Singing voice separation: a study on training data.

Laure Prétet [1 2] Romain Hennequin [2] Jimena Royo-Letelier [2] Andrea Vaglio [1 2]

[1]LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France [2]Deezer R&D, Paris, France, research@deezer.com

Summary

- **Singing Voice Separation**: A very popular topic within the **Music Information Retrieval** community
- Goal: separate a music recording into two sources: **singing voice** and **instrumental accompaniment**
- State-of-the-art systems rely on **supervised deep learning** [4]
- The **design of training datasets** is a crucial factor in the performance of such systems

Problem: Results are generally presented for a full procedure, including dataset building, data pre-processing and/or augmentation, architecture design, post-processing and sometimes a long engineering work to tune the hyperparameters of the models.

⇒ **What is the impact of the training dataset on the performances?**

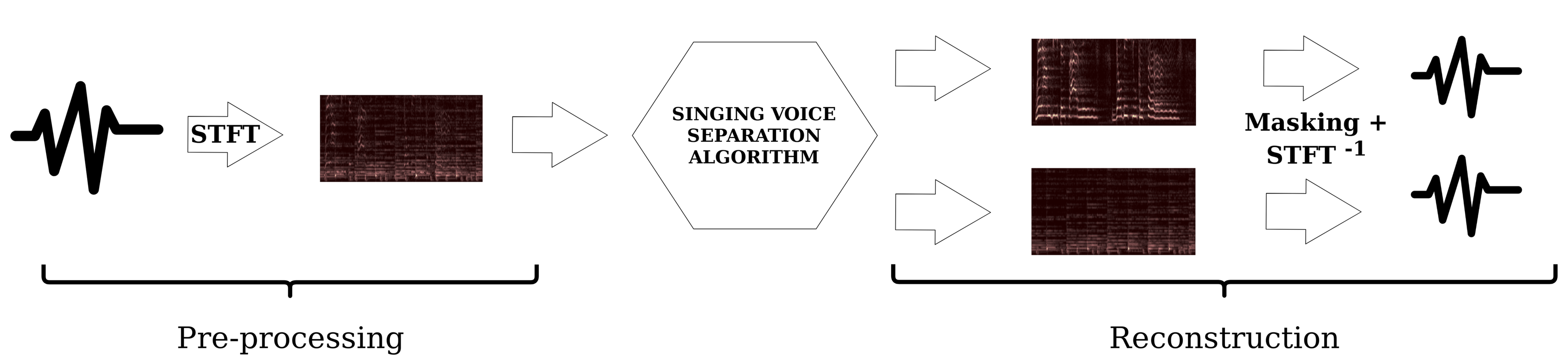
We tested the following factors:

- **Separation quality** of the dataset's tracks
- **Data diversity** (number of represented artists)
- **Data augmentation** for small datasets
- **Number of separated sources** available in the dataset

Architecture and Methodology

A standard methodology:

- Systems operate in the STFT magnitude domain.
- After separation, masks are computed from both spectrogram estimates and applied to the original mix.
- Reconstruction is performed using the phase of the original mixture.



Model architecture: the U-Net

- The U-net is a **convolutional neural network** that showed good performances for singing voice separation [2].
- **We vary the dataset while keeping the same model architecture.**
- For each experiment, we train one U-Net per source.
- We evaluate using the **MUSEVAL toolbox** (SDR, SIR, SAR) [4] on **2 test datasets**.
- We estimate the statistical significance of the results using a **Student's t-test**.

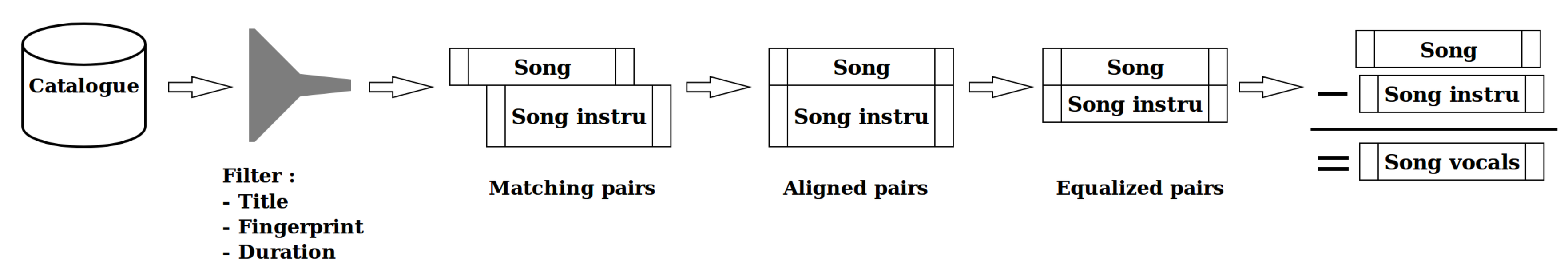
Listen to some audio examples and download our poster here !

Datasets

	MUSDB	Catalog	Bean
Diversity	150 songs	28,810 songs	24,097 songs
Quality	Separated recordings	Estimates	Separated recordings
Duration	10 hours	95 hours	79 hours

Main characteristics of the three datasets.

- **MUSDB:** Small public dataset, the reference dataset for singing voice separation.
  - 4-stems dataset: *drums, bass, vocals, other*
  - 2-stems dataset: *vocals, accompaniment*
- **Catalog:** Large dataset with estimated separated tracks built from Deezer's catalog (see [1]).
  - **Catalog Original:** Original genre distribution
  - **Catalog Balanced:** Rebalanced genre distribution
- **Bean:** Large private multi-track dataset.



How to build the Catalog dataset.

Experiment 1: Data augmentation

- We performed various data augmentation techniques on the smallest dataset (MUSDB).
- We adapted several transforms proposed by Schülter [3] for singing voice detection.

		Voice			Instruments		
Test dataset	Transform	SDR	SIR	SAR	SDR	SIR	SAR
MUSDB	*Baseline*	*4.32*	*12.62*	*4.1*	*10.65*	*13.46*	*11.51*
	Inverse Gaussian filtering	3.9	**13.35**	3.33	10.27	12.57	11.66
	Remixing	3.75	12.89	3.6	10.45	11.81	**12.05**
	Channel swapping	4.37	**13.01**	4.08	**10.69**	13.08	**11.74**
	Pitch shifting	4.0	**15.3**	3.5	10.58	12.46	**12.11**
	Loudness scaling	4.05	12.6	3.64	10.68	12.38	**11.85**
	Time stretching	4.19	**13.44**	3.57	10.96	12.76	**12.09**
	Combined	3.76	**13.86**	3.3	10.48	12.35	**11.72**
Bean	*Baseline*	*5.91*	*9.23*	*5.73*	*9.33*	*12.43*	*10.9*
	Inverse Gaussian filtering	5.58	**10.8**	5.2	9.18	11.53	10.75
	Remixing	5.7	**10.18**	5.44	9.43	11.1	**11.4**
	Channel swapping	**5.98**	**9.94**	**5.83**	**9.5**	12.25	**11.24**
	Pitch shifting	**6.06**	**11.53**	5.82	**9.57**	11.67	**11.63**
	Loudness scaling	5.87	**9.55**	5.66	9.42	11.71	**11.32**
	Time stretching	**6.12**	**10.68**	**5.94**	**9.64**	12.18	**11.35**
	Combined	5.98	**11.45**	**5.99**	9.4	11.1	**11.07**

N.S.
p < 0.05
p < 0.01
p < 0.001

Data augmentation experiment: Results of the source separation system trained on MUSDB with data augmentation (in dB). In bold are the results that significantly improve over the baseline (for $p < 0.001$). The colors represent the p-values: the darker, the more significant the results.

- Even when the improvement is statistically significant, it is very limited and hardly exceeds 0.2dB in SDR → it might not even be audible.

⇒ The various data augmentation types we tested seem to have quite a **low impact on separation results** - while being commonly used in the literature.

Experiment 2: Impact of the training dataset

		Voice			Instruments		
Test dataset	Train	SDR	SIR	SAR	SDR	SIR	SAR
MUSDB	*MUSDB (2 stems)*	*4.32*	*12.62*	*4.1*	*10.65*	*13.46*	*11.51*
	MUSDB (4 stems)	4.44	12.26	4.2	10.61	**13.7**	11.48
	Catalog Original	4.2	7.6	**7.44**	10.47	12.84	**12.03**
	Catalog Balanced	4.34	8.04	**7.05**	10.6	12.8	**12.12**
	Bean	**5.71**	**14.82**	**5.19**	**11.99**	**16.04**	**12.21**
Bean	*MUSDB (2 stems)*	*5.91*	*9.23*	*5.73*	*9.33*	*12.43*	*10.9*
	MUSDB (4 stems)	5.88	8.56	5.71	9.3	12.87	10.92
	Catalog Original	5.85	7.26	**7.16**	9.56	11.68	**12.3**
	Catalog Balanced	6.05	7.62	**6.79**	**9.74**	11.85	**12.42**
	Bean	**7.67**	**12.33**	**7.51**	**11.09**	**15.35**	**12.17**

N.S.
p < 0.05
p < 0.01
p < 0.001

Training dataset comparison experiment: Results of the source separation system trained on the 5 different datasets (in dB). In bold are the results that significantly improve over the baseline (for $p < 0.001$). The colors represent the p-values: the darker, the more significant the results.

- We expected high scores for the systems trained on Bean, since it is a large dataset with clean separated sources. And indeed, **training on the Bean dataset yields the highest scores** for most metrics on both the vocals and the accompaniment parts and on both test datasets.
- All other training datasets provide quite similar performances from one to another.
- Training the system with the Catalog dataset has a very limited impact on the separation performances compared to MUSDB alone.
- Moreover, training with Catalog Original or Catalog Balanced seems to provide very similar results.

Takeaway

For our experimental setup:

- **Data augmentation has a very limited impact** on the separation results when performed on a small training dataset.
- Using the 4 stems of MUSDB instead of vocals and accompaniment only does not improve the system performances either.
- **A large dataset with semi-automatically obtained vocal sources does not help much** the studied system compared to a smaller dataset with separately recorded sources.
- We confirmed a common belief that having a large dataset with clean separated sources improves significantly separation results over a small one.

Future work:

- Generalize these results to other state-of-the-art sources separation systems.
- Conduct perceptive tests for evaluation.

References

[1] Eric Humphrey, Nicola Montecchio, Rachel Bittner, Andreas Jansson, and Tristan Jehan. Mining labeled data from web-scale collections for vocal activity detection in music. In *Proceedings of the 18th ISMIR Conference*, 2017.

[2] Andreas Jansson, Eric J Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 323--332, 2017.

[3] Jan Schlüter. *Deep Learning for Event Detection, Sequence Labelling and Similarity Estimation in Music Signals*. PhD thesis, Johannes Kepler University Linz, Austria, July 2017.

[4] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293--305. Springer, 2018.