Electrolaryngeal Speech Enhancement based on a two stage framework with Bottleneck Feature Refinement and Voice Conversion

Yaogen Yang[a,d,*], Haozhe Zhang[a,*], Zexin Cai[a], Yao Shi[a,c], Ming Li[a,c,**], Dong Zhang[d], Xiaojun Ding[b,**], Jianhua Deng[b], Jie Wang[b]

[a] *Data Science Research Center, Duke Kunshan University, Kunshan, China*
[b] *Department of Otolaryngology, The First People's Hospital of Kunshan, Kunshan, China*
[c] *School of Computer Science, Wuhan University, Wuhan, China*
[d] *The School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China*

Abstract

An electrolarynx (EL) is a medical device that generates speech for people who lost their biological larynx. However, EL speech signals are unnatural and unintelligible due to the monotonous pitch and the mechanical excitation of the EL device. This paper proposes an end-to-end voice conversion method to enhance EL speech. We adopt a speaker-independent automatic speech recognition model to extract bottleneck features as the intermediate phonetic features for enhancement. Our system includes two stages: the bottleneck feature vectors of the EL speech are mapped by a parallel non-autoregressive model to the corresponding feature vectors of the normal speech in stage one. Then another voice conversion model maps normal speech's bottleneck feature vectors directly to normal speech's Mel-spectrogram in stage two, followed by a MelGAN-based vocoder to convert the Mel-spectrogram into waveform. In addition, we incorporate data augmentation and transfer learning to improve conversion performance. Experimental results show that the proposed method outperforms our baseline methods and performs well in terms of naturalness and intelligibility.

[*] The first two authors contribute equally.
[**] Corresponding authors: Ming Li: ming.li369@duke.edu, Xiaojun Ding: entdxj@outlook.com

The audio samples are available online[1].

Keywords: Electrolarynx speech, Voice conversion, Bottleneck features, Speech enhancement

1. Introduction

An *electrolarynx* (EL) is a medical device designed for people who lose their biological larynx. Patients with laryngeal cancer who receive full therapy have their larynx removed by the total laryngectomy surgery, and as a result, they lose the fundamental frequency generation mechanism of the human vocal tract [1]. The EL is one of the speaking-aid devices they can use to rehabilitate their speech [2]. Although laryngectomees lose their larynges, the other organs for producing speech are usually unaffected. In this case, an EL is used to substitute the function of the removed larynx. Typically, laryngectomees hold the EL against their neck to produce speech. Energy signals are transmitted to the speaker's oral cavity through the speaker's skin. Then energy signals are carried through other vocal organs that filter the signals and produce various pronunciations [3].

However, the resulting speech generated by an EL usually has several fundamental issues that make it dissimilar to natural speech [4]: 1. EL speech sounds unintelligible and unnatural due to its constant fundamental frequency (F0), 2. the continuous vibration of the EL device causes undesired noise. There are various ways to reduce the problems and improve the speech quality. Generally, noise reduction and fundamental frequency prediction are two main methods for EL speech enhancement. On the hardware side, advanced EL devices are designed with the capability to change voice intonation or pitch [1, 5, 6], while some of the devices require laryngectomees' additional practices on use. On the software side, post-processing algorithms for background noise reduction have been proposed for EL speech enhancement [7, 8]. Mathew et al. have stud-

[1] https://haydencaffrey.github.io/el/index.html

ied the effectiveness of two noise reduction algorithms, dimensional amplitude trimmed estimation (DATE) and non-negative matrix factorization (NMF) [9], which have found that NMF algorithms outperforms DATE-based algorithms on reducing acoustic noise in EL speech.

Nonetheless, recent works have shown that the voice conversion (VC) technique is more effective in EL speech enhancement [4, 10, 11, 12, 13]. VC-based approaches aim to convert the EL speech to natural speech by mapping models trained with parallel audio pairs [3, 14]. Previously, most VC-based approaches use Gaussian Mixture Models (GMM) [3, 4, 15, 16, 17] to predict the excitation parameters and adopt parametric vocoders, e.g., STRAIGHT [18] and WORLD [19], to reconstruct the enhanced speech. Among those excitation parameters, the prediction of F0 is regarded as the most important yet challenging part [6, 20]. Accordingly, Li et al. propose a hybrid approach using NMF and G-MM to estimate a smoothed F0 contour [17]. After considering the physical mechanism and the constrain of vocal phonation and speech production, the F0 estimation performance is further enhanced [15, 16]. However, the F0 contour is highly related to linguistic information [21], which motivates researchers to use phonetic features, including phoneme labels [10], phoneme embeddings [11], and phonetic posterior probabilities (PPP) [14], for predicting F0 contour. Furthermore, it is found that replacing the GMM with deep neural network models can effectively improve the quality of the converted EL speech [12, 13, 22]. Specifically, the network structure CLDNN is adopted in EL speech conversion and has shown to achieve higher naturalness and perceptual speech intelligibility than GMM models [12, 13].

However, the aforementioned approaches rely on speech features generated from conventional parametric vocoders where the phase information is discarded. As a result, acoustic artifacts are also introduced in synthesized speech. On the other hand, the end-to-end VC [23, 24, 25, 26] and neural vocoders [27, 28, 29] have dominated the voice conversion field recently with their high-fidelity synthesis performance. In addition, studies show that the bottleneck features or phoneme posterior probability features extracted by the acoustic model from an automatic speech recognition system can provide sufficient phonetic information for the VC task [30, 31]. Specifically, Ding et al. train a speech synthesizer to map PPP from the non-native speaker into the corresponding target Mel-spectrogram [32]. Such conversion techniques relieve the efforts of predicting multiple speech features, including F0, aperiodicity feature and U/V (Unvoiced/Voiced) labels, in the traditional statistical approaches. Instead, the end-to-end VC only needs to predict the spectrogram. These advancements have inspired us to investigate neural voice conversion for EL speech enhancement.

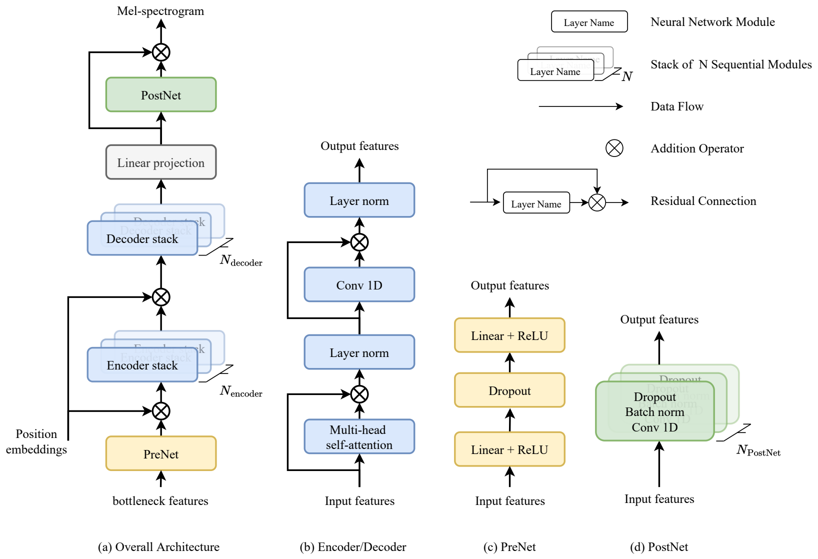
In this paper, we propose three speech-to-speech conversion systems using bottleneck features to enhance EL speech. The first system is the bottleneck feature to Mel-spectrogram (BN-MEL) VC system. This system directly maps the bottleneck features of the EL speech to the Mel-spectrogram of the normal speech using a parallel non-autoregressive model. However, considering that the bottleneck features of the EL speech may follow different distribution from those of normal speech and the parallel training data is very limited, the direct mapping from the bottleneck feature of the EL speech to the Mel-spectrogram of the normal speech is difficult. Therefore, we propose to add an intermediate conversion step to form a two-stage framework. Specifically, we first map bottleneck features of EL speech to bottleneck features of normal speech, then convert bottleneck features of normal speech to Mel-spectrogram of normal speech, which is our BN-BN-MEL VC system. In addition, by substituting the real F0 values of normal speech into constant ones and generating augmented speech through the WORLD vocoder, we generate speech signals close to EL speech. The augmented speech is then used to pre-train the BN-BN-MEL model, while the EL speech is used to finetune the model. We call this system the BN-BN-MEL-P system. In addition to our proposed systems, we also reproduce the GMM-based conversion system proposed in [14] as one of our baseline systems. The reported state-of-the-art system, the CLDNN-based conversion enhancement system proposed in [12], is also reproduced for comparison. By conducting objective and subjective evaluations, we evaluate the performance of our proposed systems. For the objective evaluation, we use an automatic speech recognition system to evaluate the intelligibility of generated speech. The performance is measured by Word Error Rate (WER). The lower the WER, the better the intelligibility of generated speech. We also use the distortion between the target Mel-spectrogram and the predicted Mel-spectrogram as another objective evaluation metric. For subjective evaluation, we ask participants to rate the generated speech on naturalness and intelligibility; then, we calculate the Mean Opinion Score (MOS). According to the subjective and objective evaluation results, our proposed BN-BN-MEL-P system achieves the best performance on naturalness and intelligibility.

Our work has the following contributions. First, our model achieves state-of-the-art performance on both naturalness and intelligibility for EL speech enhancement. Second, our proposed system works well on data-limited scenarios due to the two-stage learning process and the transfer learning strategies. Third, we incorporate the transformer and a neural vocoder to electrolarynx speech enhancement, which significantly improves the naturalness of the enhanced EL speech.

The rest of this paper is organized as follows. Section 2 describes our proposed parallel non-autoregressive VC system for EL speech enhancement. Section 3 shows the experimental setup and results, and the conclusion is given in Section 4.

2. Method

We propose three systems, namely BN-MEL, BN-BN-MEL and BN-BN-MEL-P, in this paper for EL speech enhancement. All systems utilize a parallel non-autoregressive network architecture as the conversion model, shown in Figure 1. The BN-MEL system directly adopts the conversion model to map the bottleneck feature of EL speech to the Mel-spectrogram of the corresponding target speech. The BN-BN-MEL system first adopts a conversion model to map the bottleneck feature of the EL speech to the bottleneck feature of the target



(a) Overall Architecture (b) Encoder/Decoder (c) PreNet (d) PostNet

Figure 1: (a) The overall architecture of our proposed model. (b) The Encoder and Decoder network. (c) The PreNet module. (d) The PostNet module.

speech, then use another conversion model to convert the bottleneck feature of the target speech to the Mel-spectrogram of the target speech. The BN-BN-MEL-P system applies a pre-training strategy to boost the performance of the BN-BN-MEL system.

Specifically, the input bottleneck feature is extracted by a bottleneck extractor from an automatic speech recognition (ASR) model to obtain linguistic information. The output feature of our systems is the Mel-spectrogram of target speech, which is then converted back to waveform by a neural vocoder.

2.1. Bottleneck Feature Extractor

In our proposed system, we apply a pre-trained ASR model trained with Kaldi [33] to extract linguistic information from EL speech. The acoustic model used to predict phonetic probabilities in ASR is employed as the bottleneck feature extractor. The acoustic model contains multiple time-delayed neural network (TDNN) layers, followed by a linear layer that maps the hidden feature to a lower-dimensional embedding. We refer the output of the linear layer as the speaker-independent bottleneck feature [34]. The bottleneck feature can accentuate linguistic information well since the acoustic model is trained with the genuine phonetic target for each acoustic frame. Therefore the bottleneck feature can be used as the linguistic feature for voice conversion.

2.2. The Conversion Model

The overall architecture of our conversion model is shown in Figure 1 (a), where the network structure of the Encoder and the Decoder is shown in Figure 1 (b). Basically, the architecture is used to model the mapping between input feature sequence $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l]$ to output feature sequence $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_l]$, where $l \in \mathbb{N}$ is the length of the sequence. The mapping is performed by a series of neural network components with trainable parameters θ. The conversion model is optimized according to several loss functions denoted by $\mathcal{L}$.

The input acoustic feature is mapped into a high dimensional latent space by the PreNet module, followed by an Encoder that encodes the latent representation and a Decoder that transforms encoder outputs to target acoustic features. The encoder and the decoder adopt the same structure: a stack of identical layers consisting of a multi-head attention layer, residual and normalization layers, and a convolutional layer. Particularly, residual connections are extensively used in our proposed model as the residual connection has been proven effective for deep neural networks [35]. The residual connection provides another data flow that performs identity mapping to the input. The output of the residual connection can be formulated by $F(x) + x$, where x is the identity mapping of input and F is a network component. The decoder outputs are then fed into linear layers to predict the target acoustic features, and finally the PostNet module is applied to finetune the predicted features.

2.2.1. PreNet Network

The structure of the Pre-net is shown in Figure 1 (c). The PreNet contains two fully connected hidden layers and one dropout layer. Both fully connected hidden layers have 256 units with ReLU as the activation function. During training, the linear output is dropped out with a dropout rate of 0.5.

The PreNet module transforms the input feature X into a high-dimensional representations Z as in Equation 1.

$$Z = \mathrm{PreNet}_{\theta}\left(X\right) \tag{1}$$

2.2.2. Positional Encoding

In order to utilize the context information of the input sequence, we apply the positional encoding mechanism to the feature sequence as proposed in [36]. $\mathbf{p}_j = [p_j^0, p_j^1, ...p_j^{2i}, p_j^{2i+1}, ..., p_j^{d-1}]^\top$ is added to the input feature sequence, where d is the dimension of the feature, $j \in [0, l-1]$ is the index for the input feature sequence. Specifically, the positional encoding mechanism is formulated as in Equation 2 and 3, where $i \in [0, ..., d/2-1]$ is the index of the feature dimension. The positional encoding provides explicit information of the currently processed portion in the sequence.

$$p_j^{2i} = \sin\left(j/10000^{2i/d}\right), \tag{2}$$

$$p_j^{2i+1} = \cos\left(j/10000^{2i/d}\right) \tag{3}$$

2.2.3. The Encoder and Decoder Structure

The structure of the encoder and decoder are shown in Figure 1 (b). The encoder is composed of 4 identical blocks. Each block employs a multi-head self-attention layer and a one-dimensional convolutional layer. Normalization layers are added after each of those two layers. At the same time, residual connections are applied between the input and the output of those two layers. The decoder has the same feed-forward network structure as the encoder, which significantly speeds up the training and inference process. The output of the decoder is then fed to a linear layer with 80 units to predict the corresponding target features.

The multi-head self-attention mechanism allows the model to jointly attend to information from different representation subspaces at different positions.

The outputs of the self-attention layer is obtained according to Equation 4, where Q, K and V are attention queries, keys and values. d_k is the dimension of the key. The multi-head attention layer can be formulated as Equation 5, where attention head$_i$ is calculated by Equation 6. In particular, W_i^Q, W_i^K, W_i^V, W^O are all trainable matrices for linear transformations of queries, keys, values and outputs, respectively. The number of attention heads is set to 4 in our experiments.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{4}$$

$$\text{Multihead}(Q, K, V) = \text{concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \tag{5}$$

$$\text{head}_\text{i} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

2.2.4. PostNet Network

The PostNet, shown in Figure 1 (d), is designed to refine the Mel-spectrogram predicted by the decoder. The PostNet is a convolutional neural network with a residual connection between the input and the output. It is composed of a 5 one-dimensional convolutional layers with 512 filters.

2.2.5. Loss Function

We denote the predicted Mel-spectrogram as $\hat{Y} = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_l]$ and the ground truth Mel-spectrogram is $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_l]$, where $l \in \mathbb{N}$ denotes the length of the feature sequence. As shown in equation 7, the loss function for optimizing the neural parameters of our proposed model is composed of two reconstruction losses, $\mathcal{L}_{linear}$ and $\mathcal{L}_{post}$. In particular, $\mathcal{L}_{linear}$ is the MSE loss between the Mel-spectrogram predicted by the linear projection layer before the PostNet and the target Mel-spectrogram, while $\mathcal{L}_{post}$ denotes the MSE loss between the Mel-spectrogram finetuned by the PostNet and the target Mel-spectrogram. The Mean Square Error (MSE) shown in Equation 8 is the loss function that we use to evaluate the predicted result and the target Mel-spectrogram in the training process. $\mathbf{y}_j$ is the j^{th} feature vector of the ground truth feature sequence, and $\hat{\mathbf{y}}_j$ is the j^{th} corresponding predicted feature vector.

$$\mathcal{L} = \mathcal{L}_{linear} + \mathcal{L}_{post} \tag{7}$$

$$\mathcal{L}_{MSE} = \frac{1}{l} \sum_{i=1}^{l} \left(\mathbf{y}_j - \hat{\mathbf{y}}_j\right)^2, l \in \mathbb{N} \tag{8}$$

2.3. The BN-MEL System

For the BN-MEL system, the input is the bottleneck feature extracted from the EL speech, the output is the Mel-spectrogram of the corresponding normal speech. As shown in Figure 2, in the training stage, the system learns to convert the given bottleneck feature of EL speech directly to the Mel-spectrogram of normal speech with the same content. In the test stage, the system predicts the Mel-spectrogram from the given bottleneck feature of the test EL speech, and then the neural vocoder MelGAN is used to convert Mel-spectrogram to waveform.

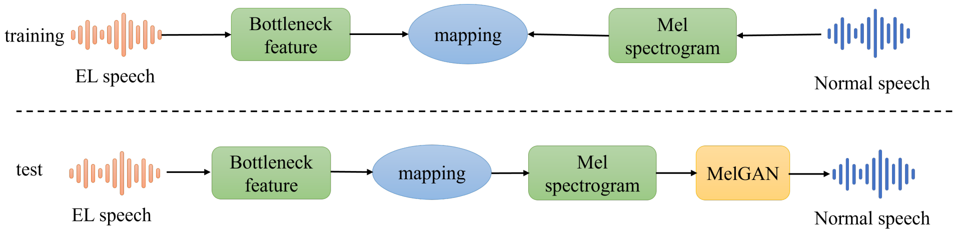


Figure 2: The overview of our proposed BN-MEL VC system

2.4. The BN-BN-MEL System

Considering that the EL speech is different from the normal speech, the bottleneck features extracted by the ASR model trained on normal speech may inevitably leads to some errors. As a result, the BN-MEL system may have accumulated errors in the output due to errors in bottleneck features, which then affects the intelligibility and naturalness of the converted speech. In order to address this issue, we propose the BN-BN-MEL system. The framework is shown in Figure 3.

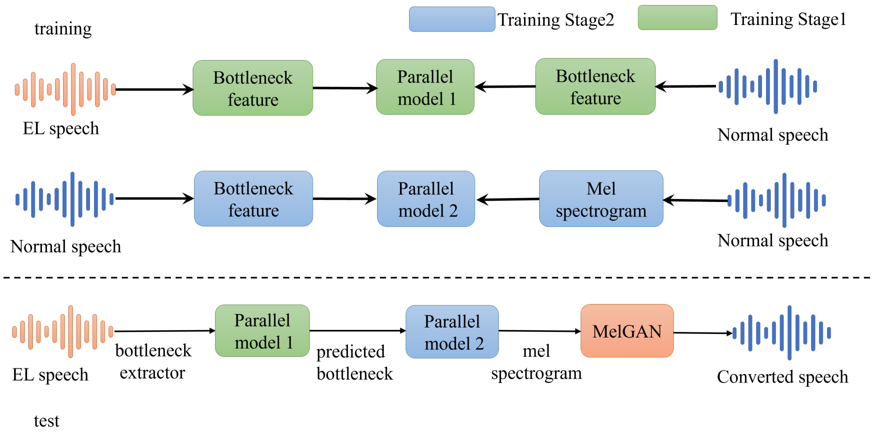


Figure 3: The overview of our proposed BN-BN-MEL system

In the training stage, the bottleneck feature of EL speech is mapped by a conversion model to the corresponding bottleneck feature of the normal speech. Let X_{EL}, X_{NL} denote the bottleneck features of parallel EL speech and normal speech with the same lexical content, respectively. Then the first conversion process can be formulated as Equation 9, where θ_1 are the model parameters of the conversion model 1. Another conversion model is followed to map normal speech's bottleneck feature vectors directly to normal speech's Mel-spectrogram. Let X_{NL}, Y denote the bottleneck features and Mel-spectrograms of normal speech. Likewise, the second conversion phase can be rendered by Equation 10, where θ_2 are the model parameters.

$$X_{NL} = \mathrm{Model}_{\theta_1}\left(X_{EL}\right) \tag{9}$$

$$Y = \mathrm{Model}_{\theta_2}\left(X_{NL}\right) \tag{10}$$

In the test stage, the bottleneck feature of EL speech is first fed to Model_1, and the output refined bottleneck feature is used as the input to conversion Model_2 to obtain the predicted Mel-spectrogram. Finally, the MelGAN-based vocoder network is adopted to convert the Mel-spectrogram into time-domain waveform.

2.5. The BN-BN-MEL-P System with Data Augmentation and Transfer Learning

By separating the mapping process from bottleneck feature to Mel-spectrogram into two stages, the BN-BN-MEL system manages to utilize other datasets to pre-train the model in each stage.

For the first model that maps the bottleneck feature of EL speech into the bottleneck feature of normal speech, we first pre-trained it on the Data-Baker DB4 dataset [2] with 11.84 hours of normal speech in Chinese Mandarin from a female speaker. Instead of directly using the DB4 dataset, we use the WORLD vocoder to decompose DB4 utterances and then substitute the F0 values with constant ones. Simulated EL utterances are synthesized by WORLD with the modified F0 contours. By extracting bottleneck features from the simulated EL speech with flat F0 and the original normal speech, we pre-train the first conversion model to map the bottleneck feature of the EL speech to that of the normal speech. Then, we finetune the pre-trained model using our collected real EL speech data.

We consider the second-stage model, which maps the bottleneck feature of normal speech into the Mel-spectrogram of normal speech, as a one-to-one voice conversion task. In this stage, we directly extract the bottleneck feature and Mel-spectrogram from the normal speech in the DB4 dataset and pre-train the model to map the bottleneck feature of normal speech to the Mel-spectrogram of normal speech. Then we finetune the pre-trained model using our collected natural speech. We denoted the system with finetuning strategy as the BN-BN-MEL-P system.

[2] https://www.data-baker.com/data/index/compose/

2.6. MelGAN Vocoder

We use the MelGAN[28] vocoder to convert the model's output back to speech waveform concerning its inference speed. MelGAN is a GAN-based network capable of generating high-quality speech from Mel-spectrograms. MelGAN, including a generator and discriminator architecture, is fully convolutional with significantly fewer parameters. Our MelGAN network is implemented based on the official open-source toolkit on Github[3].

3. EXPERIMENTS

3.1. The Conversion Model

The hyperparameters used in our conversion model is shown in Table 1. The PreNet is composed of 2 linear layers. Both the encoder and decoder have 4 blocks with two-head attention mechanism. The dimension of the key, query and value vector are set to 256. The PostNet is composed of 5 1-D CNN layers with a channel size of 80 and kernel size of 5.

Table 1: Hyperparameters of our conversion model

Module	Parameter
PreNet	2 linear layers
Encoder	block N=4, head=2 $d_k = d_q = d_v = 256$
Decoder	block N=4, head=2 $d_k = d_q = d_v = 256$
Linear	1 linear layer
PostNet	5 1D-CNN layers , channel=80 kernel size=5

[3] https://github.com/descriptinc/melgan-neurips

3.2. Data Preparation

Five hours of parallel EL speech and normal speech recorded by a healthy Chinese female speaker in [14] is used as our dataset. There are 3210 Mandarin utterances of EL speech and 3210 corresponding normal utterances. Each utterance consists of a short sentence. We randomly select 2900 utterance pairs for training and 310 utterance pairs for evaluation. The division of the dataset is the same as in [14]. For the vocoder, we use the AISHELL-3 dataset [37] to train our MelGAN model. While most studies from the literature downsample audio signals to 16kHz for experiments [12, 13, 14], we also downsample all speech utterances of aforementioned datasets to 16kHz for comparison. In addition, we preprocess the EL speech by reducing the background noises in EL speech using the WebRTC noise suppression algorithm[4]. Concerning acoustic feature extraction, 80-dimensional Mel-spectrograms were extracted every 12.5 ms with Hamming windowing of 50 ms frame length and 800-point Fourier transform. Mel-spectrograms are then normalized and scaled to range $[-4, 4]$. Since the parallel dataset mentioned above only contains one speaker, our experiments are restricted to single speaker's EL enhancement, while subject-wise and cross-gender scenarios are not studied in our work. However, our proposed method is gender-independent. The input bottleneck feature is extracted by a bottleneck extractor from a speaker-independent automatic speech recognition model, which is to obtain linguistic representations. Thus, the bottleneck feature is gender-independent. The following conversion modules are also gender-independent.

3.3. The Performance of the Bottleneck Extractor

In our systems, the bottleneck feature extractor is trained using the Mandarin dataset AISHELL-2 [38], which contains 1000 hours of clean reading-speech data for training. In addition, AISHELL-2 provides a development set containing 2500 utterances and a test set with 5000 utterances. The receipt for

[4] https://github.com/cpuimage/WebRTC_NS

training Librispeech in Kaldi is adopted here. The bottleneck extractor has 17 TDNN layers, followed by a 256-dim bottleneck layer. The frame's sub-sampling factor is set to 1 so that the output bottleneck features have the same length as the input acoustic features. The phoneme set we applied includes 52 Mandarin phonemes. The performance of our trained ASR system is shown in Table 2. The ASR model yields good performance on recognition test sets of AISHELL-2. It achieves a WER of 6.92% on the AISHELL-2 test set, which shows that the quality of this acoustic model is acceptable for linguistic feature extraction for natural speech.

Table 2: The speech recognition performance of our bottleneck feature extractor

Data Set	Word Error Rate (WER)
AISHELL-2 dev	6.20%
AISHELL-2 test	6.92%

The bottleneck extractor trained with natural utterances is unsuitable for EL speech considering acoustic differences between normal speech and EL speech. In this case, we finetune the bottleneck extractor for EL speech bottleneck extraction. The finetuned extractor trained with 2900 denoised EL utterances and 5000 randomly selected utterances from AISHELL-2. Therefore, in systems BN-BN-MEL and BN-BN-MEL-P, the bottleneck feature of EL speech is extracted by the finetuned extractor, while the bottleneck feature of normal speech is extracted by the one trained with AISHELL-2 only.

3.4. Objective Evaluation

We apply an automatic speech recognition (ASR) system for objective intelligibility estimation as introduced in [39]. The ASR system was trained using the Kaldi toolkit, and the acoustic model network uses the TDNN-f structure. Eight open source Chinese mandarin speech corpora[5] with a total of 2838 hours

[5] http://openslr.org/resources.php

of normal speech are used for training and testing. 2000 hours of normal speech are used for ASR training, while the remaining 838 hours of normal speech are used for evaluation. The WER on the evaluation set is 6.7%. In addition, half an hour of EL speech is used as an out-of-domain test set. In our experiments, we use this ASR model to evaluate the WER of the enhanced EL speech as an objective evaluation metric. A lower WER indicates higher intelligibility in the perspective of the ASR system. The ASR performance on different kinds of speech is shown in Table 3.

Table 3: The speech recognition performances of various speech utterances, including the EL, normal, and enhanced speech from our proposed systems.

Speech	ASR WER (%)
original EL speech	72.74
GMM VC enhanced speech [14]	69.85
CLDNN enhanced speech [12]	50.16
BN-MEL VC enhanced speech	84.09
BN-BN-MEL VC enhanced speech	54.39
BN-BN-MEL-P VC enhanced speech	42.62
original parallel normal speech	7.38

The enhanced EL speech obtained by the GMM conversion system achieves a WER of 69.85%, slightly lower than the WER of the original EL speech. However, the BN-MEL system makes the speech even more unintelligible than the original EL speech, as the WER of enhanced speech generated by the BN-MEL system is 84.09%. However, as we separate the conversion process into two stages, the performance is improved significantly. The BN-BN-MEL system achieves a WER of 54.39%. This indicates that converting bottleneck features of EL speech to bottleneck features of natural speech first can achieve better enhancement performance than converting to the Mel-spectrogram directly. The reported state-of-the-art CLDNN system has better performance with a WER of 50.16%. On the other hand, after adopting pre-training and transfer learning strategies, the intelligibility of the enhanced speech is further improved and the BN-BN-MEL-P system achieves the lowest WER, which is 42.62%.

We also employ the distortion of Mel-spectrogram as an objective measurement. The Mel-spectrogram Distortion (MSD) function is defined as in Equation 11, where $l \in \mathbb{N}$ denotes the sequence length, $\mathbf{m}$ denotes the Mel-spectrogram feature vector of normal speech and $\mathbf{m}^{cov}$ denotes the aligned Mel-spectrogram feature vector of converted speech, D represents the dimension of the spectral feature. The smaller the MSD value, the closer the converted speech to the target speech.

$$MSD = \frac{1}{l}\sum_{t=1}^{l} \frac{10\sqrt{2\sum_{i=1}^{D}\left(m_i - m_i^{cov}\right)^2}}{\ln 10}, l \in \mathbb{N} \tag{11}$$

The MSD results of different systems are shown Table 4. The GMM system has the highest distortion. The CLDNN system gets the second-highest distortion with a value of 16.85. Both systems rely on the traditional vocoder WORLD for synthesizing enhanced EL speech, which introduces a certain level of acoustic artifacts. The BN-BN-MEL system achieves a lower distortion value than the BN-MEL system. In addition, the speech synthesized by the BN-NB-MEL-P system achieves the lowest MSD value which demonstrate the effectiveness of data augmentation and transfer learning strategies.

Table 4: The Mel-spectrogram distortions (MSD) of various speech utterances, including the EL and enhanced speech from our proposed systems. CI denotes confidence interval.

Speech	MSD	CI ($\alpha = 0.05$)
original EL speech	18.055	[15.716, 20.393]
GMM VC enhanced speech [14]	16.851	[15.210, 18.491]
CLDNN enhanced speech	13.294	[11.539, 15.048]
BN-MEL VC enhanced speech	11.777	[10.173, 13.380]
BN-BN-MEL VC enhanced speech	11.674	[10.064, 13.283]
BN-BN-MEL-P VC enhanced speech	9.809	[8.188, 11.429]

3.5. Subjective Evaluation

To evaluate the performance of our proposed systems, we conducted subjective evaluation regarding the naturalness, speaker similarity and intelligibility of the converted speech. We ask 21 native Mandarin speakers to rate the converted speech based on the mean opinion score (MOS) scale:

- Naturalness, this evaluates how natural that the converted speech sound. Listeners evaluate the given speech in scale 1-5. The higher the value, the better the naturalness.

- Similarity, this evaluates how much the converted voice sounds like the target voice. Listeners evaluate the given synthetic voice in scale 1-5.

- Intelligibility, this evaluates how much people can understand the meaning of the converted speech. Listeners evaluate the given speech in scale 1-5. The higher the score, the more content that can be understood by evaluators.

We have 7 different systems for subjective evaluation: Source speech, GMM system, CLDNN system, BN-MEL system, BN-BN-MEL system, BN-BN-MEL-P system and Target speech. Each system has 33 sentences. Therefore, there are 231 utterances in total for evaluation. Each evaluator rates every chosen sentence regarding naturalness, similarity, and intelligibility.

Figure 4 shows the MOS results. Regarding both the naturalness and intelligibility scores, the baseline systems, which are the GMM system, the CLDNN system and the BN-MEL system, have significant improvement compared with the source EL speech. Furthermore, our proposed BN-BN-MEL system outperforms the GMM system and the BN-MEL system on both naturalness and intelligibility. The CLDNN system achieves a relatively high intelligibility s-core of around 3.51. However, its enhancement performance on naturalness is not good as the MOS is lower than 3. On the other hand, by utilizing the data augmentation and transfer learning strategies, our BN-BN-MEL-P system

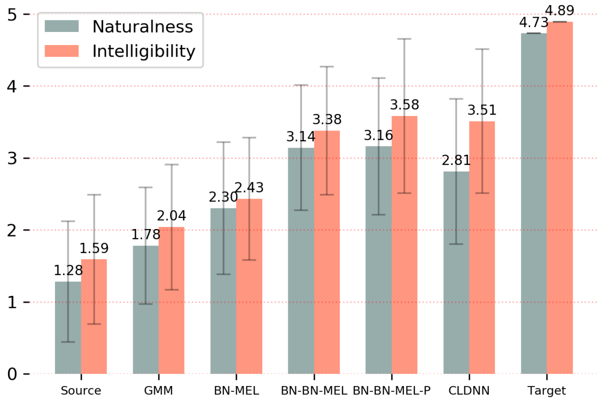
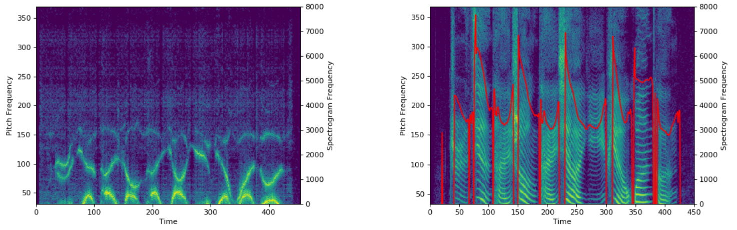


Figure 4: The MOS of the naturalness and intelligibility of speech samples from systems including Source, GMM, CLDNN, BN-MEL, BN-BN-MEL, BN-BN-MEL-P and Target, which denote the original EL speech, the enhanced speech from the GMM-based VC, the enhanced speech from the CLDNN system, the enhanced speech from the BN-MEL system, the enhanced speech from the BN-BN-MEL system, the enhanced speech from BN-BN-MEL-P system and the original parallel normal speech, respectively.

achieves further improvement concerning naturalness and intelligibility, which outperforms all other systems.

Figure 5 shows the spectrogram and the F0 contour of EL speech and normal speech. Figure 6 shows the spectrogram and the F0 information extracted from different enhancement systems. For Mandarin Chinese, each syllable contains one of four basic tones (plus a fifth, neutral one) that utilize F0 to differentiate the meaning of words with the same sound pattern. Pitch change is mainly determined by the fundamental frequency generated by the vibration of the vocal cords. We use the WORLD vocoder to extract the F0 of the converted speech and plot it with the spectrogram in the same figure.

The pinyin of the selected utterance is “zhi3 jian4 tai2 shang4 tai2 xia4 yi2 pian4 huan1 teng2”. As observed in the F0 curve, it can be seen that the F0 of



(a) original EL speech

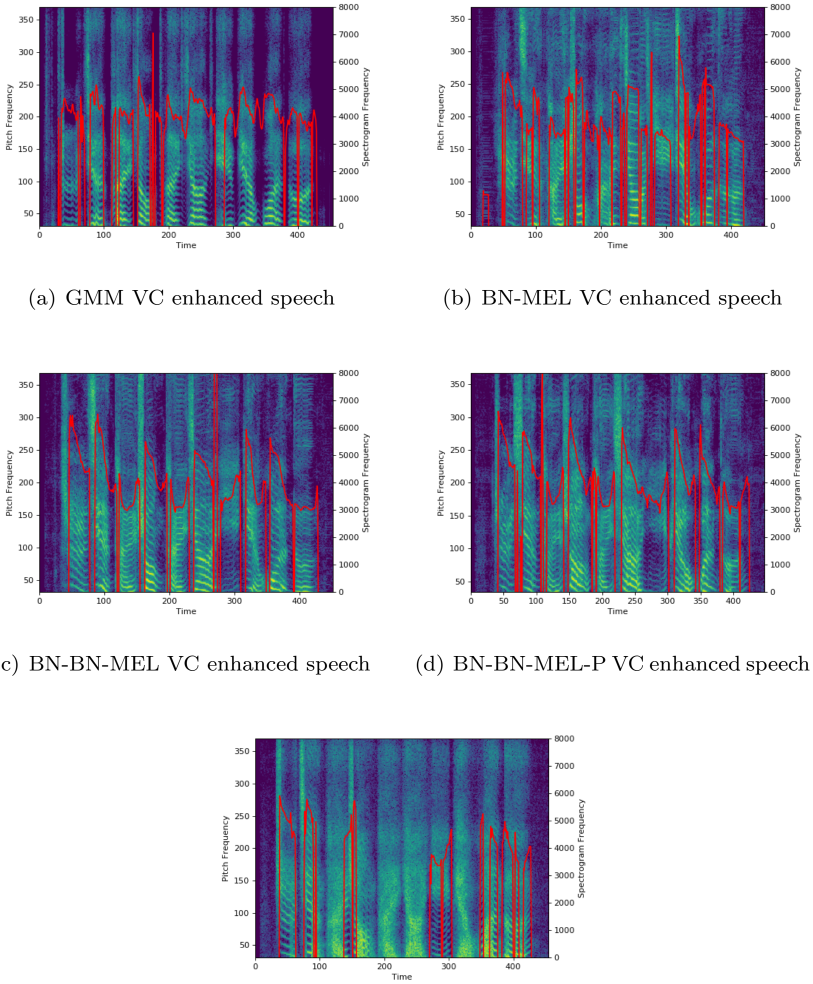
(b) original parallel normal speech

Figure 5: The extracted spectrogram and F0 contour of the original EL speech and the parallel target speech. Spectrogram and F0 are calculated using librosa and pyworld packages in Python, respectively. Text pinyin: zhi3 jian4 tai2 shang4 tai2 xia4 yi2 pian4 huan1 teng2

the original EL speech cannot be extracted. The F0 of the enhanced speech from the GMM system does not vary much, while the F0 of the enhanced speech from the BN-BN-MEL system is closer to the F0 contour of parallel normal speech. The F0 contour of the enhanced speech from the CLDNN system cannot be extracted well due to its hoarse voice. We also can observe that the enhanced speech generated by the BN-BN-MEL-P system has a very similar F0 contour to the target one.

3.6. Experimental Results with Limited Data

Our proposed systems require a parallel dataset for training. We have a corresponding normal utterance with the same content for each EL utterance. However, such parallel data is difficult and costly to collect. In this case, we conduct experiments with the BN-BN-MEL-P system finetuned by different amounts of parallel EL speech utterances to investigate the data we need for training. In particular, we randomly select 1000, 100, 50, and 10 utterance pairs from the original EL speech dataset and use the selected utterances to finetune the pre-trained BN-BN-MEL-P system. We conduct the same ASR test, MSD test and subjective evaluation on the BN-BN-MEL-P systems trained



(a) GMM VC enhanced speech

(b) BN-MEL VC enhanced speech

(c) BN-BN-MEL VC enhanced speech

(d) BN-BN-MEL-P VC enhanced speech

(e) CLDNN enhanced speech

Figure 6: The extracted spectrogram and F0 contour of the enhanced speech by different systems. Spectrogram and F0 are calculated using librosa and pyworld packages in Python, respectively. Text pinyin: zhi3 jian4 tai2 shang4 tai2 xia4 yi2 pian4 huan1 teng2

with different sizes of our EL speech dataset.

3.6.1. The ASR and MSD Test

We utilize the same ASR model mentioned in Section 3.4 to evaluate the intelligibility of the synthesized speech. The WER of the BN-BN-MEL-P system trained with different sizes of our EL dataset is shown in Table 5. The abrupt decrease of WER from the system trained with 10 utterances to the one trained with 50 utterances indicates that the amount of parallel EL speech data we need to have acceptable performance is around 50 utterances, which is about 5 minutes. In addition, the system trained with 1000 utterances achieves a WER of 41.36%, which shows that a parallel EL dataset of around 2 hours is good enough for training the BN-BN-MEL-P system. As shown in Table 6, the MSD value decreases as the data used for training increases. We can observe that the BN-BN-MEL-P system trained with only 1000 EL utterances can achieve comparable WER and MSD as the system trained with 2900 utterances.

Table 5: The ASR performance of the enhanced speech generated by BN-BN-MEL-P systems

Size of training data	WER(%)
The whole training set (2900 utterances)	42.62
1000 utterances	41.36
100 utterances	43.29
50 utterances	50.23
10 utterances	90.55

3.6.2. The MOS Score

Figure 7 shows the MOS results for BN-BN-MEL-P systems trained with different sizes of EL dataset. The results show that as the size of the EL speech dataset is reduced, the naturalness scores of generated speech reduce approximately linearly but in a minimal range. The decrease in the amount used for

Table 6: The Average MSD Values on the enhanced speech with the BN-BN-MEL-P system. CI denotes confidence interval.

Size of Datasets	MSD Value	CI ($\alpha = 0.05$)
2900 utterances	9.809	[8.188, 11.429]
1000 utterances	10.286	[8.696, 11.875]
100 utterances	11.031	[9.637, 12.424]
50 utterances	11.263	[9.822, 12.703]
10 utterances	12.779	[11.283, 14.274]

training will not lead to significant degradation in terms of the intelligibility scores when the training size exceeds 100 utterances.

However, when the dataset size is smaller than 100 utterances, decreasing the amount of data will profoundly affect the intelligibility performance. Moreover, even for the system finetuned with only 10 utterances, the intelligibility score, which is 2.94, is higher than the BN-MEL system, which is 2.43, as shown in Figure 4. Such results indicate the robustness and effectiveness of the BN-BN-MEL-P system in limited data scenarios.

4. Conclusions

This paper proposes a voice conversion-based EL speech enhancement system, which utilizes a parallel non-autoregressive model with bottleneck features as input. The bottleneck feature is extracted by a speaker-independent ASR model and considered as a linguistic representation. While adopting our proposed conversion model, we develop several EL enhancement systems, including the BN-MEL system that directly converts the bottleneck feature of EL speech to Mel-spectrogram of normal speech, the BN-BN-MEL system that uses the bottleneck feature of normal speech as an intermediate conversion feature. We also apply data augmentation and transfer learning strategies to enhance conversion performance. The objective and subjective evaluation conducted on our proposed systems and baseline systems show that our BN-BN-MEL-P system

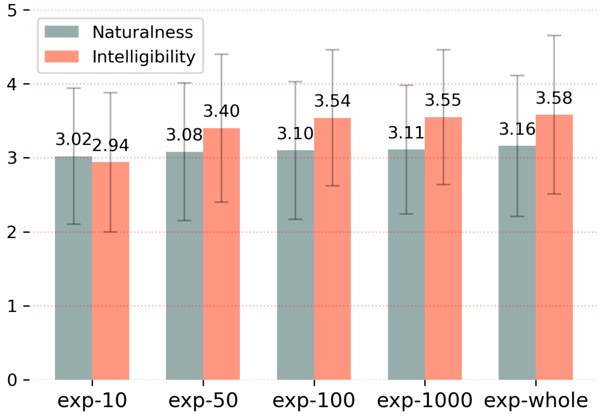


Figure 7: The MOS of the naturalness and intelligibility of speech samples from BN-BN-MEL-P systems finetuned with various size of data. exp-10 denotes the system finetuned with 10 utterances of EL speech, exp-50 is for the one with 50 utterances, exp-100 is for 100 utterances, exp-1000 is for 1000 utterances and exp-whole is for the system finetuned with the whole EL dataset (2900 utterances).

achieves impressive performance on naturalness and intelligibility and outperforms all other systems. In addition, our experiments show that the two-stage mapping method is more efficient and effective than directly converting the bottleneck feature of EL speech to the Mel-Spectrogram of corresponding normal speech. Furthermore, the augmentation by simulating EL speech from normal speech and transfer learning help improve the performance and work well on limited data scenarios. Even though the proposed method effectively improves the naturalness and intelligibility of the EL speech, there is still a notable difference between the enhanced EL speech and normal speech. In future works, we will study the zero-shot voice conversion scenario and use a multiple speakers' EL speech dataset to train a speaker-independent voice conversion model for EL enhancement.

5. Acknowledgements

This research is funded by the Kunshan Government Research Fund. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

Appendix A. Abbreviation Table

Abbreviation	Explanation
EL	Electrolarynx
DATE	dimensional amplitude trimmed estimatio
F0	fundamental frequency
NMF	non-negative matrix factorization
GMM	Gaussian Mixture Models
PPP	phonetic posterior probabilities
VC	voice conversion
BN-MEL	bottleneck feature to Mel-spectrogram
BN-BN-MEL	EL speech's bottleneck features to bottleneck features of normal speech to Mel-spectrogram of normal speech
WER	Word Error Rate
MOS	Mean Opinion Score
ASR	automatic speech recognition
TDNN	time-delayed neural network
MSE	Mean Square Error
MSD	Mel-spectrogram Distortion
BN-BN-MEL-P	BN-BN-MEL system with pre-training strategy

References

[1] A. K. Fuchs, M. Hagmüller, G. Kubin, The New Bionic Electro-larynx Speech System, IEEE journal of selected topics in signal processing 10 (5) (2016) 952–961.

[2] H. Liu, M. L. Ng, Electrolarynx in Voice Rehabilitation, Auris Nasus Larynx 34 (3) (2007) 327–332.

[3] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Speaking-aid system-s using gmm-based voice conversion for electrolaryngeal speech, Speech Communication 54 (1) (2012) 134–146.

[4] K. Tanaka, T. Toda, G. Neubig, S. Sakti, S. Nakamura, A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical Excitation Generation, IEICE Transactions on Information and Systems E97.D (6) (2014) 1429–1437.

[5] N. Uemi, T. Ifukube, M. Takahashi, J. Matsushima, Design of A New Electrolarynx Having A Pitch Control Function, in: Proceedings of 1994 3rd IEEE International Workshop on Robot and Human Communication, pp. 198–203.

[6] K. Tanaka, T. Toda, G. Neubig, S. Sakti, S. Nakamura, An Inter-speaker Evaluation through Simulation of Electrolarynx Control Based on Statistical F0 Prediction, in: 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–4.

[7] C. Y. Espy-Wilson, V. R. Chari, J. M. Macauslan, C. B. Huang, M. J. Walsh, Enhancement of Electrolaryngeal Speech by Adaptive Filtering, Journal of Speech Language and Hearing Research 41 (6) (1999) 1253–1264.

[8] H. Liu, Q. Zhao, M. Wan, S. Wang, Enhancement of Electrolarynx Speech Based on Auditory Masking, IEEE Transactions on Biomedical Engineering 53 (5) (2006) 865–874.

[9] L. R. Mathew, K. Gopakumar, Evaluation of Speech Enhancement Algorithms Applied to Electrolaryngeal Speech Degraded by Noise, Applied Acoustics 174 (2021) 107771.

[10] M. Eshghi, K. Tanaka, K. Kobayashi, H. Kameoka, T. Toda, An Investigation of Features for Fundamental Frequency Pattern Prediction in Electrolaryngeal Speech Enhancement, in: Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10), 2019, pp. 251–256.

[11] M. Eshghi, K. Kobayashi, K. Tanaka, H. Kameoka, T. Toda, Phoneme Embeddings on Predicting Fundamental Frequency Pattern for Electrolaryngeal Speech, in: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp. 572–577.

[12] K. Kobayashi, T. Toda, Electrolaryngeal Speech Enhancement with Statistical Voice Conversion Based on CLDNN, in: 2018 26th European Signal Processing Conference (EUSIPCO), IEEE, pp. 2115–2119.

[13] K. Kobayashi, T. Toda, Implementation of Low-latency Electrolaryngeal Speech Enhancement Based on Multi-task CLDNN, in: 2020 28th European Signal Processing Conference (EUSIPCO), IEEE, pp. 396–400.

[14] Z. Cai, Z. Xu, M. Li, F0 Contour Estimation Using Phonetic Feature in Electrolaryngeal Speech Enhancement, in: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6490–6494.

[15] K. Tanaka, H. Kameoka, T. Toda, S. Nakamura, Statistical F0 Prediction for Electrolaryngeal Speech Enhancement Considering Generative Process of F0 Contours within Product of Experts Framework, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5665–5669.

[16] K. Tanaka, H. Kameoka, T. Toda, S. Nakamura, Physically Constrained Statistical F0 Prediction for Electrolaryngeal Speech Enhancement, in: Proc. Interspeech 2017, pp. 1069–1073.

[17] M. Li, L. Wang, Z. Xu, D. Cai, Mandarin electrolaryngeal voice conversion with combination of gaussian mixture model and non-negative matrix factorization, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1360–1363.

[18] H. Kawahara, I. Masuda-Katsuse, A. De Cheveigne, 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 communication 27 (3-4) (1999) 187–207.

[19] M. Morise, F. Yokomori, K. Ozawa, WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications, IEICE Transactions on Information Systems E99 (7) (2016) 1877–1884.

[20] an evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement.

[21] L.-H. Chen, L.-J. Liu, Z.-H. Ling, Y. Jiang, L.-R. Dai, The USTC System for Voice Conversion Challenge 2016: Neural Network Based Approaches for Spectrum, Aperiodicity and F0 Conversion, in: Proc. Interspeech 2016, pp. 1642–1646.

[22] I. B. Othmane, J. Di Martino, K. Ouni, Enhancement of esophageal speech using voice conversion techniques, in: International Conference on Natural Language, Signal and Speech Processing (ICNLSSP), 2017.

[23] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, ACVAE-VC: Non-parallel Voice Conversion with Auxiliary Classifier Variational Autoencoder, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (9) (2019) 1432–1443.

[24] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, T. Toda, Non-Parallel Voice Conversion with Cyclic Variational Autoencoder, Proc. Interspeech 2019 674–678.

[25] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, StarGAN-VC: Non-parallel Many-to-many Voice Conversion Using Star Generative Adversarial Networks, in: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 266–273.

[26] S. Lee, B. Ko, K. Lee, I.-C. Yoo, D. Yook, Many-to-many Voice Conversion Using Conditional Cycle-consistent Adversarial Networks, in: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (I-CASSP), pp. 6279–6283.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, in: 9th ISCA Speech Synthesis Workshop, 2016, pp. 125–125.

[28] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, A. C. Courville, MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, Advances in Neural Information Processing Systems 32.

[29] J. Kong, J. Kim, J. Bae, HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, in: Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 17022–17033.

[30] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, L.-R. Dai, Sequence-to-Sequence Acoustic Modeling for Voice Conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (3) (2019) 631644.

[31] H. Zheng, W. Cai, T. Zhou, S. Zhang, M. Li, Text-independent Voice Conversion Using Deep Neural Network Based Phonetic Level Features, in:

2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2872–2877.

[32] S. Ding, G. Zhao, R. Gutierrez-Osuna, Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition, in: Proc. Interspeech 2020, pp. 776–780.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi Speech Recognition Toolkit, in: IEEE 2011 workshop on automatic speech recognition and understanding.

[34] F. Grézl, M. Karafiát, S. Kontar, J. Cernocký, Probabilistic and Bottle-Neck Features for LVCSR of Meetings, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 757–760.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[36] attention is all you need.

[37] Y. Shi, H. Bu, X. Xu, S. Zhang, M. Li, AISHELL-3: A Multi-Speaker Mandarin TTS Corpus, in: Proc. Interspeech 2021, pp. 2756–2760.

[38] J. Du, X. Na, X. Liu, H. Bu, Aishell-2: Transforming Mandarin ASR Research into Industrial Scale, arXiv preprint arXiv:1808.10583.

[39] R. K. Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian, T. Toda, Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions, in: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, pp. 99–120.