MANDARIN ELECTROLARYNGEAL SPEECH VOICE CONVERSION WITH SEQUENCE-TO-SEQUENCE MODELING

Ming-Chi Yen[1,2], Wen-Chin Huang[3], Kazuhiro Kobayashi[3], Yu-Huai Peng[2], Shu-Wei Tsai[4], Yu Tsao[5], Tomoki Toda[3], Jyh-Shing Roger Jang[1], and Hsin-Min Wang[2]

[1]National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Nagoya University, Japan
[4]Dept. of Otolaryngology, National Cheng Kung University Hospital, Taiwan
[5]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ABSTRACT

The electrolaryngeal speech (EL speech) is typically spoken with an electrolarynx device that generates excitation signals to substitute human vocal fold vibrations. Because the excitation signals cannot perfectly characterize sound sources generated by vocal folds, the naturalness and intelligibility of the EL speech are inevitably worse than that of the natural speech (NL speech). To improve speech naturalness, statistical models, such as Gaussian mixture models and deep-learning-based models, have been employed for EL speech voice conversion (ELVC). The ELVC task aims to convert EL speech into NL speech through an ELVC model. To implement a frame-wise ELVC system, accurate feature alignment is crucial for model training. However, the abnormal acoustic characteristics of the EL speech cause misalignments and accordingly limit the ELVC performance. To address this issue, we propose a novel ELVC system based on sequence-to-sequence (seq2seq) modeling with text-to-speech (TTS) pretraining. The seq2seq model involves an attention mechanism to concurrently perform representation learning and alignment. Meanwhile, TTS pretraining provides efficient training with limited data. Experimental results show that the proposed ELVC system yields notable improvements in terms of standardized evaluation metrics and subjective listening tests over a well-known frame-wise ELVC system.

Index Terms: electrolaryngeal speech, voice conversion, sequence-to-sequence learning, transformer, pretraining

1. INTRODUCTION

For people who have undergone laryngectomy or are unable to use their larynx normally, the electrolarynx is a suitable assistive device that can help them produce speech; the produced speech is termed electrolaryngeal speech (EL speech). Although EL speech provides acceptable intelligibility [1], its quality is still far from natural speech (NL speech). Statistical voice conversion (VC) methods, which convert the source type of speech into a target type with the originally underlying content, have been applied to improve the quality of EL speech [2, 3, 4, 5]. We refer to this VC task as ELVC, which requires the use of a parallel corpus containing pairs of source EL speech and target NL speech. A VC system is divided into three stages: the first stage is the feature extraction of source and target speech, the second stage is model training, and the last is the assembly of the converted feature sequences into waveform. In the first stage, the alignment between the source feature sequence and the target feature sequence is a crucial step. Especially for frame-wise VC systems, temporal alignment must be performed before model training. The dynamic time warping (DTW) approach is most widely employed to find the best alignment path of two feature sequences by estimating the similarity of acoustic features through a distance measurement (e.g., L2 distance).

The training of frame-wise ELVC systems relies heavily on accurate source-target (EL-NL) feature alignment. However, the abnormal characteristics of EL speech may cause incorrect alignment of EL and NL feature sequences. Moreover, the speaking rate of EL speakers is usually slower than that of normal speakers, which further increases the difficulty of accurate EL-NL feature alignment. The left panel of Figure 1 shows the DTW alignment result of the EL-NL mel-cepstrum (MCEP) sequences. The human labeled Mandarin syllable boundaries of EL speech and NL speech are presented by gray dash lines. Each solid black rectangle represents an area where EL speech and NL speech share the same Mandarin syllable. The blue line denotes the alignment path of EL speech and NL speech. Obviously, an alignment path that is always located in the solid black rectangles is a more accurate alignment of the EL feature sequence and the NL feature sequence. For the ideal case, the alignment path should be approximately diagonal in each syllable rectangle. However, we can see that the alignment path in the left panel of Figure 1 is not good, especially in the back section.

In this study, we propose to apply seq2seq modeling to

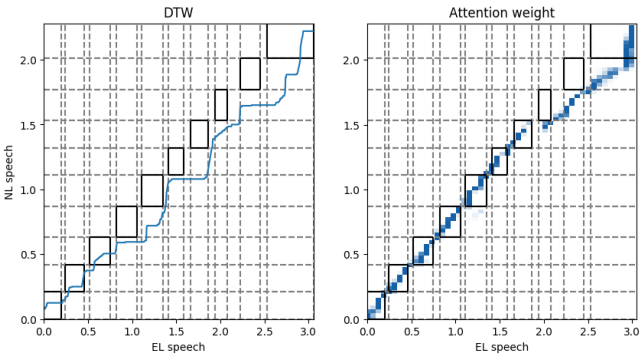

Fig. 1. A comparison of alignment results of DTW on MCEPs (the left panel) and the attention weights of seq2seq modeling (the right panel). The gray dash lines denote the human labeled Mandarin syllable boundaries. The solid black rectangles represent the regions where EL and NL share the same Mandarin syllable. The blue lines denote the alignment paths of EL/NL speech. The thicker blue line in the right panel is due to rescaling the downsampled EL/NL embedding to seconds.

overcome the aforementioned problems of frame-wise VC methods in the ELVC task. The proposed seq2seq model uses an attention mechanism [6, 7] to perform representation learning and alignment of source and target acoustic feature sequences with different lengths. The alignment result of the seq2seq model on the same EL-NL pair is shown in the right panel of Figure 1. Comparing the left and right panels in Figure 1, we note that the seq2seq model provides more accurate EL-NL feature sequence alignment. Moreover, compared with frame-wise VC systems, the seq2seq architecture typically has a larger receptive field and can capture the long-term dependency of a given input utterance. This is a unique attribute that meets the requirement of EVLC in particular. Another challenge in building a decent ELVC system is that collecting a large amount of EL speech is prohibitively expensive. To meet the data-hungry property of the seq2seq VC model, the proposed ELVC system uses the TTS pretraining technique [8]. As reported in [9], compared to the system trained from scratch, the seq2seq VC system using a pretrained TTS model provides better prosodic conversion performance and is easier to fine-tune on a resource-constrained corpus. Our experimental results show that compared with a well-known frame-wise ELVC system [10], the proposed system achieves clear improvements in terms of automatic speech recognition (ASR) accuracy, mel-cepstrum distortion (MCD), a standardized objective evaluation metric for VC, and subjective listening tests.

The major contributions of this study include:

- We employ seq2seq modeling on the ELVC task to tackle the alignment issue of the frame-wise ELVC approach.

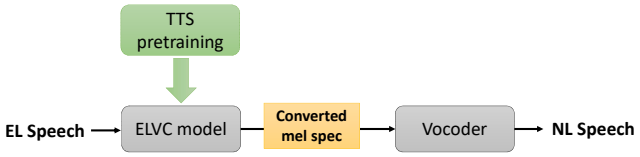

Fig. 2. System diagram of the proposed ELVC system with TTS pretraining.

- We present the first work of using seq2seq modeling to carry out comprehensive systematic measurements and analyses of ELVC.
- We use a pretrained TTS model to handle the data-insufficiency issue of the ELVC task.

The findings of this study provide useful experience in the field of ELVC research and system implementation.

2. RELATED WORK

Human speech production typically comprises two parts. The first part concerns vocal fold vibrations, which generate excitation signals and transmit tones. The second part considers the shape of the vocal tract and articulation places, which deliver timbres and other acoustic details of the speech. When a specific speech organ in the above two parts is injured, the speech signal may be disordered; therefore, speech intelligibility will be reduced. Individuals who undergo total laryngectomy are called laryngectomee and usually lose the ability to generate vocal fold vibrations to produce NL speech. The electrolarynx device can help laryngectomees produce speech signals by generating mechanical excitation signals from outside the body instead of vocal fold vibration. The excitation signals pass through the user's vocal tract and articulation places to generate EL speech. Although EL speech sounds more intelligible [1] than other types of alaryngeal speech (e.g., esophageal speech), the quality of EL speech suffers from two issues. First, the mechanical noise from the EL device notably deteriorates the quality and intelligibility of EL speech. Second, the acoustic characteristics of EL excitation signals are very different from natural excitation sources. Consequently, EL speech sounds mechanical and artificial, presenting different acoustic characteristics from NL speech [2].

In earlier works, noise suppression (NS) [11] and VC [12] are two widely adopted methods for improving quality of EL speech. Signal-processing-based NS methods, such as spectral subtraction [13, 14, 15] and Wiener filter [16], have been applied to suppress mechanical noise. Although these works can effectively reduce noise components, the main weakness of these approaches is generating unwanted accompanying musical noise in the reduction process. Moreover, these NS approaches neglect acoustic properties when modifying the

EL speech leads the enhanced speech sounds unnatural for human hearing.

Conversely, VC-based approaches focus on converting the acoustic components [12], such as spectral envelop and pitch patterns, from source to target speech. In the past, most ELVC systems are implemented in a frame-wise manner [3, 4]. For these systems, an utterance-wise parallel EL/NL speech corpus with a precise alignment of acoustic features is crucial to achieving satisfactory performance. In [1], an ELVC system based on the Gaussian mixture models was proposed and shown to effectively improve the naturalness and speaker identity of EL speech. More recently, deep-learning-based models have been applied to ELVC and achieved further improvements [2, 10]. Although these works have effectively improved EL speech, the achievable performance is actually bounded by the accuracy of the alignment. Therefore, the seq2seq model serves as a better choice for establishing ELVC systems. [17] is the first work to apply the seq2seq model to ELVC, but only provides a few converted audio samples. This paper presents the first comprehensive and systematic measurement and analysis of seq2seq ELVC. In this study, we propose to improve the seq2seq VC model by using a multitask learning criterion and adopting the TTS pretraining technique. Moreover, detailed analyses of the advantages of using the seq2seq VC model over the frame-wise VC model on the ELVC task are provided.

3. BASELINE CLDNN-BASED ELVC SYSTEM

The CLDNN [18] model was originally proposed as an acoustic model in ASR, which includes convolutional neural network (CNN) layers, long short-term memory (LSTM) recurrent layers, and fully connected (FC) layers. In [2], the CLDNN model was applied to VC and yielded satisfactory performance. This study uses a CLDNN-based VC system as the baseline system. Network architecture and training/conversion processes are introduced in this section.

3.1. Network architecture

For the CLDNN-based VC system, the CNN layers perform acoustic feature extraction, the recurrent layers model the dynamic characteristics of speech patterns, and the FC layers learn the nonlinear mapping between the source and target acoustic features. To reduce the model complexity, bidirectional gated recurrent units (Bi-GRUs) are used in the recurrent layers [2]. To further improve the performance, we follow the design of a multitask CLDNN (MT-CLDNN) VC system [10], where the objective function used to train the MT-CLDNN model is contributed by the losses of multiple acoustic features.

3.2. Preprocessing, training, and conversion processes

In the preprocessing stage, the sequences of the source and target feature vectors were extracted and aligned by DTW. The input feature vector consists of mel-cepstrum features only, whereas the output feature vector include four types of features: mel-cepstrum, aperiodicity (AP), continuous F0, and unvoiced/voiced (U/V) symbols. To capture temporal information more effectively, contextual feature vectors are used together to form the final input vector of the CLDNN model.

In the training stage, the CLDNN model is trained based on the multi-task objective function, which considers the losses of spectral features (mel-cepstrum and AP) and prosodic features (continuous F0 and U/V symbols). In the conversion stage, the CLDNN model takes the contextual mel-cepstrum features of the EL speech as the input and outputs four types of features: converted mel-cepstrum, AP, continuous F0, and U/V symbols. The final F0 is obtained by combining the information of the continuous F0 and U/V symbols. Based on the converted mel-cepstrum, AP, and F0, the MLSA method [19] is used as the synthesis filter to generate the converted speech.

Promisingly, [2, 10] confirm the effectiveness of the CLDNN-based models for improving the naturalness and speaker identity of ELVC. However, as mentioned earlier, the performance of the frame-wise VC system may become suboptimal when the alignment is not accurate. Moreover, frame-wise VC systems may not adequately capture long-term information for precise prosody conversions. These reasons motivated us to step forward to research seq2seq modeling for ELVC.

4. SEQ2SEQ ELVC WITH TTS PRETRAINING

Figure 2 shows a flowchart of the seq2seq ELVC system. The ELVC system generates the converted acoustic features, which are then converted into waveform by a vocoder. The pretrained TTS in the ELVC system aims to enable effective training and reliable prosody conversion through a small-scale ELVC corpus. Assume that we have abundant training data for TTS pretraining: $\boldsymbol{D}_{TTS}$=$\{\boldsymbol{T}_{TTS}, \boldsymbol{S}_{TTS}\}$, where $\boldsymbol{T}_{TTS}$ is the text transcription, and $\boldsymbol{S}_{TTS}$ is the corresponding speech data. Further, assume that we have collected a resource-limited ELVC corpus, consisting of sentence-wise parallel EL/NL utterances: $\boldsymbol{D}_{ELVC}$=$\{\boldsymbol{S}_{EL}^{sid}, \boldsymbol{S}_{NL}^{sid}\}$, where $\boldsymbol{S}_{EL}$ is the EL speech, $\boldsymbol{S}_{NL}$ is the NL speech, and sid denotes the speaker identity. Figure 3 shows the overall training flow of the ELVC system, which can be divided into three stages: decoder pretraining, encoder pretraining, and ELVC model training. In the following subsections, we introduce these three stages in detail.

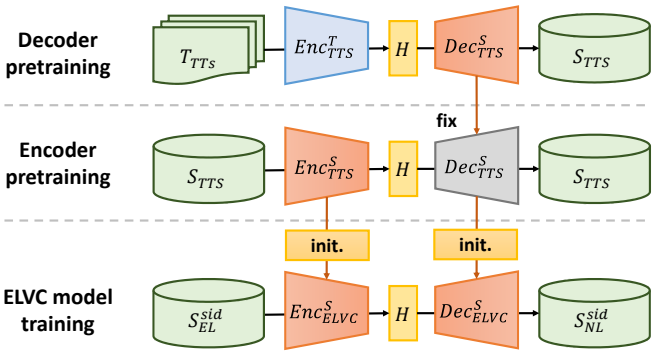

Fig. 3. Overall training flow of the ELVC system, which can be divided into decoder pretraining, encoder pretraining, and ELVC model training.

4.1. Decoder pretraining

In this stage, we train a transformer-based TTS model with an encoder-decoder architecture using $\boldsymbol{D}_{TTS}$. The input and output of the TTS model are $\boldsymbol{T}_{TTS}$ and $\boldsymbol{S}_{TTS}$, respectively. Because the text data provide linguistic information, the encoder Enc^{T}_{TTS} is forced to learn to encode text data to hidden representations, which can be converted into the target speech features by Dec^{S}_{TTS}. Because a large-scale corpus is used, the decoder, Dec^{S}_{TTS}, can be pretrained well.

4.2. Encoder pretraining

This stage trains an encoder Enc^{S}_{TTS} that can encode acoustic features to hidden representations, which can be further converted into the target acoustic features by the pretrained decoder, Dec^{S}_{TTS}. The input and output speech samples are from $\boldsymbol{S}_{TTS}$. When training Enc^{S}_{TTS}, we fix the parameters of the pretrained decoder, Dec^{S}_{TTS}.

4.3. ELVC model training

Finally, we use the EL/NL corpus, $\boldsymbol{D}_{ELVC}$, to fine tune the pretrained encoder, Enc^{S}_{TTS}, and the pretrained decoder, Dec^{S}_{TTS}, into the ELVC model. Notably, the model parameters of the pretrained encoder and decoder serve as informative priors to perform efficient model adaptation with limited ELVC data. Compared to the train-from-scratch (TFS) strategy, the ELVC system with the pretrained encoder and decoder requires less than 50% of training time to converge. In other words, the pretraining of encoder and decoder models enables the ELVC system to be an extremely efficient training process [8].

5. EXPERIMENTS

In this section, we present the experimental setup, results, and discussion of our findings.

5.1. Experimental setup

We evaluated the proposed system on a Mandarin parallel ELVC corpus. The utterances in the corpus were provided by two male speakers, denoted as s01 and s02 in the following discussion. The speaker s01 provided 320 pairs of EL/NL utterances, where the EL utterances were generated using the electrolarynx device[1]. The speaker s02 only provided 320 NL utterances. The scripts of these utterances were from the Taiwan Mandarin hearing in noise test (TMHINT) sentences [20], which were designed to have a phoneme and tone balance. Each sentence consisted of ten Chinese characters.

Based on the utterances provided by s01 and s02, we designed two tasks. The first task is to convert the EL utterances to NL ones for the same speaker s01. We denote this task as the EL01-NL01 task. The second task is to convert the EL utterances of s1 to NL ones of s2. We denote this task as the EL01-NL02 task. For both tasks, we split the 320 EL/NL utterances into 240, 40, and 40 to form the training, development, and evaluation sets, respectively. To pretrain the TTS model, we used the Mandarin continuous speech prosody corpora (COSPRO) [21], which contained 59,492 utterances (from 109 speakers, approximately 44.4 hours). The speech signals were re-sampled into 16 kHz. The TTS model was constructed in a multi-speaker transformer-based architecture.

We trained the ELVC system, as shown in Figure 3, using the open-source ESPnet toolkit [22, 23]. The speech waveforms were first converted into 80-dimensional mel-spectral features, with a window size of 1024 points and a frame shift of 256 points. We followed the Transformer.v1 configuration in [23] to pretrain the TTS model, and then implemented the ELVC system following the procedures in [8]. The reduction factors r_e and r_d were both set to 2 in the ELVC system. We used the Parallel WaveGAN (PWG)[2] [24] to generate speech waveform from spectral and prosodic features. PWG is a non-autoregressive variant of the WaveNet vocoder [25, 26]. Compared with WaveNet, PWG can generate speech waveform more efficiently while maintaining comparable speech quality. Our previous study revealed that a speaker-dependent neural vocoder outperformed a speaker-independent one [27]. Therefore, we trained a speaker-dependent PWG vocoder in the proposed ELVC system.

We evaluated the proposed ELVC system using the following objective evaluation metrics, namely the mel-cepstrum distortion (MCD), the syllable error rate (SER) of the ASR of converted speech, the F0 root mean square error (F0 RMSE), the F0 correlation coefficients (F0 CORR), and the average absolute duration difference between the converted and target utterances (DDUR). MCD is a common metric to measure the spectral envelope distortion of paired

[1] Nu-Vois III Digital™

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

Table 1. Objective evaluation results of electrolarynx speech voice conversion. TFS and PT denote the results of ELVC systems trained from scratch and pretrained with TTS, respectively.

Spk	Model	Pretraining	MCD (dB)	F0 RMSE	F0 CORR	DDUR	SER (%)
EL01-NL01	TFS	✗	8.86	24.44	0.202	**0.156**	93.3
	PT	✓	**7.10**	24.72	**0.212**	0.167	**67.5**
	MT-CLDNN		7.38	**24.38**	0.167	0.680	76.5
EL01-NL02	TFS	✗	11.17	34.41	0.365	**0.178**	99.0
	PT	✓	8.18	**33.50**	**0.458**	0.192	**75.0**
	MT-CLDNN		**7.77**	35.58	0.336	0.914	85.0

Table 2. Subjective listening test results (MOS in a five-point scale) with 95% confidence interval.

	Naturalness (MOS)		
System	PT	MT-CLDNN	Target
EL01-NL01	$\mathbf{3.38 \pm 0.71}$	2.00 ± 0.58	4.91 ± 0.12
EL01-NL02	$\mathbf{3.30 \pm 0.64}$	1.57 ± 0.65	5.00 ± 0.00

speech signals in the mel-frequency domain. To compute the MCD values, we used the WORLD vocoder [28] to extract 24-dimensional mel-cepstrum coefficients with a 5 ms frame shift, and calculated the distortion of non-silent, time-aligned frame pairs. We trained an ASR system based on the Transformer model [29] on the AISHELL-1 corpus [30]. The SER values of the NL and EL speech (treated as the upper/lower-bound ASR results) were 17.3% and 84.3%, respectively. A smaller F0 RMSE value and a larger F0 CORR value indicate more accurate F0 conversion. A smaller DDUR value indicates that the converted utterance and the target utterance have a similar length. In addition to the objective evaluation based on the above metrics, we also conducted a listening test to subjectively measure the naturalness of the converted speech. The participants were asked to evaluate the naturalness of a given speech utterance based on the mean opinion score (MOS) in a five-point scale.

5.2. Experimental results

In this section, we first present the objective and subjective evaluation results of the proposed ELVC system and the baseline MT-CLDNN system for the EL01-NL01 and EL01-NL02 tasks. Spectrogram plots are then presented to visually compare the converted speech utterances obtained from different ELVC systems.

5.2.1. Performance comparison on the EL01-NL01 task

We first conducted a systematic comparison to verify the effectiveness of TTS pretraining. The evaluation results are listed in Table 1, where the results of the EL01-NL01 task (the same speaker conversion task) are listed in the upper rows. In addition to the ELVC system with TTS pretraining (denoted as PT in Table 1), we trained another ELVC system from scratch (denoted as TFS in Table 1) without pretraining.

It can been seen from the results that TFS is worse than PT for most evaluation metrics. In particular, the SER value provided by TFS is notably higher than that of PT, showing that compared with PT, the speech converted by TFS is more difficult to be correctly recognized by the ASR system. TFS and PT achieve similar DDUR values because the same TTS model architecture was used for the two strategies. By comparing the results of TFS and PT, we confirmed the effectiveness of the TTS pretraining of the ELVC system.

Next, we compared the proposed ELVC system (PT) and the baseline MT-CLDNN system. From Table 1, we note that PT notably outperforms MT-CLDNN in terms of MCD and SER. The results confirm that the seq2seq model can generate speech signals with less spectral distortion, thus achieving better ASR results. The advantages of the seq2seq ELVC system are also prominent in the evaluation metrics related to prosody, namely F0 CORR and DDUR. The DDUR value was reduced from 0.680 to 0.167, while the F0 CORR was increased from 0.167 to 0.212.

Table 2 shows the results of the subjective listening test. For the EL01-NL01 task, it is clear that PT outperforms MT-CLDNN in terms of speech naturalness with a notable margin. The result again confirms the advantages of the seq2seq model over the frame-wise ELVC model.

5.2.2. Performance comparison on the EL01-NL02 task

In real-world scenarios, the NL speech of the laryngectomees may not always be available, and the conversion performance of the EL01-NL02 task needs to be investigated. The results of the objective evaluation metrics are listed in the lower three rows (TFS, PT, MT-CLDNN) in Table 1. We observed trends similar to those observed for the EL01-NL01 task. First, the ELVC system with TTS pretraining (PT) yields better performance than TFS with lower MCD, F0 RMSE, and SER scores

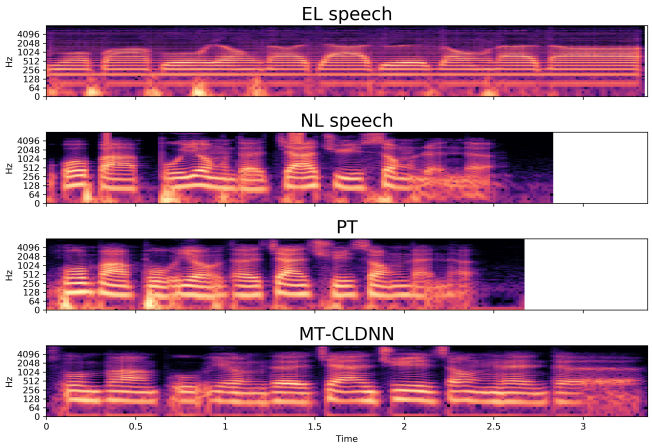

Fig. 4. Spectrogram plots of EL speech, NL speech, PT converted speech, and MT-CLDNN converted speech.

and higher F0 CORR score. Compared to MT-CLDNN, despite the slightly higher MCD scores, PT achieves considerably better performance in terms of prosody-related evaluation metrics (F0 RMSE, F0 CORR, and DDUR) and ASR results. Notably, the EL01-NL02 task aims to convert speech utterances between different speakers. Therefore, the benefits of seq2seq modeling are clearer in these prosody-related evaluation metrics than the EL01-NL01 task.

In terms of subjective listening tests, the results in the lower part of Table 2 also imply that PT outperforms the baseline MT-CLDNN, again confirming the effectiveness of the seq2seq model and the TTS pretraining technique.

5.2.3. Spectrogram analysis

To qualitatively compare the conversion results, Figure 4 demonstrates the spectrogram plots of the EL speech, NL speech, PT converted speech, and MT-CLDNN converted speech. From the figure, we can first note that some detailed speech structures are lost in the EL speech as compared to the NL speech. Further, as compared to the MT-CLDNN-converted speech, the PT-converted speech has detailed patterns that are relatively more similar to the NL speech. Moreover, the MT-CLDNN-converted speech contains obvious noise components at the end of the utterance Finally, the PT-converted speech has a length similar to the NL speech. Please note that the NL speech is at a natural speaking speed. This is another clear advantage of seq2seq modeling for ELVC over frame-wise VC systems.

6. CONCLUSIONS

This study proposes a novel ELVC system formed by a seq2seq model with a pretrained TTS to address the potential misalignment issues that often occur in frame-wise ELVC systems. The experimental results reveal that the proposed ELVC system outperforms the baseline MT-CLDNN system in terms of several objective and subjective evaluations. We also confirm the effectiveness of the pretrained TTS model trained on a large-scale TTS corpus. In the future, we will refine the current seq2seq ELVC system by including phonetic posteriorgrams (PPGs) [31] to address mispronunciation and skipped phoneme issues. ASR model [32] as an extra clue is a direction to improve the quality of converted speech.

7. ACKNOWLEDGMENTS

This work was partly supported by MOST-Taiwan Grant: 107-2221-E-001-008-MY3, Tsvoice LLC Project, and JST CREST Grant Number JPMJCR19A3, Japan.

8. REFERENCES

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 6, pp. 1429–1437, 2014.

[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, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *Proceedings of ICASSP*, 2011, pp. 5136–5139.

[5] 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.

[6] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

[7] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of EMNLP*, 2015.

[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] H. Luong and J. Yamagishi, "Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech," in *Proceedings of ASRU Workshop*, 2019, pp. 200–207.

[10] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task cldnn," in *Proceedings of 2020 28th EUSIPCO*, 2021, pp. 396–400.

[11] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," in *Proceedings of the IEEE*, vol. 67, no. 12, 1979, pp. 1586–1604.

[12] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, 1990.

[13] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] Boh Lim Sim, Yit Chow Tong, J. S. Chang, and Chin Tuan Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, 1998.

[15] K. Wojcicki, B. Shannon, and K. Paliwal, "Spectral subtraction with variance reduced noise spectrum estimates," *Proceedings of the 11th Australasian International Conference on Speech Science and Technoligy*, 01 2006.

[16] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[17] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1849–1863, 2020.

[18] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of ICASSP*, 2015, pp. 4580–4584.

[19] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[20] 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

[21] 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.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of INTERSPEECH*, 2018, pp. 2207–2211.

[23] 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.

[24] R. Yamamoto, E. Song, and J. M. Kim, "Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *Proceedings of ICASSP*, 2020.

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv e-prints*, p. arXiv:1609.03499, Sep. 2016.

[26] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proceedings of INTERSPEECH*, 2017.

[27] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proceedings of ASRU Workshop*, 2018.

[28] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[29] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proceedings of ICASSP*, vol. 2018-April, 2018.

[30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proceedings of O-COCOSDA*, 2017, pp. 1–5.

[31] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[32] J. Zhang, Z. Ling, Y. Jiang, L. Liu, C. Liang, and L. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *Proceedings of ICASSP*, 2019, pp. 6785–6789.