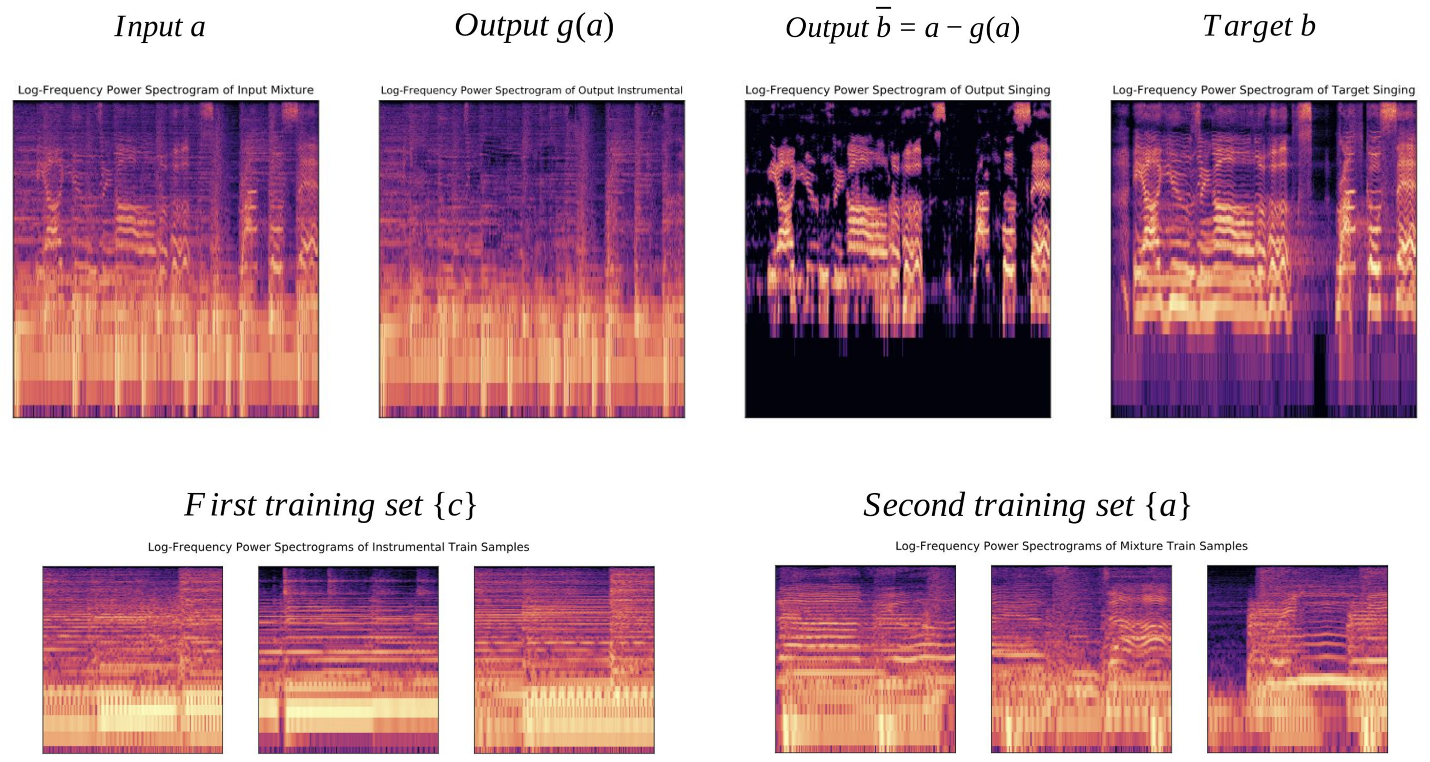
SEMI-SUPERVISED MONAURAL SINGING VOICE SEPARATION WITH A MASKING NETWORK TRAINED ON SYNTHETIC MIXTURES

Michael Michelashvili[1], Sagie Benaim[1], Lior Wolf[1,2]

[1]Tel Aviv University, [2]Facebook AI Research (FAIR)

Problem Statement

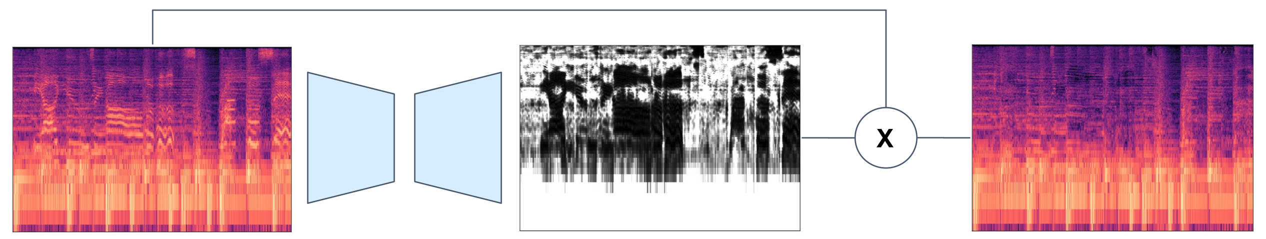
Semi-supervised source separation of singing from music, given two unmatched sets: a set of mixed music samples and a set of instrumental music.



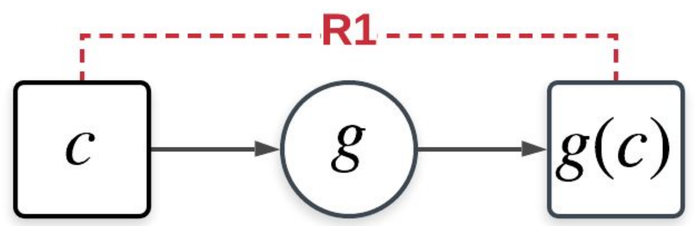
Motivation: In many source separation problems pure target channels are not available, thus methods with fully supervision cannot be used. We present a novel method in which such separation could be performed and applicable to any signal which obeys the superposition property.

Architecture & Losses

The network g is a learned autoencoder which produces a soft-mask multiplied with the network input:

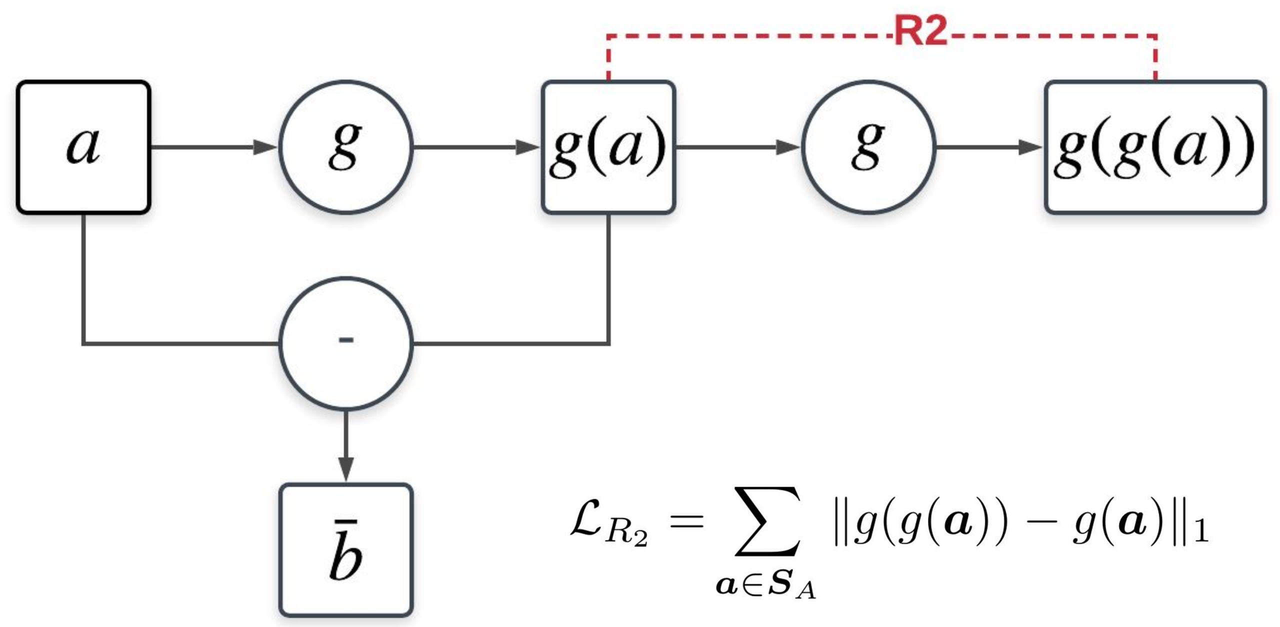


We want the network to be the identity operator on instrumental music - denoted as c, thus applying our first loss term:



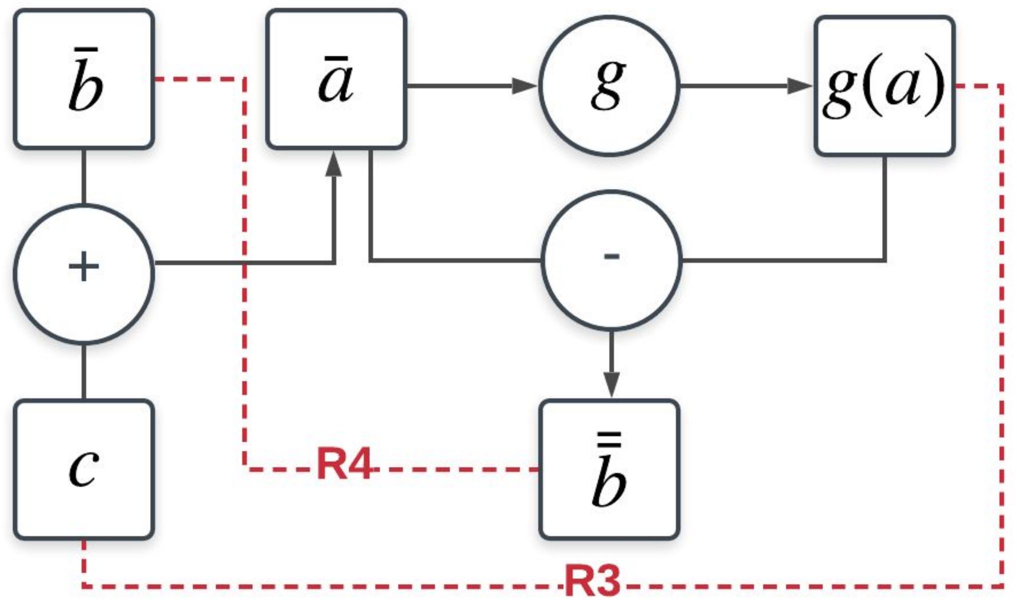
$$\mathcal{L}_{R_1} = \sum_{c \in S_C} \|g(c) - \mathbf{c}\|_1$$

The method is based on applying a learned function twice:
Phase I - applying on the mixture - denoted as a, in order to recover estimated singing voice samples - denotes as $\bar{b}$. We want the network to be idempotent (g∘g=g), meaning that applying it for the second time has no effect:



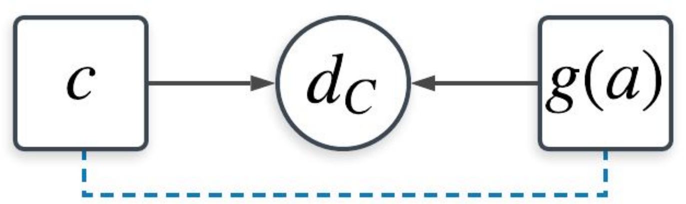
$$\mathcal{L}_{R_2} = \sum_{\mathbf{a} \in S_A} \|g(g(\mathbf{a})) - g(\mathbf{a})\|_1$$

Phase II - applying on synthetic mixes, in which the reconstructed singing samples are crossed with real instrumental samples from the training set. We want the output to be the same as the instrumental signal and the difference to be the same as the estimated singing voice. We present these requirement in losses R3 and R4:

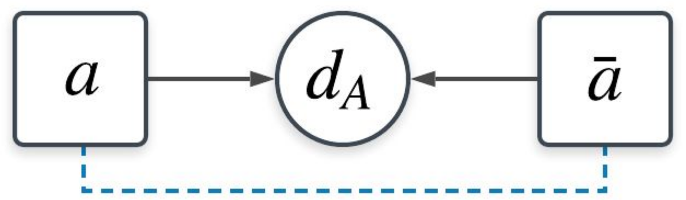


$$\mathcal{L}_{R_3} = \sum_{\bar{\mathbf{a}} \in \bar{S}_B \times C} \|g(\bar{\mathbf{a}}) - c(\bar{\mathbf{a}})\|_1 \qquad \mathcal{L}_{R_4} = \sum_{\bar{\mathbf{a}} \in \bar{S}_B \times C} \|(\bar{\mathbf{a}} - g(\bar{\mathbf{a}})) - b(\bar{\mathbf{a}})\|_1$$

GAN Losses - We utilizing GAN losses for aligning the distribution of output samples with instrumental music and the distribution of synthetic samples with real mixture samples:



$$\mathcal{L}_{\mathrm{GAN}_C} = \sum_{\mathbf{a} \in S_A} -\ell(d_C(g(\mathbf{a})), 0)$$



$$\mathcal{L}_{\mathrm{GAN}_A} = \sum_{\bar{\mathbf{a}} \in \bar{S}_B \times C} -\ell(d_A(\bar{\mathbf{a}}), 0)$$

Results

Comparison with semi-supervised baselines

We present state-of-the-art results for the semi-supervised problem setup:

Table 1. Median SDR (dB) for our method and previous semi-supervised approaches evaluated on the MUSDB18 [1] dataset. Baselines are from [2], which did not report SIR.

Approach	SDR	SIR	Approach	SDR	SIR
NMF	0.0	-	NES	0.3	-
GAN	0.3	-	NES-FT	2.1	-
GLOM	0.6	-	Ours	3.2	14.2

Comparison with supervised baselines

Comparison with fully-supervised methods shows we slightly worse than SOTA methods for SDR, but suppress all for SIR by a large gap. This is consistent with our observation - the network seems to filter out all the instrumental music very well for most samples. However, for some samples, there is a slight distortion of the voice generated:

Table 2. Median SDR and SIR (dB) values for the proposed method and previous supervised approaches, which solely deal with singing voice separation, evaluated on the evaluation subset of DSD100 [3] dataset.

Approach	Supervision	SDR	SIR
GRA3 [4]	supervised	-1.7	1.3
CHA [5]	supervised	1.6	5.2
STO2 [6]	supervised	3.9	6.7
JEO2 [7]	supervised	4.1	6.1
GRU-RIS-L [8]	supervised	4.2	7.9
MaDTwinNet [9]	supervised	4.6	8.2
Ours	semi-supervised	3.5	15.2

Ablation study

The ablation study clearly shows the importance of the proposed losses:

Table 3. Ablation study: Median SDR and SIR values for the proposed method without (w/o) selected losses evaluated on the evaluation subset of DSD100 [3].

Losses	SDR	SIR	Losses	SDR	SIR
All losses	3.5	15.2	w/o $\mathcal{L}_{R_4}$	-6.3	-4.7
w/o $\mathcal{L}_{R_1}$	-0.9	3.4	w/o $\mathcal{L}_{GAN_A}$	-6.3	-4.2
w/o $\mathcal{L}_{R_2}$	2.3	9.7	w/o $\mathcal{L}_{GAN_C}$	-4.1	-2.4
w/o $\mathcal{L}_{R_3}$	-4.3	13.3	w/o $\mathcal{L}_{GAN_A}$&$\mathcal{L}_{GAN_C}$	-17.0	-3.6

Reference

[1] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stoter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[2] Yedid Hoshen, Tavi Halperin, and Ariel Ephrat, "Neural separation of observed and unobserved distributions," in Submitted to Int. Conf. Learning Representations, 2019.

[3] Antoine Liutkus, Fabian-Robert Stöter, et al., "The 2016 signal separation evaluation campaign," in Latent Variable Analysis and Signal Separation, 2015.

[4] Emad M Grais, Gerard Roma, Andrew JR Simpson, and Mark D Plumbley, "Single-channel audio source separation using deep neural network ensembles," in Audio Engineering Society Convention 140, 2016.

[5] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gomez, "Monoaural audio source separation using deep convolutional neural networks," in Int. Conf. on Latent Variable Analysis and Signal Separation, 2017.

[6] Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau,Bernd Edler, and Paul Magron, "Common Fate Model for Unison source Separation," in ICASSP, 2016.

[7] Il-Young Jeong and Kyogu Lee, "Singing voice separation using rpca with weightedl1-norm," in Int. Conf. On Latent Variable Analysis and Signal Separation, 2017.

[8] Stylianos I. Mimilakis, Konstantinos Drossos, João F Santos, Gerald Schuller, Tuomas Virtanen, and Yoshua Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in ICASSP, 2018.

[9] Konstantinos Drossos, Stylianos Ioannis Mimilakis,et al., "MaD TwinNet: Masker-denoiser architecture with twin networks for monaural sound source separation," in Int. Joint Conf. on Neural Networks, 2018.

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant ERC CoG 725974)

Code:

Audio samples: