A STUDY ON GENERATIVE MODELS FOR SYNTHESIS OF REALISTIC VOICES USING DEEP LEARNING

Snehal Chaudhari[1], Srushti Pawar[1] and Prof. Tushar Rane[2]

[1]*Dept. of Information Technology, Pune Institute of Computer Technology Pune, Maharashtra, India*
[2]*Professor Dept. of Information Technology, Pune Institute of Computer Technology Pune, Maharashtra, India* -----

Abstract - *Every voice assistant or any application which tries to produce human-like speech, sounds unvarying. Though their voices are friendly and soothing, those voices lack in giving a natural human-like effect. These voices appear to be robotic and monotonous. This problem inspired us to model a human-like voice trained from the voices of people around us and make an effort to generate a realistic voice that sounds like natural human speech. Natural human-like voice generation is a highly desired feature for speech interfaces to have a personalized touch & to familiarize ourselves with the application due to its human-like voice. Recent studies in this field have shown impressive results in areas of text-to-speech and natural voice generation. A deep learning model needs several hours of professionally recorded speech from a single speaker, which is highly expensive as it involves a large amount of data. Amazingly,with the training of only small seconds of guiding speech, the studied neural network-based system involves three-stages that helps us to generate a voice never seen before during training. The studied model aims at passing the information learnt from different speakers to a distinctively trained speaker encoder to the multi-speaker text to speech task, and generate real sounding speech from other speakers. Wavenet aims at modelling the sync between the required target audio and human acoustic characteristics from input audio. Voice characteristics of a speaker are identified by speaker embeddings that are used to generate speech of the voices distinguishable from trained speaker voices.*

***KeyWords*: Embeddings, Mel-spectrogram, WaveNet, Vocoder, WaveRNN, etc.**

1. INTRODUCTION

Generation of realistic voices is a process for generating a particular speaker's voice while not hampering the original linguistic characteristics. The studied model is a deep neural network-based system for realistic voice generation. The system contains individually trained speaker encoders that exploit the characteristics of individual real audios. Then synthesizer networks produce excellent-standard speech by mapping exploited characteristics and they also pass the learnt information from *different speakers* to speakers not encountered during training.

2. RELATED WORK

We briefly present related work regarding audio generation. Deep learning models are dominating in generative models space. Text-to-speech (TTS), the technique of generating machine produced speech, is done using deep learning.

In the late 90's, during the beginning of Text to speech generation, the statistical parametric speech synthesis (SPSS) was considered to be productive because statistical generative models are useful in mapping the existing link between the characteristics of source text and the output acoustic characteristics. A fully built SPSS design comprises a pipeline that exploits characteristics from the text to generate audio and also redefining and shaping of waveforms of voice samples from the sound characteristics generated by vocoder.

Earlier text-to-speech techniques were based on mathematical data calculated from the audio waveform and on the acoustic characteristics of sound to compare with different models. The metrics created from this are poorly related with the human interpretation of sound and do not produce expected naturalness. But later this technique is improved by clustering the linguistic characteristics exploited from the source text with a decision tree, and per cluster a hidden Markov model is trained, it created a normal distribution over parameters of spectrogram, their derivative, second derivative and a binary flag that suggests which components of the generated audio ought to contain voice[6]. But this base technique which was considered to be the state of the artwork for SPSS, turned inferior in terms of the naturalness of the synthesized speech in comparison with the well-known concatenative approaches.

At this point feed-forward deep neural networks come into the picture. In applied machine learning, deep learning models have become predominant. Text-to-speech (TTS), the process of synthesizing artificial speech is modelled using deep learning. Recently, Deep models that produce more natural-sounding speech than the traditional con catenative approaches. There are two existing methods for Text-to-Speech(TTS) conversion.

2. 1. Concatenative TTS

This method requires voice samples with high-quality which later are then concatenated to form a full fledged speech. The audio synthesized shows clarity but it does not show emotions. One of the prominent prerequisites is that these systems need large databases and to form the words hardcoding needs to be done. [3].

2. 2. Parametric TTS

Due to the unfeasible large data and storage requirements and development time, the concatenative TTS is not flexible. Hence, a statistical mathematics based method was developed. This technique consists of the synthesis of speech by integrating parameters like magnitude spectrum, fundamental frequency, etc. and using them to generate speech. There are two stages of a parametric TTS[6] system. The starting step is to exploit linguistic characteristics after processing the text. These characteristics can be phonemes, duration, etc. The next step deals with exploiting vocoder characteristics that showcase the speech signal. These characteristics can be cepstra, fundamental frequency, spectrogram, etc. With the linguistic characteristics these parameters are fed in the mathematics based model - "Vocoder". To generate the audio waveform, vocoder takes these characteristics and multiple complex transformations are done on these features. During the generation of waveform, the parameters like prosody, phase, etc are estimated by the vocoder. Thus speech synthesized this way is more feasible and very modular. [3].

3. DIFFERENT METHODS FOR GENERATING AUDIO SAMPLES

3. 1 WaveNet

Deep learning models have proven extremely efficient in learning the inherent properties of data. As Aaron van den [3] suggested, WaveNet is primarily a deep neural network for generating raw audio waveforms. It's a probabilistic and autoregressive model. . WaveNet [3] is a convolution model that generates all p(xt|x<t) in one forward pass using causal or masked convolutions [10][11]. Each causal convolutional layer can process its inputs in parallel, which means that these architectures can be trained very quickly compared to RNNs [12], which can only be updated sequentially. With a predictive distribution for each audio sample taking into account all the previous ones; It has been shown that with tens of thousands of sound samples per second, it can be effectively trained on data. This proves substantially that it sounds much more natural when applied to text-to-speech than the current parametric and concatenative systems. It has been discovered that when trained to model music, it generates original and often quite realistic musical fragments. It has been discovered that when programmed to model music, it generates original and often quite realistic musical fragments. WaveNet can produce raw voice signals with subjective realism. It's a Generative model that works with the raw audio waveform directly. It is a fully convolutional neural network with various expansion factors that allow its receptive field to expand exponentially. The speakers record the input sequences for training. The network sample and value are derived from the probability distribution of the network. The value is then returned to the input and a new forecast is created for the next step. It produces complex, realistic-sounding sound, but requires a lot of computing power.

In [13] the authors attempted to improve the sound quality of WaveNet for use in production by sampling the sampled mixture of the logistic distribution presented in [14], also improving the fidelity by increasing the sampling rate of the audio from 16 kHz to 24 kHz and and increasing the dilated convolution filter size from 2 to 3 eating a WaveNet with a wider receiving field.

3. 2 Parallel WaveNet

But WaveNet is an autoregressive model which uses ancestral sampling; the generation of patterns remains in the inherent sequence and is therefore slow although its built-in structure allows for fast parallel learning. Thus [13] proposed an alternative approach for fast and parallel sound generation that is Parallel WaveNet. Inverse-autoregressive flows (IAFs) [15] are stochastic generative fashions whose latent variables are organized so that each one factors of an high-dimensional observable pattern may be generated in parallel. IAFs are a sort of normalising flow[16] that models a multivariate distribution. In general, stream normalization may require multiple iterations to turn unrelated noise into structured patterns, with the output produced by the stream at each iteration being input to the next [16]. For IAFs, this is less of an issue because the autoregressive latents can generate large structures in a single pass. The model's accuracy is increased by employing up to four flow iterations in[13].

3. 3 Probability Density Distillation

Probability Density Distillation which is a new approach for training a parallel feed-forward network from a already trained WaveNet with no substantial quality difference, as described in the study [13]. The claim which was made through the delving is that the system is able of manifest high-fidelity audio samples which are more than 20 times faster than real- time, and is emplaced online by Google Assistant, including serving multiple English and Japanese voices. Because the inference technique for estimating log-likelihoods is sequential and sluggish, it is impractical to train the parallel WaveNet model directly with maximum likelihood. [13] introduced a form of neural network distillation [18] that uses a previously trained WaveNet as a ' teacher ' from which a connate WaveNet "student" can efficiently learn. The elementary idea is that the student tries to compare the probabilities of his own samples according to the distribution the teacher has learned. This is similar to Generative Adversarial Networks [19] with students acting as creators and teachers acting as discriminators. But on the other hand, students are not trying to mislead the teacher in a contradictory way; rather, he cooperates by trying to match the teacher's probabilities. In the end, the teacher is kept stationary, instead of being trained in parallel with the students, and both models yield compliant homogenized distributions.

3. 4 WaveRNN

WaveRNN is the unique recurrent neural network. It consists of a dual Softmax layer that helps achieve the quality of the WaveNet model mentioned above. The WaveRNN has been modified to lower the amount of weights in order to test high-constancy sound on a CPU in a progressive manner.

Experimental results from different WaveNet implementation

For this experiment we referred to the results from [3] which are listed in Table-1. The same single-speaker speech databases from which Google's North American English TTS systems are built. The North American English dataset contains 24. 6 hours of speech data which was spoken by professional female speakers.

Method	Mean opinion score
WaveNet[3]	4. 21 ± 0. 081
Hidden Markov model-driven concatenative [3]	3. 86 ± 0. 137
Distilled WaveNet[13]	4. 41 ± 0. 078

Table -1: MOS based comparison of WaveNet distillation with the autoregressive teacher WaveNet, HMM concatenative

3. 5 Tacotron

Tacotron is an end-to-end generative text-to-speech model that synthesizes speech directly from textual content / characters, according to the study [1]. With given pairs, the model can be trained fully from a jar with haphazard initialization. We propose some fundamental strategies for optimising the sequence-to-sequence framework for this difficult endeavour. Tacotron surpasses a parametric production system of naturalness. Tacotron also generates speech at the frame level, significantly faster than sample-level autoregressive technique. Tacotron is a recurrent series-to-series version that predicts a mel spectrogram from textual content. It functions as an encoder-decoder shape that is bridged via means of a location-sensitive attention mechanism [20]. The text sequence's individual characters are initially inserted as vectors. Convolutional layers are added after that to extend the range of a single encoder frame. Then to create the encoder output frames, these frames are run through a bidirectional LSTM. An encoder, an attention-based decoder, and a post-processing net are all part of the architecture suggested in [1]. The encoder's purpose is to extract reliable text representations in a sequential order. The decoder uses a weighted sum of the encoder outputs to operate. The task of the post-processing network is to convert a sequence-to-sequence target into a composable target into a waveform.

3. 6 Tacotron 2

Google created the Tacotron 2 model, which was an improved version of the Tacotron 1. The Tacotron 2 model is in charge of generating speech synthesis directly from the characters provided. All we have to do now is give the audio and text pair to the model to train with. It determines the language rules based on the audio and text input. It creates a mel spectrogram for the supplied text and uses WaveNet to generate the words.

3. 7 Melnet

A generative model was proposed by Sean Vasquez [9] called MelNet which was primarily used for fetching the longer-term dependencies compared to present end-to-end models. Audio waveforms with complicated structures and transitory timescales is one of the challenges for generative models. Recorded local structure helps achieve high-fidelity voice, while larger-span dependencies of multiple of thousands of time steps must be recorded for globally consistent audio production. Existing generative waveform models such as SampleRNN [17] and WaveNet [3] are well suited for modeling local dependencies, but fall short in the ability to capture high-level structures that emerge in seconds. This is achieved by authors by modeling 2D time frequency portrayals, such as spectrograms, instead of 1D waveforms in the time domain. The dependencies, which include tens of thousands of time steps in waveforms, only include hundreds of time steps in spectrograms. Therefore, the spectrogram models can generate unconditional samples of voice and music with consistency for several seconds. It allows full-fledged text-to-speech conversion, a task that has not yet proven to be feasible by time-domain models.

3. 8 Wavenet Conditioning based On Mel Spectrogram Predictions

A neural network architecture for synthesizing speech directly from text is described by Jonathan Shen and others in [2] in a model called -"Tacotron 2". This model consists of a feature prediction network which is a recurrent sequence-to-sequence network in which character embeddings are mapped into mel-scale spectrograms and subsequently an improvised Wavenet Model is then used as -'Vocoder' to generate time-domain waveforms. The effects of using Mel spectrograms as the conditioning input for WaveNet are evaluated in place of F0 functions,duration,and the voice. He also explains that the use of the compact Intermediate Speaker contributes to reducing the size of the WaveNet architecture remarkably.

A sequence-to-sequence architecture[23] called -'Tacatron' [1] is used in generating huge spectrograms from a series of characters. Using single neural network which is trained from data individually and replacing the generation of these acoustic and linguistic characteristics, the pre-existing speech generation pipeline can be simplified. The Griffin-Lim algorithm[24] is used in phase estimation by Tacatron in order to vocode the obtained result. An integrated approach is described by author to generate speech which includes earlier approaches which are sequence-to-sequence Tacotron-style model [1] with improvised WaveNet vocoder [4][21][22]

An internal USA English dataset[1] is used for training a model which contains 24. 6 hours of speech voices from an individual female speaker. The texts present in these datasets are spelled out. The texts are normalised and then the models are trained on them.

Table-2 shows a comparison of the method against various prior systems[2].

System	Mean Opinion Score(MOS)
Parametric	3. 492 ± 0. 096
Tacotron (Griffin-Lim)	4. 001 ± 0. 087
Concatenative	4. 166 ± 0. 091
WaveNet (Linguistic)	4. 341 ± 0. 051
Ground truth	4. 582 ± 0. 053
Tacotron 2	4. 526 ± 0. 066

Table-2: MOS evaluations with audio intervals computed from the t-distribution for various systems.

4. TRANSFERRING LEARNING FROM SPEAKER VERIFICATION TO MULTISPEAKER TEXT-TO-SPEECH SYNTHESIS

Natural speech synthesis requires the preparation of a large number of pairs of high-quality speech transcripts, and multi-speaker support typically requires tens of minutes of training data per speaker [5]. It is not practical to record a large amount of high-quality data for many speakers.

Of all, one of the most promising approaches appeared to be that of Ye Jia et al. [5], which consists of decoupling speaker modeling from speech synthesis by independently training a discriminatory speaker embedding network that captures the space of speaker characteristics and training a high-quality TTS model on a smaller dataset conditioned on the representation learned by the first network[5]. Network decoupling facilitates them to learn unbiased facts, which eliminates the need for high-quality multi-speaker training data. To decide whether the same speaker was speaking in two different utterances, the speaker integration network is trained in a speaker verification task. . Unlike the succeeding TTS model, this network is trained on untranscribed speech with reverberation and background noise by a large number of speakers.

4. 1 Module Split-up

The studied system is composed of three independently trained neural networks illustrated in Fig-1, listed as follows:

1) A recurrent speaker encoder that computes a fixed-dimension vector from a voice signal.
2) A sequence-to-sequence synthesizer, which predicts a mel spectrogram from a succession of grapheme or phoneme inputs, conditioned on the speaker embedding vector, and
3) An autoregressive WaveNet vocoder, which converts the spectrogram into time domain waveforms.

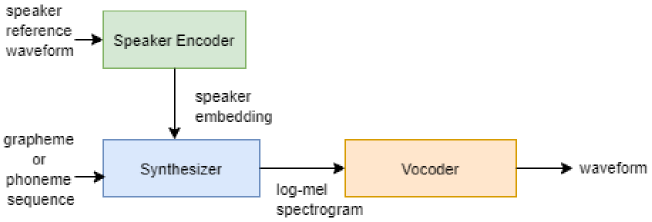

Fig -1: Model overview. Each of the three components are trained independently

4. 1. 1 Speaker Encoder

It's an encoding technique that uses unseen speakers to deduce the speaker's embedding directly. Embedding is necessary to condition the synthesis network to the identity of the speaker. This encoder converts speech frames of any length into a fixed-dimension embedding vector, known as a d-vector [7].

An N-layer LSTM with sufficient hidden nodes, followed by a projection layer, could be used as the model. It optimises a generalised end-to-end speaker verification loss, resulting in high similarity between embeddings of the same speaker while embeddings of different speakers are spread out across the embedding space. The inputs to the model are the audio window with a certain width and steps. The output is the L2 normalized hidden state of the last layer, which finally ends up during a vector.

In this implementation additional capabilities a ReLU layer is added before normalization, with the goal of making the embeddings sparse and therefore easier to interpret.

4. 1. 2 Synthesizer Network

A collection of pairs of target audio and textual content are used to train the synthesizer network[2]. A mapping of text to a series of phonemes occurs at the input stage, which ends up in quicker convergence and enhanced pronunciations. The network is trained on a transmission learning profile, where the speaker integration of the received audio is extracted using a trained speaker encoder. The embeddings received from the encoder are utilized in training a synthesizer which takes the text and phonemes as input. Then the synthesizer generates a Mel-spectrogram from the phonemes of the target sound[1].

Individual characters of a text string are integrated as vectors for the first time. Convolutional layers are added after that to extend the range of a single encoder frame To create the encoder output frames, these frames are run through a bidirectional LSTM. After that, every frame output by the encoder is concatenated with a speaker embedding.

To generate the decoder input frames, the attentiveness mechanism pays attention to the encoder output frames. The model is autoregressive because each decoder input frame is concatenated with the preceding decoder frame output and routed through a pre-net. Before being projected to a single mel spectrogram frame, this concatenated vector passes through two unidirectional LSTM layers. By projecting the same vector to a scalar and emitting a value above a specific threshold, the network may predict when it should stop creating frames on its own. The entire frame sequence passes through a residual postnet before becoming the mel spectrogram.

4. 1. 3 Vocoder

Mel-spectrogram characteristic representation is inverted into time-domain audio waveform samples using a new improvised form of the Wavenet. A ReLU activation function is used on the Wavenet output which is later passed to linear projection to estimate the parameters (log scale, mixture weight,mean) for individual components. A negative log-likelihood is used to calculate the loss. All of the related information for the high-standard generation of voice acquired by the mel spectrogram predicted using the synthesizer network. By simply training data from distinct speakers,construction of a multispeaker vocoder is allowed.

The generated mel spectrogram from synthesizer network are given as inputs to the vocoder with the ground truth audio as target. The available utterances are divided into small segments of predefined length and the synthesis is done in parallel way for all segments. Folding process is done in order to preserve the data in each segment. The folded segments are then forwarded by the model. A mel spectrogram and its related waveform in each training step are divided into the same number of segments.

Waveform segment t-1 and the spectrogram segment required to be predicted are the inputs to the model. The output expected from the model is the waveform segment t of similar length. The mel spectrogram is passed through an upsampling network inorder to be similar in length to the target waveform that means preserving the number of mel channels.

5. CONCLUSIONS

Here the study presented different approaches to generate realistic voices. We discussed the evolution of previous TTS methods which relied on statistics to the modern deep learning based generative methods. WaveNet is mainly a deep neural network for raw audio waveform generation. For generating raw speech waveforms,a neural network based WaveNet is used. The model is completely autoregressive and probabilistic. A parallel feed-forward network can be trained using already trained WaveNet, this method is called probability density distillation.

Melnet enabled the production of globally consistent audio, larger-span dependencies spanning multiple thousands of timesteps must be captured, most of the methods covered only local structures. In conditioning wavenet based on mel spectrogram predictions author describes an integrated way to speech generation which consists of earlier approaches : a sequence-to-sequence Tacotron-style model which creates mel spectrograms,with improvised WaveNet vocoder.

Amongst all, one of the most promising approaches seemed the neural network-based system for multi speaker speech synthesis discussed. The system combines an independently trained speaker encoder network with a sequence-to-sequence TTS synthesis network and neural vocoder. By using the knowledge learned by the discriminative speaker encoder, the synthesizer is able to generate high quality speech not only for speakers seen during training, but also for speakers never seen before. After multiple evaluations based on a speaker verification system as well as subjective listening tests, it is concluded that the synthesized speech is reasonably similar to real speech from the target speakers, even on such unseen speakers.

REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. “Tacotron: Towards End-to-End Speech Synthesis”. Mar. 2017

[2] Jonathan Shen et al. “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions”. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Apr. 2018

[3] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. “WaveNet: A Generative Model for Raw Audio”. Sept. 2016

[4] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder, ” in Proc. Interspeech, 2017, pp. 1118–1122.

[5] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu. “Transfer Learning from Speaker Verification to Multi Speaker TextTo-Speech Synthesis”. In: Advances in Neural Information Processing Systems 31 (2018), 4485-4495 (June 2018).

[6] Derrick Mwiti. A 2019 Guide to Speech Synthesis with Deep Learning. 2019

[7] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. “End-to-End Text-Dependent Speaker Verification”. In: (Sept. 2015).

[8] Bajibabu Bollepalli, Lauri Juvela, and Paavo Alku. “Speaking style adaptation in Text-To-Speech synthesis using Sequence-to-sequence models with attention”. In: (Oct. 2018).

[9] S. Vasquez and M. Lewis, “MelNet: A generative model for audio in the frequency domain, ”, 2019.

[10] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: masked autoencoder for distribution estimation. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 881–889, 2015

[11] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks, 2016.

[12] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In Advances in Neural Information Processing Systems, pages 4790–4798, 2016.

[13] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu. Parallel WaveNet: Fast High-Fidelity Speech Synthesis

[14] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications, 2017

[15] Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow, 2016.

[16] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2015.

[17] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model, 2016.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial

nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[20] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, KyungHyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. CoRR, abs/1506. 07503, 2015.

[21] S. O. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, ¨ W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech, " CoRR, vol. abs/1705. 08947, 2017.

[22] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, ¨ S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech, " CoRR, vol. abs/1710. 07654, 2017.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks. , " in Proc. NIPS, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. , 2014, pp. 3104–3112.

[24] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform, " IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 236–243, 1984.