applied sciences

MDPI

Article

Intelligibility Improvement of Esophageal Speech Using Sequence-to-Sequence Voice Conversion with Auditory Attention

Kadria Ezzine [1,2,*], Joseph Di Martino [3] and Mondher Frikha [2]

[1] National Engineering School of Carthage, Carthage University, 45 Rue des Entrepreneurs Charguia II, Tunis 2035, Tunisia
[2] ATISP—Advanced Technologies for Image and Signal Processing, ENET'COM, University of Sfax, Sfax 3000, Tunisia; mondher.frikha@enetcom.usf.tn
[3] LORIA—Laboratoire Lorrain de Recherche en Informatique et ses Applications, B.P. 239, 54506 Vandœuvre-lès-Nancy, France; joseph.di-martino@loria.fr
* Correspondence: kadria.ezzine@gmail.com

Abstract: Laryngectomees are individuals whose larynx has been surgically removed, usually due to laryngeal cancer. The immediate consequence of this operation is that these individuals (laryngectomees) are unable to speak. Esophageal speech (ES) remains the preferred alternative speaking method for laryngectomees. However, compared to the laryngeal voice, ES is characterized by low intelligibility and poor quality due to chaotic fundamental frequency *F*0, specific noises, and low intensity. Our proposal to solve these problems is to take advantage of voice conversion as an effective way to improve speech quality and intelligibility. To this end, we propose in this work a novel esophageal–laryngeal voice conversion (VC) system based on a sequence-to-sequence (Seq2Seq) model combined with an auditory attention mechanism. The originality of the proposed framework is that it adopts an auditory attention technique in our model, which leads to more efficient and adaptive feature mapping. In addition, our VC system does not require the classical DTW alignment process during the learning phase, which avoids erroneous mappings and significantly reduces the computational time. Moreover, to preserve the identity of the target speaker, the excitation and phase coefficients are estimated by querying a binary search tree. In experiments, objective and subjective tests confirmed that the proposed approach performs better even in some difficult cases in terms of speech quality and intelligibility.

Keywords: esophageal speech; intelligibility; voice conversion; sequence-to-sequence; attention mechanism; speech quality

Citation: 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. https://doi.org/10.3390/app12147062

Academic Editors: José A. González-López, Heidi Christensen and Inma Hernaez Rioja

Received: 15 June 2022
Accepted: 11 July 2022
Published: 13 July 2022

1. Introduction

In extensive cancers of the larynx or hypopharynx (T3 and T4 tumors), total laryngectomy remains the procedure of choice and the most reliable surgery for advanced laryngeal cancers. After this surgery, the vocal cords are completely removed and the respiratory tract is separated from the digestive tract. It is therefore essential to train laryngectomee patients to speak with an alternate voice without the vocal cords. Several techniques [1,2] exist that allow vocal rehabilitation through the acquisition of a substitute voice that is learned with the help of a speech therapist or in a specialized rehabilitation center. Among the well-known and widely used mechanisms, the esophageal voice remains the ideal replacement for the laryngeal voice according to [3]. The principle of the esophageal speech (ES) is based on the use of a pharyngo-oesophageal digestive segment as a neovibrator. To produce this voice, it is necessary to introduce oral air into the top of the esophagus and release it under control. Thus, as mentioned above, ES is the ideal substitute for the laryngeal voice. However, and as expected, the quality and intelligibility of ES are

influenced by the change in the mechanism of speech production. This change of course affects the acoustic features of the esophageal voice, which are very different from those of the laryngeal voice. Compared to normal speech, esophageal is characterized by poor intelligibility and poor quality due to chaotic fundamental frequency, specific noises that resemble belching, and low intensity.

All of these instabilities in the acoustic characteristics produce poor-quality sounds that are difficult to understand.

Due to the extensive use of the esophageal voice by laryngectomees, this type of voice has been the subject of numerous studies in the last few years. To our knowledge, the existing approaches for ES quality improvements can be summarized into three categories: approaches based on the transformation of acoustic features, such as formant synthesis [4], comb filtering [5], and smoothing of acoustic parameters [6]; approaches based on statistical techniques, where [7–9] have been carried out, and approaches based on the VC technique, which allows for the transformation of the voice of a source speaker (laryngectomee) into that of a target speaker (laryngeal) [10–16]. Although these approaches have of course improved the estimation of the acoustic characteristics to reconstruct a converted signal with better quality, the improvements in intelligibility and naturalness are still insufficient. First, most previous studies deal primarily with transforming the spectral envelope and F0 trajectories by simply adjusting them linearly in the logarithm field [5–7]. In addition, since the acoustic models were constructed frame by frame, the duration of the converted sequences was kept the same as that of the source sequences. However, the production of human speech is a highly dynamic process, and the frame-by-frame assumption makes the modeling ability of the conversion functions limited [17].

Moreover, temporal alignment is another problem when converting esophageal to laryngeal speech. As feature alignment is necessary for VC systems, the Dynamic Time Warping (DTW) algorithm [18] is most frequently used by researchers. This algorithm makes it possible to align the characteristics using a dynamic programming algorithm where the acoustic features are not taken into account, which may cause problems, especially for pathological alignments [19]. These aligned features are subsequently used in the learning stage, to train the model, which may result in poor quality and intelligibility of the converted speech. Thus, the results of the VCC2018 [20] and the VCC2020 [13] proved that there is still much research to be done to improve the quality and naturalness of the converted speech.

Recently, RNN-based Seq2Seq learning [21] has proven to be an outstanding technique for mapping one sequence to another one, especially in the field of VC [22], text-to-speech synthesis (TTS) [23], and natural language processing [24]. As far as we know, converting spectral features using a Seq2Seq model with attention has been attempted for the first time in [25]. However, in his work [26], the author mentioned that the proposed model cannot use its own predictions to generate correct output predictions. A gated CNN-based Seq2Seq spectrum conversion method has been proposed in [27]. Due to the lack of an attention mechanism in their model, they still use the DTW algorithm to obtain aligned feature sequences when preparing training data. The work [28] proposed a method based on a Seq2Seq learning with attention and context preservation mechanism for voice conversion tasks. In their method, taking into account the attention module and the context preservation losses makes it possible to further stabilize the training procedure.

A recent method was proposed in [29] for the enhancement of whisper-to-normal speech based on a Seq2Seq model. In this method, taking into account the attention technique makes it possible to further stabilize the training procedure, outperforming the conventional methods.

Nevertheless, significantly improving the intelligibility and the naturalness of the ES by overcoming the previously cited problems remains challenging.

In this work, we propose a novel esophageal-to-laryngeal VC system based on a Seq2Seq mapping model combined with an attention mechanism. Our work is inspired by the latest work [29] and the first work [25]. The strength of the proposed method is

that it can adaptively characterize the nonlinear mapping between features of the original esophageal speech and its laryngeal counterpart. Additionally, our method does not require any temporal alignment during the training phase, which avoids erroneous mappings and significantly reduces the computing time. Furthermore, to preserve the identity of the target speaker, the excitation and phase coefficients are estimated from the target training space structured as a binary search tree.

The remainder of the paper is organized as follows. Section 2 presents the proposed methods. Section 3 details the experimental setup. Section 4 presents the results and discussion. Finally, a conclusion is given in Section 5.

2. Methods

2.1. System Overview

Figure 1 shows the framework of our proposed Seq2Seq esophageal-to-laryngeal speech conversion. The conversion process is divided into two main phases: training and conversion.

First is a training phase, during which the utterances pronounced by esophageal and normal speakers, undergo a step of parameterization to extract efficient representations of both signals. Then, a standard normal distribution is adopted to normalize spectral features, and the mean and standard deviation are subsequently recorded. Next, the spectral features of ES are sent to the encoder network for training. In this step, the encoder outputs are pre-trained and a loss function is computed between the encoder output and the representation of laryngeal speech. In the end, the decoder with an attention mechanism is used to improve the quality and accuracy of the encoder outputs. Meanwhile, the encoder and decoder networks train as a whole with an optimizer for each of them, and the error is calculated and back-propagated frame-by-frame using a loss function.

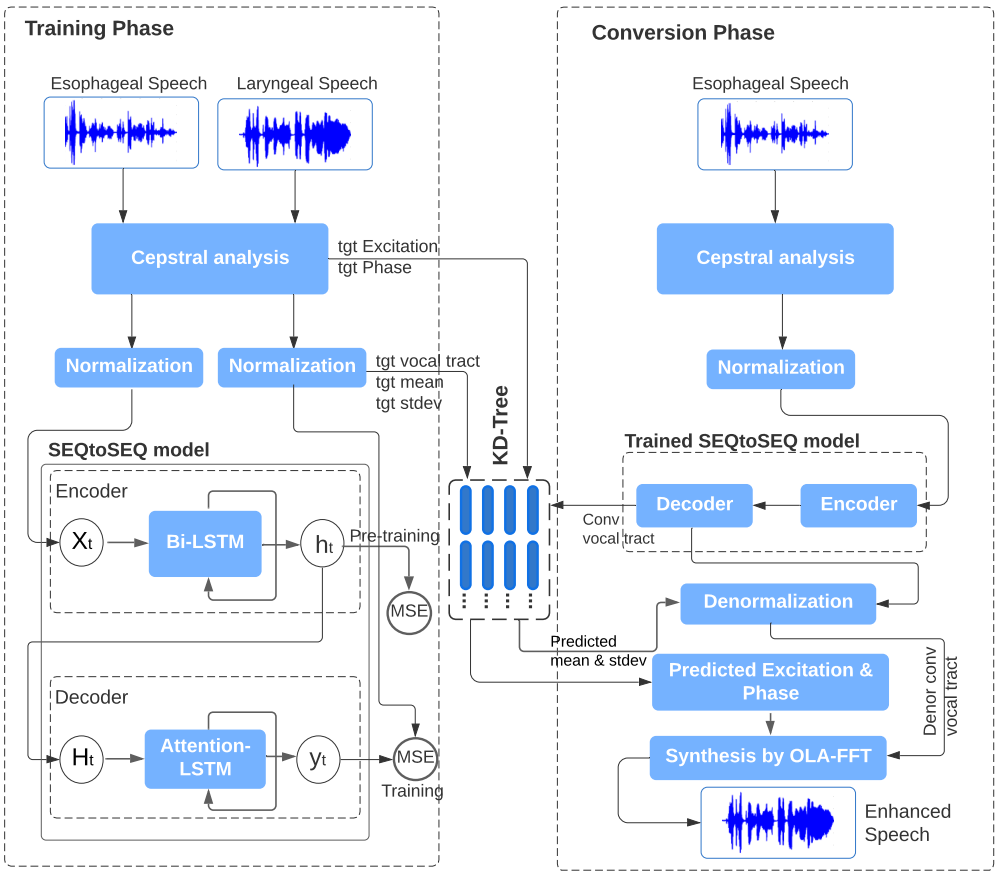

Figure 1. An overview of the proposed esophageal speech enhancement framework based on a Seq2Seq model.

In the conversion phase, cepstral coefficients are firstly extracted from each ES signal and then normalized. Next, the trained Seq2Seq model is applied to convert only the first cepstral packet (vocal tract feature vectors) from the source speaker into their approaching target. After this, to preserve the identity of the target speaker, we propose to predict cepstral excitation and phase coefficients from the target training space by using a KD-tree algorithm [30]. The binary KD-tree is constructed with the cepstral frames of the laryngeal vocal tract. Then, it is queried by the converted vocal tract ceptral vector obtained by the Seq2Seq model in order to find an index indicating the nearest target vocal tract vector. This index then serves as an index of the desired cepstral excitation and the desired phase vector. Finally, the same converted vocal tract cepstral vectors are denormalized according to the recorded mean and standard deviation. This denormalization is the reverse of the normalization process with the recovery of the original shape, which is why we utilize the recorded values. These denormalized vocal tract vectors are then used to synthesize enhanced speech.

In the resynthesis step, the magnitude and phase spectra are used to create complex spectra. Then, the enhanced speech is reconstructed by the short-term OLA-FFT (overlap-add method), which consists of applying an inverse fast Fourier transform (IFFT) to the complex spectra.

2.2. Feature Extraction and Normalization

In this article, we find it reasonable to consider the technique of cepstral analysis for feature extraction because it allows us to separate excitation from the vocal tract. Thus, we adopt Fourier cepstra at each frame that forms the input sequence $X = [X_1, \ldots, X_n, \ldots, X_N]$ of the Seq2Seq model, where N defines the frame number of the ES signal.

The real Fourier cepstra of the esophageal and target speech are obtained by computing the inverse Fourier transform (IFFT) of the logarithm of the magnitude short-time spectra. The mathematical formula for the extraction of acoustic features is given by the following equation:

$$C[n] = IFFT(Log|FFT(x[n] * H[n])|) \tag{1}$$

where $H(n)$ is a normalized Hamming window [31] of length equal to 512 in this work.

For each time frame, the linguistic contents are encoded into a set of coefficients: vocal tract cepstral vector as $[vt_0 \ldots vt_{32}]$, cepstral excitation as $[ex_{33} \ldots ex_{256}]$, and phase coefficients as $[ph_0 \ldots ph_{256}]$.

In the normalization step, we normalize the acoustic features in order to obtain a standard normal distribution and to control the amplitude of the gradients during training. For each component $x_{i,n}$ of index i at frame n, we subtract the mean and divide the result by the standard deviation according to Formula (2).

$$x'_{i,n} = \frac{(x_{i,n} - \mu_i)}{\sigma_i} \tag{2}$$

where μ_i and σ_i are, respectively, the mean and the standard deviation (stdev) of the ith component in all frames of the training sample. $x'_{i,n}$ represents the normalized vector.

2.3. Network Model

The proposed framework is a Seq2Seq model with an attention mechanism, consisting of two main components: a stack of a bidirectional LSTM encoder and an LSTM decoder based on an attention network. Figure 2 shows the overall network architecture of the designed model.

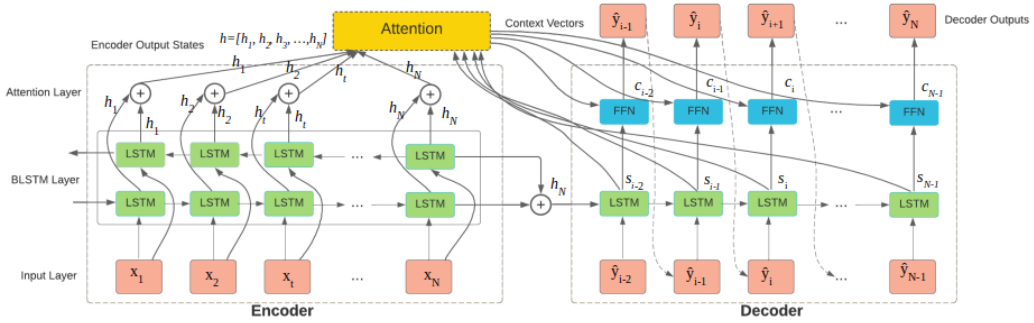

Figure 2. Structure of the proposed Seq2Seq model, where the attention layer is ignored for clarity.

Let $X^{(s)} = [x_1^{(s)}, \cdots x_{N_s}^{(s)}]$ and $Y^{(t)} = [y_1^{(t)}, \cdots y_{N_t}^{(t)}]$ represent sequences of cepstral features of the ES and laryngeal speech of non-aligned parallel utterances, where N_s and N_t denote the length (frame number) of the source and target sequences, respectively. Note that these sequences do not necessarily have the same length (i.e., generally, $N_s \neq N_t$).

We consider a Seq2Seq framework that models the mapping relationship between $X^{(s)}$ and $X^{(t)}$. As with most sequence-to-sequence models that were originally proposed by Cho et al. for machine translation applications [31], our Seq2Seq model has an encoder–decoder structure equipped with an attention mechanism.

The encoder network processes the input sequence of esophageal speech into a high-dimensional feature space in order to facilitate the decoding process. Then, it compresses and summarizes them into a fixed-length context vector, also called "hidden representation $h = [h_1, \ldots, h_t, \ldots, h_{N_h}]$", which represents all the information concerning the source features:

$$h = Encoder(X^{(s)}; \theta_{encoder}), \tag{3}$$

where $\theta_{encoder} = [W_{encoder}, U_{encoder}, b_{encoder}]$ are the parameters of the encoder model.

At each time step i, the decoder takes the output of the encoder, i.e., the hidden representations h_t, and the previously generated features y_{i-1} considered as inputs, and applies its attention mechanism network and then progressively generates the current enhanced output y_i.

$$y_i = Decoder(h_t, y_{i-1}; \theta_{decoder}), \tag{4}$$

Note that, in the typical Seq2Seq framework, the last hidden layer is used as the context vector once the entire sequence is fully encoded. However, in our case, the length of the ES feature sequence is generally longer than that of the normal speech, so the ES feature vectors of length N_x used to obtain the target feature vectors of length N_y are dynamically changed at different time steps.

To model these various nonlinear mapping relationships more accurately, we adopt an attention mechanism to obtain a self-adaptive context vector to adaptively estimate the decoder output.

2.3.1. Bidirectional-LSTM-Based Encoder

Due to the long sequences to model and the large number of time steps during training, we used a bidirectional encoder to better understand the time dependencies between the two ends of the sequences. There, our encoder network consists of three layers: a linear layer and two Bidirectional Long Short-Term Memory (BiLSTM) layers, which are arranged incrementally, as shown in Figure 2.

As the LSTM architecture shows in Figure 3, three gates control the data flow: an input gate and a forget gate, where the information is stored or vanished from the memory cell c_t, which are represented by i_t and f_t, at the tth time step, respectively; and an output gate denoted by o_t, which controls the output state (also called the "hidden state") (h_t). The LSTM propagation is formulated as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{8}$$

$$h_t = o_t \odot tanh(c_t) \tag{9}$$

where x_t is the input vector at the tth time step, c_t is the current long-term memory cell, $\sigma()$ is the sigmoid function, $\odot$ represents the element-wise multiplication, and W_* (i.e., W_i, W_f, W_o, W_c) are the model parameters that map from input dimension to hidden dimension, while U_* maps from the previous hidden dimension to the current hidden dimension.

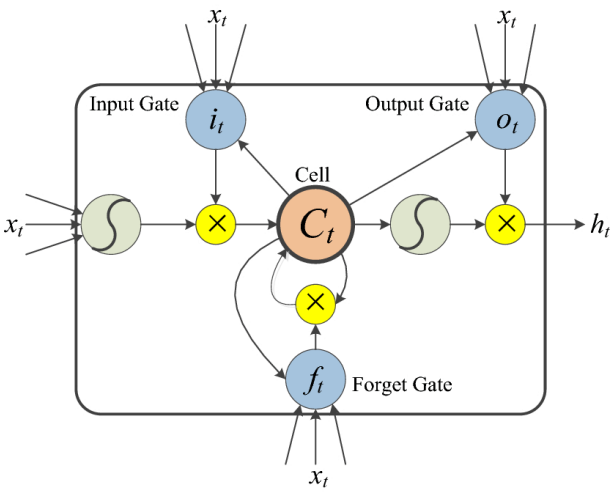

Figure 3. Internal structure of the LSTM unit [32].

After processing, each BiLSTM time step t will generate two hidden states:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{10}$$

where $\overrightarrow{h_t}$ is the forward LSTM network vector, which produces the high-dimensional features of the input signal, and $\overleftarrow{h_t}$ is the backward or inverse LSTM network output vector, given, respectively, by Formulas (11) and (12).

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(h_{t-1}, x_t, c_{t-1}) \tag{11}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t+1}, x_t, c_{t+1}) \tag{12}$$

2.3.2. Attention Mechanism-Based Decoder

Based on the hidden representation h of the encoder explained in the section above, we have redesigned the decoder that can predict the output cepstral features.

In the conventional Seq2Seq framework, the decoder adopts the output cell state c_t of the last hidden layer as its context vector. However, for esophageal speech conversion, the length of the normal feature sequence is always shorter than that of the esophageal feature sequence. In this kind of sequence, we can observe a lot of singularities, i.e., different ES phonemes can be linked to more than one phoneme at different positions of the target speech.

To cope with this problem, an attention mechanism is adopted, which more efficiently models these different nonlinear relationships. Therefore, this attention mechanism allows

for greater flexibility and corrects the alignment by adaptively estimating the decoder output through the constructed self-adaptive context vector.

In our proposed framework, we suggest that the current state of the decoder is fully connected to all hidden states of the encoder, and each hidden state has a different effect on the estimation of the decoder's current state. Therefore, to estimate the current self-adaptive context of the decoder, all backward and forward hidden states of the encoder are considered simultaneously.

Our designed decoder is illustrated in Figure 4, with step = 1 for clarity.

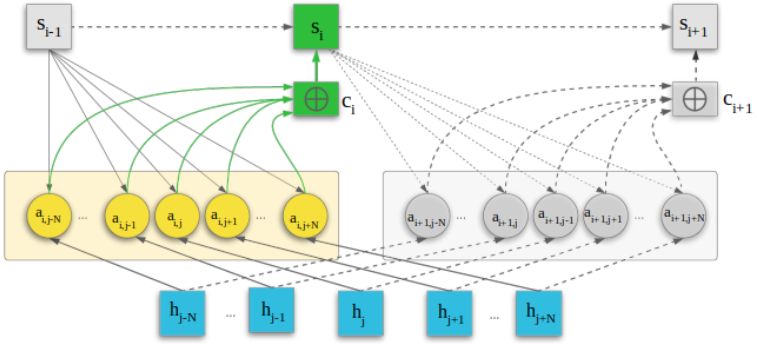

Figure 4. Illustration of attention layer, where $a_{i,j}$ = attention weight, c_i = context vector, s_i = attention target state, and $\bigoplus$ = concatenation.

At each ith decoder step, the attention output (context vector) c_i is computed using a weighted linear combination of the encoder's hidden states, as shown in Figure 4.

$$c_i = \sum_{j=1}^{j=N} \alpha_{ij} h_j, \tag{13}$$

where h_j represents the encoder's hidden state at position j, α_{ij} are the attention weights as shown by Formula (14).

$$\alpha_{ij} = \frac{\exp\left(score_{ij}\right)}{\sum_{j=1}^{N} \exp\left(score_{ij}\right)} \tag{14}$$

$score_{ij}$ is the attention score, which calculates the similarity between the encoder's hidden state and the previous cell state of the decoder, computed according to Formula (15).

$$score_{ij} = att\left(\left(s_{i-1}, h_j\right); \theta_{attention}\right) \tag{15}$$

where att is a feed-forward neural network (FFN) that produces the alignment scores between h_j and the previous state of the decoder's output s_{i-1}. $\theta_{attention}$ are the trainable parameters of the model.

Once the context vector c_i is produced, the decoder model takes this context vector, the previous decoder hidden state s_{i-1}, and the input $\hat{y}_i$ to generate the current decoder hidden state s_i.

$$s_i = LSTM\left(\left(s_{i-1}, \hat{y}_i, c_i\right); \theta_{decoder}\right) \tag{16}$$

The concatenation of the last context vector c_i and the output of decoding LSTMs s_i is linearly projected to produce the cepstrum output of the decoder network.

Thus, the decoder output is calculated according to the previous Equation (16) as:

$$\hat{y}_{i+1} = fc\left(s_i, c_i\right) \tag{17}$$

where fc is a fully connected FFN that allows mapping of the hidden dimension to the output dimension.

2.4. Loss Function

During training, the entire model was trained by the Mean Squared Error (MSE) loss function, which is calculated between the predicted and target cepstral vectors to evaluate their similarity. The formula is as follows:

$$MSE_{loss}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (18)$$

where $\hat{y}_i$ and y_i are the ith dimensions of the predicted and the target cepstral vectors, respectively. θ is the model parameters, and N is the total sample number. The decoder is optimized step by step, i.e., for each step, the loss is back-propagated to calculate the gradients of the weight parameters; then, the encoder and decoder optimizers sequentially update their own parameters. During training, the network processes speech utterances in batch style to make the calculated gradient more stable. Furthermore, the model is optimized via back-propagation using an $L1$ and $L2$ loss. The training is repeated for several epochs until convergence.

3. Experiments

3.1. Speech Datasets

For the experiments, we used our ES datasets. The recording of esophageal voices and the storage of the acoustic data were carried out at Loria laboratory in Nancy, France.

We recorded 289 phonetically balanced sentences spoken by two male French laryngectomees (PC and MH). We also recorded the same sentences spoken by a non-laryngectomee (AL), a French speaker. The audio signals were all sampled at 16 kHz and directly stored in wave files. Each file contained a single sentence and lasted between 3 and 5 s. Thus, in total, the data collected included 867 acoustic data files.

We utilized two pairs of parallel corpora (source and target speakers) for the training and evaluation of the proposed approach, namely PC and MH as source speakers and AL as the target speaker.

- PC (ES male) and AL (NS male)
- MH (ES male) and AL (NS male)

Note that we have trained our Seq2Seq mapping model for each pair of speakers, independently.

3.2. Experimental Setups

In this work, the two parallel corpora have been separated into 100 pairs of utterances for training, 20 pairs of utterances for validation, and 22 pairs of utterances for testing. We used the cepstrum analysis to obtain the Fourier cepstra of each utterance.

First, a normalized Hamming window $H(n)$ of length 512 is used to obtain the short-term temporal signals, from which the cepstral coefficients are extracted. Then, to obtain the logarithmic Fourier magnitude spectrum, a fast Fourier transformation (FFT) is applied to the 512 windowed temporal signals, followed by the calculation of the logarithm of the modulus of the complex spectrum obtained by FFT. The inverse fast Fourier transform of the logarithmic magnitude spectrum makes it possible to extract the real logarithmic cepstrum.

As already detailed in Section 2.2, for each frame, the first 33 coefficients represent the cepstral vector related to the vocal tract; the next 224 define the cepstral vector related to the excitation signal, and the phase spectrum was determined by the 257 phase coefficients.

Our model is an LSTM encoder–decoder based on an attention mechanism. The encoder consists of 2 BiLSTM layers as follows: a forward LSTM, which receives the input sequence in order (from x_1 to x_N), and a backward LSTM, which receives the same sequence in reverse order (from x_N to x_1). Each layer contains 128 hidden units. The decoder is another LSTM network combined with local attention and consists of a single decoder LSTM layer with 256 hidden units. The latter is randomly initialized by the final state of the

BiLSTM encoder to maintain long-term memory. The dense layer is a fully connected feed-forward network (FFN) that has equal input dimensionality. This dense layer produces the alignment scores between the encoder's hidden states and the previous decoder's hidden states.

The model was trained using the Adam optimizer [33] with a batch size of 32 for 500 epochs. The MSE loss function is used and the dropout regularization has also been adopted to avoid overfitting. The learning rate is initialized at 10^{-3}. In our implementation, we use the NVIDIA GTX 1050 with CUDA of 10.1. The process is stopped when the validation loss does not refine for 10 epochs.

To compare our experiments, we adopted five kinds of systems based on esophageal-to-laryngeal VC, which are JD-GMM, DNN, LSTM, BiLSTM, and Seq2Seq with an attention mechanism. For training the JD-GMM, DNN, LSTM, and BiLSTM models, we used the DTW algorithm to time-align the parallel speech corpora. Note that to properly compare the performance, we took models with similar parameters. Thus, as a typical statistical approach, JD-GMM and DNN are considered as references. These methods are described as follows:

- JD-GMM: the Joint Density GMM-based VC system was implemented based on the Sprocket toolkit introduced in VCC2018 [20] and considered as a baseline system. All the source and target parameters are directly estimated from the conversion function by the expectation-maximization (EM) algorithm. The source (x_n) and target (y_n) vectors previously aligned by the DTW algorithm are concatenated together into an extended vector $z_n = [x_n, y_n]'$ and then the GMM parameters that model the joint probability density are estimated.
- DNN: the DNN-based VC system was implemented based on the approach of [10]. The number of units at each layer is chosen in order to ensure the best network performance. The ReLU (Rectified Linear Unit) activation function was used for its good performance [34], the dropout was set to 0.5, the learning rate was 0.001, the batch size was 32, and the training epoch was set to 500. For synthesis, an overlap-add method was adopted to reconstruct the waveform of the estimated enhanced speech.
- LSTM and BiLSTM: the LSTM and BiLSTM-based esophageal-to-laryngeal speech conversion models were implemented without an attention mechanism. For LSTM, the network architecture contained two hidden stacked LSTM layers and a fully connected output layer. For BiLSTM, we used 128 hidden units for each forward and backward encoder layer, and a single LSTM layer with 256 hidden units for the decoder. This architecture aims to progressively create a higher-dimensional feature space from one layer to another, to attempt to more easily represent the speaker's information. For training the BiLSTM model, the back-propagation through time (BPTT) algorithm was adopted. Both models have been trained for around 500 epochs, with a lot size of 32 and a dropout of 0.3, until the early stop condition was reached.

3.3. Objective Performance Measures

To compare the voice quality performance of the proposed speech enhancement methods and the baseline methods, we adopted four objective measures in the temporal, frequency, and perceptual domains.

1. Cepstral distance (CD): This is used to evaluate the cepstral distance between the converted and target frames. We evaluated the source (ES-SRC) and the different types of converted stimuli, which is calculated as:

$$CD[dB] = \frac{10}{Mlog10} \sum_{(\hat{C},C)} \sqrt{2 \sum_{i=1}^{D} (\hat{C}_i - C_i)^2}, \tag{19}$$

where $\hat{C}_i$ and C_i represent the ith component of the aligned converted and target cepstral vectors, respectively. D is the dimension of the cepstral vectors and M is the number of $(\hat{C}, C)$ couples.

2. Perceptual Evaluation of Speech Quality (PESQ): referred to by the ITU-T recommendation in the P.862 standard [35]. PESQ is a suitable means to evaluate the subjective voice quality of codecs (waveform and CELP-like encoders) and end-to-end measurements [36]. The range for the PESQ score is between -0.5 and 4.5.
3. Short-Time Objective Intelligibility (STOI): this is a function that compares the temporal envelopes of normal and converted speech in segments of short duration using a correlation coefficient. A greater STOI value indicates better intelligibility of enhanced speech.
4. Segmental Signal to Noise Ratio (segSNR): it defines the average of SNRs computed from aligned converted and target cepstra and was determined by Equation (20).

$$segSNR = \frac{10}{M} \sum_{(\hat{C},C)} log_{10} \frac{\sum_{i=0}^{N-1} C_i^2}{\sum_{i=0}^{N-1} (\hat{C}_i - C_i)^2} \tag{20}$$

where $\hat{C}_i$ and C_i are, respectively, the ith component of the aligned converted and target cepstral vectors. N (512) is the cepstrum length.

4. Results and Discussion

Audio samples from this work are depicted at the demo link: https://techtech-solution.com/Kadria/seq2seq-ES-enhancement.php (accessed on 10 June 2022).

4.1. Objective Evaluations

1. Comparison between baseline and proposed methods: objective evaluations were first performed to compare the performance of CD, PESQ, STOI, and segSNR of the proposed and reference methods introduced above.
   Tables 1 and 2 summarize the objective assessment results of source esophageal speech ES-SRC, and enhanced speech obtained by JD-GMM, DNN, LSTM, BiLSTM, and Seq2Seq-based methods. First, we can see that the cepstral vectors of ES-SRC are very different from those of laryngeal speech (target). Then, since the JD-GMM is a linear model and has a poor ability to model nonlinear relationships, its performance in converting ES to laryngeal voice is poorer than all other methods. Compared to the DNN model, the LSTM and BiLSTM models have better inter-frame characterization capability because the LSTM networks can take advantage of the relationship between long-distance frames.
   As indicated in Table 1, the BiLSTM model achieved better conversion performance than the GMM, DNN, and LSTM methods. This is because we adopt the BiLSTM model in our proposed Seq2Seq since it is adequate to characterize the difference between ES and its laryngeal speech counterpart. Note that the Seq2Seq-based method has a similar inter-frame characterization ability to the BiLSTM. However, our proposed method adopts the principle of the attention mechanism to accomplish an adaptive mapping between parallel sequences of esophageal and laryngeal speech. Compared to the BiLSTM model, our proposed method has improvements in CD, PESQ, STOI, and segSNR.
2. Comparison between different variants: in this experiment, we compared our proposed model based on a BiLSTM encoder–decoder network with three variants of this model:
   - (V1) Keeping the BiLSTM encoder but excluding the attention mechanism;
   - (V2) Replacing the BiLSTM encoder by an LSTM of 256 hidden units and keeping attention;
   - (V3) Using the LSTM network but excluding attention.

Table 3 lists the evaluation results. The proposed method outperforms the three other variants, with a larger PESQ score and lower CD value. In addition, adding the attention mechanism seems to have little effect on encoder–decoder networks based on LSTM. This explains that the hidden state $\vec{h}_i$ generated by an LSTM network only considers the information in $X_{\leq i}$; consequently, the attention mechanism will be inefficient due to insufficient information in the source hidden states. Moreover, the proposed method exceeds its variant without attention in terms of PESQ and STOI, which indicates that the incorporation of the attention mechanism would improve the performance of the VC system.

Table 1. Results of the objective assessment of the reference and proposed methods on the PC speaker esophageal corpus test set.

MODELS	CD Value	PESQ Score	STOI Score	segSNR Value
ES-SRC	9.408	2.407	0.606	2.981
JD-GMM	8.795	2.102	0.518	9.706
DNN	8.257	2.570	0.544	10.099
LSTM	8.009	2.808	0.624	10.915
BiLSTM	7.311	2.914	0.641	11.943
Seq2Seq	**6.836**	**2.994**	**0.733**	**12.854**

Table 2. Results of the objective assessment of the reference and proposed methods on the MH speaker esophageal corpus test set.

MODELS	CD Value	PESQ Score	STOI Score	segSNR Value
ES-SRC	9.153	2.381	0.597	2.999
JD-GMM	8.861	2.179	0.513	9.741
DNN	8.405	2.619	0.581	10.305
LSTM	7.910	2.805	0.624	11.083
BiLSTM	7.127	2.903	0.633	11.977
Seq2Seq	**6.605**	**3.002**	**0.775**	**12.901**

Table 3. Performance comparison between different variants on the PC speaker esophageal corpus test set.

VARIANTS	CD Value	PESQ Score	STOI Score	segSNR Value
BiLSTM+Attention	**6.994**	**2.933**	**0.718**	**12.407**
BiLSTM	7.311	2.914	0.641	11.943
LSTM+Attention	**7.826**	**2.871**	**0.639**	**11.866**
LSTM	8.009	2.808	0.624	10.915

4.2. Subjective Evaluations

In addition to the objective measurements, subjective listening tests were performed to evaluate the perceptual quality of our enhanced speech samples in terms of intelligibility and speech quality. The most frequently subjective tests were conducted.

For both tests, each participant evaluated the voice quality of twenty-two sentences (11 different test utterances spoken by 2 different pairs of speakers) from each of the previous types of voice. All tests were carried out under the same conditions and based on the same principle.

1. MOS test (Mean Opinion Score): used to evaluate the speech quality and intelligibility of the resynthesized voice. In this experiment, a group of fifteen auditors (five males and ten females) listened to a set of sample utterances and judged them independently, one by one, according to a rating scale of perceived quality. This scale ranged from (1) for the poorest quality to (5) for excellent quality, ((2) poor, (3) average, and (4) good

quality)). The average score awarded, therefore, constituted the MOS, which indicated the intelligibility and the quality of the enhanced speech.
For these tests, we conducted an opinion test for intelligibility and another opinion test for naturalness. Five sets of comparative experiments were evaluated by fifteen auditors.

- ES-SRC: source esophageal speech;
- JD-GMM: conversion system based on a joint density Gaussian mixture model;
- DNN: conversion system based on feed-forward DNN model;
- BiLSTM: conversion system based on Bidirectional LSTM model;
- Seq2Seq: conversion system based on sequence to sequence with attention mechanism model.

Figures 5 and 6 show the MOS values on intelligibility and speech quality using different conversion methods.
From both figures, we can see that the method based on GMM has the lower performance when compared with all the other methods, and this method has a high refusal rate by the laryngectomees. The LSTM method achieved almost similar intelligibility and naturalness when compared with the BiLSTM method, which in turn outperformed the DNN-based method, with a slight improvement. On the other hand, we can clearly notice that the converted speech using our method achieves the highest MOS in both tests. Thus, considering the average MOS of each approach, it is obvious that listeners prefer the samples obtained by our proposed method due to its effectiveness in improving the intelligibility and naturalness of ES.

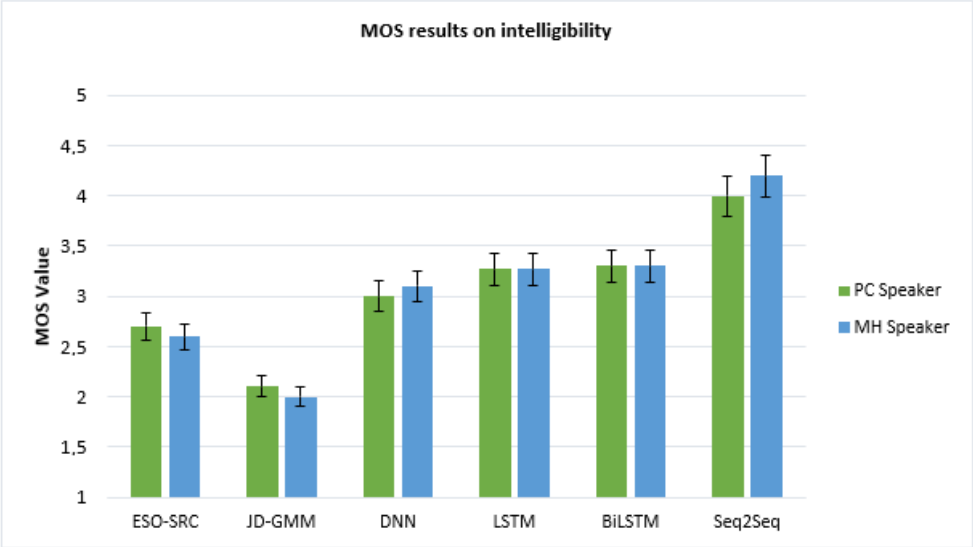

Figure 5. MOS on intelligibility test results with 95% confidence interval.

2. ABX preference test: this is a test to identify the similarity between enhanced and target sequences. In this evaluation, we presented to fifteen listeners a series of three speech samples: A, B, and X, respectively, as the source, target, and enhanced speech sample. We asked each listener to judge by a score the degree of closeness of enhanced sample X to the two other samples A and B. No preference (NP) could be selected in the event that the listener could not distinguish between two types of voice. Thus, we carried out four series of experiments: GMM with our method (Seq2Seq+Attention mechanism), DNN with our method, BiLSTM+Attention with our method, and Seq2Seq without attention with our method.
Figure 7 summarizes the results of the ABX test. The first two bars indicate that our model behaves much better than GMM and DNN. The third bar shows that our model works at similar levels with a BiLSTM+Attention. From the fourth bar, we can see that

our method performs much better when attention is applied and the speech generated by our approach is of better quality than that obtained by the other three methods. It is therefore clear that the inclusion of the attention mechanism increases the robustness of the model.

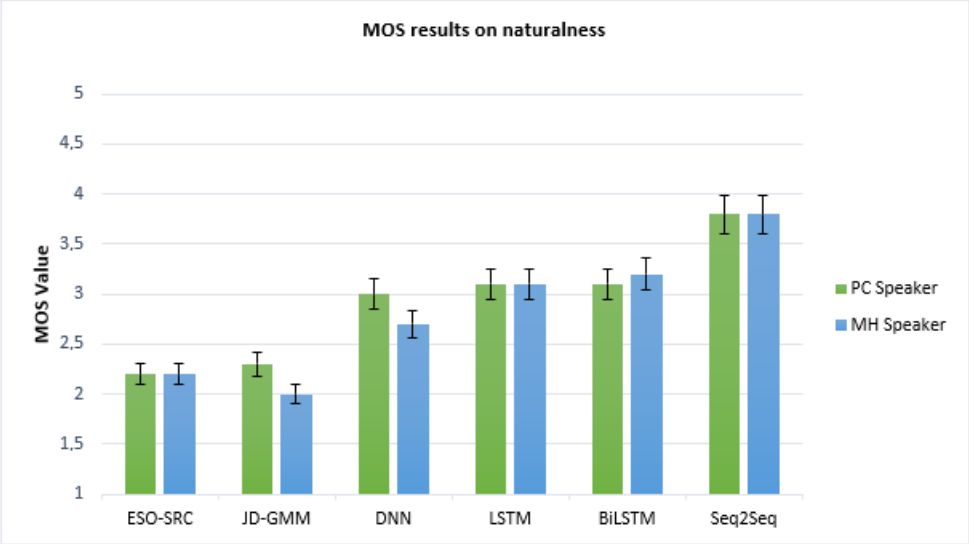

Figure 6. MOS on naturalness test results with 95% confidence interval.

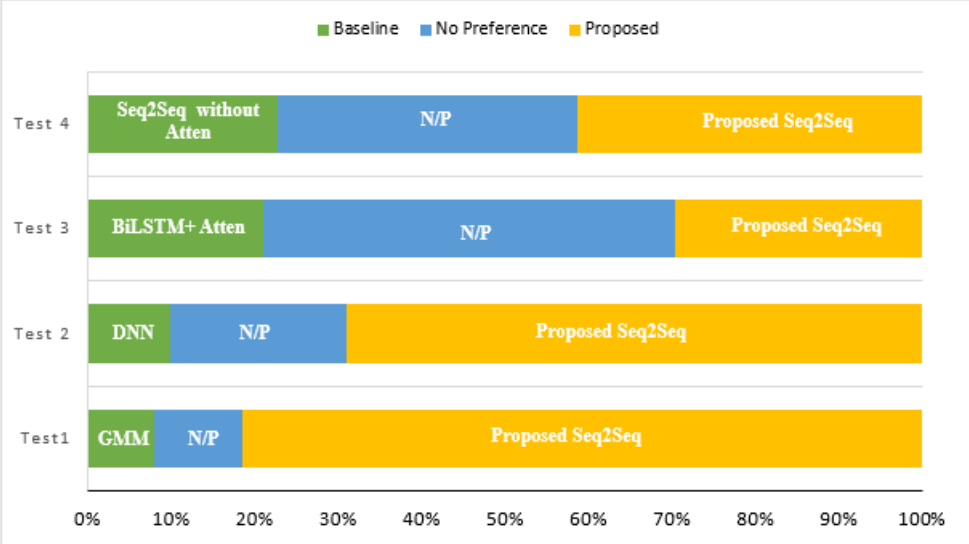

Figure 7. ABX preference test results for baseline and proposed methods. NP means no preference.

Some samples obtained from this work are depicted at the following demo link: https://techtech-solution.com/Kadria/seq2seq-ES-enhancement.php (accessed on 10 June 2022).

5. Conclusions

This paper presents a VC system for enhancing ES. A Seq2Seq mapping framework combined with an attention mechanism was proposed for esophageal-to-laryngeal VC. The proposed Seq2Seq method has a similar inter-frame characterization ability to the BiLSTM-based model. Unlike existing ES enhancement models, our method adopts the principle of the attention mechanism to accomplish an adaptive mapping between parallel sequences of esophageal and laryngeal features. It can also be used for laryngeal speech conversion. To preserve the identity of the target speaker, the excitation and phase coefficients are

estimated from the target learning space structured in the form of a binary search tree queried by the vocal tract coefficients previously predicted by the Seq2Seq model. At the resynthesis level, we applied the OLA-FFT recovery method. The experimental results show that our proposed method brings significant improvements and achieves better objective and subjective performance, even in some difficult cases. Indeed, it outperforms the reference systems based on GMM and DNN in terms of naturalness and intelligibility.

Author Contributions: Conceptualization, K.E. and J.D.M.; methodology, K.E. and J.D.M.; software, K.E.; validation, J.D.M. and M.F.; data curation, K.E.; writing—original draft preparation, K.E.; writing—review and editing, J.D.M. and M.F.; visualization, K.E.; supervision, J.D.M. and M.F. 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: Not applicable.

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

Sample Availability: Samples obtained from this study are provided at https://techtech-solution.com/Kadria/seq2seq-ES-enhancement.php (accessed on 10 June 2022).

References

1. Chalstrey, S.E.; Bleach, N.R.; Cheung, D.; Van Hasselt, C.A. A pneumatic artificial larynx popularized in Hong Kong. *J. Laryngol. Otol.* **1994**, *108*, 852–854. [CrossRef] [PubMed]
2. Diamond, L. Laryngectomy: The silent unknowns and challenges of surgical treatment. *J. Am. Acad. PAs* **2011**, *24*, 38–42. [CrossRef] [PubMed]
3. Guerrier, Y.; Jazouli, N. Vertical partial laryngectomy—Results. In *Functional Partial Laryngectomy*; Springer: Berlin/Heidelberg, Germany, 1984; pp. 145–149.
4. Matsui, K.; Hara, N.; Kobayashi, N.; Hirose, H. Enhancement of esophageal speech using formant synthesis. *Acoust. Sci. Technol.* **2002**, *23*, 69–76. [CrossRef]
5. Hisada, A.; Sawada, H. Real-time clarification of esophageal speech using a comb filter. In Proceedings of the International Conference on Disability, Virtual Reality and Associated Technologies, Veszprém, Hungary, 18–20 September 2002; pp. 39–46.
6. Desai, S.; Black, A.W.; Yegnanarayana, B.; Prahallad, K. Spectral mapping using artificial neural networks for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 954–964. [CrossRef]
7. 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]
8. Doi, H.; Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Statistical approach to enhancing esophageal speech based on Gaussian mixture models. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4250–4253.
9. Al-Radhi, M.S.; Csapó, T.G.; Németh, G. Time-Domain Envelope Modulating the Noise Component of Excitation in a Continuous Residual-Based Vocoder for Statistical Parametric Speech Synthesis. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 434–438. [CrossRef]
10. 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.* **2019**, *22*, 99–110. [CrossRef]
11. Ezzine, K.; Frikha, M. A comparative study of voice conversion techniques: A review. In Proceedings of the International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 22–24 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
12. Doi, H.; Toda, T.; Nakamura, K.; Saruwatari, H.; Shikano, K. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2013**, *22*, 172–183. [CrossRef]
13. Zhao, Y.; Huang, W.C.; Tian, X.; Yamagishi, J.; Das, R.K.; Kinnunen, T.; Ling, Z.; Toda, T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv* **2020**, arXiv:2008.12527.
14. Lachhab, O.; Di Martino, J.; Elhaj, E.I.; Hammouch, A. A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion. *SpringerPlus* **2015**, *4*, 1–14. [CrossRef]
15. 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]
16. Alers, T.J.; Fennema, B.A.; van Breukelen, J.J. Tracheo-Esophageal Speech Enhancement: Real-Time Pitch Shift and Output. Bachelor's Thesis, Delft University of Technology, Delft, The Netherlands, 2020.

17. Mohammadi, S.H.; Kain, A. An overview of voice conversion systems. *Speech Commun.* **2017**, *88*, 65–82. [CrossRef]
18. Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 43–49. [CrossRef]
19. Keogh, E.J.; Pazzani, M.J. Derivative dynamic time warping. In Proceedings of the 2001 SIAM International Conference on Data Mining, Chicago, IL, USA, 5–7 April 2001; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2001; pp. 1–11.
20. Lorenzo-Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F.; Kinnunen, T.; Ling, Z. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv* **2018**, arXiv:1804.04262.
21. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *arXiv* **2014**, arXiv:1409.3215.
22. Miyoshi, H.; Saito, Y.; Takamichi, S.; Saruwatari, H. Voice conversion using sequence-to-sequence learning of context posterior probabilities. *arXiv* **2017**, arXiv:1704.02360.
23. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4784–4788.
24. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
25. Ramos, M.V.; Black, A.W.; Astudillo, R.F.; Trancoso, I.; Fonseca, N. Segment Level Voice Conversion with Recurrent Neural Networks. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 3414–3418.
26. Ramos, M.V. Voice Conversion with Deep Learning. Masters's Thesis, Tecnico Lisboa, Lisbon, Portugal, 2016.
27. Kaneko, T.; Kameoka, H.; Hiramatsu, K.; Kashino, K. Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017.
28. Tanaka, K.; Kameoka, H.; Kaneko, T.; Hojo, N. AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6805–6809.
29. Lian, H.; Hu, Y.; Yu, W.; Zhou, J.; Zheng, W. Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention. *IEEE Access* **2019**, *7*, 130495–130504. [CrossRef]
30. Bentley, J.L. Multidimensional binary search trees used for associative searching. *Commun. Acm* **1975**, *18*, 509–517. [CrossRef]
31. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
32. Wei, H.; Zhou, A.; Zhang, Y.; Chen, F.; Qu, W.; Lu, M. Biomedical event trigger extraction based on multi-layer residual BiLSTM and contextualized word representations. *Int. J. Mach. Learn. Cybern.* **2021**, *13*, 721–733. [CrossRef]
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
34. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 21–24 January 2010.
35. Recommendation, I.T. *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*; International Telecommunication Union: Geneva, Switzerland, 2001; p. 862.
36. 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; Volume 2, pp. 749–752.