听觉选择性注意的认知神经机制与显著性计算模型

刘 扬[1,3] 张苗辉[2] 郑逢斌[1,3]
(河南大学智能技术与系统重点实验室 开封 475004)[1]
(上海交通大学图像处理与模式识别研究所 上海 200240)[2]
(河南大学计算机与信息工程学院 开封 475004)[3]

摘 要 根据听觉认知神经信息处理的结构和功能,借鉴图像处理原理实现显著性计算方法,提出了一种基于选择性注意的认知神经机制的听觉显著性计算模型。该模型兼容了自上而下和自下而上两种听觉注意机制,可很好地模拟人类的听觉注意系统。在仿真和自然音频实验中,本模型在选择性注意的显著性提取、背景音抑制等方面都取得了令人满意的结果。

关键词 听觉注意,选择性注意,听觉显著图,听觉模型,认知神经计算
中图法分类号 TP391 **文献标识码** A

Cognitive Neural Mechanisms and Saliency Computational Model of Auditory Selective Attention

LIU Yang[1,3] ZHANG Miao-hui[2] ZHENG Feng-bin[1,3]
(Key Laboratory of Intelligent Technology and Systems, Henan University, Kaifeng 475004, China)[1]
(Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China)[2]
(College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, China)[3]

Abstract According to structure and function of auditory cognitive neural information processing, a new auditory saliency computing model based on mechanisms of selective attention cognitive neural was proposed in this paper, and referring to principle of image processing, algorithms of auditory saliency were presented. This model simulates both the bottom-up and top-down human auditory attention mechanism. In selective attention saliency extraction and background noise restrain, the model has achieved satisfactory results in simulation and natural audio experiments.

Keywords Auditory attention, Selective attention, Auditory saliency map, Auditory model, Cognitive neural computing

1 引言

鸡尾酒会效应(Cocktail Patty Effect, CPE)是人类听觉系统(Human Auditory System, HAS)在处理外界环境中的声音信息时,听觉注意神经信息处理机制发挥重要作用的体现,它可帮助人类从复杂的背景噪音环境中快速精确地提取出感兴趣或重要的声音内容,并据此做出进一步的反应。了解和建立听觉注意神经信息处理计算机制对听觉场景内容自动分析与理解、语音盲源分离或独立分量分析等理论具有重要指导意义。而在多人多方对话中的语音分离、水声传感信息融合系统、飞行语音对讲等领域系统也具备重要应用价值。

HAS的听觉认知神经信息处理机制非常复杂,研究其结构与功能将有助于了解听觉感知过程和模拟。近年来,人们已开始研究利用HAS信息处理原理来改善音频信息系统性能,并取得了一定成效。文献[1]提出一种自低向上(bottom-up, BU)基于数据驱动的听觉显著性注意算法,并在音调及重音分类中取得了较好效果。文献[2]模拟人耳听觉外周的信息处理特征,根据水声信号的特点,提出了一种基于Gammatone-Meddis的听觉模型,并将其用于水声目标分类特征的提取。文献[3]针对听觉滤波器组语音信号时域滤波输出的各个子带信号,用分数阶Fourier变换方法提取声学特征,取得了比MFCC特征更好的汉字孤立词语音识别结果。但上述文献设计的听觉模型没有考虑自顶向下(top-down, TD)听觉注意机制的影响。文献[4]提出一种复杂音频场景声音信号分类和位置分析方法,并将其运用于语音和音乐的分类。文献[5]提出一种可用于声学场景自动分析的基于听觉显著性的听觉注意的计算模型,但其模型参数设置依赖于声源信息内容。

针对听觉认知的神经信息处理机制建模和仿真问题,参照听觉认知神经信息处理结构及功能,本文提出了一种基于

到稿日期:2012-08-10 返修日期:2012-12-04 本文受国家航天局遥感论证中心项目(科工技2010A03A0800),河南大学自然科研基金(2010YBZR046),河南大学第11批教学改革重点项目资助。

刘 扬(1971-),男,硕士,副教授,硕士生导师,CCF会员,主要研究方向为媒体神经认知计算、时空信息高性能处理,E-mail:ly.sci.art@gmail.com;张苗辉(1979-),男,博士生,讲师,主要研究方向为视频监控、视频目标跟踪;郑逢斌(1963-),男,教授,博士生导师,主要研究方向为空间数据处理、自然语言理解、软件工程。

选择性注意的认知神经信息处理机制的听觉显著性计算模型。该模型综合考虑自低向上 BU 和 TD 两方面的听觉注意原理，建立听觉显著性提取框架，并从信息处理角度给出了感知特征提取、听觉显著图生成以及注意控制回路的相关算法。

2 听觉认知神经信息处理机制

2.1 听觉神经信息处理结构

HAS 的听觉通路(Auditory Pathway)是指与听觉产生相关的一系列神经解剖结构。听觉通路包括听觉外周和听觉中枢。听觉外周有耳蜗的毛细胞(HC)、螺旋神经节(SG)、延脑的耳蜗复核(CN)、桥脑的上橄榄复核(SOC)、中脑的下丘(IC)、丘脑的内膝体(MGB)。听觉中枢包括大脑皮层的颞叶的 A1、A2、A3 区和岛颞区(IT)等。

听觉中枢信息传递通路比较复杂，如图 1 所示，声音在 HC 产生的神经冲动，经 SG 达到 CN。CN 发出的第二级听神经纤维大部分交叉到对侧的 SOC，小部分不交叉的纤维则终止于本侧的 SOC。SOC 接收本侧和对侧的信息，处理后的信息经 IC、MGB 至颞横回的 A1 区，完成听觉外周的信息传递。在听觉中枢中，A1 又与听皮层 A2、A3 区等听觉皮层联络区发生联系。

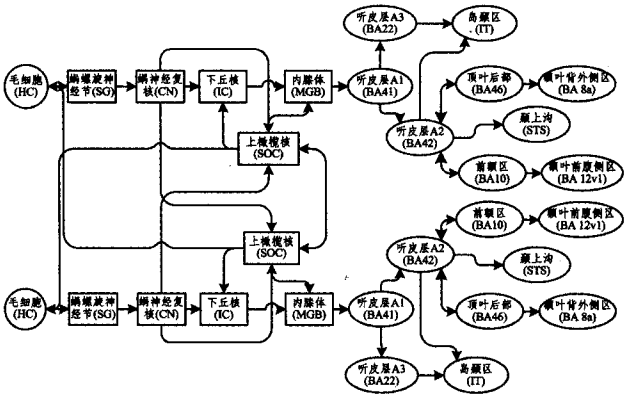


图 1 HAS 双耳听觉神经信息处理结构

2.2 听觉认知信息处理功能

在人类认知的基本架构中，认知信息处理过程包括贮存与提取、信息加工、输入与输出、知识运用 4 个步骤。由于听觉认知神经系统是巨大的，极其复杂的，对它进行建模非常困难，而神经系统是人类认知架构的基础，因此搞清 HAS 信息处理基本流程是进行认知建模的前提。认知神经相关研究表明，HAS 中随加工级别提升，参与的神经元数逐级增加。听觉认知信息处理功能分别在听觉传导路和高级听觉中枢进行加工，声音强度的分辨过程在低级听觉中枢基本完成，频率分析在皮质下各级中枢分别进行处理[6]。

听觉外周的认知信息处理功能主要是听觉信息产生、传导和初步分析。其中 HC 的主要功能是解决声音的频率拓扑的强度编码。CN 通过侧抑制增加系统的频率分辨率，整合神经脉冲序列中声强的信息，提取并通过环路延时暂存脉冲的时间信息。SOC 与 IC 对双耳信号中的强度差和时间差进行分析，执行声音的空间定位整合功能[7]。MGB 初步完成音强、频率、节奏、方位等听觉基元信息分析过程。

听觉中枢中负责声音的位置分析和内容分析。与视皮层相似，听皮层上存在着按照频率拓扑区排列的功能柱结构。A1 区的第一听觉应答野完成基音处理。A2 区的第二听觉应答野完成泛音组合处理[8]。A2 区后的信息流向存在背侧和腹侧两条通路。背侧通路由 A2、BA46、BA8a 构成了加工空间信息通路；腹侧通路由 A2、BA10、BA12v1 形成了听觉物体或模式识别的内容通路[9]。

综上所述，在 HAS 听觉神经信息处理的基础上，建立如图 2 所示的 HAS 的听觉认知框架，其在听觉神经系统的认知信息处理上兼具如下功能：

(1)声音在 HAS 中采用多频率多通道并行处理；

(2)HAS 在听觉外周解决声音基元信息的提取，听觉中枢实现声音语义对象分析；

(3)在听觉外周中，各个传导路逐级都存在空间方位与强度信息，二者被同时处理；

(4)在听觉中枢中，内容分析和方位分析分别进行，听皮层神经元对声音的频谱、时间、空间综合信息的处理，有精细的分工和高效的协同；

(5)在整个 HAS 通过多级反馈控制实现声音处理的选择性注意，其中方位、强度和频率信息在选择性注意起重要作用；

(6)HAS 存在缓存回路，时态记忆存储、音调、强度和声谱对听觉内容处理起重要作用。

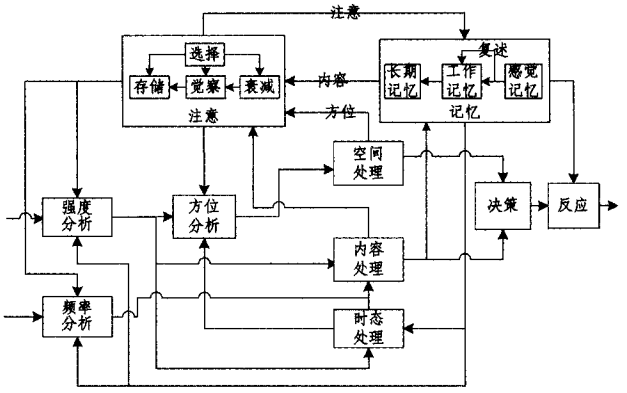


图 2 HAS 听觉认知框架

3 基于选择性注意认知神经机制的听觉显著性计算模型

考虑到 HAS 听觉认知模型的结构和计算的复杂性，下面针对听觉认知神经的选择性注意简化功能实现计算模型和算法。图 3 是本文给出的基于选择性注意认知神经机制的听觉计算模型，包括听觉信息处理通路模拟和听觉皮层信息处理模拟两部分。其中前者包括声音感知、特征抽取、显著图生成等模块，后者包括听觉信息时空内容处理以及注意振荡处理环路等模块。

注意振荡处理环路实现对被注意特征的选择，选择满足如下规则：

(1)与众不同的对象应获得较高的显著性；

(2)频繁出现的特征需要被抑制；

(3)显著频谱的分布时间应该集中；

(4)应考虑高阶特征(说话人、录音场景变化等)的影响。

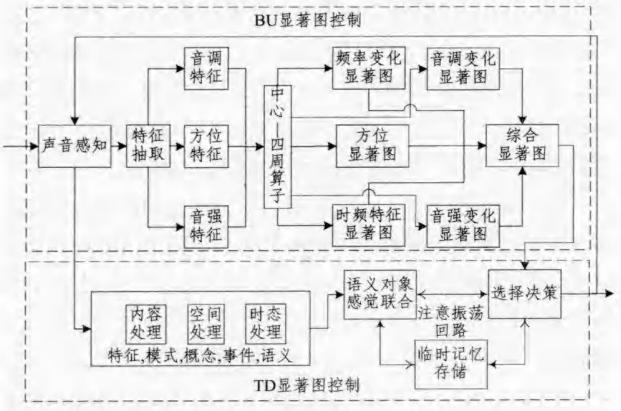


图 3　基于选择性注意认知神经机制的听觉显著性计算模型

3.1　音频的感知处理

听觉心理与生理研究表明，Mel 频率描述符合耳蜗 HC 的音频处理结构[10]，但 MFCC 特征提取通常采用基于中心频率按 Mel 刻度均匀分布的三角形滤波器组进行频域滤波，其特征提取方法并不完全符合听觉特性中的临界频带特性，且相邻频带之间的频谱能量相互泄露严重，无法反映共振特性。这里采用基于耳蜗基底膜模型的 Gammatone(GT)滤波器来模拟听觉感知处理过程。GT 滤波器仅需很少参数就能够体现基底膜尖锐的滤波特性[11]，GT 滤波器的冲击响应为：

$$g_m(t)=t^{n-1}\mathrm{e}^{-2\pi b_m t}\cos(2\pi f_m t+\phi_m)U(t),1\leqslant m\leqslant M \quad (1)$$

式中，n 为滤波器阶数，实验表明当 $n=4$ 时的 GT 滤波器能很好地模拟基底膜的滤波特性；f_m 是滤波器的中心频率，ϕ_m 为相位，b_m 为等效带宽，M 为滤波器个数，$U(t)$ 为阶跃函数。根据人的听阈范围，选择 24 个听觉临界频带的中心频率作为 GT 滤波器的中心频率，根据语音信号的采样频率即可确定 GT 滤波器的个数。根据式(1)的冲激响应函数，即可得到滤波器的频率响应特性。经过 GT 滤波器组获得的时频图 $GFT(f,t)$为：

$$GTF(f,t)=20\log_{10}(\mathrm{GTG}(\mathrm{Auditory}(t),F_s,N)) \quad (2)$$

对于如下测试函数：

$$y(x)=[\sin(x/30)(x/100+1)]\sin(100-[\sin(10+x/4+\sin(x/8))]x) \quad (3)$$

设置采样频率为 16000Hz，x 取 0～100π 产生波形 A1，x 取 100π～0 产生波形 A2，A1 与 A2 叠加产生混合波形 A3。A1，A2，A3 波形数据经 128 通道的 Gammatone 滤波组产生的频谱如图 4 所示。

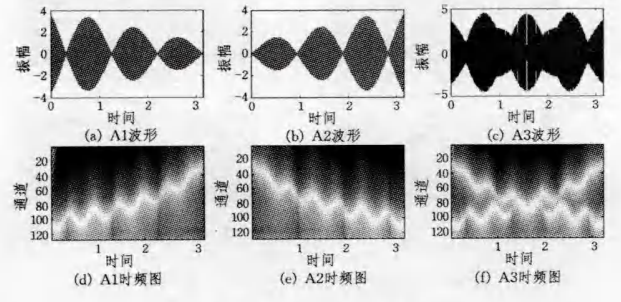


图 4　仿真实验音频波形和时频图

3.2　感知特征提取

音频处理的基本感知特征有音调、音强、方位以及相应特征随时间的变化情况。为计算音频的感知特征，实验表明通过 Daubechies-1 小波变换可模拟 MGB 和初级听觉皮层应答野的感知特征提取过程。

$$[TF(f,t),FC(f,t),IC(f,t),PC(f,t)]=DWT2(TF(f,t),"db1") \quad (4)$$

式中，近似子图 $TF(f,t)$为时频特征图，细节子图 $PC(f,t)$为音调特征图，垂直子图 $IC(f,t)$为音强变化特征图，水平子图 $FC(f,t)$为频率变化特征图。为获得音频显著图，模拟听觉皮层通路的中心-四周算法，对上述提取的感知特征 FG_j 实现金字塔多分辨率的中心-四周差分算法：

$$CSG_j(c,c+\Delta)=|FG_j(c)-FG_j(c,c+\Delta)| \quad (5)$$

式中，$c=2,3,4;\Delta=3,4;j=TF,PC,IC,FC$。经中心-四周算法处理后，式(3)测试函数所产生的仿真音频处理后的听觉感知特征如图 5 所示。

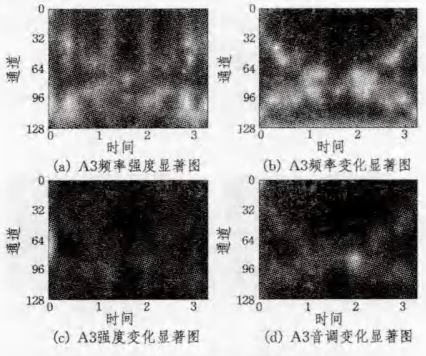


图 5　仿真音频听觉感知特征图

3.3　听觉显著图生成

根据神经竞争计算的 Winner Take All(WTA)机制[12]，对各个特征图处理得到特征显著子图：

$$ST_j(f,t)=\arg\underset{c,\Delta}{\mathrm{Max}}(CSG_j(c,c+\Delta)) \quad (6)$$

式中，$c=2,3,4;\Delta=3,4;j=TF,PC,IC,FC$。对 $ASG_{c,j}$ 作归一化计算得：

$$NST_j(f,t)=Norm(ST_j(f,t)),j=TF,PC,IC,FC \quad (7)$$

融合 4 种特征的音频显著子图，获得最后音频显著图：

$$ASG(f,t)=\mathrm{Max}(NST_{TF}(f,t),NST_{FC}(f,t),NST_{PC}(f,t),NST_{IC}(f,t)) \quad (8)$$

上述仿真音频最后生成的听觉显著子图如图 6 所示。

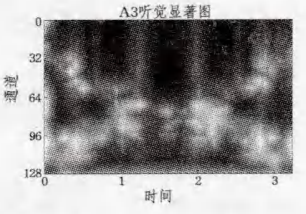


图 6　仿真音频 A3 听觉显著图

3.4　注意控制回路

上述过程模拟了 BU 机制的听觉注意的显著图生成方法。当注意焦点产生竞争，引起注意发生转移时，为实现 TD 机制的听觉注意计算，注意振荡回路将决策控制信息反馈到

输入产生抑制，进而降低非注意对象的时频特征。若当前输入时频特征 $TF(f,t)$ 由若干时频特征组成 $tf_i(f,t)$，假定在注意振荡回路中已提取和存储了听觉对象的时频特征 $tf_i(f,t)$，则：

$$TF(f,t)=\sum_{i=1}^{N} tf_i(f,t) \tag{9}$$

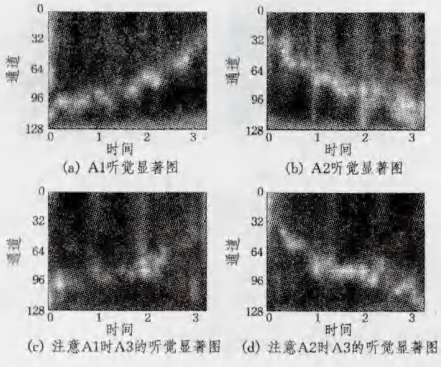
如果被注意特征为 $tf_j(f,t)$，则注意振荡回路反馈控制后的时频特征为：

$$STF(f,t)=TF(f,t)-\frac{\beta}{\Delta t}\sum_{t}^{t+\Delta t} tf_j(f,t) \tag{10}$$

式中，β 为选择性抑制参数，考虑到随着显著焦点的出现频率增加而造成其显著性降低的认知神经机制，最后被注意特征为：

$$STF(f,t)=\sum_{i=1}^{N}\left(tf_i(f,t)-\frac{\alpha}{\Delta t}\sum_{t}^{t+\Delta t} tf_i(f,t)\right) \tag{11}$$

式中，α 为注意转移参数，$\frac{1}{\Delta t}\sum_{t}^{t+\Delta t} tf_i(f,t)$ 为短时平均特征。图 7 显示了式(3)测试函数所产生的听觉显著图。图 7(a)为 A1 对象的音频听觉显著图，图 7(b)为 A2 对象的音频听觉显著图，图 7(c)为在 A1 和 A3 混合音频中注意 A1 对象的听觉显著图，图 7(d)为在 A2 和 A3 混合音频中注意 A2 对象的听觉显著图。

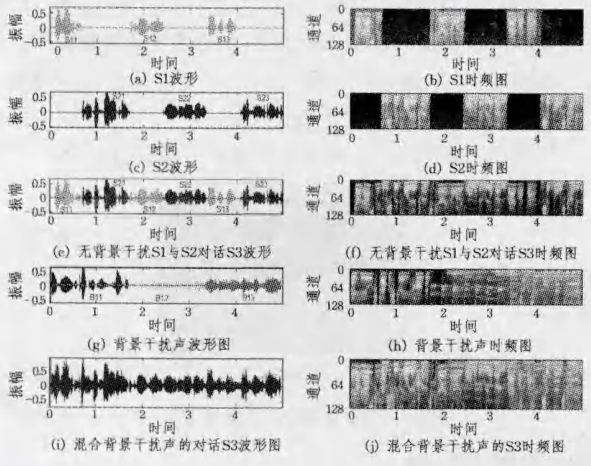


(a) A1听觉显著图　(b) A2听觉显著图

(c) 注意A1时A3的听觉显著图　(d) 注意A2时A3的听觉显著图

图 7　仿真音频选择性注意听觉显著图

4　实验与仿真



(a) S1波形　(b) S1时频图

(c) S2波形　(d) S2时频图

(e) 无背景干扰S1与S2对话S3波形　(f) 无背景干扰S1与S2对话S3时频图

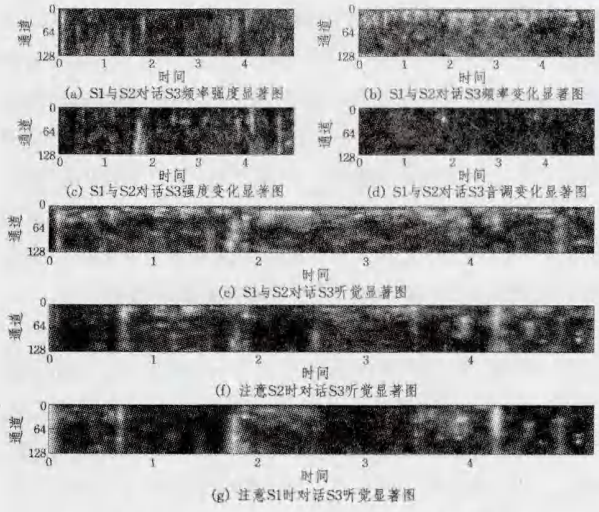
(g) 背景干扰声波形图　(h) 背景干扰声时频图

(i) 混合背景干扰声的对话S3波形图　(j) 混合背景干扰声的S3时频图

图 8　自然实验音频波形和时频图

为了验证本文计算模型的正确性和有效性，分别对 BU 和 TD 视觉注意机制进行了实验。实验音频在波士顿大学新闻广播语料库和中央电视台新闻联播录音中各随机选择 10 组，其时序同步关系如图 8 所示。图 8(a)的 S11，S12，S13 这 3 句组成女声 S1，图 8(c)的 S21，S22，S23 这 3 句组成男声 S2，二者对话组成图 8(e)。针对上述 3 句对话，背景干扰声分别由图 8(g)的女声(B11)、纯音乐(B12)、男女声合唱音乐(B13)组成。图 8 右列为对应波形的相应频谱图。

图 9 显示了无背景音干扰时自然音频选择性注意听觉显著性的处理情况。图 9 上两行分别显示了无背景干扰对话音频 S3 特征显著图。其中图 9(a)为时频特征图 TF，图 9(b)为频率变化特征图 FC，图 9(c)为音强变化特征图 IC，图 9(d)为音调特征图 PC，图 9(e)为自下而上机制的听觉显著图。图 9(f)和图 9(g)显示了分别注意 S2 和 S1 时 S3 的选择性注意听觉显著图。对比图 9(e)、图 9(f)和图 9(g)的实验结果可见，本模型算法可明显加强被注意对象的显著性。



(a) S1与S2对话S3频率强度显著图　(b) S1与S2对话S3频率变化显著图

(c) S1与S2对话S3强度变化显著图　(d) S1与S2对话S3音调变化显著图

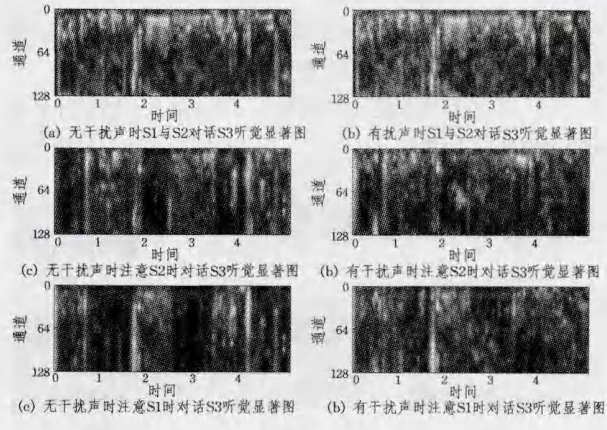
(e) S1与S2对话S3听觉显著图

(f) 注意S2时对话S3听觉显著图

(g) 注意S1时对话S3听觉显著图

图 9　无背景音干扰时自然音频选择性注意听觉显著图

为验证本文计算模型对自然音频混叠的选择性注意的显著性的有效性，图 10 示出在有背景干扰声时的处理结果。其中左列显示了在有干扰声环境下的情况，右列是对应无干扰声时的对比情况。其中图 10(a)和图 10(b)为对话 S3 的听觉显著图、图 10(c)和图 10(d)为注意 S2 时对话 S3 的听觉显著图、图 10(e)和图 10(f)为注意 S1 时对话 S3 的听觉显著图。



(a) 无干扰声时S1与S2对话S3听觉显著图　(b) 有扰声时S1与S2对话S3听觉显著图

(c) 无干扰声时注意S2时对话S3听觉显著图　(b) 有干扰声时注意S2时对话S3听觉显著图

(e) 无干扰声时注意S1时对话S3听觉显著图　(b) 有干扰声时注意S1时对话S3听觉显著图

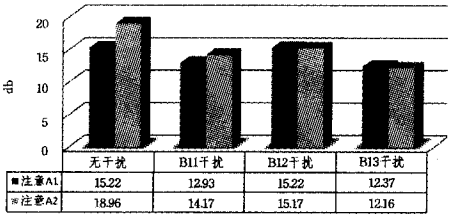
图 10　带背景干扰时自然音频选择性注意听觉显著图

为定量检测选择性注意影响，模型采用无注意对话显著

图与注意对话显著图的输出信噪比 SNR 进行定量评价：

$$SNR(f,t)=10\log\frac{\sum_{t=1}^{T}\sum_{f=1}^{F}tf_0^2(f,t)}{\sum_{t=1}^{T}\sum_{f=1}^{F}[tf_0(f,t)-tf_i(f,t)]^2} \quad (12)$$

式中，$tf_0(f,t)$为理想的对象 i 的显著图，$tf_i(f,t)$为注意第 i 个对象的显著图。具体实验结果如图 11 所示。



	无干扰	B11干扰	B12干扰	B13干扰
■注意A1	15.22	12.93	15.22	12.37
※注意A2	18.96	14.17	15.17	12.16

图 11　音频选择性注意听觉显著图输出信噪比

从图 11 中可见，背景女声干扰 B11 对被注意男声 S2 显著性影响较小，而对被注意女声 S1 的显著性影响较大，分别降低约 4.8dB 和 2.3dB；背景纯音乐干扰 B12 对被注意男声 S2 和女声 S1 的显著性影响不大；背景男女声合唱音乐干扰 B13 对注意男声 S2 和女声 S1 的显著性影响较大，分别降低约 36%和 19%。整体背景干扰对显著性平均降低约 3.4dB，频谱相近的、音强较大的影响较大。实验结果与听觉的掩蔽特性非常近似。

结束语　本文提出的基于选择性注意的认知神经机制的听觉显著性计算模型，在结构和功能上模拟了听觉认知神经信息处理机制。该模型兼容了 BU 的数据驱动和 TD 的概念反馈两种听觉注意机制，很好地模拟了人类的听觉注意过程。在自然环境下，本模型能有效增强被注意音频的显著性，降低和抑制非注意背景混叠音的干扰。

尽管本文提出的听觉认知框架考虑了双耳定位因素，但考虑到篇幅的限制，本文所给的显著性模型的计算过程仅考虑了强度和频率在选择性注意中的贡献。基于选择性注意的认知神经机制的听觉显著性计算仍有大量尚未解决的问题，有很多的研究工作要做。在今后的研究中，将进一步考虑双耳时间差、强度差和耳廓等方位因素在听觉选择性注意的影响。

参 考 文 献

[1] Kalinli O. Tone and pitch accent classification using auditory attention cues[C]// Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. Prague Congress Centre Prague, Czech Republic, 2011: 5208-5211

[2] Wang Lei, Pen Yuan, et al. The application of computational auditory peripheral model in underwater target classification[J]. Chinese journal of electronics, 2012(01): 199-203

[3] Yin Hui, Xie Xiang, Kuang Jing-ming. Acoustic features based on auditory model and adaptive fractional Fourier transform for speech recognition[J]. ACTA ACUSTICA (Chinese version), 2012, 1: 97-103

[4] Vaclav B, Rainer M, et al. A model-based auditory scene analysis approach and its application to speech source localization[C]// Acoustics, Speech and Signal Processing (ICASSP). Prague Congress Centre Prague, Czech Republic, 2011: 2624-2627

[5] De C B, Botteldooren D. A model of saliency-based auditory attention to environmental sound[C]// Proc. ICA. Sydney, Australia, 2010: 1-8

[6] Snyder J S, Pasinski A C, Devin M J. Listening strategy for auditory rhythms modulates neural correlates of expectancy and cognitive processing[J]. Psychophysiology, 2010, 48(2): 198-207

[7] Tabor K M, Coleman W L, et al. Tonotopic organization of the superior olivary nucleus in the chicken auditory brainstem[J]. Journal of comparative neurology, 2012, 520(7): 1493-1508

[8] Mutoh Y, Kashimori Y. Neural model of auditory cortex for binding sound intensity and frequency information in batys echolocation[C]// ICONIP'11 Proceedings of the 18th International Conference on Neural Information Processing-Volume Part I. Hangzhou, China, 2012: 62-69

[9] Rauschecker J P. An expanded role for the dorsal auditory pathway in sensorimotor control and integration[J]. Hearing research, 2011, 271(1/2): 16-25

[10] Chatterjee S, Kleijn W B. Auditory Model-Based Design and Optimization of Feature Vectors for Automatic Speech Recognition [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(6): 1813-1825

[11] Zhao Ya-hui, Wang Hong-li, Cui Rong-yi. An Approach to Sound Feature Extraction Method Based on Gammatone Filter [J]. Advances in Multimedia, Software Engineering and Computing, 2012, 2: 371-376

[12] Zhu Jun-mei. A Multifactor Winner-Take-All Dynamics [J]. Neural computation, 2011, 23(7): 1835-1861

(上接第 259 页)

[5] Cai Z, Vagena Z, Jermaine C, et al. Very Large Scale Bayesian Inference Using MCDB [C]// Big Learn Workshop, Neural Information Processing Systems. 2011

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[7] Porteous I, Newman D, Ihler A T, et al. Fast collapsed Gibbs sampling for Latent Dirichlet Allocation [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2008: 569-577

[8] Liu Zhi-yuan, Zhang Yu-zhou, Chang E Y, et al. Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing [J]. ACM Transactions on Intelligent Systems and Technology, special issue on Large Scale Machine Learning, 2011, 2(3): 26

[9] Smola A J, Narayanamurthy S. An Architecture for Parallel Topic models [J]. The Proceedings of the VLDB Endowment, 2010, 3(1): 703-710

[10] Newman D, Asuncion A, Smyth P. et al, Distributed Inference for Latent Dirichlet Allocation [C]// Neural Information Processing Systems. 2007

[11] 张步良. 基于分类概率加权的朴素贝叶斯分类方法[J]. 重庆理工大学学报：自然科学版, 2012, 26(7): 81-83