# Song Form Intelligence for Streaming Music across Wireless Bursty Networks

Jonathan Doherty, Kevin Curran, Paul Mc Kevitt

School of Computing and Intelligent Systems
Faculty of Engineering
University of Ulster, Magee Campus,
Derry/Londonderry, BT48 7JL, N. Ireland
{Doherty-J22, KJ.Curran, P.McKevitt}@ulster.ac.uk

**Abstract.** Preliminary research on the development of a system for streaming audio across a wireless network, whilst using Song Form Intelligence (SoFI) to correct bursty errors, is presented. Current problems identified with streaming audio across wireless networks are reviewed. Recent approaches on error concealment when bursty errors occur, and Music Information Retrieval are discussed. We propose an approach that uses an amalgamation of network approaches and music information retrieval techniques to solve streaming music problems. Initial findings indicate that this approach will benefit such problems.

## 1 Introduction

Streaming media across networks has been a focus for much research in the area of lossy/lossless file compression and network communication techniques. However, the rapid uptake of wireless communication has led to more recent problems being identified. Traffic on a wireless network can be categorised in the same way as cabled networks. File transfers cannot tolerate packet loss but can take an undefined length of time. 'Real-time' traffic can accept packet loss (within limitations) but must arrive at its destination within a given time frame.

Forward error correction (FEC) [1] which usually involves redundancy built into the packets, and automatic repeat request (ARQ) [1] are two main techniques currently implemented to overcome the problems encountered. However bandwidth restrictions limit FEC solutions and the 'real-time' constraints limit the effectiveness of ARQ. The increase in bandwidths across networks should help to alleviate the congestion problem. However, the development of audio compression including the more popular formats such as Microsoft's Windows Media Audio WMA and the MPEG group's mp3 compression schemes have peaked and yet end users want higher quality through the use of lossless compression formats on more unstable network topologies.

When receiving streaming media over a low bandwidth wireless connection, users can experience not only packet losses but also extended service interruptions. These dropouts can last for as long as 15 seconds. During this time no packets are received and, if not addressed, these dropped packets cause unacceptable interruptions in the audio stream. A long dropout of this kind may be overcome by ensuring that the buffer at the client is large enough. However, when using fixed bit rate technologies such as Windows Media Player or Real Audio a simple packet resend request is the only method of audio stream repair implemented.

### 1.1 Objectives of Song Form Intelligence (SoFI)

The principle behind the research presented here is to develop a streaming audio system called *SoFI* that uses pattern matching techniques. The core objectives of SoFI are:

- To match the current section of a song being received with previous sections.
- To identify incomplete sections and determine replacements based on previously received portions of the song.
- To use cognitive techniques to perform error concealment of the packet loss based on similarity analysis.

Satisfying these objectives requires investigation into areas including current approaches to packet loss, audio similarity analysis to satisfy the pattern matching constraint. Next, in section 2 we look at research related to network approaches to error concealment and research in the field of Music Information Retrieval that uses pattern matching techniques. In section 3 we present an overview of the architecture of SoFI. Finally, in section 4 our conclusions and future work are explained.

## 2 Related Work

Packet delay from network congestion has been partially alleviated using routing protocols and application protocols such as real-time transport protocol (RTP) that have been developed to assign a higher priority to time dependant data. However, it is also the case that some servers *automatically dump* packets that are time sensitive, so streaming applications have had to resort to 'masking' the packets by using HTTP port 80 so packets appear as normal web traffic.

The latest addition to network protocols specifically addressing 'real-time' communication include Voice over Internet Protocol (VoIP), a technology that allows telephone calls using a broadband Internet connection across a packet switched network instead of a regular (or analog) phone line.

### 2.1 Network Approaches to Error Concealment

Solutions to the inherited problems within streaming audio have included research into a number of varying techniques. The probability of packet loss across bursty

networks has been modelled where time delay is used to control the flow of packets and measure the difference between the current time and the time the packet arrives [2]. This technique can be used to predict network behaviour and adjust audio compression based on current network behaviour. Higher compression results in poorer quality audio but reduces network congestion through smaller packets. A variation of this theme has been used to create new protocols that allow scalable media streaming [3].

Randomising packet order to alleviate the large gaps associated with bursty losses was implemented, where the problem was reduced by re-ordering the packets before they are sent and reassembling the correct order at the receiver [4]. This reduced the bursty loss effect since packets lost were from different time segments. Although nothing is done to replace the missing packets, overall audio quality had improved through smaller gaps in the audio – albeit more frequent.

A number of techniques that use some form of redundancy where repetition is used to replace lost audio segments have been developed. Sending packets containing the same audio segments but with a lower bit-rate alongside the high bit-rate encoding increases the likelihood of packet arrival but at the loss of audio quality, as well as increasing the overall network bandwidth usage [1]. Another approach to using redundancy in the form of unequal error protection (UEP) was developed, where improvement is achieved with an acceptable amount of redundancy using advanced audio encoding (AAC) [5]. Segmentation of the audio into different classes such as drumbeats and onset segments allows priority to be applied to more important audio segments with ARQ applied to high priority segments and reconstruction techniques for the replacement of low priority segments based on the AAC received in previous segments.

One of the most recent methods of interpolation of low bit-rate coded voice is used where observation of high correlation of linear predictors within adjacent frames allowed descriptions to be inserted using linear spectral pairs (LSP), and then reconstruct lost LSPs using linear interpolation [6].

### 2.2 Music Information Retrieval

One of the core aspects of this research is to use pattern matching to identify similar segments within an audio file. Research into the field of music information retrieval (MIR) has gathered momentum over the past decade. With the increase of audio file sharing across heterogeneous networks, a need has arisen for more accurate search/retrieval of files. Research into the analysis of audio has led to the development of systems that can browse audio files in much the same way as search engines can browse web pages retrieving relevant data based on specific qualities [7], [8], [9].

Recent work in pattern matching within polyphonic music has shown that similarity within different sections of a piece of music can aid in both pattern matching for searching large datasets and pattern matching within a single audio file [14], [15], [17]. Results from analysis of an audio stream are stored in a similarity matrix used in [14] which can be seen in Fig 1. Using short time Fourier transform (a variation of discrete Fourier transform which allows for the influence of time as a

factor) to determine the spectral properties of the segmented audio. A chroma based spectrum analysis technique was used to identify the chorus or refrain of a song by identifying repeated sections of the audio waveform with the results also being stored in a similarity matrix [16].

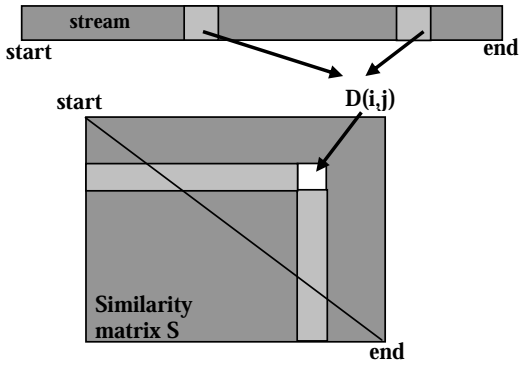


**Fig. 1.** Embedding an audio stream into a two dimensional similarity matrix [14]

### 2.3 Audio Complexity

Two inherent problems associated with MIR are the complexity of audio and the complexity of the query [10]. Music is a combination of pitch, tempo, timbre, and rhythm, making analysis more difficult than text. Structuring a query for music is made difficult owing to the varying representations and interpretations including natural transitions in music. Monophonic style queries usually perform better where simple note matching can be used whereas polyphonic audio files and queries simply compound the problem. Adding to the complexity of music structure and query structure is the method of analysis of audio.

The format of an audio file limits its type of use, different file formats exist to allow for better reproduction, compression and analysis. Hence it is also true that different digital audio formats lend to different methods of analysis. Musical Instrument Digital Interface (MIDI) files were created to distribute music playable on synthesisers of both the hardware and software variety among artists and equipment and because of its notational style allows analysis of pitch, duration and intensity [11]. An excellent tool for analysis of MIDI files is the MIDI Toolbox [12] which is based on symbolic musical data but signal processing methods are applied to cover such aspects of musical behaviour as geometric representations and short-term memory. Besides simple manipulation and filtering functions, the toolbox contains cognitively inspired analytic techniques that are suitable for context dependent musical analysis, a prerequisite for many music information retrieval applications.

However, reproduction of a MIDI file can vary greatly on different machines simply from differences between the composers and listeners equipment and it is

because of this it is not used for general audio playback. Pulse code modulation (PCM) is a common method of storing and transmitting uncompressed digital audio. Since it is a generic format, it can be read by most audio applications similar to the way a plain text file can be read by word-processing applications. PCM is used by Audio CDs and digital audio tapes (DATs). Support for WAV files was built into Windows 95 making it the de facto standard for sound on PCs. This format for storing sound in files in PCs was developed jointly by Microsoft and IBM.

One of the most common formats for audio compression is mp3, defined by the Moving Picture Experts Group (MPEG). The mp3 format uses perceptual audio coding and psychoacoustic compression to remove all the audio the ear cannot hear. It also adds a modified discrete cosine transform (MDCT) that implements a filter bank, increasing the frequency resolution 18 times higher than that of layer 2. The result in real terms is mp3 coding shrinks the original audio signal from a CD (PCM format) by a factor of 12 without sacrificing sound quality, i.e. from a bit rate of 1411.2 kbps of stereo music to 112-128 kbps. Because MP3 files are small, they can easily be transferred across the Internet. MPEG 7 [13] is a standardised description of various types of multimedia information. Where MPEG 4 defines the layout and structure of a file and codecs, MPEG 7 is a more abstract model that uses a language to define description schemes and descriptors – the Description Definition Language (DDL). Using a hierarchy of classification allows different granularity in the descriptions. All the descriptions encoded using MPEG 7 provide efficient searching and filtering of files.

## 3 Song Form Intelligence

Methods for error correction when packet loss occurs as discussed in Section 2.1 mainly try to minimise/prevent errors in the audio stream by masking missing or late packets with extra audio encoding or some method of interpolation to 'smooth' over the missing packets. Now we propose the use of pattern matching techniques within streaming audio to replace the lost segments.

The song header depicted in Fig. 2 describes a piece of music following a typical western tonal format (WTF), with a song form of intro (*I*), verse (*V*), chorus (*C*), verse (*V*), and chorus (*C*). It states that there is an introduction section of 10 seconds duration followed by a verse of 28 seconds, then a chorus of 32 seconds, then a verse of 28 seconds and finally repeats the chorus of 32 seconds.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *I* | 10 | *V* | 28 | *C* | 32 | *V* | 28 | *C* | 32 |

**Fig. 2.** Song form structure

This research proposes a novel syntax audio error concealment buffering technology, made possible by the song form structure with the possibility of developing this in the field of music semantics for replacing unidentified portions of

the song structure. Modelling of music has given rise to a number of different research angles such as modelling the human mind's conscious perception of rhythm and its syntax and semantics [17], [18], [19].

One of the problems associated with streaming audio is the time factor. The time between when a packet is received, placed into the buffer in the correct order and then used for playback can be anything between 10 seconds to as little as microseconds, depending on delays from bandwidth and congestion. It is at this point that ARQ techniques fail as there is simply no time left for a new packet to arrive. However if the segment already exists in previous sections already received replacements can be used.

Analysis of an audio file using the MPEG 7 description scheme will allow frames to be tagged with the data stored in the header file of each packet prior to broadcasting. These descriptors will be based on the similarity between different sections of the song. As can be seen in Fig. 2 an almost exact match can be obtained where the chorus is repeated. Identification of lost segments in the second chorus can simply be replaced with segments already received in the first chorus. By identifying the beginning of the missing segment, and then the end of the missing segment, a replacement can be found by matching the preceding and following segments with the same-length section in the previous chorus. To ensure smooth replacement for exact matches, the audio packets will need to be of the same duration, have the same start time, and the same end time for each segment. This should ensure error concealment occurrences within chorus sections will be inaudible to the listener. The overall SoFI system architecture can be seen in Fig. 3 showing analysis of the audio performed on the server with the packet replacement process being performed on the client.

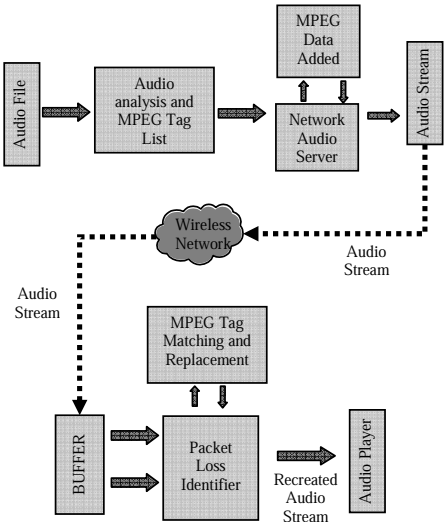


**Fig. 3.** System architecture

Not all songs have exactly repeating choruses, the underlying music and keys may be the same, but subtle differences in its pitch or even lyrics can have dramatic effects on matching segments. What appears to be the same to the human ear is very different when analysed by wavelength. With the use of 'best effort' matching, bursty losses can still be corrected with minimal perception to the listener.

Sections of the audio that contain lyrics that are different from any other section can still be repaired when bursty losses do occur. Pauses between words, phrases and sentences where only the music can be heard allow for repair in that repetition is inherent in WTF songs. Background instruments can follow the same repeated pattern throughout the entirety of the song. The following guitar chords are an extract of the chorus from R.E.M.s' "Everybody Hurts": *E minor / E minor / A / A / E minor / E minor / A / A / E minor / E minor / A / A*. Similarly the intro and verses contain a similar pattern: *D /G / D / G / D / G / D*. Pattern matching will require identification and grouping of these sections prior to streaming.

For 'best effort' matching of sections of the song with lyrics or unique sections of audio will be done using probability based on the already attached tags defining each of the sections and as near a possible match from previously received sections will be used to fill the missing segments. This will however reduce the overall quality of the audio signal, but based on the assumption that the typical length of bursty losses is no more than 1-2 seconds, it is acceptable for some degradation to occur. Studies of acceptable levels of audio quality have shown that listeners prefer to have some form of replacement rather than silence, by maintaining a rhythmic pattern using a percussive sound synthesiser to replace missing segments allows some continuity for the listener when dropouts do occur [20].

The minor increase in bandwidth from the inclusion of the MPEG 7 tags in the header sections of the packets can be justified based on the complexity of the analysis required. Calculations for real-time pattern matching will require a vast increase in processing demands on the system; to perform these at the same time as performing packet loss identification and matching lost sections, within the timescale of the buffer is not feasible.

## 4 Conclusion

In many ways the ideas presented in this paper are related to the field of FEC. However our point of departure and underlying methodology are different. Preliminary research indicates that by using similarity analysis and MPEG 7 we can identify and tag similar sections within an audio file and include the data in the audio stream when broadcast from the server. Pattern matching on the client side when receiving the audio allows error concealment through interpolation thereby placing the onus of error concealment on the client. Future research will address implementation and evaluation issues including packet replacement accuracy based on a comparative analysis of the initial song and the actual file received using both computer and human audible perception.

## References


1. Perkins, C., Hodson, O., Hardman, V.: A Survey of Packet-loss Recovery Techniques for Streaming Audio. In IEEE Network Magazine, Vol. 12, Issue 5 (1998) 40-48
2. Lee, K.K., Chanson, S.T.: Packet Loss Probability for Bursty Wireless Real-time Traffic Through Delay Model. In IEEE Transactions on Vehicular Technology, Vol. 53, Issue 3 (2004) 929 – 938
3. Mahanti, A., Eager, D.L., Vernon, M.K., Sundaram-Stukel, D.J.: Scalable On-demand Media Sreaming with Packet Loss Recovery. In IEEE/ACM Transactions on Networking, Vol. 11, Issue 2 (2003) 195 – 209
4. Varadarajan, S., Ngo, H.Q., Srivastava, J.: Error Spreading: a Perception-driven Approach to Handling Error in Continuous Media Streaming. In IEEE/ACM Transactions on Networking, Vol. 10, Issue 1 (2002) 139 – 152
5. Wang, Y., Ahmaniemi, A., Isherwood, D., Huang, W.: Content-based UEP: A New Scheme for Packet Loss Recovery in Music Streaming. In Proc. of Eleventh ACM International Conference on Multimedia, Berkeley, CA, USA. (2003) 412 – 421
6. Wah, B., Lin, D.: LSP-based Multiple-description Coding for Real-time Low bit-rate Voice Over IP. In IEEE Transactions on Multimedia, Vol. 7, Issue 1 (2005) 167 – 178
7. Leman, M., Clarisse, L., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G., Martens, J., and Van Steelant, D.: Tendencies, Perspectives, and Opportunities of Musical Audio-mining In Proc. of 3rd EAA European Congress on Acoustics, Seville, Spain (2002)
8. Gomez, E., Klapuri, A., Meudic, B.: Melody Description and Extraction in the Context of Music Content Processing. In Journal of New Music Research, Vol. 32, Issue 1 (2003)
9. Chai, W., Vercoe, B.: Structural Analysis of Musical Signals for Indexing and Thumbnailing. In Proc. of ACM/IEEE Joint Conference on Digital Libraries (2003) 27 – 34
10. Downie, J.S.: The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. In Computer Music journal, Vol. 28, Issue 2 (2004) 12 – 23
11. Doraisamy, S., Rueger, S.: A Polyphonic Music Retrieval System Using N-Grams. Presented at the 5th International Conference on Music Information Retrieval, ISMIR 2004, Barcelona, Spain (2004) 204 – 209
12. Eerola, T., Toiviainen, P.: *MIDI Toolbox: MATLAB Tools for Music Research.* University of Jyväskylä, Kopijyvä, Jyväskylä, Finland. Available at http://www.jyu.fi/musica/miditoolbox (2004)
13. Martinez, J.M., Overview of MPEG-7 Description Tools. IEEE Multimedia, Vol. 9, Issue 3 (2002) 83 – 93
14. Foote, J., Cooper, M.: Media Segmentation using Self-similarity Decomposition. In Proc. of SPIE Storage and Retrieval for Multimedia Databases, Vol. 5021 (2003) 167 – 175
15. Meredith, D., Wiggins, G.A., Lemström, K.: Pattern Induction and Matching in Polyphonic Music and other Multi-dimensional Data. In the 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI'2001), Vol. X (2001) 61 – 66
16. Bartsch, M. A., Wakefield, G. H.: To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY (2001) 15 – 19
17. Dannenberg, R. B., Hu, N.: Pattern Discovery Techniques for Music Audio. In Proc. of ISMIR 2002 M. Fingerhut, Ed., Paris, IRCAM (2002)
18. Mc Kevitt, P., O'Nuallain, S., Mulvihill, C., (Eds.): Language, Vision and Music - Selected Papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, Ireland. Amsterdam, The Netherlands: John Benjamins Publishing Company (2002)
19. Wiggins, G. A.: Music, Syntax, and the Meaning of 'meaning'. In Proc. of First Symposium on Music and Computers (1998) 18 – 23

20. Wyse, L., Wang, Y., Zhu, X.: Application of a Content-based Percussive Sound Synthesizer to Packet Loss Recovery in Music Streaming. In Proc. of 11th ACM International Conference on Multimedia (2003) 335 – 338