Descriptors for Perception of Quality in Jazz Piano Improvisation

Jeff Gregorio
Drexel University
3401 Market Street
Philadelphia, PA
jgregorio@drexel.edu

David Rosen
Drexel University
3401 Market Street
Philadelphia, PA
drosen@drexel.edu

Michael Caro
Drexel University
3401 Market Street
Philadelphia, PA
mcaro@drexel.edu

Youngmoo E. Kim
Drexel University
3401 Market Street
Philadelphia, PA
ykim@drexel.edu

Keywords

performance analysis, computational musicology and music analysis, jazz improvisation, expertise

ABSTRACT

Quality assessment of jazz improvisation is a multi-faceted, high-level cognitive task routinely performed by educators in university jazz programs and other discriminating music listeners. In this pilot study, we present a novel dataset of 88 MIDI jazz piano improvisations with ratings of creativity, technical proficiency, and aesthetic appeal provided by four jazz experts, and we detail the design of a feature set that can represent some of the rhythmic, melodic, harmonic, and other expressive attributes humans recognize as salient in assessment of performance quality. Inherent subjectivity in these assessments is inevitable, yet the recognition of performance attributes by which humans perceive quality has wide applicability to related tasks in the music information retrieval (MIR) community and jazz pedagogy. Preliminary results indicate that several musiciologically-informed features of relatively low computational complexity perform reasonably well in predicting performance quality labels via ordinary least squares regression.

1. INTRODUCTION

There is a vast amount of literature describing the underlying theory, techniques, and strategies which are paramount to achieving mastery of jazz improvisation. Broadly, the foundation of expert-level jazz improvisation relies on one's ability to deeply understand and creatively manipulate three critical aspects of a performance: rhythm, harmony, and melody[17]. The potential variations in any of these areas is virtually infinite; however, jazz improvisation is a distinct type that includes the ability to generate the unforeseen, within the pre-existing structure of a song's chord structure, carefully balancing tradition and innovation[4]. Thus, while originality and creativity are essential to achieving high-quality improvisations, jazz's rich history, genre constraints, and past performances from eminent musicians must also be considered.

In this pilot study, using a novel approach, we examine the relationship between specific rhythmic, harmonic, and melodic music features and strategies jazz musicians implement during improvisation with their perceived quality. Thus, this research proposes a systematic framework that can reliably identify discrete music features from piano improvisations via MIDI data extraction and assess their contributions in predicting the perceived quality of a given performance. Although the music features we explore do not encompass every facet of improvisation, this is the first study that has leveraged musicologically informed features, quantitatively evaluating those which have the greatest impact on listener perception with respect to jazz improvisation. This work will greatly benefit musicians and music educators, alike, furthering knowledge and gaining insights into how well-informed jazz listeners are impacted by specific note choices, articulations, variations in rhythm, dynamics, and harmonic interpretations of their playing. Furthermore, the application of this work can extend beyond jazz improvisation; however, specific feature design will vary depending on the genre/style of interest.

Acquiring precise and detailed music information of single instruments within a jazz improvisation is extremely difficult from audio alone, especially without access to multitrack recordings. Thus, we collected an original symbolic (MIDI) dataset comprised of 88 piano improvisations from trained jazz pianists. In this work, we implement features that capture several fundamental strategies and techniques of jazz improvisation, which are described later in this paper and display the ability to parse low, medium, and high-level features, identifying which aspects of performance have the greatest influence on the perceptual evaluations of expert and critical jazz listeners.

2. JAZZ IMPROVISATION DATASET

2.1 Pianists & Judges

Participants ($N = 22$) were jazz pianists recruited from local university jazz programs, seminaries, and professional organizations In Philadelphia, PA. Musicians were 19-34 years of age ($M = 24.77$, $sd = 4.39$), and participants were predominantly male (2 females). Expertise information was collected for years of music training ($M = 17.55$, $sd = 5.25$) and years of jazz training ($M = 8.09$, $sd = 4.43$). To evaluate and acquire ratings for the improvisational performances, four jazz experts were recruited to serve as judges.

NIME'15, May 31-June 3, 2015, Louisiana State Univ., Baton Rouge, LA.

These judges included a director of a collegiate jazz program, gigging professionals, and instructors, all with over 25 years of performance experience.

2.2 Design and Procedures

Trials were conducted at Drexel University in Philadelphia, PA. All performances took place in a professional sound booth, containing a 88-key semi-weighted MIDI controller keyboard, sustain pedal, music stand, and headphones. Apple's Logic Pro v.9.1.8 music software was used to record improvisations, collect MIDI performance data, and provide musicians with a bass and drums audio accompaniment to a novel 16-bar chord sequence.

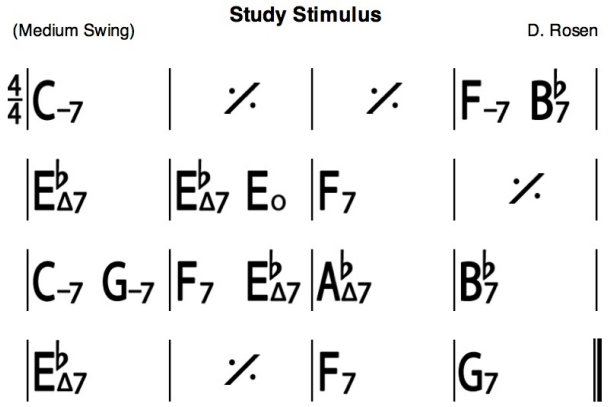

Figure 1: All jazz pianists improvised to this 16-bar chord sequence with upright bass and drums accompaniment. Each improvisation consisted of four cycles (64-bars) through the chords.

All participants completed four takes, each take was 4 times through the chord sequence (64 bars, $\approx$ 2 minutes). Upon completion of the experiment, musicians were asked to identify music features and strategies which guided their performance. Similar questions were also asked to judges after all improvisations ratings had been submitted. This data was collected as further justification for the features we analyzed. Responses were grouped in broad categories which included: harmonic alterations, rhythmic variation, increased dissonance, altered dynamics, melodic repetition, and use of "color" tones.

Performances were pseudo-randomized for judging, with the constraints that the same musician could not be heard consecutively or have more than 2 improvisations within a judging block. Each judging block consisted of 22 improvisations. While the order of the performances within each block were identical for all judges, the order of presentation of the blocks was different for each judge. Using the Consensual Assessment Technique (C.A.T.) [1], judges rated improvisations on a 7-point Likert scale for creativity, technical proficiency, and aesthetic appeal. These holistic ratings represent an evaluation of performative characteristics that contribute to the quality of an improvisation[3]. For each improvisation, ratings were averaged across categories and judges in order to arrive at a single quality of improvisation rating ($M = 4.69$, $sd = .80$). The jazz improvisation dataset ratings distribution is shown in Figure 1.

The C.A.T. has experts in a domain rate creative products relative to one another. This evaluation technique is not dependent on any particular theory of creativity and uses the same method for assessing creative production as most domains in the real world [2]. Although ratings are sample specific and not generalizable to all jazz improvisation, this pilot dataset can be used as a proof of concept for quantitatively examining the underlying music features, deviations, and choices that characterize expert-level performance. Furthermore, assessing musical creativity and improvisation via the C.A.T has precedent in past work, displaying high inter-rater reliability [9, 3, 13, 7].

2.3 Inter-rater Reliability and Correlations

The intraclass correlation coefficient model ICC(2,1) was employed to measure inter-rater reliability [22] for judges' ratings on creativity, technical proficiency, and aesthetic appeal. Reliability is calculated from single measurements rather than an average [16]. Values were computed for consistency where systematic differences between raters are considered irrelevant. ICC(2,1) and item (scale) correlations were calculated for all four judges in SPSS v.22.0.0.

Judges' reliability were calculated for creativity ($ICC = .71$), technical proficiency ($ICC = .81$), and aesthetic appeal ($ICC = .73$). Conventionally, an ICC $> .75$ is excellent, .40-.74 is adequate to good, and < 0.40 is poor [12]. Thus, all scales had excellent to very good reliability. The three scales had highly significant positive correlations across judges for creativity and aesthetic appeal ($r(86) = .84$, $p < .01$), creativity and technical proficiency ($r(86) = .89$,$p < .01$), and aesthetic appeal and technical proficiency ($r(86) = .87$, $p < .01$). Judges may have struggled differentiating between constructs due to the inherent collinearity of these characteristics in the domain of jazz improvisation, a highly technical, creative, musical performance.

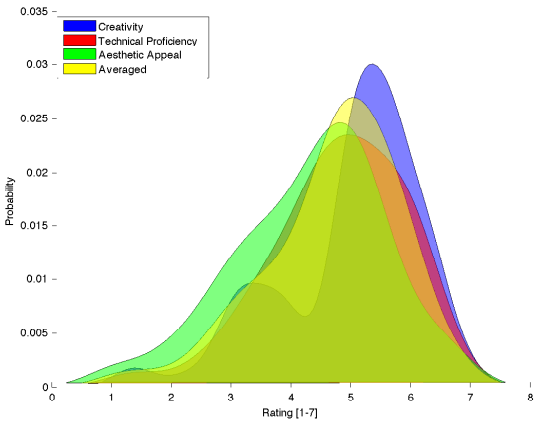

Figure 2: This is a continuous distribution of the probability for each rating scale: creativity (blue), technical proficiency (red), and aesthetic appeal (green). Averaging these three aspects of performance, we arrived at our quality score (yellow).

3. FEATURES

Feature design was guided primarily by common approaches to jazz improvisation in pedagogical literature [20] and cognitive studies [14] as well as questionnaires completed by the four judges indicating criteria by which they rated performances creative, aesthetically appealing, and technically proficient.

Many qualities noted by the judges were high-level descriptors that rely heavily on prior knowledge of the vast body of work in the jazz tradition. Assesment of each target rating is highly relative, thus inseparable from a assessment

of novelty that requires knowledge of historical context. For example, one judge noted that references, or musical quotes, of classic melodies fitting over sections of the provided chord changes and stylistic 'nods' to famous jazz musicians lead him to give higher ratings.

Obtaining a succinct feature set to adequately capture performance elements presents a challenge in that the requirement of relatively low dimensionality feature space often necessitates partially or completely disregarding temporal evolution of features. The main focus of this pilot study is the development of time-series features computed on a sliding window over the duration of the performance. Descriptors ultimately used for prediction of performance quality are statistics of each time series and its first difference, which may include min, max, range, median, mean, standard deviation, skewness, and kurtosis.

Feature computation is performed using MIDI recordings of each performance. The set of features used builds upon the MIDI Toolbox [11] for the MATLAB scientific computing language. The toolbox is used primarily for MIDI file I/O, visualization to guide feature design, and low-level distribution features from which we derive mid-level features of harmony and rhythm. The remainder of Section 3 focuses on the design of the proposed features.

We note that the use of the MIDI modality does offer several advantages over similar approaches using audio recordings. In a MIDI file, onset times, note velocity, and pitches are known precisely, thus we do not require estimates of these quantities. However, the proposed features are not exclusive to MIDI data, but can be adapted to any symbolic representation derived from estimated beat locations, onset times, loudness, and note pitches, with the additional constraint of prior knowledge of a piece's chord changes for certain features.

3.1 Preprocessing

Prior to feature computation, the set of note events is segmented into chords and an approximation of the monophonic melody. Each feature is computed threefold, that is, on the entire set of MIDI events as well as the approximated melody, and grouped chords. We note that entire set of MIDI events contains the melody and chord subsets, and that it is possible for a note to be in both a chord and the melody. We therefore expect some natural correlations in the feature space which we later eliminate via dimensionality reduction.

3.1.1 Chord Grouping

Chords are defined by onset synchrony, following a two-step grouping process detailed in [6]. In the first step, any group of notes can be considered part of a chord if their onsets fall within 50ms. In the second step, any group of chords can be merged if any two note onsets in any chord fall within 100ms.

3.1.2 Melody Extraction

A monophonic melody approximation is segmented from the polyphonic data using a naive, yet reasonably accurate [23] and computationally cheap skyline algorithm, whereby the melody note at any time is considered to be the note with the highest pitch, and any overlapping notes are truncated.

Though more recent approaches to main melody extraction from polyphonic audio and symbolic data have included exhaustive n-gram based approches geared toward query-by-humming retrieval systems [10], selection of candidate melodies based on statistical features of pitch contours [21] and melodic similarity measures from multiple versions of the same song [15], none was appropriate or possible for our task. This was due to the insuitability of multiple melody candidates, lack of MIDI examples with human-labeled melodies, and the unique nature of improvised performances.

3.2 Low-Level Expression

Musical expression is often defined as independent of the notes being played, rather how the notes are played to create timbral nuance that communicates some intention [5] [8]. We define low-level expression features as those of generally low computational complexity that are independent of melody, harmony, and rhythm. These include simple statistical descriptors that may carry information regarding expressive intention including time-series features *note density*, *mean note velocity*, *mean note duration*.

We also compute features of the notes constituting each chord including *note count*, *onset asynchrony*, *velocity asynchrony*, and *duration asynchrony*. The asynchrony features are computed as the standard deviation (respectively) of note onset times, velocities, and durations within each chord. The aformentioned statistical descriptors (min, max, range, mean, etc.) are computed on each feature over the entire chord set. If advantageous, these chord set features can be easily adapted to a time series by placing each feature value on the approximate location of the chord's onset.

3.3 Rhythmic Style

Multiple judges identified rhythm's primary importance, mentioning the importance of jazz's "swing" feel in assessing quality. Toward identifying perceptually relevant rhythmic patterns, we use an approach based on the inter-onset-interval (IOI) histogram called the Rhythmic Style Histogram Feature (RSHF) developed in [19], adapted to the MIDI modality.

The IOI is defined as the time between the onset of a note and the onset of the next note. IOI has been identified as more perceptually salient in identification of rhythmic patterns than note durations alone [18]. We use a form of the IOI histogram where the salience of each interval is weighted both by the velocity of the note, and again weighted by Parncutt's perceptually-derived durational accent model [18], which accounts for human perception of beat salience in relation to the IOI that follows the event, assigning more importance to events followed by long breaks before the next onset.

The IOI histogram can be computed with a specified number of bins per beat (quarter note). A division of 6 bins per beat is sufficient to resolve quarter note triplets (2/3 beat), eighth note triplets (1/3 beat), and dotted sixteenth notes (3/8 beat) with a maximum IOI of 4 beats. We then compute each rhythmic style feature by computing the relative contribution of each bin to the total energy for patterns:

- *Duple*: Whole, half, quarter, eighth, and sixteenth notes.
- *Triple*: Whole triplet, half triplet, quarter triplet, eighth triplet
- *Long Duration Swing*: Dotted half, quarter
- *Med Duration Swing*: Dotted quarter, eighth
- *Short Duration Swing*: Dotted eighth, sixteenth

A sample stimulus shown in figure 3, and the resulting IOI histogram and *Duple*, *Triple*, and *Short Duration Swing* time-series features are shown in figure 4. Note this IOI histogram is computed with 12 bins per beat with a maximum IOI of 2 beats.

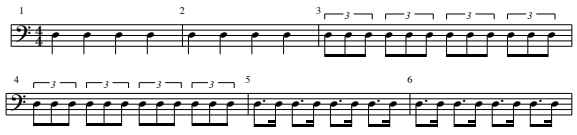

Figure 3: Sample input patterns.

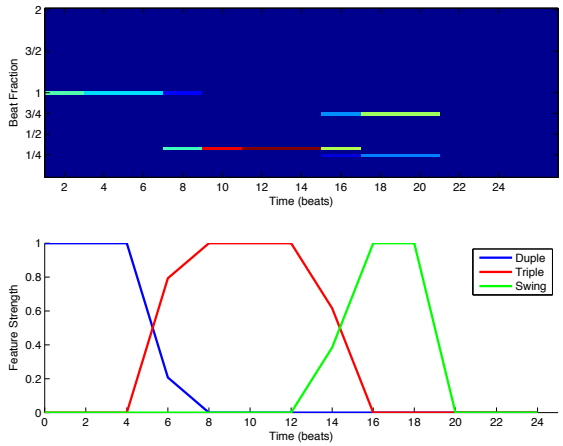

Figure 4: Top: IOI Histogram bins (y-axis) corresponding to specific beat fractions over time. Bottom: Resulting time-series features duple, triple, and swing (short duration)

3.4 Harmony and Melody

We use various descriptors dealing with harmnoic and melodic content, each derived from the pitch class distribution, or chroma, of the performer's note choices. This distribution groups the 88 possible pitches into 12 pitch classes representing the 12 semitones in an octave, with all octaves of C contributing to the first bin, all octaves of C# contributing to the second bin, and so on. The pitch class distribution is computed on a sliding window over the duration of the performance.

3.4.1 Harmonic Function Classes

We use contextual information based on knowledge of the stimulus chord changes to group pitch classes by harmonic function relative to the current chord.

We use nine broad classes of harmonic function, as follows:

- *Key Tones*: pitch classes in the piece's key
- *Chord Tones*: pitch classes in the current chord (i.e. the 1st, 3rd, 5th, and 7th scale degrees of the chord)
- *Root Tones*: 1st and 5th scale degrees of the current chord
- *Guide Tones*: 3rd and 7th scale degrees of the current chord, defining it's tonality
- *Diatonic*: pitch classes in the diatonic scale of the current chord
- *Pentatonic*: pitch classes in the pentatonic scale of the current chord
- *Avoid Tones*: 3rds and 7ths that conflict with the tonality of the current chord (e.g. flat 3rd on a major 7th chord)
- *Color Tones*: Tones in the key of the piece, excluding current chord tones and avoid tones (generally 9ths, 11ths, and 13ths)
- *Dissonant Tones*: flat 2nds (semitone from the root) and sharp 4ths

The minimum chord length used in the stimulus is two beats, so each of the features is computed on a two-beat window. For any given two beat window, we compute the pitch class distribution via the MIDI toolbox, which weights each pitch class by the duration of each occurrence modified by Parncutt's durational accent model. The seven harmonic function features are computed by summing specific bins of the pitch class distribution and dividing by the sum of all bins, yielding a value on the interval $[0, 1]$ every two beats. Like all time-series features proposed, the final features used for regression are statistical descriptors of the time series.

Table 3.4.1 illustrates two harmonic function classes, *chord tones* and *color tones*, as applied to each unique chord in the study stimulus. Currently we can assign pitch class distribution bins to harmonic function classes for an arbitrary sequence of chords including major 7th, minor 7th, dominant 7th, and diminished 7th chords. Future work will extend the analysis of harmonic function to a wider class of chords.

We note a limitation in the pitch class distribution, in that octaves are ignored. These features therefore make no distinction between 9ths and 2nds, 11ths and 4ths, 13ths and 6ths, etc.

3.4.2 Common Tone Voice Leading

The compositional and improvisational strategy of voice leading is generally defined as smooth motion between inner voices (notes) of a chord or melody through chord transitions [20]. One element of this strategy is known as *common tone* voice leading, whereby a composer or improviser will identify tones that are harmonically-related on either side of the transition. We make use again of harmonic function classes to look at the use of common-tone voice within each of the harmonic function classes.

The series of chords is parsed and the beat locations of chord transitions are identified. We then extract a two-beat window around each transition and get the relative contribution of the pitch class bins common to the same harmonic function class in both chords.

For example, at beat 12, we transition from *Cmin7* to *Fmin7*. The intersection of pitch classes in the harmonic function class *Guide Tones* for each chord is *Eb*, so the value of the *common guide tone voice leading* feature at beat 12 would be the *Eb* pitch class bin, divided by the sum of all bins. A value on the interval $[0, 1]$ is computed for each chord transition over the piece, indicating the salience of each voice leading strategy. Like the chord set features, the

	C	C#	D	Eb	E	F	F#	G	Ab	A	Bb	B
C-7	**1**	b2	2*	**b3**	3	4*	#4	**5**	b6*	6	**b7**	7
F-7	**5**	b6	6*	**b7**	7	**1**	b2	2*	**b3**	3	4*	#4
Bb7	2*	b3	**3**	4*	#4	**5**	b6	6*	**b7**	7	**1**	b2
EbΔ7	6*	b7	**7**	**1**	b2*	2	b3	**3**	4*	#4	**5**	b6
Eo	b6*	**bb7**	b7	7	**1**	b2	2*	**b3**	3	4	**b5**	5
F7	**5**	b6	6*	**b7**	7	**1**	b2	2*	b3	**3**	4*	#4
G-7	4*	#4	**5**	b6*	6	**b7**	7	**1**	b2*	2	**b3**	3
AbΔ7	**3**	4	#4*	**5**	b6	6*	b7	**7**	**1**	b2	2*	b3
G7	4*	#4	**5**	b6*	6	**b7**	7	**1**	b2*	2	b3	**3**

Table 1: Interval table of each pitch class in relation to each unique chord used in the study stimulus. Chord tones are bolded and color tones marked with an asterisk.

voice leading features are ordered, but are not a strict time series as they only have values at chord transitions, which are not uniformly spaced.

4. EXPERIMENT

In all, the full feature set includes just under 600 statistical descriptors of time series features, their first differences, and set features. Such a high feature space dimensionality with many expected correlations necessitates dimensionality reduction prior to evaluating the predictive power of the features. Prior to dimensionality reduction, the feature space is normalized to zero mean and unit variance via $Z = \frac{X - E[X]}{\sigma(X)}$ where Z is the new standardized feature set and X is the original feature set.

4.1 Dimensionality Reduction

We perform dimensionality reduction in a two-step process. We first compute Pearson's correlation coefficient and associated p-value between each feature and the target ratings. We drastically reduce the dimensionality of the feature space by omitting statistically insignificant features (p > 0.05). The feature space after this reduction step consists of 227 descriptors.

Of these 227, 49% are features of harmonic function class (with voice leading considered separately). The remaining half is made up of low-level expression features (14%), rhythmic style features (18%), and voice leading features (17%). An alternate breakdown shows the 227 being comprised of time series features (59%), first difference features (30%), and set features (11%).

Feature	ρ	p
ts rclass swing short chd skew	-0.571	6.364e-09
ts rclass swing short mean	0.564	1.027e-08
voicelead guide skew	-0.564	1.089e-08
ts rclass swing short chd kurtosis	-0.553	2.379e-08
ts rclass swing short chd mean	0.548	3.284e-08
voicelead pentatonic std	-0.546	3.676e-08
ts hclass chord std	-0.543	4.611e-08
dts voicelead guide kurtosis	-0.542	4.777e-08
ts hclass diatonic chd mean	0.536	7.164e-08
ts note density mean	0.535	8.146e-08

Table 2: Top 10 features ranked by Pearson's correlation p-value. Legend: *ts*: time series, *dts*: first difference of time series, *rclass*: rhythmic style feature, *hclass*: harmonic function class, *chd*: chords only

Table 2 shows Pearson's correlation coefficient ρ and the associated p-value for the top ten features. A positive ρ indicates that the feature is associated with high quality ratings, and a negative ρ is associated with lower quality ratings.

The ten features' correlation with high quality can be interpreted as such:

- [*Features* $1-2$]: strong use of short duration (dotted eighth $\rightarrow$ sixteenth note) swing
- [*Feature* 3]: strong use of common guide tone voice leading
- [*Features* $4-5$]: wide variation in the use of short duration swing in chord onsets
- [*Feature* 6]: consistent use of common pentatonic tone voice leading
- [*Feature* 7]: consistent use of chord tones
- [*Feature* 8]: varying rate of change in common guide tone voice leading
- [*Feature* 9]: strong use of diatonic tones in chords
- [*Feature* 10]: high mean note density

4.1.1 Principal Components Analysis

After retaining only those features significant at the $p > 0.05$ level, we perform further dimensionality reduction via principal components analysis (PCA). PCA allows us to project the feature data into a lower-dimensionality subspace such that the variance of the projected data is maximized. This allows a more compact representation that retains most of the information in the original feature space. PCA also has the property that the new basis dimensions are orthogonal, thus we can eliminate redundancy in highly-correlated features.

We use an exhaustive cross validation approach to determine the optimal number of basis dimensions to use. Through leave-one-out cross validation (LOOCV), we compute the PCA basis on the training data minus one example, then project the training data and the test example into the new feature space, and evaluate the error via Ordinary Least Squares (OLS) regression for each possible projection dimensionality. We use the projection dimensionality $M = 8$, which yielded the lowest average error over each cross validation trial and retained 63% of the variance in the original feature space.

By examining the weights of the PCA basis vectors, we can determine the relative contribution of each feature or feature group to each basis dimension, in an attempt to extract some meaning from the new, reduced feature space. However, none of the first eight basis vectors showed contributions from the feature groups differing significantly from the group contributions to the 227 most significant features.

4.2 Prediction of Quality Labels

With the optimal PCA projection dimensionality, we then perform one last round of LOOCV, where each omitted test example is scaled to unit variance based on the mean and standard deviation of the training data, and projected into the first eight basis dimensions of the training data. Again, we use OLS regression to predict the averaged quality ratings. We evaluate using mean absolute error (MAE) and root mean squared error (RMSE) which can be seen in Table 3.

Metric	Performance
MAE	0.62
RMSE	0.59

Table 3: Performance statistics from leave-one-out cross validation.

5. CONCLUSIONS

From the MIDI data, we were able to compute a set of features that captured basic salient elements of low-level expression, rhythmic style, harmonic function, and common tone voice leading. Consistently among the most informative features for predicting human quality labels were the various harmonic function classes. These features addressed a subset of qualities noted by judges as salient in their assessment of creativity, technical proficiency, and aesthetic appeal.

Other positive qualities noted by judges included high-level recognition of theme and variation. Since the study's stimulus has only chord changes and no notated melody, this problem would dependent on reliable segmentation of musical phrases, identifying themes, and subsequently detecting variations, with each sub-problem being an area of ongoing research in itself, and each sub-problem having a high potential for propagation of errors. This pilot study therefore focused on features that are more concrete mid-level features grounded in modern jazz theory and related work in musicology and music information retrieval.

Though simple statistics time series and first difference features performed reasonably well, we note the shortcomings in using descriptors that solely capture the shape of the distribution. Future work will focus on identifying periodicities in time series features and identifying key locations in chord changes where a feature would be considered more important to a human judge. One example is often noted in jazz pedagogical literature, which encourages different voicing strategies depending on the cadence, or movement of the root. Return to the tonic is of particular importance for resolution, with pianists often strategically building dissonance leading into a resolution to accentuate the effect.

We also note the limitations of the pitch class distribution, in that it makes no distinction between octaves, therefore cannot disambiguate wide from narrow intervals such as 9ths/2nds, 11ths/4ths, or 13ths/6ths. Future work will explore the efficacy of a multi-octave pitch class distribution.

6. REFERENCES

[1] T. Amabile. The social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43:997–1013, 1982.

[2] J. Baer. *The Cambridge Handbook of Creativity*, chapter Is Creativity Domain Specific?, page 489. Cambridge University Press, New York, NY, 2010.

[3] R. Beaty, B. Smeekens, P. Silvia, D. Hodges, and M. Kane. A first look at the role of domain-general cognitive and creative abilities in jazz improvisation. *Psychomusicology: Music, Mind and Brain*, 23(4):262, 2013.

[4] P. F. Berliner. *Thinking in jazz*. University of Chicago Press, Chicago, Ill., 1994.

[5] M. Bernays and T. Caroline. Expressive production of piano timbre: Touch and playing techniques for timbre control in piano performance. In R. Bresin, editor, *Proceedings of the 10th Sound and Music Computing Conference (SMC 2013)*, pages 341–346. KTH Royal Institute of Technology, Logos Verlag, Berlin, 2013.

[6] M. Bernays and C. Traube. Piano touch analysis: a matlab toolbox for extracting performance descriptors from high-resolution keyboard and pedalling data. In *Proceedings of Journées d'Informatique Musicale (JIM 2012)*, pages 55–64, Mons, Belgium, May 2012.

[7] D. Brinkman. Problem finding, creativity style and the musical compositions of high school students. *The Journal of Creative Behavior*, 33(1):62–68, 1999.

[8] S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin. Modeling and control of expressiveness in music performance. *Proceedings of the IEEE*, 92(4):686–701, Apr 2004.

[9] C. De Dreu, B. Nijstad, M. Baas, I. Wolsink, and M. Roskes. Working memory benefits creative insight, musical improvisation, and original ideation through maintained task-focused attention. *Personality and Social Psychology Bulletin*, 38(5):656–669, 2012.

[10] S. Doraisamy and S. Rüger. Robust polyphonic music retrieval with n-grams. *J. Intell. Inf. Syst.*, 21(1):53–70, July 2003.

[11] T. Eerola and P. Toiviainen. MIR in matlab: The MIDI toolbox. In *ISMIR 2004, 5th International Conference on Music Information Retrieval, Barcelona, Spain, October 10-14, 2004, Proceedings*, 2004.

[12] J. Fleiss. *The Design and Analysis of Clinical Experimnts*. Wiley, New York, NY, 1986.

[13] M. Hickey. An application of amabile's consensual assesment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, 49(3):234–244, 2001.

[14] P. N. Johnson-Laird. How jazz musicians improvise. *Music Perceptoin*, 19(3):415–442, 2002.

[15] L. Li, C. Junwei, W. Lei, and M. Yan. Melody extraction from polyphonic midi files based on melody similarity. In *Information Science and Engineering, 2008. ISISE '08. International Symposium on*, volume 2, pages 232–235, Dec 2008.

[16] K. O. McGraw and S. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30, 1996.

[17] D. Moorman. *An analytic study of jazz improvsation with suggestions for performance*. PhD thesis, New York university, 1984.

[18] R. Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, pages 409–464, 1994.

[19] M. Prockup, J. Scott, and Y. E. Kim. Representing musical patterns via the rhythmic style histogram feature. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 1057–1060, New York, NY, USA, 2014. ACM.

[20] R. Rawlins, N. Bahha, and B. Tagliarino. *Jazzology: The Encyclopedia of Jazz Theory for All Musicians*. Jazz Instruction Series. Hal Leonard, 2005.

[21] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1759–1770, 2012.

[22] P. E. Shrout and J. Fleiss. Intraclass coreelations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420, 1979.

[23] A. L. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *Proceedings of the Sixth ACM International Conference on Multimedia*, MULTIMEDIA '98, pages 235–240, New York, NY, USA, 1998. ACM.