DCT based Audio Steganography in Voiced and Un-voiced Frames

Aniruddha Kanhe
Department of Electronics and Communication Engineering
National Institute of Technology Puducherry Karaikal
kanheaniruddha@gmail.com

G. Aghila
Department of Computer Science and Engineering
National Institute of Technology Puducherry Karaikal
aghilaa@gmail.com

ABSTRACT

In this paper, a robust audio steganography method is proposed based on voiced and un-voiced frames. The key idea is to change the magnitude of Discrete Cosine Transform (DCT) coefficients of voiced and un-voiced frames separately. Taking advantage of voiced and un-voiced characteristics it is possible to embed more number of bits in unvoiced frames. The proposed method proves the high imperceptibility of 43.9dB measured in terms of signal to noise ratio, for a payload of 1.08Kbps. This paper provides the tradeoff between SNR and Embedding Bit Rate which facilitate to decide the imperceptibility requirement. The experimental results proves that the method has high capacity of 240 bps to 1800 bps and provides robustness against the common signal processing attacks such as re-quantization, re-sampling and additive white Gaussian noise

Keywords

DCT, zero crossing count, Short time energy, steganography.

1. INTRODUCTION

Steganography provides an efficient solution to the field of covert communication. It exploits the nature of digital multimedia such as audio, video and image as cover to hide the secret message. In audio steganography the limitations of Human Auditory System (HAS) is utilized effectively. Several audio steganography methods have been proposed such as low bit encoding, echo hiding, Fast Fourier Transform (FFT) based methods etc. The simple low bit encoding method provides a high payload capacity of 44.1kbps [8]. Such a method, which employ direct replacement of Least Significant Bit (LSB) of each sample by the secret message bit is high vulnerable to attacks, hence to increase its robustness various methods have been proposed by changing the embedding position from 1st LSB level 4th and 6th LSB levels [3,4]. Further, there exist the cryptographic approach also, in [11] the secret message is first encrypted using the Advanced Encryption Standards before embedding in cover audio. In [2], the embedding position are decided by genetic algorithm and the secret message is encrypted using RSA algorithm. All such approaches are the time domain approaches where the modification in signal is done in time domain. The encryption off-course increases the robustness of algorithm but the payload is compromised. This tradeoff is addressed effectively in frequency domain techniques. In [14], the high frequency and low frequency components are separated. and the secret message is embedded in the Discrete Wavelet Transform (DWT) coefficients of these components. In [7], each bit is embedded in a frame by inserting two tones. Whereas in [6], the DWT coefficient of audio signal is modified to the nearest Fibonacci number. These method provides the reasonable payload.

In this paper, we propose a robust and imperceptible audio steganography algorithm which embeds the data bits in audio with high payload and also maintains low distortion level in cover audio. Our method consists of separating the voiced and un-voiced part of the speech by using zero crossing count (ZCC) and short time energy (STE). Then computing the DCT of these parts and embedding in the DCT coefficients by modifying them. The proposed method is implemented on NOIZEUS [9,10,12] audio data base and tested against the common signal processing attacks which provides very encouraging results.

This paper is organized as follows: the background and related works are presented in Section 1. The separation of voiced and un-voiced frames is discussed in Section 2. The Section 3 consist of proposed algorithm followed by results and conclusion in Section 4.

2. VOICED AND UN-VOICED SEPARATION

The speech consist of voiced, un-voiced and silence parts [13]. In speech signal the voiced part consist of low frequency and high amplitude and the un-voiced part consist of high frequency and low amplitude. Whereas the silence part does not have such signal characteristics and usually found at the start and end of the speech. To separate these voiced and unvoiced parts, the speech signal is first divided into small frames of 10ms duration. Then the ZCC and STE are computed and using a decision boundary the frames are marked as voiced and un-voiced as shown in Fig.1.

2.1 ZCC

The ZCC calculates that how many times the signal is crossing the zero axes in time domain. This basically measures the frequency in other way. Since the voiced parts consist of low frequency component hence have a high ZCC value and similarly the un-voiced parts consist of high frequency component hence low ZCC. In proposed method the ZCC is calculated by Equ. (1).

$$D(k) = \frac{1}{N}\sum\left|s[n] - s[n+k]\right| \tag{1}$$

Where, N is the total number of samples and $s[n]$ is signal.

2.2 STE

The voiced part of the speech consists of high amplitude signal and un-voiced part have low amplitude signal. This characteristic is utilized for separating them by calculating the energy. The

ICIA '16, August 25-26, 2016, Pondicherry, India
© 2016 ACM. ISBN 978-1-4503-4756-3/16/08$15.00
DOI: http://dx.doi.org/10.1145/2980258.2980360

speech signal is divided in small frames of *N* samples and the total squared of each sample is computed to get energy. The speech is divided into small frames by multiplying with a suitable window function. In proposed algorithm the Hamming function is used, as widely used in speech processing application [13]. The hamming window $w(n)$ is given by:

$$w(n)=\begin{cases}0.54-0.46*\cos\dfrac{2\pi\,\mathrm{n}}{L-1} & \text{for } 0\le n\le L-1\\ 0 & \text{otherwise}\end{cases} \qquad (2)$$

And, if $s(n)$ denotes the signal, then the short time energy is given by Equ. (3):

$$E_m=\sum s(n)w(m-n)^2 \qquad (3)$$

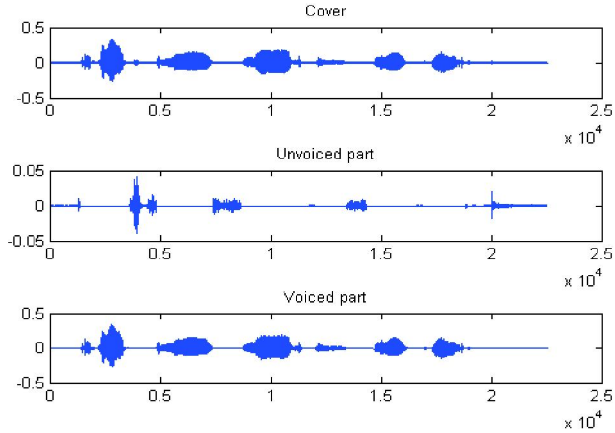

Figure 1. Voiced and un-voiced parts of the speech

These separated frames are utilized for embedding the secret message as explained next section.

3. PROPOSED ALGORITHM

In this section the embedding and extraction algorithm is discussed.

3.1 Proposed Embedding algorithm

The proposed hiding algorithm embeds variable payload in voiced and un-voiced frames. The following procedure is used for embedding in cover audio:

- Cover speech is divided into non-overlapping frames of 10ms duration, and ZCC & STE is computed for all the frames.
- To segregate the frames the ZCC and STE are compared such that if ZCC is small and STE is high then the frames is voiced frame $s_v(n)$ otherwise it is un-voiced frame $s_{uv}(n)$.
- Next the DCT of these frames are computed using :

  $$X(k)=w(k)\sum_{n=0}^{N-1}x(n)\cos\frac{(2n+1)*k\pi}{2N} \qquad \text{k=0,1...N} \quad (4)$$

  where

  $$w(k)=\begin{cases}\dfrac{1}{\sqrt{2}} & \text{if k=0}\\ 1 & \text{otherwise}\end{cases}$$

  And $x(n)$ is either $s_v(n)$ or $s_{uv}(n)$.
- The secret message is converted in to binary.
- The last *m* coefficients of $S_v(k)$ are replaced by message bit and *2m* coefficients of $S_{uv}(k)$ are replaced by the message bit, where $S_v(k)=DCT[s_v(n)]$ and $S_{uv}(k)=DCT[s_{uv}(n)]$.
- Further, if embedding bit is zero the DCT coefficient is replaced by 0 otherwise replaced by ε a non-zero value.
- The inverse DCT is performed and the modified voiced and un-voiced frames are arranged accordingly to obtain the stego audio.

Proposed Embedding Algorithm

```
Get $s_v(n)$ and $s_{uv}(n)$.
if message bit is not embedded into the frame then
      $S_v(k)=DCT[s_v(n)]$
      $S_{uv}(k)=DCT[s_{uv}(n)]$
            if embedding bit is '0'
             DCT coefficient =0
            else
            DCT coefficient= $\varepsilon$
      end
      perform IDCT
else
go to the next frame
end
combine $s_v(n)$ and $s_{uv}(n)$.
```

The threshold values of ZCC and STE for separating the voiced and un-voiced frames are computed empirically based on the result obtained after the simulation.

3.2 Extraction Procedure

The stego audio is divided into 10ms frames then ZCC and STE are computed to separate the voiced $s'_v(n)$ and un-voiced frames $s'_{uv}(n)$ as discussed in section 2. Next the DCT operation is performed on these frames and the last *m* and *2m* coefficients are checked for $S'_v(k)$ and $S'_{uv}(k)$ respectively, where $S'_v(k)=DCT[s'_v(n)]$ and $S'_{uv}(k)=DCT[s'_{uv}(n)]$. If the coefficient is less than ε, the message bit is 0 else message bit is 1. The algorithm for extraction is shown below:

Proposed Embedding Algorithm

```
Get $s'_v(n)$ and $s'_{uv}(n)$.
if message bit is not extracted from the frame then
      $S'_v(k)=DCT[s'_v(n)]$
      $S'_{uv}(k)=DCT[s'_{uv}(n)]$
      Compare the last coefficients
            if DCT coefficient < $\varepsilon$
             message bit=0
            else
            message bit=1
      end
      perform IDCT
else
go to the next frame
end
combine all the message bits to get secret message signal
```

4. RESULTS

The proposed algorithm is implemented in Matlab R2010a and tested on the NOIZEUS data base. The database consist of 30 speech files recorded by 15 male and 15 female speaker in a sound proof room using Tucker Davis instrument. The database includes all the phonemes in the American English Language from IEEE sentence database. The imperceptibility test is carried out by calculating the signal to noise ratio (SNR) and by listening test. The SNR is computed for all the 30 speech files using the Equ. (5):

$$SNR(dB) = 10\log\left[\frac{\sum |s_c(m)|^2}{\sum |s_c(m) - s_s(m)|^2}\right] \qquad (5)$$

where $s_c(m)$ is the cover audio and $s_s(m)$ is the stego audio signal [5]. The SNR values for 30 speech file obtained after performing the performing the proposed algorithm is shown in Fig.2.

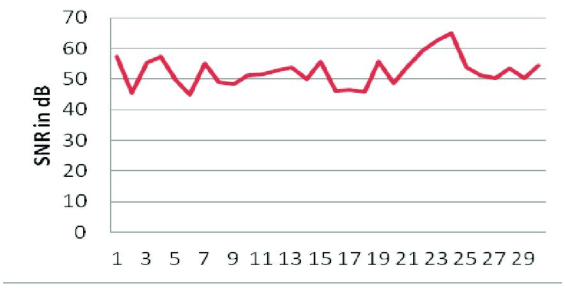

Figure 2. SNR result for proposed scheme

The highest SNR achieved is 64.86dB and the lowest SNR value is 44.78dB. The listening test is performed over 10 different person and have asked them to grade the quality of audio in a scale of 5 starting from 1 for Very annoying, 2 for Annoying, 3 for Slightly annoying, 4 for Good, and 5 for Excellent or highly imperceptible. The average grade obtained after the test, is shown in Table 1. The listening test proves that the proposed steganography algorithm is highly imperceptible to the human ears.

Table 1. Listening test result

	Grade
Proposed Algorithm	4.87

The robustness of the algorithm is tested by computing the Bit Error Rate (BER) with and without attacks. The BER is the ratio of incorrect bits retrieved by the total number of bits embedded in signal. The proposed algorithm gives the 0 BER in without attack environment. Which shows that the secret message is correctly recovered form the stego audio with 100% accuracy. The testing of algorithm against the common signal processing attacks such as re-quantization, re-sampling and AWGN is done and the BER is calculated, In re-sampling attack the stego audio is sampled at a frequency other than the original sampling frequency to corrupt the secret message. In this case the original sampling frequency of the is 16KHz and the signal is sampled at 8KHz and re-sampled back to 16KHz. Similarly in re-quantization attack the stego signal is quantized with different level other than the original and to corrupt the secret message signal. In AWGN attack a Gaussian Noise is added to stego signal to corrupt the signal.

The Table 2 summarizes the BER result for the proposed algorithm with and without attack. The proposed algorithm is tested against two levels of quantization i.e. 8 bits and 24 bits where the original quantization level is 16 bits. In re-sampling attack the stego signal is down sampled to 8 KHz and then up-sampled to original sampling frequency of 16 KHz. In AWGN attack a white noise of 15dB and 25dB is added to the stego signal. The BER values shown in the table is the average of all the BER obtained by 30 audio files. It is evident from the result that the proposed method can withstand with the signal processing attack and the information can be retrieved with high accuracy.

Table 2. BER result

Sl.No.	Types of attack	BER (average)
1.	Without attack	0.0
2.	Re-quantization (8bits)	0.0
3.	Re-quantization (16bits)	0.0
4.	Re-sampling (8 Khz)	0.51
5.	AWGN (15dB)	0.48
6.	AWGN (25dB)	0.52

The trade-off between embedding rate and SNR is shown in the Fig 3. The optimized value of SNR is obtained at a payload of 720bps with a SNR value of 51.47dB.

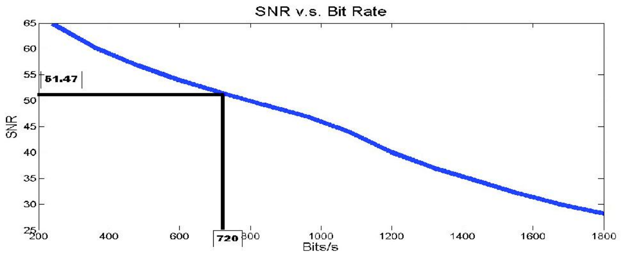

Figure 3. SNR versus payload for proposed scheme

The Table 3. shows the comparison of proposed method with sparse based steganography method presented in [1]. The SNR value and payload capacity of the proposed method is high.

Steganography Method	SNR	Payload
Sparse based [1]	36.01dB	0.4kbps
Proposed Method	51.47dB	0.7kbps

The time domain comparison and spectrogram comparison is shown in Fig 4 and Fig 5. These comparison reflects that the change is insignificant and hence the proposed algorithm is highly imperceptible.

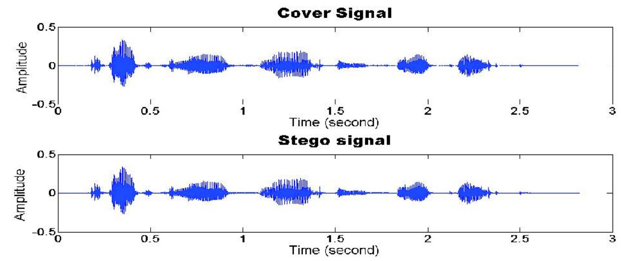

Figure 4. Time domain plot of cover signal and stego signal

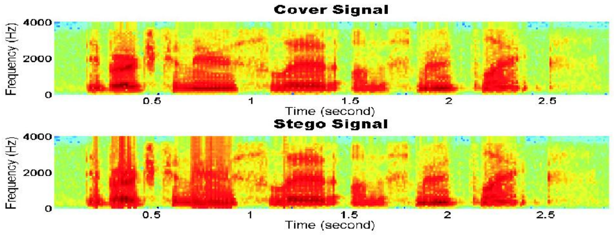

Figure 5. Spectrogram of cover and stego signal

5. CONCLUSION

In this paper we presented a new and robust method of hiding the data bits in cover audio. Our method focuses on the utilizing voiced and un-voiced speech characteristics for data hiding using DCT coefficients. The separation is done by using ZCC and STE values. This part of the algorithm is very crucial since a threshold has to be decided to segregate the frames. Hence increases the robustness of the algorithm. The algorithm is implemented on NOIZEUS database having 30 speech files. The proposed algorithm gives the high SNR value with the optimized payload. The listening test performed also proves the imperceptibility of the algorithm. The robustness of the algorithm is tested by calculating BER for with and without attack. The BER results justifies that the robustness of the algorithm. In future, the data loss will studied and robustness will be analyzed against different attacks.

6. ACKNOWLEDGMENTS

Our thanks to Rohit Anand who helped in debugging the voiced and unvoiced separation part of the algorithm.

7. REFERENCES

[1] Soodeh Ahani, Shahrokh Ghaemmaghami, and Z Jane Wang. 2014. A Sparse Representation based Wavelet Domain Speech Steganography Method. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 9290.

[2] Krishna Bhowal, Debnath Bhattacharyya, Anindya Jyoti Pal, and Tai Hoon Kim. 2013. A GA based audio steganography with enhanced security. *Telecommunication Systems* 52, 4: 2197–2204.

[3] Nedeljko Cvejic and Tapio Seppanen. 2004. Increasing robustness of LSB audio steganography using a novel embedding method. *International Conference on Information Technology: Coding Computing, ITCC*, 533–537.

[4] Nedeljko Cvejic. 2005. Increasing Robustness of LSB Audio Steganography by Reduced Distortion LSB Coding. *Journal of Universal Computer Science* 11, 1: 56–65.

[5] Fatiha Djebbar, Habib Hamamy, Karim Abed-Meraimz, and Driss Guerchix. 2010. Controlled distortion for high capacity data-in-speech spectrum steganography. *Proceedings - 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2010*: 212–215.

[6] Mehdi Fallahpour and David Megías. 2015. Audio watermarking based on Fibonacci numbers. *IEEE Transactions on Audio, Speech and Language Processing* 23, 8: 1273–1282.

[7] Kaliappan Gopalan and Stanley Wenndt. 2004. Audio steganography for covert data transmission by imperceptible tone insertion. *Proc. of the 4th IASTED, Wireless and optical communications*: 262–266.

[8] Kaliappan Gopalan. 2003. Audio steganography using bit modification. *Proceedings - IEEE International Conference on Multimedia and Expo* 1: I629–I632.

[9] Yi Hu and Philipos C Loizou. 2008. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on* 16, 1: 229–238.

[10] Yi Hu and Philipos C. Loizou. 2007. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication* 49, 7-8: 588–601.

[11] Aniruddha Kanhe, G Aghila, Ch Yaswanth Sai Kiran, Ch Hanuma Ramesh, Gabbar Jadav, and M Gowtham Raj. 2015. Robust Audio steganography based on Advanced Encryption standards in temporal domain. *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, 1449–1453.

[12] Loizou. Ma, J., Y, Hu., P. 2009. Objective measures for predicting speech intelligibility in noisy conditions vased on new band importance functions. *Journal of the Acoustical Society of America* 125, 5: 3387–3405.

[13] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Printice-hall, Inc., Upper Saddle River, NJ, USA.

[14] Siwar Rekik, Driss Guerchi, Sid-ahmed Selouani, and Habib Hamam. 2012. Speech steganography using wavelet and Fourier transforms. *EURASIP Journal on Audio, Speech, and Music Processing 2012*, 20.