RCP217 – Apprentissage profond pour les données audio

Motivations, représentations du son, réseaux de neurones pour l'audio

---

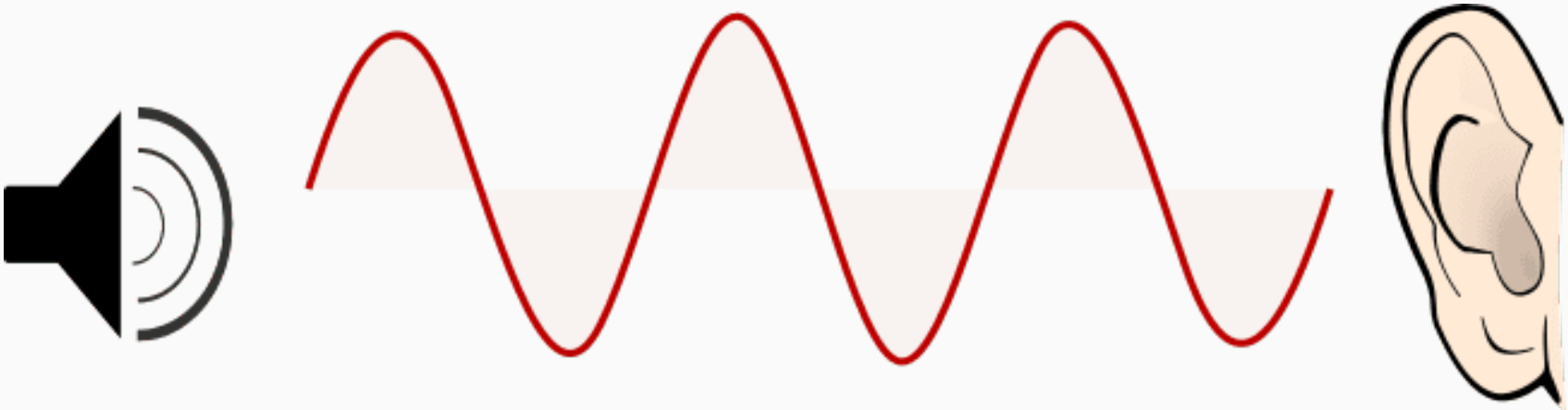
Nicolas Audebert `nicolas.audebert@lecnam.net`

22 février 2021

Conservatoire national des arts & métiers

Motivations

Qu’est-ce qu’un son ?

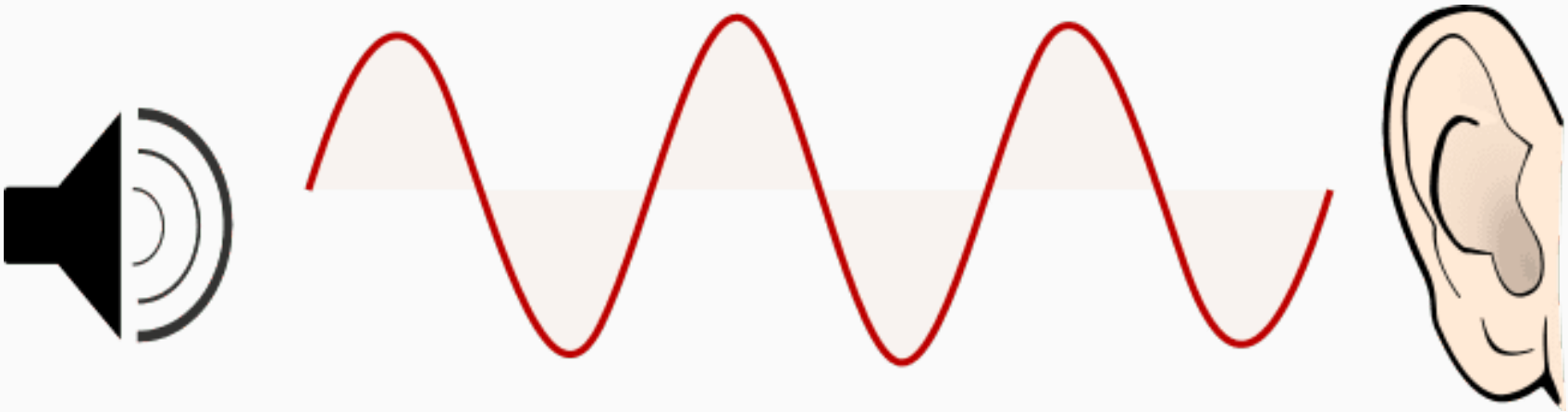


Un son est la **propagation audible d’une onde de pression** dans un fluide : alternance de compression et de dépression.

Audition humaine

Le système auditif humain est en mesure de percevoir les sons entre 20 Hz et 20 kHz (dans l’air).

Mesurer les sons



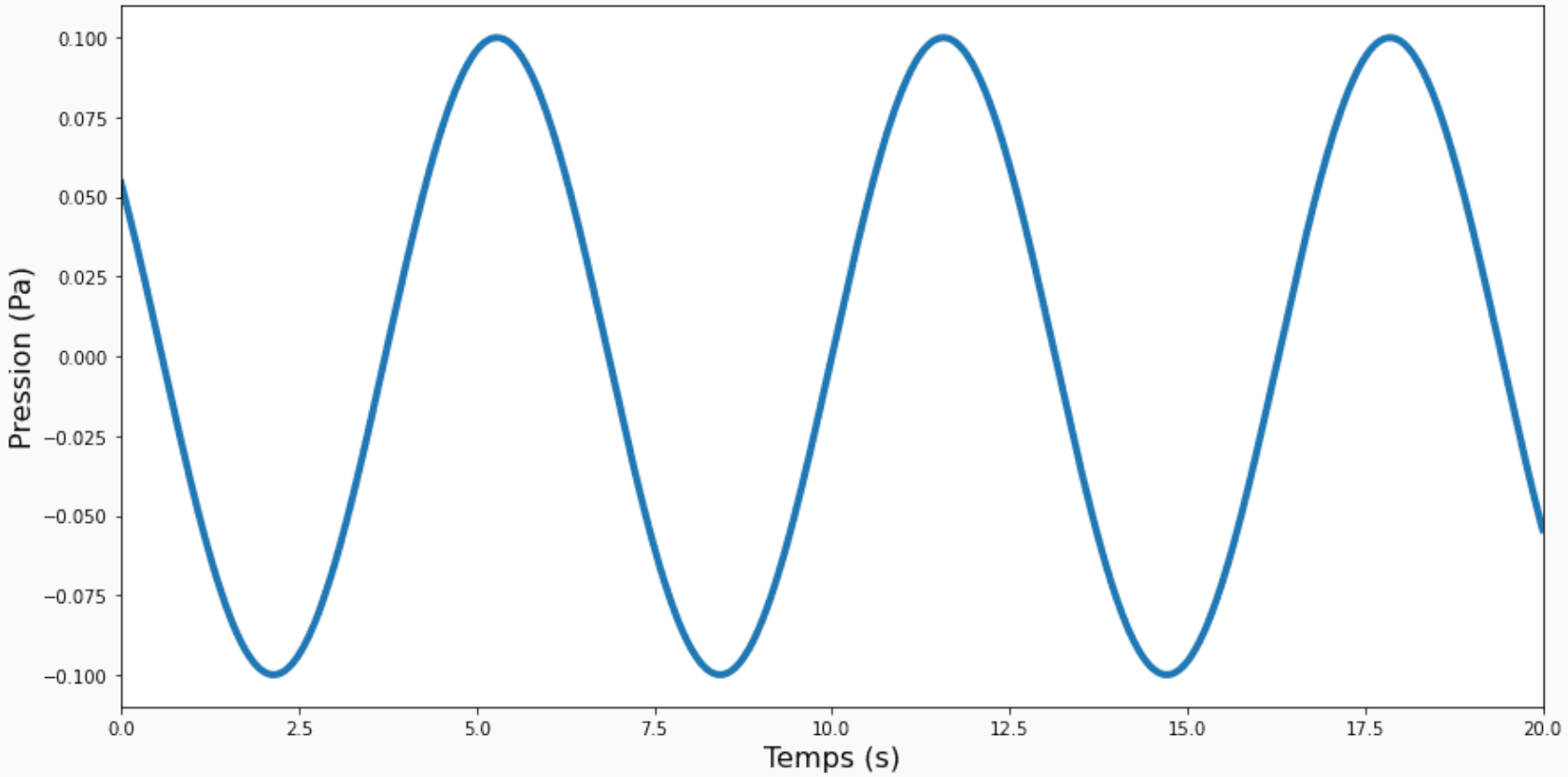
Dans l'air, l'onde de pression se déplace à $340\,\mathrm{m\,s^{-1}}$.

Le son perçu correspond à la succession de compressions/dépressions ayant lieu au point de mesure (= le tympan).

Signal sonore

On parle d'un signal sonore pour décrire l'**évolution de l'onde en un point de l'espace donné au cours du temps**.

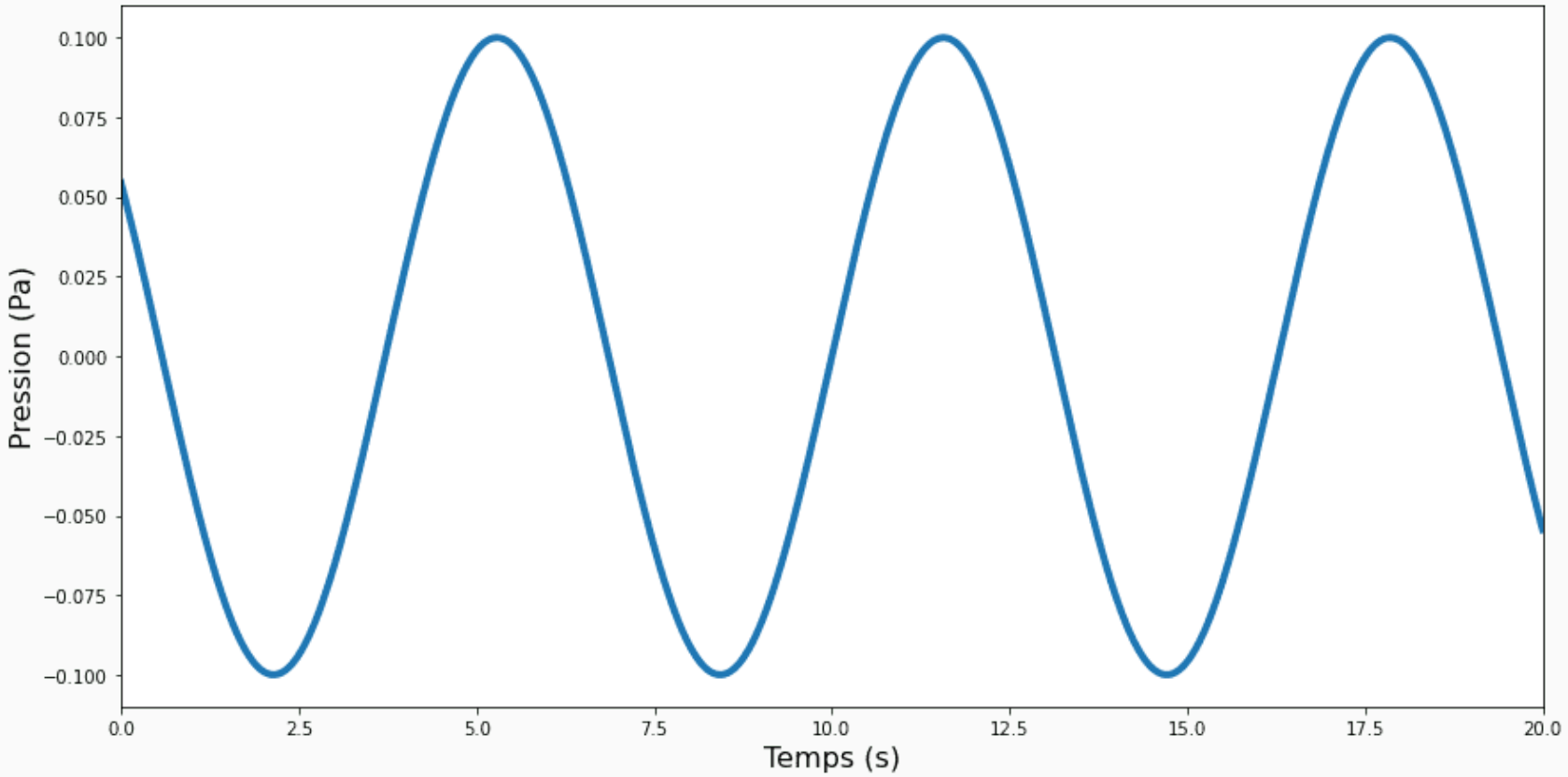
La forme d'onde



Définition

L'évolution de la valeur de pression en un point forme un signal que l'on appelle la **forme d'onde** (*waveform*).

La forme d'onde



Définition

L'évolution de la valeur de pression en un point forme un signal que l'on appelle la **forme d'onde** (*waveform*).

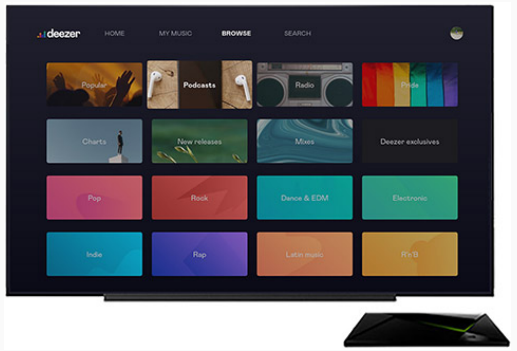
Exemples d'applications de l'IA aux sons

Langage parlé

- Commande vocale
- Transcription (sous-titrage)
- Traduction

Musique

- Catégorisation d'œuvres musicales
- Transcription en partition
- Séparation voix/instruments



Acoustique

- Débruitage
- Identification de la faune

Modélisation sous forme de problème décisionnel

Entrée

Observation $x \in \mathcal{S}$ l'ensemble des extraits sonores.

Attention ! Ces extraits peuvent être de longueur variable.

Modélisation décisionnelle

Deux possibilités :

1. Extraire des caractéristiques indépendantes de la longueur du son à l'aide d'une fonction $f : \mathcal{S} \to \mathbb{R}^n$
2. Utiliser du *padding* pour ne travailler que sur des sons de longueur constante.

Sortie : dépend de la tâche à réaliser.

Classification

Classification (ou classement) : problème décisionnel pour lequel la variable à expliquer est nominale (à valeurs discrètes).

Commande vocale

Détection (classification bruit/voix) et reconnaissance d'une commande (classification).

Problème de classification à k classes : $f : \mathcal{S} \to \{1, \ldots, k\}$

Reconnaissance du style musical, de l'instrument, du type de bruit...

Problème de classification à k classes : $f : \mathcal{S} \to \{1, \ldots, k\}$

Modélisation séquence à séquence

Modélisation séquentielle : la variable à expliquer est une séquence (de longueur variable) à valeurs discrètes.

Transcription automatique (sous-titrage)

Prédire une séquences de mots textuels correspondant aux mots parlés : $f : \mathcal{S} \to \{1, \ldots, k\}^m$

Problème inverse : synthèse vocale (voir le chapitre de RCP211 sur les modèles génératifs).

Transcription musicale

Prédire la séquence de notes (hauteur et durée) du son.

Régression

Séparation de sources

Exemples : séparer les différents instruments d'une chanson, séparer les voix d'une conversation (*cocktail party problem*)...

Problème de régression multivarié : l'entrée est une piste audio et la sortie est n pistes.

$$f(s) = \{s'_1, \ldots, s'_m\} \text{ avec } \sum_{i=1}^{m} s'(i) = s$$

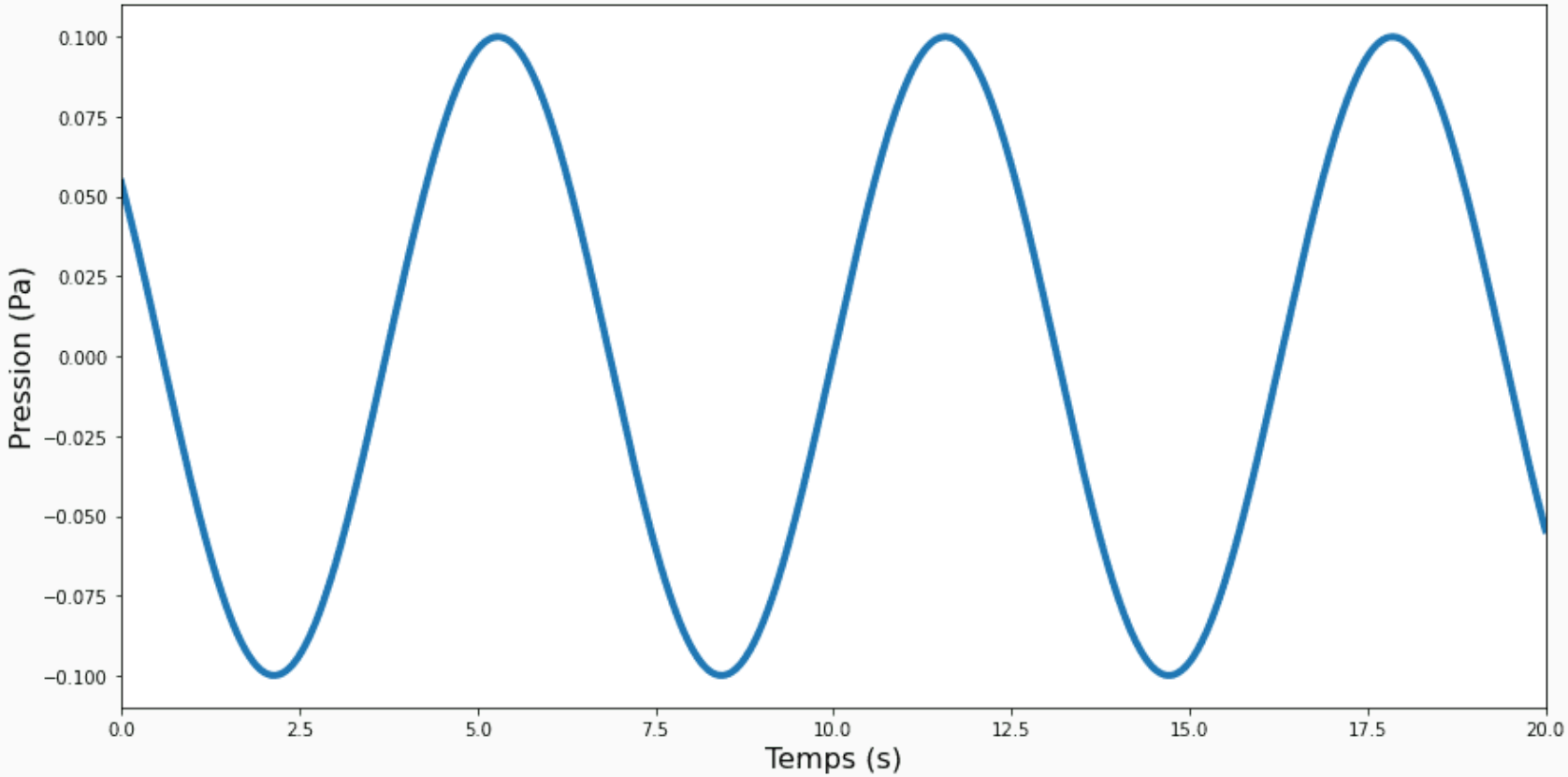
Débruitage

Retirer le bruit ϵ qui affecte le signal réel s :

$$f(\hat{s}) = f(s + \epsilon) = s$$

Représentations des signaux audio

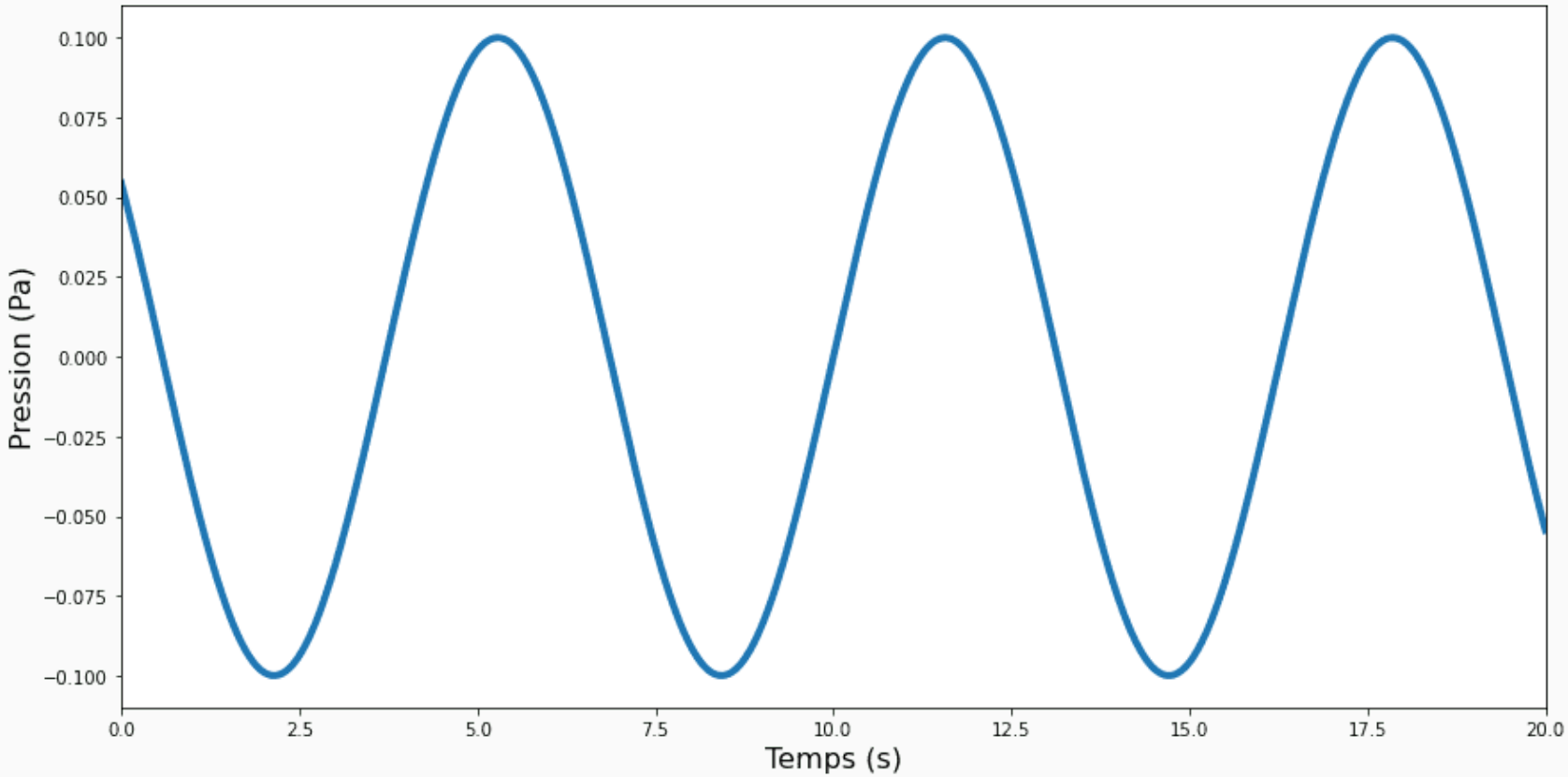
La forme d'onde



La forme d'onde est la représentation classique d'un signal audio. Elle représente l'évolution dans le temps de la pression au voisinage du point de mesure.

Il s'agit d'une **série temporelle** : $s : t \in \mathbb{R}^+ \to s(t) \in \mathbb{R}$

La forme d'onde



La forme d'onde est la représentation classique d'un signal audio. Elle représente l'évolution dans le temps de la pression au voisinage du point de mesure.

Il s'agit d'une **série temporelle** : $s : t \in \mathbb{R}^+ \to s(t) \in \mathbb{R}$

Transformée de Fourier

Définition

Soit s une fonction intégrable sur $\mathbb{R}$. On appelle transformée de Fourier de s la fonction $\hat{s}$ définie par :

$$\hat{s} : \nu \to \hat{s}(\nu) = \int_{-\infty}^{+\infty} s(t) \mathrm{e}^{-\mathrm{i}2\pi\nu t} \mathrm{dt}$$

Physiquement, $\hat{s}(\nu)$ représente l'énergie du signal s à la fréquence ν.

Interprétation

La transformée de Fourier donne une représentation fréquentielle (ou spectrale) d'un signal.

Transformée de Fourier

Définition

Soit s une fonction intégrable sur $\mathbb{R}$. On appelle transformée de Fourier de s la fonction $\hat{s}$ définie par :

$$\hat{s} : \nu \to \hat{s}(\nu) = \int_{-\infty}^{+\infty} s(t) \mathrm{e}^{-\mathrm{i}2\pi\nu t} \mathrm{dt}$$

Physiquement, $\hat{s}(\nu)$ représente l'énergie du signal s à la fréquence ν.

Interprétation

La transformée de Fourier donne une représentation fréquentielle (ou spectrale) d'un signal.

Transformée de Fourier

Définition

Soit s une fonction intégrable sur $\mathbb{R}$. On appelle transformée de Fourier de s la fonction $\hat{s}$ définie par :

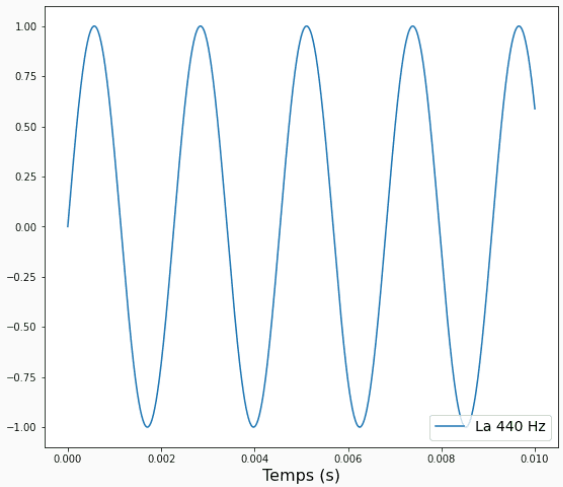
$$\hat{s} : \nu \to \hat{s}(\nu) = \int_{-\infty}^{+\infty} s(t) \mathrm{e}^{-\mathrm{i}2\pi\nu t} \mathrm{dt}$$

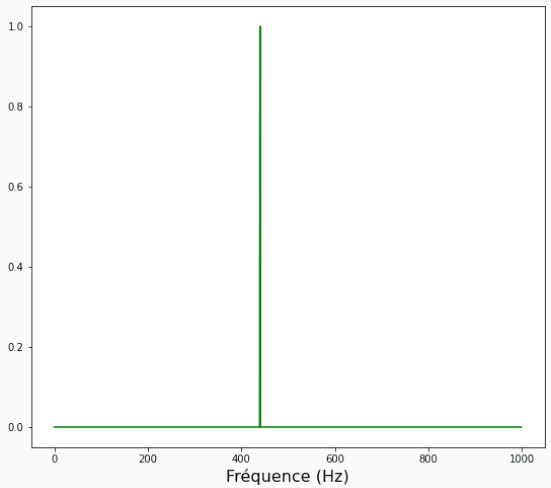
Physiquement, $\hat{s}(\nu)$ représente l'énergie du signal s à la fréquence ν.

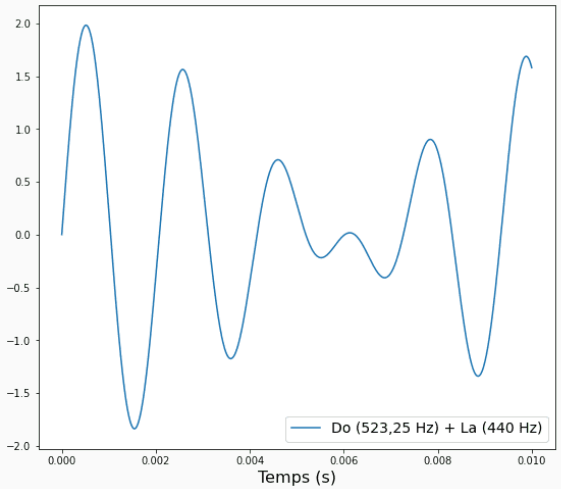
Interprétation

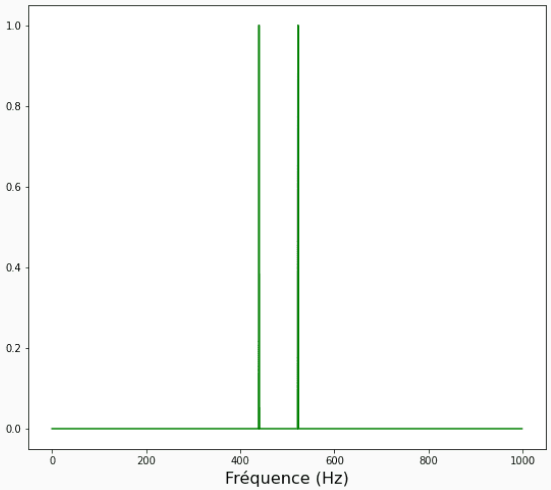
La transformée de Fourier donne une représentation fréquentielle (ou spectrale) d'un signal.

Exemple sur des notes pures





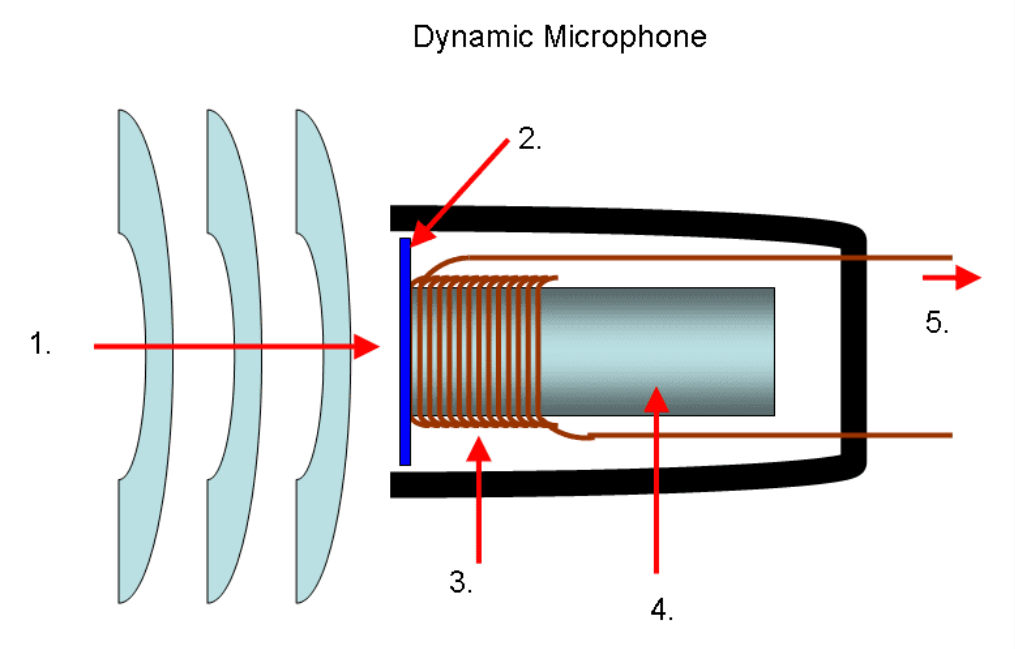




En réalité, le spectre d'un son riche est impacté par ses harmoniques, son timbre, le bruit ambiant...

Représentation numérique des signaux

En pratique, un microphone convertit le signal acoustique en pression en signal électrique.



1. onde sonore, 2. membrane, 3. bobine mobile, 4. aimant, 5. signal électrique

Banco, Wikimédia Commons

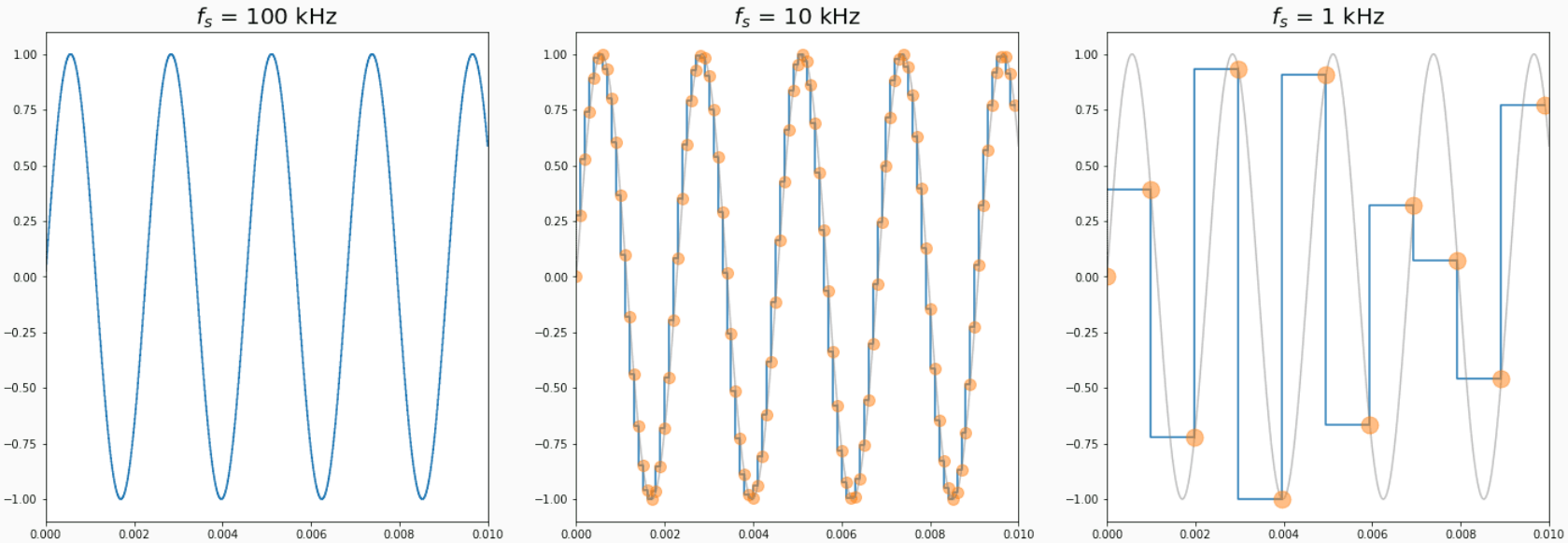
Encodage numérique

Pour numériser le son, on mesure le signal électrique toutes les T secondes, à une fréquence de $f_s = 1/T$ Hz.

f_s est appelée la **fréquence d'échantillonnage**.

Fréquence d'échantillonnage

La fréquence d'échantillonnage ou *sampling rate* correspond à l'inverse de l'intervalle de temps entre deux mesures.

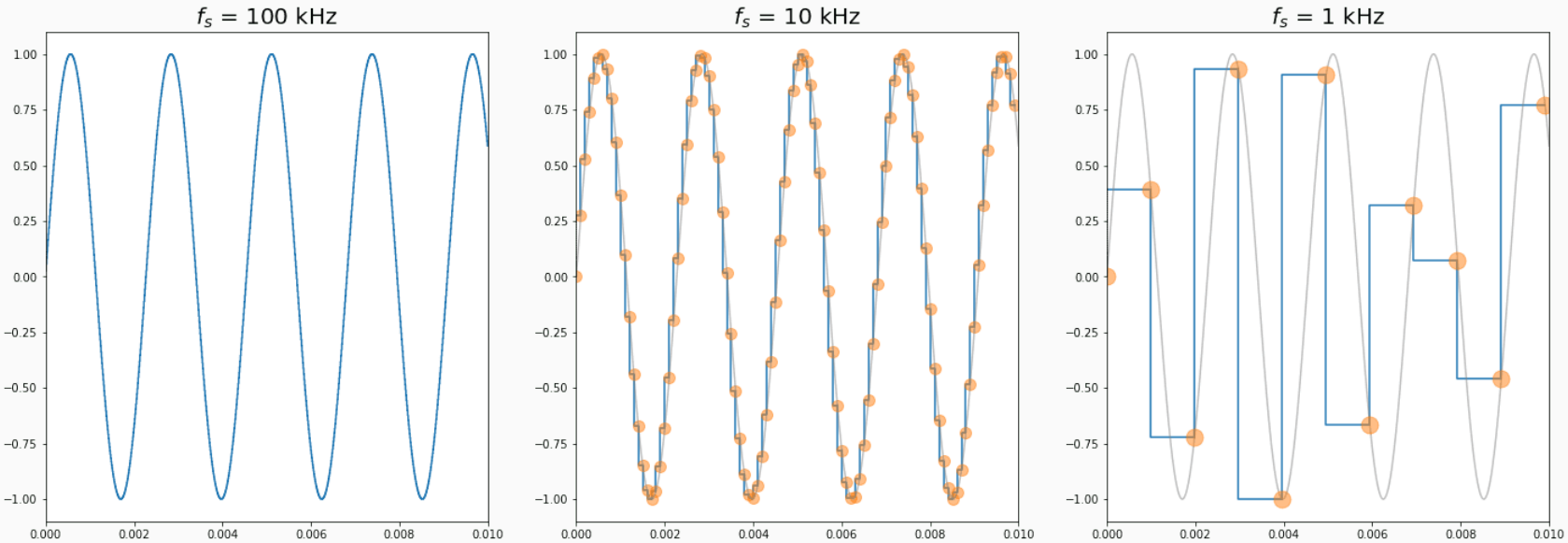


Échantillonnage d'un La (440 Hz) à différentes fréquences.

Plus f_s augmente, plus le signal numérisé est proche du signal continu (mais plus le nombre de points à stocker est grand).

Fréquence d'échantillonnage

La fréquence d'échantillonnage ou *sampling rate* correspond à l'inverse de l'intervalle de temps entre deux mesures.



Échantillonnage d'un La (440 Hz) à différentes fréquences.

Plus f_S augmente, plus le signal numérisé est proche du signal continu (mais plus le nombre de points à stocker est grand).

Équivalence de Shannon-Nyquist

Numérique ou analogique ?

Est-ce problématique d'utiliser le signal discrétisé ?

Soit un signal s contenant une plage de fréquence bornée, i.e. dont la transformée de Fourier a un support fini * $\subset [f_{\text{min}}, f_{\text{max}}]$.

Théorème de Nyquist-Shannon

Si ce signal est discrétisé à une fréquence d'échantillonnage $f_s \geq 2(f_{\text{max}} - f_{\text{min}})$, alors il est possible de reconstruire parfaitement le signal s continu initial.

$\Longrightarrow$ il y a équivalence entre les représentations analogiques et numériques d'un signal.

*. Ce qui est vrai quand on ne considère que les fréquences audibles !

Équivalence de Shannon-Nyquist

Numérique ou analogique ?

Est-ce problématique d'utiliser le signal discrétisé ?

Soit un signal s contenant une plage de fréquence bornée, i.e. dont la transformée de Fourier a un support fini * $\subset [f_{\text{min}}, f_{\text{max}}]$.

Théorème de Nyquist-Shannon

Si ce signal est discrétisé à une fréquence d'échantillonnage $f_s \geq 2(f_{\text{max}} - f_{\text{min}})$, alors il est possible de reconstruire parfaitement le signal s continu initial.

$\Longrightarrow$ il y a équivalence entre les représentations analogiques et numériques d'un signal.

*. Ce qui est vrai quand on ne considère que les fréquences audibles !

Équivalence de Shannon-Nyquist

Numérique ou analogique ?

Est-ce problématique d'utiliser le signal discrétisé ?

Soit un signal s contenant une plage de fréquence bornée, i.e. dont la transformée de Fourier a un support fini * $\subset [f_{\text{min}}, f_{\text{max}}]$.

Théorème de Nyquist-Shannon

Si ce signal est discrétisé à une fréquence d'échantillonnage $f_s \geq 2(f_{\text{max}} - f_{\text{min}})$, alors il est possible de reconstruire parfaitement le signal s continu initial.

$\Longrightarrow$ il y a équivalence entre les représentations analogiques et numériques d'un signal.

*. Ce qui est vrai quand on ne considère que les fréquences audibles !

Transformée de Fourier discrète

Définition

Soit s un signal échantillonné régulièrement représenté par une série finie de T valeurs. On appelle transformée de Fourier discrète de s la série S :

$$S(k) = \sum_{t=0}^{T-1} s[t] \mathrm{e}^{-\mathrm{i}2\pi k \frac{t}{T}} \text{ pour } 0 \leq k \leq T.$$

C'est la transposition discrète (numérique) de la définition de la transformée de Fourier continue (analogique).

Encodage de la forme d'onde : méthode PCM

Pulse-code modulation

1. Échantillonnage du signal analogique,
2. Quantification des valeurs dans un nombre fini d'intervalles,
3. Codage binaire de l'intervalle d'appartenance.

Il faut ensuite prendre en compte plusieurs informations :

- la « profondeur » des valeurs prises (*bit-depth*) : on peut encoder la hauteur du signal sur k bits (généralement une dizaine) = équilibre précision/mémoire,
- le nombre de pistes : mono (1 piste), stéréo (2 pistes), 5.1 (5 pistes + basse)...
- le pas d'échantillonnage (généralement $\geq 44\,\text{kHz}$).

Encodage de la forme d'onde : méthode PCM

Pulse-code modulation

1. Échantillonnage du signal analogique,
2. Quantification des valeurs dans un nombre fini d'intervalles,
3. Codage binaire de l'intervalle d'appartenance.

Il faut ensuite prendre en compte plusieurs informations :

- la « profondeur » des valeurs prises (*bit-depth*) : on peut encoder la hauteur du signal sur k bits (généralement une dizaine) = équilibre précision/mémoire,
- le nombre de pistes : mono (1 piste), stéréo (2 pistes), 5.1 (5 pistes + basse)...
- le pas d'échantillonnage (généralement $\geq 44\,\text{kHz}$).

Encodage de la forme d'onde : méthode PCM

Pulse-code modulation

1. Échantillonnage du signal analogique,
2. Quantification des valeurs dans un nombre fini d'intervalles,
3. Codage binaire de l'intervalle d'appartenance.

Il faut ensuite prendre en compte plusieurs informations :

- la « profondeur » des valeurs prises (*bit-depth*) : on peut encoder la hauteur du signal sur k bits (généralement une dizaine) = équilibre précision/mémoire,
- le nombre de pistes : mono (1 piste), stéréo (2 pistes), 5.1 (5 pistes + basse)...
- le pas d'échantillonnage (généralement $\geq 44\,\text{kHz}$).

Encodage de la forme d'onde : méthode PCM

Pulse-code modulation

1. Échantillonnage du signal analogique,
2. Quantification des valeurs dans un nombre fini d'intervalles,
3. Codage binaire de l'intervalle d'appartenance.

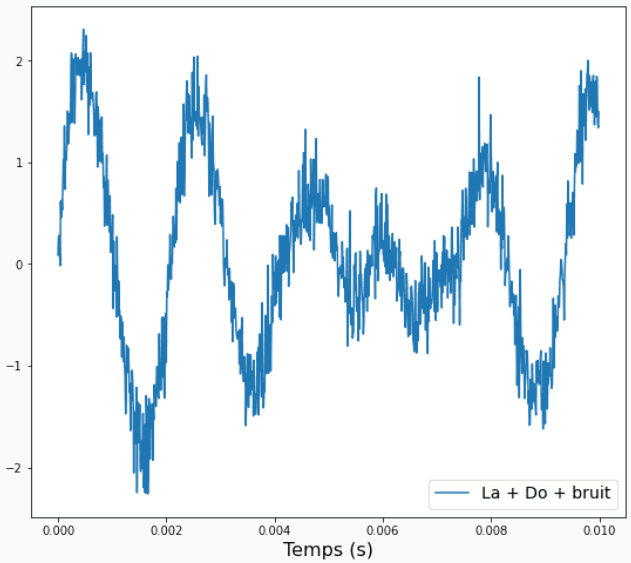
Il faut ensuite prendre en compte plusieurs informations :

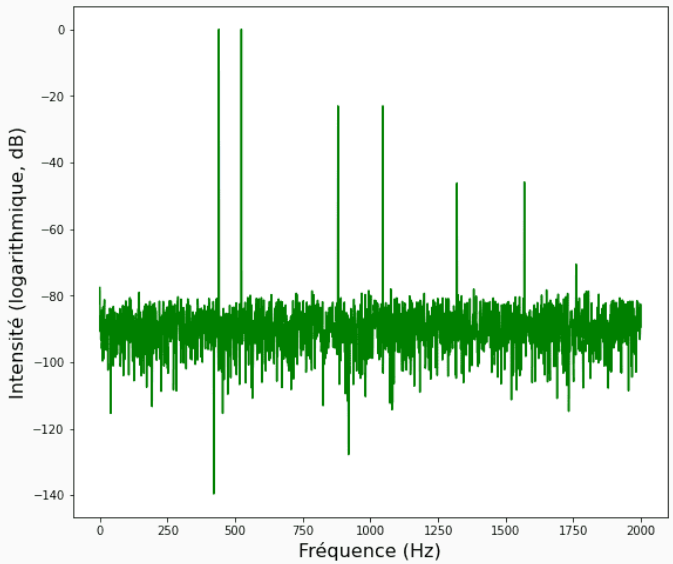
- la « profondeur » des valeurs prises (*bit-depth*) : on peut encoder la hauteur du signal sur k bits (généralement une dizaine) = équilibre précision/mémoire,
- le nombre de pistes : mono (1 piste), stéréo (2 pistes), 5.1 (5 pistes + basse)...
- le pas d'échantillonnage (généralement $\geq 44\,\text{kHz}$).

Spectrogramme

C'est le résultat de la transformée de Fourier (discrète) appliquée sur la forme d'onde.

L'énergie est exprimée en échelle logarithmique (dB) par rapport à la fréquence.





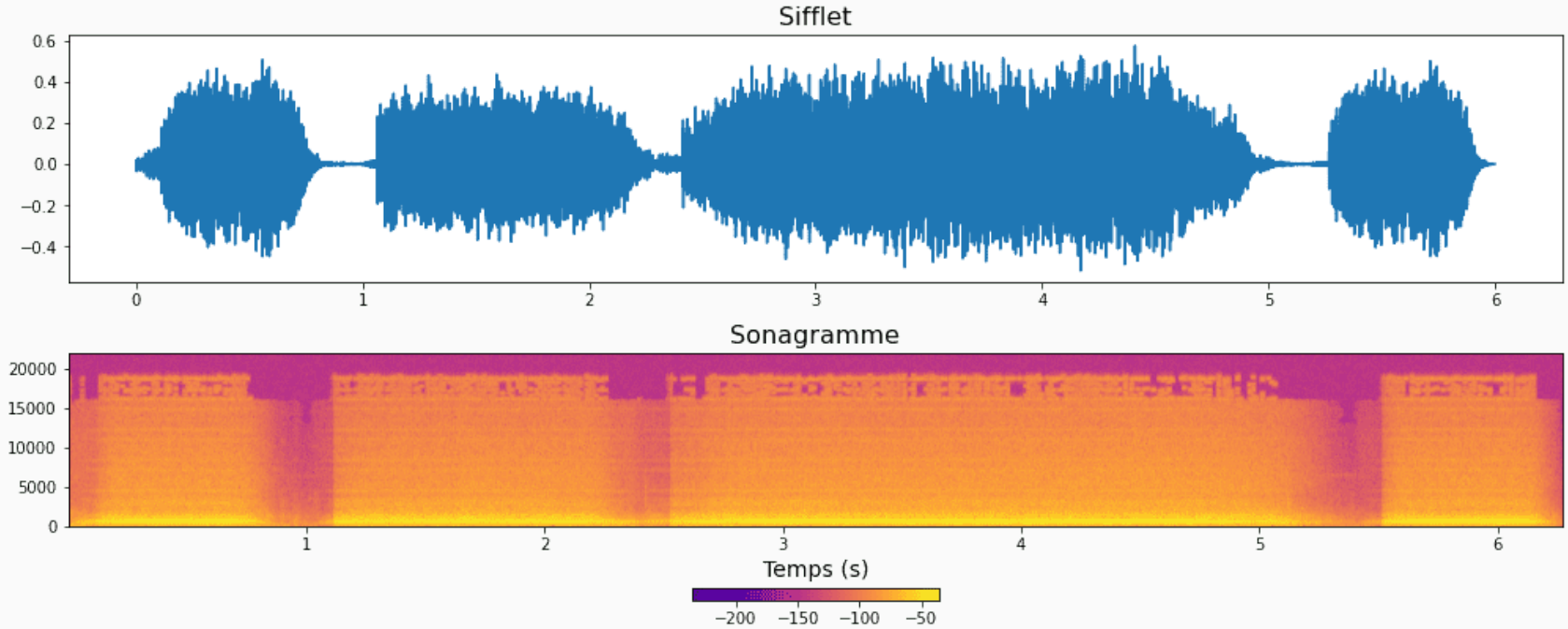
Inconvénient

L'information temporelle est cachée.

Sonagramme

Construction du sonagramme

Évolution du spectre dans le temps : transformée de Fourier sur la forme d'onde sur une fenêtre glissante $[t - \omega, t + \omega]$.

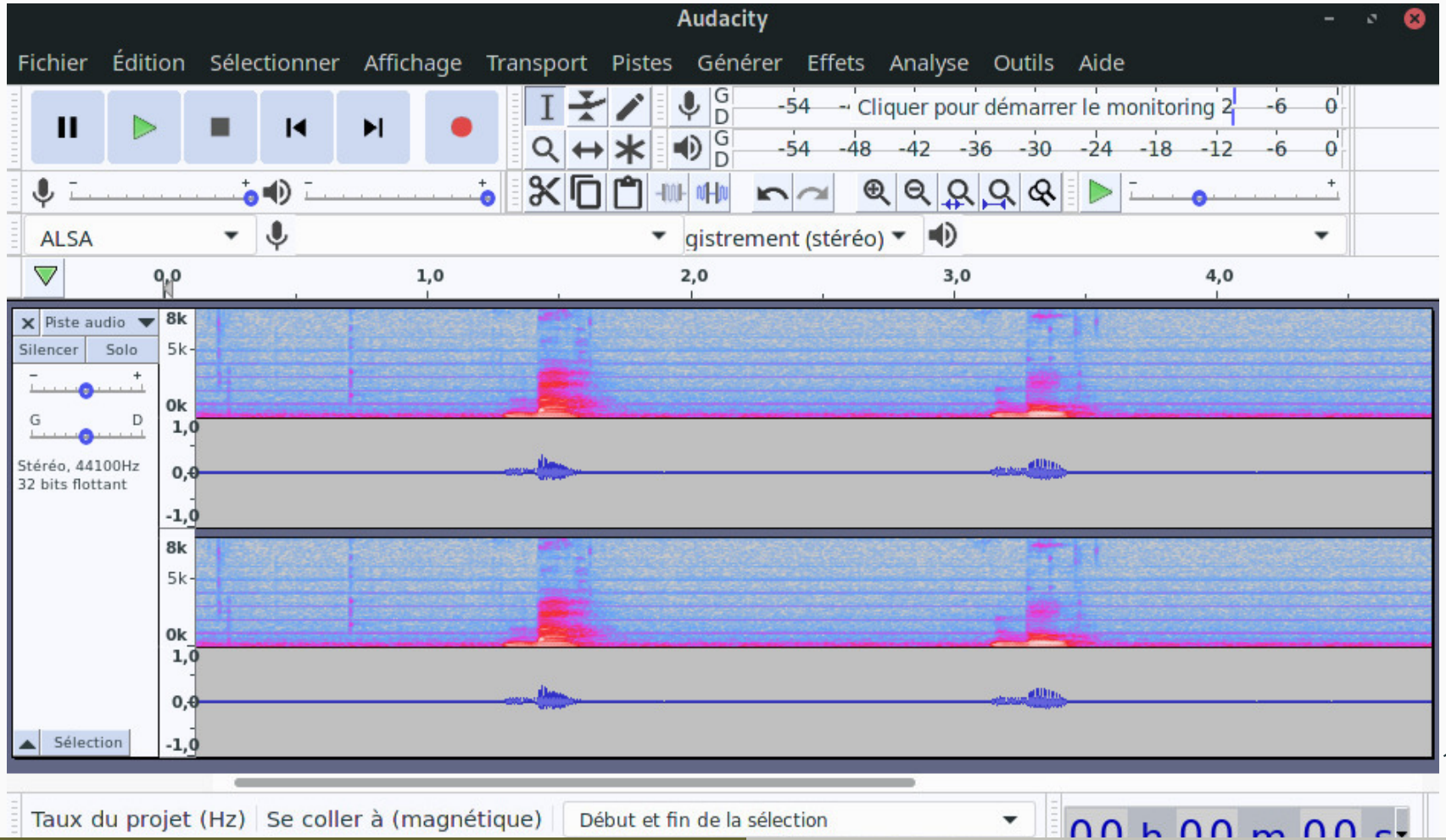


Temps en abscisse, fréquence en ordonnée. L'intensité d'une fréquence f au temps t est codée par une échelle de couleurs.

Démonstration

Audacity

Logiciel d'enregistrement audio libre et gratuit avec visualisation en forme d'onde et spectrogramme.



Échelle de Mel

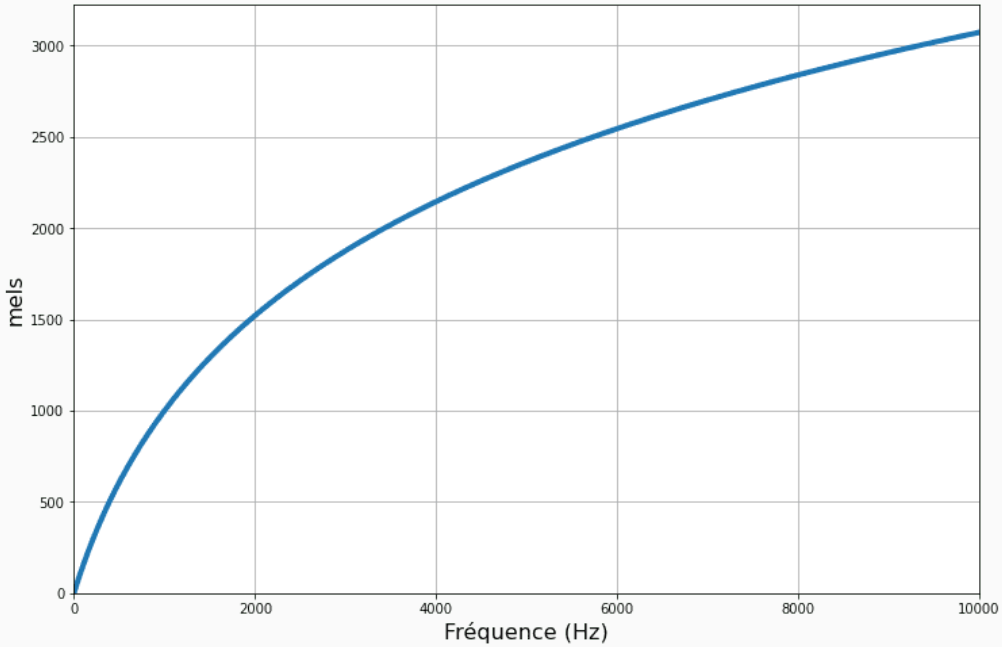
L'oreille humaine n'a pas une perception linéaire des sons : la sensibilité dépend de la fréquence.

On distingue mieux 300 Hz et 600 Hz que 10 kHz et 10.3 kHz.

mels $\leftrightarrow$ fréquence

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

$$f = 700 \cdot \left(10^{\frac{m}{2595}} - 1\right)$$



Mel-frequency cepstrum

MFCC

Les *Mel-Frequency Cepstral Coefficients* (MFCC) sont des descripteurs sonores très utilisés.

1. Appliquer la transformée de Fourier sur une fenêtre,
2. Projeter la puissance du spectre sur l'échelle de mel,
3. Calculer le log de la puissance à chaque fréquence de mel,
4. Réaliser la transformée en cosinus discrète (DCT) du signal formé par cette suite de valeurs.

Les coefficients cepstraux de mel (MFCC) sont les coefficients de la DCT.

Applications

Reconnaissance vocale dans les systèmes téléphoniques, estimation de la similarité musicale, etc.

Mel-frequency cepstrum

MFCC

Les *Mel-Frequency Cepstral Coefficients* (MFCC) sont des descripteurs sonores très utilisés.

1. Appliquer la transformée de Fourier sur une fenêtre,
2. Projeter la puissance du spectre sur l'échelle de mel,
3. Calculer le log de la puissance à chaque fréquence de mel,
4. Réaliser la transformée en cosinus discrète (DCT) du signal formé par cette suite de valeurs.

Les coefficients cepstraux de mel (MFCC) sont les coefficients de la DCT.

Applications

Reconnaissance vocale dans les systèmes téléphoniques, estimation de la similarité musicale, etc.

Formats de fichiers en pratique

D'après le théorème de Shannon-Nyquist, pour éviter la perte d'information on privilégie une fréquence d'échantillonnage d'au moins 44 kHz pour l'écoute humaine.

Format CD standard

Format CD standard : PCM 16 bits et 44.1 kHz

Divers formats audio

Non-compressés : WAV/PCM, AU...

Compressés sans perte : FLAC, ALAC...

Compressés avec perte : AAC, MP3, OGG Vorbis...

Les codecs audio utilisent généralement une notion de qualité perçue pour compresser plus les sections simples (les silences) et moins les sections riches.

Résumé

2 façons de représenter un signal audio. On passe de l'une à l'autre grâce à la **transformée de Fourier**.

Modèle temporel

Forme d'onde : la modélisation par défaut, utilisée pour coder le signal.

Modèle fréquentiel

Spectrogramme : visualise les propriétés fréquentielles mais perd l'information temporelle.

Sonagramme

Compromis entre les deux modélisations : spectrogramme calculé localement sur une courte fenêtre temporelle.

Il représente l'**évolution du spectre sonore dans le temps**.

Aller plus loin

Revue des caractéristiques pour les sons :

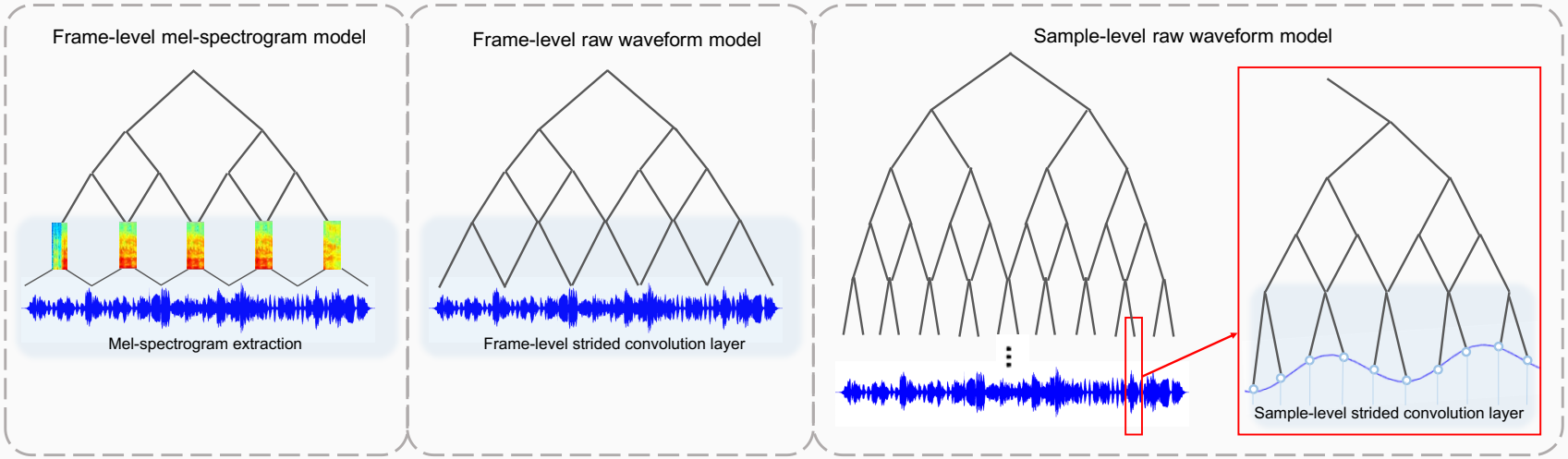
- *Environmental sound recognition : a survey*, Chachada et Kuo, 2014.
- *Speech Recognition using MFCC*, Ittichaichareon et al., 2012.
- *Musical Genre Classification of Audio Signals*, Tzanetakis et Cook, 2002.

Expérimenter avec l'audio :

- TP avec Audacity du parcours STMN de l'EICnam
- `librosa`, un paquet Python pour l'analyse audio et musicale.

Architectures de réseaux de neurones pour le son

Catégories d'architectures pour le son



Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms, Lee et al., 2017

3 grands types d'approches :

- Modèles sur le Mel-spectogram
- Modèles sur des fenêtres de la forme d'onde
- Modèles sur les échantillons de la forme d'onde

Modélisation

Principe central

La forme d'onde est une **série temporelle** : $s : \mathbb{R}^n \to \mathbb{R}$.

Les séries temp. seront examinées dans le prochain chapitre.

Ces séries unidimensionnelles sont manipulables par :

- des *Recurrent Neural Networks* (RNN),
- des *Convolutional Neural Networks* 1D (CNN),
- des perceptrons multicouches (si toutes les séries ont la même durée).

Spécificité des séries temporelles

Contrairement à un simple vecteur dans $\mathbb{R}^n$, $s(t)$ et $s(t \pm 1)$ sont fortement corrélés.

Il y a généralement un lien de causalité entre $s(0), \ldots, s(t-1)$ et $s(t)$.

Difficultés de la forme d'onde

Longueur de la séquence

$f_s \geq 44\,\text{kHz} \implies 1\,\text{s} = 44000\,\text{points}$

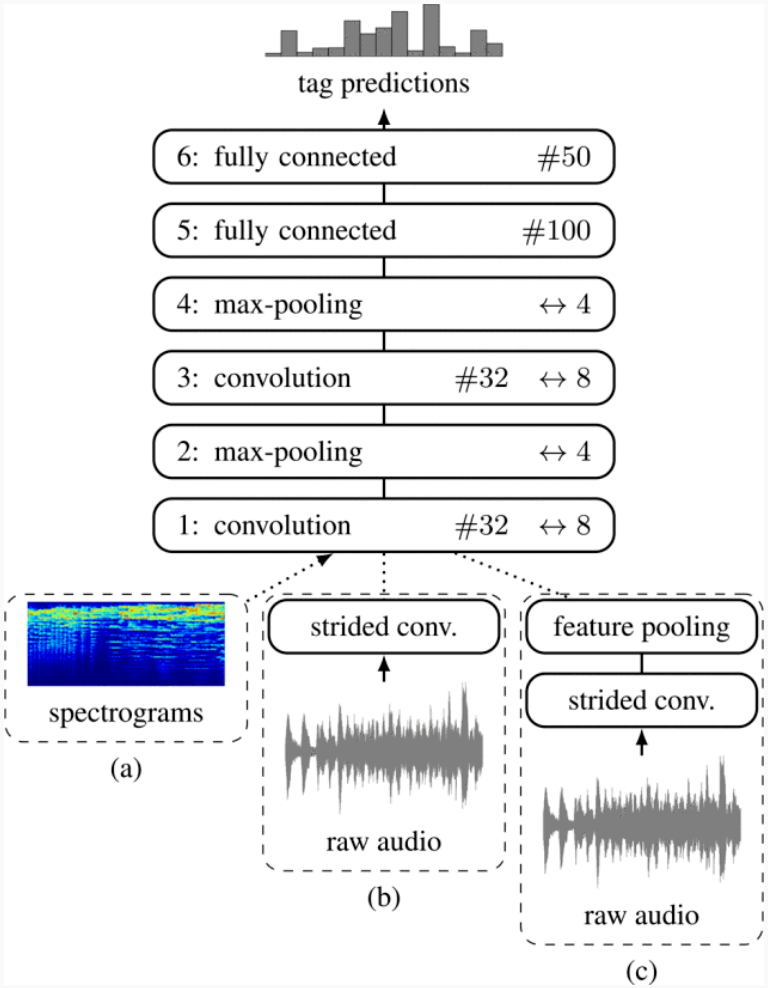
Mais l'algorithme de rétroprogation dans le temps n'est pas envisageable sur une séquence aussi longue $\implies$ difficile d'entraîner un RNN directement sur la *waveform*

Séquences de longueurs différentes

2 sons s_1 et s_2 n'ont généralement pas la même durée $\implies$ séquences de longueurs inégales.

Il est nécessaire que le réseau ne soit pas contraint à une taille d'entrée fixe (aisé pour les RNN, plus délicat pour les CNN).

Apprentissage d'une décomposition fréquentielle



End-to-end learning for music audio, Dieleman et Schrauwen, 2014.

Motivation

Les couches convolutives peuvent être considérées comme apprenant une **décomposition fréquentielle** ($\simeq$ spectrogramme) du signal.

Idée

On peut remplace le calcul du spectrogramme par une couche convolutive qui opère sur une fenêtre de même taille.

Gérer les durées différentes

Si on note i la dimension de l'entrée d'une couche convolutive, k la taille du noyau, p la longueur du *padding* et s le pas (*stride*) :

$$o = \left\lfloor \frac{i + 2p - k}{s} \right\rfloor + 1$$

$\Rightarrow$ deux séquences de longueurs i et i' produisent des *feature maps* de longueur o et o'

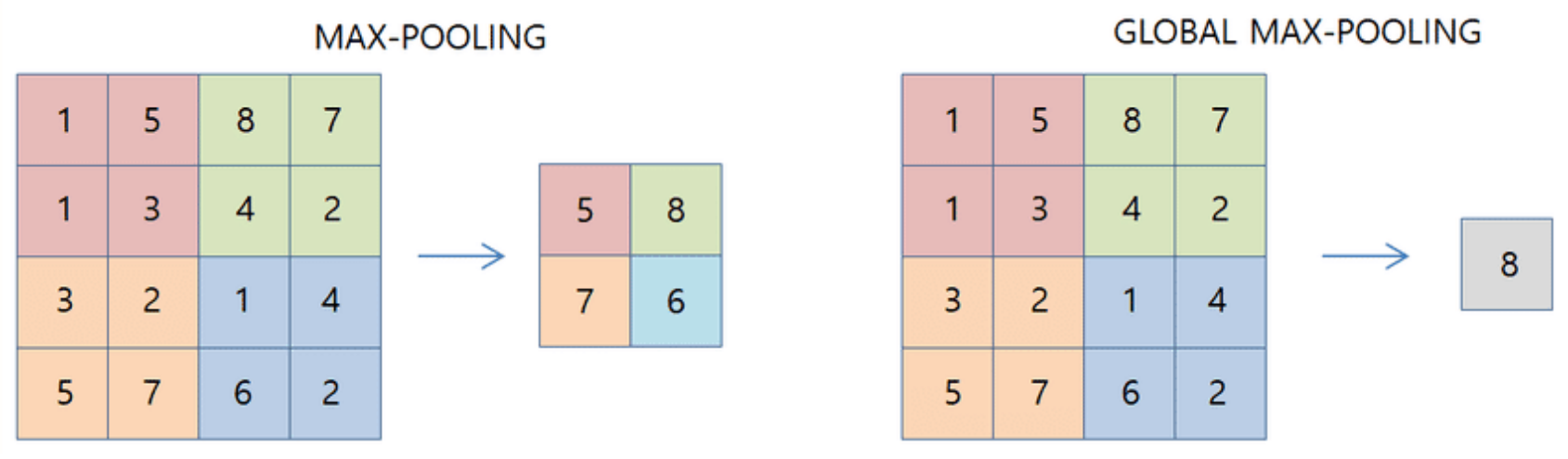
Problème

Le modèle doit pouvoir traiter un son de longueur variable.

Échantillonnage adaptatif

Adaptive pooling : on sous-échantillonne les *feature maps* à une taille fixe

Cas particulier : *Global Pooling* (max ou *average*)



Issu de Sentiment Classification Using Convolutional Neural Networks

Transforme une *feature map* $(n_c, o) \to (n_c)$ sur lequel on peut appliquer un classifieur (n_c ne dépend pas de i).

Batching

Pour la descente de gradient par *batch*, on veut un tenseur :

$$(b, n_{\text{canaux}}, \text{longueur})$$

Chaque élément du batch est une séquence s_i ($1 \leq i \leq b$).

Pour construire le tenseur, il faut que la longueur des s_i soit identique.

Solutions

- *Zero-padding* à la longueur $\max_i |s_i|$
- cf. chapitre suivant sur les séries temporelles

SampleCNN

3^9-SampleCNN Model			
59,049 Samples (2678 ms) as Input			
Layer	**Stride**	**Output**	**# of Params**
conv 3-128	3	19,683 × 128	512
conv 3-128 maxpool 3	1 3	$19{,}683 \times 128$ 6561×128	49,280
conv 3-128 maxpool 3	1 3	6561×128 2187×128	49,280
conv 3-256 maxpool 3	1 3	2187×256 729×256	98,560
conv 3-256 maxpool 3	1 3	729×256 243×256	196,864
conv 3-256 maxpool 3	1 3	243×256 81×256	196,864
conv 3-256 maxpool 3	1 3	81×256 27×256	196,864
conv 3-256 maxpool 3	1 3	27×256 9×256	196,864
conv 3-512 maxpool 3	1 3	9×512 3×512	393,728
conv 3-512 maxpool 3	1 3	3×512 1×512	786,944
conv 1-512 dropout 0.5	1 –	1×512 1×512	262,656
sigmoid	–	50	25,650
		Total params	2.46×10^6

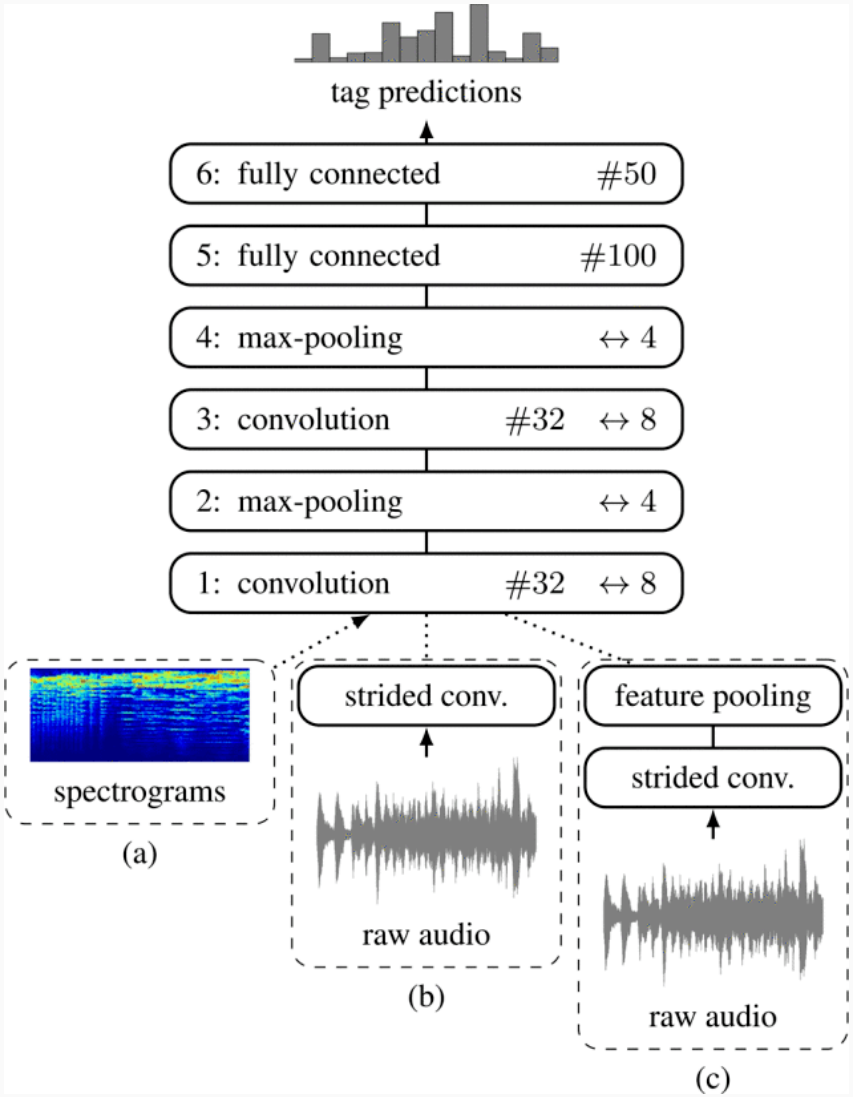
SampleCNN : End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification, Lee et al., 2018

$\rightarrow$ pour réduire la complexité, on utilise des convolutions avec une *stride* et des *maxpooling* temporels à chaque couche

$\rightarrow$ implémentation dans la dernière partie du TP

Réseaux de neurones sur le spectre

Approches unidimensionnelles



Sonagramme

Le sonagramme est une séquence de vecteurs :

$$t \in \mathbb{R} \rightarrow s(t) \in \mathbb{R}^p$$

Modèles unidimensionnels

Analogue en tous points aux CNN 1D/RNN sur la forme d'onde

End-to-end learning for music audio, Dieleman et Schrauwen, 2014.

Sonagrammes = images

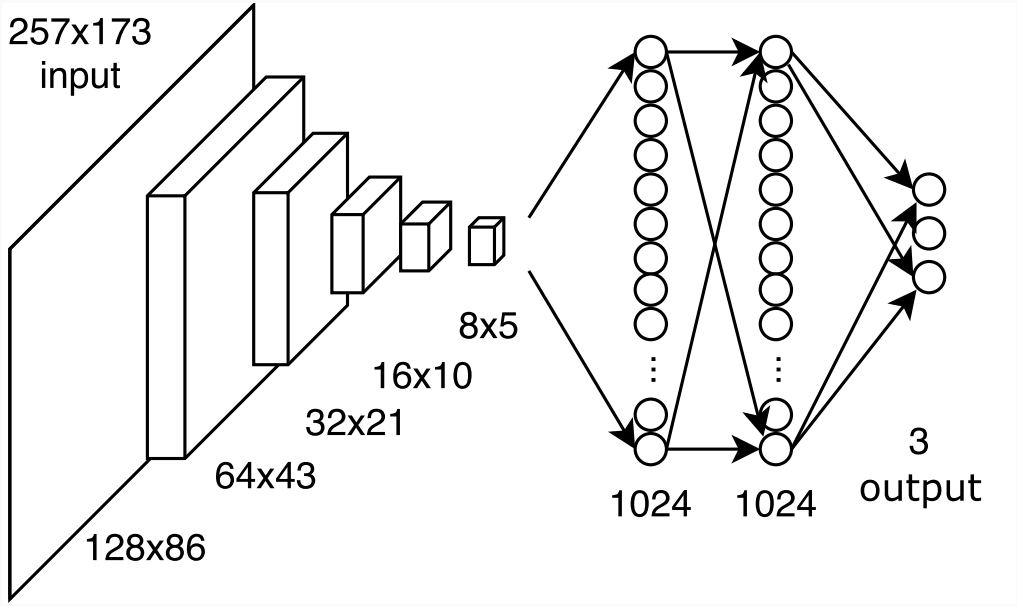
Sonagramme comme matrice 2D

Un sonagramme est un diagramme temps/fréquence représentant l'énergie de chaque point : $S : \mathbb{R}^2 \to \mathbb{R}$.

En pratique, le sonagramme discrétisé et doté d'une échelle de couleur est équivalent à une image $I : \mathbb{Z}^2 \to \mathbb{R}^3$.

Attention, contrairement à une véritable image, les axes ne sont pas interchangeables. Il peut être intéressant de choisir des noyaux rectangulaires plutôt que carrés.

CNN 2D sur le sonagramme



Explaining Deep Convolutional Neural Networks on Music Classification, Choi et al., 2016

- 5 couches convolutives 3×3 + *max-pooling* 2×2
- 2 couches entièrement connectées 1024×1024
- 1 couche entièrement connectée $1024 \times k$

Dropout appliqué sur les couches entièrement connectées

Comparaison des CNN images sur les sonagrammes

Étiquetage de vidéos YouTube à partir de l'audio uniquement :

Architectures	Steps	Time	AUC	d-prime	mAP
Fully Connected	5M	35h	0.851	1.471	0.058
AlexNet	5M	82h	0.894	1.764	0.115
VGG	5M	184h	0.911	1.909	0.161
Inception V3	5M	137h	**0.918**	**1.969**	0.181
ResNet-50	5M	119h	0.916	1.952	**0.182**
ResNet-50	17M	356h	**0.926**	**2.041**	**0.212**

Protocole

Prédiction toutes les 960 ms, étiquette d'une vidéo = moyenne des prédictions.

CNN Architectures for Large-Scale Audio Classification, Hershey et al., 2016

Résumé

CNN 1D sur le sonagramme

Approche classique, peu coûteuse, dérivée du Mel-spectogram.

CNN 1D sur la forme d'onde

2 approches :

- Remplace les MFCC par une couche convolutive,
- Convolution directement sur les échantillons (assez coûteux mais généralement le + performant).

CNN 2D sur le sonagramme

+ combine information spectrale et temporelle
+ permet de réutiliser en partie des modèles maîtrisés
– les CNN "image" ne sont probablement pas les meilleurs