DIGITAL AUDIO WATERMARKING WITH SEMI-BLIND DETECTION FOR IN-CAR MUSIC CONTENT IDENTIFICATION

Ron Healy and Joe Timoney

Department of Computer Science, National University of Ireland, Maynooth, Co. Kildare, Ireland.

rhealy@cs.nuim.ie, jtimoney@cs.nuim.ie

Keywords:

Digital Audio Watermarking, Broadcast Monitoring, Blind Detection, in car entertainment

Abstract:

Recent developments in audio watermarking techniques have gone some way towards promoting an industry-wide acceptance of digital audio watermarking as a process that will eventually be used in all audio (and video) production. The predominant focus of such watermarking research has been in the area of content protection, because the prevention of illegal copying is an area of concern for content owners. However, digital audio watermarking may also be used for other purposes, such as the added-value option of real-time content identification of music. While computer-based users of music enjoy the opportunity to identify unknown audio using online tools, identification of audio in an offline domestic or in-car scenario is not so easily achieved. This paper discusses with an area of digital audio watermarking that would facilitate real-time in-car identification of the artists, title and/or other meta-data relating to music being broadcast by radio

1.0 Introduction:

When a track is played on radio that a listener is unfamiliar with, he or she is at the mercy of the presenter or producer of the radio show for identification of the track. However, in many cases there is no identification given. This occurs when a radio presenter plays a number of tracks in sequence and does not identify any/all of them, or when a radio station is using a computerised delivery platform – particularly in the so-called dead hours where there is no human involvement in the broadcast. It can be a frustrating situation when a listener hears a track they are very interested in and wants to know more about it but is given no information with which to identify the performer or track title. Assuming the listener wants to find out more about the artist, and may even want to buy the material they are listening to, there is no current way of facilitating this discovery and so, as a consequence, sales opportunities may be lost.

Digital audio watermarking at source will overcome this limitation and not only offer listeners the information they need to research the Artist or buy the material, but also offer an 'added value' technology for a manufacturer of both domestic and in-car entertainment devices as a user-centred selling point.

An audio watermark is embedded into songs, preferably at the time of production but perhaps even later, which will allow the real-time identification of the performer, track title and/or other content such as publisher details, ownership details etc. This information can be extracted and displayed on existing screens in both domestic and in-car audio entertainment systems, in a manner similar to the way radio station data is currently displayed. The information will be part of the actual audio content, rather than meta-data such as MP3 headers, so would be transmitted with the audio, even over an analogue transmission channel. Existing technologies allow the on-screen display of information about Artist and title but only from a digital source. This paper examines how this may be achieved in a traditional analogue radio environment, or in situations where an audio source (such as an iPod etc) is being transmitted via FM to a local audio device . This is a common setup in car audio use.

The techniques used in these processes make use of well known DSP algorithms such as Fast Fourier Transform and Goertzel Algorithm so there is a wide understanding of the concepts. The system suggested is a novel use of existing, well-understood mature DSP techniques. The conceptual direction of the work discussed is inspired by [12].

The process is described as semi-blind, rather than blind, since the decode system does require knowledge of some value or parameter used in the encoding process. However, it does not need access to the original unwatermarked host audio.

2.0 Overview of Digital Audio Watermarking:

Watermarking is the addition of some form of identifying mark that can be used to prove the authenticity or ownership of a candidate item. The commonest example would perhaps be a banknote, where a watermark is added in the production process to

validate the authenticity of the specimen. Alternatively, sample images may be distributed on paper with the author's name watermarked on them to limit the use of the image for illegal copying or misrepresentation. However, in digital watermarking, one of the priorities is that the watermark should *generally* not be visible (or audible, depending on the application domain) to any user who does not know it is there. In this way, watermarking is a form of Steganography, which is defined as *"Hiding a secret message within a larger one in such a way that others can not discern the presence or contents of the hidden message"* [9].

In the case described by this paper the host is a digital audio file, exactly as it would be on a CD, but the same procedural concept could also be applied to a video file. The first step is to create the actual watermark, which is a separate item than the host file. In some cases, for example where the goal of the system is to facilitate blind or informed self-authentication, the watermark may be derived from the host or some information contained in the host but this is not strictly necessary. Once the watermark is created, it needs to be added to the host in such a way as to be invisible or inaudible to end-users. This is the area where much research into watermarking is directed. The discovery of the watermark, and subsequent identification/decoding, is dependent on the watermarking and embedding techniques.

2.1 Creating the Watermark

For the purposes of this paper, the data to be watermarked and embedded into a host file is the Artist's name and the title of the track. This choice of data was not arbitrary and was instead guided by the intended use of the watermarked audio. However, any form of information can be embedded, with restriction only on the amount of such data. It can then be decoded without any knowledge of the original host content using pre-defined watermark parameters. The simplicity of the arrangement coupled with the uniform nature of the watermarking parameters allows for the process of 'Blind detection' in real-time. 'Blind detection' is the prime motivator for the development of this system since many previous efforts to use watermarking for the monitoring and identifying of broadcast output have tended towards audio fingerprinting, which relies on the availability of the original host audio, or some representation of it, at the point of identifying and decoding the watermark [11].

2.2 Initial DTMF Concept for watermark creation:

The initial hypothesis for the creation of the watermark in this scenario is outlined in Figure 1 and was inspired by [12]. It revolves around the well-understood Dual Tone Multi-Frequency (DTMF) [13] standard as used in touch-tone and mobile telephony, amongst other areas. The idea was to reduce the data to be watermarked to a series of bit-representations of its ASCII codes. Every character has a unique ASCII code and since all of the characters of the Artist's name and track title are simply letters, characters and digits, it made sense to use the ASCII numerical values, converted into Binary representation, as the basis for the watermark.

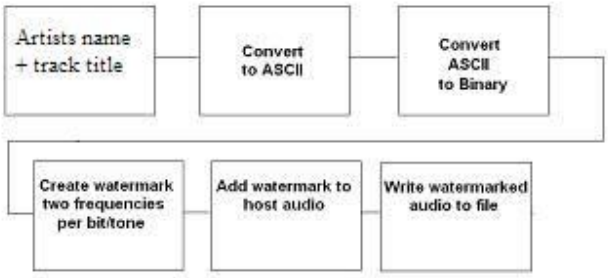


Figure 1: A block diagram of the first watermarking system

Once the Binary sequence was created, it was then to be used as the pattern for the creation of a pair of pure sine waves using the combined DTMF frequencies for 1 and 0. These are [13]:

DTMF 1 tone: 697Hz and 1209Hz combined

DTMF 0 tone: 941Hz and 1336Hz combined

Each bit (1 or 0) is represented by a very short DTMF tone exactly as can be heard on a touch-tone telephone when pressing either a 1 or 0 key on the keypad. The tones are then concatenated together to form a pattern representative of the 1s and 0s in the binary sequence representing the embedding information. The length of each tone is set to be suitably short duration of 25 milliseconds, although taking advantage of the limitations of the Human Auditory System this length is not necessarily fixed. The duration of each tone can be reduced to any length as long as it can still be detected in the decoding sampling process and this would increase the capacity of the watermarking scheme, however, shorter watermark tones may also cause more computational cost in the decoding phase. Reducing to the absolute minimum may increase the likelihood of missed tones to the point that the system simply could not decode the watermark

Initial experiments to create a watermark and then decode it again - independent of the intended host audio file - were very successful as would be expected. Encoding of the bit sequence and creation of the watermark was achieved by simply creating a pair of pure sine waves containing the two DTMF frequencies appropriate to the bit to be encoded. This process was repeated for the next bit in the sequence. Since the entire sequence was represented in bits, there were only two possible sinusoidal tones, each made from only two frequencies. To ensure that no waveform

discontinuities occurred when concatenating the sine waves for each subsequent tone the instantaneous frequencies of the pattern were first created, integrated to generate the phase, and then the sin of this phase was taken resulting in a smooth waveform. The watermark was then looped continuously to the length of the host audio. Finally, after writing to file (WAV format), reading in the file and decoding the watermark, this initial experiment resulted in a 100% recovery of the watermark, as would be expected in the absence of any corruption or attack on the watermark signal.

Once the watermark was added to the host file, however, the limitations of the system started to become obvious. Initially, adding the watermark was achieved by simple addition. This resulted in the watermark being audible. Since inaudibility of the watermark in the presence of a host signal is a constraint of any useful watermarking system, the watermark amplitude (noise level) was reduced in comparison to the amplitude of the host audio in order that the host audio might mask the presence of the watermark. Using this method, the watermark was embedded satisfactorily and a 'straw-poll' of five listeners to ten watermarked tracks suggested that the watermark could not be heard and listeners could not correctly identify which track was the original and which was watermarked.

The problem of isolating and subsequently decoding the watermark then became obvious. The initial intention was to analyse the candidate audio file in which a watermark was believed to be present to identify its actual frequency content. A simple iterative check of each 20ms 'block' of audio would determine the presence or absence of the frequencies sought. Presence of *both* frequencies meant that the corresponding bit (1 or 0) was present. Eventually, a sequence of bits was found and decoded back from ASCII to alphanumeric characters.

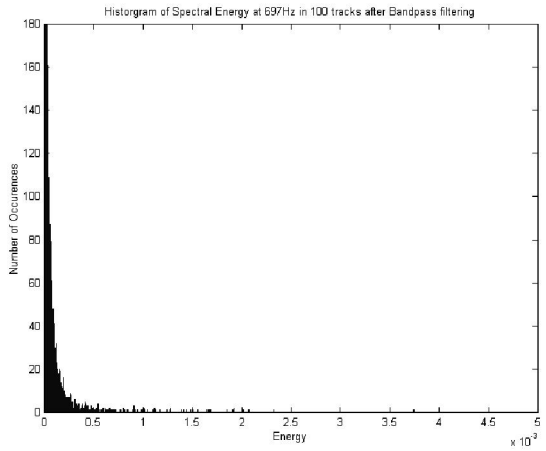
The first problem encountered was to identify which actual frequencies were present in the candidate audio. This was not as simple as might be thought. For example, the Fast Fourier Transform (FFT) method is limited in that is does not analyze a signal for the presence of particular frequencies only, rather because of its uniform sampling of the complete frequency spectrum it will only plot the strengths of components that may be close to but not at the frequency locations of interest. This problem was overcome by using the Goertzel algorithm [14] which can identify component energies at specific frequencies in a signal.

Another issue to be addressed was identifying where the bit sequence started and ended. Some form of synchronisation had to be used. It was decided to use the DTMF tone for the '*' key as a reference point to signal the start of the bit sequence. Similarly, it was necessary to identify where one bit ended and the next began because if the monitoring of the host audio did not commence from the very beginning of the track (which would be unlikely in a broadcast scenario) the tone duration would be useless as a measuring scale. Furthermore, a repeated bit in the sequence might not be easily identifiable as the decoding process had no way of discerning if the decoded tone represented one bit only or two instances of the same bit in sequence. It was decided to add the DTMF tone for the '#' key in between every bit in the bit-sequence. While this made it easy to identify where one bit-tone ended and the next one began, it also doubled the length of the watermark.

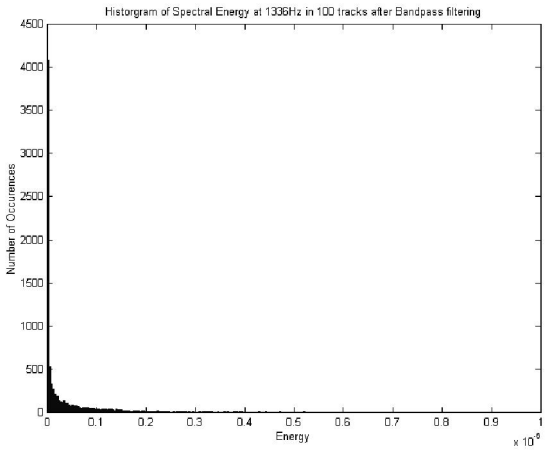
2.3 Problems encountered with pre-existing components

Once these changes were made, it became easier to identify the segments of audio that represented the bits in the watermark. However, there was another issue that needed to be addressed, namely what would happen in the case where the host audio, before being watermarked, already had high-power components at the frequencies used in the DTMF tones? This would not be a problem if the current frame of the host had only components at the frequencies needed to represent *the bit that was to be added* to it but if it already contained high powered components of the frequencies for the *other* bit, or the synchronising bit(s), this would lead to inaccurate decoding. If, for example, the frequencies associated with the '0' were inherently present in the host at a higher power than the frequencies that were deliberately added as part of the watermark for a '1' bit, the decoding process would naturally report them as being higher and assign the '0' value to that particular bit in the binary sequence, thereby altering the result.

At this point, since the matter had become an issue worthy of consideration, it was decided to perform experiments on a sample of audio files in various musical genres. Analysis of the actual frequency content at specific frequencies of 100 audio files was performed, using bandpass filtering followed by the Goertzel algorithm to try to ascertain whether there was any correlation between the powers of the desired DTMF frequencies. As can be seen from the illustrations in Figure 2(a) and 2(b), the energy of the components at the DTMF frequencies that separately identify the '1' and '0' from each other is very low. Nevertheless, there is some residual energy at each component and so no guarantee could be given before watermarking that the frequency being added would not be replaced in the decoding phase by a frequency component that was already present in the host before watermarking. It was clear that this method would not be successful in a real world scenario, mainly due to interference from existing frequencies in the host, which made the successful identification of watermark frequencies dependent on the content of the host before watermarking.



(a)



(b)

Figure 2: Histogram Plots of Spectral Energy at (a) 697Hz and (b) 1336Hz determined by bandpass filtering of 100 audio files

2.4 Weighting of frequency pairs

The next step was to weight the DTMF frequencies being used against each other so that both the frequencies for the '1' and '0' were embedded in the host audio simultaneously, rather than separately, but at powers that made it easy to decipher which one was the required bit. This experiment was based on work by [15], which illustrated the use of single frequencies to represent the bits for 1 and 0 whose powers were weighted by the total power of the frame into which they were being embedded. Various strengths and powers were considered and it was found that there was little difference between them, as long as the ratio of the powers of the desired bit to the undesired bit was significant enough to be detectable while the undesired bit was sufficiently powerful not to be 'lost'. When the weighted frequency-pairs were added to the host audio, weighted against the overall power of the audio in each host frame, the resultant decoding was significantly more successful. However, the issue still remained that in the event of a component being inherently present in the host audio before watermarking, it could negatively affect the recovery of the other bit.

Slight alterations were unsuccessfully attempted to circumvent this issues arising. For example, the actual frequencies chosen to represent the bits '1' and '0' were altered and evaluated on the premise that the higher the frequency, the easier it would be to identify due to its inherent power. The reasoning was that higher frequencies are harder to detect by the Human Auditory System (HAS) [16]. Using higher frequencies for the DTMF tones representing the bits did indeed increase the detection reliability of the system but conversely meant that the likelihood of these frequencies – and therefore the watermark – being lost in incidental 'attacks' such as DA/AD conversion, FM Transmission and perceptual encoding (e.g. MP3) was increased. Frequency pairs ranging from 8Khz to 14,635Hz were evaluated and successful decoding rates were compared. It was decided not to exceed a ceiling of 15Khz to minimise the potential for these frequencies to be lost in attacks, as mentioned above.

2.5 Applying a notch filter

By way of evolution of the technique, the idea of removing the existing frequency content of the host at the same frequencies as required by, and present in, the watermark, before actually adding the watermark to the host, was considered. Essentially, the host would have a 'hole' or 'notch' created at the two desired watermark frequencies, setting the power of any components at those frequencies to almost zero, and the aforementioned watermark with the same frequencies would then be included, weighted against each other. The effect of this would be to reduce the power of the component at the desired frequency if it exceeded the power of the same frequency in the watermark. Notch-filters were created which notched the host audio at the desired frequencies and 350 audio files from various contemporary, classical and traditional music genres were 'notched' at these separate frequencies. The watermark was then added to the notched audio. Figure 3 shows an example of a frame of audio under three conditions: (a) the original frame, (b) after notch filtering at 13875 Hz and 14625 Hz, and (c) following inclusion of the watermark signal.

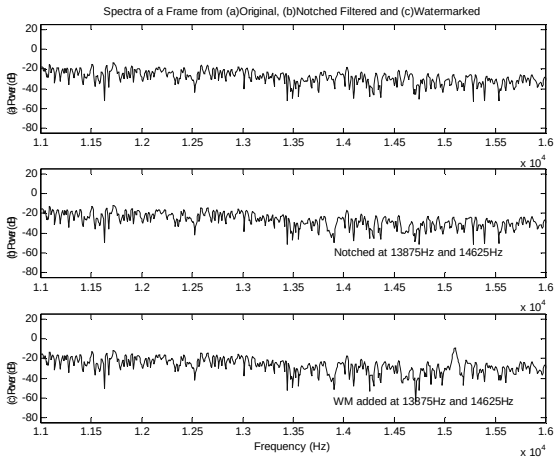


Figure 3: Spectral profile of a segment of audio, illustrating the original, the notched audio and the watermarked audio

Once this was completed, these files were analysed for watermarks, and the identified watermark decoded. Initial results were very promising and various modifications were made to increase the success rate. The most effective modification, as mentioned earlier, was the use of higher frequencies to create the watermark tones. Using frequencies in the region of 12Khz to 15Khz led to significantly increased success rates. Eventually, successful identification of the watermark was concluded for more than 98% of the 350 audio tracks analysed.

3.0 Conclusions:

The initial motivation in this work was to identify a simple yet efficient manner in which to watermark audio in such a way as to make it possible to perform blind or semi-blind detection on broadcast output and identify the audio being transmitted in real time. To this end, much progress was made. Initial attempts to use the same frequencies as specified in the DTMF standard were only partially successful but they did point towards other areas for investigation.

Altering the frequencies chosen for the watermark made the decoding response much more promising while weighting the frequency pairs for the desired and undesired bit against each other as well as against the host audio, and embedding both frequencies into the host audio also led to some improvements and pointed towards further development ideas. Creating a pair of notches in the host audio, essentially removing the components inherently present in the host at the desired frequencies before then adding those frequencies made the decoding process much more effective. Successful decoding rates of more than 98% were achieved in a sample of 350 audio files of various contemporary, classical and traditional genres.

This system is a deliberately simple implementation of the audio watermarking concept as it is envisaged that the identification of broadcast output must be possible in real-time but at low financial and computational cost. Any audio content that has been watermarked prior to broadcasting, whether by the producer, record label or Artist at the time of recording, or by the radio station before transmission, could be identified by decoding the watermark using pre-defined watermarking parameters. It would be a simple matter to then display this information on the small LCD or similar screen that is familiar in both domestic and in-car audio entertainment equipment or even on personal audio players with FM receivers such as the ubiquitous iPod etc.

Note also that the ability of a radio broadcaster to add a watermark itself, prior to broadcasting, opens up the possibility of using the watermark channel as an advertising medium, whereby information could be watermarked into the audio representing a generic, location dependent or content dependent advertisement or public service information. However, this would also remove any Artist or track information if such information *had* been previously watermarked, as the notch filtering will remove any components at the predefined frequencies, including existing watermarks. Care should be taken therefore to ensure that any additional information was embedded *along with* rather than *instead of* the Artist and track details.

3.1 Future work:

Investigation of perceptual shaping techniques and the ‘threshold of hearing’ as a means of reshaping and reformatting the watermark before addition to the host audio is an area for future research. Similarly, testing using the PEAQ audio quality assessment algorithm [17] and the more recent PEMO-Q assessment te4chnique [18] as well as subsequent listening-tests is planned on a wider range for the watermarked audio for both subjective and objective evaluation purposes.

4.0 References

[1] Cvejic, M. and Seppanen, T., “Digital Audio Watermarking Techniques and Technologies”, IGI Global, 2006

[2] Cano, P., et al., ‘A review of algorithms for audio fingerprinting’, *Proc. Of IEEE Workshop on Multimedia signal Processing*, 2002.

[3] Alsalami, M., and Al-Akaidi, M., ‘Digital audio watermarking survey’, *Proc. Of ESM 2003*, Nottingham, UK, 2003.

[4] The Recording Industry Association of America http://www.riaa.com/physicalpiracy.php

[5] World Intellectual Property Organization Copyright Treaty 1996 http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html

[6] The Economy of Culture in Europe http://www.keanet.eu/ecoculture/studynew.pdf

[7] Cano Rodríguez, G., et al, 'Analysis of Audio Watermarking Schemes'. *2nd International Conference on Electrical and Electronics Engineering (ICEEE) and XI Conference on Electrical Engineering*, Mexico, 2005.

[8] Kim, H.J.,. 'Audio Watermarking Techniques', *Pacific Rim Workshop on Digital Steganography*, Japan, July 2003.

[9] Steganography,. 'The Free On-line Dictionary of Computing'. http://dictionary.reference.com/browse/steganography

[10] International Standard Recording Code (ISRC), ISO 3901

[11] Oliveria, B., et al, 'Audio-based radio and TV broadcast monitoring', *Proceedings of the 11th Brazilian Symposium on Multimedia and the web*, 2005.

[12] Gopalan, K., et al, 'Covert Speech Communication Via Cover Speech By Tone Insertion', *Proc. of the 2003 IEEE Aerospace Conference*, Big Sky, MT, March 2003.

[13] Schenker, L, 'Pushbutton calling with a two-group voice-frequency code', *The Bell system technical journal* 39 (1): 235-255. 1960.

[14] Oppenheim, A., and Schafer, R., *Discrete-Time Signal Processing*. Prentice-Hall, 1990

[15] K. Gopalan, S. Wenndt. 'Audio Steganography For Covert Data Transmission By Imperceptible Tone Insertion', *Proc. the IASTED International Conference on Communication Systems and Applications*, Banff, Canada, July 2004.

[16] Zwicker, E. and Fastl, H., *Psychoacoustics: Facts and Models*, Springer, 1999

[17] ITU-R BS.1387, "Method For Objective Measurement of Perceived Audio Quality", 1998.

[18] Huber, R.; Kollmeier, B., **"**PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception", *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 14, Issue 6, Nov. 2006 Page(s): 1902 - 1911