

Toby Gifford & Andrew R. Brown
 Queensland University of Technology
 2 George Street
 Brisbane, 4000
 Australia
t.gifford@student.qut.edu.au
a.brown@qut.edu.au

Polyphonic Listening: Real-time accompaniment of polyphonic audio.

Abstract

This paper outlines a technique for generative musical accompaniment of a polyphonic audio stream. The process involves the real-time extraction of salient harmonic features and the generation of relevant musical accompaniment. We outline a new system for polyphonic pitch tracking of an audio signal which draws upon and extends previous pitch tracking techniques. We demonstrate how this machine listening system can be used as the basis for a generative music improvisation system with the potential to jam with a live ensemble without prior training.

Introduction

The musical collaboration between people and machines has a long history (Levenson 1994). Technologies to support this collaboration have often built upon signal processing solutions designed from an engineering perspective, most recently those dedicated to music information retrieval. The approach adopted in this paper draws on similar techniques but approaches them with a sensibility derived from music perception studies concerned with salient, or musically significant, features. Our broad objective is to enable automated accompaniment of a polyphonic audio input, and the process outlined here is a step toward that goal focusing on the identification of the harmonic context of the audio source.

Among the important steps in achieving our goal of an autonomous computational improviser is the extraction of musically salient features from polyphonic audio input, and generation of an appropriately complementary musical response. Our previous work (Gifford & Brown 2006) focused on the latter problem and this paper reports on our work toward a solution to the former.

Broadly speaking musical features can be grouped into those that relate to pitch, rhythm, and timbre. In polyphonic material beat tracking can become complex but in the case of metrical music the task is not overly confused by additional instruments. However, the identification of specific note boundaries is quite involved because, while informed by metrical concerns, it benefits from the coordinated analysis of pitch and timbre, and their continuity or change over time. It is to this task of note extraction from polyphonic audio information that we focus in this paper. We will first describe our approach to an integrated analysis of these features with a particular focus on polyphonic pitch tracking. Following this we will provide an over-

view of how we demonstrate the effectiveness of the analysis through a simple real-time accompaniment process. Outlining this process is important not only to demonstrate the veracity of our claims that the salient pitch features have adequately been identified, but also because the design of the pitch tracking necessarily takes into account the generative musical application to which it is being put when deciding amongst the necessary trade-offs and making choices with regard to degrees of filtering or data abstraction.

Background

The perception of pitch alone is multifaceted and intrinsically tied up with timbral spectra, dynamic and phase (Pierce 1999). Whilst computational analysis of monophonic material has been possible for some time the ability to distinguish between multiple sources, sometimes called the ‘cocktail party effect’ remains elusive because of physical effects resulting from interference and overlay of spectral features between parts, including resolving displacements. This is confused further when attempting stream segmentation (extraction of parts) because of ambiguities in the apparent motion of parts (Shepard 1999).

However, the salient features of musical material as required for generative musical accompaniment need not be as accurately described as those required for transcription or music information retrieval. The pitch features identified as most important in a range of studies include pitch class set, pitch range or contour, and changes in these over time. These lead to structural features including tonality, harmonic change, textual density, grouping and proximity, that can be used for generative purposes (Sloboda 1988, Temperly 2001, Cope & Hofstadter 2001).

Having identified some of the salient pitch features from the raw physical description in the audio signal, we are in a position to use these to generate a simple accompaniment based on the rules of western music theory and those outlined by musicologists such as Temperly (2001).

Generative accompaniment

Generative accompaniment systems improvise a musical part to compliment incoming data. In our system this data is the data generated by a polyphonic pitch tracking algorithm. Generative accompaniment is different to some prepared-accompaniment systems such as *SmartMusic* and *In The Chair* that track an acoustic performance and play a synchronised prepared accompaniment.

GenJam, by Al Biles (1994, 2002) is a prominent generative accompaniment system. It uses a pitch to MIDI converter to track a live performer and a combination of prepared, recombinatorial and genetic algorithm processes to create an accompaniment. Our system differs from *GenJam* in that it tracks a polyphonic audio stream and does not rely on a database for the construction of accompaniment material. Our system is similar to many computer-assisted compositions systems except that it operates in real-time and is therefore better described as an improvisation system, or an interactive music system along the lines of Robert Rowe's *Cypher* (1993) or those created by musicians including Todd Winkler (1998) or Roger Dean (2003). By comparison our current generative accompaniment systems is quite rudimentary, but it does serve to demonstrate effective musical extrapolation from data produced by our polyphonic audio tracking process; a task not attempted by the previously mentioned systems.

Beliefs and prediction

A key metaphor that drives our approach at both the signal processing and generative accompaniment stages is the maintenance of expectations or beliefs about future events. In the absence of evidence to the contrary these beliefs are upheld or states are maintained. This approach provides some inertia to the system giving it greater stability and can also improve efficiency of algorithms when searches are focused or abandoned early as a result. This approach is informed by, but not rigorously consistent with, theories in cognitive linguistics (Jackendoff 1992, 2002) and computational neuroscience (Hawkins & Blakeslee 2004).

Pitch Tracking

We wish to be able to extract pitch and timing information sufficient to recognise the key and rhythm from an audio stream in real-time.

There is a large body of work in monophonic pitch tracking, particularly from the speech analysis community. Much less attention has been devoted to polyphonic pitch tracking, which is generally regarded as a difficult and unsolved problem (Hainsworth 2004; Collins 2006 pp. 60-61).

Desired Features

- (a) Tonality analysis – we want to be able to identify the current 'key'
- (b) Timing information from pitch changes – we wish to be able to identify the times at which a new note is played, even in the case where the note changes are the result of a legato movement – for which no attack is present – so as to enable tempo, metre and rhythm to be inferred.
- (c) Cope with low frequencies – much of the important harmonic information is provided by bass instruments and many pitch tracking algorithms require large window sizes to cope ade-

quately with the lowest notes on the bass guitar.

- (d) Bias towards less False Positives – with a focus on salient features it was more important to have correct information, even at the expense of less complete data.

Onset Detection

A large variety of methods for the detection of note onsets have been discussed in the literature, a survey may be found in (Collins 2006). Typically onset detection algorithms are analysed in terms of their performance on different types of sounds. Bello (2004) considers the broad classes of Pitched Non-Percussive (PNP), Pitched Percussive (PP), Non-Pitched Percussive (NPP) and Complex Mixtures (CMIX) sounds. Broadly speaking energy based techniques are reasonably successful for Percussive sounds but fail spectacularly for Pitched Non-Percussive. Bello reports on a technique based on spectral phase information that is relatively successful in the PNP class.

The system that we present in this paper concentrates on reporting onsets for the PNP class of sounds. An example of the types of onsets that we are considering would be a wind instrument playing a legato passage. Intuitively, in order to detect note boundaries in this class of sounds, one must have an accurate estimate of the frequency of the sound through time, so that an onset may be reported when the frequency changes substantially.

The reason that we concentrate on this class is that, in accord with the intuition expressed above, accurate onset timing information in the PNP class of sounds is an added bonus from accurate frequency estimation, which is the second goal of this system. Indeed Bello's system essentially estimates the instantaneous frequencies contained in the audio signal (although he is not explicitly interested in these estimates). The system that we present here operates on a similar principal to that of Bello, producing accurate frequency information and timing information for the PNP class of sounds.

We envisage that an improvising computational agent would run a number of onset detection algorithms in parallel, of which this could be one, so as to deal most appropriately with the variety of onsets that may be encountered in a complex audio stream.

Time Resolution

If the timing information is to be useful for beat and metre induction tasks, it is our view that a high degree of time resolution for onset events is desirable. As we are proposing to utilise frequency information to report on note boundaries, a practical consideration arises regarding the size of the analysis window that we use, namely that for low frequencies the analysis window may be too short to obtain an accurate estimate of the frequency. In signal processing there is generally a trade-off between

time resolution and frequency resolution (Puckette 1998).

We are aiming at a time resolution of 1024 samples at a sample rate of 44100 Hz, corresponding to approximately 23ms. This was chosen as a reasonable goal as it is as below the generally accepted minimum perceptible time (Goebel 2001) and is a commonly available buffer size for audio hardware.

We wish to be able to determine the frequency content of pitched musical material to within a semitone across the range of frequencies commonly present in (western popular) music, say 60Hz to 16000Hz. A fundamental problem that we face is that a window size of 1024 samples represents around 1.5 cycles of a 60Hz frequency component. The low number of cycles makes it difficult to accurately estimate the frequency of low frequency components with the desired time resolution.

Harmonic Context Description

Because our goal is to isolate salient pitch material that will enable us to infer the current harmonic activity such as key or chord progression, we are not so concerned about accurately detecting all notes being produced. Our plan is as follows: Identify the spectral peaks in the analysis window – these are viewed as the *constituent frequencies*. Associate harmonically related frequencies together as being part of one *pitch*. This yields a number of distinct *pitches* present in the analysis window – this is the information that will be fed to the generative improvisation engine. The nature of overtone series for physically produced sounds suggests that an appropriate method for associating harmonically related frequencies is to start with the lowest frequency and associate with it any frequencies which are an integer multiple. To this end it is important to have an accurate estimate of the fundamental frequency as errors in this estimate become compounded with each multiple.

Harmonic Product Spectrum

A number of pitch tracking techniques attack this problem by using information from harmonics of the fundamental to refine the pitch estimate. The harmonics, being at a higher frequency, have more cycles within the analysis window and so can be more accurately estimated. Then, assuming that the sound source is harmonic, the frequency of the fundamental can be estimated as a common divisor of the frequencies of the harmonics. One such technique that operates along these lines is the Harmonic Product Spectrum (Noll 1969), which can be quite effective for pitch tracking of monophonic harmonic sources. Another technique which utilises information from a larger portion of the spectrum to infer the frequency of the fundamental is Goto's predominant-F0 estimation (2004).

The Chicken or the Egg?

However, approaches that use information from the whole spectrum to estimate the frequency of the

fundamental introduce an undesirable circularity to the analysis – whilst trying to ascertain whether or not a given frequency is a harmonic of a given fundamental frequency, the process relies upon refining the estimate of the fundamental based on the assumption that it *is* harmonically related to the given frequency. For monophonic audio sources with a known spectrum this does not pose a great difficulty, but in the case of polyphonic audio composed of an unknown number of varied sources this can be problematic.

Independent Fundamental Estimation

In this paper we describe a novel frequency estimation technique that avoids such circularity by obtaining an accurate estimate of the fundamental frequency independently of the rest of the spectrum. Having done this, the estimate of the fundamental frequency may be further refined by whole-spectrum techniques such as above, but applying our new technique first facilitates the accurate assessment of which of the constituent frequencies are indeed harmonically related to each other *before* using the harmonic relations to refine the frequency estimates.

Frequency Estimation Techniques.

There are a range of techniques commonly used to estimate the fundamental frequency of a signal. We will explore the most prominent of them and highlight ways in which they may not meet our requirements.

Time Domain:

1. Zero Crossings. A simple method for frequency estimation is to count the number of times that the signal crosses zero in a given period. For low frequency signals, where the number of cycles is a small multiple of the analysis period, the discrete nature of the zero crossing events leads to large variations in the frequency estimate depending on the phase of the signal.

2. Autocorrelation. The autocorrelation of a signal is the correlation of the signal to a time-lag of itself. A periodic signal should be most highly correlated with itself at a time-lag equal to its period. Autocorrelation frequency estimation techniques utilise this by calculating the autocorrelation as a function of time-lag (called the autocorrelation function) and searching for a maxima of this function. In practice this technique has a large variance for low frequency signals and so is unsuitable for our purposes.

Fourier Domain:

Fourier techniques generally take as a starting point the spectrum of the windowed signal, usually obtained via a Fast Fourier Transform (FFT) algorithm. The most straightforward application of the spectrum is to plot its power versus bin number and identify the bins at which the power has a 'significant peak'. The centre frequency of the bin is then identified as a *constituent frequency* in the sig-

nal. A problematic issue with this approach is that the resolution of the frequency estimates is determined by the size of the bins. In a standard FFT the bins are equally sized at SF / N where SF is the sampling frequency and N is the number of samples in the window. In our case the sampling frequency is 44100Hz and the number of samples is 1024, yielding a bin size of 43Hz. This means that frequency estimates obtained in this manner are accurate to within 43Hz, which is ample for high frequency components but insufficient for low frequency components.

A number of techniques exist for increasing the resolution of the FFT:

1. Zero Padding. Before taking the FFT the signal is padded with zeros – i.e. the signal vector is concatenated with a vector consisting of zeros producing a longer vector upon which the FFT is performed. The zeroes do not effect the frequency composition of the signal, however the resolution of the frequency estimates are increased due to the larger number of bins (Smith 2003).

2. Parabolic Interpolation. The frequency of a component that is identified by a peak in the spectrum is interpolated by fitting a quadratic function to the value of the power spectrum at the peak bin and the bins either side of the peak bin. The refined frequency estimate is given by the location of the maxima of the fitted curve (Smith 2003). Other more elaborate interpolation schemes exist also (Milivojevic 2006).

3. Constant Q transform. Rather than measuring the power at equally spaced frequency intervals as the FFT does, the Constant Q transform measures power at exponentially spaced frequency intervals (Brown 1992). This means that the frequency resolution in percentage terms is equal across the spectrum.

4. Chirp Z transform. This transform utilises equally spaced frequency intervals but concentrated in a frequency band of interest, rather than across the whole spectrum (Rabiner 1972).

All of the methods above do a good job of interpolating the frequency spectrum and result in a higher resolution frequency estimate. However, analysis of a low frequency sine wave using these methods reveals that resolution is not the only issue. Indeed all of these methods yield a similar result, and have a high variance when estimating a single sinusoidal component of known frequency. In particular, for a 60Hz sinusoid over a 1024 sample analysis window, the variance in the frequency estimate for all of these techniques exceeds a semitone; consequently these techniques are insufficient for our purposes. The fundamental issue is that for a low frequency signal the peak of the power spectrum is simply not an accurate estimator of the frequency.

Phase Based Techniques

The Fourier spectrum contains more information than the just the power spectrum; additionally it contains phase information. The efficacy of exploit-

ing this phase information for the purposes of frequency estimation was famously described by Flanagan and Golden (1966) in their description of the Phase Vocoder. The essential idea is that the instantaneous frequencies of the signal are equal to the time derivatives of the phases. The values of the instantaneous frequencies plotted against the Fourier bin number is known as the Instantaneous Frequency Distribution (IFD) (Charpentier 1986).

A number of techniques have been proposed for the calculation of the IFD. The original Phase Vocoder simply advances the signal by one sample, computes another FFT, and approximates the phase derivatives by the difference in phase divided by the sample period. This is however computationally expensive as it involves computing an FFT for every sample in the signal.

Charpentier (1986) proposed utilising symmetry properties of the Fourier Transform to approximate the FFT of a window advanced by one sample from the FFT of the original window. This way one needs only calculate one FFT per analysis window, and obtains very similar results to the Phase Vocoder technique.

Puckette (1998) expands on Charpentier's work and compares the accuracy of this technique to a similar technique but where the signal is advanced by H (the hop size) samples instead of one. The accuracy of the estimates of the IFD increase with larger hop size, but there is a trade-off. The author identifies two issues:

- (a) The frequency estimate from this technique is essentially using information from a window of size $N + H$ where N is the size of the analysis window, so the time resolution of this technique decreases with larger hop size.
- (b) The estimate becomes increasingly vulnerable to mistakes in the phase-unwrapping. The measured change in phase corresponds to some integer number of cycles plus a measured fraction of a cycle. The number of whole cycles that the phase has gone through is unknown, except by virtue of some other independent estimation of the frequency.

The technique that we present in this paper, which extends this method, addresses these two issues.

To use the IFD to determine the *constituent frequencies* of a signal, one computes the spectrum, picks the peaks of the power spectrum, and then reports the values of the IFD at the bins corresponding to the peaks of the power spectrum. A further refinement to this class of techniques has been proposed by Kahawara (1999). The idea is that when plotting instantaneous frequency against frequency, where there is a true frequency component in the signal the instantaneous frequency should equal the frequency. In other words the mapping frequency \rightarrow instantaneous frequency should have a fixed point at every true frequency in the signal. Much as the discrete Fourier spectrum can be interpolated, the phase spectrum can be interpolated. Then searching for fixed points on the

interpolated IFD yields refined frequency estimates.

Masataka Goto (2004) has utilised fixed point analysis of the IFD in conjunction with a Bayesian belief framework in his predominant-F0 estimation. A number of other authors have utilised Bayesian techniques in the context of Polyphonic transcription (Cemgil 2004; Hainsworth 2004)

Gifford-Brown Technique.

The technique that we propose is essentially a hybrid of the above techniques, tailored to suit our specific needs. It is a two stage estimation technique similar to those of Charpentier (1986) and Puckette (1998), which addresses the shortcomings of these techniques by utilising a combination of IFD fixed point analysis, belief propagation, and MQ analysis (McAuley & Quatieri 1986). The algorithm is as follows:

We analyse the input signal in non-overlapping windows of 1024 samples. Internally, we hold a number of **beliefs**, consisting of a frequency, amplitude and phase value for a component that we currently believe to be sounding. The beliefs are stored in a fixed number of 'tracks' using the terminology of MQ – analysis. Each window we iterate the following procedure:

1. Perform an FFT on the (rectangularly) windowed signal.
2. Pick the significant peaks of the power spectrum
3. The frequencies of the spectral peaks are estimated using Charpentier's technique.
4. These estimates are refined using Fixed Point Analysis.
5. The refined estimates are pair-wise matched with our current **beliefs**.
6. Any peak that is more than a quarter-tone away from the closest belief is considered to be a new component. If the refined estimate of the peak is 'sensible', then we put this estimate into a free track (if any are available) and consider it to be a *tentative* new belief. We record the frequency estimate as the frequency of this belief, the power of the spectrum at this peak as the amplitude of this belief, and we calculate the phase of this belief by performing a single component Fourier transform of the window centred on the estimated frequency. We also record a *tentative* track-birth, and if there was a component previously in this track we record a track-death.
7. Any peak that is within a quarter-tone of the closest belief is considered to be a candidate for continuation of the belief. We can now use the stored phase information for this belief to perform an enhanced frequency estimate using the N-hop technique of Puckette (1998). We can do this extremely efficiently by utilising the frequency value of our belief: we calculate just a single Fourier spectral coefficient centred at our

believed frequency. We also calculate the number of whole cycles that this component should have gone through in one window based on our believed frequency. Adding to the whole number of cycles the partial phase advance (measured as the difference between this phase of the single calculated coefficient, and the stored phase of this belief) yields a total phase advance through the window which can be converted to a refined frequency estimate.

8. If the refined estimate from step 7. is 'sensible' (in this case by sensible we mean within a quarter-tone of our belief) then we consider this to be a *definite* continuation of the believed frequency. If the belief was previously *tentative* then we retrospectively mark it as *definite*, and retrospectively mark the track-birth as *definite*. We then use an average of the old belief and the new estimate of the frequency as the frequency for our belief. We use the phase value already calculated for the phase and the power of the peak as the amplitude.
9. Finally we go through our old beliefs, and for any belief that has not been continued we record a track-death, and fill the track with a new belief as needed.

The use of a two-stage estimation, firstly utilising the 1-Hop technique of Charpentier, and secondly the N-Hop technique of Puckette, circumvents the issues raised regarding the N-Hop technique. Firstly by encapsulating the information about the previous analysis window in a set of believed frequencies and phases, we increase the time resolution of the N-Hop technique to a single analysis window, at least in so far as detecting when a steady component terminates. Whilst it will still take two windows to get an accurate frequency estimate from the N-Hop technique once a new component starts, we can in the meantime use the 1-Hop estimate of the frequency during this first window. So for the first window of a new pitch our estimate is not as good as for any following windows – however we view this as acceptable especially since during the attack phase of a new note the pitch may be unstable in any case. Furthermore, the timing information regarding the beginning of a new component has a time resolution of one window, which is the desired result. Once the frequency stabilises the frequency estimate for the first window may be retrospectively adjusted if desired.

Secondly by maintaining a belief of the sounding frequency (or in the case of a new component by having a first-stage estimate of the frequency) we can alleviate the phase-unwrapping errors inherent in the N-Hop technique.

Filtering

The output of the above algorithm is a fixed number of tracks containing frequency beliefs (along with corresponding amplitude and phase), and a series of timing events for track-births and track-

deaths. The next step in the algorithm is to filter these frequencies, and associate them into notes. Here we are not trying to reproduce the notes in the audio stream as they may have been physically produced, but rather perform a sensible data-reduction that groups frequencies into salient units.

In each analysis window we aggregate frequencies that are harmonically related. Recall that this was one motivation behind the lengths we have gone to above to obtain an accurate frequency estimation for the lowest frequencies in the signal. So roughly speaking, we loop through all of our beliefs for a given window: starting with the lowest frequency and associating with it any frequencies that are within a semitone of being a harmonic of that frequency. Such a collection we will call a pitch, and associate with it the frequency of the lowest component. Then we take the remaining beliefs and iterate this procedure, yielding a number of believed pitches each window. In actuality we modify this procedure somewhat, since as described the algorithm would aggregate the fundamentals of a bass line playing a C2 with a melodic line simultaneously playing a C4 for example. Whilst it is not our intention to segregate the audio stream into distinct parts, we make some concession towards this by requiring that harmonics have no more than half the energy of the fundamental to which they are aggregated.

Having aggregated frequencies vertically into pitches, we then aggregate pitches horizontally (in time) into notes. Notes are formed by examination of the track birth and death timing information. A note is considered to start/finish at the time of the birth/death of the track containing the fundamental. The note structure consists of a start time, an end time, a pitch envelope and an amplitude envelope. The information about the notes is known in real-time with a latency of two analysis windows, however the accuracy of the resolution of the timing information for the start/end of notes is just one analysis window.

Once the notes have been formed we perform a real-time merge of any two notes whose frequencies are within a quarter-tone and where one note finishes in the same or previous window as which the second note starts. This step eliminates a great deal of erroneous discontinuities caused by noise. On the other hand it means that this algorithm will not pick up rhythmic information from repeated pitched notes. However, since the notes are of the same pitch, for a human listener to gather rhythmic information from them they must be marked in some other way, for example with attacks or articulatory information, and these markings may be picked up by an independent onset detection algorithm such as an energy based detection system, that operates in parallel with this algorithm. Note that this information is available with a latency of two analysis windows.

Finally we eliminate from consideration notes that are too short, in our case we have chosen three analysis windows as the minimum length for a note

to be validated. Consequently the algorithm as a whole has a latency of three analysis windows, though the timing information (for feeding into a beat-tracking algorithm for example) has a resolution of one analysis window.

Extracting salient features

The results from the audio analysis are in the form of 'notes' with a frequency, amplitude, and duration. When listening to this output reproduced it is clear to hear that the transcriptions are not accurate. Often pitch slides and note ghosting are misconstrued as additional notes and jumps to related frequencies a fifth or octave displaced are not uncommon. These can be filtered by ignoring short notes or by increasing the belief threshold to provide additional stability. For the purposes of auto accompaniment, these 'notes' are considered like incoming MIDI messages, their frequencies quantized to an equally tempered scale.

The stream is generated in real-time and therefore timing and metrical placement information can be significant, however, in the trials reported here we focused on harmonic organisation. Integrating beat tracking and other temporal analysis will be a topic for further research.

Improvising accompaniment

Our intention for this accompaniment was twofold: firstly to demonstrate that the polyphonic listening process was sufficiently accurate and complete that a reasonable harmonic accompaniment could be generated from the data. Secondly, to explore methods of tracking changes in the reported salient harmonic features over time.

Data gathering

Accumulating the pitch material as it arrives is a trivial process, however there are some important aspects about the data stream that need to be accounted for when generating an accompaniment. These include:

- The data is only loosely ordered
- Stream segmentation is absent in the data.
- The data may not be clean.

The loose ordering refers to the fact that while the assessment of 'note' candidates occur within a window because of the varying length of overlapping polyphonic notes the precise order of events cannot be guaranteed. Therefore the ordering of events can only be assumed as approximate.

The data is not segmented with relation to musical parts in any way, additionally at times significant overtones may appear as discrete notes. These significant overtones are usually harmonically related and consequently their salient value to the harmonic nature of the material remains relevant. In this work the input data is accumulated as a sliding windowed cluster.

Given that the pitch analysis process is not entirely accurate it is inevitable, despite filtering ef-

forts at the analysis stage, that the data will not be clean and may contain erroneous pitches. The analysis process is sufficient that these are uncommon and so the improvisational process utilises a probability-based system that amplifies the statistical significance to further minimize these errors.

Data analysis

Our accompaniment process begins by accumulating the pitches derived from the audio analysis into a windowed histogram of pitch classes. This makes clear in a simple way the harmonic tendencies of the audio input. As incoming pitches are added their pitch class weighting is increased and the whole set is normalized reducing all other weights and acting as a form of memory loss for the system. A simple accompaniment can be generated by probabilistic selection from this histogram with reasonable effectiveness, but lacks musical structure and direction. The normalised histogram will track changes in harmonic content more or less quickly depending upon the weighting assigned to incoming notes, the higher the weighting the more rapidly the histogram reflects changes in the input data. The histogram provides a weighted pitch class set as a basis for speculation about current key and/or harmonic progression that can inform the generative accompaniment processes.

The pitch range of the incoming data is simultaneously tracked by finding the maximum and minimum pitch from the last few data. The buffer size of previous pitches can be adjusted depending upon the input material. We have found that for music where higher and lower instruments play continually, that a small buffer size (less than 10 notes) is sufficient. Knowing the range of instruments can, in simple cases, allow the accompaniment to occupy complementary pitch ranges, and otherwise acts as a basis for more sophisticated arrangement decisions based on implied instrumental functions, textural density, and larger scale structural changes.

Generating accompaniment

In demonstrating our polyphonic tracking system we have used two part audio files containing and bass and melodic parts. The generated material provides harmonic material in the form of chordal and arpeggiated accompaniment.

Pitch material for these parts is derived from a combination of direct selection from the pitch histogram within the dynamically tracked range, and from chordal and key estimations statistically derived from the pitch histogram. Compositional considerations include being more conservative about pitch choice on prominent beats and allowing passing notes to be less constrained by either the pitch class set or harmonic estimations. Textural density and accompaniment range is varied according to the distance between bass and melody lines.

At present rhythmic, and dynamic aspects of the accompaniment are unrelated to the tracking

data, but we plan to extend our research to include estimation of salient aspects of these musical features in the future providing appropriate data that can influence these aspects of the generated material.

Practical results

The resulting improvised accompaniment appeared to the authors' ears to demonstrate a reasonable degree of awareness of the harmonic context of the audio stream, though a number of 'outside' notes were also generated. An inspection of the notes produced by the listening algorithm suggested that some of the notes perceived were erroneous, particularly in the lower frequencies. The erroneous pitches tended to be short lived however, so a secondary filtering process in the improvisation algorithm yielded a much more 'inside' accompaniment.

A parameter of the improvisation algorithm is the rate at which new pitches are introduced into the pitch-set histogram, and the rate at which old pitches are forgotten. There is a balance to be struck between responsiveness and stability. If new pitches are allowed to dominate the pitch-set too quickly the system is readily thrown from the tonal centre by an outside note. This problem is exacerbated by having noisy input data. On the other hand if old pitches are maintained for too long then the system does not readily adapt to harmonic changes. In practice we found that by altering this parameter we could obtain an improvisation that sounded acceptably in key, and responded to harmonic changes within a bar.

Conclusion

We have outlined a process for the real-time extraction of salient features from a polyphonic audio stream, and discussed the application of this system to a generative accompaniment engine. This process extends and improves upon a variety of pitch tracking techniques, and has been demonstrated to extract the harmonic context from a two part musical excerpt.

The process also yields timing information, which may be supplied to a beat induction algorithm. Our intention is extend the generative accompaniment engine to also produce generative rhythmic output.

We speculate that this process may be refined by virtue of a feedback loop with a beat-induction system. By giving greater weight to candidate notes that begin on a beat erroneous pitches may be filtered more effectively.

References

- Bello, J. P., Duxbury, C., Davies, M., & Sandler, M. (2004). On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain. *IEEE Signal Processing Letters*, 11(6), 553 - 556.

- Biles, J. A. (1994). GenJam: A genetic algorithm for generating jazz solos. *International Computer Music Conference*, San Francisco.
- Biles, J. A. (2002). GenJam: Evolution of a jazz improviser. In P. J. Bentley & D. W. Corne (Eds.), *Creative Evolutionary Systems* (pp. 165-188). San Francisco: Morgan Kaufmann.
- Brown, J., & Puckette, M. (1992). An efficient algorithm for the calculation of a constant Q transform. *Journal of the Acoustical Society of America*, 92, 1394-1402.
- Brown, J., & Puckette, M. (1993). A high-resolution fundamental frequency determination based on phase changes of the Fourier Transform. *Journal of the Acoustical Society of America*, 94(2), 662-667.
- Cemgil, A. T. (2004). *Bayesian Music Transcription*. Radboud University of Nijmegen.
- Charpentier, F. J. (1986). Pitch Detection Using the Short-Term Phase Spectrum. *International Conference on Acoustics, Speech and Signal Processing*, New York.
- Collins, N. (2006). *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. Cambridge, Cambridge.
- Cope, D. a. H., D. R. (2001). *Virtual Music: Computer Synthesis of Musical Style*. Cambridge, Mass: MIT Press.
- Dean, R. (2003). *Hyperimprovisation: Computer-Interactive Sound Improvisation*: Middleton: A-R Editions.
- Flanagan, J. L., & Golden, R. M. (1966). Phase Vocoder. *Bell Systems Tech Journal*, 45, 1493-1509.
- Gifford, T., & Brown, A. (2006). The Ambidrum: Automated Rhythmic Improvisation. *Australasian Computer Music Conference*, Adelaide.
- Goebel, W., & Parncutt, R. (2001). Perception of onset asynchronies: Acoustic piano versus synthesized complex versus pure tones. *Meeting of the Society for Music Perception and Cognition*, Kingston, Canada
- Goto, M. (2004). A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communications*, 43, 311-329.
- Hainsworth, S. (2004). *Techniques for the Automated Analysis of Musical Audio*. University of Cambridge, Cambridge.
- Hawkins, J., & Blakeslee, S. (2004). *On Intelligence*. New York: Times Books.
- In the Chair*. (2006) In the Chair Pty. Ltd. Adelaide. <http://www.inthechair.com/>
- Jackendoff, R. (1992). *Languages of the Mind: essays on mental representation*. Cambridge, MASS: MIT Press.
- Jackendoff, R. (2002). *Foundations of Language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
- Kahawara, H., Katayose, H., de Cheveigne, A., & Patterson, R. D. (1999). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. *European Conference on Speech Communication and Technology*.
- Levenson, T. (1994). *Measure for Measure: a Musical History of Science*. New York: Touchstone.
- McAuley, R. J., & Quatieri, T. F. (1986). Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4), 774-754.
- Milivojevic, Z., Mirkovic, M., & Milivojevic, S. (2006). An Estimate of Fundamental Frequency Using PCC Interpolation - Comparative Analysis. *Information Technology and Control*, 35(2), 131-136.
- Noll, M. (1969). Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum, and a Maximum Likelihood Estimate. *Symposium on Computer Processing in Communications*, Polytechnic Institute of Brooklyn.
- Pierce, J. R. (1999). Introduction to Pitch Perception. In P. R. Cook (Ed.), *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics* (pp. 57-70). Cambridge, MASS: MIT Press.
- Puckette, M., & Brown, J. (1998). Accuracy of Frequency Estimates Using the Phase Vocoder. *IEEE Transactions on Speech and Audio Processing*, 6(2), 166-176.
- Rabiner, L. R., Schafer, R. W., & Rader, C. M. (1972). The Chirp z-Transform. In L. R. Rabiner & C. M. Rader (Eds.), *Digital Signal Processing* (pp. 322-328): IEEE Press.
- Rowe, R. (1993). *Interactive Music Systems: Machine Listening and Composing*. Cambridge, MASS: MIT Press.
- Scheirer, E. (1999). Towards Music Understanding Without Separation: Segmenting Music with Correlogram Comodulation. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paliz, New York.
- Shepard, R. (1999). Stream Segregation and Ambiguity in Audition. In P. R. Cook (Ed.), *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics* (pp. 117-127). Cambridge, MASS: MIT Press.
- Sloboda, J. A. (1988). *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition*. Oxford: Clarendon Press.
- Smart Music*. (2005) MakeMusic Inc. Eden Prairie, MN. <http://www.smartmusic.com/>
- Smith, J. O. (2003). *Mathematics of the Discrete Fourier Transform (DFT), with Music and Audio Applications*: W3K Publishing.
- Temperley, D. (2001). *The Cognition of Basical Musical Structures*. Cambridge, MASS: MIT Press.

Winkler, T. (1998). *Composing Interactive Music*.
Cambridge, MASS: MIT Press.