ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 1012-1021
DOI:10.16337/j.1004-9037.2017.05.019

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于 MTL-DNN 系统融合的混合语言模型语音识别方法

范正光　屈　丹　李　华　张文林

(解放军信息工程大学信息系统工程学院,郑州,450002)

摘　要:基于混合语言模型的语音识别系统虽然具有可以识别集外词的优点,但是集外词识别准确率远低于集内词。为了进一步提升混合语音识别系统的识别性能,本文提出了一种基于互补声学模型的多系统融合方法。首先,通过采用不同的声学建模单元,构建了两套基于隐马尔科夫模型和深层神经网络(Hidden Markov model and deep neural network, HMM-DNN)的混合语音识别系统;然后,针对这两种识别任务之间的关联性,采用多任务学习(Multi-task learning DNN, MTL-DNN)思想,实现 DNN 网络输入层和隐含层的共享,并通过联合训练提高建模精度。最后,采用 ROVER(Recognizer output voting error reduction)方法对两套系统的输出结果进行融合。实验结果表明,相比于单任务学习 DNN(Single-task learning DNN, STL-DNN)建模方式,MTL-DNN 可以获得更好的识别性能;将两个系统的输出进行融合,能够进一步降低词错误率。

关键词:集外词;混合模型;多任务学习结层神经网络;系统融合

中图分类号:TN912.3　　**文献标志码**:A

Hybrid Language Model Speech Recognition Method Based on MTL-DNN System Combination

Fan Zhengguang, Qu Dan, Li Hua, Zhang Wenlin

(Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou, 450002, China)

Abstract: Speech recognition system based on the hybrid language model has the advantage of recognizing the out-of-vocabulary (OOV) words, but the recognition accuracy of the OOVs is far below that of the in-vocabulary (IV) words. To further improve the performance of hybrid speech recognition, a system combination method based on complementary acoustic models is proposed in this paper. Firstly, two hybrid speech recognition systems based on hidden Markov model and deep neural network (HMM-DNN) are set up by using different acoustic modeling unites. Aiming at the relevance of these two recognition tasks, the thought of multi-task learning (MTL) is then used to share the input and hidden layers of DNN and improve the modeling accuracy by joint training. Finally, the outputs of two systems are combined with recognizer output voting error reduction (ROVER). Experimental results show that the MTL-DNN modeling method can obtain better recognition performance than the single-task learning DNN(STL-DNN) and the combining of the two systems can further reduce the final word error rates (WER).

Key words: out-of-vocabulary words; hybrid model; multi-task learning deep neural netwrok(MTL-DNN); system combination

---

基金项目:国家自然科学基金(61175017,61403415,61302107)资助项目。
收稿日期:2015-12-02;**修订日期**:2016-02-21

引 言

大词汇量连续语音识别(Large vocabulary continuous speech recognition,LVCSR)系统根据固定的词表建立发音词典和语言模型,词表以外的词称为集外词(Out-of-vocabulary,OOV)。目前,LVCSR 系统多不具备识别集外词的能力,只能将集外词识别成一个或几个发音相近的集内词(In-vocabulary,IV words),这在一定程度上影响了语音识别系统的性能。近年,随着一些机器学习技术,特别是深度学习技术在语音识别中的成功应用,语音识别的性能获得了大幅提高,但是集外词问题却仍未得到有效解决,如何有效地识别集外词仍是语音识别任务的瓶颈之一。

目前,对于集外词识别问题的研究主要集中在两方面:集外词检测和集外词恢复。前者主要通过对语音识别结果的分析来确定集外词的位置,是集外词识别的基础;后者是在前者的基础上获得集外词的正确拼写,以实现集外词的真正识别。针对集外词检测,研究人员提出了多种类型的方法,其中具有代表性的有基于音素匹配的方法[1]、基于二分类的方法[2]以及基于混合语言模型的方法[3-7]等。针对集外词恢复,往往根据检测结果采取不同的策略,其中音素字母转换[8]是较为常用的方法之一。

近年来,由于其应用灵活,集外词检测性能较好且便于恢复的特点,基于混合语言模型的语音识别方法受到许多学者的重视。该方法的基本原理是,利用词和子词构建混合字典以及混合语言模型,在解码时得到包含子词单元的混合识别结果,最终利用这些子词单元实现集外词的检测与恢复。为了提高混合模型方法的集外词识别性能,研究人员做了诸多研究,如文献[4]采用词片作为子词单元构造混合系统,并在混合识别结果的基础上使用后验概率特征对检出的集外词作进一步判决,提高了检测性能;文献[5]对比了多种不同子词单元的集外词检测与恢复性能,指出子词单元的覆盖度、长度以及类型等都会对集外词识别产生很大影响;文献[6]针对现有混合语言模型存在子词单元数据稀疏问题,提出了一种将词语言模型和子词单元语言模型先分开训练后融合的两阶段混合语言模型构造方法。虽然上述这些混合模型方法为集外词识别带来了一些改善,但是集外词的识别准确率仍然远低于集内词。

对于连续语音识别系统,将不同识别器的识别结果进行融合可以有效地降低系统的词错误率(Word error rate, WER)。具体而言,识别结果的融合可以在 1-best 识别结果上进行,如 ROVER 方法(Recognizer output voting error reduction)[9],也可以在词图结果上进行[10]。相关实验表明,通过融合可以有效利用各系统间的信息互补性来提高其识别性能。据此,本文提出了一种通过系统融合来提升混合系统识别性能的方法。该方法采用基于隐马尔科夫模型和深层神经网络(Hidden Markov model and deep neural network ,HMM-DNN)的混合模型语音识别系统作为基线系统,通过采用不同的声学建模单元(即音素和字素),训练具有描述差异的声学模型。在两套系统的基础上,结合 ROVER 融合准则,实现模型间某些薄弱环节的互相改善,从而取得更好的识别性能。同时,针对两个 DNN 任务间的关联性,提出采用多任务学习(Multitask-task learning, MTL)的思想,对两个 DNN 进行联合训练,进一步改善各子系统的性能。实验结果表明,多任务学习 DNN(Multi-task learning DNN, MTL-DNN)建模方式要明显优于单任务学习 DNN(Single-task learning DNN,STL-DNN),在此基础上的系统融合可以获得最好的识别性能。

1 基于深层神经网络的混合语音识别系统

深层神经网络(Deep neural network,DNN)是具有多隐含层的神经网络,比传统的高斯混合模型(Gaussian mixture model,GMM)具有更强的声学建模能力。DNN 与隐马尔科夫模型(Hidden Markov model, HMM)结合的方法已经成为语音识别领域的主流框架[11]。图 1 给出了基于 DNN 的混合语音识别系统框图。该系统利用 DNN 进行声学建模,代替 GMM 进行状态输出概率的计算,并采用混合解码器进行解码。该系统与传统语音识别系统的主要区别在于它可以对集外词进行识别,其中混合解码

器利用混合字典以及混合语言模型得到混合识别结果。在混合识别结果中,集内词识别成词的形式,而集外词则识别成如音素、字母音素对以及词素等子词单元形式。

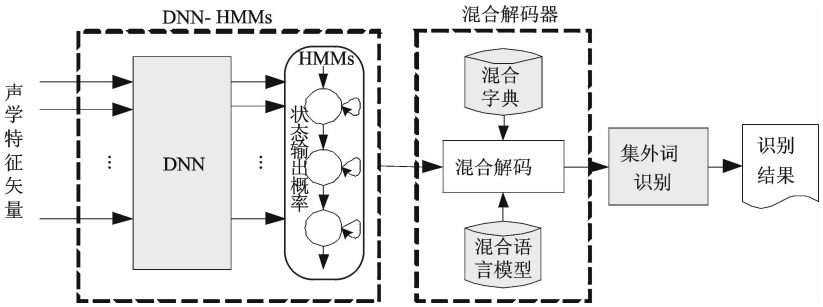

图 1 基于深层神经网络的混合语音识别系统框图

Fig. 1 Hybrid speech recognition system based on deep neural network

1.1 混合字典

混合字典由词、子词单元以及它们的发音混合构成,子词单元用于解码时表示集外词。在混合模型中,常用的子词单元包括音素、音节、字母音素对和词素等。其中,词素是构成词的最小单位的音义集合体,由于其使用方便且集外词识别性能较好,因此在混合模型中得到了广泛的应用[12]。本文采用词素作为子词单元,词素由单词的分解得到,常见的分解方法包括基于数据驱动和基于知识驱动的方法。本文采用 Morfessor[12] 工具得到词素,该工具采用数据驱动方法,基于最小描述长度(Minimum description length, MDL)准则,对给定语料实现最优分解,从而得到词素集。

1.2 混合语言模型

将语言模型训练语料中的集外词表示成相应的子词单元序列得到混合语料。为了便于集外词识别,增加词边界信息以标明单词边界。如对于语句"DUM JE ZBORENY"(捷克语),其中,"ZBORENY"为集外词,将其表示为带有词边界标记的词素序列,得到混合形式"DUM JE ZB+ ORE+ NY#."。由混合语料训练得到混合语言模型。在混合语言模型中不仅包括词的 N-gram 参数,也包括词与子词以及子词与子词的 N-gram 参数。采用该混合语言模型进行解码,即可得到混合识别结果。

1.3 集外词识别

图 2 给出了一个集外词识别示例,对混合解码器得到的混合识别结果,首先通过子词单元序列确定集外词位置,然后根据该序列以及词边界标记进行集外词恢复,从而得到正确的识别结果。

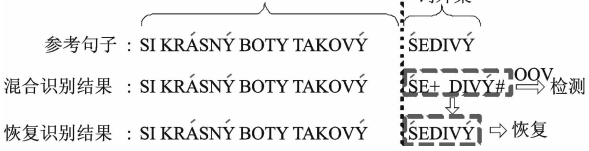

图 2 基于混合模型的集外词识别示例

Fig. 2 An example of OOV recognition based on hybrid model

2 基于 MTL-DNN 的系统融合方法

单一的混合模型系统,在进行集外词识别时往往具有较低的准确率和召回率,识别性能远远低于集

内词。本文着重探讨通过构造具有足够建模差异的声学模型来为语音集外词识别带来更多的互补信息。系统融合的关键步骤是建立若干性能相当、具有一定互补性的系统。互补性使得不同系统提供的信息具有足够的差异以取长补短;而如果两个系统性能相差太多,互补的优势反而会被其中一个系统过多的错误所掩盖,二者缺一不可。下文将详细阐述建模差异的互补系统的构造方法。

2.1 构建差异性模型系统

目前,DNN 用于声学建模时,相对于其他模型,如 GMM、子空间高斯混合模型(Subspace Gaussian mixture model, SGMM)等具有更好的区分性,在识别性能上也往往明显优于其他模型。因此,选用 DNN 与其他模型作为互补系统可能会因为性能差异而丧失互补优势。为了得到性能相匹配的互补系统,本文在采用 DNN 进行声学建模的基础上,使用不同的基本建模单元(音素和字素)进行差异性建模。

基线系统使用音素作为基本的声学建模单元。进行声学建模时,需考虑上下文相关信息,这样单音素识别基元就扩展为上下文相关的三音素基元。三音素建模是目前最常用的一种建模方式,大量研究证明,采用三音素声学建模方式对于大多数语言都可以取得较好的识别性能。但是,基于三音素的声学建模也存在一定的缺点。首先,对于识别语言需要准确确定该种语言的音素集。该音素集往往由语言学家确定,对于多资源语言如英语、汉语等,音素集的获取往往较为容易,而对于一些低资源语言,由于研究较少,音素集的确定也就较为困难。音素集不准确则难以反映语音中多种多样的声学变异现象,从而影响识别性能。此外,采用三音素建模时,需要创建精确的发音字典,实现词与音素之间的映射。字典往往由专家手动创建,因此也限制了三音素建模在一些低资源语言中的应用。

根据文献[14]的研究,采用字素(grapheme)作为基本建模单元,对于一些语言,特别是发音变化较弱的语言,往往可以得到与音素相近的识别性能,同时可以克服采用音素进行建模时存在的诸多不足。一般来说一个字母可以认为是一个字素。本文所采用的识别语料为捷克语料,根据文献[15],捷克语的发音规则较为确定,采用字素作为建模单元可以获得与音素相当的识别性能。因此,本文采用字素作为具有差异性的基本建模单元,并仿照音素引入上下文信息得到三字素(trigrapheme)基元。通过使用具有差异性的基本建模单元,在保证模型性能可比性的同时,由于 HMM 建模状态序列以及模型参数等都会发生较大改变,因此可以利用融合策略实现模型间的优势互补,从而提高集外词识别的系统性能。

2.2 系统融合方法

系统融合主要利用基于连续语音识别结果的 ROVER 方法来实现。ROVER 系统主要包括结果对齐模块和投票重打分模块。结果对齐模块实现不同系统线性输出结果的强制对齐并得到词转移网络(Word transition network, WTN)。投票重打分模块对词转移网络进行重打分,从而得到更好的识别结果,重打分公式为

$$S(W_i)=\alpha F(W_i)+(1-\alpha)C(W_i) \tag{1}$$

式中:$F(W_i)$表示词出现频率,$C(W_i)$表示词置信度得分。置信度得分有平均置信度得分和最大置信度得分两种选择。由于本文只是针对两个系统的合并,因此 α 设为 0,$C(W_i)$ 采用最大置信度得分。

系统融合结构框图如图 3 所示。首先,根据 2.1 节分别采用三音素和三字素建模得到两套差异性的系统,分别被称为 Triphone 系统和 Trigrapheme 系统;然后,分别利用两系统对识别语句进行解码,得到两个混合识别结果;最后,采用 ROVER 进行融合,并对融合结果进行集外词识别得到最终结果。在该系统融合框架中,两套系统的 DNN 为独立训练得到,用 senone 表示聚类后的绑定状态,则 Triphone 系统的 DNN 输出层为三音素 senone 的后验概率,Trigrapheme 系统的 DNN 输出层为三字素 senone 的后验概率。为与下文的 MTL-DNN 进行区分,此处的 DNN 被称之为 STL-DNN。

2.3 基于 MTL-DNN 的系统融合

多任务学习是一种通过对多个相关任务联合学习来提高各子任务性能的机器学习方法。在语音识别，特别是低资源语音识别中，基于 DNN 的多任务学习方法表现出优异的性能[16]。2.1 节在训练基于三音素以及三字素的声学模型时，各任务之间具有很强的联系性，主要体现在：(1)使用相同的输入特征；(2)具有相近的识别性能；(3)采用相同的神经网络结构等。因此，可以考虑采用多任务学习的思想对基于三音素和三字素的声学模型进行联合学习。图 4 给出了三音素和三字素声学模型联合训练的 MTL-DNN 结构以及基于该 MTL-DNN 的系统融合结构流程。在 MTL 框架下，原先的两个独立的 DNN 通过共享输入层以及隐含层进行了合并，仅保留各自的输出层。在 MTL-DNN 中，给定输入特征 x_t，则输出层第 i 个三音素 senone 的后验概率可以通过 Softmax 函数得到，即

$$p(s_i^{\mathrm{p}} \mid X)=\frac{\exp(y_i^{\mathrm{p}})}{\sum_{j=1}^{N^{\mathrm{p}}}\exp(y_j^{\mathrm{p}})} \qquad \forall\, i=1,\cdots,N^{\mathrm{p}} \tag{2}$$

式中：y_i^{p} 为激活函数的输出，N^{p} 为三音素 senone 的总数。同理，第 i 个三字素 senone 的后验概率为

$$p(s_i^{\mathrm{g}} \mid X)=\frac{\overline{\exp}(y_i^{\mathrm{g}})}{\sum_{j=1}^{N^{\mathrm{p}}}\exp(y_j^{\mathrm{g}})} \qquad \forall\, i=1,\cdots,N^{\mathrm{g}} \tag{3}$$

MTL-DNN 预训练与 STL-DNN 相同，都是通过无监督训练方法预先训练一个深置信网络(Deep belief networks, DBN)，再利用 DBN 初始化 DNN。不同的是，STL-DNN 的输出层为单一的 Softmax 层，而 MTL-DNN 的输出层为两个分离的 Softmax 层，因此在进行参数调优时，MTL-DNN 通过最小化两个任务所有输入特征向量的交叉熵之和实现，该交叉熵可以表示为

$$F_{\text{sum-ce}}=-\sum_{x}\Big(\sum_{i=1}^{N^{\mathrm{p}}}d_i^{\mathrm{p}}(x)\log P(s_i^{\mathrm{p}} \mid x)+\sum_{i=1}^{N^{\mathrm{g}}}d_i^{\mathrm{g}}(x)\log P(s_i^{\mathrm{g}} \mid x)\Big) \tag{4}$$

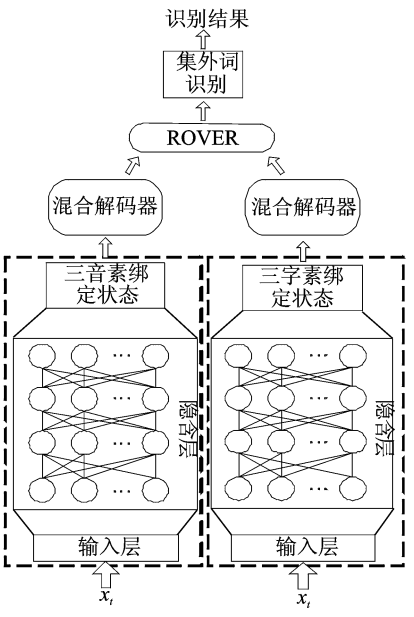

图 3 基于 STL-DNN 的系统融合结构流程图

Fig. 3 System combination architecture based on STL-DNN

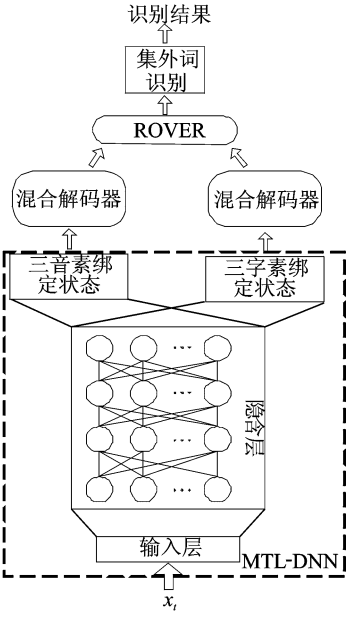

图 4 基于 MTL-DNN 的系统融合结构流程图

Fig. 4 System combination architecture based on MTL-DNN

式中:$d_i^p(x)$和 $d_i^g(x)$分别为输入 x 对应的第 i 个三音素 senone 和第 i 个三字素 senone 的目标值。在进行 MTL-DNN 训练时,各有一个三音素 senone 和一个三字素 senone 的目标值设为 1,其余设为 0。

MTL-DNN 与 STL-DNN 的解码方式一致,首先将 MTL-DNN 中各 senone 的后验概率通过除以其先验概率得到似然度,然后利用 Viterbi 解码得到最终解码结果。相比于 STL-DNN,MTL-DNN 实现了多个任务之间输入层与隐含层的参数共享,将两个独立的 DNN 训练任务联系到一起,这两个识别任务之间可以互相提供有帮助性的新信息,从而提高建模精度,改善各任务的识别性能。

3 实验结果和分析

3.1 实验数据

实验语料使用开源捷克语电话语料 Vystadial_cz[17],其中声学模型训练集(TRAIN)由 22 567 句话构成,数据总时长约 15.25 h;开发集(DEV)包括 2 000 句话,时长约为 1.23 h;测试集(TEST)的数据量与开发集基本相同,包含 2 000 句话,总时长约 1.22 h。语言模型训练语料包含两部分,一部分是训练集的标注数据,总共包含大约 1 MB 的文本数据;另一部分为欧盟官方公报提供的的开源捷克语文本语料[18],包含大约 60 MB 的数据。词汇表通过统计训练集的标注文本获取,包含 14 890 个单词,并通过 Kaldi[19]中的捷克语字母音素转换器获取发音,构造发音字典。词素集通过对语言模型训练语料中的集外词分解得到,引入词边界标记共得到 14 353 个词素。表 1 给出了不同数据集中集外词的数量及所占比例。

表 1 各数据集中集外词的数量及所占比例

Tab. 1 OOV words number and rate of different data sets

字典大小/个	开发集(DEV)		测试集(TEST)	
	数量	比例/%	数量	比例/%
14 890	787	6.69	736	6.39

3.2 实验设置

实验主要基于开源工具包 Kaldi 以及 Pdnn[20]搭建。声学特征采用 13 维的 Mel 频率倒普系数(Mel frequency cepstrum coefficient, MFCC)参数及其一阶、二阶差分,总特征维数为 39 维,帧长为 25 ms,帧移为 10 ms。在基于 HMM-GMM 的语音识别系统中,声学模型参数采用最大似然估计(Maximum likehood estimation, MLE)方法得到,其中 HMM 模型为包含 3 个发射状态的、自左向右无跨越的三音素或者三字素模型。分别对三音素和三字素模型采用决策树进行状态聚类后,系统最终分别包含 1 036 和 1 038 个不同的上下文相关状态。用 GMM 对各状态的输出概率进行建模,由于数据量不同,各状态 GMM 的混元数也不同。最终,三音素声学模型中包含 19 232 个高斯混元,三字素声学模型包含 19 239 个高斯混元。解码器采用 Kaldi 中基于加权有限状态转换器(Weighted finite-state transducer, WFST)的语音识别解码器,该解码器通过将声学模型、语言模型以及发音字典表示成 WFST 形式,并通过合成算法得到一个完整的 WFST 模型,然后采用维特比光束(Viterbi beam)搜索算法进行搜索解码得到最终识别结果,光束宽度为 1.6×10^5。

实验中所有的 DNN 模型设置 4 个隐含层,每层包含 2 048 个节点,输入为 11 帧(当前帧以及前后各 5 帧)的声学特征向量,输出节点与聚类状态数一致。最终,对于三音素声学模型,其 DNN 结构为“440-2048- 2048-2048-2048-1036”;对于三字素声学模型,其 DNN 结构为“440-2048-2048-2048-2048-1038”。在预训练过程中,mini-batch 设置为 128,对于底层的高斯-伯努利以及其余 4 个伯努-伯努利均采用 5 个 epoch 训练。高斯-伯努利 RBM 的学习速率为0.005,伯努利-伯努利 RBMs 的学习速率为

0.08。在精细调整过程中,采用 BP 算法进行参数更新,学习速率初始设为 0.08,并根据校验集的性能改善,进行折半调整。

语言模型为使用开源工具包 SriLM[21] 构建的 2-gram 语言模型,具体实现方面,首先利用 3.1 节中两部分文本语料分别训练得到两个 bi-gram 语言模型,然后采用内插方法得到最终语言模型。内插权重通过使开发集标注文本的混淆度最低确定。

3.3 评测指标

采用 F-1 值作为集外词检测性能的衡量标准。F-1 取值在 0 和 1 之间,值越高表明检测性能越好。F-1 值通过召回率和准确率计算,其计算方法为

$$F\text{-}1=\frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

式中:Recall 为集外词的召回率,Precision 为集外词检测的准确率,其计算公式分别为

$$\text{Recall}=\frac{N_{\text{OOV-right}}}{N_{\text{OOV-ref}}}\times 100\%$$

$$\text{Precision}=\frac{N_{\text{OOV-right}}}{N_{\text{detected}}}\times 100\% \tag{6}$$

式中:$N_{\text{OOV-right}}$ 为正确检测集外词数,即检测集外词中正确的个数;$N_{\text{OOV-ref}}$ 为参考文本中给定的集外词个数;N_{detected} 为检测出的集外词个数。

最终识别性能采用词错率(Word error rates,WER)指标衡量,该指标越低则说明系统性能越好。

3.4 实验结果

表 2 给出了不同混合系统在采用不同声学模型下的集外词检测的准确率和召回率。在各个系统中,分别采用了 GMM 模型,STL-DNN 模型以及 MTL-DNN 模型进行声学模型建模。从实验结果来看,在单混合系统中基于 DNN 的混合系统的集外词召回率和准确率都要高于基于 GMM 模型的混合系统。究其原因,混合系统的集外词检测性能和系统的识别准确率密切相关,系统识别性能越好,则可以更加精确地将集外词表示成子词单元序列,从而集外词的检测性能也越高,基于 DNN 的混合系统在识别性能上要好于基于 GMM 的混合系统,故而具有更好的集外词检测性能。MTL-DNN 在 STL-DNN 的基础上提高了建模精度,进一步改善了识别性能,从而改善了集外词的检测性能。以开发集 WER 作为指标,利用格点搜索法确定 ROVER 重打分公式中的参数。可以看出,通过 ROVER 融合后,由于利用了不同系统间的互补性,准确率和召回率进一步提高,这说明了利用系统融合方法对提高集外词检测性能的有效性。

表 2 不同混合系统集外词检测的准确率和召回率

Tab. 2 OOV detection precision and recall rate for different hybrid systems %

系统名称		Trigrapheme 系统			Triphone 系统			ROVER 系统		
		GMM	STL-DNN	MTL-DNN	GMM	STL-DNN	MTL-DNN	GMM	STL-DNN	MTL-DNN
DEV	Recall	11.56	12.61	14.13	12.08	12.73	14.51	13.59	14.88	**15.21**
	Precision	44.83	57.14	58.59	48.25	56.38	56.78	57.37	**60.29**	59.07
TEST	Recall	11.41	11.25	12.32	10.19	11.16	13.07	12.01	12.77	**13.19**
	Precision	39.07	45.10	52.51	41.21	49.58	51.49	49.62	53.33	**55.50**

通过 F-1 值可以更直观地对比各系统的集外词检测性能,图 5 为根据集外词检测的召回率和准确率得到的 F-1 值。由图 5 可以看出,ROVER 融合系统的 F-1 值相比于单系统都有所提高。但是各系统

的集外词检测性能都不是特别理想,*F*-1 值都较低(0.2 左右),这主要受到电话语料识别性能较差的影响。通过提高系统的识别率可以进一步改善集外词检测性能。

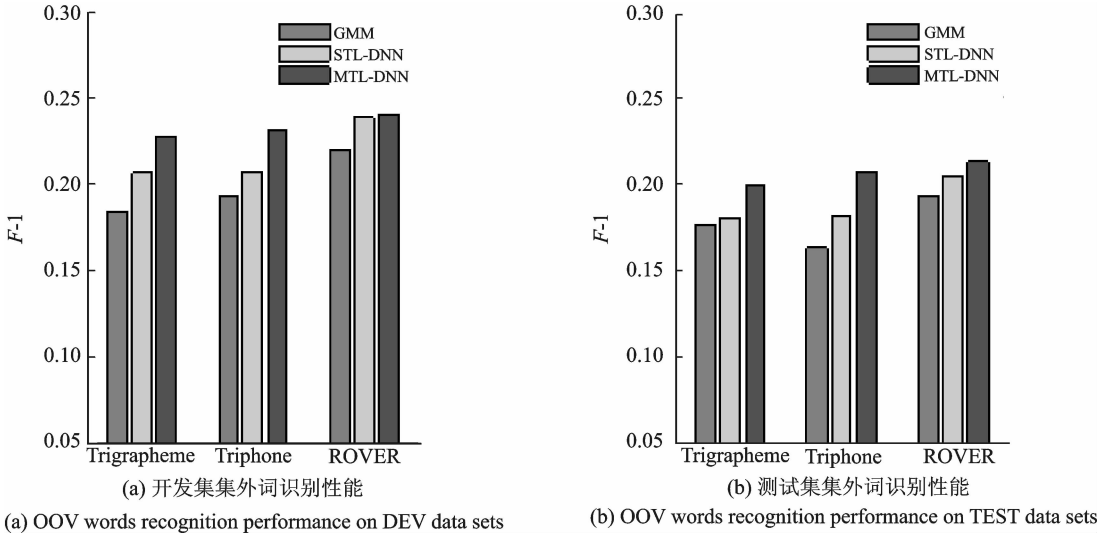

图 5 混合系统集外词识别

Fig. 5 OOV recognition based on hybrid system

表 3 给出了采用混合系统与非混合系统(采用一般字典和语言模型)进行解码得到的 WER 对比,从实验结果可以看出,混合系统(Hybrid)的识别性能普遍要高于非混合系统(Non-hybrid)的识别性能,一方面是因为非混合系统无法识别集外词,一般会把一个集外词识别成几个发音相近的集内词,而混合系统则可以将部分集外词正确识别出,同时也可以减少因为集外词错误识别而带来的插入错误;另一方面受语言模型影响,非混合系统中集外词的出现可能会引发临近集内词的识别错误,而采用混合语言模型则会减少这些识别错误。采用基于 MTL-DNN 的建模方法相比于 STL-DNN 的方法,会有 5%左右的识别性能提升,表明 MTL-DNN 模型具有更高的建模精度。Trigrapheme 系统和 Triphone 系统识别性能相近,这为两个系统的融合提供了条件。通过系统融合,WER 进一步降低,可以看出,基于 MTL-DNN 的融合方法具有最低的 WER,为 47.36%,相对 GMM 系统最佳性能(51.89%)降低了约8.7%,相对 STL-DNN 系统最佳性能(50.44%)降低了约 6.1%。

表 3 不同混合系统识别结果的词错率对比

Tab. 3 Comparison of WER for different hybrid systems %

系统名称		Trigrapheme 系统		Triphone 系统		ROVER 系统	
		DEV	TEST	DEV	TEST	DEV	TEST
Non-hybrid	GMM	53.99	53.26	53.34	52.96	52.96	52.12
	STL-DNN	51.78	51.06	51.44	50.72	51.33	50.76
	MTL-DNN	49.42	48.90	49.04	48.11	48.97	48.23
Hybrid	GMM	53.20	52.88	52.84	52.10	52.70	51.89
	STL-DNN	51.24	50.87	50.75	50.48	50.68	50.44
	MTL-DNN	48.70	48.27	48.32	47.86	48.10	47.36

为进一步研究混合模型以及系统融合方法性能提升原因,采用其他 4 个指标来具体分析。表 4 给出了上述系统中测试集的集外词和集内词错误率,其中 OOV 表示集外词的识别错误率,IV 表示集内词

的识别错误率，IV_{OOV}表示临近集外词的集内词识别错误率(由于采用 2-gram 语言模型，每个集外词前后一个集内词视为临近集内词)，$IV_{Non\text{-}OOV}$表示非临近集外词的集内词识别错误率。可以看出集外词的存在会对相邻集内词的识别产生很大影响，相比于非临近集外词的集内词，错误率要高 60%～70%左右(在文献[6]中，与集外词相邻的集内词错误率要比非临近集外词的集内词高一倍左右)。从整体来看，各系统的集外词识别错误率仍然较高，这主要受到系统整体识别性能的影响，系统识别性能越好则相应的集外词识别性能也会提高。在各混合系统中，基于 MTL-DNN 混合系统的集外词和集内词识别效果最好，进而系统的整体识别性能也相对较好。通过 ROVER 融合后，可以进一步降低集外词和集内词的识别错误率，从而使整体性能获得提升。

表 4 测试集各混合系统集外词与集内词识别错误率

Tab. 4 Misrecognition rates of test set for OOV words and IV words for different hybrid systems %

系统名称	建模方式	OOV	IV	IV_{OOV}	$IV_{Non\text{-}OOV}$
Trigrapheme	GMM	90.36	46.08	73.67	44.16
	STL-DNN	89.39	45.77	71.24	43.74
	MTL-DNN	88.92	43.92	69.64	41.32
Triphone	GMM	90.62	46.97	72.55	44.34
	STL-DNN	88.18	45.15	70.94	43.72
	MTL-DNN	**86.73**	42.89	**69.44**	40.21
ROVER	GMM	89.03	45.87	71.14	44.16
	STL-DNN	87.23	44.39	70.98	43.22
	MTL-DNN	86.76	**42.62**	**69.44**	**39.89**

4 结束语

本文提出了一种基于互补系统融合来提高混合模型语音识别系统的识别性能的方法。该方法通过采用不同的声学建模单元，构建了两套基于 HMM-DNN 的混合模型语音识别系统，并针对这两种识别任务之间的关联性，采用多任务学习思想，实现两个识别任务 DNN 网络的输入层和隐含层共享，从而提高建模精度。对套系统的输出结果，采用 ROVER 进行融合，进一步提升了系统的识别性能。实验结果表明，基于 MTL-DNN 的建模方式，可以显著提升系统的识别性能。基于 MTL-DNN 的系统融合方法获得了最低的 WER，相比于 STL-DNN 的融合方法降低了约 6.1%。

参考文献：

[1] Lin H, Bilmes J, Vergyri D, et al. OOV detection by joint word/phone lattice alignment[C]//Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Kyoto, Japan: IEEE, 2007: 478-483.

[2] Lecouteux B, Linares G, Favre B. Combined low level and high level features for out-of-vocabulary word detection[C]//Proceedings of International Speech Communication Association. Brighton, UK: [s. n.], 2009:1187-1190.

[3] Qin L. Learning out-of-vocabulary words in automatic speech recognition[D]. Pittsburgh: Carnegie Mellon University, Language Technologies Institute, 2013.

[4] Rastrow A, Sethy A, Ramabhadran B. A new method for OOV detection using hybrid word/fragment system[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Taipei, China: IEEE, 2009: 3953-956.

[5] Qin L, Sun M, Rudnicky A. OOV detection and recovery using hybrid models with different fragments[C]//Proceedings of International Speech Communication Association. Florence, Italy: [s. n.], 2011:1913-1916.

[6] Réveil B, Demuynck K, Martens J P. An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition [J]. Computer Speech and Language, 2014, 28(1): 141-162.

[7] He Y Z, Hutchinson B, Baumann P. Subword-based modeling for handling OOV words in keyword spotting[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2014: 7914-7918.

[8] Rao K, Peng F, Sak H. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks[C]//

Proceedings of International Conference on Acoustics, Speech and Signal Processing. Brisbane, Australia: IEEE, 2015: 4225-4229.

[9] Fiscus J G. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER) [C]//Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Santa Barbara, USA: IEEE, 1997: 347-354.

[10] Hoffmeister B. Bayes risk decoding and its application to system combination[D]. Aachen: RWTH Aachen University, Computer Science Institute, 2011.

[11] 戴礼荣,张仕良. 深度语音信号与信息处理:研究进展与展望[J]. 数据采集与处理,2014,29(2):171-179.
Dai Lirong, Zhang Shiliang. Deep speech signal and information processing: Research process and prospect[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 171-179.

[12] Chen N F, Ni C, Chen I F, et al. Low-resource keyword search strategies for Tamil[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Brisbane, Australia: IEEE, 2015: 5366-5370.

[13] Smit P, Virpioja S, Gronroos S, et al. Morfessor 2.0: Toolkit for statistical morphological segmentation[C]//Proceedings of European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: [s. n.], 2014:21-24.

[14] Chen D P, Mak B, Leung C C. Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 5629-5633.

[15] Pollak P, Hanzl V. Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool[C]//Proceedings of Language Resources and Evaluation. Las Palmas, Spain: [s. n.], 2002: 1264-1269.

[16] Mohan A, Rose R. Multi-lingual speech recognition with low-rank multi-task deep neural networks[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Brisbane, Australia: IEEE, 2015: 4994-4998.

[17] Matěj K, Ondřej P, Ondřej D, et al. Czech data [EB/OL]. https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-4670-6, 2014-02-21/2015-11-01.

[18] Petra G, Radovan G, Ondřej B. Czech-Slovak Parallel Corpus[EB/OL]. https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0006- AADF-0, 2012-05-15/2015-11-01.

[19] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]//Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Hawaii, USA: IEEE, 2011: 565-569.

[20] Miao Y. Kaldi+PDNN: Building DNN-based ASR systems with kaldi and PDNN [EB/OL]. http://www.cs.cmu.edu/~ymiao/kaldipdnn.html. 2014-11-01/2016-01-14.

[21] Stolcke A. SRILM—An extensible language modeling toolkit[C]// Proceedings of International Conference on Signal Processing. Beijing, China: [s. n.], 2002: 901-904.

作者简介:

范正光(1990-),男,硕士研究生,研究方向:语音识别、模式识别,Email:fanzg11@163.com.

屈丹(1974-),女,副教授,研究方向:语音识别、智能信息处理。

李华(1990-),女,硕士研究生,研究方向:语音识别、模式识别。

张文林(1982-),男,讲师,研究方向:语音信号处理、模式识别。