Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5, 256-270
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i5.1684

Deep autoencoder with gated convolutional neural networks for improving speech quality in secured communications

Hilman F. Pardede[1], Kalamullah Ramli[2*], Nur Hayati[3], Diyanatul Husna[2], Magfirawaty[4]
[1]National Research and Innovation Agency, Research Center for AI and Cybersecurity, Indonesia, hilm003@brin.go.id (H.F.P).
[2]Universitas Indonesia, Indonesia, Department of Electrical Engineering, k.ramli@eng.ui.ac.id (K.H), diyanatulhusna@ui.ac.id (D.H)
[3]Universitas Muhammadiyah Yogyakarta, Department of Electrical Engineering, Indonesia, nuha.nurhayati@umy.ac.id (N.H)
[4]Politeknik Siber dan Sandi Negara, Cryptographic Hardware Engineering Department, Indonesia, magfirawaty@poltekssn.ac.id (M).

Abstract: In this study, we introduce a speech enhancement method to improve the quality of decrypted speech signals from hand-talk devices, which are highly susceptible to security attacks. Ensuring high-quality decrypted speech is essential because traditional speech enhancement methods struggle with artifacts only present during speech due to the encryption process applied selectively. This situation limits the effectiveness of traditional methods, which assume distortion is constant and can be estimated during silent periods. Our solution involves a deep-learning approach that employs a gated convolutional neural network (GCNN). Unlike typical convolutional neural networks (CNNs) that excel in processing spatial data but falter with temporal changes, our GCNN integrates a gating mechanism to enhance handling of temporal dynamics in speech data. This method directly maps distorted speech to its clean counterpart, bypassing the need for explicit noise estimation. Our experiments indicate that this deep-learning method significantly outperforms traditional speech enhancement techniques and conventional CNNs in several key evaluation metrics, offering a promising advancement in decrypted speech quality enhancement.

Keywords: *Convolutional neural networks, Gated convolutional neural networks, Gating mechanism, Secured speech communication, Speech enhancement.*

1. Introduction

Speech communications have become more and more widely used, but they are vulnerable to unauthorized disclosure [1, 2] Therefore, providing secure speech communications is increasingly important. As such, a plethora of speech encryption techniques have been introduced. Unfortunately, the encryption process, which includes a randomization process, may leave artifacts when the signal is decrypted, which can distort speech and significantly degrade the quality of speech signals [3, 4]. In this study, our focus was on improving the quality of decrypted speech from hand-talk (or walkie-talkie) devices. For these devices, encryption is applied when speech is present, and as a consequence, the artifacts in the decryption process only exists when speech is present.

There have been many studies in the literature that aimed to develop encryption methods that are capable of preserving speech quality and intelligibility or at least producing speech with an acceptable quality. A fast Fourier transform combined with multiple chaotic maps was applied in [5], and improvements were shown in several subjective and objective metrics of speech. Chaotic encryption [6, 7] is a popular encryption method for speech, and studies have shown that it is able to maintain the quality and intelligibility of decrypted speech at some levels. However, in these approaches, the

* Correspondence: kalamullah.ramli@ui.ac.id

strategies for improving the quality of speech signals are very dependent on the communication frameworks and require some knowledge of the communication channels, which is not always available.

The other way to improve the quality of speech signals is by applying speech enhancement to the decrypted speech. Speech enhancement is a research field that aims to remove the distortions in speech signals and, hence, improve their quality and intelligibility. Applying speech enhancement improves the modularity of secure communication systems, as it may not require any knowledge of the communication channels. Before the emergence of deep learning technologies, various smoothing strategies were applied to speech signals to suppress noise. Blind source separation was implemented in [8–10], and it was shown that it was able to produce good-quality decrypted speech. Spectral subtraction [11], Wiener filtering [12, 13], and Kalman filtering [14] have been applied to speech enhancement with some success. In our previous study, we employed several traditional smoothing based speech enhancement methods to improve the quality of speech signals [3].

However, traditional speech enhancement methods may not work well for decrypted speech from hand-talk devices. Most traditional single-channel speech enhancement methods work on the assumption of slowly changing noise and/or the presence of noise all the time. If this assumption was, noise could be estimated and updated when speech was predicted to be absent. Unfortunately, the artifacts in decrypted speech signals from hand-talk devices are characteristically different. Encryption is performed only when speech is present. Therefore, the distortions occur mostly when speech is present and are minimal in non-speech periods. As a result, most noise estimators are not suitable, since they are not designed to deal with abruptly changed distortions [15–17].

Deep learning has been actively implemented for speech enhancement in recent years. Various architectures, such as deep neural networks (DNNs) [18–20], auto-encoders (AEs) [21], recurrent neural networks [22], convolutional neural networks (CNNs) [23, 24] and generative adversarial networks (GANs) [25–27] have been employed. Unlike traditional methods that rely on noise estimation, deep-learning-based methods work by mapping distorted speech to the target clean speech. With a certain amount of training data, the parameters of such models are then optimized so that the resulting models can use distorted signals to produce estimations of the clean speech signals; hence, no noise estimation is required.

In this study, we explore the use of deep learning for speech enhancement to improve the quality of decrypted speech. Using an auto-encoder with a convolutional neural network (CNN) is a popular architecture for speech enhancement; examples thereof include WaveNet [28], U-net [29], and SEGAN (speech-enhancement-based GAN) [25]. CNNs are built from convolutional layers to model spatial correlations in data, which may be beneficial for speech enhancement. However, they may not be adequate for modeling the temporal correlations in speech data. In previous studies, attention modules [30] were added for this purpose [31]. Here, we added gating mechanisms to the convolutional layers, as in [32], and used this for speech enhancement. Our experiments showed that deep learning is generally better than traditional methods in improving the quality of speech signals. In our proposed method, gated convolutional neural networks (GCNNs) were shown to be better than CNNs on several evaluation metrics. To the best of our knowledge, no effort has been made to employ deep learning to improve the quality of decrypted speech.

2. Related Studies

The objective of speech enhancement methods is to improve the intelligibility and/or overall perceptual quality of degraded speech signals caused by either environmental noise or other types of distortions. We can categorize speech enhancement methods into two types: masking-based and mapping-based methods. The first category attempts to obtain clean speech estimates via various signal-processing methods. One way to do so is by estimating the noise parts of noisy speech signals and then removing them using spectral subtraction [11], Wiener filtering [12, 13], or Kalman filtering [14]. In other words, estimated noise information is used to choose masking factors for the speech signals and,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

hence, minimize their effects. However, inaccuracies in noise estimation often produce what is called musical noise, which is often more disturbing than the original noise [33]. Various approaches have been employed to minimize the effect of musical noise. Applying flooring factors [34, 35] or nonlinear signal processing methods [36–38] can reduce musical noise. However, the reduction of musical noise levels is usually traded with the degradation of noise reduction levels. Though they are able to reduce the effect of musical noise, iterative approaches [39–41] require more computation and, hence, may not be suitable for real-time implementations.

In mapping-based methods, a particular function is used to map noisy speech to the target clean speech. Using a certain amount of training data, the mapping function is then optimized to fit with the training data. Currently, mapping-based methods are arguably more dominant for speech enhancement with the emergence of deep learning technologies. In deep learning, we can categorize the methods into two groups. Neural network architectures are trained in a discriminative manner or a generative manner; various architectures, such as deep neural networks (DNNs) [18, 19], auto-encoders (AEs) [21], recurrent neural networks [22], convolutional neural networks (CNNs) [23, 24], and generative adversarial networks (GANs) [25–27, 42], have been implemented.

For secured communications, the encryption process produces decrypted speech with degraded quality due to the randomization process. Most studies aim to improve the quality of decrypted speech by designing encryption methods that preserve the speech quality and intelligibility [6, 7, 9, 10]. Only a few studies have attempted to use speech enhancement to improve the quality of decrypted speech. Most studies in this area require multiple microphones or multi-channel communications. For this, beamforming and its variants are often proposed. For instance, in [43], distributed beamforming was shown to improve homomorphic encryption, while joint beamforming is proposed in [44].

Very few studies have attempted to implement enhancements for single microphone/single-channel secured communications. In our previous studies, we applied and proposed masking-based methods and found their limitations when dealing with distortions that were produced by encryption processes [3]. Only slight improvements in speech quality were gained because most masking-based methods assume noise to be present for all periods of the speech signals, so the estimation and the update of noise signals were performed during non-speech periods of the speech signals. However, this is not the case for speech in the encryption process. Since encryption is performed when speech is present, the energy of noise is usually low during non-speech periods, but it drastically increases when speech starts. Therefore, most noise estimators are not suitable, since they usually assume that noise changes slowly [15–17].

Applying mapping-based methods, such as those using deep learning, may be more effective for such conditions. This is because networks may able to learn complex relations between distorted and clean signals, which cannot be achieved with traditional approaches. Thus, the networks can regenerate clean signals given distorted ones when enough data are provided in the training process. In this study, we explored the use of deep learning to enhance the quality of the speech signals from the encryption process and then propose gated convolutional neural networks (GCNNs) for enhancing the quality of speech signals.

3. Proposed Method

Let us denote Ψ_θ as the mapping function with θ as its sets of parameters so that :

$$\mathcal{F}(\hat{X}) = \Psi_\theta \mathcal{F}(Y) \quad (1)$$

where F is a feature operator of speech signals. $X\hat{}$ is the target speech (estimate clean speech), and Y is noisy speech. Ψ_θ is implemented using deep learning architectures such as DNNs [18, 19], auto-encoders (AEs) [21], convolutional neural networks (CNNs) [23, 24], and GANs [25, 26]. Regarding the operating domains, the mapping can be performed in either the time [24–26, 42] or spectral (frequency) domain [18, 19, 21]. The representation of speech in the spectral domain is better at

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

showing the changes in each frequency component in time, so information such as harmonics and the energy of each phoneme may be more discriminative. However, enhancements in the spectral domain often ignore the phase and cross-term, which are very important for speech enhancement in some studies [45, 46]. This would not be the case when conducting enhancement in the time domain, as no phase information is required. In addition, the large capacity of deep learning allows complex relations between speech parts to be modeled in the time domain and lets networks learn important features on their own. For this reason, we employ our deep learning methods in the time domain for this study.

SEGAN is an example of a speech enhancement method that also performs enhancement in the time domain. Due to its effectiveness, it has been widely used, and many variants thereof have been proposed in the literature [26, 42, 47]. SEGAN works by applying a GAN, a generative model, as its core. The GAN comprises two networks: a generator (G) and a discriminator (D). G aims to learn an effective mapping to imitate the distributions of the real data so that new data samples (fake data samples) that are similar to the real ones can be produced. Meanwhile, D is a binary classifier for differentiating between real and fake samples.

Let x denotes samples from a certain distribution (unknown). To learn that distribution, a prior distribution defined as pz(z) is fed to G, and G maps it to the data space. G and D learn in an adversarial manner as follows (Minmax game objective):

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim\mathbf{p}_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim\mathbf{p}_{\mathbf{z}}(\mathbf{z})}\left[1 - \log D\left(G(z)\right)\right] \quad (2)$$

In SEGAN, G enhances speech signals, i.e., given noisy speech; it maps the enhanced speech ($\mathbf{x}\hat{} = G(\mathbf{y})$) instead of random noise, as a typical GAN would. G is built by adopting U-Net architectures [48]. It is an auto-encoder with skip connections between paired encoders and decoders.

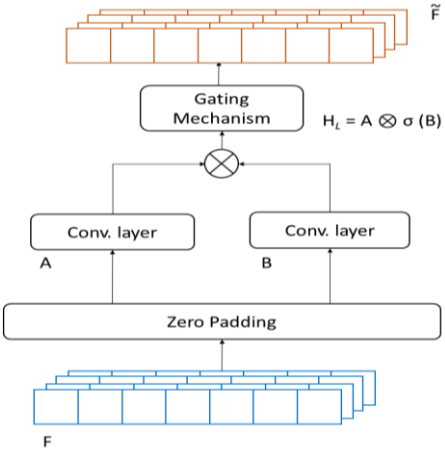


Figure 1.
A block diagram of the proposed method.

The method includes padding to ensure that the data and gating mechanism are of the same length. A gated linear unit is used as a gating mechanism by applying $\otimes$, the element-wise product between the output of the convolutional layers and its sigmoid function.

Recurrent networks and their variants, such as LSTM and BiLSTM, are often applied to many types of sequence data, such as text, video, and speech. For sequence data, contextual information may be needed to predict the next sequence. Since speech data are time-dependent, contextual information may also be beneficial for speech enhancement. Interestingly, many speech enhancement methods employ CNNs instead [23–26]. This is because of the contextual information, albeit limited, is obtained with stacked CNNs. Since CNNs are highly parallelizable, unlike LSTM and BiLSTM, they are much

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

more efficient. In CNNs, the computation of all input data can be performed in a simultaneous manner, unlike in recurrent networks, where the output of the current step depends on the previous hidden state, making them unsuitable for parallelization. Adding a gating mechanism as in [32] may be beneficial for speech enhancement without sacrificing the parallelization capabilities. By applying the element-wise product of the output from a convolutional layer to the output of the sigmoid function, the gating mechanism can control how much of it passes, similarly to LSTM.

How a GCNN works is illustrated in Fig. 1. Here, we dealt with raw speech data; therefore, 1D convolutional layers were used. To ensure that all of the speech data had the same length, zero-padding was applied. Let us denote F ∈ RT×m as a feature map that is the input for the GCNN with a size of T (time dimension) and m (the channel size). Let A = F ∗ W + a and B = F ∗ U + b be the outputs of convolutional layers so that the hidden layer HL can be written as follows:

$$H_L(F) = A \otimes \sigma(B) \quad (3)$$

where $W \in R^{k \times m \times n}$, $U \in R^{k \times m \times n}$, $a \in R^n$, and $b \in R^n$ are the parameters of the convolutional layers, σ is the sigmoid activation function, and $\otimes$ performs element-wise matrix multiplication. As a result, we can see that the output of a GCNN is a linear projection of $F * W + a$ that is modulated by $\sigma(F * U + b)$. This mechanism is called a gated linear unit (GLU) (Dauphin et al, 2017). By applying element-wise multiplication, information from A is controlled. For instance, when the output of the sigmoid function is 0, nothing is passed; in other words, it is forgotten, whereas when it is to be 1, all is passed. This is similar to the forget mechanism in LSTM. When GCNNs are stacked on top of F, the final layers are the composition of $H_L \circ H_{L-1} \ldots \circ H_1 \circ H_0(F)$, thus capturing the contextual information of speech data.

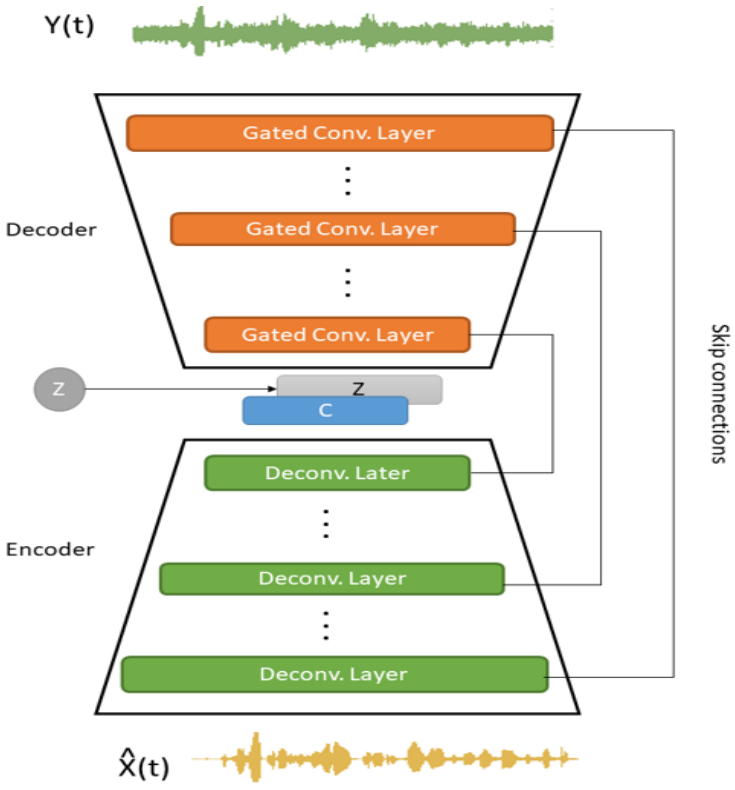


Figure 2.
A block diagram of a gated convolutional layer.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

Table 1.
The architectures of the auto-encoder for various depths of GCNN layers (N) and their kernels.

Name	N	Kernels
$Prop_{1024}$	11	[16, 32, 32, 64, 64, 128, 128, 256, 256, 512, 1024]
$Prop_{512}$	10	[16, 32, 32, 64, 64, 128, 128, 256, 256, 512]
$Prop_{256}$	9	[16, 32, 32, 64, 64, 128, 128, 256, 256]
$Prop_{128}$	7	[16, 32, 32, 64, 64, 128, 128]
$Prop_{64}$	5	[16, 32, 32, 64, 64]
$Prop_{32}$	3	[16, 32, 32{]

We employed GCNNs with the auto-encoder architecture illustrated in Fig. 2. A UNet architecture [48], which is often implemented for speech enhancement [25, 49], was adopted. We employed the GCNN on only the encoder of the auto-encoder. We replaced each convolutional layer with a GCNN. We varied the number of layers of encoders to 3, 5, 7, 9, 10, and 11 and used the corresponding kernel sizes. We also used skip connections in the autoencoder. The details of various architectures that we evaluated are shown in Table 1. *N* refers to the number of GCNN layers at the encoder.

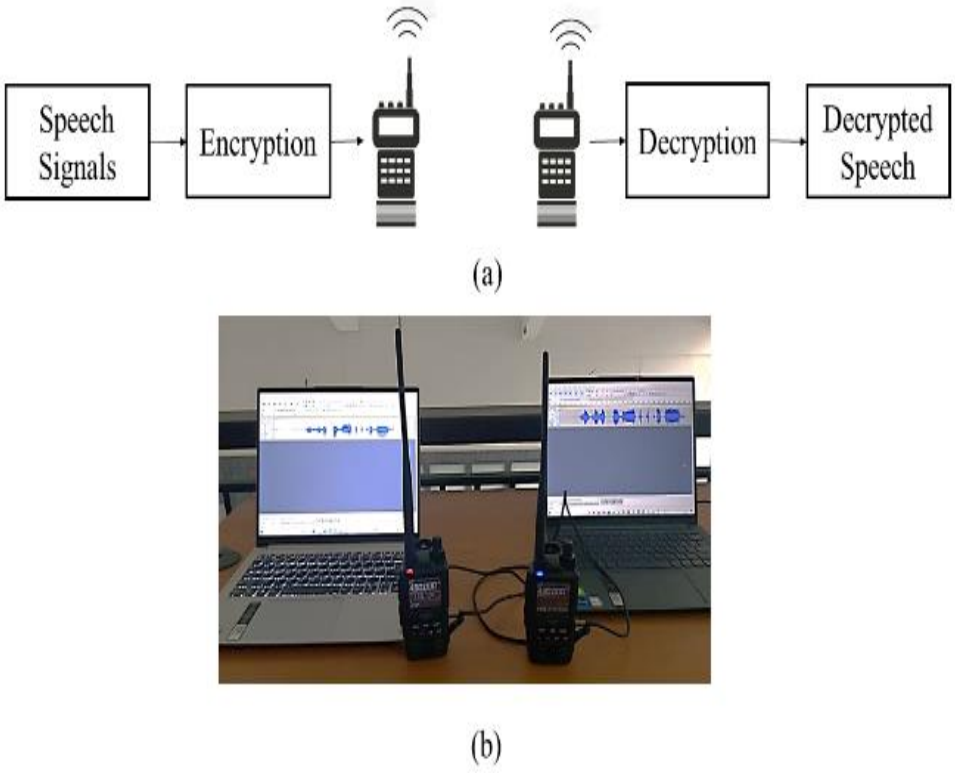


Figure 3.
Data creation process: (a) the process of creating the dataset; (b) the actual setup conditions.

4. Experimental Setup

This study was part of a project for developing secured communications through handtalk devices. We generated a dataset for this purpose from subsets of an Indonesian speech corpus [50]. This dataset was collected from 20 speakers that read around 390 phonetically balanced sentences. We selected 50

utterances from each speaker in the dataset for a total o f1000 utterances. The process of creating the experimental dataset is shown in Fig. 3.

We could not create a dataset for the whole set due to time constraints and the limited resources of the project.

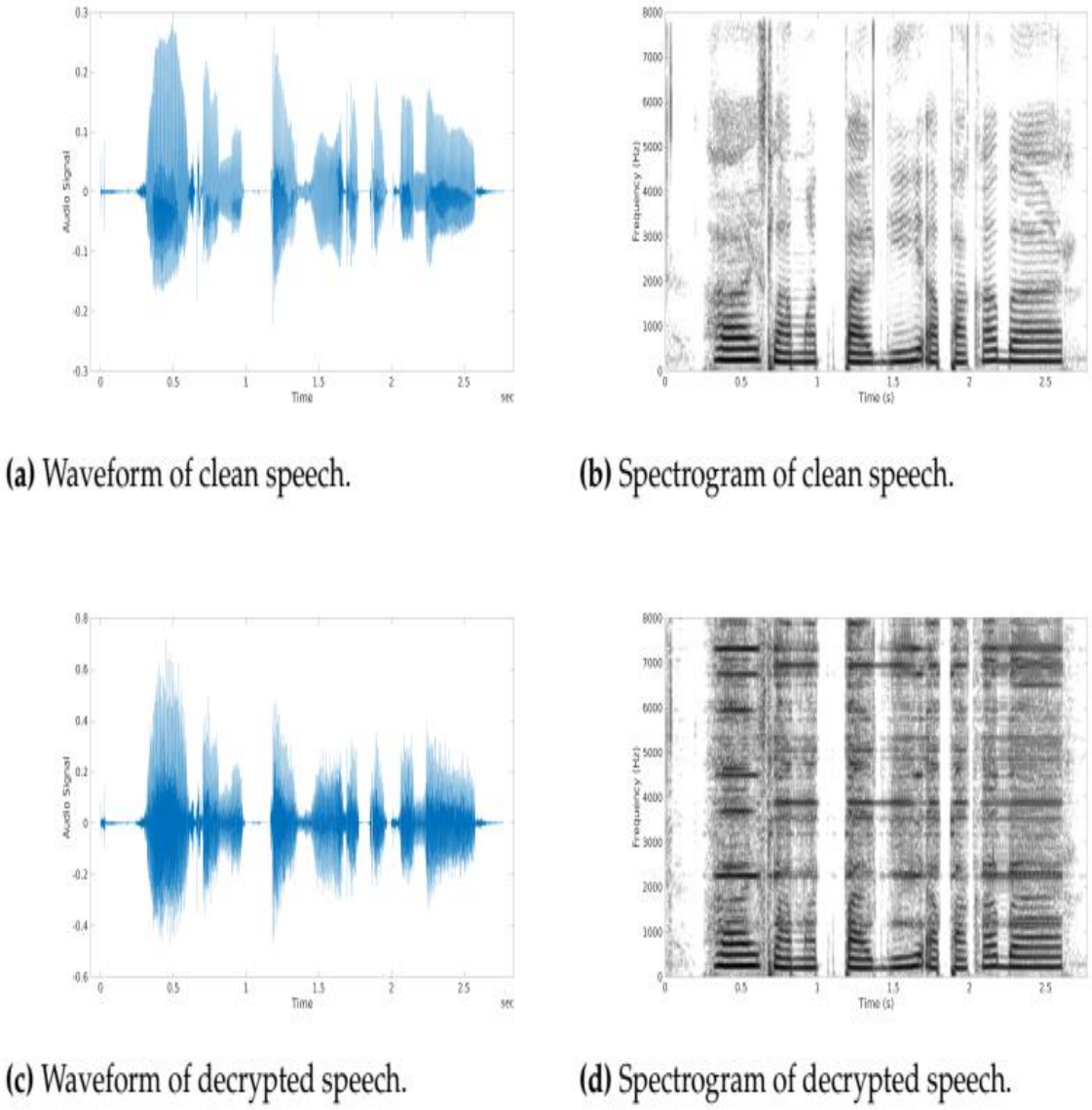
All of the selected utterances were encrypted before they were sent through handtalk devices. Then, the encrypted data received by a hand-talk device at another end were decrypted. For the hand-talk devices, we used Yaesu Model FT3DR, while a laptop with an 11th Generation Intel Core i-72 and 16 GB of memory was used to perform encryption, play the recordings of the encrypted speech from one end, and then to record the speech received from the hand-talk device. Meanwhile, a method from [51] was used for encryption. It was based on shuffling the data using the chaotic permutation of multiple circular shrinking and expanding methods. For details of the methods, the reader can refer to [51]. Samples of the decrypted speech are shown in Fig. 4. It was clear that most distortions occurred when speech was present, and when speech was absent, the distortions were minimal. This was different from the usual problems in speech enhancement, where it is assumed that noise or distortions are present all the time (during speech and non-speech periods).

To train all of the deep learning architectures, we used the same configurations to split the training and testing data. For training, we used data from 16 speakers (800 utterances), and the rest (4 speakers with 200 utterances) were used for testing. as good predictors for the MOS of the signal distortion, the background noise interference, and the overall effect, respectively. The MOS values ranged between 1 and 5, where 1 was the lowest quality and 5 was the highest; the CSIG, CBAK, and COVL also had the same range.

We used two objective metrics to evaluate our method. They were the perceptual evaluation of speech quality (PESQ) [52] and short-time objective intelligibility (STOI). For the PESQ scores, the values ranged between −0.5 and 4.5, with higher scores indicating a better quality, while a higher STOI score indicated higher quality. Due to funding limitations, the mean opinion score (MOS) from subjective listening tests could not be included in this study. Instead, we used the signal distortion (CSIG), the MOS predictor of background noise intrusiveness (CBAK), and the MOS predictor of overall signal quality (COVL) [53], which can be considered as good predictors for the MOS of the signal distortion, the background noise interference, and the overall effect, respectively. The MOS values ranged between 1 and 5, where 1 was the lowest quality and 5 was the highest; the CSIG, CBAK, and COVL also had the same range.

For comparison, we implemented several masking-based and mapping-based speech enhancement methods. For masking-based methods, we implemented spectral subtraction (SS) [54], the Karhurnen–Loeve transform (KLT) [55], the log minimum-mean-squared-error (MMSE) [56], and coupled SS and Wiener filtering (SS+WF) [3]. For mapping-based methods, several deep-learning-based speech enhancement methods were implemented. They were the CNN-based auto-encoder (CNN-AE) [57], speech-enhancement-based GAN (SEGAN) [25], SEGAN with relativistic loss (RSGAN) [26], and equilibrium recurrent neural networks (ERNNs) [58].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684



(a) Waveform of clean speech.

(b) Spectrogram of clean speech.

(c) Waveform of decrypted speech.

(d) Spectrogram of decrypted speech.

Figure 4.
Samples of decrypted speech (waveform and the spectrogram) in comparison with clean speech.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

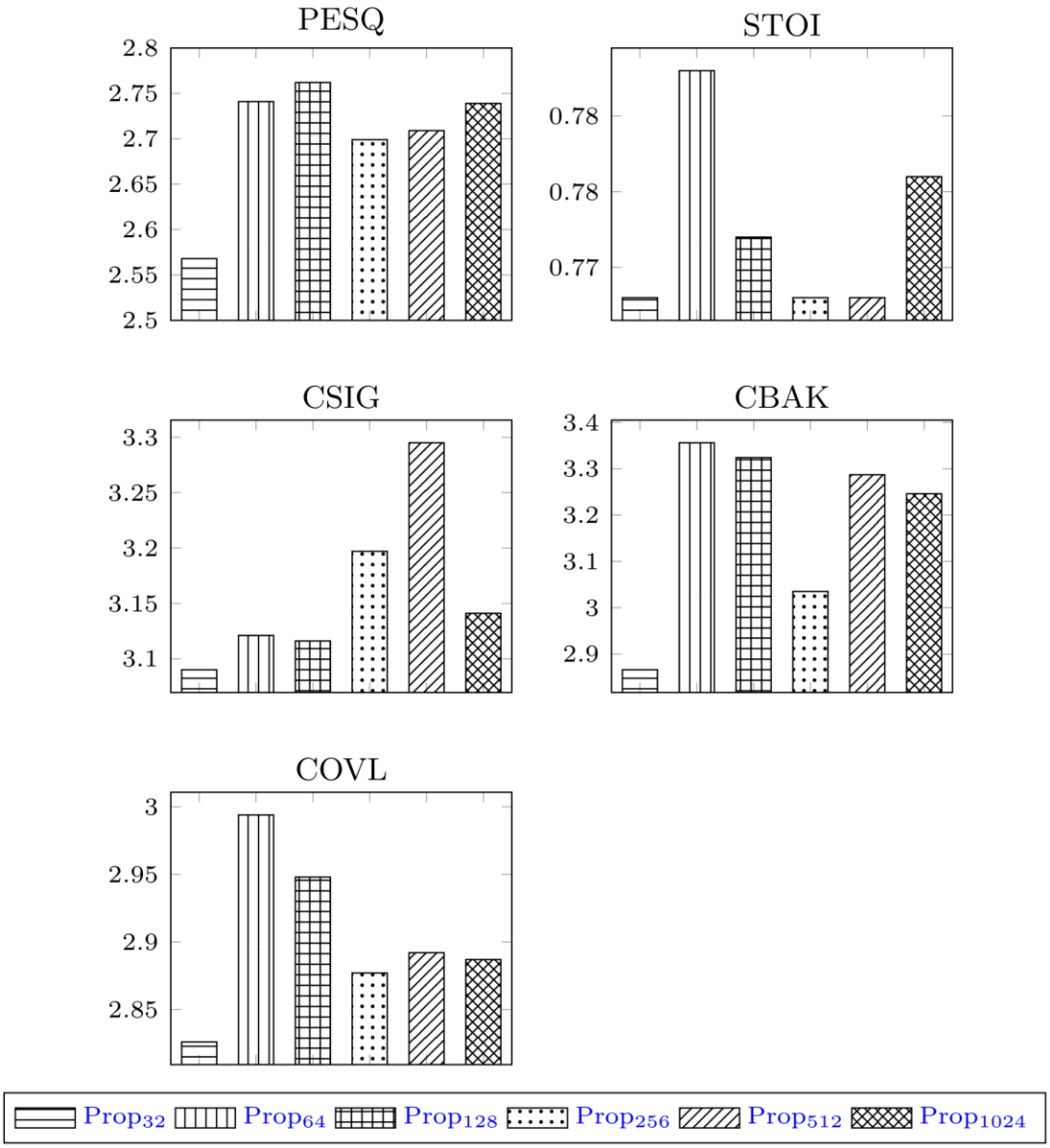


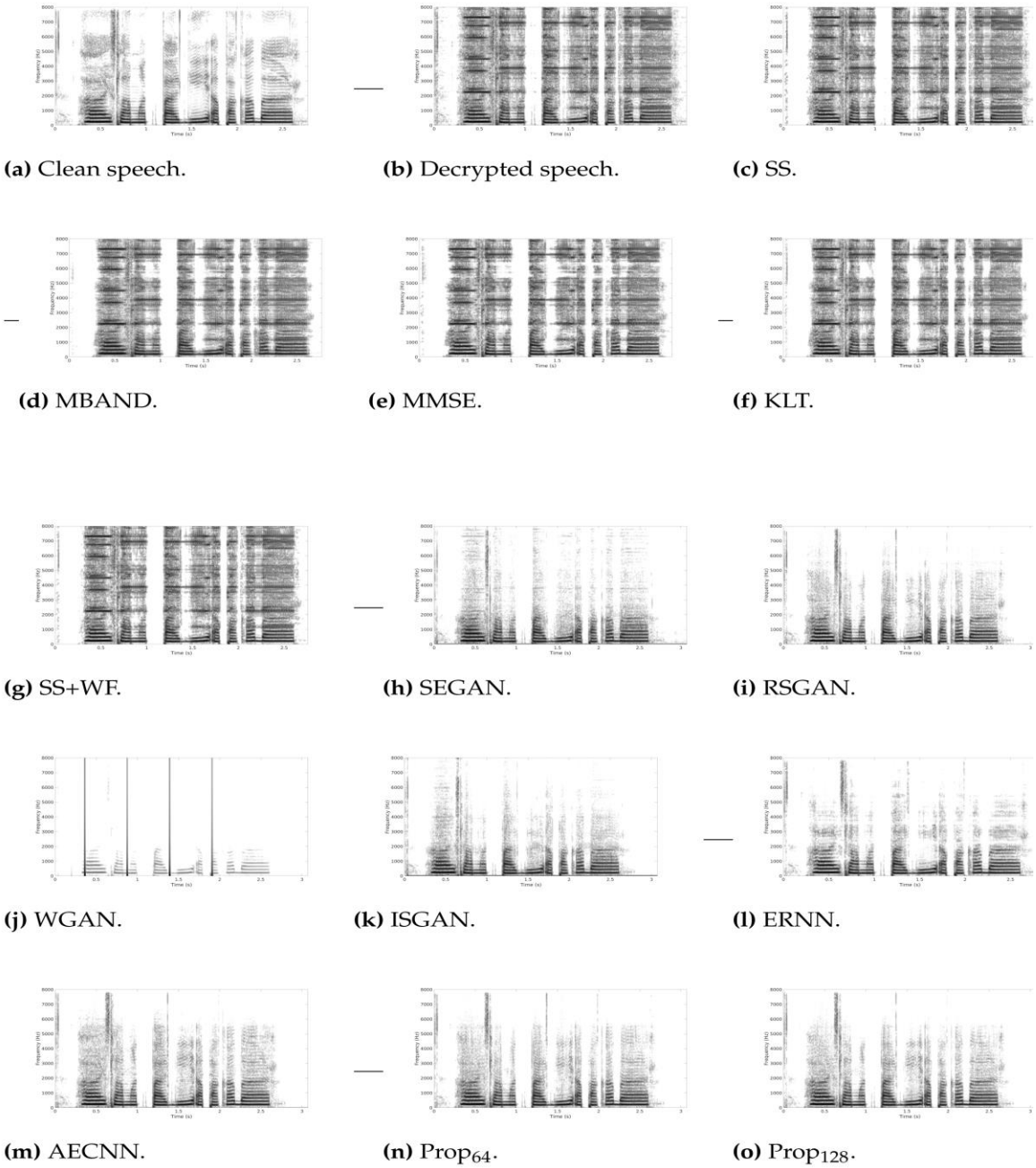
Figure 5.
Performance of the GCNN when the kernel size and number of layers were varied.

5. Results and Discussions

In Figure 5 and Figure 6 the performance of the proposed method with various depths is displayed. One can see that, while increasing the depth of the network may improve its performance, the deepest networks do not necessarily have the best performance. We noticed that the proposed method with a deeper network had less speech distortion, as indicated by the higher CSIG scores. However, the overall quality seemed to be lower. The PESQ, STOI, and COVL scores tended to be lower for Prop256, Prop512, and Prop1024. These metrics are often deemed more suitable for indicating quality and intelligibility [53] than the signal-to-noise ratio and distance-based metrics. Based on this

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

consideration, Prop64 or Prop128 may be the optimal configuration for maintaining speech with good quality and intelligibility.



(a) Clean speech. **(b)** Decrypted speech. **(c)** SS.

(d) MBAND. **(e)** MMSE. **(f)** KLT.

(g) SS+WF. **(h)** SEGAN. **(i)** RSGAN.

(j) WGAN. **(k)** ISGAN. **(l)** ERNN.

(m) AECNN. **(n)** Prop_{64}. **(o)** Prop_{128}.

Figure 6.
Samples of the spectrograms of enhanced decrypted speech from various speech enhancement methods. As a reference, the spectrograms of the clean speech and original decrypted speech are also included ((a) and (b), respectively).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

In Table 2, the proposed method is compared with other reference methods. It is clear that all traditional methods, such as KLT, SS, MMSE, and MBAND, failed to significantly remove the distortions from the decrypted speech. Their metrics only slightly improved, and in many cases, they were worse than when no enhancement was applied. This was not surprising, as we found similar results in our previous study [3]. Most smoothing-based methods can only deal with noise that changes slowly. This was not the case for our problem, where encryption was conducted when speech was present. Therefore, the distortions mostly occurred when speech was present, whereas they were nearly non-existent when speech was absent. As a consequence, the traditional speech enhancement methods failed to reduce the distortions, and little improvement could be achieved (see Figs. 3 and 6 for a comparison of the spectrogram of decrypted speech without enhancement and the spectrogram of decrypted speech after speech enhancement; not many of the artifacts from the encryption process were removed).

Table 2.
Comparison of the proposed methods with other methods. The best results are printed in bold. The results without a gating mechanism (denoted as NG) are also presented.

Methods	PESQ	STOI	CSIG	CBAK	COVL
No enh,	1.66± 0.35	0.75 ± 0.26	2.12 ± 0.75	1.98 ± 0.34	1.84 ± 0.58
KLT	1.69± 0.37	0.74 ± 0.27	1.95 ± 0.70	1.95 ± 0.29	1.75 ± 0.53
MMSE	1.65 ± 0.36	0.74 ± 0.27	2.03 ± 0.72	1.97 ± 0.36	1.79 ± 0.55
MBAND	1.64 ± 0.35	0.73 ± 0.27	2.02 ± 0.71	1.83 ± 0.27	1.77 ± 0.53
SS	1.69± 0.36	0.74 ± 0.27	2.03 ± 0.67	1.95 ± 0.38	1.80± 0.52
SS+WF	1.71 ± 0.41	0.74 ± 0.27	2.07 ± 0.68	1.97 ± 0.39	1.83 ± 0.52
SEGAN	2.20 ± 0.50	0.73 ± 0.29	2.71 ± 0.93	2.41 ± 0.77	2.38 ± 0.76
RSGAN	2.56 ± 0.82	0.77 ± 0.32	3.18 ± 1.04	3.12 ± 1.09	2.83 ± 1.02
ISEGAN	2.34 ± 0.56	0.74 ± 0.29	2.65 ± 0.96	2.41 ± 0.82	2.41 ± 0.81
WGAN	1.33 ± 0.34	0.43 ± 0.23	2.41 ± 0.49	2.49 ± 0.60	1.82 ± 0.38
ERNN	2.78 ± 0.91	0.77 ± 0.31	3.25 ± 1.31	2.76 ± 0.54	**3.00 ± 1.14**
CNN-AE	2.71 ± 0.75	0.78 ± 0.33	3.07 ± 1.12	3.23 ± 1.21	2.87 ± 1.00
$Prop_{64}$ (NG)	2.73 ± 0.80	**0.78 ± 0.32**	3.20 ± 1.13	3.29 ± 1.10	2.94 ± 1.04
$Prop_{64}$	2.74 ± 0.80	**0.78 ± 0.32**	**3.29 ± 1.13**	**3.36 ± 1.10**	2.99 ± 1.04
$Prop_{128}$ (NG)	2.69 ± 0.73	0.78 ± 0.34	3.17 ± 1.10	3.16 ± 1.10	2.89 ± 0.99
$Prop_{128}$	**2.76 ± 0.83**	0.77 ± 0.34	3.20 ± 1.04	3.32 ± 1.10	2.95 ± 1.01
$Prop_{256}$ (NG)	2.81 ± 0.81	0.79 ± 0.33	3.21 ± 1.21	3.08 ± 1.14	2.98 ± 1.07
Prop256	2.70 ± 0.77	0.77 ± 0.35	3.12 ± 1.17	3.03 ± 1.11	2.88 ± 1.03

Meanwhile, the deep learning methods performed significantly better than the smoothing based methods. For most methods, all of the metrics generally improved, confirming that the speech quality and intelligibility were improved in comparison with the original decrypted speech. The spectrogram of the speech that was enhanced using deep learning (see Fig. 6) showed that these methods were able to learn how to recreate enhanced speech that was quite close to clean speech. We also noticed that not all GAN-based methods performed well in our evaluation. WGAN appeared to be unable to regenerate the enhanced speech well. We also noticed that non-GAN-based deep learning methods (ERNN, CNN-AE, and the proposed method) were mostly better than than the GAN-based methods (SEGAN, WGAN, RSGAN, and ISGAN). This may have been related to the stability of GAN methods. Since the generator networks of GANs never see the actual clean data, the learning trajectory of a GAN may be unpredictable, and this may cause instability [25, 59]

Compared with the CNN, i.e., CNN-AE and the proposed architecture without a gating mechanism (denoted in Table 2 with NG), the GCNN was better in terms of most of the evaluation metrics. Better

PESQ, STOI, CSIG, CBAK, and COVL scores were obtained compared with those of the CNN. Compared with the ERNN, which was a variant of a recurrent neural network (RNN), the proposed method was only worse in terms of the PESQ score. It yielded better scores for other metrics. Our method outperformed all of the referenced GAN-based methods. Note that, unlike in the CNN, the changes were only applied to the CNN of the encoder, and the results showed better overall performance, which suggested that adding a gating mechanism to a CNN may improve the speech enhancement system.

We noticed that the standard deviations of our proposed methods were larger than those of traditional methods (see Table 2). The results were similar for all proposed methods ($Prop_{32}$ to $Prop_{1024}$). Larger standard deviations were also found for almost all other deep learning-based methods that we evaluated. This indicated the large variations in the quality of the decrypted speech for some test data. This might have been due to the small variations in the training data, which could have made the models prone to overfitting.

6. Conclusions and Future Works

In this study, we evaluated several deep learning methods for enhancing distorted decrypted speech from secure communication and proposed the addition of a gating mechanism to a CNN for speech enhancement. Encryption processes that include the randomization of speech components may cause distortions in the decrypted speech. However, traditional speech enhancement methods, which usually employ smoothing rules to the distorted speech, may fail to deal with this because the distortion conditions differ from the usual assumptions made by most traditional methods. Deeplearning-based methods, on the other hand, are effective in doing so. Deep learning methods are able to learn a mapping function for the distorted speech to generate clean speech estimates given the distorted samples. Our experiments confirm this. Deep-learning-based speech enhancement achieved better quality metrics than those of traditional methods. Furthermore, our proposed method was generally better than CNN-based methods. However, we need to emphasize the limitation of our study. The methods have not been evaluated by subjective listening test, which is considered the actual measure of speech quality and intelligibility. While the results on several metrics that show good relation on subjective listening tests indicate the proposed method could improve the speech quality, it is not necessarily indicate the actual quality of speech signals.

We must note, however, that while deep learning clearly produced better enhanced speech than that of smoothing-based methods, much needs to be done in this area. Our results indicate that not all deep learning approaches are effective in this task. This may be related to the limited data availability. Furthermore, deep learning-based speech enhancement methods require both the data of the distorted speech and the clean reference data. These data are often unavailable in realistic conditions. Deep learning methods also usually require much more computation to train networks and larger storage for models. This often makes it difficult to implement such methods for off-line and embedded systems. Therefore, exploring lightweight models is planned in our future work. Our study focus on distortions due to encryption process in hand-talk communications. The effectiveness for other tasks could also be explored such as for dealing with sudden noise or rapidly changing noise, which are still tasks for speech recognition.

Acknowledgment:

This work and publication was supported by Fund for Innovative-Productive Research (Riset Inovatif-Produktif – RISPRO) of the Indonesia Endowment Funds for Education (Lembaga Pengelola Dana Pendidikan, LPDP) under the contract No: PRJ-52/LPDP/2023 & No: 46/PKS/WRIII-DISTP/UI/2023.

References
[1] Katugampala, N.N., Al-Naimi, K.T., Villette, S., Kondoz, A.M.: Real time data transmission over gsm voice channel for secure voice and data applications. In: The 2nd IEE Secure Mobile Communications Forum: Exploring the Technical Challenges in Secure GSM and WLAN, 2004. (Ref. No. 2004/10660), pp. 7–174 (2004). https://doi.org/10.1049/ic.2004.0663
[2] Hamdi, M., Hermassi, H., Rhouma, R., Belghith, S.: A new secure and efficient scheme of adpcm encoder based on chaotic encryption. In: 2014 1st International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 7–11 (2014). https://doi.org/10.1109/ATSIP.2014.6834580
[3] Pardede, H., Ramli, K., Suryanto, Y., Hayati, N., Presekal, A.: Speech enhancement for secure communication using coupled spectral subtraction and wiener filter. Electronics 8(8) (2019) https://doi.org/10.3390/electronics8080897
[4] Lee, L.-s., Chou, G.-c.: Asynchronous speech encryption - formulation and simulation. In: MILCOM 1983 - IEEE Military Communications Conference, vol. 3, pp. 791–795 (1983). https://doi.org/10.1109/MILCOM.1983.4794809
[5] Sathiyamurthi, P., Ramakrishnan, S.: Speech encryption algorithm using fft and 3d-lorenz–logistic chaotic map. Multimedia Tools and Applications 79(25), 17817–17835 (2020)
[6] Matthews, R.: On the derivation of a "chaotic" encryption algorithm. Cryptologia 13(1), 29–42 (1989)
[7] Mosa, E., Messiha, N.W., Zahran, O., El-Samie, A., Fathi, E.: Chaotic encryption of speech signals. International Journal of Speech Technology 14(4), 285–296 (2011)
[8] Lin, Q.-H., Yin, F.-L., Mie, T.-M., Liang, H.-L.: A speech encryption algorithm based on blind source separation. In: 2004 International Conference on Communications, Circuits and Systems (IEEE Cat. No.04EX914), vol. 2, pp. 1013–10172 (2004). https://doi.org/10.1109/ICCCAS.2004.1346350
[9] Lin, Q.-H., Yin, F.-L., Mei, T.-M., Liang, H.: A blind source separation based method for speech encryption. IEEE Transactions on Circuits and Systems I: Regular Papers 53(6), 1320–1328 (2006)
[10] Lima, J.B., Silva Neto, E.F.: Audio encryption based on the cosine number transform. Multimedia Tools and Applications 75(14), 8403–8418 (2016)
[11] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on acoustics, speech, and signal processing 27(2), 113–120 (1979)
[12] El-Fattah, A., Marwa, A., Dessouky, M.I., Abbas, A.M., Diab, S.M., El-Rabaie, E.-S.M., Al-Nuaimy, W., Alshebeili, S.A., El-samie, A., Fathi, E.: Speech enhancement with an adaptive wiener filter. International Journal of Speech Technology 17(1), 53–64 (2014)
[13] Yelwande, A., Kansal, S., Dixit, A.: Adaptive wiener filter for speech enhancement. In: 2017 International Conference on Information, Communication, Instrumentation and Control (icicic), pp. 1–4 (2017). IEEE
[14] Paliwal, K., Basu, A.: A speech enhancement method based on kalman filtering. In: ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 12, pp. 177–180 (1987). IEEE
[15] Barnov, A., Bracha, V.B., Markovich-Golan, S.: Qrd based mvdr beamforming for fast tracking of speech and noise dynamics. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 369– 373 (2017). IEEE
[16] Lu, C.-T., Lei, C.-L., Shen, J.-H., Wang, L.-L., Tseng, K.-F.: Estimation of noise magnitude for speech denoising using minima-controlled-recursive-averaging algorithm adapted by harmonic properties. Applied Sciences 7(1), 9 (2016)
[17] He, Q., Bao, F., Bao, C.: Multiplicative update of auto-regressive gains for codebook-based speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(3), 457–468 (2016)
[18] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: An experimental study on speech enhancement based on deep neural networks. IEEE Signal processing letters 21(1), 65–68 (2013)
[19] Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.: A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(1), 7–19 (2014)
[20] Lin, S., Zhang, W., Qian, Y.: Two-stage single-channel speech enhancement with multi-frame filtering. Applied Sciences 13(8) (2023)
[21] Lu, X., Tsao, Y., Matsuda, S., Hori, C.: Speech enhancement based on deep denoising autoencoder. In: Interspeech, vol. 2013, pp. 436–440 (2013)
[22] Leglaive, S., Alameda-Pineda, X., Girin, L., Horaud, R.: A recurrent variational autoencoder for speech enhancement. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 371–375 (2020). IEEE
[23] Tawara, N., Kobayashi, T., Ogawa, T.: Multi-channel speech enhancement using time-domain convolutional denoising autoencoder. In: INTERSPEECH, pp. 86– 90 (2019)
[24] Pandey, A., Wang, D.: A new framework for cnn-based speech enhancement in the time domain. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(7), 1179–1188 (2019)

[25] Pascual, S., Bonafonte, A., Serr`a, J.: Segan: Speech enhancement generative adversarial network. Proc. Interspeech 2017, 3642–3646 (2017)
[26] Baby, D., Verhulst, S.: Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 106–110 (2019). IEEE
[27] Hao, X., Xu, C., Xie, L.: Neural speech enhancement with unsupervised pretraining and mixture training. Neural Networks 158, 216–227 (2023) https://doi. org/10.1016/j.neunet.2022.11.013
[28] Rethage, D., Pons, J., Serra, X.: A wavenet for speech denoising. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5069–5073 (2018). IEEE
[29] Wang, K., He, B., Zhu, W.-P.: Caunet: Context-aware u-net for speech enhancement in time domain. In: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5 (2021). IEEE
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
[31] Zhang, G., Yu, L., Wang, C., Wei, J.: Multi-scale temporal frequency convolutional network with axial attention for speech enhancement. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9122–9126 (2022). https://doi.org/10.1109/ICASSP43922.2022. 9746610
[32] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International Conference on Machine Learning, pp. 933–941 (2017). PMLR
[33] Capp´e, O.: Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. IEEE transactions on Speech and Audio Processing 2(2), 345–349 (1994)
[34] Cohen, I., Berdugo, B.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE signal processing letters 9(1), 12–15 (2002)
[35] Breithaupt, C., Gerkmann, T., Martin, R.: Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. IEEE Signal processing letters 14(12), 1036–1039 (2007)
[36] Udrea, R.M., Vizireanu, N., Ciochina, S., Halunga, S.: Nonlinear spectral subtraction method for colored noise reduction using multi-band bark scale. Signal Processing 88(5), 1299–1303 (2008)
[37] Boldt, J.B., Ellis, D.P.W.: A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In: 2009 17th European Signal Processing Conference, pp. 1849–1853 (2009)
[38] Pardede, H., Iwano, K., Shinoda, K.: Spectral subtraction based on non-extensive statistics for speech recognition. IEICE TRANSACTIONS on Information and Systems 96(8), 1774–1782 (2013)
[39] Yan, X., Yang, Z., Wang, T., Guo, H.: An iterative graph spectral subtraction method for speech enhancement. Speech Communication 123, 35–42 (2020)
[40] Miyazaki, R., Saruwatari, H., Inoue, T., Takahashi, Y., Shikano, K., Kondo, K.:Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Transactions on Audio, Speech, and Language Processing 20(7), 2080–2094 (2012)
[41] Yan, X., Yang, Z., Wang, T., Guo, H.: An iterative graph spectral subtraction method for speech enhancement. Speech Communication 123, 35–42 (2020) https: //doi.org/10.1016/j.specom.2020.06.005
[42] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., Gerkmann, T.: Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 2351–2364 (2023) https://doi.org/10.1109/TASLP.2023.3285241
[43] Hendriks, R.C., Erkin, Z., Gerkmann, T.: Privacy-preserving distributed speech enhancement forwireless sensor networks by processing in the encrypted domain. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7005–7009 (2013). https://doi.org/10.1109/ICASSP.2013.6639020
[44] Qin, H., Chen, X., Sun, Y., Zhao, M., Wang, J.: Optimal power allocation for joint beamforming and artificial noise design in secure wireless communications. In: 2011 IEEE International Conference on Communications Workshops (ICC), pp. 1–5 (2011). https://doi.org/10.1109/iccw.2011.5963524
[45] Paliwal, K., W´ojcicki, K., Shannon, B.: The importance of phase in speech enhancement. speech communication 53(4), 465–494 (2011)
[46] Zheng, N., Zhang, X.-L.: Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27(1), 63–76 (2018)
[47] Phan, H., Le Nguyen, H., Ch´en, O.Y., Koch, P., Duong, N.Q., McLoughlin, I., Mertins, A.: Self-attention generative adversarial network for speech enhancement. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7103–7107 (2021). IEEE
[48] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer
[49] Giri, R., Isik, U., Krishnaswamy, A.: Attention wave-u-net for speech enhancement. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 249–253 (2019). IEEE
[50] Lestari, D.P., Iwano, K., Furui, S.: A large vocabulary continuous speech recognition system for indonesian language. In: 15th Indonesian Scientific Conference in Japan Proceedings, pp. 17–22 (2006)

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684

[51] Suryanto, Y., Ramli, K., et al.: A new image encryption using color scrambling based on chaotic permutation multiple circular shrinking and expanding. Multimedia Tools and Applications 76(15), 16831–16854 (2017)
[52] Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IIEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), vol. 2, pp. 749–7522 (2001)
[53] Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, Language Process. 16(1), 229–238 (2008)
[54] Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: 1979 IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), vol. 4, pp. 208–211 (1979)
[55] Hu, Y., Loizou, P.C.: A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. 11(4), 334–341 (2003)
[56] Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33(2), 443–445 (1985)
[57] Shi, Y., Rong, W., Zheng, N.: Speech enhancement using convolutional neural network with skip connections. In: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 6–10 (2018). https://doi.org/ 10.1109/ISCSLP.2018.8706591
[58] Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., Harada, N.: Real-time speech enhancement using equilibriated rnn. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 851–855 (2020). IEEE
[59] Pandey, A., Wang, D.: On adversarial training and loss functions for speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5414–5418 (2018). https://doi.org/10.1109/ ICASSP.2018.8462614

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 5: 256-270, 2024
DOI: 10.55214/25768484.v8i5.1684