让语音识别听懂多人同时讲话——盲信号分离

拉夫伯勒的风

如今语音识别系统已经广泛应用于人们的生活中，成为人类与机器交互的主要手段之一，比如苹果的 Siri，微软的 Cortana，亚马逊的 Alexa 等等。但是，在我们实际使用语音识别系统的时候就会发现，如果在环境嘈杂、有很强干扰声的时候，语音识别系统的表现就会大打折扣。如果两个人同时问 Siri 问题，一个人问“今天是星期几”，一个人问“今天天气怎么样”，Siri 不会回答你任何一个问题，只会回答你“我没有听清”。因此，机器听觉首先要让机器“听清“，准确的获取指令，然后才能要求机器”听懂“，去执行人类的指令。

1.鸡尾酒会问题

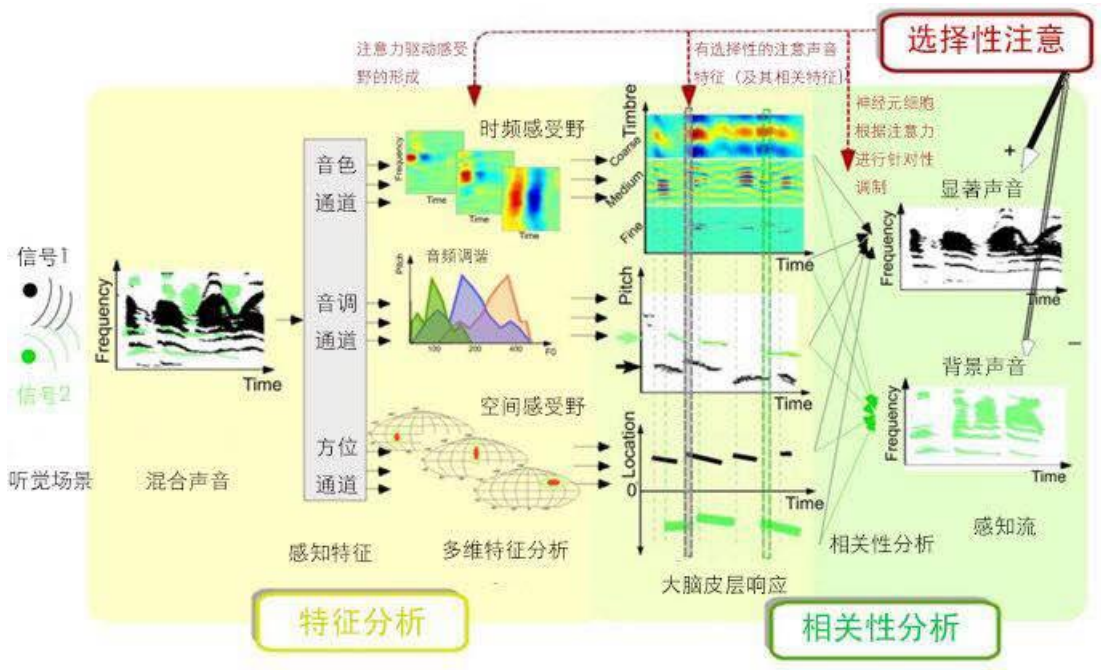
人类的听觉系统是除了视觉系统以外最重要的感觉系统，具有多种听觉功能，比如分辨声音的方位和距离，感觉声音的远近变化，选择性聆听感兴趣的声音等。其中，选择性聆听就是人类听觉能够在嘈杂环境中正常交流的根本原因。1953 年，Colin Cherry 首次提出著名的鸡尾酒会问题——“为什么人类在多个人同时说话时能够选择性的聆听，而机器却不具有这种能力？“

Colin Cherry （1914-1979）

鸡尾酒会问题

2. 计算听觉场景分析

从那时起，人们开始研究如何让机器能够像人类一样具有选择性聆听的能力。人们试图从分析人类听觉系统的感知原理入手，但是直到现在也没有完全搞明白人类为何具有这种能力。目前普遍接受的理论是听觉场景分析（Auditory Scene Analysis）。当我们聆听混合声音的时候，听觉系统会将混合信号进行一个分割、分析、再合成的基本过程，对频率、音调、音色、方位等特征信息分析分类，并重新归类合成完成声音的分离。在计算机领域，人们仿照这一过程建立了计算听觉场景分析法（Computational Auditory Scene Analysis）来解决鸡尾酒会问题。



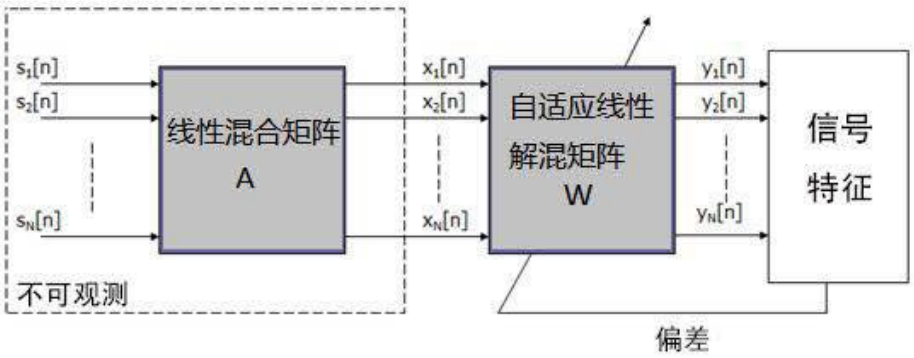
听觉场景分析过程 （原图引自参考文献[1]）

3．盲信号分离算法

而在信号处理领域，鸡尾酒会问题被转化为盲信号分离问题，我们需要在缺少源信号和传输信道先验知识的情况下，通过处理混合信号来求解源信号。

要解决这个看似不可能的问题，我们需要挖掘信号自身的特性，并利用这些特性来进行分离。一类方法是利用信号的高阶统计特性进行信号分离，即独立成分分析法（Independent Component Analysis），以及以此为基础发展而来的多种改进算法，如快速独立成分分析法（Fast ICA）、独立向量分析算法（Independent Vector Analysis）等等。另一类方法是利用信号的稀疏性进行信号分离，以稀

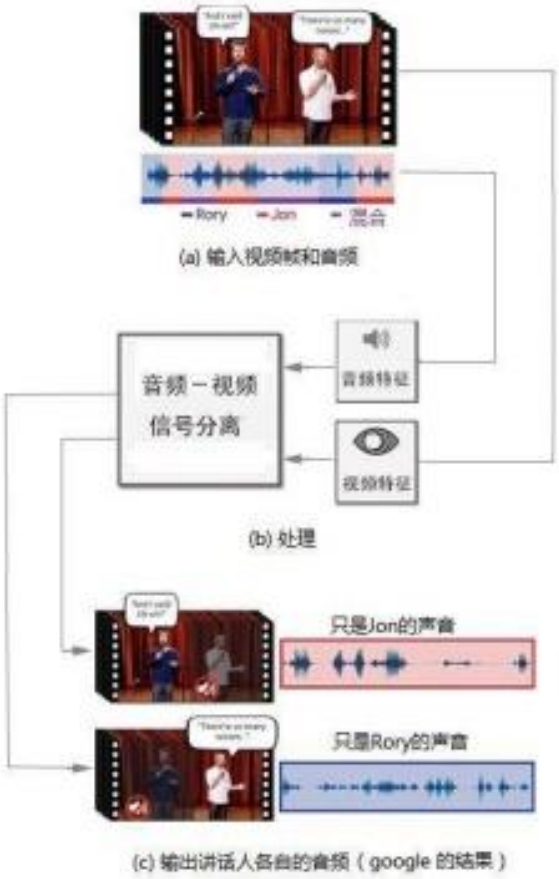
疏成分分析（Sparse Component Analysis）、非负矩阵分解（Nonnegative Matrix Factorization）和字典学习(Dictionary Learning)为代表。独立成分分析算法要求各个信号之间相互独立，且观测数要多于或等于信源数。而以稀疏性为基础的算法没有此限制，可用于解决观测数少于信源数情况下的分离问题。



盲信号分离示意图

4．盲信号分离与深度学习

随着最近深度学习的兴起，涌现了很多利用深度学习技术解决盲信号分离问题的方法，通过对大量数据的学习和训练，最终获得令人满意的分离效果。Google 最近的研究成果显示，在多人同时说话的视频中，用户可以任意选择一人的声音播放，屏蔽其他人语音。相信随着人工智能技术的发展，人类最终完全攻克鸡尾酒会问题将指日可待。



Google 解决鸡尾会问题的模型 （原图引自文献[2]）

参考文献

[1] Shihab A. Shamma, et. al. “Temporal coherence and attention in auditory scene analysis”，Trends in Neurosciences March 2011,Vol. 34, No. 3

[2] Ariel Ephrat, et. al. “Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation”，arXiv:1804.03619v1 [cs.SD] 10 Apr 2018

本文作者：拉夫伯勒的风，博士，长期从事音视频信号处理研究。

（本文原载：微信公众号：临菲信息技术港）