ISSN 1000-0054 清华大学学报（自然科学版） 2018 年 第 58 卷 第 1 期 9/17
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2018, Vol. 58, No. 1 55-60

基于迁移学习的噪声鲁棒语音识别声学建模

易江燕[1,2]， 陶建华[1,2,3]， 刘 斌[1]， 温正棋[1]

（1. 中国科学院 自动化研究所，模式识别国家重点实验室，北京 100190；2. 中国科学院大学 人工智能技术学院，北京 100190；
3. 中国科学院 自动化研究所，中国科学院脑科学与智能技术研究中心，北京 100190）

摘 要：为了提高噪声环境下语音识别系统的鲁棒性，提出了一种基于迁移学习的声学建模方法。该方法用干净语音的声学模型（老师模型）指导带噪语音的声学模型（学生模型）进行训练。学生模型在训练过程中，尽量使其逼近老师模型的后验概率分布。学生模型和老师模型间的后验概率分布差异通过相对熵（KL divergence）加以最小化。CHiME-2 数据集上的实验结果表明，该方法的平均词错率（WER）比基线的绝对下降了 7.29%，比 CHiME-2 竞赛第一名的绝对下降了 3.92%。

关键词：鲁棒语音识别；声学模型；神经网络；迁移学习

中图分类号：TP391.42；TP183 **文献标志码**：A

文章编号：1000-0054(2018)01-0055-06

DOI：10.16511/j.cnki.qhdxxb.2018.21.001

Transfer learning for acoustic modeling of noise robust speech recognition

YI Jiangyan[1,2], TAO Jianhua[1,2,3], LIU Bin[1], WEN Zhengqi[1]

（1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China;
3. CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China）

Abstract: Speech recognition in noisy environments was improved by using transfer learning to train acoustic models. The training of an acoustic model trained with noisy data (student model) is guided by an acoustic model trained with clean data (teacher model). This training process forces the posterior probability distribution of the student model to be close to the teacher model by minimizing the Kullback-Leibler (KL) divergence between the posterior probability distribution of the student model and that of the teacher model. Tests on the CHiME-2 dataset show that this method gives a 7.29% absolute average word error rate (WER) improvement over the baseline model and 3.92% absolute average WER improvement over the best CHiME-2 system.

Key words: robust speech recognition; acoustic model; deep neural network; transfer learning

近年来，深度学习已成为语音识别领域的主流技术[1-3]，基于深度神经网络的声学模型取得了突破性的进展[4-5]。目前，虽有大量语音识别系统走向市场，然而受到真实环境噪声的干扰，系统的识别准确率急剧下降。噪声是影响语音识别技术广泛实用化的一个关键因素。因此，针对噪声环境下语音识别系统的鲁棒性研究在理论和实践 2 个层面上具有重要意义。

迄今为止，已有不少旨在提高语音识别系统的环境鲁棒性的方法。这些方法大致可以概括为 3 类：1）采用自适应算法训练鲁棒声学模型；2）直接利用带噪语音数据训练声学模型；3）先对带噪语音数据进行增强处理，然后利用处理后的数据训练声学模型。

第 1 类方法主要是从特征补偿和模型补偿的层面对声学模型进行自适应。例如，蔡尚等[6]基于子带能量规整感知线性预测系数的方法对特征进行补偿；胡旭琰等[7]指出基于缺失数据技术特征补偿的方法能提高声学模型的环境鲁棒性；Gales 等[8]利用最大似然线性回归进行噪声环境自适应；Siohan 等[9]提出最大后验线性回归的模型自适应算法；Tran 等[10]基于分解的线性输入网络（LIN）对深度神经网络（DNN）的声学模型进行噪声环境自适应。

在第 2 类方法中，一般直接利用带噪语音和干净语音训练 DNN 声学模型。Seltzer 等[11]指出，与基于语音增强和模型补偿技术的 Gauss 混合模型-

收稿日期：2017-09-29

基金项目：国家"八六三"高技术项目（2015AA016305）；
国家自然科学基金面上项目（61425017，61403386）；
中国科学院战略性先导科技专项（GrantXDB02080006）

作者简介：易江燕（1984—），女，博士研究生。

通信作者：陶建华，教授，E-mail：jhtao@nlpr.ia.ac.cn

隐 Markov 模型(GMM-HMM)声学模型相比,DNN 声学模型性能更佳。这是因为 DNN 能学习到更高层次的特征表达,从而更易于捕捉到带噪语音的不变性[12]。Li 等[13]通过对特征进行扩帧,然后将其作为输入训练也能提高识别系统的鲁棒性。王青等[14]指出将各种融合的特征作为输入训练 DNN 亦能有效降低识别词错率(WER)。

第 3 类方法则是将用于 GMM-HMM 的降噪技术应用到了 DNN 声学模型中。Abe 等[15]提出了将经典谱减法和噪声估计相结合的方法。该方法先用谱减法对语音特征进行处理,再将处理后的特征和已估计噪声参数作为 DNN 的输入来进行训练。该方法被称为噪声依赖(noise-aware)训练,最早由 Xu 等[16]提出。Vincent 等[17]利用基于深度降噪自编码(DAE)模型建立带噪语音和干净语音特征间的映射关系,亦有学者提出了将噪声依赖训练和 DAE 相结合的方法[18],以及利用 DAE 和 DNN 声学模型进行多目标联合训练[19]。

上述 3 类方法固然可以有效地提高语音识别系统的噪声鲁棒性,但不少工作是在干净语音和带噪语音的平行数据已知的前提下进行的[20],且在方法上或是将干净语音直接作为训练数据,或是将其作为降噪处理的参考标准,并未最大限度地挖掘干净语音的知识。有鉴于此,本文拟从干净语音中提取知识,提高带噪语音的识别准确率。

"知识提取"的概念最早由 Bucila 等[21]提出。最近,Hinton 等[22]提出了一个更为通用的框架,即利用多个复杂神经网络模型构成的组合模型指导一个简单的神经网络模型进行训练。前者称为老师模型,后者称为学生模型。学生模型在训练的过程中,模仿老师模型的后验概率分布。在语音识别领域,不少工作正是利用这种知识提取的方法对声学模型进行压缩或简化。例如,以 Li[23]为代表的学者提出利用一个庞大复杂的 DNN 声学模型指导一个小型简单的 DNN 声学模型进行训练,从而达到模型压缩的目的。Chan 等[24]则建议利用循环神经网络声学模型指导一个 DNN 声学模型进行训练。Chebotar 等[25]亦利用多个神经网络的组合声学模型训练一个简单的声学模型。上述方法的核心思想便是采用相对熵(KL divergence)来最小化简单声学模型与复杂声学模型之间后验概率分布的差异。就本质而言,这些方法都属于迁移学习的范畴。

受到上述方法的启发,本文提出利用迁移学习的方法对带噪语音进行声学建模,以提高语音识别系统的环境鲁棒性。具体而言,本文拟将干净语音训练而得的声学模型作为"老师模型",将带噪语音训练得到的声学模型作为"学生模型"。学生模型在训练的过程中,模仿老师模型的后验概率分布。二者之间后验概率分布的差异采用相对熵来最小化。与此同时,本文亦就不同神经网络结构的老师模型对学生模型 WER 的影响等问题进行了讨论,并在 CHiME-2[26]数据集上进行了实验。

1 基于迁移学习的声学建模

1.1 迁移学习

本文中的迁移学习是指将老师模型的后验概率分布知识迁移到学生模型的训练过程中;也即学生模型在训练的过程中,尽量逼近老师模型的后验概率分布,模仿老师的行为。二者之间后验概率分布的差异用相对熵来最小化。

假设 P_c 代表老师模型的后验概率分布,Q 代表学生模型的后验概率分布,那么二者之间的后验概率分布差异可以表示为

$$D_{\mathrm{KL}}(P_c \| Q) = \sum_i P_c(s_i \mid x_c) \ln \frac{P_c(s_i \mid x_c)}{Q(s_i \mid x)}. \tag{1}$$

在学生模型训练的过程中,希望最小化式(1),可以表示为

$$D_{\mathrm{KL}}(P_c \| Q) = H(P_c, Q) - H(P_c). \tag{2}$$

其中:

$$H(P_c, Q) = \sum_i -P_c(s_i \mid x_c) \ln Q(s_i \mid x), \tag{3}$$

$$H(P_c) = \sum_i -P_c(s_i \mid x_c) \ln P_c(s_i \mid x_c). \tag{4}$$

其中:i 表示为三因子状态(senone)的下标;s_i 为第 i 个三因子状态;x_c 表示干净语音的特征;x 表示带噪语音的特征;$P_c(s_i \mid x_c)$ 表示特征 x_c 被识别为第 i 个三因子状态的后验概率,该后验概率由老师模型采用前向算法计算得到;$Q(s_i \mid x)$ 表示特征 x 被识别为第 i 个三因子状态的后验概率。然而,式(4)只与老师模型的后验概率分布有关,而与学生模型的后验概率分布无关,因此可以忽略,由此可得

$$D_{\mathrm{KL}}(P_c \| Q) \equiv \sum_i -P_c(s_i \mid x_c) \ln Q(s_i \mid x). \tag{5}$$

可以看出,求式(5)的最小值也即求交叉熵(CE)的最小值,其优化过程等同于标准交叉熵的训练过程。式(5)与标准交叉熵唯一不同的是训练所需的分类标签。标准交叉熵训练准则中的分类标签是三因子状态的由 0、1 构成的向量(硬标签),而式(5)中的分类标签为老师模型计算所得的后验概率分布(软标签)。因此,对式(5)进行优化时,只需

将标准交叉熵准则中的硬标签替换为软标签即可。这一过程是对 Hinton 等[22]提出的基于温度(temperature)知识提取方法的一种简化。

1.2 声学建模

就本文所提方法而言，老师模型和学生模型均为基于 HMM 和神经网络的混合模型。GMM-HMM 用于生成强制对齐信息(硬标签)，而神经网络则用于为给定输入特征预测其对应三因子状态的后验概率。老师模型指导学生模型的训练流程主要包括 4 个步骤：生成硬标签、训练老师模型、生成软标签和训练学生模型。具体训练流程如图 1 所示。

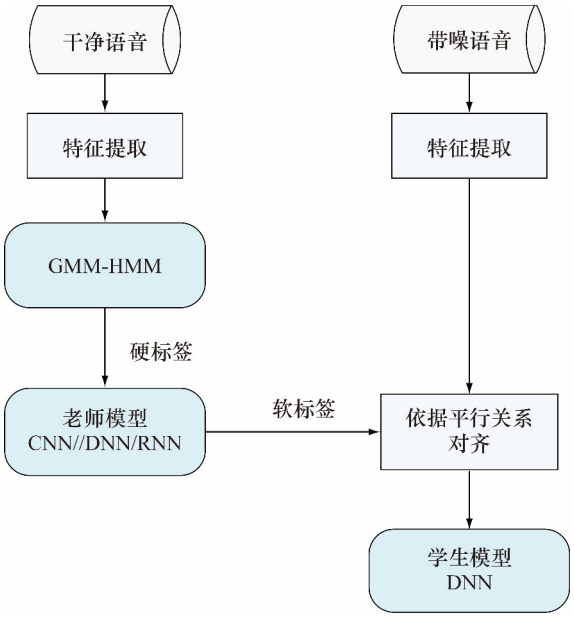

图 1 老师模型指导学生模型的训练流程

生成硬标签时，本文仅用干净语音提取特征，训练一个 GMM-HMM 模型，然后通过帧级别的强制对齐得到每帧数据的硬标签 t_{hard}。硬标签为由 0、1 构成的向量，比如某一帧数据的硬标签为[0 0 1 0 0 0]，此向量代表该帧属于标签 3 的概率为 1，属于其他标签的概率均为 0。

生成软标签时，本文用干净语音特征 x_c 作为老师模型的输入，利用前向算法计算其后验概率分布(软标签 t_{soft})。假设[0.02 0.1 0.83 0.03 0.01 0.01]为某一帧数据的软标签，此向量代表属于标签 1 的概率为 0.02，属于标签 2 的概率为 0.1，其他以此类推。

在训练学生模型阶段，其神经网络结构仅为 DNN。本文首先利用干净语音和带噪语音的平行关系，将带噪语音提取的特征 x 和上述干净语音的软标签 t_{soft} 进行对齐，得到带噪语音特征 x 的软标签。而后，利用该特征 x 和其软标签训练学生模型，学生模型的优化准则为式(5)。在学生模型训练的过程中，老师模型的参数保持不变，仅更新学生模型的参数。

在语音识别系统的测试阶段，本文仅用学生模型计算噪声数据的后验概率。此后验概率与先验概率结合得到似然值，该似然值即为标准解码器的声学似然。

2 实 验

2.1 实验数据

本文采用 CHiME-2[26]数据集进行实验，该数据集是带噪语音识别鲁棒性研究方面较为流行的数据集，它包含干净语音和带噪语音的平行数据，数据的采样率均为 16 kHz。干净语音数据集来自华尔街日报(WSJ0)，其词汇量为 5 000。带噪语音数据基于 WSJ0 数据集，随机叠加各种背景噪声生成。背景噪声的信噪比取值为 6 种：−6、−3、0、3、6、9 dB。干净语音和带噪语音各包含 3 个数据集：训练集、开发集和测试集。就干净语音而言，训练集包含 84 个说话人，共 7 138 句；开发集包含 10 个说话人，共 1 206 句；测试集包含 8 个说话人，共 330 句。就带噪语音而言，训练集包含 84 个说话人，共 7 138 句，噪声信噪比为−6 到 9 dB 之间的 6 种；开发集包含 10 个说话人和 6 种信噪比，共 2 460 句；测试集包含 8 个说话人和 6 种信噪比，共 1 980 句。

2.2 实验设置

本文在语音识别工具 Kaldi[27]的基础上进行开发和实验。实验共采用两种特征：mel 频率倒谱系数(MFCC)和 mel 标度滤波器组特征(FBANK)。提取特征的窗长为 25 ms，帧移为 10 ms。MFCC 特征为 13 维，加上其一阶和二阶差分统计量，共 39 维。FBANK 特征为 40 维，加上其一阶和二阶差分统计量，共 120 维。特征的均值方差归一化以说话人为单位进行。所有 GMM-HMM 的输入为 MFCC，所有神经网络模型的输入为 FBANK。

就本文所涉神经网络模型而言，其损失函数为交叉熵，优化准则为随机梯度下降(SGD)。DNN 和 CNN 模型采用反向传播(BP)算法进行训练。BLSTM 模型采用随时间反向传播(BPTT)算法进行训练。LSTM 模型采用截断的随时间反向传播(truncated BPTT)算法进行训练。本文实验所用语言模型为 WSJ0 提供的三元文法语言模型(lm_tgpr_5k)，词表大小为 5 000。解码的搜索空间基于加权有限状态转换器(WFST)进行构建，搜索

策略为束搜索(beam-search)算法。

2.3 基线模型

就带噪语音而言,GMM-HMM 模型的训练主要有 3 种方法:1) 用干净语音和带噪语音训练 GMM-HMM,表示为 NC-GMM;2) 仅用带噪语音训练 GMM-HMM,表示为 N-GMM;3) 仅用干净语音训练 GMM-HMM,表示为 C-GMM。所有 GMM-HMM 的 Gauss 模型数为 15 000,叶子节点数为 2 500。NC-GMM 的三因子状态数目为2 032,N-GMM 的三因子状态数目为 1 978,C-GMM 的三因子状态数目为 1 985。

由此,本文根据上述 3 种模型,用带噪语音训练 3 个 DNN 声学模型:NC-DNN、N-DNN 和 C-DNN。对于 NC-DNN,首先用 NC-GMM 生成带噪语音的硬标签,然后训练 DNN 模型。对于 N-DNN,首先用 N-GMM 生成带噪语音的硬标签,再训练 DNN 模型。对于 C-DNN,首先用 C-GMM 生成干净语音的硬标签,再根据干净语音和带噪语音的平行关系,将干净语音的硬标签与带噪语音的硬标签进行对齐,而后训练 DNN 模型。所有 DNN 模型的参数设置参照 CHiME-2 的基线系统,均含有 7 个隐层,每个隐层有 2 048 个节点。以上模型在带噪语音测试集(eval92_5k)上关于 6 种信噪比(−6 至 9 dB)的 WER 如表 1 所示。

表 1 不同声学模型在带噪语音测试集上的 WER %

模型	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	平均
NC-GMM	69.12	61.14	53.90	43.25	35.76	29.05	48.70
N-GMM	64.06	53.76	48.23	37.55	30.39	25.14	43.19
NC-DNN	56.42	45.45	36.73	28.69	23.28	19.82	35.06
N-DNN	55.33	45.36	36.41	27.44	23.05	20.33	34.65
C-DNN	49.11	38.93	31.53	24.17	20.27	17.37	30.23

从表 1 可以看出,在各种信噪比下 DNN 中 C-DNN 的 WER 最低,NC-DNN 的 WER 最高。通过对实验数据加以分析,可以发现训练集利用 3 个 GMM-HMM 模型生成硬标签的数目存在差异。NC-GMM 模型生成的对齐信息中有 634 句(约 8.88%)没有硬标签,原因在于带噪语音和 NC-GMM 的训练数据不太匹配,此外带噪语音中的音素特征亦被噪声干扰或破坏。N-GMM 模型生成的对齐信息中有 421 句(约 5.89%)没有硬标签,虽然带噪语音和 N-GMM 的训练数据很匹配,但是其中的音素特征被噪声干扰或破坏。C-GMM 模型生成的对齐信息中只有 2 句没有硬标签,这是由于干净语音中的音素特征能被模型较好地感知。

为了与本文所提方法进行对比,本文选择 WER 最低的 C-DNN 模型作为学生模型的基线。

2.4 老师模型

为了验证老师模型的 WER 对学生模型 WER 的影响,本节将尝试把老师模型设为不同结构的神经网络:CNN、DNN、LSTM 和 BLSTM。

CNN 老师模型包含 2 个卷积层和 5 个全连接层;每个卷积层采用最大池化(max-pooling)进行处理,全连接层的节点数为 2 048。DNN 老师模型包含 7 个隐层,每个隐层有 2 048 个节点。LSTM 老师模型包含 5 个隐层,每层 640 个单元。BLSTM 老师模型包含 5 个隐层,每层 320 个单元。CNN 和 DNN 老师模型的初始学习速率均设为 0.008;LSTM 和 BLSTM 的初始学习速率均设为 0.000 01,冲量值均设为 0.9。

对所有老师模型而言,干净语音的训练集用于更新模型参数,干净语音的开发集用于模型选择和超参数的确定;干净语音硬标签由 C-GMM 生成,共 1 985 个。C-GMM 模型和老师模型在干净语音的开发集(dt_05)和测试集(eval92_5k)上的 WER 如表 2 所示。

表 2 C-GMM 和老师模型在干净语音数据集上的 WER %

模型	开发集	测试集
C-GMM	22.02	5.40
CNN	20.33	3.89
DNN	19.58	3.46
LSTM	18.89	2.97
BLSTM	18.30	2.65

从表 2 可以看出,所有老师模型的 WER 都比 C-GMM 模型低。其中,BLSTM 老师模型的 WER 最低,其次是 LSTM。就 BLSTM 和 LSTM 的训练而言,本文尝试多种参数并实验多次,但是它们与 DNN 的 WER 差距不太明显。这是由于模型在测试集上的 WER 已较低,因此 BLSTM 和 LSTM 的 WER 的下降空间较小。

2.5 学生模型

本文采用基于 DNN 的声学模型作为学生模型。此模型含有 7 个隐层,每个隐层有 2 048 个节点,输出层的节点数与老师模型的相同,即为 1 985。所有学生模型的标签为老师模型计算所得的软标签,模型参数利用带噪语音的训练集进行更新,模型的选择和超参数的设置在带噪语音的开发

集上进行。在节 2.4 老师模型的指导下，所有学生模型在带噪语音的测试集(eval92_5k)上关于 6 种信噪比(－6 至 9 dB)的 WER 如表 3 所示。

表 3 中，Baseline 表示本文设置的基线，即节 2.3 中的 C-DNN；CHiME-2 表示 CHiME-2 竞赛第一名的成绩[26, 28]；CNN、DNN、LSTM 和 BLSTM 分别表示 CNN、DNN、LSTM 和 BLSTM 老师模型指导的学生模型。CHiME-2 竞赛第一名的系统采用了语音增强模块和多种模型融合的策略。

表 3 学生模型在噪声测试集(eval92_5k)上的 WER %

模型	－6 dB	－3 dB	0 dB	3 dB	6 dB	9 dB	平均
Baseline	49.11	38.93	31.53	24.17	20.27	17.37	30.23
CHiME-2	44.12	35.42	28.12	21.20	17.34	14.83	26.86
CNN	43.70	34.00	25.87	20.12	17.21	14.68	25.93
DNN	41.92	31.35	25.33	19.26	15.73	14.38	24.66
LSTM	41.06	31.03	24.92	17.63	14.68	13.99	23.89
BLSTM	40.89	30.62	22.55	16.78	14.35	12.48	22.94

从表 3 可以看出，相比本文基线和 CHiME-2 竞赛的第一名，所有学生模型的 WER 在测试集的 6 种信噪比下均有显著下降。其中，BLSTM 老师模型指导的学生模型的 WER 最低。在带噪语音的测试集上，BLSTM 老师模型指导的学生模型的平均 WER 比基线的绝对下降了 7.29%，比 CHiME-2 竞赛第一名的绝对下降了 3.92%。

从表 3 还可以发现，相比高信噪比数据，低信噪比数据的 WER 降低幅度更为明显。就－6 dB 信噪比的数据而言，BLSTM 老师模型指导的学生模型的 WER 比基线的绝对下降了 8.22%，比 CHiME-2 竞赛第一名的绝对下降了 3.23%。就 9 dB 信噪比的数据而言，BLSTM 老师模型指导的学生模型的 WER 比基线的绝对下降了4.89%，比 CHiME-2 竞赛第一名的绝对下降了 2.35%。

3 讨 论

据上述实验结果可见，本文所提方法的 WER 在测试集的 6 种信噪比下均有显著下降。

在老师模型的指导下，学生模型在测试集各种信噪比下都能获得明显的性能提升，特别是对低信噪比数据的性能提升尤为显著。其原因有二：一是老师模型对干净语音中的音素特征能较好地感知和准确地建模；二是软标签含有更为丰富的信息。就原因一而论，干净语音中的音素特征能较好地被模型感知；而带噪语音的音素特征受到干扰甚至破坏，故不能被准确感知。与带噪语音相比，干净语音利用 GMM-HMM 生成的硬标签具有更高的正确率。就原因二而论，GMM-HMM 生成硬标签为由 0、1 构成的向量。然而，老师模型的后验概率分布是一种软标签，它是概率值向量。该软标签含有更为丰富的排名信息，不仅含有每帧数据最有可能的标签，且包含潜在可能标签的概率信息。因此，学生模型不但能利用这些丰富的信息进行更好地建模，亦可依据这些信息纠正部分错误的标签。

总之，将老师模型中的后验概率信息迁移到学生模型中，能较为明显地降低学生模型的 WER。此外，学生模型的 WER 与老师模型的 WER 成正相关。这是因为老师模型的 WER 越低，生成软标签正确率越高，从而使得学生模型对带噪语音的建模更为准确。

4 结 论

本文提出了基于迁移学习的方法对带噪语音进行声学建模，即利用老师模型指导学生模型进行训练。该方法能够有效地将老师模型中的后验概率信息迁移至学生模型中，从而提高声学模型在带噪数据集尤其是低信噪比数据集上的鲁棒性。在 CHiME-2 数据集上的实验结果表明，该方法的平均 WER 比基线的绝对下降了 7.29%，比 CHiME-2 竞赛第一名的绝对下降了 3.92%。实验结果亦表明，学生模型的 WER 与老师模型的 WER 成正相关。下一步将继续探讨利用组合老师模型指导学生模型进行训练等问题，并尝试改进该方法以适用于真实环境。

参考文献 (References)

[1] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[2] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: 2013: 6645-6649.

[3] HAŞIM S, ANDREW S, FRANÇOISE B. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. Computer Science, 2014(3): 338-342.

[4] XIONG W, DROPPO J, HUANG X, et al. The microsoft 2016 conversational speech recognition system [R/OL]. (2016-09-12) [2017-02-25]. https://arxiv.org/abs/1609.03528.

[5] SAON G, SERCU T, RENNIE S, et al. The IBM 2016 English conversational telephone speech recognition system [R/OL]. (2016-04-27) [2017-02-25]. https://arxiv.org/abs/1604.08242.

[6] 蔡尚,金鑫,高圣翔,等. 用于噪声鲁棒性语音识别的子带能量规整感知线性预测系数[J]. 声学学报,2012(6):667-672.
CAI S, JIN X, GAO S X, et al. Noise robust speech recognition based on sub-band energy warping perception linear prediction coefficient [J]. Chinese Journal of Acoustics, 2012(6): 667-672. (in Chinese)

[7] 胡旭琰,邹月娴,王文敏. 基于MDT特征补偿的噪声鲁棒语音识别算法[J]. 清华大学学报(自然科学版),2013(6):753-756.
HU X Y, ZOU Y X, WANG W M. Robust noise feature compensation method for speech recognition based on missing data technology [J]. Journal of Tsinghua University (Science and Technology), 2013(6): 753-756. (in Chinese)

[8] GALES M J F, PYE D, WOODLAND P C. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation [C]// International Conference on Spoken Language. Philadelphia, USA, 1996: 1832-1835.

[9] SIOHAN O, CHESTA C, LEE C H. Hidden Markov model adaptation using maximum a posteriori linear regression [C]// Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, Finland, 1999: 147-150.

[10] TRAN D T, DELROIX M, OGAWA A, et al. Factorized linear input network for acoustic model adaptation in noisy conditions [C]// Conference of the International Speech Communication Association. San Francisco, USA 2016: 3813-3817.

[11] SELTZER M L, YU D, WANG Y. An investigation of deep neural networks for noise robust speech recognition [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 7398-7402.

[12] YU D, SELTZER M L, LI J, et al. Feature learning in deep neural networks: Studies on speech recognition tasks [J]. Computer Science, 2013(2): 329-338.

[13] LI B, SIM K C. A spectral masking approach to noise-robust speech recognition using deep neural networks [J]. IEEE/ACM Transactions on Audio, Speech & Language Processing, 2014, 22(8): 1296-1305.

[14] 王青,吴侠,杜俊,等. 基于DNN特征融合的噪声鲁棒性语音识别[C]// 全国人机语音通讯学术会议. 天津:天津大学,2015:23-29.
WANG Q, WU X, DU J, et al. DNN based feature fusion for noise robust speech recognition [C]// National Conference on Man-Machine Speech Communication. Tianjin: Tianjin University, 2015: 23-29. (in Chinese)

[15] ABE A, YAMAMOTO K, NAKAGAWA S. Robust speech recognition using DNN-HMM acoustic model combining noise-aware training with spectral subtraction [C]// Conference of the International Speech Communication Association. Dresden, Germany, 2015: 2849-2853.

[16] XU Y, DU J, DAI L, et al. Dynamic noise aware training for speech enhancement based on deep neural networks [C]// Conference of the International Speech Communication Association. Singapore, 2014: 2670-2674.

[17] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C]// International Conference on Machine Learning. Helsinki, Finland, 2008: 1096-1103.

[18] KANG H L, KANG S J, KANG W H, et al. Two-stage noise aware training using asymmetric deep denoising autoencoder [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, 2016: 5765-5769.

[19] MIMURA M, SAKAI S, KAWAHARA T. Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition [C]// Conference of the International Speech Communication Association. Dresden, Germany, 2016: 3803-3807.

[20] QIAN Y, TAN T, YU D. An investigation into using parallel data for far-field speech recognition [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, 2016: 5725-5729.

[21] BUCILU C, CARUANA R, et al. Model compression [C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535-541.

[22] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [J]. Computer Science, 2015 (7): 382-390.

[23] LI J. Learning small-size DNN with output-distribution-based criteria [C]// Conference of the International Speech Communication Association, Singapore, 2014: 2650-2654.

[24] CHAN W, KE N R, LANE I. Transferring knowledge from a RNN to a DNN [J]. Computer Science, 2015(7): 138-143.

[25] CHEBOTAR Y, WATERS A. Distilling knowledge from ensembles of neural networks for speech recognition [C]// Conference of the International Speech Communication Association. Dresden, Germany, 2016: 3439-3443.

[26] VINCENT E, BARKER J, WATANABE S, et al. The second "CHiME" speech separation and recognition challenge: Datasets, tasks and baselines [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 126-130.

[27] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit [C] // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Big Island, USA, 2011.

[28] TACHIOKA Y. Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark [C]// CHiME Workshop. Vancouver, Canada, 2013: 6935-6939.

(责任编辑 刘森)