TISMIR

Evaluating an Analysis-by-Synthesis Model for Jazz Improvisation

RESEARCH

KLAUS FRIELER

WOLF-GEORG ZADDACH

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper pursues two goals. First, we present a generative model for (monophonic) jazz improvisation whose main purpose is testing hypotheses on creative processes during jazz improvisation. It uses a hierarchical Markov model based on mid-level units and the Weimar Bebop Alphabet, with statistics taken from the Weimar Jazz Database. A further ingredient is chord-scale theory to select pitches. Second, as there are several issues with Turing-like evaluation processes for generative models of jazz improvisation, we decided to conduct an exploratory online study to gain further insight while testing our algorithm in the context of a variety of human generated solos by eminent masters, jazz students, and non-professionals in various performance renditions. Results show that jazz experts (64.4% accuracy) but not non-experts (41.7% accuracy) are able to distinguish the computer-generated solos amongst a set of real solos, but with a large margin of error. The type of rendition is crucial when assessing artificial jazz solos because expressive and performative aspects (timbre, articulation, micro-timing and band-soloist interaction) seem to be equally if not more important than the syntactical (tone) content. Furthermore, the level of expertise of the solo performer does matter, as solos by non-professional humans were on average rated worse than the algorithmic ones. Accordingly, we found indications that assessments of origin of a solo are partly driven by aesthetic judgments. We propose three possible strategies to install a reliable evaluation process to mitigate some of the inherent problems.

CORRESPONDING AUTHOR:

Klaus Frieler

Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany

klaus.frieler@ae.mpg.de

KEYWORDS:

Generative models; analysis by synthesis; jazz; improvisation; assessment; performance

TO CITE THIS ARTICLE:

Frieler, K., & Zaddach, W.-G. (2022). Evaluating an Analysis-by-Synthesis Model for Jazz Improvisation. *Transactions of the International Society for Music Information Retrieval*, 5(1), 20–34. DOI: *https://doi.org/10.5334/tismir.87*

1. INTRODUCTION

Attempts to generate music with the computer are nearly as old as the computer itself, starting with Lejaren Hiller's ILLIAC Suite in 1959 (Hiller and Isaacson, 1959). The underlying motivations are manifold, ranging from artistic experiments to proofs-of-concept, educational applications such as practising aids,[1] and commercially viable products for royalty-free music.[2] The model evaluated in this paper is based on another motivation, the development of a psychological model of jazz improvisation.

There is vast literature on jazz improvisation research, but the model by Pressing (1984, 1988) developed already in the 1980s is still state-of-the-art even though it remains largely untested. Proving the adequacy of models for jazz improvisation is problematic as the required experimental procedures are difficult. The main reason is that jazz improvisers seem to have limited conscious access to their cognitive and motor processes during improvisation. Furthermore, experimental studies on jazz improvisation must rely on production paradigms which are hard to evaluate and suffer from a sampling problem, because it is seldom possible to have improvisers play a large number of solos in a controlled lab setting. The external validity of these experiments is further often limited, as these experiments very often have to use computer-generated (e. g., with Band-in-a-Box) or prerecorded backing tracks (e. g., Aebersold play-alongs), which offer no possibilities for interaction. Confronting improvisers with their generated solos and prompting self-reflective comments has proved to be one of the most fruitful approaches so far (Norgaard, 2011), but this still has limitations, as jazz improvisers often just do not know in retrospect why they played certain notes or phrases. Another promising and recent approach is corpus-based jazz studies (Owens, 1974; Pfleiderer et al., 2017) that aim at finding phenomenological patterns in jazz improvisations of (mostly eminent) jazz players.

Psychological models of jazz improvisation seldom make sufficiently hard predictions to allow them to be tested in experiments and with corpora. Here, generative models can help, following a general analysis-by-synthesis approach. We decided that our generative model of jazz improvisation should explicitly incorporate high-level knowledge of jazz improvisation while being at the same time probabilistic in order to model inaccessible factors and genuinely probabilistic aspects of an improvisation. One obvious model choice is hierarchical Markov models, which are employed in the current study.

The paper is organized as follows. After some general consideration about evaluation methods in Section 2 for generative models of jazz improvisation and an overview of related work in Section 3, we present in Section 4 our hierarchical Markov model which is based on mid-level analysis, the Weimar Bebop Alphabet, a simple rhythm model, and chord-scale theory. In Section 5, we report on our exploratory experiment to evaluate factors influencing Turing-like solo assessment by a group of experts and non-experts. Finally, we discuss in Section 6 our findings and propose ideas for the evaluation of generative models of jazz improvisation.

2. EVALUATION OF GENERATIVE MODELS

Analysis-by-synthesis approaches are only fruitful if the models can be evaluated with respect to their adequacy. As the models serve as a "laboratory" to test various hypotheses about improvisational processes, there is a certain demand for testing often and reliably, in order to employ a continuous improvement process while ruling for or against certain modeling options.

There are two main approaches to evaluation, which are complementary and not exclusive. First, generated solos can be evaluated with a Turing-like test. Second, they can be judged against a corpus of "accepted" jazz solos, e. g., in terms of melodic features, dramaturgy, or pattern content. In this study, we decided to pursue only the first approach, due to length constraints, and also because we think that the objective approach is principally less severe and powerful, as objective features are not likely to capture the full content of a solo with all the intricate correlations between musical dimensions. Simple first and second statistics for pitch and rhythm, such as the MGEval variant used in Madaghiele et al. (2021), will not suffice for our needs, as these are partly fulfilled already by construction.

In order to evaluate a generative model using Turing-like tests with a panel of judges or raters, one has to solve several problems, which mostly relate to performance aspects. First to note is that a Turing-test needs to be designed as a signal-detection experiment, in which computer and human-generated solos are to be assessed as either human or computer-generated. This necessitates a standardization of the solos in order to minimize confounds. A judgement of computer-generated solos based only on absolute criteria is possible, but it would still need baseline measurements on human-generated solos for a complete picture.

As our (and most other) algorithms generate score-like representations, one could either let expert raters judge the score directly, or the scores could be transferred into sounding music for assessment. The first approach is seldom used as it has the disadvantage that it demands considerable skills from the raters and that imagining music from scores will always be inferior to actual listening to music. Thus, in general, a listening experiment seems preferable.

However, the transfer process introduces additional degrees of freedom in design and, most importantly,

confounds judging the actual musical content and performative aspects, which results in assessment bias.

To produce sounding music from generated scores, one can either use machine or human-generated renditions of solos, either with or without a musical context. As jazz solos without accompaniment rarely make musical sense, using an accompaniment seems mandatory. Letting human players perform the generated solo is an intriguing approach, but a very time and resource consuming method that does not scale well. The most common and most practical solution is to use machine-generated renditions over machine-generated (or prerecorded) backing tracks.

A machine-generated solo rendition can be either deadpan MIDI (a score "as is") or post-processed by humans (or further algorithms). Tweaking can be applied to performance parameters (e. g., microtiming, dynamics) and to the musical surface. The easiest solution of using deadpan MIDI data has the disadvantage that non-expressive performance might be strongly associated with a computerized, non-human performance which might likely result in rating bias as well.

Another issue is the proper selection of generated solos, as creativity is mostly conditioned on selection processes. This applies to both human- and computer-generated solos. A fair evaluation process can only be based on some form of random selection, but there are still free parameters, e. g., the choice of underlying tunes. On the human-generated side, the question is whether solos from masters or professionals or solos from non-professionals or students should be used. The choices here are likely to influence the evaluation process and need careful considerations.

Finally, there is also the question of whether to use expert or non-expert listeners for a human panel. Expert listeners are more likely to identify computer-generated jazz solos simply by having more exposure to jazz and its implicit rules. As such, an expert panel might provide a more severe test for the algorithmic model. On the other hand, jazz experts might also be (negatively) biased towards computer-generated jazz solos, as this goes against central points of ethics and aesthetics of jazz. Non-experts, on the other hand, while probably not being as sensitive to details as experts, might be less biased in this regard. In light of all these considerations, we thus felt the need that, before conducting a large scale evaluation of our algorithm, we had to address aspects of the evaluation procedure itself first, which will be the second focus of the paper.

3. RELATED WORK

3.1 GENERATION OF JAZZ SOLOS

There have been quite a few attempts to artificially create jazz solos, mostly of the monophonic type. The employed methods range from Markov models (Pachet, 2003, 2012) to rule-based models (Johnson-Laird, 1991, 2002; Quick and Thomas, 2019), (probabilistic) grammars (Keller et al., 2013; Keller and Morrison, 2007), genetic algorithms (Biles, 1994; Papadopoulos and Wiggins, 1998), agent-based approaches (Ramalho et al., 1999) and artificial neural networks (Toiviainen, 1995; Hung et al., 2019; Wu and Yang, 2020). Some of these models do not generate solos in the narrow sense, but, for instance, walking bass lines or lead sheets. Most systems work offline, whereas some are interactive and real-time. A standardized and rigorous evaluation of these models is, however, often lacking. The algorithms were most often only evaluated informally or qualitatively, either by the authors themselves or a small panel of experts. Recently, evaluation using objective features was also employed (Yang and Lerch, 2020). One notable exception is the recent work by Wu and Yang (2020), who used a rather extensive subjective listening test, which however was not the main focus of the study. The evaluation algorithm is similar to the proposed algorithm here, as it is also based on the Weimar Jazz Database and also incorporates mid-level unit annotations. However, as the model is a Transformer-variant, the implementation is quite different.

The Impro-Visor program by Keller and co-workers[3] is open source software and freely available. It seems that they moved recently from their original probabilistic grammars to RNN techniques such as LSTM and GAN-based networks for generation of solos (Trieu and Keller, 2018; Keller et al., 2013).

The JIG system by Grachten (2001) has some similarities to the model proposed here, as it also uses some form of abstract pitch motif, derived from Narmour's implication-realization model (Narmour, 1990). It has also two modes of note generation, called 'melodying' and 'motif', which are, however, more short-ranged than the mid-level units used in our model.

The most recent additions are BebopNet (Haviv Hakimi et al., 2020) and MINGUS (Madaghiele et al., 2021), which are both deep learning Transformer models. BebopNet is based on a large collection of saxophone solos, whereas MINGUS is based on the Weimar Jazz Database and the Nottingham Database. The results of BebopNet can be listened to online and are partly convincing, particularly in longer lines, which might be due to the fact that some real bebop patterns are reproduced by the model. The authors of MINGUS found a similar level of performance of their system to BebopNet.

3.2 EVALUATION OF ARTIFICIAL JAZZ SOLOS

To the best of our knowledge, no systematic work on how to set up musical Turing tests for artificial jazz solos has been undertaken so far. In a recent paper by Yang and Lerch (2020), a strong point was made about the quite sorry state of formal evaluation methods for generative models of music. They acknowledge the power of Turing-

like tests, which are not without problems though, but mainly advocate methods of objective evaluation which boil down to comparing feature distributions between original (training) and generated sets of music, similar to the idea of a "critic" proposed by Wiggins and Pearce (2001). Objective evaluation of generated music has the advantage that it can be unequivocally defined and thus reproducibly measured, but in our view, it can only be a preliminary or auxiliary step. The problem is that it has to rely on an arbitrary (though often obvious) set of features, while the space of possible features is basically infinite. At least, good care and extended domain-knowledge is necessary to devise such a system. Because of the non-trivial but crucial interaction of musical features (pitch, harmony, rhythm, meter, articulation, dynamics and micro-timing), it is hardly to expect that conforming on single feature dimensions alone can guarantee the correct conditional distribution. On the other hand, this way, true Big-C Creativity (Kaufman and Beghetto, 2009) might be precluded. However, as the Pro-c creativity problem is not solved yet, this might not be a relevant problem for the near future.

4. THE MODEL

4.1 OVERVIEW

An overview of the model can be found in ***Figure 1***. The general aim is to generate a note sequence over a given chord sequence for a prespecified number of choruses, i. e., cycles of chord sequences. This is achieved by using a hierarchical model with mid-level units (MLU) at the top level (Section 4.3). After selecting an MLU, a sequence of Weimar Bebop Alphabet (WBA) atoms (Section 4.4) is generated with a first-order Markov model conditioned on the selected mid-level unit. The pitches of these WBA atoms are then "realized" using the given chord context with the help of chord-scale theory (Section 4.6), whereas the rhythm is generated based on a first-order Markov model of duration classes likewise conditioned on the containing mid-level unit. The number of tones in an MLU is predetermined by drawing from the length distributions conditioned on the type of mid-level unit. Even though in the original mid-level annotation system a musical phrase can contain more than one MLU, we make here the simplifying assumption that an MLU always constitutes a single phrase. After a phrase is generated in this way, a short gap or break between phrases is inserted by randomly drawing from the gap duration distribution, whereupon the whole process is repeated up until the specified number of choruses is generated. All involved probability distributions are estimated by the corresponding empirical distributions from the Weimar Jazz Database.

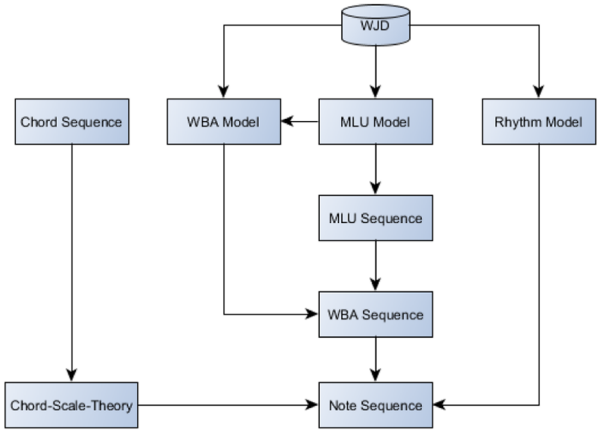

Figure 1 Overview of the generative model.

4.2 THE WEIMAR JAZZ DATABASE

The Weimar Jazz Database is a high-quality database of annotated transcriptions of monophonic jazz solos performed by eminent jazz performers from the US-American jazz canon. It covers nearly the entire history of jazz (1925–2009) and the most important tonal jazz styles, without claiming full representativeness. See ***Table 1*** for a quick overview.[4]

The WJD contains an extensive set of annotations such as metrical annotations, articulation, loudness, chords, forms, and metadata, as well as manually annotated phrases and mid-level units.

4.3 MID-LEVEL UNITS

Mid-level analysis is a content-based qualitative annotation system based on the idea that performers use short-range action plans, which cover a duration of about 2–3 seconds. The system was originally developed for jazz piano improvisations (Lothwesen and Frieler, 2012) and then modified and extended for monophonic solos (Frieler et al., 2016). Interviews with

The Weimar Jazz Database	
Solos	456
Performers	78
Top Performers	Coltrane (20), Davis (19), Parker (17), Rollins (13), Liebman (11), Brecker (10), Shorter (10), S. Coleman (10)
Styles	Traditional (32), swing (66), bebop (56), cool (54), hardbop (76), postbop (147), free (5)
Instruments	ts (158), tp (101), as (80), tb (26), ss (23), other (68)
Time range	1925–2009
Tone events	200,809
Phrases	11,802
Mid-level units	15,402 (containing 5.2 WBA atoms on average)
WBA atoms	80,123 (average length: 2.3 intervals)

Table 1: Short overview of the Weimar Jazz Database.

pianists showed that professional jazz players indeed used similar action plans (Schütz, 2015). Based on an iterative qualitative analysis of solos, a system with nine main types (and 38 subtypes) of playing ideas or design units, called mid-level units, was devised and codified. Subsequently, the entire WJD was annotated manually, with a good inter-rater agreement on section boundaries and acceptable agreement on unit labels. It could further be shown that the different MLU types indeed differ statistically on various aspects and that styles and performers differ in their application of mid-level units (e. g., Frieler, 2018, 2020). The most common MLUs are *line* and *lick* MLUs, covering about 75 % of all MLUs as well as 75 % of solo durations (Frieler et al., 2016). *Lick* MLUs are shorter and rhythmically more diverse, whereas line MLUs are rhythmically more uniform and generally longer. For the present model, only *lick* and *line* MLUs are used. See Section S1.1 in the supplementary information for a more detailed description of the two main types.

4.4 WEIMAR BEBOP ALPHABET

In an effort to find a more compact description of melodies, the Weimar Bebop Alphabet (WBA) was developed (Frieler, 2019). The guiding principle was based on identifying short melodic units that make sense in their own right, either by musical conventions or instrument rehearsal practices such as running scales and arpeggios. The system was devised based on expert knowledge and phenomenological intuition. These units are called WBA atoms and are thought to serve as basic building blocks for melodic construction. As such, they can be applied in principle to any melody, not only to jazz solos. It represents a classification system for interval sequences with six main and nine subcategories. See ***Table 2*** for an overview. The most basic categories are repetitions, scales (diatonic and chromatic), arpeggios, and trills (short oscillations of two tones). More specific is the class of 'approaches', where a target tone is 'encircled' by two other tones, one higher and one lower. Finally, there is a miscellaneous category, dubbed 'X' atoms, with the subcategory of links, which are X atoms of length 1, e. g., only one interval. The current form of the WBA should be viewed as preliminary, as, for instance, the miscellaneous category is the largest. See Frieler (2019) for more details.

Using a priority list, an interval sequence can be segmented uniquely into a sequence of non-overlapping atoms of constant direction (see Frieler (2019) for details). Hence, a WBA atom is unambiguously described by a short symbol for its category, a direction (ascending, descending, or horizontal), and a length (number of intervals). However, a single atom can have different realizations in terms of pitches, except for tone repetitions, which are the only unequivocal category. For instance, an ascending diatonic scale atom of length 3 can have the realizations [+2, +2, +2], [+2, +2, +1], [+2, +1, +2], [+1, +2, +2], [+1, +2, +1] (but not [+1,+1,+1] as this would be a chromatic scale atom). This under-determination is the main work-horse for the generative model.

In Frieler (2019), it was shown that the WBA atoms empirically follow at most a first-order Markov model. This is an interesting result, which undermined one of the original goals of the WBA—to find a significantly more compact description of melodies—, but it simplifies the generative model.

The average length of a WBA atom is, with 2.3 intervals, rather small. This means that every 2 to 3 intervals (3 to 4 tones) a change in character and/or direction takes place, which is indicative of the high variability and complexity of jazz improvisations and is an important part of jazz melodic construction.

4.5 RHYTHM MODEL

The rhythm model is currently a very simple one, basically just a first-order Markov model of inter-onset interval (IOI) classes. For this, inter-onset intervals are classified into five distinct categories (very short, short, medium, long and very long), which roughly correspond to metrical durations of sixteenth, eighth, quarter, half and whole notes. For the current model, IOI class transition probabilities for *lick* and *line* MLUs are sampled and used to generate inter-onset intervals. The current

Type	Subtype/Symbol	Description
Scales	Diatonic (D) Chromatic (C)	Diatonic scale Chromatic scale
Approaches	F	Two intervals approaching a target pitch with a direction change (e. g., –2 +1)
Trills	T	Two alternating pitches
Arpeggios	Simple (A) Jump (J)	Sequence of thirds Sequence of intervals larger than a third
Repetitions	R	
X atoms	X Link (L)	Miscellaneous category X atoms of length 1

Table 2: Overview of WBA atoms.

implementation of model uses only $\frac{4}{4}$ time with a sixteenth note resolution, so the IOI classes are here, in fact, identical to the aforementioned metrical durations. In order to get better results, two further tweaks were applied. For *line* MLUs, the durations were fixed to be either sixteenth or eighth notes, with equal probability. Markov sampling from the true IOI distribution for *line* MLUs produced too many rhythmically inhomogeneous and syncopated rhythms, which shows that rhythm is not adequately modeled by this simple approach. One reason for this is the interaction of rhythm and metrical constraints, as well as the fact that micro-timing somewhat distorts the distribution, since the metrical annotation in the WJD was done algorithmically. Likewise, for *lick* MLUs, the generated onsets needed some smoothing. Here a simple resampling was used until the number of on-beat events was twice as high as the number of off-beat events.

4.6 CHORD-SCALE THEORY

Chord-scale theory was used to fit the WBA atoms to the current chord context. Chord-scale theory was initiated by Russell (1953) and became one of the most popular harmonic theories in jazz education (Cooke and Horn, 2002; Aebersold, 1967; Coker, 1987). Tonal jazz improvisation is based on chords, and style rules demand to select pitches that sufficiently match the underlying chords. In order to achieve this goal, chord-scale theory provides a simple mapping from chords to scales which can be used by a player. For instance, a $Cmaj^7$ chord implies (amongst others) either an Ionian (major) or a Lydian scale, which differ only by the fourth scale degree which is raised in Lydian (♯11). Playing only the pitches from either scale over a $Cmaj^7$ chord will more or less guarantee a sufficient fit. Another advantage of chord-scale theory is that it is a unified framework for tonal and modal jazz alike. There is some discussion about the adequacy, benefits, and disadvantages of chord-scale theory for jazz research and practice (Ake, 2002), but this is outside the scope of this paper. It suffices to say that chord-scale theory is an approximation for modeling tonal choices in certain jazz styles, which seems to be useful for our present purposes.

In chord-scale theory, the mapping of chords to scales is not unique; typically several suitable scales are available to the player, depending on style, taste, and tonal context. For instance, a min7 chord can be mapped to a Dorian or an Aeolian scale, depending on whether it is interpreted as a ii^7 or a i^7 chord in the current context, and whether the tune is considered modal or tonal in character.

For the generative model, we used a simple and fixed mapping of chords to scales with fixed probabilities of being chosen. No attempts were made to find the most appropriate scale for a chord, which would require harmonic analysis and often external style information. This is left for future extensions of the model.

Upon realizing a WBA atom over a chord with a certain starting pitch, first, a suitable scale is randomly selected by a simple weighted sampling from the set of allowed scales in ***Table 3***. Then the pitches are generated based on the current WBA value, which has the form of an interval sequence. For diatonic and arpeggio atoms, occasionally, a link atom is inserted if the current starting pitch is not part of the chord or the corresponding scale. This is the only way link atoms are used in the generative model. The rationale behind this is that a jazz performer might also use links to get from an unsuitable pitch to a suitable one before executing an arpeggio or a diatonic scale, as these should (normally) match the current chord. For a more detailed description how the atoms are realized, see Section S2.2 in the supplementary information.

4.7 IMPLEMENTATION

The main algorithm is depicted in ***Algorithm 1***. It consists of two nested loops: the outer one generates phrases, and inner one generates pitch and rhythm sequences. Pitch sequences are generated based on a first-order Markov model of WBA atoms, conditioned on the MLU, and rhythm sequences are added to the pitch sequences also conditional to the MLU (see Section 4.5).

Input to the algorithm is a lead sheet, i. e., a chord sequence with metrical information, taken either from the iRealPro corpus or extracted from the WJD chord annotations. The model is currently constrained to $\frac{4}{4}$ time and a sixteenth note tatum resolution, but these restrictions could easily be lifted. Further input is a pitch range, in order to avoid running out of instrument ranges. This is ensured by filtering out pitches that are out of range, and by adjusting WBA directions if the current pitch is 30% below the upper or 30% above the lower pitch range limit. This is a crude but effective simulation of actual playing

Chord Type	Scales	Scale content
maj, maj^7	Ionian	[0, 2, 4, 5, 7, 9, 11]
min, min^7	Dorian	[0, 2, 3, 5, 7, 9, 10]
7	Mixolydian	[0, 2, 4, 5, 7, 9, 10]
	Major Blues	[0, 2, 3, 4, 7, 9]
	Mixolydian ♯11	[0, 2, 4, 6, 7, 9, 10]
	Altered Scale	[0, 1, 3, 4, 6, 8, 10]
m7b5	Locrian	[0, 1, 3, 5, 6, 8, 10]
	Phrygian	[0, 1, 3, 5, 7, 8, 10]
o, o^7	Octatonic Scale	[0, 2, 3, 5, 6, 8, 9, 11]

Table 3: Chord-scales used in the current model. Scale contents are given as pitch class vectors with 0 representing the root of the chord.

practice, which results in the common "regression to the mean" in melodic motion (Von Hippel and Huron, 2000).

The main loop runs until the number of specified choruses is reached by using the onset ticks (in sixteenth units) as main control condition. The generated tone events are finally converted to a proprietary CSV representation which is then converted to MIDI or Lilypond scores using the MeloSpySuite/GUI software from the Jazzomat project (Pfleiderer et al., 2017).

One example of a generated solo, that was also used in the evaluation (Algorithm-1-Original, see be-low), can be found in ***Figure 2***.

Algorithm 1: The WBA-MLU algorithm

```
input : LeadSheet, PitchRange,
        TotalNumberChoruses
initialize maxOnset from LeadSheet and
 TotalNumberChoruses;
Onset ← Draw (0:7);
MetricalPosition ← MetricalPositionFromOnset
 (Onset);
while Onset ≤ maxOnset do
    MLU ← Draw(MLUDist);
    PhraseLen ← Draw (PhraseLenDist, MLU);
    NumberOfNotesInPhrase ← 0;
    PitchSequence ← Draw (PhraseStartPitchDist);
    while NumberOfNotesInPhrase ≤ PhraseLen do
        WBA ← Draw (WBA. WBAMarkovModel,
         MLU);
        PitchSequence ← RealizeWBA (WBA, Pitch,
         Chord, PitchRange);
        Rhythm ← GetRhythm (PitchSequence,
         MLU);
        Onset ← UpdateOnset (Rhythm);
        MetricalPosition ←
         MetricalPositionFromOnset (Rhythm);
        Chord ← UpdateChord (MetricalPosition);
        NumberOfNotesInPhrase ←
         NumberOfNotesInPhrase + length
         (PitchSequence)
    end
    Onset ← UpdateOnset (Draw (0:15));
end
```

5. EVALUATION

As discussed in Section 2, we decided to address some general problems of fair evaluation of generated jazz solos using Turing-like tests instead of starting with a large-scale evaluation of the model right away. The results of our exploratory experiment will inform the design of these in the future. Furthermore, we did not explore the possibility of objective evaluations in this paper. We leave this also for the future.

5.1 PREPARATION OF STIMULI

We produced a set of stimuli along the following dimensions:

- **Good vs. bad algorithmic solos.** We generated a set of 50 solos containing one chorus of a simple jazz blues in F (over the chord sequence ‖F^7 | $B\flat^7$ | F^7 | $Cmin^7$ F^7 | $B\flat^7$ | $B\flat^7$ | F^7 | F^7 | $Gmin^7$ | C^7 | F^7 | F^7‖). After listening to the results, we selected one of the most convincing solos and one of the least convincing ones.
- **Tweaked vs. raw algorithmic solos.** For the most convincing artificial solo, we prepared two versions. One was just the solo as generated by the algorithm, for the other one, we tweaked a few notes, which seemed suboptimal to our expert ears, and manually

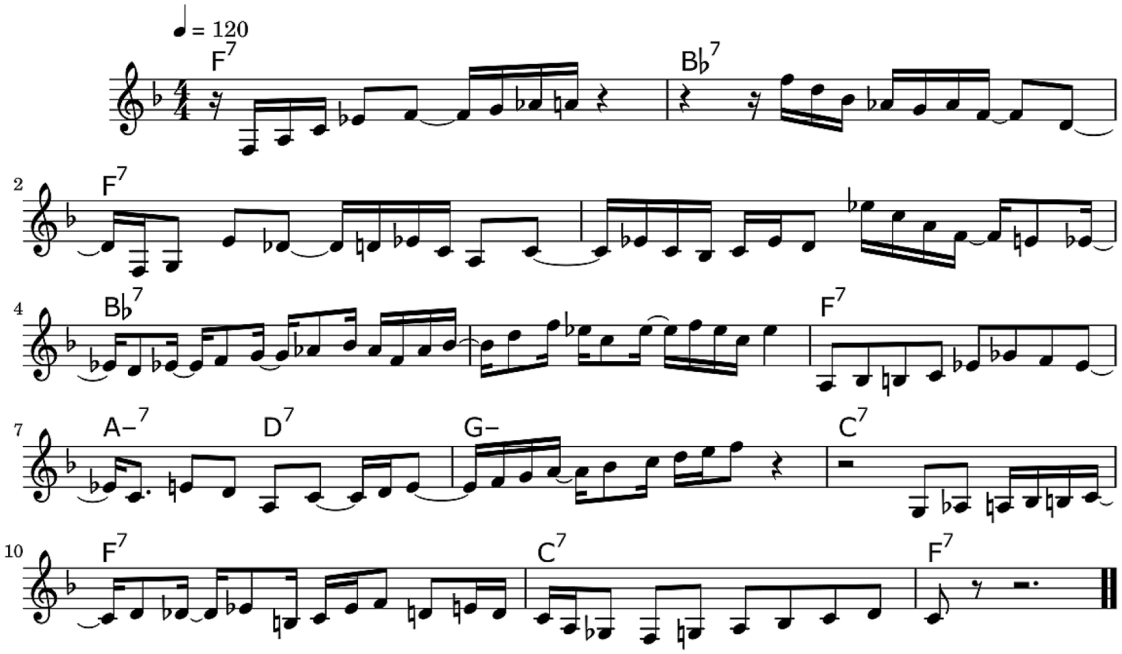

Figure 2 Example of a generated solo over an F-blues chord sequence, used in the evaluation (Algorithm-1-Original).

added microtiming variations and dynamics for a more realistic performance.

- **Human vs. algorithmic solos.** We selected three kinds of human solos to be compared to the algorithmic ones. The first set of human solos were taken at random from the WJD, using the F blues subset. This contained one solo by Charlie Parker ("Billie's Bounce"), Miles Davis ("Vierd Blues"), and Sonny Rollins ("Vierd Blues"). Next, we took four solos from a former (unpublished) study, where jazz students had improvised solos to an F blues play-along. The students had different levels of expertise (beginner, intermediate, advanced, and graduated). Thirdly, the authors recorded one solo each over the backing track used for all stimuli. The first author (AUT1) played a single line solo on a digital piano and the second author (AUT2) played a solo on electric guitar.
- **Original vs. MIDI-fied solos.** We used the original recordings of the authors and also produced two MIDI-fied versions by either using the recorded MIDI from the piano solo or an automatically converted version of the guitar solo by using the audio-to-MIDI converter of the DAW plug-in Melodyne (editor version).

All MIDI versions were rendered with a tenor sax sample over the same backing track with piano, bass, and drums (see Section S5 in the supplementary information). Only the two original solos by the authors did not use the tenor sax sound and played thus the role of a baseline condition. For all candidate solos, only the first chorus was used and rendered with tempo 120 BPM to create our stimuli, which lasted for approximately 25 seconds each (see ***Table 4***).

5.2 METHOD

We prepared an online survey using the SoSci-Survey platform with the 14 stimuli as the core items. Each solo had to be assessed on a questionnaire containing 10 Likert-like items (cf. Table S2 for a complete list) with answer options ranging from 1 = "completely disagree" to 7 = "completely agree" for all items except items 8 to 9, which had their own range but with the same polarity. The rationale was to present items that reflect typical qualitative judgments of jazz solos that do not use deep jazz-specific terminology. The sequence of solos was randomized for each participant.

We collected basic demographic data (age, gender) and asked three self-assessment questions pertaining to jazz and music expertise ("I am a jazz expert", "I am a jazz fan", "I am a music expert") using the same 7-option Likert-scale. We also asked the participants for textual feedback on the experiment itself.

Ethics approval was not required by our host institution for this study. Participants gave their informed consent before starting the experiment.

5.3 PARTICIPANTS

By advertising on social media and approaching friends and colleagues, we obtained a convenience sample of 41 participants (7 female; mean age 27.0, SD 9.9), of which

Id	Solo ID	Generator	Performance Type	Solo Sound
1	Algorithm-1-Original	WBA-MLU-Algorithm	Deadpan MIDI	tenor sax
2	Algorithm-1-Improved	WBA-MLU-Algorithm/AUT2	MIDI with microtiming	tenor sax
3	Algorithm-2-Original	WBA-MLU-Algorithm	Deadpan MIDI	tenor sax
4	WJD-Sonny Rollins	Sonny Rollins ("Vierd Blues")	MIDI with microtiming	tenor sax
5	WJD-Miles Davis	Miles Davis ("Vierd Blues")	MIDI with microtiming	tenor sax
6	WJD-Charlie Parker	Charlie Parker ("Billie's Bounce")	MIDI with microtiming	tenor sax
7	Student-Beginner	Beginner	MIDI with microtiming	tenor sax
8	Student-Intermediate	Intermediate	MIDI with microtiming	tenor sax
9	Student-Advanced	Advanced	MIDI with microtiming	tenor sax
10	Student-Graduated	Graduated	MIDI with microtiming	tenor sax
11	Author-Original	AUT2	Audio	e-guitar
12	Author-MIDI	AUT2	Converted audio-to-MIDI	tenor sax
13	Author-Original	AUT1	Audio	piano
14	Author-MIDI	AUT1	Recorded MIDI	tenor sax

Table 4: Stimuli used for the evaluation. In column Performance Type specifics of the interpretation are given. *Deadpan MIDI :* Fully-quantized MIDI without dynamics; *MIDI with microtiming:* MIDI with semiautomatically added microtiming (swing); *Audio:* Recorded audio; *Converted audio-to-MIDI :* recorded audio converted to MIDI with Melodyne, keeping microtiming and dynamics; *Recorded MIDI :* human-played MIDI with microtiming and dynamics.

29 were identified as jazz experts based on the sum of responses to the three expertise items being greater or equal to 12. The overall median value on the item "I am jazz fan" was 7 ("completely agree"). In conclusion, we can say that the sample contains a large share of jazz experts (on different levels), while all self-identified as jazz fans.

5.4 RESULTS

5.4.1 Solo characteristics

As the scores on all items (except item 10) showed various strong correlations, we reduced the variables by using a factor analysis with three factors and *oblimin* rotation, which explained 85% of the variance (see Section S2.1 in the supplementary information). The factors were named MUSICALITY (*convincing, liking, expressive, swing, inventive, expertise*), COMPLEXITY (*virtuosic, complex*), and RHYTHM_EXACT (*rhythm exact*). The number of factors was determined using standard methods (KMO, Screeplot). The factor RHYTHM_EXACT will be not considered further, as it is not very informative for our aims here.

In ***Figure 3 (A)***, MUSICALITY ratings can be found. Algorithmic solos were ranked very low, except for the enhanced solo. Two student solos were rated even lower than all algorithmic solos. As expected, the most natural-sounding solo, one of the author solos (AUT2-Original), was rated highest. Rankings by experts and non-experts are more or less similar. Note that the range of values are quite large for most of the solos. For the COMPLEXITY factor, as seen in ***Figure 3(B)***, two of the student solos and the first algorithmic solo in two versions were among the top 5. Two other student solos and the Davis and the Rollins solo were rank lowest. The MIDI-fied versions of the Author's solos were consistently rated lower here than the original versions, despite identical musical content. Again, the range of values is quite large.

5.4.2 Recognition of origin

We accounted an algorithmically generated solo as correctly (and rather confidently) recognized, if the answer on Item 10 ("Do you think the notes of this solo were generated by a computer algorithm?", variable *artificial*) had a value of 5 or larger, otherwise we accounted it as unrecognized. Conversely, a human generated solo was accounted as correctly recognized if it received a value of 3 or less on Item 10, and as unrecognized otherwise. The recognition accuracy of a solo is then defined as the proportion of responses counted as correctly recognized. The results can be found in ***Table 6***, separately for experts and non-experts and human and algorithmically generated solos. A more detailed display for all 14 stimuli can be found in ***Figure 5*** and in Table S4 in the supplement. Only four solos have a recognition accuracy whose 95 % confidence intervals do not cross the random baseline of 50 % (AUT2-Original, WJD-Sonny Rollins, Student-Intermediate, Algorithm-2-Original). For the non-jazz experts, this is only true for one solo (AUT2-Original). Some solos are consistently misclassified (AUT1-MIDI and Student-Graduated). Algorithm-2-Original, the unconvincing solo, is successfully identified as computer-generated with a mean accuracy of 69 %, though this comes mostly from the jazz experts; the non-experts are basically guessing here. Algorithm-1-Original performs better with an accuracy of 59 %. The improved version Algorithm-1-Improved is able to fool the raters with an overall mean accuracy of 44 %, but experts are

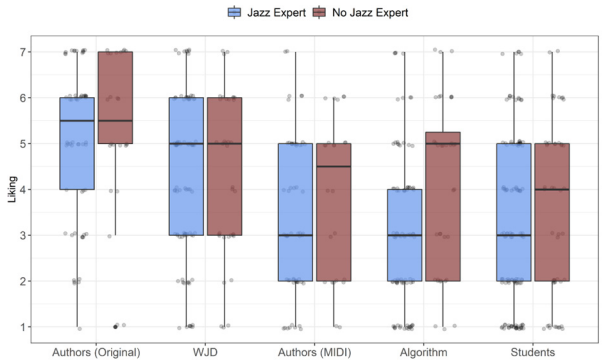

Figure 4 Boxplot of *liking* values for sources of solos, separately for rater expertise. Left boxes: jazz experts, right: non-experts.

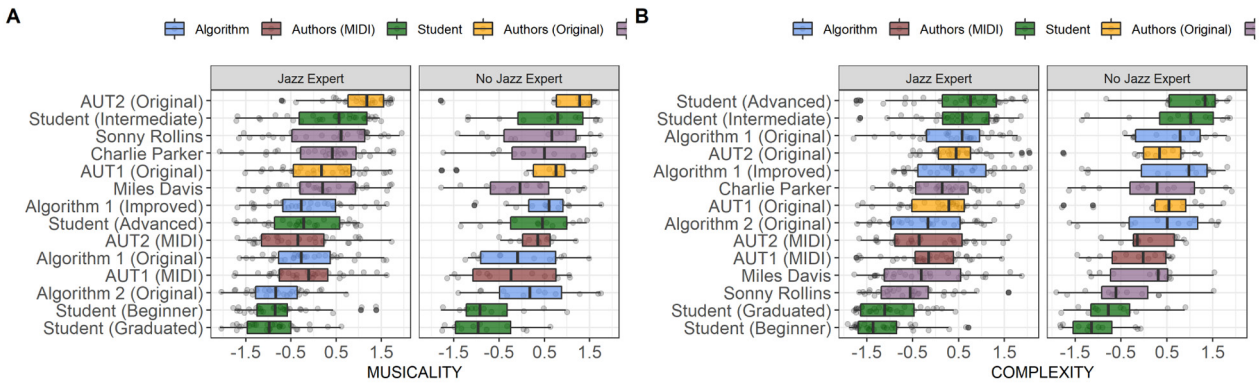

Figure 3 Boxplot of MUSICALITY (A) and COMPLEXITY (B) values for all solos, separately for rater expertise. Left panels: Jazz experts; right panels: non-experts; blue: algorithmic solos; brown: author (MIDI) solos; yellow: student solos; green: author (original) solos; violet: WJD solos.

	MUS	COMP	*artificial*
MUS	1.00	0.44	–0.59
COMP	0.44	1.00	–0.18
artificial	–0.59	–0.18	1.00

Table 5: Pearson's correlation coefficients for MUSICALITY (MUS), COMPLEXITY (COMP), and *artificial* (Item 10). All correlations $p \leq .001$.

Solo Generator	Expertise	Accuracy
Algorithm	Expert	.644
	Non-expert	.417
Human	Expert	.536
	Non-expert	.447

Table 6: Recognition accuracy for computer and human generated solos by experts and non-experts.

better with an accuracy of 53 % slightly over the random baseline, whereas non-experts are completely at loss with an accuracy of only 18 %. Interestingly, one of the student solos (Student-Graduated) is the one most clearly misclassified. The MIDI-fied versions of the author solos are regarded as computer-generated by experts and non-experts alike. This might be an effect of the different articulation and approaches by rendering piano and guitar solos with a tenor saxophone sound, e. g., due to differing attack times. They are also rated much worse on the other factors compared to the original version. This result tells a cautionary tale. For the WJD solos, Sonny Rollins's solo is most clearly identified as human-generated, whereas the other WJD solos are rated much more ambiguously. For the Charlie Parker solo this might come from the fact that the backing track had a slightly different chord sequence than the original solo and the tempo was perceivably slowed down from the original. The original version of Miles Davis's solo also has slightly different chords, but the most important factor might be that the very spacious solo of Davis works because the piano player fills in the spaces, which is, of course, not the case for the MIDI-fied version used here.

The pooled accuracies for expert/non-expert raters and human/algorithmic solos can be found in ***Table 6***. Experts had an accuracy of 64 % for correctly identifying algorithmic and 54 % for human solos, which is only slightly above chance. For non-experts the aggregated values are even below chance level, 41 % for algorithmic and 44 % for human solos, which means that non-experts tend to assume algorithms at work even when this was not the case. This might be a result of the explicit framing of the survey as a "Turing Test" for jazz, which might have raised the baseline expectation for computer-generated solos, while, in fact, only a minority (3 out of 14) were actually computer-generated.

5.4.3 Relationship of identification and characteristics

A correlation analysis of the two factors MUSICALITY and COMPLEXITY with *artificial* can be seen in ***Table 5***. MUSICALITY and COMPLEXITY are strongly positively correlated, whereas MUSICALITY and *artificial* are strongly negatively correlated. As *liking* (item 8) is the strongest contributing factor to MUSICALITY, we checked if recognition accuracy might be related to liking. Interestingly, *liking* and *recognition accuracy* are strongly positively correlated with $r = .86$ for human generated solos, but strongly negatively correlated for computer-generated solos with $r = -.66$. This suggests that the participants judge solos as human-generated on the basis of their liking of the solo, probably based on a bias against artificially generated music (Moffat and Kelly, 2006).

5.4.4 Relationship of recognition and liking

We checked further, if and to what extent the solos were aesthetically pleasing for the participants, by looking at the ratings on item 8 ("How did you like the solo excerpt? (not at all-very much)?"). First, we conducted a mixed linear regression (package `lmerTest` for `R`) to see whether jazz experts and non-expert differ in their overall liking scores. Indeed, the non-expert had slightly higher *liking* score ($\beta = 0.385$, $df = 531$, $p = .018$), with experts having a mean of 3.72 and non-experts one of 4.1, so non-experts seem to be more forgiving. There were also stark differences between subjects in the ratings. The participants mean values of liking ranged from 1.29 to 5.43 with a mean of 3.8, a median of 4.0, and a standard deviation of 0.95. The difference in *liking* with respect to the source of the solo (author original, author MIDI, WJD, student, algorithm) can be found in ***Figure 4***. As expected, the original solos were liked the best (mean *liking* = 5.14), followed by the master solos from the WJD (mean *liking* = 4.37), the MIDI-fied author solos (mean *liking* = 3.47), the algorithmic solos (mean *liking* = 3.37), and the students' solos (mean *liking* = 3.29). A linear mixed model with source and jazz expertise as fixed effects and participant as random effect showed that there is no significant difference between experts and non-experts if individual preferences are taken into account, and that only original ($\beta = 1.77$, $SE = 0.23$, $Z(503) = 7.7$. $Pr(> |t|) < .0001$) and WJD solos ($\beta = 1.00$, $SE = 0.206$, $Z(503) = 4.85$. $Pr(> |t|) < .0001$) were significantly more liked than algorithmic, MIDI-fied author, and student solos. Actually, two of the student solos were liked less than all three algorithmic solos, whereas the other two were liked considerably better. The mean liking values for all sources did not reach the neutral level of 4 except for WJD and author originals. The liking ratings of the author solo AUT2 dropped from 5.69 for the original recording to 3.51 for the MIDI-fied tenor sax version, a huge loss of $d = 2.19$ on a 7-point scale. For AUT1, the drop was from 4.59

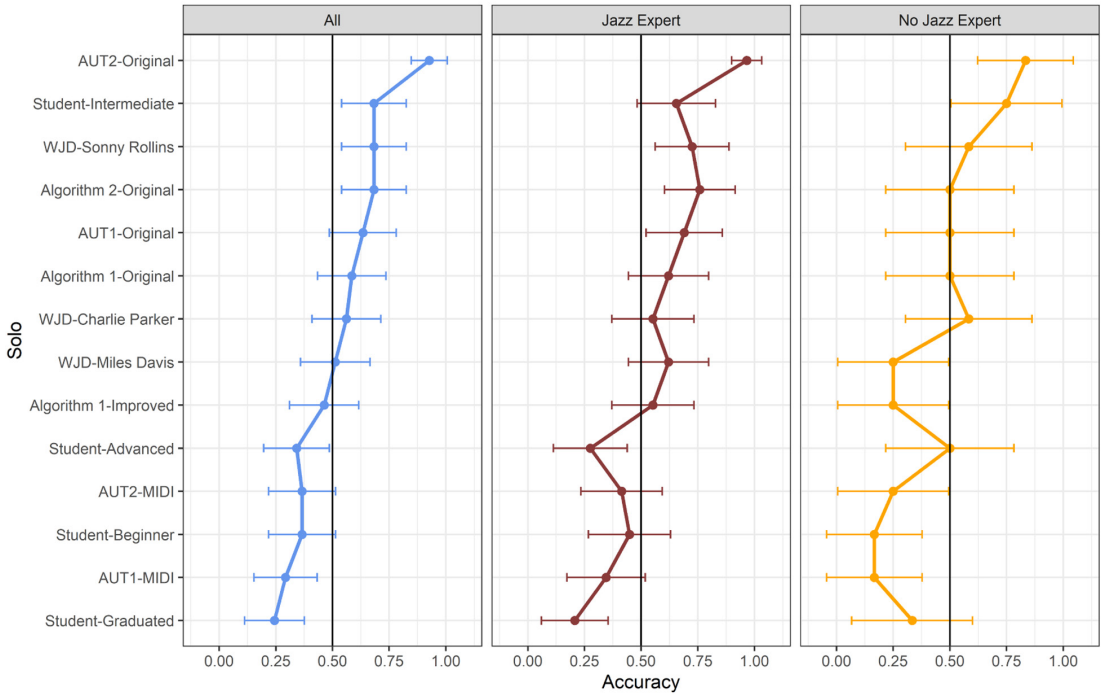

Figure 5: Recognition accuracy of all 14 stimuli by expertise level. Left: all, middle: jazz experts, right: no jazz experts. Error bars are 95% confidence interval of proportion.

to 3.44 (d = 1.15). The difference, particularly for AUT2, is striking and clearly demonstrates the influence of a natural-sounding performance on aesthetic appreciation (of jazz solos).

For the variable *inventive*, the ranking of solos is very similar to that of *liking*, with the exception that Algorithm-1-Improved is ranked fourth. In terms of source or origin, the algorithmic solos were tied with the MIDI-fied author solos (mean *inventive* = 3.77 for both), whereas the rest of the ranking was the same as for *liking* (Originals: 4.42, WJD: 3.85, students: 3.4).

6. DISCUSSION

6.1 EVALUATION OF OUR MODEL

We presented a novel algorithm to generate monophonic jazz solos over a given chord sequence. We evaluated the algorithm with a Turing-like listening test with 41 jazz-affine participants. We could show that a hand selected, edited, and expressive rendition of one of the generated solos (Algorithm-1-Improved) could fool the panel, as it was slightly more often considered to be human-generated than computer-generated. Moreover, the recognition accuracies verged on the chance level, so we can state that even the jazz experts were not entirely sure about its origin. On the other hand, the unedited, deadpan version of the same solo was successfully recognized as computer-generated, at least by the experts, but still with considerable uncertainty. The second, "bad" solo was even more clearly recognized as computer-generated, mainly by the jazz experts in the panel, whereas the non-experts were not completely sure here either. These results were anticipated, but they have to be viewed in perspective, as even solos by eminent jazz masters such as Charlie Parker and Miles Davis were frequently judged as computer-generated, when presented as deadpan MIDI over a computer-generated backing track. Only the original audio solo by the second author was unequivocally considered to be human-generated, which was expected as this was specifically included as a baseline. However, the second original solo (AUT1-Original) was not as often recognized as human-generated and the MIDI-fied version of the second original solo (AUT2-MIDI) was likewise considered mainly to be computer-generated. This clearly demonstrates that an expressive, "natural" performance is crucial for human judgements in Turing-like music tests. Furthermore, comparison of algorithmically generated solos with those of jazz greats might also not be the most important test as the computer-generated solos were recognized correctly more often than three of the four student solos. This seems to make sense, as devising a successful algorithm that is able to invent masterly solos seems a bit too much to ask for, thus a comparison with less proficient performers might be more fair.

The performance of our new algorithm is promising, as it is just in its early stages, merely a proof-of-concept, and uses a relatively simple model. This is however quite powerful, because it is based on empirical and analytical results of jazz improvisation and powered by a rather large database of solo transcriptions.

There are many possible avenues for improvement. The most obvious weakness of the model is the rhythm

model, particularly for *lick* MLUs. The simplified rhythm model for *line* MLUs works rather well. For further improvements on the rhythm model, one would need a indepth analysis of rhythm and meter and their interaction in jazz solos, similar to the WBA study, but which is unfortunately missing, as of yet. Also, the model does not allow for incorporating pre-learned patterns, which basically all jazz improvisers do (Norgaard, 2014). We plan to improve the algorithm along-side further empirical research on jazz improvisation in an iterative process.

6.2 EVALUATION OF EVALUATION

We also explored the problem space of Turing-like evaluation of computer-generated jazz solos. Along our basic distinctions, we found these results.

- **Good vs. bad algorithmic solos.** The worse (as judged by the authors) algorithmic solo was less favorably evaluated by our respondents in all aspects, and also much more often correctly identified as computer-generated.
- **Tweaked vs. raw algorithmic solos.** Even a little tweaking of the musical surface can improve the assessment of a solo. One possible reason is that even small or spurious signals of "non-authencity" can be picked up by humans to make their judgement.
- **Human vs. algorithmic solos.** Solos by jazz masters were generally better judged than solos by jazz students and algorithmic solos, but some of the algorithmic solos performed clearly better than student solos. On the other hand, deadpan synthesised solos of masters were rated worse than "natural"-sounding solos by the authors.
- **Original vs. MIDI-fied solos**. Performance seems crucial as the MIDI-fied versions of the original author solos rendered with different instrumental sounds were rated much worse and also less often recognized as human-generated solos.

Besides this, we found generally low inter-rater agreement and large variances of judgements, and saw clear differences between jazz experts and non-jazz experts. Finally, we found evidence of bias against computer-generated music, in the sense that participants seemed to expect computer-generated solos to be worse and were more likely to assume non-human origin if they did not like the solos. We also noted that by framing the experiment as a "Turing Test" people tended to expect more computer-generated solos than there actually were. Hence, their rating behaviour could have been influenced by expectations for these kinds of studies.

7. CONCLUSION AND OUTLOOK

In light of these results, we see three possible approaches for large-scale evaluation of generative models for jazz solos.

1. Using Human- and computer-generated solos performed by a single human player over a fixed accompaniment to keep all background factors constant while providing reasonably expressive performances with microtiming, articulation, dynamics, and timbre features (e. g., vibrato, slurs, bends). This procedure requires considerable effort and is probably not suitable for testing many solos at once. Interestingly, in the free test feed-back field of the survey, where we asked for further comments, many participants suggested exactly this kind of procedure (see Section S5 in the supplementary information).
2. Devising a system that is capable of generating sufficiently natural jazz performances. This would be very efficient for quick and frequent evaluations in conjunction with using a service for recruiting online participants. Such an algorithm does not exist yet and might thus require considerable development effort. It might be in reach with the current state of technology. However, in contrast to the field of classical music, not much research has gone into developing such a system, yet, but see Friberg et al. (2021); Arcos et al. (1998). Once such a system is available, large scale evaluation will become easy and cheap.
3. Using deadpan versions of human- and computer-generated solos. Because of the low baseline accuracy in this setup, a large number of raters and solos would be required to get reliable estimates. The advantage of this approach is that it is relative cheap to realize, even though the recruitment of a sufficiently large number of experts might be an issue, but using an even larger number of non-experts could remedy this problem.

Finally, there is the complementary option of evaluating solos based on objective features, as proposed by Yang and Lerch (2020), similar to the "critic" in the evaluation framework proposed by Wiggins and Pearce (2001). This is rather straightforward to implement if suitable corpora are available and we want to explore it in the future. Another option, often tacitly or explicitly part of any algorithm development, are formal or informal analyses by experts. For instance, some tweaks and design decisions that ended up in the current version of the WBA-MLU algorithm were informed by the expertise of the authors.

But, ultimately, we think that there will be no way around Turing-like tests, as objective feature distributions are not likely to provide a sufficient description of music, and clearly not of truly innovative music, whereas expert analytic evaluations do not scale. Such approaches could be useful to select the most promising candidates from a set of generated solos (if style conformity is the goal, which is often the case in analysis-by-synthesis contexts).

One last remark should be made in regard that the evaluation of human vs. computer-generated solos is driven by aesthetics and a bias against computer-generated music. This implies that for commercial (and other) applications of algorithmically generated music, it has to be very good, i. e., being unrecognizable as such, otherwise the knowledge about its origin can result in audience aversion against the product.

Here, more research is needed, as in this small pilot study, we could only find some evidence in this direction, which clearly warrants more targeted and systematic examination of this phenomenon (cf. Moffat and Kelly (2006); Chamberlain et al. (2018), who found such bias against computer-generated music and art-works).

This study had a mostly exploratory nature, both in regard to the proposed novel algorithm and an evaluation procedure for monophonic jazz solos, and presents promising and insightful first results to both aspects. Our future plans are twofold. First, we want to elaborate the WBA-MLU model further, as many simplifying assumptions were used at the moment in order to create a functioning system. Secondly, we plan to improve the evaluation framework. We think it is worth to check if the strategy that humans play artificially generated solos is feasible. Additionally, we think that the development of a system that is capable to render more natural-sounding performances, if probably only for a single type of instrument, is in reach with the current state of technology.

NOTES

1 For instance, to provide backing tracks for practising soloing, e. g., Band-in-a-Box *https://www.pgmusic.com/* or iRealPro *https://www.irealpro.com/*.

2 A list of commercial AI music generators can be found, for instance, here: *https://topten.ai/music-generators-review/*. The Top 3 names are Amper Music *https://www.ampermusic.com/*, AIVA *https://www.aiva.ai/*, and Ecrett Music *https://ecrettmusic.com/*.

3 *https://www.cs.hmc.edu/~keller/jazz/improvisor/*.

4 For a full list see *https://jazzomat.hfm-weimar.de/dbformat/dbcontent.html*.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary Material.** PDF with further information. DOI: *https://doi.org/10.5334/tismir.87.s1*

REPRODUCIBILITY

There is an accompanying OSF site for this paper: *https://osf.io/kjsdr*. It contains the R project `jazz-turing` with the survey data, the audio stimuli, and the analysis code for the evaluation. The folder `samples` provides 40 generated solos as MIDI files and scores. There is also a link to `parkR`, an `R` package for solo generation based on our model, which can be installed from *https://github.com/klausfrieler/parkR*.

ACKNOWLEDGEMENTS

We like to thank all participants in the evaluation experiment, Simon Dixon for proof-reading the manuscript, and three anomymous reviewers for their helpful comments.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

KF developed the generative algorithm, developed, conducted, and analyzed the evaluation survey, and wrote the paper. WGZ developed and conducted the evaluation survey and wrote the paper.

AUTHOR AFFILIATIONS

Klaus Frieler *orcid.org/0000-0002-6055-377X*
Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany

Wolf-Georg Zaddach
Leuphana University Lüneburg, Germany

REFERENCES

Aebersold, J. (1967). A New Approach to Jazz Improvisation.

Ake, D. A. (2002). *Jazz Cultures*. University of California Press, Berkeley. DOI: *https://doi.org/10.1525/9780520926967*

Arcos, J. L., Lopéz de Mántaras, R., and **Serra, X.** (1998). SaxEx: A case-based reasoning system for generating expressive musical performances. *Journal of New Music Research*, 27:194–210. DOI: *https://doi.org/10.1080/09298219808570746*

Biles, J. A. (1994). GenJam: Evolution of a jazz improviser. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 131–137.

Chamberlain, R., Mullin, C., Scheerlinck, B., and Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics,*

Creativity, and the Arts, 12(2):177–192. DOI: *https://doi.org/10.1037/aca0000136*

Coker, J. (1987). *Improvising Jazz*. Simon & Schuster, New York, 1st fireside edition.

Cooke, M., and **Horn, D.** (2002). *The Cambridge Companion to Jazz*. Cambridge University Press, Cambridge, UK. OCLC: 758544526. DOI: *https://doi.org/10.1017/CCOL9780521663205*

Friberg, A., Gulz, T., and **Wettebrandt, C.** (2021). Computer tools for modeling swing timing interactions in a jazz ensemble. In *16th International Conference on Music Perception and Cognition and 11th Triennial Conference of the European Society for the Cognitive Sciences of Music (ICMPC16-ESCOM11)*, Sheffield, UK.

Frieler, K. (2018). A feature history of jazz solo improvisation. In Knauer, W., editor, *Jazz @ 100: An Alternative to a Story of Heroes*, volume 15 of *Darmstadt Studies in Jazz Research*. Wolke Verlag, Hofheim am Taunus.

Frieler, K. (2019). Constructing jazz lines: Taxonomy, vocabulary, grammar. In Pfleiderer, M. and Zaddach, W.-G., editors, *Jazzforschung heute: Themen, Methoden, Perspektiven*, Berlin. Edition Emwas.

Frieler, K. (2020). Miles vs. Trane: Computational and statistical comparison of the improvisatory styles of Miles Davis and John Coltrane. *Jazz Perspectives*, 12(1):123–145. DOI: *https://doi.org/10.1080/17494060.2020.1734053*

Frieler, K., Pfleiderer, M., Abeßer, J., and **Zaddach, W.-G.** (2016). Midlevel analysis of monophonic jazz solos: A new approach to the study of improvisation. *Musicae Scientiae*, 20(2):143–162. DOI: *https://doi.org/10.1177/1029864916636440*

Grachten, M. (2001). JIG: Jazz Improvisation Generator. In *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Spain.

Haviv Hakimi, S., Bhonker, N., and **El-Yaniv, R.** (2020). BebopNet: Deep neural models for personalized jazz improvisations. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, Montréal, Canada.

Hiller, L. A., and **Isaacson, L. M.** (1959). *Experimental Music: Composition With an Electronic Computer*. McGraw-Hill, New York.

Hung, H.-T., Wang, C.-Y., Yang, Y.-H., and **Wang, H.-M.** (2019). Improving automatic jazz melody generation by transfer learning techniques. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–8, Lanzhou, China. DOI: *https://doi.org/10.1109/APSIPAASC47483.2019.9023224*

Johnson-Laird, P. N. (1991). Jazz improvisation: A theory at the computational level. In Howell, P., West, R., and Cross, I., editors, *Representing Musical Structure*, London. Academic Press.

Johnson-Laird, P. N. (2002). How jazz musicians improvise. *Music Perception*, 10:415–442. DOI: *https://doi.org/10.1525/mp.2002.19.3.415*

Kaufman, J. C., and **Beghetto, R. A.** (2009). Beyond big and little: The Four C Model of Creativity. *Review of General Psychology*, 13(1):1–12. DOI: *https://doi.org/10.1037/a0013688*

Keller, R., Schofield, A., Toman-Yih, A., Merritt, Z., and **Elliott, J.** (2013). Automating the explanation of jazz chord progressions using idiomatic analysis. *Computer Music Journal*, 37(4):54–69. DOI: *https://doi.org/10.1162/COMJ_a_00201*

Keller, R. M., and **Morrison, D.** (2007). A grammatical approach to automatic improvisation. In *Proceedings of the 4th Sound and Music Computing Conference*, pages 330–337, Lefkada, Greece.

Lothwesen, K., and **Frieler, K.** (2012). Gestaltungsmuster und Ideenuss in Jazzpiano-Improvisationen: Eine Pilotstudie zum Einfluss von Tempo, Tonalität und Expertise. In Lehmann, A., Jeßulat, A., and Wünsch, C., editors, *Kreativität: Struktur und Emotion*. Königshausen & Neumann, Würzburg.

Madaghiele, V., Lisena, P., and **Troncy, R.** (2021). MINGUS: Melodic improvisation neural generator using Seq2Seq. In *22nd International Society for Music Information Retrieval Conference*.

Moffat, D., and **Kelly, M.** (2006). An investigation into people's bias against computational creativity in music composition. In *Third Joint Workshop on Computational Creativity*, ECAI 2006, Trento, Italy. Universita di Trento.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. University of Chicago Press, Chicago.

Norgaard, M. (2011). Descriptions of improvisational thinking by artist-level jazz musicians. *Journal of Research in Music Education*, 59(2):109–127. DOI: *https://doi.org/10.1177/0022429411405669*

Norgaard, M. (2014). How jazz musicians improvise: The central role of auditory and motor patterns. *Music Perception: An Interdisciplinary Journal*, 31(3):271–287. DOI: *https://doi.org/10.1525/mp.2014.31.3.271*

Owens, T. (1974). *Charlie Parker: Techniques of Improvisation*. PhD thesis, University of California, Los Angeles.

Pachet, F. (2003). The Continuator: Musical interaction with style. *Journal of New Music Research*, 32(3):333–341. DOI: *https://doi.org/10.1076/jnmr.32.3.333.16861*

Pachet, F. (2012). Musical virtuosity and creativity. In McCormack, J. and d'Inverno, M., editors, *Computers and Creativity*, pages 115–146. Springer, Berlin, Heidelberg. DOI: *https://doi.org/10.1007/978-3-642-31727-9_5*

Papadopoulos, G., and **Wiggins, G.** (1998). A genetic algorithm for the generation of jazz melodies. In *Human and Artificial Information Processing: Finnish Conference on Artificial Intelligence (STeP'98)*, Jyväskylä, Finland.

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., and **Burkhart, B.,** editors (2017). *Inside the Jazzomat: New Perspectives for Jazz Research*. Schott Music GmbH & Co. KG, Mainz, 1. Auflage OCLC: 1015349144.

Pressing, J. (1984). *Cognitive Processes in Improvisation.* Cognitive Processes in the Perception of Art. Elsevier, North-Holland. DOI: *https://doi.org/10.1016/S0166-4115(08)62358-4*

Pressing, J. (1988). Improvisation: Method and models. In Sloboda, J. A., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*, pages 129–178, Oxford. Clarendon. DOI: *https://doi.org/10.1093/acprof:oso/9780198508465.003.0007*

Quick, D., and **Thomas, K.** (2019). A functional model of jazz improvisation. In *Proceedings of the 7th ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design (FARM 2019)*, pages 11–21, Berlin, Germany. ACM Press. DOI: *https://doi.org/10.1145/3331543.3342577*

Ramalho, G. L., Rolland, P.-Y., and **Ganascia, J.-G.** (1999). An artificially intelligent jazz performer. *Journal of New Music Research*, 28(2):105–129. DOI: *https://doi.org/10.1076/jnmr.28.2.105.3120*

Russell, G. (1953). *George Russell's Lydian chromatic concept of tonal organization.* Concept Pub. Co, Brookline, Mass. OCLC: ocm50075662.

Schütz, M. (2015). *Improvisation im Jazz: eine empirische Untersuchung bei Jazzpianisten auf der Basis der Ideenussanalyse.* Number 34 in Schriftenreihe Studien zur Musikwissenschaft. KovaČHamburg. OCLC: 915812622.

Toiviainen, P. (1995). Modelling the target-note technique of bebop-style jazz improvisation: An artificial neural network approach. *Music Perception*, 12:398–413. DOI: *https://doi.org/10.2307/40285674*

Trieu, N., and **Keller, R.** (2018). JazzGAN: Improvising with generative adversarial networks. In *Proceedings of the 6th International Workshop on Musical Metacreation (MUME 2018)*, Salamanca, Spain.

Von Hippel, P., and **Huron, D.** (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. Music Perception: *An Interdisciplinary Journal*, 18(1):59–85. DOI: *https://doi.org/10.2307/40285901*

Wiggins, G., and **Pearce, M.** (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 22–32, York.

Wu, S.-L., and **Yang, Y.-H.** (2020). The Jazz Transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures. In *21st International Society for Music Information Retrieval Conference*, pages 142–149, Montréal, Canada.

Yang, L.-C., and **Lerch, A.** (2020). On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784. DOI: *https://doi.org/10.1007/s00521-018-3849-7*

TO CITE THIS ARTICLE:
Frieler, K., & Zaddach, W.-G. (2022). Evaluating an Analysis-by-Synthesis Model for Jazz Improvisation. *Transactions of the International Society for Music Information Retrieval*, 5(1), 20–34. DOI: *https://doi.org/10.5334/tismir.87*

Submitted: 26 February 2021 Accepted: 24 November 2021 Published: 03 February 2022