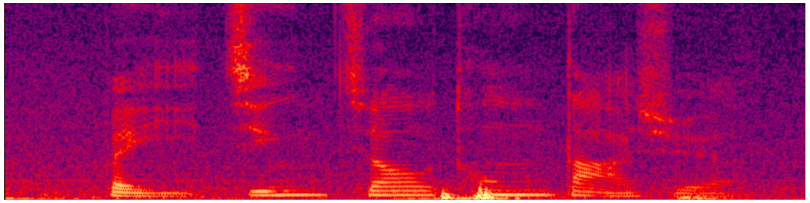
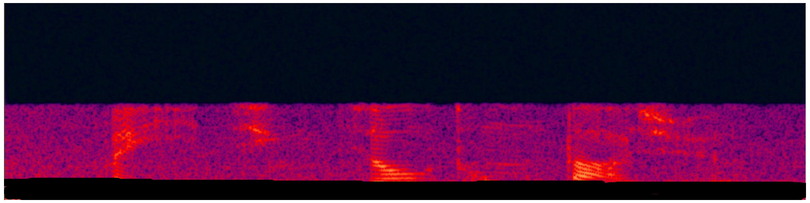
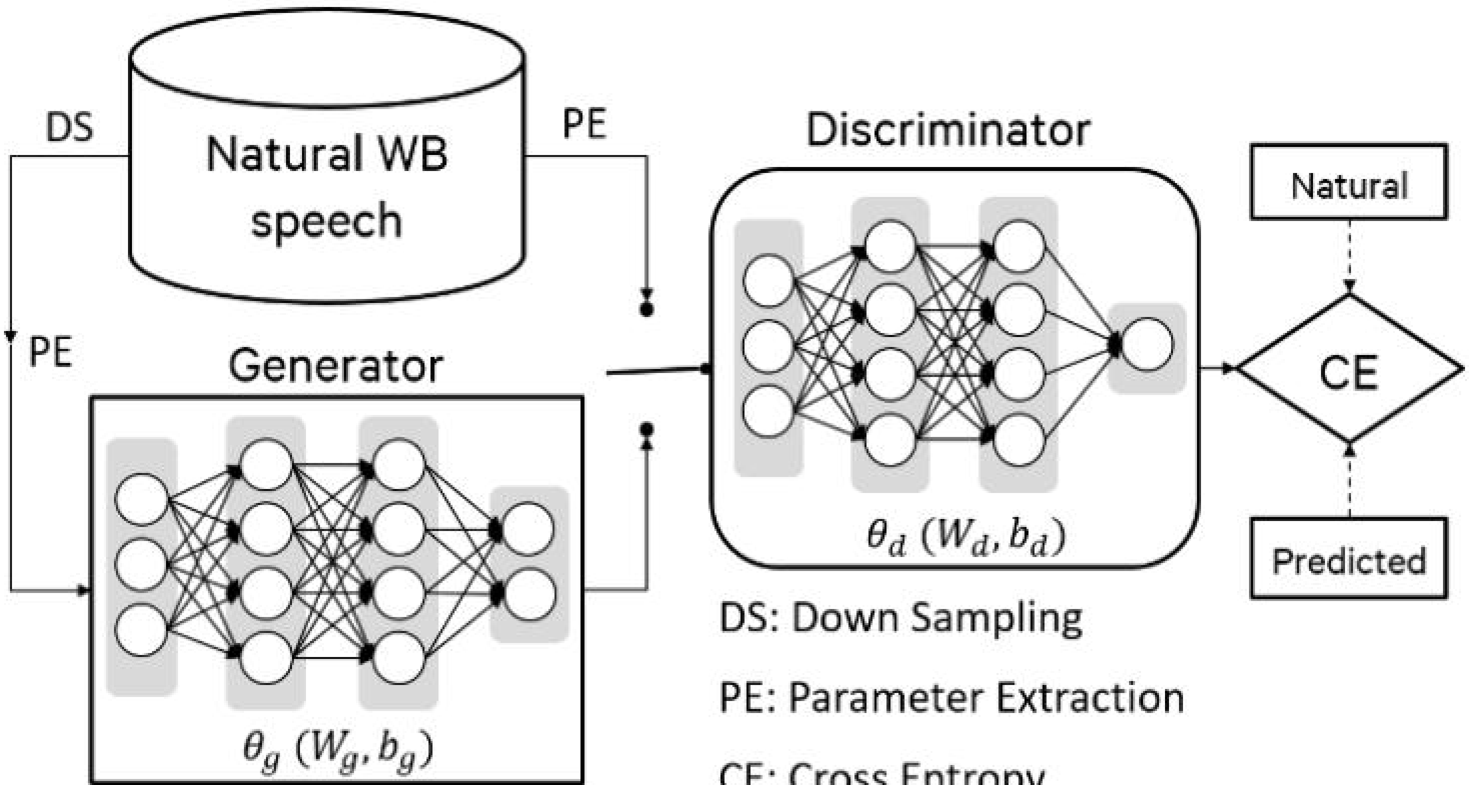
# Cross-bandwidth Train

典型带宽
(宽带：0-8kHz)



电信频带
（窄带：300hz-3.4kHz）






Speech Bandwidth Extension Using Generative Adversarial Networks , ICASSP 2018.

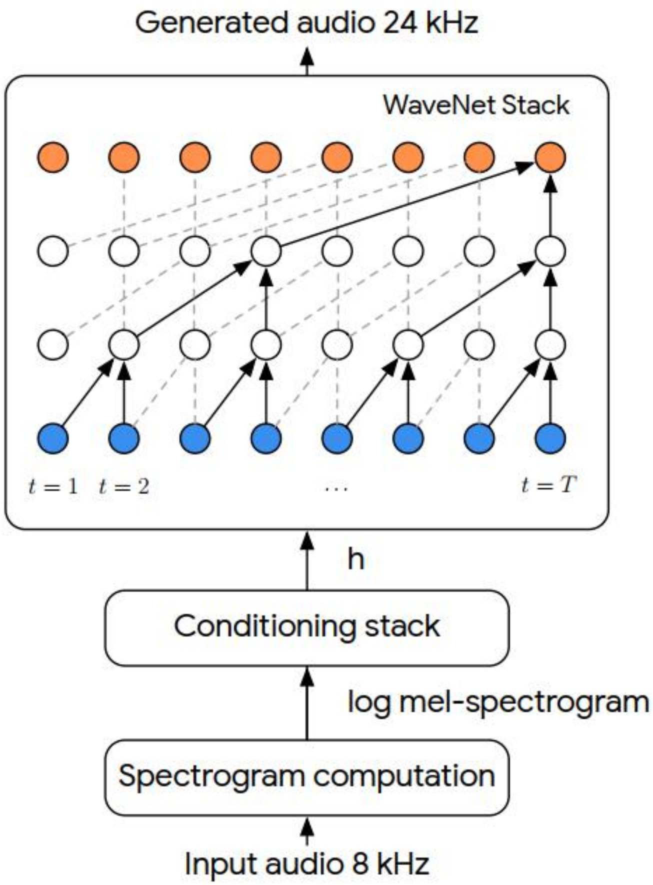



Figure 2: Illustration of the processing pipeline. The input audio, sampled at 8 kHz, is transformed to a log mel-spectrogram representation, then used as input in the conditioning stack of WaveNet. The model outputs high-sample rate 24 kHz audio with higher frequencies predicted from the rest of the signal.

$$p(\mathbf{x}_{\text{hi}}|\mathbf{x}_{\text{lo}}) = \prod_{t=1}^{T} p\left(x_{\text{hi},t}|x_{\text{hi},1}, \ldots, x_{\text{hi},t-1}, \mathbf{x}_{\text{lo}}\right).$$

where $\mathbf{x}_{\text{hi}}$ is the autoregressively modelled 24kHz waveform, and $\mathbf{x}_{\text{lo}}$ is the 8kHz band-limited version, represented as a log mel-spectrogram. The $\mathbf{x}_{\text{lo}}$ is used as input in the WaveNet conditioning stack.

Speech Bandwidth Extension Using Generative Adversarial Networks , WASPAA 2019.

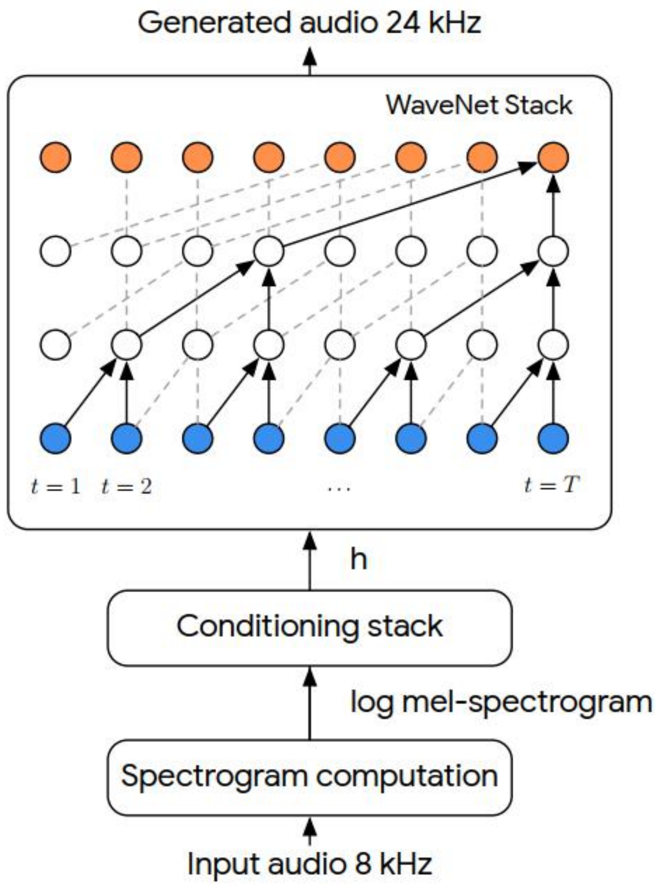



Figure 2: Illustration of the processing pipeline. The input audio, sampled at 8 kHz, is transformed to a log mel-spectrogram representation, then used as input in the conditioning stack of WaveNet. The model outputs high-sample rate 24 kHz audio with higher frequencies predicted from the rest of the signal.

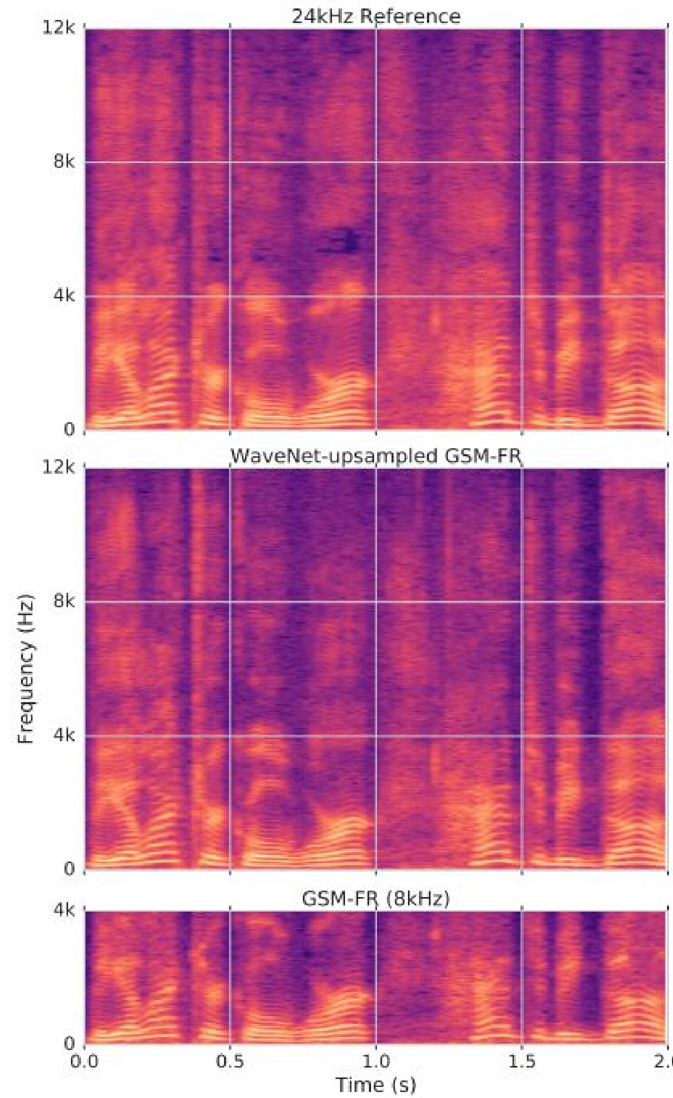



Figure 1: Spectrograms from an utterance from the LibriTTS corpus. Top: Original audio, Middle: Audio reconstructed from the WaveNet model conditioned on spectrograms derived from GSM-FR audio, Bottom: Spectrogram from GSM-FR audio.

Speech Bandwidth Extension Using Generative Adversarial Networks , WASPAA 2019.

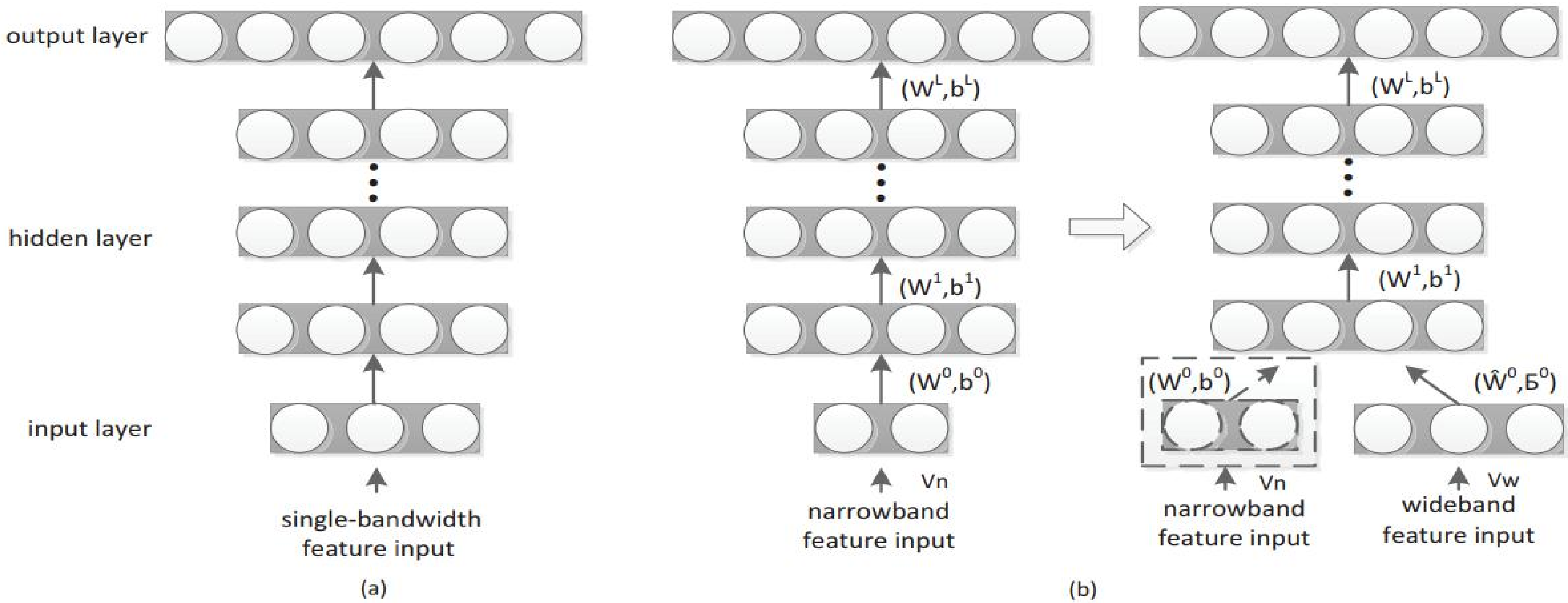



Figure 1: (a) the single bandwidth neural network and (b) the mixed-band neural network

Improving Wideband Acoustic Models Using Mixed-bandwidth Training Data Via DNN Adaptation , INTERSPEECH 2014.

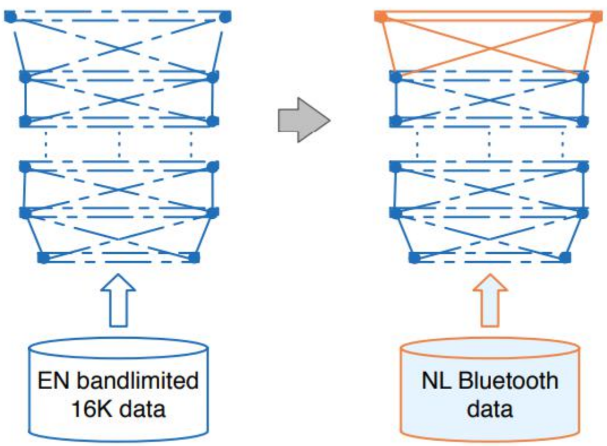



Figure 1: *Cross-lingual initialization — the hidden layers of a DNN trained on one language, e.g., English (EN), is used to initialize the DNN for a target language, e.g., Dutch (NL). The output layer is initialized with random weights and the whole network is retrained.*

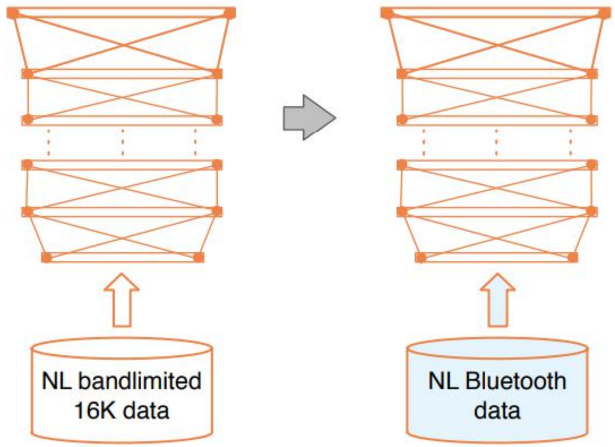
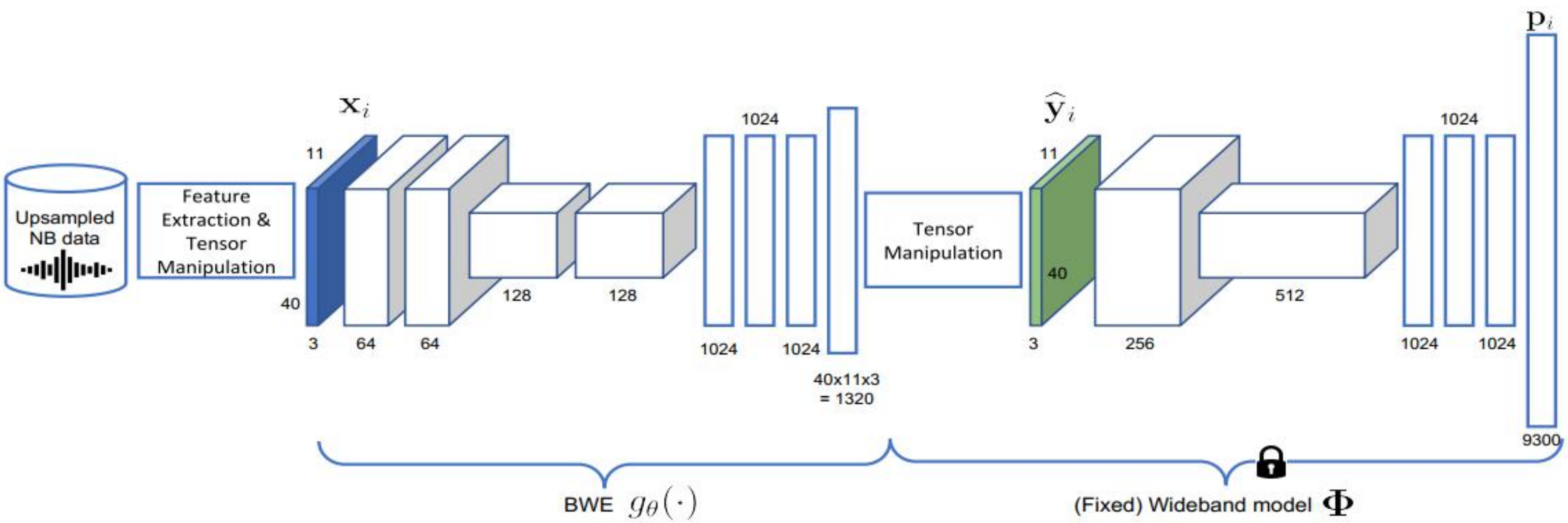
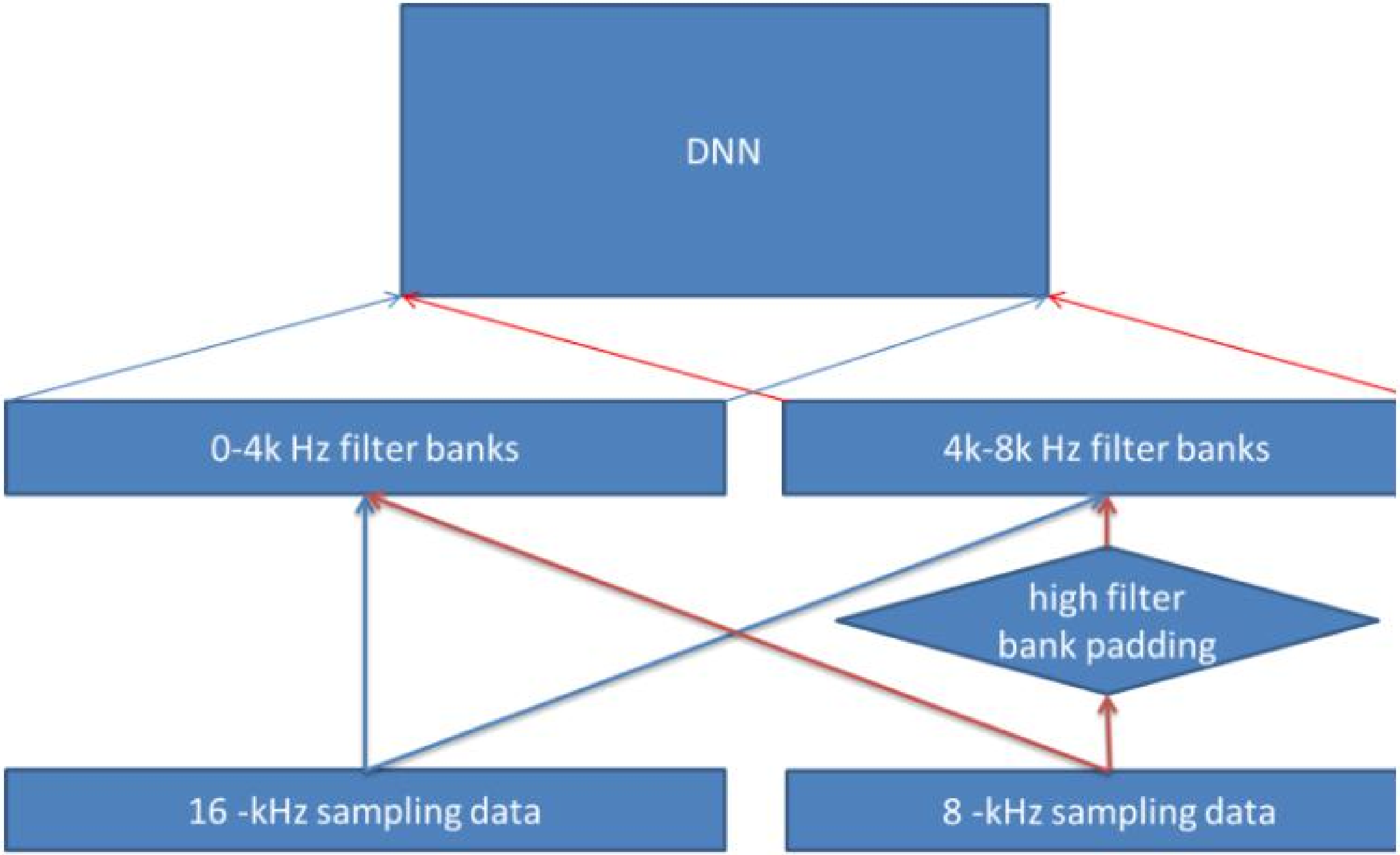



Figure 2: *Cross-bandwidth initialization — a DNN trained on bandlimited wideband audio is then further retrained on narrowband audio (using Dutch(NL) as an example language).*

Improving DNN Bluetooth Narrowband Acoustic Models by Cross-bandwidth and Cross-lingual Initialization, INTERSPEECH 2017.




Large-scale Mixed-bandwidth Deep Neural Network Acoustic Modeling For Automatic Speech Recognition , INTERSPEECH 2019.




Improving Wideband Speech Recognition Using Mixed-bandwidth Training Data In CD-DNN-HMM , SLT 2012

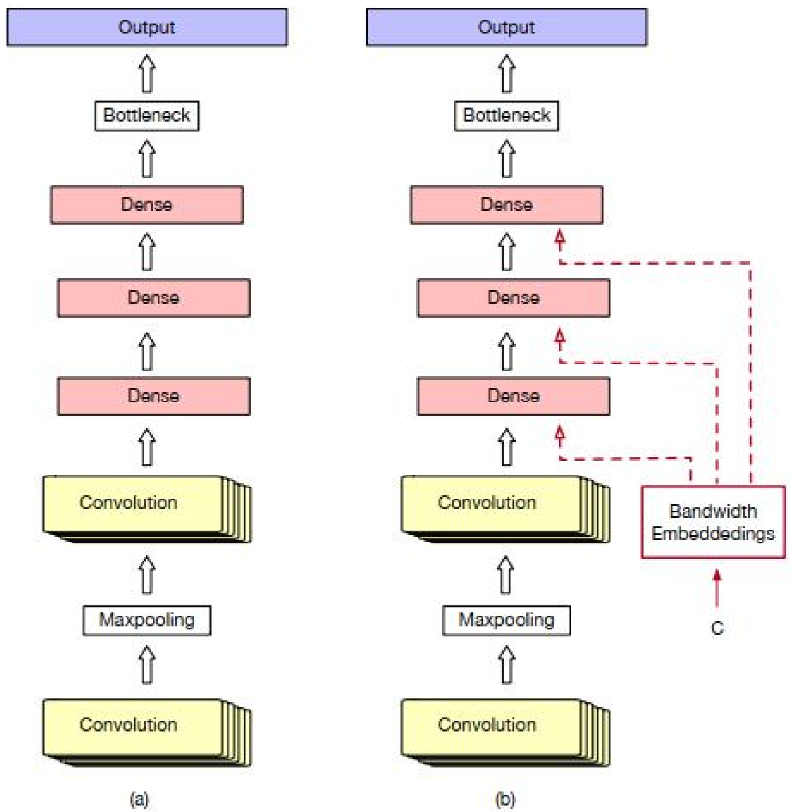
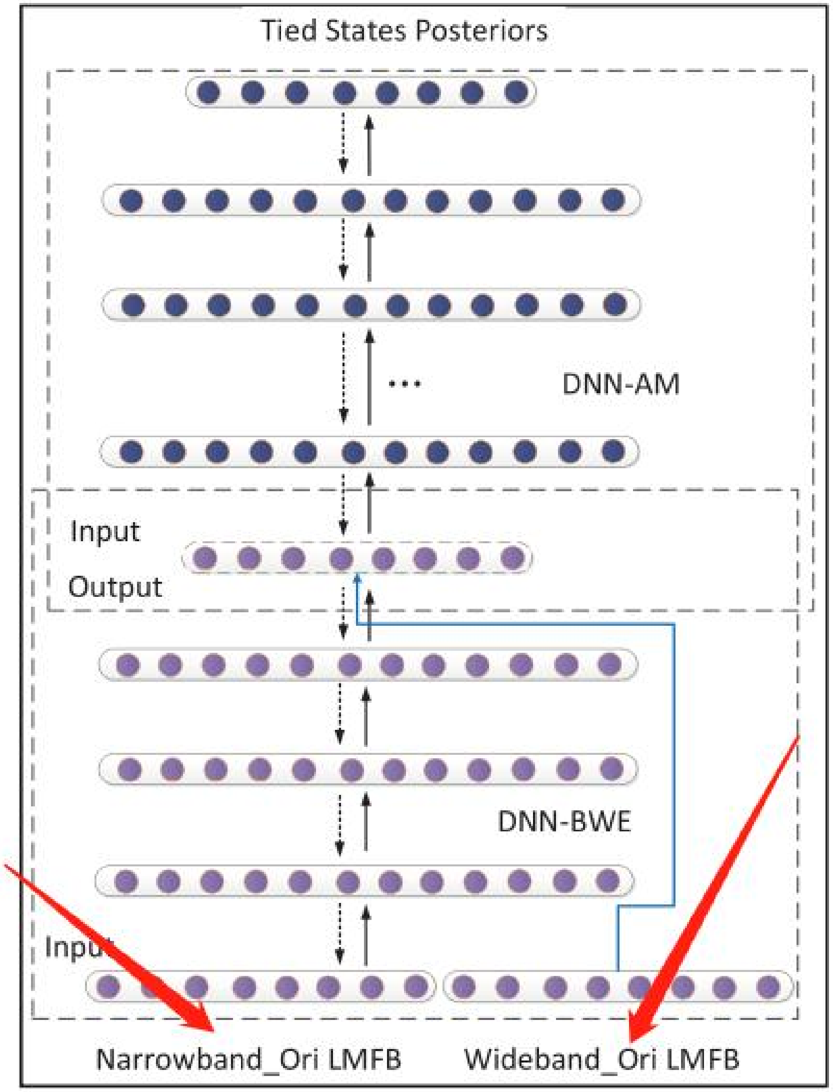



Figure 1: *(a) Baseline AM architecture containing two layers of convolution layers, 3 layers of fully connected layers, a liner bottleneck layer and then followed by an output layer, (b) Bandwidth embeddings connected to the dense layers of the baseline architecture, where c represents the type of the speech signal*

$$\begin{aligned} \mathbf{o}_l &= f(W_l \mathbf{o}_{l-1} + V_l \mathbf{e}^c + \mathbf{b}_l) \\ &= f(W_l \mathbf{o}_{l-1} + \hat{\mathbf{b}}_l), \end{aligned} \quad (2)$$

where $\hat{\mathbf{b}}_l = V_l \mathbf{e}^c + \mathbf{b}_l$. $V_l$ is the weight matrix connecting the embedding vector $\mathbf{e}^c$ to the dense layer $l$. In this paper, the bandwidth embeddings is connected to the first dense layer ($l = 3$) after two convolutional layers. $V_l \mathbf{e}^c$ is referred to as a bias correction term and thus $\hat{\mathbf{b}}_l$ can be referred to as corrected bias. This correction helps the model to differentiate and better process the narrow and wideband data. $\mathbf{e}^c$ ($c \in \{0, 1\}$) is an $n$ dimensional embedding vector and randomly initialized. During training, they are treated as model parameters and are updated during back-propogation. During decoding, the model uses the embedding vector based on the type of input speech signal and is provided by $c$.




**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.

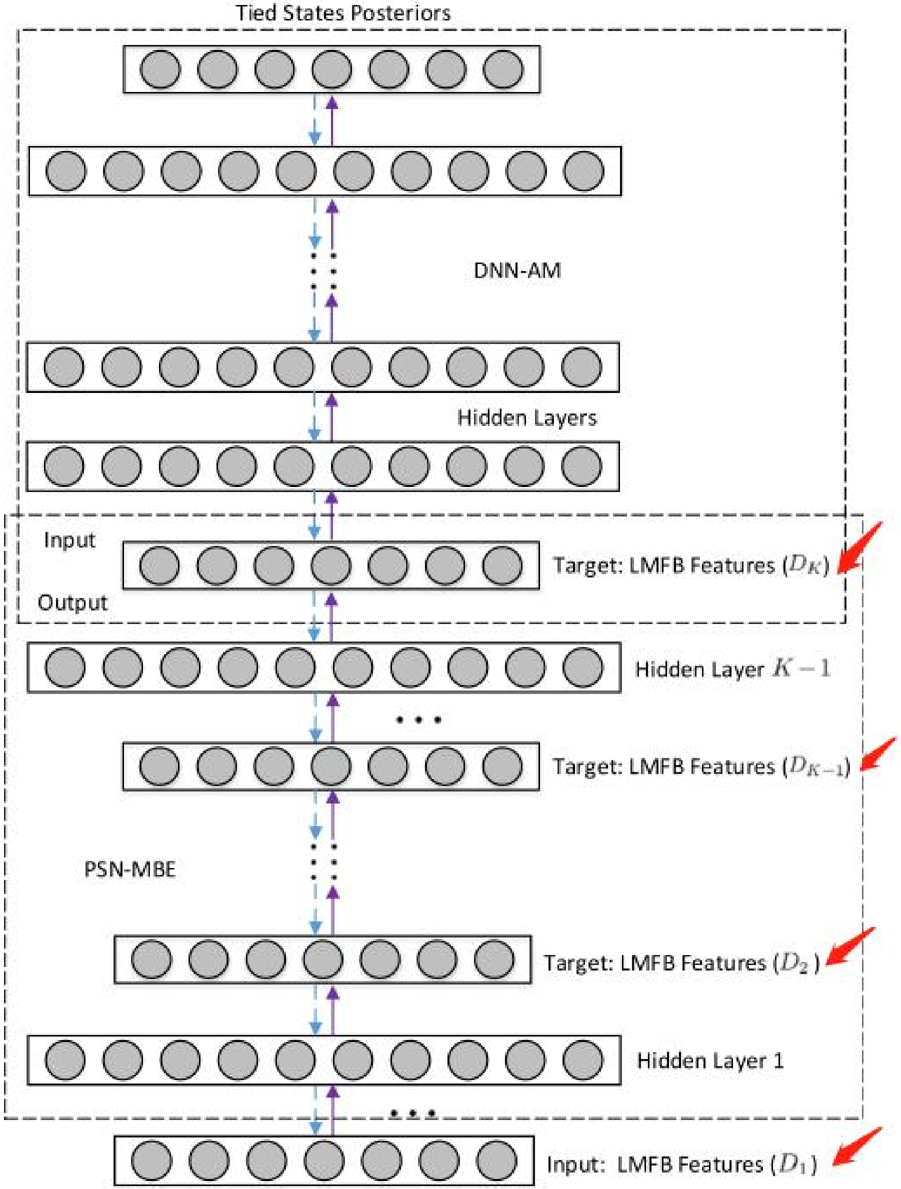
**Step3: Joint modeling**

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.

An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition , IJCNN 2016




**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.

Mixed-bandwidth cross-channel speech recognition via joint optimization of DNN-based bandwidth expansion and acoustic modeling , TASLP 2019