INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Noise-Robust Bandwidth Expansion for 8K Speech Recordings

Yin-Tse Lin[1], *Bo-Hao Su*[2], *Chi-Han Lin*[3], *Shih-Chan Kuo*[3], *Jyh-Shing Roger Jang*[3], *Chi-Chun Lee*[1,2]

[1]Institute of Communication Engineering, National Tsing Hua University, Taiwan
[2]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[3]E.SUN Financial Holding Co., Ltd., Taiwan

alexanderlin0625@gapp.nthu.edu.tw, borrissu@gapp.nthu.edu.tw,
finalspaceman-19590@esunbank.com, kakushawn-21659@esunbank.com, roger-21456@esunbank.com,
cclee@ee.nthu.edu.tw

Abstract

Speech recordings in call centers are narrowband and mixed with various noises. Developing a bandwidth expansion (BWE) model is important to mitigate the automated speech recognition (ASR) performance gap between the low and high sampling rate speech data. To further address the in-the-wild noise in call center settings, we propose an Embedding-Polished Wave-U-Net (EP-WUN) that includes an additional speech quality classifier to handle the noise and bandwidth expansion of 8k audio simultaneously. Our framework shows improved speech quality metrics on a well-known BWE dataset (Valentini-Botinhao corpus) when comparing to the current state-of-the-art noise-robust BWE model with 33% fewer parameters. It also achieves an 11.71% word error rate reduction when evaluating on a real-world interactive voice response system from the E.SUN bank.
Index Terms: Bandwidth expansion, Robust speech representation learning, Automated speech recognition

1. Introduction

Countless inbound and outbound calls are handled by call center agents every day for a variety of purposes. Most of these calls are recorded as they are valuable assets for the company. For instance, they can be utilized for agent training [1], precise telemarketing [2], and better customer service [3]. Given the large scale nature of call center recordings, sifting through each call manually is infeasible making ASR technology imperative. While many sophisticated ASR systems are available, most, if not all, of them are trained with a high-sampling rate (16k or 44.1k), these systems can not be applied directly to handle the unique characteristics of conventional telephony systems, i.e., the 8k low sampling rate and the real world noise.

Previous works have primarily concentrated on dealing with the issue of 8k narrowband through BWE. Approaches of BWE evolve from feature-based methods to model-based ones, and most current methods focus on reconstructing *super-resolution* waveform directly. For feature-based methods, Fukuda in [4] implicitly imposed narrowband information into the acoustic feature vector to form a mixed bandwidth feature. Yu et al. in [5] showed that the average distances and variances in the mix-band features are consistently smaller than the wideband ones, which helped improve ASR performance. For the model-based ones, Mantena in [6] proposed to attach a bandwidth embedding as a condition vector while learning the acoustic model . Besides, Heymans added 8k speech data to perform data augmentation [7], and it improved 8k narrowband ASR performances. Other recent works began to address this issue by directly generating the wideband speech waveform using a a variety of deep learning methods, e.g., deep neural network(DNN) [8, 9, 10], generative adversarial network(GAN) [11, 12], and neural vocoder [13].

While many of these works have achieved promising BWE results, most of the existing models ignore the effect of noise. However, real-world call center recordings are naturally noisy. There are only a few recent works focusing on noise-robust BWE; for example, UEE [14] is a joint training framework that combines BWE and speech enhancement (SE) by cascading two bi-directional long short-term memory (BLSTM) layers. Another recent state-of-the-art (SOTA) noise-robust BWE is MTL-MBE [15]. It is a multi-task framework that simultaneously learns to reconstruct the narrowband and wideband clean signal from the noisy narrowband signal. These two are one of the few studies that have demonstrated their BWE model's robustness in handling noise for 8k speech recordings.

In this work, we propose an Embedding-Polished Wave-U-Net (EP-WUN), which is a noise-robust BWE network that enhances a Wave-U-Net backbone BWE module with noise-robust learning. Specifically, besides expanding waveform along the temporal and frequency domains, our proposed EP-WUN is further trained with a modified triplet loss that encourages speech representation clean-up while performing BWE simultaneously. We evaluate our noise-robust BWE framework on the well-known English Valentini-Botinhao corpus [16]. Additionally, we perform in-the-wild ASR decoding tasks by utilizing our framework for an unseen proprietary real-world Mandarin Chinese interactive voice response (IVR) recording dataset gathered internally from the E.SUN bank. In summary, the overall contributions of our work are:

- We present a novel training method based on representation cleaning on the embedding space for BWE on 8k speech recordings.
- This is one of the first BWE work with evaluation carried out on a real-world call center recordings.
- Our noise-robust BWE not only outperforms the SOTA model but also requires 33% fewer model parameters.

Our results achieve competitive performances on all speech quality metrics, and 28.77% and 24.98% word error rate (WER) for English and Chinese compared to the SOTA MTL-MBE (38.63% and 36.69%). The following sections are organized as research methodology, experiment setup, and results, and the final section is the conclusion.

2. RESEARCH METHODOLOGY

In this section, we will describe our proposed Embedding-Polished Wave-U-Net ($EP-WUN$) for the task of noise-robust BWE. The overall framework is shown in the left side of Fig. 1. Our structure is mainly composed of two parts which are Wave-U-Net (WUN) and speech-quality classifier (SQC).

10.21437/Interspeech.2023-857

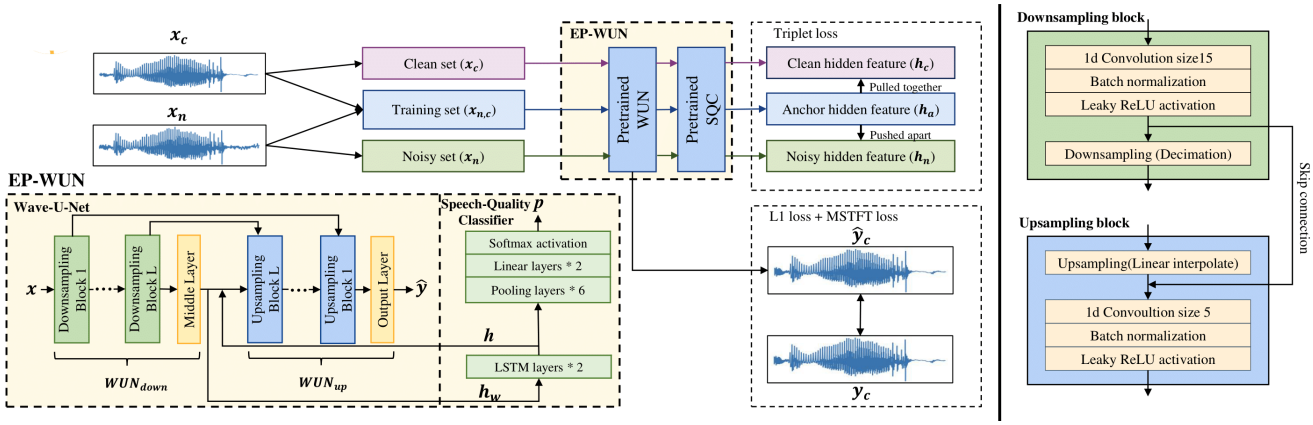

Figure 1: *Proposed $EP-WUN$ is composed of WUN and SQC. Robust learning is done by applying the modified triplet loss on the hidden feature h of SQC, where h_a is the anchor, and h_p, h_n comes from the clean and noisy speech respectively.*

WUN is a wav-to-wav model that handles the input speech in time domain, and SQC is a noisy versus clean speech classifier which helps distinguish the quality of the speech embedding. Here, we divide our method into the training stage and the evaluation stage.

2.1. Training stage

In our model, the training process is divided into three parts: WUN pre-training, SQC pre-training, and $EP-WUN$ training. We first define major symbols used in our work. The input, denoted as $x_{\{c,n\}}$, is the 8k speech signal for the BWE model, where c and n represents clean and noisy respectively. We use $y_{\{c,n\}}$ to be the corresponding 16k target speech signal for the input $x_{\{c,n\}}$, and we use $\hat{y}_{\{c,n\}}$ to represent the reconstructed 16k target output.

2.1.1. WUN Pre-Training

We use a conventional WUN [17] as our backbone model that has been shown to achieve outstanding performances in multiple speech tasks. Downsampling blocks, upsampling blocks, middle layers, and skip connections are basic components of WUN (refer to the right side of Fig. 1). The number of parameters in WUN is 1.49M.

In this pre-training stage, we train the WUN by pairing *noise_8k-noise_16k* and *clean_8k-clean_16k*. Here, we define WUN_{down} as a downsampling module, and WUN_{up} as an upsampling module. The loss used in this stage is shown below:

$$\mathcal{L}_{WUN} = \lambda\mathcal{L}_{wav}(y_{\{c,n\}}, \hat{y}_{\{c,n\}}) + \mathcal{L}_{MSTFT}(y_{\{c,n\}}, \hat{y}_{\{c,n\}}) \quad (1)$$

where λ is the weighted parameter between different losses, $\mathcal{L}_{wav}$ is L_1 loss used for time domain signal reconstruction, and the $\mathcal{L}_{MSTFT}$ is L_1 multi-resolution short-time Fourier transform loss used for the frequency domain spectrogram reconstruction. Specifically, we choose the window length from {240, 600, 1200}, fast Fourier transform (FFT) length from {512, 1024, 2048}, and hop size from {50, 120, 240} when calculating MSTFT loss.

2.1.2. SQC Pre-Training

The goal of SQC is to distinguish the quality (noisy versus clean) of an 8k speech and helps direct the hidden embedding toward *cleaned* representation. SQC consists of two layers of LSTM, followed by pooling layers, linear layers, and softmax activation (shown in Fig. 1). In this procedure, SQC would indicate whether the representation ($h_w = WUN_{down}(x_{n,c})$) is derived from the clean or noisy 8k speech data. In short, the SQC is a binary classifier for speech quality (clean versus noisy) trained with cross-entropy loss. The number of parameters in SQC is 4.09M.

2.1.3. EP-WUN Training

Finally, we combine the WUN and SQC to achieve noise-robust BWE. The goal is to train a single end-to-end network that performs wav-to-wav BWE while constraining the intermediate hidden representation to move closer to the *clean-domain* as indicated by SQC. This is done via a modified triplet loss term described below.

We leverage the ability of the pre-trained SQC for indicating whether the embedding is noisy or clean to further impose a modified triplet loss $\mathcal{L}_{m-trp}$ to force the hidden embedding of 8k speech audio to move toward a *cleaned* space. Our proposed $EP-WUN$ can be thought as a wav-to-wav BWE with an embedded representation cleaning mechanism. The proposed use of triplet loss is formulated below:

$$h_a = LSTM_{SQC}(WUN_{down}(x_{\{c,n\}})) \quad (2)$$

$$h_{c/n} = LSTM_{SQC}(WUN_{fix,down}(x_{\{c/n\}}) \quad (3)$$

$$\mathcal{L}_{m-trp} = max\{d(h_a, h_c) - d(h_a, h_n) + \delta, 0\} \quad (4)$$

where only the WUN_{down} is fine-tuned in this process, and $WUN_{fix,down}$ is set by the weight from pre-trained WUN. To clarify the difference, in the original triplet loss definition, the positive/negative samples are alternated according to the label of the anchor. In our case, the definition is slightly changed such that the positive and negative sets are always defined as clean and noisy speech for any anchor. We take the input sample through SQC to derive embedding h_a as an anchor, and force it to be close to the clean sample, h_c, and to be far from the noisy sample, h_n (equation 4). Note that h_c, h_n are embeddings derived from SQC, the margin parameter δ is set as 1, and d is the L_2 distance loss between the representations. After moving the anchor h_a representation to a clean space, we further feed this representation back to WUN_{up} to generate the bandwidth expanded signal, $\hat{y}_{\{c\}}$. The process can be written as:

$$\hat{y}_c = WUN_{up}(h_a + WUN_{down}(x_{\{c,n\}})) \quad (5)$$

Therefore, the overall training loss $\mathcal{L}_{EP-WUN}$ includes the following terms:

$$\mathcal{L}_{EP-WUN} = \lambda_t \mathcal{L}_{m-trp} + \lambda_w \mathcal{L}_{wav}(y_c, \hat{y}_c) + \mathcal{L}_{MSTFT}(y_c, \hat{y}_c) \quad (6)$$

where λ_t and λ_w are the weighted parameters between different losses, $\mathcal{L}_{wav}(y_c, \hat{y}_c)$ is the time domain L_1 loss, and $\mathcal{L}_{MSTFT}$ is the multi-resolutio frequency domain loss.

2.2. Evaluation Stage

In the evaluation stage, once we have the $EP - WUN$ model, we directly apply the model on 8k speech recordings to reconstruct the 16k speech data to be decoded by any existing 16k ASR model. The overall procedure is shown below:

$$\hat{y}_c = EP - WUN(x_{\{c,n\}}) \quad (7)$$

$$\hat{w}_1, ...\hat{w}_n = ASR(\hat{y}_c) \quad (8)$$

where $\hat{y}_c$ is the retrieved 16k signal and ASR is the speech recognition model that is trained on 16k speech data, and the $\hat{w}_n$ is the word sequence prediction.

3. EXPERIMENT

3.1. Database

3.1.1. Valentini-Botinhao

The public Valentini-Botinhao dataset is an English corpus consisting of 28/2 speakers for the training/testing set. The training set has 40 noisy conditions, and the testing set has another 20 different noisy conditions. We downsample the utterances to 16k as our training target y_c from the original 48k. For validation, we split 2 speakers from the training set. In our setup, we have an 8 hours 48 minutes training set, a 34 minutes testing set, and a 38 minutes validation set, which contains 108030, 824, 742 pieces of speech respectively. This database is used for both speech quality and ASR performance evaluation.

3.1.2. Formosa Speech in the Wild

Formosa Speech in the Wild (FSW) [20] is a large-scale Taiwanese Mandarin corpus that is collected from broadcast radio. There are around 3000 hours of spontaneous speech data which include the multi-genre shows. Besides, all the audios are sampled with 16kHz and divided into clean and other (noisy) sets. The overall dataset is split into 17 volumes with manual transcription. We randomly sample FSW to generate our training set and validation set to align with the size of Valentini-Botinhao. This dataset is used only for training BWE models and evaluation of ASR. The training set and validation set used is approximately 11 and 1 hours with 1700 and 130 pieces of speech.

3.1.3. E.SUN-IVR-cust

Our in-the-wild ASR test set is in Mandarin Chinese, called ESUN-IVR-cust. It was collected by the in-company IVR system in the call center of E.SUN bank. It contains conversations of real customers talking to the IVR chatbot about their calling intentions. These calls are recorded with 8k sampling rate. Note that each recording only contains the voice of a single customer, and it's content relates to financial services without any personal information, which has been verified several times by company employees. ESUN-IVR-cust has 1,228 recordings, whose max, min, and average durations are 42.72s, 3.02s, and 14.769s, respectively.

3.2. Experimental setup

In the following, we conduct the experiments on both English and Mandarin corpus. For English, the model is trained and evaluated on Valentini-Botinhao. For Mandarin, the model is trained on FSW and evaluated on the real-world E.SUN-IVR-cust. In both experiments, the training speech is cut into roughly 1 second segment, which follows the settings in Wave-U-Net [21]. All the models are optimized by Adam optimizer with a learning rate of 2e-4. In each stage, WUN, SQC, and $EP - WUN$ are trained with batch size $\{256, 128, 128\}$ for $\{500, 200, 500\}$ epochs respectively. λ, λ_t, and λ_w are set to $\{100, 8, 800\}$ respectively. The models and hyperparameters in each stage are selected by the validation performance. Network weights are randomly initialized. Our framework is trained on NVIDIA A100 80 GB. For each stage, the memory cost are approximately $\{1.98 \times 10^5, 6.01 \times 10^4, 2.24 \times 10^5\}$ MBs, and the time cost are about $\{2, 1, 3\}$ days respectively. The source code is also available on GitHub [1].

3.3. Baseline models

We include various baseline robust BWE frameworks as our comparison models, and each work is briefly introduced in the description. To compare fairly, all the experiments are evaluated using the common SE and BWE metrics. For perceptual quality, PESQ [22] is measured with a reference signal. STOI [23] represents speech intelligibility, and log spectral distance (LSD) measures the spectral reconstruction. For the predicted mean opinion score (MOS) of speech quality, we compute the CSIG, CBAK, and COVL scores [24]. Note that only for the value of LSD, the lower is better, otherwise the higher is better.

- **DNN** A deep neural network that estimates the spectral mapping function from narrowband to wideband.
- **DRN** A time domain U-Net structured deep residual network for audio super-resolution.
- **WUN** A classic U-Net network that is composed of repeated downsampling, upsampling, and skip connections to generate the wideband wave files.
- **UEE** An unified framework for both SE and BWE, which combines Griffin-Lim algorithm with a jointly trained model to reconstruct clean wideband speech.
- **MTL-MBE** An end-to-end time-domain framework for noise-robust BWE, which jointly optimizes mask-based SE and BWE modules with multitask learning.
- **EP-WUN -** $\mathcal{L}_{m-trp}$ $EP - WUN$ trained without the modified triplet loss term in the third stage.

4. RESULTS

4.1. Baseline comparison

All the results are summarized in Table 1, and we also indicate the training set conditions. Conventionally, BWE models were trained with clean speech only. Compared to conventional ones

[1] https://github.com/alexlinander/EP_WUN

Table 1: *Results of all metrics among different models are presented in this table, including the training condition. The upper arrow represents that higher is better, and vice versa.* $EP-WUN\ -\ \mathcal{L}_{m-trp}$ *here stands for* $EP-WUN$ *without modified triplet loss. Note that all results are evaluated under the noisy Valetini testset.* ** *and* * *indicate that p-value <0.001 and <0.05 respectively when compared to WUN.*

Model	Training Set	#Params	PESQ↑	STOI↑	LSD↓	CSIG↑	CBAK↑	COVL↑
DNN [18]	Clean	13.38M	1.79	0.92	2.8	2.45	2.32	2.09
DRN [19]	Clean	56.41M	1.74	0.92	2.97	1.18	1.97	1.38
WUN	Clean	1.49M	1.96	0.91	1.90	1.51	2.2	1.7
UEE [14]	Clean + Noisy	22.42M	2.23	0.93	2.72	2.27	2.39	2.17
MTL-MBE [15]	Clean + Noisy	6.82M	**2.55**	**0.94**	2.29	2.64	**3.21**	2.46
EP-WUN - $\mathcal{L}_{m-trp}$	Clean + Noisy	**4.58M**	2.15^{**}	0.92	$\mathbf{1.23}^{**}$	3.42^{**}	2.98^{**}	2.78^{**}
EP-WUN	Clean + Noisy	**4.58M**	2.25^{**}	0.92^{*}	$\mathbf{1.23}^{**}$	$\mathbf{3.50}^{**}$	2.94^{**}	$\mathbf{2.86}^{**}$

(DNN, DRN, WUN), our model outperforms in all the metrics, especially for CSIG (3.50), CBAK (2.94), and COVL (2.86) which are sensitive to noise. The result shows that traditional BWE models have limited capability in handling noisy speech.

While compared to the latest SOTA, MTL-MBE, our proposed $EP-WUN$ shows better performance on CSIG (3.50 v.s 2.64), COVL (2.86 v.s 2.46), and especially on LSD (1.23 v.s 2.29). The significant reduction (-1.06) in LSD also meets our expectations due to the MSTFT loss for spectral reconstruction. However, we find that metrics like PESQ (2.55), STOI (0.92), and CBAK (2.94) are slightly lower than MTL-MBE, and the reason might be the trade-off between model size and performance. It is worth mentioning that $EP-WUN$ not only achieves competitive speech quality metrics (some are even better) when compared to the current best MTL-MBE but also is a very light model. $EP-WUN$ reduces approximately 80% ($22.42M \rightarrow 4.58M$) and 33% ($6.82M \rightarrow 4.58M$) of model parameters as compared to UEE and MTL-MBE, respectively.

4.2. Ablation study

To verify the effectiveness of our proposed modified triplet loss, we further conducted an ablation study to compare the performance of $EP-WUN$ with and without the loss term during training. As shown in Table 1, we first look into the PESQ scores, the model with modified triplet loss (2.25) obtains 0.1 higher perceptual quality than the without one. For the commonly used objective quality measures for SE (CSIG, CBAK, and COVL), the model that incorporates the modified triplet loss term achieves higher performance on CSIG(3.50) and COVL(2.86), but slightly lower performance on CBAK.

Overall, it appears that the modified triplet loss term provided additional benefit to the model in terms of cleaning up the noisy embedding. This has resulted in improved performance across different speech quality metrics.

4.3. ASR performance evaluation

We further conduct ASR evaluation that is a more critical task in deploying such a technology in call-center setting. To fairly compare the ASR decoding performance, we utilize the standard ASRmodel with fglarge language model from Kaldi [25] that is pre-trained on Librispeech [26] for English, and the proprietary ASR model from E.SUN bank that is trained on multiple Chinese Corpora for Mandarin. The results are presented in Table 2, and the evaluation metrics are in word error rate (WER) and character error rate (CER).

From Table 2, the performance of our proposed $EP-WUN$ reduces approximately 10% WER than the MTL-MBE in both languages ($E: 38.63\% \rightarrow 28.77\%$, $M: 36.69\% \rightarrow 24.98\%$). Note that the testing set of Mandarin is the E.SUN-IVR-cust that is completely unseen during the training of our Chinese $EP-WUN$. It is quite promising to see our proposed model still achieves an outstanding performance in this challenging setting and implicates a higher robustness of $EP-WUN$ as compared to MTL-MBE. Another interesting observation is that the speech quality metrics in Table 1 do not directly correspond to the performance of the ASR result, i.e., higher PESQ/STOI does not translate for better WER. From the observed improved ASR results when using our proposed $EP-WUN$, we hypothesize that LSD might be the informative speech quality metric for ASR performance. A lower LSD measures indicates a better spectral reconstruction. A better spectral reconstruction during BWE potentially benefits the downstream ASR model, since most ASR relies on spectral information as input.

Table 2: *ASR performance comparison*

	English		Mandarin	
	WER	CER	WER	CER
MTL-MBE	38.63	24.59	36.69	24.21
EP-WUN	**28.77**	**17.57**	**24.98**	**13.34**

5. Conclusions

In this paper, we proposed an end-to-end network modified from the WUN to a noise-robust WUN to handle BWE for 8k speech recordings. Specifically, we imposed a modified triplet loss that adapts the intermediate WUN representation toward a clean space. This representation cleaning helps the conventional WUN BWE to handle noisy conditions on embedding domain. Our experiments showed the usefulness of the proposed loss term and consistent performance improvement across multiple speech quality metrics, and our method better translated to the ASR performances as compared to the current noise-robust SOTA in two different languages.

According to the indirect relationship between speech quality metrics and ASR results, we plan to design detailed experiments and analyses to demonstrate the hypothesis that the spectral-related metric is more feasible for ASR performance. A limitation of the current work is that we only consider the additive noise. Hence, to further extend the generalizability of our work, we would introduce other distortions such as room impulse response, reverberance, and clipping to investigate how the distortion type and intensity (SNR) affect the robustness of the model, especially in the scenario of call center 8k recordings.

6. References

[1] O. Amir, E. Kamar, A. Kolobov, and B. Grosz, "Interactive teaching strategies for agent training," in *In Proceedings of IJCAI 2016*, 2016.

[2] C. S. T. Koumétio, W. Cherif, and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," in *2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2018, pp. 1–6.

[3] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, "Automatic analysis of call-center conversations," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 453–459.

[4] T. Fukuda and S. Thomas, "Mixed bandwidth acoustic modeling leveraging knowledge distillation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 509–515.

[5] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[6] G. Mantena, O. Kalinli, O. Abdel-Hamid, and D. McAllaster, "Bandwidth embeddings for mixed-bandwidth speech recognition," *arXiv preprint arXiv:1909.02667*, 2019.

[7] W. Heymans, M. H. Davel, and C. v. Heerden, "Multi-style training for south african call centre audio," in *Southern African Conference for Artificial Intelligence Research*. Springer, 2021, pp. 111–124.

[8] S. Sulun and M. E. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 132–142, 2020.

[9] J. Abel, M. Strake, and T. Fingscheidt, "A simple cepstral domain dnn approach to artificial speech bandwidth extension," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5469–5473.

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

[11] E. Moliner and V. Välimäki, "Behm-gan: Bandwidth extension of historical music using generative adversarial networks," *arXiv preprint arXiv:2204.06478*, 2022.

[12] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5029–5033.

[13] H. Liu, W. Choi, X. Liu, Q. Kong, Q. Tian, and D. Wang, "Neural vocoder is all you need for speech super-resolution," *arXiv preprint arXiv:2203.14941*, 2022.

[14] 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.

[15] N. Hou, C. Xu, J. T. Zhou, E. S. Chng, and H. Li, "Multi-task learning for end-to-end noise-robust bandwidth extension." in *INTERSPEECH*, 2020, pp. 4069–4073.

[16] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[17] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End- to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 334–340.

[18] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4395–4399.

[19] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017.

[20] Y.-F. Liao, Y.-H. S. Chang, Y.-C. Lin, W.-H. Hsu, M. Pleva, and J. Juhar, "Formosa speech in the wild corpus for improving taiwanese mandarin speech-enabled human-computer interaction," *Journal of Signal Processing Systems*, vol. 92, no. 8, pp. 853–873, 2020.

[21] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A two-stage approach to speech bandwidth extension." in *Interspeech*, 2021, pp. 1689–1693.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[23] 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.

[24] 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, 2007.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.