Personalized Electrolaryngeal Voice Conversion with a Single Pre-operative Utterance

Devin Chang[1], *Hsin-Te Hwang*[1], *Ming-Chi Yen*[1], *Shu-Wei Tsai*[2], *Yu Tsao*[1], *Hsin-Min Wang*[1]

[1] Academia Sinica, Taipei, Taiwan
[2] National Cheng Kung University Hospital, Tainan, Taiwan

devchang918@gmail.com

Abstract

Electrolaryngeal voice conversion (ELVC) enhances the mechanical speech of laryngectomees to improve naturalness and intelligibility. While personalized ELVC typically requires extensive pre-operative recordings for supervised training, this paper investigates a one-shot scenario where only one natural utterance is available for timbre preservation. We evaluate two strategies using zero-shot VC models: (1) a cascaded approach, performing ELVC followed by secondary conversion to the target timbre; and (2) a pseudo-target strategy, using a zero-shot model to generate synthetic training data for a supervised ELVC model. Results indicate that the cascaded approach outperforms the pseudo-target strategy, remarkably improving speaker similarity and quality with minimal impact on intelligibility. We also demonstrate that direct application of zero-shot models to ELVC fails due to domain mismatch, confirming the necessity of our proposed framework for one-shot personalized voice restoration.

Index Terms: Electrolaryngeal Speech, voice conversion, timbre preservation

1. Introduction

Patients with advanced laryngeal cancer often require total laryngectomy (TL), a surgical procedure that removes the larynx, resulting in the loss of natural voice and speech [1]. To restore patients' ability to communicate after TL, electrolarynxes have been developed, which generate mechanical speech by inducing vibrations in the neck skin [1]. However, the robotic quality and low naturalness of electrolaryngeal (EL) speech often lead to poor intelligibility and listener discomfort, creating communication barriers that can significantly affect patients' social lives. [2].

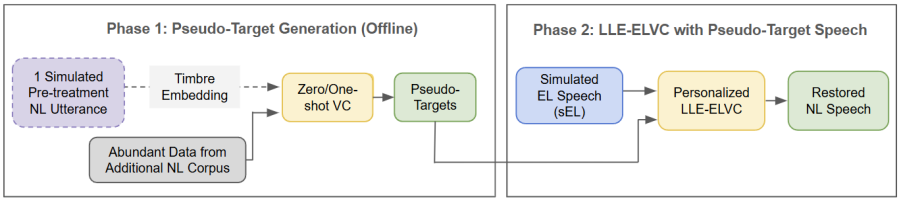
To improve communication for these patients, voice conversion (VC) techniques have been extensively studied to improve the intelligibility of EL speech. Traditional statistical methods such as Gaussian-Mixture Model [3, 4] and neural network–based approaches [3, 5, 6, 7, 8] have achieved success in improving EL speech intelligibility. Although these methods significantly improve intelligibility, the converted speech typically resembles the voice of a pre-defined target speaker rather than the patient's original voice. Consequently, most existing ELVC systems focus on converting EL speech to another speaker instead of restoring the patient's own voice.

When pre-operative recordings of a patient's natural (NL) voice are available, personalized ELVC that restores both intelligibility and speaker identity becomes highly desirable. Previous studies have explored converting EL speech to a patient's original voice using VC systems trained with parallel pre- and post-operative recordings [9]. However, such approaches typically require a substantial amount of pre-operative speech data. In real-world clinical scenarios, collecting large amounts of pre-operative recordings is often impractical. Patients are rarely recorded extensively before TL, and in many cases, only a very limited amount of NL speech is available. Recent progress in zero-shot and one-shot VC suggests that speaker identity can be inferred from a single reference utterance [10, 11, 12, 13, 14]. In related biomedical tasks such as dysarthric speech reconstruction (DSR), two-stage frameworks that separate intelligibility restoration and speaker identity transfer have shown promising results [15, 16]. These observations motivate the exploration of similar strategies for personalized ELVC under extremely limited data conditions.

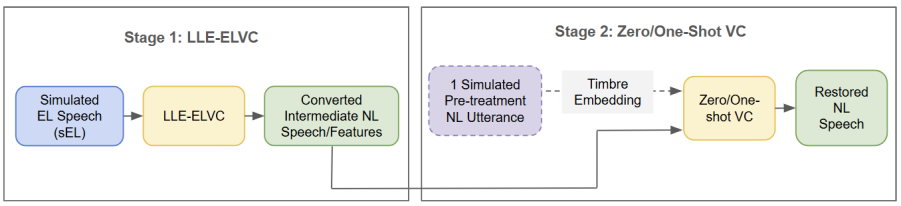
In this study, we investigated *one-shot personalized ELVC*, where only a single NL utterance from the source speaker is available as the timbre reference for voice restoration. To our knowledge, this work presents the first systematic benchmark for one-shot personalized ELVC with only a single simulated pre-operative utterance. Motivated by the failure of direct zero-shot VC on EL speech, we explored two strategies. The first is a **cascade approach** inspired by two-stage DSR [16], where EL speech is first converted to intelligible NL speech using the locally linear embedding (LLE)-based ELVC (LLE-ELVC) [17, 18] and then passed to a pretrained zero/one-shot VC model to restore speaker identity. The second is a **pseudo-target strategy** that uses pretrained VC models to synthesize pseudo training targets for personalized ELVC. We evaluated three pretrained zero/one-shot VC backbones: FreeVC [10], Vevo [11], and Seed-VC [12]. Experimental results show that directly applying zero/one-shot VC models to EL speech fails due to severe domain mismatch. In contrast, the cascaded framework significantly improves speaker similarity while maintaining comparable intelligibility. By feeding intermediate LLE-ELVC output features directly into the second-stage VC models, the cascade framework further reduces syllable error rate (SER). With supervised refinement, the best cascaded system slightly surpasses a supervised ELVC baseline that assumes access to 240 simulated pre-operative NL utterances in restoring intelligibility, while demonstrating improved speech quality and substantially higher speaker similarity.

The main contributions of this work are summarized as follows:

1. We formulated and benchmarked the problem of **one-shot personalized ELVC**, where only a single NL utterance from the target speaker is available for timbre preservation.
2. We investigated two strategies for one-shot personalized ELVC: a pseudo-target strategy and a cascaded framework that separates timbre transfer from intelligibility restoration.
3. We demonstrated that feature-level cascading and super-



(a) *The pseudo-target strategy*



(b) *The cascade approach*

Figure 1: *Illustration of two proposed personalized ELVC frameworks for restoring speaker identity using single target NL utterance while enhancing speech intelligibility.*

vised refinement enhance cross-stage compatibility, achieving strong objective and subjective improvements in speaker similarity while maintaining high intelligibility.

2. Proposed Methods

2.1. The Pseudo-target Strategy

In this work, our setup used only 240 EL-NL parallel speech utterances for ELVC training. As shown in Fig. 1a, with a pretrained zero-/one-shot VC model, we first converted the timbre of the 240 recordings of an NL speaker to match that of the single simulated pre-operative recording of the EL speaker, thereby generating 240 pseudo target utterances. LLE-ELVC was then trained to convert EL speech to pseudo targets, improving both intelligibility and speaker similarity.

2.1.1. Zero- and One-shot VC Methods

We evaluated three VC models that support single-reference timbre conversion.

- **FreeVC** [10]: A VITS-based one-shot VC model that performs timbre conversion from a single reference utterance. It uses WavLM [19] features for content representation and an information bottleneck to filter speaker characteristics.
- **Vevo** [11]: A zero-shot VC framework that decouples style, timbre, and linguistic content. Since our task only required timbre conversion, we used the flow-matching transformer component, which uses HuBERT [20] for continuous hidden features.
- **Seed-VC** [12]: A zero-shot VC model that incorporates an external timbre shifter to reduce timbre leakage and mismatches between training and inference. It utilizes Whisper-small [21] as the semantic encoder and a diffusion transformer to produce a more robust timbre representation from reference speech.

2.1.2. LLE-VC and LLE-ELVC

Exemplar-based methods such as LLE-VC [22, 23] have been used in ELVC [17, 18] since they require relatively small parallel training data. In this work, we adopted an LLE-VC framework [18] consisting of four stages: neighbor search, weight estimation, feature reconstruction, and waveform synthesis with a HiFi-GAN vocoder [24].

Following previous ELVC studies, Chinese-HuBERT features [25] were used for robust content representation during neighbor search, while WavLM-6th layer features [19] were used for weight estimation and reconstruction. The HiFi-GAN vocoder [24] was trained on LibriSpeech [26] and COSPRO Mandarin [27] corpora.

2.2. The Cascade Approach

In addition to the pseudo-target strategy, we evaluated an alternative framework that cascades LLE-ELVC with a zero/one-shot VC method (Fig. 1b). In this framework, EL speech was first converted to intelligible NL speech through LLE-ELVC. The converted waveforms were then passed to the pretrained FreeVC, Vevo, or Seed-VC models with default settings. Using the single simulated pre-operative NL utterance as reference, the zero- or one-shot VC model enhanced speaker similarity while preserving intelligibility gained from LLE-ELVC.

2.2.1. Cascading with LLE-ELVC Output Features

We further evaluated a feature-level cascade that directly feeds LLE-ELVC output features into zero- or one-shot VC models, bypassing intermediate waveform synthesis. LLE-ELVC was configured to output the same semantic features used by FreeVC, Vevo, or Seed-VC, reducing artifacts from waveform synthesis and re-encoding.

2.2.2. Supervised Fine-tuning of Zero-shot VC Methods

To further refine the cascade approach, we performed speaker-dependent supervised fine-tuning on the zero- or one-shot VC models (Fig. 2), adapting them to LLE-ELVC-generated waveforms and the target speaker's voice characteristics. LLE-ELVC outputs were generated using leave-one-utterance-out (LOUO) on the 240 EL-NL training utterance pairs. The resulting 240 utterances (features) served as source speech, while the corresponding Seed-VC generated pseudo-targets (Sec. 2.1) were used for supervision. Dynamic time warping (DTW) [28] was applied instead of the default truncation to align source and tar-

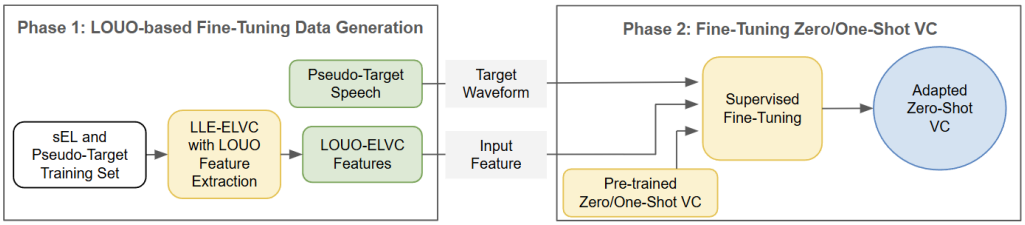


Figure 2: *Illustration for the supervised fine-tuning of zero- and one-shot VC methods using pseudo-target speech and LOUO-generated LLE-ELVC output.*

get utterances of different lengths. The best checkpoint within 600 steps was selected based on SER, speaker similarity, and model loss. During fine-tuning, the trainable Seed-VC modules were updated except for the vocoder.

3. Experimental Setup and Results

3.1. Experimental Setup

3.1.1. Datasets

Our dataset consists of three male (EL01–NL08–NL01, EL08–NL01–NL08, BYEL02–NL08–BNL02) and one female (EL06–NL05–NL06) speaker pairs. Since real pre-operative recordings from laryngectomy patients are rarely available, we simulated the one-shot personalized ELVC scenario using non-laryngectomized speakers. EL speech (EL01, EL08, BYEL02, and EL06) was produced by these speakers using an electrolarynx, and the corresponding NL recordings (NL01, NL08, BNL02, and NL06) were treated as pre-operative speech. For each speaker pair, we additionally assumed the presence of another NL speaker with more recordings available. For all speakers, we obtained 320 recordings of 10-character Mandarin sentences from the Taiwanese Mandarin Hearing in Noise Test (TMHINT) dataset [29], with all audio resampled to 16 kHz. The recordings were split into 240/40/40 utterances for training, development, and evaluation, respectively. Under the one-shot assumption, only the first utterance from the development set of the target speaker was used as the identity reference.

3.1.2. Compared Systems

We evaluated the following systems:

- **Supervised ELVC Baseline**: LLE-ELVC trained using fully paired EL–NL data, assuming access to 240 pre-operative NL utterances.
- **Direct Zero-shot VC**: applying Seed-VC, Vevo, or FreeVC directly to EL speech.
- **Pseudo-target Strategy** (Sec. 2.1): pseudo-target speech generated by zero-/one-shot VC models was used as supervision for LLE-ELVC.
- **Cascade (Waveform)** (Sec. 2.2): LLE-ELVC followed by zero-/one-shot VC on reconstructed waveforms.
- **Cascade (Feature)** (Sec. 2.2.1): intermediate acoustic features were directly passed to the second-stage VC model.
- **Refined Cascade** (Sec. 2.2.2): further improving the feature-level cascade through speaker-dependent supervised refinement.

3.1.3. Evaluation Metrics

We report four objective metrics: SpeechBERTScore (semantic fidelity) [30], SER for intelligibility, speaker similarity based on embedding cosine similarity, and UTMOS [31] for predicted perceptual quality. The embedding cosine similarity was calculated with a resemblyzer provided by the Amphion toolkit [32]. To compute the SER, we trained a syllable recognition system using the 196-hour MATBN corpus [33]. In addition, subjective listening tests were conducted to evaluate perceived similarity and intelligibility.

3.2. Results and Analysis

3.2.1. Direct Zero-Shot Failure and Structural Comparison

Table 1 summarizes the average performance across four speaker pairs. The unprocessed EL input exhibits degraded intelligibility (SER: 79.48%) and low perceptual quality (UTMOS: 1.31), confirming the need for enhancement. Directly applying pretrained zero- or one-shot VC models to EL speech leads to severe degradation. For example, FreeVC produces SER exceeding 100%, while Seed-VC and Vevo yield SER above 85%, which are worse than the unprocessed EL speech. These results indicate a strong distribution mismatch between EL speech and the NL speech assumed by pretrained VC models, motivating the need for an EL restoration stage before timbre transfer.

We then compare the two strategies for one-shot personalization. Using Seed-VC as the backbone, the cascaded framework (Waveform + Seed-VC) achieves substantially higher speaker similarity (0.82) and UTMOS (1.91) than the pseudo-target strategy (0.74 similarity and 1.64 UTMOS), although the pseudo-target approach yields slightly lower SER (65.70% vs. 69.65%). Similar trends are observed with Vevo and FreeVC. Overall, these results suggest that separating intelligibility restoration and timbre modeling is a more effective design under the one-shot constraint.

3.2.2. Improving the Cascade Approach

In the waveform-level cascade method, the second-stage VC model operates on reconstructed waveforms from LLE-ELVC, which may introduce vocoder artifacts. To mitigate this issue, we adopt feature-level cascading where intermediate acoustic representations are directly fed into the VC backbone.

For Seed-VC, this modification significantly improves intelligibility, reducing SER from 69.65% to 60.08% while maintaining high speaker similarity (0.82 $\rightarrow$ 0.81) and improving perceptual quality (UTMOS: 1.91 $\rightarrow$ 2.08). These results indicate that bypassing intermediate waveform reconstruction better preserves the acoustic structure required by the pretrained VC model. Building upon this design, we further introduce super-

Table 1: *Average performance across four speaker pairs. SER indicates Syllable Error Rate (%). Lower SER is better; higher values are better for other metrics.*

Category	System	SpeechBERT ↑	UTMOS ↑	SER (%) ↓	Similarity ↑
Input	EL (Unprocessed)	0.62	1.31	79.48	0.52
Baselines	Supervised ELVC	**0.74**	1.86	61.98	0.75
	Direct Zero-Shot (Seed-VC)	0.68	1.65	85.03	0.78
	Direct Zero-Shot (Vevo)	0.62	1.49	91.35	0.70
	Direct Zero-Shot (FreeVC)	0.62	1.96	101.48	0.68
Pseudo-Target	+ Seed-VC	0.72	1.64	65.70	0.74
	+ Vevo	0.71	1.67	66.25	0.72
	+ FreeVC	0.71	1.94	68.50	0.66
Cascade (Waveform)	+ Seed-VC	0.73	1.91	69.65	0.82
	+ Vevo	0.72	2.17	69.98	0.81
	+ FreeVC	0.69	**2.64**	79.70	0.73
Cascade (Feature)	+ Seed-VC	0.73	2.08	60.08	0.81
	+ Vevo	0.71	1.96	72.00	0.80
	+ FreeVC	0.69	2.40	82.20	0.70
Refined Cascade (Feature)	+ Seed-VC	**0.74**	2.30	**59.65**	**0.84**

vised refinement of the second-stage VC model. The refined cascade method (Feature + Seed-VC) achieves the best overall performance in Table 1, reducing SER to 59.65% (vs. 61.98%) and increasing speaker similarity to 0.84 (vs. 0.75) compared with the supervised LLE-ELVC baseline. It also maintains comparable SpeechBERTScore (0.74) while improving UTMOS (2.30 vs. 1.86), likely due to the neural speech synthesis architecture used in the second-stage VC model. Through further analysis of the cascade method, we observe that the first-stage LLE-ELVC model alone achieves SER 65.58%, while applying the second-stage VC further reduces SER, suggesting that the pretrained VC backbone can refine acoustic realization by projecting intermediate speech representations toward the natural speech manifold learned from large-scale training data. Example audio samples are available on the demo website[1].

3.2.3. Backbone-Specific Analysis

We also analyze the behavior of different VC backbones. FreeVC achieves the highest perceptual quality in the waveform cascade (UTMOS: 2.64) but suffers from poor intelligibility (SER: 79.70%), indicating sensitivity to residual EL artifacts. Vevo shows moderate performance across metrics but does not outperform Seed-VC. In contrast, Seed-VC consistently provides the most balanced trade-off between intelligibility, speaker similarity, and perceptual quality. One possible explanation is that these VC models are pretrained on different corpora and objectives, resulting in varying robustness to distribution shifts introduced by EL restoration.

3.2.4. Subjective Evaluation

We conducted a subjective evaluation comparing the supervised ELVC baseline with Refined Cascade (Feature), the best-performing system (denoted as S1 and S2, respectively). Thirteen listeners participated using headphones in a quiet environment.

Speaker similarity (ABX). Listeners were presented with two converted samples (A and B) and a reference sample (X), an NL utterance from the target speaker serving as the pre-operative reference. For each trial, listeners selected the sample that sounded more similar to the reference speaker. Across 40 trials, the cascaded system is significantly preferred over the supervised baseline (77.71% vs 9.38%), while 12.92% of the responses indicated no perceptual difference (Table 2).

Table 2: *Subjective evaluation results for S1 vs S2. MOS scores are reported as 95% confidence interval. Win/Same rate show the percentages for listener choices in the ABX test.*

System	MOS ↑	Win (%)	Same (%)
S1	2.869 ± 0.102	9.38	12.92
S2	3.090 ± 0.093	77.71	

Intelligibility MOS. Listeners also rated intelligibility on a 1–5 scale. The 95% confidence intervals in Table 2 indicate that the proposed cascade achieves slightly higher intelligibility while significantly improving perceived speaker similarity.

4. Conclusion

This paper establishes a benchmark for personalized ELVC in an extreme one-shot setting where only a single NL utterance is available to preserve speaker identity. We compared a pseudo-target strategy and a cascaded framework that separates intelligibility restoration from timbre transfer. Experimental results show that direct zero-shot VC fails due to severe domain mismatch. In contrast, the cascaded framework effectively restores intelligibility prior to timbre transfer. With feature-level cascading and supervised refinement, our system surpasses a supervised ELVC baseline in intelligibility while significantly improving speaker similarity and perceptual quality. Future work will investigate the latency and robustness of the framework on larger speaker populations, alternative evaluation methods, and more ELVC front-end models.

[1] https://devchang918.github.io/IS2026_one_shot_personalized_ELVC/

5. Acknowledgments

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

6. Generative AI Use Disclosure

Generative AI tools were used exclusively for manuscript editing and code refinement. The research design, experiments, and analysis were original and conducted by the authors.

7. References

[1] R. Kaye, C. G. Tang, and C. F. Sinclair, "The electrolarynx: voice restoration after total laryngectomy," *Medical Devices: Evidence and Research*, vol. 10, pp. 133–140, 2017.

[2] G. C. Davatz, R. Yamasaki, R. F. Sanchez, and A. N. Montagnoli, "Opinions of total laryngectomized individuals on the sound quality of electrolarynx," in *Proc. XXIX Brazilian Congress on Biomedical Engineering*, 2025, pp. 964–974.

[3] Z. Qian, H. Niu, L. Wang, K. Kobayashi, S. Zhang, and T. Toda, "Mandarin electro-laryngeal speech enhancement based on statistical voice conversion and manual tone control," in *Proc. APSIPA ASC*, 2021, pp. 546–552.

[4] M. Li, L. Wang, Z. Xu, and D. Cai, "Mandarin electrolaryngeal voice conversion with combination of gaussian mixture model and non-negative matrix factorization," in *Proc. APSIPA ASC*, 2017, pp. 1360–1363.

[5] 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, pp. 650–657.

[6] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," in *Proc. EUSIPCO*, 2018, pp. 2115–2119.

[7] D. Ma, J. Mi, F. Li, L. P. Violeta, K. Kobayashi, and T. Toda, "Improving electrolaryngeal speech enhancement via a representation learning method based on integrated text and speech representations," in *Proc. EMBC*, 2025, pp. 1–6.

[8] L. P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda, "Resolving domain mismatches in electrolaryngeal speech enhancement with linguistic intermediates," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 5, pp. 827–839, 2025.

[9] N. Nishio, K. Kobayashi, D. Ma, S. Mitani, M. Sone, and T. Toda, "A voice conversion system from electrolarynx speech to preoperative patient's speech for total laryngectomy," *OTO Open*, vol. 10, no. 1, p. e70207, 2026.

[10] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP*, 2023, pp. 1–5.

[11] 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, pp. 1–24.

[12] S. Liu, "Zero-shot voice conversion with diffusion transformers," 2024. [Online]. Available: https://arxiv.org/abs/2411.09943

[13] A. Lobashev, A. Yermekova, and M. Larchenko, "Training-free voice conversion with factorized optimal transport," in *Proc. ICASSP*, 2025, pp. 1373–1377.

[14] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, "Coding speech through vocal tract kinematics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 8, pp. 1427–1440, 2024.

[15] W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-m. Wang, and T. Toda, "A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion," in *Proc. ICASSP*, 2021, pp. 1329–1333.

[16] D. Chang, M.-C. Yen, H.-T. Hwang, F.-R. Li, C.-F. Liu, Y. Tsao, and H.-M. Wang, "Two-stage voice conversion for dysarthric speech reconstruction with speaker identity preservation," in *Proc. TAAI*, 2025.

[17] 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, pp. 1–6.

[18] 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, pp. 730–735.

[19] 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, 2022.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, p. 3451–3460, oct 2021.

[21] 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, pp. 28 492–28 518.

[22] Y.-C. Wu, H.-T. Hwang, , C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," in *Proc. INTERSPEECH*, 2016, pp. 1652–1656.

[23] H.-T. Hwang, Y.-C. Wu, Y.-H. Peng, C.-C. Hsu, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Voice conversion based on locally linear embedding," *Journal of Information Science and Engineering*, vol. 34, no. 6, pp. 1493–1516, 2018.

[24] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NIPS*, vol. 33, 2020, pp. 17 022–17 033.

[25] P. Guo and S. Liu, "chinese_speech_pretrain," 2022. [Online]. Available: https://github.com/TencentGameMate/chinese_speech_pretrain

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[27] 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, pp. 23–28.

[28] 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.

[29] L. Wong, S. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. 28, pp. 70S–74S, 05 2007.

[30] 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, pp. 4943–4947.

[31] 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, pp. 4521–4525.

[32] X. Zhang, L. Xue, Y. Gu, Y. Wang, J. Li, H. He, C. Wang, S. Liu, X. Chen, J. Zhang, Z. Fang, H. Chen, T. Y. Tang, L. Zou, M. Wang, J. Han, K. Chen, H. Li, and Z. Wu, "Amphion: an open-source audio, music, and speech generation toolkit," in *Proc. SLT*, 2024, pp. 879–884.

[33] 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.