Ciências ULisboa
Faculdade de Ciências da Universidade de Lisboa

M.Sc. *Computer Engineering*
Deep Learning class 2024/2025

Audio Source Separation

António Estêvão [58203], Diogo Venes [58216] and Guilherme Gouveia [58176]

Keywords: *STFT (Short-Time Fourier Transform), DPRNN (Dual-Path Recurrent Neural Network)*

Objectives

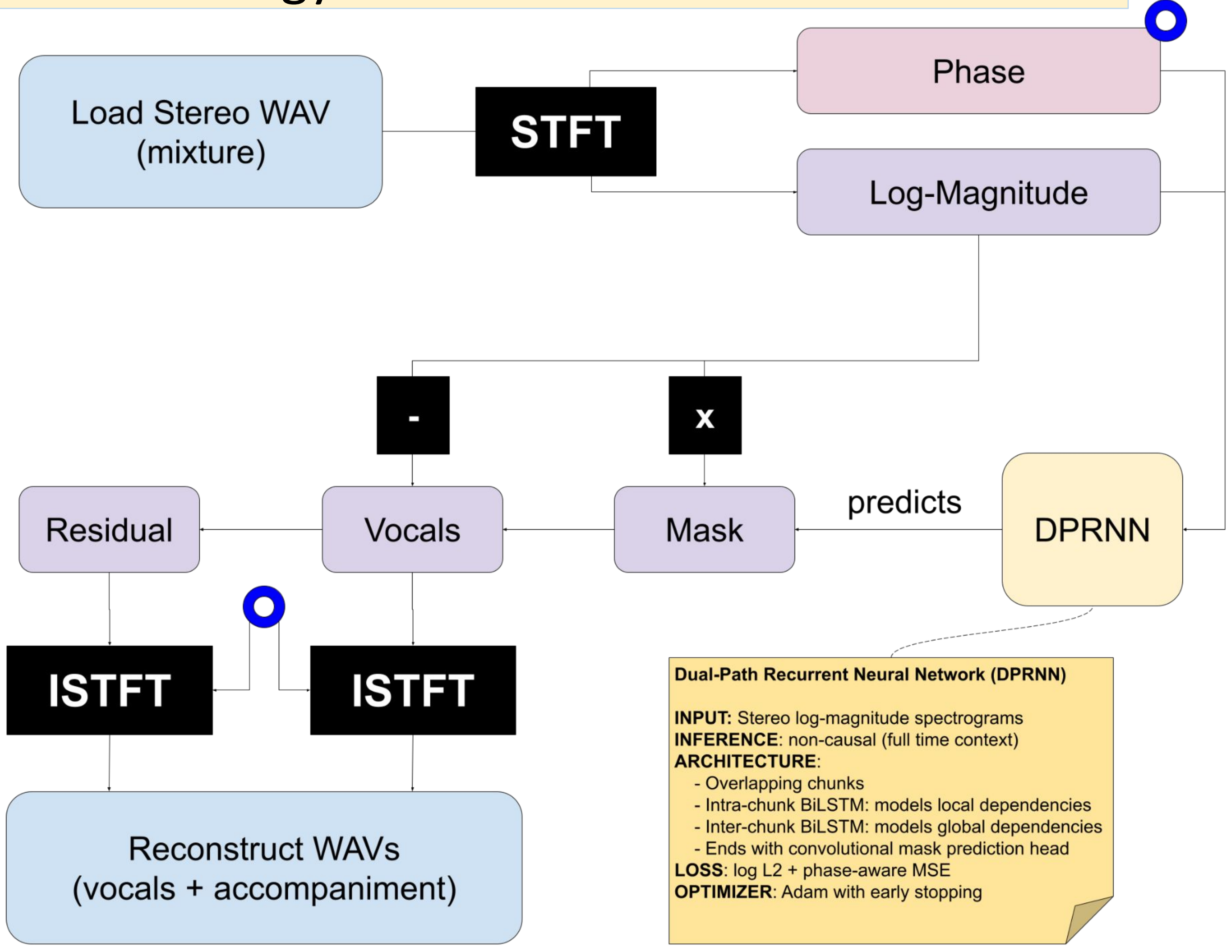
- **Build** and **evaluate** a **deep learning system** to **separate components of an audio track**: vocals and accompaniment.
- **Compare** our developed model with a popular **pre-trained model** (**Open-Unmix UMXHQ**) through the use of standard **separation metrics** that evaluate the quality of the separation as well as the lack of unwanted sources and distortions.

Problem

- **Audio source separation** is the process of isolating specific sounds from an original audio source containing a mixture of multiple elements, namely instruments and vocals.
- For example, we might want to separate the instrumental of a song from its vocals to make a karaoke track or isolate a specific instrument from a song so a musician can learn that part.
- In this type of task, we usually have 2 main models:

1. **Time-Frequency Domain**, using *STFT* to separate sources based on magnitude/phase or by predicting a mask to isolate target sources from the mixture.
   Drawbacks: Phase reconstruction can limit quality; resolution trade-offs (time vs. frequency).
2. **Time-Domain**, by learning to map raw waveforms directly to separated sources (*end-to-end*).
   Drawbacks: Often more complex, require more training data and compute resources.

Methodology

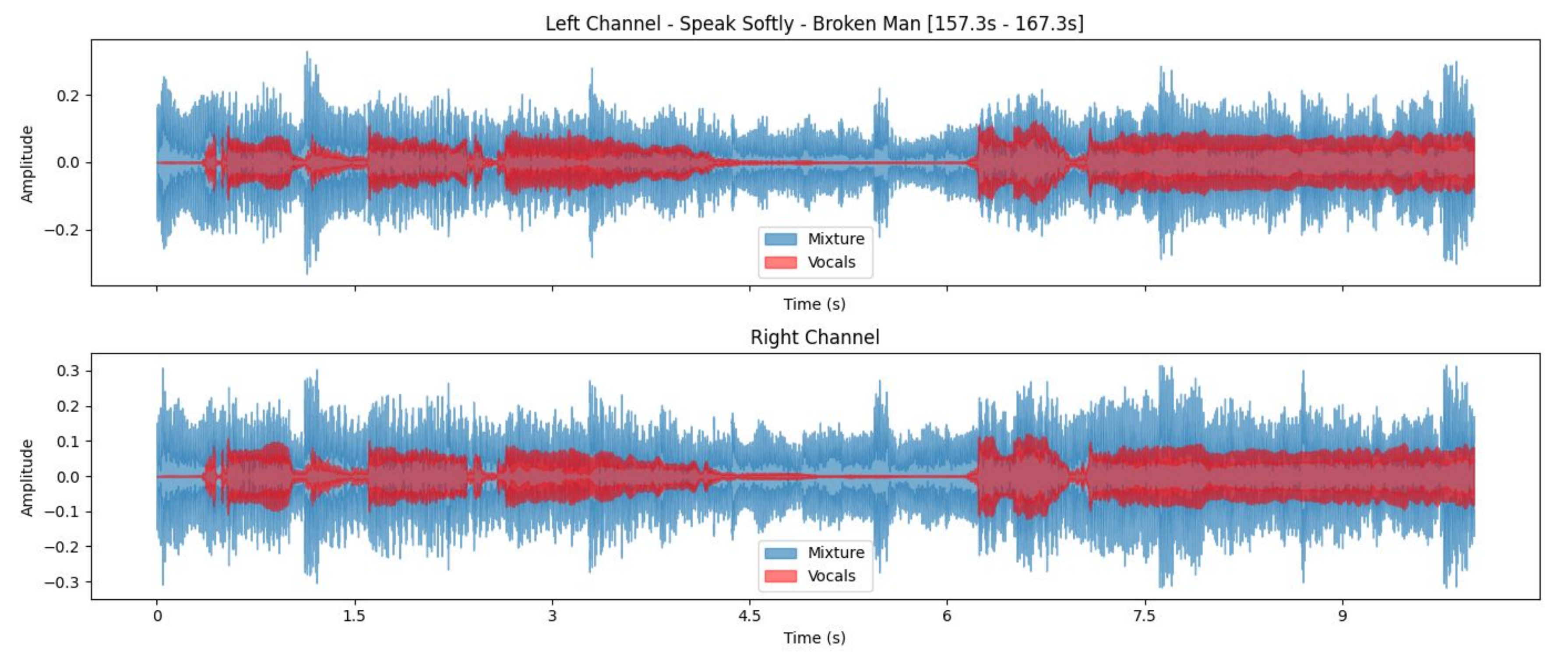


References

[1] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019). MUSDB18-HQ - An uncompressed version of MUSDB18. *Zenodo*. https://doi.org/10.5281/zenodo.3338373
[2] Stöter, F.-R., Liutkus, A., & Ito, N. (2018). The 2018 Signal Separation Evaluation Campaign. *arXiv preprint arXiv:1804.06267*. https://arxiv.org/pdf/1804.06267
[3] Araki, S., Haeb-Umbach, R., Ito, N., Wichern, G., Wang, Z-Q. & Mitsufuji, Y. (2025). 30+ Years of Source Separation Research: Achievements and Future Challenges. *arXiv preprint arXiv:2501.11837*. https://arxiv.org/pdf/2501.11837
[4] Solovyev, R., Stempkovskiy, A.& Habruseva, T. (2024). Benchmarks and Leaderboards for Sound Demixing Tasks. *arXiv preprint arXiv:2305.07489*. https://arxiv.org/pdf/2305.07489
[5] Stöter, F.-R., Liutkus, A. et al. (2019). Open-Unmix - A Reference Implementation for Source Separation. *GitHub repository*. https://github.com/sigsep/open-unmix-pytorch
[6] Butler, S.. (2024). Mono vs Stereo: What's the Difference and When Does It Matter? *How-To Geek*. https://www.howtogeek.com/mono-vs-stereo-whats-the-difference-and-when-does-it-matter/
[7] StackExchange User. (2015). What is the Importance of Phase Spectrum in Fourier Transform? *Mathematics Stack Exchange*. https://math.stackexchange.com/questions/1290620/what-is-the-importance-of-phase-spectrum-in-fourier-transform
[8] StackOverflow User. (2022). How to Calculate Metrics SDR, SI-SDR, SIR and SAR in Python. *Stack Overflow*. https://stackoverflow.com/questions/72939521/how-to-calculate-metrics-sdr-si-sdr-sir-sar-in-python
[9] Manilow, E., Seetharaman, P. & Salamon, J. (2020). Why Use Evaluation Metrics Instead of Human Reviews? *Source Separation Tutorial*. https://source-separation.github.io/tutorial/basics/evaluation.html
[10] Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). Scale-Invariant Source-to-Distortion Ratio: Optimization and Application to Speech Enhancement. *Mitsubishi Electric Research Labs Technical Report TR2019-013*. https://www.merl.com/publications/docs/TR2019-013.pdf
[11] Gusó, E., Pons, J., Pascual, S. & Serrà, J. (2022). Improving Music Source Separation Based on Deep Neural Networks through Data Augmentation and Network Blending. *arXiv preprint arXiv:2202.07968*. https://arxiv.org/pdf/2202.07968
[12] Luo, Y., Chen, Z. & Yoshioka, T. (2020). Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. *arXiv preprint arXiv:1910.06379*. https://arxiv.org/pdf/1910.06379
[13] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T. & Takahashi, N. (2017). Improving Music Source Separation Based on Deep Neural Networks through Data Augmentation and Network Blending. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 261–265). https://ieeexplore.ieee.org/document/7952158

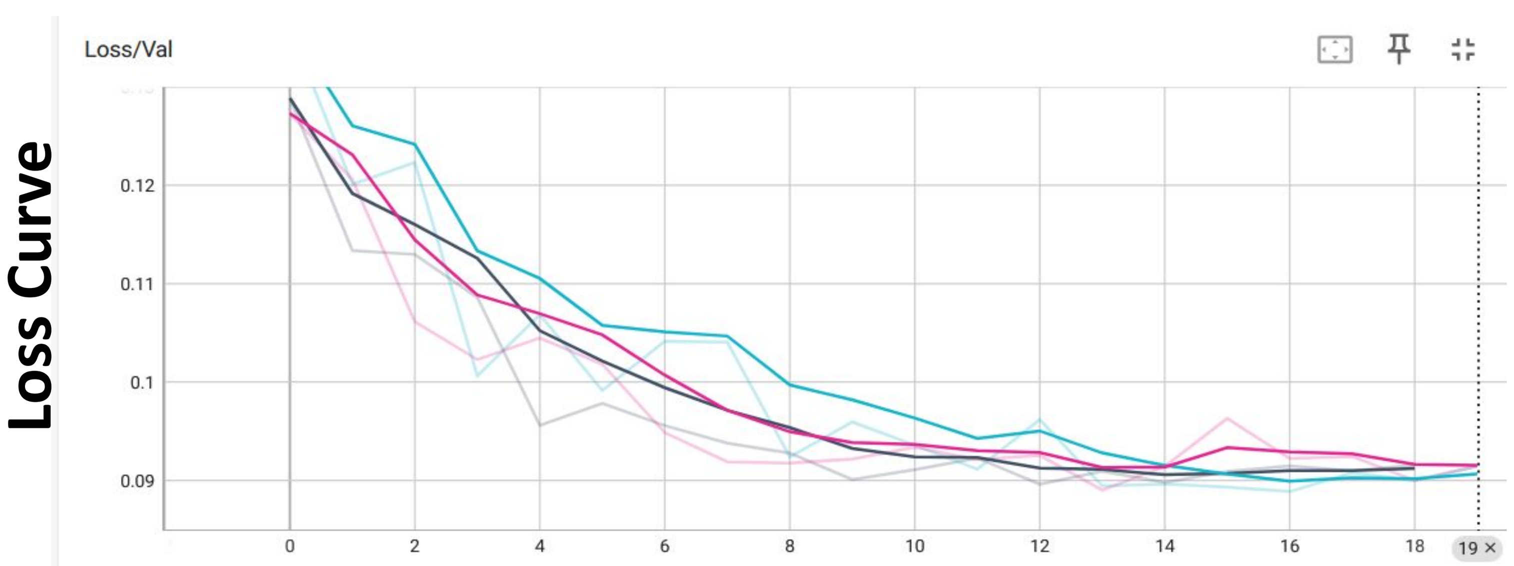
Datasets

- **MUSDB18-HQ**, an industry standard dataset, was used.
  - 150 full-length **stereo** tracks (~10 hours) of different genres split into 86 train tracks, 50 test tracks and 14 validation tracks.
  - Each track comes pre-split into **mixture**, **bass**, **drums**, **vocal** and **other** *.wav* files, of which we used the **vocal** and **mixture** files.
- Example **waveform plot** of one of the tracks (10 seconds):

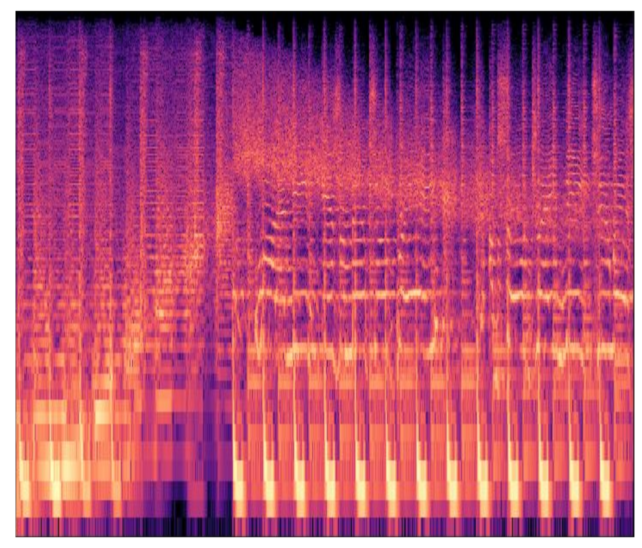


Results

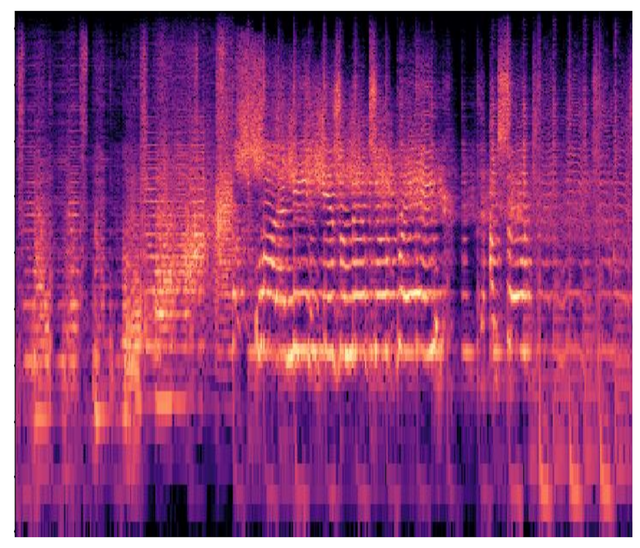
Loss Curve



Mixture Spec



Pred Vocals Spec



- Comparison of industry standard metrics between the **2 models**, tested on **30 second** chunks of **25 random** tracks:

Target	Model	SDR	SIR	ISR	SAR
Vocals	*Open-Unmix*	5.259	12.508	14.385	6.607
Accompaniment	*Open-Unmix*	**12.951**	19.901	**21.255**	**14.464**
Vocals	DPRNN-ass	2.516	5.445	9.482	6.011
Accompaniment	DPRNN-ass	6.094	**22.865**	6.941	8.449

Conclusions

- Our project demonstrates that **it is suitable to use a DPRNN-based** deep learning approach in the time-frequency domain for the audio source separation task, because the **dual intra/inter chunk RNN** balances **local** and **global context**.
- Adding the **phase-aware loss helped a lot**, even if we **don't yet have results comparable to the state of the art**.
- **Future directions** could include
  - Exploring **time-domain** architectures and **multi-target sep.**
  - Exploring more advanced **data augmentation** techniques like **pitch shifting** or **time stretching** before *STFT* and **parameter tuning** with *GridSearch*

Acknowledgements - This poster session is kindly sponsored by: NTT DATA