An Experimental Study on Joint Modeling of Mixed-Bandwidth Data via Deep Neural Networks for Robust Speech Recognition

Jianqing Gao*, Jun Du*, Changqing Kong†, Huaifang Lu†, Enhong Chen*, Chin-Hui Lee‡
*National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, Anhui, P. R. China
Email: jqgao@iflytek.com, jundu@ustc.edu.cn, cheneh@ustc.edu.cn
†iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, P. R. China
Email: cqkong@iflytek.com, hflu@iflytek.com
‡Georgia Institute of Technology, Atlanta, Georgia, USA
Email: chl@ece.gatech.edu

***Abstract*—We propose joint modeling strategies leveraging upon large-scale mixed-band training speech for recognition of both narrowband and wideband data based on deep neural networks (DNNs). We utilize conventional down-sampling and up-sampling schemes to go between narrowband and wideband data. We also explore DNN-based speech bandwidth expansion (BWE) to map some acoustic features from narrowband to wideband speech. By arranging narrowband and wideband features at the input or the output level of BWE-DNN, and combining down-sampling and up-sampling data, different DNNs can be established. Our experiments on a Mandarin speech recognition task show that the hybrid DNNs for joint modeling of mixed-band speech yield significant performance gains over both the narrowband and wideband speech models, well-trained separately, with a relative character error rate reduction of 7.9% and 3.9% on narrowband and wideband data, respectively. Furthermore, the proposed strategies also consistently outperform other conventional DNN-based methods.**

I. Introduction

In the current mobile internet era, building a universal automatic speech recognition (ASR) system leveraging upon large-scale diversified data is a popular topic, especially with the deep learning technologies. For example, with the emergence of many speech enabled applications, we could collect a huge amount of speech data with different bandwidths, namely narrowband speech (sampled at 8kHz) over the telephone channel and wideband speech (sampled at 16kHz or 44kHz) from other recording equipments. One straightforward method is to train bandwidth-dependent models separately. In this study, we aim to construct a unified model for mixed-band speech recognition with large-scale mixed-band speech data for training hybrid acoustic models via hidden Markov model (HMM).

Obviously, the simplest way is to down-sample the wideband speech and combine them with the original narrowband speech to train a narrowband model. It can boost the recognition performance of narrowband speech with the additional down-sampled data while sacrificing the performance of wideband speech as the high-frequency information is lost [1] during down-sampling. An up-sampling of narrowband speech to be combined with the original wideband speech is another alternative. Recently, speech bandwidth expansion (BWE), i.e., mapping the low-frequency spectra to high-frequency spectra, has been proposed. Statistical models [2], such as Gaussian mixture model (GMM) [3], is flexible in modeling the speech signal of both low-frequency and high-frequency bands, and estimating the high-frequency spectra with a soft decision scheme [4], [5], [6], [7], [8], [9]. In contrast, deep neural network (DNN) based approaches [2], [10], [11], have been proposed to directly learn the highly non-linear relationship between the low frequency and the high-frequency spectra, and shown to improve the ASR accuracy with high-quality simulation data [12]. However, BWE, aiming at improving the objective listening quality of narrowband speech, has not been shown to improve the recognition performance for mixed-band speech using those existing BWE techniques.

In the past few years, several approaches have been proposed to train mixed-band acoustic models. In [13], an EM algorithm is introduced to train mixed-band models for the GMM-HMM based system. However, no significant gains could be observed. Compared with the implementation in GMM-HMM based system with a complicated inference, it is easier to formulate modeling as a missing feature problem in DNN-HMM based systems, where several feature dimensions corresponding to high-frequency bands have no value and to be estimated when narrowband speech is presented [14]. Another DNN-based approach is to treat mixed-band modeling as a domain adaptation problem [15]. The DNN trained on the rich narrowband speech data can be adapted effectively to the target wideband domain with a small amount of wideband speech data.

In this study, we design a novel architecture for DNN-based mixed-band acoustic modeling. We first train a regression DNN for BWE (DNN-BWE) by mapping the log-Mel filterbank (LMFB) features from narrowband to wideband speech. Then we present several joint training strategies between

978-1-5090-0620-5/16/$31.00

DNN-BWE and the classification DNN used for acoustic modeling (DNN-AM) with a large-scale training set containing both narrowband and wideband speech. In comparison to the related work in [14], [15], the main contributions of our work can be summarized as follows: (i) our joint training approaches are conducted by explicitly learning the relationship between narrowband and wideband speech while the new architectures of the input layer are designed to accommodate both narrowband and wideband speech for the following implicit parameter learning of DNN in [14], [15]; (ii) an augmented LMFB feature vector with a higher dimension is adopted in [14] which implies that more information is used as the input of DNN while our approach uses a low-dimensional LMFB feature vector for both narrowband and wideband speech, which can be also implemented on top of the augmented features; (iii) the method in [15] seems like an adaptation scheme to wideband speech, not the joint modeling of mixed-band speech as in our approach, which can not guarantee the good performance for the narrowband speech and achieve significant gains when there is a large amount of wideband speech data; and (iv) our task is more challenging as we use equally 1000 hours of real collected narrowband and wideband speech, rather than the unbalanced data distribution of mixed-band speech and relatively smaller amount of training data in [14], [15]. Moreover, our narrowband data and wideband data varies a lot in both talking style and channel degradation, which usually leads to difficulties for joint modeling with both. Finally, the effectiveness of our proposed approach is demonstrated by yielding remarkable performance gains over the well-trained baseline systems and other conventional approaches.

The rest of the paper is organized as follows. First we briefly introduce the DNN-based BWE method in Section II. Then we describe the joint training strategy of DNN-BWE and DNN-AM for narrowband speech recognition in Section III. In Section IV, two joint training strategies for mixed-band speech recognition are presented. We report the experimental results in Section V and conclude the paper in Section VI.

II. DNN-BASED SPEECH BANDWIDTH EXPANSION

Recently, DNN has been used as a regression model to enhance noisy speech [16] followed by robust ASR [17], [18], [19]. In these works, DNN is adopted as a pre-processor to denoise the original noisy speech. During training, noisy speech is used to predict the corresponding clean speech and the regression DNN is used to model the highly non-linear denoising function mapping from noisy speech to clean speech. In this paper, a regression DNN is adopted to map the LMFB acoustic features from narrowband to wideband. This is quite different from the approach in [11], [12] where the spectra of low-frequency bands are used to predict the spectra of high-frequency bands. The approach proposed in [11], [12] is quite effective for improving the perceptual quality of narrowband speech but not always effective for improving the ASR performance, which might be due to the inconvenience of joint training for ASR. Aiming at improving the recognition performance, we propose a feature mapping architecture for bandwidth expansion which directly manipulates LMFB acoustic features for ASR.

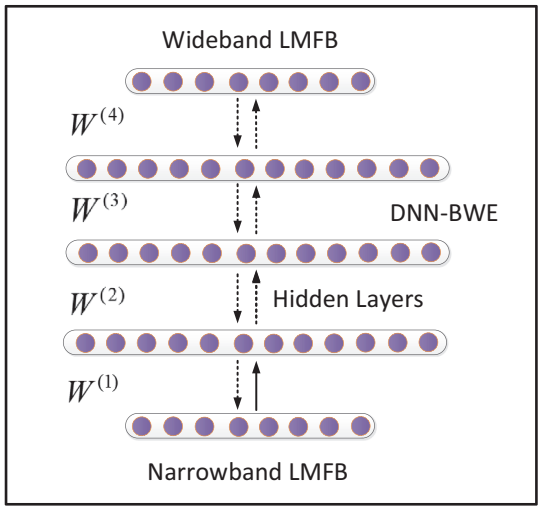

Fig. 1. DNN-based feature mapping for BWE

As illustrated in Fig. 1, our proposed DNN-BWE contains three sigmoidal hidden layers with 2048 nodes for each. The input layer consists of 11-frame narrowband LMFB features while the corresponding 11-frame wideband LMFB features are used for the output layer. For the training of this regression DNN, we need time-aligned narrowband and wideband speech data pairs, which are quite difficult to obtain in practice. To address this problem, we down-sample the original wideband speech to narrowband to obtain the corresponding data pair. Then we directly map the narrowband LMFB features to wideband LMFB features with the same dimension, which means our proposed bandwidth expansion strategy is formulated as a feature mapping problem instead of the missing feature problem in [11], [12]. In the training procedure, we aim at minimizing the following loss function with the asynchronous stochastic gradient descent (ASGD) algorithm [20] in a mini-batch mode:

$$E = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_{n-\tau}^{n+\tau}\|_2^2 + \kappa\|\boldsymbol{W}\|_2^2 \quad (1)$$

where $\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}$ and $\boldsymbol{x}_{n-\tau}^{n+\tau}$ are the n-th $D(2\tau+1)$-dimensional vectors of estimated and reference wideband features, respectively. $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$-dimensional vector of input narrowband features with neighbouring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. κ is the regularization weighting coefficient to avoid over-fitting.

During the bandwidth expansion stage, the narrowband LMFB features are fed into the BWE network to generate the estimated wideband LMFB features which are used to train the DNN-AM under cross entropy (CE) criterion.

III. Joint training strategy for narrowband speech recognition

To validate the effectiveness of the proposed DNN-BWE, we first introduce a joint training stretegy of DNN-BWE and DNN-AM for narrowband speech prior to joint modeling of mixed-band speech. As shown in Fig. 2, the framework of our proposed joint training strategy for narrowband speech recognition (denoted as JT-1 for convenience) consists two sub-models, namely "DNN-BWE" and "DNN-AM". "DNN-BWE" is the bandwidth expansion DNN as described in Section II, while "DNN-AM" stands for the regular DNN used for acoustic modeling in ASR. Before the discussion of this joint training strategy, we first define the following notations to be utilized in the rest of this article. Thus we have four datasets in total:

1) **Narrowband_Ori**:
   Original narrowband telephony speech (at 8KHz).
2) **Wideband_Ori**:
   Original wideband speech (at 16KHz).
3) **Narrowband_DS**:
   Narrowband speech (at 8KHz) down-sampled from the original wideband speech.
4) **Wideband_US**:
   Wideband speech (at 16KHz) up-sampled from the original narrowband speech.

As for feature extraction, we adopt 24-dimensional LMFB features with its first and second order derivatives for both narrowband speech and wideband speech where each filter bank carries the information of different frequency band which is different from [14], [15].

Algorithm 1 : **Training procedure of strategy JT-1**

Step1: **DNN-BWE training**

1) Train DNN-BWE with Narrowband_DS LMFB features and Wideband_Ori LMFB features under MMSE criterion with a random initialization.

Step2: **DNN-AM training**

1) Feed Narrowband_Ori LMFB features through DNN-BWE to generate BWE features.
2) Train DNN-AM with BWE features under CE criterion and regular Restricted Boltzmann Machine (RBM) based pre-training [21] is used for the initialization of DNN-AM.

Step3: **Joint training**

1) Concatenate DNN-BWE and DNN-AM as illustrated in Fig. 2.
2) Optimize the hybrid model of DNN-BWE and DNN-AM under CE criterion with Narrowband_Ori LMFB features.

The training procedure is elaborated in Algorithm 1 explicitly. As the output layer of DNN-BWE is well-matched with the input layer of DNN-AM, it is straightforward to concatenate the both. From Algorithm 1 we can see that the training datasets of DNN-BWE and DNN-AM are different, which means that the proposed bandwith expansion network trained from the simulated narrowband LMFB features and the original wideband LMFB features performs well on the mismatched but real narrowband telephony speech. After merging DNN-BWE and DNN-AM, we need to further update DNN-BWE jointly with DNN-AM under CE criterion since the original DNN-BWE is optimized under minimum mean square error (MMSE) criterion which might not be optimal in terms of recognition accuracy. Furthermore, the mismatch between the training data of DNN-AM and DNN-BWE is another underlying challenge. From this aspect, joint optimizing of DNN-BWE and DNN-AM is quite an effective and necessary strategy for acoustic modeling. Please note that the concatenation layer is a linear layer since we aim at generating estimated wideband LMFB features with this layer.

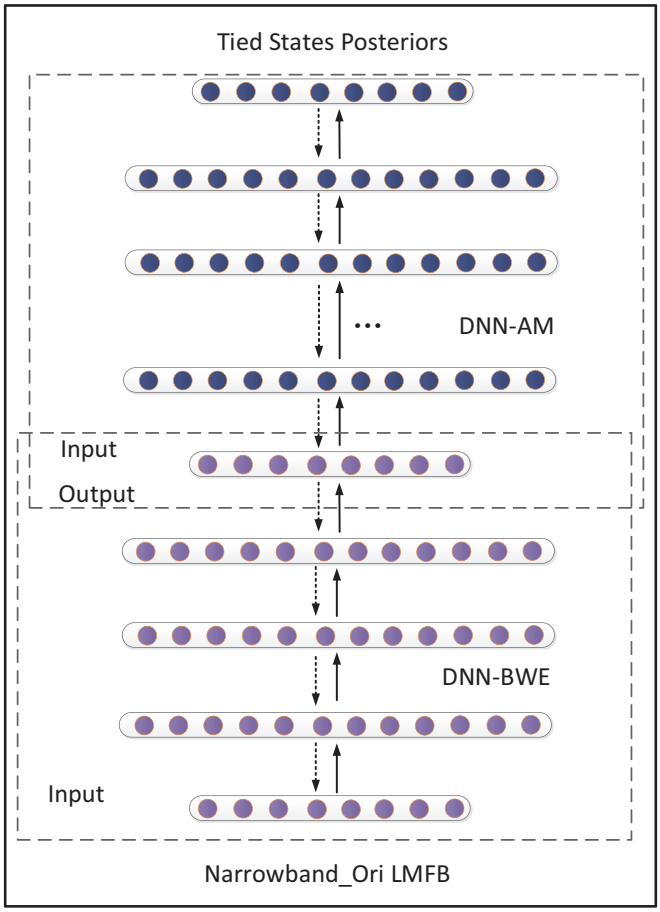

Fig. 2. Joint training strategy for narrowband speech recognition (JT-1)

IV. Joint modeling strategies for mixed-band speech recognition

In the Section III, we present a joint training strategy for narrowband speech recognition, aiming at improving the perfor-

mance of narrowband speech recognition system. Narrowband speech could be relatively easy to be collected over telephone channel while the wideband speech was more difficult in the past. With the developing of mobile internet, it becomes more and more efficient and convenient for us to collect wideband speech. So it is very important to investigate feasible modeling strategies for mixed-band speech recognition. In the following subsections, several joint modeling strategies for both DNN-BWE and DNN-AM are elaborated.

A. Same entry for mixed-band speech recognition (JT-2)

One conventional method for mixed-band speech recognition is to combine the original narrowband speech Narrowband_Ori and Narrowband_DS to train a universal narrowband DNN-AM. At the recognition stage, both Narrowband_Ori and Narrowband_DS are fed to the DNN-AM. This combination strategy might improve the recognition accuracy of narrowband speech with more training data. However, it usually leads to a performance degradation for the original wideband speech since the information in the high-frequency bands is lost after down-sampling. We therefore propose an architecture for joint modeling of both DNN-BWE and DNN-AM as demonstrated in Fig. 3, where DNN-BWE plays the role of recovering some information of high-frequency bands for both Narrowband_Ori and Narrowband_DS speech. By using this architecture, all the mixed-band speech share the same entry to DNN-BWE. The training procedures are as follows. First, the DNN-BWE is built with the Wideband_Ori and Narrowband_DS LMFB feature pair. Second, DNN-AM is initialized by RBM based pre-training, which is followed by a fine-tuning scheme with the output features of DNN-BWE using both the Narrowband_Ori and Narrowband_DS LMFB features. Finally, by concatenating DNN-BWE and DNN-AM, joint training with the CE criterion using both the Narrowband_Ori and Narrowband_DS LMFB features is performed for updating all the parameters of the hybrid DNN.

Algorithm 2 : **Training procedure of strategy JT-2**

Step1: **DNN-BWE training**

1) Train DNN-BWE with Narrowband_DS LMFB features and Wideband_Ori LMFB features under MMSE criterion as described in Algorithm 1.

Step2: **DNN-AM training**

1) Mix Narrowband_Ori and Narrowband_DS LMFB features randomly in frame-level.
2) Feed mixed narrowband LMFB features into DNN-BWE, the output features are used to train DNN-AM.

Step3: **Joint modeling**

1) Concatenate DNN-BWE and DNN-AM.
2) Jointly optimize DNN-BWE and DNN-AM under CE criterion with mixed narrowband LMFB features.

Note that DNN-BWE and DNN-AM can be seamlessly concatenated because the output of DNN-BWE as a linear layer is exactly the same as the input of DNN-AM. It usually

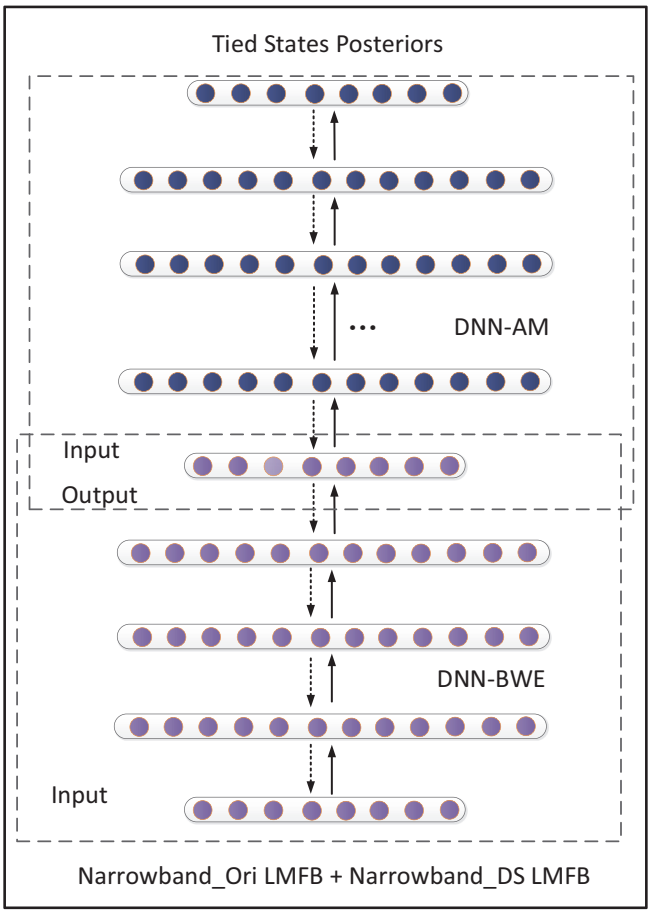

Fig. 3. Same entry for mixed-band speech recognition (Strategy JT-2)

yields a better recognition accuracy by re-optimizing DNN-BWE via the classification criterion. This strategy is referred as JT-2. To better understand this training strategy, we further illustrate the training procedure as in Algorithm 2. It is quite a novel strategy to utilize mixed-band speech compared with traditional ones. First, the narrowband speech down-sampled from wideband boosts the performance of the original narrowband system. Second, DNN-BWE helps recover some wideband information which can improve the wideband system performance. Furthermore, DNN-BWE can also deal with some channel distortion and eliminate the mismatch between windeband and narrowband speech. Wideband speech needs to be down-sampled to narrowband at the recognition stage.

B. Different entries for mixed-band speech recognition (JT-3)

Though strategy JT-2 mentioned above is quite an effective strategy for mixed-band modeling, it is worth to indicate that this strategy usually tends to boost the performance on narrowband speech more significant than wideband speech. It is clear that directly up-sampling from narrowband to wideband can rarely boost performance on wideband speech

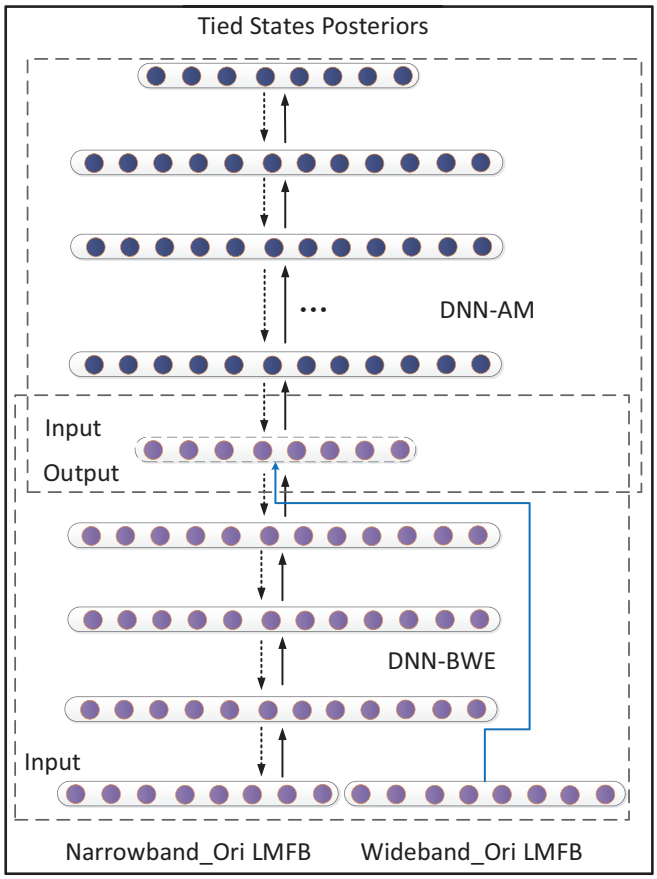

Fig. 4. Different entries for mixed-band speech recognition (Strategy JT-3)

due to the lack of high frequency information of the original narrowband speech and often suffers from the performance degradation on the original narrowband speech. To address this issue, we use the architecture in Fig. 4 and adopt the DNN-BWE rather than the simple up-sampling. Different from joint modeling strategy JT-2 described in Fig. 3, the mixed-band LMFB features are fed to different entries, wideband LMFB features (Wideband_Ori) for DNN-AM and narrowband LMFB features (Narrowband_Ori) for DNN-BWE. This architecture tends to improve the recognition accuracy for Wideband_Ori as all the high frequency information of the original wideband speech can be used at the training stage. From Fig. 4 we can see that the wideband LMFB features are directly fed into DNN-AM for classification while the narrowband features are fed into DNN-BWE instead. In [14], narrowband features are directly padded with zeros which means no extra information is involved. In this subsection, the proposed joint training strategy is more informative as we can derive the estimated wideband features through bandwidth expansion network instead of directly padding zeros.

Algorithm 3 : Training procedure of strategy JT-3

Step1: **DNN-BWE training**

1) Train DNN-BWE with Narrowband_DS LMFB features and Wideband_Ori LMFB features under MMSE criterion just as described in Algorithm 1 and Algorithm 2.

Step2: **DNN-AM training**

1) Mix Narrowband_Ori and Wideband_Ori randomly in mini-batch level.
2) Concatenate DNN-BWE and DNN-AM as illustrated in Fig. 4.
3) Feed Narrowband_Ori LMFB features into DNN-BWE and wideband_Ori LMFB features into DNN-AM, seperately, and update DNN-AM with CE criterion while fixing DNN-BWE.

Step3: **Joint modeling**

1) Jointly optimize DNN-BWE and DNN-AM as a whole part under CE criterion, using Narrowband_Ori LMFB features to update both DNN-BWE and DNN-AM, while using Wideband_Ori LMFB features to update DNN-AM only.

Step4: **Fine-tuning for narrowband speech**

1) Further optimize DNN-BWE with the Narrowband_Ori LMFB features under CE criterion while fixing DNN-AM.

The training procedure, which is elaborated in Algorithm 3, is quite different from strategy JT-2. In this architecture, narrowband LMFB features and wideband LMFB features are mixed randomly in mini-batch level, which means one mini-batch only consists of wideband LMFB features or narrowband LMFB features since the different input entry for each. This is another difference lies between JT-2 and JT-3, in JT-2, both wideband LMFB features and narrowband LMFB features can be included in the same mini-batch. In the first stage during training, only DNN-AM is updated, namely the error signals calculated on the output layer are back-propagated to the input layer of DNN-AM only. Then, the calculated errors of the output layer for wideband features are back-propagated to the input layer of DNN-AM while errors for narrowband features are back-propageted through up to the input layer of DNN-BWE to update all the parameters of the whole net. Finally, we can further update DNN-BWE only with the narrowband features as it is irrelevant to the recognition performance of wideband speech. At the recognition stage, wideband features and narrowband features are also fed into the joint trained model through different entries to generate tied states posteriors.

V. Experiments and Result Analysis

To verify the effectiveness of the proposed approaches, we conducted a series of experiments on a Mandarin speech recognition task with 1000-hour narrowband (Narrowband_Ori, 8kHz) telephony speech and 1000-hour wideband (Wideband_Ori, 16kHz) conversational speech for training.

Narrowband_Ori were collected from real life telephony conversations, suffering from channel distortion and noise disturbation. Directly mixing these multi-source mismatched speech often led to unsatisfactory system performance. For evaluation, we had one wideband dataset and one narrowband dataset, each contains about 10 hours speech. To generate the training targets for DNN-AM, a GMM-HMM system with 9004 tied states and 40 Gaussian mixtures for each state was trained with Mel-frequency cepstral coefficients (MFCCs), All the DNN-HMM systems trained under CE criterion shared the same targets derived from this model. The architecture of DNN-AM was 825-2048*6-9004, namely 11 frames of 72-dimensional LMFB features and 3-dimensional pitch features for the input layer, 2048 units for each hidden layer, and 9004 tied states for the output layer. Unsupervised RBM pre-training was used for DNN initialization.

A. Experiments and results for strategy JT-1

We firstly evaluate the effectiveness of joint training strategy for narrowband speech recognition described in Section III. Table I lists the character error rate (CER) comparision of the baseline system and strategy JT-1. "Baseline" indicates the regular DNN-AM with 6 sigmoidal hidden layers and training set is the original narrowband speech (Narrowband_Ori). The number in the brackets is the relative CER reduction over the corresponding baseline system.

From Table I we can see that our proposed joint training strategy outperforms baseline system with a 3.3% relative error reduction. Narrowband_Ori which is used to train this joint training system encounters severe channel mismatch with the training data of DNN-BWE. However, our proposed narrowband speech recognition system yields considerable performance gain over traditional baseline system. We can draw the conclusion in this experiment that strategy JT-1 can be transfered to irrelevant training data as well since we can weaken this mismatch between training data of DNN-BWE and training data of DNN-AM through joint optimizing of DNN-BWE and DNN-AM.

TABLE I
Performance (CER in %) comparison of the baseline system and strategy JT-1 on original narrowband speech.

CER(%)	Narrowband
Baseline	30.72
JT-1	29.69(3.3%)

B. Experiments and results for strategy JT-2

Table II lists the CER comparison of several mixed-band systems and baseline systems on both narrowband and wideband evaluation sets. "Baseline" stands for the narrowband and wideband models built with the corresponding training speech independently resulting in two separate models. "DS" represents the system built by randomly mixing Narrowband_Ori and Narrowband_DS, which is a conventional way of mixed-band training. The final row of Table II is our proposed joint training strategy for same entry.

TABLE II
Performance (CER in %) comparison of several mixed-band systems and baseline system on both narrowband and wideband evaluation set: "Baseline" stands for the narrowband and wideband models built with the corresponding training speech independently, "DS" represents the system built by randomly mixing Narrowband_Ori and Narrowband_DS speech, "JT-2" is the proposed joint training strategy with same input entry.

CER(%)	Narrowband	Wideband	Avg.
Baseline	30.72	28.29	29.5
DS	30.04	29.55	29.8
JT-2	28.28(7.9%)	27.48(2.9%)	27.88(5.5%)

Several observations can be made from Table II. First, down-sampling wideband speech to narrowband can slightly improve (2.2% relative reduction in CER) the performance of narrowband system after mixing with the down-sampled narrowband speech. However, a significant degradation is observed for wideband speech in DS system. Second, our proposed joint training strategy outperforms baseline system on narrowband speech with a 7.9% relative gain. By considering that the baseline system is already well-trained with a large-scale training dataset and the same amount of training data for Narrowband_Ori and Wideband_Ori speech is used, those improvements are significant on this challenging task. Third, DS system suffers from a performance degradation on average of both narrowband and wideband system while the proposed joint training method yields a 5.5% relative error reduction on average. Table II confirms that the DNN-BWE and DNN-AM structure can accommodate narrowband and wideband speech well for acoustic modeling.

C. Experiments and results for strategy JT-3

Table III illustrates the performance of our proposed joint training strategy for different entries and several other systems for comparison. First, we can observe that up-sampling helps improve the performance of wideband system while suffering from a performance degradation on narrowband system, which is opposite to DS system in Table II. Second, strategy JT-3 outperforms both US system and the system described in [14] due to our bandwidth expansion network which aims at eliminating mismatch between narrowband and wideband speech. We also find that further updating DNN-BWE with Narrowband_Ori LMFB features while keeping DNN-AM fixed yields performance gain on narrowband speech, which improves the performance on narrowband speech from 29.79% (which stands for the performance of JT-3 in step 3) to 29.39%.

Comparing Table II with Table III, it is easy to find that strategy JT-2 tends to boost the performance on narrowband speech while strategy JT-3 performs better on wideband speech. So if we focus on boosting the performance on the narrowband speech, strategy JT-2 is the recommendation. Otherwise, strategy JT-3 is a better choice. We further examine the significance of improvement when comparing our proposed joint modeling strategies with the baseline systems. Here we adopt a "matched pair test" in [22], [23], which is a two-tailed test with the null hypothesis that there is no performance

TABLE III

Performance (CER in %) comparison of the proposed joint modeling strategy and several contrast systems: "US" stands for system built by Wideband_Ori and Narrowband_US speech, [14] represents zero-padding strategy for Narrowband_Ori proposed in [14], "JT-3" stands for the training strategy described in Algorithm 3.

CER(%)	Narrowband	Wideband	Avg.
Baseline	30.72	28.29	29.5
US	30.9	27.7	29.3
[14]	30.39	27.45	28.92
JT-3	29.39(4.3%)	27.2(3.9%)	28.30(4.1%)

difference between the two systems. We use a minimum value of p to indicate a significance difference at the level of p in the statistical significance tests for the baseline and our proposed joint modeling strategies. The smaller "p" is, the more signicant differences between two systems are. We find "p" to be less than 0.001 for all cases.

VI. CONCLUSION

In this paper, we investigate several joint modeling strategies for mixed-band speech recognition via DNN-based bandwidth extension. All these approaches demonstrate the advantages for different application scenarios. By a comprehensive comparison with the existing approaches, our proposed framework yields significant improvements on a Mandarin speech recognition task leveraging upon a large-scale training set with both narrowband and wideband speech.

REFERENCES

[1] P. J. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proceedings of ICASSP*, 1994, pp. I–109.

[2] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] J. Han, G. J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[4] K.-Y. Park and H.-S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *Proceedings of ICASSP*, 2000, pp. 1843–1846.

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

[6] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," in *Proceedings of ICASSP*, 2003, pp. I–680.

[7] G.-B. Song and P. Martynovich, "A study of hmm-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, no. 10, pp. 2036–2044, 2009.

[8] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech." in *Proceedings of Interspeech*, 2005, pp. 1509–1512.

[9] Y. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544–548, 1994.

[10] 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.

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

[12] 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.

[13] 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.

[14] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm," in *Spoken Language Technology Workshop (SLT)*, 2012, pp. 131–136.

[15] Z. You and B. Xu, "Improving wideband acoustic models using mixed-bandwidth training data via dnn adaptation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[16] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[17] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," *in Proceedings of Interspeech*, 2014.

[18] A. Narayanan and D. Wangle, "Joint noise adaptive training for robust automatic speech recognition," in *Proceedings of ICASSP*, 2014, pp. 2504–2508.

[19] G. T., D. J., D. L.R., and L. C.H., "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, 2015, pp. 4375–4379.

[20] S. Zhang, C. Zhang, Z. You, R. Zheng, and B. Xu, "Asynchronous stochastic gradient descent for dnn training," in *Proceedings of ICASSP*, 2013, pp. 6660–6663.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of ICASSP*, 1989, pp. 532–535.

[23] D. S. Pallet, W. M. Fisher, and J. G. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proceedings of ICASSP*, 1990, pp. 97–100.