Clinics in Surgery

Research Article
Published: 16 Jun, 2022

Esophageal to Laryngeal Voice Conversion Using a Sequence to Sequence Mapping Model Including an Attention Mechanism

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

[1]National Engineering School of Carthage, Carthage University, Tunisia

[2]Lorraine Laboratory of Research in Computer Science and its Applications, Lorraine University, France

[3]ATISP - National School of Electronics and Telecommunications of Sfax, Sfax University, Tunisia

Abstract

Esophageal Speech (ES) can be used as an alternative speaking method for laryngectomes. Compared to laryngeal voice, ES is characterized by poor intelligibility and poor quality due to chaotic fundamental frequency, specific noises that resemble belching, and low intensity. These issues are alleviated by converting ES into more natural speech that is an effective way to improve speech quality and intelligibility. To accomplish this, we propose in this work a novel esophageal-to-laryngeal Voice Conversion (VC) system based on a Sequence-to-Sequence (Seq2Seq) technique combined with an attention mechanism. The originality of the proposed method is that it does not require any dynamic time alignment during the training phase, which avoids erroneous mappings and significantly reduces the computing time. In addition, to preserve the identity of the target speaker, excitation and phase coefficients are estimated by querying a binary search tree in the target training space through the coefficients of the vocal tract previously predicted by the proposed Seq2Seq mapping model. In experiments, we compare our approach with baseline methods using numerous measures for objective and subjective evaluations. Perceptual tests confirmed that our proposed method behaves better and achieves better performance even in some difficult cases. In fact, it consistently exceeds conventional methods as acoustic models in terms of speech quality and intelligibility.

Keywords: Esophageal-to-laryngeal conversion; Sequence-to-sequence; Attention mechanism; Intelligibility; Speech quality

OPEN ACCESS

Correspondence:
Kadria Ezzine, Department of Speech and Audio Processing, ENICarthage, Carthage University, ATISP, ENET'COM, Sfax University, National Engineering School of Carthage, Tunisia,
E-mail: kadria.ezzine@gmail.com

Received Date: *05 Apr 2022*
Accepted Date: *10 Jun 2022*
Published Date: *16 Jun 2022*

Citation:
Ezzine K, Di Martino J, Frikha M. Esophageal to Laryngeal Voice Conversion Using a Sequence to Sequence Mapping Model Including an Attention Mechanism. Clin Surg. 2022; 7: 3534.

Copyright *© 2022 Kadria Ezzine. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

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 laryngectomees patients to re-speak with an alternate voice without 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 liberal 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 the principle of the ES is based on the use of a pharyngo-oesophageal digestive segment as a neovibrator [3]. 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. Waveforms, spectrograms, and F0 contours of laryngeal and esophageal stimuli for the same sentence (Figure 1). We can observe that the pitch is chaotic and the ES is characterized by a much lower HNR (Harmonics to Noise Ratio) than the laryngeal speech, therefore, the analysis and extraction of F0 is quite difficult, if not impossible. In addition, it is clear that ES presents low intensity and high specific noises in all frequency bands, resulting in a degradation of naturalness and audio quality.

All of these instabilities in the acoustic characteristics produce poor quality sounds difficult

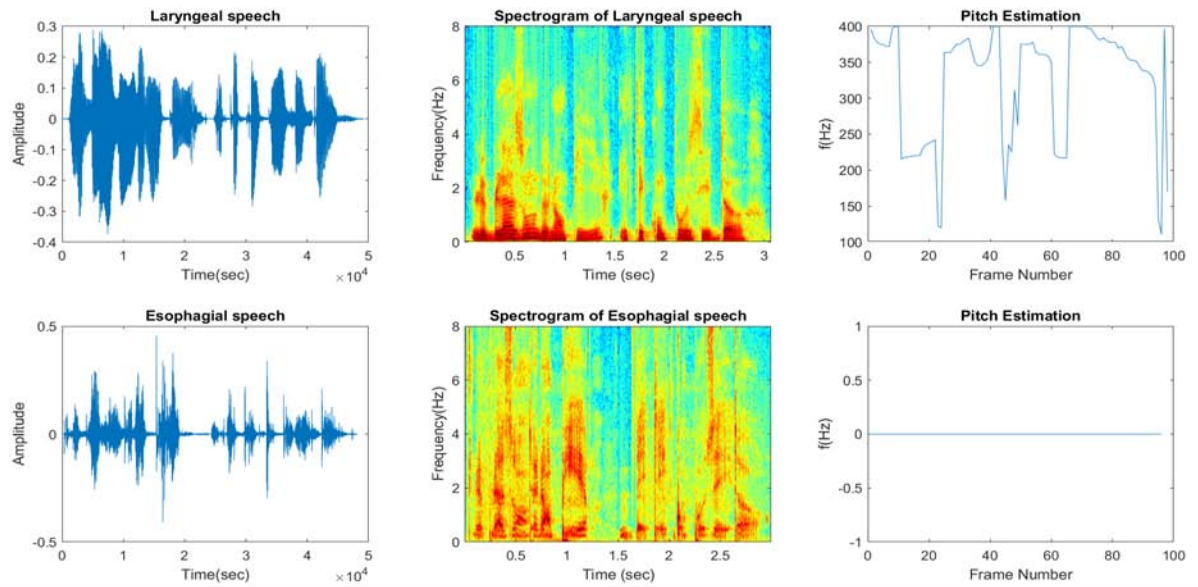

Figure 1: Example of waveforms, spectrograms and pitch contours of both laryngeal and esophageal speech.

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 have been carried out, and approaches based on VC technique which allows transforming the voice of a source speaker (laryngectomee) into that of a target speaker (laryngeal) [7-11].

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 [12].

Besides, temporal alignment is another problem when converting esophageal to laryngeal speech. As features alignment is necessary for VC systems, the Dynamic Time Warping (DTW) algorithm [13] 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 [14]. 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 [15] and the VCC2020 [16] 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 [17] has proven to be an outstanding technique for mapping one sequence to another one, especially in the field of VC [18], Text-To-Speech synthesis (TTS) [19], and natural language processing [20]. As far as we know, converting spectral features using a Seq2Seq model with attention has been attempted for the first time [21]. A recent method is proposed for an enhancement of whisper-to-normal speech based on a Seq2Seq model [22]. In this method, taking into account the attention technique makes it possible to further stabilize the training procedure that outperforms the conventional methods.

Nevertheless, how to significantly improve 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 [22] and the first work [21]. The strength of the proposed method is that 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 II presents the proposed methods. Section III details the experimental setup. Section IV presents the results and discussion. Finally, a conclusion is given in Section V.

Methods

System overview

Figure 2 shows the framework of our proposed Seq2Seq esophageal to laryngeal speech conversion. The conversion process is divided into two main phases: Training and conversion. 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 was adopted to normalize spectral features; 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

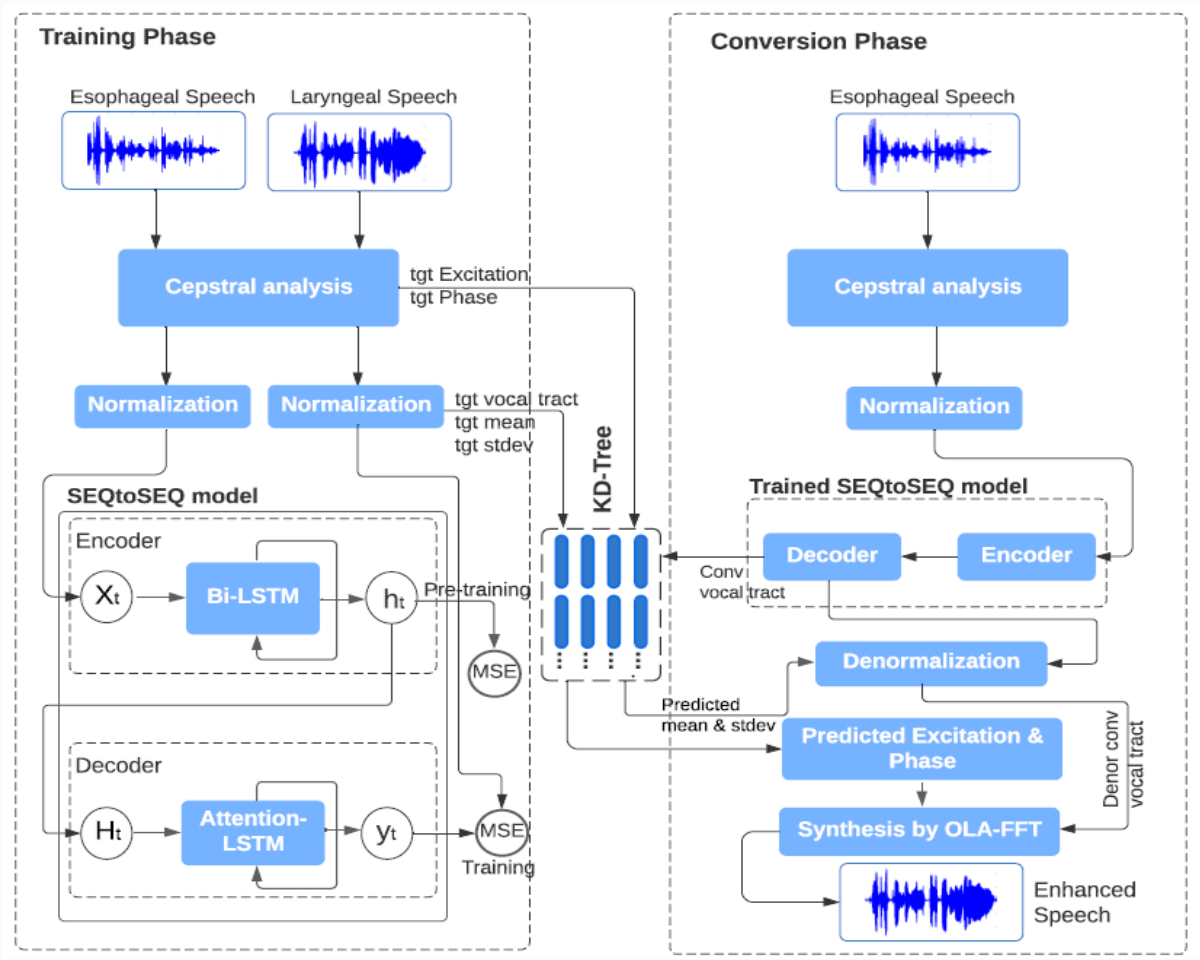

Figure 2: An overview of the proposed esophageal speech enhancement framework based on a Seq2Seq model.

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 optimized for each of them and the error is calculated and back-propagated frame by frame using a loss function.

In the conversion phase, cepstral coefficients are firstly extracted from each ES signal and then normalized. Next, the trained Seq2Seq model applied to convert only the first cepstral packet (vocal tract feature vectors) from the source speaker into their approaching target. After that, to preserve the identity of the target speaker, e propose to predict cepstral excitation and phase coefficients from the target training space by using a KD-tree algorithm [23]. The binary KD-tree is constructed with cepstral frames of the laryngeal vocal tract. Then, it is queried by the converted vocal tract cepstral 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, that's 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 in applying an Inverse Fast Fourier Transform (IFFT) to the complex spectra.

Feature extraction and normalization

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

The real Fourier cepstral 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:

$$C[n] = IFFT\ (Log|FFT\ (x[n] \times H[n])|) \quad (1)$$

where $H(n)$ is a normalized Hamming window [24] 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 $[vt_0...vt_{32}]$, cepstral excitation as $[ex_{33}...ex_{256}]$, and phase coefficients as $[ph_0...ph_{256}]$.

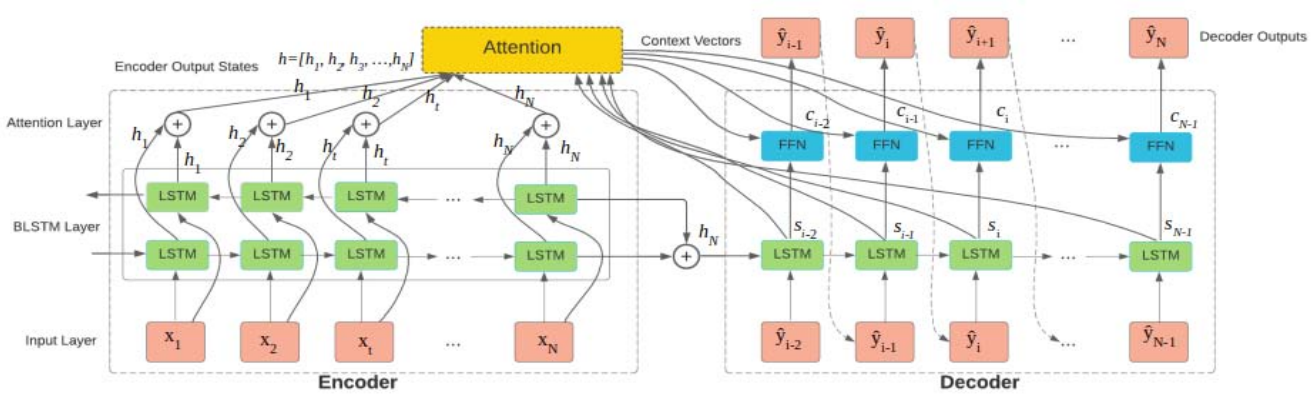

Figure 3: The network structure of a Seq2Seq model, where attention layer is ignored for clarity.

In the normalization step, we normalized 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 as:

$$\chi'_{i,n} = \frac{(\chi_{i,n} - \mu_i)}{\sigma_i} \quad (2)$$

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

Network model

The proposed framework is a Seq2Seq model with an attention mechanism, consisting of two main components: A stack of bidirectional LSTM encoder and an LSTM decoder based on an attention network. Figure 3 shows the overall network architecture of the designed model.

Let $X^{(s)} = \left[\chi_1^{(s)}, \ldots, \chi_{N_s}^{(s)}\right]$ and $Y^{(t)} = \left[y_1^{(t)}, \ldots, y_{N_t}^{(t)}\right]$, 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 $N_s \neq N_t$).

Bidirectional-LSTM based encoder

Due to the long sequences to model and the high 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) layer, which are arranged incrementally (Figure 3).

As the LSTM architecture shows (Figure 4), 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 t^{th} time step respectively; and an output gate noted by o_t, which controls the output state (also called the "hidden state") (h_t). The LSTM propagation is formulated as follow:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right), \quad (3)$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right), \quad (4)$$

$$O_t = \sigma\left(W_O x_t + U_O h_{t-1} + b_O\right), \quad (5)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh\left(W_C x_t + U_C h_{t-1} + b_C\right), \quad (6)$$

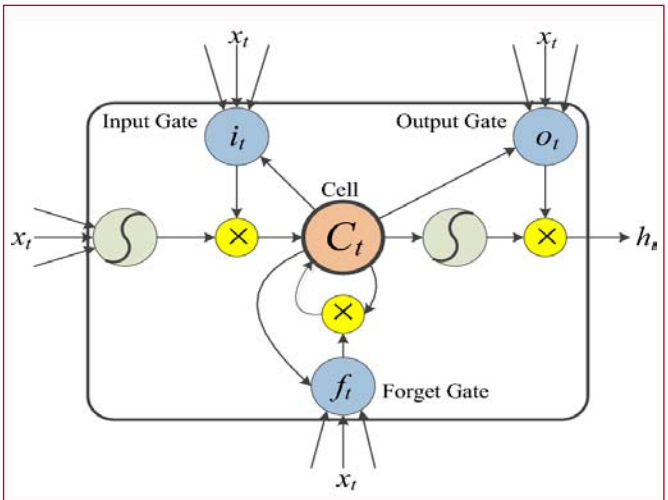

Figure 4: Internal structure of the LSTM unit.

$$h_t = O_t \odot \tanh\left(C_t\right), \quad (7)$$

where, x_t is the input vector at the t^{th} time step, c_t is the current long term memory cell, σ() 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, U_* map from the previous hidden dimension to the current hidden dimension.

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

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right], \quad (8)$$

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,

$$\overrightarrow{h_t} = \overrightarrow{LSTM}\left(h_{t-1}, x_t, c_{t-1}\right), \quad (9)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}\left(h_{t+1}, x_t, c_{t+1}\right). \quad (10)$$

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-to-laryngeal speech

conversion, the length of the laryngeal feature sequence is always shorter than that of the esophageal feature sequence. In this kind of sequences 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 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.

At each i^{th} decoder step, the attention output (context vector) c_i is computed using a weighted linear combination of the encoder's hidden states.

$$c_i = \sum_{j=1}^{j=N} \alpha_{ij} h_j \quad (11)$$

where, h_j represents the encoder's hidden state at position j, α_{ij} is the attention weights as:

$$\alpha_{ij} = \frac{\exp\left(score_{ij}\right)}{\sum_{j=1}^{N} \exp\left(score_{ij}\right)} \quad (12)$$

$score_{ij}$ is the attention score, which calculates the similarity between the encoder's hidden state and the previous cell state of the decoder, computed as follows:

$$score_{ij} = att\left(\left(s_{i-1}, h_j\right); \theta_{attention}\right) \quad (13)$$

where, att is a Feed Froward 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 current $\hat{y}_i$ to generate the decoder hidden state s_i.

$$s_i = LSTM\left(\left(s_{i-1}, \hat{y}_i, c_i\right); \theta_{decoder}\right) \quad (14)$$

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

Thus, the decoder output is calculated according to the previous equation:

$$\hat{y}_{i+1} = fc\left(s_i, c_i\right) \quad (15)$$

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

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:

$$MSE_{loss}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 \quad (16)$$

where $y^{\wedge}_i$ and y_i are the i^{th} dimensions of the predicted and the target cepstral vectors, respectively. θ is the model parameters, and N is the total samples number. The decoder is optimized step by step, that is, 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.

Experiments

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 non-Laryngectomee (AL), French speaker. The audio signals were all sampled at 16 kHz and directly stored in wave files. Each file contains a single sentence and lasts between 3 sec to 5 sec. Thus, in total, the data collected included 867 acoustic data files.

- We utilized two pairs of parallel corpora (source and target speakers) for training and evaluation of the proposed approach, namely: PC and MH as source speakers and AL as target speakerPC (ES-male) & AL (NS-male)
- MH (ES-male) & AL (NS-male)

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

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 cepstral 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 (IFFT) of the logarithmic magnitude spectrum makes it possible to extract the real logarithmic cepstrum.

As already detailed in subsection 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. Hence, the cepstral features used as input and output contained 33 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 [25] 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 over fitting. The learning rate is initialized at 10^{-3}. In our implementation, we use 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 carried out 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 the models with similar parameters. There, as a typical statistical approach, JD-GMM and DNN are considered as references. These methods are described as follows:

a) JD-GMM: The Joint Density GMM-based VC system was implemented based on the Sprocket toolkit introduced in VCC2018 [15] 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 which model the joint probability density are estimated.

b) DNN: The DNN-based VC system was implemented based on the approach of [9]. 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 [26], 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.

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 domain:

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:

$$CD[dB] = \frac{10}{M \log 10} \sum_{(\hat{C},C)} \sqrt{2 \sum_{i=1}^{D} \left(\hat{C}_i - C_i\right)^2} \qquad (17)$$

Where $\hat{C}_i$ and C_i represent the i^{th} 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.

Perceptual evaluation of speech quality (PESQ): Referred by the ITU-T recommendation in the P.862 standard [27]. PESQ is a suitable means to evaluate subjective voice quality of codecs (waveform and CELP-like encoders) and end- to-end measurements [28]. The range for the PESQ score is between –0.5 and 4.5.

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.

Segmental signal to noise ratio (segSNR): It defines the average of SNRs computed from aligned converted and target cepstral and was determined by the following equation:

$$segSNR = \frac{10}{M} \sum_{(\hat{C},C)} \log 10 \frac{\sum_{i=0}^{N-1} C_i^2}{\sum_{i=0}^{N-1} \left(\hat{C}_i - C_i\right)^2} \qquad (18)$$

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

Results and Discussion

Objective evaluations

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.

Table 1, 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 [29,30].

As indicated in Table 1, the BiLSTM model achieved better conversion performance than the GMM, DNN, and LSTM methods. This involves that 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

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	2.57	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	7.91	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: Perfermance comparaison 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

based method has a similar inter-frame characterizing ability to the BiLSTM. Though, 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.

- Comparison between different variants: in this experiment, we compared our proposed model based on BiLSTM encoder-decoder network with three variants of this modeV1) 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 a 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 (BiLSTM) in terms of PESQ and STOI which indicates that the incorporation of the attention mechanism would improve the performance of the VC system.

Subjective evaluations

In addition to the objective measurements, subjective listening tests are 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 evaluates 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 are based on the same principle.

MOS test (Mean Opinion Score): MOS test Used to evaluate the speech quality and intelligibility of the re-synthesized voice. In this experiment, a group of fifteen auditors (five males and ten females) listens to a set of sample utterances and judges independently, one by one, according to a rating scale of perceived quality. This scale goes from (1) for the poorest quality to (5) for excellent quality, ((2) poor, (3) average, and (4) good quality)). The average score awarded, therefore, constitutes the MOS, which decides 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.

Figure 5, 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 outperforms 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.

ABX preference test: It 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) can be selected in case the listener cannot distinguish between two types of voice. Thus, we carried out four series of experiments: GMM with our method (Seq2Seq + Attention mechanism), DNN

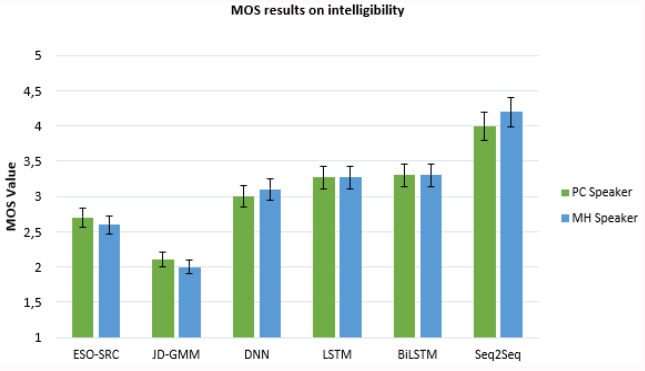

Figure 5: MOS on intelligibility test results with 95% confidence interval.

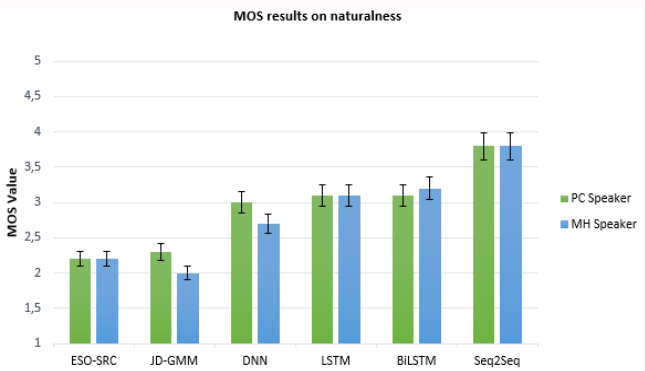

Figure 6: MOS on naturalness test results with 95% confidence interval.

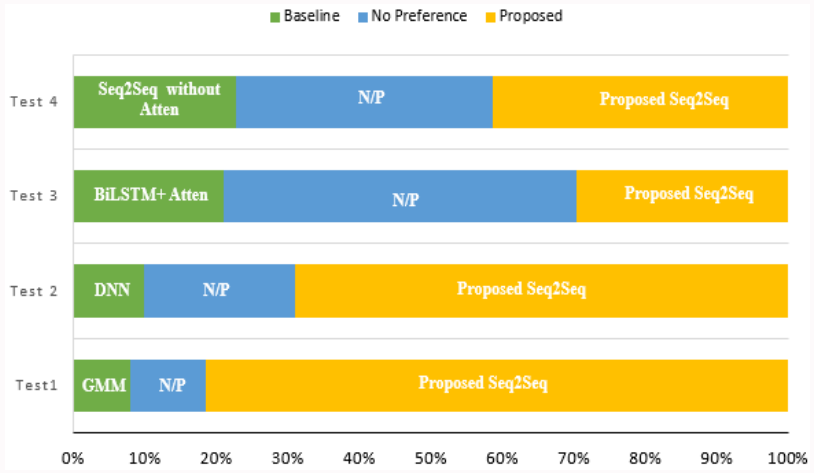

Figure 7: ABX preference test results for baseline and proposed methods. NP means no preference.

with our method, BiLSTM + Attention with our method, and Seq2Seq without Attention with our method.

The results of the ABX test summarizes in Figure 7. 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. Some samples obtained from this work are depicted in the following demo link.

Conclusion

This paper presents 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 characterizing 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.

References

1. Othmane B, 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;99-110.
2. Bentley JL. Multidimensional binary search trees used for associative searching. Commun ACM. 1975;18(9):509-17.
3. Chalstrey S, Bleach N, Cheung D, Van Hasselt C. A pneumatic artificial larynx popularized in Hong Kong. J Laryngol Otol. 1994;108(10):852-4.
4. Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Empirical Methods Natural Lang Process. 2014;1724-34.
5. Desai S, Black AW, Yegnanarayana B, Prahallad K. Spectral mapping using artificial neural networks for voice conversion. IEEE Trans Audio Speech Lang Process. 2010;18(5):954-64.
6. Diamond L. Laryngectomy: The silent unknowns and challenges of surgical treatment. J Am Acad Pas. 2011;24(8):38-42.
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. 2010a;93(9):2472-82.
8. Doi H, Nakamura K, Toda T, Saruwatari H, Shikano K. Statistical approach to enhancing esophageal speech based on Gaussian mixture models. 2010 IEEE Int Conference Acoustics Speech Signal Process. 2010b;4250-3.
9. 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(1):172-83.
10. Ezzine K, Frikha M. A comparative study of voice conversion techniques: A review. 2017 Int Conference Adv Technol Signal Image Process (ATSIP). IEEE. 2017;1-6.
11. Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Trans Acoustics Speech Signal Process. 1984;32(2):236-43.
12. Guerrier Y, Jazouli N. Vertical partial laryngectomy-results. In: Functional Partial Laryngectomy. Springer, 1984. p. 145-9.
13. Hisada A, Sawada. Real-time clarification of esophageal speech using a comb filter. Int Conference Disability Virtual Reality Assoc Technol. 2002;39-46.
14. Keogh EJ, Pazzani MJ. Derivative dynamic time warping. Proceedings of the 2001 SIAM international conference on data mining. 2001;1-11.
15. Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014.
16. Lachhab O, Di Martino J, Elhaj EI, Hammouch A. "A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion". In: SpringerPlus 4.1. 2015;1-14.
17. Lian H, Hu Y, Yu W, Zhou J, Zheng W. Whisper to normal speech conversion using sequence-to-sequence mapping model with auditory attention. 2019.
18. Lorenzo-Trueba J, Yamagishi J, Toda T, Saito D, Villavicencio F, Kinnunen T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. 2018.

19. Matsui K, Hara N, Kobayashi N, Hirose H. Enhancement of esophageal speech using formant synthesis. Acoustical Science and Technology. 2002;23(2):69-76.

20. Miyoshi H, Saito Y, Takamichi S, Saruwatari H. Voice conversion using sequence-to-sequence learning of context posterior probabilities. arXiv preprint arXiv:1704.02360 (2017).

21. Mohammadi SH, Kain A. "An overview of voice conversion systems". In: Speech Communication 88. 2017;65-82.

22. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. 2010.

23. Ramos MV, Black AW, Astudillo RF, Trancoso I, Fonseca N. Segment level voice conversion with recurrent neural networks. 2017;3414-8.

24. ITU-T Recommendation. "Perceptual Evaluation Of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs". In: Rec. ITU-T P. 862 (2001).

25. Rix AW, Beerends JG, Michael P, Hekstra AP. Perceptual Evaluation of Speech Quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE Int Conference Acoustics Speech Signal Process. 2001;749-52.

26. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. In: IEEE Transactions Acoustics Speech Signal Process. 1978;26(1):43-9.

27. Sutskever, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst. 2014;3104-12.

28. Tachibana H, Uenoyama K, Aihara S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. IEEE Int Conference Acoustics Speech Signal Process. (ICASSP). IEEE. 2018;4784-8.

29. 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;1-13.

30. Zhao Y, Huang WC, Tian X, Yamagishi J, Das RK, Kinnunen T, et al. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. 2020.