Evaluating the performance of lyrical features relative to and in combination with audio, symbolic and cultural features

Cory McKay

Central question investigated

- What kinds of information are most useful in automatically classifying music?
  - High-level symbolic musical data?
  - Low-level audio data?
  - Cultural data available on-line?
  - Lyrics?
  - Some combination of these?

Presentation overview

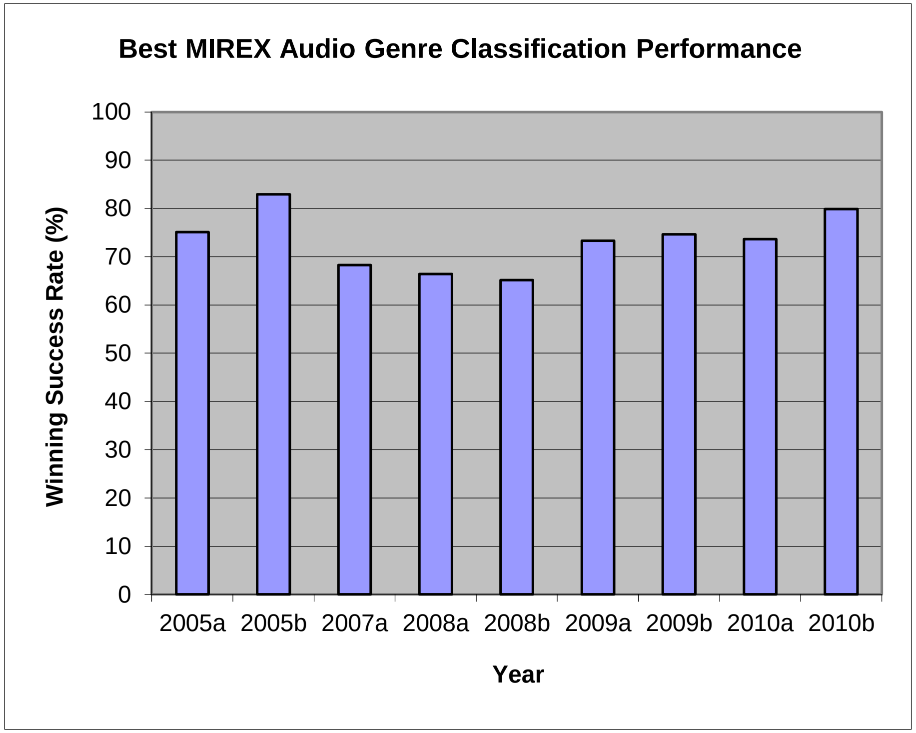
- State of the art of automatic genre classification
- Overview of the jMIR toolkit
  - Harvesting lyrics with lyricFetcher
  - Extracting features with jLyrics
- Experimental results
- Conclusions and future research

Genre classification

- Genre classification is used here as a case study
  - Its difficulty makes it a good evaluative test case
- Genre labels can be broad:
  - Jazz, classical, rock, rap, etc.
- Genre labels can be narrow
  - Microsound, chiptunes, glitch, IDM, etc.

How well can we do?

- The MIREX contest provides the best available way to compare performance
- Best audio genre classification accuracies to date:
  - 6 classes: 82.9% (2005)
  - 10 classes: 79.9% (2010)
- Differences between datasets make it different to fairly compare results, but:
  - There is no evidence of significant improvement from year to year



Note: 2005b involved 6 genres and all other runs involved 10 genres

Software tools used: jMIR

- jMIR is software suite designed for performing music classification research
  - Feature extraction
  - Machine learning
  - Dataset management
  - Data storage formats
- Priorities:
  - Encourage multimodal research
  - Increase accessibility of automatic music classification technologies
  - Standardize and facilitate communication of algorithms and data between research groups
- jMIR is free and open-source
  - Implemented in Java for platform independence

jMIR components

- jAudio: Audio feature extraction
  - 26 core features + metafeatures and aggregators
- jSymbolic: Feature extraction from MIDI files
  - 111 mostly original features
- jWebMiner: Cultural feature extraction
  - Based on web co-occurrence page counts and user tags
- lyricFetcher and jLyric: Lyric harvesting and feature extraction
- ACE: Meta-learning classification system
  - Experiments with dimensionality reduction and machine learning algorithms

- jMIR also includes other components:
  - ACE XML
  - Codaich and Bodhidharma MIDI
  - jMusicMetaManager and jSongMiner
  - jMIRUtilities

jSongMiner

- Software for automatically acquiring formatted metadata about songs, artists and albums
- Designed for use with the Greenstone digital library software
  - May also be used for other purposes, such as cultural feature extraction
- Identifies music files
  - Uses Echo Nest fingerprinting functionality and embedded metadata
- Mines a wide range of metadata tags from the Internet and collates them in a standardized way
  - Data extracted from The Echo Nest, Last.FM, etc.
  - Over 100 different fields are extracted
  - Data may be formatted into unqualified and/or qualified Dublin Core fields if desired
- Saves the results in ACE XML or text
  - Can also be integrated automatically into a Greenstone collection

lyricFetcher

- lyricFetcher automatically harvests lyrics from on-line lyrics repositories
  - LyricWiki and LyricsFly
  - Queries based on lists of song titles and artist names
- Post-processing is applied to the lyrics in order to make remove noise and make them sufficiently consistent for feature extraction
  - Deals with situations where sections of lyrics are abridged using keywords such as “chorus”, “bridge”, “verse”, etc.
  - Filters out keywords that could contaminate the lyrics
- Ruby implementation

jLyrics

- Extracts features from lyrics stored in text files
  - Automated Readability Index
  - Average Syllable Count Per Word
  - Contains Words
  - Flesh-Kincaid Grade Level
  - Flesh Reading Ease
  - Function Word Frequencies
  - Letter-Bigram Components
  - Letter Frequencies
  - Letters Per Word Average
  - Letters Per Word Variance
  - Lines Per Segment Average
  - Lines Per Segment Variance
  - Number of Lines
  - Number of Segments
  - Number of Words
  - Part-of-Speech Frequencies
  - Punctuation Frequencies
  - Rate of Misspelling
  - Sentence Count
  - Sentence Length Average
  - Topic Membership Probabilities
  - Vocabulary Richness
  - Vocabulary Size
  - Word Profile Match
  - Words Per Line Average
  - Words Per Line Variance
- Can also automatically generate word frequency profiles for particular classes if training data is provided
- Central framework implemented in Java
  - Other technologies used by third-party components

Experiment performed

- Can combining features extracted from audio, symbolic, cultural and/or lyrical sources significantly improve automatic music classification performance?
  - Intuitively, they each seem to contain very different kinds of information
- Can this help us break the seeming genre classification performance ceiling?

Experimental methodology

- Extracted features from separate audio, symbolic, cultural and lyrical sources of data
  - Corresponding to the same musical pieces
  - Using the jMIR feature extractors
- Compared ACE-based genre classification performance of each of the 15 possible subsets of these 4 feature groups
  - Audio, Symbolic + Audio, Cultural, Symbolic + Cultural + etc.
  - Applied dimensionality reduction
  - 10-fold cross-validation
    - With reserved validation set
  - Wilcoxon signed-rank significance tests were used

Musical dataset used: SLAC

- The SLAC Dataset was assembled for this experiment
  - Symbolic Lyrical Audio Cultural
  - 250 recordings belonging to 10 genres
  - Audio and MIDI versions of each recording
    - Acquired separately
  - Accompanying metadata that could be used to extract cultural features from the web
  - Lyrics mined with lyricFetcher

Genres in SLAC

- SLAC's 10 genres can be collapsed into 5 genres in order to separately evaluate performance on both moderate and small genre taxonomies
  - Also facilitates evaluation of misclassification severity

- **Blues:** Modern Blues and Traditional Blues
- **Classical:** Baroque and Romantic
- **Jazz:** Bop and Swing
- **Rap:** Hardcore Rap and Pop Rap
- **Rock:** Alternative Rock and Metal

Difficulty of SLAC

- Performances of the same song in different genres
- Performances by the same artists in different genres
- 10-genre taxonomy includes pairs of relatively similar genres
- Diverse styles of music purposely chosen within each sub-genre

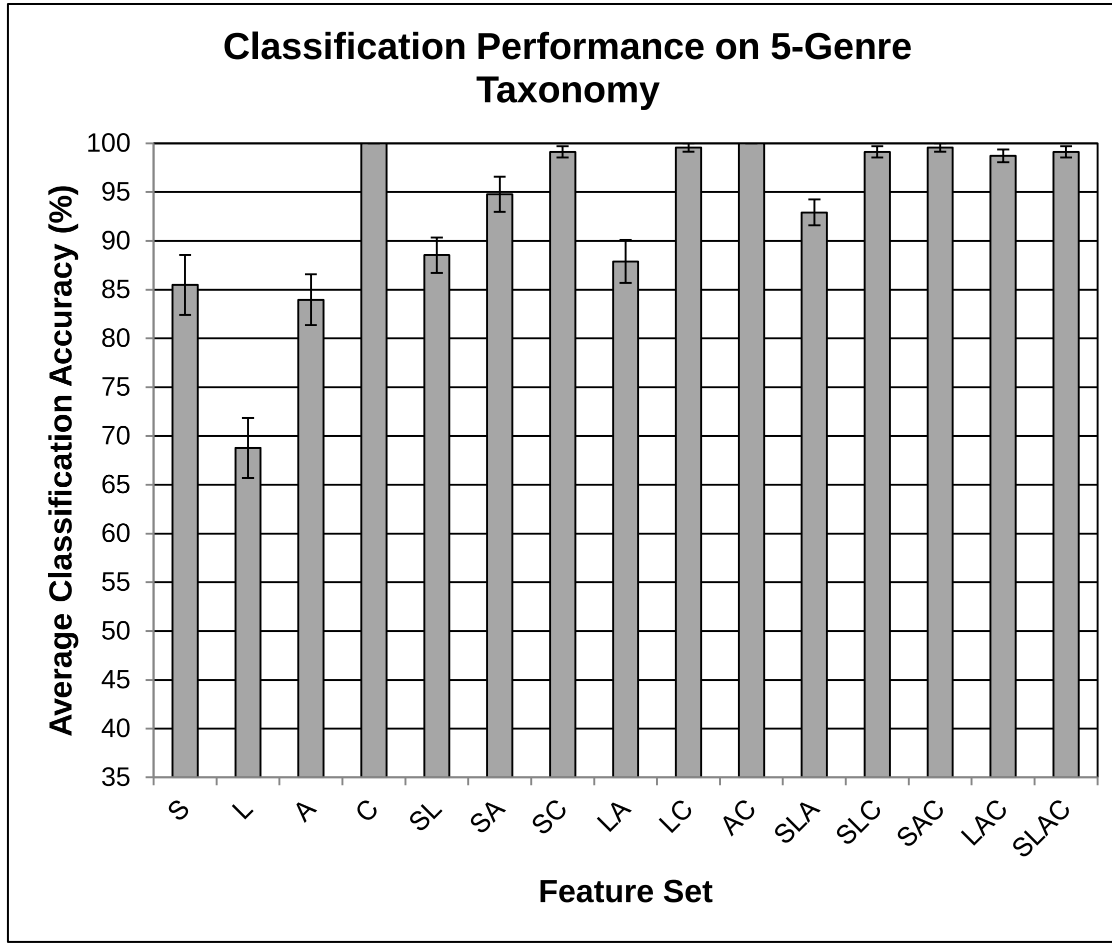
- These factors make SLAC harder than the typical MIREX datasets
  - More realistic, although still easier than real-world application would require

Comparison to 2008 experiment

- We performed a similar experiment in 2008
  - No lyrical features were used
  - Earlier versions of ACE and jWebMiner were used
- Results of these earlier experiments:
  - Combining feature types significantly improved classification results
  - No feature type dominated, although cultural features were particularly good at reducing misclassification seriousness

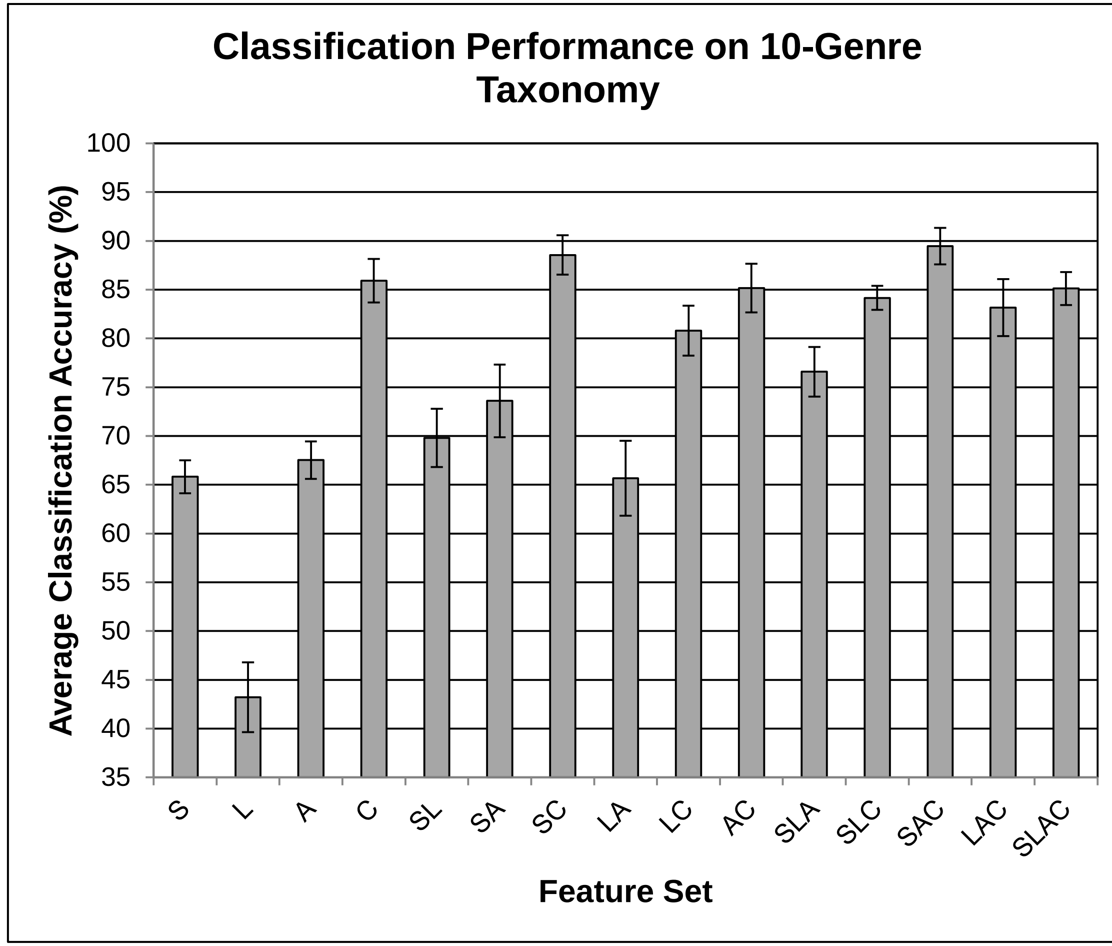
Results: 5-genre taxonomy

- All feature groups involving cultural features achieved classification accuracies of 99% to 100%
- Lyrical features alone underperformed with a classification accuracy of 69%



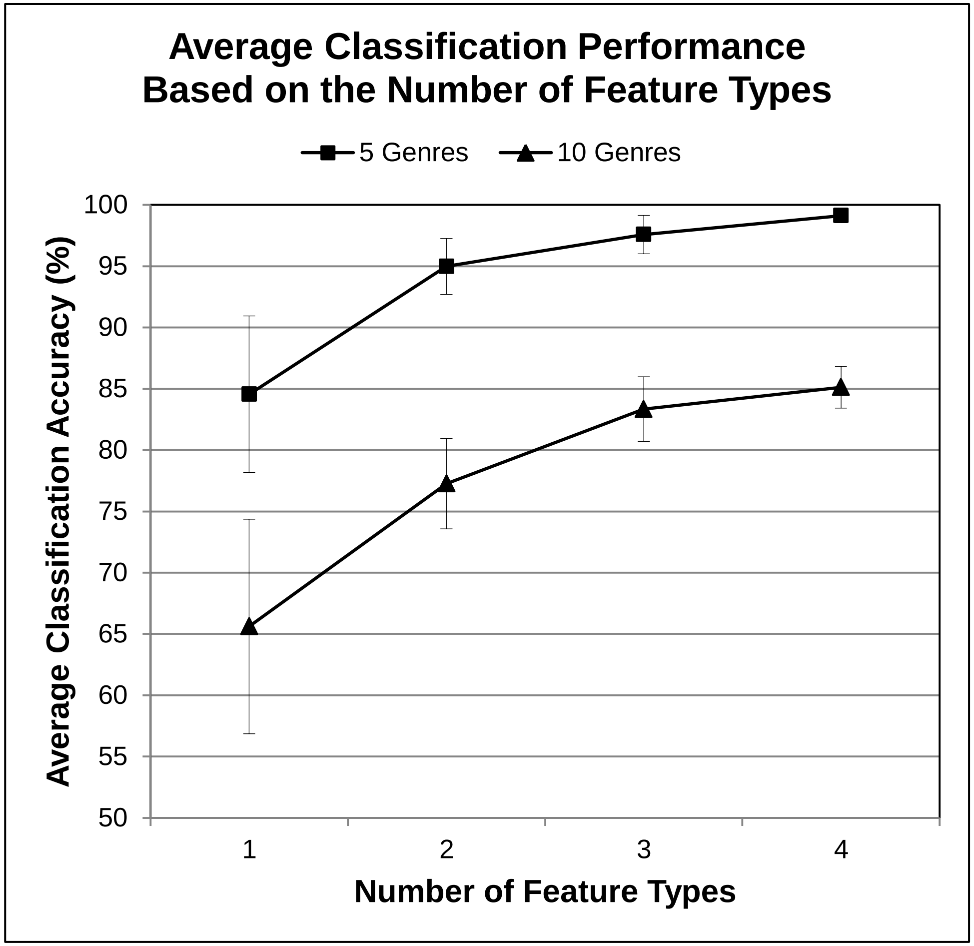
Results: 10-genre taxonomy

- SAC achieved the best classification accuracy of 89%
- All feature groups that included cultural features achieved 81% or higher
- Lyrical features alone again underperformed at 43%



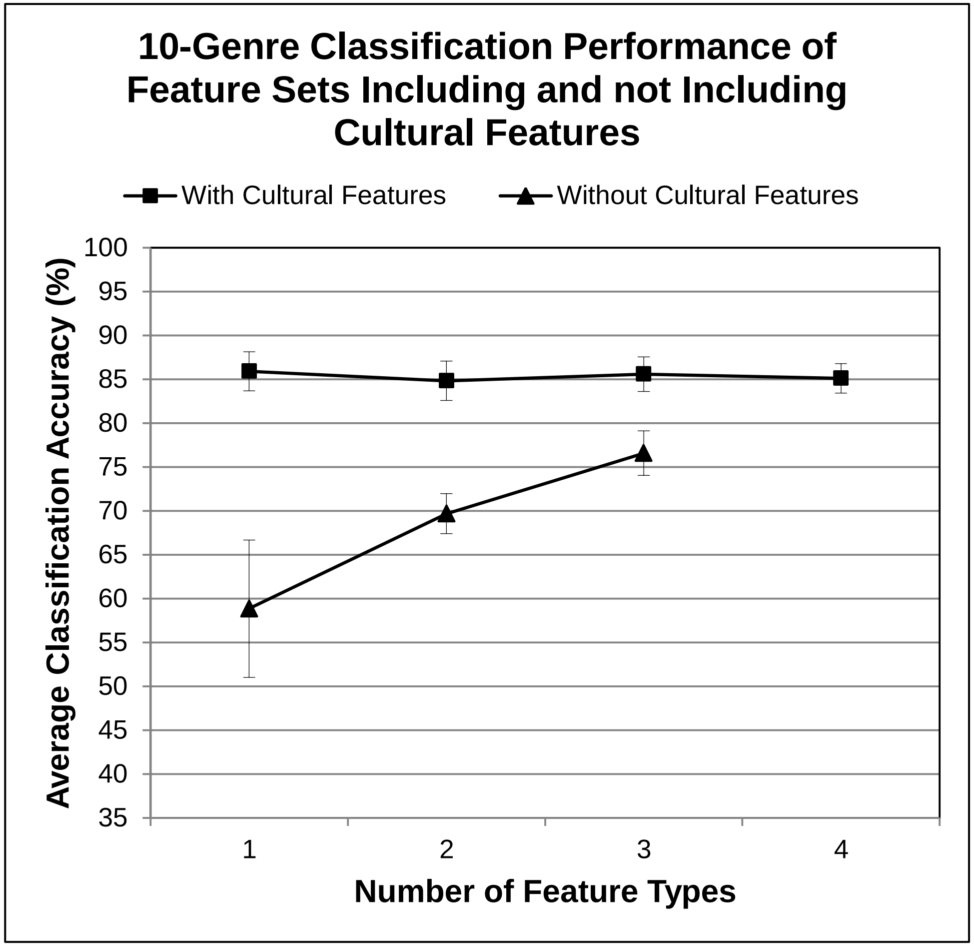
Discussion: Combining feature types

- Combining features types tended to increase classification performance on average
- However, there were exceptions
  - e.g. LC performed significantly less well than C in the 10-genre experiment



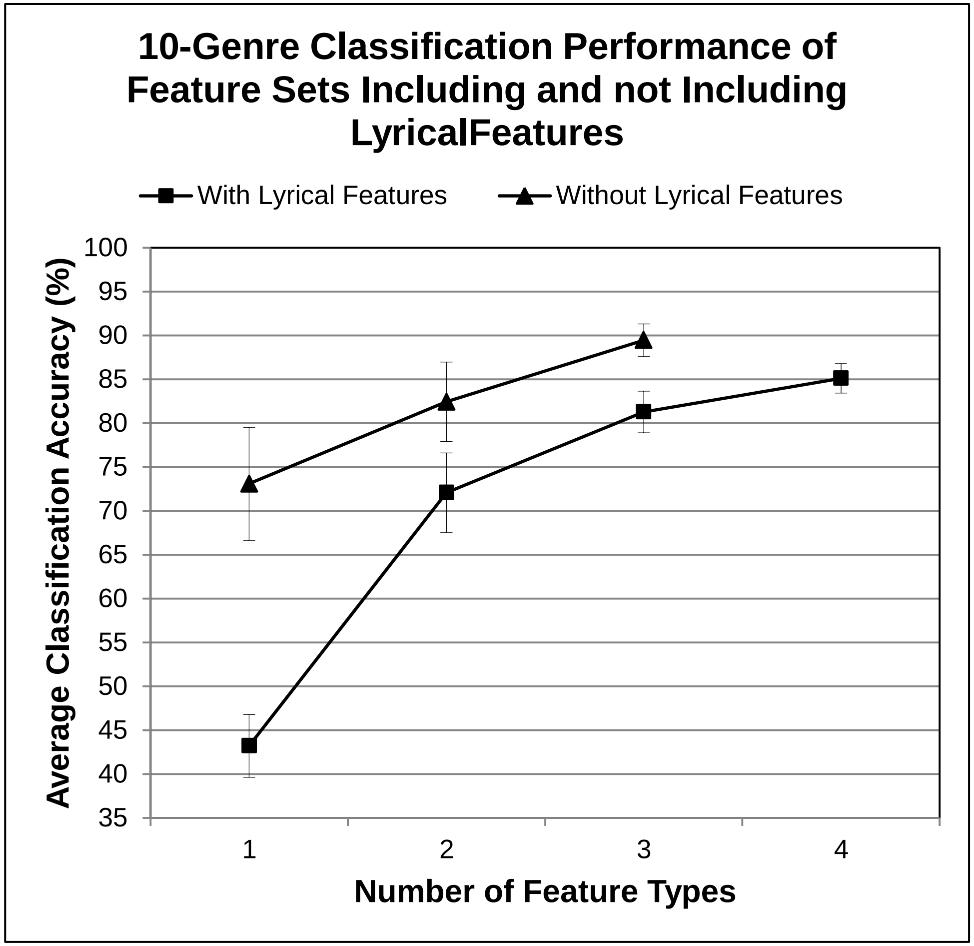
Discussion: Cultural features

- Cultural features significantly outperformed other feature types
- For the 10-genre taxonomy, all groups including cultural features outperformed all groups of the same size that did not include cultural features
- This dominance of cultural features was not evident in the 2008 SAC experiments
  - jWebMiner 2.0 (used here) substantially improved the performance of cultural features by combining search engine data with Last.FM data



Discussion: Lyrical features

- Lyrical features significantly underperformed other feature types
  - Partly explained by the necessity of classifying instrumental music
- For the 10-genre taxonomy, all groups including lyrical features underperformed all groups of the same size that did not include lyrical features
- Lyrical features did improve results in most cases where cultural features were not involved, however
  - e.g. SLA performed better than S, L, A, SL, SA or AC



Conclusions

- We obtained excellent overall genre classification results
  - 89% on 10 genres, compared to the best MIREX audio-only result to date of 80% on 10 genres
- Combining feature types often improved results
- Cultural features dominated
  - The particular jWebMiner 2.0 combination of features extracted from both web content and Last.FM user tags was extremely effective
- Lyrical features can improve results, but performed poorly individually relative to other feature types

Future research questions

- Should we focus research efforts on fingerprinting and cultural feature extraction rather than bothering with extracting features from audio and lyrics?
  - Assuming reliable fingerprinting, this could result in very high classification results
  - However, this marginalizes the musicological and music theoretical insights about musical categories that can be achieved from content-based analysis
- Can the performance of lyrical features be improved
  - Better cleaning and standardization of raw lyrics
  - More sophisticated features designed specifically with music in mind
    - The current jLyrics features consist of general-purpose text mining features

Acknowledgements

- Funding, past and present:
  - CIRMMT
  - Centre for Open Software Innovation
  - Andrew W. Mellon Foundation
  - Social Sciences and Humanities Research Council of Canada
  - Canadian Foundation for Innovation

- Contact information:
  - jmir.sourceforge.net
  - cory.mckay@mail.mcgill.ca

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

nema

McGill

Schulich School of Music
École de musique Schulich

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology

DDMAL

COSI Centre for Open Software Innovation

Greenstone