DOI：10.7652/xjtuxb201606012

借助音频数据的发音字典新词学习方法

范正光，屈丹，闫红刚，张文林
（解放军信息工程大学信息系统工程学院，450002，郑州）

摘要：针对已有的发音字典扩展方法只能从文本数据中学习新词而无法学习到音频数据中新词的问题，提出了一种基于混合语音识别系统的发音字典新词学习方法。该方法首先分别采用音节和字母音素对混合识别系统对音频数据进行集外词识别，利用系统间的互补性得到尽可能多的新词及其发音候选，然后借助感知器与最大熵模型对得到的新词及发音进行优化，降低错误率，最后实现发音字典的扩展，并利用语法语义信息完成对语言模型参数更新。基于华尔街日报（WSJ）语料库的连续语音识别实验表明：该方法可以有效学习到音频数据中的未知新词，采取的数据优化策略极大地提高了所得新词及发音的精度；在词错误率指标下，字典扩展后系统的识别性能相对基线系统提高约13.4%。

关键词：语音识别；发音字典；新词学习；集外词
中图分类号：TN912.3 **文献标志码**：A **文章编号**：0253-987X(2016)06-0075-08

Learning New Words for Pronunciation Lexicon from Audio Data

FAN Zhengguang, QU Dan, YAN Honggang, ZHANG Wenlin
(Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou 450002, China)

Abstract: A self-learning method of new pronunciation lexicons based on a hybrid speech recognition system is proposed to solve the problem that the existing self-expanding methods of pronunciation lexicons can only learn new words from text data but cannot learn from audio data. The method utilizes both the syllables and the graphones hybrid systems to recognize the out-of-vocabulary words in the audio data and then obtains as many new words with their pronunciations as possible by using the complementary information of the two systems. Then the new word and its pronunciation candidates are optimized using a perceptron model and a maximum entropy model to reduce the error rate. Finally, the lexicon is expanded and the language model parameters are updated by using syntactic and semantic information. Experimental results of continuous speech recognition on Wall Street Journal speech database show that the proposed method learns new words from audio data effectively, and the accuracy is greatly improved by using the data optimization strategies. The extended lexicon system yields a relative gain of 13.4% over the base line system in terms of word error rates.

Keywords: speech recognition; pronunciation lexicon; new words learning; out-of-vocabulary words

发音字典是搭建现代连续语音识别系统（continuous speech recognition, CSR）所必需的数据资

收稿日期：2016-01-16。 作者简介：范正光（1990—），男，硕士生；屈丹（通信作者），女，博士，副教授。 基金项目：国家自然科学基金资助项目（61175017，61403415，61302107）。
网络出版时间：2016-04-03 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20160403.1846.010.html

源,但传统的发音字典由语言学专家手动生成,需要花费较高的成本。针对这一问题,当前普遍采用发音字典自动学习来减小人工工作量。目前,常用的字典自动学习方法主要有 2 类:基于字母音素转换(grapheme to phoneme conversion,G2P)的方法[1-3]和基于网络爬取的方法[4]。基于 G2P 转换的方法是指通过对文本语料(如爬取的网络文本语料)进行统计发现新词,然后利用 G2P 转换获取这些新词的发音。常用的 G2P 转换方法有基于联合序列模型的方法[2]、基于神经网络模型的方法[3]等。基于网络爬取的方法可以认为是第一类方法的特例,该方法通过爬取一些特殊的网页(如维基字典等),直接获取带有发音的新词,从而避免了 G2P 转换带来的错误,保证了获取新词及发音的准确性。借助文本语料的发音字典扩展具有实现简单的优点,但文本语料往往存在较多的错误,如拼写错误等,这些错误会增加发音字典的混淆度进而影响识别性能[5]。此外,当文本语料较少时,该方法发现的新词数量也有限。

随着网络技术的发展,音频数据越来越成为一种较易获取的数据资源。音频数据中也会存在很多的新词,并且这些新词不在发音字典中,传统的语音识别系统无法识别。这些新词被称为集外(out-of-vocabulary, OOV)词、集内(in-vocabulary, IV)词。为了识别集外词,文献[6-9]采用不同的子词单元构建词/子词混合语音识别系统。该混合系统在解码时将集外词表示成一些被称为子词的语音单元序列,进而利用这些子词序列实现集外词的识别。混合语音识别方法虽然可以识别集外词,但在识别时同样会将部分置信度较低的集内词识别成子词形式,从而影响识别性能。此外该方法解码复杂度较高,限制了其在实际中的应用。

综合上述方法,针对音频数据中的新词,本文提出一种新的基于混合语音识别系统的发音字典新词学习方法。该方法利用混合语音识别系统的识别结果提取集外词和发音,并借助感知器以及最大熵模型等对这些新词及发音进行优化以降低错误率;针对现有的混合语音识别系统集外词召回率低,采用多个混合系统进行融合以提高新词发现率;最后提出了基于语法语义的语言模型参数估计方法。实验表明,新方法可以有效发现音频数据中的新词,采用扩展后的字典,系统性能相对基线系统也有了较大提升。

1 词/子词混合语音识别系统

图 1 给出了混合语音识别系统框图。混合语音识别系统与传统语音识别系统的主要区别在于可以对集外词进行识别。在识别时,混合语音识别系统首先采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(phones)、字母音素对(graphones)以及词素(morphemes)等子词形式。通过对混合识别结果进行处理,从而得到最终词级识别结果。

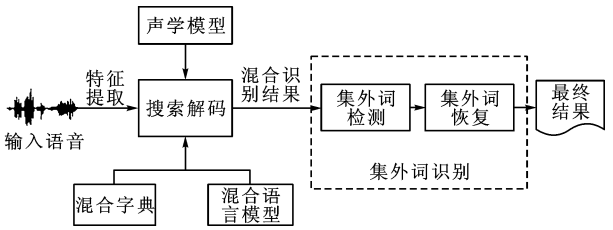


图 1 混合语音识别系统框架

1.1 混合字典

混合字典包含词和子词 2 种不同类型的语音单元,子词用于解码时表示集外词。本文讨论音节和字母音素对 2 种类型的子词。其中,音节是由一个或几个音素按一定规律组合而成的语音单位;字母音素对是在训练联合序列模型字母音素转换器时得到的,为字母序列和发音序列间的映射。本文分别使用 Festival 词典工具[10]以及 Sequitur G2P 工具[2]获取这 2 种子词。所有子词均加入词边界标记,结尾子词标记为“#”,非结尾子词标记为“+”。引入词边界标记虽然增加了子词单元数量,但使集外词的恢复变得更加简单。

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序列得到混合语料。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。训练好的混合语言模型,通过设置集外词插入惩罚因子 P_{OOV} 可以控制解码时子词单元出现的比例。如对于训练得到的语言模型参数 $p_s(s_1 \mid w_1 w_2)$,调整后的参数为 $p_t(s_1 \mid w_1 w_2) = p_{\mathrm{OOV}} \cdot p_s(s_1 \mid w_1 w_2)$,其中 s_1 为子词,w_1、w_2 为词。采用该混合语言模型进行解码,即得到混合识别结果。

1.3 集外词识别

集外词识别模块包括集外词检测和集外词恢复 2 部分。集外词检测用于通过混合识别结果,确定

集外词的位置（在混合识别结果中，子词序列出现的位置则表示集外词位置），而集外词恢复则是为了获得集外词的正确拼写。

针对集外词恢复，不同的子词有不同的恢复方法。字母音素对本身包含了单词的拼写形式可以直接用于集外词的恢复。采用音节作为子词单元时，往往先根据音节序列获取集外词的音素序列，然后通过音素字母转换（P2G）得到。图 2 给出了一个集外词识别示例，对混合解码器得到的音节混合识别结果，首先通过音节序列确定集外词位置，然后根据该序列以及词边界标记确定集外词的发音，最后经过音素字母转换获得集外词识别结果。

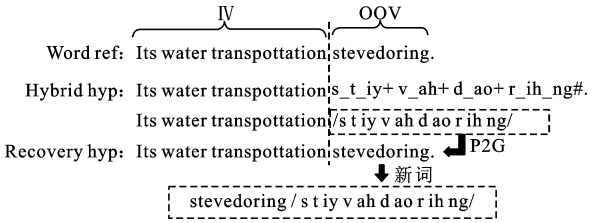


图 2　集外词识别以及新词学习示例

2　基于混合语音识别系统的发音字典新词学习

混合语音识别系统具有可以识别集外词的优点，其识别得到的集外词即为新词（如图 2 所示）。由于在识别以及音素字母转换中都可能存在一些错误，直接恢复得到的集外词及发音准确率较低，为此本文对识别得到的集外词及发音进行优化以降低错误率。同时，针对混合语音识别系统集外词召回率低的缺点，采用多个混合系统来提高新词的发现率。整个字典学习流程如图 3 所示，对于给定的音频数据，首先采用多种子词单元混合系统（本文只讨论音节混合系统和字母音素对混合系统）进行集外词识别；然后对获取的集外词及发音进行优化，降低错误率；最后将筛选结果加入发音字典中，并完成字典及语言模型参数更新。

对于获取的新词（即集外词）及发音，本文采取的优化措施归纳如下：

（1）对得到的新词，首先进行过滤去除集内词，这主要考虑到采用混合系统解码时引入的一些虚警错误；

（2）2 个不同混合系统得到的相同的新词及发音，认为可信度较大，从而直接判为正确新词；

（3）根据不同混合系统获得的新词及发音，确定不同的代价函数，并通过设定不同的门限进行筛选，将筛选结果扩充到发音字典中。

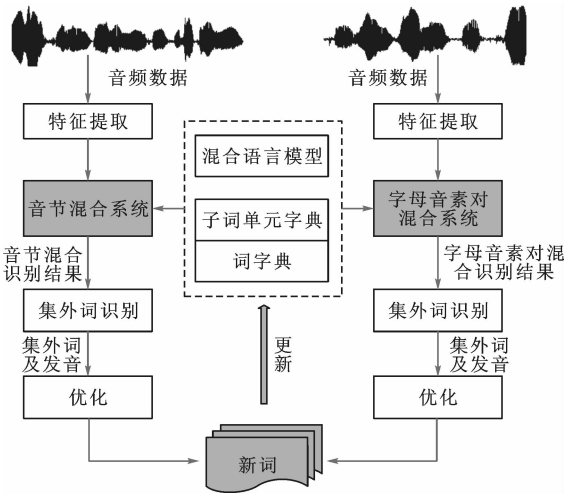


图 3　基于混合语音识别系统的字典新词学习流程

2.1　代价函数的确定

导致学习到的新词及其发音错误的原因主要有 2 个，一是识别错误，即混合识别结果中存在识别错误的子词序列，二是恢复错误，主要是在进行集外词恢复时导致的错误。因此，代价函数应包含对这 2 种错误的评估。根据在进行集外词恢复时是否需要进行 P2G 转换，本文确定了 2 种类型的代价函数，一种是基于感知器模型的代价函数，一种是基于最大熵模型的代价函数。

2.1.1　*针对音节混合系统的代价函数*　基于音节的混合系统，在进行集外词恢复时需要进行 P2G 转换。对于获取的新词及发音，借助感知器模型[11]的思想构造代价函数。首先计算多种特征值的线性加权和，即

$$g(s)=\boldsymbol{\alpha}\boldsymbol{f}(s)=\alpha_0+\alpha_1 f_1(s)+\alpha_2 f_2(s)+\alpha_3 f_3(s) \tag{1}$$

式中：s 为解码得到的音节序列；$\boldsymbol{\alpha}=[\alpha_0,\alpha_1,\alpha_2,\alpha_3]$ 为特征权重；$f_1(s)$ 为该音节序列的声学模型得分（置信度得分），是解码得到的音节序列中各音节声学模型得分的乘积，定义为

$$f_1(s)=\prod_i s_{\mathrm{AM}}(i) \tag{2}$$

其中 $s_{\mathrm{AM}}(i)$ 为第 i 个音节的声学模型得分；$f_2(s)$ 为语言模型得分，通过将词表中的单词表示成音节，从而训练得到音节语言模型并计算音节序列的得分；$f_3(s)$ 为 P2G 转换得分，由 P2G 转换工具得到。由于 $g(s)$ 是线性的，采用 Sigmoid 函数进一步将实数域上的 $g(s)$ 映射为 0 到 1，得到最终代价函数

$$\varphi(s)=\frac{1}{1+\exp(-g(s))} \tag{3}$$

对于权重 α_i，采用感知器算法进行学习，首先，对 $\varphi(s)$求导

$$\varphi'(s)\mid_{g(s)}=\varphi(s)(1-\varphi(s)) \tag{4}$$

其次，令 $d(s)$代表训练样本的正确分类，定义为

$$d(s)=\begin{cases}1, & \text{新词正确}\\ 0, & \text{其他}\end{cases} \tag{5}$$

最后，根据训练样本对权值进行迭代训练，迭代公式如下

$$\boldsymbol{\alpha}=\boldsymbol{\alpha}+\eta\varphi'(s)(\mathrm{d}(s)-\varphi(s))\boldsymbol{f}(s) \tag{6}$$

式中：η 为训练步长，本文选取固定的 η 为 1。

2.1.2 针对字母音素对混合系统的代价函数 基于字母音素对的混合系统，进行集外词恢复时不需要进行 P2G 转换，对此本文采用最大熵模型(Maximum Entropy，ME)[12]确定代价函数

$$\varphi(y\mid s)=\frac{1}{Z(s)}\exp\Big[\sum_{i=1}^{k}\lambda_i f(s,y)\Big] \tag{7}$$

$$Z(s)=\sum_{y}p(y\mid s)=\sum_{y}\exp\Big[\sum_{i=1}^{k}\lambda_i f(s,y)\Big] \tag{8}$$

式中：y 为分类标签，结果属于集合{RIGHT，WRONG}；s 为获取的字母音素对序列；$f(s,y)$为特征函数，是一个二值函数；k 为特征函数的个数；λ_i 为权重；$Z(s)$为归一化因子。

在最大熵模型中，关键是要选取合适的特征，对于得到的新词及发音，判定其正确与否的因素有该词包含的字母音素对个数、字母音素对序列的声学模型以及语言模型得分等。根据这些因素，建立特征模板，并根据训练集数据定义每个模板取值范围，如表 1 所示，模板 1～5 是决定新词是否正确的特征模板，模板 6 为一个特殊模板，表示判定结果。在表 1 定义的特征模板中，模板 2 用于判断字母音素对序列中是否含有字母音素对语言模型的二元和三元条目，目的在于确定该字母音素对序列是否符合单词的构成规则。模板 4 和 5 的定义与 2.1.1 节中的定义相似，在获取声学模型得分与语言模型得分后，计算所有得分的均值 μ 和方差 σ，并由此确定阈值 $T=\mu+\sigma$。

当模板函数取特定值时，该模板被实例化，得到具体特征。取 1～5 号中任一模板，确定模板取值，并结合当前判定结果的值(即 DEFAULT 的值)，就可以产生一个特征。定义特征格式为 $A-B=C$，其中 A 为特征模板为对新词判定时需要考虑的因素；B 为该特征模板的取值；C 为模板 DEFAULT 的取值，表示判定结果。

例如由模板 1 可以确定一个特征 ENDTAG－＃＝RIGHT，表示为二值特征函数

$$f(s,y)=\begin{cases}1, & \text{ENDTAG}=\#,y=\text{RIGHT}\\ 0, & \text{其他}\end{cases} \tag{9}$$

该特征函数表示如果新词对应的子词单元序列中最后一个子词单元的结尾标记为“＃”，并且该新词正确，则函数值为 1，否则为 0。确定特征集合后，通过训练数据(Dev93 开发集)进行参数估计。

2.2 语言模型参数的估计

加入字典中的新词及发音，只有在语言模型中包含其相关的参数，才能被识别系统正确识别。针对该问题，可以采用较大的语言模型训练语料，对语言模型进行重新训练，但在缺少训练所需的语料时，这些参数便无法通过最大似然估计有效获取。为此，本文利用语法以及语义信息来实现这些参数的估计，该方法的主要步骤如下。

步骤 1 估计新词的 unigram 参数。采用 Stanford MaxEnt POS[13]对包含新词的识别结果进行词性标注，获取新词及其上下文单词的词性信息。假设 w_i 为加入到字典中的新词，l_i 为其标注(即词性)，则该词的 unigram 得分可以表示为

$$p(w_i)=\sum_{l_i}p(w_i\mid l_i)p(l_i) \tag{10}$$

式中：$p(l_i)$是标注 l_i 的先验概率；$p(w_i\mid l_i)$为从标注为 l_i 的所有单词中观测到新词 w_i 的概率，采用下式进行估计

$$p(w_i\mid l_i)=\frac{1}{N} \tag{11}$$

表 1 特征模板及取值范围

特征模板号	模板意义	模板名称	取值范围
1	最后一个子词单元结尾标记	ENDTAG	{＃，＋}
2	包含 N-gram 字母音素对类型	LMITEM	{2GRAM，3GRAM，BOTH，OTHER}
3	字母音素对子词单元个数	NUM	{Int{1～5}，OTHER}
4	声学模型得分是否大于阈值	LMSCORE	{YES，NO}
5	语言模型得分是否大于阈值	AMSCORE	{YES，NO}
6	判定结果	DEFAULT	{RIGHT，WRONG}

其中 N 为训练集中标记为 l_i 的集内词的个数。

步骤 2　估计新词的 bigram 以及 trigram 参数。参照步骤 1，对 2 种参数的计算分别如式(12)和式(13)所示

$$p(w_i \mid w_{i-1}) = \sum_{l_i} p(w_i \mid l_i) p(l_i \mid l_{i-1}) \tag{12}$$

$$p(w_i \mid w_{i-1}, w_{i-2}) = \sum_{l_i} p(w_i \mid l_i) p(l_i \mid l_{i-1}, l_{i-2}) \tag{13}$$

式中：l_{i-1} 和 l_{i-2} 分别为第 $i-1$ 和第 $i-2$ 个位置的单词的标注。

步骤 3　借助 WordNet[14] 获取更多的语言模型参数。采用词性信息获取的新词语言模型参数数量较少，在真实条件下得到的新词可能出现在不同语境中。对于得到的新词，首先利用 WordNet 获取与该词具有相似语义的集内词(即同义集内词)；然后获得这些集内词的 bigram 以及 trigram 语言模型参数，并将这些参数中的集内词用相应的新词进行替换，从而得到更多的语言模型参数。

3　实验结果和分析

3.1　实验数据

选用华尔街日报(Wall Street Journal，WSJ)语料库作为实验语料库，其中声学模型训练集由 WSJ0 和 WSJ1 中的 37 416 句话构成，包含 284 个说话人，共约 80 h。选用 WSJ Dev93 开发集，用于新词优化中代价函数参数的训练。选用 WSJ Eval93 和 WSJ Eval92 测试集，分别用于优化过程中门限值的确定以及最终测试集。语言模型训练数据采用 WSJ 87-89 文本数据，大小约 215 MB。对上述文本进行统计得到出现频率最高的 2×10^4 个单词，并通过 CMUdict[15] 获取发音，构造发音字典。表 2 给出了采用该发音字典时不同数据集中集外词数量以及所占比例。

表 2　各数据集中集外词所占比例

字典规模	集外词数			集外词比例/%		
	Dev93	Eval92	Eval93	Dev93	Eval92	Eval93
2×10^4	239	142	117	2.54	1.99	2.00

3.2　实验设置

实验主要基于开源工具包 Kaldi 搭建。声学特征采用 13 维的 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。声学模型采用最大似然估计(MLE)方法得到，为包含 3 个发射状态的、自左向右无跨越的 3 音子 HMM 模型。采用基于决策树的三音子状态聚类，得到 3 285个不同的上下文相关状态，模型中总的高斯混元数为 2×10^4。所有的语言模型为 3-gram 语言模型。集外词插入惩罚因子 P_{OOV} 设置为 0 到 5.5，步长 0.5。

3.3　评测指标

集外词检测中常用的衡量指标为虚警概率 P_{fa} 和漏检概率 P_{miss}，定义为

$$P_{\mathrm{fa}} = \frac{N_{\mathrm{fa}}}{N_{\mathrm{IV-ref}}} \times 100\%$$

$$P_{\mathrm{miss}} = \frac{N_{\mathrm{miss}}}{N_{\mathrm{OOV-ref}}} \times 100\% \tag{14}$$

式中：N_{fa} 为虚警数，即检测集外词中包含的集内词个数；$N_{\mathrm{IV-ref}}$ 为参考文本中给定的集内词数量；N_{miss} 为漏检数，即未检测出的集外词个数；$N_{\mathrm{OOV-ref}}$ 为参考文本中给定的集外词个数。在虚警率和漏检率的基础上，可以通过检测错误折衷(detection error trade-off，DET)作为系统性能评价指标，曲线越靠近坐标原点则系统性能越好。

学习到的新词通过准确率 P_{ac} 和召回率 P_{re} 衡量，定义为

$$P_{\mathrm{ac}} = \frac{N_{\mathrm{right}}}{N_{\mathrm{filtered}}} \times 100\%$$

$$P_{\mathrm{re}} = \frac{N_{\mathrm{right}}}{N_{\mathrm{ref}}} \times 100\% \tag{15}$$

式中：N_{right} 表示筛选结果中发音正确的新词个数；N_{filtered} 为筛选后总的新词个数；N_{ref} 为音频数据中总的新词个数。此外，本文也采用综合这两者的 F 值来衡量新词学习性能

$$F = \frac{2P_{\mathrm{ac}}P_{\mathrm{re}}}{P_{\mathrm{ac}} + P_{\mathrm{re}}} \tag{16}$$

3.4　实验结果

本文建立了 3 套语音识别系统分别用于新词学习以及发音字典扩展前后识别性能的比较：①Base_20k 系统为词表大小为 2×10^4 的传统语音识别系统；②Hybrid_syllbale 系统为采用音节作为子词单元的混合语音识别系统；③Hybrid_graphone 系统为采用字母音素对作为子词单元的混合语音识别系统。

3.4.1　**基于混合系统的集外词识别**　进行集外词检测时，根据处理后的混合识别结果，子词单元出现的区域可以认为是集外词区域。图 4 是通过设置不同的集外词插入惩罚因子 P_{OOV} 对 Eval92 测试集得到的不同系统的集外词检测 DET 性能曲线。从图中可以看出，音节混合系统以及字母音素对混合系

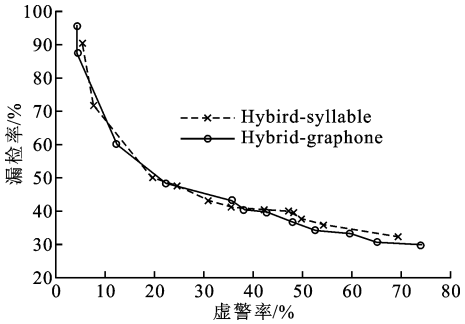
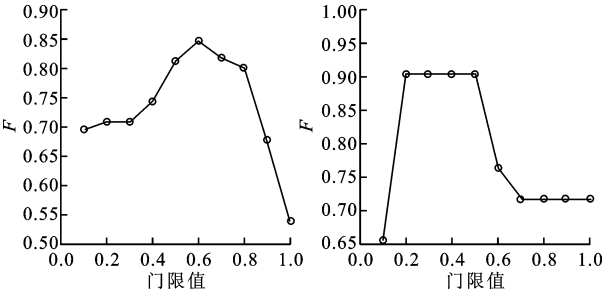


图 4 2 种系统的集外词检测性能



(a)音节混合系统 (b)字母音素对混合系统

图 5 不同门限值对新词优化的影响

统在集外词检测方面具有相近的性能,但是由于不同子词单元具有不同的特性,从而使得输出结果存在一定的互补性。

表 3 为 $P_{OOV}=1$ 的情况下,2 个混合系统的集外词检测与恢复比率(即正确检测集外词个数与正确恢复集外词个数占参考中总集外词数量的比例)。可以看出,虽然 2 个混合系统有超过一半的集外词被正确检测出,但是最终正确恢复得到的集外词仅有 30%左右,这说明即便识别音频中的新词被检测到,但由于识别得到的子词以及在恢复过程中都可能存在错误,从而导致学习到的新词以及发音的错误。这些错误加入到发音字典中,会降低字典的质量,从而对识别性能造成影响。将 2 个系统的识别结果进行融合,可以发现集外词检测以及恢复比率都有提升,从而使得学习到新词的概率大大增加。

表 3 不同系统的集外词检测与恢复比率

混合语音识别系统	检测比率/%	恢复比率/%
Hybrid_syllable 系统	55.77	31.65
Hybrid_graphone 系统	58.26	35.47
融合系统	74.23	46.00

3.4.2 *新词及发音优化* 音节混合系统以及字母音素对混合系统采用不同的代价函数进行新词优化,需要确定合理的门限值,以获取最优的系统性能。图 5 是 Eval93 测试集在不同门限值下采用不同混合系统经过筛选后的新词及其发音的 F 值。

由图 5 可以看出,2 个系统只有在选择合理门限的情况下,才能获得更好的筛选结果。如果门限值过低,则筛选后的结果中会存在较多的错误集外词及发音。门限值过高时,虽然可以保证较高的准确度,但是同样会使一些正确的集外词被过滤。根据图中结果,本文对于音节混合系统采用门限值为 0.6,对字母音素对混合系统采用门限值为 0.5。表 4 是在上述门限下,对 Eval92 测试集获取的新词及其发音筛选前后的准确率和召回率,其中,graphones NWs 表示字母音素对混合系统得到的新词及发音,syllable NWs 表示音节混合系统得到的新词及发音,same NWs 为 2 个混合系统中相同的新词及发音,all 为对上述得到的 3 类新词进行融合。可以看出优化后,新词的准确率获得了较大的提升,2 个混合系统中相同的新词具有最高的准确率 86.96%。但是,通过筛选也会使部分正确的新词及其发音被过滤掉,导致召回率下降。将 3 种筛选方式得到的结果合并到一起,可以看出准确率要略微下降。其原因在于,3 种筛选方式中可能存在不同错误的新词。但是,通过合并利用了不同系统间的互补性,召回率明显提高,此时的召回率已与优化前各单系统的召回率相当,但准确率明显高于各单系统。此外,扩展后的发音字典可以通过人工筛选来进一步提高准确率。

在运算量方面,2 种混合系统均受数据量以及数据集中集外词比例的影响。相比于音节混合系统,字母音素对混合系统采用的代价函数更为复杂,且提取的特征数量较多,但是不需要进行 P2G 转换,从实验过程中的时间消耗来看,2 个系统具有相近的运算效率。

表 4 优化前后新词及发音准确率和召回率对比

新词获取方式	准确率/%	召回率/%
graphones NWs	80.77	33.90
syllable NWs	76.00	32.20
same NWs	86.96	33.89
all	72.34	55.93

3.4.3 *扩展发音字典及语言模型在连续语音识别中的应用* 为了验证本文方法的有效性,在 Base_20k 系统的基础上,将学习到的新词加入 2×10^4 字典中,分别采用 WSJ 语言模型训练语料以及 2.2 节中所述的语言模型参数更新方法对语言模型参数进行更新,并与 Eval92 测试集的识别性能进行对比。

图 6 给出了对表 4 中 4 种不同方式得到的新词采用 3 种语言模型参数更新方法的识别性能对比。其中，WSJ-corpus LM 为采用 WSJ 语言模型训练语料重新训练的语言模型，Syntactic LM 为仅采用语法信息更新参数后的语言模型，Syntactic＋Semantic LM 为采用语法语义信息更新参数后的语言模型。可以看出，采用扩展后的字典，各系统的识别错误率相比基线系统(Base_20k)都有较为明显的下降，其中采用 2 个系统融合得到的优化新词(All)，语言模型采用 Syntactic＋Semantic LM 时的词错误率最低(7.55%)，相对基线系统的 8.72%的词错误率，降低约 13.4%。采用 WordNet 加入语义信息更新语言模型参数后，系统的识别性能并没有比单采用语法信息提高太多，这是因为虽然利用语义信息获得了更多的新词语言模型参数，但这些加入的 bigram 以及 trigram 参数，并没有出现在测试集中，但当面对新的识别任务时，加入字典的新词就可能出现一些新的上下文情况，单靠语法信息获得的语言模型参数，是无法预测这些情况的。从图中还可以看出，采用 WSJ 语言模型训练语料重新训练的语言模型，与 2.2 节的语言模型参数更新方法获得了相近的识别性能，这也验证了本文语言模型参数更新方法的有效性。但是，重新训练的语言模型可以更好地应对一些未知情况，因此采用语法语义信息进行语言模型参数的更新更多的只用在缺少语言模型训练语料时。

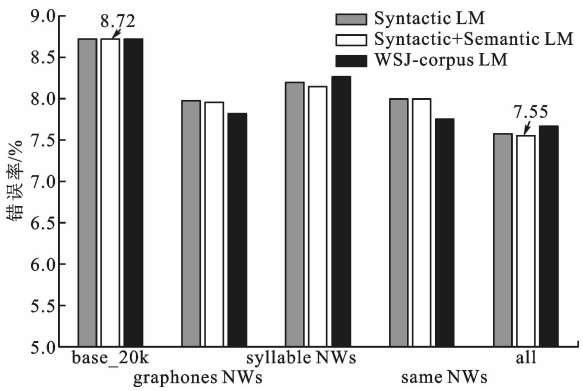


图 6　3 种语言模型参数更新方法对系统性能的影响

4　结　论

本文提出了一种针对音频数据的字典新词学习方法，可以作为现有的利用文本数据进行字典新词学习的补充。该方法利用多套混合语音识别系统进行新词学习，并通过一定的数据优化策略来提高新词的发现率以及准确率。同时，针对语言模型，采用语法语义信息完成对新词语言模型参数的更新。相关实验结果表明，本文方法能有效发现音频数据中的新词，选取的数据优化策略极大地提高了加入字典中的新词及发音的精度。

参考文献：

[1] DAVEL M, MARTIROSIAN O. Pronunciation dictionnary development in resource-scarce environments [C]// Proceedings of International Speech Communication Association. Grenoble, France: ISCA, 2009: 2851-2854.

[2] BISANI M, NEY H. Joint-sequence models for grapheme-to-phoneme conversion [J]. Speech Communication, 2008, 50(5): 434-451.

[3] RAO K, PENG F, SAK H, et al. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks [C]// Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 2015: 4225-4229.

[4] TIM S, OCHS S, TANJA S. Web-based tools and methods for rapid pronunciation dictionary creation [J]. Speech Communication, 2014, 56(1): 101-118.

[5] BERT R, KRIS D, MARTENS J. 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.

[6] 郑铁然，韩纪庆，李海洋. 基于词片的语言模型及在汉语语音检索中的应用 [J]. 通信学报，2009，30(3)：84-88.
ZHENG Tieran, HAN Jiqing, LI Haiyang. Study on performance optimization for Chinese speech retrieval [J]. Journal on Communications, 2009, 30(3): 84-88.

[7] HE Y Z, BRIAN H, PRTER B. Subword-based modeling for handling OOV words in keyword spotting [C]// Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ, USA: IEEE, 2014: 7914-7918.

[8] QIN L, RUDNICKY A I. OOV word detection using hybrid models with mixed types of fragments [C]// Proceedings of International Speech Communication Association. Grenoble, France: ISCA, 2012: 2450-2453.

[9] BASHA S, AMR M, HAHN S. Improved strategies for a zero OOV rate LVCSR system [C]// Proceedings of International Conference on Acoustics, Speech, and

Signal Processing. Piscataway, NJ, USA: IEEE, 2015: 5048-5052.

[10] BLACK A W, TAYLOR P, CALEY R. The festival speech synthesis system [EB/OL]. (2002-12-27) [2016-01-04]. http://www.festvox.org/docs/manual-1.4.3/.

[11] 韩冰, 刘一佳, 车万翔. 基于感知器的中文分词增量训练方法研究 [J]. 中文信息学报, 2015, 29(5): 49-54.
HAN Bing, LIU Yijia, CHE Wanxiang. An incremental learning scheme for perceptron based Chinese segmentation [J]. Journal of Chinese Information, 2015, 29(5): 49-54.

[12] 李素建, 王厚峰, 俞士汶. 关键词自动标引的最大熵模型应用研究 [J]. 计算机学报, 2004, 27(9): 1192-1197.
LI Sujian, WANG Houfeng, YU Shiwen. Research on maximum entropy model for keyword indexing [J]. Chinese Journal of Computers, 2004, 27(9): 1192-1197.

[13] KLEIN D, MANNING C. Feature-rich part-of-speech tagging with a cyclic dependency network [C]// Proceedings of Human Language Technology and North American Chapter of the Association for Computational Linguistics. Cambridge, MA, USA: ACL, 2003: 252-259.

[14] MILLER G. WordNet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[15] Carnegie Mellon University. The CMU pronunciation dictionary [EB/OL]. (2007-03-19)[2016-01-04]. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

(编辑 刘杨)

(上接第 14 页)

[13] AGNETIS A, BILLAUT J, GAWIEJNOWICZ S, et al. Multiagent scheduling-models and algorithms [M]. Berlin, Germany: Springer, 2014: 6-16.

[14] YIN Y, CHENG T C E, WAN L, et al. Two-agent sing-le-machine scheduling with deteriorating jobs [J]. Computers & Industrial Engineering, 2015, 81: 177-185.

[15] JIN Y C. Minimizing total weighted completion time under makespan constraint for two-agent scheduleing with job-dependent aging effects [J]. Computers & Industrial Engineering, 2015, 83: 237-243.

[本刊相关文献链接]

董春云,蔡远利.高超声速再入飞行器轨迹优化算法评估策略.2016,50(4):28-34.[doi:10.7652/xjtuxb201604005]

李国梁,张合新,周鑫,等.状态反馈事件触发控制系统的分析与设计.2016,50(5):146-150.[doi:10.7652/xjtuxb201605022]

李小虎,柳松玮,施虎,等.基于两自由度 H∞控制策略的泵控液压系统研究.2016,50(4):7-13.[doi:10.7652/xjtuxb201604002]

杨航,刘凌,阎治安,等.双闭环 Buck 变换器系统模糊 PID 控制.2016,50(4):35-40.[doi:10.7652/xjtuxb201604006]

宋汝君,单小彪,李晋哲,等.压电俘能器涡激振动俘能的建模与实验研究.2016,50(2):55-60.[doi:10.7652/xjtuxb201602010]

严惠云,张浩磊,刘小民.一种仿生鱼体自主游动的水动力学特性分析.2016,50(2):138-144.[doi:10.7652/xjtuxb201602023]

黄博,苑寿同,薛嘉伦.碳纤维气流扰动展纤器展纤过程仿真与实验.2015,49(12):19-25.[doi:10.7652/xjtuxb201512004]

陈江城,张小栋,李睿,等.利用表面肌电信号的下肢动态关节力矩预测模型.2015,49(12):26-33.[doi:10.7652/xjtuxb201512005]

连峰,王婷婷,韩崇昭.多个不可分辨目标群的联合检测与估计误差界.2015,49(11):89-95.[doi:10.7652/xjtuxb201511015]

张蕾,张爱民,景军锋,等.静止无功补偿器与发电机励磁系统的自适应鲁棒协调控制策略 2015,49(11):96-101.[doi:10.7652/xjtuxb201511016]

王海龙,王刚,陈曦,等.仿海蟹机器人游泳足水动力学分析与实验研究.2015,49(8):75-83.[doi:10.7652/xjtuxb201508013]

刘冠初,熊静琪,乔林,等.四足机器人自由步态规划建模与算法实现.2015,49(6):84-89.[doi:10.7652/xjtuxb201506014]

刘冠初,熊静琪,乔林,等.四足机器人自由步态规划建模与算法实现.2015,49(6):84-89.[doi:10.7652/xjtuxb201506014]

董恩清,刘伟,宋洋.采用三角形节点块处理无线传感器网络节点定位中节点翻转歧义的迭代方法.2015,49(4):84-90.[doi:10.7652/xjtuxb201504014]

卜祥伟,吴晓燕,张蕊,等.双曲正弦非线性跟踪微分器设计.2015,49(1):107-111.[doi:10.7652/xjtuxb201501018]

(编辑 刘杨)