Time Base Modulation: A New Approach to Watermarking Audio

Jonathan Foote, John Adcock and Andreas Girgensohn

FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue, Palo Alto, CA 94304
{foote, adcock, andreasg}@fxpal.com

ABSTRACT

A novel method is presented for inaudibly hiding information in an audio signal by subtly applying time-scale modification to segments of the signal. The sequence, duration, and degree of the time-scale modifications are the parameters which encode information in the altered signal. By comparing the altered signal with a reference copy, compressed and expanded regions can be identified and the hidden data recovered. This approach is novel and has several advantages over other methods: it is theoretically noiseless, it introduces no spectral distortion, and it is robust to all known methods of reproduction, compression, and transmission.

1. INTRODUCTION

Many techniques exist for the time-scale modification of audio[2,3]. This refers to changing the time duration of an audio sample without changing the pitch or other spectral characteristics. Because the pitch is not changed, small amounts of time scale modification are typically not noticeable. This is the basis for the watermarking approach described here. By modulating the time base of an audio signal, information can be undetectably encoded in it. As shown in Figure 1, short time regions of the signal are either compressed or expanded by an imperceptible amount (exaggerated in the figure for illustration). We call this method "time base modulation" as the underlying time basis is modulated by the watermark function. The sequence and degree of compression or expansion encode the watermark information. The watermark is detected by comparing the watermarked copy with the reference (unmarked) audio. Time-alignment of the watermarked and reference audio produces a "tempo map" that indicates how the time base of the watermarked audio has been altered. In regions of compression or expansion, the tempo map will deviate from a straight line and the embedded watermark data may be recovered from these deviations. Though this method will not work on audio with undetectable spectral change, such as silence, there are few compelling reasons to watermark such content.

The watermark can encode copyright information, a cryptographic signature, or information that specifically identifies a particular copy of the source audio. This is highly useful, for example, to hide encryption keys or for digital rights management. If each legitimate user of a copyrighted work is given a file with a unique watermark, the watermark found in illicitly distributed copies can identify the source[1]. Another application is to encode a cryptographic "hash" of the signal to verify its authenticity. Any alteration in the signal (such as inserting or deleting material) will generate a different hash value, which can be compared with the encoded value so that tampering can be detected [1].

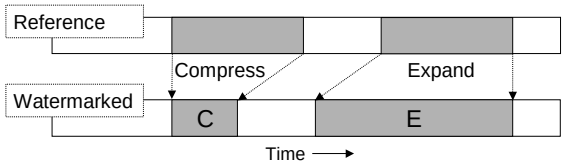

Figure 1. Watermarking a signal by time base modulation

Innumerable schemes are available for encoding a watermark as a sequence of time-scale modifications. It may be preferable to balance the amount of compression and expansion so that the watermarked version is exactly the same length as the reference signal, but this is not strictly necessary. Experiments have shown satisfactory results with compression/expansion ratios on the order of one to two percent, though this could be increased at the risk of introducing detectable artifacts. Overall, an encoding rate on the order of 8 bits per second is feasible, and is limited only by how objectionable time-scale modification becomes in the extreme. For many applications, such as speech, compression/expansion ratios of up to five or ten percent may be usable, leading to a corresponding increase in the encoded data rate.

2. PRIOR WORK

This method has several significant advantages over previous approaches. First of all, for most audio it will be virtually undetectable because of the human auditory system's insensitivity to extremely low frequency modulation. At the same time it is exceptionally robust to transmission and compression because current digital audio technology has a time precision on the order of several microseconds in an hour. Typical audio sources such as speech or music have enough natural variation that the artificial tempo changes introduced by the watermarking will be neither

[1]The authors do not believe that watermarking is a practical solution to current digital rights management problems, and are firmly opposed to the use of this or other technologies that infringe upon the Fair Use rights of the public.

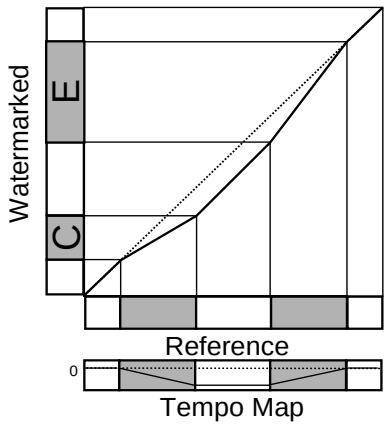

Figure 2. Recovering the "tempo map" from the dynamic alignment path of watermarked and reference signals.

audible nor easily detectable. Strictly rhythmic music produced by a computer sequencer or other mechanical device may be the exception to this rule. Tempo changes such as those inherent in analog recording and reproduction equipment will not interfere with the watermark. A consistent tempo change caused, for example, by an inaccurate playback speed will not disturb the watermark. Analog recording imperfections, commonly called "wow" and "flutter", occur on a timescale significantly shorter than the watermark tempo changes [9] and will tend to average out, leaving the watermark unaffected. A possible limitation is that the watermark can be partially obscured or degraded by intentionally changing the time scale of audio regions, or superimposing another tempo-based watermark, but this will not actually remove the watermark unless the exact inverse of the compression and expansion is known and used (which requires the reference audio to determine). If the watermark is being used as a digital signature, this alteration will invalidate both the watermark as well as the signature, and will be easily detectable. Note that few if any other watermarking schemes are robust under the application of multiple watermarks.

2.1 Existing audio watermarking techniques

Existing audio watermarking techniques can be broadly classified into data-domain and frequency-domain methods. Data-domain [7] methods work directly on the (possibly compressed) audio data and include: dithering the least significant bit in a PCM representation, hiding data within the bits of a compressed data file (compressed-domain watermarking), echo hiding [5] where a short time, low-amplitude echo is added to the signal, and amplitude modulation [8] where signal peaks are modified to fall within predetermined levels. Frequency-domain methods work by modifying the spectral content of a signal and include: frequency band modification, where frequency components are attenuated or enhanced, spread-spectrum [6] where information is disguised as low-amplitude noise, and phase coding [5] where the absolute phase offset of the signal is modified.

Compressed-domain watermarking schemes are not relevant here because they can not be applied to an uncompressed audio signal. By and large these schemes, particularly dithering and frequency domain approaches, are not robust to lossy forms of audio compression, since inaudible frequency domain modifications or dither are precisely the sort of information that is discarded when perceptual compression schemes such as MP3 are used. In contrast, time base modulation has been shown to be robust to both perceptual encoding and analog reproductions, and introduces no noise, echo, spectral or short-time phase distortion to the watermarked signal, nor does it remove or otherwise alter frequency components or signal amplitudes.

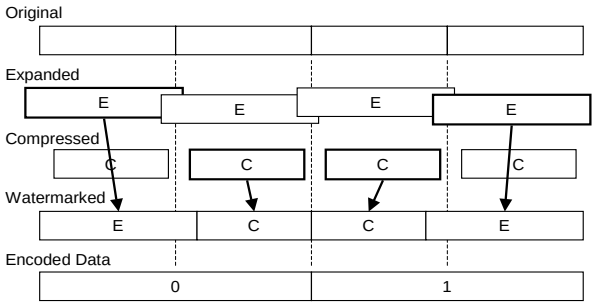

Figure 3. Construction of a 2-bit watermarked signal from compressed and expanded segments.

3. WATERMARKING VIA TIME BASE MODULATION

Techniques for changing the duration of an audio signal without altering its pitch are well-known and in common use. A common application is to play back audio at a higher rate, so that it may be auditioned or scanned in less time. A common time-scaling technique is based on the Short-Time Fourier Transform, but other methods such as the Phase Vocoder, Time Domain Harmonic Scaling, and Pitch-Synchronous Overlap Add (TD-PSOLA) are also widely used [2,3]. Any time-scaling method can be used for the time-base modulation watermarking procedure, though methods which are able to compress or expand by ratios very near one with few audible artifacts are preferred.

Consider the signal $x(t)$ expressed as the concatenation of K non-overlapping blocks, $x_{1\ldots K}$, with concatenation denoted by **C**.

$$x(t) \equiv \mathop{\mathbf{C}}_{k=1}^{K} x_k$$

The watermarked signal, $x_w(t)$, is generated by performing a compression or expansion of each block, x_k, by an amount denoted by E_k, and concatenating the modified blocks.

$$x_w(t) = \mathop{\mathbf{C}}_{k=1}^{K} TSM(x_k, E_k)$$

Where $TSM(x_k, E_k)$ indicates the time scale modification of block x_k by the amount E_k. In practice, care may be required to avoid introducing audible discontinuities at the block boundaries. This may be achieved by using a time scale modification algorithm that leaves data at or near the block boundaries unchanged, or by overlapping segments slightly and averaging data within the overlapping region during the construction of the watermarked signal.

The sequence of expansion/compression values, E_k, encodes the watermark, and can be recovered by comparing the watermarked

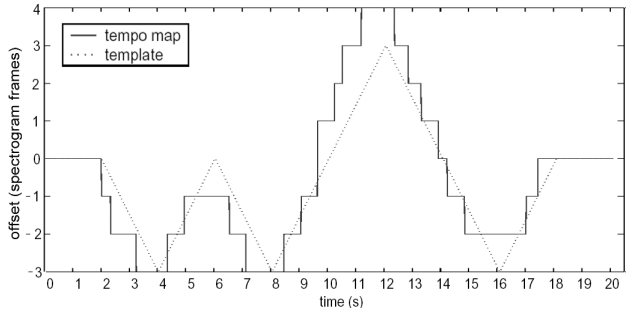

Figure 4. Recovered tempo map and expected template for binary 0010 watermark

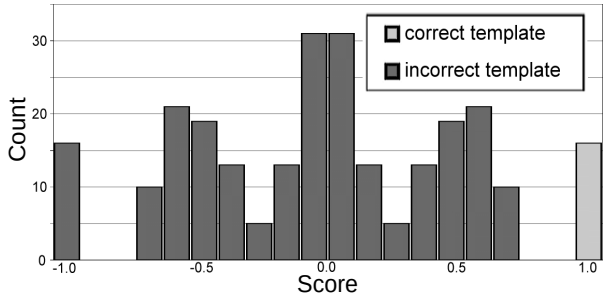

Figure 5. Histogram of similarity scores for watermark recovery experiment showing the correct template consistently outscoring all other templates.

version with the reference (unaltered) audio. This is done by finding the time-warping function that best matches the reference signal to the watermarked signal. Subtracting the linear component of the alignment path yields the "tempo map" from which the watermark information can be determined. After removing the linear component, the path has positive slope in the compressed regions, negative slope in the expanded regions, and slope zero in unaltered regions, but may be offset from zero by preceding compressions or expansions. This is illustrated in Figure 2. The slope of the recovered tempo map corresponds to the values of E_k used during encoding. This fact is used to recover watermarks later in this section; a tempo map can be estimated from the derivative of E_k and compared with the recovered watermark.

In the current implementation, reference and watermarked signals are matched using short-time spectral magnitude features. The analysis frames are 128 samples wide. For audio sampled at 22.05 kHz, this results in a 5.8 ms frame width and a frame rate of 172 frames per second. Each analysis frame is windowed with a Hamming window and converted to the frequency domain via a fast Fourier transform (FFT). The logarithm of the magnitude of the result is used as an estimate of the power spectrum of the windowed frame. The resulting vector of spectral components characterizes the spectral content of a window. The sequence of spectral vectors, represents the frequency content of the signal over time.

To find the best warping path, spectrograms for the reference and test signals are computed. The spectrograms are aligned by warping the reference to match the test signal using dynamic programming (DP, also referred to as "dynamic time warping"), which can be shown to find the optimal alignment path in quadratic time [4]. In this application the expected deviation from the diagonal is very small and the DP algorithm can be made to run in effectively linear time by only computing paths near the diagonal, or "cutting corners."

The DP algorithm is especially well suited the special conditions of this application; DP gracefully handles the case when test and reference files do not start and end at exactly the same time which would occur for example, if the watermarked signal were extracted from a continuous broadcast. Another feature is that the reference and test spectra need not match exactly, as long as they are more similar than nearby frames, thus the system is robust to modest distortion, such as low-pass filtering or additive noise. (Note that distortion or cropping severe enough to impact the watermark is self-defeating, as it will spoil the value of the original signal.) The DP method is also robust to a change in the playback rate (provided the resulting spectral distortion is small enough to allow for frame matching). The difference in playback rate is reflected as a change in the average slope of the recovered tempo map: in the case of a faster (slower) playback rate, the tempo map will have average slope less (greater) than 1.

The encoded data rate of the watermark is a trade-off with the detectability of the watermark and the degradation of the signal. We call the minimum length of a compression/expansion interval a "block." For simplicity, assume that all blocks are the same length. Each block is compressed or expanded by a factor $E_K = 1 + n_k\varepsilon$. Where n_k is a small integer, possibly negative. To reduce audible artifacts, it is advisable that the magnitude of n_k be limited to less than some small value: $-N \le n_k \le N$. Experiments have shown that reasonable values are a block length of 0.5 seconds, $\varepsilon \approx 0.01$, and $N = 2$ (*i.e.*: compression or expansion of up to 2%). Yielding a data rate of roughly $2\log_2(2N+1)$, or slightly more than 8 bits per second. While this is not a huge data rate, note that the typical 180 second popular song will encode 180 bytes; enough for song title, artist, publisher, and an ID code. When used as a watermark, 180 bytes yields more than 10^{400} individual identifiers, which is more than enough for any conceivable combination of source identifiers, device identifiers, and timestamps. We have not fully explored the space of possible window and block lengths and it is quite possible there may be more optimal values.

Informal listening tests from 13 volunteer listeners have confirmed the inaudibility of the time base modulation[1], though one "golden-eared" listener was able to detect artifacts from the time compression in direct A/B comparison with the reference audio. Superior time compression methods may have fewer artifacts. Experiments have also demonstrated that the watermark easily survives 64 kB MP3 encoding and decoding.

An experiment was performed to test the recoverability of watermarks. In this case, the reference signal was the first 20 seconds of the song "Magical Mystery Tour" by the Beatles, converted to a monophonic representation at 20,050 Hz sampling rate. A naive encoding scheme was used to encode a unique 4 bit watermark in 16 different copies of the audio. In this scheme, two, two-second blocks were used to encode one bit of information. A compression followed by an expansion represented a binary "one," while the

[1]Original and watermarked audio examples may be heard at http://www.fxpal.com/media/watermark/index.htm

reverse order indicated a "zero," as shown in Figure 3. (Obviously there are more efficient coding schemes, like those that use a region of no time scale modification to encode an additional state.)

Given this coding scheme, it is simple to generate a watermarked signal on the fly by concatenating compressed and expanded regions of the signal. Compressed and expanded versions of the signal were generated with a ratio of 2.5%. The original signal was evenly divided into 10 two second blocks. A watermarked signal was created by concatenating compressed and expanded blocks. Blocks at the beginning and end of the signal were not compressed, thus only the middle 16 seconds were altered. (In certain cases, this resulted in audible artifacts from mismatched block edges; in a practical system these could be eliminated by cross-fading the blocks over a short period.)

Given a known sequence of compression and expansion, it is straightforward to estimate what the tempo map should be in each case. Given a region of compression followed by expansion, the tempo map will speed up then slow again to zero, indicating a binary "one." Conversely, a binary "zero" will result in a tempo map feature that dips below zero in a "V" shape. Accordingly, "templates" were constructed having linear ramps corresponding to the expected tempo changes. Figure 4 shows the template and recovered tempo map for the copy watermarked with binary "0010." This corresponded to a compression/expansion sequence of CECEECCE, where "C" and "E" represent a compressed or expanded two-second block.

A tempo map was extracted for each watermarked version, and compared with the sixteen expected templates. Given a tempo map $\vec{m}$ and a template $\vec{t}$, a useful metric is the cosine of the angle between them:

$$D_C(\vec{m},\vec{t}) \equiv \frac{\vec{m} \bullet \vec{t}}{\|\vec{m}\| \|\vec{t}\|}$$

The experiment showed that the expected watermarks show a much higher cosine similarity with the expected template than with any other template. Of the 256 tempo map-template comparisons, the maximum cosine distance for the incorrect template was 0.618, while the minimum score for the correct template was 0.9094. Figure 5 shows the cosine distances for all 256 watermark-template comparisons. All correct templates had scores greater than 0.9, while all incorrect templates had scores less than 0.62. Note that it is not necessary to use templates; the watermark can be extracted by straightforward thresholding methods. For example, Figure 4 shows the binary digits could be extracted if thresholds were set at +/1 frames

4. FUTURE WORK

A particular advantage of this method is that it is computationally reasonable enough to encode the watermark on the client side, perhaps as part of a streaming media player. In this case, the server could send the watermark data as well as audio (possibly encrypted) to the client. The client would then uncompress and/or decrypt the streaming media and apply the watermark before buffering, playing, or locally storing the signal, thus reducing the computational load on the server. In fact, the watermark signal could encode information such as the time the file was played or an identifying code (such as IP or MAC address) of the client.

A drawback of this and many other watermarking methods is the need for the reference signal to decode the watermark. To use time base modulation watermarking in the absence of the reference signal the encoder and decoder must somehow identify common segments whose lengths can be used to encode the watermark. One way of doing this may be through tempo or beat analysis. Many methods exist to analyze the rhythmic tempo or speaking rate of audio [10]. Given a tempo detection method that is robust to the sorts of signal and data distortions discussed herein, the tempo or beat rate can be modulated using time-base modulation and then recovered without the reference signal.

An innovative approach to reference-less decoding is presented in a recent paper (published subsequently to the work presented here) from a group at the University of Minnesota. There, modification regions are chosen from long-scale signal changes, and are expanded or compressed to result in durations that are odd or even multiples of a smaller time unit. Watermarked data is recovered by detecting whether the regions are an odd or even length [11].

5. REFERENCES

[1] J. Zhao, E. Koch and C. Luo, "Digital Watermarking In Business Today and Tomorrow," in *Communications of ACM*. Vol. 41, No. 7, July 1998.

[2] S. Sprenger, "Time and Pitch Scaling of Audio Signals," http://www.dspdimension.com/html/timepitch.html

[3] S. Roucos, et al., "High Quality Time-Scale Modification for Speech," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1985, pp. 493-496.

[4] J. Kruskal and D. Sankoff, "An Anthology of Algorithms and Concepts for Sequence Comparison," in *Time Warps, String Edits, and Macromolecules: the Theory and Practice of String Comparison*, eds. D. Sankoff and J. Kruskal, CSLI Publications, 1999

[5] W. Bender D. Gruhl, N. Morimoto, A. Lu "Techniques for Data Hiding," in *IBM Systems Journal*, Vol 35, Nos 3&4, 1996, pp 313 -336

[6] I. Cox, J. Kilian, T. Leighton and T. Shamoon, "Secure Spread Spectrum Watermarking for Images, Audio and Video," in *Proc. of 1996 Int'l. Conf. on Image Processing (ICIP'96)*, vol. III, pp. 243-246, (1996).

[7] C. Xu & J. Wu, "Content-Based Watermarking for Compressed Audio," in *Proc. Recherche d'Informations Assistee par Ordinateur* [RIAO], April 2000.

[8] Wolosewicz, et. al., "Apparatus and method for encoding and decoding information in analog signals," United States Patent 5,828,325, Oct. 27, 1998

[9] S. J. Godsill, "Recursive Restoration of Pitch Variation Defects in Musical Recordings," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 2, pp 233-236, 1994

[10] J. Foote and S. Uchihashi, "The Beat Spectrum: A New Approach to Rhythm Analysis," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, Tokyo, 2001

[11] M. Mansour and A. Tewfik, "Audio Watermarking by Time-Scale Modification," in *Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, May 2001.