An Intelligent Music System to Perform Different "Shapes of Jazz—To Come"

Jonas Braasch, Selmer Bringsjord, Nikhil Deshpande, Pauline Oliveros and Doug Van Nort

Abstract In this chapter, we describe an intelligent music system approach that utilizes a joint bottom-up/top-down structure. The bottom-up structure is purely signal driven and calculates pitch, loudness, and information rate among other parameters using auditory models that simulate the functions of different parts of the brain. The top-down structure builds on a logic-based reasoning system and an ontology that was developed to reflect rules in jazz practice. Two instances of the agent have been developed to perform traditional and free jazz, and it is shown that the same general structure can be used to improvise different styles of jazz.

1 Introduction

Automated musical agents have a long tradition in Artificial Intelligence (AI) research. Starting first as composition tools [11, 17, 31, 18], modern computers are sufficiently fast to allow computational systems to improvise music with other performers in real time. Typically music composition/improvisation systems use a symbolic language, most commonly in form of the Musical Instrument Digital Interface (MIDI) format. Successful systems such as Lewis's Voyager system [20] and Pachet's *Continuator* [25] use MIDI data to interact with an individual per-

J. Braasch (✉) · N. Deshpande
Graduate Program in Architectural Acoustics,
Rensselaer Polytechnic Institute, 110 8th Street, Troy 12180, NY, USA
e-mail: braasj@rpi.edu

S. Bringsjord
Department of Cognitive Science, Rensselaer Polytechnic Institute,
110 8th Street, Troy 12180, NY, USA

P. Oliveros
Arts Department, Rensselaer Polytechnic Institute,
110 8th Street, Troy 12180, NY, USA

D.V. Nort
Computational Arts and Music, York University, Toronto, Canada

A. Schneider (ed.), *Studies in Musical Acoustics and Psychoacoustics*,
Current Research in Systematic Musicology 4,
DOI 10.1007/978-3-319-47292-8_12

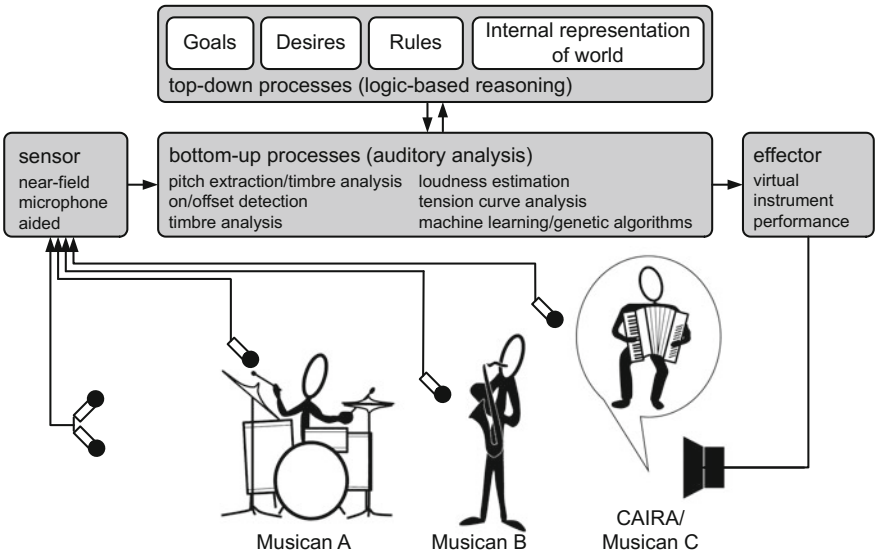

Fig. 1 Schematic of the creative artificially-intuitive and reasoning agent CAIRA

former whose sound is converted to MIDI using an audio-to-MIDI converter. The research described in this chapter stems from a larger project with the goal of developing a *Creative Artificially-Intuitive and Reasoning Agent* (CAIRA). Instead of using the simple audio-to-MIDI converter, the CAIRA uses standard techniques of Computational Auditory Scene Analysis (CASA) including pitch perception, tracking of rhythmical structures, and timbre and texture recognition (see Fig. 1). The CASA approach allows CAIRA to extract further parameters related to sonic textures and gestures in addition to traditional music parameters such as duration, pitch, and volume. This multi-level architecture enables CAIRA to process sound using bottom-up processes simulating intuitive listening and music performance skills as well as top-down processes in the form of logic-based reasoning. The low-level stages are characterized by a Hidden Markov Model (HMM) to recognize musical gestures and an evolutionary algorithm to create new material from memorized sound events. The evolutionary algorithm presents audio material processed from the input sound which the agent trains itself with during a given session, or from audio material that has been learned by the agent in a prior live session. The material is analyzed using the HMM machine listening tools and CASA modules, restructured through the evolutionary algorithms, and then presented in the context of what is being played live by the other musicians.

The logic-based reasoning system has been designed for CAIRA so it can "understand" basic concepts of music and use a hypothesis-driven approach to perform with other musicians (see top-down processes in Fig. 1). Including a logic-based reasoning system offers a significant number of benefits. The first goal is to see this multi-level approach lead to a more natural system response by trading off several

techniques; this makes the underlying processes less transparent to the human musicians without decreasing the overall responsiveness of the system. Secondly, the agent should be able to create new forms of music with the specific goal that the agent be able to develop its own concepts by expanding and breaking rules, and monitoring the outcome of these paradigm changes. Thirdly, we want to document the performance of the system—which is not easy to do—when the agent simulates intuitive listening in the context of Free Music. By adding a logic-based reasoning system, it is now possible to assess communication between the agent and human musicians by comparing the internal states of the agent and the human musicians.

This chapter focuses on the third goal for our logic-based reasoning stage. In particular, we describe a self-exploratory approach to test the performance of CAIRA within a trio ensemble. The approach, described in further detail below, is inspired by experimental ethnomusicology methods practiced by Arom [1] and others. A more detailed description of the lower- and higher-level CAIRA architecture and its ability to operate using the fundamental concepts of music ensemble interaction will precede this discussion.

1.1 Gestalt-Based Improvisation Model Based on Intuitive Listening

The artificially-intuitive listening and music performance processes of CAIRA are simulated using the *Freely Improvising, Learning and Transforming Evolutionary Recombination* (FILTER) system [28–30]. The FILTER system uses a Hidden Markov Model (HMM) for sonic gesture recognition, and it utilizes Genetic Algorithms (GA) for the creation of sonic material. In the first step, the system extracts spectral and temporal sound features on a continuous basis and tracks onsets and offsets from a filtered version of the signal. The analyzed cues are processed through a set of parallel Hidden Markov Model (HMM)-based gesture recognizers. The recognizer determines a vector of probabilities in relation to a dictionary of reference gestures. The vector analysis is used to determine parameters related to maximum likelihood and confidence, and the data is then used to set the crossover, fitness, mutation, and evolution rate of the genetic algorithm, which acts on the parameter output space [28].

1.2 Logic-Based Reasoning Driven World Model

One of the main goals of the CAIRA project was to understand how an artificially creative system can benefit from a joint bottom-up/top-down structure. CAIRA's knowledge-based system is described using first-order logic notation—for a detailed description of CAIRA's ontology see Braasch et al. [5]. For example, CAIRA knows that every musician has an associated time-varying dynamic level in

seven ascending values from *tacit* to *fortissimo*. The agent possesses some fundamental knowledge of music structure recognition based on jazz music practices. It knows what a solo is and understands that musicians take turns in playing solos while being accompanied by the remaining ensemble. The agent also has a set of beliefs. For example, it can be instructed to believe that every soloist should perform exactly one solo per piece.

One of the key analysis parameters for CAIRA is the estimation of the tension arc, which describes the currently perceived tension of an improvisation. In this context, the term 'arc' is derived from common practice of gradually increasing the tension until the climax of a performance is reached, and then gradually decreasing tension to end it. While tension often has the shape of an arc over time, it can also follow other trajectories. It is noteworthy that the focus here is not on tonal tension curves that are typically only a few bars long (i.e. demonstrating low tension whenever the tonal structure is resolved and the tonic appears). Instead, we are interested in longer structures, describing how a parameter relates to *Emotional Force* [22].

Using individual microphone signals, the agent tracks the running loudness of each distinct musical instrument using the Dynamic Loudness Model of Chalupper and Fastl [9]. The Dynamic Loudness Model is based on a fairly complex simulation of the auditory periphery that includes the simulation of auditory filters and masking effects. Additionally, the psychoacoustic parameters of roughness and sharpness are calculated according to Daniel and Weber [12] and Zwicker and Fastl [32]. In its current implementation, CAIRA estimates tension arcs for each musician from estimated psychophysical parameters. Based on these perceptual parameters and through its logic capabilities, the system recognizes different configurations for musical interplay. For example, it realizes that one of the musicians is performing an accompanied solo, by noticing that the performer is louder and has a denser texture than the remaining performers. The system can also notice that the tension arc is reaching a climax when all musicians perform denser ensemble textures. CAIRA takes action by either adapting its music performance to the analysis results or by presenting a dynamic visual score. CAIRA can, for example, suggest that a performer should end his or her solo because it is becoming too long, or it can encourage another musician to take more initiative. It can guide endings, and help an ensemble to fuse its sounds together.

Before we describe the mechanism to measure tension arcs, we briefly introduce the underlying basic concepts of jazz performance for two schools of jazz thought: traditional jazz, and free jazz.

2 Automated Music Improvisation Systems for Traditional Jazz

2.1 A Brief Overview on Traditional Jazz Practices

In this chapter, we use the term *traditional jazz* for jazz styles that precede the free jazz era—covering styles from swing to hardbop—but purposely exclude modal

jazz, which already contained numerous elements that later became characteristic features of free jazz. We will only cover the very basic fundamentals of jazz, but an extensive set of literature exists on this topic—for example Spitzer [27].

In traditional jazz, the freedom of an improviser is more constrained than one might think. Typically, each solo follows the chord progression of the song played by the rhythm section. The latter typically consists of drums, bass, and one or more chordal instruments—predominantly piano or guitar. For traditional reasons, one chord progression cycle is called a chorus.

The general repertoire of jazz tunes are called jazz standards. Most of these standards originated from Tin Pan Alley songs and pieces from Broadway musicals, in which jazz musicians performed for a living. After the theme is played the lead instruments take turns playing solos, and often players in the rhythm section take turns soloing as well. In traditional jazz, the performers are free to play over as many choruses as they want, but to end a solo before the end of the chord progression cycle is a taboo. The solo typically consists of a sequence of phrases that is chosen to match the chord progression and the intended dramaturgy. Since the two most common chord progressions in jazz are II-V and II-V-I (supertonic/dominant/tonic) combinations, professional jazz musicians train on phrases based on these progressions. Extensive literature exists with collections of standard jazz phrases.

Figure 2 shows the first eight bars of a notated saxophone solo over the 32-bar jazz standard, How High the Moon (Hamilton and Lewis 1940) to provide a practical example. Charlie Parker's Ornithology later used the same chord progression with a new bebop-style theme. Bars 3–6 consist of the typical II-V-I chord progression: G^{m7} (notes: G, B$^{\flat}$, D, F), C7 (C, E, G, B$^{\flat}$), F^{maj7} (F, A, C, E), and Bars 7 and 8 of another II-V progression Fm7 (F, A$^{\flat}$, C, E$^{\flat}$) and B$^{\flat 7}$ (B$^{\flat}$, D, F, A$^{\flat}$). Notice how in the example the saxophone initially follows the notes of the individual chords closely with additional scale-related notes—which is typical for swing. From Bar 6 on, the phrases change to bebop style with a faster eighth-note pattern. Also noteworthy is the second half of Bar 7, where the saxophone plays note material outside the chord related scale to create a dissonant effect. Whether this is appropriate depends on the agreed upon rules; In the swing era, this would

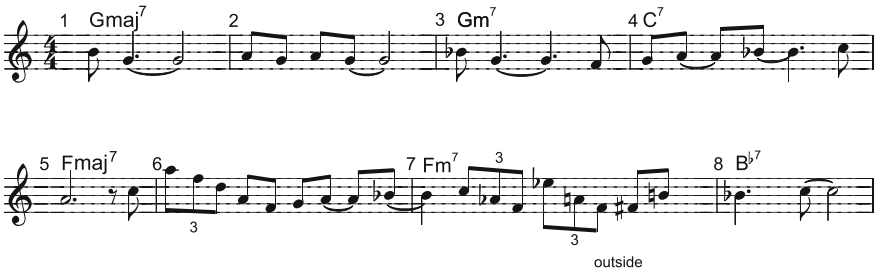

Fig. 2 Example transcription of a saxophone solo over the jazz standard How High the Moon (first 8 bars)—after Braasch [3]

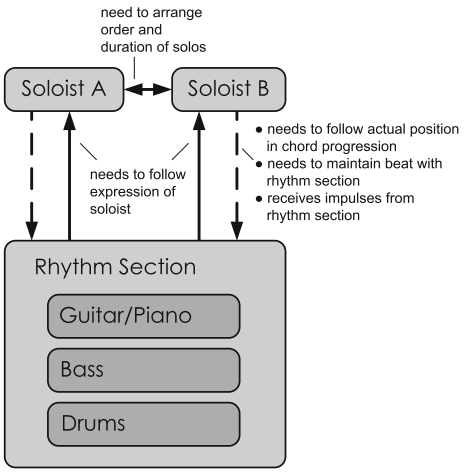

Fig. 3 Schematic communication scheme for a traditional jazz performance

have been played "incorrectly" but such techniques later became a characteristic style of players (such as Eric Dolphy) who could elegantly switch between so-called inside and outside play.

In order to play a correct solo following the rules of jazz theory, one could easily focus the attention to a very limited set of features to survive gracefully as shown in Fig. 3. Although, it should be noted that virtuoso jazz players are known to listen and respond to many details initiated by the other players. Basically, the soloist can process the rhythm section as a holistic entity, since all musicians follow the same chord progression. The tempo is quasi-stable, and the performance of the other soloist has to be observed only partially to make sure not to cut into someone else's solo. Once another soloist initiates a solo, he or she no longer needs to pay attention to the other soloists.

2.2 Rule-Based Machine Improvisation Algorithms

Numerous attempts have been made to design machine improvisation/composition algorithms to generate music material in the context of jazz and other styles [11, 17, 18, 31]. In most cases, these algorithms use a symbolic language to code various music parameters. The wide-spread MIDI (Musical Instrument Digital Interface) format, for example, codes the fundamental frequencies of sounds into numbers. Here, the note C_1 is the MIDI Number 24. Note numbers ascend in integers with the semitones. The temporal structure is also coded in numeral values related to a given rhythm and tempo structure.

By utilizing such a symbolic code, improvisation or composition can become a mathematical problem. Typically, the program selects phrases from a database according to their fit to a given chord progression (e.g., avoiding tones that are outside the musical scales for these chords, as previously discussed in context of

Fig. 2), and current position in the bar structure (e.g., the program would not play a phrase ending in the beginning of a chord structure). Under such a paradigm, the quality of the machine performance can be evaluated fairly easily by testing whether any rules were violated or not. Of course, such an approach will not necessarily lead to a meaningful performance, but the results are often in line with that of a professional musician. A system can even operate in real time as long as it has access to live music material on a symbolic level, for example, MIDI data from an electronic keyboard.

Lewis' Voyager system [20] and Pachet's Continuator [25] work using MIDI data to interact with an individual performer. The system transforms and enhances the performance of the human musician by generating new material from the received MIDI code, which can be derived from an acoustical sound source using an audio-to-MIDI converter; typically these systems fail if more than one musical instrument is included in the acoustic signal. In the case of the Continuator, learning algorithms are used based on a Hidden Markov Model help the system to copy the musical style of the human performer.

Commercial systems that can improvise jazz are also available. The program Band-in-a-Box™ is an intelligent automatic accompaniment program that simulates a rhythm section for solo music entertainers. The system also simulates jazz solos for various instruments for a given chord progression and popular music style. The system can either generate a MIDI score that can be auralized using a MIDI synthesizer, or create audio material by intelligently arranging prerecorded jazz phrases. The restricted framework of the jazz tradition makes this quite possible since the "listening" abilities of such a system can be limited to knowing the actual position within the form. Here the system needs to count along, making sure that it keeps pace with a quasi-steady beat.

3 Automated Music Improvisation Systems for Free Jazz

In contrast to traditional jazz, a formal set of rules does not exist in free jazz, although there has been a vivid tradition that has been carried on and expanded. Most of this tradition exists as tacit knowledge and is carried on in performance practice, orally and through musicological analyses. One example for tacit knowledge in free jazz is the taboo of performing traditional music material (see Jost [19]), unless it is a brief reference in the context of other adequate free music material. For the application of the informative feedback model to free jazz, it is also important to understand how the tradition progressed over time, deviating more and more from traditional jazz practice. A key moment for the development of free jazz was the introduction of modal jazz at the end of the 1950s, in which the chord progressions were replaced with fixed musical modes. In modal jazz the standard form of 12, 16 or 32 bars was initially kept, but this structure was given up in the favor of a free (variable) duration of form.

In the beginnings of free jazz, music material was fairly traditional and could be analyzed based on traditional music notation and thus easily captured using a symbolic music code like MIDI. As the field progressed musicians started to use extended techniques that shifted their performance more and more from the traditional sound production techniques of the orchestral instruments used in jazz. Albert Mangelsdorff's ability to perform multiphonics on the trombone is legendary, and so are the circular-breathed melodic streams of Evan Parker, who obtained the ability to perform arpeggio-style continuous phrases with a variable overtone structure containing both tonal and non-pitch-based elements. Peter Brötzmann's repertoire further expanded the techniques of non-pitched sounds. Among the younger generation of free jazz musicians are performers whose work focuses on complex musical textures outside the context of tonal music. Mazen Kerbaj (trumpet) and Christine Sehnaoui (saxophone) are among those who neglected the tonal heritage of their instruments in a unique way.

Initially, free jazz musicians took turns performing accompanied solos, but as time progressed it transformed into a genre where the boundaries between solos and accompaniment became blurred. While in traditional jazz a soloist has to listen to another soloist only to find a good slot for their own solo, instead performers began to pay attention all the time to other soloists. In addition, a soloist could no longer rely on the predetermined role of the rhythm section, which was now allowed to change keys, tempo and/or style. The higher cognitive load that was necessary to observe all other participants in a session led to smaller ensembles, often duos. Larger ensembles like the Willem Breuker Kollektief remained as the exception.

Figure 4 depicts a model of communication during a free jazz session. The diagram, shown here for a group of three musicians, appears to be much simpler because of the lack of rules. In contrast to the previous model for traditional jazz (Fig. 3), the distinction between rhythm section players and soloists is no longer made. While in traditional jazz the rhythm section can be represented as a holistic entity with homogeneous rhythm, tempo, and chord structure, now individual communication channels have to be built up between all musicians. Also, the feedback structure that each musician needs to enact to adequately respond to other players is fundamentally different from traditional jazz, where the communication

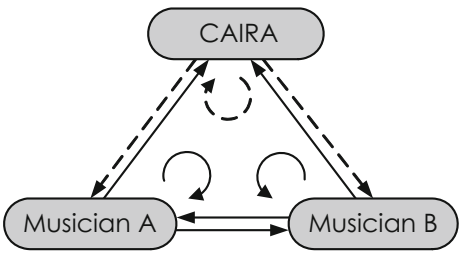

Fig. 4 Schematic of a communication scheme for a free jazz performance. The categorical distinction between soloists and rhythm section players no longer exists. Each musician has to establish individual communication channels to all other musicians, and also observe his or her own performance. The agents observation tasks are shown as *dashed arrows*

feedback loop (see Fig. 3) may simply cover a single communication stream from the ensemble (seen as a whole unit) to the soloist and back. In free music, a separate communication line has to be established between each possible pair of players, and consequently each performer has to divide his/her attention to observe all other players individually. Since the feedback from other musicians has to be detected with the ears, the multiple feedback-loop structure is not apparent in Fig. 1. However, the need to extract the information individually for each musician from a complex sound field is what makes free music improvisations a challenge. In addition, the performer always has to be prepared for unexpected changes, especially since tacit knowledge can be extended or modified within a session.

With regard to the music parameter space, for traditional jazz it is sufficient to receive the pitches of the notes played, to determine the current chord structure and melody lines, and to capture the onset and offset times of these notes to align the performance in time with the rhythm section. Commercial audio-to-MIDI converters can perform this task reliably enough if the general chord progression is known in advance. The analysis can even contain errors from information redundancy, as long as the algorithm can follow the given chord progression. In the context of an automated system that can improvise free music, machine listening demands are much higher if the system is mimicking human performance (see the Auditory Analysis box in Fig. 1). Here, it is no longer necessary to have a predetermined chord progression that serves as a general guideline. Even if there existed a system that could extract the individual notes from a complex chord cluster—which is difficult because of the complex overtone structure of the individual notes—it is not guaranteed that the musical parameter space in a session is based on traditional music notes.

To address this problem adequately, the intelligent system can be equipped with a complex model that simulates the auditory pathway. This type of model is able to extract features from acoustic signals in a similar way to the human brain (see Fig. 1). The early stages of the auditory pathway (auditory periphery, early auditory nuclei that perform spectral decomposition, pitch estimation, onset and offset detection) are thought to be purely signal driven, whereas the performance of the higher stages (e.g., timbre recognition, recognition of musical structure) are thought to be learned; these auditory features are categorized along learned patterns.

In the current CAIRA implementation, the features extracted in the Auditory Analysis stage are coded as symbolic information and passed on to the cognitive processing stage. From the symbolic information it receives, it can construct an internal representation of the world (in this case, the representation of the jazz performance). As outlined in the previous section, the art of mapping acoustic signals onto symbolic information is well defined through jazz theory for traditional jazz. Thus, if the system does not know and follows the given rules, it will be easily detected by other musicians and the audience.

In contrast, in free music, there is no longer a standardized symbolic representation of what is being played. Instead, to a greater degree, the music is defined by its overall sound. Consequently, the musicians will need to derive their own symbolic representation to classify what they have heard and experienced, and they

also need to define their own goals. For automated systems, the latter can be programmed using methods in second-order cybernetics (e.g., see Scott [26]). With regards to the symbolic music representation in humans, musicians typically draw from their own musical background, and significant differences can be found in musicians who primarily received classical music training compared to those who concentrated in jazz, or those that worked with sound textures rather than pitch and harmony. These differences extend to artists who learned in non-western music traditions. For example, if a traditionally trained musician hears a musical scale, they associate it with a scale that exists in their musical culture. This association works as long as the individual pitches of each note fall within a certain tolerance. Consequently, two people from two different cultural backgrounds could label the same scale differently, and thus operate in different symbolic worlds judging the same acoustic events. In free music, interactions between musicians of various cultural backgrounds are often anticipated, hoping that these types of collaborations will lead to new forms of music, and this precludes musicians falling into patterns. However, the communication will only work if the structures of different musical systems have enough overlap such that musicians can decipher a sufficient amount of features from other performing musicians into their own system. Furthermore, as performers, we have only indirect access to the listening ability of co-musicians through observing what they play, and in the case where something was not "perceived" correctly by others, we cannot measure their resulting response (musical action) along rules in free music, as these rules do not exist.

For cross-cultural music ensembles, examples exist where communication problems resulted from operating in different music systems. The late Yulius Golombeck once recalled when he was performing with the world music band Embryo, Charlie Mariano, and the Karnataka College of Percussion, there were certain complex Indian rhythms played by the Karnataka College of Percussion that the western trained musicians could not participate in because the rhythmical structure was too complicated to understand, despite the fact that all musicians had a tremendous experience with non-western music.[1]

While the complex communication structure in free music poses a real challenge for automated music systems, the lack of a standardized symbolic representation can be used to a system's advantage. Instead of mimicking the auditory system to extract musical features (Fig. 1), an alternative approach could be a robot-adequate design. The design could consider that as of today some parameters (e.g., complex chords) are impossible to extract in parallel for multiple musicians, especially in the presence of room reverberation. Instead, a music culture for machines can be developed that emphasizes the strengths of machines and circumvents their shortcomings. The latter is summarized in Table 1.

A directed focus on machine-adequate listening algorithms also encourages the design of machines to have their own identity, instead of focusing on making them indistinguishable from humans by passing the Turing test (e.g., compare Boden

[1]Braasch, personal communication, 1995.

Table 1 Listening strengths and weaknesses of machines compared to humans

Listening strengths	*Listening weakness*
• Absolute sense of time • Absolute sense of timbre • Absolute sense of pitch	• Difficulty to perceptually correct imperfections of other players • Difficulty to reconstruct missing information • Difficulty to extract information from multiple source and reverberant environments
Cognition strength	*Cognition weakness*
• Good at combinatorics • Absolute memory	• Has no understanding of aesthetics • Unable to develop new concepts • Cannot abstract ideas
Action strength	*Action weakness*
• Can play everything at any tempo • Redefines virtuosity	• Difficulty to adapt to other musicians • Difficulty to perform with musical expression

[2]). Man/machine communication can then be treated like a cross-cultural performance, where sufficient overlap between the various cultures is expected to allow meaningful communication. In such collaborations, the goal would not be to replace humans with machines, but to build systems that inspire human performers in a unique and creative way. A good example of machine inspired human music performance is the introduction of the drum machine, which encouraged a new generation of drummers around Dave Weckl in the 1980s to perform their instruments more accurately—almost in a machine-like style.

4 Implementation of CAIRA

In this section, we describe the different modules that were designed and implemented to operate the CAIRA agent. We first describe the bottom-up mechanisms, and then the top-down structures.

4.1 Bottom-Up Mechanisms

The bottom-up mechanisms are signal driven and include modules that simulate different functions of the auditory periphery including pitch detection [6], beat detection, loudness calculation and the calculation of tension curves [4, 8]. Further, the CAIRA system heavily uses machine learning algorithms—based on Hidden Markow Models (HMM) and Empirical Mode Decomposition to analyze sonic gestures based on different time scales. The machine learning algorithms, which are especially important for the Free Jazz Instantation of CAIRA, are subsumed in the FILTER structure and have been described thoroughly in peer-reviewed literature

[28–30]. We therefore focus on the description of the polyphonic pitch detection model and tension curve estimation in this chapter.

4.1.1 Polyphonic Pitch Perception Model

The polyphonic pitch model builds on a functional model of the auditory periphery and previous pitch perception models [10, 14, 21, 24]. In the first step, to simulate the behavior of the basilar membrane the signal is sent through a Gammatone filterbank with 128 bands to segregate sound into different auditory bands. Then, the signal frequency f_n in each band n is estimated using auto-correlation, measuring the delay τ_n between the main and the largest side peak:

$$f_n = \frac{1}{\tau_n}. \tag{1}$$

A novel aspect of the model is that both the frequency and pitch strength are measured in each frequency band, the latter calculated using the amplitude ratio a between the largest side peak and the main peak. Further, the deviation b between the estimated frequency f_n and the center of the frequency band $f_{c,n}$ is calculated. Next, all results are grouped into four categories:

1. $a > 0.90$, $b \leq 0.3$ octaves ('+' symbols)
2. $a > 0.90$, $b > 0.3$ octaves ('×' symbols)
3. $a \leq 0.90$, $b \leq 0.3$ octaves ('°' symbols)
4. $a \leq 0.90$, $b > 0.3$ octaves ('*' symbols)

The graphs on the left in Fig. 5 show the results of the pitch model for a 440-Hz sinusoid. The top graph shows the broadband autocorrelation function, the center graph the Fourier Transformation of the signal, and the bottom graph depicts the excitation of the auditory bands (solid black curve). For the curve, the energy in

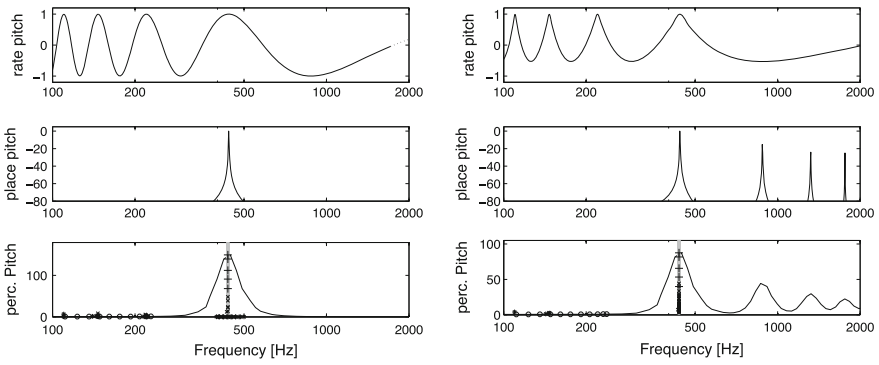

Fig. 5 Pitch estimation for a 440-Hz sinusoidal signal (*left graph*) and a 440-Hz tone complex (*right graph*)

each of the 128 bands was measured and plotted at the center frequency of the band. All values for the Group 1 are located at 440 Hz, the frequency of the sinusoid, as indicated by the gray curve. The height of the values represents the energy of the band in which the frequency was measured. The values for Group 2 also point to 440 Hz; they were measured in the adjacent side bands. All other values (Groups 3 and 4) were measured from the noise spectrum at low energies and do not represent the frequency of the sinusoid.

The right graphs of Fig. 5 depict the same context but this time for a tone complex with eight higher harmonics at integer multiples of the fundamental frequency: $f = n \cdot f_0$. The amplitude of the tone complex rolls off with $1/f$. Again, all values for Groups 1 and 2 ('+' and '×' symbols) point to the fundamental frequency of 440 Hz, even for those that belong to the higher harmonics.

Figure 6 shows the results for a $1/f$ tone complex at a lower fundamental frequency of 220 Hz (left graphs). Again, the results in all harmonics point to the fundamental, with the exception of two values in the octave region (440 Hz). It is not clear why in this case the octave is recognized; this will be further investigated. For higher harmonics, more than one overtone falls into the same auditory band. The overtones interfere with each other, and based on this interference the autocorrelation method identifies the common fundamental f_0. For the same reason, the algorithm is able to detect a missing fundamental. The right graphs show the results for the same tone complex, but this time the fundamental of 220 Hz was removed. Still, most values point to 220 Hz. Clearly, those values belong to Group 2 since there is no energy around 220 Hz and the values were computed for higher frequency bands.

Finally, chord complexes were analyzed using the model as depicted in Fig. 7. The left graph shows a triad of sinusoids with frequencies of 220, 262 and 330 Hz. The model correctly identifies all tones. The right graphs show a cluster of $1/f$ tone complexes with the following fundamental frequencies: 220, 262, 330 and 880 Hz. The model identifies all fundamental frequencies correctly, but also a number of octaves, for example at 516 Hz which is the octave of the 262-Hz tone.

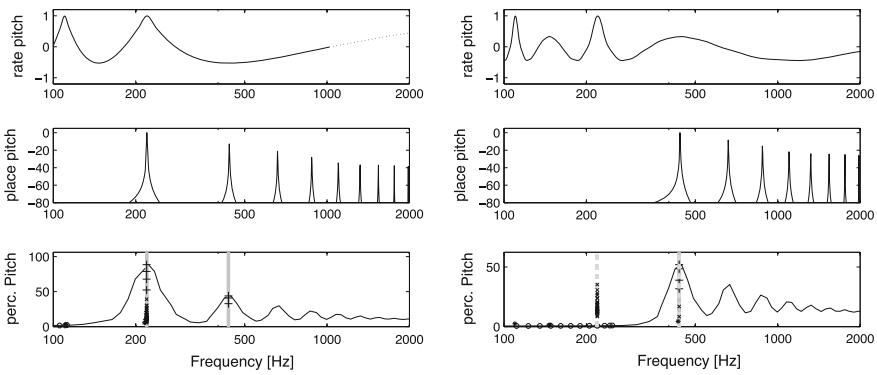

Fig. 6 Same as Fig. 5, but for 220-Hz signals

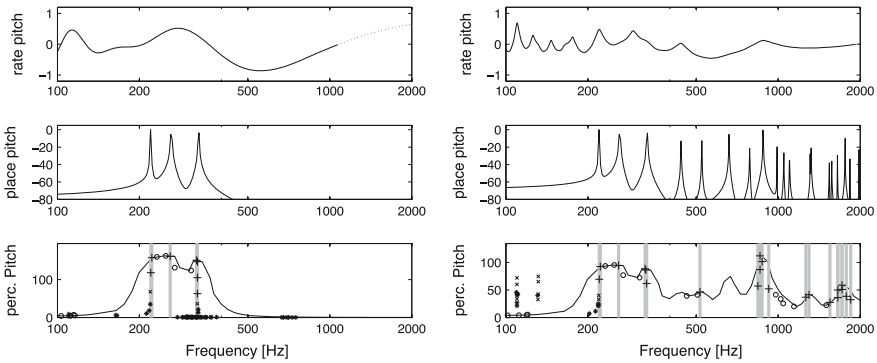

Fig. 7 Same as Figs. 5 and 6, but for polyphonic tone clusters. *Left* sinusoids with frequencies of 220, 262, and 330 Hz, *Right* tone complexes with frequencies of 220, 262, 330 and 880 Hz

4.1.2 Tension Arc Calculation

One of the key analysis parameters for CAIRA is the estimation of the tension arc, which describes the currently perceived tension of an improvisation. In this context, the term 'arc' is derived from the common practice of gradually increasing the tension until the climax of a performance section is reached, and then gradually decreasing tension to end it. Thus, tension often has the shape of an arc over time, but it can also have different time courses. It is noteworthy that we are not focusing here on tonal tension curves that are typically only a few bars long (i.e. demonstrating low tension whenever the tonal structure is resolved and the tonic appears). Instead, we are interested in longer structures, describing a parameter that is also related to Emotional Force [23].

Using individual microphone signals, the agent tracks the running loudness of each musical instrument using the Dynamic Loudness Model of [9]. The Dynamic Loudness Model is based on a fairly complex simulation of the auditory periphery including the simulation of auditory filters and masking effects. In addition, the psychoacoustic parameters of roughness and sharpness are calculated according to Daniel and Weber [12], and Zwicker and Fastl [32]. In the current implementation, CAIRA estimates tension arcs for each musician from simulated psychophysical parameters. Based on these perceptual parameters and its logic capabilities, the system recognizes different configurations for various patterns; e.g., it realizes that one of the musicians is performing an accompanied solo, by noticing that the performer is louder and has a denser texture than the remaining performers. The system can also notice that the tension arc is reaching a climax when all musicians perform denser ensemble textures. CAIRA takes action by either adapting its music performance to the analysis results or by presenting a dynamic visual score as described in more detail in the next section. CAIRA can, for example, suggest that a performer should end their solo because it is too long, or it can encourage another musician to take more initiative. It can guide endings and help an ensemble to fuse its sounds together.

In a previous study, we decided to calculate the tension arcs T from a combination of loudness L and roughness data R [5]:

$$T = L^4 + a \cdot R^3, \tag{2}$$

with an adjusting factor a. In a further study, we also suggested including information rate—e.g., as defined by Dubnov [15] and Dubnov et al. [16]—as an additional parameter for the tension arc calculation [7]. A real-time capable solution was developed to measure the rate and range of notes within each 2-s time interval. To achieve this, pitch is measured and converted to MIDI note numbers. Next, the number of notes n is counted within a 2-s interval, ignoring the repetition of identical notes. The standard deviation σ of the note sequence is then determined from the list of MIDI note numbers. Finally, the information rate I is determined from the product of *the number of notes* and *the standard deviation of MIDI note numbers*, or $I = n \cdot \sigma$. Practically, we measure values between 0 and 100. The tension curve is then calculated using the following equation:

$$T = \frac{1}{a+b}(a \cdot L + b \cdot ((1-q) \cdot R + q \cdot I)), \tag{3}$$

with the Information Rate I, Loudness L, and Roughness R. Note that all parameters, L, R, I, are normalized between 0 and 1 and the exponential relationships between the input parameters and T are also factored into these variables. The parameter q is the quality factor from the *YIN* pitch algorithm [14]. A value of one indicates a very tonal signal with a strong strength of pitch, while a value of zero indicates a noisy signal without defined pitch. The parameter is used to trade off roughness and information rate between tonal and noise-like signals. The parameters a and b are used to adjust the balance of loudness and the other input parameters for individual instruments. All tension curves are scaled integer values between zero and seven. Figure 8 shows an example of how a tension curve is estimated from the instruments' sound pressure signal.

4.2 Top-Down Mechanisms

A logic-based reasoning system is used to implement the top-down mechanism of CAIRA. The main purpose of the top-down mechanism of CAIRA was to provide the system with a rudimentary "understanding" of musical rules and concepts provided by a field-specific ontology. We hope that we will be able to expand the current architecture in the future such that CAIRA can use the externally injected knowledge to form its own concepts and ideas. The rule-based approach is also important for the system to be able to measure success by adhering to formal rules and "realizing" when these rules are broken. An important component of the CAIRA system is the interaction between the bottom-up and top-down

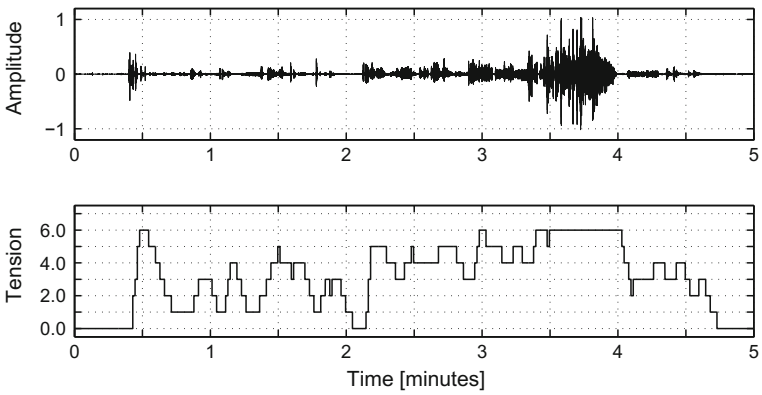

Fig. 8 Tension Arc calculation for a soprano saxophone sample. *Top* waveform of the saxophone, recorded with a closely positioned microphone. *Bottom* calculated tension arc curve—adapted from [4]

mechanisms. Of particular importance are the calculated tension arc curves, which are measured individually for each musician using closely positioned microphones, to "understand" the basics of musical interplay between the musicians. The rules of engagement for musical interplay are defined by the ontology that will be described in detail in the next section.

4.2.1 General Ontology Definitions

Our knowledge-based system is described using first-order logic notation.[2] We define that every musician has an associated dynamic level in seven ascending values from *tacit* to fortissimo, *ff*:

$$\begin{aligned}&[\forall x(Mx \rightarrow \forall t(Tt \rightarrow (d-l(x,t) = tacit \vee\ d-l(x,t)\\&\quad = pp \vee d-l(x,t) = p \vee d-l(x,t) = mp \vee d-l(x,t)\\&\quad = mf \vee d-l(x,t) = f \vee d-l(x,t) = ff)))]\\&\qquad \wedge (tacit < pp \wedge pp < p \wedge p < mp \wedge mp < mf \wedge mf < f \wedge f < ff).\end{aligned}$$

It is noteworthy here that the dynamic levels are calculated for every musician in discrete steps using the dynamic loudness model. The condition *tacit* is the case where the instrument does not produce a tone. Each moment in time is labeled through the variable S, with S_0, the start of the improvisation, and S_{END}, the end of

[2]Some of the equations have been simplified for better readability and for the reason of saving space.

the improvisation. If we have more than one musician, all existing musicians form an ensemble:[3]

$$\forall x \forall y((Mx \wedge My \wedge x \neq y) \rightarrow \exists z(Ez \wedge Ixz \wedge Iyz \wedge (\forall y(Ey \rightarrow y = z)).$$

In contrast, if we have less than two musicians an ensemble does not exist:

$$[(\neg \exists x Mx) \vee (\exists x(Mx \wedge \forall y(My \rightarrow y = x)))] \rightarrow \neg \exists z Ez.$$

4.2.2 Music Structure Recognition

Next, we define the current configuration of the ensemble. We divide the structure of the performance as a sequence of solos and ensemble parts, whereas each moment in time is characterized by a solo (of exactly one performer) or an ensemble part:

$$\begin{aligned} &\forall S(\neg \exists x Musician(x) \wedge PlaySolo(x, S)) \\ &\quad \Leftrightarrow (\forall x Musician(x) \wedge PlayEnsemble(x, S)). \end{aligned}$$

Obviously, the ensemble performance part cannot exist if we have only one performer. In this case, we can automatically conclude that the musician is performing a solo:

$$\forall x((Musician(x) \wedge \forall y(Musician(y) \Rightarrow x = y)) \Rightarrow PlaySolo(x)).$$

and the $PlayEnsemble(x, y, \ldots, S)$ mode does not exist:

$$\begin{aligned} &\forall x((Musician(x) \wedge \forall y(Musician(y) \Rightarrow x = y)) \\ &\quad \Rightarrow (PlaySolo(x) \wedge \neg PlayEnsemble(x))). \end{aligned}$$

Now, we have to decide how the agent recognizes the correct $PlaySolo(x, S)$, or the alternative $PlayEnsemble(x, y, \ldots, S)$, mode. First, our tension arc estimations for each musical instrument are based solely on their dynamic levels. This assumption needs to be refined at a later point, but high correlation values justify this initial approach. We define:

$$\forall x, S\ DynamicLevel(x, S) \Leftrightarrow TensionArc(x, S),$$

and leave it for later to refine the *TensionArc* calculation. Now we can define the solo performance mode as:

[3]Please note that we use the variable x and y for musicians, S for time, and z for an ensemble throughout this chapter.

$$\forall x, y, S\ (((Musician(x) \wedge TensionArc(x, y, S) \wedge \forall w, z \\ ((Musician(w) \wedge TensionArc(w, z, S)) \Rightarrow (y > z)))) \Rightarrow PlaySolo(x, S)).$$

Note that the solo performance mode relies on the fact that exactly one performer has to have a higher current tension arc value then all other performers. If at least two performers share the highest value, the agent recognizes the ensemble performance mode, which can be defined as:

$$\forall S\ ((\neg \exists x\ PlaySolo(x, S) \wedge \exists x\ (Musician(x) \\ \wedge \neg TensionArc(x, tacit, S))) \Rightarrow \forall x\ (Musician(x) \Rightarrow PlayEnsemble(x, S))).$$

The improvisation ends if the following condition is met:

$$\forall S\ (\forall x\ (Musician(x) \wedge TensionArc(x, tacit, S)) \Rightarrow EndOfMusic(S)).$$

We should reemphasize here that we calculate a running average dynamic level and not instantaneous values. The duration of the averaging window is crucial for the performance of the agent, but we observe similar challenges with human listeners. Take, for example, the case where an audience listens to an unknown classical composition. It always takes a certain time period until the first audience member decides when the piece is over and claps, and the audience often waits until the musicians bow. False alarms are often remembered as embarrassing incidents. Similarly, it is important that the tension arc has discrete values, otherwise, minimal tension arc differences between performers easily lead to the false detection of a solo. The integration of thresholds needs to be considered if the tension arc is calculated based on continuous values.

4.2.3 Agent Beliefs

Now we discuss the beliefs and goals of the agent. A simple example can be drawn from Jazz, where every soloist is expected to perform exactly one solo per piece:

$$\forall x, S\ (Musician(x) \wedge (NumberOfSolos(x, 0, S) \\ \vee NumberOfSolos(x, 1, S)) \Leftrightarrow DesiredState(S)).$$

In contrast, it is undesirable that a performer plays a second solo:

$$\forall x, S\ (Musician(x) \wedge PlaySolo(x, S) \wedge \neg NumberOfSolos(x, 0, S) \\ \Leftrightarrow TooManySolos(x, S)).$$

It is also an undesired state if the performer's solo gets too long:

$$\forall x, S\ ((Musician(x) \wedge PlaySolo(x, S) \\ \wedge SoloDuration(x, MaxSoloDuration, S)) \Rightarrow SoloTooLong(x, S)).$$

In this last aspect, there is much room for improvement in the agent's performance. Instead of simply assigning a threshold of what should be the maximum solo duration, the agent could observe if the performer is still producing interesting work or exhausted their creativity. An alternative method to determine if the solo becomes too long is:

$$\forall x, t, S\ [(Musician(x) \wedge PlaySolo(x, S) \wedge TensionArc(x, t, S) \\ \wedge (t < MinTensionArc) \wedge (t < MaxTensionArc)) \Rightarrow SoloTooLong(x, S)],$$

with S_{Solo} representing all moments in time of S during the performer's solo. Of course, in this case, the assumption that the tension arc simply relies on the dynamic level is very crude and the tension arc estimation should be refined, otherwise, musicians will be too restricted. Instead of simply observing the tension arc, the agent could also observe the variety of the performed solo material—e.g., via the information rate according to Dubnov et al. [16]—and the tension arc developments of the other musicians. The latter often declines, if the ensemble comes to the conclusion that the soloist should come to an end. The agent could also decide that the determination of whether a solo is too long is based on both the performance and a constant threshold. For example, John Coltrane was known to wear out the audience with long solos in Miles Davis' band, despite the excellent quality of his performance [13]. A good indicator that the solo of a performer will come to an end could be:

$$\forall x, t, S\ [(Musician(x) \wedge PlaySolo(x, S) \wedge TensionArc(x, t, S) \\ \wedge (t < AverageTensionArc)) \Rightarrow SoloMightEndSoon(x, S)],$$

which enables the agent to look ahead.

4.2.4 Action

Now we have to decide what action the agent should take if the ensemble reaches an undesired state (e.g., a solo is too long or a performer plays more than one solo within one piece). In a simple model, the agent can either

1. accept the undesired state
2. ask the other musicians ($x \neq y$) via a computer terminal if they find the solo to be too long. In this case, the agent can learn from their feedback and take further action (see below) based on the response
3. ask the performer to come to an end
4. encourage another musician to take the lead.

We can summarize this to the following proposition:

$$\begin{aligned}&\forall x, y, S[Solo\,TooLong(x, s)\\&\Rightarrow (AcceptUndesiredState(S)\\&\vee (AskMusiciansIfSolo\,TooLong(y, S) \wedge \neg(x = y))\\&\vee AskMusicianToStopSolo(x, s)\\&\vee (AskMusicianToplaySolo(y, S) \wedge \neg(x = y)\\&\wedge NumberOfSolos(y, 0.S)))].\end{aligned}$$

The agent can take similar measures if the performer plays a second solo and can also aid the musicians to end the improvisation if all performers have played a solo.

4.3 Implementation of a Free Jazz Agent

A Bayesian model is used to find an a posteriori estimation of the most likely ensemble state from the obtained tension curves. The ensemble states describe the instantaneous relationships between the musicians of an ensemble using methods in jazz ensemble practice. To keep the interaction sufficiently simple, we define six Ensemble States for a trio shown in the schematic in Fig. 4:

1. Solo A: Performer A performs a solo part
2. Solo B: Performer B performs a solo part
3. Solo C: CAIRA performs a solo part
4. Low-Tension Tutti: All ensemble members perform a tutti part with low tension
5. High-Tension Tutti: All ensemble members perform a tutti part with high tension
6. End: All musicians come to an end.

The Ensemble States are determined using a logic-based reasoning approach published in Braasch et al. [5], the practical rules that were derived in this study are given in Table 2. We cannot assume that each of the six states is performed equally long in time, but by using a Bayesian approach we can improve the Ensemble State estimation by recording how often each state occurs as a percentage over the whole training duration. To this purpose, the human performers use a foot pedal to update the Ensemble State. In addition, we can compare the states with instrumentally measured parameters. To see the general approach, let us focus on the analysis of the time-variant tension curves of Musicians A and B. We define seven discrete levels of Tension T. Curves will be computed for each participating musician and for CAIRA, so we have three tension curves: $(T_a(t), T_b(t), T_c(t))$. We can compute how often each tension level combination is observed for a given ensemble state:

Table 2 Ensemble state calculations based on logic-based reasoning

	Musician A	Musician B	CAIRA C
1 Solo A	$T_A + 1 > T_B$	$T_B - 1 < T_A$	$T_C - 1 < T_A$*
2 Solo B	$T_A - 1 < T_B$	$T_B + 1 > T_A$	$T_C - 1 < T_B$*
3 Solo C	$0 < T_A < 4$	$0 < T_B < 4$	Decision needed
4 Low Tension Tutti	$0 < T_A < 4$	$0 < T_B < 4$	Decision needed
5 High Tension Tutti	$T_B > 5$	$T_B > 5$	$T_B > 5$*
6 Ending**	$T_B = 0$	$T_B = 0$	$T_C = 0$*

The variables T_A, T_B, and T_C represent the tension curves of Musicians A, B, and CAIRA. The asterisks denote that CAIRA does not have to follow the suggestions by the other two musicians but can also respond by using a different tension curve level

$$P(E|T_{a,b}) = \frac{P(T_{a,b}|E)p(E)}{p(T_{a,b})}. \tag{4}$$

The parameter $T_{a,b}$ is the observed combined tension curve T for Musicians A and B. The Tension Curve T_c is not part of the analysis, since the intelligent agent CAIRA will observe the other two musicians to predict the current Ensemble State E. We have 49 discrete values for $T_{a,b}$, (7·7 Tension State combinations). The term $p(T_{a,b}|E)$ is the likelihood that the joint Tension Curve $T_{a,b}$ is observed for a given Ensemble State E. The term $p(E)$ is the probability that State E occurs independently of the tension curve status, and $p(T_{a,b})$ is the probability that the joint Tension Curve $T_{a,b}$ occurs independently of the ensemble state. Using the Equation given above we can compute the posterior estimate for each possible Ensemble State $E_1 - E_7$ for any Tension Curve pair $T_{a,b}$. An Ensemble State curve will be discussed further below (see also Fig. 9).

4.4 Implementation of a Traditional Jazz Agent

In this section, we describe a variation of CAIRA that is used to accompany a traditional jazz or popular music soloist instead of participating in a free improvisation. This version of CAIRA uses the aforementioned bottom-up/top-down algorithms to adjust an automated music accompany system to the live performance of a jazz soloist, for example, a trumpet player. For this purpose, the sound of the jazz soloist is captured with a microphone from a close distance, and musical features such as loudness, information rate, and musical tensions are extracted in real time. The extracted values are then used to control the probability that a certain accompany style is selected, and parameters like volume are adjusted. For example, if the soloist plays many musical notes within a short time frame (high information rate) it is much more likely that the rhythm section, performed by the CAIRA agent, will play in double time than is the case when the soloist performs a solo with only a few notes at a time.

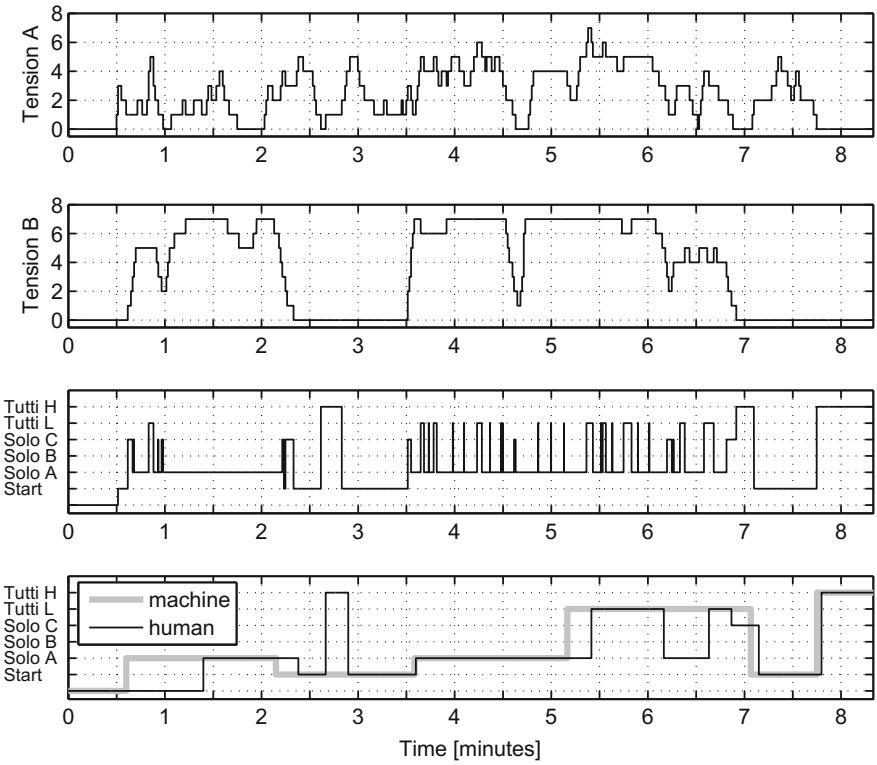

Fig. 9 Ensemble State Example for a trio session. *Top graph* tension curve for a saxophone (with variables a = 1.2 and b = 0.6); *2nd graph* tension curve for a Moog Synthesizer (a = 1.2, b = 0.4); *3rd graph* CAIRA's short term ensemble state estimations; *bottom graph* CAIRA's final (long-term) ensemble state estimations (*solid think black line*) versus human ensemble state estimations (*solid thick gray line*)—adapted from [4]

We believe that the traditional jazz version of CAIRA is particularly valuable from an educational point-of-view because it enables students to learn the jazz and popular music repertoire in a much more realistic setting than is the case for studying music with a pre-recorded backing tape because the agent offers interactive and dynamic system features.

Musical accompaniment systems have a long tradition in electronic organs used by one-man bands. Typically, the automated accompaniment produces a rhythm section (drums, bass, and a harmony instrument such as a piano) that performs in a given tempo (e.g., 120 beats-per-minute), style (e.g., Bossanova) and pre-defined chord progressions (often recorded live with the left hand of the organ player). The accompaniment system can then automatically generate a bassline and rhythmical harmonic chord structure from the performed chords and progressing chord structure. Similar systems, like Band-in-a-Box™, create a band that plays along from a manually entered chord sheet using software synthesizers for drums, bass and harmony instruments.

The problem with the current jazz/popular music accompany systems is that they are not "listening" to the performer, with the exception of systems that follow the tempo of the soloist.

Band-in-a-Box™, for example, will always perform a pre-rendered accompaniment that does not depend on the performance of the live soloist. In jazz, however, it is important that the players listen to each other and adjust their performance to the other players. For example, a good rhythm section will adjust its volume if the soloist plays with low intensity and uses sparse phrases. Often, some of the rhythm instruments rest and only part of the band accompanies the soloist. Or, the band can go into double time if the soloist plays rapidly (e.g., sequences of 16th notes). Double time is defined by playing at twice the tempo with each chord being performed twice as long in terms of musical measures such that the duration of the chord progression remains the same. In half-time, the tempo is half of the original tempo and the chord progression is half of the original metric value.

Impulses can also come from the rhythm section, for example, the rhythm section can decide to enter double time if the players believe an improvised solo could benefit from dynamic changes in structure. The adaptive performance of a rhythm section can become a real problem for a jazz student trying to practice unaccompanied. If the student is used to performing with a computerized rhythm section at home, then a live band changes this context dramatically. As a result, the jazz student is presented with a lack of experience for such situations as they may not be used to unexpected changes in the accompaniment. Likewise, it can become boring for even a highly experienced jazz player to perform with a virtual, static rhythm section that does not react to what is being played by the soloist.

Traditional jazz concretely defines the roles and basic groundwork for improvisation. In improvisation, solo instruments will introduce a composed theme and then introduce variations as elements of a solo. An improvised solo instrument will follow predetermined chord progressions that complete a phrase (such as the famous "twelve bar blues") accompanied by a rhythm section—traditionally drums, bass, and a chordal instrument such as piano or guitar. Figure 3 shows a diagram for how a traditional jazz group interacts. A standard solo will consist of an unspecified number of these complete chord progressions following the rhythm section. The more successful (and usually more famous) virtuoso jazz soloists are known to listen for and respond to details initiated by other members of their bands. In essence, the soloist sees the rhythm section as a holistic body following the chord progression, listening for details in melodic and rhythmic content from the section to incorporate into their solo. Tempo is usually held at a constant rate, and different lead instruments take cues from fellow musicians on when to introduce or fade out their own solo. In free jazz, individual solos soon blurred the line between soloists and accompaniment; in addition to distinguishing free jazz from its traditional roots, this process both increased the scope of a free jazz musician and added a layer of complexity to the improvisation process [27].

While there are rules in how to construct solos from traditional jazz theory, free improvisation gives the solo musician control over how such solos should evolve. However, as previously outlined, free improvisation carries its own history and

traditions. It seeks to eliminate almost all formal rule structures in composition. The introduction of modal jazz toward the late 1950s contributed greatly to free improvisation; chord progressions were replaced by musical modes, where instead of following pre-set changes composition instead pivoted from a musical tonic. Free improvisation at first kept the traditional structure of phrases, but soon progressed to free form in duration; this soon spilled over into other elements of performance. Popular techniques that stemmed from the experimentation of free improvisation included multiphonics, circular breathing, arpeggiated phrases, non-pitched tones, and complex textures discovered by approaching instruments in non-traditional methods [3].

Our traditional jazz agent listens to the soloist using a microphone—see Figs. 10 and 11. The system captures a number of acoustical and psychoacoustical parameters from the performed instrument including: (i) loudness, (ii) information rate (musical notes per time interval), and (iii) a tension curve based on loudness, roughness, and information rate. Alternatively, the system can compute these parameters directly from an electronic instrument (e.g., by analyzing MIDI data).

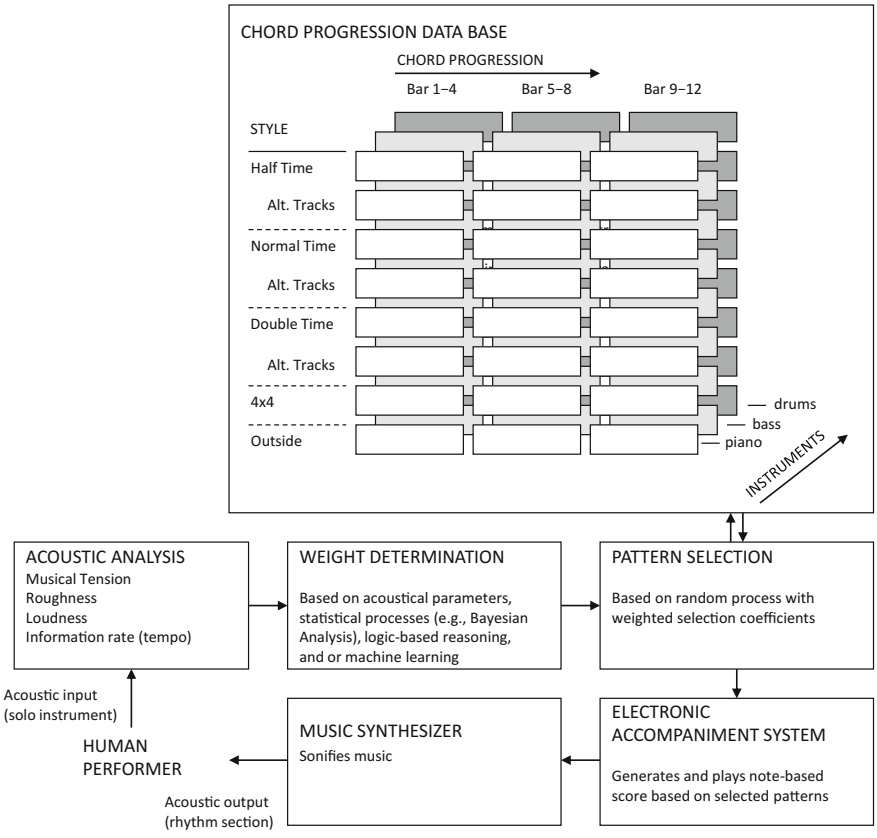

Fig. 10 System architecture for the CAIRA system for traditional jazz

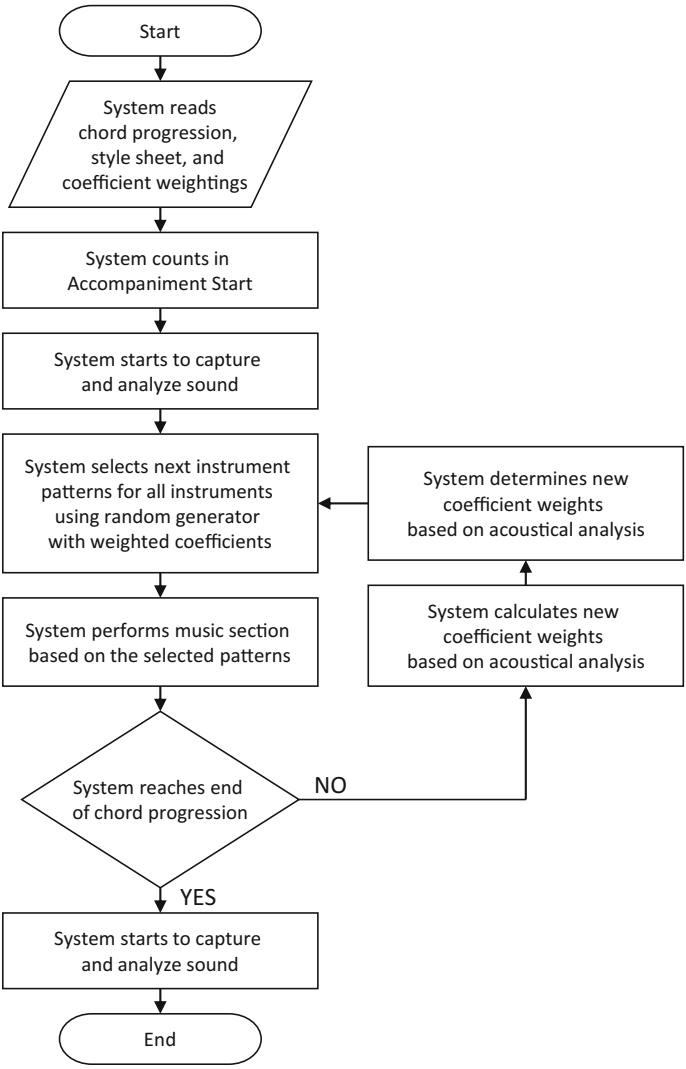

Fig. 11 System flow diagram for the traditional jazz-based CAIRA system

The accompaniment system then reacts to these measured parameters in real time making changes at strategic points in the chord progression (often at the end of four bars or the end of a phrase, or pre-specified chord structure). In particular the system will: (i) switch to double time if the soloists information rate and tension exceeds an upper threshold, (ii) perform at half time if the soloists information rate exceeds a lower threshold, (iii) return to normal time if the soloists information rate returns to in-between threshold rates, (iv) adapt the loudness of the rhythm section instruments to the loudness and tension curve of the performer, (v) play outside of the given chord structure if it detects the soloist performing outside this structure,

(iv) pause instruments if the tension curve or loudness is very low, or (vi) perform 4 × 4 between the solo instrument and a rhythm section instrument by analyzing the temporal structure of the tension curve (e.g., analyzing gaps or changing in 4-bar intervals). In a 4 × 4, the instruments take solo turns every four bars.

In addition, the rhythm section can give direction and take initiative based on a stochastic system using a random generator. For each event, a certain threshold of chance (likelihood) can be adjusted and if the internal drawn random number exceeds this threshold the rhythm section will take initiative in form of: (i) changing the style pattern, or taking a different pattern within the same style, (ii) stop instruments from changing to double time, half time, and normal time, (iii) lead into a new harmonic theme or other solos, (iv) play 4 × 4, and (v) play outside of expected structures. It should be noted that all changes can be subject to chance using a stochastic algorithm, for example by increasing the information rate to increase the likelihood for the rhythm section to change to double time, but there is no absolute threshold for these metrics.

5 Discussion and Conclusion

The purpose of this study was to develop a general framework for an intelligent music agent that can be adapted to different forms of music, in our case traditional jazz standards and free jazz. A dual bottom-up/top-down structure was chosen to simulate the creative processes needed to obtain a system that can perform live in an ensemble together with human musicians. The general architecture was identical for both types of jazz. Using the bottom-up structure an auditory scene analysis was performed, which included the estimation of pitch, loudness, information rate and beat among other parameters. A real-time tension curve was then calculated from these parameters to "understand" the intention of a soloist (traditional jazz agent) or to compare the inter-relationships between musicians (free jazz agent). A top-down structure, based on logic reasoning was used to control the agent according to specific rules of jazz.

One of the main goals for the dual bottom-up/top-down structure was to provide a mechanism where the system's response cannot be fully anticipated in advance, but at the same time to provide a framework where the human musicians who interact with CAIRA feel that the system is not responding in a random way, but "intelligently" responds to the performance of the live musicians. This can be achieved by tuning the parameter set to find the right balance between the genetic algorithms of the bottom-up stages (which can provide unexpected results) and the logic-based reasoning system, which can provide the feedback of being "understood".

One of the most critical aspects of the CAIRA project was to find solutions for the agent's ability to self-assess the success of its performance. Currently, the agent merely adheres to given rules or rejects to adhere to these rules, but it does not possess stages to assessing the musical structure according to aesthetical qualities. To achieve this is one of our long-term goals of the future.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant Nos. 1002851 and 1320059.

References

1. Arom, S.: The use of play-back techniques in the study of oral polyphonies. Ethnomusicology **20**, 483–519 (1967)
2. Boden, M.: The turing test and artistic creativity. Kybernetes **39**, 409 (2010)
3. Braasch, J.: A cybernetic model approach for free jazz improvisations. Kybernetes **40**(7/8), 972–982 (2011)
4. Braasch, J.: The μ-cos*m* project: an introspective platform to study intelligent agents in the context of music ensemble improvisation. In: Bader, R. (ed.) Sound–Perception–Performance, pp. 257–270. Springer, New York (2013)
5. Braasch, J., Bringsjord, S., Kuebler, C., Oliveros, P., Parks, A. Van Nort, D.: Caira—a creative artificially-intuitive and reasoning agent as conductor of telematic music improvisations. In: Proceedings of the 131st Convention of the Audio Engineering Society, Paper Number 8546 (2011)
6. Braasch, J., Oliveros, P. Van Nort, D.: Telehaptic interfaces for interpersonal communication within a music ensemble. In: 21st International Congress on Acoustics (ICA), Montreal, Canada (2013)
7. Braasch, J., Van Nort, D., Oliveros, P., Bringsjord, S., Govindarajulu, N. S., Kuebler, C. Parks, A.: A creative artificially-intuitive and reasoning agent in the context of live music improvisation. In: Music, Mind, and Invention Workshop: Creativity at the Intersection of Music and Computation, The College of New Jersey, Ewing, NJ. http://www.tcnj.edu/mmi/proceedings.html (2012)
8. Braasch, J., Van Nort, D., Oliveros, P., Bringsjord, S., Sundar Govindarajulu, N., Kuebler, C. Parks, A.: Music, Mind, and Invention Workshop: Creativity at the Intersection of Music and Computation, The College of New Jersey (2012)
9. Chalupper, J., Fastl, H.: Dynamic loudness model (DLM) for normal and hearing-impaired listeners. Acta Acustica United Acustica **88**, 378–386 (2002)
10. Cheveigné, A.: Pitch perception models. In: C. J. Plack A. J. Oxenham (eds.) The Psychophysics of Pitch. Springer, Berlin, Heidelberg, New York, pp. 169–233 (2005)
11. Cope, D.: An expert systemfor computer-assisted composition. Comp. Music J. **11**(4), 30–46 (1987)
12. Daniel, P., Weber, R.: Psychoacoustical roughness: implementation of an optimized model. Acustica **83**, 113–123 (1997)
13. Davis, M., Troupe, Q.: Miles, the Autobiography. Simon & Schuster, New York (1990)
14. de Cheveigné, A., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**, 1917–1930 (2002)
15. Dubnov, S.: Non-gaussian source-filter and independent components generalizations of spectral flatness measure. In: Proceedings of the International Conference on Independent Components Analysis (ICA2003), pp. 143–148. Porto, Portugal (2003)
16. Dubnov, S., McAdams, S., Reynolds, R.: Structural and affective aspects of music from statistical audio signal analysis. J. Am. Soc. Inform. Sci. Technol. **57**(11), 1526–1536 (2006)
17. Friberg, A.: Generative rules for music performance: a formal description of a rule system. Comput. Music J. **15**(2), 56–71 (1991)
18. Jacob, B.: Algorithmic composition as amodel of creativity. Organ. Sound **1**(3) (1996)
19. Jost, E.: Free Jazz. DaCapo, New York (1981)
20. Lewis, G.: Too many notes: computers, complexity and culture in voyager. Leonardo Music J. **10**, 33–39 (2000)
21. Licklider, J.: A duplex theory of pitch perception. Cell. Mol. Life Sci. **7**(4), 128–134 (1951)

22. McAdams, S., Smith, B., Vieillard, S., Bigand, E. Reynolds, R.: Real-time perception of a contemporary musical work in a live concert setting. In: Stevens, C., Burnham, D., McPherson, G., Schubert, E., Renwick, J. (eds.) Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney, Australia (2002a)
23. McAdams, S., Smith, B., Vieillard, S., Bigand, E. Reynolds, R.: Real-time perception of a contemporary musical work in a live concert setting. In: 7th International Conference on Music Perception and Cognition, Sydney, Australia, vol. 17, p. 21 (2002b)
24. Meddis, R., O'Mard, L.: A unitary model of pitch perception. J. Acoust. Soc. Am. **102**, 1811 (1997)
25. Pachet, F.: Beyond the cybernetic jam fantasy: the continuator. IEEE Comput. Graph. Appl. **24**, 31–35 (2004)
26. Scott, B.: Second-order cybernetics: an historical introduction. Kybernetes **33**(9/10), 1365–1378 (2004)
27. Spitzer, P.: Jazz Theory Handbook. Mel Bay Publications, Pacific, MO (2001)
28. Van Nort, D., Braasch, J., Oliveros, P.: A system for musical improvisation combining sonic gesture recognition and genetic algorithms. In: Proceedings of the 6th Sound and Music Computing Conference, pp. 131–136. Porto, Portugal (2009)
29. Van Nort, D., Braasch, J. Oliveros, P.: Mapping to musical actions in the filter system. In: The 12nd International Conference on New Interfaces for Musical Expression (NIME), Ann Arbor, Michigan (2012)
30. Van Nort, D., Oliveros, P. Braasch, J.: Developing systems for improvisation based on listening. In: Proceedings of the 2010 International Computer Music Conference (ICMC 2010), New York, NY (2010)
31. Widmer, G.: Qualitative perception modeling and intelligent musical learning. Comput. Music J. **16**(2), 51–68 (1992)
32. Zwicker, E., Fastl, H.: Psychoacoustics: Facts and Models, 2nd edn. Springer, Berlin (1999)

Author Biographies

Jonas Braasch is a psychoacoustician, aural architect, and experimental musician. His research work focuses on functional models of the auditory system, large-scale immersive and interactive virtual reality systems, and intelligent music systems. Jonas Braasch received a Master's Degree in Physics from the Technical University of Dortmund in 1998, and two doctoral degrees from the University of Bochum in Electrical Engineering and Information Technology in 2001 and Musicology in 2004. Afterward, he worked as Assistant Professor in McGill University's Sound Recording Program before joining Rensselaer Polytechnic Institute in 2006, where he is now Associate Professor in the School of Architecture and Director of the Center for Cognition, Communication, and Culture.

Selmer Bringsjord is a full professor in the Department of Cognitive Science at Rensselaer Polytechnic Institute. He teaches artificial Intelligence (AI), formal logic, human and machine reasoning, and philosophy of AI. Bringsjord's education includes a B.A. in Philosophy from the University of Pennsylvania, and a Ph.D. in Philosophy from Brown University. He conducts research in AI as the director of the Rensselaer AI & Reasoning Laboratory (RAIR). He specializes in the logico-mathematical and philosophical foundations of AI and cognitive science, and in collaboratively building AI systems on the basis of computational logic.

Nihil Deshpande is a doctoral student in architectural acoustics at Rensselaer Polytechnic Institute. His focus is on audio digital signal processing, specifically on computational auditory scene analysis and computer models of binaural hearing. Nikhil specializes in understanding,

developing, and extending digital audio effects, particularly reverberation algorithms. His master's research included a polyphonic pitch decomposition model and a dynamic real-time jazz accompaniment system. He has also been a programmer on Pauline Oliveros's Expanded Instrument System, for which his work was featured in the main gallery of the Whitney Museum of American Art as part of their Biennial in 2014. He received his B.S. in Electrical Engineering with a minor in Electronic Arts in 2013, and his M.S. in Architectural Acoustics in 2014, both from RPI.

Pauline Oliveros is a senior figure in contemporary American music. Her career spans fifty years of boundary dissolving music making. In the '50s she was part of a circle of iconoclastic composers, artists, poets gathered together in San Francisco. Recently awarded the John Cage award for 2012 from the Foundation of Contemporary Arts, Oliveros is Distinguished Research Professor of Music at Rensselaer Polytechnic Institute, Troy, NY, and Darius Milhaud Artist-in-Residence at Mills College. Since the 1960's she has influenced American music profoundly through her work with improvisation, meditation, electronic music, myth, and ritual. Pauline Oliveros is the founder of the Deep Listening Institute, now the Center For Deep Listening at Rensselaer.

Doug Van Nort is a computer music researcher and electronic music composer/improviser whose work is concerned with distributed agency, computational creativity, electroacoustic improvisation and sensorial immersion in technologically-mediated environments. He is Canada Research Chair in Digital Performance and an Assistant Professor in Computational Arts and Music at York University in Toronto.