DEEP LEARNING FOR SPEECH ENHANCEMENT APPLIED TO NON-CONVENTIONAL SOUND CAPTURE DEVICES

Starting date: December 1, 2021 (9 months ago)

Funding: ANR (50% ISL/50% Cnam) - AHEAD project

Artificial Intelligence for Health, Physical Models, Transportation and Defense

PhD student: Julien HAURET[☆]

Director: Éric BAVU[☆]

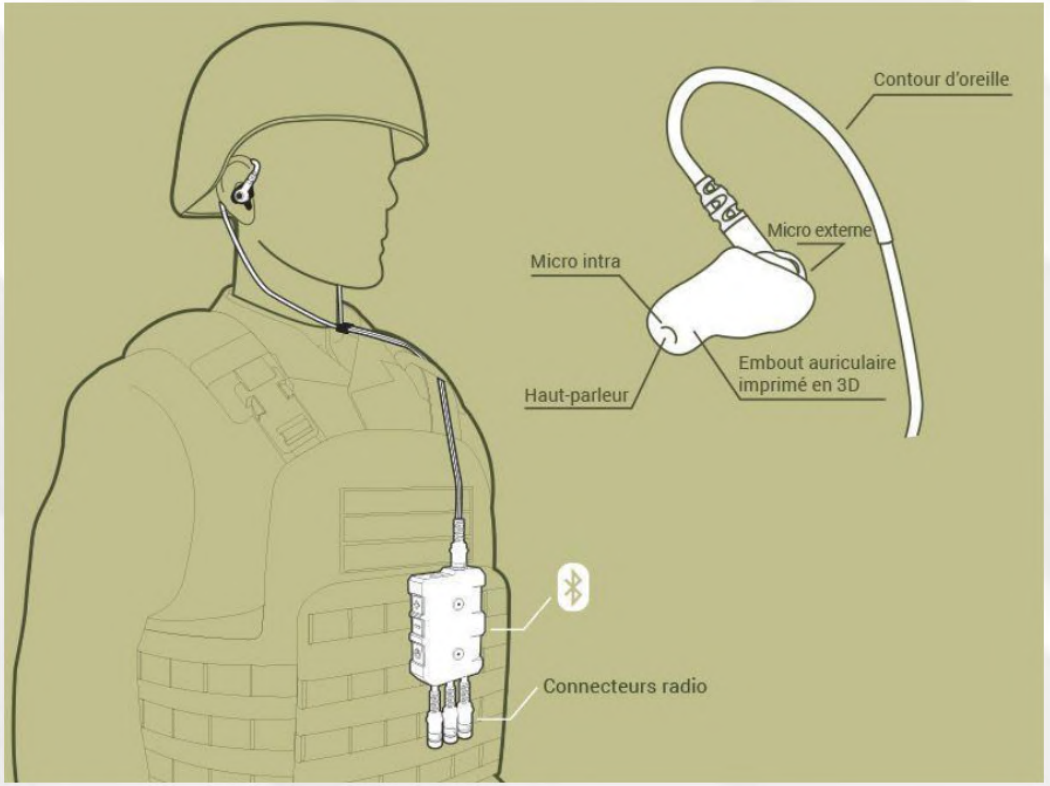
Supervisors: Thomas JOUBAUD[†]**, Véronique ZIMPFER**[†]

☆ : Structural Mechanics and Coupled Systems Laboratory, Conservatoire National des Arts et Metiers

† :Department of Acoustics and Soldier Protection, French-German Research Institute of Saint-Louis

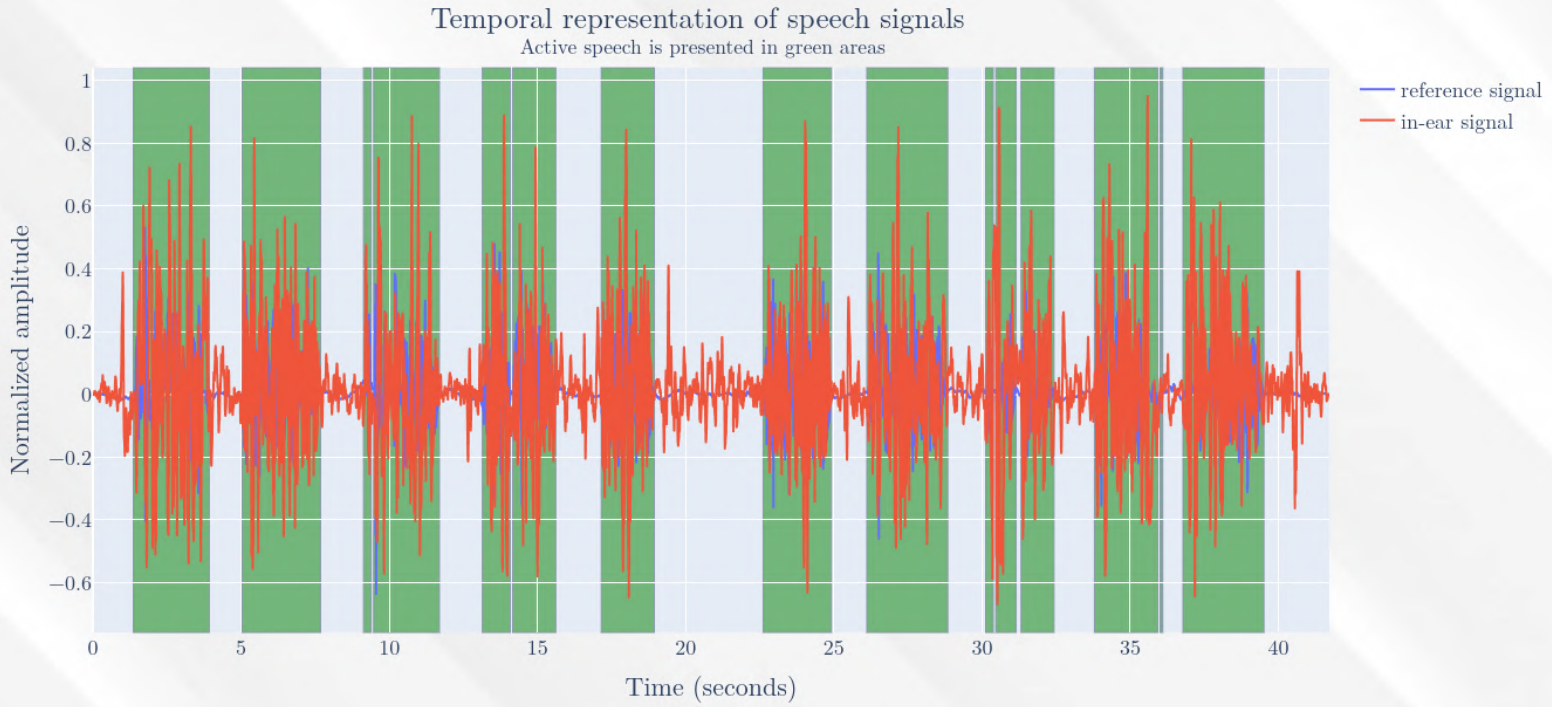
CONTEXT

BANG PROTOTYPE



Bouchon Auriculaire de Nouvelle Génération

TEMPORAL REPRESENTATION



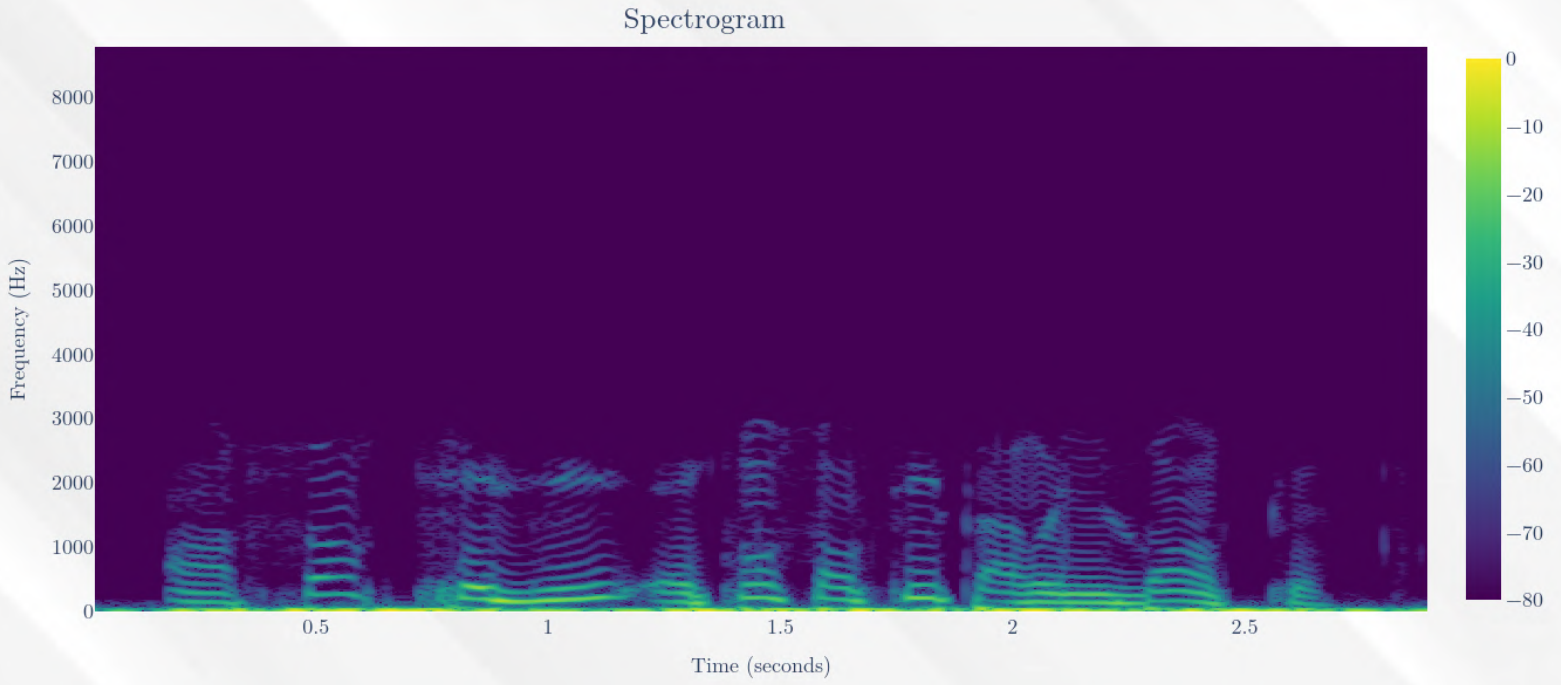
0:00 / 0:41

0:00 / 0:41

SPECTROGRAM

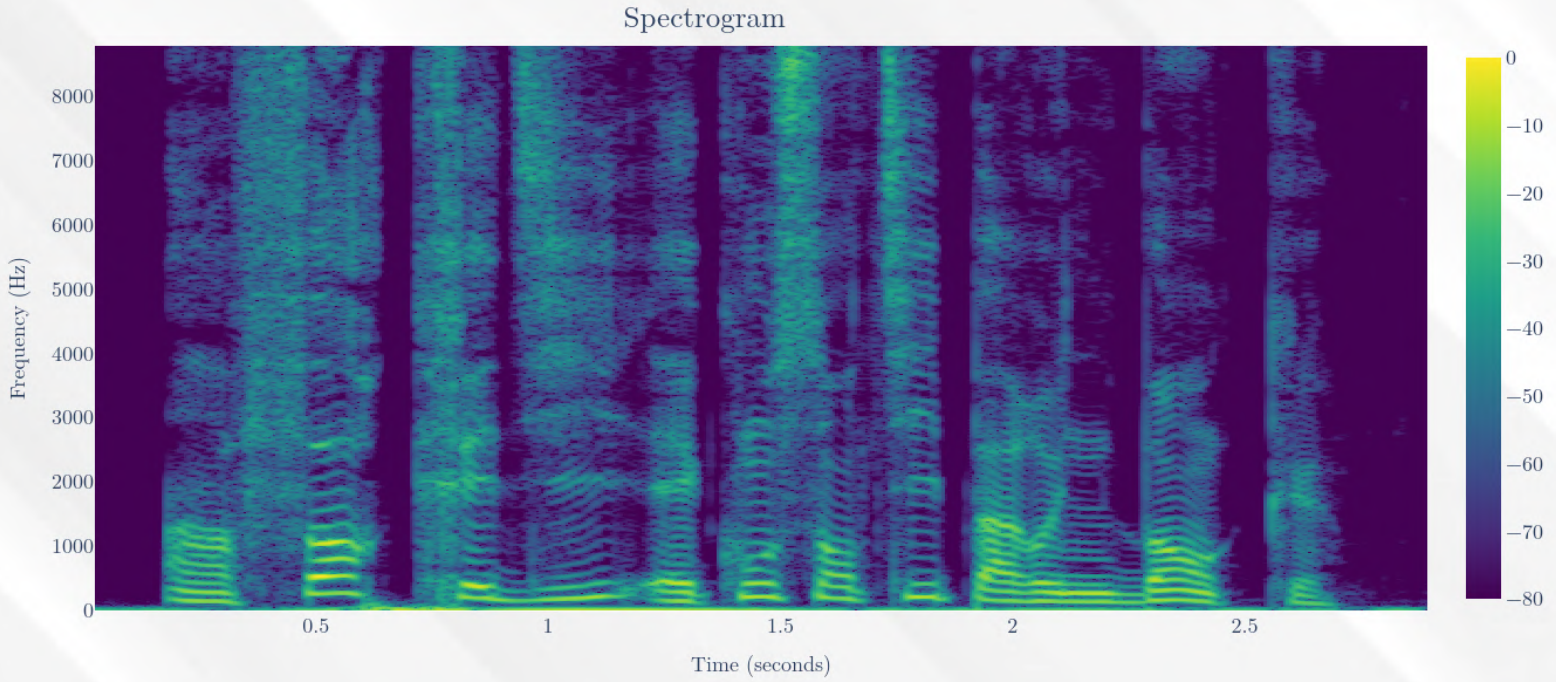
le cnam

IN-EAR SIGNAL



0:00 / 0:02

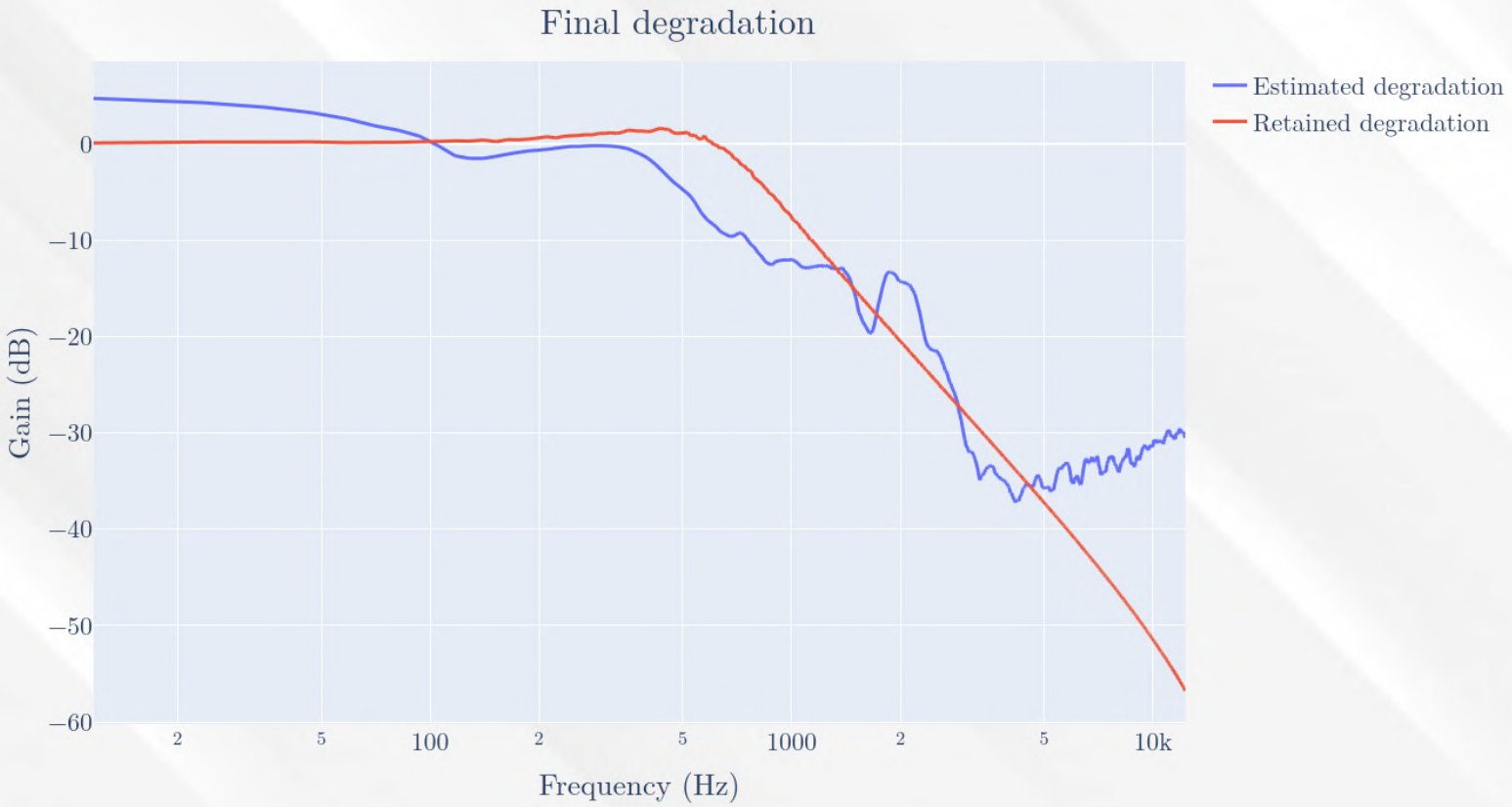
REFERENCE SIGNAL



0:00 / 0:02

le cnam

SOURCE FILTER MODEL



SUMMARY

GOAL

Improving speech intelligibility with in-the-ear transducers

CONSTRAINTS

- Light hardware
- Real-time processing
- Robustness to speaker's indentity

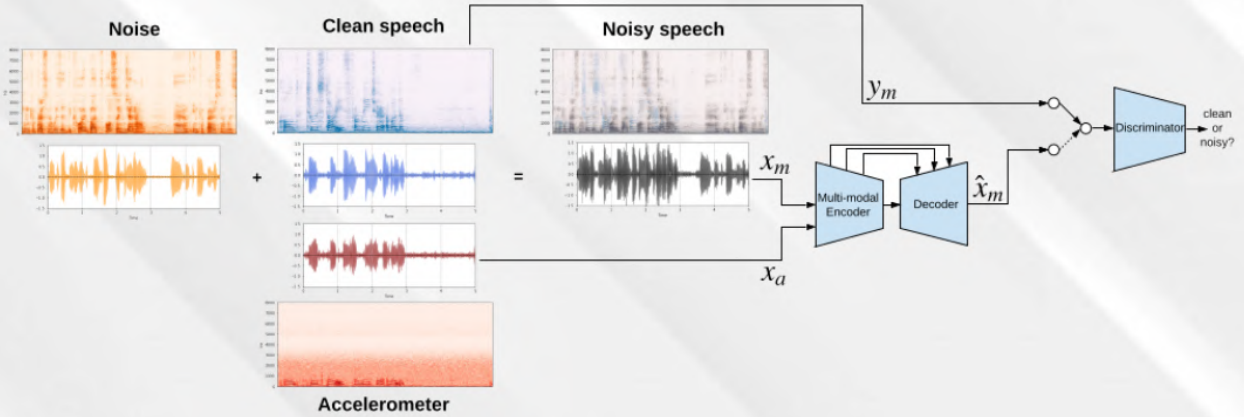
le cnam

SEANET

presented at INTERSPEECH 2020 by Google Research

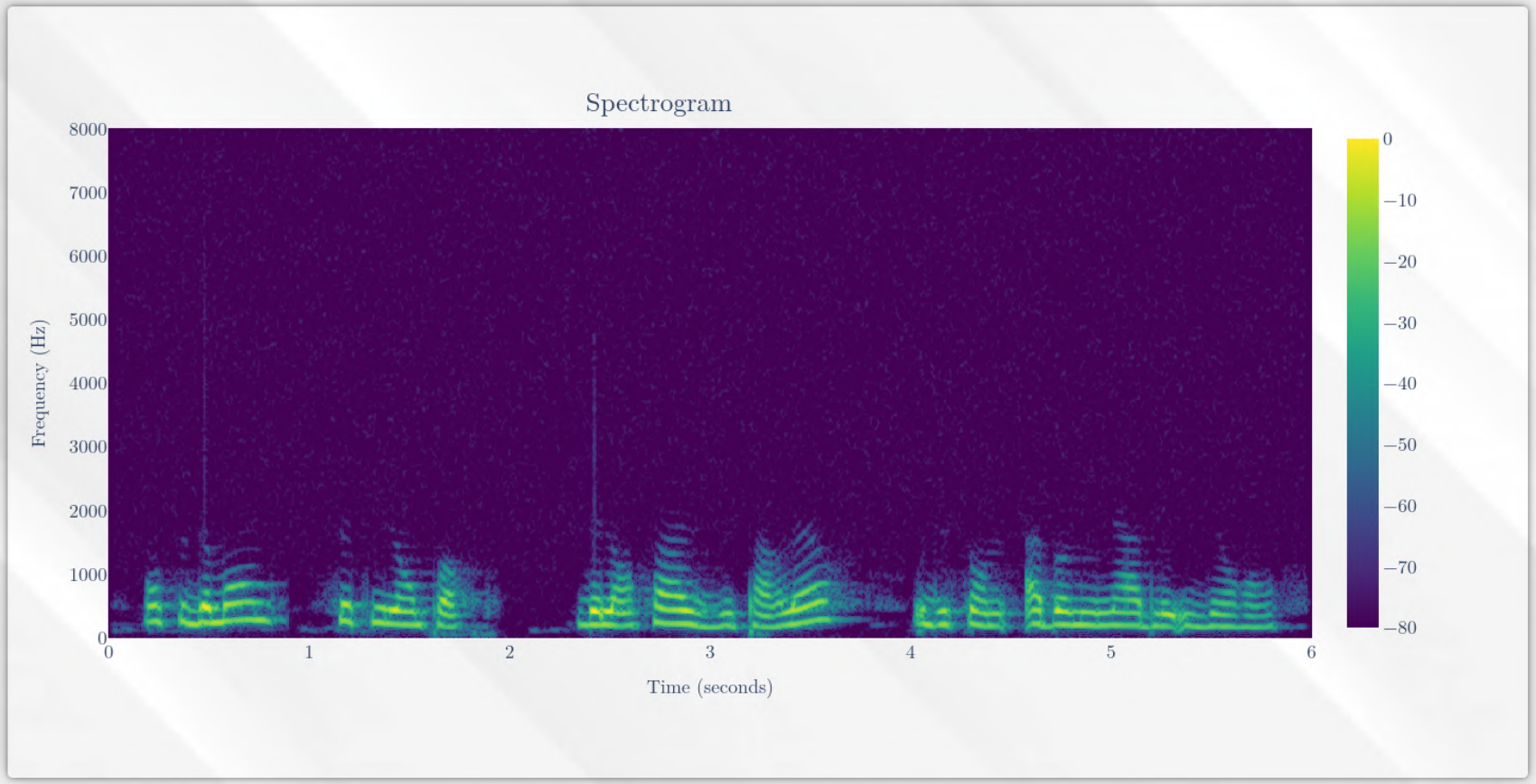
PRINCIPLE

$$\mathcal{L}_{gen} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{adv} \qquad \mathcal{L}_{dis} = \mathcal{L}_{adv\ fake} + \mathcal{L}_{adv\ real}$$



Model overview

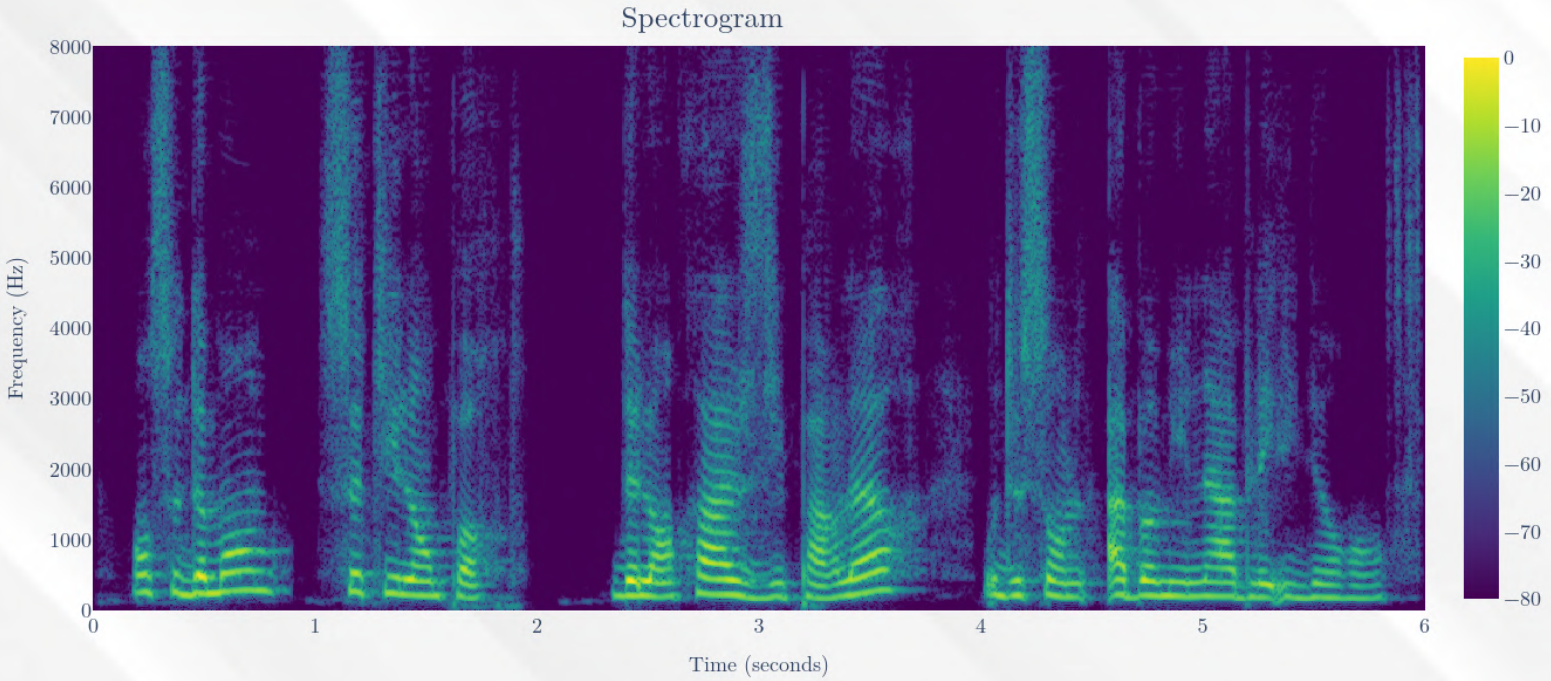
RESULTS



0:00 / 0:06

le cnam

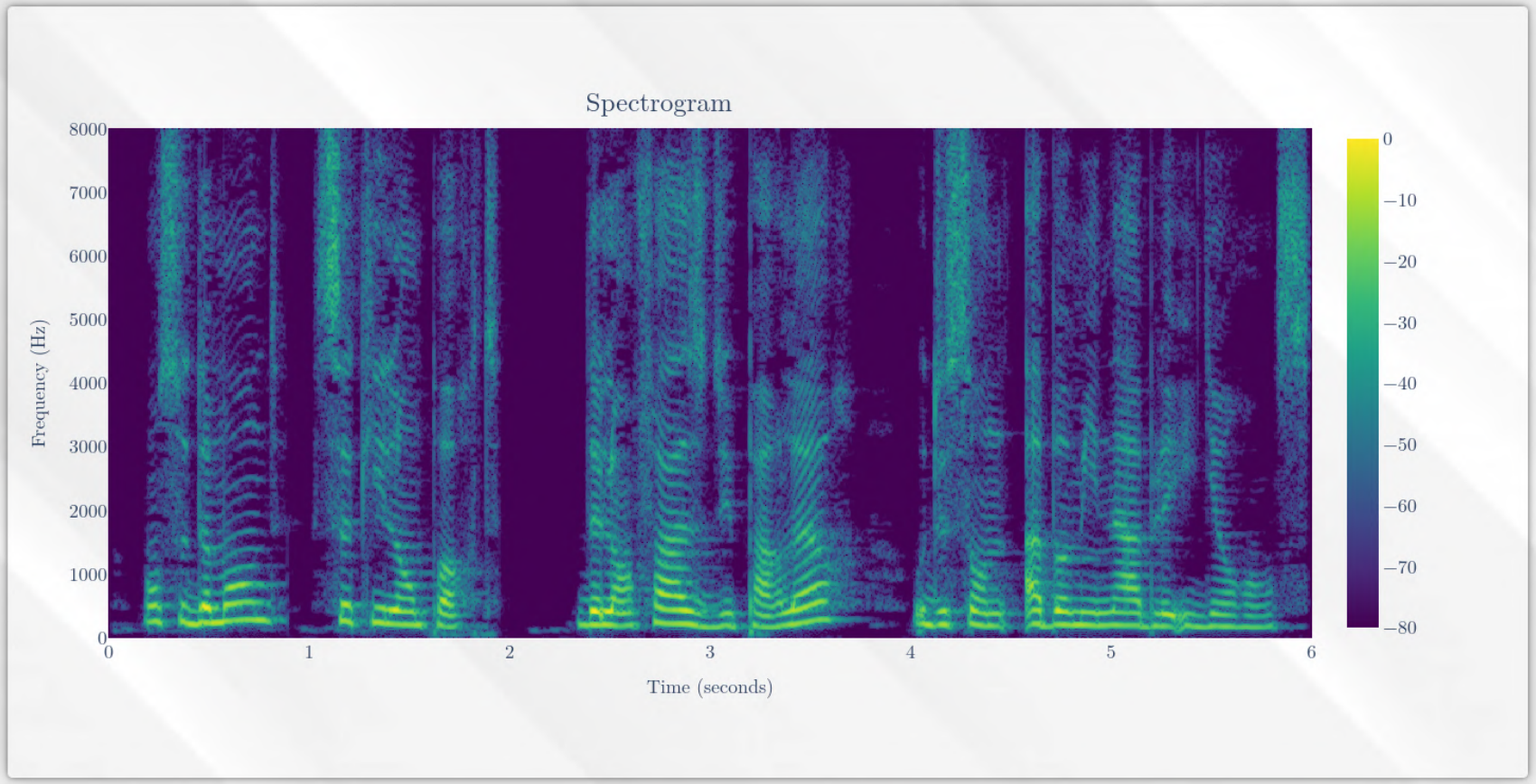
RESULTS



0:00 / 0:06

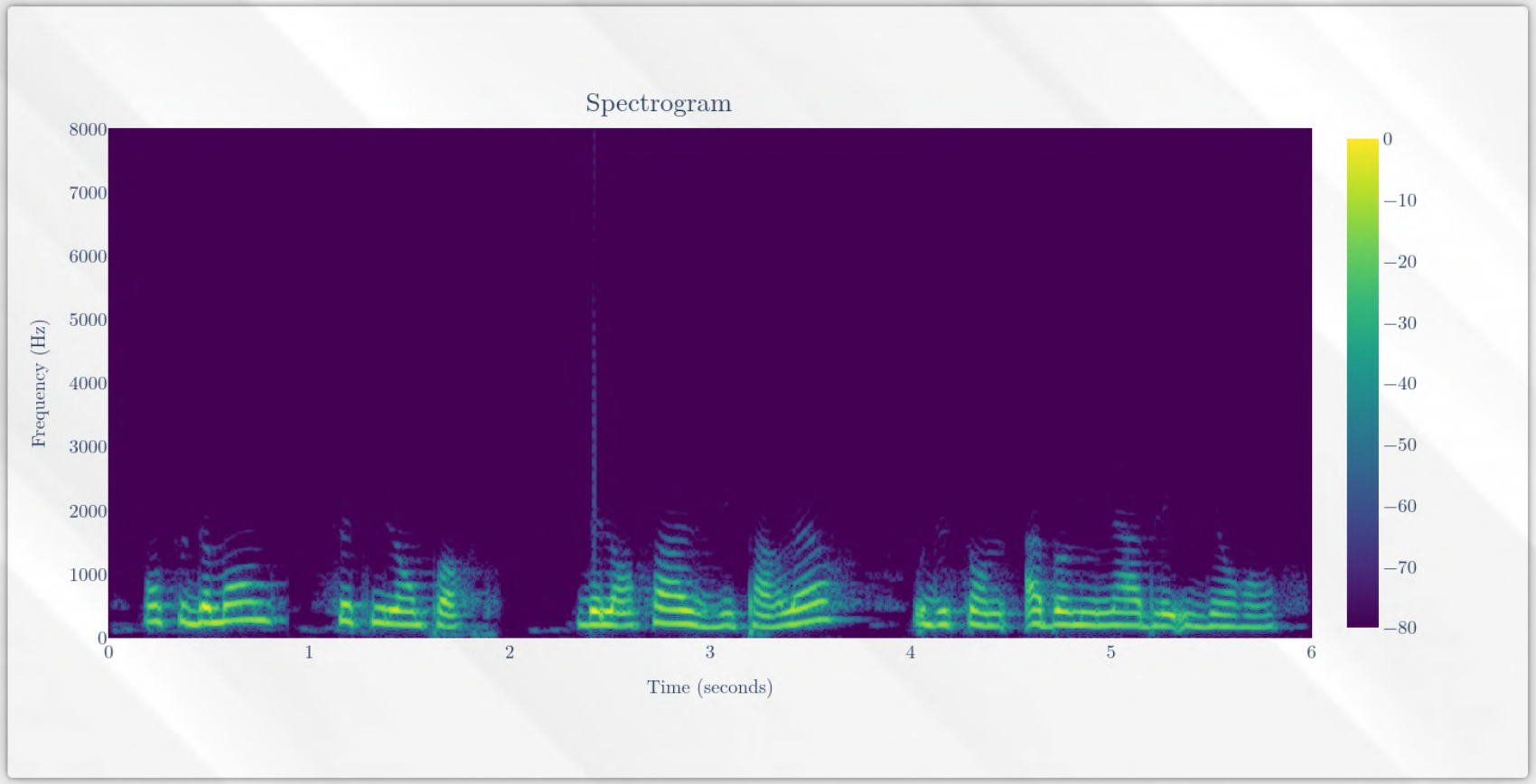
0:00 / 0:06

RESULTS



0:00 / 0:06 0:00 / 0:06 0:00 / 0:06

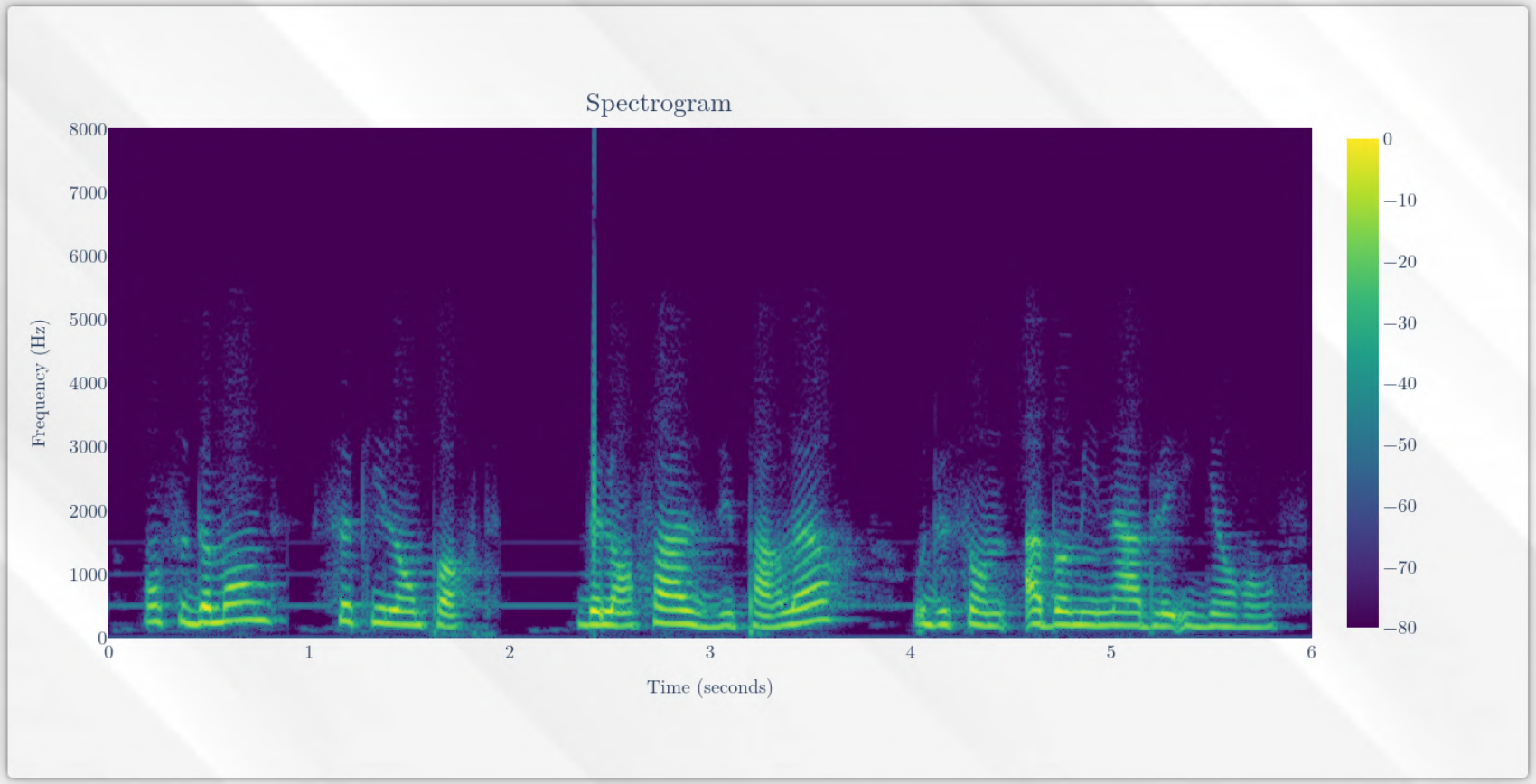
RECONSTRUCTIVE ONLY: RESULTS



0:00 / 0:06

le cnam

RECONSTRUCTIVE ONLY: RESULTS

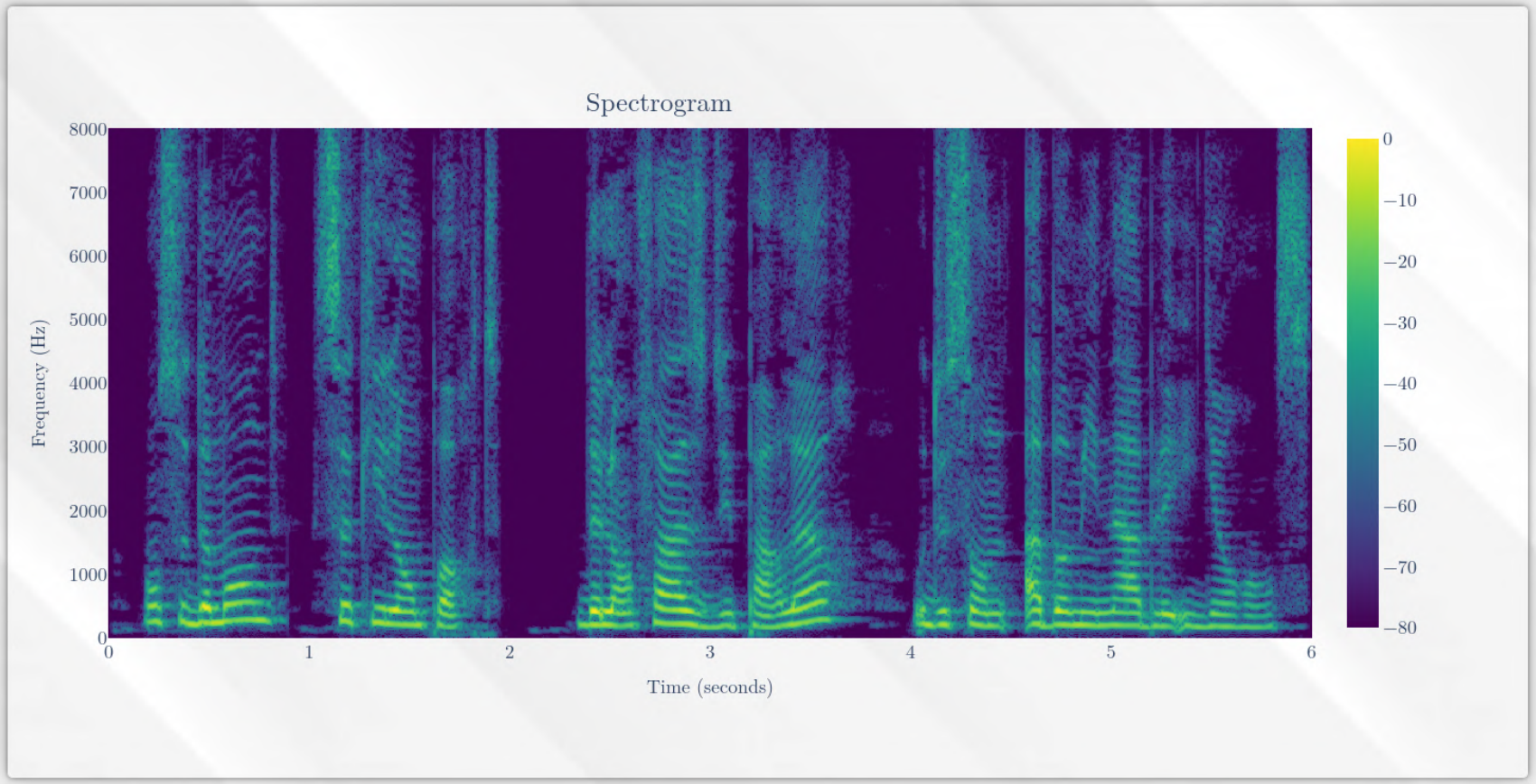


0:00 / 0:06

0:00 / 0:06

le cnam

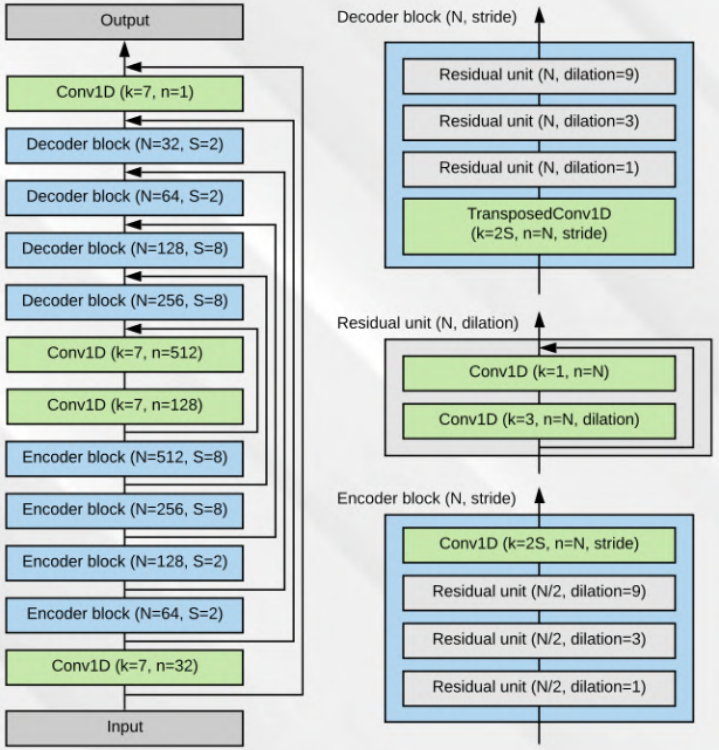
RECONSTRUCTIVE ONLY: RESULTS



0:00 / 0:06 0:00 / 0:06 0:00 / 0:06

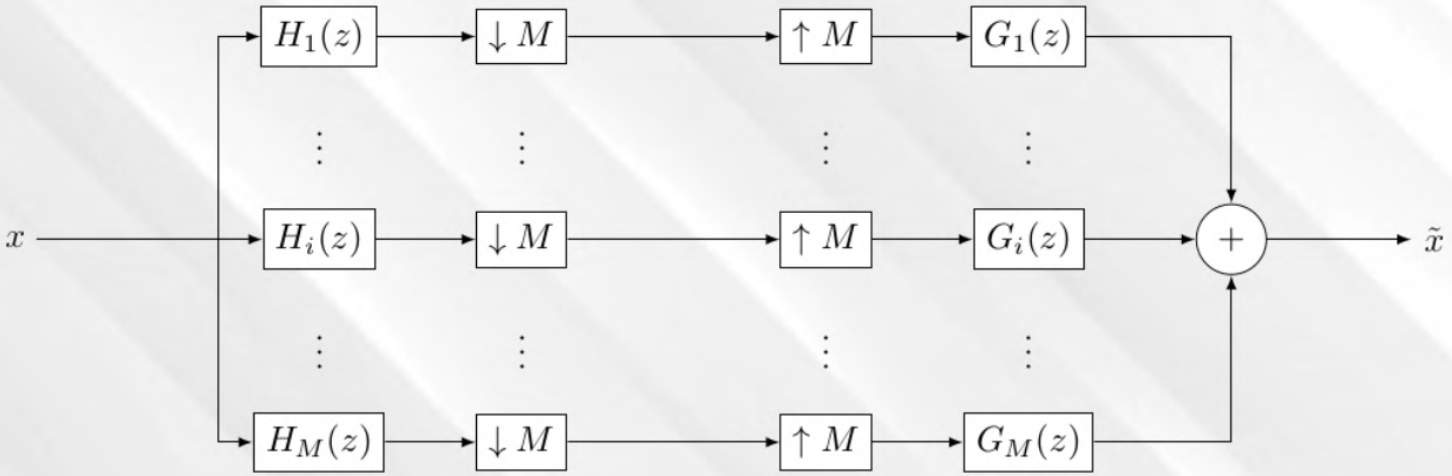
PQMF APPROACH TO LIGHTEN PROCESSING

STARTING POINT

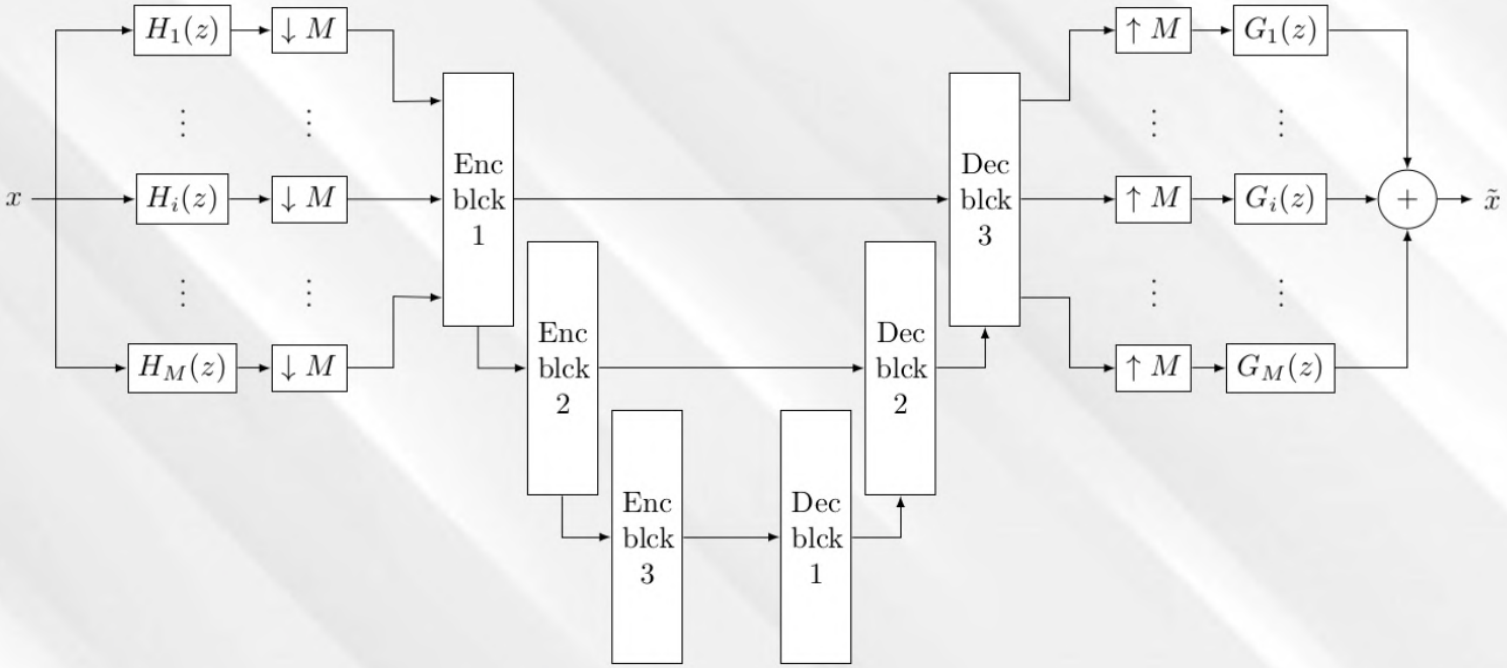


SEANet's generator architecture

le cnam

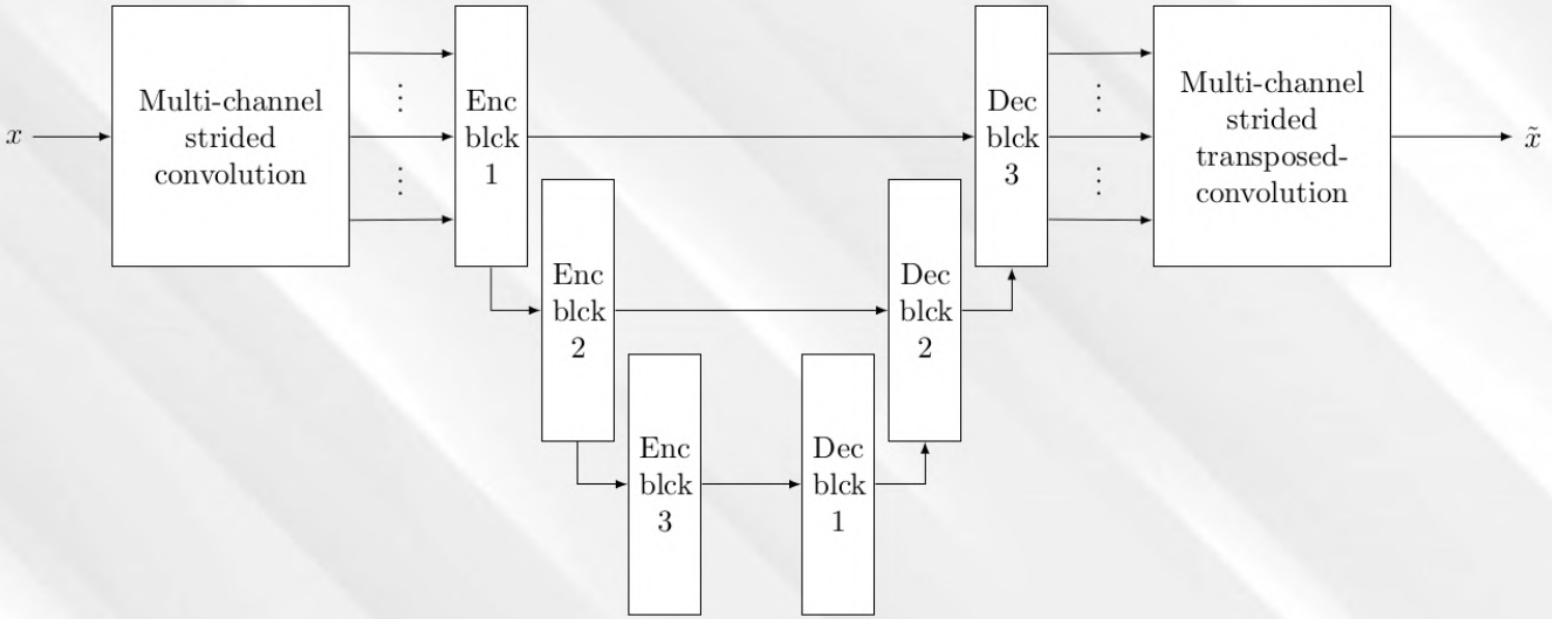


Pseudo Quadrature Mirror Filter (PQMF) banks



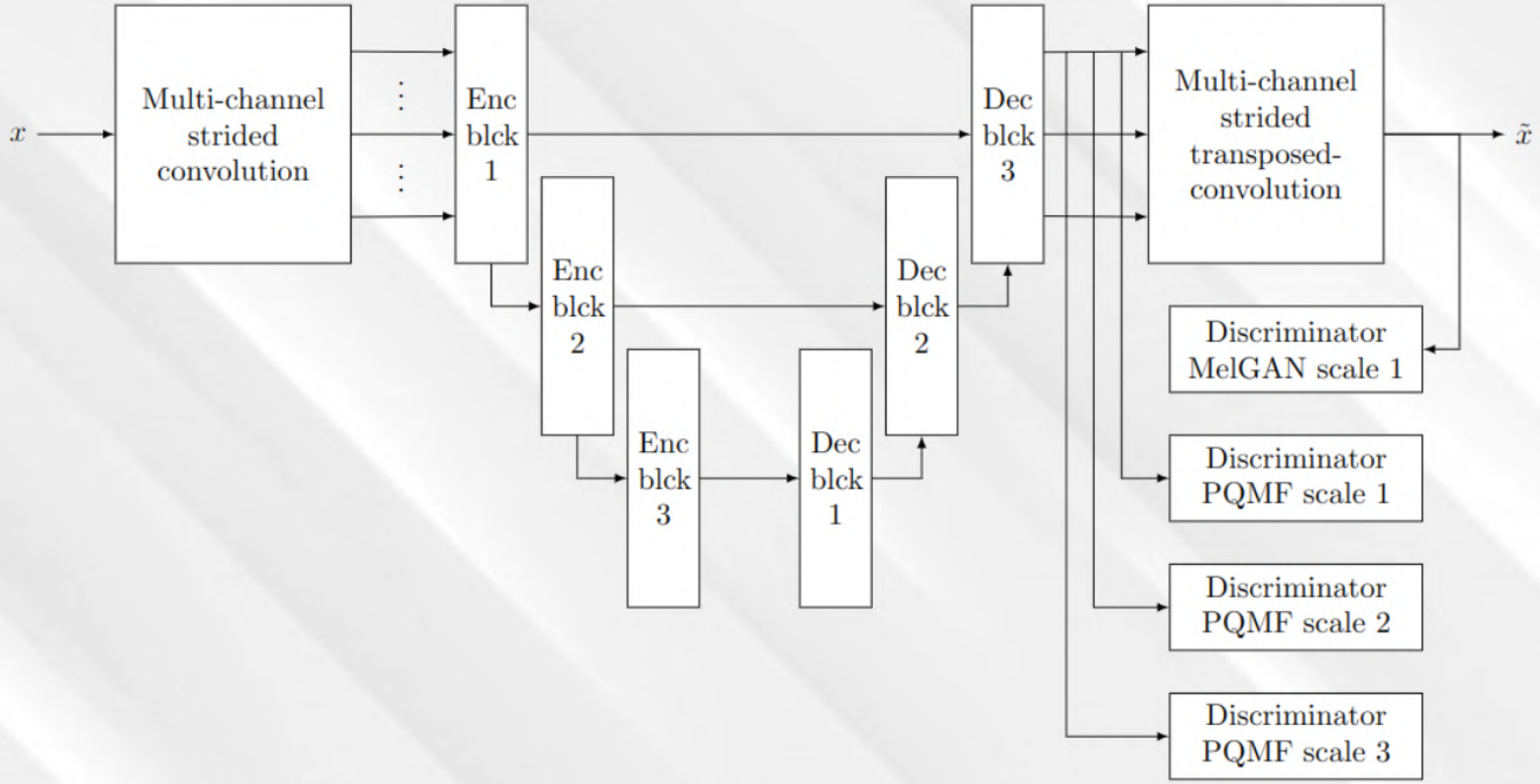
PQMF-generator

le cnam



Fast PQMF-generator

le cnam



Our PQMF-SEANet

le cnam

SUMMARY ON PQMF-SEANET

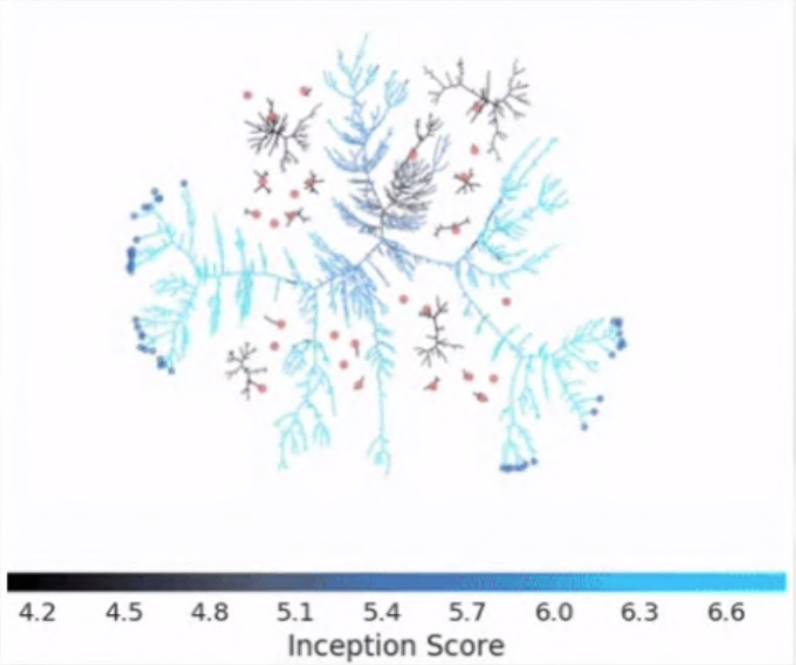
AVANTAGES

- **Memory gain** thanks to the reduction of features' temporal length
- **Gain in inference speed** thanks to the reduced number of layers
- **More flexibility** for processing and discriminators inputs

DISADVANTAGES

- Projection in a non-optimal basis for our problem
- Convolution with large kernels difficult to parallelize

AUTOMATE THE SEARCH FOR HYPERPARAMETERS



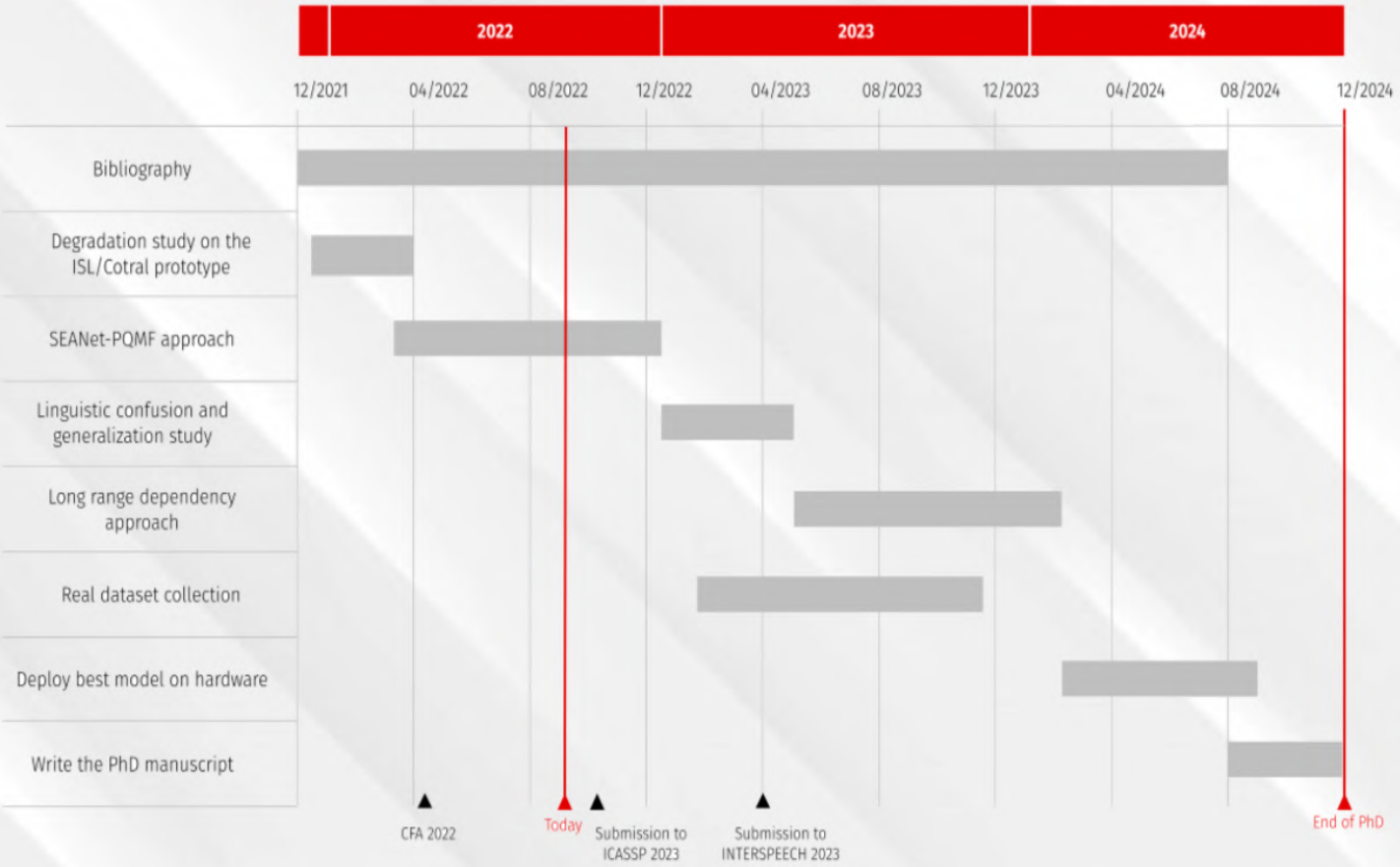
Population based training

le cnam

EVALUATION

Signals	PESQ	SI-SDR	SI-SNR	STOI
In-ear	1.246	-6.615	-7.538	0.7178
SEANet baseline	2.321	10.99	10.79	0.8985
Mix	2.573	11.25	11.05	0.8961
PQMF SEANet	2.245	11.52	11.42	0.890
PQMF SEANet + tricks	2.435	11.94	11.84	0.902

TIMELINE


2022
2023
2024
12/2021
04/2022
08/2022
12/2022
04/2023
08/2023
12/2023
04/2024
08/2024
12/2024
Bibliography
Degradation study on the ISL/Cotral prototype
SEANet-PQMF approach
Linguistic confusion and generalization study
Long range dependency approach
Real dataset collection
Deploy best model on hardware
Write the PhD manuscript
CFA 2022
Today
Submission to ICASSP 2023
Submission to INTERSPEECH 2023
End of PhD

THANK YOU FOR YOUR ATTENTION

le cnam

Title: Deep learning for speech enhancement applied to non-conventional sound capture

contact: julien.hauret@lecnam.net

acknowledgments to: Éric Bavu, Thomas Joubaud, Véronique Zimpfer, Robin Petit

le cnam lmssc

Cotral Lab

ISL

AGENCE NATIONALE DE LA RECHERCHE ANR

SIGNAL PROCESSING EQUATIONS

Source-Filter model

$$y(t) = (h * x)(t)$$

$$Y(f) = H(f).\, X(f)$$

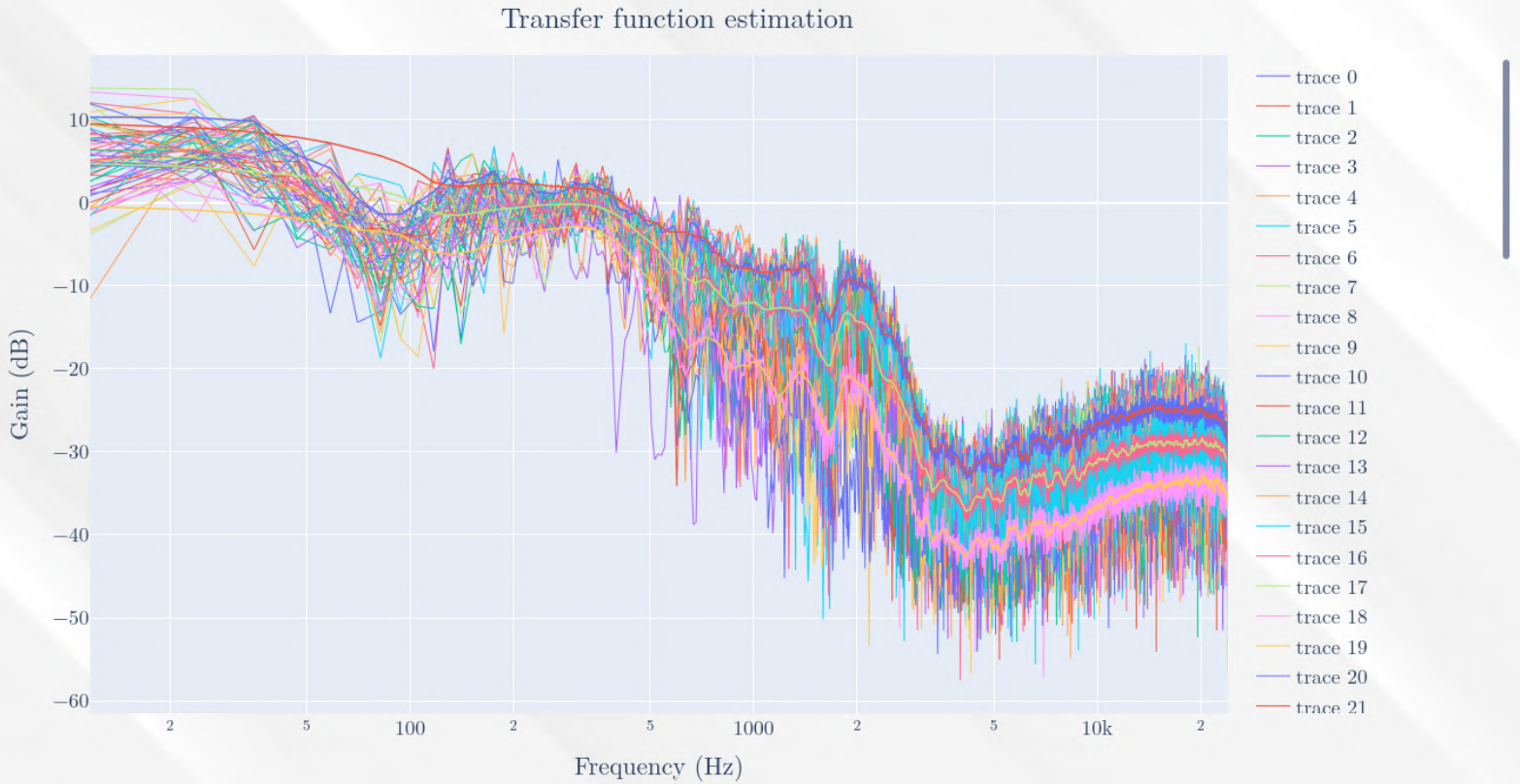
Transfer function estimation

$$H(f) = \frac{P_{xy}(f)}{P_{xx}(f)}$$

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f).\, P_{yy}(f)}$$

le cnam

BODE DIAGRAM



le cnam

LOSS FUNCTIONS : SEANET

$$\mathcal{L}_D = E_y \left[\frac{1}{K} \sum_k \frac{1}{T_k} \sum_t max(0, 1 - D_{k,t}(y)) \right] + E_x \left[\frac{1}{K} \sum_k \frac{1}{T_k} \sum_t max(0, 1 + D_{k,t}(G(x))) \right]$$

$$\mathcal{L}_G^{adv} = E_x \left[\frac{1}{K} \sum_k \frac{1}{T_k} \sum_t max(0, 1 - D_{k,t}(G(x))) \right]$$

$$\mathcal{L}_G^{rec} = E_x \left[\frac{1}{KL} \sum_{k,l} \frac{1}{T_{k,l}} \sum_t \| D_{k,t}^{(l)}(y) - D_{k,t}^{(l)}(G(x)) \|_{L_1} \right]$$

RÉFÉRENCES - 1

PQMF

- Joseph Rothweiler: Polyphase quadrature filters–a new subband coding technique. In ICASSP 1983.
- Truong Q Nguyen: Near-perfect-reconstruction pseudo-qmf banks. In 1994 IEEE.
- Yuan-Pei Lin & al: A kaiser window approach for the design of prototype filters of cosine modulated filterbanks. In 1998 IEEE.

le cnam

RÉFÉRENCES - 2

DEEP LEARNING

- Marco Tagliasacchi & al: Seanet : A multi-modal speech enhancement network. arXiv preprint, 2020.
- Mescheder, L., Geiger, A., & Nowozin, S. (2018, July). Which training methods for GANs do actually converge?. In International conference on machine learning (pp. 3481-3490). PMLR.
- Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., ... & Kavukcuoglu, K. (2017). Population based training of neural networks. arXiv preprint arXiv:1711.09846.
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.
- Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.
- Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems, 29.

RÉFÉRENCES - 3

MÉTRIQUES

- Antony W Rix & al : Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE.
- Cees H Taal & al: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE.
- Jonathan Le Roux & al: Sdr–half-baked or well done ? In ICASSP 2019.
- Yi Luo & al : time-domain audio separation network for real-time, single-channel speech separation. In 2018 IEEE.