计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(7):252−257 [doi: 10.15888/j.cnki.csa.005814]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于说话人辨识的自上而下听觉显著性注意模型①

叶于林[1], 杨　波[2], 莫建华[1], 刘　夏[1]

[1](中国人民解放军78438部队, 四川 成都 610066)
[2](中国人民解放军68108部队, 甘肃 兰州 730030)

摘　要: 为体现听觉注意神经信息处理计算机制对听觉场景内容的自动分析与理解功能, 本文基于人耳对频率变换的感知特性, 结合深度信念网络的说话人辨识与听觉显著模型, 提出了一种自上而下的听觉显著性注意提取模型. 仿真结果表明: 该模型具有可行性, 同时在利用深度信念网络的说话人辨识技术中能够有效地凸显目标说话人的显著度.

关键词: 听觉显著性注意; 显著性注意提取模型; 深度信念网络; 说话人辨识

引用格式: 叶于林,杨波,莫建华,刘夏.基于说话人辨识的自上而下听觉显著性注意模型.计算机系统应用,2017,26(7):252–257. http://www.c-s-a.org.cn/1003-3254/5814.html

Top-Down Auditory Saliency Attention Model Based on Speaker Identification

YE Yu-Lin[1], YANG Bo[2], MO Jian-Hua[1], LIU Xia[1]

[1](78438 Troops of the Chinese People's Liberation Army, Chengdu 610066, China)
[2](68108 Troops of the Chinese People's Liberation Army, Lanzhou 730030, China)

Abstract: In order to reflect the automatic analysis and understanding of the auditory scene content by the auditory attention neural information processing computational mechanism , this paper presents a top-down extraction model of the auditory saliency attention, based on the perceptual characteristics of human ear to frequency transformation, and combined with the speaker identification using the depth belief network and the auditory significant model. The simulation results show that the proposed model is feasible, and it can effectively highlight the significant degree of the target speaker in the speaker identification technology using the depth belief network.

Key words: auditory saliency attention; extraction model of saliency attention; depth belief network; speaker identification

耳朵是人体生理结构不可缺少的一部分, 在复杂的声源环境中, 人类首先通过它获取大量的听觉信息, 然后再经过大脑神经系统分析处理, 最后智能提取出我们所需的信息, 这就是人类听觉系统的选择性注意特性的具体表现. 听觉选择性注意是人类对外界声音信息进行加工处理的一项心理调节机制, 它体现了处理过程中的效率, 即在大量的声音信号中, 选择提取有用信号并抑制大部分的干扰信号以确保有用信号的进一步加工. 通过模拟人类听觉系统这种选择性注意能力, 研究探索具有一定主动性、选择性的听觉选择性注意计算模型算法, 使得计算机语音处理系统也像人类听觉系统一样具有一定的听觉主动性和选择性, 对丰富和发展计算机听觉理论及其在语音处理、人工智能等多个研究领域中都具有重要的意义, 同时对人耳听觉系统的研究也有着深远的影响, 这也是近年来国内外学者研究的热点课题.

目前国内外对于显著性注意的研究主要集中在视觉上, 近年来各大院校都相继有视觉关注度相关的文

① 收稿时间: 2016-10-11; 收到修改稿时间: 2016-11-14

献报道. 对于听觉关注度的研究尚处于起步阶段, 其主要以具有突发性的自下而上显著性声源[1]为研究对象, 即自下而上听觉显著性注意模型研究, 但在研究过程中未深入考虑听觉显著性和视觉显著性的差异. 所以, 本文在人耳听觉系统对语音信息的研究过程中, 将语音信号分别进行频率通道和时间通道处理, 并结合频率上的差异, 首先提出一种自下而上听觉显著性注意计算模型, 同时为了体现听觉注意神经信息处理计算机制对听觉场景内容有自动分析与理解功能, 在自下而上听觉显著性注意计算模型的基础上加入语音流的说话人辨识技术, 得到一种自上而下听觉显著性注意计算模型, 其目的是模拟人类听觉系统在复杂的多声源环境下智能提取感兴趣或重要的声音内容, 即“鸡尾酒会效应”[2]. 仿真结果表明: 结合了说话人辨识技术的自下而上听觉显著性注意计算模型, 能够在语音流中有效降低非目标说话人的听觉显著性, 从而提高目标说话人的听觉显著性.

1 听觉显著性注意模型

1.1 自下而上听觉显著性注意模型

自下而上显著性模型最早出现在图像研究中[3], 以Itti和Kouch提出的计算模型(即Itti模型)[4]最受肯定. Itti模型首先从原始图像中提取出颜色、方向、亮度三种特征图, 并利用中心周边差异算子提取特征的对比度, 再将三种特征显著性注意线性合并作为最终的显著性注意. 听觉显著性注意模型这一概念最早由Kayser[5]等人提出, 模型流程如图1所示, 该模型将声音信号的语谱图作为原始图像输入, 利用Itti模型的原理来提取语音信号的听觉显著性注意. 之后, Kalinli在Kasyer的基础上, 在特征提取时增加了方向和基音特征, 并采用了不同的归一化方法[6]. 随后, Durk Talsma[7]等人开始研究视、听觉关注机制与多种感知融合的交互影响, 试图建立一个融合机制的统一框架.

根据以上的模型流程图, 将语音信号的语谱图完全作为视觉显著性注意模型的输入并提取相应的特征, 这样做并未充分考虑到听觉信号和视觉信号的差异, 视觉显著度突出的是二维区域的显著度, 而声音信号显著度的重点则体现在时间和频率维度的变化上. 为了有利于突出语音信号显著度在时间和频率维度上的变化, 本文将语谱图的各个频带、各帧数据看作一个时间流, 来做相应的处理, 具体处理算法如下.

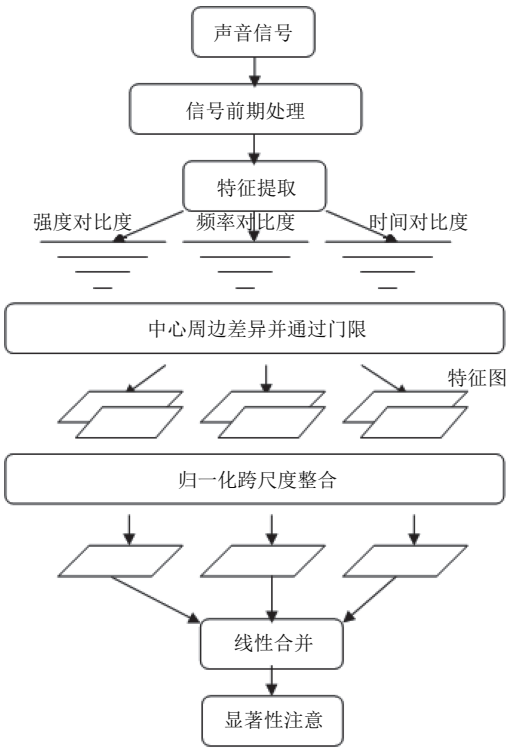

图1 Kayser模型流程图

首先将语音信号进行预处理、分帧、求得语谱图P^{tf}, t表示帧数. 再用24个不同带宽的带通滤波器将P^{tf}在频率和时间上分别划分为24个频率通道$P_i^{tf}(t)(i=1,2,...,24)$和时间通道$T_i^{tf}(t)(i=1,2,...,24)$, 这24个带通滤波器都为三角滤波器, 并且在梅尔频率下是均匀分布的, 梅尔频率与一般频率f的关系为:

$$mel(f)=2595*\log_{10}(1+\frac{f}{700}) \tag{1}$$

对每个频率通道$P_i^{tf}(t)$和时间通道$T_i^{tf}(t)$采用6个不同尺度的高斯差分滤波器滤波, 得到$R_{ij}(j=1,2,...,6)$和$S_{ij}(j=1,2,...,6)$, 其中:

$$R_{ij}(t)=P_i(t)*[\exp(\frac{-t^2}{2\sigma_j^2})-\exp(\frac{-t^2}{2\sigma_{j+1}^2})] \tag{2}$$

$$S_{ij}(t)=T_i(t)*[\exp(\frac{-t^2}{2\sigma_j^2})-\exp(\frac{-t^2}{2\sigma_{j+1}^2})] \tag{3}$$

式中i=1, 2, 3, …, 24, σ_1=2, σ_{k+1}=2×σ_k, k={1, 2, 3, 4, 5}. 将每个频率通道和时间通道不同层次的滤波结果分别线性合并得到按时间变化的听觉显著性注意模型RR和按频率变化的听觉显著性注意模型SS.

其后合并RR和SS, 本文引用图像方面的全局加强

法. 全局加强法的优点是加强突出注意目标贡献大的特征而削弱贡献小的特征. 具体策略是将各特征图的特征值归一化到同一个范围内后, 找出每一幅特征图的全局极大M和除此全局极大之外的其他局部极大的平均值$\bar{m}$, 给每一幅特征图乘以加强因子$(M-\bar{m})^2$, 这就是每幅特征图的权. 这里的"全局"体现在将每幅特征图的全局极大与其他活跃区的平均水平作比较, 差别越大, 权值就越大, 这种显著性就更加被放大; 差别越小, 权值就越小, 该特征图就越容易被忽略.

采用全局加强法合并RR和SS得到最终的自下而上听觉显著性注意模型REI为:

$$REI=(M_R-\bar{m}_R)^2RR+(M_s-\bar{m}_s)^2SS \tag{4}$$

1.2 自上而下听觉显著性注意模型

本文的自上而下听觉显著性注意模型原理框图如图2所示.

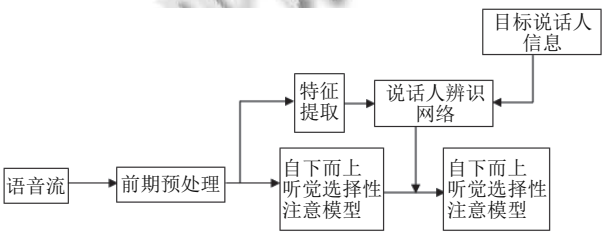

图2 自上而下听觉显著性注意模型

该模型基于时间-频率层面, 对语音流前期预处理之后, 首先采用本文提出的听觉显著性注意模型提取算法提取语音流的显著性注意模型, 即得到自下而上听觉显著性注意模型, 然后将语音流前期处理之后提取的特征参数与目标说话人的信息一起输入到说话人辨识网络中进行说话人辨识, 通过识别结果即可知道哪些时间段是目标说话人的发音, 将辨识结果与语音流的自下而上显著性注意模型线性合并, 即可得到自上而下的显著性注意模型. 本文的语音特征提取采用普遍认为能够体现人耳听觉特性的梅尔倒谱系数(mfcc)作为特征参数, 同时鉴于人体大脑结构非常复杂以及人体耳朵特殊的生理结构, 在说话人辨识部分采用基于叠层自动编码机作为基础模块的深度信念网络.

对上面的模型原理框图进行分析, 该改进的模型对先验信息有一定的依赖性, 本文先验信息主要体现在声源数量确定、噪声为白噪声、说话人语音信号没有重叠等, 其优点是可以显著提高识别性能, 缺点是识别结果明显偏向于模型中出现过的语音信号. 而现实声源环境中是非常复杂的, 如声源数量不确定、声源信号方位信息也可能在实时发生变化、嘈杂的背景噪声等多种可能性, 且这些先验信息都是无法确定的, 因此, 在实际应用中应考虑大规模声源信号的分离与识别, 同时多方面考虑影响语音信号特性的因素, 还要考虑识别过程中噪声消除、语音增强、如何处理回音等多个方面的问题, 使得该模型在现实生活中得以应用.

2 基于叠层自动编码机的深度信念网络说话人辨识

说话人辨识是说话人识别的一种, 即对目标说话人的识别过程, 识别技术目前主要有基于高斯混合模型(GMM)的说话人辨识系统[8]、利用因子分析的说话人辨识系统[9]、基于神经网络的说话人辨识系统[10]等. GMM、因子分析等方法并不能有效模拟人脑的识别过程, 对于神经网络的辨识系统, 虽然能够有效模拟人脑的神经元, 但是由于人体大脑的结构非常复杂, 所以普通神经网络一直得不到广大研究者的满足, 故深度多层的神经网络研究也就开始出现[11]. 随着研究深入, Hinton等首先提出的深度信念网络(Deep Belief Networks)[12], 采用多层结构的DBN(由波尔兹曼模型(RBM)作为每层训练的基础模块)使得深度网络在学习效率上有了突破性的进展. 之后不久, Bengio等发表文章把DBN的成功归纳为采用了逐层无监督的预训练步骤(Layer-wise Unsupervised Pre-traning)[13]. 同时另外一种名叫自动编码机(Autoencoders)的基础模块也被提出, 同样取得了很好的学习效果. 通过理论比较分析DBN和自动编码机的复杂程度和可实现性, 本文采用自动编码机作为多层DBN的基础模块.

自动编码机类似于一个含有单一隐含层的神经网络, 共有3层, 其中隐藏层为数据的特征表达, 通过最小化输入层与输出层之间的误差来校准网络权值, 基本的结构原理如图3所示.

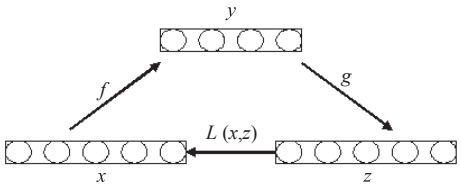

图3 自动编码机结构原理图

一般自动编码机的算法: 自动编码机的输入向量为x, 该向量通过映射函数f映射到隐藏层, 表达式为y, 即$y=f(x)=s(Wx+b)$, 其中W为权值矩阵, b为偏置向量, $s(x)=\frac{1}{1+e^{-x}}$. 之后, 中间层y通过映射函数g到输出层, 表达式为z, 即$z=s(W'y+b')$, 其中W'可以通过限定使得$W'=W^T$, b'为偏置向量. 最后通过交叉熵函数来度量x与z的距离:

$$L(x,z)=-\sum_{k=1}^{d}(x_k\log z_k+(1+x_k)\log(1-z_k)) \quad (5)$$

并通过反向传播算法来更新网络参数.

由于自动编码机只有一个隐藏层, 应用到多层的神经网络的时候, 显然是不合适的, 因此本文采用了叠层自动编码机[14], 其结构原理图如图4所示. 其基本思想就是每一层都用到自动编码机的思想, 使其输入经过网络后得到的输出尽可能的逼近输入. 与单层的自动编码机相比, 一是叠层自动编码机在自下而上的逐层训练过程中, 下层的特征可以作为上层的输入继续参加训练; 二是叠层自动编码机能进行多层次的特征提取, 提高了网络的整体表达能力. 总之, 通过叠层自动编码机对网络权值进行预训练, 能够把网络的权值限制在对后续训练有利的区域, 其后更有利于对网络权值进行进一步的整体优化调整.

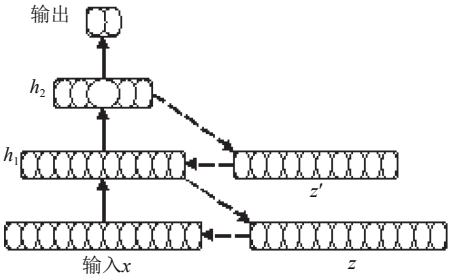

图4 叠层自动编码机结构

以图4的叠层自动编码机的结构原理图为例, x为输入数据, 输入端有12个神经单元, 由于本文实验部分的语音流是两个说话人的交替发音, 所以输出端使用了2个神经单元, h_1和h_2分别为第一个和第二个隐藏层, 神经元个数分别为10和5. 本文在自上而下的逐层预训练中, 隐藏层h_1对输入数据通过自动编码机训练, 得到的输出结果作为隐藏层h_2的输入继续训练, 完成预训练后, 网络权值对所有实例抽取隐藏层的特征, 把这些特征作为上层自动编码机的输入继续训练, 这样逐层迭代, 就构成一个深度信念网络. 在具体的训练过程中, 层与层之间的权值更新都是局部的, 也就是说隐藏层h_1和h_2之间权值的更新相互并不产生任何影响, 这样通过一层一层的训练, 使得每层的权值有一个初始值, 之后再根据具体需要采用方向传播算法对权值进行整体调优, 即可实现具体的功能, 这样做好处就是有效的防范了只采用方向传播算法所造成的局部最优问题.

3 实验仿真

为了说明本文提出的自下而上听觉显著性注意模型方法具有可行性. 实验一: 将频率分别为1500 Hz和2000 Hz的正弦信号, 频率从2500 Hz到7500 Hz以速率为800 hz/s变化的线性调频信号和频率从7500 Hz到2500 Hz以速率为-800 hz/s变化的线性调频信号组合成一个测试信号, 采用本文提出的听觉显著性注意模型算法得到的信号显著性注意和语谱图如图5所示.

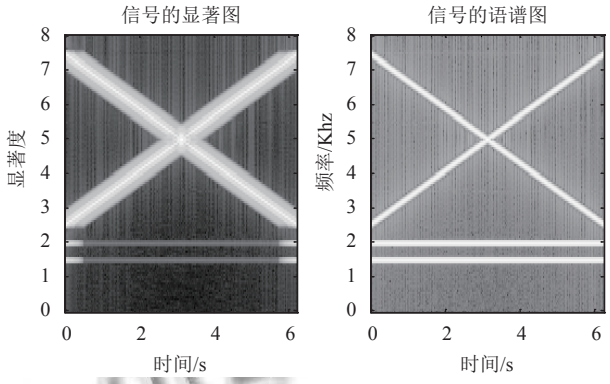

图5 测试信号语谱图和听觉显著性注意

从图5中可以看出, 调频信号因为频率在变化, 它的显著性注意大致符合其语谱图的走势; 而正弦信号由于频率恒定, 其显著性注意在开始和结束的时候比较明显, 而在中间部分比较弱, 这符合人耳的听觉特性, 故说明了本文提出的听觉显著性注意提取方法具有可行性.

实验二: 我们从语料库NIST中选取了两个说话人(一男一女), 每人10句发音, 平均每句4 s长, 各自选取其中3句, 通过女-男-女的发音顺序合成一个对话. 通过本文的听觉显著性注意模型方法, 得到其自下而上听觉显著性注意与频谱图的对比如图6所示.

实验三: 用每人另外的7句发音作为训练语音, 用来训练深度信念网络, 以男性作为目标说话人, 女性作为干扰说话人. 本文采用包含两个隐含层的网络, 其中

第一个隐藏层的节点数为10, 第二个隐藏层的节点数为5; 输出层采用的两个节点, 如果发音为男性的, 则理想输出为[1, 0], 如果发音为女性的, 则理想输出为[0, 1]. 将待识别的语音流分帧, 帧长256, 提取特征参数送入训练好的网络识别, 得到一个识别结果*REC*, 其中*REC*分布在是0到1之间的两维矩阵, 因为此处男性作为目标说话人, 所以取*REC*的第一维参数, 为其长度为帧数. 考虑到实际的语音环境中, 耳朵不可能完全屏蔽掉非关注的语音流, 所以本文设定一个阀值0.6, 对于*REC*, 小于0.6的结果默认为0.01, 大于0.6的结果默认为1. 考虑到说话人辨识系统识别不可能100%成功, 且语音流中有静音段, 即说话人说话中途的停顿, 以及说话人在发音时发音不会太短等问题, 将*REC*进行平滑处理. 平滑处理准则为: 一是如果为1的帧相连前后4帧都为0.01, 则这一帧也为0.01; 如果为0.01的帧相连前后4帧都为1, 则这帧为1; 二是如果多个连续1的长度小于20帧(由于目前听觉注意显著性模型结合说话人识别这方面的文献较少, 对这个数值还有待研究, 所以本文通过观察*REC*, 取为20帧), 则将其全部置0.01. 此时得到的*REC*已经去除了语音流中的静音段, 但是为了更好的刻画某个说话人的发音段, 再将说话人发音中途的静音段平滑, 最后结果如图7所示.

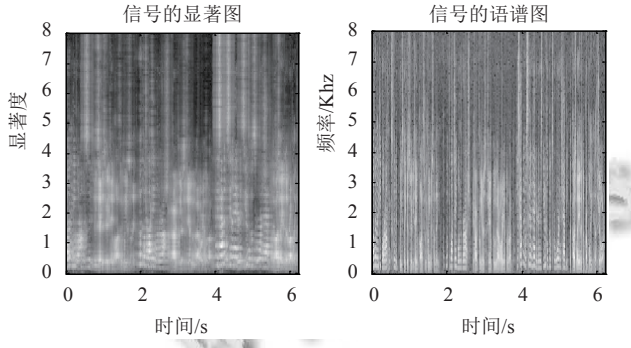

图6 语音流的自下而上听觉显著性注意与语谱图

图7中长黑线包含的部分为理想的目标说话人的发音段. 结合图6中语音流的听觉显著性注意和图7对目标说话人的识别结果可以得到目标说话人自上而下听觉显著性注意与语音流听觉显著性注意对比如图8所示.

图8说明通过深度信念网络的说话人辨识技术, 可以有效屏蔽语音流的听觉显著性注意中非目标说话人的部分, 凸显目标说话人.

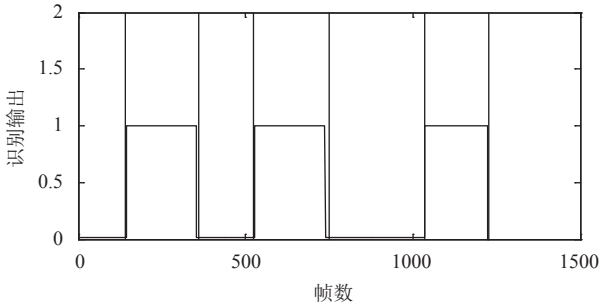

图7 语音流的识别结果

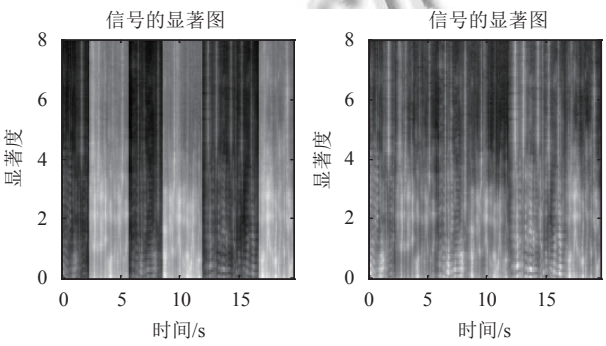

图8 语音流的自上而下听觉显著性注意

4 结语

本文首先根据听觉与视觉显著性的差异性, 提出了一种基于时间变化的自下而上听觉显著性注意模型, 该模型模拟人耳的听觉特性, 对声音按时间分频率通道进行处理, 凸显了声音随时间变化的差异. 其后与说话人辨识技术相结合设计出了自上而下听觉显著性注意模型, 该模型可以有效的屏蔽显著性注意中非目标说话人部分. 仿真实验表明: 本文提出的自下而上听觉显著性模型, 能够很好的模拟人耳的听觉特性, 在频率恒定时, 关注度低; 而在频率变化时, 关注度会随频率变化. 通过结合基于自动编码机的深度信念网络, 能够有效凸显目标说话人的显著度, 进一步体现听觉注意神经信息处理计算机制对听觉场景内容的自动分析与理解功能. 在以后的研究中, 我们希望在不知道整个语音流的时候, 可以实时的根据语音的进展提取出显著性注意并辨识出目标说话人, 屏蔽其他非目标说话人, 进而实时凸显目标说话人的显著性注意.

参考文献

1 Navalpakkam V, Itti L. Modeling the influence of task on attention. Vision Research, 2005, 45(2): 205–231. [doi: 10.1016/j.visres.2004.07.042]

2 Ainhoren Y, Engelberg S, Friedman S. The cocktail party problem [instrumentation notes]. IEEE Instrumentation & Measurement Magazine, 2008, 11(3): 44–48.
3 王彤, 滕奇志, 唐棠. EBCOT图像压缩算法中若干问题的研究. 四川大学学报(自然科学版), 2009, 46(2): 395–400.
4 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. [doi: 10.1109/34.730558]
5 Kayser C, Petkov CI, Lippert M, *et al*. Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology, 2005, 15(21): 1943–1947. [doi: 10.1016/j.cub.2005.09.040]
6 Kalinli O, Narayanan S. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium. 2007.
7 Talsma D, Senkowski D, Soto-Faraco S, *et al*. The multifaceted interplay between attention and multisensory integration. Trends in Cognitive Sciences, 2010, 14(9): 400–410. [doi: 10.1016/j.tics.2010.06.008]
8 Kenny P, Boulianne G, Ouellet P, *et al*. Speaker and session variability in GMM-based speaker verification. IEEE Trans. on Audio, Speech and Language Processing, 2007, 15(4): 1448–1460. [doi: 10.1109/TASL.2007.894527]
9 李轶杰, 郭武, 戴礼荣. 话者识别的信道补偿. 小型微型计算机系统, 2008, 29(12): 2344–2347.
10 Bengio Y. Learning deep architectures for AI. Foundations and Trends® in Machine Learning, 2009, 2(1): 1–127. [doi: 10.1561/2200000006]
11 Bengio Y, LeCun Y. Scaling learning algorithms towards AI. Bottou L, Chapelle O, DeCoste D, *et al*. Large-Scale Kernel Machines. Cambridge, London. 2007. 321–359.
12 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]
13 Le Roux N, Bengio Y. Representational power of restricted boltzmann machines and deep belief networks. Neural Computation, 2008, 20(6): 1631–1649. [doi: 10.1162/neco.2008.04-07-510]
14 Bengio Y, Lamblin P, Popovici D, *et al*. Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems, 2007, (19): 153–160.