RCP217 – Apprentissage profond pour les données audio

Techniques avancées

Nicolas Audebert `nicolas.audebert@lecnam.net`

1er mars 2021

Conservatoire national des arts & métiers

Augmentation de données audio

Augmentation de données

Principe

Générer de nouvelles données en transformant les sons existants. Il faut que :

- La transformation ne change pas l'information de supervision,
- La transformation corresponde à des distorsions "réelles".

Exemple dans le domaine image : une image de chat vue en miroir est encore une image de chat.

Comment augmenter les données sonores ?

Image

- Translation
- Bruit 2D
- Luminosité
- Redimensionnement

Son

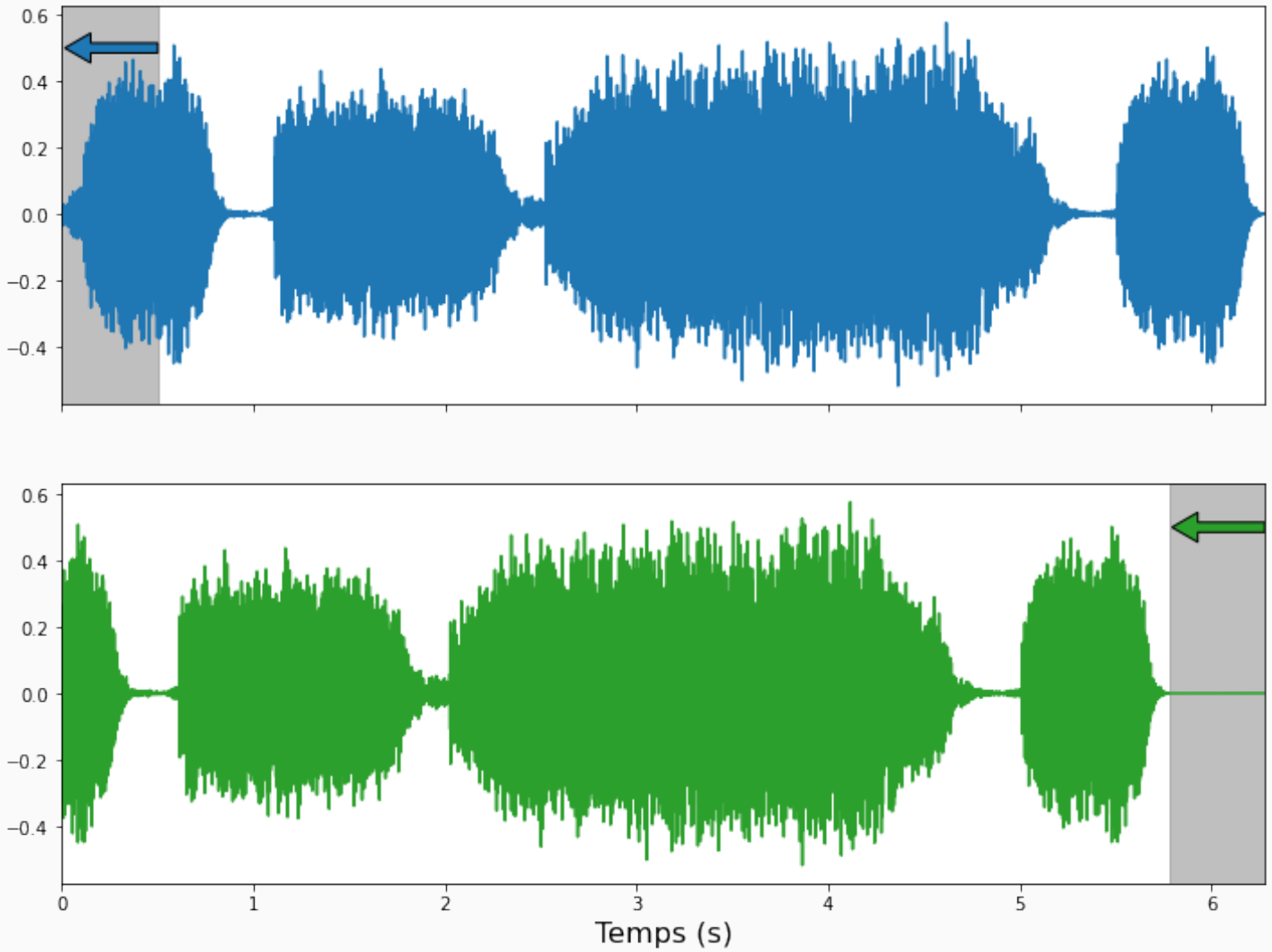
- Décalage temporel
- Bruit 1D
- Volume
- *Time stretching*

\+ transformations spécifiques au son :

- *Pitch shifting* : changer la hauteur du son
- Compresseur ($\simeq$ *contraste*)

Décalage temporel

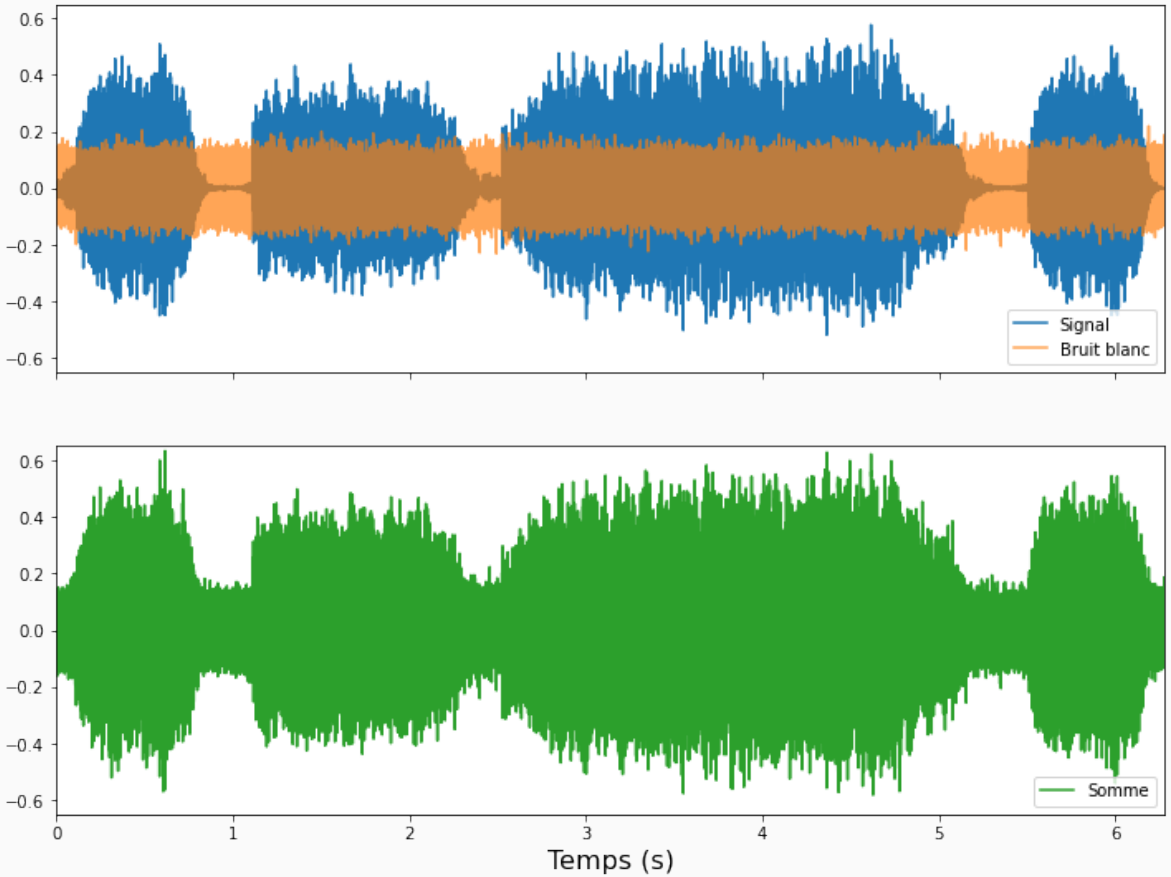
$$s[k] = s[k + \text{offset}]$$



Bruit sonore

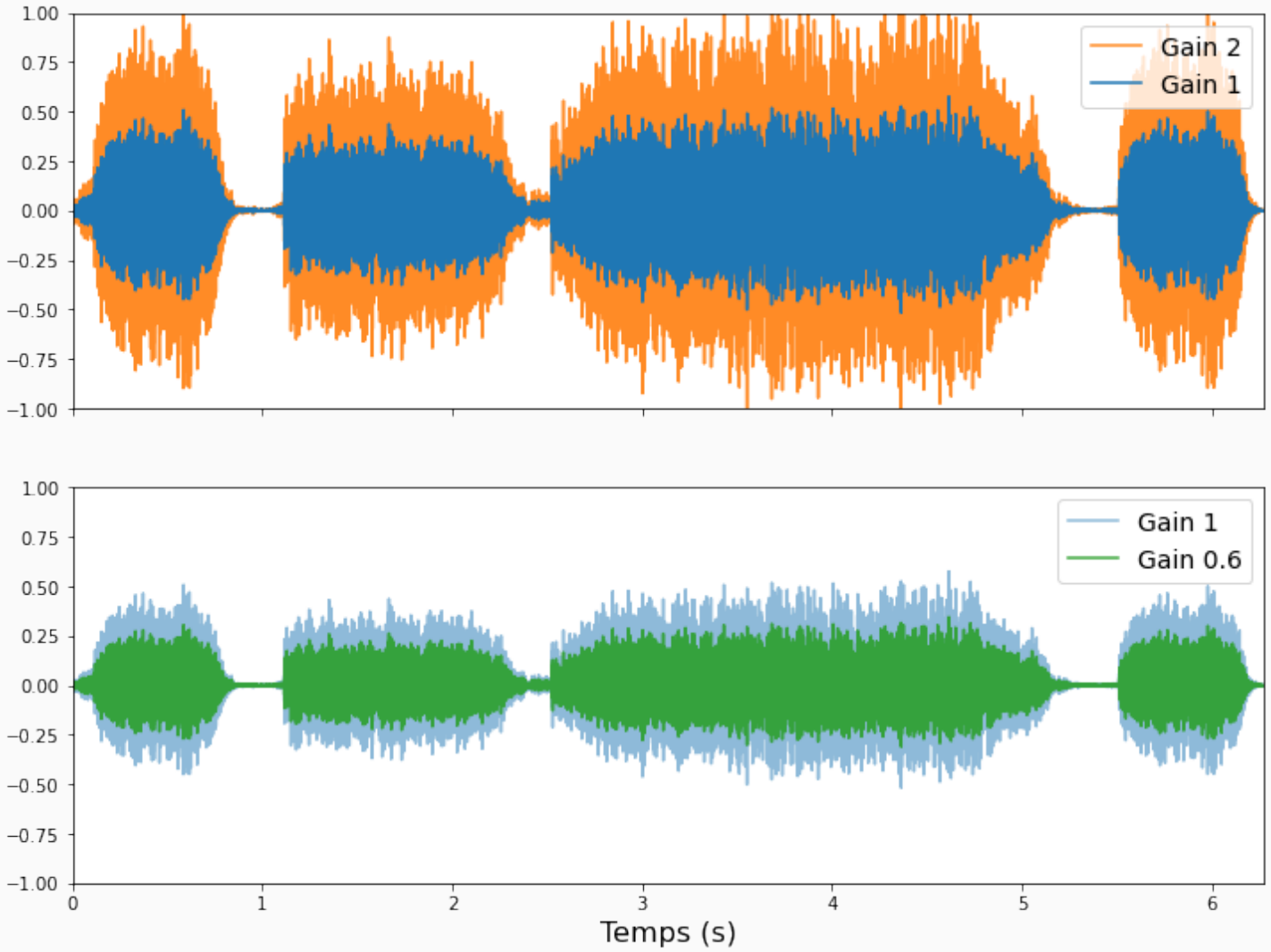
$$s[k] = s[k] + \epsilon[k]$$

où ϵ est un bruit de n'importe quel type (bruit blanc, bruit rose, voire un bruit réel pour un rendu plus plausible).



Volume

$$s[k] = \text{gain} * s[k]$$



Augmentations avancées

Étirement temporel

Modifie la durée de la séquence **sans changer la hauteur** du son.

Pitch shifting

Modifie la hauteur du son **sans changer sa durée**.

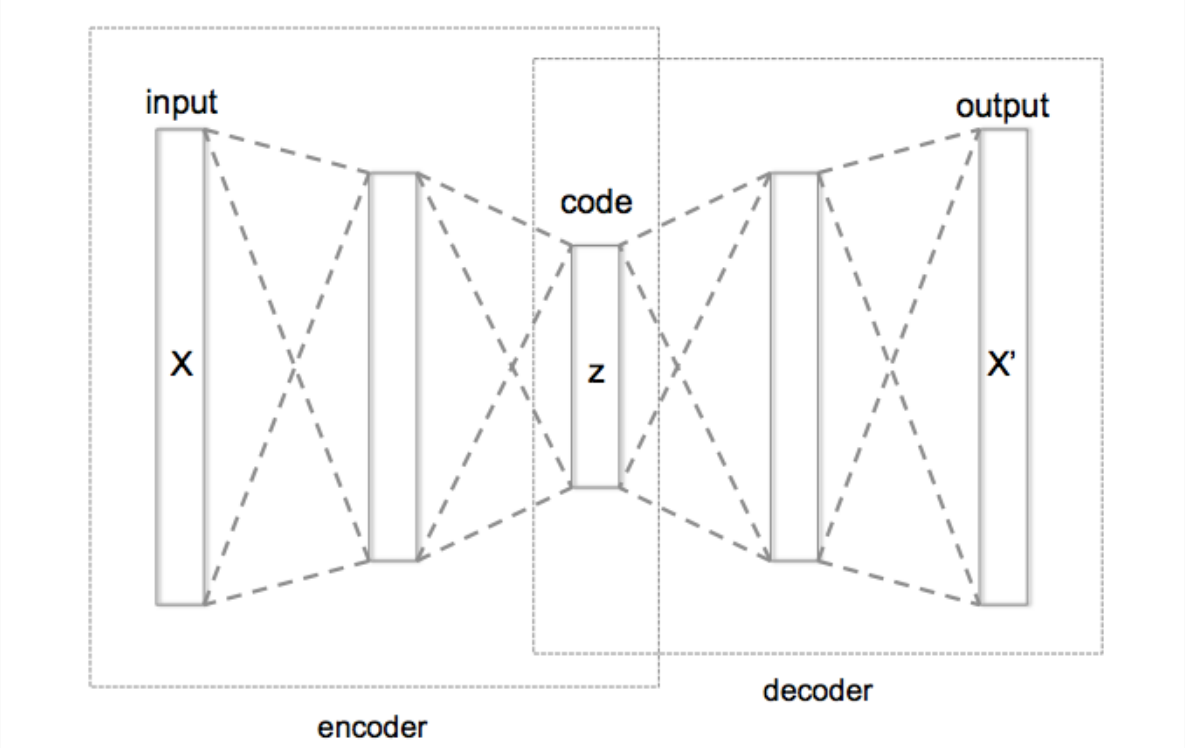
Compresseur

Modification du volume d'un gain g variable : on augmente les sons faibles et on réduit les sons forts.

$\Rightarrow$ *dynamic range compression* : réduire la dynamique du signal

Apprentissage non-supervisé sur les données audio

Autoencodeurs

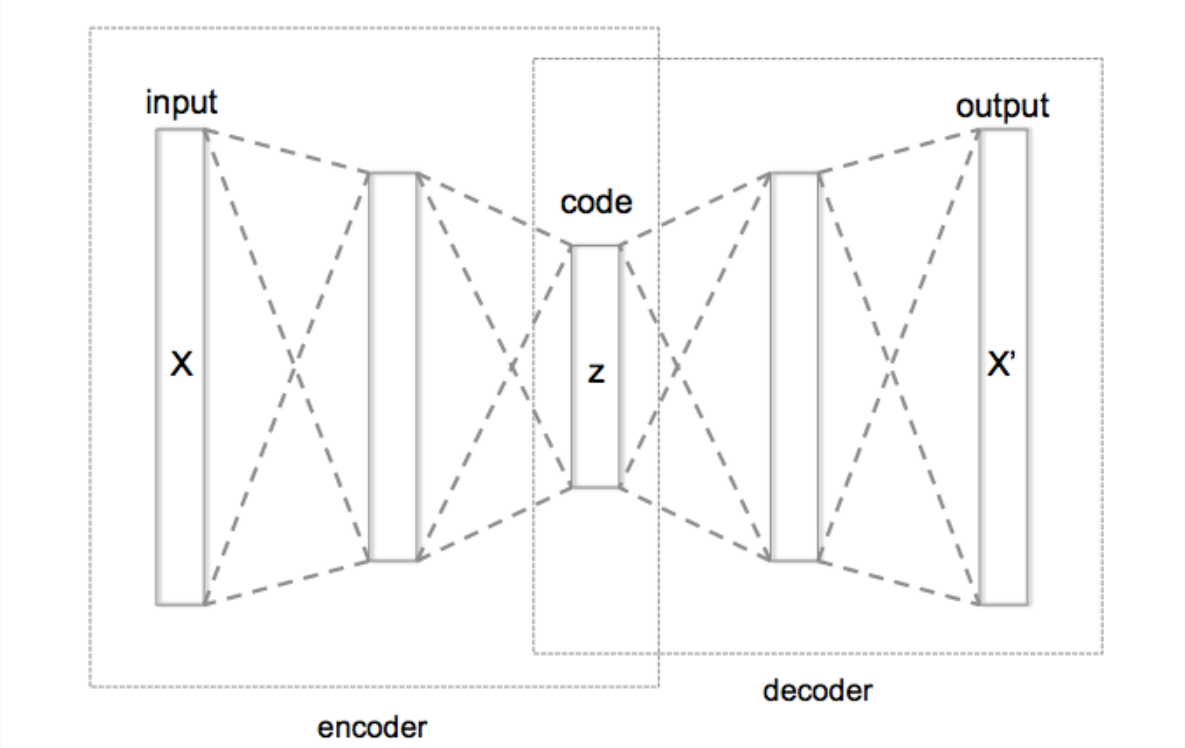


Chervinskii, Wikimédia Commons

Optimisation

Minimiser l'erreur de reconstruction entre x et x' (e.g. $||x - x'||$). $\rightarrow$ non-supervisé

Autoencodeurs



Chervinskii, Wikimédia Commons

Lien avec la théorie de l'information

Si $|code| \leq |x|$, un autoencodeur apprend à compresser les signaux d'entrée avec une perte minimale (sur le jeu d'entraînement).

Intérêt des approches non supervisées

Obtenir une **représentation vectorielle** du signal : réduction de dimension non-linéaire qui préserve l'information.

Extraction de caractéristiques

Le code peut-être utilisé comme caractéristique pour un modèle décisionnel (e.g. SVM). Les caractéristiques sont apprises sur un grand jeu de données **sans annotation**, la SVM peut être entraînée sur peu d'exemples.

Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging, Xu et al., 2016

Reconnaissance de formes/visualisation

Une fois les code obtenus, il est possible de réaliser des opérations classiques de fouille de données :

- Clustering,
- Réduction de dimension pour la visualisation.

Intérêt des approches non supervisées

Obtenir une **représentation vectorielle** du signal : réduction de dimension non-linéaire qui préserve l'information.

Extraction de caractéristiques

Le code peut-être utilisé comme caractéristique pour un modèle décisionnel (e.g. SVM). Les caractéristiques sont apprises sur un grand jeu de données **sans annotation**, la SVM peut être entraînée sur peu d'exemples.

Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging, Xu et al., 2016

Reconnaissance de formes/visualisation

Une fois les code obtenus, il est possible de réaliser des opérations classiques de fouille de données :

- Clustering,
- Réduction de dimension pour la visualisation.

Parmi ces quatre augmentations de données, laquelle vaut-il mieux éviter d'utiliser si l'on essaie d'identifier une musique à partir d'un enregistrement (type Shazam) ?

A. Décalage temporel

B. Changement de volume

C. *Pitch shift*

D. Bruit

Traitement de la parole

Spécificités

Niveau de granularité

Travail au niveau du mot ? Du phonème ? De la lettre ?

Langage

Cf. les problématiques du cours de TAL (différentes langues, synonymes, homonymes, etc.)

Temporalité

Un mot est localisé dans le temps de la séquence : il faut pouvoir “segmenter” le signal pour isoler la section pertinente.

Segmentation

Entraînement

CTC *loss* (*Connectionist Temporal Classification*, Graves et al., 2006) : recale automatiquement les prédictions du modèle pour chaque pas de temps t avec l'audio

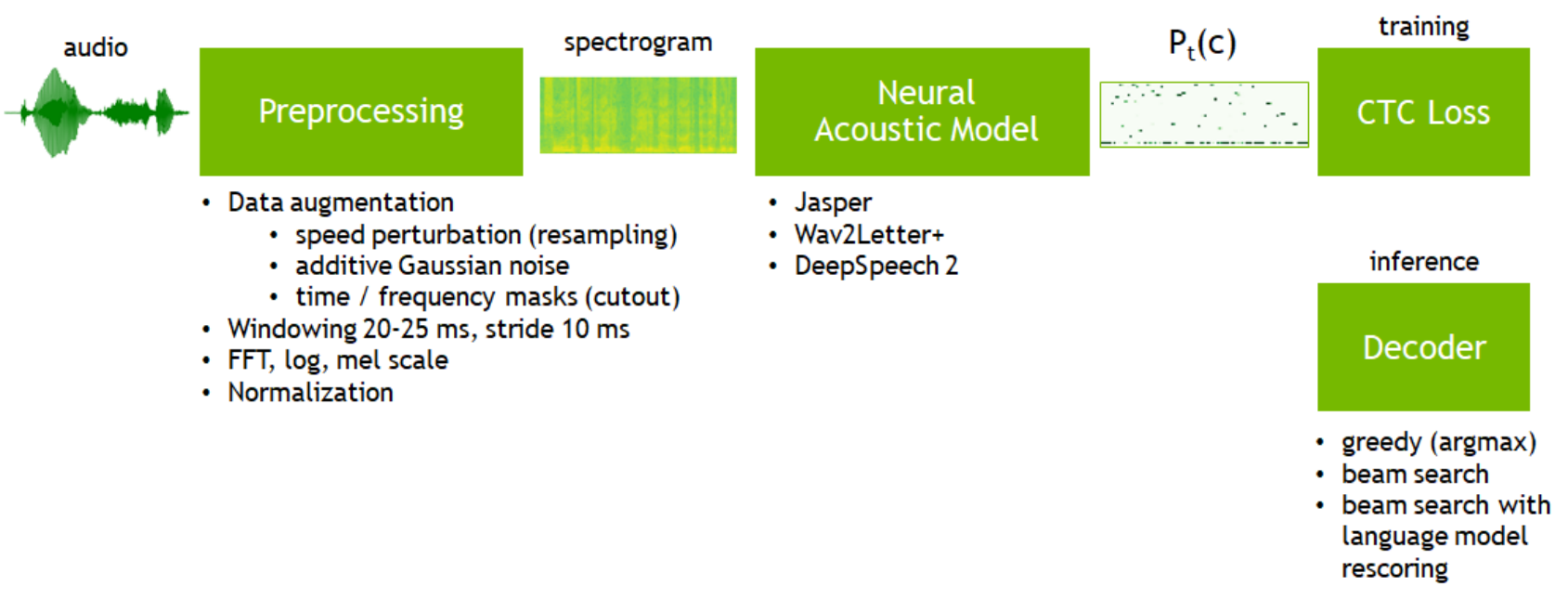
Permet d'entraîner un modèle sur de l'audio non-segmenté.

Décodage

Le modèle prédit une séquence de probabilité de présence pour chaque lettre.

- Approche *greedy* : on garde pour chaque pas de temps la lettre la plus probable.
- Ou recherche en faisceau : on examine plusieurs possibilités et on garde le n-gramme le plus probable.

Transcription



Documentation de NVIDIA OpenSeq2Seq

Cf. cours de TAL pour les modèles seq2seq/transformer

Wav2Letter

Modèle entièrement convolutif : prédit la probabilité de présence de chaque lettre pour chaque fenêtre temporelle.

Wave2Letter, 2016 et *Wave2Letter+*, 2017

Amélioration de la qualité du langage parlé

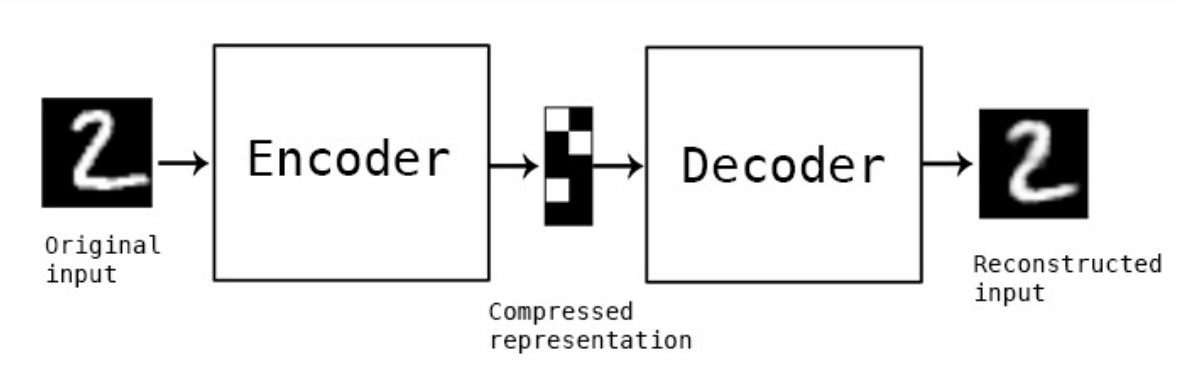
Débruitage : *denoising autoencoders*

Principe

Apprendre un autoencodeur sur des signaux bruités et optimiser la reconstruction contre le signal d'origine :

$$\|X + \epsilon, X\|$$

Les signaux bruités peuvent se générer automatiquement à l'aide d'une banque de bruits (aléatoires ou réels).



Documentation Keras sur les autoencodeurs

Audio Enhancing With DNN Autoencoder For Speaker Recognition, Plchot et al., 2016

Séparation de sources

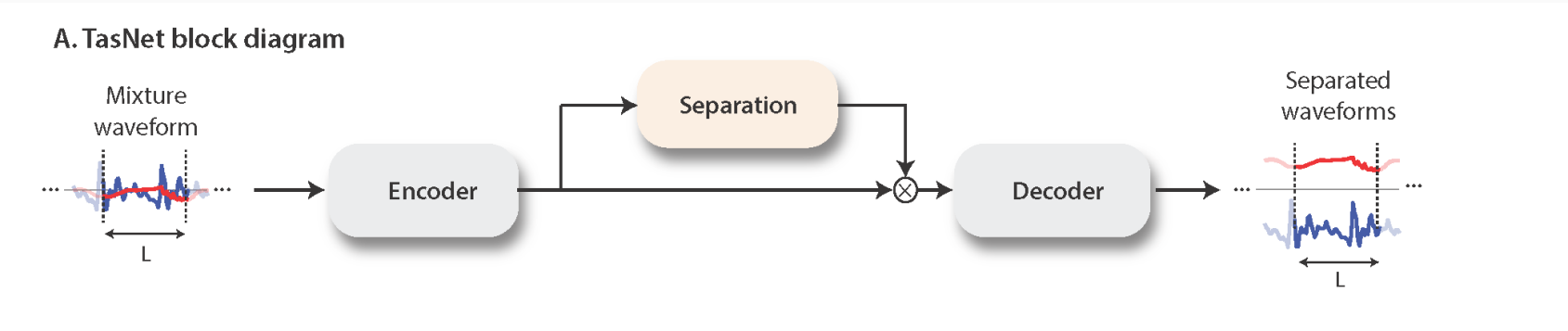
On cherche à séparer les paroles s_i prononcées par C personnes à partir de la forme d'onde x :

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

Dans le cas supervisé, on estime cherche à maximiser le *source-to-noise ratio* (SI-SNR) pour chaque s_i :

$$\begin{cases} s_{target} := \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} := \hat{s} - s_{target} \\ \text{SI-SNR} := 10\, log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \tag{1}$$

ConvTasNet



Conv-TasNet : Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, Luo et Mesgarani, 2018

Encodeur : $E : x \rightarrow \mathbf{w} \in \mathbb{R}^n$ Décodeur : $D : \mathbf{w} \rightarrow \hat{x}$

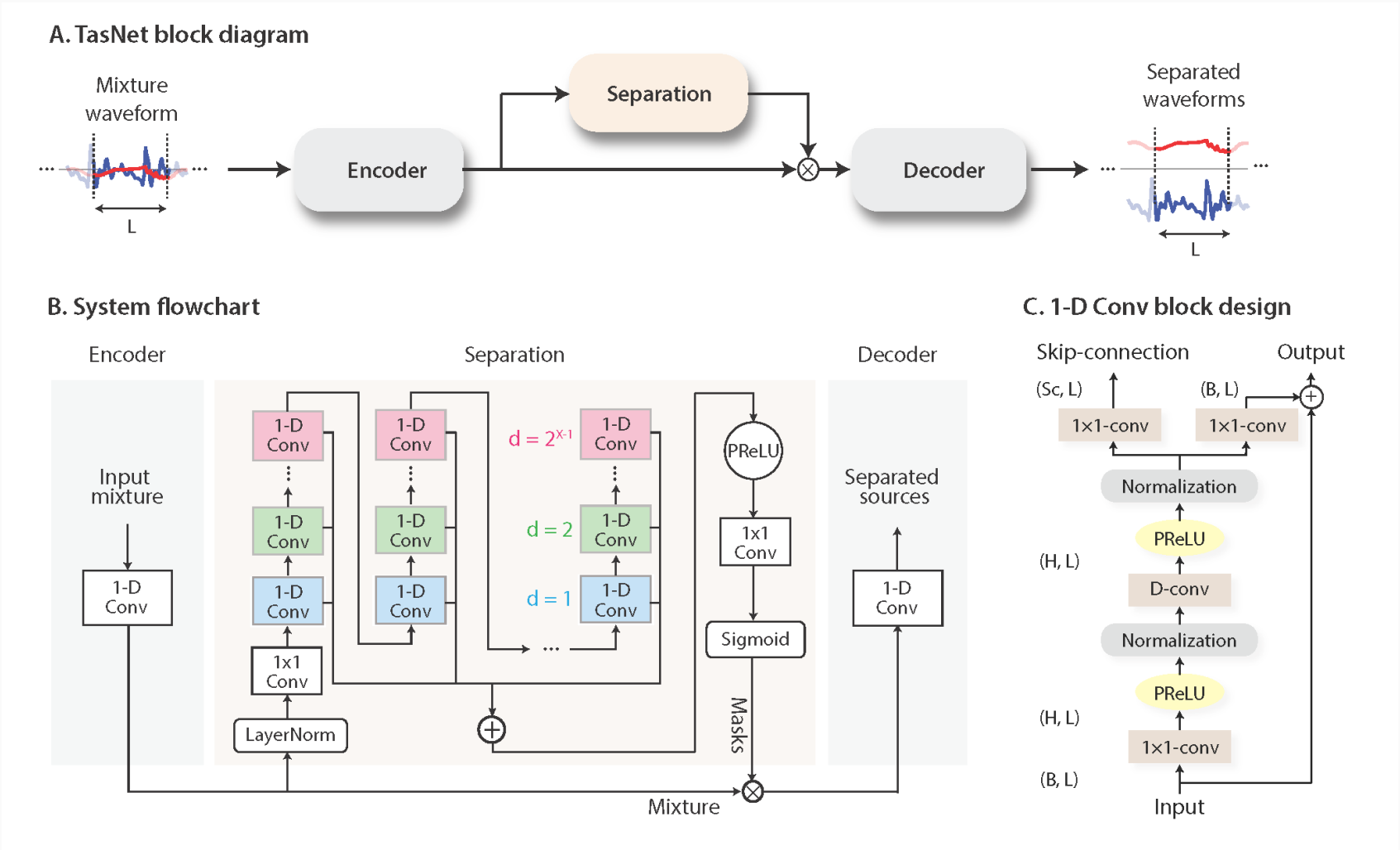
Le séparateur produit C masques $\mathbf{m}_i$ qui correspondent aux éléments du vecteur $\mathbf{w}$ correspondant à chaque source.

Les sources reconstruites sont obtenues par :

$$\hat{s}_i = D\left(\mathbf{w} \odot \mathbf{m}_i\right)$$

où $\odot$ est la multiplication matricielle élément par élément.

ConvTasNet



Conv-TasNet : Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, Luo et Mesgarani, 2018

Synthèse vocale (1/2)

La probabilité jointe d'une forme d'onde $\mathbf{x} = \{x_1, \ldots, x_T\}$ est le produit des probabilités conditionnelles :

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

Principe des modèles de synthèse vocale

Produire un x_t plausible à partir des $\{x_1, \ldots, x_{t-1}\}$.

Modèles génératifs

Chaînes de Markov, réseaux de neurones récurrents, ...

Le modèle cherche à minimiser la divergence de Kullback-Leibler entre le vecteur de probabilité prédit pour x_t et la réalité.

Synthèse vocale (2/2)

Pour générer la parole correspondant à un texte $\mathbf{h}$, on va conditionner $\mathbf{x}$ à $\mathbf{h}$:

$$p\left(\mathbf{x} \mid \mathbf{h}\right) = \prod_{t=1}^{T} p\left(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}\right)$$

Si $\mathbf{h}$ varie en fonction de t, on parle de conditionnement local.

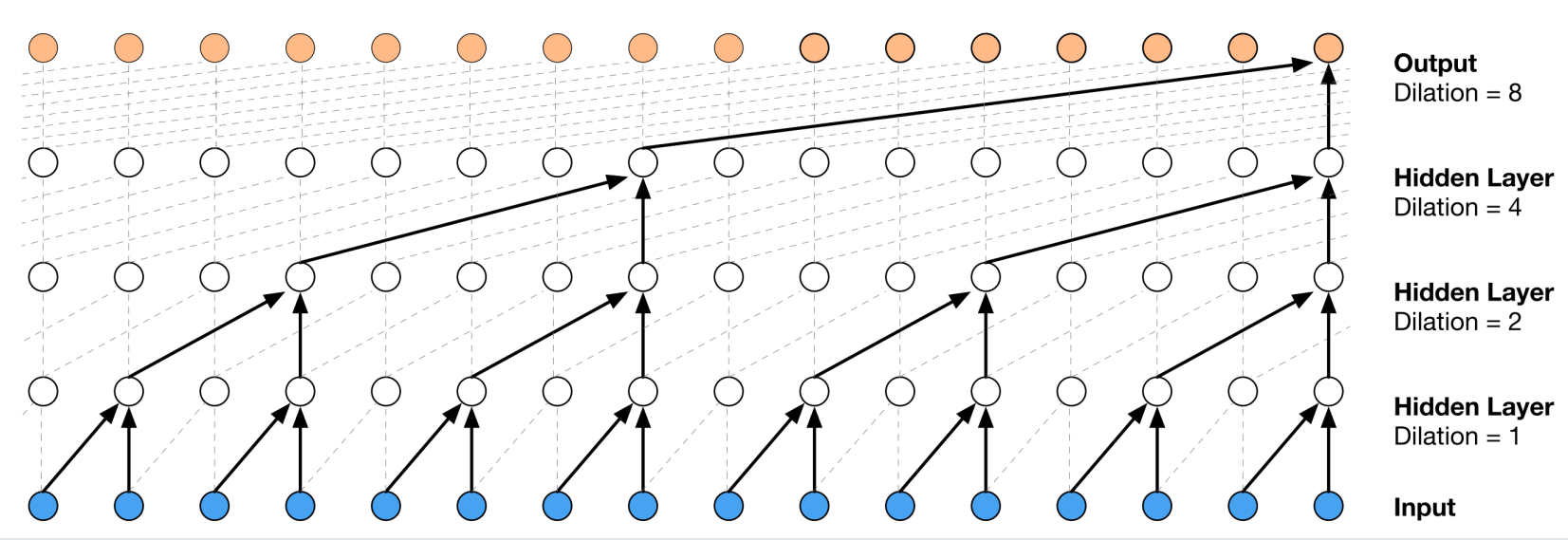
Dans le cas de la synthèse vocal, $\mathbf{h}$ représente les caractéristiques linguistiques (prononciation, accentuation, etc.).

WaveNet

WaveNet est une approche efficace basée sur les CNN pour la synthèse vocale sur la forme d'onde.

Principe

Utiliser des convolutions *à trous* (ou dilatées) **causales**, c'est-à-dire qui ne s'appliquent que sur les échantillons qui **précèdent** le pas de temps t considéré.



WaveNet : a generative model for raw audio, Oord et al., 2016

Quiz

J'entraîne un modèle de séparations de sources pour distinguer entre 2 et 6 voix. J'ai besoin de cinq modèles différents (pour $C = 2$, $C = 3$, $C = 4$, $C = 5$ **et** $C = 6$).

- Vrai
- Faux

Apprentissage multimodal son-image

Multimodalité

Définition

Mesure d'une même observation par différents modes d'acquisition (e.g. capteurs).

Exemples :

- Vidéo (microphone + caméra),
- Stéréoscopie (acquisition par deux caméras gauche/droite),
- Tweets (texte + métadonnées)
- ...

On parle d'apprentissage **multimodal** ou d'apprentissage **multi-vues**.

Intérêt

L'information sonore peut corriger des erreurs ou des ambiguités de l'information visuelle et inversement.

Spécificités de la multimodalité son-image

+ les modalités sont synchronisées temporellement,
- les modalités sont fortement hétérogènes,
- l'échantillonnage n'est pas identique (24 ips/44 kHz)

Architectures pour la fusion de données

3 paradigmes : *late, middle* et *early fusion*

L'idée est de choisir à quel endroit "mixer" les représentations des deux modalités.

En général, la combinaison se fait :

- par concaténation : il suffit que les premières dimensions des deux tenseurs soient identiques $(N_1, N_2, \ldots, H)$ et $(N_1, N_2, \ldots, H')$
- par addition : il faut que toutes les dimensions soient identiques.

Architectures pour la fusion de données

3 paradigmes : *late, middle* et *early fusion*

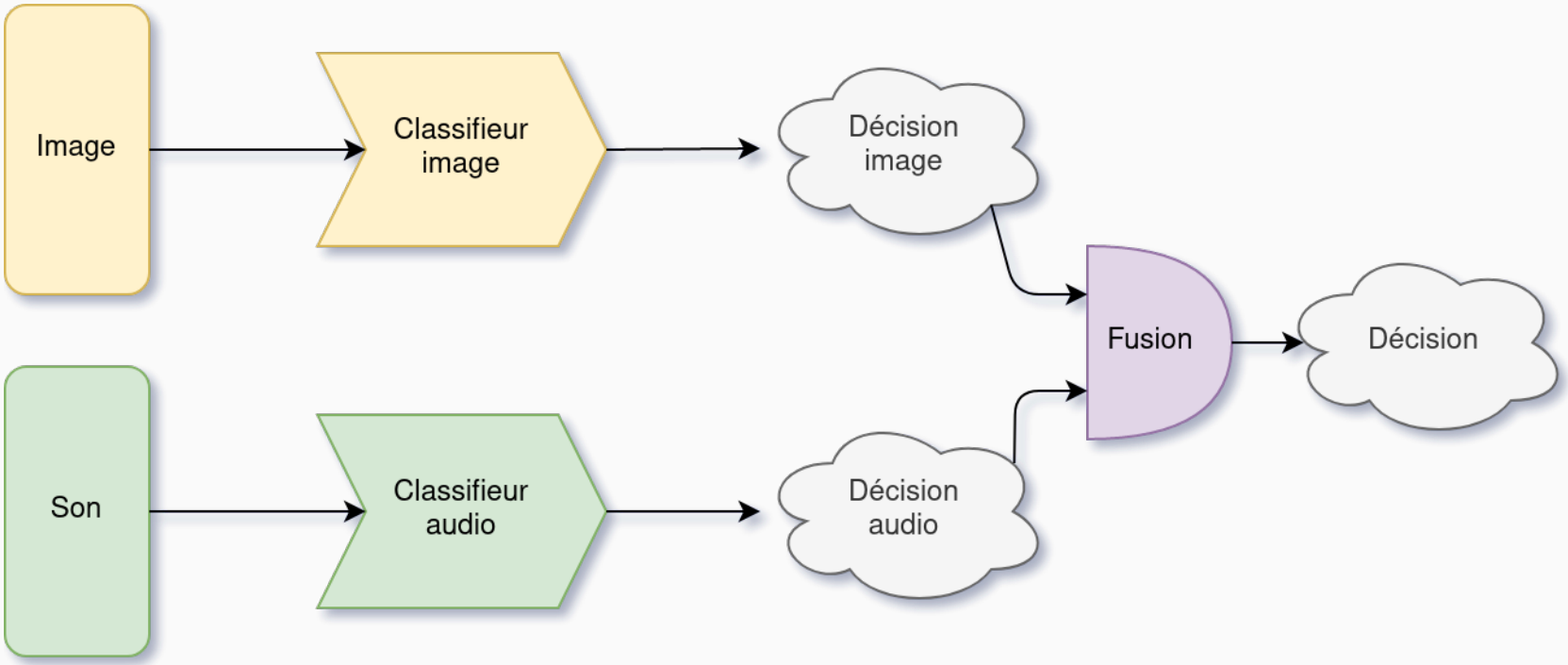
L'idée est de choisir à quel endroit "mixer" les représentations des deux modalités.

En général, la combinaison se fait :

- par concaténation : il suffit que les premières dimensions des deux tenseurs soient identiques $(N_1, N_2, \ldots, H)$ et $(N_1, N_2, \ldots, H')$
- par addition : il faut que toutes les dimensions soient identiques.

Approche historique

Fusion tardive : la combinaison des modalités se fait au niveau **décisionnel** $\Rightarrow$ pas de co-apprentissage

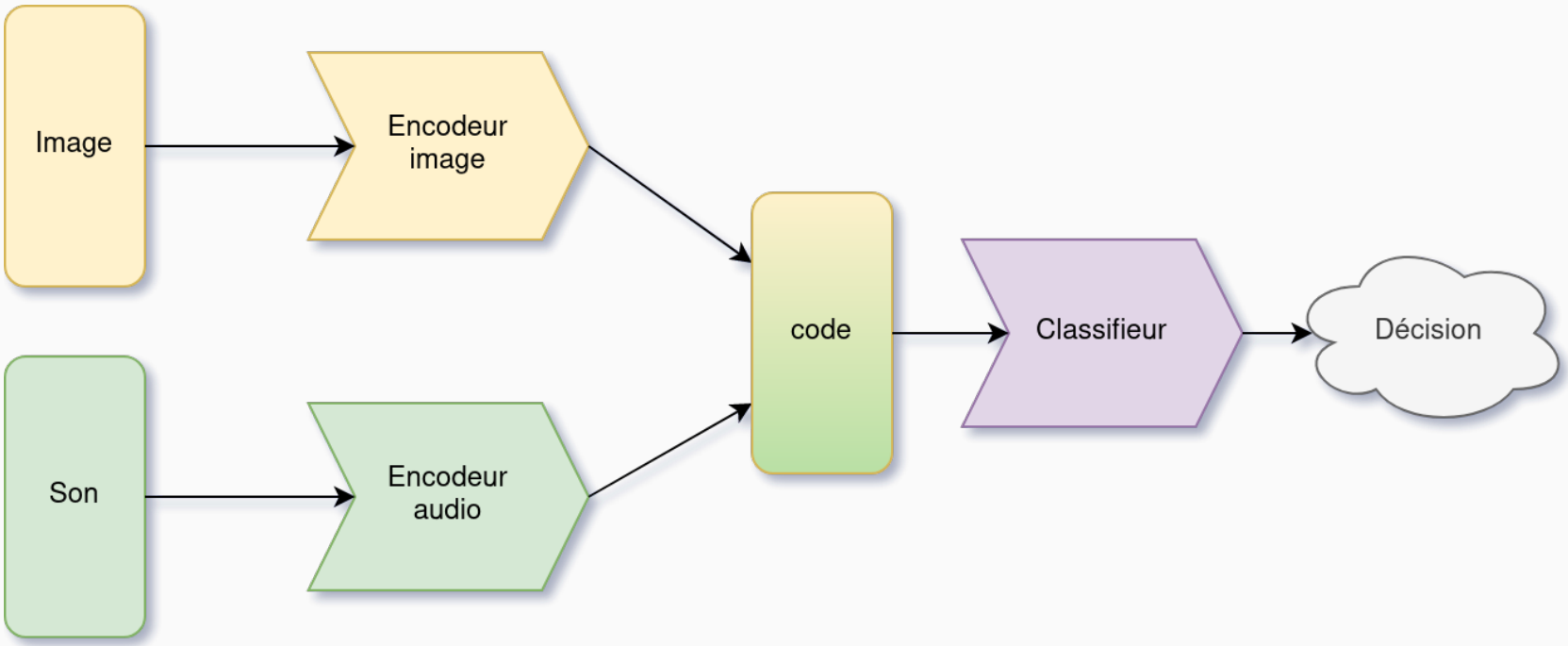


Exemples : vote majoritaire, moyenne...

A hybrid approach to news video classification multimodal features, Wang et al., 2003

Combinaison de représentations

Fusion intermédiaire : combiner des représentations internes

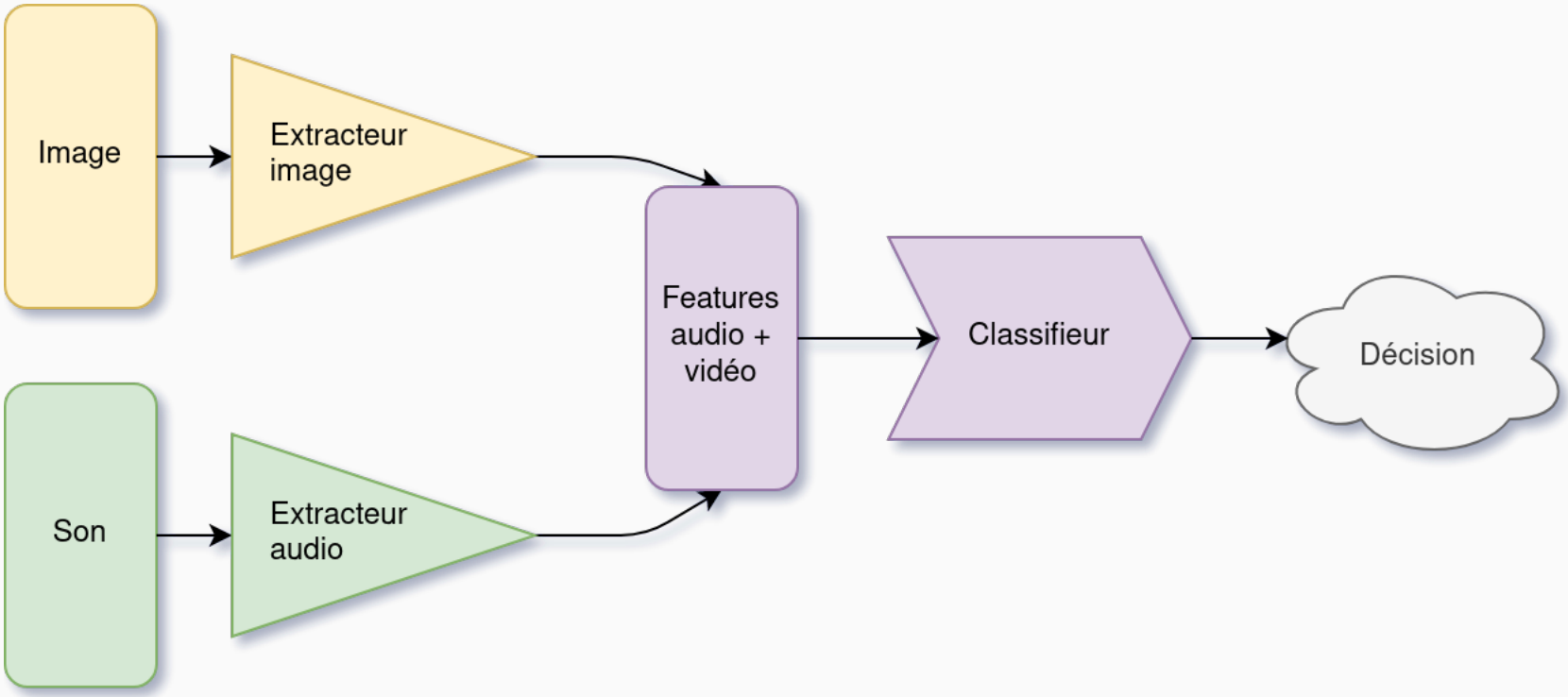


Exemples : concaténer ou additionner les vecteurs audio/vidéo.

Multimodal deep representation learning for video classification, Tian et al., 2019

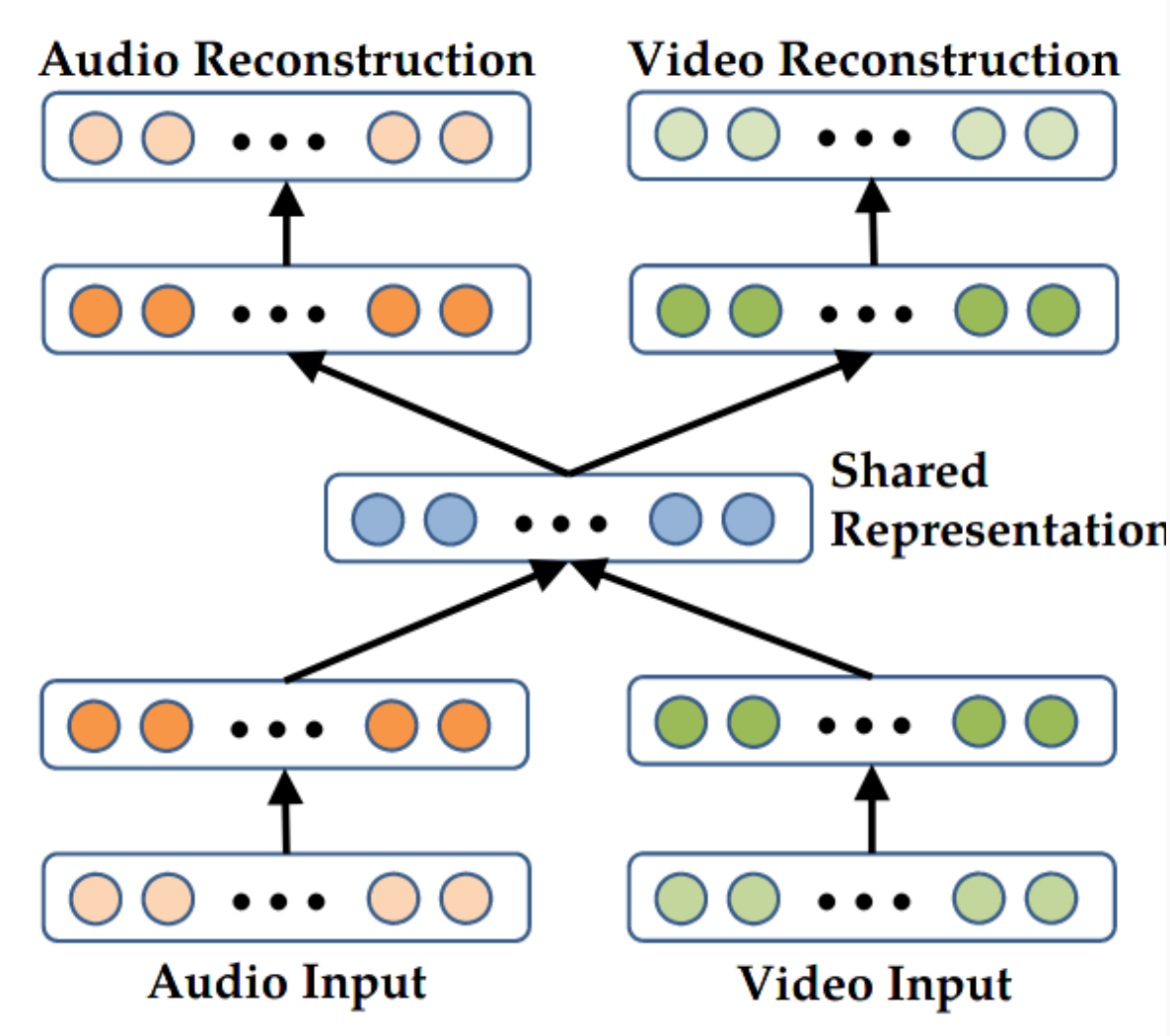
Représentation partagée

Fusion amont : une représentation unique pour les deux modalités.



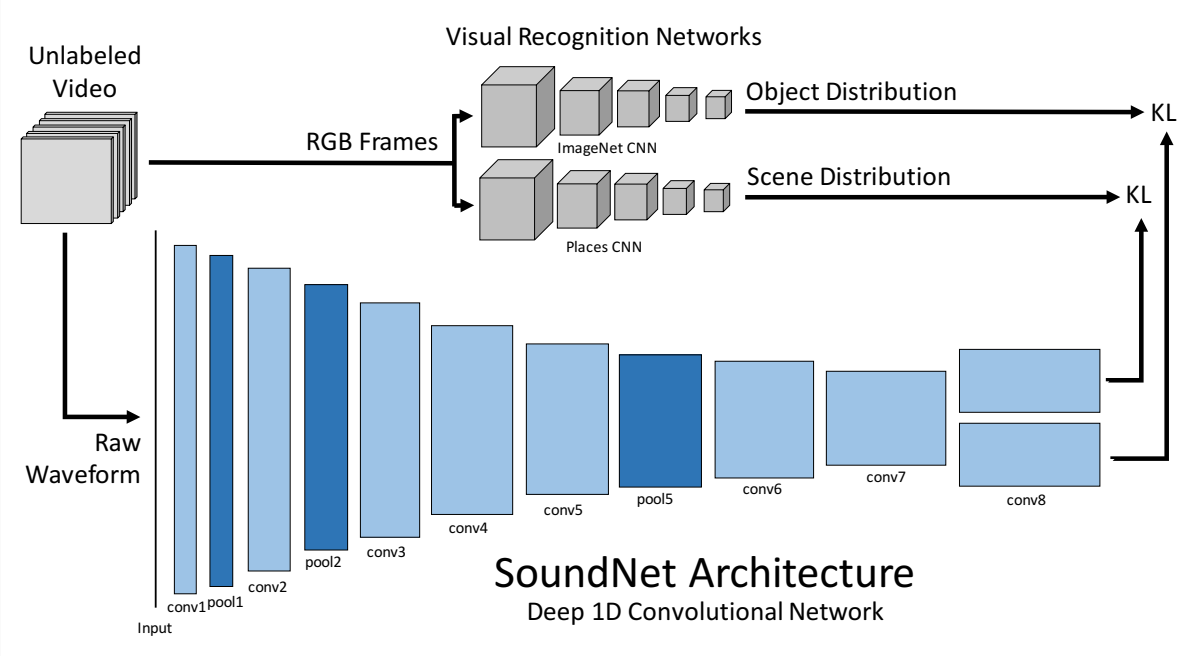
Exemple : coefficients MFCC agrégés toutes les 41 ms + *features* images concaténées.

Exemple : l'autoencodeur bimodal pour la vidéo



Multimodal Deep Learning, Ngiam et al. 2011

SoundNet : transfert image → son



`https://projects.csail.mit.edu/soundnet/`

Principe

À partir d'un jeu de données de vidéos non étiquetées :

- Extraire des informations (objets/lieux) pour chaque image par un modèle visuel préentraîné.
- Le modèle audio reproduit les prédictions du modèle visuel.

Résumé

Fusion de données

Combiner des informations venant de sources différentes mais qui décrivent la même observation.

3 approches

Fusion tardive ($\simeq$ approche ensembliste), fusion intermédiaire (apprentissage d'une représentation mixte obtenue par somme/concaténation), fusion amont (travail directement sur des caractéristiques extraites de façon multimodales).

J'ai une SVM sur les coefficients MFCC appris sur l'audio d'un jeu de données vidéo et un CNN appris sur l'image. Quelle proposition correspond à une fusion tardive ?

A. Je concatène les *features* du CNN avec les coefficients MFCC et j'entraîne un perceptron multi-couche sur levecteur qui en résulte.

B. J'additionne les probabilités de sortie de la SVM à celles du softmax du CNN et je prends le max.