基于 JSM 和 MLP 改进发音错误检测的方法

袁 桦 [1] 史永哲 [1] 赵军红 [2,3] 刘 加 [1]

摘 要 针对发音错误检测的发音字典生成提出基于联合序列多阶模型 (Joint-sequence multi-gram, JSM) 和多层神经感知 (Multi-layer perception, MLP) 的方法. 首先使用 JSM 模型对发音错误进行建模, 将标准发音和错误发音组合为发音对, 表示它们之间的对应关系, 再使用 N 元文法来统计各发音对之间的关系, 描述错误发音对上下文关系的依赖. 最后使用 MLP 对发音对之间的关系进行重新建模, 以学习到在相似的上下文条件下发生的相似的错误. 实验证明使用 MLP 对高阶模型进行概率重估能有效的平滑概率空间, 提高了发音错误检测的性能.

关键词 发音错误检测, 联合序列多阶模型, 多层神经感知, 计算机辅助语言学习

引用格式 袁桦, 史永哲, 赵军红, 刘加. 基于 JSM 和 MLP 改进发音错误检测的方法. 自动化学报, 2014, **40**(12): 2815−2823

DOI 10.3724/SP.J.1004.2014.02815

Improved Mispronunciation Detection Based on JSM and MLP

YUAN Hua[1] SHI Yong-Zhe[1] ZHAO Jun-Hong[2,3] LIU Jia[1]

Abstract In this paper, we propose a method of dictionary generation based on joint-sequence multi-gram model (JSM) and multi-layer perception (MLP) for mispronunciation detection. The JSM model is firstly used to model the mispronunciation. The canonical pronunciation and mispronunciation are combined into pronunciation pairs for representation of their corresponding relationship; then the N-gram is used to count the relationship between pronunciation pairs to describe the dependence of mispronunciations on the context. Lastly, the MLP is used to model the relationship of pronunciation pairs again, in order to capture the similar mispronunciations occurred in similar contexts. Experiments show that rescoring the probability of high-order model by MLP can effectively smooth the probability, resulting in improved mispronunciation detection.

Key words Mispronunciation detection, joint-sequence multi-gram model (JSM), multi-layer perception (MLP), computer-assisted language learning (CALL)

Citation Yuan Hua, Shi Yong-Zhe, Zhao Jun-Hong, Liu Jia. Improved mispronunciation detection based on JSM and MLP. *Acta Automatica Sinica*, 2014, **40**(12): 2815−2823

语音信号处理和自然语言处理等技术的高速发展, 推动着计算机辅助语言学习 (Computer-assisted language learning, CALL) 向着智能化、多元化和人性化发展[1−3]. 自动发音错误检测作为发音教学中的重要组成部分, 其目的是检测出学习者发音中的孤立错误. 目前的检测方法主要分为两类: 一类是使用置信度分数来判断标准文本中的每个发音是否正确[4−6]; 另一类是通过错误预测的方法检测出实际的发音串[7−9]. 由于第二类方法比第一类方法可以得到更明确的错误类型, 也就能给学习者提供更多纠正式的反馈信息, 得到了越来越多研究者的关注.

基于错误预测的方法需要首先对错误规律进行建模, 可以使用专家定义或者数据驱动的方法. 由于数据驱动的方法有很好的推广性, 所以被广泛采用. 目前采用的方法有: 1) 以上下文相关的发音规则对错误发音进行建模[8−12], 使用动态规划方法将标准发音串和错误发音串对齐后直接进行统计; 2) 基于统计机器翻译 (Statistical machine translation, SMT) 的原理, 将标准发音视为源语言, 学习者的发音视为目标语言, 分别对学习者发音串的语言模型和由学习者的发音到标准发音的转换模型进行训练[13−14]; 3) 基于字音转换 (Grapheme to phoneme, G2P) 的原理, 使用联合序列多阶模型 (Joint-sequence multi-gram, JSM) 将单词字母和相应的错误发音联合起来进行建模, 以覆盖到由单

收稿日期 2013-06-03 录用日期 2013-09-06
Manuscript received June 3, 2013; accepted September 6, 2013
国家自然科学基金 (61370034, 61005019, 61273268, 61105017) 资助
Supported by National Natural Science Foundation of China (61370034, 61005019, 61273268, 61105017)
本文责任编委 党建武
Recommended by Associate Editor DANG Jian-Wu
1. 清华大学电子工程系, 清华信息科学与技术国家实验室 北京 100084
2. 中国科学院电子学研究所, 传感技术国家重点实验室 北京 100190
3. 中国科学院大学 北京 100049
1. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084 2. State Key Laboratory of Transducer Technology,Institute of Electronics, Chinese Academy of Sciences, Beijing 100190 3. University of Chinese Academy of Sciences, Beijing 100049

词字母到发音的错误推断所带来的发音错误[15−16].

上面的方法都对标准发音和错误发音之间的对应关系, 以及错误发音对上下文的依赖进行了很好的建模. 但是发音错误还存在着一种规律, 就是在类似的上下文条件下会发生类似的发音错误. 例如在以发音/l/结尾时, 容易在/l/后插入发音/a/, 则在类似以/n/结尾时也容易插入/a/, 并且在以/l/结尾时也容易插入与/a/类似的发音/ə/. 文献 [8] 考虑到了该错误规律, 在从数据中自动提取出上下文相关的发音规则的基础上, 进一步利用专家知识对发音规则的上下文进行了扩展. 但是这种扩展方法不是数据驱动的原理, 不能在不同语言间进行灵活推广. 而这种相似性的学习是非常重要的. 因为一方面发音错误依赖于发音的上下文, 上下文信息越丰富, 对错误描述得越准确, 高阶的模型能够取得更好的效果; 而另一方面语言学习数据库的采集和标注是非常耗时耗力的, 一般的数据集很难覆盖到错误发生的所有情况. 在这种情况下, 如果能够对发音错误的相似性进行建模学习, 可以有效地弥补数据集不足的缺陷.

本文提出基于 JSM 和多层神经感知 (Multi-layer perception, MLP) 的发音建模方法, 在使用 JSM 模型将标准发音串和错误发音串表示为发音对序列的基础上, 首先使用 N 元文法来对发音错误的上下文建模, 再使用 MLP 对发音错误之间的相似性进行建模. 由于 MLP 模型将不同的词映射到一个连续的特征空间中, 能够学习到这种在相似上下文条件下出现的词的相似性, 可以很好的平滑高阶模型的概率空间.

1 自动发音错误检测和反馈系统

本文实现的自动发音错误检测和反馈系统如图 1 所示. 系统的流程为: 1) 系统自动生成学习文本, 学习者按照指定的文本发音; 2) 提取出学习者发音中的声学特征; 3) 根据发音文本和错误发音字典生成包含各种发音错误的检测网络; 4) 利用检测器, 结合声学模型和声学特征, 从检测网络中找出与学习者发音最接近的音素串; 5) 将检测出的音素串与标准音素串进行对比, 得到不匹配的部分, 也就是发音错误信息; 6) 将错误信息输入到错误反馈中, 以可视化语音的方式表现出正确发音与错误发音的差别, 反馈给学习者.

2 发音字典的生成

从上文描述的系统处理过程中可以看出, 在这种基于错误预测的检测系统中, 发音字典是非常关键的, 只有在发音字典中相应的包含了发音错误类型, 才可能会被检测出. 发音字典中描述的错误规律与实际越相符, 检测的结果就会越准确. 下文首先对发音错误的规律进行分析, 再介绍目前被广泛采用的发音规则的方法, 最后提出本文的发音字典生成方法.

2.1 发音错误的规律

基于错误预测的检测方法的出发点是, 学习者的发音错误存在一定的规律性, 这种规律在语言学的研究中被称作错误的语言迁移[17]. 对于中国人学习英语而言, 就是会将汉语的发音习惯引入到英语的学习当中, 但是却与英语的正确发音不一致. 对中国人的发音错误研究表明, 一般容易发生的错误主要有:

1) 替换错误, 在发音容易混淆时以一个发音替换另一个发音, 例如/l/与/n/, /v/与/w/, /ð/与/z/, /æ/与/a:/, 这类错误也是三类错误中发生最多的;

2) 插入错误, 在/ʌ/、/ɔ/等元音后插入/r/发音, 以及在以辅音结尾或者出现两个及两个以上的辅音时, 在辅音后插入元音, 例如以/t/结尾时插入/ə/;

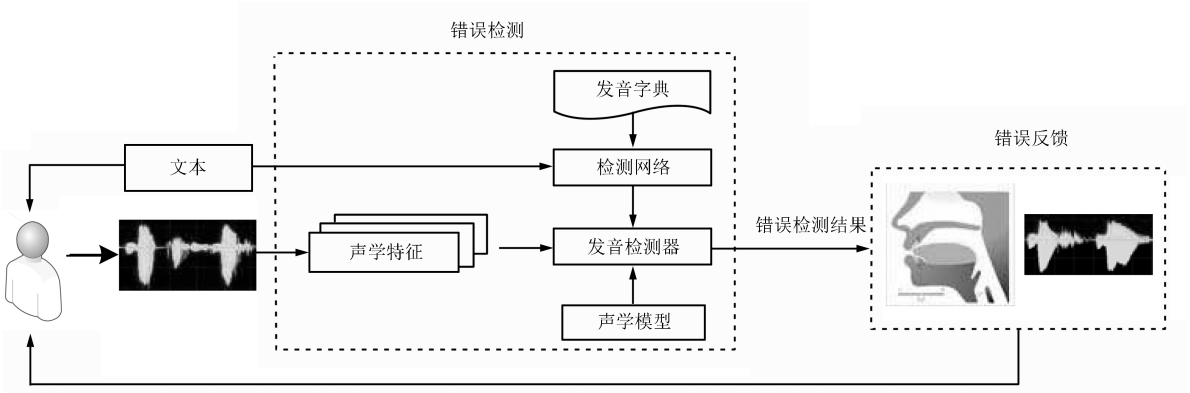


图 1 自动发音错误检测系统

Fig. 1 The system flow of automatic mispronunciation detection

3) 删除错误, 在多个辅音相连时, 其中的/r/、/t/发音就会比较容易被删除, 在以辅音结尾时也会将辅音删除.

总结这几类发音错误具有以下特点:

1) 对应性, 每种发音只会对应几种错误情况, 不会错成任意的发音 (例如/l/会被错读成/n/, 但是不会被错读成/z/);

2) 关联性, 发音错误的发生是与上下文相关的;

3) 相似性, 在类似的上下文条件下会发生类似的错误.

2.2　基于发音规则的发音字典生成方法

发音规则是目前被广泛采用的对发音错误建模的方法, 本文使用该方法生成的发音字典作为本文方法的一个参考. 发音规则一般采用上下文相关的表示形式

$$\alpha \rightarrow \beta \ / \ \lambda \_ \ \omega \tag{1}$$

表示在前一个发音为 λ 且后一个发音为 ω 时, 当前发音 α 会被错误发音为 β. 例如, 如果单词 invite (/ih n v ay t/) 被错发音为/ ih n w ay r t /, 则提取出的发音规则为 v → w / n _ ay 和 0 → r / ay _ t. 如果上下文为单词边界, 则以 # 表示. 当前规则若为替换错误, 则 α 与 β 互不相同, 若为插入错误, 则将 α 取值为 0, 若为删除错误, 则将 β 取值为 0.

发音规则的提取是使用动态规划算法 (Dynamic programming, DP) 将标准发音串和人工标注的发音串进行强制对准, 然后提取出不匹配的部分得到.

对发音规则计算相应的先验概率, 可以减小对检测网络引起的虚警, 同时利于对发音字典的优化. 先验概率的计算如下式所示:

$$p = \frac{N_{\text{occur}}}{N_{\text{pattern}}} \tag{2}$$

其中, N_{occur} 表示该发音规则所描述的错误实际出现的总数, N_{pattern} 表示匹配该规则模式的所有情况.

使用发音规则对标准发音字典进行扩展, 就能得到错误检测的发音字典. 以单词 live 为例, 通过查询标准发音字典, 得到两个标准发音串 /l ay v/ 和 /l ih v/, 构建音素级的标准网络, 在此基础上使用发音规则 (如 v → f / ay _ #) 对网络进行扩展, 则得到包含错误的网络 (如图 2 所示). 单词的网络构建中使用了 openFST 的开源工具包[18], 具体实现参照文献 [19]. 最后将错误网络中的每条路径展开, 就得到该单词所有可能的错误发音串以及相应的概率.

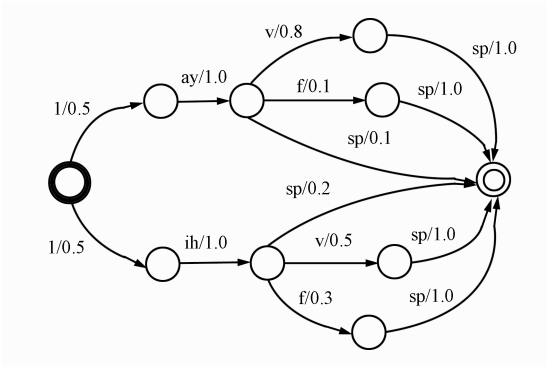


图 2　单词 live 的扩展网络示意图

Fig. 2　The diagram of extended network for word "live"

2.3　基于 JSM 和 MLP 的发音字典生成方法

从第 2.2 节对发音规则的描述中可以看到, 发音规则的建模方法简单直观, 可以对第 2.1 节中描述的发音错误的对应性和关联性进行建模, 但是不易学习到隐含在数据中的相似性. 本文针对第 2.1 节中描述的发音错误规律, 在文献 [15−16] 将 JSM 引入对发音错误进行建模的基础上, 提出基于 JSM 和 MLP 相结合的发音字典生成方法, 挖掘发音错误所隐含的相似性, 用于进一步平滑发音概率空间.

JSM 模型由 Bisani 和 Ney 首先提出, 用于语音合成和语音识别中对新单词生成对应的发音[20]. 它采用了基于统计计算的方法描述字母串和发音串之间的多对多的映射关系, 在建模过程中将单个字母和对应的发音首先组合成一个"字", 再使用 N 元文法来表示各字之间的相关性. 同时, 基于 MLP 的语言模型已成功的应用在大词汇量连续语言识别和统计语言翻译中[21−22]. 与 N 元文法语言模型相比, N 元文法语言模型中使用最大似然估计对词发生的概率进行统计, 词被表示到字典这么一个离散的空间中, 而 MLP 使用隐含的方式对模型进行平滑, 将整个字典投影到一个连续空间再进行概率估计, 使得未出现过的高阶模型可以得到更好的估计. 并且 MLP 将每个词映射到一个实值的特征空间, 能够捕捉到出现在相同上下文情况中的词的相似性. 所以将 JSM 和 MLP 模型结合起来用于对发音错误建模, 很好的契合了上述发音错误的三个规律.

本文首先利用 JSM 模型将标准发音和相应的错误发音分别组合起来, 以发音对的形式来描述 (例如 (/l/, /n/) 表示会将/l/错成/n/, (/l/, NULL) 表示会将/l/删除), 这是对发音错误的对应性建模; 然后用常见的 N 元文法模型来统计各个组合对的出现概率和它们相互关联的条件概率, 这是对发音错误的关联性建模; 最后用 MLP 对高阶模型的概率重估, 这是对发音错误的相似性进行建模. 具体的字

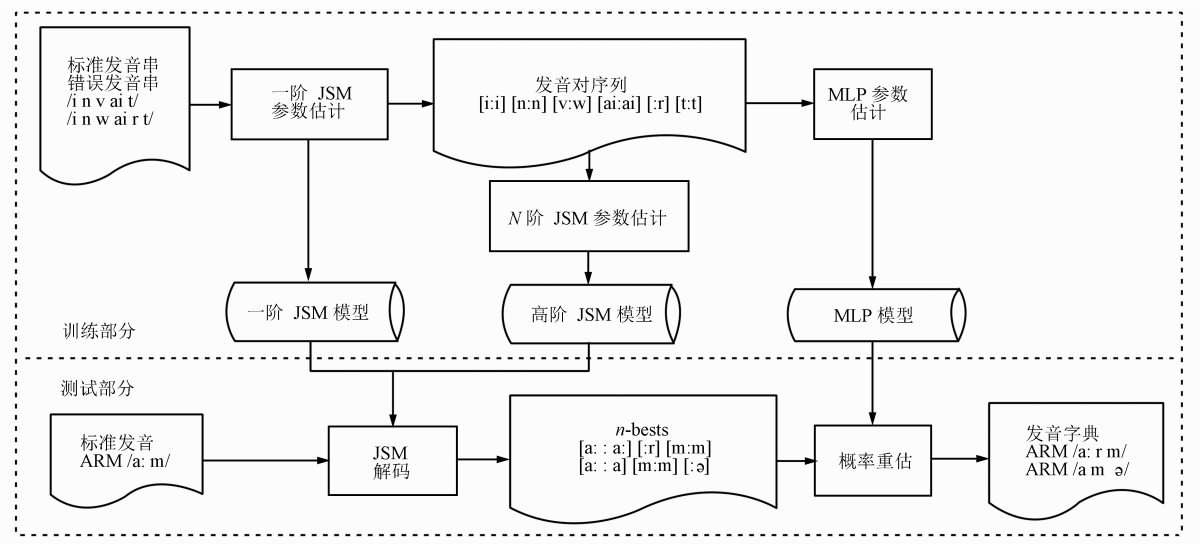


图 3 发音字典生成流程图

Fig. 3 The generation flow of pronunciation dictionary

典生成方法如图 3 所示, 分为训练和测试两部分. 在训练部分, 训练数据为标准发音串和错误发音串 (/i n v ai t/ → /i n w ai r t/). 标准发音串是通过查询标准发音字典得到的, 对于一个单词存在多个标准发音的情况, 则寻找出与当前错误发音匹配代价最小的标准发音. 准备好发音串后, 进行一阶 JSM 模型参数估计, 在得到 JSM 的模型的同时将发音串转换成发音对序列, 再分别进行高阶的 JSM 模型参数估计和 MLP 模型参数估计. 在测试部分, 首先查询标准发音字典得到单词的标准发音, 使用 JSM 模型求解得到 n 个概率最高的发音, 再使用 MLP 模型重新计算概率, 得到最终的发音字典.

JSM 模型的参数估计使用了期望最大化算法 (Expectation maximization, EM) 进行最大似然估计 (Maximum likelihood estimation, MLE). 首先考虑上下文无关的情况, 用 $\boldsymbol{g}$ 表示正确发音序列, $\boldsymbol{\varphi}$ 表示对应的错误发音序列, $\boldsymbol{q}$ 表示发音对序列, 定义 $S(\boldsymbol{g},\boldsymbol{\varphi})$ 表示 $\boldsymbol{g}$ 和 $\boldsymbol{\varphi}$ 所有可能的发音对集合为

$$S(\boldsymbol{g},\boldsymbol{\varphi}) := \left\{ \boldsymbol{q} \left| \begin{array}{c} g_{q1} \cup \cdots \cup g_{qk} = \boldsymbol{g} \\ \varphi_{q1} \cup \cdots \cup \varphi_{qk} = \boldsymbol{\varphi} \end{array} \right. \right\} \tag{3}$$

其中, $\cup$ 表示字符连接, $k = |\boldsymbol{q}|$ 表示 $\boldsymbol{q}$ 中发音对的总数. 则一阶 JSM 模型参数 $\boldsymbol{\vartheta}$ 的重估算法可以表示为:

$$p(\boldsymbol{q};\boldsymbol{\vartheta}) = \prod_{j=1}^{|\boldsymbol{q}|} p(q_j;\boldsymbol{\vartheta}) \tag{4}$$

$$e(q;\boldsymbol{\vartheta}) := \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}\,|\boldsymbol{g}_i,\boldsymbol{\varphi}_i;\boldsymbol{\vartheta}) n_q(\boldsymbol{q}) = \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} \frac{p(\boldsymbol{q};\boldsymbol{\vartheta})}{\sum_{\boldsymbol{q}'\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}';\boldsymbol{\vartheta})} n_q(\boldsymbol{q}) \tag{5}$$

$$p(q;\hat{\boldsymbol{\vartheta}}) = \frac{e(q;\boldsymbol{\vartheta})}{\sum_{q'} e(q';\boldsymbol{\vartheta})} \tag{6}$$

其中, N 为训练的发音序列个数, $n_q(\boldsymbol{q})$ 表示在序列 $\boldsymbol{q}$ 中某个字音组合 q 发生的概率, $e(q;\boldsymbol{\vartheta})$ 表示在当前参数 $\boldsymbol{\vartheta}$ 下在训练样本中 q 的发生次数.

对于高阶的模型 ($M > 1$), 就可以在低一阶模型的基础上进行重估. 先引入 h 来表示之前的 $M-1$ 个发音对的序列 $h_j = (q_{j-M+1},\cdots,q_{j-1})$, 用 $n_{q,h}(\boldsymbol{q})$ 表示在序列 $\boldsymbol{q}$ 中发音对组合 $q_{j-M+1},\cdots,q_{j-1}$ 的出现次数, 那么参数重估可以表示为

$$p(\boldsymbol{q};\boldsymbol{\vartheta}) = \prod_{j=1}^{|\boldsymbol{q}|} p(q_j\,|h_j\,;\boldsymbol{\vartheta}) \tag{7}$$

$$e(q,h;\boldsymbol{\vartheta}) := \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}\,|\boldsymbol{g}_i,\boldsymbol{\varphi}_i;\boldsymbol{\vartheta}) n_{q,h}(\boldsymbol{q}) = \sum_{i=1}^{N} \sum_{\boldsymbol{q}\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} \frac{p(\boldsymbol{q};\boldsymbol{\vartheta})}{\sum_{\boldsymbol{q}'\in S(\boldsymbol{g}_i,\boldsymbol{\varphi}_i)} p(\boldsymbol{q}';\boldsymbol{\vartheta})} n_{q,h}(\boldsymbol{q}) \tag{8}$$

$$p(q\,|h\,;\boldsymbol{\vartheta}') = \frac{e(q,h;\boldsymbol{\vartheta})}{\sum_{q'} e(q',h;\boldsymbol{\vartheta})} \tag{9}$$

由于 MLE 估计不能对训练数据中未出现的模型进行估计, 并且训练数据中可能存在噪声, 因此文献 [19] 中分别采用了证据折扣和证据修剪来应对这两个问题. 证据折扣是指将训练集中出现次数很多的 n-gram 的数量取出一部分来, 赋值给那些出现次数很少 (甚至是没有出现) 的 n-gram. 证据修剪是指将出现次数小于设定的门限的 n-gram 当做噪声丢弃.

将已经训练好的 JSM 模型用于对单词求解最可能的 n 个错误发音序列, 后验概率计算为:

$$p(\boldsymbol{\varphi}\,|\boldsymbol{g}) = \frac{\sum_{\boldsymbol{q}\in S(\boldsymbol{g},\boldsymbol{\varphi})} p(\boldsymbol{q})}{p(\boldsymbol{g})} \approx \frac{\max_{\boldsymbol{q}\in S(\boldsymbol{g},\boldsymbol{\varphi})} p(\boldsymbol{q})}{p(\boldsymbol{g})} \quad (10)$$

其中,

$$p(\boldsymbol{g}) = \sum_{\boldsymbol{\varphi}} p(\boldsymbol{g},\boldsymbol{\varphi}) = \sum_{\boldsymbol{\varphi}|g(\boldsymbol{q})=\boldsymbol{g}} p(\boldsymbol{q}) \quad (11)$$

MLP 模型的训练和测试则使用了由 Schwenk 实现的连续空间语言模型 (Continuous space language model, CSLM) 工具[23], 它的结构如图 4 所示, 为全连接的多层神经网络结构, P 表示一个投影的大小, H 和 D 分别是隐含层和输出层的大小. 神经网络的输入是前 $n-1$ 个字在字典 $h_j = w_{j-n+1}, w_{j-n+2}, \ldots, w_{j-1}$ 中的索引, 输出是字典中所有字的后验概率:

$$p(w_j = i|h_j), \qquad \forall i \in [1, D] \quad (12)$$

其中, D 表示字典的大小. 输入使用了 1 对多的编码, 也就是将字典中的第 i 个字编码为一个第 i 个元素为 1 而其他元素都为 0 的向量. $D*P$ 维投影矩阵的第 i 行就是第 i 个字的连续表示. 神经网络的训练使用了标准的反向传播算法.

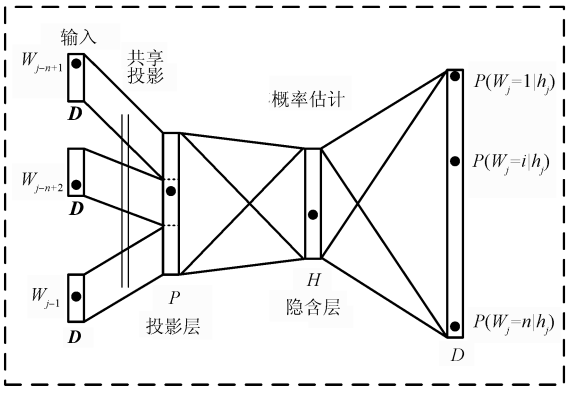


图 4　基于 MLP 的语言模型结构示意图

Fig. 4　The structure diagram of MLP based language model

3　实验

3.1　数据准备

实验是在 CU-CHLOE 数据库上进行的. 该数据库是为中国人学习英语而采集的, 一共包含了 111 人 (50 名女性和 61 名男性) 的语音. 每个录音者所使用的参考文本相同, 总共包括 86 句话. 在数据库的采集过程中, 我们关注在由学习者本身的发音缺陷造成的错误上, 所以在录制过程中, 严格控制了由心理精神状态所造成的随机错误, 也就是偶然的 "失误". 当学习者发音存在明显的 "失误" 时, 会被要求重新录制该语音. 本文随机选择其中 55 人 (25 名女性和 30 名男性) 的语音作为训练集, 共 4 727 句话, 约 9.6 h; 其余的作为测试集, 共 4 815 句话, 约 10 h. 数据库由经过训练的语言学专家进行了音素级的标注.

本文使用包含 45 个音素的音素集. 在 CU-CHLOE 训练集上进行声学模型训练, 使用 39 维的 PLP 特征 (13 维的基本特征和一阶差分以及二阶差分), 得到的模型包含了 1 074 个状态, 每个状态有 8 个高斯混元分量.

而在对发音错误进行建模时, 由于训练集和测试集的参考文本相同, 所包含的上下文情况完全一致, 本文采取了 10 倍交叉验证的方法. 将单词集随机分为 10 组, 每次使用其中一组作为测试词, 其余作为训练词, 使用训练集中对应训练词的数据用于发音规则的提取以及 JSM 模型和 MLP 模型的训练. 数据库中正确发音和错误发音的统计情况如表 1 所示.

进行错误检测时, 直接将使用发音规则和 JSM 方法生成的扩展发音字典用于识别, 识别器使用的是 HTK 工具包中的 HVite.

3.2　发音字典性能评价

对字典性能的评价指标为虚警率 (False alarm rate, FAR) 和漏报率 (Miss rate, MR).

首先定义一个单词 w 的发音集合 D_w 在测试集 T 上的引起的漏报率 $MR(D_w)$ 和虚警率 $FAR(D_w)$ 分别为:

$$\begin{aligned} MR(D_w) &= \textstyle\sum_{m\in T, m\notin D_w} \varphi(m) \\ FAR(D_w) &= \textstyle\sum_{m\in D_w, m\notin T} p(m) \end{aligned} \quad (13)$$

其中, $\varphi(m)$ 表示单词发音 m 在 T 中的归一化概率, $p(m)$ 表示单词发音 m 在 D_w 中的归一化概率. 计算整个字典的性能则是按照每个单词的概率对单词错误率进行加权平均:

表 1 训练集和测试集的数据分布

Table 1 The data distribution of training and testing set

	单词正确发音个数	单词错误发音个数	单词错误发音类别	正确发音个数	错误发音个数
训练集	20 202	14 468	4 523	95 153	22 652
测试集	20 115	15 181	4 672	96 246	23 837

$$\begin{aligned} MR &= \textstyle\sum_w p(w)MR(D_w) \\ FAR &= \textstyle\sum_w p(w)FAR(D_w) \end{aligned} \tag{14}$$

其中, $p(w)$ 表示单词 w 在测试集 T 上的发生概率.

3.2.1 发音字典在测试集上的性能

本文首先对每个单词生成 1 000 个发音 (音素较少的单词会相应生成得少一些), 以尽可能将学习到的错误规律都展现出来. 在没有使用 MLP 模型进行概率重估的情况下, 表 2 给出了不同模型阶数时 JSM 发音字典在在测试集上的性能. 从表中可以看到, 2 阶字典的性能是最好的, 随着模型阶数的升高或者降低, 性能都会变差. 这与发音错误分析的结果是比较一致的, 发音错误的所关联的上下文情况大多数为上下文无关或者与前一个发音相关, 还有少量的与前两个或者多个发音相关. 但是高阶的模型由于训练数据不足而使得概率空间不够平滑, 对未知的 N 元文法的概率估计不准确. 下文的实验将在模型阶数为 2 的基础上进行.

表 2 JSM 发音字典在测试集上的性能

Table 2 The performance of JSM pronunciation dictionary on testing set

模型阶数	FAR (%)	MR (%)
1	64.86	12.09
2	59.18	9.71
3	68.67	12.39
4	71.09	10.07
5	71.43	9.35
6	71.41	9.35

3.2.2 选择合适的 Top-N

对于发音错误检测而言, 字典中发音越多, 必然会对检测器造成越多的混淆, 所以需要从 1 000 个发音中选取出合适的 Top-N 个发音. 我们将 N 从 1 增大到 1 000, 统计出上面得到的 2 阶字典在测试集上的性能, 图 5 给出了 FAR 和 MR 随 N 的变化曲线. 可以看到, 随着 N 增大, FAR 增大, MR 减小, 在 N 为 200 时, FAR 和 MR 都趋于平稳. 则本文将 200 作为一个合适的 Top-N 值.

3.2.3 平滑发音概率空间

在使用 JSM 模型生成 1 000 个发音的基础上, 再用 MLP 模型对 1 000 发音进行概率重估, 分别对每个单词取 Top 200 个概率最大的发音, 得到的性能如表 3 所示. 同时也给出了分别使用发音规则和 JSM 模型生成的发音字典性能. 表 3 中第 1 行表示的是直接使用从数据中提取出的上下文相关 (Context-dependent, CD) 的规则生成的发音字典的性能, 可以看到, FAR 和 MR 都非常高. 这一方面是 CD 发音规则对上下文条件的限制太多, 而本文训练集对三元文法的覆盖不够全面, 错误规则不能够有效的在测试集单词上应用, 导致高 MR; 另一方面则是没有对发音规则进行概率平滑和去噪, 导致高 FAR.

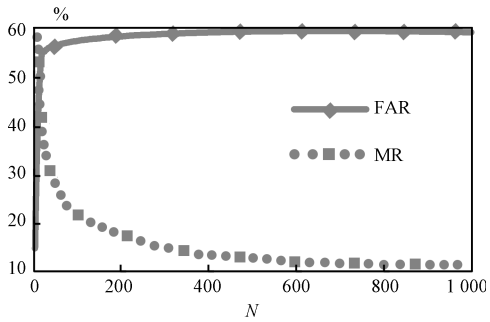


图 5 FAR 和 MR 随 N 的变化曲线

Fig. 5 The FAR and MR versus N curves

表 3 Top 200 字典的性能对比

Table 3 Comparison of performance between different Top 200 dictionaries

发音字典	FAR (%)	MR (%)
CD 发音规则 (未处理)	92.35	68.22
CD 发音规则	66.3	67.3
LCRC 发音规则	74.29	35.51
JSM	58.44	17.68
JSM+MLP	52.78	15.31
JSM+MLP(TIMIT)	49.24	13.46

本文首先对 CD 发音规则进行概率平滑和去噪, 除去发生次数非常少的规则, 并对概率特别高的规则适当降低起其概率, 使得 FAR 得到了有效的下降 (表 3 中第 2 行). 为了能够与二阶的 JSM 模型平等的进行比较, 本文在概率平滑的基础上, 进一步的提取出只与上文或者下文相关 (Left-context or right-context, LCRC) 的发音规则, 使得规则的应用条件更加宽泛, MR 得到了很大的下降 (表 3 中第

3 行).

与发音规则方法相比, JSM 方法则大大降低了 FAR 和 MR. JSM 的优势在于采用了统计估计的方法, 能够更准确的描述错误的对应关系和相应的概率. 同时 JSM 方法在模型训练中采用的证据折扣和证据修剪也是非常有用的, 能够解决部分由于数据分布不均所造成的概率不平衡, 并去除部分数据中包含的噪声 (数据库中仍然包含了少量的随机错误). 在使用本文提出的 MLP 进行概率重估后, 发音概率空间被平滑, FAR 和 MR 则得到了进一步的下降. MLP 的概率重估是在 JSM 框架中 n-gram 模型的基础之上进行的, 它对性能的提高也印证了[24] 中 MLP 与 n-gram 模型的互补性. 在 MLP 的训练中, MLP 模型的单个投影大小设置为 200, 隐含层大小也设置为 200, 输出层根据字典大小设置为 602. 也就是说从 602 个 "字" 中除去 45 个正确音素与正确音素组成的 "字", 数据集中一共存在 557 种单个发音的错误 (包括单个发音的插入、删除和替换错误).

考虑到 TIMIT 标准发音字典中包含了丰富的 n-gram 种类, 我们将 (标准发音序列, 标准发音序列) 也加入到训练集中, 用于训练 JSM 模型和 MLP 模型, 使得性能得到了进一步的提高, 与只使用了 JSM 模型的基线字典相比, FAR 下降了 9.20 %, MR 下降了 4.22 %. TIMIT 字典和 MLP 都更好地平滑了发音错误的概率空间, 有效地对发音错误进行了建模.

发音字典的性能与发音字典的大小是密切相关的. 在表 3 中 Top 200 字典的基础之上, 本文改变字典的大小, 取 Top 2 到 Top 200 的字典进行测试, 性能如图 6 所示. 在发音字典较小时, 包含的错误发音种类有限, 同时虚警也会比较低. 字典变大时, MR 会随之降低, 同时 FAR 随之升高. 几种不同的字典的性能差异也非常明显, 使用了 MLP 模型和 TIMIT 字典的方法是很有效的.

3.3　发音错误检测

本文采用的错误检测结果的评价指标为: 1) 音素正确率 (Correctness, CORR), 表示识别结果中与人工标注一致的音素所占的比率; 2) 错误类型的正确率 (Diagnostic accuracy, DA), 表示诊断为发音错误的音素中与人工标注的错误类型完全一致的音素所占的比率; 3) 错误接受率 (False acceptance rate, FACR), 表示人工标注为错误发音的音素中被诊断为正确发音的比率; 4) 错误拒绝率 (False rejection rate, FRR), 表示人工标注为正确发音的音素中被诊断为错误发音的比率. CORR 和 DA 越高, FACR 和 FRR 越低, 说明系统性能越好. 对于发音错误检测而言, 错误拒绝比错误接受所带给学习者的负面影响更大, 因此 FRR 的重要性高于 FACR.

我们将图 6 中不同大小的字典用于发音错误检测, 得到测试集的测试词上的性能如图 7～9 所示. 可以看到, 基于 JSM 模型的方法远远优于发音规则的方法, 使用 TIMIT 标准发音字典辅助训练和使用 MLP 模型进行概率重估对发音字典的生成都是有利的, 使错误检测的性能得到了明显的提高. 对于图 7～9 的 4 种方法而言, 随着字典中发音数量的增多, FRR 缓慢上升, FACR 快速下降, 这是因为数据集中错误发音总数只有正确发音总数的五分之一左右, FACR 的变化比 FRR 的变化更明显. 而由于字典变大, 对识别网络造成的混淆也越大, DA 和 CORR 都会随之下降. 图 5 中 MLP 模型和 TIMIT 字典对发音字典的改进是不随字典大小变化的, 那是理想状态下的性能, 但是在实际结合了声学模型的错误检测中, 受到声学模型的性能的限制, 随着字典中发音数量的增多, 解码网络的混淆越来越大, 概率平滑的作用也就越来越不明显.

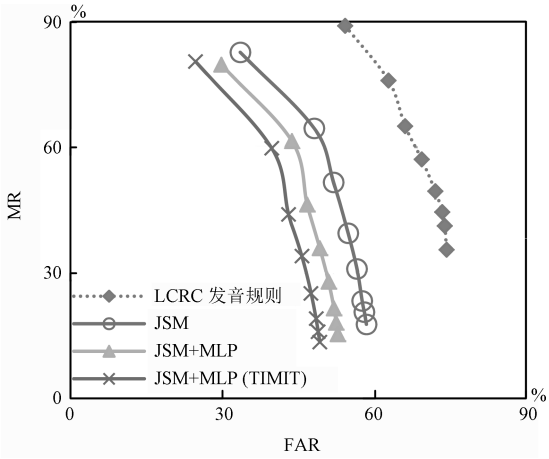


图 6　不同大小的发音字典在测试集上的 FAR 和 MR 性能曲线

Fig. 6　The FAR and MR curves on testing set for pronunciation dictionary of different sizes

4　总结

本文针对发音错误具有的对应性、关联性和相似性的特点, 使用 JSM 和 MLP 相结合的方法来改进发音错误检测的发音字典生成. JSM 模型能够有效地描述发音错误的对应性和关联性, MLP 模型则能够对发音错误的相似性建模. 由于 MLP 将不同的词映射到连续的空间中, 可以捕捉到相似上下文条件下发生的相似的错误发音, 对有限数据集上统计出的高阶模型的概率空间进行平滑. 实验证明, 本文提出的方法有效地提高了发音字典对错误的预测性能. 而系统性能的进一步提高, 则需要声学模型对各发音之间有更好的区分性.

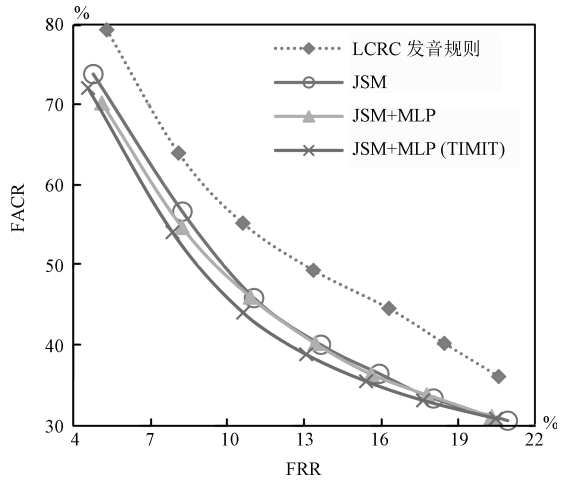


图 7 错误检测的错误权衡 (Detection error trade-off, DET) 曲线

Fig. 7 The DET curves of mispronunciation detection

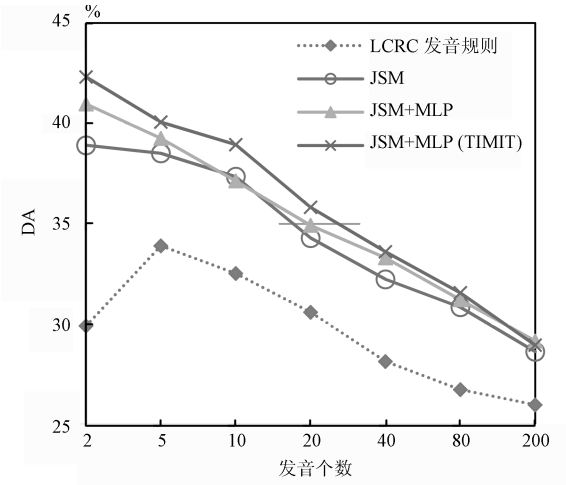


图 8 错误检测的 DA 性能曲线

Fig. 8 The DA curves of mispronunciation detection

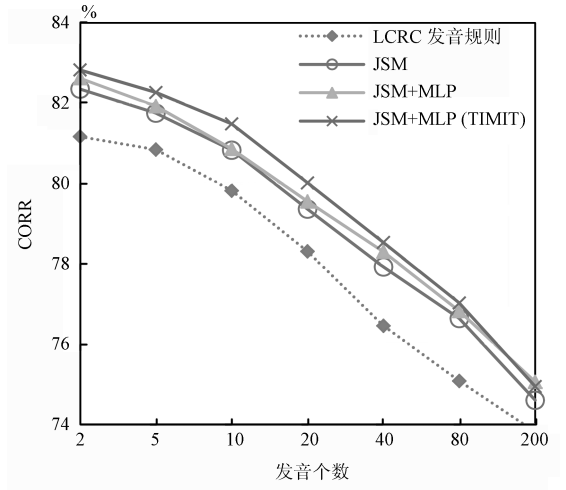


图 9 错误检测的 CORR 性能曲线

Fig. 9 The CORR curves of mispronunciation detection

References

1 Eskenazi M. An overview of spoken language technology for education. *Speech Communication*, 2009, **51**(10): 823−844

2 Ito A, Lim Y L, Suzuki M. Pronunciation error detection method based on error rule clustering using a decision tree. In: Proceeding of the 6th Annual Conference of the International Speech Communication Association. Tohoku University, Japan: ISCA, 2005. 173−176

3 Yoon S Y, Hasegawa-Johnson M, Sproat R. Landmark-based automated pronunciation error detection. In: Proceeding of the 11th Annual Conference of the International Speech Communication Association. Tokyo: ISCA, 2010. 614−617

4 Strika H, Truongb K, Wet F D, Cucchiarini C. Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 2009, **51**(10): 845−852

5 Zhang F, Huang C, Soong F K, Chu M, Wang R H. Automatic mispronunciation detection for Mandarin. In: Proceeding of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, Nevada, USA: IEEE, 2008. 5077−5080

6 Wei S, Hu G P, Hu Y, Wang R H. A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 2009, **51**(10): 896−905

7 Wang H C, Waple C J, Kawahara T. Computer Assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition. *Speech Communication*, 2009, **51**(10): 995−1005

8 Luo D, Yang X S, Wang L. Improvement of segmental mispronunciation detection with prior knowledge extracted from large *L*2 speech corpus. In: Proceeding of the 12th Annual Conference of the International Speech Communication Association. Florence, Italy: ISCA, 2011. 1593−1596

9 Yuan H, Zhao J H, Liu J. A two-stage mispronunciation detection approach for computer-assisted pronunciation training. In: Proceeding of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference 2011. Xi′an, China: Asia-Pacific Signal and Information Processing Association, 2011. 972−976

10 Meng H, Lo Y Y, Wang L, Lau W Y. Deriving salient learners′ mispronunciations from cross-language phonological comparisons. In: Proceeding of the 2007 Automatic Speech Recognition and Understanding Workshop. Kyoto, Japan: IEEE, 2007. 437−442

11 Lo W K, Zhang S, Meng H. Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. In: Proceeding of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Chiba, Japan: ISCA, 2010. 765−768

12 Harrison A M, Lau W Y, Meng H, Wang L. Improving mispronunciation detection and diagnosis of learners′ speech with context-sensitive phonological rules based on language transfer. In: Proceeding of the 9th Annual Conference of the International Speech Communication Association. Brisbane: ISCA, 2008. 2787−2790

13 Stanley T, Hacioglu K, Pellom B. Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system. In: The 2011 Speech and Language Technology in Education Workshop. Venice, Italy: ISCA, 2011. 125−128

14 Stanley T, Hacioglu K. Improving *L*1-specific phonological error diagnosis in computer assisted pronunciation training. In: Proceeding of the 13th Annual Conference of the International Speech Communication Association. Portland, Oregon: ISCA, 2012. 826−829

15 Qian X J, Meng H, Soong F F. On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training. In: Proceeding of the 12th Annual Conference of the International Speech Communication Association. Italy, Florence: ISCA, 2011. 865−868

16 Qian X J, Meng H, Soong F. Capturing *L*2 segmental mispronunciations with ioint-sequence models in computer-aided pronunciation training (CAPT). In: Proceeding of the 7th International Symposium on Chinese Spoken Language Processing. Taiwan, China: IEEE Computer Society, 2010. 84−88

17 Gass S M, Selinker L. *Language Transfer in Language Learning*. Philadelphia, USA: John Benjamins Publishing Company, 1993. 87−101

18 Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 2002, **16**(1): 69−88

19 Harrison A M, Lo W K, Qian X J, Meng H. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In: The 2009 Speech and Language Technology in Education Workshop. Warwickshire, England: ISCA, 2009. 45−48

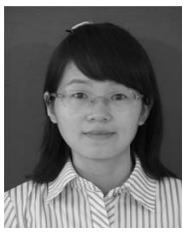
20 Bisani M, Ney H. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 2008, **50**(5): 434−451

21 Schwenk H. Continuous space language models. *Computer Speech and Language*, 2007, **21**(3): 492−518

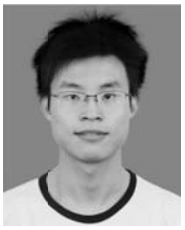
22 David T, Miles O. Randomised language modelling for statistical machine translation. In: Proceedings of the 45th Prague, Czech Republic Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic: ACL, 2007. 512−519

23 Schwenk H. Continuous-space language models for statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 2010, **93**(1): 137−146

24 Oparin I, Sundermeyer M, Ney H, Gauvain J. Performance analysis of neural networks in combination with *n*-gram language models. In: Proceeding of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012. 5005−5008



袁 桦 清华大学电子工程系博士研究生. 主要研究方向为发音错误检测. 本文通信作者. E-mail: yuanh08@mails.tsinghua.edu.cn
(**YUAN Hua** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. Her research interest covers mispronunciation detection. Corresponding author of this paper.)



史永哲 清华大学电子工程系博士研究生. 主要研究方向为语音识别, 语言模型和音频检索.
E-mail: shiyz09@gmail.com
(**SHI Yong-Zhe** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition, language modelling, and audio indexing and searching.)

赵军红 中国科学院电子学研究所博士研究生. 主要研究方向为韵律检测和表现力音视频语音合成.
E-mail: junhong.iecas@gmail.com
(**ZHAO Jun-Hong** Ph. D. candidate in the State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences. Her research interest covers prsodic event detection and expressive visual-speech synthesis.)

刘 加 清华大学电子工程系教授. 主要研究方向有信号处理, 语音识别, 语音合成, 语音编码和多媒体通信.
E-mail: liuj@tsinghua.edu.cn
(**LIU Jia** Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers signal processing, speech recognition, speech synthesis, speech coding and multimedia communication.)