

文章编号: 1000-5862(2019)04-0394-08

# 语音共振峰包络的增量频移字典学习方法

霍颖翔, 滕少华\*

(广东工业大学计算机学院, 广东 广州 510006)

**摘要:** 数字语音在当今应用非常广泛,大量的语音流产生了巨大的网络带宽和服务器存储空间的消耗。因此,在保持听觉效果基本不受影响的前提下,对语音进行有损压缩,降低其比特率是非常重要的。针对压缩语音的共振峰包络提出了一种新颖的在线字典学习方法。不同于一般的线性方法,该方法通过对字典中的原子进行频移,使其能更好地进行共振峰拟合。通过使用希尔伯特变换,能快速并精确地确定最优频移量。实验结果表明,在还原近似度下限为 99.5% 的前提下,经过该方法压缩后,比特数比原包络平均减少了 99%。因此,该方法能适用于对传输带宽或存储空间有严格要求的场合,同时保证解压后的语音听觉比较自然。

**关键词:** 语音流; 有损压缩; 在线字典学习

**中图分类号:** TP 301.6    **文献标志码:** A    **DOI:** 10.16357/j.cnki.issn1000-5862.2019.04.11

## 0 引言

语音是重要的日常交流方式,数字语音为人类提供了以更低的价格传输、存储和回放高质量语音的可能。大量的实际需求推动着语音压缩方法的发展。首先,对一般网络应用和移动语音通话而言,为了更充分地利用网络带宽资源,并带来更好的用户体验,需要对语音进行压缩。而一些(如卫星电话)带宽和存储空间极其昂贵应用,更是要求极低码率的语音流压缩。除此之外,对于一些移动设备而言,计算量与电池续航能力密切相关,因此语音压缩的计算复杂度也是一个需要考虑的因素。

常用的有损语音压缩方法一般通过移除语音中的冗余信息、语义不相关信息,以及人类听觉不灵敏的信息,或降低这些信息的数值编码精度以获得更低的码率。本文着重研究的语音幅值包络谱的压缩是语音压缩的重要一环。通过字典学习,并每次选用字典中的少量原子进行包络谱拟合,能利用少量参数表示完整包络,从而达到压缩的目的。大量文献给出了字典学习的广泛应用,如语音增强<sup>[1]</sup>、音频压

缩<sup>[2-3]</sup>、语音分类<sup>[4]</sup>、音乐分析<sup>[5]</sup>、盲源分离<sup>[6]</sup>、图像降噪<sup>[7-8]</sup>、图像分类<sup>[9-11]</sup>、图像升采样<sup>[12]</sup>和纹理合成<sup>[13]</sup>等。在这些应用中稀疏学习模型能够较好地适用于自然信号。

在实际应用中,由于实时产生的信号和已记录但规模庞大的信号无法在压缩时观察到其全貌,因此需要一种动态的压缩方法,通过对信号进行增量式的学习而进行压缩,而在线字典学习(ODL)方法则做到了这一点。ODL 算法<sup>[14-15]</sup>能根据输入信号动态调整字典中的基,而另一些算法<sup>[16-17]</sup>则着重于在计算效率和内存要求上的优化,它们为处理大数据提供了可能。

字典学习通常分为线性和非线性 2 种。主成分分析(PCA)<sup>[18]</sup>是一种线性方法,PCA 和它的多种变体通过使用正交基快速而唯一地分解信号,使信号能通过少数原子的线性混合进行表示与重构。然而,在处理现实中的非线性信号时,这些算法可能并不一定能完全适应,因此催生了一些非线性方法,例如基于神经网络<sup>[19]</sup>、核方法<sup>[20]</sup>和主测地线分析(principal geodesic analysis,PGA)<sup>[21]</sup>的方法。PCA 和 K-SVD<sup>[22]</sup>方法通过统计学的方式生成字典,它们能

收稿日期: 2019-03-10

基金项目: 国家自然科学基金(61772141, 61402118, 61673123, 61603100, 61702110), 广东省科技计划(2016B010108007), 广东省教育厅(粤教高函[2018]179 号, 粤教高函[2018]1 号, 粤教高函[2015]113 号, 粤教高函[2014]97 号)和广州市科技计划(201802030011, 201802010026, 201802010042, 201604020145, 201604046017)资助项目。

通信作者: 滕少华(1962-), 男, 江西南昌人, 教授, 博士, 主要从事大数据、数据挖掘、数字音频分析与处理、网络安全方面的研究。E-mail: shteng@gdut.edu.cn

够较好地表示多数的模式,但可能忽略了不常见的模式以致于对这种模式的数据不能较好地拟合.另外,通过使用最小平方误差(MSE)或与其等效的方法,一些算法能获得极高的平均拟合度.然而,当应用于语音共振峰压缩时,这些算法并不能保证共振峰结果自然并符合人类声学特性,从而可能导致较差的感知质量.若将同一人在不同时刻对同一字的发音的幅值谱包络交换,即使这些包络有一定差别,也不会对其听觉质量产生较大的影响.但若将这2个包络进行线性叠加,听觉质量却会明显下降.主要的原因是,单个包络的峰和谷锐利且明确,而线性叠加混合后使得包络变得模糊而影响听觉.因此,在字典拟合过程中,使用一个或少数的原子进行拟合比使用多个原子互相修正有更好的听觉效果.多数字典学习算法追求最高平均拟合相似度,这使它们不重视这些短时冲击,不着重避免因拟合误差造成的短时冲击.但人类听觉系统对瞬时冲击比较敏感,这对感知质量影响较大.

针对已有方法的不足,本文提出一种新颖的方法.通过使用实际的共振峰包络作为字典原子,同时使用尽量少的原子进行拟合使得还原的共振峰包络更为锐利和清晰.本文首先介绍了原子的选择和原子权值的基本计算方法;然后通过频率方向上移动原子,使其能更好地拟合目标包络,通过使用希尔伯特变换,该算法能高效地计算出原子在频率方向上的最佳移动量,从而保证了信号分解算法的效率;最后还探讨了单趟与多趟的信号扫描方法,在要求低编码延迟的情况下可使用前者,而使用后者则能以可接受的编码延迟作为代价,进一步降低码率.通过同时支持单趟与多趟方案,本文的算法能更灵活地应对不同应用场合.

## 1 增量频移字典学习方法

### 1.1 线性原子匹配与字典学习方法

假设目前要拟合的目标是一些 $n$ 维的向量,记需要构造的字典和它的原子为 $D = [d_0, d_1, \dots, d_{r-1}] \in \mathbf{R}^{n \times r}$ ,需被拟合的一个目标向量记为 $\lambda \in \mathbf{R}^{n \times 1}$ ,有 $\xi$ 个原子从 $D$ 中被选用于拟合 $\lambda$ ,它们在字典中的编号记为 $\Omega = [\omega_0, \omega_1, \dots, \omega_{\xi-1}] \in \mathbf{R}^{n \times \xi}$ , $d_{\omega_i}$ 所对应的权值为 $\theta_i$ , $\Theta = [\theta_0, \theta_1, \dots, \theta_{\xi-1}] \in \mathbf{R}^{1 \times \xi}$ ,记 $\lambda'$ 和 $\theta'$ 、 $\theta''$ 分别为临时向量和2个临时权值, $i$ 、 $j$ 、 $k$ 为计数器.将 $\lambda$ 分解为字典中原子的过程如表1

所示.

表1 目标函数 $\lambda$ 的分解过程

输入: 目标函数 $\lambda(x)$ 的离散采样 $\lambda$ ,字典 $D$	
输出: 选用原子编号 $\Omega$ 及其权值 $\Theta$ ,更新后的字典 $D$	
1	$\omega_0 \leftarrow \operatorname{argmax}_i \  \operatorname{proj}(\lambda, d_i) \ $
2	$\theta_0 \leftarrow \  \operatorname{proj}(\lambda, d_{\omega_0}) \  / \  d_{\omega_0} \ ^2$
3	$\lambda' \leftarrow \lambda - \theta_0 d_{\omega_0}$
4	for $j \leftarrow 1$ to $r$ do
5	$\omega_j \leftarrow \operatorname{argmax}_i \  \operatorname{proj}(\lambda', \{d_i\}) \ $ s. t. $\omega_i \notin \{\omega_k   k \in [0, j]\}$
6	$\{\theta', \theta_j\} \leftarrow \operatorname{argmin}_{\{\theta', \theta_j\}} \  \theta' \lambda' + \theta_j d_{\omega_j} - \lambda \ ^2$
7	$\lambda' \leftarrow \operatorname{proj}(\lambda, \{\lambda', d_{\omega_j}\})$
8	$\forall_{k \in [0, j]} \theta_k \leftarrow \xi \theta_k$
9	if $\operatorname{sim}(\lambda', \lambda) \geq \text{threshold}$ then exit for
10	end for
11	输出 $\forall_{k \in [0, j]} \{\theta_k, \omega_k\}$

在表1中 $\operatorname{proj}(a, b)$ 表示向量 $a$ 在向量 $b$ 上的投影,另外,其同名扩展 $\operatorname{proj}(a, \{b_1, b_2\})$ 表示向量 $a$ 在向量 $b_1$ 和 $b_2$ 所构成的(超)平面上的投影.需要注意的是 $\theta'$ 、 $\theta''$ 、 $\lambda'$ 和 $\Theta$ 在循环中会被更新和覆盖.表1的第6行中 $\operatorname{argmin}$ 可通过几何关系快速求得而不需要复杂的寻优过程,因为字典中的原子不一定互相正交,一些原子之间甚至有一定的相似度,所以在拟合过程中,即使有一些原子与目标函数相似也不一定会使用.而第9行中的 $\operatorname{sim}()$ 函数代表拟合度,其计算方式为

$$\operatorname{sim}(\lambda', \lambda) = \frac{\lambda' \cdot \lambda}{\|\lambda'\| \|\lambda\|}, \quad (1)$$

其中运算符“ $\cdot$ ”代表向量的内积.事实上,稀疏字典学习的目标是在每次拟合时只选取少量的原子进行拟合.原子的选取方式可能是多解的,而表1中的算法以比较低的计算量为代价给出了较为合理的解.

表1中的算法并不能保证其拟合结果误差一定在允许范围内.若尝试拟合后发现误差较大,应将当前目标函数作为新的原子加入字典,从而保证其一定能成功拟合.如图1所示,这是构造字典的唯一机会,并会同时存储/传输该原子.为了让第5行中的 $\operatorname{argmax}()$ 函数效率更高,需通过限制字典容量来减少其穷举次数.在字典中的原子数量超过上限时,本文算法使用了最小最近使用(least recently used, LRU)方法对字典进行换入换出.此外,第5行的 $\operatorname{argmax}()$ 过程还使用了并行方式计算,以增加计算效率.

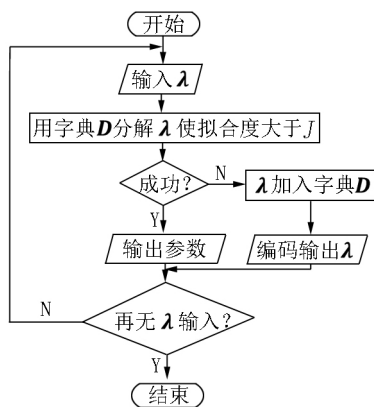


图1 字典构造方法

图2和图3展示了编码过程与字典训练程度之间的关系。如图2所示,由于在初始时字典内容较少,表现能力较弱,容易发生拟合度 $J$ 达不到最低阈值要求的情况。因此,会较频繁发生向字典添加新原子的情况。每当添加完1个原子,拟合度就能回到100%。在一小段时间后,因为内容变化,又有新的内容使当前字典无法较好地表示,因此拟合度再度下降,最后又触发添加原子操作,如此反复。如图3所示,经过一段时间的编码后,字典得到了训练,含有比较完整的原子以应对多数情况,此时拟合度保持较好,原子更新也变得不再频繁,达到了较好的压缩效果。

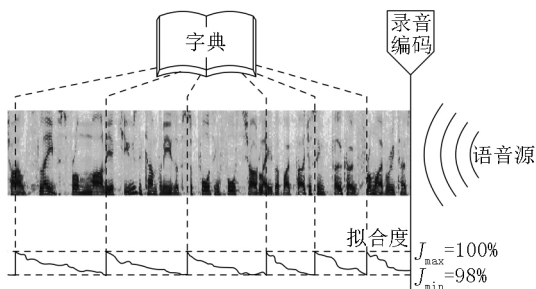


图2 初始的编码示例

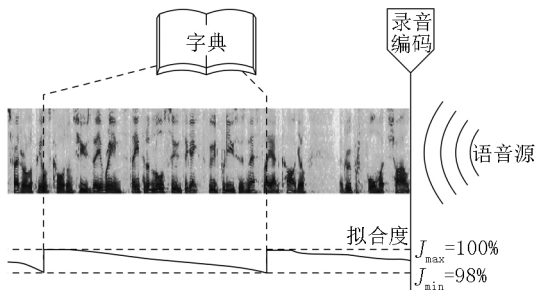


图3 在字典训练稳定后的编码示例

## 1.2 针对共振峰包络的在线字典学习方法

1.2.1 共振峰包络特点 图4展示了一组共振峰包络随时间变化的例子。由图4可知,在大多数情况下共振峰不会突变,上一时刻的共振峰形状,与紧邻的下一时刻非常相似。因此,可采用时间差分记录的

方式对该特性进行更优化的编码。此外,与图4中的(b)段为例,相邻的2个窗口可以近似通过在频率方向上的平移来互相表示,可通过改进字典学习方法,尝试将原子进行平移或近似平移后对目标函数进行拟合。

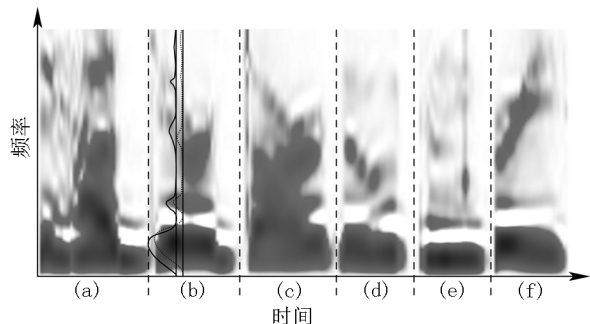


图4 共振峰包络例子

1.2.2 希尔伯特变换 希尔伯特变换是一种特殊的线性操作。希尔伯特变换将原函数的傅里叶变换频谱中的各个频率分量的初相旋转 $90^\circ$ 。记实数值输入信号为 $\lambda(t)$ ,希尔伯特变换的实数值输出信号为 $H(\lambda)(t)$ ,则希尔伯特变换可以表示为函数 $\lambda(t)$ 和函数 $\pi^{-1}t^{-1}$ 的卷积,即

$$H(\lambda)(t) = \pi^{-1} \int_{-\infty}^{\infty} \lambda(\tau) / (t - \tau) d\tau, \quad (2)$$

由于函数 $\pi^{-1}t^{-1}$ 在 $t \rightarrow 0$ 处有一个无穷间断点,导致上述积分不收敛。因此,积分结果仅取柯西主值。在应用于离散信号的实现上,则是通过直接忽略当累加时 $t - \tau = 0$ 的项来实现。

1.2.3 近似平移模型 尝试以如下的形式使用原函数 $\lambda(t)$ 及其希尔伯特变换的线性混合对函数 $\hat{\lambda}(t)$ 进行函数拟合:

$$\hat{\lambda}(t) \approx \lambda(t) \theta \cos w + H(\lambda)(t) \theta \sin w. \quad (3)$$

图5给出了一个简单的例子,该图将 $\theta$ 设为1,并通过如下轻微平移避免各曲线重叠

$$w + \lambda(t) \cos w + H(\lambda)(t) \sin w. \quad (4)$$

图6以 $\pi/10$ 为步进间隔,显示了当 $w \in [-\pi, \pi]$ 时(4)式呈现的曲线。其中,靠近下端的粗线刚好为原函数 $\lambda(t)$ ,而位置靠上的粗线为其希尔伯特变换后再上移 $w$ 的结果 $H(\lambda)(t) + w$ 。从图6可以清晰看到,随着 $w$ 的增长,(4)式的几个较显著的峰值和谷值均有左移的趋势。因此,可以通过调整(3)式中的 $w$ 达到近似将函数进行水平方向平移的目的。

图5是使用调整 $w$ 进行近似平移拟合的效果。目标函数为原函数轻微左移(子图(a))/右移(子图(b))40个采样点后的结果(右/左侧补零)。从图5可见,拟合结果的各极大极小值点的 $x$ 轴位置与目

标函数非常接近,但函数有轻微走样使得其  $y$  轴方向的值产生了一定偏差. 因为它们的大小关系基本

能够保持,所以若用此方式拟合共振峰,听觉效果将基本不受影响.

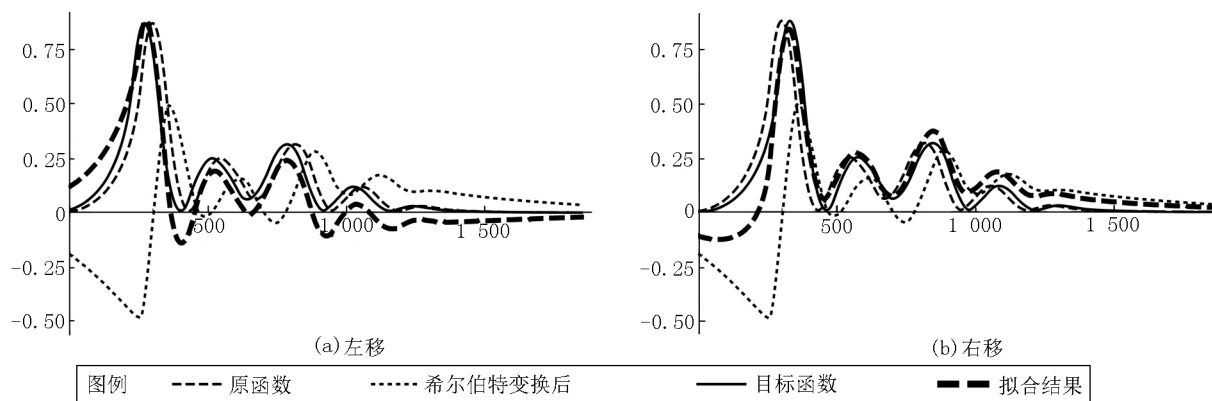


图5 近似平移拟合效果示意

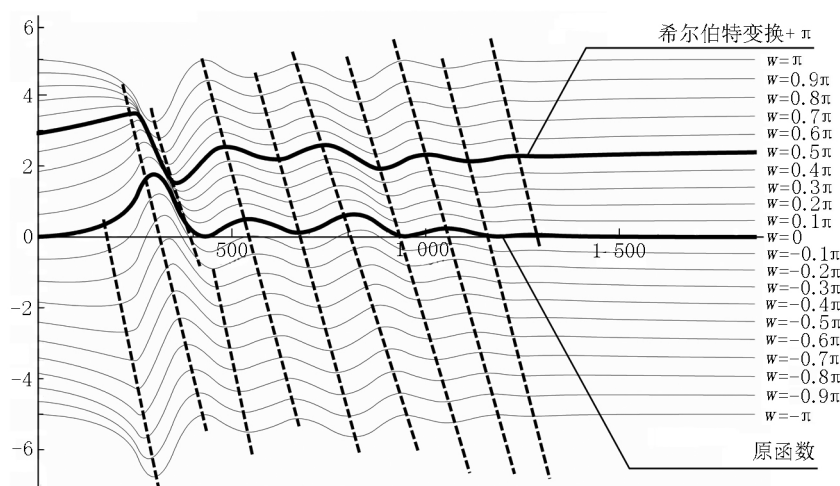


图6 希尔伯特变换例子

由于函数与它的希尔伯特变换结果为正交关系,若要用函数  $\lambda(t)$  和  $H(\lambda)(t)$  通过(3)式进行线性混合以拟合  $\hat{\lambda}(t)$ ,  $w$  可通过以下方式直接确定:

$$w(\lambda(t), \hat{\lambda}(t)) = \arg\left(\int_{-\infty}^{\infty} \lambda(t) \hat{\lambda}(t) dt + i \int_{-\infty}^{\infty} H(\lambda)(t) \hat{\lambda}(t) dt\right), \quad (5)$$

在实际 PCM 信号的应用中, (5) 式变为累加, 即

$$w(\lambda(t), \hat{\lambda}(t)) = \arg\left(\sum_t \lambda(t) \hat{\lambda}(t) + i \left(\sum_t H(\lambda)(t) \hat{\lambda}(t)\right)\right). \quad (6)$$

1.2.4 严格平移模型 为完全避免近似平移模型中函数走样的问题,提出通过自变量方向的平移进行拟合的方法. 利用近似平移特性,该方法能快速地确认最适合的平移量. 使用图6中的原函数  $\lambda(t)$  作为例子,用(6)式计算使用  $\lambda(t)$  和  $H(\lambda)(t)$  拟合  $\hat{\lambda}(t) = \lambda(t + \Delta t)$  的最佳  $w$ ,将得到如图7所示的结果.

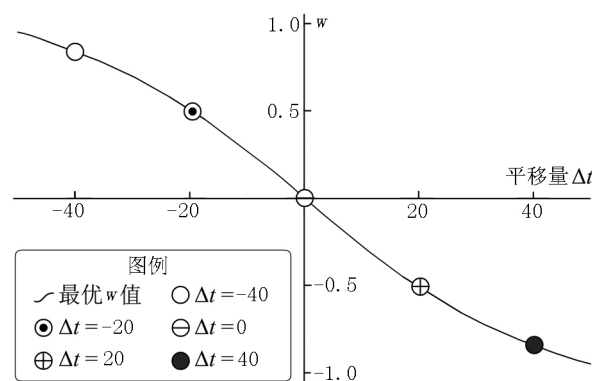


图7 平移量  $\Delta t$  与  $w$  值之间的关系例子

由图7可见,在  $\Delta t$  接近0时,  $\Delta t$  和  $w$  之间呈负相关关系,特别地,当  $\Delta t = 0$  时,  $w = 0$ . 而且该函数在  $\Delta t \in [-40, 40]$  的区间内,斜率变化不大. 实验表明,若  $\lambda$  为共振峰包络,则上述性质基本不变,而且斜率在一个较窄的范围内变动. 因此,该斜率可取得一个经验平均值  $Q$ ,使得

$$w(\lambda(t), \lambda(t + \Delta t)) \approx Q \Delta t. \quad (7)$$

因为共振峰包络有当  $t$  接近上下限时幅值接近0的特性,所以  $w(\lambda(t), \lambda(t + \Delta t))$  与  $w(\lambda(t - \Delta t), \lambda(t))$

$\lambda(t)$  结果相似. 在假设  $\lambda(t)$  能通过  $t$  方向的平移拟合  $\hat{\lambda}(t)$  的前提下, 可以使用确定的斜率  $Q$  值和简化的牛顿迭代法求解  $\Delta t$ :

$$\hat{\Delta t} \xleftarrow{\text{迭代更新}} \hat{\Delta t} - Qw(\lambda(t - \hat{\Delta t}) - \hat{\lambda}(t)). \quad (8)$$

1.2.5 平移模型与字典学习的结合 若将 1.2.1 与 1.2.4 结合, 则可以通过对若干原子进行平移并线性叠加来拟合新的函数. 记函数

$$\text{shift}(\lambda(x), w) = \begin{cases} \lambda(x+w) & x+w \text{ 在 } \lambda \text{ 定义域内} \\ 0 & \text{其他} \end{cases} \quad (9)$$

表 1 的过程可被扩展为表 2 中所示的形式. 可以看出, 扩展后增加了参数  $w_j$ , 但实际上由于这会导致更早地在第 9 行处退出循环, 减少了拟合所需的原子数, 并有效减少拟合失败的可能, 减少了原子的传输, 因此这反而会有利于降低压缩后的比特率.

表 2 目标函数  $\lambda$  的平移分解过程

输入:	目标函数 $\lambda(x)$ 的离散采样 $\lambda$ , 字典 $D$
输出:	选用原子编号 $\Omega$ 及其权值 $\Theta$ , 各原子平移量 $w_j$ , 更新后的字典 $D$
1	$(\omega_0, w_0) \leftarrow \arg\max_{i, w} \  \text{proj}(\lambda, \text{shift}(d_i, w)) \ $
2	$\theta_0 \leftarrow \  \text{proj}(\lambda, d_{\omega_0}) \  / \  d_{\omega_0} \ $
3	$\lambda' \leftarrow \theta_0 d_{\omega_0}$
4	for $j \leftarrow 1$ to $r$ do
5	$\{\omega_j, w_j\} \leftarrow \arg\max_{i, w} \  \text{proj}(\lambda, \{\lambda' \text{shift}(d_i, w)\}) \ $ s. t. $\omega_i \notin \{\omega_k   k \in [0, j]\}$
6	$\{\theta', \theta_j\} \leftarrow \arg\min_{\theta', \theta_j} \  \theta'^{\lambda'} + \theta_j' \text{shift}(d_{\omega_j}, w_j) - \lambda \ $
7	$\lambda' \leftarrow \text{proj}(\lambda, \{\lambda', \lambda' d_{\omega_j}\})$
8	$\forall_{k \in [0, j]} \theta_k \leftarrow \xi \theta_k$
9	if $\min(\lambda', \lambda) \geq \text{threshold}$ then exit for
10	end for
11	输出 $\forall_{k \in [0, j]} \{\theta_k, w_j\}$

### 1.3 多趟扫描法

上文提到的字典算法实际上是单趟扫描算法. 该算法在拟合失败时直接将当前数据加入字典, 优点是直接而且编码延时较小, 但缺点是, 这样增加的原子更像是在临时补救, 使得字典中的原子选择不够优化. 这将导致较高的比特率, 并且容易导致语音质量不稳定. 例如, 若使用当前字典直接拟合一个输入数据, 可能需要使用 10 个原子进行混合. 但是如果使用 100 ms 后的字典, 可能只需要 1 个原子就能拟合当前数据. 而且后者拟合度甚至比前者好. 因此, 通过提前构造字典可使拟合效果更佳. 实际流式信号不能预知未来信号, 但可以通过推迟编码时刻来达到字典预扫描的效果. 因此, 提出一种改进方

法, 以轻微增加编码延时为代价, 通过提前预判优化原子选择. 记  $K$  为当前扫描趟数,  $K$  为最大扫描趟数,  $J$  为目标拟合度, 则算法流程如图 8 所示.

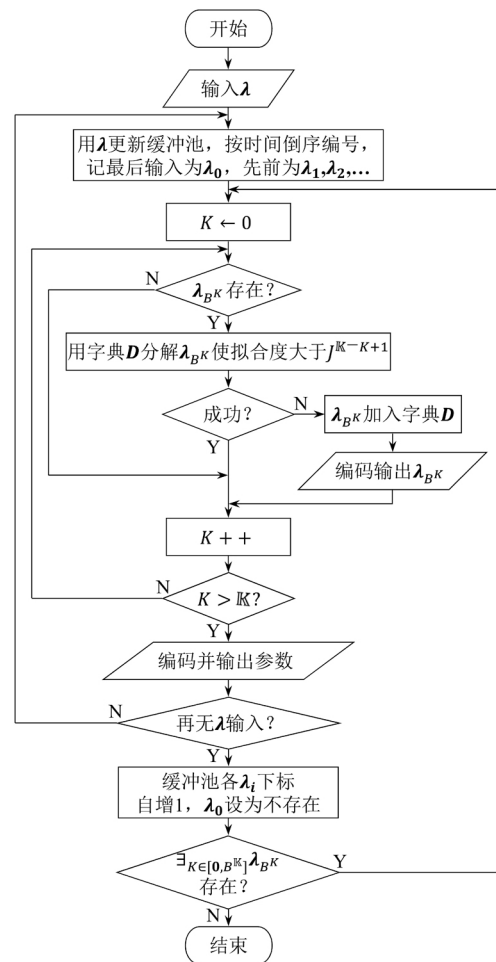


图 8 多趟字典构造方法

以最简单的 2 趟方案为例, 在信号输入的同时进行第 1 趟分析. 其字典的原子更新规则与 1 趟方案相同, 但拟合度阈值比较宽松. 在第 1 趟中, 若有原子更新, 则需要记录, 但丢弃一切拟合权值. 在第 2 趟分析中, 拟合度阈值比较严格, 与目标拟合度相同, 而且既记录原子更新, 又记录所有权值. 这种做法有利于在第 1 趟中优先选取关键性的原子, 以便使其可用于第 2 趟扫描中. 第 2 趟扫描的严格阈值则保证了最终拟合信号的质量.

图 9 展示了在使用 2 趟压缩编码时拟合度的变化情况. 由图 9 可见, 当  $J$  下降到 96% 时, 在预扫描中更新了字典, 而第 2 趟编码中又更新了 2 次字典. 由于有了一定的预判能力, 2 趟压缩编码能使拟合度在刚编码开始就比较稳定. 而且由于字典构造比较有针对性, 所以其原子的添加数量也比 1 趟算法少.

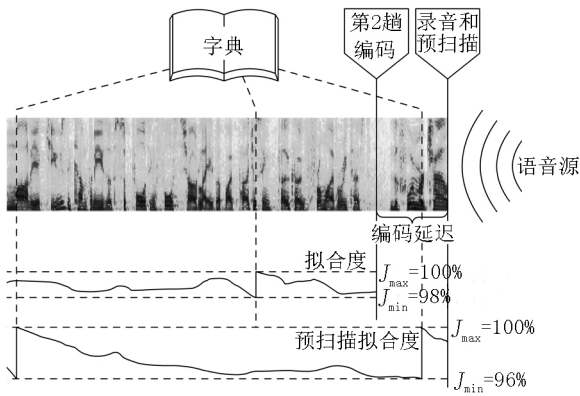


图9 2趟压缩编码示例

## 2 实验结果

首先将本文方法与 ODL 方法<sup>[15]</sup>进行比较. 然后给出本文方法使用不同参数的结果. 实验中使用到的所有音源数据均来自于 PTDB-TUG 数据集<sup>[23]</sup>. PTDB-TUG 包含了 20 人的录音片段, 所有录音片段均使用了麦克风和喉头仪进行同时录制, 并且数据集已经给出了通过喉头仪信号测得的准确基频值. 该算法首先通过对麦克风信号进行短时傅里叶变换获得了信号的幅值谱, 然后通过光滑连接各整数倍基频位置的幅值得到幅值谱包络. 由于原音频采样率为 48 000 Hz, 能表示的最高频率为 24 000 Hz, 而仅保留 0~4 000 Hz 的频谱范围即可使得语音被无障碍地辨别, 因此本实验先对原信号进行 4 096 点的短时傅里叶变换, 再通过仅使用低频的 512 个幅值点生成共振峰. 相当于该实验的处理信号的频率范围为 0~6 000 Hz. 因为 ODL 要求事先设定字典的规模, 为了更公平地与其比较, 在实验中, 本算法的字典规模与 ODL 一致.

### 2.1 压缩比计算

假设原幅值包络占用了  $A$  比特, 而压缩后仅占用  $A'$  比特, 则称  $A/A'$  为压缩比. 为简化实验, 本文直接使用 32 比特单精度浮点数来记录包络中与字典原子的每个元素, 并使用最少的比特数的无符号整数表示原子 ID. 本文的实验使用 (1) 式作为重构质量的目标. 该标准简称  $R_Q$  (reconstruction quality). 其中, 对本文提出的算法使用  $R_Q$  下限作为压缩目标, 而对 ODL 和 PCA 则使用平均  $R_Q$  作为压缩目标. 这种差异是根据算法本身的特性而决定的.

通过随机选取 PTDB-TUG 数据集中 5 男 5 女的全部录音片段, 并在本文算法中使用支持频率平移

的 3 趟扫描, 在重构质量  $R_Q = 98\%$  的情况下, 对比了本文算法与 PCA 和 ODL 之间的压缩比. 对所有录音, PCA 都能通过使用 6 个原子来使平均  $R_Q \geq 98\%$ . 这表明 PCA 的压缩比大约为  $6/512 \approx 0.011719$ . 而本文算法与 ODL 的压缩比则随输入样本而变, 将 ODL 的平均重构质量设为 98%, 其结果如图 10 和图 11 所示. 从图 10 和图 11 可见, ODL 的压缩比与 PCA 接近.

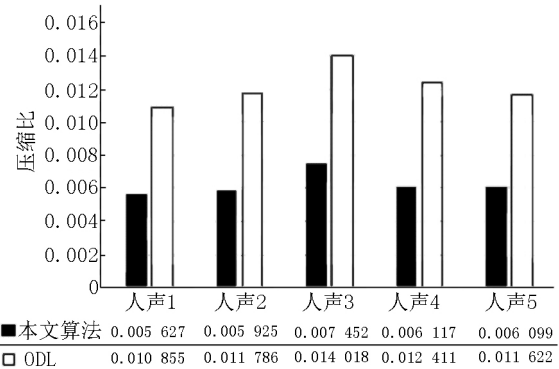


图10 女声压缩比

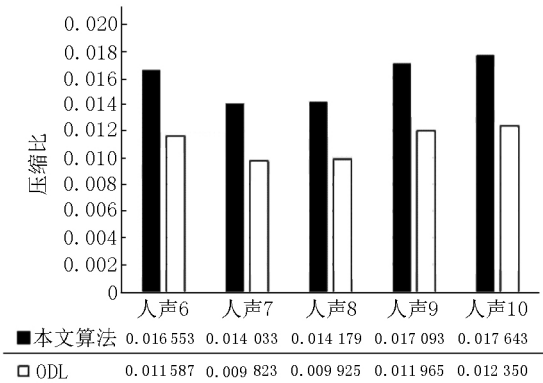


图11 男声压缩比

当压缩女声时, 本文算法的压缩比明显低于 ODL 和 PCA, 因此优于此 2 种算法. 但当压缩男声时, 本文算法压缩比看似比 ODL 和 PCA 更高. 这是由于本文算法的  $R_Q$  目标和 PCA、ODL 不同. 本文算法  $R_Q$  为不低于 98%, 而后 2 者则是平均 98%. 因此, 就平均还原质量而言, 本文算法更优. 图 12 使用了光滑频数密度图的方式直观地展示了本文算法与 ODL 还原质量的区别. 尽管 ODL 的确做到了平均  $R_Q \geq 98\%$ , 但其无法保证所有时刻的  $R_Q \geq 98\%$ . 反观本文算法, 则是各窗口无一例外地符合  $R_Q \geq 98\%$ . ODL 给出  $R_Q$  偶尔低于 98% 的结果分散于整个音频中的多个时间点, 这将导致解压还原后的音频在这些时间点出现可感知的瑕疵, 例如产生咔嚓声. 因此, 本文算法由于能控制  $R_Q$  下限而更适用于语音压缩.

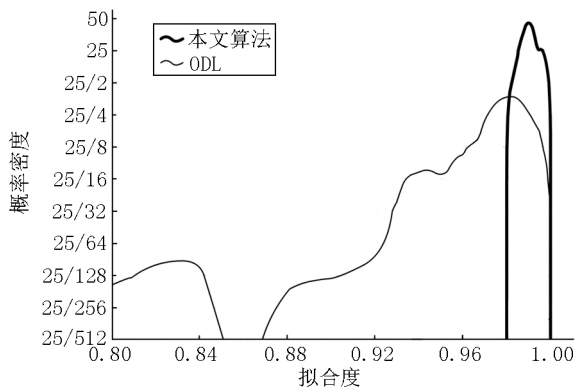
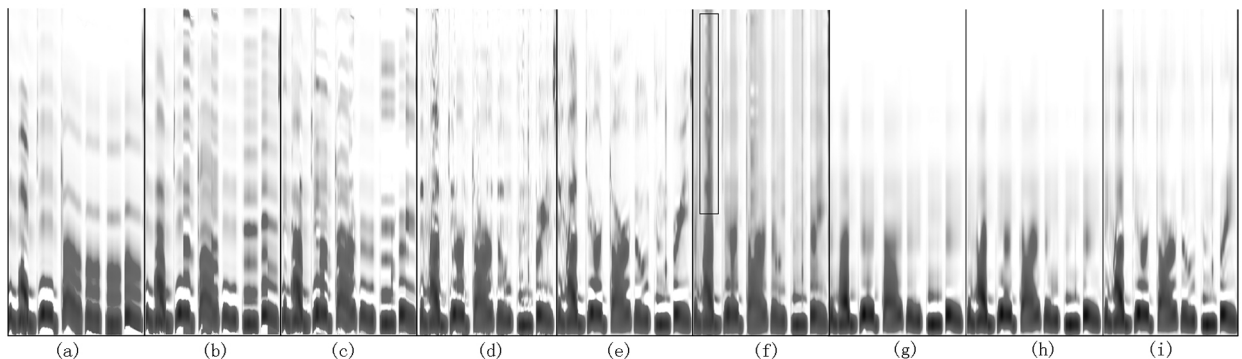


图 12 本文算法与 ODL 对人声 1 的重构质量频数密度对比

## 2.2 重构效果示例

以几段语音为例,对比了本文算法与 ODL 和 PCA 对幅值包络谱的还原结果.从图 13(a)~(d) 中可见,随着  $R_0$  下限的提升,包络谱的还原结果变得越来越接近原包络谱.虽然在较低  $R_0$  时(如

子图 13(a)) 本文算法还原结果在一些区间内与原包络谱有一定区别,但由于其所有还原结果均基于真实的字典原子在频率方向上的平移而实现,因此在听觉上不会有不像人声的声音出现.从低  $R_0$  的结果能较清楚地看到本文算法中原子平移的效果.还原的共振峰在频率方向上的峰和谷很锐利,而且峰和谷的位置随时间很平滑地变化.这种特性有助于得到流畅且咬字清晰的语音还原结果.与之相比,如子图 13(f) 中方框部分所示,即使  $R_0$  已经达到 0.980,ODL 在此处产生了不合乎语音特性的连续峰值.PCA 的每一个结果无论是频率方向还是时间方向上均较光滑,以致于在高频部分丢失了太多纹理信息.事实上,由于 PCA 每一时刻的结果都是来自于平均统计,导致了在频率方向上的光滑感,并将最终导致不自然的听觉感受.



(a)~(d) 为本文算法在平均  $R_0$  分别为 0.900 0.950 0.980 和 0.995 时的结果 (e) 为原包络 (f) 是 ODL 在  $R_0$  下限为 0.980 时的结果 (g)~(i) 分别是 PCA 在平均  $R_0$  为 0.950 0.980 0.995 下的结果.

图 13 原幅值包络及本文算法、ODL 和 PCA 的重构结果

## 3 总结

本文提出了一种新颖的语音共振峰包络的增量频移字典学习方法,它是整个语音压缩研究的重要一环,为最终的低码率的语音流编码打下了坚实基础.该方法能通过增量字典学习动态地适应语音流的变化,并保证其所有重构结果均不低于阈值.该算法能通过频率方向上平移字典原子来拟合包络谱,通过利用希尔伯特变换的近似平移特性,平移量能被高效且准确地估算出.该算法支持使用多趟扫描降低码率.其参数调整较为灵活,能适用于不同带宽要求、计算量要求和延迟要求.在后续的研究中,将着手压缩语音信号的另一重要组成部分——相位信息.

## 4 参考文献

- [1] Sunnydayal V, Kumar T K. Speech enhancement using posterior regularized NMF with bases update [J]. Computers and Electrical Engineering 2017 62: 663-675.
- [2] Gunawan T S, Khalifa O O, Shafie A A, et al. Speech compression using compressive sensing on a multicore system [EB/OL]. [2019-02-05]. <http://ieeexplore.ieee.org/document/5937130#>.
- [3] Al-Azawi M K M, Gaze A M. Combined speech compression and encryption using chaotic compressive sensing with large key size [J]. IET Signal Processing 2018, 12(2): 214-218.
- [4] Grosse R, Raina R, Kwong H, et al. Shift-invariant sparse coding for audio classification [EB/OL]. [2019-01-13]. <https://arxiv.org/abs/1206.5241>.
- [5] Févotte C, Bertin N, Durrieu J L. Nonnegative matrix factorization with the itakurasaito divergence: with application to music analysis [J]. Neural Computation 2009 21(3): 793-830.
- [6] Zibulevsky M, Pearlmutter B A. Blind source separation by sparse decomposition in a signal dictionary [J]. Neural Computation 2001 13(4): 863-882.

- [7] Elad M ,Aharon M. Image denoising via sparse and redundant representations over learned dictionaries [J]. IEEE Transactions on Image Processing ,2006 ,15 ( 12 ) : 3736–3745.
- [8] Mairal J ,Elad M ,Sapiro G. Sparse representation for color image restoration [J]. IEEE Transactions on Image Processing ,2008 ,17 ( 1 ) : 53–69.
- [9] Mairal J ,Bach F ,Ponce J ,et al. Supervised dictionary learning [J]. Advances in Neural Information Processing Systems ,2009 ,21: 1033–1040.
- [10] Bradley D M ,Bagnell J A. Differentiable sparse coding [J]. Advances in Neural Information Processing Systems ,2009 ,21: 113–120.
- [11] Yang Jianchao ,Yu Kai ,Gong Yihong ,et al. Linear spatial pyramid matching using sparse coding for image classification [C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition ,2009: 1794–1801.
- [12] Lu Xuan ,Wang Dingwen ,Shi Wenxuan ,et al. Group-based single image super-resolution with online dictionary learning [J]. Geomatics and Information Science of Wuhan University ,2016 ,2016 ( 1 ) : 84.
- [13] Peyré G. Sparse modeling of textures [J]. Journal of Mathematical Imaging and Vision ,2009 ,34 ( 1 ) : 17–31.
- [14] Warmuth M K ,Kuzmin D. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension [J]. Journal of Machine Learning Research ,2008 ,9: 2287–2320.
- [15] Mairal J ,Bach F ,Ponce J ,et al. Online dictionary learning for sparse coding [C]. Proceedings of the 26th Annual International Conference on Machine Learning ,2009: 689–696.
- [16] Mensch A ,Mairal J ,Thirion B ,et al. Stochastic subsampling for factorizing huge matrices [J]. IEEE Transactions on Signal Processing ,2017 ,66 ( 1 ) : 113–128.
- [17] Liu Jialin ,Garcia-Cardona C ,Wohlberg B ,et al. Online convolutional dictionary learning [C]. 2017 IEEE International Conference on Image Processing ( ICIP ) ,2017: 1707–1711.
- [18] Jolliffe I T. Principal component analysis [M]. New York: Springer-Verlag ,2005.
- [19] Hinton G E ,Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science ,2006 ,313 ( 5786 ) : 504–507.
- [20] Schölkopf B ,Smola A ,Müller K-R. Kernel principal component analysis [C]. Artificial Neural Networks-ICANN ,1997 ,1997: 583–588.
- [21] Schmitz M A ,Heitz M ,Bonneel N ,et al. Wasserstein dictionary learning: optimal transport-based unsupervised nonlinear dictionary learning [J]. SIAM Journal on Imaging Sciences ,2018 ,11 ( 1 ) : 643–678.
- [22] Aharon M ,Elad M ,Bruckstein A. K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representations [J]. IEEE Transactions on Signal Processing ,2006 ,54 ( 11 ) : 4311–4322.
- [23] Pirkker G ,Wohlmayr M ,Petrik S ,et al. A pitch tracking corpus with evaluation on multipitch tracking scenario [C]. INTERSPEECH 2011 ,12th Annual Conference of the International Speech Communication Association ,2011: 1509–1512.

## The Online Dictionary Learning Method of Incremental Frequency Shift for Speech Formant Envelopes

HUO Yingxiang ,TENG Shaohua<sup>\*</sup>

( School of Computer Science and Technology ,Guangdong University of Technology ,Guangzhou Guangdong 510006 ,China)

**Abstract:** Digital voice is widely used nowadays. The uncountable and continuous generating voice streams consume huge amount of network bandwidth and hard disk space. Hence ,under the pre-requisition of keeping the perceptual quality ,it is important to compress these voice signals to achieve minimum bit rates. Therefore ,a novel lossy compression algorithm for speech formant envelopes based on online dictionary learning method is proposed. By utilizing Hilbert transform ,the proposed method efficiently shifts the atoms of the dictionary over the frequency for better fitting the formant envelopes. Experimental results show that when the minimum reconstruction quality threshold is 99.5% ,in comparison of the uncompressed envelope data ,the proposed method achieves 99% bit rate reduction on average. Hence ,this method is applicable to scenarios that has limited band width and storage capacity ,and meanwhile can keep good perceptual quality.

**Key words:** voice stream; lossy compression; online dictionary learning

( 责任编辑: 冉小晓)