# Prediction of the Distribution of Perceived Music Emotions Using Discrete Samples

Raphaël Romero

ENS Paris-Saclay

13/03/2018

# Plan

1. Reconnaissance automatique d'émotions dans la musique
2. Approches existantes
3. Modèle présenté dans l'article
4. Evaluations

# Plan

1. Reconnaissance automatique d'émotions dans la musique

# Reconnaissance automatique des émotions

- Attribut "haut niveau" des morceaux musicaux.
- Problématique étudiée depuis longtemps dans la communauté.
- Plusieurs catégories d'émotions "exprimées", "perçues" et "ressenties".
- Ici il s'agit d'étudier l'émotion "perçue" par une approche machine learning.

# Problématique

- Caractéristique fondamentalement subjective.
- Approche "universelle" : tous les individus perçoivent la même émotion
- Limites : ne permet pas de faire de la prédiction avec une bonne précision
- Solution proposée : considérer une distribution probabiliste des émotions associée à un morceau donné.

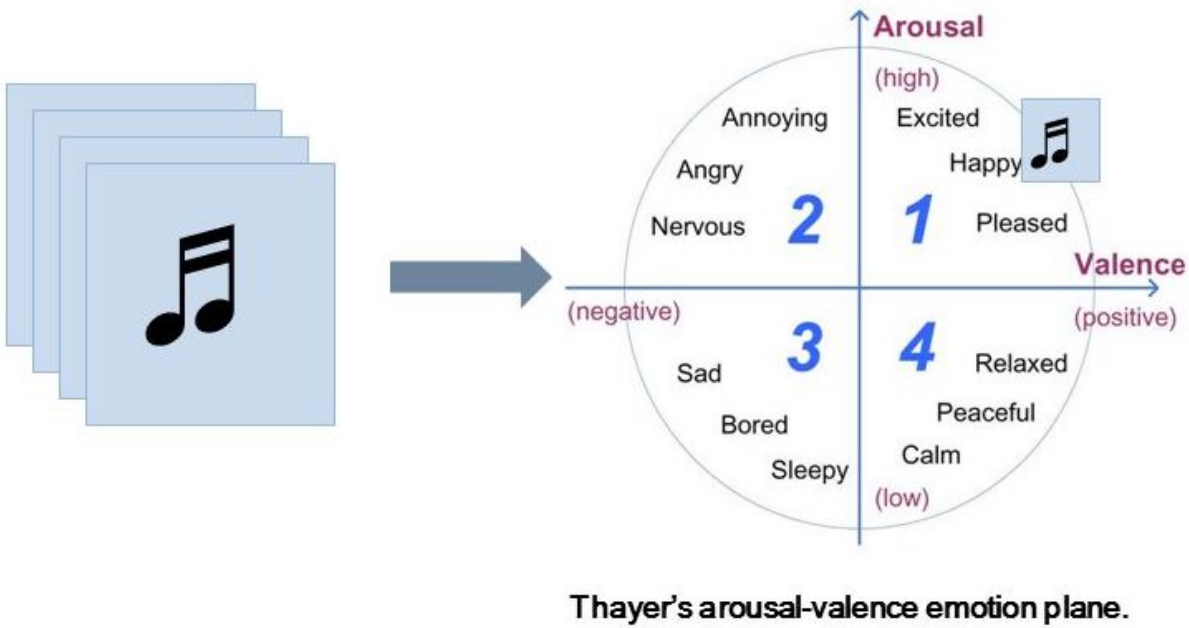
# Plan

1. Reconnaissance automatique d'émotions dans la musique
2. Approches existantes

# Approches existantes

- Morceau $\longrightarrow$ Features $\longrightarrow$ Emotion (catégories ou valeurs continues)
- On apprend la fonction $f$ : *Features* $\mapsto$ *Emotion* sur des valeurs collectées auprès d'anotateurs humains.
- Approche catégorielle : assigner un ou plusieurs label à chaque morceau. $\Rightarrow$ Classification
- Approche dimensionnelle : une émotion est un vecteur (ex : [Valence(humeur),Arousal(énergie)]) $\Rightarrow$ Régression

# Approches existantes




Thayer's arousal-valence emotion plane.

FIGURE 1 – Plan Valence-Arousal

# Plan

1. Reconnaissance automatique d'émotions dans la musique
2. Approches existantes
3. Modèle présenté dans l'article

# Modèle présenté

- Morceau $(d_s) \longrightarrow$ features $(x_s) \longrightarrow$ distribution $(p)$ sur le plan VA.
- Quantification : On restreint p à ses valeurs sur une grille uniforme de taille $G \times G$.
- $(i,j) \in [|1,G|]^2 \mapsto (V,A)_{i,j} = ((i-\frac{1}{2})\Delta - 1, (j-\frac{1}{2})\Delta - 1)$ où $\Delta = \frac{2}{G}$ est le pas de quantification.
- On modélise la sortie comme la distribution discrète $(y_{i,j}^{(s)})$ des émotions associées au clip $d_s$.

# Formulation du problème

- On cherche à estimer la fonction

$$f : x_s \mapsto y_s = [y_{1,1}^{(s)}, ..., y_{G,G}^{(s)}]$$

- Sous l'hypothèse que les $G^2$ valeurs d'émotions sont indépendants, cela revient à estimer les $G^2$ fonctions $f_{i,j} : x_s \mapsto y_{i,j}^{(s)}$.
- Points à préciser : collecte d'une réalité terrain, algo de régression, features utilisées.

# Réalité terrain

- On demande à chaque utilisateur de cliquer un point de l'espace V-A.
- On obtient une liste de $U_s$ points ($U_s$ est le nombre d'utilisateurs) $(q_{su})_{s=1,...,U_s}$ dont les positions ne sont pas forcément sur la grille : $\{p_{i,j} = (i - \frac{1}{2})\Delta - 1, (j - \frac{1}{2})\Delta - 1) i, j = 1, ..., G\}$

# Exemples d'annotations

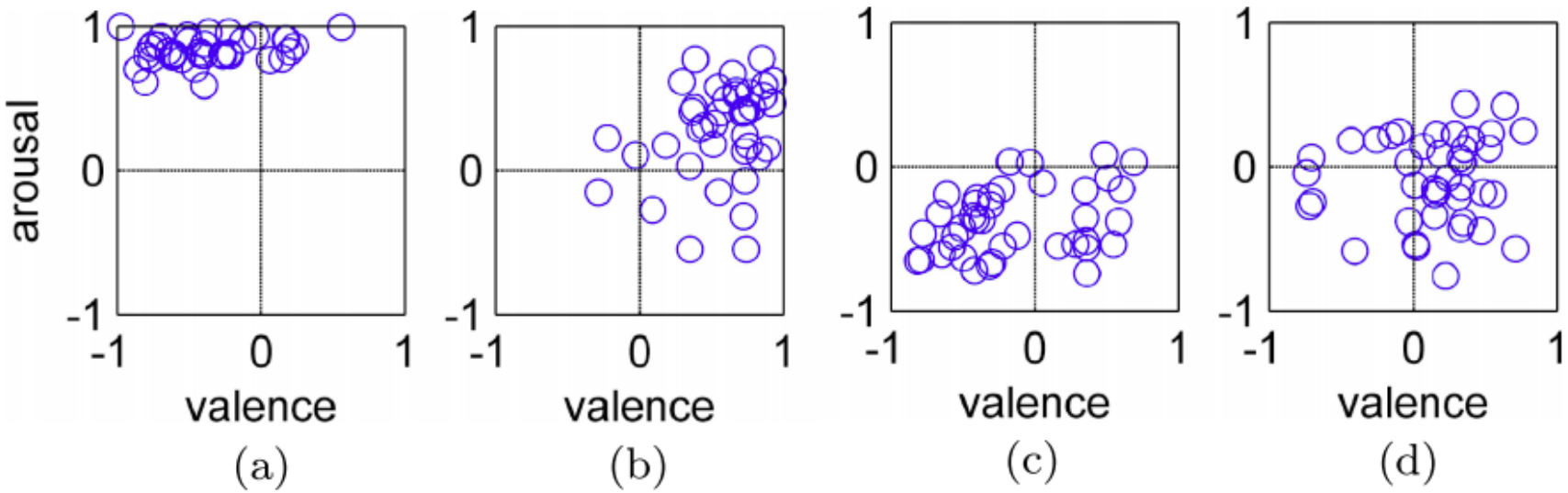



FIGURE 2 – Annotations récoltées par les auteurs

# Réalité terrain

- Les poids sont estimés à l'aide d'un noyau gaussien :

$$y_{i,j}^{s} = \frac{1}{U_s} \sum_{u=1}^{U_s} K(p_{i,j} - q_{s,u})$$

- $K(z) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2} z^T \Sigma^{-1} z)$ , $\Sigma = diag(h_a, h_v)$.
- Pour chaque point $p_{i,j}$, chaque utilisateur a une contribution au poids $y_{i,j}^{s}$ à hauteur d'un coefficient qui décroît exponentiellement en $||p_{i,j} - q_{su}||_{\Sigma}^{2}$.

# Réalité terrain

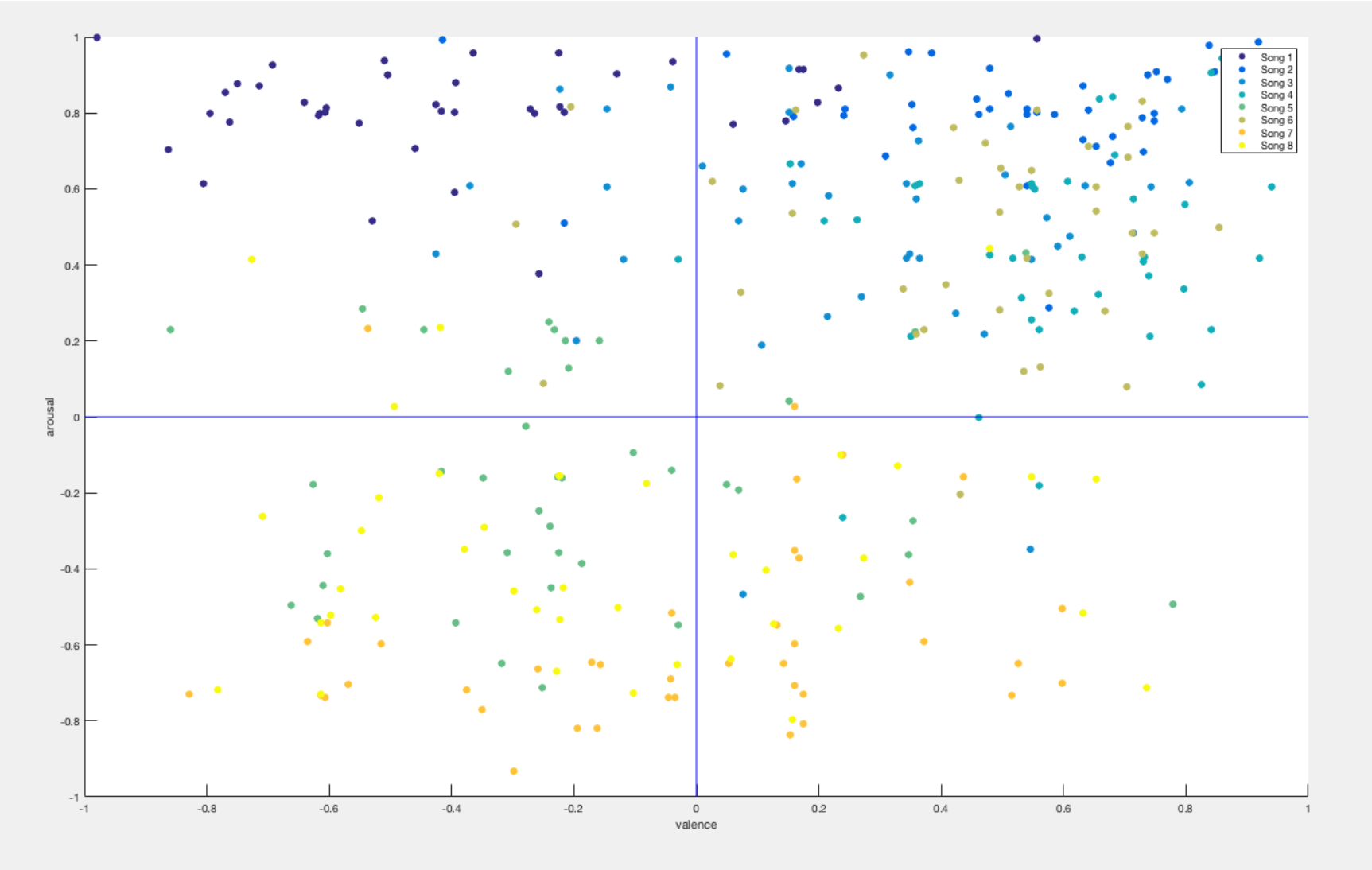



FIGURE 3 – Exemple d'annotations données dans l'article

# Régression

- On entraîne $G^2$ régresseurs.
- Chaque $f_{i,j}$ est entraîné sur la base de donnée $(x_s, y^s_{i,j})_{s=1,\dots,N}$, construite précédemment.
- Critère à minimiser :

$$\epsilon = \frac{1}{N} \sum_{s=1}^{N} (y^s_{i,j} - f_{i,j}(x_s))^2$$

(Mean squared-error)

# Régression : méthode utilisée

- Support vector regression : $f$ est modélisée par :

$$f(x_s) = m^T \phi(x_s) + b$$

- $\phi$ est une transformation de l'espace des features dans un espace de redescription.
- Le noyau associé est $K(x_p, x_q) = \phi(x_p)^T \phi(x_q)$.
- On prend généralement $K(x_p, x_q) = exp(-\gamma ||x_p - x_q||^2)$ (noyau RBF) en ajustant bien le paramètre $\gamma$.
- Les paramètres $m$ et $b$ sont obtenus en résolvant le problème d'optimisation associé au SVM.

# Régression : Fusion de régresseurs

- Plusieurs features peuvent être extraites d'un même morceau.
- Chaque représentation explique un aspect de l'émotion perçue.
- Idée : aggréger linéairement les régresseurs obtenus pour chaque clip.

# Régression : Fusion de régresseurs

- Les T représentations du morceau $d_s$ sont notées $x_s^t, t = 1, ..., T$
- Pour chaque représentation on entraîne un régresseur $f_{ij}^t$ sur une base de donnée de morceaux annotés $(x_s, y_{i,j}^s)_{s=1,...,N}$
- Le régresseur final sera une moyenne pondérée de ces régresseurs :

$$\hat{y_{ij}^s} = \sum_{t=1}^{T} w_{ij}^t f_{ij}^t(x_s^t)$$

.

# Régression : Choix des pondérations

- Exemple de modèle : EQM gaussienne :

$$||f^* - \hat{f}|| \sim \mathcal{N}(\lambda, \sigma^2)$$

- On estime lambda par l'inverse de l'EQM empirique

$$\lambda_{ij}^t = \frac{N}{\sum_{s=1}^{N} (y_{i,j}^s - f_{i,j}^t(x_s^t))^2}$$

- On en déduit les poids $w_{ij}^t = \frac{\lambda_{ij}^t}{\sum_{t'=1}^{T} \lambda_{ij}^{t'}}$
- Autre possibilité : coefficient de détermination ($R^2$)

# Choix des features

| Feature Set | # | Features |
|---|---|---|
| Melody/ Harmony | 10 | Salient pitch, chromagram center, key clarity, mode, and harmonic change [60]. Take mean and standard deviation for temporal integration [61]. |
| Spectral | 10 | 32 spectral flatness measures, 32 spectral crest factors [62], and 26 Mel-scale frequency cepstral coefficients [63]. Take the first ten principal components for dimension reduction [47]. |
| Temporal | 6 | Mean and standard deviation of zero-crossing rate, temporal centroid, and log attack time [64]. |
| Rhythmic | 10 | 60-bin rhythm histogram and average tempo [65]. Take the first ten principal components. |
| Lyrics | 10 | Probability distribution over ten latent topics [66]. |

FIGURE 4 – Tableau récapitulatif des features utilisées

# Evaluation

- Plusieurs mesures de performances :
- Coefficient de détermination : indique à quel point le modèle explique la variance des observations. $\leftarrow R^2 \approx 0.54 enmoyenne$
- Distance entre distribution prédite et réalité terrain.

# Commentaires

- On aurait pu comparer le SVR avec d'autres régresseurs.
- Un certain nombre de paramètres sont à adapter empiriquement dans le modèle.
- Approche "spécifique" au genre considéré (pop anglaise).

# Conclusion

- Première méthode modélisant l'émotion comme une distribution.
- Les résultats sont relativement proches de la réalité-terrain.
- Répond en partie au problème de subjectivité en reconnaissance d'émotions.

Merci de votre attention