Spotify

Deep Architectures & Methods for Separating Vocals in Recorded Music

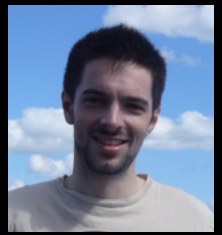
Eric J. Humphrey
ejhumphrey@spotify.com

DL Summit Montreal – 11 October 2017

MiQ @ Spotify

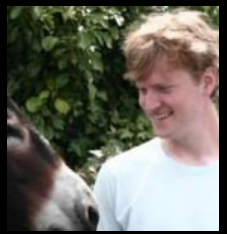
Advance the state of the art in **understanding music** at scale.

Rachel Bittner



Simon Durand

Eric Humphrey



Andreas Jansson

Tristan Jehan

Aparna Kumar

Nicola Montecchio

Music + AI?

Music is a fundamentally intelligent behavior.

Nature has produced some remarkable creatures:

- Cats and dogs recognize objects…
- Birds and whales sing…
- Bees dance, elephants never forget!

Only humans (as far as we can tell) make **music**.

Can we really study music without inadvertently studying **human** intelligence?

so what have we been up to?

“Source Separation,” you say?

Let’s decompile a music recording!

“Source Separation,” you say?

We could remove the vocalist ...

“Source Separation,” you say?

...or try to isolate them.

A quick recap

Spoiler alert: deep learning works well.

Approaches operate on time-frequency (spectral) representations and attempt to “mask” the components relevant to a given source.

Traditional methods included non-negative matrix factorization; recently, deep learning dominates the literature, e.g. Huang 2014, Simpson 2015, Luo 2017.

A quick recap

Spoiler alert: deep learning works well.

Approaches operate on time-frequency (spectral) representations and attempt to “mask” the components relevant to a given source.

Traditional methods included non-negative matrix factorization; recently, deep learning dominates the literature, e.g. Huang 2014, **Simpson 2015**, Luo 2017.

A View from **Emerging Technology from the arXiv**

Deep Learning Machine Solves the Cocktail Party Problem

not quite...

Separating a singer’s voice from background music has always been a uniquely human ability. Not anymore.

April 29, 2015

The cocktail party effect is the ability to focus on a specific human voice while filtering out other voices or background noise. The ease with which humans perform this trick belies the challenge that scientists and engineers have faced in reproducing it synthetically. By and large, humans easily outperform the best automated methods for singling out voices.

A particularly challenging cocktail party problem is in the field of music, where humans can easily concentrate on a singing voice superimposed on a musical background that includes a wide range of instruments. By

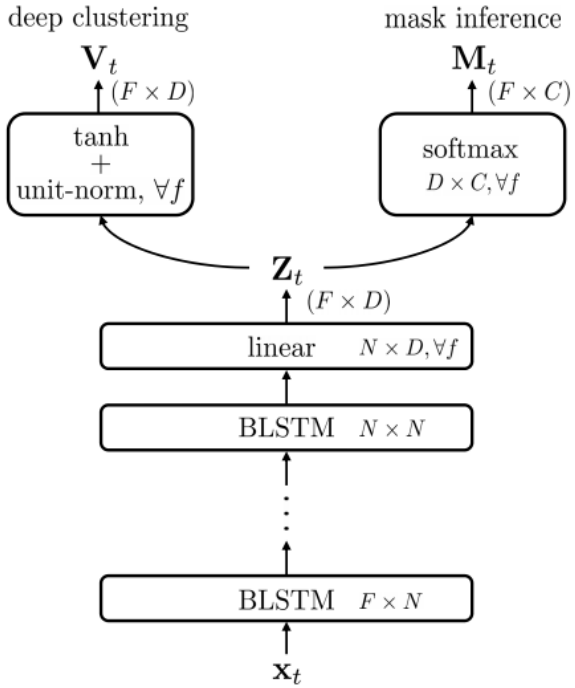
A quick recap

Spoiler alert: deep learning works well.

Approaches operate on time-frequency (spectral) representations and attempt to “mask” the components relevant to a given source.

Traditional methods included non-negative matrix factorization; recently, deep learning dominates the literature, e.g. Huang 2014, Simpson 2015, **Luo 2017.**

Luo’s “Chimera” model (right) combines multiple learning objectives to achieve state of the art results.



So ... what's the hold up?

There are still a few challenges.

Data: Most research is achieved with only a few hours of recorded audio (right).

Metrics: Source separation measures are imperfectly aligned with human perception [Emiya 2011].

Models: DL approaches are data hungry; innovative architectures and designs are actively being explored, e.g. resnets, skip-connections, recurrence, etc.

Standard datasets for music source separation.	
DSD100 (SISEC)	100 professionally mixed tracks, variety of genres (≃5hr)
MedleyDB	70 professionally mixed tracks, variety of genres (≃4hr)
iKala	252 30-second Karaoke tracks with amateur performers (≃2hr)

…but maybe there is data?

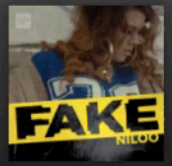
Browse
Radio

YOUR LIBRARY
Your Daily Mix
Recently Played
Songs
Albums
Artists

New Playlist

spotify:track:6GK0EvOj5Ai28E

Eric Humphrey



SINGLE

Fake

By Niloo

November 27, 2013 • 4 songs, 14 min

PLAY SAVE ···

#	TITLE	
1	Fake	3:34
2	Fake (Instrumental)	3:34
3	Fake (Acapella)	3:34
4	Fake (Acoustic)	3:28

More from this label

More by Niloo

Fake
Niloo

0:15 3:33

Oh right – instrumental versions!

Turns out instrumental (and acapella) versions are … not uncommon.

Find track sets with the following logic:

- A and B are recorded by the same artist.
- “instrumental” does not appear in the title of A.
- “instrumental” does appear in the title of B.
- The titles of A and B are fuzzy matches.
- The track durations differ by less than a second.

Use fingerprinting algorithm to improve results:

- ≃ 1 in 10 are metadata confusions
- Discard extreme cases, too similar / different
- Results in 12k pairs, or ≃750 hours of audio (10^2x increase!)

Genre	Percentage
Pop	26.0%
Rap	21.3%
Dance & House	14.2%
Electronica	7.4%
R&B	3.9%
Rock	3.6%
Alternative	3.1%
Children’s	2.5%
Metal	2.5%
Latin	2.3%
Indie Rock	2.2%
Other	10.9%

Input features

Short-time Fourier Transform (STFT) features

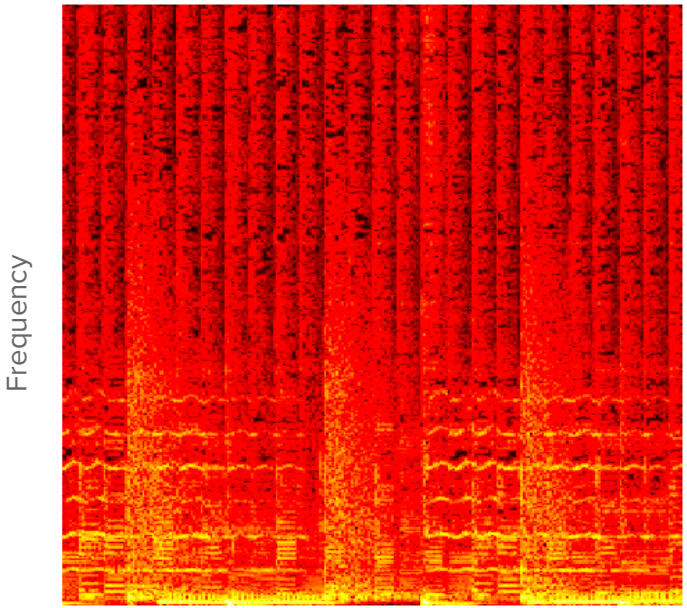
Transforms audio into an **invertible** complex valued time-frequency representation, called a **spectrogram**

Phase is preserved; only magnitudes are processed by the model, making resynthesis straightforward.

Approximate the target vocals from track pairs:

```
voice = max(mix - instrumental, 0)
```

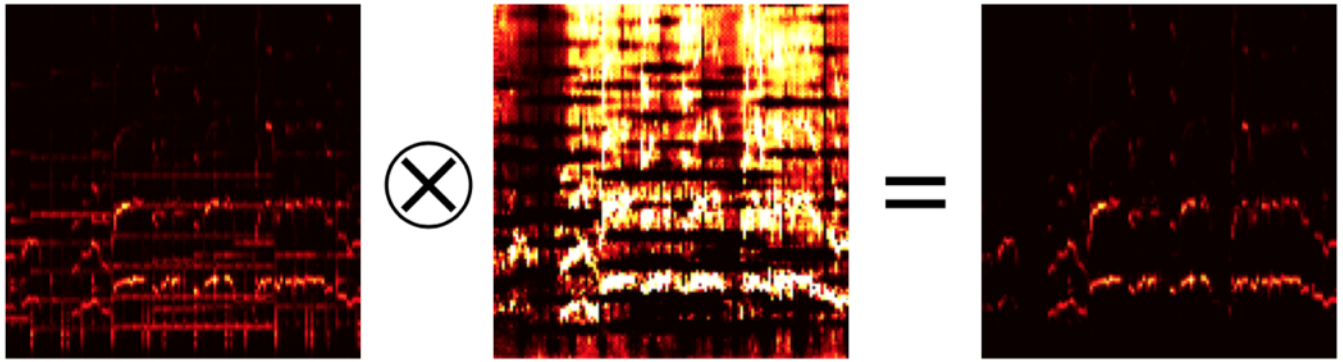
Initial experiments used low resolution features for speed (8kHz), but scales well higher quality audio.



Learning to mask

Goal: Input ⊗ Mask = Target Source

Learn a function that suppresses time-frequency coefficients that don't correspond to the target source.



Input Estimated Mask Vocals

Learning to mask

Adapting the U-Net architecture

Draws inspiration from modern architectures in computer vision [Ronneberger 2015, Isola 2016]

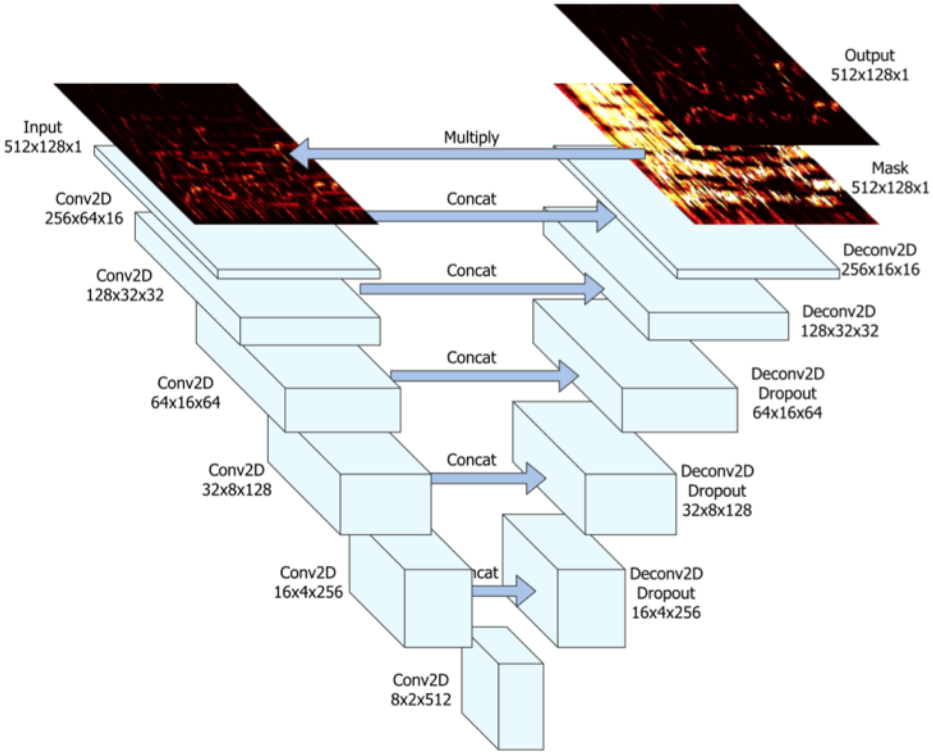
Operates on STFT patches of ≂10 seconds.

Fed through convolutional “hourglass” architecture, e.g. commonly used for auto-encoders.

Skip connections concatenate encoder feature maps with decoder feature maps.

Trained by minimizing the L1 loss between the target and masked input (estimate).

Independent networks are trained per source, i.e. vocal and background.

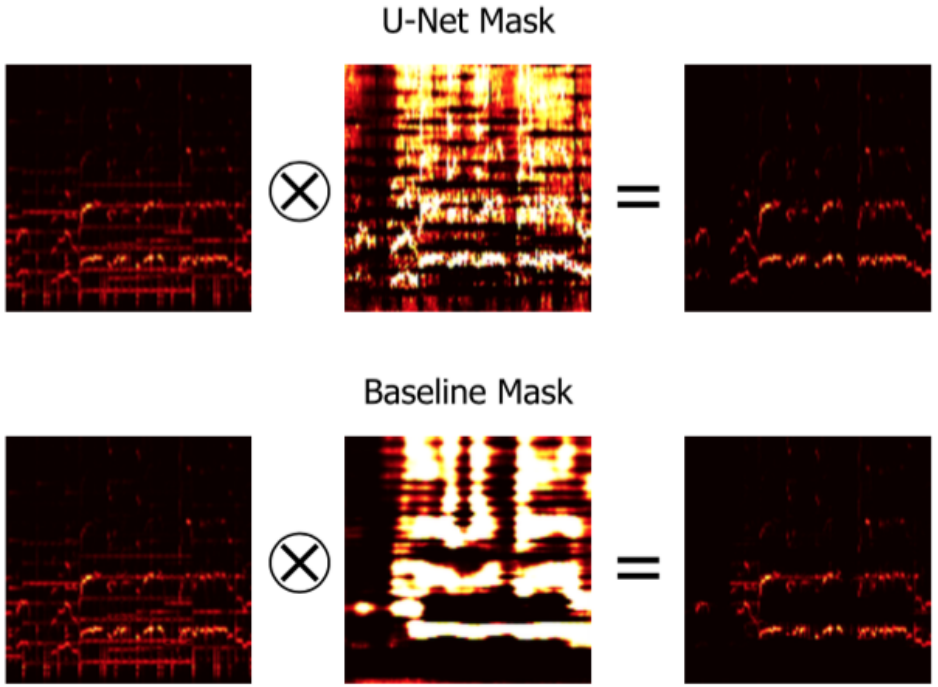


Architectural insights

Do skip connections really matter?

It turns out that, **yes**, skip connections provide additional detail for masking vocals in a mix.

A traditional auto-encoder architecture struggles to produce high-resolution masks necessary for manipulating acoustic scenes.

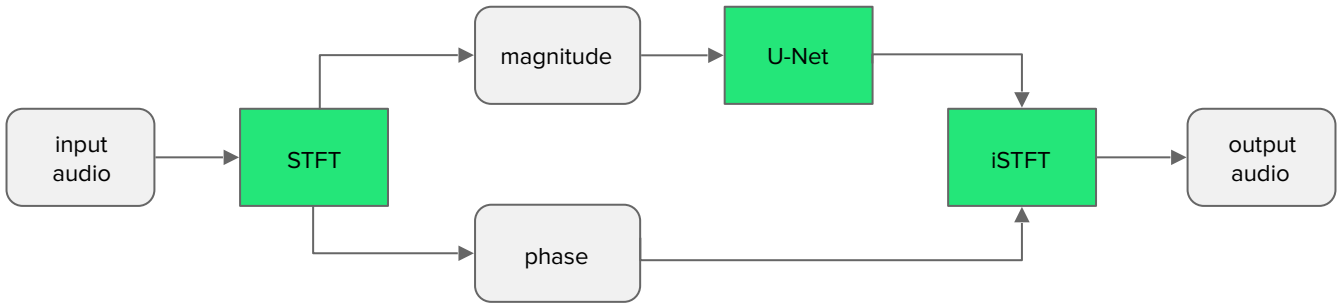


Inference

How is the model used to process new audio?

A trained model is applied convolutionally over time to an input spectrogram.

The masked spectrogram is then recombined with the input phase and transformed back into the time domain.



Quantitative evaluation

Signal-level statistics provide insight into how well a source separation algorithm performs.

Here we consider three common measures:

- Signal to Distortion Ratio (SDR)
- Signal to Interference Ratio (SIR)
- Signal to Artifact Ratio (SAR)

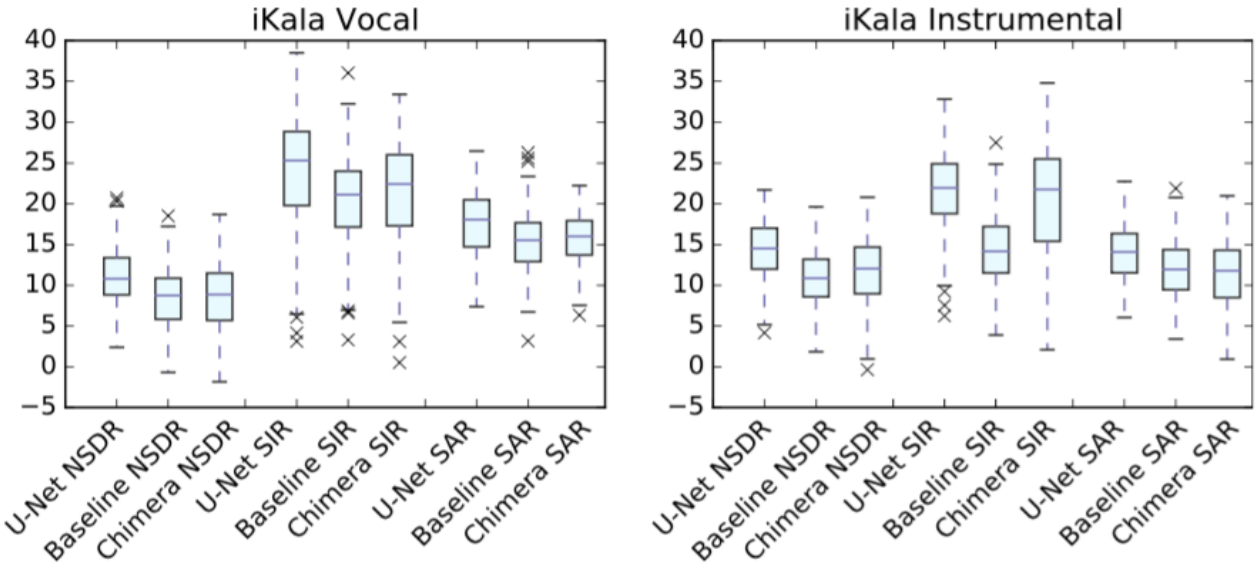
$$\mathrm{SDR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}}\|^2}{\|e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}\|^2}$$

$$\mathrm{SIR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}}\|^2}{\|e_{\mathrm{interf}}\|^2}$$

$$\mathrm{SAR} := 10 \log_{10} \frac{\|s_{\mathrm{dist}} + e_{\mathrm{interf}} + e_{\mathrm{noise}}\|^2}{\|e_{\mathrm{artif}}\|^2}$$

Quantitative evaluation

tl;dr: Slightly surpasses state of the art, and considerably better than the auto-encoder baseline.



Qualitative evaluation

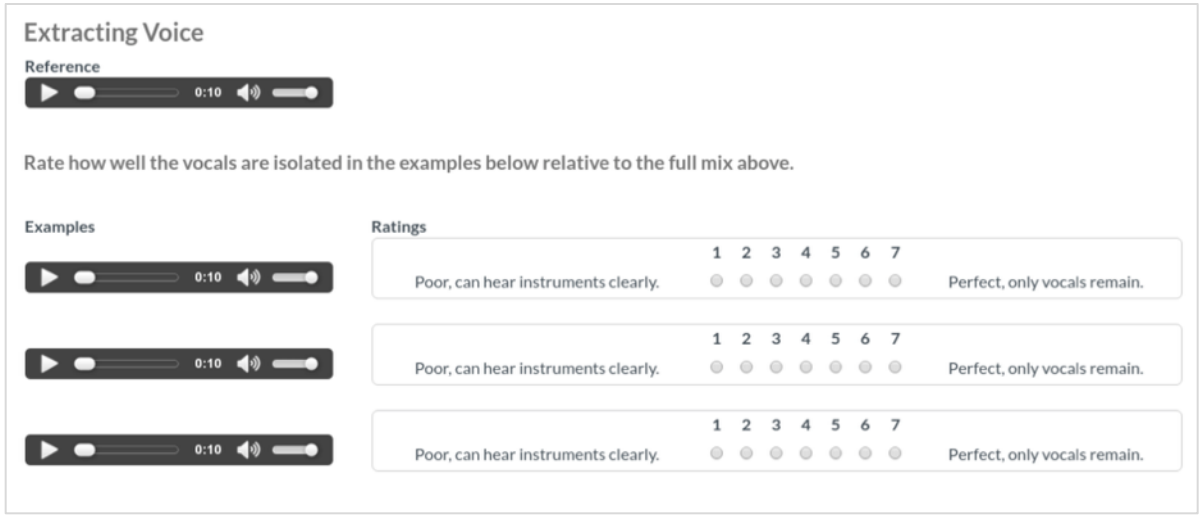
Metrics are better / on par with state of the art, but...

SDR, SIR, and SAR are actually not that great performance measures. They don't really match human perception very well.

We complement our quantitative offline evaluation with a subjective survey on Crowdflower, a paid crowdsourcing platform.

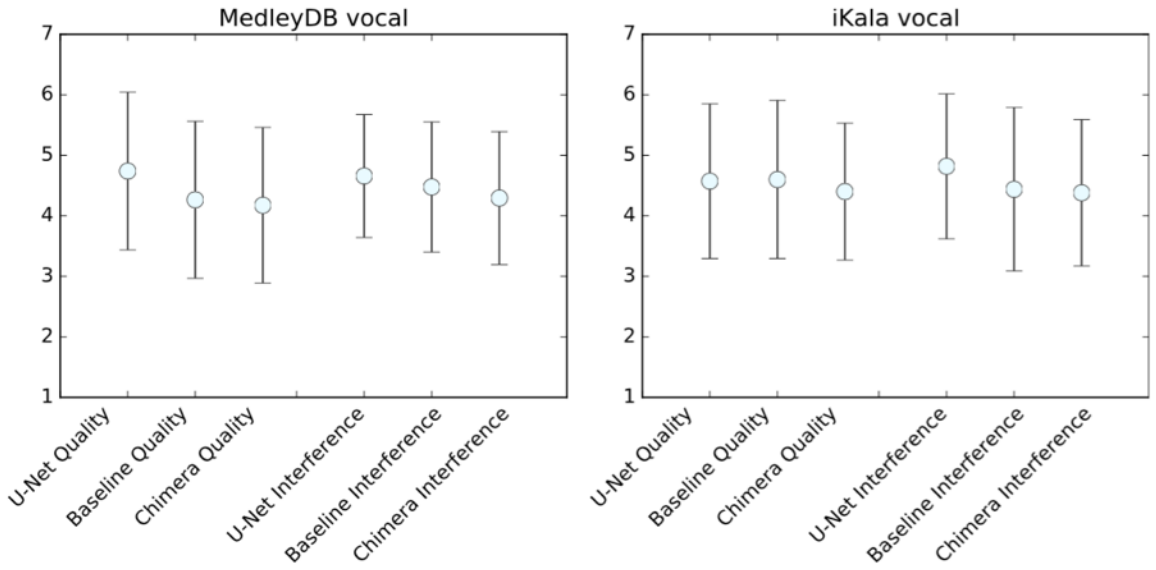
Qualitative evaluation

Around 100 contributors provided 1500 ratings.



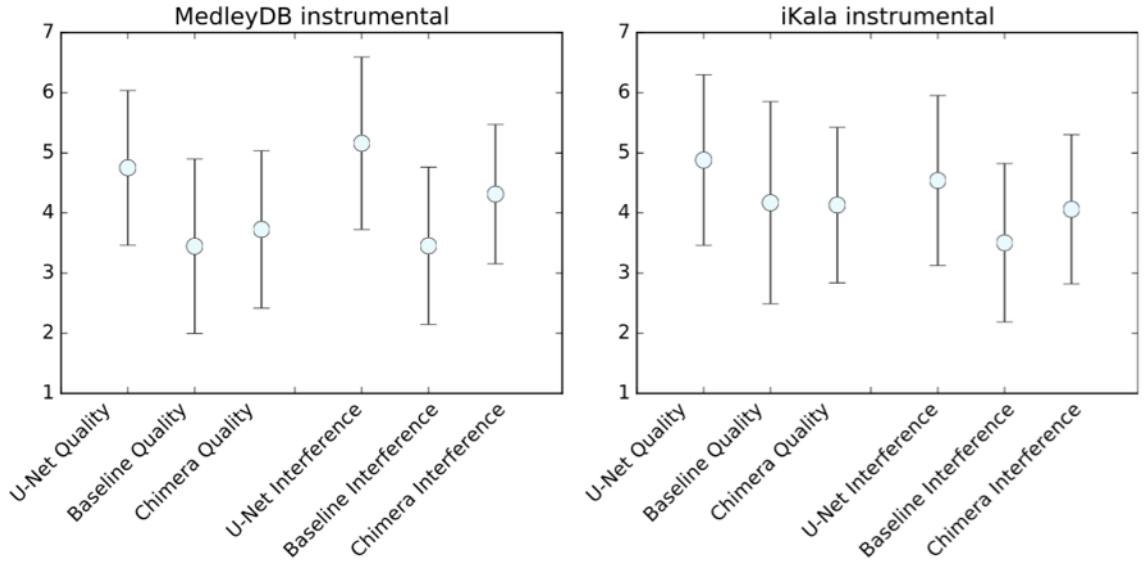
Qualitative evaluation

Ratings for isolated vocals are inconclusive...



Qualitative evaluation

..but instrumental separation scores much higher.



examples

takeaways

If you forget everything else...

Source separation in music is **not solved (just) yet**, but the future is bright.

Machine learning research that generates audio still needs **good perceptual measures** of quality.

There is plenty of data to use for some deep learning applications – if you **know where to look**.

Interesting problems **won't always have data** lying around for supervised learning – and then what?

The path to **humane artificial intelligence** almost certainly goes through music.

Spotify

thanks! / questions?

(also we're growing 😉)

@ejhumphrey // ejhumphrey@spotify.com