IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Deep Fake Voice Synthesizer

D. Arvind Sai, D. Shiva Sai, C. Sathya Narayana, Mr. Christal Anand
Department of Computer Science and Engineering
Prathyusha Engineering College, Poonaamallee, Thiruvallur, Chennai, India

Abstract: *With the rapid-fire development of computer technology, voice technology has become an exploration hotspot in the field of deep literacy. Allowing computers to hear, see, speak and feel is the unborn development direction of mortal- computer commerce. Among them, voice will become the most promising mortal- computer commerce system in the future which has further advantages than other commerce styles. Voice cloning is one of the branches of voice technology, introducing a language modelling approach for textbook to speech conflation(TTS). Specifically, we train a neural codec language model using separate canons deduced from an out- the- shelf neural audio codec model, and regard TTS(Text to Speech) as a tentative language modelling task rather than nonstop signal retrogression as in former work. During the pre- training stage, we gauge the TTS training data to 1k hours of English speech which is hundreds of times larger than systems. Software emerges in- environment literacy capabilities and can be used to synthesise high- quality substantiated speech with only a 10- second enrolled recording or written textbook of an unseen speaker as an aural advice. trial results show that our software significantly outperforms the state- of- the- art zero- shot TTS system in terms of speech light heartedness and speaker similarity. In addition, we find our software could save the speaker's emotion and the aural terrain of the aural advisement in conflation and there are diligence similar as audio books, filmography(dubbing) and people with disabilities facing problems through audio. To break the problem of furnishing a large number of sample lines to reduplicate a voice and the long waiting time for voice cloning.*

Keywords: Voice Clone; A Few Samples; Real Time

I. INTRODUCTION

Voice cloning is to keep the current semantic information, and only change the speaker's voice personality characteristics to make it sound like another speaker.

Through the study of voice cloning, we can further strengthen the study of speech- related parameters, explore the pronunciation medium of humans, control the personality characteristic parameters of speech signals. advantage is that the network and data set used in each model can be flexibly selected and changed, which is convenient for later optimization.

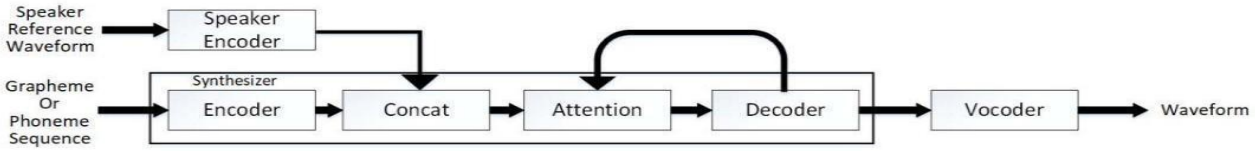
1.1 Research Background and Significance

Now generally use the following two methods to achieve voice cloning. The first one is speech synthesis which turns text to speech (TTS). The second is voice conversion which converts voice into voice with target features (VC). In this study, I use TTS technology, so I will not talk too much about VC. Early TTS was based on the Hidden Markov Model (HMM) statistical parameter speech synthesis method that can generate continuous, smooth, and intelligible speech, but the generated speech spectrum is too smooth, the sound quality is relatively low, very different from natural speech. In recent years, deep learning has been used to clone voices. The speech synthesis method based on deep learning is WaveNet [1], DeepVoice [2], DeepVoice2 [3], DeepVoice3 [4], Tacotron [5] and so on. Voice cloning technology has been applied in various fields such as personalised voice interface, advertisement, robot, game, broadcasting and so on.

1.2 System Overview

The system in this study is divided into three modules: encoder, synthesiser and vocoder, as shown in Figure 1.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023



1.3 Network structure

Speaker encoder

In this study, high quality input audio is not necessary. So a large corpus composed of many different speakers is used to train the encoder. This choice has the advantages of strong anti- noise capability and convenience in capturing humans. Among them, the encoder module converts the speaker's voice into speaker embedding. The vocoder module converts mel-spectrogram to waveform verification tasks. The task specifies the way of output embedding.

Consider a voice data set grouped by speakers. Represent the j^{th} speech of the i^{th} speaker as, and use to represent the logarithmic Mel spectrum of speech. The log- mel spectrum is a deterministic and irreversible function that can extract speech features from the waveform.

The encoder ε calculates the embedding corresponding to the speech , where is the parameter of the encoder. In addition, the speaker embedding is defined as the centroid of the speaker's speech embedding,

The model is a 3-layer LSTM with 256 hidden nodes, followed by a projection layer of 128 units. The input of this model is 40-dimensional MFCC, with a window width of 25 ms and a step size of 10 ms. Then there is the ReLU layer, whose purpose is to make the embedding sparse for easy understanding. The output is generated by L2 norm regularisation of the last layer, which is a vector of 256 elements.

Speaker verification is to verify a person's identity through voice and create a template for each person. This process is called registration. The embedding of different voices spoken by the same speaker is highly correlated, while the embedding of different speakers will be divided into different spaces.GE2E loss simulated this process to optimise the model. During training, the model calculates the embedded of M fixed-duration speech from N speakers (1 i N, 1 j M). The speaker embedding is derived for each speaker. The similarity matrix is the result of a two-by-two comparison between all embedded and each speaker embedded (1 k N) in the batch. This metric is the scaled cosine similarity, see formula 1:

$$s_{,} = \omega \cdot \cos e_{,} c + b = \omega \cdot e \cdot |c| + b \quad (2)$$

Where ω and b are learnable parameters. The whole process is shown in Figure 2. From a computational point of view, the cosine similarity of two L2 normed vectors is just their dot product. The loss is the sum of the loss of the normalised exponential function line by line.

When calculating the loss, each speech is contained in the centroid of the same speaker. This will prejudice the right speaker. In order to avoid this situation, the speech of the right speaker for comparison will be deleted from the speaker embedding data. The definition of the similarity matrix is as formula 3:

$$s_{,} = \omega \cdot \cos e_{,} c + b \quad \text{if } i = k \quad (3)$$

The exclusive centroid is defined as the following formula 3:

$$c = \sum e \quad (4)$$

The data set used by this module is ST-CMDS. The data set was opened by Beijing surf Technology Company with a total of 855 people, each speaking 120 sentences, a total of 10260 sentences of speech.

Synthesiser

The synthesiser is based on optimised Tacotron 2. Replace Wavenet in Tacotron 2 with a modified network. Each character in the text sequence is first embedded as a vector and then convolved to increase the span of a single encoder frame. Meanwhile, input the corresponding phoneme sequence, which can quickly converge and improve pronunciation. These encoder frames are transmitted via bidirectional LSTM to produce encoder output frames. The attention mechanism focuses on the encoder output frame to generate the decoder input frame. Each decoder input frame is connected to the previous decoder frame output, thus making the model autoregressive. This cascading vector passes through two one-way LSTM layers before being projected onto a single MEL spectrum frame. Another projection

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

prediction network of the same vector into a scalar emits a value above a certain threshold to stop frame generation. The entire sequence of frames passes through a residual network before becoming a MEL spectrum. This architecture is shown in Figure 2.

The target MEL spectrogram of the synthesiser has more acoustic characteristics than the speaker encoder. They are calculated in 12.5ms steps from a 50ms window and fed into 80-dimensional MFCC.

The data set used by this module is thchs30. Thchs30 is a 30-hour data set of Tsinghua university. The data set is chosen because it has pinyin, which is convenient for processing. To process it into lib speech format, one step needs to process each transcript to get alignment, which takes a lot of time.

Vocoder

Speech synthesis usually uses WaveNet, and the speech it generates has a high naturalness. However, it takes too long to train and use. The vocoder in this paper adopts the LPCNET model improved by Wavernn. Figure 4 shows General structure of it. On its left is a frame rate network and on its right is a sampling rate network. The synthesised input is limited to the features of 18 Bark scale cepstrum coefficients and 2 pitch parameters. For low bit rate coding applications, the above features need to be quantized. For text to speech, the method of J. Shen et al. [7] will be used.

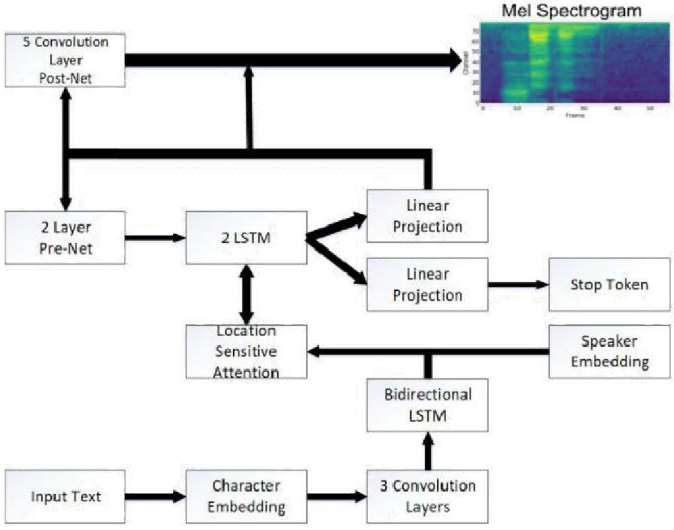


Figure 2. The structure of the Synthesiser

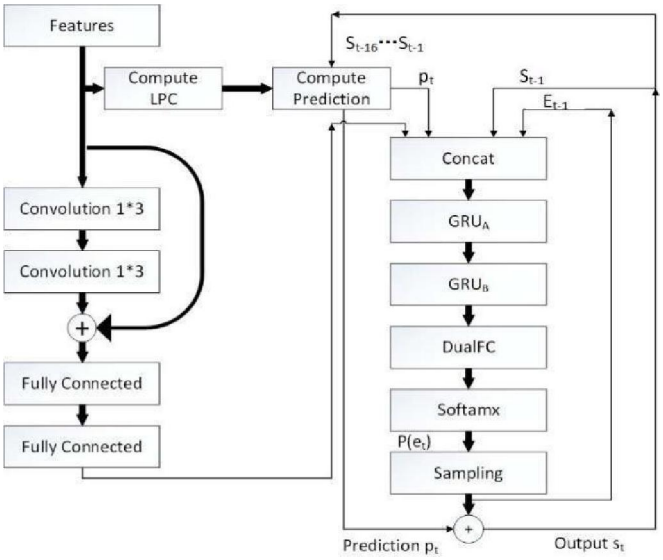


Figure 3. The structure of the vocoder

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Conditional parameters

Two convolution layers with filter size of 3 convert the 20- dimensional features in the frame rate network into receptive fields of 5 frames. The receptive field is added to the residual connection followed by the two full connection layers. The frame rate network outputs a 128-dimensional condition vector F to the sampling rate network

b) Pre-weighting and quantification

WaveNet and other composite models use the 8-bit μ-law quantization [8] to reduce the possible output sample value to 256. Speech signals tend to be concentrated in low frequencies, while μ-law white quantization noise is concentrated in high frequencies, which will reduce speech quality. Therefore, the output is usually extended to 16 bits [9] to solve this problem. The first-order pre emphasis filter ( ) = 1 − is applied to the training data. The synthesised output is filtered by an inverse filter D(z) to reduce its power by 16db at Nyquist rate. This significantly reduces noise and improves the quality of synthesised speech. D(z) is defined in formula 5 below:

c) Linear prediction

The neural networks in many speech synthesis methods [10-13] must simulate the entire speech generation process, including glottic impulses, noise excitation, and channel response. A simple all-pole linear filter well represents the channel response. When is set as the signal at time t, its linear prediction based on previous samples is shown in formula 6 below:

$$p = \Sigma\ a\ s \qquad (6)$$

Where is the linear prediction coefficient (LPC) of ŭ order of the current frame. For vocoders, the input is the feature of one frame, so the autocorrelation function needs to be calculated from the feature. Starting from autocorrelation, Levinson-Durbin algorithm is used to calculate the prediction factor. Cepstrum calculation of prediction factors ensures that no other information will be transmitted or synthesised.

The output layer

In order to calculate the output probability more easily without increasing the size of the previous layer, two fully connected layers are combined in an element-by-element weighted sum manner. A layer that can be called a dual full connection (or DualFC) is defined as formula 7 below:

$$\text{dual } c(x) = a * \tanh(W\ x) + a\ * \tanh(W\ x) \quad (7)$$

Where and are the weight matrix, and are the weight vectors. The output of DualFC generates a probability distribution through the softmax activation function, and then the excitation signal (or linear prediction residual) is sampled from this distribution, and the excitation signal plus the linear prediction signal obtain the final output signal. The DualFC layer can judge whether a value is in the μ-law quantization interval.

Sparse matrix

The first largest GRU (in figure 3) is represented by a sparse matrix to aid vectorization. After a certain iter is normally trained in the initial stage, weight is multiplied by a binary mask at regular intervals to force thinning, and the number of non-zero weights decreases as the training progresses. At the same time, the weight of the diagonal position is always kept. The main purpose of sparse is to reduce the computational complexity during testing.

Embedding and algebraic simplification

First, the non-cyclic part is pre-calculated. Obviously, there are only 256 possibilities for μ-law input, so μ-law Embedding is first converted into 128 dimensions, and then the multiplication results of 256 possible embedding and GRU related acyclic matrices are stored, so that this part of calculation can be completed through lookup table during synthesis. At the same time, the condition vector f is only calculated once per frame, so the multiplication results of the corresponding matrices in f and GRU can also be stored and multiplexed in the frame. Through these steps, synthesised speech can be accelerated.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

a) Inject noise

When synthesising speech, the quality of speech is reduced because the resulting samples are different from those used in training. After synthesis, the result will be even worse. To make the network more robust, you can add noise to the input during training.

The noise has the same shape as the synthetic filter ( ). By adding the noise shown in figure 4, Linear prediction and excitation are both based on noisy speech recalculation, which makes it better for network input and reference excitation to introduce noise synchronously. This can improve the quality of synthesised speech.

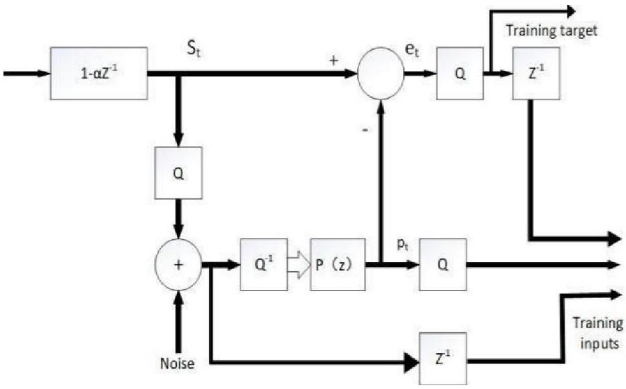


Figure 4. Noise injection during training

Q is the quantization of -law, and is the transformation from -law to linear. The predictive filter is defined in formula 8 below:

$$p(z) = \sum a\, z \qquad (8)$$

Applied to the quantized input of noise.

The data set used by this module is AISHELL. AISHELL is a Chinese speech data set published by Beijing hill company, which contains 178 hours of open source data. The dataset contains the voices of 400 people from different parts of China with different accents.

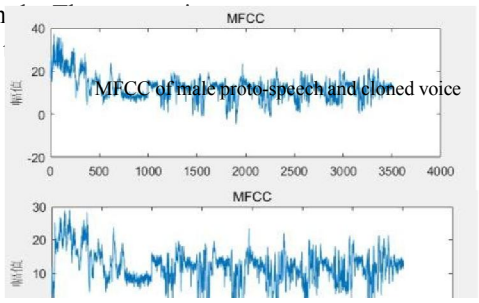
II. EXPERIMENTAL RESULTS AND ANALYSIS

The test and evaluation of the performance of the voice clone method is one of the important parts of the voice clone research. It is of great significance to design a credible and efficient evaluation scheme to improve the performance of the voice clone. Nowadays, the performance of the voice clone method is tested and evaluated mainly by objective and subjective means.

2.1 Objective evaluation and analysis

The test generated cloned speech was compared with the original speech in terms of MFCC and spectrum:

Take STCMD 00044A as an exam
"the man asked me if I would like



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

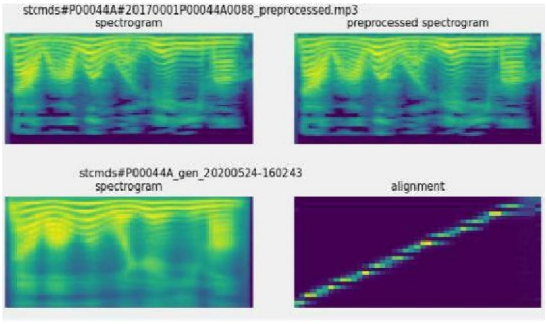


Figure 6. Comparison and alignment of male speech spectra

Take STCMD00052I as an example, the content is: "prepare the stock gap in advance" for women.

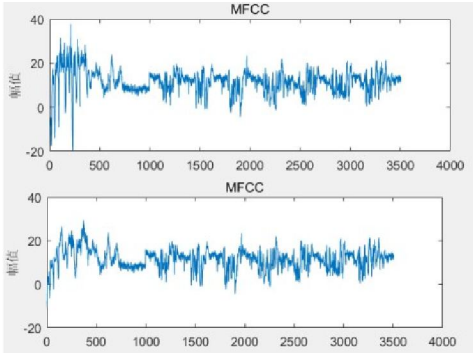


Figure 7. MFCC of Female proto-speech and cloned voice

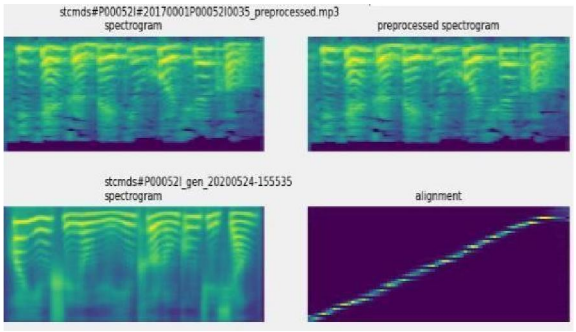


Figure 8. Comparison and alignment of Female speech spectra

It can be seen from the figure above that the original speech and the cloned speech approximate each other in the middle and back parts, but the beginning part is not true. This aspect can be further optimised in the future improvement. Female voice clones performed better than male ones because they were trained with less male voice data. High spectral similarity, high alignment between speech and text

2.2 Subjective evaluation and analysis

The subjective evaluation is testing the pronunciation through people's subjective feelings. The subjective method generally evaluates the clone effect from the two perspectives of speech quality and the similarity of speaker characteristics, and the method adopted is mainly Mean opinion score (MOS) [14].

MOS test: the main principle of MOS test is to ask the reviewer to score the subjective feelings of the test speech according to five grades, which can be used for the subjective evaluation of the speech quality and the similarity of the speaker's characteristics. The MOS score is a composite average of all test statements and all reviewers.

Generally, it can be divided into 5 levels, 1 point corresponds to the worst unintelligible, 5 points correspond to the best

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

close to nature.

We got 10 people on the Internet to make subjective evaluations

TABLE I. FEMALE VOICE MOS TEST SCORES

No.	1	2	3	4	5	6	7	8	9	10
Score	3	3	5	4	3	5	4	4	5	3

TABLE II.MALE VOICE MOS TEST SCORES

NO.	1	2	3	4	5	6	7	8	9	10
Score	4	3	3	3	4	3	4	5	4	3

It can be seen from the above data that there is still a gap between the effect of male voice cloning and female voice cloning. There are several reasons for this: 1. the female voice is generally more sharp and has a stronger penetration which is easier to extract data by computer. 2. In this study, the training database of male voice data is insufficient, so the effect of male voice cloning is not as good as that of female voice.

III. CONCLUSION

In view of the fact that speech cloning used to require a large amount of user data, this paper presents a speech cloning based on a few samples, which is not only faster than the traditional method, but also helpful for the future deployment to low- performance devices. In this study, a new structure is adopted to train the modelling on three modules of encoder, synthesiser and vocoder. In this method, it can be optimised and improved more quickly in the future. At present, it is difficult for the effect of Chinese speech cloning to catch up with English speech cloning, mainly because the Chinese data is extremely scarce compared with the English data. On the other hand, Chinese prosody is much more complicated than English, so there is still a lot to be improved in order to achieve perfect speech cloning. In the future work will continue to improve the structure of the network, improve the effect of cloning

REFERENCES

[1]. Aaron van den Oord, Sander Dieleman, Heiga Zen, et al. Wavenet: A Generative Model for Raw Audio[J/OL]. arXiv Preprint arXiv:1609.03499, 2011,
[2]. Sercan O. Arik, Mike Chrzanowski, Adam Coates, et al. Deep Voice:real-time neural text-to-speech[J/OL]. arXiv preprint arXiv:1702.07825, 2017.
[3]. Sercan Arik, Gregory Diamos, Andrew Gibiansky, et al. Deep Voice Multi-Speaker Neural Text-to-Speech. arXiv Preprint arXiv: 1705.08947, 2017.
[4]. Wei Ping, Kainan Peng, Andrew Gibiansky, et al. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. arXiv Preprint arXiv: 1710.07654.
[5]. Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, et al. Tacotron: Towards End-to-End Speech Synthesis. arXiv Preprint arXiv: 1703.10135, 2017.
[6]. Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalised end-to-end loss for speaker verification. arXiv Preprint arXiv:1710.10467, 2017.
[7]. J. Shen, R. Pang, R. J. Weiss, et al. Natu- ral tts synthesis by conditioning wavenet on mel spectrogram predictions, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp. 4779– 4783.
[8]. N. Kalchbrenner, E. Elsen, K. Simonyan, et al. Efficient neural audio syn- thesis. arXiv Preprint arXiv:1802.08435, 2018.
[9]. W. B. Kleijn, F. SC Lim, A. Luebs, et al. Wavenet based low rate speech coding, in Proc. International Conference on Acous- tics, Speech, and Signal Processing (ICASSP), 2018, pp. 676– 680.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

[10]. Xavi Gonzalvo, Siamak Tazari, Chun-an Chan, et al. Recent advances in google real-time hmm-driven unit selection synthesiser. Interspeech, 2016,pp. 2238–2242.

[11]. Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, Fftnet: A real- time speaker-dependent neural vocoder, in Proc. Interna- tional Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp. 2251–2255.

[12]. S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, Courville, and Y. Bengio, SampleRNN: An unconditional end-to- end neural audio generation model, arXiv:1612.07837, 2016.

[13]. Fu Qiang. Research on speech parameter representation and objective quality evaluation [D]. Xidian University 2001