# 基于 AR-HMM 在线能量调整的语音增强方法

何玉文,鲍长春,夏丙寅
(北京工业大学电子信息与控制工程学院语音与音频信号处理实验室,北京 100124)

**摘 要:** 针对单通道语音增强技术对非平稳噪声的跟踪不准确、噪声抑制效果较差的问题,本文提出一种基于在线能量调整的语音增强方法.该方法以归一化临界带能量为特征,采用高斯混合模型对背景噪声进行分类,利用对应类型噪声的自回归隐马尔可夫模型(Auto-Regressive Hidden Markov Model,AR-HMM)和纯净语音的 AR-HMM,在最小均方误差准则下估计语音和噪声的功率谱.考虑到非平稳环境中训练集和测试集的差异性,需在线调整语音模型和噪声模型中的能量,语音模型的能量调整采用迭代的期望最大化算法;噪声模型的能量调整则利用的是模型训练过程中的能量重估方法,并以最小值控制的递归平均算法确定噪声能量调整的初始值.在 ITU-T G.160 标准下对算法进行性能测试,测试结果表明,本文方法对非平稳噪声的跟踪效果较好,对噪声衰减量较大,收敛时间较短.

**关键词:** 语音增强; 非平稳噪声; 隐马尔可夫模型; 高斯混合模型
**中图分类号:** TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2014)10-1991-07
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.10.019

# Online Energy Adjustment Using AR-HMM for Speech Enhancement

HE Yu-wen, BAO Chang-chun, XIA Bing-yin
(*Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing* 100124, *China*)

**Abstract:** Because the existing single channel speech enhancement technologies perform not well in the tracking and suppression of non-stationary noise, the speech enhancement method based on online energy adjustment is proposed. The normalized critical band energy parameters are employed as the feature in Gaussian mixture model (GMM) to distinguish the background noises. Based on the AR-HMM of clean speech and the noise of corresponding type, the power spectrums of speech and noise are estimated under minimum mean square error (MMSE) criteria. When the differences between the training data and test data are considered in the non-stationary noise environment, the online adjustment method for the speech and noise models is necessary. The scaling factor of speech energy is estimated with the iterative expectation maximization (EM) algorithm and the one of noise energy is estimated with the re-estimation approach similar to the training stage. And the initial scaling factor of noise energy is obtained by minima-controlled recursive averaging (MCRA) algorithm. The evaluation of the proposed method is performed under the standard of ITU-T G.160. The test results reveal that, comparing with the two reference methods, the proposed method performs well in non-stationary noise environments, including larger noise reduction and shorter convergence time.

**Key words:** speech enhancement; non-stationary noise; hidden Markov model; Gaussian mixture model

## 1 引言

现实中,语音信号不可避免地会受到噪声的污染,因此语音增强技术一直是语音信号处理的重要内容.自上世纪 70 年代以来,语音增强经历 40 多年的发展,单通道语音增强算法已形成较为经典的几种算法.根据是否使用参数描述模型可以分为无参数模型法和有参数模型法,无参数模型法包括维纳滤波法,谱减类算法,基于统计模型算法等,有参数模型法包括基于隐马尔可夫模型、高斯混合模型以及码本的语音增强方法.

无参数模型法中的噪声估计多通过对含噪语音功率谱的最小值搜索或递归平滑实现,噪声谱的更新依赖话音激活检测或是语音存在概率的估计,主要包括话音激活检测算法,最小值统计方法和最小值控制的递归平均算法等.在实际的非平稳噪声环境中采用此方法估计噪声时会产生噪声低估,估计噪声较平稳以及残留音乐噪声等问题.而基于参数模型的语音增强算法通过对噪声信号建模,可以较准确的描述噪声的非平稳特性,

收稿日期:2013-06-28;修回日期:2013-09-23;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61072089);北京市教育委员会科技发展计划重点项目(No.KZ201110005005)

从而达到更好的增强效果.

1992 年 Y Ephraim 将隐马尔可夫模型(HMM)引入到语音增强领域中,构造最小均方误差和最大后验概率估计器[1,2],形成语音增强的基础架构.随后邓力等将该算法扩展,噪声类型由白噪声扩展为几类噪声,以适应不同类型的非平稳噪声[3],每类噪声模型由单个状态扩展为多个状态、多个混合.考虑到语音和噪声训练集和测试集能量的匹配问题,Zhao 和 Kleijn 等将激励能量当作随机变量并建立模型[4~6],但求解过程较复杂,不易实现.

本文考虑不同类型的非平稳噪声环境,利用高斯混合模型来判断当前的噪声类型,并结合经典的最小值控制的递归平均算法在线调整语音和噪声的能量,方法易于实现,适用于多种非平稳噪声环境.

## 2 基于在线能量调整的增强方法

图 1 为基于在线能量调整的语音增强算法的原理框图,首先分别对语音和噪声的自回归(Auto-Regressive,AR)系数和激励方差进行线下训练,得到语音和噪声的自回归隐马尔可夫模型(Auto-Regressive Hidden Markov Model,AR-HMM),其中噪声模型为不同类型噪声的模型列表.将含噪语音通过傅里叶变换转换到频域,进而利用高斯混合模型(Gaussian Mixture Model,GMM)判断当前噪声环境类型.然后,根据含噪语音功率谱和语音、对应噪声类型的 AR-HMM,基于最小均方误差准则增强含噪语音可以估计出语音和噪声的功率谱.最后,在线调整语音和噪声功率谱的能量并构造频域的维纳滤波器,增强含噪语音,增强后的信号经过频谱综合转化为时域信号输出.下面具体介绍本文方法的各个环节.

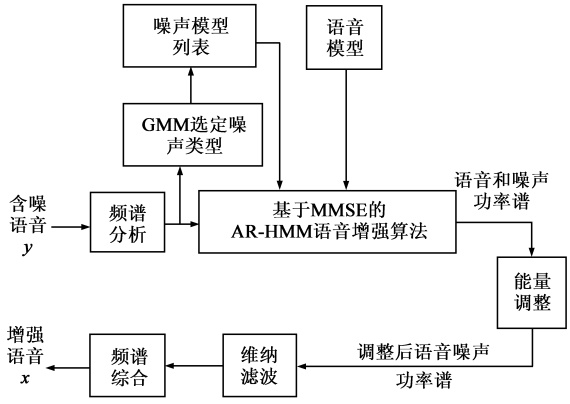



图1 基于在线能量调整的语音增强方法原理框图

### 2.1 信号建模

含噪语音可以表示为纯净语音和加性噪声的叠加,$K$ 维含噪语音可以表示为:

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{w}_t \tag{1}$$

式中 $\boldsymbol{x}_t \in R^K$ 为 $K$ 维纯净语音矢量,$\boldsymbol{w}_t \in R^K$ 为 $K$ 维噪声矢量.

纯净语音[5]可以用为一阶 $\bar{M}$ 个状态的自回归隐马尔可夫模型来描述,混合为零均值高斯自回归分布的矢量子源,"-"表示语音模型参数,$\boldsymbol{x}_0^{T-1} = \{\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{T-1}\}$ 表示 0 到 $T$-1 帧的语音信号的实现序列,其概率 $f(\boldsymbol{x}_0^{T-1})$ 可以写成:

$$f(\boldsymbol{x}_0^{T-1}) = \sum_{\bar{s}} \prod_{t=0}^{T-1} a_{\bar{s}_{t-1}\bar{s}_t} b(\boldsymbol{x}_t \mid \bar{s}_t) \tag{2}$$

式中,$\bar{s}$ 为纯净语音的状态,$\bar{S}$ 为语音状态个数,$a_{\bar{s}_{t-1}\bar{s}_t}$ 为语言模型状态转移矩阵,$b(\boldsymbol{x}_t \mid \bar{s}_t)$ 为观测概率.每个状态下的观测概率由多个高斯自回归的混合组成:

$$b(\boldsymbol{x}_t \mid \bar{s}_t) = \sum_{\bar{m}_t} c_{\bar{m}_t \mid \bar{s}_t} b(\boldsymbol{x}_t \mid \bar{s}_t, \bar{m}_t) \tag{3}$$

式中 $\bar{m}$ 为高斯混合成分,$\bar{M}$ 为斯混合数,$c_{\bar{m}_t \mid \bar{s}_t}$ 为高斯混合的权值.

$b(\boldsymbol{x}_t \mid \bar{s}_t, \bar{m}_t)$ 为给定状态和混合$(\bar{s}_t, \bar{m}_t)$时,得到高斯自回归输出矢量 $\boldsymbol{x}_t$ 的概率密度函数.$b(\boldsymbol{x}_t \mid \bar{s}_t, \bar{m}_t)$ 是关于 AR 系数和激励方差的函数,

$$b(\boldsymbol{x}_t \mid \bar{s}_t, \bar{m}_t) = \frac{\exp\{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}(\Sigma_{\bar{m}_t \mid \bar{s}_t})^{-1}\boldsymbol{x}\}}{(2\pi)^{K/2}\left|\Sigma_{\bar{m}_t \mid \bar{s}_t}\right|^{1/2}} \tag{4}$$

式中,$\Sigma_{\bar{m}_t \mid \bar{s}_t}$ 为状态 $\bar{s}_t$ 混合 $\bar{m}_t$ 的协方差矩阵,$\Sigma_{\bar{m}_t \mid \bar{s}_t} = \sigma^2_{\bar{m}_t \mid \bar{s}_t}(\boldsymbol{A}^{\mathrm{T}}_{\bar{m}_t \mid \bar{s}_t}\boldsymbol{A}_{\bar{m}_t \mid \bar{s}_t})^{-1}$,$\boldsymbol{A}_{\bar{m}_t \mid \bar{s}_t}$ 为 $K \times K$ 阶下三角 Toeplitz 矩阵,第一列前 $P+1$ 个元素由语音 AR 系数 $\bar{\alpha} = [1 \quad \bar{\alpha}_1 \quad \bar{\alpha}_2 \quad \cdots \bar{\alpha}_P]^{\mathrm{T}}$ 组成,$|\cdot|$ 表示矩阵行列式,T 表示转置.类似语音的建模过程可以得到噪声的 HMM 模型,"··"表示相应的噪声模型参数.不同类型噪声若采用统一的噪声模型会增加搜索复杂度,不能精确描述非平稳噪声,这里根据不同噪声建立不同噪声 AR-HMM.

语音和噪声信号依赖 AR-HMM 来描述,而两个 AR 过程的和不一定是 AR 过程,故含噪语音信号通过语音和噪声模型的组合来描述.语音和噪声状态组合成含噪语音的状态,转移概率为转移到组合状态对应的语音和噪声模型的转移概率的乘积.例如,纯净语音模型有 $\bar{S}$ 个状态,噪声模型有 $\ddot{S}$ 个状态,对应含噪语音模型共有 $\bar{S} \times \ddot{S}$ 个状态,当含噪语音在 $t$ 时刻的组合状态$(\bar{s}_i, \ddot{s}_k)$跳转到 $t+1$ 时刻的组合状态$(\bar{s}_j, \ddot{s}_l)$时,含噪语音状态间的转移概率为 $a_{\bar{s}_{t-1}\bar{s}_t}a_{\ddot{s}_{t-1}\ddot{s}_t}$,同理可得高斯混合的组合权值 $c_{\bar{m}_t \mid \bar{s}_t}c_{\ddot{m}_t \mid \ddot{s}_t}$.

假设语音和噪声是独立的,两者之间的协方差矩阵为 0,故含噪语音的协方差矩阵为语音和噪声的协方

差矩阵之和 $\Sigma_{m_t|s_t}=\Sigma_{\bar{m}_t|\bar{s}_t}+\Sigma_{\ddot{m}_t|\ddot{s}_t}$.可以得到含噪语音的观测概率:

$$b(\boldsymbol{y}_t|\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t)=\frac{\exp\{-\frac{1}{2}\boldsymbol{y}_t^{\mathrm{T}}(\Sigma_{\bar{m}_t|\bar{s}_t}+\Sigma_{\ddot{m}_t|\ddot{s}_t})^{-1}\boldsymbol{y}_t\}}{(2\pi)^{K/2}\left|\Sigma_{\bar{m}_t|\bar{s}_t}+\Sigma_{\ddot{m}_t|\ddot{s}_t}\right|^{1/2}} \tag{5}$$

用语音和噪声组合的含噪模型描述含噪信号时,观测概率的 $K\times K$ 阶矩阵操作复杂度较高,当协方差矩阵为循环矩阵时,将有关协方差的矩阵相乘操作转化到频域计算[7],能减少复杂度.

## 2.2　模型训练

语音和噪声的模型训练过程相同,训练采用的模型是遍历连续型 AR-HMM,模型参数 $\lambda$ 为 $\{\pi,\boldsymbol{A},\boldsymbol{C},\boldsymbol{h}\}$. $\pi$ 为初始状态概率,$\boldsymbol{A}$ 为转移概率,$\boldsymbol{C}$ 为高斯混合权,$\boldsymbol{h}$ 为提取的特征参数矢量,包括归一化的 LSF 系数和激励方差.信号模型初始化后通过 Baum-Welch 方法优化模型参数,继而 Viterbi 解码计算两次迭代间输出概率的相对误差,当该误差达到收敛容限或重估次数达到最大迭代次数时,判断模型已收敛,此时的模型为所求模型.

训练中的语音训练数据选自 NTT 标准语音库中的中文、英文和美音的数据子库,信号采样率为 16kHz,采用 16bit 量化,每句话持续约 8s,将所有语句连接起来并去掉静音段得到训练的语音,时长在 22 分钟左右,每帧 512 样点(32ms)预处理时采用 50% 的加窗叠接,窗函数为归一化的汉明窗.模型共有 8 个状态,16 个混合,语音 AR 阶数为 16.噪声的训练数据选自 NOISEX-92[8]和 ITU 噪声数据库,包括嘈杂人声(Babble)、工厂噪声(Factory)、街道噪声(Street)、车内噪声(Volvo)和白噪声(White)5 种噪声情况,采样率为 16kHz,每类噪声的训练数据持续时长在 21 分钟左右,噪声的预处理与语音相同,AR 阶数均为 16,模型共有 3 个状态,3 个混合.

图 2 和图 3 分别为语音和噪声模型随着训练次数增加 Viterbi 输出概率变化的趋势图,从图中可以看出,随着训练迭代次数的增加,训练模型的 Viterbi 概率逐渐趋于平稳,其中 Street 噪声由于训练数据较短,出现了过拟合的现象.训练次数多或少都不能使训练模型较精确的描述信号,语音和噪声训练选取的 Baum-Welch 最大迭代重估次数[1]为 20 次.

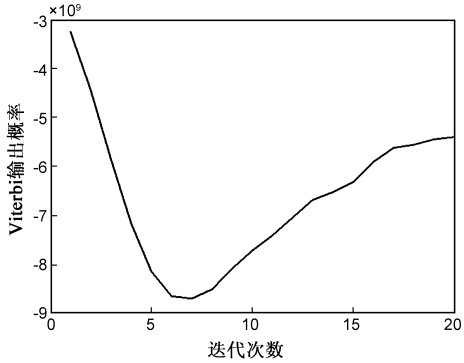



**图2　语音AR-HMM的Viterbi输出概率随训练迭代次数变化趋势图**

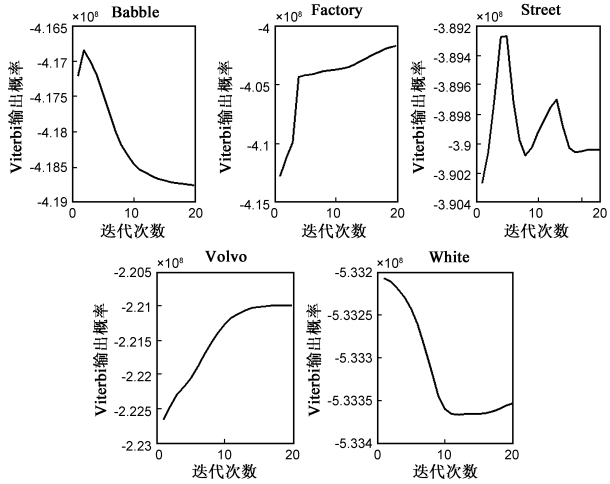



**图3　不同噪声环境下的噪声AR-HMM的Viterbi输出概率随训练迭代次数变化趋势图**

## 2.3　噪声类型选择

在语音增强时需要首先确定当前的噪声环境类型以及相应的噪声模型,噪声模型选择的精确与否严重影响着增强语音质量.这里假定含噪语音开始的 160ms 为纯噪声段,利用基于人耳听觉模型的归一化临界带能量建立高斯混合模型判断噪声处于哪种噪声类型,下面详述噪声类型选择过程.

将采样率为 16kHz 的噪声信号分为 21 个临界带[9],每个归一化临界带能量为每个临界带的能量占整个频带谱能量的比值,

$$Br_i=\sum_{\omega=bl_i}^{\omega=bh_i}\frac{P(\omega)}{P_n} \tag{6}$$

式中,$bh_i$、$bl_i$ 分别为第 $i$ 个临界带的上限和下限,$Br_i$ 为第 $i$ 个归一化临界带能量,$i$ 取值范围为 1 到 21,$P_n$ 为噪声整个频带的谱能量.

图 4 为提取每帧不同噪声特征矢量的示例,从图中可以看出,Babble 噪声能量集中在较宽泛的低频段,白噪声的能量分布较均匀,Factory 噪声的临界带能量比重在 500Hz 以下分布较分散,Street 和 Volvo 噪声能量集中在不同的频段.因此根据归一化临界带能量这一特征矢量可以将 5 种噪声区别出来.

图 5 为 Factory 噪声、Volvo 噪声和白噪声的归一化临界带能量经过主成分分析降维得到的三维特征图,由图可知,利用归一化临界带能量特征能够将该三类噪声区分开.

得到训练的特征矢量后,使用 GMM 训练特征参数,得到高斯混合模型的均值、方差、权重等参数. Babble, Factory, Street, Volvo 和 White 噪声源的时长均较短,将其各自复制 $N$ 遍后构成 20 分钟左右的训练数据. 高斯混合模型中的混合数设为 3,迭代次数为 100,收敛误差为 0.001.

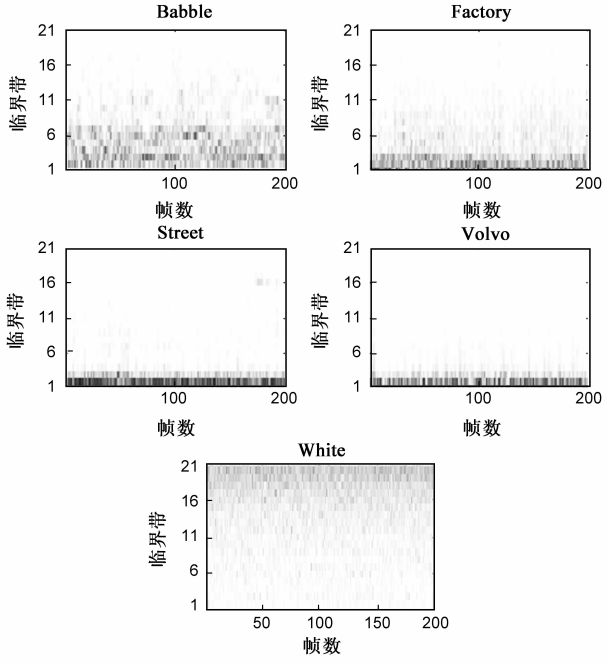



图4 噪声的归一化临界带能量特征

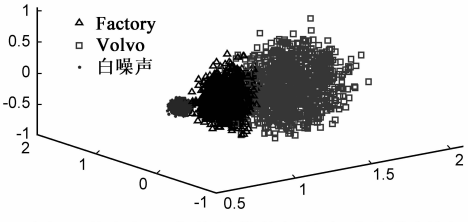



图5 噪声归一化临界带能量的主成分分析图

训练好每种噪声的高斯混合模型后,通过噪声对该几个模型进行测试,表 1 给出了分类结果. 表中左侧栏为实际输入的不同类型的测试噪声数据,表的主体数据为不同输入类型的噪声计算归一化临界带能量特征,经过 GMM 判断为不同类型噪声的输出结果. 从表中可以看出,每类噪声的正确率在 84% 以上,除了 Street 噪声其他类噪声的正确率在 95% 以上,区分度较好.

## 2.4 能量调整

给定含噪语音 $\boldsymbol{y}_0^{\tau}=\{\boldsymbol{y}_0,\cdots,\boldsymbol{y}_{\tau}\}$ 和语音噪声的 AR-HMM 后,可以得到增强语音功率谱的最小均方误差估计(Minimum Mean-Squared Error, MMSE)[10],

表 1 噪声分类结果

| | | 输出噪声类型 | | | | |
|---|---|---|---|---|---|---|
| | | Babble | Factory | Street | Volvo | White |
| 输入噪声类型 | Babble | 97.7% | 2.3% | 0 | 0 | 0 |
| | Factory | 0.6% | 95.6% | 2.6% | 1.2% | 0 |
| | Street | 0 | 7.1% | 84.6% | 8.3% | 0 |
| | Volvo | 0 | 0 | 4.8% | 95.2% | 0 |
| | White | 0 | 0 | 0 | 0 | 100.0% |

$$\left|\boldsymbol{X}_{t,k}\right|^2=\sum_{\ddot{s},\bar{m},\ddot{s},m} f(\ddot{s},\bar{m},\ddot{s},\ddot{m}\mid\boldsymbol{y}_0^{\tau})\cdot E\left[\left|\boldsymbol{X}_{t,k}\right|^2\mid\boldsymbol{y}_0^{\tau},\ddot{s},\bar{m},\ddot{s},\ddot{m}\right] \tag{7}$$

式中 $\left|\boldsymbol{X}_{t,k}\right|^2$ 为纯净语音 $t$ 时刻,$k$ 频点的功率谱,$f(\ddot{s},\bar{m},\ddot{s},\ddot{m}\mid\boldsymbol{y}_0^{\tau})$ 为给定 0 到 $\tau$ 帧的含噪语音,处于语音状态 $\ddot{s}$、混合 $\bar{m}$ 和噪声状态 $\ddot{s}$、混合 $\ddot{m}$ 的概率,$E\left[\left|\boldsymbol{X}_{t,k}\right|^2\mid\boldsymbol{y}_0^{\tau},\ddot{s},\bar{m},\ddot{s},\ddot{m}\right]$ 为给定 0 到 $\tau$ 帧的含噪语音,在语音状态 $\ddot{s}$、混合 $\bar{m}$ 和噪声状态 $\ddot{s}$、混合 $\ddot{m}$ 条件下,纯净语音功率谱的期望.

能量调整分为语音能量调整和噪声能量调整部分,利用经典的递归平均算法从含噪语音功率谱中估计出噪声的功率谱,以该噪声功率谱为目标,第 1 次调整噪声能量,得到第一次噪声能量调整因子 $g_w'$. 在此基础上,根据已知的含噪功率谱,通过期望最大化(Expectation Maximization, EM)迭代估计算法来调整语音功率谱的能量,得到语音能量调整因子 $g_s'$. 确定语音的能量调整值后对噪声的能量再次调整,得到第 2 次噪声能量调整因子 $g_w''$. 上述操作是每组语音和噪声状态和混合的操作,这样调整后的语音和噪声的功率谱可以构成语音和噪声的状态和混合的维纳滤波器组及相应的权值.

噪声能量调整部分是根据目标噪声源,对噪声模型中的激励能量进行调整. 调整公式参考 Y Ephraim 在增益自适应的模型中对语音激励能量的训练方法[2],将估计的噪声功率替代文献中已知的语音功率来估计噪声能量调整因子 $g_{w,t}$:

$$g_{w,t}=\sum_{\ddot{s}_t,\ddot{m}_t} f_{\lambda}(\ddot{s}_t,\ddot{m}_t\mid\boldsymbol{w}_t)\frac{1}{K}\boldsymbol{w}_t^{\mathrm{T}}\Sigma_{\ddot{m}_t\mid\ddot{s}_t}^{-1}\boldsymbol{w}_t \tag{8}$$

给定噪声估计时,每个噪声状态和混合的概率为转移概率和混合权的乘积 $f_{\lambda}(\ddot{s}_t,\ddot{m}_t\mid\boldsymbol{w}_t)=a_{\ddot{s}_{t-1}\ddot{s}_t}c_{\ddot{m}_t\mid\ddot{s}_t}$, $\frac{1}{K}\boldsymbol{w}_t^{\mathrm{T}}\Sigma_{\ddot{m}_t\mid\ddot{s}_t}^{-1}\boldsymbol{w}_t$ 为噪声激励方差的期望估计,可近似在频域计算.

噪声的两次能量调节过程采用同样的方法确定噪声能量调整因子,不同的是估计的噪声目标不同,第一次的噪声功率目标是根据最小值控制的递归平均算法(Minima-Controlled Recursive Averaging, MCRA)估计的噪声功率,第二次的噪声目标为语音能量调整后,含噪语

音通过维纳滤波得到的噪声功率.其中 MCRA 算法估计出的噪声功率较贴近非平稳噪声能量,避免迭代估计噪声能量调整因子复杂度高的问题.

噪声的能量初步调整后,假定噪声功率不变,根据含噪语音功率谱需要调整语音的能量.文献[2]中的模型没有考虑能量信息,利用 EM 算法估计纯净语音能量,但语音能量的变化范围较大,不易估计.本文将语音能量引入到模型中,作为能量调整的初始值,降低能量估计的动态范围,将噪声的状态由单状态扩展为多状态多混合,使之更精细描述噪声功率.

该方法估计变量为语音能量调整因子 $g_x$,通过求取后验概率 $f_\lambda(\boldsymbol{y}\mid g_x)$最大得到,

$$\max_{g_x} f_\lambda(\boldsymbol{y}\mid g_x) \tag{9}$$

式中 $\lambda$ 为纯净语音模型和噪声模型组合成的含噪语音模型.

后验概率 $f_\lambda(\boldsymbol{y}\mid g_x)$关于 $g_x$ 的梯度方程是非线性的,利用 EM 算法迭代估计语音能量调整因子可以得到其重估公式,

$$g_{x,t}(n+1)=\sum_{\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t} f_\lambda(\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t\mid \boldsymbol{y}_t,g_{x,t}(n))\cdot \frac{1}{K}E\{\boldsymbol{x}_t^{\mathrm{T}}\Sigma^{-1}_{\bar{m}_t\mid\ddot{s}_t}\boldsymbol{x}_t\mid \bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_{x,t}(n)\} \tag{10}$$

式中 $g_{x,t}(n)$为 $t$ 时刻第 $n$ 次迭代的语音能量调整值,条件均值可以计算为:

$$E\{\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}_{\bar{m}_t\mid\ddot{s}_t}\boldsymbol{x}\mid \bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_{x,t}(n)\} = \mathrm{tr}\{R_{\boldsymbol{x}\mid\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_t(m)}\Sigma^{-1}_{\bar{m}_t\mid\ddot{s}_t}\} \tag{11}$$

式中,$R_{\boldsymbol{x}\mid\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_{x,t}}\triangleq E\{\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\mid\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_{x,t}\}$为给定$(\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,\boldsymbol{y}_t,g_t)$时,纯净语音 $\boldsymbol{x}$ 样本方差的 MMSE 估计值,tr(.)表示矩阵的迹.因为纯净语音和噪声为高斯自回归过程,可以得到:

$$R_{x\mid\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,y_t,g_{x,t}} = h_{\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,g_{x,t}}\Sigma_{\ddot{m}_t\mid\ddot{s}_t} + [h_{\bar{s}_t},\bar{m}_t,\ddot{s}_t,\ddot{m}_t,g_x,\boldsymbol{y}_t][h_{\bar{s}_t},\bar{m}_t,\ddot{s}_t,\ddot{m}_t,g_x,\boldsymbol{y}_t]^{\#} \tag{12}$$

式中 $h_{\bar{s}_t,\bar{m}_t,\ddot{s}_t,\ddot{m}_t,g_{x,t}}\triangleq g_{x,t}\Sigma_{\bar{m}_t\mid\ddot{s}_t}(g_{x,t}\Sigma_{\bar{m}_t\mid\ddot{s}_t}+g_w'\Sigma_{\ddot{m}_t\mid\ddot{s}_t})^{-1}$为给定 $g_{x,t}$时,在语音第 $\ddot{s}_t$ 个状态和第 $\bar{m}_t$ 个混合,噪声第 $\ddot{s}_t$ 个状态第 $\ddot{m}_t$ 个混合下输出的维纳滤波器.语音能量在时域实现复杂度较大,这里近似在频域实现语音能量调整.

语音能量调整因子的初始值设为含噪模型的真实值与估计值的比,

$$g_{x,t}(0)=\frac{1}{K}\sum_{k=1}^{K}\frac{|\boldsymbol{Y}_k|^2}{\sum_{\bar{s},\bar{m}}f(\bar{s},\bar{m})\bar{\boldsymbol{P}}_k(\bar{s},\bar{m})+\sum_{\ddot{s},\ddot{m}}f(\ddot{s},\ddot{m})\ddot{\boldsymbol{P}}_k(\ddot{s},\ddot{m})} \tag{13}$$

## 3　性能测试

本文方法的测试项目是在 G.160 标准的要求下进行的,G.160 标准是 ITU-T 提出的一项针对移动网络中语音增强设备的国际标准.

在白噪声环境下的性能测试主要包含以下两个方面:

(1)信号电平衰减测试.该部分测试噪声的期望衰减量 $Q_m$,语音信号衰减 $Q_s$和实际噪声衰减量 $Q_n$,需满足 $-3<Q_s<2$,$Q_m-3<Q_n<Q_m+3$.

(2)收敛性测试.该部分测试语音增强方法对三段噪声能量突变处理的衰减时间,收敛时间应保证在 3s 内.

在有色噪声环境下用三种客观测度来描述噪声衰减的作用,分别是信噪比提高(Signal-to-Noise Ratio Improvement,SNRI),整体噪声电平衰减量(Total Noise Level Reduction,TNLR)和语音电平衰减量(DSN).

在 G.160 标准外本文采用语音质量感知评价提升(Perceptual Evaluation of Speech Quality Improvement,PESQI)[11]测试语音增强算法对语音客观质量的影响.

纯净语音信号从 NTT 标准语音库中的中文子库中选择,采样频率为 16kHz,序列长度为 8s.噪声信号从 ITU-T 噪声数据库中选择,包括嘈杂人声(Babble)、工厂噪声(Factory)、车内噪声(Volvo)、街道噪声(Street)和高斯白噪声(White).

本文方法的参考算法有两种,参考算法 1 是基于最小均方误差的 AR-HMM 语音增强方法,该方法没有对语音和噪声能量进行在线调整的过程,噪声类型选择利用 VAD 确定噪声存在段,对噪声存在段判断噪声类型并调整能量;参考算法 2 是最小均方误差的幅度谱语音增强方法[12],该方法通过最小化幅度谱的真实值与估计值的均方误差得到.该方法假设信号的傅里叶变换系数服从高斯分布,没有利用先验的信号特征信息,噪声估计部分采用 MCRA 算法,与本文方法中的噪声估计参数设置相同.

图 6 为白噪声环境下语音和噪声的衰减测试,从图中可以看出,本文方法的 $Q_{n1}$、$Q_{n2}$远远大于参考算法 1,比参考算法 2 高 2dB 左右,同时本文方法和参考算法 1 的语音衰减量比参考算法 2 大,仍在[-3,2]范围内,符合 G.160 标准要求.

表 2 为不同噪声环境下三次突变噪声出现后的收敛测试时间 $T_1$、$T_2$和 $T_3$,从表中可以看出,参考算法 2 的收敛时间从 $T_1$到 $T_3$呈现递减的趋势,而本文方法和参考算法 1 的收敛时间对噪声突变的跟踪速度较快,收敛时间总体较短,特别在 $T_2$时,本文方法和参考算法 1 的收敛时间在 ±0.1s 以内.

表 3 为有色噪声环境下的测试结果,本文方法的信噪比提升(SNRI)远远高于参考算法 1,低于参考算法 2.参考算法 2 为谱细节上的调整,而本文方法与参考算法

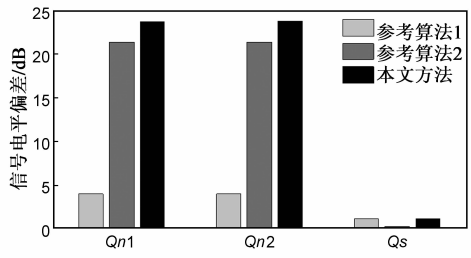



**图6 白噪声环境下的信号电平衰减测试**

1 均是在谱包络能量上的调整,信噪比提升略低.本文方法的整体噪声衰减(TNLR)较大,语音衰减(DSN)三种方法均是负值,表明语音信号被放大,参考算法 1 只对噪声能量进行了调整,语音放大量较少,本文方法和参考算法 2 的 DSN 差距较小,语音被放大的程度相当.

**表 2 不同噪声环境下的收敛时间测试结果**

| | | Babble | Factory | Street | Volvo | White |
|---|---|---|---|---|---|---|
| $T_1$ (s) | 参考算法 1 | 0.16 | 0.16 | 0.15 | 0.15 | 0.21 |
| | 参考算法 2 | 0.64 | 0.73 | 0.67 | 0.69 | 0.72 |
| | 本文方法 | 0.14 | 0.28 | 0 | 0.31 | 0.13 |
| $T_2$ (s) | 参考算法 1 | 0 | 0 | 0 | 0 | 0 |
| | 参考算法 2 | 0 | 0.81 | 0 | 0.78 | 0.81 |
| | 本文方法 | 0.03 | 0 | 0 | 0 | 0.01 |
| $T_3$ (s) | 参考算法 1 | 0 | 0 | 0 | 0 | 0 |
| | 参考算法 2 | 0 | 0 | 0 | 0 | 0 |
| | 本文方法 | 0.1 | 0.47 | 0.04 | 0 | 0 |

**表 3 有色噪声环境下的性能测试结果**

| | 参考算法 1 | 参考算法 2 | 本文方法 |
|---|---|---|---|
| SNRI (dB) | 1.08 | 14.32 | 10.77 |
| TNLR (dB) | -2.89 | -19.68 | -23.57 |
| DSN (dB) | -1.11 | -2.18 | -2.20 |

**表 4 客观语音质量提高测试结果**

| SNR | 参考算法 1 | 参考算法 2 | 本文方法 |
|---|---|---|---|
| 6dB | -0.0065 | 0.34 | 0.27 |
| 12dB | -0.0012 | 0.27 | 0.22 |
| 18dB | -0.0042 | 0.08 | 0.21 |

表 4 为 6dB、12dB 和 18dB 三种信噪比情况下对语音质量的测试结果,参考算法 1 的 PESQI 得分为负值,表明质量下降,本文方法和参考算法 2 的 PESQI 为正值,语音质量有所提升.在信噪比为 6dB、12dB 时,本文方法与参考算法 2 相比,PESQI 得分低 0.05 左右;信噪比为 18dB 时,本文方法比参考算法 2 高 0.1 左右,总体上与参考算法 2 的客观质量相当.本文方法在不同信噪比环境下的质量提升较稳定,参考算法 2 的语音增强结果极大的依赖 MCRA 噪声估计算法中设定的参数,而该参数为经验值,不易调节.

## 4 结论

本文在非平稳噪声环境下,提出一种基于在线能量调整的 AR-HMM 语音增强方法,该方法利用高斯混合模型来判断当前的噪声类型,并将经典的最小值控制递归平均算法引入到在线调整语音和噪声能量中,提高对非平稳噪声的适应能力.根据 G.160 标准测试结果,本文方法与参考算法 1 相比,噪声衰减、信噪比和话音质量提升均有较大提升,与调整谱细节的参考算法 2 相比,能够快速跟踪能量的变化,收敛时间缩短约 7 倍,且噪声衰减量较大,语音质量提升较稳定,较适用于能量变化剧烈的非平稳噪声环境.

**参考文献**


[1] Ephraim Y. A Bayesian estimation approach for speech enhancement using hidden Markov models[J]. IEEE Transactions on Signal Processing, 1992, 40(4): 725 - 735.

[2] Ephraim Y. Gain-adapted hidden Markov models for recognition of clean and noisy speech[J]. IEEE Transactions on Signal Processing, 1992, 40(6): 1303 - 1316.

[3] Sameti H, Sheikhzadeh H, Deng L, Brennan R L. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(5): 445 - 455.

[4] Srinivasan S, Samuelsson J, Kleijn W B. Codebook-based Bayesian speech enhancement[A]. IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. IEEE, 2005. 1077 - 1080.

[5] Zhao D Y, Kleijn W B. HMM-based gain modeling for enhancement of speech in noise[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 882 - 892.

[6] Zhao D Y, Kleijn W B, Ypma A, et al. Online noise estimation using stochastic-gain HMM for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(4): 835 - 846.

[7] Srinivasan S, Samuelsson J, Kleijn W B. Codebook-based Bayesian speech enhancement for nonstationary environments [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(2): 441 - 452.

[8] Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12 (3): 247 - 251.

[9] Johnston J D. Transform coding of audio signals using perceptual noise criteria[J]. IEEE Journal on Selected Areas in Communications, 1988, 6(2): 314 - 323.

[10] Ephraim Y. A minimum mean square error approach for speech enhancement[A]. International Conference on Acoustics, Speech, and Signal Processing[C]. IEEE, 1990. 829 – 832.

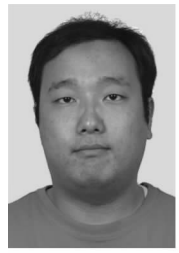
[11] ITU-T Recommendation P. 862. Perceptual Evaluation of Speech Quality(PESQ): An Objective Method for End-to-end Speech Quality Assessmen of Narrow-band Telephone Networks and Speech Codecs[S]. 1996.

[12] Loizou P. Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2005, 13(5): 857 – 869.

## 作者简介

**何玉文**　女,1988 年生于北京,北京工业大学硕士研究生,主要研究方向为语音增强.
E-mail: iamhyw@emails.bjut.edu.cn

**鲍长春**　男,1965 年生于内蒙古赤峰,博士,北京工业大学教授、博士生导师,IEEE 高级会员,国际语音通信学会(ISCA)会员,亚太信号与信息处理学会(APSIAP)会员,中国电子学会理事,中国声学学会理事,信号处理学会委员.主要研究方向为语音与音频信号处理.
E-mail: chchbao@bjut.edu.cn



**夏丙寅**　男,1986 年生于北京,北京工业大学博士生,主要研究方向为语音编码与增强.
E-mail: xby-abc@emails.bjut.edu.cn