Mandarin Electrolaryngeal Speech Voice Conversion with Speech Encoder Loss Learning and Seq2seq Modeling

Ming-Chi Yen, Chia-Hua Wu, Shu-Wei Tsai, Jyh-Shing Roger Jang, Yu Tsao, Amir Hussain, and Hsin-Min Wang

***Abstract*—Electrolaryngeal (EL) speech utilizes excitation signals generated by an electrolarynx instead of human vocal vibrations. In daily communication, EL speech is less natural and more difficult to understand than natural (NL) speech due to mechanical vibration noise and fixed pitch. Different methods have been proposed to improve the quality and intelligibility of EL speech, but limited training data and atypical acoustic characteristics pose challenges. Voice conversion (VC) is one popular method, and the task is called EL speech VC (ELVC). Sequence-to-sequence (seq2seq) modeling with pretraining strategies has been proposed for ELVC. However, seq2seq ELVC still faces the problem of incomplete and missing phonemes. Furthermore, although previous work has evaluated simulated EL (sEL) speech produced by healthy speakers using electrolarynxes, the effectiveness of seq2seq ELVC on patient EL (pEL) speech has not been studied. In this paper, we propose three approaches to address the issues of ELVC implementation. First, we utilize sEL speech in the pretraining stage to close the gap between pEL speech and NL speech. Second, we adopt a speech encoder loss to solve the problem of incomplete and missing phonemes. Third, we introduce waveform similarity overlap-and-add to augment pEL training speech. We conduct systematic experiments on pEL speech to evaluate our approaches. Ablation studies show that incorporating our approaches improves the converted speech in both objective and subjective evaluations compared to the baseline model.**

***Index Terms*—electrolaryngeal speech, voice conversion, pretraining, sequence-to-sequence learning**

I. Introduction

SPEECH is a fundamental communication method for human interaction in our daily life. According to the source-filter model [1], human speech is generated through two main stages: the airflow from the lungs forms an excitation signal (source) through the opening and closing of the vocal folds, and then passes through the vocal tract (filter) to determine the spectral structure of the output speech. The speech produced by healthy people through the above mechanism is called natural (NL) speech. However, patients who undergo total laryngectomy for pharyngeal malignancies such as laryngeal cancer will lose the ability to generate excitation signals, resulting in a complete loss of the ability to speak. For these laryngectomees, an alternative method of generating excitation signals is to use an electrolarynx, which generates mechanical excitation signals. The speech produced by people using this device is called electrolaryngeal (EL) speech. EL speech is less understandable and natural than NL speech due to fixed pitch and mechanical vibration noise.

The goal of voice conversion (VC) is to convert one speech sound into another without changing the linguistic content. It has been applied to enhance EL speech to approximate the quality and intelligibility of NL speech [2], [3], [4]. This type of VC task is called EL speech VC (ELVC). Current ELVC methods [2], [3], [4] usually require the use of a parallel corpus of EL speech and NL speech pairs with the same linguistic content for modeling. ELVC systems typically include feature extraction and alignment, mapping function learning, and waveform reconstruction. Feature extraction and alignment are crucial for ELVC systems to accurately convert the attributes from the source to the target, especially for frame-based ELVC. Dynamic time warping (DTW) [5]

is a well-known method for finding the best alignment path of two feature vector sequences based on distance measures (e.g., L2 distance) of source-target feature vector pairs. However, the atypical acoustic characteristics of EL speech are very different from NL speech, resulting in incorrect feature extraction, making accurate feature alignment difficult. Mismatched source-target feature pairs will lead to incorrect mapping function learning, resulting in poor ELVC performance. Furthermore, compared to other VC tasks, the corpus of the ELVC task is relatively small due to the challenges of collecting EL speech corpora from patients.

To cope with the data scarcity problem, the authors of [6] used text-to-speech (TTS) to augment training data. By combining the generated data with the original data for pretraining, the accuracy of automatic speech recognition (ASR) of the converted EL speech is effectively improved. However, for ELVC tasks, the main challenge in training EL-TTS is the requirement for certain quality standards for EL speech and the cost of additional training resources. Furthermore, the EL

Manuscript received January 15, 2025

Ming-Chi Yen is a research assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan and a PhD. student in Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Chia-Hua Wu is a research assistant in Institute of Information Science, Academia Sinica, Taipei, Taiwan

Shu-Wei Tsai is a Doctor in Department of Otolaryngology Head and Neck Surgery, National Cheng Kung University Hospital, Tainan, Taiwan

Jyh-Shing Roger Jang is a professor in Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

Yu Tsao is a research fellow in Research Center for Information Technology Innovation, Academia Sinica, Nankang, Taipei, Taiwan

Amir Hussain is a professor in School of Computing, Edinburgh Napier University, Scotland, UK

Hsin-Min Wang is a research fellow in Institute of Information Science, Academia Sinica, Taipei, Taiwan

speech generated is far from ideal. Due to these shortcomings, the EL-TTS method may not be suitable for the Mandarin ELVC task in this paper on real patients.

To address these issues, we proposed alignment-free sequence-to-sequence (seq2seq) ELVC [7], which involves a pretraining strategy on a large-scale NL speech corpus to alleviate the insufficient data problem of ELVC tasks. The seq2seq mechanism can learn feature mapping and alignment simultaneously, which can reduce errors caused by incorrect alignment. However, due to the architecture of seq2seq modeling, the output often suffers from incomplete and missing phonemes, resulting in syllable errors in the converted speech. Furthermore, previous work has evaluated simulated EL (sEL) speech produced by healthy speakers using electrolarynxes, and the effectiveness of seq2seq ELVC on patient EL (pEL) speech has not been studied. To go one step further, we propose three techniques to improve seq2seq ELVC. The major contributions of this study include:

- We add an additional pretraining stage using sEL speech to the seq2seq ELVC task, alleviating the gap between pEL speech and NL speech.
- We use the speech encoder loss for ELVC model training to solve the problem of incomplete and missing phonemes in seq2seq modeling.
- We apply efficient speech rate-based data augmentation to pEL speech, which can alleviate the problem of insufficient data and improve the performance of ELVC systems.

The remainder of this paper is organized as follows. Section II presents the proposed model. Sections III and IV present the experimental setup and results, respectively. Section V provides some discussion. Finally, Section VI concludes the paper.

II. Proposed approach

Our ELVC model is developed on top of the voice transformer network (VTN) for the speaker VC task [8]. Figure 1 shows the overall training process. VTN pretraining includes decoder pretraining and encoder pretraining. Our electrolaryngeal speech transformer network (ETN) pretraining includes decoder pretraining, encoder pretraining, and ELVC pretraining on sEL speech. Finally, the ELVC model is trained on pEL speech. L1 loss is used in all training stages.

Compared to the previous seq2seq ELVC model based on VTN [7], we integrate three new methods into the improved model. The first is to use sEL speech for pretraining, which can reduce the gap between pEL speech and NL speech in the ELVC task. The second is to use the speech encoder loss for ELVC model training, which can extend the advantages of a large-scale NL speech corpus to a very small amount of EL training speech, solve the problem of incomplete and missing phonemes in seq2seq modeling, and improve the intelligibility of the converted speech. The third is to use speech rate-based data augmentation on pEL training speech. To reduce the mismatch between pEL training speech and NL training speech, we change the speech rate of pEL training speech to be closer to NL training speech, which can reduce the difficulty of aligning EL and NL feature sequences. We describe these methods in detail below.

A. ETN: Electrolaryngeal speech transformer network

One of the focuses of this paper is to explore how sEL speech can be leveraged to facilitate model training for pEL speech in ELVC. The main concept of VTN involves pretraining with a large amount of NL speech to obtain good speech generation for the decoder and speech feature disentanglement for the encoder. Furthermore, after the encoder learns how to extract speech features from NL speech, we extend it by training a sEL-to-NL VC model using a relatively large amount of sEL speech. This pretrained model is then utilized to train the pEL-to-NL VC model using a small amount of pEL speech. As illustrated in Figure 1, compared with VTN, it becomes ETN after adding the third stage of sEL-to-NL pretraining.

There are two main benefits of using sEL-to-NL pretraining. First, sEL speech and pEL speech are both challenging atypical speech. As each patient's pathology differs, pEL speech exhibits variability, whereas sEL speech of healthy speakers are relatively more consistent. Since sEL speech is closer to NL speech than pEL speech, sEL-to-NL is a relatively easier ELVC task to handle than pEL-to-NL. By first training on a relatively simple task, the model acquires the ability to disentangle EL speech, making it easier to achieve good conversion results in the subsequent relatively challenging pEL-to-NL task. Second, more data are available for training the sEL-to-NL VC model. Given the general lack of ELVC training data, the advantage of more training data is obvious. In summary, using a larger amount of data in ELVC pretraining is expected to provide better model initialization and training outcomes for the pEL-to-NL VC task.

B. Speech encoder loss

We integrate the encoder of a pretrained ASR model into the ELVC training stage to provide the speech encoder loss. This allows the ELVC model to continuously benefit from the rich information of a large NL speech dataset when trained on a limited amount of EL speech. In addition, using the ASR model can also guide the ELVC model to retain phoneme information, thereby solving the problem of incomplete and missing phonemes and enhancing intelligibility. The concept is shown in the lower part of Figure 1. The converted EL speech and the target NL speech are respectively input into the encoder of the ASR model to obtain the corresponding latent representations. The L1 loss between the two representations is minimized together with the traditional L1-based reconstruction loss to further reduce the difference between the converted EL speech and the target NL speech.

C. Data augmentation

Limited pEL training speech is a significant challenge for pEL-to-NL VC tasks. To address this issue, we use a data augmentation method based on the waveform similarity overlap-and-add (WSOLA) algorithm [9] to increase the amount of

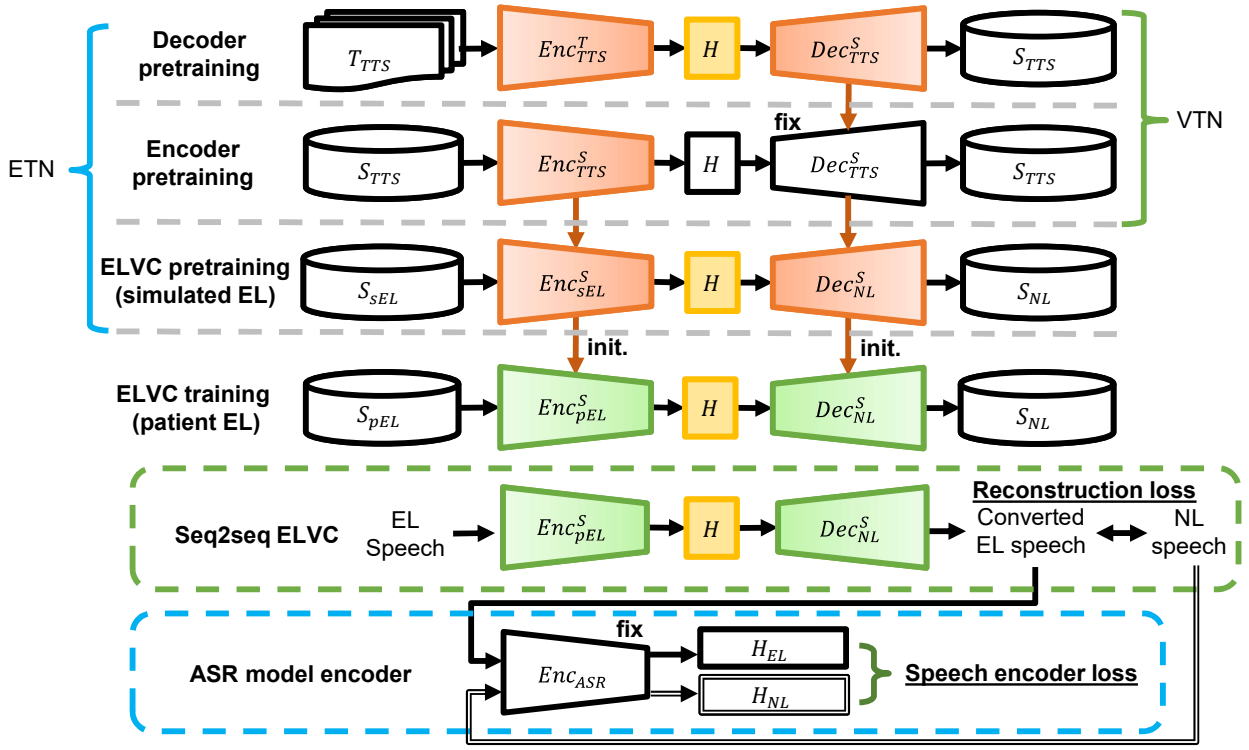

Fig. 1. The upper part is the training flow of our ELVC model, which can be divided into decoder pretraining, encoder pretraining, ELVC pretraining on simulated EL speech, and ELVC training on patient EL speech. The lower part depicts the ELVC training loss. When combining the speech encoder loss to train the ELVC model, the converted EL speech and the target NL speech are respectively input into the encoder of an ASR model to obtain the corresponding latent representations, and the difference between the two representations is minimized together with the traditional reconstruction loss to further reduce the difference between the converted EL speech and the target NL speech.

pEL training speech. Since pEL speech tends to be slower than NL speech, we shortened the duration of pEL speech, starting from the original duration (100%) and gradually reducing it by 5% each time until reaching 80% of the original duration. In this way, each pEL speech utterance was augmented into five versions (including the original version). The main advantage of this approach is its rapid and simple implementation. Unlike TTS-based data augmentation [6], which requires training a TTS model and fine-tuning it on EL speech to produce synthetic EL speech, this method requires less computational resources and requires no training. Furthermore, the EL speech produced by TTS-based methods may be significantly different from the EL speech, and using WSOLA to appropriately modify the duration will not cause this difference.

III. Experimental Setup

A. Datasets

We adopted the TMHINT text set [10] as the script for recording the sEL speech dataset. The TMHINT text set was designed taking into account diverse and balanced Mandarin phonemes. There are 320 sentences in total, and each sentence has 10 Chinese characters. Seven healthy speakers (four males and three females) participated in the recording process. All audio files were recorded in a studio environment. The sEL speech was recorded using a Nu-Vois III Digital electrolarynx. Each speaker recorded sEL speech and NL speech separately for each of the 320 sentences. Therefore, the sEL dataset is a parallel ELVC corpus. The pEL speech dataset only contains the pEL speech of two male patients, which were also recorded based on the TMHINT sentences using a Nu-Vois III Digital electrolarynx, with 320 pEL speech utterances per patient. Unlike the sEL dataset, all pEL audio files were recorded in the hospital's treatment clinic room.

We used the TMSV dataset [11] to augment NL speech in ELVC pretraining (see Figure 1). This dataset contains the NL speech of 16 speakers, also recorded based on the TMHINT sentences. We only used the speech utterances corresponding to the first 240 TMHINT sentences in the sEL, pEL, and TMSV datasets for ELVC pretraining and training. The speech utterances corresponding to the next 40 sentences were used as the development set, and the speech utterances corresponding to the last 40 sentences were used as the test set.

Additionally, the COSPRO dataset [12] was used for VTN pretraining (and the first two pretraining stages of ETN) and vocoder training. It contains Taiwanese-accented Mandarin read or spontaneous speech of 109 speakers, approximately 44.4 hours. The MATBN dataset [13] was used to train the ASR model, which was used to evaluate the syllable recognition error rate of the converted speech and provide the speech encoder loss in ELVC training. It is a Taiwanese-accented Mandarin broadcast news corpus consisting of 196 hours. A sampling rate of 16 kHz was used uniformly for all corpora used.

B. Training settings

As shown in Figure 1, in VTN pretraining, the COSPRO dataset was first used for decoder pretraining (training the TTS model consisting of Enc^{T}_{TTS} and Dec^{S}_{TTS}), and then used for encoder pretraining through self-reconstruction of audio input/output (training the speech encoder Enc^{S}_{TTS} under the fixed decoder Dec^{S}_{TTS}). Note that we only used NL speech in VTN pretraining.

The first two stages of ETN pretraining are the same as VTN pretraining. We evaluated two setings for the third pretraining stage of ETN (i.e., ELVC pretraining in Figure 1). In the first setting, we selected two males and two females from the sEL dataset for ELVC pretraining (denoted as ETN: from 4 source sEL speakers to 4 target NL speakers, for a total of 16 speaker pairs). In the second setting, we used the entire sEL dataset and the TMSV NL dataset for ELVC retraining (denoted as L-ETN: from 7 source sEL speakers to 23 target NL speakers, for a total of 161 speaker pairs).

For the ELVC task, we selected two male speakers as VC targets for the two male patients. In this work, we focused on training ELVC models in a one-to-one VC manner. We have two pEL speakers and two NL speakers, so four different ELVC models were developed. The reported experimental results are the average of the four ELVC models.

We implemented our system using the open-source ESPnet toolkit [14]. Both ELVC and ASR models are based on the transformer encoder-decoder architecture with multi-head self-attention. The feature used is the 80-dimensional Mel-spectra. The window length is set to 1024, and the frame shift is 256. VTN pretraining follows the transformer.v1 configuration outlined in [14]. We adopted Parallel WaveGAN (PWG) as the vocoder[1].

C. Evaluation metrics

We evaluated the proposed model with both objective and subjective evaluation metrics.

The objective metrics include Mel-cepstrum distortion (MCD), F0 root mean square error (F0 RMSE), F0 correlation coefficient (F0 CORR), and average absolute duration difference between the converted and target utterances (DDUR), and ASR error rate. MCD is a widely used metric for assessing the spectral envelope distortion between paired speech signals in the Mel-frequency domain. To calculate the MCD values, we use the WORLD vocoder[2] to extract 40-dimensional Mel-cepstral coefficients with a 5 ms frame shift, and then calculate the distortion for non-silent, time-aligned frame pairs. A smaller F0 RMSE value and a larger F0 CORR value indicate more accurate F0 conversion. A smaller DDUR value indicates that the converted speech has a similar duration to the target speech. We used three ASR systems to perform speech recognition evaluation on the converted speech, namely Google ASR[3], Whisper[4], and our self-trained Mandarin syllable recognition system. Considering the characteristics of Mandarin, we provide two error rates: syllable error rate (SER) and character error rate (CER). SER was evaluated by our own Mandarin syllable recognition system. CER evaluated by Google ASR and Whipser are denoted as G-CER and W-CER, respectively. The average SER, G-CER, and W-CER are 8.4%, 5.1%, and 2.6% for NL speech (NL07 and NL08), 88.5%, 95%, and 81% for sEL speech (sEL07 and sEL08), and 91.5%, 99.3%, and 111.8% for pEL speech (pEL01 and pEL02). It is clear that the error rates for the NL speech of two healthy speakers (NL07 and NL08) are low, while the error rates for the simulated EL speech of two healthy speakers (sEL07 and sEL08) and the EL speech of two patients (pEL01 and pEL02) are very high, almost completely unrecognizable. Given the critical importance of intelligibility in ELVC tasks, our evaluation prioritizes the ASR error rate as the primary metric, while also considering other metrics to ensure a comprehensive evaluation.

We conducted two listening tests. The first involves transcription and intelligibility assessment. Participants were required to complete two tasks: (1) listen to the speech and transcribe the speech content, similar to manual speech recognition; (2) listen to the speech and rate the mean opinion score (MOS) for intelligibility on a five-point scale, with 1 being the lowest and 5 being the highest. Even difficult-to-understand EL speech becomes easier to understand if listeners hear the corresponding normal speech first. In order to avoid this bias, through a special design, the content of the test utterances (including normal and pEL utterances and pEL utterances processed by various ELVC models) presented to each listener is different. The second one is AB testing. In this test, participants were presented with the conversion results of two different ELVC systems on each trial and were asked to choose the better one. Additionally, we used MOSA-Net+[5], a neural speech assessment model, to provide quality and intelligibility assessments for comparison with human listening tests. The quality score is also measured on a five-point scale, while the intelligibility score ranges from 0 to 1[6]. For both scores, a higher value indicates better performance.

D. Baseline models

We compared our model with two baseline models: CDVAE [15] and the VTN-based seq2seq ELVC model [7]. CDVAE is a representative frame-based ELVC model combined with self-supervised learning (SSL) features. Our model is developed on top of the VTN-based seq2seq ELVC model by replacing VTN pretraining with ETN pretraining and integrating multiple methods for improvement.

[1]We followed the open-source implementation at https://github.com/kan-bayashi/ParallelWaveGAN

[2]We adopted the Python wrapper for World Vocoder at https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[3]We used Google Cloud Speech API at https://github.com/Uberi/speech_recognition

[4]We adopted Whisper-large model at https://github.com/openai/whisper

[5]https://github.com/dhimasryan/MOSA-Net-Cross-Domain/tree/main/MOSA_Net+

[6]MOSA-Net+ is trained on the results of human listening tests in terms of the percentage of correctly recognized characters in an utterance, rather than the standard MOS score on a 5-point scale.

TABLE I
COMPARISON OF DIFFERENT PRETRAINING STRATEGIES FOR ELVC MODELS. VTN IS THE BASELINE WITHOUT ELVC PRETRAINING, ETN INDICATES ELVC PRETRAINING USING THE SMALL SEL-NL DATASET, AND L-ETN INDICATES ELVC PRETRAINING USING THE FULL SEL-NL DATASET.

Pretraining	MCD ↓	F0 RMSE ↓	F0 CORR ↑	DDUR ↓	SER (%) ↓	G-CER (%) ↓	W-CER (%) ↓
VTN	8.745	**29.712**	0.164	**0.397**	79.0	87.1	86.3
ETN	8.544	30.005	0.168	0.438	74.0	82.2	**78.9**
L-ETN	**8.471**	30.275	**0.176**	0.436	**71.5**	**80.3**	81.9

TABLE II
EXPERIMENTAL RESULTS USING EL DATA AUGMENTATION IN DIFFERENT STAGES. COMPARISONS WERE MADE BETWEEN NO USE, USE IN THE ELVC PRETRAINING STAGE, AND USE IN THE ELVC TRAINING STAGE.

EL data augmentation		MCD ↓	F0 RMSE ↓	F0 CORR ↑	DDUR ↓	SER (%) ↓	G-CER (%) ↓	W-CER (%) ↓
ELVC training (pEL)	ELVC pretraining (sEL)							
✗	✗	8.471	30.275	0.176	0.436	71.5	80.3	81.9
✗	✓	8.549	**30.061**	0.146	0.434	71.9	79.8	85.6
✓	✗	**8.377**	30.146	**0.194**	0.442	**67.1**	**74.9**	**81.3**
✓	✓	8.861	30.121	0.133	**0.385**	72.3	82.1	85.1

IV. EXPERIMENTAL RESULTS

Through objective evaluation, we first examine the effect of ETN pretraining. Next, we evaluate the effectiveness of speech rate-based augmentation of EL speech. Then, we show the results of incorporating the speech encoder loss into ELVC training. Finally, we provide ablation study results for all proposed methods. For subjective evaluation, we provide the results of two sets of human listening tests and the evaluation results of the neural assessment model MOSA-Net+.

A. Pretraining strategy

We first compare the effectiveness of different pretraining strategies for pEL-to-NL ELVC. The results in Table I show that ETN using sEL-NL data for an additional third pretraining stage outperforms VTN using only NL speech, achieving lower MCD and lower error rates in all ASR systems. Compared to ETN, L-ETN further improves the performance by using more sEL-NL data in the third pretraining stage, resulting in lower MCD and lower error rates in two of the three ASR systems. The results of three different pretraining methods (i.e., VTN, ETN, and L-ETN) show a clear trend: as the amount of simulated EL speech in pretraining increases, the MCD and ASR error rates gradually decrease. This verifies the effectiveness of incorporating an appropriate amount of simulated EL speech into pretraining to bridge the gap between natural speech and patient EL speech.

B. Data augmentation

Based on the results of the first experiment, the second experiment evaluates data augmentation of EL speech in the L-ETN pretraining setting. From Table II, we can see that, with limited pEL training speech, a simple speech rate-based data augmentation approach helps ELVC training, achieving lower MCD and lower error rates across all ASR systems. But this approach does not bring significant benefits when applied to augment sEL speech used in ELVC pretraining. The reason may be that the amount of sEL speech is relatively sufficient compared to pEL speech, and adding too much homogeneous training data with this simple approach may lead to overfitting, negatively affecting the effectiveness of ELVC modeling on patient EL speech. It can also be seen from Table II that using EL data augmentation simultaneously in ELVC pretraining and ELVC training will not further improve performance.

C. Speech encoder loss

The third experiment focuses on the impact of using the speech encoder loss in ELVC training in the L-ETN pretraining setting. As shown in the last two rows of Table III, the additional use of speech encoder loss in ELVC training does reduce ASR error rates, although the MCD is not reduced. By aligning the latent representations of the target NL speech and the output of the ELVC model through the speech encoder, the ELVC model can better ensure the integrity of phonemes and the accuracy of intonation in the converted speech. The speech encoder used here is the encoder of the ASR model used to objectively evaluate the syllable error rate. The additional use of speech encoder loss in ELVC training not only reduces the SER, but also reduces the CER of Google ASR and Whisper ASR, neither of which participate in ELVC training. This further shows the effectiveness of additionally using the speech encoder loss in ELVC training.

D. Ablation study

The results of the ablation study are shown in Table III. The first row represents the baseline VTN-based seq2seq ELVC system. From the second to the fourth row, we sequentially merge L-ETN pretraining, pEL data augmentation, and speech encoder loss, respectively. Comparing row 2 and row 1 shows that L-ETN outperforms VTN. Comparing row 3 and row 2 shows that pEL data augmentation is effective. Comparing row 4 and row 2 shows that additional use of the speech encoder loss helps improve performance. The ELVC model combining all the above methods achieves the lowest error rates (see the last row). The results show that these methods not only improve performance independently, but also complement each other to create synergistic effects.

TABLE III
ABLATION STUDY RESULTS. DA DENOTES PEL DATA AUGMENTATION, VC DENOTES RECONSTRUCTION LOSS, AND SE DENOTES SPEECH ENCODER LOSS.

Training loss	DA	Pretraining	MCD ↓	F0 RMSE ↓	F0 CORR ↑	DDUR ↓	SER (%) ↓	G-CER (%) ↓	W-CER (%) ↓
VC	✗	VTN	8.745	**29.712**	0.164	**0.397**	79.0	87.1	86.3
VC	✗	L-ETN	8.471	30.275	0.176	0.436	71.5	80.3	81.9
VC	✓	L-ETN	**8.377**	30.146	**0.194**	0.442	67.1	74.9	81.3
VC+SE	✗	L-ETN	8.474	30.103	0.183	0.410	69.0	78.1	78.3
VC+SE	✓	L-ETN	8.412	29.910	0.193	0.417	**66.4**	**74.4**	**76.1**

TABLE IV
LISTENING TEST RESULTS IN CER AND MOS OF INTELLIGIBILITY (LEFT), AB TEST RESULTS (MIDDLE), AND ASSESSMENT SCORES OF MOSA-NET+ (RIGHT).

	Intelligibility			AB test			MOSA-Net+	
Model	CER (%) ↓	MOS of Int. ↑		Prefer (%) ↑	Prefer (%) ↑		Quality ↑	Intelligibility ↑
pEL	95.0	1.075		-	-		1.679	0.455
CDVAE	99.6	1.181		2.08	-		2.228	0.668
VTN	88.2	1.996		**89.58**	10.00		3.259	0.947
Ours	**79.8**	**2.130**		-	**26.67**		**3.382**	**0.960**
NL	6.8	4.850	Nearly	8.33	63.33	NL	4.280	0.993

E. Subjective evaluation

The first subjective evaluation involves transcription and intelligibility assessment. Participants were required to complete two tasks: (1) listen to the speech and transcribe the speech content; (2) listen to the speech and rate the MOS for intelligibility on a five-point scale, with 1 being the lowest and 5 being the highest. The first task asked participants to perform speech recognition, so CER was evaluated in the same way as ASR. There were 33 participants in this listening test. The audio files included pEL speech, NL speech, and pEL speech converted by CDVAE, VTN-based model, and our best model. The results are shown in the left part of Table IV. For the average listener, transcribing EL speech is a challenging task. This difficulty arises primarily from non-speech noise signals, which can severely impair comprehension. Compared with the 6.8% error rate of NL speech, the error rate of pEL speech is as as high as 95.0%, reflecting the extremely low intelligibility of pEL speech. Comparing three ELVC systems revealed a surprising result: CDVAE achieved an error rate of 99.6%, even higher than that of pEL speech. In comparison, our model shows better performance than the baseline VTN-based model, with a relative error reduction of 9.5% (from 88.2% to 79.8%). As for intelligibility scores, pEL speech had the lowest score, as expected. CDVAE was unable to significantly improve the intelligibility score compared to pEL speech, possibly due to the noise introduced during its conversion process. The other two seq2seq ELVC models performed significantly better than CDVAE, with our model achieving the highest MOS score among all tested models.

We conducted two sets of AB tests: the first pairing CDVAE with the VTN-based model, and the second pairing the VTN-based model with our best model. Fifteen participants took part in the AB test. The results are shown in the middle part of Table IV. In the first set of AB tests, we observed that the VTN-based model showed a clear preference over CDVAE, with a difference of 87.5%. In the second set of AB tests, our best model outperformed the VTN-based model by 16.67%, although 63.33% of the ratings deemed the two indistinguishable. The results show that our model outperforms the VTN-based model, which in turn outperforms CDVAE. The results also show that, in addition to intelligibility, our model improves the overall listening experience.

The quality and intelligibility assessment results of MOSA-Net+ on the speech data used for subjective evaluation are shown in the right part of Table IV. The automated quality assessment results of MOSA-Net+ and the AB test show a consistent trend that our model outperforms the VTN-based model, which in turn outperforms CDVAE. Furthermore, the intelligibility scores of MOSA-Net+ align with the trend of CER and MOS of intelligibility in the left part of Table IV, with our model outperforming the VTN-based model, which in turn outperforms CDVAE. However, the intelligibility scores of the VTN-based model and our best model are very close to the intelligibility score of NL speech, which is different from the human evaluation results. The reason for this difference may be that the training data of MOSA-Net+ does not include good-quality but difficult-to-understand speech like converted EL speech. Although MOSANet+ may not precisely predict absolute subjective evaluation scores and ASR performance, its trends in relative improvement remain consistent. This consistency makes it a valuable reference, potentially reducing the need for costly subjective evaluations.

F. Spectrogram analysis

To qualitatively compare the conversion results, Figure 2 shows the spectrograms of EL speech, speech converted by CDVAE, VTN, and our best system, and NL speech. From the first row, we can notice that the horizontal pattern (the mechanical noise of the EL device) fills the EL speech. Compared with NL speech, EL speech loses detailed speech structure, and is much longer than NL speech. Compared to CDVAE-converted speech, the results from both seq2seq-based models have more detailed patterns in high-frequency bins and are closer in duration to NL speech. Comparing VTN-converted

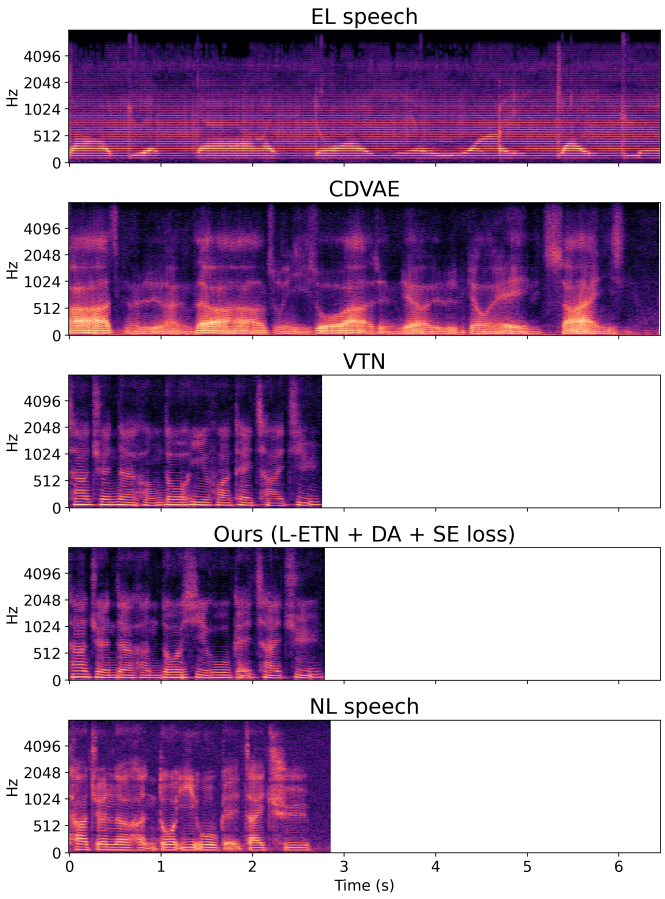

Fig. 2. Spectrograms of EL speech, speech converted by CDVAE, VTN, and our best system, and NL speech.

speech with speech converted by our best system utilizing the proposed techniques, our converted speech exhibits very similar detailed patterns to NL speech, highlighting its superior performance.

V. Discussion

While the proposed ELVC model significantly improves the intelligibility and quality of Mandarin EL speech, there is still a large gap between the current results and the acceptable levels for humans and ASR models. The challenge of maintaining recording quality in uncontrolled patient environments remains a significant issue. Since each patient's pathology is different, evaluation on a large number of patients is absolutely necessary. Although it is difficult to obtain authorization to use patient EL speech, we are working hard to collect more data through doctors. Due to lack of data, we are currently unable to verify the generalization of the model across languages. However, considering that the original VTN model architecture was applied to the English speaker VC task, and we applied it to the Mandarin ELVC task, there should be no difficulty in applying it to other languages. Exploring the characteristics of different patients and languages is a good experimental direction, but also presents significant challenges.

While we are currently focused on improving ELVC performance, i.e., improving the quality and intelligibility of EL speech, our ultimate goal is indeed to provide this service to laryngectomy patients via smartphones or other customized edge devices. Our ELVC system processes speech on an utterance-by-utterance basis, processing short utterances or segmenting long utterances into multiple shorter ones. Although completely real-time processing is not achieved, the delay is acceptable and does not significantly affect the user experience. The model has approximately 30 million parameters and currently runs on a standard GPU server. Since the large pre-trained model does not participate in the inference process, the model can be deployed on high-performance IoT devices. With continued advancements in chip technology, we anticipate that the proposed model can be capable of running on devices with more limited computing power and smaller memory in the future. It is also worth mentioning that with the popularization of network technology, many communications have gradually turned to video conferencing. Our system can run smoothly on ordinary computers, further improving the communication experience of electrolarynx users in video conferences and making conversations smoother.

VI. Conclusions

In this paper, we have introduced three methods to improve the existing seq2seq ELVC framework. Most previous work has evaluated simulated EL speech produced by healthy speakers using electrolarynxes, but our goal was to enhance EL speech of real patients. We performed additional pretraining using a simulated EL speech dataset to improve ELVC modeling and close the gap between patient EL speech and natural speech. Additionally, we augmented the patient EL training data using a simple but effective speech rate-based data augmentation method. We also incorporated an additional speech encoder loss into ELVC model training, thereby alleviating the problem of incomplete and missing phonemes of seq2seq ELVC. Through objective and subjective evaluations, we have confirmed that these methods not only improve performance independently but also complement each other to create synergistic effects.

Joint training that combines audio and visual information is a valuable direction for ELVC, especially since the performance of audio-only models is currently unsatisfactory. Future work will consider combining visual information, such as facial expressions and lip movements, with audio to assist in speech signal processing. The application scenario is to allow patients to use the electrolarynx to give speeches or communicate verbally in video conferences. In addition to using more simulated EL speech to pretrain the ELVC model and applying more effective data augmentation to augment patient EL training data, employing robust representations of larger-scale pretrained speech models (trained via self-supervised learning or supervised learning) is a promising direction to improve the quality and intelligibility of converted speech.

References

[1] G. Fant, *Acoustic theory of speech production*. The Netherlands: Mouton: The Hague, 1960.

[2] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," in *Proceedings of EUSIPCO*, 2018, pp. 2115–2119.
[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.
[4] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2014.
[5] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
[6] D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion." [Online]. Available: http://arxiv.org/abs/2210.10314
[7] M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. R. Jang, and H.-M. Wang, "Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling," in *Proceedings of ASRU*, 2021, pp. 650–657.
[8] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *arXiv e-prints*, p. arXiv:1912.06813, Dec. 2019.
[9] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proceedings of ICASSP*, 1993, pp. 554–557.
[10] M.-W. Huang, "Development of Taiwan Mandarin hearing in noise test," Master's thesis, National Taipei College of Nursing, Dept. Speech and Hearing Disorders and Sciences, 2005. [Online]. Available: http://140.131.94.7/handle/987654321/1917
[11] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processin*, vol. 30, pp. 1345–1359, 2022.
[12] C.-Y. Tseng, Y.-C. Cheng, and C.-H. Chang, "Sinica COSPRO and Toolkit: Corpora and platform of Mandarin Chinese fluent speech," in *Proceedings of O-COCOSDA*, 2005, pp. 23–28.
[13] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, Jun. 2005. [Online]. Available: https://aclanthology.org/O05-3004
[14] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *arXiv e-prints*, p. arXiv:1910.10909, Oct. 2019.
[15] H.-H. Chen, Y.-L. Chien, M.-C. Yen, S.-W. Tsai, Y. Tsao, T.-S. Chi, and H.-M. Wang, "Mandarin electrolaryngeal speech voice conversion using cross-domain features," in *Proceedings of INTERSPEECH*, 2023, pp. 5018–5022.