

IFT 4030/7030,
Machine Learning for Signal Processing
Week12: Speech/Audio

Cem Subakan



UNIVERSITÉ
Laval



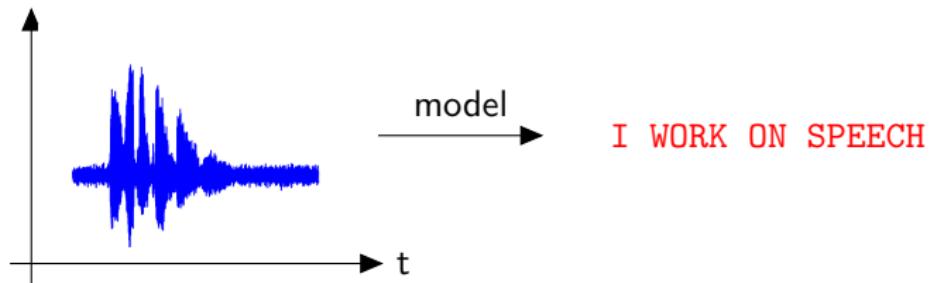
- The last lab is this week!
 - ▶ Le dernier lab est ce vendredi!
- The last homework is out. It will be done with your team.
 - ▶ Le dernier devoir va se faire en équipe.
- Don't forget to sign-up for a project presentation slot!
https://docs.google.com/spreadsheets/d/1jWw0zGoaixzYXJwDCsK0LaF_z8J9-jY1be-d9WzG9GY/edit?usp=sharing
 - ▶ N'oubliez pas d'inscrire votre projet sur le fichier excel.
- I will assign a peer-review schedule. Don't ignore it!
 - ▶ Je vais assigner un peer-review schedule. Faites attention!
- This week it's on speech and audio!
 - ▶ Cette semaine, on a le dernier cours, et c'est sur speech et audio!

Today

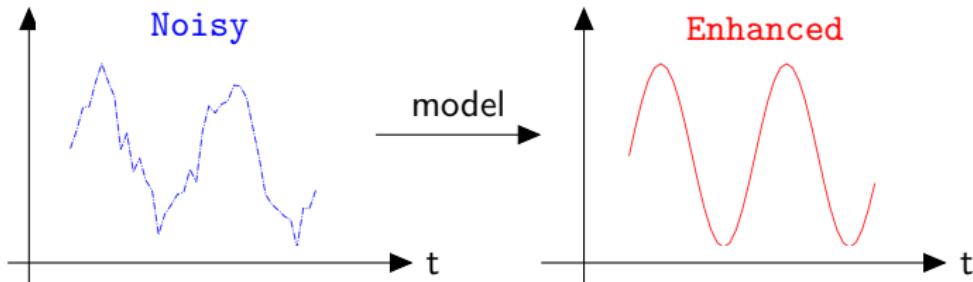
- ASR, TTS, Speech Separation / Enhancement, Text-Audio Representations, Interpretability
- The first part will be more like our usual classes. / La première partie va être comme business as usual.
- It will get more like a research talk afterwards. / Je vais parler plus comme si c'était une présentation de recherche dans la deuxième partie.

Typical Applications in Speech and Audio Modeling

■ Speech Recognition

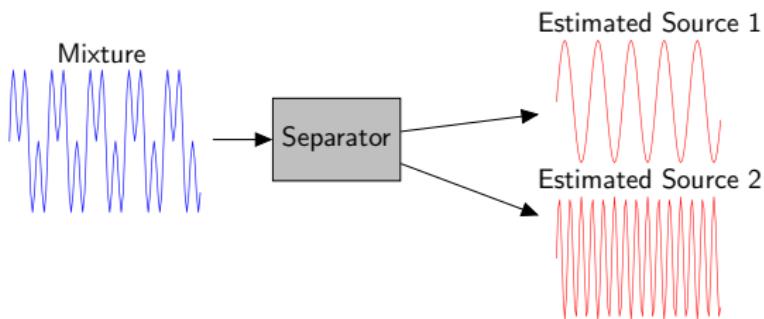


■ Speech Enhancement

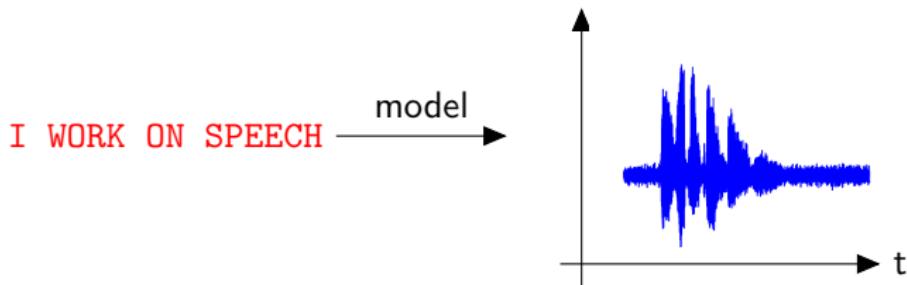


Speech and Audio Modeling

■ Speech Separation

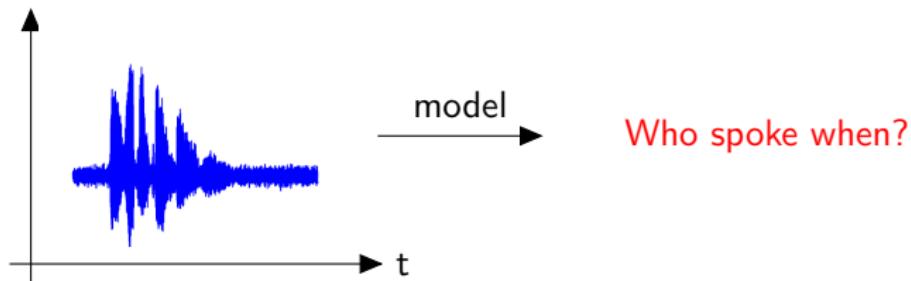


■ Text-to-Speech

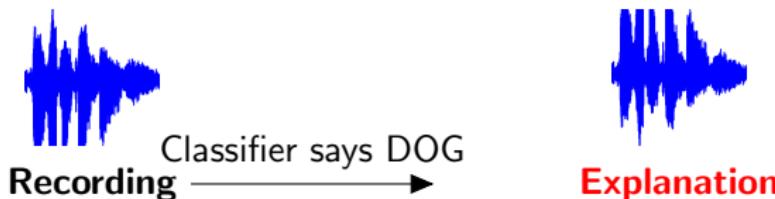


Speech and Audio Modeling

- Speaker Diarization



- Neural Network Explanation



- Other problems: Generating Deep fakes, Detecting deep fakes, Music Source Separation, Music Transcription, Sound Event Detection/Classification...

Table of Contents

Speech Recognition

RNN Based ASR

Transformer ASR

CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

Problem Definition

Source Separation Methods

SepFormer

Moving towards real-life source separation

Data collection

Performance estimation

Evaluating the SI-SNR Estimator

User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

Posthoc Explanations for Audio Classifiers

A little bonus

Automatic Speech Recognition

- In Automatic Speech Recognition (ASR) the goal is to convert a an audio signal into text.
 - ▶ Dans la tache reconnaissance vocale le but est de convertir un signal audio au texte.
- We already implemented a very simple ASR system in this class by the way!
 - ▶ On a déjà fait un exercice dans le lab1 pour un système d'ASR simple.

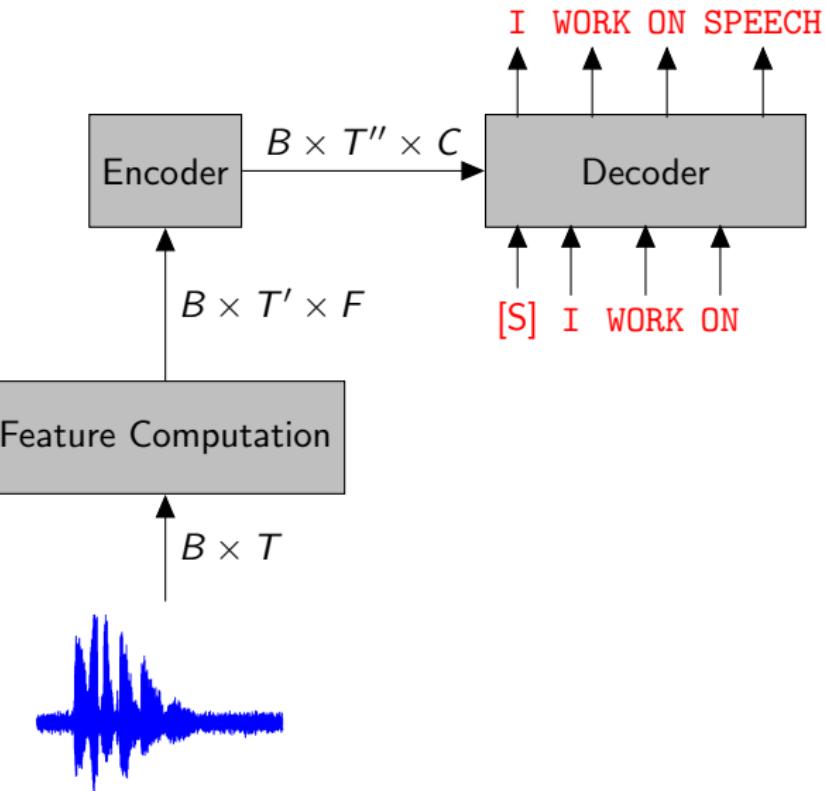
Automatic Speech Recognition

- In Automatic Speech Recognition (ASR) the goal is to convert a an audio signal into text.
 - ▶ Dans la tache reconnaissance vocale le but est de convertir un signal audio au texte.
- We already implemented a very simple ASR system in this class by the way!
 - ▶ On a déjà fait un exercice dans le lab1 pour un système d'ASR simple.
- We also talked about HMMs which used to be the SOTA for speech recognition in the 90s - 2000s.
 - ▶ On a aussi parlé des HMMs qui était l'approche SOTA pour la reconnaissance vocale dans les 90s - 2000s.

Automatic Speech Recognition

- In Automatic Speech Recognition (ASR) the goal is to convert a an audio signal into text.
 - ▶ Dans la tache reconnaissance vocale le but est de convertir un signal audio au texte.
- We already implemented a very simple ASR system in this class by the way!
 - ▶ On a déjà fait un exercice dans le lab1 pour un système d'ASR simple.
- We also talked about HMMs which used to be the SOTA for speech recognition in the 90s - 2000s.
 - ▶ On a aussi parlé des HMMs qui était l'approche SOTA pour la reconnaissance vocale dans les 90s - 2000s.
- Today we will talk about more modern approaches.
 - ▶ On va parler des approches modernes.

Encoder-Decoder Sequence-to-Sequence Learning



Feature Computation Block

- We turn an input waveform $x \in \mathbb{R}^T$, into a time-frequency representation $x' \in \mathbb{R}^{T' \times F}$.
 - ▶ On transforme un waveform x à une représentation x' temps-fréquence.

Feature Computation Block

- We turn an input waveform $x \in \mathbb{R}^T$, into a time-frequency representation $x' \in \mathbb{R}^{T' \times F}$.
 - On transforme un waveform x à une représentation x' temps-fréquence.
- ASR systems typically use mel-transformed spectra
 - Les systèmes d'ASR typiquement utilisent des spectrogrammes en échelle-mel

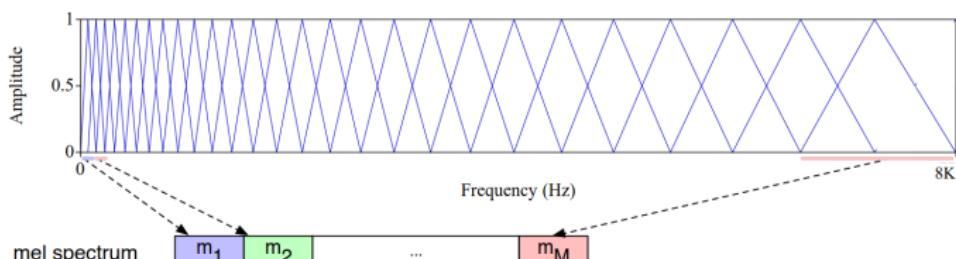


Image taken from Speech and Language Processing, Jurafsky, Martin

Feature Computation Block

- We turn an input waveform $x \in \mathbb{R}^T$, into a time-frequency representation $x' \in \mathbb{R}^{T' \times F}$.
 - On transforme un waveform x à une représentation x' temps-fréquence.
- ASR systems typically use mel-transformed spectra
 - Les systèmes d'ASR typiquement utilisent des spectrogrammes en échelle-mel

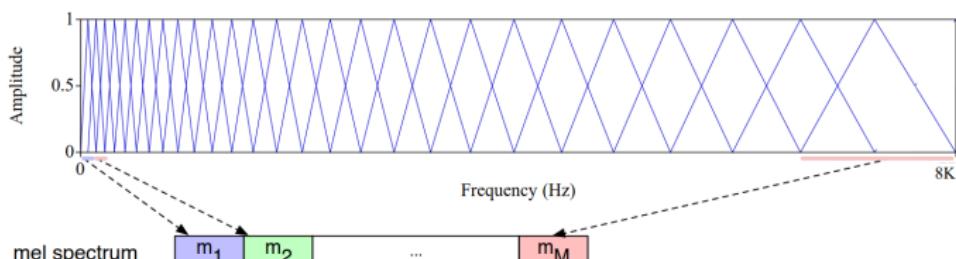


Image taken from Speech and Language Processing, Jurafsky, Martin

- You remember this? (slide 31)

Table of Contents

Speech Recognition

RNN Based ASR

Transformer ASR

CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

Problem Definition

Source Separation Methods

SepFormer

Moving towards real-life source separation

Data collection

Performance estimation

Evaluating the SI-SNR Estimator

User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

Posthoc Explanations for Audio Classifiers

A little bonus

The Encoder Block

- What we typically do in the encoder block is to reduce the sampling rate, while applying convolutions / RNNs.
 - ▶ On typiquement réduit le taux d'échantillon avec des convolutions / RNNs.

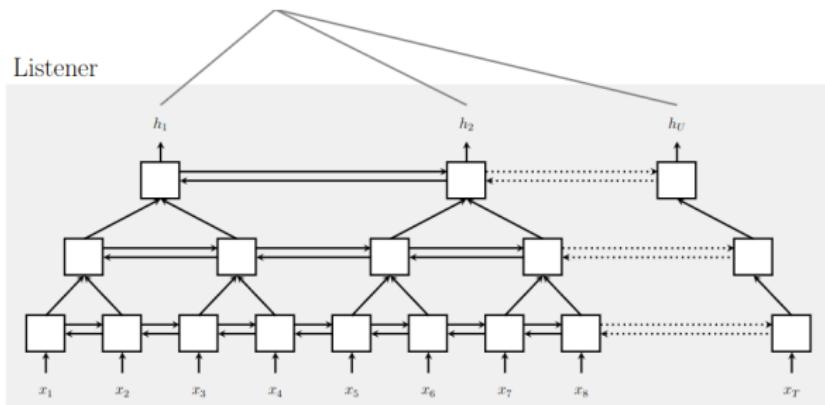


Image taken from the paper, Listen, Attend, Spell <https://arxiv.org/pdf/1508.01211.pdf>

- We can denote $h =: \text{Encoder}(x)$.

The Decoder Block

- The goal is to maximize the following probability distribution / Le but est de maximiser la distribution suivante,

$$\max \sum_{t=1}^T \log p(y_t | y_{t-1}, \dots, y_1, x)$$

The Decoder Block

- The goal is to maximize the following probability distribution / Le but est de maximiser la distribution suivante,

$$\max \sum_{t=1}^T \log p(y_t | y_{t-1}, \dots, y_1, x)$$

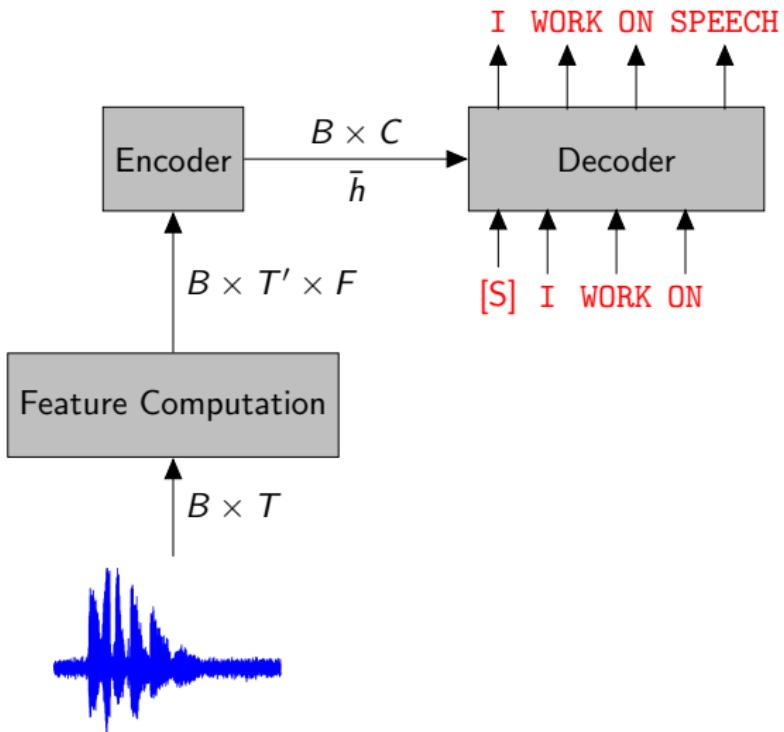
- Here's a naive way of doing it / Une facon naive pour le faire:

$$\bar{h} = \frac{1}{T} \sum_t h_t$$

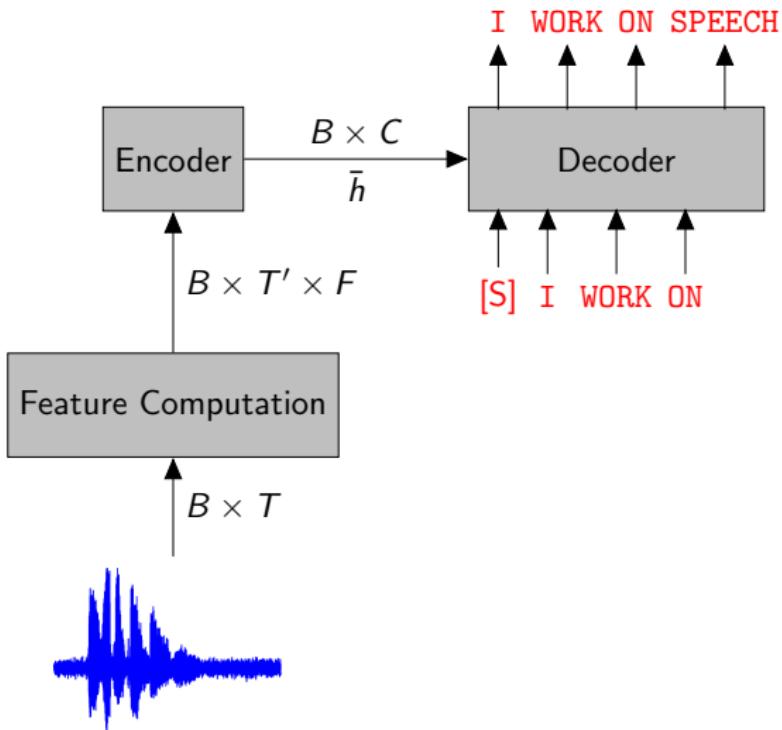
$$\hat{y}_t = \text{RNN}(y_{1:t-1}, \bar{h}) = \text{RNN}(y_{t-1}, s_{t-1}, \bar{h})$$

We inject the average representation \bar{h} to the autoregressive RNN model. / On injecte une representation average au modele autoregressif.

Naive encoder-decoder



Naive encoder-decoder



This system however shrinks the input signal resolution too much. / On perd trop de résolution temporelle avec ce système.

Encoder-Decoder with Attention in between

- The additional attention block / le bloc d'attention:

$$c_i = \text{AttentionContext}(s_i, h_{1:T''})$$

$$s_i = \text{RNN}(s_{i-1}, y_{i-1}, c_{i-1})$$

$$p(y_i|x, y_{1:i-1}) = \text{OutputDistribution}(s_i, c_i)$$

- Here's how it works / Voici comment ça fonctionne:

$$e_{i,u} = \langle \phi(s_i), \psi(h_u) \rangle$$

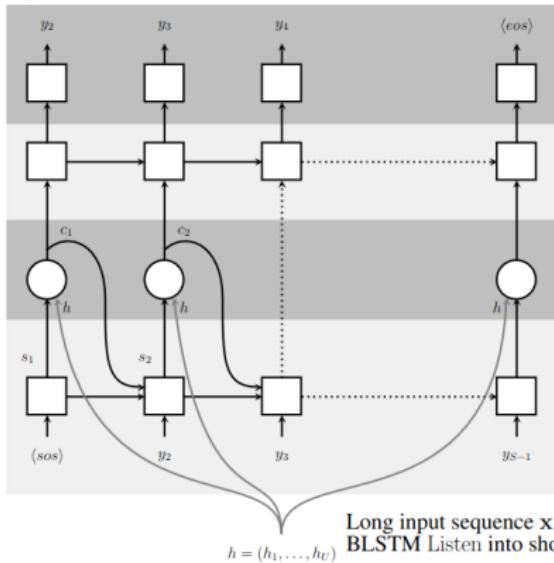
$$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_{u'} \exp(e_{i,u'})}$$

$$c_i = \sum_u \alpha_{i,u} h_u$$

- $\phi(\cdot), \psi(\cdot)$ are MLPs.

Listen, Attend, Spell Decoder

Speller



Grapheme characters y_i are modelled by the CharacterDistribution

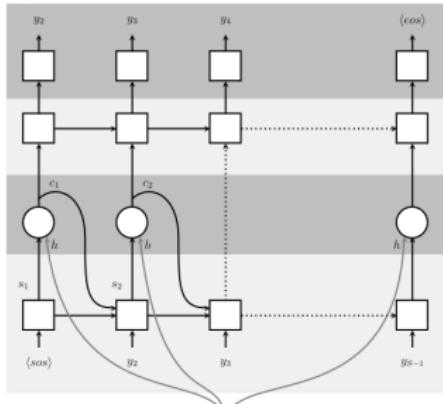
AttentionContext creates context vector c_i from h and s_i

Long input sequence x is encoded with the pyramidal BLSTM Listen into shorter sequence h

$$h = (h_1, \dots, h_V)$$

Listen, Attend, Spell All picture

Speller



Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from \mathbf{h} and s_i

Long input sequence x is encoded with the pyramidal BLSTM Listen into shorter sequence \mathbf{h}
 $h = (h_1, \dots, h_V)$

Listener

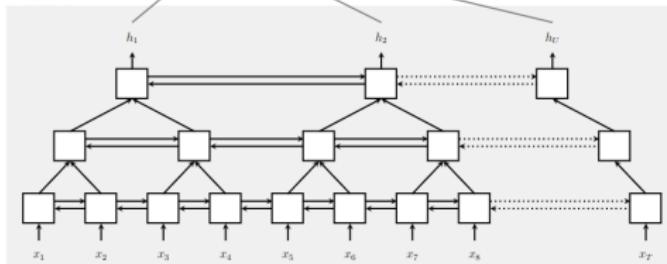


Table of Contents

Speech Recognition

RNN Based ASR

Transformer ASR

CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

Problem Definition

Source Separation Methods

SepFormer

Moving towards real-life source separation

Data collection

Performance estimation

Evaluating the SI-SNR Estimator

User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

Posthoc Explanations for Audio Classifiers

A little bonus

Query-Key-Value Attention

- Attention is the basis of the transformer architecture which obtains state-of-the art results in several domains such as NLP, computer vision, speech recognition. / Attention est base de l'architecture transformer qui obtient SOTA dans plusieurs domaines.

$$\text{Attention}(X_1, X_2, X_3) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

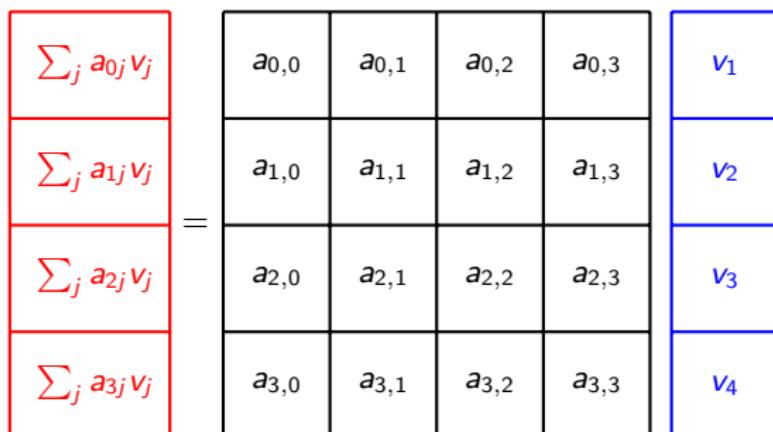
$$Q = X_1 W^Q, K = X_2 W^K, V = X_3 W^V$$

Query-Key-Value Attention

- Attention is the basis of the transformer architecture which obtains state-of-the art results in several domains such as NLP, computer vision, speech recognition. / Attention est base de l'architecture transformer qui obtient SOTA dans plusieurs domaines.

$$\text{Attention}(X_1, X_2, X_3) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

$$Q = X_1 W^Q, K = X_2 W^K, V = X_3 W^V$$

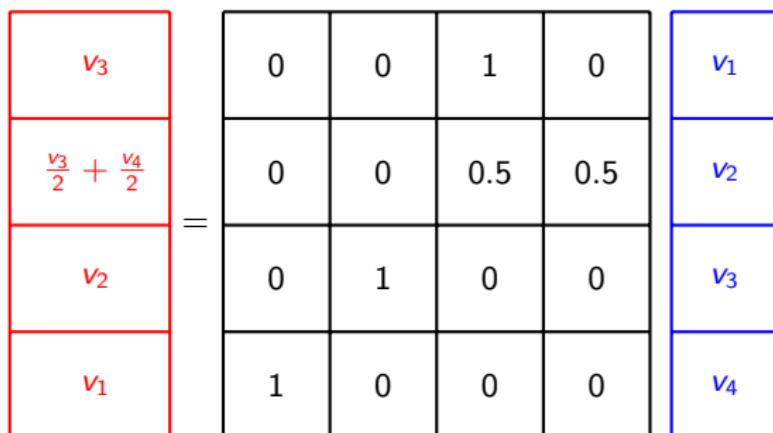


Query-Key-Value Attention

- Attention is the basis of the transformer architecture which obtains state-of-the art results in several domains such as NLP, computer vision, speech recognition. / Attention est base de l'architecture transformer qui obtient SOTA dans plusieurs domaines.

$$\text{Attention}(X_1, X_2, X_3) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

$$Q = X_1 W^Q, K = X_2 W^K, V = X_3 W^V$$



Multi-head Attention

- We calculate parallel attentions, and then combine them. / On calcule des attentions parallèles, et les combine.

$$\text{MHA}(X_1, X_2, X_3) = \text{Concatenate}(\text{Attention}_1, \dots, \text{Attention}_h) W^O$$

$$\text{where } \text{Attention}_i = \text{softmax} \left(\frac{X_1 W_i^Q (X_2 W_i^K)^\top}{\sqrt{d_k}} \right) X_3 W_i^V$$

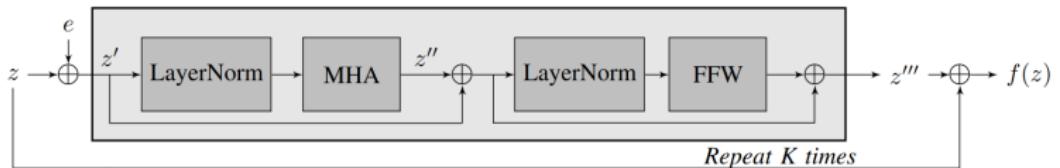
$$X_1, X_2, X_3 \in \mathbb{R}^{T \times L}, W^Q, W^K, W^V \in \mathbb{R}^{L \times L}$$

Self-Attention and the Transformer Encoder

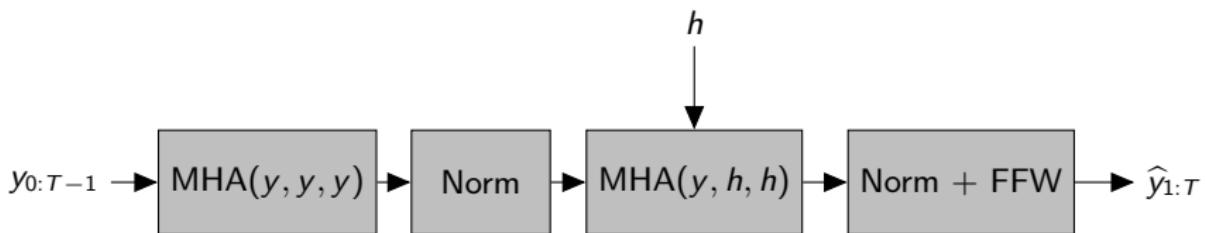
- If the inputs are the same sequence, this is called self-attention / Si les séquences d'entrée sont les mêmes, on appelle ça self-attention

$$\text{Self-Attention}(X) = \text{MHA}(X, X, X)$$

- In addition to Multihead attention, we also have, normalizations, a feed-forward layer, positional embeddings, and skip connections. / On a aussi des normalisations, des connections qui sautent, et une couche feed-forward, et les embeddings positionnel.



The Transformer Decoder



- I am ignoring the skip connections for simplicity. / Je n'inclus pas les connections qui sautent pour la simplicité dans la figure.
- The idea is very simple to RNN with attention, but we do it via the multi-head attention layers.

Replacing the Attention Layer with Multi-Head Attention

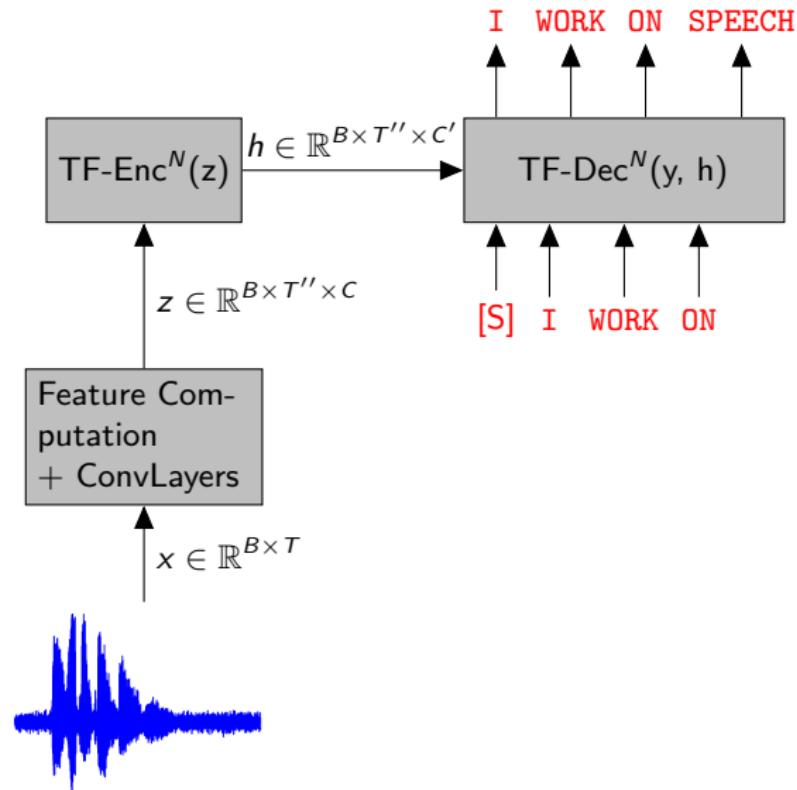


Table of Contents

Speech Recognition

RNN Based ASR

Transformer ASR

CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

Problem Definition

Source Separation Methods

SepFormer

Moving towards real-life source separation

Data collection

Performance estimation

Evaluating the SI-SNR Estimator

User Study and Further Evaluation

Cross-Modal Representation Learning

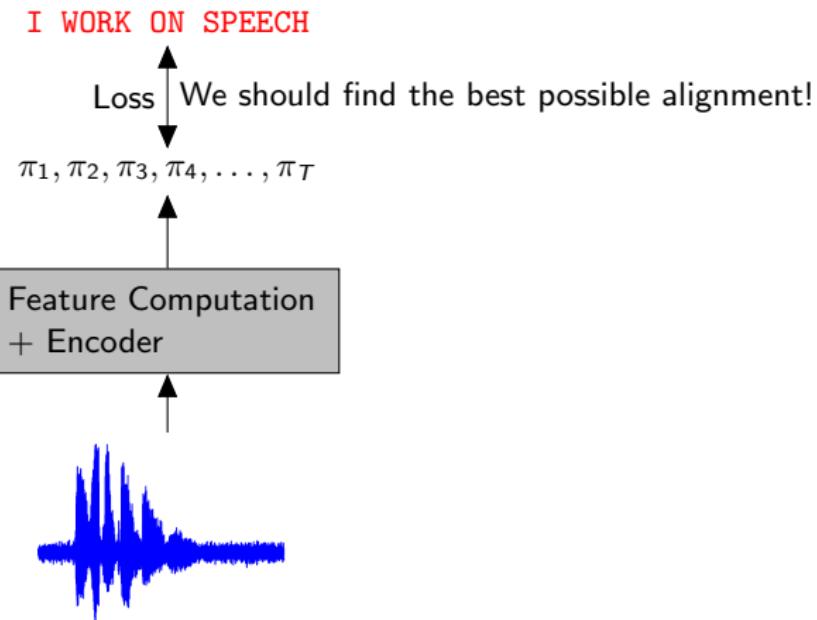
Interpretability

Posthoc Explanations for Audio Classifiers

A little bonus

Another Approach for Alignment

- ASR is a sequence-to-sequence problem. We need to resolve an alignment between the network output, and the target sequence. / ASR est un problème de séquence-à séquence. Il faut résoudre l'alignement entre la sortie du network, et la séquence des caractères à la sortie.
- We saw the encoder - decoder architecture. Decoder takes care of the alignment. / Le decoder trouve une résolution pour ce problème.



Naive alignment approach

- Maybe we can just merge the repeated characters? / Peut-être on peut merger les caractères qui répètent?

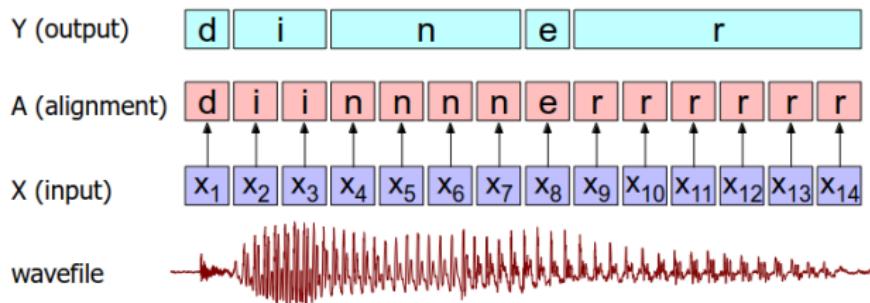


Image taken from Speech and Language Processing, Jurafsky, Martin

Naive alignment approach

- Maybe we can just merge the repeated characters? / Peut-être on peut merger les caractères qui répètent?

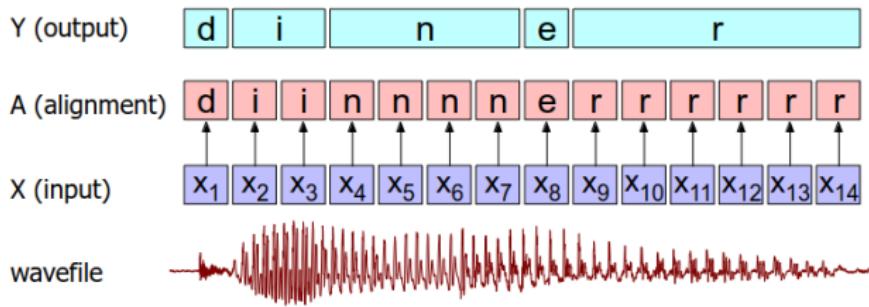


Image taken from Speech and Language Processing, Jurafsky, Martin

- What's wrong with this? / Qu'est-ce qui ne fonctionne pas ici?

Naive alignment approach

- Maybe we can just merge the repeated characters? / Peut-être on peut merger les caractères qui répètent?

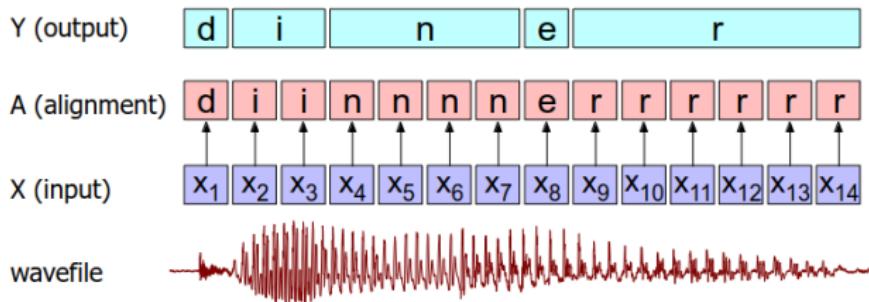


Image taken from Speech and Language Processing, Jurafsky, Martin

- What's wrong with this? / Qu'est-ce qui ne fonctionne pas ici?
- Dinner vs diner ??? , silences?

Connectionist Temporal Classification

- We can however find alignments by inserting a separator character. / On peut trouver un alignment en trouvant un meilleur alignment.

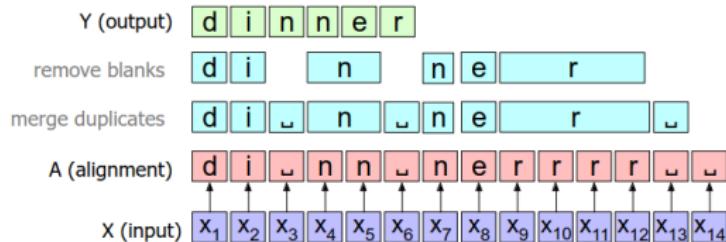


Image taken from Speech and Language Processing, Jurafsky, Martin

Connectionist Temporal Classification

- We can however find alignments by inserting a separator character. / On peut trouver un alignment en trouvant un meilleur alignment.

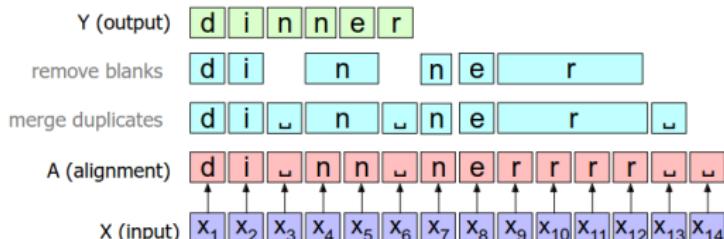


Image taken from Speech and Language Processing, Jurafsky, Martin

- How do we find this alignment though? Alignments which would work for the example above. / Comment trouve-t-on un alignment? Ces alignements suivants fonctionnerait par exemple pour l'exemple en haut:

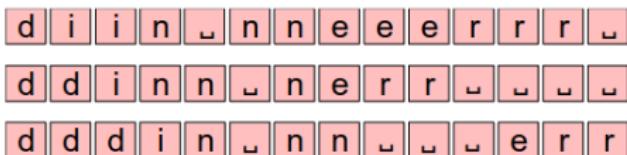


Image taken from Speech and Language Processing, Jurafsky, Martin

Finding the best possible alignment

- Training objective:

$$\max_{\theta} p(y_{1:T} | \pi_{1:T}, \theta) = \max_{\theta} \sum_{A_{1:T}} p(y_{1:T}, A_{1:T} | \pi_{1:T}, \theta)$$

- The CTC assumes a Markov chain on $A_{1:T}$ / CTC suppose un chaine Markov:

$$p(y_{1:T}, A_{1:T}) = \prod_t p(y_t | A_t) p(A_{t+1} | A_t) = \prod_t \pi_{A_t} p(A_{t+1} | A_t)$$

Finding the best possible alignment

- Training objective:

$$\max_{\theta} p(y_{1:T} | \pi_{1:T}, \theta) = \max_{\theta} \sum_{A_{1:T}} p(y_{1:T}, A_{1:T} | \pi_{1:T}, \theta)$$

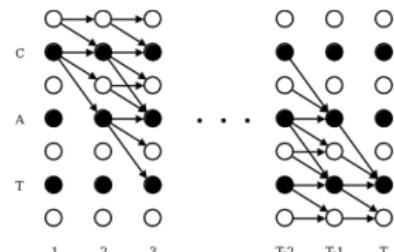
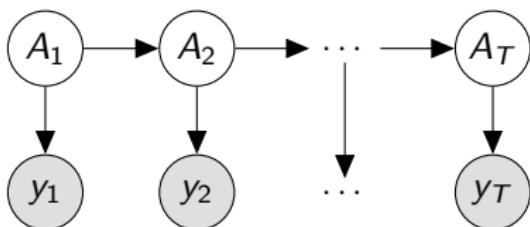
- The CTC assumes a Markov chain on $A_{1:T}$ / CTC suppose un chaine Markov:

$$p(y_{1:T}, A_{1:T}) = \prod_t p(y_t | A_t) p(A_{t+1} | A_t) = \prod_t \pi_{A_t} p(A_{t+1} | A_t)$$

- This is an HMM! (with a specific transition structure) / C'est un HMM avec une structure de transition spécifique.

Emission Model: $p(y_t | A_t) = \pi_{A_t}$

Transition Model:



How do we calculate CTC then?

- We need to calculate / On doit calculer:

$$p(y_{1:T}) = \sum_{A_{1:T}} p(y_{1:T}, A_{1:T})$$

$$\alpha(A_t) = p(y_t|A_t) \sum_{A_{t-1}} p(A_t|A_{t-1}) p(y_{t-1}|A_{t-1}) \dots p(y_2|A_2) \underbrace{\sum_{A_1} p(A_2|A_1)p(y_1|A_1)}_{\alpha(A_2)} \underbrace{p(A_1)}_{\alpha(A_1)}$$
$$\underbrace{\alpha(A_{t-1})}_{\alpha(A_{t-1})}$$

$$p(y_{1:T}) = \sum_{A_T} \alpha(A_T)$$

How do we calculate CTC then?

- We need to calculate / On doit calculer:

$$p(y_{1:T}) = \sum_{A_{1:T}} p(y_{1:T}, A_{1:T})$$

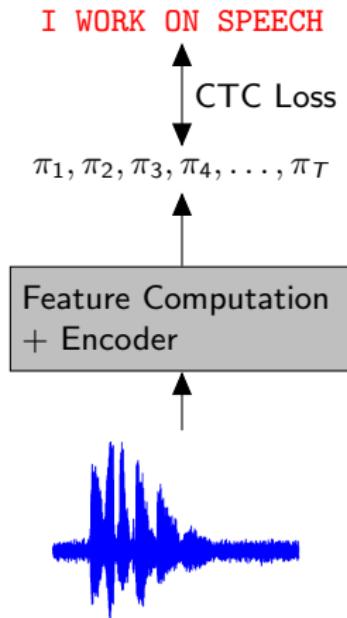
- Dynamic Programming to the rescue (the forward pass we talked about two weeks ago) / Programmation Dynamique va nous sauver. (la recurrence en avancant dont on a parlé il y a deux semaines)

$$\alpha(A_t) = p(y_t|A_t) \sum_{A_{t-1}} p(A_t|A_{t-1}) p(y_{t-1}|A_{t-1}) \dots p(y_2|A_2) \underbrace{\sum_{A_1} p(A_2|A_1)p(y_1|A_1)}_{\alpha(A_2)} \underbrace{p(A_1)}_{\alpha(A_1)}$$
$$\underbrace{\alpha(A_{t-1})}_{\alpha(A_2)}$$

$$p(y_{1:T}) = \sum_{A_T} \alpha(A_T)$$

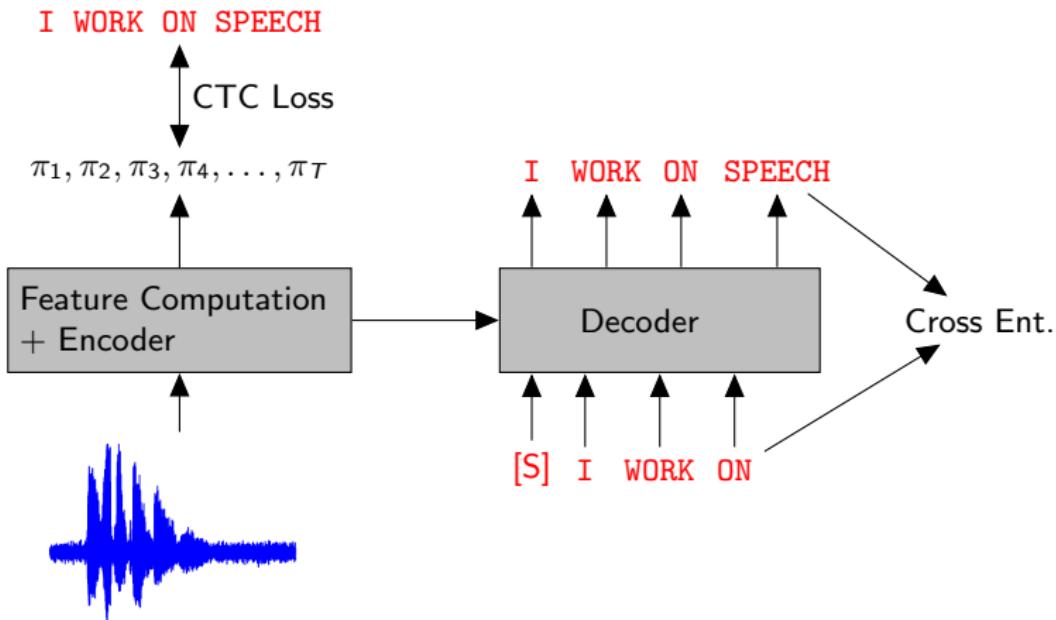
Of course you can also just call `torch.nn.CTCLoss.` :)

CTC Based ASR



CTC + Encoder-Decoder

- It's also possible to use both CTC and a decoder! / On peut aussi utiliser CTC et un décodeur.



Incorporating a Language Model

- In addition to the probability $p(y_{1:T}|X)$, we can also include a language model score $\mathcal{LM} = \prod_t p(y_t|y_{1:t-1})$. The final score function when decoding is then as follows:

- ▶ On peut aussi incorporer un language model quand on fait le decodage. Dans ce cas ci le score final est comme le suivant:

$$\hat{Y} = \arg \max_Y [\lambda \log p_{encdec}(Y|X) + (1 - \lambda) \log p_{CTC}(Y|X) + \gamma p_{LM}(Y)]$$

Performance Evaluation

■ Word-Error Rate:

$$WER = 100 \times \frac{Insertions + Substitutions + Deletions}{Total N.Words}$$

REF:	i	***	**	UM	the	PHONE	IS	i	LEFT	THE	portable	****	PHONE	UPSTAIRS	last	night
HYP:	i	GOT	IT	TO	the	*****	FULLEST	i	LOVE	TO	portable	FORM	OF	STORES	last	night
Eval:	I	I	S	D	S		S	S	I	S	S					

This utterance has six substitutions, three insertions, and one deletion:

$$\text{Word Error Rate} = 100 \frac{6+3+1}{13} = 76.9\%$$

Image taken from Speech and Language Processing, Jurafsky, Martin

Typical Performance under LibriSpeech

- LibriSpeech is a 16kHz speech dataset with over 1000 hours of audio books. Sentences are aligned at the sentence level. / LibriSpeech est un jeu de données large qui a au-delà de 1000 heures. Les phrases sont alignées de niveau phrase.
 - ▶ The transformer model in SpeechBrain obtains 2 % WER on test-clean.
 - ▶ With a wav2vec pretrained encoder trained with CTC, we obtain 1.9 % WER.
 - ▶ An RNN based encoder-decoder model obtains 3 \$ WER.
- If you want to check yourself / Si vous voulez voir vous même
<https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech>

Real-life Generalization

- Deep learning methods have been shown to work extremely well on synthetic benchmarks.
 - ▶ ~ 2% Word-error rate on LibriSpeech test set.
 - ▶ >20dB SNR for speech separation on WSJ0-2Mix.
 - ▶ >3 PESQ for speech enhancement on Voicebank.
- It is not clear that these models generalize well on real-data.
 - ▶ **Recording 1:**

Real-life Generalization

- Deep learning methods have been shown to work extremely well on synthetic benchmarks.
 - ▶ ~ 2% Word-error rate on LibriSpeech test set.
 - ▶ >20dB SNR for speech separation on WSJ0-2Mix.
 - ▶ >3 PESQ for speech enhancement on Voicebank.
- It is not clear that these models generalize well on real-data.
 - ▶ **Recording 1:**
 - ▶ **HE'LL BEEN RUTH MORAL** (WER 3.0 on LS)

Real-life Generalization

- Deep learning methods have been shown to work extremely well on synthetic benchmarks.
 - ▶ ~ 2% Word-error rate on LibriSpeech test set.
 - ▶ >20dB SNR for speech separation on WSJ0-2Mix.
 - ▶ >3 PESQ for speech enhancement on Voicebank.
- It is not clear that these models generalize well on real-data.
 - ▶ **Recording 1:**
 - ▶ **HE'LL BEEN RUTH MORAL** (WER 3.0 on LS)
 - ▶ **PESTEIN THIS MORAL** (WER 2.2 on LS)

Real-life Generalization

- Deep learning methods have been shown to work extremely well on synthetic benchmarks.
 - ▶ ~ 2% Word-error rate on LibriSpeech test set.
 - ▶ >20dB SNR for speech separation on WSJ0-2Mix.
 - ▶ >3 PESQ for speech enhancement on Voicebank.
- It is not clear that these models generalize well on real-data.
 - ▶ **Recording 1:**
 - ▶ **HE'LL BEEN RUTH MORAL** (WER 3.0 on LS)
 - ▶ **PESTEIN THIS MORAL** (WER 2.2 on LS)
 - ▶ **TESTING THIS MOLO** (pretr. wav2vec2 backbone, fine-tuned on CommonVoice)

Real-life Generalization

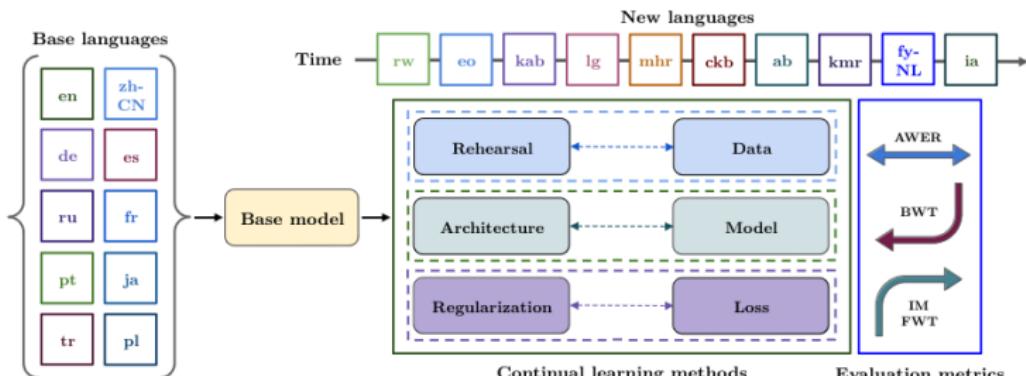
- Deep learning methods have been shown to work extremely well on synthetic benchmarks.
 - ▶ ~ 2% Word-error rate on LibriSpeech test set.
 - ▶ >20dB SNR for speech separation on WSJ0-2Mix.
 - ▶ >3 PESQ for speech enhancement on Voicebank.
- It is not clear that these models generalize well on real-data.
 - ▶ **Recording 1:**
 - ▶ HE'LL BEEN RUTH MORAL (WER 3.0 on LS)
 - ▶ PESTEIN THIS MORAL (WER 2.2 on LS)
 - ▶ TESTING THIS MOLO (pretr. wav2vec2 backbone, fine-tuned on CommonVoice)
 - ▶ **Recording 2:**
 - ▶ HE WAS MALLOW (WER 3.0 on LS)
 - ▶ PESSIM WAS MODEL (WER 2.2 on LS)
 - ▶ TESTING THIS MODEL (pretr. wav2vec2 backbone, fine-tuned on CommonVoice)

Continual Learning of Multi-Lingual ASR Models

1

CL-MASR: A Continual Learning Benchmark for Multilingual ASR

Luca Della Libera*, Pooneh Mousavi*, Salah Zaiem, Cem Subakan, Mirco Ravanelli



The paper: <https://arxiv.org/pdf/2310.16931.pdf>

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Tacotron

- Again a sequence-to-sequence architecture / Encore une fois une architecture séquence-à-séquence

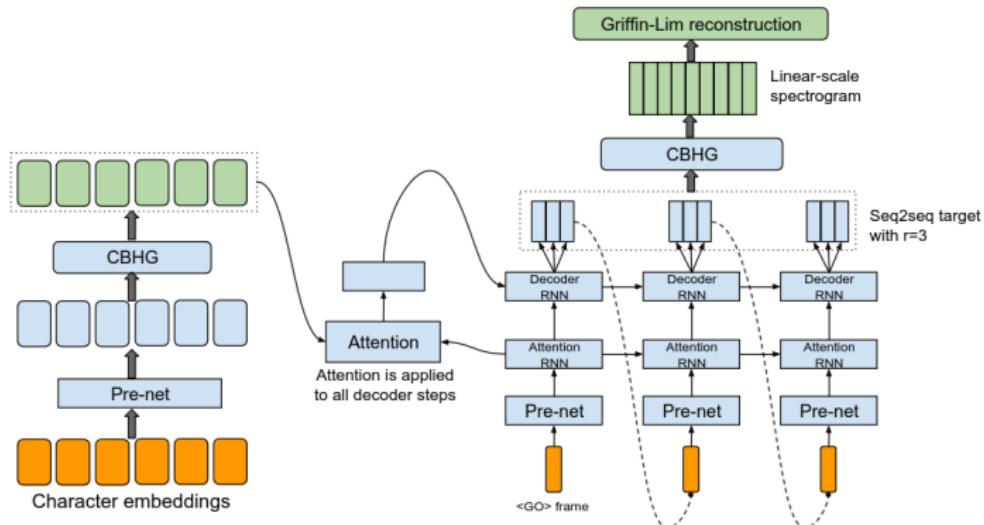


Image taken from the original paper <https://arxiv.org/pdf/1703.10135.pdf>

A typical modern TTS pipeline

- First sequence of characters is converted into a series of representations (encoder) / L'encodeur transforme la texte à une serie de vecteurs.
- Then, a decoder, with the help of an attention mechanism, predicts the next mel-spectrogram column. / Le decodeur avec l'aide d'une mechanisme d'attention prédit la colonne suivante d'un mel-spectrogramme.
- The mel-spectrogram is turned into audio with a vocoder. / Le mel-spectrogramme est transformé à l'audio en utilisant un vocoder.

Tacotron 2

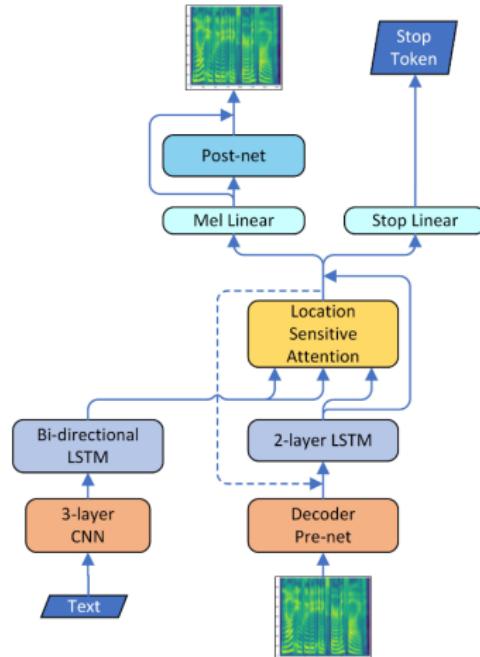


Image taken from <https://arxiv.org/pdf/1809.08895.pdf>

Transformer for TTS

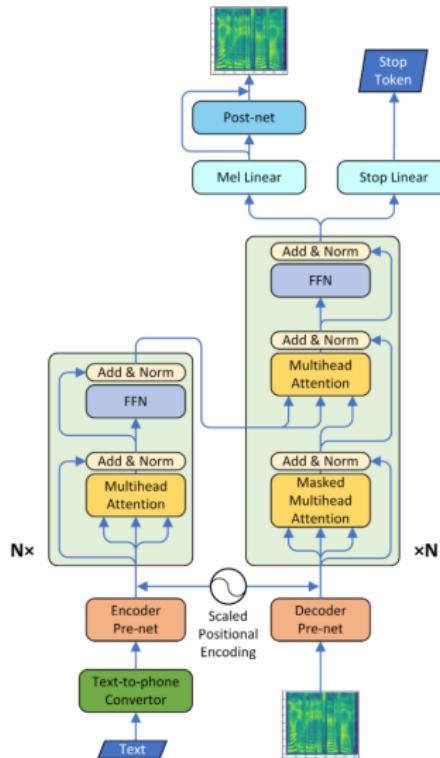
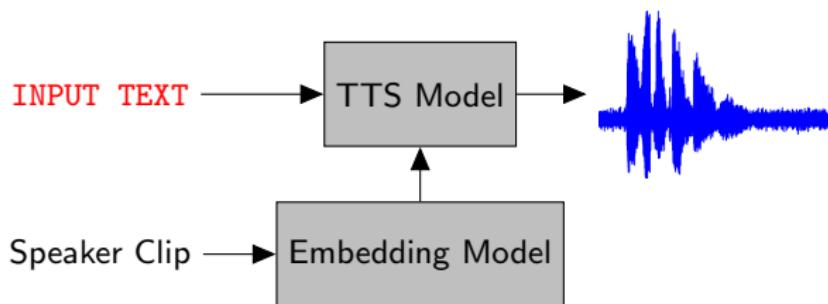


Image taken from <https://arxiv.org/pdf/1809.08895.pdf>

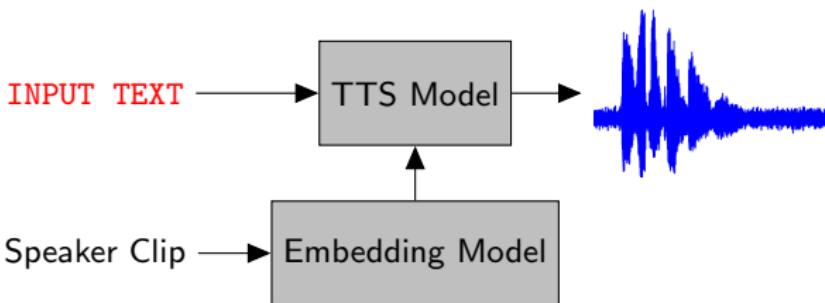
Speaker Embeddings for Zero-Shot Speaker Generalization in Multi-Speaker TTS

- Adaptation of a multi-speaker TTS system with voice snippets.



Speaker Embeddings for Zero-Shot Speaker Generalization in Multi-Speaker TTS

- Adaptation of a multi-speaker TTS system with voice snippets.



- Examples (Unseen speakers):
 - ▶ Speaker1 Clip, Speaker1 UnseenPhrase
 - ▶ Speaker2 Clip, Speaker2 UnseenPhrase
 - ▶ Speaker3 Clip, Speaker3 UnseenPhrase
- Goals: Improving speaker generalization, Increasing clip invariance through better speaker embeddings
 - ▶ Speaker1 Clip, Speaker1 UnseenPhrase

ECAPA-TDNN for speaker embeddings

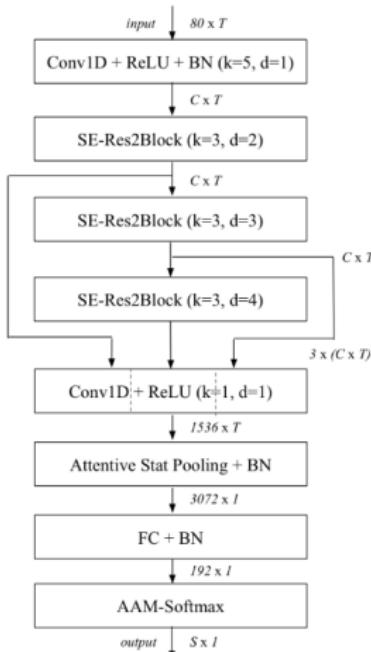


image taken from the original paper
<https://arxiv.org/pdf/2005.07143.pdf>

- Model pretrained on speaker-id works well for speaker embeddings.
- Pretrained model on SpeechBrain Huggingface <https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>
- There's also the X-Vector model, which is a smaller. <https://huggingface.co/speechbrain/spkrec-xvect-voxceleb>

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

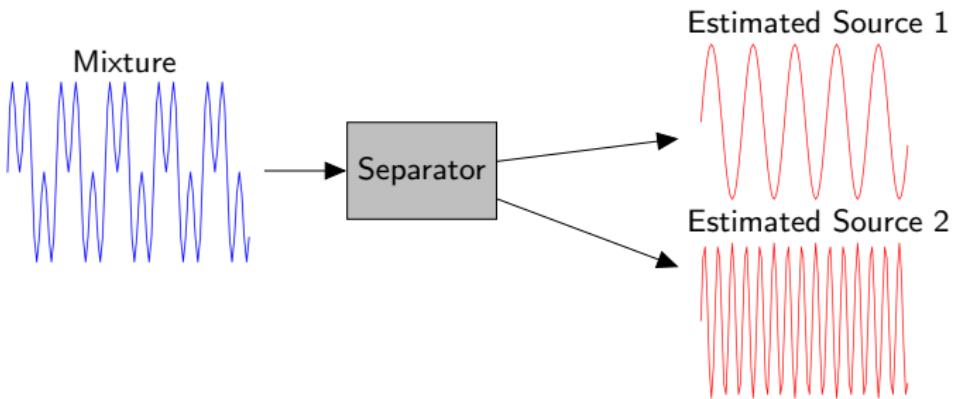
Cross-Modal Representation Learning

Interpretability

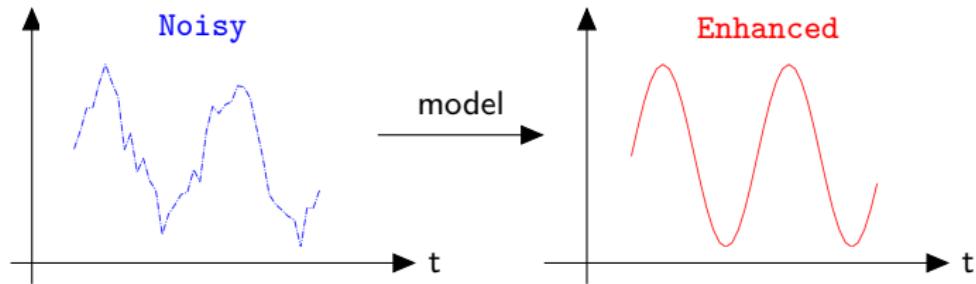
- Posthoc Explanations for Audio Classifiers

A little bonus

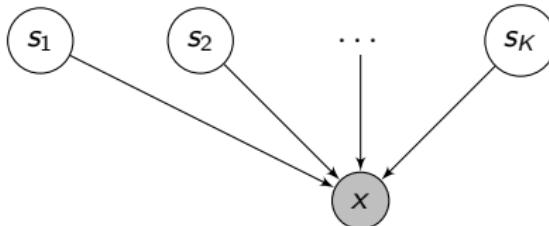
Source Separation



Speech Enhancement



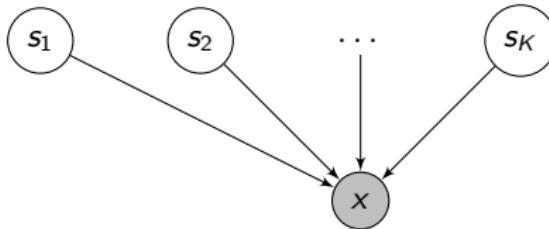
Source Separation



- The observation x is dependent on latent factors s_1, s_2, \dots, s_K .
 - ▶ Technical definition:

$$s_1 \sim p(s_1) \dots s_K \sim p(s_K)$$
$$x \sim p(x|s_1, \dots, s_K)$$

Source Separation



- The observation x is dependent on latent factors s_1, s_2, \dots, s_K .
 - ▶ Technical definition:

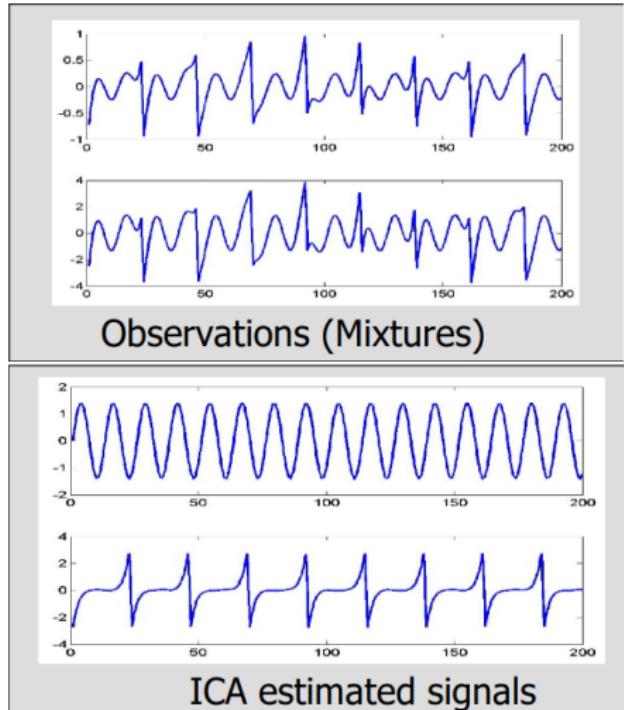
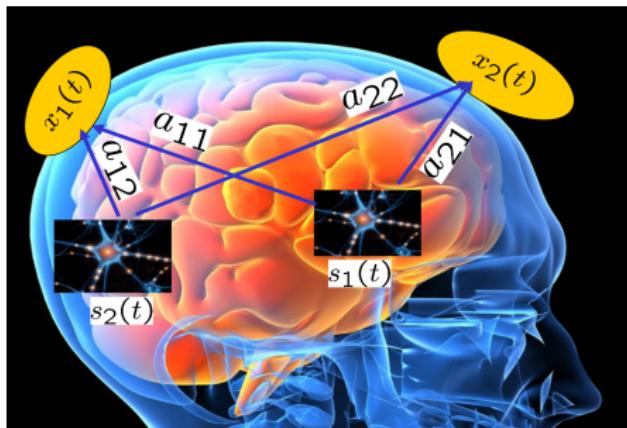
$$\begin{aligned}s_1 &\sim p(s_1) \dots s_K \sim p(s_K) \\ x &\sim p(x|s_1, \dots, s_K)\end{aligned}$$

- Additive separation model:

$$x(t) = \sum_{k=1}^K a_k s_k(t)$$

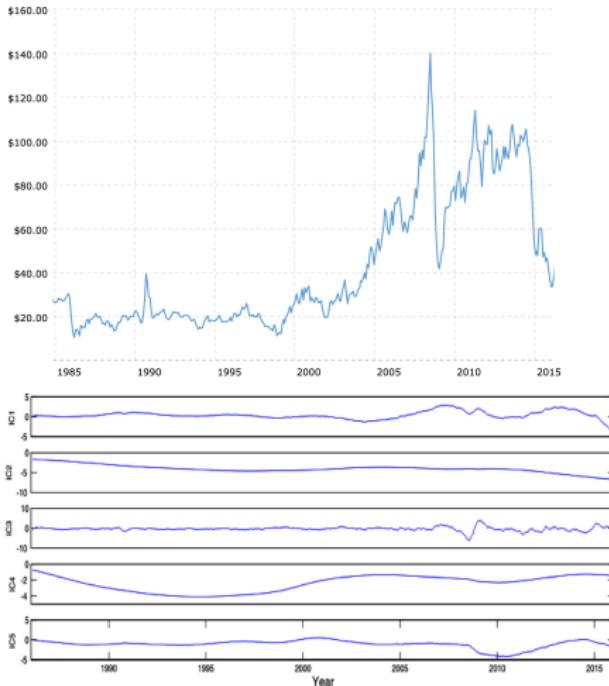
- ▶ This is a very general formulation and captures several different models / algorithms.
E.g. PCA, ICA, Factor Analysis, Mixture models, NMF, HMMs, Linear Dynamical Systems (Kalman filters)

Source Separation Application



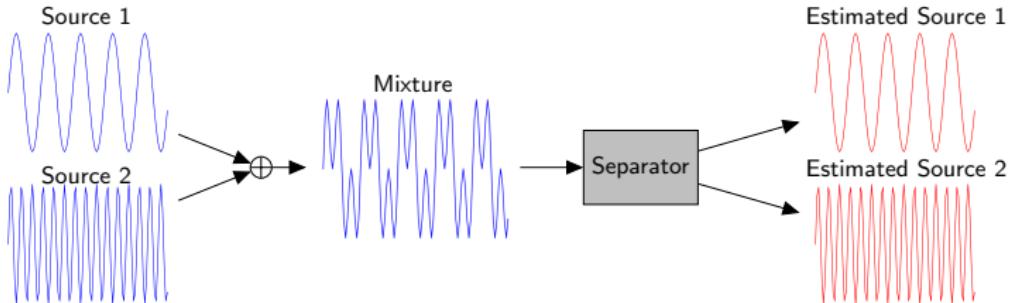
[images taken from https://www.cs.cmu.edu/~bapoczos/other_presentations/ICA_26_10_2009.pdf]

Source Separation for Financial Data



- In the paper **Factor analysis of financial time series using EEMD-ICA based approach** the authors decompose oil prices using an ICA variant.
- They claim:
 - ▶ IC1 is correlated to USD.
 - ▶ IC2 is correlated to oil supply and demand.
 - ▶ IC3 is correlated to political and extreme events.
 - ▶ IC4 reflects cyclical nature of oil prices.
 - ▶ IC5 is correlated with stock, gold markets.

Single-Microphone Source Separation Problem



- **Goal:** To recover the original sources from the observed mixture
- **Applications:** Music production, hearing devices, meeting analysis, editing software, and more...
- **Some of my contributions**
 - ▶ Hierarchical tensor factorizations
 - ▶ Globally optimal unsupervised source separation with FHMM.
 - ▶ Neural network analogs to matrix factorization (**best paper award**)
 - ▶ GANs in source separation
 - ▶ **SepFormer**, a self-attention based source separation architecture and obtain state-of-the-art results on multiple datasets.
 - ▶ **REAL-M** dataset and evaluation framework

Table of Contents

Speech Recognition

RNN Based ASR

Transformer ASR

CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

Problem Definition

Source Separation Methods

SepFormer

Moving towards real-life source separation

Data collection

Performance estimation

Evaluating the SI-SNR Estimator

User Study and Further Evaluation

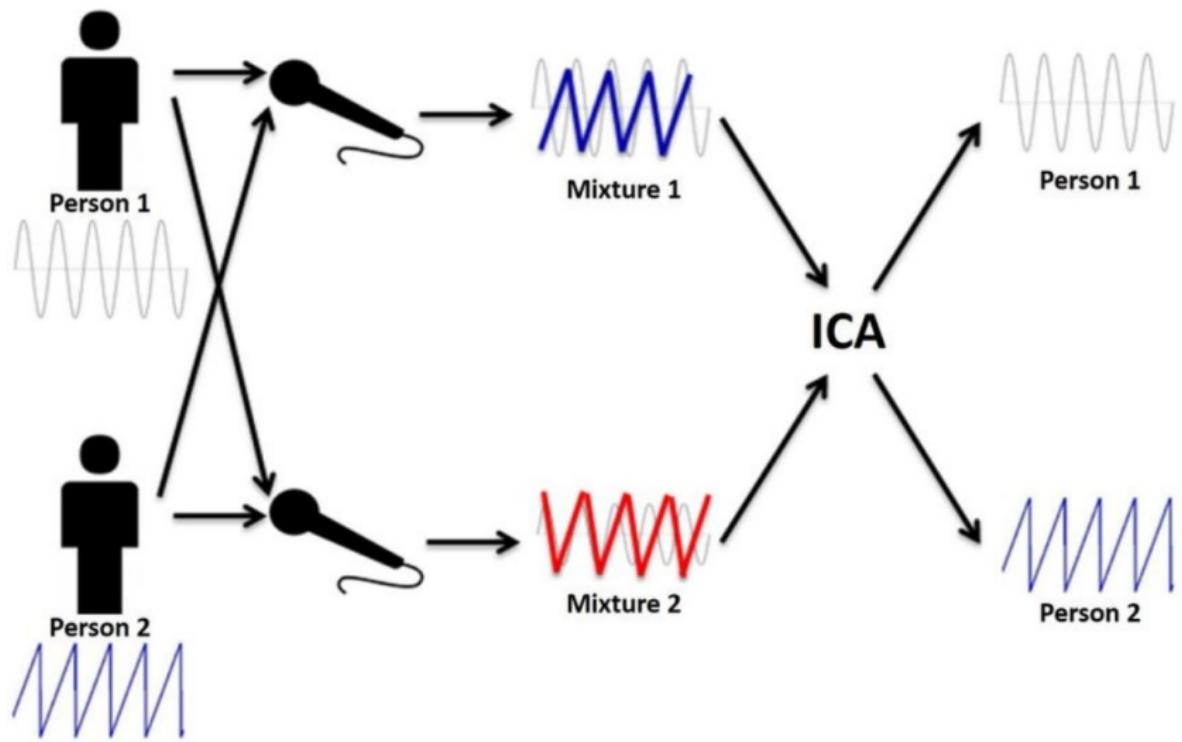
Cross-Modal Representation Learning

Interpretability

Posthoc Explanations for Audio Classifiers

A little bonus

Independent Component Analysis



Independent Component Analysis

$$x_{1:T} = \sum_{k=1}^K a_k s_{1:T}^k$$
$$\underbrace{\begin{pmatrix} \cdots & x_{1:T}^1 & \cdots \\ \cdots & x_{1:T}^2 & \cdots \\ \vdots & \vdots & \cdots \\ \cdots & x_{1:T}^N & \cdots \end{pmatrix}}_{\text{Mixtures}} = \underbrace{\begin{pmatrix} a_{1,1} & \dots & a_{1,K} \\ a_{2,1} & \dots & a_{2,K} \\ \dots & \dots & \dots \\ a_{N,1} & \dots & a_{N,K} \end{pmatrix}}_{\text{Mixing Matrix}} \underbrace{\begin{pmatrix} \cdots & s_{1:T}^1 & \cdots \\ \cdots & s_{1:T}^2 & \cdots \\ \vdots & \vdots & \cdots \\ \cdots & s_{1:T}^K & \cdots \end{pmatrix}}_{\text{Sources}}$$

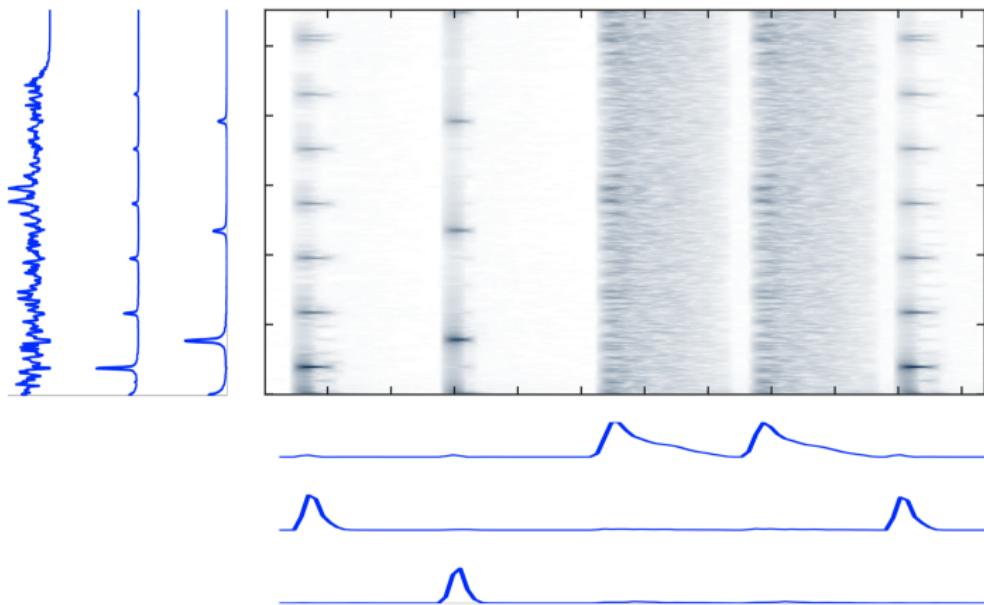
- Unsupervised!
- We try to estimate statistically independent components $s^{1:K}$.
- 2×2 case:

$$\begin{pmatrix} \cdots & x_{1:T}^1 & \cdots \\ \cdots & x_{1:T}^2 & \cdots \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} \cdots & s_{1:T}^1 & \cdots \\ \cdots & s_{1:T}^2 & \cdots \end{pmatrix}$$

- Works in time domain, and requires $N \geq K$.

Non-Negative Matrix Factorization

[Lee, Seung 1999][Smaragdis 2003, Non-Negative Matrix Factorization for Polyphonic Music Transcription]



Popular NMF model: $X = WH$

$$\min_{W,H} \|X - WH\|, \text{ s.t. } W \geq 0, H \geq 0$$

Early deep learning approaches

[Huang et al. 2014, Deep Learning for Monoaural Speech Separation], figure taken from the paper.

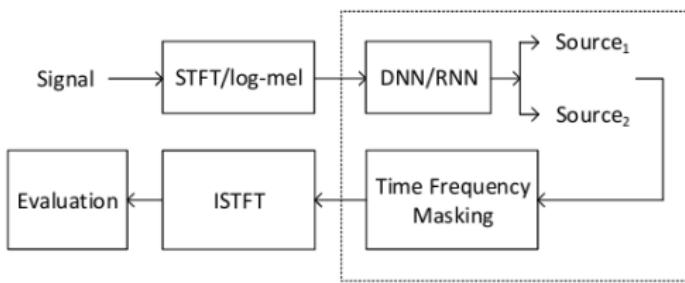


Fig. 1: Proposed framework

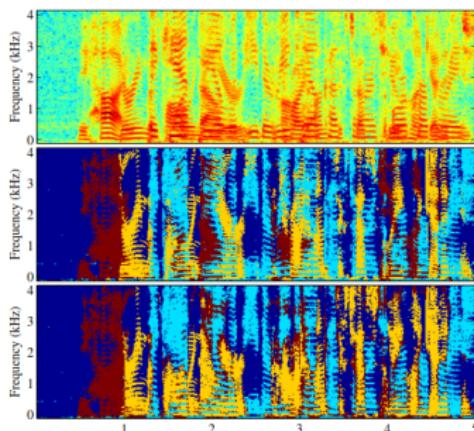
- Works on magnitude STFTs.
- Uses the mixture phase to reconstruct

Deep Clustering

[Hershey et al. 2015, Deep Clustering], The idea is to find an embedding for the affinity matrix.

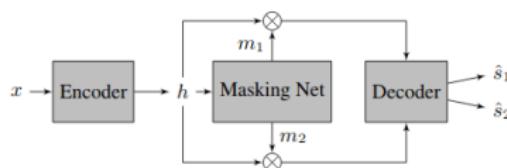
$$\min_{\theta} \|YY^T - f_{\theta}(x)f_{\theta}(x)^T\| \quad (1)$$

They learn an embedding for each time-frequency bin, and minimize this affinity based loss. In test time, they cluster the embeddings.



End-to-End training: Masking based architecture

- The masking based architecture [Luo, Mesgarani 2018, ConvTasNet],



- Encoder:** A time-transformation representation is calculated by passing via the encoder. Special case: STFT
- Masking Network:** A masking network estimates an element-wise mask m_i for each source.
- Decoder:** For each source i , we reconstruct the estimated source by passing the filtered representation $h * m_i$ through the decoder.

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer**

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

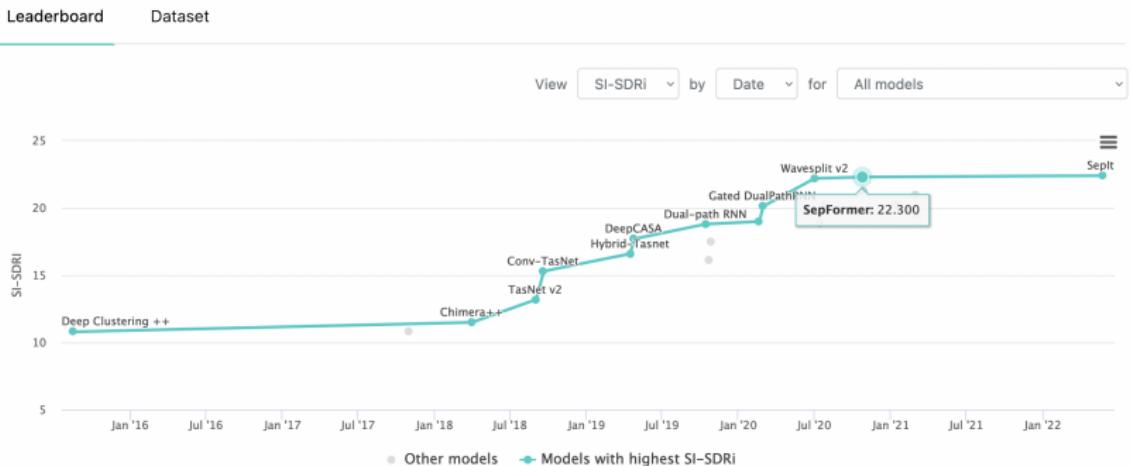
Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

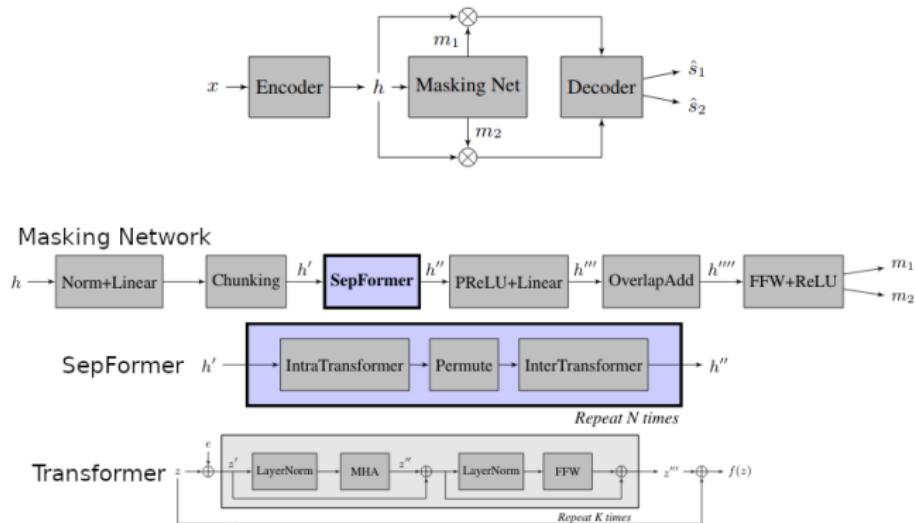
A little bonus

WSJ0-2Mix Leaderboard

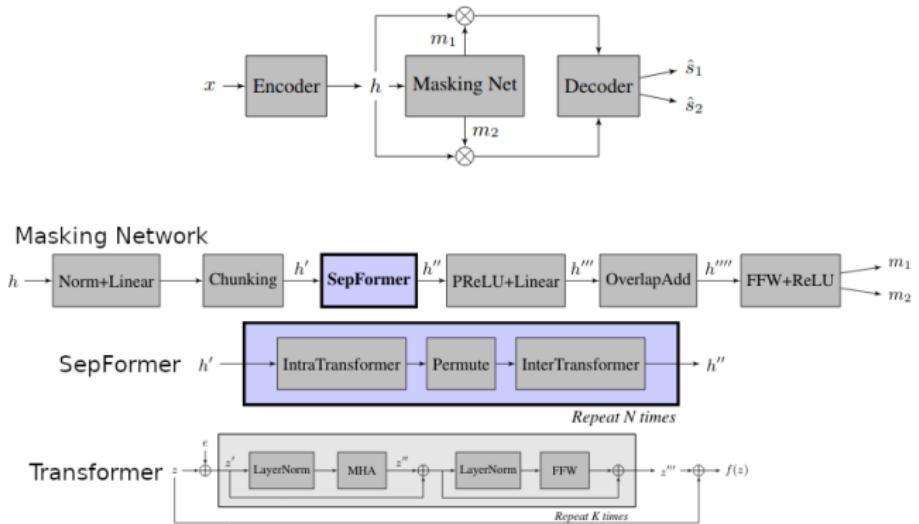


- Taken from <https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix> on October 2022. SepFormer stayed state of the art on WSJ0-2Mix from October 2020-September 2022.

The SepFormer Architecture



The SepFormer Architecture

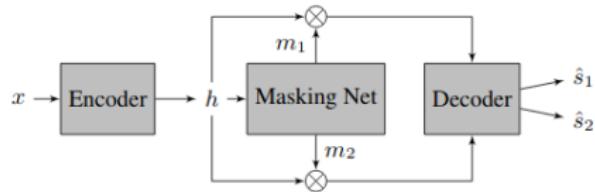


We train this architecture with permutation invariant SI-SNR.

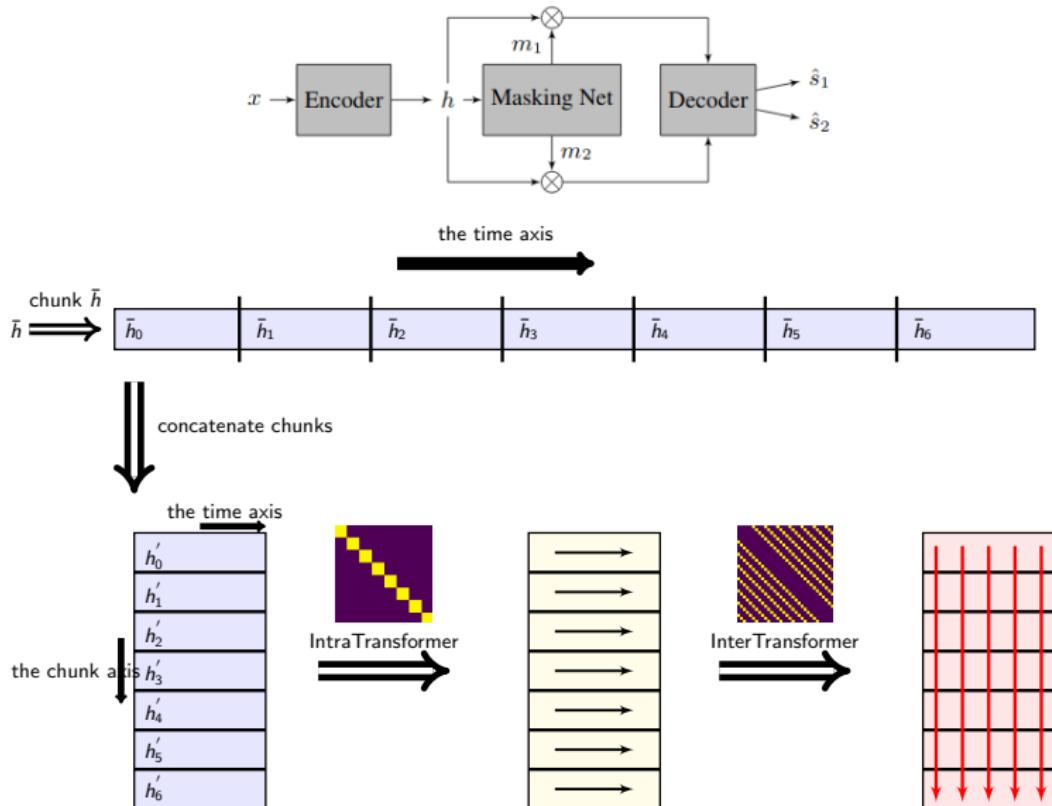
$$s_{\text{target}} := \frac{\hat{s}^\top s}{\|s\|^2} s, \quad e_{\text{noise}} := \hat{s} - s_{\text{target}}, \quad \text{SI-SNR} := 10 \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right)$$

$$\text{PIT-SISNR} = \sum_k \min_{k' \in \mathcal{P}} 10 \log_{10} \left(\frac{\|s_{\text{target}}^k\|^2}{\|\hat{s}^{k'} - s_{\text{target}}^k\|^2} \right)$$

SepFormer Architecture



SepFormer Architecture



Best Results on Mixtures of 2 speakers (WSJ0-2Mix)

Model	SI-SNRi	SDRi	# Param	Stride
Tasnet	10.8	11.1	n.a	20
SignPredictionNet	15.3	15.6	55.2M	8
ConvTasnet	15.3	15.6	5.1M	10
Two-Step CTN	16.1	n.a.	8.6M	10
DeepCASA	17.7	18.0	12.8M	1
FurcaNeXt	n.a.	18.4	51.4M	n.a.
DualPathRNN	18.8	19.0	2.6M	1
sudo rm -rf	18.9	n.a.	2.6M	10
VSUNOS	20.1	20.4	7.5M	2
DPTNet	20.2	20.6	2.6M	1
Wavesplit	22.2	22.3	29M	1
SepFormer	22.3	22.4	26M	8

$$SNR \propto 10 \log \left(\frac{\text{Ener. Signal}}{\text{Ener. Noise}} \right)$$

Best Results on Mixtures of 3 Speakers (WSJ0-3Mix)

Model	SI-SNRi	SDRi	# Param
ConvTasnet	12.7	13.1	5.1M
DualPathRNN	14.7	n.a	2.6M
VSUNOS	16.9	n.a	7.5M
Wavesplit	17.8	18.1	29M
Sepformer	19.5	19.7	26M

Example Results on Test Set:

[Click for Mixture](#)

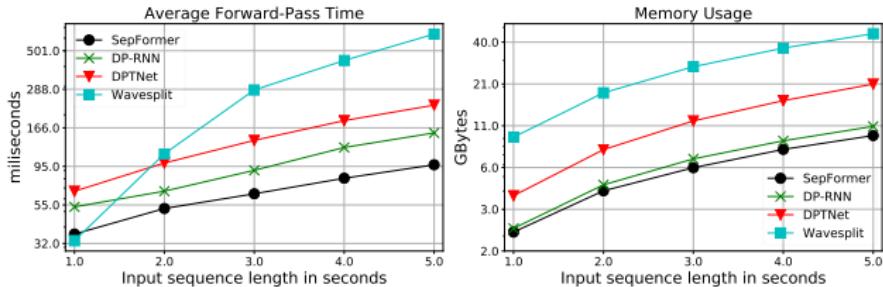
[Click for Estimated Source1](#)

[Click for Estimated Source2](#)

[Click for Estimated Source3](#)

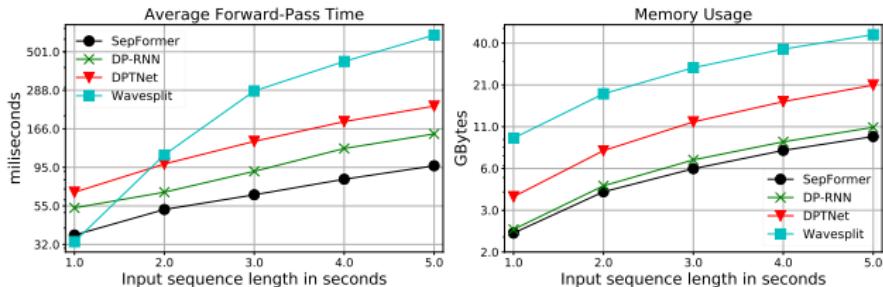
Speed/Memory Comparison with Other Methods

Speed and Memory Comparison on Forward Pass:

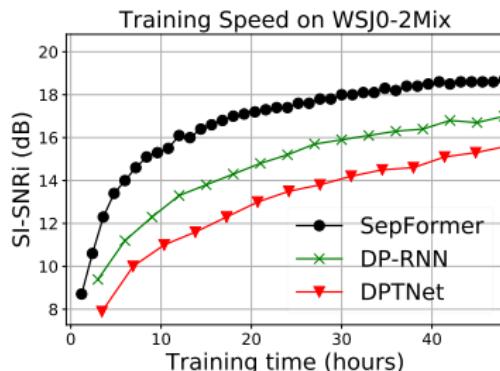


Speed/Memory Comparison with Other Methods

Speed and Memory Comparison on Forward Pass:



Training Curve Comparison:



Environmental Corruption

We try our model with environmental noise / reverberation.

Best results on the WHAM dataset (noise).

Model	SI-SNRi	SDRi
ConvTasnet	12.7	-
Learnable fbank	12.9	-
Wavesplit	16.0	16.5
Sepformer	16.4	16.7

Best results on the WHAMR (noise + reverb) dataset.

Model	SI-SNRi	SDRi
ConvTasnet	8.3	-
BiLSTM Tasnet	9.2	-
Wavesplit	13.2	12.2
Sepformer	14.0	13.0

Cross-Dataset Experiment

We test our model trained on WSJ0-2Mix on LibriMix.

Model	SI-SNRI	SDRi
ConvTasnet	14.7	-
Sepformer trained on WSJ0-2Mix	17.0	17.5
Wavesplit	20.5	20.7
Sepformer	20.2	20.5
Sepformer + FT	20.6	20.8

We test our model trained on WSJ0-3Mix on LibriMix.

Model	SI-SNRI	SDRi
ConvTasnet	10.4	-
Sepformer trained on WSJ0-3Mix	15.0	15.6
Wavesplit	17.5	18.0
Sepformer	18.2	18.6
Sepformer + FT	18.7	19.0

Note: We release our pretrained models, training scripts on SpeechBrain!

Synthetic vs Real Life Mixture

Synthetic: WSJ0-2Mix test set

[Click for Mixture](#)

[Click for Estimated Source1](#)

[Click for Estimated Source2](#)

Real-life: One mic, two people speaking, reverberant environment

[Click for Mixture](#)

[Click for Estimated Source1](#)

[Click for Estimated Source2](#)

[Click for Mixture](#)

[Click Estimated Source 1](#)

[Click Estimated Source 2](#)

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Real-life machine learning

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

- A lot of machine learning focuses on improving on benchmarks
- This is good, but it's likely that there is a reality-gap.

Reality GAP on MNIST



- Small, fixed size images
- Perfectly aligned images,
- Uniform backgrounds
- **We need at least an evaluation set for evaluating models on real-life data.**

Closing the reality gap

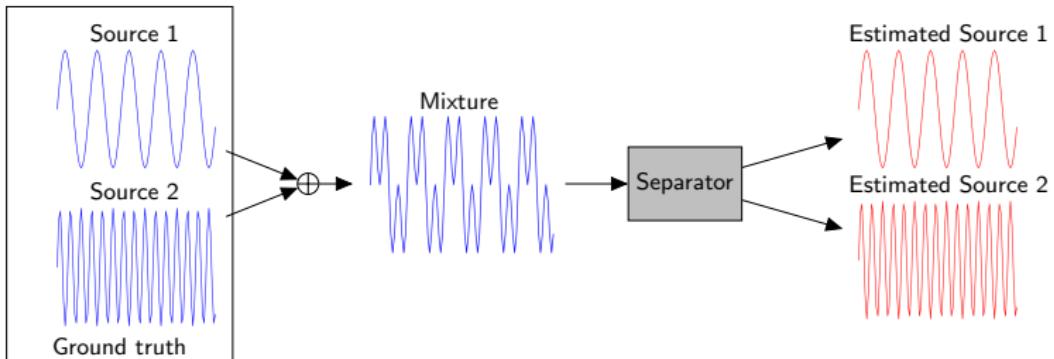
- We need evaluation sets that represent the challenges of real-life so that researchers can more meaningfully benchmark their performance.
- We can then design data augmentations, and models to improve performance on real-life data.

Closing the reality gap

- We need evaluation sets that represent the challenges of real-life so that researchers can more meaningfully benchmark their performance.
- We can then design data augmentations, and models to improve performance on real-life data.
- An important hurdle: Ground truth data.

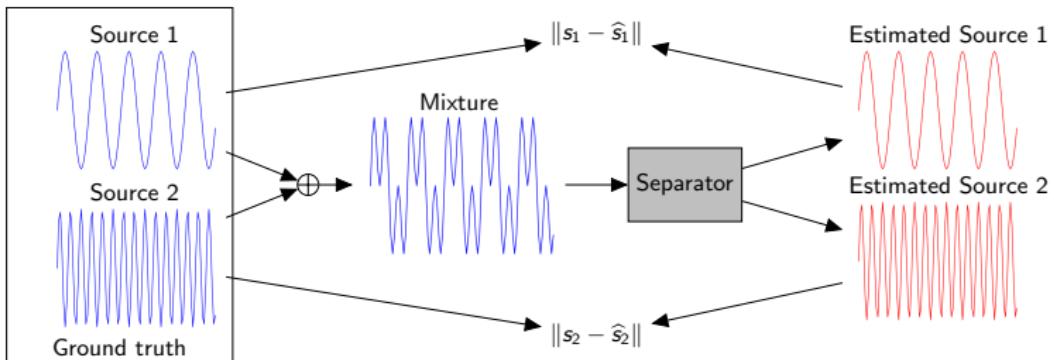
Lack of ground truth in real-life separation

Synthetic source separation datasets



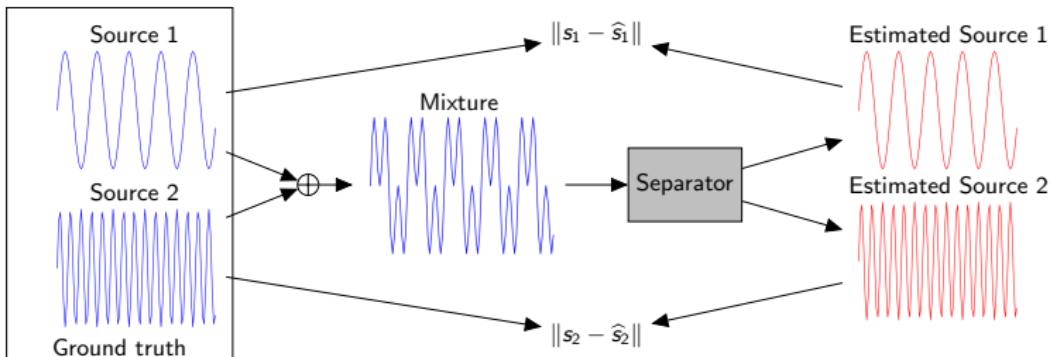
Lack of ground truth in real-life separation

Synthetic source separation datasets

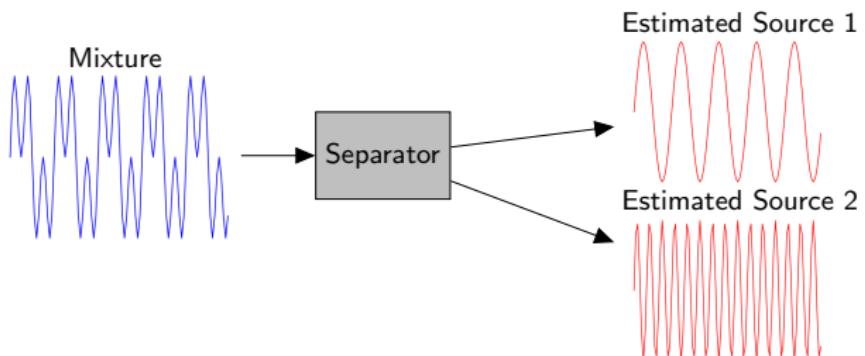


Lack of ground truth in real-life separation

Synthetic source separation datasets

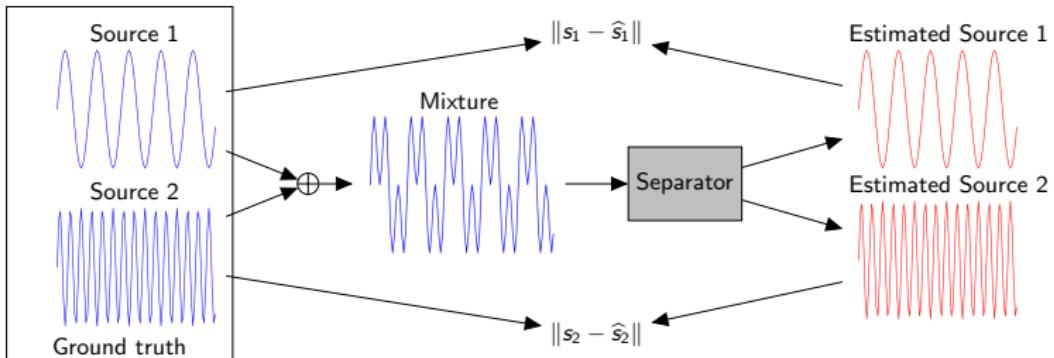


Real-life source separation

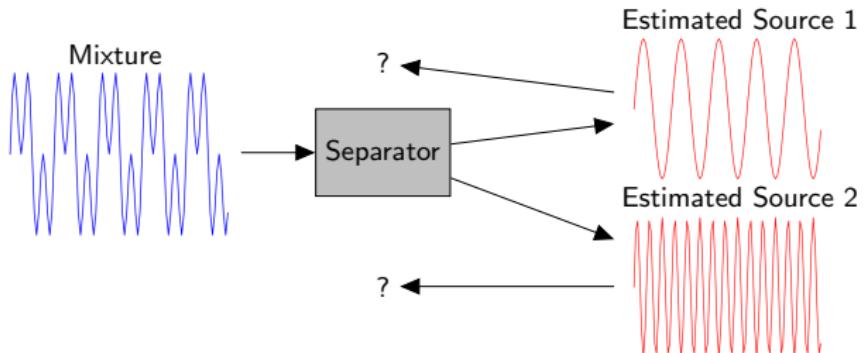


Lack of ground truth in real-life separation

Synthetic source separation datasets



Real-life source separation

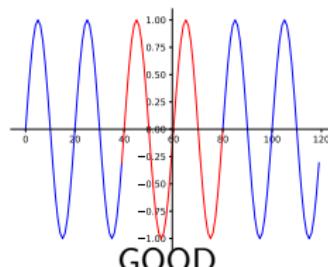
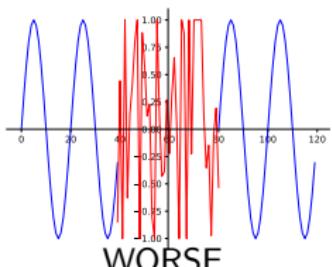


Tackling the lack of ground truth

- In real-life datasets we often do not have the ground truth information.
- Examples:
 - ▶ Image in-painting
 - ▶ Speech enhancement and separation
 - ▶ Image super-resolution
 - ▶ Evaluating the performance of chatbots
 - ▶ Website design optimization to maximize user retention
- The lack of ground truth prevents evaluating estimation quality on real-life data.

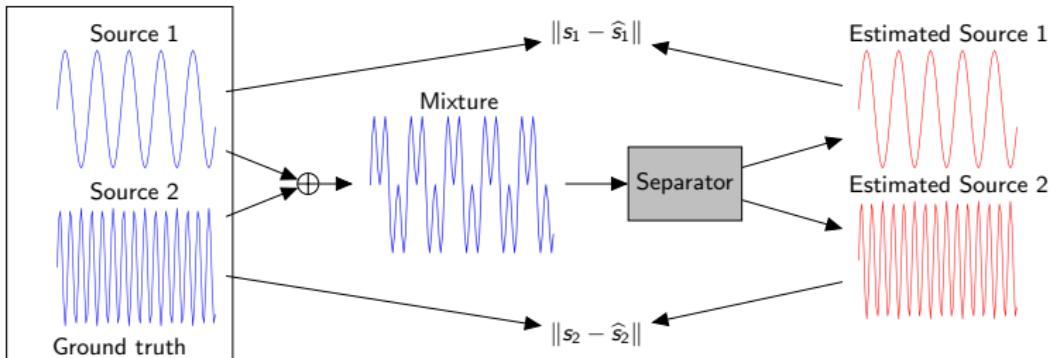
Tackling the lack of ground truth

- In real-life datasets we often do not have the ground truth information.
- Examples:
 - ▶ Image in-painting
 - ▶ Speech enhancement and separation
 - ▶ Image super-resolution
 - ▶ Evaluating the performance of chatbots
 - ▶ Website design optimization to maximize user retention
- The lack of ground truth prevents evaluating estimation quality on real-life data.
- We can however **estimate** the performance!
- We can train a model to estimate the performance.

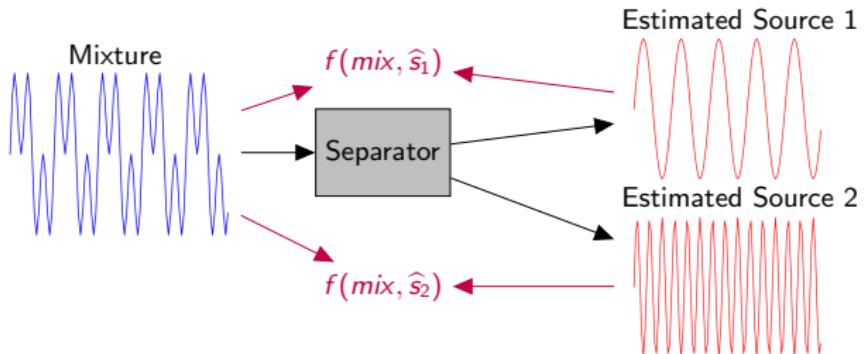


Tackling the lack of ground truth

Synthetic source separation datasets



Real-life source separation



REAL-M: Towards Speech Separation on Real Mixtures

- Goal: Systematic Evaluation of Speech Separation Models on Real-Life Speech Mixtures.
- Contributions:
 - ▶ We propose a dataset for **real-life speech separation**. The dataset is **crowdsourced**, hence **scalable and diverse** in acoustic conditions, recording hardware, speakers.
 - ▶ We show that **blind SI-SNR estimation** is a feasible way to evaluate real-life speech separation.
 - ▶ Therefore, this opens up a scalable methodology for large-scale real-life source separation evaluation.
 - ▶ The 5th most viewed poster in ICASSP 2022! (out of 1900 posters)

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. [Click to hear examples.](#)

Choose the genders of the speakers

- 1 male and 1 female speaker
- 2 male speakers
- 2 female speakers

Choose whether the speakers are native English speakers

- 2 native English speakers
- 2 non-native English speakers
- 1 native English speaker, 1 non-native English speaker

Please write your native language(s) if you are not native English speaker.

The collected utterances will be used in a dataset for developing a speech separation system.
Your recording might be publicly released with this dataset in an anonymous way.
Please check the box which signifies that each person in the recording accept this.

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. Click to hear examples.

Choose the genders of the speakers

- 1 male and 1 female speaker
- 2 male speakers
- 2 female speakers

Choose whether the speakers are native English speakers

- 2 native English speakers
- 2 non-native English speakers
- 1 native English speaker, 1 non-native English speaker

Please write your native language(s) if you are not native English speaker.

The collected utterances will be used in a dataset for developing a speech separation system. Your recording might be publicly released with this dataset in an anonymous way. Please check the box which signifies that each person in the recording accept this.

Read the sentences

Sentence 1:

the crampness and the poverty are all intended

Sentence 2:

do you think so she replied with indifference

Record Audio

[Click the "Start Recording" button to start recording.](#)

[Start recording](#) [Stop recording](#)

After recording, please re-listen and make sure you can tell what is being said by both of the speakers.. We are looking for VERY clear and natural pronunciations. If not, please re-record.

Make sure relative levels for each speaker are roughly the same (one speaker should not be louder than the other).

You CAN NOT record by yourself. Sentences need to be read by two different people, reading at the same time, at the same room (no playback through loudspeaker)! Listen to this [example](#).

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. Click to hear examples.

Choose the genders of the speakers

- 1 male and 1 female speaker
- 2 male speakers
- 2 female speakers

Choose whether the speakers are native English speakers

- 2 native English speakers
- 2 non-native English speakers
- 1 native English speaker, 1 non-native English speaker

Please write your native language(s) if you are not native English speaker.

The collected utterances will be used in a dataset for developing a speech separation system. Your recording might be publicly released with this dataset in an anonymous way. Please check the box which signifies that each person in the recording accept this.

Sentence 1:

the crampness and the poverty are all intended

Sentence 2:

do you think so she replied with indifference

Record Audio

Read the sentences

Sentence 1:

the crampness and the poverty are all intended

Sentence 2:

do you think so she replied with indifference

After recording, please re-listen and make sure you can tell what is being said by both of the speakers.. We are looking for VERY clear and natural pronunciations. If not, please re-record.

Make sure relative levels for each speaker are roughly the same (one speaker should not be louder than the other).

You CAN NOT record by yourself. Sentences need to be read by two different people, reading at the same time, at the same room (no playback through loudspeaker)! Listen to this [example](#).

Received Audio

Select your file to upload to Mechanical Turk
Then Click the "Upload" button
Upload successful!

Your total number of submissions: 1

Your work stamp:

ZJZJ/0001/3NaEN_A8_2022-01-2611:16:00,13/101+00:00/01/01

Do not forget to copy-paste the work stamp we show on Mechanical Turk before going on to the next sentence!

Data collection

- We crowdsource a dataset to construct an evaluation of audio mixtures in real-life. Click to hear examples.

Choose the genders of the speakers

male and female speaker
 2 male speakers
 2 female speakers

Choose whether the speakers are native English speakers

2 native English speakers
 2 non-native English speakers
 1 native English speaker, 1 non-native English speaker

Please write your native language(s) if you are not native English speaker.

The collected utterances will be used in a dataset for developing a speech separation system. Your recording might be publicly released with this dataset in an anonymous way. Please check the box which signifies that each person in the recording accepts this.

Read the sentences

Sentence 1:
the crampness and the poverty are all intended

Sentence 2:
do you think so she replied with indifference

Record Audio

Start recording | Stop recording

After recording, please re-listen and make sure you can tell what is being said by both of the speakers.. We are looking for VERY clear and natural pronunciations. If not, please re-record.

Make sure relative levels for each speaker are roughly the same (one speaker should not be louder than the other).

You CAN NOT record by yourself. Sentences need to be read by two different people, reading at the same time, at the same room (no playback through loudspeaker)! Listen to this example.

In this task, we are asking you to record audio mixture with someone else, while you are in the same room. You should read the shown sentences at the same time (and wait one after the other). [Click to hear an example for what your recordings should sound like.](#)
We have developed a website which will allow you a series of two sentences that you will be asked to read and record with someone else in your household.
For each audio recording, we ask you to copy and paste the Work Stamp, that you will see on the website after uploading the mixture, in order to get paid.
Please note that we will be choosing the submitted mixture before accepting your work. If you submit empty recordings, or do not follow the rules specified in the website, we might need to reject your submission. So, please try to do high-quality work!
You will be asked to fill in a short questionnaire in the website. After that you can start submitting your recordings!
Do not click go back on the website during your online session!
You can go to the data collection website by clicking on the link below. Do not forget to read the instructions on the website!
<https://newrecorder.auditionresearch.com/>
After uploading your each mixture, submit the information you get from the website on mechanical turk, in order to get paid.
Work Stamp:
Enter the Work Stamp you see on the website after your upload.

- Contributors are asked simultaneously read the shown sentences.
- This gives a way to collect real-life speech mixtures in a scalable way. We interface our platform with Mechanical Turk.
- We collected 3 hours of speech, from 50 unique speakers, with various native (e.g. US, UK) and non-native (e.g. French, Italian, Persian, Indian, African) accents, in various conditions, with various recording equipment.

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation**
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(\cdot)$ such that,

$$f(x, \hat{s}) \approx \mathbf{SI-SNR}(s, \hat{s}),$$

where $f(\cdot)$ is a neural network, x is the mixture, \hat{s} is the source estimate, s is the true source.

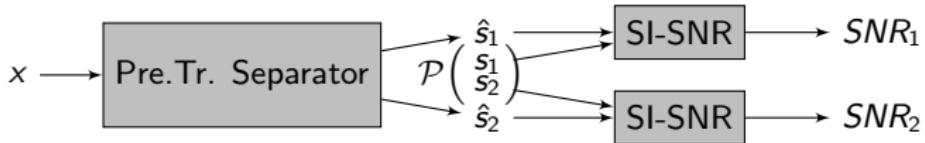
SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(\cdot)$ such that,

$$f(x, \hat{s}) \approx \text{SI-SNR}(s, \hat{s}),$$

where $f(\cdot)$ is a neural network, x is the mixture, \hat{s} is the source estimate, s is the true source.

- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)



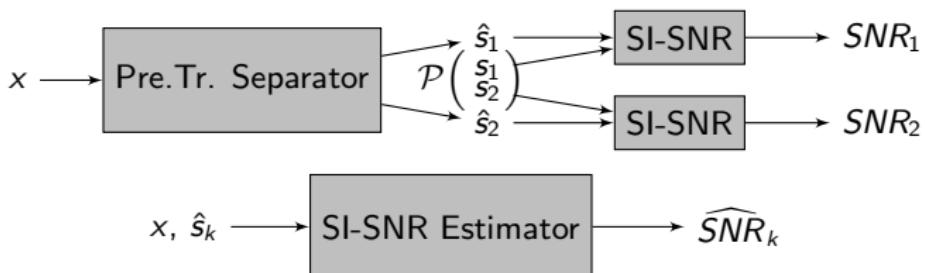
SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(\cdot)$ such that,

$$f(x, \hat{s}) \approx \text{SI-SNR}(s, \hat{s}),$$

where $f(\cdot)$ is a neural network, x is the mixture, \hat{s} is the source estimate, s is the true source.

- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)



$$\mathcal{L} = \|SNR_1 - \widehat{SNR}_1\| + \|SNR_2 - \widehat{SNR}_2\|.$$

- SI-SNR Estimator is a 5-layer convolutional NNet in the time domain.

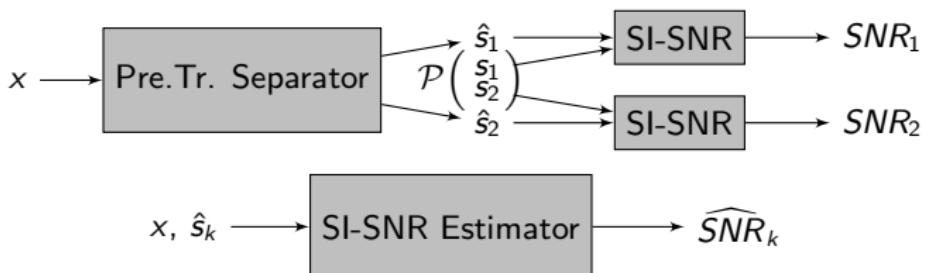
SI-SNR Estimator Training Pipeline

- We construct a performance estimator $f(\cdot)$ such that,

$$f(x, \hat{s}) \approx \text{SI-SNR}(s, \hat{s}),$$

where $f(\cdot)$ is a neural network, x is the mixture, \hat{s} is the source estimate, s is the true source.

- We train the estimator on synthetic mixtures on which ground truth information is available. (Also a lot of data!)



$$\mathcal{L} = \|SNR_1 - \widehat{SNR}_1\| + \|SNR_2 - \widehat{SNR}_2\|.$$

- SI-SNR Estimator is a 5-layer convolutional NNet in the time domain.
- Important Question:** Is this estimator going to work well (generalize to) real-mixtures?

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator**
- User Study and Further Evaluation

Cross-Modal Representation Learning

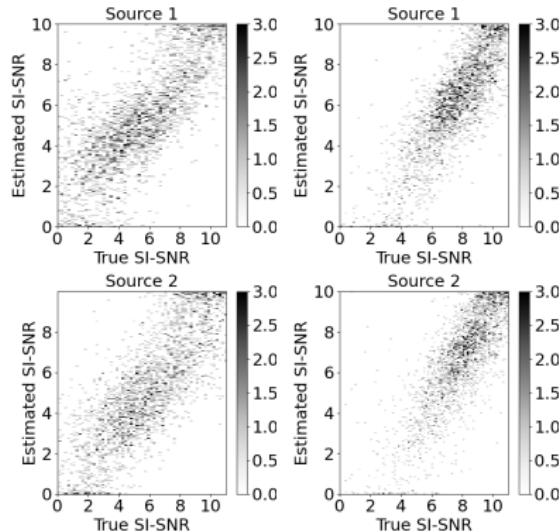
Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Evaluating the SI-SNR Estimator

- We first evaluate the SI-SNR Estimator on synthetic data.



- Evaluating on (left) LibriMix, (right) WHAMR!
- Both scatter plots correspond to Pearson correlation coefficient of 0.8.

Improving the SI-SNR Estimator

- We train with multiple separators.
- We observe improvement in the pearson correlation coefficient when we train with multiple separators.

Model	SI-SNR-Estimator 1 (<i>single</i>)		SI-SNR-Estimator 2 (<i>pool</i>)	
	LibriMix	WHAMR!	LibriMix	WHAMR!
SF	0.80	0.81	0.82	0.87
DPRNN	0.80	0.80	0.83	0.84
CTN	0.81	0.79	0.85	0.86

Improving the SI-SNR Estimator

- We train with multiple separators.
- We observe improvement in the pearson correlation coefficient when we train with multiple separators.

Model	SI-SNR-Estimator 1 (<i>single</i>)		SI-SNR-Estimator 2 (<i>pool</i>)	
	LibriMix	WHAMR!	LibriMix	WHAMR!
SF	0.80	0.81	0.82	0.87
DPRNN	0.80	0.80	0.83	0.84
CTN	0.81	0.79	0.85	0.86

- **Important Question:** Is this estimator going to work well (generalize to) real-mixtures?

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Evaluating the SI-SNR Estimator on REAL-M

- We validate the SI-SNR estimator with a user study on real-life data.
- We presented 50 random mixtures and the separation results to 5 users.
- We asked the users to rate the presented separation result between 1-5.

Mixture

Estimate for Source 1

Estimate for Source 2

How good is the separation in your opinion for source 1? (Level of Separation + sound quality)

- bad
- poor
- fair
- good
- excellent

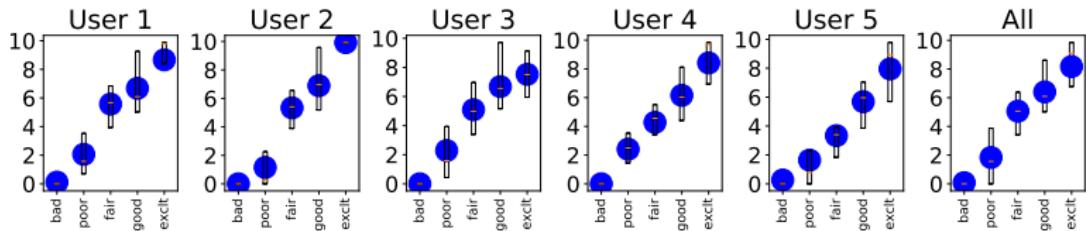
How good is the separation in your opinion for source 2? (Level of Separation + sound quality -- Note that if you hear the same source twice, this means the level of separation bad, so you should vote 'bad' in this case.)

- bad
- poor
- fair
- good
- excellent

Submit

Results of User Study

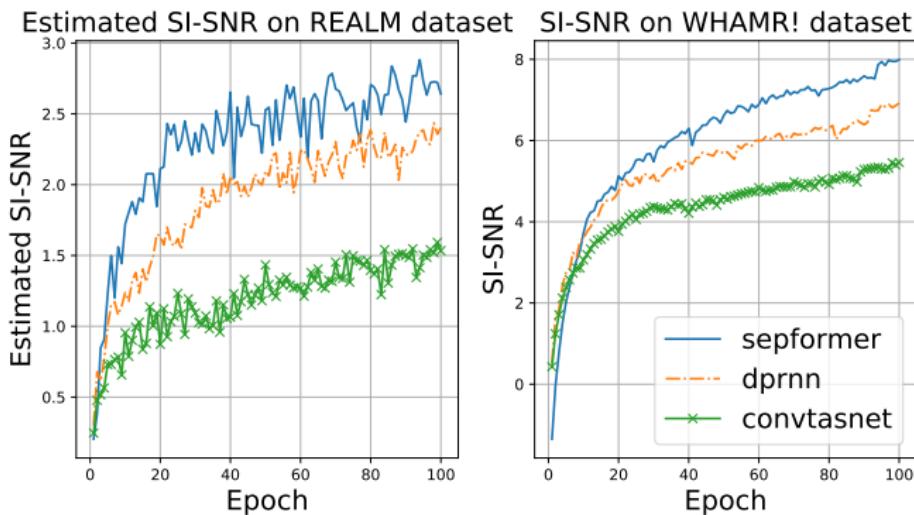
- Results of the user study suggest that on average the opinion scores correlate-well with SNR estimates.



- Y-axes show the estimated-SNR, X-axes show the user rating.

Further evaluation of SI-SNR Estimator

- The performance rankings of models on synthetic data holds true for REAL-M as well.
- We also observe that with training epochs performance on REAL-M dataset improves.



Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
 - ▶ Scalable data collections
 - ▶ Variability
 - ▶ Blind Performance estimation

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
 - ▶ Scalable data collections
 - ▶ Variability
 - ▶ Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	\widehat{SNR} Synth	\widehat{SNR} Real
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
 - ▶ Scalable data collections
 - ▶ Variability
 - ▶ Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	\widehat{SNR} Synth	\widehat{SNR} Real
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

- Next steps:
 - ▶ Casual talking, Meeting settings
 - ▶ Scaling up

Conclusions

- With the REAL-M framework, we are moving towards working with real-mixtures. It provides,
 - ▶ Scalable data collections
 - ▶ Variability
 - ▶ Blind Performance estimation
- The performance on real-life data is far worse than synthetic benchmarks.

Separator	$\widehat{\text{SNR Synth}}$	$\widehat{\text{SNR Real}}$
SepFormer	8.40	2.88
DPRNN	7.04	2.43
CTN	5.49	1.59

- Next steps:
 - ▶ Casual talking, Meeting settings
 - ▶ Scaling up
 - ▶ Improving the generalization (Data augmentations, Using the performance estimators, Using pretrained models, ...)

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

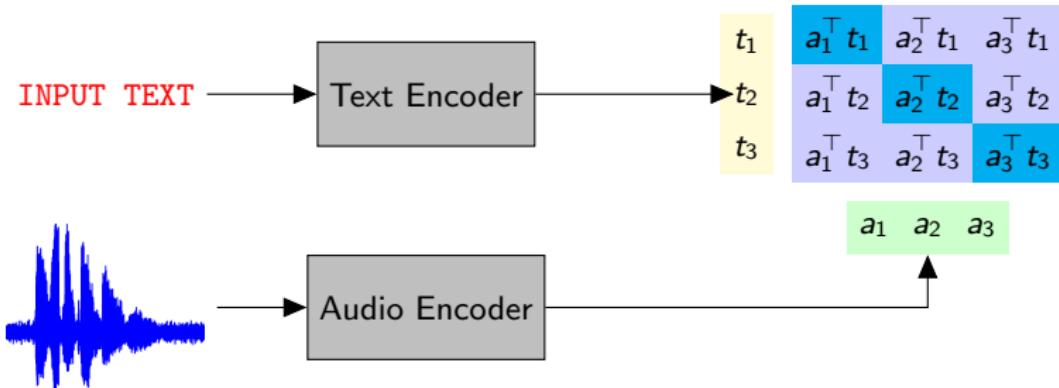
Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Cross-Modal Representation Learning

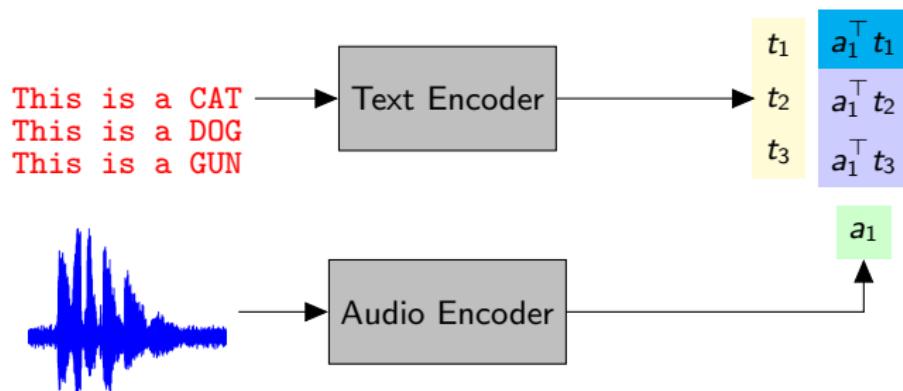


■ CLAP: Contrastive Language-Audio Pretraining

- ▶ We maximize $a_i^T t_j$ for $i = j$, and minimize for $i \neq j$.
- ▶ This enables text-based audio retrieval, zero-shot classification.

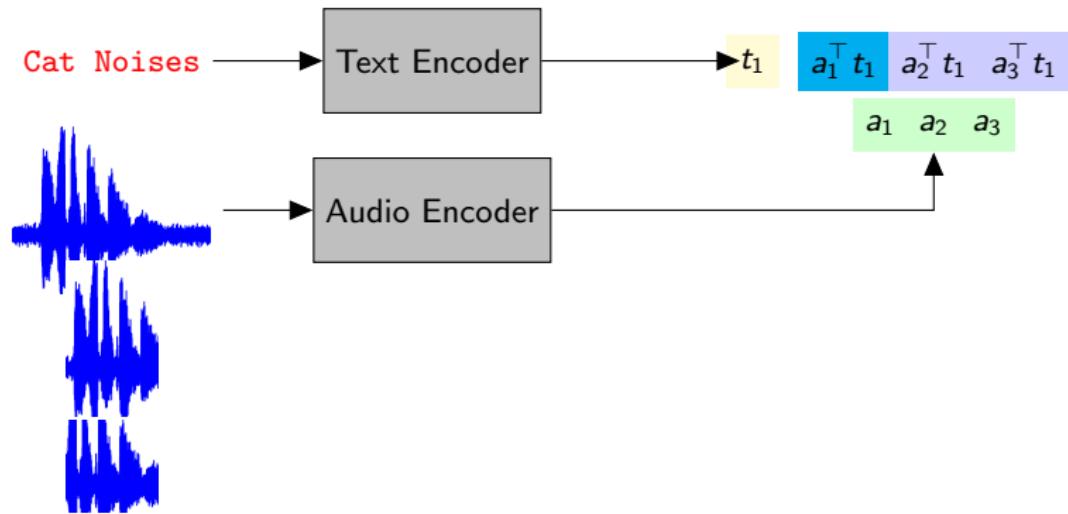
Cross-Modal Representation Learning

■ Zero-shot evaluation

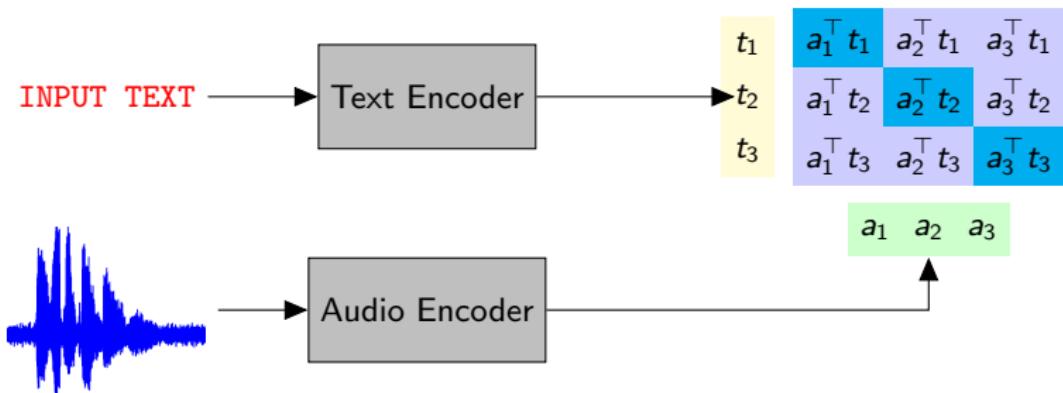


Cross-Modal Representation Learning

■ Audio Retrieval



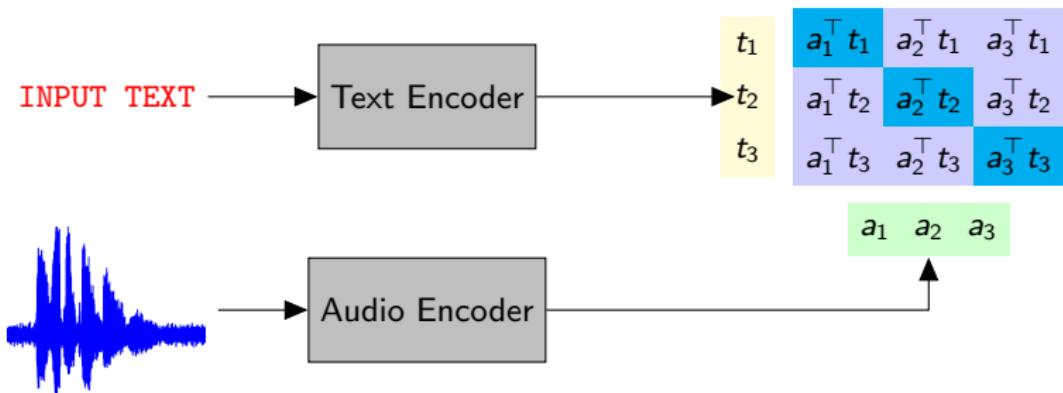
Cross-Modal Representation Learning



■ CLAP

- ▶ Training this model requires large number of paired data.

Cross-Modal Representation Learning



■ CLAP

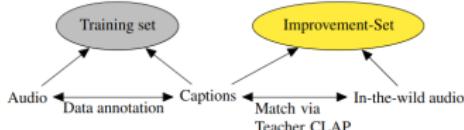
- ▶ Training this model requires large number of paired data.
- ▶ **WASPAA 2023 paper:** We published a paper where we improve the zero-shot classification performance using unpaired text and audio.

CLAP training objective

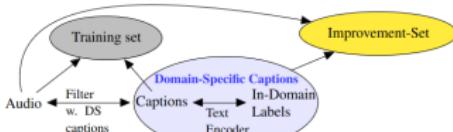
- Audio and text encoders $L_a = f_a(X_a)$, $L_t = f_t(X_t)$.
- The joint space: $a = \text{MLP}_a(L_a)$, $t = \text{MLP}_t(L_t)$
- We maximize the diagonal here $C = ta^\top$, through this loss function:

$$\mathcal{L}(C) = \frac{1}{2} \sum_{i=1}^N \left(\log(\text{Softmax}_t(C/\tau)_{i,i}) + \log(\text{Softmax}_a(C/\tau)_{i,i}) \right),$$

- We bootstrap CLAP by self-training. We explore different strategies for bootstrapping.



Model	Zero-Shot Evaluation Set		
	ESC-50	UrbanSound8K	TUT17
CLAP teacher	81.9 ± 0.9	74.8 ± 1.2	29.8 ± 1.3
SL	83.1 ± 1.2	73.9 ± 2.6	30.1 ± 2.1
DU	82.4 ± 1.4	73.9 ± 0.2	31.5 ± 1.0
DU+SL	83.0 ± 0.5	74.9 ± 1.4	29.9 ± 1.9
DS	78.8 ± 0.5	73.2 ± 1.1	29.8 ± 2.6
DS+SL	83.5 ± 0.6	75.5 ± 1.4	31.8 ± 2.6
ADS	84.2 ± 0.5	74.2 ± 2.1	32.5 ± 1.0
ADS+SL	85.1 ± 0.7	77.4 ± 0.6	36.0 ± 1.8



Model	Zero-Shot Evaluation Set		
	ESC-50	UrbanSound8K	TUT17
CLAP teacher (full-dataset)	81.9 ± 0.9	74.8 ± 1.2	29.8 ± 1.3
CLAP teacher (subset)	74.2 ± 1.3	73.5 ± 2.0	30.9 ± 1.6
DU + SL	78.9 ± 0.3	73.7 ± 1.3	28.8 ± 1.1
ADS + SL	81.3 ± 1.0	74.5 ± 0.7	31.3 ± 0.5

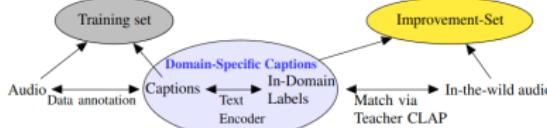


Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

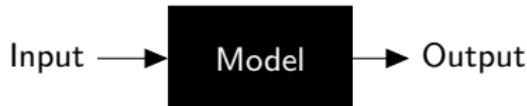
Interpretability

- Posthoc Explanations for Audio Classifiers

A little bonus

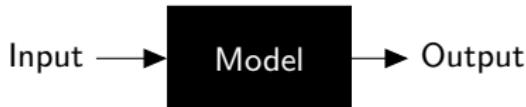
Explainable Machine Learning

- Black-box models

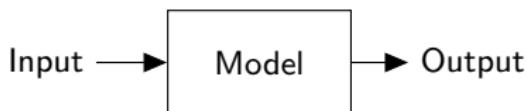


Explainable Machine Learning

- Black-box models

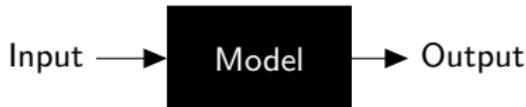


- Explainable Models

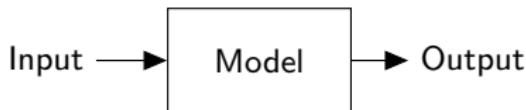


Explainable Machine Learning

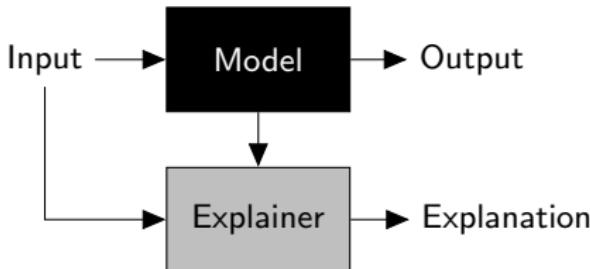
- Black-box models



- Explainable Models

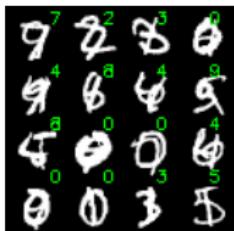


- Posthoc Explanations



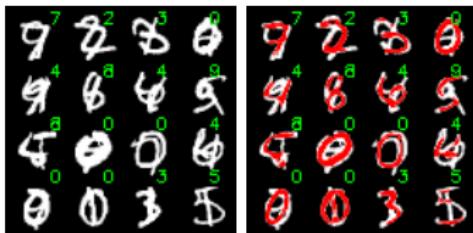
Neural Network Explanation

- Why does this particular input lead to that particular output?



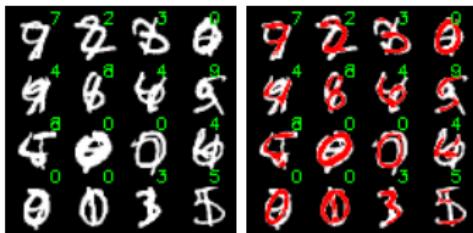
Neural Network Explanation

- Why does this particular input lead to that particular output?



Neural Network Explanation

- Why does this particular input lead to that particular output?



Recording, Classified as DOG

Neural Network Explanation

- Why does this particular input lead to that particular output?



Recording, Classified as DOG
Interpretation

- Posthoc Explanation vs Explainable Models: Posthoc Explanation produces explanations for already trained models. Explainable are by design so.

Listen-to-Interpret

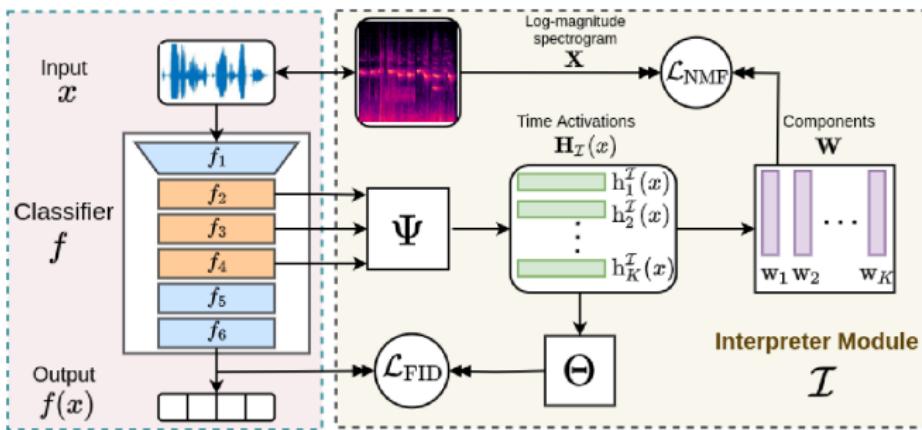


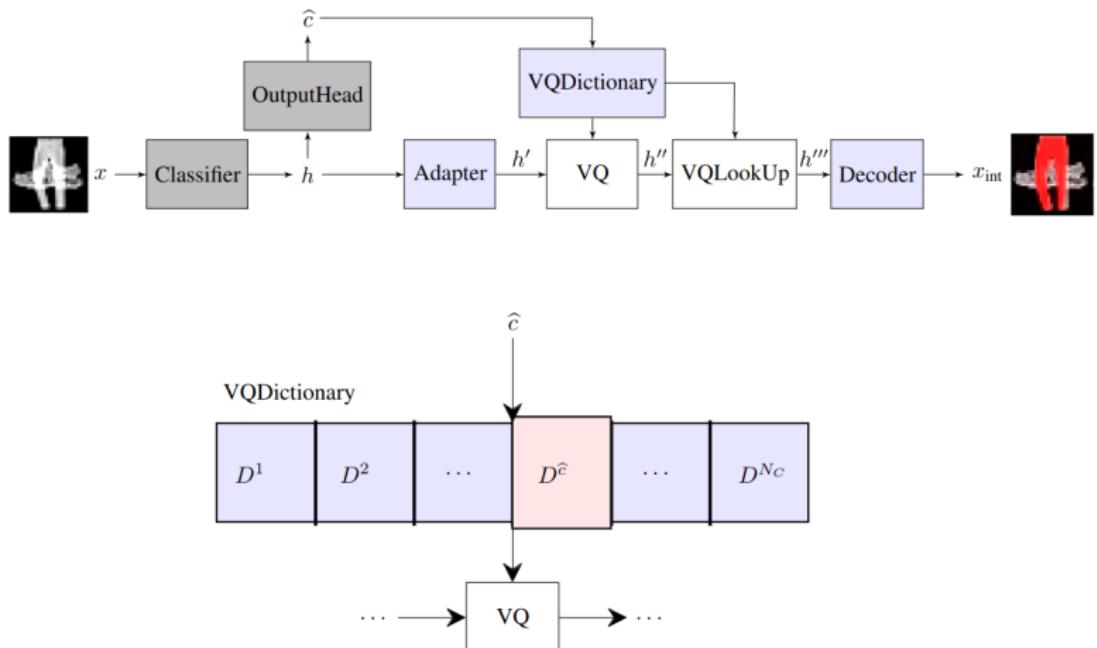
Image taken from <https://arxiv.org/pdf/2202.11479.pdf>

Posthoc Interpretation via Quantization

- We have developed a method that learns “high-level” concepts for each class in form of latent VQ dictionary, and then reconstructs the input using this VQ dictionary conditioned on the class information.

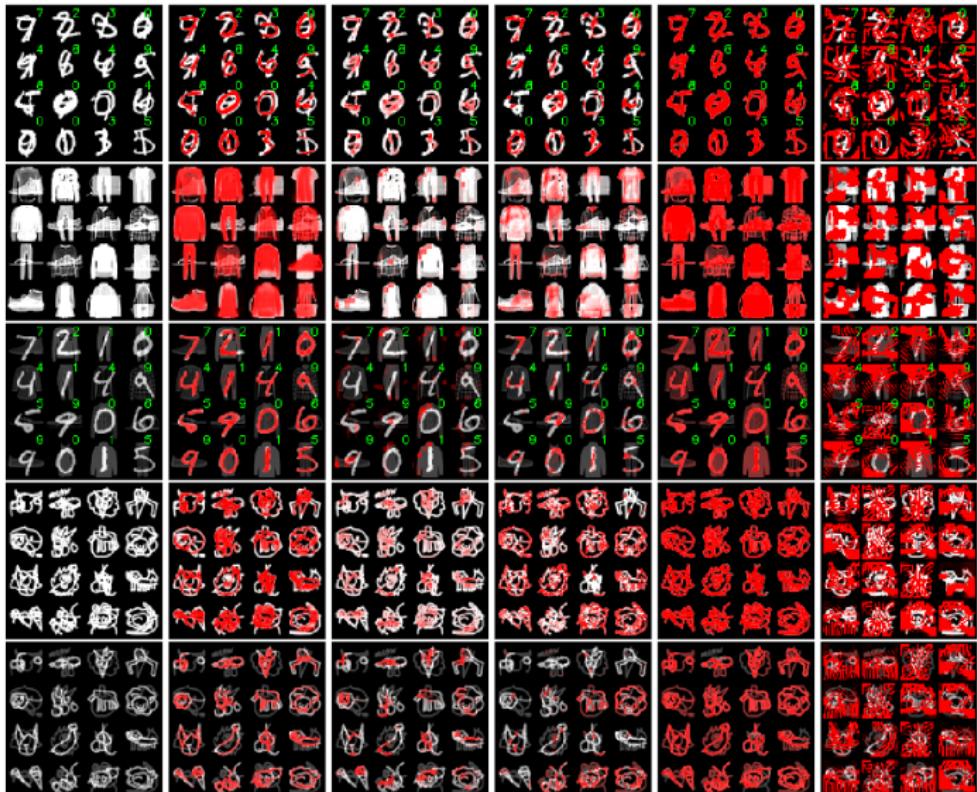


Posthoc Interpretation via Quantization



Above shows the inference time. In training, we only use images with single classes. NOT mixtures.

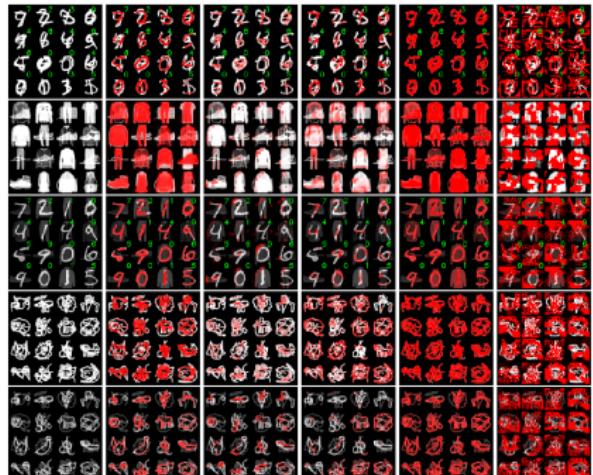
Qualitative Results on Images



Left-to-Right: Input, PIQ (ours), VIBI, L2I, LIME, FLINT

Mean-Opinion-Scores on Images

DATASET	METHOD	MOS (\uparrow)
MNIST B1 (CASE 1)	PIQ (OURS)	4.04 ± 0.48
	VIBI	1.77 ± 0.68
	L2I	2.4 ± 0.66
	FLINT	1 ± 0
	LIME	2 ± 1.34
MNIST B2 (CASE 1)	PIQ (OURS)	3.95 ± 0.72
	VIBI	1.86 ± 0.71
	L2I	1.86 ± 0.56
	FLINT	1.04 ± 0.21
	LIME	2.13 ± 1.21
FMNIST Mix (CASE 2)	PIQ (OURS)	4.87 ± 0.50
	VIBI	1.37 ± 0.50
	L2I	3.18 ± 0.91
	FLINT	1.12 ± 0.50
	LIME	1.37 ± 0.89
MNIST+FMN (CASE 3)	PIQ (OURS)	4.78 ± 0.43
	VIBI	1.14 ± 0.47
	L2I	2.18 ± 0.96
	FLINT	1.09 ± 0.47
	LIME	3.23 ± 0.72
QUICKDRAW1 (CASE4-I)	PIQ (OURS)	2.6 ± 1.67
	LIME	2.35 ± 1.46
QUICKDRAW2 (CASE4-II)	PIQ (OURS)	3.55 ± 1.0
	LIME	3 ± 1,38



Quantitative Results on Images

Dataset	MNIST			FMNIST		
Metric	Fidelity-In (\uparrow)	Faithfulness (\uparrow)	FID (\downarrow)	Fidelity-In (\uparrow)	Faithfulness (\uparrow)	FID (\downarrow)
PIQ (ours)	98.03 ± 0.05	0.588 ± 0.00021	0.029 ± 0.0004	81.3 ± 0.2	0.773 ± 0.004	0.030 ± 0.0004
VIBI	73.90 ± 16.08	0.369 ± 0.002	0.710 ± 0.962	42.4 ± 17.8	0.578 ± 0.073	0.395 ± 0.104
L2I	96.56 ± 2.66	0.453 ± 0.002	0.160 ± 0.010	68.3 ± 1.5	0.343 ± 0.011	0.188 ± 0.011
FLINT	10.9	0.361	0.677	15.37	-0.097	0.482

Dataset	Quickdraw		
Metric	Fidelity-In (\uparrow)	Faithfulness (\uparrow)	FID (\downarrow)
PIQ (ours)	60.89 ± 0.60	0.675 ± 0.005	0.034 ± 0.0001
VIBI	26.36 ± 3.01	0.341 ± 0.031	0.388 ± 0.032
L2I	25.97 ± 0.82	0.340 ± 0.031	0.397 ± 0.020
FLINT	15.62	-0.057	0.672

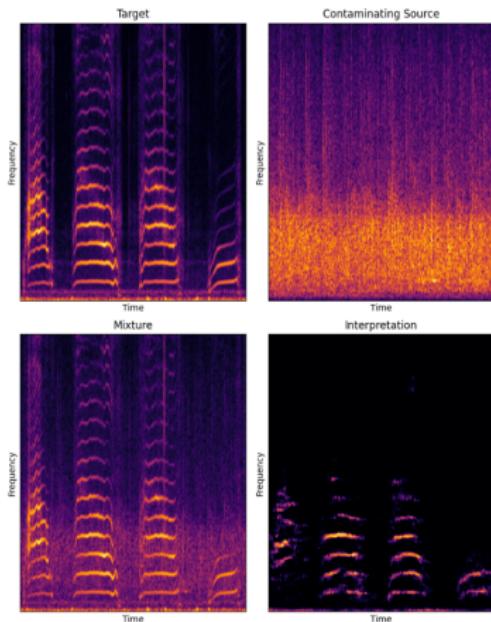
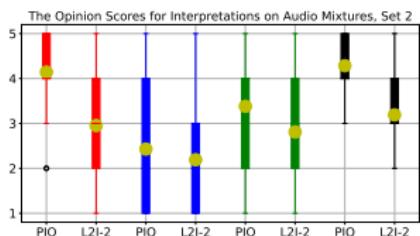
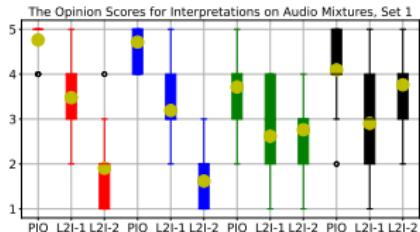
■ Input Fidelity

$$\text{FID-I} = \frac{1}{N} \sum_{n=1}^N \left[\arg \max_c f_c(x_n) = \arg \max_c f_c(x_{\text{int},n}) \right],$$

■ Faithfulness

$$\text{Faithfulness} = f_c(x) - f_c(x - x_{\text{int}}),$$

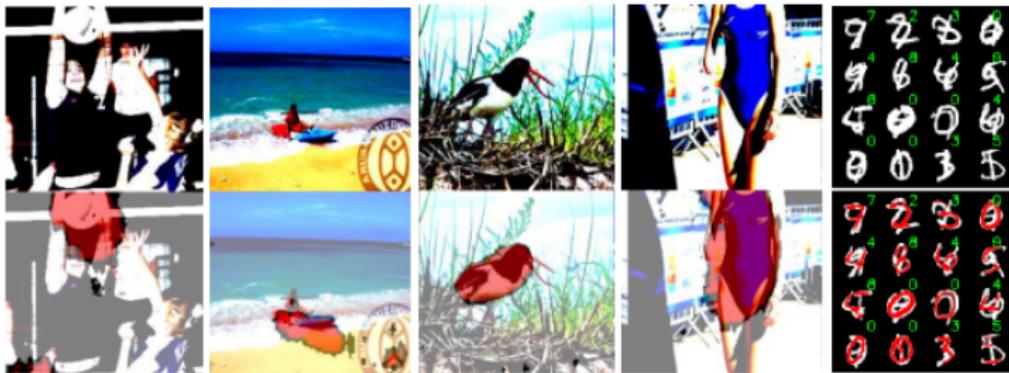
Mean-Opinion Scores on Audio



Click for More Example Results

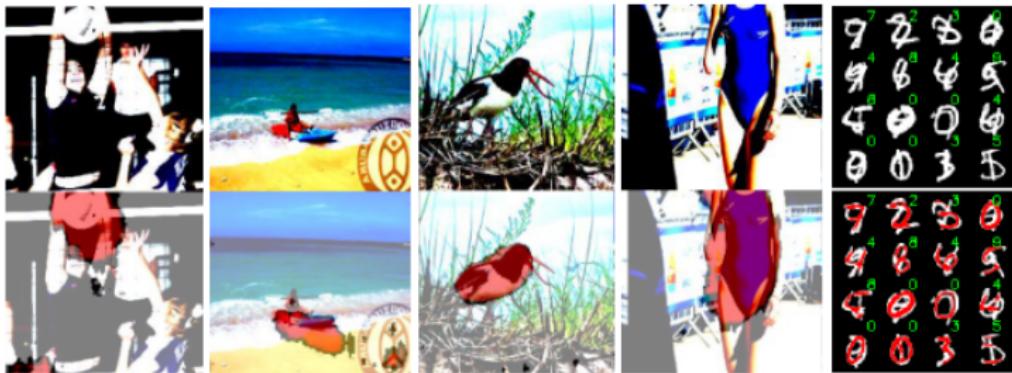
Explanations

- Saliency maps are commonly used in computer vision for producing explanations.



Explanations

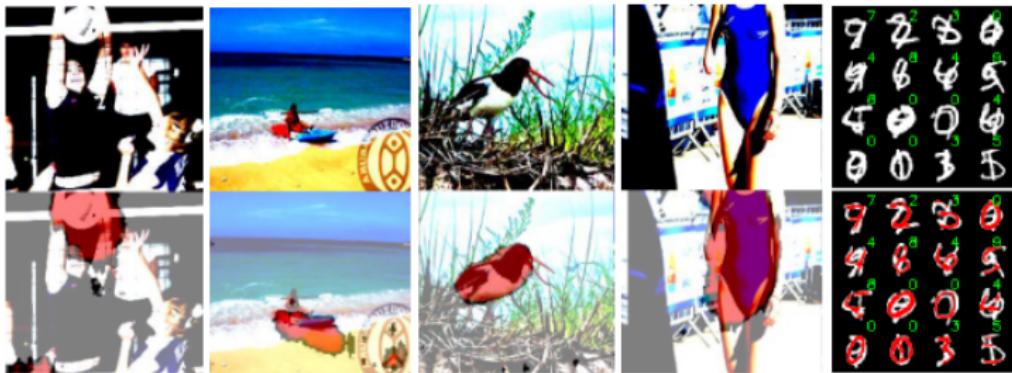
- Saliency maps are commonly used in computer vision for producing explanations.



- The explanations should **faithfully** follow the original model.

Explanations

- Saliency maps are commonly used in computer vision for producing explanations.

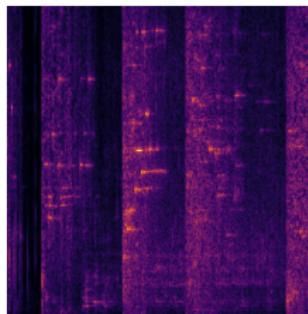


- The explanations should **faithfully** follow the original model.
- Faithful** and **understandable** explanations are important for domains where decisions are critical!

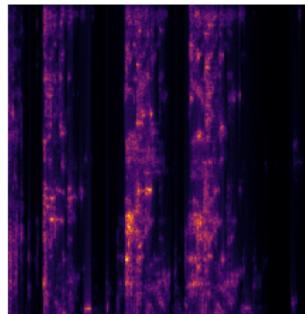
Contributions

- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

Input Audio

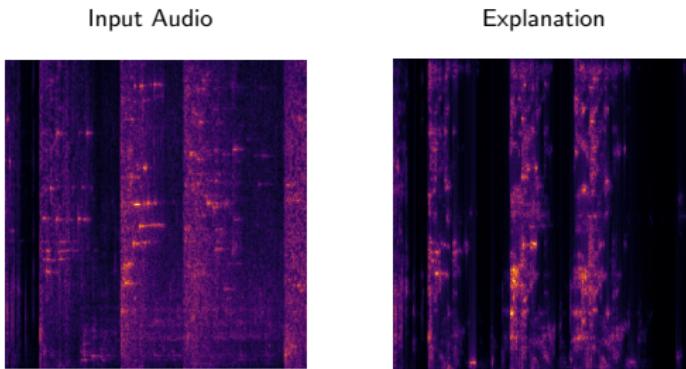


Explanation



Contributions

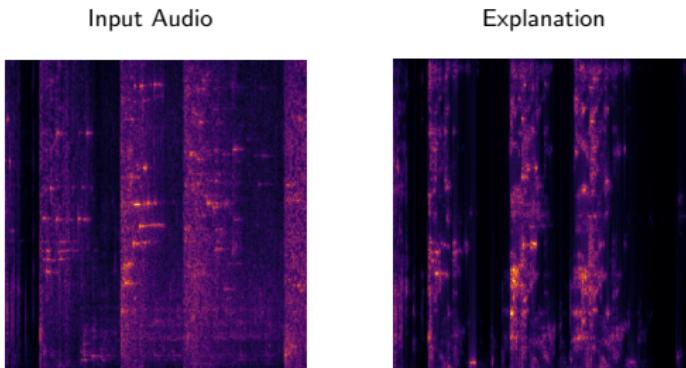
- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.



- Our method is agnostic to classifier input domain, and generates **listenable** explanations.

Contributions

- We develop an **understandable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.



- Our method is agnostic to classifier input domain, and generates **listenable** explanations.
- We propose a fine-tuning strategy that improves understandability/faithfulness trade-off.

Considerations

We would like to obtain

Considerations

We would like to obtain

- Faithful,

Considerations

We would like to obtain

- Faithful,
- Listenable,

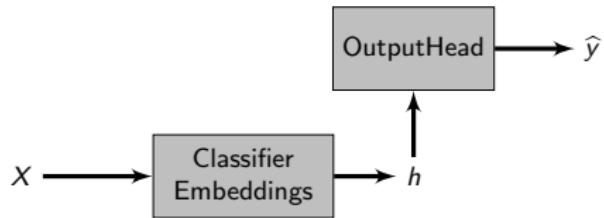
Considerations

We would like to obtain

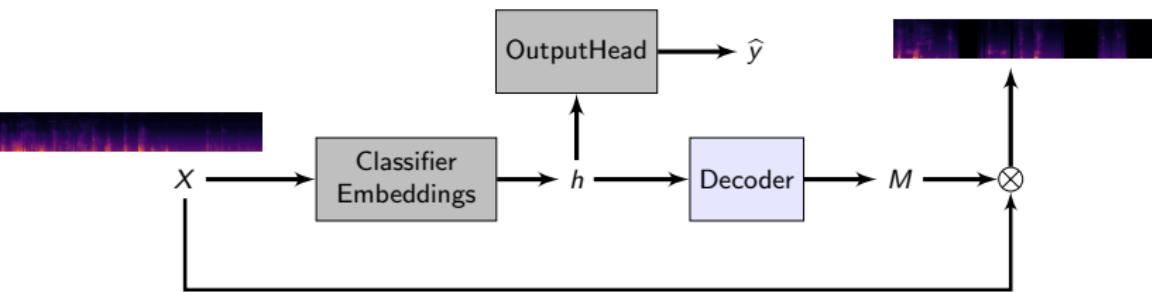
- Faithful,
- Listenable,
- Understandable

Posthoc Explanations for Audio Classifiers

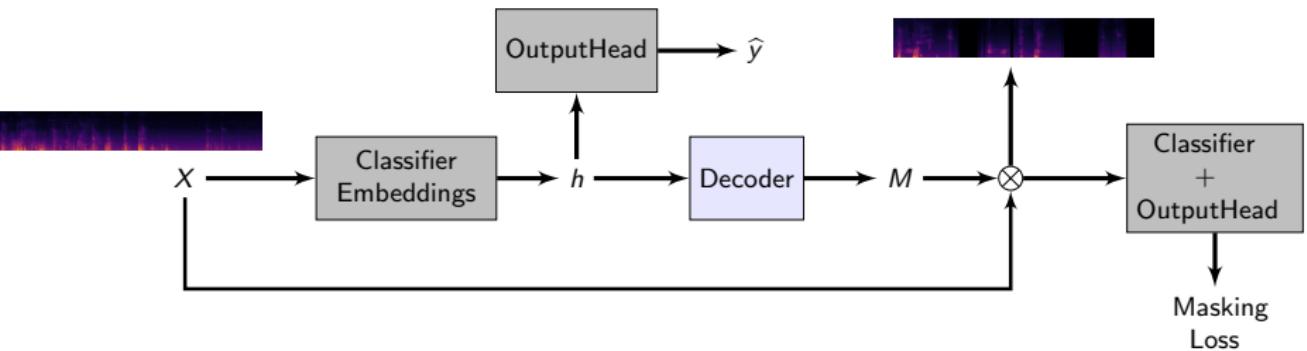
Listenable Maps for Audio Classifiers (ICML 2024 - Oral)



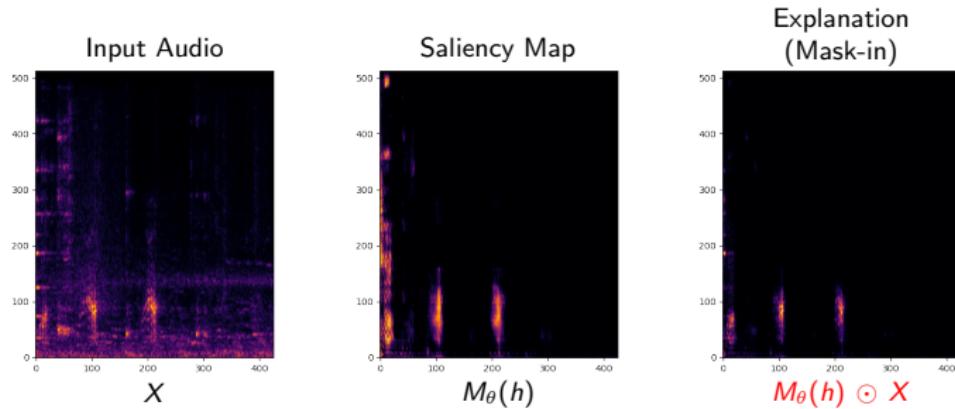
Listenable Maps for Audio Classifiers (ICML 2024 - Oral)



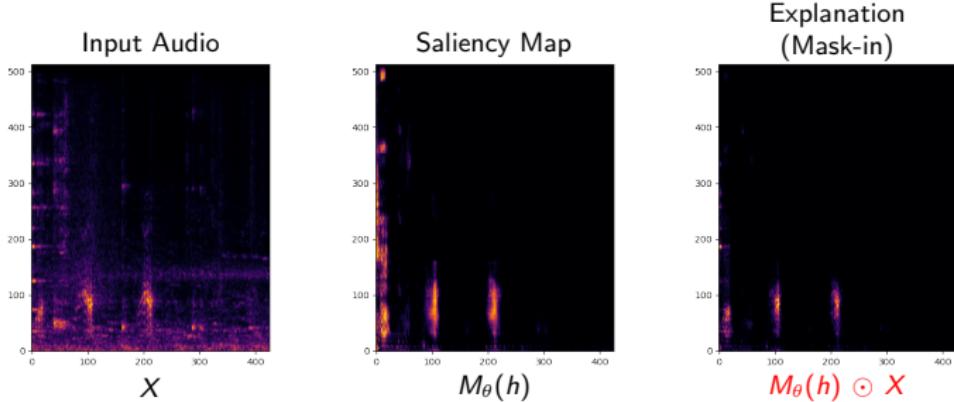
Listenable Maps for Audio Classifiers (ICML 2024 - Oral)



Optimization objective



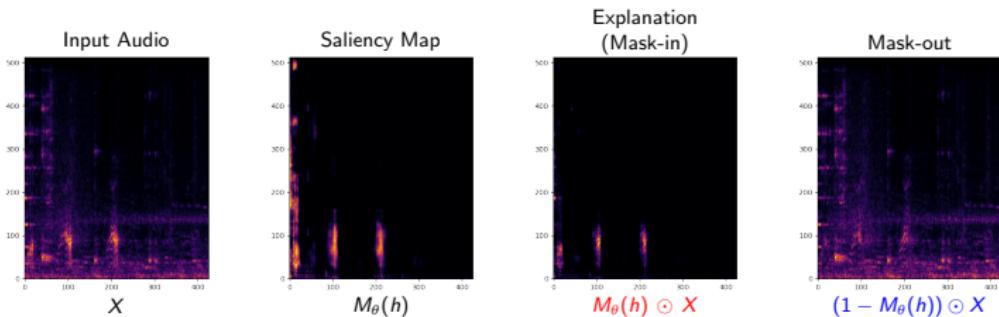
Optimization objective



$$\min_{\theta} \overbrace{\lambda_{in} \mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \hat{y})}^{\text{Mask-in}}$$

Maximizes the classifier agreement between the input and the explanation.

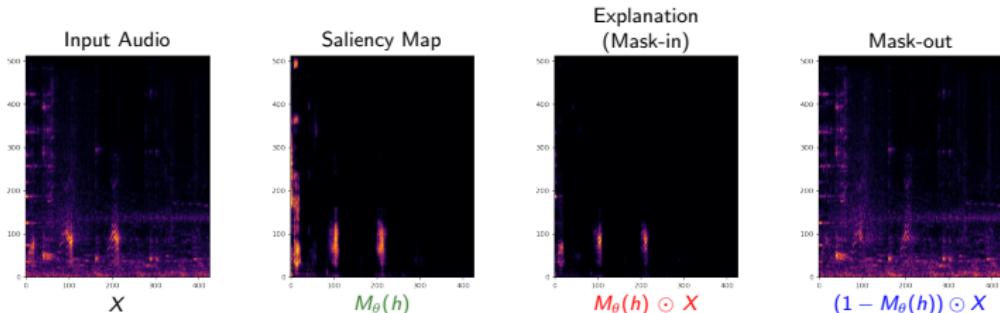
Optimization objective



$$\min_{\theta} \underbrace{\lambda_{in} \mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \hat{y})}_{\text{Mask-in}} - \underbrace{\lambda_{out} \mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \hat{y})}_{\text{Mask-out}}$$

Minimizes the classifier agreement of what is not in the explanation and the input.

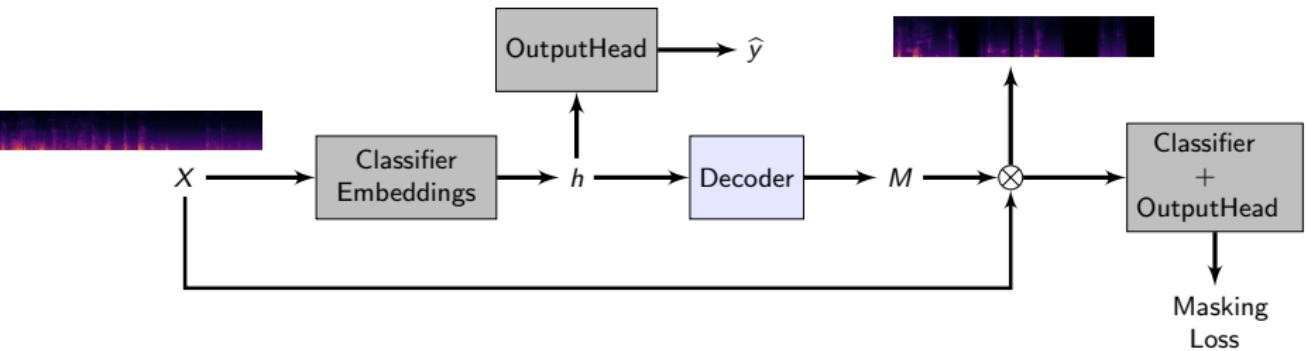
Optimization objective



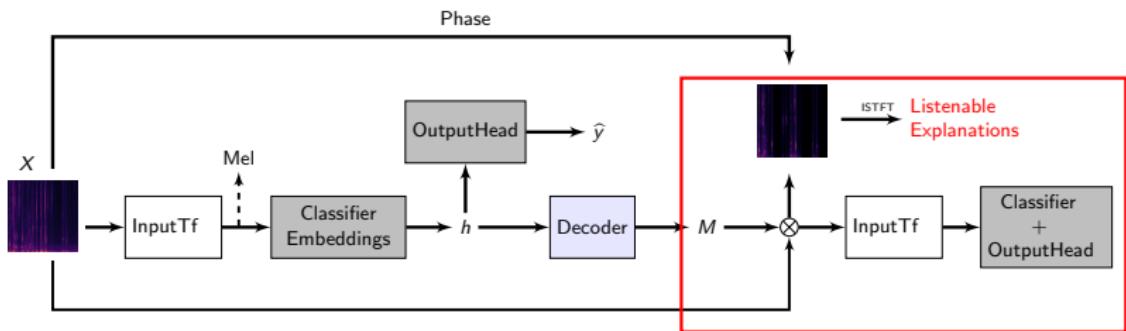
$$\min_{\theta} \underbrace{\lambda_{in} \mathcal{L}_{in}(\log f(M_\theta(h) \odot X), \hat{y})}_{\text{Mask-in}} - \underbrace{\lambda_{out} \mathcal{L}_{out}(\log f((1 - M_\theta(h)) \odot X), \hat{y})}_{\text{Mask-out}} + \underbrace{|M_\theta(h)|}_{\text{Mask Reg}}$$

Avoids trivial solutions.

Producing Listenable Explanations



Producing Listenable Explanations



$$\text{Listenable Explanation} = \text{ISTFT} \left((M_\theta(h) \odot X) e^{jX_{\text{phase}}} \right)$$

Measuring faithfulness and understandability

- **Faithfulness:** Measures importance of explanations for classifier decisions
 - ▶ L2I-Faithfulness
 - ▶ Average-Increase
 - ▶ Average-Gain
 - ▶ Average-Drop
 - ▶ Input Fidelity
- **Structural metrics:** Measures the understandability of the explanations
 - ▶ Sparseness
 - ▶ Complexity

Understandability

$$\begin{aligned} & \min_{\theta} \lambda_{in} \mathcal{L}_{in}(\log f(M_{\theta}(h) \odot X), \hat{y}) \\ & - \lambda_{out} \mathcal{L}_{out}(\log f((1 - M_{\theta}(h)) \odot X), \hat{y}) + \overbrace{|M_{\theta}(h)|}^{Regularizer} \end{aligned}$$

Understandability

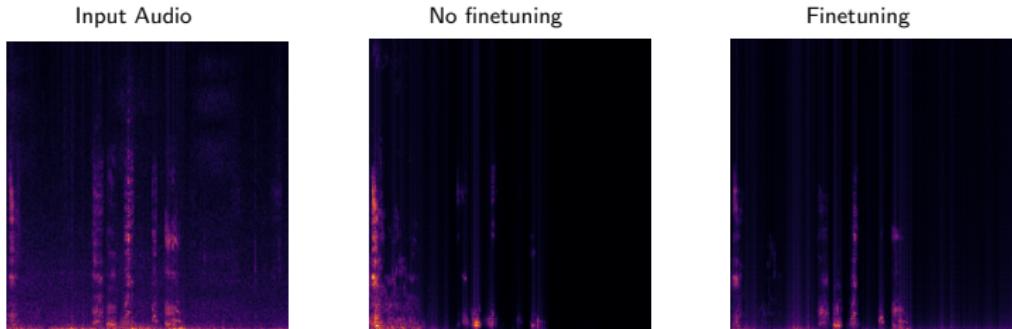
$$\begin{aligned} \min_{\theta} & \lambda_{in} \mathcal{L}_{in}(\log f(M_{\theta}(h) \odot X), \hat{y}) - \lambda_{out} \mathcal{L}_{out}(\log f((1 - M_{\theta}(h)) \odot X), \hat{y}) \\ & + \lambda_s \underbrace{\|M_{\theta}(h)\|_1}_{L_1} + \lambda_g \underbrace{\|M_{\theta}(h) \odot X - X_{clean}\|}_{Finetuning} \end{aligned}$$

- L_1 : Avoids trivial solutions (e.g. all 1s).
- *Finetuning*: Improves Understandability.
 - ▶ Used in a second stage, selectively.

Understandability

$$\begin{aligned} \min_{\theta} & \lambda_{in} \mathcal{L}_{in}(\log f(M_{\theta}(h) \odot X), \hat{y}) - \lambda_{out} \mathcal{L}_{out}(\log f((1 - M_{\theta}(h)) \odot X), \hat{y}) \\ & + \lambda_s \underbrace{\|M_{\theta}(h)\|_1}_{L_1} + \lambda_g \underbrace{\|M_{\theta}(h) \odot X - X_{clean}\|}_{Finetuning} \end{aligned}$$

- L_1 : Avoids trivial solutions (e.g. all 1s).
- *Finetuning*: Improves Understandability.
 - ▶ Used in a second stage, selectively.



Experiments

- We produce explanations for classifiers trained on Sound Event Classification Datasets (**ESC50**, **US8k**).

Experiments

- We produce explanations for classifiers trained on Sound Event Classification Datasets (**ESC50**, **US8k**).
- We examine explanations on In-Domain (**ID**) and Out-of-Domain (**OOD**) cases.
 - ▶ ID: Plain datasets with data augmentation
 - ▶ OOD: Mixtures with different contaminating sources

Quantitative Results - ID

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT \rightarrow MeI)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
	L-MAC (ours)	36.25	1.15	23.50	0.20	0.42	0.47	4.71
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	0.21	0.41	0.43	5.20
Not Listenable (MeI)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	36.13	1.28	21.15	0.23	0.42	0.32	4.71

Quantitative Results - ID

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT \rightarrow MeI)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
L-MAC (ours)		36.25	1.15	23.50	0.20	0.42	0.47	4.71
L-MAC, FT, $\lambda_g = 4$ (ours)		32.37	1.98	18.74	0.21	0.41	0.43	5.20
Not Listenable (MeI)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
L-MAC (ours) FT, $\lambda_g = 4$		36.13	1.28	21.15	0.23	0.42	0.32	4.71

Quantitative Results - ID

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT→Mel)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
L-MAC (ours)		36.25	1.15	23.50	0.20	0.42	0.47	4.71
L-MAC, FT, $\lambda_g = 4$ (ours)		32.37	1.98	18.74	0.21	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
L-MAC (ours) FT, $\lambda_g = 4$		36.13	1.28	21.15	0.23	0.42	0.32	4.71

- Finetuning does not harm faithfulness significantly.
- Generating listenable explanations does not decrease the alignment with the classifier.

Quantitative Results - ID

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT \rightarrow MeI)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
Not Listenable (MeI)	L-MAC (ours)	36.25	1.15	23.50	0.20	0.42	0.47	4.71
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	0.21	0.41	0.43	5.20
	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	36.13	1.28	21.15	0.23	0.42	0.32	4.71

- Finetuning does not harm faithfulness significantly.
- Generating listenable explanations does not decrease the alignment with the classifier.
- We have comparable structural metrics.

Quantitative Results - OOD (Audio Mixtures)

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT \rightarrow Mel)	Saliency	0.62	31.73	0.07	0.06	0.12	0.76	11.06
	Smoothgrad	0.12	31.84	0.00	0.06	0.13	0.83	10.66
	IG	0.37	31.15	0.03	0.12	0.26	0.87	10.22
	L2I	5.00	25.65	1.00	0.20	0.35	0.52	10.99
	GradCAM	14.12	17.62	7.46	0.25	0.00	0.91	9.66
	Guided GradCAM	0.00	31.74	0.00	0.07	0.11	0.89	10.24
	Guided Backprop	0.63	31.73	0.07	0.06	0.11	0.76	11.06
	SHAP	0.00	31.81	0.00	0.07	0.14	0.84	10.58
	L-MAC (ours)	60.63	4.82	35.85	0.39	0.81	0.94	9.61
	L-MAC FT, $\lambda_g = 4$ (ours)	50.75	6.73	26.00	0.39	0.78	0.84	10.51
Not Listenable (Mel)	Saliency	0.38	31.64	0.01	0.15	0.12	0.77	9.17
	Smoothgrad	0.25	31.66	0.01	0.14	0.11	0.79	9.03
	IG	0.12	31.52	0.01	0.19	0.19	0.84	8.62
	GradCAM	19.88	18.85	4.67	0.34	0.69	0.66	9.49
	Guided GradCAM	0.00	31.68	0	0.14	0.12	0.89	10.24
	Guided Backprop	0.38	31.64	0.01	0.15	0.12	0.77	9.16
	SHAP	0.25	31.60	0.00	0.17	0.15	0.82	8.81
	L-MAC (ours)	60.25	4.84	34.72	0.44	0.80	0.90	8.29
	L-MAC - FT, $\lambda_g = 4$ (ours)	60.75	4.84	29.34	0.44	0.83	0.64	9.38

Quantitative Results - OOD (Audio Mixtures)

Metric	AI (\uparrow)	AD (\downarrow)	AG (\uparrow)	FF (\uparrow)	Fid-In (\uparrow)	SPS (\uparrow)	COMP (\downarrow)	
Listenable (STFT \rightarrow Mel)	Saliency	0.62	31.73	0.07	0.06	0.12	0.76	11.06
	Smoothgrad	0.12	31.84	0.00	0.06	0.13	0.83	10.66
	IG	0.37	31.15	0.03	0.12	0.26	0.87	10.22
	L2I	5.00	25.65	1.00	0.20	0.35	0.52	10.99
	GradCAM	14.12	17.62	7.46	0.25	0.00	0.91	9.66
	Guided GradCAM	0.00	31.74	0.00	0.07	0.11	0.89	10.24
	Guided Backprop	0.63	31.73	0.07	0.06	0.11	0.76	11.06
	SHAP	0.00	31.81	0.00	0.07	0.14	0.84	10.58
	L-MAC (ours)	60.63	4.82	35.85	0.39	0.81	0.94	9.61
	L-MAC FT, $\lambda_g = 4$ (ours)	50.75	6.73	26.00	0.39	0.78	0.84	10.51
Not Listenable (Mel)	Saliency	0.38	31.64	0.01	0.15	0.12	0.77	9.17
	Smoothgrad	0.25	31.66	0.01	0.14	0.11	0.79	9.03
	IG	0.12	31.52	0.01	0.19	0.19	0.84	8.62
	GradCAM	19.88	18.85	4.67	0.34	0.69	0.66	9.49
	Guided GradCAM	0.00	31.68	0	0.14	0.12	0.89	10.24
	Guided Backprop	0.38	31.64	0.01	0.15	0.12	0.77	9.16
	SHAP	0.25	31.60	0.00	0.17	0.15	0.82	8.81
	L-MAC (ours)	60.25	4.84	34.72	0.44	0.80	0.90	8.29
	L-MAC - FT, $\lambda_g = 4$ (ours)	60.75	4.84	29.34	0.44	0.83	0.64	9.38

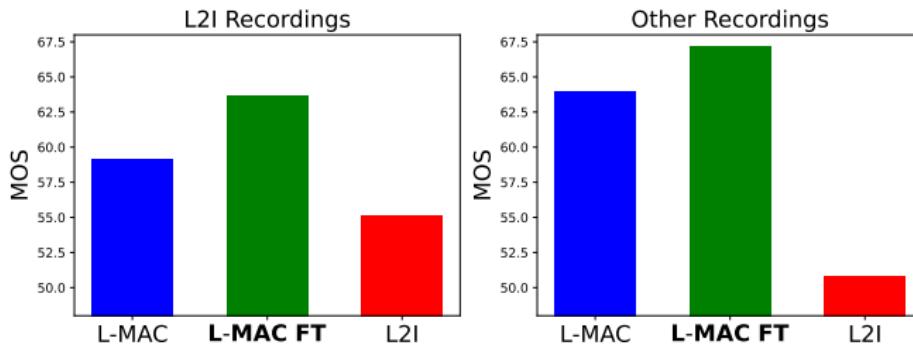
We observe the same outcome on US8k as well! (See the appendix)

User Study

1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.

User Study

1. How well does the explanation correspond to the part of the input audio associated with the given class?
2. While evaluating, please pay attention to audio quality also.



- Recording 1:
 - ▶ L-MAC
 - ▶ L2I [NeurIPS'22]

- OOD (Speech):
 - ▶ L-MAC
 - ▶ L2I [NeurIPS'22]

Conclusions

- We proposed a SOTA **posthoc explanation** method for audio classifiers.
- Our method is agnostic to classifier input representation.
- Our method provides **understandable, listenable and faithful** explanations both in ID and OOD cases.
- Our code is available in SpeechBrain.

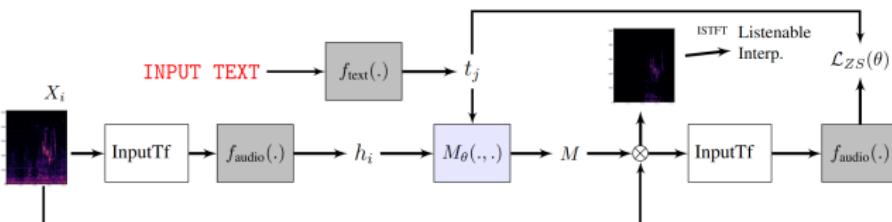


Conclusions

- We proposed a SOTA **posthoc explanation** method for audio classifiers.
- Our method is agnostic to classifier input representation.
- Our method provides **understandable, listenable and faithful** explanations both in ID and OOD cases.
- Our code is available in SpeechBrain.



Follow-up Work: LMAC ZS: Listenable Maps for Zero-Shot Classifiers
(NeurIPS 2024)

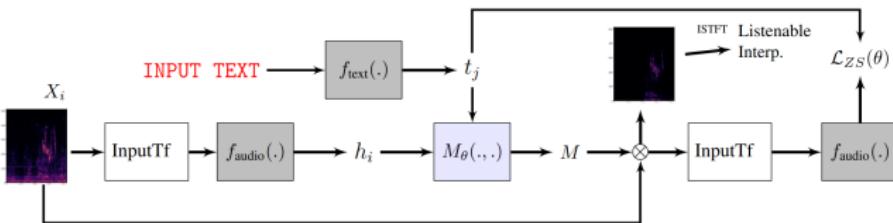


Thanks! Questions?

- We proposed a SOTA **posthoc explanation** method for audio classifiers.
- Our method is agnostic to classifier input representation.
- Our method provides **understandable, listenable** and **faithful** explanations both in ID and OOD cases.
- Our code is available in SpeechBrain.



Follow-up Work: LMAC ZS: Listenable Maps for Zero-Shot Classifiers



Ongoing projects in Interpretability

- Exploring Understandability / Faithfulness Tradeoffs
- Interpretable Detection of Fake vs Real Audio
- Listenable Recommendation Systems
- Interpretable Baby Cry Analysis

Table of Contents

Speech Recognition

- RNN Based ASR
- Transformer ASR
- CTC Based ASR

Text-to-Speech

Speech Separation / Enhancement

- Problem Definition
- Source Separation Methods
- SepFormer

Moving towards real-life source separation

- Data collection
- Performance estimation
- Evaluating the SI-SNR Estimator
- User Study and Further Evaluation

Cross-Modal Representation Learning

Interpretability

- Posthoc Explanations for Audio Classifiers

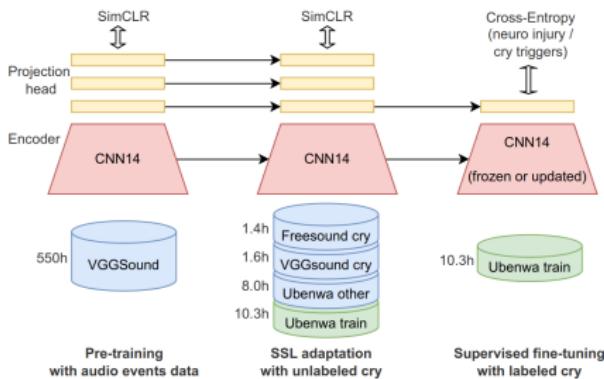
A little bonus

ML for Infant Cry Analysis

- Collaboration with UbenwaAI, a Mila based startup.
- The goal is to develop machine learning methods for Infant Cry Analysis.
(Representation Learning / Interpretability, Efficient Learning, ...)
Self-Supervised Learning for Cry Analysis,
ICASSP 2023 workshop paper in the Self-Supervision in Audio, Speech
and Beyond Workshop.

SELF-SUPERVISED LEARNING FOR INFANT CRY ANALYSIS

Arsenii Gorin, Cem Subakan[✉], Sajjad Abdoli*, Junhao Wang*, Samantha Latremouille*, Charles Onu[✉]*



The paper <https://arxiv.org/abs/2305.01578>

Baby Identification Challenge: CryCeleb



ubenwa.ai

CryCeleb

A machine learning challenge for speaker verification using infant cry sounds.

MAY 1 - JUNE 30

REGISTER HERE: bit.ly/crychallenge



Same? → Yes/No

Powered by:  Ubenwa ×  SpeechBrain

The CryCeleb2023 Challenge!

SpeechBrain



SpeechBrain

An Open-Source Conversational AI Toolkit

[Get Started](#) [GitHub](#)   [Star](#) 5,739

 [SpeechBrain](#) 126 member

The call for Sponsors 2023 is open!



Recap

- ASR: We have covered different modern ASR techniques. / On a couvert différentes techniques de reconnaissance vocale moderne.
 - ▶ Encoder-decoder with RNN, transformer, CTC, combination.
- TTS: Conceptually similar sequence-to-sequence techniques as ASR. / Des techniques qui ressemblent en concepte à ceux de ASR.
- Source separation/enhancement / Séparation de sources, amélioration du son
 - ▶ End-to-end separation, going towards real-life mixtures
- Learning text-audio representations, zero-shot audio classification
- Interpretability for Audio

Suggested Reading

- ASR
 - ▶ Book on Speech Processing:
<https://web.stanford.edu/~jurafsky/slp3/>
 - ▶ Listen-Attend-Spell: <https://arxiv.org/pdf/1508.01211.pdf>
 - ▶ On CTC: https://www.cs.toronto.edu/~graves/icml_2006.pdf,
<https://distill.pub/2017/ctc/>
- TTS
 - ▶ Tacotron: <https://arxiv.org/pdf/1703.10135.pdf>
 - ▶ Transformer TTS: <https://arxiv.org/pdf/1809.08895.pdf>
 - ▶ Tacotron2: <https://arxiv.org/pdf/1712.05884.pdf>
 - ▶ ECAPA-TDNN: <https://arxiv.org/pdf/2005.07143.pdf>
- Speech Separation
 - ▶ Sepformer: <https://arxiv.org/abs/2010.13154>,
<https://arxiv.org/abs/2202.02884>
 - ▶ REAL-M: <https://arxiv.org/pdf/2110.10812.pdf>
- Cross Model Representations
 - ▶ CLAP: <https://arxiv.org/pdf/2206.04769.pdf>
 - ▶ UI-CLAP: <https://arxiv.org/abs/2305.01864>
- Interpretability for Audio
 - ▶ L2I: <https://arxiv.org/pdf/2202.11479v2.pdf>
 - ▶ L-MAC: <https://arxiv.org/abs/2403.13086>

Next week

- It's your turn now. Don't forget to sign-up: https://docs.google.com/spreadsheets/d/1jWwOzGoaixzYXJwDCsK0LaF_z8J9-jY1be-d9WzG9GY/edit?usp=sharing
 - ▶ Le show est à vous maintenant!

Thanks!/Merci!