语音增强算法综述

王晶 傅丰林 张运伟
（西安电子科技大学通信工程学院 西安 710071）

摘要 噪声降低了语音的信噪比和可懂度，严重时会使语音处理系统无法正常工作。对带噪语音信号进行语音增强处理，是一个亟待解决的课题。本文对目前主要几种语音增强算法做了分析研究，结果表明各种方法都有不足，在实际应用中应根据具体环境和系统要求，结合各种算法以达到语音增强的最佳效果。

关键字 语音增强 噪声抑制 可懂度 识别率

1 引言

语音信号是人类传播信息和感情交流的重要媒体，是听觉器官对声音传媒介质的机械振动的感知，也是人类最重要、最有效、最常用、最方便的通信方式。但人们在语音通信过程中不可避免地会受到来自周围环境、传输媒介引入的噪声、通信设备内部电噪声、乃至其它讲话者的干扰，这些干扰最终将使接收到的语音信号并非纯净的原始语音信号，而是受噪声污染的带噪语音信号。这里的“噪音”定义为所需语音信号以外的所有干扰信号。干扰信号可以是窄带或宽带的、白噪声或有色噪声、声学的或电学的、加性的或乘性的，甚至可以是其它无关的语音。为了从带噪语音信号中获得尽可能纯净的语音信号，减少噪音的干扰，就需要进行语音增强。

语音增强有着广泛的应用，因此寻求一种有效的算法-对带噪语音信号进行处理以达到较高抗噪效果的研究意义很大。在一般情况下干扰信号是随机信号，要完全排除噪音是不现实的，所以语音增强的目标对收听人而言主要是改善语音质量，提高语音可懂度，减少疲劳感；对语音处理系统（识别器、声码器、手机）而言是提高系统的识别率和抗干扰能力。目前语音增强的方法很多，一些主要的方法都有各自的优缺点，在应用过程中要根据系统实际情况来选用.

2 语音增强的依据

语音增强与语音信号处理理论有关，而且涉及到人的听觉感知和语音学。噪声来源众多，随应用场合不同而特性各异，因此难以找到一种通用的语音增强算法可以适用于各种噪声环境，必须针对不同环境下的噪声采取不同的语音增强策略。要语音增强首先要了解语音和噪声的有关特性[3]。

2.1 语音特性

语音是时变的、非平稳、非遍历的随机过程。语音发声是一个时变过程，很多因素造成了发声系统的时变性，例如声道的面积随着时间和距离改变，气流速度随着声门处压力变化而变化等。但是声道形状有相对稳定性，在一段时间内（10ms~30ms），人的声带和声道形状是相对稳定的，可认为其特征是不变的，因而语音的短时谱具有相对稳定性，在语音分析中可以把语音信号分为若干分析帧，每一帧的语音可以认为是准稳定的。

语音可以分为周期性的浊音和非周期的清音。浊音和清音经常在一个音节中同时出现。浊音部分和音质关系密切，在时域上呈现出明显的周期性，在频域上有共振峰结构，而且能量大部分集中在较低频段内，是语音中大幅度高能量的部分；清音则具有明显的时域和频域特征，类似于白噪声，能量较小，在强噪声中容易被掩盖，但在较高信噪比时能提供较多的信息。在语音增强中，可以利用浊音的周期性特征，采用梳状滤波器提取语音分量或者抑制非语音信号，而清音则难以与宽带噪声区分。

语音感知对语音增强研究有重要作用，人耳对语音的感知主要是通过语音信号频谱分量幅度获取的，对各分量相位则不敏感，对频率高低的感受近似与该频率的对数值成正比。人耳有掩蔽效应，即强信号对弱信号有掩盖的抑制作用，掩蔽的程度是声音强度与频率的二元函数，对频率的临近分量的掩蔽要比频差大的分量有效的多。

2.2 噪声特性

噪声来源取决于实际的应用环境，因而噪声特性可以说变化无穷。噪声可以是加性的，也可以是非加性的。对于非加性噪声，有些可以通过变换转变为加性噪声。例如，乘性噪声可以通过同态变换成为加性噪声。某些与信号相关的量化

噪声可以通过伪随机噪声扰动的方法变换成信号独立的加性噪声。加性噪声大致上有：周期性噪声、脉冲噪声、宽带噪声和同声道的其他语音干扰等。

周期性噪声主要来源于发动机等周期性运转的机械，电气干扰，特别是电源交流声也会引起周期性噪声，其特点是有许多离散的窄谱峰。脉冲噪声来源于爆炸、撞击和放电等，表现为时域波形中突然出现的窄脉冲。宽带噪声的来源很多，包括热噪声、气流（如风、呼吸）噪声及各种随机噪声源，量化噪声也可视为宽带噪声。平稳的宽带噪声，通常也可认为是白色高斯噪声。同声道语音干扰是指当多个语音叠加在一起在单信道中传输时，双耳信号因合并而消失。

3 语音增强算法

如前所述，由于噪声特性各异，语音增强方法各有不同。40 多年来，人们针对加性宽带噪声研究了各种语音增强方法。目前应用的算法大致可以分为四种：参数方法、非参数方法、统计方法和其它方法。下面对这几类方法进行简介和分析。

3.1 参数方法

此类方法主要依赖于使用的语音生成模型（例如 AR 模型），需要提取模型参数（如基音周期、LPC 系数），常常使用迭代方法。如果实际噪音或语音条件与模型有较大的差距，或提取模型参数有困难，则此类方法容易失效。采用滤波器模型典型的有梳状滤波器、维纳滤波器[3]、卡尔曼滤波器[1,3]等。

在人类发声器官和语音产生的基本声学理论研究的基础上，建立起了离散时域的语音信号模型。语音信号被看成是线性时变滤波器在激励源激励下的输出。激励源分为浊音和清音两个分支，在浊音情况下，激励信号由一个周期脉冲发生器产生。在清音情况下，激励信号由一个随机噪声发生器产生；通常认为声道模型是一个全极点时变滤波器，滤波器参数可以通过线性预测分析得到。显然，如果能够知道激励参数和声道滤波器的参数，就能利用语音生成模型合成得到“纯净”的语音。这种方法的关键在于如何从带噪语音中准确地估计语音模型的参数(包括激励参数和声道参数)，这种增强方法称为分析—合成法。另一种方法则是鉴于激励参数难以准确估计，而只利用声道参数构造滤波器进行滤波处理。

利用语音信号浊音段有明显周期性的特点，可采用梳状滤波器提取语音分量来抑制噪声。滤波器输出信号是输入信号的延时加权和的平均值，当延时与信号的基音周期一致时，这个平均过程使周期性分量加强，而非周期分量或周期不同于信号的其他周期分量被抑制或消除。这种方法的关键是要准确估计出语音信号的基音周期。在基音变化的过渡段和强噪声背景干扰下无法精确估计时，这种方法的应用受到限制。

维纳滤波方法采用最小均方误差准则设计一个数字滤波器，带噪语音信号通过此滤波器便得到语音信号的估计。这个最佳滤波器就是维纳滤波器。带噪语音模型为 $y(n)=s(n)+d(n)$，式中 $y(n)$ 是带噪语音，$s(n)$ 是纯净语音，$d(n)$ 是噪声。维纳滤波器频域表达式为 $H(\omega)=\dfrac{P_s(\omega)}{P_s(\omega)+\lambda_d(\omega)}$（其中 $P_s(\omega)$ 和 $\lambda_d(\omega)$ 分别是纯净语音和噪声的功率谱密度）。维纳滤波器是在平稳条件的最小均方误差意义下的最优估计。但语音是非平稳的，实际环境中的噪声也是非平稳的。而且维纳滤波方法也没有完全利用语音生成模型。

卡尔曼滤波器是在已知状态方程和噪声统计特性的条件下，用线性预测（LPC）分析参数实现波形最小均方误差意义下的最佳估计器。卡尔曼滤波弥补了维纳滤波的两个缺陷，它是基于语音生成模型的，且在非平稳条件下也可以保证最小均方误差意义下的最优，故适合于非平稳噪声干扰下的语音增强。它的优点是不需要假定噪声的平稳性，对非平稳噪声也能运用。其缺点是：(a)需要叠代估计模型参数，在噪声强时误差大；(b)语音生成模型中假定激励是白噪声源，这仅对清音成立而对浊音是不成立的；(c)计算量较大；(d)优化标准是时域的波形误差最小，对语音信号而言此标准不够合理。

3.2 非参数方法

非参数方法不需要从带噪信号中估计模型参数，因此这种方法的应用范围较广。但由于没有利用可能的语言统计信息，故结果一般不是最优化的。这类方法包括谱减法[2,3]、自适应滤波法[4]等。

语音是非平稳随机过程，但在 l0ms~30ms

的分析帧内可近似看成是平稳的。如果能从带噪语音的短时谱中估计出“纯净”语音的短时谱，即可达到语音增强的目的。由于人耳对语音的感知主要是通过语音信号中各频谱分量幅度获得的，对各分量的相位不敏感。因此，此类语音增强方法将估计的对象放在短时谱幅度上。

谱减法的基本出发点是直接从带噪语音信号中减去噪声谱，并利用带噪语音相位重建增强后的“纯净语音”。它的优点是比较简单，只需要进行正反傅立叶变换，而且实时实现较容易。但是谱减法是一种最大似然估计，它存在的一个缺点就是放弃了对语音频谱的分析假设，然而对人耳来说，频谱分量的幅度才是最重要的。谱减法适用的信噪比范围较窄，在信噪比较低时对语音的可懂度损伤较大，这是因为信噪比主要代表了由浊音决定的大信号能量，而语音可懂度主要取决于元音和相对较小的代表辅音的信号。所以除了要降低噪声外，还要兼顾语音的可懂度和自然度。而且由于频谱相减会产生一种具有一定节奏感的残余噪声，一般称为“音乐噪声”。

自适应滤波法是通过双话筒分别采集噪声和带噪语音信号，从带噪语音幅度谱中减去经过自适应滤波器后的噪声分量幅度谱，然后加上带噪语音频谱的相位，经过傅里叶反变换就得到增强的语音信号。自适应滤波器通常采用 FIR 滤波器，系数采用最小均方（LMS）误差准则来迭代估计。这种方法的问题是如何得到与带噪语音中的噪声一致的噪声。利用双话筒实时采集到两路相同段的噪声，而且不受回声及其它衰变特性影响是很困难的。系统要求双话筒采集信号，这就限制了这种方法的适用范围。对于只允许单话筒采集的，一般是在语音间歇期间利用采集的带噪语音来估计噪声，但是这样会影响语音增强效果。此外如同谱减法一样还有一个缺点就是增强语音中含有明显的“音乐噪声”。

3.3 统计方法

统计方法较充分的利用了语音和噪音的统计特性，一般要建立模型库，需要训练过程获得初始统计参数，它与语音识别系统的联系很密切。如最小均方误差估计 (MMSE-minimum mean square error)[5~7]、听觉掩蔽效应[8~10]等。

语音特性的分析告诉我们要了解语音短时谱幅度分布，可以通过两个途径：一是假设一个合理的概率分布模型；另一个则是通过实际统计的方法去获得。对于语音增强来说，听觉意义上的失真准则与给定噪声情况下语音频谱的后验分布是无法知道的，因此，对于特定的失真准则和后验概率不敏感的估计方法是很有用处的。

最小均方误差（MMSE）估计正是一种对特定的失真准则和后验概率不敏感的估计方法。它是利用已知的噪声功率谱信息，从带噪语音频谱分量中估计出纯净语音频谱分量，借助带噪语音相位得到增强的语音信号。考虑到大部分语音的变化是比较缓慢的，帧与帧之间的频谱有着一定的相似性，其相应频谱分量之间存在某种相关性。这种相关性可以反映在前一帧的频谱值对后一帧频谱的分布产生一种约束影响。由此，产生了基于帧间频谱分布约束的 MMSE 估计方法。人耳对声音强度的感受是与谱幅度的对数成正比的，而且，语音处理的实践也表明，采用对数失真准则更为适合一些。为此，将上述 MMSE 估计式进行推广，得到频域分布约束下的短时对数谱的 MMSE 估计。

MMSE 算法达到了语音可懂度和降噪比的折衷，适用信噪比的范围较广，但是由于需要统计各种参数，算法运算量大，实时性不好。

当两个能量不等的声音作用于人的听觉系统时，能量较高的信号可以使较低的信号不易察觉，这就是人耳听觉系统的掩蔽效应。应用听觉掩蔽效应进行语音增强，语音信号能够掩蔽与其同时进入听觉系统的一部分能量较小的噪声信号，而使得这部分噪声不为人感知，利用一个功率谱域的基于听觉掩蔽门限的不等式准则，动态选择一个参数自适应变化的非线性函数估计语音短时谱幅度从而实现语音增强。这种方法在进行语音增强时，不需要把噪声完全抑制掉，只要使残留的噪声信号不被人感知即可，所以这样在消噪的同时可以减少不必要的语音失真。但是噪声掩蔽门限的计算是在纯净语音基础上得到的，在实际中一般只能用带噪语音来估计掩蔽门限，这样估计的结果误差很大。

3.4 其它方法

如小波变换[11,12]、卡亨南-洛维变换(KLT)[13,15]、离散余弦变换（DCT）[14,15]、人工神经网络[16,17]等。这些方法不像前三类方法那样成熟，可以概括地称为非主流方法。

利用之前的各种方法进行语音增强，需要知道噪声的一些特征或统计性质。在没有噪声先验知识的情况下，从唯一带噪语音信号中分离出语音信号，这非常困难。小波变换能将信号在多个

尺度上进行小波分解，各尺度上分解所得到的小波变换系数代表信号在不同分辨率上的信息。语音信号和噪声之间具有不同的 Lipschitz 指数，即信号具有正奇异性，而随机噪声具有负奇异性。这种性质在小波变换中，表现为信号的变换模值随尺度的增加而增加，随机噪声的变换模值随尺度的增加而减小。我们可以利用信号和噪声在小波下的这种截然不同的表现，提出一种有效的语音信号去噪方法。对输入带噪信号的小波系数设置一个合理阈值，仅让那些超过阈值的显著的小波系数用于小波逆变换来重构信号。这个阈值的选择确定对信号的去噪和恢复是有很重要影响的，因为这个门限阈值的确定直接影响信号去噪的效果和重构信号的失真程度。所以，用小波变换进行信号去噪，门限阈值的选择是关键。

Karhunen–Loeve 变换用于语音增强，这种算法是把带噪语音沿着经过 KLT 变换的纯净语音向量空间进行分解，得到特征向量，修正每一个向量使得当剩余噪声功率被限制在一特定值，然后经 KLT 反变换合成输出增强后的语音。

离散余弦变换(Discrete Cosine Transform)的语音消噪方法与小波变换类似，通过对带噪信号进行离散余弦变换后用阈值函数处理，再进行离散余弦反变换就可以得到增强的语音信号。同样，阈值的选择是这类方法的关键，也是不断研究改进的重要内容。

语音增强方法可以看作是从语音中区分出背景噪声的一种说话人区分方法。所以可以利用人工神经网络（例如反向传播 BP 网络）,用纯净语音信号作为网络训练信号形成一个语音数据库，带噪语音时间样值与纯净语音时间样值相比较并计算误差，然后基于误差最小准则利用 BP 算法调整网络权值，从而就可以提取出增强的语音信号。这种方法最适合语音识别领域。

4 总结

上述各种方法各有优缺点，分别适用于不同情况。参数方法对语音的模型参数依赖性强，但在低信噪比条件下不容易得到正确的模型参数；非参数方法由于频谱相减会产生一种具有一定节奏感的残余噪声，即 “音乐噪声”；统计方法需要大量数据进行训练以得到统计信息；小波变换以及离散余弦变换的阈值选取困难，运算量大。

实际使用中常常根据具体的环境噪声和语音特性将不同方法结合起来应用，通过方法互补取得更好的语音增强效果。例如 MMSE 谱估计方法中需要通过带噪语音估计出语音方差和噪声方差，这些未知参数可以通过用谱减法处理带噪语音部分得到的增强语音来计算。

为了达到减少对语音可懂度的损伤，可以结合 MMSE 谱估计法和听觉掩蔽效应 [10]。又由于小波阈值的语音增强算法中很难选取合适的阈值，一般达不到理想的效果，可以与 MMSE 相结合，根据一定的均方误差准则自适应寻找到最佳阈值，或者基于神经网络选取合适阈值[18]。还有，一般通过小波阈值消噪处理，达不到最佳语音增强效果，有必要在此基础上进一步进行噪声处理，可以再通过 KLT 变换进行噪声处理[13]，或者在小波处理之前先用谱减法进行预处理等等。这样结合各种方法对语音进行增强就可以达到比较理想的效果。

参考文献

[1] Gabrea M. Adaptive Kalman filtering-based speech enhancement algorithm. Canadian Conference on Electrical and Computer Engineering, Toronto， 2001

[2] Ogata S, Shimamura T. Reinforced spectral subtraction method to enhance speech signal. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, Singapore， 2001

[3] 杨行峻, 迟惠生等. 语音信号数字处理[M]. 北京:电子工业出版社, 1995

[4] 胡啸， 胡爱群， 赵力． 一种新的自适应语音增强系统[J] . 电路于系统学报, 2003, 10:72~75

[5] Pei Ding, Zhigang Cao.Combining MMSE enhancement with LA model adaptation for robust automatic speech recognition[J]. Electronics Letters , 2001,37(8):539 ~ 540

[6] Martin R. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando，2002

[7] ChangHuai You, SooNgee Koh, Rahardja S. Adaptive /spl beta/-order MMSE estimation for speech enhancement. IEEE International Conference on Acoustics, Speech, and Signal Processing, Singapore，2003

[8] 刘海滨, 吴镇扬, 赵力,等. 非平稳环境下基于人耳听觉掩蔽特性的语音增强[J]. 信号处理, 2003, 8：303~307

[9] Dai Qijun, Chen Yanpu, Bian Zhengzhong. Optimizing speech enhancement based on noise masked probability.

2002 6th International Conference on Signal Processing, , Xi'an，2002

[10] 张金杰, 曹志刚, 马正新. 一种基于听觉掩蔽效应的语音增强方法[J]. 清华大学学报(自然科学版), 2001,41(7) ：34~37

[11] Yi Hu, Loizou P C. Speech enhancement based on wavelet thresholding the multitaper pectrum[J]. IEEE Transactions on Speech and Audio Processing, 2004, 12(1):59 ~ 67

[12] 胡广书编著. 数字信号处理–理论、算法与实现[M]. 北京: 清华大学出版社, 2001

[13] Rezayee A, Gazor S. An adaptive KLT approach for speech enhancement[J] . IEEE Transactions on Speech and Audio Processing, 2001, 9(2): 87 ~ 95

[14] Hasan M K, Zilany M S A, Khan M R. DCT speech enhancement with hard and soft thresholding criteria[J]. Electronics Letters , 2002,38(13):669 ~ 670

[15] 楼红伟, 胡光锐. 基于简化的 KLT 和小波变换的非平稳宽带噪声语音增强[J]. 控制与决策, 2003, 9：577~589

[16] Liew Ban Fah, Hussain A, Samad S A. Speech enhancement by noise cancellation using neural network. TENCON 2000. Proceedings, Kuala Lumpur， 2000

[17] Murakami T, Namba M, Hoya T. et al. Speech enhancement based on a combined higher frequency regeneration technique and RBF networks. TENCON '02. Kawasaki，2002

[18] Medina C A, Alcaim A. Wavelet denoising of speech using neural networks for threshold selection[J]. Electronics Letters , 2003,39(25):1869 ~ 1871

(上接第 15 页)

参考文献

[1]S C Butler, F A Tito . A Broadband Hybrid Mangetostrictive/Piezoelectric Transducer Array. Oceans 2000 MTS/IEEE conference proceeding,Washington, 2000

[2] Patrick R Downey, Marcelo J Dapino, Ralph C Smith. Analysis of Hybrid PMN/Terfenol Broadband Transducers in Mechanical Series Configuration. SPIE 10th Annual International Symposium on Smart Structures and Materials, Barcelona, 2003

3 路德明 . 水声换能器原理[M]. 青岛:青岛海洋大学出版社,2001

4 Göran . Handbook of Giant Magnetostrictive Materials[M]. SanDiego: Academic Press,2000