Improving Wideband Acoustic Models Using Mixed-bandwidth Training Data via DNN Adaptation

Zhao You, Bo Xu

Interactive Digital Media Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
{zhao.you, xubo}@ia.ac.cn

Abstract

In the past few years, deep neural networks (DNNs) have achieved great successes in speech recognition. The deep network model can be viewed as a series of feature transform s followed by a log-linear classifier. For input of speeches from different bandwidths, although the hidden layer transform and log-linear classification can be shared, the input layer transforms should be specially designed respectively. So, training DNNs directly on different bandwidth speeches is intractable. In this paper, we treat the problem of training DNNs on mixed bandwidth data as an domain- adaptation problem. Upon our adaptation approach, DNNs trained on the rich narrowband speech can be adapted effectively to the target wideband domain, and meanwhile shows good performance on the wideband speech. We evaluate this approach on the wideband clean7k and noise360 speech. Experimental results show that the DNNs adaptation approach can reduce character error rate (CER) range from 5% to 15%, relatively, over the baseline DNNs trained only on the limited wideband data.

Index Terms: DNN adaptation, CD-DNN-HMM, mixed-band DNN, transfer learning

1. Introduction

In the past few years, deep neural networks (DNNs) were introduced to speech recognition tasks and gained great successes. Specially, DNN-HMM acoustic models achieved significant recognition error reduction over discriminatively trained GMM-HMM models [1, 2, 3].

The automatic speech recognition (ASR) systems can perform very well when sufficient training data is provided. Obviously, collecting narrowband speech (8-kHz) is easier than wideband speech (16-kHz), since there is an efficient way to collect large amounts of data through recording speech over the telephone. Therefore, training wideband acoustic models using a little wideband speech while a large scale of narrowband speech has vital practical significance. Naturally, the aim of training this mixed-band model is to improve wideband speech recognition using sufficient narrowband speech.

Obviously, an simple approach is to down sample the wideband training data and testing data. However, this method is suboptimal since wideband speech drops some useful information during the sampling process. Recently, several approaches [4, 5, 6] have been proposed to train wideband acoustic models on mixed-bandwidth data. Typically, Michael [4] introduced the EM algorithm to train mixed-band models in the GMM-HMM based system, and further incorporated it into the exiting training schemes of HMM speech recognizers. However, this procedure is quite complicated and have no obvious gains.

An alternative approach is to use the CD-DNN-HMM framework. Exploiting mixed-bandwidth training data in the CD-DNN-HMMs framework [5] requires much simpler training procedure than the feature bandwidth extension method [4]. In this framework, the feature of narrow speech can be considered as wideband speech with some feature dimensions missing.

In this paper, we present the DNN adaptation for training wideband acoustic models using mixed-bandwidth training data. Considering the differences between the filter features of narrowband speech and wideband speech from differing mel frequency filter locations and filter number, we adopt the Mel-scale filter banks used in [4]. Concretely, we first trained the original deep network with narrowband speech and then adapt it to the target domain by training on the wideband speech. We evaluated the effectiveness of our approach on the two test sets and demonstrated that the DNN adaptation can largely reduce character error rate over the baseline DNNs trained using only the target speech.

The rest of the article is organized as follows. We will first briefly describe two conventional methods for training wideband acoustic models using mixed-bandwidth data in Section 2. In Section 3 we give a formulated introduction on how to train wideband acoustic models with mixed-bandwidth training data via DNN adaptation. We report the experimental results in Section 4 and conclude the paper in Section 5.

2. Review of improving wideband speech recognition using mixed-band training data

In this section, we briefly describe two conventional methods for training wideband acoustic models using mixed-bandwidth data.

2.1. Feature bandwidth extension

In [4] Michael proposed an alternative approach in which the narrowband features are converted to wideband features using the feature bandwidth extension (FBE). For the wideband speech, the log mel spectrum vector \( x \) of a speech frame is converted to a cepstral vector \( y \) using discrete-cosine transform (DCT)

\[ y = Ax \]  

where \( A \) is the DCT matrix. And the corresponding log mel spectrum vector \( x \) spans 0-8kHz. While for the narrowband speech, the log spectrum vector spans only 0-4kHz, and the residual 4-8kHz is missing. In order to distinguish the original
components from the missing components, the log mel spectrum vector $x$ can be expressed as

$$x = [x^0 \ x^m]$$ (2)

where $x^0$ denotes the original components and $x^m$ denotes the missing components. For the narrowband speech, the missing subvector corresponds to the high frequency components (4-8kHz). For wideband speech, $x^0 = x$ and $x^m = [\,]$. Analogously, the cepstral vector $y$ can be partitioned into two sections

$$y = y^0 + y^m$$ (3)

In the training stage, the bandwidth-extended features are combined with the wideband speech features to train the acoustic models using a modified forward-backward algorithm.

2.2. CD-DNN-HMMs mixed-bandwidth training

Comparing with the bandwidth extension procedure, training CD-DNN-HMMs with mixed-bandwidth data [5] is much simpler. The high frequency feature of narrowband speech is considered missed when training CD-DNN-HMMs with wideband data. Specially, the missed feature dimensions of narrowband speech is padded with zeros. The key question of this CD-DNN-HMM framework is how to design the feature of different bandwidths data. The Mel-scale log filter-bank feature is considered missed when training CD-DNN-HMMs with mixed-bandwidth data [5]. 22 filter banks are used for narrowband speech and 29 filter banks are used for wideband speech. Specially, the 7 upper filter banks for the narrowband speech are padded with 0s. This treatment makes the training process with mixed-bandwidth data set much simpler since it does not involve the bandwidth extension.

3. Exploiting DNN adaptation for training mixed-bandwidth data

In this section, we describe an adaptation method for training context-dependent deep-neural-network Hidden Markov Models using mixed-bandwidth data. Instead of training DNNs solely on the target speech, we first trained the original deep networks with narrowband speech, in order to provide a proper initial state for the DNN. Then, to capture the data property in the wideband speech, we adapt the pretrained DNNs to the target domain by training on the wideband speech. The training procedure is presented concretely as follows.

We denote the input vector of narrowband speech as $v_n$, the input vector of wideband speech as $v_w$ and the hidden layer at layer $l$ ($l = 0, 1, 2, ..., L$) as $h^l$. Here we set $b^0 = v_n$. The input layer is transformed into a feature vector through non-linear transform:

$$h^1 = \sigma(W^0 v_n + b^0)$$

$$h^l = \sigma(W^{l-1} h^{l-1} + b^{l-1}) \quad (l = 2, 3, ..., L)$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function. $W^l$ and $b^l$ are the weight matrix and bias of layer $l$, respectively. Specially, $(W^0, b^0)$ denote the parameters between the input vector of narrowband speech and first hidden layer. The posterior probability of context-dependent phone state $s$ is estimated using softmax function as

$$P(s|h^L) = \frac{\exp(W_L h^L + b_L)}{\sum_s \exp(W_L h^L + b_s^L)}$$ (6)

where $s = 1, 2, 3, ..., S$ is a senone id. $S$ is the number of the senones. Here $h^L$ is equivalent to $o$. The HMM state emission probability is estimated as

$$P(o|s) = \frac{P(s|o)P(o)}{P(s)}$$ (7)

The parameters $(W^l, b^l)$ of DNNs are typically updated to maximize the negative cross entropy, i.e.

$$L = \frac{1}{N} \sum_{t=1}^{N} L(o(t)) = \frac{1}{N} \sum_{t=1}^{N} \sum_{s=1}^{S} P_T(s|o(t)) \log P(s|o(t))$$ (8)

where $N$ is the total number of the training samples and $P_T(s|o(t))$ is the target probability. Stochastic gradient ascent is used to update the parameters of DNNs, i.e.

$$W^l_{new} = W^l_{old} + \epsilon_w \frac{\partial \log L}{\partial W^l_{old}} \quad (l = 0, 1, 2, ..., L)$$

$$b^l_{new} = b^l_{old} + \epsilon_b \frac{\partial \log L}{\partial b^l_{old}} \quad (l = 0, 1, 2, ..., L)$$

where $\epsilon_w$ and $\epsilon_b$ are learning rate.

After the DNN is trained using narrowband speech, we use a modified neural network in which the input layer is replaced by the wideband speech as shown in Fig.1(b). The weights between the input and first hidden layer are randomly initialized, i.e.

$$h^1 = \sigma(W^0 v_w + b^0)$$

(11)

where $(W^0, b^0)$ denote the parameters between the input vector of wideband speech and first hidden layer. The parameters of other layers keep fixed. Once the new DNN has been initialized, we update all the parameters on wideband speech via back propagation (BP).

4. Experiments

4.1. Experimental setup

4.1.1. Data

To evaluate the efficiency of our approach, we perform a series of experiments using the 130-hour narrowband speech (telephone speech) and 30-hour wideband speech. The feature extraction is similar. Speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. The speech feature vector was generated by a Fourier-transform-based filter bank. For the narrowband speech, we use 22 filter banks which spans 0-4 kHz. For the wideband speech, we use 29 filter banks which spans 0-8 kHz. The systems use the filter bank features appended with the first and second order derivatives. Mean and variance normalization is performed on per utterance case. The systems are evaluated on two individual test sets of wideband speech, namely clean7k and noise360, which are collected through mobile microphone under clean and noise environments, respectively. We hold out about 5 hours wideband speech development set for cross-entropy (CE) loss evaluation.

4.1.2. Training tools

All the deep networks in this paper are trained using the modified Kaldi toolkit [7]. For fast training, we use an effective parallel approach-asynchronous stochastic gradient descent (ASGD) based our prior work [8]. Regarding hardware, the NVIDIA GeForce GTX 690 cards are used for pre-training and fine-tuning.
4.1.3. Deep neural network

Before training DNN, we first train a GMM-HMM system with maximum likelihood (ML) and boosted maximum mutual information (BMMI) [9] criteria. The GMM system is served as the label generator at the frame level. The deep networks used for our experiments have five hidden layers each containing 2048 hidden units and an output layer with 10217 senones. In DNN-HMM, we use long concatenated feature vectors, stacking from all consecutive frames within a context window (5+1+5), as DNN’s inputs. In BP, the iteration stops if the development set frame accuracy increment is smaller than a small constant $\lambda$ (we set $\lambda = 0.05\%$). Note that we use different minibatch size during the training iterations. For the first iteration, we use very small minibatch (64) for training DNN with partial (about 10%) training dataset. For the next two iterations, large minibatch (256) is used. For the subsequent iterations, we tend to use larger minibatch (1024).

4.2. Baseline systems

There are three baselines. One is that using only the wideband (WB) training data (marked as B1 in the table). In this base we throw away the narrowband (NB) training data. The structure of single bandwidth DNN is showed in the Fig.1(a). In the second baseline, we down sample both the wideband training data and the testing data so that the wideband speech is treated as the narrowband speech. This baseline is marked as B2 in the table. In the third baseline, we refer to the experimental configuration that is described in [5]. The filter bank design of narrow and wide bandwidth data is described in section 4.1.1. Additionally, the 7 upper filter banks (spans 4-8kHz) are padded with 0s for the narrow bandwidth data. This baseline is marked as B3 in the table. For all the experiments, we used different sizes of wideband training data (5, 10 and 30 hours) to evaluate the recognition performance.

4.3. Training Mixed-band DNN with DNN adaptation

In this subsection, we introduce the training approach for the DNN systems on narrowband and wideband speech. The network is trained using the following steps.

1. Train the DNN on 130 hours narrowband speech. For the DNN architecture, the input layer uses 22 Mel-scale log filter banks. The input layer of DNN has $22 \times 3 \times 11 = 726$ units. We firstly initialized DNNs with stacked RBMs [10] that are pre-trained on the narrowband speech. The deep network is then trained discriminatively via BP.

2. Adapt the network to wideband speech. For the wideband speech, the phone set, HMM topology, and context-dependent states are the same as the narrowband speech setting. Incompatible filter number is a challenge for adapting the deep network using mix-bandwidth training data. We replace the first layer of the original deeper network by the wideband speech and train the network on the wideband speech. The dotted line in the Fig.1(b) expresses that the first layer of original deeper network is replaced by the wideband speech input. We initialize the weights of the first layer using random weights, while keeping the other layers fixed. Finally, we update all the parameters using the wideband speech via BP algorithm.
5. Conclusions

In this paper, we showed a simple and effective approach for training acoustic models based CD-DNN-HMM framework using mixed-band training data. Our approach is based on the observation that DNN has the flexibility of reducing the mismatch in utterances from different bandwidths. We thus can formulate and reduce the mixed-bandwidth training problem into a domain adaptation problem.

In our approach, the limited amount of wideband training data is augmented with narrowband training data in order to train a speech recognizer for the recognition of wideband speech. Through a series of experiments using wideband and telephone speech, we demonstrated that our method significantly outperforms a wideband recognizer trained with a limited wideband data. It suggests the possibility to quickly build a high-performance CD-DNN-HMM system for wideband speech from an existing narrowband DNN.

6. RELATION TO PRIOR WORK

The motivation of this study is to use knowledge learned from narrowband speech to improve the performance of wideband speech. Most of previous works generally focus on the learning algorithms of FBE based on GMM framework [4, 6]. However, the modeling power of Gaussian mixtures is significantly less than that of DNNs and so the features learned from the GMMS would be much less selective and invariant than that from the DNNs. A similar work on mixed-bandwidth training is described in [5], which uses narrowband speech downsampled from wideband speech. Experiments operate on rich wideband speech. This paper focus on the learning algorithm of DNN adaptation based on DNN framework. In our work, we directly use the narrowband speech collected from the telephone and consider the wideband speech with limited resource. Moreover, we use a different mixband neural network structure.

7. Acknowledgement

This work is partly supported by National Program on Key Basic Research Project (973 Program) under Grant 2013CB329302 and National Natural Science Foundation of China grant 61103152.

8. References


