INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-task Learning for End-to-end Noise-robust Bandwidth Extension

Nana Hou[1], *Chenglin Xu*[1,4], *Joey Tianyi Zhou*[3], *Eng Siong Chng*[1,2], *Haizhou Li*[4,5]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Temasek Laboratories, Nanyang Technological University, Singapore
[3]Institute of High Performance Computing (IHPC), A*STAR, Singapore
[4]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[5]Machine Listening Lab, University of Bremen, Germany

`nana001@e.ntu.edu.sg`

Abstract

Bandwidth extension aims to reconstruct wideband speech signals from narrowband inputs to improve perceptual quality. Prior studies mostly perform bandwidth extension under the assumption that the narrowband signals are clean without noise. The use of such extension techniques is greatly limited in practice when signals are corrupted by noise. To alleviate such problem, we propose an end-to-end time-domain framework for noise-robust bandwidth extension, that jointly optimizes a mask-based speech enhancement and an ideal bandwidth extension module with multi-task learning. The proposed framework avoids decomposing the signals into magnitude and phase spectra, therefore, requires no phase estimation. Experimental results show that the proposed method achieves 14.3% and 15.8% relative improvements over the best baseline in terms of perceptual evaluation of speech quality (PESQ) and log-spectral distortion (LSD), respectively. Furthermore, our method is 3 times more compact than the best baseline in terms of the number of parameters.

Index Terms: Noise-robust bandwidth extension, multi-task learning, time-domain masking, temporal convolutional network

1. Introduction

Speech signals with broader bandwidth provide higher perceptual quality and intelligibility. Bandwidth extension aims to recover the high-frequency information from narrowband signals, which is found useful in hearing aids design [1,2], speech recognition [3–5] and speaker verification [6,7].

Speech bandwidth extension methods, such as deep neural networks (DNN) [8,9], fully convolutional network [10,11], generative adversarial network (GAN) [12], and wavenet [13], mostly perform extension under ideal conditions with clean narrowband signals as inputs. This is called ideal bandwidth extension. However, in practice, speech signals are always corrupted by channel or ambient noise, for example, the received pilot speech via ultra high frequency (UHF) radio for air traffic control. Without addressing the noise issue, ideal bandwidth extension techniques are greatly limited in real-world applications.

A typical way to address the noise problem is to perform speech enhancement on the noisy narrowband signal first (Step 1), and ideal bandwidth extension next (Step 2), as illustrated in Figure 1. For example, there was a study to apply the iterative Vector Taylor Series (VTS) approximation algorithm [14] for feature enhancement, which is followed by a Gaussian mixture models or maximum a posterior models to reconstruct the wideband signals [15,16].

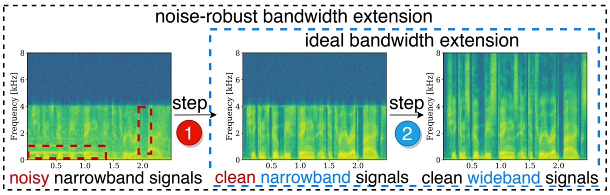


Figure 1: *The work flow of noise-robust bandwidth extension. In Step 1, the noisy narrowband signal is enhanced to remove noise. In Step 2, the enhanced narrowband signal is bandwidth-extended to generate the clean wideband signal.*

With the advent of deep learning, recent studies suggest [17] an unified approach that combines speech enhancement and bandwidth extension (UEE) in a joint training neural network. As shown in Figure 2(a), the UEE approach firstly applies a bi-directional long-short-term-memory (BLSTM) layer as the speech enhancement module to map the noisy narrowband input to enhanced narrowband features. Then, another BLSTM layer is applied as the ideal bandwidth extension module [18] to recover the missing high-frequency information from the enhanced narrowband features. The speech enhancement and bandwidth extension module are first trained separately as the pre-training, which are then fine-tuned with a single mean square error (MSE) loss between the clean wideband ground-truth and enhanced-plus-extended output. Overall, the UEE approach is implemented with a two-stage training scheme, and it also faces phase estimation difficulty just like other frequency domain techniques.

In this paper, we propose an end-to-end time-domain framework for noise-robust bandwidth extension, which is achieved by jointly optimizing mask-based speech enhancement and ideal bandwidth extension modules with a multi-task learning (MTL-MBE). As a time-domain technique, the proposed method inherently avoids phase estimation issues. Specifically, the noisy narrowband signal is firstly encoded into acoustic features instead of the short time Fourier transform (STFT). The speech enhancement module takes the acoustic features to estimate a mask and obtains the enhanced narrowband features for subsequent bandwidth extension. Two speech decoders are trained to reconstruct the enhanced narrowband and enhanced-plus-extended features into time-domain signals, in a similar way like what inverse STFT (iSTFT) does. The network is optimized with a multi-task learning [19–21] over both narrowband and wideband signals. To the best of our knowledge, this is the first work to explore noise-robust bandwidth extension in the time domain.

http://dx.doi.org/10.21437/Interspeech.2020-2022

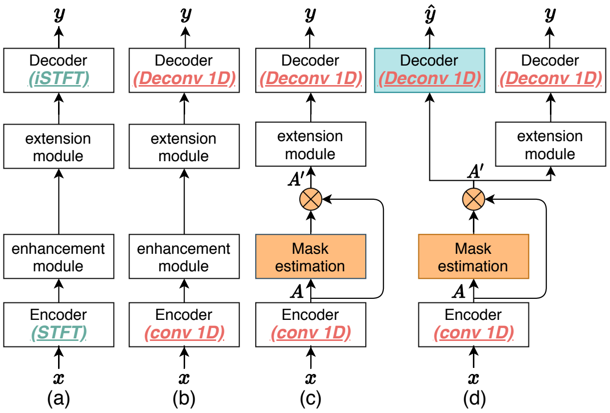


Figure 2: *Block diagrams of (a) frequency-domain noise-robust bandwidth extension, (b) time-domain noise-robust bandwidth extension, (c) time-domain mask-based noise-robust bandwidth extension (MBE), and (d) time-domain mask-based noise-robust bandwidth extension with multi-task learning (MTL-MBE).* $\otimes$ *is an operator that refers to the element-wise multiplication.*

2. Enhancement and Extension Multi-Task Learning

We now propose a time-domain masking for noise-robust bandwidth extension with multi-task learning (MTL-MBE), which is illustrated in Figure 2 (d).

We first examine a noise-robust bandwidth extension network in the time domain, which consists of a 1-D convolutional encoder to extract acoustic features from input speech, and a 1-D de-convolutional decoder to reconstruct waveforms from enhanced-plus-extended features, as shown in Figure 2(b). Such convolutional encoder-decoder-like structure is widely used in enhancement and separation tasks [22, 23]. The enhancement and extension are implemented as a pipeline of two similar regression, or mapping-based, neural networks. If trained jointly, their individual functions of the respective network are not clear. If trained separately, we face the same issue as other two-stage training schemes do.

2.1. Time-domain masking

To address the problem in the pipeline scheme of Figure 2(b), we propose a time-domain masking module to replace the mapping-based enhancement module, as shown in Figure 2(c), which has a unique architecture different from the extension module and is called MBE.

The time-domain masking aims to reduce the additive noise in noisy narrowband signals prior to extension. As shown in Figure 2(c), the input narrowband signal $x(t) \in \mathbb{R}^{1\times T}$ is encoded to a representation $A \in \mathbb{R}^{K\times M}$ by a 1-D CNN with $M(=512)$ filters and a filter size of $L(=16)$ samples with a stride of $L/2$ samples followed by a rectified linear unit (ReLU) activation function. Then, a time-domain masking W is estimated to suppress the additive noise in encoder representation A. It can be formulated as

$$A' = W \otimes A \tag{1}$$

where the estimated mask has the constraint $W \in [0,1]$, $\otimes$ denotes element-wise multiplication, and A' is the enhanced representation output from the mask estimation module.

The mask estimation module consists of a temporal convolutional network (TCN), which is illustrated in Figure 3. TCN is not the first time to be explored in speech enhancement. Prior work [24] utilized TCN as a regression module to map noisy input to clean signals, but their mapping-based framework is not suitable as an enhancement module here because it still suffers from the same problem as two-stage training schemes do. Therefore, we utilize TCN as a mask estimation module, which is a unique architecture different from the extension module.

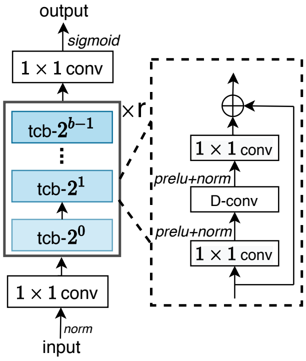


Figure 3: *Block diagram of temporal convolutional network (TCN). "tcb-*2^{b-1}*" denotes a temporal convolutional block (TCB) with the dilation of* 2^{b-1}*, where b is the total number of the TCB. "D-conv" is the dilated convolutional layers stacked in several TCBs to exponentially increase the dilation factors.* $\bigoplus$ *is the residual connection.*

As shown in Figure 3, the encoder representation A is firstly normalized by its mean and variance on channel dimension scaled by the trainable bias and gain parameters [25]. Then, a 1×1 CNN with $N(=128)$ filters is performed to adjust the number of channels for the inputs. To capture the long-range temporal information of the speech with a manageable number of parameters, dilated convolutional layers are stacked in several temporal convolutional blocks (TCB) by exponentially increasing the dilation factor. Each TCB, as shown in dot box of Figure 3, consists of two 1×1 CNNs and one dilated convolutional layer with a parametric rectified linear unit (PReLU) [26] activation function and normalization operation. The first 1×1 CNN (with 512 filters and 1×1 kernel size) determines the number of the input channels and the second 1×1 CNN (with 128 filters and 1×1 kernel size) adjusts the output channels from the dilated convolutional layer (with 512 filters and 1×3 kernel size). We form $b(=8)$ TCBs as a batch and repeat the batch for $r(=3)$ times in the TCN of mask estimation module. In each batch, the dilation factors of the deptwise convolutions in the b TCBs will be increased as $[2^0,\dots,2^{b-1}]$. To keep the estimated mask W in a consistent dimension with the encoder representations A, one 1×1 CNN (with 512 filters and 1×1 kernel size) is applied with a sigmoid activation function for ensuring that the estimated mask W ranges within $[0,1]$.

2.2. Multi-task learning

To provide cogent constraints for the enhancement module training, we further propose a multi-task loss for MBE as shown in Figure 2(d), that is designed for two training objectives: enhancement ("en") and extension ("ex"). It can be formulated as

$$\mathcal{L}_{total} = \lambda\mathcal{L}_{ex}(y,z) + (1-\lambda)\mathcal{L}_{en}(\hat{y},\hat{z}) \tag{2}$$

where y denotes the enhanced-plus-extended signal, while z is its corresponding clean wideband signal as ground-truth target for training; similarly $\hat{y}$ denotes the enhanced narrowband signal, while $\hat{z}$ is its corresponding clean narrowband signal as ground-truth target. All signals are sampled at 16kHz. λ is a trainable weighting parameter to balance the two loss functions. $\mathcal{L}_{ex}$ and $\mathcal{L}_{en}$ are optimized by the scale-invariant signal-to-distortion ratio (SI-SDR) loss function [27–29].

As shown in Figure 2(d), two loss functions are applied at different places of the processing pipeline. For enhancement objective, the enhanced representation A' is reconstructed to form an enhanced narrowband signal $\hat{y}$ by a 1-D de-convolutional decoder, which is supervised by $\hat{z}$. For extension objective, A' is taken by the extension module to form an enhanced-plus-extended signal y, which is supervised by the clean wideband signals z. The proposed network in Figure 2(d) is referred to as multi-task learning for mask-based bandwidth extension, or MTL-MBE.

The extension module consists of a TCN as shown in Figure 3, which is similar to the mask estimation module, except that there is no element-wise multiplication $\otimes$, and we use ReLU as the activation function for the last 1×1 CNN instead of a sigmoid function.

3. Experiments and Results

3.1. Database

We conduct evaluations on the public dataset by Valentini et al. [30], which is widely used for speech enhancement and bandwidth extension [31–35]. This dataset consists of 11,572 mono audio samples for training and 824 mono audio samples for testing. The speech is sampled at 16kHz. The training dataset has 40 noisy conditions (10 noise types × 4 signal-to-noise (SNR) values). The test dataset has 20 noise types that are different from the training set (5 new noise types × 4 new SNR values). The 2 speakers in the test dataset do not overlap the 28 speakers in the training dataset. We prepare both narrowband and wideband noisy data at 16kHz. We also prepare the clean wideband signals as ground-truth for the extension training and the clean narrowband signals as ground-truth for the enhancement training.

3.2. Experimental setup

3.2.1. Network configuration

During the training stage, the noisy narrowband waveforms were cut to 2-second long segments ($T = 32,000$ samples) for batch training. The network was optimized by the Adam algorithm [36]. The learning rate started from 0.001 and was halved when the loss increased on the development set for at least 3 epochs. Early stopping scheme was applied when the loss increased on the development set for 20 epochs.

Table 1: *PESQ, CSIG, CBAK, COVL, STOI and LSD in a comparative study of the proposed time-domain masking and multi-task loss.*

Methods / Metrics	Single-loss		Multi-loss
	MBE w/o mask	**MBE**	**MTL-MBE**
PESQ	2.02	2.46	2.55
CSIG	2.13	2.52	2.64
CBAK	2.11	3.14	3.21
COVL	2.04	2.38	2.46
STOI	0.92	0.94	0.94
LSD	2.82	2.44	2.29

3.2.2. Reference baselines

We implement three reference baselines. Two of them [8, 11] are for ideal bandwidth extension under noisy conditions. The other [17] is designed particularly for noise-robust bandwidth extension.

- **LSM** [8]: a 3-layer network that predicted the missing high-frequency components from the low-frequency log-spectrum in frequency domain. The missing high-frequency phase was recovered by the imaged phase of low-frequency signals.
- **DRCNN** [11]: a fully convolutional encoder-decoder framework that mapped narrowband signals to wideband in the time domain. To increase the time dimensions during upscaling, subpixel shuffling layers were introduced in the upsampling blocks. The skip connections were utilized to speed up training.
- **UEE** [17]: a unified speech enhancement module and bandwidth extension module in one frequency-domain framework that recovered the high-frequency signals from noisy narrowband signals, as shown in Figure 2(a).

We use the following metrics to evaluate the results. PESQ [37] stands for perceptual evaluation of the speech quality, ranging from -0.5 to 4.5. Three objective metrics that approximate mean opinion scores (MOSs) [38]: CSIG, CBAK and COVL. They are designed for signal distortion evaluation, noise distortion evaluation, and overall quality evaluation, respectively. Short-time objective intelligibility (STOI) [39] reflects the improvement of speech intelligibility. Log-spectral distortion (LSD) [40] is to measure the distance between reconstructed and target spectrum. Except LSD, higher scores are better for all metrics.

3.3. Results

3.3.1. Effect of the proposed time-domain masking

We first investigate how the proposed time-domain masking contributes to the framework MBE in Figure2(c) by experimenting with and without (w/o) the time-domain mask. For fair comparison, the single loss is utilized in this experiment and the results are summarized in Table 1. We observe that the performances of MBE w/o time-domain masking decrease sharply because the noise issue is not addressed. Under the constraint of the single loss, the MBE achieves 21.8% and 13.5% relative improvements in terms of PESQ and LSD, compared with MBE w/o mask. The experiment also confirms the need to perform enhancement prior to bandwidth extension operation.

3.3.2. Effect of the proposed multi-task loss

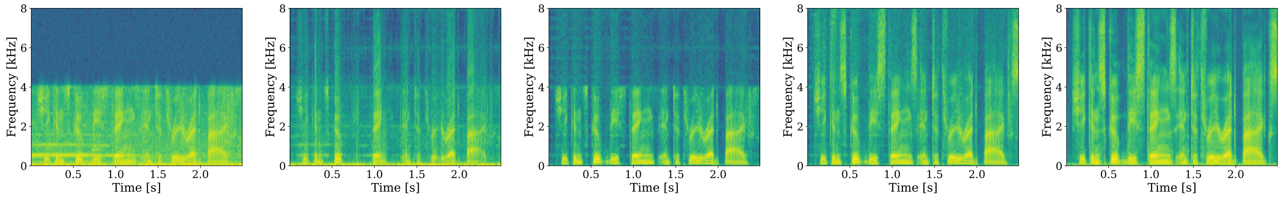
We further investigate how the proposed multi-task learning contributes to the noise-robust bandwidth extension. The comparative results of the MBE in Figure 2(c) and the MTL-MBE in Figure 2(d) are shown in Table 1. We observe that the performances are improved by utilizing the multi-task loss. Compared with the MBE, the MTL-MBE achieves 3.7% and 6.1% relative improvements in terms of PESQ and LSD. Such experiments show the performances of noise-robust bandwidth extension can be further improved by providing constraints for the enhancement module.

3.3.3. Overall comparisons

Table 2 summarizes the comparison between the proposed MTL-MBE in Figure 2(d) and other baselines in terms of PESQ,

Table 2: *A comparison of different techniques. "Designed conditions" refers to the conditions the method is designed for (clean or noisy). We perform all tests under noisy conditions. "#Paras" denotes the number of parameters of the model. "Feature type" denotes the types of narrowband inputs. "Spectrum" means that the approach is performed in frequency domain, while "waveform" means that time-domain signals are directly taken as inputs.*

Designed conditions	**Methods**	**#Paras**	**Feature type**	**PESQ**	**CSIG**	**CBAK**	**COVL**	**STOI**	**LSD**
clean	LSM [8]	13.38M	spectrum	1.79	2.45	2.32	2.09	0.92	2.80
clean	DRCNN [11]	56.41M	waveform	1.74	1.18	1.97	1.38	0.92	2.97
noisy	UEE [17]	22.42M	spectrum	2.23	2.27	2.39	2.17	0.93	2.72
noisy	**MTL-MBE**	**6.82M**	waveform	**2.55**	**2.64**	**3.21**	**2.46**	**0.94**	**2.29**



(a) (b) (c) (d) (e)

Figure 4: *The spectrograms of a sample (p232_005.wav) in the test set for (a) noisy-narrowband input, (b) the best baseline UEE, (c) enhanced narrowband result of MTL-MBE, (d) the enhanced-plus-extended result of MTL-MBE and (e) wideband signal (ground-truth).*

CSIG, CBAK, COVL, STOI and LSD. "LSM" and "DPRNN" are designed for bandwidth extension under clean conditions but we evaluate them under noisy conditions in this experiment. Their results reveal the limitation when working under noisy conditions. We observe that the proposed MTL-MBE achieves the best performance. Comparing with the UEE method [17], MTL-MBE achieves 14.3% and 15.8% relative improvements in terms of PESQ and LSD. Meanwhile, the parameter size of MTL-MBE is 3 times smaller than that of UEE.

We extract one speech sample from the test set to illustrate the differences of recovered enhanced-plus-extended signal between the best baseline UEE and the proposed MTL-MBE, as shown in Figure 4. We observe that MTL-MBE (see Figure 4(d)) produces cleaner signal at low-frequency and richer high frequency content than UEE (see Figure 4(b)). The intermediate enhanced-narrowband magnitude spectrum is also shown in 4(c). We also observe that the enhanced-narrowband representations constrained by multi-task supervision provide well-presented features for subsequent extension operation.

3.3.4. *Subjective evaluation*

Since the UEE presents the best baseline performances in the objective evaluation in Table 2, we only conduct an A/B preference test between the UEE and the proposed MTL-MBE to evaluate the signal quality and intelligibility for listening. We randomly select 20 pairs of listening examples and invite 10 subjects to choose their preference according to the quality and intelligibility. The percentage of the preferences is shown in Figure 5. We observe that the listeners clearly preferred the proposed MTL-MBE with a preference score of 84% to the best baseline UEE with a preference score of 10%. Most subjects significantly preferred the reconstructed wideband signals by the MTL-MBE with a significance level of $p < 0.05$, because the MTL-MBE produces cleaner signals at low-frequency and richer high-frequency content. Some listening examples are available at Github[1].

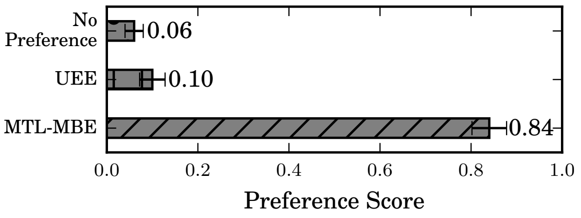


Figure 5: *The A/B preference test result of the recovered speech between the best baseline UEE and the proposed MTL-MBE. We conducted t-test using a significance level of $p < 0.05$, which is depicted with error bars.*

4. Conclusions

In this paper, we propose an end-to-end mask-based noise-robust bandwidth extension framework with multi-task learning (MTL-MBE). As a time-domain technique, the proposed MTL-MBE inherently avoids decomposing signals into magnitudes and phase spectra, and therefore requires no phase estimation. Experimental results show that the proposed MTL-MBE outperforms the prior work UEE in terms of PESQ and LSD with 3 times fewer parameters.

5. Acknowledgements

This work was supported by Air Traffic Management Research Institute of Nanyang Technological University, Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore; Neuromorphic Computing Programme (Grant No. A1687b0033), RIE2020 Advanced Manufacturing and Engineering Programmatic Grant. The work by H. Li is also partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

[1] `https://nanahou.github.io/mtl-mbe/`

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] C. Liu, Q.-J. Fu, and S. S. Narayanan, "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. EL77–EL83, 2009.

[3] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] P. S. Nidadavolu, C.-I. Lai, J. Villalba, and N. Dehak, "Investigation on bandwidth extension for speaker recognition." in *Interspeech*, 2018, pp. 1111–1115.

[5] D. Haws and X. Cui, "Cyclegan bandwidth extension acoustic modeling for automatic speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6780–6784.

[6] R. Kaminishi, H. Miyamoto, S. Shiota, and H. Kiya, "Investigation on blind bandwidth extension with a non-linear function and its evaluation of x-vector-based speaker verification," in *Proc. INTERSPEECH*, 2019, pp. 4055–4059.

[7] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," *Proc. Interspeech 2019*, pp. 406–410, 2019.

[8] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *ICASSP*. IEEE, 2015, pp. 4395–4399.

[9] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 71–83, 2017.

[10] Y. Gu and Z.-H. Ling, "Waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension." in *INTERSPEECH*, 2017, pp. 1123–1127.

[11] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *ICLR*, 2017.

[12] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *CoRR*, vol. abs/1903.09027, 2019. [Online]. Available: http://arxiv.org/abs/1903.09027

[13] A. Gupta, B. Shillingford, Y. Assael, and T. C. Walters, "Speech bandwidth extension with wavenet," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 205–208.

[14] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 733–736.

[15] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," in *EUSIPCO*, 2005.

[16] H. Seo, H.-G. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *ICASSP*. IEEE, 2014, pp. 6087–6091.

[17] B. Liu, J. Tao, and Y. Zheng, "A novel unified framework for speech enhancement and bandwidth extension based on jointly trained neural networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 11–15.

[18] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias icml," *Google Scholar Google Scholar Digital Library Digital Library*, 1993.

[20] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[21] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[22] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 327–334.

[28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[29] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *arXiv preprint arXiv:2004.08326*, 2020.

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, 2016, pp. 352–356.

[31] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *INTERSPEECH*, 2017.

[32] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *APSIPA ASC*. IEEE, 2018, pp. 1246–1251.

[33] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[34] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 667–672.

[35] X. Hao, C. Xu, N. Hou, L. Xie, E. S. Chng, and H. Li, "Time-domain neural network approach for speech bandwidth extension," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 866–870.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union*, 2005.

[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[40] L. R. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," *Englewood Cliffs, NJ, USA: Prentice-Hall*, 1993.