Music Similarity and Cover Song Identification: The Case of Jazz

Simon Dixon and Peter Foster

s.e.dixon@qmul.ac.uk

Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

Queen Mary University of London

centre for digital music

Outline

- Musical similarity and cover song detection
- Standard approaches to cover song detection
- Information-theoretic measures of similarity
  (Foster, Dixon and Klapuri, IEEE/ACM Trans. ASLP 2015)
- Concluding thoughts

MIR and Similarity

- The music industry thinks you will buy music that is similar to music that you have bought in the past
- This has inspired the Music Information Retrieval community (computer scientists, music psychologists/librarians/experts, engineers, etc.) to invest considerable effort investigating similarity in music
- Assessing the similarity of pairs of audio recordings is of particular interest, as it solves the "cold-start" problem (lack of data concerning new or unknown music items)
- For music recommendation and playlist generation, the task is often expressed as: "given a seed song, return a song or songs that is/are (most) similar"
- What is music similarity? What makes two recordings similar?

Defining Similarity

- Music has many dimensions or aspects: melody, rhythm, harmony, instrumentation, timbre, lyrics, genre, style, mood
- Assessing similarity along any one dimension is subjective
- Reducing similarity to a scalar value makes it extremely ill-posed
- Casey et al. (Proc. IEEE, 2008) describe a spectrum of *specificity* in MIR tasks
  - Highly specific: identification of specific recordings (fingerprinting) for copyright monitoring, royalty assignment
  - Remixes, versions and imitations
  - Performances of the same piece
  - Pieces by the same artist or composer, that sound similar, or match the same user's listening profile
  - Low specificity: similar mood, genre, instrumentation; influence
- The context or application defines what is meant by similarity

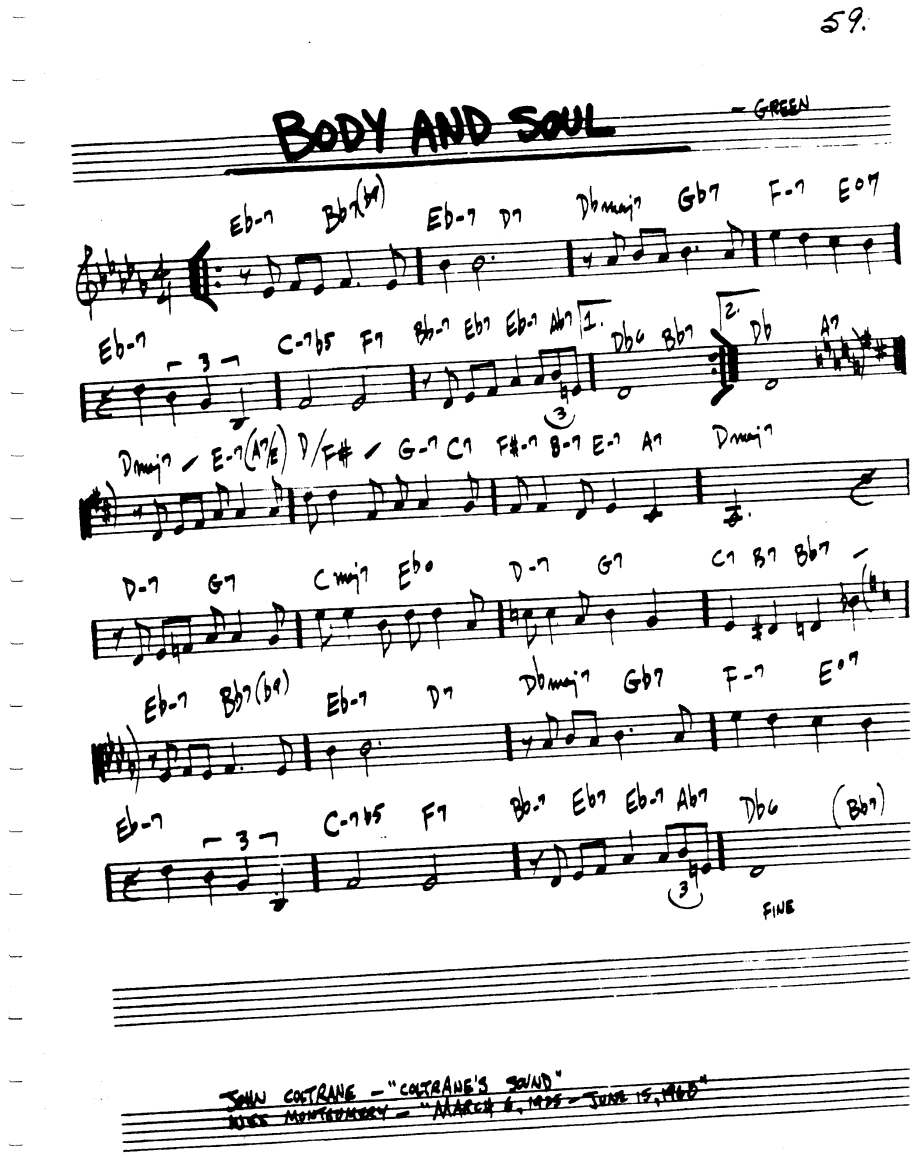
Cover Songs

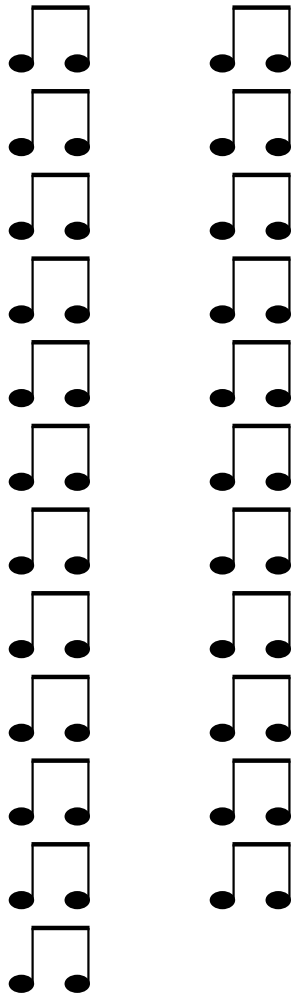
- Most pop songs have a canonical “original” recorded version
- Other musicians who perform the song are creating a “cover version” or just “cover”
- Motivation may be as tribute, parody, or means of artistic expression, or to obtain recognition
- Covers range from being almost indistinguishable from the original to being unrecognisable
- Some aspects of the original are preserved, some are modified
- When looking for covers, we don’t know in advance *which* aspects are going to be preserved
- Genre dependence: within pop, harmony and (to a lesser extent) melody are likely to be preserved, along with lyrics

Standards and Covers

- In jazz and many traditional/folk music styles (e.g. flamenco, Irish), there is a shared repertoire of commonly-performed works, allowing musicians who have never met to perform together
- These may be passed on orally, or captured in notation (usually no more than melody, chords and lyrics, i.e. the *lead sheet*)
- Collections of lead sheets appear in real/fake books
- Performances of such *standards* can be considered as cover versions, even where there is no definitive original version
- Jazz allows/expects ornamentation and transformation of the melody as well as substitution of chords in the harmony
- Good MIR task: the ground truth is relatively easy to determine

Example


59.
BODY AND SOUL
- GREEN
JOHN COLTRANE - "COLTRANE'S SOUND"
WES MONTGOMERY - "MARCH 6, 1925 - JUNE 15, 1968"



Standard Approaches in MIR

- Bag-of-features approaches are suitable mainly for low-specificity tasks (e.g. genre, mood)
- Temporal features can represent time-varying tonal content
  - frequency, pitch or chroma
  - predominant melody or harmony (pitches or chords)
- Adapt features to allow for variation in key or tempo
  - circular shift of chroma vectors
  - beat synchronous features
- Perform pairwise sequence matching
  - dynamic programming (edit distance) on similarity matrix
  - correlation
  - local alignment: as there is no guarantee that different versions share the same structure

Information-Theoretic Approach to Similarity

- Similarity can be viewed as *predictability*
- That is, given information about one piece, how predictable does the other piece become
- If the underlying composition is the same, the given information should increase the predictability of the second piece
- Music psychologists have reflected on modelling prediction with information theory
- We compare various approaches: discrete valued vs continuous data; compression vs prediction; correlation and bag-of-feature baselines
- Our approach is based on representing pieces by sequences of features encoding the harmonic content (chroma)

Discrete-Valued Approaches

- In information-thoeretic terms, predictability means redundancy
- Predictability can be estimated by data compression: an optimal compression algorithm removes all redundancy, leaving only the information content
- Joint compressibility quantifies the similarity between pairs of sequences
- Normalised compression distance (NCD) approximates algorithmic information content (Kolmogorov complexity):

$$\mathrm{NCD}(x, y) = \frac{\max\{C(xy) - C(x), C(yx) - C(y)\}}{\max\{C(x), C(y)\}}$$

- Extend with interleaving of sequences rather than concatenation, with circular shift to maximise the correlation of the sequences

centre for digital music

Queen Mary University of London

Continuous-Valued Approaches

- Predictive approach based on previous context (self-prediction) or model of other sequence (cross-prediction), or both (conditional self-prediction)
- Temporal context is encoded with time-delay embedding, and prediction performed using the nearest neighbour
- The sequence of prediction errors is normalised and statistics of the sequence are computed
- This approach can also be applied to discrete-valued sequences
- Alternative approach using conditional entropy (predictability) instead of compressibility

Data

- Jazz box sets from `zweitausendeins.de`
- With metadata in CSV file (!)
- 300 recordings of 97 pieces
- Relaxed definition of “cover”: we do not attempt to distinguish the original, nor the artist (self-covers are allowed)
- Two recordings are a “cover pair” iff their titles are identical
- Further (large-scale) experiments were performed on the Million Song Dataset: see our paper for details

Features

- 12-dimensional beat-synchronous chroma (preferred beat rate: 240 BPM)
- pitch adjusted within ± 0.5 semitones to allow for reference frequencies other than A4 = 440 Hz
- one sequence transposed to maximise inner product of global average chroma between the pair
- K-means applied (with various codebook sizes up to 48) for discrete methods

Methods

- Compression: off-the-shelf standard algorithms (LZ, BW, PPM)
- Prediction: PPMC, LZ78
- Continuous prediction: various parameterisations of time-delay embedding
- Normalisation to remove hubs
- For larger scale experiments, a filter-and-refine approach was used: fast histogram-based approach then the temporal sequence methods on the best matches

Results: Discrete-Valued Approaches

- Evaluated in terms of mean average precision (MAP)
- Numbers are in the paper: not comparable across datasets
- Compression-based: Interleaving of sequences helped, except for block-based compression algorithms, but ...
- Histogram (BoF) baseline outperformed compression-based techniques (but not on the larger dataset)
- Discrete cross-prediction was even better in most cases

Results: Continuous-Valued Approaches

- Continuous cross-prediction was better than conditional self-prediction and better than all discrete approaches
- Baseline cross-prediction method performed equally well
- Combination of our approach with the baseline gave significant improvement
- Baseline cross-correlation approach also performed well on the jazz set, but not on the extended set
- State-of-art results were obtained on the Million Song Dataset

Conclusions and Future Work

- Information-theoretic measures do capture some aspects of medium-specificity music similarity
- Codebook for discrete-valued approach is not musically motivated
- Future work: compare with a more “musical” representation: automatically generated chord symbols
- Continuous-valued vectors are also uninterpreted in these approaches
- Extend experiments to complete jazz dataset (10000 recordings, 1–31 covers)
- Full methods are slow: test filter-and-refine for jazz

Any Questions?

- Acknowledgements / references
  - Peter Foster, Simon Dixon and Anssi Klapuri: *Identifying Cover Songs Using Information-Theoretic Measures of Similarity*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 6, 2015, pp 993-1005
  - Peter Foster: *Information-Theoretic Measures of Predictability for Music Content Analysis*, PhD Thesis, Queen Mary University of London, School of Electronic Engineering and Computer Science, 2015