IMAGE HIDING ON HIGH FREQUENCY SPEECH COMPONENTS USING WAVELET PACKET TRANSFORM

[1] Asst. Prof. Dr. Mohammed Nasser Hussein Al-Turfi
[1]*Building and Construction Engineering Department, University of Technology, Baghdad, Iraq*
alaa.qassim1967@gmail.com

Abstract - This paper propose a method for security threw hiding the image inside the speech signal by replacing the high frequency components of the speech signal with the data of the image where the high frequency speech components are separated and analyzed using the Wavelet Packet Transform (WPT) where the new signal will be remixed to create a new speech signal with an embedded image.

The algorithm is implemented on MATLAB 15 and is designed to achieve best image hiding where the reconstruction rate was more than 94% while trying to maintain the same size of the speech signal to overcome the need for a powerful channel to handle the task. Best results were achieved with higher speech resolution (higher number of bits per sample) and longer periods (higher number of samples in the media file).

Keywords - Image Hiding, High Frequency Speech Components (HFSC), Wavelet Packet Transform (WPT).

I. INTRODUCTION

Hiding data is the most important issue that people who work in security must keep in mind, must keep an eye to it because now a day's all type of wars are changing from the traditional war with gun and powder to wars of water, economic and information which may destroy without a noticeable feature or a touchable action.

An image may stand for millions of words so we must hide these words in a safe place where no one can see it but the authorized. As much as we can hide this image inside an un-expectable place the more become safer, one of these places is the media files where sound files and specially the high frequency components of such files that contains the minimum amount of intelligent information are where we are going to hide the image inside it [1].

In the ends of the 20th & the earlier of the 21st century the Wavelet Transform was worldwide tool used in nearly all engineering and scientific researches because of some special properties that any other frequency transform don't provide some of these properties are the easiness of coefficients calculations and multi-resolution analysis for the frequency components' at specific times where a three dimensional frequency spectrum can be provided, orthogonality, & orthonormaityl .

Its well known that wavelet transform decompose the signal in the first level into high (H) and low (L) frequency components by using bank of band pass filters and a down sampler. Same process is repeated at the second stage where the (L) part will produce low high (LH) and low low (LL) frequency components while the (H) part produces high high (HH) and high low (HL) frequency components and so on in the decomposition process to the last stage where if the total number of samples is 2N then the number of decomposition stages (levels) is N[2].

The most usable or the intelligent part of these components is the (LL) branch "if we have two decomposition levels only" where all intelligent information are inside this part while the (HH) part doesn't since the majority of the speech spectrum lies in the low frequency region.

But we are in need for them if we are seeking for 100% perfect reconstruction which is not necessary in most applications or it may be within the acceptable levels of noise[3].

In this paper the (HH) part (for two level of decomposition and HHH if we have three level of decomposition) will be applied on the speech signal and the image as well to hide the image inside the speech signal where the signal will be decomposed and replacing the (HH) part of the speech signal with the image and then the reconstruction process will be applied to inform a new speech signal which is as similar as possible to the original one. At the receiving side the same process will be applied to extract the image by decomposing to obtain the (HH) part and then re-create the image again from this part [2,3].

The similarity between the original speech signal and the re-created one in the sending side exceeds 95% where the reconstruction process for the image in the receiving side was more than 92%. These results are affected depending upon the level of decomposition, the size of the speech file (number of samples), the resolution of the speech signal (number of bits per sample), the type of the image (RGB or B&W), and the resolution of the image (number of bits per pixel)..

II. . IMAGE VERSES SPEECH

Talking about images means talking about different types of compression – decompression techniques, drivers, colors and application forms with various extensions and resolutions depending upon the place of usage whether if it's a standstill (snap shoot) image or a movie (mpeg media application). Hence it must be notified that the image which will

be hidden must meet conditions some of them are necessary to increase the security and the system fidelity but others are sufficient and must be satisfied in order to achieve best results with minimum distortion in the received image or else some changes must be applied on the source image or the speech file before starting with hiding this image inside the speech file[4].

Speech is something different, where two major factors must be taken in consideration for the speech sample that may affect the process, first is the number of bits per samples "which may stand as the same of the brightness level for images", where all speech signals are represented using single dimension vector (not like images may be represented as a two or three dimensional matrix). Second, is the number of samples per seconds "which may stay as the number of pixels in the image for representing the image resolution"[5].

Both speech and images have the same property that the intelligent information are kept inside the low frequency components (part) where the high frequency part contains only the complimentary information which may be some times represents only noise, therefore replacing these information "if its achieved under certain circumstances" may give high results specially if the sound files are implemented using high resolution (16 or 32 bit per sample) with CD Quality recordings (44100 sample per second) and the image with normal resolution (8 or 16 bit per pixel) and in gray scale not color image.

An important property exists in this approach is that both the image and the speech may be operated using the same media player drivers like Windows Media Player by Microsoft, RealPlayer by Apple,…etc and they are became more usable in our daily life specially when they are built in some Multi-Media players and mobile communications like I-Pads and I-Phones by Apple and Android[6].

This paper uses the HFSC that produced from the decomposition process Applying Mallat Multi-Resolution Analysis in the Wavelet Packet Transform (WPT) to the N^{th} level (depending on the properties of both the image and the speech signal) as a decoyed since the amount of information (and hence the accumulative power) is small comparing to other parts, so an exchange process may replace it with the new components of the image keeping in mind the overall exchanged frequency components of both the image and the speech must be normalized in order not to affect the speech signal and hence may be embedded inside a movie or a passing chat without a noticeable effect[7]..

III. IOT SECURITY

The wavelet packet method is a generalization of wavelet decomposition that offers a richer range of possibilities for signal analysis. In wavelet analysis, a signal is split into an approximation (L) (Low Frequency components) and a detail (H) (High Frequency components). The approximation is then itself split into a second-level approximation (LL) and detail (LH), and the process is repeated. For an n-level decomposition, there are n+1 possible ways to decompose or encode the signal as shown in fig(1)[4,6].

In wavelet packet analysis, the details (H) as well as the approximations can be split. This yields more than $2^{\wedge}(2^{(n-1)})$ different ways to encode the signal. This is the wavelet packet decomposition tree. where a strictly binary tree is achieved as shown in fig(2)[3].

The wavelet decomposition tree is a part of this complete binary tree. For instance, wavelet packet analysis allows the signal S to be represented as A1 + AAD3 + DAD3 + DD2. This is an example of a representation that is not possible with ordinary wavelet analysis. Therefore in this paper the speech components that exist in the DDD3 part (if we are using only three levels of decomposition) will be replaced with the pixels

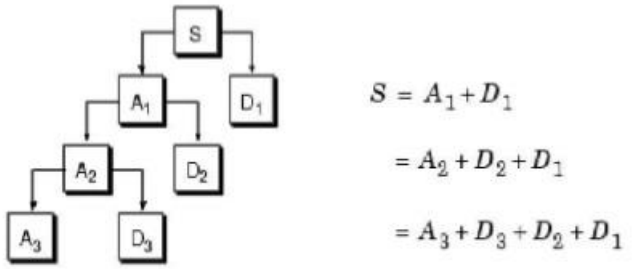

Fig. 1: Wavelet Transform Decomposition tree

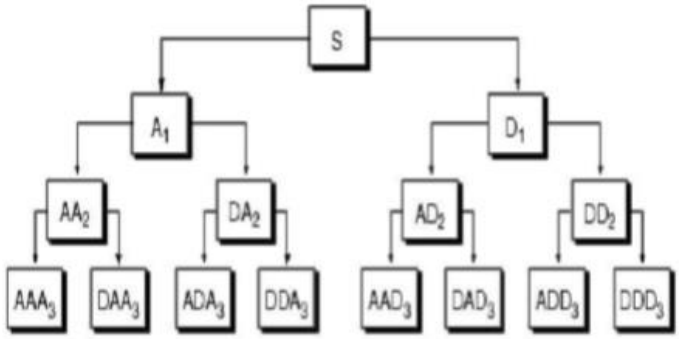

Fig.2: Wavelet Packet Transform Decomposition tree

of the image that the system will hide in it keeping in mind that the overall energy In the part DDD3 must be equal or as near as possible to the energy in the image in-order to reduce distortion to the minimum while the correlation factor will rise to its maximum achievable value and the number of speech samples in this part must be as near as possible to the number of pixels in the image so that we can have as near as possible perfect reconstruction in the receiving side[4,6].

IV. THE PROPOSED ALGORITHM

The algorithm contains two phases where both of these phases may be shown as in figure (3) which may considered as one of the strength points in this algorithm since the construction and reconstruction architecture are as same as

possible which reduces the system complexity and rises the system performing rate and ease system maintenance.

The Construction phase which is applied to hide the image inside the high frequencies of the speech signal while the Re-Construction phase is applied to extract the image from the speech signal in the receiving side.

In this paper the algorithm is applied for a digital speech signal sampled at 11Ksample/sec with a resolution of 8bit/sample mono type with duration of 2 seconds where a color image of size 400*300 with 24 bit/pixel of brightness will be hidden inside the speech signal as the secret image.

It must be noticed that a speech signal with higher sampling frequency and resolution may hide a larger image with higher levels of brightness, and here it will be obvious that the algorithm will be in need for a longer speech or an assay if the process going to hide more than one image or the system going to use more than one speech to hide a fragment of an image so the system is in need for a powerful channel to transfer data.

Before starting the process, the system must check the sound file whether it can fit the image or not, since the system takes each 8 sound samples and decompose it too the 3rd level so for each 8speech samples there will be one image pixel, assuming the number of bits per speech sample is equal to the number of bits per image pixel. This means "assuming the above operating conditions" that the sound file must have samples 8 times larger than the number of pixels in the image.

A. . the Construction Phase

First of the first both the image pixels (p) and the sound samples (s) must be normalized since the image pixels occupy the range of $0 \le p \le 255$ while the sound sample occupies the range of $-1 \le s \le 1$ in Matlab.

The pixels of the image will multiplied by 4 so the overall range will be [0 → 1000] while the speech samples will be multiplied by 500 and add to 500 so each sample will be in the range of [0 → 1000].

The speech samples will be approximated to three significant figures (after using the function **ceil** which magnifies the number to the nearest larger integer) which is the same as the pixels range. This means that the speech samples will be written in three significant figures after comma, which is very acceptable, where higher resolution can be achieved by normalizing to 10000 leading for four significant figures after comma.

Now both the pixels and the samples are on the same range which consists of 10 bit binary representation where both size and accuracy are acceptable where this represents the first step in data hiding which raises the standard of coding and reduces the distortion rate.

Second, the system will start to evaluate the WPT decomposition of 8-points at a time for both the sound samples and the image pixels where each output of the process of evaluating the WPT decomposition for the image pixels will be in the place of the HHH output only of the process of evaluating the WPT for the sound samples. That means that we are in need for 64 sound sample and 8 times of 8-points WPT processes on the sound file to hide the output of one WPT process on 8-points WPT image pixels.

Third, the exchange process will be implemented by changing the HHH of the WPT of the speech samples by the WPT of the image pixels, where each HHH of the speech sample is replaced by the image pixel where the effect of the HHH change is at the minimum effect.

Fourth, evaluating the IWPT of 8-points reconstruction process for the new speech samples after data embedding in-order to reconstruct the speech file.

At the end, the speech file must be reconstructed to be as same as possible to its origin file shape by subtracted from 500 and divided by 500 to be within the range of $-1 \le s \le 1$ which is in the same format of the input file before processing for changing the file format from image to media player.

The rate of change in the embedded sound file from the original one is evaluated as in the following formula where the error E(e)={(Real-Embedded)/Real}. So the error doesn't exceed 0.01% which is very acceptable

B. the Re-Construction Phase

The Re-Construction phase started from the moment of receiving the encrypted speech file where the process begins by normalizing the speech file where the samples will be multiplied by 500 and add to 500 so each sample will be in the range of [0 → 1000] just like in the Construction phase (where the processes in the transmitting side must follow the same standard in the receiving side so if the standard is raised to 10000 then the process will be multiplied by 5000 and added to 5000 and so on in order to unify the system operation base).

Segmenting the sound file into 8-samples group is the second process which is implemented to evaluate the 8-points WPT decomposition to the third level to evaluate the HHH part of each segment. The HHH of each segment is extracted from the WPT evaluation to inform the image file where arranging each (8) HHH from the speech file to reconstruct a single pixel in the 8-points IWPT in the image file. This means that from 64 point of WPT from the speech file the system will be able to obtain and reconstruct a single pixel of the image file.

After finishing the construction of the WPT of the image file from the speech file, a segment of each (8) point gathered together to implement a (8) point IWPT block which represents the image pixels in the special domain and they must be retrieved by evaluating each pixel in the time domain again but it will be in the range of [0 → 1000] just like in the Construction phase after normalization to hold the system standardization where retrieving the original image shape is achieved by dividing each pixel by 4 and chose the **floor** function from the Matlab software to retrieve the original image shape in the range $0 \le p \le 255$to reconstruct the original image brightness levels.

More than 94 % of the retrieved pixels are the same as in the original image while the other pixels are suffering from very little distortion which may not affect the image since the overall rate of change in the distorted pixels not exceed 0.8% on most cases (for example if the real value R is 223 then the reconstructed one E^s will be in the range of $221 \le E^s \le 225$

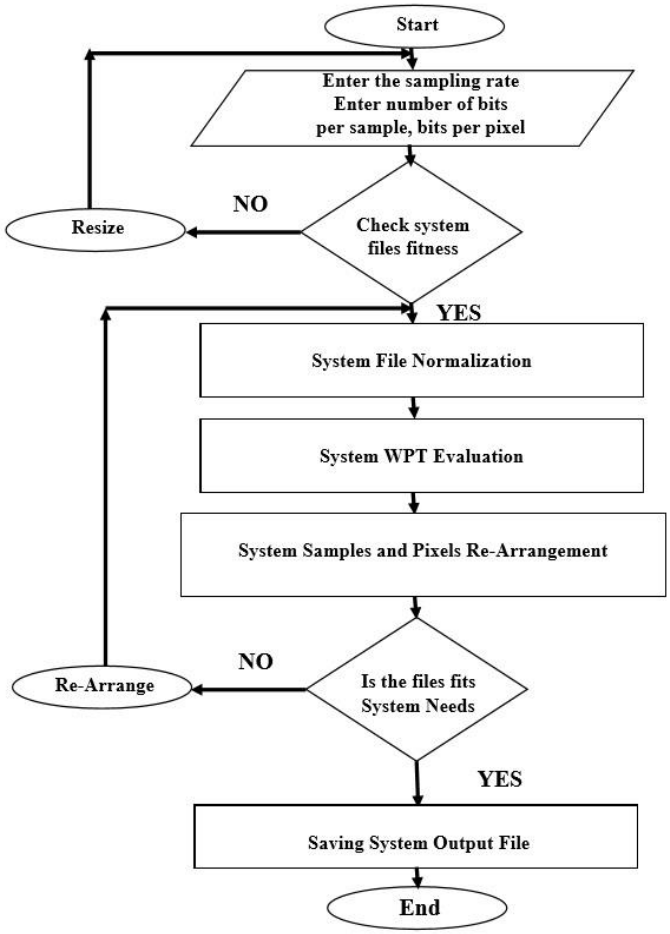

Figure 3:- System operation flow chart for both phases

V. RESULTS AND CONCLUSIONS

The algorithm is very simple and efficient since Mallat Multi resolution Analysis is used in implementing the WPT where transforming the image and the speech to the frequency domain doesn't contains the complex part (imaginary part) and multiplication process (convolution in the BPF bank of filters) use only 1 & -1 coefficients of the wavelet transform (Haar Coefficients), therefore multiplication is tend to be either addition or subtraction, therefore the processing time is very small and the application is real time even if the image is tend to be a movie and the sound is tend to be a long time speech.

On the other hand the system accuracy is very acceptable since the maximum **theoretical** distortion is 12.5% (because the maximum rate of change is one of eight and in case of total difference, i.e. if the speech HHH is 1011010110 and the image is 0100101001 which is 100% different). While in practical cases the distortion doesn't exceed 6% where 94% image reconstruction rate is achieved where the probability of bit-matching is 50% which raises the construction probability by reducing the total distortion rate which is 12.5% to 50% which will be 6.25% and raises the reconstruction rate to 93.75%.

Higher rate may be achieved if the image contains edges and corners (not a flat image like a desert picture or a snow field picture for example), as well as if the decomposition and hence reconstruction reaches to the fourth level, i.e. we will get HHHH, where the maximum distortion rate will be 6.25% and the probability bit-match will be 3.125%, hence the practical construction rate may be 97% or higher.

Side by side, the sound file is not recognized by the human eye like the image therefore the sound file security is much higher than the image file especially if the sound file is not in (wav) extension where in this case an extension change process must be accomplished which is very easy on Matlab with a negligible processing time.

An important result is achieved where the overall size of the sound file is not changed, that's mean that the transmission channel remain unchangeable and there is no need for more powerful media to hold-up the file update which may be noticed by third party.

REFERENCES

[1] J. "Space Time Coding Techniques for Wireless Communications Using MIMO System for Channel Estimation", International Journal of Engineering Science and Technology (IJEST) ISSN : 0975-5462 Vol. 4 No.07 July 2012.

[2] . "A robust encryption method for speech data hiding in digital images for optimized security", **IEEE *Xplore*:** conference, **DOI:** 10.1109/PERVASIVE.2015.7087134 , IEEE, Pune, India, 16 April 2015.

[3] "Speech steganography using wavelet and Fourier transforms", Siwar Rekik1,2*, Driss Guerchi2, Sid-Ahmed Selouani3 and Habib Hamam4 , doi:10.1186/1687-4722-2012-20 , EURASIP Journal on Audio, Speech, and Music Processing 2012.

[4] Y Hu, P Loizou, Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun 49, 588–601 (2007)

[5] Y Hu, P Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Speech Audio Process. 16(1), 229–238 (2008)

[6] " Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions".J Ma, Y Hu, P Loizou, J. Acoust. Soc. Am. **IEEE *Xplore*.** 125(5), 3387–3405 (2009)

[7] "Unified phase and magnitude speech spectra data hiding algorithm Authors", Fatiha Djebbar, UAE University, Al Ain, UAE - Beghdad yad, Canadian University of Dubai - UAE, Karim Abed-Meraim, Telecom Paris Tech, Paris, France - Habib Hamam , Faculty of Engineering, University de Moncton, Moncton, NB, Canada, Springer, 11 January 2013, DOI: 10.1002/sec.644