ISSN 1000-0054 清华大学学报（自然科学版）2001 年 第 41 卷 第 1 期 11/32
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2001, Vol. 41, No. 1 41－44

语音增强用于抗噪声语音识别

徐义芳， 张金杰， 姚开盛， 曹志刚， 王勇前

(清华大学 电子工程系，微波与数字通信国家重点实验室，北京 100084)

摘 要：语音识别系统通常是将在安静的环境下训练得到的参数应用于实际环境中。如果实际环境也是安静的，则语音识别系统可以令人满意地工作。然而，当实际环境中有噪声存在时，语音识别系统性能急剧下降。为了让语音识别系统在安静的环境和有噪声的环境中都获得令人满意的工作性能，研究了一个将语音增强器和语音识别器级连起来的系统。该系统中，语音增强作为前端处理用于提高识别器输入端信号的信噪比。通过 3 种不同的增强算法用于纯净语音和 3 种类型带噪语音的实验结果分析比较表明，这一方法对纯净语音的识别精度几乎没有任何改变而大大提高了系统的抗噪声性能。

关键词：语音增强；谱相减；语音识别

中图分类号：TN 912.3 **文献标识码**：A

文 章 编 号：1000-0054(2001)01-0041-04

Speech enhancement applied to speech recognition in noisy environments

XU Yifang, ZHANG Jinjie, YAO Kaisheng,
CAO Zhigang, WANG Yongqian

(State Key Laboratory on Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Speech recognition systems work in practical environments using parameters train in quiet environments. The system performance is satisfactory when the environment is also quiet, but degrades quickly in noisy environments. This paper presents a noisy speech recognition system with added background noise so that the recognition system can work well both in quiet and noisy environments. Speech enhancement as the front-end processing module is used to improve the Signal-to-Noise Ratio (SNR) of the input signal for recognition in the latter stages. Three different speech enhancement algorithms were tested with six types of noise. Experimental results show that the cascading of the speech enhancer and a Hidden Markov Model (HMM) based speech recognizer can significantly improve recognition accuracy in noisy environments without performance degradation for clean speech.

Key words: speech enhancement; spectral subtraction; speech recognition

目前的语音识别系统一般都是基于隐马尔可夫模型(hidden markov model, HMM)的，其中模型参数是在一定的环境下采集的语音库训练得到。实验表明，系统在应用环境和训练环境匹配的情况下可以令人满意的工作，然而当识别、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。最常见的一种情形是将安静环境下训练的模型应用于实际有背景噪声的环境中。噪声背景环境中的语音识别技术长期以来一直受到人们的关注。研究了一个将语音增强和语音识别级连起来的抗噪声语音识别系统。该系统在前端的语音增强模块中，采用了 3 种语音增强算法提高语音识别模块输入端信号的信噪比。

1 语音增强算法和基本识别平台的构筑

将语音增强和语音识别级连起来的抗噪声语音识别系统原理框图如图 1 所示。图中 $y(n)$ 是带噪语音，$\hat{s}(n)$ 表示经过语音增强得到的估计语音，W 是识别器的输出。语音增强作为该级连系统的前端模块。采用了两种不同的谱相减法和对数谱最小均方误差估计法。

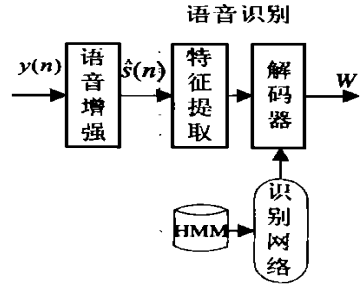


图 1 抗噪声语音识别系统框图

1.1 语音增强算法

1) 谱相减法

收稿日期：2000-02-17
作者简介：徐义芳(1975-)，女(汉)，湖北，硕士研究生。

谱相减法[1]的基本框图如图 2 所示。图中 $s(n)$ 表示纯净语音, $d(n)$ 表示加性噪声, $\lambda_n(k)$ 表示噪声功率谱系数, 通常在语音中的无声段估计而得。Y_k 和 S_k, $k=0,1,\cdots$ 分别表示带噪语音 $y(n)$ 和纯净语音 $s(n)$ 的频谱系数。

$$y(n)=s(n)+d(n)$$

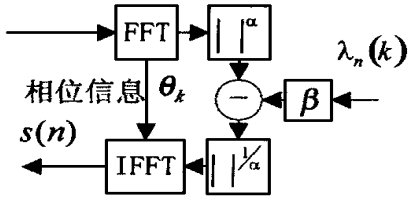


图 2 减谱法框图

增强后的语音 $\hat{s}(n)$ 的谱幅度系数 $|\hat{S}_k|$ 由下式得

$$|\hat{S}_k|=[\,|Y_k|^{\alpha}-\beta\lambda_n^{\alpha}(k)\,]^{1/\alpha}. \tag{1}$$

其中 α 和 $\beta(\beta>=1)$ 是两个参数。当 $\alpha=2$, $\beta=1$ 时, 这个算法就是传统的谱相减法。适当调整 α 和 β 的值可以得到更好的增强效果。实验表明在 $\alpha=2$, $\beta=5$ 时的增强效果大大优于传统的谱相减法。在第 3 部分的实验结果中, 称这种算法为改进谱相减法。

2) 对数谱最小均方误差估计法

如果让 A_k 表示纯净语音的谱幅度系数, 那么 A_k 的最小均方误差估计由下式给出

$$\hat{A}_k=\arg\min\{(A_k-\hat{A}_k)^2\,|y(n),\ 0\le n\le N-1\}=E\{A_k\,|Y_0,Y_1,\cdots\}. \tag{2}$$

考虑到人耳对频谱强度的感应与幅度对数成正比, 引进对数谱最小均方误差估计[2]

$$\hat{A}_k=\exp(E\{\ln(A_k)\,|Y_k\}). \tag{3}$$

1.2 语音识别算法

1) 特征参数提取

我们的识别器采用 Mel 频率倒谱系数(mel-frequency cepstral coefficients, MFCC)[3]加上它的一阶差分向量 ΔMFCC 作为特征参数。MFCC 的提取过程如图 3 所示。

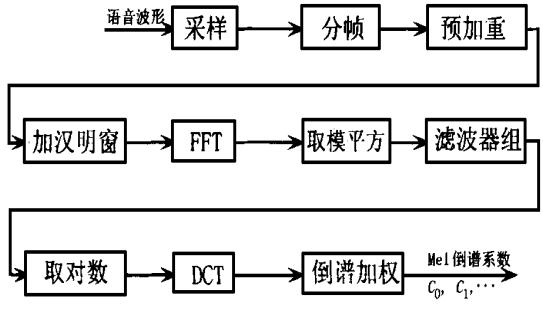


图 3 MFCC 的提取过程

模拟语音信号经过 16 kHz 采样数字化, 分帧, 预加重, 加窗, 快速付立叶变换(fast fourier transform, FFT), 得到频谱能量系数, 再通过一组 Mel 尺度的三角形滤波器组, 求得该滤波器组能量系数取对数再经过离散余弦变换(discrete cosine transform, DCT), 取前 13 维系数, 将这 13 维系数加权得到MFCC。所有帧的MFCC 求得以后, 再计算一阶差分MFCC。这样一来, 每帧语音由一个 26 维的观测向量(13 维 MFCC 和 13 维 ΔMFCC)表示。

2) 隐含马尔可夫模型 HMM

一个 HMM 由许多状态和状态之间的转移弧组成。识别器中, 采用 10 个状态的从左到右 HMM, 如图 4 所示。起始状态必然以概率 1 跳转到别的状态和终止状态是一个吸收态, a_{ij} 表示任意一对状态 i, j 之间的转移概率, 转移概率矩阵为 10 行 10 列, 除起始状态和终止状态外, 每个状态 j 有一个与之相关联的观测向量概率密度分布函数 $b_j(O_t)$, 它表示在 t 时刻产生观测向量 O_t 的概率密度, 由 4 个高斯分布概率密度函数线性叠加而成。这种情况下, $b_j(O_t)$ 由下式给出

$$b_j(O_t)=\sum_{m=1}^{M_j}c_{jm}N(O_t;\ u_{jm},\ \Sigma_{jm}). \tag{4}$$

这里, c_{jm} 表示第 m 个高斯分布函数的权重, $N(\cdot;u,\Sigma)$ 表示均值向量为 u, 方差矩阵 Σ 为的多维高斯分布函数。我们的识别器中, O_t 是 26 维向量, $N(\cdot;u,\Sigma)$ 是 26 维的高斯分布函数。

3) 识别网络

识别网络的功能是向解码器提供一个搜索空间, 识别器的输出结果一定是识别网络中的一条路径。识别网络由一系列节点和转移弧组成。在连接词识别系统中, 识别网络如图 5 所示。每个节点代表一个 HMM 模型, 各个 HMM 本身又是由状态和转移弧连接而成。这样一来, 识别网络最终可看作是由一系列的 HMM 状态和转移弧连接而成。图中"1","2",…,"0"表示各个数字, 它代表的 HMM 如图 4 所示; 无声段的节点代表语音起始段和终止段的静音模型, 与表示数字的 HMM 相比较, 无声段的模型较简单, 它由 3 个状态组成(也包括一个起始状态和一个终止状态), 与第 2 个状态相关联的观测向量概率密度分布函数是单个高斯分布函数。

4) 解码器

对于一段输入语音, 每条从识别网络起始节点到终止节点的路径都是一个可能的识别结果。解码

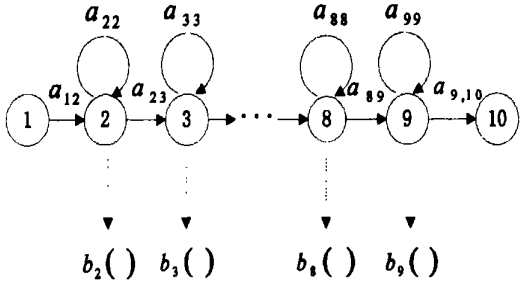


图 4 隐含马尔可夫模型

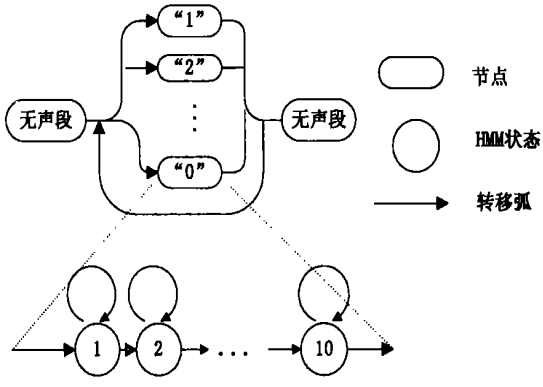


图 5 连接词识别网络

器的功能就是要从这些路径中找到一条最可能产生这段输入语音的路径作为识别器的输出。衡量最可能路径的标准是看该路径产生这段语音的观测向量序列的对数似然值是否最大。路径的对数似然值定义为这条路径经过的所有转移弧的对数似然值与它经过的所有 HMM 状态相关联的观测向量的对数概率密度之和。为了找到识别网络中的最佳路径, 解码器采用 token passing 算法[4]。起始时刻, 所有可能的起始节点都初始化一个 token。每经过一个时刻, 各个 token 沿着与它相连的转移弧延伸直到下一个具有输出概率密度函数的HMM 状态。当一个状态或节点有许多个出口时, token 沿着所有的出口延伸过去, 这样保证 token 经历所有可能的路径。当 token 穿过节点和转移弧时, 它的对数似然值随之增加。Token 在网络中游走的同时保存它历经路径的历史记录。最后在输出节点中找到对数似然值最大的那个 token, 根据它的历史记录可以回朔找到这个 token 曾经历过的所有节点和转移弧, 输出识别结果。

2 实验结果与分析

用图 2 的级连系统进行了一系列的非特定人数字串识别的实验。实验中用于训练HMM 参数的训练语音库是来自 TI-Digits 的纯净语音, 包括 16 个男子说的 500 个连续发音数字串语音文件。识别测试的语音库由 TI-Digits 中不同于训练库中的 4 个男子说的 100 个连续发音数字串组成, 系统共有 12 个 HMM, 其中 11 个代表 10 个数字("0"有两个 HMM 与它对应), 1 个代表无声段静音模型。噪声来自Noisex92。

用于测试的带噪语音由测试库的纯净语音加上噪声组成。信噪比

$$r = 10\lg\left(\frac{\sum_{i=1}^{L}\sigma_{s_i}^2}{\sigma_n^2}\right). \qquad (4)$$

其中: $\sigma_{s_i}^2$ 表示第 i 个语音信号的功率, σ_n^2 是噪声功率, L 是测试语音文件的数目。识别结果以识别精度来衡量。识别精度为

$$E = \left(1 - \frac{S + D + I}{N}\right) \times 100\%. \qquad (5)$$

其中: S 表示替换错误单词数, D 表示漏识错误单词数, I 表示插入错误单词数, N 表示测试数据单词的总数。例如, 如果数字串"123"被识别成了"1345", 则"2"是漏识错误, D 增加 1, "4"和"5"是插入错误, I 增加 2, 没有替换错误, S 不变; 若"123"被识别成了"124", 则"4"是替换错误, S 增加 1, 没有漏识错误和插入错误, D, I 不变。

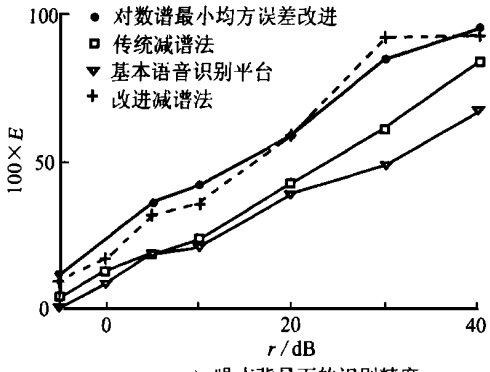
为了便于比较并考察级连系统对纯净语音的损伤程度, 同时给出了纯净语音通过这个级连系统的结果, 如表 1 所示。图 6a～6c 分别给出了系统在白噪声, 汽车噪声, 人群讲话噪声, 飞机噪声, 工厂噪声和坦克噪声下的实验结果。表 1 中的第 1 行"基本语音识别平台"表示语音只通过语音识别模块, 没有经过任何增强处理的结果, 第 2 行至第 4 行表示语音通过图 2 所示的级连系统的识别结果, 各行的不同之处在于语音增强模块中采用了不同的算法, 如表中所示。

表 1 纯净语音的识别精度

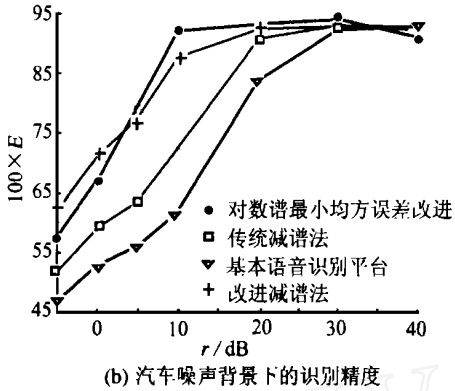
方法	识别精度 E/%
基本语音识别平台	92.3
传统减谱法	92.3
改进减谱法	92.7
对数谱最小均方误差估计	91.3

实验结果说明:

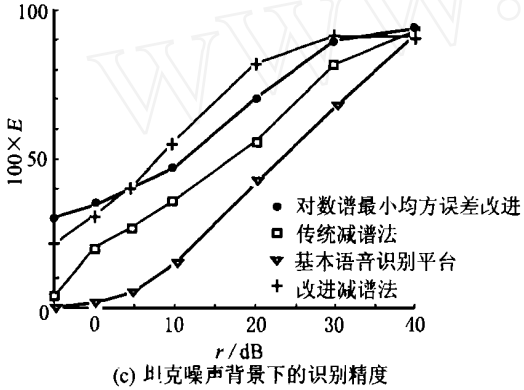
1) 这个系统对纯净语音的识别精度并没有下降。从表 1 可以看出, 纯净语音在不通过任何增强处理只经过识别器和通过增强再通过识别器的识别结



(a) 噪声背景下的识别精度



(b) 汽车噪声背景下的识别精度



(c) 坦克噪声背景下的识别精度

图 6 各种噪声背景下的识别结果

果几乎相同, 说明前端的语音增强模块并没有降低系统在安静环境下的工作性能。

2) 加性噪声确实大大降低了系统的性能, 尤其是宽带高斯白噪声。即使在信噪比为 40 dB 的情况下, 系统在不做任何增强处理时的识别精度只有 67%, 见图 6a。

3) 语音增强对提高系统的抗噪声性能是有效的。例如, 在汽车噪声环境下, 当识别精度在 90% 以上时, 系统至少能获得 10 dB 的信噪比增益, 见图 6b。

4) 对数谱最小均方误差估计和改进的减谱法比传统减谱法性能要好的多, 见图 6a～6c。

5) 信噪比在 5 dB 到 20 dB 这一段内, 识别精度提高的幅度较大, 见图 6a～6c。这一段信噪比也是实际中经常遇到的情形, 所以这种方法对实用很有效。

3 结 论

研究了一个抗噪声语音识别系统。这个系统中, 将语音增强作为前端处理以提高识别器输入端信号的信噪比。实验结果表明, 将语音增强用于抗噪声语音识别是有效的。

参考文献 (References)

[1] Lim J S, Oppenheim A V. Enhancement and bandwidth compression of noisy speech [J]. Proc of IEEE, 1979, 67(12): 1586-1604.

[2] 曹志刚, 郑文涛. 基于短时谱最小均方误差估计的语音增强和剩余噪声衰减 [J]. 电子学报, 1993, 21(4): 7-12.
CAO Zhigang, ZHENG Wentao. Speech enhancement based on minimum mean-square error short-time spectral estimation and residual noise reduction [J]. Acta Electronica Sinica, 1993, 21(4): 7-12. (in Chinese)

[3] Young S J. The HTK Book [M]. Version 2.1, 1997. 72-75. http://svr-www.eng.cam.ac.uk.

[4] Young S J, Russell N H, Thornton J H S. Token passing: a simple conceptual model for connected speech recognition systems [EB/OL]. http://svr-www.eng.cam.ac.uk. 1989-06.