement de comportement lorsque la taille N des mélodies considérées augmente. Favorisant d'abord les GT, puis les EL pour les requêtes plus longues, les moteurs ne "basculent" pas tous pour la même valeur de N.

Le tableau 6.6 récapitule les valeurs illustrées. Comme nous pouvons le voir, les moteurs h ont l'avantage pour les requêtes courtes : $h1$ pour N compris entre 3 et 6, et $h2$ pour N égal à 6 et 7. Au delà, c'est le moteur $i1$ qui est le plus proche de la valeur unité recherchée. C'est seulement à partir de N = 31 que le moteur $i2$ se place en première position.

Selon le critère de correspondance entre GT et EL, les moteurs $h1$ et $h2$ sont donc plus adaptés aux requêtes courtes, le moteur $i1$ étant performant pour les requêtes de taille moyenne et grande.

6.5.4 Relation entre GT et RT

Dans cette section, nous allons observer comment les GT sont traités par rapport aux RT. Nous considèrerons donc l'expression suivante :

$$\frac{D_X(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_X(\bar{d}+\bar{b}_{RT(a_2,\alpha)}\ ,\ \bar{d})} \tag{6.8}$$

X désignant toujours le moteur de comparaison concerné, $X = \{h1, h2, i1, i2\}$.

Voici les expressions représentatives des différences de traitement entre GT et RT, pour nos quatre moteurs de comparaison.

Moteur $h1$: Relation entre GT et RT

1. Si N est impair

$$\begin{aligned}
\frac{D_{h1}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h1}(\bar{d}+\bar{b}_{RT(a_2)}\ ,\ \bar{d})} &= \frac{N^2-1}{4(\alpha-1)} * \left|\frac{a_1}{a_2}\right| && \text{si} \quad \alpha \in [2; \tfrac{N+1}{2}] \\
&= \frac{N^2-1}{4(N-\alpha+1)} * \left|\frac{a_1}{a_2}\right| && \text{si} \quad \alpha \in [\tfrac{N+3}{2}; N-1]
\end{aligned}$$

2. Si N est pair

$$\begin{aligned}
\frac{D_{h1}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h1}(\bar{d}+\bar{b}_{RT(a_2)}\ ,\ \bar{d})} &= \frac{N^2}{4(\alpha-1)} * \left|\frac{a_1}{a_2}\right| && \text{si} \quad \alpha \in [2, \tfrac{N}{2}] \\
&= \frac{N}{2} * \left|\frac{a_1}{a_2}\right| && \text{si} \quad \alpha = \tfrac{N}{2}+1 \\
&= \frac{N^2}{4(N-\alpha+1)} * \left|\frac{a_1}{a_2}\right| && \text{si} \quad \alpha \in [\tfrac{N}{2}+2, N-1]
\end{aligned}$$

Moteur $h2$: Relation entre GT et RT

$$\frac{D_{h2}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{h2}(\bar{d}+\bar{b}_{RT(a_2)}\ ,\ \bar{d})} = \frac{N^2(N^2-1)}{12(\alpha-1)(N-\alpha+1)} * \left(\frac{a_1}{a_2}\right)^2 \tag{6.9}$$

Moteurs i : Relation entre GT et RT

$$\frac{D_{i\gamma}(\bar{d}+\bar{b}_{GT(a_1)}\ ,\ \bar{d})}{D_{i\gamma}(\bar{d}+\bar{b}_{RT(a_2)}\ ,\ \bar{d})} = (N-1) * \left|\frac{a_1}{a_2}\right|^\gamma \tag{6.10}$$

avec $\gamma = \{1,2\}$.

Une fois de plus, contrairement aux moteurs i, les moteurs h présentent des expressions dépendantes de l'emplacement α de la RT.

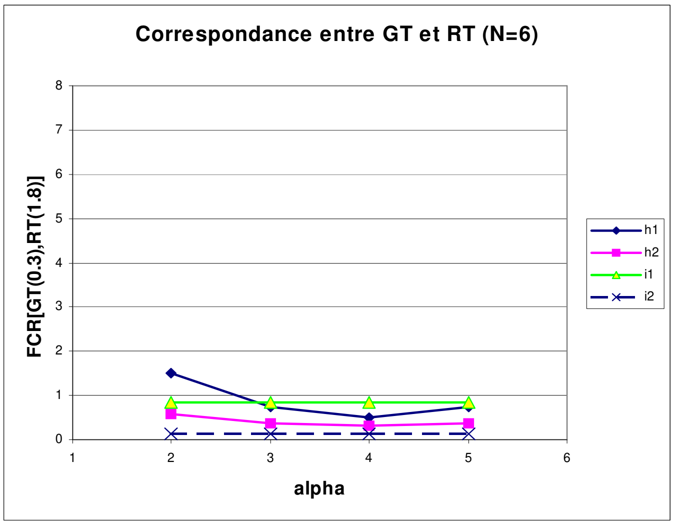
Correspondance entre GT et RT

Puisque EL et RT ont été jugées équivalentes, les amplitudes d'erreurs définissant la correspondance sont identiques au cas précédent. Nous avons donc $a_1 = 0.3$ (GT) et $a_2 = 1.8$ (RT).

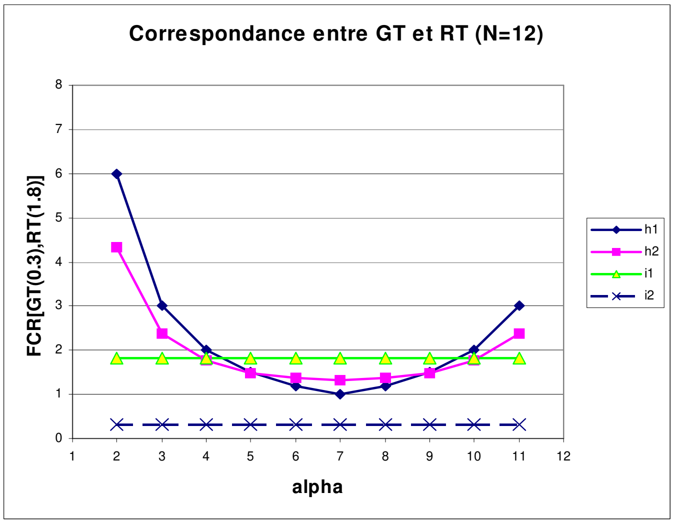
La figure 6.4 illustre le comportement des expressions que nous venons de présenter en fonction de α, pour trois valeurs de N (avec $a1 = 0.3$ et $a2 = 1.8$).

Sur ce troisième critère, les expressions s'approchent de la valeur unité recherchée, mais nous voyons qu'une fois de plus, les moteurs h présentent de fortes variations qui les en éloignent rapidement.

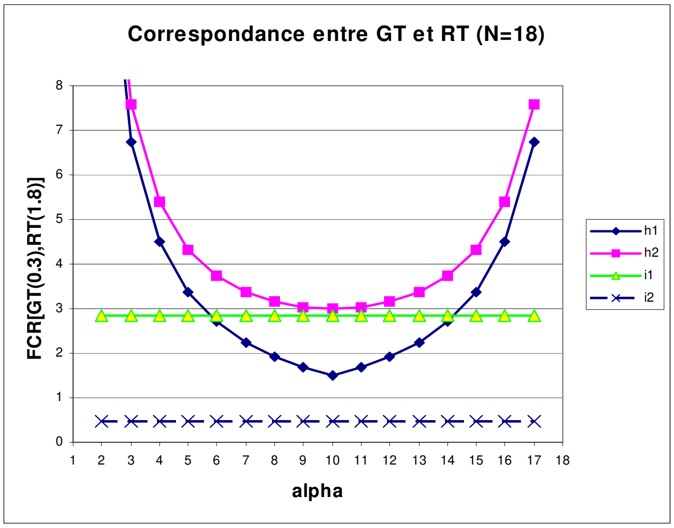
Pour les requêtes courtes, le moteur $i1$ donne les meilleurs résultats. Pour les requêtes de taille moyenne, la suprématie est partagée. Lorsque α emprunte les valeurs extrêmes (début et fin de requête), le moteur $i1$ reste le plus conforme au critère de correspondance. Par contre, pour les α situés en milieu de requête, les moteurs h sont les meilleurs (et $h1$ meilleur que $h2$). Concernant les grandes requêtes, $i2$, meilleur que $i1$, est moins bon que $h1$ sur les α medium.



(a)



(b)



(c)

figure 6.4 : *Influence de l'emplacement des erreurs sur la correspondance entre GT et RT, pour (a) N = 6 ; (b) N = 12 ; (c) N = 18.*

		$\frac{D_X(d+b_{GT(0.3)}, d)}{D_X(\bar{d}+\bar{b}_{RT(1.8,\alpha)}, \bar{d})}$			
moteur		$h1$	$h2$	$i1$	$i2$
$N = 6$	$\alpha = 2$	1.50	0.58	**0.83**	0.14
	$\alpha = 3$	0.75	0.36	**0.83**	0.14
	$\alpha = 4$	0.50	0.32	**0.83**	0.14
	$\alpha = 5$	0.75	0.36	**0.83**	0.14
$N = 12$	$\alpha = 2$	6.00	4.33	**1.83**	0.31
	$\alpha = 3$	3.00	2.38	**1.83**	0.31
	$\alpha = 4$	2.00	**1.77**	1.83	0.31
	$\alpha = 5$	1.50	**1.49**	1.83	0.31
	$\alpha = 6$	**1.20**	1.36	1.83	0.31
	$\alpha = 7$	**1.00**	1.32	1.83	0.31
	$\alpha = 8$	**1.20**	1.36	1.83	0.31
	$\alpha = 9$	1.50	**1.49**	1.83	0.31
	$\alpha = 10$	2.00	**1.77**	1.83	0.31
	$\alpha = 11$	3.00	2.38	**1.83**	0.31
$N = 18$	$\alpha = 2$	13.50	14.25	2.83	**0.47**
	$\alpha = 3$	6.75	7.57	2.83	**0.47**
	$\alpha = 4$	4.50	5.38	2.83	**0.47**
	$\alpha = 5$	3.38	4.33	2.83	**0.47**
	$\alpha = 6$	2.70	3.73	2.83	**0.47**
	$\alpha = 7$	2.25	3.36	2.83	**0.47**
	$\alpha = 8$	**1.93**	3.15	2.83	0.47
	$\alpha = 9$	**1.69**	3.03	2.83	0.47
	$\alpha = 10$	**1.50**	2.99	2.83	0.47
	$\alpha = 11$	**1.69**	3.03	2.83	0.47
	$\alpha = 12$	**1.93**	3.15	2.83	0.47
	$\alpha = 13$	2.25	3.36	2.83	**0.47**
	$\alpha = 14$	2.70	3.73	2.83	**0.47**
	$\alpha = 15$	3.38	4.33	2.83	**0.47**
	$\alpha = 16$	4.50	5.38	2.83	**0.47**
	$\alpha = 17$	6.75	7.57	2.83	**0.47**

TAB. 6.7 : *Influence de l'emplacement des erreurs sur la correspondance entre GT et RT, pour N = 6 ; N = 12 ; N = 18.*

6.5.5 Classement des moteurs de comparaison proposés

Nous venons d'appliquer les conclusions des tests subjectifs en distinguant, lorsque cela était nécessaire différentes tailles de requêtes et zones d'emplacements d'erreur. Les observations effectuées ne permettent pas de désigner un gagnant unique parmi les quatre moteurs proposés. Comme le montre le tableau 6.8, qui rassemble les moteurs les plus performants, aucun moteur ne réunit à lui seul toutes les qualités attendues. Le choix d'une méthode particulière ne découle pas directement de l'application des critères dégagés par nos tests.

Correspondance	Vainqueur(s)	Domaine
EL/RT	$h2$ $i1$ et $i2$	$N = 6$ $N = 12, 18$
GT/EL	$h1$ $h2$ $i1$	$3 \leq N \leq 6$ $N = 7, 8$ $9 \leq N \leq 30$
GT/RT	$i1$ $h1$ puis $h2$ puis $i1$ $h1$ puis $i2$	$N = 6$ selon α ; $N = 12$ selon α ; $N = 18$

TAB. 6.8 : *Récapitulatifs des vainqueurs pour chacun des critères objectifs.*

Les moteurs fondés sur les *séquences d'intervalles* sont peu sensibles à la taille des requêtes et aux emplacements d'erreur. Cela leur permet de ne pas trop s'éloigner des critères recherchés, mais sans les atteindre, bien souvent. Les moteurs fondés sur les *profils de hauteurs* remplissent quasi-systématiquement les critères, mais généralement de manière ponctuelle. Ailleurs, leur comportement, très sensible à la taille des requêtes et aux emplacements d'erreur, est souvent bien pire que les moteurs *i*.

Une manière d'établir un classement est de considérer les douze expressions présentées dans 6.5.2, 6.5.3, et 6.5.4, et de les calculer pour $N = 9$ et $\alpha = 5$. Ces deux valeurs constituent une sorte de "point de fonctionnement" de nos tests subjectifs. En effet, elles correspondent respectivement à la taille moyenne des mélodies diffusées, ainsi qu'à l'emplacement moyen des EL/RT injectées. Les résultats sont présentés tableau 6.9.

		moteur			
Correspondance	Expression	$h1$	$h2$	$i1$	$i2$
EL/RT	$\frac{D_X(\bar{d}+\bar{b}_{EL(a,5)}\ ,\bar{d})}{D_X(\bar{d}+\bar{b}_{RT(a,5)}\ ,\bar{d})}$ pour $N = 9$	0.25	0.40	<u>**2.00**</u>	<u>**2.00**</u>
GT/EL	$\frac{D_X(\bar{d}+\bar{b}_{GT(0.3)}\ ,\bar{d})}{D_X(\bar{d}+\bar{b}_{EL(1.8,5)}\ ,\bar{d})}$ pour $N = 9$	3.33	**1.86**	<u>**0.67**</u>	0.11
GT/RT	$\frac{D_X(\bar{d}+\bar{b}_{GT(0.3)}\ ,\bar{d})}{D_X(\bar{d}+\bar{b}_{RT(1.8,5)}\ ,\bar{d})}$ pour $N = 9$	<u>**0.83**</u>	**0.75**	**1.33**	0.22

TAB. 6.9 : *Valeurs des expressions témoignant des correspondances établies par nos tests subjectifs, au "point de fonctionnement" de ces derniers (i.e. $N = 9$, et $\alpha = 5$).*

Au regard de ce dernier résultat, il semble que le moteur $i1$ ait l'avantage. Classé deux fois premier et une fois deuxième, il se place devant $h1$ (une fois premier, et deux fois troisième), $h2$ (deux fois deuxième et une fois quatrième), puis $i2$ (une fois premier, et deux fois quatrième).

6.5.6 Conclusion

A la lumière des critères issus des tests subjectifs, aucun des moteurs proposés ne prend franchement le pas sur les autres. Selon ces critères, les moteurs $h1$ et $h2$ ont montré leur capacité à traiter avec succès des requêtes plutôt courtes et moyennes. Cependant, la dépendance de leurs performances vis-à-vis de la taille des requêtes et de l'emplacement des erreurs les rend moins stables que $i1$.

Par ailleurs, la qualité des critères objectifs définis n'est pas assez grande (perturbations simples, résultats moyennés) pour motiver la mise en place d'un moteur de comparaison hybride (i.e. sélection automatique du moteur utilisé en fonction de la taille de la requête soumise). S'il fallait désigner un unique gagnant, ce serait le moteur $i1$.

6.6 Conclusion

Face au manque de mesures objectives exploitables, nous avons procédé à des tests subjectifs inédits afin de dégager des critères de jugement des moteurs de comparaison développés au Chapitre 5. En faisant identifier des mélodies dégradées à des sujets, nous avons établi des correspondances entre les dissimilarités causées par trois types d'erreurs fréquentielles simples :

- Erreurs Locales - EL,
- Ruptures de Ton - RT,
- Glissement de Ton - GT.

Nos tests ont montré qu'à amplitude égale EL et RT provoquaient des dissimilarités comparables. Par ailleurs, ils ont également révélé qu'un GT progressant de 0.3 demi-ton/note causait une perturbation équivalente à une EL de 1.8 demi-tons, et conformément au premier critère, à une RT d'amplitude identique.

Ces trois critères objectifs ont été appliqués aux expressions analytiques des mesures de similarité des quatre moteurs de comparaison proposés. Ainsi, nous avons pu en déduire des expressions témoignant de la qualité de nos moteurs. L'observation des ces dernières nous a montré qu'*aucune méthode* ne satisfaisait à l'*ensemble des critères objectifs* définis par nos tests.

Les moteurs fondés sur les *profils de hauteurs*, plutôt performants sur les requêtes courtes et moyennes, présentent une sensibilité aux emplacements d'erreurs et aux tailles de requêtes qui les pénalise. Les moteurs fondés sur les *séquences d'intervalles*, peu sensibles à la taille des requêtes et aux emplacements d'erreur, sont plus homogènes dans leurs performances, mais ces dernières sont généralement moins bonnes que celles atteintes par les autres moteurs.

Les résultats obtenus pourraient encourager la mise en place d'un moteur de comparaison hybride qui sélectionnerait automatiquement le moteur utilisé en fonction de la taille de la requête soumise. Cependant, cette démarche serait prématurée compte-tenu de la qualité des critères objectifs définis. En effet, l'utilisation de perturbations simples, ainsi que la moyenne des résultats causée par le manque de données exploitables, nous ont simplement permis d'obtenir un ordre d'idée de ce que nous souhaitions.

Les stimuli utilisés dans nos tests ne reflètent pas la complexité des erreurs rencontrées dans des requêtes chantonnées. Sur ce point, une perspective d'amélioration consisterait à déterminer la manière dont les types d'erreur simples se combinent dans les mélodies chantonnées. La démarche que nous avons adoptée (i.e. jugement des moteurs de comparaison à partir de l'expression analytique de leur mesure de similarité) serait enrichie, aboutissant à un jugement plus élaboré.

Ce type d'expérience gagnerait également à être renouvelé à plus grande échelle, afin de disposer de davantage de données exploitables. Seulement, l'augmentation du nombre de sujets s'accompagne généralement d'une diversification des cultures musicales représentées, synonyme d'un affaiblissement du nombre de mélodies connues de tous les sujets.

Une alternative aux tests réalisés consisterait à faire classer différentes versions d'une mélodie dégradée par ordre de similarité à l'original. Les sujets constitueraient ainsi des "classements-références" permettant de juger de la qualité de moteurs de comparaison. Cela permettrait de traiter des dégradations plus réalistes (en ignorant cependant la contribution propre des erreurs présentes). La difficulté de mise en œuvre consisterait à proposer des mélodies que les sujets pourraient effectivement ordonner.

Ne sachant pas quelle confiance accorder au jugement que nous venons de prononcer concernant nos moteurs de comparaison, nous allons adopter, dans le chapitre suivant, une autre manière de qualifier leur efficacité.

Chapitre 7

Evaluation de la Qualité de SR Mélodiques

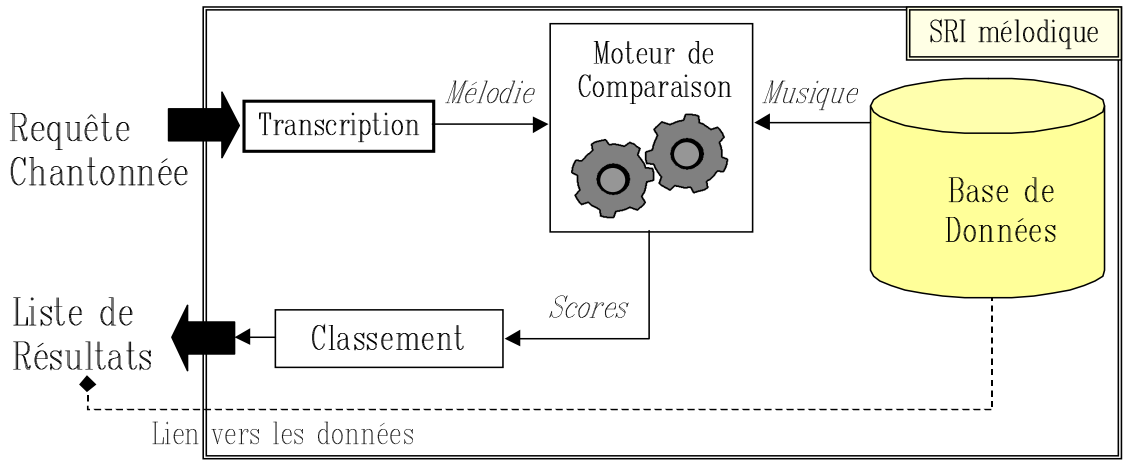


figure 7.1 : *Système de recherche de documents musicaux par chantonnement.*

7.1 Introduction

Dans le chapitre précédent, nous avons tenté de mesurer objectivement la qualité des moteurs de comparaison (i.e. descripteurs mélodiques associés à une mesure de similarité) proposés au chapitre 5. Seulement, il s'avère difficile de comparer les similarités mesurées par les moteurs avec les similarités humainement perçues.

Afin d'assurer une qualification des moteurs proposés, nous allons passer par la qualification de systèmes complets (cf. figure 7.1). Cependant, les systèmes considérés ne diffèreront que par leurs moteurs de comparaison. Nous pourrons donc comparer la qualité de ces derniers puisque les autres éléments des systèmes testés (transcription automatique de la requête, base de données) seront les mêmes.

Dans un premier temps, nous allons voir les critères sur lesquels nous allons nous appuyer pour évaluer la qualité des systèmes de recherche étudiés. Ceux-ci nous permettront de présenter l'application que nous proposons. Nous présenterons ensuite les différentes configurations (i.e. moteurs) testées, puis nous exposerons les résultats obtenus.

7.2 Critères de qualité

Dans [SM83] (p. 162), six critères ont été dégagés pour l'évaluation de la qualité de systèmes de recherche d'information. Nous les reprenons ci-dessous.

1. L'*étendue* de la collection de documents constituant la base de données.
2. L'*effort*, physique ou intellectuel, que doit fournir l'utilisateur pour formuler sa requête, guider la recherche, et consulter la réponse qui lui est fournie ;
3. La *présentation* du résultat, qui influence l'exploitation des éléments qui la constituent ;
4. Le *rappel*, qui témoigne de la capacité d'un système à présenter tous les éléments pertinents vis-à-vis de la requête formulée par l'utilisateur ;
5. La *précision*, qui témoigne de la capacité d'un système à ne présenter que les éléments pertinents ;
6. Le *temps* d'attente entre la soumission de la requête et l'obtention du résultat ;

Le jugement suivant les trois premiers critères sera le même pour tous les systèmes testés puisque ces derniers ne diffèrent que par leurs moteurs de comparaison. Nous n'avons pas mis en place des tests spécifiques à ces trois critères, mais nous avancerons néanmoins certains éléments de réponse. Ceux-ci font l'objet de la sous-section 7.2.1.

Au contraire, les trois derniers critères sont directement liés aux moteurs de comparaison utilisés. Les résultats devraient donc varier d'une configuration de système à l'autre. Le traitement des critères 4, 5 et 6 sera abordé dans la sous-section 7.2.2.

7.2.1 Qualité de l'application proposée

Concernant l'étendue de la base de données musicales (critère 1), nous avons vu en 5.3.5 que la base de données dont nous disposions pour notre système de recherche contenait de nombreux documents (19.282 fichiers) aux styles musicaux variés. Chaque instrument (i.e. piste MIDI) possédant sa description propre, l'information mélodique accessible est très importante (près de 40 millions de notes).

Au stade de prototype, notre interface d'utilisation fournit des voies de requête et de consultation des résultats conviviales. Comme illustré figure 7.2, le module d'enregistrement de la requête est simple : un bouton pour commencer l'enregistrement (*Record*), un bouton pour le clore (*Stop*). Deux indicateurs permettent de contrôler le niveau du signal d'entrée et la durée de la requête.

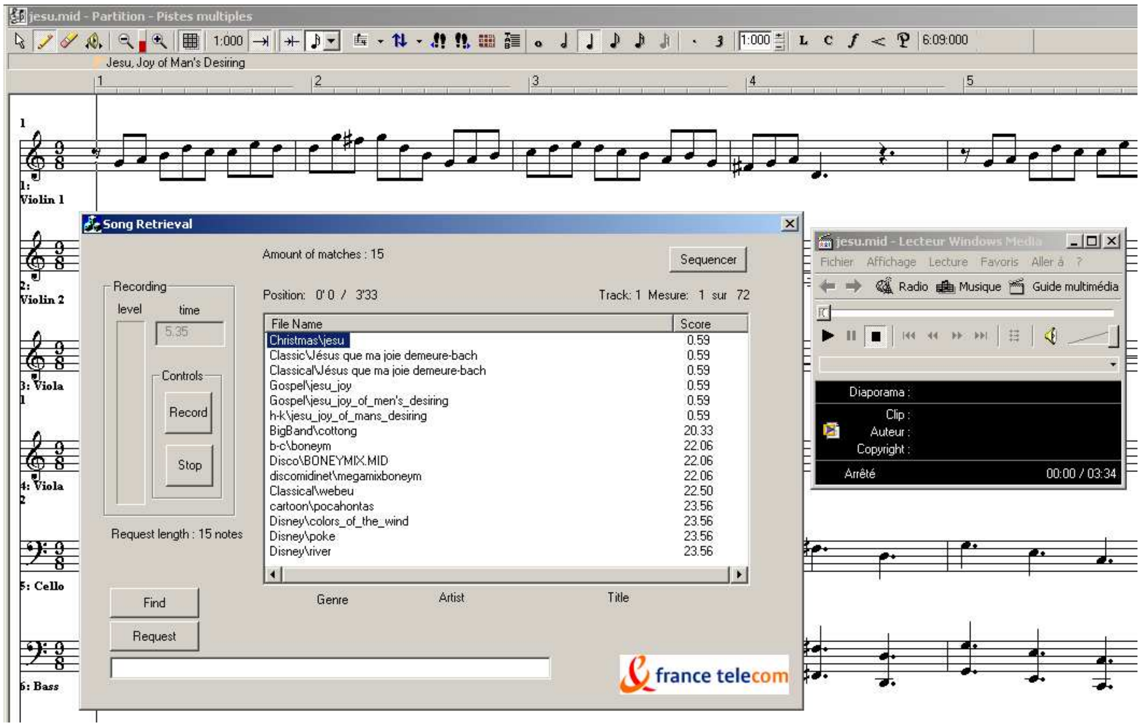


figure 7.2 : *Interface d'utilisation du moteur de recherche, accompagnée d'applications complémentaires pour la consultation des résultats proposés.*

L'effort que l'utilisateur doit fournir (cf. critère 2) consiste donc à chantonner la mélodie recherchée en prononçant un "ta" pour chaque nouvelle note. Aucune information complémentaire à la mélodie soumise n'est requise.

Une fois sa requête réalisée, l'utilisateur peut, s'il le souhaite, vérifier que la mélodie extraite à la transcription est bien celle qu'il recherche. Ainsi, il peut choisir de lancer le moteur de comparaison sur la requête effectuée (bouton *Find*) ou bien en produire une nouvelle (bouton *Record*). La vérification de la requête peut s'effectuer visuellement par l'affichage d'une fenêtre graphique (raccourci clavier) présentant la requête transcrite dans le plan temps/fréquence (voir p. 62 pour un exemple). L'utilisateur peut également écouter la mélodie transcrite resynthétisée (bouton *Request*).

La figure 7.2 le montre également, le résultat est présenté sous la forme d'une liste de noms de fichiers. Chaque fichier est associé à un score correspondant au meilleur score rencontré dans toutes les mélodies qu'il contient. Dans le système présenté, la liste est limitée aux quinze meilleurs éléments. En effet, nous pensons que l'écoute nécessaire à la vérification de la pertinence d'un document a tendance à limiter la consultation aux premiers éléments de la liste. Par ailleurs, ce choix permet d'améliorer la rapidité du système.

Les différents modes de consultation proposés montrent que le critère 3 n'est pas négligé :

– Double-clic droit sur le nom du fichier : écoute de la portion mélodique donnant son score au fichier ;
– Double-clic gauche sur le nom du fichier : écoute du fichier MIDI (ou MP3 associé) grâce à une application externe (cf. à droite sur la figure 7.2). Une indication sur l'interface permet

d'aller placer le curseur de jeu à l'*instant* de la meilleure portion ;
- Sélection du nom du fichier et clic sur "Sequencer" : ouverture du ficher MIDI par un séquenceur qui permet, entre autres, de visualiser et de jouer la partition de chaque instrument. Une indication permet d'aller jouer le fichier à la *mesure* où commence la meilleure portion.

7.2.2 Choix de critères objectifs pour la qualification des réponses fournies

Le critère 6 est facilement mesurable puisqu'il s'agit du temps d'attente entre soumission de la requête et résultat. Nous le considérerons après avoir étudié la pertinence des documents renvoyés par les différents systèmes en testés.

Cette pertinence est prise en compte par les critères 4 et 5. Ils sont moins faciles à mettre en œuvre car la notion de pertinence peut être interprétée de différentes manières (cf. 3.6). Pour notre part, nous avons choisi de considérer comme pertinents pour une requête donnée, les documents retournés avec un score idéal (i.e. une distance nulle) lorsque la portion mélodique *visée* est injectée dans l'un des quatre systèmes proposés. Compte-tenu de la nature de ces derniers, les documents pertinents sont donc ceux qui comportent une séquence de hauteurs/intervalles *identique* à celle visée.

Cette démarche permet de ne pas considérer comme indésirable, un document différent du document visé, mais contenant une mélodie proche de la mélodie chantonnée (au sens de la mesure de similarité proposée).

Les performances des systèmes testés sont donc mesurées grâce aux critères de rappel et de précision, dont les formules sont rappelées ci-dessous [SM83] :

$$Rappel = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ total\ de\ documents\ pertinents}$$

$$Prcision = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ total\ de\ documents\ retrouvés}$$

Dans nos systèmes, le nombre de documents retrouvés est fixé aux 15 documents les plus similaires à la requête. Par conséquent, le critère de précision, renseignant sur la présence d'éléments indésirables, est lié au critère de rappel par la formule suivante :

$$Précision = Rappel * \frac{nombre\ total\ de\ documents\ pertinents}{15}$$

Toutes les requêtes n'ont pas le même nombre de documents pertinents. En effet, ceux-ci sont plus ou moins nombreux selon la taille et l'originalité des mélodies considérées. Le critère de rappel implique des résultats indépendants de ce facteur. La performance maximale est 1, quelle que soit la requête considérée.

7.3 Qualification objective de différentes configurations de systèmes

Dans un premier temps, nous allons juger les systèmes sur les critères de rappel et de précision. Le critère de rapidité viendra dans un second temps, pour compléter notre jugement. Nous commencerons par comparer les quatre moteurs de comparaison proposés au Chapitre 5. Nous comparerons ensuite trois méthodes de quantification sur la base d'un système représentant l'état-de-l'art. Enfin, nous verrons l'influence de la quantification sur les meilleurs moteurs des deux familles proposées dans cette thèse.

Les requêtes utilisées pour la stimulation sont celles qui ont été recueillies au Chapitre 4. Nous disposons donc d'un corpus de 500 requêtes. Chaque motif est utilisé plusieurs fois pour un système donné. En effet, nous désirons observer les critères de précision et de rappel en fonction du *nombre de notes* soumises. Les requêtes disponibles sont donc tronquées afin de ne disposer que des premières notes (de 5 au minimum, et jusqu'à 20 notes lorsque la requête le permet).

Le lourd travail d'identification des mélodies visées par les requêtes a été réalisé lors de l'étude du Chapitre 4[1]. La détermination des réponses pertinentes pour chaque requête (motif et taille donnés) est donc facilitée.

Pour chacune des 500 requêtes-tests disponibles, et pour chaque taille de requête considérée, nous adoptons le processus suivant :

1. Afin de déterminer les documents pertinents pour la requête considérée, le motif visé par la requête est soumis au système référence[2]. Ce motif correspond à une requête parfaitement réalisée ;
2. La requête elle-même est à son tour injectée dans le système testé.
3. La comparaison des réponses obtenues avec les documents pertinents recherchés permet de calculer les taux de rappel et de précision.

Pour une taille de requête donnée, le taux de rappel final est la moyenne des taux de rappel obtenus pour l'ensemble des requêtes de cette taille. Il en est de même pour le critère de précision final, moyenne des taux de précision collectés. La base de données utilisée est celle citée en 7.2.1.

7.3.1 Profils de hauteurs vs. Séquences d'intervalles

Nous allons comparer les quatre moteurs de comparaison proposés dans cette thèse. Deux d'entre eux, fondés sur les profils de hauteurs, utilisent un critère mathématique afin de s'affranchir du ton des mélodies comparées. La mesure de similarité employée gère de manière globale l'invariance aux tons des mélodies (cf. 5.4.2). Nous allons comparer l'efficacité cette démarche par rapport à celle garantissant l'indépendance au ton dès la description. Fondée sur les séquences d'intervalles, celle-ci gère localement l'invariance aux tons par l'utilisation d'une information relative.

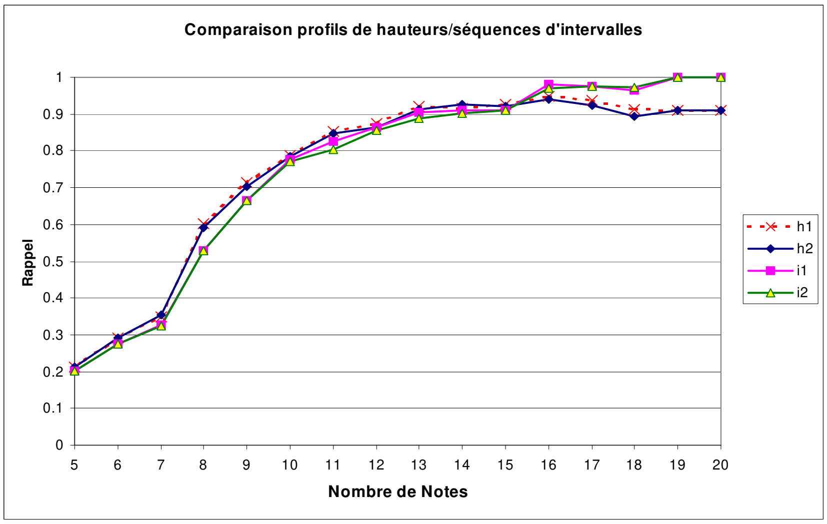
Le tableau 7.1 rappelle les caractéristiques des moteurs de comparaison testés.

[1] Compte-tenu de l'adaptation des motifs visés effectué au Chapitre 4, le contexte de test dans lequel nous nous inscrivons ne considère pas les erreurs de type insertions/omissions de notes.

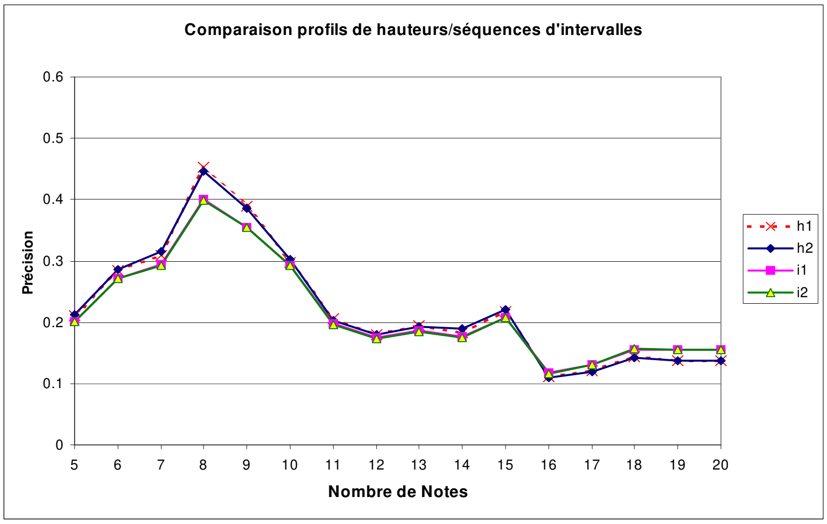
[2] Quasiment tous les systèmes présentés ici peuvent jouer ce rôle. En effet, les hauteurs du motif visé étant tempérées, les documents ayant un score idéal (i.e. distance nulle) seront les mêmes quel que soit le système utilisé, excepté UDS.

Nom	Type de descripteur	Quantification	Mesure de similarité
*h*1	profil de hauteurs	non	L1 avec ajustement
*h*2	profil de hauteurs	non	L2 avec ajustement
*i*1	séquence d'intervalles	non	L1
*i*2	séquence d'intervalles	non	L2

TAB. 7.1 : *Profils de hauteurs vs. Séquences d'intervalles : Systèmes en compétition.*



(a)



(b)

figure 7.3 : *Comparaison des performances des quatre moteurs de comparaison proposés. (a) rappel ; (b) précision.*

La figure 7.3 présente les performances des quatre systèmes proposés. Nous pouvons voir que, globalement, les performances de rappel s'améliorent avec l'augmentation du nombre de notes soumises. Par ailleurs, les taux obtenus sont élevés, à l'image des 90% obtenus dès 13 notes. Les systèmes proposés sont donc tous performants.

Les moteurs fondés sur les séquences d'intervalles atteignent les 100% de rappel pour les requêtes longues, alors que ceux fondés sur les profils de hauteurs présentent des performances légèrement décroissantes. Cela est sans doute dû à la difficulté qu'ont les utilisateurs à maintenir un ton stable sur des requêtes longues. En contrepartie, ces moteurs fondés sur les profils de hauteurs obtiennent plus rapidement de meilleurs taux de rappel. A l'instar des conclusions dégagées au Chapitre 6, les systèmes fondés sur les profils de hauteurs donnent de meilleures performances pour les requêtes courtes et moyennes (jusqu'à 15 notes), au delà, la hiérarchie s'inverse. Nous retrouvons également le fait qu'au sein d'une même famille, les mesures de similarités utilisant la distance L1 sont (très légèrement) meilleurs.

Il apparaît donc que l'ordre d'idée obtenu grâce à nos tests subjectifs soit fidèle à la réalité, bien que ces derniers fussent fondés sur des perturbations simples et une moyenne des données exploitables.

La figure 7.3 nous permet de conclure qu'une mesure de similarité peut assurer l'indépendance au ton des mélodies comparées (moteurs $h1$ et $h2$), autrement qu'en déclinant la requête selon les 12 tons possibles (cf. 3.3.6). Pour les requêtes inférieures à 16 notes, les résultats obtenus sont même supérieurs à ceux donnés par les moteurs dont l'indépendance au ton est assurée à la description (par l'utilisation d'intervalles : moteurs $i1$ et $i2$).

Lors de la réalisation des tests, nous avons remarqué la plus grande lenteur des moteurs fondés sur la distance L1 (notamment, $h1$ avec son ajustement par calcul de valeur médiane). C'est pourquoi, bien que les moteurs $i1$ et $h1$ présentent des résultats légèrement meilleurs, nous leur préfèrerons les moteurs $i2$ et $h2$, pour leur rapidité. Ce choix témoigne de l'influence du critère 6 (cf. p. 146).

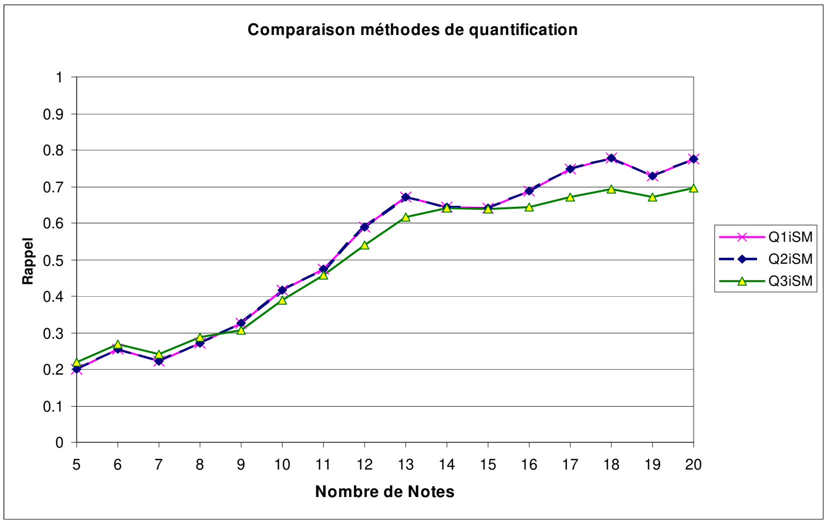
7.3.2 Quantifications

A notre connaissance, tous les systèmes développés jusqu'à présent appliquent une quantification des hauteurs de la requête. Dans le chapitre 5, nous en avons considéré plusieurs sortes (cf. 5.2.2), nous allons maintenant comparer trois d'entre elles. Pour cela nous allons prendre comme base un système représentant l'état-de-l'art. Dans celui-ci, les mélodies sont décrites par des séquences d'intervalles quantifiés, la mesure de similarité est fondée sur un algorithme de string matching flexible. A la différence d'un calcul de distance, ce dernier pénalise les substitutions indépendamment de leur amplitude. Une erreur d'un demi-ton sera donc autant pénalisée qu'une erreur de 2 tons.

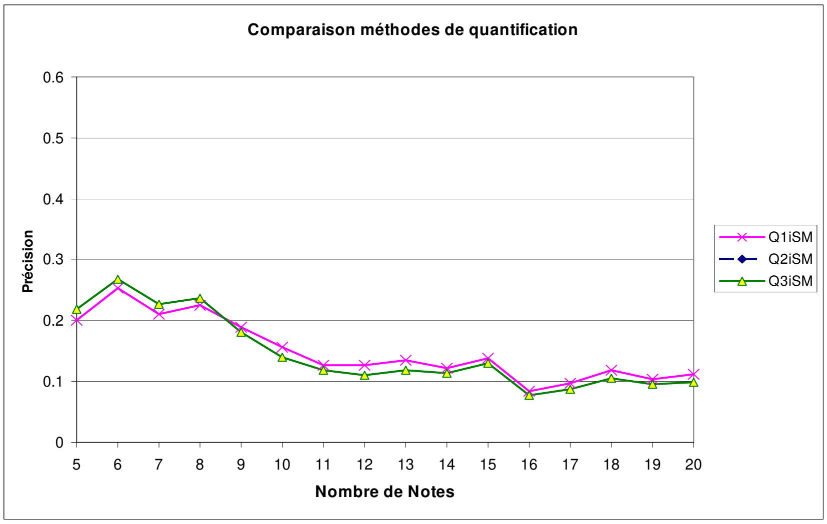
Le tableau 7.2 rappelle les caractéristiques des moteurs de comparaison testés. Chacun d'eux possède sa méthode de quantification propre. La première consiste en une quantification brutale appliquée sur les intervalles (cf. p. 108), la seconde est celle proposée par McNab (quantification adaptative, cf. p. 107), et la dernière est l'alternative que nous avons proposée (quantification fondée sur l'extraction automatique de la tonalité, cf. p. 107).

Nom	Type de descripteur	Quantification	Mesure de similarité
Q1iSM	séquence d'intervalles	Quant1 (directe)	string matching
Q2iSM	séquence d'intervalles	Quant2 (McNab)	string matching
Q3iSM	séquence d'intervalles	Quant3 (Carré)	string matching

TAB. 7.2 : *Systèmes en compétition pour la meilleure quantification.*



(a)



(b)

figure 7.4 : *Comparaison des performances de trois méthodes de quantification. (a) rappel ; (b) précision.*

La figure 7.4 présente les performances des trois systèmes testés. La quantification proposée par McNab donne des résultats identiques à ceux d'une quantification directe sur les intervalles. Les glissements de tonalité que cette méthode est censée compenser ne sont donc pas si courants ou pas assez importants pour perturber la requête quantifiée ou, si cela était le cas, pour modifier le nombre d'éléments pertinents dans la liste des réponses du système.

Notre méthode de quantification donne des résultats légèrement meilleurs pour les requêtes courtes. A partir de 9 notes, les deux autres méthodes testées lui sont supérieures.

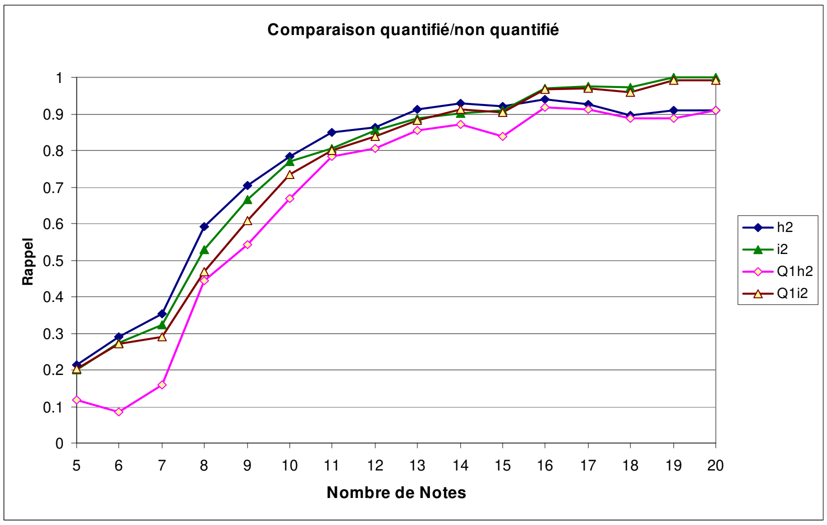
La quantification directe sur les intervalles, à la fois simple et efficace, semble donc être la meilleure solution. Nous allons maintenant voir l'influence de cette quantification sur nos moteurs de comparaison.

7.3.3 Quantifié vs. Non quantifié

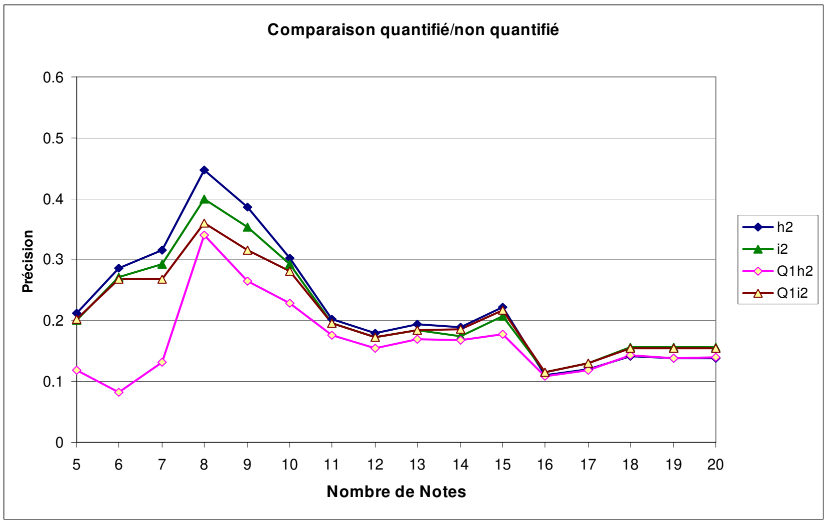
L'une des originalités de cette thèse est l'utilisation de hauteurs/intervalles non quantifiés pour représenter la requête au sein du moteur de comparaison. Afin de montrer l'intérêt de notre démarche, nous allons comparer les moteurs $i2$ et $h2$ (sélectionnés pour leurs performances en terme de précision/rappel et de rapidité) avec leurs versions "quantifiées". Le tableau 7.3 présente leurs caractéristiques.

Nom	Type de descripteur	Quantification	Mesure de similarité
$h2$	profil de hauteurs	non	L2 avec ajustement
$i2$	séquence d'intervalles	non	L2
$Q1h2$	profil de hauteurs	Quant1 (directe)	L2 avec ajustement
$Q1i2$	séquence d'intervalles	Quant1 (directe)	L2

TAB. 7.3 : *Quantifié vs. Non quantifié : Systèmes en compétition.*



(a)



(b)

figure 7.5 : *Comparaison des performances de deux des moteurs de comparaison proposés et leurs versions quantifiées. (a) rappel ; (b) précision.*

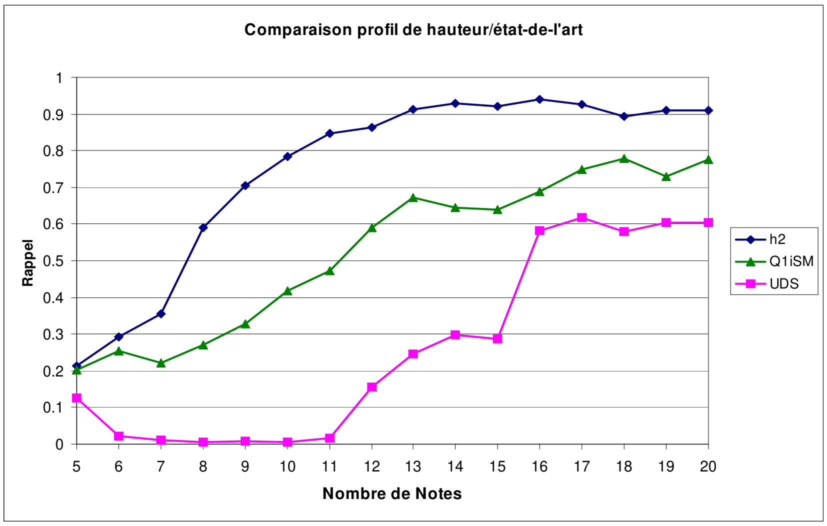
La figure 7.5 confirme la diminution des performances due à la quantification de la requête. La quantification pénalise d'avantage le moteur fondé sur les profils de hauteurs que celui fondé sur les séquences d'intervalles. On peut penser que l'approche du second, comparant les intervalles deux à deux, est plus robuste à la quantification, processus également appliqué à l'échelle de l'intervalle. L'approche du premier, reconstruisant un profil de hauteurs après quantification sur les intervalles, serait plus sensible à ce genre de traitement.

7.3.4 Système choisi vs. Etat-de-l'art

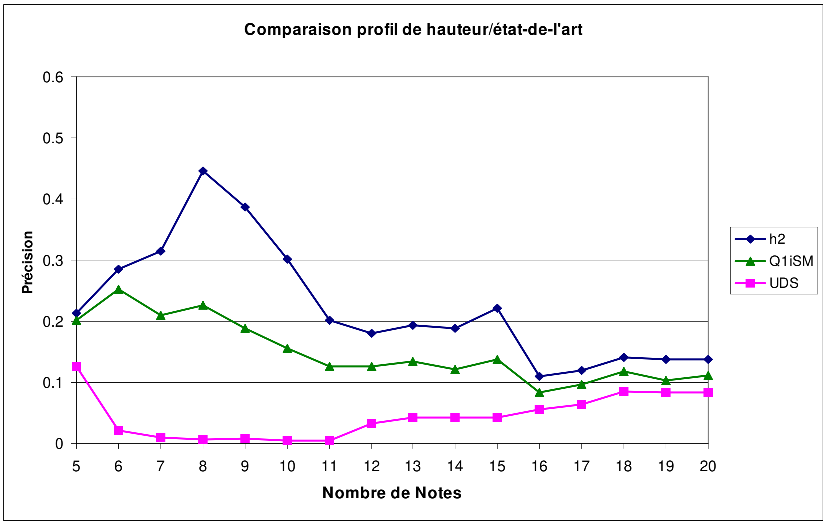
Nous venons de montrer que l'absence de quantification permettait une amélioration des performances. Par ailleurs, la figure 7.3 montre qu'un système hybride donnerait d'excellents résultats. Fondé sur la taille des requêtes injectées, il utiliserait $h2$ pour les requêtes courtes et moyennes, et commuterait sur $i2$ pour les requêtes de taille de supérieure à 15 notes.

Le but de ce chapitre étant de déterminer le meilleur des systèmes proposés, nous allons nous prononcer sur un système unique. Donnant les meilleurs résultats (rappel/précision et rapidité) pour la majorité des requêtes injectées, nous désignons $h2$ comme étant le meilleur système proposé.

La figure 7.6 illustre ses performances, accompagnées de celles du système représentant l'état-de-l'art, et de celles du système UDS. Leurs caractéristiques sont rassemblées dans le tableau 7.4.



(a)



(b)

figure 7.6 : *Comparaison des performances du système proposé avec celles de l'état-de-l'art. (a) rappel ; (b) précision.*

Nom	Type de descripteur	Quantification	Mesure de similarité
$h2$	profil de hauteurs	non	L2 avec ajustement
$Q1iSM$	séquence d'intervalles	Quant1 (directe)	string matching
UDS	séquence UDS	Quant1 (directe)	string matching

TAB. 7.4 : *Système proposé vs. Etat-de-l'art : Systèmes en compétition.*

Il est donc évident que, dans notre contexte d'étude, les performances du système que nous proposons sont nettement supérieures à l'état-de-l'art. Nous espérons que les travaux futurs concernant par exemple la prise en compte des insertions/omissions de notes permettront de conserver un tel niveau de performance.

Enfin, puisque la rapidité est un élément fondamental des systèmes de recherche d'information, nous allons nous intéresser au temps de traitement associé au système choisi.

7.3.5 Rapidité du système choisi

La qualité d'une réponse dépend aussi du temps qu'elle met à arriver. Nous avons voulu observer le temps d'attente moyen en fonction de la taille de la base de données consultée. Pour chaque taille de base considérée, nous avons mesuré le temps de traitement de nos 500 requêtes. Nous en avons tiré le temps *moyen* que prend le système choisi pour traiter *une* requête, en fonction de la base de données. Ces durées sont présentées tableau 7.5.

Taille du corpus	Temps de recherche moyen pour une requête
250	0.27 secondes
500	0.32 secondes
1.000	0.44 secondes
5.000	1.49 secondes
10.000	2.42 secondes
19.282	4.18 secondes

TAB. 7.5 : *Temps de recherche moyen pour le système $h2$, sur processeur PentiumIII - 933MHz.*

Pour les bases de données utilisées, le temps d'attente est tout à fait acceptable. La complexité de notre système n'est donc pas rédhibitoire. Notons que les durées présentées donnent un ordre de grandeur des performances que l'on peut attendre d'un algorithme non optimisé. Un travail sur ce point ainsi que sur l'indexation des descripteurs[3] pourrait améliorer sensiblement la vitesse du système. Ce gain serait appréciable car il permettrait de s'adresser sans lourdeur à des volumes de données plus importants.

[3]Pour le moment, un seul fichier recueille tous les descripteurs mélodiques (redondances comprises), et ceux-ci sont parcourus linéairement lors de la recherche.

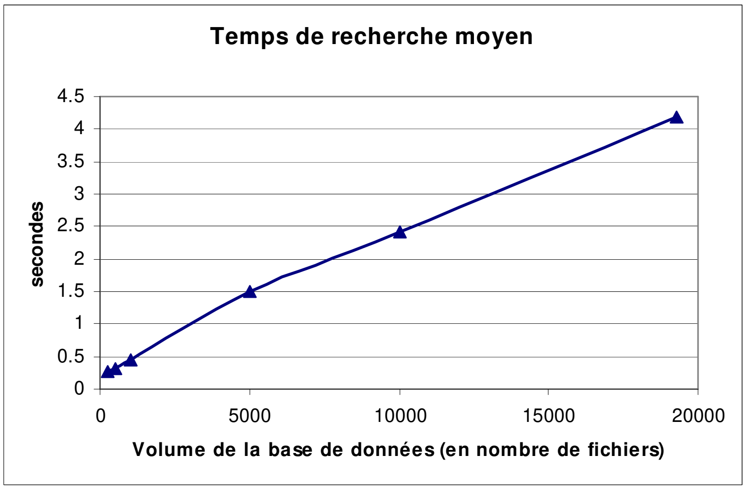


figure 7.7 : *Temps de recherche moyen (sur processeur PentiumIII - 933MHz) en fonction du volume de données consulté.*

La figure 7.7 reprend les valeurs du tableau 7.5. On peut y constater l'augmentation quasi-linéaire du temps de recherche avec la taille de la base de données. Avec la plus volumineuse d'entre elles, la recherche d'une mélodie prend environ 4 secondes. Notons que ce temps serait divisé par deux sur une machine plus récente (type PentiumIV - 2GHz).

7.4 Conclusion

Afin de parvenir à une comparaison des moteurs proposés au Chapitre 5, nous venons de procéder à une évaluation de la qualité de systèmes *complets*. Nous avons tout d'abord considéré des critères concernant la base de données, la soumission d'une requête, et la présentation des résultats. Ces critères nous ont permis d'aborder les éléments communs à tous les systèmes évalués, en particulier l'interface logicielle que nous avons développé.

Concernant la spécificité des configurations testées (i.e. leur moteur de comparaison), les taux de rappel et de précision ainsi qu'un critère de rapidité nous ont permis de qualifier objectivement les performances de recherche. Nous avons ainsi pu désigner $h2$ comme le meilleur des quatre moteurs de comparaison proposés. Fondé sur les profils de hauteurs non tempérées et une distance L2, ce moteur assure un taux de rappel supérieur à 90% dès 13 notes, pour un temps de recherche moyen de 4.18 secondes sur une base de données contenant près de 40 millions de notes (sur processeur PentiumIII - 933MHz).

Idéalement, un système *hybride*, commutant entre $h2$ et $i2$ selon la taille de la requête soumise, donnerait les meilleurs résultats. En effet, fondé sur les séquences d'intervalles, $i2$ est plus performant pour les requêtes longues. Surpassant $h2$ à partir de 16 notes, le moteur $i2$ atteint les 100% de rappel pour 19 notes.

Le fait que deux familles de moteurs se partagent les meilleures performances montre que l'indépendance au ton, nécessaire à la comparaison mélodique, peut être assurée tant par le descripteur mélodique (utilisation des intervalles) que par la mesure de similarité (minimisation de la distance entre profils de hauteurs). Cette conclusion ouvre le champ des solutions fondées sur un *appariement global*, du type profil de hauteurs, qui se démarque de l'approche fragmentée qu'implique l'utilisation des intervalles.

Par ailleurs, la comparaison de trois méthodes de quantification nous a montré que la méthode classique (i.e. quantification directe sur intervalles) était à la fois simple et efficace. Notre méthode, fondée sur l'extraction automatique de la tonalité, s'avère légèrement meilleure sur les requêtes courtes (au plus 8 notes). Nous avons utilisé la méthode classique (globalement meilleure) pour montrer qu'une quantification de l'information fréquentielle diminuait effectivement les performances de recherche.

Dans un contexte prenant uniquement en compte les erreurs de transposition (elles sont systématiques), les critères de rappel et de précision nous ont permis de démontrer la supériorité de notre approche par rapport à l'état-de-l'art. L'absence de quantification de la requête couplée à l'utilisation d'une distance constitue donc une voie d'amélioration sérieuse pour les systèmes de recherche de documents musicaux par chantonnement.

La prudence témoignée lors de la qualification des moteurs de comparaison du Chapitre 6 nous a poussé à juger ces derniers via l'évaluation de la qualité de systèmes de recherche complets. En effet, le premier jugement obtenu s'appuyait sur des tests subjectifs concernant la similarité mélodique qui investissaient des phénomènes simples et manquaient de données exploitables. Or, les résultats obtenus par ces tests subjectifs se voient confirmés par les résultats de ce chapitre. Il s'agit de la supériorité des moteurs fondés sur les profils de hauteurs pour les requêtes courtes et moyennes, et les meilleurs résultats (au sein d'une même famille) des mesures de similarité fondées sur la distance L1. Le type de démarche proposé au Chapitre 6 (i.e. définition de critères objectifs de qualité pour les moteurs de comparaison) est donc encouragé par les résultats de ce chapitre.

Les tests que nous avons effectués dans ce chapitre sont proches de l'utilisation réelle de systèmes, ils renseignent d'avantage sur les performances attendues à l'usage. En contrepartie, ils nécessitent un important corpus de requêtes dont les mélodies visées doivent être identifiées (afin de déterminer les réponses pertinentes nécessaires aux critères de rappel et de précision). Pour notre corpus de requêtes, ce lourd travail d'identification avait été effectué dans le Chapitre 4, nous avons donc pu en bénéficier pour ce chapitre.

Dans le chapitre suivant, nous proposons une méthode pour disposer plus facilement de requêtes tests qui, à la différence des portions directement extraites de la base de données, assurent une stimulation réaliste des systèmes évalués.

Chapitre 8

Stimulations Artificielles pour la Qualification Objective de SR Mélodiques

8.1 Introduction

L'imprécision des hauteurs est systématiquement présente dans les mélodies chantonnées. Phénomène non négligeable (cf. Chapitre 4), elle doit être prise en compte pour la qualification de systèmes de recherche musicale par chantonnement. Or, comme nous l'avons vu en 3.6, les requêtes-tests généralement utilisées pour stimuler les systèmes sont directement issues de la base de données. Elle sont donc parfaitement tempérées et ne comportent pas ce type d'imprécisions.

Une alternative consiste à utiliser des requêtes produites par des sujets. Cependant, comme nous l'avons vu, cette voie est lourde à mettre en œuvre. En effet, il est fastidieux de recueillir des requêtes-tests dont les propriétés soient maîtrisées (mélodie visée connue, nombre de notes suffisant, présence d'erreurs).

Dans ce chapitre, nous proposons une modélisation de l'imprécision fréquentielle, qui nous permettra de synthétiser des requêtes artificielles réalistes à partir de motifs mélodiques issus de la base de données.

Nous commencerons par *modéliser* le phénomène d'imprécision en fréquence observé sur les requêtes étudiées au Chapitre 4, ce qui nous permettra ensuite de *synthétiser* nos requêtes-tests. Celles-ci serviront à *évaluer* une nouvelle fois la qualité de cinq des systèmes qualifiés au chapitre 7. Nous pourrons ainsi *comparer* les performances obtenues avec celles issues de la stimulation par requêtes réelles.

8.2 Modélisation de l'imprécision en fréquence

Au Chapitre 4, nous avons vu que l'imprécision fréquentielle dépendait des intervalles visés. Seulement, au delà d'une amplitude de 5 demi-tons, le comportement des erreurs commises reste mal connu. Dans notre modèle, nous pourrions considérer que les erreurs issues des intervalles d'amplitude supérieure ou égale à 6 se comportent comme celles des intervalles d'amplitude 5. En effet, parmi eux, ceux correspondant aux amplitudes 6, 8, 9, 10, 11 sont fort peu représentés au sein de la base de données. Etant moins sollicités, l'erreur de modélisation ne serait sans doute pas dramatique.

Cependant, nous préférons regrouper toutes les données disponibles. Ainsi, les 5345 valeurs d'erreur interviendront dans le modèle qui sera, par ailleurs, plus facile à utiliser[1]. Nous verrons si ce choix, revenant à considérer que l'imprécision est *indépendante* de l'intervalle visé, donne satisfaction pour la stimulation de systèmes lors de l'évaluation de leurs performances.

De la même manière que dans le chapitre 4, nous regroupons les erreurs de signe opposé. Le modèle choisi sera donc symétrique.

Concernant le choix de ce dernier, nous avons, dans un premier temps, étudié une forme de gaussienne simple, dont la formule est donnée ci-dessous.

$$P_x(x) = \frac{1}{\sqrt{2\pi}.\sigma_x} \cdot e^{-\frac{(x-m_x)^2}{2.\sigma_x^2}} \tag{8.1}$$

Nous en avons tiré deux modèles distincts. La gaussienne de la figure 8.1(a) possède le même écart-type que les données modélisées, soit 7.16e-1, ainsi qu'une moyenne identique, soit 5.31e-18. Celle de la figure 8.1(b) a été ajustée à la densité de probabilité de l'imprécision selon le critère des moindres carrés. Son écart-type vaut 4.27e-1, et sa moyenne 3.96e-9.

Nous pouvons voir que la première sur-représente les imprécisions de "forte" amplitude, et que la seconde, au contraire, les sous-représente.

Afin d'affiner la représentation, nous avons considéré un modèle plus complexe : la gaussienne généralisée. Sa formule est donnée ci-dessous.

$$P_x(x) = G_g(x, \gamma) = a.e^{-|b.(x-c)|^\gamma} \tag{8.2}$$

avec

$$a = \frac{b.\gamma}{2.\Gamma(1/\gamma)} \quad \text{et} \quad b = \frac{1}{\sigma_x}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}$$

où $\Gamma(.)$ est la fonction Gamma, soit

$$\Gamma(n) = \int_0^{+\infty} e^{-x}.x^{n+1}dx \tag{8.3}$$

Le paramètre γ, appelé *exposant de la gaussienne généralisée*, détermine l'allure de la densité de probabilité en la rendant plus ou moins saillante.

Comme pour la deuxième gaussienne, les paramètres de la gaussienne généralisée ont été déterminés grâce au critère des moindres carrés[2]. Les paramètres obtenus sont les suivants :

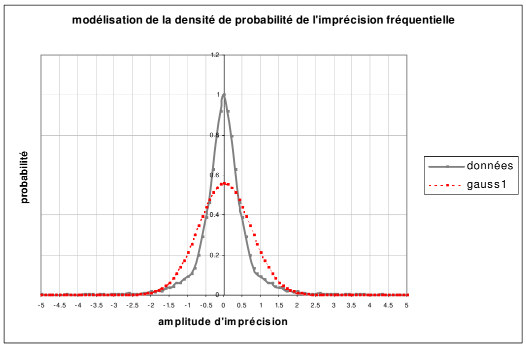
- a = 1.06
- b = 1.98

[1]Nous n'aurons pas à adapter le traitement à l'amplitude des intervalles considérés.

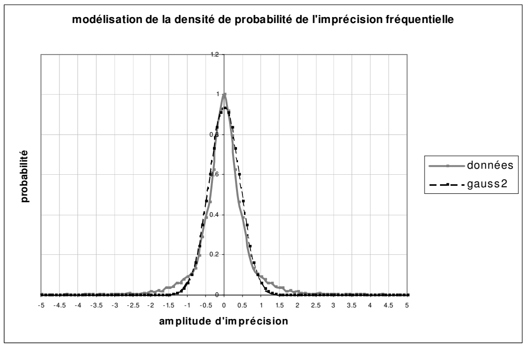
[2]La racine carrée de la somme du carré des différences est égale à 0.35 pour la gaussienne de la figure 8.1(b) et 0.15 pour la gaussienne généralisée, figure 8.1(c)

– c = 0.002
– $\gamma = 1.23$

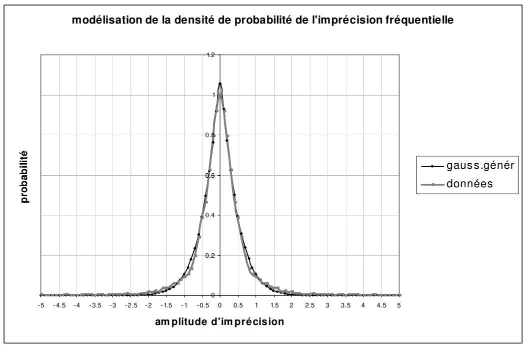
Le modèle finalement choisi est donc celui de la figure 8.1(c) qui représente le plus fidèlement la densité de probabilité de l'imprécision fréquentielle.



(a)



(b)



(c)

figure 8.1 : *Illustration de trois modélisations de l'imprécision fréquentielle, (a) et (b) : par des gaussiennes, (c) : par une gaussienne généralisée.*

Notre modèle étant déterminé, nous allons maintenant l'utiliser pour générer des requêtes aux imprécisions fréquentielles artificielles.

8.3 Qualification objective de systèmes par requêtes-tests artificielles

Le modèle que nous venons de définir nous permet de transformer une mélodie parfaitement tempérée en une mélodie comportant des imprécisions fréquentielles. Concernant la mise en forme du bruit perturbateur à partir de la densité de probabilité 8.2, le lecteur pourra se référer à [PTVF].

Pour vérifier le caractère réaliste de la perturbation, nous allons renouveler les tests du chapitre précédent, en substituant un corpus de requêtes artificiellement bruitées au corpus de requêtes réelles.

Le corpus servant pour la stimulation des systèmes comportera donc 500 requêtes-tests artificielles dont les motifs visés sont strictement identiques à ceux du corpus de requêtes réelles. Ainsi, les différences de résultats témoigneront uniquement de la qualité de la perturbation injectée, puisque ni le choix des mélodies recherchées, ni la taille des motifs utilisés seront différents.

Dans nos tests, les motifs utilisés pour servir de base à la constitution des requêtes artificielles sont donc fixés par le désir de comparer nos résultats à ceux obtenus précédemment avec des requêtes réelles.

Cependant, dans l'optique de tests isolés (i.e. sans comparaison prévue avec des résultats issus d'un corpus de requêtes réelles), les motifs mélodiques pourraient être aléatoirement extraits de la base de données. En nombre suffisant, ils témoigneraient de la variété des mélodies indexées.

Ce dernier point concerne la notion d'originalité d'une mélodie qui constitue un facteur important dans l'indexation mélodique. En effet, plus un motif est original (dans le sens où il ne ressemble pas aux autres), plus la requête le visant pourra être courte et/ou comporter des erreurs sans que cela n'empêche un système de le retrouver. Par conséquent, les tests réalisés doivent autant que possible utiliser une stimulation représentative des mélodies disponibles (ou recherchées), afin de ne pas trop orienter les résultats obtenus.

Le processus de test est similaire à celui emprunté pour les requêtes réelles. Les motifs considérés permettent, dans un premier temps, d'identifier les musiques pertinentes qui serviront pour les critères de rappel. Ils sont ensuite bruités conformément au modèle proposé et injectés dans les systèmes testés. La comparaison des réponses à chacune des requêtes artificielles avec les documents pertinents attendus permet de calculer le taux de rappel.

La robustesse du système vis-à-vis (d'un modèle) de l'imprécision en fréquence des requêtes chantonnées peut ainsi être qualifiée sans avoir à traiter manuellement un corpus de requêtes réelles.

8.3.1 Systèmes fondés sur les profils de hauteurs

La figure 8.2 présente les résultats obtenus pour le test des systèmes fondés sur les profils de hauteurs.

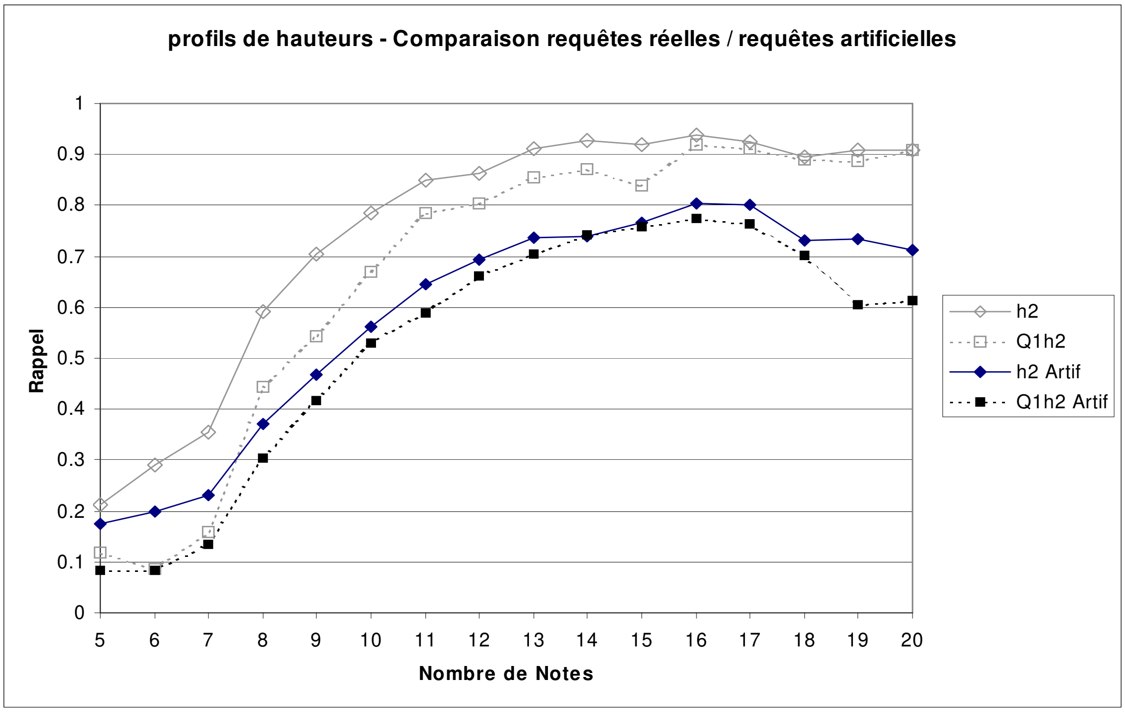


figure 8.2 : *Systèmes fondés sur les profils de hauteurs. En foncé : performances de rappel sur requêtes artificielles ; en clair : performances avec requêtes réelles.*

Pour cette famille de systèmes, la hiérarchie est conservée (supériorité de l'approche non quantifiée) mais les valeurs de rappel sont sous-estimées.

Pour les systèmes fondés sur les profils de hauteurs, nos requêtes artificielles ne permettent donc pas d'estimer le comportement de systèmes en situation réelle, mais informent néanmoins sur la qualité relative des configurations testées.

Nous allons voir s'il en est de même avec les systèmes fondés sur les séquences d'intervalles.

8.3.2 Systèmes fondés sur les séquences d'intervalles

La figure 8.3 présente les résultats obtenus pour le test des systèmes fondés sur les séquences d'intervalles.

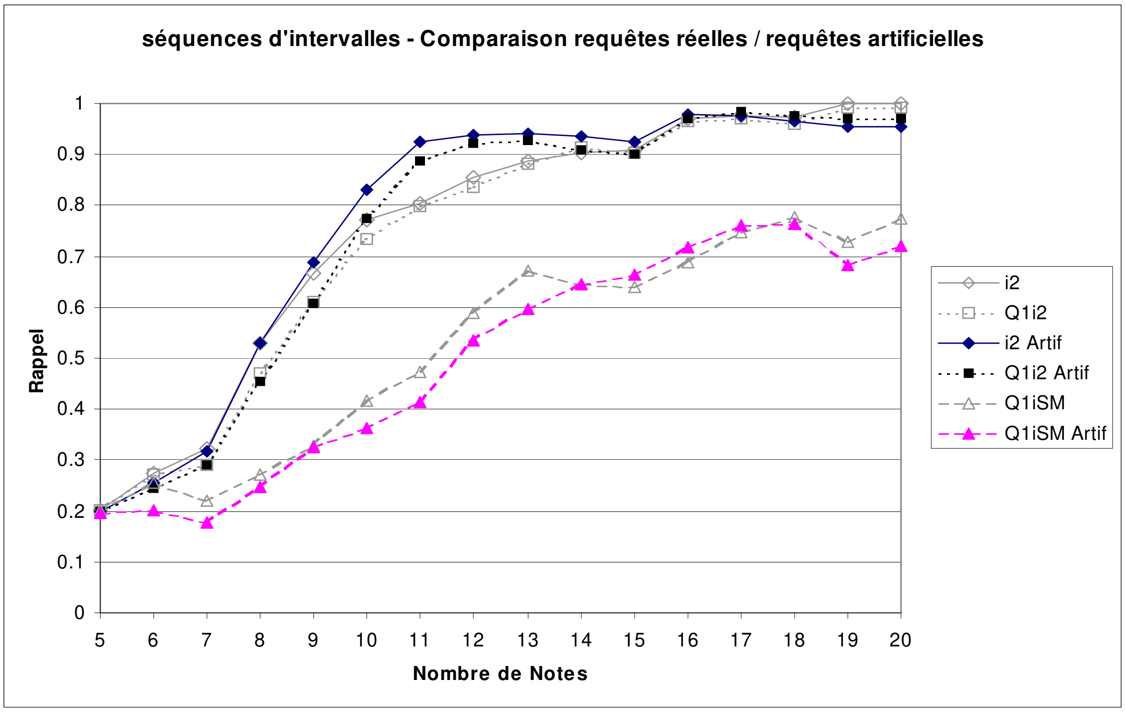


figure 8.3 : *Systèmes fondés sur les séquences d'intervalles. En foncé : performances de rappel sur requêtes artificielles ; en clair : performances avec requêtes réelles.*

Pour cette famille de systèmes, nous constatons que les performances de rappel obtenues sur requêtes artificielles sont proches de celles obtenues sur requêtes réelles. Notre modèle de bruit permet donc d'une part d'établir une hiérarchie correcte entre les systèmes testés, et d'autre part d'estimer le comportement de ces derniers en situation réelle.

8.3.3 Jugement sur la qualité des requêtes-tests artificielles

Nos requêtes artificielles n'ont pas le même impact sur les deux familles de moteurs de comparaison proposés. La stimulation permise par notre outil ne peut donc se substituer purement et simplement à un corpus de requêtes réelles.

La sous-estimation des performances des systèmes fondés sur les profils de hauteurs semble due à la simplicité du modèle proposé. Reposant sur une indépendance des imprécisions injectées vis-à-vis de leur contexte (amplitudes des intervalles visés, imprécisions voisines), notre modèle pénalise les moteurs procédant à une comparaison "globale" des mélodies.

Les moteurs fondés sur les profils de hauteurs sont donc désavantagés par un ajustement des motifs sollicitant *l'ensemble* de l'information contenue dans les descripteurs. Au contraire, les moteurs fondés sur les séquences d'intervalles sont moins sensibles aux limitations du modèle. En effet, leur comparaison mélodique, somme de contributions indépendantes, est moins affectée par la simplicité de la perturbation injectée.

Notre modélisation des imprécisions est donc plus adaptée à l'un des deux types de systèmes testés. Une comparaison de moteurs de familles différentes serait donc faussée. Un modèle traitant de manière plus équitable les différents systèmes serait donc appréciable. Notre outil permet néanmoins de comparer de manière réaliste plusieurs configurations d'une même famille de moteurs de comparaison, sachant que pour la famille "séquences d'intervalles", les résultats obtenus constituent une bonne estimation des performances réelles.

8.4 Conclusion

La stimulation de systèmes pour la qualification de leurs performances pose certains problèmes. D'un côté, l'utilisation de portions mélodiques directement extraites de la base de données est irréaliste puisque l'imprécision fréquentielle, phénomène à la fois systématique et non négligeable, n'y est pas représentée. D'un autre côté, la collecte de requêtes réelles (i.e. chantonnées par des sujets) implique un lourd travail de dépouillement.

Dans ce chapitre, nous avons proposé une alternative pour la stimulation de systèmes lors de leur qualification objective. Une modélisation simple des imprécisions en fréquence observées sur les requêtes chantonnées (cf. Chapitre 4) nous a permis de générer des requêtes aux imperfections artificielles. Celles-ci permettent d'éviter la lourde mise en place d'un corpus de requêtes réelles tout en conservant une part de réalisme dans la stimulation de systèmes.

La comparaison des résultats issus de la stimulation par requêtes artificielles avec ceux issus de la stimulation par requêtes réelles (cf. Chapitre 7) a montré l'intérêt du modèle proposé, tout en révélant ses limites.

Efficaces pour le classement de différentes configurations de systèmes appartenant à une même famille (typiquement *profils de hauteurs* ou *séquences d'intervalles*), nos requêtes-tests artificielles s'avèrent très utiles pour le développement d'un système, dont différentes variantes seraient à départager. Par contre, la stimulation proposée ne permet pas la comparaison de systèmes de familles différentes. En effet, les hypothèses simplificatrices sur lesquelles s'appuie notre modèle d'imprécision sont moins adaptées à l'approche "globale" des profils de hauteurs, qu'à celle, plus "locale" des séquences d'intervalles. Cela se traduit par une sous-estimation des performances des systèmes fondés sur les profils de hauteurs, empêchant un traitement équitable des deux familles. Les systèmes fondés sur les séquences d'intervalles voient leurs performances correctement estimées. La solution que nous proposons constitue donc, pour ces systèmes, une alternative de qualité à l'utilisation de requêtes réelles.

Notre outil offre ainsi la possibilité de constituer facilement un corpus de stimulation conséquent et varié, pour la qualification objective de systèmes de recherche par chantonnement. Par sa facilité de mise en œuvre, il permet de disposer de requêtes représentatives de la variété des mélodies indexées, et de représenter chaque motif par un grand nombre de réalisations. Cela est appréciable car un trop petit nombre de requêtes-tests introduit un biais dans la qualification de systèmes. Notre outil autorisant des requêtes de longueur quelconque, l'étude du comportement des systèmes en fonction du nombre de notes soumis n'est pas limitée. Enfin, le fait de partir de portions mélodiques extraites de la base de données évite la phase d'identification des mélodies visées, ce qui permet d'obtenir facilement les réponses attendues pour une requête donnée (i.e. les documents pertinents nécessaires au critère de précision/rappel).

L'approche proposée dans ce chapitre s'avère donc tout à fait pertinente. L'utilisation d'un modèle plus élaboré (sans l'hypothèse d'indépendance des imprécisions injectées aux intervalles visés et/ou avec la prise en compte de relations entre imprécisions voisines) permettrait de s'approcher d'une méthode universelle pour la stimulation de systèmes lors de la qualification objective de leur performances.

Chapitre 9

Conclusion Générale et Perspectives

Le manque de maturité de l'indexation mélodique se traduit par une méconnaissance de domaines essentiels. En particulier, la notion de similarité mélodique et, par conséquent, les moyens d'en témoigner automatiquement restent à approfondir. La question de la requête est également cruciale. Constituant l'information de départ d'une recherche, elle influence fortement la comparaison mélodique, et donc la similarité témoignée. En particulier, la requête chantonnée n'a pas reçu une grande attention, malgré une nature qui la destine aux systèmes de recherche grand public.

Dans cette thèse, nous avons contribué à l'amélioration des connaissances sur ces deux points. Concernant la requête chantonnée, nous nous sommes penchés sur l'imprécision fréquentielle des notes provenant de 500 requêtes. Ainsi, nous avons pu révéler des tendances à la compression/extension de certains intervalles. Nous avons également observé une imprécision dépendant du rang au sein de la mélodie, ainsi que de l'amplitude de l'intervalle visé.

Nous avons également réalisé des tests subjectifs concernant la similarité mélodique. En faisant identifier des mélodies dégradées à des sujets, nous avons établi des correspondances entre les dissimilarités causées par trois types d'erreurs fréquentielles. Ces correspondances nous ont permis de juger objectivement de la qualité des moteurs de comparaison proposés dans cette thèse.

La conception de ces moteurs (i.e. la définition de descripteurs associés à une mesure de similarité) a été détaillée en prenant soin, à chaque étape, d'envisager les différentes voies possibles. Désirant réaliser un système de recherche dédié au grand public, nous nous sommes prononcés pour les voies offrant une convivialité d'utilisation maximum. Ainsi, nos moteurs de comparaison sont adaptés à la requête chantonnée, sans qu'aucune information complémentaire à la mélodie soumise ne soit requise.

L'intérêt porté à la nature de la requête chantonnée nous a conduits à refuser la quantification de ses hauteurs lors de la description. Nos moteurs se démarquent ainsi de l'état de l'art en évitant l'inéquité de jugement qui en découle. Ce choix se traduit par une amélioration des performances de recherche.

Ces performances ont été évaluées via la qualification de systèmes de recherche complets. Le meilleur des quatre moteurs de comparaison proposés (fondé sur les profils de hauteurs non tempérées et une distance L2) assure un taux de rappel supérieur à 90% dès 13 notes. Ce moteur offre un temps de recherche réaliste, puisque l'attente moyenne est de l'ordre de 4 secondes pour une base de données contenant près de 40 millions de notes[1].

[1] Ces performances ont été mesurées sur un processeur PentiumIII de fréquence 933MHz.

La qualification de systèmes de recherche complets renseigne d'avantage sur les performances attendues à l'usage que la qualification directe des moteurs de comparaison. En contrepartie, elle nécessite un important corpus de requêtes dont les mélodies visées doivent être identifiées (afin de déterminer les réponses pertinentes nécessaires aux critères de rappel et de précision). Pour que des évaluations futures puissent éviter la lourde mise en place d'un corpus de requêtes réelles, nous avons proposé un outil permettant de générer des requêtes aux imperfections artificielles. Cet outil permet de constituer facilement un corpus de stimulation conséquent et varié. S'appuyant sur une modélisation des imprécisions fréquentielles que nous avons étudiées, les requêtes constituées sont plus réalistes que les portions directement extraites de la base de données, souvent utilisées dans les publications disponibles.

Au cours de cette thèse, nous avons donc conçu un système permettant la recherche de documents musicaux par chantonnement. Des perspectives d'amélioration pourraient découler d'études complémentaires, notamment sur la similarité mélodique et sur la requête chantonnée. Par exemple, la connaissance de la stabilité rythmique des requêtes pourrait permettre une meilleure description de l'aspect temporel des mélodies. Celle-ci permettrait la gestion de phénomènes tels que insertions/omissions de notes, mais aussi des phénomènes plus complexes, tels que fragmentations/consolidations de notes (si ceux-ci s'avéraient incontournables).

La convivialité des systèmes actuels de recherche par chantonnement devrait également être améliorée. En effet, la transcription automatique des requêtes chantonnées impose à l'utilisateur de prononcer un "ta" pour chaque note. Cette contrainte classique a pour but de faciliter la segmentation en notes du flux acoustique soumis. La connaissance des notes d'une mélodie permet le calcul des intervalles qui, à leur tour, permettent une description mélodique indépendance du ton emprunté. Or, les moteurs de comparaison que nous avons développés ont montré leur capacité à gérer cette indispensable invariance par la mesure de similarité (les descripteurs associés étant fondés, non pas sur les intervalles, mais sur les hauteurs).

Une description mélodique issue d'une segmentation fondée sur un échantillonnage temporel régulier (et non sur les instants d'apparition des notes) serait plus proche du matériau fourni par l'utilisateur[2], tout en soulageant ce dernier du "ta" imposé à chaque note. La requête serait ainsi libérée d'une contrainte génante.

La conception de systèmes de recherche de documents musicaux se heurte aux même limitations que l'indexation audio en général. Dans l'état actuel des recherches, les systèmes assurant les meilleures performances nécessitent une importante intervention humaine. Pour que celle-ci diminue, les recherches doivent d'avantage se tourner vers l'étude de la perception humaine afin d'en tirer la part réalisable par la machine.

Ce travail de recherche a été valorisé par la publication d'un article dans la revue Annales des Télécommunications [CP00], ainsi qu'une contribution dans la conférence internationale DAFx01 [CPA01]. Par ailleurs, une démonstration de notre application a influencé le choix des descripteurs mélodiques normalisés par le standard international ISO/MPEG-7 [Qua00]. Une démarche

[2]en assurant le suivi de *glissendo*, par exemple.

de vulgarisation a également été adoptée, concrétisée par une participation au festival Agora2001 organisé par l'IRCAM [3], ainsi que par une présentation de nos travaux dans le magazine "L'ordinateur individuel" [Ord01]. Enfin, une équipe de France Télécom R&D travaille actuellement à la mise en ligne du système de recherche développé.

[3]http ://www.ircam.fr/departements/creation/agora/jpo/ligne.html

Bibliographie

[AN89] J. Astola and Y. Neuvo.
Optimal median type filters for exponential noise distribution.
Signal Processing, 17 :95–104, 1989.

[AT00] M. Alghoniemy and A.H. Tewfik.
User-defined music sequence retrieval.
In *Proc. ACM Multimedia*, pages 356–358, 2000.

[Bag94] Paul Christopher Bagshaw.
Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching.
PhD thesis, University of Edinburgh, 1994.

[BD00] S. Blackburn and D. DeRoure.
Musical part classification in content based systems.
In *6th Open Hypermedia Systems Workshop*, pages 66–76, 2000.

[Bee97] Doug Beeferman.
Qdp : Query by pitch dynamics, indexing tonal music by content, 1997.
15-829 Course Project.

[Bla99] Steven Blackburn.
Content based retrieval and navigation of music, 1999.
Mini-Thesis submitted for transfer of registration from Mphil to Ph.D.

[BNMW+99] D. Bainbridge, C.G. Nevill-Manning, I.H. Witten, L.A. Smith, and R.J. McNab.
Towards a digital library of popular music.
In *Proc. of the Fourth ACM Conference on Digital Libraries*, pages 161–169, 1999.

[BYP92] R.A. Baeza-Yates and C.H. Perleberg.
Fast and practical approximate string matching.
In *Combinatorial Pattern Matching, 3rd Annual Symp.*, pages 185–192, 1992.

[CC98] James C. C. Chen and Arbee L. P. Chen.
Query by rhythm : An approach for song retrieval in music databases.
In *Proc. of 8th Int. Workshop on Research Issues on Data Engineering*, pages 139–146, 1998.

[CCC+00] Arbee L.P. Chen, Maggie Chang, Jesse Chen, Jia-Lien Hsu, Chih-How Hsu, and Spot Y.S. Hua.
Query bu music segments : an efficient apporach for song retrieval.
In *Proc. IEEE Int. Conf. on Multimedia and Expo (II)*, pages 873–876, 2000.

[CCL96] Ta-Chun Chou, Arbee L.P. Chen, and Chih-Chin Liu.
Music databases : Indexing techniques and implementation.
In *Proc. Int. Workshop on MultiMedia Database Management Systems*, pages 46–53, 1996.

[CP00] Matthieu Carré and Pierrick Philippe.
Indexation audio : un état de l'art.
Annales des Télécommunications, 55(9-10) :507–525, 2000.

[CPA01] Matthieu Carré, Pierrick Philippe, and Christophe Apélian.
New query-by-humming music retrieval system conception and evaluation based on a query nature study.
In *DAFx01, Digital Audio Effects*, pages 227–231, 2001.
URL : http ://www.csis.ul.ie/dafx01/proceedings/papers/carre2.pdf.

[CS98] E.J. Coyle and I. Shmulevich.
A system for machine recognition of music patterns.
IEEE, pages 3597–3600, 1998.

[CV00] Wei Chai and Barry Vercoe.
Using models in music information retrieval systems.
In *Proc. International Symposium on Music Information Retrieval*, 2000.

[Dix00] S. Dixon.
A lightweight multi-agent musical beat tracking system.
In *Proc. 6th Pacific Rim Int. Conf. on Artificial Intelligence*, volume 1886, pages 778–788, 2000.

[DN00] J. Stephen Downie and Michael Nelson.
Evaluation of a simple and effective music information retrieval method.
In *Proc. SIGIR2000*, pages 73–80, 2000.

[Dow78] W.J. Dowling.
Scale and contour : Two components of a theory of memory for melodies.
Psychological review, 85 :341–354, 1978.

[EK00] Antti Eronen and Annsi Klapuri.
Musical instrument recognition using cepstral coefficients and temporal features.
In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 753–756, 2000.

[Foo97] Jonathan Foote.
A similarity measure for automatic audio classification.
In *Proc. AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, 1997.

[Foo99] Jonathan Foote.
An overview of audio information retrieval.
Multimedia Systems, 7(1) :2–10, 1999.

[Foo00] Jonathan Foote.
Automatic audio segmentation using a measure of audio novelty.
In *IEEE International Conference on Multimedia and Expo*, volume 1, pages 452–455, 2000.

[GLCS95] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith.
Query by humming : Musical information retrieval in an audio database.
In *ACM Multimedia*, pages 231–236, 1995.

[Gui96] M. Guillemot, editor.
Dictionnaire de la Musique.
Larousse-Bordas, 1996.
ISBN 2-03-750013-0.

[Hes83] W.H. Hess.
Pitch determination of speech signals.
Algorithms and Devices - Springer-Verlag, 1983.

[HLC01] J.L. Hsu, C.C. Liu, and L.P. Chen.
Discovering nontrivial repeating patterns in music data.
IEEE Transaction on Multimedia, 3(3) :311–325, 2001.

[KCGV00] Youngmoo Kim, Wei Chai, Ricardo Garcia, and Barry Vercoe.
Analysis of a contour-based representation for melody, 2000.

[Kla98] A. Klapuri.
Automatic transcription of music.
Master's thesis, Tampere University of Technology, 1998.

[KMT93] T. Kageyama, K. Mochizuki, and Y. Takashima.
Melody retrieval with humming.
In *Proc. ICMC'93*, pages 349–351, 1993.

[Kru90] C.L. Krumhansl.
Cognitive Foundations of Musical Pitch.
New York Oxford University Press, 1990.

[LC00] Wegin Lee and Arbee L.P. Chen.
Efficient multi-feature index structures for music data retrieval.
In *Proc. of SPIE Conf. on Storage and Retrieval for Media Databases*, pages 177–188, 2000.

[LHC99] Chih-Chin Liu, Jia-Lien Hsu, and Arbee L.P. Chen.
Efficient theme and non-trivial repeating pattern discovering in databases.
In *ICDE'99, Proc. 15th Int.Conf. on Data Engineering*, pages 14–21, 1999.

[LHU98] Kjell Lemström, Atso Haapaniemi, and Esko Ukkonen.
Retrieving music - to index or not to index.
In *Proc. ACM Multimedia Conf.*, pages 64–66, 1998.

[Li00] Stan Z. Li.
Content-based audio classification and retrieval using the nearest feature line method.
IEEE Transactions on Speech and Audio Processing, 8(5), 2000.

[Lin96] A. Lindsay.
Using contour as a mid-level representation of melody.
Master's thesis, Massachusetts Institute of Technology, 1996.

[LL98] Kjell Lemström and Pauli Laine.
Musical information retrieval using musical parameters.
In *Proc. ICMC98*, pages 341–348, 1998.

[LMB97] Elizabeth Lyon, Jon Maslin, and Bob Baker.
Audio and video on-demand for the performing arts : Project patron.
In *Proc. of the 4th UK/Int. Conf. on Electronic Library and Visual Information Research*, pages 177–185, 1997.

[Mar96] K.D. Martin.
A blackboard system for automatic transcription of simple polyphonic music.
Technical report, MIT Media Laboratory, Perceptual Computing Section, 1996.

[Mar01] José M. Martinez.
Overview of the mpeg-7 standard (version 6.0), 2001.
http ://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

[MD90] Mongeau M. and Sankoff D.
Comparison of musical sequences.
Computer and the Humanities, 24 :161–175, 1990.

[MSWH00] Rodger J. McNab, Lloyd A. Smith, Ian H. Witten, and Clare L. Henderson.
Tune retrieval in the multimedia library.
Multimedia Tools and Applications, 10(2/3) :113–132, 2000.

[MYC91] Y. Medan, E. Yair, and D. Chazan.
Super resolution pitch determination of speech signals.
IEEE Transactions on Signal Processing ASSP, 39(1) :40–48, 1991.

[Nol67] P. Noll.
Cepstrum pitch determination.
JASA, 41(2) :293–309, 1967.

[Ord01] L'ordinateur individuel, no. 132, Oct. 2001.

[PHMW98] David Pye, Nicholas J. Hollinghurst, Timothy J. Mills, and Kenneth R. Wood.
Audio-visual segmentation for content-based retrieval.
In *5th International Conference on Spoken Language Processing (ICSLP'98)*, 1998.

[PRF$^+$01] C. Papaodysseus, G. Roussopoulos, D. Fragoulis, T. Panagopoulos, and C. Alexiou.
A new approach to the autimatic recognition of musical recordings.
J. Audio Eng. Soc., 49(1/2) :23–35, 2001.

[PTK98] A. Pikrakis, S. Theodoridis, and D. Kamarotos.
Recognition of isolated musical patterns using discrete observation hidden markov models.
In *IX EUropean SIgnal Processing Conference (EUSIPCO'98)*, volume 4, pages 2357–2360, Sept. 1998.

[PTVF] Press, Teukolsky, Vetterling, and Flannery.
Numerical Recipies in C.
Cambridge University Press, ? ? ? ?
second edition, ISBN 0-0521-43108-5.

[QL01] S. Quackenbush and A. Lindsay.
Overview of mpeg-7 audio.
IEEE Transactions on Circuits and Systems for Video Technology, 11(6) :725–729, 2001.

[Qua00] S. Quackenbush.
Audio subgroup report for the 54th mpeg meeting, 2000.

[Rol99] P.-Y. Rolland.
Discovering patterns in musical sequences.
Journal of New Music Research, 28 :334–350, 1999.

[Sau96] John Saunders.
Real-time discrimination of broadcast speech/music.
In *Proc. ICASSP '96*, pages 993–996, 1996.

[SGM98] Tomonari Sonoda, Masakata Goto, and Yoichi Muraoka.
A www-based melody retrieval system.
In *Proc. ICMC98*, pages 349–352, 1998.

[SJ98] P. Salosaari and K. Jarvelin.
Musir - a retrieval model for music.
Technical Report RN-1998-1, University of Tampere, Department of Information Studies, 1998.

[SM83] G. Salton and M. J. McGill.
Introduction into Modern Information Retrieval.
McGraw-Hill, 1983.

[SS97] E. Scheirer and M. Slaney.
Construction and evaluation of a robust multifeature speech/music discriminator.
In *Proc. ICASSP '97*, pages 1331–1334, Munich, Germany, 1997.

[SW97] Stephen W. Smoliar and Lynn D. Wilcox.
Indexing the content of multimedia documents.
In *Proc. of VISual'97, 2d Int. Conference on Visual Information Systems*, pages 53–60, 1997.

[Tse99] Y.H. Tseng.
Content-based retrieval for music collections.
In *Proc. of the Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 176–182, 1999.

[TYK00] M. Tang, C.L. Yip, and B. Kao.
Selection of melody lines for music databases.
In *Proc. Computer Software and Applications Conf.*, pages 243–248, 2000.

[UZ98] A. Uitdenbogerd and J. Zobel.
Manipulation of music for melodic matching.
In *Proc. of ACM Multimedia Conf.*, pages 235–240, 1998.

[UZ99] A. Uitdenbogerd and J. Zobel.
Melodic matching techniques for large music databases.
In *Proc. of ACM Multimedia Conf.*, pages 57–66, 1999.

[WBKW99] Earling Wold, Thom Blum, Douglas Keislar, and James Wheaton.
Classification, search, and retrieval of audio.
CRC Handbook of Multimedia Computing, 1999.

[ZK98] T. Zhang and C. Kuo.
Content-based classification and retrieval of audio.
In *Proc. of the SPIE - The Int. Soc. For Optical Engineering*, volume 3461, pages 432–443, 1998.

[ZK99] T. Zhang and C. Kuo.
Heuristic approach for generic audio data segmentation and annotation.
In *Proc. of ACM Multimedia*, pages 67–76, 1999.

[ZK01] Tong Zhang and C.-C. Jay Kuo.
Audio content analysis for online audiovisual segmentation and classification.
IEEE Transactions on Speech and Audio Processing, 9(4), 2001.

Annexe A

Analyse de Mélodies Chantonnées

A.1 Précisions sur les sujets et les mélodies

Sujet	Age	Pratique musicale	Pratique vocale
1 (AB)	25 ans	1 an (Guitare)	-
2 (CC)	24 ans	(Flûte)	-
3 (GC)	28 ans	-	0.5 an (Chant)
4 (GF)	26 ans	10 ans (Guitare)	-
5 (JC)	31 ans	-	7 ans (Chant lyrique)
6 (MM)	29 ans	7 ans (Guitare&Solphège)	1.5 ans (Initiation)
7 (MC)	27 ans	10 ans (Guitare)	2.5 ans (Chant lyrique)
8 (PC)	29 ans	-	-
9 (YB)	23 ans	-	-

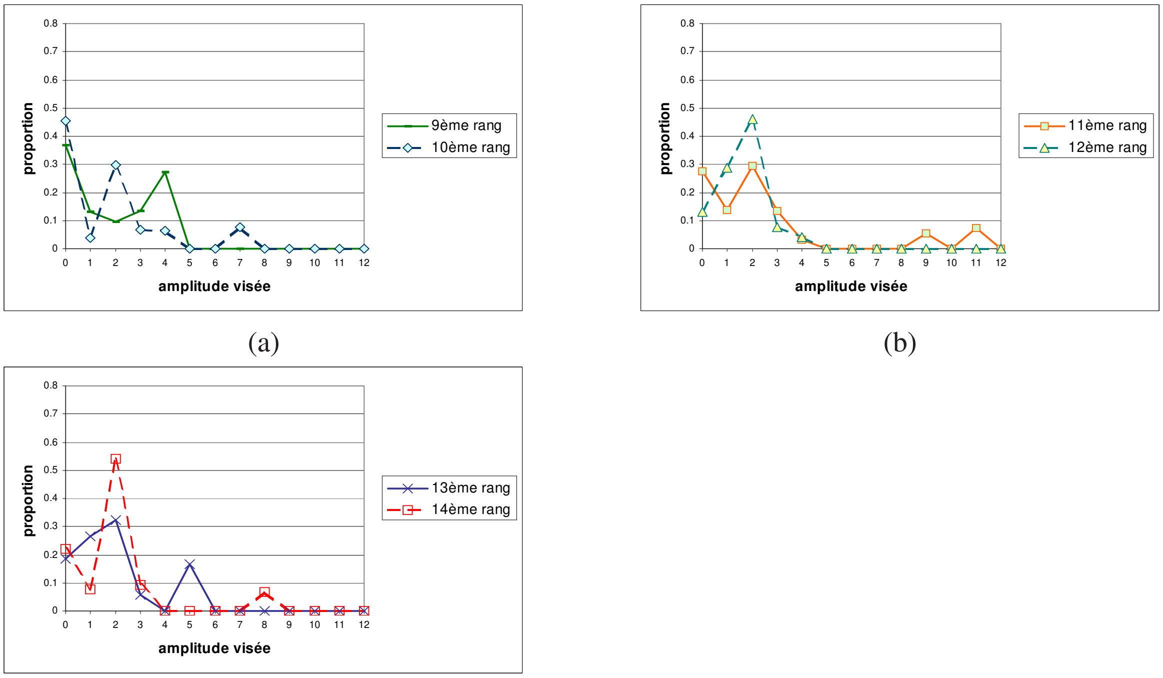
TAB. A.1 : *Détail de la pratique musicale (instrumentale et vocale) des sujets.*

Sujet	Taille de requête moyenne
1 (AB)	13.8 notes (sur 51 requêtes)
2 (CC)	15.7 notes (sur 51 requêtes)
3 (GC)	9.9 notes (sur 24 requêtes)
4 (GF)	13.0 notes (sur 48 requêtes)
5 (JC)	13.7 notes (sur 46 requêtes)
6 (MM)	13.8 notes (sur 37 requêtes)
7 (MC)	14.9 notes (sur 51 requêtes)
8 (PC)	11.6 notes (sur 51 requêtes)
9 (YB)	12.0 notes (sur 36 requêtes)
Tous	13.4 notes (sur 395 requêtes)

TAB. A.2 : *Taille moyenne des requêtes "libre" (i.e. hors "jingles" dont la taille est fixe) pour chacun des sujets.*

A.2 Dépendance de l'imprécision au rang

La figure A.1 présente la répartition des amplitudes d'intervalles pour les 6 rangs non présentés au Chapitre 4.



(a) (b) (c)

figure A.1 : *Répartition des intervalles présents pour chacun des 6 derniers rangs. (a) rangs 9 et 10 ; (b) rangs 11 et 12 ; (c) rangs 13 et 14.*

Annexe B

Extraction Automatique d'une Base Temporelle

B.1 Principe

Le principe de la méthode mise au point consiste à regrouper les durées voisines, et choisir comme base temporelle, le représentant (centroïde) du groupe le plus nombreux (sous une contrainte de compacité).

Voici la description de l'algorithme que nous avons implémenté, il se divise en trois étapes :

1. Initialisation :
   (a) Les valeurs disponibles sont ordonnées par taille croissante ;
   (b) Calcul de moyennes glissantes (taille de fenêtre : 4, pas d'avancement :1) ;
   (c) Constitutions de groupes par détection des ruptures du profil de moyennes (seuil sur les variations = 0.01) ;
   (d) Calcul du centroïde de chaque groupe : moyenne des valeurs regroupées.

2. Proposition de candidats à la base temporelle : processus itératif en trois étapes décrites ci-dessous.
   (a) Découpage de l'espace en fonction des centroïdes précédemment définis : l'espace séparant deux centroïdes contigüs est équitablement partagé entre ces derniers ;
   (b) Calcul des nouveaux centroïdes représentant les populations présentes dans les différentes zones de l'espace découpé,
   (c) Test de convergence sur les centroïdes obtenus :
      – Si les centroïdes obtenus sont différents de ceux obtenus à l'itération précédente, alors retour en 2a ;
      – Sinon, le processus itératif prend fin, les derniers centroïdes obtenus sont déclarés *candidats* à la base temporelle.

3. Sélection du candidat gagnant : Les critères de choix portent sur la taille des populations représentées, ainsi que sur les variances associées. Après le classement des candidats, par ordre décroissant des tailles de populations représentées, le processus itératif suivant prend place :

(a) Sélection du premier candidat (si ex aequo, sélection de celui, parmi les ex aequo, qui possède la plus faible variance associée, et élimination des autres)

(b) Test sur la valeur de la variance :
  – Si le candidat est associé à une variance supérieure à un seuil donné (0.005), il est éliminé (c'est le maillon faible !), et retour en 3a.
  – Sinon, le candidat est élu (c'est le maillon fort !), sa valeur devient la *base temporelle* du motif rythmique traité

Nous allons détailler les deux exemples cités dans 5.2.1. Le premier correspond au cas où l'ajustement est effectué avec succès (cf. points 1 à 4, page 91), le second illustre l'échec de l'ajustement lorsque la requête présente une anticipation de certaines notes (cf. page 94).

B.2 Exemple 1 : "Jésus que ma joie demeure"

Le tableau B.1 présente les hauteurs et instants d'apparition des notes pour les deux motifs mélodiques (requête et référence), avant et après application de la base temporelle extraite. Ces informations sont illustrées figure 5.4.

Rang de la note	Hauteur (# MIDI)		Localisation initiale (en s)		Localisation finale (en tps musical)	
	Requête	Référence	Requête	Référence	Requête	Référence
1	57.5	67	0.00	0.00	1.00	1.00
2	59.1	69	0.28	0.32	1.96	2.95
3	31.0	71	0.54	0.66	2.89	2.98
4	64.5	74	0.82	0.99	3.87	3.98
5	62.5	72	1.10	1.33	4.83	5.02
6	62.6	72	1.38	1.66	5.81	5.99
7	66.7	76	1.66	1.98	6.80	6.97
8	64.9	74	1.97	2.32	7.87	7.98
9	64.7	74	2.27	2.66	8.92	9.01
10	68.8	79	2.55	2.98	9.90	9.98
11	68.4	78	2.82	3.32	10.84	10.99
12	69.2	79	3.12	3.63	11.89	11.94
13	64.4	74	3.42	3.98	12.94	13.00
14	60.9	71	3.72	4.31	13.99	13.98
15	57.7	71	4.01	4.65	15.02	15.01

TAB. B.1 : *Hauteurs et instants d'apparition des notes pour les deux motifs mélodiques (requête et référence) "Jésus que ma joie demeure", avant et après application de la base temporelle extraite.*

Les durées séparant les instants de détection des notes de la requête sont présentées tableau 5.1, page 93. Cette information est injectée dans le programme qui suit les étapes suivantes :

1. L'initialisation ne fournit qu'une seule valeur (0.286). Cela n'est pas surprenant puisque la mélodie chantée par l'utilisateur est composée de notes régulièrement espacées. Le centroïde obtenu représente donc les 14 durées.

2. Aucun découpage de l'espace n'est nécessaire, la première itération du processus de proposition de candidats confirme la valeur issue de l'initialisation. La variance associée a pour valeur 1.62e-4.

3. L’unique candidat posséde une variance inférieure à 0.005, il est donc déclaré base temporelle du motif rythmique traité.

Pour le motif mélodique visé, le processus est similaire (candidat unique dès l’initialisation) : les 14 durées du tableau 5.1 (deuxième colonne) fournissent un unique centroïde de valeur 0.332. Etant associé à une varaince de 1e-4, ce dernier est déclaré base temporelle du motif rythmique traité.

B.3 Exemple 2 : "Mission impossible"

Cet exemple concerne le cas d’anticipation de notes vu en 5.2.1, page 94. Le motif mélodique correspond à un motif de flûte du générique de la série "Mission impossible".

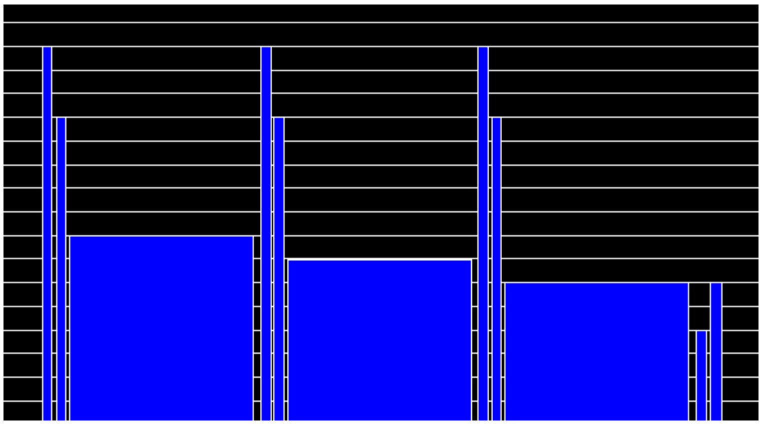


figure B.1 : *Profil temps/fréquence issu de la transcription de la mélodie visée (motif de flûte dans "Mission impossible").*

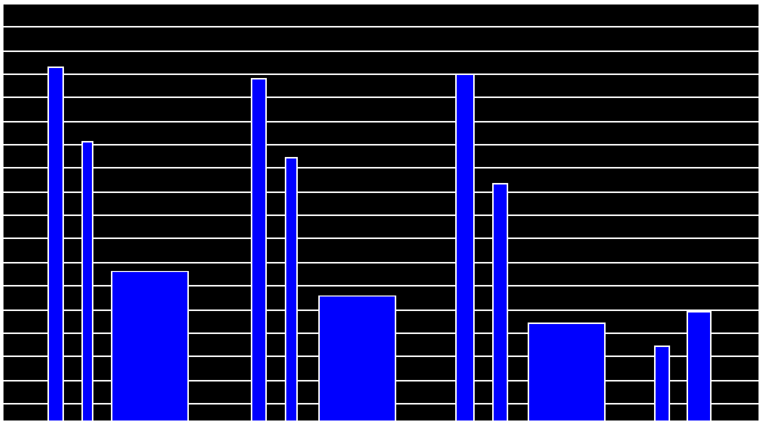


figure B.2 : *Profil temps/fréquence d’une requête réelle visant le motif de flûte dans "Mission impossible".*

Les figures B.1 et B.2 permettent de comparer les profils temps/fréquence du motif visé et de la requête. On peut y voir que les notes les plus aigües [1], sont plus larges dans la requête que dans le motif visé. Cela témoigne du tempo plus lent de la première par rapport au second. Cependant, ces notes sont en même temps moins espacées dans la requête que dans le motif visé, alors que le tempo du premier, plus faible, aurait du induire un éloignement supérieur.

Cette incohérence vis-à-vis du motif visé traduit le phénomène d'anticipation présent dans la requête : les notes de rang 3, 6 et 9 ne sont pas menées à leur terme, les notes suivantes sont placées plus tôt que dans la musique originale recherchée.

Les durées inter-débuts de notes de la requête et du motif visé sont présentées tableau B.2

Rang	Durées inter-débuts de notes	
	Requête	Référence
1	0.225	0.158
2	0.205	0.158
3	0.940	2.216
4	0.225	0.168
5	0.235	0.126
6	0.920	2.227
7	0.250	0.142
8	0.240	0.168
9	0.850	2.221
10	0.220	0.311

TAB. B.2 : *Durées inter-débuts de notes de la requête et du motif visé ("Mission impossible").*

B.3.1 Requête

1. L'initialisation fournit deux valeurs de centroïdes présentées tableau B.3.

Nom	Valeur	Population	Variance
C1	0.229	7	1.84e-4
C2	0.903	3	1.49e-3

TAB. B.3 : *Centroïdes issus de l'initialisation du processus d'extraction d'une base temporelle ("Mission impossible"-Requête).*

2. Le découpage de l'espace induit n'entraîne pas, à la première itération, un changement des valeurs de centroïdes. Les candidats proposés à l'initialisation sont donc confirmés (les variances associées sont présentées tableau B.3).
3. La variance associée au candidat représentant le plus grand nombre de durées étant inférieure à 0.005, la base temporelle extraite est égale à 0.229. Les 7 durées qui lui ont permis d'être élue correspondent aux notes de rang 1, 2, 4, 5, 7, 8 et 10.

B.3.2 Référence

1. L'initialisation fournit trois valeurs de centroïdes présentées tableau B.4.

[1] Il s'agit des notes occupant les rangs 1, 4, et 7.

Rang de la note	Hauteur (# MIDI)		Localisation initiale (en s)		Localisation finale (en tps musical)	
	Requête	Référence	Requête	Référence	Requête	Référence
1	65.3	82	0.00	0.00	1.00	1.00
2	62.1	79	0.23	0.16	1.98	1.90
3	56.6	74	0.43	0.32	2.88	2.80
4	64.8	82	1.37	2.53	6.98	15.39
5	61.4	79	1.60	2.70	7.97	16.34
6	55.6	73	1.83	2.83	8.99	17.06
7	65.0	82	2.75	5.05	13.01	29.71
8	60.4	79	3.00	5.20	14.10	30.52
9	54.4	72	3.24	5.36	15.15	31.47
10	53.4	70	4.09	7.58	18.86	44.09
11	54.9	72	4.31	7.90	19.82	45.86

TAB. B.5 : *Hauteurs et instants d'apparition des notes pour les deux motifs mélodiques (requête et référence), avant et après application de la base temporelle extraite.*

Initialisation			Itérations 1&2			
Nom	Valeur	Population	Nom	Valeur	Population	Variance
C1	0.146	4	C'1	0.176	7	3.24e-3
C2	0.716	4	C'2	2.221	3	2.02e-5
C3	2.224	2				

TAB. B.4 : *Centroïdes issus du processus d'extraction d'une base temporelle ("Mission impossible"-Référence).*

2. Le découpage de l'espace induit entraîne, lors de la première itération, un changement des valeurs de centroïdes. Le centroïde C2 a vu sa population répartie entre C1 et C3, rebaptisés C'1 et C'2. Ces derniers, confirmés par l'itération suivante (cf. tableau B.4), sont proposés comme candidats à la base temporelle.
3. C'1 représente la population la plus large. Sa variance associée étant inférieure à 0.005, il donne sa valeur (0.176) à la base temporelle extraite. Notons que les 7 durées qui ont permis à C'1 d'être élu, correspondent exactement à celles qui ont élu la base temporelle de la requête (à savoir les notes de rang 1, 2, 4, 5, 7, 8 et 10).

Annexe C

Extraction Automatique de la Tonalité

C.1 Principe

Notre méthode s'inspire de la méthode d'extraction de la tonalité de Krumhansl [Kru90]. Cette dernière a établi une hiérarchie concernant l'importance des douze notes de l'échelle tempérée en fonction d'un contexte tonal donné. Le tableau illustre les poids assignés aux différentes notes pour les tonalités do majeur et do mineur. La permutation des éléments de la colonne "Note" permet d'obtenir la hiérarchie pour les 24 tonalités possibles (2 modes, majeur et mineur, pour chacun des douze tons possibles).

Note	Contexte tonal	
	do majeur	do mineur
do	**6.35**	**6.33**
do♯/ré♭	2.23	2.68
ré	3.48	3.52
ré♯/mi♭	2.33	**5.38**
mi	**4.38**	2.60
fa	4.09	3.53
fa♯/sol♭	2.52	2.54
sol	**5.19**	**4.75**
sol♯/la♭	2.39	3.98
la	3.66	2.69
la♯/si♭	2.29	3.34
si	2.88	3.17

TAB. C.1 : *Hiérarchie concernant l'importance des douze notes de la gamme tempérée, pour un ton de* do*, dans les deux modes fondamentaux (majeur et mineur).*

Cette pondération permet de déterminer le niveau de représentation de chacune des 24 tonalités possibles, au sein d'un ensemble de notes.

Si l'on note $p(h)$, le poids correspondant à la hauteur h pour la tonalité étudiée (e.g. $p(h) = 3.48$, pour $h =$ré, dans la tonalité do majeur). Le score de cette tonalité, au regard des N hauteurs $[h_1, h_2, \ldots, h_{N-1}]$ dont les durées respectives sont $[d_1, d_2, \ldots, d_{N-1}]$, est égal à :

$$\sum_{i=0}^{N-1} p(h_i) * d_i \tag{C.1}$$

Grâce à cette mesure, il est donc possible de déterminer la tonalité prépondérante pour un ensemble de notes donné.

Cet algorithme a été repris par Coyle et Shmulevich afin d'être appliqué à la comparaison mélodique [CS98]. Ne disposant, à l'époque, que de cette référence (présentant sommairement le principe de l'algorithme original), nous en avons développé une version simplifiée, que nous avons, par ailleurs, adapté aux hauteurs non tempérées. Cependant, nos observations rejoignant celles de Krumhansl, nous pensons que notre algorithme est représentatif des performances (et des limitations) que l'on peut attendre de l'algorithme original sur des mélodies courtes.

Nous avons défini la "hiérarchie" illustrée tableau C.2. En fait, il s'agit plutôt d'une pondération destinée à ne prendre en compte que les hauteurs les plus stables de la tonalité étudiée. On peut noter une cohérence avec l'algorithme original puisque les hauteurs sélectionnées sont celles qui possèdent les scores les plus élevés dans la hiérarchie de Krumhansl (en gras dans le tableau C.1).

Ces hauteurs étant également les plus fréquentes [Kru90], cette pondération permet de faire ressortir les tonalités les plus plausibles pour le groupe de notes considérées.

Note	Contexte tonal	
	do majeur	do mineur
do	1	1
do♯/ré♭	0	0
ré	0	0
ré♯/mi♭	0	1
mi	1	0
fa	0	0
fa♯/sol♭	0	0
sol	1	1
sol♯/la♭	0	0
la	0	0
la♯/si♭	0	0
si	0	0

TAB. C.2 : *Pondérations utilisées dans notre méthode d'extraction de la tonalité, pour un ton de* do*, dans les deux modes fondamentaux (majeur et mineur).*

C.1.1 Adaptation aux hauteurs non tempérées

Ne disposant pas de valeurs de hauteurs discrètes, il n'est pas possible d'appliquer directement notre pondération. Par conséquent, la sélection des hauteurs s'effectue de manière relative. Pour trouver la tonalité la plus probable, chaque note de la requête est tour à tour prise comme ton potentiel. Dans chaque cas, on recense le nombre de notes situées à moins d'un quart-de-ton des 3 notes les plus stables de la tonalité étudiée.

Il s'agit de la *fondamentale*, la *quinte*, et la *tierce* (mineure ou majeure selon le mode considéré). La *fondamentale* donne son nom au ton, les deux se démarquent de cette dernière par des intervalles précis :

– +3 demi-tons entre *fondamentale* et *tierce mineure*
– +4 demi-tons entre *fondamentale* et *tierce majeure*
– +7 demi-tons entre *fondamentale* et *quinte*

Ces intervalles s'entendent *modulo 12*, ainsi, une note dont la hauteur est à $7 - 12 = -5$ demi-tons de la *fondamentale* est également considérée comme une *quinte*.

Rang	# MIDI	Mode	Score
1	56.11	m	1.86
		M	2.32
2	55.75	m	4.66
3	60.42	m	3.12
		M	2.84
4	60.46	m	3.28
		M	2.84
5	62.43	m	1.92
6	63.49	M	3.74
7	65.73	?	1.84
8	64.21	M	1.88
9	62.67	m	2.4
10	58.51	M	4.46
11	58.63	M	**4.94**
12	58.70	M	4.82

TAB. C.3 : *Extraction de tonalité d'*Amsterdam*, Requête 1 (hauteurs non tempérées).*

Par exemple, pour la tonalité "sol majeur", on a :
– *fondamentale* = sol
– *tierce mineure* = sol+3 demi-tons = la♯
– *tierce majeure* = sol+4 demi-tons = si
– *quinte* = sol+7 demi-tons = ré

Pour chaque ton envisagé[1], un score est calculé en fonction de la répartition des notes avoisinant les pôles (Fond., 5te et 3ce majeure ou mineure). La contribution $C(i)$ d'une note de hauteur h_i située à moins d'un quart de ton de l'un des trois pôles (noté h_p) est telle que :

$$C(i) = 1 - 2|h_p - h_i| \tag{C.2}$$

Ainsi, si $h_p = h_i$, la contribution est maximale (=1), sinon cette dernière diminue de manière linéaire, jusqu'à s'annuler lorsque l'écart entre les deux hauteurs est égal à un quart de ton (soit 0.5 demi-ton).

Comme on peut le voir sur l'expression C.2, une autre simplification par rapport à l'algorithme original, consiste à considérer uniquement le nombre d'occurences d'une hauteur au sein de la mélodie étudiée, et non sa durée totale d'apparition.

Le score d'une tonalité étudiée dépend donc du nombre de notes avoisinant les valeurs stables, et également de leur position par rapport à ces dernières.

C.2 Application

Le motif mélodique recherché par la requête que nous allons traiter est le début du chant de "Amsterdam" (J. Brel). La valeur des hauteurs transcrites est présentée tableau C.3.

[1] Il y en a *a priori* autant que de notes, puisque dans une requête chantonnée, il est rare que deux notes aient exactement la même hauteur.

Rang	Ecart	Ecart (bis)	Identification
1	-2.52	9.48	-
2	-2.88	9.12	-
3	1.79	-	-
4	1.83	-	-
5	3.80	-	tierce majeure
6	4.86	-	-
7	7.10	-	quinte
8	5.58	-	-
9	4.04	-	tierce majeure
10	-0.12	-	fondamentale
11	0.00	-	fondamentale
12	0.07	-	fondamentale

TAB. C.4 : *Exemple de recensement des notes contribuant au score d'une tonalité (hypothèse : ton = hauteur de la onzième note).*

Nous allons détailler le calcul de la note finalement désignée comme étant le ton le mieux représenté, soit la onzième note. Le tableau C.4 présente l'écart des hauteurs traitées par rapport à cette onzième note prise comme ton potentiel. La troisième colonne ramène (+12 demi-tons) les valeurs inférieures à -0.5 demi-ton afin de les comparer avec les valeurs d'intervalles présentées plus tôt (0, +3, +4, +7).

N'ayant recensé qu'un seul type de tierce (majeure), seul le mode majeur est envisagé. Lorsque deux types de tierce sont recensés, deux calculs distincts sont effectués, aboutissant au score des deux tonalités correspondant aux deux modes du ton considéré.

Les 6 hauteurs sélectionnées contribuent, de la manière suivante, au score de la tonalité, dont le ton est la hauteur de la note de rang 11, et dont le mode est majeur :

$$\begin{aligned} Score[note11, majeur] &= (1-2|4-3.80|)+(1-2|7-7.10|)+(1-2|4-4.04|) \\ &\quad +(1-2|0+0.12|)+1+(1-2|0-0.07|) \\ &= 4.94 \end{aligned}$$

Toutes les hauteurs sont ainsi successivement considérées comme ton potentiel, menant aux scores des différentes tonalités représentées. Ceux-ci, présentés tableau C.3, permettent d'élir la tonalité prépondérante, et donc de disposer des informations de mode, et surtout de ton.

Annexe D

Application des Mesures Objectives issues des Tests Subjectifs

Nous rappelons l'expression des trois requêtes comportant chacune une des trois erreurs traitées :

- Erreur Locale :
$$\bar{r} = \bar{d} + \bar{b}_{EL(a,\alpha)}$$
avec $\bar{b}_{EL(a,\alpha)} = [0,\ldots,0,a,0,\ldots,0]$, $\alpha \in [1,N]$ étant le rang auquel l'erreur apparaît ;

- Rupture de Ton :
$$\bar{r} = \bar{d} + \bar{b}_{RT(a,\alpha)}$$
avec $\bar{b}_{RT(a,\alpha)} = [0,\ldots,0,a,\ldots,a]$;

- Glissement de Ton :
$$\bar{r} = \bar{d} + \bar{b}_{GT(a)}$$
avec $\bar{b}_{GT(a)} = [0,a,2a,\ldots,(N-1)a]$.

Nous désignons par *Fonction Comportementale Elémentaire*, ou *FCE*, l'expression de la dissimilarité engendrée par un type d'erreur donné.

D.1 Calcul des Fonctions Comportementales Elémentaires

Tirées des équations 5.6, 5.9, et 5.11, les *FCE* suivantes sont donc représentatives de la dissimilarité mélodique mesurée par nos trois moteurs de comparaison :

$$FCE_{h1} = N * D_{h1}(\bar{r},\bar{d}) = \sum_{j=0}^{N-1} |b_j - median(\bar{b})| \tag{D.1}$$

$$FCE_{h2} = N * D_{h2}(\bar{r},\bar{d}) = \sum_{j=0}^{N-1} (b_j - moyenne(\bar{b}))^2 \quad \text{(D.2)}$$

$$FCE_{i\gamma} = N * D_i(\bar{r},\bar{d}) = \sum_{j=0}^{N-2} \left|b_{j+1} - b_j\right|^{\gamma} \quad \text{(D.3)}$$

avec $\gamma = \{1,2\}$.

Contrairement aux mesures de similarités, les *FCE* ne sont pas normalisées par le nombre de notes N. En effet, il ne s'agit pas ici d'observer l'erreur moyenne par note mise en jeu, mais bien la dissimilarité mesurée entre deux mélodies séparées par une erreur donnée.

D.1.1 Erreur Locale - EL

Pour ce type d'erreur, la requête est la suivante : $\bar{r} = \bar{d} + \bar{b}_{EL(a,\alpha)}$.

Moteur $h1$
Le vecteur $\bar{b}_{EL(a,\alpha)} = [0,\ldots,0,a,0,\ldots,0]$ ayant une valeur médiane nulle [1], l'expression D.1, enrichie d'un paramètre identifiant le type d'erreur considéré, nous donne :

$$FCE_{h1}[EL(a,\alpha)] = N * D_{h1}(\bar{d} + \bar{b}_{EL(a,\alpha)}, \bar{d}) = \sum_{j=1}^{\alpha-1} |0| + |a| + \sum_{j=\alpha+1}^{N} |0|$$

d'où

$$\boxed{FCE_{h1}[EL(a,\alpha)] = |a|} \quad \text{(D.4)}$$

On note une indépendance à α, ainsi qu'à N. Ainsi, que la requête soit longue ou courte, et que l'erreur se situe en debut ou fin de celle-ci, $FCE_{h1}[EL]$ reste égale à l'amplitude de la perturbation (en valeur absolue). Dans ce cas, l'ajustement α n'empêche pas le moteur de comparaison de témoigner de l'erreur elle-même. En effet, le centrage de l'erreur par sa valeur médiane n'a aucune incidence.

Moteur $h2$
Le vecteur $\bar{b}_{EL(a,\alpha)} = [0,\ldots,0,a,0,\ldots,0]$ ayant pour valeur moyenne $\frac{a}{N}$, l'expression D.2 nous donne :

$$FCE_{h2}[EL(a,\alpha)] = N * D_{h2}(\bar{d} + \bar{b}_{EL(a,\alpha)}, \bar{d}) = \sum_{j=1}^{\alpha-1} (0 - \frac{a}{N})^2 + \sum_{j=\alpha}^{N} (a - \frac{a}{N})^2 \quad \text{(D.5)}$$

[1] Nous considérons que $N \geq 3$. Cela allège le cheminement sans affecter sensiblement le champ d'application des résultats.

d'où

$$\boxed{FCE_{h2}[EL(a,\alpha)] = \frac{(N-1)a^2}{N}} \tag{D.6}$$

Dans ce cas, l'indépendance à α est constatée. $FCE_{h2}[EL]$ augmente avec le carré de l'amplitude a de l'erreur. Ici, l'influence de l'ajustement est d'autant plus grande que N est petit. Si l'on considère a^2 comme la dissimilarité "attendue" pour le moteur $h2$, on peut dire que l'ajustement diminue la mesure effectuée.

Moteur *i*

Le vecteur $\bar{b}_{EL(a,\alpha)} = [0,\dots,0,a,0,\dots,0]$ implique que seuls deux éléments de la somme de l'expression D.3 sont non nuls. On a

$$FCE_{i\gamma}[EL(a,\alpha)] = N * D_{i\gamma}(\bar{d}+\bar{b}_{EL(a,\alpha)},\bar{d}) = \sum_{j=0}^{\alpha-2}|0|^{\gamma}+|a|^{\gamma}+|-a|^{\gamma}+\sum_{j=\alpha+1}^{N-2}|0|^{\gamma}$$

$$\boxed{FCE_{i\gamma}[EL(a,\alpha)] = 2|a|^{\gamma}} \tag{D.7}$$

avec $\gamma = \{1,2\}$.

Comme pour $FCE_{h1}[EL]$, $FCE_{i\gamma}[EL]$ reste fonction de l'amplitude a de l'erreur, quel que soit l'emplacement, que la requête soit longue ou courte. Comme remarqué précédemment, une EL provoque deux erreurs d'intervalles, la FCE résultante correspond au double de la dissimilarité "attendue".

D.1.2 Rupture de Ton - RT

Pour ce type d'erreur, la requête est la suivante : $\bar{r} = \bar{d}+\bar{b}_{RT(a,\alpha)}$.

Moteur *h1*

Selon l'emplacement α de l'erreur, et le nombre de notes de la requête, N, la valeur médiane de $\bar{b}_{RT(a,\alpha)} = [0,\dots,0,a,\dots,a]$ peut emprunter plusieurs valeurs[2] :

1. Si N est impair,

$$median(\bar{b}_{RT(a,\alpha)}) = a \quad \text{si} \quad \alpha \in [2;\tfrac{N+1}{2}]$$

$$median(\bar{b}_{RT(a,\alpha)}) = 0 \quad \text{si} \quad \alpha \in [\tfrac{N+3}{2};N-1]$$

[2]Bien que α soit défini de 1 à N, nous ne considèrerons que ses valeurs comprises entre 2 et $N-1$. En effet, une RT au rang $\alpha = 1$ ne cause aucune rupture (elle ne fait que changer le ton sur l'ensemble de la requête), et une RT au rang N n'entre pas dans le cadre des erreurs traitées ici (ni RT, ni EL).

2. Si N est pair,

$$median(\bar{b}_{RT(a,\alpha)}) = a \quad \text{si} \quad \alpha \in [2; \tfrac{N}{2}]$$

$$median(\bar{b}_{RT(a,\alpha)}) = \tfrac{a}{2} \quad \text{si} \quad \alpha = \tfrac{N}{2} + 1$$

$$median(\bar{b}_{RT(a,\alpha)}) = 0 \quad \text{si} \quad \alpha \in [\tfrac{N}{2} + 2; N - 1]$$

Les trois valeurs médianes que nous venons de recenser entraînent les trois valeurs de fonction, issues de l'expression D.1 :

1. Avec $median(\bar{b}_{RT(a,\alpha)}) = a$, on a

$$FCE_{h1}[RT(a,\alpha)] = N * D_{h1}(\bar{d} + \bar{b}_{RT(a,\alpha)}, \bar{d}) = \sum_{j=1}^{\alpha-1} |a| + \sum_{j=\alpha}^{N} |a - a|$$

d'où

$$\boxed{FCE_{h1}[RT(a,\alpha)] = (\alpha - 1)|a|} \tag{D.8}$$

2. Avec $median(\bar{b}_{RT(a,\alpha=\frac{N}{2}+1)}) = \frac{a}{2}$, on a

$$FCE_{h1}(RT(a,\alpha = \frac{N}{2} + 1)) = \sum_{j=1}^{\alpha-1} |\frac{a}{2}| + \sum_{j=\alpha}^{N} |a - \frac{a}{2}|$$

d'où

$$\boxed{FCE_{h1}(RT(a,\alpha = \frac{N}{2} + 1)) = \frac{N|a|}{2}} \tag{D.9}$$

3. Avec $median(\bar{b}_{RT(a,\alpha)}) = 0$, on a

$$FCE_{h1}[RT(a,\alpha)] = \sum_{j=1}^{\alpha-1} |0| + \sum_{j=\alpha}^{N} |a|$$

d'où

$$\boxed{FCE_{h1}[RT(a,\alpha)] = (N - \alpha + 1)|a|} \tag{D.10}$$

En résumé : $FCE_{h1}[RT(a,\alpha)]$ s'exprime, selon N et α, de la manière suivante :

1. Si N est impair

$$\begin{aligned} FCE_{h1}[RT(a,\alpha)] &= (\alpha-1)|a| && \text{si} \quad \alpha \in [2; \tfrac{N+1}{2}] \\ &= (N-\alpha+1)|a| && \text{si} \quad \alpha \in [\tfrac{N+3}{2}; N-1] \end{aligned}$$

2. Si N est pair

$$\begin{aligned} FCE_{h1}[RT(a,\alpha)] &= (\alpha-1)|a| && \text{si} \quad \alpha \in [2, \tfrac{N}{2}] \\ &= \tfrac{N}{2}|a| && \text{si} \quad \alpha = \tfrac{N}{2}+1 \\ &= (N-\alpha+1)|a| && \text{si} \quad \alpha \in [\tfrac{N}{2}+2, N-1] \end{aligned}$$

La dissimilarité mesurée dépend donc à la fois de α et de N. Elle est égale à l'amplitude de l'erreur, modulée par un facteur correspondant au nombre de notes constituant le ton le moins représenté. $FCE_{h1}[RT]$ obtient donc son maximum pour une erreur située en milieu de requête. Par ailleurs, ce maximum est d'autant plus élevé que N est grand.

Moteur $h2$

Le vecteur $\bar{b}_{RT(a,\alpha)} = [0,\ldots,0,a,\ldots,a]$ ayant pour valeur moyenne $\frac{(N-\alpha+1)a}{N}$, l'expression D.2 devient :

$$FCE_{h2}[RT(a,\alpha)] = N * D_{h2}(\bar{d}+\bar{b}_{RT(a,\alpha)}, \bar{d}) = \sum_{j=1}^{\alpha-1}\left(0-\frac{(N-\alpha+1)a}{N}\right)^2 + \sum_{j=\alpha}^{N}\left(a-\frac{(N-\alpha+1)a}{N}\right)^2$$

d'où

$$\boxed{FCE_{h2}[RT(a,\alpha)] = \frac{(\alpha-1)(N-\alpha+1)a^2}{N}} \tag{D.11}$$

Comme pour le moteur $h1$, la dissimilarité mesurée dépend à la fois de α et de N. Le carré de l'amplitude de la RT est modulé par un facteur de type $\frac{AB}{A+B}$, A et B représentant la taille des portions mélodiques de ton stable (on a donc $A+B=N$).

$FCE_{h2}[RT]$ ayant pour dérivée par rapport à N, l'expression $[\frac{a}{N}(\alpha-1)]^2$ strictement positive, elle augmente avec la taille N de la requête.

$FCE_{h2}[RT]$ a pour dérivée par rapport à α, l'expression $\frac{a^2}{N}[\frac{N}{2}-\alpha+1]$. Cette dernière étant positive pour $\alpha \in [2; \frac{N}{2}+1]$, et négative pour $\alpha \in [\frac{N}{2}+1; N-1]$, $FCE_{h2}[RT]$ augmente sur le premier intervalle et diminue sur le second. Son maximum est donc atteint en $\alpha = \frac{N}{2}+1$, soit pour une perturbation située en milieu de requête.

Moteur i

Le vecteur $\bar{b}_{RT(a,\alpha)} = [0,\ldots,0,a,\ldots,a]$ implique qu'un seul élément de la somme de l'expression D.3 est non nul. On a

$$FCE_{i\gamma}[RT(a,\alpha)] = N * D_{i\gamma}(\bar{d}+\bar{b}_{RT(a,\alpha)},\bar{d}) = \sum_{j=1}^{\alpha-1}|0|^{\gamma}+|a|^{\gamma}+\sum_{j=\alpha+1}^{N-1}|0|^{\gamma}$$

$$\boxed{FCE_{i\gamma}[RT(a,\alpha)] = |a|^{\gamma}} \tag{D.12}$$

avec $\gamma = \{1,2\}$.

Comme pour $FCE_{i\gamma}[EL]$, $FCE_{i\gamma}[RT]$ reste identique, que la requête soit longue ou courte, et quel que soit l'emplacement de l'erreur. La dissimilarité est donc uniquement fonction de l'amplitude a de l'erreur.

Comme attendu, les moteurs fondés sur les séquences d'intervalles sont moins pénalisants vis-à-vis des RT que des EL.

D.1.3 Glissement de Ton - GT

Pour ce type d'erreur, la requête est la suivante : $\bar{r} = \bar{d}+\bar{b}_{GT(a)}$.

Moteur $h1$

1. Si N est impair, $median(b_{GT})$ correspond à l'élément central de $b_{GT} = [0,a,2a,\ldots,(N-1)a]$ soit $\left(\frac{N-1}{2}\right)a$. Et puisque $b_{GT}(j) = j*a$, $\forall j \in [0,N-1]$, la FCE devient :

$$\begin{aligned} FCE_{h1}[GT(a)] = N * D_{h1}(\bar{d}+\bar{b}_{GT(a)},\bar{d}) \quad &= \sum_{j=0}^{\frac{N-1}{2}-1}\left|j*a-\left(\frac{N-1}{2}\right)a\right|+\left|\left(\frac{N-1}{2}\right)a-\left(\frac{N-1}{2}\right)a\right| \\ &+ \sum_{j=\frac{N-1}{2}+1}^{N-1}\left|j*a-\left(\frac{N-1}{2}\right)a\right| \end{aligned}$$

$$
\begin{aligned}
FCE_{h1}[GT(a)] \quad = \quad & \left|0-\left(\frac{N-1}{2}\right)a\right|+\left|a-\left(\frac{N-1}{2}\right)a\right|+\ldots \\
& +\left|\left(\frac{N-1}{2}-1\right)a-\left(\frac{N-1}{2}\right)a\right|+0+\left|\left(\frac{N-1}{2}+1\right)a-\left(\frac{N-1}{2}\right)a\right|+\ldots \\
& +\left|(N-2)a-\left(\frac{N-1}{2}\right)a\right|+\left|(N-1)a-\left(\frac{N-1}{2}\right)a\right|
\end{aligned}
$$

ce qui donne

$$
\begin{aligned}
FCE_{h1}[GT(a)] \quad = \quad & \left|-a\left(\frac{N-1}{2}\right)\right|+\left|-a\left(\frac{N-3}{2}\right)\right| \\
& +\left|-a\left(\frac{N-5}{2}\right)\right|+\ldots+|-2a|+|-a| \\
& +|a|+|2a|+\ldots+\left|a\left(\frac{N-5}{2}\right)\right| \\
& +\left|a\left(\frac{N-3}{2}\right)\right|+\left|a\left(\frac{N-1}{2}\right)\right|
\end{aligned}
$$

d'où

$$
\begin{aligned}
FCE_{h1}[GT(a)] \quad & = \quad 2*\left[|a|+|2a|+\ldots\left|a\left(\frac{N-3}{2}\right)\right|+\left|a\left(\frac{N-1}{2}\right)\right|\right] \\
& = \quad 2|a|*\left[1+2+\ldots+\frac{N-3}{2}+\frac{N-1}{2}\right] \\
& = \quad 2|a|*\left[\sum_{j=1}^{\frac{N-1}{2}} j\right] \\
& = \quad 2|a|*\left[\frac{N-1}{2}*\left(\frac{N-1}{2}+1\right)\right]
\end{aligned}
$$

ce qui donne finalement

$$
\boxed{FCE_{h1}[GT(a)] = \frac{(N^2-1)|a|}{4}} \tag{D.13}
$$

2. Si N est pair, $median(b_{GT})$ correspond à la moyenne des deux éléments centraux de b_{GT} soit $\left(\frac{N-1}{2}\right)a$.
   Cette valeur médiane, identique à celle obtenue au cas N impair, entraîne un calcul très similaire au précédent. On obtient l'expression suivante :

$$\boxed{FCE_{h1}[GT(a)] = \frac{N^2|a|}{4}} \tag{D.14}$$

$FCE_{h1}[GT(a)]$ augmente donc avec le carré de $2E[\frac{N}{2}]$ (avec $E[x]$, partie entière de x), et est proportionnel à l'incrément a de l'erreur.

Moteur $h2$

Que N soit pair ou impair, $moyenne(b_{GT})$ est égal à $\left(\frac{N-1}{2}\right)a$.

Cette valeur moyenne, identique aux valeurs médianes obtenues avec le moteur $h1$, entraîne un calcul très similaire aux deux précédents. On aboutit à l'expression suivante :

$$\begin{aligned} FCE_{h2}[GT(a)] = N * D_{h2}(\bar{d}+\bar{b}_{GT(a)},\bar{d}) &= 2a^2 * \left[1+2^2+\ldots\left(\frac{N-3}{2}\right)^2+\left(\frac{N-1}{2}\right)^2\right] \\ &= 2a^2 * \left[\sum_{j=1}^{\frac{N-1}{2}} j^2\right] \\ &= 2a^2 * \frac{\left(\frac{N-1}{2}\right)\left[\left(\frac{N-1}{2}\right)+1\right]\left[2\left(\frac{N-1}{2}\right)+1\right]}{6} \end{aligned}$$

ce qui donne finalement

$$\boxed{FCE_{h2}[GT(a)] = \frac{N(N^2-1)a^2}{12}} \tag{D.15}$$

$FCE_{h2}[GT(a)]$ augmente donc avec le carré de l'amplitude a de la perturbation, et quasiment avec le cube de N.

Moteur i

La construction du vecteur $\bar{b}_{GT(a)} = [0, a, 2a, \ldots, (N-1)a]$ implique que tous les éléments de la somme de l'expression D.3 sont identiques :

$$FCE_{i\gamma}[GT(a)] = N * D_{i\gamma}(\bar{d}+\bar{b}_{GT(a)},\bar{d}) = \sum_{j=0}^{N-2}[(j+1)|a|^{\gamma} - j|a|^{\gamma}] = \sum_{j=0}^{N-2}|a|^{\gamma}$$

d'où

$$\boxed{FCE_{i\gamma}(GT(a)) = (N-1)|a|^{\gamma}} \tag{D.16}$$

avec $\gamma = \{1,2\}$.

$FCE_{i\gamma}(GT(a))$ augmente donc avec l'amplitude a de la perturbation, ainsi qu'avec N.

Ainsi, l'ensemble des moteurs considérés témoignent d'une dissimilarité d'autant plus grande que N est élevé.

Les *FCE* que nous venons de calculer, pour chaque moteur et pour chacun des trois types d'erreurs, nous serviront de base pour la construction des *FCR*. Ces dernières nous permettront par la suite d'estimer le moteur le plus conforme aux résultats de nos tests subjectifs.

D.2 Calcul des Fonctions Comportementales Relatives

Les résultats des tests subjectifs effectués portent sur les trois comparaisons EL/RT, GT/EL et GT/RT. Nous allons donc calculer les *FCR* correspondantes.

D.2.1 Relation entre EL et RT

Afin d'observer la manière dont EL et RT sont traitées l'une par rapport à l'autre, nous allons étudier le rapport existant entre leurs deux *FCE*.

Dans les tests, pour un motif mélodique donné, EL et RT étaient systématiquement injectées au même endroit afin de conserver le contexte mélodique de l'erreur (et ainsi procéder à des comparaisons). Il en sera donc de même ici : le rang α de l'erreur injectée sera le même pour EL et RT.

La convention suivante simplifiera les expressions manipulées :

$$FCR_X[EL(a_1,\alpha), RT(a_2,\alpha)] = \frac{FCE_X[EL(a_1,\alpha)]}{FCE_X[RT(a_2,\alpha)]} = \frac{N * D_X(\bar{d}+\bar{b}_{EL(a_1)},\bar{d})}{N * D_X(\bar{d}+\bar{b}_{RT(a_2)},\bar{d})} \tag{D.17}$$

X désignant le moteur de comparaison concerné, $X = \{h1, h2, i\}$.

Remarque : Puisque la *FCR* représente un rapport, on a la propriété suivante :

$$FCR_X[Erreur1, Erreur2] = \frac{1}{FCR_X[Erreur2, Erreur1]} \tag{D.18}$$

Par exemple

$$FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{1}{FCR_{h1}[RT(a_2),EL(a_1)]} \tag{D.19}$$

Moteur $h1$: Relation entre EL et RT

Les équations D.4, D.8, D.9, D.10 nous donnent les expressions suivantes :

1. Si N est impair
   (a) pour $\alpha \in [2, \frac{N+1}{2}]$

   $$FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{|a_1|}{(\alpha-1)|a_2|}$$

   d'où

   $$\boxed{FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{1}{\alpha-1} * \left|\frac{a_1}{a_2}\right|} \tag{D.20}$$

   (b) pour $\alpha \in [\frac{N+3}{2}, N-1]$

   $$FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{|a_1|}{(N-\alpha+1)|a_2|}$$

   d'où

   $$\boxed{FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{1}{N-\alpha+1} * \left|\frac{a_1}{a_2}\right|} \tag{D.21}$$

2. Si N est pair
   (a) pour $\alpha \in [2, \frac{N}{2}]$

   $$\boxed{FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{1}{\alpha-1} * \left|\frac{a_1}{a_2}\right|} \tag{D.22}$$

   (b) pour $\alpha = \frac{N}{2}+1$:

   $$FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = |a_1| * \frac{2}{N|a_2|}$$

   d'où

   $$\boxed{FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{2}{N} * \left|\frac{a_1}{a_2}\right|} \tag{D.23}$$

(c) pour $\alpha \in [\frac{N}{2}+2, N-1]$:

$$\boxed{FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{1}{N-\alpha+1} * \left|\frac{a_1}{a_2}\right|} \tag{D.24}$$

$FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)]$ emprunte donc des expressions différentes selon la taille N de la requête ainsi que l'emplacement α des erreurs.

Le domaine de variation de $FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)]$ en fonction de α est le suivant :

$$\boxed{\frac{2}{N-1} * \left|\frac{a_1}{a_2}\right| \leq FCR_{h1}[EL(a_1,\alpha),RT(a_2,\alpha)] \leq \left|\frac{a_1}{a_2}\right|} \tag{D.25}$$

pour $N \leq 3$, et $\forall\, \alpha \in [2; N-1]$.

Quel que soit α, la fonction est donc inférieure ou égale au rapport des amplitudes d'erreur, sachant que, plus N est grand, plus la borne inférieure de la fonction diminue.

Moteur $h2$: Relation entre EL et RT

Comme pour le moteur $h1$, nous allons voir, grâce au rapport $FCR_{h2}[EL(a_1,\alpha),RT(a_2,\alpha)]$, comment EL et RT sont traitées l'une par rapport à l'autre. Les équations D.6 et D.11 nous donnent :

$$FCR_{h2}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{(N-1){a_1}^2}{N} * \frac{N}{(\alpha-1)(N-\alpha+1){a_2}^2}$$

d'où

$$\boxed{FCR_{h2}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{N-1}{(\alpha-1)(N-\alpha+1)} * \left(\frac{a_1}{a_2}\right)^2} \tag{D.26}$$

Le domaine de variation de $FCR_{h2}[EL(a_1,\alpha),RT(a_2,\alpha)]$ en fonction de α est le suivant :

$$\boxed{\frac{4(N-1)}{N^2} * \left(\frac{a_1}{a_2}\right)^2 \leq FCR_{h2}[EL(a_1,\alpha),RT(a_2,\alpha)] \leq \left(\frac{a_1}{a_2}\right)^2} \tag{D.27}$$

pour $N \leq 3$, et $\forall\, \alpha \in [2; N-1]$.

On arrive donc à un résultat similaire à celui obtenu pour $h1$. la fonction calculée est inférieure au carré du rapport des amplitudes d'erreur, et également supérieure à une borne diminuant lorsque N grandit.

Moteur i : Relation entre EL et RT

Comme pour les deux moteurs précédents $h1$ et $h2$, nous allons voir comment EL et RT sont traitées l'une par rapport à l'autre. Les équations D.7 et D.12 nous donnent :

$$FCR_{i\gamma}[EL(a_1,\alpha),RT(a_2,\alpha)] = \frac{2|a_1|^\gamma}{|a_2|^\gamma}$$

$$\boxed{FCR_{i\gamma}[EL(a_1,\alpha),RT(a_2,\alpha)] = 2 * \left|\frac{a_1}{a_2}\right|^\gamma} \tag{D.28}$$

avec $\gamma = \{1,2\}$.

A la différence des moteurs $h1$ et $h2$, $FCR_{i\gamma}[EL(a_1,\alpha),RT(a_2,\alpha)]$ est indépendante de l'emplacement α de l'erreur, ainsi que de la taille N de la requête. Autre différence, le facteur modulant le rapport des amplitudes est supérieur à 1.

On observe donc deux tendances selon la nature des moteurs considérés. Pour ceux fondés sur les hauteurs, le facteur modulant est inférieur à 1, et pour ceux fondés sur les intervalles, il est supérieur à 1.

D.2.2 Relation entre GT et EL

Dans cette section, nous allons observer comment les GT sont traités par rapport aux EL. Nous considèrerons l'expression suivante :

$$FCR_X[GT(a_1),EL(a_2,\alpha)] = \frac{N * D_X(\bar{d}+\bar{b}_{GT(a_1)},\bar{d})}{N * D_X(\bar{d}+\bar{b}_{EL(a_2)},\bar{d})} \tag{D.29}$$

X désignant toujours le moteur de comparaison concerné, $X = \{h1,h2,i\}$.

Moteur $h1$: Relation entre GT et EL

Concernant $FCR_{h1}(GT(a_1),EL(a_2,\alpha))$, deux cas sont recensés en fonction de la valeur de N :

1. Si N est impair, alors les expressions D.13 et D.4 nous donnent

$$FCR_{h1}[GT(a_1),EL(a_2,\alpha)] = \frac{(N^2-1)|a_1|}{4} * \frac{1}{|a_2|}$$

d'où

$$\boxed{FCR_{h1}[GT(a_1),EL(a_2,\alpha)] = \frac{N^2-1}{4} * \left|\frac{a_1}{a_2}\right|} \tag{D.30}$$

2. Si N est pair, alors les expressions D.14 et D.4 nous donnent

$$FCR_{h1}[GT(a_1),EL(a_2,\alpha)] = \frac{N^2|a_1|}{4} * \frac{1}{|a_2|}$$

d'où

$$\boxed{FCR_{h1}[GT(a_1),EL(a_2,\alpha)] = \frac{N^2}{4} * \left|\frac{a_1}{a_2}\right|} \tag{D.31}$$

La détermination du domaine de variation est directe, puisque $FCR_{h1}(GT(a_1),EL(a_2,\alpha))$ n'emprunte que deux valeurs (indépendantes de α) :

$$\boxed{\frac{N^2-1}{4} * \left|\frac{a_1}{a_2}\right| \leq FCR_{h1}[GT(a_1),EL(a_2,\alpha)] \leq \frac{N^2}{4} * \left|\frac{a_1}{a_2}\right| \quad , \ \forall\, \alpha \in [2;N-1]} \tag{D.32}$$

La FCR calculée est donc systématiquement supérieure au rapport des amplitudes d'erreur. Par ailleurs, l'emplacement de l'EL n'y a aucune influence.

<u>**Moteur $h2$: Relation entre GT et EL**</u>

Concernant $FCR_{h2}(GT(a_1),EL(a_2,\alpha))$, les expressions D.15 et D.6 nous donnent :

$$FCR_{h2}[GT(a_1),EL(a_2,\alpha)] = \frac{N(N^2-1){a_1}^2}{12} * \frac{N}{(N-1){a_2}^2}$$

d'où

$$\boxed{FCR_{h2}[GT(a_1),EL(a_2,\alpha)] = \frac{N^2(N+1)}{12} * \left(\frac{a_1}{a_2}\right)^2} \tag{D.33}$$

On peut noter une indépendance de la FCR vis-à-vis de α, ainsi qu'un facteur modulant supérieur à 1. La FCR est donc supérieure au carré du rapport des amplitudes d'erreur.

<u>**Moteur i : Relation entre GT et EL**</u>

Concernant $FCR_{i\gamma}(GT(a_1),EL(a_2,\alpha))$, les expressions D.16 et D.7 nous donnent :

$$FCR_{i\gamma}[GT(a_1),EL(a_2,\alpha)] = \frac{(N-1)|a_1|^\gamma}{2|a_2|^\gamma}$$

d'où

$$\boxed{FCR_{i\gamma}[GT(a_1),EL(a_2,\alpha)] = \frac{(N-1)}{2} * \left|\frac{a_1}{a_2}\right|^\gamma} \tag{D.34}$$

avec $\gamma = \{1,2\}$.

Dans ce cas également, le FCR est indépendance de α, et le facteur modulant est supérieur à 1.

Cette fois, l'ensemble des moteurs s'accorde à mesurer une dissimilarité supérieure au rapport des amplitudes d'erreur (ou à son carré selon les distances utilisées).

D.2.3 Relation entre GT et RT

Dans cette section, nous allons observer comment les GT sont traités par rapport aux RT. Nous considèrerons l'expression suivante :

$$FCR_X[GT(a_1), RT(a_2,\alpha)] = \frac{N * D_X(\bar{d} + \bar{b}_{GT(a_1)}, \bar{d})}{N * D_X(\bar{d} + \bar{b}_{RT(a_2)}, \bar{d})} \tag{D.35}$$

X désignant toujours le moteur de comparaison concerné, $X = \{h1, h2, i\}$.

Moteur $h1$: Relation entre GT et RT

Concernant $FCR_{h1}(GT(a_1), RT(a_2,\alpha))$, cinq cas sont recensés selon la valeur de N et la valeur médiane de $\bar{b}_{RT(a,\alpha)} = [0, \ldots, 0, a_2, \ldots, a_2]$:

1. Si N est impair
   (a) pour $\alpha \in [2, \frac{N+1}{2}]$, les expressions D.13 et D.8 nous donnent :

   $$FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{(N^2-1)|a_1|}{4} * \frac{1}{(\alpha-1)|a_2|}$$

   d'où

   $$\boxed{FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2-1}{4(\alpha-1)} * \left|\frac{a_1}{a_2}\right|} \tag{D.36}$$

   (b) pour $\alpha \in [\frac{N+3}{2}, N-1]$, les expressions D.13 et D.10 nous donnent :

   $$FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{(N^2-1)|a_1|}{4} * \frac{1}{(N-\alpha+1)|a_2|}$$

   d'où

   $$\boxed{FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2-1}{4(N-\alpha+1)} * \left|\frac{a_1}{a_2}\right|} \tag{D.37}$$

2. Si N est pair

(a) pour $\alpha \in [2, \frac{N}{2}]$, les expressions D.14 et D.8 nous donnent :

$$FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2|a_1|}{4} * \frac{1}{(\alpha-1)|a_2|}$$

d'où

$$\boxed{FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2}{4(\alpha-1)} * \left|\frac{a_1}{a_2}\right|}$$

(b) pour $\alpha = \frac{N}{2}+1$, les expressions D.14 et D.9 nous donnent :

$$FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2|a_1|}{4} * \frac{2}{N|a_2|}$$

d'où

$$\boxed{FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N}{2} * \left|\frac{a_1}{a_2}\right|} \tag{D.38}$$

(c) pour $\alpha \in [\frac{N}{2}+2, N-1]$, les expressions D.14 et D.10 nous donnent :

$$FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2|a_1|}{4} * \frac{1}{(N-\alpha+1)|a_2|}$$

d'où

$$\boxed{FCR_{h1}(GT(a_1), RT(a_2,\alpha)) = \frac{N^2}{4(N-\alpha+1)} * \left|\frac{a_1}{a_2}\right|} \tag{D.39}$$

Le domaine de variation de $FCR_{h1}[GT(a_1), RT(a_2,\alpha)]$ en fonction de α est le suivant :

$$\boxed{\frac{N}{2} * \left|\frac{a_1}{a_2}\right| \leq FCR_{h1}[GT(a_1), RT(a_2,\alpha)] \leq \frac{N^2}{4} * \left|\frac{a_1}{a_2}\right|} \tag{D.40}$$

pour $N \leq 3$, et $\forall\, \alpha \in [2; N-1]$.

Comme pour la correspondance entre GT et EL, on observe un facteur modulant supérieur à 1. Par contre, la FCR calculée ici est dépendante de l'emplacement α de la RT.

Moteur $h2$: Relation entre GT et RT

Concernant $FCR_{h2}[GT(a_1), RT(a_2,\alpha)]$, les expressions D.15 et D.11 nous donnent :

$$FCR_{h2}[GT(a_1), RT(a_2,\alpha)] = \frac{N(N^2-1){a_1}^2}{12} * \frac{N}{(\alpha-1)(N-\alpha+1){a_2}^2}$$

d'où

$$\boxed{FCR_{h2}[GT(a_1), RT(a_2,\alpha)] = \frac{N^2(N^2-1)}{12(\alpha-1)(N-\alpha+1)} * \left(\frac{a_1}{a_2}\right)^2} \tag{D.41}$$

Le domaine de variation de $FCR_{h1}[GT(a_1), RT(a_2, \alpha)]$ en fonction de α est le suivant :

$$\boxed{\frac{N^2-1}{3} * \left(\frac{a_1}{a_2}\right)^2 \leq FCR_{h2}[GT(a_1), RT(a_2, \alpha)] \leq \frac{N^2(N+1)}{12} * \left(\frac{a_1}{a_2}\right)^2} \quad \text{(D.42)}$$

pour $N \leq 3$, et $\forall\, \alpha \in [2; N-1]$.

A l'instar du moteur $h1$, le moteur $h2$ présente un facteur modulant supérieur à 1, à la fois pour la correspondance entre GT et EL, et pour celle entre GT et RT. Deuxième similarité avec le moteur $h1$, on note également ici la dépendance de la FCR calculée ici à l'emplacement α de la RT.

Moteur *i* : Relation entre GT et RT

Concernant $FCR_{\,i\gamma}(GT(a_1), RT(a_2, \alpha))$, les expressions D.16 et D.12 nous donnent :

$$FCR_{\,i\gamma}(GT(a_1), RT(a_2, \alpha)) = \frac{(N-1)|a|\gamma}{|a|\gamma}$$

d'où

$$\boxed{FCR_{\,i\gamma}(GT(a_1), RT(a_2, \alpha)) = (N-1) * \left|\frac{a_1}{a_2}\right|^{\gamma}} \quad \text{(D.43)}$$

avec $\gamma = \{1, 2\}$.

Les moteurs i voient leur FCR indépendante à α. Par contre, ils se ralient aux moteurs h, concernant le facteur modulant supérieur à 1.

Nous disposons donc des FCR permettant de comparer les dissimilarités mesurées pour les quatre moteurs de comparaison proposés.

Grâce à elles, nous allons pouvoir interpréter les résultats des tests subjectifs en terme de performances des moteurs considérés. Ainsi, nous pourrons porter un jugement sur les qualités présumées de ces derniers.

Notons que ces fonctions ont été établies de manière indépendante aux résultats des tests subjectifs effectués. Cela implique qu'elle pourraient être réutilisées pour des résultats différents (inéquité des EL et RT, pas exemple).

Annexe E

Programmes Réalisés

Les programmes réalisés ont été écrits en langages C, et C++.

- Extraction automatique d'une base temporelle
- Extraction automatique de la tonalité
- Programme de constitution de la base de données indexée (Interpréteur de fichiers MIDI - Sélection des pistes et réduction de la polyphonie - Description - Encodage)
- Description de la requête à partir des résultats du module d'analyse de fréquence fondamentale (ce dernier ayant été récupéré)
- Mesures de similarité
- Interface GUI (Graphic User Interface)