EXEMPLAR-BASED METHODS FOR MANDARIN ELECTROLARYNGEAL SPEECH VOICE CONVERSION

Hsin-Te Hwang[1], Chia-Hua Wu[2], Ming-Chi Yen[2], Yu Tsao[1], Hsin-Min Wang[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
hwanght@citi.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw
[2]Institute of Information Science, Academia Sinica, Taiwan
whm@iis.sinica.edu.tw

ABSTRACT

Electrolarynx (EL) can restore speech to patients without vocal cords, but the resulting speech often lacks naturalness and intelligibility. Inspired by recent advances in exemplar-based methods for speaker voice conversion (VC), this study explores their application to electrolaryngeal speech voice conversion (ELVC). As a baseline, we apply the k-nearest neighbors VC (KNN-VC) method with self-supervised learning (SSL) representations to ELVC. To circumvent the SSL feature mismatch between EL speech and natural speech when applying KNN-VC to ELVC, we propose two methods under the parallel VC setting: the parallel KNN-ELVC method and the parallel locally linear embedding (LLE)-based ELVC method. Experimental results demonstrate that while directly applying KNN-VC to ELVC improves the naturalness of EL speech, our proposed methods, especially LLE-ELVC, significantly improve both naturalness and intelligibility. Furthermore, LLE-ELVC is comparable to our previous deep-learning-based ELVC method, further demonstrating the potential of exemplar-based methods in ELVC tasks.

***Index Terms*—** electrolaryngeal speech, voice conversion, self-supervised learning, exemplar-based, k-nearest neighbors, locally linear embedding

1. INTRODUCTION

Laryngectomy, a surgery often performed for laryngeal cancer, removes the larynx and vocal cords, resulting in the loss of natural voice. Patients undergoing this procedure typically speak using an electrolarynx (EL), a device that generates mechanical signals to simulate vocal cord vibrations. However, EL speech often has poor sound quality, low intelligibility, and unnatural intonation, which seriously hinder effective communication.

To address these challenges, a number of electrolaryngeal speech voice conversion (ELVC) techniques have been developed to enhance EL speech by converting it into more natural-sounding speech while preserving the original linguistic content [1–9]. Early ELVC methods relied on frame-level Gaussian mixture models (GMMs) and made initial progress in improving speech quality [1, 2]. With the advent of deep learning, more sophisticated methods have been introduced, including deep neural networks (DNNs) [3–5] and sequence-to-sequence (seq2seq) models [6, 7], to further advance the field by improving frame alignment and generating more natural-sounding speech. Additionally, representations of large-scale self-supervised learning (SSL) speech models have recently been applied to the ELVC task and achieved significant improvements [7–9].

Exemplar-based methods that exploit SSL features have recently received attention for their effectiveness in speaker voice conversion (VC) and related tasks. A prominent example is k-nearest neighbors (KNN)-based voice conversion (KNN-VC) [10], which uses SSL features as shared content representation. Through simple KNN regression, the target SSL features corresponding to the source SSL features are identified, thereby converting the source speaker's speech into the target speaker's speech. This method stands out for its simplicity, ability to perform VC using only target data, and strong performance in any-to-any VC scenarios. In addition to speaker VC, it has also been applied to other tasks such as singing VC [11], dysarthric speech VC [12], and speech emotion recognition [13].

Motivated by these recent successes, we investigate exemplar-based methods for ELVC. Unlike typical speaker VC or related tasks [10–13], EL speech presents unique challenges due to the significant acoustic differences between EL speech and natural (NL) speech. Directly applying any-to-any KNN-VC to ELVC may result in poor performance due to the mismatch in SSL features between EL speech and NL speech. To circumvent the mismatch problem when directly applying KNN-VC to ELVC, we propose a parallel KNN-ELVC method under the parallel VC setting, which performs SSL feature matching in the EL speech domain to select target EL feature neighbors and then uses the corresponding target NL exemplars for conversion. In addition, we also propose a locally linear embedding (LLE)-based method for ELVC,

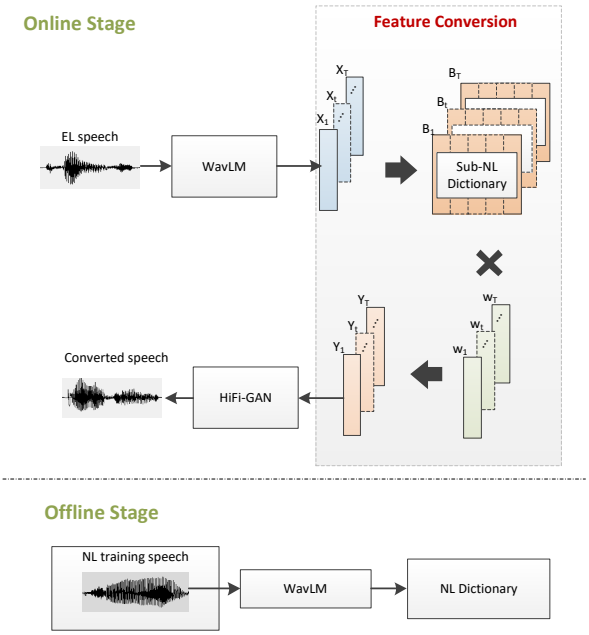


Fig. 1. The **KNN-ELVC** system.

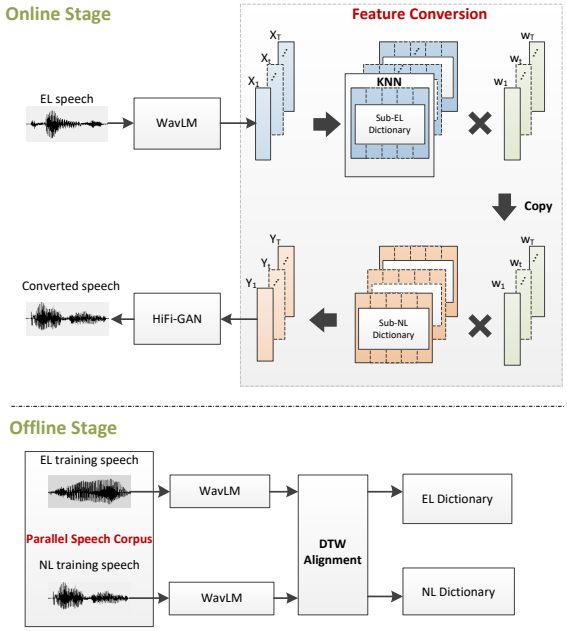


Fig. 2. The proposed parallel ELVC systems.

which combines ideas from our previous work on speaker VC [14] and KNN-VC.

The main contributions of this study are:

1. We propose two novel parallel ELVC methods based on KNN-VC and LLE-VC, which improve performance over baseline methods.
2. Our results demonstrate that the proposed methods, especially the LLE-ELVC method, improve the naturalness and intelligibility of EL speech. These methods show competitive performance to our previous deep-learning-based method [9], highlighting the potential of exemplar-based methods for ELVC.

The remainder of this paper is organized as follows. Section 2 introduces the application of KNN-VC to ELVC. Section 3 presents the proposed parallel ELVC methods. Sections 4 and 5 present the experimental setup and results, respectively. Finally, Section 6 concludes the paper with insights and future directions.

2. KNN-VC FOR ELVC

KNN-VC was originally proposed for any-to-any speaker VC [10]. Here, we describe its adaptation to ELVC in a one-to-one VC setting, with the overall framework shown in Fig. 1. In the offline stage, the NL speech data is processed to build a reference corpus. The WavLM-Large model [15] is used to extract SSL feature sequences from NL speech, which are then compiled into an NL feature dictionary (referred to as a matching set in [10]). A HiFi-GAN vocoder [16] is trained using these NL speech features to facilitate high-quality waveform synthesis.

In the online stage, feature extraction, conversion, and waveform synthesis are performed in sequence.

1. **Feature Extraction**: The WavLM-Large model extracts an SSL feature sequence $\{X_t\}_{t=1}^{T}$ from the input EL speech utterance, where T denotes the number of speech frames, and X_t represents the EL feature vector at frame t.
2. **Feature Conversion**: Each EL feature vector X_t is converted independently on a frame-by-frame basis. For each X_t, the model retrieves its k nearest neighbors from the NL dictionary (yielding Sub-NL Dictionary in Fig. 1) based on cosine distance. The k nearest neighbors are uniformly averaged to obtain the converted feature vector Y_t. The process produces the converted feature vector sequence $\{Y_t\}_{t=1}^{T}$.
3. **Waveform Synthesis**: The converted SSL feature vector sequence is fed into the HiFi-GAN vocoder to synthesize the final speech waveform.

To address potential discontinuities in the feature sequence resulting from frame-independent conversion, we incorporate the dynamic features proposed in [14]. Delta and delta-delta features are calculated from static features during feature extraction. The Maximum Likelihood Parameter Generation (MLPG) algorithm [17] is then applied to ensure temporal coherence. The ELVC model that directly applies KNN-VC is called **KNN-ELVC**.

3. PROPOSED METHODS

Fig. 2 shows the architecture of the proposed parallel ELVC systems in a one-to-one VC setting. In the offline stage, a parallel corpus of EL and NL speech is prepared in advance. The WavLM-Large model is then used to extract SSL feature vector sequences from both EL and NL speech. Next, Dynamic Time Warping (DTW) aligns feature sequences of EL speech and NL speech. Finally, a pairing dictionary is built based on the aligned EL and NL features, allowing the model to directly map EL features to their corresponding NL counterparts.

In the online stage, feature extraction and waveform synthesis follow the same process as the KNN-VC method. However, the feature conversion process differs in the following ways:

1. **Neighbor Selection**: For each EL feature vector X_t, the model identifies its k nearest neighbors from the EL dictionary (forming Sub-EL Dictionary in Fig.2) based on cosine distance.
2. **Weight Computation**: A weight vector W_t is derived based on these k nearest neighbors according to

$$W_t = \arg\min_{W} ||X_t - \sum_{i=1}^{k} w^i \times b^i||^2, \qquad (1)$$

   where b^i is the i-th neighbor of X_t, and w^i is the weight.
3. **Feature Conversion**: The converted feature vector Y_t is obtained by linearly combining the k NL feature vectors corresponding the k nearest EL neighbors (Sub-NL Dictionary in Fig.2) using the weight vector W_t.

The weight calculation in step 2 differentiates between KNN-based and LLE-based methods. The proposed KNN-based parallel ELVC (**pKNN-ELVC**) employs uniform weighting similar to KNN-VC, while the proposed LLE-based parallel ELVC (**LLE-ELVC**) computes weights to best reconstruct X_t from its k nearest neighbors according to Eq. (1). The dynamic features with MLPG introduced in Section 2 can also be applied to both models.

In contrast to **KNN-ELVC** (see Section 2), in **pKNN-ELVC** and **LLE-ELVC**, the **Neighbor Selection** step is performed entirely within the EL speech domain. This approach can mitigate potential performance degradation caused by the different acoustic characteristics of EL and NL speech. But the prerequisite for using this approach is that the target speaker must provide parallel EL and NL speech.

4. EXPERIMENTAL SETUP

4.1. Experimental dataset

Our experimental dataset was recorded by a physician reading the Taiwanese Mandarin Hearing in Noise Test (TMHINT) phonetically balanced sentences [18] with and without an electrolarynx. There were 320 EL and 320 NL speech utterances. The speech utterances corresponding the first 240 sentences were used to build the sample dictionary for exemplar-based methods or train the DNN model, while the speech utterances corresponding to the remaining 80 sentences were used as the test set. All speech signals were recorded at a sampling rate of 16 kHz.

4.2. Reference systems

To evaluate the effectiveness of exemplar-based ELVC, we compared six exemplar-based ELVC systems with one DNN-based ELVC system. These systems include: **KNN-ELVC-s** and **KNN-ELVC-d**, which use the KNN-VC method with static and dynamic features for feature conversion, respectively (Section 2); **pKNN-ELVC-s** and **pKNN-ELVC-d**, which represent the proposed KNN-based parallel ELVC methods, respectively using static and dynamic features for feature conversion (Section 3); **LLE-ELVC-s** and **LLE-ELVC-d**, which present the proposed LLE-based parallel ELVC methods with static and dynamic features (Section 3). The DNN-based ELVC system used the **CDVAE** model in [9].

4.3. Settings for exemplar-based systems

For exemplar-based systems, we followed the settings in [10] for SSL feature extraction and waveform synthesis. Specifically, we used the output of layer 6 of the pretrained WavLM-Large model [15] as SSL features, and each 20 ms of 16 kHz audio was represented as a 1024-dimensional vector (static feature). As reported in [10], these SSL features are crucial for achieving good performance in speaker VC tasks. Dynamic features (delta and delta-delta) were calculated for systems using dynamic features, resulting in a 3072-dimensional feature vector. For waveform synthesis, we employed a pretrained HiFi-GAN vocoder[1], which takes static SSL features as input. The vocoder was trained on the English multi-speaker LibriSpeech train-clean100 dataset [19]. The prematched technique was used during vocoder training, i.e., the KNN-VC method was used to generate vocoder training data to ensure consistency between feature conversion and vocoder generation. We chose this HiFi-GAN version because our preliminary experiments showed that it provided better sound quality than the regular HiFi-GAN vocoder (without the prematched technique). For more details on the prematched HiFi-GAN vocoder, see [10].

4.4. Settings for the DNN-based baseline system

In [9], the CDVAE-based ELVC system utilized both traditional spectral features and WaveLM features for pretraining. Then, DTW was used to align the EL and NL speech signals

[1] https://github.com/bshall/knn-vc

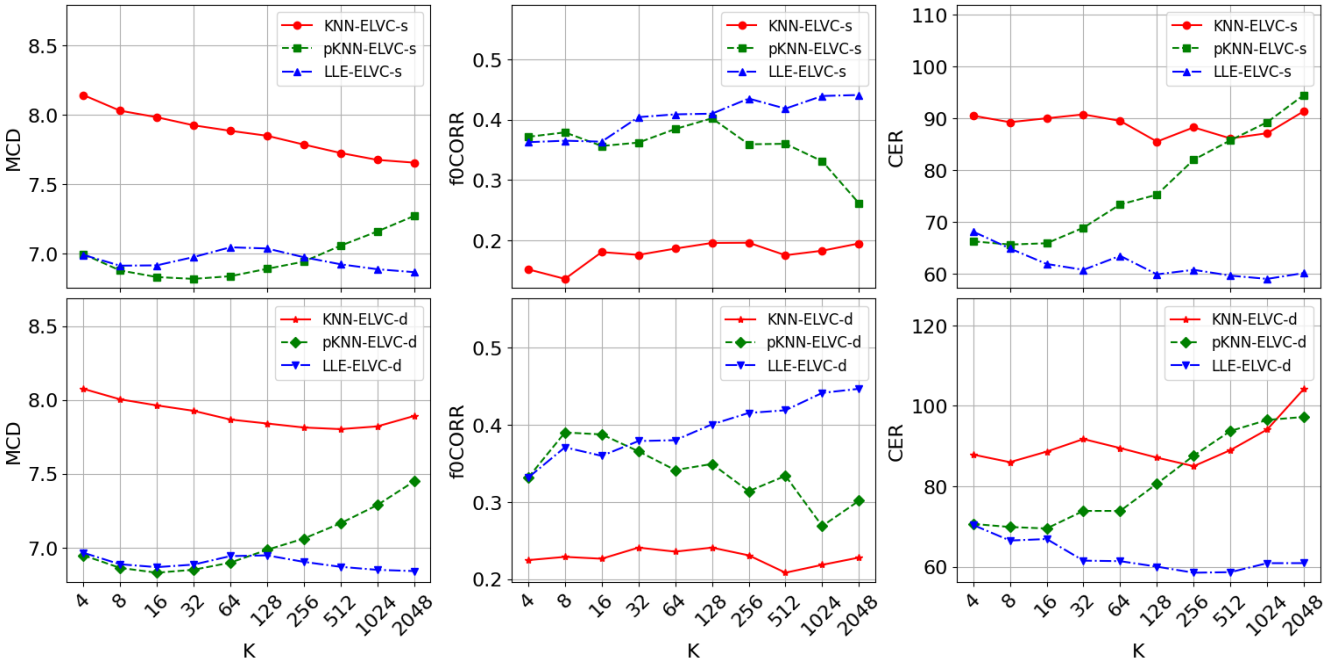


Fig. 3. Performance trends of exemplar-based ELVC methods with increasing k value.

based on the WaveLM features, and the model was trained accordingly to convert the EL speech into the NL speech. A Parallel WaveGAN vocoder [20] trained with the WaveLM features was used for waveform synthesis. For a fair comparison, in our experiments, the SSL features and vocoder used in **CDVAE** are the same as those used in the exemplar-based methods.

4.5. Experimental protocols

We used four objective metrics to evaluate the converted speech obtained by each system. Mel-cepstral distortion (MCD) in dB measures spectral distortion. When calculating MCD, the 0th order of the mel-cepstral coefficients was excluded. Fundamental frequency root mean square error (f0RMSE) in Hz measures the accuracy of the F0 information. Fundamental frequency correlation coefficient (f0CORR) measures F0 feature correlation. Character Error Rate (CER) was obtained using the pretrained Whisper large model [21]. In addition, a subjective listening test was conducted using the mean opinion score (MOS) test. Participants, all native Chinese speakers with normal hearing, rated the naturalness and intelligibility of the converted speech on a five-point scale, with 1 being the worst and 5 being the best. For intelligibility, the scoring scale is as follows: 5 means complete understanding of every word, 4 means difficulty with a few words but generally understandable, 3 means understanding about half of the words, 2 means understanding only a few words, and 1 means no understanding at all. Twenty speech utterances from each ELVC system were selected for the listening test. A total of 25 subjects participated. For each test sample, participants were unaware of which ELVC system generated it.

5. EXPERIMENTAL RESULTS

5.1. Selection of the best exemplar-based system

This section aims to identify the best exemplar-based systems under different settings. Given the critical importance of intelligibility in ELVC tasks, our evaluation prioritizes CER as the main metric, while also considering other metrics such as MCD and f0CORR to ensure a comprehensive quality assessment. Fig. 3 shows the performance trends of these methods at different values of k (i.e., the number of neighbors in the k-nearest neighbors method).

The main observations from Fig. 3 are as follows. First, the KNN-ELVC, pKNN-ELVC, and LLE-ELVC methods achieved the lowest CER in the following settings: **KNN-ELVC-d** with $k = 256$, **pKNN-ELVC-s** with $k = 8$, and **LLE-ELVC-d** with $k = 256$. Systems under these configurations are also competitive in other metrics, with **LLE-ELVC-d** not only achieving the lowest CER but also maintaining high scores in other metrics, highlighting its overall effectiveness. Second, the proposed parallel ELVC systems consistently outperform KNN-ELVC on most metrics across k values, emphasizing their suitability for ELVC tasks. Among the proposed methods, **pKNN-ELVC** exhibits performance degradation with increasing k, whereas **LLE-ELVC** improves or maintains stable performance in f0CORR and CER with increasing k while maintaining relatively consistent MCD values.

Table 1. Comparison of different ELVC systems. Subjective test (MOS with 95% confidence interval).

		Objective Test				Subjective Test (MOS: 1-5)	
Type	System	MCD (dB)↓	f0RMSE↓	f0CORR↑	CER (%)↓	Naturalness↑	Intelligibility↑
EL	-	10.80	27.74	−0.021	73.68	1.21 ± 0.04	1.91 ± 0.09
Exemplar-based	**KNN-ELVC-d**	7.81	24.09	0.231	86.15	2.24 ± 0.07	2.04 ± 0.09
	pKNN-ELVC-s	**6.88**	24.30	0.379	64.00	2.97 ± 0.08	$\mathbf{2.96 \pm 0.10}$
	LLE-ELVC-d	6.90	**23.91**	0.416	**56.22**	2.99 ± 0.07	2.93 ± 0.10
DNN-based	**CDVAE**	7.07	24.02	**0.481**	64.96	$\mathbf{3.12 \pm 0.09}$	2.61 ± 0.10
NL	-	-	-	-	4.9	4.95 ± 0.03	4.95 ± 0.03

5.2. Overall results

Table 1 shows a comprehensive comparison of representative exemplar-based systems selected in Section 5.1), the **CDVAE** baseline system, unprocessed EL speech, and NL speech. In objective tests, **LLE-ELVC-d** shows superior performance in f0RMSE (23.91 Hz) and CER (56.22%), while achieving a comparable MCD (6.90 dB) to **pKNN-ELVC-s** (6.88 dB). The **CDVAE** baseline outperforms all exemplar-based systems in terms of f0CORR (0.481). In subjective tests, it is interesting to note that although **CDVAE** is worse than **pKNN-ELVC-s** and **LLE-ELVC-d** in objective metrics CER and MCD, it achieves the highest naturalness score (3.12 ± 0.09), significantly better than other methods ($p < 0.05$, t-test). **LLE-ELVC-d** and **pKNN-ELVC-s** follow closely in terms of naturalness (2.99 ± 0.07 and 2.97 ± 0.08, respectively), with no statistically significant difference between them ($p = 0.6581$). In terms of intelligibility, **pKNN-ELVC-s** scores 2.96 ± 0.10, which is better than other methods, but the difference from **LLE-ELVC-d** (2.93 ± 0.10) is not statistically significant ($p = 0.7202$). Although **CDVAE**'s naturalness score is high, its intelligibility score is low (2.61 ± 0.10). All ELVC systems demonstrate significant improvements over unprocessed EL speech on all metrics. However, there is still a large gap between converted speech and natural speech, especially in subjective evaluations of naturalness and intelligibility.

In summary, our comprehensive analysis reveals that while exemplar-based systems, particularly **LLE-ELVC-d**, provide robust performance on a variety of objective and subjective metrics, **CDVAE** maintains an advantage in terms of perceived naturalness. However, considering all metrics, **LLE-ELVC-d** provides the best trade-off between naturalness, intelligibility, and objective accuracy. This demonstrates the potential of **LLE-ELVC** for practical applications in ELVC tasks and highlights future research directions to further bridge the gap between converted speech and natural speech. Furthermore, these findings highlight how our proposed parallel methods effectively address the limitations observed when directly applying KNN-VC to ELVC. The key improvement lies in resolving the mismatch between EL speech and NL speech, a potential issue with the KNN-VC baseline due to the assumption of similar SSL feature distributions. By performing neighbor selection within the same EL speech domain, the proposed methods avoid performance degradation caused by the different acoustic characteristics of EL speech and NL speech. This leads to more effective conversion results, as evidenced by the improved subjective and objective metrics, especially for **LLE-ELVC**. Audio samples are publicly available[2].

6. CONCLUSIONS

In this paper, we study exemplar-based methods for ELVC and propose two parallel methods, namely parallel KNN-ELVC and parallel LLE-ELVC. Our findings show that both methods effectively address the mismatch problem between EL speech features and NL speech features, which is a key limitation in applying KNN-VC to ELVC. Among these methods, LLE-ELVC is the most balanced, performing well in terms of intelligibility and objective metrics while also delivering competitive naturalness scores. Although the CDVAE baseline system leads in terms of perceived naturalness, LLE-ELVC exhibits robust overall performance, making it a strong candidate for practical ELVC applications. Future work will focus on refining the components of these systems using Mandarin datasets, exploring prematched techniques to further improve exemplar-based methods, and investigating more flexible any-to-any ELVC methods to enhance patient usability.

7. ACKNOWLEDGEMENTS

The authors would like to thank Professor Wen-Chin Huang of Nagoya University, Japan, for valuable discussions and suggestions that greatly enhanced this work.

[2]https://xam219.github.io/exemplar-elvc/

8. REFERENCES

[1] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech communication*, vol. 54, no. 1, pp. 134–146, 2012.

[2] P Malathi, GR Sureshw, and M Moorthi, "Enhancement of electrolaryngeal speech using frequency auditory masking and GMM based voice conversion," in *Proceedings of AEEICB*. IEEE, 2018, pp. 1–4.

[3] Kazuhiro Kobayashi and Tomoki Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," in *Proceedings of EUSIPCO*. IEEE, 2018, pp. 2115–2119.

[4] Zhaopeng Qian, Kejing Xiao, and Chongchong Yu, "Mandarin electro-laryngeal speech enhancement using cycle-consistent generative adversarial networks," *Applied Sciences*, vol. 13, no. 1, pp. 537, 2022.

[5] Yaogen Yang, Haozhe Zhang, Zexin Cai, Yao Shi, Ming Li, Dong Zhang, Xiaojun Ding, Jianhua Deng, and Jie Wang, "Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion," *Biomedical Signal Processing and Control*, vol. 80, pp. 104279, 2023.

[6] Ming-Chi Yen, Wen-Chin Huang, Kazuhiro Kobayashi, Yu-Huai Peng, Shu-Wei Tsai, Yu Tsao, Tomoki Toda, Jyh-Shing Roger Jang, and Hsin-Min Wang, "Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling," in *Proceedings of ASRU*. IEEE, 2021, pp. 650–657.

[7] Lester Phillip Violeta, Wen-Chin Huang, Ding Ma, Ryuichi Yamamoto, Kazuhiro Kobayashi, and Tomoki Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders," in *Proceedings ICASSP*. IEEE, 2024, pp. 10961–10965.

[8] Yung-Lun Chien, Hsin-Hao Chen, Ming-Chi Yen, Shu-Wei Tsai, Hsin-Min Wang, Yu Tsao, and Tai-Shih Chi, "Audio-visual Mandarin electrolaryngeal speech voice conversion," in *Proceedings of INTERSPEECH*, 2023, pp. 5023–5026.

[9] Hsin-Hao Chen, Yung-Lun Chien, Ming-Chi Yen, Shu-Wei Tsai, Tai-Shih Chi, Hsin-Min Wang, and Yu Tsao, "Mandarin electrolaryngeal speech voice conversion using cross-domain features," in *Proceedings of INTERSPEECH*, 2023, pp. 5018–5022.

[10] Matthew Baas, Benjamin van Niekerk, and Herman Kamper, "Voice conversion with just nearest neighbors," in *Proceedings of INTERSPEECH*, 2023, pp. 2053–2057.

[11] Binzhu Sha, Xu Li, Zhiyong Wu, Ying Shan, and Helen Meng, "Neural concatenative singing voice conversion: rethinking concatenation-based approach for one-shot singing voice conversion," in *Proceedings of ICASSP*. IEEE, 2024, pp. 12577–12581.

[12] Dong Liu, Yueqian Lin, Hui Bu, and Ming Li, "Two-stage and self-supervised voice conversion for zero-shot dysarthric speech reconstruction," in *Proceedings of IALP*, 2024, p. 428–432.

[13] Pravin Mote, Berrak Sisman, and Carlos Busso, "Unsupervised domain adaptation for speech emotion recognition using k-nearest neighbors voice conversion," in *Proceedings of INTERSPEECH*, 2024.

[14] Yi-Chiao Wu, Hsin-Te Hwang, Chin-Cheng Hsu, Yu Tsao, Hsin-Min Wang, et al., "Locally linear embedding for exemplar-based spectral conversion.," in *Proceedings of INTERSPEECH*, 2016, pp. 1652–1656.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of NeurIPS*, 2020, vol. 33, pp. 17022–17033.

[17] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[18] M Huang, "Development of taiwan mandarin hearing in noise test," *Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of ICASSP*. IEEE, 2015, pp. 5206–5210.

[20] Ryuichi Yamamoto, Eunwoo Song, and Jae Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proceedings of ICASSP*, 2020.

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of ICML*. PMLR, 2023, pp. 28492–28518.