What Sounds So Good? Maybe, Time Will Tell.

Steven Crawford
University of Rochester
steven.crawford@rochester.edu

ABSTRACT

One enduring challenge facing the MIR community rests in the (in)ability to enact measures capable of modelling perceptual musical similarity.

In this paper, we examine 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, and probability models. This discarded information has subsequently been shown to be of critical relevance to musical perception and cognition. To this end, we elucidate the fundamental need for the inclusion of temporal information within a similarity model.

We propose a novel content-based approach emphasizing the sequential repetition of perceptually relevant expressive musical features and compare with results obtained from several instantiations of spectral-based MFCC methods.

1. INTRODUCTION

1.1 Perception

How can we define music similarity? Such a subjective and abstract idea is challenging to articulate. Colloquially speaking, music similarity references a laundry list of notions. From performance style to rhythmic complexity, perhaps harmonic progression or melodic variation, quite possibly timbral content and tempo; and the list goes on. Still, behind this ambiguity, we can be sure that the pith of any similarity judgment is rooted in cognition.

In order to conceive an effective computational model of what music similarity is, human cognition and perception must be taken into consideration. What information is essential to our formation of complex auditory scenes capable of affecting temperament and disposition?

Further exploration into these questions will inevitably bring about a richer and more perceptually relevant computational model of music. Attempting to actualize machines capable of auditioning music similar to humans can only enhance our endeavor of understanding what it means to hear music.

1.2 Motivation

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 [35]. For instance, sudden onsets of rapidly dynamic sounds trigger feelings of anxiety and unpleasantness [11]. Our brains are hardwired to interpret expeditiously occurring patterns of sound as indicative of dangerous or threating circumstances [14]. Sensitivities to temporally fluctuating aural information has thus proven beneficial to our survival [5].

We must therefore recognize the importance that temporal information might play in our perception of music; a phenomenon based entirely in and of sound. Rhythm organizes the movement of musical patterns linearly in time and repetitive sequences, absolutely dependent on temporal relationships, are vital for perceived musical affect [15]. In fact, sequential repetition has been shown to be of critical importance for emotional engagement in music [28]. The perceptual bases of musical similarity judgments correlate strongly with temporal tone patterns and spectral fluctuations of said tones through time (ASDR) [13], while significant musical repetitions are crucial to metrical and contrapuntal structure [32].

2. "BAG-OF-FRAMES"

2.1 Rationale

Mel Frequency Cepstral Coefficients are standard operating procedure for speech processing. They essentially present the spectral shape of a sound. Through some basic domain manipulation and a Fourier-related transform (DCT), the MFCC can drastically reduce the overall amount of raw data, while maintaining the information most meaningful to human perception (i.e. Cepstrum is approximately linear for low frequencies and logarithmic for higher ones) [24]. However, speech and music, while similar in certain communicative aspects, differ widely in most dimensions [21]. So how has the MFCC become the leading contender to model our music?

The MFCC is a computationally inexpensive model of timbre [33]. Studies have shown that there is a strong connection between perceptual similarity and the (monophonic) timbre of single instrument sounds [17]. Polyphonic timbre has also been shown to be perceptually significant in genre identification and classification [12]. However, most MFCC models disregard temporal ordering, they're static. They describe the audio as a global distribution of short term spectral information [3], much

like a histogram would describe the distribution of colors used in a painting.

2.2 Evolution

Initially proposed by Jonathan Foote in 1997, use of the MFCC as a representative measure of musical similarity [10] has seen several innovative modifications. Refining Foote's global clustering approach, Logan and Salomon propose a localized technique where the distance between two spectral distributions (mean, covariance, and cluster weight) is seen as a similarity measurement and computed via Earth Movers Distance (EMD) [19]. EMD evaluates the amount of work ($d_{p_i q_j}$) required to convert one model into the other as well as the cost of performing said conversion ($f_{p_i q_j}$). Here, work is defined as the symmetrized KL-divergence [31].

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{p_i q_j} f_{p_i q_j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{p_i q_j}} \quad (1)$$

The next prominent contribution, set forth by Aucouturier and Pachet, uses a Gaussian Mixture Model (GMM) in synchrony with Expectation Maximization (initialized by k-means) for frame clustering. Ultimately, a song is modelled with three 8-D multivariate Gaussians fitting the distribution of the MFCC vectors. Similarity is assessed via a symmetrized log-likelihood of (Monte-Carlo) samples from one GMM to another [2].

$$sim_{a,b} = \frac{1}{2}\,[p(a|b) + p(b|a)] \quad (2)$$

Mandel and Ellis simplified the aforementioned approach and modelled a song using a single multivariate Gaussian with full covariance matrix [20]. The distance between two models, considered the similarity measurement, is computed via symmetrized KL-divergence. In the following equation, $\theta_{i,j}$ are Expectation Maximized parameter estimations of the mean vectors and full covariance matrices [23].

$$D[p(x|\theta_i), p(x|\theta_j)] = \int_{-\infty}^{\infty} p(x|\theta_i)\log\left(\frac{p(x|\theta_i)}{p(x|\theta_j)}\right) dx + \int_{-\infty}^{\infty} p(x|\theta_j)\log\left(\frac{p(x|\theta_j)}{p(x|\theta_i)}\right) dx \quad (3)$$

2.3 Glass-Ceiling

The MFCC-based similarity model has since seen several parameter modifications and subtle algorithmic adaptations, yet the same basic architecture pervades [3]. Front end adjustments (e.g. dithering, sample rate conversions, windowing size, etc.) have been implemented in conjunction with primary system variations (e.g. number of MFCC's used, number of GMM components to a model, alternative distance measurements, Hidden Markov Models over GMM's, etc.) in an attempt towards optimization [1].

Nonetheless, these optimizations fail to provide any significant improvement beyond an empirical glass-ceiling [4] [25]. The simplest model (single multivariate Gaussian) has actually been shown to outperform its more complex counterparts [20]. It would appear that results from this approach are bounded which suggests the need for an altogether new interpretation.

Moreover, bag-of-frames systems inadequately attempt to model perceptual dependencies as statistical occurrences. It is quite possible, even likely, that a frame appearing with very low statistical significance contains information vital for perceptual discernment. Hence, this engineering adaptation towards modeling human cognition is not ideally equipped for polyphonic music and future enhancements will ultimately result from a more complete perceptual and cognitive understanding of human audition [3].

3. IMPLEMENTATION

3.1 Progression

The following sections will explore the implementations and results of three MFCC based bag-of-frames similarity systems. Beginning with a wholly static instantiation (i.e. zero temporal or fluctuating spectro-temporal information), we proceed to system instantiations incorporating ΔMFCC's and fluctuation patterns as enhancements to the standard MFCC's. As a novel and contrasting perspective, the section closes with our temporally dependent sequential model.

3.2 Single Multivariate Gaussian

In the initial, completely static model, we segment the audio into 512-point frames with a 256-point hop (equating to a window length of 23ms at a sample rate of 22050Hz) [20]. From each segmented frame, we extract the first 20 MFCC's as follows:

I. Transform each frame from time to frequency via DFT [$S_i(k)$], where $s_i(n)$ is our framed, time based signal, h(n) is an N sample long Hanning window, and K is the DFT length.

$$\mathrm{S_i(k)} = \sum_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{s_i(n)h(n)e^{-j2\pi kn/N}} \quad 1 \le \mathrm{k} \le \mathrm{K} \quad (4)$$

II. Obtain the periodogram-based power spectral estimate [$P_i(k)$] for each frame.

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \quad (5)$$

III. Transform the power spectrum along the frequency axis into its Mel representation consisting of triangular filters [M(f)], where each filter defines the response of one band. The center frequency of the first filter should be the starting frequency of the second, while the height of the triangles should be 2/(freq. bandwidth).

$$M(f) = 1125 log_{10}(1 + {}^{f_{Hz}}/_{700}) \quad (6)$$

IV. Sum the frequency content in each band and take the logarithm of each sum.
V. Finally, take the Discrete Cosine Transform on the Mel value energies, which results in the 20 MFCC's for each frame.

Next, we compute a 20x1 mean vector and a 20x20 covariance matrix and model a single multivariate Gaussian from the data [20]. This process is repeated over every song in the database while a for loop iterates over each model, calculating the symmetrized KL-divergence.

3.3 Two Multivariate Gaussians

This model, proposed by de Leon and Martinez, attempts to enhance baseline MFCC performance with the aid of some dynamic information (i.e. its time derivative, the ΔMFCC) [7]. The novelty here is in the modelling approach towards the ΔMFCC's. As opposed to directly appending the time derivatives to the static MFCC information, an additional multivariate Gaussian is employed [7].

The motivation here being the simplification of distance computations required to ultimately quantify similarity (i.e. symmetrized KL-divergence). For a d-dimensional multivariate normal distribution (N) described by an observation sequence (x), a mean vector (μ), and a full covariance matrix (Σ),

$$N(x|\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} e^{-\frac{1}{2}(x\text{-}\mu)^T \Sigma^{-1}(x\text{-}\mu)} \quad (7)$$

there exists a closed form KL-divergence of distributions p and q [27]:

$$2KL(p||q) = \log\frac{|\Sigma_q|}{|\Sigma_p|} + Tr\left(\Sigma_p^{-1}\Sigma_p\right) + (\mu_p\text{-}\mu_q)^T\Sigma_p^{-1}(\mu_p\text{-}\mu_q)\text{-}d \quad (8)$$

where $|\Sigma|$ now denotes the determinant of the covariance matrix.

In this approach, audio is segmented into 23ms frames and the first 19 MFCC's are extracted in the same fashion as previously described. A single multivariate normal distribution is then modelled on the 19x1 mean vector and the 19x19 covariance matrix. The ΔMFCC (d) at time (t) is then computed from the cepstral coefficient (c) using a time window (Θ).

$$d_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta}\text{-}c_{t\text{-}\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2} \quad (9)$$

An additional Gaussian is modelled on the ΔMFCC's and the song is ultimately characterized by two single multivariate distributions. Symmetrized KL-divergence is used to compute two distance matrices; one for the MFCC's and one for the ΔMFCC's. Distance space normalization is applied to both matrices and a full distance matrix is produced as the result of a weighted, linear combination of the two [7].

3.4 Fluctuation Patterns

To augment the standard MFCC effectiveness, Elias Pampalk suggests the addition of some dynamic information correlated with the musical beat and rhythm of a song [26]. Fluctuation patterns essentially attempt to describe periodicities in the signal and model loudness evolution (frequency band specific loudness fluctuations over time) [25]. To derive the FP's, the Mel-spectrogram is divided into 12 bands (with lower frequency band emphasis), and a DFT is applied to each frequency band to describe the amplitude modulation of the loudness curve [25]. The conceptual basis of this approach rests on the notion that perceptual loudness modulation is frequency dependent [26]. Implementation of the system is as follows:

I. The Mel-spectrum is computed using 36 filter banks and the first 19 MFCC's are obtained from 23ms Hanning windowed frames with no overlap.
II. The 36 filter banks are mapped onto 12 bands. Fluctuation patterns are obtained by computing the amplitude modulation frequencies of loudness for each frame and each band via DFT.
III. Finally, the song is summarized by the mean and covariance of the MFCC's in addition to the median of the calculated fluctuation patterns.

3.5 Sequential Motif Discovery

Attempting to coalesce spectrally extracted features with temporal information, our approach characterizes a song by frequently recurring chronological patterns (motifs). These patterns are encoded into strings of data describing extracted features and serving as a stylistic representation of a song. The encoded string format enables us 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; audio segmentation, feature extraction, and quantization / pattern analysis [30].

3.5.1 Audio Segmentation

In this module, with the aid of an automatic beat tracking algorithm, we segment the audio into extraction windows demarcated by musically rhythmic beat locations [9]. Each window subsequently consists of the audio interval between two beat locations [30].

I. Estimate the onset strength envelope (the energy difference between successive Mel-spectrum frames) via:
    a. STFT
    b. Mel-spectrum transformation
    c. Half-wave rectification
    d. Frequency band summation
II. Estimate global tempo based on onset curve repetition via autocorrelation.

III. Identify beats as the locations with the highest onset strength curve value. Ultimately, the beat locations are decided as a compromise between the observed onset strength locations and the maintenance of the global tempo estimate [9].

3.5.2 Feature Extraction

Essentially, each extraction window serves as a "temporal snapshot" of the audio, from which quantitative measurements corresponding to perceptually relevant (loudness, vibrato, timing offsets) features are extracted. Each of these features is chosen in hopes of a qualitative representation of genre and/or expressive performance style (e.g. the abiding loudness levels pervading rock and hip-hop, vibrato archetypical of the classical styles, the syncopations of jazz, reggae, and the blues).

- Loudness here $[v_L]$ is defined as the sum of all constituent frequency components in an STFT frame $[S_M(i,k)]$ and computed as the time average (i.e. the total number of frames in the extraction window, M) of logarithmic perceptual loudness.

$$v_L = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left[\sum_{k=1}^{K} S_M^2(i,k)\right]} \qquad (10)$$

- The vibrato detection algorithm (an instantiation of McAulay-Quatieri analysis) begins by tracking energy peaks in the spectrogram. From these peaks, several conditional statements are imposed onto the data. Upon conditional satisfaction, vibrato is recognized as being present in the extraction window. In short, the algorithm begins from a peak frequency frame(i)/bin(m) location [i.e. f(i,m)], and compares values in subsequent frame/bin locations [f(j,n)] in an attempt to form a connection path (rising or falling) identifiable as vibrato. A more detailed algorithmic explanation can be found in [22].
- Timing offsets are identified as deviations between the aforementioned derived beat locations and significant spectral energy onsets. They symbolize the superimposition of rhythmic variations upon the inferred beat structure. Each extraction window is segmented into four equal-length evaluation sections. The upbeat is identified by the left, outermost edge, while the downbeat is located on the boundary shared between the second and third sections. Onsets located in either of the first two sections are correlated with the upbeat, while onsets occupying either of the latter sections are downbeat associated. The timing offset is seen as the Euclidean distance from the up(down) beat location to the onset. A cursory illustration can be seen in *Fig.1*, while a comprehensive description of the implementation can be found in [30].

3.5.3 Quantization / Pattern Analysis

In this module, the extracted features are discretized into symbolic strings and segmented further to facilitate motif discovery. Additionally, a dimensionality compression algorithm converts our symbol strings into the 1-D sequences required by our bioinformatics alignment tools.

- To produce the quantization codebook, we take our continuous feature sequence *s(n)*, sort the values in ascending order *s'(n)*, and apportion them into *Q* equal sets. The max/min values of each set dictate the thresholds for each quantization division.[1] The discretized feature sequence is equal in length to the number of extraction windows obtained in the audio segmentation module.
- A sliding mask *M* is applied to the feature sequence, creating multiple sub-sequences, expediting pattern analysis.[2] The sliding value is 1 data point and the overlap value of each subsequence is 3 data points.
- To convert our 3-D feature value strings into 1-D symbols while maintaining chronological evolution, the following transform (where *d* is the feature dimension and S_q is the quantized value of said data point) is used:

$$S_{qi} = (d-1)Q + S_q \qquad (11)$$

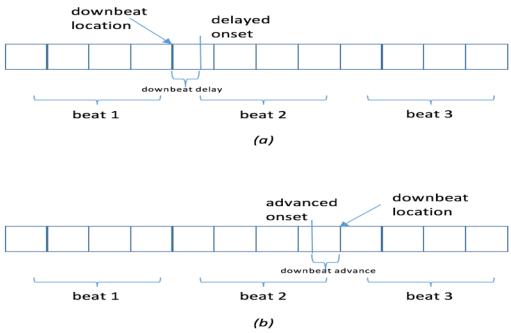

Fig.1 *Measurement of timing offsets. The delayed onset in (a) is given a positive value while the advanced onset in (b) is given a negative value [30].*

- At this point, we have, on average, ~10K, 1-D, sub-sequences. Here, we use the sequence alignment tools of Hirate and Yamana [16]. Succinctly, each sub-sequence (of length=*M*) is compared to every other sub-sequence, to verify if and when the pattern recurs. If the motif recurs more than a minimum threshold, this motif is accepted into the motif bank. The motif bank is the end model of the song and corre-

[1] In our implementation, we set *Q* to 3 as a compromise between computational efficiency and satisfactory data representation.
[2] Our sliding mask window *M* is set to 4 extraction windows.

spondence between motif banks could be a signal of similarity. The algorithm is highly customizable, allowing for various support values and time span intervals.

4. EVALUATION

4.1 "Ground-Truth"

Musical similarity is recognized as a subjective measure, however there is consistent evidence signifying a semi-cohesive similarity experience pervading diversified human listening groups [6] [19] [26]. This is implicative of there being validity in utilizing human listening as an evaluation of similarity. Nonetheless, objective statistics are revealing and must also be incorporated into the system appraisals.

To provide an equitable qualitative assessment of the systems' performance, we adopt a multifaceted scoring scheme comprising three branches:

I. Human Listening (BROAD score)[3]
II. Genre Similarity (Mean % of Genre matches)[4]
III. F-measure over top 3 candidates[3]

4.2 Database and Design

The musical repository used in this research consists of 60 songs spanning the following genres; Rock, Singer/Songwriter, Pop, Rap/Hip-Hop, Country, Classical, Alternative, Electronic/Dance, R&b/Soul, Latino, Jazz, New Age, Reggae, and the Blues. A geometrically embedded visualization of a portion of the artist space according to their Erdös distances[5] can be seen in *Fig.2*. The pool of 20 participants engaging in the listening experiments spans multiple contrasting musical preferences, age groups, and backgrounds.

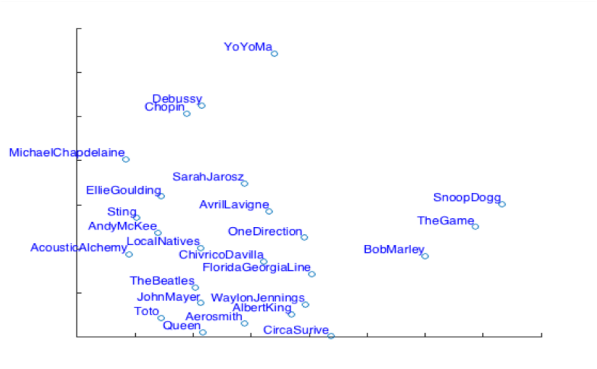

Fig.2 *Erdös distance, a function of transitive similarity, evaluates the similitude of two performers (A&B) as the number of interposing performers required to create a connection from A to B [8]. Above orientation derived via Multidimensional Scaling, optimized by gradient descent..*[6]

[3]Adopted from MIREX. See [29] for a detailed explanation.
[4]Genres are allocated according to iTunes® artist descriptions.
[5]For a complete description of the Erdös measure, see [8].

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 moderately 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.

Table 1. *BROAD scale used by listening participants.*

The database is analyzed using each of the 4 systems and distance matrices are computed correspondingly. Each song from the set is used, in turn, as a seed query. Following the query (each listener hears 3 seed queries in total), the participants hear the top candidate from each system (i.e. each listener hears a total of fifteen, 30-second song snippets). The seed queries presented to the participants were randomized, as was the order in which the top system candidates were played. In the instance that multiple systems returned the same top candidate, the second candidates were used instead.

For homogeneity, the chorus, 'hook', or section containing the main motive of the song was used as playback to the participants. Each participant was asked to rate the candidate return from each system, for each query according to Table 1.

4.3 Results

In regards to our BROAD scores, the performance of each algorithm is established as the mean value rating computed over every top candidate return from each system. Performance according to this metric is displayed in *Fig.3*.

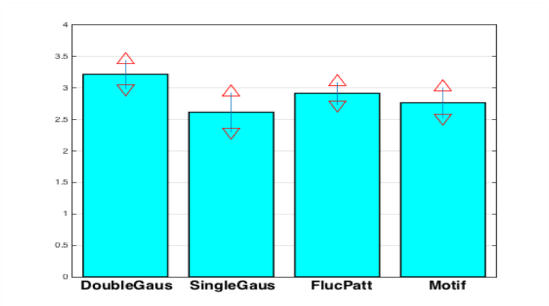

Fig.3 *Average assessment for each system, computed across all listening participants. The red arrows indicate a 95% confidence interval, calculated utilizing the compensatory formula for sample size less than 30 (T-score).*

System performance according to genre similarity is computed as the ratio of seed genre to candidate genre matches to the top candidate returned. This metric is averaged for all seed queries over the entire database. Performance according to this metric is displayed in *Fig.4*.

F-measure, the harmonic mean of two classic information retrieval metrics (precision and recall), communi-

cates information regarding the accuracy and propriety of returned responses to a given query. To qualify as a relevant return item, a candidate must satisfy at least one of the following conditions:

i. Be of the same genre as the seed.
ii. Be of the same artist as the seed.
iii. The seed and query share at least one similar artist.[6]

Our F-measure metric is viewed as the average score computed over all possible seed queries. Performance according to this metric is displayed in *Fig.5*.

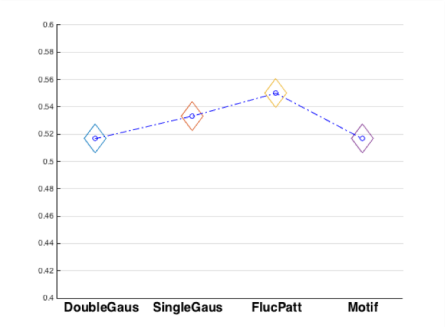

Fig.4 *Average system performance according to genre matches between seed query and 1st candidate returns.*

5. CONCLUSIONS

5.1 Discussion

We have presented four different approaches towards establishing a musical similarity estimate and compared each approach using three evaluation schemes. While the data does carry some implications, we must keep in mind that no recommendation system can perpetually placate the sentiment of every listener. Anticipating an individual's musical penchant is a highly variable undertaking, regulated by a multitude of psychological, psychoacoustic, cultural and social components.

However, what can be unambiguously interpreted from the data is the fact that the inclusion of dynamic, temporal information increases system performance. This is what we hoped to find. Our lives unfold in time; as music is a reflection of the life experience, the pervading temporal aspect of its cognition is intuitively observed and understood.

A further curious implication arising from the data can be seen with regards to the idea of a 'genre'. Results from the F-measure and human listening are essentially congenial, however, this is not mirrored in the genre similarity metric. This might suggest the 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.' This trend can be witnessed at present with companies like Spotify®, Last.fm®, and Allmusic.com®.

5.2 Future Trajectory

At its inception, the sequential motif system was designed with the aim of identifying, quantifying, and ultimately extracting expressive, humanistic lineaments from performed music. Upon successful identification and extraction, a myriad of potentialities arises. One of the more interesting pursuits being the superimposition of said extracted features onto a generic MIDI composition with the intention of "bringing it to life." Accurately extracting the subtle, expressive nuances intrinsic to a performance and mapping them to tractable MIDI parameters could reveal a deeper comprehension of human audition.

We have yet to reach this end, but as an unexpected waypoint en-route to our destination, we found that our system might be able to offer an additional interpretation as to what musical similarity means. Our research into perceptually salient feature identification, extraction, and quantization is currently advancing.

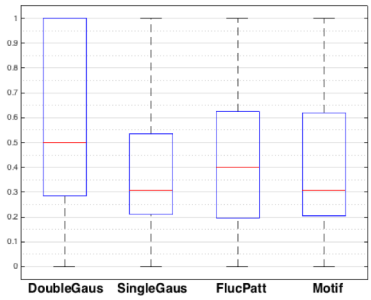

Fig.5 *F-measures of each system. On each boxplot, the red line represents the median, the ends of each box denote the 25th and 75th percentiles, and the whiskers extend to the most extreme data points.*

6. REFERENCES

[1] Aucouturier, Jean-Julien and Francois Pachet: "Improving Timbre Similarity: How High's the Sky?," *Negative Results Speech Audio Sci.*, vol. 1, 2004.

[2] Aucouturier, Jean-Julien and Francois Pachet: "Music Similarity Measures: What's The Use?," In *ISMIR*, 2002.

[3] Aucouturier, Jean-Julien, Boris Defreville and Francois Pachet: "The Bag-of-frames Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes But Not For Polyphonic Music," *Acoust. Soc. Am.*, Vol. 122, No. 2, pp. 881–91, 2007.

[4] Aucouturier, Jean-Julien and Francois Pachet: "Finding Songs That Sound the Same,", 2000.

[5] Burt, Jennifer L., Debbie S. Bartolome, Daniel W. Burdette and J. Raymond Comstock Jr.: "A Psychophysiological Evaluation of the Perceived Urgency of Auditory Warning Signals," *Ergonomics,* Vol. 38, pp. 2327–40, 1995.

[6] Artist similarity data used in our measure was extracted from Last.fm®.

[6] Berenzweig, Adam, Beth Logan, Daniel P.W. Ellis, and Brian Whitman: “A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures,” 2003.
[7] De Leon, Franz, and Kirk Martinez: “Enhancing Timbre Model Using MFCC and its Time Derivatives for Music Similarity Estimation,” In 20th European Signal Processing Conference.
[8] Ellis, Daniel P.W., Brian Whitman, Adam Berenzweig, and Steve Lawrence: "The Quest for Ground Truth in Musical Artist Similarity," 2002. (Whitman)
[9] Ellis, Daniel P.: “Beat tracking by dynamic programming,” *Journal of New Music Research*, Vol. 36, No. 1, pp. 51–60, 2007.
[10] Foote, J. T.: “Content-based Retrieval of Music and Audio,” In *SPIE*, pp. 138–147. 1997
[11] Foss, John A., James R. Ison, and James P. Torre: “The Acoustic Startle Response and Disruption of Aiming: I. Effect of Stimulus Repetition, Intensity, and Intensity Changes,” *Human Factors,* pp. 31:307–18, 1989.
[12] Gjerdingen, Robert O. and David Perrott: “Scanning the Dial: The Rapid Recognition of Music Genres,” *Journal of New Music Research*, Vol. 37, No. 2, pp. 93-100, 2008.
[13] Grey, J.M.: “Multidimensional Perceptual Scaling of Musical Timbres,” *Journal of the Acoustical Society of America*, Vol.61, No.5, pp. 1270-1277, 1977.
[14] Halpern, D. Lynn, Randolph Blake, and James Hillenbrand: “Psychoacoustics of a Chilling Sound,” *Perception and Psychophysics,* Vol. 39, pp. 77–80, 1986.
[15] Hevner, Kate: “Experimental Studies of the Elements of Expression in Music,” *The American Journal of Psychology*, Vol. 48, No. 2, pp. 246-268, 1936.
[16] Hirate, Yu, Hayato Yamana: “Generalized Sequential Pattern Mining with Item Intervals,” *Journal of Computers,* Vol. 1, No. 3, 2006.
[17] Iverson, Paul, and Carol L. Krumhansl: “Isolating the dynamic attributes of musical timbre,” *Journal of the Acoustical Society of America*, Vol. 94, pp. 2595-2603, 1993.
[18] Li, Tao, Mitsunori Ogihara, and George Tzanetakis. *Music Data Mining*. Boca Raton: CRC, 2012. Print.
[19] Logan, Beth, and A. Salomon: “A Music Similarity Function Based on Signal Analysis,” *Multimedia and Expo, 2001*. ICME 2001. IEEE International Conference on, pages 745 748, 2001.
[20] Mandel, Michael and Dan Ellis: “Song-level Features and Support Vector Machines for Music Classification,” *In ISMIR*, pp. 594-599, 2005.
[21] Margulis, Elizabeth Hellmuth: “Repetition and Emotive Communication in Music Versus Speech,” *Frontiers in Psychology* 4, 2013.
[22] McAulay, Robert J., and Thomas F. Quatieri: “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” IEEE ICASSP, vol. 34, no. 4, pp. 744–754, 1986.
[23] Moreno, Pedro J., Purdy P. Ho, and Nuno Vasconcelos: “A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications,” 2004.
[24] Oppenheim, Alan V.: “A speech analysis-synthesis system based on homomorphic filtering,” *Journal of the Acoustical Society of America*, Vol. 45, pp. 458–465, 1969.
[25] Pampalk, Elias, Arthur Flexer, and Gerhard Widmer: "Improvements of Audio Based Music Similarity and Genre Classification.”
[26] Pampalk, Elias: *Computational Models of Music Similarity and their Application in Music Information Retrieval.* Diss. Vienna University of Technology, Austria, March 2006.
[27] Penny, W.D.: “Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart Densities,” 2001.
[28] Pereira, Carlos Silva, Joao Teixeira, Patricia Figueiredo, Joao Xavier, and Sao Luis Castro: “Music and Emotions in the Brain: Familiarity Matters,” 2011.
[29] Raś, Zbigniew, and Alicja A. Wieczorkowska: "The Music Information Retrieval Evaluation EXchange: Some Observations and Insights." *Advances in Music Information Retrieval*. Berlin: Springer Verlag, 2010. 93-114. Print.
[30] Ren, Gang, and Mark Bocko. *Computational Modelling of Musical Performance Expression: Feature Extraction, Pattern Analysis, and Applications*. Diss. U of Rochester, 2015.
[31] Rubner, Yossi, Carlo Tomasi, and Leonidas Guibas: “The Earth Mover's Distance as a Metric for Image Retrieval,” *International Journal of Computer Vision*, Vol. 40, pp. 99-121, 2000.
[32] Temperley, David: *The Cognition of Basic Musical Structures*, Cambridge, MA: MIT Press, 2001.
[33] Terasawa, Hiroko, Malcolm Slaney, and Jonathon Berger: “Perceptual distance in timbre space” *Proceedings of International Conference on Auditory Display,* pp. 61 - 68. Limerick: International Community for Auditory Display.
[34] Wolpoff, Mieford H., Fred H. Smith, Geoffrey Pope; David Frayer: “Modern Human Origins,” *Science*, Vol. 241, No. 4867, pp. 772-774, 1988.