CARNATIC SINGING VOICE SEPARATION USING COLD DIFFUSION ON TRAINING DATA WITH BLEEDING

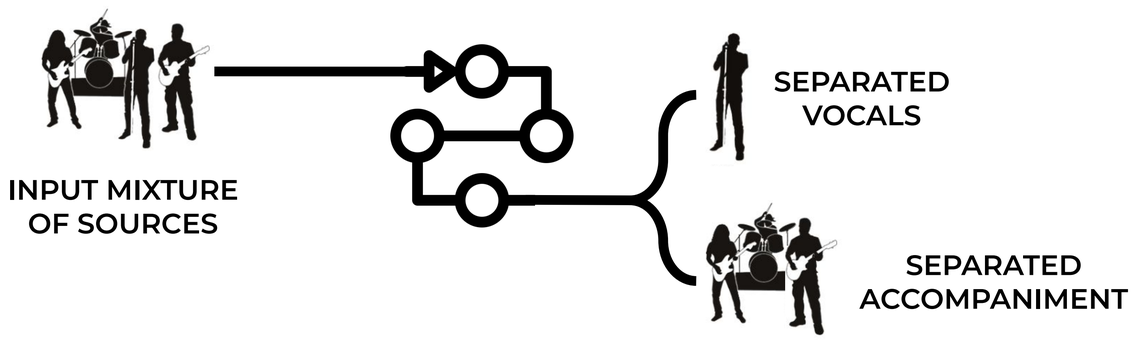
CODE AND MODELS

Genís Plaja-Roglans, Marius Miron, Adithi Shankar and Xavier Serra
Universitat Pompeu Fabra, Music Technology Group, Barcelona, Spain

ISMIR 2023

CONTEXT

SINGING VOICE EXTRACTION



The SOTA on the singing voice extraction task is lead by deep learning models

PROBLEM: what happens if...

- 🪘 Our instrument line-up is out of domain?
- 📁 No open and fully-isolated multi track data is available for our use case.

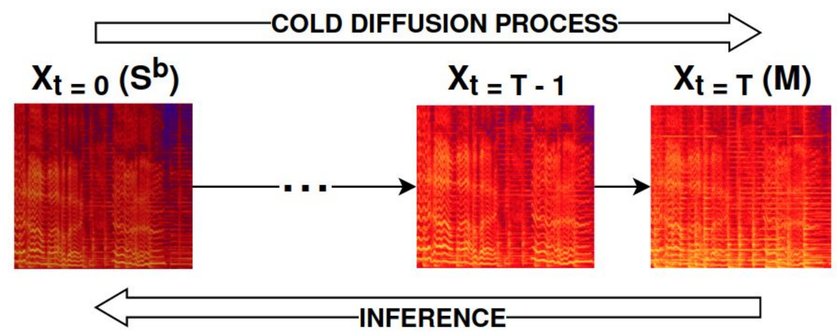
PROPOSED APPROACH

SEPARATION DATA W/ LEAKAGE

- Bleeding or leakage: the sound of the rest of the instruments being present in the background in the stem of a given source
- Can be recorded easily, e.g. in live shows
- The Saraga Dataset [1]: +34h of multi track audio data with leakage of Carnatic Music concerts

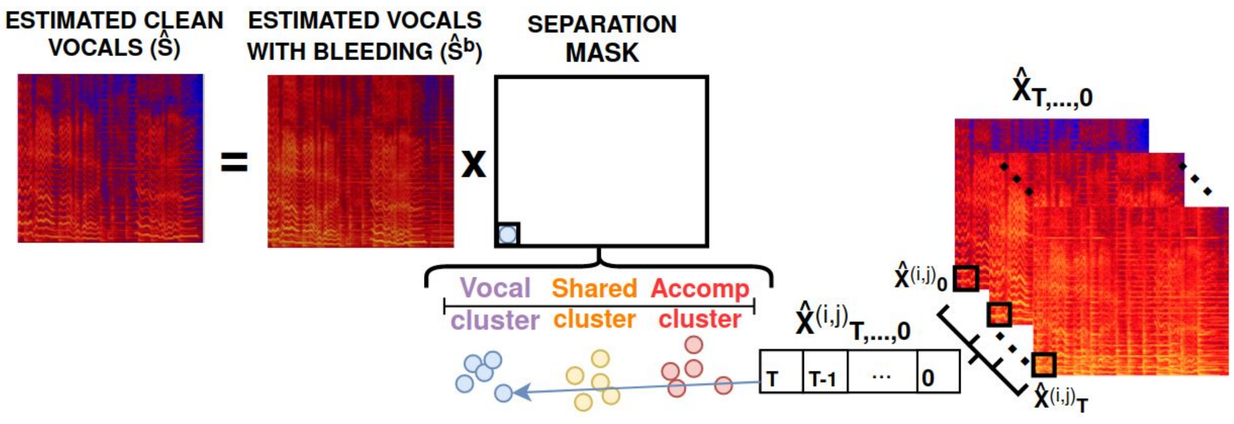
🎯GOAL: use these data to develop a singing voice extraction model
🤷‍♀️WHY: to take advantage of domain knowledge in the data for improved performance
🛠️CHALLENGE: leakage in the vocal stem

COLD SPECTROGRAM DIFFUSION (inspired by [2]



- We iteratively transform, in T steps, a mixture spectrogram to the corresponding vocals
- We train a convolutional U-Net network with skip connections, conditioned on T, following [3]
- In inference, we stack the intermediate steps in a feature matrix.

SEPARATION MASK ESTIMATION



- Given feature matrix, we cluster the frequency bins to build the separation mask
- **Vocal bins: +** energy overall, **-** energy change
- **Accomp bins: -** energy overall, **+** energy change
- **Shared bins:** middle range (we can select and remove more for more restrictive separation)

EVALUATION

- We record a clean testing set with two different singers and recording and mixing settings
- We run a controlled MUSHRA test on 25 subjects
- We select 6 tracks from the private Dunya database [4] with artist and audio quality diversity

	Mean Opinion Scores (MOS)	
	Vocal quality	Vocal isolation
Ours (*C*=5, *F*=4)	2.80 ± 0.29	**3.72** ± 0.31
Spleeter	**3.73** ± 0.17	1.97 ± 0.19

Our most *restrictive* model

Comparison of our system with *T*=8					
Config		vs. No clustering		vs. Spleeter	
C	*F*	SIRd	SARd	SIRd	SARd
2	1	+4.70	**-3.43**	+0.14	**-4.09**
3	1	+4.31	-3.56	+0.52	-4.22
3	2	+5.46	-4.49	+0.64	-5.16
4	2	+5.55	-4.71	+0.72	-5.38
4	3	+6.41	-5.53	+1.60	-6.20
5	2	+5.61	-4.69	+0.78	-5.35
5	3	+6.45	-5.59	+1.63	-6.26
5	4	**+7.14**	-6.25	**+2.32**	-6.91

The model is *adaptable*!

NOTE: By using and removing more clusters, we obtain a more restrictive separation, at the expense of source quality.

[1] A. Srinivasamurthy, S. Gulati, R. Caro, X. Serra, **"Saraga: Open Datasets for Research on Indian Art Music"**. Empirical Musicology Review, vol. 16, no. 1, pp. 85-98, 2021.
[2] A. Bansal, Em Borgnia, H. Chu, Jie S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein, **"Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise"**. Pre-print [under review], 2022.
[3] J. Ho, A. Jain, and P. Abbeel, **"Denoising Diffusion Probabilistic Models"**. In Proc. of the 33th Advances in Neural Information Processing Systems (NeurIPS), Online, pp. 6840–6851, 2020.
[4] A. Porter, M. Sordo and X. Serra, **"Dunya: A System to Browse Audio Music Collections Exploiting Cultural Context"**. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), 2013.

This work was carried out under the projects Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033 and NextCore - RTC2019-007248-7 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades (MCIU) and the Agencia Estatal de Investigación (AEI). We would like to thank the 40 participants that took the perceptual test for this work.

Universitat Pompeu Fabra Barcelona — MTG Music Technology Group — MUSICAL AI_ — nextcore — GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES