CSLT TECHNICAL REPORT-20180807 [Wednesday 26th December, 2018]

语种识别的相关研究

Qi Zhaodi

Correspondence:
zdqi0707@163.com
School of Physics and Electronic Engineering, Jiangsu Normal University, 221116 Xuzhou,Jiangsu, China
Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

语种识别是利用机器学习算法对输入到计算机的语音自动判别语言种类的一种技术，是多语言智能语音系统不可或缺的关键组合部分之一。目前世界上现存的语言种类大约6909种，其中拥有书面文字的语言多达2000多种，并且大部分语言之间差别很大。而伴随着全球化的推进，多语言交流日益被需要，人类迫切地需要多语言智能语音技术提供技术支持。本文对语种识别近些年使用方法，如基于音素识别器、基于底层声学特征以及基于深度学习的语种识别进行了一定的调研。

Keywords: 语种识别; 机器学习; 音素识别器; 声学特征; 深度学习

1 引言

语种识别最早是由Leonard和Doddington在1974年提出的[1]。由于当时数据的缺乏，使研究结果具有局限性，结论不具有普遍性。直到美国国家标准技术研究院在上世纪90年代开始组织的语种识别评测比赛，语种识别技术才得以快速发展。目前主要的语种识别方法主要分为基于音素识别器、基于底层声学特征和基于深度学习三种。本文的研究内容主要围绕当前语种识别技术，并简要阐述其基本概念、理论及算法。

2 语种识别简介

对计算机来说，语种识别可以看作为模式识别问题。一个典型的语种识别系统主要分为三方面，其内容分别是特征提取、模型建立和模型分类，如Figure 1 所示。

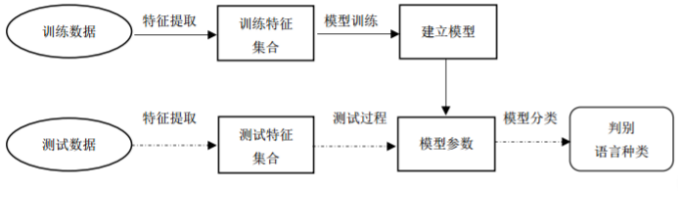

Figure 1 语种识别基本流程图

研究结果表明，人类自身在进行语种识别时常用的区分性特征按照从低层到高层可以划分为底层声学特征（MFCC、SDC、PLP等）、韵律特征（时长、基频、重读等）、音素（音素级别的N-gram）、词法和语法[2]。底层声学特征包括频谱、倒谱、谱包络以及共振峰，被认为是声学信号所固有的物理特征，其他特征都是建立在底层声学特征的基础之上的更高层次的特征，来自于其低层信息的分析和提取。越往高层，特征区分语种的能力越强，然而特征提取的代价也越高。越往底层，特征中所含的冗余信息越多，特征区分语种的能力越弱，但是特征提取的代价越低。语言学上的研究表明，语言中可用于语种分类的特性在每一层信息中都有着不同的体现。目前的大部分语种识别研究方法集中在底层的声学特征和音素特征上。词法和语法特征因其对语言学知识和语音识别器的依赖性而鲜有研究。而韵律特征的提取虽然也比较简单直接，但是由于其对于语种信息的反映较弱和性能差而不被推广。

3 存在的挑战

语种识别虽然已经发展了近六十年，在性能上也取得了很大的进步，但其仍然面临着严峻的挑战，主要表现在以下三个方面：

(1) 外界噪声

随着技术的突飞猛进，不同噪声、信道环境下的语音数据呈爆炸性增长。语音数据可能来自于电话信道、广播信道以及不同频段的移动电话信号等，噪声环境也有可能来自于各种复杂恶劣的背景环境。语种识别的研究从传统的单一信道、单一语音环境转向了现代的复杂信道、恶劣噪声环境。传统的语种识别特征和建模方法下的性能迅速恶化，已经无法达到实用要求。

(2) 短时语音的需求

语种识别在做多语言语音技术前端处理或做其它需要系统能够在短时语音下迅速准确地做出语种类别判决时，对实时性要求很强。而目前的方法都是基于统计的思想，识别结果很大程度上依赖于统计量的精确程度，因此短时语音的识别效果无法达到实用标准，更无法满足多语言人机交互的需求。

(3) 易混方言或者口音识别任务

目前语种识别的任务已经不满足于区分汉语、英语和俄语等这些极易区分的语言种类，逐渐转向了易混的方言或者口音识别任务，NIST在2011年组织的LRE任务就明确地从传统的多类语种识别任务转变成了易混语种对的语种识别任务。对于这类问题，不论是声学特征还是音素搭配反应出来的差异都很弱。在智能语音技术中，针对这类任务也往往需要单独构建语料库设计系统，因此对这类语种识别问题的研究越来越迫切。

从上述三种挑战发现，语种识别任务正向着复杂场景、实时性以及精细语言种类识别转变。基于现有的声学特征建模方法与基于音素识别的建模方法都难于取

得重大进展，主要原因在于底层声学特征对于噪声的鲁棒性的影响极易受到说话人、信道、环境噪声以及特定说话内容的干扰和语音识别器又受限于对这些场景条件下语音信号的识别准确率，从而影响特征对语种的区分效果。最近几年，深度学习(Deep Learning)理论在语音识别和图像识别领域取得了令人振奋的性能提升，迅速成为了当下学术界和产业界的研究热点，为处在瓶颈期的语音和图像等模式识别领域提供了一个强有力的工具，所以产生了基于深度学习的语种识别的热潮。

4 语种识别的研究方法

本节内容将对目前语种识别领域的主要研究方法进行调研，包括基于音素识别器的语种识别方法、基于底层声学特征的语种识别方法，并在此基础上详细介绍主流识别方法。

4.1 基于音素识别器的语种识别方法

基于音素识别器的语种识别方法认为不同的语言之间相同的音素搭配体现的统计特性是有差别的，因此可以用来进行语种识别。具体实现而言，首先通过音素识别器将语音信号转换为音素序列，然后根据音素序列提取N-gram单元统计量作为特征，最后根据这些统计特性建立每个语种的N-gram 语言模型(Language Model，LM)。该方法在语种识别领域中被称之为音素识别器结合语言模型(Phone Recognizer followed by Language Model，PRLM) [3] [4]。

为了充分利用不同阶数的描述能力，克服语种中不同阶N-gram因语料过少而造成的稀疏性问题，研究者在语言模型的基础上，又提出了二叉决策树(Binary Tree，BT)模型。称为音素识别器结合二叉决策树(Phone Recognizer followed by Binary Tree，PRBT)方法。由于每个语种的音素单元不一致，因此语种的N-Gram差异在不同的语种音素识别器上反映的差异有所不同。那么采用多个不同语种的语音识别器，分别将同一语音段转换成不同语种音素集合下的音素序列，然后在判决得分上对所有子系统进行融合。该方法称之为并行音素识别器(Parallel Phone Rcognizer，PPR) 方法，相应的就有并行音素识别器结合语言模型(PPRLM)和并行音素识别器结合二叉树模型(PPRBT)。

之后，MIT的研究者们认为传统的1-best识别序列易受识别精度的影响，丢失掉了许多次优路径信息，因此提出了利用词图(Lattice)代替最优识别序列来提取N-gram统计量的方法。

随着支持向量机模型(SVM) [5]的发展，研究者们又提出利用N-gram 基元构建词袋矢量bag-of-Ngram，然后利用SVM来构建模型的PPRSVM 方法。由于SVM是区分性模型，较语言模型而言建模能力更强，并且SVM模型对于小样本情况鲁棒性更强，因此成为了目前基于音素识别器方法中的主流系统。在SVM之后，大量的研究工作都集中在如何解决N-gram的稀疏性以及如何挑选出最具有区分性的N-gram的单元等问题。近年来，受全差异建模方法在基于声学特征的成功应用

的启发，研究者们提出了利用bag-of-ngram提取i-vector的方法，该方法能够将高维的bag-of-Ngram矢量变换到低维的表示空间中，从而降低分类器的建模难度，提升系统性能的目的。

4.2 基于底层声学特征的语种识别方法

基于底层声学特征的语种识别方法是利用底层声学特征所能够描述的声学单元的统计特性差异来对语种进行分类。由于该方法所需要的特征直接通过底层的谱参数得到，不需要音素识别器作为支撑，因此一直以来都是语种识别研究的热点。

第一个突破就是2002年提出的SDC特征[6]结合混合高斯模型-通用背景模型(Gaussian Mixture Model-Universal Background Model，GMM-UBM)的方法。在这之后，基于SDC特征和GMM模型的改进方法不断涌现出来。为了克服底层声学特征易受说话人、信道、噪声以及内容信息差异影响的问题，在特征域层面提出了抑制卷积信道噪声影响的倒谱域减均值(Cepstral Mean Subtraction，CMS)、去掉说话人影响的声道长度规整(Vocal Tract Length Normalization，VTLN)、特征高斯化以及RASTA滤波等方法。

第二个突破就是区分性建模方法的引入。传统的GMM-UBM模型是一个典型的生成性模型，它并不能很好的对不同类别之间的易混部分进行分类。因此在GMM的基础上，研究者们分别提出了利用区分性训练准则最大互信息准则(Maximum Mutual Information，MMI)训练语种GMM模型的GMM-MMI方法，利用区分性的SVM模型来对对每段语音的GMM均值超矢量进行建模的GSV-SVM方法以及利用GSV-SVM方法反推语种GMM模型的model pushing方法。

第三个突破就是因子分析方法(Factor Analysis，FA)。该方法借鉴于说话人识别当中得到成功应用的联合因子分析方法(Jiont Factor Analysis，JFA)，在底层声学空间中对信道噪声进行子空间建模，然后通过特征域的去噪或者模型域的补偿来去除噪声的影响。近年来，研究者们意识到构建的子空间中仍然包含着目标的有效分类信息，因此在此基础上提出了基于全差异空间建模的方法(Total Variability，TV)，该方法围绕样本的GMM超向量与均值超向量之间的差异，将每个样本视为独立的个体，训练得到每个样本之间的全差异空间，然后得到样本之间差异的低维表示称之为i-vector之后通过线性区分性分析或者类内协方差规整等技术对i-vector进行类内类间差异补偿和降维，再采用SVM或者快速余弦距离来进行建模。目前，TV[7]方法(或者称为i-vector方法)因其低维的语音段表示以及良好的性能成为了语种识别领域的主流系统。

以下分5个点详细介绍上面提及的主流算法。

4.2.1 SDC特征

SDC主要是在底层谱参数特征MFCC或者PLP的基础上通过移位差分扩展而来。一个典型的SDC特征提取流程如Figure 2所示。

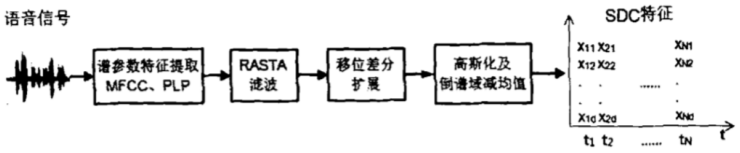

Figure 2 典型的SDC特征提取流程框图

首先对语音加窗提取其底层声学谱参数特征，在语种识别任务中一般提取MFCC特征或者PLP特征。在提取完谱参数特征后，采取RASTA 滤波(Relative Spectral filtering)来抑制参数表示中非语音频谱部分的影响。对于提取出来的第t帧静态谱特征N维静态谱参数特征c(t)，其对应的第k个一阶差分矢量计算表达式为

$$\Delta c(t,k) = c(t+(k-1)P+d) - c(t+(k-1)P-d) \tag{3.1}$$

此时，得到的SDC特征就是将静态特征和k个移位差分矢量$\Delta c(t,k)$拼接起来，形成最终的SDC特征x_t。

$$x = \begin{bmatrix} c(t) \\ \Delta c(t,0) \\ \Delta c(t,1) \\ \cdots \\ \Delta c(t,k-1) \end{bmatrix} \tag{3.2}$$

一个典型的移位差分计算如Figure 3所示，它的计算主要由4个参数N-d-P-k描述。其中N表示提取出来的静态参数特征维数，d表示计算一阶差分的帧与参考帧的时移距离，P表示参考帧的跳帧长度，i表示参考帧的个数，由于其在每个参考帧下都会得到一个差分矢量，因此也表示为一阶差分矢量单元数目。

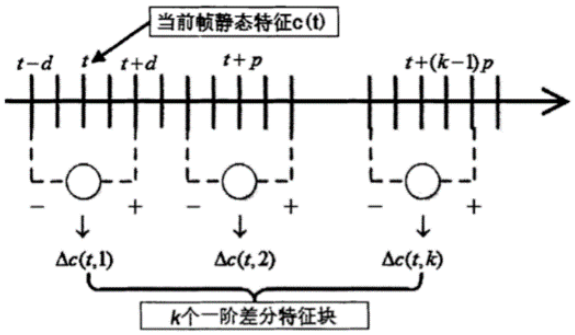

Figure 3 移位差分计算示意图

在得到SDC特征后，一般还采用倒谱域件均值(CMS)来去除信道的卷积噪声以及特征高斯化技术对语音的参数进行规整，这些技术就构成了目前SDC特征提取的标准流程。这样就得到了一段语音s的D维帧级(Frame-level)特征表示X=X_1，...，$X_T n$，其中T_s表示语音段s的总帧数。

4.2.2 GMM-UBM语种识别方法

UBM其实就是一个大型的GMM模型，用来训练表示与说话人无关的特征分布。它的训练数据是某一信道下的所有人的语音数据，而不是想target模型只是反映某一个人的特征分布。说白了，只是一个大的GMM，那么训练UBM也就是训练GMM，所用算法采用的是EM算法。

GMM中，从说话人语音抽出来的D维特征矢量对应的似然率可用K个高斯分量表示：

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{3.3}$$

其中是第K个高斯分量的权重，$\sum_{i=1}^{M} w_i = 1$

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^{'}(\sum_i)^{-1}(x-\mu_i)\right\} \tag{3.4}$$

所以，整个高斯模型可以用模型参数$\lambda = \{w_i, \mu_i, \sum_i\}$，i=1,2,...,m来表示。

SDC和GMM模型的出现，给促进了基于底层声学特征语种识别方法。由于它不像PR方法那样对识别器有很强的依赖，建模方法的实现更加容易。由此，大量的研究者们开始致力于基于声学特征的语种建模方法研究，这期间区分性建模方法和因子分析方法的引入是两个最具影响力的进展。

4.2.3 GMM-MMI语种识别方法

由语音识别中区分性训练的成功应用启发，研究者们开始将区分性训练准则引入到GMM模型的训练领域，其中基于最大互信息准则(MMI)的训练方法最为成功。最大互信息准则实质上是最大化训练数据真实类别的后验概率。假设训练集合语音段共有N句S=s_1 , ...，s_n，对应的每句语音段提取出来的特征集合表示为X=X_1, ...，X_n，其中Xn=$x_n, 1, ..., x_n, Tn$)，那么MMI准则的目标函数可以表示为

$$L_{MMI}(\Lambda_l|\chi) = \frac{1}{N}\sum_{n=1}^{N}\frac{p(\chi_n|\Lambda_{ln})|p(l_n))}{\sum_l p(\chi_n|\Lambda_l)p(l)} = \frac{1}{N}\sum_{n=1}^{N} p(\Lambda_{ln}|\chi_n) \tag{3.5}$$

其中l_n表示语音段s_n真实的语种标记，$p(\lambda_{ln}|X_n)$表示语音段s_n 的属于真实类别ln的后验概率，p(l)表示语种l的先验概率，一般而言认为语种的是先验是等概率的，因此一般可以忽略掉。

4.2.4 GSV-SVM语种识别方法

高斯均值超矢量(GMM Super Vector, GSV)来源于GMM-UBM。与SVM组成GSV-SVM模型，广泛用于语种及说话人识别。GSV以GMM模型的均值（或者方差）超矢量作为输入特征序列，避免了直接使用带噪语音信号的特征参数，在实际应用中获得了更好的实验效果。

利用前文GMM-UBM模型，通过MAP自适应获得代表语音段的GSV特征。GMM均值超矢量的计算过程如Figure 4：

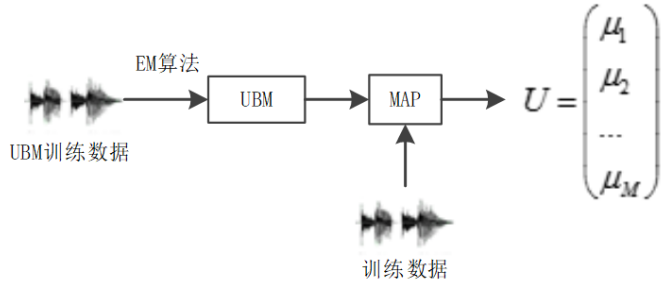

Figure 4 GSV计算过程

MAP自适应的过程，是根据目标语种的训练特征向量与UBM的匹配程度，将UBM的各个高斯混元向目标语种模型“拉近”的过程。对于目标语种的训练数据O=(o_1,o_2,...,o_T)，首先计算o_t 与UBM中每个高斯模型的匹配似然度，见式(3.6)：

$$P(m|O_t,\lambda_{UBM})=\frac{w_m P_m(O_t|\mu_m,\sum_m)}{\sum_{i=1}^{M} w_i P_i(O_t|\mu_i,\sum_i)} \tag{3.6}$$

然后利用P(m—o_t ，λ_{UBM})和o_t分别计算对混合权重、均值矢量和均方值的充分估计，见式(3.7)：

$$\begin{cases} n_m=\sum_{t=1}^{T} P(m|O_t,\lambda_{UBM}) \\ E_m(O)=\frac{O_t}{n_m}\sum_{t=1}^{T} P(m|O_t,\lambda_{UBM}) \\ E_m(O^2)=\frac{O_t^2}{n_m}\sum_{t=1}^{T} P(m|O_t,\lambda_{UBM}) \end{cases} \tag{3.7}$$

然后利用这些充分统计和修正因子对第个高斯的参数进行修正，具体过程见式(3.8)：

$$\begin{cases} \bar{W_m}=[a_m^w n_m/T+(1-a_m^w n_m)w_m]\gamma \\ \bar{\mu_m}=a_m^\mu E_m(o)+(1-a_m^\mu)\mu_m \\ \bar{\sigma_m^2}=a_m^v E_m(o^2)+(1-a_m^v))(\sigma_m^2+\mu_m^2)-\bar{\mu_m^2} \end{cases} \tag{3.8}$$

其中，γ为权重的规整因子，用来保证$barW_m$的和为1,a_m^w、a_m^u、a_m^v 为第m个高斯的权重、均值和方差的修正因子，见式(3.9)：

$$a_m^\rho = \frac{n_m}{n_m + r^\rho}, \rho \in w, \mu, v \tag{3.9}$$

其中，r^ρ为常数，用来控制修正因子的变化尺度，一般取为16。由于自适应语音不够长，无法准确描述每个高斯的权重和方差，MAP过程中，一般只更新均值，权重和方差保持和原来的UBM模型一致。

4.2.5 I-vector特征

语种语音的帧级局部特征MFCC作为i-vector特征提取步骤的前端输入。为了抑制信道噪声，对MFCC差分特征进行谱均值方差归一化处理(Cepstral Mean Variance Normalization, CMVN)。具体提取步骤如Figure 5所示：

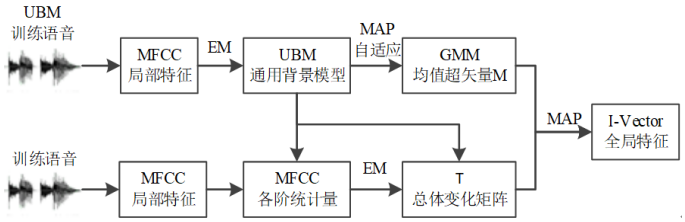

Figure 5 i-vector特征提取流程图

首先利用部分训练数据通过期望最大化(Expectation Maximum, EM) 得到UBM。然后利用最大后验概率(Maximum A Posterior, MAP)自适应得到GMM，根据JFA理论，重新定义GMM均值超矢量，见式(3.10)：

$$M = m + Tw \tag{3.10}$$

其中，M表示GMM均值超矢量，m表示一个与特定目标方言和信道都无关的超矢量，通常由UBM均值超矢量替代。全差异载荷矩阵(Total Variability Matrix)的估计是关键性环节，首先计算语音MFCC特征相对于UBM均值超矢量m的零阶N_c、一阶F_c^1及二阶F_c^2统计量，见式(3.11)、(3.12)、(3.13)：

$$N_c = \sum_{t=1}^{L} P(c/y_t, \theta_{UBM}) \tag{3.11}$$

$$F_c^1 = \sum_{t=1}^{L} P(c/y_t, \theta_{UBM})(y_t - m_c) \tag{3.12}$$

$$F_c^2 = \sum_{t=1}^{L} P(c/y_t, \theta_{UBM})(y_t - m_c)(y_t - m_c)^T \tag{3.13}$$

其中，c=1,2,…, C，m_c与$P(c/y_t, \theta_{UBM})$分别为UBM第c个高斯子分布的均值及后验概率。其次，利用各阶统计量，通过EM算法随机初始化全差异矩阵，在最大似然准则(Maximum Likelihood, ML)下估计w(即I-vector)的一阶和二阶统计量，见式(3.14)、(3.15)：

$$E_s^1(w) = L_s^{-1} T^T \sum^{-1} F_s^1 \tag{3.14}$$

$$E_s^2(ww^T) = E_s^1(w) E_s^1(w^T) + L_s^{-1} \tag{3.15}$$

其中，L_s是临时变量，具体表示见式(3.16)：

$$L_s = I + T^T \sum^{-1} N_s T \tag{3.16}$$

N_s是由N_c作为主对角元拼接得到的矩阵，F_s^1是由F_c^1直接拼接得到的矢量，I是单位矩阵，$\sum$是UBM的协方差矩阵。T 和$\sum$的更新见式(3.17)、(3.18)：

$$\sum_s N_s T E_s^2(ww^T) = \sum_s F_s^1 E_s^1(w) \tag{3.17}$$

$$\sum = N^{-1} \sum_s F_s^2 - N^{-1} diag\left\{\sum_s F_s^1 E_s^1(w^T T^T)\right\} \tag{3.18}$$

其中，$N = \sum N_s$，F_s^2由F_c^2进行矩阵拼接得到。上述步骤反复迭代6 8次后，近似认为T和$\sum$收敛。假定GMM的高斯子分布数为C，MFCC特征的维数为D，i-vector的维数为K，那么超矢量M和m的维数是C x D，全差异空间T就是CD X K的矩阵。i-vector特征矩阵计算见式(3.19)：

$$w = (I + T^T \sum^{-1} N_c T)^{-1} T^T \sum^{-1} F_c^1 \tag{3.19}$$

通过TV法提取的i-vector特征整合了帧级局部特征MFCC，以语音段为单位表征信息且与语音段长度无关。

5 基于深度学习的语种识别方法

2006年，加拿大多伦多大学Hinton教授及其学生Salakhutdinov在顶级学术刊物《科学》上发表文章提出深度神经网络模型的训练方法，掀起了深度学习在学术界及工业界的热潮。深度学习方法在语音识别、图像处理、机器翻译等领域广泛应用，各种深度网络改进算法也相继提出，为相关学术领域提供新的思路和模型，成为研究学者关注的热点。特别是在语音识别领域，DNN模型给处在瓶颈阶段的传

统的GMM-HMM 模型带来了巨大的革新，使得语音识别的准确率又上了一个新的台阶，目前国内国外知名互联网企业(谷歌、讯飞以及百度等)的语音识别算法都采用的是DNN方法。本节主要介绍近些年关于语种识别的一些方法，以及介绍其所用模型的优缺点。

5.1 DNN模型

研究者们早在2014年就发现当提供大量训练数据时，使用DNN来解决自动语音识别（LID）任务中Cavg的百分比可高达70[24]，具体实现网络结构如Figure 6 所示。

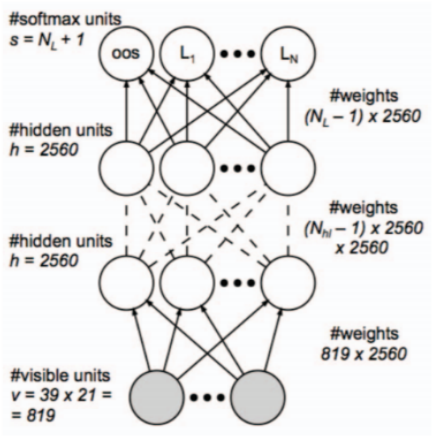

Fig. 1. DNN network topology

Figure 6 DNN network topology，图来自[24]

除了将DNN用于语种的分类模型域，由Maryam Najafian，Sameer Khurana等人[13]还利用DNN进行特征提取，知基于两个连续的深度神经网络（DNN）ASR模型可提取i-vector特征，流程图如Figure 7 所示，第一个DNN的输入包括从梅尔滤波器组获得的23个临界频带能量，第二个DNN的输入特征是从第一个DNN输出的SSD层。

5.2 RNN模型

(1) RNN

在[14]中研究了端到端的RNN及其变种LSTM,GRU的语种识别。RNN 存在收敛问题，训练网络经常面临消失梯度和爆炸梯度问题的问题。传统的基于RNN神经结构的端到端LID系统如Figure 8所示。

(2) LSTM

同样，在[15]中，研究者们也利用LSTM进行语种分类，LSTM网络结构如Figure 9所示，其解决了梯度爆炸问题。

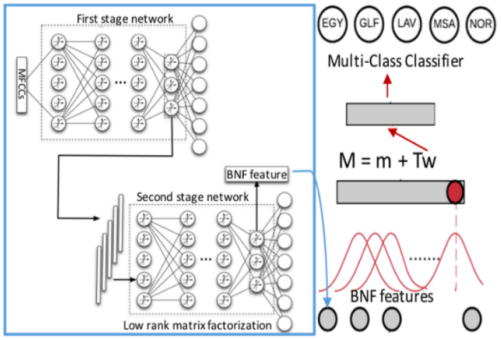

Figure 3: *I-vector based DID system*

Figure 7 DID中提取i-vector模型，图来自[12]

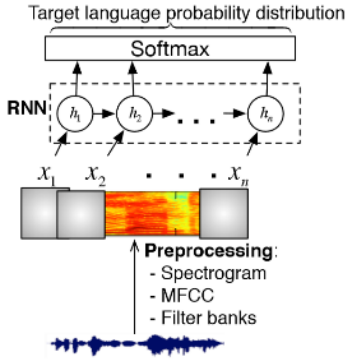

Figure 1: *Design for an end-to-end LID system using RNN.*

Figure 8 RNN结构的端到端LID模型，图来自[14]

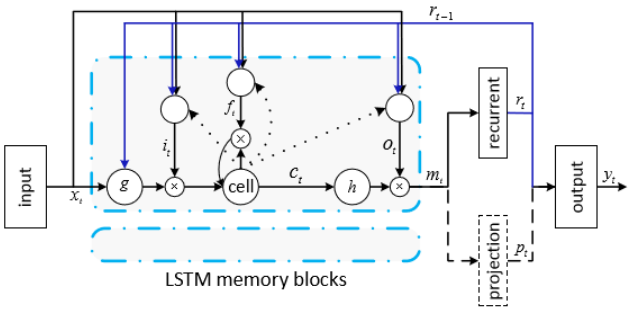

Figure 9 LSTM模型，图来自[17]

随着近些年关注机制被大量应用于自然语言处理中，以及由于其基于大脑注意力模型原理被提出，会着重于重要的信息。相关关注机制的基于关注机制的端到

端语种识别[16]也随及被提出，且实验结果比RNN明显有了提升。下面将会着重讲解。

(3) GRU

GRU的性能可与LSTM相媲美，但其设计显着减少了要估算的参数数量[14]。在整体性能方面，LSTM的表现优于GRU，准确率上表现的并不明显，而且GRU的计算效率更高。因此，我们使用GRU来构建更深层的架构，因为这个优点。

5.3 TDNN模型

近期TDNN也被应用于语种辨识，取得的结果要优于LSTM。该网络有结构多层，每层对特征有较强的抽象能力；有能力表达语音特征在时间上的关系；具有时间不变性；学习过程中不要求对所学的标记进行精确的时间定位以及通过共享权值，方便学习等优点。

5.4 CNN模型

由Maryam Najafian，Sameer Khurana等人提出CNN进行语种识别[12] ，Yu-Wen Lo, Yih-Liang Shen等人提出了一种嵌入式NN用于语音处理的生成听觉模型[18] 进行说话人识别都验证了CNN在语音分类问题的成功应用。

Yu-Wen Lo, Yih-Liang Shen等人提出CNN模型可分为两个阶段[18] ，第一阶段由一维卷积模拟的耳蜗过滤，第二阶段是由二维卷积模拟皮质过滤，如Figure 10所示。

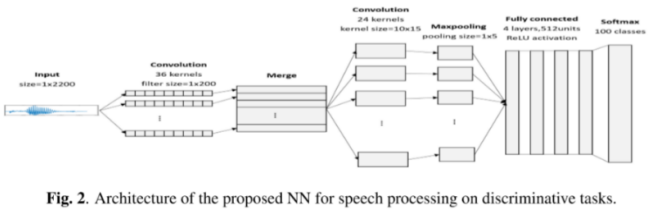

Fig. 2. Architecture of the proposed NN for speech processing on discriminative tasks.

Figure 10 CNN应用于语种识别系统, 图来自[18]

最近，在[19],[20]中，研究者们在神经网络体系结构中采用时间平均pooling层（TAP），利用TAP的优点，神经网络能够训练具有持续时间的输入段。基于此，研究者们又提出了CNN-LDE系统[21]，该系统中将CNN-TAP系统中平均pooling层用LDE层替换, 与简单的TAP不同，它依赖于可学习的字典，LDE与TAP相比具有优越性和互补性。

以及分层思想的模型被Saad Irtza, Vidhyasaharan Sethu等人提出[22]，也给语种识别提供了一个新的发展方向。文献[22]论证了在分层框架中开发出现有的语言模型能够比非分层方法更好地拒绝未知语言。

Masschusetts等人还提出了一种端到端的方言辨识，利用卷积神经网络可直接将原始波形直接映射相应的方言[23]。

5.5 Attention模型

在注意力机制的推动下，Geng, Wang等人在[16]中第一次提出了基于注意力机智的递归神经网络结构，用以实现端到端语种识别的话语水平分类。这种语种识别模块受机器翻译的启发，与其他基于注意力机制的模型类似，将LSTM RNN用在编码输入序列的长跨度连接。但是基于注意力机制的序列到标签结构中的模型训练与推理与之前应用于序列到序列结构的模型不一致。[16]提出通过查找表操作所提供的注意力机制向量参与编码的方法得到高级特征，然后从中选择输入序列中的关键特征代表帧级输入。根据是否对所有源框架或仅在少数源框架上“注意”，[16]中还开发了两种注意方法：软注意和硬注意方法。流程图如Figure 10 所示。

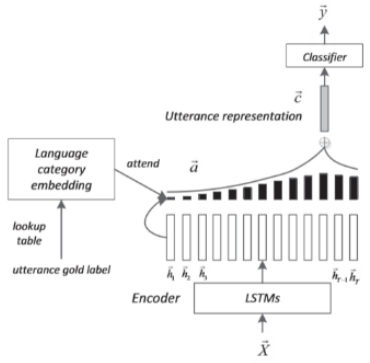

Figure 1: *The Architecture of attention-based recurrent neural network.*

Figure 11 Attention-based RNN LID系统框架, 图来自[25]

5.6 PTN模型

由Zhiyuan Tang等人提出的PTN模型[17]是一种由音素判别DNN产生的phonetic特征作为输入，而不是原始声学特征的LSTM-RNN LID系统。这个新模型类似于传统的phonetic LID方法，但这里的phonetic知识更丰富，它通过帧级判别训练可以学习短时语音信息且涉及由所有作用的音素信息。与传统的Based-token方法有所不同，首先，PTN方法中的语音信息是帧级的，而在传统的Based-token的方法中，该信息是单元级的。因此，PTN方法可以在更短的时间分辨率下表示语音属性。其次，传统的based-token的方法将语音信息表示为源自音素识别的序列，而PTN方法将语音信息表示为涉及由所有音素的信息的特征向量，因此表示更详细的语音信息。最后，传统的基于token方法的后端模型是基于离散token的n-gram LM并且使用最大似然（ML）标准训练，而PTN方法的后端模型是RNN。总之，PTN利用DNN phonetic特征和强大的LSTM模型来获取区分语

种信息属性。PTN方法令许多LID研究人员重新认识到语音时间信息在语种识别中非常有价值。流程图如Figure 11所示。

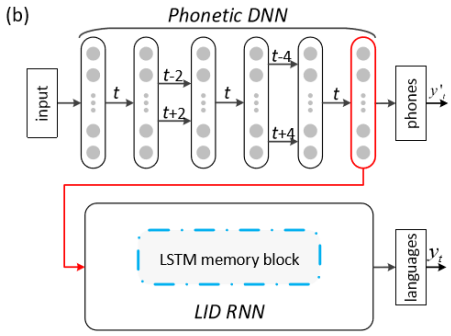

Figure 12 PTN-LID系统框架, 图来自[17]

6 实验与结果

6.1 运行lre-baseline基线遇到的问题

(1) 数据准备(AP17-OLR)

在本语种辨识的基线系统中需要准备训练集与测试集并放置于data/train,test中，其中需要手动创建的文件包括：wav.scp, utt2lang, spk2utt 和utt2spk。在数据准备阶段遇到的问题是不知道utt2lang文件的内容，从一开始以为此文件与语音识别中lang文件相关到最终的正确认知及是话语与语种相对应文件走了很多弯路，以下写出utt2lang的文件形式：

utt1 language1

utt2 language2

ect

(2) LSTM基线

在运行基线中lstm网络部分时遇到srand=None造成参数类型不正确(srand应为整数)，出现此问题的原因是脚本steps/nnet3/train_raw_rnn.py 的第379行所示代码中common_train_lib.prepare_initial_network(args.dir, run_opts, args.input_model)缺少参数srand。解决方法为将其改为common_train_lib.prepare_initial_network(args.dir, run_opts, args.srand, args.input_model)及添加srand参数的传递。

6.2 脚本介绍

6.2.1 run_ivector.sh

写在前面的话：Computer_Cavg.sh比17年要简要很多很多，减少了自己配置语种的关键词等信息。

(1) local/prepare_trials.py：准备trials文件，此文件用于最后的余弦打分。

(2) steps/make_mfcc.sh：提取mfcc特征

lid/compute_vad_decision.sh：进行端点检测

(3) lid/train_diag_ubm.sh

lid/train_full_ubm.sh：训练UBM

(4) lid/train_ivector_extractor.sh：训练i-vector提取器

(5) lid/extract_ivectors.sh：提取i-vector

(6) 改进：

对i-vector进行lda

对i-vector进行plda

(7) 余弦距离进行打分(eer与cavg)

对i-vector、lda-ivector与plda-ivector三种方法分别进行余弦距离打分。

6.2.2 run_nnet.sh

(1) local/prepare_trials.py：准备trials文件，此文件用于最后的余弦打分。

(2) steps/make_fbank.sh：提取fbank特征

(3) local/lang_ali.py：将每个话语与语种的关键词(整形)之间进行一一对应。

(4) 神经网络辨识部分

tdnn

lstm

(5) 余弦距离进行打分(eer与cavg)

对tdnn、lstm两种方法分别进行余弦距离打分。

6.3 基线运行结果

6.3.1 参数配置

数据集：AP17-OLR

num_gauss=1024

ivector_dim=400

lda_dim=9

covar_factor=0.1

6.3.2 结果

Table 1 test on original length

test set: ap17-olr test_all		
LID system	Cavg	EER%
i-vector	0.0617	5.99
i-vector + lda	0.0446	4.50
i-vector + plda	0.0554	5.49
tdnn	0.1401	15.17
lstm	0.1463	16.16

Table 2 test on short duration (1s)

test set: ap17-olr test_1s		
LID system	Cavg	EER%
i-vector	0.1627	19.23
i-vector + lda	0.1536	19.05
i-vector + plda	0.1446	18.60
tdnn	0.1301	15.14
lstm	0.1329	16.06

6.4 提升

6.4.1 SVM

在近年来神经网络火热之前，SVM一直是明星算法。它是一种典型的具有区分性的建模方法。由以下测试结果可知，用i-vector特征结合核函数为rbf的SVM效果最佳。

Table 3 SVM:test on original length

test set: ap17-olr test_all		
LID system	Cavg	EER%
i-vector + linerSVM	0.0502	4.37
i-vector + polySVM	0.0398	4.14
i-vector + rbfSVM	0.0353	3.41
i-vector + lda + linerSVM	0.0403	3.94
i-vector + lda + polySVM	0.0436	4.04
i-vector + lda + rbfSVM	0.0416	3.90

6.4.2 PTN(phonetic temporal neural)

文献[17]中提出一种用于语种辨识的PTN方法，识别率相对于lstm有较大的提升，并证明了语种辨识中语音的长时信息比原始声学特征信息更有利。论文还指出PTN方法显著优于现有的声学神经模型，在短话语和嘈杂条件下甚至优于传统的i-vector方法。以下表格记录了实验结果，所示结果符合论文结论。

Table 4 PTN:test on AP17-OLR

LID system : PTN		
test set	Cavg	EER%
test_all	0.0692	8.01
test_1s	0.0775	9.14

依据论文中介绍可知模型域是基于LSTM网络上的，所以在本实验中PTN模型框架是在local/nnet3/run_lstm.sh上进行修改的。

6.5 改进

6.5.1 CNN

大量研究者们已经证明卷积神经网络（CNN）模型对于许多语音和语言处理应用是有效的[12]。于是将先前参加科大讯飞方言辨识完成的基于CNN的方言辨识

系统(本地测试acc为83%)应用于本次语种识别中，该模型输入为MFCC,FBANK或者语谱图，网络结构为4层1-d CNN、一层全局pooling 层以及2 层全连接层最后用softmax进行判别。但实验结构并不如意，可能原因，参数没调节好，也可能网络结构不适合语种辨识系统。

Table 5 CNN1:test on original length

test set: ap17-olr test_all			
LID system	Cavg	EER%	ACC%
CNN	—	—	53%

6.5.2 Phonetic-based CNN

语音phonetic特征表示比声学特征更高水平的信息[17]，因此在噪声和信道方面更加不变。基于此特征以及CNN的局部性提出Phonetic-based CNN语种辨识系统。

Table 6 CNN2:test on original length

test set: ap17-olr test_all			
LID system	Cavg	EER%	ACC%
Phonetic-based CNN	—	—	TODO

总结

通过这两周对语种识别方向进行大量文献的阅读，对语种识别有了新的认识，同时也对科研有了新的认识。前段时间参加科大方言辨识初赛也发现由于前期读文献太少而引起了除了改参数改模型没有其它想法的局限。在接下来的时间还会继续跟进最新论文看，读，想。

写于20180807

References

1. Yeshwant K Muthusamy, Etienne Barnard, and Ronald A Cole, "Automatic language identification: A review/tutorial," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, 1994.
2. Yeshwant Kumar Muthusamy, "A segmental approach to automatic language identification," *IEEE Signal Processing Magazine*, 1993.
3. Marc A Zissman and Elliot Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *icassp*. IEEE, 1994, pp. 305–308.
4. Marc A Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, pp. 31, 1996.
5. Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
6. Pedro A Torres-Carrasquillo, Elliot Singer, Mary A Kohler, Richard J Greene, Douglas A Reynolds, and John R Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Seventh International Conference on Spoken Language Processing*, 2002.
7. Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
8. William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.
9. Colin Raffel and Daniel PW Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv preprint arXiv:1512.08756*, 2015.
10. Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
11. KV Mounika, Sivanand Achanta, HR Lakshmi, Suryakanth V Gangashetty, and Anil Kumar Vuppala, "An investigation of deep neural network architectures for language recognition in indian languages.," in *INTERSPEECH*, 2016, pp. 2930–2933.
12. Maryam Najafian, Sameer Khurana, Suwon Shan, Ahmed Ali, and James Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5174–5178.
13. Alicia Lozano-Diez, Oldrich Plchot, Pavel Matejka, and Joaquin Gonzalez-Rodriguez, "Dnn based embeddings for language recognition," in *Proceedings of ICASSP*, 2018.
14. Trung Ngo Trong, Ville Hautamäki, and Kong Aik Lee, "Deep language: a comprehensive deep learning approach to end-to-end language recognition," in *Odyssey: the Speaker and Language Recognition Workshop*, 2016.
15. Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor, "Classifying asr transcriptions according to arabic dialect," in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 2016, pp. 126–134.
16. Wang Geng, Wenfu Wang, Yuanyuan Zhao, Xinyuan Cai, Bo Xu, et al., "End-to-end language identification using attention-based recurrent neural networks," in *INTERSPEECH*, 2016, pp. 2944–2948.
17. Zhiyuan Tang, Dong Wang, Yixiang Chen, Lantian Li, and Andrew Abel, "Phonetic temporal neural model for language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 134–144, 2018.
18. Yu-Wen Lo, Yih-Liang Shen, Yuan-Fu Liao, and Tai-Shih Chi, "A generative auditory model embedded neural network for speech processing," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5179–5183.

19. Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.
20. David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.
21. Weicheng Cai, Zexin Cai, Xiang Zhang, Xiaoqi Wang, and Ming Li, "A novel learnable dictionary encoding layer for end-to-end language identification," *arXiv preprint arXiv:1804.00385*, 2018.
22. Saad Irtza, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Haizhou Li, "End-to-end hierarchical language identification system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5199–5203.
23. Suwon Shon, Ahmed Ali, and James Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv preprint arXiv:1803.04567*, 2018.
24. Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno, "Automatic language identification using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5337–5341.