Investigation of Vowel Generation Method in Low-resource Pathological Voice Database

Jinyang Qian, Yihua Zhang, Denghuang Zhao, Xiaojun Zhang, Yishen Xu, and Zhi Tao

Abstract—**The number of voices in commonly used pathological voice databases such as Massachusetts Eye and Ear Infirmary (MEEI) and Saarbruecken Voice Database (SVD) is insufficient and imbalanced. This may make the classification results lack credibility and robustness. Our works investigate the direct generation of normally pitched vowel /:a/ to expand the low-resource pathological voice database. A framework for generating vowels of different lengths using improved WaveNet and Generative Adversarial Network (GAN) is proposed in this work. Long and short vowel segments can be generated by an improved WaveNet model and a model based on Stationary Wavelet Transform and Wasserstein GAN with gradient-penalty (SWT-WGANGP) in our framework respectively. The generated voice segments are added to the original imbalanced database to improve the classification performance. Besides, we propose the accuracy-score and the diversity-score to evaluate the generated voice. By adding our generated data based on the traditional classification pipeline, accuracy on the two databases increased by 1.81% and 4.38% respectively, and the recall has achieved more than 10% improvement. Compared with the other data generation method, our method improves the classification results the most while using the same feature and classifier. Besides, using our proposed framework based on existing advanced pathological voice detection methods can further improve performance. Our results show that deep generative models with optimized structure can be used for direct vowel generation in low-resource pathological voice databases to expand the raw database, which has been ignored in previous research. Our framework can be used as a generic pre-processing module to improve the detection of pathological voice.**

Index Terms—**pathological voice, data augmentation, improved WaveNet, SWT-WGANGP**

I. Introduction

Voice is a primitive natural tool for human communication. However, around 25% of the world population suffers from pathological voice [1]. Hence, many researchers [2] have investigated pathological voice detection or voice restoration [3], voice production system modeling, and so on. Most of the reported breakthroughs [4]-[6] are, however, based on some low-resource pathological voice databases, such as Massachusetts Eye and Ear Infirmary (MEEI) database, Saabruechen Voice Database (SVD), Arabic Voice Pathology Database (AVPD) and Voice ICar fEDerico II Database (VOICED). In these databases, the number of voice recordings is imbalanced and insufficient. For example, SVD, a comparative big database of pathological voice, has 687 healthy voice recordings of normal pitch vowel /:a/ and 1354 pathological recordings with a duration of 1-3 seconds. This is problematic for some data-consuming works where more labelled data should be involved.

A natural solution to the problems mentioned above is to collect more labelled voice recordings. But the fact is that the collection of voice data is constrained by many factors. For example, MEEI database is recorded in a professional voice recording environment. Besides, the data annotation is quite a time, labor, and money-consuming task. Hence, it is difficult to collect more voice recordings directly.

Some works [7], [8] used an indirect method to generate more data by over-sampling the features extracted from raw data. Since feature extraction is an essential procedure in traditional pathological voice detection methods, generating extra features is reasonable. Other works [9] used a generative model called Generated Adversarial Network (GAN) [10] to generate extra data to train classifiers. The data augmentation [11] has also been used to create synthetic features. However, these methods all have a drawback which they can only handle with the features extracted from raw voice. Compared with generating more synthesised features of raw voice, direct generation of voice has more scenes to be used.

To the best of our knowledge, this is the first attempt to directly expand the pathological voice database by synthesizing voice, so we refer to methods from much more than the pathological voice field. As we know, speech can be directly generated by voice conversion methods [12] or voice synthesis [13]-[16] methods in other fields such as Text-To-Speech Synthesis (TTS). Some previous works [17], [18] conversed voice from specific features using voice conversion methods. For example, Gaussian Mixture Model (GMM) was used to transform healthy vowel /a:/ to pathological vowel /a:/ in the work of Amara et al. [17]. The Line Spectrum Pair feature was used to repair the vowel of pathological voice in the work of Zhang et al. [18]. Due to the big development of deep generative models in recent years, there also exist some works using deep generative models for voice conversion. YiShan Jiao et al. [19] used GAN to convert a healthy phrase into a dysarthric phrase to overcome the problem of the lack of dysarthric speakers. Except for the conversion of vowels and phrases, the

Manuscript received May 19, 2022; revised December 31, 2022.

This research was funded by the National Natural Science Foundation of China under Grant No. 61271359.

Jinyang Qian is a postgraduate student of Soochow University, Suzhou 215000, China. (e-mail: jyqian20@stu.suda.edu.cn).

Yihua Zhang is a postgraduate student of Soochow University, Suzhou 215000, China. (e-mail: 20204239034@stu.suda.edu.cn).

Denghuang Zhao is a postgraduate student of Soochow University, Suzhou 215000, China. (e-mail: 20214239035@stu.suda.edu.cn).

Xiaojun Zhang is an associate professor of Soochow University, Suzhou 215000, China. (e-mail: zhangxj@suda.edu.cn).

Yishen Xu is a professor of Soochow University, Suzhou 215000, China. (Corresponding author, e-mail: xys2001@suda.edu.cn).

Zhi Tao is a professor of Soochow University, Suzhou 215000, China. (Corresponding author, e-mail: taoz@suda.edu.cn).

conversion of sentence has also aroused the interest of many researchers. In TTS field, Berrak Sisman et al. [20] used Mel-spectrograms followed by a WaveNet [21] vocoder to utilize sequence-to-sequence conversion with different speakers. But such works are hard to apply in our pathological voice field because they need too many voice recordings for training. As for voice synthesis methods associated with pathological voice databases, we don't find any strongly related works, but there are many similar works in other fields such as TTS and pronounce model. TTS methods usually use some vocoders to transform acoustic features into waveforms. STRAIGHT [13] and WORLD [14] are two conventional vocoders. With the development of deep learning, many neural vocoders which are used for voice generation have demonstrated better performance than conventional vocoders. In [15], a neural source-filter waveform model was proposed for TTS. The model depended on the source-filter model of voice generation, using a neural network as a filter to generate voice. Yang et al. [16] proposed HiNet vocoder which considers both the amplitude spectrum and phase spectrum for better naturalness of the voice. Angrick et al. [22] directly synthesize speech from electrocorticography. Cataldo et al. [23] designed a stochastic model of voice generation. The model can generate different voiced sounds by different parameters of the stochastic model.

Inspired by the problems in the pathological voice field and the works mentioned above, we investigate the feasible method to directly generate voice segments from limited pathological voice recordings. The main goal of this work is to generate the voice of the vowel /a:/ directly with different lengths and to objectively evaluate the generated data. In previous works, every vowel recording can be used as a whole input or divided into many frames to extract different features. For example, the statistical features [24], spectrograms, and formants [2] are extracted from a whole vowel recording. The zero crossing rate (ZCR) [25], one order Mel Frequency Cepstral Coefficient (MFCC) [26], and one order Gammatone Frequency Cepstral Coefficient (GFCC) [27] are extracted from every frame. This is the reason why we want to generate vowel with different lengths.

In this work, we propose a framework in which an improved WaveNet model is used for long voice segment generation and a GAN-based model is used for short voice segment generation. Despite there being little similar work for the same purpose of this work, WaveNet has been widely used in other fields such as the generation of electrocardiograms [28], the emulation of the sound of an audio processor reference device [29], and the fast simulation of seismic waves [30]. However, we find it difficult to train the raw WaveNet model using our database, so we improve it by changing the network structure. We simplify the residual block in the raw WaveNet model and change the skip-connection in different layers. Because the voice is usually framed into hundreds of points, and the WaveNet model is not suitable for modeling such short voice segments, we also investigate the method for short voice segment generation. The GAN has been widely used in recent works [31], [32] in the field of figure generation. It can also be used in voice conversion [19] and emotional speech generation [33] by training a GAN which can generate the mel-cepstral features of specific emotional speech. We propose a Stationary Wavelet Transform and Wasserstein GAN with gradient-penalty (SWT-WGANGP) model to generate the short voice segments. This model is a two-stage model inspired by the voice conversion methods we mentioned above. We first transform the waveform to SWT coefficients. Then GAN is used to generate new SWT coefficients. Inverse Stationary Wavelet Transform (ISWT) is used to generate voice segments using the generated new SWT coefficients. After the generation of long and short voice segments, they are added to the raw imbalanced databases to improve the classification results. In addition, we propose a quantitative method called accuracy-score and diversity-score to evaluate the performance of generated data.

The rest of this paper is structured as follows. In section II, we detail the proposed framework based on the improved WaveNet and SWT-WGANGP. In this section, the databases and the evaluation methods we proposed will also be presented. In section III, the experimental setup will be described. The experimental results will be summarized in section IV. In section V, we will give a discussion and analysis of our methods and experiments compared with other methods. In section VI, we will conclude our work.

II. Materials and Methods

The whole architecture of the proposed voice augmentation method is shown in Fig. 1. Two networks are used for direct voice generation with different lengths, which is inherited from our previous works. We improve the WaveNet model for the generation of long voice segments and a GAN-based model for the generation of short voice segments. After the generation of long and short voice segments, they are both used for imbalance testing and evaluation. In the imbalanced test, the generated data are added to the raw imbalanced datasets to improve the classification results. In the evaluation, the accuracy and diversity of the generated data will be presented using two scores. In the improved WaveNet, we first transform the data and then quantize it into 256 possible values with one-hot encoding. The one-hot encoding is then sent to the improved WaveNet model followed by a decompression operation. In the model of SWT-WGANGP, we first enframe the raw waveforms to clip them into a relatively short length. Then, we transform the waveforms to get the coefficient arrays. The coefficient arrays are then fed into GAN for generating new coefficient arrays. The newly generated coefficient arrays will finally be inverse-transformed to new voice segments.

A. Databases

The databases used in this work were MEEI database and SVD. These two databases are commonly used in voice pathology detection. The MEEI database is commercially available and SVD is open source. The number of voices in MEEI is less than SVD. There are 53 healthy vowel /:a/ and 657 pathological vowel /:a/ in MEEI. As for SVD, there are more than 2000 speakers in it, which contains 687 healthy vowel /:a/ and 1354 pathological vowel /:a/ from these speakers. The environment for recording SVD is not as

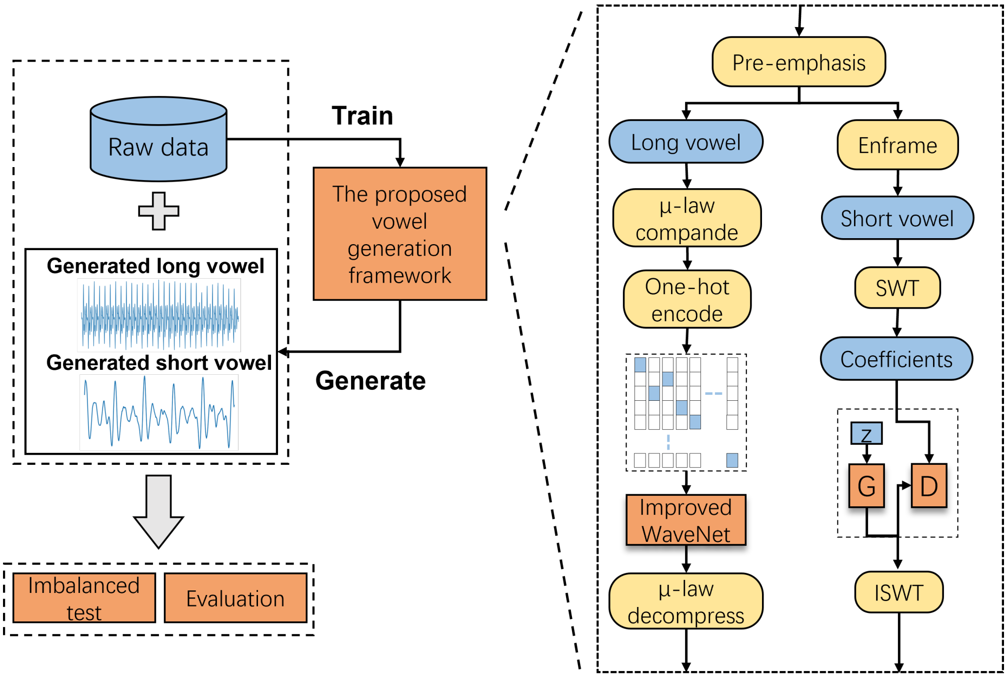


Fig. 1. The architecture of the whole voice generation method.

strict as MEEI's. Part of the databases is shown in Table I, which contains the healthy vowel /:a/ and four common types of pathological vowel /:a/. Different types of voices in the two databases are shown in Fig. 2.

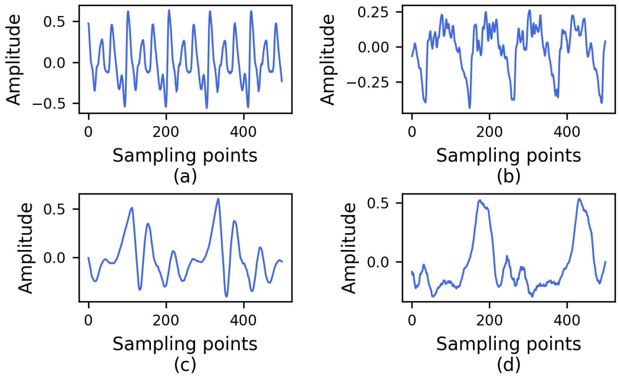


Fig. 2. different types of vowel /:a/: (a) healthy vowel in MEEI, (b) pathological vowel in MEEI, (c) healthy vowel in SVD, (d) pathological vowel in SVD.

TABLE I
THE NUMBER OF DIFFERENT TYPES OF VOICES IN MEEI AND SVD

Databases	Healthy	Pathological			
		Nodule	Polyp	Cyst	Paralysis
MEEI	53	19	20	43	67
SVD	687	17	48	68	64

We can see that the number of different types of voices is quite imbalanced. Hence, there are always obvious differences in sensitivity and specificity when some traditional features are used for classification experiments [26]. This phenomenon implies a problem: Since the number of different samples is imbalanced, even if the classifier judges all samples as the majority samples, it can obtain an accuracy of more than 50%, but this classification result is of no value. This will be important when the pathological voices are in the minority classes because we do not want to classify a pathological voice as a normal voice. Besides, the number of samples in the pathological voice database is much less than that in other fields because of the difficulty in establishing the database. Our works are very meaningful since we direct expand the raw database with generated vowels. We used 53 healthy voices and 149 pathological voices in MEEI. In SVD, we have just selected a random part of the data to train the improved WaveNet for the generation of long voice segments due to the computational complexity when using data with 50 kHz. When we train the SWT-WGANGP, 53 healthy voices, 149 pathological voices in MEEI, 687 healthy voices, and 1290 pathological voices in SVD are used.

B. Improved WaveNet Model for The Generation of Long Vowel Segments

WaveNet is a raw audio generation model with complex architecture. It can extract the long-term correlation of voice signals. However, the training of the raw WaveNet is very difficult [34] especially using our databases with limited data. We improve the raw WaveNet by simplifying the two different activation gates in every residual block and only using skip-connection inside the residual block for our purpose. We improved the network by this way because the raw different activation gates are responsible for the shape and amplitude of the generated waveform respectively, but the amplitude of the waveform is limited to the range of -1 to 1 in our methods, so the tanh activation gate is s redundant for our purpose theoretically. The improved WaveNet is also an autoregressive generative model that can model complex distributions of raw audio. The model can predict the next sample from all the input, which means the probability of the next sample can be described as follows:

$$p(x) = \prod_{T}^{t=1} p\left(x_t \middle| x_1,\ldots,x_{t-1}\right) \tag{1}$$

where the *p(x)* represents the probability value of the next sample. The model is an autoregressive model in which each sample x_t is conditioned on all past samples.

Before being fed into the network, the data first needs to be preprocessed. μ-law companding and one-hot encoding are used for the preprocessing of the input data. The audio data in our databases are stored in 16-bit integer values for every point. If we predict the value directly, the output should have 65536 probabilities. μ-law companding can reduce the dynamic range of an audio signal and reduce the quantization error. After the μ-law companding, we quantize the data into 8-bit integer values and apply one-hot encoding to the values. Hence, the network can predict the value of the next point by classification task.

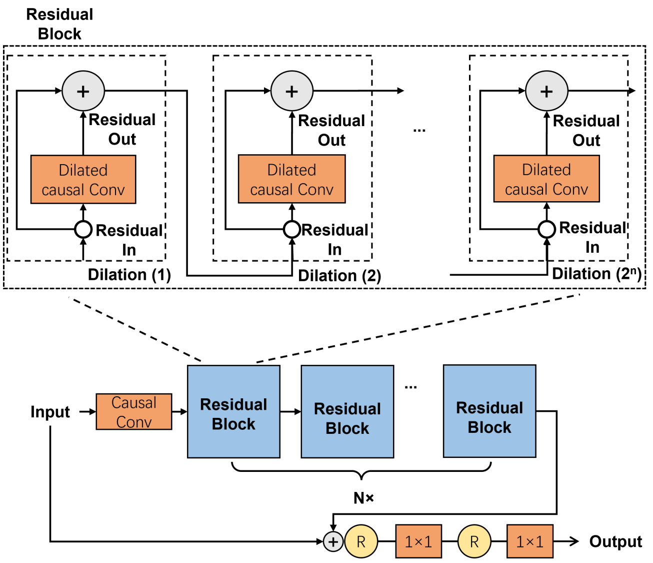


Fig. 3. The architecture of residual block and improved WaveNet.

The residual block is the basic component of the improved WaveNet. The architecture of the residual block is shown in the upper part of Fig. 3. Every residual block is composed of many sub-layers with different dilation rates. For example, the first to tenth sub-layers have 1, 2, 4, ..., 512 dilation rates respectively. The input of every sub-layer is the output of the previous sub-layer. Every sub-layer has a gated activation unit:

$$z = W_{f,k} * x + x \tag{2}$$

where x denotes the input data, k represents the layer number, $W_{f,k}$ denotes the learnable parameters, and * represents a convolution operator. The addition of the two parts is the residual operation. This is much simpler than in the original WaveNet model, and it is more reasonable for our task since all data is normalized in the first step.

The whole architecture of the improved WaveNet is shown in the lower part of Fig. 3. The input data is first fed into a causal convolution layer. It means the outputs of the layer are all conditioned on the previous points. Then, the residual block is repeated N times one by one. Every residual block has the same architecture. The output of the last residual block will be added to the input of the whole network. The R represents the relu activation function. The symbol *1x1* in the figure represents a one-dimensional convolution with a kernel size of 1. The dilated convolution and residual operation can speed up the convergence and can avoid the vanishing of the gradient.

The input of the improved WaveNet model is two-dimensional data, in which the first dimension represents the length of the voice segment and the second dimension represents the one-hot encoding of every point. The output of the improved WaveNet model is the probability of the next point of the input. The probability conforms to the softmax distribution.

C. SWT-WGANGP Model for The Generation of Short Vowel Segments

The improved WaveNet is an end-to-end model in our experiments. However, in pursuit of product quality and length of the vowel /:a/, the time consumption is very high. Because the vowel signals have short-term stability, and many features are extracted from every frame of the vowels, it is reasonable to generate such vowels, especially for our investigation purpose.

GAN can generate waveforms in a one-shot manner. The core idea of GAN is an equilibrium of the game theory. GAN consists of two parts, a generator G and a discriminator D. The purpose of the G is to generate data distribution similar to the original data distribution, and the D is responsible for distinguishing the data generated by the G from the real data. The input of G is random variables z, and the input of D is real data X and generated data G(x). The loss function of G and D are written as:

$$L_D(x) = -E_{x\sim P_r}[\log D(x)] - E_{x\sim P_g}[\log(1 - D(x))] \tag{3}$$

$$L_G(x) = E_{x\sim P_g}[\log(1 - D(x))] \tag{4}$$

where $x{\sim}P_r$ means the x is real data, and $x{\sim}P_g$ means the x is generated data. The *D(x)* is the output of the discriminator. When training the discriminator, if x conforms to the distribution of P_r, the D needs to judge x as a positive sample as much as possible, which means the value of *D(x)* should be 0. Correspondingly, when x conforms to the distribution of P_g, the value of *D(x)* should be 1. When training the generator, the G should deceive the D as much as possible, so the loss of G should be contrary to D. The two parts begin to fight with each other and find the equilibrium according to G's ability and D's ability. The loss function of G can also be written as:

$$L_G(x) = -E_{x\sim P_g}[\log(D(x))] \tag{5}$$

where all parameters are the same in Equation 4. In this formula, if the G is good enough, the value of *D(x)* should be close to 1, which means the loss of G is 0.

The above GAN model also has some drawbacks. Due to the neural networks having strong capabilities of data fitting and classification capabilities, the training of D is always easier than G. When the D is the optimal D, calculating the minimum value of Equation 4 and 5 can be equivalent to calculating the minimum value of Equation 6 and 7 respectively.

$$L_G(x) = 2JS\left(P_r \| P_g\right) - 2\log 2 \tag{6}$$

$$L_G(x) = KL\left(P_r \| P_g\right) - 2JS\left(P_r \| P_g\right) \tag{7}$$

where $JS(P_r||P_g)$ denotes Jensen-Shannon (JS) divergence of

P_r and P_g, $KL(P_r||P_g)$ denotes Kullback–Leibler (KL) divergence of P_r and P_g. In Equation 6, we can calculate $JS(P_r||P_g)$ is log 2 when P_r and P_g are two completely non-overlapping distributions, which means G can't get any gradient to optimize itself, and it is a high probability event since the output of G is a high-dimensional data while the input is relatively a low-dimensional data. In Equation 7, we can find that G is trying to minimize the KL divergence of P_r and P_g but to maximize the JS divergence of P_r and P_g, which is intuitively unreasonable. Besides, the KL divergence is not a symmetrical measure of P_r and P_g, and this will make the generated samples lack diversity. So if we use Equation 4 or 5 as the loss of G, it needs to balance the training level of the D and the G carefully.

Wasserstein GAN (WGAN) [35] has solved the problem of unstable training of GAN by using Wasserstein distance. It is no longer necessary to carefully balance the training level of the D and the G. The WGAN uses the loss functions as follows:

$$L_D(x) = -E_{x\sim P_r}\left[f_w(x)\right] + E_{x\sim P_g}\left[f_w(x)\right] \tag{8}$$

$$L_G(x) = -E_{x\sim P_g}\left[f_w(x)\right] \tag{9}$$

where f is a continuous function with the Lipschitz constraint, the w is a parameter, the fw denotes a discriminator in which the last layer is not a non-linear activation layer. The $fw(x)$ is similar to the $D(x)$ mentioned above. The other parameters are the same to the equations above. By the operation of fw, all the weights are clipped to the range of w.

However, the weight clipping operation which ensures the Wasserstein distance between Pr and Pg can be calculated can also limit the gradient of the discriminator, causing the problem of gradient vanishing. In WGAN-GP [36], the Lipschitz constraint can be realized by adding a new loss to the equation:

$$\begin{aligned} L_D(x) = -E_{x\sim P_r}[D(x)] + E_{x\sim P_g}[D(x)] + \\ \lambda E_{x\sim P_{\hat{x}}}\left[\|\nabla_x D(x)\|_p - 1\right]^2 \end{aligned} \tag{10}$$

where the first two parts are similar to Equation 8. The third part is a new loss. $||\nabla_x D(x)||_p$ represents the Lp-norm of $D(x)$. The $x \sim P_{\hat{x}}$ represents x is limited to the concentration area of the generated samples, the concentration area of the real samples, and the area between them.

Based on the WGAN-GP, we use Wavelet Transform (WT) to transform the waveform before we feed the data into the WGAN-GP. We do this based on the following two considerations. Firstly, the generated short vowel /:a/ in this way is more diverse. Secondly, it is easier to train the network in this way. We have also considered short-term Fourier transform (STFT) as the input of GAN, and reconstructed the waveform by the spectrogram just like many other methods [19], [33]. But the spectrogram extracted by STFT ignores the phase information, and can't get a satisfying waveform reconstructed by the spectrogram with Griffin-Lim algorithm [37] in our experiments. Besides, the multiresolution of WT can help us avoid the setting of the window size which is important in STFT. The WT of a signal f(t) is defined as:

$$T(a,b) = \frac{1}{a}\int_R f(t)\times y\left(\frac{t\text{-}b}{a}\right)dt \tag{11}$$

where a is the scale factor, b is the translation, and ψ represents wavelet. With a large a, we can have an overall view of the signal because of the expansion of the wavelet, and with a small a, we can have a localized and detailed view of the signal because the wavelet is shrunk in the time domain.

We chose the Stationary Wavelet Transform (SWT) [38] as the voice feature because we don't need to change the architecture of the network using the SWT coefficients. SWT can make up for the loss of translation-invariant using Discrete Wavelet Transform (DWT) due to down-sampling operation. SWT is different from DWT, mainly in that after each order of the high-pass filter and low-pass filter, the results are up-sampled to the same length of original data, instead of the down-sampling operation in DWT after the filter. We trained different GANS for the coefficients with different levels, and this makes the training of the network easier because the SWT coefficients use more data to characterize the raw data. In the reconstruction of the voice segments, inverse stationary wavelet transform (ISWT) was used for the SWT coefficients, and this makes the result diverse since the final waveform can be reconstructed with the random combination of the generated coefficients with different levels.

D.Imbalanced Test and Evaluation

Due to the imbalanced distribution of positive and negative samples in MEEI database and SVD, we have performed a classification experiment in which we use our generated data to expand the original imbalanced database into a balanced database just like some other works [7]. The results are shown in a confusion matrix like Table II.

TABLE II
THE CONFUSION MATRIX OF THE BINARY CLASSIFICATION

Actual classes	Predicted classes	
	Positive class	Negative class
Positive class	TP	FN
Negative class	FP	TN

In the Table II, TP, FP, TN and FN represent the positive sample classified as positive, the negative sample classified as positive, the negative sample classified as negative, and the positive sample classified as negative respectively.

In the imbalanced test, we use the accuracy, recall, precision and F1-score as the main objective evaluation metrics, which are defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}\times 100\% \tag{12}$$

$$Recall = \frac{TP}{TP+FN}\times 100\% \tag{13}$$

$$Precision = \frac{TP}{TP+FP}\times 100\% \tag{14}$$

$$F1\text{ - }score = 2\times\frac{Pression\times Recall}{Presion+Recall}\times 100\% \tag{15}$$

where the parameters are the same as in Table II. The Equation 13 shows the recall of positive samples since the positive samples are minority samples in our experiments.

To evaluate the performance of our methods for generating voice in a low-resource pathological voice database, we should test them in at least two aspects. The first is the similarity between the generated data and the true data. The second is whether the generated data is diverse. Different from other fields such as TTS, the human listening test is not very suitable for our task because it is hard for common listeners to distinguish between different types of vowels and the true labels of raw vowels are clinically determined. Our goal is not only to get natural-sounding vowels, but more importantly, they must have healthy or pathological characteristics. For the first aspect, we can test whether the distribution of the generated data is similar to the original data. The Time domain waveforms of the generated vowels and the original vowels will be shown to see their similarity directly. Besides, we can observe their distribution from the feature space of the generated voices and the original voices. For the objective evaluation, we take evaluation methodology called accuracy-score and diversity-score. Where accuracy-score means the accuracy of our generated healthy and pathological voices using the classifier trained with original healthy and pathological voices. The diversity-score represents the accuracy of original healthy and pathological voices using the classifier trained with our generated healthy and pathological voices. If the accuracy-score is high, we can infer that the generated data conforms to the distribution of the original data. It is similar to the Mode Score (MS) [39] and Inception Score (IS) [40] in the evaluation of GAN in other fields. If the diversity-score is high, we can consider that the generated data is diverse because the original data can be classified with high accuracy only when the generated data is diverse. The proposed accuracy-score and diversity-score can evaluate the accuracy and diversity of the generated data.

III. Experiments

The improved WaveNet for the generation of long vowel segments and SWT-WGANGP for the generation of short vowel segments were implemented by Pytorch [41] deep learning framework with Nvidia 3060Ti GPU. Our experiments were based on MEEI and SVD databases which are two commonly used databases in the pathological voice field.

A. Experiments of the Generation of Long Vowel Segments

To find the best set of parameters, we have implemented a lot of experiments. Due to the WaveNet is an autoregressive model, we made the dimensions of output as the same as input except for the length of data. For example, if the input data have a dimension of (batch size, 256, n), which means the input of WaveNet is composed of batch size pieces of data of length n. 256 is the result of one-hot encoding. The dimension of output is (batch size, 256, output length). In our experiments, the output length is 16, the n is defined as follows:

$$n = receptive\ \ field + output\, length\ \text{-}\ 1 \qquad (16)$$

where the *receptive field* is associated with the dilation rate as we have introduced in section II A. We use the teacher forcing strategy when training the improved WaveNet.

When training the improved WaveNet, the loss is calculated by the data (x_t, x_{t+1}, ……, x_{t+16}) and the output of the network. Then the x_t is added into the input to train the next series of data. When inferencing the network, we predict one point per time step.

The learning rate was set to 0.0001 based on our pilot experimental results. Cross entropy was used for the loss function, and the optimizer was Adam. We first adjusted the structure of the network by directly observing the generated waveform. Part of the experimental parameter settings and preliminary experimental results are shown in Table III.

The 'Data nums' column in Table III represents the number and category of the voices we used. For example, '1 h' represents 1 healthy voice, '149 p' represents 149 pathological voices. The 'Dilation depth' column represents the n in the residual block of Fig. 3, and the column 'Repeats' represents the N in the improved WaveNet of Fig. 3. The ‘Shuffle’ column means whether the order of input data is randomly disrupted. The 'useful' column means whether the trained network can generate different types of voices using different seeds.

In experiment 1, the network was overfitting with the trained voice, so it can’t be used for the other voices without training. In experiment 2, though we used more voice segments for training with the input in order, the network can only predict the waveform of the last trained voice well. Therefore, we inferred that the order of the input needs to be shuffled. In experiments 3 and 4, the structure of the network was adjusted, and we came to a preliminary conclusion that the size of the receptive field is about 50 to 100 milliseconds for the vowel /:a/. In experiments 6 to 8, we used MEEI database and design the network and strategy of training based on previous experiments. The data generated in experiments 7 to 10 is used for the following imbalanced test and evaluation.

B. Experiments of the Generation of Short Vowel Segments

In the experiments of SWT-WGANGP, we framed the voice to a length of 512 points. We chose 53 healthy voices and 149 pathological voices in MEEI, 687 healthy voices, and 1290 pathological voices in SVD. We down-sampled the voice to 25 kHz to get more periods at a fixed length of 512 points in all used voices. We used a rectangular window with a length of 512 and a shift of 25% for every voice. There were 30438 frames of healthy voice and 34427 frames of pathology voice in MEEI, 174012 frames of healthy voice, and 314792 frames of pathology voice in SVD. Each frame has 512 points. SWT with 2 levels was used for all frames. That means we trained 16 WGAN-GP models in the two databases. Each model took about half an hour to train, which is very efficient compared to the improved WaveNet. SWT can effectively reduce the complexity of the original data since we use 2048 points to describe the original 512 points, and this can effectively reduce the difficulty for GAN to learn the data distributions of different voices. Daubechies 1 wavelet is chosen as the mother wavelet to implement the wavelet transform. The architecture of WGAN-GP is shown in Table IV.

TABLE III
EXPERIMENTS OF THE IMPROVED WAVENET

Num	Sr(kHz)	Database	Data nums	Dilation depth	Repeats	Shuffle	Useful
1	50	SVD	1 h	10	4	No	No
2	50	SVD	10 h	10	4	No	No
3	50	SVD	10 h	10	4	Yes	No
4	50	SVD	10 h	10	2	Yes	Yes
5	25	SVD	10 h	10	2	Yes	No
6	25	MEEI	10 h	9	2	Yes	Yes
7	25	MEEI	53 h	9	2	Yes	Yes
8	25	MEEI	149 p	9	2	Yes	Yes
9	50	SVD	100 h	10	4	Yes	Yes
10	50	SVD	50 p	10	4	Yes	Yes

TABLE IV
THE ARCHITECTURE OF SWT-WAGNGP

Generator	Output Shape	Discriminator	Output shape
Noise	1 × 100	Input data	1×512
Liner + Relu	1 × 512	Liner + Relu	1×256
Liner + Relu	1 × 512	Liner + Relu	1×256
Liner + Relu	1 × 512	Liner + Relu	1×256
Liner	1 × 512	Liner	1×1

Preliminary experiments show that 100 training epochs in MEEI and 50 training epochs in SVD is suitable from the view of the generated waveforms. Hence, the training epochs are 100 and 50 in MEEI and SVD respectively. After the training of each model, we used the model to generate new data. SWT with Daubechies 1 wavelet was used to reconfigure a voice segment. In the last, we generated 53*20 healthy voice segments and 149*20 pathological voice segments in MEEI. In SVD, we also generated 20 times the number of voices in the original database.

IV. RESULTS

A. Results of The Generation of Long Vowel Segments

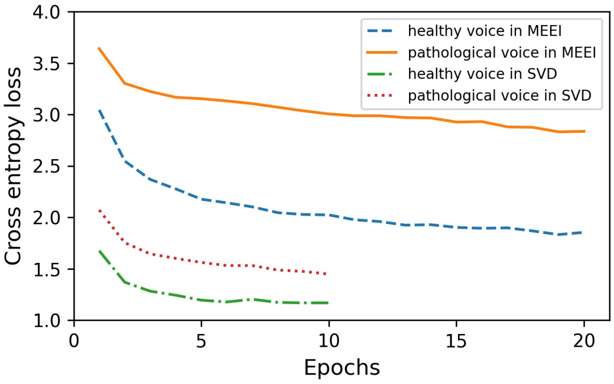


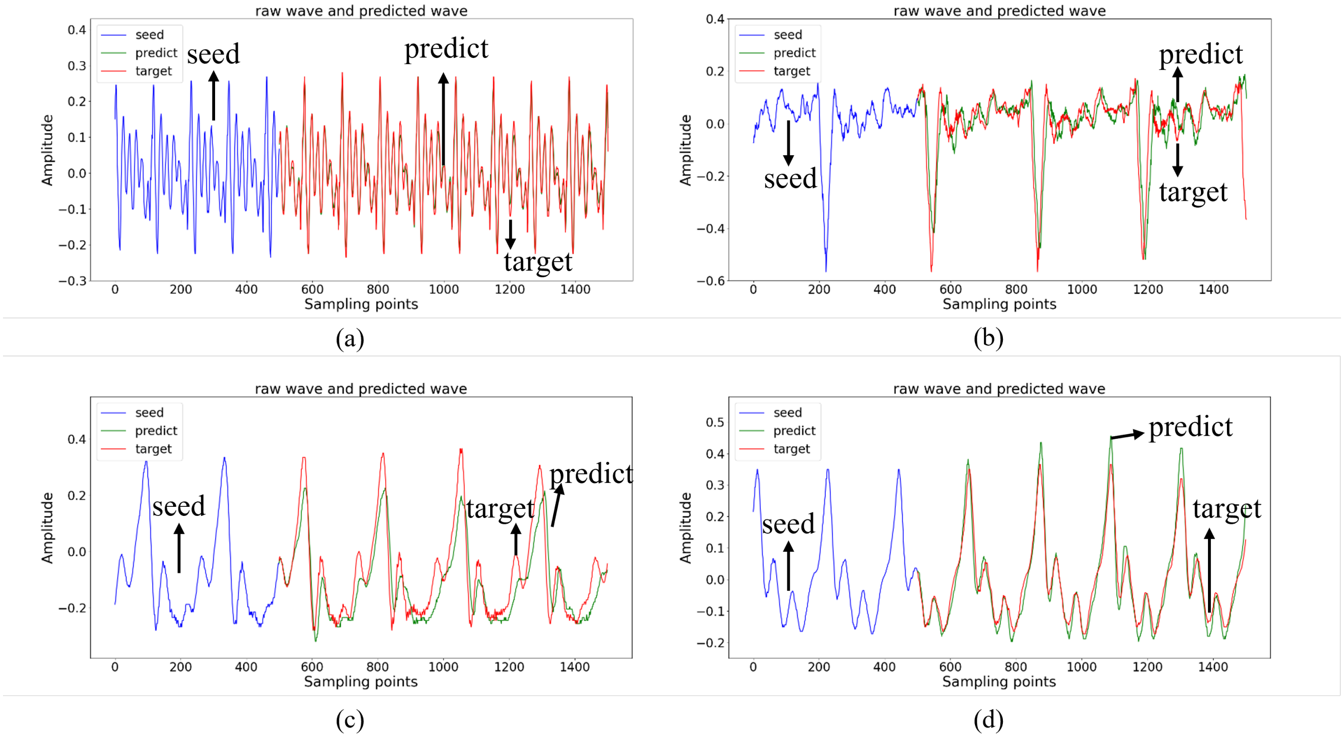
Fig. 4. The losses in experiments 7 to 10.

The Fig. 4 is the training loss in experiments 7 to 10 respectively. We used 53 healthy voices and 149 pathological voices in MEEI. In SVD, we used 100 healthy voices and 50 pathological voices. We trained 20 epochs in MEEI and 10 epochs in SVD. When we trained the models in MEEI, every epoch took about 20 minutes. When we trained the models in SVD, it took about 4 hours to use the healthy voices and 2 hours to use the pathological voices every epoch. We can see that the loss of pathological voices is higher than healthy voices in both databases. It is reasonable considering that healthy voices are more regular than pathological voices. Besides, the overall loss of MEEI is higher than SVD, which is mainly because we used a batch size of 4 for SVD due to the GPU memory limit, but a batch size of 16 for MEEI. As we know, with a smaller batch size, the training loss will drop faster, so the loss in SVD will be lower than in MEEI from the first epoch to the last epoch. We found the loss of first iteration of the first epoch in both MEEI and SVD was almost 5, the loss in SVD dropped fast than MEEI.

The Fig. 5 shows the generated voices in experiments 7 to 10 respectively. The blue line is the seed we used, the red line is the true data after the seed, and the green line is the predicted data. We predicted 200 milliseconds of points for both MEEI and SVD. It took about 30 seconds to generate every voice in MEEI, and 4 minutes in SVD. We can find that the healthy and pathological voices in the two databases are both quasi-periodical. Due to the original periodicity in many pathological voices having been broken [28], there will be some high-frequency disturbances in the waveforms which will make it more difficult to predict. This is the reason why the network is more difficult to converge during the training of pathological voices. In Fig. 5 (a), the predicted line and the target line almost overlap, which shows that our predicted waveform is very close to the real data. We can see the data in MEEI can be predicted better than the data in SVD. We believe the reasons are as follows. Firstly, the MEEI database has less noise and fewer samples, so the network is easier to converge. Secondly, the improved WaveNet model used for SVD is more complicated than the model used for the MEEI since the receptive fields are 4093 and 1023 respectively. The second reason is likely to cause the results that the length of the predicted voice will not be long enough. In MEEI, the generated voices tended to amplitude attenuation., so we need to weigh the model complexity and prediction accuracy to prevent the model from being unable to predict the data or failing to converge.

B. Imbalanced Test and Evaluation of the Generated Long Vowel Segments

We have mentioned above that the imbalanced distribution of positive and negative samples can cause meaningless classification results when most samples are classified into the category of majority samples. To explore whether this situation will be improved after balancing different types of samples by adding the generated data, we designed an experiment in both MEEI database and SVD like other papers [7], [24]. We used the trained models to generate extra data and added them to the original



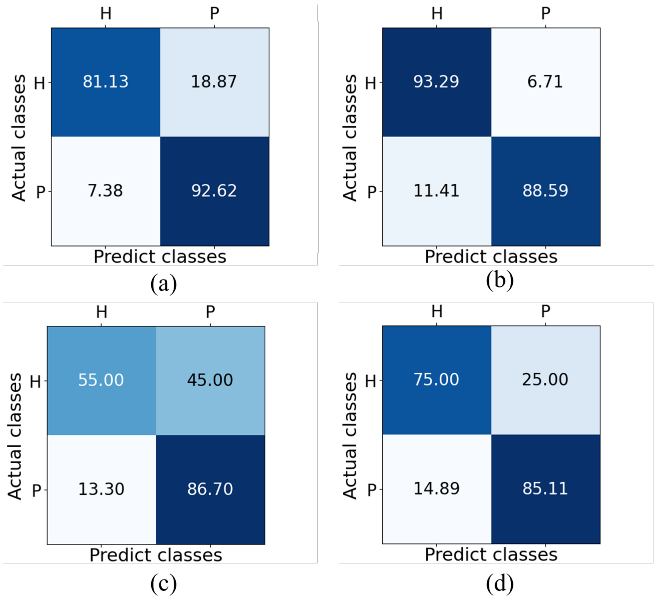
(a)

(b)

(c)

(d)

Fig. 5. The generated waveform with different seeds: (a) healthy voice in MEEI, (b) pathological voice in MEEI, (c) healthy voice in SVD, (d) pathological voice in SVD.



(a)

(b)

(c)

(d)

Fig. 6. The confusion matrixes of the imbalanced tests using long vowels: (a) imbalanced MEEI, (b) balanced MEEI, (c) imbalanced SVD, (d) balanced SVD.

imbalanced databases. As for MEEI database, we used the model with a receptive field of 1023 to generate 96 healthy voices since there are 53 healthy voices and 149 pathological voices in the original MEEI database. As for SVD, we used the model with a receptive field of 4097 to generate 88 healthy voices since we only chose 100 healthy voices to train the network, and there are 188 pathological voices with the types of the vocal nodule, polyp, cyst, and paralysis in SVD. After we got the generated vowels, we extracted MFCC parameters of the true labelled voices and the generated voices to make the imbalanced test. The classifier we used is support vector machine (SVM). 10-fold cross-validation was used in the experiments. We show the confusion matrix in Fig. 6.

In Fig. 6, there are four confusion matrixes which are binary classification results of imbalanced MEEI, balanced MEEI, imbalanced SVD, and balanced SVD respectively, so there is a comparison between sub-figure a and b, and another comparison between c and d. As we can see in the confusion matrixes, the classification results can be improved using the generated data to expand the raw databases. The recall of minority samples in the original MEEI database and SVD is 81.13% and 55% respectively. After we balanced the databases, the recall of minority samples increased to 93.29% and 75% respectively. The recall of minority samples improved by 12.16% and 20.00% respectively.

TABLE V
FOUR DIFFERENT EVALUATION RESULTS USING AND WITHOUT USING OUR IMPROVED WAVENET IN MEEI AND SVD

Database	Parameters	Baseline	Improved WaveNet
MEEI	Accuracy (%)	89.60	90.93
	Recall (%)	81.13	93.29
	Precision (%)	79.63	74.63
	F-1 score (%)	80.37	82.92
SVD	Accuracy (%)	75.69	80.09
	Recall (%)	55.00	75.00
	Precision (%)	68.75	72.82
	F-1 score (%)	61.11	72.91

Apart from the confusion matrixes which can show the accuracy and recall directly, we also presented the four different parameters in Equation 12 to 15. In Table V, the baseline is the result without using our method. For a fair comparison, our proposed method uses the same features and classifier as the baseline except for the data generated by the improved WaveNet model.

For the four objective evaluation metrics in Equation 12 to 15, shown in Table V, almost all of them have been improved by our method except for the precision in MEEI. In the two databases, the overall accuracy was improved by 1.33% and 4.40% respectively. The classification results showed that adding our generated voice data to the original imbalanced database can effectively improve the experimental performance. The four parameters improved by 2.76% and 10.07% on average in our two databases.

We augmented the original database with different proportions of generated data. In Fig. 7, we showed the accuracy and recall of MEEI and SVD by adding different proportions of generated data. The X-axis represents the proportion of the generated data used for all the generated data. We can see that with the increase in the amount of data, the accuracy rate and recall rate have increased significantly. Among them, the accuracy of the MEEI increased by 1.33% with 100% data augmentation, which means we added the same number of generated data as in Table V. The recall of MEEI increased by 12.16% with 100% expended data. In SVD, the biggest improvement in accuracy and recall is also using the full amount of data we generated.

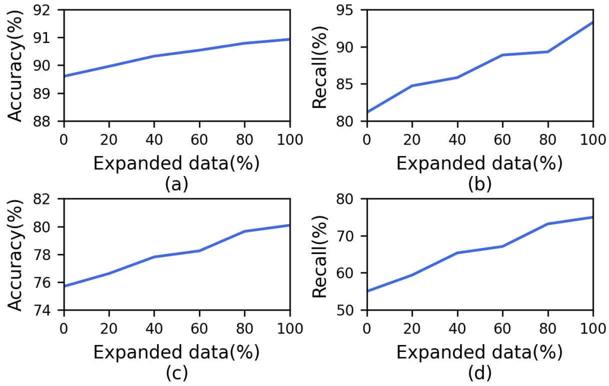


Fig. 7. The classification accuracy and recall of MEEI and SVD using improved WaveNet: (a) accuracy of MEEI, (b) recall of MEEI, (c) accuracy of SVD, (d) recall of SVD.

TABLE VI
ACCURACY-SCORE AND DIVERSITY-SCORE OF THE GENERATED LONG VOWEL SEGMENTS IN MEEI AND SVD

Database	Raw accuracy	Accuracy-score	Diversity-score
MEEI	89.60%	77.22%	85.64%
SVD	75.70%	62.00%	63.33%

Besides, we calculated the accuracy-score and diversity-score of our generated voices. It has been introduced in section II D and it is used to evaluate the accuracy and diversity of our generated data. The features used were MFCC parameters, and the classifier was SVM. The results are shown in Table VI.

In Table VI, the raw data in MEEI were 53 healthy voices and 149 pathological voices, and there were 100 healthy voices and 50 pathological voices in SVD. The raw accuracy represents the classification accuracy of the raw data. The generated voices had the same number as the raw data. We can find the accuracy-score and diversity-score are both higher in MEEI than in SVD. The lower accuracy-score and diversity-score in SVD were partly because the generated data itself was not as good as the generated data in MEEI. Besides, the relatively low classification accuracy of baseline in SVD is also an important factor. The 77.22% accuracy-score and 85.64% diversity-score in MEEI are already at a very high level, indicating that the generated data is very close to the original data in accuracy and diversity.

C. Results of the Generation of Short Vowel Segments

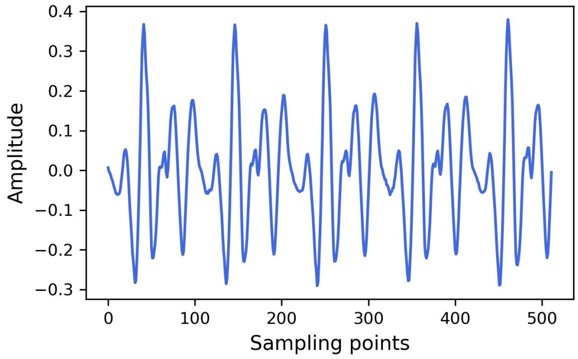


Fig. 8. The generated healthy voice in MEEI.

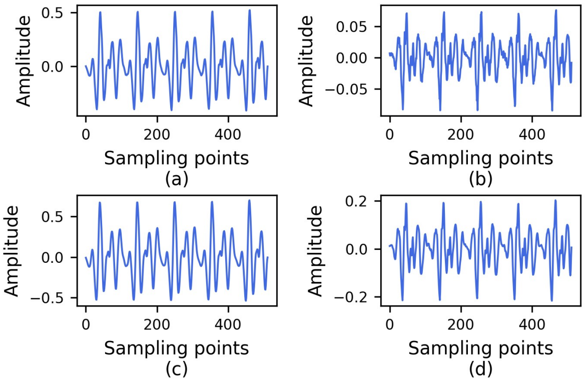


Fig. 9. The SWT coefficients of the data in Fig. 9: (a) first-order approximation SWT coefficient, (b) first-order detail SWT coefficient, (c) second-order approximation SWT coefficient, (d) second-order detail SWT coefficient.

Different from the data generated by the improved WaveNet, the length of the data generated by SWT-WGANGP is much shorter, but it has the advantage of training time and difficulty. The short vowel segments generated by SWT-WGANGP are shown in Fig. 8, Fig. 9 and Fig. 10.

We used SWT to transform raw voice segments and used ISWT to reconstruct new voice segments. From Fig. 8 and Fig. 9, we can directly see the generated voice segment and its SWT coefficients. The generated voice segment in Fig. 8 is very similar to the voice segments in the original database shown in Fig. 2. In Fig. 10 (b), the generated waveform is very smooth. In Fig. 10 (a) and (c), the generated voice segments are mixed with some noise. When the waveforms change greatly such as the parts framed by the boxes, and their periodicity is not perfect as the waveform in Fig. 8 and Fig. 10 (b). This is very similar to the characteristics of true voice segments. The SWT-WGANGP method has the advantage of using less training time.

D. Imbalanced Test and Evaluation of the Generated Short Vowel Segments

To evaluate the performance of SWT-WGANGP, we also took an imbalanced test and evaluation as we did for the generated long vowel segments. The voices with a length of

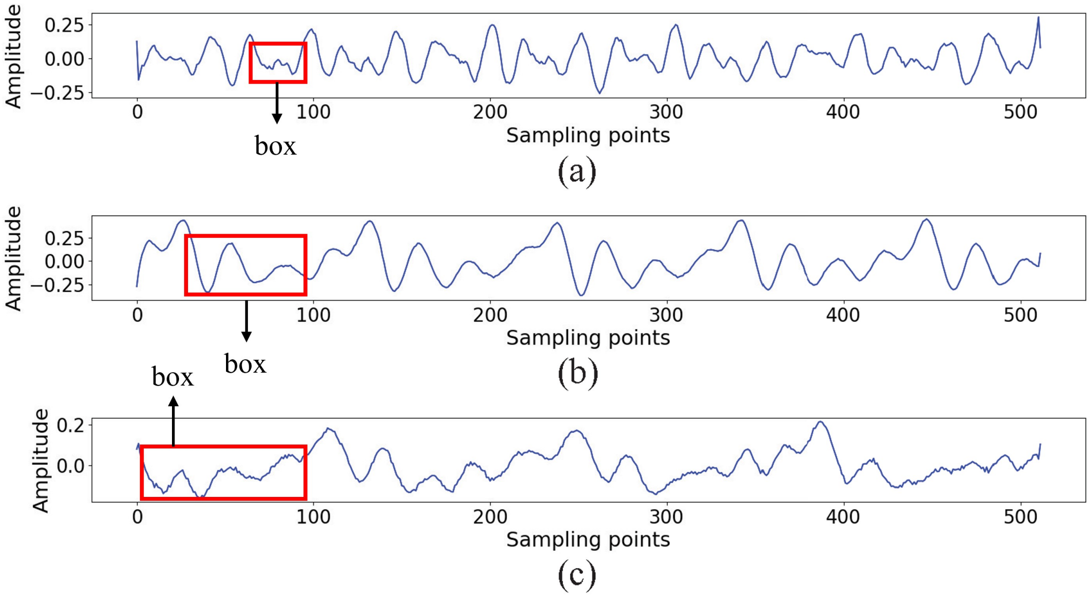


Fig. 10. The voice segments generated by SWT-WGANGP: (a) pathological voice in MEEI, (b) healthy voice in SVD, (c) pathological voice in SVD.

512 points were used as the input of the neural network classifier. The classifier is a fully connected network with one input layer, two hidden layers, and one output layer. The input layer has 512 nodes followed by a Relu activation function. Each hidden layer has 256 nodes and is followed by a Relu activation function. The output layer has 2 nodes to be used to perform binary classification. The learning rate is 0.0001 with an Adam optimizer.

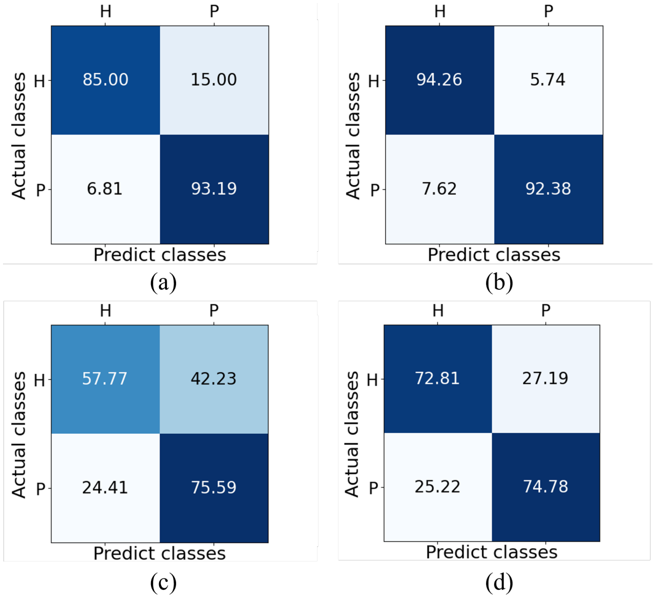


Fig. 11. The confusion matrix of the imbalanced test using short vowel: (a) imbalanced MEEI, (b) balanced MEEI, (c) imbalanced SVD, (d) balanced SVD.

In the imbalanced test, we extracted voice segments at equal intervals in every raw voice. In MEEI, we extracted 20 voice segments of every voice. In SVD, 10 voice segments of every voice were chosen. We used 90% data for training and the remaining 10% data for testing. The confusion matrixes are shown in Fig. 11.

The structure of Fig. 11 is similar to that of Fig. 6. The data used in the original MEEI database and SVD were all the data we have introduced in the above paragraph. The classification accuracy and recall of minority samples of the imbalanced MEEI database are 91.04% and 85% respectively. The two parameters improved by 2.28% and 9.26% respectively after we added the generated data to the original MEEI. In SVD, we added 6030((1290-687)×10) healthy voice segments to the original database. The classification accuracy and recall of minority samples improved by 4.7% and 15.04% respectively after using our generated data.

TABLE VII
FOUR DIFFERENT EVALUATION RESULTS USING AND WITHOUT USING OUR IMPROVED WAVENET IN MEEI AND SVD

Database	Parameters	Baseline	SWT-WGANGP
MEEI	Accuracy (%)	91.04	93.32
	Recall (%)	85.00	94.26
	Precision (%)	81.61	81.48
	F-1 score (%)	83.27	87.41
SVD	Accuracy (%)	69.40	74.10
	Recall (%)	57.77	72.81
	Precision (%)	55.76	58.67
	F-1 score (%)	56.75	64.98

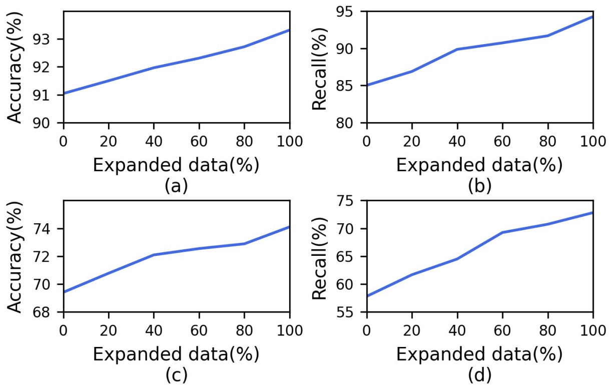


Fig. 12. The classification accuracy and recall of MEEI and SVD using SWT-WGANGP: (a) accuracy of MEEI, (b) recall of MEEI, (c) accuracy of SVD, (d) recall of SVD.

Similar to Table V, we also evaluated the parameters from Equation 12 to 15 for our generated short vowel segments. The results are shown in Table VII. The overall accuracy improved by 2.28% and 4.70% in the two databases respectively. In MEEI, the accuracy of baseline performed better when using short vowels compared with using long vowels. This also explains the necessity of using long and short vowels in our experiments. The four parameters improved by 3.89% and 7.72% on average in our two databases.

The Fig. 12 is similar to Fig. 7. When more generated data is used, the improvement is more obvious. 0% expanded data represents the baseline in Table VII, and 100% expanded data means the SWT-WGANGP in Table VII.

TABLE VIII
ACCURACY-SCORE AND DIVERSITY-SCORE OF THE GENERATED SHORT VOWEL SEGMENTS IN MEEI AND SVD

Database	Raw accuracy (%)	Accuracy-score (%)	Diversity-score (%)
MEEI	91.04	93.17	83.76
SVD	69.40	69.27	52.90

Except for the imbalanced test, we also evaluated the accuracy-score and diversity-score of our generated data. In MEEI, we had 1060(53×20) true healthy voice segments, 2980(149×20) true pathological voice segments, 1060 generated healthy voice segments, and 2980 generated pathological voice segments. In SVD, we had 6870(687×10) true healthy voice segments, 12900(1290×10) true pathological voice segments, 6870 generated healthy voice segments, and 12900 generated pathological voice segments. The accuracy-score and diversity-score in different databases are shown in Table VIII.

As we can see in Table VIII, the accuracy-score in both MEEI and SVD is close to the raw accuracy, which means the generated data can be classified correctly using the classifier trained by true labelled data. The accuracy-score is higher than diversity-score in both MEEI and SVD. This shows that the diversity of the generated data is lower than its accuracy, and it's a common drawback of GAN. The scores in SVD are quite low compared with MEEI because the scores are determined both by the quality of the generated data and the raw accuracy of the imbalanced databases, and the raw accuracy in MEEI is much higher than SVD. This result is similar to the analysis in Table VI.

E. Comparison with traditional methods

Traditional methods include noise addition and spectral augmentation. In the noise addition, white noise will be added to the vowels to generate more vowels. As for spectral augmentation, the extracted spectral features will be masked to generate more robust features. Since our improved WaveNet can generate signals long enough to perform both noise addition and spectral augmentation, we compared our improved WaveNet with these two methods. Since the segmental speech generated by the second method is insufficient to extract sufficiently long spectral features, we compared the short speech generated in our framework with the noise addition method only. The experimental setup is consistent with the above experiments. In the noise addition method, we randomly added white noise to the original vowels [42], and the energy of white noise is about 10% to 50% of the energy of the original vowels. In the spectral augmentation method, we randomly masked 10% to 20% of the extracted spectral features. The results are in Table IX.

In the MEEI, we used short vowels for comparison because our experiments above show that short vowels perform better with our approach. Similarly, in the SVD, we used long vowels for our experiments. The results for the long vowels in SVD achieved the highest average score compared with the other methods, and the accuracy, recall, and f1-score are all the highest in our experiments. As for the short vowels in MEEI, compared with the noise addition method, our proposed framework has the highest three out of four scores and the average score is 3.30% higher.

To verify the effectiveness of our method for improving the performance of AVPD, we not only compared our method with other traditional methods but also verified our method on the state-of-the-art (SOTA) method. In Table X, we used our generated long vowel signals to balance the raw databases and used the Gammatone spectral latitude features (GTSL) [27] with the SVM classifier. The GTSL has been proven to be very useful on AVPD.

As we can find in Table X, the classification results of using GTSL features improved significantly after data augmentation using our method. Specifically, in the MEEI, the accuracy increased by 1.59%, and in the SVD, the accuracy increased by 1.65%. In addition to accuracy, several other indicators have also improved. In MEEI, the average improvement of the four indicators is 2.52%. In SVD, the average boost is 2.09%. The above experiments prove that even on some SOTA methods, the proposed data augmentation method is still suggestive of the overall classification performance.

V. DISCUSSION

Our method can directly generate voices to help solve the problem of imbalanced conditions in pathological voice databases and expand the raw databases. The generated voices with different durations can have access to a wide range of applications according to actual conditions. For example, the generated data can be added into the raw imbalanced and insufficient databases to make the classification results have a higher recall of minority samples and robustness. The improved WaveNet model in our framework can generate voice signals with hundreds of milliseconds. Considering the true voice signals in databases are only about 1 to 3 seconds, we believe the generated data can meet the needs of use. The proposed SWT-WGANGP model in our framework can generate relatively short voice segments in a one-shot manner. No similar works have been taken to expand the raw MEEI database and SVD before.

Fig. 5, 8, and 10 directly show the waveforms generated by our framework and can give us subjective evaluations. From the time domain waveform, our method can preserve the health or pathological information of the original voice signal very well. Table VI and VIII are the objective evaluations of our generated waveforms. Due to the innovation of our work, there is no good method for evaluating generated vowels, so we first work on the above-mentioned subjective and objective evaluation methods to ensure the quality of the generated voices. Our evaluation methods are the basis of our work and can provide a strong guide for future work since they are the steps required for all such work on vowel generation. The Fig. 6 and Table V are

TABLE IX
COMPARISON WITH TRADITIONAL METHODS

	Methods	Accuracy (%)	Recall (%)	Precision (%)	F1-score(%)	Average(%)
Long vowels	Noise addition	78.13	66.00	69.47	67.69	70.32
	spectral augmentation	79.17	65.00	72.22	68.42	71.20
	Proposed framework	80.09	75.00	72.82	72.91	75.21
Short vowels	Noise addition	89.60	77.36	82.00	79.61	82.14
	spectral augmentation	\	\	\	\	\
	Proposed framework	90.93	93.29	74.63	82.92	85.44

TABLE X
EXPERIMENTS BASED ON THE SOTA METHODS

Databases	Methods	Accuracy (%)	Recall (%)	Precision (%)	F1-score(%)	Average(%)
MEEI	GTSL [27]	93.56	90.57	85,71	88.07	89.48
	Proposed framework	95.05	94.33	87.72	90.91	92.00
SVD	GTSL [27]	79.29	77.29	68.43	72.59	74.40
	Proposed framework	80.94	79.77	70.44	74.81	76.49

TABLE Ⅺ
COMPARISON WITH OTHER WORKS

Methods	Models	Expand of raw data	Type of generated data	Viability in pathological voice database
Santos et al. [43]	Physical model	No	Voice	Yes
Fan et al. [7]	FC-SMOTE	Yes	Features of voice	Yes
Chui et al. [9]	GAN	Yes	Features of voice	Yes
Hwang et al. [44]	WaveNet	Yes	Voice	No
Our methods	Generative models	Yes	Voice	Yes

the results of the improved WaveNet in our framework, and the Fig. 11 and Table VII are the results of the SWT-WGANGP in our framework. The best results in our framework achieve average relative improvements of 4.55% and 16.85% in four metrics in MEEI and SVD respectively. Such results show that our method is promising because our data amplification method can be used as a pre-processing method for other researchers. Based on these best results, we have done comparative experiments in Table IX to verify the superiority of our method. In the generation of long vowels, our method is higher than noise addition and spectral augmentation by an average of 4.89% and 4.01% respectively. This is mainly due to the usage of dilated convolutions in our improved WaveNet, which can better capture the information of voice signals. In the generation of short vowels, our method is higher than the noise addition method by an average of 3.30%.

We show the novelty of our idea in Table Ⅺ. We compare our method with other works in three main aspects. The method [43] used a physical model of vocal polyp to generate extra vowel signals with or without vocal polyp. However, this work only explored the model of vocal polyp, and there are many other diseases such as vocal nodules, and vocal paralysis. Some other works [7], [9] explored the expansion of the features extracted from voices rather than the voice itself. This leads to a limited application scenario for them, as they can only generate specified features instead of generic raw voices. The raw WaveNet model can be trained directly in the other field [44]. However, due to the complexity of the model and the number of raw voice recordings, it is difficult to train the raw WaveNet in our databases. Our proposed method performs well in the areas analyzed above.

Whilst these results are encouraging and indicate that the neural generative model can directly generate voices to overcome the drawbacks in the commonly used databases, there are further challenges when generating longer voices with high quality in a faster way. We believe further research can help us to understand whether more efficient deep learning models can be used in the task of voice generation using pathological voice databases and discuss these aspects in more detail below:

- We have improved the raw WaveNet model for our specific task, but the core idea of the improved model is dilated causal convolution as same as the raw model. The attention mechanism [45] has been considered as convolution operation in a broad sense, and it can also have a large receptive field. So we should investigate the probability and effects of using the attention mechanism in our models.
- A compromise method has been adopted when we designed the SWT-WGANGP model due to the limitation of GAN. We have tried the GAN model [46] which can generate long voice segments, but we find it can't be trained using our databases. Future work can investigate the more effective voice generation GAN model.
- The data used in our improved WaveNet was raw voice data, and the data used in SWT-WGANGP was the SWT coefficients. Except for the raw voice data and SWT coefficients, there exist many other types of input data for voice generation, such as the STFT sequence and Mel spectrogram used in other works [15], [47]. The effect of these types of input data should also be investigated.

VI. CONCLUSION

This paper investigates the use of deep generative models to synthesize vowel signals in low-resource pathological voice databases for the better detection of pathological voice. A framework which can generate both long and short vowel segments is presented. We improve the original WaveNet for the generation of vowel signals with 200 milliseconds in our framework with less computational complexity. A

SWT-WGANGP model is proposed to generate vowel segments with 512 sample points in our framework. The SWT-WGANGP model can generate vowel segments with a short length in a one-shot manner. Our imbalanced test and evaluation results show that the generated data resembles the real data to a large extent. The generated voice segments can be used to expand the raw database to improve classification accuracy, recall of minority samples, and F1-score.

Our main contribution is to show that we can expand the original imbalanced and insufficient pathological voice databases by directly generating new voices. To the best of our knowledge, no analytical solutions exist, which we believe is a positive step for better application of pathological voice detection. We can achieve a relative average boost of up to 4.55% and 16.85% on the two databases respectively. The advantage of our method is that it can be used in any other relevant research work, as it can be used as a pre-processing to expand the quantity of original data. We also discuss the challenges for the improvement of our methods and future research directions of our work.

REFERENCES

[1] Al-Nasheri, Ahmed, et al. "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions." *Ieee Access* 6 (2017): 6961-6974.

[2] Hegde, Sarika, et al. "A survey on machine learning approaches for automatic detection of voice disorders." *Journal of Voice* 33.6 (2019): 947-e11.

[3] Janak Kapoor, Ajita Pathak, Manish Rai, and G.R Mishra, "Speech Quality Enhancement through Noise Cancellation using an Adaptive Algorithm," *IAENG International Journal of Computer Science*, vol. 49, no.3, pp653-665, 2022

[4] Dankovičová, Zuzana, et al. "Machine learning approach to dysphonia detection." *Applied Sciences* 8.10 (2018): 1927.

[5] Orozco-Arroyave, J.R.; Vásquez-Correa, J.C.; Vargas-Bonilla, J.F. et al., "NeuroSpeech: An open-source software for Parkinson's speech analysis." *Digital Signal Processing* 2018, 77, 207–221.

[6] Kim, H.; Jeon, J.; Han, Y.J. et al.; "Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy." *Journal of Clinical Medicine* 2020, 9, 3415.

[7] Fan, Z.; Wu, Y.; Zhou, C.; Zhang, X.; Tao, Z. "Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method." *Applied Sciences* 2021, 11, 3450.

[8] Soltanzadeh, P.; Hashemzadeh, M. "RCSMOTE: range-controlled synthetic minority over-sampling technique for handling the class imbalance problem." *Information Sciences* 2021, 542, 92–111.

[9] Chui, K.T.; Lytras, M.D.; Vasant, P. "Combined generative adversarial network and fuzzy C-means clustering for multi-class voice disorder detection with an imbalanced dataset." *Applied Sciences* 2020, 10, 4571.

[10] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu et al.; "Generative adversarial nets." *Advances in Neural Information Processing Systems* 2014, 27.

[11] Aicha, A.B. "Contribution of data augmentation for the prenventive detection of vocal fold precancerous lesions." *Procedia Computer Science* 2019, 159, 212–220.

[12] Childers, D.; Yegnanarayana, B.; Wu, K. "Voice conversion: Factors responsible for quality. *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*." *IEEE*, 1985, Vol. 10, pp. 748–751.

[13] Kawahara, H.; Masuda-Katsuse, I.; De Cheveigne, 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 Communication* 1999, 27, 187–207.

[14] Morise, M.; Yokomori, F.; Ozawa, K. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications." *IEICE Transactions on Information and Systems* 2016, 99, 1877–1884.

[15] Wang, X.; Takaki, S.; Yamagishi, J. "Neural source-filter waveform models for statistical parametric speech synthesis." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2019, 28, 402–415.

[16] Ai, Y.; Ling, Z.H. "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2020, 28, 839–851.

[17] Amara, A.B.; Jebara, S.B. "Contribution on Gaussian Mixture Model Order Determination for Voice Conversion." *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC). IEEE*, 2018, pp. 87–92.

[18] Zhang, T.; Shao, Y.; Wu, Y.; Pang, Z.; Liu, G. "Multiple vowels repair based on pitch extraction and line spectrum pair feature for voice disorder." *IEEE Journal of Biomedical and Health Informatics* 2020, 24, 1940–1951.

[19] Jiao, Y.; Tu, M.; Berisha, V.; Liss, J. "Simulating dysarthric speech for training data augmentation in clinical speech applications." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018, pp. 6009–6013.

[20] Sisman, B.; Zhang, M.; Li, H. "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2019, 27, 1085–1097.

[21] Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O. et al.; "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 2016.

[22] Angrick, M.; Herff, C.; Mugler, E. et al.; "Speech synthesis from ECoG using densely connected 3D convolutional neural networks." *Journal of Neural Engineering* 2019, 16, 036019.

[23] Cataldo, E.; Soize, C. "A stochastic model of voice generation and the corresponding solution for the inverse problem using Artificial Neural Network for case with pathology in the vocal folds." *Biomedical Signal Processing and Control* 2021, 68, 102623.

[24] Wu, Y.; Zhou, C.; Fan, Z. et al.; "Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection." *IEEE Access* 2020, 9, 30–44.

[25] Bachu, R.; Kopparthi, S.; Adapa, B.; Barkana, B. "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal." *American Society for Engineering Education (ASEE) zone conference proceedings. American Society for Engineering Education*, 2008, pp. 1–7.

[26] Mohammed, A.; Mansour, A.; Ghulam, M. et al.; "Automatic speech recognition of pathological voice." *Indian Journal of Science and Technology* 2015, 8, 1–6.

[27] Zhou, C.; Wu, Y.; Fan, Z. et al.; "Gammatone spectral latitude features extraction for pathological voice detection and classification." *Applied Acoustics* 2022, 185, 108417.

[28] Wulan, N.; Wang, W.; Sun, P. et al.; "Generating electrocardiogram signals by deep learning." *Neurocomputing* 2020, 404, 122–136.

[29] Ramírez, M.M.; Benetos, E.; Reiss, J.D. "Deep learning for black-box modeling of audio effects." *Applied Sciences* 2020, 10, 638.

[30] Moseley, B.; Nissen-Meyer, T.; Markham, A. "Deep learning for fast simulation of seismic waves in complex media." *Solid Earth* 2020, 11, 1527–1549.

[31] Radford, A.; Metz, L.; Chintala, S. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 2015.

[32] Elgammal, A.; Liu, B.; Elhoseiny, M.; Mazzone, M. "Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms." arXiv preprint arXiv:1706.07068 2017.

[33] Jia, N.; Zheng, C.; Sun, W. "A Model of Emotional Speech Generation Based on Conditional Generative Adversarial Networks." *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). IEEE*, 2019, Vol. 1, pp. 106–109.

[34] Oord, A.; Li, Y.; Babuschkin, I. et al.; "Parallel wavenet: Fast high-fidelity speech synthesis." *International Conference on Machine Learning. PMLR*, 2018, pp. 3918–3926.

[35] Arjovsky, M.; Chintala, S.; Bottou, L. "Wasserstein generative adversarial networks." *International Conference on Machine Learning. PMLR*, 2017, pp. 214–223.

[36] Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. "Improved training of wasserstein gans." arXiv preprint arXiv: 1704.00028 2017.

[37] Griffin, D.; Lim, J. "Signal estimation from modified short-time Fourier transform." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 1984, 32, 236–243.

[38] Nason, G.P.; Silverman, B.W. "The stationary wavelet transform and some statistical applications." *Wavelets and Statistics; Springer*, 1995; pp. 281–299.

[39] Barratt, S.; Sharma, R. “A note on the inception score.” arXiv preprint arXiv:1801.01973 2018.

[40] Che T, Li Y, Jacob A P, et al. “Mode regularized generative adversarial networks.” arXiv preprint arXiv:1612.02136, 2016.

[41] Paszke, A.; Gross, S.; Massa, F. et al.; “Pytorch: An imperative style, high-performance deep learning library.” *Advances in Neural Information Processing Systems* 2019, 32, 8026–8037.

[42] Ruxue Guo, Tao Jiang, Qingyun Wang, Ruiyu Liang, and Cairong Zou, "An Improved Low-Complexity Echo Suppression Algorithm Based on the Acoustic Coloration Effect," *IAENG International Journal of Computer Science*, vol. 49, no.3, pp637-643, 2022.

[43] Santos, J.; Montalvao, J.; Santos, I. “Improved Model for Vocal Folds with a Polyp with Potential Application.” *INTERSPEECH*, 2020, pp. 1386–1390.

[44] Hwang, M.J.; Yamamoto, R.; Song, E.; Kim, J.M. “TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis.” *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2021, pp. 6598–6602.

[45] Vaswani, A.; Shazeer, N.; Parmar, N. et al.; “Attention is all you need.” *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[46] Engel, J.; Agrawal, K.K.; Chen, S. et al.; “Gansynth: Adversarial neural audio synthesis.” arXiv preprint arXiv:1902.08710 2019.

[47] Kumar, K.; Kumar, R.; de Boissiere, T. et al; “Melgan: Generative adversarial networks for conditional waveform synthesis.” arXiv preprint arXiv:1910.06711 2019.