applied sciences

MDPI

Article

Mandarin Electro-Laryngeal Speech Enhancement Using Cycle-Consistent Generative Adversarial Networks

Zhaopeng Qian [1,*], Kejing Xiao [2] and Chongchong Yu [1]

[1] School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
[2] School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China
* Correspondence: qianzhaopeng@btbu.edu.cn

Abstract: Electro-laryngeal (EL) speech has poor intelligibility and naturalness, which hampers the popular use of the electro-larynx. Voice conversion (VC) can enhance EL speech. However, if the EL speech to be enhanced is with complicated tone variation rules in Mandarin, the enhancement will be less effective. This is because the source speech (Mandarin EL speech) and the target speech (normal speech) are not strictly parallel. We propose using cycle-consistent generative adversarial networks (CycleGAN, a parallel-free VC framework) to enhance continuous Mandarin EL speech, which can solve the above problem. In the proposed framework, the generator is designed based on the neural networks of a 2D-Conformer-1D-Transformer-2D-Conformer. Then, we used Mel-Spectrogram instead of traditional acoustic features (fundamental frequency, Mel-Cepstrum parameters and aperiodicity parameters). At last, we converted the enhanced Mel-Spectrogram into waveform signals using WaveNet. We undertook both subjective and objective tests to evaluate the proposed approach. Compared with traditional approaches to enhance continuous Mandarin EL speech with variable tone (the average tone accuracy being 71.59% and average word error rate being 10.85%), our framework increases the average tone accuracy by 12.12% and reduces the average errors of word perception by 9.15%. Compared with the approaches towards continuous Mandarin EL speech with fixed tone (the average tone accuracy being 29.89% and the average word error rate being 10.74%), our framework increases the average tone accuracy by 42.38% and reduces the average errors of word perception by 8.59%. Our proposed framework can effectively address the problem that the source and target speech are not strictly parallel. The intelligibility and naturalness of Mandarin EL speech have been further improved.

Keywords: Mandarin electro-laryngeal speech; CycleGAN; WaveNet

check for updates

Citation: Qian, Z.; Xiao, K.; Yu, C. Mandarin Electro-Laryngeal Speech Enhancement Using Cycle-Consistent Generative Adversarial Networks. *Appl. Sci.* **2023**, *13*, 537. https://doi.org/10.3390/app13010537

Academic Editor: Agnese Sbrollini

Received: 13 November 2022
Revised: 18 December 2022
Accepted: 26 December 2022
Published: 30 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

An electro-larynx is the most commonly used assistive device for laryngectomees' speech recovery. An electro-larynx is easy to understand and operate and suitable for people of all ages and health conditions. However, due to flattened fundamental frequency (F0) and mechanical and radiation noise, electro-laryngeal (EL) speech is insufficiently intelligible and natural, especially for tonal languages such as Mandarin Chinese [1–3]. Due to the inflexible variation of F0, Mandarin EL speech sounds mechanical. That is why EL speech is less natural. Large radiation noises result in the poor quality of continuous EL speech. Meanwhile, laryngectomees lack airflow and air pressure, which hampers their pronouncing process. As a result, EL speech is poor in intelligibility. Mandarin Chinese is a tonal language whose intelligibility is greatly influenced by tone variation. Therefore, the inflexible variation of F0 also leads to the poor intelligibility and rhythm of Mandarin EL speech. Although using an F0-controllable electro-larynx, a speaker can pronounce the four tones of Mandarin Chinese speech, the inflexible F0 variation still seriously affects the naturalness of Mandarin EL speech.

To improve the intelligibility and naturalness of EL speech, researchers have tried multiple approaches. During the early stage, researchers mainly tried to improve the electro-larynx, exploring an electro-larynx controlled by biophysical signals. For instance, researches tried to improve the flexibility of F0 variation through applying vocal air pressure signals [4], neck strap muscle electromyographically activity signals [5,6], finger pressure signals [7] or finger sliding signals [8,9]. Meanwhile, radiation noise is still an important factor that influences the intelligibility of EL speech. In this regard, researchers tried to reduce the radiation noise through various approaches including noise cancelling based on adaptive filter [10–12] and spectral subtraction [13–15]. The enhancement of Mandarin EL speech is different from the enhancement of normal speech with background noise [16–18], for that the Mandarin EL speech has the drawbacks of radiation noise and tone errors [1–3], which make it more difficult to deal with. Signal processing methods can effectively reduce the radiation noise of Mandarin EL speech [10–15]. However, traditional approaches of improving the F0-controllable electro-larynx (this approach cannot reduce the radiation noise) or reducing radiation noise through signal processing (this approach cannot improve the flexibility of F0 variation) still have deficiencies in making EL speech more intelligible and natural. Recently, researchers applied frame-by-frame voice conversion (VC, parallel-dependent VC) to improve the intelligibility and naturalness of the EL speech [19–22]. However, this parallel-dependent VC approach is still deficient in enhancing Mandarin EL speech. This deficiency is due to the complicated tone-variation rules of Mandarin EL speech and is more prominent when the Mandarin EL speech is pronounced using a finger-pressed electro-larynx at fixed tone [23]. Meanwhile, the pipeline approach obtaining semantic content from Mandarin EL speech through automatic speech recognition (ASR) [24] can also help improve the intelligibility and naturalness of Mandarin EL speech.

Although the pipeline approach [25] effectively improves the intelligibility and naturalness of Mandarin EL speech, the approach includes multiple steps that may lead to hard-to-improve deficiencies such as high error rate of syllables or low tone accuracy. Therefore, both the parallel-dependent VC and the pipeline approach have their own drawbacks in dealing with Mandarin EL speech. Kaneko et al. [26] proposed parallel-data-free VC based on cycle-consistent generative adversarial networks (CycleGAN). This CycleGAN-based VC does not require parallel source-target content but has poor performance in dealing with the Mel-Spectrogram [27]. The CycleGAN-VC3 (VC3 in this paper) proposed by Kaneko et al. [27] incorporates a 2-1-2 dimension (2D-1D-2D) generator based on time-frequency adaptive normalization (TFAN), an improved version of CycleGAN-VC2 [28]. However, VC3 is still weak in processing Mandarin EL speech with complicated tone variations. To solve the problem, we design neural networks with 2D-conformer-1D-transformer-2D-transformer (2D-Con-1D-Tran-2D-Con) architecture as the generator for our CycleGAN, which is used to enhance Mandarin EL speech. The transformer proposed by Vaswani, et al. [29] and conformer proposed by Gulati, et al. [30] achieve good performance in processing sequential data, and in particular, the transformer can capture long-distance dependencies. The conformer and transformer both have been successful in speech recognition and separation. The conformer in 2D-Con-1D-Tran-2D-Con architecture is used to process the input Mel-Spectrogram and the convolutional layer can capture both frequency and temporal information. The transformer in the proposed architecture can capture the dependencies of the input sequence. The proposed architecture is used to take advantage of both contextual semantic information and frequency information.

WaveNet proposed by Oord et al. [31] is an advanced speech synthesis technology. As a neural vocoder, WaveNet can synthesize high-quality speech based on low-dimension acoustic features such as the Mel-Spectrogram. WaveNet can effectively make up for the defects of traditional vocoders. In this paper, WaveNet is used to directly convert Mel-Spectrogram parameters into waveform signals.

To our knowledge, we are the first to use CycleGAN-based VC to enhance Mandarin EL speech. The main contributions of this paper are as follows: The 2D-Con-1D-Tran-2D-Con neural networks are designed as the generator in CycleGAN to process the input, and

the 2D-Conformer is designed as the discriminator in CycleGAN to discriminate whether the predictions approximate the outputs. This architecture is used to capture contextual semantic information to enhance the tone variation of continuous Mandarin EL speech. The CycleGAN-based VC is applied to address the problem that the source speech (Mandarin EL speech) and the target speech (normal speech) are not strictly parallel. This is because the Mandarin EL speech has some tone errors. Objective and subjective evaluations are designed to test the proposed approach. In particular, average tone accuracy and word perception error rate are used in subjective evaluation to explain the effect of enhancement.

The rest of this paper is organized as follows. Section 2 introduces the proposed methods to further enhance Mandarin EL speech. Section 3 introduces the experimental results of evaluation for enhancement. Section 4 discusses the advantages and limitations of the proposed methods. Section 5 give the conclusions of our research.

2. Methods

CycleGAN-based VC has two stages: the training stage and the converting stage, as shown in Figure 1.

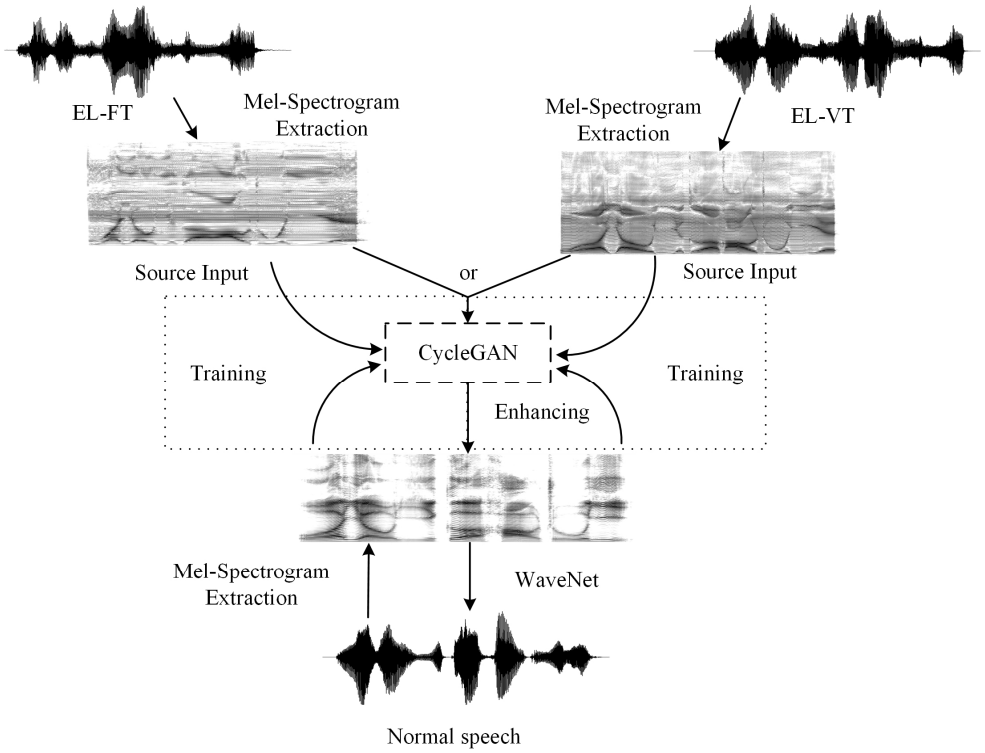


Figure 1. Procedure for enhancing Mandarin EL speech using CycleGAN and WaveNet.

During the training stage, the Mel-Spectrogram parameters are used to train the CycleGAN. In addition, the target Mel-Spectrogram parameters and waveform signals are used to train the WaveNet. During the converting stage, Mel-Spectrogram parameters of Mandarin EL speech are converted to the Mel-Spectrogram parameters of ordinary speech. At last, WaveNet synthesizes ordinary speech based on the Mel-Spectrogram parameters. The whole procedure of CycleGAN-based VC for enhancing Mandarin EL speech is shown in Figure 1.

In Figure 1, EL-FT represents the Mandarin EL speech with fixed tone (Tone 1) pronounced using a finger-pressed electro-larynx. EL-VT represents the Mandarin EL speech with variable tone (some tone errors exist in the continuous Mandarin EL speech), pronounced using a finger-touch pad electro-larynx.

2.1. Architecture of CycleGAN for Enhancing the Mandarin EL Speech

According to Kaneko, et al. [27], CycleGAN-VC and CycleGAN-VC2 were originally designed for Mel-cepstrum conversion. VC3 uses 2D-1D-2D TFAN-based architecture as its generator, and patch-consistent generative adversarial networks (PatchGAN) as its discriminator [27]. Although this design achieves good performance in dealing with Mel-spectrogram parameters, enhancing Mandarin EL speech still requires more semantic information, especially when the tone variation rules are complicated. In this paper, we use the conformer and transformer bolocks to replace the TFAN networks. The architecture of our proposed CycleGAN is shown in Figure 2.

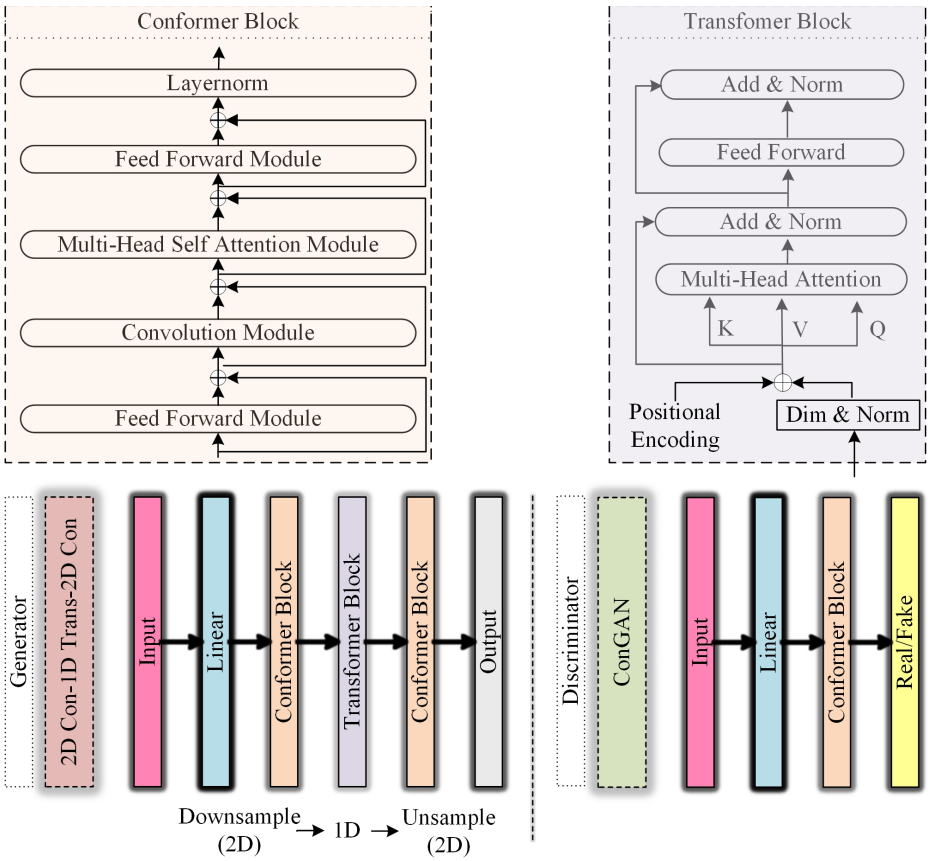


Figure 2. The architecture of CycleGAN including the 2D-Con-1D-Tran-2D-Con-based generator and 2D-Conformer-based discriminator.

In Figure 2, the conformer and transformer are both blocks in the CycleGAN. The architectures of the two blocks are shown at the top part of Figure 2. The left bottom part of Figure 2 shows the architecture of the generator of CycleGAN. The right bottom part of Figure 2 shows the architecture of the discriminator of CycleGAN. "Dim" means dimension of input. "Norm" means layer normalization of input, which is equivalent to the Layernorm. "Add" denotes the residual block, where in this block data performs the operation of residual connection. The 2D-Con-1D-Tran-2D-Con neural-based generator and the 2D-Conformer-based discriminator of our CycleGAN are designed to capture the long dependences of Mandarin EL speech.

2.1.1. Generator Architecture

The proposed CycleGAN uses 2D-Con-1D-Tran-2D-Con as the generator to capture temporal and frequency information. This design follows the framework of VC3, whose generator is designed based on 2D-1D-2D TFAN [27].

Particularly, the network is composed of downsampling, residual, and upsampling blocks. In this proposed CycleGAN framework, the 2D-Conformer in the downsampling and upsampling blocks is used to process the Mel-Spectrogram parameters; the 1D-Transformer in the residual blocks is used to process the transformed Mel-Spectrogram parameters. The former is used to extract the time-frequency structure while preserving the original structure. The latter is used to capture the contextual semantic information.

The conformer block contains two feed-forward Networks modules sandwiching the multi-headed self-attention module and the convolutional module, as shown in Equations (1)–(4).

$$\widetilde{x}_i = x_i + \frac{1}{2}\mathrm{FFN}(x_i) \tag{1}$$

$$x\prime_i = \widetilde{x}_i + \mathrm{MHSA}(\widetilde{x}_i) \tag{2}$$

$$x''_i = x\prime_i + \mathrm{Conv}(x\prime_i) \tag{3}$$

$$y_i = \mathrm{Layernorm}(x''_i) \tag{4}$$

where x_i is the ith input; y_i is the ith output; FFN refers to the feed-forward networks module; MHSA refers to the multi-headed self-attention module, and Conv refers to the Convolutional module as shown in Figure 2.

The Transformer block contains the multi-headed attention module (the first sublayer) and the feed-forward networks (the second sublayer), as shown in Equations (5)–(7). Surrounding every two sublayers, one keep-residual connection layer and one normalization layer are used.

$$h^t = f\Big(\frac{g}{\sigma^t} \odot (a^t - \mu^t) + b\Big) \tag{5}$$

$$\mu^t = \frac{1}{H}\sum_{i=1}^{H} a_i^t \tag{6}$$

$$\sigma^t = \sqrt{\frac{1}{H}\sum_{i=1}^{t}(a_i^t - \mu^t)^2} \tag{7}$$

where b and g are defined as the bias and gate parameters with the same dimension of h^t, respectively. In Equation (5), $a^t = W_{hh}h^{t-1} + W_{xh}x^t$, where W_{hh} represents the weights of the recurrent hidden layer, and W_{xh} represents the weights of line that from the input layer to the hidden layer. Please note that, the outputs of every sublayer are $\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x))$, where $\mathrm{Sublayer}(x)$ is achieved by every sublayer themselves.

2.1.2. Discriminator Architecture

Differently from in VC3, we use a 2D-Conformer instead of 2D convolutional neural networks (CNN) as the discriminator to focus on the contextual information in the spectral texture. This structure not only captures the temporal and frequency information of the spectrum, but also draws upon the semantic information of the spectrum.

The conformer block (described in Section 2.1.1) is composed of four modules stacked together: feed-forward, a self-attention module, a convolution module, and a second feed-forward module in the end. In the proposed framework a 2D-Conformer directly processes the Mel-Spectrogram parameters, as shown in Figure 2.

2.1.3. Training Objectives

During the training stage, the Mel-spectrogram parameters are used to train the transformation function. The training process is shown in Figure 3.

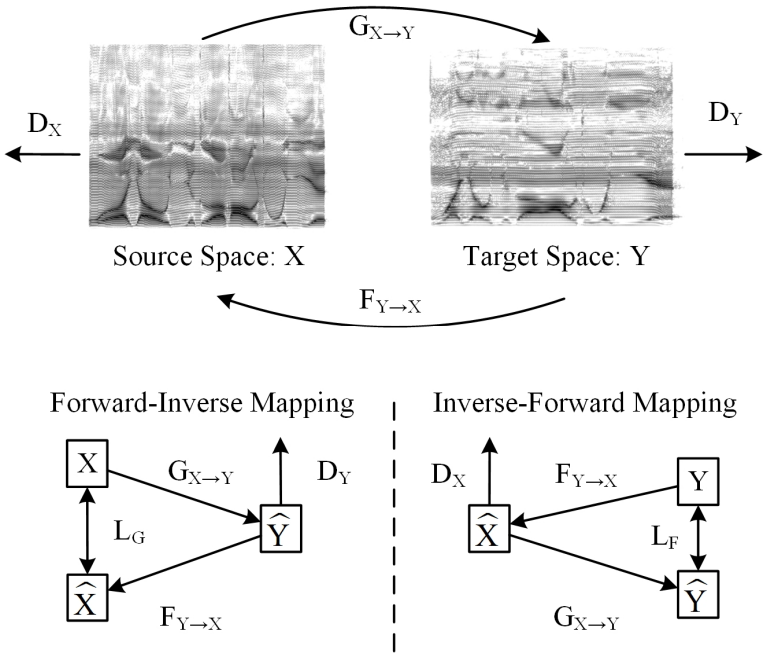


Figure 3. The data flow between the source space and the target space of CycleGAN.

The top part of Figure 3 contains the acoustic features (plot using praat) of source speech and target speech. The top part of Figure 3 shows the data flow mapping from the source to the target and the data flow mapping from the target to the source. The left bottom part of Figure 3 shows how the discriminator D_Y discriminates the generated $\hat{Y}$ using the forward generator $G_{X\to Y}$ of CycleGAN. The right bottom part of Figure 3 shows how the discriminator D_X discriminates the generated $\hat{X}$ using the inverse generator $F_{Y\to X}$ of CycleGAN. In CycleGAN shown as Figure 3, the mapping function $G_{X\to Y}$ represents the forward generator; $F_{Y\to X}$ represents the inverse generator; D_X represents the forward discriminator; D_Y represents the inverse discriminator; $\hat{X}$ represents the predicted output of Y using F(Y); and $\hat{Y}$ represents the predicted output of X using G(X). In particular, D_X is used to discriminate the X and F(Y); D_Y is used to discriminate the Y and G(X). Consistence loss and inverse consistence loss are both used to evaluate this model. The consistence loss is used to match the generated and target acoustic features. The inverse consistence loss is used to avoid the contradiction between mapping functions G and F.

The process of computing losses takes reference from Zhu et al. [32]. The adversarial loss $L_G(G, D_Y, X, Y)$ is computed as in Equation (8), and the inverse adversarial loss $L_F(F, D_X, Y, X)$ is computed as in Equation (9).

$$L_G(G, D_Y, X, Y) = E_{P_{Data}(Y)}(\ln D_Y) + E_{P_{Data}(X)}(\ln(1 - D_Y(G(X))) \quad (8)$$

where $P_{Data}(X)$ and $P_{Data}(Y)$ represent the distribution probability of acoustic features for source and target speakers respectively; E represents the expected value of the acoustic feature distribution. Log (ln) represents the logarithmic calculation. The smaller the loss is, the narrower the gap between the acoustic features of the converted and target speech.

$$L_F(F, D_X, Y, X) = E_{P_{Data}(X)}(\ln D_X) + E_{P_{Data}(Y)}(\ln(1 - D_X(F(Y))) \quad (9)$$

The final adversarial loss of CycleGAN equals to the mean of Equation (8) plus Equation (9). The final consistence loss $L_{cyc_con}(G, F)$ is computed as in Equation (10).

$$L_{cyc_{con}}(G, F) = E_{P_{Data}(X)}(\|F(G(X)) - X\|_1) + E_{P_{Data}(Y)}\|(G(F(Y)) - Y\|_1) \quad (10)$$

where $\| \ \|_1$ represents the 1-norm. The loss of the whole model L_{full} is computed as in Equation (11).

$$L_{full} = L_G(G, D_Y, X, Y) + L_F(F, D_X, Y, X) + \lambda \cdot L_{cyc_{con}}(G, F) \tag{11}$$

where λ represents the control parameter used to balance the weights of the forward and inverse training process.

We take reference from VC3 [27] to design the generator and discriminator of the proposed CycleGAN. We use the Adam algorithm for gradient descent [33,34] and set the parameters β_1 and β_2 as 0.9 and 0.99 respectively.

2.2. Neural Vocoder Based on WaveNet

WaveNet is a new generative model operating directly on raw audio waveforms. WaveNet assumes that samples are interrelated. The input of WaveNet is $X = \{x_1, x_2, \ldots, x_{t-1}, x_t\}$, where t means the number of frames. The joint probability of a waveform $W = \{w_1, \ldots, w_T\}$ is factorized as a product of conditional probabilities, which is shown in Equation (12).

$$p(X) = \prod_{t=1}^{T} p(x_t | x_1, x_2, \ldots, x_{t-1}) \tag{12}$$

Therefore, the sample at time t depends on all of the previous samples. The details of the dilated causal convolutions framework can be found in the researches of Oord et al [31] in 2016. We use SoftMax distribution to compute the outputs of the WaveNet. To trace the probability per time step, we apply a μ-law companding transformation to the data and then quantify the data into 256 specific values. The companding transformation is shown in Equation (13).

$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu|x_t|)}{\ln(1 + \mu)} \tag{13}$$

where $\text{sign}(*)$ denotes the "signum function", $-1 < x_t < 1$ and $\mu = 255$. The gated activation unit is defined in Equation (14), which is similar to the gated PixelCNN [35].

$$z = \tanh(W_{f,k} * x) \odot \sigma\left(W_{g,k} * x\right) \tag{14}$$

where $*$ denotes the convolution operator; $\odot$ denotes the element-wise multiplication operator; $\sigma(\cdot)$ denotes the sigmoid function; k is the layer index; f and g denote the filter and gate respectively; and W is a linear convolution filter.

2.3. Experiment

2.3.1. Experiment Setup Conditions

In this paper, we undertake both objective and subjective evaluation of the convolutional, long short term memory, fully connected deep neural networks (CLDNN)-based [19,23], VC3-based [27] and our CycleGAN-based approaches. In the evaluations, we use traditional features (including F0, Aperiodicity, Mel-Cepstrum parameters) and Mel-Spectrogram parameters, respectively as acoustic features. In addition, the hyper-parameters are adjusted manually according to our engineering experience. The main parameters of the proposed CycleGAN framework are shown in Table 1.

The hyper-parameters in Table 1 show the main parameters of the proposed framework. In addition, the models were implemented using PyTorch (version 1.11.0). All of the experiments are implemented via python program using PyCharm (version professional 2022.1). And the models are trained on the GPU server with 4 pieces of NVIDIA Tesla T4 GPU cards.

Table 1. Hyper Parameters of proposed CycleGAN-WaveNet.

Main Parameters Name	Main Parameters Value
Attention heads of 2D-Conformer in generator	4
Attention dimension of 2D-Conformer in generator	256
Convolution kernel size of 2D-Conformer in generator	32 × 32
Attention heads of 1D-Transformer in generator	4
Attention dimension of 1D-Transformer in generator	256
Attention heads of 2D-Conformer in discriminator	4
Attention dimension of 2D-Conformer in discriminator	256
Convolution kernel size of 2D-Conformer in discriminator	32 × 32
Kernel size of WaveNet	3
Dropout of WaveNet	0.05
Residual channels of WaveNet	512
Gated channels of WaveNet	512
Skip-out channels of WaveNet	256
Batch size	16
Epochs	100
Intial learning rate	0.001

2.3.2. Listeners

We enrolled 10 Chinese listeners (5 men and 5 women) with an average age of 25 years (23 to 28 years). All the listeners had good listening ability and were native Mandarin Chinese speakers. Listeners are asked to write down the pinyin with tones according to the content they heard. Additionally, the listeners were required to give a mean opinion score (MOS) on the intelligibility and naturalness of the audio files. All of the listeners finished the test using headphones in the listening test room without being disturbed.

2.3.3. Data Preparation

In this paper, we only used 1000 utterances of Chinese text from the THCHS-30 database [36] and 250 utterances of daily Chinese as the speech materials to collect 1250 pairs of continuous Mandarin EL and normal speech. The 1000 pairs were used to train the model and the remaining 250 pairs were used for statistical tests (including objective and subjective evaluations). In the silent recording room, 1250 utterances of continuous Mandarin EL speech with fixed tone (EL-FT) were recorded by a speaker using a finger-pressed electro-larynx; 1250 utterances of continuous Mandarin EL speech with variable tone (EL-VT) were recorded by a speaker using a finger-sliding Electro-Larynx; and 1250 utterances of continuous Mandarin normal speech are recorded as the target speech. The THCHS-30 database is an open access database offered by Wang and Zhang [36]. This database of Mandarin Chinese speech covers about 71.5% of bi-phones and 14.3% of tri-phones. On average, one utterance for training is 35 s, and one utterance for testing is 5 s. This is because the CycleGAN model requires a large amount of data to train. However, daily Chinese speech does not feature sentences that long. All the audio files were recorded at the sampling rate of 44,100 Hz through Adobe Audition software and a Roland professional sound collection device. However, to conveniently extract the acoustic features from audio files, we transformed all the audio files at the sampling rate of 16,000 Hz during the training (and testing) stage. Some example files can be find in the Supplementary Materials.

2.3.4. Objective Evaluation Conditions

During the objective test, we used Mel-Cepstral distortion (MCD) and the root mean square error (RMSE) of CodeAP (CodeAP RMSE, CodeAP is an aperiodicity parameter extracted using WORLD [37]) to evaluate the spectrum features, we also used the correlation

coefficient of F0 (F0 CC) and RMSE of logarithmic F0 (log F0 RMSE) to evaluate the F0 pattern of enhanced speech. The MCD can be calculated as in Equation (15).

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(c_d - c\prime_d)^2} \quad (15)$$

where c_d and $c\prime_d$ are the dth coefficients of the target and predicted Mel-cepstra respectively. The coefficients' dimension was set as 24 during the test.

2.3.5. Subjective Evaluation Conditions

Listeners gave MOS on the intelligibility and naturalness of the following speech:

EL-FT: EL-FT represents Mandarin EL speech with fixed tone (Tone 1), pronounced using a finger-pressed electro-larynx [8].

EL-VT: EL-VT represents Mandarin EL speech with variable tone (some tone errors exist in continuous Mandarin EL speech), pronounced using a finger-touch pad electro-larynx [8].

CLDNN-FT: CLDNN-FT represents enhanced Mandarin EL speech with fixed tone using CLDNN [23].

CLDNN-VT: CLDNN-VT represents enhanced Mandarin EL speech with variable tone using CLDNN [23].

VC3-FT: VC3-FT represents enhanced Mandarin EL speech with fixed tone using CycleGAN-VC3 [27].

VC3-VT: VC3-VT represents enhanced Mandarin EL speech with variable tone using CycleGAN-VC3 [27].

2C1T2C-FT: 2C1T2C-FT represents enhanced Mandarin EL speech with fixed tone using our proposed method.

2C1T2C-VT: 2C1T2C-VT represents enhanced Mandarin EL speech with variable tone using our proposed method.

Please note that the sequence of the audio is random, and listeners did not know the content of the audio beforehand. Moreover, the subjective evaluation was on a standard MOS scale where 5 stands for excellent; 4 for good; 3 for common; 2 for poor and 1 for bad.

3. Results

The results for objective evaluation in Section 3.1 include F0 pattern evaluation and spectrum augmentation evaluation. The results of U/V analysis, F0 CC and log F0 RMSE are used to evaluate the F0 pattern. MCD and CodeAP RMSE are used to evaluate the spectrum augmentation. The results for subjective evaluation in Section 3.2 include the analysis of tone accuracy, influence of semantic information for tone accuracy, word perception error rate, MOS of intelligibility and naturalness.

3.1. Objective Evaluation

The results of objective evaluation listed in Tables 2–4 show the effect of different approaches in enhancing Mandarin EL speech. The U/V confusion matrixes in Table 2, the log F0 RMSE and F0 CC in Table 3 show the F0-based enhancement; the MCD and CodeAP RMSE in Table 4 show spectrum enhancement.

In Table 2, U_{tar} represents the unvoiced part of target speech; U_{est} represents the unvoiced part of converted speech; V_{tar} represents the voiced part of target speech; and V_{est} represents the voiced part of converted speech. The results of U/V analysis were calculated according to the testing data (referring to the work [23]). Table 2 leads to three findings. Firstly, the difference between traditional features and Mel-Spectrogram parameters is very small. Secondly, different approaches are close in the accuracy of voiced segments. Thirdly, our approach has the highest accuracy of unvoiced segments and the accuracy of enhanced speech with variable tone is higher than that of enhanced speech with fixed tone.

Table 2. U/V analysis for continuous Mandarin EL speech under different approaches.

Speech Types	U/V Type	Traditional Features		Mel-Spectrogram	
		U_{est}	V_{est}	U_{est}	V_{est}
CLDNN-FT	U_{tar}	83.42%	16.58%	84.35%	15.65%
	V_{tar}	2.45%	97.55%	2.88%	97.12%
VC3-FT	U_{tar}	90.21%	9.79%	92.13%	7.87%
	V_{tar}	2.32%	97.68%	2.05%	97.95%
2C1T2C-FT	U_{tar}	90.58%	9.42%	93.1%	6.90%
	V_{tar}	1.68%	98.32%	1.75%	98.25%
CLDNN-VT	U_{tar}	85.44%	14.56%	84.89%	15.11%
	V_{tar}	3.89%	96.11%	3.78%	96.22%
VC3-VT	U_{tar}	92.43%	7.57%	93.86%	6.14%
	V_{tar}	2.22%	97.78%	1.33%	98.67%
2C1T2C-VT	U_{tar}	92.15%	7.85%	95.12%	4.88%
	V_{tar}	1.17%	98.83%	1.18%	98.82%

Table 3. Objective evaluations of converted F0 pattern.

Speech Types	Traditional Features		Mel-Spectrogram	
	log F0 RMSE	F0 CC	log F0 RMSE	F0 CC
CLDNN-FT	0.1321	0.6783	0.1300	0.7247
VC3-FT	0.1214	0.7223	0.1205	0.7218
2C1T2C-FT	0.1131	0.6287	0.1121	0.6025
CLDNN-VT	0.1633	0.7472	0.1131	0.7794
VC3-VT	0.1037	0.7122	0.1027	0.7081
2C1T2C-VT	0.1011	0.6275	0.1005	0.5934

Table 4. Objective evaluations of spectrum features.

Speech Types	Traditional Features		Mel-Spectrogram	
	MCD (dB)	CodeAP RMSE (dB)	MCD (dB)	CodeAP RMSE (dB)
CLDNN-FT	5.5294	4.3523	5.6625	4.8072
VC3-FT	4.9205	4.1055	4.7032	3.7805
2C1T2C-FT	4.8135	3.9522	4.6305	3.7625
CLDNN-VT	5.3215	4.5778	5.5441	4.7931
VC3-VT	4.7720	3.9815	4.5625	4.3505
2C1T2C-VT	4.7026	3.3765	4.5211	3.5922

In Table 3, log F0 RMSE and F0 CC are obtained based on the comparison of the converted and target F0. If log F0 RMSE is low and F0 CC is high, the enhancement of F0 pattern would be effective. Our proposed approach performs better than the two baseline (CLDNN-based and VC3-based) approaches. In addition, the F0 pattern enhancement of speech with variable tone is more effective than that of speech with fixed tone. This is because the difference between EL-VT and normal speech is much less than the difference between EL-FT and normal speech.

Table 4 shows the results of spectrum enhancement. For both traditional features and Mel-Spectrogram, our proposed approach performs better than baseline approaches. For CycleGAN-based approaches (VC3 and our approach), Mel-Spectrogram parameters lead to better performance than traditional features. This is because CycleGAN-based approaches can effectively transform one Mel-spectrogram into another as a whole, not frame-by-frame. In particular, our proposed approach can capture more information, such as semantic contextual information and frequency information. However, if the Mel-Spectrogram is

merely processed using CLDNN-based frame-by-frame VC, the enhancement through Mel-Spectrogram parameters would be less effective than using the traditional features.

In addition, waveform, pitch and spectrogram analysis were also applied to evaluate the enhancement effect of Mandarin EL speech. The results can be found in Figures 4–6.

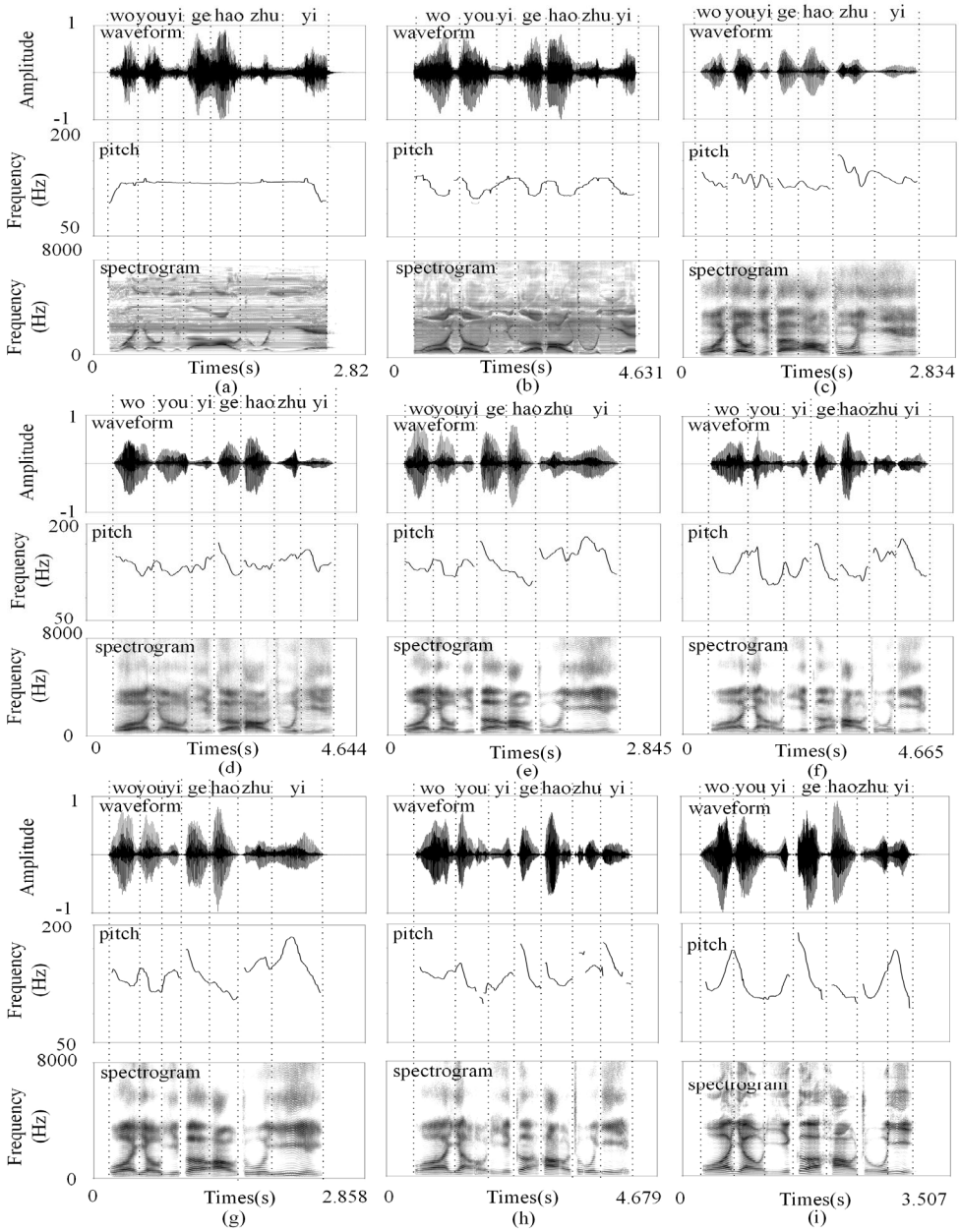


Figure 4. The acoustic features analysis of (**a**) EL-FT speech; (**b**) EL-VT speech; (**c**) CLDNN-FT speech; (**d**) CLDNN-VT speech; (**e**) VC3-FT speech; (**f**) VC3-VT speech; (**g**) 2C1T2C-FT speech; (**h**) 2C1T2C-VT speech; and (**i**) normal speech.

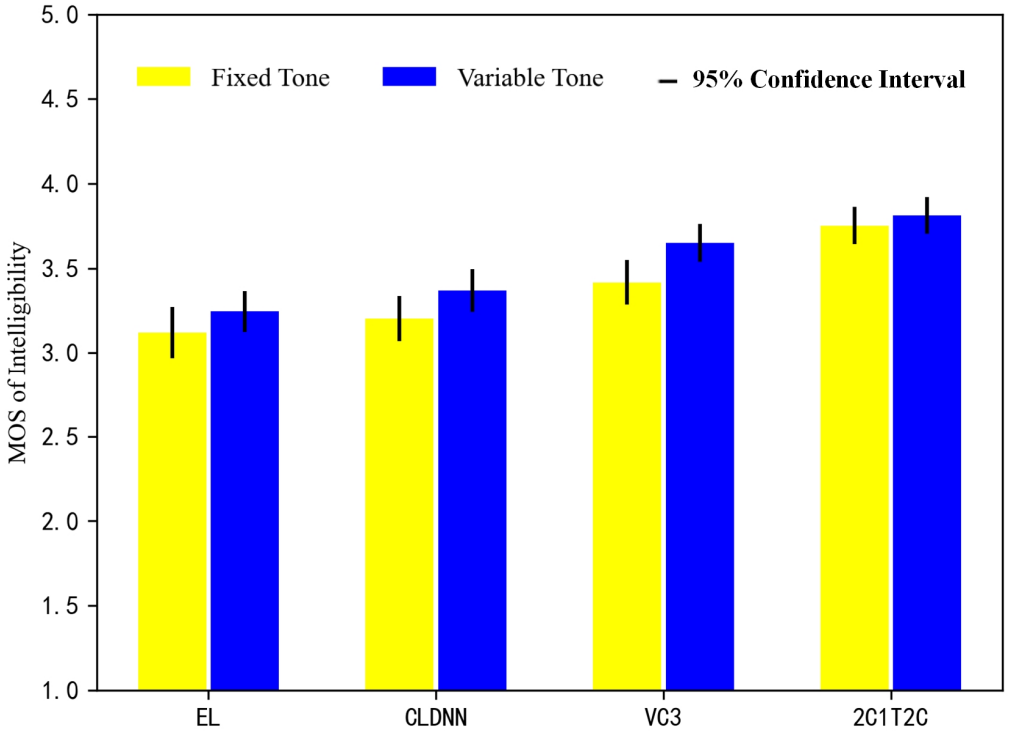


Figure 5. MOS for intelligibility with 95% confidence intervals.

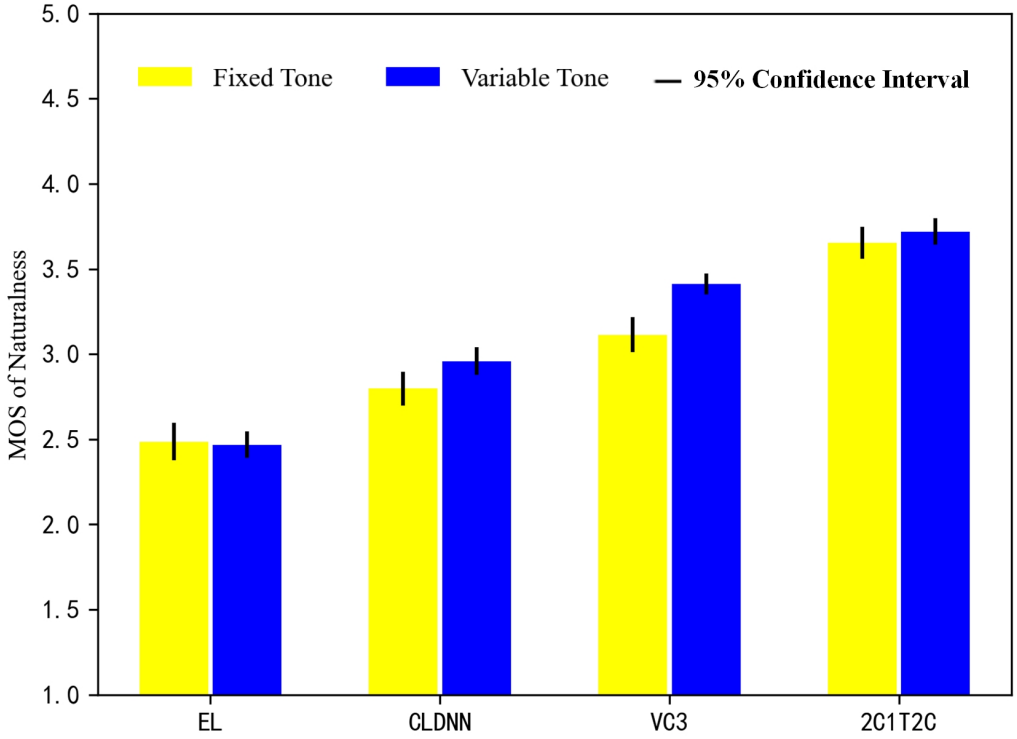


Figure 6. MOS for naturalness with 95% confidence intervals.

Figure 4a shows the acoustic features analysis of Mandarin EL speech with fixed tone (EL-FT); Figure 4b shows the acoustic features analysis of Mandarin EL speech with variable tone (EL-VT); Figure 4c shows the acoustic features analysis of enhanced Mandarin EL speech with fixed tone using CLDNN (CLDNN-FT); Figure 4d shows the acoustic features analysis of CLDNN-VT; Figure 4e shows the acoustic features analysis of enhanced Mandarin EL speech with fixed tone using VC3 (VC3-FT); Figure 4f shows the acoustic

features analysis of VC3-VT; Figure 4g shows the acoustic features analysis of enhanced Mandarin EL speech with fixed tone by our proposed approach (2C1T2C-FT); Figure 4h shows the acoustic features analysis of 2C1T2C-VT; and Figure 4i shows the acoustic features analysis of normal speech.

In Figure 4, the top part of every subfigure shows waveform signals; the middle part shows pitch analysis; and the bottom part shows the narrow-band spectrogram analysis. For waveform analysis, based on the EL-FT and EL-VT shown in Figure 4a,b, the radiation noise is all effectively cancelled in enhanced speech. For pitch analysis, the pitch lines of 2C1T2C-FT and 2C1T2C-VT are closest to that of the normal speech. For example, from the comparison of the Figure 4c,e,g, the utterance of VC3-FT has three correct tones ("ge4", "zhu2" and "yi4") and four tones of 2C1T2C-FT are correct ("yi2", "ge4", "zhu2" and "yi4"). However, the utterance of CLDNN-FT has only one correct tone ("yi4"). In addition, based on the Figure 4d,f,h, there are six correct tones in all of the three utterances. This result illustrates that all approaches can modify the tones of EL speech. However, the CLDNN-based approach has a worse performance in enhancing continuous Mandarin EL speech with fixed tone. In addition, narrow-band spectrogram analysis is applied to show the enhancement of temporal details. The results show that our approach has the best performance in enhancing Mandarin EL speech (even including the EL-FT).

Furthermore, comparing with Figure 4a,b, from Figure 4c–h, we can find that the radiation noise of Mandarin EL speech was cancelled. This result show that all VC methods can cancel the radiation noise with high performance.

3.2. Subjective Evaluation

3.2.1. Results and Analysis of Tone Accuracy

Tone accuracy is an important factor to analyze the enhancement of continuous Mandarin EL speech. The accuracy of all tones is listed in Table 5.

Table 5. The tone accuracy of unenhanced/enhanced continuous Mandarin EL speech.

Speech Types	Tone1	Tone2	Tone3	Tone4	Mean	STD
EL-FT	95.44%	10.37%	0.61%	13.14%	29.89%	0.3813
CLDNN-FT	77.90%	49.03%	44.31%	66.50%	59.44%	0.1349
VC3-FT	84.22%	65.33%	53.32%	75.38%	69.56%	0.1151
2C1T2C-FT	92.15%	68.22%	53.86%	74.85%	72.27%	0.1583
EL-VT	87.64%	55.49%	56.04%	87.20%	71.59%	0.1583
CLDNN-VT	83.25%	66.47%	67.28%	78.50%	73.88%	0.0720
VC3-VT	88.75%	76.58%	78.74%	85.32%	82.35%	0.0490
2C1T2C-VT	91.00%	77.84%	79.96%	86.05%	83.71%	0.0518

From Table 5, for the EL-FT, the accuracy of Tone2, Tone3 and Tone4 is not zero. This is because listeners would inevitably correct the tone according to what they have heard. In addition, in one utterance, the first tone sounds like Tone2 and the last tone sounds like Tone4. This is because the speaker pressed the EL button at the beginning of pronouncing and released the button at the end. The process is shown in Figure 4a. For EL-VT, after being enhanced using the CLDNN-based approach, the accuracy of Tone1 and Tone4 decreases and the accuracy of Tone2 and Tone3 increases. After being enhanced using our proposed approach, the accuracy of Tone2, Tone3 and Tone4 increases while the accuracy of Tone1 decreases slightly. For EL-FT, the average tone accuracy increases by 29.55% through the CLDNN-based approach, by 39.67% through the V3-based approach and by 42.38% through our approach. For EL-VT, the improvement of different tones is not stable through CLDNN-VT and CycleGAN-VT. For CLDNN-VT, the accuracy of Tone1 and Tone4 decreases dramatically, while for VC3-VT and 2C1T2C-VT, only the accuracy of Tone4 decreases slightly. For EL-VT, the average tone accuracy increases 2.29% through the CLDNN-based approach, 11.16% through the V3-based approach and 12.12% through our

approach. In summary, our approach performs better than others in enhancing the tone variation of continuous Mandarin EL speech.

3.2.2. Results of Semantic Information Evaluation

In our proposed framework, the generator is designed based on conformer and transformer neural networks. The generator can thus capture semantic and frequency information to improve Mandarin EL speech. To evaluate this performance, we prepared special testing data including "Chinese Characters (single syllable)", "Chinese Words (two syllables)", "Chinese Idioms (four syllables)", "short sentences (six syllables on average)" and "long sentences (over ten syllables)". The results are shown in Table 6.

Table 6. Tone accuracy of Mandarin EL speech with different amount of semantic information (average accuracy %/STD) [1].

Speech Types	Chinese Characters	Chinese Words	Chinese Idioms	Short Sentences	Long Sentences
CLDNN-FT	40.00/0.49	45.00/0.38	48.75/0.40	53.28/0.24	26.22/0.14
VC3-FT	35.00/0.48	47.50/0.40	63.75/0.27	64.45/0.15	63.78/0.12
2C1T2C-FT	30.00/0.46	42.50/0.33	68.75/0.25	71.74/0.16	73.69/0.16
CLDNN-VT	-	-	87.5/0.15	71.56/0.15	73.98/0.09
VC3-VT	-	-	87.5/0.15	80.46/0.10	78.13/0.11
2C1T2C-VT	-	-	88.75/0.15	83.05/0.08	85.93/0.12

[1] Please note that in Table 6 every type of testing data has 20 items, such as 20 characters, 20 words, 20 idioms and 20 sentences. "-" means no results.

Obtained from the special testing data, the results from Table 6 are the average values calculated based on each utterance. The tones of Chinese characters and Chinese words pronounced using an EL touch panel are all correct. In this case, statistical analysis is unnecessary for the tone accuracy of CLDNN-VT, VC3-VT and 2C1T2C-VT. The results of enhanced EL-FT show that if the testing data (especially Chinese characters) contain little semantic information, the enhanced speech has low tone accuracy. Under our proposed approach, if only one or two characters are tested, the results are random. In this case, our approach is even less effective than CLDNN-based or VC3-based approaches. However, with sufficient semantic information, our approach achieves much higher tone accuracy than CLDNN-based or VC3-based approaches. For instance, under our proposed approach, the tone accuracy of Chinese idioms as well as short and long sentences is much higher than that of Chinese characters or Chinese words. The above results show that our proposed approach takes advantage of semantic information to improve the tone accuracy of Mandarin EL speech.

3.2.3. Results of Word Error Perception Rate

The word error perception rate (WER, or maybe syllable or pinyin error perception rate) is another important metric to evaluate the enhancement effect of Mandarin EL speech. WER can be used to evaluate the enhancement effect of the consonants and vowels of Mandarin EL speech. The WER in Table 7 is calculated based on the comparison of the results written by the listeners and the real contents. Please note that here WER means the pinyin/syllable error rate and does not include tonal errors.

Table 7 shows the WER of unenhanced/enhanced continuous Mandarin EL speech. Interestingly, tone variation does not significantly influence the WER of Mandarin EL speech. In addition, for EL-FT, the WER of CLDNN-FT decreases by 3.07%, the WER of VC3-FT by 6.41% and that of our approach by 8.59%. For EL-VT, the WER of CLDNN-VT decreases by 5.26%, the WER of VC3-VT by 6.53% and that of our approach by 9.15%. The results show that tone variation has substantial influence on the enhancement effect of the CLDNN-based approach but little influence on the VC3-based and our approach.

Table 7. The statistical WER of unenhanced/enhanced continuous Mandarin EL speech.

Speech Types	WER	STD
EL-FT	10.74%	0.1081
CLDNN-FT	7.61%	0.1111
VC3-FT	4.33%	0.1933
2C1T2C-FT	2.15%	0.0495
EL-VT	10.85%	0.0949
CLDNN-VT	5.59%	0.0978
VC3-VT	4.32%	0.1620
2C1T2C-VT	1.70%	0.0521

3.2.4. MOS of Intelligibility and Naturalness

The results of subjective evaluation were obtained from the 250 test utterances. The subjective evaluation includes MOS on speech's intelligibility and naturalness. Mandarin Chinese is a typical tonal language where tone influences the listeners' perception of speech's intelligibility and naturalness. The higher the tone accuracy, the higher the MOS of intelligibility and naturalness. The intelligibility and naturalness MOS are shown in Figure 5 and Figure 6, respectively.

In Figure 5, EL represents the intelligibility MOS of EL-FT and EL-VT; CLDNN represents the intelligibility MOS of CLDNN-FT and CLDNN-VT; VC3 represents the intelligibility MOS of VC3-FT and VC3-VT; and 2C1T2C represents the intelligibility MOS of 2C1T2C-FT and 2C1T2C-VT. The confidence interval here is 95%.

The experiment results from Figure 5 show the perceptual intelligibility of test speech. The intelligibility of EL-VT speech is higher than that of EL-FT speech because tone variation influences the comprehension of Mandarin speech. Our approach achieves higher intelligibility of enhanced speech than other approaches. Meanwhile, the intelligibility of enhanced speech with variable tone is higher than that of enhanced speech with fixed tone.

In Figure 6, EL represents the naturalness MOS of EL-FT and EL-VT; CLDNN represents the naturalness MOS of CLDNN-FT and CLDNN-VT; VC3 represents the naturalness MOS of VC3-FT and VC3-VT; and 2C1T2C represents the naturalness MOS of 2C1T2C-FT and 2C1T2C-VT. The confidence interval here is 95%.

The experiment results in Figure 6 show the naturalness of enhanced speech. The naturalness of EL-FT and EL-VT are close. Both enhancement and tone variation can effectively improve the naturalness of continuous Mandarin EL speech. Tone variation makes the enhancement less difficult and tone accuracy of enhanced speech with variable tone is much higher than that of enhanced speech with fixed tone. In addition, the naturalness of enhanced speech under our approach is higher than that under other approaches.

The results of both objective and subjection evaluation show that our proposed approach can improve both the intelligibility and naturalness of continuous Mandarin EL speech. The better the proposed approach performs, the more effective the enhancement is.

4. Discussion

VC is a powerful tool to improve the intelligibility and naturalness of EL speech. Traditional VC approaches are on a parallel-dependent basis and are based on the belief that the mapping relationship from source to target acoustic features can be represented using statistical machine learning. However, parallel-dependent approaches are weak in enhancing Mandarin EL speech, especially EL-FT, because such approaches cannot take advantage of contextual semantic information during processing. CycleGAN is a powerful tool in image transformation [32], from which we take reference to develop a more effective approach to enhance Mandarin EL speech. Traditionally, parameters including F0, Mel-Cepstrum, and band aperiodicity (extracted using STRAIGHT [38], or CodeAP extracted using WORLD) are used as acoustic features. In this paper, we use the Mel-Spectrogram parameters instead of a combination of features. The Mel-Spectrogram parameters have been widely applied in the text-to-speech conversion [39–41]. The experiment results show

that Mel-Spectrogram parameters and the combination of features do not differ greatly from each other in enhancing continuous Mandarin EL speech. Furthermore, we use WaveNet as the neural vocoder to directly resynthesize the Mel-Spectrogram parameters into waveform signals. Our research can be applied to improve the communication ability of the laryngectomees. Additionally, this proposed method can also be used to popularize the application of the Electro-Larynx.

The objective and subjective evaluation results show that the proposed approach performs better than traditional parallel-dependent VC in enhancing Mandarin EL speech. In particular, the proposed approach substantially improves tone accuracy and reduces WER. Moreover, the intelligibility and naturalness MOSs of 2C1T2C-FT are much higher than those of EL-FT, CLDNN-FT and VC3-FT; the intelligibility and naturalness MOSs of 2C1T2C-VT are much higher than those of EL-VT, CLDNN-VT and VC3-VT. This is because the proposed approach takes advantage of contextual semantic information based on the whole input speech. Our approach improves both the frequency and temporal information of Mandarin EL speech.

Although the proposed approach effectively improves the intelligibility and naturalness of Mandarin EL speech, the training process still takes up more than 10 GB of graphics processing unit (GPU) memory and at least 168 h. The time cost of our proposed method, CLDNN-based VC and VC3 are close. In the future, it would be an important task to explore a quicker training process requiring smaller GPU memory. Delay of processing is one major shortcoming of our approach. In the future, we would explore ways to reduce the delay, which would be more useful for laryngectomees.

5. Conclusions

In this paper, CycleGAN framework is applied to enhance Mandarin EL speech, where the generator of CycleGAN is designed based on 2D-Con-1D-Tran-2D-Con neural networks. The experiment results show that this proposed approach can effectively improve the intelligibility and naturalness of Mandarin EL speech. In particular, tone accuracy improves substantially. This proposed approach greatly helps laryngectomees communicate in Mandarin Chinese. In the future, we will explore how to expand the application of the proposed approach, which can also be useful for improving EL speech in other tonal languages.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app13010537/s1.

Author Contributions: Conceptualization, Z.Q.; methodology, Z.Q.; formal analysis, Z.Q.; writing—original draft preparation, K.X.; writing—review and editing, K.X. and C.Y.; funding acquisition, Z.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Humanity and Social Science Youth Foundation of Ministry of Education of China, grant number 21YJCZH117.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guo, L.; Nagle, K.F.; Heaton, J.T. Generating tonal distinctions in Mandarin Chinese using an electrolarynx with preprogrammed tone patterns. *Speech Commun.* **2016**, *78*, 34–41. [CrossRef] [PubMed]
2. Liu, H.; Ng, M.L. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx* **2007**, *34*, 327–332. [CrossRef] [PubMed]
3. Watson, P.J.; Schlauch, R.S. Fundamental Frequency Variation with an Electrolarynx Improves Speech Understanding: A Case Study. *Am. J. Speech Lang. Pathol.* **2009**, *18*, 162–167. [CrossRef] [PubMed]
4. Uemi, N.; Ifukube, T.; Takahashi, M.; Matsushima, J. Design of a new electrolarynx having a pitch control function. In Proceedings of the 1994 3rd IEEE International Workshop on Robot and Human Communication, Nagoya, Japan, 18–20 July 1994; pp. 198–203. [CrossRef]

5. Goldstein, E.A.; Heaton, J.T.; Kobler, J.B.; Stanley, G.B.; Hillman, R.E. Design and Implementation of a Hands-Free Electrolarynx Device Controlled by Neck Strap Muscle Electromyographic Activity. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 325–332. [CrossRef] [PubMed]
6. Goldstein, E.A.; Heaton, J.T.; Stepp, C.E.; Hillman, R.E. Training Effects on Speech Production Using a Hands-Free Electromyographically Controlled Electrolarynx. *J. Speech Lang. Hear. Res.* **2007**, *50*, 335–351. [CrossRef] [PubMed]
7. Choi, H.-S.; Park, Y.J.; Lee, S.M.; Kim, K.-M. Functional Characteristics of a New Electrolarynx "Evada" Having a Force Sensing Resistor Sensor. *J. Voice* **2001**, *15*, 592–599. [CrossRef]
8. Wang, L.; Qian, Z.; Feng, Y.; Niu, H. Design and Preliminary Evaluation of Electrolarynx with F0 Control Based on Capacitive Touch Technology. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 629–636. [CrossRef]
9. Wan, C.; Wang, E.; Wu, L.; Wang, S.; Wan, M. Design and Evaluation of an Electrolarynx with Tonal Control Function for Mandarin. *Folia Phoniatr. Logop.* **2012**, *64*, 290–296. [CrossRef] [PubMed]
10. Espy-Wilson, C.Y.; Chari, V.R.; Huang, C.B. Enhancement of alaryngeal speech by adaptive filtering. In Proceedings of the 4th International Conference on Spoken Language Processing ICSLP'96, Philadelphia, PA, USA, 3–6 October 1996; pp. 764–767. [CrossRef]
11. Espy-Wilson, C.Y.; Chari, V.R.; Macauslan, J.M.; Huang, C.B.; Walsh, M.J. Enhancement of Electrolaryngeal Speech by Adaptive Filtering. *J. Speech Lang. Hear. Res.* **1998**, *41*, 1253–1264. [CrossRef] [PubMed]
12. Niu, H.J.; Wan, M.X.; Wang, S.P.; Liu, H.J. Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis. *Med. Biol. Eng. Comput.* **2003**, *41*, 670–678. [CrossRef]
13. Cole, D.; Sridharan, S.; Moody, M.; Geva, S. Application of noise reduction techniques for alaryngeal speech enhancement. In Proceedings of the IEEE TENCON'97, IEEE Region 10 Annual Conference, Speech and Image Technologies for Computing and Telecommunications, Brisbane, QLD, Australia, 4 December 1997; pp. 491–494. [CrossRef]
14. Liu, H.; Zhao, Q.; Wan, M.; Wang, S. Enhancement of electrolarynx speech based on auditory masking. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 865–874. [CrossRef]
15. Pandey, P.C.; Bhandarkar, S.M.; Bachher, G.K.; Lehana, P.K. Enhancement of alaryngeal speech using spectral subtraction. In Proceedings of the 2002 14th International Conference on Digital Signal Processing, Santorini, Greece, 1–3 July 2002; pp. 591–594. [CrossRef]
16. Mahmmod, B.M.; Abdulhussain, S.H.; Naser, M.A.; Alsabah, M.; Mustafina, J. Speech Enhancement Algorithm Based on a Hybrid Estimator. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1090*, 012102. [CrossRef]
17. Mahmmod, B.M.; Ramli, A.R.; Baker, T.; Al-Obeidat, F.; Abdulhussain, S.H.; Jassim, W.A. Speech Enhancement Algorithm Based on Super-Gaussian Modeling and Orthogonal Polynomials. *IEEE Access* **2019**, *7*, 103485–103504. [CrossRef]
18. Wang, D.L.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef] [PubMed]
19. Kobayashi, K.; Toda, T. Electrolaryngeal Speech Enhancement with Statistical Voice Conversion based on CLDNN. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 2115–2119. [CrossRef]
20. Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Commun.* **2012**, *54*, 134–146. [CrossRef]
21. Yang, Y.G.; Zhang, H.Z.; Cai, Z.X.; Shi, Y.; Li, M.; Zhang, D.; Ding, X.J.; Deng, J.H.; Wang, J. Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion. *Biomed. Signal Process. Control* **2023**, *80*, 104279. [CrossRef]
22. Kobayashi, K.; Toda, T. Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 396–400. [CrossRef]
23. Qian, Z.; Niu, H.; Wang, L.; Kobayashi, K.; Zhang, S.; Toda, T. Mandarin Electro-Laryngeal Speech En-hancement based on Statistical Voice Conversion and Manual Tone Control. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 546–552.
24. Qian, Z.; Wang, L.; Zhang, S.; Liu, C.; Niu, H. Mandarin Electrolaryngeal Speech Recognition Based on WaveNet-CTC. *J. Speech Lang. Hear. Res.* **2019**, *62*, 2203–2212. [CrossRef]
25. Qian, Z.; Xiao, K.; Liu, C.; Sun, Y. Voice Conversion for Enhancing Mandarin Electro-Laryngeal Speech Based on Semantic Information. *Acta Electron. Sin.* **2020**, *48*, 840–845. [CrossRef]
26. Kaneko, T.; Kameoka, H. Parallel-data-free voice conversion using cycle-consistent adversarial networks. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 2114–2118.
27. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion. *arXiv* **2020**, arXiv:2010.11672. [CrossRef]
28. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. Cyclegan-VC2: Improved Cyclegan-based Non-parallel Voice Conversion. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6820–6824. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. At-tention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

30. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020.
31. Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
32. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
33. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
34. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
35. Oord, A.; Kalchbrenner, N.; Espeholt, L.; Kavukcuoglu, K.; Vinyals, O.; Graves, A. Conditional image gen-eration with pixelcnn decoders. In Proceedings of the NIPS'16: 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4797–4805.
36. Wang, D.; Zhang, X. THCHS-30: A free Chinese speech corpus. *arXiv* **2015**, arXiv:1512.01882.
37. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
38. Kawahara, H.; Masuda-Katsuse, I.; de Cheveigné, A. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.* **1999**, *27*, 187–207. [CrossRef]
39. Li, N.; Liu, S.; Liu, Y.; Zhao, S.; Liu, M. Neural Speech Synthesis with Transformer Network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6706–6713. [CrossRef]
40. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural Tts Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
41. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.