Audio Steganography by Cepstrum Modification

Kaliappan Gopalan

Department of Electrical and Computer Engineering
Purdue University Calumet
Hammond, IN 4323

email: gopalan@calumet.purdue.edu

Abstract: A method of embedding information in the cepstral domain of a cover audio signal is described for audio steganography application. The proposed technique combines the commonly employed psychoacoustical masking property of human auditory system with the decorrelation property of speech cepstrum, and achieves imperceptible embedding, large payload, and accurate data retrieval. Results of embedding using a clean and a noisy hot utterance show the embedded information is robust to additive noise and bandpass filtering.

1. Introduction

Audio steganography, or information hiding in audio signals, is gaining widespread importance for secure communication of information such as covert battlefield data and banking transactions via open audio channels. On another level, watermarking of audio signals for digital rights management is becoming vital to prevent illegal copying, file sharing, etc. A steganography system, in general, is expected to meet three key requirements, namely, imperceptibility of embedding, correct recovery of embedded information, and large payload. Practical audio embedding systems, however, face hard challenges in fulfilling all three requirements due to the large power and dynamic range of hearing, and the large range of audible frequency of the human auditory system (HAS). These challenges are more difficult to surmount than those faced by image and video steganography systems due to the relatively low visual acuity and large cover image/video size available for embedding.

One of the commonly employed techniques to overcome audio embedding limitations due to the acute sensitivity of the HAS is to alter the cover speech spectrum in the auditorily masked regions [1 - 3]. Frequency masking phenomenon of the psychoacoustic masking property of the HAS renders weaker tones in the presence of a stronger tone (or noise) inaudible. A large body of embedding work has appeared recently with varying degrees of imperceptibility, data recovery and payload, all exploiting the frequency masking effect of the HAS for watermarking and authentication applications. Several steganography methods using indirect exploitation of frequency masking have been recently proposed with varying degrees of success [4-6]. These methods typically alter speech samples by a small amount so that inaudibility is achieved without explicitly modifying masked regions.

This paper reports on embedding in the cepstral domain of a cover audio signal by altering the speech cepstrum at frequencies that are in the spectrally masked regions. In the following sections we briefly review the cepstral domain speech processing and embedding, and present a method of modification of cepstra with and without combining frequency masking for efficient and imperceptible embedding.

2. Cepstral Domain Speech Processing and Embedding

Cepstral domain features have been used extensively in speech and speaker recognition systems, and speech analysis applications. Complex cepstrum $\hat{x}[n]$ of a frame of speech $\{x[n]\}$ is defined as the inverse Fourier transform of the complex logarithm of the spectrum of the frame, as given by

$$\hat{x}[n] = IDFT\left[\ln\left\{DFT\left(x[n]\right)\right\}\right] \qquad (1)$$

Since

$$\ln\left[X\left(e^{j\omega}\right)\right] = \ln\left[E\left(e^{j\omega}\right)\right] + \ln\left[H\left(e^{j\omega}\right)\right], \qquad (2)$$

where

$$X\left(e^{j\omega}\right) = DFT\{x[n]\} = E\left(e^{j\omega}\right)H\left(e^{j\omega}\right), \qquad (3)$$

The work on this paper was supported by the Air Force Research Laboratory, Air Force Material Command, USAF, under research grant/contract number F30602-03-1-0070)

and $e(n) \Leftrightarrow E\left(e^{j\omega}\right)$ and $h(n) \Leftrightarrow H\left(e^{j\omega}\right)$ are the discrete Fourier transform pairs from the speech signal production model, $x(n) = e(n)*h(n)$. The cepstrum in (1) effectively separates the excitation source $e(n)$ from the vocal tract system $h(n)$.

This additive separation indicates that modification for data embedding can be carried out in either of the two parts of speech. Imperceptibility of the resulting cepstrum-modified speech from the original speech may depend upon the extent of changes to the pitch and/or the formants, for instance. Any modification carried out in the cepstral domain in accordance with data alters the speech source, system, or both, depending on the quefrencies involved.

Prior work employing cepstral domain feature modification for embedding includes statistical mean manipulation [7], and adding pseudo random noise sequence for watermarking [8]. More recently, Hsieh and Tsou [9] showed that by modifying the cepstral mean values in the vicinity of rising energy points, frame synchronization and robustness against attacks can be achieved. Based on the robustness of cepstrum to signal processing operations, the present work proposes embedding by altering the cepstrum – rather than the mean – in regions that are psychoacoustically masked to ensure imperceptibility and data recovery.

3. Cepstrum Modification for Embedding at Masked Frequencies

To combine the masking effect on hearing, initially the log spectrum – Eq. (2) above – was modified in accordance with data as the log spectrum is closely related to the cepstral domain representation. Log spectral values at two common masked frequency points *f1* and *f2* were altered by one of two fixed ratios from their masking threshold values. For bit 1, log spectrum at *f1* was raised to 80 % of threshold and that at *f2* was raised to 40 %, regardless of their original (masked) values. Complementary values were used for bit 0. Data at the receiver were retrieved based on the relative values of log spectrum at *f1* and *f2* – both in the masked region – compared to their masking thresholds. Although the method was successful in imperceptible embedding and correct data recovery, payload was low due to fewer frames having the same pair of masked spectral points. In the next method, the mean of the cepstrum of a selected range of quefrencies was modified in a nonreturn-to-zero mode – increased by a fraction of the peak cepstrum for embedding a bit 1, or decreased for 0. Successful embedding in terms of inaudibility and data recovery with low BER was observed in this case with one bit per frame [10].

To improve BER further and to embed at specific points in a host audio, a two-step procedure is proposed. In the first step, a pair of masked frequencies that occur most frequently in a given host audio is obtained as follows. For each frame of cover speech, power spectral density (PSD) and masking threshold (in dB) are determined and the frequency indices at which the PSD is below a set dB are obtained. To avoid altering silence intervals between phonemes or before plosives, or low energy fricatives, only those frames that have a minimum number of masked points are considered. (Alternatively, frames with low energies may be excluded.) For the entire length of cover speech, a count of the number of occurrences of each frequency index in the masked region of a frame is obtained. From this count, a pair of the two most frequently occurring spectral points are chosen for modification.

Alternatively, the spectral points that are the farthest from the masking threshold of each frame are obtained. These points have the largest leeway in modifying the spectrum or cepstrum in most of the frames of the cover speech.

In the second step, complex cepstrum of a sinusoid at each of the two selected frequencies *f1* and *f2*, which form a key, are obtained with the maximum amplitude of the sinusoid set to the full quantization level of the given cover speech. For each frame of speech that is to be embedded (that is, the frame does not correspond to silence or low energy speech, as determined in the first step with fewer masked points), its complex cepstrum is modified as follows.

Initialize: *Spectrum at f1 and f2 = mean of frame spectrum at f1 and f2*

To embed a 1: $mod\_cep = cep + \alpha.c1 - \beta.c2$ (4a)

To embed a 0: $mod\_cep = cep - \alpha.c1 + \beta.c2$, (4b)

where

cep = original cepstrum of frame
c1 = cepstrum of sinusoid at frequency f1, and
c2 = cepstrum of sinusoid at frequency f2

The parameters α and β are set to low values (one-tenth, empirically, for example), or based on a fraction of frame power. Since the two frequencies are in the masked regions of most frames, adding or subtracting low level cepstra at these frequencies ensures that the modification results in minimal perceptibility in hearing. If no bit is to be embedded, as in the case of excluded frames, only the initialization step is carried out.

Modified frame cepstrum is transformed to time domain and quantized to the same number of bits as the cover speech for transmission.

At the receiver, embedded information in each frame is recovered by the spectral ratio at the two frequencies, *f1* and *f2*. Since an unembedded frame is transmitted with the

same spectral magnitude at *f1* and *f2*, the ratio at the receiver is close to unity; hence, no bit is retrieved. Thus the recovered bit *rb* is given by

$$rb = \begin{Bmatrix} 1, \text{ if } \log\left|\dfrac{X(f1)}{X(f2)}\right| \geq b1 \\ 0, \text{ if } \log\left|\dfrac{X(f2)}{X(f1)}\right| \geq b0 \\ -1 \ (no\ data), \text{ else} \end{Bmatrix}, \qquad (5)$$

where b0 and b1 are set close to unity. The indices of the embedded frames need not be transmitted or specified at the receiver. Additionally, by embedding only in selected frames, a second key can be incorporated for added security.

4. Experimental Results

The above two-step procedure was applied to (a) a clean host speech from the TIMIT database, and (b) a noisy utterance from an air traffic controller (ATC) database. For the clean speech sampled at 16,000 per second with 16 bits per sample, the first step of finding masked spectral points yielded a set of eight frequencies that were in the masked region in 100 frames or more out of a total of 208 frames (with 512 points per frame and 256-point overlap). The frame PSD at these masked frequencies was at least 3 dB down from their corresponding threshold sound pressure levels. Two of the eight frequencies were chosen for cepstrum modification. From the alternative set of masked frequencies – those that occurred with at least five other frequencies – frames that had fewer than six masked points were excluded from embedding. This exclusion ensures that any small change in the embedded PSD at the two selected frequencies is not likely to be noticeable in audibility or spectrogram as being different from other masked points.

With *f1= 906.25 Hz* and *f2 = 1218.8 Hz* (which occurred in the masked regions along with five other frequencies), and excluding 29 frames from embedding (because of low energy and/or in V/UV transition), the remaining 179 frames were embedded with (a) bit 0, (b) bit 1, (c) -1, i.e., no data, and (d) a random set of 179 bits, using $\alpha = \beta = 0.1$ in Eq. (4). This gives an embedding rate of approximately 54 bits/s for the cover speech used (including unembedded frames). Employing *b1 = b0 = 1.1* in Eq. (5), all the bits were retrieved correctly from the embedded frames that were quantized to 16 bits. No audible difference was detected between the original cover speech and the embedded speech. However, the reconstructed time waveform – the stego – showed a slightly noticeable difference.

A reason for the small difference in the waveform and spectrogram is that the chosen pair of frequencies is in the masked region of only 24 frames with a difference of 6 dB or more lower than the masking threshold and PSD. At other frames, these frequencies may have lower than 6 dB margin, or not at all masked.

As an alternative, the frequencies *f1= 1937.5 Hz* and *f2 = 1062.5 Hz*, which occurred in 95 out of the 208 frames, albeit with only a 3 dB margin were selected for cepstrum modification. The results of embedding 205 values – a random set of 0, 1, -1 – showed no discernible difference in audibility. (Three low energy frames that did not have *f1* and *f0* in the masked region were excluded.) Waveform and/or spectrogram indicated no noticeable difference for random bit streams. If a continuous stream of 1's or 0's was embedded, the increase in the strength of spectrum at *f1* or *f2* resulted in visible difference relative to the original waveform or spectrogram. Due to the low power, however, the difference was not audible. All the embedded data were recovered correctly. Fig. 1 depicts the spectrograms of the host and stego for the case of embedding a random set of 205 values that consisted of 61 with bit 0, 84 with 1 and 63 with -1.

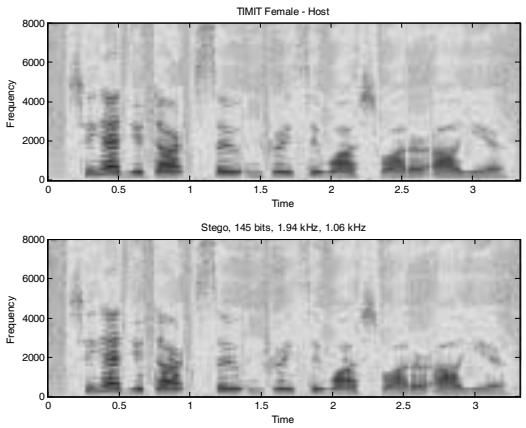

Fig. 1 Host and Stego spectrogram for cepstrum modified at f1= 1937.5 Hz and f2 = 1062.5 Hz

Using a noisy cover speech utterance – from the ATC data base at a sampling rate of 8000 samples/s – similar results were observed for data recovery with no bit error and imperceptibility. Fig. 2 shows the spectrograms of the noisy host and stego with 316 bits embedded at f1= 2500 Hz and f2 = 2625 Hz.

5. Data Robustness

Data retention in the presence of noise after cepstrum modification was studied by adding Gaussian noise at

varying power levels as a ratio of stego frame power. At a signal-to-noise (frame) power ratio of approximately 25 dB, for example, a BER of 5 to 8 out of 205 bits of random data was observed. Higher noise levels proportionally increased the BER. The variability in BER at any given SNR resulted

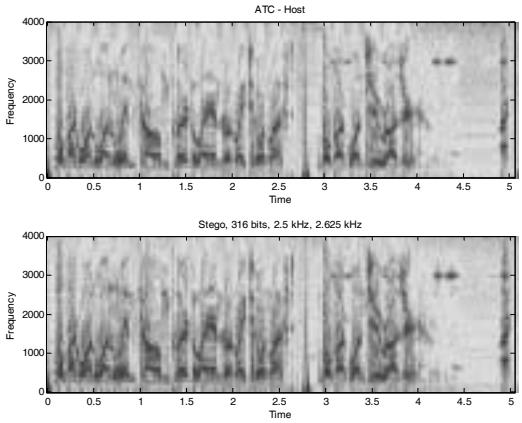

Fig. 2 Host and Stego spectrogram using a noisy host

due to differences in data. This suggests that careful adjustment of the threshold for bit detection may alleviate the problem. Similar BER was observed for the noisy host.

Since the frequencies *f1* and *f2* are chosen in the midband region, bandpass filtering over telephone bandwidth, clearly, has no bearing on BER.

6. Conclusion

A method of hiding information on a cover audio signal by modifying the cepstrum of the signal has been described. Altering the cepstrum at frequencies that are in the perceptually masked regions of most of the frames of the cover speech ensures inaudibility of the resulting stego. Hidden data are retrieved using a spectral threshold at the receiver. A high embedding rate with zero BER has been observed in the experiments conducted on the proposed technique. Depending on the choice of frequencies used for cepstrum modification, embedding may be slightly noticeable in the waveform, spectrogram or in hearing. This effect can be minimized by selecting the frequencies based on their power levels relative to masking thresholds. Altering cepstrum at higher masked frequencies has shown to result in inaudible stego that is also more robust to additive noise.

Noise arising from intentional or unintentional attacks on the stego appears to raise the BER only slightly. With large embedding capacity, however, BER can be minimized using bit duplication or spreading.

References

1. W. Bender, D. Gruhl, N. Morimoto and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, Vol. 35, Nos. 3 & 4, pp. 313-336, 1996.

2. M. D. Swanson, M. Kobayashi, and A.H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, Vol. 86, pp. 1064-1087, June 1998.

3. R.J. Anderson and F.A.P. Petitcolas, "On the limits of steganography," *IEEE Journal of Selected Areas in Communications*, Vol. 16, No. 4, pp. 474-481, May 1998.

4. N. Cvejic, A. Keskinarkaus, and T. Seppanen, "Audio watermarking using m-sequences and temporal masking," *Proc. 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp.227 – 230, Oct. 2001.

5. K. Gopalan, S. Wenndt, A.Noga, D Haddad,and S. Adams, "Covert speech communication via cover speech by tone insertion," *Proc. 2003 IEEE Aerospace Conference*, Vol. 4, pp. 4_1647 -- 4_1653, March 2003.

6. K. Gopalan, "Audio steganography using bit modification," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03)*, Vol. 2, pp. 421-424, April 2003.

7. X. Li and H.H. Yu, "Transparent and robust audio data hiding in cepstrum domain," *Proc. IEEE International Conference on Multimedia and Expo, (ICME 2000)*, New York, NY, 2000.

8. S.K. Lee and Y.S. Ho, "Digital audio watermarking in the cepstrum domain," *IEEE Trans. Consumer Electronics*, Vol. 46, pp. 744-750, August 2000.

9. C.-T. Hsieh and P.-Y. Tsou, "Blind cepstrum domain audio watermarking based on time energy features," *14th International Conference on Digital Signal Processing*, 2002, vol. 2, pp. 705-708, July 2002.

10. K. Gopalan, "Cepstral Domain Modification of Audio Signals for Data Embedding: Preliminary Results," *Proc. of 16th Annual Symposium on Electronic Imaging -- Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, January 2004.