# 基于环境特征判别学习的顽健语音识别方法

韩纪庆,高 文
(哈尔滨工业大学计算机科学与工程系,哈尔滨 150001)

**摘 要**: 提出一种基于环境特征判别学习的顽健语音识别方法,它首先通过使用一个简单的分类器和梯度下降法迭代地学得环境特征,接着利用得到的环境特征从观测到的混噪语音特征中估计出纯净的语音特征,然后将估计出来的纯净语音特征用到后端的 HMM 分类器中. 使用所提出的方法对不特定话者小词表进行实验,其系统误识率与基本 HMM 系统相比下降了 33.3 %.

**关键词**: 顽健语音识别; 环境特征; 判别学习
**中图分类号**: TP391.42 **文献标识码**: A **文章编号**: 0372-2112 (2001) 02-0196-03

# Robust Speech Recognition Based on Discriminative Learning of Environmental Features

HAN Ji-qing, GAO Wen
*(Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)*

Abstract: A discriminative learning method of environmental features is proposed for robust speech recognition, in which a simple classifier and the gradient descent algorithm are firstly used to iteratively learn the environmental features, and then the pure speech features are estimated from the observed speech features. Finally the estimations of the pure speech features are used in the back-end HMM classifier. Using the proposed method, a 33.3 % reduction in word error rate is obtained relative to conventional HMM system for a speaker-independent, small vocabulary recognition task.

Key words: robust speech recognition; environmental features; discriminative learning

## 1 引言

语音识别系统往往受到加性噪声和通道畸变等环境特征的影响,为补偿这些影响研究者提出了很多方法[1~3],它们大都是基于最大似然估计准则的. 近年来,人们提出了判别学习方法[4],它使用 GPD(Generalized Probabilistic Descent)算法[5]按照最小分类错误 MCE(Minimum Classification Error)准则迭代地调整分类器参数,因而达到直接最小化误分类数的目的. 基于 MCE 的判别学习方法,首先被用到语音识别器中的参数优化问题上,诸如基于 DTW(Dynamic Time Warping)的参数优化[6]、基于 HMM(Hidden Markov Model)的参数优化[7]. 由于这种判别学习方法就是利用梯度下降法优化参数值的过程,如将特征参数嵌入到代价函数中,也可以利用该方法来优化特征参数. 因而它后来被用于语音识别中的特征提取上,称为判别特征提取 DFE (Discriminative Feature Extraction)[8],并被分别用于优化动态倒谱参数[9],以及寻找 Mel 频率对数频谱上的最优线性变换[10]. DFE 在迭代地优化特征参数的同时,也优化了分类器参数. 在所有的情况下都证明 DFE 是一种减少错误率的有效方法. 然而 DEF 需要同步地估计前端特征参数和后端 HMM 分类器参数.

本文提出一种新的类似于 DFE 的方法来学习环境的统计特征,并基于学得的环境特征来估计纯净语音信号的特征. 文中首先将 VTS(Vector Taylor Series)方法[3]扩展应用到倒谱域上. 而后通过使用一个简单的分类器和 GPD 算法迭代地学得环境特征. 接着利用得到的环境特征从观测到的混噪语音特征中估计出纯净的语音特征,然后将估计出来的纯净语音特征用到后端的 HMM 分类器中.

## 2 一个倒谱域上的环境模型

环境模型可以假设为受到噪声污染后的语音信号是由一个纯净语音信号经过一个通道畸变滤波后又受到一个与之不相关的加性噪声的影响[1,2]. 如令 $Y(\omega)$ 为混噪语音的频谱, $N(\omega)$ 为加性噪声的频谱, $X(\omega)$ 为纯净语音的频谱,而 $H(\omega)$ 为通道畸变滤波器的频率响应;则对第 $k$ 个 Mel 频带($\omega_k$ 为其中心频率, $1 \le k \le B$, $B$ 为 Mel 频带的个数),环境特征对语音的影响可表达如下:

$$Y(\omega_k) = H(\omega_k)\, X(\omega_k) + N(\omega_k) \qquad (1)$$

收稿日期:1999-12-15;修回日期:2000-04-25
基金项目:国家 863 计划(No. 863-306ZT03-02-3),黑龙江省留学回国人员基金(No. L99-2)

对上式两边分别取对数,则可将其转化到对数频谱域:

$$\log(Y(\omega_k)) = \log\{H(\omega_k)X(\omega_k) + N(\omega_k)\} \quad (2)$$

如果用 $y[k]$、$x[k]$、$h[k]$和 $n[k]$分别表示混噪语音信号、纯净语音信号、通道畸变滤波器,以及加性噪声的 Mel 频带对数频谱,则可以推导出:

$$x[k] = y[k] - h[k] + \log(1 - \exp(n[k] - y[k])) \quad (3)$$

当在倒谱域上观察环境特征影响时,语音信号、噪声和通道畸变倒谱向量之间存在着一个比较复杂的非线性关系:

$$X = Y - h + C\{\log(I - \exp(C^{-1}(n - Y)))\} \quad (4)$$

其中 X、Y、n 和 h 分别代表纯净语音信号、混噪语音信号、加性噪声和通道畸变的倒谱向量,I 是一个单位向量,C 和 $C^{-1}$ 分别代表从 Mel 频率对数频谱到其倒谱变换时 $B \times B$ 的离散余弦变换矩阵及其逆矩阵.

环境中的噪声 n 是变化的,显然上式没有精确的形式解. Pedro 提出了 VTS 方法[3],将对数频谱域上的环境模型用有限项的台劳级数展开来近似,并且演示了用二阶台劳级数就可以充分近似环境函数,这里将其扩展应用到倒谱域上.

公式(4)可以写成如下的形式:

$$X = Y + f(Y, n, h) \quad (5)$$

这里称 $f(Y, n, h)$为倒谱域上的环境函数,它的形式为:

$$f(Y, n, h) = C\{\log(I - \exp(C^{-1}(n - Y)))\} - h \quad (6)$$

假定通道畸变 h 是由反映环境通道畸变的先验估计和当前发音的通道畸变相互混合而成,这可用如下的泄漏积分器来实现:

$$h = (1 - \alpha)\cdot h_{wh} + \alpha\cdot h_{cu} \quad (7)$$

其中 $h_{wh}$和 $h_{cu}$分别代表环境通道畸变的先验估计和从当前发音中估计出来的通道畸变,而 $\alpha$是一个反映以 $h_{wh}$或 $h_{cu}$作为通道畸变 h 的可信度的参数.

$h_{cu}$可以从当前发音中直接估计出来,这样 $f(Y, n, h)$的形式变为 $f(Y, n, h_{wh})$:

$$f(Y, n, h_{wh}) = C\{\log(I - \exp(C^{-1}(n - Y)))\} - ((1 - \alpha)h_{wh} + \alpha h_{cu}) \quad (8)$$

不妨假定噪声 n 服从高斯分布 $N_n(\mu_n, \sum_n)$,为简化计算,进一步假设 $\sum_n$ 为对角阵. 将 $f(Y, n, h_{wh})$在 $\mu_n$ 点进行二阶台劳级数展开,则有:

$$X \approx Y + f(Y, \mu_n, h_{wh}) + \nabla_n f(Y, \mu_n, h_{wh})(n - \mu_n) + (1/2)\nabla_n^2 f(Y, \mu_n, h_{wh})(n - \mu_n)^2 \quad (9)$$

其中 $\nabla_n f(Y, \mu_n, h_{wh})$和 $\nabla_n^2 f(Y, \mu_n, h_{wh})$分别代表 $f(Y, n, h_{wh})$对 n 的一阶偏导数和二阶偏导数在 $\mu_n$ 点的值.

注意到 $(n - \mu_n)^2$ 与 $\sum_n$ 存在着某种关系,不妨用 $\rho\sum_n$ 来近似 $(n - \mu_n)^2$,其中 $\rho$为一个可调整的比例系数,因此式(9)可进一步简化为:

$$X \approx Y + f(Y, \mu_n, h_{wh}) + \nabla_n f(Y, \mu_n, h_{wh})(n - \mu_n) + (1/2)\nabla_n^2 f(Y, \mu_n, h_{wh})\rho\sum_n \quad (10)$$

这样就得到一个倒谱域上的环境模型. 若用 $\Phi$来代表环境特征 $\mu_n$, $\Sigma_n$, $\rho$和 $h_{wh}$,则利用估计到的 $\Phi$及观测到混噪信号 Y,就可以估计出纯净的语音信号 X.

## 3　环境特征的判别学习

设有一个观测到的训练语音信号的倒谱序列集合$\{Y_1, Y_2, \cdots, Y_M\}$和一个类别集合$\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$,假定所有属于类别 $\lambda_i$ 的纯净语音信号 $X_m(m = 1, 2, \cdots, M)$可被等分为相等的几段,相同类别的每一段服从一个高斯分布 $N_{X_{m,a}}(\mu_{i,a}, \sum_{i,a})$(对 $X_m$ 的第 $a$ 段,$a = 1, 2, \cdots, A$). 本文定义的用于学习环境特征的分类器,将估计出来的纯净语音信号分布模型化为服从如下的概率密度函数的形式:

$$P(X_m|\lambda_i) = \prod_{a=1}^{A} P(X_{m,a}|\lambda_i) = \prod_{a=1}^{A} \frac{1}{(2\pi)^{d/2}|\sum_{i,a}|^{1/2}} \exp\left(-\frac{1}{2}(X_{m,a} - \mu_{i,a})^t \sum_{i,a}^{-1}(X_{m,a} - \mu_{i,a})\right) \quad (11)$$

$d$ 是 $X_{m,a}$的维数,均值向量 $\mu_{i,a}$和协方差矩阵 $\Sigma_{i,a}$分别为:

$$\mu_{i,a} = E_i(X_{m,a}) \quad (12)$$

$$\sum_{i,a} = E_i((X_{m,a} - \mu_{i,a})(X_{m,a} - \mu_{i,a})^t) \quad (13)$$

环境特征的判别学习可用如下的步骤来实现:

(1)定义判别函数:按照公式(11)给出的模型,判别函数可定义为,

$$g_i(X_m, \Phi) = \log P(X_m|\lambda_i) \quad (14)$$

其隐含的分类规则为:

$$X_m \in \lambda_i, \text{当 } g_i(X_m, \Phi) = \max_j g_j(X_m, \Phi) \quad j = 1, 2, \cdots, N \quad (15)$$

(2)定义误分类测度:对上述给定的判别函数,其误分类测度定义为,

$$d_i(X_m, \Phi) = -g_i(X_m, \Phi) + \max_{j \neq i} g_j(X_m, \Phi) \quad (16)$$

$d_i(X_m, \Phi) > 0$ 隐含着误分类,而 $d_i(X_m, \Phi) \leq 0$ 意味着正确的分类.

(3)定义代价函数:代价函数定义为 $d_i(X_m, \Phi)$的 Sigmoid 函数,

$$\zeta_i(X_m, \Phi) = 1/[1 + \exp(-\xi d_i(X_m, \Phi))] \quad (17)$$

(4)平均代价函数:对所有的训练样本 $X_m(m = 1, 2, \cdots, M)$,其平均代价函数为,

$$L(\Phi) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\zeta_i(X_m, \Phi)1(X_m \in \lambda_i) \quad (18)$$

其中 $1(\ell)$是一个逻辑值 $\ell$的 indicator 函数.

(5)参数 $\Phi$可通过最小化平均代价函数 $L(\Phi)$迭代地求得,

$$\Phi^u = \Phi^{u-1} - \eta(u)\nabla L(\Phi^{u-1}) \quad (19)$$

其中 $\Phi^u$ 是第 $u$ 次迭代时的环境特征,$\nabla L(\Phi^{u-1})$是平均代价函数的梯度,而 $\eta(u)$是为了控制训练过程的收敛速度而采用的步长.

这样,环境特征可以通过沿着平均代价函数梯度下降方向不断地进行迭代优化. 在估计到了环境特征后,就可以在此基础上进行纯净语音特征的估计.

## 4　实验情况

通过对韩国语 84 个孤立词的电话语料库进行实验评测了所提出方法的性能. 其中训练语料库中包含了 11381 个发音,测试语料库中包含了 8036 个发音. 基本识别系统是一个不特定话者的 HMM 系统,HMM 模型结构中的状态数是与该词中所包含的音素有关的. 输入的语音信号先进行 Mel 频谱分析,求得 40 个 Mel 频带上的对数频谱,然后求其 12 个 MFCCs 系数,以及一阶差分 MFCCs 系数.

由于使用语料库中全部语音数据进行环境特征学习时训

练周期较长,所以使用了其中一小部分的语音数据来学习有关的特征. 考虑到判别学习方法中,只有易混淆的词才对参数的调整贡献较大. 因此,选取了全词表中 18 个易混词的语音数据来进行特征学习. 然后再将这种实验中获得的特征应用到全词表的实验中.

对式(19)中的收敛调整步长 $\eta(u)$,取 $\eta(u) = 1/(T_c + 2u)$,其中 $T_c$ 是一个预先定义的较大值(在实验中采用了 50[8]),因此,$\eta(u)$ 是随着迭代次数的增加而逐渐减少的. 对当前发音中噪声 n 的估计,采用了保留该发音前几帧无声段,使用它们的平均来代表其估计. 对式(7)中的 $\alpha$ 和 $h_{cu}$ 它们是根据倒谱平均减的假设通过当前的发音来估计的,其内容另文讨论. 前面反映环境函数的公式中有一个离散余弦变换矩阵的逆矩阵,实验中,使用 MATLAB 先将该矩阵中的各元素求出,并写成一个文件的形式,在使用的时候是通过查表的方式获得各个元素值的.

通过实验探讨了所定义的分类器中不同样本分段对性能的影响,比较了将小词表中的词全都等分为相同的段,与对词表中不同的词,依据其所包含的音素个数的不同而等分为不同段时的系统性能,其误识率列于表 1 中.

表 1 使用小词表不同分段情况下的误识率

| 分段情况 | 等分为两段 | 等分为三段 | 等分为四段 | 基于音素的分段 |
|---|---|---|---|---|
| 误识率 | 8.39 % | 9.04 % | 8.10 % | 7.73 % |

可以看出,基于音素的分段优于将全部词都等分为相同段的情况. 使用全部语音库中的数据对所提出的这种 DEFE (Discriminative Environment Feature Extraction) 方法进行了性能评价,实验中采用了基于音素的分段方法. 并将这种方法与使用先前的对加性噪声和通道畸变的补偿方法 LIN-LOG RASTA[11] ($J = 10^{-6}$ 时的情况) 进行了性能比较. 表 2 给出了对全部测试语音采用上述方法时的系统误识率,以及它们与基本识别系统 (Baseline) 相比较误识率的下降情况. 可以看出,DEFE 的性能优于基本识别系统,以及 LIN-LOG RASTA 系统. 因此它能对同时受加性噪声和通道畸变影响的语音进行较好的补偿.

表 2 使用全词表 DEFE 方法的识别结果

| 方 法 | 误识率 | 误识率的下降率 |
|---|---|---|
| Baseline | 9.9 % | — |
| LIN-LOG RASTA | 7.0 % | 29.3 % |
| DEFE | 6.6 % | 33.3 % |

与先前的学习环境特征的方法相比,DEFE 的方法是从最小分类错误的角度来学得环境特征的,从方法上与原来的工作有本质的不同. 与 DFE 相比,DEFE 与之有两点不同,其一,它是利用 MCE 准则和梯度下降法来学习环境特征,而不是来学习语音特征;其二,它通过使用一个简单的分类器,而不是使用后端的 HMM 分类器来学得环境特征,这样环境特征的学习过程并不影响后端的 HMM 分类器的结构. 因此其学习环境特征的过程要比直接使用 DFE 简单.

## 5 结束语

本文提出了一种类似于 DFE 的方法来学习环境特征. 不同于 DFE,所提出的方法利用一个简单的分类器和梯度下降法迭代地学习环境特征. 在获得这些特征后,利用观测到的混噪语音特征来估计纯净语音特征,并将估计出来的纯净语音特征用到后端的 HMM 分类器中.

**致谢**:感谢徐近需教授的积极讨论和帮助. 感谢韩国科学院系统工程研究所的 Park Gyu-Bong、Han Munsung、Park Jeongue 等协助完成本实验.

**参考文献**:

[1] Acero A, Stern R. Environmental robustness in automatic speech recognition [A]. Proceedings of ICASSP90 [C], New Mexico: Lonnie Ludeman, 1990: 849 - 852.

[2] Pedro J, Bhiksha R, Evandro G, Stern R. Multivariate-gaussian-based cepstral normalization for robust speech recognition [A]. Proceedings of ICASSP95 [C], Michigan: Diane Drago, 1995: 137 - 140.

[3] Pedro J, Bhiksha R, Stern R. A vector taylor series approach for environment-independent speech recognition [A]. Proceedings of ICASSP96 [C], Atlanta: Vijay K, Madisetti, 1996: 733 - 736.

[4] B, Katagiri S. Discriminative learning for minimum error classification [J]. IEEE Trans. On Signal Processing, 1992, 40(12): 3043 - 3054.

[5] Amari S. A theory of adaptive pattern classifiers [J]. IEEE Trans. on Electronic Computers, 1967, 16(3): 299 - 307.

[6] Chang P, Juang B. Discriminative temple training for dynamic programming speech recognition [A]. Proceedings of ICASSP92 [C], San Francisco: Les Niles, 1992, (1): 493 - 496.

[7] Chou W, Juang B, Lee C. Segmental GPD training of HMM based speech recognizer [A]. Proceedings of ICASSP92 [C], San Francisco: Les Niles, 1992, (1): 473 - 476.

[8] Biem A, Katagiri S. Feature extraction based on minimum classification error/ generalized probabilistic descent method [A]. Proceedings of ICASSP93 [C], Minnesota: Barry J. Sullivan, 1993, (2): 275 - 278.

[9] Bacchiani M, Aikawa K. Optimization of time-frequency masking filters using the minimum classification error criterion [A]. Proceedings of ICASSP94 [C], Adelaide: Lever K. 1994, (2): 197 - 200.

[10] Rathinavelu C, Deng L. HMM-based speech recognition using state-dependent linear transforms on mel-warped DFT features [A]. Proceedings of ICASSP96 [C], Atlanta: Vijay K. Madisetti, 1996: 9 - 12.

[11] Hermansky H, Morgan N, Hirsch H. Recognition of speech in additive and convolutional noise based on RASTA spectral processing [A]. Proceedings of ICASSP93 [C], Minnesota: Barry J. Sullivan, 1993: 83 - 86.

**作者简介**:

**韩纪庆** 1964 年出生,副教授,教研室常务副主任. 分别于 1987 年 7 月,1990 年 3 月在哈尔滨工业大学电气工程系模式识别与智能控制专业获学士和硕士学位,1998 年 7 月在哈尔滨工业大学计算机系计算机应用技术专业获博士学位. 主要从事语音信号处理方面的研究.

**高 文** 1957 年生,教授,博士生导师,主要从事智能计算机接口和人工智能应用研究.