EmoDiderot : Reconnaissance des émotions dans la musique

Massinissa Hamidi*, Hassane Gaci, and Van Luan Nguyen

Université Paris Diderot

13 décembre 2016

*hamidi@informatique.univ-paris-diderot.fr

Table des matières

1 Introduction

L'un des aspects indéniables que l'on peut reconnaître à la musique est l'immense influence que celle-ci peut avoir sur nos émotions et sa grande capacité à modeler nos ressentiments. En mettant en place un système de recommandation de musique, on peut penser à suggérer à l'utilisateur, des contenus du même artiste ou d'artistes similaires. On peut penser aussi à suggérer des titres en fonction du genre musical ; pop, rock, ... mais pourquoi ne pas prendre en compte l'aspect émotion et de proposer des morceaux qui seraient plus en adéquation avec son humeur.

Le but du projet est donc d'aboutir à un système capable de suggérer du contenu audio à un utilisateur en fonction de son humeur du moment. Les données audio brutes de notre système seront analysées en "off-line" pour les regrouper selon la (ou les) émotion(s) qu'elles véhiculent. Les suggestions se feront alors, à partir d'un ou de quelques morceaux écoutés durant une session. L'intérêt du projet est double. En effet, nous aurons à découvrir de nouveaux concepts liés au traitement des données sonores (traitement du signal) et à explorer les liens entre musique et humeur.

2 Reconnaissance des émotions

Une des premières questions qui se pose et de savoir comment représenter les émotions et la manière de les formaliser. En effet, l'émotion est une notion subjective difficile à cerner mais il existe tout de même dans la littérature en psychologie musicale, deux paradigmes principaux pour représenter les émotions. Le premier est une représentation catégorielle, le deuxième et une représentation dimensionnelle qui définissent un espace émotionnel.

2.1 Représentation catégorielle

La représentation catégorielle consiste à diviser les émotions en catégories. Chaque émotion est étiquetée avec un ou plusieurs adjectifs. L'un des plus pertinents travaux dans ce domaine est l'étude de Hevner (1936) et son cercle d'adjectif comme le montre la figure 2.1. La liste des adjectifs de Hevner est composée de 67 mots arrangés en huit clusters. Cette représentation est beaucoup moins utilisée ; d'après une étude de certains morceaux de musique, il se peut qu'on tombe sur un morceau qui appartient à plusieurs clusters en même temps, ce qui rends difficile de la détermination de son coté émotionnel.

2.2 Représentation dimensionnelle

L'approche dimensionnelle est dûe à Russell[5] qui en 1980 a introduit le "circumplex model of affect". Cette représentation est basée sur le principe que les émotions résultent d'un nombre fixé de concepts que l'on peut représentér dans un espace bidimensionnel, les dimensions peuvent être un axe de plaisir et de déplaisir, d'éveil ou d'ennui. Cette approche s'accorde sur les deux premières dimensions : la valence et l'activation comme le montre la figure 2.2. La valence permet de distinguer les émotions positives, agréables comme la joie, des émotions négatives, désagréable, comme la colère. L'activation représente le niveau

FIGURE 1 – Cercle d'adjectifs de Hevner [2]

d'excitation corporelle, qui transparait, comme l'accélération du cœur, la transpiration. Ce modèle a eu beaucoup de sucées car il permet de représenter une infinité d'émotions, il permet également la représentation facile des émotions nuancées.

3 Données

La première étape du projet a été de chercher une base de données qui soit la plus compète pour notre étude. Notre première idée était de rassembler le

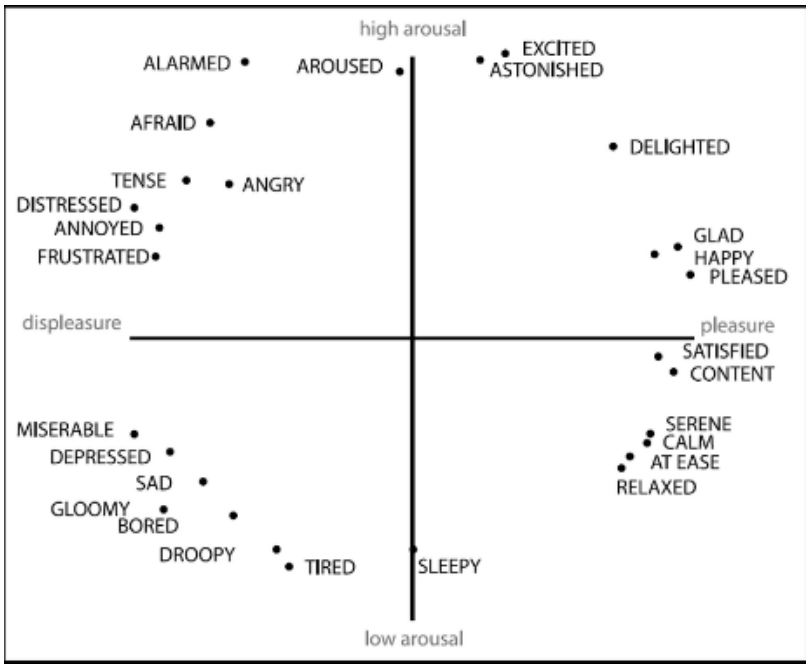

FIGURE 2 – Représentation dimensionelle de Russell [5]

maximum possible de chansons et de ne pas s'intéresser au genre de la music ni au style, nous expliquons ci-dessus les déférente base de données qu'on a exploré.

3.1 Million Song Dataset (MSD)

The Million data set est une collection librement disponible elle contient environ un million de morceau musical, ces chansons sont du format meta, chaque morceau est une composition d'un certain nombre de champs (id, name_Artist, name_track, etc.) et de quelques caractéristique extraite des signaux a partir des signaux musicaux brute, mais on s'est pas basé sur cette collection de music en raison qu'elle se limite juste en genre et au style de la music et ce qui n'est pas notre objectif de ce projet donc on a décidé de ne la pas prendre comme une source pour nos traitement.

Dans le même esprit on peut citer deux autres dataset qui se basent sur le million song data set et qui peuvent etre exploité en parallèle.

3.2 Second Hand Song

3.3 MusiXmatch

c'est une des plus grande bibliothèque de paroles au monde, il dispose de 50 millions d'utilisateurs et plus de 14 millions de paroles, cette bibliothèque donne aux utilisateurs la possibilité de commenter en donnant leur avis et leur sensations sur les textes des chansons et ce qui pourra nous aider à décrire l'émotion de la chanson.

FORMAT:

```
#     #   - comment, to ignore
#     \%  - list of top words, comma-separated
#         - normal line, contains track\_id, mxm track id,
#           then word count for each of the top words, comma-separated
#           word count is in sparse format -> ...,<word idx>:<cnt>,...
#           <word idx> starts at 1 (not zero!)
```

Un exemple d'une entrée dans le fichier à notre disposition

TRAABRX12903CC4816,1548880,2:19,4:7,5:6,10:1,12:13,13:6,17:4, ...

4 Circonvenir aux limitations du MSD

4.1 tuwien-ifs project

Des chercheurs dans un laboratoire informatique a l'université tu wein de viennes en Autriche ont conclu Que l'ensemble de morceaux musicaux compote millions Song data sont toutefois pas possible de les télécharger ainsi qu'on est limité aux fonctionnalités fournis par ces données, a cet effet ils ont proposé une solution d'utilisé un content provider, ils ont téléchargé des échantillons audio de 30 a 60 secondes afin d'analyser leur contenu et d'extraire de nouvelles caractéristique. Pour ce faire ils ont utilisé deux outlis a savoir jaudio digital signal project et marsays.

4.2 `jAudio` Digital Signal Processing Project

jaudio est un outil qui fournit un programme facile pour extraire les propriétés d'un document sonor telles que les points de battement, des résumés statistique ainsi que de nombreuses autres propriétés utiles. Ces propriétés peuvent ensuite êtres introduites pour l'apprentissage automatiques.

4.3 `Marsyas`

MARSYAS est un logiciel pour le prototypage rapide d'applications audio, son objectif de base est de fournir une architecture flexible et des algorithmes très performants qui seront utiles dans le développement en temps réel d'analyse audio.

5 À la recherche des labels

À ce stade nous ne disposons d'aucune information exploitable en relation avec le thème que nous avons retenu afin d'effectuer de la classification. En effet, nous avons en main un certain nombre de features liées à la structure haut niveau et bas niveau d'un morceau muical, mais on ne peux rien faire avec du moment qu'on ne connais quelles sont les émotions associées. C'est pour cela que nous nous sommes lancés à la recherche d'un set de labels associés aux données qu'on a déja et dont le vocabulaire couvre celui utilisé dans les représentations utilisées en psychologie.

5.1 lastfm

Lastfm est un réseau social musical regroupant une communauté de plus de 30 millions de fans de musique. La plateform permet aux utilisateurs de rechercher de la musique, d'écouter des extraits, constituer des playlists et le plus interessant dans notre cas, annoter les musiques. Cette annotation étant libre, les termes utilisés couvrent donc un plus grand espace que celui formé par les termes propres aux études menées par les psychologues. C'était donc ce qui nous fallait ; chaque chanson lui est associée une liste de tags ordonés du plus fréquement utilisé au plus rare. De plus, à chaque tag est associée une liste des tags similaires.

La prochaine étape à considérer, après la collecte des labels, est la construction de notre espace de représentation des émotions en accord avec les représentations proposées dans la littérature par les psychologues. En d'autres termes, passer du tag, à une position concrète, la composante en x, la composante en y, dans le plan de Russell. À cet effet, nous avons mené une petite revue de la littérature qui nous a fait aboutir à un certain nombre de publications qui parlent de ce passage et qui se basent sur le dataset lastfm.

6 Modèles de representation à partir des réseaux sociaux

Dans cette partie, nous allons présenter succinctement les travaux menés à l'université Pompeu Fabra à Barcelone[3] et ceux d'une équipe de recherche finlandaise[6] qui ont pour but de traiter le problème de reconnaissance des émotions dans la musique à l'aide des données collectées dans le contexte de la vie de tous les jours (et non pas dans le cadre controllé des laboratoires qui se fait d'habitude dans les études psychologiques). Ces travaux se basent sur les informations qui peuvent être inférées à partir d'un réseau social tel que lastfm[3]. Le processus se base sur la collecte de chansons et des tags qui leur sont associés dans la base de données de lastfm et procède ensuite, à une analyse sémantique de ces données afin d'inférer une représentation spatiale des tags qui soit la plus proche du ressentie des membre de ce réseau social.

6.1 Collecte des données à partir de lastfm

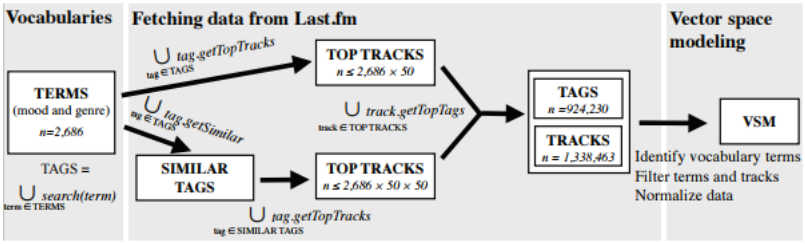

FIGURE 3 – Processus de collecte des chansons et des tags [6]

Dans notre cas, vu que le web service `tag.getSimilar` ne répond pas, nous avons opté pour la solution suivante ; pour tous les tags que nous avons en main

(101 tags), nous récupérons la liste des tracks qui ont été le plus de fois annotés avec un tag particulier à l'aide de `tag.getTopTracks`. Ensuite, pour tous les tracks obtenus, nous récupérons les tracks similaires. Il faut noter toute fois que les tracks similaires sont déterminés par lastfm sur la base des écoutes des utilisateurs ce qui n'écarte pas la possibilté d'obtenir des tracks qui n'ont pas dutout le même mood mais apporte quand même une information additionnelle sur une possible similarité. Au final, pour toutes les chansons obtenues, nous allons collecter les tags qui lui sont associés.

6.2 Document-Term Matrix

À partir des données collectées précedement, on procède à la construction d'une matrice qui se compose des tags en lignes et des chansons en colonnes. Pour une chanson donnée, on aura les tf-idf des tags qui lui sont associés[4].

6.3 Singular Value Decomposition (SVD)

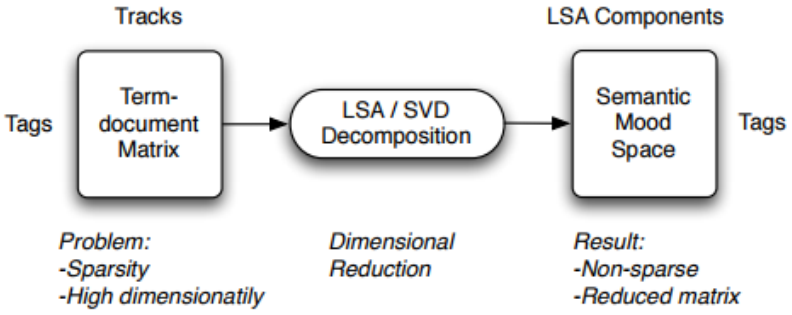

FIGURE 4 – Réduction de dimensionalité de la Term-Document Matrix [3]

6.4 Affective Circumplex Transformation (ACT)

Après avoir obtenu le Semantic Mood Space par réduction de dimensionnalité, nous devons passer vers une représentation en deux dimensions en accord avec l'étude de Russell. Cette transformation est effectué par l'intérmédiaire de l'ACT[6] qui concrètement, prend le plan de Russell, avec les différents positions "théoriques" des tags et applique une transformation basée sur le Semantic Mood Space obtenu à partir du ressentiment général du réseau social. On obtient finalement le plan de la figure 5 où les tags se décalent de leurs positions initiales vers de nouvelles positions suivant ce que ressent la majorité des gens.

7 Conclusion

Nous avons hélas pas pu poursuivre notre pipline de prédiction jusqu'au bout vu que la constitution de notre dataset nous a pris beaucoup plus de temps que prévu à cause principalement de l'éparpillement des bouts d'informations dont nous avions besoins. Cette recherche d'information s'est étalée tout au long du projet et nous avons du procéder par tatônements jusqu'à arriver à la

FIGURE 5 – Résultat de l'Affective Circumplex Transformation [6]

publication [6] dont nous avons contacté les auteurs afin de nous fournir des données manquantes. Cette denière nous a éclairci les choses vu que nous étions restés bloqués sur la construction de l'espace de représentation beaucoup plus longtemps que prévu.

Ceci étant dit, ce qui a été fait jusqu'à présent pose les fondements pour un future travail plus rigoureux et plus précis qui pourra être mené à terme pour avoir des prédictions pour n'importe quelle chanson en se basant sur l'étude de sa structure bas et haut niveau. Les données de ce types sont déja à notre dispostion, et les techniques utilisées pour les calculer sont mentionnées dans ce présent rapport. Nous encourageons d'ailleurs les étudiants de la prochaine promotion à tenter l'aventure.

A Structure des fichiers `.hdf5` du Million Song Dataset

```
>> h5disp('TRAAAAW128F429D538.h5')
HDF5 TRAAAAW128F429D538.h5
Group '/'
    Attributes:
        'TITLE':  'H5 Song File'
        'CLASS':  'GROUP'
        'VERSION':  '1.0'
        'PYTABLES_FORMAT_VERSION':  '2.0'
        'FILTERS':  65793
    Group '/analysis'
        Attributes:
            'TITLE':  'Echo Nest analysis of the song'
            'CLASS':  'GROUP'
            'VERSION':  '1.0'
            'FILTERS':  65793
        Dataset 'bars_confidence'
            Size:  83
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of confidence of bars'
                'EXTDIM':  1922
        Dataset 'bars_start'
            Size:  83
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of start times of bars'
                'EXTDIM':  1921
        Dataset 'beats_confidence'
            Size:  344
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of confidence of sections'
                'EXTDIM':  1920
        Dataset 'beats_start'
            Size:  344
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of start times of beats'
                'EXTDIM':  1919
        Dataset 'sections_confidence'
            Size:  10
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of confidence of sections'
                'EXTDIM':  1918
        Dataset 'sections_start'
            Size:  10
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of start times of sections'
                'EXTDIM':  1917
        Dataset 'segments_confidence'
            Size:  971
            MaxSize:  Inf
            Datatype:  H5T_IEEE_F64LE (double)
            ChunkSize:  1024
            Filters:  shuffle, deflate(1)
            Attributes:
                'CLASS':  'EARRAY'
                'VERSION':  '1.0'
                'TITLE':  'array of confidence of segments'
                'EXTDIM':  1911
```

```
Dataset 'segments_loudness_max'
    Size:  971
    MaxSize:  Inf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  1024
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of max loudness of segments'
        'EXTDIM':  1914
Dataset 'segments_loudness_max_time'
    Size:  971
    MaxSize:  Inf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  1024
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of max loudness time of segments'
        'EXTDIM':  1915
Dataset 'segments_loudness_start'
    Size:  971
    MaxSize:  Inf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  1024
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of loudness of segments at start time'
        'EXTDIM':  1916
Dataset 'segments_pitches'
    Size:  12x971
    MaxSize:  12xInf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  12x85
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of pitches of segments (chromas)'
        'EXTDIM':  1912
Dataset 'segments_start'
    Size:  971
    MaxSize:  Inf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  1024
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of start times of segments'
        'EXTDIM':  1910
Dataset 'segments_timbre'
    Size:  12x971
    MaxSize:  12xInf
    Datatype:  H5T_IEEE_F64LE (double)
    ChunkSize:  12x85
    Filters:  shuffle, deflate(1)
    Attributes:
        'CLASS':  'EARRAY'
        'VERSION':  '1.0'
        'TITLE':  'array of timbre of segments (MFCC-like)'
        'EXTDIM':  1913

        [...]
```

B Format des fichiers Million Song Dataset to AcousticBrainz mapping

The json mapping files have the following format[1] :

```
{
  "query": { // MSD metadata used for the query
    "song_id": "SOTUGDX12A8C13E5F7",
    "title": "Aground",
    "artist_name": "Fresh Moods",
    "track_id": "TRMRLVN128F42AA35E",
    "release": "Exhale",
    "duration": "377.02485",
    "artist_mbid": "cba5dbef-14f8-47a7-8632-a63e7a9738e2"
  },
  "match": [ // A list of all results from MusicBrainz which match on artist id and title
    {
```

```
      "length": 375000,
      "title": "Aground",
      "id": "eb301e57-5c6d-49a4-bb0d-2d963ca5a59b",
      "releases": [ // Releases on which this recording appears. Could be more than 1
        {
          "id": "24e38551-44ab-4aed-81c6-b60447dbfd0d",
          "title": "Campari Lounge II"
        }
      ],
      "artists": [ // Artists from MusicBrainz. Could be more than 1
        {
          "id": "cba5dbef-14f8-47a7-8632-a63e7a9738e2",
          "name": "Fresh Moods"
        }
      ]
    },
    {
      "length": 380026,
      "title": "Aground",
      "id": "47ce77c9-9296-4bc3-a878-7390a3303e0c",
      "releases": [
        {
          "id": "6ca97f9c-b764-4c98-b512-3ddf6e51db79",
          "title": "Exhale"
        }
      ],
      "artists": [
        {
          "id": "cba5dbef-14f8-47a7-8632-a63e7a9738e2",
          "name": "Fresh Moods"
        }
      ]
    }, //... More matches
  ],
  "matchtypes": { // Some data about how the MusicBrainz data matches the MSD data
    "duration": "withindur",
    "release": "",
    "type": "exact"
  }
}
```

Références

[1] Million song dataset to acousticbrainz mapping. `http://labs.acousticbrainz.org/million-song-dataset-mapping/`, 2016. accessed :2016-12-06.

[2] Paul R Farnsworth. A study of the hevner adjective list. *The Journal of Aesthetics and Art Criticism*, 13(1) :97–103, 1954.

[3] Cyril François Laurier et al. *Automatic Classification of musical mood by content-based analysis*. PhD thesis, Universitat Pompeu Fabra, 2011.

[4] Mark Levy and Mark Sandler. Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37(2) :137–150, 2008.

[5] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 :1161–1178, 1980.

[6] Pasi Saari and Tuomas Eerola. Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, 26(10) :2548–2560, 2014.