A Real-time Implementation for the Speech Steganography using Short-Time Fourier Transform for Secured Mobile Communication

Rajeev Shrivastava¹, Mangal Singh*, Rakhi Thakur³, Kalluri Saidatta Subrahmanya Ravi Teja⁴

¹Associate Professor, Department of Electronics and Communication Engineering, Bharat Institute of Engineering and Technology, Hyderabad, India.
²Associate Professor, Department of Electronics and Communication Engineering, Institute of Technology, Nirma University, Ahmedabad, India.
³Email id: mangal.etce@gmail.com
⁴Lecturer, Department of Electronics and Telecommunication Engineering, Kalaniketan Polytechnic College, Jabalpur, India.
⁴Research Scholar, Department of Electronics and Communication Engineering, Bharat Institute of Engineering and Technology, Hyderabad, India.

Abstract. Steganography can be described as an approach of masking an undisclosed message with a normal message which is known as the Carrier message signal. DSP techniques, such as LSB encoding, have historically been implemented for secret information hiding. Utilization of steganography functions of deep neural networks for voice data is something this paper will present. This paper also demonstrate that the steganography techniques suggested for vision are less suitable for speech signals this paper present a implementation technique that involves the use of ISTFT and STFT as differentiable layers in the network. Empirically, the efficacy of the proposed methods based on multiple datasets of speech should be demonstrated and the outcome are examined quantitatively and qualitatively. Using of multiple decoders or a single conditional decoder helps to hide multiple signals in a single carrier signal. Finally, under various channel distortion situations, this model Qualitative studies indicate that human listeners cannot detect changes made to the carrier and hence the decoded messages are highly intelligible.

Index Terms—Speech Steganography, Digital Signal Process, Fourier Transform, STFT

1. Introduction
Steganography can be described as an approach of masking an undisclosed message with normal public message which is known as the Carrier data. It is described as digital steganography when we implement these techniques over the digital data signals [1,2]. Cover objects use for steganography may vary, for example steganography algorithms for image based data are widely developed, however the steganography algorithms or steganography techniques for audio signals are relatively very less in number.

Implementation of steganography to the audio archive is not as simple when compared with implementing steganography to image archive, in general the raw sound data files are larger when
compared to the raw image data, the raw image file with 1280x800 resolution, has approximately 3MB data for a 24 bit colour image file has. While the raw audio files with:
1. Sampling frequency 44.1 KHZ,
2. 16 bit stereo channels
3. 4-minute duration has a size of 40 MB approximately.

The variation between both the image and audio is very large, hence the use of steganography techniques in audio data becomes difficult when compared to image signals. When Discrete Fourier Transform is used for data domain conversion, then the audio data files require more cost as number of samples that require transformation are much more when compared to image signals. When we implement the LSB Method the noise generated at the sound file is very high. 24 bits are used to encode the Pixel, however only 15 bits are required for encoding the sound signal, the usage of the raw audio files is low when compared to the image files because of large size, hence we should implement such a technique need that allow us to save the secret messages, even after compressing the audio data. Modern steganography technique includes the techniques for hiding data into digital files in network level in the packets that get transmitted [3,4]. The following elements are required to hide the information into media files.

- C - Cover media or a masking signal to hide the secret data
- M - Hidden message.
- Fe – Stego function and inverse stego function Fe⁻¹ which are used to hide and unhide the information.
- K – an optional stego key to hide and unhide the message

Stegofunction operates on the cover media or the masking signal and the message to be hidden along with the stegokey to produce the stegomedia S.

The outline of a stenographic operation is shown below.

![Fig. 1. The Stenographic Process](image)

Steganography with cryptography is considered as a perfect combination for information security. Modern techniques of steganography

The modern techniques in steganography include exploitation of media file to convey the message. Media that are widely used for embedding message digitally.

- Plain text
- Images
- Audio and Video
- IP datagram

Proposal for usage of deep neural networks as a stenographic function[1,2] to hide the image inside other image, unlike the traditional steganography techniques [3,4], in the proposed technique network try to retrieve the secret data hidden in the cover signal with no manual specification of the specific check of redundancy, the above outcomes present acceptable outcomes on the image data, however the implementation of such models on speech data is not highly developed due limitations like size of the data and noise in the audio data.

In comparison to dealing with raw image files in the vision processing in time domain, the most commonly implemented approach when understanding speech data in frequency domain is to use STFT and then monitor the spectral changes over period of time. The output of the STFT is the complex matrix of the FT at distinct time intervals, the most commonly used approach is to use absolute value or magnitude value of Short Time Fourier Transform measurement, and maintain a acceptable
overlapping in between adjacent intervals of time[5]. This approach also further complicates the restoration of signal in time domain. This paper proves that steganography techniques implemented for vision are less suitable when implemented with voice data. We propose a new method which include implementing STFT and ISTFT as differentiable layers in the network[6,7].

The process of hiding the written text in the audio file is easy when compared to the hiding the audio signal inside the audio file, and this requires many other additional features, for example the hidden information may reveal the identity of the speaker, speaker sentimental state, etc. These features later used for identification and authentication of message. Similarly, [8,9] the proposed model contain three verticals, first vertical is to learn how to encode the hidden data or data in the carrier signal, next vertical is DSTFT and ISTFT layers that simulate transformations between time domain and frequency domain. Third vertical is to learn decoding the hidden message from a generated carrier [10,11], in addition we demonstrated how the above method allow us to use single carrier signal to hide multiple secret messages, each with a intended recipient who is authorised to retrieve the message [12,13]. Further analysis proves including of layers in STFT produce a highly sophisticated method which is highly efficient for various distortions and compressions techniques, which include MP3 encoding, AWGN,. Qualitative analysis suggests that modification to carrier is unpredictable by human and the messages decoded preserve other important semantic content like speaker identity [14,15].

2. Methodology

A. **STFT**

Only a small variation exists in STFT and Fourier Transform. Sampling is done on the signal to divide the signal into small enough portions, these signal portions are assumed to be stationary; “W” is the function of window that is chosen in such a way the window function width is equal to the signal partition where signal is stationary[16,17]. At t=0, i.e., at the initial position of the signal “W” a window function is located, let us assume that at the t=0 the window function width be “T” s, with initial T/2 seconds the W will overlap [18]. Then the signal is multiplied with window function, only initial T/2 seconds of the signal is chosen, with appropriate window weighting, then the resultant product is just other signal, whose Fourier transform is to be considered [19,20]. The transformation result is the FT of the initial T/2 seconds of the signal., then there will be no problem if this portion of the signal is constant and the result generated is true frequency representation of the initial T/2 seconds of the signal. The next step would be scaling this window function to a new position then multiply this with the signal and take FT of the product. By performing the shifting operation on the window with “t1” second interval this is repeated until the end of the signal.

\[
STFT^{(w)}(t', f) = \int [x(t) * w(t - t')] e^{-j2\pi ft} dt
\]  

(4)

We can understand from the previous equation, that the term x(t) is the representation of the signal, the term w(t) represent the window function which is used, and finally * is the complex conjugate. 

For every t’ and f, the signal’s FT is multiplied with the signal’s window function STFT, as shown in the previous equation. New STFT coefficient is computed better:

\[w^*(t-t')\]  
\[w^*(t-t')\]  
\[w^*(t-t')\]  

Fig.2. Spectra of Short Time Fourier transform
Windowing functions are the Gaussian like function as shown in the Fig 2 with different colours as:

- The red one represents the window at \( t = t_1 \) location
- Blue colour represents the window at \( t = t_2 \) location
- Green one shows the window at \( t = t_3 \) location,

These are three different Fourier Transforms at three different time intervals, so a true time frequency representation (TFR) of the signal is obtained. The transform will be two dimensional. Since our transform is a function of both frequency and time consider a non-stationary signal, shown in fig 3.

![Fig. 3. Non stationary signal](image)

The above signal comprises four components of frequency at different intervals of time, The interval between 0 and 250 ms is a basic sinusoidal signal of 300 Hz, the other 250 ms intervals are sinusoid signals of 200 Hz, 100 Hz, and 50 Hz, respectively. Let's take a look at this non-stationary signal, STFT. [21, 22].

![Fig. 4. STFT Non stationary signal](image)

This is a plot of two dimensions. The 'x' axes — as planned — reflect time. The axes on the 'y' axis reflect frequency. Note that with respect to centre of the 'y' axis which is frequency axis the graph is symmetric. Though it was not shown Symmetric form always represent the Fourier Transform of a real signal as it is known that windowing FT will generate STFT, so we can conclude that STFT is also frequency symmetric. Negative frequencies represent the symmetric part, which is difficult to comprehend, fortunately it can be avoided as it is not important, hence, we can conclude that both Short Time Fourier Transform and Fourier Transform are symmetric. We can observe that to four different frequency components there are four peaks corresponding, each peak is separated by a time interval. Unlike FT, thereal signal had four components of spectrum situated at various time intervals. You may wonder why we need the wavelet transformation, TFR of the signal is given by STFT, the STFT's implicit problem is not apparent. In the case above. To explain the concept [23, 24], It's impossible to say what spectral components are present at any given time. In a nutshell, the signal's precise time-frequency representation is unknown, i.e. we can only obtain the intervals of time during which a specific band of frequencies exists, which is a issue of resolution. [25, 26].
the window function width is some way related to STFT function. The window width element is selected as the window's support. It is compactly supported when the narrow window feature is present, this often used in wavelet world.

In the frequency domain, we remember that there is no resolution problem in FT, That is to say, we have precise knowledge of the frequencies that exist. In the same way, in time domain there is no time resolution problem as we have at every instant we have information about the signal's value. In contrast, we can deduce that the Fourier Transform’s time resolution and frequency resolutions in the time domain are both zero because no information related to them is available. The window used in the FT is its kernel this is the fact that gives the perfect frequency resolution in FT , the exp(jwt) range is from + infinity to - infinity, there is a function that lasts at all times. ShortTimeFourierTransform's window function is finite in duration and only covers a part of the signal, reducing frequency domain resolution. That is, we no longer have the knowledge of exact frequencies of the signal's frequency components, but we do know the frequency band in which they reside.

Because the kernel function in FT has an infinite duration window, it aids in obtaining an accurate frequency resolution. Because the STFT window is finite in size, Perfect frequency resolution is no longer possible. Why don't we make the STFT window length infinite to achieve perfect frequency resolution? just like the FT? In that case, we may All time information is lost, and the FT is used instead of the STFT. To put it another way, we're faced with the following conundrum: We get the FT when we use an infinite length window, which provides complete frequency resolution with out information about time.. We will need a window which is almost short enough in which the signal is constant to achieve stationarity. [27,28]. The smaller the window, better the resolution of time and presumption of stationarity; hence, the frequency resolution suffers. [29] when comparing a narrow and a wide window The time resolution of a narrow window is excellent, but the frequency resolution is poor, and the frequency resolution of a wide window is good but the time resolution is poor.

3. Simulation Results

We used MATLAB to run all of the simulations, and we used a variety of audio samples to test both current and proposed methods with different cr values. Figure 5 shows audio steganography based on STFT, in which the stegoaudio differs from the original audio, indicating that the unauthorised entity does not have a choice. Figure 5 shows the difference between original and stegoaudio, which results in a highly secure system. We can see that both original and stegoaudio appear to be identical, which is achieved by our proposed STFT approach. and retrieved signal in figure 6.

![Fig.5. STWT Steganography signal](image)
4. Conclusion

It has been concluded that the suggested speech steganography produced the superior results. The proposed method proved to be resistant to audio manipulation and very secure, with very little noise as a result. The most important feature of this technique is that it reduces the number of computations needed and does not require the use of complex equations, making it simple and straightforward to use in real-time situations.

References

[1] Shumeet Baluja, “Hiding images in plain sight: Deep steganography,” in Advances in Neural Information Processing Systems, 2017, 2069–2079.
[2] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei, “Hidden: Hiding data with deep networks,” in European Conference on Computer Vision. Springer, 2018. 682–697.
[3] Tayana Morkel, Jan HP Eloff, and Martin S Olivier, “An overview of image steganography.,” in ISSA, 2005, pp. 1–11.
[4] GC Kessler, “An overview of steganography for the computer forensics examiner. retrieved february 26, 2006,” 2004.
[5] Jae Soo Lim and Alan V Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604, 1979.
[6] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi, “Sft phase retrieval: Uniqueness guarantees and recovery algorithms,” IEEE Journal of selected topics in signal processing, vol. 10, no. 4, pp. 770–781, 2016.
[7] E Hofstetter, “Construction of time-limited functions with specified autocorrelation functions,” IEEE Transactions on Information Theory, vol. 10, no. 2, pp. 119–126, 1964.
[8] Daniel Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.
[9] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallet, “Darpatimit acoustic-phonetic communications speech corpus cd-rom. nist speech disc 1-1.1,” NASA STI/Recon technical report n, vol. 93, 1993.
[10] Joseph P Campbell, “Testing with the yohocd-rom voice verification corpus,” in ICASSP, 1995.
[11] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, “Language modeling with gated convolutional networks,” in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 933–941.
[12] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, “End-to-end text-dependent speaker verification,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5115–5119.

[13] P Jayaram, HR Ranganatha, and HS Anupama, “Information hiding using audio steganography—a survey,” The International Journal of Multimedia & Its Applications (IJMA) Vol, vol. 3, pp. 86–96, 2011.

[14] Xiaoxiao Dong, Mark F Bocko, and Zeljko Ignjatovic, “Data hiding via phase manipulation of audio signals,” in Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP’04). IEEE International Conference on. IEEE, 2004, vol. 5, pp. V–377.

[15] Walter Bender, Daniel Gruhl, and Norishige Morimoto, “Method and apparatus for echo data hiding in audio signals,” Apr. 6 1999, US Patent 5,893,067.

[16] Jamie Hayes and George Danezis, “Generating steganographic images via adversarial training.” in Advances in Neural Information Processing Systems, 2017, pp. 1954–1963.

[17] Yinlong Qian, Jing Dong, Wei Wang, and Tieniu Tan, “Deep learning for steganalysis via convolutional neural networks,” in Media Watermarking, Security, and Forensics 2015. International Society for Optics and Photonics, 2015, vol. 9409, p. 94090J.

[18] Lionel Pibré, Jérôme Pasquet, Dino Ienco, and Marc Chaumont, “Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover source mismatch,” Electronic Imaging, vol. 2016, no. 8, pp. 1–11, 2016.

[19] Pin Wu, Yang Yang, and Xiaoqiang Li, “Stegnet: Mega image steganography capacity with deep convolutional network,” Future Internet, vol. 10, no. 6, pp. 54, 2018.

[20] Weixuan Tang, Shunquan Tan, Bin Li, and Jiwu Huang, “Automatic steganographic distortion learning using a generative adversarial network,” IEEE Signal Processing Letters, vol. 24, no. 10, pp. 1547–1551, 2017.

[21] Nameer N El-emam, “Embedding a large amount of information using high secure neural based steganography algorithm,” International Journal of Information and Communication Engineering, vol. 4, no. 2, pp. 2, 2008.

[22] Haichao Shi, Jing Dong, Wei Wang, Yinlong Qian, and Xiaoyu Zhang, “Ssgan: secure steganography based on generative adversarial networks,” in Pacific Rim Conference on Multimedia. Springer, 2017, pp. 534–544.

[23] Wenhui Cui, Shaoxiong Liu, Feng Jiang, Yongliang Liu, and Debin Zhao, “Multi-stage residual hiding for image-into-audio steganography,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 2832–2836.

[24] Shivam Agarwal and Siddarth Venkatraman, “Deep residual neural networks for image in speech steganography,” arXiv preprint arXiv:2003.13217, 2020.

[25] Ru Zhang, Shiqi Dong, and Jianyi Liu, “Invisible steganography via generative adversarial networks,” Multimedia tools and applications, vol. 78, no. 7, pp. 8559–8575, 2019.

[26] Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne, “A digital watermark,” in Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference. IEEE, 1994, vol. 2, pp. 86–90.

[27] Raymond B Wolfgang and Edward J Delp, “A watermark for digital images,” in ICIP (3), 1996, pp. 219–222.

[28] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin’ichi Satoh, “Embedding watermarks into deep neural networks,” in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. ACM, 2017, pp. 269–277.

[29] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet, “Turning your weakness into a strength: Watermarking deep neural networks by backdooring,” USENIX, 2018.