基于稀疏字典的听觉显著性计算①

陈 曦, 夏秀渝

(四川大学 电子信息学院, 成都 610064)

摘 要: 听觉注意显著性计算模型是研究听觉注意模型的基本问题, 显著性计算中选择合适的特征是关键, 本文从特征选择的角度提出了一种基于稀疏字典学习的听觉显著性计算模型. 该模型首先通过 K-SVD 字典学习算法学习各种声学信号的特征, 然后对字典集进行归类整合, 以选取的特征字典为基础, 采用 OMP 算法对信号进行稀疏表示, 并直接将稀疏系数按帧合并得到声学信号的听觉显著图. 仿真结果表明该听觉显著性计算模型在特征选择上更符合声学信号的自然属性, 基于基础特征字典的显著图可以突出噪声中具有结构特征的声信号, 基于特定信号特征字典的显著图可以实现对特定声信号的选择性关注.

关键词: 听觉选择性注意; 听觉显著图; 显著性; 字典学习

Auditory Saliency Calculation Based on Sparse Dictionary

CHEN Xi, XIA Xiu-Yu

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China)

Abstract: Auditory attention saliency computation model is one of the fundamental problems in the study of auditory attention model, and the key of this model is the selection of appropriate features. In this paper, an auditory significance calculation model based on sparse dictionary learning is proposed from the view of feature selection. The first step is getting the characteristics of a variety of acoustic signals by the K-SVD dictionary learning algorithm. Then the dictionary set is classified and integrated. Based on a selected feature dictionary, OMP algorithm is used for signal sparse representation. And the sparse coefficients are combined frame by frame to obtain the auditory saliency map. The simulation results show that this auditory saliency map computation model can achieve better correspondence characteristic with the nature attribute of acoustic signal in feature selection. The saliency map based on dictionary of basic characteristics can highlight the structure characteristics of noisy acoustic signal. The saliency map based on dictionary of special characteristics can achieve selective attention for certain signals.

Key words: auditory selective attention; auditory saliency map; saliency; dictionary learning

注意是一种在指定时间内大脑关注某种特定信息的能力, 根据参与器官的不同, 可以分为听觉注意、视觉注意等. 注意有两个基本特征: 指向性和集中性. 指向性主要指选择出现在同一时间的各种刺激; 集中性主要指对干扰的抑制, 其产生的范围以及持续时间取决于外部刺激的特点和人的主观因素. 学界普遍认为听觉注意是由自底向上(Bottom-Up)外源性听觉注意和自顶向下(Top-Down)内源性听觉注意的两种因素所驱动[1-3]. 自顶向下的注意因受到具体任务和人的主观意识的影响, 其研究结果往往呈现出较大的差异性[4-7], 并依赖自底向上的注意方式对信息的提取加工起作用.

研究听觉注意计算模型无论对生理心理学和计算机科学都具有重要的理论意义和实用价值. 目前国内外对听觉注意计算模型的研究主要集中在外源性听觉注意上, 即 Bottom-Up 听觉显著性模型. 现有的听觉

① 基金项目:四川省科技支撑项目(2011SZ0123,2013GZ1043)
收稿时间:2015-08-12;收到修改稿时间:2015-09-21

显著性模型主要参考了经典 Itti 视觉显著图计算模型. 该模型提取图像的三个初级视觉特征(颜色, 强度和方位), 然后对每一个特征进行中央周围差和标准化得到视觉显著图. 近年来国内外还提出用傅里叶变换、小波分析等算法对图像的纹理特征及运动显著图进一步强化[8,9]. Kayser 等人[10]借鉴 Itti 模型首先提出了一个听觉显著图计算模型, 他们将声音信号通过听觉外周计算模型得到听觉图谱, 然后对听觉图谱进行不同尺度的高斯滤波提取图像的强度、时间对比度、频率对比度等特征, 整合各种特征得到听觉显著图, 该计算模型初步实现了显著图的计算. Emine 等人[11]在 Kayser 模型的基础上增加了波形包络, 谱图, 速度, 带宽, 和音高等特征信息, 同时对频率通道分别处理得到声音信号的显著图, 将特征提取方法加以细化计算显著图. 文献[12]提出在音乐背景中对声音显著性特征进行提取的方法, 该方法对声源信号限定了内容. 目前的听觉显著性计算模型主要采用了人工选取的时域、频域、能量等各种声学特征, 采用一定的合并策略合成最终显著图, 但在表示声学信号自然属性的准确性和反映听觉感知特性的完整性上存在缺陷.

本文对自底而上和自顶向下听觉显著性计算模型进行研究. 根据 Itti 模型结构框架, 其中特征选取是关键, 但听觉显著性经典模型多来源于视觉模型, 选取的特征不一定符合声音自然属性. 因而, 本文提出通过稀疏字典学习算法从自然声音中自动学习各种声学特征, 并选择性地利用这些特征计算听觉显著图, 从而得到更符合听觉感知特性的显著性表示.

1　听觉显著性计算模型

1.1 经典听觉显著性计算模型

自底向上听觉显著性计算主要是对环境中“突兀”声音的响应, 最后以显著图的方式凸显值得关注的声音. 已有的听觉显著性提取模型大多都基于 Itti 的图像显著模型框架, 以 Kayser[10]提出的听觉显著性提取模型为例, 其原理流程图如图 1 所示.

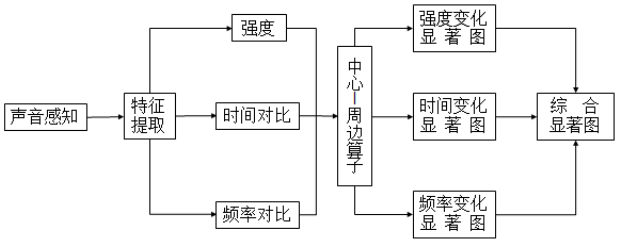

图 1　自底向上听觉注意模型框图

Kayser 模型在听觉前期处理中将声音信号转换成语谱图, 再通过二维高斯滤波器在不同尺度上提取语谱图的强度、频率对比度、时间对比度等特征, 利用中心-周边差(center-surround difference)算子计算各特征的显著度并进行跨尺度整合, 最后通过各特征显著度的线性合并得到声学信号的听觉显著图. 该方法一度成为计算听觉显著图的基础模型, 随后 Kalinli[13]等人在 Kayser 模型的基础上提出依赖词法和句法信息的使用概率作为判别条件, 采用不同的显著图归一化算法计算声学信号的听觉显著图. Duangudom[14]模型主要利用了信号的时频能量和时频调制特性, 考虑听觉谱图中时频接受域的输出并计算出信号的听觉显著图.

上述文献通过增加符合听觉特性的特征对模型加以完善, 这些特征都是基于经验人工选取的, 选取工作比较困难且计算的有效性较低. 为突出听觉特性和声音的自然属性, 本文将声音信号送入更贴近人耳听觉特性的 Mel 滤波器组得到声音信号的听觉图谱, 然后通过字典学习算法从自然声音中自动学习各种典型的听觉特征, 进而形成听觉显著图, 该方法避免了传统人工选取特征的困难, 更符合听觉感知特性.

1.2 基于稀疏字典的听觉显著性计算

显著图计算关键是选取合适的特征, 除根据先验知识人工选择外, 还可通过学习的方法获得自然声音的听觉特征, 利用其特征计算得到的显著图更符合声源的信息特点, 减小人工干涉的影响, 本文利用稀疏字典 K-SVD 学习算法获取声学信号特征.

1.2.1 信号稀疏表示及 K-SVD 算法

语音信号具有典型的稀疏性, 利用稀疏表示就能提取出信号特点, 仅用数个特征值即可. 目前 K-SVD 算法在处理稀疏信号上应用较广, 能够简便高效地获得信号的特征原子.

设听觉谱信号为 Y, 字典为 D 和系数矩阵 A, 则 K-SVD 的目标函数为:

$$\min_{D,A}\left\{\left\|Y-DA\right\|_F^2\right\}\ s.t.\ \forall i,\left\|a_i\right\|_0\le T_0 \tag{1}$$

为了使式(1)成立, 在每次迭代的过程中需要进行对 $\left\|Y-DA\right\|_F^2$ 进行 SVD 分解, 其分解共分两个步骤: 稀疏分解和字典更新.

① 稀疏分解

假设字典 D 是固定的, 预设定 $\left\|Y-DA\right\|^2=e$ 使得

$e \le \varepsilon$. 基于字典 D 对 Y 进行稀疏分解得到系数矩阵 A, 惩罚项为:

$$\left\|Y-DA\right\|_F^2=\sum_{i=1}^{n}\left\|y_i-Da_i\right\|_2^2 \tag{2}$$

解决式(2)可用正交匹配追踪(OMP)算法.

② 字典更新

基于 SVD 字典更新原理, 在字典更新阶段, 对字典中的原子逐个进行更新, 每次更新一个字典原子和相应的系数, 惩罚项为:

$$\left\|Y-DA\right\|_F^2=\left\|Y-\sum_{j=1}^{k}d_ja_T^i\right\|_F^2=\left\|E_k-d_ka_k^i\right\|_F^2 \tag{3}$$

其中 $E_k=Y-d_ja_T^i$ 表示去掉第 k 个原子 d_k 后稀疏表示的误差, 对式(3)中的 E_k 进行 *SVD* 分解得到更新的原子 $\hat{d}_k$ 及系数矩阵 A_k. 为了使信号得到稀疏表示, 对 A_k 进行补偿, 得到新的 E_k^r, 使得(3)式成立, 将 E_k^r 通过 *SVD* 分解:

$$E_k^r=U\nabla V \tag{4}$$

其中∇中的奇异值是由大到小排列的, 则得到 U 的第一列为第一个原子 $\hat{d}_k$, 表示 V 的第一列乘以∇(1,1)为更新后系数矢量 a_k^i 的解. 接着进行下一次迭代, 第一步和第二步交替完成, 得到信号的稀疏表示矩阵及相应的字典原子.

1.2.2 基于稀疏字典的听觉显著性计算方法

通过 K-SVD 算法来学习自然声音的典型特征, 经学习得到的字典中每个原子都是声音信号的一个典型特征, 根据显著图计算模型可以考虑用字典原子作为显著图的特征, 用这些特征滤波器对声音进行滤波, 经整合后可形成听觉显著图. 但由于训练出的初始字典原子个数较多, 多特征显著图合并时还有相互抵消的现象, 导致模型计算量较大且合并后的显著图区分度不够明显. 为此本文提出直接利用声学信号稀疏表示系数矩阵来计算声学信号显著度的方法, 即直接将信号单元的稀疏系数相加得到最终的听觉显著度. 这里利用系数绝对值直接相加形成显著度曲线, 具有比通过直接滤波整合计算显著图方法更低的算法复杂度, 提升了算法的效率和信号的区分度. 本文听觉显著图的计算总体框架如图 2 所示.

具体步骤如下:

① 将一维的声音信号通过短时傅里叶变换转换为语谱图, 考虑人耳听觉特性及字典原子大小对 K-SVD 算法学习效率的影响, 进一步采用 Mel 滤波器组(24 个三角滤波器组)滤波将语谱图转换为听觉谱图. 普通图像和听觉谱图虽然都是二维图像, 但两者是有区别的, 普通图像两个维度物理意义完全一样, 均表示空间分布. 而听觉图谱第一维是时域维度, 第二维是频域维度, 这两维的物理意义完全不同.

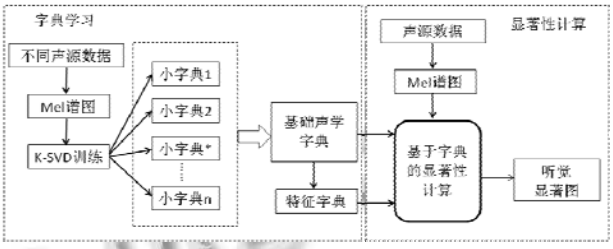

图 2 本文听觉显著图计算模型结构

② 选取不同声源的输入信号通过 K-SVD 算法进行字典学习. 图像应用中原子形状常取正方形, 而声学信号更多地表现为频谱随时间的变化, 所以提取声学信号特征时, 本文考虑时频特征, 以每帧信号的 Mel 谱为基础进行稀疏字典学习, 原子大小设定为 24*1. 通过学习得到对应不同种类声信号的小字典集, 然后通过统计分析挑选出部分原子合成基础声学特征库 D_n.

③ 将该字典固定, 对实际输入信号在字典 D_n 上进行稀疏分解, 可以获得分解系数矩阵 A_i, 然后将每帧信号稀疏系数绝对值 $\left|a_i\right|$ 叠加就得到该输入信号最终的听觉显著度曲线. 由于字典原子具有结构化特征, 该基础声学字典可用于区分具有结构特征的自然声和不具有结构特征的噪音.

④ 另外统计分析发现各原子在每类声信号稀疏表示中出现的概率不同, 即不同种类的声音具有不同的特征, 因此也可从基础声学特征库 D_n 中挑选部分原子构成某类特殊声音的特征字典, 用于特定声音的显著性计算, 从而实现稀疏分解的简化计算和有偏向的显著性计算.

2 仿真实验及结果分析

实验选取语音、猫叫、鸟鸣、风扇声各一段, 所用纯净语音选自 TIMIT 语音库, 其中男女声各 2 句, 信号采样频率 16 kHz, 帧长为 512 个样点, 将一帧的 Mel 谱维数作为字典原子大小 24*1.

2.1 字典学习及其统计分析

我们将语音、猫叫、鸟鸣、风扇声分别作为输入

信号, 采用 K-SVD 算法学习得到各种声音的小字典. 每种声音时频原子大小为24*1,每个字典大小为256个原子, 图 3 为以上类型声音的时频字典, 为了节省空间这里只给出了各声音字典的部分典型原子.

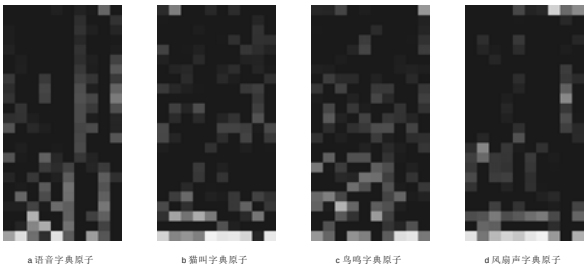

图 3　各种声学信号特征字典

从图 3 可以看出时频字典能很好地反映出各种声信号的时频结构, 字典中的每个原子能够表示声学信号的局部时频域特征. 不同类信号学习得到的字典也不同, 他们的主要特征结构不同. 另外还统计了每类声学信号字典原子在稀疏矩阵中出现的概率(如图4所示).

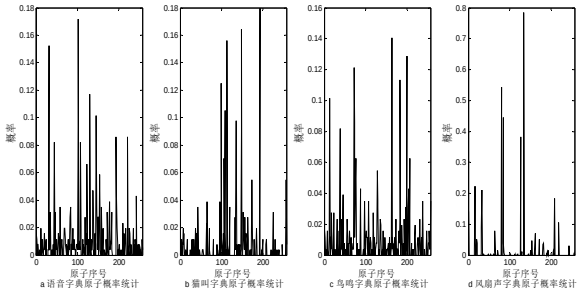

图 4　不同信号字典原子概率分布

通过图 4 可以看出每类声学信号各个字典原子在稀疏矩阵中出现的概率区别较大, 可以认为某些原子就是信号重要特征. 为提高信号稀疏表示的计算效率, 可挑选出现概率大的原子构成特征小字典. 图 3 表示了各声学信号引用概率大的 10 个原子.

2.2 基于字典的听觉显著图

通过 2.1 节的实验分析可知, 用于显著度提取的特征字典可以从小字典集中挑选部分原子构成. 首先我们从小字典集中挑选出现概率大的原子构成一个基础特征字典, 通过对语音、猫叫、鸟鸣、风扇声进行字典学习(各字典原子个数设定为256个), 进而对每个小字典的原子特征进行统计, 提取出现概率大于 3 倍概率均值的原子作为该小字典的特征原子, 每个小字典分别可以挑出 10~20 个原子, 而后将几组小字典组成为具有结构化特性的基础声学字典(含原子 49 个). 该字典的原子皆具有结构化特点, 所以用基础声学字典来计算信号显著度时, 可以区分结构化声音与非结构化噪声. 截取前述四种声音并和白噪声拼接后送入显著性模型计算显著度, 实验结果如图 5.

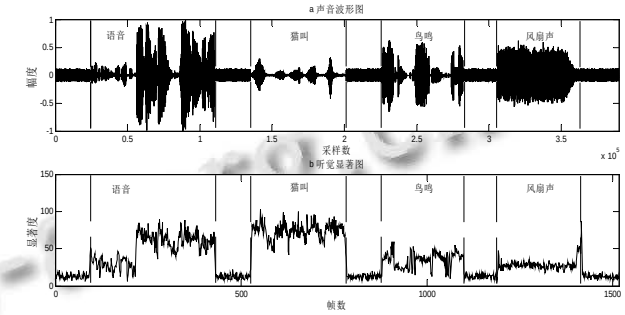

图 5　声音显著图

由图 5 看出, 白噪声不具有结构化特性, 其稀疏分解系数较小, 所以对应的显著度曲线取值小, 而语音、猫叫、鸟鸣、风扇声这些具有结构化特性的声音显著度取值都比较大.

另外对于该字典, 每类声音对各原子的使用概率也不同, 所以还可以挑选部分在某类声音中出现概率高, 而在其他类声音中出现概率低的原子构成特征字典, 以此得到具有偏向性的显著图. 如我们想选择性关注语音, 则从基础声学字典中挑选语音稀疏表示中使用概率大的原子构成语音特征字典, 然后基于该语音特征字典计算输入信号的显著度, 就可以实现对语音的选择性注意.

实验中我们从基础声学字典中挑选出 18 个原子构成语音特征字典, 12 个原子构成猫叫声特征字典, 下图分别是基于语音和猫叫声特征字典得到的具有偏向性的显著图.

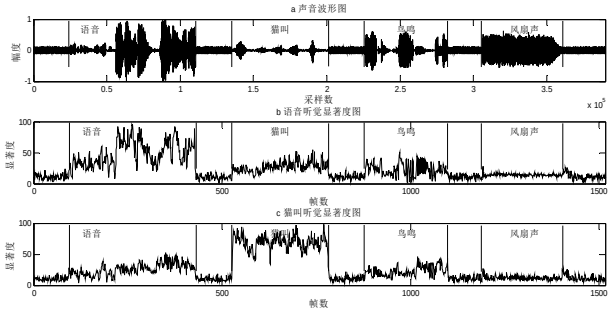

图 6　选择性注意显著图

图 6-b 中只有语音段具有相对较高显著度，图 6-c 中猫叫声段显著值明显偏高，基于语音和猫叫特征字典的显著图表现出对两种特定声音的偏向性，不仅非结构化的噪声显著度低，而且和特性不同的另三种声音显著度也低. 从而验证了分类小字典具有一定的选择特性，可用于实现人类听觉自顶向下选择性注意.

3 结论

本文针对人的听觉注意方式提出了一种基于特征字典的听觉显著图计算模型. 该模型在稀疏字典学习的基础上提取了信号的特征字典，并利用稀疏系数矩阵表示出了信号的显著图. 实验结果表明，本文方法实现了对具有结构特性声音的显著性注意计算并通过对特征原子的提取应用实现了对声源的指向性注意. 本文模型兼具自顶向下和自底向上的选择性注意功能，降低了对显著图计算的复杂度和计算量，提取特性未加以人工干涉，在实现构建听觉选择性注意模型方面更具现实意义. 在以后的研究中，希望能用更多的声学特征来完善本文模型，针对声源的分类识别上有进一步的研究.

参考文献

1 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254–1259.

2 Itti L, Koch C. Computational modeling of visual attention. Nature Reviews Neuroscience, 2001, 2(3): 194–203.

3 Tsotsos J, Culhane S, Kei WW, et al. Modeling visual attention via selective tuning. Artificial Intelligence, 1995, 78 (1): 507–545.

4 Borjia IL. State-of-the-art in visual attention modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(1): 185–207.

5 Yarbus A. Eye Movements and Vision. Plenum Press, 1967.

6 Foulsham T, Under WG. What can saliency models predict about eye movements spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision, 2008, 8(2).

7 Hayhoe M, Ballard D. Eye movements in natural behavior. Trends in Cognitive Sciences, 2005, 9(4): 188–194.

8 徐贵力,毛罕平.利用傅里叶变换提取图像纹理特征新方法. 光电工程,2004,31(11):55–58.

9 张焱,张志龙,沈振康.一种融入运动特性的显著性特征提取方法.国防科技大学学报,2008,30(3):109–115.

10 Kayser C, Petkov CI, Lippert M, et al. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology. 2005. 1943.

11 Kaya EM, Elhilali M. A temporal saliency map for modeling auditory attention. Department of Electrical and Computer Engineering.

12 Vaclav B, Rainer M, et al. A model-based auditory scene analysis approach and its application to speech source localization. Acoustics, Speech and Signal Proccessing (ICASSP). Prague Congress Centre Prague, Czech Republic. 2011. 2624–2627.

13 Kalinli O, Member S, Narayanan S. Prominence detection using auditory attention cues and task-dependent high level information. IEEE Trans. on Audio, Speech, And Language Processing, 2009, 17(5).

14 Duangudom V, Anderson DV. Using auditory saliency to understand complex auditory scenes. Proc. of the 15th European Signal Processing Conference(EUSIPCO 2007). 2007. 1206–1210.

15 王雪君,夏秀渝,张欣,何培宇.新的听觉注意显著图计算模型研究.信号处理,2013,29(9):1142–1147.