A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures

Adam Berenzweig
Dan Ellis

LabROSA
Columbia University

Beth Logan

HP Labs

Brian Whitman

Music, Mind & Machines
MIT Media Lab

Motivation

- Similarity is at the heart of:
  - Classification
  - Content-based Music Information Retrieval
  - Recommendation
  - Similarity Browsing
- Similarity? says who? (Evaluation is hard.)
  - Subjective
  - Context-dependent (mood, time of day)
  - Similarity how? rhythm, melody, singing voice, lyrics?

Evaluation

- Music IR needs TREC-like framework: Common corpus, common evaluation.
  - Acoustic data. Copyright Hell.
  - Evaluation methodology. "The quest for ground truth continues"
- Our solution:
  - Truth = Aggregate various sources of human subjective judgments
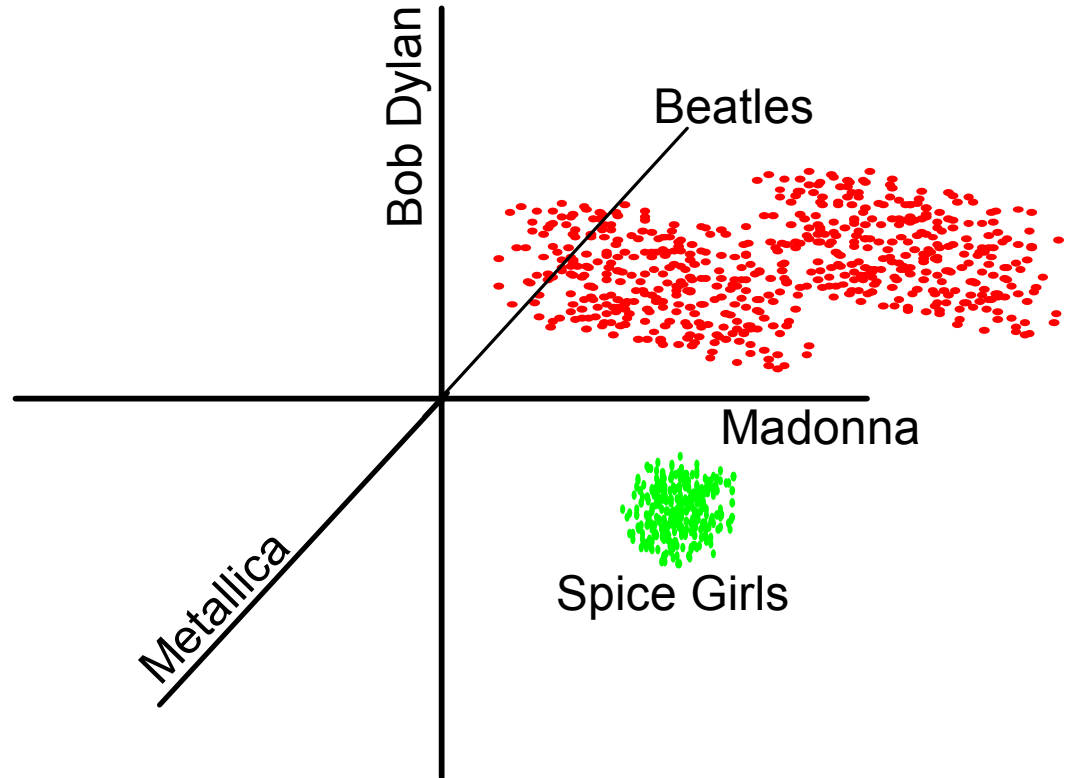 - Share features, not music.

- **Acoustic Measures**
- Subjective Measures
- Scoring Methods
- Results

Feature Space

- Artists/Songs are distributions, not points.
  - Model with GMMs
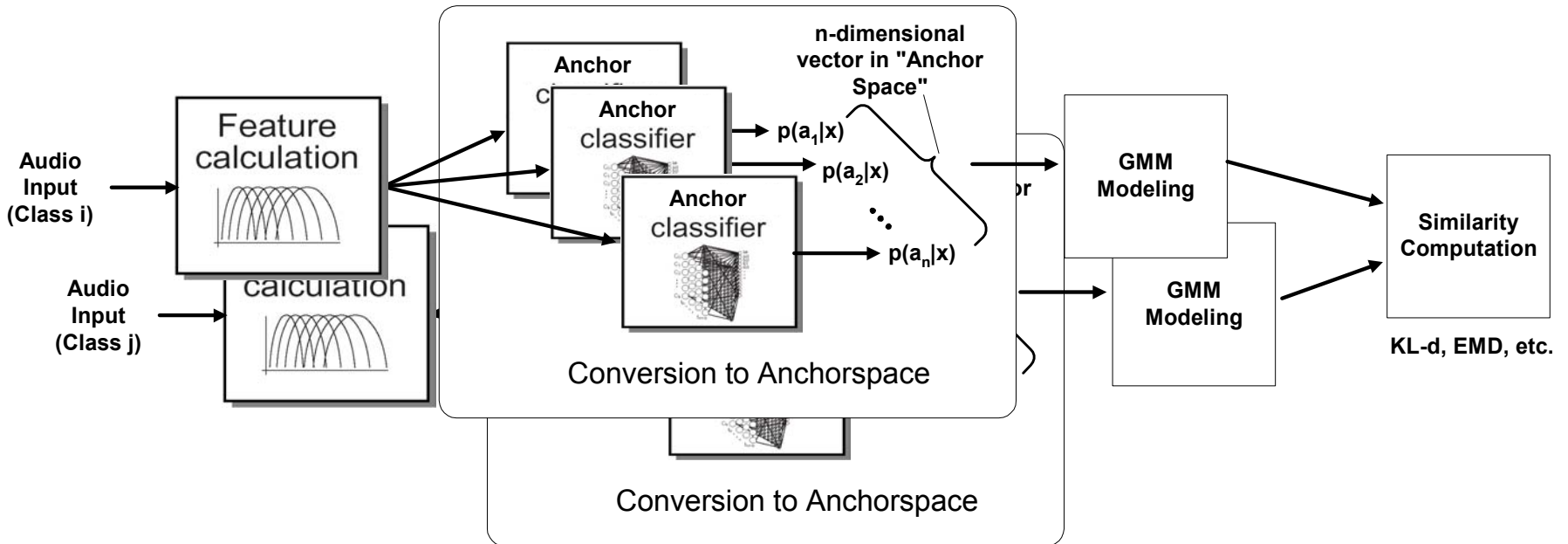 - Each frame of audio (32 milliseconds) is a point.

MFCC Clustering

- Logan & Salomon, ICME 2001
- MFCC features
- K-means clustering as pseudo-EM, per song or artist
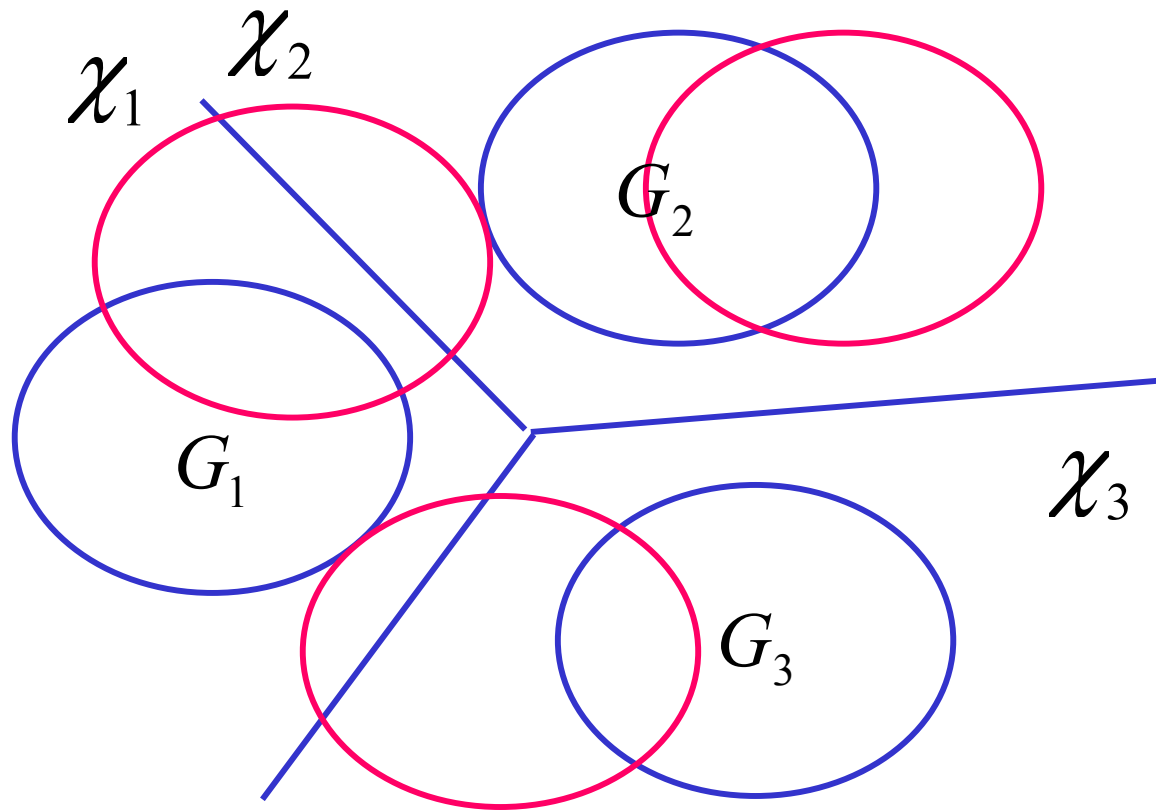- Earth-mover's distance (EMD) to compare distributions

Anchor Space

Comparing Clouds

- Centroid distance

- GMMs
  - KL-divergence? No closed form. So:
  - Likelihood of samples
  - Earth Mover's Distance (Rubner 98)
  - Asymptotic Likelihood Approximation (Vasconcelos 01)

Berenzweig et al

ALA

Berenzweig et al

Lab ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Acoustic Data

- www.ee.columbia.edu/~dpwe/research/musicsim

- 400 artists
  - Most popular artists on OpenNap mid-2002
  - Overlap with "Art of the Mix" playlist data early 2003

- 8827 songs, average 22 per artist
  - Coverage not equal for all artists
  - ~35G of mp3, 11G of MFCC data

Sharing Data

- Due to copyright, share MFCC features, not audio.

- Can add new features in future:
  - authors submit code for feature extraction
  - Columbia runs it over the data, shares feature output

- Acoustic Measures

- Subjective Measures

- Scoring Methods

- Results

- Study Proposal

Sources of Human Opinion

- ## Ask Directly
  - Survey
  - Experts: All Music Guide

- ## Infer from co-occurrence
  - User collections: OpenNap servers
  - Playlists: Art of the Mix

- ## Text
  - Web sites that discuss, describe artists

Please help us create a music similarity database which will be publicly available.

Which artist is most similar to:
Metallica?

1. Jimi Hendrix Experience
2. Marilyn Manson
3. Pink Floyd
4. Kingdoms Farm
5. Nirvana
6. Enya
7. Mayilyn Manson
8. Aaron Carter
9. Rage Against The Machine
0. Offspring
U. Unknown

Select U if you are unfamiliar with Metallica, or if you are unfamiliar with most of the options.
If your browser supports it, you can use the 1-9, 0, and U keys.
Answer as many questions as you like. You can quit at any time.
Return to mode selection menu.

Fleshing out expert opinion: Erdos distance

Ani Difranco

Liz Phair

Tracy Chapman

Beck

P.J. Harvey

R.E.M.

Velvet Underground
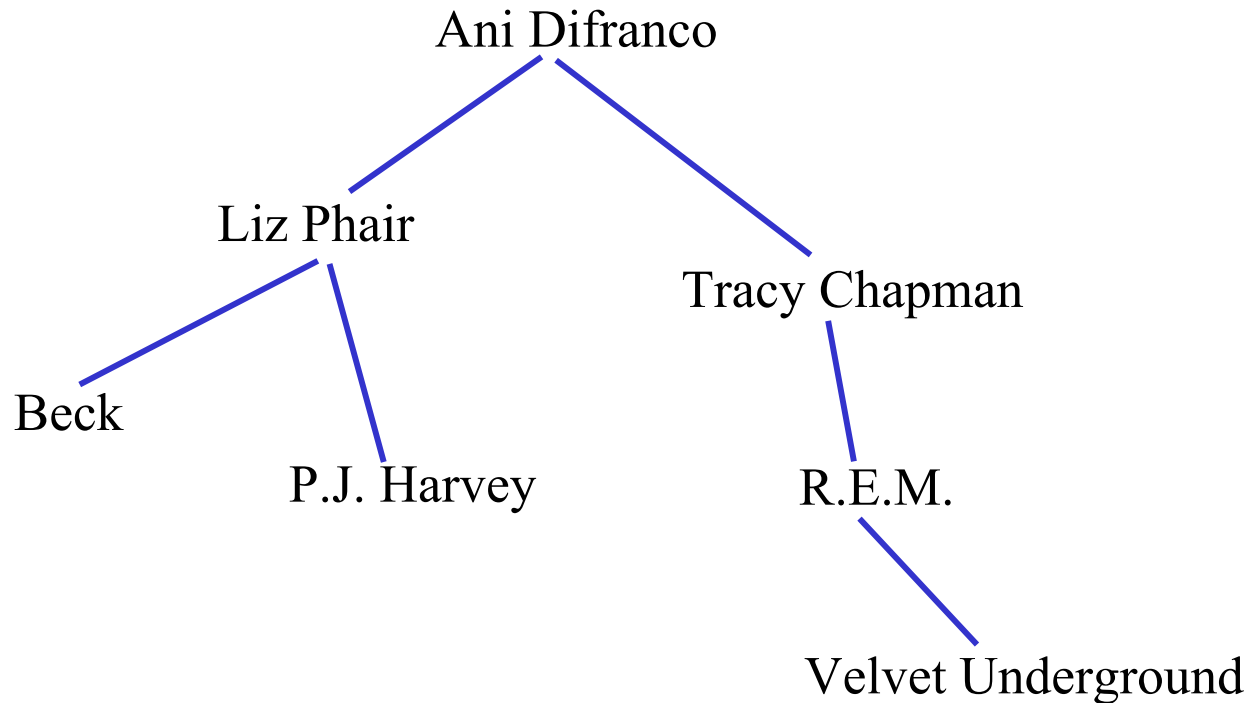
WebText

- Compare the language used to describe artists.

- Whitman & Lawrence, ICMC 2002
  - Google search for band name
  - Bag-of-words vector similarity

Lab ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.2-4-7-music.com/reviews/reviewsearch/long.asp?ArtistName=FRANK%20BLACK&FeatureType=REVIEW Go Links »

Google ▾ "frank black" review ▾ Search Web ▾ 145 blocked AutoFill Options frank b »

crud magazine   alternative satisfaction
                                                    website

MUSIC REVIEWS :: LATEST NEWS :: INTERVIEWS :: NEW R

THERE IS NO PEACE IN SILE

SEARCH

artist preview

ARTIST PREVIEW
Artist:
  NADA SURF
  TITLE :
  HIGH SPEED SOUL
  [MP3]
  LABLE :
  BARSUK/HEAVENLY
  PLAY   BUY

crud newsletter

CRUD
newsletter
sign-up
FREE/AUTOMATIC
ENTRY INTO ALL
PRIZE DRAWS

crud reviews

FRANK BLACK
EVERYTHING IS NEW [S
Label:
COOKING VINYL
Rating: ***

Frank Black limbers up for the release o
for release on 8th September on Cookin
– cover versions of Reid Paley's "Take W
The Hole" - the single is full of the same
that made 'Black Letter Days' so black.

Who knows, what with the election of the

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.allmusic.com/cg/amg.dll?p=amg&uid=CASS70308151747&sql=B0pfpxqy5ldje Go Links »

Google ▾ "frank black" review ▾ Search Web ▾ 145 blocked AutoFill Options frank black re »

music
guide   ● artists ○ albums ○ songs ○ styles ○ labels

Frank Black

Born	1965 in Long Beach, CA
Years Active	10s | 20s | 30s | 40s | 50s | 60s | 70s | 80s | 90s | 00s
Genres	Rock
Styles	Indie Rock, Alternative Pop/Rock
Instruments	Vocals, Guitar
Tones	Energetic, Playful, Acerbic, Aggressive, Quirky, Theatrical, Visceral, Cerebral
Labels	spinArt (3), Elektra (3), What Are Records? (2), Spin Art (2), Epic (2), Badd (2), 4AD/Elektra (2)
Charts & Awards	Click here for Billboard Chart Positions & GRAMMY Awards
See Also	All Movie Guide Entry
Internet links	On-Tour - provided by musictoday.com

AMG BIOGRAPHY   expand bio

by **Stephen Thomas Erlewine**

Inverting his stage name from Black Francis to Frank Black, the former Pixies lead singer/songwriter embarked on a solo career after he broke up the band in early 1993; actually, he began recording his solo album *before* he told the band the news. Working with former Pere Ubu member Eric Drew Feldman, Black occasionally heads into the ferocious post-punk guitar territory that marked such landmark albums as *Surfer Rosa* and *Doolittle*, but more frequently he plays up his considerably underrated melodic side. His self-titled 1993 debut album was an adventurous sketchbook of pop styles ranging from surf rock to heavy metal, from Beatlesque pop to new wave. Black's second album, 1994's *Teenager of the Year*, was a sprawling and diverse album that amplified all the best points of Frank

go to www.amazon.com   Local intranet

Lab
ROSA   Berenzweig et al
Laboratory for the Recognition and
Organization of Speech and Audio

User Collections and Playlists

- OpenNap servers.
- Art of the Mix
- Related to Collaborative Filtering
  - users who have X also have Y

File   Edit   View   Favorites   Tools   Help

Back   Search   Favorites   Media

Address http://www.artofthemix.org/findamix/GetContents.asp?strMixID=47104   Go

Google   CNN bought AOL   Search Web   Search Site   News   PageRank   Page Info   Up

Actions:
Return To Previous Page
Add This Mix To Your Favorites
View Feedback About This Mix
View Other Mixes By coyote23

Name: coyote23
Category: Theme
Format: CD
Date: 4/3/2003
View Previous Mix
View Next Mix

Santarchy III: Attack of the Christmas Tree Worms

Track	Artist:	Song:
1	(Spoken)	A Very Special Family Guy Freakin' Xmas Pageant
2	Wesley Willis	Merry Christmas
3	Jimmy Eat World	Last Christmas
4	Grandaddy	Alan Parsons in a Winter Wonderland
5	Luscious Jackson	Let It Snow
6	Poe	Grandma Got Run Over By A Reindeer
7	Belle and Sebastian	O Come, O Come Emmanuel
8	Joni Mitchell	River
9	Bright Eyes	Blue Christmas
10	Candy Butchers	All I Want For Christmas (Is You)
11	Saint Etienne	Driving Home For Christmas
12	Sugarcubes	Christmas Eve
13	Arab Strap	Xmas (Baby, Please Come Home)
14	The Cruxshadows	Happy Xmas (War Is Over)

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Berenzweig et al

19

Co-occurrence Similarity

- Compute co-occurrence matrix, rows are p(x|y)
- Normalize by prior p(x): related to Mutal Information

$$\mathrm{co}(x, y) = \frac{\mathrm{p}(x|y)}{\mathrm{p}(x)} = \frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)}$$

$$\mathrm{I}(X;Y) = \mathrm{E}_{p(x,y)} \log \frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)} = \mathrm{E}_{p(x,y)} \log \mathrm{co}(x, y)$$

- Used for playlists (AOTM) and collections (OpenNap)

Berenzweig et al

Data Stats

- **Survey Data**
  - 22, 300 responses from users to questions about the 400 artists
  - `Given artist a, which of the 10 presented artists is closest ?'
- **Expert Opinion**
  - similar artist lists from All Music Guide (www.allmusic.com)
  - average of 5.4 similar artists per list
- **OpenNap User Collections**
  - co-occurrence data from 3200 user collections
  - 175, 000 user-to-artist relations
- **Art of the Mix Playlists**
  - co-occurrence data from 23000 playlists
  - average of 4.4 entries per playlist
- **Other data**
  - again, we encourage other groups to submit data

Sparsity

Source	# obs	art/obs	> 0 obs	≥ 10 obs	med#art
Survey	17,104	5.54	7.49%	0.36%	23
Expert	400	5.41	1.35%	-	5
Playlist	23,111	4.38	51.4%	11.4%	213
Collection	3,245	54.3	94.1%	72.1%	388

- Only some subset of pairs are directly compared.
  - Too dissimilar
  - Artist unknown
  - Exception: acoustic data! can do all n^2 compares.
- How does it affect results?
  - Evaluation method should be agnostic wrt sparsity.

Berenzweig et al

- Acoustic Measures

- Subjective Measures

- Scoring Methods

- Results

- Study Proposal

Scoring Methods

- Survey-based: Ask the metric the same questions we asked users
  - Average rank agreement
  - First-place agreement
- Cross-reference Evaluation
  - Compare two similarity matrices
  - Any similarity matrix can be considered truth.
  - Top-N agreement

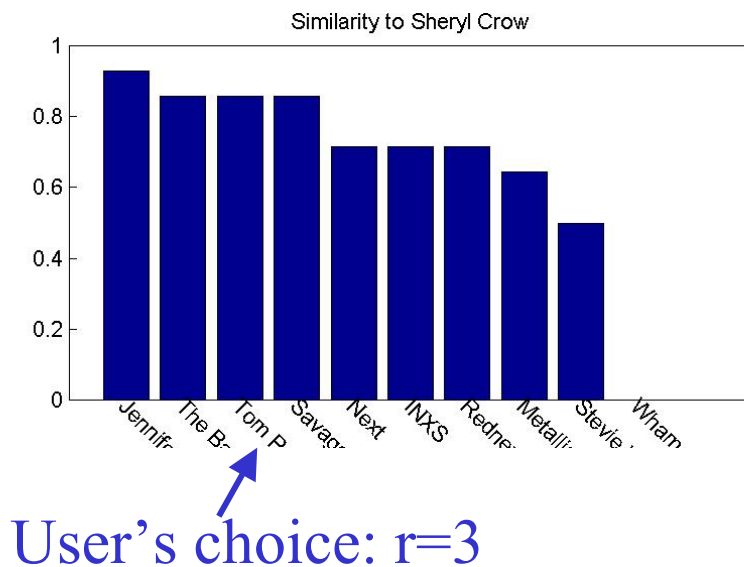# Survey-based Evaluation: Average Rank

Which artist is most similar to Sheryl Crow?

Wham

Metallica

Savage Garden

Rednex

Stevie Wonder

Tom Petty ←

The Bangles

INXS

Jennifer Paige

Next

Ordered by the metric

Similarity to Sheryl Crow
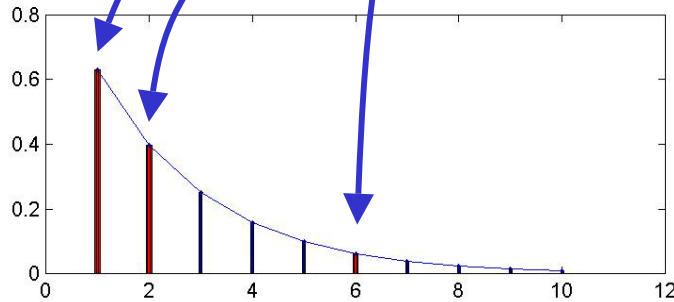
User's choice: r=3

- Normalize to [1,10]

$$R = 1 + \frac{(r-1)(10-1)}{N-1}$$

- Then average over all judgements.

- Random=5.5

- Optimal ~ 2.13

Top-N Reference Ranking

$$s_i = \sum_{r=1}^{N} \alpha_R^r \alpha_C^{k_r}$$

Jennifer Paige
The Bangles
Tom Petty
Svg Garden
Next
INXS
Rednex
Metallica
Stevie
Wonder
Wham

Top N sorted by
experimental metric

Top N sorted by
reference

Scoring Methods

- ## Survey-based: Ask the metric the same questions we asked users

  - Average rank agreement

  - First-place agreement

    - statistical significance: one-tailed binomial test (1% at 5%)

- ## Cross-reference Evaluation

  - Compare two similarity matrices

  - Any similarity matrix can be considered truth.

  - Top-N agreement

- Acoustic Measures

- Subjective Measures

- Scoring Methods

- Results

- Study Proposal

Searching Parameter Space: MFCC

	#mix	c0?	Independent		Pooled		
			ALA	EMD	ALA	Cntrd	EMD
EM	8	y	4.76 / 16%	4.46 / 20%	4.72 / 17%	4.66 / 20%	4.30 / 21%
	8	n	-	4.37 / 22%	-	-	4.23 / 22%
	16	n	-	4.37 / 22%	-	-	4.21 / 21%
K-means	8	y	-	4.64 / 18%	-	-	4.30 / 22%
	8	n	4.70 / 16%	4.30 / 22%	4.76 / 17%	4.37 / 20%	4.28 / 21%
	16	y	-	4.75 / 18%	-	-	4.25 / 22%
	16	n	4.58 / 18%	4.25 / 22%	4.75 / 17%	4.37 / 20%	4.20 / 22%
	32	n	-	-	4.73 / 17%	4.37 / 20%	4.15 / 23%
	64	n	-	-	4.73 / 17%	4.37 / 20%	4.14 / 23%
Optimal			2.13 / 53.5%				
Random			5.50 / 11.4%				

- Pooled covariance, no c0 (energy), more mixture components are better, up to 32.
- K-means comparable to EM, and computationally simpler.
- EMD is best; but for Anchor Space, ALA.
  - ALA assumptions fail in MFCC space?

Searching Parameter Space: Anchorspace

- Full, independent covariance
- All 14 dimensions
- ALA

Results

- Compare two different acoustic distance measures
  - Local K-means clustering of MFCC features (Logan & Salomon)
  - GMM clustering of features in Anchor Space
- Search parameter space with survey as ground truth
- Scoring is survey-based
  - Average rank response / % 1st-place agreement

#mix	MFCC	Anchor
8	4.28/21.3%	4.25/20.2%
16	4.20/22.2%	4.20/19.8%

Cross-Reference Results

- What's best ground truth? pairwise comparisons between measures

	1st place	survey	expert	playlist	collection	webtext
Random	11.8%	0.015	0.020	0.015	0.017	0.012
Anchor	19.8%	0.092	0.095	0.117	0.097	0.041
MFCC	22.2%	0.112	0.099	0.142	0.116	0.046
Survey	53.5%	0.874	0.249	0.204	0.331	0.121
Expert	27.9%	0.267	0.710	0.193	0.182	0.077
Playlist	26.5%	0.222	0.186	0.985	0.226	0.075
Collection	23.2%	0.355	0.179	0.224	0.993	0.083
Webtext	18.5%	0.131	0.082	0.077	0.087	0.997
mean*		0.197	0.148	0.160	0.173	0.074

- Natural asymmetry because $\alpha_r \neq \alpha_c$
- Diagonal<1 because of random tiebreaker, sparsity
- 53% reflects low agreement between subjects

Berenzweig et al

Cross-Reference Results

- What's best ground truth? pairwise comparisons between measures

	1st place	survey	expert	playlist	collection	webtext
Random	11.8%	0.015	0.020	0.015	0.017	0.012
Anchor	19.8%	0.092	0.095	0.117	0.097	0.041
MFCC	22.2%	0.112	0.099	0.142	0.116	0.046
Survey	53.5%	0.874	0.249	0.204	0.331	0.121
Expert	27.9%	0.267	0.710	0.193	0.182	0.077
Playlist	26.5%	0.222	0.186	0.985	0.226	0.075
Collection	23.2%	0.355	0.179	0.224	0.993	0.083
Webtext	18.5%	0.131	0.082	0.077	0.087	0.997
mean*		0.197	0.148	0.160	0.173	0.074

- (survey,collection) .343 [surprising!]
- (survey,expert) .258 [explicit judgments]
- (playlist,collection) .225 [co-occurrence data]
- (survey,playlist) .213

Lab ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Berenzweig et al

33

Cross-Reference Results

- What's best ground truth? pairwise comparisons between measures

	1st place	survey	expert	playlist	collection	webtext
Random	11.8%	0.015	0.020	0.015	0.017	0.012
Anchor	19.8%	0.092	0.095	0.117	0.097	0.041
MFCC	22.2%	0.112	0.099	0.142	0.116	0.046
Survey	53.5%	0.874	0.249	0.204	0.331	0.121
Expert	27.9%	0.267	0.710	0.193	0.182	0.077
Playlist	26.5%	0.222	0.186	0.985	0.226	0.075
Collection	23.2%	0.355	0.179	0.224	0.993	0.083
Webtext	18.5%	0.131	0.082	0.077	0.087	0.997
mean*		0.197	0.148	0.160	0.173	0.074

- Respectable performance of acoustic measures
- Survey is best scoring mean, but sparse.
- Collection is next, and high agreement w/ survey.

Berenzweig et al

34

Invitation

- Hosted at Columbia
  www.ee.columbia.edu/~dpwe/research/musicsim/

- Acoustic and Subjective data
  - 400 artists, 8827 songs, 11G of MFCC, OpenNap, Art of the mix, AMG, Survey, Webtext

- Sharing features is viable for corpus sharing. We welcome feature contributions.

Thanks!