The Million Song Dataset

AUDIO FEATURES

Features for Audio and Music Classification

M.F. McKinney, J. Breebaart, ISMIR 2003 (2023: 520 citations)

General Audio Class	Classical Music	Popular Music	Speech	Noise	Crowd Noise
# of Files	35	188	31	25	31

Popular Music Classes	Jazz	Folk	Electro nica	R&B	Rock	Reggae	Vocal
# of Files	38	23	27	43	37	11	9

Low Level Features (Li et al. 2001)

- root-mean-square (RMS) level
- zero-crossing rate
- band energy ratio
- Pitch
- ..

Mel-frequency cepstral coefficients (MFCC) derived Features (Slaney et al. 1998)

- MFCC

Psycho Acoustic Features

- Roughness, sdev roughness, loudness, sharpness, modulations of them

Filterbank temporal envelopes

Features for Audio and Music Classification

M.F. McKinney, J. Breebaart, ISMIR 2003 (2023: 520 citations)

Low Level Features
Root Mean Square (RMS) level
Spectral Centroid
Bandwidth
Zero-Crossing Rate
Spectral roll-off frequency (harmonics vs noise)
Band energy ratio
Delta spectrum magnitude
Pitch
Pitch strength

- Fast implementations.
- Often computed in the time domain.
- Pitch detection using autocorrelation in time domain.

$$R_x(m) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m)$$

Spectral roll-off frequency:
- a cutoff frequency under which some percentage of the spectrum is contained
- harmonic sounds below cutoff
- noise above roll-off)

Features for Audio and Music Classification

M.F. McKinney, J. Breebaart, Features for Audio and Music Classification, ISMIR 2003

(2023: 520 citations)

General Audio Class	Classical Music	Popular Music	Speech	Noise	Crowd Noise
Number of Files	35	188	31	25	31

Popular Music Class	Jazz	Folk	Electronica	R&B	Rock	Reggae	Vocal
Number of Files	38	23	27	43	37	11	9

Table 1: Audio database by class: number of audio files in each class.

Low Level Features (Li et al. 2001)
- root-mean-square (RMS) level
- Spectral centroid
- Bandwidth
- zero-crossing rate
- Spectral roll-off frequency (harmonics vs noise)
- band energy ratio
- delta spectrum magnitude,
- Pitch
- pitch strength

Mel-frequency cepstral coefficients (MFCC) derived Features (Slaney et al. 1998)
- MFCC
- Modulation Energy of MFFC
- Note: in Speech Recognition MFFC, delta MFCC, delta² MFCC are used

Psycho Acoustic Features
- Roughness, sdev roughness, loudness, sharpness, modulations of them

Filterbank temporal envelopes

Mel-frequency cepstral coefficients (MFCC) derived Features (Slaney et al. 1998)

Audio Input

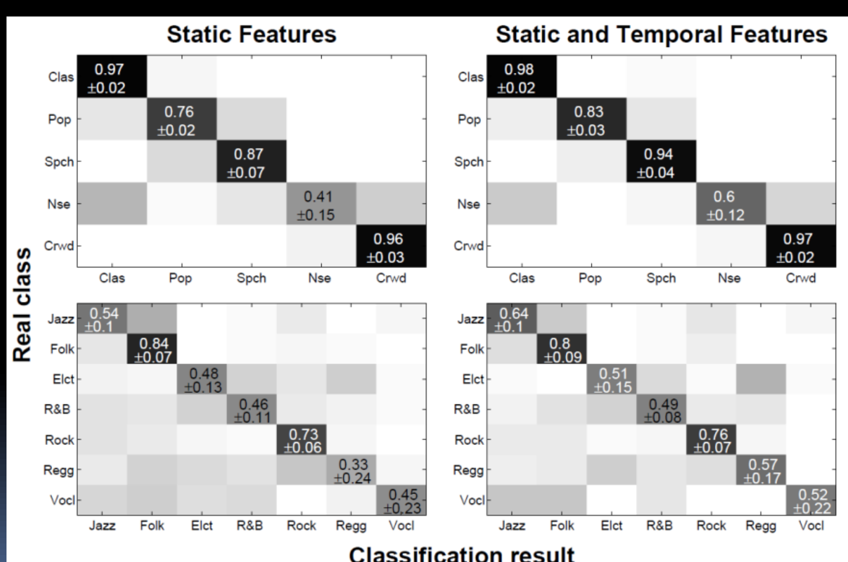
- Pre-Emphasis
- Framing
- Windowing
- Discrete Fourier Transform
- Mel Filter Banks
- Discrete Cosine Transform

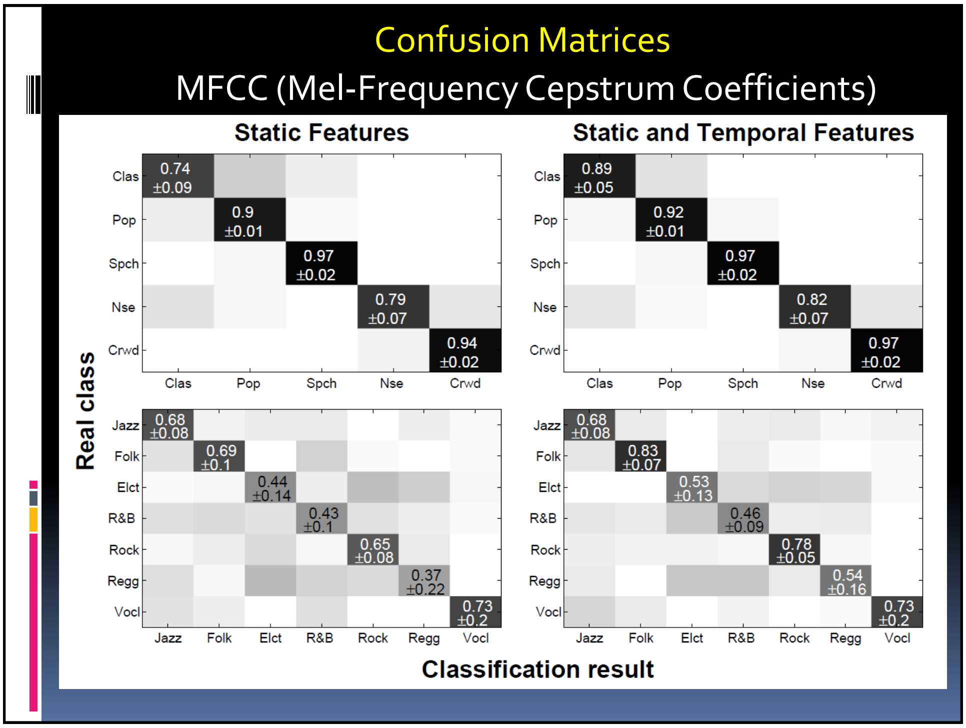
MFFC’s

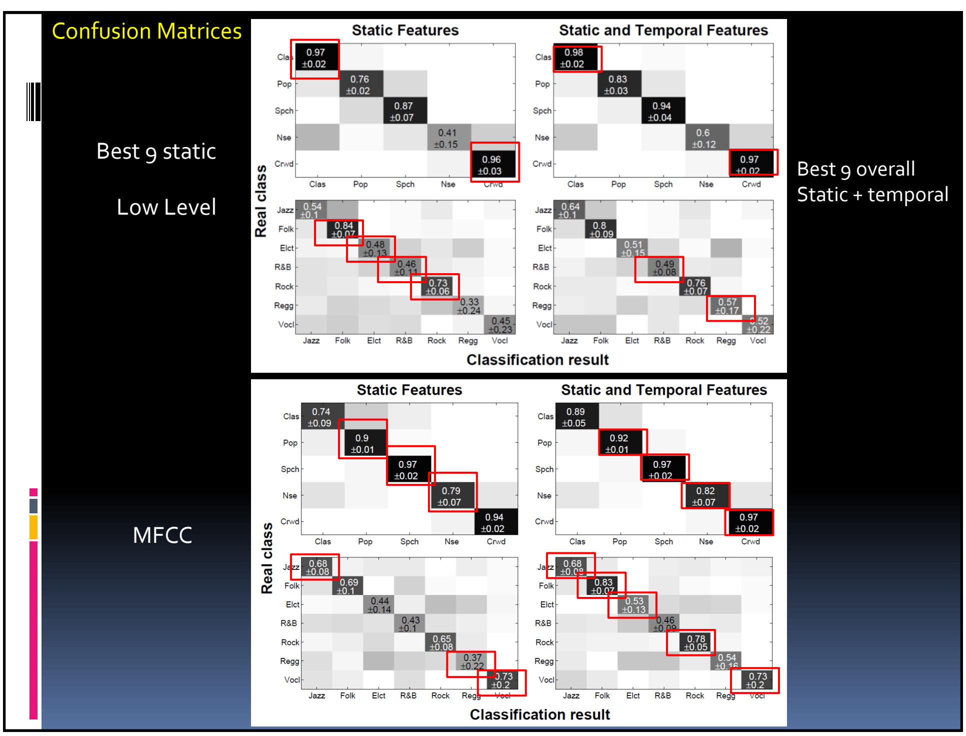
Confusion Matrices

Best 9 static Low Level Features

Best 9 overall Static + temporal




Confusion Matrices
MFCC (Mel-Frequency Cepstrum Coefficients)
Static Features
Static and Temporal Features
Real class
Classification result


Confusion Matrices
Best 9 static
Low Level
MFCC
Best 9 overall
Static + temporal
Static Features
Static and Temporal Features
Real class
Classification result

The Million Song Dataset

"There is no data like more data" Bob Mercer of IBM (1985).

T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman, P. Lamere, **The Million Song Dataset**, In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

(2023: 1655 citations)

The Million Song Dataset (MSD)

metadata and extracted audio features for a million songs from The Echo Nest.

Licensing
- GZTAN a smaller dataset
- Magnatagatune
- MSD Legally available

Other audio data sets:
- https://www.audiocontentanalysis.org/datasets
- http://www.ismir.net/resources/datasets/

Audio Data Sets

- The Million Song Dataset (MSD)
  - metadata and extracted audio features for a million songs from The Echo Nest.
  - GZTAN a smaller dataset
  - Magnatagatune

Other audio data sets:

- https://www.audiocontentanalysis.org/datasets
- http://www.ismir.net/resources/datasets/

MIREX 2021 **http://www.music-ir.org/mirex/wiki/MIREX_HOME**

- **Chord Estimation, Cover Song Detection, Melody Extraction,**
- **Lyrics Transcription, Drum Transcription, Music Detection**
- **Query by Singing, Humming**
- **Set List Identification: determine the song sequence in a live concert**

Previous challenges on MIREX:

- **Multiple Fundamental Frequency Estimation and Tracking**
- **K-POP Mood and Genre Classification**
- **Singing Transcription, Lyrics Transcription**
- Audio Key detection, Audio Fingerprinting, and Mood-, Genre-, Tag-Classification, etc,

MSD Goals: Reference Benchmark Dataset

- Scale MIR related research to commercial sizes
- Provide reference dataset for research evaluation
- Alternative shortcut for The Echo Nest's API
  - >=2016 only Spotify https://en.wikipedia.org/wiki/The_Echo_Nest
  - https://acousticbrainz.org/ (data collection stopped 2022-02-16)
  - https://musicbrainz.org/
- API of the 7digital service, 30-s audio previews
- Kick start new MIR researchers

MIR Datasets Critical Requirements

- Algorithms should be scalable
- Realistically sized datasets are necessary

dataset	# songs / samples	audio
RWC	465	Yes
CAL500	502	No
GZTAN genre	1,000	Yes
USPOP	8,752	No
Swat10K	10,870	No
Magnatagatune	25,863	Yes
OMRAS2	50,000?	No
MusiCLEF	200,000	Yes
MSD	1,000,000	No

G. Tzanetakis et al. 2002

MusiCLEF 2012: http://www.cp.jku.at/datasets/musiclef/index.html

MSD Creation

- The Echo Nest API with Python wrapper pyechonest. (*)
- Echo Nest provided:
  - Metadata: artist, title, etc.
  - Audio Features: short time scale – global scale
  - Defined by Echo Nest Analyze API (per segment)
- Additional info from musicbrainz server
- 5 Threads during 10 days
- Code available (not relevant anymore)

*) 'Retired' since 2016

Alternative: http://acousticbrainz.org/ (data collection stopped 2022-02-16)

MSD Content

- 280 GB of data
- 1, 000, 000 songs/files
- 44, 745 unique artists
- 7, 643 unique terms (Echo Nest tags)
- 2, 321 unique musicbrainz tags
- 43, 943 artists with at least one term
- 2, 201, 916 asymmetric similarity relationships
- 515, 576 dated tracks starting from 1922

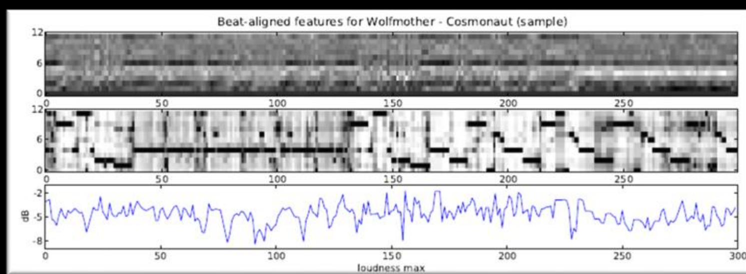
MSD Content

- HDF5 format
- 55 fields per song
- Audio Features
  - Timbre
  - Pitches
  - Loudness max
  - Beats
  - Bars (~3 – 4 beats)
  - Note onsets/tatum

analysis_sample_rate	artist_7digitalid
artist_familiarity	artist_hotttnesss
artist_id	artist_latitude
artist_location	artist_longitude
artist_mbid	artist_mbtags
artist_mbtags_count	artist_name
artist_playmeid	artist_terms
artist_terms_freq	artist_terms_weight
audio_md5	bars_confidence
bars_start	beats_confidence
beats_start	danceability
duration	end_of_fade_in
energy	key
key_confidence	loudness
mode	mode_confidence
num_songs	release
release_7digitalid	sections_confidence
sections_start	segments_confidence
segments_loudness_max	segments_loudness_max_time
segments_loudness_start	segments_pitches
segments_start	segments_timbre
similar_artists	song_hotttnesss
song_id	start_of_fade_out
tatums_confidence	tatums_start
tempo	time_signature
time_signature_confidence	title
track_7digitalid	track_id
year	

MSD Audio Features

Wolfmother, Cosmonaut (2009)



Timbre

Pitch

Loudness

- Timbre, Pitches (both 12 elements per segment) and Loudness max for one song.

MSD Integration

- Echo Nest identifiers
  - (track, song, album, artist) => updates on dynamic values: popularity, familiarity, etc.
- Yahoo Music Ratings Datasets provides user ratings for 97 954 artists
  - 15 780 artists in MSD (91% overlap with the more popular artists in MSD)
  - At the time one of the largest benchmarks for evaluating content-based music recommendation
- Identifiers
  - Artist, album, song names
  - Echo Nest id
  - Musicbrainz id
  - MusiXmatch id => lyrics
  - 7digital identifiers > 30sec samples

Note: Spotify and others use ISRC (International Standard Recording Code)

MSD Usage Examples

- Metadata Analysis
- Artist Recognition
- Automatic Music Tagging
- Recommendation
- Cover Song Recognition
  - SecondHandSong Dataset 18 196 covers of 5 854 songs
  - Most methods based on chroma features
- Lyrics
  - Mood prediction
- Year Prediction

Metadata Analysis

- Are all good artist names already taken?

"Tim and Sam's Tim and the Sam Band with Tim and Sam"

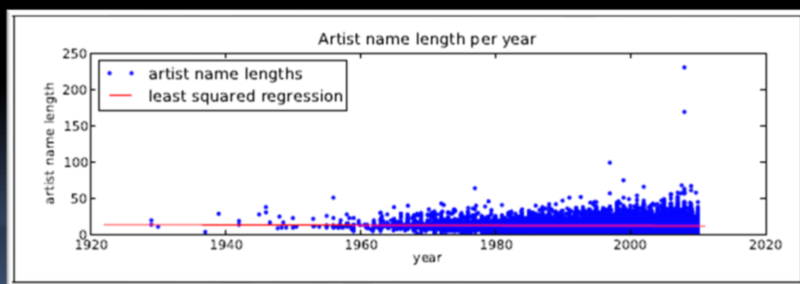
- Do newer bands have to use longer names?
  - ...
- Etc.

Metadata Analysis

- Are all good artist names already taken?

"Tim and Sam's Tim and the Sam Band with Tim and Sam"

- Do newer bands have to use longer names?
  - Seems false, apart from outliers. See graph.
- Etc.



Artist Recognition

- 18 073 artists with at least 20 songs in MSD
- 2 standard training/test datasets
  - 20 songs/artist
  - 15 songs/artist
- Benchmark k-NN algorithm resulted in an accuracy of 4% !

  => much room for improvement?

Automatic Music Tagging

- Core of MIR research for many years
- 300 most popular terms in The Echo Nest
- Split all artists in training/test sets according to terms
- Correlations between artist names and genre, or year and genre etc.

artist	EN terms	musicbrainz tags
Bon Jovi	adult contemporary arena rock 80s	hard rock glam metal american
Britney Spears	teen pop soft rock female	pop american dance

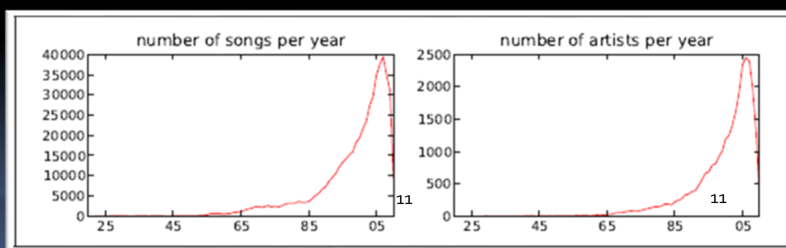
Music Recommendation

- Music recommendation and music similarity have high commercial value.
- Content based systems underperform when compared to collaborative filtering methods (2011)
  - Also novelty and suprise are important.
- Integration with Yahoo Music Ratings
  - Enables large scale experiments
  - Clean ground truth
- Similar Artists according to Echo Nest:

Ricky Martin	Weezer
Enrique Iglesias Christina Aguilera Shakira Jennifer Lopez	Death Cab for Cutie The Smashing Pumpkins Foo Fighters Green Day

Year Prediction

- Little studied
- Practical applications in music recommendation
- Years-of-release field (1922 – 2011)
  - 515 576 tracks of 28 223 artists
  - Errors
  - Non-uniformity over the years



Year Prediction

- K-NN: the predicted year is the average of the k nearest training songs
- Vowpal & Wabbit (**VW**): regression by learning a linear transformation **T** of the features using gradient descent => predicted year is equal to the application of **T** on the features of the song
- Table shows
  - average absolute difference between predicted and actual year
  - the square root of the average squared difference between predicted and actual year.
- Benchmark average release year predicted from the training set. VW improves this baseline.

method	diff	sq. diff
constant pred.	8.13	10.80
1-NN	9.81	13.99
50-NN	7.58	10.20
VW	**6.14**	**8.76**

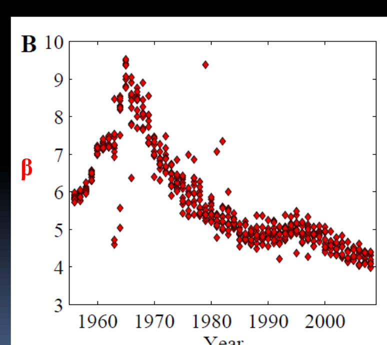
Smaller is better

Evolution of Pop Music

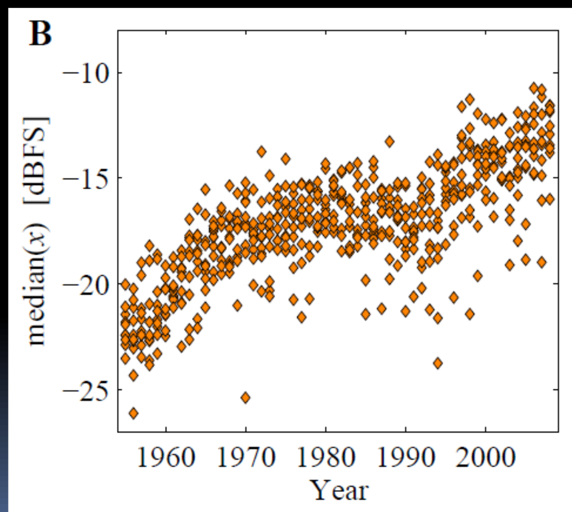
Measuring the evolution of contemporary western popular music, J. Serra, A. Corral, M. Boguna, M. Haro and J.L. Arcos, *2012*

Timbre of Pop Music

- The distributions of timbre codewords are fitted to a power-law distribution with parameter β.
- Lower β indicates less timbre variety, i.e., frequent code words become more frequent and infrequent ones less frequent.
- More homogeneity in timbre



Loudness of Pop Music



MSD Limitations

- No or limited access to original audio
  - Novel audio feature analysis and acoustic features
- Lack of album and song level meta data and tags
- Limited Diversity
  - World, ethnic, and classic music almost not represented
- Accurate time stamps problematic
  - No guarantee that audio features have been computed using the same audio track
  - As a result from many official releases, different ripping and encoding schemes, etc

the Million Song Dataset Challenge

B. McFee, et al., WWW 2012 Companion, April 16-20 2012, Lyon, France.

Personalized music recommendation challenge.

Goal:

- predict the songs that a user will listen to, given the user's listening history and full information (including meta-data and content analysis) for all songs.

the Million Song Dataset Challenge (2012)

http://www.kaggle.com/c/msdchallenge

"What is the task in a few words?" You have:

1) the full listening history for 1M users,
2) half of the listening history for 110K users (10K validation set, 100K test set), and
3) you must predict the missing half. .."

Winner: *aio* with a MAP@k score of 0.17910
(MAP@k = Mean average precision over k queries)

Future (of 2012)

- Success? Time will tell.
- Hopefully used as one of the default benchmarks
- Depends on efforts of research community
- Preserving commonality and comparability
- Important for visibility of MIR research
- Subsets on UCI Machine Learning Repository

2021: Number of citations 1211.
2022: Number of citations 1378 (March); 1481(October); 2023: 1659
Recent citations in work on recommender systems, etc.
Example: https://zenodo.org/record/1240485#.W78ZtPloSUk
MSD-I: Million Song Dataset with Images for Multimodal Genre Classification

Multimodal Deep Learning for Music Genre Classification. Transactions of the International Society for Music Information Retrieval

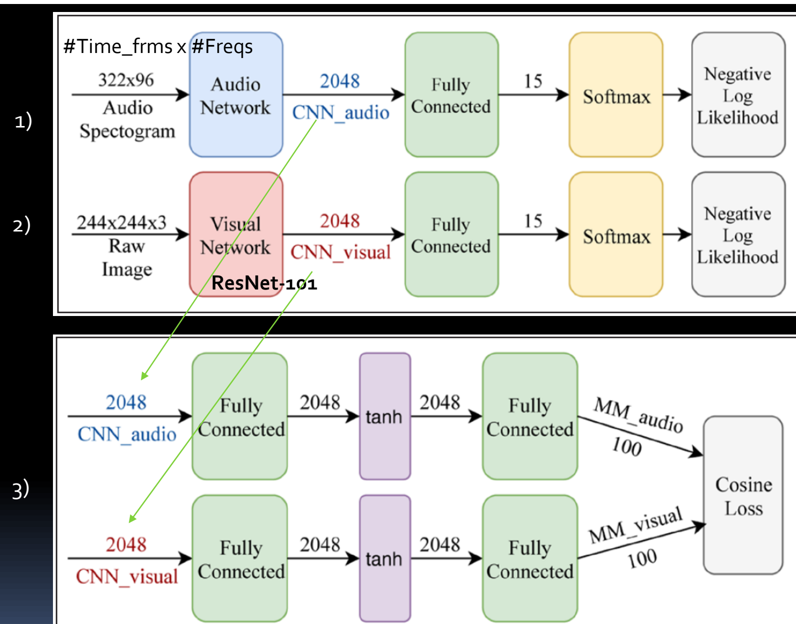
Oramas, S., et al. (2018)

- learn and combine multimodal data representations for music genre classification
- deep neural networks are trained with:
  - audio tracks
  - text reviews
  - cover art images
- single label genre classification (only A + V)
  - using Million Songs Data set (MSD-I)
- multi label genre classification (A + V + T)
  - using their Multimodal Music dataset (combines Amazon Review dataset and the Million Song Dataset)

Cover Art

New Aged misclassified as Heavy Metal

Genre Heat-Maps



CNN's and Feature Space Network

Genre Classification

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

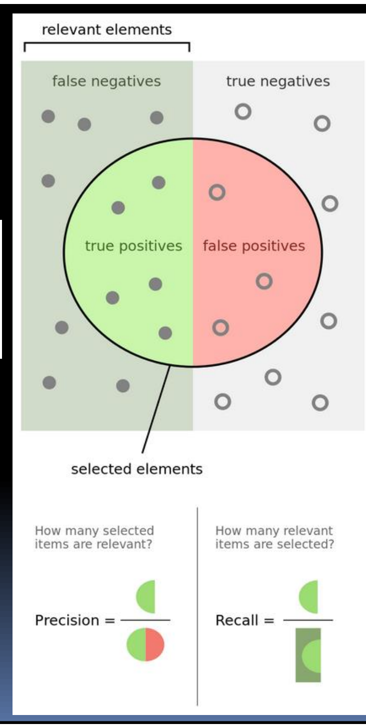
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

F1 →
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

https://en.wikipedia.org/wiki/Precision_and_recall



Genre Classification

Input	Model	Precision	Recall	F1
Audio	CNN_Audio	0.385 ± 0.006	0.341 ± 0.001	0.336 ± 0.002
	MM_Audio	0.406 ± 0.001	0.342 ± 0.003	0.334 ± 0.003
	CNN_Audio + MM_Audio	0.389 ± 0.005	0.350 ± 0.002	**0.346 ± 0.002**
Video	**CNN_Visual**	0.291 ± 0.016	0.260 ± 0.006	**0.255 ± 0.003**
	MM_Visual	0.264 ± 0.005	0.241 ± 0.002	0.239 ± 0.002
	CNN_Visual + MM_Visual	0.271 ± 0.001	0.248 ± 0.003	0.245 ± 0.003
A + V	CNN_Audio + CNN_Visual	0.485 ± 0.005	0.413 ± 0.005	0.425 ± 0.005
	MM_Audio + MM_Visual	0.467 ± 0.007	0.393 ± 0.003	0.400 ± 0.004
	ALL	0.477 ± 0.010	0.413 ± 0.002	$\boxed{\mathbf{0.427 \pm 0.000}}$

Genre	Human Annotator			Neural Model		
	Audio	Visual	A + V	Audio	Visual	A + V
Blues	0	0.50	0.67	0.05	0.36	0.42
Country	0.40	0.60	0.31	0.37	0.21	0.40
Electronic	0.62	0.44	0.67	0.64	0.44	0.68
Folk	0	0.33	0	0.13	0.23	0.28
Jazz	0.62	0.38	0.67	0.47	0.27	0.49
Latin	0.33	0.33	0.40	0.17	0.08	0.13
Metal	0.80	0.43	0.71	0.69	0.49	0.73
New Age	0	0	0	0	0.12	0.10
Pop	0.43	0.46	0.42	0.39	0.43	0.49
Punk	0.44	0.29	0.46	0.04	0	0.30
Rap	0.74	0.29	0.88	0.73	0.39	0.73
Reggae	0.67	0	0.80	0.51	0.34	0.55
RnB	0.55	0	0.46	0.45	0.31	0.51
Rock	0.58	0.40	0.40	0.54	0.20	0.58
World	0	0.33	0	0	0	0.03
Average	**0.41**	**0.32**	**0.46**	**0.35**	**0.25**	**0.43**

MSD-I: Million Song Dataset with Images for Multimodal Genre Classification

For data see:

https://zenodo.org/record/1240485#.XamLyn9S-Uk

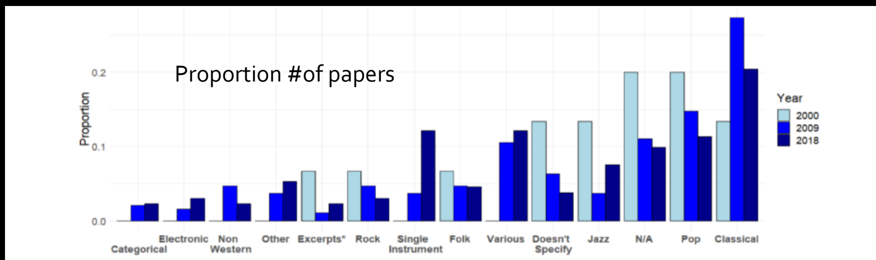
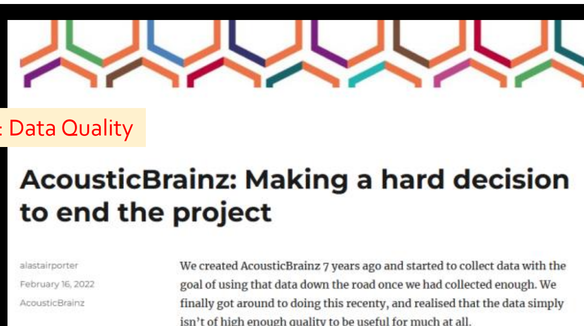


Figure 1: Proportion [3] of the number of papers that use different genres of data from first ISMIR conference in 2000 [1] to the 10th ISMIR in 2009 [2], to the 19th ISMIR in 2018 [3]. "Excerpts*" refers to music excerpts under 3 seconds, and "categorical" refers to music selected for a non-genre category such as mood. The "non-Western" category does not include genres such as J-pop and K-pop, which were classified as solely "pop".

W. Chen et al., DATA USAGE IN MIR: HISTORY & FUTURE RECOMMENDATIONS. ISMIR 2019.

Problem: Data Quality



AcousticBrainz: Making a hard decision to end the project

alastairporter
February 16, 2022
AcousticBrainz

We created AcousticBrainz 7 years ago and started to collect data with the goal of using that data down the road once we had collected enough. We finally got around to doing this recenty, and realised that the data simply isn't of high enough quality to be useful for much at all.

Goals

- Musical characteristics of audio recordings: musical key, bpm
- Use extracted data to predict: instrumentation, genre, mood, etc.
- Source of features to build and train models for prediction

Problems

- Musical key accurate on some styles but not on the full range
- BPM worked well but on many recordings incorrect, also no confidence levels available
- Existing models for genre not working very well and not covering the full range
- AcousticBrainz data extractor has not high enough resolution for Deep Learning
- Content-based similarity methods by AcousticBrainz did not work well

https://mtg.github.io/acousticbrainz-genre-dataset/

MSD Related publications

https://www.researchgate.net/publication/220723656_The_Million_Song_Dataset

Some examples:

H. Eghbal-Zadeh, M. Dorfer, G. Widmer, A Cosine-Distance based Neural Network for **Music Artist Recognition** using Raw I-vector Features, Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16), Brno, Czech Republic, September 5–9, 2016

K. Choi, G. Fazekas, M. Sandler, K. Cho, Convolutional Recurrent Neural Networks for **Music Classification**, arXiv:1609.04243v1 [cs.NE] 14 Sep 2016

Oramas S., Nieto O., Sordo M., & Serra X. (2017) A Deep Multimodal Approach for Cold-start **Music Recommendation**. https://arxiv.org/abs/1706.09739

Music recommendation system approaches in machine learning
SN Pasha, D Ramesh, S Mohmmad... - AIP Conference ..., 2022 - aip.scitation.org
... For this investigation we used the **million song dataset** available from kaggle.We used both the content based filtering and collaborate filtering algorithms to provide best music ...
☆ Opslaan 99 Citeren Verwante artikelen

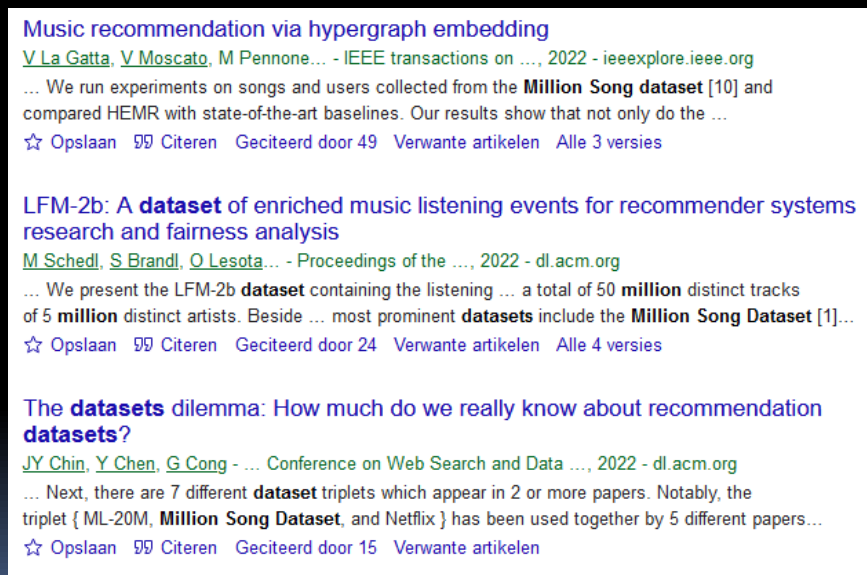
Discussion of 'Multi-scale Fisher's independence test for multivariate dependence'
A Schrab, W Jitkrittum, Z Szabó, D Sejdinovic... - ..., 2022 - academic.oup.com
... the **Million Song Dataset** with X consisting of 90 **song** features and Y being the **song's** release ... (a) Power experiment using the **Million Song Dataset** with |X| consisting of 90 **song** ...
☆ Opslaan 99 Citeren Geciteerd door 1 Alle 9 versies

Music Recommendation via Hypergraph Embedding
V La Gatta, V Moscato, M Pennone... - ... on Neural Networks ..., 2022 - ieeexplore.ieee.org
... We run experiments on songs and users collected from the **Million Song dataset** [10] and compared HEMR with state-of-the-art baselines. Our results show that not only do the ...
☆ Opslaan 99 Citeren Geciteerd door 4 Verwante artikelen Alle 3 versies

[PDF] **Datasets** Finders and Best Public **Datasets** for Machine Learning and Data Science Applications
R Marappan, S Bhaskaran - COJ Rob Artificial Intel, 2022 - sunbeltsport.com
... Hence it is necessary to learn the different public **datasets** before one starts the project. **Dataset** Finders This section explores the best public **datasets** finders available for data science ...
☆ Opslaan 99 Citeren Geciteerd door 7

LFM-2b: A **Dataset** of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis
M Schedl, S Brandl, O Lesota... - ACM SIGIR Conference ..., 2022 - dl.acm.org
... We present the LFM-2b **dataset** containing the listening records of over 120,000 users of the ... **datasets** include the **Million Song Dataset** [1], Spotify's Music Streaming Sessions **Dataset** [2...
☆ Opslaan 99 Citeren Geciteerd door 8 Verwante artikelen Alle 2 versies

Improved self-attentive Musical Instrument Digital Interface content-based music recommendation system
N Yadav, A Kumar Singh, S Pal - Computational Intelligence, 2022 - Wiley Online Library
... over a real-world **dataset**, that is, the **million song dataset**, to evaluate the performance of our ... **dataset**, we use the **Million Song dataset**, 39 which also contains The Lakh MIDI **dataset**. ...
☆ Opslaan 99 Citeren Verwante artikelen



API Student Paper Selection

Due: Monday October 23rd 2023

Each student has to select a research paper on an audio related subject, that they would like to present during one of the 4 Student Paper Presentation Sessions and submit the pdf of the paper to Brightspace before October 23rd 2023, 23.59h.

Note:

- The subject may be related to your project but this is not mandatory.
- Always select a paper that has been refereed, i.e., is from a scientific journal or scientific conference/workshop proceedings.

- For research papers see for example:
  - ISMIR https://dblp.org/db/conf/ismir/index.html
    - Proceedings: https://www.ismir.net/conferences/
  - Interspeech https://dblp.org/search?q=interspeech
    - Proceedings: https://www.isca-speech.org/archive/
  - Eurasip https://dblp.org/db/journals/ejasmp/index.html
  - And the API-website for further journals

Audio Features Workshop

Available on Wednesday October 18th 2023 (late)