汽车噪声中自动语音的识别技术*

韦晓东, 朱 杰, 胡光锐†
(上海交通大学与贝尔实验室通信与网络联合实验室)
(† 上海交通大学电子工程系)

摘 要 汽车中的话音拨号系统是自动语音识别技术的应用热点. 自动语音识别系统是一个基于训练的系统. 在汽车噪声中, 由于实际应用环境与形成系统参数的训练环境的失配, 传统语音识别系统的性能会大幅度地下降, 从而无法实用. 为了提高语音识别系统在特定环境下的识别率及实用性, 首先根据汽车环境中语音的失真模型分析了系统性能下降的原因, 然后针对加性汽车噪声与信道失真对系统的影响, 讨论了在汽车噪声中改善语音识别系统性能的方法. 提出了在识别系统中用基于子带的语音增强算法和倒谱均值相减算法相结合的方法. 对大量的多人连续数字串语音的识别实验表明, 这一方法大大提高了系统在汽车噪声环境中的识别率, 它还可以简便、实时的实现, 具有一定的实用性.

关键词 语音识别; 语音增强; 倒谱均值相减
中图法分类号 TN 912.34

Techniques for Automatic Speech Recognition in Car Noise

Wei Xiaodong, *Zhu Jie*, *Hu Guangrui*†
Shanghai Jiaotong University & Bell Labs Communication and Networks Joint Laboratory, Shanghai, China
† Department of Electronic Engineering, Shanghai Jiaotong University

Abstract Voice dialing system is a hot application issue for automatic speech recognition (ASR). Speech recognition system is always based on training, so its performance degrades significantly when there is a mismatch between testing and training. To improve the robustness of ASR in car noise, this paper analyses the main reason why the performance degrades. For the effect of additional noise and channel distortion, some techniques to improve ASR in car noise are described. The main two methods—— speech enhancement based on subband filtering and cepstral mean subtraction are proposed. Both of them can be real-time implemented to fulfill the requirement of speech recognition system in cars. The experiments show that higher accuracy recognition rate can be achieved in car noise.

Key words speech recognition; speech enhancement; cepstral mean subtraction

---

汽车中的话音拨号系统是语音识别技术的一个应用和研究热点. 目前的语音识别系统一般都是以隐马尔可夫模型(HMM)作为系统的基本模型. 这种系统中的HMM 参数都是利用对语音库的训练得到, 而语音库往往是在一定的环境下采集到的. 所以这些系统都对形成系统的训练环境具有敏感性. 实验表明, 语音识别系统在识别、测试环境与训练环境不一致时, 识别性能明显下降直至无法工作. 汽车中

收稿日期: 1998-04-04
* 美国贝尔实验室上海分部资助项目
韦晓东: 男, 1970 年生, 博士生. 邮编: 200030

的话音拨号系统, 由于有各种汽车噪声的存在, 语音信号会发生明显的失真, 造成环境失配, 从而严重影响了识别性能. 为了在汽车环境下能够利用语音进行电话拨号、设备控制, 必须在系统中采用相应的环境补偿技术. 汽车噪声中语音识别系统的设计需要注意以下几个问题: ① 算法能够实时实现, 用户希望在发出指令后, 系统马上作出反映. ② 计算量与内存不宜太大, 否则硬件花费过大, 增加系统成本. ③ 识别系统输出结果有一定的可靠度, 避免出现一些烦人的误操作. 本文着重讨论汽车噪声中的语音识别技术.

1　语音识别系统基本结构与汽车噪声的特点

语音识别系统基于语音的子词单元(类似于音素), 以 HMM 作为系统的基本模型. 每个模型有 3 个状态、8 个高斯密度函数混和而成. 输入的语音 16 位 PCM 采样、量化后被分成每 30 ms 一帧, 相邻帧有 20 ms 的重叠, 即帧移为 10 ms, 每帧输入的数据通过高通滤波和汉明窗, 然后计算一个 25 维的特征向量用以训练与识别. 25 维向量分别由 12 阶 LPC 倒谱系数、12 阶 LPC 倒谱系数的一阶导数、短时能量的一阶导数组成. HMM 的参数通过对标准的多人干净语音库 TIDIGITS 训练得到. TIDIGITS 主要包括连续数字串语音. 搜索算法采用 Viterbi 算法结合一个语法模型. 用干净的多人语音语音库做识别测试, 词识别率为 98. 43%, 句子(数字串)识别率为 94. 81%.

在开汽车的过程中, 会出现各种背景噪声. 本文主要处理汽车本身的噪声, 包括发动机声响、轮胎摩擦等, 这些噪声一般为加性噪声. 汽车噪声不是白噪声[1]. 它的能量主要集中在低频段, 与语音信号类似. 在功率谱空间, 加性汽车噪声造成了语音信号谱的加性失真. 而用于训练与识别的语音信号的特征空间, 是倒谱向量空间, 功率谱空间的简单加性失真, 在特征向量空间, 形成一种复杂的非线性变换失真. 由于汽车的行驶情况的变化, 语音信号的失真也是变化的. 带有汽车噪声的平均 5 dB 的语音信号做测试, 语音识别系统的词识别率为 73. 6%, 串识别率为 19. 20%. 出现了许多插入与替代错误. 另外在实际的应用环境中, 信号还会有信道失真, 这也严重影响了系统的识别性能. 汽车中的语音识别系统必须作出相应的处理, 否则就无法实用. 在汽车环境下, 语音信号的失真模型如图 1 所示. 图中噪声 1 指前面提到的汽车环境噪声, 噪声 2 指系统电路等引起的噪声, 一般可忽略. $H(\omega)$ 指各种信道失真, 主要是麦克风失真等.

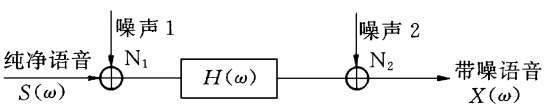

图 1　语音信号失真模型

Fig. 1　Speech signal distortion model

2　带噪语音预处理——语音增强

由于语音识别系统的 HMM 的参数是通过对相对比较干净的语音资料训练得到的, 所以希望从带有汽车噪声的语音信号中提取出“干净”的语音信号, 然后再计算特征向量用于识别. 汽车中的语音识别系统需要对输入系统的带噪语音进行噪声压缩预处理, 即语音增强. 在汽车上的语音识别系统中不宜采用较复杂、昂贵的麦克风设备. 本文采用了一种基于单麦克风的语音增强方案. 单信道的语音增强方法需要估计信号中的噪声谱特性. 最常见的方法是谱相减法, 这种方法简便易行, 但主要的缺陷是增强后语音信号往往含有“音乐”噪声. “音乐”噪声往往会引起识别结果的插入错误; 影响系统的识别率. 而用基于维纳滤波的语音增强方法时, 输出结果中“音乐”噪声就小得多. 但是由于信号失真的存在, 输出信号的自然度仍需要提高.

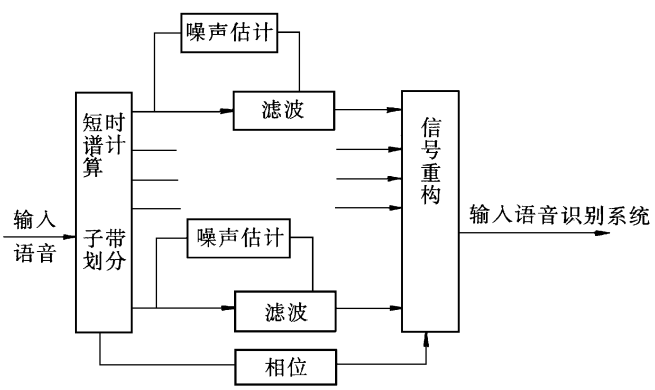

图 2　子带语音增强方案

Fig. 2　Subband speech enhancement block diagram

本文提出了一种子带滤波的语音增强方法, 为了提高语音增强输出结果的自然度, 这种方法首先根据人的听觉特性, 把语音信号

谱分成一些子带. 在每个子带中分别估计噪声特性和滤波, 然后得到增强后的语音. 此方法的框图如图 2 所示, 它的几个关键部分讨论如下:

(1) 子带划分. 实际上, 人的耳蜗把输进来的声音划成了一个个的带限信号, 然后再输入到听觉系统中做进一步处理. 这一现象可以应用到语音增强算法中, 以提高增强效果. 在文献[2]中, 用来模拟耳蜗特性的对数模型为 $F(x) = A(10^{ax} - k)$ Hz, 其中, $A = 165.4$, $a = 2.1$, $k = 0.88$[3]. 根据此式, 把输入信号的频谱分成一个个的频段. 本文采用 256 点 FFT 求频谱, 子带数为 20. 每个子带中噪声谱估计不同, 所以每个子带滤波器的系数也不同. 实验证明, 这样划分子带后, 滤波输出的语音有更高的自然度.

(2) 噪声估计. 单信道的噪声压缩算法都要首先估计信号中的噪声特性. 大多数的噪声估计算法, 通常根据短时能量或其他特征先判断语音段和噪声段, 然后用噪声段的平均来得到噪声谱的估计值, 这种方法叫做话音判决估计法. 但是在汽车噪声环境下, 信号的信噪比较低, 语音与噪声并不容易明显区分. 另外话音判决估计法忽略了对语音帧中噪声的估计. 这里用另外一种噪声的估计方法. 通过对背景汽车噪声的观察和测量, 汽车噪声特性的变化一般慢于语音特性的变化, 利用这一点, 可得到噪声谱的平滑估计. 这一平滑估计通过递归的方法得到, 递归公式为 $\hat{N}_i(k) = (1-\alpha)X_i(k) + \alpha\hat{N}_i(k-1)$. 其中: $X_i(k)$ 表示第 k 帧语音信号在第 i 个子带的幅度谱; $\hat{N}_i(k)$ 表示第 k 帧语音信号中在第 i 个子带的噪声幅度谱估计; 时间常数 $\tau = 1/(1-\alpha)$, 它影响整个算法的性能, 一般 α 取 0.95 左右. 如果语音信号自身的谱变化速率与噪声谱变化可以明显区分, 那么就可以得到噪声谱的较好估计. 这种方法非常简单, 易于实时实现, 并且不需要做语音帧与噪声帧的判决. 另外即使在语音帧也可以进行噪声谱估计值的更新. 所以即使在汽车行驶速度变化时, 即噪声电平变化时, 这种估计方法也能适用.

(3) 维纳滤波系数. 子带中的维纳滤波系数 $G_i(k) = [(|X_i(k)|^2 - |N_i(k)|^2)/|X_i(k)|^2]^{\beta}$, 各个子带噪声特性估计结果不一致, 所以各个子带的维纳滤波系数不同. 子带内部的各个谱分量通过加窗(如三角窗)合并在一起. β 是一个压缩因子, 目的是降低谱的动态范围.

(4) 语音信号重构. 对语音信号而言, 相位信息一般来说不太重要, 因此在信号重构过程, 利用滤波后的幅度谱和带噪相位来恢复“干净”语音信号. 因为在前面的噪声估计中, 并没有考虑语音中的清音信号与噪声信号的相似之处, 所以在压缩噪声时, 同时也压缩了清音信号, 从而引起了语音信号的失真. 另外维纳滤波结果中也可能含有一些具有固定频率的噪声信号. 这些残留噪声引起了增强后语音自然度下降, 也影响了系统识别率. 这里用加入附加噪声的方法来提高输出结果的自然度. 附加小噪声尽管稍稍降低了信噪比, 但是它不仅掩蔽了残留的具有固定频率的噪声, 对清音信号的丢失也有补偿作用. 所以明显提高了重构语音信号的自然度. 加入的附加小噪声 n 采用对数自然分布, 这是因为对数自然分布更接近自然界的背景声音[4]. 加入方式是在频域直接相加. 即 $10\lg n \sim N(\mu, \sigma^2)$.

为了测试这种增强方法的性能, 用带有汽车噪声的语音库做实验, 语音库包括 2 000 句连续发音的数字串, 信噪比平均为 4.97 dB. 这种语音增强方法可以得到较好的信噪比改善, 信噪比平均提高了 9.56 dB. 对一些句子的抽样主观视听, 没有明显的“音乐”噪声. 这种语音增强方法的另一个优点是运算简便、便于实时实现.

3 信道补偿——倒谱均值相减

汽车中的语音识别系统信号往往存在信道失真. 比如在汽车中, 话筒自身的失真、说话者距话筒的远近、话筒的种类不同、环境(如车壁、车窗)对声音的反射等都会使语音信号产生失真. 语音识别技术中, 用于信道失真补偿的算法很多, 如 CDCN、MLLR 等. 但根据汽车中的语音识别系统的具体要求, 不宜采用太复杂的处理算法. 这里采用倒谱均值相减方法[5]. 采用这一方法的根据是引起失真的信道特性随时间变化同语音信号本身的变化相比非常缓慢, 甚至可以认为在一定时间段内, 信道脉冲响应是时不变的.

参照图 1 所示的语音信号失真模型, 噪声 N_2 一般情况下可以忽略, 噪声 N_1 假定在语音增强处理后, 也能够忽略. 这时, 信号的频谱可表示为

$$X(\omega) = S(\omega)H(\omega) \tag{1}$$

前面已经提到, 倒谱系数相当于对信号频谱的对数求逆傅里叶变换得到. 那么在倒谱域, 式(1)变成

$$C^X = C^S + C^H \tag{2}$$

式中: $C^X = \mathrm{IDFT}[\log X(\omega)]$; $C^S = \mathrm{IDFT}[\log S(\omega)]$; $C^H = \mathrm{IDFT}[\log H(\omega)]$.

在一定时间段内, 如若干相邻帧内, 取倒谱向量的时间平均, 即

$$E[C^S] = \frac{1}{M}\sum_{k=1}^{k=M} C_k^S, \quad E[C^H] = \frac{1}{M}\sum_{k=1}^{k=M} C_k^H \tag{3}$$

然后各个帧的倒谱系数减去这平均值, 即倒谱均值相减 可得

$$C_k^{\wedge X} = C_k^X - (E[C^S] + E[C^H]) \tag{4}$$

式中, k 为帧号.

倒谱均值相减算法去掉了倒谱向量中的时不变部分, 从而大大减小了倒谱向量由信道引起的失真. 实验证明, 倒谱均值相减算法可以大大提高识别率.

4　识别结果的后处理和实验结果

汽车上的语音识别系统, 如话音拨号系统, 对识别输出结果的可靠性有一定的要求, 用户不希望由于识别结果错误, 而引起尴尬的拨错号现象. 所以系统在输出识别结果之前, 需要对搜索算法得出结果的可靠性进行判决. 丢弃(Rejection)后处理方法可以提高输出结果的可靠度, 这种方法事先用语音库训练得到一个"Garbage"模型, 然后利用它结合系统语法模型, 并根据识别错误发生概率, 统计得到一个判决似然比门限. 在识别时, 根据这一门限, 确定是否对 Viterbi 搜索结果进行丢弃. 这一处理过程提高了语音识别结果的可靠性, 使系统的实用性得到提高, 避免了汽车中语音识别系统的一些误操作.

实验中用于训练 HMM 参数的训练语音库是干净语音库, 包括 8 500 个连续发音数字串语音文件, 说话者为 54 个男人和 57 个女人. 用于识别测试的语音库有两个. 第一个(SET1)是由 2 000 句数字串连续发音组成的干净语音库, 说话者为不同于训练数据库说话者的另外 56 个男人和 57 个女人. 第二个(SET2)是由 SET1 加入了现场录制的汽车噪声而组成. SET2 的平均信噪比为 4. 97 dB, 与实际汽车噪声环境情况类似. 识别实验的结果如表 1 所示.

表 1　汽车噪声中语音的识别实验结果

Tab 1　The experiment results of speech recognition in car noise

语音库+ 处理方法	词识别率	串识别率
SET1+ None	98. 43	94. 81
SET2+ None	73. 63	19. 20
SET2+ SE+ CMS	93. 15	81. 15

注: SE 表示语音增强, CMS 表示倒谱均值相减

实验表明, 在平均信噪比仅有 5 dB 的汽车噪声中进行语音识别是一项十分困难的任务. 采用子带语音增强和倒谱均值相减处理后, 系统的识别率有很大程度的改善. 但是仅有这两项处理, 识别率还不够. 通过结合其他一些处理方法(如 N-Best 处理), 识别率会提高到 90% 左右, 基本能够满足需求, 但是仍然需要进一步的改进. 需要说明的是, 本文中用到的处理方法都是基于系统的简便、实时实现的考虑, 所以没有采用复杂的环境补偿算法.

致谢　本课题得到了美国贝尔实验室上海分部的大力支持, 尤其得到了贝尔实验室 Sunil K. Gupta 博士的指导.

参　考　文　献

1　Lecomte I, Lever M, Boudy J, *et al*. Car noise processing for speech input. '89 ICASSP, 1989, 1: 512～515

2　Ghitza O. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans on Speech and Audio Processing, 1994, 2(1): 115～132

3　Allen J B. How do humans process and recognize speech? IEEE Trans on Speech and Audio Processing, 1994, 2(4): 567～577

4　Compernolle D. Noise adaptation in a hidden Markov model speech recognition system. Computer Speech and Language, 1989, 3: 151～167

5　Gupta S, Soong F, Haimi-Cohen R. High-accuracy connected digit recognition for mobile application. '96 ICASSP, 1996, 1: 57～60