INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Audio-Visual Mandarin Electrolaryngeal Speech Voice Conversion

Yung-Lun Chien[1,2], *Hsin-Hao Chen*[1,2], *Ming-Chi Yen*[2], *Shu-Wei Tsai*[3],
Hsin-Min Wang[2], *Yu Tsao*[2], *and Tai-Shih Chi*[1]

[1] National Yang Ming Chiao Tung University, [2] Academia Sinica
[3] National Cheng Kung University Hospital

ajul1230@gmail.com,123ggg3304@gmail.com,ymchiqq@iis.sinica.edu.tw,tsaisuwei@gmail.com,
whm@iis.sinica.edu.tw,yu.tsao@citi.sinica.edu.tw,tschi@mail.nctu.edu.tw

Abstract

Electrolarynx is a commonly used assistive device to help patients with removed vocal cords regain their ability to speak. Although the electrolarynx can generate excitation signals like the vocal cords, the naturalness and intelligibility of electrolaryngeal (EL) speech are very different from those of natural (NL) speech. Many deep-learning-based models have been applied to electrolaryngeal speech voice conversion (ELVC) for converting EL speech to NL speech. In this study, we propose a multimodal voice conversion (VC) model that integrates acoustic and visual information into a unified network. We compared different pre-trained models as visual feature extractors and evaluated the effectiveness of these features in the ELVC task. The experimental results demonstrate that the proposed multimodal VC model outperforms single-modal models in both objective and subjective metrics, suggesting that the integration of visual information can significantly improve the quality of ELVC.

Index Terms: Electrolaryngeal speech, voice conversion, lip images, multimodal learning, feature extractor.

1. Introduction

The ability to speak and communicate is fundamental for human life. However, individuals who undergo laryngectomy lose the ability to produce excitation signals because of the removal of their vocal cords. This loss significantly affects their ability to speak normally, decreasing their overall quality of life. To address this issue, the use of the electrolarynx is the primary method for speech recovery. However, this device often produces a relatively flat fundamental frequency (F0) and generates noise that affects the voice quality, highlighting the need for improved electrolaryngeal (EL) speech techniques.

Voice conversion (VC) is a technique that converts a human voice from a source speaker to target speaker without changing the underlying content. One of the applications of VC is to improve the naturalness and intelligibility of EL speech [1, 2]; this VC task is called electrolaryngeal speech voice conversion (ELVC). A typical ELVC approach first extracts the acoustic features of EL speech and target natural (NL) speech and then trains a conversion model. When in use, the converted features are synthesized back into a waveform using a vocoder. For frame-based VC, aligning the acoustic features of paired EL and NL speech is critical before training the conversion model. Dynamic time warping (DTW) is the most commonly used algorithm for determining the best alignment path over two feature sequences based on a predefined distance (e.g., the Euclidean distance). However, in ELVC, the DTW algorithm often fails to find the correct alignment path and causes the model to fail in learning the correct conversion function, which seriously affects the performance of ELVC. To address this issue, Liou *et al.* used lip images instead of acoustic features for alignment [3]. Although this method achieved better ELVC results, it was not the best alignment method. In this study, we explored different alignment methods to improve the performance of ELVC.

In addition to its role in alignment, the lip shape may play an important role in speech signal processing [4]. Although users of the electrolarynx cannot speak normally, their lip movements are similar to those of healthy people. Therefore, the use of lip-shape information to improve the ELVC model is worth studying. Multimodal training methods have been employed in many speech-processing studies [5, 6], including the VC task [7]. In this study, we evaluated different visual feature extractors and determined the best one for the ELVC task. The main contributions of this study are twofold: i) the proposal of a new feature-alignment method suitable for frame-based ELVC, and ii) a novel multimodal VC architecture that uses both acoustic and visual features.

The remainder of this paper is organized as follows. Section 2 introduces the alignment methods, including the traditional and proposed methods. Section 3 introduces different lip-image feature extractors and their uses. Section 4 presents the experimental setup and various objective and subjective evaluations. Finally, Section 5 presents the conclusions of this study and directions for future research.

2. Alignment methods

In this section, we will introduce previous and our alignment methods for ELVC.

2.1. Previous alignment methods

As shown in Fig. 1, EL speech is generally longer than NL speech, even with the same linguistic content. Differences in the speech length can cause distortion of NL speech owing to the stretching of length during alignment. In addition to the very different acoustic properties of EL and NL speech, the length difference is one of the key challenges in aligning these two types of speech.

As a baseline, we used the WORLD vocoder [8] to decompose EL and NL speech into acoustic features, such as mel-cepstral coefficients (MCC). Subsequently, an alignment was performed based on the DTW algorithm using MCC. This method is referred to as DTW-MCC. The path calculation is based on the mel-cepstral distortion (MCD).

Liou *et al.* used the lip images of EL speech and NL speech to align both [3]. This approach involves first obtaining 20 lip landmarks using the dlib library [9], relocating the coordinates according to their centroid, and then calculating the Euclidean distance between the source and target landmark sets. Although the DTW-lip-landmark method was shown to outperform the

10.21437/Interspeech.2023-866

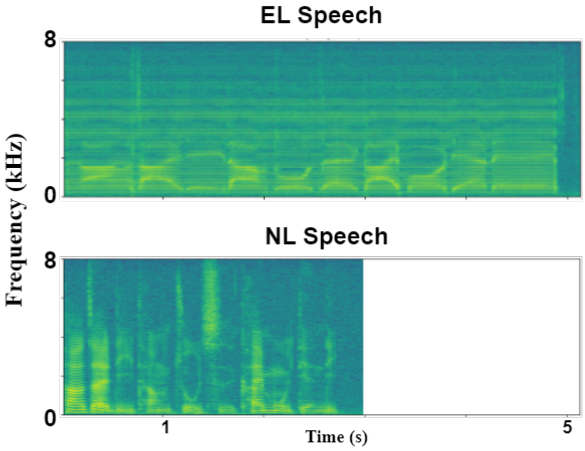

Figure 1: *Spectrogram plots of EL speech and NL speech.*

DTW-MCC method in the ELVC task in [3], the room for improvement exists.

2.2. Proposed alignment method

To address the misalignment caused by the difference in length of EL and NL speech, we applied the waveform similarity overlap-and-add (WSOLA) algorithm [10], which is a time-scale modification method that can adjust the speed of speech while preserving F0. Specifically, we used WSOLA to adjust the length of NL speech to match that of the EL speech, thereby reducing the distortion caused by the length difference. The modified DTW-MCC method that uses length-adjusted NL speech is referred to as the DTW-WSOLA method. We conducted preliminary listening tests and confirmed that the intelligibility of the NL speech was not compromised after length adjustment.

3. Multimodal system architecture

The overall architecture of the proposed multimodal ELVC system, which consists of a VC model and lip image feature extractor, is illustrated in Fig. 2. The VC model and lip-image feature extractor are described in detail in the following sections.

3.1. Voice conversion model

The VC model is implemented based on the CLDNN model proposed in [11]. CLDNN has been used in ELVC with satisfactory results in [12]. Using the MCC features as the model input, three independent CLDNN models were trained to predict the target speaker's MCC, aperiodicity (AP), and F0 and unvoiced/voiced (U/V) symbols. To reduce the experimental variability, we changed the input to a logarithmic Mel spectrogram (LMS) and trained a single CLDNN to convert the input LMS into the target LMS. To synthesize the waveform from the LMS, we used parallel WaveGAN [13] as the vocoder in our experiments.

3.2. Lip image feature extractor

The compressed visual features were obtained using a lip-image feature extractor. The feature extractor can be completely removed during the training phase. The lip-image feature extractors used in this study are described below.

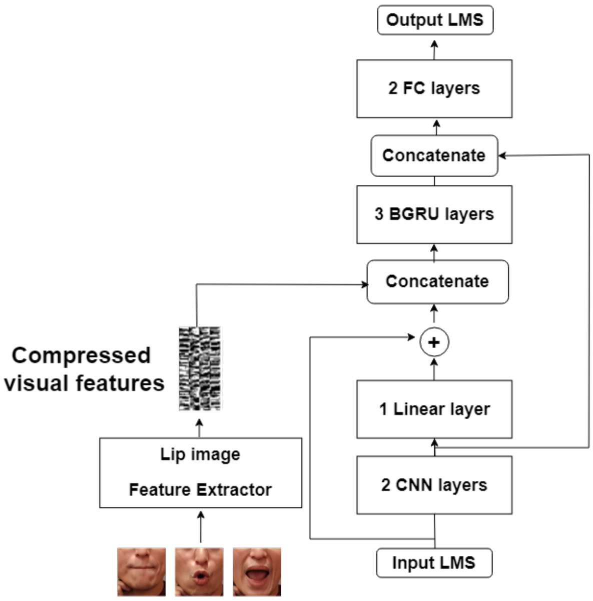

Figure 2: *Overall architecture of the proposed multimodal ELVC system.*

3.2.1. CNN encoder

The overall architecture of the CNN-based lip-image feature extractor includes an encoder and a decoder [14]. The encoder consists of three 2D convolutional layers and one linear layer. The decoder architecture is similar to that of the encoder; however, the convolutional layers are replaced by 2D transposed convolutional layers. The CNN-based model was trained in a self-supervised manner by reconstructing input lip images. Then, the lip images were processed by the pre-trained encoder to obtain latent representations of dimension 768, and these representations were used as the visual features for the multimodal VC model.

3.2.2. Vision Transformer

Vision Transformer (ViT) [15] is an image classification model with Transformer [16] as the backbone. We used a pre-trained ViT model[1] as a lip image feature extractor. The lip images were processed using the ViT model, and 768-dimensional representations of the last hidden layer were used as the visual features for the multimodal VC model.

3.2.3. AV-HuBERT

In recent years, many model architectures for self-supervised learning (SSL) have been developed, including AV-HuBERT [17], which inputs both acoustic features and lip images during training. AV-HuBERT enables the model to learn better features through the complementarity of information provided between the two modalities, leading to better results for downstream tasks that utilize lip information. We used a pre-trained AV-HuBERT model [2] as the lip image feature extractor.

[1] https://github.com/google-research/vision_transformer

[2] https://github.com/facebookresearch/av_hubert

Table 1: *Results using different DTW methods.*

Method	MCD (dB)
DTW-MCC	7.46
DTW-lip-landmark	7.21
DTW-WSOLA	**6.83**

When using AV-HuBERT as a feature extractor, it is possible to analyze whether the output of each layer of the transformer encoder is helpful for the ELVC task. Inspired by [18], a weighted-sum (WS) method was used for the output of each layer to combine the best-fit features. During VC model training, the AV-HuBERT model was fixed, but the weights were learned and updated. To balance the values of the output features of each layer, the output features were normalized and multiplied by the weight values. In our experiments, we compared the performance of the output features of the last hidden layer (LL) with that of the features using the WS method.

4. Experiments

This section presents the experimental setup, including the data and evaluation metrics, and the experimental results.

4.1. Datasets and evaluation metrics

We conducted experiments on the Mandarin parallel ELVC corpus, which was recorded by a doctor imitating a total laryngectomy patient using an electrolaryngeal device. The doctor read each sentence in the phonetically balanced TMHINT [19] dataset with and without the use of electrolarynx, while the audio and video were simultaneously recorded. We used 288 and 18 utterances as training and test data, respectively. All the speech utterances were sampled at a frequency of 16 kHz. Each speech waveform was converted into an 80-dimensional LMS with a window size of 512 points and frame shift of 160 points. The layer parameters of the CLDNN model architecture in Fig. 2 are similar to those in [12], except for the last fully connected layer. Since the input acoustic feature is an 80-dimensional LMS, the number of hidden units in the last fully connected layer is set to 80 to ensure that the input and output dimensions are consistent. The parallel WaveGAN used to synthesize the LMS back into a waveform was trained using the TMSV dataset [14].

The frame rate of the video was 50 FPS, and we downsampled the frame rate to 25 FPS, such that one image corresponded to four acoustic frames. Lip images were acquired by the lip-image extractor in [20] and converted into lip-image features using a lip-image feature extractor. In the experiments, the lip-image feature sequence was aligned with the acoustic frame sequence for model training. The batch size was 16, the learning rate was set to 0.0005, and the Adam optimizer was used.

Three objective metrics were used to evaluate the ELVC systems, including MCD, the syllable error rate (SER) measured by an ASR system trained on the MATBN dataset [21], and the estimated mean opinion score (MOS) of the pre-trained MOSA-Net [22][3]. The SER and predicted MOS values were 7.3% and 3.052 for NL speech and 82.3% and 1.556 for EL speech, respectively. These values were considered the upper and lower bounds of the performance of the ELVC models.

Table 2: *Results using different lip image feature extractors.*

Feature Extractor	MCD (dB)	SER (%)	MOS
None	6.83	73	1.965
CNN encoder	6.81	72.6	1.972
ViT	6.76	71.7	1.977
AV-HuBERT(LL)	6.45	69.2	2.073
AV-HuBERT(WS)	**6.32**	**66.7**	**2.077**

4.2. Experimental results

Experiments were conducted in two stages. First, the ELVC results obtained using different alignment methods were compared, and the best alignment method for use in subsequent experiments was determined. Subsequently, we compared the ELVC results obtained using different lip-image feature extractors.

4.2.1. Comparison of alignment methods

Table 1 lists the results obtained by applying different alignment methods to ELVC. The best-performing method was DTW-WSOLA, which stretched the target speech length so that more corresponding acoustic frames were aligned with the EL speech. While this could lead to distortion, it performed better than the DTW-lip-landmark method, which uses lip images for alignment. DTW-WSOLA cannot fully solve the alignment problem caused by the large difference in the acoustic characteristics of EL and NL speech; however, it is much better than other alignment methods. Therefore, DTW-WSOLA was used as the alignment method in subsequent experiments.

4.2.2. Comparison of visual feature extractors

Table 2 lists the results of applying different lip image feature extractors to ELVC. The visual features extracted by the CNN encoder and ViT showed no notable improvement in all three metrics. However, the visual features extracted by AV-HuBERT, both LL and WS, had a significant improvement in MCD, and the WS visual features were more helpful than the LL visual features. Compared with the CNN encoder and ViT, AV-HuBERT used both acoustic features and lip images as model input, which can extract meaningful features and provide more information to better train the conversion model.

4.2.3. Fine-tuning visual features

In our previous experiments, we concatenated the visual features extracted using a lip-image feature extractor with the acoustic features and trained a conversion model. In this experiment, we aimed to improve the conversion ability by fine-tuning (FT) the extracted visual features. We fed the extracted visual features to a unidirectional GRU layer and maintained the dimensionality of the features, enabling the model to learn dynamic information between images. The GRU module was trained together with the VC model. Comparing the results in Tables 2 and 3, it is found that the simple FT method can effectively improve the usability of the visual features extracted by all the lip image feature extractors.

4.2.4. Subjective evaluation

For subjective evaluation, an intelligibility test was conducted. During testing, one converted EL speech item was played for each question, and the subjects were asked to rate intelligibility

[3] https://github.com/dhimasryan/MOSA-Net-Cross-Domain

Table 3: *Results using different fine-tuning visual features.*

Method	MCD (dB)	SER (%)	MOS
CNN encoder+FT	6.62	70.3	2.055
ViT+FT	6.56	68	2.047
AV-HuBERT(WS)+FT	**6.28**	**62.7**	**2.113**

Table 4: *Subjective evaluation of intelligibility.*

System	Intelligibility
Audio-only CLDNN	2.586
AV-HuBERT(WS)	2.951
AV-HuBERT(WS)+FT	**3.218**

on a scale of 1–5, regardless of the speech quality. The evaluation criteria are as follows: 5 means that every word in the sentence can be understood; 4 means that a few words in the sentence cannot be understood, but it does not affect the understanding of the sentence; 3 means that nearly half of the words in the sentence can be understood, and the content of the sentence can be roughly judged; 2 means that only a few words in the sentence can be understood, but not the whole sentence; and 1 means that the sentence cannot be understood at all.

Table 4 presents the subjective evaluation results of three ELVC systems. The listening test was conducted on 12 untrained but experienced normal hearing subjects. Among them, 8 were male, and 4 were female. The average age of these 12 subjects was 24 years old. For each test sample, participants were not informed which ELVC system was used to generate it. We selected 18 speech utterances converted from each ELVC system to conduct the subjective test. Both audio-visual systems (AV-HuBERT(WS) and AV-HuBERT(WS)+FT) using the AV-HuBERT features achieved higher intelligibility than the Audio-only CLDNN system; and the system with fine-tuned visual features (AV-HuBERT(WS)+FT) achieved the best intelligibility. The subjective evaluation results confirm that multimodal learning can help with the ELVC task.

5. Conclusions and future work

In this study, we proposed a multimodal ELVC approach. The experimental results show that the quality and intelligibility of converted EL speech can be improved. The features of the SSL models that have been frequently used in recent years also play a pivotal role in our model. In future research, we will attempt to fine-tune the pre-trained AV-HuBERT model to generate more useful features for ELVC. We will also leverage the features of AV-HuBERT to help align EL and NL speech for better ground truth when training the conversion model.

6. References

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

[2] 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.

[3] Y.-S. Liou, W.-C. Huang, M.-C. Yen, S.-W. Tsai, Y.-H. Peng, T. Toda, Y. Tsao, and H.-M. Wang, "Time alignment using lip images for frame-based electrolaryngeal voice conversion," in *Proc. APSIPA ASC*, 2021.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[5] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[6] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. CVPR*, 2021.

[7] J. Zhou, Y. Hu, H. Lian, H. Wang, L. Tao, and H. K. Kwan, "Multimodal voice conversion under adverse environment using a deep convolutional neural network," *IEEE Access*, vol. 7, pp. 170 878–170 887, 2019.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[9] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, 1993.

[11] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, 2015.

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

[13] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020.

[14] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, "Lite audio-visual speech enhancement," *Proc. Interspeech 2020*, pp. 1131–1135, 2020.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[17] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=Z1Qlm11uOM

[18] K.-H. Hung, S.-w. Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting self-supervised embeddings for speech enhancement," in *Proc. Interspeech*, 2022.

[19] M.-W. Huang, "Development of Taiwan Mandarin hearing in noise test," *Master's thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[20] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. ICASSP*, 2020.

[21] 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, 2005.

[22] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.