The ISC Int'l Journal of Information Security

July 2017, Volume 9, Number 2 (pp. 131–145)

http://www.isecure-journal.org

Cipher Text Only Attack on Speech Time Scrambling Systems Using Correction of Audio Spectrogram

Hamzeh Ghasemzadeh [1,2,*], Mehdi Tajik Khasss [3], and Hamed Mehrara [4]

[1] *Department of Communicative Sciences and Disorders, Michigan State University, Michigan, USA*
[2] *Department of Computational Mathematics Science and Engineering, Michigan State University, Michigan, USA*
[3] *Department of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran*
[4] *Department of Electrical Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran*

ARTICLE INFO.

Article history:
Received: 1 August 2016
First Revised: 27 February 2017
Last Revised: 18 June 2017
Accepted: 22 July 2017
Published Online: 26 July 2017

Keywords:
Cryptanalysis, Cipher text only attack, Audio scrambling system, Multimedia encryption systems, Jigsaw puzzle, Spectrogram.

ABSTRACT

Recently permutation multimedia ciphers were broken in a chosen-plaintext scenario. That attack models a very resourceful adversary which may not always be the case. To show insecurity of these ciphers, we present a cipher-text only attack on speech permutation ciphers. We show inherent redundancies of speech can pave the path for a successful cipher-text only attack. To that end, regularities of speech are extracted in time and frequency using short time Fourier transform. We show that spectrograms of cipher-texts are in fact scrambled puzzles. Then, different techniques including estimation, image processing, and graph theory are fused together in order to create and solve these puzzles. Conducted tests show that the proposed method achieves accuracy of 87.8% and intelligibility of 92.9%. These scores are 50.9% and 34.6%, respectively, higher than scores of previous method. Finally a novel method, based on moving spectrogram distance, is proposed that can give accurate estimation of segment length of the scrambler system.

1 Introduction

Nowadays multimedia signals are present in every aspect of our daily life. Therefore, an increased amount of work has been published on their applications and their security aspects including encryption systems, steganography [1], and steganalysis [2]. One possible security aspect is confidentiality, which speech scrambler systems try to address. High speed, low power consumption, excellent voice recognition capabilities, independency of the quality of recovered signal from the language and speaker, possibility of directly coupling them to any handset, removing the need for speech compression and modem, and most important of all, compatibility with the existing analog and narrowband communication channels are among the most important advantages of speech scrambling systems [3].

Reviewing the literature shows that there are different methods for realizing this task. The hopping-window time domain scrambler permutes segments of audio signal in the time domain. On the other hand, the band splitting scrambler, breaks the spectrum of the signal into several sub-bands and then performs permutation and frequency reversion on them [4]. In another work, chaotic system was used to interleave voice packets into different frames of network for real-time voice over IP (VoIP) applications [5]. Owing to

* Corresponding author.
Email addresses: ghasemza@msu.edu (H. Ghasemzadeh), email@mail.com (M. Tajik Khass), email@mail.com (H. Mehrara)
ISSN: 2008-2045

the dawn of high-speed signal processing hardware, recently, more complex methods have been proposed. This class of scramblers works in the transform domains. First, they apply some transformation on the digitized signal. Then, after permuting the resultant coefficients, the inverse transform is applied to create the scrambled signal. Hadamard matrices were used for speech scrambling in [6]. Sakurai et al. [7] proposed a method based on re-arrangement of the coefficient of fast Fourier transform (FFT). To improve security of the system, adaptive dummy spectrum insertion technique was used [8]. A method based on discrete cosine transform (DCT) was suggested in [9]. Goldburg et al. investigated the potency of audio scrambling in transform domain [10]. They considered discrete Fourier transform, DCT, Walsh-Hadamard transform, and discrete prolate spheroidal transform. Later, it was shown that DCT is the best transform to be used for transform-based encryption [11]. A practical system based on ITU-T G.723.1 speech codec and DCT permutation was proposed in [12]. To increase the effective number of permutations, an scrambling algorithm based on the wavelet packets was proposed in [13]. Another method based on the parallel structure of two different types of wavelet with the same decomposition level was proposed in [14]. Tseng et al. argued that most scrambling methods preserve the signal energy, talk spurts, and the original intonation. To solve these problems, they proposed a method based on orthogonal frequency division multiplexing [15]. Li et al. increased the dimension of sample points coordinates and used arbitrary matrices to scramble audio signals [16]. Another work used intractability of the problem of underdetermined blind source separation for speech encryption [17]. The idea of progressive scrambling/descrambling in the wavelet domain and MP3 files was proposed in [18]. This idea allows a person to retrieve different qualities of audio signal, based on the amount of key that he knows. Different approaches for joint multimedia compression and encryption were discussed in [19]. Zeng et al. used compressed sensing ideas to construct a robust scrambling method against active attacks [20]. Linear feedback shift register (LFSR) was used for selective encryption of compressed audio signals [21]. Hierarchical selective encryption of G.729 standard was investigated in [22]. According to this method, bit stream is partitioned into the most sensitive and the least sensitive parts. Then, different chaotic maps are used for encryption of each part. Finally, joint scrambling and watermarking of MP3 was discussed in [23, 24].

1.1 Related Work

While different audio encryption and scrambling methods have been proposed, very few works have been published on their security evaluation. First, Goldburg et al. employed a frequency domain vector codebook to estimate a model for space of clear speeches [11]. Then, this model was used to implement a ciphertext-only attack on fixed and varying permutation frequency domain scramblers. Later, this method was extended to cryptanalysis of DFT-based systems [25]. Spectral distance measure between the end of each segment and the start of all other segments was employed for cryptanalysis of time scrambling systems in [26]. Finally, Zhao et al. proposed a technique for solving rectangular jigsaw puzzles and mentioned audio cryptanalysis as a possible application, but the paper neither provided the means for converting the speech into jigsaw puzzles, nor presented any results on such cryptanalysis system [27].

Recently, security of permutation based multimedia encryption systems were studied [28]. The work showed that "most (if not all) permutation-only multimedia ciphers" can completely be broken in a chosen-plaintext scenario. In a chosen-plaintext scenario, the attacker can obtain the ciphertext of any plaintext he wishes [29]. Apparently, this model of attack is very powerful and does not hold in many practical situations. In fact, this model is mostly applicable to the public key cryptography systems [29]. Another model of attack is ciphertext-only attack. In this model, the attacker has only access to a set of ciphertexts. Not only this model is more realistic and applicable to any cryptosystem, but also this vulnerability leads to vulnerability to chosen-plaintext model whereas the reverse may not hold. In order to show that permutation based systems are insecure in the practical setting, we present a ciphertext-only attack on these systems. We show that time domain permutation based audio encryption systems can be broken very efficiently. Also, the proposed attack is independent from key generation mechanism; therefore, no better pseudorandom permutation mapping can be realized to offer a higher level of security against the proposed attack. Through different simulations we show that the proposed method results in very high level of intelligibility even if the scrambling system uses small segment lengths and high frame sizes. In addition, the proposed system can retrieve quite intelligible speeches even when the value of signal to noise ratio (SNR) is low. The main contributions of this work are:

- Continuing on our work [30], we present a practical ciphertext-only attack on time domain au-

dio scrambling system. Therefore, these systems can be considered totally broken. The attack is based on converting scrambled audio signals into two dimensional image puzzles. The suitable transformation and its optimum parameters are also presented.

- We show that the proposed transform has a windowing effect which causes a discontinuity between correct segments. This phenomenon may drastically reduce performance of the system. To solve this issue, we present a novel solution where the boarder samples are estimated. We show that this technique can mitigate the aforementioned problem.
- Jigsaw solving algorithms rely on a distance metric for assessing how well two pieces match each other. Because an ideal estimation does not exist, even for true neighbours, a small discrepancy between the borders would remain. A new distance metric is introduced that can mitigate this effect.
- Both objective and subjective tests are conducted to quantify efficacy of the proposed system. Also, effect of different parameters of scrambling system including its segment length, its key size, and SNR of the channel on the performance of the proposed method are investigated.
- Behaviour of distance between two consecutive small windows of scrambled speech is analysed and it is shown that this behaviour can be exploited for finding the correct value of segment length.

The rest of this paper is organized as follows. Section 2 presents concepts of hopping window time domain scrambler and STFT. Then, it elaborates on converting scrambled signal into jigsaw puzzles, solving the puzzles, and estimating segment length of scrambler system. Section 3 presents results of simulations and tests. Finally, discussion and conclusion are presented in Section 4 and Section 5.

2 Proposed Cryptanalysis Method

The hopping window time-domain scrambler maintains confidentiality through permutation of signal in the time domain. To that end, first, audio signal is split into frames with predefined duration. Then, each frame is further split into N non-overlapping segments, where N is called the frame size. Afterwards, segments within each frame are re-arranged according to a permutation key [4]. In order to improve the security, different keys should be used to permute segments of different frames [25].

Speech signals can be represented very efficiently using STFT transform. This transform is a 2D transform from time domain to frequency-time domain and it can visualize any time varying signal versus its spectrum of frequencies. Principally, the signal is multiplied by a sliding window (typically Hamming). This window sweeps the signal in time domain (usually with some overlap) and then magnitude of the frequency spectrum is calculated using FFT. Considering signal $x(n)$, its STFT representation is defined as [31]:

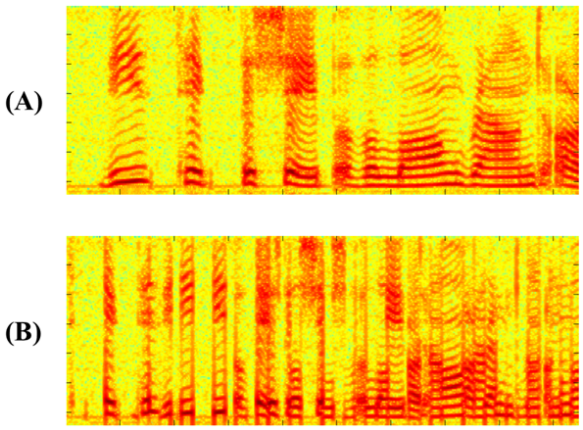

Figure 1. Comparing continuity of STFT of clear and scrambled speech in the time domain. A: Clear speech, B: Scrambled speech.

$$X(m,w) = \sum_{n=-\infty}^{+\infty} x[n].W_m(n).e^{-jwn} \qquad (1)$$

where $W_m(n)$ is a sliding window with size of δ and overlap of γ samples. This window has non-zero values in the interval of $[m,\ m+\delta-1]$ and it is zero elsewhere. By changing the window size (δ), it is possible to trade-off between time and frequency resolutions. If we use an M-point FFT for calculation of Equation 1 and $x(n)$ has B samples, the size of $X(m,w)$ can be calculated as:

$$1 \le m \le \lfloor \frac{B-\gamma}{\delta-\gamma} \rfloor, \qquad 1 \le w \le \frac{M}{2} \qquad (2)$$

Speech signal can be predicted both accurately and efficiently [32]. Speech coding techniques, exploit this feature to achieve high compression rates. Investigating different clear speeches shows that these characteristics are reflected very clearly in their spectrograms. In this manner, spectrogram of a clear voice is smooth in both directions (time and frequency) and it exhibits no abrupt transitions. On the other hand, spectrogram of a scrambled voice shows abrupt transitions in the border of segments. Figure 1 demonstrates these phenomena.

Regarding Figure 1, it is possible to distinguish between clear and scrambled signals. This observation can be exploited for cryptanalysis of scrambled voices. In this manner, cryptanalysis of the scrambled signal

can be considered as re-arranging its time segments in a fashion that the resulting signal produces a regular and smooth spectrogram. This problem is equivalent to solving the jigsaw puzzle of the scrambled signal. Cutting spectrogram into its pieces is the prerequisite for this step.

2.1 Transforming Scrambled Signal into Jigsaw Puzzles

According to Equation 1, a sliding window is applied on the signal to produce the spectrogram. Although this windowing technique adds time resolution to the FFT, but it makes the transition area between two adjacent segments blurry. Figure 2.A shows this fact more clearly. Let *x(n)* and *X(m,w)* denote a scrambled signal and its spectrogram, respectively. Extracting pieces of the puzzle from *X(m,w)* is highly problematic. First, exact border of the segments are not clear in the *X(m,w)*. Second, since the segments are derived from *X(m,w)*, it is very likely that solution of the puzzle solving algorithm be the same as *X(m,w)*. To discuss these phenomena more clearly, assume STFT is implemented using a sliding window with duration of δ samples and the maximum value of overlap ($\gamma = \delta - 1$). Furthermore,

$$x(n) = \{x_{i,j}\},\ 1 \le i \le T,\ 1 \le j \le L,\ n = j + (i-1) \times L \quad (3)$$

where *i*, *j*, *L*, and *T* denote segment index, sample index, segment length, and total number of segments, respectively. While the hamming window is sliding from segment *k* to *k+1*, some transition steps occur. In these steps, head of the sliding window covers the initial samples of segment *k+1* whereas its tail is on the remaining samples from segment *k*. These transition steps are:

$$Lk - \delta + 1 \le m \le Lk - 1, \qquad 1 \le k \le N - 1 \quad (4)$$

According to Equation 4, the width of this transition region is equal to $\delta - 1$ pixels. Furthermore, as *m* is moving from *Lk-δ+1* to *Lk-1*, gradually impact of samples from segment *k* on the *X(m,w)* decreases and impact of samples from segment *k+1* increases. Consequently, instead of occurring in a single pixel which would result a sharp and clear border for each segment, the transition from segment *k* to *k+1* would be smooth and would result in a blurred region. Therefore, the exact borders of segments are not clear. This phenomenon is highlighted in the Figure 2.A.

Reviewing literature on solving jigsaw puzzles shows that border pixels play a vital role in the solution of puzzle. That is, these algorithms choose the arrangement that has the smoothest overall transition in the boarders. As we argued, STFT smooths down

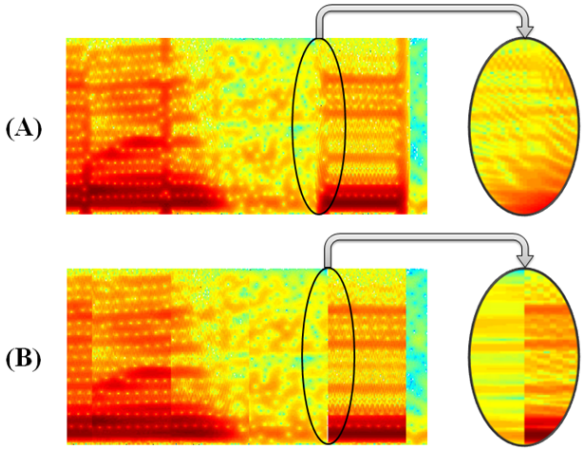

Figure 2. Cutting spectrogram of scrambled speech into pieces of jigsaw puzzle. A: Conventional spectrogram, B: Segmented spectrogram

spectrogram of the encrypted signal. Since the criteria for puzzle solving is also the smoothness, STFT may mislead the algorithm and the solved puzzle may be the same as the spectrogram of encrypted speech.

To solve these problems, first, we cut the scrambled signal into distinct segments. Then STFT of each segment is calculated separately. In this fashion both of aforementioned problems are solved. Figure 2 compares segmented spectrogram and conventional spectrogram of a portion of scrambled speech.

Segmented spectrogram can solve the abovementioned problems, but it also has its own drawback. Let $x_{k,1}, x_{k,2}, \ldots, x_{k,L}$ denote samples in the segment *k*. Investigating mathematical representation of the segmented spectrogram shows that middle samples and border samples act differently. To be more specific, if $\gamma = \delta - 1$, then middle samples contribute in δ sliding windows (*Wm(n)*). That is, if a middle sample is changed then values of δ columns of segmented spectrogram will change. However, in the border samples, this effectiveness gradually reduces to one column. Let *j* denotes index of a sample in the segment k, Equation 5 shows value of effectiveness in the segmented spectrogram. In other words, for sample *j*, $\phi(j)$ shows the number of columns which it is affecting.

$$\phi(j) = \begin{cases} j & \text{if } j < \delta \\ \delta & \text{if } \delta \le j \le L - \delta \\ L - j + 1 & \text{if } j > L - \delta \end{cases} \quad (5)$$

Different values of effectiveness in the border samples causes a discontinuity between consecutive segments even if they are re-arranged correctly. This effect is illustrated in Figure 3.A. In the next section a novel method is proposed to mitigate this problem.

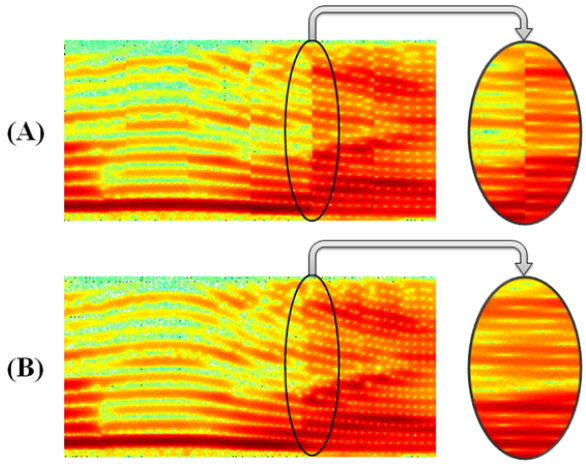

Figure 3. Removing discrepancies between correct neighbors in the segmented spectrogram: A: without estimation, B: with ideal estimation

2.2 Estimating Border Samples

Let *E(x,±l)* denote an ideal algorithm that can accurately predict both *l*-past and *l*-future samples of the sequence *x(n)*. Now we calculate:

$$y = [E(x, -(\delta-1)), x, E(x, \delta-1)] \\ = [x'_{-\delta+1}, ..., x'_{-1}, x_1, ..., x_L, x'_{L+1}, ..., x'_{L+\delta-1}] \quad (6)$$

which is a sequence with *L+2δ-2* samples. Effectiveness values of the middle indices *(1≤j≤L)* of this new sequence are:

$$\phi(j) = \delta, \qquad 1 \le j \le L \quad (7)$$

In other words, in the segmented spectrogram of sequence *y(n)*, all samples of sequence *x(n)* will have the same value of effectiveness. A simulation is conducted to justify that ideal estimation removes discontinuity between segmented spectrogram of the sequences *x(n)*. Figure 3 compares the segmented spectrogram of the correct arrangement of pieces with ideal estimation and without it.

Unfortunately, an ideal estimation does not exist in practice; therefore, we have adopted recursive least squares estimation (RLS) method for this purpose. Let *x(n)* and *d(n)* denote input and output samples of a system, RLS is an adaptive transversal filter of the form [33]:

$$\hat{d}(n) = \sum_{k=0}^{P} W_n(k).x(n-k) \quad (8)$$

such that x_n=[*x(n),x(n-1),...,x(n-P)*] is the vector containing the *P+1* most recent samples of input signal and *P* is the order of filter. Design objective is, estimating weights of the filter at each time *n* (W_n), in a way that the cost function is minimized. The cost function is defined as [33]:

$$C(W_n) = \sum_{i=0}^{n} f^{n-i}.e^2(i) \quad (9)$$

f is called the forgetting factor of the algorithm and it reduces the influence of old data. Furthermore, the error signal (*e(n)*) is calculated as [33]:

$$e(n) = d(n) - \hat{d}(n) \quad (10)$$

2.3 The Proposed method

According to Kerckhoffs's principle [34], it is assumed that security of the system relies only on its key. Therefore, all the parameters of the scrambling system, including its segment length and frame size, are considered to be known. Block diagram of the proposed method is presented in Figure 4.

First, according to the frame length, audio signal is split into non-overlapping frames. Then, each frame is further divided into non-overlapping segments. After that, both $(\delta-1)$-past and $(\delta-1)$-future samples of the border samples in each segment are estimated, where δ is the length of window in the STFT transform. Then, segmented spectrogram is applied on each estimated segment to transform them to time-frequency domain. Going back to Equation 6, if signal *y(n)* is used there would be some overlaps between border samples. Therefore, after converting *y(n)* to its spectrogram, it was trimmed from both left and right to only represent *x(n)* signal. At this stage, each segment has transformed into a two dimensional signal. The resulting signal is quantized and then all segments within the same frame are jointly mapped onto values between 0 and 255 (corresponding to 256 grey scale levels). At this point, a scrambled jigsaw is obtained for each frame of the scrambled speech. These puzzles are fed into puzzle-solving algorithm, where a combination of different image processing techniques and search methods are employed to find the best solution of the puzzle. Finally, the key of each frame is extracted from the solved puzzle and it is used to descramble the corresponding encrypted frame.

2.4 Solving the Puzzle

Puzzle solving is the art of pattern recognition and combinatorial optimization. There are many different practical applications that can be reduced to jigsaw puzzles [35]. Rectangular jigsaw puzzles are a class of puzzles where the borders of all pieces have rectangular shape. Solving rectangular jigsaw puzzles can be broken into two main tasks. First, a proper metric should be devised to assess how well two pieces match each other. This problem is tackled by using signal processing techniques and it results in a set of distance values. Second, the space of all possi-

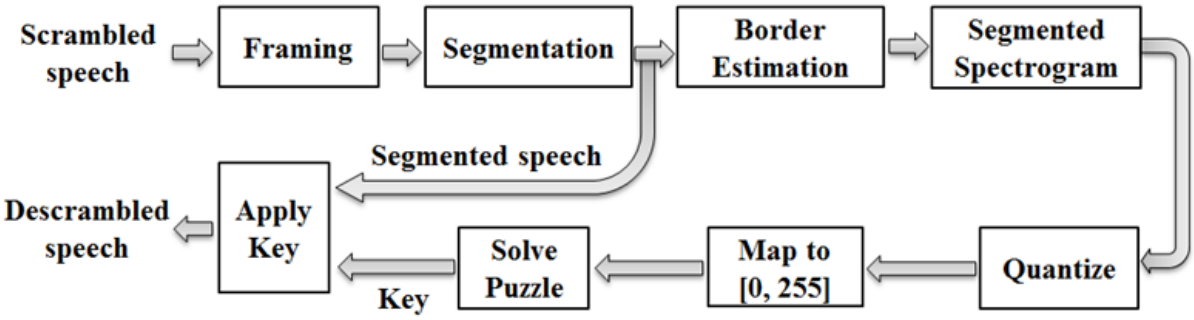

Figure 4. Block diagram of the proposed cryptanalysis method.

ble arrangements should be searched efficiently. This problem is a combinatorial problem and leads to an optimization task.

2.4.1 Distance Function

To find the best arrangement of puzzle pieces, it is necessary to formulate a measure of correctness. This task can be accomplished by comparing the shape, colour, texture, and etc of border pixels. Let $Ii(x, y)$ be the spectrogram image of an audio segment with size of $X \times Y$ pixels. A suitable distance function between I_1 and I_2 is the root mean square (RMS) value of differences between their border pixels. In the previous section we showed that an ideal estimation removes discontinuity of border in the segmented spectrogram of true neighbours. Unfortunately, such ideal algorithm does not exist in practice; thus, even for true neighbours, a small discrepancy between the borders would remain. To mitigate this problem, we let image I_1 to slide along image I_2 vertically for small number of pixels ($\beta \in \{0, 1, \ldots, 7\}$). Also, we let two pieces penetrate into each other for small number of pixels ($\alpha \in \{0, 1, 2, 3\}$). This concept is illustrated in Figure 5. We use the lowest distance as an estimation of true distance between two pieces I_1 and I_2.

$$D = \min_{\alpha,\beta}\{\sqrt{\sum_{y=1}^{y=Y-\beta}(I_1(X-\alpha, y+\beta) - I_2(\alpha, y))^2}\} \tag{11}$$

Also, Equation 11 and Figure 5 consider only when I_2 is slid upward along I_1 piece. The actual distance was calculated as the minimum value of both upward and downward sliding. Figure 6 represents a better insight about the proposed distance metric.

2.4.2 Solving Puzzle Algorithm

Distance function of Equation 11 was employed to calculate total distance for a possible arrangement of pieces. This cost is equal to:

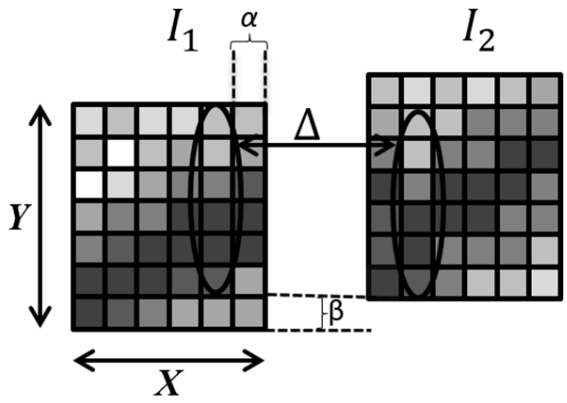

Figure 5. For caculating distance between two images, they are slid horizontaly and vertically untill the best match is found.

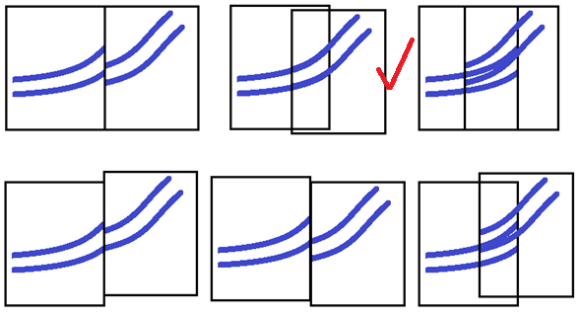

Figure 6. Two pieces of the puzzle are slid along each other, and penetrated until the best match is found.

$$Dist = \sum_{i=1}^{N-1} D(I_i, I_{i+1}) \tag{12}$$

Now, solving the jigsaw can be expressed as an optimization problem [36]:

$$S = min\{Dist\}_{\Omega} \tag{13}$$

where, N is the frame size and Ω is the space of all possible arrangements of all pieces.

Let B and L denote duration of scrambled speech (in minutes) and segment length, respectively. If the scrambling key is changed for each frame, Equation 14 calculates cardinality of the search space.

$$|\Omega| = \frac{B \times 60}{L \times N} \times N! \tag{14}$$

To put the size of key space ($|\Omega|$) into a better perspective, this value was calculated for B=5 minutes, L=40ms, and different frame sizes (N). The result is presented in Table 1.

Table 1. Size of key space

N	$\|\Omega\|$(bit)
6	20
8	26
10	32
12	39
14	46

According to Table 1 the size of key space is very huge; thus, finding the accurate answer in an efficient manner is very crucial for this method to work. Reviewing literature shows that this task can be accomplished in different ways. For example, different metaheuristic algorithms were employed to solve this problem [36]. Unfortunately, these methods may get trapped in local minima of Equation 13. Recently, a method based on branch and bound (B&B) technique was proposed [35]. It was shown that this algorithm finds the global minima of the Equation 13with acceptable level of search complexity. This algorithm employs a tree for systematic and efficient search of key space. At each step, it selects one node and expands the tree from that node. Furthermore, it uses both upper and lower bounds to prune the tree in the most effective way. The best solution found so far (Opt) and the minimum weight arborescence of directed graph [37] were used as the upper and the lower bounds. A simplified schematic of this method is presented in Figure 7.

2.5 Estimating Segment Length

Initially, we assumed prior knowledge of segment length. In this section, we propose a novel metric called moving spectrogram distance (MSD) for accurate estimation of segment length in time scrambling systems. To that end, we use two consecutive moving windows with length of τ called cells. Then, we convert each cell into its spectrogram with window size of δ and maximum value of overlap ($\gamma = \delta - 1$). After that, we calculate the distance between these two spectrograms employing Equation 11. Finally, we move the cells one sample forward and repeat the same procedure until we reach the end of signal.

We show that MSD has periodic property and its period is equal to segment length. Let $x =$

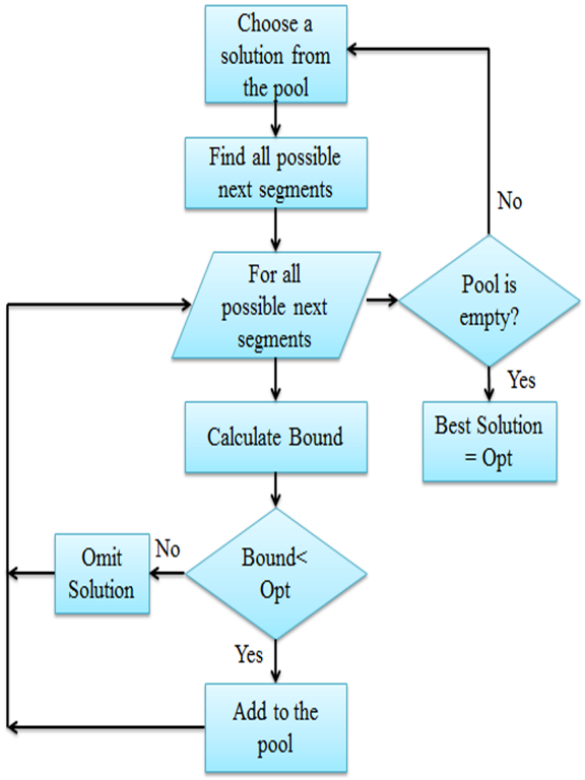

Figure 7. Solving jigsaw puzzle based on B&B method [35]

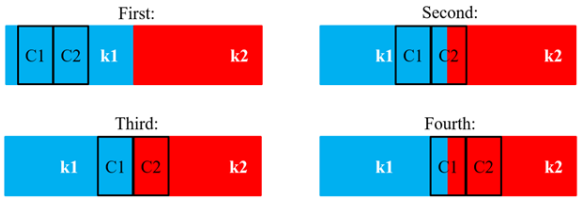

Figure 8. four different situations that could happen while calculating MSD

$[x_1, x_2, x_3, \dots]$, c_1, c_2 and L denote scrambled signal, first and second cell, and segment length of scrambler system, respectively. In this manner, c_1 and c_2 start from samples x_i and $x_{i+\tau}$ end at samples $x_{i+\tau-1}$ and $x_{i+2\tau-1}$, respectively. Also, let k_1 and k_2 denote segment index corresponding to sample i and the next segment of scrambled speech, respectively. If τ is selected such that $\tau < L/2$, as value of i is changing four different situations could happen.
First: both c_1 and c_2 are completely in segment k_1,
Second: c_1 is completely in k_1 but c_2 is both in k_1 and k_2,
Third: c_1 is completely in k_1 and c_2 is completely in k_2,
Fourth: c_1 is in both k_1 and k_2 and c_2 is completely in k_2.
Figure 8 shows these four situations. In the first case, both cells are residing in the same segment, so their distance would be low. In the second case, c_2 is gradually moving into another segment, so the dis-

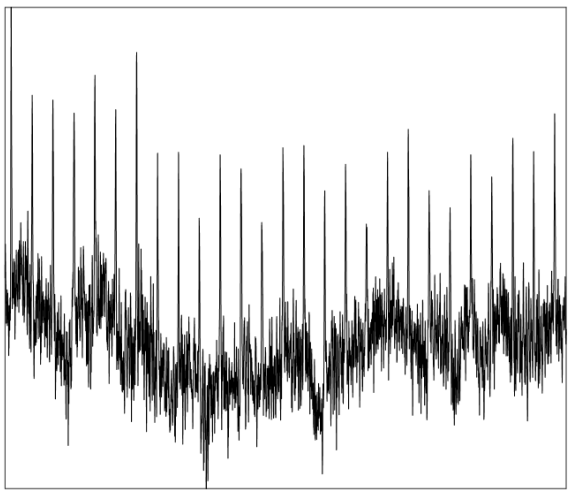

Figure 9. autocorrelation of MSD

tance would increase. In the third case, the distance would be maximum and finally in the last case the distance would decrease. Because, this pattern would repeat every L samples, we expect to see a periodic behaviour in the MSD.

To capture, this periodicity we used autocorrelation of MSD and found all of its maxima. Then we used distance between two consecutive maxima as the estimation of L. Figure 9 shows a portion of autocorrelation of MSD. The block diagram of the proposed method is presented in Figure 10.

3 Experimental Results

This section investigates efficacy of the proposed method. To that end, time domain hopping window scrambler was simulated using MATLAB. Table 2 presents parameters of the simulated system.

Table 2. Parameters of scrambling system

Parameter	value
Frame Size (N)	8
Segment Length (L)	40ms
Fs	8KHz

3.1 Performance Criteria

The core of the proposed method is solving the resulted puzzle. Any error in the solution of the puzzle is directly reflected into the descrambled speech. Our main objective is to achieve the highest value of intelligibility in the descrambled signal. Intelligibility is a subjective measure but intuitively we expect that if a puzzle is solved more accurately, its intelligibility would be higher. Therefore, for objective measurement of performance we have used the accuracy metric proposed in [35]. This metric progressively divides the found solution into different sub-blocks and compares them with sub-blocks of correct solution. In this manner, concurring between larger sub-blocks of found solution and correct solution will result in higher score. Let us denote sub-block that starts from position i in the correct solution and found solution with B_{Ci} and B_{Fi}, respectively. Now, define S_n as the total number of B_{Ci} and B_{Fi} that are the same:

$$S_n = \sum_{\forall i} \sum_{\forall j} \delta(B_{Fi} = B_{Cj}) \tag{15}$$

where δ is Dirac delta function. Finally, accuracy was calculated as the weighted sum of S_n for all possible values of n [35].

$$Ac = \frac{\sum_{n=1}^{N} S_n \times n}{\sum_{n=1}^{N} (N - n + 1) \times n} \tag{16}$$

According to Equation 16, accuracy is calculated as the weighted sum of the number of sub-blocks that are correct. That is, for blocks with larger size (larger values of n) the numerator is multiplied with a larger number; therefore, correct larger blocks result in higher values of accuracy. Finally, denominator is for normalizing the value of score between 0 and 1.

3.2 Tests Methodology

The tests were conducted on two different databases. For the first database, 1000 utterances were selected randomly from TIMIT [38] database. These utterances were divided into two sets of 400 and 600 excerpts for parameter optimization and conducting objective evaluation. For the second database, we used seven Persian speech utterances with total duration of 90 seconds. They included both male and female speakers and they were recorded in a low-noise room with sampling frequency of 44100 Hz and resolution of 16 bits. The second database was used for conducting the subjective evaluation of intelligibility. It is noteworthy that, both databases were subjected to the following processes. All excerpts were downsampled from their original frequency to 8000 Hz. Also, silent portions of the speech have no intelligibility; therefore, a simple energy based voice activity detection (VAD) method was employed to detect and remove such portions. To that end, signals were segmented into chunks of 50ms duration with 25ms of overlaps. Then, energy of all segments was calculated. If energy of a segment was lower than a threshold it was considered as a silent one and therefore it was removed. The threshold was defined as the minimum of energy of all segments plus α% of difference between

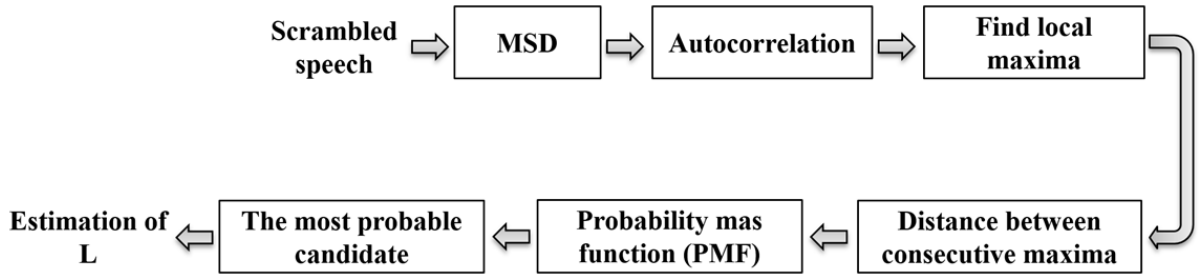

Figure 10. Estimating segment length of scrambling system

maximum and minimum of energy of all segments. We set $\alpha = 1$ for this purpose.

Scrambled excerpts were produced by feeding each signal into simulated scrambling system of Table 2. Also, a different permutation key was used for each frame. Finally, the produced cipher voices were descrambled with different methods. To show efficacy of the proposed method, two previously proposed methods of frequency weighted log spectral distance (FWLSD) and basic puzzle were also implemented. FWLSD method was proposed in [26]. This method calculates a spectral distance between the first and the last K samples of two consecutive speech segments. To that end, the whole spectrum is divided into N_f frequency bands and linear prediction coefficient (LPC) of each band is calculated. Unfortunately, work of [26] did not report the optimum values of these parameters; therefore, we have adopted parameters reported in [30] for this task. We proposed our basic puzzle method in [30]. Table 3 provides optimum values of parameters for FWLSD and basic puzzle methods.

Table 3. Parameters of scrambling system

	parameter	value
FWLSD	K	80
	Sub-band No.	44
	LPC order	16
Basic Puzzle	Win size	55
	Overlap value	54

3.3 Optimizing Parameters

There are some parameters associated with the proposed method that need to be optimized. STFT has two important parameters: window size and length of overlap. Also, filter order and forgetting factor in the RLS estimation should be determined. An exhaustive search over these four parameters was conducted on the first set of files from TIMIT to achieve the highest value of accuracy. The optimized parameters are shown in Table 4. According to Table 4 the best results are achieved when STFT is used with window size of 60 and a high value of overlap. Also, the estimation system uses 52 samples for estimating the new sample and it uses a high value for forgetting factor.

Table 4. The Optimum value of parameters

	parameter	value
STFT	Win size	60
	Overlap value	51
RLS	Order	52
	Forgetting factor	0.97

3.4 Constructing Jigsaw Puzzles

Before descrambling the samples, their corresponding puzzles should be constructed. In Section 2.1 and 2.2 we presented three possible approaches of conventional spectrogram, segmented spectrogram, and estimated spectrogram for this purpose. We also argued that conventional spectrogram and estimated spectrogram would be the least and most efficient ways, respectively. To further justify that claim and put those arguments into better perspective, a set of tests were carried out. After scrambling all samples of second part of first database with simulated system, they were converted into their jigsaw puzzles with all of these three methods. Then, the distance between all pieces were calculated using metric of Equation 11. Finally, these distances were sorted and index of correct pieces were extracted. In the ideal case, index of the correct piece would always be 1. Probability mass function (PMF) of index of correct pieces are plotted in Figure 11. Figure 11 shows that if we use conventional spectrogram most of the time the minimum distance does not correspond with the correct answer. On the other hand, in the estimated spectrogram 87% of the times the correct piece had the minimum distance. It is noteworthy that these results agree with discussions of Section 2.1 and 2.2.

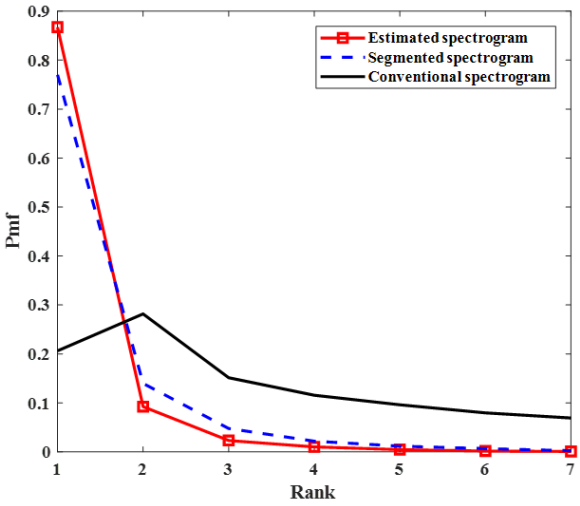

Figure 11. Probability mass function (PMF) of index of correct pieces for different approaches for converting scrambled speech into jigsaw puzzles

3.5 Intelligibility of Descrambled Speech

To measure efficacy of the proposed system, both accuracy of solving the puzzle and intelligibility of descrambled samples are measured. In the accuracy tests, the second set from TIMIT database is used. We use Equation 16 for this purpose. Table 5 compares average of these values for different methods.

Table 5. Accuracy of descrambled speeches

Method	Accuracy(mean±std)	Reference
Puzzle + RLS	87.8±23.0%	This work
Basic puzzle	73.9±31.8%	[30]
FWLSD	36.9±37.3%	[26]

According to Table 5, the proposed method achieves score of 87.8% while this value for puzzle without estimation and FWLSD methods are 73.9% and 36.9%, respectively.

For measuring actual values of intelligibility a set of subjective tests is conducted. To this end, all seven Persian excerpts are descrambled by basic puzzle [30], FWLSD [26], and the proposed method. This results in a total of 21 different samples. We devide these files into tree distinct sets, such that each set contains three excerpts descrambled with one method and two samples from each of the remaining methods. Also, every speech is used in each set only once. Then, each set is randomly given to five different persons and they are asked to listen to all samples and transcribe only the words that they can understand. With the help of data gathered from these fifteen people (6 females and 9 males), subjective intelligibility was calculated. First, gathered data is divided into three categories according to their method of descrambling. Then, subjective intelligibility of each method is calculated as the average value of correct transcription of words over that category. Table 6 presents the results.

Table 6. Subjective Intelligibility of descrambled speeches

Method	Accuracy(mean±std)	Reference
Puzzle + RLS	92.9±4.5%	This work
Basic puzzle	83.5±10.3%	[30]
FWLSD	58.3±14.2%	[26]

According to Table 6 on average, 92.9% of the words descrambled with the proposed method are intelligible. On the other hand, only 83.5% and 58.3% of the words descrambled with puzzle without estimation and FWLSD methods are intelligible.

3.6 System Parameters and Accuracy of Descrambled Speech

To measure effect of scrambler parameters on performance of different descrambling methods, a set of tests is carried out. These tests can demonstrate limitations and potency of each method. In the first test, frame size (N) of scrambler is varied and then scrambled samples are descrambled with different methods. Finally, accuracy of descrambled samples is measured. In the second test, a scrambler with 8 segments per frame with different segment lengths (L) is simulated. After descrambling samples with different methods, their accuracies are measured. Figure 12 shows results of these simulations.

According to Figure 12.A when the frame size increases from 6 to 16, accuracy of RLS+puzzle, Puzzle, and FWLSD methods drops to 0.7, 0.48, and 0.23. Also, accuracy of FWLSD method decreases faster than RLS+puzzle method. These observations justify that estimation improves performance of the system considerably. Referring to Figure 12.B, we observe that the proposed method (RLS+puzzle) with segment length of 10ms can provide better accuracy than FWLSD method with segment length of 40ms. The same comparison can be made between RLS+puzzle and puzzle methods.

Finally, let 0.5 be the minimum level of acceptable accuracy. Figure 12.A shows that FWLSD method is acceptable for very small frame sizes ($N<6$). On the other hand, the proposed method remains in the acceptable range even for big frame sizes ($N>14$). Also, Figure 12.B shows that FWLSD needs large amount of data ($L>40ms$) to achieve an acceptable level of accuracy. On the other hand, RLS+puzzle needs only small amount of data ($L\geq 10ms$) to achieve an accept-

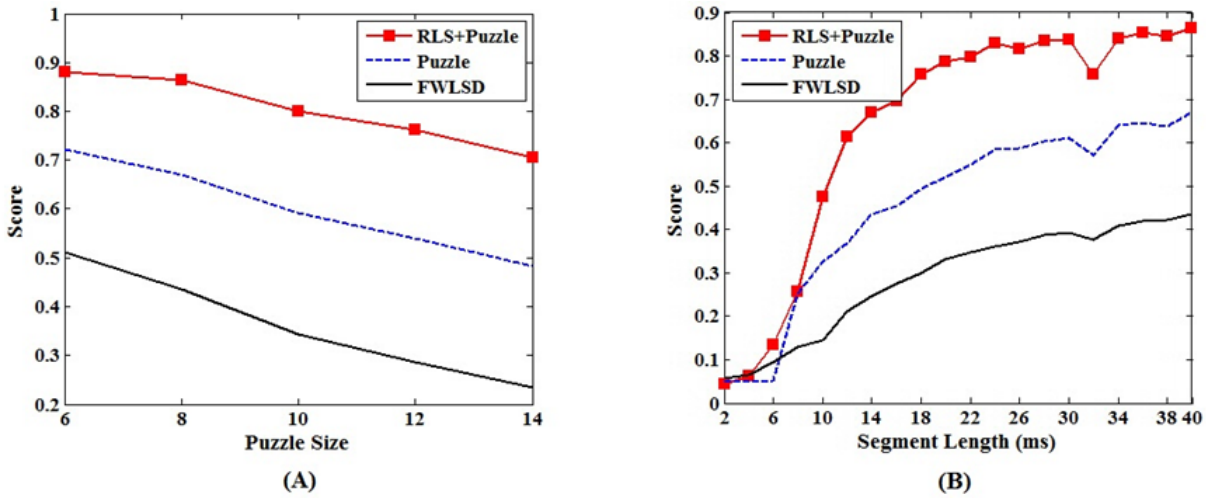

Figure 12. Effect of scrambler parameters on accuracy of descrambled speech. A: Effect of frame size, B: Effect of segment length

able level of accuracy. Based on these observations we can conclude that the proposed method (RLS+puzzle) has fewer limitations than other methods.

3.7 Effect of Noise on Accuracy of Descrambled Speech

In practical situations different types of noise are added to the voice. These noises may be categorized into two different types. First, source noises that mix with the voice before it is scrambled. These noises represent ambient noise and noise of the scrambler system. Second, channel noise mixes with the scrambled voice as it is transmitted through the communication channel. To measure the effect of both noises, a set of tests was carried out. In this regard, noise was modelled as an additive white Gaussian noise. Figure 13 shows accuracy of different methods for different power of noises. Comparing Figure 13.A and Figure 13.B shows that the effect of the channel and the source noise is equal on accuracy of descrambled speech. We know that white noise is a stationary random process. Therefore, its statistics remains constant over the time and hence both noises would have the same effect on the accuracy of descrambled speeches. Comparing results of different methods demonstrates that accuracy of the proposed method (RLS+puzzle) with signal to noise ratio (SNR) of 10db is almost equal to the FWLSD method with SNR of 40db. Also, if 0.5 is the minimum level of acceptable accuracy, then Figure 13.A shows that accuracy of FWLSD method is acceptable only at very high SNRs (SNR$\geq$ 40db). On the other hand, the proposed method remains in the acceptable range even for low values of SNRs (SNR$\geq$ 15db). Based on these observations we can conclude that the proposed method has less limitations than other methods.

3.8 Performance of Segment Length Estimation

First, there are some parameters associated with MSD that need to be optimized. These parameters include cell size (τ) and window size of STFT (δ). To that end, a system with segment length of 40ms was simulated. Then, we use value of probability mass function (PMF) at the correct segment length (i.e., 320 samples) as an indication of performance of the system. Result of this analysis is shown in Figure 14.A. Based on these results the best performance was achieved for $\tau = 26$, $\delta = 13$. Plot of PMF for these optimum parameters is shown in Figure 14.B.

In order to measure performance of the proposed system, different scrambler systems with segment lengths of 20ms, 25ms, 30ms, 35ms, 40ms, 45ms, and 50ms were simulated. We used system depicted in Figure 10 for estimating segment length. Figure 15 shows the results. Referring to Figure 15, it is evident that in all tests, the correct segment length always corresponds to the maximum of PMF. Therefore, the proposed method can accurately find the correct segment length.

4 Discussion

Audio signals have high amount of redundancies. These redundancies could be exploited for cryptanalysis of scrambling systems. According to Figure 1 these redundancies are reflected in the spectrogram very well. Comparing spectrogram of clear and cipher samples shows that, there is no abrupt transition in the spectrogram of clear samples and they are smooth images in both dimensions. Based on these observations, cryptanalysis of audio scrambling system was formulated as a puzzle solving problem. The first step of this method was creating the jigsaw puzzles from

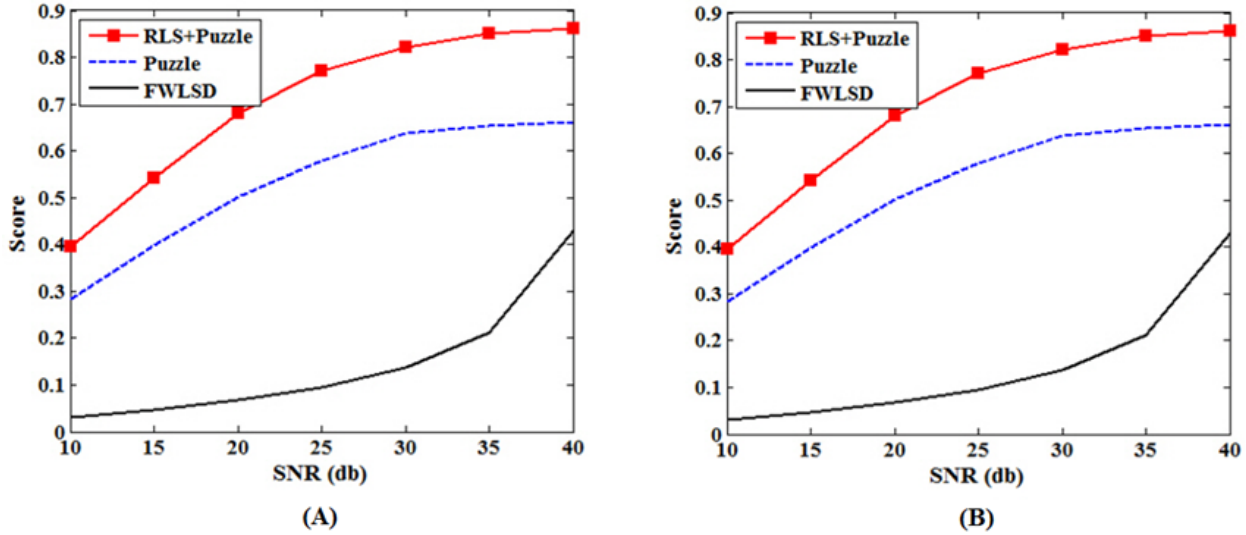

Figure 13. Effect of noise on accuracy of descrambled speech. A: Effect of source noise, B: Effect of channel noise

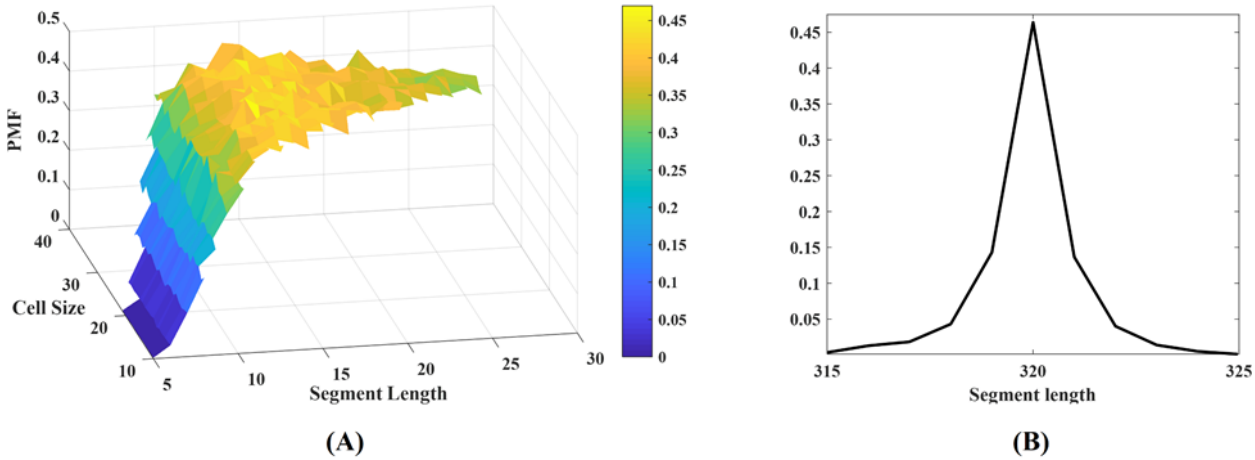

Figure 14. Optimizing parameters of MSD. A: PMF for different values of parametrs, B: PMF for $\tau = 26$, $\delta = 13$

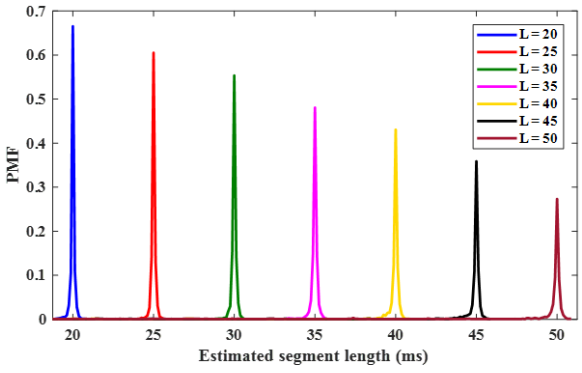

Figure 15. Results of the proposed method for scramblers with different segment lengths

scrambled samples. To that end, it was argued that STFT should be applied on each segment separately. Later, it was shown that if segmented spectrogram is used, some discrepancies between borders of true pieces exist. These discrepancies were the result of different values of effectiveness in the border samples. We incorporated concept of estimation in the system to ameliorate these discrepancies. Furthermore, to mitigate effect of non-ideal estimation, distance function was modified to further alleviate effect of any remaining discrepancies. In this regard, the selected pieces were slid in front of each other to find their best match.

Our experiments showed that the proposed method has better performance. According to Table 5 in the objective test, the proposed method achieves accuracy of 87.8% while this score is 73.9% and 36.9% for puzzle without estimation and FWLSD methods. Furthermore, accuracy of the proposed method has lower variances; thus, not only it descrambles the samples more accurately, but also its accuracy does not deviate very much. We believe that higher values of accuracy in the puzzle-based methods are direct consequence of STFT transform. Investigating formula of

FWLSD shows that, it only extracts frequency properties of different sub-bands of audio signal. On the other hand, spectrogram transform reveals both time domain and frequency domain redundancies of the signal. In other words, proposed method exploited redundancies of the speech signal more efficiently.

Comparing Table 5 and 6 shows that intelligibility is higher than actual accuracy. Conducted accuracy tests had two important properties. First, it considered one frame at a time. Second, only orders of the segments were investigated. However, in the subjective tests both ear and brain plays important roles. For example, post-masking and pre-masking properties of human auditory system [39] may mask some of the erroneous patterns. Furthermore, brain knows structure of the language and can easily predict and guess portions of sentences in everyday life. Thus, it is very likely that brain of test subjects have used its ability to fill some of the gaps and to find the most suitable words. Therefore, higher value of subjective intelligibility is justified.

Finally, if 0.5 is defined as the minimum level of acceptable accuracy, then according to results of Figure 8 and Figure 9 the proposed method has the following limitations. In the proposed method, segment length should be larger than 10ms and SNR of scrambled speeches should be at least 15db. It is noteworthy that these parameters should be higher than 40ms and 40db in the FWLSD method. As the future work we are working to extend this method to other scrambling techniques.

5 Conclusion

This paper addressed the absence of a through security analysis of scrambling systems. To that end, security of hopping window time domain scrambler was investigated. It was shown that high redundancies of audio signals were reflected in their spectrograms very efficiently. Based on this transformation, cryptanalysis problem was transformed into solving rectangular jigsaw puzzles. Furthermore, integrating estimation technique into the system improved results of the system. Finally, the proposed method was compared with FWLSD which, to the best of our knowledge, is the only existing cryptanalysis method for these systems. Both subjective and objective tests were carried out to compare performance of different methods. Objective and subjective tests showed that the proposed method achieved accuracy and intelligibility of 87.8% and 92.9%, respectively. These scores were 50.9% and 34.6% higher than scores achieved by FWLSD method. Finally, a novel method based on moving spectrogram distance was proposed that gave accurate estimation of segment length of the scrambler system.

References

[1] H. Ghasemzadeh and M. H. Keyvanrad, "Toward a Robust and Secure Echo Steganography Method Based on Parameters Hopping," in Signal Processing and Intelligent Systems, 2015.

[2] H. Ghasemzadeh, M. T. Khass, and M. K. Arjmandi, "Audio steganalysis based on reversed psychoacoustic model of human hearing," Digital signal processing, vol. 51, pp. 133-141, 2016.

[3] S. Sridharan, E. Dawson, and B. Goldburg, "Fast Fourier transform based speech encryption system," IEE Proceedings I (Communications, Speech and Vision), vol. 138, pp. 215-223, 1991.

[4] M. A. Brandau, "Implementation of a real-time voice encryption system," 2008.

[5] J. Guo, J.-C. Yen, and H.-F. Pai, "New voice over Internet protocol technique with hierarchical data security protection," in Vision, Image and Signal Processing, IEE Proceedings-, 2002, pp. 237-243.

[6] V. enk, V. Deli, and V. Miloevi, "A new speech scrambling concept based on Hadamard matrices," Signal Processing Letters, IEEE, vol. 4, pp. 161-163, 1997.

[7] K. Sakurai, K. Koga, and T. Muratani, "A speech scrambler using the fast Fourier transform technique," Selected Areas in Communications, IEEE Journal on, vol. 2, pp. 434-442, 1984.

[8] A. Matsunaga, K. Koga, and M. Ohkawa, "An analog speech scrambling system using the FFT technique with high-level security," Selected Areas in Communications, IEEE Journal on, vol. 7, pp. 540-547, 1989.

[9] E. Dawson, "Design of a discrete cosine transform based speech scrambler," Electronics letters, vol. 27, pp. 613-614, 1991.

[10] S. Sridharan, E. Dawson, and B. Goldburg, "Speech encryption in the transform domain," Electronics Letters, vol. 26, pp. 655-657, 1990.

[11] B. Goldburg, S. Sridharan, and E. Dawson, "Design and cryptanalysis of transform-based analog speech scramblers," Selected Areas in Communications, IEEE Journal on, vol. 11, pp. 735-744, 1993.

[12] A. Jameel, M. Y. Siyal, and N. Ahmed, "Transform-domain and DSP based secure speech communication system," Microprocessors and Microsystems, vol. 31, pp. 335-346, 2007.

[13] A. S. Bopardikar, "Speech encryption using wavelet packets," 2005.

[14] S. Sadkhan, N. Abdulmuhsen, and N. F. Al-Tahan, "A proposed analog speech scrambler based on parallel structure of wavelet transforms," in Radio Science Conference, 2007. NRSC 2007. National, 2007, pp. 1-12.

[15] D. Tseng and J. Chiu, "An OFDM speech scrambler without residual intelligibility," in TENCON 2007-2007 IEEE Region 10 Conference, 2007, pp. 1-4.
[16] H. Li, Z. Qin, L. Shao, and B. Wang, "A novel audio scrambling algorithm in variable dimension space," in Advanced Communication Technology, 2009. ICACT 2009. 11th International Conference on, 2009, pp. 1647-1651.
[17] Q.-H. Lin, F.-L. Yin, T.-M. Mei, and H. Liang, "A blind source separation based method for speech encryption," Circuits and Systems I: Regular Papers, IEEE Transactions on, vol. 53, pp. 1320-1328, 2006.
[18] W.-Q. Yan, W.-G. Fu, and M. S. Kankanhalli, "Progressive audio scrambling in compressed domain," Multimedia, IEEE Transactions on, vol. 10, pp. 960-968, 2008.
[19] C.-P. Wu and C. J. Kuo, "Design of integrated multimedia compression and encryption systems," Multimedia, IEEE Transactions on, vol. 7, pp. 828-839, 2005.
[20] L. Zeng, X. Zhang, L. Chen, Z. Fan, and Y. Wang, "Scrambling-based speech encryption via compressed sensing," EURASIP Journal on Advances in Signal Processing, vol. 2012, pp. 1-12, 2012.
[21] S. James, S. George, and P. Deepthi, "Secure selective encryption of compressed audio," in Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy (AICERA/ICMiCR), 2013 Annual International Conference on, 2013, pp. 1-6.
[22] Z. Su, J. Jiang, S. Lian, G. Zhang, and D. Hu, "Hierarchical selective encryption for G. 729 speech based on bit sensitivity," Journal of Internet Technology, vol. 11, pp. 599-607, 2010.
[23] G.-R. Kwon, C. Wang, S. Lian, and S.-s. Hwang, "Advanced partial encryption using watermarking and scrambling in MP3," Multimedia Tools and Applications, vol. 59, pp. 885-895, 2012.
[24] K. Datta and I. S. Gupta, "Partial encryption and watermarking scheme for audio files with controlled degradation of quality," Multimedia tools and applications, vol. 64, pp. 649-669, 2013.
[25] B. Goldburg, S. Sridharan, and E. Dawson, "Cryptanalysis of frequency domain analogue speech scramblers," IEE Proceedings I (Communications, Speech and Vision), vol. 140, pp. 235-239, 1993.
[26] B. Goldburg, E. Dawson, and S. Sridharan, "The automated cryptanalysis of analog speech scramblers," in Advances in CryptologyEUROCRYPT91, 1991, pp. 422-430.
[27] Y.-X. Zhao, M.-C. Su, Z.-L. Chou, and J. Lee, "A puzzle solver and its application in speech descrambling," in Proc. 2007 WSEAS Int. Conf. Computer Engineering and Applications, 2007, pp. 171-176.
[28] A. Jolfaei, X.-W. Wu, and V. Muthukkumarasamy, "On the Security of Permutation-Only Image Encryption Schemes," Information Forensics and Security, IEEE Transactions on, vol. 11, pp. 235-246, 2016.
[29] H. C. Van Tilborg and S. Jajodia, Encyclopedia of cryptography and security: Springer Science & Business Media, 2014.
[30] H. Ghasemzadeh, H. Mehrara, and M. Tajik Khas, "Cipher-text only attack on hopping window time domain scramblers," in 4th. International Conference on Computer and Knowledge Engineering (iccke2014), 2014.
[31] K. Grchenig, Foundations of time-frequency analysis: Springer Science & Business Media, 2013.
[32] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," The Journal of the Acoustical Society of America, vol. 50, pp. 637-655, 1971.
[33] S. S. Haykin, Adaptive filter theory: Pearson Education India, 2008.
[34] D. Kahn, The Codebreakers: The comprehensive history of secret communication from ancient times to the internet: Simon and Schuster, 1996.
[35] H. Ghasemzadeh and M. Khalil Arjmandi, "Optimum solution and evaluation of rectangular jigsaw puzzles based on branch and bound method and combinatorial accuracy," Multimedia Tools and Applications, pp. 1-25, 2017.
[36] H. Ghasemzadeh, "A metaheuristic approach for solving jigsaw puzzles," in Intelligent Systems (ICIS), 2014 Iranian Conference on, 2014, pp. 1-6.
[37] J.-C. Fournier, Graphs Theory and Applications: With Exercises and Problems vol. 72: John Wiley & Sons, 2010.
[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Technical Report N, vol. 93, 1993.
[39] E. Zwicker and H. Fastl, Psychoacoustics: facts and models: Springer Science & Business Media, 2013.

Hamzeh Ghasemzadeh was born in Tehran in 1984. He received his B.S. degree in Electrical Engineering from Ferdowsi University of Mashhad in 2007. He received his M.S. degree in Communications Engineering from Malek-e-Ashtar University of Technology in 2011. He was an adjunct professor at department of Electrical Engineering at Islamic Azad University until 2016. He is now pursuing a dual Ph.D. degree in both "Communication Sciences and Disorders" and "Computational Mathematics Science and Engineering" at Michigan State University. He has been working on different aspects of audio signals, covering acoustic analysis of pathological impaired voices to security driven applications of audio signals. His primary research interests are applying statistical signal processing and machine learning techniques for solving different speech/voice related problems. Right now, he is working on statistical signal processing and data mining on videos acquired through high speed video endoscopy (HSV) from vocal folds.

Mehdi Tajik Khass was born in 1985 in Tehran, Iran. He got his B.S. from the University of Tabriz and M.S. degree from Malek-e-Ashtar University of Technology both in Telecommunication Engineering. His research interests are in signal and image processing, cryptography, human voice processing, and bio-engineering especially in research regarding human auditory system.

Hamed Mehrara received his B.S. degrees in electrical engineering from Shahrood University of Technology, Shahrood, Iran, in 2008 and M.S. and Ph.D. degrees in electrical engineering from Malek Ashtar University of Technology, Tehran, Iran, in 2012 and 2017, respectively. His research interests include wide band imaging systems: fabrication and processing.