

DATA-HIDING IN AUDIO USING FREQUENCY-SELECTIVE PHASE ALTERATION

Rashid Ansari, Hafiz Malik, Ashfaq Khokhar

Dept. of Electrical and Computer Engineering University of Illinois at Chicago, Illinois, USA

ABSTRACT

A novel perception-based data hiding technique for digital audio is proposed. It exploits lower sensitivity of human auditory system (HAS) to phase distortion in audio compared with magnitude distortion. Audio is decomposed into subband signals some of which are selected for embedding data with a controlled alteration of phase using suitable allpass digital filters. The proposed scheme is robust to standard data manipulations yielding less than 2% error probability against compression, re-sampling, re-quantization, random chopping and noise addition. The proposed method is also robust to the desynchronization attacks.

1. INTRODUCTION

Digital media production and distribution has witnessed dramatic growth in recent years. The availability of the Internet, low-cost and reliable storage devices, and high-speed networks has made it possible to replicate and distribute digital information easily. This has created a need for protection and enforcement of intellectual property rights for digital media to prevent its illegal copying/reproduction. The need has spurred research in data hiding in digital media due to its applications in watermarking, annotation, and steganography [2].

The development of data-hiding methods requires many design and quality tradeoffs. An important requirement is that embedded data should be imperceptible. In addition data-embedding should be robust to standard digital data manipulations such as lossy compression, noise addition, and sampling rate conversion. Moreover, embedded data should be tamperproof against any adversary attacks.

Perception-based data hiding schemes for audio are influenced by properties of the human auditory system (HAS). In the past, different perception-based algorithms were proposed for data hiding/watermarking in audio data [4 -10]. These algorithms can be broadly classified according to the underlying technique of data embedding: perceptual masking [3, 4, 5, 9], direct sequence spread spectrum (DSSS) [3,4], and phase coding [3,7,8]. Algorithms based on phase coding [3, 7, 8] work well as far as imperceptibility of the embedded data is concerned, but suffer from some limitations e.g. some of them do not perform well against standard data manipulations and

most of them carry small payloads i.e. the amount of information embedded. For example, the phase coding technique proposed in [3] can embed only 16-32 bits of data in one-second duration audio samples. The algorithm based on echo-based coding [6] can embed about 40-50 bits of data in one-second duration of an audio signal.

In this paper, a novel method is proposed to exploit the HAS property that human auditory perception is largely insensitive to audio phase distortion [1] in a certain range of audible frequencies. In this method audio is decomposed into subband signals some of which are selected for embedding data with a controlled alteration of phase. The full audible frequency range i.e. 20 Hz~ 20 kHz is unsuitable for such data embedding. In the higher frequency range it is hard to detect phase changes reliably. The higher frequency range contains insignificant signal energy and perceptual auditory model based compression techniques, such as MP3 [11], generally discard information in these frequency bands due to perceptually inaudible distortion. Moreover, low signal energy makes detection susceptible to error due to additive noise in the higher frequency range. In order to make embedded data more robust and resilient to standard data manipulations, the frequency range should be carefully selected to ensure imperceptibility as well as robustness of the embedded information.

In our proposed method, a frequency range suitable for data hiding is first selected. The signal content in this range is partitioned into subbands using discrete wavelet packet analysis filter bank (DWPA-FB). Next the data is embedded in audio by modifying its phase in selected frequency bands using suitable allpass digital filters (APF) [1]. For synchronization, input audio data is analyzed to extract salient points characterized by fast tonal transitions. These salient points are attack sensitive regions too [9]. Therefore, synchronization based on these salient points can withstand desynchronization attacks such as random sample chopping, or intentional attacks such as randomly added or deleted samples from audio with embedded data. The frequency-selective phase alteration (FS-PA) technique can reliably embed more than 1000 bits of data in an audio segment of one-second duration, which is 10-15 times more compared with the existing methods. The proposed technique is robust to standard data manipulations yielding less than 2% error probability for lossy compression, noise addition, etc.

2. ANALYSIS FOR SALIENT POINT EXTRACTION

Frame level content features of audio data such as short-term energy ratio or fast energy climbing points, zero crossing rate and spectral flatness measure (also known as salient points) can be used to withstand against desynchronization attacks. Salient points are attack sensitive locations [9] where one can embed synchronization code that a detector can use for resynchronization and blind detection. Salient points in the audio signal are extracted by examining audio features to which HAS is sensitive [1]. If an adversary alters these features, audible distortion is introduced. A desirable attribute of a salient point extraction method is that approximately the same salient points should be extracted before and after common information hiding attacks.

The salient point extraction technique proposed in [9] based on fast energy transition feature is computationally demanding. An efficient and effective approach is to use the spectral flatness measure (SFM) for salient point extraction. SFM helps to distinguish tone-like or noise-like nature of an audio signal. SMF is defined as the ratio of geometric mean to arithmetic mean of the energy per critical band and generally expressed in dBs.

$$SFM = 10 \log_{10} \left\{ \frac{\left(\prod_{b=1}^{b_t} E_b \right)^{1/b_t}}{\left(\frac{1}{b_t} \sum_{b=1}^{b_t} E_b \right)} \right\} dB \quad (1)$$

where b_t is the total number of critical bands in the signal and E_b is the energy in each critical band.

SFM varies between 0 and 1, i.e. $0 \leq SFM \leq 1$. SFM close to 1 corresponds to pure-tone like characteristics while SFM value close to 0 corresponds to noise-like characteristics. The procedure to extract salient points based on SFM is outlined here:

Consider an audio frame with N samples: $x(n)$, $n = 1, 2, \dots, N$. Let $X(k)$, $k = (0, 1, \dots, N/2)f_s/N$ be the corresponding discrete fourier transform (DFT), where f_s denotes the sampling rate.

A salient point based on SFM is estimated by checking if: $|d(SFM(n))| > Th_1$. Here $d(\cdot)$ is the first difference.

This index is then labeled as fast tonality transition point. It has been observed that these fast transition points generally appear in a group [9]. If two groups are separated by less than a threshold, Th_2 , then they are merged into a larger group and the strongest tonality transition point in i -th group is labeled as a salient point SP_i . Thresholds Th_1 and Th_2 are selected such that on average 4-6 salient points are extracted per second in order to allow adequate synchronization rate.

3. DATA EMBEDDING

The salient points extracted before and after audio processing such as compression are generally not exactly

identical. Therefore, data recovery and resynchronization would be hard if data is embedded in time domain. This problem can be effectively addressed by embedding data in the frequency domain [9].

The proposed data embedding scheme is transformed domain based. Audio data is segmented into non-overlapping blocks of P samples (10 msec duration), starting from the first salient point in the salient point list. Each block is then decomposed into 2^l subbands using an l -level DWPA-FB. Twelve-subbands in the middle frequency range i.e. starting from subband number 4 (sb_4) to subband number 15 (sb_{15}) are selected for data embedding, to meet the perceptibility and robustness constraints of the data embedding scheme.

Channel encoded data is embedded in the selected subband signal by simply passing the signal through a suitable configuration of APFs with distinct patterns of pole-zero pairs. Bit '0' or '1' is embedded by passing a signal through an n^{th} order APF with transfer functions $H_0(z)$ or $H_1(z)$ respectively. This n^{th} order ($n=0,2,3\dots$ for our implementation) APF is implemented using cascaded realization of 2^{nd} order complex APF, which is characterized by parameter p_i defined as: $p_i = re^{j\omega_i}$,

where $0 < r < 1$, $0 < \omega_i < \pi$ and $i = 0, 1$ and the transfer function $H_i(z)$ is given by:

$$H_i(z) = \frac{[(z^{-1} + p_i)(z^{-1} + p_i^*)]}{[(p_i z^{-1} + 1)(p_i^* z^{-1} + 1)]} \quad (2)$$

This is 2^{nd} order APF with poles at p_i and p_i^* and zeros at $1/p_i$ and $1/p_i^*$. Figure 1 illustrates the pole-zero layouts of the 2^{nd} order APFs H_0 and H_1 used for data hiding.

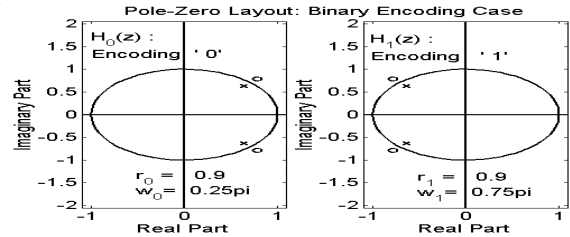


Figure 1: Pole-Zero layouts of APFs $H_0(z)$ and $H_1(z)$.

Consider the processing of the sequence $x_{k,j}[n]$ corresponding to the j^{th} subband of k^{th} block for embedding bit '0'. Let $y_{k,j}[n]$, denote the output, and given as: $y_{k,j}[n] = x_{k,j}[n] * h_0(n)$ (3)

Here '*' stands for convolution, $4 \leq j \leq 15$ and $k=1,2\dots M$. where M is the cardinality of the salient point list. And $h_0(n)$ is the impulse response of $H_0(z)$.

APFs have constant magnitude response and a phase response which is controlled by pole-zero placement. These properties are useful for perception-based data hiding by the controlled alteration of the phase of an audio signal within a suitable frequency range.

In practice, however, the use of APFs has some limitations due to truncated support of signals after processing. For example APF can introduce audible distortion that depends on factors such as the parameters

p_i , the order of the APF, and the length of the input audio block which ultimately influence the probability of error P_e at receiver. Placing pole-zero pairs of a high order APF close to the unit circle can cause magnitude distortion in a finite-duration of the output. That leads to an undesirable distortion, but to a better detection performance. On the other hand, poles placed close to the origin do not introduce audible distortion but are generally hard to detect. Longer blocks improve performance in estimating APF parameters but reduce the amount of data that is embedded into the given audio clip. The selection of APF parameter and sequence length is determined by the tradeoff between the amount of data embedded, perceptibility of the data embedded in the audio and the probability of error of the received data.

The following steps summarize the scheme:

- A list of salient points is extracted for a given audio signal using analysis discussed in section 2.
- Starting with the first salient point, the audio signal is segmented into frames of 10ms duration.
- Each frame is decomposed using a 5-level DWPA-FB and 12 subbands from sb_4 to sb_{15} are selected for data embedding.
- One bit data is embedded in each selected subband in a frame by processing it with the filter $H_0(z)$ for information bit ‘0’ or $H_1(z)$ for bit ‘1’.
- All frames that contain salient points are embedded with synchronization code consisting of a specific sequence, e.g. all ones.
- After data embedding each frame is resynthesized using a discrete wavelet packet synthesis filter bank.

Detailed block diagram for data embedding algorithm is given in Figure 2.

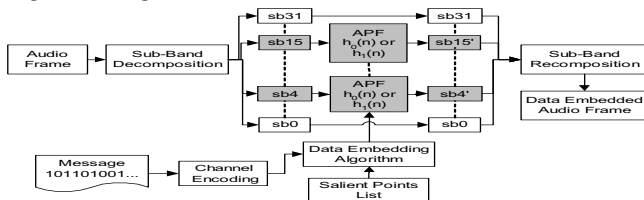


Figure2: Block Diagram of the Data Embedding Scheme

4. DATA DETECTION

The recovery of embedded data requires the detection of APF parameter p_i used for data embedding in each subband. The first step in data detection is the analysis of the data embedded audio to extract the set of salient points (as discussed in Section 2). Starting from first salient point in the set, the audio signal is segmented into non-overlapping frames of P samples. Each frame is decomposed using 5-level DWPA-FB. Then twelve subbands i.e. from sb_4 to sb_{15} are selected for data detection. The detector evaluates finite length Z-transform of the selected subband at all possible values of APF parameter (i.e. at p_0 and p_1 for our implementation). The

detector in advance does not know which APF was used for data embedding in the received subband sequence; however, detector has knowledge of the parameters of APFs used for data embedding i.e. p_0 and p_1 . The decision for bit ‘0’ or bit ‘1’ is made based on calculating Z-transform of the sub-band sequence at $1/p_0$ and $1/p_1$ (zero-tracking) and then estimating the local minima from the magnitude spectrum.

In practice zero-tracking is generally used for APF parameter estimation because theoretically output of a stable and causal APF is an infinite sequence, i.e. $y_{k,j}(n)$ output of the APF in Eq. 3, which is the convolution of a finite length input sequence $x_{k,j}(n)$ and an infinite sequence $h_0(n)$ (as $H_i(z)$ a rational function of z). But for APF parameter estimation only a finite length sequence is available at the detector input.

Let $\check{y}_{k,j}(n)$ be a finite length approximation of an infinite sequence $y_{k,j}(n)$ by dropping higher indexed terms, available at the detector input. This approximation is valid only if $Y_{k,j}(z)$ converges $\forall z \in C$ which is true if $z = 1/p_0$ (zeros of APF). This fact is illustrated in Figure 3, Figure 3 (right, up) shows the plot of chirp z-transform (CZT) of a finite length subband sequence $x_{4,6}(n)$ before and after passing through APF $H_0(z)$, calculated at $r = 1/0.9$, clearly minima in Figure 3 (right, bottom) occurs at $\omega = \omega_0 \sim 0.25\pi$; similarly, Figure 3 (left, up) shows the CZT of the same sequence before passing through APF $H_0(z)$ calculated at $r = 0.9$, and 3(left bottom) after passing through $H_0(z)$ but there is no maxima at $\omega = \omega_0 \sim 0.25\pi$, this might be the reason that finite length approximation of an infinite length sequence at $z = p_0$ is not accurate.

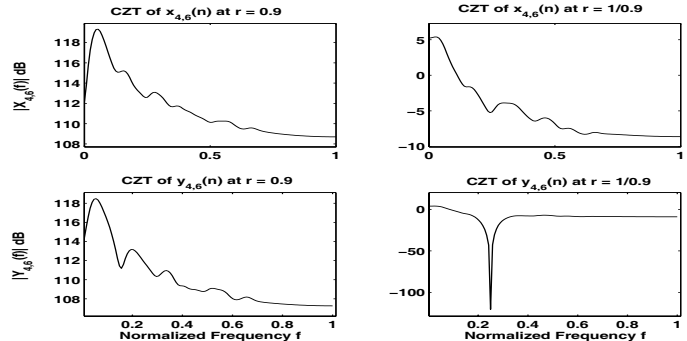


Figure 3: Magnitude spectrum of CZT of the subband sequence $x_{4,6}(n)$ before and after passing through $H_0(z)$ i.e. $y_{4,6}(n)$, at $r = 0.9$ (right) and at $r = 1/0.9$ (right).

Therefore detector uses zero-tracking for APF parameter estimation (ω_i , as $r = 0.9$ which is fix in our implementation) for data detection. For parameter estimation we need to estimate only ω which is done by estimating local minima magnitude response of CZT of the selected subband calculated at $r = 1/0.9$, and then based on the nearest neighborhood hypothesis bit ‘0’ or bit ‘1’ is decided. Finally received data is channel decoded to recover the embedded information. The block diagram in figure 4 illustrates the data detection process in detail.

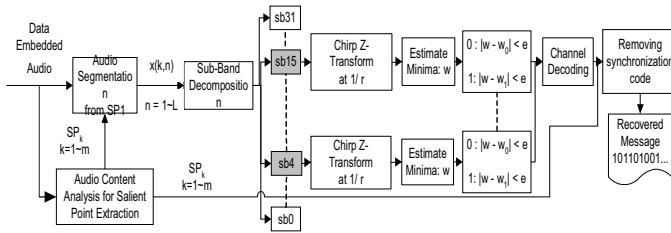


Figure 4: Block Diagram of the Data Detection Process

5. SIMULATION RESULTS

The proposed data hiding scheme divides input audio into 10 msec audio frame. Twelve-subbands (as discussed in Section 3) are selected after subband decomposition for data hiding. Hence embedded data = $12 \times 100 = 1200$ bps, which is 10-15 times more than the existing data embedding methods [5- 9]. We applied the proposed data hiding scheme to the variety of music clips having diverse frequency characteristics. Imperceptibility and robustness are the two benchmarks used for performance evaluation of the proposed data hiding scheme. Robustness is measured based on the probability of error in the received data under different constraints, these constraints include: a) noise addition, b) lossy compression, c) random chopping and d) resampling.

For robustness test we have following observations:

- White Gaussian noise with 0% to 30% of the audio power is added into data embedded audio signal. The probability of error of the recovered data for different values of signal to noise ratio (SNR in dB) is plotted in Figure 5.

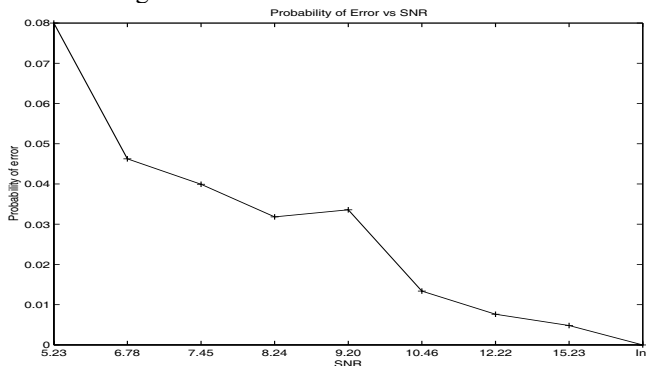


Figure 5: Probability of Error for different SNR values.

- Data embedded audio signal is compressed using MPEG layer III coder. Despite the lossy compression the p_e value of the recovered data was below 2%.
- To test robustness against desynchronization attacks, one sample out of every 100 samples of the data-embedded audio is dropped. Probability of error was 1.69%.
- In this test data-embedded audio is down-sample to 22.05 kHz and then interpolated to 44.1 kHz. Probability of error of the recovered data after resampling was 2.01%.

6. CONCLUSION

We have proposed a novel method of data hiding based on phase alteration in selected signal subbands. The proposed technique is robust to standard data manipulations yielding less than 2% error probability against noise addition, compression, random chopping and re-sampling. The error probability performance can be improved further by using channel coding scheme with higher error correction capability. Data detection results show that proposed scheme can embed data 10-15 times more data in a unit duration audio samples compared with existing schemes [3 - 9], but this was based on informal tests for imperceptibility. The performance will be evaluated using formal listening tests that are being set up.

We are currently investigating data hiding techniques based on FS-PA for other types of multimedia data such as, images and videos. We are also working of FS-PA based digital watermarking technique for audio data.

7. REFERENCES

- [1] D. A. Nelson, and R.C. Bilger, "Pure-Tone Octave Masking in Normal-Hearing Listeners," *J. of Speech and Hearing Research*, Vol. 17 No. 2, June 1974.
- [2] F. A.P. Petitcolas, R. J. Anderson, M. G. Kuhn, "Information Hiding - A Survey," *Proc. of IEEE*, vol. 87, No. 7, pp. 1062-1078, July 1999.
- [3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol.35, nr. 3/4, 1996.
- [4] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, pp. 337-355, 1998.
- [5] P. Bassia and I. Pitas, "Robust audio watermarking in the time domain," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP99)*, 1999.
- [6] M. F. Mansour, and A. H. Tewfik, "Time-scale invariant audio data embedding," *Proc. IEEE International Conference on Multimedia and Expo, ICME, Japan, August 2001*.
- [7] Y. Yardimci, A. E. Cetin, and R. Ansari, "Data hiding in speech using phase coding," *Proc. Eurospeech Conference, 1997*.
- [8] D. Radakovic, "Data hiding in speech using phase coding," *Master thesis, ECE Dept. UIC, 1999*.
- [9] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo, "Robust Audio Watermarking for Copyright Protection," *SPIE's 44th Annual Meeting Advanced Signal Processing Algorithms, Architectures, and Implementations*, July 18-23, 1999.
- [10] C. I. Podilchuk and E. J. Delp, "Digital watermarking algorithms and applications," *IEEE Signal Processing Magazine*, pp. 33-45 July, 2001.
- [11] D. Pan, "A Tutorial on MPEG / Audio Compression," *IEEE Multimedia Magazine*, 2(2):60-74, 1995.