A Preclinical Study of Electrolaryngeal Voice Conversion for a Novel Nasal Electrolarynx: Feature Choice and Data Augmentation

Qi-Yan Chen[1,*], *Ming-Chi Yen*[1,2,*], *Fo-Rui Li*[3], *Hsin-Te Hwang*[1], *Ching-Hung Lai*[4], *Shu-Wei Tsai*[4], *Ping-Cheng Yeh*[2], *Jyh-Shing Roger Jang*[2], *Yu Tsao*[1], *Hsin-Min Wang*[1]

[1] Academia Sinica, Taipei, Taiwan
[2] National Taiwan University, Taipei, Taiwan
[3] National Central University, Taoyuan, Taiwan
[4] National Cheng Kung University Hospital, Tainan, Taiwan

ymchiqq@iis.sinica.edu.tw

Abstract

Electrolaryngeal (EL) speech often exhibits reduced intelligibility and naturalness due to imperfect excitation and fixed pitch. In this preclinical study, we present the first systematic investigation of electrolaryngeal voice conversion (ELVC) for speech produced by a newly invented nasal electrolarynx (NEL), extending previous work that mainly targets cervical EL (CEL) speech. We examine the choice of features under NEL's distinctive acoustics by comparing Mel-spectrograms and WavLM features as input to a seq2seq ELVC system. To address NEL data scarcity, we propose an LLE-VC-based augmentation strategy to synthesize paired sNEL–sNL speech for VC pretraining. Experiments reveal device-dependent trends: WavLM features benefit CEL-oriented conversion and augmentation, whereas Mel-spectrogram inputs better preserve NEL-specific spectral characteristics and yield better intelligibility-related metrics and listener preferences for NEL-to-NL conversion.

Index Terms: electrolaryngeal speech, voice conversion, sequence-to-sequence, pretraining, locally linear embedding

1. Introduction

Human speech production follows a source-filter mechanism: compressed airflow from the lungs passes through the vocal folds to form an excitation signal, which is then filtered by the articulators. In this paper, the speech produced by healthy people is termed natural speech (NL speech). Patients who have undergone laryngectomy lose the ability to generate excitation via vocal-fold vibration. The most commonly used assistive device is the cervical electrolarynx (CEL) (see Fig. 1, left), which generates a constant-frequency mechanical excitation for the vocal tract to modulate. However, CEL speech suffers from (1) loud mechanical noise and (2) a lack of pitch variation, which limits naturalness and intelligibility. Traditional CEL devices require the vibrator to be pressed against the neck, which may not be suitable for patients with surgical wounds or restricted neck movement. A nasal electrolarynx (NEL) has recently emerged that places the excitation source at the entrance of the nasal cavity and uses the nasopharynx as a coupling point (see Fig. 1, right) [1]. NEL provides (1) improved coupling stability due to a fixed intra-nasal excitation location and (2) reduced mechanical noise because the excitation is coupled internally rather than transmitted through neck skin and muscles, while also allowing hands-free speech. Nevertheless, NEL speech exhibits distinct acoustic characteristics compared to CEL and NL speech, including mid–high-frequency resonances, low-frequency attenuation, and elevated vowel formants due to nasal cavity geometry, leaving open the question of whether ELVC systems optimized for CEL speech can be effectively applied to NEL speech.

Most prior ELVC systems relied on parallel EL–NL corpora with aligned sentences [2, 3], using feature extraction, alignment (often dynamic time warping [4]), and mapping followed by vocoder synthesis. Gaussian mixture models [5] improved naturalness while preserving speaker identity, while later deep learning methods [6] further enhanced quality. However, frame-level alignment errors still limited CEL-to-NL mapping. Seq2seq models offered a solution by jointly learning alignment and mapping through an attention mechanism [7]. Pretraining on large NL corpora alleviated the data scarcity issue [8], and subsequent seq2seq ELVC systems achieved significant improvements in CEL speech intelligibility [9]. More recently, advanced strategies have incorporated speech encoder loss and duration adjustment for patient ELVC [10]. Self-supervised learning (SSL) and automatic speech recognition (ASR) techniques have also been applied to ELVC. For example, in [11], ASR is used for temporal alignment, and SSL features are used for mapping. K-nearest neighbors VC (KNN-VC) [12] and locally linear embedding ELVC (LLE-ELVC) [13, 14] also use SSL features. Additionally, fine-tuning a pre-trained VC model for ELVC still requires a certain amount of EL–NL paired data [15].

Based on these insights, this study aims to develop an ELVC system to improve the quality and intelligibility of NEL speech. Due to limited NEL data, we leveraged LLE-VC [14] for data augmentation and investigated the effects of different input features (Mel-spectrograms and features from layer 6 of the WavLM-Large model [16]) within a seq2seq ELVC system [9]. The contributions of this study are as follows:

- We present the first study of ELVC for a newly invented NEL and conduct controlled comparisons with CEL within a unified framework.
- We show that feature choice is device-dependent: WavLM features are more effective for CEL-to-NL ELVC, whereas Mel-spectrograms are more effective for NEL-to-NL ELVC.
- We propose an LLE-VC-based NEL-oriented data augmentation strategy that combines WavLM features and Mel-spectrograms to better preserve NEL-specific acoustic features in synthesized NEL speech.

[*] These authors contributed equally.

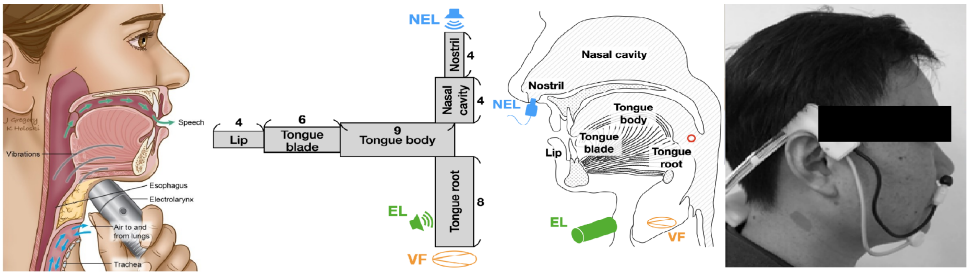


Figure 1: *Overview of CEL and NEL speech production. The leftmost panel shows the use of a traditional cervical EL (CEL) device [17], the rightmost panel shows the use of a nasal EL (NEL) device, and the two middle panels illustrate the excitation pathways of the two EL devices [1].*

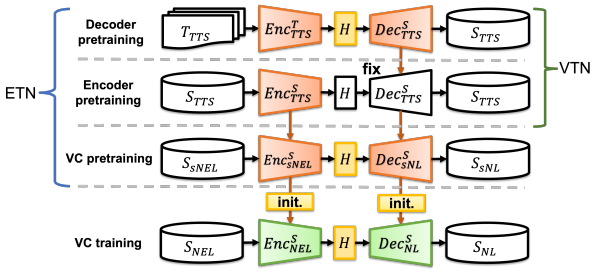


Figure 2: *The training process of VTN-VC and ETN-VC models.*

2. Methodology

We build ELVC models based on the transformer-based seq2seq VC model in [8]. The baseline model follows the VTN-VC model in [9]. The advanced model follows the ETN-VC model in [10].

2.1. VTN-VC

As shown in Fig. 2, VTN-VC is trained in three stages: two pretraining stages followed by VC training. Here, $D_{TTS} = \{T_{TTS}, S_{TTS}\}$ denotes the TTS dataset, where T and S indicate the data types of text and speech, respectively; the superscript of Enc/Dec denotes the input/output data type, and H denotes the hidden representations of the encoder. First, a TTS model is trained on D_{TTS} to obtain a decoder Dec^S_{TTS} with a strong speech generation capability. Second, the input is switched to the audio features of S_{TTS} while the decoder Dec^S_{TTS} is fixed, so that the encoder Enc^S_{TTS} learns hidden representations H that are compatible with Dec^S_{TTS}. Third, the encoder Enc^S_{NEL} and the decoder Dec^S_{NL} are trained on an ELVC training set $D_{ELVC} = \{S_{NEL}, S_{NL}\}$ to map the S_{NEL} features to the S_{NL} features, using the good initialization provided by the large-scale NL pretraining.

2.2. ETN-VC

As shown in Fig. 2, ETN-VC extends the advantages of the VTN-VC framework by performing an additional VC pretraining stage with an external simulated sNEL–sNL paired dataset after the two pretraining stages [10]. In this study, we construct a synthetic paired dataset $D_{syn} = \{S_{sNEL}, S_{sNL}\}$ using the data augmentation procedure in Section 2.3. Following the strategies in [10] and [18], we combine D_{syn} with the paired NEL–NL dataset D_{ELVC} for third-stage VC pretraining, and then use D_{ELVC} alone for final-stage VC training.

2.3. Data Augmentation

We used a pretrained F5-TTS model [19] to generate synthetic NL (sNL) speech from the TWnews corpus (containing 10,000 phonetically balanced Mandarin Chinese sentences) [20]. The first NL training utterance was used as the reference speech for F5-TTS. We also tried Breezy Voice [21], which is based on the CosyVoice architecture [22] and trained on a Mandarin Chinese dataset. However, for our target text, the generated Mandarin speech had excessively long pauses in utterances. Therefore, we ultimately chose F5-TTS.

For synthetic NEL (sNEL) speech, we followed the LLE-VC [13] process to convert sNL speech into paired sNEL speech. The NEL–NL paired training utterances were used to construct the dictionary for LLE-VC. We used features from layer 6 of the WavLM-Large model [16] for the neighbor search in LLE-VC, since kNN-VC identifies this layer as the best-performing for NL speech VC [12], and LLE-ELVC has also proven its effectiveness for EL speech VC [13]. We then used these neighboring WavLM features to reconstruct the converted WavLM features, or used the Mel-spectrograms corresponding to these neighboring WavLM features to reconstruct the converted Mel-spectrograms. Compared to TTS-based data augmentation [18], our method does not require training a dedicated EL TTS model; and unlike [10], which augments its limited set of 240 utterances by altering speech duration, our method can generate diverse paired sNEL–sNL pairing data from the 10,000-sentence TWnews corpus while preserving device-specific acoustic characteristics.

3. Experimental Setup

3.1. Datasets

The dataset used in this study contains five types of Mandarin speech (16 kHz): NL, CEL, NEL, sNL, and sNEL. This study is positioned as a preclinical stage and uses ELVC to examine differences between NEL speech and CEL speech under controlled conditions. NL, CEL, and NEL utterances were recorded by a single healthy speaker reading 320 TMHINT sentences (10 characters each) [23], reducing variability from device usage and speaker differences and serving as a reference for future studies on patients. The NL and CEL speech was recorded in a studio, while the NEL speech was recorded in a hospital ward environment to reflect anticipated clinical recordings and practical use of NEL. Although ward recordings may introduce environment-related factors (e.g., background noise and room acoustics), we control major confounding factors by keeping the speaker, sentence set, and parallel NL–CEL and NL–NEL pairing fixed, so that the observed trends primarily reflect device-

and representation-related differences under a realistic recording condition. Importantly, patient data are expected to be collected in the same ward setting for direct comparison. Of these, 240/40/40 paired CEL–NL (or NEL–NL) utterances were used for training/development/evaluation. The first two pretraining stages used 45k utterances (44 hours) from the COSPRO dataset [24], and VC pretraining used 10k sNEL–sNL utterance pairs (see Section 2.3). The MATBN dataset (196 hours) [25] was used to train the syllable recognizer for evaluation.

3.2. Training Settings

Both the VTN-VC and ETN-VC models were implemented using the seq2seq-VC framework [8]. Each model was trained with either 80-dim Mel-spectrograms (16 kHz) or 1024-dim WavLM features (the 6th layer of WavLM-Large). Data augmentation was applied only to ETN-VC: we first performed VC pretraining using sNEL–sNL and NEL–NL utterance pairs, and then fine-tuned the model using only NEL–NL utterance pairs. For generating synthetic NEL data, we followed the default LLE-VC configuration [14]. For waveform reconstruction, we used the ParallelWaveGAN toolkit [26] to train HiFi-GAN vocoders [27] on the COSPRO dataset, separately for each feature type. The vocoders trained for the two feature types achieved very similar objective performance on the test set, and thus are unlikely to introduce evaluation bias.

3.3. Evaluation Metrics

We used both objective and subjective metrics for the experimental evaluation. Objective metrics included mel-cepstral distortion (MCD), fundamental-frequency root mean squared error (F0 RMSE), fundamental-frequency correlation (F0 CORR), average absolute duration difference (DDUR), character error rate (CER), and syllable error rate (SER) to assess speech quality, prosody, and intelligibility. CER was calculated using the Whisper-Large model [28], and SER was calculated using our Mandarin syllable recognizer. For subjective evaluation, we conducted a randomized A/B test to compare the intelligibility of converted utterances from two ELVC models. We also report three off-the-shelf neural assessment scores: Speech-BERTScore [29] for semantic fidelity, and UTMOS [30] and MOSA-Net+ [31] for speech quality and intelligibility.

4. Experimental Results

4.1. NEL-to-NL Conversion vs CEL-to-NL Conversion

As shown in Table 1, although the NEL speech yields worse metric values and higher ASR errors (CER/SER) than the CEL speech, the overall conversion difficulty is comparable. We observe opposite feature preferences across devices: Mel-spectrograms work better for NEL-to-NL, whereas WavLM features are more effective for CEL-to-NL. We attribute this discrepancy to the acoustic uniqueness of NEL speech. As shown in Fig. 3, NEL speech exhibits distinct mid-high-frequency resonance and low-frequency attenuation, and WavLM features, pretrained largely on NL speech, may under-represent these NEL-specific spectral cues. This uniqueness may also explain the consistently poorer metric values of NEL than CEL. We also report the conversion performance of several zero-/few-shot VC systems on CEL and NEL speech. The SER results are as follows: SeedVC [32], 71.8% and 84.8%; MKL-VC [33], 85.0% and 99.5%; and Vevo [34], 84.0% and 101.5%. These results suggest that general-purpose zero-shot VC is insufficient for CEL/NEL speech, motivating the need to tailor and adapt VC models to their unique acoustic characteristics.

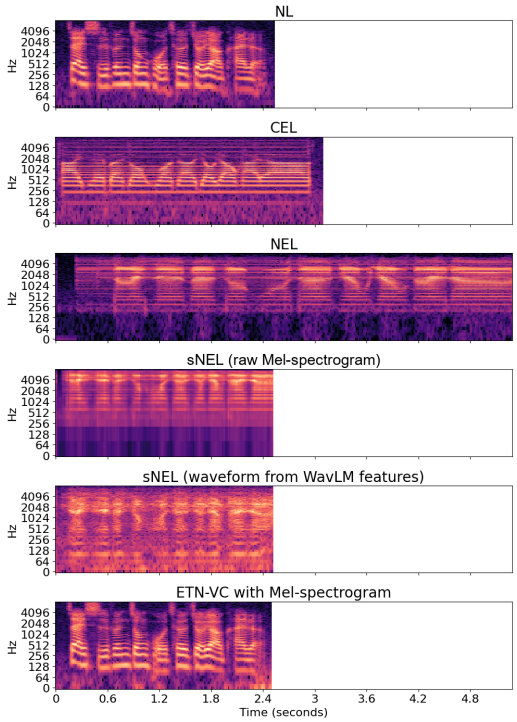


Figure 3: *Mel-spectrograms of NL, CEL, NEL, sNEL, and ETN-VC-converted speech. NEL speech is longer, while sNEL and converted speech follow the NL duration due to frame-wise conversion and seq2seq alignment, respectively.*

4.2. Comparison of VTN-VC and ETN-VC

Table 2 shows that ETN-VC (with data augmentation) outperforms VTN-VC, regardless of whether Mel-spectrograms or WavLM features are used, and the performance improves with increasing data augmentation. Just as VTN-VC with Mel-spectrograms outperforms VTN-VC with WavLM features, ETN-VC with Mel-spectrograms also outperforms ETN-VC with WavLM features. Note that when training the ETN-VC model with Mel-spectrograms, the Mel-spectrograms synthesized by LLE-VC are directly used for data augmentation. Compared to the VTN-VC model with Mel-spectrograms, the best ETN-VC model with Mel-spectrograms reduces the CER from 72.0% (see Table 1) to 63.0% and the SER from 58.8% (see Table 1) to 53.8%. As shown in the lower part of Fig. 3, the speech converted by ETN-VC exhibits a Mel-spectrogram that is similar to that of the NL speech.

4.3. Spectrogram Analysis

Based on the previous results and our Mel-spectrogram analysis of NEL speech, we explored different strategies for generating sNEL speech for data augmentation. The first approach uses WavLM features in LLE-VC for both neighborhood search and feature reconstruction. The second approach still uses WavLM features for neighborhood search, but reconstructs the corresponding Mel-spectrogram features. For spectrogram analysis, the reconstructed WavLM features of the first approach must be vocoded to waveforms and then converted to a Mel-

Table 1: *Performance comparison of VTN-VC models on CEL/NEL-to-NL tasks. Metrics of unprocessed raw speech are included for reference. (↑ indicates higher is better, ↓ indicates lower is better.)*

Model	Feature	EL Type	MCD ↓	f0RMSE ↓	f0CORR ↑	DDUR ↓	CER (%) ↓	SER (%) ↓
✗	✗	CEL	10.79	34.10	-0.08	0.88	75.3 ± 6.15	69.3 ± 5.2
		NEL	13.75	48.81	0.02	2.68	90.8 ± 2.41	89.5 ± 6.3
VTN-VC	Mel-spectrogram	CEL	5.88	34.46	0.27	**0.11**	73.3 ± 6.33	64.5 ± 5.6
		NEL	**5.66**	34.62	0.25	0.25	72.0 ± 6.64	58.8 ± 6.4
	WavLM	CEL	5.97	33.57	0.30	0.20	**69.0 ± 6.65**	**57.5 ± 6.3**
		NEL	6.12	**33.32**	**0.31**	0.29	71.5 ± 7.68	62.0 ± 5.6

Table 2: *ETN-VC trained on the TMHINT training set with 1k/5k/10k TWNews utterances.*

Model	Feature	VC Pretraining Data		MCD ↓	f0RMSE ↓	f0CORR ↑	DDUR ↓	CER% ↓	SER% ↓
		TWnews	TMHINT						
ETN-VC	Mel-spectrogram	1K	✓	5.64	33.97	0.27	**0.14**	66.5 ± 7.72	58.8 ± 5.1
		5K	✓	5.63	**33.23**	0.30	0.16	66.3 ± 7.62	59.3 ± 5.4
		10K	✓	**5.56**	34.11	0.28	**0.14**	**63.0 ± 7.37**	**53.8 ± 4.8**
	WavLM	1K	✓	6.01	33.39	**0.31**	0.27	71.3 ± 7.01	61.8 ± 5.6
		5K	✓	6.07	34.18	0.29	0.28	68.8 ± 6.75	58.0 ± 5.8
		10K	✓	6.09	34.00	0.29	0.27	65.8 ± 7.39	54.3 ± 5.6

Table 3: *SpeechBERTScore, UTMOS, and MOSA-Net+ scores for NEL, NL, and converted speech (mean ± 95% confidence interval).*

Model	SpeechBERT Score ↑	UTMOS ↑	MOSA-Net+	
			Quality ↑	Intelligibility ↑
NEL Speech	0.523 ± 0.006	1.299 ± 0.005	2.153 ± 0.062	0.642 ± 0.022
VTN-VC Mel	0.674 ± 0.008	2.469 ± 0.071	3.511 ± 0.076	0.969 ± 0.004
VTN-VC WavLM	**0.678 ± 0.008**	2.944 ± 0.092	3.783 ± 0.093	0.980 ± 0.004
ETN-VC Mel	0.675 ± 0.009	2.666 ± 0.084	3.584 ± 0.079	0.973 ± 0.004
ETN-VC WavLM	0.674 ± 0.007	**2.998 ± 0.094**	**3.788 ± 0.073**	**0.981 ± 0.003**
NL Speech	1.000 ± 0.000	3.261 ± 0.069	4.510 ± 0.076	0.996 ± 0.001

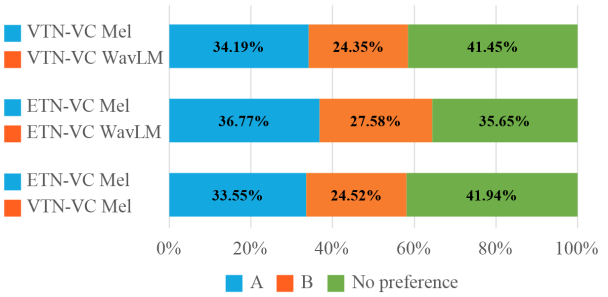


Figure 4: *A/B test results on intelligibility. The bars represent the percentage of subjects' voting for each system.*

spectrogram.

As shown in Fig. 3, Mel-spectrograms synthesized directly by LLE-VC preserve the mid-high-frequency resonance of NEL speech much better. In contrast, sNEL speech generated from reconstructed WavLM features fails to preserve this resonance and introduces low-frequency noise. The "Mel-spectrogram → waveform → Mel-spectrogram" route exhibits a similar issue, but to a lesser extent.

4.4. Subjective Evaluation

As shown in Fig. 4, we conducted three sets of A/B intelligibility tests with 30 participants (native Mandarin speakers, 18 males and 12 females, aged 20–50, self-reported normal hearing). The results show that Mel-spectrogram-based VTN-VC and ETN-VC both outperform their WavLM-based counterparts, further confirming the effectiveness of Mel-spectrograms for nasal ELVC. We also find that ETN-VC with Mel-spectrograms outperforms VTN-VC with Mel-spectrograms, confirming the benefit of data augmentation for nasal ELVC.

Table 3 shows that WavLM-based VC models achieve higher non-intrusive neural assessment scores (e.g., UTMOS and MOSA-Net+), suggesting smoother prosody and more natural signal flow. In contrast, Mel-spectrogram-based models perform better on intelligibility-related metrics (CER/SER) and in A/B tests, highlighting the trade-off between naturalness and intelligibility. SpeechBERTScore differences are small across systems, indicating good semantic preservation after conversion. Given the medical context of NEL speech, intelligibility is the focus of this study; therefore, we prioritize CER, SER, and human listening evaluation over non-intrusive neural quality metrics. Detailed results and audio samples are available on our demo website[1].

5. Conclusion

This paper presents a preclinical, systematic study of ELVC for a newly invented NEL device, extending previous ELVC work that mainly targets CEL speech. To mitigate NEL data scarcity, we propose an LLE-VC-based data augmentation strategy that synthesizes paired sNEL–sNL speech for VC pretraining, yielding consistent gains in NEL-to-NL conversion. Our experiments further show that feature choice is device-dependent: WavLM features benefit CEL-to-NL ELVC, whereas Mel-spectrogram inputs better preserve NEL-specific spectral characteristics and lead to improved intelligibility-related metrics and listener preferences for NEL-to-NL conversion. Because this preclinical study used the speech of a single healthy speaker, validation with real patient NEL speech remains an important next step. Future work will collect patient NEL data for formal perceptual evaluation, including naturalness MOS, and will examine broader SSL features, such as layer sweeps and alternative encoders.

[1]Demo site: `https://ymchiqq.github.io/nelvc_demo/`

6. Acknowledgments

This work was partially supported by the Taiwan National Science and Technology Council under Grant No. NSTC 114-2221-E-001-006-MY3.

7. Generative AI Use Disclosure

Generative AI tools were used solely for manuscript refinement; all research design, experiments, and analyzes were carried out independently by the authors.

8. References

[1] C.-H. Lai, S.-W. Tsai, C. Chiu, D. S.-Y. Hung, M. H.-C. Hsieh, I.-P. Lee, and T. Tsai, "The acoustic analysis of nasal electrolarynx," Society for Psychophysiological Research (SPR) Annual Meeting, 2024.

[2] 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.

[3] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *Proc. ICASSP*, 2011.

[4] 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.

[5] 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.

[6] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN," in *Proc. EUSIPCO*, 2021.

[7] H. Kameoka, K. Tanaka, D. Kwasńy, 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.

[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," in *Proc. INTERSPEECH*, 2020.

[9] 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 *Proc. ASRU*, 2021.

[10] M.-C. Yen, C.-H. Wu, S.-W. Tsai, J.-S. R. Jang, Y. Tsao, A. Hussain, and H.-M. Wang, "Mandarin electrolaryngeal speech voice conversion with speech encoder loss learning and seq2seq modeling," *IEEE Internet of Things Magazine*, vol. 8, no. 4, pp. 22–28, 2025.

[11] L. P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders," in *Proc. ICASSP*, 2024.

[12] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Proc. INTERSPEECH*, 2023.

[13] H.-T. Hwang, C.-H. Wu, M.-C. Yen, Y. Tsao, and H.-M. Wang, "Exemplar-based methods for Mandarin electrolaryngeal speech voice conversion," in *Proc. O-COCOSDA*, 2024.

[14] F.-R. Li, H.-T. Hwang, M.-C. Yen, M.-T. Lo, Y. Tsao, and H.-M. Wang, "Improving exemplar-based electrolaryngeal speech voice conversion via robust content representations," in *Proc. APSIPA ASC*, 2025.

[15] B. Mayrhofer, F. Pernkopf, P. Aichinger, and M. Hagmüller, "Lightweight and perceptually-guided voice conversion for electro-laryngeal speech," in *Proc. ICASSP*, 2026.

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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, Oct. 2022.

[17] IMU University, "Facts and information on laryngeal cancer," Nov. 2015, accessed: 2026-06-15. [Online]. Available: https://imu.edu.my/imunews/facts-information-laryngeal-cancer/

[18] 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," in *Proc. SLT*, 2022.

[19] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv:2410.06885*, 2024.

[20] Y.-W. Chen, H.-M. Wang, and Y. Tsao, "BASPRO: A balanced script producer for speech corpus collection based on the genetic algorithm," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, e15, 2023.

[21] C.-J. Hsu, Y.-C. Lin, C.-C. Lin, W.-C. Chen, H. L. Chung, C.-A. Li, Y.-C. Chen, C.-Y. Yu, M.-J. Lee, C.-C. Chen *et al.*, "BreezyVoice: Adapting TTS for Taiwanese Mandarin with enhanced polyphone disambiguation–challenges and insights," *arXiv:2501.17790*, 2025.

[22] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv:2407.05407*, 2024.

[23] 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.

[24] C.-Y. Tseng, Y.-C. Cheng, and C.-H. Chang, "Sinica COSPRO and Toolkit: Corpora and platform of Mandarin Chinese fluent speech," in *Proc. O-COCOSDA*, 2005.

[25] 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.

[26] T. Hayashi, "ParallelWaveGAN: PyTorch implementation of Parallel WaveGAN," 2019. [Online]. Available: https://github.com/kan-bayashi/ParallelWaveGAN

[27] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[29] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics," in *Proc. INTERSPEECH*, 2024.

[30] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022," in *Proc. INTERSPEECH*, 2022.

[31] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," in *Proc. ICME*, 2024.

[32] S. Liu, "Zero-shot voice conversion with diffusion transformers," *arXiv:2411.09943*, 2024.

[33] A. Lobashev, A. Yermekova, and M. Larchenko, "Training-free voice conversion with factorized optimal transport," in *Proc. ICASSP*, 2025.

[34] X. Zhang, X. Zhang, K. Peng, Z. Tang, V. Manohar, Y. Liu, J. Hwang, D. Li, Y. Wang, J. Chan, Y. Huang, Z. Wu, and M. Ma, "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement," in *Proc. ICLR*, 2025.