基于 DNN 特征融合的噪声鲁棒性语音识别*

王青[1]，吴侠[2]，杜俊[1]，戴礼荣[1]
(1. 中国科学技术大学，合肥 230027；2. 重庆邮电大学，重庆 400000)

文　摘：本文比较了几种不同的特征融合方法。在有噪声影响的情况下，深度神经网络可以极大的降低语音识别的错误率。本文基于倒谱均值规整（Cepstral Mean Normalization，CMN）、均值方差规整（Mean Variance Normalization，MVN）、功率规整倒谱系数（Power Normalized Cepstral Coefficients，PNCC）、ETSI AFE 标准前端四种不同的特征规整和提取方式，比较了神经网络输入层融合、隐藏层融合、输出层融合三种特征融合以及后验概率平均和 ROVER（Recognizer Output Voting Error Reduction）的方法，希望可以进一步降低识别错误率。在 Aurora4 数据库上的实验表明，融合方法还是有一定效果的。

关键词：参数规整；鲁棒性特征；特征融合
中图分类号： 分类号 1；分类号 2

随着自动语音识别（ASR：Automatic Speech Recognition）技术的发展，语音识别的噪声鲁棒性越来越受到人们的关注。各种各样的噪声鲁棒性技术可以分为两大类，一类是特征域方法，一类是模型域方法[1][2]。模型域方法通过对声学模型进行变换以适应特定实际环境，对运算复杂度要求更高，所以本文主要关注特征域，包括进行参数规整和提取鲁棒性特征，减少训练和识别环境之间的不匹配。

近年来，深度神经网络（Deep Neural Network，DNN）在噪声鲁棒性语音识别领域得到了广泛应用[3][4][5][6]。基于深度学习的特征融合方法也越来越受到关注，文章[7]把 Cochleogram 和 Spectrogram 特征结合起来训练卷积神经网络（Convolutional Neural Network，CNN）和 DNN，有比较大的性能提升。文章[8]提出了一种集成深度学习的方式，对递归神经网络（Recurrent Netural Network，RNN）、DNN 和 CNN 的输出后验特征进行线性融合，可以一定程度提高系统的识别性能。本文利用不同参数规整和鲁棒性特征之间存在的互补性，比较了几种基于 DNN 的融合方式。输入层融合对不同的输入特征做拼接处理，然后利用拼接后的特征训练 DNN；隐藏层融合在某一层对该层输出值拼接并训练一个与融合层数有关的 DNN；输出层融合对不同系统输出的后验概率做做非线性融合处理，训练一个多层感知器（Multi-layer perceptron，MLP）；后验概率平均（Posterior Average，Post-Avg）是对两个或多个系统的后验概率直接求取平均；ROVER（Recognizer Output Voting Error Reduction,，ROVER）利用不同识别结果的出现频次和置信度得分来进行判决。CMN 和 MVN 属于比较简单的规整方法，可以补偿信道畸变和加性噪声的影响，PNCC 利用人的听觉特性，ETSI AFE 标准前端通过维纳滤波做降噪处理，都在一定程度上增加了语音特征的鲁棒性。本文主要研究和比较了特征之间的互补性，验证融合方法的有效性。

本文组织如下：第 1 节介绍参数规整和特征提取的方法；第 2 节介绍 DNN 的五种融合方法；第 3 节是实验结果和分析；第 4 节是结束语。

1　参数规整和特征提取

1.1　倒谱均值规整（CMN）

CMN 是比较鲁棒的方法之一，该方法的基本原理是通过对特征参数的均值进行规整，来消除时不变信道的影响[9]。

$x_t = y_t - \bar{y}$ 设带噪语音特征序列（一般指一句话）为 $Y = [y_0, y_1, \ldots, y_{T-1}]$，则 CMN 算法公式如下：

$$x_t = y_t - \bar{y} \tag{1}$$

$$\bar{y} = \frac{1}{T}\sum_{t=0}^{T-1} y_t \tag{2}$$

其中，下标 t 表示第 t 帧，T 是一句话的总

*基金项目：国家自然科学基金（61305002）
作者简介：王青（1989-），女（汉），河北，博士研究生。
通讯联系人：杜俊，副教授，jundu@ustc.edu.cn

帧数，$\bar{y}$ 是一句话的均值。

1.2 均值方差规整（MVN）

类似地，MVN 不仅对特征参数的均值规整，还对其方差也进行规整，这样既能消除时不变信道的影响，也可以减少加性噪声的影响[10]。MVN 计算公式为：

$$x_t = \frac{y_t - \bar{y}}{\sigma_y} \tag{3}$$

$$\sigma_y = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(y_t - \bar{y})^2} \tag{4}$$

其中,$\bar{y}$ 和 σ_y 分别是一句话的均值和方差。

1.3 功率规整倒谱系数（PNCC）

在存在单一说话人干扰和背景音乐等更复杂的噪声环境中，语音识别错误率上升很快。PNCC[11]是为了减少环境变化对语音识别影响而提出的，保证在语音不失真时性能不下降的情况下来使得特征更加具有鲁棒性。

PNCC 提取过程中比较重要的几个技术点有：用幂律非线性替代 MFCC 提取中使用的对数非线性；在短时分析的基础上增加 50ms-120ms 的“中时”处理；使用不对称非线性滤波来估计噪声；采用了时域掩蔽的方式。

1.4 ETSI AFE 标准前端

ETSI AFE[12]标准前端针对噪声信号做了一系列处理，降低噪声带来的影响。在特征提取部分，首先进行降噪过程，使用二阶维纳滤波器；然后对降噪之后的语音信号做波形处理，并计算频谱；最后对频谱信号进行盲规整得到原始的 AFE。这样提取得到的AFE具有较强的噪声鲁棒性，训练环境和识别环境不匹配时依然具有比较好的性能。

2 特征融合

2.1 输入层融合

DNN 输入层融合并不改变网络的结构，而是对输入的特征直接做拼接处理，然后使用拼接后的特征训练 DNN。输入层融合这种方式实现最为简单，不需要进行额外的训练，实现原理如下图 1 所示。

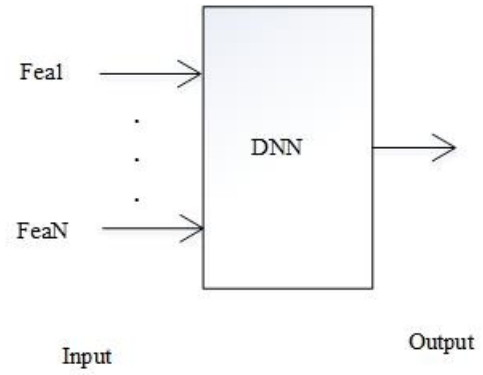

图 1 输入层融合实现框架图

2.2 隐藏层融合

DNN 隐藏层融合是在某一个隐藏层后对该层网络值进行拼接，然后按照训练传统 DNN 的流程对剩下的网络层数训练，图 2 是隐藏层融合的实现框架图，需要在已有 DNN 基础上额外训练一个结构稍浅的 DNN。

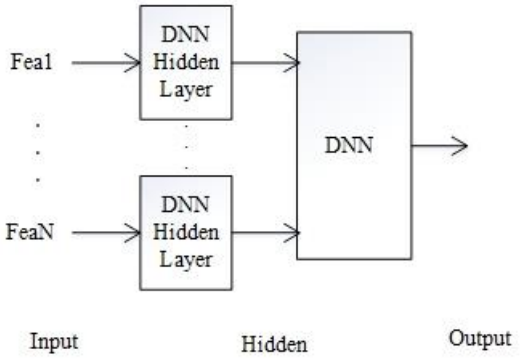

图 2 隐藏层融合实现框架图

2.3 输出层融合

DNN 输出层融合是对 DNN 输出状态的后验概率特征进行拼接，然后训练一个 MLP，利用融合得到的后验概率特征解码进行识别，图 3 是输出层融合的实现框架图，需要在已有 DNN 基础上额外训练一个 MLP。

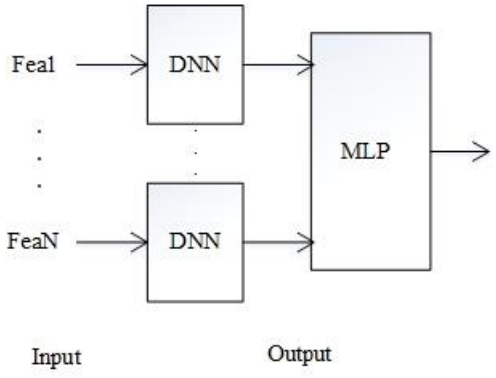

图 3 输出层融合实现框架图

2.4 后验概率平均

后验概率平均[13]是最简单的一种后端处理方式，通过计算两个或者多个系统后验概率的平均值，得到一个性能更优的识别特征。实现原理框图如下：

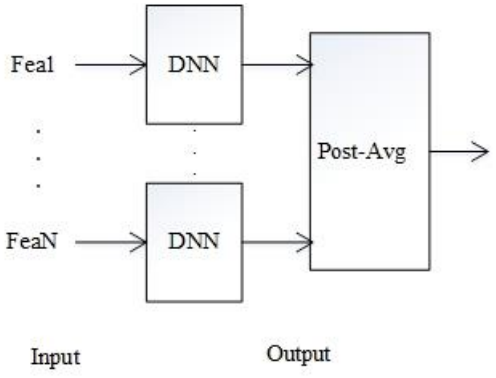

图 4 后验概率融合实现框架图

ROVER[14]融合是识别文本层面上的融合，不涉及到训练部分，与 Post-Avg 都属于比较简单的融合方式。

3 实验结果及分析

3.1 实验配置

本文使用 Aurora4[15][16]数据库验证特征融合的有效性。Aurora4 是人工加入噪声和信道影响的英文数据库，它规定了两种声学模型训练模式：一种模式是用第一种麦克风录制的 7138 句干净语音训练（称为 Clean），另一种模式是用干净语音和带噪语音混合训练（称为 Multi）。对 Multi 训练集，一半语音是用第一种麦克风录制的，一半语音是用第二种麦克风录制的，并且这两种语音都包含了来自干净训练集合的语音以及混合了六种噪声（街道、火车站、汽车、人声、餐馆、飞机场）且信噪比在 10-20dB 的带噪语音。两类原始测试集合是用第一、二种麦克风录制的干净语音，然后分别混合训练集中使用到的六种噪声且信噪比为 5-15dB，所以一共 14 个测试集合。我们把这 14 个测试集合分为 4 类：干净（set1）、带噪（set2-set7）、干净并且信道不匹配（set8）、带噪并且信道不匹配（set9-set14），分别记为 A、B、C、D。

对于前端处理部分，每一帧特征的帧长是 25ms，帧移是 10ms。本文提取了 3 种类型的声学特征，第 1 种是 13 维 MFCC 特征及其一、二阶动态差分，按照规整方法的不同，分别记为 CMN 和 MVN；第 2 种是 13 维 PNCC 特征及其一、二阶动态差分；第 3 种是 13 维 AFE 特征及其一、二阶动态差分。

对于 GMM-HMM，每个 3 音素模型有 3 个状态，每个状态 16 个混合高斯。静音（silence）模型有 3 个状态，短暂停顿（short pause）模型只有 1 个状态，并且这个状态和静音模型的中间状态绑定在一起。最终基于决策树绑定的状态是 3296 个。对于 DNN-HMM，输入层扩展 11 帧（39*11=429），7 个隐藏层，每个隐藏层为 2048 个节点，输出层为 3296 个节点的 soft-max 层，与 HMM 绑定后的状态保持一致。其他的参数设置参考文献[17]。

对输入层融合来说，输入节点数目根据融合的特征种类成倍增加；对于隐藏层融合，融合后训练的 DNN 隐藏层节点数依然为 2048，隐层层数根据融合所在层数确定；对于输出层融合的 MLP，只有 1 个隐藏层且节点数目为 2048。

3.1 实验结果及分析

在这一小节中，我们将先给出四种特征的基线性能，然后比较五种不同特征融合方式。

表 1 给出了四种特征的基线识别结果。从实验结果可以看出来，在训练环境和测试环境不匹配的情况下（即 Clean-condition training），PNCC 和 AFE 识别 WER 更低，说明这两种特征对噪声具有更好的鲁棒性。在 Multi 训练下，四种特征的 WER 总体相差不多，都相对 Clean 训练情况有了很大提高，其中 CMN 的 WER 是最低的，这表明了简单的均值规整就很有效果。由于 Multi 训练的 WER 远低于 Clean 训练的 WER，因此融合实验室基于 Multi 训练完成的。

表 1 四种特征的基线识别结果（WER）

Feature	A	B	C	D	Avg.
Clean-condition Training					
CMN	4.6	29.6	23.1	45.9	34.3
MVN	5.0	24.3	20.1	39.9	29.3
PNCC	5.3	20.0	18.1	38.1	26.6
AFE	5.0	20.8	22.2	36.1	26.1
Multi-condition Training					
CMN	5.2	9.4	8.9	20.1	13.7
MVN	5.9	9.8	9.7	21.1	14.4
PNCC	5.0	9.8	9.4	20.4	14.0
AFE	5.0	10.5	9.6	20.6	14.4

表 2 给出了在输入层融合的结果，可以看出，CMN 分别与 PNCC 和 AFE 融合是有一定效果的，并且 MVN 和 AFE 融合的相对提升最大，三种和四种特征融合的结果比两种特征融合稍微差一点，这是因为 CMN 和 MVN 都属于 MFCC 特征，

只是参数规整方式稍有差异，所以特征会存在信息冗余，说明并不是拼接特征越多对识别性能越好，而是跟融合的特征有关系。

表 2 网络输入层融合识别结果（WER）

	A	B	C	D	Avg
CMN-PNCC	5.0	9.3	8.9	19.9	13.4
CMN-AFE	4.8	9.0	8.8	19.3	13.1
MVN-AFE	5.1	9.2	8.6	20.0	13.5
AFE-MVN-CMN	4.6	9.0	8.6	19.6	13.2
FOUR	4.9	9.1	8.7	19.7	13.3

表 3 是隐藏层融合的结果，隐藏层的融合结果表明 4 隐藏层融合比 1 隐藏层融合 WER 下降很多，网络深一些，提取的特征噪声鲁棒性更好一些。

表 3 隐藏层融合识别结果（WER）

	A	B	C	D	Avg.
MVN-AFE_ Hidden1	6.9	9.9	9.1	20.3	14.1
MVN-AFE_ Hidden4	5.3	9.1	8.6	19.3	13.2

表 4 是网络输出层的融合结果，比较输入层融合和输出层融合，平均 WER 几乎一样。比较 MVN 和 AFE 的融合，隐藏层融合 WER 比输入层融合和输出层融合稍低一些。深度神经网络可以看做一个非线性特征提取器，能把低层特征形成更加抽象的高层表示属性类别或者特征，对输入层或者输出层，要么是原始的低层特征，要么是类别信息很强的高层特征，而中间隐藏层则既保留了原始低层特征又包含了类别信息，所以融合的性能是相对比较好的。

表 4 网络输出层融合识别结果（WER）

	A	B	C	D	Avg.
CMN-PNCC	4.9	9.2	8.9	19.8	13.4
CMN-AFE	5.1	9.2	8.9	19.2	13.2
MVN-AFE	5.9	9.3	8.2	19.7	13.4
FOUR	4.9	9.2	8.4	19.6	13.3

表 5 是 Post-Avg 融合的结果，两种特征融合时，Post-Avg 优势不是很明显，但是三种或者四种特征融合，Post-Avg 性能提升很大。

表 5 Post-Avg 融合识别结果（WER）

	A	B	C	D	Avg.
CMN-PNCC	4.9	9.1	8.9	19.2	13.1
CMN-AFE	7.6	9.5	8.1	18.2	13.2
MVN-AFE	7.5	9.4	8.0	18.4	13.0
AFE-MVN-CMN	6.4	9.1	8.0	18.3	12.8
FOUR	5.8	8.8	7.9	18.1	12.5

表 6 是使用 ROVER 融合的结果，ROVER 利用不同识别结果中单词的出现频次和置信度得分来判决，因此融合的系统越多，性能会越好。在两种或者三种特征融合时，输入层融合和输出层融合要优于 ROVER 融合，说明特征之间的互补性是存在的。

表 6 ROVER 融合识别结果（WER）

	A	B	C	D	Avg.
CMN-PNCC	5.5	9.6	8.9	19.9	13.7
CMN-AFE	5.2	9.5	8.4	19.8	13.5
MVN-AFE	5.5	9.6	8.6	20.0	13.7
AFE-MVN-CMN	4.9	9.0	8.6	19.5	13.2
FOUR	4.9	8.7	8.4	19.2	12.9

综合比较上述几种融合方法，Post-Avg 最简单并且最有效，ROVER 融合适用于多个系统融合，输入层、隐藏层、输出层融合性能相差不多，隐藏层融合稍微好一点。

4 结束语

本文比较了五种基于 DNN 的特征融合的方法，简单容易实现，在 Aurora4 数据库上的实验表明，输入层融合和输出层融合可以达到相当的水平，融合效果很接近，隐藏层融合性能稍好，Post-Avg 和 ROVER 最容易实现，并且 Post-Avg 融合性能最好。这说明利用特征融合的方式，可以进一步提高系统的噪声鲁棒性。

参 考 文 献

[1] 丁沛，曹志刚．基于语音增强失真补偿的抗噪声语音识别技术[J]．中文信息学报，2004, 18(5)．

[2] Gong Y.. speech recognition in noisy environments：a survey [J]. Speech Communication, 1995, 16(3).

[3] Jun Du, Qing Wang, Tian Gao, et al. Robust speech recognition with speech enhanced deep neural networks[A]. Proc. INTERSPEECH, 2014

[4] Mohamed A, Dahl G, and Hinton G. Acoustic modeling using deep belief networks[J], IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 14-22, 2012.

[5] Hinton G, Deng Li, Yu Dong, et al. Deep neural networks for acoustic modeling in speech recognition[J], IEEE Signal Processing Magazine, Vol. 29, No. 6, pp.82-97, 2012.

[6] Narayanan A. and Wang De-Liang. Ideal ratio mask estimation using deep neural networks for robust speech recognition[A], Proc. ICASSP, 2013, pp. 7092-7096.

[7] Tjandra A, Sakti S, Neubig G, et al. Combination of two-dimensional cochleogram and spectrogram features for deep learning-based ASR[A]. Proc. ICASSP, 2015

[8] Deng Li and John C. Platt. Ensemble deep learning for speech recognition[A]. Proc. INTERSPEECH, 2014

[9] 杜俊，胡郁，王仁华．鲁棒性语音识别中的一种特征参数规整

的优化算法[A]. 第八届全国人机语音通讯学术会议论文集. 2005.

[10] Viikki O and Laurila K. Cepstral domain segmental feature vector normalization for noise robust speech recognition [J]. Speech Communication, 1998, 25(1): 133-147.

[11] Chanwoo Kim and Richard M. Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition[A]. Proc. ICASSP, 2012

[12] ETSI. Speech processing, transmission and quality aspects(STQ); distributed speech recognition; advanced front-endfeature extraction algorithm; compression algorithms[S]. Tech. Rep. ETSI ES 202 050, ETSI, October 2002.

[13] Li Bo and Khe Chai Sim. Improving robustness of deep neural networks via spectral masking for automatic speech recognition[A], in Automatic Speech Recognition and Understanding(ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 279–284

[14] Fiscus J G.. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)[A], in Proceedings IEEE Automatic Speech Recognition and Understanding Workshop, pp. 347–352, Santa Barbara, CA, 1997

[15] Hirsch H. G.. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task[J], ETSI STQ Aurora DSR Working Group, Version 2.0, 2002.

[16] Parihar N and Picone J, DSR front end LVCSR evaluation [J], Aurora Working Group, 2002.

[17] Seltzer M L, Yu Dong, and Wang Y-Q. An investigation of deep neural networks for noise robust speech recognition[A], Proc. ICASSP, 2013, pp. 7398-7402.

Deep neural network based feature fusion for noise robust speech recognition

Qing Wang[1], Xia Wu[2], Jun Du[1], Li-Rong Dai[1]

1. University of Science and Technology of China, Hefei, 230027

2. Chongqing University of Posts and Telecommunications, Chongqing, 400000

Abstract: We compare several feature fusion methods in this paper. Deep neural network (DNN) can greatly reduce word error rate (WER) of speech recognition in noisy environments. Based on four parameter normalization and feature extraction methods namely cepstral mean normalization (CMN), mean variance normalization (MVN), power normalized cepstral coefficients (PNCC), and ETSI advanced front end (AFE), we compare three fusion methods in DNN in input layer, hidden layer and output layer besides posterior average and ROVER (Recognizer Output Voting Error Reduction) method in order to reduce WER further more. Experimental results on Aurora4 task indicate that feature fusionhave some effect.

Key words: parameter normalization; robust feature; feature fusion