HAJIM
SCHOOL OF ENGINEERING
& APPLIED SCIENCES
UNIVERSITY of ROCHESTER

What Sounds So Good? Maybe, Time Will Tell.

EASTMAN
SCHOOL OF MUSIC
UNIVERSITY of ROCHESTER

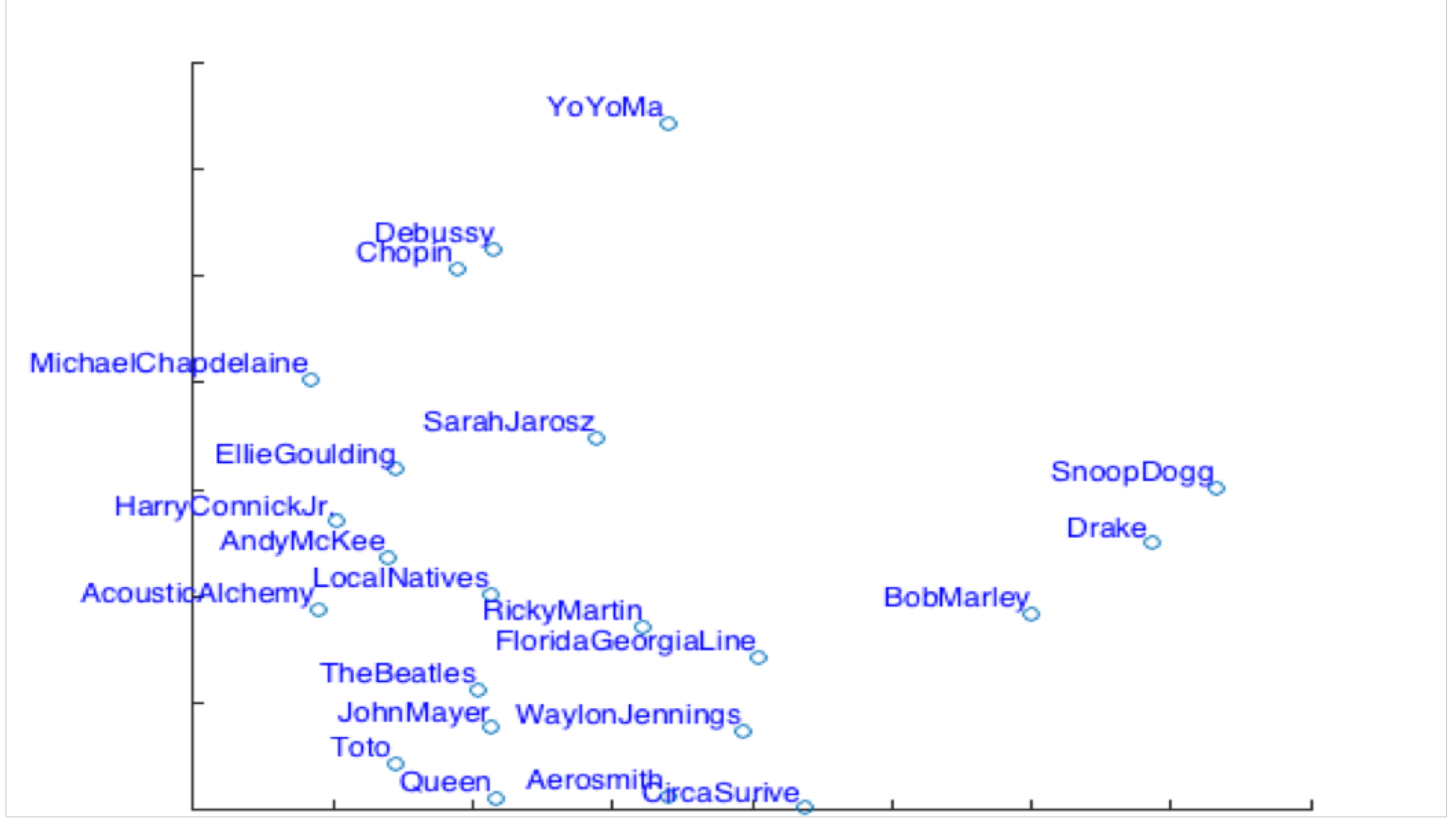
Steven Crawford
University of Rochester, Department of Electrical and Computer Engineering - Music Research Lab

Abstract

One enduring challenge facing the MIR community rests in the (in)ability to enact measures capable of modelling perceptual musical similarity. This research examines techniques for assessing musical similarity. More specifically, we explore the notion of designing a system capable of modeling the subtle nuances intrinsic to particular performances. Presently, the pervading method for establishing an indication of musical similarity is via the Mel Frequency Cepstral Coefficient. However, some de-facto MFCC methods jettison pertaining temporal information with first moment calculations & frame clustering. The discarded information has subsequently been shown to be of critical relevance to musical perception/cognition. To this end, we elucidate the fundamental need for the inclusion of temporal information within our models. We propose a novel approach emphasizing sequential repetition of perceptually relevant expressive features and compare with results obtained from several instantiations of spectral-based MFCC methods.

Introduction: Perception / Motivation

1. Sound has always been an integral component in the successful proliferation of our species. Our auditory systems have evolved over hundreds of thousands of years with specific temporal acuities.
2. We must therefore recognize the importance that temporal information might play in our perception of music; a phenomenon based entirely in and of sound.
3. Rhythm organizes the movement of musical patterns linearly in time and repetitive sequences, absolutely dependent on temporal relationships, are vital for perceived musical affect.
4. Sequential repetition has been shown to be of critical importance for emotional engagement in music.



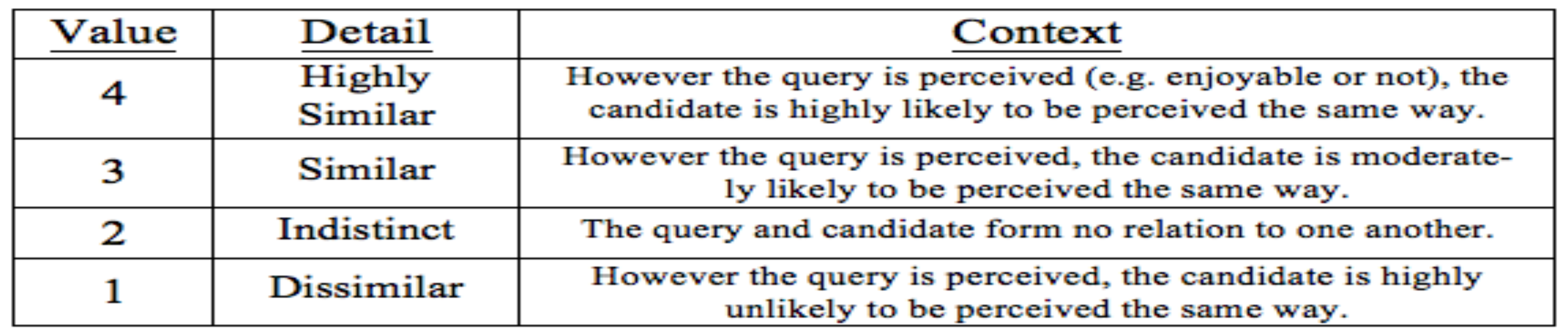
Erdös distance, a function of transitive similarity, evaluates the similitude of two performers (A&B) as the number of interposing performers required to create a connection from A to B. Above orientation derived via gradient descent with Multidimensional Scaling (MDS).

"Bag-of-Frames"

- MFCC's are a computationally inexpensive model of timbre.
- Studies have shown timbre to be perceptually significant in genre identification and classification.
- Some MFCC models disregard temporal ordering and describe the audio as a global distribution of short term spectral information much like a histogram would describe the distribution of colors used in a painting.
- Typical MFCC feature extraction procedure:
  - i. FT of signal.
  - ii. Map log amplitudes of spectrum to the Mel-scale.
  - iii. DCT of Mel log-amplitudes.
  - iv. MFCC's are amplitudes of resulting spectrum.

MFCC Models

- Single Multivariate Gaussian – completely static model based on mean and covariance of MFCC's.
- Double Multivariate Gaussian – addition of temporal information in the form of ΔMFCC's.
- Fluctuation Patterns – an additional mapping of initial 36 Mel-bands into 12 "perceptually salient" sub-bands. Track frequency specific loudness modulations over time.



Value	Detail	Context
4	Highly Similar	However the query is perceived (e.g. enjoyable or not), the candidate is highly likely to be perceived the same way.
3	Similar	However the query is perceived, the candidate is moderate-ly likely to be perceived the same way.
2	Indistinct	The query and candidate form no relation to one another.
1	Dissimilar	However the query is perceived, the candidate is highly unlikely to be perceived the same way.

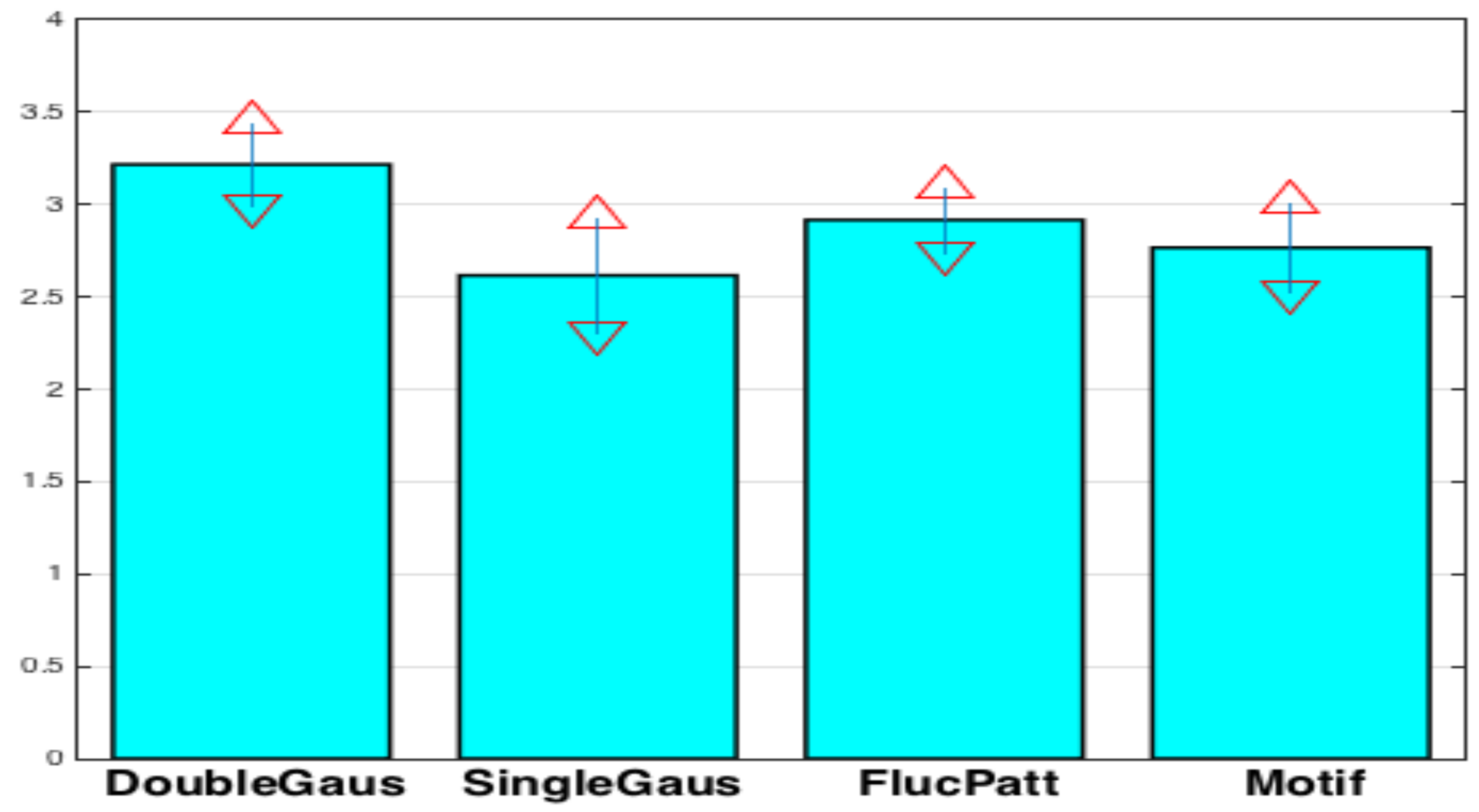
BROAD scale used by listening participants

Sequential Motif Discovery

- Songs are modelled by frequently recurring chronological patterns (motifs).
- Patterns are encoded into strings of data describing extracted features and serving as a stylistic representation of a song.
- Encoded string format enables the luxury of sequence alignment tools from bioinformatics.
- Similarity is quantified as the amount of overlapping motifs between songs.
- The system is composed of three major units:
  - i. Audio Segmentation (according to musical beats).
  - ii. Feature extraction (loudness, vibrato, timing offsets).
  - iii. Quantization / pattern analysis (extracted features are discretized and segmented. Dimensionality compression converts our symbol 4D strings into 1D sequences required by our bioinformatics alignment tools).

Evaluation

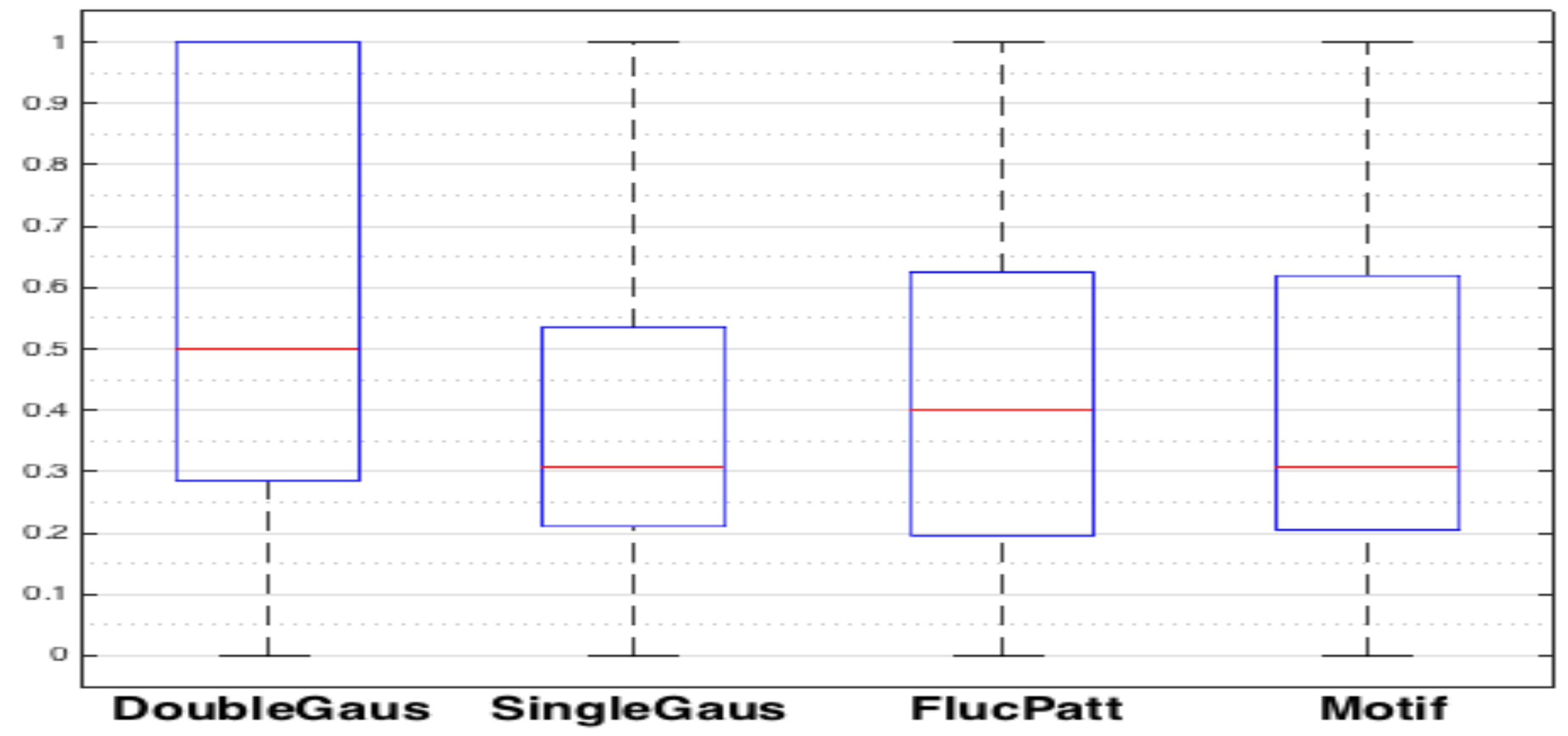
I. Human Listening (BROAD score).
II. Genre Similarity (Mean % of Genre matches).
III. F-measure over top 3 candidates.



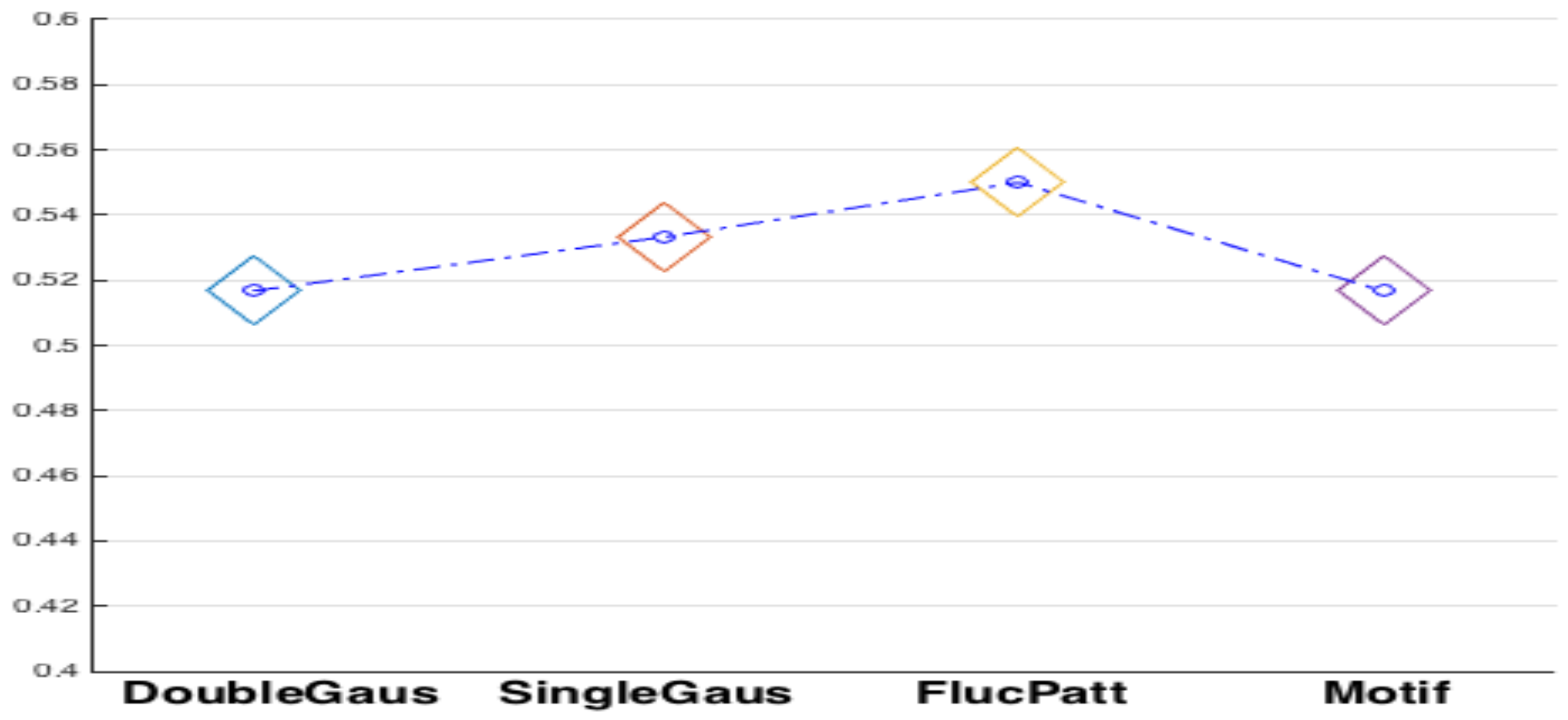
Average BROAD assessment for each system, computed across all listening participants.

Discussion

- Inclusion of dynamic, temporal information increases system performance.
- Overwhelming variability of musical expression has outgrown the classification bandwidth of the 'genre.' Perhaps a more authentic approach at describing (and recommending) similar music would be in terms of 'mood' or 'occasion.'



F-measures of each system. *On each boxplot, the red line represents the median, the ends of each box denote the 25th and 75th percentiles, the whiskers extend to the extreme data points.*



Average system performance according to genre matches between seed query and 1st candidate returns.