鲁棒性语音识别中的一种特征参数规整的优化算法

杜俊，胡郁，王仁华

中国科学技术大学　电子工程与信息科学系　合肥　230027

jdu3@mail.ustc.edu.cn

摘要

为了提高语音识别系统的鲁棒性，本文提出了一种特征参数规整的优化算法。整个算法由环境选择、MFCC 差分扩展、均值方差规整(Mean and Variance Normalization, MVN)和 ARMA 滤波器平滑四个模块组成。首先我们对扩展和平滑这两个模块进行了一系列的优化，然后再加入环境选择的思想进一步提高了性能。在 Aurora2 数据库上总识别率的相对提升达到了 53.23%，要明显优于传统的各种参数规整方法，并且和 ETSI AFE 标准前端的性能基本持平。

关键词：鲁棒性语音识别；参数规整；环境选择；　参数优化

1. 引言

在实际应用时，有许多原因可能会导致语音识别系统的识别率显著下降，这些原因包括语音采集环境的影响(如加性噪声，录音设备，信道畸变等)和说话人的影响(如说话风格，口音，以及环境影响引起的说话风格的变化等)。为了使语音识别系统在面对这些不利条件时也能具有较好的性能，采用了许多方法来增强系统的鲁棒性(Robustness)。这些方法总的来说可以分为两大类：第一类是自适应方法，主要是着眼于对声学模型进行变换以适应特定的使用环境；第二类是参数规整方法，主要通过对语音特征参数的变换来减小训练和使用环境之间的不匹配程度。

倒谱均值规整(Cepstral Mean Normalization, CMN)方法是规整方法的一个典型代表，但是一般只能用来补偿信道畸变的影响，这是它的局限。MVN 方法[1]则是同时规整特征矢量的均值和方差，因而对加性噪声也有一定的效果。直方图均衡方法[1]，是一种利用特征参数的累积直方图的规整方法，取得了比 MVN 更好的结果。此外也有人将直方图均衡方法进一步发展，提出了基于分位数的直方图均衡方法[2][3]，这种方法只用少量的数据便可获得数据分布的累积直方图；或者把它与其他方法结合起来，比如谱相减[4]，矢量泰勒级数(Vector Taylor Series, VTS)[5]等；还有对语音段和噪声段分别计算累积直方图的均衡方法[6]，但这种方法需要一个 VAD 算法来很好的区分语音段和噪声段，有时这并不容易做到；将直方图均衡作为特征矢量的自适应变换方法实验也取得了相当好的效果[7]。

本文中我们在 MVN 规整方法的基础上，对与之相关的扩展和平滑两个模块进行了优化；同时引入了环境选择的思想，因为在实际环境中，噪声类型和级别是多种多样的，如果我们的降噪策略也围绕变化的噪声有针对性的变化，那么必然会带来性能的提高。

下面的内容是这样安排的。在第 2 节中，将阐述特征参数规整的重要原理—累积分布函数匹配原理，并简单介绍 MVN 和直方图均衡方法。在第 3 节中，我们将具体说明算法的组成。在第 4 节中，一方面通过实验结果来说明算法各模块的优化过程，另一方面将比较我们的算法和其他一些典型算法在 Aurora2 数据集上的结果。结论将在第 5 节中给出。

2. 特征参数规整方法

2.1. 累积分布函数匹配原理

目前语音识别方法的概率统计框架要求系统的训练环境和测试环境之间应该尽可能匹配，二者之间的失配将严重影响系统性能，甚至使得系统完全不具有实用性。

为了减小二者之间不匹配的程度，一个非常直接的想法就是对训练或者测试的语音参数进行某种变换，以使得它们的概率分布能够比较接近，从而减小训练和测试的失配程度。我们也可以通过使得二者的概率密度函数的积分—累积分布函数(Cumulative Distribution Function, CDF)匹配，来做到这一点。根据这个原理，变换函数可以由数据的累积分布函数获得，如下：

设参数变换函数为 $x = T[y]$，y 是规整前的特征参数，x 是规整变换后的特征参数。

再设 x 的累积分布函数为 $C_X(x)$，y 的累积分布函数是 $C_Y(y)$，则参数变换函数应该使得：

$$C_Y(y) = C_X(x) \tag{1}$$

由此可以得到：

$$x = T[y] = C_X^{-1}(C_Y(y)) \tag{2}$$

上述方法也被称为参数补偿，实际应用中，为了算法实现的方便，经常把训练和测试的数据概率分布都变换到同一个事先给定的标准分布(通常是标准高斯分布)，这称为参数规整。

2.2. 均值方差规整方法(MVN)

MVN 是目前比较有效的鲁棒性方法之一。该方法的基本原理是通过特征参数的均值和方差来对其进行规整，它可以

看成是累积分布函数匹配原理中参数满足标准高斯分布的一个应用。

设带噪语音特征序列（一般是指一句话）为 $Y=[y_0,y_1,\ldots,y_{T-1}]$，则 MVN 的算法公式如下：

$$x_t=\frac{y_t-\bar{y}}{\sigma_y} \tag{3}$$

$$\bar{y}=\frac{1}{T}\sum_{t=0}^{T-1}y_t \quad \sigma_y=\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}(y_t-\bar{y})^2} \tag{4}$$

其中，下标 t 表示第 t 帧，T 是一句话的总帧数，$\bar{y}$ 和 σ_y 分别是一句话的均值和方差。

2.3. 直方图均衡方法

直方图均衡方法是累积分布函数匹配原理应用的另一个例子。这类方法被广泛的应用在图象处理中，最近被应用到鲁棒性语音识别中，取得了较好的结果。与 MVN 不同的是，直方图均衡使用非参数方法来估计累积分布函数。实际应用中，通常是用参数的累积直方图来近似表示累积分布函数，所以称为直方图均衡。

此外，也有人提出了将谱相减方法和直方图均衡方法结合起来的方法[4]，这是目前直方图均衡方法中效果较好的一个。首先用谱相减方法在频域中减小加性噪声的影响，然后在倒谱域中用直方图均衡方法对残留的由加性噪声和信道引起的非线性畸变进行补偿。

3. 特征参数规整的优化算法

3.1. 算法框图

我们的算法仍然是基于累积分布函数匹配原理的，因为它的核心模块之一是 MVN。整个算法的框图如下图所示，输入的是 13 维 MFCC，输出的是规整之后的 39 维 MFCC。对每个模块我们将一一加以说明。

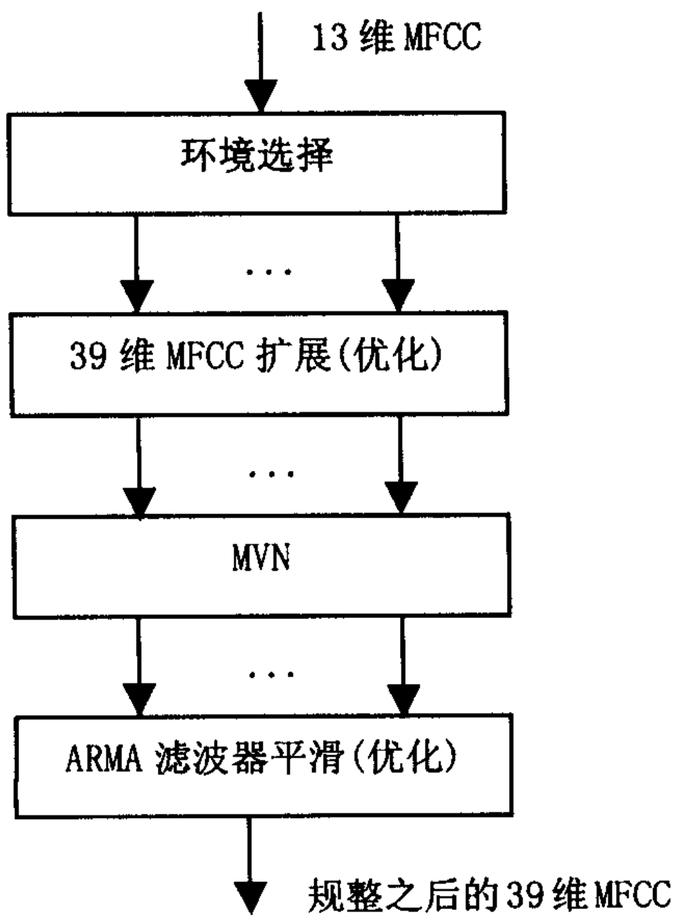

图 1　优化算法框图

3.2. 环境选择模块

环境选择的思想是这样的：对于实际的一个识别系统的应用环境，我们总可以按噪声类型和级别事先对环境细分成很多种小环境，在每个小环境定义了不同的降噪策略。然后我们用每个小环境的数据分别训练一个高斯混合模型 GMM(Gaussian Mixture Model)，从而这个 GMM 就表征了这个小环境的特性。当测试时，我们对每个 GMM 计算当前一句话的似然值，选择似然值最大的那个做为当前的环境，然后进行相应的操作。用 E_i 表征第 i 个环境，则似然值的计算如下：

$$L(Y\mid E_i)=\prod_{t=0}^{T-1}P(y_t\mid E_i) \tag{5}$$

$$P(y_t\mid E_i)=\sum_{j=1}^{M}w_{ij}N(y_t;\mu_{ij},\sigma_{ij}) \tag{6}$$

这里我们假设 GMM 的方差矩阵都是对角阵，其中 w_{ij}、μ_{ij}、σ_{ij} 分别是权值、均值和方差，混合高斯数对每个环境都定义为 M；T 是一句话的总帧数。

假设环境数有 K 个，理论上说应该对这 K 个环境都定义不同的操作，而这些操作的定义其实一定程度上又会依赖于应用的环境。也就是说，对于给定应用环境的语音识别，可以通过环境选择很好的提高性能。当前环境选择的结果只在 ARMA 滤波器平滑模块中加以使用，细节将在后面说明。

3.3. 39 维 MFCC 扩展模块

这个模块是对输入的 13 维特征扩展得到其一阶差分(13 维)和二阶差分(13 维)，这样总共就是 39 维特征。一阶差分在 HTK 中[10]定义如下：

$$\Delta y_t=\frac{\sum_{n=1}^{N}n^2\dfrac{(y_{t+n}-y_{t-n})}{2n}}{\sum_{n=1}^{N}n^2} \tag{7}$$

其中，N 表示差分窗口大小，下标 t 表示第 t 帧，$\mathbf{y}_t$ 表示第 t 帧的 MFCC，$\Delta\mathbf{y}_t$ 表示第 t 帧 MFCC 的一阶差分。

二阶差分只要将一阶差分代入 Eq.(7)即可得到。当前很多识别系统中差分扩展缺省使用的都是 Eq.(7)，下面给出一种优化的差分定义：

$$\Delta y_t=\frac{\sum_{n=1}^{N}(N-n+1)\dfrac{(y_{t+n}-y_{t-n})}{2n}}{\sum_{n=1}^{N}(N-n+1)} \tag{8}$$

Eq.(8)其实就是将 Eq.(7)中的权重 n^2 改成了 $N-n+1$，这样之所以会更优那是因为存在这样的经验：越

靠近当前帧的帧和当前帧的相关性越大，也就是影响越大，在这里表现为公式中的权重应该随 n 的变大而变小。Eq. (7) 不满足这一点，而 Eq. (8) 满足。

下面讨论一下一阶和二阶差分窗口 N 的选取。我们知道，差分的作用一方面是使得语音成分更加明显的表现出来，另一方面可以去除一些比较稳定的噪声成分，因此差分后的特征比原来的特征更加鲁棒。但是窗口 N 应该选的适当，N 太小则不能很好的体现帧与帧之间的相关性，N 太大会破坏当前帧的语音成分同时也不能较好的去除一些非稳态噪声，所以存在一个平衡的最优值。

3.4. MVN 模块

MVN 在 2.2. 节中已经提及，这里就不再赘述了。

3.5. ARMA 滤波器平滑

经过 MVN 模块之后，加性噪声和信道畸变都得到了一定的补偿，但是由于噪声引起的一些毛刺对性能仍有影响，因此这里使用 ARMA 滤波器进行平滑，经典的 ARMA 滤波器如下：

$$\hat{y}_t = \frac{\sum_{l=1}^{L} \hat{y}_{(t-l)} + \sum_{l=0}^{L} y_{(t+l)}}{2L+1} \tag{9}$$

其中，L 表示平滑窗口大小，$\mathbf{y}$ 和 $\hat{\mathbf{y}}$ 分别表示平滑前后的特征。对 Eq. (9) 同样存在优化问题，下面给出一种优化的 ARMA 滤波器定义：

$$\hat{y}_t = \frac{\sum_{l=1}^{L} (L+1-l)\hat{y}_{(t-l)} + \sum_{l=0}^{L} (L+1-l)y_{(t+l)}}{(L+1)^2} \tag{10}$$

Eq. (10) 更优的原因和前面分析差分模块时很相似，离当前窗口中心位置越近的帧与当前帧的相关性越大，也就是权重应该越大，Eq. (10) 正是使用了这样的加权策略。

下面讨论一下窗口 L 的选取。我们知道，干净语音中的一些突变的峰值往往代表着很重要的信息，而带噪语音中的一些毛刺则常是由噪声引起的，因此平滑时应该兼顾两方面。L 太大，虽有较好的抗噪性，但同时会牺牲语音信息；L 太小，则无法很好滤除噪声，所以存在一个平衡的最优值。

4. 实验结果

4.1. 实验数据和相关配置

我们的实验是在ETSI 制定的Aurora2 数据库上进行的。Aurora2 是人工加入噪声和信道影响的 TI 数字串数据集。规定了两种声学模型训练模式：一种模式是用干净语音训练(称为 Clean)，另一种是用干净语音和带噪语音混合训练(称为 Multi)。对每种训练模式，都要进行三个集合的测试：A 集，测试和训练噪声类型相同；B 集，测试和训练噪声类型不同；C 集，不仅有加性噪声还有信道不匹配的影响。

Aurora2 对语音识别的后端配置也有一些标准[8]，规定使用 HTK 来进行训练和测试；每个数字模型有 16 个状态，每个状态 3 个混合高斯；静音(silence)模型有 3 个状态，每个状态 6 个混合高斯；短暂停顿(short pause)模型只有 1 个状态，并且这个状态和静音模型的中间状态绑定在一起。这样，Aurora2 就确定了一个规定训练和测试数据集以及识别后端的数字串识别系统，它的目的就是在完全相同的训练和测试条件下，比较不同前端鲁棒性方法的效果。

Aurora2 的基线结果在[8]中给出，下面提到的所有实验结果都是相对于基线结果的错误率下降(或识别率提升)。实验中所有算法对训练集和测试集都同时进行处理。

前面算法框图中提到的 13 维 MFCC 特征是指 C0-C12，基本按照[8]计算，唯一需要修改的是将 FFT 之后计算幅度谱部分改成计算功率谱，这样会使识别率有一定的提升。

4.2. 环境数的确定

由于Aurora2 的Multi 训练集按噪声类型和级别共细分了 17 个环境(实际有 20 个，但是有 4 个 clean 环境只看成 1 个)，1 个是 clean；其它 16 个都是带噪环境，其中噪声类型有 4 种，噪声级别也有 4 种(20dB，15dB，10dB，5dB)，所以共 4×4=16。显然这里取环境数 K 为 17 最为方便，因为直接用这 17 个环境的数据即可训练得到各个环境的 GMM。GMM 中的混合高斯数 M 设为 16。

4.3. 未进行优化的实验

未进行优化是指在算法框图中去掉环境选择模块，扩展采用 HTK 的缺省配置 Eq. (7)，使用经典的 ARMA 滤波器 Eq. (9)。其中一阶差分窗口为 3，二阶差分窗口为 2，ARMA 滤波器平滑窗口为 4，这些值在未优化之前的算法中也是最优的。从表 1 结果来看，识别率提升已经很高了。

表 1 未进行优化的性能

Aurora2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	32.29%	37.23%	50.06%	38.52%
Clean	59.33%	65.88%	55.73%	61.62%
Average	45.81%	51.56%	52.89%	50.07%

下面我们将进行各模块的优化实验。优化分两步，第一步是对 ARMA 滤波器模块进行优化，包括滤波器本身优化，利用环境选择进行窗口优化；第二步是对扩展模块优化，包括差分定义和差分窗口的优化。

优化过程中的一些实验(如表 2)我们将只给出类似于表 1 最后一列的结果，其中 Clean，Multi 和 Average 这三项分别和表 1 中的相对应。

4.4. 优化第一步---ARMA 滤波器平滑模块

前面说过在 ARMA 滤波器平滑模块会使用环境选择结果，当前的策略如下：

$$L=\begin{cases}L_1 & \text{环境选择判为clean} \quad L_1<L_2 \\ L_2 & otherwise\end{cases} \qquad (11)$$

其中 L 是平滑窗口，clean 是 4.2.节中提到的 17 种环境中的一种。Eq.(11)的原理如下：对于信噪比高的语音，峰值多表示语音成分；而对于信噪比低的语音，噪声也会产生峰值。如果不用环境选择，对各种信噪比的语音都采用同样的平滑程度(即窗口相同)，总体自然会有一些效果，但是明显不是最优的。Eq.(11)正是从这个思路提供了一种方案，对信噪比高的语音平滑程度要比信噪比低的小一些(即 L1<L2)。

另外，从 Eq.(11)可以看出，虽然定义了 17 种环境，但实际上只有两种操作。从理论上说，如果将操作划分的更细一些，性能会更好。但是这里之所以没有这样做主要是因为当前我们环境选择的正确率暂时只能对 clean 的判决达到 100%,其它各种信噪比级别下总会存在一些误判，故采用了这种保守的策略。

表 2 是 ARMA 滤波器优化前后的性能对比，平滑窗口 L=4。滤器器优化前的结果其实就是表 1 中最后一列的结果，优化后 Clean 和 Multi 上的性能都有了一致的提高。

表 2 ARMA 滤波器优化前后的性能(L=4)

	Clean	Multi	Average
优化前	61.62%	38.52%	50.07%
优化后	62.03%	40.51%	51.27%

定义(L1,L2)表示 Eq.(11)中的窗口组合，表 3 是窗口组合的环境选择实验。从 Average 识别率来看，(L1,L2)=(3,4)时是最优的，这正是保留语音成分和平滑噪声平衡后的结果。这样，在 ARMA 滤波器平滑模块中，错误率下降从 50.07%上升到 51.83%，有较明显的性能提升。

表 3 窗口组合的环境选择实验

(L1,L2)	Clean	Multi	Average
(3,4)	62.04%	41.62%	51.83%
(2,4)	61.24%	41.29%	51.27%

4.5.优化第二步—差分扩展模块

这步优化是在 4.4.节中优化之后的基础上进行的。

定义(N1,N2)表示一阶差分窗口是 N1，二阶差分窗口是 N2。

表 4 是差分算法优化前后的性能对比。其中一阶差分窗口 N1=3，二阶差分窗口 N2=2。优化后 Clean 和 Multi 上的性能都有了一致的提高。

表 4 差分算法优化前后的性能(N1=3,N2=2)

	Clean	Multi	Average
优化前	62.04%	41.62%	51.83%
优化后	62.86%	41.96%	52.41%

表 5 是差分窗口的选择实验。当差分窗口(N1,N2)=(3,3)时，Average 识别率达到最佳。这样，在 39 维扩展模块中，错误率下降从 51.83%上升到 53.23%，性能提升也很明显。

表 5 差分窗口的选择实验

(N1,N2)	Clean	Multi	Average
(2,3)	63.02%	38.21%	50.61%
(3,3)	63.89%	42.57%	53.23%
(3,4)	63.39%	42.14%	52.76%
(4,3)	63.07%	42.64%	52.85%

4.6.新算法与其它算法分析比较

本节中将我们的新算法和当前一些文献中较好的方法加以对比。各种性能列于表 6-8。

从复杂度来看，ETSI AFE 最大，因为它是一个以维纳滤波为核心模块综合其它多种方法的前端；谱相减和直方图均衡方法其次；我们的算法最小。但从性能来看，新算法的结果要远远优于谱相减和直方图均衡组合方法，而和 ETSI AFE 的结果也基本上持平。

此外，从表 1 和表 7 对比来看，我们的新算法优化前后在各种环境和训练模式下的识别率都是一致提升的，这进一步说明了我们的优化方案具有很好的普适性。

表 6 谱相减和直方图均衡组合方法的性能[4]

Aurora2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	9.85%	17.51%	22.36%	15.42%
Clean	50.15%	64.27%	49.13%	55.59%
Average	30.00%	40.89%	35.74%	35.51%

表 7 新算法(优化后)的性能

Aurora2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	36.98%	41.01%	53.60%	42.57%
Clean	61.81%	68.39%	56.84%	63.89%
Average	49.40%	54.70%	55.22%	53.23%

表 8 ETSI AFE 的性能

Aurora2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	37.09%	42.47%	41.82%	40.39%
Clean	67.71%	70.74%	57.26%	67.28%
Average	52.40%	56.60%	49.54%	53.84%

5. 结论和展望

本文提出了一种特征参数规整的优化算法，性能相当不错。优化主要体现在差分扩展模块和 ARMA 平滑模块。优化过程中使用的环境选择思想值得我们注意，本文中它只在 ARMA 滤波器模块进行了简单的使用。其实环境选择只是提

出了一种框架，它并不拘泥于某种方法，可以应用于各种各样的前端方法，这会是我们以后继续研究的方向之一。此外，环境选择模块本身也能进行更深入的研究，比如可以尝试其他判决准则如 MCE 等，对用于环境选择的 MFCC 特征先进行一些处理从而提高判决正确率等等。现在我们也正在考虑将一些参数补偿方法加入到当前框架中，进一步体现环境选择的优势。

6. 参考文献

[1] A. de la Torre, J. C. Segura, C. Benitez, etc. Non-linear transformations of the feature space for robust speech recognition. ICASSP 2002, pp.401-404. 2002

[2] F. Hilger, H. Ney. Quantile based histogram equalization for noise robust speech recognition. EUROSPEECH 2001, pp. 1135-1138. 2001.

[3] F. Hilger, S. Molau, H. Ney. Quantile based histogram equation for online application.ICSLP 2002, pp. 237-240. 2002

[4] J. C. Segura, M. C. Benitez, A. de la Torre, A. J. Rubin. Feature extraction combining spectral noise reductionand cesptral histogram equalization for robust ASR. ICSLP 2002, pp. 225-228. 2002

[5] J. C. Segura, M. C. Benitez, A. de la Torre, etc. VTS residual noise compensation. ICASSP 2002, pp. 409-412. 2002

[6] S. Molau, F. Hilger, D. Keysers, H. Ney.Enhanced histogram normalization in the acoustic feature space. ICSLP 2002, pp. 1421-1424. 2002

[7] S. Dharanipragada, M. Padmanabhan, A nonlinear unsupervised adaptation technique for speech recognition. ICSLP 2000, pp. 556-559.2000

[8] H. G. Hirsch, D. Pearce. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. ISCA ITRW ASR 2000, 2000

[9] ETSI. Speech Processing, Transmission and Quality Aspects(STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms". Tech. Rep. ETSI ES 202 050, ETSI, October 2002.

[10] S. Young and etc. The HTK book(V3.0), July 2000