ON TECHNIQUES FOR CONTENT-BASED VISUAL ANNOTATION TO AID INTRA-TRACK MUSIC NAVIGATION

Gavin Wood
University of York
York YO10 5DD
United Kingdom
gav@cs.york.ac.uk

Simon O'Keefe
University of York
York YO10 5DD
United Kingdom
sok@cs.york.ac.uk

ABSTRACT

Despite the fact that people are increasingly listening to music electronically, the core interface of the common tools for playing the music have had very little improvement. In particular the tools for intra-track navigation have remained basically static, not taking advantage of recent studies into the field of audio jisting, summarising and segmentation.

We introduce a novel mechanism for musical audio linear summarisation and modify a widely used open source media player to utilise several music information retrieval techniques directly in the graphical user interface. With a broad range of music, we provide a qualitative discussion on several techniques used for content-based music information retrieval and perform quantitative investigation to their usefulness.

1 INTRODUCTION

In recent years the techniques for content based analysis of musical audio have improved dramatically. Moore's law continues steadily to provide software with ever-greater resources with respect to processing power, and the extra storage available for media has meant that we are able to store our entire music collection for digital playback. Graphical interfaces to media players have become more elaborate and most mainstream software now supports some sort of visualisation of the music as it plays.[1]

In the original generation of the graphical media player, a typical user interface feature would be the "time bar". This allowed the user to visualise how far through the current track they were, in relation to the length of the track. This was, in many ways, similar to the progress bar in order to show the user how much of a particular task is completed, with the exception that time bars may be used to directly navigate through a track by clicking some way along it. The player would resume playing at the corresponding point through the track. However, as a navigation tool its use is limited due to the fact that the user had to know in advance the approximate place on the bar to deliver the wanted moment of the track.

In order to improve the usefulness of this "time bar", some extra information must be added to it, providing the user with some visual cues. This allows the user to better guess which point along it maps to the particular moment they are trying to find. Many studies (e.g., Bocker et al., 1986) have shown visual cues to be a simple and effective means to convey information to the user without confusing a novice or distracting one already familiar. We call this visual annotation a "mood bar", referring to the varying shades to depict the music content.

There is work aplenty in the field of IR with respect to analysing and classifying individual segments of musical audio, perhaps with a view to archiving, retrieval, grouping, large-scale exploration and browsing. Extensive work has been carried out into forming the user interface to deal with this functionality. Comparatively little has been done to the interface once the necessary segment has been located and is ready to be played. One might assume that once the user has their segment of data—be it a music track, a monolithic compilation (e.g. live performance) or perhaps a radio broadcast—they are happy to have it play throughout.

The concept of *user-directed* navigation is essential to this work. We set out not to produce a visual annotation by which some (however small) absolute truth may be gained from one single sample. Instead we take a more holistic approach and free ourselves from the constraint that the annotation must mean something absolute and concrete. We allow our visualisation to take on any abstract form, and judge performance as to what, as a human, we are able to ascertain from the final depiction. We go on to measure how *useful* these depictions are for searching tasks with a broad range of music.

The task is an acute *acid test*; i.e. we give the participants an absolute minimum of learning time. As such the results will heavily favour the annotation methods with more obvious visual cues to those with a more complex visual representation. This is because we wish to test realistic casual usage; people should not have to suffer a significant learning curve to use a media player. In particular

[1] Though in many cases the correspondence between audio and video leaves much to be desired.

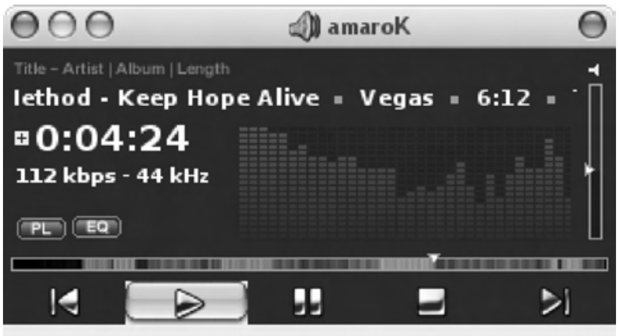

amaroK
Title – Artist | Album | Length
lethod - Keep Hope Alive ▪ Vegas ▪ 6:12 ▪
0:04:24
112 kbps - 44 kHz
PL
EQ

Figure 1: The amaroK music player with the *moodbar* operational.

we test whether the addition of colour improves performance when the participants have no prior experience of the new interface.

1.1 Related Work

Little work has been openly published specifically regarding intra-track navigation. The most connected work to that presented here is by Tzanetakis and Cook (2000a), who have discussed the Marsyas augmented sound editor. This is a basic sound editor that can "colour-in" the edited waveform according to some particular acoustic characteristics. Apparently, this line of research was not continued any further since it remains around only in the original niche application (the Marsyas augmented sound editor). The technology presented here is comparative; however we evaluate it in more depth and with user studies. We also utilise several different methods for calculating the colour and present a novel method. Tzanetakis and Cook (2000b) later provided some insight into segmentation and possible methods in their article on Marsyas.

There is significant literature in the related arena of context based segmentation of audio; Raphael (1999) has presented a segmentation method with Markov models. Foote (1999), then later in Foote and Cooper (2003) and Cooper and Foote (2003), present a mechanism for calculating musical novelty, with a view to segmentation and jisting. The work doesn't go so far as to quantitatively evaluate the usefulness, instead presenting the techniques and discussing the output.

Couprie (2004) has given a most interesting discussion on possible intuitive graphical representations of music. He suggests the display is made of discrete elements rather than a continuous form (that might be easier when working with source audio). The discussion does argue well that navigational aids are helpful in numerous situations. Brazil et al. (2002) discussed intra-collection browsing through track features as did Tzanetakis (2003) for musical audio. A similar problem was approached by Blackburn and DeRoure (1998) for MIDI navigation between tracks using pitch contours.

2 USER-INTERFACE DESIGN

Following the principle of least surprise, we changed the playback interface as little as possible. The media player we set about to augment, amaroK (2005), already provided a highly intuitive interface with the now-standard track slider bar. In amaroK's case a triangular pointer scrolls across the top of the bar denoting the current position of the player through the track. The only change we made to the interface was to have the internal portion of the bar coloured (with vertical lines) according to some analysis metric. The colour changes along the x-axis, which represents time. Because we allowed ourselves use of colour each point on the x axis corresponds to a 3D value using the RGB components of the colour. As before the user is free to click anywhere on the bar to warp the player to the corresponding position in the song.

3 ANALYSIS TECHNIQUES

Several main techniques are used in order to populate the "mood bar". These techniques can be split into two groups—those that result in only one value and those that result in three values. Those that result in one value were transformed into a colour by using it as the luminosity of a shade of gray. Those giving three values were transformed into a colour by assigning each value to either the red, green or blue component.

The spectra were calculated by using a series of FFTs over the signal. The window size used was 1024 samples, with a 50% overlap. With the input signal being at the CD standard 44100Hz, this puts the lowest frequency to be detected at around 43Hz with windows being around 11ms apart. The stereo signals were first downmixed into mono, to prevent any problematic stereo seperation effects.

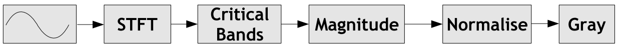
A psychoacoustic variant of the mechanisms was used, where the spectra were first summed into the critical bands on the Bark scale. This significantly cuts down on the computation cost in many areas since the 512 bands of the FFT output is reduced to only 24 critical bands. Initial experimentation backed up by previous studies (such as Wood and O'Keefe, 2003) showed that it made little discernable difference to the final performance.

Also, the full gamut of output for the track was normalised before being changed into a colour. The normalisation technique used was a simple min/max stretch given by:

$$value_{normalised} \equiv \frac{(value - value_{min})}{value_{max} - value_{min}}$$

3.1 Spectral Magnitude

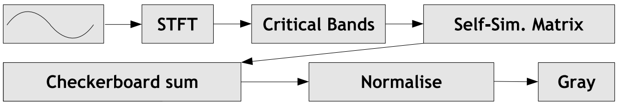
We used the spectral magnitude calculation to provide us with a 1 dimensional representation of the audio track:



Because of its simplicity it was used as a benchmark to which other techniques could be compared.

3.2 Novelty

Foote (1999) first detailed the "novelty" score: It provides a value determined by the cross dissimilarity of the portions of signal before and after the moment in time, as shown in the workflow diagram:



It relies upon a prior abstraction of the signal known as a self-similarity matrix, which is calculated simply by evaluating the similarity of the signal to itself at varying intervals (given by $x - y$). The similarity between the two signals is the cosine of the angle between the two spectra when expressed as vectors, as suggested in the literature.

$$D_C(\mathbf{S}_x, \mathbf{S}_y) \equiv \frac{\mathbf{S}_x \bullet \mathbf{S}_y}{\|\mathbf{S}_x\| \, \|\mathbf{S}_y\|}$$

where $\mathbf{S}$ is the spectrum in question.

A “checkerboard” weighting is applied to the matrix with a Guassian taper weighting and the values summed. This is the novelty score of the moment at the centre of the input series of spectra. The spectra were calculated in the same manner as the technique above. The kernel (with the Guassian taper) is given by:

$$K(x,y) \equiv \left\{ \begin{array}{l} G(x,y), (x > 0) = (y > 0) \\ -G(x,y), (x > 0) \neq (y > 0) \end{array} \right.$$

where

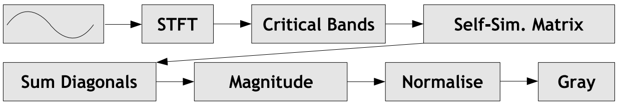
$$G(x,y) \equiv Gaussian(\| \left(\frac{2x}{s}, \frac{2y}{s} \right) \|)$$

Where x and y both fall in the range $[-\frac{s}{2}, \frac{s}{2}]$ and the kernel matrix is of width s.

The size of the self-similarity matrix and accompanying checkerboard kernel were experimented with and qualitatively evaluated. We found a value of around 128 spectra (1.49 seconds) provided a good balance between time precision and larger scale feature presentation. Figure 4 demonstrates the differences in matrix size on several tracks.

3.3 Rhythm Magnitude

The rhythm magnitude is a novel technique to deliver the “rhythmicity” of audio at a particular point. It is calculated by using the rhythm spectrum (also known as beat spectrum) as a vector and taking its magnitude:



We use the algorithm proposed by Foote (1999) for calculating the rhythm spectra, which involves populating a self-similarity matrix and summing across the superdiagonals.

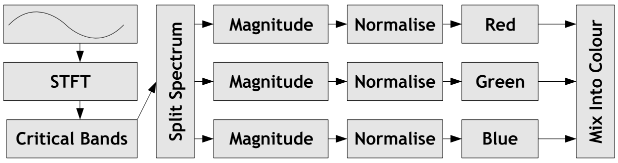
$$B(l) \equiv \sum_{k=0}^{s-l} \mathbf{M}(k, k+l)$$

where s is the size of the self-similarity matrix M. Tzanetakis and Cook (2000a) have discussed other techniques such as the beat histogram.

A higher value (i.e. lighter shade) is caused by having more power in the rhythm spectrum. A higher lag-correlation denotes a stronger rhythm which will cause a lighter shade to be output. If there is little correlation, or it is compromised between two successive and unique rhythms then it should have a lower overall power and thus be darker in shade.

3.4 Bandwise Spectral Magnitude

We devised the spectral magnitude ratio metric as a novel way to introduce colour into the mood bar. Part of the precessing pipeline is split into three separate channels for red, green and blue respectively. The point at which the split takes place is directly after the Bark critical banding; here we take the 24 bands and split them into 3 8-band sub spectra. Each spectra is then used as an 8-dimensional vector to which the magnitude is calculated (as the Euclidean distance from 0). These magnitudes are normalised across the track and used as each of a red, green and blue component of the final colour:



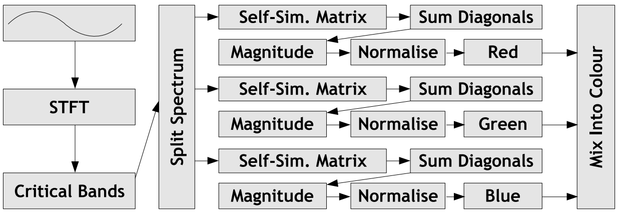
The hue of the colour should therefore be an indication of the “brightness” of the sound. A more red hue will denote more power in the low frequency portion. A more green hue denotes more mid-range content and a bluer hue would denote high-range. The lightness denotes overall power as in the standard spectral magnitude measurement. Finally the saturation of the colour would denote the balance of power in the spectrum. A spectrum that contains much of its power in a particular place should give rise to a very saturated colour, since it is likely the power will all be engulfed into one of the three subspectra.

3.5 Bandwise Rhythm Magnitude

As before, this is a extension to the standard rhythm magnitude technique done to provide colour. The output of the critical banding is split into three subspectra, a rhythm magnitude for each one is found. Each are normalised individually and used as their corresponding red/green/blue component in the final colour.

The brightness of the colour relates to the simplicity of the rhythm at that point, whereas the hue describes where in the spectrum that simplicity lies; if the rhythmic simplicity is most affected by voices in the upper part of the frequency spectrum, the hue will be more purple, in the lower part and the hue will be more “orangey”.

The work flow diagram for the bandwise rhythm magnitude technique follows:



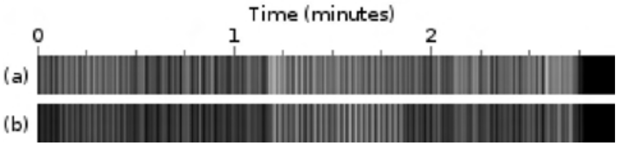


Figure 2: The track *Green Onions* by *Booker T. and the MG's* displayed with the metrics (a) spectral magnitude and (b) bandwise spectral magnitude.

4 IMPLEMENTATION

All signal processing code was implemented under the open source Exscalibar framework for audio signal information retrieval using the Geddei library. As such it took only one hour for the signal processing part to be programmed to completion. Due to the transparent and efficient concurrency that the Geddei provides, all code developed was efficient and concurrent ready to take maximum advantage of dual-core processors, hyper-threading, SMP and other forms of hardware parallelism. More information can be found at the project website (Exscalibar, 2005).

The open source media player amaroK was used as the mainstream media player on which to add the functionality. It is an advanced music player for use under the Unix desktop environment KDE. Despite being relatively young, it is currently used by many in the Linux community, having had around 50,000 downloads in total from its site. It was chosen due to its usable interface, its emphasis on new technology and the quality of the source code. The code changes necessary to make it work alongside the processing software and display the mood bar took around two hours in total.

In many instances the data generated from the IR techniques would not fit completely into the limited space for the slider bar in amaroK. In these cases all points that would be in any given pixel's space were simply averaged; this was then used for that pixel's colour.

5 EVALUATION

5.1 Discussion

5.1.1 Bandwise Spectral Magnitude

In many instances the addition of colour to the spectral magnitude display appears to much better describe the music and allow more and easier discerning of particular features in the track. The well known jazz track *Green Onions* is shown in figure 2. The bandwise variant (b) clearly identifies where the funk guitar can be heard in three parts (30s, 1:10 to 1.50 and 2:35 onwards) where the original version (a) does not. This is represented by the thin highlights (of a turquoise hue), especially evident between 1:30 and 1:50, where the guitar steps up a key.

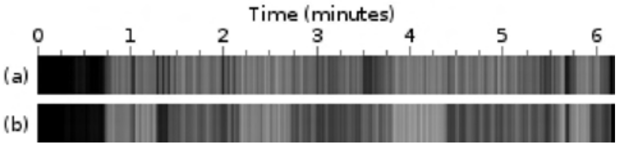


Figure 3: The track *Keep Hope Alive* by *The Crystal Method* displayed with the metrics (a) spectral magnitude and (b) bandwise spectral magnitude.

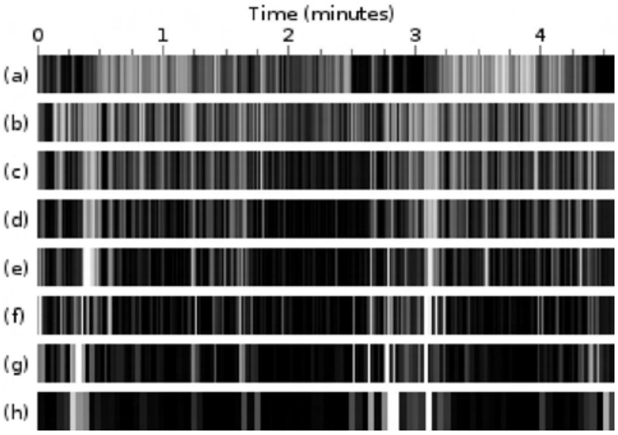


Figure 4: The track *Fantasia on Greensleeves* by *Solisti di Zagreb* displayed with the metrics (a) spectral magnitude and novelty with a kernel size of (b) 92ms, (c) 185ms, (d) 371ms, (e) 743ms, (f) 1.49s, (g) 2.97s and (h) 5.94s.

The track *Keep Hope Alive*, shown in figure 3 gives an even greater demonstration of the difference between the bandwise magnitude and the basic version. For those unfamiliar with this particular track, it is a complex piece of progressive techno music, which for the most part switches between two moods (at approximately 0:48, 1:21, 2:08, 2:44, 3:35, 4:23 and 5:41). The track is generally "loud", regardless of the mood it is in (the quiet bits are found when it changes between them).

Without the use of colour, as in the original variant, the only parts of the track that are identifyable are the bridges where the loudness dies down for some time; these appear as darker spots, such as at 2:00, 3:30 and 5:40. The moods themselves are virtually indistinguishable. The bandwise version (b) is able to distinguish between both portions allowing the user to better separate the parts; even in this desaturated image, the difference is pronounced.

5.1.2 Novelty

The novelty algorithm, due to the fact it determines a value for each particular moment by using some number of seconds' data around it, means that it will have less temporal precision. The amount of precision can be controlled by the kernel size. Figure 4 shows a rendition of the well known theme *Greensleeves* with varying kernel sizes. As is discussed later, the processing required for the very large kernel sizes (i.e. over 128) results considerably more processing time. For instance (g) and (h) take respectively

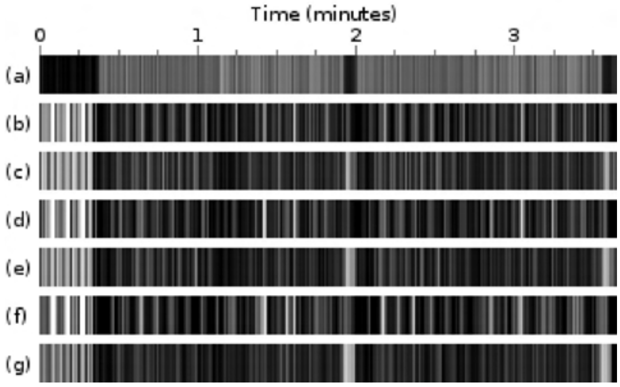


Figure 5: The track *Time is the Enemy* by *Quantic* displayed with the metrics (a) spectral magnitude and rhythm magnitude interlaced with bandwise rhythm magnitude with a rhythm spectrum formed from a matrix of size (b/c) 371ms, (d/e) 743ms, (f/g) 1.49s.

33% and 100% longer to compute than (f).

The novelty output is visibly different to the spectral magnitude, since it is one level of indirection away; rather than showing the track directly and allowing the user to determine when the metric changes enough to denote a feature, it instead shows the changes directly, essentially providing a differential view.

The performance of the metric is however quite interesting; the start of the track, which takes place between 20 to 30 seconds into the track is detected at drastically different times by different kernel widths. (g) and (h) undershoot and have their main strokes at around 23 seconds in. (f) is less clear and has several strokes around the time (accurate, if not precise). The second start to the main theme at 3:05 to 3:10 is well depicted by (e) to (h). The change in theme at 1:16 and 3:07 is also clear in (e) to (h). (f) to (h) all picked up on the key change at 4:00. We finally picked (f) as a good compromise between visual clarity and processing needed.

5.1.3 Rhythm Magnitude

Like the novelty algorithm, we can already deduce that the rhythm magnitude metric will have less time precision, since each value it produces is based upon a number of spectra from either side of the moment in question. The number of spectra used, and thus the imprecision of the metric is equivalent to the number of bands of the rhythm spectrum (or the cardinality of the vector we measure). Determining the optimum size of the spectrum is rather a black art; a smaller size results in better time precision and less processing. A larger size should allow higher-level features to be captured.

Figure 5 depicts the track *Time is the Enemy*, a grandiose, if slightly repetitive piece of electronic music. In the figure, the top row (a) is the basic spectral magnitude of the track, and we can easily pick out the two sections of the track with the gap at 1:56 to 2:01. Aside from the start and finish, little else is visible.

Looking at the rhythm magnitude depictions (b), (d) and (f), we can see that the start of the track isn't nearly as uninteresting as (a) makes out. Clearly visible are the four echoey repetitions of a theme. Arguably we could also make out the frequency and strength of the vertical bars changes in each of the two halves. This can be seen at approximately 1:07 and then in the other half at about 2:45; these relate to the change in theme each half goes through.

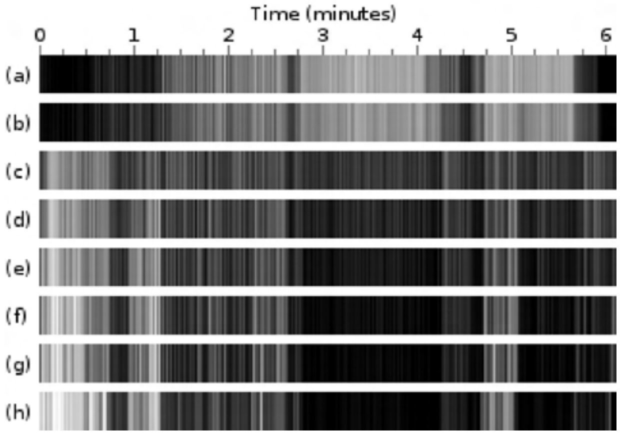


Figure 6: The track *They're Hanging Me Tonight* by *Red Snapper* displayed with the metrics (a) spectral magnitude, (b) bandwise spectral magnitude, and bandwise rhythm magnitude with rhythm spectrum formed from a matrix of size (c) 92ms, (d) 185ms, (e) 371ms, (f) 743ms, (g) 1.49s and (h) 2.97s.

Comparing (b) to (f), between which the size of the rhythm spectrum increased by a factor of 4, we can see roughly the same features are visible, though in (f) they appear to be better defined, with less overall noise; it would appear that the time precision (with this track, at least) is negligible.

5.1.4 Bandwise Rhythm Magnitude

In figure 5 we can consider the differences between the initial method and the bandwise variant. All the bandwise variants are better able to depict the bridge at 1:55 to 2:00 and clearly distinguish between the initial 20 seconds and elsewhere in the track with the extreme change of hue "orangey" green to cyan, red and blue. However (e) with a matrix size of 743ms appears on the whole the best, pushing cyan strokes into the first portion of each half and then magenta strokes into the latter portion.

Figure 6 depicts the track *They're Hanging Me Tonight*, a short electronic-acoustic symphonic (or perhaps cacophonic) piece. There are numerous instruments in the track, and sampling is used considerably to make the track quite complex listening. The basic spectral magnitude (a) doesn't really show a lot of information other than the rough start and finish times (around 1:15, with breaks at 2:40 and 4:10, finishing at 5:40).

The bandwise spectral magnitude (b) helps distinguish a little more; we see a large cyan bar at 4:05 where a guitar takes over for 10 seconds before the mellower bridge is reached. However due to the fact the track is loud at most parts anyway, despite the music differences, we see a largely monotonic picture.

When we utilise the bandwise rhythm magnitude, we

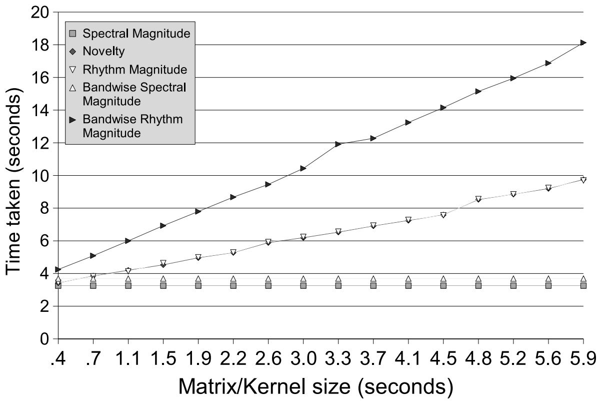


Figure 7: The total CPU time spent analysing the audio data in order to provide the relevant visualisation. Timing conducted on Intel Pentium-M 1.7GHz system with 512MB RAM on 171s of audio.

see a different picture. In this instance two portions of the track that used to be quite indistinguishable (from 1:15 to 2:35 and then from 2:45 to around 4:00) are now readily identifyable as green and dark red respectively. Even more, the portions can be re-identified later on in the track at 4:45 to 5:00 (green) and 5:00 to 5:35 (dark red). The initial 45 seconds of the track, which appears as a constant very dark grey in the both (a) and (b) has a bright pink/yellow/cyan hue which changes to dull green, and then again to dark purple. The first three hue changes don't appear to correspond to anything useful in the music, however the latter two changes correspond precisely to the addition one and then another repeating samples.

Varying the matrix size between 100ms to 1.5s gives somewhat different output. The visual features become clearer and less noisy, though problematic "red herrings" (like the meaningless hue changes early on in the track) also become better defined. The processing time also increases significantly as is shown later. A good compromise appears to be with a matrix size of 64 or 128 samples, which correspond to 743ms and 1.49s respectively.

5.2 Computational Performance

From figure 7 we can see that all methods that rely upon a self-similarity kernel are $O(n)$ where n is the size of the kernel. The bandwise rhythm magnitude takes around twice as long to compute for a given kernel size than either of the other two non-bandwise matrix-based methods. The non-matrix based measures were shown as a baseline only.

We can see that with modern hardware, and choosing a reasonable size of matrix, computing the annotation for a given audio track would be quite trivial. This could be done either on-demand, having the annotation computed on play and displayed a few seconds into the track or perhaps precomputed (as we did in this study) so that the annotation is stored ready for immediate use.

5.3 User Study

For the user study, we attempted to best simulate practical and real-world conditions, rather than previous experiments where we focused mainly on mathematical tractability. We initially selected five tracks from a reasonably broad range of music. The tracks were selected to give a good range of different types of music and of different difficulties of problem. Tracks were chosen for their interesting features that would best test the systems. Mood changes, both subtle and blunt, instrument changes, vocal changes and rhythm diversity are among the features we attempted to utilise to best examine the systems. Table 1 shows the tracks that were chosen.

Track	Genre	Times
(1) D. McMurray *Walk in the Night* Reasonable, consistant beat structure and loudness. Minimally defined transitions.	Jazz	14
(2) Muse *Plug In Baby* Simple rock ballad with clear verse/chorus structure.	Rock	9
(3) Shivaree *Goodnight Moon* Fluid pop song with little beat structure and hazy verse/chorus structure.	Pop	8
(4) Plaid *Prague Radio* Structurally complex, highly dynamic with multiple moods and well defined beats.	Abstract	6
(5) Crystal Method *Keep Hope Alive* Structurally simple, well defined beats, consistently loud, few moods.	Dance	13

Table 1: The five tracks chosen for the user study, their respective genres and the number of significant changes.

5.3.1 Method

We formed a base truth about our data by allowing a "music lover" to dictate where in each track the main musical changes took place. Around 7 such points were allocated to each track. To prevent the effects of suggestion the individual who was given this task was not previously subject to any annotation of the tracks in question. When listening to the tracks, no visualisation at all was provided.

With our base truth established, we conducted the study proper. We conducted the study with 18 subjects, each subject was given five trials—a trial for each of the five tracks. We rotated through each of the five analysis techniques and a control with no annotation.

Each subject was given an initial period of training (some required more than others), until they felt familiar with the controls of the player. Aside from getting to grips with the "look and feel" of the application, they were given no specific information on the mood bar algorithms. For each trial, the subject was given 60 seconds with the player incorporating the given analysis technique with the track loaded. They were allowed to skip back and forth through the track at will, and could utilise the mood bar as they saw fit. Their task was to find each of the times where the music "changed" most. When the minute was up, the music was stopped and they had to finish writing.

5.3.2 User Commentry

The overall feeling of those interviewed was that they preferred the spectral magnitude visualisations over any of the others. Having utilised all five of the methods, many also indicated that they intuitively related the intensity of a point with the loudness at that point. Some went on to suggest that they would then intuitively relate the colour of a point with any instruments playing at that point. Only one candidate suggested that brighter parts in the visualisation might mean increased dynamics and otherwise less constancy.

The standard rhythm magnitude measure as well as the novelty measure were generally disliked. Specific comments were "daunting" and "less predictable". The general feeling was that differential measurement (ala novelty) was unsuitable for intuitive learning; people expected to see "chunks" of similar sounding portions of time, rather than specific points at which the music changed. Those who commented felt that the rhythm magnitude measure simply looked overly populated and excessively contrasting, and thus determined it too "daunting" for general use.

Cosmetically almost all people preferred colour over monochromatic visualisations (one even went so far as to say it was "pretty"). The majority of those suggested that they found colour to be the better visualisation in respect to usability also. They found it "easier to distinguish", and "more informative". The opinion of colour in the bandwise rhythm magnitude was somewhat more divided. While nobody made it out to be worse that the mono-chrome variant, most favoured the look of the bandwise spectral magnitude, finding the rhythm magnitude less well defined.

As for usability and comfort, the participants were quite polarised on their opinions as to whether adding colour was more helpful in the experiment. While some decided that it gave them more information and thus was more useful, others felt the addition of colour increased the learning curve too much. Most went on to suggest that perhaps, given enough time to learn, the colour might be better eventually anyway.

One participant suggested the three colour components used to create the colour form the low, mid and high portions of the spectrum be switched. Apparently they expected redder hues to relate to "warmer" (i.e. bassier) sounds while bluer hues related to harsher, sharper (presumably higher) sounds.

5.3.3 Quantitative Results

The score for each candidate was computed simply by summing the number of times they gave which fell within 3 seconds of a time from our ground-truth. In order to prevent the results for those that were better overall from biasing the general trend, every candidates score was zero-mean normalised. The means and standard deviations of the scores for each analysis method taken over all the trials conducted are presented in figure 8.

From the figure we can see that the best two methods were clearly the two based upon the spectral magnitude. Their means are far too close to distinguish which was actually better though the bandwise variant had a higher variance suggesting that its use was somewhat changeable depending upon the person using it and/or the type of track being used upon.

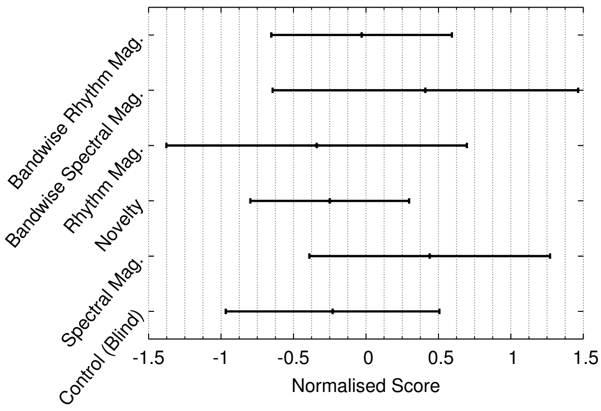


Figure 8: Mean and σ of the score of each analysis method.

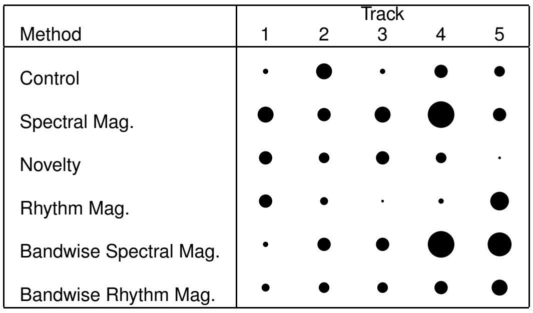


Figure 9: Method results on a track by track basis. Larger is higher score.

The bandwise rhythm magnitude far outperforms the rhythm magnitude method, increasing in both mean score and constancy. The novelty method was barely different to the control (blind) trials, and the basic rhythm magnitude slightly worse. This suggests that the false positives provided by these annotations was highly detrimental to their overall usefullness, and in the case of rhythm magnitude possibly caused more harm than good.

Though the mean of the bandwise rhythm magnitude is somewhat higher than that of the control, the variances are so high that the results lose statistical significance. Further tests would need to be conducted to determine what degree of usefullness, if any, the metric provided over the control.

From figure 9 we can see matrix showing how each method performed on each track, allowing us to compare between specific tracks and methods. We can determine that the bandwise spectral magnitude method is most problematic with track number 1, the Jazz piece. Perhaps the most surprising sign is the performance of track 2, the rock music; the performance is similar across methods, though the control actually comes slightly ahead of all the others. The two rhythm magnitude methods do reasonably well at track 5, the abstract piece, though their performance is below par elsewhere.

Interestingly the bandwise variation of the spectral magnitude causes the (quite uniform) jazz piece (1) to be even less distinguishable and the cleaner, better defined dance piece (5) to be more distinguishable. This is not surprising; while the addition of colour to a track that is musically dynamic and heavy will help further define segmentations, it may easily hinder a track whose segmentations are questionable and ill-defined by providing the user with even more false-positive cues.

6 CONCLUSION

We demonstrated that automatic, content-based visual annotation, in general, makes a positive addition to modern music playing tools. We presented findings, through a user study, that people find such visual annotations both usable and aesthetically pleasing. We also found that people were able to immediately utilise the visualisation with an absolute minimum amount of training.

We demonstrated several annotations, and found each had a particular niche under which it worked reasonably well. We found the spectral magnitude and bandwise variant to be the overall winners, for their consistant performance. The bandwise variant was declared favourite for its visual charm as well as general performance. We showed that the addition of the novel *bandwise* technique for introducing colour to two of the methods helps under many circumstances with quantitative evidence, and that users typically prefer to use the colour annotation.

This study didn't take into account various accessibility problems, not least colour blindness. However it would seem unlikely that colour blindness in itself would have such a detrimental impact as to render the colour variants worse than the original monochromatic versions.

7 FURTHER WORK

In this study only a quick and simple test was carried out giving the user a minimum of time to learn. It is unclear whether a greater learning time would change the relative performances of the metrics.

With proper standardisation such a technique could be used to "jist" music tracks at browsing in music stores. Such a "fingerprint" may describe music adequately enough to allow a potential purchaser to determine their interest in a record. Initially, electronic music players such as amaroK could have such a fingerprint added to their playlist for each track, giving the user a visual cue and providing an automatic iconification a music track. In theory the fingerprint information could be precomputed and embedded into the digital file itself. Enough technologies exist to encapsulate such metadata inside a track such as the Ogg Vorbis comment system or the MP3 id3v2 tag system.

These techniques, or others like them could easily be combined with machine learning algorithms to convert from the continuous form presented here to a discretely annotated form with specific semantics, in an automatic version of the work by (Couprie, 2004).

In so far as the bandwise mechanism was implemented, this study conducted only a simple 3-way fair division of the critical bands. Initial experimentation suggests that uneven division could significantly improve the fidelity of the resultant visualisation. Furthermore, such a graphic has the potential to show more information; varying the width of the stripes would allow further metrics to be encoded.

REFERENCES

amaroK. amarok.kde.org/, 2005. URL `http://amarok.kde.org`.

S. Blackburn and D. DeRoure. A tool for content based navigation of music. In *Proc. ACM Multimedia*. ACM, Bristol, UK, 1998. ISBN 1-58113-036-8.

H. D. Bocker, G. Fischer, and H. Nieper. The enhancement of understanding through visual representations. In *CHI '86: Proc. SIGCHI conference on Human factors in computing systems*, pages 44–50, New York, NY, USA, 1986. ACM Press. ISBN 0-89791-180-6.

E. Brazil, M. Fernstrom, G. Tzanetakis, and P. R. Cook. Enhancing sonic browsing using audio information retrieval. In *Proc. International Conference on Auditory Display*. ICAD, 2002.

M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *Proc. ACM Multimedia*, pages 364–373. ACM, Berkeley, CA, 2003.

P. Couprie. Graphical representation: An analytical and publication tool for electroaccoustic music. In *Organised Sound*, volume 9, pages 109–113. Cambridge University Press, 2004.

Exscalibar. exscalibar.sf.net/, 2005. URL `http://exscalibar.sf.net`.

J. Foote. Methods for the automatic analysis of music and audio. Technical report, 1999. URL `citeseer.nj.nec.com/foote99methods.html`.

J. Foote and M. Cooper. Media segmentation using self-similarity decomposition. In *Proc. SPIE Storage and Retrieval for Multimedia Databases*, volume 5021, pages 167–175. SPIE, San Jose, California., 1 2003.

C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 31. IEEE, 4 1999.

G. Tzanetakis. Musescape: An interactive content-aware music browser. In *Proc. Int. Conference on Digital Audio Effects*. DAFx-03, London, UK, 9 2003.

G. Tzanetakis and P. R. Cook. Audio information retrieval tools. In *Proc. International Symposium on Music Information Retrieval*. ISMIR, 2000a.

G. Tzanetakis and P. R. Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 2000b.

G. Wood and S. E. O'Keefe. Quantitative comparisons into content-based music recognition with the self-organising map. In *Proc. Int. Symposium on Music Information Retrieval*. ISMIR 03, Baltimore, US, 10 2003.