



Large-Scale Mixed-Bandwidth Deep Neural Network Acoustic Modeling for Automatic Speech Recognition

Khoi-Nguyen C. Mac¹, Xiaodong Cui², Wei Zhang² and Michael Picheny²

¹ Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, IL 61801, USA

² IBM Research AI

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

knmac@illinois.edu, {cuix, weiz, picheny}@us.ibm.com

Abstract

In automatic speech recognition (ASR), wideband (WB) and narrowband (NB) speech signals with different sampling rates typically use separate acoustic models. Therefore mixed-bandwidth (MB) acoustic modeling has important practical values for ASR system deployment. In this paper, we extensively investigate large-scale MB deep neural network acoustic modeling for ASR using 1,150 hours of WB data and 2,300 hours of NB data. We study various MB strategies including downsampling, upsampling and bandwidth extension for MB acoustic modeling and evaluate their performance on 8 diverse WB and NB test sets from various application domains. To deal with the large amounts of training data, distributed training is carried out on multiple GPUs using synchronous data parallelism.

Index Terms: speech recognition, deep neural networks, mixed-bandwidth, bandwidth extension, parallel computing

1. Introduction

Wideband (WB) and narrowband (NB) speech signals are two types of input signals that widely exist in speech-related applications. In automatic speech recognition (ASR), acoustic models are usually separately trained for WB and NB speech data given their distinct spectral characteristics under different sampling rates. From the system deployment's perspective, one acoustic model for both WB and NB speech would be greatly preferred. In this paper, we investigate mixed-bandwidth (MB) acoustic modeling using neural networks with deep architectures.

The goal of MB acoustic modeling is to converge the WB and NB speech to one bandwidth from which acoustic modeling is carried out. This could be accomplished either by downsampling or upsampling. In this paper we are interested in exploring MB acoustic modeling using deep neural networks (DNNs) and we are interested in seeking answers to the following questions: **(1) To converge to one bandwidth, which strategy is better, downsampling or upsampling? (2) how would direct mixing perform? (3) how would bandwidth extension (BWE) help in this case?** Furthermore, we are interested in real-world cases where large amounts of WB and NB training data are available and their amounts may be unbalanced. Specifically, we investigate MB deep convolutional neural network (CNN) acoustic models using 1,150 hours of WB speech and 2,300 hours of NB speech. We evaluate the ASR performance on a wide variety of WB and NB test sets collected from diverse scenarios.

To study the impact of BWE to MB acoustic modeling, we use a CNN with a VGG architecture [1] to map the upsampled NB speech to WB speech. The CNN-based BWE network has its output connected to the input of an existing CNN acoustic

model. It is discriminatively trained under the cross-entropy (CE) criterion.

Training deep CNN acoustic models using approximately 3,500 hours of speech data is computationally demanding. We resort to parallel computing for stochastic gradient descent (SGD) based network optimization with multiple GPUs. The system design and engineering consideration will be addressed in the system implementation.

The remainder of the paper is organized as follows. Section 2 will discuss the related work on MB acoustic modeling. Section 3 gives the mathematical formulation of BWE. Section 4 is devoted to the system implementation including the model architectures and parallel computing. Experimental results are provided in Section 5 followed by a discussion and summary in Section 6.

2. Related Work

MB acoustic modeling for ASR has been investigated previously under various conditions [2, 3, 4, 5].

In [2], the NB data is used to leverage the training of WB acoustic models in a GMM-HMM framework and the missing components in upsampled NB speech features are dealt with using the expectation-maximization (EM) algorithm [6] for the MB GMM-HMM acoustic models. MB training in [3] follows a similar argument of [2] as a missing feature problem but the problem is addressed in a DNN-HMM framework where no explicit BWE is assumed. Acoustic models in both [2] and [3] are trained on a small amount training data (< 100 hours).

A joint MB training scheme is studied in [4, 5] where BWE is used for MB. Specifically, a fully connected (FC) feedforward DNN is used to capture the BWE mapping from NB to WB speech with MMSE-based pretraining based on parallel speech data. Also, joint training is conducted to further improve the performance. There are 1,000 hours of WB data and 1,000 hours of NB data used in [4] where experiments are conducted on one WB and one NB test sets. There are 900 hours MB training data used in [5] (300 hours for 6KHz, 8KHz and 16KHz data respectively).

One difference among the previous literature is whether to assume an explicit BWE mapping and how it would help to the ultimate performance of MB models. Since DNNs are universal approximators, DNN-based BWE has been used for mapping between NB and WB speech. However, if we assume no explicit BWE network but only increase the capacity of MB DNNs, would it give rise to similar performance? We would like to design the experiments and evaluate with a competitive BWE network in this work.

3. Bandwidth Extension

BWE has been an active research topic in communication and acoustics processing. NB speech signals, such as telephony speech signals, suffer from degraded quality and intelligibility due to the lack of high frequency spectral information eliminated by the low-pass band limitation of communication channels. Over the years, extensive research has been carried out on BWE to compensate this degradation so as to improve the speech quality and intelligibility [7, 8, 9, 10, 11, 12]. BWE aims to estimate the missing high frequency spectral components and, therefore, effectively “extend” the bandwidth of the speech signals. However, most of the work on BWE is optimized towards intelligibility, which may not be well aligned with ASR performance. In ASR-oriented BWE, we are interested in the mapping from a NB feature sequence to a WB feature sequence such as it improves the word error rate (WER).

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ denote a sequence of n NB features in certain feature domain, $\mathbf{x}_i \in \mathbb{R}^{d_x}$. We want to estimate a mapping function f_θ with some parameter θ to map the NB feature sequence \mathbf{X} to a WB feature sequence $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_n\}$, $\hat{\mathbf{y}}_i \in \mathbb{R}^{d_y}$, where $\hat{\mathbf{y}}_i = f_\theta(\mathbf{x}_i)$. A loss function

$$\mathcal{L}(l_i, \hat{\mathbf{y}}_i) \triangleq \mathcal{L}(l_i, f_\theta(\mathbf{x}_i)) \quad (1)$$

is defined to measure the closeness of the mapped WB features $\hat{\mathbf{y}}_i$ and their labels l_i . We want to optimize the parameter θ such that it minimizes the following empirical risk

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(l_i, f_\theta(\mathbf{x}_i)). \quad (2)$$

Depending on whether the problem is viewed as a regression or classification problem, the labels l_i are chosen differently.

BWE is typically treated as a regression problem and the mapping function is estimated under the MMSE criterion as follows [4, 5, 13]

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - f_\theta(\mathbf{x}_i)\|_2^2 \quad (3)$$

where the labels $l_i = \mathbf{y}_i$, which is the ground truth WB counterpart of the NB speech features. This requires parallel WB and NB data and is usually accomplished by downsampling the WB speech to create the feature pairs. The mapping functions are learned in a reconstructive way to minimize the L_2 distance between the mapped NB features and their WB counterparts.

Since ASR is a classification problem by nature, it is desirable to have the BWE mapping estimated with a matched objective. In this paper, we choose another way of estimating the BWE mapping function. Suppose we have a WB neural network acoustic model Φ which takes the WB speech features as input and outputs the posterior probabilities p_{ik} with respect to context-dependent phone classes after the softmax layer for feature i and class k . We use the BWE mapping function g_θ to map the upsampled NB speech features to WB features which are directly fed into the WB acoustic model Φ to generate posteriors $\mathbf{p}_i = \Phi(g_\theta(\mathbf{x}_i))$ where Φ is fixed and g_θ is subject to optimization. In this case, the mapping function is $f_\theta \triangleq \Phi \circ g_\theta$. The CE loss function is defined between the posterior probabilities \mathbf{p}_i and the class labels l_i :

$$\theta^* = \arg \min_{\theta} \frac{1}{n} \sum_i \mathcal{L}(l_i, \Phi(g_\theta(\mathbf{x}_i))) \quad (4)$$

$$= \arg \min_{\theta} \frac{1}{n} \sum_{i,k} l_{ik} \log \frac{1}{p_{ik}} \quad (5)$$

The labels l_i are generated by aligning the upsampled NB features against the WB acoustic model. The mapping function involves a composite of two CNNs. One is used to map the upsampled NB speech to WB which is subject to optimized and the other is an existing WB acoustic model which is fixed. This is illustrated in Fig 1.

In our pilot simulation experiments on 50-hours Broadcast News (BN50), we found that a DNN-based MMSE BWE did not help ASR performance when using a WB CNN model to decode the upsampled NB features mapped by it, which is worse than the NB baseline. Moreover, MMSE BWE relies on parallel data. Although this can be done by downsampling WB speech, it is an artificial setup. In real world such parallel data is difficult to collect.

4. System Implementation

4.1. Feature Space

The WB speech is sampled at 16KHz while the NB speech at 8KHz. The input feature space consists of 40-dimensional logmel features after application of first a global cepstral mean normalization (CMN) and then an utterance-based CMN. There are three input feature maps to the CNNs, the static logmel features and their delta and double delta, all with a temporal context of 11 frames. For upsampled NB speech signals, they go through the WB Mel filter banks after upsampling in the time domain.

4.2. Models

Acoustic models CNN acoustic models are used for WB baseline, NB baseline and MB models, which have the same configuration. There are 2 convolutional layers and each convolutional layer is followed by a max-pooling layer. The first convolutional layer uses 5×5 kernels with a stride is 1×1 and padding 2×2 . The second convolutional layer uses the same kernel, stride and padding sizes as those of the first convolutional layer. Both max-pooling layers use a kernel of 2×2 and stride of 2×2 . On top of the convolutional and pooling layers are 3 FC layers with 1,024 hidden units. All activation functions are Relu except the last FC layer which uses sigmoid. The output softmax layer has 9,300 output units. We investigate two model capacities in the experiments, one with 128 and 256 feature maps for the two respective convolutional layers and the other 256 and 512 feature maps.

BWE Models The BWE mapping network is also a CNN. The design of the network follows the VGG architecture that uses small convolutional kernels, small stride and small pooling kernels but, in the meantime, uses increased depth of the convolutional layers and reduced max-pooling layers. Specifically, we use 4 convolutional layers, 2 max-pooling layers and 3 FC layers. Every 2 convolutional layers are followed by one max-pooling layer. The first 2 convolutional layers use 3×3 kernels with a stride 1×1 and padding 1×1 . The second 2 convolutional layers again use 3×3 kernels with a stride 1×1 and padding 1×1 . The 2 max-pooling layers use 2×2 kernels with a stride 1×1 . The 3 FC layers have 1,024 hidden units. All activation functions are Relu except the last FC layer which uses tanh. We also investigate two model capacities, one with 64 feature maps for the two convolutional layers and 128 feature maps for the next two convolutional layers, and the other 128 and 256 features maps. They are indicated when reporting the results.

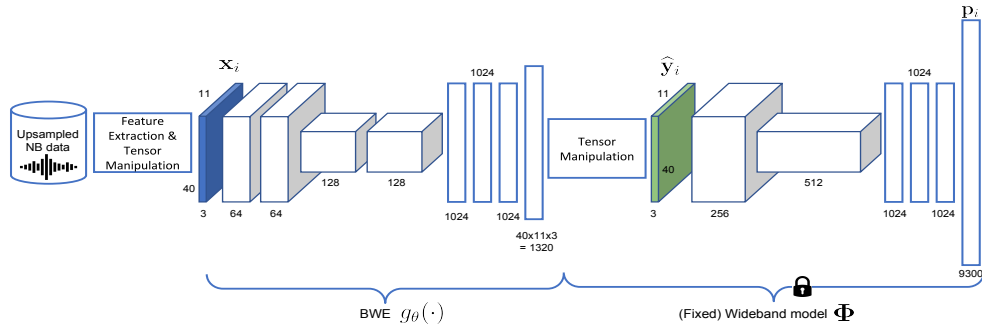


Figure 1: Illustration of the training of the BWE mapping. The mapping is realized as a CNN with a VGG architecture. Its output is connected to the WB CNN acoustic model after tensor manipulation. The WB CNN is fixed and the BWE CNN is optimized under the CE criterion.

4.3. Distributed Training

The networks are optimized under the CE criterion using the SGD algorithm. Learning rate starts as 0.01 for 10 epochs and is annealed by half every epoch for the next 10. We apply synchronous data parallelism on multi-GPUs (8 Nvidia v100s), within the same server, to accelerate training [14]. To minimize parameters/gradients copy, we remap each trainable layer’s gradient buffer to one consecutive region upfront. For each iteration, after each learner finishes backward propagation, we use NCCL[15] to sum all learners’ gradients, via an Allreduce call, back to each learners’ gradients region, before weights update. By doing this, 8-GPU training achieves almost 8x speedup. In our experiments, each GPU receives a mini-batch of size 512, bringing the total batch size per iteration to 4096. Another benefit of using multi-GPU is each learner effectively pre-load training data for others, which reduces I/O time to negligible.

5. Experimental Results

There are 1,150 hours of WB training data which consists of 420 hours of Broadcast News data, 450 hours of internal dictation data, 100 hours of meeting data, 140 hours of hospitality (travel and hotel reservation) data and 40 hours of accented data. There are 2,300 hours of NB training data which consists of 2,000 hours of Switchboard data and 300 hours of IBM call center data. In practice, we often find that the amounts of available WB and NB training data are unbalanced. In our case, the amounts of NB data is larger than that of the WB data.

We choose 4 WB test sets and 4 NB test sets for our experiments whose description and statistics are given in Table 1. The decoding vocabulary comprises of 250K words and the language model (LM) is a 4-gram LM with modified Kneser-Ney Smoothing consisting of 200M n-grams. The LM training data is selected from a broad variety of sources. Stronger language modeling techniques such as RNN LMs will be investigated in the future. The word error rates (WERs) of various models on the 8 test sets are shown in Table 2.

Baselines The WERs of WB and NB CNN baselines are shown in the first two rows in Table 2. The numbers of feature maps in the convolutional layers are indicated in the parentheses (e.g. [128,256] vs. [256, 512]). The underlined WERs indicate a change of sampling rate of the test data in order to be decoded by the acoustic model. For instance, the NB test sets are upsampled to decode against the WB CNN and vice versa. Obviously, without any compensation, mismatched test data and acoustic model give rise to significant degradation of the performance.

Table 1: WB and NB datasets used for evaluation.

		Description	Hours
WB	WS1	Dev04f test set from Broadcast News	2.21
	WS2	Commercial services help desk	0.34
	WS3	Hospitality domain 1	1.21
	WS4	Hospitality domain 2	0.81
NB	NS1	Hub5-2000 test set from Switchboard	2.10
	NS2	Technical support	4.09
	NS3	Commercial services help desk	3.01
	NS4	Multi-domain command and control	12.78

(17.2% \rightarrow 23.6% for WB test sets and 17.8% \rightarrow 25.0% for NB test sets on average.) We also carry out an experiment (third row) where only the WB training data is downsampled to train a CNN acoustic model which is used to decode the NB test sets. This model also gives a 25% average WER which is significantly worse than the matched training with NB data only. The performance gap may be due to the mismatched data but also to the mismatched domains.

Direct Mixing The second block of Table 2 presents the performance of MB models trained using the direct mixing strategy where WB data and upsampled NB data are mixed for the training of a CNN acoustic model. The CNN model with [128,256] feature maps obtains about the same average WER as the NB CNN baseline (17.8%) but slightly worse average WER than the WB CNN baseline (17.6%). Since the amount of training data increases after mixing, it is reasonable to increase the capacity of the MB model. With doubled feature maps in the convolutional layers ([256,512]) the MB CNN has a lower average WER (17.4%) on the NB test sets and only 0.3% absolute worse on the WB test sets. On the other hand, however, if the MB model is trained using NB data and downsampled WB data, its performance is far inferior on both WB and NB test sets. Therefore, from the table we can tell that upsampling the NB data and then mixing with the WB data appears to be a better strategy for MB modeling.

BWE The third block of Table 2 presents the performance of the proposed BWE approach. The VGG architecture of the BWE network is discriminatively trained with respect the WB CNN baseline model. With 64 feature maps in the first two convolutional layers and 128 feature maps in the second two convolutional layers, the BWE can significantly improve the average WER from 25.0% to 18.9%. If increase the network capacity with doubled feature maps, the average WER can be further improved to 18.6%. The last row of this block shows the performance of BWE trained in a denoising manner (denoted nBWE)

Table 2: Word error rates (WERs) of WB, NB and MB models on 8 test sets. The WERs are reported in 5 blocks from top to bottom representing various experimental conditions.

	WB					NB				
	WS1	WS2	WS3	WS4	Avg	NS1	NS2	NS3	NS4	Avg
WB baseline ([128,256])	15.4	14.9	9.1	29.2	17.2	25.1	39.0	13.7	22.0	25.0
NB baseline ([128,256])	<u>21.3</u>	<u>16.8</u>	<u>15.6</u>	<u>40.5</u>	23.6	13.5	25.0	12.8	19.7	17.8
WB↓ ([128,256])	<u>17.9</u>	<u>17.5</u>	<u>10.9</u>	<u>33.9</u>	20.1	26.0	39.0	12.9	21.9	25.0
DirectMix (WB+NB↑,[128,256])	17.1	13.0	12.2	27.9	17.6	13.8	25.5	12.2	19.6	17.8
DirectMix (WB+NB↑,[256,512])	16.5	12.8	11.8	28.8	17.5	13.4	25.2	11.8	19.2	17.4
DirectMix (WB↓+NB,[128,256])	18.9	17.2	13.3	35.9	21.3	14.0	26.2	12.5	19.1	18.0
BWE ([64,128])	-	-	-	-	-	15.2	27.8	12.4	20.2	18.9
BWE ([128,256])	-	-	-	-	-	14.9	27.4	12.2	20.0	18.6
nBWE ([64,128])	-	-	-	-	-	15.0	27.6	12.4	19.6	18.7
Mix (WB+NB↑+BWE, [128,256])	16.5	14.2	10.1	29.9	17.7	13.6	25.6	12.2	19.7	17.8
Mix (WB+NB↑+BWE, [256,512])	16.0	14.6	9.7	29.9	17.6	13.7	25.4	12.2	19.6	17.7
Mix (WB+NB↑+nBWE, [128,256])	16.4	14.3	10.0	30.9	17.9	13.7	25.6	12.1	19.5	17.7
MixFT (WB+NB↑+BWE, [128,256])	16.6	14.8	9.9	29.2	17.6	13.6	25.6	12.5	19.7	17.9
MixFT (WB+NB↑+BWE, [256,512])	16.1	15.1	9.7	29.3	17.6	13.7	25.5	12.4	19.6	17.8
MixFT (WB+NB↑+nBWE, [128,256])	16.2	14.4	9.8	30.3	17.7	13.6	25.4	12.0	19.6	17.7

where zero-mean Gaussian noise with variance of 0.01 is added to the input upsampled NB logmel features. It indicates that the denoising BWE can improve the generalization of the mapping and gives better performance under the same model capacity (18.9% \rightarrow 18.7%). Note that the BWE achieves improvement across all the four NB test sets against the WB CNN model compared to simple upsampling. In some test sets, BWE also yields better performance than the NB baseline. In the following experiments, we will stick to the BWE CNN configuration with the [64, 128] feature maps.

Mixing with BWE The fourth block of Table 2 shows the performance of the mixing strategy of using WB data and BWE-mapped upsampled NB data from which the MB CNN models are trained. As shown by the table, the MB models further improve the WERs from the BWE alone. Using larger model capacity helps. Overall, it is slightly better than the direct mixing strategy when the model capacity is [128, 256] on the NB test sets. Denoising BWE helps the NB test sets but hurts the WB test sets.

Fine-tune The last block of Table 2 shows the WERs after fine-tuning the mixing with BWE. In the fine-tuning, the output of the BWE CNN is connected to the input of the MB CNN which is fixed. The BWE CNN is fine-tuned with a smaller learning rate for 6 epochs. After that, another MB CNN is trained using the fine-tuned BWE with a smaller learning rate (1/10 of the original learning rate) for another 6 epochs. The improvement given by this finetuning, as can be observed from the table, is only marginal.

6. Discussion and Summary

As can be observed from the breakdown performance in Table 2, consistent improvements of one technique across all test sets are rarely observed. The conclusion drawn from one particular test set by one technique may not generalize to other test sets, although the average WERs can give us a good idea on the overall performance of certain technique. That is the reason we believe it is important to evaluate the BWE and mixed strategies extensively on diverse test sets from various domains and conditions.

In ASR applications, it is desirable to use one unified acoustic model for both WB and NB speech data. The experimental results in Sec.5 show that it is possible to train a MB model

with an appropriate strategy. Upsampling the NB data appears to be more helpful than downsampling the WB data. In addition, direct mixing with appropriately increased model capacity, due to increased training data after mixing, can give competitive ASR performance compared to separate WB and NB models individually. In our investigated case, the best MB model yields lower average WER than the NB baseline and only slight degradation over the WB baseline. Although direct mixing assumes no explicit BWE, one would expect the DNNs will implicitly learn the mapping from the zero-padded upper frequency bins of NB speech to WB speech.

In our pilot experiments, MMSE-based BWE turned out not to be very helpful. Compared to the MMSE-based BWE, the proposed discriminatively trained BWE can significantly help when mapping NB speech to WB speech and decoded against WB models. Therefore, it is a competitive mapping function. However, when mixing WB speech and BWE-mapped NB speech for MB training, using BWE is not consistently better than direct mixing with increased model capacity. In addition, when considering BWE-based MB modeling, the BWE network should be treated as part of the MB model. Hence, its performance should be compared with direct mixing using networks with equivalent model capacity.

In summary, we have investigated in this paper the large-scale MB acoustic modeling with deep architectures for ASR. Extensive experiments were carried out to evaluate a variety of mixing strategies, including downsampling, upsampling and BWE, on diverse WB and NB test sets from various domains. Looking forward, with the success of deep generative models [16] and its applications to BWE [17], we hope to further improve BWE in the context of MB modeling for consistent superior performance.

7. References

- [1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." in *International Conference on Learning Representations (ICLR)*, 2015.
- [2] M. L. Seltzer and A. Acero, "Training wideband acoustic models using mixed-bandwidth training data for speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 15, no. 1, pp. 235–245, 2007.
- [3] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-

- DNN-HMM,” in *IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 131–136.
- [4] J. Gao, J. Du, C. Kong, H. Lu, E. Chen, and C.-H. Lee, “An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition,” in *International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 588–594.
- [5] J. Gao, J. Du, and E. Chen, “Mixed-bandwidth cross-channel speech recognition via joint optimization of DNN-based bandwidth expansion and acoustic modeling,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 559–571, March 2019.
- [6] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [7] B. Iser, W. Minker, and G. Schmidt, *Bandwidth extension of speech signals*. Springer, 2008.
- [8] N. Prasad and T. Kumar, “Bandwidth extension of speech signals: a comprehensive review,” *International Journal of Intelligent Systems Technologies and Applications*, vol. 2, no. 2, pp. 45–52, 2016.
- [9] F. Nagel and S. Disch, “A harmonic bandwidth extension method for audio codecs,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 145–148.
- [10] H. Pulakka and P. Alku, “Bandwidth extension of telephone speech using a neural network and a filter bank implementation for high-band Mel spectrum,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2170–2183, 2011.
- [11] C. Liu, Q.-J. Fu, and S. Narayanan, “Effect of bandwidth extension to telephone speech recognition in cochlear implant users,” *The Journal of the Acoustical Society of America*, vol. 26, no. 5, pp. 77–83, 2018.
- [12] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, “Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 125, no. 2, pp. 883–894, 2009.
- [13] 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 *Interspeech*, 2015, pp. 2578–2582.
- [14] W. Zhang, M. Feng, Y. Zheng, Y. Ren, Y. Wang, J. Liu, P. Liu, B. Xiang, L. Zhang, B. Zhou, and F. Wang, “Gadei: On scale-up training as a service for deep learning,” in *The IEEE International Conference on Data Mining (ICDM)*, 2017.
- [15] “Nvidia collective communications library (NCCL),” <https://developer.nvidia.com/nccl>, accessed: 2018-10-26.
- [16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [17] D. Haws and X. Cui, “CycleGAN bandwidth extension acoustic modeling for automatic speech recognition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6780–6784.