We propose to develop a new comprehensive generative model for musical creativity, utilizing deep learning analysis of canonical jazz improvisation and informed by input from musicians and lay listeners. We choose jazz improvisation as our target area since it features real-time musical creativity both on the symbolic level (what is being played) and the rendition level (how it is being played). Therefore, unlike previous efforts in machine learning and computational creativity, our system will be trained with both symbolic information using transcribed musical scores as well as with data extracted from audio recordings of these improvisations utilizing Music Information Retrieval (MIR) methods. Based on score annotation by experts, we will develop a novel hierarchical deep architecture designed to capture semantically meaningful symbolic features such as phrasing and structure. Furthermore, the system will be designed to model semantically meaningful performance features, informed by audio analysis such as micro-timing and dynamics. In addition, the network will be constrained by input from humans aimed at capturing perceptual elements of the music listening experience. The proposed model, based on a constrained LSTM (Long Short Term Memory) architecture, bears the promise of generating semantically and perceptually informed musical improvisations that follow a well-structured musical arch using rich performance articulations, leading to novel expressive and creative musical outcomes. The project could, therefore, capture and generate nuanced and expressive human-like musical creativity, leading to engaging and meaningful musical experiences that are hitherto unmatched by any other artificial creativity systems. It will be systematically evaluated using both subjective and objective methods.

Intellectual Merit: The project will advance our current understanding of musical creativity — both in generation and perception — by creating four main intellectual contributions: (i) for the first time, both symbolic musical structure and expressive performance elements will be analyzed and used for musical generation; (ii) novel hierarchical architectures for recurrent neural networks will be designed to capture higher-level semantics of transcribed and recorded jazz improvisations; (iii) new methods for addressing insufficient training data and high-dimensional input data will be developed and systematically evaluated; (iv) input from human musicians and listeners will be integrated into the deep learning model as constraints, which could provide new insights into the mostly uncharted territory of aesthetic, emotional, and perceptual aspects of listening to and creating music.

Broad Impact: Our proposed generative system would lead to broad impact on the field of computational music generation by introducing a novel approach that combines both symbolic and performance parameters. Moreover, the recursive neural network we will develop could lead to broad impact in other domains that maintain structural hierarchies but where training data is limited. In particular, our approach would be useful for domains with hierarchical sequential data such as natural language modeling and the generation of speech. Insights that will be gained from our improvisational computational creativity model may lead to impact in other domains that include real-time decision making. The project will also lead to broad impact by creating a large annotated corpus of transcribed improvisations and recordings that could serve as a reference ground truth for future research in human and artificial musicianship. The project will also help in bringing deep learning and datasets with performances from underrepresented African-American musicians to the general public through workshops and high visibility concerts, aimed at capturing the interest and imagination of students who are not regularly drawn to the STEM fields. Lastly, the project will serve as a pivotal point for the recently approved BS program in Music Technology at the Georgia Tech Center for Music Technology, allowing undergraduate students to engage in creative interdisciplinary research.

A novel deep learning model for musical creativity

1 Introduction & Overview

Computational creativity is a research area that "enable[s] us to understand human creativity and to produce programs for creative people to use, where the software acts as a creative collaborator rather than a mere tool" [1]. Music, as a performing art, is an intriguing research subject as it entails both the generation of the underlying idea (composition) and its rendition into an acoustic signal (performance). More specifically, music improvisation combines these two aspects in a real-time context, as the performer, in an ad hoc scenario, conceives of the pitches and rhythms to play (composition) and decides how to perform them while varying timing, dynamics, and other parameters (performance). Two main computational approaches have been explored so far to study, model, and generate musical improvisations: (i) rule-based systems, for which musicological analysis of human improvisation is used to inform rule-based improvisational or performance generation algorithms (see, for example, [2, 3]), and (ii) data-driven probabilistic systems, where datasets of transcribed improvisations are used to train models such as Hidden Markov Models (HMMs) for the generation of improvisations based on probabilistic models (see, for example, [4, 5, 6, 7]). Both of these approaches, however, have fallen short in producing human-like, rich and expressive improvisations that could pass the Turing test [8] or lead to new creative improvisations that resonate with human listeners [9].

We hypothesize that the failure of current systems to generate high quality, rich, and expressive musical improvisations is due to the following reasons:

(i) Existing generative systems are not based on *perceptual* musical attributes. Experts tend to disagree about such attributes as multiple perspectives from music theory and perception research have never been reconciled to inform generative improvisations.

(ii) No current generative system combines compositional modeling (sequences of pitches, note lengths, etc.) with performance modeling (micro-timing deviation, loudness variation, pitch intonation, etc.). This leads to stale "robotic" generative improvisations that cannot project the richness of human creativity and expression.

(iii) Previous improvisational systems have only used short term analysis of human improvisations, either on the note-to-note level (most probabilistic approaches) or the phrase level (most rule based approaches). Improvisations generated by such systems, therefore, fail to produce long structural musical arches, which are crucial for creating meaningful and aesthetically pleasing musical forms leading to an engaged listening experience.

The goal of the proposed work (see overview in Fig. 1), is to dramatically improve the quality of generative computational improvisation systems by basing our system on the identification and annotation of fundamental musical perceptual dimensions, combining symbolic "compositional" analysis with audio-based performance analysis, and by using deep learning networks to infer musical phrase boundaries, hierarchies, and forms.

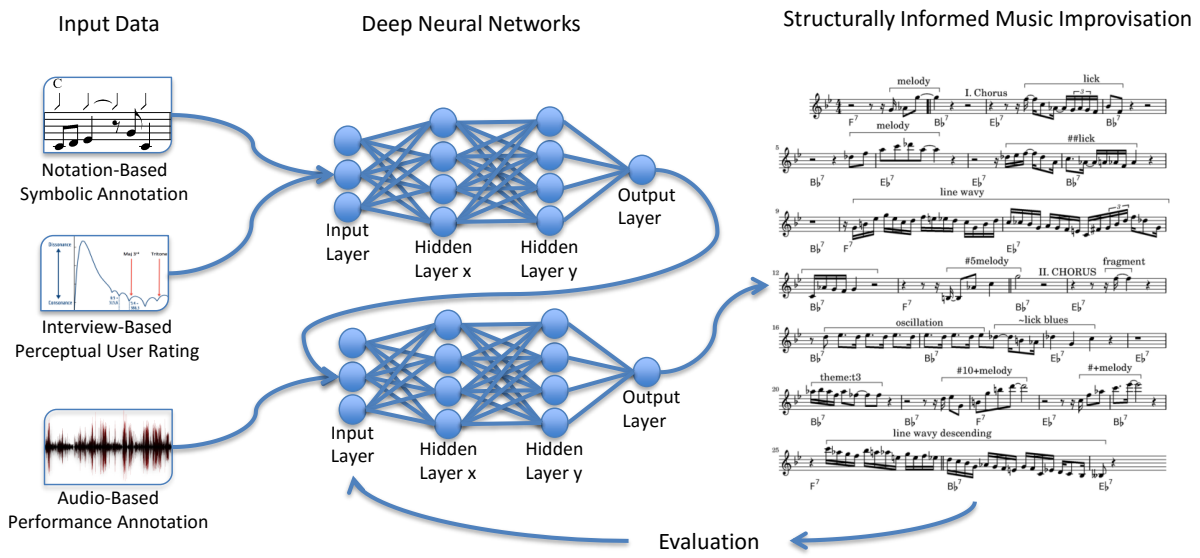

Figure 1: Overview of proposed work (architecture (ii) in Sect. 7.3.2)

2 Research Questions

Addressing the shortfalls of current efforts in this field, we have formulated three main research questions which motivate our proposed work, and could lead to a transformational paradigm shift in our understanding of real-time musical creativity:

(i) What are the perceptual attributes that experts use to characterize jazz solos? Could computational modeling of these attributes help generate meaningful and aesthetically pleasing musical improvisations?

(ii) Can the combination of performance parameters and score-based parameters, extracted from human improvisations, inform a generative model that will lead to high quality human-like musical improvisations and a sublime listening experience?

(iii) Can a deep neural network architecture that integrates high level score and performance data be successfully used to create long form rich musical improvisations? What insights into music improvisation can we gain by such a model?

3 Proposed Contributions

In pursuing these research questions, our proposed novel paradigm for computational creativity will lead to the following contributions:

(i) *Identification of the most salient perceptual dimensions* for the description of jazz solos through expert interviews. These attributes could aid music performers, teachers, and researchers as they describe how music is perceived.

(ii) *Annotation of seminal recordings* of jazz improvisations with both objective data (pitches, rhythm, chords, etc.) and subjective data (listener annotation according to the previously extracted perceptual dimensions). The annotations could give new insights into relationships between music structure and perceived responses, and reveal how improvisers manipulate musical features in real-time to communicate with others.

(iii) *Publication of the complete dataset* in order to foster future research on music improvisation. This dataset could be useful for research in multiple fields including musicology, perception and cognition, computational creativity, among others.

(iv) *Design of a generative system* based on Deep Neural Networks (DNNs) taking into account not only score-based and performance parameters, but also long-term musical phrasing relationships. This system will be the first to combine training data from symbolic (score-based) and performance (audio-based) analysis.

(v) *Evaluation of the generative system* using both subjective (listening tests) and objective (statistical analysis) methods.

(vi) *Generation of knowledge* and insight about the creation and perception of music based on the collected data as well as the output of the generative system.

4 Related Work

The related work for this interdisciplinary project can be divided into four main categories, the creation and perception of music, the automatic analysis of audio signals, the current advances in deep neural networks, and an overview of generative models in music.

4.1 Music Perception, Improvisation and Structure

Improvisers and composers manipulate elements of music to create and violate the expectations of the listener [10]. For example, a composition may repeat a motif multiple times only to violate expectations by altering the pitch or rhythm in the last iteration. Independent of training, listeners within a musical culture learn to decode expectations and violations much the same way they learn their native language [11]. Furthermore, it appears that expectation violations may elicit emotions in the listener [12, 13]. Therefore, listeners appear to prefer music in styles that contain a balance of predictability and novelty as related to the listeners' background [14].

Like composers, improvisers manipulate listeners' expectations on various hierarchical levels using musical elements such as pitch, timing, and dynamics. On a lower level, improvisers may repeat motifs or introduce tension by employing notes from outside the dominant tonality. On a higher level, improvisers describe organizing their entire solo around an architectural design. Such a design could use various elements such as register or note density to build intensity over several choruses [15]. Interestingly, improvisers' descriptions of their own thinking label these high-level designs as their primary concern during improvisation [16].

No previous work has asked expert improvisers to annotate existing canonical jazz solos according to perceived structure and emotional response. A large body of research has investigated links between emotional response and musical structure (for a review, see [17]). This research, however, was based on listeners' responses to western classical music and gathered from a limited number of responses per musical excerpt. With respect to jazz, there are only few examples of perceptual studies. Li and Ogihara investigated the emotion in jazz recordings, although not solos [18]. Recently, it has been proposed to record continuous responses to longer pieces; Schubert, for example, did such a study for classical music [19]. To the best of our knowledge, there is only one study in which continuous responses were recorded while listening to *two* jazz recordings for one predetermined dimension of tension and without a specific focus on improvisation [20]. In the proposed study, time-based annotations with data from four perceptual dimensions (see Sect. 7.1.3) will add a layer of information to each improvisation in a large collection of solos. Incorporating this information into the proposed machine-learning algorithm will allow the system to create output containing multi-level musical information previously absent in automated generation models.

4.2 Audio Content Analysis and Music Performance Analysis

While music or musical ideas are usually notated in a score or a lead sheet, there is much more to a music performance than the basic score-based parameters such as pitch and note length [21]. Even if two performers play the very same piece, the result may be perceived vastly different, depending on different choices for performance parameters such as tempo and timing, loudness and its variation, intonation and vibrato, and timbral properties [22]. The categories of temporal and loudness variations are usually considered to be the most important performance parameter categories [23]. These parameters have been shown to be important for emphasizing the listener's perception of, for example, musical phrases [24], groove [25], and structure [26]. On a higher level, these parameters convey emotion and mood [27, 28, 29]. These variations are inherent to the recording and can thus only be extracted from the audio signal. The research field of Music Information Retrieval (MIR) deals with the automatic extraction of information from the audio signal [30] and has produced approaches that are sufficiently reliable to serve as a starting point for the required audio analysis. Music Information Retrieval approaches enable performance researchers to analyze the recorded audio signal instead of being restricted to sensor-based recordings done in the lab as it traditionally used to be common [21].

In order to perform a detailed timing analysis, the exact onset times of the notes being played are of interest. The extraction of micro-timing in music performance analysis traditionally utilized sensors for accurate onset time data [31, 32] or manual annotation [33]; nowadays, more and more researchers use detection methods as described above to extract the onset times from the audio either automatically or semi-automatically with a human verifying the results [21]. Automatic onset detection is a well-known MIR task which is usually implemented as a 2-stage process: first, a *novelty function* is extracted that indicates the onset likelihood over time; and second, the onset positions are determined by using an adaptive threshold and peak picking [34]. Over the past few decades many methods for onset detection have been proposed [35, 36, 37, 38, 39], focusing mostly on various methods to compute the novelty function [30]. Although onset detection is often considered a "solved" problem, ongoing research shows that results can still be improved [40].

Musical dynamics have a high correlation with the perceived loudness, although other factors might have an impact as well [41]. There are different methods for estimating or modeling the perceptual loudness [42, 43], but in recent years the EBU norm R128, based on ITU recommendation BS.1770 [44] has been widely accepted as a simple algorithm with perceptual relevance [45] and has been used successfully in music performance analysis [46].

With respect to the analysis of audio recordings of jazz, one active German research group recently started to cover a range of research topics such as phrasing [47, 48], intonation [49], and dynamics [50]. Their research does not involve deep learning and is, utilizing traditional approaches, expected to be inferior in capturing the subtleties of human performance.

4.3 Deep Neural Networks and Semantic Embeddings

The successful application of Deep Neural Networks (DNNs) during the last decade to various problems make them one of the most promising approaches in the field of machine learning. DNNs have been successfully applied to artistic tasks [51] as well as for tasks with similar properties as music due to their inherent sequential and hierarchical nature such as natural language and speech processing [52, 53, 54, 55]. More generally speaking, deep learning has been shown to be effective in projecting a high dimensional sparse feature space into a low-dimensional semantic space [56, 57, 58]. Long Short Term Memory (LSTM) networks [59] have proven successful in modeling temporal dependencies in sequential data such as text in Natural Language Processing (NLP)

[60, 61, 54]. Palangi et al. described a method to sequentially pass each word in a sentence through an LSTM, while the resulting hidden layer of the network provides a semantic representation of the whole sentence [62]. The so-called Deep Structured Semantic Model, described in [63], allows for a supervised projection of low-level features into a common higher-level semantic space. It is particularly useful for ranking tasks as the relevance can be computed as the distance between a given input and a candidate within that space. Recently, Wen et al. described a methodology to semantically condition an LSTM in [64]. In this work, a single LSTM network is learned such that the LSTM cell jointly optimizes for the sentence mapping and structure realization components of natural language generation. They show that the semantics improve the model on objective and subjective qualities including informativeness and naturalness. Similarly, a contextual-LSTM is presented by Ghosh et al. [65], which demonstrates that in an NLP-related task, including additional features at the topic-level (contextual features) can improve the model at the word and sentence levels.
The LSTM approach seems well suited to the task of jazz solo creation, given the similarities in sequential representation of language and music. Several frameworks are available for the work with deep network architectures such *TensorFlow* [66], *Theano* [67], *Torch* [68], and *Caffe* [69]. TensorFlow in particular seems to be well suited for our purpose as it has been used for *Magenta* — a library by Google designed for music learning and generation.[1]

4.4 Computer-Generated Music

The field of computer generated music can be roughly categorized into rule based systems [70, 71, 2] and data-driven (probabilistic) methods, which typically employ n-gram Markov models [72, 73, 74, 75, 76] or (recurrent) neural networks [77, 78, 79]. Other less popular categorization methodologies exist, compare [80]. Rule-based systems have been explored in detail by groups such as the KTH Speech, Music and Hearing Group.[2] They formalize musicological concepts and use empirical results to construct a complex system of weighted rules, implemented as computer instructions. While some of these systems produce compelling musical results, they tend to suffer from subjective interpretation by the designer, and might still lack the subtleties and richness of music composed and performed by human musicians. Other rule-based approaches, such as Recombinant Music by David Cope [81], suffer from lack of clarity regarding the role of automation vs. the role of human composer in the generation process. On the other hand, data-driven models such as Markov-based approaches generally reduce the model complexity to simple note-to-note transitions (typically bi-gram or tri-gram note models) and therefore fail to address the higher level structure and semantics that is at the core of the temporal art form music. One attempt to address this issue was the definition and training of a general grammar for jazz solos [82]. While still not a common approach, modeling jazz improvisation with neural networks has already been proposed two decades ago [83]. More recent non-recursive generative approaches with neural networks include, for instance, melody modeling with a Time-Convolutional Restricted Boltzmann Machine [84] or Deep Belief Networks (DBNs) [85]. Recursive approaches such as Recurrent Neural Networks (RNNs) seem to be a better general fit to the modeling of sequential data. For instance, an RNN-RBM has been successfully used to model sequential information from a piano-roll-like representation [78]. Basic RNNs, however, model mostly note-to-note transitions and fail to capture higher level semantics and long term dependencies, similar to Markov-based methods. Using a Long Short Term Memory (LSTM), on the other hand, can address this problem by allowing higher level temporal structure to be learned [86] and produce encouraging preliminary results [87].

[1] `https://magenta.tensorflow.org/welcome-to-magenta`, accessed: Sep 8, 2016
[2] `http://www.speech.kth.se/music/performance/performance_rules.html`, accessed: Sep 29, 2016

Research has also been conducted on computer models for performance analysis and generation addressing elements such dynamics, micro timing, articulation and note ornamentation. Scholars have explored both rule-based approaches [88] and statistical approaches [89] to address this challenge, leading to frequently convincing results. As of yet, however, there has been no research attempting to utilize machine learning in general (and deep learning in particular) to analyze and generate both symbolic and performance musical parameters. We believe that this approach bears the promise of leading to hitherto unmatched creative and expressive musical results.

5 Intellectual Merit

The project will advance our current understanding of musical creativity — both in generation and perception — by creating four main intellectual contributions:

(i) For the first time, both symbolic musical structure and expressive performance elements will be analyzed and used for musical generation.

(ii) Novel hierarchical architectures for recurrent neural networks will be designed to capture higher-level semantics of transcribed and recorded jazz improvisations.

(iii) New methods for addressing insufficient training data and high-dimensional input data will be developed and systematically evaluated.

(iv) Input from human musicians and listeners will be integrated into the deep learning model as constraints, which could provide new insights into the mostly uncharted territory of aesthetic, emotional, and perceptual aspects of listening to and creating music.

6 Broad Impact

Our proposed generative performance system would lead to broad impact on the field of computational music generation, by introducing a novel approach that combines both symbolic and performance parameters. Moreover, the recursive neural network we will develop could lead to broad impact in other domains that maintain structural hierarchies but where training data is limited. In particular, our approach would be useful for domains with hierarchical sequential data such as natural language modeling and the generation of speech. Insights that will be gained from our improvisational computational creativity model may lead to impact in other domains that include real-time decision making. The project will also lead to broad impact by creating a large annotated corpus of transcribed improvisations and recordings that could serve as a reference ground truth for future research in human and artificial musicianship. The project will also help in bringing deep learning and datasets from underrepresented African-American musicians to the general public through workshops and high visibility concerts, aimed at capturing the interest and imagination of students who are not regularly drawn to the STEM fields. Lastly, the project will serve as a pivotal point for the recently approved BS program in Music Technology at the Georgia Tech Center for Music Technology, allowing undergraduate students to engage in creative interdisciplinary research.

7 Research Plan

Our research plan is divided into the following main steps: (i) the creation and annotation of the dataset containing both symbolic and perceptual annotation and the analysis of correlations between symbolic and perceptual data, (ii) the extraction of performance parameters from the audio signals, (iii) the development of a generative model using deep neural networks that are informed by symbolic and audio analysis, and (iv) the evaluation of the previous steps.

7.1 Dataset Creation & Analysis

A major effort in this project is the creation (and publication) of a large, well-designed dataset of jazz improvisations comprising symbolic annotations by experts, perceptual data from listening tests, audio data, and bibliographical information for recording and artist.

7.1.1 Data Selection and Symbolic Annotation

For the dataset we will choose 600–1200 jazz solos which have been played by prominent jazz masters. The selected recordings will have to be (i) considered seminal (studied and performed today in academic jazz programs), (ii) played by monophonic instruments to allow for robust automated audio analysis, and (iii) documented in both high quality audio recording as well as score transcript. These constraints will lead to a range of recording years approximately from the 1920s to the late 20th century. Several hundred of the selected solos can probably be found in online databases such as the Weimar Jazz database [90]. However, many of these transcriptions may require additional annotation of chord progression and phrase structure. Expert consultants such as jazz instructors and musicians will be hired to provide these annotations as well as for transcribing solos without available score.
The annotated transcriptions will be archived as score-based annotation (e.g., in MusicXML), text data, and MIDI files. The audio recordings will be archived as uncompressed audio files, while each file will be linked to its corresponding annotation.

7.1.2 Expert Interviews

Interviews will be conducted with 5–15 specialists in the area of jazz performance to identify perceptual dimensions that represent essential characteristics of jazz improvisation. Specialists will include musicians, journalists, and musicologists. Musicians will be "artist-level," defined as having over 10 years of experience recording and performing jazz. Journalists must be "nationally recognized," defined as having a body of work published nationally related to jazz spanning at least ten years. Musicologists must be employed by a research university and have published articles related to jazz history and analysis in major research journals. The participants will be interviewed in separate hour-long sessions. Prior to each session, participants will be asked to select two favorite audio recordings from our collection of canonical jazz improvisations. During the interview, participants will listen to the selected solos in short segments while describing their reactions. These reactions could range from analytical descriptions to emotional evaluations. The comments may be specific to the segment or more general observations concerning the entire solo and the performing artist. Only comments related to the specific segment will be used in the subsequent analysis. Using responsive interviewing, the researcher may ask follow-up questions related to themes already introduced by the participants [91]. The only prior instruction to the participant will be to "describe defining characteristics to each segment of the solo that is played back." We plan to use the standard interview format in which one person interviews the participant in an isolated setting [91] because it is a format familiar to musicians and journalists and it only requires participants to commit to one study session. Other methodologies including concept mapping [92] and the Delphi technique [93] might also be considered though they require repeated engagement with participants.
After completion of all the interviews, each interview will be transcribed and analyzed for themes. The thematic analysis follows standard analysis procedure based on the Grounded Theory framework in which all themes are directly linked to data [94]. Initially, each interview will be analyzed separately by coding the most salient ideas in each sentence, phrase, or paragraph. Later, conceptual categories will emerge in an iterative process of code comparison [95]. Of particular interest are conceptual

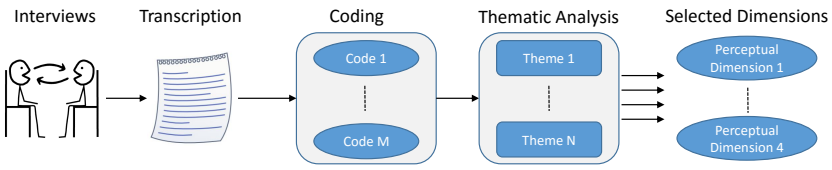

Figure 2: Identification of the perceptual dimensions from expert interviews

categories that describe features that change over time for the duration of the solo. We hypothesize that this will include structural features such as melodic contour, note density and high-level properties such as tension. These features were often mentioned in previous interviews with jazz experts concerning their own solos [16]. We also expect to identify categories related to the listener's emotional response. Central to this step is to find clearly definable terms that are used by a majority of the participants. Each term should represent a different perceptual dimension that signifies an essential feature of jazz improvisation that may be assigned a value descriptor. For example, one could say that the "level of tension" is high or low. The result will be the first systematically derived set of systematically derived perceptual categories unique to improvised jazz solos. The dimensions will be used to annotate the large corpus of solos introduced above. An overview over the complete process is shown in Fig. 2.

7.1.3 Perceptual Annotation

In order to annotate each solo with perceptual dimensions that were identified in the interview section, we will collect continuous self-report data from a large number of listeners through an online interface. For recruitment, advertising will appear in traditional and online media. We will collect information about subjects' musical background and demographics so that we can track and control for age, level of music proficiency, gender, and other variables. Continuous self-report has been used successfully to study emotion in music (for a review, see [96]). In recent research, computer interfaces have been used to collect data (e.g., [97]). Although in previous work some self-reported emotional responses to music have been collected through online tools [98], no online tools have been used to collect continuous responses to improvised music. In addition, prior research using continuous responses has mostly utilized predetermined dimensions such as valence and arousal [99]. Our proposed work would be the first in which domain specific dimensions are defined in expert interviews prior to collection of continuous response data as opposed to pre-defined dimensions such as valence and arousal.

A reduced set of four key perceptual dimensions will be used to collect continuous responses to the canonical jazz improvisations. This reduction is necessary to limit the listening time for the online test. The dimensions will be assigned to two pairs of orthogonal axes so participants can annotate four dimensions by listening to an audio file twice. According to Schubert [96], participants should be able to annotate two dimensions in a single listening session. An online interface will be created that has the following features:

- The participants will be asked to create a unique user-name to allow them to return to the study and listen to different selections at different times.
- After the initial sign in, the participants will complete required permissions, will be asked to fill in demographic and music background information, and will be given instructions.
- A recorded improvisation will be randomly selected.
- Participants will be asked to listen to and annotate the entire selection, in order for the data to be recorded and used.

- Before each listening, participants will be assigned two pairs of the perceptual dimensions.
- While they listen, subjects will manipulate an on-screen two-dimensional space where left/right and up/down represent values of each dimension, respectively. An example of such a screen interface where the dimensions are valence (horizontal) and arousal (vertical) was used in [97].
- The values will represent participant's perceived continuous coordinates in the two-dimensional parameter space.

Prior to statistical analysis, the data stream will be down-sampled equivalent to one value per notated bar or annotated phrase and outliers will be removed. In addition, the first annotation by each subject will be discarded as it will be considered training for the use of the interface. For a full review of technical issues and statistical analysis related to continuous self-report, see [96]. Our goal is to have each jazz improvisation annotated approximately 30 times for each dimension pair.

7.1.4 Annotation Analysis

The transcribed improvisations in combination with the annotated data provide a wealth of information that will enable us to gain insights about the perception of jazz improvisation by investigating the relationship between perceptual, high level parameters and descriptors derived from the score and the symbolic annotations. The results of the analysis will help closing the semantic gap between low level features (technical descriptions extracted from score and audio) and high level descriptions (human descriptions and characterizations of music). More specifically, descriptors such as pitch distribution, rhythmic density and variation, the relation of solo pitch to underlying chord, and the variation of the loudness, among others, could be extracted and related to listening test results for the perceptual dimensions. Standard approaches for statistical analysis will be used to identify the relation between the extracted descriptors and the annotations, including regression and ANOVA analysis.

7.2 Extraction of Performance Parameters from Audio

The performance parameters related to micro-timing and loudness will be extracted directly from audio recordings, and will be used to inform our generative model. For this purpose, we will devise new approaches and adapt established approaches from the field of Music Information Retrieval (MIR).

7.2.1 Pre-processing

The main improvising instrument is expected to be salient and dominant in the mix, however, other instruments in the background will impact the reliability of results negatively. Therefore, a pre-processing step is required in order to avoid inaccurate and inconsistent descriptors. For this purpose, we will use state-of-the-art algorithms for separating the dominant solo instrument from the background mix. Such algorithms detect the presence and frequency content of the dominant voice and then apply a time-varying spectral mask to subtract the target signal [100, 101, 102]. Although the resulting overall quality is often relatively poor [103], it will be sufficient for our purpose as we only extract a reduced feature description from the pre-processed audio signal.

7.2.2 Extraction of Onset Times

The challenge we face in extracting micro-timing is that a high accuracy of onset detection is required for meaningful time analysis. The time spans we are interested in (5–80 ms, [104]) are in the typical accuracy range of current onset detection systems [105, 38]. Fortunately, the available symbolic annotation of the improvisation in a score format will allow us to extend the presented 'blind' approach with additional domain knowledge: the number and approximate location of onsets. This could allow for to two possible extraction approaches —which might also be combined— for increasing onset detection reliability: (i) create an additional likelihood weighting based on onset locations from the transcription and perform peak picking on a combined novelty function, or (ii) apply an audio-to-score alignment algorithm that finds an optimal overall solution for the alignment of the two sequences (audio and score) [106, 107, 108, 109, 110, 111, 112]. The deviation of the onset time from the 'score' time can then be computed through a simple difference.

7.2.3 Extraction of Loudness

The extraction of the loudness as a representative feature for dynamics in music is well known and is frequently performed in the fields acoustics and audio processing. Existing implementations (e.g., [113, 114]) will be used for the feature extraction.

7.2.4 Feature Aggregation

The descriptors mentioned above will be extracted on different time scales. The onset time deviations will be represented per note, while the loudness will be extracted in equidistant windows. It is, however, straight-forward to compute the loudness per note as the EBU recommendation provides the procedure for computing an average loudness from each window in order to compute the loudness per note or per measure.

7.3 Generative Model

Jazz improvisation combines the generation of musical ideas on the fly with their immediate acoustic rendition in performance. The combination of score-based and performance-based parameters is at the core of the musical experience, however, it is not addressed by current generative models, which focus either on score OR performance generation. A generative model able to address both score and performance data could enable the creation of musically rich and meaningful improvisations, which would expand the current boundaries of computational musical creativity. Moreover, a significant shortcoming of existing models is that they either use probabilistic models, which tend to become too computationally complex when taking into account long-term relationships spanning more than simple note-to-note transitions, or that they are based on expert-defined rules with definitions and weightings that are subjectively, and often arbitrarily, defined. Hitherto, a deep hierarchical model that combines note, phrase, and performance parameters has not been proposed, particularly not in the context of deep architectures. We believe that two main challenges have prevented researchers from successfully addressing this task: (i) insufficient amount of training data and (ii) high dimensionality of the parameter space. We are addressing these challenges using various solutions including automatic data generation and augmentation, the addition of perceptual constraints, and a novel design of deep learning architectures as outlined below. Furthermore, using regularization as outlined in Sect. 7.3.3 we attempt to break with the paradigm of treating the network as a black box and try to ensure the perceptual interpretability of the trained embedded representation.

7.3.1 Step 1: Baseline System

The first step in a chain of increasingly complex sub-tasks towards the complete generative model is the creation of a comparably simple baseline system for the generation of single-voiced jazz solos. A recurrent neural network, more specifically a LSTM with One Hot encoding of the input "states" will be used, which means only one state can be active at each time. Each state in the model represents an individual note that is defined by its pitch and duration. The model will then be trained to predict the next note given a sequence of preceding notes. It is important to note that a jazz solo exists only in the harmonic context provided by the other instruments or through an established harmony progression. For this baseline system, we will incorporate the harmonic context by normalizing the pitch values to the current tonal center, normalizing the length values to the current tempo, and weighting the output with chord-note probability distributions derived from the training data.

We will look into ways of augmenting our data (shifting the pitch, changing the tempo, etc.) in order to address potential problems with insufficient training data. A further reduction of pitch values, possibly even an octave-independent representation of pitch will also be explored as additional potential solutions to this challenge.

7.3.2 Step 2: Enhanced Model with Performance Parameters

In the second step, the number of input and output states will be extended by adding loudness and time deviation information. Since the number of possible states grows dramatically, we might need to modify our baseline architecture. The solution candidates that will be explored include:

(i) embedded One-Hot Encoding: adding a pre-processing layer that maps the 4-dimensional input data (pitch, duration, loudness, time deviation) to a One-Hot Encoding that is used in the same way as in the baseline system,

(ii) separation of the symbolic model and the performance model (as shown in Fig. 1): the resulting two networks, one for the symbolic information and one for the performance data in a series to add timing and loudness to the generated symbolic stream of notes, and

(iii) encoding rhythm and time deviation implicitly rather than explicitly by splitting every beat into small equidistant sections (e.g., 128 sections); a beat-length note would then be represented by 128 neighboring instances.

The advantage of the first solution is that it would possibly take into account inter-dependencies between the input dimensions, while these would be separated in the second solution. The second alternative would be much simpler to implement and probably easier to train and control due to the reduced state-space. The same is true for the third option, which offers two additional advantages: integrated representation of rhythm and timing and a simple MIDI-like input (each segment would be represented by a MIDI tick). The disadvantage of the third option is that the system would not be able to differentiate between multiple fast note repetitions and one long note, similar to Google's Magenta.[3]

The insufficient amount of training data could turn out to be a challenge nevertheless as with many music-related deep learning tasks. One possible solution to this problem could be to use collections of MIDI data of performances, easily accessible online [115], as performance (pre-)training data. We will attempt to address this issue with negative sampling [116]. If the amount of training data given the model complexity remains to be a problem, the state space can be reduced by allowing fewer variations/steps in timing and dynamics. In case the high dimensionality of the output space impedes the model, a hierarchical softmax output layer [117] will be explored as a fitting solution

[3] `https://magenta.tensorflow.org/welcome-to-magenta`, accessed: Sep 8, 2016

for our model. The groupings in this hierarchical tree would also provide additional insight into interval and phrase probabilities in the data.

7.3.3 Step 3: Auto-encoding High-Level Perceptual Attributes

For this step, our goal is to use the perceptual annotations that were created in Sect. 7.1.3 to influence the embedded representation. We will constrain this representation so that it can become more interpretable and usable by applying domain knowledge from the annotated perceptual attributes. We hope to thus be able to control the generative model outcome by adjusting parameters in a perceptually relevant space. The setup we envision is an auto-encoder setup; an auto-encoder learns a compressed embedded representation with reduced dimensionality. Traditionally, the learned embedded representation in auto-encoders, which are trained to reconstruct their own input, is not interpretable by or meaningful to humans. By applying proper regularization with domain knowledge during the training, we intend to make this representation both more interpretable and controllable on the decoder side. More specifically, it will be enforced that, given a pair of human annotations, the distance between those annotations is similar to the distance between the embedded representations of the two improvisations. To give an example: we will compute a vector distance measure between each pair of perceptual annotations, and will modify the loss function to be small if the distance between the embedded representation is similar to the vector distance. More specifically, we will modify the network's loss function in a way that it includes a term approximating the distance between the current phrase and one or more randomly selected phrases, normalized by the sum of the distances to multiple other phrases. Thus, the embedded representation will be forced to span a perceptually meaningful parameter space (note that this applies only to distances, not absolute values). After successful training, the decoder could arguably be controlled in a perceptually meaningful way by adding distances to the embedded representation. This will allow a potential user to have access to high level musical control over the output. Input from evaluation results (see below) will be needed to adjust the network and to regularization parameters properly.

7.3.4 Step 4: Phrase-based Model

As mentioned above, the commonly used note-by-note generative models lack important context that would enable the generation of more convincing and musically meaningful results. Therefore, the model will be extended to a hierarchical model with the higher levels responsible for beat and phrase boundaries. This approach is inspired by recent approaches to word and sentence boundary detection in a natural language processing [118]. Our intention is to use an auto-encoder model that will hierarchically build an embedding for a phrase from embeddings for notes, and then will decode this embedding to reconstruct the phrase as shown in Fig. 3. This end-to-end strategy will allow the network to learn all hierarchical levels simultaneously, a general approach that has shown encouraging results in other contexts [119].
We expect the amount of available training data for phrase boundaries to be insufficient for training the network properly as there are per definition less data points. Therefore, we intend to explore a learning approach in which a manually designed expert system is used to generate training data. More specifically, the Generative Theory of Tonal Music (GTTM) defines frequently-used rules for grouping notes into phrases and higher structural levels [120]. We will utilize an implementation of this system (e.g., [121]) to automatically generate training data to pre-train the network. We propose that using training data generated by an expert system could be applied to other deep learning tasks with insufficient amounts of training data. We will therefore closely investigate

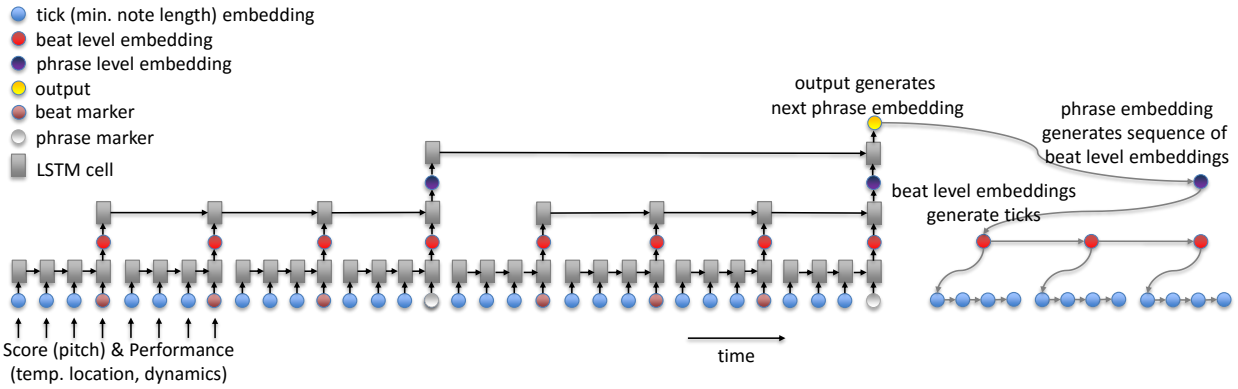

Figure 3: Hierarchical model for phrase boundary encoding (shown here with minimum note length input: architecture (iii) in Sect. 7.3.2)

the applicability of this approach to other scenarios. In this context, we will also investigate reinforcement learning [122] as a potentially valuable training strategy by letting subjects make a quick binary acceptable/unacceptable decision concerning the generated sequences.

7.4 Evaluation

7.4.1 Dataset Validation

Evaluation of the thematic analysis of the expert interviews will include member checks and triangulation [123]. Member checks will be conducted after the thematic analysis has revealed a limited number of perceptual dimensions. The expert interviewees will receive a short questionnaire that lists the dimensions and asks them to validate their applicability to jazz improvisation. They will also be asked to add additional dimensions and other information as appropriate. Triangulation will be done through comparison of the dimensions with standard vocabulary used in scholarly writings about jazz improvisation. The perceptual annotations' inter-rater reliability will be evaluated with statistical methods such as Krippendorff's alpha coefficient [124]. The subsequent data analysis will implicitly extract effects entered by multiple subjects.

7.4.2 Audio Processing

To evaluate our approach for audio processing we will use a number of online datasets,[4] which provide ground truth for the accuracy of onset tracking systems. The typical metrics used in such evaluations are F-Measure and standard deviation or average absolute error [105]. Our loudness measurement implementation will be verified with publicly available test signals provided by the EBU.[5]

A significantly more interesting and potentially transformational evaluation study would be to investigate the relationship between the extracted parameters with the high-level category ratings from the listeners. The procedure will be similar to the analysis proposed in Sect. 7.1.4. This evaluation would allow us to find whether and what kind of correlations exist between analyzed performance parameters and listeners' perception of the performance. To find such correlations

[4] `http://www.audiocontentanalysis.org/data-sets` accessed: Sep 30, 2016

[5] `https://tech.ebu.ch/publications/ebu_loudness_test_set` accessed: Oct 12, 2016

we plan to perform a statistical analysis of derived features, for example, features describing the distribution of the onset time deviations as input of a regression model to predict the assessments (compare, e.g., [19]). The metrics to measure the fit of the model would be standard metrics such as R^2, the correlation coefficient, and the standard error. We propose that this step could provide transformational insights into the relation between low-level performance characteristics and high-level perceptual characteristics.

7.4.3 Generative Model

The process of tuning a recursive neural network is complex and sometimes difficult to control, and an understanding of what is is actually learned (and not learned) is crucial for success. Visual analysis of the network will be used to gain a better understanding of hidden state dynamics with tools such as LSTMVis [125].

The assessment of the main goal of our proposed work — to generate more convincing, natural, and musically meaningful improvisations than current generative models — is inherently subjective. We, therefore, intend to undertake an extensive listening study as a major element of our evaluation. The first study will focus on expert listeners. Based on the RENCON tests for evaluating expressive music performance [126], experts will be asked to rate various categories such as naturalness and expression as well as overall quality on a Likert scale [127]. This will be a blind test with both human generated and randomly generated improvisations as hidden references and anchors, respectively. An alternative method would be to use a ranking test [128]. A second listening study will look specifically into how well the network models the perceptual dimensions by modifying the embedded representation. Listeners will then be asked to annotate the generated improvisations using the same process as in the perceptual annotations. Here too, we will use standard metrics such as R^2, the correlation coefficient, and the standard error to measure how well the output of the model fits the input.

In addition to the subjective studies, we will also use objective measures. For example, melodic distance measures between expected result and output would be of special interest in the case of the auto-encoder models. Established measures for melodic similarity include an edit distance approach counting, for instance, the number of note transformations for a match (compare [129, 130]), n-gram based distance measures that represent the melodies as pitch n-gram distributions and essentially count the number of matching n-grams [131]. We will also explore other evaluation methods including probabilistic approaches [132], geometric approaches [133], and network-based distance measures [134].

8 Project Timeline

Figure 4 shows the timeline for the proposed project, which closely follows the tasks outlined above. The first year of the project will be dedicated to dataset creation, including expert interviews, perceptual and symbolic annotation, audio analysis and annotation analysis. The next two years will focus on the four steps of the generative model, and the last year will be dedicated to evaluation, dissemination, and outreach.

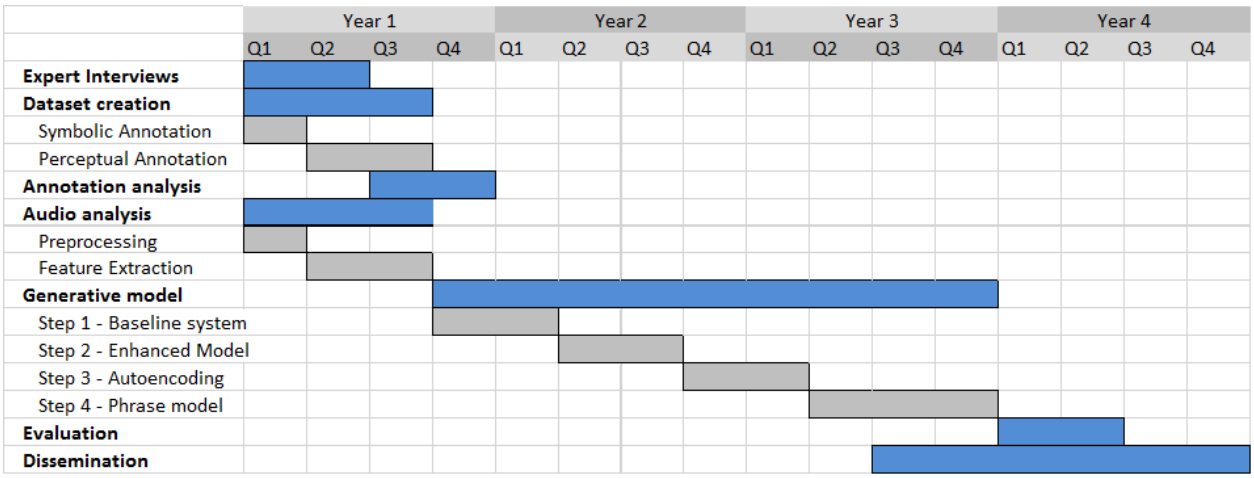

Figure 4: Proposed timeline for the described tasks

9 Dissemination & Outreach

We plan to disseminate the outcome of this project using our robotic musician platforms, which will utilize the proposed improvisation algorithms in concerts and demonstrations locally, nationally, and internationally. Some of the venues we have already performed with our robots include The Kennedy Performing Art Center, Atlanta Science Festival and US Science and Engineering Festival. We have strong relationships with these organizations and will continue to explore performance and workshop opportunities with these and other presenters. To extend the reach of the project, we will also develop an online demonstration that would allow musicians, researchers and the general public to listen to the generated improvisations and browse through our annotated data sets. Our detailed results will also be disseminated in conference proceedings (e.g., NIPS, AAAI, ISMIR) and journals (e.g., TASLP, JNMR).

In addition, we plan to conduct performances, workshops and demonstrations in high schools in the Atlanta area through a collaboration with the CEISMC program at Georgia Tech (see letter of support). We have also initiated collaborative efforts with institutions such as the Atlanta Symphony Orchestra Education Center and with elementary schools in the Atlanta area such as the Centennial Place Elementary School and the Galloway School. We will conduct workshops with our improvising robots and software designed to enhance scientific and musical literacy, involving students from groups that have been traditionally underrepresented in sciences and the arts. Design meetings with the directors of educational activities in these institutions have been conducted and further collaborations will be formed to develop and implement the pedagogy and curriculum. The workshops, which will be designed in collaboration with teachers in these organizations, will be based on Constructionist Learning approaches and designed to facilitate learning experiences by programming and interacting with our improvising robotic musicians.

The project will also provide research opportunities to incoming undergraduate students, who will join Georgia Tech's approx. 40 Music Technology students. The potential impact of this program to the US Southeast region can be substantial, as there are no other institutions in the region where undergraduate and graduate students can conduct research and receive an academic degree in an interdisciplinary field that combines music, computer science, and engineering. The proposed project will, therefore, serve as a focal point for students and researchers who will be actively engaged in computational music research and in dissemination of this work in the community.

References

[1] S. Colton, R. López de Mantaras, and O. Stock, "Computational Creativity: Coming of Age," *AI Magazine*, vol. 30, no. 3, pp. 11–14, 2009. [Online]. Available: https://digital.csic.es/handle/10261/31477

[2] A. Friberg, R. Bresin, and J. Sundberg, "Overview of the KTH rule system for musical performance," *Advances in Cognitive Psychology, Special Issue on Music Performance*, vol. 2, no. 2-3, pp. 145–161, 2006.

[3] J. Gillick, K. Tang, and R. M. Keller, "Machine learning of jazz grammars," *Computer Music Journal*, vol. 34, no. 3, pp. 56–66, 2010.

[4] D. M. Franz, "Markov Chains as Tools for Jazz Improvisation Analysis," thesis, Virginia Tech, Apr. 1998. [Online]. Available: https://vtechworks.lib.vt.edu/handle/10919/36831

[5] R. Nikolaidis and G. Weinberg, "Playing with the masters: A model for improvisatory musical interaction between robots and humans," in *19th International Symposium in Robot and Human Interactive Communication*, Sep. 2010, pp. 712–717.

[6] M. Norgaard, J. Spencer, and M. Montiel, "Testing cognitive theories by creating a pattern-based probabilistic algorithm for melody and rhythm in jazz improvisation." *Psychomusicology: Music, Mind, and Brain*, vol. 23, no. 4, pp. 243–254, 2013. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/pmu0000018

[7] G. L. Ramalho, P.-Y. Rolland, and J.-G. Ganascia, "An Artificially Intelligent Jazz Performer," *Journal of New Music Research*, vol. 28, no. 2, pp. 105–129, jun 1999. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1076/jnmr.28.2.105.3120

[8] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950. [Online]. Available: http://www.jstor.org/stable/2251299

[9] S. Senturk, "Computational modeling of improvisation in Turkish folk music using Variable-Length Markov Models," Thesis, Georgia Institute of Technology, Aug. 2011. [Online]. Available: https://smartech.gatech.edu/handle/1853/42761

[10] D. Huron, *Sweet anticipation*. Cambridge, MA: The MIT Press, 2006.

[11] J. F. Hay, B. Pelucchi, K. G. Estes, and J. R. Saffran, "Linking sounds to meanings: Infant statistical learning in a natural language." *Cognitive psychology*, vol. 63, no. 2, pp. 93–106, jul 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21762650

[12] E. Brattico, B. Bogert, and T. Jacobsen, "Toward a neural chronometry for the aesthetic experience of music." *Frontiers in psychology*, vol. 4, no. May, pp. 1–21, jan 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3640187{\&}tool=pmcentrez{\&}rendertype=abstract

[13] V. N. Salimpoor, D. H. Zald, R. J. Zatorre, A. Dagher, and A. R. McIntosh, "Predictions and the brain: how musical sounds become rewarding," *Trends in Cognitive Sciences*, pp. 1–6, dec 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364661314002538

[14] M. T. Pearce and G. a. Wiggins, "Auditory expectation: the information dynamics of music perception and cognition." *Topics in cognitive science*, vol. 4, no. 4, pp. 625–52, oct 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22847872

[15] P. F. Berliner, *Thinking in jazz*. Chicago, IL: University Of Chicago Press, 1994.

[16] M. Norgaard, "Descriptions of improvisational thinking by artist-level jazz musicians," *Journal of Research in Music Education*, vol. 59, no. 2, pp. 109–127, jun 2011. [Online]. Available: http://jrm.sagepub.com/cgi/doi/10.1177/0022429411405669

[17] A. Gabrielsson and E. Lindstrom, "The role of structure in the musical expression of emotions," in *Handbook of music and emotion: Theory, research, applications*, P. N. Juslin and J. A. Sloboda, Eds. Oxford, England: Oxford University Press, 2010, pp. 367–400.

[18] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5. IEEE, 2004, pp. 705–708. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1327208

[19] E. Schubert, "Modeling Perceived Emotion With Continuous Musical Features," *Music Perception*, vol. 21, no. 4, pp. 561–585, 2004.

[20] W. E. Fredrickson and J. C. Coggiola, "A comparison of music majors' and nonmajors' perceptions of tension for two selections of jazz music," *Journal of Research in Music Education*, vol. 51, no. 3, pp. 259–270, 2003.

[21] A. Lerch, *Software-Based Extraction of Objective Parameters from Music Performances*. München: GRIN Verlag, 2009. [Online]. Available: http://opus.kobv.de/tuberlin/volltexte/2008/2067/pdf/lerch_alexander.pdf

[22] C. Palmer, "Music Performance," *Annual Review of Psychology*, vol. 48, pp. 115–138, 1997.

[23] S. Dixon, W. Goebl, and G. Widmer, "The Performance Worm: Real Time Visualisation of Expression based on Langner's Tempo Loudness Animation," in *Proceedings of the International Computer Music Conference (ICMC)*, Göteborg, Sep. 2002.

[24] B. H. Repp, "Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists." *Journal of the Acoustical Society of America (JASA)*, vol. 88, no. 2, pp. 622–641, Aug. 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2212286

[25] J. Frühauf, R. Kopiez, and F. Platz, "Music on the timing grid: The influence of microtiming on the perceived groove quality of a simple drum pattern performance," *Musicae Scientiae*, vol. 17, no. 2, pp. 246–260, Jun. 2013. [Online]. Available: http://msx.sagepub.com/content/17/2/246

[26] F. Shifres, "Three expressive strategies of expert performance of an excerpt by J.S. Bach," in *MikroPolyphonie*, vol. 17, 2001. [Online]. Available: http://www.mikropol.net

[27] P. N. Juslin, "Cue Utilization in Communication of Emotion in Music Performance: Relating Performance to Perception," *Journal of Experimental Psychology*, vol. 26, no. 6, pp. 1797–1813, 2000.

[28] J. Kantor-Martynuska, "Emotion-relevant characteristics of temperament and the perceived magnitude of tempo and loudness of music," in *Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC)*, Bologna, Aug. 2006.

[29] R. Timmers, M. Marolt, A. Camurri, and G. Volpe, "Listeners' emotional engagement with performances of a Scriabin étude: an explorative case study," *Psychology of Music*, vol. 34, no. 4, pp. 481–510, 2006.

[30] A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*. Hoboken: Wiley-IEEE Press, 2012.

[31] L. H. Shaffer, "Timing in Musical Performance," *Annals of the New York Academy of Sciences*, vol. 423, no. 1, pp. 420–428, 1984.

[32] W. Goebl, S. Dixon, G. D. Poli, A. Friberg, R. Bresin, and G. Widmer, "Sense in Expressive Music Performance: Data Acquisition, Computational Studies, and Models," in *Sound to Sense, Sense to Sound: A State-of-the-Art*, M. Leman and D. Cirotteau, Eds. Logos Berlin, Nov. 2005.

[33] B. H. Repp, "A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 2, pp. 1085–1100, 1998.

[34] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1495485

[35] W. A. Schloss, "On the Automatic Transcription of Percussive Music – From Acoustic Signal to High-Level Analysis," Dissertation, Stanford University, Center for Computer Research in Music and Acoustics (CCRMA), Stanford, 1985.

[36] C. Duxbury, M. B. Sandler, and M. Davies, "A Hybrid Approach to Musical Note Onset Detection," in *Proceedings Of The 5th International Conference On Digital Audio Effects*, Hamburg, 2002.

[37] A. Röbel, "Onset Detection in Polyphonic Signals by means of Transient Peak Classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2005. [Online]. Available: http://www.music-ir.org/evaluation/mirex-results/articles/all/roebel.pdf

[38] S. Böck, F. Krebs, and M. Schedl, "Evaluating the Online Capabilities of Onset Detection Methods," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 49–54. [Online]. Available: http://ismir2012.ismir.net/event/papers/049-ismir-2012.pdf/at_download/file

[39] C.-Y. Liang, L. Su, and Y.-H. Yang, "Musical Onset Detection Using Constrained Linear Reconstruction," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2142–2146, Nov. 2015.

[40] J. Schlüter and S. Böck, "Improved musical onset detection with Convolutional Neural Networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6979–6983.

[41] T. Nakamura, "The communication of dynamics between musicians and listeners through musical performance," *Perception & Psychophysics*, vol. 41, no. 6, pp. 525–533, 1987.

[42] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness and Partial Loudness," *Journal of the Audio Engineering Society (JAES)*, vol. 45, 1997.

[43] G. A. Soulodre and S. G. Norcross, "Objective Measures of Loudness," in *Proceedings of the 115th Audio Engineering Society Convention (Preprint No. 5896)*. New York: Audio Engineering Society, 2003.

[44] I.-R. BS.1770:2006, "Algorithms to measure audio programme loudness and true-peak audio level," ITU, Recommendation, 2006.

[45] E. Skovenborg and S. H. Nielsen, "Evaluation of Different Loudness Models with Music and Speech Material," in *Proceedings of the 117th Audio Engineering Society Convention (Preprint No. 6234)*. San Francisco: Audio Engineering Society, 2004.

[46] Z. Schärer Kalkandjiev and S. Weinzierl, "The Influence of Room Acoustics on Solo Music Performance: An Empirical Case Study," *Acta Acustica united with Acustica*, vol. 99, no. 3, pp. 433–441, May 2013.

[47] K. Frieler, W.-G. Zaddach, and J. Abeßer, "Exploring Phrase Form Structures. Part II: Monophonic Jazz Solos," in *Proceedings of the International Workshop on Folk Music Analysis (FMA)*, Istanbul, 2014.

[48] K. Frieler, M. Pfleiderer, W.-G. Zaddach, and J. Abesser, "Midlevel analysis of monophonic jazz solos: A new approach to the study of improvisation," *Musicae Scientiae*, vol. 20, no. 2, pp. 143–162, Jun. 2016. [Online]. Available: http://msx.sagepub.com/lookup/doi/10.1177/1029864916636440

[49] J. Abeßer, E. Cano, K. Frieler, and W.-G. Zaddach, "Score-Informed Analysis of Intonation and Pitch Modulation in Jazz Solos," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015. [Online]. Available: https://www.researchgate.net/publication/284019684_Score-Informed_Analysis_of_Intonation_and_Pitch_Modulation_in_Jazz_Solos

[50] J. Abeßer, E. Cano, K. Frieler, and M. Pfleiderer, "Dynamics in Jazz Improvisation - Score-informed Estimation and Contextual Analysis of Tone Intensities in Trumpet and Saxophone Solos," in *9th Conference on Interdisciplinary Musicology (CIM14)*, Berlin, 2014. [Online]. Available: https://www.researchgate.net/publication/267584396_Dynamics_in_Jazz_Improvisation_-_Score-informed_Estimation_and_Contextual_Analysis_of_Tone_Intensities_in_Trumpet_and_Saxophone_Solos

[51] L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv:1508.06576 [cs, q-bio]*, Aug. 2015, arXiv: 1508.06576. [Online]. Available: http://arxiv.org/abs/1508.06576

[52] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390177

[53] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of Deep Belief Networks for Natural Language Understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 778–784, Apr. 2014.

[54] M. Sundermeyer, H. Ney, and R. Schlüter, "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 3, pp. 517–529, Mar. 2015. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2015.2400218

[55] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[56] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "DeViSE: A Deep Visual-Semantic Embedding Model," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2121–2129. [Online]. Available: http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf

[57] A. Karpathy, A. Joulin, and F. F. F. Li, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1889–1897. [Online]. Available: http://papers.nips.cc/paper/5281-deep-fragment-embeddings-for-bidirectional-image-sentence-mapping.pdf

[58] H. Wu, M. R. Min, and B. Bai, "Deep Semantic Embedding." in *SMIR@ SIGIR*, 2014, pp. 46–52. [Online]. Available: https://www.cl.uni-heidelberg.de/courses/ws14/deepl/WuETAL14.pdf

[59] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[60] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000. [Online]. Available: http://dx.doi.org/10.1162/089976600300015015

[61] A. Graves, "Generating Sequences With Recurrent Neural Networks," *arXiv:1308.0850 [cs]*, Aug. 2013, arXiv: 1308.0850. [Online]. Available: http://arxiv.org/abs/1308.0850

[62] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 694–707, Apr. 2016.

[63] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 2333–2338. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505665

[64] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems," *arXiv:1508.01745 [cs]*, Aug. 2015, arXiv: 1508.01745. [Online]. Available: http://arxiv.org/abs/1508.01745

[65] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck, "Contextual LSTM (CLSTM) models for Large scale NLP tasks," *arXiv:1602.06291 [cs]*, Feb. 2016, arXiv: 1602.06291. [Online]. Available: http://arxiv.org/abs/1602.06291

[66] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *arXiv:1603.04467 [cs]*, Mar. 2016, arXiv: 1603.04467. [Online]. Available: http://arxiv.org/abs/1603.04467

[67] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio, "Theano: new features and speed improvements," *arXiv:1211.5590 [cs]*, Nov. 2012, arXiv: 1211.5590. [Online]. Available: http://arxiv.org/abs/1211.5590

[68] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011. [Online]. Available: http://infoscience.epfl.ch/record/192376/files/Collobert_NIPSWORKSHOP_2011.pdf

[69] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 675–678. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654889

[70] G. E. Lewis, "Interacting with latter-day musical automata," *Contemporary Music Review*, vol. 18, no. 3, pp. 99–112, Jan. 1999. [Online]. Available: http://dx.doi.org/10.1080/07494469900640381

[71] R. Rowe, *Machine Musicianship*. MIT Press, 2001, google-Books-ID: kSKC1QQarvwC.

[72] I. Simon, D. Morris, and S. Basu, "MySong: Automatic Accompaniment Generation for Vocal Melodies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 725–734. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357169

[73] F. Pachet and P. Roy, "Markov constraints: steerable generation of Markov sequences," *Constraints*, vol. 16, no. 2, pp. 148–172, Sep. 2010. [Online]. Available: http://link.springer.com/article/10.1007/s10601-010-9101-4

[74] P. Chordia, A. Sastry, and S. Şentürk, "Predictive Tabla Modelling Using Variable-length Markov and Hidden Markov Models," *Journal of New Music Research*, vol. 40, no. 2, pp. 105–118, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1080/09298215.2011.576318

[75] C.-i. Wang and S. Dubnov, "Guided Music Synthesis with Variable Markov Oracle," in *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, Sep. 2014. [Online]. Available: http://www.aaai.org/ocs/index.php/AIIDE/AIIDE14/paper/view/9053

[76] T. Collins, R. Laney, A. Willis, and P. H. Garthwaite, "Developing and evaluating computational models of musical style," *AI EDAM*, vol. 30, no. 1, pp. 16–43, Feb. 2016. [Online]. Available: https://www.cambridge.org/core/journals/ai-edam/article/developing-and-evaluating-computational-models-of-musical-style/2D13038AEC3BB894F1345C63F74F6CF4

[77] A. E. Coca, R. A. F. Romero, and L. Zhao, "Generation of composed musical structures through recurrent neural networks based on chaotic inspiration," in *The 2011 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2011, pp. 3220–3226.

[78] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription," *arXiv:1206.6392 [cs, stat]*, Jun. 2012, arXiv: 1206.6392. [Online]. Available: http://arxiv.org/abs/1206.6392

[79] K. Goel, R. Vohra, and J. K. Sahoo, "Polyphonic Music Generation by Modeling Temporal Dependencies Using a RNN-DBN," in *Artificial Neural Networks and Machine Learning – ICANN 2014*, ser. Lecture Notes in Computer Science, S. Wermter, C. Weber, W. Duch, T. Honkela, P. Koprinkova-Hristova, S. Magg, G. Palm, and A. E. P. Villa, Eds. Springer International Publishing, Sep. 2014, no. 8681, pp. 217–224, dOI: 10.1007/978-3-319-11179-7_28. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-11179-7_28

[80] G. Papadopoulos and G. Wiggins, "AI methods for algorithmic composition: A survey, a critical view and future prospects," in *AISB Symposium on Musical Creativity*. Edinburgh, UK, 1999, pp. 110–117. [Online]. Available: http://www.academia.edu/download/3433841/AI_Methods_for_Algorithmic_Composition_A_Survey_a_Critical_View_and_Future_Prospects.pdf

[81] D. Cope, "Recombinant music: using the computer to explore musical style," *Computer*, vol. 24, no. 7, pp. 22–28, Jul. 1991.

[82] R. M. Keller and D. R. Morrison, "A grammatical approach to automatic improvisation," in *Proceedings, Fourth Sound and Music Conference, Lefkada, Greece, July."Most of the soloists at Birdland had to wait for Parker's next record in order to find out what to play next. What will they do now*, 2007. [Online]. Available: http://www.smc-conference.net/smc07/SMC07%20Proceedings/SMC07%20Paper%2055.pdf

[83] P. Toiviainen, "Modeling the Target-Note Technique of Bebop-Style Jazz Improvisation: An Artificial Neural Network Approach," *Music Perception: An Interdisciplinary Journal*, vol. 12, no. 4, pp. 399–413, Jul. 1995. [Online]. Available: http://mp.ucpress.edu/content/12/4/399

[84] A. Spiliopoulou and A. Storkey, "Comparing Probabilistic Models for Melodic Sequences," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Springer Berlin Heidelberg, Sep. 2011, no. 6913, pp. 289–304, dOI: 10.1007/978-3-642-23808-6_19. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-23808-6_19

[85] G. Bickerman, S. Bosley, P. Swire, and R. Keller, "Learning to Create Jazz Melodies Using Deep Belief Nets," *All HMC Faculty Publications and Research*, Jan. 2010. [Online]. Available: http://scholarship.claremont.edu/hmc_fac_pub/643

[86] D. Eck and J. Schmidhuber, "Finding temporal structure in music: blues improvisation with LSTM recurrent networks," in *Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing, 2002*, 2002, pp. 747–756.

[87] J. A. Franklin, "Recurrent Neural Networks for Music Computation," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 321–338, Aug. 2006. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/ijoc.1050.0131

[88] A. Friberg, "Generative Rules for Music Performance: A Formal Description of a Rule System," *Computer Music Journal*, vol. 15, no. 2, pp. 56–71, 1991. [Online]. Available: http://www.jstor.org/stable/3680917

[89] T. H. Kim, S. Fukayama, T. Nishimoto, and S. Sagayama, "Statistical Approach to Automatic Expressive Rendition of Polyphonic Piano Music," in *Guide to Computing for Expressive Music Performance*, A. Kirke and E. R. Miranda, Eds. Springer London, 2013, pp. 145–179, dOI: 10.1007/978-1-4471-4123-5_6. [Online]. Available: http://link.springer.com/chapter/10.1007/978-1-4471-4123-5_6

[90] K. Frieler, J. Abeßer, W.-G. Zaddach, and M. Pfleiderer, "Introducing the Jazzomat Project and the Melo(S)py Library," in *Proceedings of the Third International Workshop on Folk Music Analysis (FMA2013)*, P. van Kranenburg, C. Anagnostopoulou, and A. Volk, Eds., Amsterdam, 2013. [Online]. Available: http://www.academia.edu/download/31583096/fma2013_proceedings.pdf

[91] H. J. Rubin and I. S. Rubin, *Qualitative interviewing: The art of hearing data*, 2nd ed. Thousand Oaks, California: Sage Publications, 2005.

[92] I. G. J. H. Wopereis, S. Stoyanov, P. a. Kirschner, and J. J. G. Van Merriënboer, "What makes a good musical improviser? An expert view on improvisational expertise." *Psychomusicology: Music, Mind, and Brain*, vol. 23, no. 4, pp. 222–235, 2013. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/pmu0000021

[93] C.-c. Hsu and B. Sandford, "The delphi technique: making sense of consensus," *Practical Assessment, Research & Evaluation*, vol. 12, no. 10, pp. 1–8, 2007.

[94] B. G. Glaser and A. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. New York, NY: Aldine de Gruyter, 1967.

[95] K. Henwood and N. Pidgeon, "Grounded theory in psychological research," in *Qualitative research in psychology: Expanding perspectives in methodology and design*, P. M. Camic, J. E. Rhodes, and L. Yardley, Eds. Washington, DC: American Psychological Association, 2003, pp. 131–155.

[96] E. Schubert, "Continuous self-report methods," in *Handbook of music and emotion: Theory, research, applications*, P. N. Juslin and J. A. Sloboda, Eds. Oxford, England: Oxford University Press, 2010, pp. 223–253. [Online]. Available: http://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:Continuous+self-report+Methods{\#}3

[97] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmuller, "EMuJoy : Software for continuous measurement," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[98] D. Griffiths, S. Cunningham, and J. Weinel, "A self-report study that gauges perceived and induced emotion with music," *2015 Internet Technologies and Applications, ITA 2015 - Proceedings of the 6th International Conference*, pp. 239–244, 2015.

[99] Y. E. Kim, E. M. Schmidt, and L. Emelle, "MoodSwings: A Collaborative Game for Music Mood Label Collection," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Philadelphia, 2008. [Online]. Available: http://dblp.uni-trier.de/rec/html/conf/ismir/KimSE08

[100] Z. Rafii and B. Pardo, "Music/Voice Separation Using the Similarity Matrix," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Porto, 2012. [Online]. Available: http://www.cs.northwestern.edu/~zra446/doc/Rafii-Pardo%20-%20Music-Voice%20Separation%20using%20the%20Similarity%20Matrix%20-%20ISMIR%202012.pdf

[101] I. Y. Jeong and K. Lee, "Vocal Separation from Monaural Music Using Temporal/Spectral Continuity and Sparsity Constraints," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1197–1200, Oct. 2014.

[102] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 574–578.

[103] U. Gupta, E. Moore II, and A. Lerch, "On the Perceptual Relevance of Objective Source Separation Measures for Singing Voice Separation," in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz: IEEE, 2015.

[104] A. Friberg and A. Sundströöm, "Swing Ratios and Ensemble Timing in Jazz Performance: Evidence for a Common Rhythmic Pattern," *Music Perception: An Interdisciplinary Journal*, vol. 19, no. 3, pp. 333–349, Mar. 2002. [Online]. Available: http://mp.ucpress.edu/content/19/3/333

[105] S. Dixon, "Onset Detection Revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFX)*, Montreal, Sep. 2006.

[106] M. Müller, F. Kurth, and T. Röder, "Towards an Efficient Algorithm for Automatic Score-to-Audio Synchronization," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, 2004.

[107] C. Raphael, "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, 2004.

[108] A. Cont, "Realtime Audio to Score Alignment for Polyphonic Music Instruments using Sparse Non-Negative Constraints and Hierarchical HMMs," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, Toulouse, 2006.

[109] H. Kaprykowsky and X. Rodet, "Globally Optimal Short-Time Dynamic Time Warping, Application to Score to Audio Alignment," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2006.

[110] B. Niedermayer, "Towards Audio to Score Alignment in the Symbolic Domain," in *Proceedings of the 6th Sound and Music Computing Conference (SMC)*, Porto, 2009.

[111] C. Joder, S. Essid, and G. Richard, "A Conditional Random Field Viewpoint of Symbolic Audio-to-Score Matching," in *Proceedings of the International Conference on Multimedia (MM)*, 2010.

[112] Z. Duan and B. Pardo, "A State Space Model for Online Polyphonic Audio-Score Alignment," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Prague: IEEE, 2011.

[113] J. Timoney, T. Lysaght, M. Schoenwiesner, and L. MacManus, "Implementing Loudness Models in Matlab," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Naples, Italy, Oct. 2004. [Online]. Available: http://eprints.maynoothuniversity.ie/4141/

[114] N. Jillings, D. Moffat, B. De Man, J. D. Reiss, and R. Stables, "Web Audio Evaluation Tool: A framework for subjective assessment of audio," in *Proceedings of the Web Audio Conference (WAC)*. Atlanta: Georgia Institute of Technology, Apr. 2016. [Online]. Available: https://smartech.gatech.edu/handle/1853/54595

[115] C. Raffel and D. P. Ellis, "Extracting Ground Truth Information from MIDI Files: A Midifesto," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York, 2016. [Online]. Available: http://colinraffel.com/publications/ismir2016extracting.pdf

[116] N. A. Smith and J. Eisner, "Contrastive Estimation: Training Log-linear Models on Unlabeled Data," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 354–362. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219884

[117] A. Mnih and G. E. Hinton, "A Scalable Hierarchical Distributed Language Model," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 1081–1088. [Online]. Available: http://papers.nips.cc/paper/3583-a-scalable-hierarchical-distributed-language-model.pdf

[118] J. Li, M.-T. Luong, and D. Jurafsky, "A Hierarchical Neural Autoencoder for Paragraphs and Documents," *arXiv:1506.01057 [cs]*, Jun. 2015, arXiv: 1506.01057. [Online]. Available: http://arxiv.org/abs/1506.01057

[119] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv:1609.03499 [cs]*, Sep. 2016, arXiv: 1609.03499. [Online]. Available: http://arxiv.org/abs/1609.03499

[120] F. Lerdahl and R. Jackendorf, *A Generative Theory of Tonal Music*. Cambridge: MIT Press, 1983.

[121] M. Hamanaka, K. Hirata, and S. Tojo, "Implementing "A Generative Theory of Tonal Music"," *Journal of New Music Research*, vol. 35, no. 4, pp. 249–277, Dec. 2006. [Online]. Available: http://dx.doi.org/10.1080/09298210701563238

[122] L. P. Kaelbling, M. K. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996. [Online]. Available: http://www.jair.org/papers/paper301.html

[123] K. Charmaz, *Constructing grounded theory*. London: Sage, 2006.

[124] K. Krippendorff, "Computing Krippendorff's Alpha-Reliability," *Departmental Papers (ASC)*, Jan. 2011. [Online]. Available: http://repository.upenn.edu/asc_papers/43

[125] H. Strobelt, S. Gehrmann, B. Huber, H. Pfister, and A. M. Rush, "Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks," *arXiv:1606.07461 [cs]*, Jun. 2016, arXiv: 1606.07461. [Online]. Available: http://arxiv.org/abs/1606.07461

[126] H. Katayose, M. Hashida, G. D. Poli, and K. Hirata, "On Evaluating Systems for Generating Expressive Music Performance: the Rencon Experience," *Journal of New Music Research*, vol. 41, no. 4, pp. 299–310, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1080/09298215.2012.745579

[127] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22 140, p. 55, 1932.

[128] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*. Chichester, England; Hoboken, NJ: John Wiley & Sons, 2006.

[129] A. Habrard, J. M. Iñesta, D. Rizo, and M. Sebban, "Melody Recognition with Learned Edit Distances," in *Structural, Syntactic, and Statistical Pattern Recognition*, ser. Lecture Notes in Computer Science, N. d. V. Lobo, T. Kasparis, F. Roli, J. T. Kwok, M. Georgiopoulos, G. C. Anagnostopoulos, and M. Loog, Eds. Springer Berlin Heidelberg, Dec. 2008, no. 5342, pp. 86–96, dOI: 10.1007/978-3-540-89689-0_13. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-89689-0_13

[130] D. Müllensiefen and K. Frieler, "Measuring melodic similarity: Human vs. algorithmic judgments," in *Proceedings of the Conference on Interdisciplinary Musicology (CIM04) Graz/Austria*, 2004, pp. 15–18. [Online]. Available: https://www.researchgate.net/profile/Daniel_Muellensiefen/publication/228876095_Measuring_melodic_similarity_Human_vs_algorithmic_judgments/links/00463519a48a2b54e1000000.pdf

[131] I. S. Suyoto and A. L. Uitdenbogerd, "Simple efficient n-gram indexing for effective melody retrieval," *Proceedings of the Annual Music Information Retrieval Evaluation exchange*, 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.9371&rep=rep1&type=pdf

[132] N. Hu, R. Dannenberg, and A. Lewis, "A Probabilistic Model of Melodic Similarity," in *Proceedings of the International Computer Music Conference (ICMC)*, Jan. 2002. [Online]. Available: http://repository.cmu.edu/compsci/509

[133] G. Aloupis, T. Fevens, S. Langerman, T. Matsui, A. Mesa, Y. Nunez, D. Rappaport, and G. T. Toussaint, "Algorithms for Computing Geometric Measures of Melodic Similarity," *Computer Music Journal*, vol. 30, no. 3, pp. 67–76, 2006. [Online]. Available: https://muse.jhu.edu/article/202595

[134] S. Ferretti, "On the Modeling of Musical Solos as Complex Networks," *Information Sciences*, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025516311550