计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(1):54−62 [doi: 10.15888/j.cnki.csa.007732]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

概率线性判别分析在语音命令词置信度判决中的应用①

闫宏宸, 肖　熙

(清华大学 电子工程系, 北京 100084)
通讯作者: 闫宏宸, E-mail: yhc17@mails.tsinghua.edu.cn

摘　要: 置信度判决用于确定语音数据与模型之间的匹配程度, 可以发现语音命令系统中的识别错误, 提高其可靠性. 近年来, 基于身份矢量 (identity vector, i-vector) 以及概率线性判别分析 (Probabilistic Linear Discriminant Analysis, PLDA) 的方法在说话人识别任务中取得了显著效果. 本文尝试将 i-vector 以及 PLDA 模型作为一种命令词识别结果置信度分析方法, 其无需声学模型、语言模型支撑, 且实验表明性能良好. 在此基础上, 针对 i-vector 在刻画时序信息方面的不足, 尝试将该系统与 DTW 融合, 有效提升了系统对音频时序的鉴别能力.

关键词: 置信度; 身份矢量; 概率线性判别分析; 时序信息; 动态时间规整

引用格式: 闫宏宸,肖熙.概率线性判别分析在语音命令词置信度判决中的应用.计算机系统应用,2021,30(1):54–62. http://www.c-s-a.org.cn/1003-3254/7732.html

Application of Probabilistic Linear Discriminant Analysis in Voice Command Confidence Measures

YAN Hong-Chen, XIAO Xi

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Confidence measures represent the degree of match between speech data and models, and thus can be utilized to spot errors in voice command systems, improving their reliability. In recent years, systems based on identity vector (i-vector) and Probabilistic Linear Discriminant Analysis (PLDA) have been proven effective in the task of Speaker Verification (SV). This study proposes i-vector and PLDA as a confidence measure for voice command system without the need for acoustic or language models and demonstrates fair performance. Furthermore, in consideration of the deficiency of such i-vectors in modeling temporal information, this study proposes a fusion approach of such system with DTW, enhancing its time sequence discrimination ability.

Key words: confidence measure; identity vector (i-vector); Probabilistic Linear Discrimination Analysis (PLDA); temporal information; dynamic time warping

计算机技术的发展为人类生活带来了极大便利, 基于语音的人机交互已经以命令词识别系统的形式在智能家居、可穿戴设备等平台得到了应用. 命令词识别系统是一种“*N* 选 1”的识别系统, 将输入语音识别为预先设定的命令词之一, 系统的错误主要来自对集内命令词的错识和对集外语音或噪声的误识. 有鉴于此类系统使用环境的多样性, 通过某种手段拒绝错误的识别结果, 特别是拒绝环境噪声和集外语音引发的错误识别结果, 对提高命令词系统的可靠性极为重要.

对语音识别结果的置信程度加以检验和判决是一种比较理想的做法. 在数理统计中, 置信度分析是分析一个随机变量落在某个区间的概率, 而在语音识别中,

① 收稿时间: 2020-05-21; 修改时间: 2020-06-16; 采用时间: 2020-06-28; csa 在线出版时间: 2020-12-31

置信度分析通常用于衡量模型与数据之间匹配的可信程度. 置信度分析方法大致可以分为基于预测特征的组合、基于后验概率和基于似然值比值 (似然比) 等 3 大类方法[1]. 其中基于似然比的置信度分析方法将置信度问题转化为统计假设检验问题, 设定数据由某一模型产生 (零假设) 和数据非由该模型产生 (对立假设) 两种假设, 通过两种假设上的似然比检验以及阈值判断是否接受零假设. 已有的置信度分析方法包括基于词网格生成后验概率的置信度[2]、基于逆模型建模对立假设计算似然比的置信度[3] 等.

本文提出了一种无需声学模型、语言模型支撑的命令词置信度分析方法. 调研发现, 身份矢量 (identity vector, i-vector) 特征[4] 和概率线性判别分析 (Probabilistic Linear Discriminant Analysis, PLDA) 方法[5] 已经在说话人识别中得到了广泛应用, 但是将 i-vector 特征与 PLDA 应用于命令词语音识别的置信度分析中尚未有文献报导. i-vector 的原理是通过对所有语音数据训练, 建立通用背景模型 (Universal Background Model, UBM)[6], 将语音表示为高维的均值超矢量 (supervector), 然后通过因子分析将其投影为低维、定长的矢量表示, 其特点在于它能在较大的粒度范围内提取语音特征, 可以作为一段语音信号的整体描述, 这使得 i-vector 作为无需语言模型支持的置信度判决的输入特征成为了可能. 另一方面, PLDA 方法最早发源于图像领域的人脸识别应用, 通过增大类间差异, 达到补偿识别过程中无关因素的作用. 由于在判决阶段, PLDA 通过计算假设检验的比值 (也即似然比) 打分, 因此其可以自然地作为一种置信度分析手段.

本文首先对汉语的 1254 个全音节孤立字以及连接词进行了置信度实验, 考察了基于 i-vector 和 PLDA 方法在置信度判决中的有效性. 在连接词实验中分析发现, i-vector 特征对语音在全局层面上的刻画能力较强, 但是对于语音中的时序特征的辨识, 例如音节发音顺序辨识, 其存在一定的模糊性, 而时序信息是语音语义的重要成分, 这在命令词识别中是不可回避的问题. 针对此缺陷, 本文在实验验证的基础上, 尝试提出了改进方法, 较好地解决了此问题.

1 置信度与基线系统概述

1.1 置信度及其评价方法

语音识别系统的性能在过去几十年中取得了长足的进步, 但环境噪声、非对话内容等干扰因素依然是语音控制这类系统在实际应用中面临的一大挑战. 引入置信度模型, 通过后处理排除识别结果中的无关内容, 是提高系统可靠性的一个有效思路.

在语音识别中, 置信度代表某一语音 X 来自模型 W 的可信程度. 文献 [1] 中对置信度予以综述, 其中将置信度估计方法大致分为 3 类: 1) 基于预测特征的组合, 即收集解码过程中各环节的相关特征并融合为判据; 2) 基于后验概率, 即使用识别过程中的后验概率; 3) 基于似然值比值 (似然比): 将置信度转换为一个假设检验问题处理, 零假设$\mathcal{H}_0$表示语音 X 来自模型 W, 对立假设$\mathcal{H}_1$反之. 根据 Neyman-Pearson 准则, 对上述假设的最优检验为似然比检验:

$$LRT = \frac{p(X|\mathcal{H}_0)}{p(X|\mathcal{H}_1)} \geq \tau ? \tag{1}$$

其中, τ为与虚警概率相关的阈值.

在置信度估计中一般会遇到两类错误: 第 1 类错误 (漏报), 即实际情况符合零假设$\mathcal{H}_0$时, 检验结果拒绝$\mathcal{H}_0$; 第 2 类错误 (虚警), 即实际情况不符合零假设$\mathcal{H}_0$时, 检验结果接受$\mathcal{H}_0$. 两类错误以及它们衍生出的接收者操作特征 (Receiver Operating Characteristic, ROC) 曲线、检测错误权衡 (Detection Error Tradeoff, DET) 曲线、等错误率 (Equal Error Rate, EER) 等均为评价置信度的统计手段. 根据置信度在语音识别中的应用场景, 可以在帧搜索阶段就融入置信度得分信息, 达到实时剪枝提高识别率的作用, 也可以作为后处理方法, 对识别结果的正确性进行检验. 对于后者, 在实际应用中更关注根据置信度进行拒识后对系统性能的影响, 可以采用拒绝率 (Rejection Rate, RR) 和拒绝后的识别准确率 (Accuracy after Rejection, AR) 来考察置信度在语音命令识别中的作用:

$$\begin{aligned} RR &= \frac{\text{被拒绝的语音数}}{\text{语音总数}} \\ AR &= \frac{\text{被正确接受的语音数}}{\text{被接受的语音数}} \end{aligned} \tag{2}$$

1.2 基线系统

在基于 GMM-HMM 的语音识别系统的识别过程中, 语音识别器对每次输出能给出 N-best 候选的似然值得分. 在基线系统中我们采用首选输出的似然值得分与次优候选的似然值得分之比来作为置信度判断的依据, 对识别结果进行后处理, 简单易行且应用广泛.

$$LR = \frac{p(X|w_1)}{p(X|w_2)} \tag{3}$$

其中, $p(X|w_1)$为首选似然值, $p(X|w_2)$为次优候选的似然值, 二者均已根据帧长度做归一化. 似然比$LR \geq \tau$则接受w_1作为首选识别结果.

2 基于 i-vector 和 PLDA 的置信度判决方法

2.1 通用背景模型

传统的语音识别系统常常是通过训练一个高斯混合模型 (Gaussian Mixture Model, GMM), 对其语音特征的分布进行建模, 通过求取并比较测试语音在不同 GMM 上的似然值确认其相似程度, 完成识别. 但是实际应用中, 用于训练特定 GMM 的语音往往长度较短或语料较少, 导致训练数据不足, 无法训练出高质量的 GMM 模型; 另一方面也存在大量未标注的语料, 其中的信息无法被利用. Reynolds 等人提出的通用背景模型 (Universal Background Model, UBM)[6] 利用所有数据训练得到一个混合分量数较高的 GMM 模型, 其代表了全局语音特征的分布情况. 训练得到 UBM 模型之后, 通过自适应算法适应特定语句的数据, 可以得到各语句的 GMM 模型, 其特征分布随语句内容而不同, 可用于识别确认.

UBM 的训练采用传统的 EM 算法, 反复迭代更新 UBM 各分量的权重w_i、均值μ_i、方差Σ_i. 在自适应阶段, 对于给定语音数据$x = x_1, x_2, \cdots, x_t, \cdots, x_T$, 实际应用中一般采用最大后验概率 (Maximum A Posteriori, MAP) 算法, 且只更新 UBM 的均值. 首先计算数据x_t与 UBM 中第 i 个分量的相似度:

$$\begin{cases} \Pr(i|x_t) = \dfrac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)} \\ p_i(x_t) = \dfrac{\exp\left(-\dfrac{1}{2}(x_t-\mu_i)^T \Sigma_i^{-1}(x_t-\mu_i)\right)}{\sqrt{(2\pi)^D |\Sigma_i|}} \end{cases} \tag{4}$$

然后计算充分统计量:

$$\begin{cases} N_i = \sum_{t=1}^{T} \Pr(i|x_t) \\ F_i = \sum_{t=1}^{T} \Pr(i|x_t)\, x_t \end{cases} \tag{5}$$

最后计算新均值$E_i(x)$, 并与原均值μ_i加权融合:

$$\begin{cases} E_i(x) = \dfrac{F_i}{N_i} \\ \hat{\mu}_i = \alpha_i E_i(x) + (1-\alpha_i)\mu_i \\ \alpha_i = \dfrac{N_i}{N_i + r} \end{cases} \tag{6}$$

其中, α_i称作自适应系数, 用于控制新旧参数对 UBM 的影响. 在特征空间中, x_t的分布只能覆盖到 UBM 的部分分量, 这些分量的N_i较高, 相应地α_i也较高, 更新的均值$\hat{\mu}_i$倾向于在数据x上训练得到的$E_i(x)$; 类似地, 未被覆盖到 (数据量不足) 的分量, 其$\hat{\mu}_i$倾向于 UBM 中经充分背景数据训练得到的μ_i. 通过根据数据分布情况有选择地调整 UBM 参数, 能够获得与数据相匹配且高质量的 GMM 模型.

2.2 i-vector 模型

前述 GMM-UBM 方法得到的特定 GMM 模型可以用于常规的 GMM 似然值得分确认, 但考虑到各 GMM 的均值足以代表特征的分布情况, 因此可以将均值拼接起来, 称为均值超矢量, 作为反映变长语音特性的一种定长特征, 其同样包含了说话内容等信息. 常见的利用此超矢量的方式包括将其送入支持向量机 (Support Vector Machine, SVM) 等分类器中训练判别[7], 或通过联合因子分析 (Joint Factor Analysis, JFA)[8] 对超矢量建模并进行分解:

$$M = m + Vy + Ux + Dz \tag{7}$$

其中, M 为语音的超矢量, m 一般取 UBM 的均值超矢量; V为本征语音 (eigenvoice) 矩阵, y为语音因子; U为本征信道矩阵, x为信道因子; D 为残差矩阵 (对角阵), z 为残差因子. y、x、z 均服从标准高斯分布. 通过训练 V、U、D 矩阵, 对语音和信道空间分别建模并求解, 理论上可以得到仅包含有用信息的因子y作为新的语音特征.

然而在文献 [9] 中, Dehak 等人通过实验发现上述分离方法较为理想化, 在信道因子中同样存在语音信息, 并在文献 [4] 中提出了 i-vector 模型:

$$M = m + Tw \tag{8}$$

其中, T 表示的全局差异空间 (total variability space) 包含了说话内容、信道等各方面的信息, w为全局因子, 服从标准高斯分布, 又称为身份矢量 (identity vector, i-vector). i-vector 模型可以看做 JFA 的简化, 不再试图完全分离无关信息, 而是使用全局差异空间同时予以刻画. i-vector 主要起对均值超矢量的降维作用, 与均值超矢量同样包含说话内容相关的信息, 文献 [10] 等已有研究中通过实验证明其确实对内容具有一定的鉴别能力; 另一方面, 由于均值超矢量代表语句整体的特征分布, 未包含语句中音节内容的时间顺序信息, 因此基于 UBM 均值超矢量产生的 i-vector 特征类似地具

备对语音片段的全局刻画能力, 而对内容的时序信息缺乏更精确的描述.

使用 EM 算法训练 T 矩阵[11]. 对于给定语音数据 $\boldsymbol{x}=x_1,x_2,\cdots,x_t,\cdots,x_T$, 由式 (5) 中的充分统计量$N_i$、$F_i$得到中心化一阶统计量:

$$\tilde{F}_i = F_i - N_i\mu_i \tag{9}$$

将x在各分量i上的统计量拼接为$N(x)$、$\tilde{F}(x)$.

令 UBM 的均值超矢量、方差为m、Σ, 并随机初始化矩阵 T. E 步骤中, 更新隐变量w的后验分布:

$$\begin{cases} l(x) = I + T^{\mathrm{T}}\Sigma^{-1}N(x)T \\ E[w(x)] = l^{-1}(x)T^{\mathrm{T}}\Sigma^{-1}\tilde{F}(x) \\ E\left[w(x)w^{\mathrm{T}}(x)\right] = l^{-1}(x) + E[w(x)]E[w(x)]^{\mathrm{T}} \end{cases} \tag{10}$$

M 步骤中, 更新矩阵 T:

$$\begin{cases} A_i = \sum_x N_i(x)E\left[w(x)w^{\mathrm{T}}(x)\right] \\ C = \sum_x \tilde{F}(x)E[w(x)]| \\ T = \begin{bmatrix} A_1^{-1}C_1 \\ A_2^{-1}C_2 \\ \vdots \\ A_C^{-1}C_C \end{bmatrix} \end{cases} \tag{11}$$

反复迭代更新得到矩阵 T 后, 语音 x 的 i-vector 为:

$$w = \left(I + T^{\mathrm{T}}\Sigma^{-1}N(x)T\right)^{-1}T^{\mathrm{T}}\Sigma^{-1}\tilde{F}(x) \tag{12}$$

2.3 概率线性判别分析模型

概率线性判别分析 (Probabilistic Linear Discriminant Analysis, PLDA) 最早由 Prince 等在文献 [5] 中提出, 应用于图像识别中的人脸识别任务. PLDA 的原始形式如下:

$$w_{ij} = \mu + Vy_i + Ux_{ij} + z_{ij} \tag{13}$$

其中, w_{ij}为第i个人的第j次采样特征, μ为全局均值, V表示类间差异空间, U表示类内差异空间, z_{ij}为残差. $\mu+Vy_i$是w_{ij}的信号分量 (只与i相关), $Ux_{ij}+z_{ij}$是噪声分量.

与此前常用的线性判别分析 (Linear Discriminant Analysis, LDA)[12] 相比, PLDA 同样试图寻找数据的某种低维投影, 使得投影后类间差异最大, 但 PLDA 是一种生成式 (generative) 模型, 考虑了图像由信号与噪声两部分组成并予以显式建模, 噪声模型更为完备, 因而取得了更好的效果.

在原始 PLDA 模型的基础上, 由于在语音相关任务中无需求解类内差异, 文献 [13] 中引入了简化的 PLDA 模型:

$$w_{ij} = \mu + Vy_i + z_r \tag{14}$$

其中, 隐变量y_i服从标准高斯分布, 类内差异被合并为z_r, 其协方差为Σ.

使用 EM 算法训练 PLDA 模型, 迭代优化完全数据的对数似然函数的期望Q得到最合理的参数$\theta=\{\mu,V,\Sigma\}$:

$$Q(\theta,\theta_{t-1}) = E_y\left\{\sum_{i,j}\log\left[p\left(w_{ij}|y_i,\theta\right)p(y_i)\right]|w,\theta_{t-1}\right\} \tag{15}$$

随机初始化矩阵V、Σ. E 步骤中, 更新隐变量y的后验分布只需估计其均值和方差:

$$\begin{cases} E(y_i) = \left(nV^{\mathrm{T}}\Sigma^{-1}V + I\right)^{-1}V^{\mathrm{T}}\Sigma^{-1}w_i \\ E\left(y_iy_i^{\mathrm{T}}\right) = \left(nV^{\mathrm{T}}\Sigma^{-1}V + I\right)^{-1} + E(y_i)E(y_i)^{\mathrm{T}} \end{cases} \tag{16}$$

M 步骤中, 更新矩阵V、Σ:

$$\begin{cases} V = \left(\sum_i w_iE(y_i)^{\mathrm{T}}\right)\left(\sum_i E\left(y_iy_i^{\mathrm{T}}\right)\right)^{-1} \\ \Sigma = \frac{1}{N}\sum_i\left(w_iw_i^{\mathrm{T}} - VE(y_i)w_i^{\mathrm{T}}\right) \end{cases} \tag{17}$$

其中, N为训练 i-vector 总数, n为y_i所属的语句对应的 i-vector 总数.

测试阶段, 给定两条待比对的 i-vector w_1、w_2, 假设$\mathcal{H}_s$表示二者由相同的因子y生成, $\mathcal{H}_d$表示二者由不同的因子y生成, PLDA 模型通过计算两种假设的似然值得分给出w_1、w_2之间的相似度:

$$\begin{aligned} score &= \log\frac{p(w_1,w_2|\mathcal{H}_s)}{p(w_1,w_2|\mathcal{H}_d)} \\ &= \log\mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix};\begin{bmatrix} \mu \\ \mu \end{bmatrix},\begin{bmatrix} \Sigma_{\mathrm{tot}} & \Sigma_{\mathrm{ac}} \\ \Sigma_{\mathrm{ac}} & \Sigma_{\mathrm{tot}} \end{bmatrix}\right) \\ &\quad -\log\mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix};\begin{bmatrix} \mu \\ \mu \end{bmatrix},\begin{bmatrix} \Sigma_{\mathrm{tot}} & 0 \\ 0 & \Sigma_{\mathrm{tot}} \end{bmatrix}\right) \end{aligned} \tag{18}$$

其中, $\Sigma_{\mathrm{tot}}=VV^{\mathrm{T}}+\Sigma$, $\Sigma_{\mathrm{ac}}=VV^{\mathrm{T}}$.

与 1.1 节中基于似然比的置信度对比可以发现, PLDA 可以比较自然地作为一种置信度计算方法, 以语音整体的 i-vector 作为输入, 不依赖声学模型和语言模型即可完成似然比检验.

2.4 孤立字语音识别置信度检验实验

音节是汉语发音的基本单元, 因此考察 i-vector 特

征对音节的置信度的检测能力, 是该方法能否成功应用于连接词识别置信度检验的基础. 本实验采用 IsoWord 孤立字数据集, 其包含了 50 名男性、50 名女性、每人 1254 个有调音节, 覆盖了汉语的全部具有实义的音节, 采样率 16 kHz. 随机选取 1 名男性的语音样本作为测试集, 其余作为训练集. 使用 1.1 节中的拒绝率和拒绝后的识别准确率评价系统的性能.

本文采用 45 维 MFCC 特征, 对输入的单帧语音信号, 去除直流, 预加重 (系数取 0.98), 加汉明窗 (帧长 20 ms、帧移 10 ms) 后, 提取 14 维 Mel 倒谱系数, 对相邻帧计算一阶、二阶差分系数, 并加入三者的归一化能量系数.

为了确定理想的模型参数, 本文首先在识别率有代表性的数据样本上进行了调参实验, 调整的参数包括 UBM 混合分量数和 i-vector 维度. 可以观察到不同参数的组合对性能有微小的影响. 以第 25 号孤立字男声样本为例, 实验结果见表 1 和表 2.

表 1 UBM 混合分量数对性能的影响

UBM混合分量数	i-vector维度	AR(%)
32		83.80
64		84.47
128	100	**84.47**
256		83.63
512		83.88

表 2 i-vector 维度对性能的影响

UBM混合分量数	i-vector维度	AR(%)
	50	84.30
	75	84.30
128	**100**	**84.47**
	150	84.13
	200	83.54

根据调参实验结果, 本文在孤立字的置信度判决实验中, UBM 模型混合数取 128, i-vector 维数取 100.

在确定了模型参数后, 在训练阶段, 首先训练 UBM 模型, 对每条语音计算所需的充分统计量, 然后训练 i-vector 模型的 T 矩阵. 参照文献 [14] 中的建议, 使用已知的语音对应的说话内容作为训练标签, 对 i-vector 预先做 LDA 降维, 从而初步补偿类间差异. 由于文献 [13] 中发现 i-vector 具有较强的非高斯性, 为使其符合前述基于高斯假设的 PLDA 模型, 参照文中建议对 i-vector 做白化与长度规整后, 再训练 PLDA 模型, PLDA 因子维度与 LDA 维度相同 (不再做进一步降维). 每条语音的代表 i-vector 取该语音所有说话人语音样本对应的 i-vector 的均值. 测试阶段, 将测试语音通过训练集上训练好的 UBM 模型、T 矩阵、LDA 矩阵, 得到测试 i-vector, 在 PLDA 模型上与每条语音的代表 i-vector 逐对计算似然值得分. 部分实验流程使用 MSR Identity Toolbox[15] 完成.

图 1、图 2 为采用基线系统和 i-vector+PLDA 对随机两组男声孤立字各 1254 个发音样本置信度检测的 RR-AR 曲线. 可以看出, 不论是对于原本识别率较低还是较高的男性语音样本, i-vector+PLDA 都能通过拒识提高其性能, 且效果较基线系统有一定的提升.

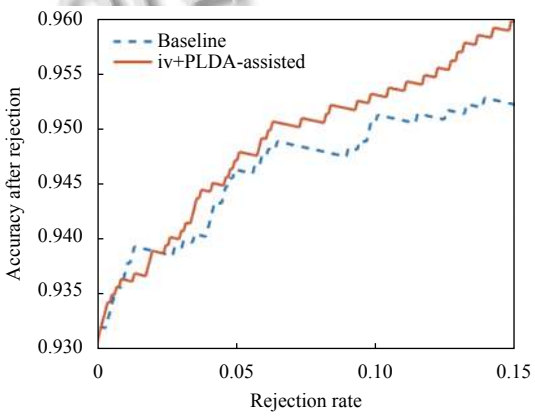

图 1 第 17 号孤立字男性语音样本上的 RR-AR 曲线

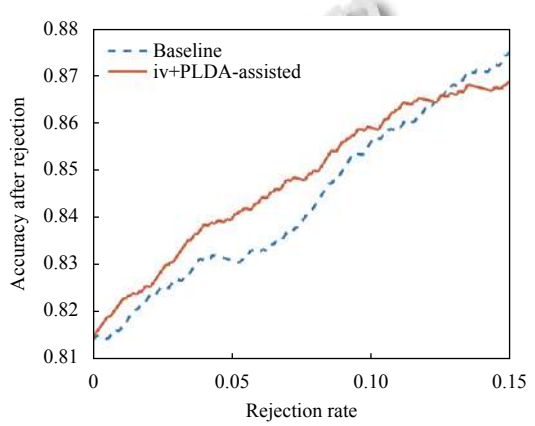

图 2 第 50 号孤立字男性语音样本上的 RR-AR 曲线

表 3 为将 RR 固定为 5% 时, 各系统在所有男声样本上轮流训练测试的平均性能, 其中原始无置信度辅助的 GMM-HMM 孤立字识别系统的平均识别率是 89.81%. 可以看出置信度的拒识使系统输出的正确率绝对提高约 2%, 且 i-vector+PLDA 相比基线系统再绝对提高约 0.3%.

表 3 置信度辅助的系统在 IsoWord 数据集上的性能

系统	正确率(%)
GMM-HMM(无置信度辅助)	89.81
基线置信度系统	91.67
i-vector+PLDA 置信度系统	91.92

2.5 连接词短语置信度检验与拒识实验

连接词实验采用的 SbPhrase 短语语料数据库包含了 699 条四字短语, 较为均衡地覆盖了所有的汉语音节及音节间的连接关系, 可以较为客观地评测命令词系统的一般性能. 该数据库包含了 50 名男性、50 名女性的录音样本, 采样率 16 kHz, 每条短语时长约 1 s.

置信度检验实验中, 以前 25 名男性的所有短语语音作为训练集, 训练 GMM-HMM 系统, 其余男性语音作为测试集, 并使用置信度对识别结果做后处理. MFCC 特征提取与 2.4 节相同, 根据实验调整 i-vector 提取参数为 512 分量 UBM、200 维 i-vector.

图 3、图 4 为随机两个男声连接词短语样本置信度检测的 RR-AR 曲线.

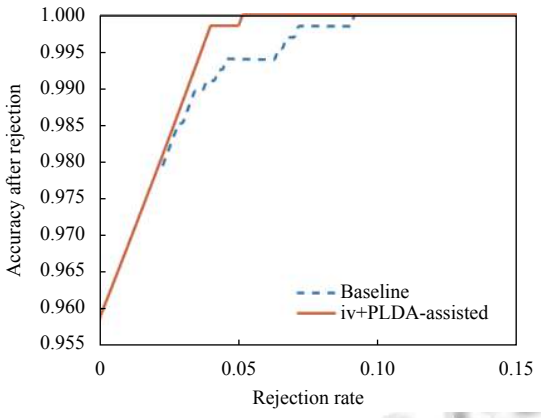

图 3 第 30 号男性连接词语音样本上的 RR-AR 曲线

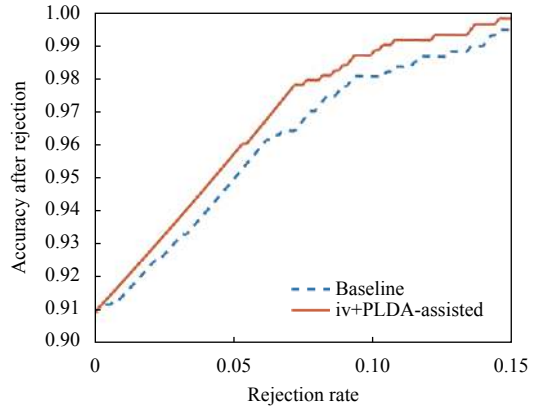

图 4 第 41 男性连接词语音样本上的 RR-AR 曲线

表 4 为将 RR 固定为 5% 时, 各系统在所有样本上的平均性能, 其中原始 GMM-HMM 连接词识别系统的平均识别率是 95.97%. 与孤立字类似地, 置信度的引入提高了系统的识别性能, 而 i-vector+PLDA 的效果更佳.

表 4 置信度辅助的系统在 SbPhrase 数据集上的性能

系统	正确率(%)
GMM-HMM(无置信度辅助)	95.97
基线置信度系统	99.10
i-vector+PLDA置信度系统	99.34

在命令词识别系统中, 对集外词或噪声的有效拒识至关重要, 我们通过实验单独测试了系统的拒识性能. 仍然使用 SbPhrase 数据库的男声部分, 取数据库中的前 300 条短语作为集内词训练 PLDA 模型并确定其阈值, 其余作为集外词进行实验. 除此之外, 采用从 CMU NoiseX-92 数据集[16] 中截取的噪声片段考察系统对噪声的抵抗能力, 该数据集包含了白噪声、工厂噪声、背景说话声等常见噪声类型. 使用虚警率评价系统的性能.

表 5 中的结果表明, i-vector+PLDA 系统性能良好, 不论对语音类的集外词还是非语音类的干扰噪声都具有较高抗性, 保证了系统的稳健性.

表 5 i-vector+PLDA 系统的集外词、噪声拒识性能

系统	集外虚警率(%)	噪声虚警率(%)
i-vector+PLDA	0.65	0.52

3 融合 DTW 的置信度判决方法

3.1 i-vector 时序鉴别能力分析

在 2.1 节中已经指出, GMM-UBM 模型通过自适应得到每条语音对应的 GMM 模型, 这种建模方式的一个缺陷是不包含时序信息: 对于仅字序、词序不同的语音, 由于使用了相同或相近的音素, 全局上看, 各自的特征集内其特征分布彼此相似, 因而在这类系统上会体现为相似度较高. 换言之, 虽然 i-vector 的全局描述能力较好, 但缺乏对其中时序信息的描述, 理论上, 若单独使用 i-vector 特征, 对于较长的命令词导致鉴别力下降的可能性会增大. 在实际应用中, 这会导致部分与命令词在字序、词序上相似的集外词无法被系统有效拒识, 引发不必要的虚警.

有鉴于 i-vector 的上述特点, 一种解决方法是利用命令词识别系统在识别时给出的最佳音节分割点, 对组成命令词的每个音节或是单词分别进行确认, 如在

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.

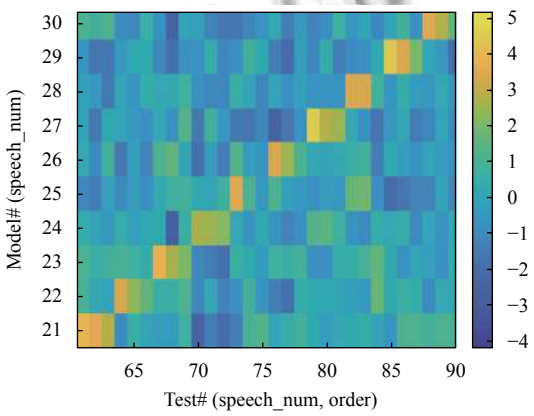

图 5 原系统的混淆矩阵 (部分)

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.

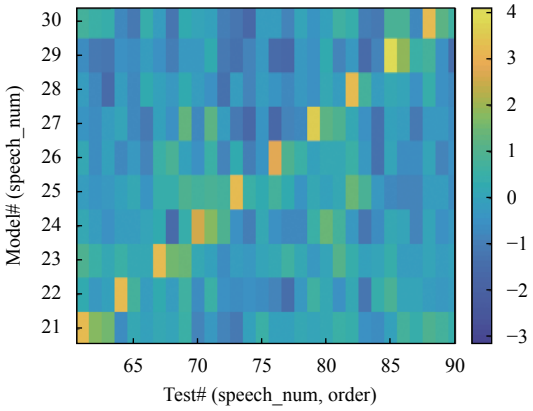

图 6 新系统的混淆矩阵 (部分)

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]