Beyond Reflexivity: Mediating between imitative and intelligent action in an interactive music system

Toby Gifford
Griffith University, Brisbane
QLD, Australia 4101
http://www.griffith.edu.au
t.gifford@griffith.edu.au

Andrew R. Brown
Griffith University, Brisbane
QLD, Australia 4101
http://www.griffith.edu.au
andrew.r.brown@griffith.edu.au

In this paper we describe some interaction strategies employed by the Jambot – an interactive music system that we have recently developed. In particular we outline three approaches that the Jambot uses to mediate between imitative and intelligent action. The approaches are (i) mode switching based on confidence of understanding, (ii) filtering and elaboration of imitative actions and (iii) measured deviation from imitative action according to a salient parametrisation of the action space. We provide an abstract description of the properties required to apply these approaches in more general computational agents, and suggest that they may be useful outside of the music domain, particularly in non-verbal communication for affective conversational agents.

Interactive Music Systems, Reflexive, Jambot, Conversational Agent, Human-Computer Interaction

1. INTRODUCTION

We have recently developed an interactive music system called the Jambot. The Jambot is a computational music agent that listens to an audio stream and produces improvised percussive accompaniment in real-time. In this paper we describe some interaction strategies employed by the Jambot, and comment on the broader applicability of these strategies in Human-Agent Interaction.

Human-Agent Interaction (HAI) is a subfield of HCI in which the interface between the human and the computer is a computational agent. Agent interfaces are "shifting users' view of information technology from tools to actors" (Persson et al. 2001:349). From an HCI perspective these interfaces are interesting because they potentially provide a more natural mode of engagement with the computer.

Two common paradigms in HAI are identifiable:

1. the use of knowledge representations, symbolic processing, and other techniques of Good Old Fashioned Artificial Intelligence to generate actions based on 'understanding'.
2. the use of imitative actions, where the computational agent mirrors the actions of the human, typically with a twist to obfuscate the direct relationship.

A key issue then is how to seamlessly combine these two paradigms. This paper describes three abstract approaches to mediating between imitative and intelligent action, and their implementations in an interactive music system.

Interactive music systems are systems in which a human interacts with a computational agent in musical performance. Interactive music systems provide an interesting context for research into HAI due to the non-representational nature of music, which brings into focus extra-lingual aspects of interaction, particularly in improvisation.

A burgeoning field of research within HAI is the study of affective conversational agents, which combine linguistic competence with non-verbal communication mechanisms such as facial expressions. We suggest that the approaches to combining imitative and intelligent action outlined in this paper may have bearing for affective conversational agents.

2. THE JAMBOT

The Jambot is a computational music agent, designed for real-time ensemble improvisation. It listens to a musical audio stream, such as an audio feed from a live band or DJ, and improvises percussive accompaniment.

Talk about parsing the input signal into three percussive streams.

Talk about beat tracking, metre induction, and other examples of musical 'understanding'.

Talk about salient parametrisation of rhythm space.

For the purposes of this paper we are considering 'intelligent' actions to be actions taken on the basis of some musical 'understanding'.

Introduce the language of reactive and proactive generation for imitative and 'intelligent'.

3. INTERACTIVE MUSIC SYSTEMS

The Jambot is an interactive music system. Interactive music systems are computer systems for musical performance, in which a human performer interacts with the system in live performance. The computer system is responsible for part of the sound production, whether by synthesis or by robotic control of a mechanical instrument. The human performer may be playing an instrument, or manipulating physical controllers, or both. The system's musical output is affected by the human performer, either directly via manipulation of synthesis or compositional parameters through physical controllers, or indirectly through musical interactions.

There exists a large array of interactive music systems, varying greatly in type, ranging from systems that are best characterised as hyperinstruments to those that are essentially experiments in artificial intelligence. The type of output varies from systems that perform digital signal processing on input from an acoustic instrument, through systems that use techniques of algorithmic improvisation to produce MIDI output, to systems that mechanically control physical instruments. More detailed surveys of interactive music systems may be found in Rowe (1993), Dean (2003) and Collins (2006).

3.1. Transformative vs Generative Systems

Interactive music systems can be broadly classified into two categories, transformative and generative (Rowe 1993). Transformative systems transform incoming musical input (generally from the human performer playing an instrument) to produce output. Generative systems utilise techniques of algorithmic composition to generate output. Rowe also discusses a third category of sequencing, however in this paper we consider sequencing as a simple form of generation.

Transformative systems tend to be relatively robust to a variety of musical styles. They benefit from inheriting musicality from the human performer, since many musical features of the input signal may be invariant under the transformations used. A limitation of transformative systems is that they can generally only produce output that is stylistically similar to the input. From a philosophical point of view, transformative systems can be argued to not be displaying musical 'understanding'.

Generative systems, on the other hand, use algorithmic composition techniques to produce output. The appropriateness of the output to the input is achieved through more abstract musical analyses, such as beat tracking and chord classification. Generative systems are able to produce output that has a greater degree of novelty than transformative systems. These systems can be argued to have some understanding of the musical context. They are often limited stylistically by the pre-programmed improvisatory approaches, and may not be robust to unexpected musical styles.

3.2. Reflexive Systems

Sitting between generative and transformative systems is another class of interactive music system, called reflexive (Pachet 2006). Reflexive systems are transformative in the sense that they manipulate the input music to produce an output. The manipulations that they perform are more complex than is typical of transformative systems. Reflexive systems aim to model the style of the input material, for example using Markov models trained on a short history of the input.

Reflexive systems enjoy the benefits of transformative systems, namely inheriting musicality from the human input, and so are robust to a variety of input styles. The use of more abstracted transformations means that they can produce surprising and novel output whilst maintaining stylistic similarity to the input. Reflexive systems, despite giving the appearance of sensitivity to style, do not demonstrate musical 'understanding' in the sense of abstract analyses such as beat-tracking. Like transformative systems they are limited to producing output stylistically similar to the input.

3.3. Beyond Reflexivity

The Jambot is designed to combine transformative and generative approaches. In this way achieves the flexibility and robustness of a transformative system, whilst allowing for aspects of musical understanding to be inserted into the improvisation. The idea is to operate from a baseline of transformed imitation, and to utilise moments of confident understanding to deviate musically from this baseline.

Another limitation of many reflexive and generative systems is that they rely on sub-symbolic representations of music, such as Markov models, neural nets, genetic algorithms and the like. The difficulty with sub-symbolic representations is that they do not directly model salient musical features. This means that the parameters of these models do not afford intuitive control over the musical features of their output. The Jambot utilises a representation of musical rhythm that parametrises rhythm space into musically salient features. This way the Jambot's generative processes may be controlled intuitively.

In the next two sections we outline the Jambot's reactive and proactive generation techniques. The reactive (or imitative) technique we call *transformational mimesis*. The proactive (or 'intelligent') technique relies upon a salient parametrisation of the action space.

4. TRANSFORMATIONAL MIMESIS

Transformational mimesis is the term that we use for the Jambot's imitative approach to musical improvisation. Transformational mimesis involves imitating the percussive onsets in the musical stream as they are detected.

For transformational mimesis to sound musical it is essential that onsets in the signal are detected with very low latency. This way the Jambot can play a percussive sample as soon as an onset is detected and the timing difference between the actual onset and the Jambot's response will be imperceptible to a human listener.

Transformational mimesis seeks to transform the onsets, so that the imitation is not too readily identifiable as being an imitation. The Jambot uses several approaches to transforming the detected onsets.

One simple way in which the detected onsets are transformed is that the timbre of the percussive samples the Jambot plays differ from the original onsets. Indeed, the Jambot itself does not have any synthesis capacity, but rather sends out MIDI note-on messages which may be used to fire samples from any synthesiser. Used in this way transformational mimesis may be thought of as a real-time timbral remapping technique.

The timbral remapping is made more effective due to the Jambot's ability to discriminate between three streams of percussive onsets, tuned to the hi-hat, snare and kick drum sounds of a standard drum-kit. Because of this the Jambot is able to selectively highlight any of these streams, which again helps to obscure its direct relationship to the source signal.

Another simple transformation is to filter the onsets in some fashion, such as by a threshold amplitude. This can have the effect of highlighting important musical events. Transformations that select certain events from the original are reminiscent of Aristotle's discussion of mimesis in the context of drama:

> At first glance, mimesis seems to be a stylizing of reality in which the ordinary features of our world are brought into focus by a certain exaggeration ... Imitation always involves selecting something from the continuum of experience, thus giving boundaries to what really has no beginning or end. Mimesis involves a framing of reality that announces that what is contained within the frame is not simply real. Thus the more "real" the imitation the more fraudulent it becomes. (Aristotle in Davis 1999:3)

A limitation of the purely reactive approach is that it is difficult (or musically dangerous) for the Jambot to take any action other than when an onset is detected. For music that is very busy (i.e. has a lot of percussive onsets) simple reactivity can be quite effective. For music that is more sparse this can render the purely reactive approach ineffective. Transformational mimesis need not, however, be a purely reactive approach. Indeed, the Jambot uses musical understanding gathered from its perceptual algorithms to help transform its imitation.

5. PROACTIVE TECHNIQUES

The Jambot's proactive generation technique uses 'understanding' of the musical context to produce appropriate and complementary musical actions. The Jambot performs beat-tracking, metre induction and rhythmic analyses on the input signal, which constitute its understanding of the musical context.

The proactive generation technique operates by targeting values for various rhythmic analyses. The rhythmic analyses are performed continuously on the combined rhythm of the human performer(s) and the Jambot's own musical actions. At each decision point the Jambot searches the space of possible rhythmic actions for the action that will move the ensemble rhythm most closely to the desired value for each rhythmic analysis.

The rhythmic analyses used are (i) what periodicities are present in the rhythm, (ii) how consonant the periodicities are, (iii) how well aligned to the metre the rhythm is, (iv) how syncopated the rhythm is, and (v) how densely the rhythm fills the bar. More detail on the proactive generation technique is given in (Gifford & Brown 2010).

6. COMBINING REACTIVE AND PROACTIVE

We suggest that both imitative behaviour and behaviour grounded in understanding are crucial aspects of creating an impression of agency. The Jambot utilises imitation, but also takes actions based on its musical understanding of what would be an appropriate and complementary action.

The critical point here is a question of baseline. From the perspective of classical AI a musical improvising agent would operate from a blank slate - any actions taken would be on the basis of parsing the incoming music into some higher order understanding, and utilising its musical knowledge to generate an appropriate response.

We suggest, on the other hand, taking direct imitation as a baseline, and utilising musical understanding to deviate artfully from this baseline. This way the agent can communicate its musical understanding in a manner that minimises the cognitive dissonance with the human performer(s).

The Jambot utilises three approaches to combining reactive and proactive interaction strategies:

1. Switching based on confidence
2. Filtering and elaborating imitative actions
3. Measured deviation from imitative actions according to the dimensions of a salient parametrisation of action space

In the next sections we describe these approaches in more detail.

6.1. Switching based on confidence

The first approach is a simple switching mechanism according to the Jambot's confidence in its understanding of the musical context. Below a threshold confidence the Jambot operates entirely by transformational mimesis, and above the threshold it uses proactive generation.

Give an example

6.2. Filtering and Elaborating

The second approach uses the detected onsets as a 'decision grid'. Each time an onset is detected the Jambot considers its potential actions. If the onset is tolerably close to a beat then it will make a stochastic decision as to whether to take an action. If it does take an action, that action may be to play a note (which note depends on whether this beat is believed to be the downbeat or not), or to play a fill. A fill consists of playing a note followed by a series of notes that evenly subdivide the gap between the current beat and the next beat. If the onset is not tolerably close to the beat then no action is taken.

The musical understanding, which in this case consists of knowing where the beat and downbeat are, is thus incorporated by affecting the distributions for choosing whether or not to play a note and which note to play, and for the timing of the subdivisions of the fills. The fills mean that the Jambot is not restricted to playing only when an onset is detected, but anchoring the decision points to detected onsets provides a good deal of robustness.

Using this second approach, if no onsets are detected then the Jambot does not play. Although in some musical circumstances it would be desirable to have the Jambot playing in the absence of any other percussion (such as taking a solo), in practice this is frequently a desirable property. It means that the Jambot doesn't play before the piece starts or after it finishes, and allows for sharp stops; the most the Jambot will ever spill over into a pause is one beat's worth of fill. It also means that it is highly responsive to tempo variation, and can cope with sudden extreme time signature changes – especially in combination with confidence thresholding described above.

6.3. Measured Deviation

The third approach is to make measured deviations from a baseline of imitative action according to a salient parametrisation of rhythm space. As discussed above, the Jambot performs a number of rhythmic analyses, which together form a salient representation of rhythm space. This means that any rhythm can be represented in terms of a collection of musically significant parameters. This type of representation contrasts with sub-symbolic representations (such as markov models and neural nets) commonly used in interactive music systems, whose internal parameters do not correspond directly to musically significant features.

Representing rhythm space via a salient parametrisation facilitates deviation from a baseline of imitative action in a musically appropriate fashion. For example, the Jambot's understanding of the musical context might suggest that a lower rhythmic density would be musically appropriate. It can then transform

its imitative action along the dimension of density, whilst holding the other rhythmic properties constant.

In this way the generated rhythm still benefits from inheriting musicality from the human performer along some rhythmic dimensions, whilst having the flexibility to incorporate musical decisions along other dimensions.

7. BROADER IMPLICATIONS FOR HCI

We suggest that the approaches to combining imitative and intelligent action, outlined above, may have bearing more broadly in Human-Agent Interaction, particularly for non-verbal communication in affective conversational agents.

There is a long history of conversational agents (reference), and recognition of the potential role for conversational agents in Human-Computer Interaction is growing (reference). More recently it has been recognised that non-verbal communication mechanisms play an important role in human conversation, and that simulation of these mechanisms can enhance the effectiveness of conversational agents (Foster 2007).

The non-verbal communication mechanisms most commonly simulated are facial expressions (references). Kopp et al. (2004:437) note that imitative approaches to facial expression generation can be effective, but that implementation of imitative approaches is rare, with most affective conversational agents utilising one-to-one mappings between perceived emotional content and appropriate gestures.

Given that both imitative and intelligent strategies of gestural action can be effective in affective conversational agents, we suggest that the abstract approaches to combining imitative and intelligent action, described in this paper, may have relevance for conversational agents. For example, a conversational agent might be imitating the humans facial expressions, when it perceives that something surprising has been said. How best can it combine the goals of imitating with the goal of looking surprised? In the next paragraphs we outline how the three approaches outlined above might function in this context.

The first approach, switching based on confidence, would simply choose between imitation or looking surprised based on how confident the agent was of the perceived emotional content.

The second approach, filtering and elaborating ... hmmm.

The third approach, measured deviation from imitation along the dimensions of a salient parametrisation, would suggest striking a facial expression somewhere between imitation and surprise. For this approach to function it would be necessary to have an underlying parametrisation of the space of gestures in terms of salient parameters. The salient properties for affective agents are generally considered to be emotional states, and a salient parametrisation of emotional space could be the two-dimensional Valence-Arousal model of (reference) or the three dimensional Valence-Arousal-Stance model of (reference).

Kopp et al. note that “most current systems do not provide any means to model gestural images explicitly based on their inner structure” (2004:438), and that this is an impediment to escaping from the one-to-one mapping of emotional state to gesture characteristic of most affective conversational agents. Furthermore, they note that many of these agents utilise sub-symbolic representations of gesture, such as markov models and neural nets, whose internal parameters do not correspond to emotional properties, much as the use of sub-symbolic representations of rhythm in interactive music systems does not provide for intuitive musical parameters.

8. REFERENCES

Rowe, R. (1993) *Interactive Music Systems*. The MIT Press, Cambridge MA.

Dean, R. (2003) *Hyperimprovisation: Computer-Interactive Sound Improvisation*. A-R Editions, Madison WI.

Collins, N. (2006) ‘Towards Autonomous Agents for Live Computer Music: Realtime machine listening and interactive music systems’, PhD thesis, Cambridge University, Cambridge.