面向内容的语音信号压缩感知研究

高 畅 李海峰 马 琳
(哈尔滨工业大学计算机学院,黑龙江 哈尔滨 150001)

摘 要:压缩感知理论依据信号的稀疏性质进行压缩测量,将信号的获取方式从对信号的采样上升为对信息的感知,是信号处理领域的一场革命。本文提出一种基于非确定基字典(Uncertainty Basis Dictionary, UBD)对语音信号进行稀疏表示的方法,将压缩感知理论应用于对语音信号稀疏表示的压缩,并提出了基于求解线性规划问题的方法重构语音信号的算法。通过语音识别、话者识别和情感识别实验,从面向内容分析的角度,研究这种基于压缩感知理论的信息感知方法是否保留了语音信号的主要内容。实验结果表明,语音识别、话者识别和情感识别的准确率,与目前这些领域研究方法得到的结果基本一致,说明基于压缩感知理论的信息感知方法能够很好地获取语音信号的语义、话者和情感方面的信息。

关键词:压缩感知;语音信号;稀疏表示;线性规划;信息感知
中图分类号:TP391.42 **文献标识码**:A **文章编号**:1003-0530(2012)06-0851-08

Content-based Compressive Sensing for Speech Signal

GAO Chang LI Hai-feng MA Lin
(School of computer science and technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Compressive sensing theory compress measurements using sparsity of signal, changes the method of signal obtaining from signal sampling to information sensing, and is a revolution of signal processing. The speech signal is sparse represented based on Uncertainty Basis Dictionary proposed in this paper, the sparse representation of speech signal is compressed by compressive sensing theory, and proposes an speech signal reconstruction algorithm based on the method of solving linear programming problem. Through the experiments of audio, speaker and emotion recognition, we research that this information sensing method based on compressive sensing theory weather preserves the main content from the angle of content-based analysis. Experiment results show that the precision of audio, speaker and emotion recognition is general the same with methods in these research domain, and proves that it can acquire the audio, speaker and emotion information of speech signal using the information sensing method based on compressive sensing theory.

Key words: compressive sensing; speech signal; sparse representation; linear programming; information sensing

1 引言

在数字信号处理中,对信号进行稀疏表示,有利于信号的后续处理,从本质上降低了信号处理的成本,具有非常重要的意义,已经被应用于信号处理的各个领域[1]。

传统的信号稀疏表示方法是对信号基于正交变换基进行展开,离散小波变换就属于这种方法,能比较好地表示信号的局部特性,但是当信号与正交变换基不相适应时,往往达不到很好的稀疏表示效果。因此,在图像的稀疏表示中,提出了多尺度几何分析[2]的方法,主要解决了高维空间数据稀疏表示的问题,被认为是

收稿日期:2011-12-27;修回日期:2012-05-10
基金项目:语言语音教育部-微软重点实验室开放基金资助项目(HIT. KLOF. 2011XXX);中央高校基本科研业务费专项资金(HIT. NSRIF. 2012047)以及国家自然科学基金项目(61171186)的支持

小波兴起后的又一场革命,多尺度几何分析属于前沿的研究领域,其理论和算法还处在发展之中。过完备字典下的信号稀疏表示方法,最早由 Mallat 和 Zhang 在 1993 年首次提出[3],它采用过完备的冗余函数代替传统的基函数,为信号自适应地稀疏表示提供了方便,已经成为一个新的研究方向和热点,将会是继小波与多尺度几何分析方法后的又一个研究高潮。

2004 年 Donoho 与 Candes 提出了利用少量的可压缩性或稀疏性数据精确或近似精确地重构原始信号的压缩感知(Compressed Sensing,CS)理论,实验验证信号越稀疏,通过压缩感知重构的信号越精确[4-6]。CS 理论建立在信号稀疏表示的基础上,对信号的处理从信号采样的层面上升到对信息感知的层面,是信号处理领域的一场革命,它将信号稀疏表示的重要性提升到了一个新的高度,掀起了信号稀疏表示的一个新的研究热潮,并被应用于很多领域,如:压缩成像[7]、医学成像[8]、压缩编码[9]、通信[10]、雷达成像[11]、模拟/信息转换[12]和生物传感[13]等领域。

信号越稀疏,CS 理论重构的信号越精确,对 CS 理论在信号处理中的应用也越有利。对于语音这种变化比较复杂的信号,目前的稀疏表示方法并不能得到理想的效果,对基于 CS 理论的语音信号处理会产生影响。本文针对基于 CS 理论对语音信号压缩与重构后,是否保留了主要的语义、话者和情感信息等内容进行了研究。

目前,CS 理论在语音信号处理中的应用,主要集中在对语音信号的压缩与重构和语音编码等方面,如:Peyré 提出了基于最优基的 CS 算法,构建一个基于离散余弦变换的树形结构冗余字典,通过搜索少数几个最优基重构信号,对动物发出的声音进行重构,取得了很好的重构效果[14];Griffin 和 Tsakalides 利用不同的"函数基"对声音信号进行稀疏表示,对压缩感知在不同"函数基"对语音信号的重构性能进行了研究,其中在离散余弦变换基下的重构性能最佳[15]。Ciacobello 等人提出了基于 CS 理论的语音信号编码算法[16]。

当前对 CS 在语音信号处理中的应用研究从理论和实验结果上,都体现出其所具有的巨大潜力。CS 理论中随机性、稀疏性和不相关性等性质,在语音信号处理中有着重要的应用价值,将 CS 理论与语音信号处理技术结合具有广阔的前景。特别是,在 CS 理论框架下,将其对信息的感知特性,应用于语音编码和语音识别等领域,会对语音信号处理方法产生巨大的影响,从传统的基于信号采样的处理方式,转变为基于信息感知的方式。虽然 CS 理论在语音信号处理中的应用取得了很大的进展,但是目前还无人对基于 CS 理论重构语音信号,在面向语音内容方面保留的信息进行研究。本文提出了基于 CS 理论的语音信号的压缩与重构方法,通过语音识别、话者识别和情感识别实验,验证了基于 CS 理论的语音信息获取方法对语义、话者和情感信息等语音内容的影响。本文对语音、话者和情感的识别实验均采用经典的识别方法,其目的在于通过这些经典算法,对基于 CS 理论的语音信息感知方法在压缩与重构性能上进行评价。

本文的组织结构如下:第二部分简单介绍了 CS 理论的基本原理,第三部分提出了基于 CS 理论的语音信号重构算法,第四部分从主客观两个方面对重构语音信号的效果进行了评价,通过语音识别、话者识别和情感识别实验验证了基于 CS 理论的重构信号保留了语音信号的主要内容,最后对本文做了总结。

2 CS 理论基本原理

CS 理论将采样与压缩合并进行,以达到直接获取采样数据压缩表示的目的,这包括信号稀疏表示、压缩测量和重构算法三方面的内容[17],其中信号的稀疏表示是其他两个方面研究的前提和基础。

(1) 信号稀疏表示即是将信号经过某种数学变换,把信号投影到正交变换基上。若绝大多数的变换系数的绝对值都为零,则把变换以后的信号称为稀疏信号;若较小的变换系数占变换系数总数的 50% 以上,则把变换以后的信号称为近似稀疏信号,或可压缩信号,它是信号的一种简洁表示[18]。

一个长度为 N 的一维离散信号 $X=[x_1,x_2\ldots,x_N]^{\mathrm{T}}$,它的稀疏表示如公式(1):

$$X=\sum_{i=1}^{N}s_i\psi_i \quad \text{or} \quad X=\Psi S \tag{1}$$

其中,$\Psi=[\psi_1,\psi_2,\ldots,\psi_N]$,$\psi_i$ 为列向量,Ψ 即为稀疏基,可以根据原始信号特征灵活选取。常用的有离散余(正)弦变换基,离散小波变换基、Chirplet 基和 Fourier 变换基等。S 即为稀疏信号,是信号 X 在正交变换

基Ψ上的投影,是信号 X 的等价表示。采用合适的稀疏基Ψ,可以使信号的稀疏度尽量小,可以大大提高采样速度。但是对于语音信号,经过某种变换后并不是绝对稀疏的,而是可压缩的,所以对基于 CS 理论的语音信号后续处理会产生影响。

(2)CS 理论压缩测量是得到信号的压缩表示。这个过程通过原始信号向一个矩阵做投影来实现,该矩阵称为测量矩阵:$\Phi \in R^{M\times N}(M<<N)$,其中 M 为压缩感知过程中的测量次数,N 为信号的长度。未知信号 X 在该测量矩阵下的测量值为 y,测量值 y 即是信号的压缩表示,即:

$$y=\Phi X \tag{2}$$

由公式(1)和(2)将 y 进一步表示为公式(3)

$$y=\Phi X=\Phi\Psi S=\tilde{\Phi}S \tag{3}$$

其中 $\tilde{\Phi}$ 是由测量矩阵与稀疏基计算得到,被称为感知矩阵,是 $M\times N$ 维的,并且 M 要满足条件:$M\geqslant O(K\ln(N))$。同时,为了精确重构语音信号,感知矩阵要满足约束性等价(Restricted Isometry Property, RIP)条件[19]和不相关特性[17]。

压缩测量是对信号中信息的感知,决定了重构信号中保留的信息的多少,随着测量次数的减少,感知过程中所保留的信息也有所减少,当测量次数 $M<O(K\ln(N))$ 时,感知信息会丢失信号的主要内容。在本文中,我们用测量频率来表示本文提出的重构算法在每秒内的测量次数。

(3)CS 理论中由重构算法获得信号 X 或其稀疏表示。若重构的信号是稀疏的或可压缩的,并且测量向量 y 和感知矩阵 $\tilde{\Phi}$ 满足 RIP 条件,则可以通过求解最优 l_0 范数问题精确或近似精确地重构原始信号 X,如公式(4)所示

$$\hat{X}=\min\|X\|_0 \quad \text{s. t.} \quad \Phi\Psi^{\mathrm{T}}X=y \tag{4}$$

式中 $\|\ \|_0$ 为 l_0 范数,表示向量 X 中非零元素的个数。在最优化理论中,求解公式(4)是一个 NP-hard 问题,通常转化为 l_1 范数问题来求解。

目前,重构算法主要有匹配追踪(Match Pursuit, MP)类算法和基追踪(Basis Pursuit, BP)类算法,前者重构速度快,但是对类似语音信号这种结构复杂的信号,重构效果不好,后者虽然重构速度较前者慢,但是重构效果比前者好。

3　基于 CS 理论的语音信号压缩与重构方法

由前所述,语音信号在某种正交基或冗余字典下具有近似稀疏的特性,CS 理论可以直接获取近似稀疏信号的压缩表示并精确重构语音信号,针对具体问题,我们提出了基于 CS 理论的语音信号压缩与重构方法。实现过程如图 1 所示。

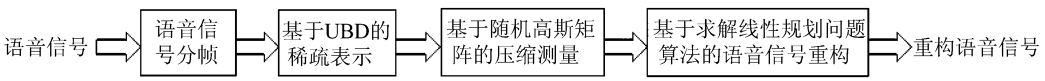


图 1　基于压缩传感理论语音信号重构过程

Fig. 1　The process of reconstructing speech signal based on compressive sensing theory

基于 CS 理论的重构语音信号面向内容的分析,主要包含四个部分。首先,对语音信号分帧;其次,基于聚类形成的 UBD,对语音信号进行稀疏表示;再次,通过随机高斯矩阵得到每一帧稀疏信号的压缩表示,即感知每一帧语音信号的信息;最后重构语音信号,通过对重构信号的语音、话者和情感识别实验,分析重构信号的内容。

3.1　语音信号的稀疏表示

对于语音这种比较复杂的信号,可以通过某种正交变换得到其近似稀疏表示,这种方法计算复杂度低,运行时间短,但是稀疏表示的效果比较差;此外,还可以通过基于某种冗余字典的非正交分解,得到其稀疏表示,这种方法对语音信号稀疏表示的效果,比正交变换的方法好,但是通常冗余字典包含大量的原子,因此稀疏分解过程计算量较大,需要较长时间,这些缺点阻碍了其在对信号稀疏表示中的应用推广。

针对上文论述中提到的对语音信号稀疏表示方法中存在的问题,本文提出了一种基于 UBD 的语音信号稀疏表示方法,该方法首先对训练集中的语音信号进行分帧,然后对分帧后的语音信号进行聚类,将每一类的聚类中心作为字典的原子,最后通过正交匹配追踪(Orthogonal Matching Pursuit, OMP)算法,从 UBD 中寻

找原子对语音信号进行稀疏表示。

下面给出本文提出的对语音信号稀疏表示方法中非确定基字典的构造算法：

UBD 构造算法（Uncertainty Basis Dictionary Construction Algorithm）：

1）构造训练集 X，$X=[x_1\ x_2\cdots x_n]$，x_i 为列向量，由一帧语音信号构成，并设定字典中原子数量 k，距离变化率阈值 t，本文将其设为0.005，计算训练集中的均值向量 m，作为初始聚类中心；

2）将每一个聚类中心按式 $m+e\times m$ 和 $m-e*m$ 分解为新的聚类中心，本文将权值 e 设为0.01；

3）计算 X 中每个列向量与聚类中心的距离，根据最小距离准则将其划分到相应类，重新计算聚类中心，并计算 X 中所有聚类的类内距离之和，记为 td；

4）如果是重新划分聚类后的首次循环，则初始类内距离 $pd=td$，返回2），否则，转入5）；

5）计算类内距离变化率 $ct=(pd-td)/pd$，如果 $ct<t$，并且聚类数量小于 k，重新设置距离变化率阈值 $t=w\times t$，并且重置 $e=w\times e$，本文中权值 $w=0.75$，返回2）；如果 $ct<t$，并且聚类数量大于 k 时，舍弃多余的类，返回3），根据各类类内距离和，从大到小舍弃；如果 $ct\geqslant t$，则 $pd=td$，返回2）；

6）如果 $ct<t$，并且聚类数量等于 k，则聚类结束。

UBD 由语音信号自适应形成，字典中包含的原子数量远小于一般冗余字典（如 Gabor 字典）中的原子数量，因此计算量更小，运行时间更短，克服了基于冗余字典的信号稀疏表示方法的缺点，本文的实验中使用的 UBD 包含 1024 个原子。分别使用基于 UBD 和 Gabor字典，在相同的环境下对同一段语音信号进行稀疏表示，稀疏分解得到的原子重构语音信号的信噪比达到 20dB 时，基于 Gabor 字典的稀疏表示方法所用时间，是本文提出的基于 UBD 的稀疏表示方法的 40 倍，并且二者所用原子数量基本相同。

3.2 感知矩阵的选择

CS 理论中，所选测量矩阵要与正交变换基组成的感知矩阵满足 RIP 条件。选择合适的感知矩阵，主要是选择与正交变换基不相关的测量矩阵。

本文选择 $M\times L$ 的随机高斯矩阵为测量矩阵（其中，L 为信号的长度），它的每一个元素都满足 $N(0,1/L)$ 的独立正态分布，任意列都不相关，若正交变换基是一个单位矩阵 I，那么传感矩阵 $\tilde{\Phi}=\Phi I=\Phi$，它的任意列不相关，满足 RIP 条件。Baraniuk 在文献[17]中证明，选择满足 $N(0,1/L)$ 的独立正态分布的高斯随机矩阵作为测量矩阵，构成的感知矩阵 $\tilde{\Phi}$ 能以很大的概率满足 RIP 条件。

3.3 基于求解线性规划问题的语音信号重构算法

语音信号是一种随时间变化的非平稳随机信号，而在 10～30ms 内，可以认为该时间段内的语音信号是平稳的。所以对语音信号进行重构时，先对语音信号进行分帧处理，然后对每一帧语音信号进行重构。

对压缩测量获取的测量值重构语音信号，可以转化为一个求解最小 l_1 范数问题，如公式(5)。

$$\hat{X}=\min\|X\|_1\ \text{ s. t. }\ \Phi\Psi^{\mathrm{T}}X=y \tag{5}$$

解公式(5)这样的方程，可以归为解线性规划问题的最优化计算的过程，一般可以用单纯形法和内点法等方法求解[20]。单纯形法是解线性规划问题的通用解法，在基本可行解有限的情况下，通过有限次迭代，一定可以得到最优解，因此，本文采用单纯形法。

如公式(6)所示，是线性规划问题的标准形式

$$\min\quad z=c_1x_1+c_2x_2+\ldots+c_nx_n\quad\text{subject to}\quad\begin{cases}a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1\\a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2\\\ldots\ldots\\a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n=b_m\\x_j\geqslant 0(j\geqslant 1,2,\ldots,n)\end{cases} \tag{6}$$

即

$$\min C^{\mathrm{T}}X\quad\text{subject to}\quad AX=b\ \text{and}\ \ X\geqslant 0 \tag{7}$$

其中 X 称为决策向量，即线性规划问题的解，通过它得到重构语音信号，C 是目标函数的系数向量，X 和 C 都是 n 维列向量。b 为 m 维列向量，是约束方程组的常数向量，A 为 $m\times n(m\leqslant n)$ 的矩阵，是约束方程组的系数矩阵，它由一个 $m\times m$ 的基矩阵 B 和一个 $m\times(n-m)$ 的非基矩阵 W 构成，即 $A=(B,W)$。

下面给出用求解线性规划问题的单纯形法重构语音信号算法：

LPCS 重构算法（linear programming compressive sensing reconstruction algorithm）：

1）传感矩阵 $\tilde{\Phi}$ 构成系数矩阵，$A=(\tilde{\Phi},-\tilde{\Phi})$，$x^{(0)}=(b_1^{(0)},b_2^0,\ldots b_m^0,0,\ldots,0)^{\mathrm{T}}$ 为初始常数向量，其中 b_i 为 CS 理论框架下的测量值，m 为测量次数；

2）对所有非基系数对应部分，计算 $\sigma_j=c_j-$

$\sum_{i=1}^{m} c_i a_{ij}^{(0)}, j = m+1, m+2, \ldots, n$, 构成向量 Δ,其中最大值为 σ_k,如果 $\sigma_k \leqslant 0$,$X^{(0)}$ 为最优解,否则转入 3);

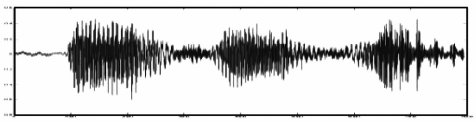
3) 如果 $a_{ik}^{(0)} \leqslant 0 (i = 1, 2 \ldots, m)$,无最优解,生成一个新的感知矩阵,重新获得压缩测量值,并返回到 1),否则转入 4);

4)计算 $\min\left\{\frac{b_i^{(0)}}{a_{ik}^{(0)}} \mid a_{ik}^{(0)} > 0\right\} = \frac{b_l^{(0)}}{a_{lk}^{(0)}} (l = 1, 2, \ldots, m)$,将系数矩阵的第 l 列和第 k 列互换,形成新的系数矩阵,决策向量 X 中的第 l 个和第 k 个互换,得到新的基本可行解 $X^{(1)}$;

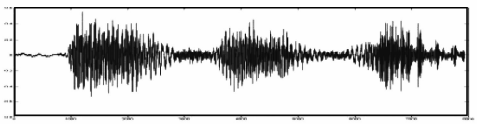
5)重复步骤 2)至 4),直至满足所有的 $\sigma_j \leqslant 0$,得到最终的可行解 X^*;

6)可行解 X^* 要求 $x_i \geqslant 0 (i = 1, 2, \ldots, m, m+1, \ldots, n)$,故将可行解 X^* 中的基变量 $x_i (i = 1, 2, \ldots, n/2)$ 减去非基变量 $x_j (j = n/2+1, n/2+2, \ldots, n)$,得到最终的重构语音信号。

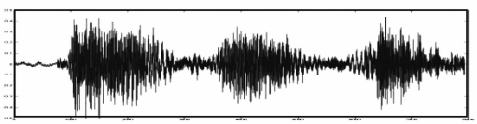
图 2 是在不同测量频率下对语音信号重构得到的波形图。



(a) 采用经典方法测量得到的语音信号(相当于测量频率=16KHz)
(a) speech signal measured using classical method (measuring frequency=16KHz)



(b) 重构语音信号(测量频率=8KHz)
(b) Reconstructed speech signal (measuring frequency=8KHz)



(c) 重构语音信号(测量频率=4KHz)
(c) Reconstructed speech signal (measuring frequency=4KHz)

图 2　不同测量频率下基于 CS 理论的重构语音信号

Fig. 2　The reconstructed speech signal under different measure frequency based on compressive sensing theory

在图 2 中,测量频率为 4KHz 时,重构信号已经比较接近采用经典方法测量得到的语音信号,测量频率为 8KHz 时,基于求解线性规划问题的语音信号重构算法可以精确重构语音信号,直接比传统的方法节约 50% 的采样数据。

4　实验结果及性能分析

为了验证基于 CS 理论的语音信息感知方法在内容保持方面的效果,我们设计了四个实验。通过基于 CS 理论重构语音信号的效果评测,对本文提出的重构算法的有效性进行评价。从面向语音信号内容的角度,通过基于 CS 理论的重构语音信号语义信息、话者信息和情感信息保留效果的评测实验,验证了基于 CS 理论的语音信息感知方法是否保留了语音信号中的主要信息。在本文的实验中,情感识别实验采用的是作者所在实验室录制的情感语料库,其他实验采用截取自新闻联播播音员播音的音频文件,音频文件的格式为 wav,采样频率为 16 KHz(等价于 CS 理论框架下的测量频率为 16KHz),16 bit 量化,重构时的测量频率范围为 8KHz ~ 4KHz,识别实验所使用的特征为 39 维 MFCC 特征(12 维 MFCC 和对数能量,以及其对应的一阶、二阶差分)。

4.1　基于 CS 理论重构语音信号的效果评测实验

本实验对采用经典方法测量得到的语音信号,用本文的方法进行重构,在主客观两方面对重构语音信号进行了评价。

根据语音信号的短时平稳的特性,对语音信号进行分帧重构,对重构后语音信号质量采用分段信噪比(SNRSEG)进行评价,具体计算方法如公式(8)所示

$$SNRSEG = \frac{1}{FN}\sum_{i=1}^{FN} 10\log_{10}\left(\frac{\|x_i\|_2^2}{\|x_i - \hat{x}_i\|_2^2}\right) \qquad (8)$$

其中 FN 为整个语音信号的总帧数,x_i 为一帧原始语音信号,$\hat{x}_i$ 为对应帧的重构语音信号。

实验中,使用总时长为 30 分钟的语音信号,在不同帧长和不同测量频率下,对采用经典方法测量得到的语音信号进行重构,计算重构语音信号分段信噪比,如表 1 所示。

表 1　重构语音信号的分段信噪比

Tab. 1　The SNRSEG of reconstructed speech signal

帧长(ms)	测量频率(KHz)				
	8	7	6	5	4
30	22.01	19.31	16.36	16.10	12.62
20	23.21	21.13	18.71	16.00	12.49
16	24.24	21.20	18.26	15.18	12.09
10	23.82	20.52	16.82	14.02	11.01

从表 1 中可以看出,在帧长为 20ms、16ms 和 10ms,测量频率为 7KHz 时,重构语音信号分段信噪比达到 20 以上,本文算法可以精确的重构语音信号。

本实验用诊断押韵测试(diagnostic rhyme test, DRT)作为重构语音信号质量的主观度量,简称 DRT 得分,它是反映话音清晰度或可懂度的一种测试方法。实验中使用 96 对截取自新闻联播播音员播音中同韵母同音调的字测试,如“对”和“退”等,帧长为 16ms。测试中让测试者每次听到一对韵字中的某一个,判断是哪一个字,全体测试者判断正确的百分比作为 DRT 得分,具体得分如表 2 所示。

表 2 重构语音信号的 DRT 得分

Tab. 2 The DRT of reconstructed speech signal

测量频率(KHz)	8	7	6	5	4
DRT 得分	100.0	100.0	99.5	99.5	98.5

一般认为, DRT 得分为 90 以上时,重构语音信号的可懂度达 100%。从表 2 可以看出,在帧长为 10ms,测量频率为 4KHz 时,整个句子的可懂度达 100%,重构语音信号在主观听感上,可以达到让人满意的效果。

4.2 基于 CS 理论重构语音信号的语义信息保留效果评测实验

本实验使用作者所在实验室搭建的基于隐马尔科夫模型(Hidden Markov Model, HMM)的大词表连续语音识别平台,对重构语音信号做了语音识别的实验,验证基于 CS 理论的语音信息感知方法是否保留了语音信号的语义信息。

在实验中,选取了 902 句非特定女播音员和 956 句非特定男播音员的音频文件,重构这些语音信号,作为训练语料,再从这些重构的语音信号中选取 136 句非特定女播音员和 127 句非特定男播音员的语音信号作为测试语料。识别结果如表 3 所示。

表 3 不同测量频率下重构语音信号的语音识别率

Tab. 3 The accuracy of speech recognition of reconstructed speech signal under different measuring frequency

测量频率(KHz)	16	8	6	4
女播音员	97.92%	97.91%	98.22%	98.21%
男播音员	97.93%	97.83%	98.46%	98.05%
女+男播音员	98.03%	98.01%	98.42%	98.31%

从表 3 中可以看出,在 CS 理论框架下,用较低的测量频率(低于 Nyquist 采样定理要求的采样频率),重构的语音信号可以达到采用经典方法测量得到的语音信号的识别效果,保留了主要的语义信息。

4.3 基于 CS 理论重构语音信号的话者信息保留效果评测实验

本实验通过作者所在实验室搭建的话者识别实验平台,验证基于 CS 理论的语音信息感知方法是否保留语音信号的话者信息。该实验平台采用基于高斯混合模型(Gaussian Mixture Model, GMM)的说话人识别技术,高斯混合数为 32。

本实验分别选取了截取自新闻联播中 300 句男播音员和 300 句女播音员的语音信号作为训练语料,30 句男播音员和 30 句女播音员的语音信号作为测试语料。训练阶段采用期望最大化(Expectation Maximization, EM)方法,为每种类型的样本建立一个混合高斯模型。识别阶段计算语音特征与每个模型的似然概率,似然概率最大的即判断为相对应的类型。识别准确率如表 4 所示。

表 4 不同测量频率下重构语音信号的话者识别率

Tab. 4 The accuracy of speaker recognition of reconstructed speech signal under different measuring frequency

测量频率(KHz)	16	8	6	4
女播音员	100%	100%	100%	99.97%
男播音员	100%	100%	100%	100%
女播音员+男播音员	100%	100%	100%	99.98%

在表 4 中可以看出,在测量频率为 6KHz 和 8KHz 时,基于 CS 理论重构的语音信号,与采用经典测量方法得到的语音信号在话者识别准确率上一样,都为 100%,随着测量频率的降低,对女播音员的话者识别准确率略有下降,男播音员的话者识别准确率仍为 100%,可以认为,在 CS 理论框架下,重构后的语音信号可以达到理想的话者识别准确率,保留了主要的话者信息。

4.4 基于 CS 理论重构语音信号的情感信息保留效果评测实验

本实验使用作者所在实验室搭建的基于支持向量机(Support Vector Machine, SVM)的情感识别平台,验证基于 CS 理论的语音信息感知方法是否保留了语音信号的情感信息。

实验使用作者所在实验室录制的情感语料库,分为男性话者和女性话者,包含高兴、悲伤、生气和惊奇四种情感类型,每一类都有 100 个音频文件,选取其中

的 80 句作为训练语料,共 320 句。采用 SVM 方法,利用训练样本为每一类训练一个模型。每类中余下的 20 句作为测试语料,共 80 句。利用模型将提取的特征对应并产生返回值,与四种情感类型进行比较,最相近的情感类型作为待测音频的情感类别。实验进行 5 次,每次实验都随机地选择 80 句训练语料和 20 句测试语料,取 5 次实验识别率的平均值作为最终结果。识别准确率如表 5 所示。

表 5　不同测量频率下重构语音信号的情感识别准确率

Tab. 5　The accuracy of emotion recognition of reconstructed speech signal under different measuring frequency

测量频率(KHz)	16	8	6	4
男性话者	63.8%	63.5%	62.4%	62.9%
女性话者	66.3%	67.7%	65.1%	67.3%
男+女性话者	65.1%	65.3%	63.7%	65.2%

从表 5 中可以看出,在 CS 理论框架下,用较低测量频率重构的语音信号在情感识别准确率上,与采用经典测量方法得到的语音信号的情感识别准确率基本一致。可以认为,重构语音信号可以达到采用经典方法测量得到的语音信号的识别效果,保留了主要的情感信息,并且与目前该领域得到的识别结果基本一致。

5　结束语

本文在 CS 理论框架下,提出一种基于 UBD 的语音信号稀疏表示方法,并给出了感知矩阵的选择方法,用解决最优化问题中线性规划问题的单纯形法重构语音信号,从主客观两方面对重构效果进行了评价。并且,本文对重构语音信号做了语音识别、话者识别和情感识别实验,验证了重构语音信号保留了语音信号的主要信息。今后还需要继续改进重构算法,使重构语音信号在主观听感上,能更加接近原始语音信号。此外,在 CS 理论框架下,感知矩阵存在信号压缩表示的特性,保留了信号的主要信息,可将这一特性运用到对信号的特征提取中,通过这个特性可以加快信号后续处理的速度。

参考文献

[1] M. Elad. Sparse and redundant representations from theory to application in signal and image processing[M]. New York, USA: Springer Press, 2010.

[2] S. Mallat. A wavelet tour of signal processing[M]. Third Edition. San Diego, USA: Academic Press, 2009.

[3] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries[J]. IEEE Transactions on Signal Process, 1993,41(12):3397-3415.

[4] D. Donoho. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006,52(4):1289-1306.

[5] E. Candès and M. Wakin. An introduction to compressive sampling[J]. IEEE Signal Processing Magazine, 2008, 25(2):14-20.

[6] 石光明,刘丹华,高大化,等. 压缩感知理论及其研究进展[J]. 电子学报. 2009,37(5):1070-1081.
Shi Guangming, Liu Danhua, Gao Dahua, et al. Advances in Theory and Application of Compressed Sensing [J]. Acta Electronica Sinica. 2009,37(5):1070-1081. (in Chinese)

[7] M. F. Duarte, M. A. Davenport and D. Takhar, et al. Single-pixel imaging via compressive sampling[J]. IEEE Transactions on Signal Processing, 2008,25(2):83-91.

[8] M. Lustig, D. L. Donoho and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging[J]. Magnetic Resonance in Medicine, 2007,58(6): 1182-1195.

[9] E. Candes and T. Tao. Decoding by linear programming [J]. IEEE Transactions on Information Theory, 2005,52 (12):4203-4215.

[10] G. Taubock and F. Hlawatsch. A compressed sensing technique for OFDM channel estimation in mobile environments: Exploiting channel sparsity for reducing pilots [C]. IEEE International conference on Acoustics, Speech and Signal Processing. Washington D. C. ,USA,2008: 2885-2888.

[11] R. Baraniuk and P. Steeghs. Compressive radar imaging [C]. IEEE Proceedings od the Radar Conference. Washington D. C., USA, 2007:128-133.

[12] S. Kirolos, J. Laska, and M. Wakin, et al. Analog-to-information conversion via random demodulation[C]. Processing of the IEEE Dallas Circuits and Systems Workshop. Washington D. C., USA, 2006:71-74.

[13] M. Sheikh, O. Milenkovic and R. Baraniuk. Designing compressive sensing DNA microarrays[C]. Processing of

IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing. Washington D. C., USA. 2007:141-144.

[14] G. Peyré. Best basis compressed sensing[J]. IEEE Transactions on Signal Processing,2010,58(5):2613-2622.

[15] A. Griffin and P. Tsakalides. Compressed sensing of audio signals using multiple sensors[C]. Proc. of EUSIPCO 2008,2008.

[16] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen and M. Moonen. Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction[J]. IEEE Signal Processing Letters, 2010,17(1):103-106.

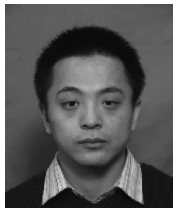
[17] R. Baraniuk. Compressive sensing[J]. IEEE Signal Processing Magazine. 2007,24(4):118-121.

[18] B. A. Olshausen, D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural image[J]. Nature, 1996,381(6583):607-609.

[19] E. Candès, T. Tao. Decoding by linear programming[J]. IEEE Transactions on Information Theory. 2005,51(2): 4203-4215.

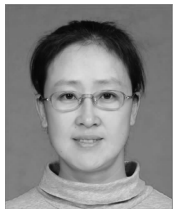
[20] 陈宝林. 最优化理论与算法[M]. 第 2 版. 北京:清华大学出版社. 2005:182-204.
Chen Baolin. Optimization Theory and Algorithms[M]. Beijing: Tsinghua University Press. 2005:182-204. (in Chinese)

作者简介



高 畅(1981-),男,河北秦皇岛,哈尔滨工业大学博士研究生,研究方向为压缩感知和音频信息检索。
E-mail:gaochang1981@yahoo.cn

李海峰(1969-),男,黑龙江哈尔滨,哈尔滨工业大学教授,博士生导师。1997 年毕业于哈尔滨工业大学,获得电磁测量技术及仪器博士学位;2002 年毕业于法国巴黎第六大学,获得计算机、通讯与电子科学博士学位。主要研究领域包括:数字媒体信息处理技术、智能信息处理与人工神经网络、认知科学与方法、自然人机交互技术、智能化测量技术等。
E-mail: lihaifeng@hit.edu.cn



马 琳(1967-),女,辽宁沈阳。哈尔滨工业大学计算机学院副教授,硕士生导师。主要研究方向为智能信息处理、图像处理和认知科学。
E-mail:malin-li@hit.edu.cn