语音分离技术的研究现状与展望

李从清[1,2], 孙立新[1], 龙 东[1], 任晓光[1]
(1. 河北工业大学机械工程学院, 天津 300130; 2. 天津城市建设学院能源与机械工程系, 天津 300384)

摘要: 从计算听觉场景分析和盲源分离两种方法综述了当前语音分离技术的研究现状和发展方向。计算听觉场景分析是用计算机来模拟人类听觉系统的处理机制。它可分为两大类:一类是数据驱动型,特点是信息由低级向高级的单向流动;另一类是图式驱动型,特点是信息由低级向高级和由高级向低级结合的双向流动。最后指出信息双向互流的混合语音分离模式将是未来计算听觉场景分析研究的主要模式; 基于听觉和视觉的结合来改善语音分离效果的研究将是未来研究方向之一。此外,盲源分离的欠完全问题,非线性混叠信号的可分离性、非平稳混叠信号的盲分离问题都将需要进一步研究;基于 CASA 和 BSS 联合进行语音分离将是未来研究的热点。

关键词: 语音分离;听觉场景分析;计算听觉场景分析;盲源分离
中图法分类号:TN912.3,TN911.7 **文献标识码**: A **文章编号**: 1000-3630(2008)-05-0779-09

State and frontiers of research in speech separation

LI Cong-qing[1,2], SUN Li-xin[1], LONG Dong[1], REN Xiao-guang[1]
(1. *School of Mechanical Engineering, Hebei University of Technology , Tianjin* 300130, *China;* 2. *Tianjin Urban Construction Institute, Department of Energy Source and Mechanical Engineering, Tianjin* 300384, *China*)

Abstract: The state and the direction of research in speech separation technique, including computational auditory scene analysis (CASA) and blind source separation (BSS), are reviewed. Computational auditory scene analysis is a processing mechanism that simulates human auditory system by computer. CASA is classified as data-driven CASA and schema-driven CASA. The information direction flowing of data-driven CASA is bottom-up and that of schema-driven CASA is integrating bottom-up and top-down. Finally, it is pointed out that the speech separation model of information bi-direction flowing of CASA is the dominant pattern of CASA in the future, and the speech separation method based on integrating audition and vision in order to improve effect of speech separation is one of the directions of research in the future. Furthermore, it is pointed out that the under-complete, the nonlinear and the non-stationary BSS problems are needed to study further, and the speech separation technique based on integrating CASA and BSS is the hot issue of research in the future.

Key words: speech separation; auditory scene analysis; computational auditory scene analysis; blind source separation

1 引 言

语音分离问题源于著名的“鸡尾酒会效应”,即在复杂的混合声音中,人类能有效地选择并跟踪某一说话人的声音。语音分离的研究在语音通信、声学目标检测、声音信号增强等方面有着重要的理论意义和实用价值。而将语音分离技术应用到智能机器人中,让机器人具有智能的听觉,实现声源定位和分离,确定说话人个数,进行人机对话等方面更具有广阔的应用前景。目前,这方面的研究主要有听觉场景分析[1](Auditory Scene Analysis, ASA)和盲源分离(Blind Source Separation, BSS)两类方法。对于听觉场景分析的研究有两种方法:一种是从人的听觉生理及心理特性出发,研究人在声音识别过程中的规律,即听觉场景分析;另一种是利用对人听觉感知的研究成果建立模型,对模型进行数学分析并

收稿日期:2007-06-26;修回日期:2008-09-10
基金项目:河北省科学技术与发展计划项目(05547003D-2);河北省自然科学基金(F2007000118)
作者简介:李从清(1962-),男,天津人,博士生,副教授,研究方向为机器人听觉系统。
通讯作者:李从清, E-mail: congqing6206@vip.sina.com

用计算机来实现它,这是计算听觉场景分析[2](Computational Auditory Scene Analysis, CASA)所要研究的内容。盲源分离是指在源信号、传输通道特性未知的情况下,仅由观测信号和源信号的一些先验知识(如概率密度)来估计出源信号各个分量的过程。盲源分离的独立分量分析方法首先是由Comon提出的[3],它是基于神经网络和统计学的基础发展起来的一种技术,是一个十分活跃的前沿领域。该技术在通信、生物医学信号处理、语音分离、图像处理、阵列信号处理以及通用信号分析等方面有着广泛的应用。

2 听觉场景分析

2.1 听觉场景分析

听觉场景分析源于Cherry在1953年的发现,经加拿大心理学家Albert. Bregman总结20年研究成果,提出了听觉场景分析的一系列理论,并就人类听觉系统多信息流的检测分离给出了一系列的准则。Bregman认为[1],人对声音的感知过程是对声音分量的组合过程。正是这种组合过程使得人可以将混合声音中同一人所发出的声音分量组合到同一个声音流中。一般有两种组合原则。(1)同时组合(Simultaneous Organization)。这种组合是将同时存在的不同频率范围的声音分量组合在一起。影响组合的因素有:第一是谐波之间的调和度。人的听觉系统倾向于将具有相同基音的谐波组成一组。第二是音调之间的同步性。如果两音调的开始与结束都是同步的,那末人的听觉系统就趋向于把它们认为是一个单独的复音。第三是音调之间的接近程度。它们之间在频率上越接近,听者就越容易把它们听成一个独立的声音流。第四是声音的强度。来自不同声源的声音往往具有不同的声音强度。第五是声音的方位。人的听觉系统倾向于将来自相同空间方位的声音组合在一起。(2)序列组合(Sequential Organization),这种组合是将一串声音分量按时间先后组合到一个或多个声音流中。每一个声音流中的声音分量都被认为是来自同一个声源。影响人对声音流的时序分割与组合有两个主要因素:第一是频率差。在声音流的循环速度一定时,高、低音调间的频率差越大,声音流分离的现象越明显。第二是高低音间变化的速度。在高、低音调间的频率差一定时,高低音间变化的速度越快,声音流分离的现象越明显。同时,还可采用完形心理学的邻近组合原则,即指人主观上会根据一些独立的感知因素(声音的音调、响度,音色间的差异)的共同变化来将声音分量按照一定原则组合成一个声音流。Bregman认为,人对声音的序列组合是一种后天性的,而对声音的同时组合则是先天的能力。

在听觉心理特性的基础上,Bregman总结了听觉场景分析中的一些准则:独占分配准则和闭包或连续准则。独占分配准则指人的听觉系统认为任何一个感官要素在某个时刻只能属于某一个声音流,而不能属于两个或多个声音流。闭包或连续准则是指人脑能够将进入人的感觉器官的非连续的信息合理地组合起来,得到对事物或现象的整体认识。正是这些定性准则为今后的计算听觉场景分析提供了有效的理论基础,出现了许多有效的计算听觉场景分析方法。

2.2 计算听觉场景分析

计算听觉场景分析(CASA)是利用计算机技术将人类听觉对声音的处理过程(听觉场景分析)建模,使计算机具备从混合声音中分离各物理声源并做出合理解释的能力。综观语音分离的各种方法,可将计算听觉场景分析概括为三大阶段:

第一个阶段是将达到人耳的混合声音信号分解为一组感官元素(sensory elements),称为分解(Segmentation)。初始信号在时域和频域上的观测量,往往是多个源声音信号的混合。CASA一般模仿人耳采用耳蜗模型,用近似耳蜗频响的听觉滤波器组,将混合信号变换到一个可以区分混合信号中各个分量的变换域中。

第二个阶段是将这些感官元素按照声源不同进行分组,形成可以对某个声源信号进行感知的"听觉流"(auditory stream),这一阶段称为组合(Grouping)。这一阶段是计算听觉场景分析的难点。Bregman认为人具有利用这些分类感官元素得到目标场景的能力。难点之一是如何找出定量的声音分离线索,现有的分离线索有:基音、声源的起始/结束时刻、双耳时差、双耳声压差、触发、频率过渡等。难点之二是如何利用线索组合出属于某个声源的信号,现有的计算听觉场景分析利用了人工智能的成果,如专家系统中的黑板模型,以及神经网络等。

第三个阶段就是重新合成。即将来自同一个声源的感官元素重新合成,重建该源声音信号。这一阶段称为合成(Integration)。

计算听觉场景分析系统可分为两大类。一类称为数据驱动型(Data-driven)CASA,特点是信息由

低级向高级单向流动，这种方法符合传统的信息处理过程。另一类是图式驱动型(Schema-driven) CASA，特点是信息由低级向高级和由高级到低级结合的双向流动，它更接近于人类对听觉感知的机制。

2.2.1　数据驱动型 CASA

数据驱动型的 CASA 研究是从 20 世纪 80 年代中期开始的。最早是 1983 年 Scheffers 的双元音分离模型[4]，接着 1985 年 Weintraub 分离和重构两个同时说话人声[5]。系统将信号经过耳蜗滤波器组后，再利用类似自相关函数的“共生函数(coincidence function)”识别每个通道的明显的周期性。将各个通道的共生函数迭加检测音调，用检测得到的音调“神经事件”组合为“群目标(group object)”来分离两个讲者的语音。以后相继出现许多数据驱动型 CASA 系统。1991 年 Mellinger 的基于探测 onset 和频率的变化，增加相似形，确定频率调制，产生分离的谱图的方法[6]。1992 年 Meddis 和 Hewitt 的混叠语音基频提取算法 [7]，M.P.Cooke 提出对听觉模型表达进行组织的系统[8]。1994 年 G.J.Brown 和 M.P.Cooke 试图建立一个与生理学和心理学发现相吻合用于将语音从各种干扰声源(包括语音)中分离出来的系统[2]。1995 年 B.Markus 提出一种基于人的双耳听觉特性的双耳模型 CASA 方法[9]，并把它用于噪声环境中语音识别的前端，实验结果表明采用了前端处理后，语音识别的效果大为改善。1996 年 D.L.Wang 和 G.j.Brown 用自相关谱检测各频道的调幅，用振荡器标记并抑制各周期，再用神经振荡器网络进行元音分离[10]，1999 年他们对模型进行了改进，取得了很好的效果[11]。1999 年 N. Roman，D. L. Wang，G. J. Brown 直接利用空间位置线索(如 ITD，IID)，结合鸡尾酒会效应进行语音分离[12]，并在 2002 年对系统进行了改进[13]。他们提出一种“理想”的时频二元的掩蔽概念，该系统性能比以前方法有显著改善。近几年，相继出现了利用声音线索间的结合及听觉和视觉的结合来改善声源分离的方法。1999 年 Nakatani 等开发了从双耳输入信号分离语音流的 BiHBSS 系统[14]。系统首先通过使用谐波结构作为线索来提取谐波片段，然后根据声源方向和基频的连续性组合它们。为了改善识别性能，BiHBSS 系统被用到识别两个同时语音上[15]。同年，他们利用视觉和听觉信息的联合来克服声源方向的模糊性[16]。2001 年 H.G.Okuno 等提出了通过联合听觉和视觉处理的方向通道滤波(direction-pass-filter)方法来分离三个同时语音[17]。此后，他们又对系统进行了多次改善[18-20]。2004 年 G. Hu and D. Wang提出基于基音跟踪和调幅的单耳听语音分离系统[21]。2006 年 G. Hu and D. Wang 提出通过分析听觉事件的起始和结束时刻来进行听觉分离的系统[22]。2007 年 S. N. Wrigley and G. J. Brown 结合使用基音和空间位置线索来改善声源分离[23]。

综观各种语音分离方法，数据驱动型 CASA 系统由四个部分组成。第一部分听觉外围系统模型。它模仿人类外围听觉结构(包括外耳、耳膜及耳蜗)对声音进行处理，并提取特殊的听觉特征(即线索)。第二部分听觉图的表述。利用前一阶段所得到的声音处理结果，再加上对另外的声音归类线索进行处理，形成了 4 个听觉计算图：自相关图、互相关图、频率转移图、声源起始/结束图。第三部分场景分析(声源分离)。结合第二部分得到的各种信息图便可以进行听觉场景分析。Brown & Cooke 将场景分析分三步进行：(1)定声音元素，(2)声音元素的归类，(3)听觉场景搜索。第四部分声源的再合成及定量评估：可以通过对各类声源的再合成技术来定性地评估 CASA 的性能，也可以方便地进行一些定量的评估，如信噪比以及各种 CASA 的结果比较。此类系统，在文献[24]中指出了局限性：线索不充分，算法不能扩展，固定的证据集成，不能处理模糊数据。

2.2.2　图式驱动型 CASA

数据驱动型 CASA 系统中信息的流向是由低级传向高级的单向流动过程，信息由低级系统收集后逐级处理，低一级的处理结果作为高一级的输入，最后得到各物理声源的分解结果。这种系统被 Slaney称为“纯听觉”系统[25]，它是从 Marr 的视觉计算理论(“纯视觉”理论)借鉴过来的[26]。Charchland 等人对此提出了质疑，并证明视觉过程是信息双向流动的过程，称之为“交互视觉”[27]。

目前，在许多语音识别系统中，研究人员已经利用信息双向流系统来进行语音识别和语音分离。Varga和 Moore 提出用两个 HMM 模型来进行语音和噪音的分离[28]，Brown and Cook 等用 Kohonen 神经元及 HMM 对信息不全的语音进行识别[29]，还有利用黑板系统建立 CASA 系统，允许用期望来控制感知过程，这类双向流动信息的系统被 Bregman 称为“图式驱动型”系统。目前，许多图式驱动型的 CASA 系统都是基于黑板结构。黑板结构的构想起源于 70 年代美国 APPA 发起的语音理解系统研究计划，灵感来源于对诸如聆听并理解语音这类感知活动的观察。信息存在于各个层次。比较成功的图式

驱动型 CASA 系统有：

1994年 Nakadai 等人提出了被称为“误差驱动型”的系统[30]。该系统利用双耳模型抽取声音的谐波结构和声源空间位置信息实行声源分离。他们设计了一个多代理系统(Multi-Agent System)，用残余波形触发新的代理，用一个代理对一种声音进行谐波跟踪，并将结果反馈。采用谐波跟踪进行 CASA 的方法。1998 年他们对多代理系统进行了改进，将每一代理动态分配给每一声音流，并用各个代理间的相互作用来解决不同声音流之间的干涉[31]。

1995年 Kashino 等人进行了找出音乐中乐符和和弦的研究[32]。将乐器模板与谐波跟踪匹配，用 Bayesian 网络组合分散的证据，提出音乐的“状态”影响可能的乐符，得到了分级的音乐描述。

1996年 Dan Ellis 提出了基于预测驱动的计算听觉场景分析结构[33]。该系统使用了比较成功的专家系统模型：“黑板”模型。

1997年 Unoki Masashi 以小波滤波后的信号的振幅包络，及输入、输出的相位偏移为基础提出了一种混合声音分离模型[34]，该模型对混合后的声音能很好地分离。

1998年 Abe Mototsugu 等人用波型算子来提取子带信号的瞬时频率、频率幅度的变化[35]；并用在概率分配中的投票系统来提取各个子带信号中相同的瞬时频率、频率幅度的变化。

1998年 Klassner 和 Lesser 等提出 IPUS (Integrated Processing and Understanding of Signals)黑板结构[36]。系统以信号处理与理解集成模块为框架，构造前端信号处理算法与后续的信号理解过程之间信息双向交互作用的动态结构，可以动态地改变前端的信号处理算法以适应各种变化。该系统将注意力集中在观测到的信号特征与目前所得到的解释之间是否存在差异，如果有差异，则采用两种方法来解决：一是修改解释；二是修改用以产生信号特征的前端信号处理算法中的有关参数，使差异问题得到满意解决。

1999年 D. J. Godsmark 和 G. J. Brown 进一步发展了一种“黑板”模型的计算听觉场景分析模型，该模型将从多种成组原则(Grouping Principe)中所提取的特征组合在一起，并对各个成组原则之间的竞争采用了一种与人的心理生理特性相符的管理方法[37]，同时，该黑板结构可以用启发式知识来影响其对听觉场景的组织。相对以前的 CASA 模型，该黑板结构可以更好地模拟人对听觉场景的分析。

2004年 Stephen Grossberg 等人提出了一种称作 ARTSTREAM 的神经网络听觉场景分析模型[38]。模型阐明了符合一给定声源的频率分量是怎样被基于基音周期和空间位置线索组成一独特声流的，同时，也阐明了多种声流是怎样由人脑区分和分离的。2005 年 D.L.Wang 等提出一种基于图式驱动的音素修复模型[39]。该模型使用了缺少数据的语音识别系统来解码完整部分的语音，而后激活包含掩蔽音素相应词的模板。该系统能高度自然地修复不完全或缺少的浊音和清音。

我国听觉场景分析的研究尚处于起步阶段。文献[40]是国内比较早提出听觉场景分析这一概念的。文中提出，人耳除了可以感受声音的强度、音调、音色和空间方位外，还可以在两个人以上的讲话环境中分辨出所需的声音，这种分辨能力是人体内部语音理解机制具有的一种感知能力。人类的这种分离语音的能力与人的双耳输入效应有关，称为"鸡尾酒会效应"。听觉场景分析正是利用语音信号的基音、谐波、起止点、幅度以及频率等信息，来追踪或重建某路语音信号。文献[41]提出了一种重叠语音信号分离的方法。该方法通过频域的 SHS(分谐波累加法)算法和小波变换算法两种方法来提取源信号的基音频率轨迹。然后，将基音频率轨迹作为依据，获取相应的各次谐波的频率、幅度以及相位信息，利用这些参数送入 SIN 模型来合成分离的语音信号。文献[42]提出基于谐波增强与压缩的同声道语音分离系统。它利用设计好的听觉模型中所提取的基音频率设计出一个滤波器，该滤波器从混叠语音粗略地分离出各个源信号。分离出的源信号再进一步在它谐波点和共振峰处增强，同时压缩其它不属于它的频率成分以获得目标信号。文献[43]提出采用两层振荡器神经网络的 CASA 改进模型语音分离算法结构，有效地分离出混叠输入语音。文献[44]提出了一种新的基于双基频多带激励分离模型的元音分离系统。该系统利用两语音信号的基音轨迹特性提取多带激励分离模型中两基频对应的语音的各次谐波参数，将两组参数代入多带激励合成模型从而得到两个分离的语音信号。文献[45,46]也都对语音分离研究进行了有益探索。

3 盲源分离技术

盲源分离(Blind Source Separation, BSS)是指在不知道源信号和传输通道的参数的情况下，根

据输入源信号的统计特性，仅由观测信号恢复出源信号各个独立成分的过程。盲源分离的研究是到 20 世纪 80 年代末才真正发展起来。它的另一个名称是独立分量分析（Independent Component Analyaia，ICA），意思是提取混合信号中的独立分量。

1986 年法国学者 Jeanny Herault 和 Christian Jutten 的一篇题为“Space or time adaptive signal processing by neural network models”的开创性的研究报告，揭开了盲源分离研究的新的一章。

1987 年 Giannakis 等人提出了 BSS 的可辨识问题，且引入了三阶累积量[47]。

1988 年、1989 年 Linsker 发表了对盲处理研究具有重要作用的几篇文章[48,49]，提出了最大互信息准则，该准则的基本形式非常适合建立自组织模型和特征映射。

1991 年 C. Jutten 和 J. Herault、P. Comon 和 E. Sorouchyari 发表了关于盲信号分离的三篇经典文章[50-51]，标志着盲分离研究的重大进展。他们不但提出了盲分离中著名的 H-J 算法，而且设计了专用的 CMOS 集成芯片来实现他们的算法。考虑到人类有很好的分离多个音源的能力，所以他们的算法是基于神经网络对耳神经进行模拟。

1992 年 G. Burel 提出了用 BP 神经网络实现线性和非线性混合信号盲分离方法[53]。

1994 年 P. Comon 将数据处理与压缩的主分量分析（PCA）加以扩展而成为 ICA，阐述了独立分量这一概念[54]，界定了解决 BSS 问题的 ICA 方法的基本假设条件。同年，Cichocki 等人提出了著名的自然梯度学习算法[55]。此外，Oja 与 Karhunent 提出采用高阶统计量的非线性主分量分析盲源分离算法[56,57]。

1995 年 A. J. Bell 和 T. J. Sejnowski 发表了 ICA 发展史上的里程碑文献[58]。他们利用神经网络的非线性特性来消除观测信号中的高阶统计关联，并把盲分离问题归入信息论框架，用信息最大化准则建立目标函数，从而将信息论方法与 ICA 结合起来。文中给出了神经网络式的最优权矩阵迭代学习算法，仿真实验成功地对 10 个人谈话的“鸡尾酒会”问题给出了很好的分离效果。同年，Delfosses 和 Loubaton 提出一种紧缩（Deflation）算法[59]。该算法首先从混合信号中分离出一个源信号，从而使问题降低。另外，K.Matsuoka 提出了一种用于非平稳随机信号（如语音信号）的算法[60]，该算法实用性很大。

1996 年 Cardoso 和 Laheld 提出了 ICA 学习算法中的“相对梯度”、“等变化性”和有关稳定性及分离精度等重要思路和方法[61]。同年，Pearlmutte 在 ICA 中引入以最大似然为准则的目标函数[62]，Cardoso 于 1997 年证明了最大信息方法和最大似然估计方法是等价的[63]。

1997 年 Girolami 和 Fyfe 受探测寻踪（EPP）的启发，使用边缘负熵作为投影指数[64]，而且证明了寻找峭度的投影寻踪法可以抽取一个线性混合的源信号。接着，他们实现了对所有信号的提取。同年，Pham 和 Garat 通过准最大似然途径对 ICA 的学习算法、稳定性、分离精度和源信号概率密度函数（pdf）的确定作了进一步讨论[65]。

从 1999 年至 2004 年，分别在法国、芬兰、美国、日本和西班牙成功召开了五届“独立分量分析与盲信号分离”国际学术会议。会议期间，各国科学家发表了大量的论文，内容涉及盲信号处理技术的各个领域。其中，Harri Valpola 等人于 2000 年在 ICA 会议上发表了两篇关于非线性独立分量分析（NICA）的论文[66,67]，提出利用贝叶斯集合学习方法进行 NICA。2001 年，Ki-Seok Cho 和 Soo-Young Lee 利用 CMOS 芯片实现了经典的基于 Infomax 理论的自然梯度算法[68]。2003 年，Cardoso 提出将 ICA 应用于天文图像[69]。2004 年，Zhijian Yuan 和 Erkki Oja 提出了一种用于非负独立分量分析的 FastICA 算法[70]，指出了该算法在非负 ICA 这种特殊情况下的应用。

最近几年的国际声学、语音和信号处理大全（IEEE Int. Conf. Acoust Speech，Signal Processing-ICASSP）上，每次都有关于盲信号处理的专题，而且信号处理界的权威刊物 IEEE Transaction on Signal Processing 以及 Signal Processing 中，盲信号处理的文章也频繁出现。十几年来，对于盲信号分离问题，学者们提出了很多成功的算法。从算法的角度而言，BSS 算法可分为批处理算法和自适应算法；从代价函数或准则的角度而言，又分为基于神经网络的方法、基于高阶统计量的方法、基于互信息量的方法、基于非线性函数的方法等。

国内对于盲信号分离技术的研究相对较晚。清华大学的张贤达在其 1996 年出版的《时间序列分析--高阶统计量方法》一书中，介绍了有关盲分离的基本理论，并给出了相关的算法。直到近几年，国内关于盲分离技术的研究才渐渐多起来。冯大政、史维祥等提出了一种具有抑制噪声作用的有效自适应学习算法，研究了算法的有界性和渐进稳定性[71]。胡光锐、虞晓等研究了盲信号处理在语音分离与识别领域的应用[72]，在分析最大熵法（ME）与最小互信

息量法(MMI)的基础上,提出了一种利用反馈结构的输出信号概率密度函数估计的增强 ME 算法。通过计算机模拟表明,新算法比传统 ME 算法在解决卷积混合输入的盲信号分离问题时,具有更好的分离性能。在无线通讯领域,何振亚、汪军等提出基于高阶谱的信号盲分离[73],利用三阶和四阶累计量研究了瞬时混叠信号盲分离问题,并提出了两种分离算法,模拟实验证实了其有效性。

已有的盲分离算法大致归纳为三类。在第一类方法中,信号经过变换后,使不同信号分量之间的相依性(Dependency)最小化。这类方法称为独立分量分析,由 Pierre Comon 于 1994 年提出。当信号之间的独立性采用不同的准则测度时,即得到不同的算法。第二类方法是用非线性传递函数对输出进行变换,使得输出分布包含在一个有限的超立方体中,然后熵的最大化将迫使输出分量可能在超立方体中均匀散布。这类方法称为熵最大化, 是由 Bell 和 Sejnowski 于 1995 年提出的。熵最大化最终也得到信号分量独立性的类似测度。第三类方法为非线性主分量分析,由 Oja 和 Karhunen 等人于 1994 年提出,它是线性主分量分析方法的推广。目前,前两类方法已逐渐趋于成熟,对于第三类方法的研究也已初步开展起来。

4 语音分离技术展望

目前,听觉场景分析和计算听觉场景分析的研究还是处于起步阶段,特别是计算听觉场景分析研究中所建立的模型只能用于验证听觉场景分析研究中的一些不够明了的理论,即人脑处理听觉信号的机制。如 Slaney 的耳蜗模型,Loughborough 研究组的集成心理学模型等,这些都是听觉外周模型和低层处理,另一些单位则侧重于研究谐波跟踪或双耳效应等。未来基于 CASA 的语音分离研究方向主要有以下几个方面。

(1)在听觉表达上,需要进一步深入研究听觉生理学和听觉心理学, 寻找恰当的中层听觉表达。目前, 在 CASA 中仅仅利用了语音的周期特征,还无法用计算机去模拟人的某一些心理特征。因此,还需要对人的听觉心理学进行更广泛的研究。

(2)在语音分离模型上,信息双向互流的混合分离模式是将来 CASA 研究的主要模式。从以上的讨论可以看到,CASA 的研究已从较早期的数据驱动型(被动型)逐渐转向图式驱动型(主动型),事实上图式驱动型(信息双向流动)CASA 也包含了数据驱动型的结构。愈来愈多的证据显示,人类听觉视觉的感知过程是信息从下往上及自上往下的双向交互的过程。因此,在未来声音/语音流分离过程中,信息从下往上及自上往下双向交互处理将是热点研究方向之一。

(3)研究计算听觉场景分析的快速算法以及实时实现的方法。现有算法的计算量普遍较大,期望找出快速算法,并用通用 DSP 或 ASIC 实现,使听觉场景分析走向实际应用。

(4)视觉听觉联合模型用于语音分离的研究。通过使用两耳麦克或麦克阵列提供的空间信息可改善语音分离性能。可是,这种方法超过一定方向角度时仍然有模糊性。因为视觉信息能提供比较精确和准确的位置信息,分离性能将彻底改善。此外,使用接近的方向信息会彻底改善单一视觉系统目标跟踪的准确性,反过来,会改善视觉系统的性能。因此,视觉听觉联合作用会大大改善各自感觉任务功能, 即声源分离和目标跟踪。所以,用于语音分离的视觉听觉联合模型的研究也将是未来方向之一。

(5)基于 CASA 和 BSS 联合进行语音分离将是未来的研究热点之一。由于 CASA 和 BSS 分离技术各自有不同的条件性和缺陷,联合两种方法将是非常有益的。确切地说,由 CASA 系统(像 F0 的连续性和空间定位)所利用的场景分析的启发将由盲分离算法所利用,以便改善实时声音混合的性能。相反,盲分离技术也能帮助 CASA 分离在时-频域范围的混叠语音。有关 CASA 和 BSS 分离技术的各自的优缺点在文献[74,75]中做了分析研究比较。

当前盲源分离的研究是非常活跃的。盲源分离是从所谓“鸡尾酒会”问题引出的,但是这个问题还没有得到很好的解决。这要涉及到多通道卷积混叠系统和盲反卷积系统的稳定性及相位不确定性问题, 尤其是源的数目未知时盲反卷积问题以及带噪声的情况。因此,我们认为今后还应该对以下问题做进一步研究。

①带噪声混叠信号的盲分离问题。现在研究的大部分盲源分离或盲反卷积算法,大都假设在无噪声情况下,或把噪声看做一个独立源信号。在高阶统计方法中,由于高斯信号高阶累积量为零,因此可以假设加性高斯信号存在。将已有的盲分离方法应用到一般的噪声混叠模型,是有待解决的问题。

②欠完全(under-complete)问题。现在的研究都假设传感器的个数不少于源信号的个数,如何解

决当源信号个数多于传感器个数的情况是又一困难的问题。这方面问题已有文献[76,77]做了研究。

③现在的非线性混叠信号盲源分离方法研究的只是在极特殊情况下的非线性混叠的情况，对更一般的非线性混叠信号的可分离性以及分离的充分必要条件需做进一步的研究。

④非平稳混叠信号的盲分离问题。许多情况下源信号可能是非平稳的，如源信号可能消失一段时间后再出现。如何利用信号的非平稳特性进行盲分离是解决这类问题的一个方向。

参 考 文 献

[1] Bregman A S. Auditory scene analysis: the perceptual organization of sound[M]. Cambridge, MA: The MIT Press, 1990.

[2] Brown G J, Cooke M. Computational auditory scene analysis[J]. Computer Speech and Language, 1994, (8): 297-336.

[3] Common P. Independent component analysis, a new concept[J]. Signal Processing, 1994, (36): 287-314.

[4] Scheffers M. T. M. Sifting vowels: auditory pitch analysis and sound segregation[D]. University of Gronigen, 1983.

[5] Weintraub M. A theory and computational model of auditory monaural sound separation[D]. E. E, Department, Stanford University, 1985.

[6] Mellinger D K. Event formation and separation in musical sound[D]. CCRMA, Stanford, 1991.

[7] Meddis R, Hewitt M. J. Modeling the identification of concurrent vowels with different fundamental frequencies [J]. Acoustic Society of America, 1991, **89**(6): 2866-2882.

[8] Cooke M P. Modeling auditory processing and organization[D]. CS Dept, Univ. of Sheffield, 1991.

[9] Markus B Binaural modeling and auditory scene analysis[A]. IEEE ASSP Workshop[C], 1995, 15-18.

[10] Brown G J, Wang D L. Modeling the perceptual segregation of vowels with a network of neural oscillators [R]. Memo CS-96-06, Sheffield Univ, 1996.

[11] Wang D. L. and Brown G. J. Separation of speech from interfering sounds based on oscillator correlation[J]. IEEE Trans. Neutal Networks, 1999, **10**(3): 684-697.

[12] Roman N, Wang D L and Brown G L. Speech segregation based on sound localization[J]. Proceeding of IJCNN-01, 2001, 2861-2866.

[13] Roman N, Wang D L, Brown G L. Speech segrega-tion based on sound localization[J]. J. Acoustical So-ciety of America, 2003, **114**(4) 2236-2252.

[14] Nakatani T, Okuno H G. Harmonic sound stream segregation using localization and its application to speech stream segregation[J]. Speech Communication, 1999, **27**: 209-222.

[15] Okuno H G, Nakatani T, Kawabata T. Listening to two simultaneous speech[J]. Speech Communication, 1999, **27**: 299-310.

[16] Nakagawa Y, Okuno H G., and H Kitano. Using vision to improve sound source separation [A]. Proceeding of 16th National Conference on Artificial Intelligence(AAAI-99)[C]. 1999, 768-775.

[17] Okuno H G., Nakadai K, Lourens T. Separating three simultaneous speech with two microphones by integrating auditory and visual processing[A]. Proc. Eurospeech[C]. 2001, 2643-2646.

[18] Nakadai K, Hidai K, Mizoguchi H. Real-time auditory and visual multiple-object tracking for robots[A]. Proc. of the 17 th Int. Jiont Conf. on Artificial Intelligence (IJCAI-01), MIT Press[C]. 2001, 1424-1432.

[19] Nakadai K, Kitano H. Exploiting auditory fovea in humanoid-human interaction[a]. Proceeding of 18th National Conference on Artificial Intelligence[C]. AAAI, 2002, 431-438.

[20] Nakadai K, Matsuura D, Okono H G. Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots [J]. Speech Communication, 2004, **44**: 97-112.

[21] Hu Guoning, Wang Deliang. Monaural speech segre-gation based on pitch tacking and amplitude modulation [J]. Transactions on Neural Networkes, 2004, **15**(5): 1135-1150.

[22] HU Guoning, WANG Deliang. Auditory segmentation based on onset and offset analysis [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, **15**(2):396-405.

[23] Wrigley S N, Brown G J. Recurrent timing neural networks for joint f0-localization based speech separa-tion [C]. Proceeding of ICASSP2007, Honolulu, Hawaii, 2007, 157-160.

[24] Ellis D P W. Computer implementation of psychoacoustic grouping rules[A]. Proc 12th Conf on Pattern Recognition[C]. Jerusalem: 1994

[25] Slaney M. A Critique of pure audition[A]. Proceedings of the First Workshop on CASA, Int Jiont Conf AI[C]. Montreal: 1995.

[26] Marr D. Vision[M]. New York: Freeman W H. 1982.

[27] Churchland P, Ramachandran V S, Sejnowski P. A critique of pure vision[M]. Davis Jeds. Large-scale Neuronal Theories of the Brain. Cambridge, MA: MIT Press, 1994.

[28] Varga A P, Moore R K. Hidden markov model decomposition of speech and noise[A]. Proceedings of the ICASSP'90[C]. 1990.

[29] Brown G. J. and Cooke M. Temporal synchronization in a neural oscillator model of primitive auditory stream segregation[A]. D. Rosenthal & H. Okuno. Computational Auditory Scene Analysis[C]. Lawrence Erlbaum, 1998, 87-101.

[30] Nakatani T, Okuno H G, Kawabata T. Auditory stream segregation in separation in auditory scene analysis with a multi-agent system [A]. Proc.Am.Assoc. Artif.Intel. Conf.[C]. Seattle, 1994, 100-107.

[31] Nakatani T, Okuno H G, Goto M and Ito T. Multiagent based binaural sound stream segregation [A]. D. Rosenthal & H. Okuno. Computational Auditory Scene Analysis[C]. Lawrence Erlbaum, 1998, 195-214.

[32] Kashino K, et al. Application of bayesian probability network to music scene analysis[A]. Workshop on Comp. Aud.Scene Anal, Int. Jint Conf. on Artif. Intell [C]. Montreal, 1995.

[33] Ellis D P. Prediction-driven computational auditory scene analysis [D].Ph.D. dissertation, EECS Dep, M. I. T. 1996.

[34] Masashi Unoki, Masato Akagi. Method for signal extraction from noise-added signals[A]. Electronic & Comunications in Japan[C]. 1997,11: 1-11.

[35] Mototsugu, Abe, Shigeru Ando. Auditory scene analysis based on time-frequency integration of shared FM and AM[A]. Icassp98[C], 1998, 4: 2421-2424.

[36] Klassner F, V Lesser, and Nawab S H. The IPUS blackboard architecture as a framewoek for computational auditory scene analysi[A]. Rosenthal D & Okuno H. Computational Auditory Scene Analysis[C], Lawrence Erlbaum, 1998, 105-114.

[37] Godsmark D, Brown J. A black board architecture for computational auditory scene analysis[J]. Speech Communication, 1999, **27**: 353-366.

[38] Stephen Grossberg, et al. ARTSTREAM: A neural network model of auditory scene analysis and source segregation[J]. Neural Networks, 2004, **17**: 511-536.

[39] Soundararajan Srinivasan, WANG Deliang. A schema-based model for phonemic restoration[J]. Speech Communication, 2005, **45**: 63-87.

[40] **任永川. 混合语音信号的盲源分离和声场景分析**[D]. **南京: 东南大学**, 2000.
REN Yongchuan. Blind source separation and sound scene analysis of mixed speech signals[D]. Nanjing: Southeast University, 2000.

[41] **朱美虹. 基于小波变换的重叠语音信号分离**[D]. **苏州: 苏州大学**, 2001.
ZHU Hongmei. Separation of overlapping speech signals based on wavelet tansform[D]. Suzhou: Shuzhou University, 2001.

[42] **陈雪勤. 基于计算声场景分析的混叠语音分离研究**[D]. **苏州: 苏州大学**, 2002.
CHEN Xueqin. Separation of overlapping speech signals based on computational auditory scene analysis[D]. Suzhou: Shuzhou University, 2002.

[43] **胡光锐, 虞晓. 一种采用振荡器神经网络的** CASA **计算模型语音分离算法**[J]. **上海交通大学学报**. 2001, **135**(11): 1640-1644.
HU Guangrui., YU Xiao. Computational auditory scene analysis separation algorithm using oscillator neural network[J]. Journal of Shanghai Jiaotong University. 2001, **135**(11): 1640-1644.

[44] **黄秀轩, 韦岗. 基于双基频多带激励分离模型的元音分离**[J]. **计算机工程与应用**, 2003, (10): 73-76.
HUANG Xiuxuan, WEI Gang. Concurrent vowels separation based on double-pitch multi-band excitation separation model[J]. Computer Engineering and Application, 2003, (10): 73-76.

[45] **赵鹤鸣, 葛良, 陈雪勤, 俞一彪, 等. 基于声音定位和听觉掩蔽效应的语音分离研究**[J]. **电子学报**, 2005, **33**(1): 158-160.
ZHAO Heming, GE Liang, CHEN Xueqin, YU Yibiao. et al.Speech separation based on sound localization and auditory masking effect[J]. ACTA ELECTRONICA SIGNICA, 2005, **33**(1): 158-160.

[46] **黄秀轩, 季飞, 韦岗. 基于分频带自相关函数的混叠语音基频分离提取新算法**[J]. **信号处理**, 2004, **20**(5): 490-493.
HUANG Xiuxuan, JI Fei, WEI Gang. Pitch extration new method for mixed speeches based on multi-band autocorrelation function[J]. Signal Processing, 2004, **20**(5): 490-493.

[47] Giannakis G B, Swami A. New results on stte-space and input-out-put identification of non-Gaussian processing using cumulants[A]. Proc. SPIE'87[C]. San Diego, CA, 1987, 826: 199-205.

[48] Linsker R. An application of the principle of maximum information preservation to linear systems[A]. Adv. Neural Inform. Processing Systems[C]. 1989.

[49] Linsker R. Self-organization in a perceptual network[J]. Computer, 1988, **21**: 105-117.

[50] Jutten C, Herault J. Blind separation of sources, Part **Ⅰ**: An adaptive algorithm based on neuromimatic architecture[J]. Signal Processing, 1991, **24**(1): 1-10.

[51] Comon P. Blind separation of sources, Part **Ⅱ**: Problem statement[J]. Signal Processing, 1991, **24**(1): 11-24.

[52] Sorouchyari E. Blind separation of sources, Part **Ⅲ**: Stability analysis[J]. Signal Processing, 1991, **24**(1): 21-29.

[53] Burel G. Blind separation of sources: A nonlinear neural algorithm[J]. Neural Network, 1992, **5**: 937-947.

[54] Comon P. Independent component analysis, a new concept[J]. Signal Processing, 1994, (36): 287-314.

[55] Cichocki A, Unbehauen R, Moszczynski L, et al. A new on-line adaptive learning algorithm for blind separation of source signal[A]. ISANN94[C]. Taiwan, 1994: 406-411.

[56] Oja E, Karhunen J, Wang L, et al. Principle and independent components in neural networks -Recent developments[A]. Proc. 7th Italian Workshop Newural Networks, WIRN'95[C]. Vietri, Italy, 1995: 20-26.

[57] Karhunen J, Joutsensalo J. Representation and separation of signals using nonlinear PCA type learning[J]. Neural Networks, 1994, **7**: 113-127.

[58] Bell A J, Sejnowski T J. An information-maximization approach to blind separation and blind deconvolution [J]. Neural Computation, 1995, **7**(6): 1004-1034.

[59] Delfosse N, Loubaton P. Adaptive blind separation of independent source: A deflation approach[J]. Signal Processing, 1995, (45): 59-83.

[60] Matsuoka K. A neural net for blind separation of nonstationary signals[J]. Neural Network, 1995, 3: 311-419.

[61] Cardoso J F, Laheld B H. Equivariant adaptive source separation[J]. IEEE Ttrans. Signal Processing, 1996, **44**(10): 3017-3030.

[62] Pearlmutter B A, Parra L C. A context-sensitive-generalization of ICA[A]. Proc. Int. Conf. Neural Information Processing, ICONIP'96[C]. Hong Kong, 1996: 151-157.

[63] Cardoso J F. Informax and maximum likelihood for source separation[J]. IEEE. Signal Processing Letters, 1997, 4: 109-111.

[64] Girolami M, Fyfe C. Kurtosis extrema and identification of independent components: a neural network approach [A]. Proc. ICASSP'97[C]. 1997, 4: 3329-3333.

[65] Pham D T, Garat P. Blind separation of independent sources through a quasi-maximum likelihood approach [J]. IEEE Trans. Signal Processing, 1997, **45**(7): 1712-1725.

[66] Valpola H. Nonlinear independent component analysis using ensemble learning: Theory[A]. Proc. ICA2000[C]. Helsinki, Finland, 2000: 251-256.

[67] Valpola H, Giannakopoulos X, Honkela A, et al. Nonlinear independent component analysis using ensemble learning: Experiments and discussion[A]. Proc. ICA2000 [C]. Helsinki, Finland, 2000: 351-356.

[68] Cho K S, Lee S Y. Implementation of Informax ICA algorithm with analog CMOS circuits[A]. Proc. ICA2001[C]. San Diego, CA, USA, 2001: 70-73.
[69] Cardoso J F, Delabrouille J, Patanchon G. Independent component analysis of the cosmic microwave background[A]. Proc. ICA2003[C]. Nara, Japan, 2003: 1111-1116.
[70] Yuan Z, Oja E. A FastICA Algorithm for non-negative independent component analysis[A]. Proc. ICA2004[C]. Granada, Spain, 2004: 1-8.
[71] 冯大政, 史维祥. 一种自适应信号盲分离和盲辨识的有效算法[J]. 西安交通大学学报. 1998, **32**(5): 76-79.
FENG Dazheng, SHI Weixiang. An efficient adaptive algorithms for blind separation and blind identification of signals[J]. Jounals of Xian Jiaotong University, 1998, **32**(5): 76-79.
[72] 胡光锐, 虞晓. 基于二阶前向结构和信息最大理论的语音增强算法[J].上海交通大学学报. 2000, **34**(7): 877-880.
HU Guangrui, YU Xiao. Speech enhancement based on second-order architecture and informax thoery[J]. Jounals of Xi′an Jiaotong University, 2000, **34**(7): 877-880.
[73] 汪军, 何振亚. 瞬时混叠信号盲分离[J]. 电子学报, 1997, **25**(4): 1-5.
WANG Jun, HE Zhenya. Blind source separation of instantaneous signals[J]. ACTA Electronica Sinica, 1997, **25**(4): 1-5.
[74] Van A J W der Kouwe, WANG D. A Comparison of auditory and blind separation technique for speech segregation[J]. IEEE Tran. on Speech and Audio Proc, 2001, **9**(3): 189-195.
[75] Okuno H G, Ikeda S, Nakatani T. Combining independent component analysis and sound stream segregation [A]. Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA′99)[C]. 1999, 92-98.
[76] SHI Zhenwei, TANG Huangwen, LIU Wenyu. Blind source separation of more sources than mixtures using generalizaed exponential mixture models[J]. Neurocomputing, 2004, 61: 461-469.
[77] SHI Zhenwei, TANG Huangwen, LIU Wenyu. Blind source separation of more sources than mixtures using sparse mixture models[J]. Pattern Recognition Letters, 2005, 26: 2491-2499.

高强度治疗超声(HITU)测量标准化的进展

高强度聚焦超声(HIFU)的医学应用研究和临床实践,经过 20 多年的发展,不断提高,逐步完善,取得长足进步。各种高强度治疗超声(HITU)设备产品纷纷涌现。促成了 2005 年我国国家标准《 GB/T 19890–2005 高强度聚焦超声(HIFU)声功率和声场特性的测量》的制定,也引起了国际电工委员会(IEC)87 技术委员会(TC87:超声学)第六工作组(WG6:聚焦换能器)和第八工作组(WG8:超声场测量)以及国际同行专家的重视。在得到了由笔者翻译的英文副本后,2006 年 2 月英国国家物理实验室(NPL)的专家 Adam Shaw 和英国癌症研究所(ICR)的 HIFU 专家 Gail ter Haar(国际治疗超声学会首任主席)联名发表技术报告 NPL REPORT DQL AC 015 "Requirements for Measurement Standards in High Intensity Focused Ultrasound (HIFU)Fields ",在其附录 C 中全文引用了我国国家标准。该技术报告被递交 IEC/TC87 后,经修改后随即成为 IEC 的技术报告《对高强度治疗超声 (HITU) 设备的测量标准的要求》("Requirements for measurement standards for High Intensity Therapeutic Ultrasound(HITU)Devices "),报告详细叙述了除碎石机和理疗仪以外的用于医学治疗的高强度聚焦超声(HIFU)声场的测量和规范的要求。其中附录 C 仍为 Chinese National Standard(18 pages)。此报告现已分发到各成员国,期望今后成为一个制定 HITU 标准的指导性文件。

由 Adam Shaw 负责起草标准 IEC 62555《超声学–功率测量–高强度治疗超声(HITU)换能器与系统的功率测量》(Ultrasonics-Power measurement-Power measurement for High Intensity Therapeutic Ultrasound(HITU)transducers and systems)。该标准的内容包括:规范测定频率为 0.5MHz~5MHz 超声换能器在水中的总发射声功率(可达 500W)的两种主要方法:基于测量液体靶的热膨胀的量热法和基于测量换能器处的反冲力的辐射力法。建立在 HITU 场中做声功率测量的三项原则:使用辐射力天平(RFB)、量热法、用水听器的平面扫描。对测量中避免声空化和对不同结构和几何形状的换能器(发散的、聚焦的和多元换能器)的测量提供相应的指南;提供综合测量不确定度的估计方面的信息。

由美国 ONDA 公司的专家 Sam Howard 起草的标准 IEC 87/382NP《超声学-外科系统-高强度治疗超声(HITU)换能器与系统的场参数的规范和测量》(Ultrasonics-Surgical systems-Specification and measurement of field parameters for High Intensity Therapeutic Ultrasound (HITU) Transducers and systems,draft 3)。该标准的内容包括:规范 HITU 场的定量参数;建立在相对低功率水平上测量场参数并将其外推到更高的功率水平的场参数的方法;对旁瓣和焦点前的最大声强的考虑;运动的换能器和多个声源的相关参数;"热指数"的定义等。此文件在正文第二条规范性的引用文件(2 Normative references)里,把 Chine HIFU Standard 正式列入其中,使我国国家标准首次成为 IEC 标准(草案)的有效条款。

2008 年 9 月 10 日在美国 Minnesoapolis 举行了 IEC/TC87/WG6 会议,讨论了上述 3 个文件,部署了进一步的修改补充工作进程。可以预期,在各国的共同努力下,HITU 的国际标准化工作将会取得较快的进展,在几年内完成它的制定标准使命。

上海交通大学生物医学工程系　寿文德