

doi: 10.12052/gdutxb.180043

一种增量式学习的语音字典构造方法

滕少华, 宋 欢, 霍颖翔, 张 巍

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 爆炸式增长的语音数据为存储与传输带来极大困难, 现有方法难以实时应对海量语音频域数据. 因此本文提出一种增量式学习的语音字典构造方法, 该方法先将语音时域信号经短时傅里叶变换处理后转换为各窗频谱幅值, 再将高维空间向量投影到低维空间, 并以字典中的少数基向量线性拟合当前窗向量. 进而通过存储基向量的标识和拟合系数完成对当前窗向量的存储, 把无法拟合的窗向量经处理后加入字典, 实现增量式学习. 解压过程依据用户请求将字典中指定条目经线性拟合实现. 实验结果表明, 本方法能大幅度压缩语音频谱包络, 适用于受带宽限制下实时高采样率的流式语音数据, 与同类算法相比, 在保证还原质量的情况下, 能对信号的存储空间以及传输带宽进行大幅度的压缩.

关键词: 语音压缩; 语音解压; 实时处理; 流式数据; 增量学习; 稀疏字典学习

中图分类号: TP301 **文献标志码:** A **文章编号:** 1007-7162(2018)03-0029-08

An Incremental Learning Approach in Voice Compression via Sparse Dictionary Learning

Teng Shao-hua, Song Huan, Huo Ying-xiang, Zhang Wei

(School of Computers, Guangdong University of Technology, Guangzhou, 510006, China)

Abstract: The explosive growth of audio streams brings difficulties in storage and transmission; however, many methods could not give high compression ratio while keeping the quality. In order to solve this problem, the proposed method compresses amplitude spectrum of voice by constructing a dynamic sparse voice dictionary based on incremental learning. It calculates amplitude envelopes spectrums via Short-Time Fourier Transform (STFT) firstly, and then it uses a dictionary to fit each envelope by projecting high dimensional vectors to several 2D planes. In addition, it minimizes the number of dictionary items and therefore can store the parameters of linear interpolation instead of spectrums. Otherwise, if the fitting step above fails, it will store this window of spectrum directly. By using dictionary and parameters of linear interpolation, it can reconstruct the spectrum efficiently in decompressing process. The results of experiments show that comparing with other methods, the proposed method gives high compression ratio as well as better accuracy in decompressing, and adapt to live voice stream encoding with high sampling rate.

Key words: voice compression; voice decompression; real-time processing; streaming data; incremental learning; sparse dictionary learning

各种各样的数据爆炸式地产生, 图片、语音、视频等海量多媒体数据给远程传输、存储带来了巨大的困难和挑战. 因此在传输和存储之前对数据进行压缩很有必要. 通过压缩, 大幅减少信号存储空间以及传输带宽, 并保证高重构质量, 是急需解决的研究

与工程应用热点^[1-6].

考虑到现实多媒体数据所具有稀疏性质, 一种可行的思路为利用稀疏字典学习方法对语音数据进行压缩. 稀疏字典学习^[7]包括稀疏表示和字典学习两阶段, 稀疏表示的本质是利用尽可能少的资源表示

收稿日期: 2018-03-07

基金项目: 国家自然科学基金资助项目(61402118, 61673123, 61603100, 61702110); 广东省科技计划项目(2015B090901016, 2016B010108007); 广东省教育厅项目(粤教高函〔2018〕1号, 粤教高函〔2015〕113号, 粤教高函〔2014〕97号); 广州市科技计划项目(201604020145, 2016201604030034, 201508010067, 201604046017)

作者简介: 滕少华(1962-), 男, 教授, 博士, 主要研究方向为数据挖掘、网络安全、协同计算、大数据.

通信作者: 张巍(1964-), 女, 副教授, 主要研究方向为协同计算、数据挖掘、网络安全、大数据. E-mail: weizhang@gdut.edu.cn

更多的知识;字典是数据特征的集合.目前,稀疏字典学习方法被广泛运用在多媒体数据的识别、降噪、增强、检索^[8-11]等方面.

针对不断增长的实时语音数据,本文结合稀疏字典学习方法,提出一种面向增量式学习的语音字典构造方法,增量学习有助于大规模数据的特征学习^[12].该方法的核心是通过增量式学习构建一个字典,令字典中的基向量编号 $\mathbf{G}$ 结合相应拟合系数 $\mathbf{W}$ 组成参数对稀疏表示大部分频域信号,从而达到压缩目的.具体流程为:

(1) 先将语音时域信号经短时傅里叶变换STFT (Short-Time Fourier Transform)处理后转换为各窗频谱幅值,然后基于喉头仪信号找到各窗内的共振峰,通过多项式拟合方法得到各窗频谱的包络.

(2) 将包络作为高维空间向量投影到低维空间,并以字典中的少数基向量线性拟合当前窗向量.进而通过存储基向量编号 $\mathbf{G}$ 和拟合系数 $\mathbf{W}$ 完成对当前窗向量的存储;把无法拟合的窗向量利用LRU的策略加入字典,实现增量式学习.

(3) 解压过程依据用户请求将字典中指定基向量经线性拟合实现.

论文共分为五部分.第一部分为引言;第二部简单介绍了稀疏字典学习的内容.第三部分详细介绍本文提出的字典学习方法以及压缩应用.第四部分列出了在公开数据集上的实验结果.实验结果表明,该方法能大幅度压缩语音频谱包络的数据量,适用于受带宽限制下实时高采样率的流式PCM语音信号.该算法与同类算法相比,有更好的压缩效果.最后,第五部分对全文进行了总结并阐述了下一步工作.

1 相关工作

关于稀疏字典学习,国内外研究学者进行了大量的研究工作,已有许多研究成果.稀疏字典学习的目标是找到一个字典 $\mathbf{D} \in \mathbb{R}^{n \times z}$,能够稀疏表示原信号 $\mathbf{Y} \in \mathbb{R}^{n \times t}$:

$$\mathbf{Y} \approx \mathbf{D}\mathbf{X}. \quad (1)$$

其中 $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_z] \in \mathbb{R}^{n \times z}$, $\mathbf{d}_i \in \mathbb{R}^{n \times 1}$ 称为字典中的一个基. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t] \in \mathbb{R}^{z \times t}$ 是 $\mathbf{Y}$ 在字典 $\mathbf{D}$ 上的稀疏因子, $\mathbf{x}_i$ 是严格 L 稀疏的,即 $\|\mathbf{x}_i\|_0 \leq L$.这表明原信号 $\mathbf{Y}$ 中每一个采样点都能通过字典中的少数基的线性拟合来近似表示.

在稀疏字典学习中,求解稀疏因子的方法一般是匹配追踪(MP)或者正交匹配追踪(OMP).同样该方法也被用于字典训练过程中,字典训练的目的是

在当前稀疏度的标准下获得一个最小期望的字典 $\hat{\mathbf{D}}$,它能够稀疏表示当前训练数据^[13]:

$$\hat{\mathbf{D}} = \underset{\mathbf{D}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad \text{s.t. } \forall j \in \{1 : z\}, \|\mathbf{d}_j\|_2 = 1. \quad (2)$$

常见的字典学习方法是K-SVD^[14]、在线字典学习(ODL)^[15]等.文献[16]基于OMP算法和KSVD算法,在稀疏参数因子中提出池化概念和抽样选择方法,结合SVM对语音数据进行分类,分类结果比利用MFCCs作为特征的情况下准确率高,但未考虑实时在线应用的情况.文献[17]提出一种基于图像块数据的交替迭代方案,通过严格的收敛分析证明该方法有全局收敛的特性,并且通过实验证明获得相同效果下相比KSVD具有更低的计算开销.文献[18]提出一种分层学习结构,通过贡献字典原子能够减少训练一个大的混合字典的计算开销,但是尚未考虑到标签内在特性,文献[19]提出一种新的距离测量方法,联合KSVD稀疏字典学习方法运用于图片相似度检测,利用一个已知图片信号的信息去显著压缩另一个与之相似的图片信号的信息,两者越相似压缩比越高.文献[20]则通过K-SVD分解字典学习方法,改进了单通道语音增强算法.文献[21]提出了基于过完备字典稀疏表示的自适应语音增强方法,利用K-SVD算法训练字典,对含噪语音采用正交匹配追踪算法进行系数分解,再通过系数重构得到语音信号.

由于K-SVD算法在更新字典过程中,每次迭代都需要所有的训练样本集进行计算,所以基于K-SVD算法的字典更新操作过程对计算空间、存储空间要求较高.而ODL方法则是在每次迭代过程中使用一个训练样本去更新字典,文献[15]在彩色图像恢复中提出了在线字典学习算法.联合LARS-Lass算法,在字典的每次迭代更新只处理一个样本,从而达到计算速度快的目的.文献[22]在降噪、提取脉冲信号的应用中,借助压缩感知理论,通过学习自适应字典稀疏提取混合在大量噪声背景下的脉冲信号,该方法相比传统方法具有自适应调整的能力以及高稀疏性等优势.文献[23]提出利用MFCCs特征以及结合在线字典学习方法的分类算法,该算法通过字典去除了使用不同分类器带来的冗余信息,降低了高维空间向量的运算复杂度,对音乐和语音信号达到了完全分离的效果.

考虑到当前互联网研究与应用中,语音数据大量的产生,本文综合了相关理论研究,提出一种增量式语音字典构造学习方法,该方法基于语音信号的稀疏性^[24],利用冗余稀疏字典学习和增量学习的策

略对语音信号频谱包络进行流式压缩编码与解码的研究实验。

2 增量式字典构造方法

2.1 预备知识

本文涉及了几个概念,统一在此描述。

定义1 窗: 引用STFT^[25]中定义,设音频PCM序列为 L ,将 L 按照固定采样点数 $2n$ 和步长 θ 划分成多个时间片 $l_1, l_2, l_3, \dots, l_t$,其中步长 θ 表示两个PCM序列的起始位置之差。且定义 l_i 为第 i 窗。

定义2 包络: 设通过STFT获取第 i 窗 l_i 的峰值谱为 p_i ,将 p_i 中各共振峰用光滑曲 m_i 线连接起来,则定义曲线 m_i 为第 i 窗的频谱包络。

定义3 空间权值向量: 向量 $V = (v_1, v_2, \dots, v_n)$ 定义为 n 维空间权值向量, v_i 是 n 维空间上第 i 维的权值。

由于人耳对不同频率的音频信号听觉灵敏度不同,因此在语音频域模值包络中,不同频段的模值应具有不同的权值。

定义4 带权值的向量点积计算公式: 假设 n 维空间中各维权值为 $V, a = (a_1, a_2, \dots, a_n), b = (b_1, b_2, \dots, b_n)$ 为 n 维空间中的任意两个向量,则 a, b 的点积定义为

$$\langle a, b \rangle_V = \sum_{i=1}^n a_i \cdot b_i \cdot v_i. \quad (3)$$

定义5 带权值的向量模值公式: 假设 n 维空间中各维权值为 $V, a = (a_1, a_2, \dots, a_n)$ 是 n 维空间上的向量,则 a 的带权值的向量模值为

$$\|a\|_{(V)} = \langle a, a \rangle_V^{1/2}. \quad (4)$$

定义6 带权值的余弦公式: 假设 n 维空间中各维权值为 V, a, b 为 n 维空间中的任意两个向量,则 a, b 之间夹角的余弦值的计算式定义为

$$\cos_V(a, b) = \frac{\langle a, b \rangle_V}{\|a\|_{(V)} \cdot \|b\|_{(V)}}. \quad (5)$$

定义7 稀疏向量的压缩表示方式: 设有 n 维的稀疏向量 μ ,若有向量 G, W 和函数 f ,

$$\begin{aligned} G &= (g_i | i \in \{1 : \mu_0\}, \forall i \neq \tau, g_i \neq g_\tau). \\ f(G, W, i) &= \begin{cases} 0, & i \notin G; \\ \lambda_{\argmin_{\beta} |i - g_\beta|}, & i \in G \end{cases} \\ W &= (\lambda_i | i \in \{1 : \mu_0\}). \end{aligned} \quad (6)$$

使得 $\forall i, f(G, W, i) = \mu_i$,则称为使用编号 G 与系数 W 对 μ 进行压缩表示。

2.2 包络压缩总体框架

本节将结合语音压缩描述字典构造方法的总体

框架以及数据流向问题。首先需要说明的是,本方法直接求解编号 G 与系数 W ,从物理意义上,相对于ODL算法,相当于使用编号 G 与系数 W 间接表示 X 中的一个元素 x_i 。并且本方法不一次性处理整体 Y ,而是通过依次处理 Y 中的第 α 个包络 m_α ,来实现增量式压缩。对语音时域信号经STFT处理后得到了各窗频谱幅值,然后按照每个窗的输入顺序,利用字典中基向量进行拟合。若拟合匹配成功,则记录字典编号 G 和对对应系数 W ;若无法拟合则加入字典并写入文件,若此时字典未满,则直接加入字典;若此时字典容量已满,则选用LRU策略进行字典更新,然后再进行一次拟合匹配。总体框架图如图1所示。

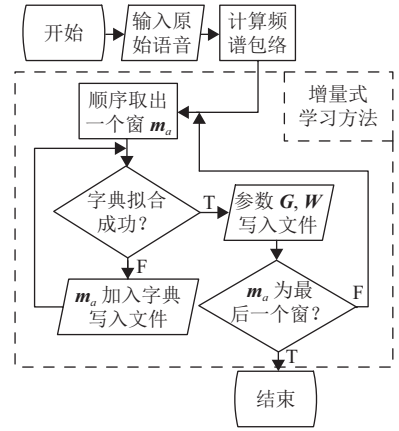


图1 总体框架图

Fig.1 System frame diagram

2.3 匹配与拟合

本部分描述本方法在字典学习和更新过程中利用所使用的高维向量个性化匹配算法,以下是基本实现方式。

2.3.1 二向量拟合单向量

在高维向量拟合的步骤中,因利用到了高维向量与高维空间中平面的投影关系,每次使用两个高维向量拟合一个向量,所以这个过程可视为待拟合向量在两个高维向量构成的平面上的分解拟合过程。

定义8 投影分解系数: 设 n 维空间上存在 $e = a' - a = j'b + k'c - a$,有

$$\begin{aligned} j, k &= \argmin_{j', k'} e_{(V)}. \\ \text{s.t. } \cos_V(e, b) &= 0, \cos_V(e, c) = 0. \end{aligned} \quad (7)$$

则称 j, k 为 a 在 b, c 构成超平面上的投影分解系数, a' 是 a 在 b, c 构成超平面上的投影。 e 表示误差。

求解投影分解系数的具体方法是: 首先求 b 施密特正交化于 c 的投影 b' ,然后用正交基 b' 和 c 来表示 a' 。那么可得

$$\mathbf{a}' = \frac{\langle \mathbf{a}', \mathbf{b}' \rangle_V}{\|\mathbf{b}'\|_V^2} \cdot \mathbf{b}' + \frac{\langle \mathbf{a}', \mathbf{c} \rangle_V}{\|\mathbf{c}\|_V^2} \cdot \mathbf{c}. \quad (8)$$

根据几何意义,可以求得

$$j = \frac{\langle \mathbf{a}', \mathbf{b}' \rangle_V}{\|\mathbf{b}'\|_V^2}, \quad k = \frac{\|\mathbf{a}' - j\mathbf{b}\|}{\|\mathbf{c}\|_V}. \quad (9)$$

综合式(8),(9)和施密特正交化过程,设函数为 $\mathbb{J}$, $\mathbb{K}$, j 和 k 的计算式为

$$j = \mathbb{J}(\mathbf{a}, \mathbf{b}, \mathbf{c}), \quad k = \mathbb{K}(\mathbf{a}, \mathbf{b}, \mathbf{c}). \quad (10)$$

利用三维向量说明上述向量拟合的图解形式,如图2所示.

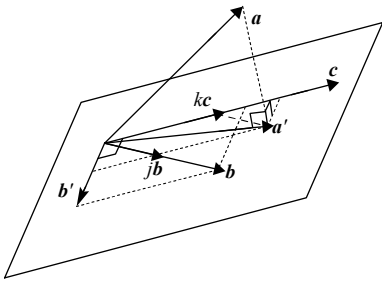


图 2 三维向量拟合
Fig.2 The example of vector fitting on 3D space

2.3.2 多向量拟合单向量

通过递归调用2.3.1中的拟合规则,本小节将描述使用多个基向量拟合一个向量的方法.

假设在 n 维空间中,已知给定基向量组 $\Omega = (\omega_1, \omega_2, \dots, \omega_s)$ 和向量 $\mathbf{a}$;拟合时所允许的最大误差阈值为 ε ,若存在一组系数

$$\mathbf{W} \in \mathbb{R}^{1 \times s}, \mathbf{W} = (\lambda_1, \lambda_2, \dots, \lambda_s)$$

使得

$$\mathbf{a} = \mathbf{W} \cdot \Omega^T + \mathbf{e}, \text{ s.t. } \|\mathbf{e}\|_V < \varepsilon, \quad (11)$$

则称向量 $\mathbf{a}$ 能由向量组 Ω 线性拟合.

2.3.3 多向量拟合的系数计算与拟合匹配

实际利用上述方法对高维向量的拟合匹配流程如图3所示.首先 $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_s)$ 表示使用到的基向量集合,与 $\mathbf{W}$ 组成参数对.图中 $\mathbf{m}_a$ 为待拟合向量; Γ 表示该过程最多能使用的基向量个数.

假设使用了 $s, s \in \{1: \Gamma\}$ 个基向量来拟合向量 $\mathbf{m}_a$,约定第 $i, i \in \{1: s\}$ 次加入的基向量的编号为 $\mathbf{g}_i$,其最终系数为 λ_i ,记当前轮次的系数为 k_i .则

$$g_i = \begin{cases} \operatorname{argmax} \cos_V(\mathbf{m}_a, \mathbf{d}_\tau), & i = 1; \\ \operatorname{argmax}_{\tau \notin \{g_j | j \in (1, i-1)\}} \cos_V(\varphi_{i-1}, \mathbf{d}_\tau), & i \in \{2: s\}. \end{cases} \quad (12)$$

式(12)中, $\omega_i = \mathbf{d}_{g_i}$,当前轮次下所有选中的基向量的临时合成向量为 φ_i ,

$$\varphi_i = \begin{cases} k_i \omega_i, & i = 1; \\ j_i \varphi_{i-1} + k_i \omega_i, & i \in \{2: s\}. \end{cases} \quad (13)$$

式(13)中,套用式(10), k_i 和 j_i 为

$$k_i = \begin{cases} \frac{\langle \mathbf{m}_a, \mathbf{d}_{g_1} \rangle_V}{\|\mathbf{d}_{g_1}\|_V}, & i = 1; \\ \mathbb{K}(\mathbf{m}_a, \varphi_{i-1}, \mathbf{d}_{g_i}), & i \in \{2: s\}. \end{cases} \quad (14)$$

$$j_i = \begin{cases} 1, & i = s + 1; \\ \mathbb{J}(\mathbf{m}_a, \varphi_{i-1}, \mathbf{d}_{g_i}), & i \in \{2: s\}. \end{cases} \quad (15)$$

那么,最终 $\mathbf{W}$ 中的各元素 λ_i 为

$$\lambda_i = k_i \cdot \prod_{u=i}^s j_{u+1}. \quad (16)$$

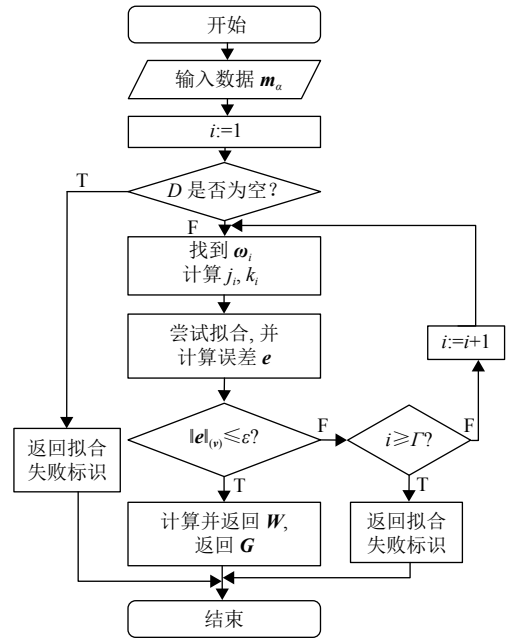


图 3 高维向量拟合匹配过程
Fig.3 The process of high dimension vector fitting

2.4 文件格式

本文提出的方法可用于语音频谱模值包络的压缩,将压缩后文件分为文件头和数据两部分.文件头记录压缩相关的参数大小,数据部分则分别记录字典和稀疏表示后的参数.字典标记后会接一个窗口长度的字典数据;参数标记后接一个参数块个数,表示包含多少块单块信息,一个参数块最多能存255个单块,因此参数块数使用1字节表示,若在实际使用中需要超过255个连续单块,则拆分为多个参数块;而每块单块信息包含了用于拟合一个向量的编号 $\mathbf{G}$ 以及系数 $\mathbf{W}$.更具体的文件格式如图4所示.



图4 文件格式
Fig.4 File format

2.5 字典构建与压缩过程

按 α 增序, 逐个处理待拟合向量 m_α , 压缩流程图如图5所示. 图5中的拟合匹配使用图3的拟合方法. 若图3返回拟合失败, 则需要更新字典. 更新时, 当字典未达到容量上限 z 时直接加入向量 m_α , 当达到容量时当前采用的是最近最少使用(Least Recently Used, LRU)策略替换旧的基向量. 在写文件过程中, 写入文件格式参照图4. 记标记位变量为 Δ , 如果 $\Delta=0$, 则表示记录字典到文件, 那么将写入字典标记和字典数据; 如果 $\Delta=1$, 则表示写参数到文件. 那么先写参数块个数为0并记下参数位置, 之后再写入单块信息, 直到再次遇到 $\Delta=0$, 则将统计好的单块信息总数回填至参数位置; 若单块信息的实际统计个数超过1字节(255)存储容量, 则将255回填至参数位置并重新开一个参数块记录余下的单块信息.

2.6 还原过程

本文提出的语音字典在压缩过程中被碎片式地记录于压缩文件中, 压缩过程中标志位变量记为 Δ .

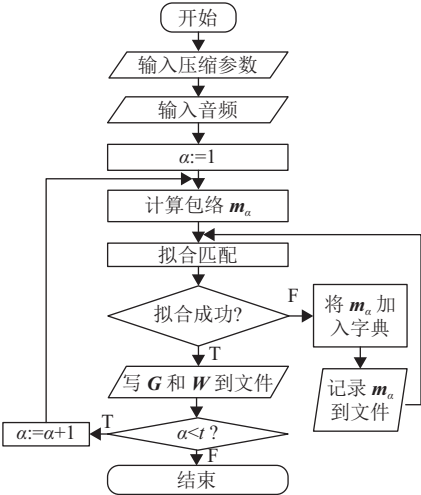


图5 压缩过程
Fig.5 The process of compressing

整个解压还原的流程图如图6所示. 首先获取文件头, 根据文件头的参数对解码器做好初始化设置. 然后按顺序处理数据部分的每个块, 读取 Δ . 若 $\Delta=0$, 则读取基向量, 并按照LRU的规则加入字典; 若 $\Delta=1$, 则读取参数块, 通过获取一个单块信息中的 G 和 W 并使用

$$\Omega = (d_i | i \in \{G\}). \tag{17}$$

$$m_\alpha' = W \cdot \Omega^T. \tag{18}$$

得到对应的近似还原包络.

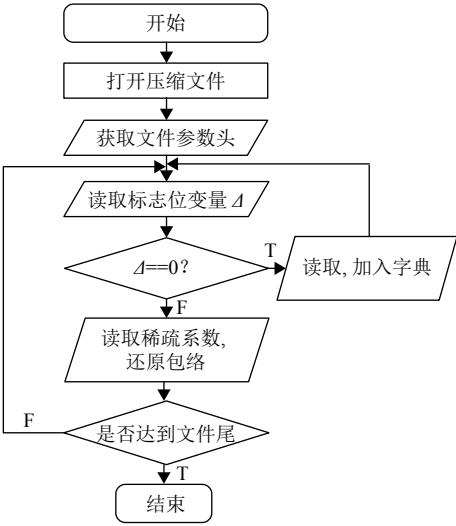


图6 还原过程流程图
Fig.6 The process of decompressing

3 实验结果及分析

实验在 win10 系统下进行, 利用 Visual Studio2015 编码. 开源语音数据集 PTDB-TUG 包含有 10 位女生 10 位男生的声音片段集. 从中随机选取 5 名

女生和5名男生的录音集,每个人的录音大约有220个片段,采样率是44.1 kHz,按照窗长度为4 096点进行STFT实验获取频谱包络,每个窗口长度为512,表示频率范围是0~6 kHz. 最终每个人大约有70 000~90 000个窗口包络,每个人的语音量可以模拟数据不断增长的过程.

目前流行的字典学习方法是训练一个过完备字典. 过完备字典是指字典中条目的数量大于每个条目的维度^[15],过完备字典一般从样本数据随机抽取 K 条,利用KSVD或在线字典学习(Online Dictionary learning, ODL)等字典学习算法进行训练,然后利用系数分解算法求解出样本数据的系数.

相比较而言ODL算法比KSVD训练字典更迅速,本次实验选择与ODL算法进行对比,ODL算法源码在文献[16]的作者主页下载,并在Matlab环境下进行运算使用. 通过调整参数使得两种算法在数据的平均相似度上尽可能一致,然后比较两种算法的压缩比和具体每一个窗的还原质量.

3.1 压缩比对比

定义9 压缩比计算公式:假设 A 是原始数据大小; A' 是压缩后数据的大小,大小单位为bit. 那么压缩比为 $A'A^{-1}$. 压缩比越小表明压缩效果越好.

在本文实验中,原始频谱包络 $Y \in R^{n \times l}$ 以浮点(float)的形式存储,一列为一个窗口的频谱包络. 压缩后的数据中,字典中所有系数以及基向量以浮点(float,32bit)形式存储,记字典大小为 z ,每个基的长度

为 n ;计本方法生成的文件中,共有 ς_1 个参数块, ς_2 个单块信息, ς_3 对参数对. 最终字典写入的基向量个数为 ζ , G 中每一个元素 g_i 所占位数均为 $\lceil \log_2 z \rceil$ bit,那么本方法中的压缩数据大小 A' 为

$$A' = (32 + 1)\zeta n + (8 + 1)\varsigma_1 + \varsigma_2 \lceil \log_2 \Gamma \rceil + \varsigma_3 (\lceil \log_2 z \rceil + 32).$$

(19)

对于ODL,为公平对比,在每次测试中,字典容量均与本算法的 ζ 相同,而且从直接存储 X ,或者 G 和 W 表示 X 的这两种方法中,选取最小者作为压缩数据大小,记为 B' . 记 X 中非零元素的个数为 ς_4 ,则

$$B' = \min \left\{ \begin{array}{l} 32\zeta n + \varsigma_4 (\lceil \log_2 z \rceil + 32) \\ 32(\zeta n + zl) \end{array} \right\}.$$

(20)

在本实验中,通过调整ODL中的lambda参数使得两种方法得到的平均相似度近似.

实验结果分成男女两组,分别记录两种方法对同一数据处理后的相似度和压缩比两个结果,详情见表1和表2. 但是在具体实验中没有针对男女的区别进行特殊处理. 由表中可得,在相同的平均近似度结果下,本文方法的压缩比均大于ODL算法的压缩比. 另外,实际应用中ODL算法一般需要预先定义字典大小,如果字典大小定义不合适,则会影响压缩效率,并且ODL算法的字典需要在已知部分数据下进行训练得到. 而本文中提出的方法能够根据流式数据进行增量式字典学习,既保证较高的压缩效率,又有着在字典训练过程不必预先获取部分数据进行训练的优势.

表1 女声实验结果对比
Tab.1 Experiment of female speech

	女1		女2		女3		女4		女5	
	相似度	压缩比	相似度	压缩比	相似度	压缩比	相似度	压缩比	相似度	压缩比
本方法	0.986 73	0.008 039	0.987 21	0.008 465	0.986 00	0.010 647	0.986 95	0.008 739	0.986 32	0.008 713
ODL	0.982 29	0.010 855	0.987 99	0.011 786	0.984 54	0.014 018	0.987 59	0.012 411	0.986 30	0.011 622

表2 男声实验结果对比
Tab.2 Experiment of male speech

	男1		男2		男3		男4		男5	
	相似度	压缩比	相似度	压缩比	相似度	压缩比	相似度	压缩比	相似度	压缩比
本方法	0.985 85	0.016 553	0.98 463	0.014 033	0.985 09	0.014 179	0.984 52	0.017 093	0.985 21	0.017 643
ODL	0.983 76	0.017 211	0.98 790	0.019 231	0.987 60	0.019 037	0.983 14	0.018 692	0.985 08	0.021 425

3.2 还原质量对比

定义10 平均相似度计算公式:假设 Y 是原始频谱包络; Y_i 是还原后频谱包络. 那么平均相似度 C 为

$$C = t^{-1} \sum_{i=1}^t \cos v(m_i, m'_i).$$

(21)

从PTDB-TUG数据集中随机选择一个人的全部录音,将两种方法的各个窗口具体还原质量绘制出相似度散点图. 对比图7和图8可看出,本文方法各窗相似度均在所设阈值以上,整体还原质量较高;而ODL算法的各窗相似度分布较为分散,这表明部分

窗口包络存在较大的缺失,这会对听觉质量产生较大影响.

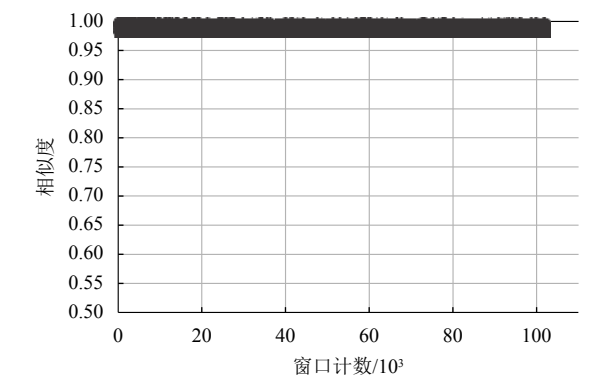


图7 本方法还原质量
Fig.7 The performance of our method

此外通过频谱包络图可以观察还原质量,将原包络以及两种方法还原得到的包络转换成频谱包络图(见图9),颜色越深表示数值越大.

在图9中,(a)是原始频谱包络图,(b)是本方法还原后频谱包络图,(c)是ODL还原后的频谱包络图.从

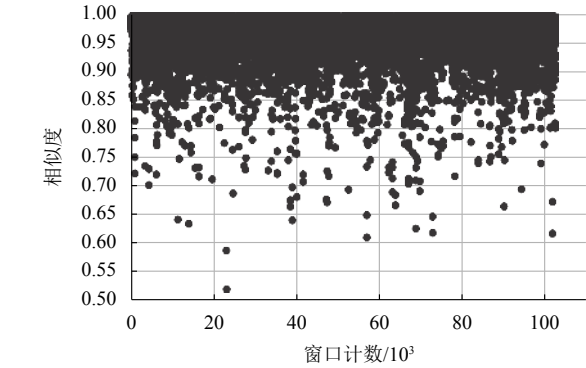


图8 ODL还原质量
Fig.8 The performance of ODL

图中黑色方框可以看出,图9(c)中ODL方法对语音模值的高频部分包络过于平滑,没有保留原始包络的细节部分,例如图9(c)方框中每个窗口内数值很接近,而实际原始包络(a)方框中每个窗口内数值存在变化.相反,本文提出的方法则能够较好地保留一些细节部分,且与图9(a)中的纹样相似.这些细节对人耳听觉会产生一定的影响,因此如果能够正确还原细节将对语音质量有所保证.



(a) 原始频率包络图 (b) 本方法还原后频率包络图 (b) ODL 还原后频率包络图

图9 还原质量对比
Fig.9 Reconstruction quality

4 结束语

针对大量产生的稀疏性模值信号包络,本文提出一种较低压缩比,易于重构还原以及还原度较高的稀疏数据压缩方法.该方法利用稀疏字典学习的方法迭代学习提取稀疏数据的重要特征,减少存储数据中的冗余信息,同时结合增量学习策略来解决不断产生的数据流带来的特征学习困难的问题.实验结果表明该方法在语音数据流的压缩与还原方面能取得较好的效果.

在下一步工作中,将继续研究如何降低压缩比同时保持较高的还原率以及如何提高计算效率,尤其是编码时的运算效率的问题.

参考文献:

[1] GIBSON J. Speech compression [J]. Information, 2016,

7(2): 32-54.
[2] DIETZ M, MULTRUS M, EKSLER V, *et al.* Overview of the EVS codec architecture[C]//Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. Brisbane: IEEE, 2015: 5698-5702.
[3] AURISTIN F N, MALI S. New Ieee standard for advanced audio coding in lossless audio compression: a literature review [J]. International Journal of Engineering and Computer Science, 2016, 5(4): 16167-16171.
[4] GEORGE M L, TOMAR G S, SHARMA T, *et al.* Hardware implementation of adaptive-differential pulse code modulation speech compression algorithm [J]. International Journal of Smart Device and Appliance, 2014, 2(2): 1-10.
[5] DUTA C L, GHEORGHE L, TAPUS N. Real time implementation of MELP speech compression algorithm using Blackfin processors[C]//Image and Signal Processing and Analysis (ISPA), 2015 9th International Symposium on. Zagreb: IEEE, 2015: 250-255.

- [6] YANG L, GUO Z Y, YONG S S, *et al.* A hardware implementation of real time lossless data compression and decompression circuits[J]. *Applied Mechanics and Materials*, 2015, 719: 554-560.
- [7] Wikipedia Encyclopedia. Sparse dictionary learning[EB/OL]. (2017-12-19)[2018-01-29]. https://en.wiki-pedia.org/wiki/Sparse_dictionary_learning
- [8] YANG M, DAI D, SHEN L, *et al.* Latent dictionary learning for sparse representation based classification[C]//Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 4138-4145.
- [9] LI F, ZHANG X, ZHANG H, *et al.* An AFK-SVD sparse representation approach for speech signal processing [C]//PAN J S, TSAI P W, HUANG H C. *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Kaohsiung, Taiwan: [s. n.], 2018: 177-184.
- [10] 叶向荣, 刘怡俊, 陈云华, 等. 基于 $L_{(1/2)}$ 自适应稀疏正则化的图像重建算法[J]. *广东工业大学学报*, 2017, 34(6): 43-48.
YE X R, LIU Y J, CHEN Y H, *et al.* A super-resolution image reconstruction algorithm with adoptive $L_{(1/2)}$ sparse regularization [J]. *Journal of Guangdong University of Technology*, 2017, 34(6): 43-48.
- [11] YAN Y, YANG Y, MENG D, *et al.* Event oriented dictionary learning for complex event detection [J]. *IEEE Transactions on Image Processing*, 2015, 24(6): 1867-1878.
- [12] 杨婷, 滕少华. 改进的贝叶斯分类方法在电信客户流失中的研究与应用[J]. *广东工业大学学报*, 2015, 32(3): 67-72.
YANG T, TENG S H. Research and application of improved Bayes algorithm for the telecommunication customer churn [J]. *Journal of Guangdong University of Technology*, 2015, 32(3): 67-72.
- [13] NADERAHMADIAN Y, BEHESHTI S, TINATI M A. Correlation based online dictionary learning algorithm [J]. *IEEE Transactions on Signal Processing*, 2015, 64(3): 592-602.
- [14] AHARON M, ELAD M, BRUCKSTEIN A. SVD: an algorithm for designing overcomplete dictionaries for sparse representation [J]. *IEEE Transactions on Signal Processing*, 2006, 54(11): 4311-4322.
- [15] MAIRAL J, BACH F, PONCE J, *et al.* Online learning for matrix factorization and sparse coding [J]. *Journal of Machine Learning Research*, 2010, 11(1): 19-60.
- [16] ZUBAIR S, YAN F, WANG W. Dictionary learning based sparse coefficients for audio classification with max and average pooling [J]. *Digital Signal Processing*, 2013, 23(3): 960-970.
- [17] BAO C, JI H, QUAN Y, *et al.* Dictionary learning for sparse coding: algorithms and convergence analysis [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2016, 38(7): 1356-1369.
- [18] SHEN L, WANG S, SUN G, *et al.* Multi-level discriminative dictionary learning towards hierarchical visual categorization[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. Portland: IEEE, 2013: 383-390.
- [19] GUHA T, WARD R K. Image similarity using sparse representation and compression distance [J]. *IEEE Transactions on Multimedia*, 2014, 16(4): 980-987.
- [20] 李轶南, 张雄伟, 曾理, 等. 改进的稀疏字典学习单通道语音增强算法[J]. *信号处理*, 2014, 30(1): 44-55.
LI Y N, ZHANG X W, ZENG L, *et al.* An improved Monaural speech enhancement algorithm based on sparse dictionary learning [J]. *Journal of Signal Processing*, 2014, 30(1): 44-55.
- [21] SUN L H. Speech enhancement based on data-driven dictionary and sparse representation [J]. *Signal Processing*, 2011, 27(12): 1793-1800.
- [22] CHEN X, DU Z, LI J, *et al.* Compressed sensing based on dictionary learning for extracting impulse components [J]. *Signal Processing*, 2014, 96(5): 94-109.
- [23] SRINIVAS M, ROY D, MOHAN C K. Learning sparse dictionaries for music and speech classification[C]//Digital Signal Processing (DSP), 2014 19th International Conference on. Hong Kong: IEEE, 2014: 673-675.
- [24] HSIEH C T, HUANG P Y, CHEN T W, *et al.* Speech enhancement based on sparse representation under color noisy environment[C]//Intelligent Signal Processing and Communication Systems (ISPACS), 2015 International Symposium on. Nusa Dua: IEEE, 2015: 134-138.
- [25] ALLEN J. Short term spectral analysis, synthesis, and modification by discrete Fourier transform [J]. *IEEE Transactions on Acoustics Speech & Signal Processing*, 2003, 25(3): 235-238.