applied sciences

MDPI

Article

Assessment of Self-Supervised Denoising Methods for Esophageal Speech Enhancement

Madiha Amarjouf [1,*], El Hassan Ibn Elhaj [1], Mouhcine Chami [2], Kadria Ezzine [3] and Joseph Di Martino [3]

[1] Research Laboratory in Telecommunications Systems: Networks and Services (STRS), Research Team: Multimedia, Signal and Communications Systems (MUSICS), National Institute of Posts and Telecommunications (INPT), Av. Allal Al Fassi, Rabat 10112, Morocco; ibnelhaj@inpt.ac.ma

[2] Research Laboratory in Telecommunications Systems: Networks and Services (STRS), Research Team: Secure and Mixed Architecture for Reliable Technologies and Systems (SMARTS), National Institute of Posts and Telecommunications (INPT), Av. Allal Al Fassi, Rabat 10112, Morocco; chami@inpt.ac.ma

[3] LORIA-Laboratoire Lorrain de Recherche en Informatique et ses Applications, B.P. 239, 54506 Vandœuvre-lès-Nancy, France; kadria.ezzine@gmail.com (K.E.); joseph.di-martino@loria.fr (J.D.M.)

* Correspondence: amarjouf.madiha@doctorant.inpt.ac.ma

Abstract: Esophageal speech (ES) is a pathological voice that is often difficult to understand. Moreover, acquiring recordings of a patient's voice before a laryngectomy proves challenging, thereby complicating enhancing this kind of voice. That is why most supervised methods used to enhance ES are based on voice conversion, which uses healthy speaker targets, things that may not preserve the speaker's identity. Otherwise, unsupervised methods for ES are mostly based on traditional filters, which cannot alone beat this kind of noise, making the denoising process difficult. Also, these methods are known for producing musical artifacts. To address these issues, a self-supervised method based on the Only-Noisy-Training (ONT) model was applied, consisting of denoising a signal without needing a clean target. Four experiments were conducted using Deep Complex UNET (DCUNET) and Deep Complex UNET with Complex Two-Stage Transformer Module (DCUNET-cTSTM) for assessment. Both of these models are based on the ONT approach. Also, for comparison purposes and to calculate the evaluation metrics, the pre-trained VoiceFixer model was used to restore the clean wave files of esophageal speech. Even with the fact that ONT-based methods work better with noisy wave files, the results have proven that ES can be denoised without the need for clean targets, and hence, the speaker's identity is retained.

Keywords: esophageal speech; self-supervised denoising; speech enhancement; DCUNET; DCUNET-cTSTM; STFT; VoiceFixer

check for updates

Citation: Amarjouf, M.; Ibn Elhaj, E.H.; Chami, M.; Ezzine, K.; Di Martino, J. Assessment of Self-Supervised Denoising Methods for Esophageal Speech Enhancement. *Appl. Sci.* **2024**, *14*, 6682. https://doi.org/10.3390/app14156682

Academic Editors: Edoardo Piana and Douglas O'Shaughnessy

Received: 16 April 2024
Revised: 5 July 2024
Accepted: 23 July 2024
Published: 31 July 2024

1. Introduction

Certainly, speech serves as the primary means of communication for humans. However, people suffering from pathological voice disorders encounter obstacles in both communication and social interaction, which have a considerable impact on their overall quality of life [1–3]. One of the multiple causes of voice dysfunctions is laryngeal or hypo-pharyngeal cancer [4]. In the advanced stages of this disease, the larynx needs to be removed entirely [5–7], which is a crucial apparatus for speech generation using vocal folds [4,8,9]. However, with the help of a speech therapist [8–10], patients can communicate through either esophageal speech (ES), tracheo-esophageal speech (TES), or electro-larynx speech (EL Speech) [1,3,4]. ES is the most commonly used communication tool and sounds more natural than EL Speech [6]. Still, it is often characterized by harsh speech with specific noises, low pitch range and intensity [11,12], weak intelligibility, and being difficult to understand [13]. Therefore, this study aims to find suitable methods to enhance ES and retain the speaker's identity so that laryngectomy patients can have a better quality of life.

Since obtaining recordings of laryngectomy patients before the ablation operation is challenging, most researchers used voice conversion (VC) technology in their studies to enhance esophageal speech [2–4,6,11,12,14]. VC consists of manipulating the speech characteristics of the ES source speaker to replicate the vocal qualities of another healthy target speaker without altering the linguistic information, meaning, or content [15].

Many studies have been conducted to investigate this. An approach to increase the clarity and naturalness of ES using a statistical voice Esophageal-Speech-to-Speech conversion technique has been proposed in [6]. The objective of this method was to retain the linguistic content while changing esophageal speech into speech that sounds natural by applying probabilistic approaches. Another study proposed a voice conversion system based on Gaussian mixture models (GMMs) and deep neural networks (DNNs) to enhance the quality of ES [14]. This research aimed to lower speech noise and maintain the unique features of the esophageal utterer using the time-dilated Fourier cepstra on the source vocal tract. To accomplish this objective, GMMs were applied as the primary voice conversion technique, and the deep neural network was trained to function as a nonlinear mapping function. In addition, the study used a frame selection algorithm to predict the excitation and phase separately. The research conducted in [11] utilized a neuromimetic statistical method to enhance the naturalness and lucidity of ES. This method involved estimating the vocal tract and excitation cepstral coefficients independently. DNN model and GMMs were trained with the original cepstral vocal tract, while the phase and cepstral excitation were estimated by tracing the target training space through a KD-tree. The system proposed in [12] used an attention mechanism specifically for esophageal-to-laryngeal voice conversion by incorporating adaptive mapping between esophageal and healthy features. Also, the speaker's identity was preserved by determining the excitation and phase coefficients from a target-learning space, organized as a binary search tree, and queried using the vocal tract coefficients predicted by the Seq2Seq model. As per another research paper [4], a phase-based approach was introduced to improve ES. The method involves aligning the source and target phases using the Fast Dynamic Time Warping method (FastDTW) and then using the best alignment path as input to the deep neural network. Additionally, this technique preserves the identity of the source speaker by incorporating the cepstral coefficients of their vocal tract in the reconstruction of the improved sound file.

Similar to supervised learning, voice conversion models require copious amounts of data and a substantial quantity of labeled data to achieve high performance. However, acquiring these resources is costly and may result in the development of large models that are inefficient to utilize [16,17]. In addition, the speaker's identity may not be retained.

Moving to unsupervised learning for ES enhancement, the most used methods are based on traditional filtering, such as the Kalman filter [18] and the Wiener filter combined with wavelet-based methods [13]. These classical methods are known for producing artifacts called musical noises [19].

To address these challenges, our contribution involves utilizing Only-Noisy-Training (ONT) [20], an approach that enables the denoising of noisy wave files without requiring clean counterparts. To the best of our knowledge, this marks the first application of a self-supervised denoising technique to enhance ES while preserving the speaker's identity. In this paper, we have conducted experiments using two ONT-based methods, DCUNET and DCUNET-cTSTM [20], to determine whether these methods are effective for denoising and enhancing ES. Additionally, to compare and calculate the loss and metrics of the ONT systems, the pre-trained VoiceFixer model (VF) [21] was used to restore the clean sound files of ES. Our results demonstrate the feasibility of denoising tasks without conventional targets. Furthermore, DCUNET-cTSTM exhibited particularly notable performance.

The remainder of this manuscript is structured as follows: Section 2 describes the materials and methodologies used for denoising and comparison tasks. Section 3 provides a comprehensive summary of the experimental results. Section 4 engages in a detailed discussion of the outcomes. Lastly, Section 5 concludes the work.

2. Materials and Methods

2.1. Only-Noisy Training

In conventional speech-denoising tasks, it is common practice to use clean audio signals as a target for the training. However, acquiring such signals can be challenging due to their high cost or the need for specialized equipment or studio environments with strict requirements. To overcome this challenge, an end-to-end self-supervised method for speech denoising, called Only-Noisy-Training (ONT), was introduced [20]. This innovative approach aims to reduce reliance on clean signals during training. Unlike traditional methods, ONT utilizes only noisy audio signals for training, eliminating the need for additional conditions. The ONT framework comprises two key components: a module for generating training pairs and a module for speech enhancement. The first part employs a random audio sub-sampler to extract training pairs from each noisy audio signal. These pairs are then processed by a complex-valued speech denoising component, as illustrated in Figure 1.

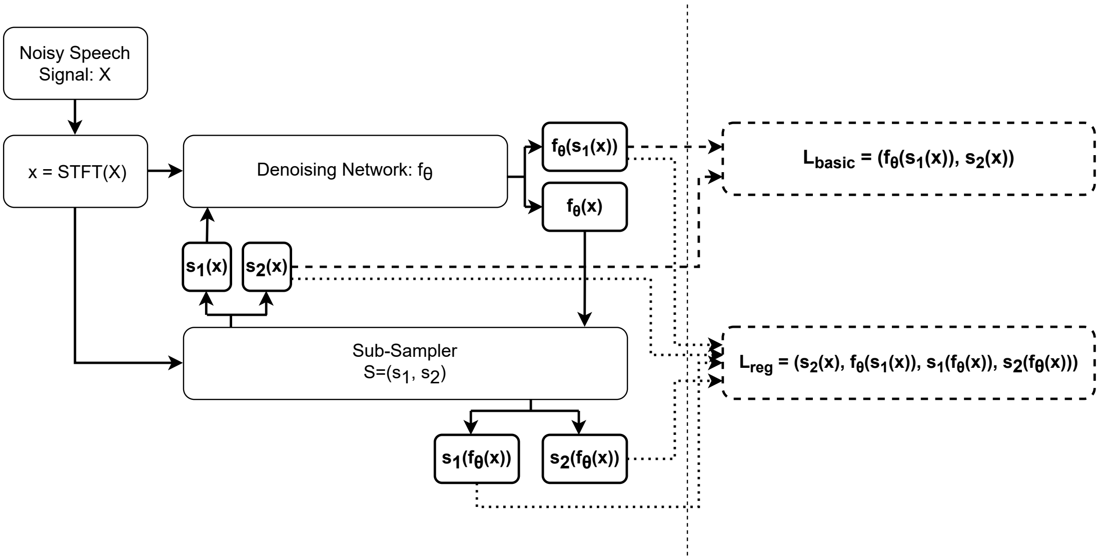
Noisy Speech Signal: X
x = STFT(X)
Denoising Network: f_θ
$f_\theta(s_1(x))$
$f_\theta(x)$
$L_{basic} = (f_\theta(s_1(x)), s_2(x))$
$s_1(x)$
$s_2(x)$
Sub-Sampler $S=(s_1, s_2)$
$L_{reg} = (s_2(x), f_\theta(s_1(x)), s_1(f_\theta(x)), s_2(f_\theta(x)))$
$s_1(f_\theta(x))$
$s_2(f_\theta(x))$

Figure 1. Schema of ONT model.

The denoising network incorporates state-of-the-art encoder and decoder modules based on the groundbreaking architecture of DCUNET-10.

2.1.1. Deep Complex UNET (DCUNET)

Deep Complex UNET architecture combines the advantages of deep complex networks [22] and UNET models, basically known as a convolutional autoencoder with skip connections [23]. DCUNET has been optimized to perform complex domain operations more efficiently and effectively [24]. The convolutional layers in the UNET architecture were replaced with complex convolutional layers, initialized according to Glorot initialization [25]. Additionally, complex-batch normalization was applied to every convolutional layer, excluding the last layer in the network. Each encoder of the network was composed of stridden complex convolutional layers, and the decoder was made of strided complex deconvolutional operations in order to restore the input size. To prevent the loss of spatial information, max-pooling operations were not used. For both the encoder and decoder, Leaky ReLU was the activation function. Figure 2 resumes the functioning of DCUNET.

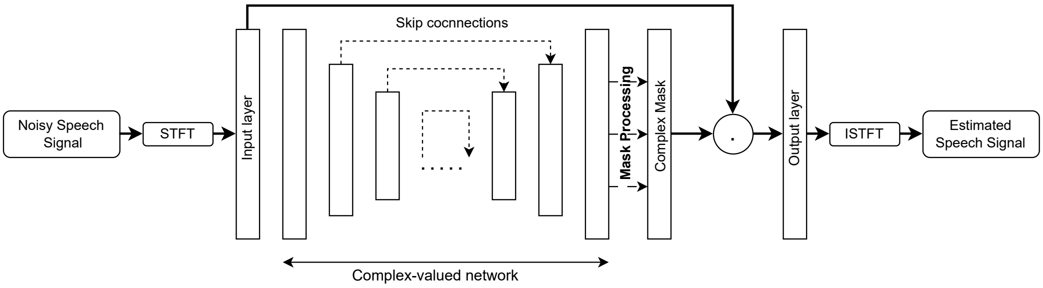

Figure 2. Schema of DCUNET architecture.

2.1.2. Deep Complex UNET with Complex Two-Stage Transformer Module (DCUNET-cTSTM)

The proposed complex-valued speech system uses a complex encoder and decoder. Each complex 2D convolution and 2D transposed convolution [22] used in the encoder and decoder comprises four conventional convolution operations, as shown in Equation (1).

$$X \odot Z = (X_r * Z_r - X_i * Z_i) + j(X_r * Z_i + X_i * Z_r) \tag{1}$$

where Z is the complex convolution kernel, and X is the complex input.

This model has a complex two-stage transformer (cTSTM) module that connects the encoder to the decoder. This architecture enables more precise processing and reconstruction of amplitude and phase information from spectrograms while preserving the contextual information of speech. Equation (2) denotes the operation to obtain the cTSTM's output.

$$O_{cTSTM} = (O_{\mathrm{rr}} - O_{\mathrm{ii}}) + j(O_{\mathrm{ri}} + O_{\mathrm{ir}}) \tag{2}$$

where O_{cTSTM} denotes the outcome of the cTSTM module, $O_{\mathrm{rr}} = TSTM_r(X_r)$, $O_{\mathrm{ii}} = TSTM_i(X_i)$, $O_{\mathrm{ri}} = TSTM_i(X_r)$, and $O_{\mathrm{ir}} = TSTM_r(X_i)$.

Figure 3 presents the DCUNET-cTSTM architecture.

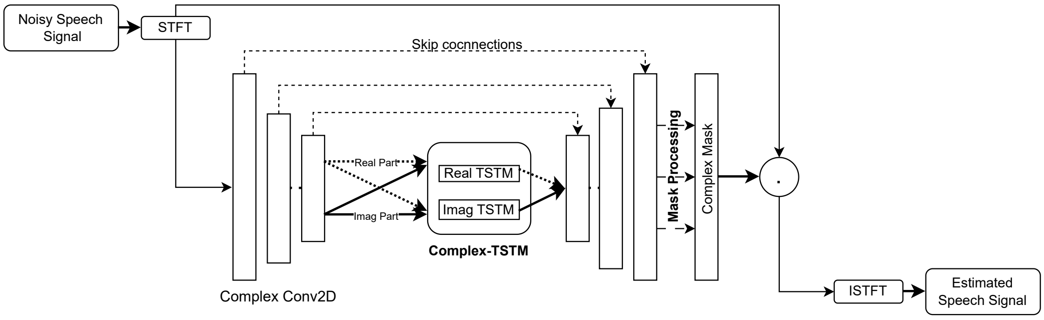

Figure 3. Schema of DCUNET-cTSTM architecture as proposed in [20].

2.2. Loss Functions

To calculate the training loss of the two models above, two different loss functions were used: the basic loss and the regularization loss, which were defined and described in [20] as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{basic}} + \gamma \cdot \mathcal{L}_{\mathrm{reg}} \tag{3}$$

where γ denotes a tunable parameter.

As formulated in Equation (4), the basic loss function consists of weighted SDR loss $\mathcal{L}_{\mathrm{wSDR}}$ [24], frequency domain loss $\mathcal{L}_{\mathrm{F}}$, and time domain loss $\mathcal{L}_{\mathrm{T}}$.

$$\mathcal{L}_{\mathrm{basic}} = (\alpha \cdot \mathcal{L}_{\mathrm{F}} + (1-\alpha) \cdot \mathcal{L}_{\mathrm{T}}) \cdot \beta + \mathcal{L}_{\mathrm{wSDR}} \tag{4}$$

α and β are hyper-parameters that regulate the strength of $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{T}}$.

The measure of the loss in the time domain is the average of the squared differences between the output and the clean waveforms.

$$\mathcal{L}_{\mathrm{T}} = \frac{1}{N} \sum_{i=0}^{N-1} (s_i - \hat{s}_i)^2 \tag{5}$$

In this context, s_i and $\hat{s}_i$ represent the i-th instance of unprocessed speech and its corresponding denoised version, with N referring to the total number of audio samples.

The model can gather additional information which can lead to improved speech intelligibility and better-perceived quality using frequency domain loss, which is given by Formula (6) as follows:

$$\mathcal{L}_{\mathrm{F}} = \frac{1}{T_F} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} \left[\left(|S_r(t,f)| + |S_i(t,f)| \right) - \left(|\hat{S}_r(t,f)| + |\hat{S}_i(t,f)| \right) \right] \tag{6}$$

S and $\hat{S}$ symbolize the clean and processed spectrogram, while r and i represent the real and imaginary components of the complex variable. T and F, on the other hand, denote the respective quantities of frames and frequency bins.

Also, $\mathcal{L}_{\mathrm{wSDR}}$, given by Equation (7), helps to improve the well-known assessment metrics specified in the temporal domain:

$$\mathcal{L}_{wSDR}(x, y, \hat{y}) = -\alpha \frac{\langle y, \hat{y} \rangle}{\|y\| \|\hat{y}\|} - (1 - \alpha) \frac{\langle x - y, x - \hat{y} \rangle}{\|x - y\| \|x - \hat{y}\|} \tag{7}$$

and

$$\alpha = \frac{\|y\|^2}{\|y\|^2 + \|x - y\|^2} \tag{8}$$

x represents a noisy sample, while y and $\hat{y}$ represent the target sample and the predicted outcome. The symbol α denotes the energy ratio between the target speech and the noise.

2.3. VoiceFixer

VoiceFixer (VF) is a proposed system [21] that aims to mimic speech analysis and comprehension of the human auditory system. VF uses advanced algorithms to simultaneously eliminate noise, distortions, and other issues in speech signals. Additionally, it can improve speech's intelligibility and clarity by enhancing the signal's spectral and temporal features. This restoration system consists of an analysis stage and a synthesis stage. The analysis stage was modeled by a ResUNet model (Mel spectrograms were utilized during this stage), and neural vocoder, which is a generative adversarial network (GAN) that operates in both time and frequency domains (TFGAN), was used to model the synthesis stage. VoiceFixer generalizes well to severely degraded real speech recordings. Figure 4 presents an overview of its architecture. VF was trained (as described in another paper [21]) on two speech databases (CSTR VCTK corpus and AISHELL-3) and two noise databases (VCTK-Demand and TUT).

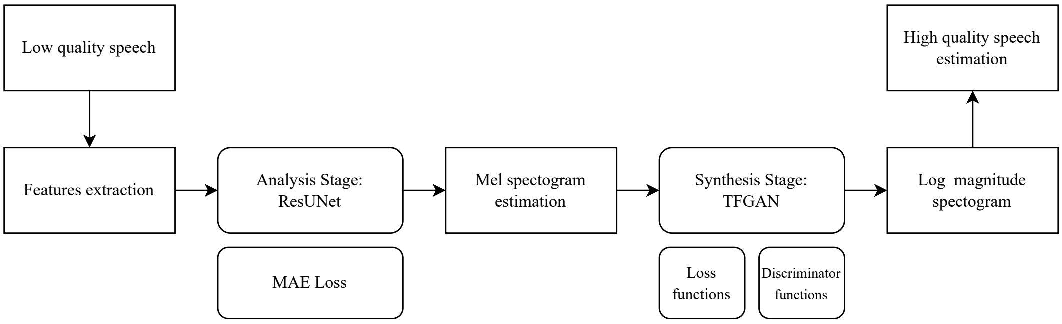

Figure 4. Overview of VoiceFixer architecture.

2.4. Dataset

This study aimed to evaluate the effectiveness of the ONT approach and DCUNET-based methods in processing esophageal speech. To accomplish this, a French database consisting of three parallel datasets uttered by three French male laryngectomy patients was used. Each corpus comprised 289 phonetically balanced sentences. This database will be referred to as (ES) in subsequent experiments.

In addition, the Voice Bank Dataset (VBD) [26] was used to train some of the experiments in this study. This database comprises around 400 English sentences spoken by 84 healthy males and females with different accents from both England and the United States.

The noises used in this assessment are white noise and one type of noise from the UrbanSound8K [27] dataset, which comprised nine different types of real-world noise.

In the experiments, VF, WN, and Ur8 denote the restored ES file dataset using the VoiceFixer, the white noise, and the UrbanSound8K dataset, respectively.

2.5. Evaluation Metrics

Various objective measures were used to assess the effectiveness of noise reduction. One of the simplest measures available is the signal-to-noise ratio (SNR), which aims to gauge the level of distortion in the waveform coders that reproduce an input waveform. The SNR is determined by the following formula:

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i)-y(i))^2} \tag{9}$$

where N is the total number of samples, $x(i)$ denotes the signal at sample i, and $y(i)$ represents the noise or interference at sample i.

Instead of working on the entire signal, the segmental signal-to-noise ratio (SSNR) computes the mean of the SNR values of the small sections (15–20 ms). It is given by

$$\mathrm{SSNR} = 10\log_{10}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i)-y(i))^2} \tag{10}$$

where N is the total number of samples, $x(i)$ represents the original signal at sample i, and $y(i)$ denotes the reconstructed or noisy signal at sample i.

The Perceptual Evaluation of Speech Quality (PESQ) is an objective method for assessing the quality of speech as received by a listener [28]. It evaluates speech quality by calculating the total loudness difference between the processed and clean signals [29]. The PESQ score fluctuates in the range of -0.5 to 4.5 [12]. Figure 5 presents a schema of the structure of PESQ.

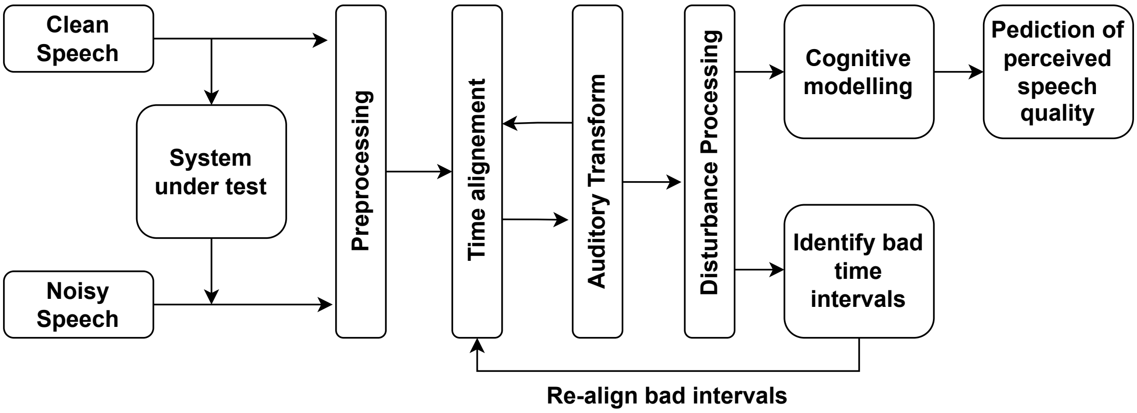

Figure 5. PESQ architecture.

The Wide-Band Perceptual Evaluation of Speech Quality (PESQ-WB) is an extension of the original PESQ. It assesses the quality of wide-band speech signals based on perceptual evaluation. It covers a wider frequency range, typically 50–7000 Hz, to accommodate modern wideband telephony and VoIP services [28].

The Narrow-Band Perceptual Evaluation of Speech Quality (PESQ-NB) is a metric used to evaluate the quality of narrow-band speech signals, such as those typically encountered in telephone networks and lower-bandwidth communication channels, mainly operating within the 300–3400 Hz frequency range [28].

The Short-Time Objective Intelligibility (STOI) metric is used to predict the intelligibility of the processed speech through enhancement or separation algorithms. STOI operates on short-time segments of the speech signal, typically around 400 milliseconds, and provides a score correlating with the intelligibility of the processed speech as perceived by human listeners [30]. Improved speech intelligibility is indicated by a higher STOI rating [12]. Figure 6 presents the structure of STOI.

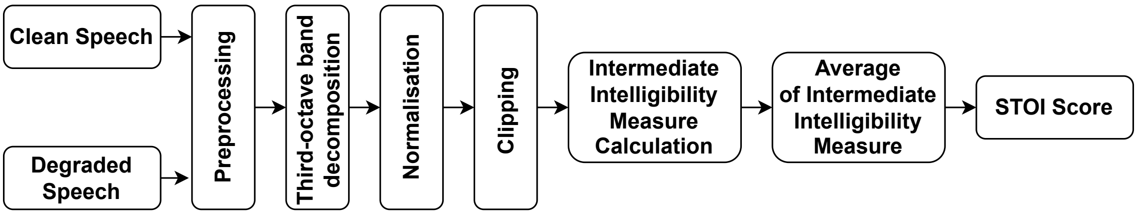
Clean Speech
Degraded Speech
Preprocessing
Third-octave band decomposition
Normalisation
Clipping
Intermediate Intelligibility Measure Calculation
Average of Intermediate Intelligibility Measure
STOI Score

Figure 6. STOI architecture.

We used the “pypesq” and “pystoi” PyTorch (1.10.0 + CPU on Python 3.9) packages to calculate the evaluation metrics.

2.6. Experimental Parameters

A comparative study was conducted between the DCUNET and DCUNET-cTSTM models using different databases. Initially, the pre-trained VF model was utilized to generate the restored speech files of ES files. These files served as clean wave files in the experiments and were also employed to calculate the loss and metrics of the ONT model.

Table 1 presents the experimental parameters used in the two models, and Table 2 gives the experiments realized within these two.

Table 1. Experimental parameters.

	DCUNET	DCUNET-cTSTM
Sampling rate	16 kHz	16 kHz
Window size	64 ms Hamming window	64 ms Hamming window
Number of layers	10	10
Number of channels	(45, 90, 90, 90, 90, 90, 90, 90, 45, 1)	(32, 64, 64, 64, 64, 64, 64, 64, 32, 1)
Kernel size	(3, 3)	(3, 3)
Step size	(2, 2)	(2, 2)
Step size for the middle two layers	(2, 1)	(2, 1)
Loss function parameters	$\alpha = 0.8$, $\beta = 1/200$, $\gamma = 1$	the same
Number of TSTM layers	None	6

Table 2. Different experiments on the DCUNET and DCUNET-cTSTM model using different training and test datasets.

Experiment	Training	Test
Experiment 1	Input and target: VBD + Ur8-9	Clean:VBD Noisy: VBD + Ur8-9
Experiment 2	Input and target: ES + WN	Noisy: ES + WN Clean:ES
Experiment 3	Input and target: VF + WN	Noisy: VF + WN Clean:VF
Experiment 4	Input and Target: ES	Noisy: ES Clean:VF

3. Results

Figure 7 shows the spectrograms of the original ES and VF wave files and their corresponding white noise wave files.

The spectrograms of both the original ES and the VF test files show clear frequency bands with high power density in yellow, suggesting a strong signal presence. For the noised ES file and the noised VF wave file, the spectrograms show a more uniform power distribution across frequencies with less defined bands than the original. This indicates the presence of noise, which has spread the signal power more uniformly, making the frequency bands less distinct.

Table 3 resumes the test outcomes of the different experiments described in Table 2, using the trained models of DCUNET and DCUNET-cTSTM. These test results were evaluated by objective metrics: SNR, SSNR, PESQ-WB, PESQ-NB, and STOI. The bold numbers denote the best results of the metrics.

After the training phase, we conducted a mismatch condition (for example, trained on VB8 and tested with ES), using the pre-trained models from the four experiments to denoise various speech test signals. Figure 8 illustrates the resulting spectrograms of an ES test wave file, while Figure 9 displays the testing outcomes with cleaned VF wave files. Audio samples of the results are presented at the following demo link: https://madipraise.github.io/ES/ (accessed on 12 April 2024).

Table 3. Results of the four experiments using the DCUNET and DCUNET-cTSTM.

Training	Model	SNR	SSNR	PESQ-NB	PESQ-WB	STOI
Experiment 1	DCUNET	−19.25 ± 21.93	−10 ± 15.84	1.01 ± 4.15	1.01 ± 3.15	0.17 ± 0.97
	DCUNET-cTSTM	**4.05 + 22.87**	**0.14 + 16.36**	**1.09 + 3.73**	**1.02 + 2.21**	**0.28 + 0.99**
Experiment 2	DCUNET	0.48 ± 14.82	−7.17 ± 10.17	1.16 ± 3.33	1.05 ± 2.36	0.43 ± 0.97
	DCUNET-cTSTM	**0.81 ± 15.61**	−**6.91 ± 10.17**	**1.24 ± 3.42**	**1.071 ± 2.61**	**0.44 ± 0.97**
Experiment 3	DCUNET	−2.97 ± 19.17	−7.02 ± 10.98	1.22 ± 3.09	1.02 ± 2.08	0.42 ± 0.99
	DCUNET-cTSTM	−**0.59 ± 19.81**	−3.91 ± 10.63	**1.25 ± 3.05**	1.036 ± 2.191	0.28 ± 0.98
Experiment 4	DCUNET	−15.73 ± −0.01	−6.44 ± −0.58	1.02 ± 1.55	1.01 ± 2.09	−0.17 ± 0.50
	DCUNET-cTSTM	−16.10 ± −0.01	−6.92 ± −0.60	1.02 ± 2.246	**1.020 ± 1.27**	−0.18 ± 0.50

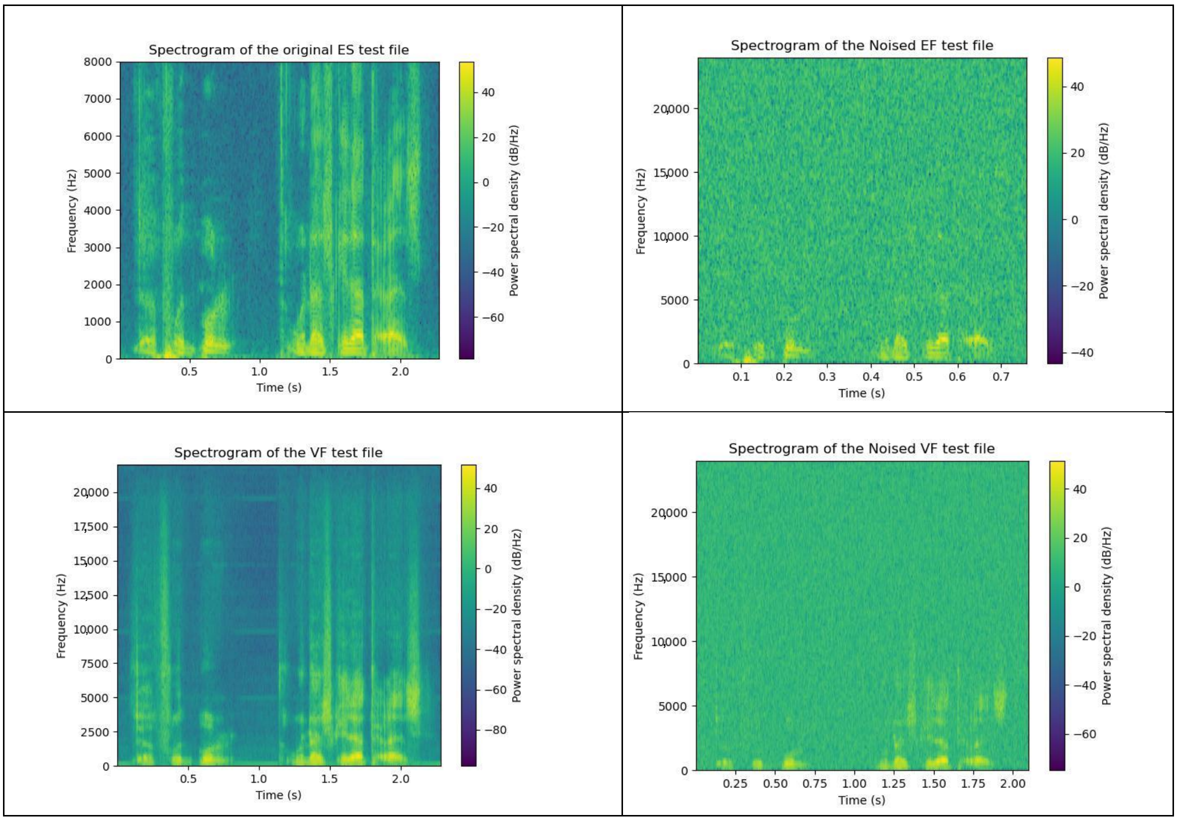

Figure 7. Spectrograms of ES and VF wave files, and of their corresponding noised wave files.

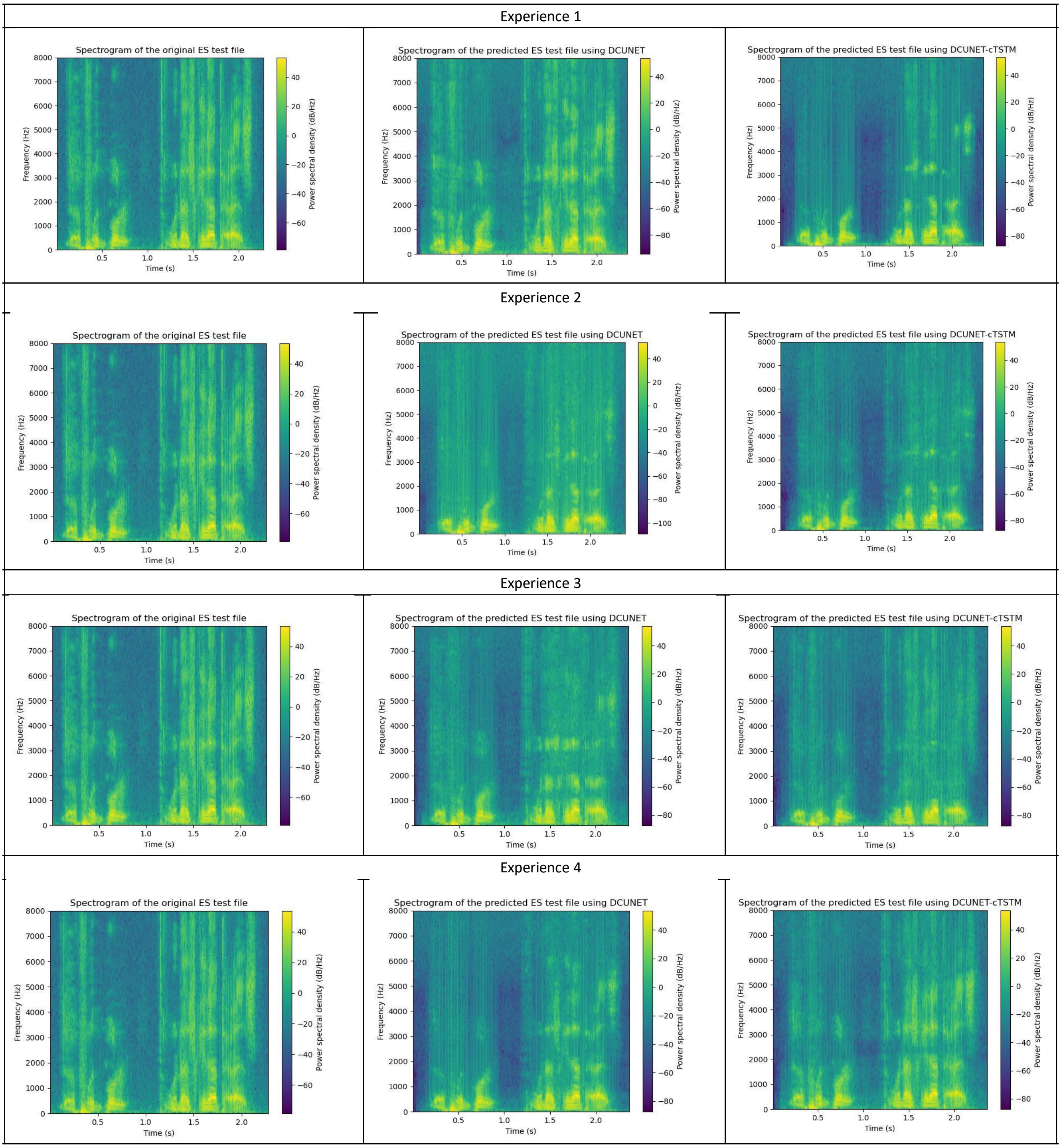

Figure 8. Spectrograms of the ES test file using the pre-trained models of the four experiments.

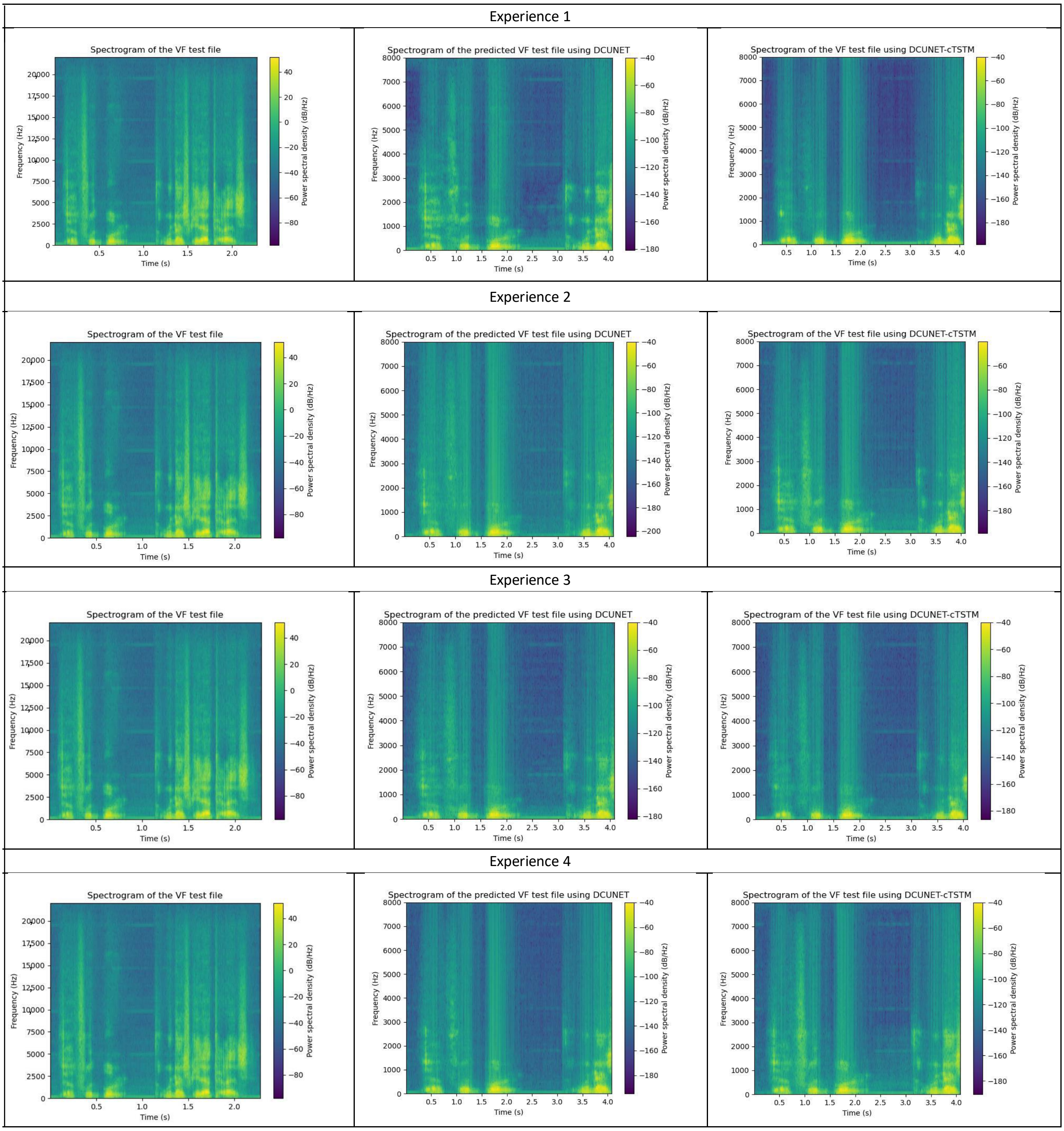

Figure 9. Spectrograms of the VF test file using pre-trained models of the four experiments.

4. Discussion

Based on the results shown in Table 3, it is evident that in experiments 1 and 2, the DCUNET-cTSTM model consistently outperformed the DCUNET model across all metrics. What is particularly remarkable is its exceptional performance when handling ES mixed with white noise and tested with clean ES. In experiment 3, the DCUNET-cTSTM model shows improved SNR and PESQ-NB compared to the DCUNET model, while other metrics remained broadly similar. Experiment 4 revealed slightly varied metric results for the two models, with the PESQ-WB of the DCUNET-cTSTM model notably surpassing that of the other model.

Moving to the results depicted in Figure 8, we start with the first experiment, which was trained with the VBD database mixed with noise number 9 of the Ur8 database (which corresponds to street music) and tested with the ES test file. The spectrogram of the original ES test file shows distinct frequency bands with high power density in yellow, indicating strong signal presence. Also, the spectrogram of the predicted ES test file using DCUNET shows improved clarity with distinct frequency bands, though some noise is still present. The power spectral density in the low-frequency regions remains similar to the original, but the higher frequencies seem slightly more uniform, indicating some noise reduction. The spectrogram of the predicted ES test file using DCUNET-cTSTM shows further improvement in clarity with more defined frequency bands. The power spectral density appears to be more similar to the original ES test file, suggesting better noise reduction compared to the DCUNET alone.

For experiment 2, in which the training was performed using ES wave files with additional white noise, the traces of the uniform noise distribution in the higher frequencies are still visible in the DCUNET result. In contrast, the DCUNET-cTSTM result exhibits a slightly clearer spectrogram. The low-frequency regions of both spectrograms resemble those of the original test file.

In the third experiment, which was trained using noisy VF sound files, some diffuse noise is still present in the DCUNET results. On the one hand, DCUNET appears to preserve more of the yellow regions in the low-frequency area than the second model's output. On the other hand, DCUNET-cTSTM seems to preserve more of the yellow information in the high-frequency region than the first model's outcome.

Lastly, in the final experiment, which was trained with the ES dataset without additional noise, the DCUNET's result shows that the voiced segments are clearer and more distinct, suggesting an improved SNR compared to the original ES wave file. The DCUNET-cTSTM's outcome demonstrates that the voiced segments are even brighter, substantially improving SNR compared to the original and the DCUNET prediction.

Referring to Figure 9, it can be seen that the original VF test file possesses a broad range of frequencies with considerable transient signals and noise distributed throughout the spectrum, in addition to high power density at both low and high frequencies. By employing DCUNET, noise reduction is effective, primarily in high-frequency regions. Consequently, the power density is concentrated below 2000 Hz, and transient noise is diminished, resulting in a clearer signal.

A comparison of DCUNET with DCUNET-cTSTM reveals that the latter exhibits superior noise reduction compared to the former. Furthermore, DCUNET-cTSTM displays a more focused energy distribution with clearer transient signals and minimal high-frequency noise. Therefore, DCUNET-cTSTM outperforms DCUNET in enhancing signal clarity and reducing noise. It should be noted that the VF wave file is the restored version of the ES wave file using VoiceFixer. As the DCUNET-based methods target noise reduction within sound files, it is a fact that the models eliminate a considerable amount of signal information, assuming that such segments represent noises.

5. Conclusions

In this study, self-supervised denoising methods were used to tackle the challenge of enhancing esophageal speech using the Only-Noisy-Training method. For this purpose, two DCUNET-based architectures were trained in different experimental scenarios. Also, for comparison purposes and to calculate the metrics, the pre-trained VoiceFixer model was used to restore clean versions of the original ES wave files. In order to obtain different pre-trained models, four scenarios were realized in the experiments. In the first scenario, the models were trained with the normal English VBD database mixed with street music. The second one consisted of the ES database mixed with white noise for training. The VF database mixed with white noise was used in the third scenario. And in the last one, the training was performed using the ES database without any additions. After having the pre-trained models, we used mismatched conditions for testing. The results show that DCUNET-cTSTM outperforms the results of DCUNET when testing with an ES sound file. The testing with a VF wave file shows that the DCUNET-based methods remove some information because those parts of the signal were considered noises. To sum up, this study has proven the effectiveness of the ONT approach along with the DCUNET-cTSTM model in denoising ES sound files. Moreover, this approach has successfully preserved the speaker's identity and proven that we can dispose of having healthy target wave files. An improved system based on these methods is under investigation, aiming for an even better enhancement for ES.

Author Contributions: Conceptualization, M.A. and E.H.I.E.; methodology, M.A. and E.H.I.E.; software, M.A.; supervision, E.H.I.E. and M.C.; validation, E.H.I.E. and M.C.; visualization, M.A.; writing—original draft, M.A.; writing—review and editing, E.H.I.E., M.C., K.E. and J.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hui, T.; Cox, S.; Huang, T.; Chen, W.-R.; Ng, M. The Effect of Clear Speech on Cantonese Alaryngeal Speakers' Intelligibility. *Folia Phoniatr. Logop.* **2021** , *74*, 103–111. [CrossRef] [PubMed]
2. Raman, S.; Sarasola, X.; Navas, E.; Hernaez, I. Enrichment of Oesophageal Speech: Voice Conversion with Duration–Matched Synthetic Speech as Target . *Appl. Sci.* **2021**, *11*, 5940. [CrossRef]
3. Dinh, T.; Kain, A.; Samlan, R.; Cao, B.; Wang, J. Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 4781–4785.
4. Amarjouf, M.; Bahja, F.; Di-Martino, J.; Chami, M.; Ibn-Elhaj, E.H. Predicted Phase Using Deep Neural Networks to Enhance Esophageal Speech. In Proceedings of the 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023), Marrakesh, Morocco, 5–7 March 2023; Lecture Notes on Data Engineering and Communications Technologies; Springer Nature: Cham, Switzerland, 2023; Volume 164, pp. 68–76.
5. Huang, T.Y.; Lin, B.S.; Lien, C.F.; Yu, W.H.V.; Peng, Y.Y.; Lin, B.S. A Voice-Producing System with Naturalness and Variable Multi-Frequency Vocalization for Patients Who Have Undergone Laryngectomy. *IEEE Access* **2023**, *11*, 30619–30627 [CrossRef]
6. Doi, H.; Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Esophageal Speech Enhancement Based on Statistical Voice Conversion with Gaussian Mixture Models. *IEICE Trans. Inf. Syst.* **2010**, *93*, 2472–2482. [CrossRef]
7. Yamamoto, K.; Toda, T.; Doi, H.; Saruwatari, H.; Shikano, K. Statistical Approach to Voice Quality Control in Esophageal Speech Enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4497–4500.
8. Caeiros, A.V.M.; Meana, H.M.P. Esophageal Speech Enhancement Using a Feature Extraction Method Based on Wavelet Transform. In *Modern Speech Recognition Approaches with Case Studies*; IntechOpen: London, UK, 2012.
9. Serrano García, L.; Raman, S.; Hernáez Rioja, I.; Navas Cordón, E.; Sanchez, J.; Saratxaga, I. A Spanish Multispeaker Database of Esophageal Speech. *Comput. Speech Lang.* **2021**, 66, 101168. [CrossRef]
10. Ouattassi, N.; Benmansour, N.; Ridal, M.; Zaki, Z.; Bendahhou, K.; Nejjari, C.; Cherkaoui, A.; El Alami, M.N.E.A. Acoustic Assessment of Erygmophonic Speech of Moroccan Laryngectomized Patients. *Pan Afr. Med. J.* **2015**, 21, 270. [CrossRef] [PubMed]
11. Ben Othmane, I.; Di Martino, J.; Ouni, K. Enhancement of Esophageal Speech Using Statistical and Neuromimetic Voice Conversion Techniques. *J. Int. Sci. Gen. Appl.* **2018**, *1*, 10.
12. Ezzine, K.; Di Martino, J.; Frikha, M. Intelligibility Improvement of Esophageal Speech Using Sequence-To-Sequence Voice Conversion with Auditory Attention. *Appl. Sci.* **2022**, 12, 7062. [CrossRef]
13. Amarjouf, M.; Bahja, F.; Martino, J.D.; Chami, M.; Elhaj, E.H.I. Denoising Esophageal Speech Using Combination of Complex and Discrete Wavelet Transform with Wiener Filter and Time Dilated Fourier Cepstra. In *ITM Web of Conferences, Proceedings of the 4th International Conference on Computing and Wireless Communication Systems (ICCWCS), Tangier, Morocco, 21–23 June 2022*; EDP Sciences: Les Ulis, France, 2022; Volume 48, p. 03004.

14. Ben Othmane, I.; Di Martino, J.; Ouni, K. Enhancement of Esophageal Speech Obtained by a Voice Conversion Technique Using Time Dilated Fourier Cepstra. *Int. J. Speech Technol.* **2018**, 22, 99–110. [CrossRef]
15. Zhang, M.; Wang, X.; Fang, F.; Li, H.; Yamagishi, J. Joint Training Framework for Text-To-Speech and Voice Conversion Using Multi-Source Tacotron and WaveNet. *arXiv* **2019**, arXiv:1903.12389.
16. Huang, Z.; Watanabe, S.; Yang, S.; Garcia, P.; Khudanpur, S. Investigating Self-Supervised Learning for Speech Enhancement and Separation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6837–6841.
17. Walczyna, T.; Piotrowski, Z. Overview of Voice Conversion Methods Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 3100. [CrossRef]
18. Ruiz, I.; Garcia, B.; Mendez, A.; Villanueva, V. Oesophageal speech enhancement using Kalman filters. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, Giza, Egypt, 15–18 December 2007; pp. 1176–1179.
19. Lan, C.; Wang, Y.; Zhang, L.; Liu, C.; Lin, X. Research on Speech Enhancement Algorithm of Multiresolution Cochleagram Based on Skip Connection Deep Neural Network. *J. Sens.* **2022**, *2022*, e5208372. [CrossRef]
20. Wu, J.; Li, Q.; Yang, G.; Li, L.; Senhadji, L.; Shu, H. Self-Supervised Speech Denoising Using Only Noisy Audio Signals. *Speech Commun.* **2023**, *149*, 63–73. [CrossRef]
21. Liu, H.; Kong, Q.; Tian, Q.; Zhao, Y.; Wang, D.; Huang, C.; Wang, Y. VoiceFixer: Toward General Speech Restoration with Neural Vocoder. *arXiv* **2021**, arXiv:2109.13731.
22. Trabelsi, C.; Bilaniuk, O.; Zhang, Y.; Serdyuk, D.; Subramanian, S.; Felipe Santos, J.; Santos, F.; Mehri, S.; Rostamzadeh, N.; Bengio, Y.; et al. Deep Complex Networks. In Proceedings of the ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
24. Kashyap, M.; Tambwekar, A.; Manohara, K.; Natarajan, S. Speech Denoising Without Clean Training Data: A Noise2Noise Approach. *arXiv* **2021**, arXiv:2104.03838.
25. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; JMLR Workshop and Conference Proceedings; JMLR: New York, NY, USA, 2010; pp. 249–256.
26. Valentini-Botinhao, C. *Noisy Speech Database for Training Speech Enhancement Algorithms and TTS Models*; Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh: Edinburgh, UK, 2017. [CrossRef]
27. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 1041–1044. Available online: https://urbansounddataset.weebly.com/urbansound8k.html (accessed on 2 February 2024).
28. 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 Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.
29. Ma, J.; Hu, Y.; Loizou, P.C. Objective Measures for Predicting Speech Intelligibility in Noisy Conditions Based on New Band-Importance Functions. *J. Acoust. Soc. Am.* **2009**, *125*, 3387–3405. [CrossRef] [PubMed]
30. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.