Séparation de sources appliquée à la musique

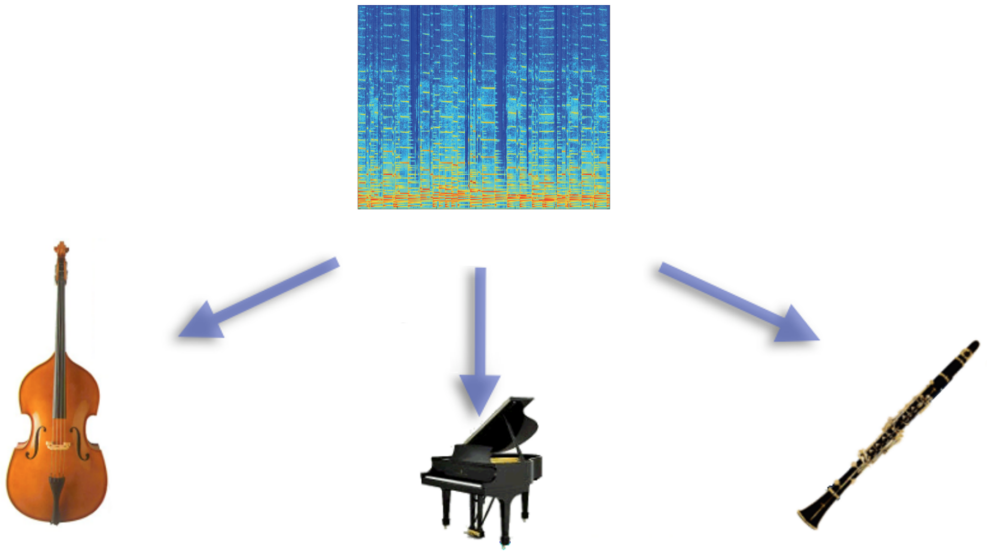
Signal et IA

Romain Hennequin

ENSEA, 2019

Deezer Research & Developement

Formalisation de la séparation de sources



On dispose d'un signal de mélange x contenant I différentes sources s_i (généralement des instruments de musique).

On suppose généralement que le mélange est linéaire :

$$x(t) = \sum_{i=1}^{I} s_i(t)$$

Formalisation de la séparation de sources

La TFCT étant linéaire, on dispose de la même formule dans le domaine temps/fréquence :

$$X(f,t) = \sum_{i=1}^{I} S_i(f,t)$$

La séparation de sources musicales est très largement traitée dans la littérature par l'estimation des $S_i(f,t)$.

Formalisation de la séparation de sources

Masquage : on peut chercher à estimer des masques $M_i(f,t) \in [0,1]$ à appliquer à $X(f,t)$.

Les sources estimées $s_i^{\text{est}}(t)$ sont alors recalculés par TFCT inverse (overlap and add) de $S_i^{\text{est}}(f,t) = M_i(f,t)X(f,t)$

Remarque : $S_i^{\text{est}}(f,t)$ n'est généralement pas la TFCT d'un signal (en général : $\text{TFCT}(\text{TFCT}^{-1}(S_i^{\text{est}})) \neq S_i^{\text{est}}$).

Masque

Les masques peuvent être binaires ($M_i(f,t) \in \{0,1\}$) ou continue ($M_i(f,t) \in [0,1]$)

Si on impose la contrainte que $\sum_i M_i(f,t) = 1$ alors :

$$\sum_{i=0}^{I} s_i^{\text{est}}(t) = x(t)$$

Masque

Une façon standard d'estimer des masques est d'estimer l'amplitude de $|S_i(f,t)^{\text{est}}| \approx |S_i(f,t)|$ et d'utiliser le masque de Wiener :

$$M_i(f,t) = \frac{|S_i^{\text{est}}(f,t)|^2}{\sum_{i=0}^{I} |S_i^{\text{est}}(f,t)|^2}$$

La phase de la TFCT des sources étant généralement difficile à estimer, il est pratique de se limiter à l'estimation de l'amplitude.

NMF

NMF d'un spectrogramme d'amplitude :

$$|X|^{\odot 2} = \mathbf{V} \approx \mathbf{WH}$$

Application en séparation de sources :

$$\mathbf{V} \approx \sum_{i=1}^{I} \mathbf{W}_i \mathbf{H}_i$$

On crée alors un masque pour chaque instrument :

$$M_i = \frac{(\mathbf{W}_i \mathbf{H}_i)}{(\mathbf{WH})} = \frac{(\mathbf{W}_i \mathbf{H}_i)}{(\sum_{i=1}^{I} \mathbf{W}_i \mathbf{H}_i)}$$

NMF

Approche non supervisée :

- Apprentissage des bases spectrales $\mathbf{W}$ et des activations $\mathbf{H}$.
- Clustering des $\mathbf{W}$ en I clusters correspondant aux instruments.

NMF

Approche semi-supervisée :

- Apprentissage des bases spectrales $\mathbf{W}_i$ pour chaque instrument sur une base d'apprentissage.
- Les $\mathbf{W}_i$ sont ensuite fixés et on apprend uniquement les activations $\mathbf{H}$ sur les signaux à séparer.

NMF

Approche par modèle acoustique :

- Décomposition avancée reposant sur un modèle acoustique
- Généralement, modèle spécifique à un instrument.

Exemple : modèle source/filtre :

$$\mathbf{V}_{\text{voix}} = (\mathbf{W}_{\text{source}}\mathbf{H}_{\text{source}}) \odot (\mathbf{W}_{\text{filtre}}\mathbf{H}_{\text{filtre}})$$

$\mathbf{W}_{source}$: motifs harmoniques à toutes les fréquences fondamentales d'intérêt. $\mathbf{W}_{filtre}$: filtres lisses.

Le signal de mélange est modélisé par $\mathbf{V} = \mathbf{V}_{\text{voix}} + \mathbf{V}_{\text{reste}}$ où $\mathbf{V}_{\text{reste}} = \mathbf{W}_{\text{reste}}\mathbf{H}_{\text{reste}}$

Approche supervisée

Si on ne s'intéresse qu'à un ensemble fixé d'instruments et si on dispose d'une base d'apprentissage avec les pistes séparées pour ces instruments, le problème de séparation peut être reformulé comme un problème purement supervisé.

- Estimation des S par minimisation d'une distance entre $f_i(X) = S_i^{\text{est}}$ et S_i sur la base d'apprentissage
- La fonction f_i est typiquement un réseau de neurones.

Variante : f_i estime directement un masque.

Approche supervisée

Démo