汉语系统中可以检验组成命令词的单字, 此时系统对这些单元的分辨能力则至关重要, 这点在 2.4 节中已经予以验证. 然而, 此种实现依赖于上游的分割结果, 为系统带来了新的困难. 除此之外, 另一种思路则是尝试增强系统本身的时序鉴别能力.

例如, 在 i-vector 框架下, 一般通过隐马尔科夫模型 (Hidden Markov Model, HMM)、长短期记忆网络 (Long Short-term Memory, LSTM) 等时序相关的模型建模时序特征, 产生新的 i-vector 或作为已有 i-vector 的补充. 文献 [10] 对比了 i-vector、d-vector、s-vector 三种特征对不同语音特性 (如说话人身份、说话速度等) 的刻画能力. 其中对于词序特性, 该文通过在两段拼接顺序不同的语音上的分类任务予以验证, 在此实验中 i-vector 的鉴别效果较差, 接近随机猜测, 说明其几乎没有时序鉴别能力, 而基于 LSTM 的 s-vector 效果突出, 因此该文通过拼接二者得到所谓 i-s-vector, 在包括词序区分的大部分任务上均取得了最优结果. Hossein 等[17] 则提出使用 HMM 代替 GMM 作为 UBM 模型的基础, 通过对每个音素训练 HMM 并拼接, 得到特定语句的 HMM模型, 由此模型产生的 i-vector 与语句的相关性更强.

上述方法通过引入其它时序相关的模型增强 i-vector 的时序鉴别性能, 其共同局限性在于需要与语句相关的信息, 如每段语句的音素标签, 用于训练对应的 HMM 或神经网络模型, 而实际应用中我们希望在仅具备录入语音, 没有关于语音内容知识的情况下, 完成系统的训练. 动态时间规整 (Dynamic Time Warping, DTW) 算法[18] 是语音领域的经典方法之一, 其通过对语音序列进行非线性扭曲实现序列间对齐, 从而求取相似度, 算法直观且易于实现, 其约束条件决定其适于衡量时序差异, 且不依赖语音以外的信息. 因此, 本文提出将 DTW 与原有 i-vector+PLDA 系统融合, 期望二者融合而成的系统可以兼顾 i-vector+PLDA 的低错误率和 DTW 的时序鉴别能力.

3.2 得分计算、似然比校准与系统融合

DTW 算法产生两段序列之间的相似度得分, 而在很多命令词系统中, 单个词语对应存在多个模板 (训练语音片段). 本文中将目标语音在某词语下所有模板上的 DTW 得分的平均值作为该语音与此词语的相似度.

尽管上述得分与对数似然比同为相似度的体现, 但由于计算方式、统计特性上的差异, 数学上二者并不相容. 本文采用文献 [19] 中的逻辑回归校准方法, 通过在同源、不同源得分上训练二元逻辑回归模型得到模型系数, 并校准原始得分 s, 使其等价于对数似然比:

$$\log(LR) = \alpha + \beta s \tag{19}$$

系统融合采用两系统似然比的连乘, 即对数似然比的简单相加:

$$\log(LR_m) = \log(LR_{\mathrm{DTW}}) + \log(LR_{\mathrm{PLDA}}) \tag{20}$$

3.3 逆序短语拒识实验

第 2.5 节实验中使用的 SbPhrase 数据集不含有实验所需的音素相近但字序不同的短语对, 因此为 SbPhrase 中前 50 条短语重新采集语音, 构建小型子数据集 SbPhrase-T. 对于每条短语, 除其正序 (如“曼彻斯特”) 外, 另行采集部分逆序 (如“斯特曼彻”) 和完全逆序 (如“特斯彻曼”) 两份语音. 将 SbPhrase 中前 50 条短语作为集内词训练 i-vector+PLDA 系统, 将两种逆序语音作为集外词进行拒识实验.

图 5 为短语的 3 种字序在原系统上对数似然比得分的混淆矩阵 (confusion matrix), 展示了所有语音在所有正序短语的 PLDA 模型上的相似度情况. 其中, 为方便横向比较, 横轴每 3 列对应一条短语, 其下三列依次对应正序、部分逆序、完全逆序语音的得分. 观察对角线可以发现, 两种逆序语音在其对应序号正序模型上的得分总体较高, 说明系统不能将其有效拒识, 再次确认了前述 i-vector 在时序鉴别能力方面的弱点.

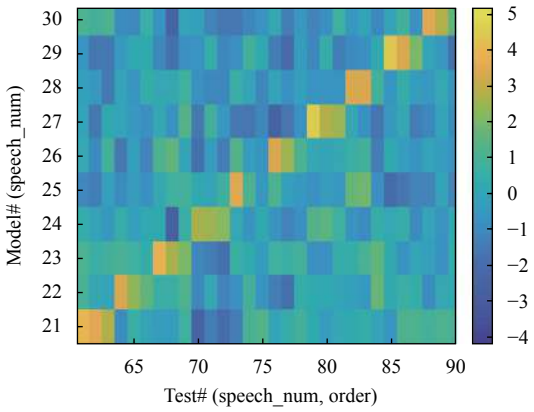

图 5 原系统的混淆矩阵 (部分)

图 6 为 DTW 与 i-vector+PLDA 系统融合后, 新系统上得分的混淆矩阵, 经 DTW 修正后, 混淆矩阵的对角线更加清晰, 两种逆序语音的得分明显降低, 接近背景 (短语不匹配情况) 水平.

表 6 为两种系统对逆序语音拒识的量化实验结果. 数据表明, 相比单 i-vector+PLDA 系统, 融合系统有效

降低了系统在逆序语音上的虚警, 说明 DTW 得分的引入提高了系统的时序鉴别能力.

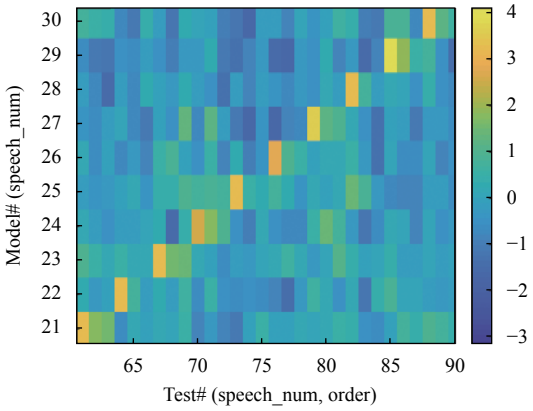

图 6 新系统的混淆矩阵 (部分)

表 6 不同系统在 SbPhrase-T 数据集上的拒识性能

系统	虚警率(%)
i-vector+PLDA	1.98
融合	0.72

4 i-vector+PLDA 置信度应用意义分析

相比传统的置信度估计方法, 上文提出的基于 i-vector 和 PLDA 以及融合 DTW 的方法具有两点优势:

其一, 无需训练声学模型及语言模型. 传统方法, 特别是基于后验概率的置信度判决方法, 依赖基本语音识别单元 (如音素或音节) 声学模型的似然值得分和相应的声学模型. 这些信息常常与特定系统及其使用的声学模型、语言模型相关, 迁移至传统语音识别系统的诸多变种以及未来更新颖的语音识别框架中存在困难. 本文方法训练过程则仅需语音及对应的类别标签, 外部系统不额外提供其他先验的声学和语言模型信息, 一方面使得系统结构直观、易于实现, 另一方面因为无需考虑前端系统的实现细节, 可以独立测试与部署, 达成一定程度的模块化, 使用更加灵活广泛.

其二, 无需提供语句内容相关信息. 实际应用中, 很多命令词系统通过非确定性的命令词加强安全性或保证用户体验. 例如, 用户可以根据个人喜好为智能音箱、手环等智能设备录入自选的唤醒词, 后续通过该词唤醒设备进入工作状态. 此类场景中, 设备在录入阶段无法获知命令词的内容, 因此文献 [10,17] 中的方法缺乏训练所需的标签. 本文方法通过 DTW 完成时序信息的补充, 避免了对此类“标签”的依赖, 可以应对较为复杂多变的命令词. 在电话银行、智能家居等应用中, 通过本文方法对语音识别系统的结果进行验证, 既有助于降低错误, 提升用户体验, 同时仍不失原系统交互过程中的灵活性, 对命令词系统的改进具有实际价值.

此外, 第 2 节的置信度检验实验结果中, 本文方法辅助语音识别系统对连接词识别率的提升相比孤立字更为显著. 越长的语音片段, 其中包含的语音内容信息越丰富, 通过相应增加 UBM 混合数和 i-vector 维度, 得到的 i-vector 能够充分包含此信息, 而特征信息量的增加也有益于 PLDA 对有用信息的分离与鉴别. 因此, 相比孤立字, 本文方法更适合用于词语、短句等较长的语音.

5 结束语

本文提出将 i-vector 以及 PLDA 模型用于置信度判决. i-vector 语音特征包含了包括说话内容在内的各种差异信息, 利用 PLDA 可以中和其他信息的影响, 有效鉴别说话内容, 且其形式上符合基于似然比的置信度分析, 在孤立字、连接词实验中体现出了良好潜力. 通过与 DTW 融合, 补充缺失的时序信息, 得到不依赖声学、语言模型以及语句标签的置信度分析方法, 在应用中较传统的置信度分析方法有其独特优势.

参考文献

1 Jiang H. Confidence measures for speech recognition: A survey. Speech Communication, 2005, 45(4): 455–470. [doi: 10.1016/j.specom.2004.12.004]

2 Wessel F, Schluter R, Macherey K, *et al*. Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing, 2001, 9(3): 288–298. [doi: 10.1109/89.906002]

3 Rahim MG, Lee CH, Juang BH. Discriminative utterance verification for connected digits recognition. IEEE Transactions on Speech and Audio Processing, 1997, 5(3): 266–277. [doi: 10.1109/89.568733]

4 Dehak N, Kenny PJ, Dehak R, *et al*. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788–798. [doi: 10.1109/TASL.2010.2064307]

5 Prince SJD, Elder JH. Probabilistic linear discriminant analysis for inferences about identity. 2007 IEEE 11th International Conference on Computer Vision. Rio de

Janeiro, Brazil. 2007. 1–8.
6 Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10(1–3): 19–41.
7 Campbell WM, Sturim DE, Reynolds DA. Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 2006, 13(5): 308–311. [doi: 10.1109/LSP.2006.870086]
8 Kenny P, Boulianne G, Ouellet P, *et al*. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1435–1447. [doi: 10.1109/TASL.2006.881693]
9 Dehak N. Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification[Ph.D. thesis]. Montreal, QC: École de Technologie Supérieure, 2009.
10 Wang S, Qian YM, Yu K. What does the speaker embedding encode? Interspeech 2017. Stockholm, Sweden. 2017. 1497–1501.
11 Dehak N, Shum S. Low-dimensional speech representation based on factor analysis and its applications. Interspeech 2011. Florence, Italy. 2011.
12 Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7(2): 179–188. [doi: 10.1111/j.1469-1809.1936.tb02137.x]
13 Garcia-Romero D, Espy-Wilson CY. Analysis of i-vector length normalization in speaker recognition systems. 12th Annual Conference of the International Speech Communication Association. Florence, Italy. 2011. 249–252.
14 Matějka P, Glembek O, Castaldo F, *et al*. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic. 2011. 4828–4831.
15 Sadjadi SO, Slaney M, Heck AL. MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research. Speech and Language Processing Technical Committee Newsletter, 2013, 1(4): 1–32.
16 Varga A, Steeneken HJM. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication, 1993, 12(3): 247–251. [doi: 10.1016/0167-6393(93)90095-3]
17 Zeinali H, Sameti H, Burget L. HMM-based phrase-independent i-vector extractor for text-dependent speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(7): 1421–1435. [doi: 10.1109/TASLP.2017.2694708]
18 Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 43–49. [doi: 10.1109/TASSP.1978.1163055]
19 Morrison GS. Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. Australian Journal of Forensic Sciences, 2013, 45(2): 173–197. [doi: 10.1080/00450618.2012.733025]