Improving Exemplar-Based Electrolaryngeal Speech Voice Conversion via Robust Content Representations

Fo-Rui Li*†, Hsin-Te Hwang†, Ming-Chi Yen‡, Men-Tung Lo*, Yu Tsao†, Hsin-Min Wang‡,
* Dept. of Biomedical Science and Engineering, National Central University, Taoyuan, Taiwan
E-mail: f323red@gmail.com, mzlo@ncu.edu.tw
† Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: {hwanght, yu.tsao}@citi.sinica.edu.tw
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {ymchiqq, whm}@iis.sinica.edu.tw

***Abstract*—Electrolaryngeal speech voice conversion (ELVC) aims to improve the naturalness and intelligibility of electrolaryngeal (EL) speech. A key challenge lies in the limited availability of EL data. We previously proposed a locally linear embedding-based ELVC (LLE-ELVC) method that leverages features from the sixth layer of WavLM Large and exemplar-based techniques to reduce reliance on large corpora. Building on recent findings that using advanced content representations can improve speech quality, we extend LLE-ELVC by incorporating robust content features for neighbor search while retaining prosody-rich features for speech reconstruction. Evaluated on a Mandarin corpus, our method consistently improves both objective and subjective metrics. Among the foundation models evaluated, WavLM Large, Chinese-HuBERT Large, and Whisper Large V3 show the most robust performance across different speakers and recording conditions. These results confirm the effectiveness of integrating content representations into LLE-ELVC.**

I. Introduction

Electrolaryngeal (EL) speech is produced by people who have lost their ability to produce natural (NL) speech due to total laryngectomy using an external electrolarynx. However, EL speech often sounds robotic and is difficult to understand due to the absence of natural glottal and prosodic cues. Electrolaryngeal speech voice conversion (ELVC), which converts EL speech to NL speech, has clinical and social value. Early ELVC studies used traditional acoustic features and statistical or neural models, but achieved limited improvements in speech quality [1]–[3]. Although deep learning-based methods [4]–[7] have shown promise, they usually require large parallel corpora to train accurate mapping functions.

To address these issues, we previously proposed a locally linear embedding-based ELVC (LLE-ELVC) approach [8] that integrates self-supervised learning (SSL) representations into an exemplar-based framework. LLE-ELVC avoids model training and comprises four stages: neighbor search in the EL domain, LLE weight estimation, feature reconstruction using aligned NL exemplars, and waveform synthesis. All components use prosody-rich features from the sixth layer of WavLM Large [9] (hereafter referred to as WavLM-6th features), as suggested in [10]. This design reduces the reliance on large datasets and achieves higher intelligibility than neural network-based methods [11]. However, in this study, our analysis shows that the conversion quality degrades significantly when the recorded EL speech is highly unintelligible.

Recent work on voice conversion has shown that advanced content representations such as those from ContentVec [12] and Whisper [13] can improve conversion accuracy [14], [15]. Notably, [16] reported improved performance in singing voice conversion by using the average of the features from the last five layers of WavLM Large for matching, highlighting the advantage of using richer content features in their exemplar-based system. Inspired by this, we extend LLE-ELVC by introducing content features in the neighbor search stage while retaining WavLM-6th features for LLE weight estimation and reconstruction. This flexible design allows the integration of advanced foundation models to improve intelligibility and stability in ELVC.

We evaluate five content models on a Mandarin EL/NL corpus: WavLM Large, HuBERT Large [17], Chinese-HuBERT Large [18], ContentVec [12], and Whisper Large V3 [13]. The proposed system consistently improves both objective and subjective metrics. WavLM Large, Chinese-HuBERT Large, and Whisper Large V3 show the most robust performance. Notably, Chinese-HuBERT Large outperforms HuBERT Large, highlighting the benefits of training specifically for Mandarin.

Our main contributions are as follows:

1) We propose a feature-agnostic modification of LLE-ELVC that supports arbitrary content representations without the need for additional network training.
2) Our comprehensive experiments show that stronger content features can improve quality, intelligibility, and stability in ELVC.
3) Our analysis reveals that Mandarin-focused pre-training and clear EL speech articulation contribute substantially

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

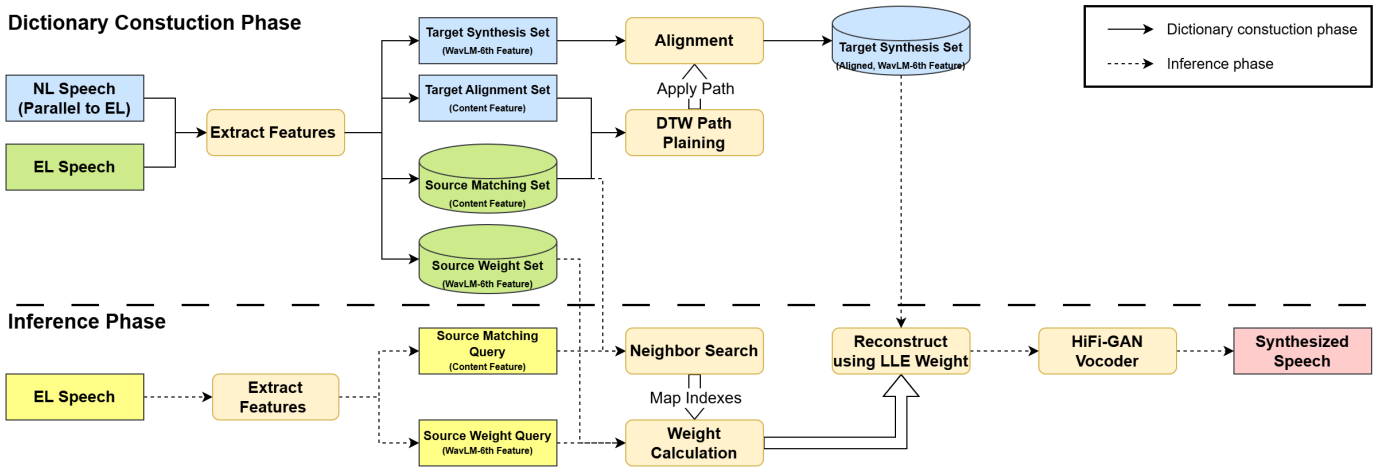


Fig. 1. Improved LLE-ELVC architecture.
Colored rectangles: waveform/features; Rounded rectangles: operations; Flattened cylinders: fixed dictionaries.

to performance improvements, suggesting important directions for future research.

II. Methodology

This section presents the improved LLE-ELVC framework, which extends our previous work [8] to enable flexible integration of content representations. As shown in Fig. 1, the system operates in two phases: dictionary construction and inference.

A. Dictionary Construction

In the dictionary construction phase, four feature vector sets are built using paired EL and NL utterances: *Source Matching Set*, *Source Weight Set*, *Target Alignment Set*, and *Target Synthesis Set*. Feature extraction is performed using pretrained foundation models to obtain content features (e.g., the last layer of HuBERT Large) and prosody-rich features (e.g., WavLM-6th), depending on the role of each feature set.

- **Source Matching Set**: contains content features from EL speech and is used for nearest-neighbor search.
- **Source Weight Set**: contains WavLM-6th features from EL speech and is used for LLE weight estimation.
- **Target Alignment Set**: contains content features from NL speech and is used as references for dynamic time warping (DTW) [19] alignment.
- **Target Synthesis Set**: contains WavLM-6th features from NL speech and is used for final waveform reconstruction.

DTW is applied to the *Source Matching Set* and the *Target Alignment Set* to obtain pairs of aligned content feature vectors. After the dictionary construction process, each vector in the *Source Matching Set* corresponds to a vector in the *Source Weight Set* and a vector in the *Target Synthesis Set*. No model training is required except for the dictionary construction process.

B. Inference

During inference, EL speech is processed by two pretrained foundation models to produce two parallel queries:

- **Source Matching Query**: contains frame-wise content features and is used to search for neighbors in the *Source Matching Set* frame-by-frame.
- **Source Weight Query**: contains WavLM-6th features for frame-by-frame LLE weight estimation.

Let $\mathbf{x}_t^{\mathrm{M}} \in \mathbb{R}^d$ denote the content vector of the t-th frame of *Source Matching Query* and $\mathbf{x}_t^{\mathrm{W}} \in \mathbb{R}^d$ denote the corresponding WavLM-6th feature vector of *Source Weight Query*. The inference process is as follows:

1) **Neighbor Search**: For each $\mathbf{x}_t^{\mathrm{M}}$, we search for its k nearest neighbors from the *Source Matching Set* and obtain the index set $\mathcal{N}_t = \{n_1, \ldots, n_k\}$.
2) **Weight Estimation**: According to the index set $\mathcal{N}_t$, we collect the corresponding vectors $\mathbf{b}_1, \ldots, \mathbf{b}_k$ from the *Source Weight Set*. The weight vector $\mathbf{w}_t = [w_1, \ldots, w_k]^\top \in \mathbb{R}^k$ is obtained by minimizing the following error:

$$\mathbf{w}_t = \arg\min_{\mathbf{w}} \left\| \mathbf{x}_t^{\mathrm{W}} - \sum_{i=1}^{k} w_i \mathbf{b}_i \right\|^2 \quad \text{s.t.} \quad \sum_{i=1}^{k} w_i = 1. \tag{1}$$

3) **Feature Reconstruction**: According to the index set $\mathcal{N}_t$, we collect the corresponding vectors $\mathbf{y}_1, \ldots, \mathbf{y}_k$ from the *Target Synthesis Set* and use $\mathbf{w}_t$ to reconstruct the converted feature vector $\hat{\mathbf{y}}_t$ as:

$$\hat{\mathbf{y}}_t = \sum_{i=1}^{k} w_i \mathbf{y}_i. \tag{2}$$

Finally, the reconstructed features are converted to waveform using a HiFi-GAN vocoder [20] trained on the LibriSpeech [21] and COSPRO Mandarin [22] corpora.

C. Relationship to Previous LLE-ELVC

When all components in our framework (Fig. 1) are set to use WavLM-6th features, the system is simplified to the previous LLE-ELVC [8]. The original configuration leverages

TABLE I
AVERAGE EVALUATION RESULTS OF THE PROPOSED METHOD USING DIFFERENT CONTENT FEATURES. **MOSA-NET+ I** AND **MOSA-NET+ Q** DENOTE THE INTELLIGIBILITY AND SPEECH QUALITY SCORES, RESPECTIVELY.

Content Feature	MCD (dB)↓	F0 RMSE↓	F0 CORR↑	CER (%)↓	SpeechBERTScore↑	MOSA-Net+ I↑	MOSA-Net+ Q↑
WavLM	6.06	40.82	0.22	54.0	0.71	0.87	2.95
HuBERT	6.18	38.96	0.23	73.0	0.70	0.80	2.70
Chinese-HuBERT	**6.02**	41.30	0.21	**50.8**	0.71	0.88	2.97
ContentVec	6.16	40.08	0.23	67.9	0.70	0.82	2.78
Whisper	6.03	40.73	0.23	55.3	**0.72**	**0.89**	**2.99**
Baseline	6.32	**38.77**	**0.24**	68.0	0.69	0.79	2.68
EL speech	10.95	68.05	0.03	83.5	0.59	0.68	2.27
NL speech	0.0	0.0	1.0	2.4	1.00	1.00	4.48

the strong prosodic and speaker-related cues embedded in the WavLM-6th features, which improve synthesis quality [10] but may hinder accurate content-based matching when the input EL speech is severely degraded.

In contrast, the proposed framework decouples the roles of these components, allowing different types of features to be applied at different stages. Specifically, we introduce content-rich representations for alignment and neighbor search, while retaining WavLM-6th features for weight estimation, feature reconstruction, and waveform synthesis. As a result, the proposed method generalizes the previous LLE-ELVC while preserving its original inference structure.

III. EXPERIMENTAL SETUP

A. Experimental Dataset

Our experimental dataset contains six healthy speakers, each reading a fixed set of 320 phonetically balanced sentences from the Taiwanese Mandarin Hearing in Noise Test (TMHINT) [23], both with and without an EL device. When recording EL speech, the speakers imitated a total laryngectomy patient using the EL device to produce an artificial voice. This results in 320 pairs of utterances for EL and NL speech per speaker. The first 240 pairs were used to construct the exemplar dictionary, while the remaining 80 pairs were used for testing. All speech signals were recorded in a studio at a sampling rate of 16 kHz.

B. Content Models

We evaluated five foundation models: WavLM Large (hereafter WavLM) [9], HuBERT Large (hereafter HuBERT) [17], Chinese-HuBERT Large (hereafter Chinese-HuBERT) [18], ContentVec [12], and Whisper Large V3 (hereafter Whisper) [13]. In our framework, these models are used as alternative content representations for DTW alignment and neighbor search. Specifically, we used the features of the final layer of each model, which were commonly used in previous work to capture abstract content information. In the improved LLE-ELVC framework, all content features and WavLM-6th features are extracted with a 20ms hop size, ensuring temporal alignment across components without additional resampling.

C. Evaluation Metrics

We use both objective and subjective metrics to comprehensively evaluate the effectiveness of each content representation. The volume of all generated audio was normalized to -24 dB before evaluation.

Mel-Cepstral Distortion (MCD) evaluates the spectral distortion between the converted speech and the reference speech, excluding the 0th-order mel-cepstral coefficient to reduce the impact of volume. F0 Root Mean Square Error (F0 RMSE) and F0 Correlation Coefficient (F0 CORR) evaluate pitch prediction accuracy and contour similarity, respectively. Character Error Rate (CER), calculated with transcriptions of the Whisper Large V3 [13] model, measures speech intelligibility.

In addition, we use two neural network-based metrics aligned with human assessment: SpeechBERTScore [24], a reference-based metric that quantifies semantic fidelity; and MOSA-Net+ [25], a non-intrusive model to estimate speech quality and intelligibility. In addition to these objective metrics, we also conducted a subjective A/B test to evaluate the most critical metric in the ELVC task–intelligibility. A total of 23 native Mandarin speakers were recruited for this evaluation.

IV. EXPERIMENTAL RESULTS

We provide complete details of the tables and audio samples available on our demo site[1]. The number of nearest neighbors k (Section II-B) is empirically set to 512 for all systems.

A. Overall Performance across Metrics

Table I summarizes the average performance over all speaker pairs. Compared to the baseline system (previous LLE-ELVC) [8], the proposed model shows notable improvements when using several content models, especially WavLM, Chinese-HuBERT, and Whisper. These models consistently outperform the baseline except for the F0 metrics. Notably, the proposed model with Chinese-HuBERT achieves the lowest CER (50.8%), which is a relative reduction of 25.3% compared to the baseline (68.0%).

The results evaluated by MOSA-Net+ show similar trends. In terms of MOSA-Net+ I (intelligibility), Whisper, Chinese-HuBERT, and WavLM achieve scores of 0.89, 0.88, and 0.87, respectively, compared to the baseline score of 0.79. For MOSA-Net+ Q (quality), the same models achieve 2.99, 2.97, and 2.95, respectively, surpassing the baseline score of 2.68. These results show that the improvement in perceived quality aligns with the observed reduction in CER.

[1]Demo site: https://f323red.github.io/improved-lle-elvc/

TABLE II
OVERVIEW OF TRAINING DATA FOR EACH CONTENT MODEL.

Model	Training Data
WavLM	Libri-light, VoxPopuli, GigaSpeech
HuBERT	Libri-light, LibriSpeech
Chinese-HuBERT	WenetSpeech-L
ContentVec	LibriSpeech
Whisper	Multilingual web-scale corpora

Training data references: Libri-light [26], VoxPopuli [27], GigaSpeech [28], LibriSpeech [21], WenetSpeech-L [29].

The baseline performs slightly better than the proposed method in terms of F0 RMSE and F0 CORR, likely due to its use of prosody-rich features during alignment and neighbor search, which may help preserve pitch continuity. However, the performance differences are small, indicating that reconstruction using WavLM-6th features remains effective across all models.

In terms of spectral similarity, WavLM, Chinese-HuBERT, and Whisper also achieve lower MCD values than the baseline, indicating higher reconstruction accuracy. They achieve slightly higher SpeechBERTScore values than the baseline, suggesting slightly improved semantic fidelity.

In summary, WavLM, Chinese-HuBERT, and Whisper achieve the most robust results in all objective metrics evaluated. These findings support the effectiveness of decoupling matching from reconstruction in our LLE-ELVC framework, i.e., using content-rich representations for alignment and neighbor search while retaining WavLM-6th features for weight estimation, feature reconstruction, and waveform synthesis.

B. Model-wise Insights and Comparisons

While Section IV-A focused on the performance of all content models across various metrics, we now turn to model-specific insights to better understand the factors that contribute to performance differences. This section highlights five key observations regarding the content models evaluated.

First, ContentVec and HuBERT perform on par with or slightly below the baseline in most metrics. As shown in Table II, both are trained only on English audiobooks with limited language and domain diversity. Their content embeddings, while usable, may lack sufficient alignment with Mandarin EL speech, limiting their generalizability.

Second, comparing the last and sixth layers of WavLM, we found that deeper representations are more effective for content-based matching. Importantly, WavLM is pretrained with noisy or mixed speech using denoising and masked prediction objectives, which enhances its robustness under acoustic distortions and background interference. In contrast, models like HuBERT were trained primarily on clean speech data. WavLM's stronger noise resilience may be the reason for its superior spectral and intelligibility performance under degraded EL conditions.

Third, although Chinese-HuBERT and HuBERT share the same architecture and training method, the performance gap between them highlights the value of Mandarin-specific pretraining. Chinese-HuBERT is trained on a diverse and large-scale Mandarin corpus, which probably accounts for its superior performance in both intelligibility and spectral accuracy.

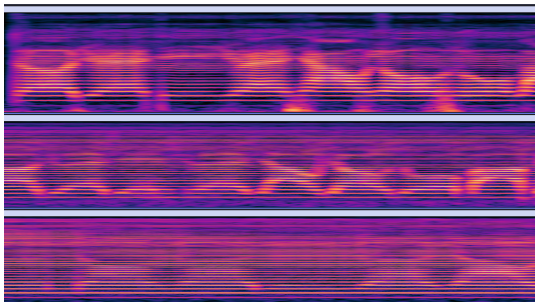


Fig. 2. Examples of EL speech spectrograms for three groups of EL speakers. Top: Easy group; Middle: Moderate group; Bottom: Difficult group.

Fourth, while ContentVec introduces disentanglement mechanisms on top of HuBERT-Base, its performance remains close to HuBERT-Large. This suggests that its potential may be better realized when combined with a larger or Mandarin-specific architecture. Future work can explore applying ContentVec-style training to more powerful or multilingual SSL backbones.

Fifth, Whisper presents a unique case. Unlike other models trained with masked prediction, it uses a weakly supervised encoder-decoder design optimized for speech recognition and translation. Despite the differences in architecture and training, its representations perform robustly in our framework, which may benefit from the large-scale multilingual and task-diverse pretraining. This result highlights the adaptability of our LLE-ELVC framework to a variety of content models.

C. Speaker-Level Analysis by EL Speech Intelligibility

To investigate how input speech quality affects model performance, we divided the six EL speakers into three levels based on the intelligibility of their EL speech (i.e., CER): Easy (Pair B02), Moderate (Pairs 01, 06, 07, 08), and Difficult (Pair 05). Fig. 2 shows representative spectrograms for each level. As illustrated, the formant structure in the Easy case is clear and distinct, while formant in the Difficult case is severely distorted or masked due to electrolarynx signal interference. These spectrogram differences align well with subjective impressions: speakers in the Easy case are generally easier to understand, while speakers in the Difficult case produce EL speech that is difficult to understand even for human listeners. Table III summarizes the objective evaluation results of each group under different models.

In the Easy case, all systems performed well. The speaker's slow and clear articulation and the quiet recording environment resulted in clear formant structures. Even without conversion, the CER of EL speech reached 55.4%.

In the Moderate group, all systems show moderate and consistent gains. Whisper, Chinese-HuBERT, and WavLM again achieved the best results, maintaining stable performance across all metrics.

In the Difficult case, the quality of EL speech was severely degraded due to mechanical noise and unstable articulation,

TABLE III
OBJECTIVE EVALUATION RESULTS FOR DIFFERENT GROUPS OF EL SPEAKERS.

Content Feature	Difficulty	MCD (dB)↓	F0 RMSE↓	F0 CORR↑	CER (%)↓	SpeechBERTScore↑	MOSA-Net+ I↑	MOSA-Net+ Q↑
WavLM	Easy	5.88	29.33	0.25	29.0	0.81	0.98	3.63
	Moderate	6.04	41.27	0.22	57.1	0.70	0.87	2.86
	Difficult	6.30	50.52	0.23	66.5	0.64	0.80	2.60
HuBERT	Easy	5.95	29.49	0.23	39.4	0.80	0.97	3.52
	Moderate	6.12	40.12	0.22	72.7	0.69	0.81	2.65
	Difficult	6.68	43.81	0.30	107.9	0.61	0.61	2.09
Chinese-HuBERT	Easy	5.93	28.28	0.28	30.4	0.80	0.98	3.60
	Moderate	6.00	42.53	0.18	54.1	0.70	0.87	2.89
	Difficult	6.20	49.44	0.27	58.1	0.64	0.82	2.65
ContentVec	Easy	5.93	28.54	0.28	30.7	0.81	0.98	3.62
	Moderate	6.06	41.12	0.21	63.0	0.70	0.83	2.74
	Difficult	6.79	47.43	0.29	124.6	0.60	0.60	2.08
Whisper	Easy	5.96	28.36	0.28	31.6	0.80	0.98	3.61
	Moderate	5.99	42.31	0.19	57.9	0.71	0.89	2.94
	Difficult	6.31	46.80	0.33	68.8	0.65	0.79	2.56
Baseline	Easy	6.01	29.00	0.26	25.9	0.80	0.97	3.59
	Moderate	6.22	39.98	0.22	61.6	0.69	0.80	2.65
	Difficult	7.06	43.72	0.32	135.9	0.58	0.53	1.90
EL speech	Easy	14.16	29.55	0.01	55.4	0.67	0.83	2.75
	Moderate	10.41	64.37	0.02	83.6	0.59	0.67	2.20
	Difficult	9.90	121.29	0.00	111.1	0.53	0.61	2.04
NL speech	Easy	0.0	0.0	1.0	1.4	1.00	1.00	4.64
	Moderate	0.0	0.0	1.0	2.5	1.00	1.00	4.42
	Difficult	0.0	0.0	1.0	3.1	1.00	1.00	4.56

and all systems performed poorly. The CER of the original EL speech, the speech converted by the systems using HuBERT and ContentVec, and the speech converted by the baseline system exceeded 100%. Despite this, the systems using Whisper, Chinese-HuBERT, and WavLM still produced relatively intelligible output, demonstrating their robustness.

These findings confirm that robust content models, especially Chinese-HuBERT, Whisper, and WavLM, have better generalization capabilities under adverse input conditions. They also noted that speaker factors (such as articulation and EL device proficiency) significantly affect ELVC performance. Therefore, improving the intelligibility of the original EL speech through user training may be beneficial, even if it is not related to voice conversion technology.

D. Subjective Evaluation

We conducted an A/B test with 23 native Mandarin speakers to evaluate the intelligibility of utterances converted by LLE-ELVC using only WavLM-6th features (baseline) and LLE-ELVC using Chinese-HuBERT (achieving the lowest CER in Table I) and WavLM-6th features (proposed). We randomly selected a pair of speakers from the Easy group (pair-b02, CER: 55.4%), the Moderate group (pair-07, CER: 84.7%), and the Difficult group (pair-05, CER: 111.1%), respectively, with 15 utterances in each group.

From the results of the A/B test, we observed that listeners did not show a strong preference for either system in the Easy group (Proposed: 29.6%, Baseline: 27.8%, Same: 42.6%), as both systems produced similarly intelligible outputs under favorable EL speech input conditions. In the Moderate group, listeners began to show a moderate preference for the output produced using Chinese-HuBERT features (Proposed: 34.5%, Baseline: 23.2%, Same: 42.3%), indicating that it is easier to understand. Finally, in the Difficult group, listeners clearly preferred the output produced using Chinese-HuBERT features (Proposed: 79.1%, Baseline: 5.5%, Same: 15.4%), highlighting that LLE-ELVC using Chinese-HuBERT yielded noticeably higher intelligibility under the most challenging EL speech conditions. These findings are consistent with the trends observed in the objective metrics and further demonstrate the robustness of the proposed model when dealing with degraded input speech.

V. CONCLUSION

This study shows that incorporating robust content features (particularly from WavLM, Chinese-HuBERT, and Whisper) significantly improves the stability of the LLE-ELVC framework and the intelligibility and quality of its output speech. These content models help the proposed system consistently outperform the baseline system in both objective and subjective metrics, highlighting the importance of language-aligned and noise-robust content embeddings in exemplar-based ELVC.

Our speaker-level analysis further shows that articulation clarity and device proficiency substantially affect ELVC performance. These user-side factors are particularly important when the quality of input EL speech is severely degraded. In addition, the results show that content models pretrained on Mandarin (e.g., Chinese-HuBERT) have a clear advantage over models pretrained only on English, underscoring the importance of training data alignment in the low-resource ELVC task. For future work, we suggest exploring multi-model integration, audio-visual features (e.g., AV-HuBERT [30]), and developing vocoders tailored to these representations.

REFERENCES

[1] 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.

[2] P. Malathi, G. Sureshw, and M. Moorthi, "Enhancement of electrolaryngeal speech using frequency auditory masking and GMM based voice conversion," in *Proc. AEEICB*, 2018.

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

[4] Z. Qian, K. Xiao, and C. Yu, "Mandarin electrolaryngeal speech enhancement using cycle-consistent generative adversarial networks," *Applied Sciences*, vol. 13, no. 1, p. 537, 2022.

[5] Y. Yang, H. Zhang, Z. Cai, *et al.*, "Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion," *Biomedical Signal Processing and Control*, vol. 80, p. 104 279, 2023.

[6] M.-C. Yen, C.-H. Wu, S.-W. Tsai, *et al.*, "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.

[7] 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.

[8] 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.

[9] S. Chen, C. Wang, Z. Chen, *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.

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

[11] H.-H. Chen, Y.-L. Chien, M.-C. Yen, *et al.*, "Mandarin electrolaryngeal speech voice conversion using cross-domain features," in *Proc. INTERSPEECH*, 2023.

[12] K. Qian, Y. Zhang, H. Gao, *et al.*, "ContentVec: An improved self-supervised speech representation by disentangling speakers," in *Proc. ICML*, vol. 162, 2022.

[13] 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.

[14] S. Liu, "Zero-shot voice conversion with diffusion Transformers," *arXiv preprint arXiv:2411.09943*, 2024.

[15] X. Zhang, Z. Fang, Y. Gu, *et al.*, "Leveraging diverse semantic-based audio pretrained models for singing voice conversion," in *Proc. SLT*, 2024.

[16] B. Sha, X. Li, Z. Wu, Y. Shan, and H. Meng, "Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion," in *Proc. ICASSP*, 2024.

[17] 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, pp. 3451–3460, 2021.

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

[19] 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.

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

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

[22] C.-Y. Tseng, Y.-C. Cheng, and C.-H. Chang, "Sinica COSPRO and toolkit — corpora and platform of Mandarin Chinese fluent speech," 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:204792394.

[23] L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M. W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and Hearing*, vol. 28, no. 2 Suppl, 70S–74S, 2007.

[24] 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.

[25] 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.

[26] J. Kahn, M. Riviere, W. Zheng, *et al.*, "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proc. ICASSP*, 2020.

[27] C. Wang, M. Riviere, A. Lee, *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. IJCNLP*, 2021.

[28] G. Chen, S. Chai, G.-B. Wang, *et al.*, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Proc. INTERSPEECH*, 2021.

[29] B. Zhang, H. Lv, P. Guo, *et al.*, "WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition," in *Proc. ICASSP*, 2022.

[30] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proc. ICLR*, 2022.