Ircam AudioPrint : Evaluation et résultats de la reconnaissance d'extraits audio dégradés

Rémi Mignot

Ircam AudioID :

Evaluation et résultats de la reconnaissance d'extraits audio dégradés

Rémi Mignot

IRCAM -CNRS, rapport interne, projet BeeMusic

31 décembre 2015

Résumé

Ce document présente l'évaluation de l'indexation audio mise en œuvre lors du projet BeeMusic. Ici la problématique testée est la reconnaissance d'un extrait musical dégradé parmi une grande base de morceaux de référence. Les résultats présentés permettront ou pas de valider notre approche pour les deux objectifs donnés : robustesse aux altérations sonores et possibilité de passer à l'échelle. A partir d'un grand nombre d'extraits testées, la robustesse est évaluée en utilisant un grand nombre de dégradations différentes, avec plusieurs forces, et le passage à l'échelle sera validé par l'étude de la mémoire utilisée et du temps de réponse pour une grande base, constituée de plus de 700 milles extraits musicaux de 30 secondes.

1 Introduction

L'un des buts de l'indexation audio est la recherche d'extraits sonores donnés, parmi une base de référence contenant un très grand nombre d'items sonores, qui sont des morceaux de musique par exemple. Cette opération est en général basée sur le calcul d'empreintes sonores obtenues séparément sur ces différents signaux. Ainsi, après avoir construit une base de données d'empreintes sonores des signaux de référence, items, la recherche de l'extrait se fait en quelques sortes, par une comparaison des valeurs.

Dans le cadre du projet BeeMusic, nous avons développé un nouveau système complet pour l'identification audio, cf. le rapport global en [1]. Nous nous sommes focalisés sur deux objectifs : robustesse aux dégradations sonores et passage à l'échelle.

Selon l'utilisation, le flux audio peut éventuellement être altéré par une ou plusieurs dégradations sonores. Par exemple : bruits additifs (synthétiques ou environnementaux), égalisation, filtrage, encodage (formats MP3 ou GSM par exemple), saturation, compression des dynamiques (mono ou multi-bandes), réverbération, changement d'échelles temporelle ou fréquentielle. Le problème est que toute modification du son peut engendrer une modification des empreintes sonores, et ainsi empêcher la reconnaissance de l'extrait. Dans ce contexte, les empreintes développées dans ce travail ont pour vocation à être robuste, dans le sens ou leurs valeurs varient très peu en présence de dégradation.

De plus, afin d'utiliser une base de référence aussi grande que possible, il est nécessaire de s'intéresser à l'algorithme de recherche de sorte à constituer une structure de données aussi petite que possible, et à avoir un très faible temps de réponse, de l'ordre de quelques secondes seulement. Pour cette raison, nous avons aussi développé un nouveau système de recherche, basé sur un hachage approximatif, permettant à la fois un possible passage à l'échelle et une bonne tolérance aux erreurs. Un aperçu complet du système est donné en [1].

Le présent document s'intéresse à l'évaluation de notre méthode dans le cadre de la reconnaissance d'un flux dégradé. Son organisation est la suivante : en section 2 nous présentons la base des morceaux de référence et comment nous

avons sélectionner les fichiers audio. Un résumé du système de reconnaissance est donné en sec. 3, avec une présentation des différentes structures de données construites. En section 4 est présenté la procédure d'évaluation ainsi que les différents types de dégradations. Enfin la section 5 présente l'ensemble des résultats et la section 6 conclut ce document.

2 Filtrage de la base

Dans cette évaluation, nous avons dans un premier temps fait une sélection des extraits musicaux de la base BeeMusic. En effet dans cette base contenant plus de 2 millions d'extraits de 30 secondes en moyenne, beaucoup de morceaux apparaissent plusieurs fois ; nous parlons alors de doublons. Puisque dans cette évaluation nous ne traitons pas du dédoublonnage, pour éviter d'avoir des performances réduites à cause des morceaux dupliqués, nous avons réalisé un dédoublonnage basé sur les ISRC quand ceux là sont donnés. Les extraits ne comportant pas d'ISRC sont exclus. Rappelons que l'ISRC (*International Standard Recording Code*) est un code unique identifiant un enregistrement d'un morceau.

Le genre musical de chaque morceau de la base est renseigné la plupart du temps. Nous y retrouvons les styles : *ambiance, blues, classique, country, techno/dance, enfance, jazz, metal, monde, soul/funk, rock, film, variété/pop, variété, rap/hip hop, raggae, blues/gospel.* Cependant, un style donné est *texte* qui correspond à des enregistrements de voix seule. Puisque nous nous intéressons ici uniquement à des sons musicaux, nous avons exclu ces extraits en ce basant sur le genre annoté. Aussi, tous les morceaux sans genre annoté ont été exclus.

Enfin, les extraits n'ont pas tous une durée de 30 secondes. Certains ont une durée de moins de 5 secondes, et d'autres de près de 10 minutes. Pour éviter des problèmes dus à la longueur des morceaux, nous avons conservés les extraits d'une durée comprise entre 25 et 35 secondes uniquement.

En conséquence de ce filtrage, **la base** totale utilisée pour l'évaluation **contient 755,762 extraits musicaux**, d'une durée comprise en 25 et 35 secondes et sans doublons.

Notons qu'il persiste malgré tout un très petit nombre de doublons, avec des ISRC différents. Aussi, dans de rares cas, l'extrait choisi peut ne pas contenir de musique. Par exemple, s'il s'agit d'un enregistrement de concert et si l'extrait est en début ou en fin de piste, le son peut être celui de la foule et/ou de la voix de l'artiste parlant. Malgré cela, l'évaluation est faite sur ces 755,762 extraits restant.

3 Structures de données et procédure de la reconnaissance

Pour la reconnaissance d'un flux musical, deux tables sont construites : la première est la table de hachage, se comportant comme un index inversé. Elle renvoie pour un code de hachage donné, les identifiants de tous les morceaux l'ayant produit. La seconde table est un index direct revoyant pour un morceau donné, tous les codes de hachage produits, ainsi que l'indice temporel du temps

d'analyse auquel le code a été calculé. L'identification audio est réalisée en deux étapes : une première étape permet de sélectionner parmi tous les extraits de référence de la base, un petit nombre de candidats pour une recherche raffinée en seconde étape.

3.1 Première phase

Dans la première phase, étant donné les codes de hachage du flux à reconnaître, nous en avons en moyenne 4×5×L = 1020 par seconde (avec 4 points d'ancrage par seconde, 5 bandes et L=51), nous déterminons pour chaque morceau de référence le nombre de codes reconnus. Les morceaux candidats sont alors ceux ayant le plus de détections. Voir [1] pour plus de détails sur ces nombres.

Remarquons que pour de bonnes performances, la répétition d'un code pour un morceau de la base de référence n'est pas prise en compte dans le calcul, alors qu'elle l'est pour le flux à reconnaître. Par exemple, si un code est répété 10 fois dans un morceau de référence et le même code est répété 5 fois dans le flux à reconnaître, alors le compteur de ce morceau est incrémenté de 5, et non de 10 ou de 50.

Grâce à l'utilisation d'une table de hachage, index inversé, cette opération est très rapide. Si la table est entièrement chargée en mémoire vive, le calcul du nombre de codes détectés, pour tous les 755,762 extraits, ne dure pas plus d'une seconde pour un flux de 7 secondes à reconnaître.

Avec des codes de 24 bits, la table de hachage contient 2^{24} = 16,777,216 entrées. Remarquons que grâce à la décorrélation faite lors de la réduction des empreintes sonores les tailles des cases de la table, *buckets*, ont une faible variance, c'est-à-dire qu'elles ont toutes des tailles relativement proches d'une valeur moyenne. Avec 755,762 extraits d'environ 30 secondes, à raison de 4 points d'ancrage par seconde en moyenne, et 5L' = 50 codes par point d'ancrage, la table contient un total de 755762×30×4×50 ≈ 4 milliards d'identifiants, soit en moyenne 270 identifiants par case. En codant chaque identifiant sur 32 bits, la table de hachage occupe environ 16 giga octets.

Remarquons aussi que 1 minute de son dans la base de référence occupe 60×4×50×32/8 ≈ 48 kilo octets. Il s'agit donc bien d'une représentation condensée du signal musical, puisqu'une minute de son encodé en MP3 à 128kpbs occupe environ 960 kilo octets.

3.2 Seconde phase

Alors que la première phase ne s'intéresse uniquement au nombre de codes détectés pour chaque extrait de référence, la seconde phase prend compte de la cohérence temporelle des correspondances de codes. Puisque cette étape est plus longue à réaliser, elle se réalise sur un petit nombre de candidats.

Dans l'évaluation réalisée durant ce travail, cette sélection se fait de la manière suivante : avec M le nombre maximal de codes détectés pour un morceau de référence, les candidats sont alors ceux ayant un minimum de M/2 codes

communs avec le flux à reconnaître. Ce ratio de ½ a été déterminé empiriquement, et conduit visiblement à de bons résultats. Cependant, nous imposons un nombre minimal de 10 candidats et maximal de 500 ; la sélection se faisant toujours par ordre décroissant du nombre de codes détectés.

Parmi les candidats sélectionnés, la recherche du meilleur morceau de référence se fait alors par cohérence temporelle des codes reconnus via la méthode basée sur la cohérence temporelle, cf. [1]. Rappelons qu'en plus de renvoyer le candidat qui a le plus de chance d'être le morceau à reconnaître, cette méthode renvoie aussi le temps de départ estimé du flux à reconnaître.

Pour réaliser cette phase, il est nécessaire de connaître pour un candidat donné, les codes de hachage ainsi que l'indice temporel où ils ont été produits. Ces informations sont stockées dans la seconde table, *index direct*. Afin de gagner de la place en mémoire, nous avons stocké dans cette table les codes binaires originaux de K bits, et non les L, ou L', codes de hachage obtenus par LSH. Les L codes de hachage sont facilement obtenus pour les comparaisons.

Avec 755,762 morceaux de références de 30 secondes en moyenne, à raison de 4 points d'ancrage par seconde, 5 bandes de fréquences, et en stockant les codes de K=40 bits sur 64 bits, et les indices temporels sur 32 bits, la seconde table occupe : 755762×30×4×(5×64 + 32)/8 ≈ 4 giga octets au total, ou bien environ 10 kilo octets par minutes.

3.3 Temps de calcul

Lors de l'évaluation, nous avons fait une estimation du temps de calcul de chaque étape. Cette estimation a été faite sur le serveur de l'IRCAM *inverno*, qui possède un processeur *Intel Xeon CPU E5-2630* à 2.3 GHz, et grâce au 128Go de la mémoire vive, les deux tables ont été chargée dans leur totalité pour un total d'environ 20Go.

A partir d'un flux de 7 secondes à reconnaître, le calcul complet des points d'ancrage, empreintes sonores et des codes LSH, prend en moyenne 120 millisecondes. Le calcul des nombres de codes reconnus de tous les morceaux de références, étape 1, prend environ 890 millisecondes, et la seconde étape prend en moyenne 57 millisecondes pour chaque candidat.

4 Procédure d'évaluation

Après l'analyse complète des 755,762 morceaux de références sélectionnés, pour la construction des deux tables, nous avons lancé une évaluation de la reconnaissance en présence d'altérations sonores. Nous avons ici testé 18 types d'altérations sonores, et chacune avec 3 forces d'altérations : la plus faible correspond à de faibles dégradations, comme il est souvent testé dans la littérature. La force intermédiaire correspond à des altérations relativement fortes compatibles avec la réalité, et la dernière correspond à des dégradations exagérément fortes. Remarquons que toutes ces altérations ont été produites grâce à la librairie BeeAlter, cf. [2]. Voici la liste des types de dégradations testées.

a) Equaliseur graphique 10 bandes, avec gain alterné : $\pm\alpha$ dB. Les valeurs de α valent pour chacune des 3 forces : 3, 6 et 9 dB. Remarquons que pour la force extrême, il y a des écarts 18 dB dans la réponse.

b) Bruit environnemental enregistré de *restaurant*, avec pour rapport signal-à-bruit (RSB ou SNR) : 12dB, 6dB et 0dB, pour chacune des 3 forces testées respectivement. Remarquons que pour la force extrême, le bruit ajouté a la même énergie que le signal musical à reconnaître.

c) Bruit environnemental enregistré de *bus*, avec les mêmes RSB.

d) Bruit environnemental enregistré de *rue*, avec les mêmes RSB.

e) Bruit blanc synthétique, avec les même RSB.

f) Bruit rose synthétique, avec les même RSB.

g) Transposition des fréquences vers les aigus. Nous utilisons ici le logiciel *SuperVP* avec préservation des formants et des attaques, et avec pour transposition : ¼ ton, ½ ton et 1 ton, respectivement pour chacune des 3 forces.

h) Transposition des fréquences vers les grave. Il s'agit du même principe, mais avec une transposition vers le bas : -¼ ton, -½ ton et -1 ton.

i) Dilatation temporelle (*ralentissement*). Ici aussi nous utilisons le logiciel SuperVP avec les mêmes options, mais pour changer l'échelle de temps, avec des valeurs de : 15 %; 30 % et 45 % respectivement.

j) Compression temporelle (accélération). Il s'agit du même principe, mais pour une accélération du temps, avec pour ratio : -15 %; -30 % et -45 % respectivement.

k) Encodage MP3. Via le logiciel *lame*, cette dégradation consiste à simuler la dégradation d'un encodage MP3. Nous avons testé ici les taux : 32kbps, 24kbps, et 16kbps.

l) Distorsion. Il s'agit ici d'altérer le signal via une caractéristique non-linéaire du type *arctan*. La force est donnée par un gain préalable qui vaut : 5dB, 12dB et 24dB respectivement.

m) Compression 4-bandes des dynamiques. Pour régler la force, nous avons agit simultanément sur le ratio et les temps d'attaque/relâche. Valeurs des ratio : 2, 8 et 50 ; valeurs des temps caractéristiques : 100ms, 10ms, et 1ms.

n) Trémolo de modulation 4Hz, et d'amplitude : ±3db, ±6db, ±9db.

o) Réverbération par convolution avec une réponse impulsionnelle d'un hall de taille moyenne, pré-enregistrée. Ici la force est réglée par le ratio entre la réverbération et le signal direct : 9dB, 3dB, et 0dB.

p) Scenario GSM. Il s'agit d'une succession d'altérations représentant un scénario donné : filtrage par la réponse impulsionnelle d'un microphone de téléphone (*NexusOne*), ajout de bruit blanc (RSB de 18dB, 12dB et 6dB respectivement selon la force), et encodage au format GSM qui est ici la principale dégradation testée (sans paramètre).

q) Scenario ralentissement. Il s'agit aussi d'une succession d'altérations : ralentissement temporel avec pour ratio 4 %, 8 % et 12 % selon la force, égalisation 10 bandes à ±3dB, compresseur 4-bandes avec ratio de 2, encodage MP3 à 32kbps, réverbération à 3dB et ajout de bruit de restaurant avec un RSB à 18dB.

r) Scenario bruit. Ce scenario comprend plusieurs dégradations à la suite, avec un test principal sur le bruit ajouté. Les dégradations sont : ralentissement à 4%, égalisation 10 bandes à ±3dB, compresseur 4-bandes avec ratio de 2, encodage MP3 à 32kbps, réverbération à 3dB et ajout de bruit de restaurant avec un RSB à 18dB, 12dB et 6dB selon la force.

Le flux à reconnaître est alors dégradé 18×3=54 fois, voir les paramètres précédemment donnés, et le recherche, décrite en section précédente, est lancée pour chaque dégradation. Ce test étant répété quelques milliers de fois sur différents morceaux, la section suivante donne le taux de réussite, c'est-à-dire le pourcentage de flux correctement reconnu.

5 Résultats

Deux évaluations ont été réalisées : la première est faite sur la base complète de 755,762 morceaux de références, et la seconde sur un sous-ensemble de 42,040 morceaux. En effet, plus la taille de la base est faible, plus la reconnaissance est facile. Afin de comparer nos résultats avec ceux de la littérature qui utilise en général une base de quelques milliers de morceaux seulement, nous avons donc testé notre méthode sur une base volontairement réduite. Alors que la première démontre la possibilité de passer à l'échelle, la seconde permet de montrer de bonnes performances en comparaison avec la littérature.

5.1 Données affichées

Dans les sections suivantes nous donnons pour deux évaluations similaires les données suivantes :

- ✔ *Bit Error Rate* (BER) : il s'agit du taux d'erreur moyen des K (=40) bits formant les codes binaires avant le LSH. Notons que dans les cas d'altérations extrêmes, où les codes dégradés sont indépendant des codes orignaux, le taux moyen est statistiquement égal à 0.5. La valeur du BER est donc comprise entre 0, pour une absence de dégradations des codes, et 0.5 pour une dégradation extrême.

- ✔ Taux de réussite de l'étape 1 (*STEP 1 SUCCESS*) : il s'agit pour l'étape 1 du pourcentage de tests réussis, où le morceau ayant le maximum de

correspondances de codes de hachage, parmi tous les morceaux de référence de la base, est bien celui utilisé pour le flux dégradé à reconnaître.

- ✔ Intervalle de confiance à 95 % (*int95*) : il s'agit de la valeur du rang, donné par ordre décroissant du nombre de codes détectés, qui contient 95 % des flux testés. Par exemple, pour une valeur de 6, cela signifie que dans 95 % des tests, le morceau cible fait parti des 6 morceaux de la base ayant les plus forts nombres de correspondances de codes.

- ✔ Taux d'échec (*fail*) : la valeur affichée correspond au pourcentage de tests pour lesquels le morceau cible n'est pas parmi les candidats sélectionnés pour l'étape 2.

- ✔ Taux de réussite de l'étape 2 (*STEP 2 SUCCESS*) : il s'agit du pourcentage de tests réussis pour l'étape 2. Cette fois-ci un test est considéré comme réussi si le morceau cible fait parti des candidats et si il a le meilleur score lors de l'étape 2.

- ✔ Intervalle de confiance à 95 % (*int95*) : il s'agit aussi de l'intervalle à 95 %, mais cette fois-ci nous considérons le rang par ordre décroissant des scores de l'étape 2. Remarquons que si le taux d'échec, *fail*, est supérieur à 5 %, alors l'intervalle de confiance à 95 % ne peut pas être donné, il est alors affiché le symbole : *Inf.*

- ✔ Erreur de positionnement temporel (*dt*) : cette valeur est l'écart moyen en valeur absolue entre l'indice temporel de départ du flux testé et l'indice temporel détecté par l'algorithme de l'étape 2, affichée en millisecondes. Remarquons que seuls les tests ayant réussis sont utilisés pour cette estimation.

5.2 Évaluation sur la base complète de 755,762 extraits

Lors de la première évaluation faite sur la base complète de 755,762 extraits de référence, nous avons testé 7,146 flux audio dégradés de 7 secondes pris à partir de la 7^{ième} seconde d'un extrait de référence. Les valeurs des résultats, pour chacun des 18 types de dégradation et chacune des 3 forces, sont données en tableaux tab. 1 et tab. 2.

	BER	STEP 1 SUCCESS	int95	fail	STEP 2 SUCCESS	int95	dt
Equalizer					10 bands equalizer: ±3dB, ±6dB, ±9dB		
force 1	0.10	**98.7** %	1	0.0 %	**98.7** %	1	13.4 ms
force 2	0.18	**96.4** %	1	0.1 %	**97.7** %	1	13.4 ms
force 3	0.24	**71.2** %	154	5.7 %	**87.9** %	Inf	13.4 ms
Background noise: Restaurant					SNR: 12dB, 6dB, 0dB		
force 1	0.14	**98.5** %	1	0.1 %	**98.6** %	1	13.4 ms
force 2	0.24	**96.8** %	1	0.4 %	**97.0** %	1	13.4 ms
force 3	0.35	**65.4** %	11642	12.1 %	**68.8** %	Inf	13.2 ms
Background noise: Bus					SNR: 12dB, 6dB, 0dB		
force 1	0.11	**98.5** %	1	0.1 %	**98.6** %	1	13.4 ms
force 2	0.19	**96.7** %	1	0.5 %	**97.4** %	1	13.4 ms
force 3	0.30	**70.9** %	3058	11.3 %	**81.0** %	Inf	13.2 ms
Background noise: Street					SNR: 12dB, 6dB, 0dB		
force 1	0.12	**98.5** %	1	0.1 %	**98.5** %	1	13.4 ms
force 2	0.21	**96.5** %	1	0.6 %	**97.0** %	1	13.4 ms
force 3	0.32	**67.5** %	18721	12.7 %	**71.5** %	Inf	13.1 ms
White noise					SNR: 12dB, 6dB, 0dB		
force 1	0.10	**98.7** %	1	0.1 %	**98.9** %	1	13.4 ms
force 2	0.16	**98.4** %	1	0.2 %	**98.5** %	1	13.4 ms
force 3	0.25	**92.3** %	4	2.2 %	**95.8** %	1	13.4 ms
Pink noise					SNR: 12dB, 6dB, 0dB		
force 1	0.10	**98.6** %	1	0.1 %	**98.7** %	1	13.4 ms
force 2	0.18	**96.9** %	1	0.6 %	**97.6** %	1	13.4 ms
force 3	0.28	**74.5** %	469	11.0 %	**84.5** %	Inf	13.3 ms
Pitch Shifting plus					SuperVP: $+\frac{1}{4}$, $+\frac{1}{2}$, +1 ton,		
force 1	0.14	**98.1** %	1	0.0 %	**97.9** %	1	13.4 ms
force 2	0.18	**95.5** %	1	0.1 %	**97.0** %	1	13.4 ms
force 3	0.24	**75.2** %	80	3.7 %	**88.2** %	7	13.5 ms
Pitch Shifting minus					SuperVP: $-\frac{1}{4}$, $-\frac{1}{2}$, -1 ton,		
force 1	0.13	**98.2** %	1	0.0 %	**98.1** %	1	13.4 ms
force 2	0.18	**95.9** %	1	0.1 %	**97.2** %	1	13.4 ms
force 3	0.24	**77.4** %	55	3.1 %	**89.2** %	4	13.5 ms
Time Stretching slower					SuperVP: +15%, +30%, +45%		
force 1	0.19	**97.3** %	1	0.0 %	**97.4** %	1	13.4 ms
force 2	0.26	**93.5** %	2	0.4 %	**95.4** %	1	13.4 ms
force 3	0.32	**82.5** %	16	1.7 %	**86.4** %	4	13.3 ms
Time Stretching faster					SuperVP: -15%, -30%, -45%		
force 1	0.19	**97.6** %	1	0.0 %	**97.8** %	1	13.4 ms
force 2	0.27	**95.3** %	1	0.1 %	**96.1** %	1	13.4 ms
force 3	0.32	**91.5** %	2	0.9 %	**87.8** %	3	13.3 ms

Tab. 1: Résultats de l'évaluation de l'identification (10 premières dégradations)

	BER	STEP 1 SUCCESS	int95	fail	STEP 2 SUCCESS	int95	dt
MP3					Bitrate: 32kbps, 24kbps, 16kbps		
force 1	0.05	**98.9** %	1	0.0 %	**98.8** %	1	13.4 ms
force 2	0.08	**98.8** %	1	0.0 %	**98.8** %	1	13.4 ms
force 3	0.13	**98.5** %	1	0.1 %	**98.5** %	1	13.4 ms
Distortion					Pre-gain: 5dB, 12dB, 24dB		
force 1	0.13	**98.0** %	1	0.0 %	**98.0** %	1	13.4 ms
force 2	0.19	**95.9** %	1	0.2 %	**96.9** %	1	13.4 ms
force 3	0.25	**82.4** %	46	3.0 %	**89.0** %	5	13.3 ms
Multi-band Compressor					4 bands, ratio: 2, 8, 50, time: 100ms, 10ms, 1ms		
force 1	0.08	**98.5** %	1	0.0 %	**98.5** %	1	13.4 ms
force 2	0.20	**92.2** %	2	0.8 %	**95.5** %	1	13.3 ms
force 3	0.24	**80.4** %	178	4.8 %	**88.1** %	59	13.2 ms
Tremolo					Modulation 4Hz, depth: ±3dB, ±6dB, ±9dB		
force 1	0.09	**98.7** %	1	0.0 %	**98.7** %	1	13.4 ms
force 2	0.15	**97.9** %	1	0.2 %	**98.0** %	1	13.4 ms
force 3	0.20	**92.6** %	2	1.8 %	**95.4** %	1	13.4 ms
Reverberation					Reverberation gain: -9dB, -3dB, 0dB		
force 1	0.10	**98.8** %	1	0.0 %	**98.7** %	1	13.4 ms
force 2	0.18	**97.5** %	1	0.0 %	**97.8** %	1	13.4 ms
force 3	0.23	**88.3** %	4	0.5 %	**95.8** %	1	13.4 ms
Scenario GSM					Filter, white noise: 18, 12, 6dB; GSM encoding		
force 1	0.24	**88.6** %	6	1.3 %	**95.3** %	1	13.4 ms
force 2	0.25	**86.3** %	11	1.9 %	**94.0** %	2	13.4 ms
force 3	0.27	**77.2** %	54	4.3 %	**90.9** %	3	13.4 ms
Scenario speed					Speed: 4, 8, 12%; equalizer, compressor, MP3, reverb, noise		
force 1	0.20	**93.5** %	2	0.3 %	**96.6** %	1	13.4 ms
force 2	0.23	**83.1** %	18	1.9 %	**93.0** %	2	13.4 ms
force 3	0.27	**61.9** %	321	7.1 %	**80.3** %	Inf	13.5 ms
Scenario noise					Speed, equalizer, compressor, MP3, reverb, noise 18, 12, 6dB		
force 1	0.20	**93.6** %	2	0.3 %	**96.5** %	1	13.4 ms
force 2	0.22	**90.3** %	3	0.9 %	**95.2** %	1	13.4 ms
force 3	0.28	**76.6** %	69	3.7 %	**86.6** %	8	13.4 ms

Tab. 2: Résultats de l'évaluation de l'identification (8 dernières dégradations)

Nous observons d'assez bons résultats. En particulier avec l'encodage MP3, la distorsion, le changement d'échelle temporelle et l'ajout de bruit synthétique. En revanche l'ajout de bruit environnemental donne plutôt de mauvais taux de reconnaissance avec un RSB de 0dB, mais donne de très bons résultats pour des RSB plus élevés. Aussi, malgré nos efforts pour la robustesse, l'identification rencontre des difficultés avec la transposition des fréquences, puisque pour une transposition de seulement 1 ton, le taux de succès n'est que de 80 % environ.

Ces tableaux montrent par ailleurs l'intérêt d'utiliser la cohérence temporelle en deuxième phase. Par exemple, pour le scénario GSM avec les altérations les plus fortes, alors que le taux de reconnaissance n'est que de 77 % à l'étape 1, il monte à 90 % en étape 2. De manière générale, le taux de l'étape 2 est bien supérieur à celui l'étape 1, excepté pour l'accélération du temps.

Enfin, on peut noter la très bonne précision de l'estimation du temps de départ qui est en moyenne de 13.4ms, plus faible que le pas d'avancement du spectrogramme qui est de 20ms.

Remarquons que le BER est probablement faussé dans le cas du changement d'échelle temporelle, en raison d'un problème de synchronisation des temps d'analyse pour l'évaluation.

5.3 Évaluation par genre musical

En comparant les résultats par genre musicaux nous observons que les meilleurs résultats apparaissent pour la musique classique et les moins bons pour le rock. Une des raisons possibles est du fait que la base de référence contient plus de 40 % de morceaux classique et seulement 7.4 % de rock. Ainsi peut-être la phase d'apprentissage pour la réduction des empreintes, cf. [1], a privilégié les directions de faibles variations temporelles pour favoriser les variations fréquentielles. En effet, la musique rock a pour propriété d'être en général plus *rythmée* que la musique la classique, et on observe alors dans la représentation de Fourier en 2D, de dimension 1056, de plus fortes valeurs pour les fréquences du log-temps.

La liste qui suit donne le pourcentage de chaque genre dans la base complète :

- AMBIANCE 0.4 %
- BLUES 0.5 %
- CLASSIQUE 40.1 %
- COUNTRY 0.1 %
- TECHNO / DANCE 1.8 %
- ENFANT 0.9 %
- JAZZ 7.7 %
- METAL 0.6 %
- MONDE 1.7 %
- SOUL / FUNK 0.6 %
- ROCK 7.5 %
- BOF 1.8 %
- VARIETE / POP 14.3 %
- VARIETE 18.3 %
- RAP / HIP HOP 2.0 %
- RAGGAE 0.6 %
- BLUES GOSPEL 0.2 %

Remarquons que nous avons pu vérifier que d'une part l’ensemble utilisé pour l'apprentissage de la réduction et d'autre part les flux utilisés pour les tests de l'évaluation, sont représentatifs de la base puisque chaque genre musical apparaît dans les mêmes proportions.

Pour observer les résultats style par style, les deux tableaux suivant donnent les résultats en ne considérant uniquement les flux d'un style donné. Notons que la recherche se fait malgré tout sur la base complète contenant tous les genres musicaux. Pour chaque style, chaque dégradation et chaque force, les tableaux donnent les taux de succès à l'étape 1 et à l'étape 2.

	CLASSIQUE		ROCK, METAL		JAZZ		VARIETE / POP		VARIETE	
Equalizer							10 bands equalizer: ±3dB, ±6dB, ±9dB			
force 1	98.9	98.8	98.5	98.3	98.4	99.0	99.5	99.5	97.3	97.8
force 2	97.1	98.2	94.9	97.9	95.0	96.3	96.8	97.6	94.5	95.8
force 3	78.6	92.6	51.4	77.4	78.3	90.9	61.4	82.2	63.7	81.8
Background noise: Restaurant							SNR: 12dB, 6dB, 0dB			
force 1	98.5	98.6	97.7	97.7	99.0	98.9	99.5	99.1	97.2	97.4
force 2	96.7	97.2	96.8	96.6	96.1	96.6	98.1	98.1	95.7	95.7
force 3	71.0	76.3	46.1	47.6	66.8	70.5	64.0	65.8	62.5	63.8
Background noise: Bus							SNR: 12dB, 6dB, 0dB			
force 1	98.4	98.5	98.5	98.3	98.1	98.9	99.8	99.5	97.4	97.3
force 2	96.8	97.7	95.8	96.4	95.5	96.5	98.3	98.3	95.8	96.0
force 3	72.8	84.4	63.9	72.3	63.9	76.1	75.1	81.7	67.8	77.4
Background noise: Street							SNR: 12dB, 6dB, 0dB			
force 1	98.3	98.5	98.3	98.3	98.7	98.7	99.4	99.4	97.4	97.4
force 2	96.6	97.4	95.3	95.4	95.8	97.0	97.9	97.7	94.9	95.1
force 3	72.3	78.2	53.5	56.9	66.0	68.9	68.1	69.8	61.1	65.3
White noise							SNR: 12dB, 6dB, 0dB			
force 1	98.7	98.7	98.7	98.7	99.0	99.5	99.7	99.7	97.3	97.8
force 2	98.3	98.4	97.5	98.1	98.7	99.0	99.4	99.5	97.3	97.3
force 3	94.8	97.1	82.5	92.4	92.9	95.7	93.6	97.0	88.4	92.5
Pink noise							SNR: 12dB, 6dB, 0dB			
force 1	98.5	98.6	98.7	98.9	98.6	99.0	99.5	99.4	97.7	97.7
force 2	97.0	97.7	95.1	97.7	95.7	96.5	97.8	98.5	96.0	96.3
force 3	82.2	91.5	54.3	64.7	71.4	83.3	71.3	82.7	67.9	76.7
Pitch Shifting plus							SuperVP: $+\frac{1}{4}$, $+\frac{1}{2}$, +1 ton,			
force 1	98.3	98.1	96.6	95.8	97.8	97.8	98.5	98.5	97.4	96.9
force 2	95.2	97.3	95.8	95.8	93.9	95.5	97.0	97.4	94.4	96.5
force 3	71.4	87.5	71.3	84.4	77.2	87.3	82.9	91.1	78.5	88.9
Pitch Shifting minus							SuperVP: $-\frac{1}{4}$, $-\frac{1}{2}$, -1 ton,			
force 1	98.4	98.4	97.7	96.8	97.8	97.1	98.8	98.8	97.1	97.2
force 2	95.9	97.4	94.3	96.2	93.6	95.3	97.0	97.7	95.4	96.7
force 3	74.6	89.5	72.5	85.6	78.7	86.8	83.4	90.4	80.6	89.0
Time Stretching slower							SuperVP: +15%, +30%, +45%			
force 1	98.1	98.1	95.6	96.0	96.5	96.1	97.3	97.5	95.7	95.8
force 2	96.9	97.5	89.2	93.9	92.8	95.5	91.3	93.0	88.9	92.2
force 3	91.1	94.4	71.2	79.5	83.1	89.1	72.2	77.1	74.6	78.1
Time Stretching faster							SuperVP: -15%, -30%, -45%			
force 1	98.3	98.2	96.6	96.2	97.0	97.6	97.6	98.1	96.3	96.4
force 2	97.4	97.5	92.4	95.1	95.5	96.6	92.8	95.0	93.5	93.7
force 3	96.7	95.1	85.0	78.4	92.6	89.1	86.4	81.1	85.8	79.3

Tab. 3: Résultats de l'évaluation style/style (10 premières dégradations)

	CLASSIQUE		ROCK, METAL		JAZZ		VARIETE / POP		VARIETE	
MP3							Bitrate: 32kbps, 24kbps, 16kbps			
force 1	98.8	98.8	98.7	98.3	99.4	99.0	99.7	99.7	98.1	98.0
force 2	98.7	98.7	98.1	98.5	98.9	99.2	99.7	99.7	97.8	97.6
force 3	98.5	98.5	97.9	98.5	99.2	99.0	99.6	99.0	97.2	97.2
Distortion							Pre-gain: 5dB, 12dB, 24dB			
force 1	98.3	98.3	97.7	97.7	97.9	97.4	99.0	98.5	96.5	96.8
force 2	96.4	97.6	97.2	96.4	93.4	96.3	96.2	96.7	94.4	95.1
force 3	84.7	92.2	83.1	88.8	76.9	83.6	84.2	89.2	79.7	83.4
Multi-band Compressor							4 bands, ratio: 2, 8, 50, time: 100ms, 10ms, 1ms			
force 1	98.7	98.7	98.5	98.5	98.4	98.6	99.1	99.1	97.3	96.9
force 2	91.9	96.5	95.4	95.4	91.3	93.7	94.0	94.6	90.3	93.6
force 3	82.2	91.1	87.3	90.7	77.4	85.6	82.1	86.9	76.2	82.2
Tremolo							Modulation 4Hz, depth: ±3dB, ±6dB, ±9dB			
force 1	98.7	98.7	98.3	98.1	99.4	99.4	99.6	99.5	97.4	97.2
force 2	98.3	98.4	96.8	97.2	98.6	98.2	98.2	98.3	96.0	95.9
force 3	95.3	97.1	78.7	88.2	95.3	96.8	90.4	94.3	89.0	92.7
Reverberation							Reverberation gain: -9dB, -3dB, 0dB			
force 1	98.6	98.6	98.3	98.3	99.0	99.0	99.7	99.4	97.8	97.6
force 2	97.4	98.0	96.8	97.0	96.6	97.9	98.3	98.3	96.9	96.3
force 3	89.2	96.5	86.9	94.7	83.3	94.9	87.9	96.3	87.1	94.0
Scenario GSM							Filter, white noise: 18, 12, 6dB; GSM encoding			
force 1	88.3	95.5	80.5	93.2	89.9	95.0	89.7	94.7	89.8	95.1
force 2	85.2	94.2	76.7	89.9	88.6	93.6	88.9	95.1	87.5	93.1
force 3	76.5	91.5	65.3	84.3	78.5	91.3	80.9	91.8	77.9	89.8
Scenario speed							Speed: 4, 8, 12%; equalizer, compressor, MP3, reverb, noise			
force 1	93.1	97.0	93.2	96.8	91.8	95.3	95.2	96.3	93.2	95.8
force 2	81.0	93.8	80.6	91.3	79.6	90.2	88.2	93.5	84.9	91.2
force 3	55.8	79.1	59.2	76.7	63.4	79.3	71.2	84.1	68.5	81.3
Scenario noise							Speed, equalizer, compressor, MP3, reverb, noise 18, 12, 6dB			
force 1	93.1	96.9	92.4	96.4	92.1	95.7	95.4	96.7	94.0	95.4
force 2	89.2	95.5	89.9	94.3	87.6	93.9	92.4	95.4	91.6	94.5
force 3	74.7	87.3	72.9	82.4	73.7	85.4	82.5	88.3	78.5	85.2

Tab. 4: Résultats de l'évaluation style/style (8 dernières dégradations)

5.4 Évaluation sur une sous-base de 42,040 extraits

Dans la littérature, la plupart des évaluations sont basés sur des ensembles de morceaux de petites tailles en général ; de l'ordre de quelques milliers, voir quelques dizaines de milliers. Le fait d'utiliser une petite base améliore artificiellement les résultats. L'évaluation précédente montre de relativement bons résultats obtenus avec passage à l'échelle, ici nous donnons les résultats sur une base plus petites, de 42,040 morceaux, afin de faire une comparaison avec la littérature. Dans cette nouvelle expérience, 9,293 flux audio représentatifs ont été testés avec la même procédure que précédemment. Le tableau Tab. 5 donne les résultats.

	STEP 1	STEP 2
Equalizer		
force 1	**99.8** %	**99.8** %
force 2	**99.4** %	**99.7** %
force 3	**84.4** %	**95.4** %
Background noise: Restaurant		
force 1	**99.8** %	**99.8** %
force 2	**99.1** %	**99.3** %
force 3	**77.1** %	**81.6** %
Background noise: Bus		
force 1	**99.8** %	**99.8** %
force 2	**98.9** %	**99.3** %
force 3	**81.2** %	**88.9** %
Background noise: Street		
force 1	**99.8** %	**99.8** %
force 2	**98.9** %	**99.0** %
force 3	**76.1** %	**81.7** %
White noise		
force 1	**99.8** %	**99.9** %
force 2	**99.6** %	**99.7** %
force 3	**97.5** %	**98.7** %
Pink noise		
force 1	**99.8** %	**99.8** %
force 2	**99.3** %	**99.4** %
force 3	**85.3** %	**91.7** %
Pitch Shifting plus		
force 1	**99.7** %	**99.8** %
force 2	**99.1** %	**99.4** %
force 3	**89.6** %	**96.2** %
Pitch Shifting minus		
force 1	**99.7** %	**99.8** %
force 2	**99.2** %	**99.6** %
force 3	**91.1** %	**97.0** %
Time Stretching slower		
force 1	**99.5** %	**99.5** %
force 2	**97.7** %	**98.7** %
force 3	**92.6** %	**94.8** %
Time Stretching faster		
force 1	**99.6** %	**99.6** %
force 2	**98.5** %	**99.0** %
force 3	**96.5** %	**95.2** %

	STEP 1	STEP 2
MP3		
force 1	**99.9** %	**99.9** %
force 2	**99.9** %	**99.9** %
force 3	**99.7** %	**99.7** %
Distortion		
force 1	**99.7** %	**99.7** %
force 2	**99.3** %	**99.5** %
force 3	**92.1** %	**95.6** %
Multi-band Compressor		
force 1	**99.8** %	**99.8** %
force 2	**97.6** %	**98.8** %
force 3	**89.7** %	**94.7** %
Tremolo		
force 1	**99.8** %	**99.8** %
force 2	**99.6** %	**99.7** %
force 3	**96.7** %	**98.5** %
Reverberation		
force 1	**99.9** %	**99.9** %
force 2	**99.7** %	**99.7** %
force 3	**96.6** %	**99.1** %
Scenario GSM		
force 1	**96.0** %	**98.7** %
force 2	**94.9** %	**98.5** %
force 3	**90.1** %	**97.0** %
Scenario speed		
force 1	**98.8** %	**99.3** %
force 2	**94.3** %	**98.2** %
force 3	**81.3** %	**93.3** %
Scenario noise		
force 1	**98.5** %	**99.3** %
force 2	**97.1** %	**98.9** %
force 3	**90.3** %	**95.4** %

Tab. 5: Résultat d'identification sur la sous-base

Sous certaines réserves liées aux procédures d'évaluation qui ne sont pas identiques, nous pouvons observer de très bons résultats par rapport à la littérature. Cela valide la méthode ici mise en œuvre.

6 Conclusion

L'évaluation, présentée dans ce document, montre non seulement la très bonne robustesse de notre méthode à différentes dégradations sonores ; mais aussi le possible passage à l'échelle avec un temps de réponse correct, et une utilisation de mémoire compatible avec la plupart des serveurs de calcul. Cela prouve la validité de la méthode mise en œuvre par rapport aux objectifs fixés en début de projet : indexation audio robuste aux dégradations et passage à l'échelle.

7 Références internes

[1] Rémi Mignot, "RAPPORT GLOBAL – Ircam AudioID: Indexation Audio avec Robustesse aux Dégradations Sonores," rapport interne IRCAM – CNRS, projet BeeMusic. Décembre 2015.

[2] Rémi Mignot, "BeeAlter Toolbox Boîte à Outils de Dégradations Sonores, pour Tests," rapport interne IRCAM – CNRS, projet BeeMusic. Décembre 2015.