Research White Paper

WHP 166

June 2008

Music Information Retrieval in Broadcasting: Some Visual Applications

Andrew Mason, Michael Evans, and Alia Sheikh

BRITISH BROADCASTING CORPORATION

BBC Research White Paper WHP 166

Music Information Retrieval in Broadcasting: Some Visual Applications

Andrew Mason, Michael Evans, and Alia Sheikh

Abstract

The academic research field of music information retrieval is expanding as rapidly as the MP3 collection of a stereotypical teenager. This could be no coincidence : the benefit of an automated genre classifier increases when the music collection contains several thousand tracks. Of course, there are other applications of music information retrieval. Here we highlight a few that make use of a simple, visual, representation of an audio signal, based on three easy-to-calculate audio features. The applications range from simple navigation around consumer recordings of broadcasts, to a music video production planning tool, to a short term "Listen Again" eye-catching display.

This document was originally published at the 123rd Audio Engineering Society Convention, New York, USA, October 2007

Additional key words:

White Papers are distributed freely on request.

Authorisation of the Head of Broadcast/FM Research is required for publication.

Music Information Retrieval in Broadcasting: Some Visual Applications

Andrew Mason[1], Michael Evans[2], and Alia Sheikh[3]

[1] BBC Research, Tadworth, Surrey, KT20 6NP, UK
andrew.mason@rd.bbc.co.uk

[2] BBC Research, Tadworth, Surrey, KT20 6NP, UK
michael.evans@rd.bbc.co.uk

[3] BBC Research, Tadworth, Surrey, KT20 6NP, UK
alia.sheikh@rd.bbc.co.uk

ABSTRACT

The academic research field of music information retrieval is expanding as rapidly as the MP3 collection of a stereotypical teenager. This could be no coincidence : the benefit of an automated genre classifier increases when the music collection contains several thousand tracks. Of course, there are other applications of music information retrieval. Here we highlight a few that make use of a simple, visual, representation of an audio signal, based on three easy-to-calculate audio features. The applications range from simple navigation around consumer recordings of broadcasts, to a music video production planning tool, to a short term "Listen Again" eye-catching display.

1. ADVANTAGES OF VISUALISATION

That navigating through an audio recording using only the sound can be time-consuming and inefficient is obvious. Listening using fast-forward or rewind soon becomes of limited benefit as the replay speed is increased much more than two times. The frequency shift makes it difficult to recognise the audio for which one is searching, and finite human reaction time makes it difficult to stop the search before it is too late.

Digital audio systems that deal with blocks of samples sometimes provide fast-forward replay by playing non-contiguous blocks: playing every third block provides replay at triple speed. The advantage of this is that it preserves the pitch (and some other properties) of the audio, but a lot is also lost.

Digital audio workstations (DAWs) have evolved over the years from early systems where one had to listen to the audio to edit[1][2][3], into ones where the image on the screen is as important. It is usual now for such a system to present the user with a picture of the waveform envelope, or the waveform itself, to assist in the location of edit points. This can work quite well, especially when points of interest are marked by significant changes in amplitude of the signal – gaps

between utterances, or the onset of isolated musical notes, for example.

However, the information conveyed by a waveform envelope is limited: there is nothing about pitch or timbre. To make the situation worse, audio is often subjected to aggressive dynamic range compression before being provided to the consumer, so the envelope is a rectangle.

Figure 1a shows the waveform envelope of an extract from a radio programme containing speech and music. One might guess that the middle section contains music, primarily because of its limited dynamic range.

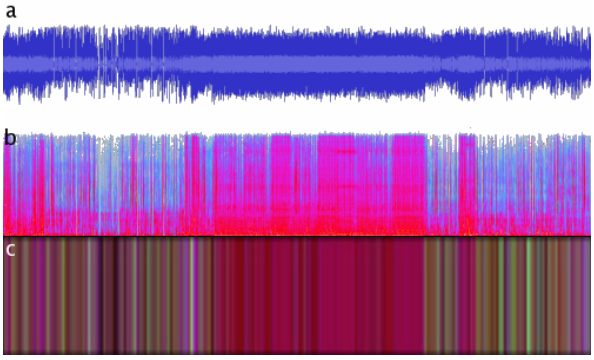

Figure 1a, b, and c From top: waveform envelope, spectrogram, and visualisation, of 8 minutes of BBC Radio 2 “Jeremy Vine” show

Numerous DAWs now offer a spectrogram visualisation as well as a waveform envelope. A spectrogram shows the temporal variation in energy in a multiplicity of frequency bands. This contains more information than the waveform envelope and can, to the trained eye, convey information about such things as pitch or timbre. The difficulty becomes one of interpretation because the spectrogram can contain too much information to absorb when looking looking at a long recording.

Figure 1b shows the full spectrogram of the same eight-minute long audio extract as Figure 1a. The spectrogram was calculated using the “Audacity” audio editing software, set to 4096 bands, showing the full bandwidth. We can see that the spectral character of the central section is distinct from the rest, with significantly more energy in the upper part of the spectrum. Some variation can be seen within the music, four short sections being followed by three longer sections. Figure 1c shows the same extract again, using our visualisation generated by mapping three audio feature vectors onto a red, green, and blue colour space. The music clearly stands out in the centre, as does the jingle that is played about 30s after the music finishes. A shorter jingle, a few seconds before the music starts, also shows up.

It is already clear that the visualisation presents a different kind of view, one that is more informative than the waveform envelope, and more readily comprehensible than the spectrogram.

Figure 2a, 2b, and 2c show the same kind of figures for an eight minute long extract of the "Today" news and current affairs programme from BBC Radio 4. Although the waveform envelope in Figure 2a shows much more variation than that in Figure 1a, it is not easy to navigate using it.

The spectrogram is good at showing up the low bandwidth telephone contribution near the centre. Our visualisation shows much more clearly that there are different speakers in the programme: the female speaker shows as pale blue, the male speakers as purple.

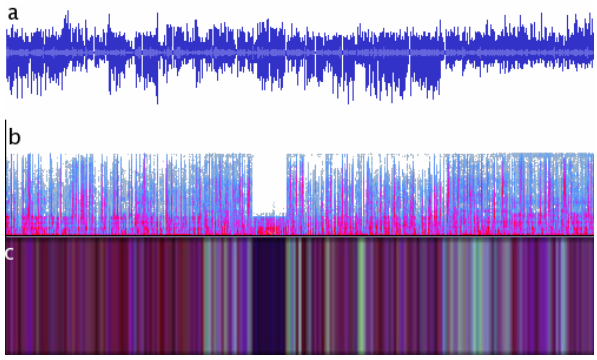

Figure 2a,b, and c From top: waveform envelope, spectrogram, and visualisation, of 8 minutes of BBC Radio 4 “Today” programme.

The third example, shown in Figure 3a, 3b, and 3c shows an edition of "From Our Own Correspondent" in which a number of foreign correspondents each contribute a section of the programme. There is a presenter, in this case Bridget Kendall, who introduces the programme and each section in turn.

The waveform envelope, figure 3a, shows almost nothing about the structure of the programme. The spectrogram, Figure 3b, reveals rather more, in that there appear to a number of sections with different uses

of bandwidth. Figure 3c, our visualisation, shows very clearly the overall structure of the programme: an overall introduction and 5 contributions each with an introduction.

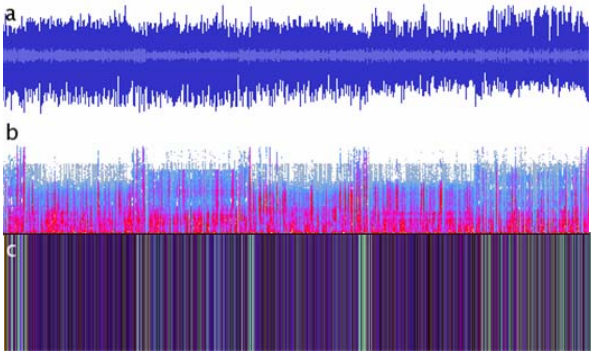

Figure 3a, b, and c From top: waveform envelope, spectrogram, and visualisation, of 8 minutes of BBC Radio 4 "From Our Own Correspondent" programme

It can be concluded from these examples that even a relatively simply generated visualisation can help greatly in the process of navigating around an audio recording, providing significant advantages over waveform envelopes and spectrograms.

2. VISUALISATION APPROACH

2.1. Background

The core of our approach is the direct visualisation of numerical features extracted from the audio signal. This contrasts with the alternative, two-stage approach of performing an audio classification - using an analysis of the audio feature to discriminate between, for example, speech and music, or male and female speech - and then mapping the detected classifications onto visual image features. Visualising feature-classified audio is a potentially useful technology in many applications but, for this research, we are primarily interested in harnessing users' abilities to perceive transitions, differences and similarities between visual features.

Our visualisations do not include a key or legend identifying the meaning of each colour or other visual feature. We process the audio to compute a number of low-level audio features, and then map variation in these features onto variation in fundamental visual parameters. If this mapping is achieved effectively, then viewers of the visualisation should be able to detect contrasts and transitions in the image which are analogous to those that would be perceived by listening to the audio.

2.2. Colour Mapping

2.2.1. Colours and Audio

Colours are the fundamental features of our visualisations. Three audio features are chosen and mapped on to red, green and blue in the image. In our experimental system we have chosen three audio features which, whilst not fully orthogonal to each other, are largely independent and exhibit contrasting relationships across different types of audio signal. These audio features are:

- Zero Crossing Rate (*ZCR*)
- The Third Central Moment of the Zero Crossing Rate ($\mu 3Z$)
- The 95th Percentile of the Amplitude Spectrum ($\omega 95$)

Our algorithm segments the audio signal into a sequence of equal duration *blocks*. For each of these, a numerical value derived from each of the three audio features is computed. The analysis also uses a subdivision of each block into a sequence of equal duration *frames*. We typically use a block duration of 1s and a frame duration of 8⅓ms (48000 and 400 samples respectively, for audio sampled at 48kHz).

2.2.2. Zero Crossing Rate (ZCR)

ZCR is computed by counting the number of times the amplitude of the audio signal changes sign over a block, and then dividing by the block length. It can be used as an indication of the dominant frequency in the signal during this block and, therefore, as a strong indicator of perceived pitch[4].

In our algorithm, the ZCR of the *K*th block is given by

$$Z_K = \frac{1}{N} \sum_{n=NK}^{N(K+1)-1} [a_n a_{n-1} < 0] \qquad (1)$$

where Z_K is the ZCR of the *K*th block, *N* is the number of samples per block and *a* is the sampled audio signal. (Note that an Iverson bracketed expression of the form

[*P*] is substituted with the value 1 when the proposition *P* is true and the value 0 otherwise.)

At frame level $z_{k,K}$, the ZCR of the *k*th frame of the *K*th *block*, is

$$z_{k,K} = \frac{1}{M} \sum_{m=NK+Mk}^{NK+M(k+1)-1} [a_m a_{m-1} < 0] \quad (2)$$

where *M* is the number of samples per frame. Note that *N* must be an integer multiple of *M* and, therefore, *N/M* is the number of frames per *block*.

2.2.3. Third Central Moment of the Zero Crossing Rate (μ3Z)

The Third Central Moment ($\mu 3$) of a variable's distribution indicates how skewed (i.e. non-symmetric) its distribution is around the mean. Calculating this variable for ZCR data is a measure of asymmetry in the audio spectrum and, as such, has been used for speech-music discrimination[5]. $\mu 3Z_K$, the third central moment of framewise ZCR data for the *K*th block, is given by

$$\mu 3Z_K = \frac{M}{N} \sum_{k=0}^{\frac{N}{M}-1} \left(z_{k,K} - Z_K \right)^3 \quad (3)$$

2.2.4. 95th Percentile of the Amplitude Spectrum (ω95)

Computed in the frequency domain, this feature is used to indicate the spectral roll-off point of the audio in each block. It is a strong indicator of the relative bandwidth of audio signals and is used in audio classification tasks including the discrimination between voiced and unvoiced speech and between speech and music[6]. We determine $\omega 95_K$, the 95th percentile of the amplitude spectrum of the *K*th block, by computing the lowest frequency, ω_r, which satisfies the relation

$$\sum_{\omega=0}^{\omega_r} \left| A_K[\omega] \right| \geq 0.95 \left(\sum_{\omega=0}^{\omega_{MAX}} \left| A_K[\omega] \right| \right) \quad (4)$$

where $A_K[\omega]$ is the Discrete Fourier Transform (DFT) of the *K*th block of the audio signal, and ω_{MAX} is the highest frequency bin in the DFT.

2.2.5. Feature Ranges and Normalisation

The numerical values of the three audio features for each block combine to create a single colour. The value of *ZCR* controls the amount of blue, $\mu 3Z$ the green, and $\omega 95$ the red. To make best use of the available colour space, the ranges of each audio feature should be normalised to encompass the full range of level of each of the three colour components.

The simplest normalisation algorithm determines the lowest value of one of the audio features across the audio as a whole and maps that to a zero value of the related colour component. Similarly, the maxium value observed in the audio feature is mapped to the highest level of the colour component. Between this minimum and maximum value, a linear increase in colour component intensity can be used, but, whilst this produces useful results, the use of the colour gamut is not optimal. It is also difficult to do this if the visualisation is of a continuing stream of audio: one can never know the maximum or minimum that will be reached (unless one has already encountered the theoretical limit of a feature).

In order to develop better scaling, a study has been started into the distribution of the three features in normal broadcast output. It is hoped that this will lead to a better mapping of features into colour space. Figure 4 shows the distributions of the three features over a period of 48 hours for four of the BBC's national radio networks, as received from digital satellite transmissions. The three features are, from left to right, *ZCR*, $\mu 3Z$, and $\omega 95$. The four radio networks are, from top to bottom, BBC Radio 3 (classical music), BBC Radio 4 (speech, news, drama), BBC Radio 5 Live (sports), and BBC 6 Music (pop music).

Clearly, the distributions are not uniform, but it is not clear that distorting them to make them more uniform would be optimal. It seems probable that better discrimination between the more likely values would be beneficial. On the other hand, there might be unlikely feature values that should remain as outliers, and not be brought closer to the rest of the distribution.

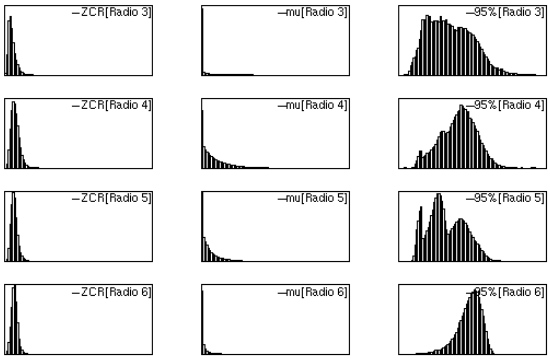

Figure 4 Distributions of feature values over 48 hours for four BBC national radio networks.

The optimal mapping of features to colour space remains to be explored more fully.

2.3. Beyond the Triple Feature Vector

Visualisation using colour component mapping is limited to three audio features at a time. A much wider range of audio features could be extracted from the audio signal and then the three that give the best visualisation chosen automatically or interactively.

Further, we could explore extra visual features in addition to colour and visualise a vector of four or more audio features simultaneously. Additional audio features could be mapped onto visual features such as image texture, luminosity, 3-D depth, or transparency.

3. CONSUMER AUDIO PLAYER

The first application to be built using our visualisation was a simple audio player. This was constructed as a web service: the user uses a web browser to upload an audio file and the web server returns a page containing a player (a Flash[1] movie) to play the file, including the visualisation. A simple HTML form, generated by a perl script[7], is used to submit the audio file. The uploaded file is handled by the same script, a visualisation generated, and an HTML page sent back to the browser. An example of a page returned by the server is shown in Figure 5.

[1] "Flash" and "ActionScript" are trademarks of Adobe Systems Incorporated.

Figure 5 On-line "Radio" player with visualisation

The web page actually contains very little, just the HTML to load the Flash movie, together with the parameters to get audio and visualisation files. The Flash movie contains an image of a radio[2], and some ActionScript to load an audio file, a visualisation file, and provide some audio replay control. The coloured bar near the top contains the visualisation, together with a cursor indicating the current playing position. Clicking the left-hand mouse button in the visualisation moves the cursor to the position of the mouse, allowing random access to the audio.

ID3 tags from the MP3 file are displayed in a text box, in a way reminiscent of dynamic labels on DAB. Simple "skip forward", "skip back", and "play/pause" buttons are provided on the radio. Volume control is effected by clicking on the loudspeakers.

The example shown in Figure 5 used a podcast of "From Our Own Correspondent" from BBC Radio 4 (as was shown in Figure 3). The ease with which the user can now see the structure of the programme, allowing very quick navigation to an item of interest, is apparent. The cursor and time indicator also provide a sense of place within the recording that can be useful in some circumstances.

4. RADIO ON TELEVISION

A further application of the technique is for use with broadcasts over digital terrestrial (DTT) and digital

[2] Drawn by Kevin Claydon of BBC Research

satellite (DSat) television. A PVR (personal video recorder) can record audio programmes broadcast using these systems, and is very simple to operate if an appropriate electronic programme guide (EPG) is incorporated. On DTT, for example, the MHEG engine[8] can be used to display images and text transmitted on an ancilliary data service. If the programme being broadcast is pre-recorded, then a visualisation of the whole programme can be shown, together with a marker that is periodically updated to show progress through the programme. For live programmes this is, of course, not possible, but the visualisation can be updated to reflect what has just happened.

Figure 6 shows a screenshot with two visulisations. The upper visualisation shows the entire 30 minutes of the current programme and was generated in advance of the broadcast, from the file on the server. This visualisation is present for the duration of the pre-recorded programme: the cursor that is superimposed moves from left to right to indicate progress. The lower visualisation is generated live and scrolls from right to left. The most recently generated part of the visualisation is at the right of the figure. Because the screenshot depicts a time just over half-way through the current programme, the lower visualisation still shows a substantial amount of the preceeding programme.

The visualisations are shown superimposed on the normal MHEG generated display for BBC 7, one of the BBC's digital radio services.

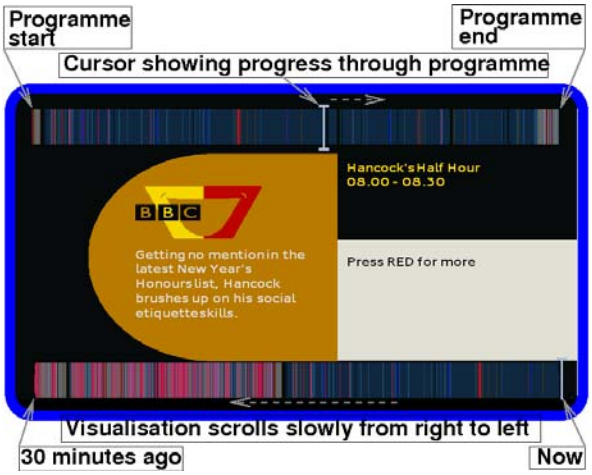

Figure 6 Digital terrestrial TV audio service with visualisations showing progress through current programme (top), and live history (bottom).

At the time of writing, not all PVRs record the data stream necessary to re-run the MHEG applications when replaying recordings, and there are some technical difficulties to be resolved with the generation and timely delivery of the data stream.

For on-line delivery of programmes, such as has been done in the past with "Listen Again"[1] and with BBC iPlayer, the incorporation of facilities using visualisations would be, in some ways, rather simpler.

5. RADIO NETWORK SHOWCASE

In public areas of BBC buildings it is conventional to have publicity material displayed. This takes several forms but still images and moving images are by far the most prevalent. Moving images take the form of live output from the various television networks, and it is easy, by having several monitors, to show all the television networks' pictures simultaneously. Radio is at a disadvantage here: to have all the radio networks presented simultaneously over loudspeakers would be unbearable.

Before the refurbishment of the BBC's Broadcasting House in London there was an area there where several radio networks were displayed on a bank of Peak Programme Meters (PPM). This went some way to demonstrating the continuity and multiplicity of the BBC's output. As a replacement for this, the idea of using live visualisations is being considered. A variant of the visualisation generation sofware has been created that analyses a live stream and continuously updates a visualisation representing the last two hours of the stream.

The visualisation is presented as a spiral that gradually winds its way into the centre as it gets older. The spiral is continuously updated at the outside with the latest value, causing the whole to rotate slowly.

This could be used in the new Broadcasting House reception area, as shown in an artist's impression in Figure 7.

[1] "Listen Again" is the name used for the BBC's service whereby listeners can listen on-line to individual radio programmes on demand for up to seven days after broadcast.

Figure 7 Artist's impression of BBC Broadcasting House reception with live radio network visualisations.

The five different radio network visualisations, shown projected on the front of the reception desk in the figure, exhibit different visual characteristics: From left to right, Radio 1 is pop music, Radio 2 is mixed speech and pop music, Radio 3 is classical, Radio 4 is news, current affairs, and drama, Radio 5 is sports.

The application that was developed to do this buffers the audio from the last two hours so that, when run on their own computer, the user is able to listen to the audio from a given point in time by clicking a mouse at the desired point on the visualisation. This provides a very simple, short term, "Listen Again" facility, that can be useful if one misses the news or a traffic report.

6. THE USE OF MUSIC BREAKDOWNS IN BROADCAST PRODUCTION

A music breakdown is a textual description of an audio track, which lists the timing of events in the music. Breakdowns are used as part of the broadcast production workflow for both live and scripted music productions. The structure and complexity of a breakdown is a function of the use to which it will be put and the person who creates it. A music breakdown can be incredibly detailed, showing what occurs at each beat of the music: for example beat and bar number, lyrics, and which instruments are being played. An example is shown in Figure 8.

Ray & Camilla
"It's My Life"
Dur: 1'29"
Paso Doble

1 2 3 4

2 2 3 4

3 2 3 4

4 2 3 4

This ain't a song for the
1 2 3 4

broken-hearted
1 2 3 4

1 2 3 4

2 2 3 4

A silent prayer for th'
1 2 3 4

Figure 8 An extract from a music breakdown for "Strictly Come Dancing": a live studio dance show.

At its most broad, a music breakdown can be an overview, listing events such as verse/chorus/drum fill and usually assigning them a length in bars and beats. An example of this is shown in Figure 9.

Orson				
"Happiness"				
Dur: 3'26"				
4	Bars		Keys & Gtr	Drum Fill on 4,3
3	Bars	+1,2	Full Instrumental	
3,4 +7	Bars	+1,2	Verse (with Gtr)	
3,4 +7	Bars	+1,2	Verse (with 2nd Gtr)	
3,4 +7	Bars		Bridge	
8	Bars		Chorus (with BV "Ah" on 4)	
3	Bars	+1,2	Inst (Gtr) - builds on 3	
3,4 +8	Bars		Verse (2 Drum hits at very top)	Drum Fill on 8,4
4	Bars		Bridge (with BVs)	
8	Bars		Chorus (with BV "Ah" on 4)	
5	Bars		Inst (Gtrs)	Gtr & DF on 5
8	Bars		Chorus (with BV "Ah" on 4)	Drum Fill on 8,3
4	Bars	+1	Inst (Gtr) & Vox Top Lines (to the end)	

Figure 9 The music breakdown for "Happiness" by Orson, for the Leeds and Reading Festival, 2007, created by Tony Grech, the PA for the production

It is usually the role of the production assistant (PA) otherwise known as the script supervisor, to create and use a music breakdown. Studio-based productions can have either live or pre-recorded musical segments for both of which the camera direction is usually pre-rehearsed, and then filmed 'for real'. Alternatively a

production might be on location - for example a music festival for which there would be only one opportunity to film the performance and no rehearsal.

For studio performances, the camera direction is usually pre-scripted by the director and rehearsed by the crew prior to the final performance. Not only is the breakdown used by the director to plan the script, but when it comes to rehearsing and filming the performance, the breakdown is used by the PA to 'call out' the point that the artist(s) have reached in the music, the current shot number, and which camera should prepare for the next shot: essentially providing a point of reference that is used by the entire crew to carry out their roles successfully.

In some cases a director may decide to 'busk' a studio performance. A 'busk' is the name given to a performance that does not have pre-scripted camera shots. In these situations, having a PA call out the music breakdown (the bars, beats and structure of the song) is invaluable, because it is the only information the director receives which indicates what is coming next in the performance.

A detailed music breakdown is only useful when it is certain that an artist or group will play a musical number sufficiently similarly to the available recorded performance on which the breakdown is based. In some cases there is a strong chance that a performance will be very different from a pre-recorded or 'album' version, or that the director will want to film spontaneous events that cannot be planned for. This is usually the case when filming music festivals, as the artists will have a strong tendency to freestyle and a significant part of the compelling footage will be comprised of the interaction of the performers with the crowd. In these situations the rules are similar to those for a 'busk' with the director calling shots on the fly. Again, some sort of music breakdown is invaluable. To balance providing helpful cues and risking information overload, a 'race read' or minimal breakdown is used, in which the PA calls only major events in the music and their durations, such as the one shown in Figure 9. These breakdowns are more robust in terms of remaining helpful if the band decides to change something in the live performance.

It should be noted however that whilst the presence of a PA calling out a race-read breakdown makes a very real difference to the crew of a live broadcast, it is not guaranteed that it will be available. In order for race-read breakdowns to be produced, there needs to be a sufficient availability of music PAs who receive the music in enough time to produce the breakdown. Dedicated music PAs are both highly skilled and rare, so there is always more demand for them than can be met. Budgetary or timetabling constraints can also can lead to situations where a PA does not have enough time to create breakdowns, with PAs sometimes breaking down music minutes before a performance begins.

It is obvious therefore that anything which improves the process of breaking down music for a PA, will have a positive impact across the rest of the production team. The same can be said for any tool that can be used by a PA inexperienced in working in music productions, for both training purposes and in order to more effectively create breakdowns.

6.1. Music Marker

When producing a breakdown the PA will listen through the entire track, identifying the beat patterns and noting down the lengths of logical musical segments (e.g. intro/verse/chorus/instrumental/middle eight) as well as the timing of significant events (e.g. drum fill/cymbal crash/pause in the music/backing vocals). The disadvantage of this approach is that the PA must listen to the music in a linear fashion and has no visibility of when a segment will start or end or when a significant event is about to occur. It therefore takes a skilled PA more than one pass to create even a simple race-read breakdown, in about 10-15 minutes. A novice PA may take twice as long, using three or four passes. A non-music PA would find the process unintuitive and, unless they were already fairly 'musical', difficult.

The current method for creating a music breakdown is for the PA to play the music on either a portable CD player, or as an MP3 on a software player, making notes using pen and paper which are later typed up. A software media player offers advantages over a compact disc player as these often provide a more intuitive interface for skipping back and forth in the track, however neither can be said to be ideal.

Our goal was to create a web based software tool to assist in the creation of race-read breakdowns, in such a way as to improve the experience for a skilled PA and to make the process more intuitive and accessible for a novice or non-music PA. It would be counterproductive to create the breakdown entirely automatically with no input from the PA as PAs rely on this process as a way of familiarising themselves with the music. This in

itself is an essential prerequisite for the PA's role in the production gallery.

The software we have created is accessed via a webpage and takes as its input an MP3 of the music to be broken down. Once the MP3 has been uploaded, the user is presented with a visualisation of the entire length of the music. Buttons and keyboard shortcuts for adding mark-up and an option to print out or save a correctly formatted music breakdown are also provided. The upload and visualisation mechanism is the same as for the player in Section 3. Further scripts handle the storage of markers and the generation of formatted output.

It is immediately apparent that the Music Marker tool presents a far richer view of the audio than a standard software player. As well as the standard play controls, the music itself is presented, not as a featureless timeline but as two vertical stacks of horizontal coloured stripes which represents the music. The first visualisation always shows the entire length of the audio and also indicates the section that is denoted by the second visualisation which can be zoomed in to and out of. A change in the music shows up as a colour change in the visualisation and a short event such as a cymbal crash will show up as a coloured line, whilst a recurring segment (such as a verse or chorus) shows up as a recurring segment of colour in the visualisation.

Figure 10 shows the view presented by the tool for the song 'Steady As She Goes' by the band The Raconteurs. The verses and the chorus show up as different colours in the visualisation, deep pink and deep purple respectively, and the opening instrumental segment shows clear distinction between the different instruments present.

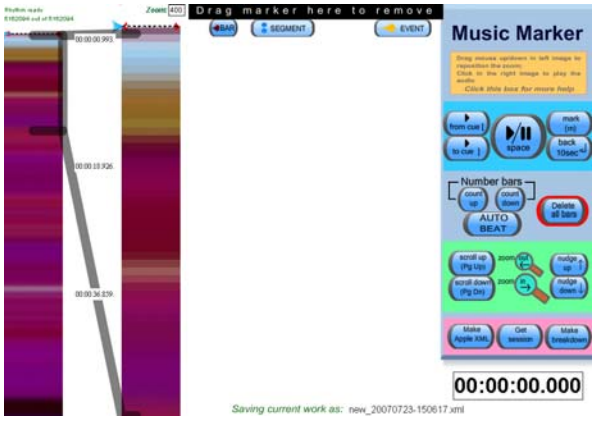

Figure 10 The initial view presented by the Music Marker tool. The visualisation has been generated from the song "Steady As She Goes" by The Raconteurs.

During the upload process the beat structure of the audio is analysed. If the user inputs the beat pattern for a single bar (tapped out using keyboard shortcuts) in a portion of the music that has a strong beat, our tool can extrapolate the positions of the remaining beats and bars. Variations in tempo are tracked in this process using a beat tracking algorithm[9]. At the time of writing, this process does not cope with changes in time signature, however, this facility can be added, requiring the user only to mark a second bar when the time signature change occurs. Alternatively, the user has the option to mark the beats and bars manually, in which case any change in time signature can be represented.

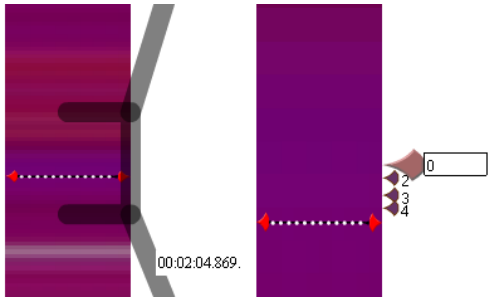

Figure 11 The beats of a single bar marked up alongside the right-hand visualisation. The remaining bars and beats are extrapolated from this when the "auto beat" button is pressed.

Once the beat pattern has been extrapolated, the user plays the music track from start to finish, using keyboard shortcuts and onscreen buttons to mark up the music in real-time. The music can be paused at any

time to allow a more detailed textual description to be added and markers can be resized and moved in an intuitive fashion. Bar and beat durations are calculated for and added to each segment that is marked up. These are automatically recalculated if the segment is resized by dragging the upper or lower handle. As can be seen in Figure 12, distinct events, in this case when certain instruments come in, are clearly apparent. The snare drum shows up as a pale blue, the bass drum as brown and a high pitched note on the electric guitar as a deep purple.

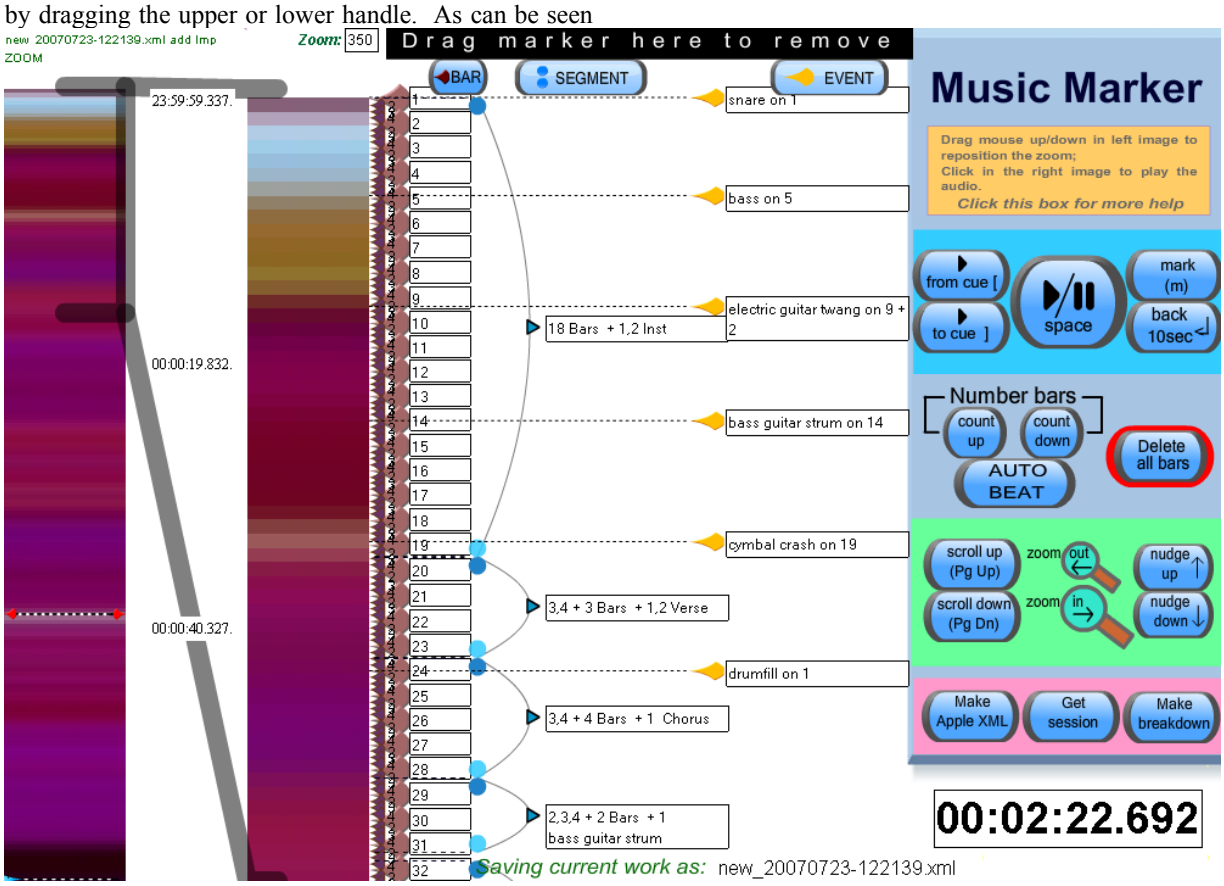

Figure 12 Events and segments in the music marked up against automatically generated beats

Once familiar with the system, a user is able to complete a breakdown in two complete passes, one pass to mark up the song and a second pass to check it through. The final breakdown is generated as a simple text file, an example of which is shown in Figure 13.

```
"STEADY AS SHE GOES" (THE RACONTEURS) (00:03:35)
                                                    snare on 1
18 Bars + 1,2 Inst
                                                    bass on 5
                                                    electric guitar twang on 9 + 2
                                                    bass guitar strum on 14
                                                    cymbal crash on 19
3,4 + 3 Bars + 1,2 Verse
3,4 + 4 Bars + 1 Chorus
                                                    drumfill on 1
2,3,4 + 2 Bars + 1 bass guitar strum
4 + 3 Bars + 1,2 Verse
3,4 + 12 Bars + 1 Chorus
                                                    drum fill on 1
                                                    cymbals on 5
                                                    cymbals on 11
2,3,4 + 2 Bars + 1,2,3 bass guitar strum
4 + 8 Bars + 1 Verse
2,3,4 + 0 Bar + 1,2,3 Electric twang
4 + 15 Bars Chorus
                                                    cymbals on 5
                                                    drumfill on 9
                                                    cymbals on 11,4
                                                    bass guitar strum on 14,2
8 Bars Middle 8 (Backing Vocals & bass guitar)
2,3,4 + 3 Bars + 1,2,3 Electric twang
                                                    drumfill on 5, 4
4 + 8 Bars + 1 Chorus
2,3,4 + 7 Bars + 1,2 Chorus/Outtro
3,4 + 5 Bars + 1,2 (Quiet) Bass Guitar
```

Figure 13 Breakdown created using Music Marker. The format follows that used in Figure 9 that was generated by hand (and ear).

6.2. Benefits of Music Marker

Music Marker offers several advantages over the traditional approach to breaking down music. Once a user has become familiar with the system, creating a breakdown is extremely efficient. For the inexperienced PA it offers a very definite speed advantage. For an experienced PA, (once upload and analysis times are taken into account) using Music Marker may not be much faster than using pen and paper. However, the feedback we have had from PAs indicates that it is a more pleasant experience to use and it also eliminates the final typing up stage.

Music Marker sessions can be stored and accessed anywhere there is a computer connected to the internet. This means that a breakdown created in one location can be completed in another, by a completely difference person if necessary. It is easier for breakdowns to be shared and reused as changes and additions can easily be made to the original breakdown. Since outside broadcast events such as music festivals now usually have a good internet connection, the tool is available on location, offering a way of making last minute breakdowns in the gallery (or outside broadcast truck) itself. Music Marker also offers a solution to the problem of marking up music with no strong beat pattern. Completely un-percussive music is rare in pop but breaking down such music is extremely difficult using traditional methods as segments in the music have to be identified and then listed against time code. Since Music Marker by default visualises and displays segments in the music against time code it makes it very easy to create a music breakdown for this type of music.

7. ASSISTED SEGMENTATION FOR PODCAST PRODUCTION

As rights negotiation continues, more and more radio programmes are being made available as podcasts, or for conventional download. An example of a programme where part of it is made available is the "Today" news progamme on BBC Radio 4. There is a high-profile, live, interview scheduled for 8.10 each day, that is later made available as a podcast.

> 0710: The schools inspector OFSTED has criticised the way history is being taught.
>
> 0730: The results of the by-election are out and Labour have held onto both their Sedgefield and Ealing South seats. Meanwhile the Conservatives came third in both. Caroline Spellman
>
> 0750: The modern assumption that repeat cot deaths are natural is being challenged by a new report.
>
> 8.10 Interview. After 16 months and 136 interviews no-one is to face charges in the cash for honours investigation. John McTernan and Sir Menzies Campbell

Figure 14 Example of partial running order from an edition of BBC Radio 4 "Today" programme.

Of course, there are several ways in which a segmented version of the otherwise continuous programme can be made, but the use of a visualisation could speed up one where a complete recording and an intended running order are used.

In this case, the recording cannot be segmented automatically based on the intended running order, since there are certain to be variations in actual timing of the different segments: human intervention will be required. A variant of the Music Marker application, such that the intended running order is presented aligned to the audio visualisation, can aid the segmentation. The visualisation can provide clues as to where the actual segment junctions are in relation to the intended junctions. Once the operator has marked up the visualisation, the process of segmentation, including generation of appropriate meta-data annotations can take place at the click of a button.

8. FURTHER WORK

For all the applications described here, the simple player, the MHEG display, the spiral, segmentation, and annotation there is clearly much that can be done to improve the effectiveness of the visualisation. Discrimination between different audio segments is the key. What constitutes "different" depends on context.

For some applications it might be of great help to discriminate between different people speaking; for others, the difference between instrumental and vocal sections of music might be more important; for others, the difference between symphony and chamber orchestra; yet other may wish to discriminate between country and western music.

The three features that we have used were chosen because they were simple to implement, yet demonstrably effective – but many other features are known. One strategy for improvement would be to calculate more features and map a subset to a visualisation. The subset could be chosen automatically for best contrast, or with assistance from the user, by, for example, context-based indication : "speech mode", "vocal/instrumental mode", "classical music mode".

Used in Music Marker, we have found that useful discrimination is obtained in about 75% of pop music tracks. Where there is no clear percussive or instrumental difference between verse and chorus, the discrimination is lessened. We can foresee a situation where the music can be segmentated automatically sufficiently well for an initial breakdown to be presented for the PA to refine. This might actually be disadvantageous if it means that the PA does not get the familiarisation with the music that will be required.

One director has already proposed using the music visualisation in the gallery itself, with the display automatically scrolling to keep in time with the live performance. This would help in situations where a PA was able to create breakdowns, but was not available on the day.

9. SYNAESTHETIC CONTENT NAVIGATION

In this work we have investigated using a visual representation of audio to assist in navigation. It is easy to process a visual representation of something changing over time in a shorter time frame than the event itself took to occur (for example using graphs to analyse stock market trends): a fact that our navigational tools exploit. The same cannot be said for processing information in the audio domain, as the temporal nature of audio defines the way in which we experience it. In order to hear the change over time of a piece of audio, it must be experienced in a linear fashion.

Research into synaesthetic techniques offers us an exciting approach to the problem of improving data analysis, by mapping information which is normally processed by one sense, into information that can be processed by another. There are obvious accessibility advantages to be had, with the prospect of making information accessible to those who have lost, or have reduced functionality in, the sense that usually handles a given type of data. There may also be scope for revealing 'hidden' properties of the information we are trying to analyse by changing the way in which we perceive it.

The most important point to remember when creating any sort of representation of audio data is the nature of the information the end user requires. A music PA will be more interested in the structure of a piece of audio (and will therefore be interested in being able to see when a 'change' happens) whereas a sound engineer would be more concerned about the quality of a piece of audio at different points along its length. Someone wishing to play a piece of music recreationally would be interested in representations of the music that simply enhance the overall experience for them - see as an example the rise of VJs ("video jockeys") in nightclubs. A person with impaired hearing may be more interested in representations of audio that go some way towards making accessible the information that they have lost.

There is scope for investigating whether mapping the audio features we analyse, not to a visualisation, but once again to audio, might provide another useful representation of the original source.

It was mentioned earlier that other audio features could be mapped to image properties such as luminosity, texture, depth, and so on. The next obvious step from mapping audio features into a virtual 3D space, is to map them into actual 3D space. The concept of using haptic feedback as a navigational aid is not a new one but the technology is an obvious candidate for enhancing the visualisation methods we have presented here. Our coloured visualisations could be rendered on a flexible touch-sensitive display which uses either vibration or mechanical deformation (for example, a backing of a pin-array which can be raised or withdrawn) to convey a texture. One could imagine in the sci-fi world of tomorrow, the scroll wheel on a portable music player applying a texture across its surface as the user scrolls through a music track. Variations in the temperature of a surface could also be used to indicate some parameter of the music. Cool.

Music already represents itself in a haptic sense, as the vibrational energy of the music is felt not only by our eardrums but also through our skin and the soles of our feet. There is potential for mapping audio features to a vibration that is applied across a surface - whether this could be used as a navigational aid, especially when fast forwarding through a track, remains to be seen. Refining this further, haptic feedback could be used to present lyrics represented in Morse or Braille in time to the music. It is not clear though that this information could be processed quickly enough by the user.

Ultimately it can be seen that there are advantages in using the kinds of multisensory signal processing that we have described here. Representations of data normally processed by one sense (audio in this case) mapped to the other senses of the body offer us a way of distinguishing information and trends that may be less apparent in the original source. We have used this approach to streamline the job of marking up music, segmenting audio files for playback, and as a navigational aid. There is however, much that can still be explored.

10. ACKNOWLEDGEMENTS

The authors would like to thank the BBC's Head of Research for permission to publish this paper. The authors would also like to thank Tony Grech for his willlingness to try, and constructive criticism of early versions of the Music Marker.

11. REFERENCES

[1] McNally, G. W.; Gaskell, P. S.; Stirling, A. J. "Digital Audio Editing" presented at the 77th AES

Convention (February 1985), Preprint Number 2214

[2] McNally, G. W. "Fast Edit Point Location and Cueing in Disc-Based Digital Audio System" presented at the 78th AES Convention (April 1985), Preprint Number 2232

[3] McNally, G. W. "Variable Speed Replay of Digital Audio with Constant Output Sampling Rate" presented at the 76th AES Convention (September 1984), Preprint Number 2137

[4] Kedem B., "Spectral analysis and discrimination by zero-crossings", Proc. IEEE, vol.74, pp.1477-1493, Nov. 1986

[5] Saunders J., "Real-time Discrimination of Broadcast Speech/ Music" Proc. ICASSP96, vol.11, pp.993-996, Atlanta, May, 1996

[6] Scheirer E., Slaney M., "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97), vol.2, pp.1331-1334, April 21-24, 1997

[7] L. Stein, "Official Guide to Programming with CGI.pm", Wiley Computer Publishing, 1998

[8] International Standard ISO/IEC 13522-5:1997 "Information Technology – Coding of Multimedia and Hypermedia Information – Part 5: Support for base level interactive applications", International Organisation for Standardisation.

[9] Evans, M. J., "Interactive Beat Tracking for Assisted Annotation of Percussive Music", paper to be presented at 123rd AES Convention (October 2007)