

UGOSA

User-guided one-shot deep model adaptation for music source separation

GIORGIA CANTISANI¹, ALEXEY OZEROV², SLIM ESSID¹, GAËL RICHARD¹

¹LTCl, Télécom Paris, Institut Polytechnique de Paris, 91120, Palaiseau, France

²InterDigital R&D France, 35510, Cesson Sevigne, France



MIPFrontiers



**INSTITUT
POLYTECHNIQUE
DE PARIS**



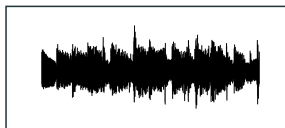
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No.765068.



IP PARIS



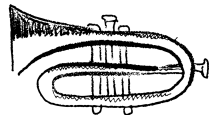
Audio Source Separation



Music mixture

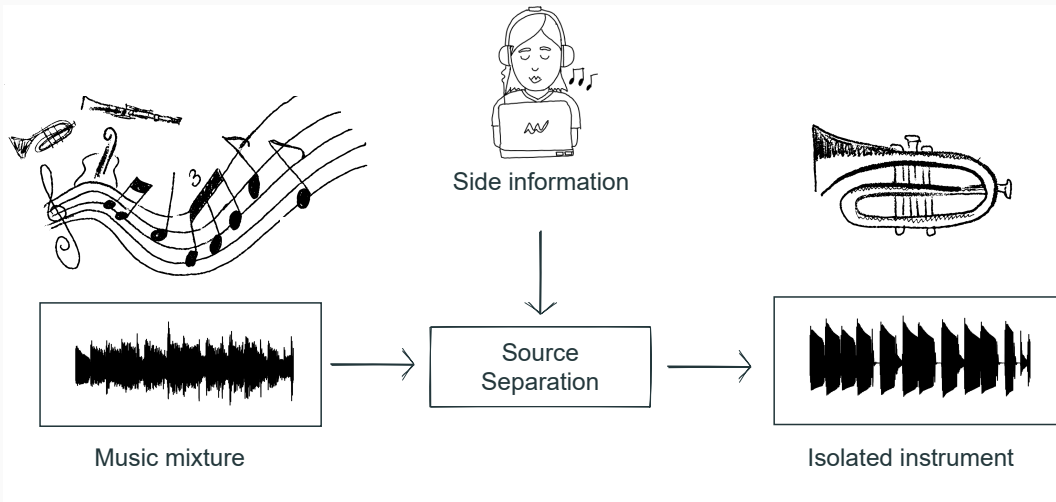


Source
Separation



Isolated instrument

User-driven Audio Source Separation



NMF/NTF-based:

- Time activations [Laurberg et al., 2008, Ozerov et al., 2011, Duong et al., 2014a]
- TF activations [Lefevre et al., 2012, Jeong and Lee, 2015, Rafii et al., 2015]
- Interactive frameworks [Bryan and Mysore, 2013, Duong et al., 2014b]
- Humming [Smaragdis and Mysore, 2009], sing/play [FitzGerald, 2012], F0 [Durrieu and Thiran, 2012]

DL-based:

- Multi-task learning [Stoller et al., 2018, Hung and Lerch, 2020, Nakano et al., 2020]
- Class conditioning
[Swaminathan and Lerch, 2019, Slizovskaia et al., 2019, Seetharaman et al., 2019, Karamatli et al., 2019]

However:

- ✗ require **large datasets** of mixtures and corresponding isolated sources;
- ✗ if informed, require the **side information** also during training;

One-shot deep model adaptation

Fully-supervised DNNs:

- ✗ require **large datasets** of mixtures and corresponding isolated sources;
- ✗ if informed, require the **side information** also during training;
- ✗ do **not generalize well** to test data with significant timbre variation.

Proposal: use the *side information* + *specific training strategies*

- *Adaptation*: fine-tuning of a pre-trained DNN using the side info
- *One-shot*: adaptation to one mixture and not a dataset
- *Side information*: time activations indicating where each instrument is active

Minimize difference between the estimated and the ground truth sources

$$L = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^I |\hat{s}_{i,n} - s_{i,n}|;$$

which represents the average absolute error between waveforms

- $\hat{s}_{i,n}$ estimated source i at time frame n ;
- $s_{i,n}$ ground truth source i at time frame n .

Minimize the energy of the silent sources while forcing the mix reconstruction:

$$L = \frac{1}{N} \sum_{n=1}^N \left[\underbrace{\sum_{i=1}^I |(h_{i,n} \cdot \hat{s}_{i,n}) - y_n|}_{\text{reconstruction loss}} + \lambda \underbrace{\sum_{i=1}^I |(1 - h_{i,n}) \cdot \hat{s}_{i,n}|}_{\text{activations loss}} \right].$$

where $h_{i,n}$ are the binary activations of each instrument i at time frame n :

$$\begin{cases} h_{i,n} = 1 & \text{if instrument } i \text{ is active} \\ h_{i,n} = 0 & \text{if instrument } i \text{ is not active} \end{cases}$$

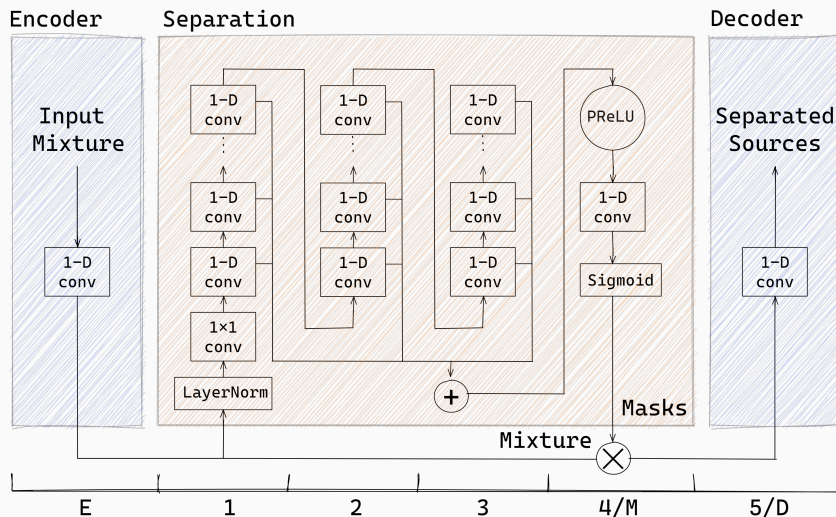
Data

- MUSDB18 [Rafii et al., 2017]
- 4 classes: bass, drums, vocals and other;
- first 10 songs of the test set only;
- binary activations of the ground truth sources.

Model:

- ConvTasnet adapted for music source separation [Luo and Mesgarani, 2019, Défossez et al., 2019]
- 10 epochs of fine-tuning using Ranger and $\text{lr } 10^{-5}$

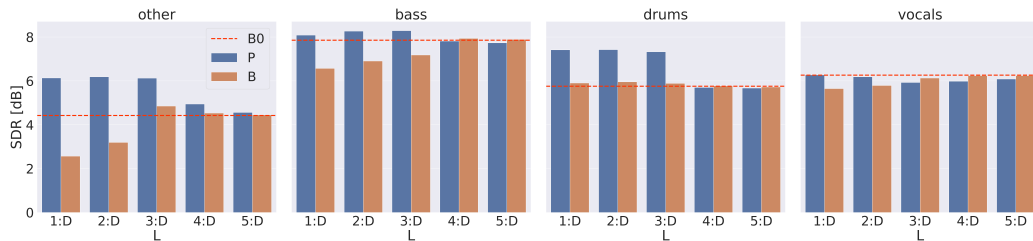
Fine-tuning strategies



Example P-L2:D: from the 2nd block to the last one with the proposed loss;

Results

SDR (dB)	#TP	other	bass	drums	vocals
P-L1:D	8.2M	6.1	8.1	7.4	6.3
P-L2:D	5.6M	6.2	8.3	7.4	6.2
P-L3:D	2.9M	6.1	8.3	7.3	5.9
P-L4:D	0.4M	4.9	7.8	5.7	6.0
P-L5:D	0.01M	4.6	7.7	5.7	6.1
B0	-	4.4	7.9	5.8	6.3



Conclusions

Take-home

- ✓ no need for the side info during training, **adaptation directly at test time**;
- ✓ **improvement of the separation**, especially for underrepresented instruments;
- ✓ **general approach** that can be applied to other tasks and DL architectures.




However:





- ✗ need at least a **weak guiding signal**;
- ✗ the **sources cannot be always activated**.





Resources:




-  <https://github.com/giorgiacantisani/ugosa>
-  <https://giorgiacantisani.github.io/projects/ugosa>

Thank you for the attention!




-  Bryan, N. and Mysore, G. (2013).
An efficient posterior regularized latent variable model for interactive sound source separation.
In *ICML*.
-  Défossez, A., Usunier, N., Bottou, L., and Bach, F. (2019).
Music source separation in the waveform domain.
arXiv preprint:1911.13254.
-  Duong, N. Q., Ozerov, A., and Chevallier, L. (2014a).
Temporal annotation-based audio source separation using weighted nonnegative matrix factorization.
In *IEEE ICCE-Berlin*.

-  Duong, N. Q., Ozerov, A., Chevallier, L., and Sirot, J. (2014b).
An interactive audio source separation framework based on non-negative matrix factorization.
In *IEEE ICASSP*.
-  Durrieu, J.-L. and Thiran, J.-P. (2012).
Musical audio source separation based on user-selected f0 track.
In *Int. Conf. LVA/ICA*. Springer.
-  FitzGerald, D. (2012).
User assisted separation using tensor factorisations.
In *Proc. 20th EUSIPCO*.
-  Hung, Y.-N. and Lerch, A. (2020).
Multitask learning for instrument activation aware music source separation.
In *ISMIR*.

-  Jeong, I.-Y. and Lee, K. (2015).
Informed source separation from monaural music with limited binary time-frequency annotation.
In *IEEE ICASSP*.
-  Karamatlı, E., Cemgil, A. T., and Kırılmaz, S. (2019).
Audio source separation using variational autoencoders and weak class supervision.
IEEE Signal Process. Lett.
-  Laurberg, H., Schmidt, M. N., Christensen, M. G., and Jensen, S. H. (2008).
Structured non-negative matrix factorization with sparsity patterns.
In *IEEE 42nd Asilomar Conf. on Signals, Systems and Computers*.
-  Lefevre, A., Bach, F., and Févotte, C. (2012).
Semi-supervised NMF with time-frequency annotations for single-channel source separation.
In *ISMIR*.

-  Luo, Y. and Mesgarani, N. (2019).
Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation.
IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP),
27(8):1256–1266.
-  Nakano, T., Koyama, Y., Hamasaki, M., and Goto, M. (2020).
Interactive deep singing-voice separation based on human-in-the-loop adaptation.
In *Proc. 25th Int. Conf. on Intelligent User Interfaces (IUI)*.
-  Ozerov, A., Févotte, C., Blouet, R., and Durrieu, J.-L. (2011).
Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation.
In *IEEE ICASSP*.

-  Rafii, Z., Liutkus, A., and Pardo, B. (2015).
A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds.
In *IEEE ICASSP*.
-  Rafii, Z., Liutkus, A., Stöter, F.-R., Mimitakis, S. I., and Bittner, R. (2017).
The MUSDB18 corpus for music separation.
-  Seetharaman, P., Wichern, G., Venkataramani, S., and Le Roux, J. (2019).
Class-conditional embeddings for music source separation.
In *IEEE ICASSP*.
-  Slizovskaia, O., Kim, L., Haro, G., and Gomez, E. (2019).
End-to-end sound source separation conditioned on instrument labels.
In *IEEE ICASSP*.

-  Smaragdis, P. and Mysore, G. J. (2009).
Separation by “humming”: User-guided sound extraction from monophonic mixtures.
In *IEEE WASPAA*.
-  Stoller, D., Ewert, S., and Dixon, S. (2018).
Jointly detecting and separating singing voice: A multi-task approach.
In *Int. Conf. on LVA/ICA*. Springer.
-  Swaminathan, R. V. and Lerch, A. (2019).
Improving singing voice separation using attribute-aware deep network.
In *IEEE MMRP*.