Data in brief 26 (2019) 104537

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Encrypted audio dataset based on the Collatz conjecture

Diego Renza*, Sebastian Mendoza, Dora M. Ballesteros L

Universidad Militar Nueva Granada, Colombia

ARTICLE INFO

Article history:
Received 30 April 2019
Received in revised form 27 August 2019
Accepted 9 September 2019
Available online 17 September 2019

Keywords:
Audio
Encryption
Security
Cryptanalysis
Privacy
Collatz conjecture
Dataset

ABSTRACT

In information security, one way to keep a secret content is through encryption. The objective is to alter the content so that it is not intelligible, and therefore only the intended user can reveal the secret content. With the aim to provide examples of encrypted audio data, we applied a novel method of encryption based on the Collatz conjecture in five hundred speech recordings (50 speakers, 10 different messages), and then five hundred encrypted audio files were obtained. The main characteristics of our encrypted recordings are as follows: the spectrogram is quasi-uniform, histograms have a repetitive pattern, average of samples is around −0.4, standard deviation is around 0.55; Shannon entropy is around 7.5 (for 8-bits per sample). The novelty of the results consists in obtaining a completely different behavior than natural speech recordings, i.e.: spectrogram with higher energy in low frequencies, histogram with Gaussian behavior, average of samples around 0, standard deviation around 0.11, entropy around 5.5. A more comprehensive analysis of our encrypted signals may be obtained from the article "*High-uncertainty audio signal encryption based on the Collatz conjecture*" in the Journal of Information Security and Applications.

* Corresponding author.
E-mail address: diego.renza@unimilitar.edu.co (D. Renza).

https://doi.org/10.1016/j.dib.2019.104537

Specifications Table

Subject area	*Computer science*
More specific subject area	*Speech processing; security*
Type of data	*Audio files and a spreadsheet*
How data was acquired	*The encrypted audio files were obtained through the Matlab implementation of the algorithm proposed in Ref.* [2]
Data format	*Raw: encrypted audio files (wav)* *Analyzed data (xlsx)*
Experimental factors	*50 speakers, 10 different messages*
Experimental features	*All encrypted data were obtained with the same encryption process*
Data source location	*Colombia*
Data accessibility	*Repository name: Mendeley* *Data name: Encrypted audio files* [1] *Direct URL to data:* https://doi.org/10.17632/3vwwv3xhhc.1
Related research article	*D. Renza, S. Mendoza, D.M. Ballesteros L., High-uncertainty audio signal encryption based on the Collatz conjecture, Journal of Information Security and Applications, Volume 46, 2019, Pages 62–69* [2]

Value of the Data

- This data set can be used for cryptanalysis purposes in order to try to break the encryption method proposed in Ref. [2].
- It is useful for comparing the quality of encrypted audios in terms of their statistics like average, standard deviation, kurtosis, and entropy. Our encrypted audio files have the following values: average around −0.4, standard deviation around 0.55, kurtosis around 2.03, and entropy of 7.5 (for 8 bits).
- In addition, it can be used to compare the behavior of the encrypted audio signal in terms of its spectrogram (quasi-uniform) and histogram (repetitive pattern).

1. Data

The shared data contain 500 audio files that have been encrypted using the algorithm proposed in Ref. [2]. The original audio recordings are 500 audio files corresponding to 50 speakers and 10 different messages per speaker. The encrypted files are new audio signals with unintelligible content that can be used to test cryptanalysis techniques; they have a length in the range [85 295] s, a sampling frequency of 8 kHz and 32 bits/sample (128 kbps).

Statistical analysis of the audio encrypted data are provided in Fig. 1 to Fig. 4, using radial plots. Fig. 1 shows the average, Fig. 2 the standard deviation, Fig. 3 the kurtosis and Fig. 4 the entropy.

2. Experimental design, materials, and methods

The encrypted audio data were obtained from 500 speech recordings using the method presented in Ref. [2]. The statistical analysis of the 500 encrypted audio files is provided in the file titled *Encrypted audios.rar*.

The average, μ, is calculated with the equation $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, where x_i are the audio samples (original or encrypted) and n is its number of samples. Fig. 1 shows the results of average for every group of audio file (i.e. 500 original audios, 500 encrypted audios). According to Fig. 1, the average of the original audios is around 0, but for the encrypted audio files it is around −0.4. The difference in this statistical metric between the original and its encrypted audio using the Collatz Conjecture is remarkable.

Standard deviation (σ) is obtained as $\sigma = \sqrt{\sum_{i=1}^{n} |x_i - \mu|^2/(n-1)}$. Fig. 2 shows the results for this parameter. It is remarkable that σ values of the original audio signals are not nearly constant, but they are for the encrypted audios.

Kurtosis is obtained through the equation, $k = E(x-\mu)^4/\sigma^4$, where *E(.)* represents the expected value of data. Fig. 3 shows the results of original recordings and their encrypted files. Again, the behavior between these two groups is completely different.

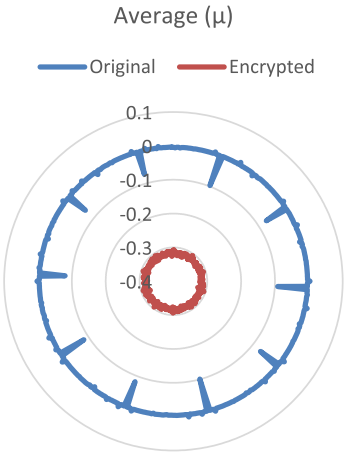

Fig. 1. Radial plot of the mean. The blue line corresponds to the original audio files, the red line to the encrypted data.

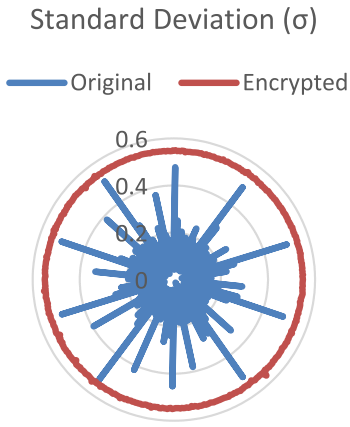

Fig. 2. Radial plot of the standard deviation. The blue line corresponds to the original audio data, the red line to the encrypted data.

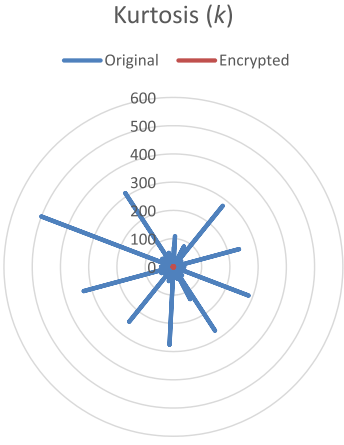

Fig. 3. Radial plot of the kurtosis. The blue line corresponds to the original audio files; the red line to the encrypted data.

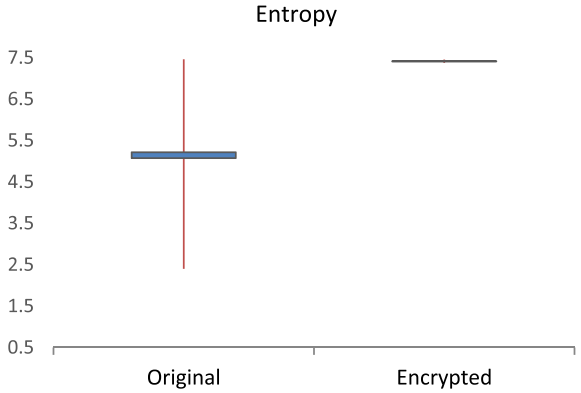

Fig. 4. Confidence plot (95%). The left part corresponds to the entropy of the original data; the right part to the encrypted data. These values are associated to the 500 audio files of this dataset.

In terms of the Shannon entropy, H(x), is obtained as $H(x) = -\sum_i^n P(x_i) log_2(P(x_i))$, where P (.) is the probability of data. For data with uniform distribution, i.e. where all values are equally likely, the expected entropy value is equal to the number of bits per sample [3]. Otherwise, entropy decreases. Fig. 4 shows the entropy comparison.

Given that, the encrypted audio files have 8 bits per sample, the theoretical highest value of entropy is 8. According to Fig. 4., the entropy of the encrypted data is around 7.5, whereas for natural audio signals it is around 5.5, in most of the cases. The entropy obtained in the encrypted audio files suggests that the level of uncertainty is very close to the highest possible.

With the encrypted recordings, the histogram and spectrogram can be obtained. If you use Matlab, the following code can help you to plot the figures:

```
[y,Fs] = audioread('name.wav'); % read the encrypted audio.
h = histogram(y, 256); % histogram of the encrypted audio with 256 bins.
s = spectrogram(x); % obtain the spectrogram of the encrypted audio.
spectrogram(y,'yaxis') % plot the spectrogram
```

Funding sources

This work was funded by the “Universidad Militar Nueva Granada - Vicerrectoría de Investigaciones” under the grant IMP-ING-2936.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Renza, S. Mendoza, D.M. Ballesteros L, Encrypted audio files, Mendeley Data 1 (2018). https://doi.org/10.17632/3vwwv3xhhc.1.
[2] D. Renza, S. Mendoza, M. Dora, L. Ballesteros, High-uncertainty audio signal encryption based on the Collatz conjecture, J. Inform. Sec. Appl. 46 (2019) 62–69. https://doi.org/10.1016/j.jisa.2019.02.010.
[3] D.W. Robison, Entropy and uncertainty, Entropy 10 (4) (2008) 493–506. https://doi.org/10.3390/e10040493.