生理学报 *Acta Physiologica Sinica*, December 25, 2019, 71(6): 935–945
DOI: 10.13294/j.aps.2019.0060 http://www.actaps.com.cn

综 述

连续语音的神经编码——脑电图、脑磁图研究进展

潘珣祎[1]，邹家杰[2]，金培清[2]，丁 鼐[2,*]

浙江大学[1]外国语言文化与国际交流学院；[2]生物医学工程教育部重点实验室，生物医学工程与仪器科学学院，杭州 310027

摘 要：语音是人类交流的主要方式，语音理解也是人脑特有的核心认知功能。人脑中的动态神经活动如何编码连续语音流的声学特征，并解析出多个层级的语言结构是认知神经科学领域的重要问题。近年来，一系列脑电图、脑磁图研究通过包络跟踪响应、层级跟踪响应等新指标来刻画连续语音的神经加工机制。本文对这些研究进行综述，并聚焦于对两个语音加工问题：一是大脑如何编码语音中连续变化的声学特征。这方面的研究表明，大脑中的低频神经活动可以动态跟踪语音包络并且受到高级认知功能调节。二是大脑如何表征语音中不同大小的语言单元，比如音节、词、短语、语句。研究显示，大脑皮层中不同时间尺度的神经活动分别跟踪不同大小的语言单元，构成对多层级语言单元的并行表征。综合上述，近期研究初步揭示了大脑如何表征连续语音的声学特征并构建不同层次的语言单元，为进一步研究大脑如何加工连续语音提供了新的思路。

关键词：语音；语言；神经活动；包络跟踪响应；层级跟踪响应
中图分类号：Q427

The neural encoding of continuous speech – recent advances in EEG and MEG studies

PAN Xun-Yi[1], ZOU Jia-Jie[2], JIN Pei-Qing[2], DING Nai[2,*]

[1]*School of International Studies;* [2]*Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 310027, China*

Abstract: Speech comprehension is a central cognitive function of the human brain. In cognitive neuroscience, a fundamental question is to understand how neural activity encodes the acoustic properties of a continuous speech stream and resolves multiple levels of linguistic structures at the same time. This paper reviews the recently developed research paradigms that employ electroencephalography (EEG) or magnetoencephalography (MEG) to capture neural tracking of acoustic features or linguistic structures of continuous speech. This review focuses on two questions in speech processing: (1) The encoding of continuously changing acoustic properties of speech; (2) The representation of hierarchical linguistic units, including syllables, words, phrases and sentences. Studies have found that the low-frequency cortical activity tracks the speech envelope. In addition, the cortical activities on different time scales track multiple levels of linguistic units and constitute a representation of hierarchically organized linguistic units. The article reviewed these studies, which provided new insights into the processes of continuous speech in the human brain.

Key words: speech; language; neural activity; envelope tracking response; concurrent hierarchical tracking

语音理解是人脑特有的高级认知功能，是研究人脑复杂序列加工能力的经典案例[1–3]。语音理解分多个阶段进行：语音是一种声音，所以大脑要首先编码语音的声学特征，并根据这些声学特征识别出音素、音节等语音基本单元，随后根据语音信息识别词汇，再根据语法、语义及韵律信息将词整合成

Received 2019-04-08 Accepted 2019-08-15

Research from the corresponding author's laboratory was supported by grants from the National Natural Science Foundation of China (No. 31771248, 31500873) and Zhejiang Provincial Natural Science Foundation of China (No. LR16C090002, LY20C090008).

[*]Corresponding author. Tel: +86-571-87951086; E-mail: ding_nai@zju.edu.cn

为短语、语句等更大的结构。这些阶段可以按照上述顺序进行，但是各阶段之间也存在明显的交互作用[4-6]。

语音理解的不同阶段分别对应于怎样的神经电生理过程呢？大量研究通过脑电图的事件相关电位技术 (event-related potential, ERP) 和脑磁图的事件相关磁场技术 (event-related field, ERF) 对这个问题进行了探讨[7–16]。事件相关响应是研究语言理解认知神经机制的有力工具，其特点是依赖于对事件的定义 (比如某个词语的起始时间就可以看作一个事件)。然而，对于动态变化的连续语音，往往难以对事件进行准确定义，因而利用该技术研究连续语音加工存在一些困难。

近年来提出了多种新方法来研究动态神经活动如何加工动态语音信息。本文针对近年来研究较多的两种神经响应进行综述——包络跟踪响应(envelope tracking responses)和语言层级结构跟踪响应(cortical tracking of hierarchical linguistic structures)。包络跟踪响应指的是大脑在聆听语音时产生的与语音包络同步变化的神经响应。一般来说，包络跟踪响应与语音的声学特征相关，刻画了语音理解的听觉加工阶段。语音不仅包含声学特征，还承载了语言信息，语流中的语言信息按照分层级的语言单元进行组织 —— 词之类小的语言单元可以依据句法规则组合成为语句等大的语言单元。大脑对语音中包含的不同层次的句法结构产生的神经表征称为语言层级结构跟踪响应。层级跟踪响应与抽象的句法语义知识相关，刻画了语音理解的语言加工阶段。句法结构之外，语音中还存在韵律结构，包括韵律词、韵律短语、语调短语等大小不同的韵律单元[17–19]。本文对于语言层级结构跟踪响应的讨论集中于对句法结构响应的讨论，而且只关注大脑利用内隐句法知识进行句法结构加工的过程。对于韵律结构加工的认知神经科学研究可以参见相关文献[15, 16, 20, 21]。

本文综述的两种神经响应分别对应语音理解过程中的两个方面：动态声音特征的编码和语言结构的加工，这两部分研究汇总起来可以较为完整地刻画大脑对连续语音的加工和理解过程。下文首先简介语音的动态特征和时间结构，然后综述两种神经响应的基本现象和性质。

1 语音的动态特征与时间结构

1.1 语音的声学特征

在时间尺度上，语音特征可以分为语音精细结构 (fine structure) 和语音包络 (envelope)。语音精细结构包含音调、共振峰等信息，这些信息通常属于较高频率的声学特征 (一般在 100 Hz 以上)。语音包络则反映了声强随时间产生的低频变化，一般将语音包络定义为 50 Hz 以下的声强变化。语音包络与音节边界相关，因为一般而言音节中心部位的元音声强较高，而音节边界的声强较低；语音的精细结构则与音节的内部组成成分 (比如是 /ba/ 还是 /pa/) 及音调有关[22–24]。语音包络又可以分为宽带语音包络和窄带语音包络，宽带语音包络是指语音能量随时间的缓慢变化[25](图 1*A*)。宽带语音包络的一种常见的提取方法是首先计算每一时刻的语音能量 (定义为语音幅度的平方或绝对值)，然后再进行低通滤波 (截止频率为 50 Hz 或更低)。窄带语音包络则先将语音滤波到某一中心频率附近，然后再提取语音包络，后续也可以将各个中心频率附近的窄带包络进行平均，得到平均的窄带包络。语谱图中每个频率对应的横截面可以视为一个窄带包络 (图 1*B*)。图 1 以中文语句“老师在上课”为例，呈现了语音波形、宽带语音包络、窄带语音包络等语音特征，以及该语句的包络跟踪响应[26]。对于汉语、英语、法语等多种语言，语音包络的能量均集中在 4~8 Hz 这一范围[26](图 1*C*)。

本文主要讨论人脑对连续语音包络的神经响应。动物听觉皮层也可以加工声音包络，具体工作可以参见相关综述[27]。对于语音的精细结构，因为其主要反映音节内特征，其神经表征可以利用音节为单位进行研究，相关研究也已经有文章进行综述[28, 29]。

1.2 语言结构

语音中包含音素、音节等基本语音单元，这些基本语音单元可以组合为词素。词素是语义的基本单元，每个词素的意义均储存在长时记忆中。词素可以进一步组合构成语句和篇章，而且这一步骤至关重要，因为词素的数量有限，而有限数量的词素可以构成的语句是无限量的。正因为词素可以灵活地按照语法规则构成语句，人类语言才具有灵活表意的能力，才能具备以有限单元组合产生无限表达形式的多产性，以及对不同层次结构进行嵌套组合的递归性[30]。从理论语言学角度，一般认为词素首先组合成为短语，短语进一步迭代组合，形成具有层级组织结构的短语结构，最终构成语句[31, 32]。这一理论语言学假设是否真实描述了大脑中的语言加

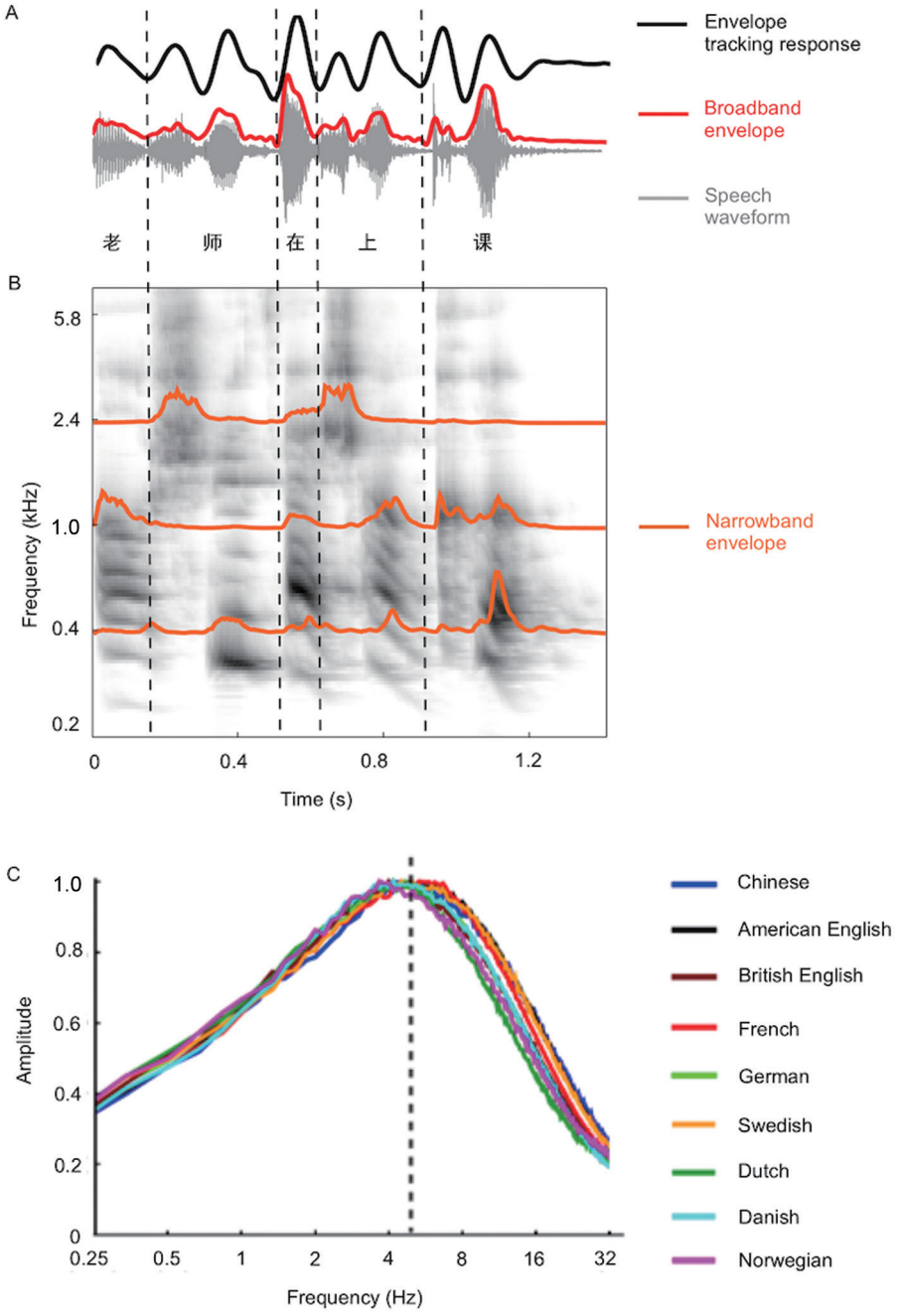

图 1. 语音包络及语音包络跟踪响应

Fig. 1. Speech envelope and envelope tracking response. *A*: The sound wave and spectrogram of the sentence “The teacher is giving a lesson” (in Chinese). The broadband spectrum envelope describes intensity fluctuations over time. Boundaries between syllables are indicated by dotted lines, corresponding to low intensity in speech. In speech listening, cortical activity recorded using EEG and MEG tracks the speech envelope (the line in black). *B*: Auditory spectrogram of speech and narrowband envelope. *C*: The spectra of speech envelope for different languages. The peak frequency of the speech envelope spectra ranges between 4 and 8 Hz for all tested languages (Adapted from Ding *et al*., 2017 [26] with permission).

工过程一直存在争议，但是下文第 3 节介绍的语言结构跟踪响应为这一假设提供了有力支持，研究发现至少对于简单短句，大脑确实形成了词、短语、语句这三个层次的神经表征。

2 动态听觉特征加工

2.1 连续语音的包络跟踪响应

大脑在聆听连续语音时会产生与语音包络同步变化的神经响应 (如图 1*A* 所示)[33, 34]，这种神经响

应一般称作包络跟踪响应，或者包络同步响应，可以利用脑电图、脑磁图等多种方法观测。包络跟踪响应是一种低频神经活动，主要集中在神经活动的 delta (1~4 Hz) 和 theta (4~8 Hz) 频段 [35, 36]。人类语言的声学结构包含多个时间尺度上的特征，跨语言分析发现大多语言的音节时长往往在 125~250 ms 之间；而更长时间尺度上的的声学特征则反映了短语和语句层面的韵律特征。大脑中 theta 波的时间窗口恰好与平均音节时长相对应，可能在语音的早期听觉分析和音节加工阶段起重要作用；而 delta 频段的神经响应则可能反映了大脑对短语、语句等更大的语言单元的加工。包络跟踪响应的出现说明听觉皮层编码了语音包络信息，而且编码的形式是与包络同步的低频神经振荡。包络跟踪响应产生于广泛的脑区，既包括位于 Heschl 氏回的听觉皮层 [37, 38]，也包括颞叶、额叶、顶叶的广泛区域 [39, 40]。

包络跟踪响应有多种量化方法。一种方法通过计算试次间神经响应的相关性 (inter-trial correlation, ITC) 来刻画包络跟踪响应。应用这一方法时需要多次播放同一段语音，由于每次播放的语音都具有相同的包络，每次语音聆听所产生的包络跟踪响应也应该基本相同。因此，可以通过计算多次聆听同一段语音所产生的神经响应之间的相关性来刻画包络跟踪响应 [35, 39]。一般将每次聆听一段语音的过程称为一个试次 (trial)，而多次聆听同一段语音时神经响应之间的相关性称作 ITC。ITC 可以利用神经响应波形之间的相关系数来计算，也可以在频率域利用神经响应的相位一致性来计算 (试次间相位一致性，inter-trial coherence)。计算上述两种 ITC 需要重复播放语音刺激，如果语音刺激长度较长 (比如 1 min 以上)，脑磁图实验中 2~3 次重复就可以计算出较为稳定的 ITC [41]。如果语音长度较短 [35]，或者实验工具是信噪比较低的脑电图 [42]，那么重复次数则需要大幅度增加。

包络跟踪响应的另一种刻画方法是计算神经响应与语音包络的相关性，这里称为刺激 - 响应相关性 (stimulus-response correlation)。神经响应的产生和传导需要时间，因而包络跟踪响应并不是实时地跟踪语音包络，而是具有一定延时的跟踪。考虑到这种延时效应，通常采用互相关函数 (cross-correlation function) 来计算语音包络与神经响应之间的相关度 —— 互相关函数计算不同延迟条件下语音包络与神经响应的相关系数 [33, 34, 37, 42, 43]。与互相关函数法类似的是时间响应函数法 (temporal response function, TRF) [36, 44]。时间响应函数可以视为互相关函数的扩展 [45]，它描述了语音包络中单位能量变化引起的神经响应变化 [36]。还有一种刻画神经响应与语音包络相关的方法是神经解码法 (neural decoding)，这种方法试图通过整合不同通道和不同延时的神经响应来解码语音包络波形 [36]。神经解码法的结果一般可以理解为全脑范围的动态神经响应到底以什么样的精度编码语音包络信息。刺激 - 响应相关性分析并不需要重复播放同一段语音，但是如果重复播放语音并平均其神经响应则可以起到去除随机噪音的效果，从而得到更稳定的刺激 - 响应相关性。

以上综述的两种包络跟踪响应刻画方法各有利弊。试次间相关性考察的是对同一语音多次刺激的响应中的共有成分，这个共有成分不仅限于对语音包络的响应，还可以包含对语音中其他特征的响应。刺激 - 响应相关性 (包括互相关函数、时间响应函数和神经解码法) 则针对包络与神经响应之间的关系进行建模，只考虑包络诱发的响应。但是由于采用线性模型，这些刺激 - 响应模型只能考察包络与神经响应之间的线性对应关系。

2.2 包络跟踪响应对应的语音加工过程

包络跟踪响应是一种可以稳定观测到的实验现象，但是其代表的语音加工过程仍存在争议。造成争议的根源在于语音包络与多种语音特征存在很强的相关性，因此很难判别神经响应是跟踪语音包络还是与之相关的其它特征。语音包络跟踪响应所对应的语音加工过程有多种假说，下面对几种主流假说 [46] 进行介绍与分析。

第一种假说可称为“声音边界跟踪”假说 (onset tracking hypothesis)。众所周知，如果单独呈现一个语音音节，大脑会产生听觉诱发响应 [7]。这个假说认为在连续语音中，每个显著的声音边界 (比如音节边界) 也可以诱发事件相关响应，而一连串声音边界所诱发的一连串的事件相关响应就构成了语音的包络跟踪响应。该假说认为包络跟踪响应实际上是大脑对语音边界离散响应迭加的结果 [47]。

第二种假说可称为“复合声学特征跟踪”假说 (collective feature tracking hypothesis)。人的听觉系统会把接收到的声音分解为基本特征，比如声强、音调、声源位置等 [48]。声强随时间产生的变化就是语音包络，而其他特征随时间变化的趋势大多也与

语音包络相关[49]，因此跟踪这些声学特征的神经活动会表现出与语音包络的同步性。该假说认为，脑电图、脑磁图等宏观电生理手段观测到的语音跟踪响应，实际上是跟踪不同语音特征的神经活动的复合响应[36]。

第三种假说可称为“音节切分”假说 (syllabic parsing hypothesis)。语音识别的过程中，听者需将连续语音信号切分为音素、音节、词等不同的语言单位，其中音节的边界在语音包络中有明显体现。如果神经活动能够跟踪音节中元音的位置，或者两个音节的边界，那么这种跟踪音节单元的神经活动也与语音包络具有很强的相关性[50]。该假说认为包络跟踪响应并不仅仅被动跟踪语音的声学特征，其主要作用在于主动切分与表征音节，传递语言信息[51]。

以上三种假说均获得了一些证据支持。“声音边界跟踪”假说可以看作是“复合声学特征跟踪”假说的一种特例 —— 将复合声学特征简化为声音边界这一种特征。其优势在于可以与传统的事件相关响应建立直接联系，劣势在于在连续变化的语音信号中，往往很难对声音边界进行准确定义。作为这种假说的证据支持，已有研究发现把语音中显著的声学边界去除之后，包络跟踪响应就消失了[52]。“复合声学特征跟踪”假说则涉及相对更广泛的声学特征，将包络跟踪响应视为对多项声学特征响应的叠加。其优势在于可以与动物模型电生理研究中的时频感受野模型 (spectrotemporal receptive field, STRF) 建立直接联系，但是这个假说仅限于提供了一个框架，没有明确说明神经响应到底跟踪哪种声学特征。近期研究显示除包络信息之外，脑电图、脑磁图响应也可以跟踪音素信息[53]、词汇信息[54]、以及语言结构[55]。“声音边界跟踪”和“复合声学特征跟踪”两种假说都认为包络跟踪响应反映了大脑对基本声学特征的编码，属于初级听觉加工，与语音理解并没有直接关系。换而言之，这两种假说认为包络跟踪响应反映了“听到语音”的过程，而并不反映“听懂语音”的过程。作为这两种假说的证据支持，已有研究表明雪貂、猕猴等动物初级听觉皮层记录到的电生理响应也可以很好地跟踪语音包络信息[55, 56]，而动物很显然并没有理解语音。

“音节切分”假说与前两个假说不同，它认为包络跟踪响应反映了大脑对音节这种离散语音单元的主动加工，反映了大脑将连续变化的声学特征转换为离散语言单元的过程。因为音节边界与语音包络密切相关 (参见 1.1 节)，所以很难直接将这一假说与上述两个假说区分开[57]。以上三种假说并不完全互斥，它们可能同时存在于大脑中，分别对应着语音加工的不同阶段。“声音边界跟踪”假说和“复合声学特征跟踪”假说很可能描述了初级听觉皮层对语音的加工，“音节切分”假说则很可能描述了更高级脑区将声学特征整合成为语言单元的过程。脑电图、脑磁图测量到的包络跟踪响应是产生于多个脑区的复合响应，因此它很可能包含了性质不同的多个响应成分。

上述三种假说关注的是包络跟踪响应所表征的语音特征，从神经生理学角度而言，对于包络跟踪响应的产生机制也有两种主要假说。一种假说认为包络跟踪响应是语音中特定特征所诱发的神经响应 (evoked response)，与大脑中的自发神经振荡 (instrinsic neuronal oscillation) 无关。另一种假说则认为语音信号中的特征可以重置大脑中自发神经振荡的相位 (phase resetting)，使之与语音包络同步。具体来说，在没有外部刺激的情况下，大脑中存在自发神经振荡；当语音特征出现的时候，这些自发神经振荡的相位被重置到特定值，然后再从这个特定相位开始振荡。这两种假说之间的争论在探讨 ERP 产生机制时同样存在[58]，它们的区分非常困难，对雪貂初级听觉皮层的研究表明，不同神经元集群中包络跟踪响应的产生机制可能不同[55]。

2.3 注意及语音可懂度对包络跟踪响应的调节作用

包络跟踪响应虽然是对包络这种基本声学特征的编码，但是它产生于大脑皮层，不仅仅是被动的编码机制，而是受到注意等高级认知功能的调节。在一个包含多个语音流的复杂环境中，听者可以选择性地注意并理解某一个语音流，但是不能理解其它未注意的语音流 —— 这一现象通常称为“鸡尾酒会效应”(cocktail party effect)[59]。研究显示，大脑对于注意的语音流的包络跟踪响应强于对未注意的语音流的包络跟踪响应[36, 38, 43, 60, 61]；并且注意与未注意的语音流的包络跟踪响应产生于不同脑区[38]，注意的调节作用在高级脑区的体现更加明显[39]。在包含多个语音流的复杂听觉场景中，注意对包络跟踪响应的调节作用非常稳定，但是在安静环境中播放单一语音流的条件下，注意对包络跟踪响应的调节作用非常有限[42]。

除了注意的调节作用之外，另一个关于包络跟

踪响应的热点研究问题是其与语音可懂度 (speech intelligibility) 的关系。在复杂听觉场景中，听者可以理解注意的语音流却不能理解非注意的语音流，而注意的语音流也诱发了更强的包络跟踪响应，这一现象说明包络跟踪响应有可能和语音理解相关。但是语音可懂度受到很多因素影响，因此仅根据这一现象还不能说明两者间的直接联系。目前已有大量研究探讨了包络跟踪响应与语音可懂度之间的关系，根据调节语音可懂度的方法，这些研究大致可以分为三类。

第一类研究通过改变语音的声学特征来调节语音的可懂度。研究一致发现当引入噪音干扰时，语音可懂度降低，包络跟踪响应也减弱 [41, 43, 62–64]。语音可懂度的降低还可以由多种其它因素导致，比如倒序播放语音、破坏语音精细结构、加快语速等。当这些因素导致语音可懂度降低时，一些研究发现包络跟踪响应减弱 [33, 35, 52, 53, 65, 66]，而另一些研究发现包络跟踪响应并不发生改变 [37, 46, 47, 67]。综合这些研究可以发现，在噪音太强的情况下，语音难以听到，包络跟踪响应也减弱；当其它声学特征发生改变时，语音依然可以清晰地听到，只是难以理解，这些情况下包络跟踪响应是否减弱在文献中还没有一致的结论。

第二类研究通过改变语言层面的上下文信息来调节语音可懂度。一项研究比较了由伪词构成的语音与由真词构成的语音诱发的响应。该研究发现在 delta 频段上，对伪词的包络跟踪响应要强于对真词的包络跟踪响应 [68]。另一项研究比较了粤语母语者和不懂粤语的听者对粤语语段的响应，研究发现粤语母语者的包络跟踪响应比不懂粤语者更弱 [62]。这两项研究均表明语言层面的上下文信息减弱了包络跟踪响应。造成这一现象的原因可能有两个，一是在缺乏语言上下文信息的情况下，听者只能依赖于声学线索进行语音识别，因此更加注意语音包络等声学特征；二是存在语言上下文信息时，大脑中产生了语言结构跟踪响应 (3.1 节)，而语言结构跟踪响应与包络跟踪响应之间可能存在对加工资源的争夺，因而是互相抑制的竞争关系 [62]。

第三类研究比较包络跟踪响应与个体语音理解之间的关系。即使听力正常的青年在噪音环境下的语音理解能力也有很大个体差异 [69]。研究表明包络跟踪响应更强的个体能够更好地理解语音 [41, 43, 46, 52, 64]。然而，当比较不同听者群体时，研究发现老龄化引起了噪音环境下语音识别能力的下降，但是却增强了包络跟踪响应 [70]。

综上，包络跟踪响应与语言可懂度之间的关系非常复杂，并不是简单的正相关关系。另外，即使两者存在相关也很难确定其间的因果关系 —— 更强的包络跟踪响应导致了更好的语音理解还是更好的语音理解可以反馈增强包络跟踪响应 [46]。

3 层级语言结构加工

3.1 多层级语言结构跟踪响应

现有研究中，多层级语言结构跟踪响应的诱发大多依赖于特定的语音材料，因此下文首先对这种语音材料进行简要描述。一方面，语言结构跟踪响应关注于基于脑中的语言知识进行的语言结构加工，而不是对语音中语言结构相关的韵律线索的加工，因此，在实验材料中需要去除与语言结构边界相关的韵律线索。实验中通过采用逐个音节独立合成语音的方法来实现这一目的 —— 当每个音节均独立合成的时候，音节的韵律特征不随上下文而发生改变，因此与语言结构无关。另一方面，为了简化频率域的数据分析，实验中往往将音节按照恒定速率播放，例如 Ding 等 [40] 的研究中，语音材料每秒钟匀速播放 4 个音节 (即 4 Hz 的播放频率)。在这种设计中，音节的长度均衡设定为 1/4 s (250 ms)，对于这样匀速播放的音节，音节相关的神经响应也锁定在同样的单一频率 ——4 Hz。

实验通过排列音节来嵌入更高层次的语言结构 (图 2*A*)。例如，在 Ding 等 [40] 的研究中，相邻的音节可以组成一个双字词或者短语 (绵羊、吃草等)，而相邻短语可以组成一个短句 (绵羊吃草、老师上课等)。对于一个包含 2 个音节的短语，其持续时间是 1/2 s，所以每秒钟匀速播放 2 个短语 (播放速率为 2 Hz)，对应的神经响应的频率为 2 Hz。与此相似，对于一个包含 4 个音节的短句，其持续时间是 1 s，播放速率为 1 Hz，对应的神经响应的频率也为 1 Hz。基于这样的实验设计，1 Hz 和 2 Hz 的神经响应分别代表了大脑对 4 字结构 (该实验中为短句) 和 2 字结构 (该实验中为短语) 的响应，这些响应统称为语言结构跟踪响应。一般而言，如果每秒播放 N 个音节，那么对于包含 M 个音节的语言结构的跟踪响应的频率为 N/M 赫兹。正常语音中的音节速率大约为 4~5 Hz，所以语言结构跟踪响应的频率范围集中在 4 Hz 以下的 delta 频段。

在聆听上述四字短句序列的过程中，脑磁图和脑电图中观测到了 1、2、4 Hz 的神经响应 (图 2*B*)，表明大脑确实将匀速播放的语音流整合成为多个语言层级。其中 4 Hz 的神经响应与音节节奏相对应，也与语音包络的频率相对应，可以解释为包络跟踪响应；而 1 Hz 和 2 Hz 的响应则不能解释为包络跟踪响应，它们反映了大脑对短语、语句等高层级语言结构的响应。与语言结构跟踪响应密切相关的是事件相关响应中的终止正漂移 (closure positive shift, CPS) 成分，CPS 成分受到韵律层级的调制 —— 音节、音系短语和语调短语诱发不同波形的 CPS [20, 71]。语言结构跟踪响应和 CPS 的区别在于语言结构跟踪响应是由不包含韵律线索的语音刺激诱发的，而 CPS 一般认为是音调等韵律线索诱发的响应 [20]。

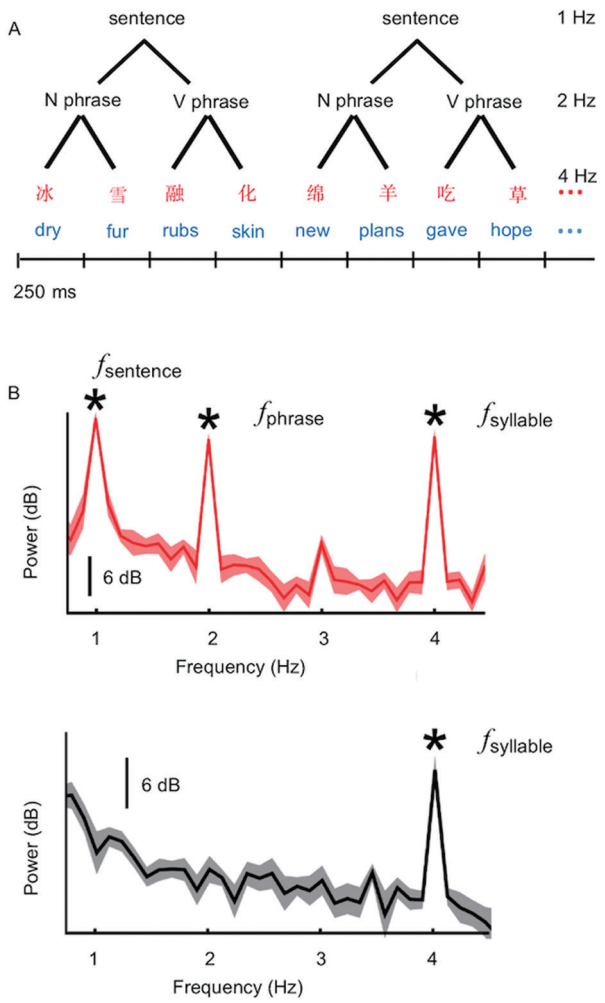

图 2. 语音中的语言结构及语言结构跟踪响应

Fig. 2. Neural tracking of hierarchical linguistic structures. *A*: Sequences of 4-syllable sentences are presented as stimuli, in which the first 2 syllables form a noun phrase and the last 2 syllables form a verb phrase. The syllables are isochronously presented at a rate of 4 Hz. *B*: MEG response spectra to the stimuli used in *A*. The upper panel: Three spectral peaks can be seen in the neural response spectrum for native Chinese listeners, at the corresponding frequencies of the sentence, phrase and syllable levels, respectively. This indicates that cortical activity can track linguistic structures of different hierarchies based on linguistic knowledge. The lower panel: The neural response spectrum of non-native listeners shows the tracking at the syllabic level only. Similar responses can be observed by using EEG [41, 72]. Adapted from Ding *et al*., 2016 [40] with permission.

近期一项研究进一步表明，不但大脑皮层的神经活动可以跟踪语句结构，这种语句跟踪响应在眼部肌肉活动中也有所反映 [73]。该研究使用的语音材料与上述实验相同 —— 每秒匀速播放 4 个音节，前两个音节可组合为名词短语，后两个音节可组合为动词短语，四个音节组合为语句，因而语句频率为 1 Hz。实验观测对象为对应不同层次结构播放频率的脑电和眼电响应。结果显示，无论被试在聆听过程中保持睁眼还是闭眼，他们的垂直眼电信号中均表现出了 1 Hz 的语句跟踪响应。在睁眼情况下，眼动仪数据表明被试的眨眼与语句节奏相同步。这一结果说明对语句的加工很可能涉及运动皮层，而运动皮层中的语句跟踪响应导致了眼部肌肉的节奏性收缩。那么，不同层级的语言结构跟踪响应发生于哪些脑区呢？颅内脑电 [40] 和脑磁图溯源结果 [74] 表明每个层级的加工都涉及颞叶的多个脑区，不同层级加工的具体脑区在微观层面上交织在一起，但并不完全重合。颞叶之外，音节跟踪响应还激活了右侧的运动皮层，语句跟踪响应则激活了左侧额叶。

3.2 任务与注意对语言结构跟踪响应的调节

在 Ding 等 [40] 的研究中，实验任务为判断是否出现主谓语义搭配不当的短句，比如“公鸡开车”。具体来说，大多数试次中所有短句均为主谓搭配合理的短句，比如“公鸡打鸣”、“司机开车”，一小部分试次中偶尔出现一个“公鸡开车”这类搭配不当的短句。被试需要对这两类试次做出不同的按键反应。这一任务需要被试将主语和谓语的信息整合为语句才能完成，因此需要被试同时注意短语和语句层面的信息。在 Jin 等 [73] 的研究中，一个实验任务为判断一段连续语流末尾播放的语句是否完整，假如末尾播放的双音节名词后缺少与其搭配成句的双音节动词，则判断为语句不完整，且做出相应的按键反应。这一实验任务也要求被试将双音节名词 (主语) 和双音节动词 (谓语) 的信息进行整合加工。

近期的一项研究进一步表明，任务和注意对语言结构跟踪响应的产生至关重要[75]。该研究实验材料的设计同样遵循上文所述原则，单独合成每个音节以去除韵律线索，且语流同样以每秒 4 个音节的恒定速率播放，其中一些相邻的音节组成双音节名词，因而语音材料中的音节频率为 4 Hz，双音节词语的频率为 2 Hz。实验观测对象是对应音节和词语两个不同层级语言结构的神经响应，其频率也分别锁定在 4 Hz 和 2 Hz。实验中的一个任务要求被试判断一个试次中有生命性名词更多还是非生命性名词更多，这个任务需要被试关注词语层面的信息。实验中的另一个任务要求被试检测某一个音节的音调是否发生改变，这个任务需要被试关注语音的基本声学特征。结果显示，相比于关注语音声学特征，关注词汇生命性的任务要求诱发了更强的对应双字词频率 (2 Hz) 的语言结构跟踪响应，然而对应音节频率的响应则没有受到任务的显著调节。这说明实验任务对加工深度的要求对于词语跟踪响应具有显著的调节作用，但是对于音节跟踪响应的调节作用不明显。该研究还考察了对不同类型干扰刺激的加工是否影响语言结构跟踪响应。实验过程中，在被试聆听语音材料的同时播放不同的干扰刺激，构成不同的实验条件。结果显示，无论干扰刺激为语音 (朗读的故事)、复杂视觉刺激 (无声电影) 还是无意义的简单视听觉刺激，只要被试注意的是干扰刺激，而不是实验材料，在其脑电信号中均未能检测到双字词频率的语言结构跟踪响应，只能检测到对应音节频率的响应，说明双字词跟踪响应的产生需要注意调节。

另一项研究比较了清醒和睡眠状态下的语言结构跟踪响应[76]。该研究在以色列完成，使用了希伯来语的语句作为刺激材料。语音材料中嵌入了音节、词、短语、语句四个层级的语言结构，分别以恒定速率呈现。此外，还使用了结构相似却无意义的希伯来语流、以及被试无法理解的外语语句作为对照材料。结果显示，无论是清醒还是睡眠状态下，也无论使用何种语音材料，都可以稳定观测到对应音节频率的脑电响应。然而，只有被试在清醒状态下聆听可理解语流时，才可以检测到词、短语和语句频率的语言结构跟踪响应，睡眠状态下 (无论是快速眼动睡眠还是二期或三期睡眠) 仅能检测到音节频率的脑电响应，而没有检测到词以及更高级语言结构的跟踪响应。

这两项研究表明，对音节这一基本语音单元的神经加工在很大程度上是一个自动化的过程，在非注意以及睡眠状态均可观测到音节频率的神经响应，但是注意可以进一步增强音节响应的幅度[75]。与之相反，将音节整合成为双音节词、短语等更高层级结构的过程则强烈依赖于自上而下的注意调节，在非注意以及睡眠状态下，不能稳定观测到对应双音节词或更大的语言单元频率的响应。

4 总结和展望

包络跟踪响应和语言结构跟踪响应分别是研究连续语音听觉加工和语言结构加工的有力工具。这两种响应不是在特定时刻点由单个事件诱发的响应波形，而是描述了连续变化的神经电生理信号与连续变化的语音包络或语音结构之间的耦合关系。这两种神经响应表明大脑在加工语音的过程中，delta 和 theta 频段的低频神经活动表征语音中的包络信息，delta 频段活动则进一步参与将音节等基本语音单元整合为高级语言结构的过程。研究表明包络跟踪响应可以看作语音理解的必要条件。当听觉环境中包含噪声时或者当语音特征遭到破坏时，只要语音依然可懂，脑电图、脑磁图信号中就会出现稳定的包络跟踪响应。然而，包络跟踪响应并不是语音理解的充分条件，很多情况下，虽然语音不可懂，但是大脑中依然出现稳定的包络跟踪响应。语言结构跟踪响应反映了大脑将小的语言结构整合为大的语言结构的过程。在现有实验中，语言结构跟踪响应与语音理解存在直接对应关系，只有注意聆听可懂语言时才产生跟踪多音节词及更高层级语言结构的神经响应。

综上，包络跟踪响应和语言结构跟踪响应分别刻画语音加工的不同阶段，为研究连续语音神经编码提供了有力工具。后续研究可以进一步探索大脑可否在复杂听觉环境中构建多层级的语言结构表征，分析表征不同层级语言结构的脑区之间的交互作用，以及语义、语境等因素如何调控层级语言结构的构建，从而对大脑的语音加工过程进行更完整的刻画。

参考文献

1 Lashley KS. The problem of serial order in behavior. In: Jeffress LA (ed.). Cerebral Mechanisms in Behavior: The Hixon Symposium. New York: John Wiley Press, 1951, 112–136.

2 Hauser MD, Chomsky N, Fitch WT. The faculty of language: What is it, who has it, and how did it evolve? Science 2002; 298(5598): 1569–1579.
3 Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C. The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. Neuron 2015; 88(1): 2–19.
4 Ganong WF. Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform 1980; 6(1): 110–125.
5 Warren RM. Perceptual restoration of missing speech sounds. Science 1970; 167(3917): 392–393.
6 Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. J Exp Psychol 1951; 41(5): 329–335.
7 Näätänen R, Picton T. The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. Psychophysiology 1987; 24(4): 375–425.
8 Wagner M, Shafer VL, Martin B, Steinschneider M. The phonotactic influence on the perception of a consonant cluster /pt/ by native English and native Polish listeners: A behavioral and event related potential (ERP) study. Brain Lang 2012; 123(1): 30–41.
9 Kutas M, Federmeier KD. Electrophysiology reveals semantic memory use in language comprehension. Trends Cogn Sci 2000; 4(12): 463–470.
10 Lau EF, Phillips C, Poeppel D. A cortical network for semantics: (de)constructing the N400. Nat Rev Neurosci 2008; 9: 920–933.
11 Friederici AD, Pfeifer E, Hahne A. Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. Cogn Brain Res 1993; 1: 183–192.
12 Hagoort P, Brown CM. ERP effects of listening to speech: Semantic ERP effects. Neuropsychologia 2000; 38: 1518–1530.
13 Hagoort P, Brown CM. ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. Neuropsychologia 2000; 38: 1531–1549.
14 Friederici AD. Towards a neural basis of auditory sentence processing. Trends Cogn Sci 2002; 6(2): 78–84.
15 Steinhauer K, Alter K, Friederici AD. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nat Neurosci 1999; 2(2): 191–196.
16 Li W, Yang Y. Perception of prosodic hierarchical boundaries in Mandarin Chinese sentences. Neuroscience 2009; 158(4): 1416–1425.
17 Selkirk EO. Phonology and Syntax: The Relationship between Sound and Structure. Cambridge, MA: MIT Press, 1986.
18 Turk A, Shattuck-Hufnagel S. What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. Lab Phonol 2013; 4(1): 93–118.
19 Shattuck-Hufnagel S, Turk AE. A prosody tutorial for investigators of auditory sentence processing. J Psycholinguist Res 1996; 25(2): 193–247.
20 Li WJ (李卫君), Yang YF. The cognitive processing of prosodic boundary and its related brain effect in quatrain. Acta Psychol Sin (心理学报) 2010; 42(11): 1021–1032 (in Chinese with English abstract).
21 Yang YF (杨玉芳). Psycholinguistics. Beijing: Science Press, 2015, 214–227 (in Chinese).
22 Greenberg S. Speaking in shorthand — A syllable-centric perspective for understanding pronunciation variation. Speech Commun 1999; 29: 159–176.
23 Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech — A syllable-centric perspective. J Phon 2003; 31(3–4): 465–485.
24 Stevens KN. Toward a model for lexical access based on acoustic landmarks and distinctive features. J Acoust Soc Am 2002; 111: 1872–1891.
25 Rosen S. Temporal information in speech: Acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 1992; 336(1278): 367–373.
26 Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. Temporal modulations in speech and music. Neurosci Biobehav Rev 2017; 81: 181–187.
27 Joris PX, Schreiner CE, Rees A. Neural processing of amplitude-modulated sounds. Physiol Rev 2004; 84(2): 541–577.
28 Johnson K, Nicol T, Kraus N. Brain stem response to speech: A biological marker of auditory processing. Ear Hear 2005; 26(5): 424–434.
29 Chandrasekaran B, Kraus N. The scalp-recorded brainstem response to speech: Neural origins and plasticity. Psychophysiology 2010; 47(2): 236–246.
30 Humboldt WV. On Language: The Diversity of Human Languages — Construction and its Influence on the Mental Development of the Human Species. Translated by Peter Heath, Cambridge: Cambridge University Press, 1999.
31 Chomsky N. Syntactic Structures. Berlin/New York: Mouton de Gruyter, 1957.
32 Everaert MB, Huybregts MA, Chomsky N, Berwick RC, Bolhuis JJ. Structures, not strings: Linguistics as part of the cognitive sciences. Trends Cogn Sci 2015; 19(12): 729–743.
33 Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with

temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 2001; 98(23): 13367–13372.
34 Aiken SJ, Picton TW. Human cortical responses to the speech envelope. Ear Hear 2008; 29(2): 139–157.
35 Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 2007; 54(6): 1001–1010.
36 Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 2012; 107(1): 78–89.
37 Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Howard MA, Brugge JF. Temporal envelope of time-compressed speech represented in the human auditory cortex. J Neurosci 2009; 29(49): 15564–15574.
38 Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 2012; 109(29): 11854–11859.
39 Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron 2013; 77(5): 980–991.
40 Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 2016; 19(1): 158–164.
41 Ding N, Simon JZ. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 2013; 33(13): 5728–5735.
42 Kong YY, Mullangi A, Ding N. Differential modulation of auditory responses to attended and unattended speech in different listening conditions. Hearing Res 2014; 316: 73–81.
43 Kong YY, Somarowthu A, Ding N. Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. J Assoc Res Otolaryngol 2015; 16(6): 783–796.
44 Lalor EC, Power AJ, Reilly RB, Foxe JJ. Resolving precise temporal processing properties of the auditory system using continuous stimuli. J Neurophysiol 2009; 102(1): 349–359.
45 Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network 2001; 12(3): 289–316.
46 Ding N, Simon JZ. Cortical entrainment to continuous speech: Functional roles and interpretations. Front Hum Neurosci 2014; 8: 311.
47 Howard MF, Poeppel D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J Neurophysiol 2010; 104(5): 2500–2511.
48 Shamma S. On the role of space and time in auditory processing. Trends Cogn Sci 2001; 5: 340–348.
49 Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. Trends Neurosci 2011; 34(3): 114–123.
50 Ghitza O. The theta-syllable: A unit of speech information defined by cortical function. Front Psychol 2013; 4: 5.
51 Giraud AL, Poeppel D. Cortical oscillations and speech processing: Emerging computational principles and operations. Nat Neurosci 2012; 15(4): 511–517.
52 Doelling K, Arnal L, Ghitza O, Poeppel D. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage 2014; 85: 761–768.
53 Di Liberto Giovanni M, O'Sullivan James A, Lalor Edmund C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 2015; 25(19): 2457–2465.
54 Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 2018; 28: 1–7.
55 Ding N, Simon JZ, Shamma SA, David SV. Encoding of natural sounds by variance of the cortical local field potential. J Neurophysiol 2016; 115(5): 2389–2398.
56 Steinschneider M, Nourski KV, Fishman YI. Representation of speech in human auditory cortex: Is it special? Hear Res 2013: 57–73.
57 Cummins F. Oscillators and syllables: A cautionary note. Front Psychol 2012; 3: 364.
58 Shah AS, Bressler SL, Knuth KH, Ding M, Mehta AD, Ulbert I, Schroeder CE. Neural dynamics and the fundamental mechanisms of event-related brain potentials. Cereb Cortex 2004; 14(5): 476–483.
59 Cherry EC. Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 1953; 25(5): 975–979.
60 O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex 2015; 25(7): 1697–1706.
61 Wang Y, Zhang J, Zou J, Luo H, Ding N. Prior knowledge guides speech segregation in human auditory cortex. Cereb Cortex 2019; 29(4): 1561–1571.
62 Zou J, Feng J, Xu T, Jin P, Luo C, Zhang J, Pan X, Chen F, Zheng J, Ding N. Auditory and language contributions to neural encoding of speech features in noisy environments.

NeuroImage 2019; 192: 66–75.
63 Keitel A, Gross J, Kayser C. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol 2018; 16(3): e2004473.
64 Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T. Speech intelligibility predicted from neural entrainment of the speech envelope. J Assoc Res Otolaryngol 2018: 1–11.
65 Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol 2013; 11(12): e1001752.
66 Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 2013; 23(6): 1378–1387.
67 Zoefel B, VanRullen R. EEG oscillations entrain their phase to high-level features of speech sound. NeuroImage 2016; 124(Pt A): 16–23.
68 Mai G, Minett JW, Wang WS. Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. NeuroImage 2016; 133: 516–528.
69 Ruggles D, Bharadwaj H, Shinn-Cunningham BG. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proc Natl Acad Sci U S A 2011; 108(37): 15516–15521.
70 Presacco A, Simon JZ, Anderson S. Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. J Neurophysiol 2016; 116(5): 2346–2355.
71 Li W, Yang Y. Perception of Chinese poem and its electrophysiological effects. Neuroscience 2010; 168(3): 757–768.
72 Ding N, Melloni L, Yang A, Wang Y, Zhang W, Poeppel D. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front Hum Neurosci 2017; 11: 481.
73 Jin P, Zou J, Zhou T, Ding N. Eye activity tracks task-relevant structures during speech and auditory sequence perception. Nat Commun 2018; 9(1): 5374.
74 Sheng J, Zheng L, Lyu B, Cen Z, Qin L, Tan LH, Huang MX, Ding N, Gao JH. The cortical maps of hierarchical linguistic structures during speech perception. Cereb Cortex 2018. doi: 10.1093/cercor/bhy191.
75 Ding N, Pan X, Luo C, Su N, Zhang W, Zhang J. Attention is required for knowledge-based sequential grouping: Insights from the integration of syllables into words. J Neurosci 2018; 38(5): 1178–1188.
76 Makov S, Sharon O, Ding N, Ben-Shachar M, Nir Y, Golumbic EZ. Sleep disrupts high-level speech parsing despite significant basic auditory processing. J Neurosci 2017; 37(32): 7772–7781.