Combined Speech Compression and Encryption using Contourlet Transform and Compressive Sensing

Maher K.M. Al-Azawie
Department of Electrical Engineering,
College of Engineering
Al-Mustansiriya University/
Baghdad

Ali M. Gaze
Department of Electrical Engineering,
College of Engineering
Al-Mustansiriya University/
Baghdad

ABSTRACT
This paper introduces a new technique for Speech compression and encryption in one-step. Speech compression is the process of Converting human speech signals into a form that is compact and is reliable for communication and storage by reducing the size of data without losing quality of the original speech. Speech encryption is the process of converting the normal form of speech into unrecognized form to increase the security of communication through an insecure channel. Compressive sensing theory is used to apply the compression and encryption in one-step; in addition, the contourlet transform is used to prove the principle of Compressive Sensing (CS) (i.e. Spars structure) that is one of the most important aspect of the compressive sensing theory..

General Terms
Speech compression, speech encryption.

Keywords
Compressive sensing, Contourlet Transform.

1. INTRODUCTION
Speech is a vocal signal by nature and it is an efficient medium for face-to-face communication and telephony application such as mobile networks (GSM, CDMA, LTE and ….etc.). The compression process, in general, is the operation that lowers the stream of the data in a specific way without losing quality of the original speech [1]. One of the main purpose of compression algorithms is to remove the redundancy of the data [2]. This lowering in data redundancy will be imparted less space and time for storage of the data in memory. From the point of view of communication the compression process, produces streams of data that is compact with limited channel bandwidth and increased data rate of communication. If speech compression process is used, more users can be accommodated at a given time because of lesser bandwidth for each one. Hence, mobile employee can reduce the cost [3]. Because the data are transmitted through insecure channels of communication, the users need to protect this data from the third party. Therefore, the encryption process must be applied to the data to prove security portion [4]. The encryption is the process that converts the format of speech from normal format into unrecognized (random) format. Therefore, the encryption is the most important part in nowadays communication. The previous works of speech compression and encryption were done by separated algorithms (i.e. The compression process is complete then encryption process is applied to compressed speech). Therefore, in traditional methods of compression and encryption separated two algorithms did the two processes. This paper will explain how to do the two processes in one-step (single algorithm). This is achieved by compressive sensing algorithm. The compressive sensing is a new pattern and a new tactic to simultaneously provide two jobs in the same step, (i.e. Sampling and compression) with the influence of encryption of the containing information. Using CS, signal reconstruction performance can be commercialized with the willing processing power at the receiver. Regular rate digitization is followed by compression is overwhelmingly used form of transmission or storage [5-6]. CS has attracted attention of researchers since "2006". The classic method of sampling signals such as Nyquist theorem declared for the exact reconstruction of sampled signals must be "twice-maximum" signal frequency (i.e. FS $\geq$ 2FM) to guarantee the exact reconstruction of the signal. CS offers acquisition of signal under Nyquist rate, (i.e. taking the largest elements in original signal only). The primary theory of CS declared that if the signal is familiar to be sparse in representation, CS could do the sampling and compression under Nyquist rate with encryption benefits [6]. The sparse possession is a measure of signal redundancy and CS statement to use this property in the best way at the signal sampling port [5]. Furthermore, many signals, such as speech and audio signals, which are, include the sparse possession in some linear transform domain of the signal such as wavelet transform (WT) or Discrete Cosine Transform (DCT), Contourlet Transform (CT). This paper is organized as follows. Section 2 briefly explains the compressive sensing theory, section 3 briefly explains contourlet transform, section 4 discusses the proposed algorithm, section 5 gives simulation and results, and section 6 gives the conclusion of the work.

2. COMPRESSIVE SENSING THEORY
Compressive sensing (CS) is a new way to simultaneous sensing and compression of sparse or compressible signals, i.e. speech signal. Candes et al [9], and Donoho [8], introduced the theory of compressive sensing (CS) in "2004". The CS theory showed in this section is appropriate from [7]. In classical methods the signals are sampled according to Nyquist theorem, whereas by using "CS" it is possible to sample the signals under Nyquist rate. It is possible to convert the signal in particular transform domain that has spares exemplification such as (DCT, DWT, and CT). By using, many of optimization techniques one can reconstruct the spares s signal. The basic block diagram of a compressive sensing is shown in Figure 1.

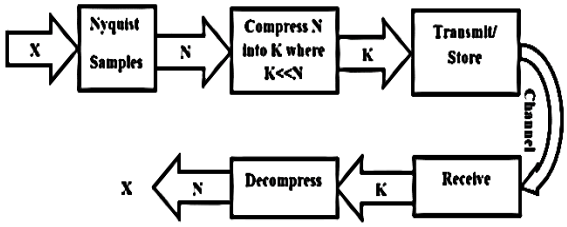


Fig1. Basic block of compressive sensing

To comprehend the theory of CS, let X to be speech signal where X∈ $\boldsymbol{R^n}$, and let Ψ to be a basis vector to span in $\boldsymbol{R^n}$ and Ψ is:-

Ψ =[$\Psi_1, \Psi_2, \Psi_3, \ldots \ldots \ldots \Psi_N$]. The speech signal is said to be spares if :-

$$\boldsymbol{x_i = \sum_{i=1}^{k} s_i \Psi_i , [n_1, n_2, .. n_k] c [1, 2, \ldots, N]} \quad (1)$$

Where $\boldsymbol{S_i}$, are scalar coefficients and k ≪N, i.e. $\boldsymbol{s_i}$ or simply S is the sparse vector with only k non-zero components. Speech signal has been sampled by dropping onto the random basis and at "receiver side"; the signal is reconstructed by full information on the random basis. In other words, the sampling step is done as:-

$$\boldsymbol{yi = \sum_{i=1}^{k} \emptyset_{m(i_{M\times N})} x_{(i_{N\times N})} \quad 1 < m < M \ll N} \quad (2)$$

Where Y=∅ × $\boldsymbol{x}$; and $\boldsymbol{\emptyset_{m\times N}}$ is a sensing matrix. The ∅ is made up of orthonormal random basis vector ∅m, and Y is a measurement vector. To illustrate the idea of CS see figure (2).

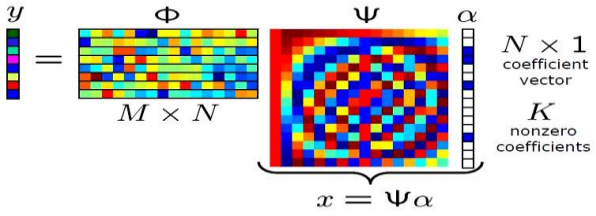


Fig2. Principle of compressive

If ∅ and Ψ **are** incoherent to satisfy RIP condition**, Y** can be exactly reconstructed if **(** $\boldsymbol{M > k\, Log(N)}$ **),** as presented in **[**10]. Convex optimization, can then be utilized as follows :-

$$\boldsymbol{\hat{s} = arg\ min ||s||_1} \textbf{ Subjected to Y=} \emptyset \times \Psi \times \boldsymbol{x} \textbf{ and } \hat{x}\textbf{=}\Psi\times\hat{s} \quad \textbf{(3)}$$

Where $||\bullet||_1$is the $\boldsymbol{l_1}$ norm**.** The algorithm above is also known as basis pursuit **(**BP**)** since a subset of the column vector of [∅Ψ] is being particular. One of the efficient algorithm to solve CS is "orthogonal matching pursuit" **(**OMP**), [11],** which can be formulated as follows:

$$\boldsymbol{\hat{s} = arg\ min ||y - \emptyset\ \Psi\ S\_ ||_2\ and\ ||s||_0 = k} \quad \textbf{(4)}$$

Many solutions to sparse approximation have been put, "such as Matching Pursuit" **(**MP), "Least Absolute Shrinkage and Selection Operator" **(**LASSO**), "**Basis Pursuit" **(**BP), "Gradient Pursuit" **(**GP**),** in which its rendering shows some interdependence among the numbers of measurement, "measurement noise", "signal sparsity", and "the reconstruction algorithm itself" [12, 13]. Due to the time varying property of speech signal, a continuous speech signal must be framed with a frame length of about (10_25) **ms**. Then to get sparse vectors for each frame, orthonormal transform may be used, such as, Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Contourlet Transform (CT). Each of these can prove sparse representation of speech vector, but contourlet transform **(**CT**)** proves more spares than other transformations. Therefore, the best transform that provides a higher spares representation index be selected **[**14**].**

3. CONTOURLET TRANSFORM

The contourlet transform is one of the new geometric 2D data transforms, which can expeditiously represent 2D data elements contours and textures. It is an efficient impersonation of signals that demands the coefficients of functions, to be sparse. Wavelets can elect up discontinuities of one-dimensional piecewise sleek functions very efficiently and represent them as point discontinuity. 2D WT executed by a tensor product of one-dimensional wavelets are good to depose discontinuities, at edge points but cannot recognize softness along contours. This transform uses a structure similar and based on curvelets that is a stage of sub band decomposition followed by a directional transform. Numerous methods have been sophisticated to solve this by "adaptive series" [16], or "filter bank-based techniques" [17]. In the contourlet CT transform a "Laplacian pyramid" (LP) [18, 19], is utilized for the first step to decompose the 2D data, while "directional filter banks" (DFBs) [20], are used in the angular decomposition stage to give directional shapes. The contourlet transform CT is built as a combination of the "Laplacian pyramid" (LP) and "the directional filter banks" (PDFB). Conceptually, the flow of procedure is illustrate in Figure (3), where the "Laplacian pyramid" (LP) iteratively decomposes a 2-D image into low pass and high pass sub bands, and the "directional filter banks" (DFBs) are applied to the high pass parts to give the frequency spectrum [15]. CT may be decomposed into many levels and each level is divided into a number of directions. The CT directions number is different from level to others.

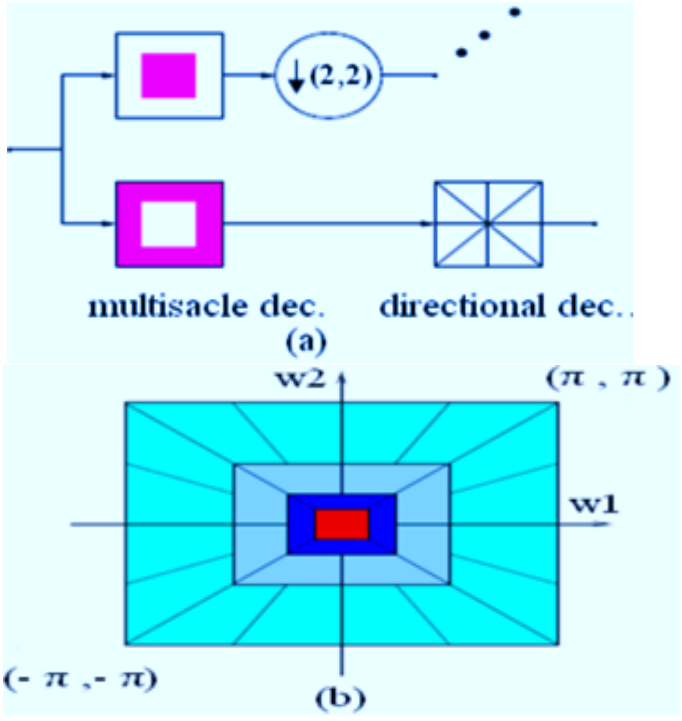


Fig 3. The Original Contourlet Transform. (a) Block Diagram. (b) Resulting Frequency Division

The preference to contourlet on wavelet is that the high frequency sub bands of contourlet do not expose to down sampling process as what occurs in wavelet that makes the high frequency scrambled. Contourlet Transform (CT) can capture softness transitions and edges exist in speech signals [21], as shown in figure (3). The contourlet transform proves good performance in high compression ratio (i.e. when the remaining size of original signal below 50%)? Due to CT gives, high spares structure of the 2D data with the full advantage from k coefficients to apply the compressive sensing theory with high compression ratio (CR)**.**

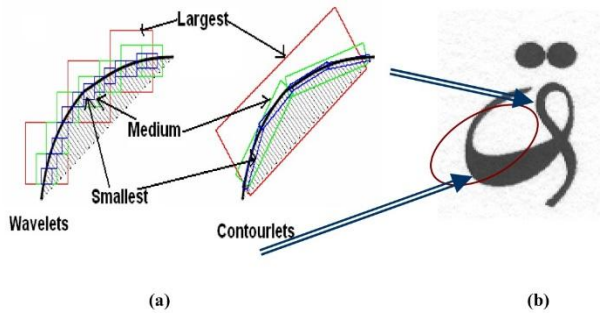


Fig 4. Wavelets vs. Contourlet

The high sparsity of CT is used in this work as a transform domain for speech signal**.** However, CT deals with 2D data only, so must look about some way to convert speech signal from 1D into 2D. Using spectrogram of speech, the problem is fixed and can read the spectrogram by CT, so the problem of dimensionality of speech is solved that prevents it to be used by contourlet transform CT.

4. COMBINED SPEECH COMPRESSION AND ENCRYPTION USING CONTOURLET TRANSFORM AND COMPRESSIVE SENSING

Proposed Algorithm

The proposed algorithm is illustrated in figure (5). Where the speech file is framed from the start up to the end of the file and the frames are arranged in some way to convert the speech into 2D (spectrogram) initialization to be used by CT**.** The CT acts as transform domain to produce spares structures of the spectrogram (2D), so that the spares structure is achieved and the data compatible with the principle of compressive sensing **(**CS**).** Next step is to use CS for encryption and compression in one-step by applying each sub band of CT to one or different sensing matrices (∅) for encryption and compression. The compression process is done by removing the redundancy of signal and reducing the spectrogram size by sampling important samples only. The encryption process is based on sensing matrix(∅), which is Gaussian independent and identical distribution **(**IId**)** that has overwhelmed probability of restricted isometric property **(**RIP**).** RIP condition proves perfect reconstruction of spares vectors and this condition indicates high incoherence between sensing matrix and bases of CT**.** At the receiver side, **L1** minimization algorithm is applied to the spares vector reconstruction because it gives spares representation of the vector and all other steps are done inverse of the steps at the transmitter side. These steps of the algorithm are shown in the figure below:-

Fig 5. System flow chart

5. SIMULATION RESULTS

The tested speech file is loaded from ***NOIZEUS*** database. Two files are used one of them contains only vowels (voiced) speech, the other is continuous speech containing voiced and unvoiced. The sampling frequency is **16KHZ** and time duration of frames is **10ms.** Some speech quality measures are used such as **(**SNR, SSSNR, and MSE**),** their definitions are given by:-

[1]. Signal to Nosie Ratio (SNR):-

This test gives the value of speech signal energy to noise energy [22], it's calculated as:-

$$\boldsymbol{SNR = 10log_{10}\left[\frac{6x^2}{6z^2}\right]} \quad [\boldsymbol{dB}] \qquad \textbf{(5)}$$

Where $6x^2$ is the mean square of the speech signal, $6z^2$ is the mean square difference between the original and reconstructed signal.

[2]. Segmental Spectral Signal to Noise Ratio (SSSNR):-

This test shows the Encryption strength of the system and it takes positive or negative values. Smaller values of SSSNR indicate highest encryption strength. It's is given by:-

$$\boldsymbol{SSSNR = 20\ log_{10}\frac{\sum_{k=1}^{n}|X(k)|}{\sum_{k=1}^{n}|X(k)-Y(k)|}\ [dB]} \qquad \textbf{(6)}$$

Where X (k) is the DFT of original signal, and Y (k) is the DFT of encrypted signal [22].

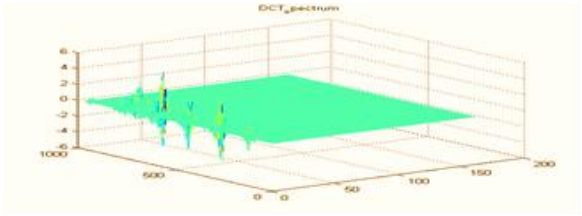
[3]. Mean Square Error (MSE):-

This test shows the difference (error) between the original signal and reconstructed signal. The low value of MSE meaning high quality of the reconstructed signal [22], MSE is given by:-

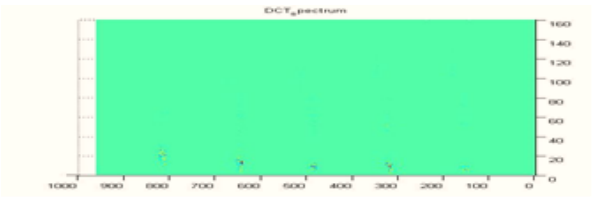
$$MSE = \frac{\sum(y-x)^2}{n} \quad (7)$$

Where x is original signal and y is reconstructed signal and n is the length of the signal.

"**Compressive Sensing based spares vector with Contorlet Transform**" (CSCT) carried out the test**.** In this, test the spectrogram of speech is shown in figure (6), where the frames are reshaped vertically one beside to others to produce this (2D) structure that is analyzed by CT with one level with eight directions as shown in figure (7) below:-



(a)



(b)

Fig 6. Spectrogram of speech.
(a)(3D) spectrogram. (b) (2D) spectrogram

Fig 7. CT sub bands analysis of spectrogram of speech (2D)

These sub bands of CT (the low frequency sub band is the square located at the upper left corner of figure (7), while the high frequency sub bands are located elsewhere with rectangular shapes). Each of these sub bands is sampled either by one sensing matrix $[\emptyset_{i(m\times N)}]$ or if it is required to increase the security of the system, different sensing matrices are used for each sub band. The outputs from the CS are the measurement vectors $[\boldsymbol{Y_{i(m\times N)}} = \emptyset_{i(M\times N)} \times \boldsymbol{X_{i(N\times N)}}]$, where the size of $[\boldsymbol{X_{i(N\times N)}}]$ is reduced from $[N \times N]$ into $[M \times N]$ where $[M < N]$. Since the elements of $\emptyset_{i(M\times N)}$ are selected from a Gaussian random variable $\emptyset_i$ will scramble into Yi**.**

At the receiver side, L1 minimization algorithm **(**convex optimization**)** is used to ensure convergent and accurate reconstruction of $\boldsymbol{X_{i(N\times N)}}$ from received $\boldsymbol{Y_{i(N\times N)}}$**.** These $\boldsymbol{X_{i(N\times N)}}$ will construct the spectrogram of speech at receiver side by Inverse Contorlet Transform (ICT). Finally, the spectrogram (2D) is used to get original speech frames (1D).

The simulation results are listed in (1) and (2) for vowels speech (voiced) and continuous speech (voiced and unvoiced) with compression ratio **(**CR**)** expressed as:-

$$CR = \frac{size\ of\ remained\ \ vectors}{size\ of\ orignal\ speech} \quad (8)$$

Table 1. Results Of Vowels Speech (Voiced)

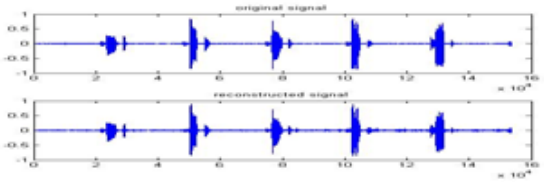
CR	SNR(dB)	SSSNR(dB)	MSE
80%	20.30	-24.31	6.5e-05
60%	14.61	-22.35	0.00020
50%	12	-24.16	0.00037
40%	8.6	-24.27	0.00074
30%	7.04	-28.32	0.001
20%	2.81	-28.45	0.003

Table 2 Results Of Continuous Speech (Voiced And Unvoiced)

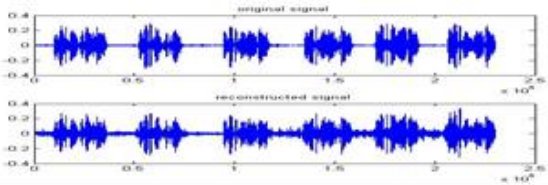
CR	SNR(dB)	SSSNR(dB)	MSE
80%	13.45	-26.78	0.00019
60%	9.17	-20.06	0.0004
50%	7.01	-22.40	0.0006
40%	4.25	-24.92	0.0011
30%	3.49	-24.49	0.0013
20%	1.40	-24.79	0.0021

In both tables, the SSSNR measure gives the encryption strength as a measure of residual intelligibility while the remaining measures give the quality of the reconstructed speech, With CR as a parameter.

The Figure shows the original and reconstructed signals of vowels speech (voiced) and continuous speech (voiced and unvoiced) with **[CR=50%].**



(a)



(b)

Fig 8. Original and reconstructed signals with CR=50% , (a) vowels speech (voiced) and (b) continuous speech (voiced and unvoiced).

6. CONCLUSION

A combined speech compression and encryption are done in one-step using Contorlet Transform (CT) and Compressive Sensing (CS). The CT gives sparsity to CS that provides both compression and encryption. Simulation results showed that a reasonable quality of reconstructed speech and encryption strength are obtained for both Vowels speech (voiced) and continuous speech (voiced and unvoiced) signals with compression ratio (CR) being a parameter. The next work will include the chaos theory to improve the high security level of the system. The chaos will be added as a radix for sensing matrix of the compressive sensing.

7. ACKNOWLEDGMENTS

All thank for those who had helped in this work and give support and first of them professorial Mr. Maher Al-Azzawi, and thanks to the family and thank you for your magazine estimable.

8. REFERENCES

[1] Lawrence R Rabiner 'Digital Processing of Speech Signals' (2nd Ed),Pearson Education, 2005 ISBN 81-297-0272-X.

[2] Kalid Sayood 'Introduction to Data Compression' (2nd Ed), Morgan Kaufmann Publishers, 2005.ISBN 81-8147-191-1.

[3] S.Haykin, Communication Systems, (4th Edn) John Wiley & Sons, New York, 2001.ISBN 0-471-17869-1.

[4] Changgui Shi,Bharat Bhargara, (1998). "Fast MPEG Video Encryption Algorthim", Department of computer Sciences, Purdue University.

[5] E. J. Candes and M. B. Wakin, "An Introduction to Compressive Sampling," IEEE, Signal Processing Magazine, pp. 21-30, 2008.

[6] R. G. Baraniuk, "Compressive sensing," IEEE Signal Processing Magazine, vol. 24, pp. 118-121, 2007.

[7] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4125-4128, 2009.

[8] D.L. Donoho, "Compressed Sensing," IEEE Transactions on Information Theory, vol. 52, pp.1289-1306, 2006.

[9] E. Candes, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information," IEEE Transaction on Information Theory, vol. 52, pp. 489–509, 2006.

[10] E. J. Candes and M. B. Wakin, "An Introduction to Compressive Sampling," IEEE, Signal Processing Magazine, pp. 21-30, 2008.

[11] J. A. Tropp and A. C. Gilbert, "Signal Recovery from Random Measurements via Orthogonal Matching Pursuit," IEEE Transactions on Information Theory, vol. 53, pp. 4655-4666, 2007.

[12] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction," IEEE Journal of Selected Topics in Signal Processing, vol. 1, pp. 586-597, 2007.

[13] H. Rauhut, K. Schnass, and P. Bandergheynst, "Compressed sensing and redundant dictionaries," IEEE Transactions on Information Theory, vol. 54, pp. 2210-2219, 2008.

[14] N. Hurley and S. Rickard, "Comparing Measures of Sparsity," IEEE Transactions on Information Theory, vol. 55, pp. 4723-4741, 2009.

[15] Zanartu M., 2005," Audio Compression using Wavelet Techniques", University Purdue, Electrical and computer engineering.

[16] E.L. Pennec and S. Mallat. Image Compression with Geometric Wavelets. IEEE International Conference on Image Processing, 2000.

[17] M. Do. Directional Multiresolution Image Representations. Ph.D. Thesis, Department of Communication Systems, Swiss Federal Institute of Technology Lausanne, November 2001.

[18] P. J. Burt, E. H. Adelson. The Laplacian pyramid as a compact image coder. IEEE Trans. Commun., Vol.31 (4):532-540, April 1983

[19] M. Do and M. Vetterli. Framing Pyramids. IEEE Trans. On Signal Processing, VOL. 51, NO.9, September 2003.

[20] M. Do and M. Vetterli. Contourlets. In: J. Stoeckler, G. V. Welland (Eds.), Beyond Wavelets, pp.1-27., Academic Press, 2002.

[21] M. N. Do and M. Vetterli, "Contourlets," in Beyond Wavelets, Academic Press, New York, 2003.

[22] Journal of Engineering and Development, Vol. 17, No.4, October 2013, ISSN 1813- 7822 Speech Scrambling Employing Lorenz Fractional Order Chaotic System.

[23] J. Benesty, M. Sondhi, Y. Huang (Ed.), Springer Handbook of Speech Processing. Berlin Heidelberg: Springer-Verlag, 2008.

[24] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Speech and Audio Processing, 16(1), pp. 229-238, 2008.