# 一种新的基于数据场的语音增强算法

黄建军 张雄伟 张亚非 淦文燕 邹 霞
(解放军理工大学指挥自动化学院, 南京 210007)

**摘 要**: 语音增强是消除噪声干扰的主要手段, 在语音处理系统中得到广泛应用。传统语音增强算法认为相邻帧语音幅度谱之间是相互独立的, 而研究表明语音幅度谱时频点之间存在相互依赖关系。缺乏对邻域时频结构信息的利用使得传统增强算法的性能难以进一步提高。本文首次将数据场引入到对语音的听觉感知领域, 用数据场对语音的时频依赖性进行建模, 提出一种新的基于数据场的语音增强算法。该算法通过最小化势场分布的熵确定了时频点之间的相互作用力程, 在带噪语音数据场中估计噪声的最小统计量得到二值时频掩蔽值, 最后利用二值时频掩蔽消除噪声干扰。实验测试表明, 与 Martin 算法相比, 基于数据场的语音增强算法在提高去噪效果的同时能有效减少语音的失真。

**关键词**: 语音增强; 数据场; 时频掩蔽; 噪声估计
**中图分类号**: TN912.3 **文献标识码**: A **文章编号**: 1003-0530(2011)08-1200-06

# A Novel Speech Enhancement Algorithm Based on Data Field

HUANG Jian-jun ZHANG Xiong-wei ZHANG Ya-fei GAN Wen-yan ZOU Xia
(Institute of Command Automation, PLA Univ. of Sci. & Tech., Nanjing 210007, China)

**Abstract**: In traditional speech enhancement algorithm the speech spectral amplitude is assumed to be mutually independent. Little work has been done to incorporate the time and frequency dependencies of speech. Without exploring the structure information of the time and frequency neighbors limit the performance of traditional speech enhancement algorithms. In this paper, we propose a novel speech enhancement algorithm based on data field theory, which is capable of modeling the time and frequency dependencies of speech. Data field defines the distribution of the magnitude of speech spectral samples conditioned on the values of their time and frequency neighbors. This formulation allows the explicit inclusion in the amplitude estimation model of both time and frequency dependencies that exist among the amplitudes of speech spectral. The proposed algorithm is evaluated by applying it to enhance noisy speech at various noise levels. Systematic evaluation shows that the proposed algorithm offers improved signal to noise ratio and presents an enhanced ability in preserving the weaker speech spectral components compared to Martin's algorithm.

**Key words**: Speech enhancement; Data Field; Time-Frequency masking; Noise estimate

## 1 引言

在实际应用中,语音增强作为语音处理系统的前端处理手段已经成为不可缺少的一部分,许多研究者在该领域做了大量工作。目前常用的谱减法及其改进算法[1]具有一定的去噪效果,这类算法的研究工作主要集中在噪声的短时谱幅度估计、先验信噪比和无语音概率的计算上,但是语音和噪声的非平稳性使得谱估计并不准确,从而产生了音乐噪声和语音的失真。为了克服传统增强算法在去除噪声的同时造成的语音失真问题,Ding[2]提出了在整个频域内重新生成浊音和清音谱的后处理技术。该技术首先采用非线性变换方法得到激励信号和平滑的谱包络估计,然后结合传统去噪算法得到激励和谱包络的加权乘积以生成综合后的语音。实验证明这种后处理技术能产生更加自然的语音,并且实验还发现通过重新生成语音成分的方法还可以有效地减少残余的音乐噪声。为了从理论上分析去噪和语音失真问题,Seokhwan[3]将语音增强转化为一个约束最优化问题,提出一种心理声学约束和失真最小化的语音增强算法。通过约束优化问题的求解,该算法不仅能使残余噪声保持在掩蔽阈值之下而且增强后的语音失真最小。然而该算法中的约束最优化问题在某些条件下不可解,限制了其在实际语音通信中的应用。总之,目前对语音增强的研究主要集中在残留噪声和语音失真两者之间的权衡问题上[3],要尽可能地消除残留噪声常常得以牺牲语音质量为代

收稿日期: 2011 年 6 月 3 日; 修回日期: 2011 年 7 月 1 日
基金项目: 江苏省自然科学基金项目(BK2009059)资助

价,反之亦然。

对语音信号特性的研究表明,缺少对语音结构信息的利用是语音增强算法性能不能进一步提高的主要原因。目前语音增强算法大多基于单个或多个短时帧考虑,只利用了时域上相邻帧的信息。而语音是一种具有很强谐波结构的信号,特别地,对于浊音来说,在一个短时帧内相隔一个基音频率的时频点之间也存在相互依赖关系,对频率轴信息的进一步利用有利于提高语音增强系统的性能[4]。近来兴起的计算听觉场景分析(CASA, Computational Auditory Scene Analysis)[5,6]应用听觉感知分段来进行语音分离的研究也表明:对时间轴和频率轴上相邻时频信息的利用可以更好地表示语音的内在规律。CASA 是基于人类感知和组织声音机制中的潜在规则来建立模型。CASA 系统中使用的听觉感知分段就是将可能来自同一个信源的时频单元聚集在一起,形成时频区域,这个时频区域对应着基本的听觉感知单元。相比于单个时频单元,听觉感知分段作为由时频单元组成的区域包含更多的关于信源的全局信息,例如谱包络和时域包络。这些全局的信息对区分来自不同信源的混合信号至关重要。同样,对邻域语音结构信息的利用可以更好地区分语音和噪声,从而提高语音增强系统的性能。

Andrianakis 在文献[4]中首次提出了基于马尔科夫随机场(Markov Random Fields,MRF)的语音增强算法,该算法在利用相邻时频信息来指导语音增强方面做出了有意义的尝试。该算法认为语音谱幅度之间并不是相互独立的,相邻时频点之间具有相互依赖关系,并进一步利用 MRF 来构建语音谱幅度的条件先验分布从而得到语音谱幅度估计。但是该算法存在两个问题:一是需要对每一帧带噪语音做基音估计。而对含噪语音做基音估计本身就是一个棘手的问题,现有算法很难准确估计带噪语音的基音周期,因此增强算法的性能很大程度上依赖于基音估计的准确性;二是在构建马尔科夫随机场模型时只考虑了前后帧和上下频率点的信息,对语音时频点相邻信息的利用有限。

为了充分利用语音时频域的结构信息来提高语音增强算法的去噪能力并尽可能地减少语音失真,本文提出了用数据场对带噪语音的时频点之间的依赖关系进行建模,从而利用邻域帧时频结构信息更好地区分语音和噪声,形成对基本听觉感知单元轮廓的刻画,利用二值时频掩蔽消除噪声干扰,有效地减少了残留噪声和语音的失真。

## 2　基于数据场的语音增强

### 2.1　数据场的引入

数据场概念是李德毅在文献[7]中首次提出的,数据场将物质粒子间的相互作用及其场描述方法引入抽象的数域空间。已知空间 $\Omega \subseteq R^p$ 是包含 $n$ 个数据对象的数据集 $D=\{x_1,x_2,\ldots,x_n\}$,我们将每个数据对象视为 $p$ 维空间中具有一定质量的粒子,其周围存在一个虚拟作用场,位于场内的任何其他对象都将受到场力的作用,由此所有对象的联合作用就在空间上确定了一个数据场。

数据场描述了人类从数据到信息再到知识的认知和思维过程,在数据处理中得到了广泛应用,例如掌声同步中的涌现计算[8]、聚类分析[9]、空间数据挖掘[10]等等。数据场认为样本空间中的每个观测数据都不是独立存在的,都对数域空间中的每个点具有影响力,数据的作用可以遵循距离衰减的原则辐射。数据场刻画了每个数据对知识发现任务的不同作用,能够把所有处于规则或散乱状态的数据的能量都扩展到规则的数据场域中,既考虑了数据不确定性,又使得数据挖掘易于理解和操作。

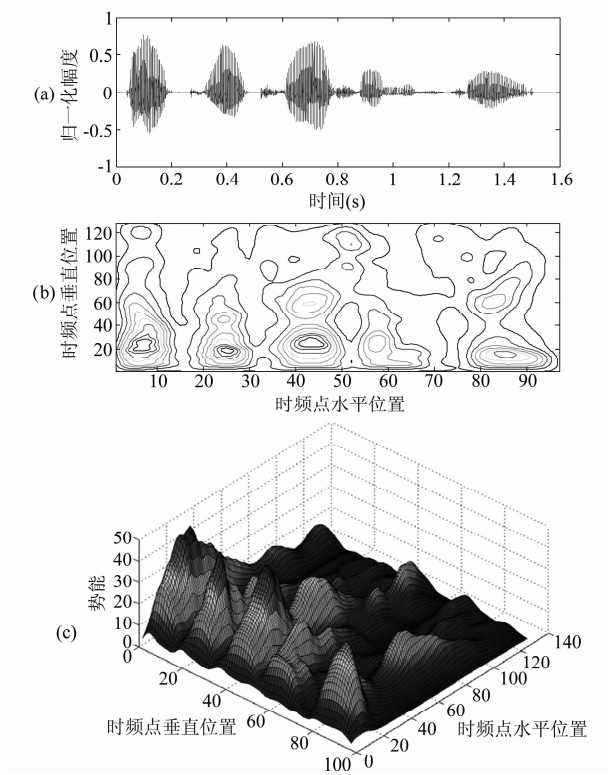


图 1　(a)语音的时域波形 (b)语音的数据场势能分布 (c) 语音三维势场分布

Fig. 1　(a) speech signal in time domain, (b) equipotential lines plot of speech data field, and (c) three-dimensional view of a two-dimensional potential field

文献[4]表明语音时频点之间并不是独立的,相邻时频点之间具有相互依赖关系,而数据场则可以用来对时频点间的依赖关系进行建模。将数据场应用到语音的听觉感知中,可以使得语音时频域中不规则的数

据点有序的组织起来,形成一个基本的听觉感知单元,这个基本单元可以对应着一个音素、一个音节、或是一个字,甚至一个句子,这可以通过定义数据场的影响因子来实现不同尺度的组织,达到不同粒度的听觉感知。图 1(a)(b)显示了一段纯净语音的时域波形及其数据场分布。从图中可以看出,在场分布的密集区正好对应这一个音节,二维势场分布的形状能很好地刻画出语音的时频分布的轮廓。图 1(c)二维势场分布的三维视图的各个山峰更能形象地表示基本语音事件的分布特点。

### 2.2 带噪语音数据场势函数的定义

设 $x(n)$ 和 $d(n)$ 分别表示干净语音信号和加性噪声信号,则带噪语音信号 $y(n)=x(n)+d(n)$。其中,$x(n)$ 和 $d(n)$ 相互独立。对带噪语音信号分帧,加窗后变换到频域。$y(n)$ 的 DFT 变换可以表示为:

$$Y(k,l)=X(k,l)+D(k,l)=\sum_{n=0}^{L-1}y(lR+n)h(n)e^{-j2\pi kn/L},\quad 0\le k\le L-1 \tag{1}$$

其中,$h(n)$ 为归一化窗,$k$ 为频带序号,$l$ 为帧序号,$R$ 为帧间重叠长度,$L=2R$ 为帧长,一般设为 256。选择 $M$ 个分析帧作为分析的对象,则这 $M$ 个帧就形成一个 $L\times M$ 的二维时频空间,若将频域带噪语音的时频点视为二维空间中的一个数据对象,将频率点幅度 $\rho_{kl}=\|Y(k,l)\|$ 作($k=0,12,\ldots,L\text{-}1;l=0,12,\ldots,M\text{-}1$)为数据对象的质量。由于人耳对语音的感知是通过语音信号中各频率分量幅度获取的,对各分量的相位则不敏感,因此可以选择语音的频率分量幅度作为数据对象的质量。根据数据场的势函数公式[7],数据场任一点 z 的势函数可以计算为:

$$\varphi(z)=\sum_{k=0}^{L-1}\sum_{l=0}^{M-1}\varphi_{kl}(z)=\sum_{k=0}^{L-1}\sum_{l=0}^{M-1}\left(\rho_{kl}\times e^{-\left(\frac{\|z-z_{kl}\|}{\sigma}\right)^2}\right) \tag{2}$$

其中,$\|z-z_{kl}\|$ 为对象 $z_{kl}$ 与场点 z 间的距离,$\sigma\in(0,\infty+)$ 用于控制对象间的相互作用力程,称为影响因子。选择合适的影响因子有利于语音和噪声的分离。

### 2.3 语音数据场中影响因子的优化

对于给定的势函数形态,影响因子 $\sigma$ 的取值会对数据场的空间分布产生极大的影响。当 $\sigma$ 值很小时,时频点间的相互作用力程很短,每个时频点周围的势值很小;反之,如果 $\sigma$ 值很大,时频点间的相互作用很强,每个时频点对周围时频点影响很大。所以不合适的 $\sigma$ 值下的势场分布显然不能产生有意义的总体估计。因此影响因子 $\sigma$ 的选取应使语音数据场分布尽可能体现时频点的内在分布。

语音数据场的空间分布主要取决于时频点之间的相互作用力程。那么如何确定语音数据场中时频点之间相互作用力程?从人耳听觉的角度来看,基本的听觉感知单元可以认为是一个音素,处于同一个音素的时频点之间才会相互影响;从计算听觉场景分析的角度,时频点之间的相互作用力程局限于同一个“分段”中,这个分段来自某个独立的声学事件。而不论音素还是分段都是不断变化的,此时不能给出一个合适的相互作用力程,即不能确定影响因子的大小。但是在语音和噪声分离的场合中,可以从另一个角度来考虑这个问题。语音增强的目的是尽量地扩大语音和噪声的差异从而消除噪声,因此使语音和噪声分布尽可能的不对称有利于消除噪声。

为了选取合适的影响因子,文献[7]引入了势熵的概念来衡量势场分布的合理性。根据信息论,熵为系统不确定性的度量,熵越大,系统的不确定性越大。对于空间中的势场分布来说,如果每个对象所处位置的势值近似相等,数据分布的不确定性最大,即具有最大的熵;反之,如果对象所处位置的势值非常离散或很不对称,则不确定性最低,具有最小的熵。因此,通过寻找具有最小熵的势场分布,就可以获得优化的 $\sigma$ 值。

如果语音数据场每个时频点的势值为 $\varphi(k,l)$,则整个语音数据场的势熵可以定义为:

$$\mathrm{H}=-\sum_{k=0}^{L-1}\sum_{l=0}^{M-1}\left(\frac{\varphi(k,l)}{\Theta}\log\left(\frac{\varphi(k,l)}{\Theta}\right)\right) \tag{3}$$

其中 $\Theta=\sum_{k=0}^{L-1}\sum_{l=0}^{M-1}\phi(k,l)$ 为一个标准化因子。

显然有 $0\le H\le\log(K\times M)$,当且仅当所有 $\varphi(k,l)$ 相等时,$H=\log(K\times M)$。优化 $\sigma$ 本质上是一个单变量非线性函数 $H(\sigma)$ 的最小化问题,即

$$\min\ \mathrm{H}(\sigma)=\min-\sum_{k=0}^{L-1}\sum_{l=0}^{M-1}\left(\frac{\phi(k,l)}{\Theta(\sigma)}*\log\left(\frac{\phi(k,l)}{\Theta(\sigma)}\right)\right) \tag{4}$$

这里选择黄金分割法来求解最小化问题,其基本思想是通过选取试探点和比较函数值,使包含极小点的搜索区间不断减小,直至获得满足精度要求的函数极小点。具体的算法描述见算法 1。

通过获得优化的 $\sigma$ 值,可以得到具有最小熵的势场分布,此时语音数据场中语音和噪声的分布具有最大的不对称性,也就是说通过优化影响因子可以在强化语音成分的同时弱化噪声的影响,即扩大了噪声与语音的差异,使得下一步的时频掩蔽更具鲁棒性,有利于消除噪声。

**算法 1　影响因子 $\sigma$ 的优化算法**

输入: 数据集 $D=\{|Y(k,l)|,k=1,2,\ldots,L;l=1,2,\ldots M\}$

输出: 优化的 $\sigma$

算法步骤:

**begin**:

置 $a=\frac{\sqrt{2}}{3}\min\limits_{k\neq k',l\neq l'}\left\|\,|Y(k,l)|-|Y(k',l')|\,\right\|$, $b=\frac{\sqrt{2}}{3}\max\limits_{k\neq k',l\neq l'}\left\|\,|Y(k,l)|-|Y(k',l')|\,\right\|$, 置精度要求 $\varepsilon$

令 $\sigma_l=a+(1-\tau)(b-a)$, $\sigma_r=a+\tau(b-a)$, $\tau=\frac{-1+\sqrt{5}}{2}$

计算 $H_l=H(\sigma_l)$ 和 $H_r=H(\sigma_r)$

**while** $|b-a|>\varepsilon$ **do**

　　**if** $H_l<H_r$ **then**{

　　令 $b=\sigma_r,\sigma_r=\sigma_l,\ H_r=H_l$;

　　计算 $\sigma_l=a+(1-\tau)(b-a)$, 和 $H_l=H(\sigma_l)$;

　　}

　　**else**{

　　令 $a=\sigma_l,\sigma_l=\sigma_r,\ H_l=H_r$;

　　计算 $\sigma_r=a+\tau(b-a)$ 和 $H_r=H(\sigma_r)$;

　　}

**end while**

**if** $H_l<H_r$ **then** $\{\sigma=\sigma_l\}$

**else** $\{\sigma=\sigma_r\}$

**return** $\sigma$

**end**

### 2.4　带噪语音数据场中的噪声估计

R. Martin 在文献[11]中提出了一种新的最小统计量噪声估计算法,算法认为,不论有话无话,带噪语音信号的功率谱在一定的分析窗内总会降低到一定的底限,此底限的统计最小值即为噪声功率的最小值,通过相应的偏差补偿,则可以得到噪声功率的统计平均值。因此,噪声的更新就也可以在有话的语音帧更新。将 Martin 提出的功率谱最小统计量噪声估计扩展到带噪语音数据场,可以得到数据场中的最小统计量噪声估计。算法步骤如下:

(1)计算数据场中带噪语音 $y(n)$ 的势能估计 $\hat{\varphi}(k,l)$

$$\hat{\varphi}(k,l+1)=\alpha\hat{\varphi}(k,l)+(1-\alpha)\,|\varphi(k,l)|^2 \tag{5}$$

其中 $\varphi(k,l)$ 表示数据场时频点的势值, $|\varphi(k,l)|^2$ 表示该时频点的势能; $\alpha$ 表示平滑因子,一般取值在 0.95 ~0.98 之间[11]。

(2)噪声的势能估计。

得到平滑的势能估计 $\hat{\varphi}(k,l)$ 后,在连续的 M 帧的时间窗内,寻找带噪语音势能的最小值,这个最小值可近似作为噪声的势能估计。

$$\hat{\varphi}_{\min}(k,l)=\min\{\hat{\varphi}(k,l-M-1),\ldots,\hat{\varphi}(k,l-1),\hat{\varphi}(k,l)\} \tag{6}$$

(3)偏差补偿。

考虑到带噪语音的最小值与真实噪声的偏差,需要引入偏差补偿因子 $B_{min}$ 对进行修正,其取值范围在 1.1 ~1.4 之间[11]。

$$\hat{\varphi}_d(k,l)=B_{\min}\hat{\varphi}_{\min}(k,l) \tag{7}$$

### 2.5　二值时频掩蔽计算

在得到噪声的势能估计 $\hat{\varphi}_d(k,l)$ 之后,如果带噪语音势能 $\hat{\varphi}(k,l)$ 大于 $\hat{\varphi}_d(k,l)$,则二值掩蔽值 $M(k,l)$ 为 1,反之为 0。见下式。

$$M(k,l)=\begin{cases}1, & if\ \hat{\varphi}(k,l)>\hat{\varphi}_d(k,l)\\ 0, & otherwise\end{cases} \tag{8}$$

利用噪声谱信息得到二值时频掩蔽值后就可以从带噪语音谱 $|Y(k,l)|$ 中估计出语音信号谱 $\hat{X}(k,l)$。由于人耳对相位不敏感,因此只要估计出语音谱幅度 $|\hat{X}(k,l)|$,添加上带噪语音谱的相位,再进行逆 STFT 分析就可以得到增强后的语音 $\hat{x}(n)$。

增强后的语音信号谱为:

$$\hat{X}(k,l)=M(k,l)\cdot|Y(k,l)|\cdot\angle Y(k,l) \tag{9}$$

傅立叶逆变换后得到各帧时域增强语音 $\hat{x}_l(n)=\sum_{k=0}^{L-1}\hat{X}(k,l)e^{j2\pi kn/L}$, 将各帧恢复时域信号重叠相加即可得到连续的语音信号。

基于数据场的语音增强算法的总体框图如图 2 所示。

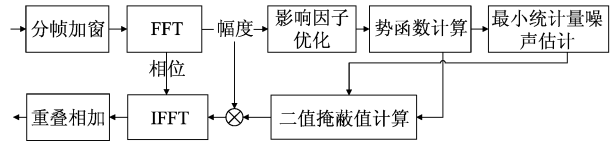


图 2　本文算法总体框图

Fig. 2　Diagram of the proposed enhancement algorithm

## 3　算法仿真与性能评估

实验时以 NOIZEUS 带噪语音库[12]来评估语音增强算法的性能。该语音增强算法评估库包含 8 种类型的噪声,分别为 suburban train 噪声、babble 噪声、car 噪声、exhibition hall 噪声、restaurant 噪声、street 噪声、airport 噪声和 train-station 噪声。每种类型的带噪语音的信噪比等级又分为 0dB、5dB、10dB、15dB。实验中选择三种噪声类型进行评估,分别是 babble 噪声、car 噪声和 airport 噪声,以 Martin 算法作为对照,语音增强算法的性能的评价标准选择信噪比和对数频谱距离。

信噪比是衡量信号中所含噪声能量大小的重要标

准,主要通过输入信噪比和输出信噪比的对比来检验语音增强的效果。我们用输入信噪比($\mathrm{SNR}_{in}$)和输出信噪比($\mathrm{SNR}_{out}$)分别表示输入带噪语音的信噪比和输出增强语音的信噪比,它们分别定义如下:

$$\mathrm{SNR}_{in} = 10\lg\frac{\sum_{n=1}^{N}\sum_{l=1}^{L}x_n^2(l)}{\sum_{n=1}^{N}\sum_{l=1}^{L}d_n^2(l)} \tag{10}$$

$$\mathrm{SNR}_{out} = 10\lg\frac{\sum_{n=1}^{N}\sum_{l=1}^{L}x_n^2(l)}{\sum_{n=1}^{N}\sum_{l=1}^{L}[\hat{x}_n(l)-x_n(l)]^2} \tag{11}$$

式中 $x_n(l)$ 和 $d_n(l)$ 分别表示纯净语音和噪声信号,$\hat{x}_n(l)$ 表示增强语音,$L$ 表示帧长,$N$ 表示帧的总数。

对数频谱距离(LSD)比较两个信号的频谱相似度,与人的主观听觉有很强的关联性,是评价语音质量的重要标准。对数频谱距离越小,表明增强语音失真越小,增强语音质量越好;反之,则增强语音质量越差。对数频谱距离计算公式如下:

$$LSD = \left(\frac{1}{K}\sum_{l=0}^{K-1}\frac{1}{L}\sum_{k=0}^{L-1}\left(10\lg\frac{|\hat{X}(k,l)|^2}{|X(k,l)|^2}\right)^2\right)^{1/2} \tag{12}$$

式中 $X(k,l)$ 和 $\hat{X}(k,l)$ 分别表示分帧纯净语音和增强语音的傅立叶变换。

表 1 给出了不同信噪比条件下不同噪声类型的消噪效果。表中的每列数据为输出信噪比,括号内的数据表示对数频谱距离,从表中数据可知,相比 Martin 算法,基于数据场的语音增强算法具有更高的输出信噪比和更低的对数频谱距离,这说明本文算法能更好的抑制噪声,同时语音失真更小。

表 1 输入信噪比、输出信噪比与对数频谱距离的对应关系

Tab. 1 Performance evaluation using SNR and LSD measure

| 算法 | 噪声类型 | $\mathrm{SNR}_{in}=0$ | $\mathrm{SNR}_{in}=5$ | $\mathrm{SNR}_{in}=10$ | $\mathrm{SNR}_{in}=15$ |
|---|---|---|---|---|---|
| Martin 算法 | Babble | 7.15(10.55) | 10.33(9.12) | 12.78(7.43) | 16.55(6.32) |
| 本文算法 | | 9.95(8.60) | 12.61(7.53) | 13.28(5.97) | 16.90(5.29) |
| Martin 算法 | Car | 9.33(8.87) | 13.43(7.40) | 17.69(6.53) | 22.03(5.60) |
| 本文算法 | | 12.24(6.43) | 15.83(5.30) | 19.58(4.79) | 23.21(4.47) |
| Martin 算法 | Airport | 10.24(7.73) | 14.57(6.15) | 18.67(5.24) | 21.73(4.60) |
| 本文算法 | | 13.74(5.36) | 17.54(4.45) | 20.79(3.61) | 23.54(2.94) |
| Martin 算法 | White | 10.40(7.4) | 14.87(5.94) | 19.21(5.06) | 22.02(4.51) |
| 本文算法 | | 13.98(5.06) | 17.81(4.29) | 21.06(3.48) | 23.65(2.85) |

信噪比和对数频谱距离能从整体上衡量信号中所含噪声能量大小和语音的失真程度,但是无法描述残余噪声的细节信息。时域波形、语谱图则是观察语音细节信息的很好的工具。为了说明数据场消除噪声的有效性,我们同时在图中加入了各种信号的语谱场描述。

图 3 ~ 图 5 从时域波形、语谱图、数据场方面对两个算法进行比较,给出了在高斯白噪声且 $\mathrm{SNR}_{in}=5$ dB 条件下,两种算法增强的效果。图 3 从上至下分别为干净语音、带噪语音、Martin 算法增强后的语音、本文算法增强后语音的时域波形,图 4 为图 3 对应信号的语谱图,图 5 为图 3 对应信号的数据场。

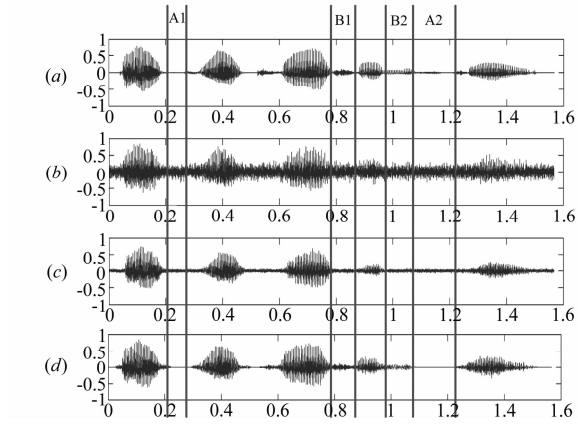


图 3 语音的时域波形(a 干净语音、b 带噪语音、c Martin 算法增强后的语音、d 本文算法增强后语音)

Fig. 3 Time waveforms of speech. (a) Original speech. (b) Noisy speech. (c) Speech enhanced with Martin's algorithm. (d) Speech enhanced with proposed algorithm

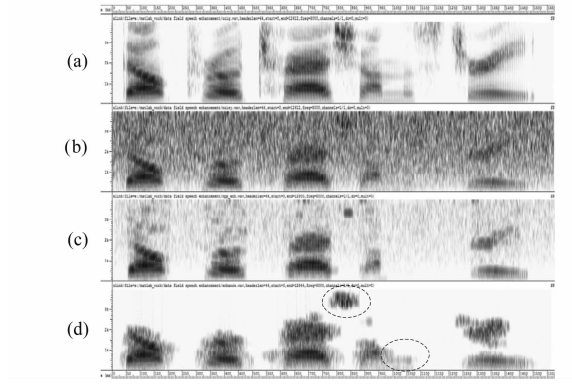


图 4 语音的语谱图(a 干净语音、b 带噪语音、c Martin 算法增强后的语音、d 本文算法增强后语音)

Fig. 4 Speech spectrograms: (a) Original speech; (b) Noisy speech; (c) Speech enhanced with Martin's algorithm; (d) Speech enhanced with proposed algorithm

从图中可以看出,在无声段,如图 3 中的 A1、A2 处,Martin 算法仍然有噪声残留,而本文算法几乎消除了噪声;在清音段,如图 3 的 B1、B2 处,Martin 算法几乎把清音切除了,而本文算法则能保留部分清音段,在一定程度减少了语音失真。由于清音具有类噪声特性,在低信噪比条件下传统语音增强算法很容易将清音切除,因此降低了语音的可懂度,造成了语音失真。

从恢复出来的时域波形可知,基于数据场的语音增强算法能够在消除噪声的同时减少语音失真。这主要是由于在计算带噪语音数据场的势能分布时通过优化的影响因子可以更好的区分语音和噪声,从而可以更好的消除噪声。图 4 语音的语谱图可以看出 Martin

算法增强后的语音仍然残留有大量的残余噪声,而本文算法增强后的语音则几乎完全消除了噪声,这表明本文算法在消除噪声方面具有一定的优越性;同时采用数据场描述语音的时频分布时,由于清音段往往位于势能较大的浊音段周围,即清音段与浊音段相邻,当清音段在浊音段的作用力程之内时,清音段区域的时频点会受到浊音段区域时频点的影响,其势能也将得到提升,故在进行二值时频掩蔽时不易被切除,使得部分清音段得以保留,减少了语音失真。这一结果可以从图 4、5 中的椭圆虚线框中得到体现。不可否认,采用数据场方法增强后的语音在弱的高频成分上有一定的失真,但主观听觉表明这并不影响增强后语音的可懂度和清晰度。

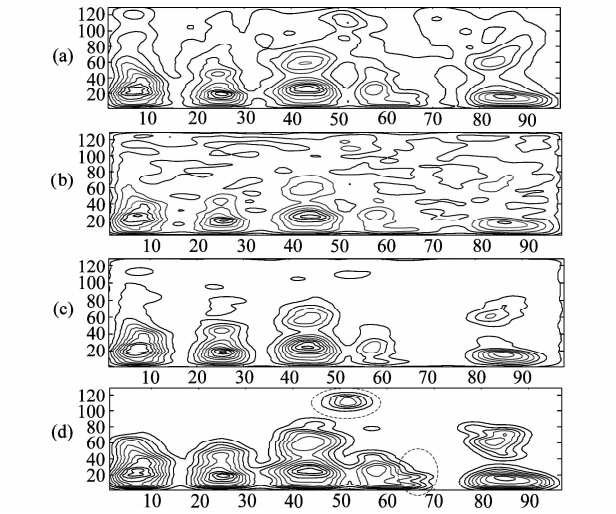


图 5　语音数据场分布图(a 干净语音、b 带噪语音、c Martin 算法增强后的语音、d 本文算法增强后语音)

Fig. 5　Speech signal in Data Field: (a) Original speech; (b) Noisy speech; (c) Speech enhanced with martin's algorithm; (d) Speech enhanced with proposed algorithm.

## 4　结论

本文首次将数据场的概念引入到语音信号处理中,提出了一种新的基于数据场的语音增强算法,实验表明语音数据场表示可以充分利用相邻时频点之间的信息,具有类似听觉感知分段的能力,能够更好的区分来自不同信源的信号,有利于区分语音和噪声,从而更好地消除噪声。语音测试表明,该算法确实可行有效,对数据场在语音信号处理其他方面的应用具有借鉴作用。例如在盲语音分离场合中可以用数据场对混合语音进行建模,然后将其融入到传统的分离算法框架当中(如独立分量分析等)以提高语音分离效果。

### 参考文献

[1] Lu. Y. and P. C. Loizou. A geometric approach to spectral subtraction[J]. Speech Communication, 2008, 50: 453-466.

[2] Ding H., I. Y. Soon, and C. K. Yeo. Over-Attenuated Components Regeneration for Speech Enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2004: 2014.

[3] S. Jo and C. D. Yoo. Psychoacoustically Constrained and Distortion Minimized Speech Enhancement [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2099-2110.

[4] I. Andrianakis. Bayesian algorithms for speech enhancement [D][Doctor thesis]. University of Southampton, 2007.

[5] D. L. Wang and G. J. Brown. Computational Auditory Scene Analysis: Principles, Algorithms and Applications [M]. IEEE Press/Wiley-Interscience, 2006: 2-18.

[6] Y. Shao, S. Srinivasan, Z. Jin, etal. A computational auditory scene analysis system for speech segregation and robust speech recognition[J]. Computer Speech and Language, 2010, 24: 77-93.

[7] 李德毅,杜鹢. 不确定性人工智能[M], 国防工业出版社, 2005: 207-212.
Li D. Y., Du Y. Artificial Intelligence with Uncertainty[M]. National Defense Press, 2005, 207-212. (in Chinese)

[8] Li D. Y., Liu K., Sun Y., etal. Emerging Clapping Synchronization From a Complex Multiagent Network With Local Information via Local Control [J]. IEEE Transactions on Circuits and Syetems-II: Express Brief, 2009, 56(6): 504-507.

[9] 淦文燕,李德毅,王建民. 一种基于数据场的层次聚类方法[J]. 电子学报, 2006, 34(2): 258-262.
Gan W. Y., Li D. Y., Wang J. M.. An Hierarchical Clustering Method Based on Data Fields [J]. ACTA ELECTRONICA SINICA, 2006, 34(2): 258-262. (in Chinese)

[10] 王树良,基于数据场与云模型的空间数据挖掘和知识发现[D][博士论文]. 武汉大学, 2002.
WANG S. L., Data Field and Cloud Model Based Spatial Data Mining and Knowledge Discovery[D][ PhD ]. Wuhan University, 2002. (in Chinese)

[11] R. Martin. 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.

[12] Hu Y. and P. Loizou. Subjective comparison of speech enhancement algorithms [C]. Proceedings of ICASSP, Toulouse, France, May 2006, 153-156.

### 作者简介

黄建军(1984-),男,解放军理工大学指挥自动化学院博士研究生,主要研究方向为语音增强和语音分离等。
E-mail: hjj954@ gmail. com

张雄伟(1965-),男,解放军理工大学指挥自动化学院教授,博士生导师,中国通信学会理事、学术委员会委员,主要研究方向为数字通信和多媒体信号处理等。

张亚非(1955-),男,解放军理工大学教授、博士生导师,主要研究方向为视觉信息处理。

淦文燕(1971-),女,解放军理工大学指挥自动化学院副教授,研究数据挖掘、数字水印和复杂网络等。

邹　霞(1979-),男,讲师,研究低速率语音编码和基于听觉感知的噪声抑制技术。