一种基于实时噪声估计的改进谱减法

程 塬 郭 雷 贺 胜 赵天云
(西北工业大学自动化学院 西安 710072)

摘 要 针对非平稳噪声环境和低信噪比下的语音增强,提出了一种基于实时噪声估计的改进谱减法。该方法首先利用临界带特征矢量距离进行端点检测,然后利用低频区和高频区带噪语音特性定义一个时变的调节系数,该系数结合端点检测可以实时地对噪声的估计值进行更新,从而达到快速跟踪外界环境变化的目的。仿真结果表明,该方法在抑制背景噪声、提高信噪比、减少语音失真等方面优于传统的语音增强方法。

关键词 噪声估计,端点检测,信噪比,谱减法,语音增强
中图法分类号 TP391 **文献标识码** A

Improved Spectral Subtraction Based on Real-time Noise Estimation

CHENG Gong GUO Lei HE Sheng ZHAO Tian-yun
(College of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Aiming at speech enhancement under non-stationary noise environment and low SNR, an improved spectral subtraction based on real-time noise estimation was proposed. First, voice activity detection was carried out based on selected sub-bands vector distance, and then a real-time adjustment coefficient was defined through noisy speech properties of low-frequency regions and high-frequency regions. The combination of coefficient and voice activity detection could realize the updating of noise estimate. The method allowed tracking the time variation of noise environment. The experiment showed that the method was more effective in reducing background noise, improving SNR and decreasing speech distortion than traditional speech enhancement methods.

Keywords Noise estimation, Voice activity detection, Signal to noise ratio (SNR), Spectral subtraction, Speech enhancement

谱减法以其算法简单和普适性强在语音增强中得到了广泛应用。基本谱减法在平稳的声学环境及较高信噪比时能取得较好的效果,然而在非平稳噪声环境及低信噪比下的增强结果由于语音失真和残留噪声的影响而不能令人满意。

近年来,为了减少语音失真和残留噪声对听觉效果的影响,进一步提高谱减法的性能,文献[1-6]提出了一系列改进算法。与传统的增强算法相比,噪声明显减少,残留噪声也得到了一定抑制,取得了较好的去噪效果。但上述几种算法在非平稳噪声环境及低信噪比情况下,残留噪声和语音失真现象依然存在。为此本文提出了一种基于实时噪声估计的改进谱减法。该方法首先利用基于临界带特征矢量距离的端点检测算法进行端点检测,然后利用低频区和高频区带噪语音特性定义一个时变的调节系数,该系数结合端点检测可以实时更新噪声的估计值,从而达到快速跟踪外界环境变化的目的。仿真结果表明,对于输入为低信噪比的带噪语音,本文提出的方法与其他增强方法相比,在提高信噪比、抑制背景噪声、减少语音畸变等方面取得了较好的效果。

1 谱减法理论分析

基本谱减法是一种基于短时谱幅度估计的算法。其思想是假设在加性噪声与短时平稳的语音信号相互独立的条件下,从带噪语音的功率谱中减去噪声功率谱,从而得到较为纯净的语音频谱。假设含噪语音信号 $y(t)$ 表示为

$$y(t)=s(t)+d(t) \tag{1}$$

式中,$s(t)$ 为纯净语音信号,$d(t)$ 为加性噪声,$s(t)$ 和 $d(t)$ 统计不相关。用 $Y(\omega)$,$S(\omega)$,$N(\omega)$ 分别表示 $y(t)$,$s(t)$,$d(t)$ 的傅里叶变换,对式(1)进行傅里叶变换,可得

$$Y(\omega)=S(\omega)+N(\omega) \tag{2}$$

假定语音信号与加性噪声是相互独立的,有:

$$|Y(\omega)|^2=|S(\omega)|^2+|N(\omega)|^2 \tag{3}$$

如果用 $P_y(\omega)$,$P_s(\omega)$,$P_n(\omega)$ 分别表示 $y(t)$,$s(t)$,$d(t)$ 的功率谱,则有

$$P_y(\omega)=P_s(\omega)+P_n(\omega) \tag{4}$$

由于平稳噪声的功率谱在发音前和发音期间可以认为基本没有变化,这样就可以通过发音前的"寂静段"(即认为在这一段里只有噪声存在)来估计噪声的功率谱 $P_n(\omega)$,从而由 $P_s(\omega)=P_y(\omega)-P_n(\omega)$ 计算出的功率谱即可认为是较纯净的语音信号的功率谱。频域处理过程中只考虑了功率谱的变换,而最后 IFFT 变换中需要借助相位谱来恢复增强后的语音时域信号。根据人耳对相位变换不敏感这一特点,可以用

到稿日期:2009-12-17 返修日期:2010-03-08 本文受航空科学基金(20080153002)资助。
程 塬(1984-),男,博士生,主要研究方向为语音信号处理及模式识别等,E-mail:isuccess@126.com;郭 雷(1956-),男,教授,博士生导师,主要研究方向为模式识别等;贺 胜(1986-),男,硕士生,主要研究方向为模式识别等;赵天云(1970-),男,讲师,主要研究方向为模式识别等。

原始带噪语音信号 $y(t)$ 的相位谱来代替估计之后的语音信号的相位谱，从而可以得到降噪后的语音时域信号。具体运算时，为防止出现负功率的情况，谱减时如 $P_y(\omega)<P_n(\omega)$，令 $P_s(\omega)=0$。所以，完整的谱减法计算公式如下

$$P_s(\omega)=\begin{cases}P_y(\omega)-P_n(\omega), & P_y(\omega)\geqslant P_n(\omega)\\ 0, & P_y(\omega)<P_n(\omega)\end{cases} \tag{5}$$

2 改进谱减法

在基本谱减法中，假定噪声是局部平稳的，即带噪语音中的噪声具有和语音段开始前的“寂静段”相同的统计特性，且在整个语音段中保持恒定值不变，从而采用对整个语音段减去相同噪声功率谱的办法进行语音增强。然而在现实世界中，这种理想的平稳噪声几乎是不存在的，因此在利用基本谱减法进行语音增强时效果并不是很好。我们有必要对估计噪声进行不停的更新，从而提高噪声估计的准确性。本文提出的语音增强系统框图如图 1 所示。

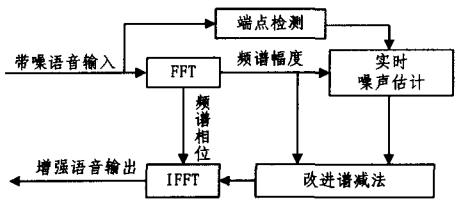

图 1 改进谱减法系统框图

2.1 语音端点检测

在语言增强算法中，通常用端点检测来更新噪声功率谱。本文拟采用文献[7]中提出的具有较高鲁棒性和正确率的基于临界带特征矢量距离的端点检测算法。

由于检测及增强是按帧进行的，因此可把语音模型写成帧的形式

$$y(m,n)=s(m,n)+d(m,n),m=1,2,\cdots,$$
$$n=0,1,\cdots,N-1 \tag{6}$$

式中，m 为帧号，N 为帧长。对其进行傅里叶变换，得

$$Y(m,k)=S(m,k)+D(m,k) \tag{7}$$

端点检测的算法流程为：

1)计算每一帧加窗语音 $y(m,n)$ 的功率谱 $|Y(m,k)|^2$。

2)划分临界带。在 $0\sim0.5f_s$（f_s 为采样频率）中确定 $\hat{f}_1$，$\hat{f}_2$，$\hat{f}_3$ …若干个临界带频率分割点。确定的方法是将 $i=1,2,3,\cdots$ 代入下式，即可求出相应的 $\hat{f}_i$，单位为 Hz。

$$i=\frac{26.81\times\hat{f}_i}{1960+\hat{f}_i}-0.53 \tag{8}$$

式中，由 $\hat{f}_1$ 和 $\hat{f}_2$ 构成第一临界带，$\hat{f}_2$ 和 $\hat{f}_3$ 构成第二临界带，依次类推。语音信号的采样频率通常为 8kHz，这样在 100Hz 至 4000Hz 范围内需要安排 16 个临界带。

3) 求临界带特征矢量。对每一个临界带中的功率谱 $|Y(m,k)|^2$ 求和，即可得到相应的临界带特征矢量。如果用 $H=[h_1,h_2,\cdots h_l,\cdots,h_L]$ 表示临界带特征矢量，则每一个分量可以通过下式得到，从而每帧都可以得到一个十六维的临界带特征矢量。

$$h_l=\sum_{\hat{f}_l<f_k\leqslant\hat{f}_{l+1}}|Y(m,k)|^2 \tag{9}$$

4) 求临界带特征矢量距离。假设前几帧信号是背景噪声，对这几帧的临界带特征矢量求均值，即可得到噪声的平均特征矢量值。然后利用下式对每一帧的特征矢量求其与噪声平均特征矢量的均方距离，即可得到特征矢量距离轨迹。

$$d_{cep}^2=\sum(c_i-c_0)^2 \tag{10}$$

式中，c_i 表示当前帧的临界带特征矢量，c_0 表示噪声的平均特征矢量。

5) 设定阈值，进行端点检测。设定一个阈值 D，逐帧进行比较。如果第 $i-1$ 帧、第 $i-2$ 帧的特征矢量距离都小于 D，而第 $i+1$ 帧、第 $i+2$ 帧的特征矢量距离都大于 D，我们就认为第 i 帧为语音段的起始位置。同样，如果第 $i-1$ 帧、第 $i-2$ 帧的特征矢量距离都大于 D，而第 $i+1$ 帧、第 $i+2$ 帧的特征矢量距离都小于 D，我们就认为第 i 帧为语音段的结束位置。

2.2 实时噪声估计[1,2]

众所周知，人们的语音信号主要分布在 50Hz 到 4.0kHz 之间，所以，我们可以认为 0～50Hz 低频区和 4.0kHz～0.5 f_s 高频区的语音信号只有噪声存在。于是，低频区和高频区的噪声频谱和 $|D^{Low}(m,k)|+|D^{High}(m,k)|$ 就等价于低频区和高频区带噪语音信号的频谱和 $|Y^{Low}(m,k)|+|Y^{High}(m,k)|$。也就是

$$|D^{Low}(m,k)|+|D^{High}(m,k)|\approx|Y^{Low}(m,k)|+|Y^{High}(m,k)|$$

但是，由于 0～50Hz 低频区和 4.0kHz～0.5f_s 高频区的范围有限，不能正确估计噪声的变化，为此，笔者引入一个变量 $|W|$，它表示通过端点检测得到的“寂静段”的噪声频谱的平均估计，用来实现对噪声频谱估计值的更新。同时，定义一个变化的调整系数 λ_m，λ_m 可以通过下式计算得到

$$\lambda_m=\frac{\sum\limits_k|D^{Low}(m,k)|+\sum\limits_k|D^{High}(m,k)|}{\sum\limits_k|W^{Low}|+\sum\limits_k|W^{High}|}=\frac{\sum\limits_k|Y^{Low}(m,k)|+\sum\limits_k|Y^{High}(m,k)|}{\sum\limits_k|W^{Low}|+\sum\limits_k|W^{High}|} \tag{11}$$

λ_m 的大小反映了噪声频谱的变化趋势，从而可以实现对噪声的实时估计。这样噪声的估计值为

$$|\hat{D}(m,k)|=\lambda_m\cdot|W| \tag{12}$$

于是，改进的谱减法计算公式为

$$|\hat{S}(m,k)|=\begin{cases}|Y(m,k)|-|\hat{D}(m,k)|, & |Y(m,k)|\geqslant|\hat{D}(m,k)|\\ \gamma|\hat{D}(m,k)|, & |Y(m,k)|<|\hat{D}(m,k)|\end{cases} \tag{13}$$

式中，参数 γ 通常取 0.01～0.5，可根据实际情况进行设定。对上式进行反傅立叶变换即可得到增强后的语音。

3 实验与分析

为了验证笔者提出的语音增强方法的有效性，将其与基本谱减法、文献[2]中提出的方法进行比较，并分别进行主观和客观评价。实验中，纯净语音选自 863 中文语音识别语料库，噪声取自 NOISEX-92 数据库中喷气式飞机驾驶舱的非平稳噪声信号，并将语音和噪声按比例线性相加生成不同信噪比(−5dB，0dB，5dB)的带噪语音。增强实验中使用的语音样本经过 8kHz 采样，16 位线性量化。在增强处理前对含噪语音信号加汉明窗进行分帧，每帧 256 个采样点，帧间重叠 128 个采样点，并且在增强处理后对语音信号进行恢复。

(下转第 222 页)

[5] Karypis G, Han EH, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling[J]. Computer, 1999, 32(8): 68-75

[6] Ester M, Kriegel HP, Sander J, et al. A density-based algorithm for discovering clusters in large spatial database with noise[C]// Simoudis E, Han J, Fayyad UM, eds. Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226-231

[7] Ankerst M, Breuning M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure[C]// Delis A, Faloutsos C, Ghandeharizadeh S, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Philadelphia: ACM Press, 1999: 49-60

[8] Polikar R. Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features[J]. IEEE Signal Processing Magazine, 2007, 24(4): 59-72

[9] Kuncheva L I. Combining Pattern Classifiers, Methods and Algorithms[M]. New York, NY: Wiley Interscience, 2005

[10] Asuncion A, Newman D J. UCI Machine Learning Repository [OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, School of Information and Computer Science, 2007

[11] 雷小锋，谢昆青，林帆，等. 一种基于 KMeans 局部最优性的高效聚类算法[J]. 软件学报，2008，19(7)：1683-1692

[12] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

(上接第 213 页)

图 2 为 0dB 的带噪语音经不同方法增强后的时域波形图。从图中不难发现，对含噪语音在使用本文方法增强后，比采用基本谱减法、文献[2]中提出的方法增强后都有明显的改善。同时，本文对不同信噪比条件下增强前后的语音进行主观试听实验。试验结果表明，采用基本谱减法对带噪语音进行增强后虽然噪声已明显减少，但又产生了有节奏的音乐噪声；采用文献[2]中提出的方法对带噪语音进行增强后，不仅噪声明显减少，而且音乐噪声也得到了一定的抑制；采用本文方法对带噪语音进行增强后，效果较基本谱减法有了很大改善，较文献[2]中提出的方法也有一定程度的改善，尤其在低信噪比时较为明显，所得结果更加容易让人接受。

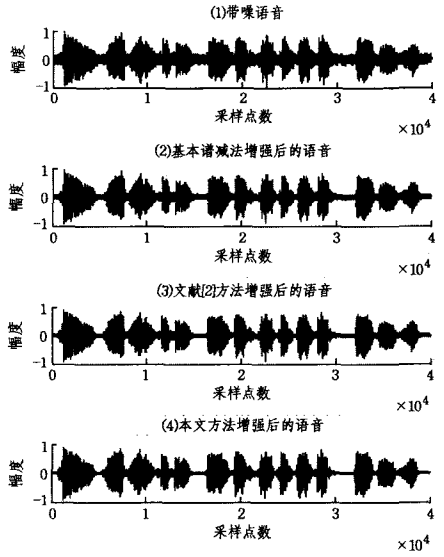

图 2 带噪语音和 3 种不同方法增强后的语音

客观评价选择信噪比(SNR)作为衡量指标，信噪比的定义为

$$SNR=10\lg\frac{\sum_{i=1}^{n}[s(i)]^2}{\sum_{i=1}^{n}[y(i)-s(i)]^2} \tag{14}$$

式中，n 为采样点数，$s(n)$为增强后的语音，$y(n)$为带噪语音。对输入信噪比为−5dB，0dB，5dB 的带噪语音分别采用基本谱减法、文献[2]中提出的方法以及本文方法进行语音增强。各增强方法的输出信噪比如表 1 所列。

表 1 3 种增强方法输出信噪比对比表 (单位：dB)

输入信噪比	基本谱减法输出信噪比	文献[2]方法输出信噪比	本文方法输出信噪比
−5	2.85	3.94	8.12
0	5.76	7.37	10.95
5	8.92	11.06	13.26

从表 1 中可以看出，本文方法的增强效果优于基本谱减法和文献[2]中提出的方法，噪声被更好地消除，信噪比得到了进一步提高。尤其是输入为低信噪比语音时，本文方法对信噪比的改善效果更加明显。

结束语 语音增强是语音信号处理的前沿领域，也是语音识别和语音合成等方向的基础。目前已存在众多的针对平稳噪声环境下的语音增强技术。然而，许多环境下的噪声都是非平稳的，本文提出了基于实时噪声估计的改进谱减法。通过与其他方法比较可知，本文提出的方法具有良好的降噪性能，较大程度地抑制了音乐噪声，减少了语音失真，提高了语音质量，特别是对于信噪比较低的情况，笔者方法优势明显。

参 考 文 献

[1] Yamauchi J, Shimamura T. Nonstationary Noise Estimation Using High Frequency Regions for Spectral Subtraction[J]. IEICE Transactions on Communications, 1998, 70(3): 335-349

[2] Yamashita K, Shimamura T. Nonstationary Noise Estimation Using Low-frequency Regions for Spectral Subtraction[J]. IEEE Signal Processing Letters, 2005, 12(6): 465-468

[3] Virag N. Single Channel Speech Enhancement Based on Masking Properties of Human Auditory System[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(2): 126-137

[4] Cohen I, Loizou P. Speech enhancement based on wavelet th - resholding and multitaper spectrum[J]. IEEE Signal Processing Letters, 2002, 9(1): 12-15

[5] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(5): 504-512

[6] 陈国明，赵力，邹采荣. 一种基于短时谱估计和人耳掩蔽效应的语音增强算法[J]. 电子与信息学报，2007，29(4)：863-866

[7] 武文娟，顾宏斌，潘秀林. 基于临界带特征矢量距离的端点检测算法[J]. 计算机科学，2009，36(2)：220-221