Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures

Zafar Rafii[1], François G. Germain[2], Dennis L. Sun[2,3], and Gautham J. Mysore[4]

[1]Northwestern University, Department of Electrical Engineering & Computer Science
[2]Stanford University, Center for Computer Research in Music and Acoustics
[3]Stanford University, Department of Statistics
[4]Adobe Research
zafarrafii@u.northwestern.edu, {fgermain,dlsun}@stanford.edu, gmysore@adobe.com

Introduction

Context

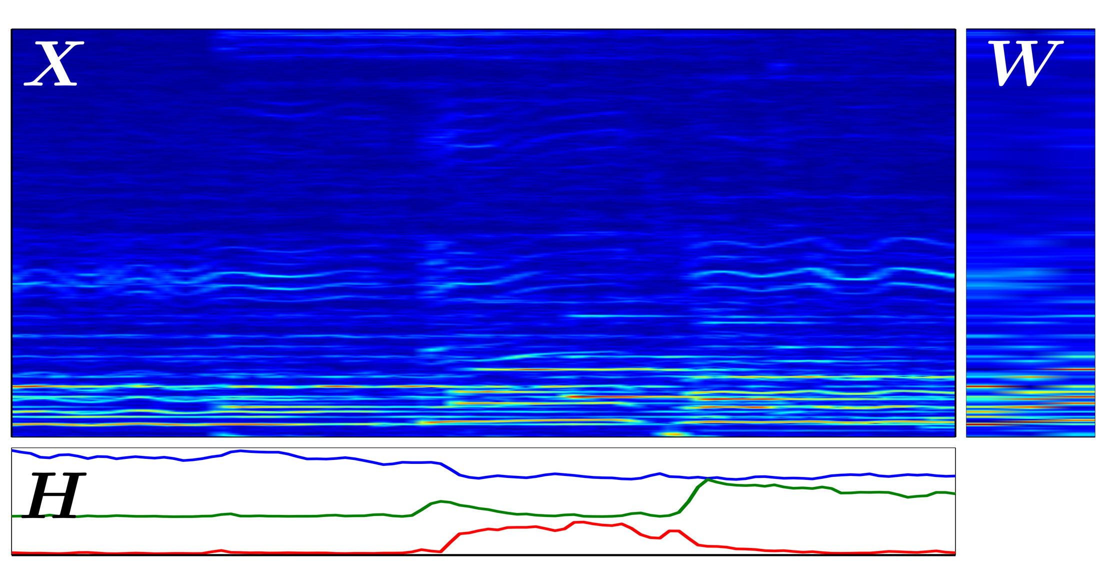
Many music/voice separation techniques focus on modeling one source and explain the other one as the residual, often resulting in degraded performance for the latter.

Proposed method

We propose to combine:

- A new Non-negative Matrix Factorization (NMF) based technique that explicitly models the voice without requiring singer-dependent training examples.
- An existing technique that explicitly models the background music.

Modeling the singing voice



We model the non-negative magnitude spectrogram $\boldsymbol{X}$ as a linear combination of non-negative basis vectors stored as columns of $\boldsymbol{W}$, with the rows of a matrix $\boldsymbol{H}$ representing the activations of the basis vectors over time, such that:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$$

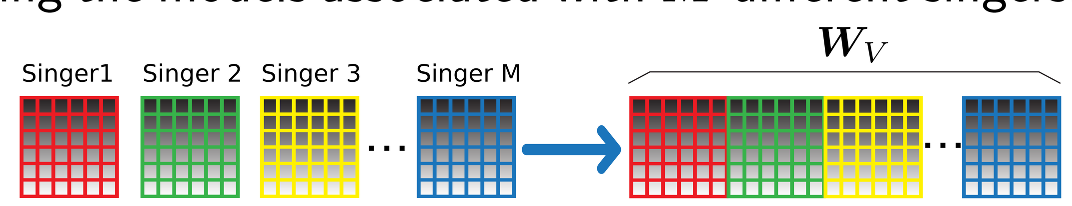
The factorization is learned from $\boldsymbol{X}$ through the minimization of a cost function $D\left(\boldsymbol{X}||\boldsymbol{W}\boldsymbol{H}\right)$.

The separation algorithm happens in two steps:

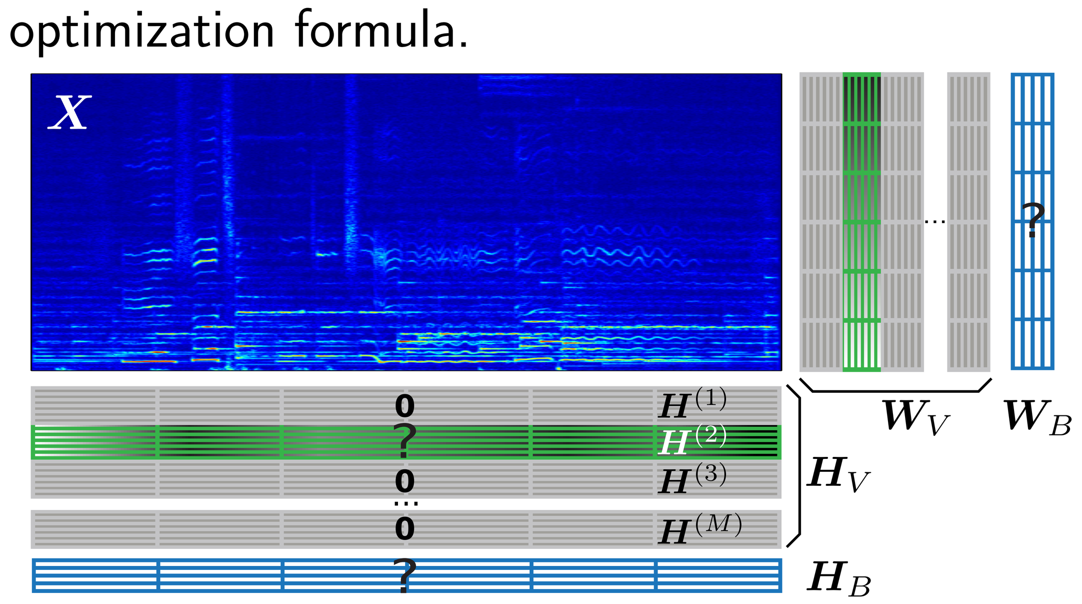
1. Train the basis vectors for the source models in $\boldsymbol{W}$, using *clean isolated* samples of each source.
2. Estimate the activations $\boldsymbol{H}$ of each source from the test samples, keeping $\boldsymbol{W}$ fixed.

Universal Voice Models (UVM)

Gathering isolated training samples for either the singer or the music is usually impractical. This limitation can be alleviated in the proposed approach by learning a ***universal voice model*** concatenating the models associated with M different singers.

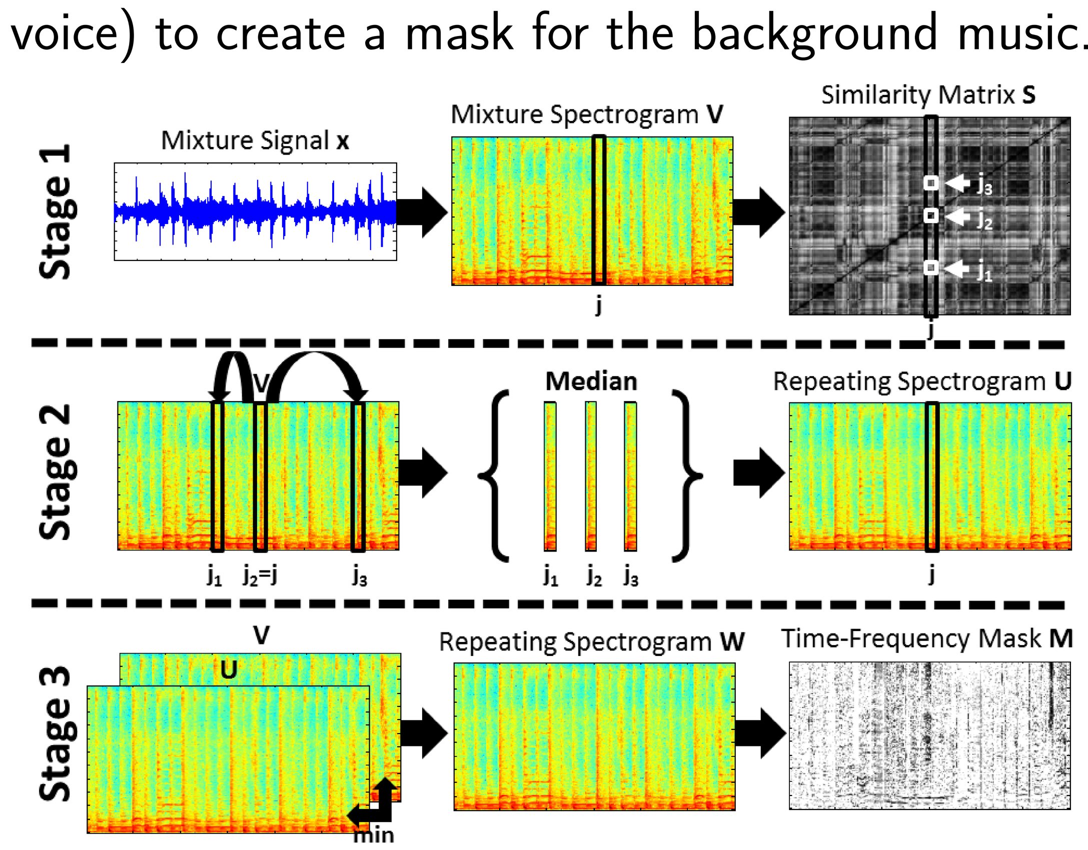


To capture the intuition that only a few singer models should be sufficient to explain an unseen singer, we assume that only a few of them in $\boldsymbol{W}$ are active by adding block sparsity penalty on $\boldsymbol{H}$ to the optimization formula.



Modeling the background music

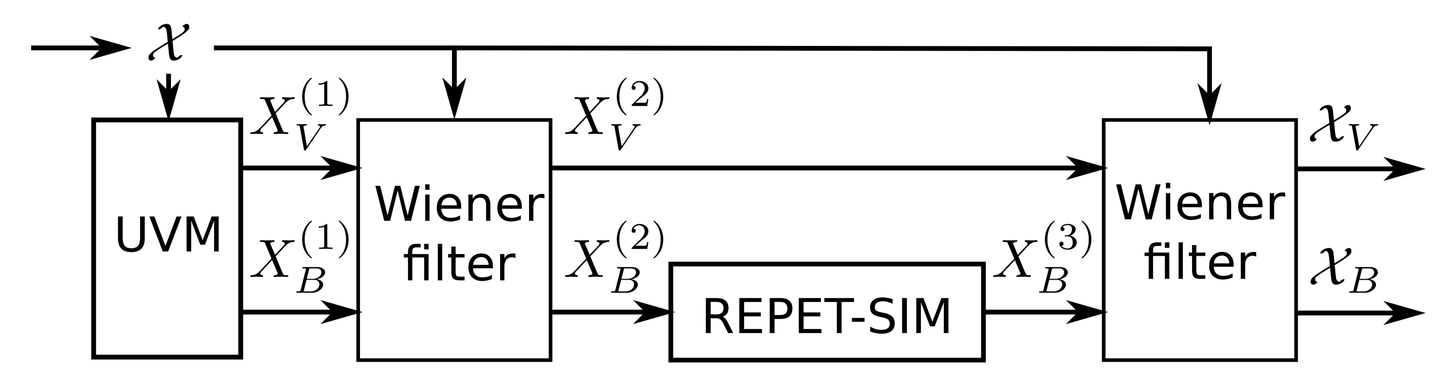
We use the REPET-SIM method, which focuses on extracting the background music. The separation is done by building a similarity matrix and smoothing the non-repeating patterns (assumed to be the voice) to create a mask for the background music.



Combined approach

The two presented methods are such that:

- The UVM approach explicitly models the singing voice.
- REPET-SIM explicitly models the background music.



We propose to combine those methods in order to leverage their ability to model each source in the mixture as follows:

1. Estimate $\boldsymbol{X}_V^{(1)} = \boldsymbol{W}_V\boldsymbol{H}_V$ and $\boldsymbol{X}_B^{(1)} = \boldsymbol{W}_B\boldsymbol{H}_B$ from the magnitude spectrogram $\boldsymbol{X}$.
2. Build new magnitude spectrogram estimates through Wiener filtering of the complex spectrogram $\mathcal{X}$:

$$\boldsymbol{X}_V^{(2)} = \left|\frac{\boldsymbol{X}_V^{(1)}}{\boldsymbol{X}_V^{(1)} + \boldsymbol{X}_B^{(1)}} \odot \mathcal{X}\right| \qquad \boldsymbol{X}_B^{(2)} = \left|\frac{\boldsymbol{X}_B^{(1)}}{\boldsymbol{X}_V^{(1)} + \boldsymbol{X}_B^{(1)}} \odot \mathcal{X}\right|$$

3. Refine the music estimate using REPET-SIM as $X_B^{(3)}$.
4. Build the final complex estimates through Wiener filtering:

$$\mathcal{X}_V = \frac{\boldsymbol{X}_V^{(2)}}{\boldsymbol{X}_V^{(2)} + \boldsymbol{X}_B^{(3)}} \odot \mathcal{X} \qquad \mathcal{X}_B = \frac{\boldsymbol{X}_B^{(3)}}{\boldsymbol{X}_V^{(2)} + \boldsymbol{X}_B^{(3)}} \odot \mathcal{X}$$

Experimental setup

Dataset

We evaluate our algorithm on the 1,000 song clips (at 16kHz) of the MIR-1K dataset, from 19 individual singers. The performance of the algorithm is done using a leave-one-out cross-validation:

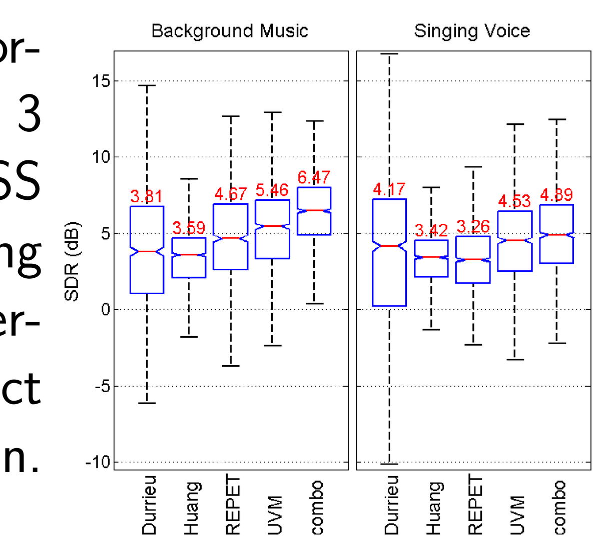
- The clips from one given singer are withheld as test set.
- The clips from the other singers are used to train a universal voice models with 18 dictionaries.

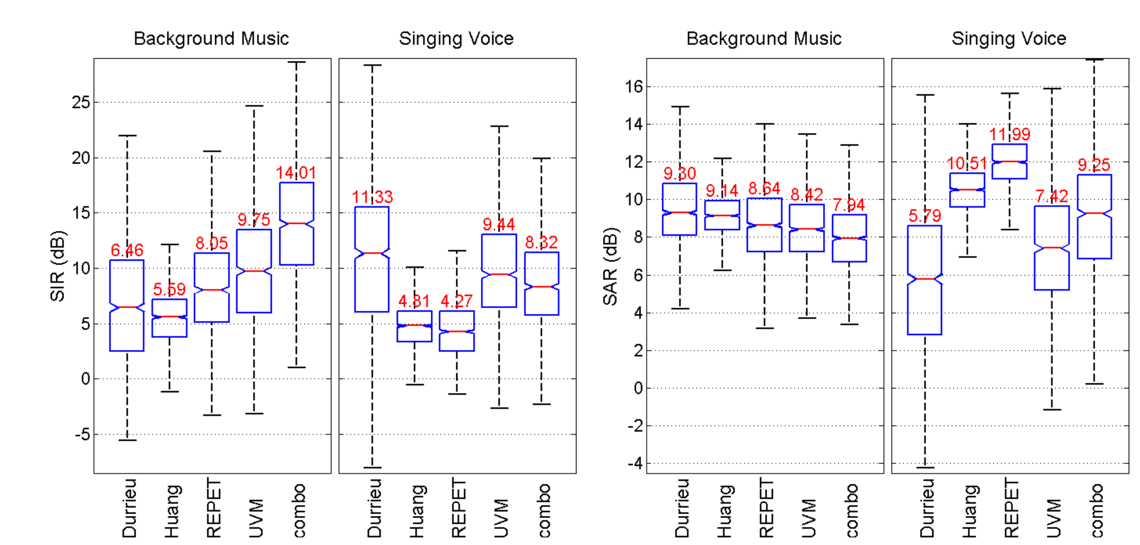
Competitor methods

- *Durrieu*: Source/filter voice model associated with an NMF-based background model learned iteratively from $\boldsymbol{X}$.
- *Huang*: Robust Principal Component Analysis assuming music as a low-rank component and voice as a sparse component.
- *REPET*: REPET-SIM alone.

Comparative results

To evaluate the performance, we use the 3 metrics from the BSS Eval toolbox measuring overall (SDR), interference (SIR) and artifact (SAR) related distortion.





Statistical analysis of the results shows that:

- The UVM approach achieves state-of-the-art results for music/voice separation in terms of Signal-to-Distortion Ratio.
- The combined approach further improves the Signal-to-Distortion Ratio scores.

Conclusion

We present here:

- The Universal Voice Models, a new NMF approach for music/voice separation modeling the voice without the need for specific training samples.
- A combined approach using the UVM and the REPET-SIM methods in order to leverage their complementary properties.
- Experimental results demonstrate that the two proposed methods achieve state-of-the-art results for overall music/voice separation performance.

This work was supported in part by NSF grant number IIS-0812314.