A TWO-STEP BACKWARD COMPATIBLE FULLBAND SPEECH ENHANCEMENT SYSTEM

Xu Zhang, Lianwu Chen, Xiguang Zheng, Xinlei Ren, Chen Zhang, Liang Guo, Bing Yu

Kuaishou Technology, Beijing, China

ABSTRACT

Speech enhancement methods based on deep learning have surpassed traditional methods. While many of these new approaches are operating on the wideband (16kHz) sample rate, a new fullband (48kHz) speech enhancement system is proposed in this paper. Compared to the existing fullband systems that utilize perceptually motivated features to train the fullband speech enhancement with a single network structure, the proposed system is a two-step system ensuring good fullband speech enhancement quality while backward compatible to the existing wideband systems.

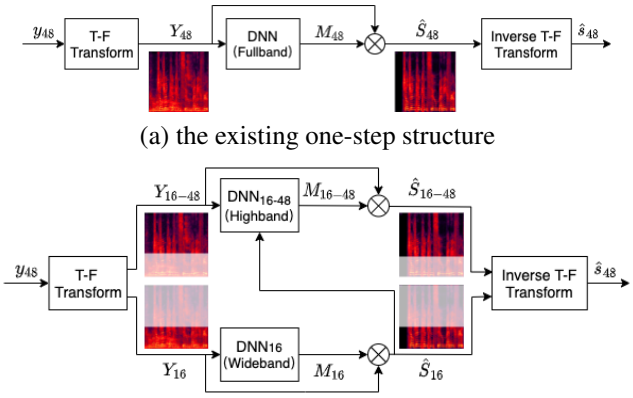
Index Terms— speech enhancement, noise suppression

1. INTRODUCTION

In recent years, deep learning based speech enhancement (SE) approaches have achieved significant improvement over the traditional signal processing based methods and become the main stream. While recent systems [1, 2, 3, 4, 5, 6] are performed well for the wideband (16kHz) speech [7, 8], only few of them [1, 2] have considered the fullband (48kHz) speech processing. Extending the speech bandwidth from wideband to super-wideband (32kHz) and fullband (48kHz) has widely practiced in contemporary speech codecs [9, 10] demonstrating improved speech quality [11]. The fullband speech compression system is usually built on-top of the existing wideband speech compression methods to ensure scalability under limited bitrates and backward compatibility of the existing wideband speech systems. In the remainder of this paper, the sampling frequency is abbreviated to '16k', '32k' and '48k' to avoid confusion with the signal frequency.

For the fullband SE systems, employing the psychoacoustically motivated features (such as the Bark-frequency cepstral coefficients (BFCC) in [1] and equivalent rectangular bandwidth (ERB) features in [2]) can significantly reduce the frequency-wise feature dimension and thus reduce the overall complexity of the fullband system. The drawback of using these features is the smeared frequency resolution in the high frequency region (8-24kHz). This leads to degraded performance in the high frequency region especially for the SE problem where the signal to noise ratio (SNR) in the high frequency region is usually much lower than the low frequency region (0-8kHz) as discussed in this paper.

To solve these problems, a two-step backward compatible fullband speech enhancement system is proposed. Compared to [1, 2], the contribution of the proposed method can be summarized by: a) the two-step system is proposed consisting of a pre-trained wideband (0-8kHz) SE system (the first step) and a highband (8-24kHz) SE system (the second step). Backward compatibility is ensured as any existing wideband SE systems can be used. b). the highband SE system takes the estimated wideband clean speech as the additional input alongside with the 8-24kHz noisy speech to provide extra guidance. Another practical benefit for this arrangement is that the publicly available speech and noise datasets are commonly released under 16k sample rate. Hence, dividing the fullband SE system into a wideband and a highband subsystem can optimally utilize the available resources ensuring good generalization ability.



(a) the existing one-step structure

(b) the proposed two-step structure

Fig. 1. Fullband speech enhancement architectures.

The proposed system is evaluated using two commonly used SE structures and compared with different one-step fullband training strategies using both psychoacoustically motivated and the short-time Fourier transform (STFT) features. As shown in the objective and subjective results, the proposed two-step system outperforms the other conditions.

2. RELATED WORKS

2.1. Problem formulation

A time domain noisy-reverberant mixture $y(t)$ is formed by:

$$y(t) = h(t) * s(t) + n(t) \tag{1}$$

where $s(t)$ is the speech signal, $h(t)$ is the transfer function from the talker to the microphone, $*$ denotes the convolution,

and $n(t)$ is the noise signal. The purpose of the SE system is to estimate $s(t)$ from $y(t)$ by removing $n(t)$ and $h(t)$.

2.2. One-step fullband SE system

Figure 1 (a) presents the existing one-step SE system. The time-frequency (T-F) transform is performed on the input noisy speech y to obtain the T-F domain representation Y. A common choice of the T-F domain features for a wideband SE system is the STFT feature as in [3, 4, 5, 6], and psychoacoustically motivated features such as the BFCCs in [1] and the ERB features in [2]. For the existing one-step fullband SE systems, the T-F domain noisy-reverberant speech Y_{48} is directly fed into a deep neural network (DNN) to estimate an ideal ratio mask or an ideal complex mask M_{48}. The T-F domain estimated clean speech $\hat{S}_{48}$ can be obtained by,

$$\hat{S}_{48} = Y_{48} \cdot M_{48} \tag{2}$$

The time-domain estimated clean speech $\hat{s}_{48}$ is formed using the inverse T-F transform, such as using the Inverse short-time Fourier transform.

3. PROPOSED APPROACH

3.1. Two-step fullband SE system

The proposed two-step fullband SE system is shown in Figure 1 (b). After the T-F transform, Y_{48} is divided into two parts, namely, Y_{16} and Y_{16-48}, representing the wideband and high-band time-frequency component below 8kHz and between 8kHz and 24kHz, respectively. The T-F domain estimated clean speech $\hat{S}_{48}$ can be obtained by,

$$\hat{S}_{48} = concat(\hat{S}_{16}, \hat{S}_{16-48}) \tag{3}$$

$$= concat(Y_{16} \cdot M_{16}, Y_{16-48} \cdot M_{16-48}) \tag{4}$$

where $\hat{S}_{16}$ is the final output signal of the standalone wideband SE network (DNN_{16}), M_{16} is the intermediate mask. Backward compatibility is ensured as any existing wideband SE networks can be employed, such as SE networks with the encoder-decoder structure [4, 5, 6] and the stacked LSTM structure [1, 2, 12]. Compared to the existing one-step systems which require 48k datasets, datesets with 16k sample rate can be fully utilized to ensure good generalization of the wideband SE. M_{16-48} is the mask of the higher frequency bands Y_{16-48}, which can be obtained by,

$$M_{16-48} = DNN_{16-48}(Y_{16-48}, \hat{S}_{16}) \tag{5}$$

The estimated wideband speech signal $\hat{S}_{16}$ is fed to the high-band SE network (DNN_{16-48}) alongside with Y_{16-48}. Using the estimated signal $\hat{S}_{16}$ to assist producing $\hat{S}_{16-48}$ can significantly improve the highband SE quality. Figure 2 shows the frequency dependent SNR (fSNR) of the fullband noisy speech signals based on the VCTK [13] test set, where the red and the yellow curves indicates 95% confidence interval (CI). As shown in the right side of the dashed vertical line,

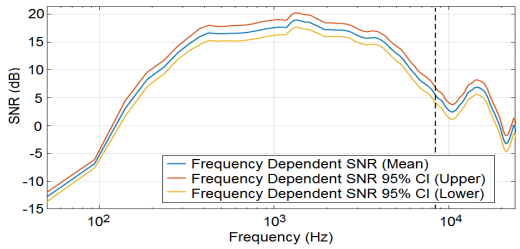


Fig. 2. Frequency dependent SNR for the VCTK 48k testset

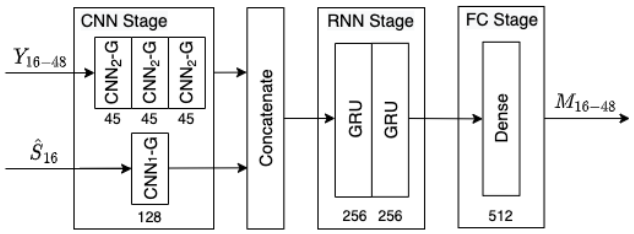


Fig. 3. The proposed CRNN-based highband SE network

the highband (8-24kHz) region has significantly lower fSNRs compared to the wideband (0-8kHz) region (left side of the vertical line). This demonstrates the motivation of employing $\hat{S}_{16}$ to assist the $\hat{S}_{16-48}$ estimation as shown in Figure 1 (b).

3.2. Highband SE network

As shown in Figure 1 (b) and 3, the proposed highband SE network DNN_{16-48} is based on the convolution recurrent network (CRN) [14] that serialized the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN).

The highband frequency component Y_{16-48} is the input to the upper CNN stream containing three 2-dimensional CNN groups (CNN_2-G) to extract the highband features. Each CNN_2-G is consisted by a 2D convolutional layer producing the feature maps with the 2D convolutional filters. The feature maps are fed to the exponential linear unit (ELU) activation function followed by a batch normalization layer and a dropout layer with the dropout rate set to 0.25. The input to the lower CNN stream of the CNN Stage is the estimated wideband speech $\hat{S}_{16}$ obtained from the pre-trained DNN_{16} network. The lower CNN stream contains a 1-dimensional CNN group formed by the identical structure as in CNN_2-G except the 2D-CNN is replaced by a 1D-CNN. The output of the upper 2D-CNN stream is reshaped and concatenated with the output of the lower CNN stream, and then fed into the recurrent layers. A dropout layer with 0.25 dropout rate is also employed after each recurrent layer. The final output tensor is produced using the output of the last recurrent layer by a feed-forward layer with the sigmoid activation function. Details of the hyper-parameters are presented in Table 1.

3.3. Loss function

As proposed in our earlier work [6], the ideal amplitude mask (IAM) weighted Mean Absolute Logarithmic Error (MALE) loss is applied to better suppress the noise for the low-SNR time-frequency bins. The loss is calculated by,

Table 1. Hyper-parameters for proposed DNN_{16-48}

Layer	CNN			RNN	FC
	Channels	Size	Stride	Units	Units
Conv2D	45	[4,3]	[1,2]		
Conv2D	45	[1,3]	[1,2]		
Conv2D	45	[1,3]	[1,2]		
Conv1D	128	1	1		
GRU				256	
GRU				256	
Dense					512

Table 2. The Encoder-Decoder Structure

Layer	Encoder (Decoder)			RNN	FC
	Channels	Size	Stride	Units	Units
Conv2D	45	[4,3]	[1,2]		
Conv2D	45	[1,3]	[1,2]		
Conv2D	45	[1,3]	[1,2]		
GRU				256	
GRU				256	
de-Conv2D	8	[1,5]	[1,2]		
de-Conv2D	1	[1,3]	[1,1]		
Dense					D

$$\mathcal{L}_{\text{IAM-MALE}} = \sum_t \sum_f W_{\text{IAM}}(t,f) \cdot |ln(X'_{mag}(t,f)+1) - ln(X_{mag}(t,f)+1)| \tag{6}$$

where,

$$W_{\text{IAM}}(t,f) = e^{a/(b+M(t,f))} \tag{7}$$

$$M(t,f) = (\frac{X_{mag}(t,f)}{Y_{mag}(t,f)})^{\gamma} \tag{8}$$

$X'_{mag}(t,f)$, $X_{mag}(t,f)$ and $Y_{mag}(t,f)$ are the predicted, clean and noisy speech amplitude spectrum, respectively. $M(t,f)$ is the IAM. According to [6], the value of γ, a and b are set to 1, 2 and 1, respectively.

4. DATASETS, EXPERIMENTS AND RESULTS

4.1. Datasets

The 48k clean speech and noise dataset of the experiments is from the DNS Challenge (INTERSPEECH 2021) which contains multiple languages and various transient and stationary noise types. We synthesize the training data with randomly selected SNRs chosen from -5 to 30 dB. To further improve the robustness in reverberated environments, the noisy and target signals are convolved with measured or simulated room impulse responses. Various EQ filters are also applied to simulate the frequency response of different microphones. To avoid speech distortion introduced by dereverberation, speech with 75ms early reflection is used as training target. This results 400 and 100 hours of noisy speech signals for training and validation. The 48k VCTK [13] test set containing 824 sentences under different SNRs is employed for testing.

4.2. Experimental setups

Four one-step baseline systems are compared in the experiments. Condition FFT_{768} is a one-step system (Figure 1 (a)) that employs the T-F representation of the input 48k noisy speech signal using a 1536-point STFT with 480 point stride (769 frequency bins for each frame). Conditions Mel_{48}, Mel_{64}, Mel_{80} are all one-step systems with the settings identical to Condition FFT_{768}, except the mel-scaled spectrogram with corresponding resolutions (i.e. 48, 64 and 80). It should be noted that condition Mel_{48} is designed to provide similar psycho-acoustically motivated frequency resolution compared to the existing works in [1, 2]. Condition Mel_{80} is selected as it is a common choice for fullband Text to Speech (TTS) [15] and singing voice synthesis [16].

Three two-step systems are evaluated. 1536-point STFT with 480-point stride is employed. Them resulting 769 frequency bins are divided into a wideband part (1-257 frequency bins) and a highband part (258-769 frequency bins). Condition $\text{TS-FFT}^{e16k}_{768}$ is implemented using the two-step structure as in Figure 1 (b) with the aid of estimated 16k wideband signal $\hat{S}_{16}$ for highband SE. Condition TS-FFT_{768} is the condition $\text{TS-FFT}^{e16k}_{768}$ without the aid of $\hat{S}_{16}$, and condition $\text{TS-FFT}^{n16k}_{768}$ is the the condition $\text{TS-FFT}^{e16k}_{768}$ with the estimated $\hat{S}_{16}$ replaced by noisy Y_{16}.

We use the Adam optimizer to minimize the IAM weighted loss for all systems. The learning rate is set to 0.001 and decays by multiplying a factor of 0.5 when the validation loss does not decrease for 5 epochs. Two kinds of SE structures are studied in the experiments for the one-step fullband DNN systems and the two-step wideband DNN_{16} systems.

- Exp1: Both fullband DNN and wideband DNN_{16} employ the encoder-decoder structure proposed in [6]. The hyper-parameters are shown in Table 2, where the number of FC units (D) depends on the corresponding frequency dimension for each condition.
- Exp2: Both fullband DNN and wideband DNN_{16} employ three sequentially stacked layers of LSTM[12], with the hidden units all set to 256.

PESQ-WB, STOI, SiSNR and SDR are employed as evaluation metrics, where PESQ-WB and STOI are only used for wideband evaluation. For wideband and highband evaluation, the fullband output signal of SE systems is first converted to the wideband and highband signal by a 0-8kHz low-pass filter and a 8-24kHz high-pass filter respectively.

4.3. Backward compatibility for wideband enhancement

The wideband (0-16k) SE performance is firstly evaluated as it provides the major contribution to the speech quality. As shown in Table 3, for the wideband region, the proposed $\text{TS-FFT}^{e16k}_{768}$ condition (TS-FFT_{768} and $\text{TS-FFT}^{n16k}_{768}$ has the same result in the wideband region) achieves the highest PESQ-WB, STOI, SiSNR and SDR scores. Compared to condition Mel_{48}, Mel_{64}, Mel_{80}, this result is reasonable as

Table 3. 0-16k results for Exp1

	PESQ-WB	STOI	SiSNR	SDR
Noisy	2.00	0.96	8.32	8.41
Mel_{48}	2.20	0.95	11.79	13.45
Mel_{64}	2.55	0.96	14.41	16.26
Mel_{80}	2.65	0.97	16.28	18.27
FFT_{768}	2.39	0.95	14.23	16.71
$\text{TS-FFT}_{768}^{e16k}$	2.70	0.97	16.63	19.44

Table 4. 16-48k & 0-48k results for Exp1

	highband(16-48k)		fullband(0-48k)	
	SiSNR	SDR	SiSNR	SDR
Noisy	−8.54	−3.40	8.27	8.30
Mel_{48}	−1.97	0.73	11.75	12.98
Mel_{64}	−0.91	1.23	14.33	15.62
Mel_{80}	−0.84	1.51	16.16	17.68
FFT_{768}	0.37	2.73	14.17	16.14
TS-FFT_{768}	−9.66	−1.27	16.39	18.51
$\text{TS-FFT}_{768}^{n16k}$	0.11	2.19	16.50	18.57
$\text{TS-FFT}_{768}^{e16k}$	2.14	3.55	16.51	18.58

the frequency resolution of condition $\text{TS-FFT}_{768}^{e16k}$ is higher. Compared to condition FFT_{768}, $\text{TS-FFT}_{768}^{e16k}$ is more effective since a standalone model is used for wideband SE. When training the fullband system FFT_{768} using the original STFT features, the feature dimension for the 8-24kHz frequency region is twice of the feature dimension for wideband region, resulting in sub-optimal performance in wideband region which is more perceptual important than the 8-24kHz frequency region. The results also justified the importance of using the psycho-acoustically motivated features in a one-step system. It should also be noted that the objective score of condition Mel_{80} is only slightly lower than condition $\text{TS-FFT}_{768}^{e16k}$. This demonstrates the effectiveness of the common choice of the mel-scaled feature as in [15] and [16].

4.4. Superior performance for fullband enhancement

The highband (16-48k) as well as the fullband (0-48k) speech quality is also evaluated. As shown in Table 4, $\text{TS-FFT}_{768}^{e16k}$ outperformed TS-FFT_{768} and $\text{TS-FFT}_{768}^{n16k}$ in highband indicating the importance of the $\hat{S}_{16}$ to the DNN_{16-48} network. The proposed two-step system $\text{TS-FFT}_{768}^{e16k}$ also outperformed other one-step systems for highband and fullband enhancement. Among these one-step systems, Condition FFT_{768} achieved the highest score due to its biased frequency resolution toward the 8-24kHz region as discussed earlier.

4.5. Generalization on different network structures

To further evaluate the generalization ability of the proposed two-step fullband SE system, the encoder-decoder structure in Section 4.3 and 4.4 is replaced by three stacked LSTM layers. The identical evaluation as in Table 3 and 4 is performed. The results are listed in Table 5 and 6. As shown, similar conclusions as in Section 4.3 and 4.4 can be reached using the stacked LSTM structure.

Table 5. 0-16k results for Exp2

	PESQ-WB	STOI	SiSNR	SDR
Noisy	2.00	0.96	8.32	8.41
Mel_{48}	2.11	0.95	11.69	13.20
Mel_{64}	2.50	0.96	14.42	16.14
Mel_{80}	2.55	0.97	16.24	18.05
FFT_{768}	2.40	0.96	16.20	18.30
$\text{TS-FFT}_{768}^{e16k}$	2.65	0.97	17.83	20.07

Table 6. 16-48k & 0-48k results for Exp2

	highband(16-48k)		fullband(0-48k)	
	SiSNR	SDR	SiSNR	SDR
Noisy	−8.54	−3.40	8.27	8.30
Mel_{48}	−0.63	1.19	11.65	12.76
Mel_{64}	−0.58	1.46	14.35	15.54
Mel_{80}	−0.22	1.73	16.14	17.53
FFT_{768}	0.54	2.89	16.13	17.73
TS-FFT_{768}	−9.66	−1.27	17.17	18.89
$\text{TS-FFT}_{768}^{n16k}$	0.11	2.19	17.32	18.95
$\text{TS-FFT}_{768}^{e16k}$	2.06	3.51	17.33	18.96

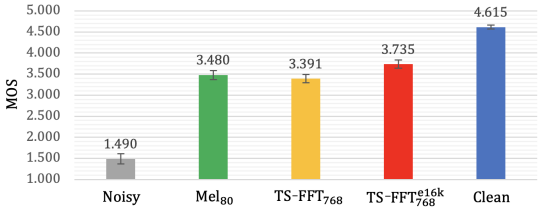


Fig. 4. Subjective Mean Opinion Score Test

4.6. Subjective Mean Opinion Score Test

A subjective MOS test is also conducted with 10 randomly selected noisy speech files processed by 3 of the representative conditions in Table 4. The noisy and the clean speech conditions are also considered resulting 5×10=50 48kHz audio samples. 15 listeners participated the MOS test. The results with 95% confidence interval are shown in the Figure 4. As shown, the proposed system (Condition $\text{TS-FFT}_{768}^{e16k}$) achieved statistically higher subjective MOS score than the one-step system (Mel_{80}) and the two-step system without using the aid of $\hat{S}_{16}$ (TS-FFT_{768}), demonstrating the effectiveness of the proposed system in Figure 1 (b).

5. CONCLUSIONS

A two-step backward compatible fullband speech enhancement system is proposed. Compared with the existing one-step structure, the proposed system can achieve comparable performance for the wideband speech quality while significantly outperformed the existing structure for the highband speech quality and ensured backward compatibility with the existing wideband SE approaches.

6. REFERENCES

[1] Jean-Marc Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.

[2] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *Interspeech 2020*, 2020, pp. 2482–2486.

[3] N. Westhausen and B. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," in *Proc. Interspeech 2020*, 2020, pp. 2477–2481.

[4] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network," in *ICASSP 2021*, 2021, pp. 6628–6632.

[5] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *Proc. Interspeech 2021*, 2021, pp. 2801–2805.

[6] X. Zhang, X Ren, X. Zheng, L. Chen, C. Zhang, L. Guo, and B. Yu, "Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss," in *Proc. Interspeech 2021*, 2021, pp. 2826–2830.

[7] C. Reddy, H. Dubey, K. Koishida, A.Asokan Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *Interspeech 2021*, 2021.

[8] C. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *ICASSP 2021*, 2021, pp. 6623–6627.

[9] Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry, "Definition of the opus audio codec," *RFC*, vol. 6716, pp. 1–326, 2012.

[10] Martin Dietz, Markus Multrus, Vaclav Eksler, Vladimir Malenovsky, Erik Norvell, Harald Pobloth, Lei Miao, Zhe Wang, Lasse Laaksonen, Adriana Vasilache, Yutaka Kamamoto, Kei Kikuiri, Stephane Ragot, Julien Faure, Hiroyuki Ehara, Vivek Rajendran, Venkatraman Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, "Overview of the evs codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5698–5702.

[11] Richard V. Cox, Simao Ferraz De Campos Neto, Claude Lamblin, and Mostafa Hashem Sherif, "Itu-t coders for wideband, superwideband, and fullband speech communication [series editorial]," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 106–109, 2009.

[12] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," pp. 91–99, 2015.

[13] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[14] Ke Tan and DeLiang Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, 2018, pp. 3229–3233.

[15] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *International Conference on Learning Representations*, 2018.

[16] Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, and Tie-Yan Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *ArXiv*, vol. abs/2009.01776, 2020.