Spotify

Learning a Large-Scale Vocal Similarity Embedding for Music

Aparna Kumar
aparna@spotify.com
MiQ Lab : Spotify

Music Discovery - ICML 2017

MiQ @ Spotify

To advance the state of the art in understanding music content at scale

Aparna
Kumar

Rachel
Bittner

Nicola
Montecchio

Andreas
Jansson

Maria
Panteli

Eric
Humphrey

Tristan
Jehan

Vocal styles in music are diverse

What makes each vocal style similar, or different from the previous?

Vocal Similarity

*Can we **characterize** vocal style?*

*Can we **measure** vocal style similarity at scale?*

Learning a Large-Scale Vocal Similarity Embedding for Music

1. Vocal style representations
2. Train a supervised embedding
3. Evaluate the embedding space
4. New labels for improving the embeddings

In this talk...

Do distances between groups of songs change across vocal style embedding spaces?

Do distances in the embedding spaces represent perceptually distinguishable vocal styles?

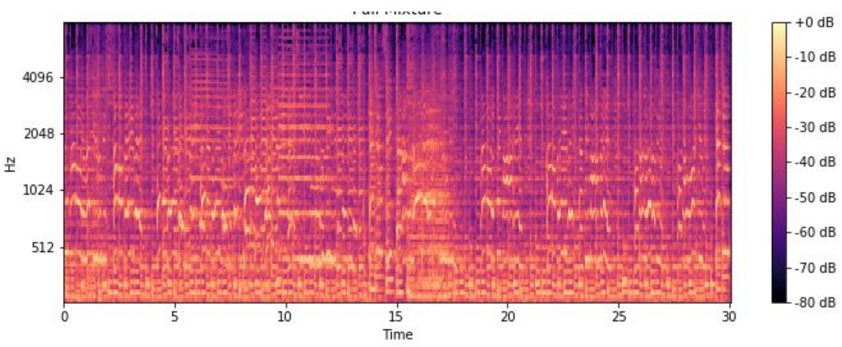
Are we under-evaluating vocal style by using artist similarity?

Vocal Representations

Vocal representations to learn singing style embeddings

1. **Full Mixtures**
2. Source Separated Vocals
3. Source Separated Instrumentals/Backgrounds
4. Vocal F0 + Timbre
5. Vocal F0

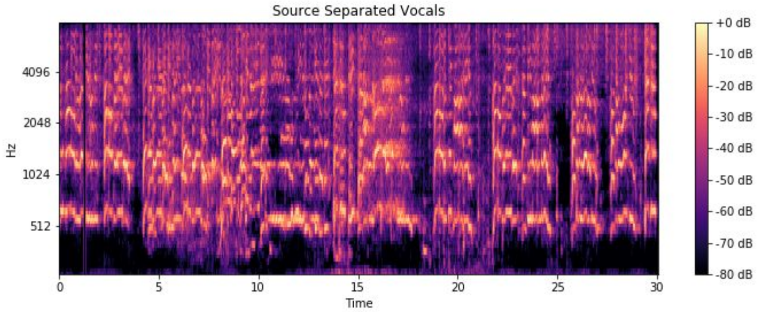
Full mixture



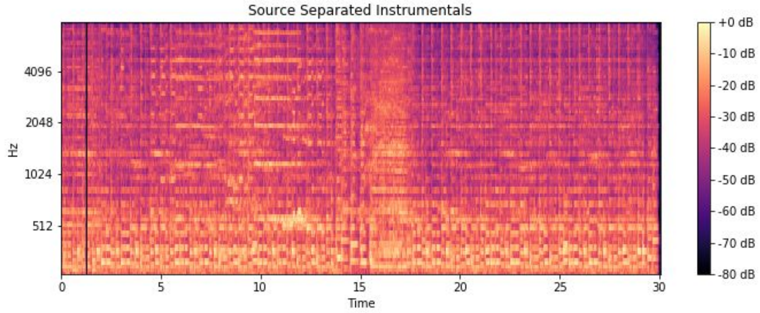
Vocal representations to learn singing style embeddings

1. Full Mixtures
2. **Source Separated Vocals**
3. **Source Separated Instrumentals/Backgrounds**
4. Vocal F0 + Timbre
5. Vocal F0

Vocals

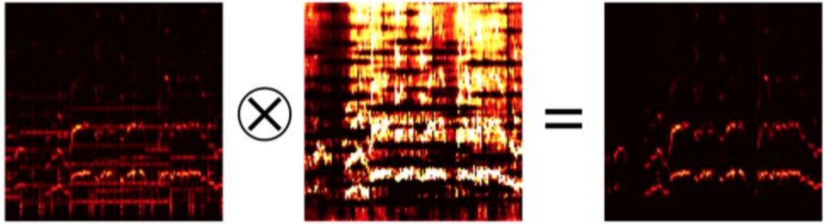


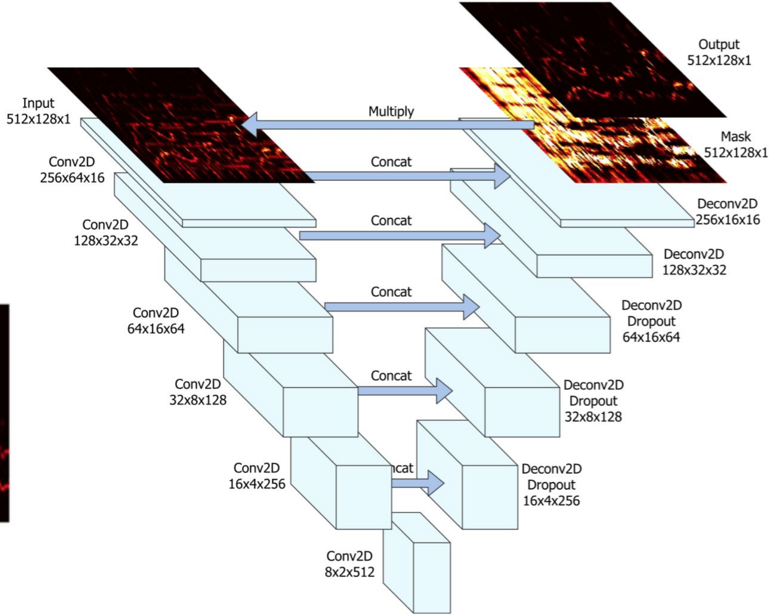
Instrumental



Vocal Source Separation

Convolutional encoder-decoder with skip connections is trained to predict spectral mask



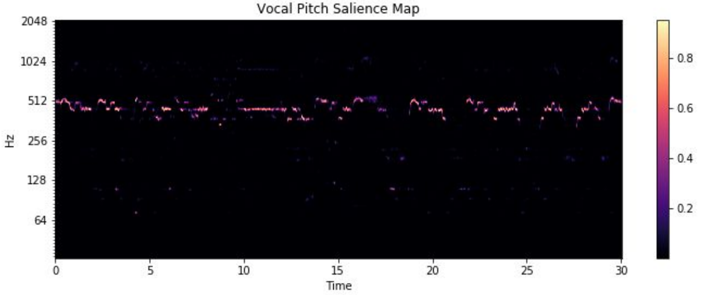


[1] Singing Voice Separation with Deep U-Net Convolutional Networks
Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar and Tillman Weyde
18th International Society for Music Information Retrieval (ISMIR) conference

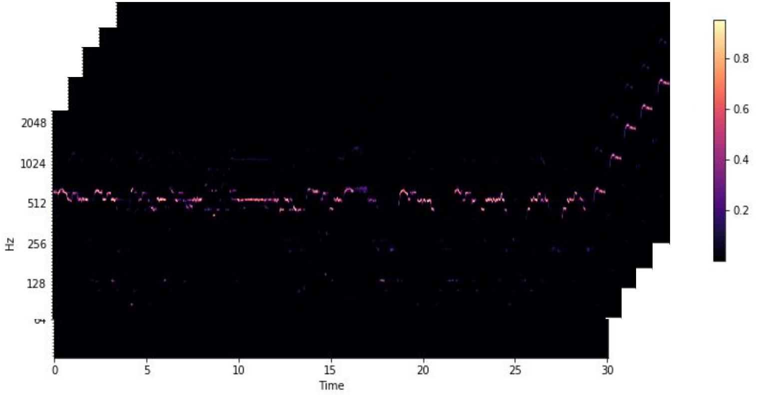
Vocal representations to learn singing style embeddings

1. Full Mixtures
2. Source Separated Vocals
3. Source Separated Instrumentals/Backgrounds
4. **Vocal F0 + Harmonics**
5. **Vocal F0**

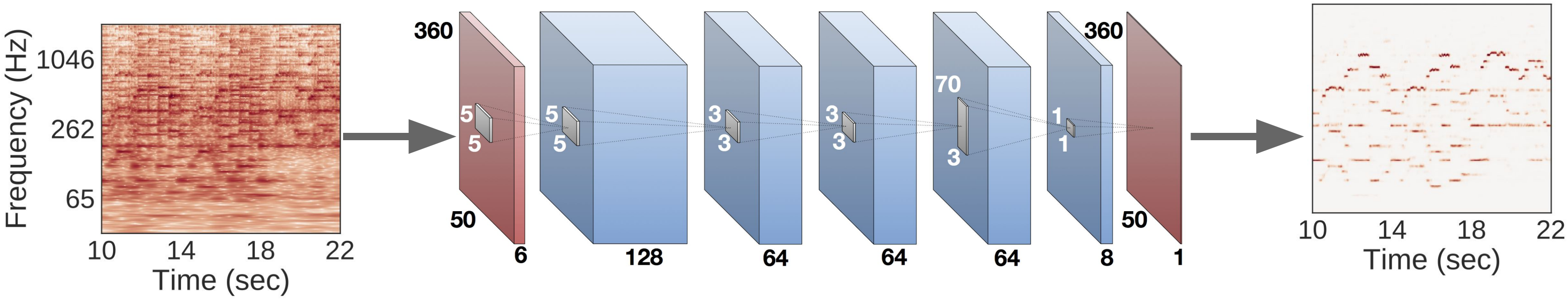
VPSM

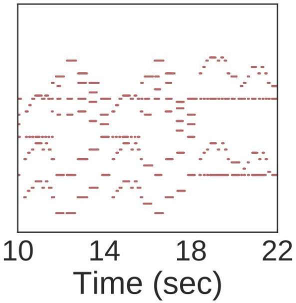


H-VPSM



Vocal F0 Estimation





[2] Deep Salience Representations for F0 Estimation in Polyphonic Music
Rachel M. Bittner, Brian McFee, Justin Salamon, Peter Li and Juan Pablo Bello
18th International Society for Music Information Retrieval (ISMIR) conference

Vocal representations to learn singing style embeddings

Vocal Representations

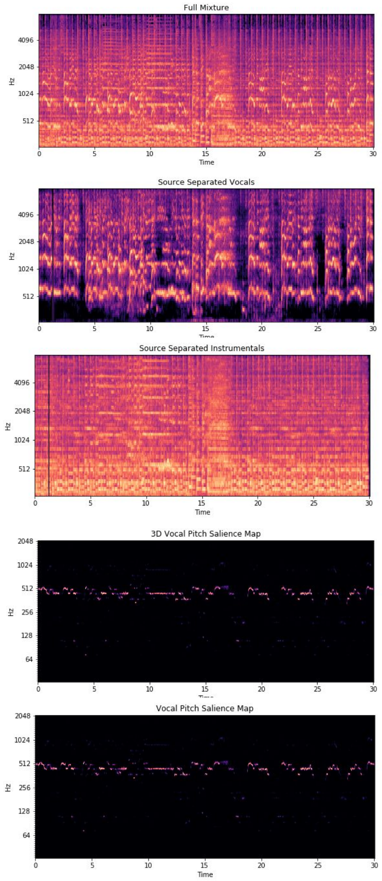
Full mix

Vocal

Instrumental

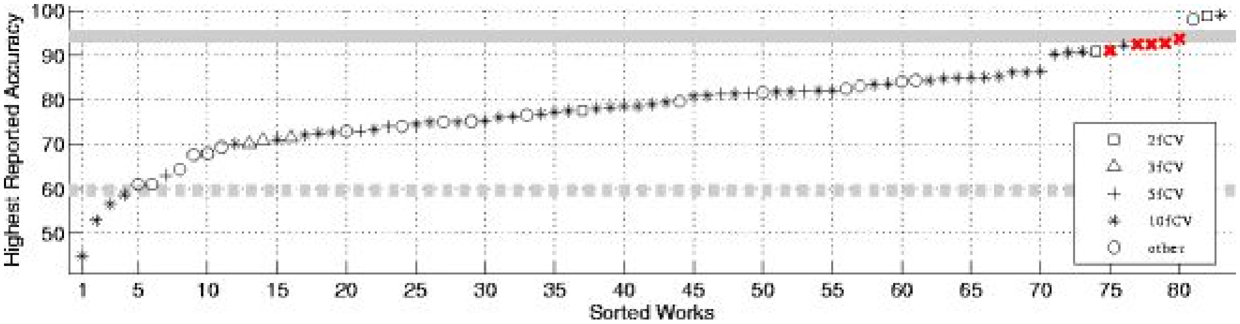
H-VPSM

VPSM



Genre as a Proxy for Style

Genre Classification, eh?



[4] Classification is Not Enough
Bob L. Sturm
Journal of Intelligent Information Systems, Vol 41, Iss 3, pp 371-406, 2013

LABELS

Supervised learning for a vocal style embedding

Let’s say, “genre” is an expert-defined cluster of artists who perform similar styles of music.

ASSUMPTION : singing voice characteristics are correlated with genre

DATA

10 genres, 100 artists each, 20 tracks each

Chicago Soul	**Rock-and-roll**
Jazz Blues	**Skate punk**
Pop	**Soul**
R&B	**Southern hip hop**
Rock	**Trap music**

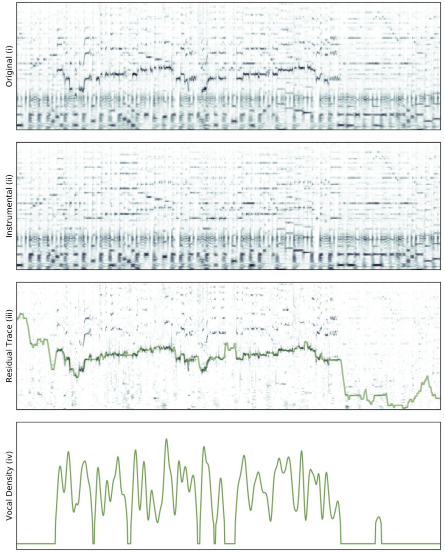
Mining Vocal Activity from the Catalogue

Idea: Exploit pairs of original and instrumental recordings for MIR tasks.

- Form pair candidates from metadata
- Align feature representations via DTW
- Track positive difference (residual) with Viterbi decoding
- Use amplitude of best path as a vocal *density* estimate

Evaluate quality of signal by benchmarking on voice activity detection.

- Strongly labeled set of 12*k* pairs
- Matches state of the art without manual labeling
- Creates ML opportunities for various vocal tasks



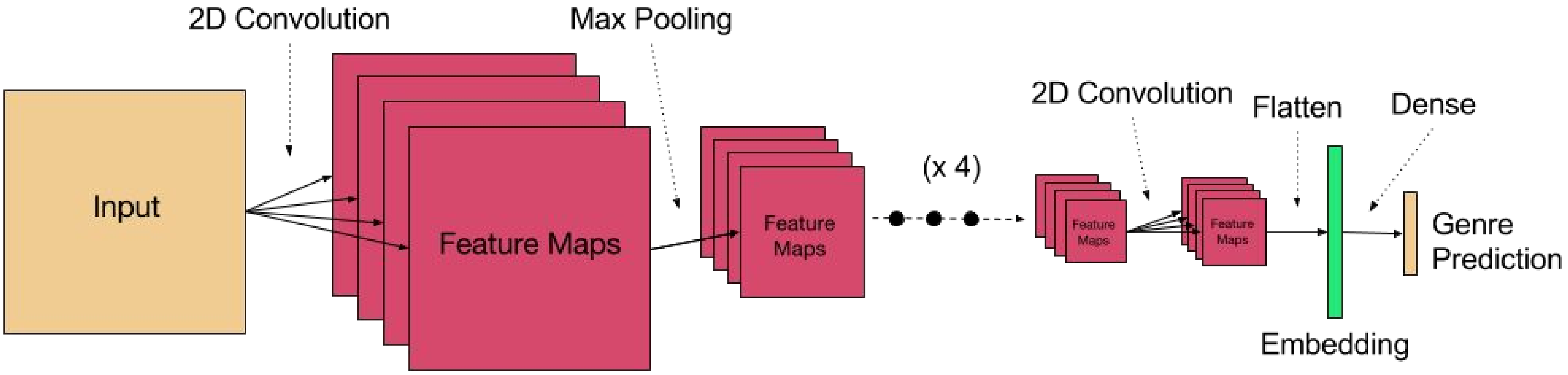
[3] Mining Labeled Data from Web--Scale Collections for Vocal Activity Detection in Music
Eric J. Humphrey, Nicola Montecchio, Andreas Jansson, Rachel M. Bittner and Tristan Jehan
18th International Society for Music Information Retrieval (ISMIR) conference

Convolutional network for learning genre-constrained vocal style embeddings

Understanding the singing voice

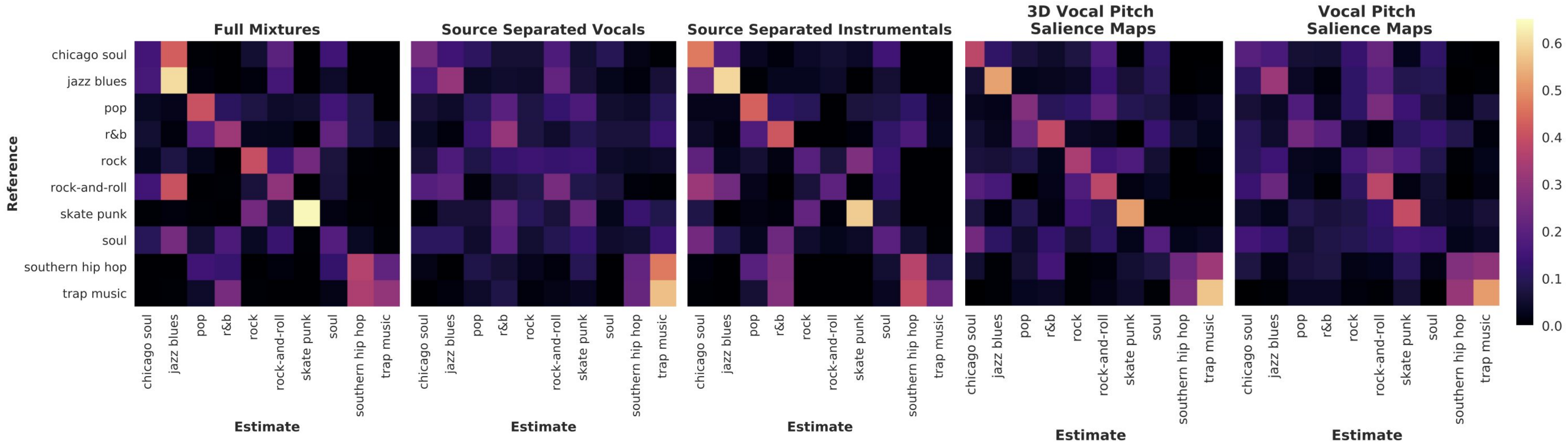
Convolutional net for learning genre-constrained vocal style embeddings

The embedding layer



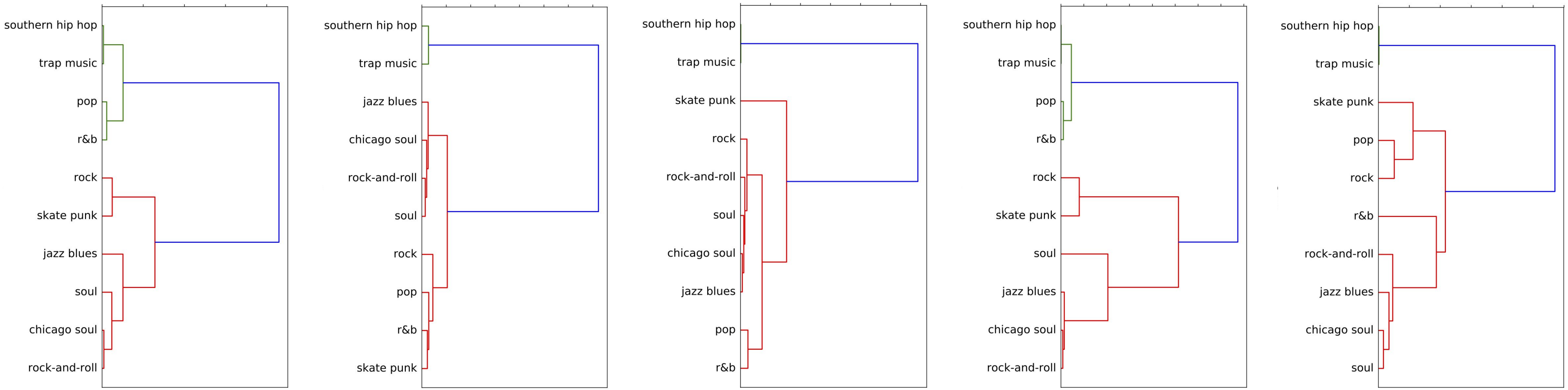
Embedding spaces

Learning genre-constrained embeddings



Embedding spaces capture many different relationships between genres

Genre relationships across the embedding spaces :



Evaluating the embeddings

Understanding the singing voice

Validating the embeddings : is artist retrieval enough?

Objective evaluation

Artist retrieval is used in MIR to evaluate vocal style similarity [4,5,6]

Embeddings perform at least 10 times better than random, measured by artist retrieval.

[4] Kim & Whitman, 2002;
[5] Nakano et al., 2014;
[6] Fujihara et al., 2010

Validating the embeddings : is artist retrieval enough?

Perceptual evaluation via crowdsourcing

1. Randomly select a subset of artists whose tracks are disparate.

2. In triplicate comparisons : query whether a track has vocal styles like its neighbors, or whether it is more like non-neighboring tracks from the same artist

Validating the embeddings : is artist retrieval enough?

Perceptual evaluation via crowdsourcing

1. Randomly select a subset of artists whose tracks are disparate.

2. In triplicate comparisons : query whether these tracks have vocal styles like their neighbors, or like non-neighboring tracks from the same artist

Findings

1. Tracks share similar vocal styles with neighbors that are not present in non-neighboring tracks by the same artist, as assess through crowdsourcing.

Conclusions

Artist retrieval scores are not enough to evaluate vocal style similarity.

Tracks have similar vocal styles with their neighbors, which are not present in non-neighboring tracks by the same artist.

R.E.M. is in Cluster 1

R.E.M. also is in Cluster 2

Are vocal characteristics really correlated with genre? Can we do better?

Understanding the singing voice

Vocal attributes in music

Raspy	Nasal	Simple	Clear	Thin
Robotic	Natural	Deep	Shrill	Spoken
Rhythmic	Heavy Audio Effects	Shrill	Ornamented	Choppy
Whispering	Shouting	Full	Breathy	Calm
Old	Young	Controlled	Somber	Energetic

Rank consistency to measure the quality of *crowdsourced* vocal attribute labels

Pairwise comparisons and rank aggregation result in consistent labels for sorted, discretized, and binary lists.

Consistency score of the recovered rank over over three experiments is used to assess attribute label quality.

Poor reproducibility suggests noisy labels.

Attribute	Rank consistency
Raspy	0.83
Masculine	0.84
Feminine	0.85
Rhythmic talking	0.85
Powerful	0.72
Natural	0.81

Active learning and pairwise rank aggregation to label 10K tracks with vocal attributes

Ranking at scale is expensive!

Active learning to improve the pairwise rank aggregation

- identify query with highest expected information gain.
- send the query to the crowd platform.
- crowd platform pushes query responses back to the server.
- re-rank and iterate

[4] Pairwise rank aggregation in a crowdsourced setting
Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, Eric Horvitz
13th International Conference on Web Search and Data Mining (WSDM)

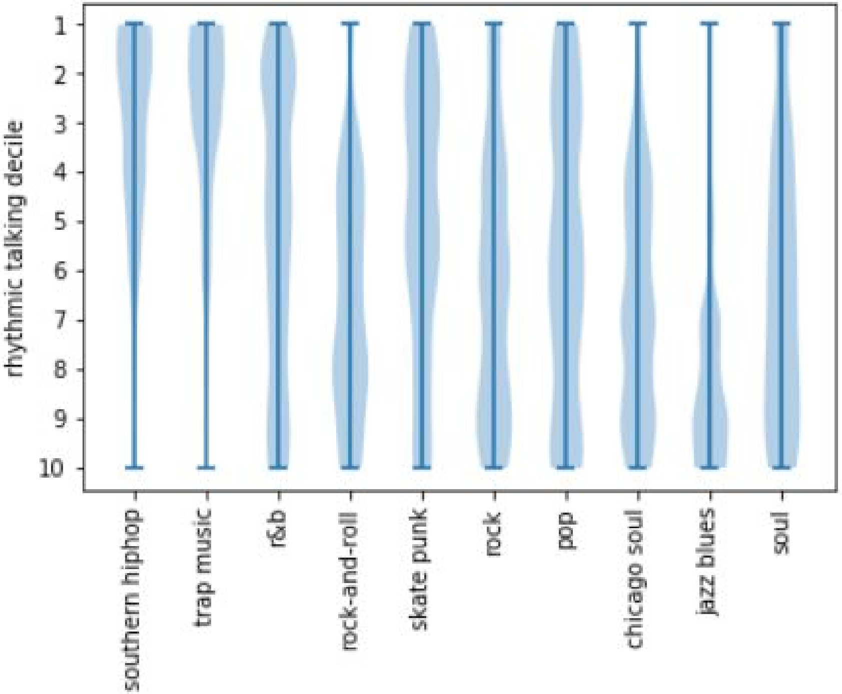
Genre distributions for 10K tracks ranked by *“Rhythmic Talking”*

“Rhythmic talking” appears in many tracks of *Skatepunk*, *Trap music*, *Southern hip-hop* and *R&B*.

As expected, *Soul, Jazz, Chicago soul* have few occurrences of *“rhythmic talking”*.

Vocal attribute labels can provide additional information to refine embedding spaces.

We can rank tracks by attributes that are not correlated with genre, like “*raspy*”, not shown.

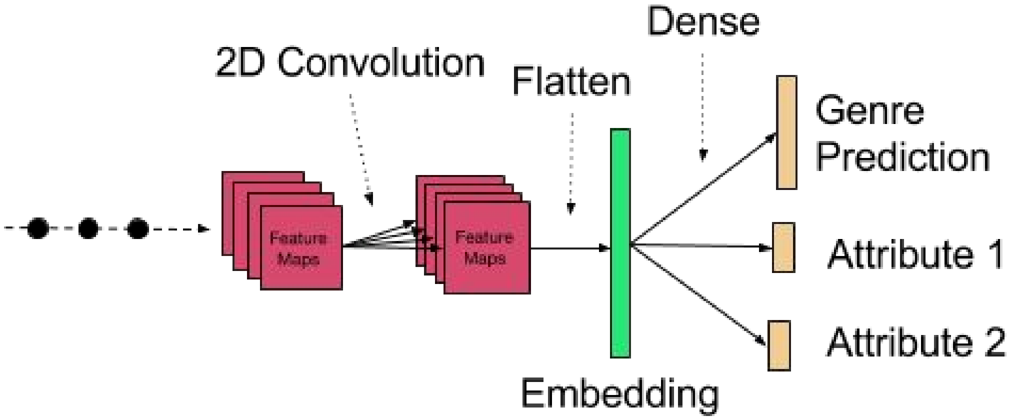


Ongoing work...

Understanding the singing voice

How will genre-independent attributes change the embedding space?

Ongoing work...



Vocal style similarity in action

Ongoing work...

“Spotify, play me a song by a ***female*** artist with a ***powerful***, ***raspy voice*** who is not ***American***”

Search

Playlist ordering

Recommendations

Categorization

THANK YOU!!

Learning a Large-Scale Vocal Similarity Embedding for Music

Aparna Kumar
aparna@spotify.com

MiQ Lab : Spotify