

说话人确认中基于无监督聚类的得分规整

古斌 郭武

(中国科学技术大学语音及语言信息处理国家工程实验室, 合肥, 230026)

摘要: 在说话人确认任务中, 得分规整可有效调整测试得分分布, 使每个说话人的得分分布接近同一分布, 从而提升系统整体性能。直接从开发集中获得针对待识别目标说话人的大量冒认者得分, 利用无监督聚类手段对这些得分进行筛选, 并采用混合高斯模型来拟合得分分布, 挑选均值最大的高斯单元作为得分规整的参数并将其应用于说话人的得分规整。在 NIST SRE 2016 测试集上的测试结果表明, 相对于其他得分规整算法, 采用无监督聚类得分规整的方法可有效提升系统性能。

关键词: 说话人确认; 得分规整; 无监督聚类

中图分类号: TN912 **文献标志码:** A

Unsupervised Clustering Score Normalization in Speaker Verification

Gu Bin, Guo Wu

(University of Science and Technology of China, National Engineering Laboratory for Speech and Language Information Processing, Hefei, 230026, China)

Abstract: In the speaker verification (SV) task, score normalization can improve the system performance by adjusting the score distribution of each speaker to a similar distribution. Here, a large number of imposter scores for the target speakers are obtained from the development set firstly, then these scores are clustered by unsupervised clustering algorithm and the Gaussian mixture models (GMM) are used to fit the score distribution. The mean and standard deviation of Gaussian component with maximum mean value are used in the SV score normalization method. Experiments are conducted on the NIST SRE 2016 test set and results show that compared with the conventional score normalization methods, the proposed method can effectively improve the system performance.

Key words: speaker verification; score normalization; unsupervised clustering

引言

说话人确认(Speaker verification, SV)是判断一段测试语音与其所声明身份是否一致的过程, 是一个典型的二分类模式识别任务。在说话人确认中, 需要计算每段测试语料与其声明说话人模型之间的得分, 然后将得分与设定的判决门限比较并作出判决, 该门限对所有测试语音都是同一个数值^[1]。受语料之间信道、语种、时长、性别等因素的影响, 不同的说话人模型、不同冒认者的得分分布往往不同, 这种固定数值的门限会导致整个系统的性能受到严重的影响。

为了使得不同说话人模型、不同冒认者的得分分布趋于一致,从而使得固定的门限能够比较好地区分目标说话人和冒认者,一般采用得分规整的方法来实现这一目的。通常有2种规整的途径,一种是规整目标说话人模型的得分分布,例如零规整(Z-norm),另一种是规整冒认者得分分布,例如测试规整(T-norm),在这二者的基础上,又衍生出了一系列得分规整方法,包括ZT-norm^[2]、S-norm^[3]、Top-norm^[4]等。目前得分规整已在说话人确认系统中的得到广泛运用,并在高斯混合模型-通用背景模型(Gaussian mixture models, universal background model, GMM-UBM)^[5]、i-vector^[6]、x-vector^[7]等系统中取得显著效果。

近几年的研究中,如何挑选合适的测试语料与其声明身份组成测试从而得到接近真实测试条件的得分分布参数是一关键问题^[8]。由于测试集的得分分布实际上是无法事先获知的,为了尽量缩小估计分布与真实分布之间的偏移,研究者们提出了自适应得分规整方法。一些研究者从说话人特征层面来挑选规整数据^[9-10],挑选与测试数据尽可能相似的数据,还有一些研究者选择从得分层面来挑选规整数据,对规整集合得分只取值较大的若干个得分用于统计分布参数^[11-12]。近几年的应用中,从得分层面挑选规整数据的自适应得分规整方法应用更加广泛^[13-15]。

得分规整总体逻辑思路就是把所有冒认者得分规整到同一分布上,消除说话内容、语速等因素的影响^[16],从而通过一个合理的门限来获得最好的识别性能。NIST SRE 2016测试集出现了以往数据集中未曾出现过的语种,开发集中只有部分无标签数据与测试集标签信息相匹配,因此无法针对每个说话人或测试语句挑选合适的规整数据集,根据以往的研究,此时需要使用自适应得分规整的方法。在选择得分最高的一些冒认者得分来估计规整参数时,对每个待规整对象(说话人模型或测试语料)都取相同数量的较大得分,这些数值较大的得分不服从高斯分布,因此必然与测试集的分布存在偏差。

本文提出了一种新的自适应得分规整方法,利用无监督聚类对每个待规整对象的得分集合进行聚类,然后对数据进行筛选,并且用混合高斯函数来拟合经过筛选后的冒认者得分分布,取均值最大的高斯函数的均值和标准差作为规整参数,可以缩小与测试集的真实得分分布之间的偏差,减小通过固定数量的得分统计分布参数时丢失的分布信息,从而提高整体识别性能。本文在NIST SRE 2016数据集的测试集上进行实验验证,从结果上来看,明显优于传统的方法。

1 得分规整

1.1 基础得分规整方法

Z-norm和T-norm是使用最为广泛的2种得分规整方法,Z-norm利用实际说话人模型 $\{e_1, e_2, \dots, e_M\}$ 对大量冒认者语料 $\{t_1^*, t_2^*, \dots, t_L^*\}$ 进行测试,得到测试得分 $s(e_m, t_l^*)$,其中 e_m 表示注册集中第 m 个说话人模型, t_l^* 表示规整数据集中第 l 条冒认者语料,统计每个说话人模型得分均值 $\mu(e_m)$ 和标准差 $\sigma(e_m)$ 。

然后利用这2个分布参数对含有说话人模型 e_m 的实际测试得分 $S(e_m, *)$ 进行规整

$$S(e_m, *)_{Z\text{-norm}} = \frac{S(e_m, *) - \mu(e_m)}{\sigma(e_m)} \quad (1)$$

Z-norm可以较好地消除说话人模型之间的分布差异性,属于离线计算,在模型训练阶段就可以获得分布参数,在测试阶段不会额外占用计算时间。

T-norm与Z-norm相似,不同的是它利用每段测试语料与不同冒认说话人模型进行测试得到得分,

从而统计每段测试语句的得分分布参数并对测试得分进行规整。T-norm可以较好地消除测试语音环境对输出评分分布的影响,但是得分分布参数需在实际测试时获取,属于在线计算,会降低模型识别速度。基于Z-norm、T-norm的规整方法,又衍生出了S-norm、ZT-norm、TZ-norm等一系列方法,其中S-norm是将Z-norm和T-norm规整后的得分进行加权求和,该方法凭借其优异的性能在近些年获得了广泛应用。

1.2 自适应得分规整方法

自适应得分规整方法在基础得分规整方法之上对规整数据进行了选择。在这类方法中,只有部分被选中的规整数据用于统计得分的均值和标准差,被选中的这部分数据可能随着每一个规整对象而变化。以Top-norm为例。在Top-norm中,规整数据集中的每条语句都参与得分计算,但是只有得分值最大的 N 个得分将被用于统计规整时所使用的均值和方差,图1以某个说话人模型在规整数据集上的得分分布为例,展示了自适应规整数据选择的过程,其中横轴表示得分的数值大小,纵轴表示得分在单位区间内出现的次数。

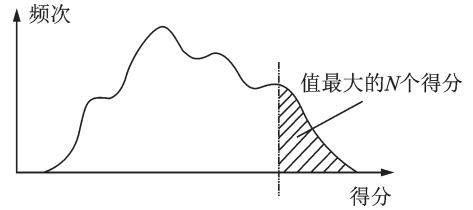


图1 自适应规整数据选择

Fig.1 Adaptive normalization data selection

该种自适应得分挑选的方法如今已广泛应用于Z-norm、T-norm、S-norm等基础得分规整方法中,本文所提出的自适应得分规整方法将与该方法进行对比。在本文中使用该方法时会在相应的基础规整方法前用“Top”标注。

2 基于无监督聚类的得分规整

本节以Z-norm为例,说明基于聚类的得分规整。

2.1 冒认者得分筛选

对于注册集中第 m 个说话人模型 e_m ,与包含 L 条语料的规整数据集测试后可得到 L 个测试得分 $\{s(e_m, t_l^*), l \in [1, L]\}$,利用K-means算法对 L 个得分进行聚类,然后仅保留均值靠前的若干类得分作为筛选后得分。具体算法如下:

步骤1 随机选择 K 个初始中心点 $\{\mu_1, \mu_2, \dots, \mu_K\}$ 。

步骤2 计算每个测试得分到 K 个中心点欧式距离,根据最小欧式距离划分每个测试得分所属类别 $C(\mu_k)$,当 $s(e_m, t_i^*) \in C(\mu_k)$ 则对于 $\forall k' \in [1, K]$ 有

$$\left[s(e_m, t_i^*) - \mu_k \right]^2 \leq \left[s(e_m, t_i^*) - \mu_{k'} \right]^2 \quad (2)$$

步骤3 根据每个类别的得分更新每个类别中心点,其中 $|C(\mu_k)|$ 表示 $C(\mu_k)$ 类中的得分个数

$$\mu_k = \frac{1}{|C(\mu_k)|} \sum_{s(e_m, t_i^*) \in C(\mu_k)} s(e_m, t_i^*) \quad (3)$$

步骤4 当每个类别更新后的中心点相较未更新中心点的距离小于设定阈值或迭代次数大于设定值,此时认为聚类数据已收敛,否则跳转至步骤2。

聚类完成后,可以认为K-means聚类获得的 K 个类别的得分代表着不同匹配程度的测试得分,类中心数值越大的类得分所对应的冒认者测试语料与声明说话人在性别、语种等方面信息匹配程度

越高。

由于得分值较大的冒认者在实际测试时最难以分辨,需要更多地关注这类较大值得分,因此舍弃中心值较小的得分,保留前 K' 个聚类中的得分,如图2所示。在本文Z-norm中 K 取6, K' 取3;T-norm中 K 取3, K' 取2。

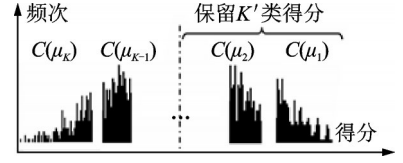


图2 得分筛选示意图

Fig.2 Score screening schematic

2.2 GMM拟合得分分布

不同于以往假设每个待规整对象对于冒认者得分分布服从单高斯分布,本文假设其得分分布满足更为复杂的混合高斯分布,在2.1节剔除部分冒认者得分的基础上,引入GMM模型拟合保留得分的分布。

鉴于2.1节已经获得了 K' 个得分,可以按这 K' 个类的均值和标准差初始化 K' 个高斯的GMM模型均值和标准差,权重初始值 w_i 按照式(4)初始化

$$w_i = \frac{|C(\mu_i)|}{|C|} \quad (4)$$

式中: $|C|$ 为保留的总得分个数, $|C(\mu_i)|$ 表示第 i 类中的得分个数。利用EM算法对保留的得分进行进一步聚类直至满足收敛条件,GMM模型的参数按式(5)–(8)进行更新:

步骤1 计算每个得分在每个高斯的占有率

$$p(i|s(e_m, t_l^*)) = \frac{w_i N(s(e_m, t_l^*); \mu_i, \sigma_i^2)}{\sum_j w_j N(s(e_m, t_l^*); \mu_j, \sigma_j^2)} \quad (5)$$

步骤2 根据每个高斯的占有率更新GMM参数

$$w'_i = \frac{1}{|C|} \sum_{l=1}^{|C|} p(i|s(e_m, t_l^*)) \quad (6)$$

$$\mu'_i = \frac{\sum_{l=1}^{|C|} s(e_m, t_l^*) p(i|s(e_m, t_l^*))}{\sum_{l=1}^{|C|} p(i|s(e_m, t_l^*))} \quad (7)$$

$$\sigma_i'^2 = \frac{\sum_{l=1}^{|C|} [s(e_m, t_l^*) - \mu'_i]^2 p(i|s(e_m, t_l^*))}{\sum_{l=1}^{|C|} p(i|s(e_m, t_l^*))} \quad (8)$$

步骤3 当模型参数收敛或达到最大迭代次数时终止参数更新,否则跳转至步骤1。

当GMM模型收敛后,取均值最大的高斯单元均值 $\mu(e_m^{\text{top-gauss}})$ 和标准差 $\sigma(e_m^{\text{top-gauss}})$ 作为当前说话人模型的得分规整参数,如图3所示,其中纵轴表示得分在单位区间内出现的概率。

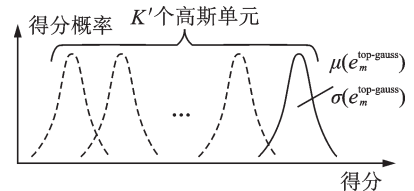


图3 GMM得分规整示意图

Fig.3 GMM score normalization Schematic

3 实验配置与结果

3.1 实验数据与评价指标

训练数据集包括NIST SRE 2004、2005、2006、2008、2010, SwitchBoard以及Mixer6数据集中的部

分数据,共计 87 457 条数据,数据时长主要分布在 30 s~3 min,大部分内容为网络语音、室内录音以及电话中的英文对话。这些数据主要用于 UBM、i-vector 因子分析的载荷矩阵、PLDA 的载荷矩阵训练。

开发集中包含一份无标签数据,为广东话与菲律宾塔加路语两种语言的混合数据,开发集中只有该部分数据语种与测试集匹配,以往研究证明若测试集与规整数据集语种不匹配得分规整将很难获得效果,因此这部分数据将用于得分规整,共计 2 272 条语音。从这些无标签数据中随机挑选 3/4 数据作为 Z-norm 集,剩余作为 T-norm 集,将说话人模型与 Z-norm 集、测试语料与 T-norm 集组合构造测试用于得分规整。

本文使用 NIST SRE 2016 测试集作为系统性能评估数据集,测试集为广东话与菲律宾塔加路语 2 种语言的混合数据集。测试集中说话人注册语音为时长 60 s 的语料,说话人模型由 1 段或 3 段语料注册得到;测试语料时长均匀分布在 10~60 s,目标说话人测试 37 058 条,冒认者测试 19 494 662 条。

本文采用 NIST SRE 2016 官方计划中的等错误率(Equal error rate, EER),最小错误代价函数(Minimal detection cost function, Min_DCF)和实际的错误代价函数(Actual detection cost function, Act_DCF)作为评价指标^[17]。

为了体现估计的得分分布参数与真实分布参数的差异,另外给出均值和标准差 2 个偏差参数的定义

$$\mu_{\text{bias}} = \frac{1}{N} \sum_{n=1}^N |\hat{\mu}_n - \mu_n| \quad (9)$$

$$\sigma_{\text{bias}} = \frac{1}{N} \sum_{n=1}^N |\hat{\sigma}_n - \sigma_n| \quad (10)$$

式中: N 为测试集中待规整对象的数量,对于 Z-norm, N 为注册说话人数量,对于 T-norm, N 为测试语料数量, $\hat{\mu}_n$ 、 $\hat{\sigma}_n$ 为根据规整集得分估计的均值和标准差, μ_n 、 σ_n 为根据真实测试得分统计的结果。通过这 2 个参数来比较不同算法对于真实得分分布的拟合能力,越小则代表拟合能力越强。

3.2 系统描述

本文采用基于 i-vector 和概率线性判别分析(Probabilistic linear discriminant analysis, PLDA)的说话人确认系统。在得到 PLDA 测试得分之后,采用各种得分规整方法进行得分规整。整个流程主要使用 Kaldi toolkit^[18]开源代码实现。

在特征提取阶段,首先提取加上一阶二阶差分后共计 60 维的梅尔频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)特征,使用 3 s 的平滑窗作均值方差规整,然后通过基于能量的语音活动检测(Voice activity detection, VAD)算法去除静音帧。

在模型训练阶段使用 NIST SRE 2016 开发集中的无标签数据训练具有全对角协方差矩阵的 UBM 模型,该模型有 2 048 个高斯成分,利用训练数据集与 UBM 模型训练 i-vector 生成矩阵,提取 600 维的 i-vector 向量,然后训练集测试集各自进行均值中心化,通过线性判别分析(Linear discriminant analysis, LDA)将 i-vector 降维至 400 维,利用无标签数据作主成分分析(Principal component analysis, PCA)白化,所有数据进行长度规整后作概率形式的线性判别分析(Probabilistic linear discriminant analysis, PLDA)得到测试得分,训练和测试过程均为性别、文本无关。

本文根据不同的得分规整方法,构建了不同的得分规整系统进行对比:

系统 0:得分不做规整,这是基线(Baseline)系统。

系统 1~3:使用所有得分进行 Z-norm、T-norm、S-norm;这些算法是目前广泛采用的算法,也是对比系统。

系统 4~6:使用数据自适应选择的方法进行 Z-norm、T-norm、S-norm。这是参考文献[12]所采用的

算法。本文中 Z-norm 中 N 取 150, T-norm 中 N 取 100, 在这两个参数上本文系统性能最佳, 后面的描述中按此种做法所做的规整将在规整方法前加 Top, 例如 Top Z-norm, 这些属于对比系统。

系统 7~9: 利用基于 GMM 的 Z-norm、T-norm、S-norm。后面的描述中按此种做法所做的规整将在规整方法前加 GMM, 例如 GMM Z-norm。

3.3 实验结果与分析

从表 1 中各项指标可看出, 系统 9 GMM S-norm 整体性能最佳, EER 为 13.69, 而 Min_DCF 和 Act_DCF 为 0.716 7 和 0.721 4, 这 2 项指标较 Baseline 分别有 7.1% 和 22.0% 的相对改进, 较自适应得分规整方法也有 3.3% 和 6.3% 的相对优化。

从基本的规整方法来看, S-norm 效果要好于 Z-norm、T-norm, 说明 S-norm 充分结合了二者的优点, 实现了性能互补。对比系统 1~3、4~6 以及系统 7~9 可知, 从自适应选取规整得分策略来看, 选取靠前的得分统计规整参数性能更优, 这一定程度上是因为靠前的得分集合相对目标说话人真实得分有更加稳定的均值和标准差, 而靠后的得分相对目标说话人得分波动性比较大, 将使得每个说话人的理想最佳门限和统一的判决门限相差较远。在此基础上, 使用基于聚类的变长数量得分的策略要优于选定固定数量得分, 这得益于 GMM 模型在刻画数据分布时的优势, 从而使用该模型进行聚类能够得到更加稳定的均值和方差, 这也表明利用 GMM 模型能有效地解决在选取固定数量的得分计算规整参数时分布信息丢失的问题。

为了更近一步说明 GMM 模型在获得规整参数时的优势, 表 2 给出式(9, 10)定义的均值和标准差偏差参数。

从表 2 可以看出, 无论是 Z-norm 还是 T-norm, 基于聚类的得分规整方法的均值偏差都要明显小于选取固定数量个得分的规整方法, 二者的方差偏差基本属于同一水平。本文基于 GMM 的规整方法首先剔除了靠后的得分, 可以缩小与真实分布均值上的差距, 然后利用 GMM 软聚类的方法使得保留的每个得分点都参与规整参数的计算, 这样可以缩小标准差与真实分布的偏差, 因此基于聚类的得分规整可更加有效地拟合真实的得分分布。

4 结束语

本文采用聚类手段, 替代了以往自适应规整算法中直接取得分最高的特定数量个得分统计分布参数这一做法, 选取变长个得分参与分布参数计算, 获得了很好的性能提升。进一步的工作可以将从特征端规整数据自适应挑选和本文从得分端的规整数据自适应挑选结合起来以获得进一步的改进。

表 1 不同得分规整方法实验结果

Tab. 1 Results of different scoring normalization methods

序号	规整方法	EER	Min_DCF	Act_DCF
0	Baseline	13.94	0.771 6	0.925 0
1	Z-norm	14.61	0.787 1	0.833 9
2	T-norm	14.27	0.769 4	0.813 3
3	S-norm	14.17	0.768 5	0.811 8
4	Top Z-norm	14.23	0.752 6	0.799 3
5	Top T-norm	13.93	0.744 5	0.773 2
6	Top S-norm	13.72	0.741 3	0.770 1
7	GMM Z-norm	14.04	0.741 1	0.744 8
8	GMM T-norm	13.87	0.729 2	0.738 7
9	GMM S-norm	13.69	0.716 7	0.721 4

表 2 不同得分规整方法偏差参数

Tab. 2 Bias of different score normalization methods

规整方法	μ_{bias}	σ_{bias}
Z-norm	93.6	21.1
Top Z-norm	36.7	34.2
GMM Z-norm	15.3	13.2
T-norm	72.6	13.8
Top T-norm	22.9	27.7
GMM T-norm	17.4	17.1

参考文献:

- [1] Auckenthaler R, Carey M, Lloyd-Thomas H. Score normalization for text-independent speaker verification systems[J]. Digital Signal Processing, 2000, 10(1/2/3): 42-54.
- [2] Kenny P. Bayesian speaker verification with heavy tailed priors[C]// Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic: [s.n.], 2010: 1-41.
- [3] Aronowitz H, Irony D, Burshtein D, et al. Modeling intra-speaker variability for speaker recognition[C]// European Conference on INTERSPEECH. Lisbon, Portugal: IEEE, 2005: 2177-2180.
- [4] Zigel Y, Wasserblat M. How to deal with multiple-targets in speaker identification systems?[C]// 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop. [S.l.]: IEEE, 2006: 1-7.
- [5] Sturim D E, Reynolds D A, Dunn R B, et al. Speaker verification using text-constrained Gaussian mixture models[C]//2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE, 2002, 1: 1-677-1-680.
- [6] Kenny P, Ouellet P, Dehak N, et al. A study of interspeaker variability in speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(5): 980-988.
- [7] Snyder D, Ghahremani P, Povey D, et al. Deep neural network-based speaker embeddings for end-to-end speaker verification [C]// 2016 IEEE Spoken Language Technology Workshop (SLT). [S.l.]: IEEE, 2016: 165-170.
- [8] Khemiri H, Petrovska-Delacretaz D. Cohort selection for text-dependent speaker verification score normalization[C]// 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). [S.l.]: IEEE, 2016: 689-692.
- [9] Skorkovská L, Zajíč Z, Müller L. Comparison of score normalization methods applied to multi-label classification[C]// 2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). [S. l.]: IEEE, 2014: 000433-000437.
- [10] Swart A, Brummer N. A generative model for score normalization in speaker recognition[J]. arXiv preprint arXiv:1709.09868, 2017.
- [11] Sturim D E, Reynolds D A. Speaker adaptive cohort selection for Tnorm in text-independent speaker verification[C]// Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE, 2005, 1: 1/741-1/744 Vol. 1.
- [12] Matejka P, Novotný O, Plchot O, et al. Analysis of score normalization in multilingual speaker recognition[C]// INTERSPEECH. Stockholm, Sweden: IEEE, 2017: 1567-1571.
- [13] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 5329-5333.
- [14] Li L, Mak M W. Unsupervised domain adaptation for gender-aware PLDA mixture models[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 5269-5273.
- [15] Nidadavolu P, Lai C I, Villalba J, et al. Investigation on bandwidth extension for speaker recognition[C]// INTERSPEECH. Hyderabad, India: IEEE, 2018: 1111-1115.
- [16] Rouzi A, Dong W, Lantian L I, et al. Score domain speaking rate normalization for speaker recognition[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(4): 337-341.
- [17] Sadjadi S O, Kheyrkhah T, Tong A, et al. The 2016 NIST speaker recognition evaluation[C]// INTERSPEECH. Stockholm, Sweden: IEEE, 2017: 1353-1357.
- [18] Povey D, Ghoshal A, Boulianne G, et al. The kaldi speech recognition toolkit[C]// IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. [S.l.]: IEEE Signal Processing Society, 2011.

作者简介:



古斌(1996-),男,硕士研究生,研究方向:声纹识别,
E-mail: 545088212@qq.com。



郭武(1973-),男,博士,副教授,研究方向:语音信号处理。