Classification à grande échelle de morceaux de musique en fonction de la présence de chant

Yann Bayle, Pierre Hanna, Matthias Robine

CLASSIFICATION A GRANDE ECHELLE DE MORCEAUX DE MUSIQUE EN FONCTION DE LA PRESENCE DE CHANT

Yann Bayle *Pierre Hanna* *Matthias Robine*
LaBRI, CNRS, France
prenom.nom@labri.fr

RÉSUMÉ

Le chant est un élément remarquable d'une chanson et sa détection automatique au sein d'un morceau est un défi largement étudié. Cet article propose une approche permettant de discriminer les titres musicaux comportant du chant dans une base de données musicales conséquente.

L'approche précédemment proposée par Ghosal et al.[9] fonde la prise de décision sur l'analyse des descripteurs à l'échelle de la chanson. Nous générons ici une probabilité de présence de chant à l'échelle de la trame afin de prendre une décision globale. Une première méthode proposée pour cette classification utilise la densité de probabilité des prédictions et une seconde des n-grammes sur les trames supposées contenir du chant.

Les résultats de ces nouvelles méthodes améliorent ceux obtenus par [9] et montrent une meilleure robustesse lorsque la taille de la base musicale augmente. La précision de la classification chute ainsi de 3.6% seulement contre 13.1% pour [9] lorsque la base de test est multipliée par 16.

1. INTRODUCTION

Dans le domaine musical, la voix et le chant constituent des éléments essentiels d'une chanson. Plusieurs champs d'application requièrent de pouvoir distinguer les pistes contenant de la voix chantée des chansons purement instrumentales. Cette tâche est par exemple utile pour l'indexation et la description des grandes bases de données musicales [32]. Les plateformes de streaming en ligne requièrent notamment ce type d'informations afin de proposer à leurs utilisateurs une sous-partie de l'importante base de données musicales dont elles disposent. La classification de genres musicaux peut également utiliser l'information concernant la présence de voix dans les pistes musicales [19, 33]. Dans cet article, les morceaux instrumentaux sont définis par l'absence de sons produits par les cordes vocales humaines, provenant d'un enregistrement ou d'une synthèse par ordinateur.

La majeure partie des travaux qui tentent de détecter le chant utilisent des descripteurs à l'échelle de la chanson [9, 10, 21, 36]. Cette approche est cependant mise en difficulté lorsque des chansons strictement instrumentales contiennent des solos ou sont élaborées avec des instruments imitant la voix, tels que les guitares munies de wah-wah, les synthétiseurs et les didgeridoos. Par ailleurs, ces différentes méthodes ont été testées sur un nombre restreint d'extraits de pistes musicales. Le présent article étudie les performances de ces méthodes au sein d'une base de données musicales de taille plus conséquente.

De nombreux travaux ont été menés à l'échelle de la trame afin de distinguer les trames contenant de la voix de celles purement instrumentales [3, 12, 14, 16]. Ces travaux, menés afin d'étudier la présence de chant dans les trames, ont montré que certains descripteurs pouvaient être particulièrement pertinents [18, 35]. Au niveau temporel, les descripteurs tels que le taux de passage du signal sonore à zéro [4, 34] et l'énergie du signal à court terme [5, 29] ont été introduits. Les trois descripteurs principalement utilisés au niveau fréquentiel ont été la fréquence fondamentale [37], l'énergie spectrale du signal [1, 3] et les *Mel Frequency Cepstral Coefficients* (MFCC) [6, 8, 23]. D'autres descripteurs plus proches de la perception humaine ont aussi été utilisés, à savoir la mesure de sonie [7, 31] et l'énergie contenue dans les sous-bandes [12]. Parmi l'ensemble de ces descripteurs, les MFCC sont apparus comme étant les plus pertinents afin de détecter le chant dans une piste musicale [25, 27]. Plusieurs classifieurs tels que la Machine à Vecteurs de Support (SVM) [12, 24, 28], le Perceptron Multicouches (MLP) [9], Random Forest [17,18] et l'utilisation d'un seuillage [29] se sont également révélés adaptés à la détection du chant.

Les bases de données expérimentales utilisées pour les études présentées dans cet article sont décrites en section 2 et les limites de la méthode proposée par Ghosal *et al.* [9] vis-à-vis d'une base de données musicales conséquente sont étudiées en section 3. Une nouvelle approche est ensuite proposée en section 4 afin de répondre à ces limitations. Elle consiste à prendre des décisions concernant la présence de chant à l'échelle d'une trame, afin de conclure quant au contenu vocal d'une chanson. La justification du choix des descripteurs ainsi que la procédure de prise de décision globale sont également explicitées dans la section 4. La nouvelle méthode proposée est enfin comparée à l'existant dans la section 5 afin de montrer l'amélioration qu'elle apporte à la tâche de détection de morceaux chantés inclus dans de grandes bases de données.

2. BASES DE DONNEES

Dans cet article, deux bases de données sont utilisées afin de mener les expériences. La première est composée de 1080 pistes musicales et est utilisée pour paramétrer la nouvelle approche proposée. La seconde base est constituée de plus de 8000 pistes musicales et est utilisée pour les tests à plus grande échelle. Les deux bases sont composées d'environ 75 % de Pop/Rock, de 10 % de Blues/Jazz, de 10 % de Métal, de 5 % de Classique et de moins d'un pourcent de Reggae/Punk/Musique du monde. Les propriétés de ces deux bases sont décrites dans cette section.

2.1. Base de données utilisée pour le paramétrage

Cette base de données est composée de 540 pistes instrumentales (I) et de 540 pistes vocales (V). Elle comprend des morceaux ayant déjà été utilisés par d'autres études et des morceaux provenant d'une bibliothèque musicale personnelle. Le tableau 1 indique la provenance des morceaux et le nombre d'éléments de chaque classe I / V.

Nom	**Nombre de I**	**Nombre de V**
CCMixter	50	50
MedleyDB	100	63
Quasi	10	9
RWC	0	100
Jamendo	0	93
Personnel	380	225
Total	540	540

Tableau 1. Description des bases de données utilisées.

La base de données CCMixter a été mise à disposition par [30] afin d'être utilisée pour détecter le chant à l'échelle de la trame. Elle contient 50 morceaux de musique stéréo et trois pistes sont présentes pour chacun d'entre eux. La première piste comprend uniquement la voix, la deuxième piste contient uniquement l'accompagnement musical et la dernière piste mixe ces deux premières pistes. Seules les deux dernières pistes ont été utilisées dans cette étude.

La base de données MedleyDB [2] a été constituée afin de répondre à la problématique de séparation des différentes sources audio composant une œuvre musicale [20]. Cette base comprend 122 pistes audio et l'intégralité des pistes de chaque instrument et de chaque voix est fournie pour chacune d'elles. Certaines pistes ont pu être fusionnées afin d'obtenir 100 pistes contenant uniquement de la musique instrumentale. 63 pistes contenant de la musique instrumentale ainsi que du chant ont pu être constituées.

La base de données QUASI[1] est composée de onze pistes audio. Les différentes pistes individuelles sont proposées pour chaque morceau. Plusieurs effets musicaux, tels que de l'égalisation, de la compression ainsi que de la balance panoramique ont été appliqués à chaque piste. Un fichier sonore correspondant à chaque effet appliqué est disponible. Le mix final de toutes les pistes ainsi que de leurs effets est également fourni. Dix pistes instrumentales sont utilisées dans cette étude. Ces dernières sont le résultat de la fusion de différentes pistes instrumentales ainsi que de leurs effets. Parmi ces dix morceaux musicaux, neuf d'entre eux présentaient une piste vocale disponible dans la base de données. Cette piste vocale a donc été ajoutée au mix des chansons instrumentales afin de générer neuf nouvelles chansons.

La base de données RWC [11] comprend 100 chansons contenant de la musique instrumentale accompagnée de chants. Ces chansons ont été utilisées dans la nouvelle base de données sans avoir été modifiées.

La base de données Jamendo[2] contient 93 chansons provenant du site web de Jamendo.fr. Cette base de données est utilisée dans le cadre de la détection du chant à l'échelle de la trame. Les chansons sont ici composées de musique instrumentale ainsi que de voix chantée. Ces chansons sont utilisées sans modification dans la nouvelle base de données proposée.

Les morceaux provenant de la bibliothèque personnelle sont des fichiers stéréo au format WAVE, échantillonnés à 44.1 kHz, représentant cinq genres musicaux à savoir du Rock, du Blues, du Jazz, de l'Electro et du Classique ainsi que plusieurs artistes différents.

Des annotations ont été créées par Bernhard Lehner (communication personnelle) pour la base de données RWC et par [24] dans le cas de Jamendo, afin d'indiquer le contenu en chant de chaque trame. Ces annotations ont permis de créer la base d'apprentissage dans le cadre de la nouvelle approche proposée dans cet article.

2.2. Base de données de test

La base de données de test contient des pistes musicales obtenues de manière synthétique grâce à l'augmentation du nombre de données [30]. Onze modifications ont été appliquées à 782 pistes uniques qui n'ont pas été utilisées dans la base de données de paramétrage. Les fichiers audio, d'une durée comprise entre une et six minutes, proviennent d'une bibliothèque personnelle et sont au format WAVE, échantillonnés à 44.1 kHz en stéréo et représentent cinq genres musicaux, à savoir du Rock, du Blues, du Jazz, de l'Electro et du Classique ainsi que plusieurs artistes différents.

Type de modification	Valeur de modification
Vitesse	x0.5, x1.5, x2
Fréquence	+1 demi-ton, +7 demi-tons
Tempo	x0.5, x1.5, x2
Volume	-6dB, -3dB, +3dB

Tableau 2. Liste des modifications appliquées aux fichiers audio.

Les modifications, qui consistent à faire varier la vitesse, le tempo, le volume et la fréquence du morceau, ne sont pas cumulées. La version 2.1.1 du logiciel Audacity a été utilisée afin de réaliser ces modifications présentées dans le Tableau 2.

3. TEST DES PERFORMANCES DE LA METHODE EXISTANTE SUR UNE GRANDE BASE DE DONNEES

Ghosal *et al.* [9] ont récemment proposé un algorithme de classification des chansons I / V à partir d'une base de données comprenant 540 extraits de morceaux de musique, 270 contenant du chant, les autres composés uniquement de musique instrumentale. Ces extraits, d'une durée de 40 à 45 secondes, sont échantillonnés à 22050 Hz sous un format mono, 16-bits et comprennent trois différents styles musicaux à savoir du Classique, du Folk et du Rock. L'algorithme proposé par [9] calcule les treize premiers coefficients MFCC pour chaque trame d'une chanson, puis les moyenne dans le but d'obtenir treize descripteurs par morceau. Le logiciel Weka [13] est utilisé afin de classer les chansons à l'aide de trois classifieurs, à savoir SVM, MLP et *Random Sample Consensus* (RANSAC), dans un mode de validation croisée en deux temps.

Cette méthode a été reproduite afin de vérifier la robustesse de la précision de classification lorsque celle-ci est appliquée à une base de données plus importante, constituée de morceaux de musique entiers et non d'extraits musicaux. Afin d'implémenter l'algorithme de [9], la version 7.14.0.739 du logiciel Matlab ainsi que la MIRToolBox [15] sont utilisés. Le classifieur RandomForest du logiciel Weka [13] est utilisé en apprentissage sur une base de 500 morceaux musicaux. La Figure 1 représente l'évolution de la précision de l'algorithme de [9] en fonction de la taille de la base de données de test qui varie de 1000 à 8000 morceaux par pas de 1000 morceaux sélectionnés aléatoirement dans la grande base de données. Les résultats sont exprimés en pourcentage de précision. Celui-ci est défini comme le pourcentage total de pistes instrumentales et de pistes contenant de la voix qui ont été correctement classées.

La précision de classification de la méthode de Ghosal *et al.* [9] chute lorsque la taille de la base de données augmente. Cette diminution est de 13.1% lorsque la base de données augmente de 1000 (précision à 77.2 %) à 8000 pistes musicales (précision à 67.1 %). On pourrait supposer qu'en augmentant la taille de la base d'apprentissage, cette chute de précision ne serait pas observée, or il s'agit d'une condition nécessaire à l'application désirée. L'objectif de cet article est en effet de proposer une approche qui puisse être entraînée avec une base de données musicales de taille réduite, afin d'être utilisée à grande échelle sur la base musicale d'un site de streaming par exemple. La diminution de la précision observée provient de la réduction d'informations provoquée lors du calcul de la moyenne des coefficients MFCC. La moyenne des coefficients MFCC n'est en effet pas en mesure de discriminer la voix des instruments qui l'imitent, notamment lors d'un solo instrumental. Il est cependant nécessaire de conserver les coefficients MFCC calculés pour chaque trame afin que l'algorithme d'apprentissage puisse en extraire les informations pertinentes de distinction entre le chant et les instruments qui l'imitent.

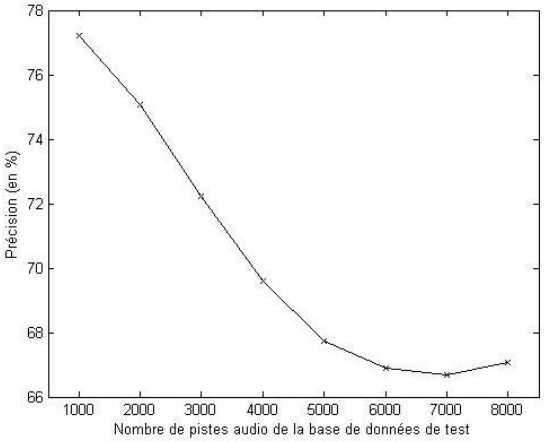

Figure 1. Evolution de la précision de l'algorithme de Ghosal *et al.* [9] en fonction de la taille de la base de données.

Une nouvelle approche, fondée sur une prise de décision concernant la présence de chant à l'échelle de la trame, est donc proposée dans la section suivante afin de pallier cette limite.

4. PROPOSITION D'UNE NOUVELLE APPROCHE DE CLASSIFICATION

Une nouvelle approche hiérarchique est proposée afin de réaliser la tâche de détection du chant. Cette approche détermine d'abord si chaque trame contient du chant et utilise ensuite les prédictions effectuées pour chaque trame afin de déterminer si l'intégralité du morceau musical contient du chant. Dans le cas idéal, toutes les trames sont classées correctement en fonction de la présence de chant et il est uniquement nécessaire de vérifier la présence d'une trame contenant du chant pour affirmer que la chanson en contient. Les approches qui tentent de prédire la présence de voix à l'échelle de la trame ne sont actuellement pas fiables à 100%, il est donc nécessaire d'élaborer une stratégie de prise de décision fondée sur les trames et qui prend en compte ce manque de précision.

Cet article propose des solutions afin de pallier ce problème. Il est d'abord nécessaire de constituer une base de vérité indiquant la présence du chant à l'échelle de la trame et ce pour un nombre important de morceaux musicaux. Cette base de vérité à l'échelle de la trame permet au modèle de classification de distinguer efficacement les trames vocales de celles qui contiennent des instruments capables d'imiter le chant. Les deux bases de données musicales utilisées afin de fournir cette base de vérité sont Jamendo [22] et RWC [13]. Pour chacune d'elles, les annotations de présence

du chant ont été générées à la main pour chaque trame de 200 ms, par [26] pour Jamendo et par Bernhard Lehner pour RWC.

Un modèle de classification à l'échelle de la trame est ensuite généré à partir de la base de vérité. Pour cela, les annotations concernant chacune des trames sont utilisées afin de constituer une base d'apprentissage à l'échelle de la trame. Le logiciel Weka permet de générer un tel modèle d'apprentissage. Il est donc possible de prédire la présence de chant dans chaque trame de chaque morceau musical de la base de données à l'aide du modèle généré par Weka. Une prédiction comprise entre 0 et 1 est fournie par ce logiciel pour chaque trame de chaque chanson et indique la probabilité pour qu'une trame contienne de la voix.

Pour chaque morceau, un fichier contenant le vecteur des prédictions pour chaque trame est créé. Les sections suivantes décrivent l'utilisation de ce vecteur dans le cadre de deux méthodes de prise de décision qui permettent de passer d'une classification à l'échelle de la trame à une classification à l'échelle du morceau.

4.1. Décision fondée sur la probabilité de présence de la voix chantée

La méthode présentée dans cette section utilise le vecteur de prédiction de chaque morceau contenant les probabilités de présence de voix pour chaque trame. L'hypothèse envisagée par cette méthode repose sur la présence de davantage de trames possédant une probabilité élevée de contenir du chant pour les morceaux vocaux. En ce qui concerne les morceaux instrumentaux, davantage de trames possédant une faible probabilité de contenir du chant sont attendues.

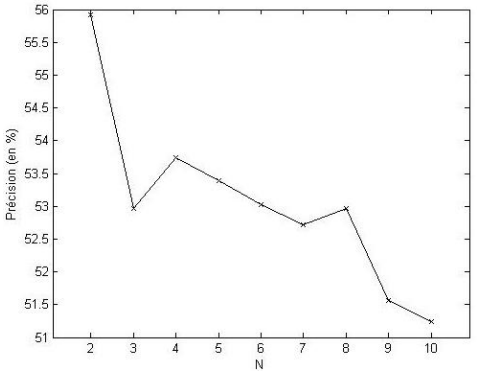

Figure 2. Analyse de l'impact du choix de la valeur de N sur la précision de classification des morceaux musicaux.

Un histogramme de dimension N représentant la répartition des probabilités entre 0.5 et 1 est créé pour chaque morceau et les probabilités inférieures ou égales à 0.5 ne sont pas prises en compte puisqu'elles ne sont pas censées contenir du chant. Le rapport entre la dernière valeur de l'histogramme et la somme des autres valeurs est alors utilisé pour informer la méthode de prise de décision de la présence de chant puisque ce rapport s'est révélé discriminant dans cette tâche. Le nombre de classes N considérées par l'histogramme a un impact sur la précision de classification.

C'est pourquoi l'évolution des performances de classification I / V en fonction des valeurs de N a été étudié, les résultats obtenus sont présentés dans la Figure 2. La précision est obtenue avec Weka en utilisant une validation croisée sur deux ensembles à l'aide du classifieur RandomForest.

La Figure 2 montre que les meilleures performances sont obtenues en utilisant un histogramme à 2 valeurs (N = 2). Il est intéressant de comparer les résultats obtenus avec l'utilisation de ce seul descripteur avec ceux obtenus pour l'ensemble des valeurs de N. Le Tableau 3 représente les résultats comparatifs de cette étude. Le premier cas affiche les résultats pour N = 2 et le second pour la combinaison des N compris entre 2 et 10. Pour chacun des deux cas, la précision pour quatre classifieurs utilisés dans le logiciel Weka est fournie en validation croisée en deux temps. Ces quatre classifieurs sont ceux décrits dans la Section 2 comme étant les plus appropriés pour la tâche de détection de chant. En plus des quatre classifieurs représentés dans le tableau, la valeur moyenne des quatre classifieurs pour chacun des cas est indiquée. La base de paramétrage est utilisée afin d'obtenir les résultats présentés dans le Tableau 3.

Classifieur	N = 2	N = 2 à 10
RandomForest	55.9	**67.4**
SVM	60.1	66.8
MLP	54.4	57.7
J48	58.2	64.1
Moyenne	57.2	64.0

Tableau 3. Comparaison du pourcentage de précision obtenu dans le cas où le nombre de classes de l'histogramme est de deux (N=2) et dans le cas où les classes de deux à dix sont combinées (N = 2 à 10) pour quatre classifieurs.

Le Tableau 3 indique qu'il existe d'importantes différences de performances en fonction du classifieur utilisé dans le cas où le nombre de classes de l'histogramme est de deux (N = 2) et dans le cas où les classes de deux à dix sont combinées (N = 2 à 10). Le descripteur MLP est en moyenne le descripteur le moins performant. De meilleures performances sont observées dans le second cas puisqu'elles sont en moyenne supérieures de 6.8% à l'utilisation d'un seul descripteur. Les résultats de cette méthode sont encourageants mais n'atteignent pas 68%, ce qui montre que cette méthode ne peut pas être utilisée comme telle. C'est pourquoi une autre méthode de prise de décision, utilisant différemment le vecteur de prédiction de chaque morceau, est proposée dans la section qui suit.

4.2. Décision utilisant des n-grammes de trames vocales

Cette section définit une méthode de prise de décision utilisant des n-grammes à partir du vecteur de prédiction des morceaux musicaux. Pour cela, les trames pour lesquelles la probabilité de présence du chant est comprise dans l'intervalle]0.5 ;1] sont sélectionnées et les probabilités inférieures ou égales à 0.5, qui dénotent une trame ne contenant pas de chant, ne sont pas prises en compte. Elles ne sont en effet pas discriminantes pour les deux classes I/V puisqu'elles contiennent toutes deux des parties instrumentales. Les trames qui ne sont pas censées contenir de la voix ne sont donc pas utilisées puisque la présence d'instruments n'est pas corrélée avec la présence de chant.

Un vecteur de n-grammes est créé et comptabilise le nombre de fois consécutives pour lesquelles une trame a été décelée comme contenant de la voix. Il est supposé ici qu'un morceau musical contenant du chant possédera davantage de trames vocales consécutives qu'un morceau strictement instrumental. Un histogramme de ces n-grammes est créé en accord avec cette réflexion. La dimension de l'histogramme est fixée à 30 valeurs par construction et tous les n-grammes de valeur supérieure sont ajoutés à la trentième valeur de l'histogramme. Un exemple d'histogramme des n-grammes obtenu pour une chanson est présenté en haut dans la Figure 3, tandis que le graphique du bas représente celui d'un morceau instrumental.

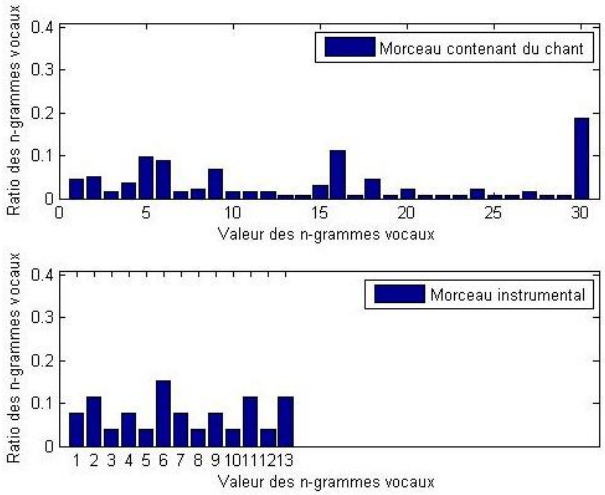

Figure 3. Histogrammes des n-grammes obtenus pour un morceau contenant de la voix (en haut) et pour un morceau instrumental (en bas).

Les histogrammes de n-grammes des morceaux contenant du chant sont différents de ceux obtenus pour des morceaux strictement instrumentaux car les morceaux contenant du chant possèdent davantage de n-grammes de haut rang que les précédents. Il est donc possible d'utiliser les valeurs des histogrammes en tant que descripteurs afin de différencier les morceaux instrumentaux de ceux contenant du chant. Un descripteur plus simple puisque composé d'une unique dimension peut par ailleurs être calculé en effectuant le rapport entre le nombre de n-grammes de bas rang et celui de n-grammes de haut rang. Dans ce dernier cas, le choix de la limite de séparation R entre des n-grammes de haut rang et ceux de bas rang impacte la précision de classification. Cet impact est donc étudié et illustré par la Figure 4 dans laquelle la précision de classification en pourcent est représentée en fonction de la valeur de R choisie. La précision est obtenue avec Weka en utilisant une validation croisée sur deux ensembles à l'aide du classifieur RandomForest. La figure 4 affiche des valeurs jusqu'à R = 40, dans le but de faciliter l'analyse de la tendance de la précision en fonction de R.

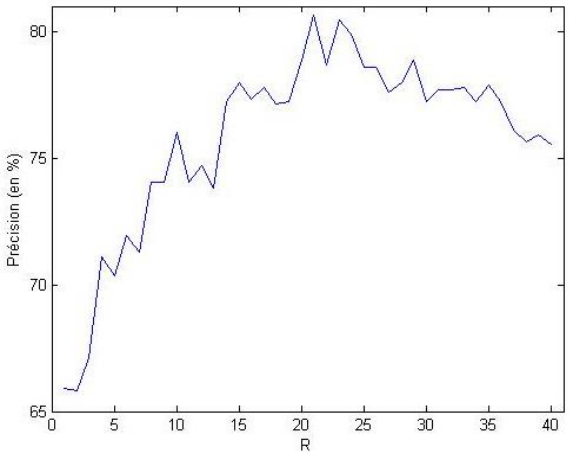

Figure 4. Graphique représentant l'évolution de la précision du résultat de classification des morceaux musicaux V/I en fonction de R entre les n-grammes de bas et de haut rang.

La Figure 4 indique que la courbe présente un maximum lorsque R = 21, puisque celui-ci possède une précision de 80.7%, tandis que cette précision diminue pour des valeurs de R inférieures ou supérieures à 21. Un descripteur D est donc obtenu en divisant la somme des valeurs supérieures à ce rang R = 21 par la somme de celles qui lui sont inférieures ou égales, comme défini par l'équation (1).

$$D = \frac{\sum_{i=1}^{R-1} r_i}{\sum_{i=R}^{30} r_i} \quad (1)$$

La valeur du descripteur sera généralement nulle pour un morceau instrumental et généralement positive si le morceau contient de la voix. De plus, si l'on considère que le nombre de descripteurs ne constitue pas un problème, il est possible d'obtenir une meilleure précision en utilisant un descripteur de dimension 30, composé de l'ensemble des rapports compris entre 1 et 30. Le Tableau 4 indique les résultats obtenus pour le

descripteur unique proposé et pour le descripteur de dimension 30 utilisant l'ensemble des rapports combinés. Ces résultats sont également comparés à la colonne « histogramme » contenant les valeurs brutes de l'histogramme de n-grammes comme décrit précédemment. Quatre classifieurs utilisés dans le logiciel d'apprentissage Weka sont représentés dans chacun des trois cas. Les performances ont été obtenues avec ce logiciel en utilisant la configuration de traitement de la base de données en validation croisée sur deux ensembles.

D'importantes différences de performances sont observées en fonction du classifieur utilisé. La faible précision du classifieur MLP lorsque R = 21 provient du fait que celui-ci n'est pas optimisé lorsqu'il est appliqué à un seul descripteur. Les performances sont meilleures pour ce classifieur dans les deux autres cas puisqu'ils contiennent en revanche un plus grand nombre de descripteurs.

Classifieur	Histogramme	R = 21	R = 1 à 30
RandomForest	**87.3**	80.7	85.9
SVM	67.2	75.3	82.2
MLP	82.1	50.1	66.8
J48	81.8	83.1	84.4
Moyenne	79.6	72.3	79.9

Tableau 4. Comparaison du pourcentage de précision obtenu dans les cas de l'utilisation de l'histogramme brut « histogramme », d'une unique valeur de rapport « R=21 » et de l'utilisation de tous les rapports « R=1 à 30 », pour quatre classifieurs.

Le descripteur MLP est en moyenne le moins performant par rapport aux deux classifieurs à base d'arbre de décision qui sont J48 et RandomForest. Dans le cas où R = 21, la moyenne des précisions des quatre classifieurs est de 72.3%, soit 7,3 et 7,6% inférieure aux moyennes respectives de l'histogramme et de toutes les valeurs de R entre 1 et 30. Il n'existe pas de différences significatives de précision entre le cas considérant l'histogramme et celui considérant les valeurs de R entre 1 et 30. Il ne semble donc *a priori* pas nécessaire de réaliser l'étape intermédiaire de création du descripteur D fondé sur l'histogramme et défini au début de cette section.

Cette méthode de prise de décision présente une meilleure précision que celle utilisant la probabilité de présence de chant. Ces deux méthodes sont cependant suffisamment différentes afin d'être utilisées simultanément. La première a en effet recours à la probabilité de présence du chant de chaque trame, tandis que la seconde se fonde sur une information provenant de plusieurs trames consécutives. La section suivante décrit donc les résultats obtenus dans le cas de l'utilisation de l'ensemble des descripteurs fournis par les deux méthodes de prise de décision.

4.3. Combinaison des méthodes proposées

Les deux méthodes proposées et décrites dans les parties 4.1 (*proba*) et 4.2 (*ngram*) analysent des aspects différents d'un même morceau de musique. La première fournit une analyse de la présence de chant à l'échelle d'une trame et la seconde se fonde sur l'étude des blocs de trames. Il semble intéressant d'étudier l'impact de l'utilisation combinée de ces deux méthodes sur les résultats de classification. Le Tableau 5 présente la différence de performances entre la combinaison des méthodes *proba* et *ngram* pour deux conditions. Dans les deux conditions, l'ensemble des descripteurs fournis par la méthode *ngram* est utilisé. Dans la première condition, l'ensemble des rapports de la méthode *ngram* est utilisé, tandis que les rapports issus de l'histogramme brut de la méthode *ngram* sont utilisés dans la seconde condition.

De même que pour les Tableaux 3 et 4, il existe d'importantes différences de performances en fonction du classifieur utilisé. Les deux classifieurs à base d'arbre de décision, représentés par J48 et RandomForest, sont les plus performants dans les deux conditions.

Classifieur	*proba* & *ngram* « R = 1 à 30 »	*proba* & *ngram* « Histogramme »
RandomForest	**89.3**	88.0
SVM	76.4	68.5
MLP	74.1	82.2
J48	85.1	82.1
Moyenne	81.2	80.2

Tableau 5. Comparaison de deux combinaisons différentes de descripteurs provenant des méthodes *proba* et *ngram*.

Les résultats obtenus lors de l'utilisation combinée des deux méthodes de prise de décision sont supérieurs à ceux obtenus lorsque chaque méthode est utilisée individuellement. Cette augmentation de précision est supérieure dans le cas de l'utilisation des méthodes *proba* & *ngram* « R = 1 », une augmentation de la précision de 1.3% est en effet observée contrairement au cas « histogramme » dans lequel cette augmentation n'est que de 0.8%. Les descripteurs provenant de l'ensemble des rapports de la méthode *ngram* sont donc utilisés de manière conjointe avec ceux issus de la méthode *proba* dans la version finale de la nouvelle approche. La section suivante compare les performances obtenues grâce à l'utilisation des descripteurs de la nouvelle approche aux performances obtenues suite à l'implémentation de l'algorithme de [9].

5. TEST DE LA ROBUSTESSE DE L'APPROCHE PROPOSEE

L'approche décrite dans la partie précédente est testée sur une importante base de données et les résultats obtenus sont comparés à ceux obtenus au moyen de l'implémentation de [9]. Pour cela, le même protocole

que celui défini dans la partie 3 est observé, la précision de la méthode proposée est donc évaluée en fonction de la taille de la base de données de test. Celle-ci varie de 1000 à 8000 morceaux suivant un pas de 1000 morceaux sélectionnés aléatoirement dans la grande base de données (voir Section 2). Les résultats sont exprimés en pourcentage de précision.

La Figure 5 montre que la précision des deux approches étudiées chute lorsque la taille de la base de données augmente. Si la chute atteint cependant 13.1% pour [9] lorsque la base de données augmente de 1000 à 8000 morceaux, la nouvelle approche ne subit qu'une diminution de 3.6% dans les mêmes conditions. Les précisions de 82.8% et de 80.6% sont en effet obtenues respectivement pour 1000 et 8000 morceaux. La nouvelle approche proposée est par conséquent plus robuste à un passage à l'échelle puisque la précision demeure stable pour une base de test comprenant entre 2000 et 8000 morceaux musicaux.

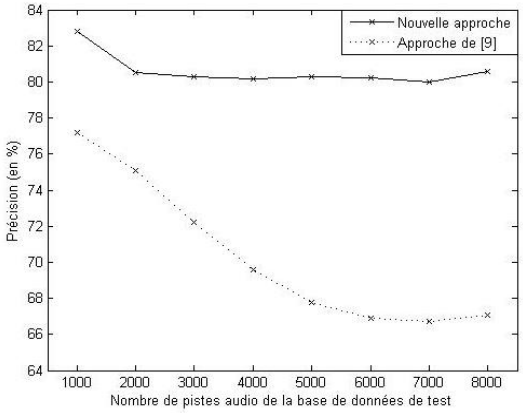

Figure 5. Précision de l'approche proposée en comparaison avec celle de [9], en fonction de la taille de la base de données de test.

Cette différence de robustesse semble provenir de l'utilisation d'une approche hiérarchique, qui utilise d'abord une description à l'échelle de la trame avant d'inférer la présence de chant à l'échelle de la chanson. La précision obtenue, bien que supérieure à celle de l'approche de [9], demeure inférieure à 83% et n'est donc pas transférable au domaine industriel.

6. DISCUSSION ET PERSPECTIVES

L'article propose une nouvelle approche permettant de discriminer les morceaux musicaux comportant du chant de ceux qui n'en présentent pas, au sein d'une base de données musicales conséquente.

La méthode de prise de décision utilisant les trames en tant que n-grammes fournit des résultats de classification plus robustes à un passage à l'échelle que l'algorithme existant qui, à notre connaissance, possède les meilleures performances. Bien que la méthode de prise de décision fondée sur la probabilité de présence de chant apparaît peu performante lorsqu'elle est utilisée seule, elle permet néanmoins d'améliorer les résultats de classification lorsqu'elle est utilisée conjointement avec la méthode utilisant des n-grammes.

Cet article propose principalement une nouvelle approche concernant la prise de décision quant à la détection du chant dans un morceau musical. Cette approche est possible grâce à l'utilisation, à l'échelle de la trame, de descripteurs existants. Elle pourrait par la suite être utilisée afin de comparer les performances obtenues avec d'autres descripteurs à l'échelle de la trame. Les deux méthodes utilisées sont en cours d'amélioration, il est par exemple prévu d'utiliser le pourcentage de prédiction d'une trame afin de pondérer l'impact d'un n-gramme sur une prédiction. Cette dernière amélioration est toutefois limitée par la précision des descripteurs à l'échelle de la trame. C'est pourquoi il est également prévu d'étudier d'autres descripteurs, nouveaux et existants tels que les *i-vectors*, à l'échelle de la trame.

De plus, la base de données de test utilisée et composée de 8000 morceaux musicaux a été obtenue artificiellement, une nouvelle base de données est donc en cours de création. La précision obtenue dans le cas de l'augmentation de la base de données de test sera d'autant plus pertinente qu'à la base de données contenant la base de vérité seront ajoutés des morceaux musicaux contenant des instruments imitant le chant.

Une dernière démarche consistera à reconsidérer le système de classification en deux classes. L'objectif est actuellement de classifier correctement un maximum de morceaux comme contenant du chant ou non. Cet objectif n'est cependant pas nécessaire par exemple pour un site de streaming musical. Un tel site possède en effet une base de données de plusieurs millions de morceaux musicaux et le classement de leur intégralité n'est pas indispensable, contrairement au caractère certain de ce classement. Il est donc nécessaire de proposer un système capable de classifier certainement un nombre restreint de morceaux musicaux dans une base de données et donc de garantir une précision de classification de 100% pour chacun de ces morceaux. Le nouvel objectif consiste par conséquent à augmenter la proportion de morceaux de la base de données qui sont correctement classés par notre approche.

7. REMERCIEMENTS

Les auteurs remercient Bernhard Lehner de l'Université Johannes Kepler en Autriche pour le partage du code de calcul de ses descripteurs ainsi que pour ses annotations concernant la base de données RWC.

8. REFERENCES

[1] H. S. Beigi, S. H. Maes, U. V. Chaudhari, et J. S. Sorensen, « A hierarchical approach to large-scale speaker recognition. », in *Eurospeech*, 1999.

[2] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, et J. Bello, « MedleyDB: a multitrack

dataset for annotation-intensive MIR research », in *15th International Society for Music Information Retrieval Conference*, 2014, p. 155–160.

[3] W. Chou et L. Gu, « Robust singing detection in speech/music discriminator design », in *Acoustics, Speech, and Signal Processing, ICASSP Proceedings. IEEE.*, 2001, vol. 2, p. 865–868.

[4] J. S. Downie, « The scientific evaluation of music information retrieval systems: Foundations and future », *Computer Music Journal*, vol. 28, n° 2, p. 12–23, 2004.

[5] K. El-Maleh, M. Klein, G. Petrucci, et P. Kabal, « Speech/music discrimination for multimedia applications », in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 6, p. 2445–2448.

[6] A. Eronen et A. Klapuri, « Musical instrument recognition using cepstral coefficients and temporal features », in *Acoustics, Speech, and Signal Processing, ICASSP Proceedings*, 2000, vol. 2, p. II753–II756.

[7] H. Fastl et E. Zwicker, *Psychoacoustics: Facts and models*, vol. 22. Springer Science & Business Media, 2007.

[8] J. T. Foote, « Content-based retrieval of music and audio », in *Voice, Video, and Data Communications*, 1997, p. 138–147.

[9] A. Ghosal, R. Chakraborty, B. C. Dhara, et S. K. Saha, « Song/instrumental classification using spectrogram based contextual features », in *Proceedings of the CUBE International Information Technology Conference*, 2012, p. 21–25.

[10] A. Ghosal, R. Chakraborty, B. C. Dhara, et S. K. Saha, « A hierarchical approach for speech-instrumental-song classification », *SpringerPlus*, vol. 2, n° 1, p. 1–11, 2013.

[11] M. Goto, H. Hashiguchi, T. Nishimura, et R. Oka, « RWC Music Database: Popular, Classical and Jazz Music Databases. », in *ISMIR*, 2002, vol. 2, p. 287–288.

[12] G. Guo et S. Z. Li, « Content-based audio classification and retrieval by support vector machines », *Neural Networks, IEEE Transactions on*, vol. 14, n° 1, p. 209–215, 2003.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten, « The WEKA data mining software: an update », *ACM SIGKDD explorations newsletter*, vol. 11, n° 1, p. 10–18, 2009.

[14] Y. Ikemiya, K. Yoshii, et K. Itoyama, « Singing Voice Separation », in *MIR Exchange*, 2014.

[15] O. Lartillot, P. Toiviainen, et T. Eerola, « A Matlab Toolbox for Music Information Retrieval », in *Data Analysis, Machine Learning and Applications*, C. Preisach, P. D. H. Burkhardt, P. D. L. Schmidt-Thieme, et P. D. R. Decker, Éd. Springer Berlin Heidelberg, 2008, p. 261‑268.

[16] S. Leglaive, R. Hennequin, et R. Badeau, « Singing voice detection with deep recurrent neural networks », in *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, p. 121–125.

[17] B. Lehner, G. Widmer, et R. Sonnleitner, « On the reduction of false positives in singing voice detection », in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, p. 7480–7484.

[18] B. Lehner et G. Widmer, « Monaural blind source separation in the context of vocal detection », présenté à 16th International Society for Music Information Retrieval Conference (ISMIR), At Malaga, Spain, 2015.

[19] T. Lidy et A. Rauber, « Evaluation of feature extractors and psycho-acoustic transformations for music genre classification », in *International Society for Music Information Retrieval Conference*, 2005, p. 34–41.

[20] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, et L. Daudet, « Kernel additive models for source separation », *Signal Processing, IEEE Transactions on*, vol. 62, n° 16, p. 4298–4310, 2014.

[21] D. Y. Loni et S. Subbaraman, « Extracting Acoustic Features of Singing Voice for Various Applications Related to MIR: A Review », 2013.

[22] M. E. Markaki, A. Holzapfel, et Y. Stylianou, « Singing voice detection using modulation frequency feature. », in *SAPA@ INTERSPEECH*, 2008, p. 7–10.

[23] M. McVicar et T. De Bie, « An ensemble method for learning to extract vocals from polyphonic musical audio », in *16th International Society for Music Information Retrieval Conference*, 2015.

[24] T. L. Nwe et H. Li, « On fusion of timbre-motivated features for singing voice detection and singer identification », in *Acoustics, Speech and Signal Processing, ICASSP, IEEE*, 2008, p. 2225–2228.

[25] L. Rabiner et B.-H. Juang, « Fundamentals of speech recognition », *Prentice Hall*, 1993.

[26] M. Ramona, G. Richard, et B. David, « Vocal detection in music with support vector machines », in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, p. 1885–1888.

[27] M. Rocamora et P. Herrera, « Comparing audio descriptors for singing voice detection in music audio files », in *Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007, vol. 26, p. 27.

[28] S. O. Sadjadi, S. M. Ahadi, et O. Hazrati, « Unsupervised speech/music classification using one-class support vector machines », in *Information, Communications & Signal Processing, 2007 6th International Conference on*, 2007, p. 1–5.

[29] J. Saunders, « Real-time discrimination of broadcast speech/music », in *icassp*, 1996, p. 993–996.

[30] J. Schlüter et T. Grill, « Exploring data augmentation for improved singing voice detection

with neural networks », in *International Society for Music Information Retrieval Conference*, 2015.
[31] M. S. Spina et V. W. Zue, « Automatic transcription of general audio data: Preliminary analyses », in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 2, p. 594–597.
[32] Y.-H. Tseng, « Content-based retrieval for music collections », in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, p. 176–182.
[33] G. Tzanetakis et P. Cook, « Musical genre classification of audio signals », *Speech and Audio Processing, IEEE transactions on*, vol. 10, n° 5, p. 293–302, 2002.
[34] K. West et S. Cox, « Finding An Optimal Segmentation for Audio Genre Classification. », in *International Society for Music Information Retrieval Conference*, 2005, p. 680–685.
[35] K. West et S. Cox, « Features and classifiers for the automatic classification of musical audio signals. », in *International Society for Music Information Retrieval Conference*, 2004.
[36] T. Zhang et C.-C. J. Kuo, « Content-based classification and retrieval of audio », in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, 1998, p. 432–443.
[37] T. Zhang, « Semi-automatic approach for music classification », in *ITCom 2003*, 2003, p. 81–91.