一种基于噪声对消与倒谱均值相减的鲁棒语音识别方法

王振力[1]，裴凌波[2]，于元斌[1]
（1.南京国际关系学院 博士后流动站，江苏 南京 210039；2.工程兵指挥学院 训练部，江苏 徐州 221004）

摘 要：提出一种基于语音增强算法的噪声鲁棒语音识别方法.在语音识别预处理阶段，通过噪声对消语音增强法来抑制噪声提高信噪比.然后对增强语音提取Mel频段倒谱特征参数，并在倒谱域应用倒谱均值相减处理来补偿增强语音中的失真成分和剩余噪声.实验结果表明，在低信噪比（-12～0 dB）条件下，该方法对于数字语音识别具有较好的识别率，其性能明显优于基本的Mel频段倒谱参数识别器、传统的谱减法和噪声对消语音增强法.

关键词：自适应噪声对消；语音增强；谱减法；噪声鲁棒语音识别；倒谱均值相减法
中图分类号：TN912.34 **文献标识码**：A **文章编号**：1673-4785(2008)06-0552-05

A robust speech recognition method by combining noise cancelling and cepstral mean subtraction

WANG Zhen-li[1], PEI Ling-bo[2], YU Yuan-bin[1]
(1. Postdoctoral Station, Nanjing University of International Relations, Nanjing 210039, China; 2. Training Department, The Command Academy of Engineer Corps, Xuzhou 221004, China)

Abstract: A noise resistant speech recognition method based on a speech enhancement algorithm was implemented. First, it obtains the denoised speech, with significant SNR (signal-to-noise ratio) improvement, by applying adaptive noise cancelling (ANC) to the pre-treatment stage of speech recognition. Then Mel-frequency cepstral coefficients (MFCC) are computed from the enhanced speech. Then cepstral mean subtraction (CMS) is used to compensate for components of distortion and the residual noise of the enhanced speech in the cepstral domain. When speech samples have a low SNR, ranging from 0 to 12 dB, experimental results indicate that the proposed method performs better than a standard MFCC recognizer, conventional spectral subtraction (SS) and the ANC speech enhancement for digital speech recognition.

Keywords: adaptive noise cancelling; speech enhancement; spectral subtraction; noise robust speech recognition; cepstral mean subtraction

当前，噪声鲁棒性是语音识别技术实用化过程中最迫切需要解决的关键问题之一.提高语音识别系统在噪声环境中的识别性能，常见的方法是用语音增强在前端预处理中来消除噪声.语音增强[1-4]虽然可以抑制带噪语音信号中的噪声成分，但同时也带来了语音失真和残余噪声；使得测试环境和训练环境产生失配，最终恶化了语音识别系统性能.如果单纯地采用抗噪语音特征如线性预测倒谱参数（linear prediction cepstral coefficients，LPCC）[5]、Mel频段倒谱参数（Mel-frequency cepstral coefficients，MFCC）[6]、RASTA（relative spectra）[7]和PLP（perceptual linear predictive）[8]等，在一定程度上能提高识别系统的噪声鲁棒性；但在低信噪比条件下很难遏制系统性能的恶化.基于时不变线性信道假设的倒谱均值相减法（cepstral mean subtraction，CMS）[9]

收稿日期：2008-03-06.
基金项目：江苏省博士后科研基金资助项目(0701008C)；中国博士后科学基金资助项目(20070420561).
通信作者：王振力. E-mail：dowen3619@sina.com.

是一种常用的信道补偿方法,主要用于消除信道卷积噪声和加性噪声对识别特征在倒谱域造成的偏差;但其改善系统的性能有限. 本文提出将一种噪声对消语音增强算法用于语音识别系统的前端预处理,接着对增强语音计算 MFCC 参数,然后应用 CMS 方法消除语音增强带来的语音失真和残余噪声,最后在低信噪比条件下研究了数字语音识别系统的性能.

1　自适应噪声对消语音增强算法

当环境噪声的统计特性未知,并且可能不断发生变化时,一般采用自适应噪声对消技术来完成对带噪语音的增强处理. 噪声对消器通常将麦克风采集的未知噪声信号输入到自适应滤波器中,通过滤波器自适应地调节其权系数,按照均方误差最小(minimum mean square error, MMSE)的准则,以尽可能地接近主信道带噪语音中的噪声成分并从中减去. 图 1 为自适应噪声对消器的原理框图. 设参考信道接收到干扰 n_1,由于传输路径不同,n_0(主信道接收噪声)和 n_1 是不同的. 但因二者都来自同样的噪声源,所以它们是相关的.

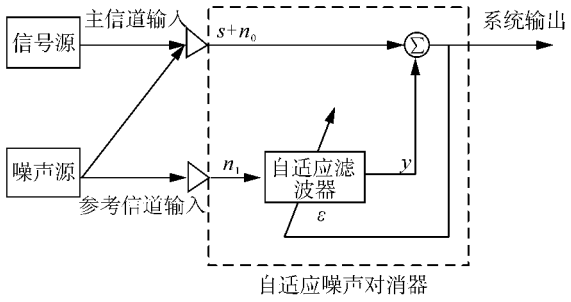


图 1　自适应噪声对消原理

Fig. 1　The principle of adaptive noise cancelling

在图 1 中,主信道输入的带噪语音即为自适应滤波器的期望信号 d($d = s + n_0$,s 表示主信道接收到的语音信号),系统输出则为误差信号 ε,则

$$\varepsilon(n) = d(n) - y. \tag{1}$$

假设参考信道中滤波器 n 时刻输入矢量 $\boldsymbol{X}(n) = [n_1(n), n_1(n-1), \cdots, n_1(n-M+1)]^{\mathrm{H}}$($M$ 表示滤波器阶数,H 表示共轭转置),对应的权系数矢量为 $\hat{\boldsymbol{W}}(n) = [w(n), w(n-1), \cdots, w(n-M+1)]^{\mathrm{H}}$,将自适应滤波器输出 $y = \hat{\boldsymbol{W}}^{\mathrm{H}}(n)\boldsymbol{X}(n)$ 代入式(1)得

$$\varepsilon(n) = d(n) - \hat{\boldsymbol{W}}^{\mathrm{H}}(n)\boldsymbol{X}(n). \tag{2}$$

从矢量 $\boldsymbol{X}(n)$ 中去除与 $\boldsymbol{X}(n-1)$ 相关的部分,并定义为新矢量:

$$\boldsymbol{U}(n) = \boldsymbol{X}(n) - \gamma(n)\boldsymbol{X}(n-1). \tag{3}$$

其中:$\gamma(n) = \dfrac{\boldsymbol{X}^{\mathrm{H}}(n)\boldsymbol{X}(n-1)}{\boldsymbol{X}^{\mathrm{H}}(n-1)\boldsymbol{X}(n-1)}$,表示输入矢量 $\boldsymbol{X}(n)$ 在 n 和 $n-1$ 时刻的相关系数. 将矢量 $\boldsymbol{U}(n)$ 看作为滤波器 n 时刻的输入,可得到自适应噪声对消语音增强算法权系数更新公式[10]为

$$\hat{\boldsymbol{W}}(n+1) = \hat{\boldsymbol{W}}(n) + \frac{\tilde{\mu}}{\| \boldsymbol{U}(n) \|^2 + \delta}\boldsymbol{U}(n)\varepsilon^*(n). \tag{4}$$

其中:$\tilde{\mu}$ 为自适应常数;$\delta > 0$ 为实数,用于控制算法的稳态失调大小和收敛速度的快慢. 文献[10]给出了该算法的证明,即当迭代次数无限增加时,滤波器的权系数均值趋向于广义 Wiener-Hopf 解,如式(5)所示.

$$\begin{aligned}\lim_{n\to\infty} E\{\hat{\boldsymbol{W}}(n)\} &= \boldsymbol{Q}\lim_{n\to\infty}(\boldsymbol{I}-\Lambda)^n\boldsymbol{Q}^{\mathrm{T}}E\{\hat{\boldsymbol{W}}(0)\} + \\ &\bar{\boldsymbol{P}}\lim_{n\to\infty}\sum_{i=0}^{n-1}\boldsymbol{Q}(\boldsymbol{I}-\Lambda)^i\boldsymbol{Q}^{\mathrm{T}} = \\ &\bar{\boldsymbol{P}}\boldsymbol{Q}\sum_{i=1}^{\infty}(\boldsymbol{I}-\Lambda)^i\boldsymbol{Q}^{\mathrm{T}} = \\ &\bar{\boldsymbol{P}}\boldsymbol{Q}\Lambda^{-1}\boldsymbol{Q}^{\mathrm{T}} = \bar{\boldsymbol{P}}(\bar{\boldsymbol{R}}_U)^{-1} = \hat{\boldsymbol{W}}_{\mathrm{opt}}.\end{aligned} \tag{5}$$

对于实际的 ANC 系统,主信道、参考信道的两个麦克风之间的距离都是固定的. 因此,可以通过对参考信号进行固定时延处理,很容易解决信道间信号的时间同步问题. 考虑到语音识别系统的实时性,对噪声进行统计平均的次数取为 1. 这样虽然噪声对消的性能下降了,但是满足了系统的实时性要求. 由于统计次数为最低,使得自适应噪声对消器权值估计 $\hat{\boldsymbol{W}}(n)$ 对随机性噪声的变化不能够进行统计意义上的自适应更新;故输出增强语音中不但包含了未彻底对消的剩余噪声,而且增强语音也发生了失真. 此外,参考信道接收的噪声信号中往往也会混有少量的信号源发出的语音信号;因此,在进行噪声对消时会部分地对消有用的语音信号,导致增强语音信号的频谱畸变,最终恶化了识别系统的性能.

通常,纯净语音信号在倒谱域的均值接近于 0[9],那么带噪语音信号倒谱特征均值主要表现为语音信号在信道传输过程中信道卷积噪声和加性噪声在倒谱域的均值,这种非期望均值使得识别特征与纯净训练语音的倒谱特征产生了失配. 在 ANC 方法中,作者将上述情况下的残留噪声和语音失真均假定为信道卷积噪声和加性噪声对语音信号的影

响. 对这种混合残留噪声和失真信号的增强语音在倒谱域采用 CMS 处理,可以使其在倒谱域的均值近似为 0,从而消除了残留噪声和失真语音在倒谱域对特征提取的影响. 为了取得更好的识别率,必须对纯净语音的倒谱特征训练也加入 CMS 处理,以获得训练特征和识别特征更好的匹配性. 后续的实验结果进一步表明了 CMS 处理可以在语音特征层较好地消除前端预处理残留噪声及降低畸变信号对后续倒谱域特征提取的影响.

2 倒谱均值相减法

在缓慢变化的噪声环境中,CMS 对于消除信道的卷积和加性噪声所造成的训练和测试识别特征间的失配,是一种非常简单有效的方法. CMS 方法通过在倒谱域减去估计的信道噪声均值,使得带噪语音特征尽可能接近于 0,从而消除信道的不利影响. 假设 $\boldsymbol{D}_t$ 表示带噪语音信号的倒谱特征矢量,$\boldsymbol{m}_t$ 表示倒谱均值矢量,其中 t 表示采样时刻. 考虑到识别系统的实时性,作者根据迭代计算法[11]来求取倒谱均值 $\boldsymbol{m}_t$.

对于最初的 N 个特征矢量(N 即为用于计算倒谱均值的窗宽),完成 $\boldsymbol{m}_t$ 和标准方差的初始化,如式(6)所示.

$$\begin{cases} \boldsymbol{m}_t(i) = \dfrac{1}{N}\sum_{t=1}^{N}\boldsymbol{D}_t(i); \\ \boldsymbol{\sigma}_t(i) = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}[\boldsymbol{D}_t^2(i)] - [\boldsymbol{m}_t(i)]^2} = \\ \qquad \sqrt{\overline{\boldsymbol{s}_t^2(t)} - [\boldsymbol{m}_t(i)]^2}. \end{cases} \tag{6}$$

对于后续的特征矢量,向前滑动窗宽 N,根据式(7)更新 $\boldsymbol{m}_t$ 和采样均方估计.

$$\begin{cases} \boldsymbol{m}_t(i) = \lambda \cdot \boldsymbol{m}_{t-1}(i) + (1-\lambda)\cdot \boldsymbol{D}_t(i); \\ \overline{\boldsymbol{s}_t^2(i)} = \lambda \cdot \overline{\boldsymbol{s}_{t-1}^2(i)} + (1-\lambda)\cdot \boldsymbol{D}_t^2(i). \end{cases} \tag{7}$$

其中 λ 是更新步长,λ 与 N 的关系如式(8)所示.

$$1 - \lambda^N = \frac{1}{\sqrt{2}}. \tag{8}$$

在窗宽 N 内的特征矢量根据式(9)减去倒谱均值,同时完成归一化:

$$\hat{\boldsymbol{D}}_t(i) = \frac{\boldsymbol{D}_t(i) - \boldsymbol{m}_t(i)}{\boldsymbol{\sigma}_t(i)}. \tag{9}$$

3 实验结果及分析

为了研究比较本文方法、传统谱减法和基本 MFCC 识别器的性能,实验中所用的语音数据为汉语普通话数字 0 ~9 的发音语句,每个数字被朗读了 41 次,前 20 次数据用于训练,后 21 次数据用于识别. 语音的采样频率为 8kHz,帧长为 256 点(32ms),帧移为 80 点(10ms),加窗函数为 Hamming 窗,量化精度为 16 位,3 种方法中识别特征均为 MFCC 及其一阶差分参数 ΔMFCC,两者维数都是 12. 噪声来自 Noisex-92 数据库,识别系统中包含 4 个隐 Markov 模型(hidden Markov model,HMM),由左至右无跳转,每个状态由 2 个高斯正态分布叠加而成. 在语音识别预处理阶段,噪声对消语音增强算法(adaptive noise cancelling,ANC)的参数设置为:$\bar{\mu}=1$,$M=12$,$\delta=0.01$,噪声统计平均数为 1 次. 同时假定其参考信号中无串音影响,这样使得该方法抑制噪声的性能尽可能的好;在识别阶段,CMS 处理中的窗宽 $N=20$,更新步长 $\lambda=0.94$.

表 1 ~4 分别给出了数字语音识别系统分别在白噪声、驱逐舰机舱噪声、F16 噪声和 Babble 噪声环境中的识别精度. 由实验数据可知,本文方法(MFCC + ANC + CMS)比未经 CMS 处理的噪声对消语音增强法(MFCC + ANC)性能优异,这主要得益于 CMS 可以部分地消除增强语音中的畸变成分和剩余噪声. 在低信噪比($\gamma_{SNR} \leqslant 0$ dB)条件下,无论参考信道接收的噪声信号中是否混入少量串音(即信号源发出的语音信号),本文方法与基本的 MFCC 识别器、谱减法和噪声对消语音增强法相比,仍然具有良好的识别性能. 当参考信号无串音且 $\gamma_{SNR} \leqslant 0$ dB时,本文方法对应的平均识别率比基本的 MFCC 识别器、谱减法和噪声对消语音增强法在白噪声环境中分别提高了 55.24%、48.41% 和 20.64%;在驱逐舰机舱噪声环境中分别提高了 42.86%、28.25% 和10.63%;在 F16 噪声环境中分别提高了 55.39%、44.60% 和21.27%;在 Babble 噪声环境中分别提高了 52.54%、39.05% 和 14.76%. 当 $\gamma_{SNR} \leqslant 0$ dB时,对传统的谱减法在倒谱域采用 CMS 处理(MFCC + SS + CMS),在 4 种噪声环境中其识别性能仍然明显劣于MFCC + ANC + CMS. 以上数据或者结论说明了在所测试的噪声环境中,本文方法具有比基本的MFCC识别器、谱减法和噪声对消语音增强法更好的抗噪性能,因而更适合于低信噪比噪声环境中的语音识别.

应当说明的是,本文方法与传统的谱减法相比,

二者都具有实时性;但是由于本文方法是基于双信　　的谱减法有所增加.
道噪声对消系统,因而硬件实现的成本会比单信道

表 1　白噪声环境下的语音识别率

Table 1　Speech recognition accuracy under white Gaussian noise background

γ_{SNR}/dB	MFCC/%	MFCC + SS/%	MFCC + SS + CMS/%	MFCC + ANC (无串音)/%	MFCC + ANC + CMS/%		
					无串音	5% 串音	10% 串音
-12	10.00	18.57	16.67	40.48	65.24	65.71	66.19
-6	13.81	22.86	26.19	53.33	70.95	70.48	68.57
0	29.52	32.38	35.61	63.33	82.86	80.00	75.71
6	42.38	52.38	54.28	77.14	89.05	88.10	81.43
12	54.29	71.90	83.33	87.62	91.90	83.81	61.90

表 2　驱逐舰机舱噪声环境下的语音识别率

Table 2　Speech recognition accuracy under destroyer engine room noise background

γ_{SNR}/dB	MFCC/%	MFCC + SS/%	MFCC + SS + CMS/%	MFCC + ANC (无串音)/%	MFCC + ANC + CMS/%		
					无串音	5% 串音	10% 串音
-12	16.67	21.43	22.38	42.86	55.24	54.76	52.86
-6	20.48	32.86	30.48	55.24	65.24	62.38	60.00
0	29.52	56.19	60.48	65.24	74.76	71.43	66.19
6	38.10	76.19	86.19	84.76	90.95	85.71	76.19
12	64.76	79.52	88.09	90.95	93.33	87.62	79.05

表 3　F16 噪声环境下的的语音识别率

Table 3　Speech recognition accuracy under F - 16 noise background

γ_{SNR}/dB	MFCC/%	MFCC + SS/%	MFCC + SS + CMS/%	MFCC + ANC (无串音)/%	MFCC + ANC + CMS/%		
					无串音	5% 串音	10% 串音
12	10.00	13.81	14.76	36.67	57.14	57.62	59.05
-6	10.48	20.95	23.81	45.71	69.52	68.10	64.76
0	20.95	39.05	44.76	61.43	80.95	76.67	76.19
6	33.81	62.38	70.00	82.86	91.43	89.52	83.81
12	63.81	75.24	89.53	90.48	93.33	89.05	79.52

表 4　Babble 噪声环境下的的语音识别率

Table 4　Speech recognition accuracy under Babble noise background

γ_{SNR}/dB	MFCC/%	MFCC + SS/%	MFCC + SS + CMS/%	MFCC + ANC (无串音)/%	MFCC + ANC + CMS/%		
					无串音	5% 串音	10% 串音
-12	10.00	14.76	15.24	44.29	49.52	48.10	45.24
-6	12.38	25.71	24.76	51.90	69.52	59.52	56.19
0	20.00	42.38	43.81	59.52	80.95	69.52	64.76
6	42.86	65.71	75.71	73.81	91.43	81.90	77.14
12	70.95	80.48	87.15	93.33	93.33	87.14	80.48

4 结束语

本文研究了一种噪声对消语音增强算法在语音识别前端预处理中的应用,并通过CMS处理来消除语音增强过程中所产生的语音失真和剩余噪声.研究表明:当信噪比不大于零分贝时,在4种所测试的噪声环境中,无论参考信道接收的噪声信号是否混入串音信号,本文方法均具有比基本的MFCC识别器、谱减法和噪声对消语音增强法更高的识别率,这主要得益于本文方法中CMS处理在一定程度上避免了畸变信号在倒谱域对特征提取的不利影响,从而可以更好地满足低信噪比噪声环境中较高数字语音识别率的需要.本文方法可否较好地提高连续大词汇量语音识别系统在低信噪比条件下的性能,这是下一步的研究工作.

参考文献:

[1]STEVEN F B. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Trans on Speech and Audio Processing, 1979, 27(2): 113-120.

[2]徐义芳, 张金杰, 姚开盛,等. 语音增强用于抗噪声语音识别[J]. 清华大学学报:自然科学版, 2001, 41(1): 41-44.
XU Yifang, ZHANG Jinjie, YAO Kaisheng, et al. Speech enhancement applied to speech recognition in noisy environments[J]. Journal of Tsinghua University: Science and Technology, 2001, 41(1): 41-44.

[3]丁 沛,曹志刚. 基于语音增强失真补偿的抗噪声语音识别技术[J]. 中文信息学报, 2004, 18(5): 64-69.
DING Pei, CAO Zhigang. Robust speech recognition based on the compensation of speech enhancement distortion [J]. Journal of Chinese Information Processing, 2004, 18(5): 64-69.

[4]王振力,张雄伟,郑 翔, 等. 一种新的子波域语音增强方法[J]. 信号处理, 2006, 22(3): 325-328.
WANG Zhenli, ZHANG Xiongwei, ZHENG Xiang, et al. A new wavelet domain speech enhancement method [J]. Signal Processing, 2006, 22(3): 325-328.

[5]MAMMONE R J,ZHANG Xiaoyu,RAMACHANDRAN R P. Robust speaker recognition: a feature-based approach [J]. IEEE Signal Processing Magazine, 1996, 13(5): 58.

[6]DAVIS S B, MERMELSTEIN P. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences [J]. IEEE Trans on Speech and Audio Processing, 1980, 28(4): 357-366.

[7]HERMANSKY H, MORGAN N. RASTA processing of speech [J]. IEEE Trans on Speech and Audio Processing, 1994, 2(4): 578-589.

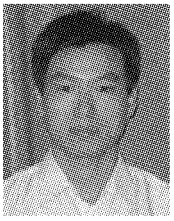
[8]HERMANSKY H. Perceptual linear predictive (PLP) analysis of speech [J]. J Acoust Soc Am,1990, 87(4): 1738-1752.

[9]LIU F H, ACERO A, STERN R. Efficient joint compensation of speech for the effects of additive noise and linear filtering[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. San Francisco, USA, 1992 (1): 257-260.

[10]王振力,张雄伟,杨吉斌, 等. 一种新的快速自适应滤波算法的研究[J]. 通信学报, 2005, 26(11): 1-6.
WANG Zhenli, ZHANG Xiongwei, YANG Jibin, et al. Study of a new fast adaptive filtering algorithm [J]. Journal of China Institute of Communications, 2005, 26(11): 1-6.

[11]VIILDU O, BYE D, IAURILA K. A recursive feature vector normalization approach for robust speech recognition in noise [C]//Proceedings ICASSP 98. Seattle, WA, USA: IEEE Acoustics, Speech and Signal Processing Society, 1998: 733-736.

作者简介:



王振力,男,1977年生,工程师,博士后,主要研究方向为人工智能、多媒体信息处理等.发表学术论文20余篇,被SCI、EI、ISTP收录10余篇.

裴凌波,男,1972年生,讲师,主要研究方向为网络测量、网络性能建模和智能化信息检索等.发表学术论文20余篇.

于元斌,男,1973年生,讲师,博士后.主要研究方向为作战指挥.发表学术论文10余篇,出版专著3部.