深层神经网络语音识别自适应方法研究*

邓 侃, 欧智坚
(清华大学 电子工程系, 北京 100084)

摘 要: 为了解决语音识别中深层神经网络的说话人与环境自适应问题,从语音信号中的说话人与环境因素的固有特点出发,提出了使用长时特征的自适应方案。基于高斯混合模型建立说话人—环境联合补偿模型,对说话人与环境参数进行估计,将此参数作为长时特征,将估计出来的长时特征与短时特征一起送入深层神经网络进行训练。Aurora4 实验表明,该方案可以有效地对说话人与环境因素进行分解,并提升自适应效果。

关键词: 语音识别; 声学模型自适应; 深层神经网络
中图分类号: TP391.42 **文献标志码**: A **文章编号**: 1001-3695(2016)07-1966-05
doi:10.3969/j.issn.1001-3695.2016.07.009

Adaptation method for deep neural network-based speech recognition

Deng Kan, Ou Zhijian
(*Dept. of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: To handle the speaker and noise adaptation problem in deep neural network-based speech recognition system, this paper studied the inherent characters of speaker and noise random factors and proposed a new adaptation method using long term features. Firstly, it built a joint adaptation model based on Gaussian mixture models and estimated and used the parameters of speaker and noise factors as long term features. Then, it used these long term features in deep neural network together with traditional short term features. Experiment results on Aurora4 database show that this method can effectively factorize speaker and noise factors, and improve adaptation performance.

Key words: speech recognition; acoustic model adaptation; deep neural networks

0 引言

语音识别的声学模型自适应关注噪声环境下以及说话人与训练集有差异时如何提升识别率。在语音识别实际应用场景中,这两类随机因素是固然存在的,抗噪稳健性与说话人自适应成为提升识别率的瓶颈,声学模型自适应的研究受到广泛关注。从自适应模块在系统中的位置上看,自适应研究的思路有前端自适应和模型域自适应两种。前端自适应方法对语音特征进行变换,达到补偿训练与测试数据失配的目的。前端方法具有运算复杂度低,方便快速的优点。

基于高斯混合模型(Gaussian mixture models,GMM),研究者提出了以下方法:针对说话人因素,采用声道长度归一化(vocal track length normalization,VTLN)[1];针对环境因素,采用倒谱均值减(cepstrum mean normalization,CMN)[2],此外,通用的前端变换方法,如 fMLLR(feature maximul likelihood linear regression)[3]和 HLDA(heteroscedastic linear discriminant analysis)[4]。在高斯混合模型中,采用变换过的特征进行训练与识别,可以取得较好的自适应效果[5]。

基于深层神经网络(deep neural networks,DNN),除了对已有前端变换进行实验验证[6]之外,研究者尝试引入其他的长时信息作为特征,送入前馈神经网络进行特征学习[7]。考虑时间窗长度的影响,研究者考察了 STC(split temporal context)特征在深层神经网络中的应用[8,9]。考虑句子层面的随机因素,研究者尝试将噪声作为扩展特征送入神经网络,进行带噪的训练和识别[10]。考虑跨句的长时特征,研究者考察说话人信息的利用。为了量化地表示说话人信息,研究者采用 iVector[11]作为特征提取工具,然后将提取得到的 iVector 作为扩展特征送入神经网络中[12,13],相比单纯采用短时特征,加入 iVector 扩展特征后识别率均得到了提升。

模型域自适应方法对模型参数进行变换,这一方案通常需要更新整个声学模型参数,比较耗时,但是自适应效果优于前端方法。

对于高斯混合模型,其参数有物理意义:均值代表特征分布的中心点,方差代表特征分布的范围。对高斯混合模型进行模型域自适应,需要对参数进行细致的建模。作为一般的自适应方法,研究者提出了最大似然线性回归(maximum likelihood linear regression,MLLR)[14]可以用于说话人自适应,也可以用于环境自适应。针对说话人因素,研究者提出了本征音(eigenvoice)方法[15]。本征音有效地利用了训练数据中的说话人先验知识,实验证明,自适应得到的本征音系数具有表征说话人信息的物理意义[15]。针对环境因素,尤其是加性噪声和卷积干扰,研究者提出了环境模型[16]。为了处理环境模型的非线性,矢量泰勒级数(vector taylor series,VTS)[14]被用于环境模型的局部线性近似,这一近似方法避免了繁琐的非线性积分,可以更加简便地从带噪语音中估计环境参数模型参数。

本文沿用在深层神经网络前端送入长时特征这一思路,选

收稿日期: 2015-04-10; **修回日期**: 2015-05-25 **基金项目**: 国家自然科学基金资助项目(61075020,61473168)

作者简介:邓侃(1989-),女,云南昆明人,硕士研究生,主要研究方向为稳健语音识别(karendeng@outlook.com);欧智坚(1975-),男,副研究员,博导,主要研究方向为语音处理与机器智能.

取说话人与环境这两类长时特征进行考察。其中,说话人是跨句的长时特征,环境特征是句子内部的长时特征。为了从带噪语音中估计这两类特征,需要对说话人与环境进行联合建模。

Wang 等人的说话人—环境联合建模研究[18,19]将 MLLR 用做干净语音模型,由于 MLLR 变换的通用性,这一方法未能实现有效分解[18]。在此基础上,笔者提出了本征音与环境联合建模方法[20],采用这一方法实现了对说话人与环境因素的有效分解,从而达到使用长时特征的目的。

基于上述说话人—环境长时特征估计方案,本文考察深层神经网络的长时特征使用效果。首先基于高斯混合模型,从带噪语音特征中估计得到噪声矢量和本征音矢量两类特征,实现环境因素与说话人因素的分离;然后将这两类长时特征与 MFCC(mel frequency cepstrum coefficient)短时特征一起,作为深层神经网络的输入特征进行特征学习。与单独送入 iVector 或者噪声矢量相比,这一方法能够综合考察噪声因素与说话人因素对于深层神经网络的影响。延续这一思路,本文完成了 Aurora4 数据集的实验。

1　基于高斯混合模型的说话人—环境特征估计方法

1.1　说话人建模:本征音

本征音方法[15]利用训练语音对说话人相关性进行建模。训练阶段,训练 S 个说话人相关模型,其中第 s 个模型的高斯均值可以拼接成超矢量 $\boldsymbol{\mu}^{(s)}=[\boldsymbol{\mu}_{11}^{\mathrm{T}},\boldsymbol{\mu}_{12}^{\mathrm{T}},\cdots,\boldsymbol{\mu}_{MN}^{\mathrm{T}}]^{\mathrm{T}}$,其中 $\boldsymbol{\mu}_{ij}^{\mathrm{T}}$ 为第 i 个高斯混合模型的第 j 个分量的均值。对 S 个超矢量 $\boldsymbol{\mu}^{(1)},\cdots,\boldsymbol{\mu}^{(s)}$ 进行 PCA 降维,得到一组基矢量 $\boldsymbol{e}_r(r=0,1,2,\cdots,R)$,称为本征音,其中 $\boldsymbol{e}_0=(\boldsymbol{\mu}_1+\cdots+\boldsymbol{\mu}_s)/S$ 为均值超矢量的算术平均值。识别阶段,假设目标说话人模型的均值超矢量是这些本征音的线性加权 $\boldsymbol{\mu}_x=\boldsymbol{e}_0+\sum_{r=1}^{R}w_r\boldsymbol{e}_r$,首先用自适应数据估计得到加权系数 w_r,然后计算加权组合后的高斯均值 $\boldsymbol{\mu}_x$,最后进行识别。

识别阶段,给定自适应语音数据 $\boldsymbol{\chi}(s)$ 和本征音 $\boldsymbol{E}=\{\boldsymbol{e}_0,\boldsymbol{e}_1,\cdots,\boldsymbol{e}_R\}$,可以用最大似然准则求解加权系数:

$$\hat{w}_{ML}=\arg\max_{w}P(\boldsymbol{\chi}(s)\mid\boldsymbol{E},w) \tag{1}$$

由于自适应数据的隐状态不可观测,需要用 EM 算法迭代求解。每次迭代需要求解关于 $\hat{w}$ 的线性方程组[15]:

$$\boldsymbol{A}\hat{\boldsymbol{w}}=\boldsymbol{b} \tag{2}$$

其中:矩阵 $\boldsymbol{A}$ 的维度是 $R\times R$,矢量 $\boldsymbol{b}$ 是 R 维矢量,对于 $1\leqslant r,1\leqslant R$,有

$$\begin{cases}\boldsymbol{A}(r,1)=\sum_{jk}\sum_{t}\gamma_{jk}(t)\boldsymbol{e}_{r,jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{e}_{l,jk}\\ \boldsymbol{b}(r)=\sum_{jk}\sum_{t}\gamma_{jk}(t)\boldsymbol{e}_{r,jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{jk}^{-1}(\boldsymbol{x}_t-\boldsymbol{e}_{0,jk})\end{cases} \tag{3}$$

其中:$\gamma_{jk}(t)$ 为第 j 个高斯混合模型的第 k 个分量在时刻 t 的状态占有概率;$\boldsymbol{\Sigma}_{jk}$ 为该分量的协方差矩阵。

求解上述线性方程组,可以得到本征音的权重矢量 $\hat{\boldsymbol{w}}$。接下来利用线性加权关系 $\boldsymbol{\mu}_x=\boldsymbol{e}_0+\sum_{r=1}^{R}w_r\boldsymbol{e}_r$,可以得到自适应之后的高斯均值。

在本征音模型中,基矢量 $\boldsymbol{E}=\{\boldsymbol{e}_0,\boldsymbol{e}_1,\cdots,\boldsymbol{e}_R\}$ 称为说话人空间,体现了人声差异性。权重矢量 $\boldsymbol{w}$ 则是目标说话人在这一线性空间中的位置,可以作为表征说话人的长时特征。已有实验表明,本征音权重矢量 $\boldsymbol{w}$ 的取值与说话人本身的特征,如性别、年龄具有相关性[15]。

1.2　环境因素建模:VTS

在倒谱域建立非线性环境模型[16]:

$$\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{h}+\boldsymbol{C}\ln\{1+\exp(\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}))\} \tag{4}$$

其中:$\boldsymbol{C}$ 代表 DCT 变换矩阵;$\boldsymbol{x}$ 代表干净语音特征;$\boldsymbol{n}$ 是加性噪声在倒谱域的表示,对于每个句子,将 $\boldsymbol{n}$ 建模为高斯分布,其参数为 $\boldsymbol{\mu}_n$、$\boldsymbol{\Sigma}_n$;$\boldsymbol{h}$ 是卷积信道冲激响应的倒谱域表示,为确定而未知的常数。

如果直接对这一非线性模型进行参数估计,会涉及复杂的非线性积分,利用矢量泰勒级数(VTS)将上述模型在干净语音码本均值附近展开,得到近似的线性关系式[17]如下。

$$\boldsymbol{y}\approx\boldsymbol{y}|_{(\boldsymbol{\mu}_{x,jk},\boldsymbol{\mu}_{n,jk})}+\boldsymbol{G}_{x,jk}(\boldsymbol{x}-\boldsymbol{\mu}_{x,jk})+\boldsymbol{G}_{n,jk}(\boldsymbol{n}-\boldsymbol{\mu}_n) \tag{5}$$

其中:线性项的系数 $\boldsymbol{G}_{x,jk}$,$\boldsymbol{G}_{n,jk}$ 为 Jacobi 矩阵。

基于上述线性模型,用 EM 算法估计环境模型参数 $\boldsymbol{\Theta}=\{\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n,\boldsymbol{\mu}_h\}$,给出噪声均值 $\boldsymbol{\mu}_n$ 的参数重估式如下[17]:

$$\hat{\boldsymbol{\mu}}_n=\boldsymbol{\mu}_n+[\boldsymbol{\Sigma}_{jk}\gamma_k\boldsymbol{G}_{n,jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{x,jk}^{-1}\boldsymbol{G}_{n,jk}]\boldsymbol{\Sigma}_{jk}\boldsymbol{G}_{n,jk}^{\mathrm{T}}\boldsymbol{\Sigma}_{x,jk}^{-1}\boldsymbol{c}_{x,jk} \tag{6}$$

其中:γ_{jk},$\boldsymbol{c}_{x,jk}$ 是带噪语音特征的两个充分统计量:

$$\begin{cases}\gamma_{jk}=\sum_t\gamma_{jk}(t)\\ \boldsymbol{c}_{x,jk}=\sum_t\gamma_{jk}(t)(\boldsymbol{y}_t-\boldsymbol{\mu}_{\gamma,jk})\end{cases} \tag{7}$$

信道参数 $\boldsymbol{\mu}_h$ 的重估公式参考文献[17]。噪声方差 $\boldsymbol{\Sigma}_n$ 的重估采用的是 Gauss-Newton 方法,参见文献[21]。

1.3　说话人—环境长时特征估计模型

在联合分解模型中,干净语音不再用说话人无关模型建模,而是用本征音描述:$\boldsymbol{\mu}_x=\boldsymbol{e}_0+\sum_{r=1}^{R}w_r\boldsymbol{e}_r$,其中 $w_1,\cdots,w_R$ 是待估计的权重系数,需要对每个说话人估计一组参数。环境因素仍然用非线性环境模型进行建模,采用 VTS 作线性近似。需要考虑的随机因素是加性噪声和卷积干扰 $\boldsymbol{H}=\{\boldsymbol{n},\boldsymbol{h}\}$,对于每句话来说,加性噪声 $\boldsymbol{n}$ 服从高斯分布,卷积信道的冲激响应是个确定但未知的参数。

建立联合分解模型后,对于每个说话人的语音数据,待估计的参数集合为 $\boldsymbol{\Theta}=\{\boldsymbol{\mu}_n^u,\boldsymbol{\Sigma}_n^u,\boldsymbol{\mu}_h^u,u=1,\cdots,U;w\}$,其中 U 为该说话人的语音句子数目,参数集合 $\boldsymbol{\Theta}$ 分为两部分:

a)本征音系数 w;

b)环境模型参数 $\boldsymbol{\Lambda}=\{\boldsymbol{\mu}_n^u,\boldsymbol{\Sigma}_n^u,\boldsymbol{\mu}_k^u,u=1,\cdots,U\}$。

由于噪声的非平稳性特性,对每句话,需要单独估计一组环境模型参数 $\boldsymbol{\mu}_n^u$、$\boldsymbol{\Sigma}_n^u$、$\boldsymbol{\mu}_k^u$;说话人因素相比起噪声来说变化较为缓慢,应当对来自同一说话人的一批语音数据进行批处理,共同估计一个本征音系数。

将干净语音 $\boldsymbol{\mu}_x$ 的本征音模型公式 $\boldsymbol{\mu}_x=\boldsymbol{e}_0+\sum_{r=1}^{R}w_r\boldsymbol{e}_r$ 代入环境模型,再做 VTS 局部线性展开,推导得到第 u 句带噪语音 $\boldsymbol{y}$ 的 GMM 码本参数如下:

$$\begin{cases}\boldsymbol{\mu}_{y,jk}^{(u)}=g(\boldsymbol{e}_{o,jk}+\sum_{r=1}^{R}w_r\boldsymbol{e}_{r,jk},\boldsymbol{\mu}_n^{(u)},\boldsymbol{\mu}_h^{(u)})\\ \boldsymbol{\Sigma}_{y,jk}^{(u)}=\boldsymbol{G}_{x,jk}^{(u)}\boldsymbol{\Sigma}_{x,jk}(\boldsymbol{G}_{x,jk}^{(u)})^{\mathrm{T}}+(1-\boldsymbol{G}_{x,jk}^{(u)})\boldsymbol{\Sigma}_n^{(u)}(1-\boldsymbol{G}_{x,jk}^{(u)})^{\mathrm{T}}\end{cases} \tag{8}$$

对参数 $\boldsymbol{\Theta}=\{w,\boldsymbol{\Lambda}\}$ 进行最大似然估计的过程,就是从带噪语音中学习说话人特征 w 与环境特征 $\boldsymbol{\Lambda}=\{\boldsymbol{\mu}_n^{(u)},\boldsymbol{\Sigma}_n^{(u)},\boldsymbol{\mu}_k^{(u)},u=1,\cdots,U\}$ 的过程。为了处理 HMM 的隐含状态,仍然要用 EM 算法进行迭代求解。

在 E 步,写出关于 $\boldsymbol{\Theta}$ 的辅助函数如下:

$$Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})=\sum_{u,t}\sum_{j,k}\gamma_{jk}^{(u)}(t)\log P\{\boldsymbol{y}_t|j,k,\hat{\boldsymbol{\mu}}_{y,jk}^{(u)}(\hat{\boldsymbol{\theta}}),\hat{\boldsymbol{\Sigma}}_{y,jk}^{(u)}(\hat{\boldsymbol{\theta}})\} \tag{9}$$

在求解 M 步时,为了实现 w 与 $\boldsymbol{\Lambda}$ 的解耦合,采取坐标轮换的迭代求解方法,交替地固定 w 求解 $\boldsymbol{\Lambda}$,然后固定 $\boldsymbol{\Lambda}$ 求解 w。固定 w 时,可以根据公式 $\boldsymbol{\mu}_x=\boldsymbol{e}_0+\sum_{r=1}^{R}w_r\boldsymbol{e}_r$ 求出干净语音的

GMM 模型参数,从而 Λ 的估计公式与经典的 VTS 环境参数估计一致。下面推导固定 Λ 时,本征音系数 w 的求解公式。

首先处理带噪语音码本 $\hat{\mu}_{y,jk}^{(u)}$ 与 $\hat{w}$ 的非线性关系,再次采用 VTS 线性近似方法,将这一非线性函数在 w 附近展开:

$$\hat{\mu}_{y,jk}^{(u)} \approx \hat{\mu}_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} (\hat{w}_r - w_r) e_{r,jk} \tag{10}$$

基于这一近似关系,可以求解关于 w 的最大化问题,先求偏导数 $\partial Q/\partial \hat{w}$:

$$\frac{\partial Q}{\partial \hat{w}_r} = \sum_{u,t} \sum_{j,k} \gamma_{jk}^{(u)}(t) e_{r,jk}^{\mathrm{T}} (\gamma_{x,jk}^{(u)})^{\mathrm{T}} (\hat{\Sigma}_{y,jk}^{(u)})^{-1} \cdot \{ y_t - [\mu_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} (\hat{w}_r - w_r) e_{r,jk}] \} \tag{11}$$

令上述偏导数等于 0,得到关于 $\hat{w}$ 的线性方程组:

$$A\hat{w} = b \tag{12}$$

对于 $1 \leqslant r, 1 \leqslant R$,线性方程组系数表示如下:

$$\begin{cases} A(r,l) = \sum_{u,t} \sum_{j,k} \gamma_{jk}^{(u)}(t) e_{r,jk}^{\mathrm{T}} (G_{x,jk}^{(u)})^{\mathrm{T}} (\hat{\Sigma}_{y,jk}^{(u)})^{-1} G_{x,jk}^{(u)} e_{l,jk} \\ b(r) = \sum_{u,r} \sum_{j,k} \gamma_{jk}^{(u)}(t) e_{r,jk}^{\mathrm{T}} (G_{x,jk}^{(u)})^{\mathrm{T}} (\hat{\Sigma}_{y,jk}^{(u)}) (y_t - \\ \quad \mu_{y,jk}^{(u)} + G_{x,jk}^{(u)} \sum_{r=1}^{R} w_r e_{r,jk}) \end{cases} \tag{13}$$

求解上述线性方程组,即可实现固定 Λ、估计本征音系数的目的。

综合以上步骤,得到采用坐标轮换法求解 EM 算法辅助函数最大化的算法流程:

a)对每句话,用首尾 20 帧带噪语音特征的平均值作为噪声参数 $\mu_n^{(u)}$、$\Sigma_n^{(u)}$ 的初始化估计,对于信道参数,设置 $\mu_n^{(u)} = 0$,对于本征音模型,设置 $w = 0$,即采用平均说话人模型作为初始的说话人模型。

b)对于噪声的初始化值,对式(8)进行 VTS 展开,更新带噪语音模型,计算 HMM 隐含状态的充分统计量。

c)基于当前估计的环境模型参数 $\mu_n^{(u)}$、$\Sigma_n^{(u)}$、$\mu_h^{(u)}$ 与干净语音模型,利用本节描述的参数估计方法,利用带噪语音特征估计本征音展开系数,更新干净语音模型。

d)基于当前估计的本征音系数 w,利用式 $\mu_x = e_0 + \sum_{r=1}^{R} w_r e_r$ 得到干净语音模型的参数,然后利用经典的 VTS 方法估计环境模型参数 $\mu_n^{(u)}$、$\Sigma_n^{(u)}$、$\mu_h^{(u)}$,最后利用式(8)的 VTS 近似展开更新带噪语音模型。

2 深层神经网络的说话人—环境联合自适应方法

2.1 深层神经网络结构与训练流程

深层神经网络由多层感知器级联而成,经典的前馈神经网络结构如图 1 所示。

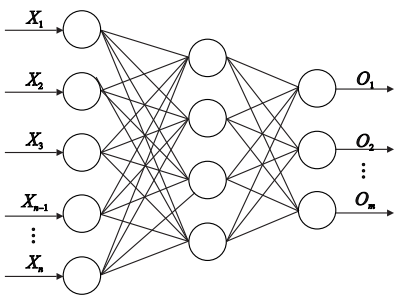

图1 经典的前馈神经网络结构图

其中属于同一层的神经元相互没有连接,层与层之间神经元的连接关系可以用权重矩阵 W 表示。

神经网络训练是一个典型的最优化问题,给定训练数据及其标注,通过反向传播(back propagation,BP)算法训练权重。为了解决 BP 算法容易陷入局部极小的问题,Hinton 提出了用受限玻尔兹曼机(restricted Boltzmann machine,RBM)对神经网络参数进行无监督预训练的算法[22]。

加入预训练的深层神经网络声学模型训练流程如下[23]:

a)按照标准流程训练 GMM-HMM 模型,通过决策数据类确定三音子状态集,设计神经网络结构。

b)不考虑标注,用训练语音特征无监督训练 RBM 模型参数,将其作为前馈神经网络的权重初值。

c)利用 GMM 对训练数据进行强制对齐(forced alignment),将词层面的标注对齐到状态层面。

d)利用 BP 算法对神经网络的权重与阈值进行有监督训练。

训练得到 DNN 模型后,可以用 DNN 模型替代 GMM 更新状态标注,重复步骤 c)、d)。

2.2 基于说话人—环境随机因素分解的深层神经网络自适应方法

在深层神经网络训练中,输入特征的设计至为关键。MFCC、FBANK、PLP 都是典型的短时特征,只能包含前后若干帧内的信息。如何超越短时特征,利用句子层面甚至是跨句的信息,成为研究此关注的焦点。

说话人与环境是语音信号中变化相对缓慢的因素,现有的深层神经网络自适应方法中,这两类随机因素仍然被分开处理。单独处理说话人,研究者用 iVector 提取说话人特征,作为长时特征,与短时特征一起送入神经网络[12]。单独处理环境因素,研究者用噪声数据的均值作为噪声估计,对短时语音特征加以扩展[10]。

上述研究分别考虑了说话人因素与环境因素,基于前面的研究成果,本文提出对这两类随机因素进行联合特征学习:首先利用 GMM-HMM 联合补偿模型,从语音信号中提取能代表说话人与环境因素的特征,然后将这些特征和短时特征一起送入神经网络进行学习,其中的关键点在于,说话人与环境因素的学习需要用产生式模型 GMM,而非鉴别式模型 DNN,DNN 仅仅作为特征学习工具。

基于上述阐述,本文对神经网络联合自适应方法进行如下实验:

a)单独考察噪声因素的实验。利用环境模型的 VTS 展开方法,从带噪语音中估计噪声参数,将每句话的噪声参数作为各帧噪声样本的估计,与短时特征一起送入神经网络。

b)单独考察说话人因素的实验。对于测试集,从平行的干净语料中提取 eigenvoice,将权重矢量作为说话人特征送入神经网络中。这一方法仍然是从干净语音中提取 eigenvoice 权重,只利用了说话人信息。

c)考察说话人与环境因素联合建模的实验。利用 GMM 产生式模型的联合自适应方法,从带噪语音数据中分解得到噪声参数与本征音参数,作为特征送入神经网络,实现在线无监督联合自适应。

3 Aurora4 带噪数据集实验结果分析

3.1 Aurora4 带噪数据集介绍

Aurora4[24] 是带噪的大词汇量英文语音识别实验平台,其数据由华尔街日报数据(Wall Street Journal)人工加噪得到。Aurora4 的测试集有 14 个,其语音内容完全相同,考虑卷积干

扰和加性噪声,对 Nov'92 测试集加噪得到 Aurora4 的 14 个测试集。按照加噪方式的不同,把 14 个测试集划分为集合 A、B、C、D 四个部分,test 01 无信道干扰、无噪声,记为 setA;test 02 ~ 07 无信道干扰、有噪声,记为 set B;test 08 有信道干扰、无噪声,记为 setC;test 09 ~ 14 有信道干扰、有噪声,记为 SetD。

Aurora4 的训练集分为干净训练和多场景(multi-condition)训练。在 GMM 基线系统与自适应实验中,由于是对干净语音建模,采用的是干净训练的方法;在 DNN 基线系统与自适应实验中,为了考察 DNN 对带噪特征的学习能力,采用的是多场景的带噪训练方法。

3.2　基于 GMM 的说话人—环境联合分解实验

作为对照实验,这里给出 MLLR + VTS 和本文提出的 eigenvoice + VTS 实验结果。在实验中,MLLR 仍然采用两个回归类的配置,固定迭代次数为 2,eigenvoice 采用 20 个本征音作为线性空间的基。实验步骤如前面所述,实验识别错误率结果如表 1 所示。其中 VTS 第二遍噪声参数更新为单独进行 VTS 自适应的结果,在 MLLR 与 eigenvoice 联合补偿方法中,均采用两轮坐标轮换对说话人与环境的参数进行更新。

表 1　Aurora4 数据集的 GMM-HMM 联合补偿模型实验(干净训练)

自适应方法	setA	setB	setC	setD	总平均
GMM-HMM 基线实验系统	8.93	54.38	48.35	69.89	57.35
VTS 第二遍噪声参数更新	7.97	16.21	13.68	23.38	18.51
MLLR + VTS 联合补偿实验	7.15	14.69	12.59	21.72	17.02
eigenvoice + VTS 联合补偿实验	6.85	13.91	11.65	21.44	16.47

从实验结果可见,采用 eigenvoice + VTS 方法可以取得比 MLLR + VTS 更好的效果。可见在大词汇量连续语音识别任务中,同时考虑说话人因素和环境因素,能为识别率带来进一步的提升,这一提升的来源是因为 VTS 的环境模型是对干净语音进行加噪,而本征音恰好只针对干净语音建模。MLLR 是一种通用的参数估计方法,已有文献的实验结果表明,从带噪语音中估计出的 MLLR 变换矩阵并不完全代表干净语音,也包含一部分噪声因素,所以本征音建模更加符合 VTS 环境模型的基本假设,二者联合补偿能带来识别率的提升。

为了验证 VTS + eigenvoice 联合补偿模型的有效性,本文设计联合分解实验,对这两类随机因素进行分解。联合分解系统如图 2 所示。

联合分解实验的目的在于:验证从带噪语音中提取的说话人特征,即本征音分解系数,真正能够代表其背后的干净语音说话人信息。为了保证联合分解实验的正确性,用于进行联合补偿自适应的带噪语音数据,必须与最后用于对本征音自适应模型进行测试的干净语音完全一致。由于 Aurora4 是人工加噪,各个测试集的语音内容完全一致,为平行语料,保证了实验设计的正确性。在下述联合分解实验中,采用三种方法估计本征音:用经典的本征音估计方法,直接从干净数据中估计本征音;用联合补偿模型从 test01(干净语音)中估计本征音;用联合补偿模型从 test04(加入 Restaurant 噪声)测试集中估计本征音。下面依次给出这三种本征音估计方式的实验结果,作为对照,本文也列出了用自适应后的模型识别带噪数据的实验结果。联合分解实验识别错误率结果如表 2 所示。

从上述实验结果可以看出,对于本征音的估计,最为精确的方式应当是从干净语音中估计,这种方法与本征轴的训练场景最为匹配,估计出来的本征音自适应效果最好(eigenvoice 行,setA 列);在实际场景下,往往难以采集干净的语音,需要采用联合补偿模型进行本征音估计(第 2、3 行)。由于受噪声模型的干扰,本征音估计得不够准确,自适应后的结果没有直接进行 eigenvoice 好,但是自适应后的错误率为 8.17%,相对自适应前的基线系统 8.93% 的错误率有 8% 的相对错误率下降。因此,本文提出的联合补偿模型是一种从噪声中提取说话人信息的行之有效的方案。

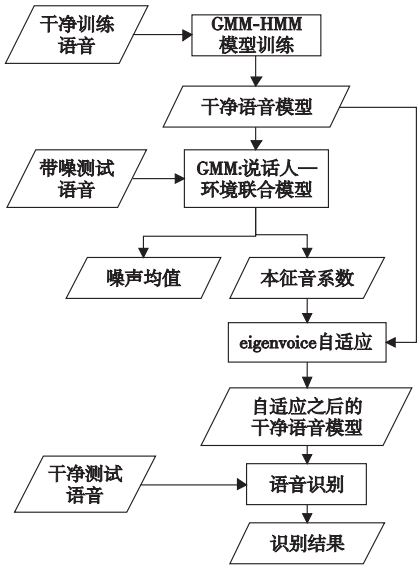

图2　Aurora4 联合分解实验系统设计图

表 2　Aurora4 数据集的说话人—环境因素分解实验(干净训练)

自适应方法	setA	setB	setC	setD	总平均
GMM-HMM 基线系统	8.93	54.38	48.35	69.89	57.35
eigenvoice	7.12	51.04	47.65	68.54	55.16
从 test01 中估计本征音	7.59	50.29	47.87	67.76	54.55
从 test04 中估计本征音	8.17	49.85	48.29	66.79	54.02

3.3　深层神经网络自适应实验结果

深层神经网络的基线识别系统采用开源工具包 Kaldi 搭建,使用 Kaldi 训练 DNN 模型的步骤如第 2 章所述。

对于 Aurora4 数据集,采用带噪数据集训练深层神经网络,以考察 DNN 对于多样化的训练样本的学习能力。利用 Kaldi 工具包,对语音数据提取 39 维 MFCC 与 72 维 FBANK 特征,用 MFCC 模型训练基线 GMM 系统,三音子模型的状态数为 3 026 个,采用鉴别式训练进一步提高识别率,用做状态标注模型。神经网络采用经典的 7 × 2048 隐层结构,输入层送入前后 11 帧特征,输出层为 softmax,训练准则是交叉熵。为了达到避免过拟合的目的,在训练集中随机抽取 1/5 的数据用做交叉验证集合,一旦误差下降不大时,就停止训练。设置学习速率为 0.08,较小的学习速率需要经历多遍迭代才能收敛,但能避免验证误差剧烈的振荡。

在提取噪声与说话人特征时,以下实验暂时用前后 20 帧特征作为噪声均值的初始化估计,考虑到 Aurora4 提供了干净语音,为了验证想法,暂时从干净语音中估计本征音系数,把权重矢量作为说话人特征。用上述特征,初步验证 DNN 联合补偿模型的想法是否可行,识别错误率结果如表 3 所示。

从以上实验结果可以看出,从平均识别率来看,带噪训练的错误率从 DNN 基线的 15.12% 下降到了 14.82%,验证了带

噪训练的有效性。

从总体实验结果来看,向 DNN 模型中送入说话人特征 eigenvoice 系数,对于干净语音 setA 效果显著,这与现有文献的结果[24]是一致的。但是,对于带噪语音 setB,setC,setD,加入 eigenvoice 后效果反而变差了,由于这里的 eigenvoice 是从平行语料中提取的,可以排除 eigenvoice 提取不够准确的原因。对于带噪语音,如何把 eigenvoice 这一上层特征与底层特征 MFCC 加以融合,还有待探索。

表 3 Aurora4 数据集的说话人—环境因素分解实验(干净训练)

DNN 特征自适应方法	setA	setB	setC	setD	总平均
GMM 基线实验	7.04	12.73	12.67	26.13	18.06
DNN-MFCC	3.84	8.78	10.65	24.08	15.12
DNN-MFCC + noise	3.75	8.77	9.83	23.54	14.82
DNN-MFCC + speaker	3.64	8.84	11.71	24.3	15.3
DNN-MFCC + noise + speaker	3.56	8.71	11.69	24.21	15.2

4 结束语

本文根据语音信号中噪声与说话人因素固有的特性,提出了一种从带噪语音中提取本征音系数与噪声均值,作为长时特征送入深层神经网络进行训练的方法。Aurora4 数据集上的实验表明,这一方法实现了说话人与环境特征的有效分解,能够从语音信号中提取具有物理意义的长时特征,并在干净测试集上得到了识别率的提升。

在今后的研究中,还要考虑如何修改神经网络模型结构,使得噪声与说话人这两类长时特征可以更好地与短时特征进行融合,进一步提升带噪语音的识别率。

参考文献:

[1] Lee L, Rose R C, Richard C. Speaker normalization using efficient frequency warping procedures[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]: IEEE Press, 1996:353-356.

[2] Liu Fuhua, Stern R M, Huang Xuedong, *et al*. Efficient cepstral normalization for robust speech recognition[C]//Proc of Association for Computational Linguistics Workshop on Human Language Technology. 1993:69-74.

[3] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition[J]. Computer Speech & Language, 1998,12(2):75-98.

[4] Duin R P W, Loog M. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2004,26(6):732-739.

[5] Gong Yifan. Speech recognition in noisy environments: a survey[J]. Speech Communication,1995,16(3):261-291.

[6] Seide F, Li Gang, Chen Xie, *et al*. Feature engineering in context-dependent deep neural networks for conversational speech transcription[C]//Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. [S. l.]:IEEE Press,2011:24-29.

[7] Li Jinyu, Deng Li, Gong Yifan, *et al*. An overview of noise-robust automatic speech recognition[J]. IEEE Trans on Audio, Speech and Language Processing,2014,22(4):745-777.

[8] Siniscalchi S M, Yu Dong, Deng Li, *et al*. Speech recognition using long-span temporal patterns in a deep network model[J]. Signal Processing Letters,2013,20(3):201-204.

[9] Baccouche M, Besset B, Collen P, *et al*. Deep learning of split temporal context for automatic speech recognition[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]: IEEE Press,2014:5422-5426.

[10] Seltzer M L, Yu Dong, Wang Yongqiang. An investigation of deep neural networks for noise robust speech recognition[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]:IEEE Press,2013:7398-7402.

[11] Dehak N, Kenny P, Dehak R, *et al*. Front-end factor analysis for speaker verification[J]. IEEE Trans on Audio, Speech, and Language Processing,2011,19(4):788-798.

[12] Saon G, Soltau H, Nahamoo D, *et al*. Speaker adaptation of neural network acoustic modeling using I-vectors[C]//Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. [S. l.]: IEEE Press,2013:55-59.

[13] Senior A, Lopez M I. Improving DNN speaker independence with I-vector inputs[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]:IEEE Press,2014:225-229.

[14] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models [J]. Computer Speech & Language,1995,9(2):171-185.

[15] Kuhn R, Junqua J C, Nguyen P, *et al*. Rapid speaker adaptation in eigenvoice space[J]. IEEE Trans on Speech and Audio Processing,2000,8(6):695-707.

[16] Moreno P J, Raj B, Stern R M. A vector Taylor series approach for environment-independent speech recognition[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]: IEEE Press,1996:733-736.

[17] Li Jinyu, Deng Li, Yu Dong, *et al*. A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions [J]. Computer Speech & Language,2009,23(3):389-405.

[18] Wang Yongqiang, Gales M J F. Speaker and noise factorization for robust speech recognition[J]. IEEE Trans on Audio, Speech, and Language Processing,2012,20(7):2149-2158.

[19] Karanasou P, Wang Yongqiang, Gales M J F, *et al*. Adaptation of deep neural network acoustic models using factorised I-vectors[C]//Proc of the 15th Annual Conference of International Speech Communication Association. 2014:155-159.

[20] Ou Zhijian, Deng Kan. Combining eigenvoice speaker modeling and VTS-based environment compensation for robust speech recognition [C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]:IEEE Press,2012:4673-4676.

[21] Zhao Yong, Juang B H. Non-linear noise compensation for robust speech recognition using Gauss-Newton method[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]: IEEE Press,2011:4796-4799.

[22] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation,2006,18(7):1527-1554.

[23] Seltzer M L, Yu Dong, Wang Yongqiang. An investigation of deep neural networks for noise robust speech recognition[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal. [S. l.]:IEEE Press,2013:7398-7402.

[24] Parihar N, Picone J. Aurora working group: DSR front end LVCSR evaluation AU/384/02[R]. Starkville: Institution for Signal and Information Process, Mississippi State University,2002.