Mixed-Bandwidth Cross-Channel Speech Recognition via Joint Optimization of DNN-Based Bandwidth Expansion and Acoustic Modeling

Jianqing Gao, Jun Du, and Enhong Chen, *Senior Member, IEEE*

***Abstract*—Automatic speech recognition (ASR) systems are often built using scene related speech data due to large variations of transmission channels and sampling rates in different scenarios. In this study, we propose a general framework that establishes a unified model for diversified speech data with different sampling rates and channels. The framework is a joint optimization of deep neural network (DNN)-based bandwidth expansion and acoustic modeling to exploit a large amount of diversified training data. First, we design two novel DNN architectures to map the acoustic features from narrowband to wideband speech through direct mapping and progressive mapping. The learning targets of the direct mapping DNN (DNN-DM) are the acoustic features extracted from speech with the largest bandwidth, while the acoustic features from speech with all the other bandwidths are used as input. A progressive stacking network (PSN) gradually maps the features from the low sampling rates to the highest sampling rate through the design of intermediate target layers via multitask training. Then, in addition to these bandwidth expansion networks, we investigate several joint training strategies for DNN-based acoustic models. Our experiments conducted on three diversified large-scale Mandarin speech datasets with different recording channels and sampling rates (6, 8, and 16 kHz) show that the proposed unified model using PSN for bandwidth expansion not only is a more flexible and compact design than conventional multiple acoustic models with each bandwidth for a specific sampling rate, but also yields consistent and significant improvements over bandwidth-dependent models with an average relative word error rate reduction of 6.2%, indicating that the proposed model can fully utilize the diversified cross-channel speech data with multiple bandwidths. Moreover, the proposed methods are verified to be robust on different realistic scenes and can be effectively extended to a long short-term memory framework.**

***Index Terms*—Deep neural network, bandwidth expansion, acoustic modeling, joint optimization, mixed-bandwidth speech recognition.**

I. Introduction

AUTOMATIC speech recognition (ASR) as a core technique has been widely used in many speech-enabled fields, e.g., call centers, medical services, and mobile application [1]–[3]. Large variabilities across different transmission channels and sampling rates are challenging problems in acoustic modeling [4]. Accordingly, most off-the-shelf ASR systems are channel or bandwidth dependent. For example, narrowband speech (e.g., with 6 kHz or 8 kHz sampling rates) is often used in telephone speech recognition [5], while wideband speech (e.g., with 16 kHz or 44 kHz sampling rates) is adopted in mobile phone speech recognition [6]. However, rebuilding an ASR system to recognize the speech well from unseen transmission channels and bandwidths is obviously not flexible and often requires expensive data collecting and labeling. Therefore, how to build a unified acoustic model to exploit diversified mixed-bandwidth data is an important research topic.

Historically, there have been two heuristic methods for mixed-bandwidth acoustic modeling. The first method [7] downsamples all the speech to the lowest sampling rate and combines them to train a unified model. The recognition performance of speech with a low sampling rate might be slightly increased due to the additional downsampled data. However, the performance of speech with a high sampling rate is likely to be decreased because speech with a high sampling rate drops some useful information during the downsampling process [8], [9]. The second method [7] upsamples all the speech to the highest sampling rate and trains a hybrid model, which cannot improve the performance over the bandwidth-specific ASR system because no useful information is gained by upsampling.

In the past few years, several approaches have been proposed to construct mixed-bandwidth acoustic models with the aim to train a hybrid acoustic model with a large amount of narrowband data and a small amount of wideband data. In [10], a feature bandwidth extension (FBE)-based approach was proposed to transform narrowband features into wideband features. Then, the original features extracted from the wideband speech were combined with the generated wideband features using FBE and were utilized to train the mixed-bandwidth Gaussian

Manuscript received May 15, 2018; revised October 30, 2018; accepted December 4, 2018. Date of publication December 13, 2018; date of current version December 28, 2018. This work was supported in part by the National Key R&D Program of China under Contract 2017YFB1002202, in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, and in part by the Key Science and Technology Project of Anhui Province under Grant 17030901005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Eric Fosler-Lussier. *(Corresponding author: Jun Du.)*

J. Gao is with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230022, China, and also with iFlytek Research, iFlytek Co., Ltd., Hefei 230027, China (e-mail: jqgao@mail.ustc.edu.cn).

J. Du and E. Chen are currently with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China (e-mail: jundu@ustc.edu.cn; cheneh@ustc.edu.cn).

Digital Object Identifier 10.1109/TASLP.2018.2886739

mixture model-based hidden Markov models (GMM-HMMs) by using the modified expectation maximization (EM) algorithm. Recently, deep learning techniques have been widely adopted for acoustic modeling in ASR systems [11]–[14]. Consequently, many researchers have also investigated the mixed-bandwidth speech recognition in the framework of deep neural network-based hidden Markov models (DNN-HMMs). In [15], mixed-bandwidth speech modeling was formulated as a missing feature problem where several feature dimensions corresponding to high-frequency bands were set to zero when the narrowband speech was presented. The DNN-based approach in [9] treated mixed-bandwidth speech modeling as a domain adaptation problem where the DNN-HMMs trained with a very large amount of narrowband speech could be adapted effectively to the targeted wideband acoustic models with a small amount of wideband speech.

Another type of approach adopts the strategy of bandwidth expansion (BWE), which extends narrowband signals to wideband signals and then combines the extended and original wideband speech to train the acoustic models. Early work employed statistical models [16], such as GMM [17], [18], to learn the mapping between low-frequency bands and high-frequency bands [19]–[23]. Recently, DNN has been used to learn the nonlinear mapping between low-frequency signals and high-frequency signals, which can obviously improve the subjective listening quality of narrowband speech [16], [24]–[26]. This latest approach [27], which causes the expanded narrowband speech to follow the same distribution as wideband speech by using a generative adversarial network (GAN), shows a sharp increase in the quality of narrowband speech. However, the improvement to the listening quality of speech does not necessarily improve the ASR performance [28]. Accordingly, we proposed a mixed-bandwidth acoustic modeling method based on BWE and the strategy of joint training [29]. A DNN-BWE model was applied to learn the mapping of narrowband mel-filterbank features and wideband mel-filterbank features while jointly training a DNN-BWE network and the acoustic model was used to improve the ASR performance. A unified model could be used by speech with different bandwidths, which could yield average relative performance gains of 4.1% on the narrowband and wideband data.

In this study, we extend our previous work [29] with the following new contributions. First, a general framework of mixed-bandwidth cross-channel speech recognition is presented by considering the input speech signals with arbitrary bandwidths rather than only two bandwidths ($K = 2$) as in most previous research such as [29]. As for the practical values of our proposed framework, it is common that there are many speech-enabled applications requiring a speech recognizer fed by speech with different sampling rates ($K > 2$). It is a repetitive and time-consuming work to rebuild the acoustic model for each application, and the amount of training speech data with a certain sampling rate is very limited. Our proposed framework provides a solution to train a unified acoustic model for different applications and fully utilize the diversified training speech data with different sampling rates. For example, in our experiments, based on three ($K = 3$) realistic ASR tasks, we demonstrate the effectiveness of our approach. Second, we design two novel DNN architectures for multiple bandwidth expansion (MBE), namely, direct mapping and progressive mapping. The learning targets of the direct-mapping DNN (DNN-DM) are the acoustic features extracted from speech with the largest bandwidth, while the acoustic features from speech with all the other bandwidths are used as input. The progressive stacking network (PSN) arranges time-synchronized speech features from the lowest bandwidth to the highest bandwidth and then progressively learns the relationship between the features extracted from speech of each pair of ordered bandwidths using one hidden layer by multitasking. Then, in addition to these bandwidth expansion networks, we investigate several joint modeling strategies with DNN-based acoustic models. Third, our experiments conducted on three diversified large-scale Mandarin speech datasets with different recording channels and sampling rates (6 kHz, 8 kHz, and 16 kHz) show that the proposed unified model using PSN for the bandwidth expansion not only is a more flexible and compact design than conventional multiple acoustic models with each bandwidth for a specific sampling rate but also yields consistent and significant improvements over bandwidth-dependent models with an average relative word error rate reduction of 6.2%, indicating that the proposed model can fully utilize the diversified cross-channel speech data with multiple bandwidths. Moreover, as a unified model, the proposed approach with a more compact design also outperforms the conventional approaches including downsampling, upsampling, and zero-padding.

II. DNN-Based Multiple Bandwidth Expansion

In recent years, DNN-based regression networks as nonlinear mapping functions have been widely used in speech areas, e.g., speech enhancement [30] and noise-robust speech recognition [31], [32]. Therefore, it is natural to adopt the regression DNN model to learn the mapping between low-frequency signals (0$\sim$4 kHz) and high-frequency signals (4$\sim$8 kHz). In [28], DNN was used to extend narrowband speech into wideband speech by mapping low-frequency spectra to high-frequency spectra, yielding better subjective listening quality. However, only the ASR performance for narrowband speech was improved; the performance of wideband speech was not improved. In our recent work [29], DNN was used to map the log-mel filter bank (LMFB) features from narrowband speech to wideband speech so that the extended features could be fed into the acoustic model directly, and joint training of a DNN-based BWE model and acoustic model could be conveniently conducted. However, the unified model could only be shared by two types of bandwidth data in the previous work. In this study, we investigate a general framework to solve the problem of mixed modeling for multiple bandwidth data.

A. The Direct-Mapping DNN for MBE

The strategy of MBE used in this paper aims to extend the speech features of various bandwidths to the highest bandwidth. Suppose we have K speech datasets $\{D_1, D_2, \ldots, D_K\}$ with different sampling rates ($B_1 < B_2 < \cdots < B_K$). The MBE network should extend all speech data to the highest sampling

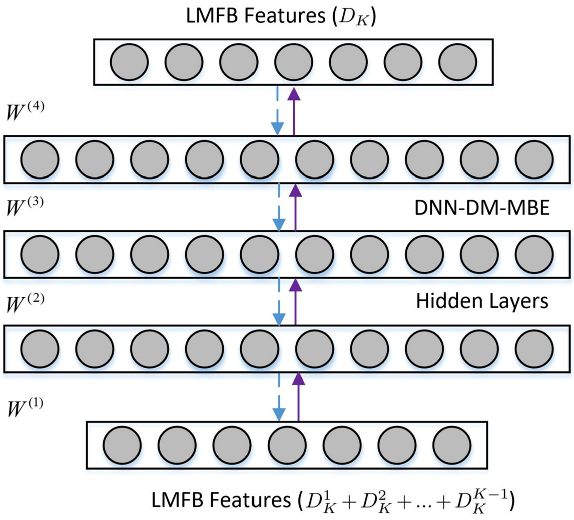


Fig. 1. The illustration of the DNN-DM-MBE.

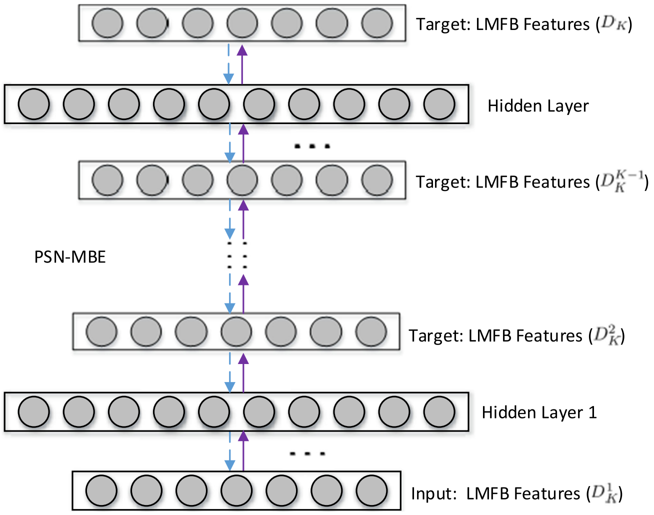

Target: LMFB Features (D_K)
Hidden Layer
Target: LMFB Features (D_K^{K-1})
PSN-MBE
Target: LMFB Features (D_K^2)
Hidden Layer 1
Input: LMFB Features (D_K^1)

Fig. 2. The illustration of the PSN-MBE.

rate of B_K. Accordingly, we first design a feedforward neural network shown in Fig. 1 for bandwidth expansion, including several sigmoidal hidden layers and a linear output layer. We denote this MBE network as DNN-DM-MBE. Considering the direct connection with the DNN-based acoustic model for ASR, the input of the DNN-DM-MBE should be the LMFB features of multiband speech $(B_1, B_2, \ldots, B_{K-1})$, and the output of the DNN-DM-MBE should be the corresponding LMFB features of speech with the highest sampling rate B_K. The DNN-DM-MBE performs the direct mapping between the narrowband LMFB features and the wideband LMFB features with the same dimension, which means that the bandwidth expansion strategy is designed to predict the entire wideband features instead of the missing high-frequency band. Training the DNN-DM-MBE requires multiband parallel data, which are very difficult and expensive to collect. Therefore, we artificially generate a parallel dataset with a variety of bandwidth data by downsampling the dataset D_K. We define the speech datasets as $\{D_K^j | j = 1, 2, \ldots, K-1\}$ where D_K^j denotes the downsampled dataset with the sampling rate B_j generated from D_K with the sampling rate B_K.

With the dataset pairs $\{(D_K^j, D_K) | j = 1, 2, \ldots, K-1\}$, we optimize the parameters of the DNN-DM-MBE using the minimum mean square error (MMSE) criterion as follows:

$$E = \frac{1}{N(K-1)} \sum_{n=1}^{N} \sum_{j=1}^{K-1} \left(||\hat{\boldsymbol{x}}_{n\pm r}^{j,K} - \boldsymbol{x}_{n\pm r}^{K}||_2^2 + \alpha ||\boldsymbol{W}||_2^2 \right) \tag{1}$$

$$\hat{\boldsymbol{x}}_{n\pm r}^{j,K} = \mathcal{F}(\boldsymbol{x}_{n\pm r}^{j}, \boldsymbol{W}, \boldsymbol{b}) \tag{2}$$

where $\boldsymbol{x}_{n\pm r}^{K}$ are the n^{th} $D(2r+1)$-dimensional vectors of the reference wideband LMFB features (extracted from D_K with B_K) with the neighboring left and right r frames, as the acoustic context. $\hat{\boldsymbol{x}}_{n\pm r}^{j,K}$ is a vector estimated by the DNN-DM-MBE, where the input of the DNN-DM-MBE is $\boldsymbol{x}_{n\pm r}^{j}$, a vector of input LMFB features with the acoustic context extracted from the downsampled speech dataset $D_K^j (j = 1, 2, \ldots, K-1)$. N denotes the number of training samples. $\mathcal{F}$ is the neural network function, and $(\boldsymbol{W}, \boldsymbol{b})$ represent all the weight and bias parameters. α is the regularization weighting coefficient to avoid overfitting. We minimize this loss function with the asynchronous stochastic gradient descent (ASGD) [33]. The output of the DNN-DM-MBE network, namely, the wideband features, can be fed to the acoustic network of the ASR system.

B. The Progressive Stacking Network for MBE

The DNN-DM-MBE network directly expands the features of multiple sampling rates to the highest sampling rate. However, the training of the DNN-DM-MBE network might be difficult when the inputs with multiple sampling rates are diverse. It is challenging to expand the bandwidth effectively, especially for the data with lower sampling rates. Progressive learning based on the idea of curriculum learning [34] is used for speech enhancement [35], where each hidden layer of the DNN is designed to learn an intermediate target with an increased SNR. Besides, a work in [36] retains a pool of pre-trained models throughout training and learns lateral connections from these to extract useful features for new tasks. In this paper, we propose a progressive stacking network for MBE (PSN-MBE) similar to [35], which gradually maps the features from the low sampling rate to the highest sampling rate through the design of intermediate target layers. It has the potential to make the bandwidth expansion network easier to train by learning progressively. The PSN-MBE is illustrated in Fig. 2. For example, if there are four types of speech data with the sampling rates 4 kHz, 6 kHz, 8 kHz, and 16 kHz, then two intermediate layers are designed to represent the 6 kHz and 8 kHz sampling rates in addition to the input layer with 4 kHz and output the layer with 16 kHz. As shown in Fig. 2, all target layers, including the intermediate targets and the final output targets, use a linear activation function, while the hidden layers adopt the sigmoid activation function. This stacking style network can learn multiple targets progressively. In the forward process of the PSN-MBE, the features of each intermediate target layer are expanded as the input for learning the next target layer with a higher sampling rate. Then, backpropagation of the PSN-MBE is adopted with a weighted MMSE criterion in terms of multitask learning to update the randomly initialized parameters as follows:

$$E^{\mathrm{PL}} = \sum_{j=2}^{K} \alpha_j E_j^{\mathrm{PL}} + \alpha ||\boldsymbol{W}^{\mathrm{PL}}||_2^2 \tag{3}$$

$$E_j^{\mathrm{PL}} = \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{F}_j^{\mathrm{PL}}\left(\hat{\boldsymbol{x}}_{n\pm r}^{j-1}, \boldsymbol{W}_j^{\mathrm{PL}}, \boldsymbol{b}_j^{\mathrm{PL}}\right) - \boldsymbol{x}_{n\pm r}^{j}\|_2^2 \tag{4}$$

where $\hat{\boldsymbol{x}}_{n\pm r}^{j}$ and $\boldsymbol{x}_{n\pm r}^{j}$ are the n^{th} $D(2r+1)$-dimensional vectors of the estimated and reference target, respectively, of the LMFB feature vectors for the target layer representing $B_j (j > 1)$. α_j is the weighting factor of an objective function for the j^{th} target layer. $\hat{\boldsymbol{x}}_n^1$ is equivalent to $\boldsymbol{x}_n^1$, denoting the n^{th} $D(2r+1)$-dimensional vector of input LMFB features extracted from D_K^1 with B_1. $\mathcal{F}_j^{\mathrm{PL}}(\hat{\boldsymbol{x}}_{n\pm r}^{j-1}, \boldsymbol{W}_j^{\mathrm{PL}}, \boldsymbol{b}_j^{\mathrm{PL}})$ is the neural network function for the j^{th} target using the previously learned intermediate target $\hat{\boldsymbol{x}}_{n\pm r}^{j-1}$, and $(\mathbf{W}_j^{\mathrm{PL}}, \boldsymbol{b}_j^{\mathrm{PL}})$ represents the weight matrices and bias vectors of the hidden layer $j-1$, which are optimized using backpropagation with ASGD. Obviously, the gradients of E_j^{PL} only affect the parameter update of its previous layers.

The PSN-MBE has three advantages compared with the DNN-DM-MBE. First, the PSN-MBE explicitly realizes what is learned from each hidden layer, while the DNN-DM-MBE acts similar to "a black box". Only the final goal of the network is designed, but not the learning goal of the middle layer. Second, we can clearly decompose the mapping among multiple bandwidths into a series of mappings between pairs of bandwidths progressively. Furthermore, the training process is simpler via multitask learning, that is, mappings between different pairs of bandwidths can be trained at the same time. Finally, different from the DNN-DM-MBE mixing all the network parameters together to learn multiple mappings between the lower bandwidths and the highest bandwidth, a more compact model can be designed through the PSN-MBE.

III. Multibandwidth Mixed Acoustic Modeling

The DNN-DM-MBE or PSN-MBE can extend the speech data of a low sampling rate to a high sampling rate, but the bandwidth expansion network trained by the MMSE criterion does not necessarily play a positive role in the performance improvement of the ASR system. Therefore, we investigate jointly training the MBE model and the acoustic model for multibandwidth mixed speech recognition. First, the last layer of the MBE network uses a linear activation function so that the output of the MBE network can be directly fed into the ASR acoustic network. Second, the LMFB features with context frame expansion are adopted as the output of the MBE network. Finally, the joint training of the MBE network and the acoustic network can be performed under the cross-entropy (CE) criterion for improving the recognition performance. In this study, three strategies of multibandwidth joint training (MBJT) are presented by exploiting different MBE models and data usages to fully utilize the multibandwidth mixed speech data.

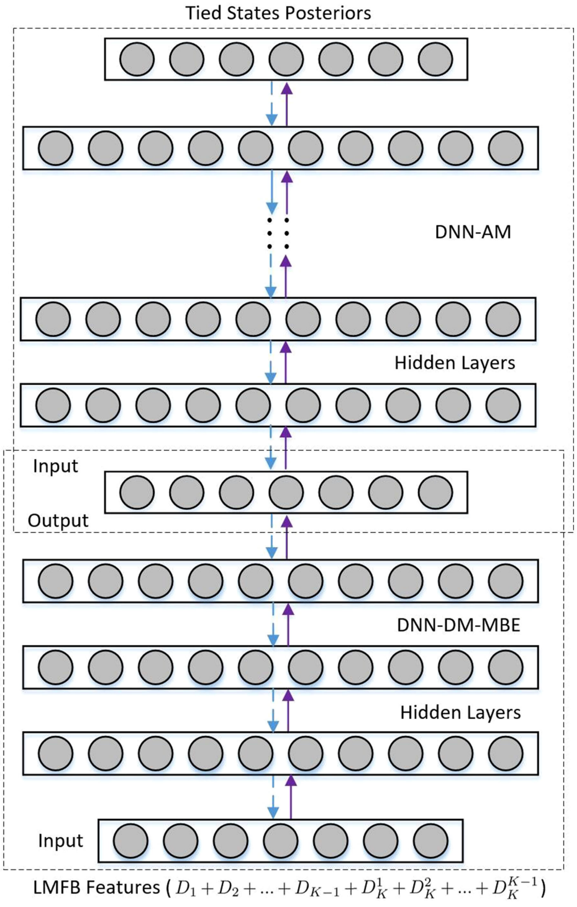


Fig. 3. Same entry using DNN-DM-MBE (MBJT-1).

A. Same Entry Using the DNN-DM-MBE (MBJT-1)

As shown in Fig. 3, after joint training, speech features of all bandwidths share the same entry, namely, the input of the DNN-DM-MBE. The detailed procedure of MBJT-1 is demonstrated in Algorithm 1. First, the DNN-DM-MBE is built with the LMFB features extracted from D_K and $D_K^1, D_K^2, \ldots, D_K^{K-1}$ using the algorithm depicted in Section II-A. Second, the DNN-based acoustic model (DNN-AM) is initialized by a restricted Boltzmann machine (RBM)-based pretraining [37], which is followed by a fine-tuning stage with the output features of the DNN-DM-MBE using the LMFB features from both datasets $\{D_1, D_2, \ldots, D_{K-1}\}$ and $\{D_K^1, D_K^2, \ldots, D_K^{K-1}\}$. Finally, by connecting the DNN-DM-MBE and DNN-AM, joint training with the CE criterion is performed for updating all the parameters of the hybrid DNN.

By using the above algorithm, speech data with arbitrary K-bandwidth data can be used to establish a unified acoustic model. In the decoding stage, the LMFB features of the speech data with $\{B_1, B_2, \ldots, B_{K-1}\}$ are directly fed into the hybrid DNN model. For the speech data with the highest sampling rate B_K, we downsample from B_K to B_{K-1}, and then, the corresponding LMFB features with B_{K-1} are used for

Algorithm 1: Training Procedure of the MBJT-1 Strategy.

```
Step 1: DNN-DM-MBE training
    Train the DNN-DM-MBE with the LMFB features extracted from the dataset pairs
    $\{(D_K^j, D_K)|j = 1, 2, \ldots, K-1\}$ under the MMSE criterion as in Eq. (1).
Step 2: DNN-AM training
    Combine the LMFB features from both datasets $\{D_1, D_2, \ldots, D_{K-1}\}$ and
    $\{D_K^1, D_K^2, \ldots, D_K^{K-1}\}$ randomly in the mini-batch level. Then, feed the combined
    LMFB features into the DNN-DM-MBE; the output features are used to train DNN-AM.
Step 3: Joint training
    Concatenate the DNN-DM-MBE and DNN-AM. Jointly optimize the DNN-DM-MBE and DNN-AM
    under the CE criterion with the combined LMFB features in Step2.
```

decoding. This strategy has the following advantages. First, the sharing of multibandwidth data is efficiently implemented, and the amounts of training data are effectively increased compared to the single-bandwidth modeling case. Therefore, the recognition performance of the speech data with lower sampling rates $\{B_1, B_2, \ldots, B_{K-1}\}$ should be improved. Second, if the DNN-DM-MBE can well learn the feature mapping between the lower bandwidths and the highest bandwidth, the recognition performance of the speech data with the highest sampling rate B_K should also be potentially improved. Finally, the DNN-DM-MBE can also learn channel information during the joint training stage and alleviate the performance loss due to the channel mismatch of different bandwidth data.

B. Different Entries Using the DNN-DM-MBE (MBJT-2)

Though the MBJT-1 mentioned above is an effective strategy for mixed-bandwidth modeling, it is worth noting that this strategy tends to improve the performance for the low sampling rate speech. For the high sampling rate speech, the performance improvement mainly depends on the recovery capability of the high-frequency band using the DNN-DM-MBE. Therefore, we propose another strategy aimed at improving the recognition performance of the speech with the highest sampling rate B_K, as shown in Fig. 4.

Unlike MBJT-1, there are two different entries for the input features with different sampling rates in the MBJT-2. The LMFB features with lower sampling rates $\{B_1, B_2, \ldots, B_{K-1}\}$ are still fed into the DNN-DM-MBE, while the LMFB features with the highest sampling rate B_K are directly sent to the DNN-AM for acoustic modeling. The training procedure is demonstrated in Algorithm 2. First, the DNN-DM-MBE is built similar to MBJT-1. Second, DNN-AM is trained using two datasets, namely, one set of D_K with the highest sampling rate B_K, and the other set of $\{D_1, D_2, \ldots, D_{K-1}\}$ with the lower sampling rates $\{B_1, B_2, \ldots, B_{K-1}\}$, which should be passed to the DNN-DM-MBE for the bandwidth expansion. Then, by concatenating the DNN-DM-MBE and DNN-AM, the two networks are updated with the CE criterion in which D_K is only used to update the DNN-AM and $\{D_1, D_2, \ldots, D_{K-1}\}$ is adopted to update both the DNN-AM and DNN-DM-MBE. In addition, after the joint training of the DNN-DM-MBE and DNN-AM, we perform a fine-tuning of the DNN-DM-MBE using $\{D_1, D_2, \ldots, D_{K-1}\}$ and correct the DNN-AM to further improve the recognition performance of speech data with lower sampling rates.

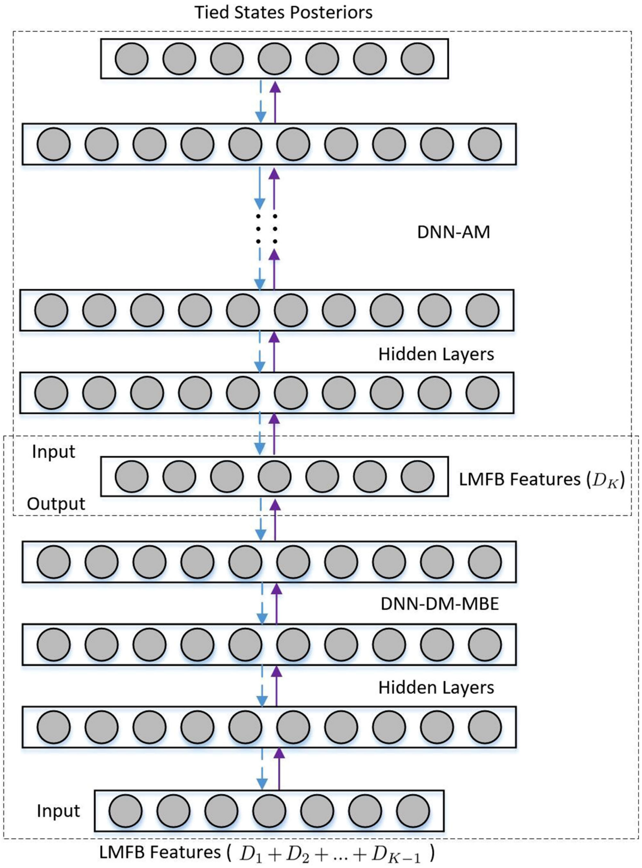


Fig. 4. Different entries using the DNN-DM-MBE (MBJT-2).

The main difference between MBJT-2 and MBJT-1 is that the D_K with the highest sampling rate B_K in the MBJT-2 strategy can be used directly to train the DNN-AM; thus, unlike MBJT-1, the downsampling and bandwidth expansion operations do not result in the loss of information for the speech data with the highest sampling rate at the decoding. In the DNN-AM training and joint training stages, since the speech data for the highest sampling rate and the lower sampling rates are handled differently, the training data are randomized at the mini-batch level, namely, containing the speech features with the same sampling rate in a mini-batch. In the decoding stage, the features with the lower sampling rates are expanded by DNN-DM-MBE and then decoded by DNN-AM, while the features with the highest sampling rate are directly decoded by the DNN-AM. We expect better data sharing can be achieved between lower sampling rates and the highest sample rate without information loss, which will potentially yield better performance for all bandwidths.

C. Different Entries Using the PSN-MBE (MBJT-3)

The MBJT-2 ensures that there is no information loss for speech data with the highest sampling rate, and the speech data

Algorithm 2: Training Procedure of the MBJT-2 Strategy.

```
Step 1: DNN-DM-MBE training
  Train the DNN-DM-MBE with the LMFB features
  extracted from the dataset pairs
  $\{(D_K^j, D_K) | j = 1, 2, \ldots, K-1\}$ under the MMSE
  criterion as described in Algorithm 1.
Step 2: DNN-AM training
  Combine the LMFB features from datasets
  $\{D_1, D_2, \ldots, D_{K-1}, D_K\}$ randomly in the mini-batch
  level. Then, feed the combined LMFB features of
  $\{D_1, D_2, \ldots, D_{K-1}\}$ into the DNN-DM-MBE and the
  LMFB features of $D_K$ into the DNN-AM, and then update
  the DNN-AM with the CE criterion while fixing the
  DNN-DM-MBE.
Step 3: Joint training
  Jointly optimize the DNN-DM-MBE and DNN-AM under
  the CE criterion, using the LMFB features of
  $\{D_1, D_2, \ldots, D_{K-1}\}$ to update both the DNN-DM-MBE
  and the DNN-AM, while using the LMFB features of $D_K$
  to update the DNN-AM only.
Step 4: Fine-tuning of the DNN-DM-MBE
  Further optimize the DNN-DM-MBE with the LMFB
  features of $\{D_1, D_2, \ldots, D_{K-1}\}$ under the CE criterion
  while fixing the DNN-AM.
```

with lower sampling rates can be expanded by the DNN-DM-MBE. However, the DNN-DM-MBE might not be well-trained with multiple bandwidth data since the features of the multiple sampling rates data need to be directly extended to the highest sampling rate. Therefore, we propose a mixed-band modeling approach based on the PSN-MBE, namely, MBJT-3, as shown in Fig. 5. The MBJT-3 improves the MBJT-2 in two aspects. First, unlike MBJT-2 where different speech features with lower sampling rates are fed to the DNN-DM-MBE via the same entry and only speech features with the highest sampling rate are directly sent to the DNN-AM, speech features with K bandwidths are located in different entries of the PSN-MBE (the input layer, intermediate target layers and the output layer) in the MBJT-3. Second, the PSN-MBE in MBJT-3 is trained progressively using the speech features with increased bandwidths as in Section II-B, which can potentially achieve a better convergence property. The detailed training procedure is depicted in Algorithm 3.

The MBJT-3 strategy is quite appealing. During the decoding stage, the speech features with different sampling rates are fed into the hybrid network through different entries so that the decoding time can be reduced for the higher sampling rate. More importantly, the novel design of the progressive network architecture makes the training of the network easier. Furthermore, the high-frequency information expansion can be achieved more effectively, potentially yielding better recognition performance.

D. Initialization of the Mixed-Bandwidth Model

For the above three strategies, the mixed-bandwidth speech data are shared to train a unified acoustic model. One important issue is initializing the unified model, namely, the generation of the state-level labels for different bandwidth data and the parameter initialization of the DNN-AM. In this study, we first use D_K with the highest sampling rate B_K to train a GMM-HMM acoustic model to generate the state-level labels of D_K for the DNN-AM training. Then, the DNN-AM can be initialized using the LMFB features of D_K and the output features of the well-trained DNN-DM-MBE/PSN-MBE with the input LMFB features of the datasets $\{D_K^1, D_K^2, \ldots, D_K^{K-1}\}$. Here, we assume that the output features of the DNN-DM-MBE/PSN-MBE use the same labels as D_K because all datasets $\{D_K^1, D_K^2, \ldots, D_K^{K-1}\}$ are downsampled and frame-synchronized versions of D_K. Finally, with both initialized DNN-DM-MBE/PSN-MBE and DNN-AM, the forced-alignment step is performed to generate the labels for $\{D_1, D_2, \ldots, D_{K-1}\}$ using the unified model.

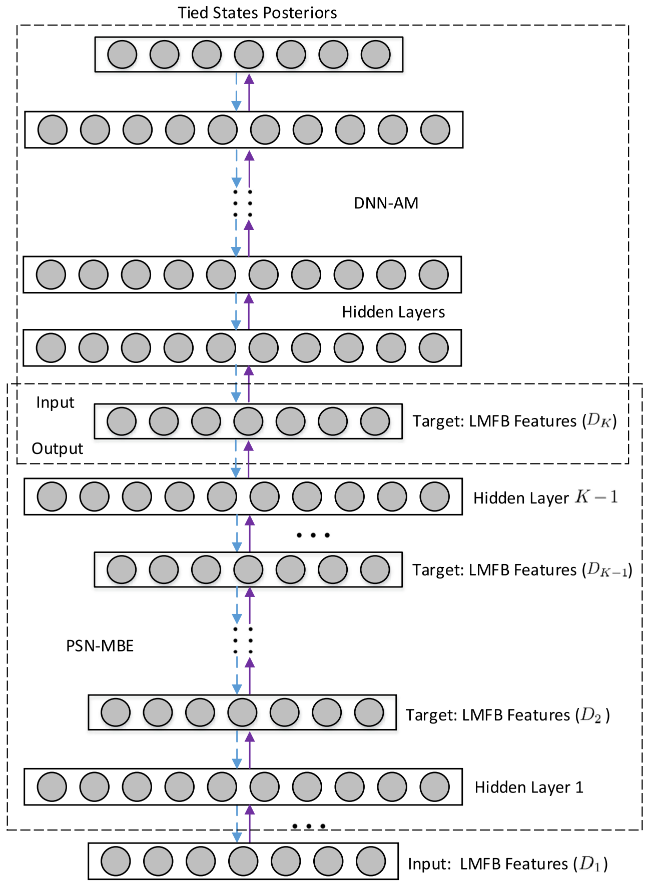

Tied States Posteriors
DNN-AM
Hidden Layers
Input
Output
Target: LMFB Features (D_K)
Hidden Layer $K-1$
Target: LMFB Features (D_{K-1})
PSN-MBE
Target: LMFB Features (D_2)
Hidden Layer 1
Input: LMFB Features (D_1)

Fig. 5. Progressive stacking network based mixed-band modeling (MBT-3 strategy).

IV. Experiments and Result Analysis

To verify the effectiveness of the proposed approach, we defined three Mandarin large vocabulary continuous speech recognition (LVCSR) tasks based on collected realistic cross-channel speech data with different sampling rates (6 kHz, 8 kHz and 16 kHz), as illustrated in Table I. For the purpose of data balance, all the training sets of different bandwidths consisted of 300 hours of speech data, which was comparable to the training set size of the widely used Switchboard task for English

TABLE I
THE DETAILS OF THREE MANDARIN SPEECH RECOGNITION TASKS

Sampling rate	Channel	Speaking style	Microphone	Training set size	Testing set size	Vocabulary	Language model
6 kHz	Telephone	Conversational	Near-field	300 hours	11 hours	68,000	3-gram
8 kHz	Telephone	Conversational	Near-field	300 hours	20.8 hours	86,000	3-gram
16 kHz	Unknown	Spontaneous	Near-field, Far-field	300 hours	26.6 hours	100,000	3-gram

Algorithm 3: Training Procedure of the MBJT-3 Strategy.

```
Step 1: PSN-MBE training
  Train the PSN-MBE under the MMSE criterion as in
  Eq. (3) by feeding the input layer with the LMFB features
  of $D_K^1$ with the lowest sampling rate $B_1$, the intermediate
  target layers with the LMFB features of $\{D_K^2, \ldots, D_K^{K-1}\}$
  with the sampling rates $\{B_2, \ldots, B_{K-1}\}$, and the output
  layer with the LMFB features of $D_K$ with the highest
  sampling rate $B_K$.
Step 2: DNN-AM training
  Combine the LMFB features from datasets
  $\{D_1, D_2, \ldots, D_{K-1}, D_K\}$ randomly in the mini-batch
  level. Then, feed the LMFB features of
  $\{D_1, D_2, \ldots, D_{K-1}\}$ into the PSN-MBE via different
  entries and the LMFB features of $D_K$ into the DNN-AM,
  and then update the DNN-AM with the CE criterion while
  fixing PSN-MBE.
Step 3: Joint training
  Jointly optimize the PSN-MBE and the DNN-AM under
  the CE criterion, using the LMFB features of
  $\{D_1, D_2, \ldots, D_{K-1}, D_K\}$ to update both the DNN-AM
  and PSN-MBE. Please note that only the succeeding
  parameters after each entry for one sampling rate are
  updated.
Step 4: Fine-tuning of the PSN-MBE
  Further optimize the PSN-MBE with the LMFB features of
  $\{D_1, D_2, \ldots, D_{K-1}\}$ under the CE criterion while fixing
  the DNN-AM.
```

LVCSR [38]. The amount of data in each test set for the different bandwidths was listed in Table I. The frame length was 25 ms, while the frameshift was 10 ms. The utterance-level mean normalization was applied to the LMFB features.

For the 6 kHz task, the data were recorded from the call center over the telephone channel with a special VOX compression format. The utterances were the conversations between the client and customer service. Part of the data included serious accents. For ASR setup, the vocabulary size was approximately 68,000 words, and the language model was a 3-gram.

For the 8 kHz task, the data were also recorded over the telephone channel. The speaking style of the speech data was more spontaneous than the Switchboard task [39] because the subject of the conversation was completely open. For the ASR setup, the vocabulary size was approximately 86,000 words, and the language model was a 3-gram.

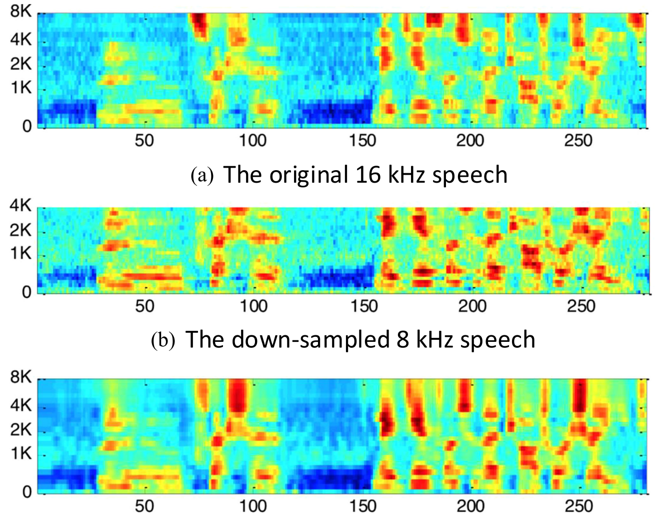
For the 16 kHz task, the data were recorded over various unknown channels. The speaking style of the data was also spontaneous. Part of the data was collected with far-field microphones in noisy environments. For the ASR setup, the vocabulary size was approximately 100,000 words, and the language model was a 3-gram.

To build the acoustic models for both the single-bandwidth and multiple-bandwidth tasks, we adopted the tied-state triphones as the HMM units for Mandarin speech. To generate the aligned state labels, the GMM-HMM system with 9004 tied states and 40 Gaussian mixtures for each state was trained using the mel-frequency cepstral coefficient (MFCC) features. For the single-bandwidth systems, we adopted three separate GMM-HMMs for data of different sampling rates; each system had a different decision tree, but the numbers of states of the three systems were the same. For the mixed-bandwidth system, we used the procedure described in the Section III-D to obtain unified tied states with 16 kHz model. The architecture of the DNN-AM was 825-2048*6-9004, namely, 11 frames of 72-dimensional LMFB features and 3-dimensional pitch features for the input layer, 6 hidden layers with 2048 nodes for each layer, and 9004 tied states for the output layer. Unsupervised RBM pretraining was used for the DNN-AM initialization. The learning rate was 0.004, and 10 iterations were performed under the CE criterion for the DNN-AM fine-tuning.

For the bandwidth expansion network, the architecture of the DNN-DM-MBE was 825-2048*3-825. The input and output layers of the DNN-DM-MBE were 825-dimensional LMFB features with 11-frame expansion, while 3 hidden layers were adopted with 2048 nodes for each layer. For the PSN-MBE ($K = 3$), the architecture was 825-2048-825-2048-825, α_3 and α_2 in Eq. (3) were set to 0.5 and 0.5, respectively. For both the DNN-DM-MBE and the PSN-MBE, the learning rate was set to 0.001, and 30 iterations were performed with the random initialization for training. For the joint training, the learning rates for the DNN-DM-MBE&DNN-AM and the PSN-MBE&DNN-AM were 0.002 and 0.0015 respectively. In the final step of the MBJT-2 and the MBJT-3, the learning rate was set to 0.0015 for the fine-tuning of the DNN-DM-MBE/PSN-MBE. All of the experiments were performed by using our own tools for better parallel computing by multiple machines and multiple GPUs.

Based on the description in Table I, the ASR task seemed more challenging with a higher sampling rate in terms of more spontaneous speaking styles, more adverse environments and a larger vocabulary. Specifically, due to quite limited domain data of the call center in the 6 kHz task, the number of 3-gram models for the 6 kHz task was much smaller than for the 8 kHz and the 16 kHz tasks, the recognition performance of the 6 kHz task was always significantly better than the 8 kHz and the 16 kHz tasks in the subsequent experiments.

In the following subsections, we first examined the effectiveness of the bandwidth expansion network. Then, the recognition performance of mixed-band speech modeling was analyzed.



(a) The original 16 kHz speech

(b) The down-sampled 8 kHz speech

(c) The recovered 16 kHz speech using DNN-DM-MBE

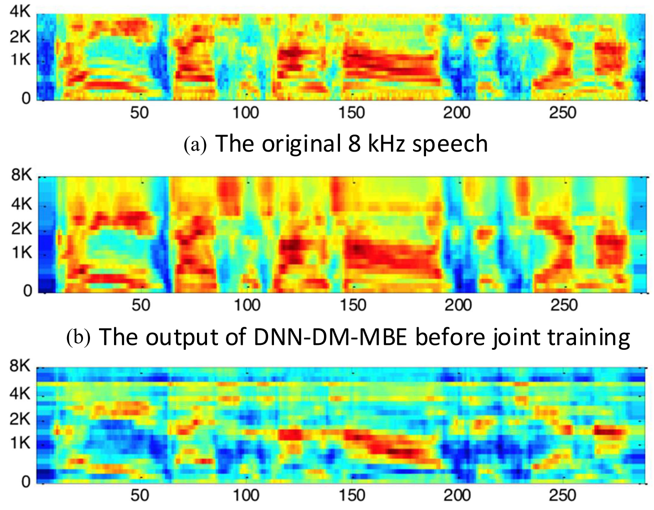
Fig. 6. An example comparison of the LMFB features using the DNN-DM-MBE with sampling rates from 8 kHz to 16 kHz ($K = 2$).

Moreover, the robustness issues of the proposed approach, including accents, speaking styles, microphone settings, and multiple speakers, were comprehensively discussed. Finally, the expansive experiments based on LSTM were performed.

A. *Effectiveness of the Bandwidth Expansion Network*

In this section, we examined the effectiveness of the bandwidth expansion networks using the visualization of the LMFB features from three aspects. First, the experiment on the simulation data was conducted for the proof of concept. Second, the experiment on the realistic data was designed to show the differences of the bandwidth expansion network output before and after the joint training described in Section III. Third, we compared the DNN-DM-MBE and the PSN-MBE.

1) *Proof of Concept on the Simulation Data:* As shown in Fig. 6, we listed an utterance example from the 16 kHz test set. We first downsampled the 16 kHz waveform to an 8 kHz waveform and then compared the corresponding LMFB features among the 16 kHz, downsampled 8 kHz and recovered 16 kHz waveforms using the DNN-DM-MBE. The horizontal axis was the frame index, while the vertical axis was the mel-frequency index (also applied for Figs. 7 and 8). The DNN-DM-MBE was trained only with the 16 kHz training set and its downsampled 8 kHz version ($K = 2$). From Fig. 6, we observed that the high-frequency bands between 4 kHz and 8 kHz of the recovered LMFB features were similar to those of the reference 16 kHz case, which demonstrated the effectiveness of the DNN-DM-MBE.

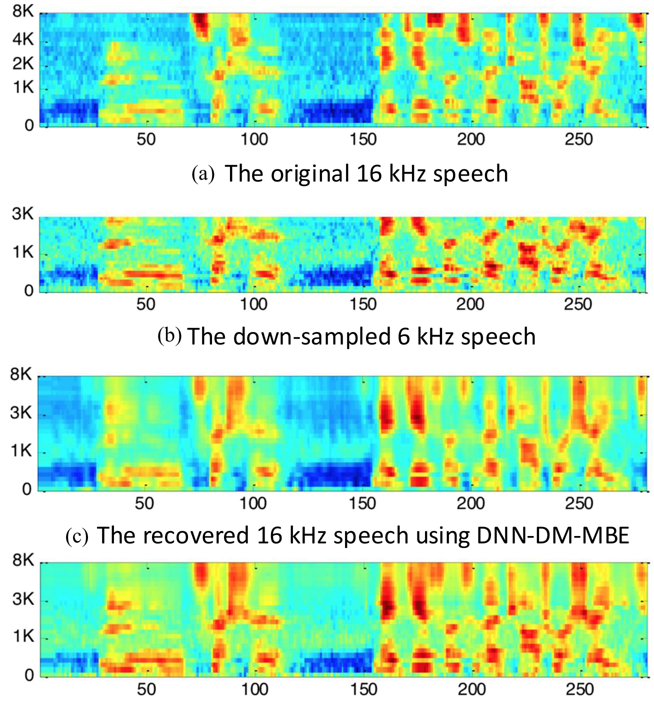


(a) The original 8 kHz speech

(b) The output of DNN-DM-MBE before joint training

(c) The output of DNN-DM-MBE after joint training

Fig. 7. An example comparison of the DNN-DM-MBE outputs before/after joint training with the sampling rate extended from 8 kHz to 16 kHz ($K = 2$).



(a) The original 16 kHz speech

(b) The down-sampled 6 kHz speech

(c) The recovered 16 kHz speech using DNN-DM-MBE

(d) The recovered 16 kHz speech using PSN-MBE

Fig. 8. An example comparison between the DNN-DM-MBE and the PSN-MBE with the sampling rate extended from 6 kHz to 16 kHz ($K = 3$).

2) *The DNN-DM-MBE Outputs Before/After Joint Training:* In Fig. 7, we showed a realistic example of the bandwidth expansion from 8 kHz to 16 kHz using the utterance from the 8 kHz test set as the input, different from the downsampled 8 kHz version in Fig. 6. Both the outputs of the DNN-DM-MBE before/after the joint training in MBJT-1 were compared. The DNN-DM-MBE before joint training created high bands of the LMFB features between 4 kHz and 8 kHz. After joint training, the DNN-DM-MBE output generated quite different results from those before training because the optimizing of the DNN-DM-MBE was guided by the CE criterion for the ASR rather than the MMSE criterion. We discussed the recognition performance in Section IV-B.

3) *DNN-DM-MBE vs. PSN-MBE:* Finally, we maked a comparison between the DNN-DM-MBE and the PSN-MBE for multiple bandwidth expansion including 6 kHz, 8 kHz and

TABLE II
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-1 STRATEGY AND OTHER METHODS ON 8 kHz AND 16 kHz TEST SETS

CER(%)	8 kHz	16 kHz	Avg.
Baseline	31.8	31.8	31.8
DS	31.4	32.6	32.0
MBJT-1-Init	31.8	33.2	32.5
MBJT-1	30.0	31.1	30.6

TABLE III
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-1 STRATEGY AND OTHER METHODS ON THE 6 kHz AND 16 kHz TEST SETS

CER(%)	6 kHz	16 kHz	Avg.
Baseline	14.7	31.8	23.3
DS	15.2	35.0	25.1
MBJT-1-Init	15.6	35.9	25.8
MBJT-1	14.7	33.8	24.2

TABLE IV
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-2 STRATEGY AND OTHER METHODS ON 8 kHz AND 16 kHz TEST SETS

CER(%)	8 kHz	16 kHz	Avg.
Baseline	31.8	31.8	31.8
US	31.9	31.2	31.6
MBJT-1	30.0	31.1	30.6
MBJT-2-Init	32.2	30.4	31.3
MBJT-2	30.6	29.6	30.1

16 kHz ($K = 3$), as illustrated in Fig. 8. Here, we only showed the expansion results from 6 kHz to 16 kHz. We also used the same 16 kHz waveform as in Fig. 6 to generate the downsampled 6 kHz waveform. As the downsampled 6 kHz waveform lost more details of the high-frequency bands compared with the downsampled 8 kHz waveform, it was more challenging to recover the contents of the high-frequency bands as shown in the comparison between Fig. 8(c) and Fig. 6(c). However, by using the PSN-MBE, the recovered feature in Fig. 8(d) was more similar to the reference 16 kHz feature than that in Fig. 8(c), especially in the high-frequency bands. We calculated the mean squared error between the recovered features and the reference features for this utterance. The mean squared error of 55.9 shown in Fig. 8(d) was smaller than the mean squared error of 59 shown in Fig. 8(c). For the entire 16 kHz testing set, the mean squared error of 78.1 with PSN-MBE was also smaller than the mean squared error of 93.6 with DNN-DM-MBE.

B. Comparison of Mixed-Bandwidth Speech Recognition

In this section, we validated the effectiveness of several proposed mixed-bandwidth modeling approaches, namely, MBJT-1, MBJT-2, and MBJT-3. First, the experiments with MBJT-1 and MBJT-2 were conducted for mixed-bandwidth modeling cases with $K = 2$ including (6 kHz, 16 kHz) and (8 kHz, 16 kHz). Second, MBJT-2 and MBJT-3 were compared for the mixed-bandwidth modeling case with $K = 3$ including all three bandwidths. The baseline single-bandwidth models were always trained on 300 hours of speech, and all other mixed-bandwidth models were trained on 600 hours ($K = 2$) or 900 hours ($K = 3$).

1) $K = 2$: Table II lists the character error rate (CER) comparison of the MBJT-1 strategy and several conventional approaches for mixed-bandwidth modeling of 8 kHz and 16 kHz data. “Baseline” refers to the single-bandwidth acoustic models built using 8 kHz and 16 kHz data. “DS” denotes the unified acoustic model trained using both 8 kHz data and downsampled 8 kHz data from 16 kHz data, which is one heuristic mixed-bandwidth modeling strategy. “MBJT-1” represents the MBJT-1 approach in Algorithm 1, and “MBJT-1-Init” is the version without Step 3 in Algorithm 1, which means that joint training of the DNN-DM-MBE and DNN-AM is not performed. In Table II, the strategy of downsampling (DS) improved the performance of the 8 kHz test set. However, the performance degraded for the 16 kHz test set because the high-frequency band of the 16 kHz data was lost after downsampling. The “MBJT-1-Init” without joint training even led to worse results than the “Baseline”, which might be explained because the data for training the DNN-DM-MBE did not match the training data used for the DNN-AM. Therefore, the joint training in “MBJT-1” could alleviate the mismatch, yielding consistent improvements on both the 8 kHz and 16 kHz test sets. For example, the CER of “MBJT-1” decreased from 31.8% to 30% compared with the “Baseline” for the 8 kHz case, demonstrating the effectiveness of the MBJT-1 strategy for mixed-bandwidth modeling.

We also performed another set of experiments on the mixed-bandwidth modeling of 6 kHz and 16 kHz data as shown in Table III. “MBJT-1” still outperformed “DS” and “MBJT-1-Init”. One main difference observed was that in Table II, “MBJT-1” could not improve the performance over the “Baseline.” One reason was the bandwidth expansion from 6 kHz to 16 kHz was obviously more difficult than the expansion from 8 kHz to 16 kHz. Another reason was that the speaker styles of the 6 kHz and 16 kHz data were quite different.

In both Tables II and III, the performance gain of “MBJT-1” over “Baseline” for speech with the highest sampling rates (16 kHz) was less significant than that for speech with lower sampling rates (8 kHz or 6 kHz). To further improve the performance of speech with the highest sampling rate, the MBJT-2 strategy proposed to directly feed the features of the highest sampling rate into the DNN-AM during the training and testing stages.

Table IV lists the CER comparison of the MBJT-2 strategy and several conventional approaches for mixed-bandwidth modeling of 8 kHz and 16 kHz data. “US” denotes the unified acoustic model built using both the 16 kHz data and the upsampled 16 kHz data from the original 8 kHz data, which is another heuristic mixed-bandwidth modeling strategy. “MBJT-2” represents the MBJT-2 approach in Algorithm 2 while “MBJT-2-Init” is the version without Step 3 and Step 4 in Algorithm 2. Based on Table IV, the “US” strategy improved the performance of the 16 kHz test set. Nevertheless, the performance was slightly degraded for the 8 kHz test set. Obviously, this was an opposite observation compared with “DS” in Table II. Similar to “MBJT-1-Init” in Table II, “MBJT-2-Init” did not perform well on both the 8 kHz and 16 kHz tasks without joint training of the bandwidth expansion model and the acoustic model. Mean-

TABLE V
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-2 STRATEGY AND OTHER METHODS ON THE 6 KHZ AND 16 KHZ TEST SETS

CER(%)	6 kHz	16 kHz	Avg.
Baseline	14.7	31.8	23.3
US	15.5	32.0	23.8
MBJT-1	14.7	33.8	24.2
MBJT-2-Init	15.7	30.9	23.3
MBJT-2	14.5	31.1	22.8

TABLE VI
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-2/MBJT-3 STRATEGIES AND OTHER METHODS ON THE TEST SETS OF ALL THREE BANDWIDTHS. $N_{\rm M}$ AND $N_{\rm T}$ ARE THE MODEL SIZE AND RUN-TIME LATENCY, RESPECTIVELY, NORMALIZED BY THE BASELINE SYSTEM

CER(%)	6 kHz	8 kHz	16 kHz	Avg.	$N_{\rm M}$	$N_{\rm T}$
Baseline	14.7	31.8	31.8	26.1	1	1
DS	14.9	31.8	34.8	27.2	0.33	1
US	15.2	31.4	31.8	26.1	0.33	1
Zero-padding [16]	15.2	31.2	31.2	25.9	0.34	1.01
MBJT-2-Init	15.8	33.8	31.1	26.9	0.43	1.29
MBJT-2	14.1	30.1	29.9	24.7	0.43	1.29
MBJT-3-Init	15.2	31.5	31.3	26.0	0.39	1.16
MBJT-3	14.0	29.5	30.0	24.5	0.39	1.16

while, “MBJT-2” consistently and remarkably improved the performance on the two types of data. One main difference from “MBJT-1” was that the performance gain on the 16 kHz data was more significant due to the preservation of high-frequency information of the 16 kHz data. In addition, the average result on the two sets was better than that of “MBJT-1”.

The corresponding experimental results on the 6 kHz and 16 kHz tasks are shown in Table V. The observation was similar to that in Table IV. Although the improvements were less significant than those on the 8 kHz and 16 kHz tasks due to more information loss from the 16 kHz to 6 kHz data and more data diversity between the 16 kHz and 6 kHz tasks, “MBJT-2” still significantly outperformed “MBJT-1” in Table III.

2) $K = 3$: The above experiments proved that the strategy of MBJT-2 could build a unified acoustic model and achieve performance gains for mixed-bandwidth modeling with two types of sampling rates. In this section, we increased the modeling difficulty and designed the experiments on all three bandwidths to verify the scalability of the proposed strategy for mixed-bandwidth acoustic modeling. Since the average performance of the MBJT-2 strategy was better than that of the MBJT-1 strategy on two types of data, we only compared the MBJT-2 and MBJT-3 strategies.

Table VI lists the results of the MBJT-2/MBJT-3 strategies and other approaches on the test sets of all three bandwidths for mixed-bandwidth speech recognition. “Baseline” refers to the independent acoustic models built using the 6 kHz, 8 kHz and 16 kHz data. “DS” and “US” represent the unified acoustic models using the simple downsampling and upsampling strategy, respectively, where all data are downsampled to 6 kHz or upsampled to 16 kHz. For “zero-padding” proposed in [16], the high-frequency parts of the LMFB features were padded to zero for the data with lower sampling rates. “MBJT-2-Init” and “MBJT-2” are similar to those in Tables IV and V. “MBJT-3” represents the MBJT-3 approach in Algorithm 3, while “MBJT-3-Init” is the version without Step 3 and Step 4 in Algorithm 3.

From Table VI, “DS” and “US” achieved the goal of unified modeling, but the recognition performance was not stable for the data with different sampling rates. Specifically, compared with the “Baseline” systems, the performance of “DS” was significantly degraded on the 16 kHz data due to the loss of the high-frequency information while the CER of “US” slightly increased on the 6 kHz data. “Zero-padding”, as one mixed-bandwidth modeling approach, slightly outperformed “US” on average. None of the above mentioned approaches could generate consistent gains over “Baseline” for all types of sampling rates. However, our proposed “MBJT-2” and “MBJT-3” yielded consistent and remarkable improvements, demonstrating the powerful modeling capability of diversified data with different speaker styles and channel variations. For example, the CER reductions of ‘MBJT-2” over “Baseline” were 0.6%, 1.7%, and 1.9% for 6 kHz, 8 kHz, and 16 kHz, respectively. Moreover, “MBJT-3” with a compact design of the bandwidth expansion network via the progressive mapping among the LMFB features with different sampling rates achieved a slight gain over “MBJT-2”, indicating the reasonableness of the progressive architecture. Based on the comparison between Table IV and Table VI, “MBJT-2” was still quite effective with the increase of K implying the good generalization capability to multiple bandwidths.

We also compared the model size and run-time latency in Table VI. It was clear that the proposed “MBJT-2” and “MBJT-3” required additional memory and computational costs compared to “Baseline” (for one bandwidth), “DS”, “US”, and “zero-padding” due to the use of the bandwidth expansion network. Therefore, one concern might be the performance when using more parameters in the DNN-AM for those systems not using the bandwidth expansion model. The experiments confirmed that increasing the hidden layers of the DNN-AM did not obtain the performance gains. For example, the average CER of the “Baseline” system using 10 hidden layers was 26.11%, which was almost the same as that using 6 hidden layers (26.1%) as shown in Table VI. This finding implied that the DNN-AM could not implicitly model the diversified data well by simply using more parameters, while the DNN-DM-MBE or the PSN-MBE can explicitly learn the bandwidth expansion relationship to improve the performance of ASR.

In Table VI, compared with “MBJT-2”, the model size and run-time latency of “MBJT-3” were both smaller. Accordingly, we showed the learning curves of the DNN-DM-MBE and the PSN-MBE on the cross-validation set in Fig. 9. We used only 6 kHz and 16 kHz data pairs to train the DNN-DM-MBE and the PSN-MBE, and compared the 6 kHz to 16 kHz part of the PSN-MBE loss with DNN-DM-MBE loss to ensure comparability. Obviously, the PSN-MBE achieved a faster convergence than the DNN-DM-MBE, indicating that the progressive architecture in the PSN-MBE was easier to train. This result was because the design of feeding the features with different sampling rates to the PSN-MBE via different entries aimed to progressively recover the lost information of the high-frequency bands, which

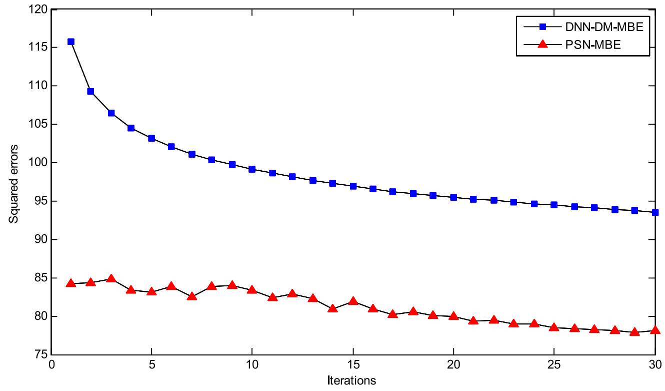


Fig. 9. The learning curves of the DNN-DM-MBE and the PSN-MBE on the cross-validation set ($K = 3$).

potentially decomposed a challenging bandwidth expansion problem into several easier subproblems.

C. Robustness Analysis

In previous experiments, we only discussed the overall performance of each test set with one sampling rate. In this section, we conducted a deep analysis of the robustness issues of the proposed MBJT-3 approach ($K = 3$) in realistic scenes, including the accents, speaking styles, microphone settings, and multiple speakers.

Table VII shows the performance comparison of the "MBJT-3" ($K = 3$) and "Baseline" systems on well-designed test subsets of all three tasks to examine the robustness issues. For the 6 kHz task, we designed two test subsets with 2-hour data from the whole test set to verify the impact of the accent. It is well known that there are many types of dialects in China. Although we included the accented speech data in the training stage, the CER of the "Baseline" system on accented speech data was still much worse than that on the Mandarin speech data (17.5% vs. 13.9%). Our proposed "MBJT-3" was quite effective on the Mandarin speech, yielding a relative CER reduction (CERR) of 10%. However, the performance gain of the "MBJT-3" over the "Baseline" on the accented speech was halved compared with that on the Mandarin speech.

For the 8 kHz task, one subset with 2-hour data consisted of the formal speech in business calls, while the other subset with 1.6-hour data was composed of the conversational speech in family calls. Different from the accented case in the 6 kHz task, "MBJT-3" was more robust for both the formal and conversational speaking styles, with relative CERRs of 7.8% and 7.5%, respectively.

Finally, we examined the environmental issues on the 16 kHz task using three test subsets. The test subset with 9-hour data recorded by a close-talking microphone was referred to as the press interview scenario. The 2-hour far-field data were the speech of teachers with the background noises in the classroom. The third subset with 3-hour data was collected in a meeting scenario with multiple speakers. Obviously, the "Baseline" performance sharply declined for the low SNR speech in the far-field and the overlapping speech of multiple speakers, generating double or triple CERs compared with the close-talking case. However, for all three environments, "MBJT-3" achieved consistent performance gains. In addition, for those adverse acoustic environments, the relative CERRs were slightly better than that in the close-talking case. In Table VII, we examined almost all possible issues in realistic scenes, and the proposed "MBJT-3" demonstrated its high robustness and stability. This result implied that the proposed "MBJT-3" without using additional training data, compared to the "Baseline" system, may be deployed as a practical solution for mixed-bandwidth speech recognition.

D. LSTM Acoustic Model Experiments

All of the above experiments were conducted on the DNN model, but it is well-known that the LSTM model is a stronger acoustic model structure for ASR since the long-term correlation of speech can be modeled well by the LSTM model [40]. Therefore, in this section, some mixed-bandwidth modeling experiments were conducted based on the LSTM model to verify the scalability of our proposed strategies.

We experimented on all three bandwidths data ($K = 3$) and used the LSTM for both the acoustic models and the two bandwidth extension networks, denoted as LSTM-AM, LSTM-DM-MBE, and LSTM-PSN-MBE. The architecture of LSTM-AM was 75-2048(512)*3-9004, namely, 72-dimensional LMFB features and 3-dimensional pitch features for the input layer, 3 hidden layers with 2048 nodes for each layer and 512 nodes for each project layer, and 9004 tied states for the output layer. LSTM-AM was trained under the CE criterion with random initialization, the learning rate was 0.08, and 10 iterations were performed.

For the bandwidth expansion network based on LSTM, the architecture of LSTM-DM-MBE was 75-1024(512)*3-75. The input and output layers of the LSTM-DM-MBE were 75-dimensional LMFB and pitch features, while 3 hidden layers were adopted with 1024 nodes for each layer and 512 nodes for each project layer. For LSTM-PSN-MBE ($K = 3$), the architecture was 75-1024(512)-75-1024(512)-825, and α_3 and α_2 were set to 0.5 and 0.5, respectively. For both the LSTM-DM-MBE and the LSTM-PSN-MBE, the learning rate was set to 0.02, and 15 iterations were performed with the random initialization for training. For the joint training, the learning rates for the LSTM-DM-MBE&LSTM-AM and the LSTM-PSN-MBE&LSTM-AM were 0.04. In the final step of MBJT-2 and MBJT-3, the learning rate was set to 0.02 for the fine-tuning of the LSTM-DM-MBE/LSTM-PSN-MBE.

Table VIII lists the results of the MBJT-2/MBJT-3 strategies based on the LSTM on test sets of all three bandwidths. "DNN" refers to the independent DNN acoustic models built using the 6 kHz, 8 kHz and 16 kHz data. "LSTM" refers to the independent LSTM acoustic models. 'MBJT-2-LSTM" and "MBJT-3-LSTM" represent the MBJT-2 and MBTJ-3 approach similar to Table VI, while the LSTM is used for the bandwidth extension network and the acoustic network.

From Table VIII, the LSTM-based acoustic model achieved significant performance improvements over the DNN-based

TABLE VII
PERFORMANCE (CER IN %) COMPARISON OF THE "MBJT-3" ($K = 3$) AND "BASELINE" SYSTEMS ON WELL-DESIGNED TEST SUBSETS OF ALL THREE TASKS TO EXAMINE THE ROBUSTNESS ISSUES

CER(%)	6 kHz		8 kHz		16 kHz		
	No Accent	With Accent	Formal	Conversational	Close-talking	Far-field	Multispeakers
	2 hours	2 hours	2 hours	1.6 hours	9 hours	2 hours	3 hours
Baseline	13.9	17.5	27.7	30.8	12.7	25.0	36.3
MBJT-3	12.5	16.7	25.5	28.5	12.1	23.4	34.1
Relative CERR	10.0%	4.6%	7.8%	7.5%	4.9%	6.2%	6.0%

TABLE VIII
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-2/MBJT-3 STRATEGIES ON THE TEST SETS OF ALL THREE BANDWIDTHS. N_{M} AND N_{T} ARE THE MODEL SIZE AND RUN-TIME LATENCY, RESPECTIVELY, NORMALIZED BY THE BASELINE LSTM SYSTEM

CER(%)	6 kHz	8 kHz	16 kHz	Avg.	N_{M}	N_{T}
DNN	14.7	31.8	31.8	26.1	-	-
LSTM	13.8	28.2	27.6	23.2	1	1
MBJT-2-LSTM	13.5	27.0	26.4	22.3	0.47	1.42
MBJT-3-LSTM	13.3	26.8	26.4	22.2	0.40	1.20

TABLE IX
PERFORMANCE (CER IN %) COMPARISON OF THE MBJT-2/MBJT-3 STRATEGIES ON THE TEST SETS OF ALL THREE BANDWIDTHS UNDER THE SDT CRITERION

CER(%)	6 kHz	8 kHz	16 kHz	Avg.
LSTM	13.5	26.9	26.1	22.1
MBJT-2-LSTM	13.2	25.8	25.0	21.3
MBJT-3-LSTM	13.1	25.7	25.0	21.2

model, with relative CERR of 11.8%. Our proposed "MBJT-2-LSTM" and "MBJT-3-LSTM" yielded consistent improvements, although the baseline LSTM system had better performance. The experimental results further verified the scalability of the proposed mixed-bandwidth strategies. Moreover, "MBJT-3-LSTM" achieved a slight performance gain over "MBJT-2-LSTM" with the smaller model size and run-time latency similar to Table VI (relative CERRs of 4.4% and 3.9%), indicating the rationality of the progressive mapping.

Furthermore, we verified the performance after sequence discriminative training (SDT), and the state-level minimum Bayes risk (sMBR) criterion was adopted [41]. For the baseline LSTM system, LSTM-AM was trained under the sMBR criterion, the learning rate was 0.002 and 1 iteration was performed. For the mixed-bandwidth system, only the joint model of the LSTM-DM-MBE/LSTM-PSN-MBE and LSTM-AM was trained under the sMBR criterion, and the learning rates for the LSTM-DM-MBE&LSTM-AM and LSTM-PSN-MBE&LSTM-AM were 0.002 and 0.001 respectively.

Table IX listed the results of the MBJT-2/MBJT-3 strategies based on the LSTM on the test sets, the configuration was the same as in Table VI, except that the SDT criterion was used. From Table IX, the SDT criterion achieved consistent performance improvements in various configurations. Moreover, the SDT criterion also worked for our proposed mixed-bandwidth modeling strategies. Finally, the MBJT-3 achieved a CERR of 4% over the baseline LSTM under the SDT criterion. These experimental results indicated that the proposed approaches can be extended to more complicated acoustic modeling structures and stronger training criteria.

V. CONCLUSION

In this study, we first designed two DNN architectures, namely, direct mapping and progressive mapping, for bandwidth expansion among speech data with multiple sampling rates. Accordingly, three strategies of mixed-bandwidth speech recognition by jointly optimizing the DNN-based bandwidth expansion and acoustic model were proposed. We showed that the bandwidth expansion network can be better trained by using progressive mapping for multiple bandwidth data. The joint modeling approach with the progressive mapping network for bandwidth expansion can fully utilize the diversified cross-channel data with different sampling rates. The proposed strategies are quite effective and robust to many realistic issues and can be extended to LSTM framework, which can potentially become a practical solution for mixed-bandwidth speech recognition.

REFERENCES

[1] A. Chandel *et al.*, "Sensei: Spoken language assessment for call center agents," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2008, pp. 711–716.

[2] G. Muhammad, "Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system," *Cluster Comput.*, vol. 18, no. 2, pp. 1–8, 2015.

[3] J. Schalkwyk *et al.*, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition*, New York, NY, USA: Springer, 2010, pp. 61–90.

[4] Y. He and J. Han, "Gaussian specific compensation for channel distortion in speech recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 10, pp. 599–602, Oct. 2011.

[5] G. Saon, H. K. J. Kuo, S. Rennie, and M. Picheny, "The IBM 2015 english conversational telephone speech recognition system," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, pp. 1–15, 2015.

[6] M. Schuster, "Speech recognition for mobile devices at Google," in *Proc. Pacific RIM Int. Conf. Artif. Intell.*, 2010, pp. 8–10.

[7] X. Huang and A. Acero, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[8] P. J. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 1, pp. I–109.

[9] Z. You and B. Xu, "Improving wideband acoustic models using mixed-bandwidth training data via DNN adaptation," in *Proc. INTERSPEECH*, 2014, pp. 2204–2208.

[10] M. L. Seltzer and A. Acero, "Training wideband acoustic models using mixed-bandwidth training data for speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 235–245, Jan. 2007.

[11] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 145–154, Jan. 2011.

[12] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4688–4691.

[13] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[14] G. E. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[15] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. Spoken Lang. Technol. Workshop*, 2012, pp. 131–136.

[16] Y. Wang, S. Zhao, W. Liu, M. Li, and J. Kuang, "Speech bandwidth expansion based on deep neural networks," in *Proc INTERSPEECH*, 2015, pp. 2593–2597.

[17] J. Han, G. J. Mysore, and B. Pardo, "Language informed bandwidth expansion," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2012, pp. 1–6.

[18] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden Markov model," in *Proc. IEEE Workshop Speech Coding*, 2000, pp. 133–135.

[19] K.-Y. Park and H.-S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, pp. 1843–1846, 2000.

[20] H. Seo, H.-G. Kang, and F. K. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6087–6091.

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

[22] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1509–1512.

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

[24] B. Liu, J. Tao, Z. Wen, Y. Li, and D. Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Proc. INTERSPEECH*, 2015, pp. 2598–2602.

[25] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4395–4399.

[26] A. Johannes, S. Maximilian, and F. Tim, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5469–5473.

[27] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5029–5033.

[28] 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 *Proc. INTERSPEECH*, 2015, pp. 2578–2582.

[29] 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 *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 588–594.

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

[31] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2504–2508.

[32] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4375–4379.

[33] S. Zhang, C. Zhang, Z. You, R. Zheng, and B. Xu, "Asynchronous stochastic gradient descent for DNN training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, 2013, pp. 6660–6663.

[34] Y. Bengio, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[35] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Proc. INTERSPEECH*, 2016, pp. 3713–3717.

[36] A. A. Rusu *et al.*, "Progressive neural networks," arXiv:1606.04671, 2016.

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

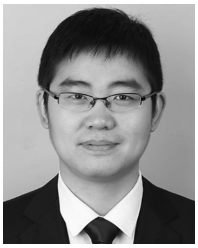
[38] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2011, pp. 24–29.

[39] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, 2011, pp. 437–440.

[40] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Comput. Sci. INTERSPEECH*, 2014, pp. 338–342.

[41] V. Karel, G. Arnab, B. Lukas, and P. Daniel, "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.

Jianqing Gao received the B.S. degree from Hefei University of Technology, Hefei, China, in 2006. He is currently working toward the Ph.D. degree at the University of Science and Technology of China, Hefei, China,. Since 2006, he has been with iFlytek Research on speech recognition and spoken dialogue system. His general area of research includes speech recognition, text summarization, dialogue systems, and machine learning. He has obtained more than 20 patents in the domain of speech recognition.



Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2004 to 2009, he was with the iFlytek Speech Lab of USTC. During the above period, he worked as an intern twice for nine months at Microsoft Research Asia (MSRA), Beijing. In 2007, he also worked as a Research Assistant for six months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he was with iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC.

Enhong Chen (SM'07) received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China. He is currently a Professor and the Vice Dean of the School of Computer Science, USTC. His general areas of research include data mining and machine learning, social network analysis, and recommender systems. He has published more than 100 papers in refereed conferences and journals, including the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Mobile Computing, the IEEE Transactions on Industrial Electronics, the *ACM Transactions on Knowledge Discovery From Data*, ACM SIGKDD, IEEE International Conference on Data Mining, and Neural Information Processing Systems. He has served on the program committees of numerous conferences including SIGKDD, ICDM, and SDM. He received the Best Application Paper Award on KDD-2008, the Best Research Paper Award on ICDM-2011, and the Best of SDM-2015. His research is supported by the National Science Foundation for Distinguished Young Scholars of China.