语音音节节律的声学相关因素：

时间包络的调制谱还是局部特征

Yuran Zhang[a]、Jiajie Zou[a]、Nai Ding[a], *

[a]生物医学工程与仪器科学学院,

中国浙江省杭州市，浙江大学

***通讯作者：**

电子邮箱：ding_nai@zju.edu.cn（N. Ding）。浙江大学生物医学工程与仪器科学学院，中国浙江杭州，310027

摘要

音节是语音中一种具有显著感知特征的单位。由于音节及其声学相关物（即语音包络）的节律性偏好范围均为 4 至 8 赫兹，因此有研究假设，θ频段神经振荡在基于语音包络提取音节的过程中发挥重要作用。然而，文献综述显示，语音包络与音节之间的关系尚无一致证据。本研究通过分析大型语音语料库重新探讨这一问题。研究表明，只有在对数分钟的语音录音进行合并分析时，以调制谱为特征的语音包络中心频率才会与音节速率稳定相关。相比之下，在时域中，语音包络的一个成分与音节起始稳定锁相。根据说话人无关模型，音节起始的时间可解释语音包络约 24% 的方差。这些结果表明，相较于调制谱，语音包络中的局部特征是音节更可靠的声学相关物。

关键词：语音包络；调制谱；音节；神经同步 nt

1. 引言

语音信息以层级结构组织成不同大小的单位，例如音位、音节、语素、词语，以及短语和句子等更大的单位。音节是一种中尺度语音单位，其时间尺度介于音位和语素之间；音位和语素分别是最基本的语音单位和语义单位。过去二十年来，音节神经表征背后的神经机制受到了广泛关注：研究者假设，θ频段神经振荡（4–8 赫兹）可能提供一种机制，将连续语音流解析为离散的音节单位（Assaneo 和 Poeppel，2018；Ghitza，2013；Giraud 和 Poeppel，2012；Greenberg，1999；Hovsepyan 等人，2020；Hyafil 等人，2015；Poeppel等人，2008）。根据这一假设，音节是一种基本单位，可将语音的听觉表征转化为语言表征。这一假设主要基于两项观察。首先，与音位和语素相比，音节在知觉上更为显著，即使语言知识相对有限，人们也能感知到音节（Liberman等人，1974；Mehler 等人，1981）。音节具有知觉显著性的可能原因是，人们认为音节具有可靠的声学相关物，即语音包络（Greenberg 等，2003）。语音包络被定义为声功率的低频（通常低于 16 赫兹）起伏（Rosen，1992），既可以是在窄频带内，也可以是跨越所有频率的起伏（图 1）。声包络是一种基本的声学特征，在听觉皮层中有良好表征（Shamma，2001），并且对语音可懂度至关重要（Chi 等，2005，1999；Elhilali 等，2003；Elliott 和 Theunissen，2009）。由于人们通常认为语音包络与音节边界之间存在简单关系，因此有研究提出，大脑可能先通过分析语音包络确定音节之间的边界，然后再进一步解码语音内容

基于对时频听觉特征更为细致的分析（Doelling 等人，2014；Ghitza，2013）。

其次，音节和语音包络都呈准节律性：不同语言中的平均音节速率通常介于 4 至 8 赫兹之间（Coupé 等人，2019；Greenberg 等，2003；Pellegrino et al., 2011），而语音调制谱（即语音包络的功率谱）的功率在不同语言中也集中于 4–8 赫兹之间（Ding 等人，2017；Varnet 等人，2017）。更重要的是，聆听语音时，通过脑磁图和脑电图测得的大尺度 θ 频段皮层活动受语音包络驱动的程度，强于受音位和词汇特征驱动的程度（Ding 和 Simon，2012；Lalor 和 Foxe，2010；Brodbeck 等人，2018；Daube 等人，2019）。这种 θ 频段包络追踪反应也常被视为音节加工的神经标记（Ding 等人，2016；Doelling 等人，2014），并广泛用于探究语音编码背后的神经机制（Brodbeck 和 Simon，2020；Ding 和 Simon，2014），以及刻画特殊人群的语音加工（Di Liberto 等人，2022；Palana 等，2021）。

尽管语音包络被视为音节节律的一个主要声学相关物，但仔细审视相关文献便会发现，这种关系并不像通常所认为的那样确凿。例如，一方面，人们普遍认为音节速率与语音调制谱的峰值频率相符（Greenberg 等，2003）。另一方面，有研究指出，音节速率与峰值调制频率具有不同的特性：一些研究表明，不同语言的平均音节速率差异很大（Coupé et al., 2019; Pellegrino et al., 2011），而另一些研究则表明，峰值调制频率在不同语言中

高度一致（Ding 等人，2017; Varnet 等人，2017）。其次，只有在去除 1/f 趋势后，语音调制谱才会在 4–5 赫兹附近达到峰值（Ding 等人，2017）。否则，峰值调制频率约为 0 赫兹（Voss 和 Clarke，1975）。相比之下，计算平均音节速率时无需去除 1/f 趋势。因此，音节速率究竟只是碰巧与峰值调制频率相符，还是两种信号确实相关，仍不清楚。文献中的结果之所以不一致，还因为以往研究常常分析不同的语音语料库，并采用不同的方法计算音节速率和峰值调制频率。

音节速率与峰值调制频率一样，都是一段录音的汇总统计量。语音包络与音节之间关系密切的进一步证据，来自语音包络与音节之间的瞬时锁相：人们普遍认为，音节边界对应于语音包络中的局部极小值（Poeppel 和 Assaneo，2020），而且可以通过分析语音包络检测出音节边界（Hovsepyan 等人，2020；Hyafil 等人，2015；Villing 等人，2006）。然而，显然语音包络编码的信息并不局限于音节边界。例如，窄带包络能够很好地编码语音信息（Stevens，2002），而韵律信息（包括时长信息和间隔）则会显著影响窄带和宽带包络（Li 和 Yang，2009）。目前尚不清楚音节边界在多大程度上能够解释语音包络，以及与音节边界对应的包络特征是否会随语言和说话风格而变化。

在此，我们通过分析大型语音语料库（>1000 小时的录音），重新探讨调制谱与音节速率之间的关系。首先，我们利用大型语料库重新评估总平均调制频谱的峰值频率是否与不同语言和说话风格中的平均音节速率相符。接下来，为检验峰值调制频率是否确实与音节速率相关，我们量化个体说话人之间峰值调制频率与音节速率的相关性。峰值调制频率和音节速率分别是从语音包络和音节序列中提取的汇总统计量。最后，我们直接考察时域语音包络和音节起始序列，并使用计算模型量化这些信号之间的相位锁定。

2. 语音调制与音节之间的频域关系

2.1. 方法

2.1.1. 语音语料库

分析纳入了七个语音语料库（表1），即 DARPA-TIMIT（Garofolo 等，1993年）、Chinese-TIMIT（Yuan 等，2017）、AISHELL-1（Bu 等，2017）、WenetSpeech（Zhang 等，2022年）、GigaSpeech（Chen 等，2021年）、TED-LIUM（Rousseau 等，2012年）和 Common Voice（Ardila 等，2020）。所有语料库均经过转录。这些语料库包含四种说话风格，即朗读句子、有声书朗读、公开演讲和访谈（表1）。所有公开演讲均由单人发表，而访谈则包含多位说话人之间的对话。对于访谈，录音是从 GigaSpeech 中的 YouTube 类别手动选取的，筛选标准是录音中不含背景音乐。对于包含多种语言的语料库 Common Voice，我们选取了说话人平均时长

超过 512 秒的六种语言，即泰语、法语、西班牙语、波兰语、俄语和德语。对于包含篇章级录音的语料库，我们根据语料库提供的句子边界将录音切分为句子。所有句子的时长均不超过 20 秒，并重采样至 16 千赫。语料库处理脚本可在 https://github.com/austin-365/ms-tools 获取。

2.1.2. 调制谱

语音包络和调制谱采用与 Ding 等人（2017）相同的方法提取，简要说明如下（图1）。利用听觉模型提取包络，该模型将声音信号滤波为128个按对数分布的频带，范围为180赫兹至7千赫兹。在每个频带中，对滤波后的声音信号进行半波整流、平滑处理，并降采样至200赫兹。听觉模型在每个频带的输出称为窄带包络，128个窄带包络的平均值称为宽带包络。窄带包络表征窄频带内的声功率波动，而宽带包络更能反映跨频带的相干功率波动。通过对语音包络的平方应用离散傅里叶变换（DFT），得到调制谱。窄带谱为128个窄带包络的调制频谱的均方根（RMS），宽带谱则为宽带包络的调制谱。对每个句子分别计算调制谱，并通过零填充将每个句子的包络调整至相同的时长，即20秒。说话人的峰值调制频率采用两种方法计算，依据的是单个句子的调制谱。方法1根据每个

句子计算峰值调制频率，然后对所有句子的峰值调制频率取平均。方法2首先计算所有句子的调制频谱的均方根（RMS），然后从均方根谱中提取峰值。

2.1.3. 音节速率

音节起始是自动提取的，并基于可进行人工标注的语料库对该方法进行了验证（见附录）。根据音频和转录文本，使用蒙特利尔强制对齐器（MFA）（McAuliffe 等，2017）提取音位之间的边界。对于中文，MFA 还直接提供了音节边界，因为音节直接对应于书写系统中的主要单位，即汉字。对于英语，MFA 提供了词语和音位之间的边界，随后根据词典（即 Unisyn词典）（Fitt，2001）将音位组合为音节，以确定音节边界。对于其他语言，由于音节结构多样，未提取音节边界。

音节速率采用三种指标进行表征。第一种指标是音节速率，即音节数除以录音时长；第二种指标是发音速率，即音节数除以音节的总时长。音节速率与发音速率的区别在于，前者会受到语音中静音时段的影响，而后者不会。第三种指标是音节众数，即单个音节时长倒数直方图的众数（Greenberg 等，2003），该指标采用高斯核密度法估算，窗口宽度则依据斯科特规则确定（Scott，2015）。

2.2. 结果

我们首先比较了语音调制谱的形状和音节时长分布。在此分析中，各项指标均在每个语料库内取平均值。平均而言，宽带和窄带调制谱的峰值分别出现在 3.9 赫兹（SD = 0.7 赫兹）和 4.8 赫兹（SD = 0.7 赫兹）。音节速率分布的模式，以及平均音节速率和平均发音速率，均与调制谱的峰值频率一致（图 2A）。不同音节速率指标与峰值调制频率之间的平均差异见图 3A，通常在 1 赫兹以内（M = 0.5 赫兹，SD= 0.3 赫兹）。发音速率与采用方法2计算的窄带调制谱峰值频率之间的差异最小。

接下来，我们考察了不同个体说话人的音节速率与峰值调制频率是否相关。不同测量指标之间的相关系数取各语料库的平均值，结果见图3B。发音速率与采用方法1计算的宽带调制谱峰值频率之间的相关性最高，图2B进一步展示了各个语料库的这一相关性。两个规模较小的语料库，即 TIMIT 和 Chinese-TIMIT（< 6小时录音）的相关性最低，而 AISHELL（179小时录音）的相关性最高。值得注意的是，Chinese-TIMIT 和 AISHELL 均包含中文朗读句子的录音，两者的主要区别在于每位说话人的录音时长。为检验录音时长是否会限制峰值调制频率与发音速率之间的相关性，我们基于录音子集计算了相关系数。结果发现，

随着个体说话人的录音时长增加，相关性也随之增强；这一结果在中文和英语中均成立（图4），在另外六种语言中也同样如此（图S5）。当每位说话人的录音时长为4秒时，峰值调制频率与发音速率之间的相关性较低（各语料库的平均值为$R = 0.25$，标准差为$= 0.05$）；而对于大多数语料库，渐近相关性则高于0.5（各语料库的平均值为$R = 0.61$，标准差为 = 0.11）。根据 Sigmoid 拟合，对于大多数语料库，当录音时长超过512秒时，相关系数达到最大值的95%。

3. 语音调制与音节之间的时域关系

峰值调制频率与音节速率之间的相关性可用于判断包络节律与音节节律是否具有相同的中心频率。然而，具有相同的频率范围并不能保证两个序列在时间上锁定。此外，即使两个信号不具有相同的中心频率，它们仍可能包含相位锁定的成分。因此，在下文中，我们使用两个预测模型，即时间响应函数（TRF）模型和长短期记忆网络（LSTM）（图5A），直接分析语音包络与音节起始之间的相位锁定。

3.1. 方法

3.1.1. 时间响应函数

使用 TRF 和 LSTM 模型表征语音包络与音节起始之间的相位锁定。TRF 使用线性时不变系统模型描述两个序列之间的关系（Crosse 等人，2016；Ding 和 Simon，2012）。简而言之，TRF 是一种

用于将一个信号转换为另一个信号的最优线性滤波器。音节起始由一个二进制序列编码，该序列仅在音节起始处取值为 1（采样率为 50 赫兹）；此外，为便于训练，宽带包络也被降采样至 50 赫兹。如果分别用 $\sigma(n)$ 和 包络(n) 表示音节起始序列和宽带包络，则二者的关系可用以下方程表示：

$$env(n) = \sum_{m} TRF(m)\sigma(n-m) + \varepsilon(n),$$

其中，$TRF(m)$ 和 $\varepsilon(n)$分别表示模型的 TRF 和残差。模型所考虑的时间滞后范围，即m，为 -0.5 s 至 0.5 s。采用带十折交叉验证的岭回归估计 TRF。TRF 的预测能力定义为预测包络与实际宽带包络之间的相关系数，并对 10 折的结果取平均。我们采用说话人无关的 TRF 模型，并基于每个语料库训练 TRF，将训练集中的所有句子串接起来。初步分析表明，说话人无关的 TRF 对语音包络的预测准确度与基于每位说话人训练的说话人相关时间响应函数相近（数据未显示）。

3.1.2. 长短期记忆

LSTM（Hochreiter 和 Schmidhuber，1997）是一种深度神经网络模型，可将输入序列转换为输出序列（见图 S1）。TRF 是一种时不变模型：每个音节起始都会触发相同的语音包络。然而，LSTM模型是上下文相关的：音节起始所触发的语音包络取决于上下文，即相邻音节的起始；上下文范围由模型自动确定。我们采用一个 8 层双向长短期记忆网络，将音节起始

序列转换为语音包络。我们基于每个语料库训练了一个说话人无关 LSTM 模型。由于训练 LSTM模型耗时较长，我们将每个语料库划分为训练集、验证集和测试集。训练集包含从语料库中随机选取的 80% 句子，验证集和测试集则各包含剩余 20% 句子中的一半。

3.1.3 统计分析

为了评估音节起始是否能够通过时间响应函数或 LSTM模型显著预测语音包络，我们采用置换策略估计机会水平预测能力。具体而言，我们将一个句子的音节起始序列与另一个句子的语音包络随机配对，并基于时间响应函数或 LSTM模型计算预测结果。此过程重复了 1000 次，得到 1000 个机会水平预测能力值。如果实际预测能力低于A 个机会水平预测结果中的A个，则显著性水平为（+1）/（1000+1）。

3.2. 结果

TRF在音节起始后 32 ± 8 ms的潜伏期处显示一个波谷（图5B），表明音节起始后宽带包络出现局部最小值。TRF的预测能力，即实际包络与根据音节起始预测的包络之间的相关系数，对于每个语料库通常介于0.25和0.35之间（表2，$p < 10^{-3}$，每个语料库均经假发现率校正）。预测能力的平方，即R^2, ，每位说话人对应的数值为10.6 ± 4.4%，可解释为音节起始所解释的包络方差百分比。对于上下文相关 LSTM 模型，其预测能力，即R，通常介于0.4和0.6之间

对于每个语料库（表2，$p < 10^{-3}$，每个语料库均经假发现率校正），约为TRF预测能力的1.5倍。预测能力的平方，即R^2, ，每位说话人对应的数值为24.2 ± 6.9%。此外，在两个较小语料库上训练的模型（即TRF和LSTM）能够很好地泛化到较大的语料库，甚至能跨语言泛化（表2，$p < 10^{-3}$，每个语料库均经假发现率校正）。最后，为了分析是否需要较长的录音时长才能揭示语音包络与音节起始之间的相位锁定，我们计算了两个模型在大型语料库子集上的预测能力。结果发现，预测能力并未随录音时长发生明显变化（见图S2）。这些结果表明，语音包络与音节起始之间的相位锁定不受录音时长影响，且两个模型均能泛化至不同风格和语言的语音。

4. 讨论

本研究考察了语音包络与音节边界之间的关系。与普遍认为峰值调制频率是音节速率的一种声学相关物（Greenberg 等，2003）的观点相反，研究表明，对于个体说话人而言，峰值调制频率与音节速率的相关性并不强，尤其是在每位说话人仅分析几秒录音时。因此，峰值调制频率不能作为判断单个句子音节速率的有效线索。相比之下，TRF/LSTM 分析显示，在时域中，宽带语音包络与单个音节的起始时点稳定锁相。因此，为了编码音节节律，大脑可能必须从语音包络中提取与音节相关的时间特征，而不是仅仅被动地跟随语音包络。

4.1. 语音调制谱

窄带调制谱已被广泛用于描述语音的声学节律，而且在不同语言和常见说话风格中形状相似，例如成人朗读和访谈（Ding 等人，2017；Varnet 等人，2017）。然而，包括婴儿导向语音（IDS）在内的一些更特殊的说话风格，可能会改变调制谱的形状。下文首先概述有关常见说话风格的既往研究，然后讨论有关更特殊说话风格的发现（图6）。英语电话语料库 Switchboard 的调制谱平均峰值频率为 4.5 赫兹（Greenberg 等，2003）；不同说话风格的英语语料库中，该值介于 4.4 至 4.8 赫兹之间（Ding 等人，2017）；九种语言的自然语篇层面语音中，该值介于 4.3 至 5.4 赫兹之间（Ding 等人，2017）；而十种语言的半朗读语音（SRS）语料库中，该值介于 4.3 至 5.5 赫兹之间（Varnet 等人，2017）。本研究中，各语料库窄带调制谱的峰值频率介于 4.2 至 5.9 赫兹之间，平均为 4.9 赫兹（表3）。宽带调制谱也呈现峰值，在六种语言中，其峰值频率介于 3.5 至 4.5 赫兹之间（Poeppel 和 Assaneo，2020）。本研究中，各语料库宽带调制谱的峰值频率介于 3.3 至 5.0 赫兹之间，平均为 4.3 赫兹（表3）。无论在本研究还是既往研究中，宽带调制谱的峰值频率都低于窄带调制谱的峰值频率（Poeppel 和 Assaneo，2020）。因此，无论是宽带还是窄带调制谱，其峰值频率在不同语言和说话风格中都高度一致，研究间的差异不超过 1.7 赫兹。

当说话人试图以特别清晰的方式说话时，语音的调制谱会发生变化。具体而言，在这类说话情境中，4 赫兹以下的低频时间调制都会增强，包括对婴儿（Leong 等，2017）、儿童（Leong 和 Goswami，2015；Pérez-Navarro 等，2022）或听力障碍者（Krause 和 Braida，2004）说话，以及在嘈杂环境中说话（Bosker 和 Cooke，2018）。换言之，当听者因听力状况受损或缺乏语言知识而可能难以理解语音时，说话人往往会调整说话风格，以增强低频调制；即使音节速率几乎没有变化，也能做到这一点（Bosker 和 Cooke，2018；Krause 和 Braida，2004；Pérez-Navarro 等，2022）。低频调制和音节速率变化最显著的情形是婴儿导向语音（图 6），因为此时听者的语音理解能力被认为最低。

4.2. 音节速率

我们还在图6中汇总了文献报道的音节速率。一项研究发现，八种语言的平均音节速率介于5.2至7.8赫兹之间（Pellegrino et al., 2011）；另一项研究发现，十七种语言的平均音节速率介于4.7至8.0赫兹之间（Coupé et al., 2019）。在本研究中，八种语言的音节速率介于3.5至6.1赫兹之间（表3）。同样，根据文献报道，发音速率随说话风格而有很大差异。英语中不同说话风格的发音速率介于3.1至5.6赫兹之间（Jacewicz et al., 2009），德语中则介于3.8至7.2赫兹之间（Jessen, 2007）。在本研究中，英语不同说话风格的发音速率介于4.3至5.2赫兹之间，中文则介于4.3至6.0赫兹之间（表3）。最后，音节众数此前仅有关于Switchboard的报告，即5.2赫兹

（Greenberg et al., 2003）；本研究中，不同语料库的音节众数介于3.5至5.3赫兹之间。总之，在不同语言和说话风格中，文献报道的音节速率介于3.5至8.0赫兹之间，发音速率则介于3.1至7.2赫兹之间。

文献中报告的结果基于不同的语料库，因此难以跨研究比较结果。例如，按定义，发音速率快于音节速率，而文献中报告的发音速率却往往低于音节速率。一个可能的原因是，报告较高音节速率的研究会要求说话人在熟悉材料后再朗读（Coupé et al., 2019）。考虑到说话风格可能产生的影响，语言之间的比较尤其困难。例如，一项研究发现，中文的发音速率高于英语（Ann Burchfield 和 Bradlow，2014年），而其他研究则报告了相反的结果（Coupé et al., 2019; Pellegrino et al., 2011）。本研究分析了两个中文朗读句子语料库和一个英文朗读句子语料库，结果发现，英语语料库的发音速率介于两个中文语料库的发音速率之间。因此，尽管据报告，不同语言的平均音节速率存在差异，但这些差异仍需通过规模更大的受控研究加以验证。

4.3. 调制谱与音节速率之间的关系

尽管人们通常认为调制谱反映了音节速率分布（Greenberg 等，2003），但这里表明，窄带调制谱的峰值频率与音节速率分布的模式仅呈弱相关（R 约为 0.3，图 3B）。因此，窄带

调制谱的峰值频率并非主要受音节速率影响，而可能源自人类发音器官的生物物理特性（Chandrasekaran 等人，2009）。宽带调制谱的峰值频率与每位说话人的音节速率相关性更强（图 3B），尤其是在分析每位说话人数分钟的录音时。结果表明，宽带包络反映了跨频带一致的强度波动（Poeppel 和 Assaneo，2020），因此相比各个窄频带内的包络，它更能指示音节节律。不过，如果只分析数秒的语音录音，宽带调制谱与音节速率的相关性也较弱。

当语音信号经过均匀的时间压缩以提高音节速率时，峰值调制频率会准确地跟踪音节速率（图 7A，左）。换言之，如果快速语音只是慢速语音的时间压缩版本，那么音节速率与峰值调制频率之间就会呈现完美相关性（$R = 1$）。那么，为什么真实句子中的相关性如此之低？一种可能是，语音中的低频时间调制实际上比接近音节速率的调制更强（Voss 和 Clarke，1975）。强烈的低频功率可能会使调制谱中的音节速率峰值在不同话语间高度变化，即使校正了 1/f 趋势也是如此。第二个原因是，语速改变时，不同语音片段并不会按相同比例缩放。例如，既往研究也已广泛记录到，语速提高会伴随元音弱化（Lindblom，1963；Taylor 等人，2014年）、辅音弱化（van Son 和 Pols，1999）、协同发音增加（Agwuele 等人，2008），以及辅音／元音比的轻微变化（Kessinger 和 Blumstein，1998 ）。

最后，另一种可能是语音内容也很重要，因为在本研究分析的语料库中，不同说话人的语音内容各不相同。为了检验语音内容是否确实能够解释为何单个句子的峰值调制频率与发音速率之间没有可靠相关性，我们进一步分析了 TIMIT语料库中所有说话人都说过的两个句子。然而，对于这两个句子，不同说话人的发音速率与峰值调制频率之间仍然只有较弱的相关性（图 7B 左侧面板，$R = 0$，两个句子分别为 0.28 和 0.11）。这一结果表明，即使语音内容保持不变，单个句子的发音速率与峰值调制谱之间仍然没有很强的相关性。不过，对这两个句子的分析进一步揭示了调制谱反映的内容，以及为何短语句的峰值调制频率与发音速率之间没有可靠相关性。在图 7 的中间面板中，我们展示了每位说话人的调制谱，并按照发音速率对说话人进行排序。图中还展示了快速、典型和慢速说话人的平均调制谱（图 7，右侧面板）。这些结果表明，两个句子的调制谱形状各不相同，而且这种差异在不同说话人中稳定存在。句子的调制谱有别于语料库平均调制谱，并且往往具有更多窄峰；这些窄峰是与语音内容相关的特征。此外，对于每个句子，说话人的调制谱都明显受到发音速率的影响。实质上，发音速率提高时，调制谱中的所有峰都会向更高的调制频率移动（图 7 中间列）。峰值调制频率与发音速率之间相关性不强的一个原因是，

句子的调制谱有多个谱峰，且最强峰因说话人而异。因此，尽管调制谱中的每个峰在一定程度上都反映了发音速率，但与最高峰相关的频率在不同说话人之间差异很大。总之，由于语音内容会显著影响句子的调制谱并产生多个谱峰，最高峰的频率与发音速率的相关性较弱。当对大量话语的调制谱取平均值时，无论这些话语来自同一说话人还是不同说话人，调制谱都会呈现出典型的钟形，而这个单一谱峰的频率与发音速率稳定相关。

4.4. 音节起始与语音包络之间的关系

时间响应函数分析显示，即使语音录音只有短短4秒，语音的宽带包络也会与单个音节的起始可靠地相位锁定（见图S2）。时间响应函数显示，包络往往在音节起始后约32毫秒处出现局部最小值，并在音节起始后约160毫秒处出现局部最大值。这些结果与先前的发现一致：可以从语音包络中提取音节边界（Hovsepyan 等人，2020；Hyafil 等人，2015；Villing 等人，2006）；一个音节通常在包络中有一个局部最大值，对应于音节核（Greenberg 等，2003；Hooper，1976；Oganian和Chang，2019），而音节边界则与包络中的波谷更为吻合（Mermelstein，1975）。然而，本研究通过量化音节起始能够解释语音包络方差的百分比，拓展了先前的研究。此外，先前利用语音包络提取音节边界的研究通常认为，只要提取出的边界与

人工标注的边界相差不超过50毫秒，结果便是正确的（Hovsepyan 等人，2020；Hyafil 等人，2015；Räsänen等人，2015；Villing 等人，2006）；而本研究则量化了包络波谷与音节起始之间的时间差。

TRF模型是一种时不变模型，它预测每个音节起始都会触发相同的语音包络，例如相同的时长、振幅和时间进程，而不受语音内容和语境影响。尽管这一假设显然过于简单，但根据TRF模型，单个音节的起始可以解释宽带语音包络约11%的方差，即R^2。此外，考虑语境信息时，也就是说，音节起始触发的语音包络会受到其他音节出现时间的影响，音节起始可以解释语音包络约24%的方差，即由长短期记忆网络模型得到的R^2, 。这一比例之所以较高，是因为该模型不考虑任何语音信息。此外，TRF和长短期记忆网络模型都能在英语和中文之间，以及不同说话风格的语音语料库之间良好泛化，这表明语音包络与音节起始之间的时域锁相可能是一种普遍特征。

4.5. 音节边界声学线索的意义

语音包络和调制谱都是语音的声学特征，而音节则是一个由语言学定义的单位。如何将语音的声学特征与语言单位联系起来，是心理语言学和神经语言学中最核心的问题之一（Poeppel等人，2008；Poeppel和Embick，2017）。音位的可靠声学特征难以识别（Liberman等人，1967），而文盲对音位缺乏意识（Morais等人，1979），这促使研究者探究其他语言单位是否具有更可靠的声学相关物。由于语音包络和

调制谱被认为能够为音节提供直接明了的声学相关物（Greenberg等人，2003），因此音节被视为连接听觉加工与语言加工的有效接口（Ghitza，2013；Giraud 和 Poeppel，2012）。

本研究证实，语音包络与音节起始锁相（Poeppel 和 Assaneo，2020），但也表明，语音包络中只有一个成分在追踪音节起始：根据长短期记忆网络模型，音节起始只能解释包络约四分之一的方差（R^2）。同样，调制谱会受到多种因素影响，包括语音和韵律特征，而且只有在对较长的话语进行平均时，才能可靠地追踪音节速率。这些结果表明，大脑仍需进行额外运算，才能根据语音包络提取音节边界。另一个相关之处在于，当听者理解语音的效率可能较低时，说话人往往会调整说话风格（见第4.1节）。在此类情况下产生的语音（包括婴儿/儿童导向语音）可能具有更明确的语音包络与音节边界之间的对应关系，使听者更容易将语音的声学特征映射到语言单位上（Goswami，2019）。此外，近期有研究指出，若干临床人群在语音包络的皮层编码方面存在缺陷（综述见 Palana 等，2021）；这些人群包括患有阅读障碍的儿童和成人，以及亚临床阅读困难者（Di Liberto 等，2018；Lizarazu 等，2021；Molinaro 等，2016）。对于这些人群，语音包络的神经编码受损可能会使音节等语言单位更难提取。未来研究有必要检验，这些人群是否能从语音包络与音节之间对应关系更明确的语音中获益。

最后，尽管本研究关注语音包络与音节之间的关系，包络（尤其是低于4赫兹的时间调制）也能为其他层级的语言单位提供有用线索，例如重读音节（Greenberg 等，2003）和语调短语（Inbar 等，2020）。尤其对于面向儿童的语音，研究发现，约2赫兹的极低频调制可为韵律重音提供线索，而约4赫兹的调制则可为音节提供线索（Leong 和 Goswami，2015）。因此，时间调制的层级可能对应于语音韵律的层级；有假设认为，这种声学信息与语言信息之间的映射对于促进语言习得起着关键作用（Goswami，2019）。未来研究有必要量化语音包络与完整韵律单位层级之间的关系，并考察这种关系在婴儿导向、儿童导向和成人指向语音之间是否有所不同。

总之，语音包络中的一个成分与音节起始锁相。然而，这一成分并不总是语音中最强的成分，因此，对于通常持续数秒的单个句子，峰值调制频率与音节速率的相关性较弱。不过，当对数分钟的语音录音进行调制谱平均时，单个句子语音内容相关的特征会变得模糊，峰值调制频率则能很好地反映平均音节速率。

致谢

本研究部分受到国家自然科学基金委员会（32222035）和浙江省重点研发项目（编号：2022C03011）的资助。

参考文献

Agwuele, A., Sussman, H.M., Lindblom, B., 2008. 语速对辅音-元音协同发音的影响。Phonetica 65, 194–209. https://doi.org/10.1159/000192792 Ann Burchfield, L., Bradlow, A.R., 2014年。普通话和英语语音中的音节弱化。美国声学学会期刊 135, EL270–EL276. https://doi.org/10.1121/1.4874357 Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L

.,

Tyers, F., Weber, G., 2020. Common Voice：一个大型多语言语音语料库。发表于《第12届语言资源与评测会议》（LREC 2020）论文集，第4218–4222页。Assaneo, M.F., Poeppel, D., 2018. 听觉皮层与运动皮层之间的耦合受速率限制：内在语音运动节律的证据。《科学进展》4，eaao3842。https://doi.org/10.1126/sciadv.aao3842 Boersma, P., Weenink, D., 2016. Praat：用计算机进行语音学研究 [计算机程序]。Bosker, H.R., Cooke, M., 2018. 在噪声环境中说话时，说话者会产生更显著的振幅调制。美国声学学会期刊 143，EL121–EL126。https://doi.org/10.1121/1.5024404 Brodbeck, C., Presacco, A., Simon, J.Z., 2018. 连续刺激引发的脑反应的神经源动态：从声学到理解的语音加工。《神经影像》172, 162–174。https://doi.org/10.1016/j.neuroimage.2018.01.042 Brodbeck, C., Simon, J.Z., 2020. 连续语音处理。Curr. Opin. Physiol. 18，25–31。https://doi.org/10.1016/j.cophys.2020.07.014 Bu, H., Du, J., Na, X., Wu, B., Zheng, H., 2017. AISHELL-1：一个开源普通话语音语料库及语音识别基线。发表于国际语音数据库与语音输入／输出系统及评估协调委员会东方分会第20届会议（O-COCOSDA 2017），IEEE，首尔，第58–62页。https://doi.org/10.1109/ICSDA.2017.8384449 Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. 视听语音的自然统计特性。PLoS Comput. Biol. 5，e1000436。https://doi.org/10.1371/journal.pcbi.1000436 Chen, G., Chai, S., Wang, G.-B., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., Jin, M., Khudanpur, S., Watanabe, S., Zhao, S., Zou, W., Li, X., Yao, X.,

Wang, Y., You, Z., Yan, Z., 2021. GigaSpeech：一个不断演进的多领域自动语音识别语料库，包含10,000小时的转录音频。发表于 Interspeech 2021，国际语音通信协会，第3670–3674页。https://doi.org/10.21437/Interspeech.2021-1965 Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. 频谱-时间调制传递

r

功能与语音可懂度。《美国声学学会期刊》106，2719–2732。Chi, T., Ru, P., Shamma, S.A., 2005. 复杂声音的多分辨率时频分析。《美国声学学会期刊》118，887–906。https://doi.org/10.1121/1.1945807 Coupé, C., Oh, Y.M., Dediu, D., Pellegrino, F., 2019. 不同语言，相似的编码效率：人类交际生态位中可比的信息速率。《科学进展》5，eaaw2594。https://doi.org/10.1126/sciadv.aaw2594 Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C., 2016. 多变量时间响应函数（MTRF）工具箱：一种用于关联神经信号与连续刺激的 MATLAB 工具箱。《人类神经科学前沿》10，604。https://doi.org/10.3389/fnhum.2016.00604 Daube, C., Ince, R.A.A., Gross, J., 2019. 简单的声学特征可以解释基于音位的语音皮层反应预测。《当代生物学》29，1924–1937。https://doi.org/10.1016/j.cub.2019.04.067 Di Liberto, G.M., Hjortkjær, J., Mesgarani, N., 2022. 编者按：神经追踪——缩小神经生理学与转化医学之间的差距。《神经科学前沿》16，872600。https://doi.org/10.3389/fnins.2022.872600 Di Liberto, G.M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., Lalor, E.C., 2018. 右半球对语音的非典型皮层同步是阅读障碍音位缺陷的根源。《神经影像》175，70–79。https://doi.org/10.1016/j.neuroimage.2018.03.072 Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. 连贯语音中层级语言结构的皮层追踪。《自然神经科学》19，158–164。https://doi.org/10.1038/nn.4186 Ding, N., Patel, A.D., Chen, L., Butler,H., Luo, C., Poeppel, D., 2017. 时间调制 i

n

语音与音乐。《神经科学与生物行为评论》81，181–187。https://doi.org/10.1016/j.neubiorev.2017.02.011

Ding, N., Simon, J.Z., 2014. 连续语音的皮层同步：功能作用与诠释。《Front. Hum. Neurosci.》8，311。https://doi.org/10.3389/fnhum.2014.00311 Ding, N., Simon, J.Z., 2012. 聆听相互竞争的说话者时，听觉对象神经编码的出现。《Proc. Natl. Acad.Sci.》109，11854–11859。https://doi.org/10.1073/pnas.1205381109 Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. 声学地标驱动δ-θ振荡，通过促进知觉解析实现语音理解。《神经影像》85，761–768。https://doi.org/10.1016/j.neuroimage.2013.06.035 Elhilali, M., Chi, T., Shamma, S.A., 2003. 用于评估语音可懂度的谱时调制指数（STMI）。《语音通信》41，331–348。https://doi.org/10.1016/S0167-6393(02)00134-6 Elliott, T.M., Theunissen, F.E., 2009. 语音可懂度的调制传递函数。《PLoS Comput. Biol.》5，e1000302。https://doi.org/10.1371/journal.pcbi.1000302 Fitt, S., 2001. Unisyn词典发布版（1.3版）。Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. DARPA TIMIT：声学语音连续语音语料库 CD-ROM，NIST语音光盘 1-1.1（NIST IR 4930号）。美国国家标准与技术研究院，马里兰州盖瑟斯堡。https://doi.org/10.6028/NIST.IR.4930 Ghitza, O., 2013. θ音节：由皮层功能定义的语音信息单位。《Front. Psychol.》4，138。https://doi.org/10.3389/fpsyg.2013.00138 Giraud, A.-L., Poeppel, D., 2012. 皮层振荡与语音加工：新兴的计算原理与运算。《Nat. Neurosci.》15，511–517。https://doi.org/10.1038/nn.3063 Goswami, U., 2019. 语音节律与语言习得：振幅调制相位层级视角。《Ann. N. Y. Acad. Sci.》1453，67–78。https://doi.org/10.1111/nyas.14137 Greenberg, S., 1999. 简略表达——理解发音变异的以音节为中心的视角。《语音通信》29，159–176。https://doi.org/10.1016/S0167- 6393(99)00050-3

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2003年。自发语音的时间特性——以音节为中心的视角。《语音学杂志》31，465–485。https://doi.org/10.1016/j.wocn.2003.09.005 Hochreiter, S., Schmidhuber, J., 1997。长短期记忆。《神经计算》9，1735–1780。https://doi.org/10.1162/neco.1997.9.8.1735 Hooper, J.B., 1976。自然生成音系学导论。《语言》54，667–674。https://doi.org/10.2307/412793 Hovsepyan, S., Olasagasti, I., Giraud, A.-L., 2020。结合预测编码与神经振荡，可实现对自然语音的在线音节识别。《自然·通讯》11，311

https://doi.org/10.1038/s41467-020-16956-5 Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., Giraud, A.-L., 2015. 由耦合的皮层θ和γ振荡实现语音编码。eLife 4, e06213. https://doi.org/10.7554/eLife.06213 Inbar, M., Grossman, E., Landau, A.N., 2020. 语调单位序列形成一种 ~1 赫兹节律。Sci. Rep. 10, 15846. https://doi.org/10.1038/s41598-020-72739-4 Jacewicz, E., Fox, R.A., O'Neill, C., Salmons, J., 2009. 不同方言、年龄和性别群体的发音速率。Lang. Var. Change 21, 233–256. https://doi.org/10.1017/S0954394509990093 Jessen, M., 2007. 德语发音速率的法证参考数据。Sci. Justice 47, 50–67. https://doi.org/10.1016/j.scijus.2007.03.003 Kessinger, R.H., Blumstein, S.E., 1998. 语速对起声时间和元音发音的影响：对知觉研究的一些启示。J. Phon. 26, 117–128. https://doi.org/10.1006/jpho.1997.0069 Krause, J.C., Braida, L.D., 2004. 正常语速下自然产生的清晰语音的声学特性。美国声学学会期刊 115, 362–378. https://doi.org/10.1121/1.1635842 Lalor, E.C., Foxe, J.J., 2010. 可通过精确的时间分辨率提取对连续自然语音的神经反应。Eur. J. Neurosci.31, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x Leong, V., Goswami, U., 2015. 面向儿童的语音振幅包络中的声学涌现音系学。PLoS ONE 10, e0144411. https://doi.org/10.1371/journal.pone.0144411

Leong, V., Kalashnikova, M., Burnham, D., Goswami, U., 2017. 婴儿导向语音的时间调制结构。Open Mind 1, 78–90. https://doi.org/10.1162/OPMI_a_00008Li, W., Yang, Y., 2009. 普通话句子中韵律层级边界的知觉。Neuroscience 158, 1416–1425. https://doi.org/10.1016/j.neuroscience.2008.10.065 Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. 语音编码的知觉。Psychol. Rev.74, 431. Liberman, I.Y., Shankweiler, D., Fischer, F.W., Carter, B., 1974. 幼儿对音节和音位的显性切分。J. Exp. Child Psychol. 18, 201–212. https://doi.org/10.1016/0022-0965(74)90101-5 Lindblom, B., 1963. 元音弱化的声谱分析研究。美国声学学会期刊 35, 1773– 1781. https://doi.org/10.1121/1.2142410 Lizarazu, M., Lallier, M., Bourguignon, M., Carreiras, M., Molinaro, N., 2021. 阅读障碍者对语音边界的神经反应受损。Cortex 135, 207–218. https://doi.org/10.1016/j.cortex.2020.09.033 McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M., 2017. 蒙特利尔强制对齐器：使用 kaldi 进行可训练的文本-语音对齐。发表于 Interspeech 2017，国际语音通信协会，页 498–502。https://doi.org/10.21437/Interspeech.2017-1386 Mehler, J., Dommergues, J.Y., Frauenfelder, U., Segui, J., 1981. 音节在语音切分中的作用。J. Verbal Learn. Verbal Behav. 20, 298–305. https://doi.org/10.1016/S0022-5371(81)90450-3 Mermelstein, P., 1975. 将语音自动切分为音节单位。J. Acoust. Soc. A

m.

58, 880–883. https://doi.org/10.1121/1.380738 Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., Carreiras, M., 2016. 发展性阅读障碍中的语音同步 entrainment 异常：阅读障碍中的皮层语音追踪改变。Hum. Brain Mapp. 37, 2767–2783. https://doi.org/10.1002/hbm.23206 Morais, J., Cary, L., Alegria, J., Bertelson, P., 1979. 语音作为音段序列的意识会自发产生吗？Cognition 7, 323–331. https://doi.org/10.1016/0010- 0277(79)90020-9

Oganian, Y., Chang, E.F., 2019. 人类颞上回中用于音节编码的语音包络标志。Sci. Adv. 14. Palana, J., Schwartz, S., Tager-Flusberg, H., 2021. 评估利用皮层同步测量非典型语音处理的做法：系统综述。Neurosci. Biobehav.Rev. 133, 104506. https://doi.org/10.1016/j.neubiorev.2021.12.029 Pellegrino, F., Coupé, C., Marsico, E., 2011. 跨语言视角下的语音信息率。《语言》87, 539–558. https://doi.org/10.1353/lan.2011.0057 Pérez-Navarro, J., Lallier, M., Clark, C., Flanagan, S., Goswami, U., 2022. 西班牙语面向儿童的语音中的局部时间规律。《言语、语言与听力研究杂志》65, 3776– 3788. https://doi.org/10.1044/2022_JSLHR-22-00111 Poeppel, D., Assaneo, M.F., 2020. 语音节律及其神经基础。《自然综述：神经科学》21, 322–334. https://doi.org/10.1038/s41583-020-0304-4 Poeppel, D., Embick, D., 2017. 论语言学与神经科学之间关系的界定，载于：《二十一世纪心理语言学：四大基石》。Routledge, pp. 103–118. Poeppel, D., Idsardi, W.J., van Wassenhove, V., 2008. 神经生物学与语言学交界处的语音感知。《英国皇家学会哲学汇刊B辑：生物科学》363, 1071–1086. https://doi.org/10.1098/rstb.2007.2160 Räsänen, O., Doyle, G., Frank, M.C., 2015. 利用自动切分为类音节单位，从语音中无监督地发现词语。发表于 Interspeech 2015, 国际语音通信协会, pp. 3204–3208. https://doi.org/10.21437/Interspeech.2015-645 Rosen, S., 1992. 语音中的时间信息：声学、听觉与语言学方面。《伦敦皇家学会哲学汇刊B辑：生物科学》336, 367–373. https://doi.org/10.1098/rstb.1992.0070 Rousseau, A., Deléglise, P., Estève, Y., 2012. TED-LIUM：一个专用于自动语音识别的语料库。发表于第八届语言资源与评估国际会议论文集（LREC'12）, pp. 125–129. Scott, D., 2015. 多变量密度估计：理论、实践与可视化（第2004,16号）。柏林洪堡大学，应用统计与经济学中心（CASE），柏林。Shamma, S., 2001. 论空间与时间在听觉加工中的作用。《认知科学趋势》5, 340–348. https://doi.org/10.1016/S1364-6613(00)01704-6

Stevens, K.N., 2002. 基于声学地标和区别性特征的词汇通达模型探索。美国声学学会期刊 111, 1872–1891. https://doi.org/10.1121/1.1458026 Taylor, S., Theobald, B.-J.,Matthews, I., 2014年. 语速对听觉和视觉语音的影响。发表于 IEEE声学、语音与信号处理国际会议（ICASSP 2014），IEEE，佛罗伦萨，意大利，pp. 3037–3041. https://doi.org/10.1109/ICASSP.2014.6854158 van Son, R.J.J.H., Pols, L.C.W., 1999. 辅音弱化的声学描述。《语音通信》28, 125–140. https://doi.org/10.1016/S0167-6393(99)00009-6 Varnet, L., Ortiz-Barajas, M.C., Erra, R.G., Gervain, J., Lorenzi, C., 2017. 语音调制谱的跨语言研究。美国声学学会期刊 142, 1976–1989. https://doi.org/10.1121/1.5006179 Villing, R., Ward, T., Timoney, J., 2006. 基于包络的自动音节切分的性能上限。发表于 IET爱尔兰信号与系统会议（ISSC 2006），IEE，都柏林，爱尔兰，pp. 521–526. https://doi.org/10.1049/cp:20060489 Voss, R.F., Clarke, J., 1975. 音乐和语音中的“1/f噪声”。《自然》258, 317–318. https://doi.org/10.1038/258317a0

Yuan, J., Ding, H., Liao, S., Zhan, Y., Liberman, M., 2017. 中文TIMIT：标准中文的类TIMIT语料库。发表于国际语音数据库与语音输入/输出系统及评估协调委员会东方分会第20届会议（O-COCOSDA 2017），IEEE，首尔，第1–5页。https://doi.org/10.1109/ICSDA.2017.8384463

Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., Xu, X., Bu, H., Chen, X., Zeng, C., Wu, D., Peng, Z., 2022. WENETSPEECH：一个包含 10000+ 小时的多领域普通话语音识别语料库。发表于 IEEE声学、语音与信号处理国际会议（ICASSP 2022），IEEE，新加坡，新加坡，第 6182–6186 页。https://doi.org/10.1109/ICASSP43922.2022.9746682

图注

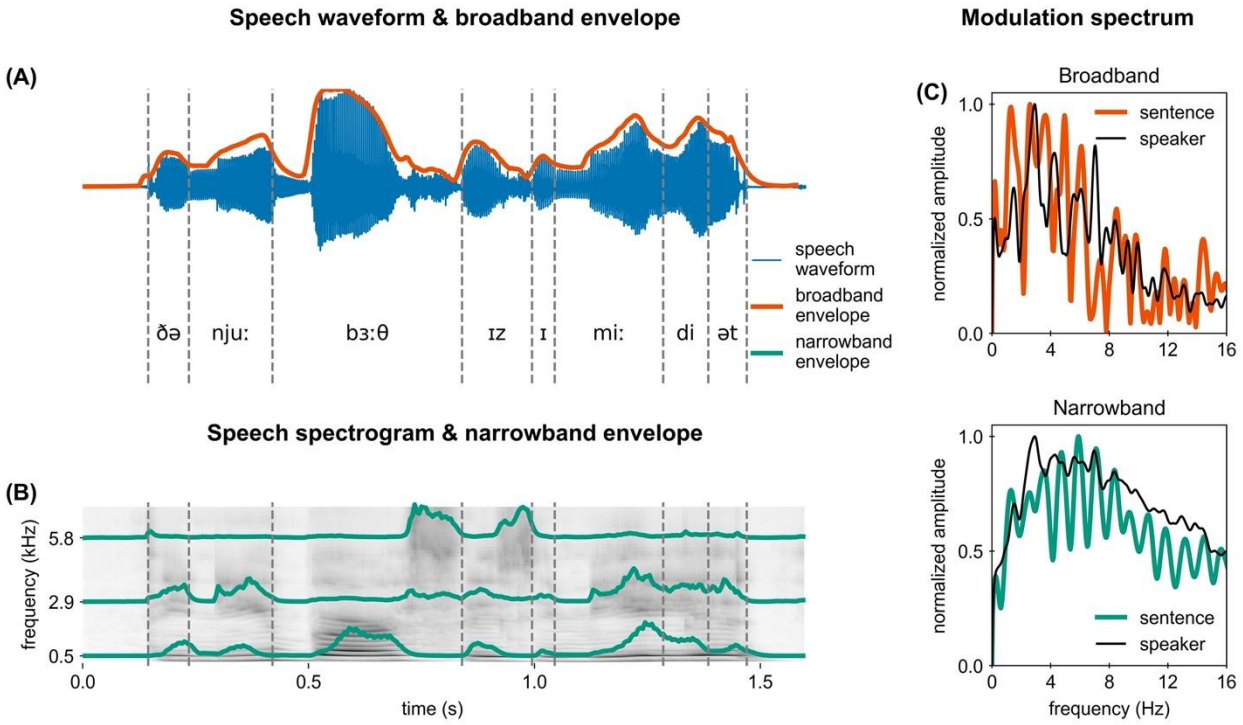

图 1. 语音包络和调制谱。(A) 语音波形和宽带

来自一位说话人的句子（“新生即刻降临”）的包络。虚线标出音节边界。（B）语音频谱图和窄带包络。频谱图中的每一行都构成一个窄带包络。（C）归一化调制谱。宽带和窄带调制谱分别基于宽带和窄带包络计算。彩色线和黑线分别基于该句子和该说话人的所有句子计算。

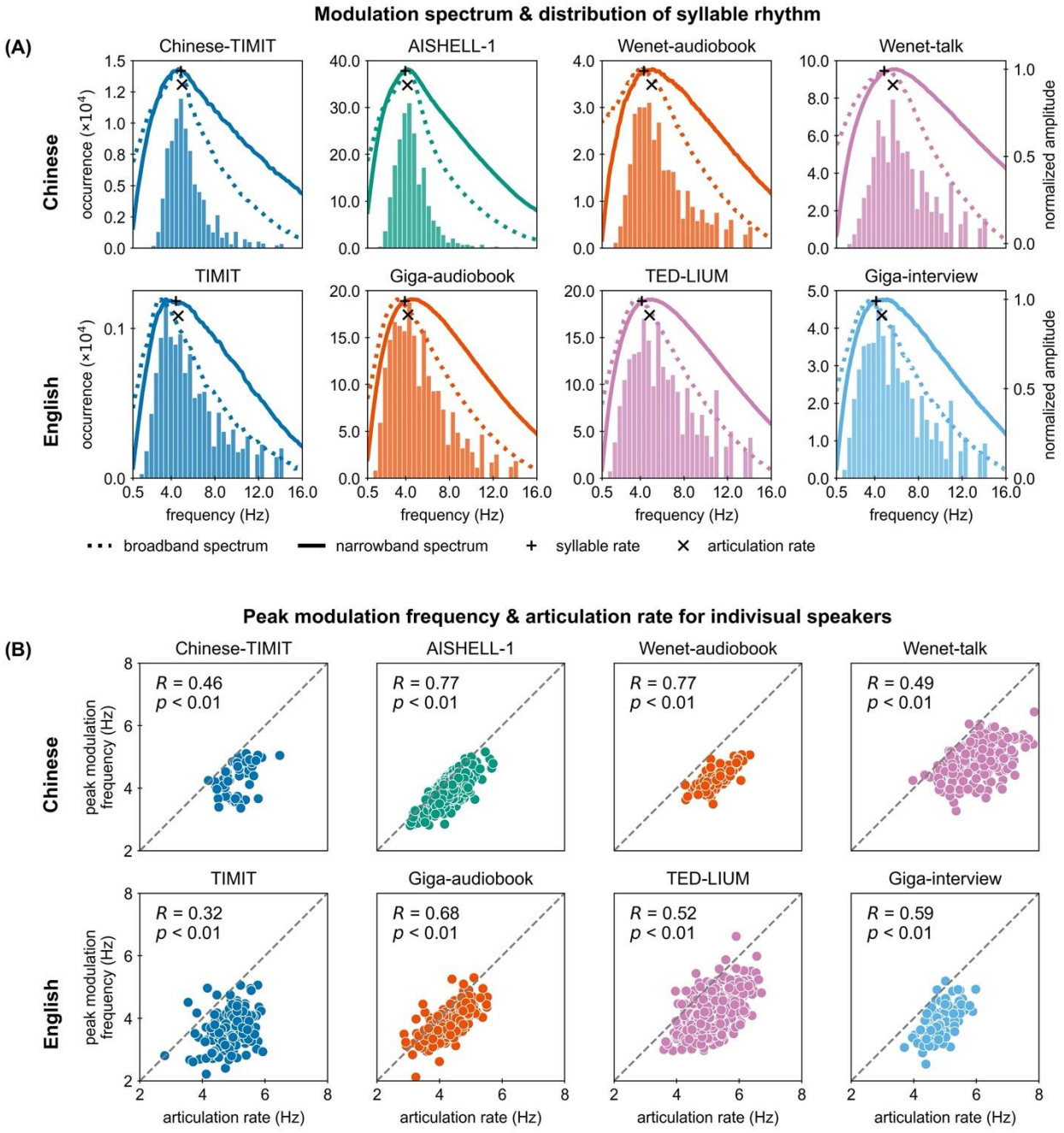

图 2。峰值调　　制频率、音节速率及其相关性。(A) 宽带

并展示每个语音语料库的窄带调制谱，以及音节时长倒数的直方图。黑色标记表示平均音节率和平均发音速率。音节众数是直方图的峰值。对于每种语言，每个语料库都用一种独特的颜色表示，并且在所有图中保持一致。（B）

个体说话人的峰值调制频率与发音速率之间的相关性（*R*）。使用方法1从宽带调制谱中提取峰值调制频率。每个点代表一位个体说话人。

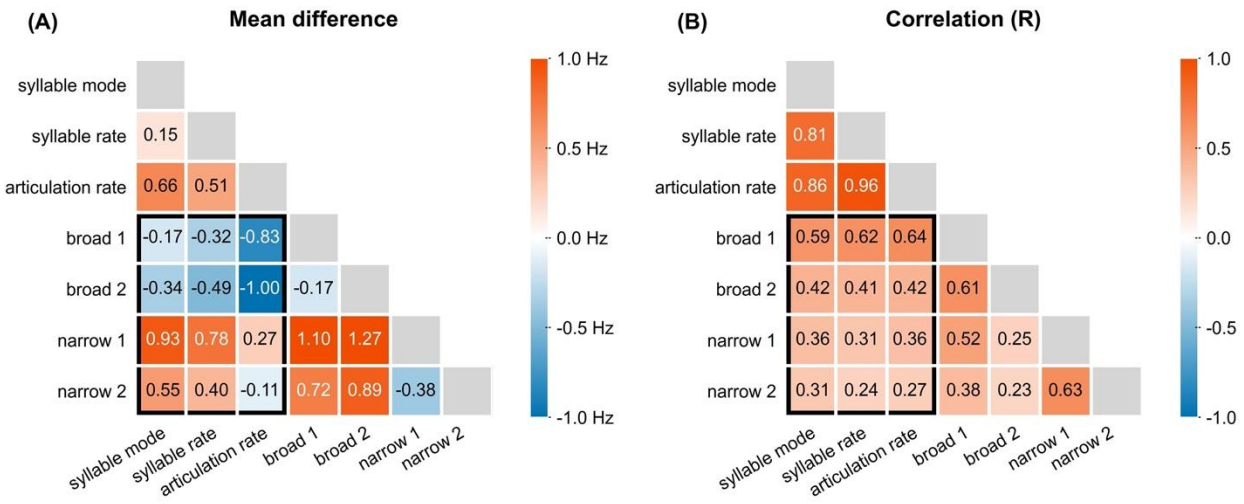

图3. 描述语速的不同测量值之间的平均差异和相关性

音节与峰值调制频率。(A) 对各语料库中的每项测量取平均值，并显示各测量之间的差异。(B) 在每个语料库内计算不同测量之间的成对相关性（*R*），并以语料库的总时长为权重，对各语料库的相关系数取加权平均。黑框突出显示了音节速率测量与峰值调制频率测量之间的关系。

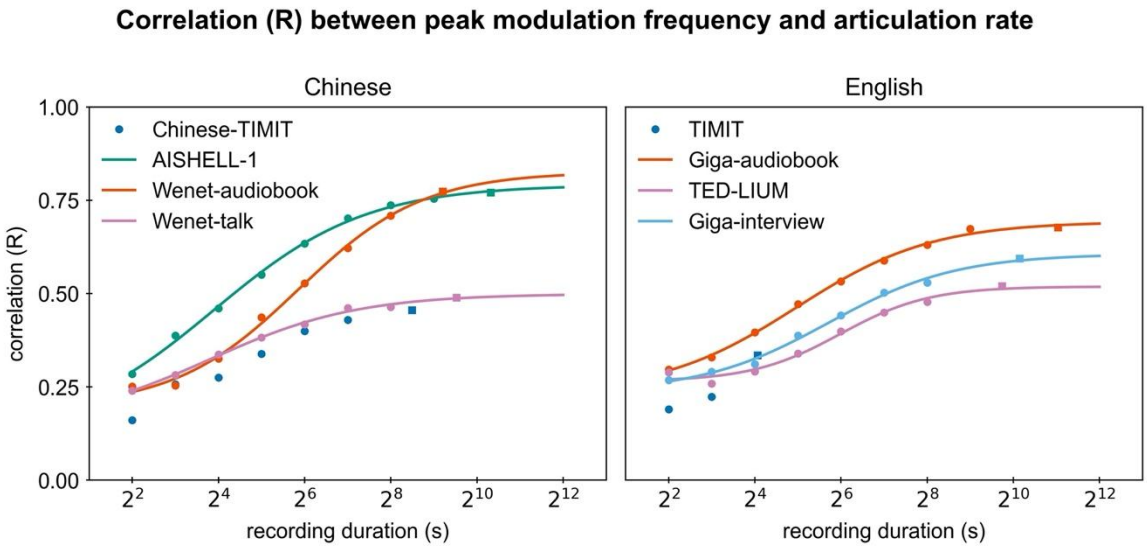

图 4. 宽带中的峰值调制频率与发音速率之间的相关性（R）

作为录音时长的函数。录音时长是每位说话人可用语音录音的时长。圆点和方块标记分别表示基于语音语料库子集或整个语料库计算出的相关性。对每个大型语料库的结果采用 S 形函数拟合。

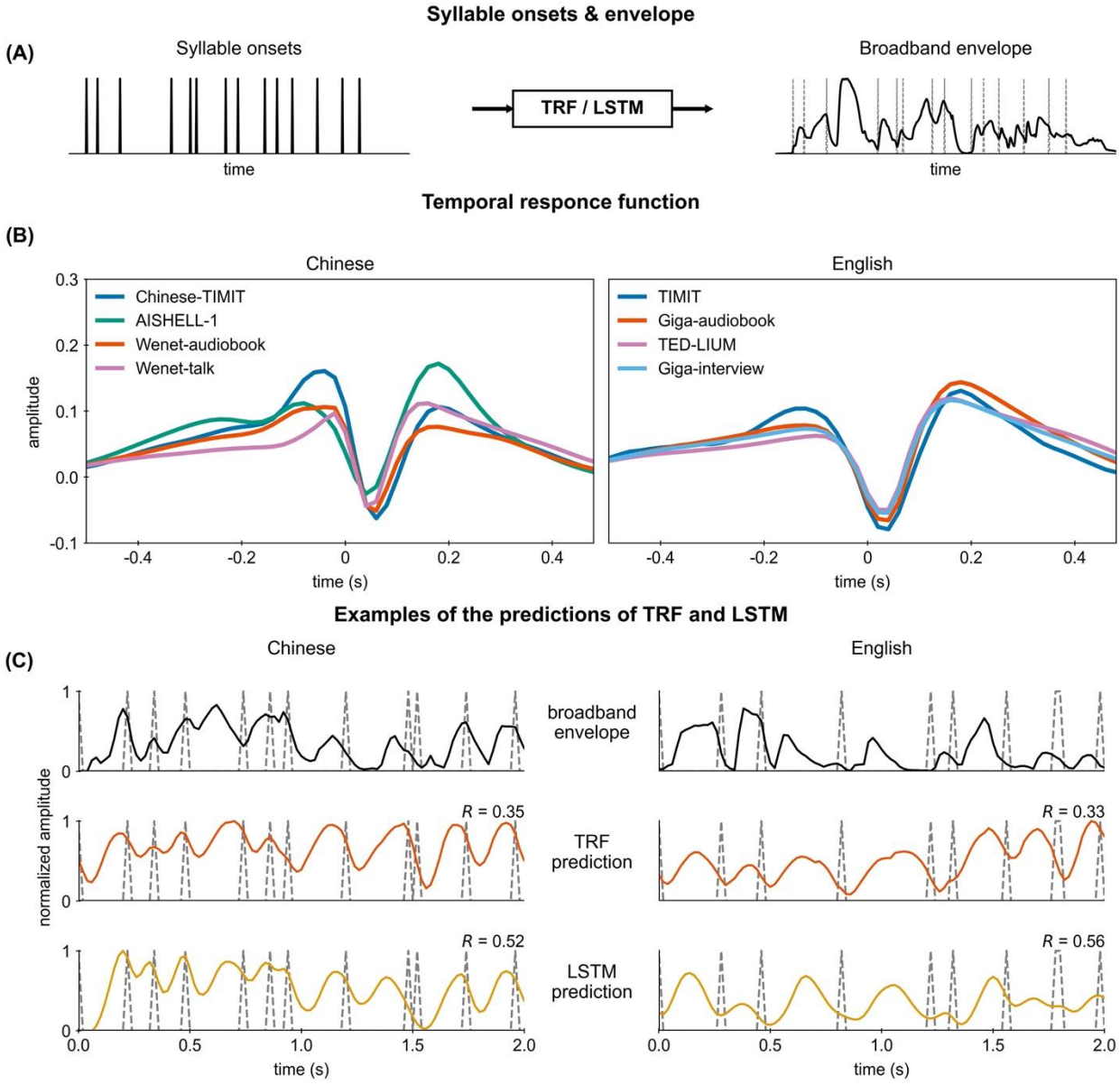

图 5。 基于音节起始预测语音包络。（A）音节起 t

始序列通过时间响应函数或 LSTM模型转换为宽带包络。（B）各语料库的时间响应函数以不同颜色标示。（C）时间响应函数和 LSTM模型预测的宽带包络示例。音节起始以灰色虚线标出。所选句子的预测能力与整个语料库的平均预测能力大致一致。

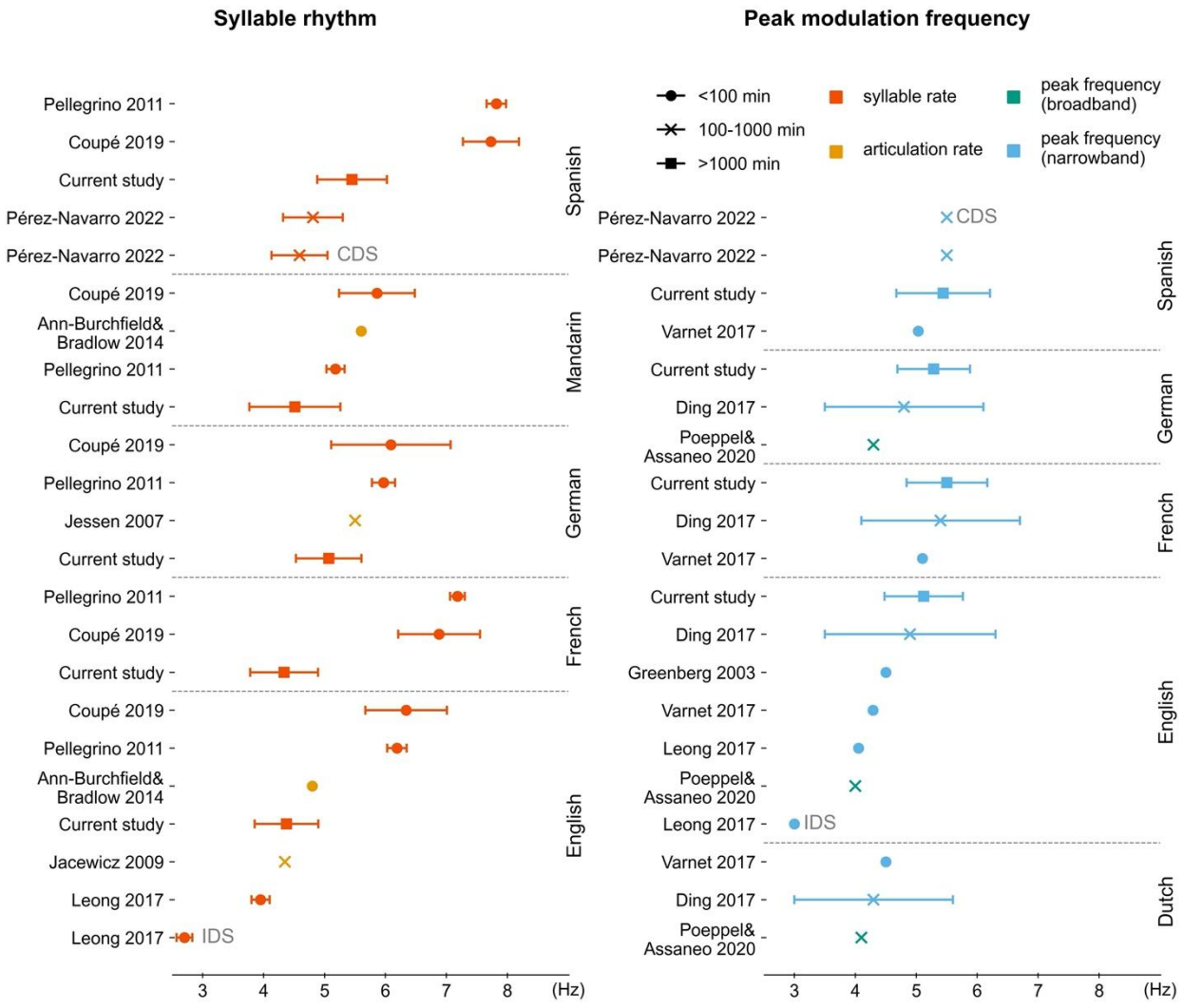

图 6. 不同研究中音节速率与峰值调制频率的比较。

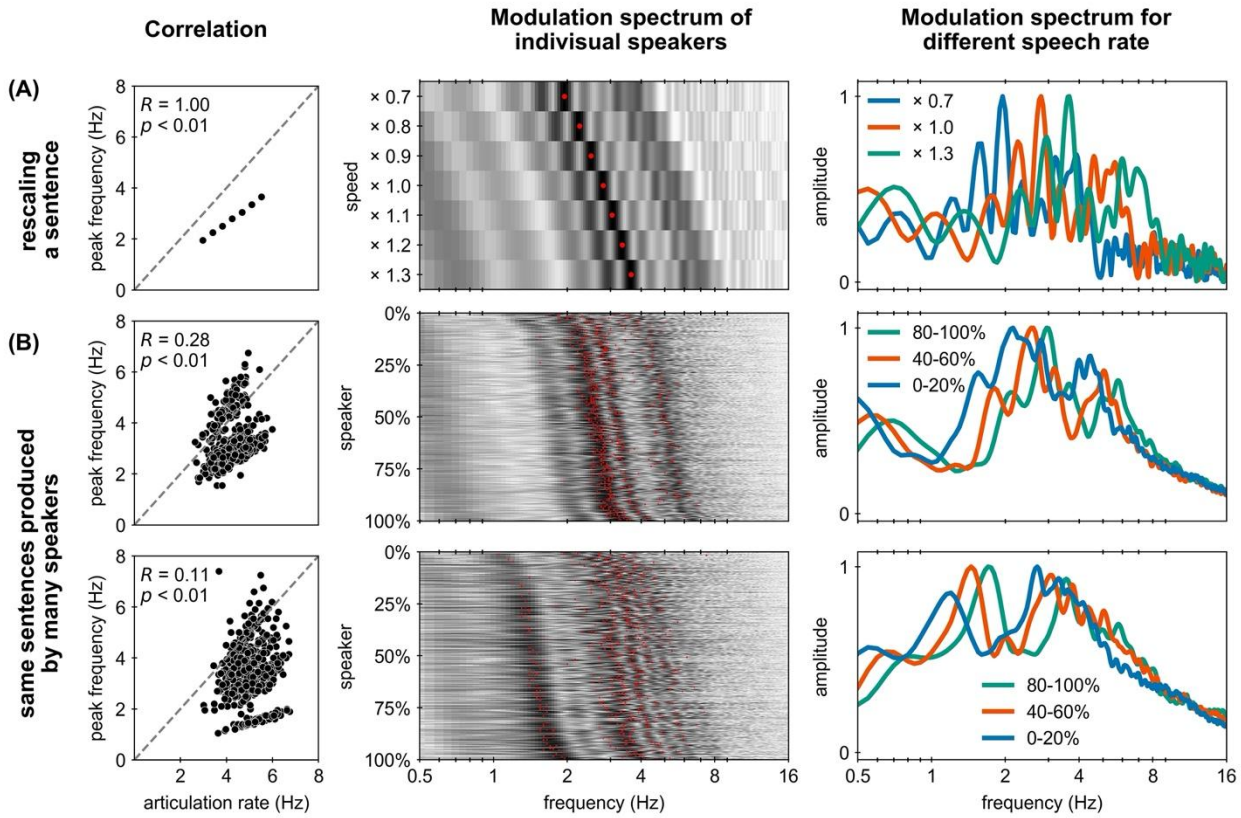

图 7. 示例句子的发音速率与调制频谱之间的关系。

A 面板分析了 TIMIT 中由单个说话人说出的一个句子（句子 ID：SA1，说话人 ID：MJAC0）。使用 PSOLA算法（Boersma和Weenink，2016）对该句子进行人为的时间压缩或扩展。B 面板分析了 TIMIT 中由许多说话人说出的两个句子（句子 ID：SA1 和 SA2）。左侧面板显示发音速率与峰值调制频率之间的相关性。中间面板将每位说话人的宽带调制谱依次排列，并按句子的发音速率对说话人排序。每位说话人的峰值调制频率以红点标出。右侧面板显示具有代表性的宽带调制谱（A），或不同说话人组的平均调制谱（B）。

表格

表1

语音语料库。

Name	来源	《语言》	说话风格	# 句子	句子时长（秒，均值 ± 标准差）	# 说话人	说话人时长（秒，均值 ±标准差）	总时长 (h)
Chinese-TIMIT	Chinese-TIMIT	中文	朗读句子	6 k	3.5 ± 0.6	50	415.1 ± 32.9	5.8
AISHELL-1	AISHELL-1	中文	朗读句子	142 千	4.6 ± 1.4	400	1610.6 ± 181.2	179
Wenet 有声读物	WenetSpeech	中文	有声书朗读	27 k	3.2 ± 2	135	636.6 ± 201.5	23.9
Wenet 对话	WenetSpeech	中文	Talk	81 k	2.4 ± 1.1	238	820.7 ± 254.9	54.3
TIMIT	DARPA-TIMIT	英语	朗读句子	1 k	2.9 ± 0.8	178	19.1 ± 2.5	0.9
Giga有声书	GigaSpeech	英语	有声书朗读	159 k	3.8 ± 2.3	264	2303.8 ± 931.6	168.9
TED-LIUM	TED-LIUM	英语	Talk	99 k	6.3 ± 2.9	646	963.6 ± 157.7	172.9
Giga-interview	GigaSpeech	英语	访谈	31 k	4.7 ± 3.4	121	1214.6 ± 736.2	40.8
CV-Thai	Common Voice	Thai	朗读句子	11 k	4.9 ± 1.7	38	1434.5 ± 3024.1	15.1
CV-法语	Common Voice	法语	朗读句子	76 k	4.8 ± 1.5	307	1188.8 ± 442.2	101.4
CV-西班牙语	Common Voice	西班牙语	朗读句子	32 k	5.2 ± 1.5	174	961.1 ± 466.1	46.5
CV-波兰语	Common Voice	波兰语	朗读句子	70 k	4.3 ± 1.5	190	1593.2 ± 3100.6	84.1
CV-俄语	Common Voice	俄语	朗读句子	77 k	5.3 ± 1.7	226	1809.9 ± 4012.6	113.6
CV-German	Common Voice	德语	朗读句子	54 k	5.1 ± 1.6	400	693.9 ± 146	77.1

表2TRF或长短期记忆网络的预测能力（M ± SD）。

	TRF			LSTM		
用于评估	**训练于自身**	**训练于 TIMIT**	**在……上训练 Chinese-TIMIT**	**在……上训练 自身**	**在……上训练 TIMIT**	**训练于 Chinese-TIMIT**
Chinese-TIMIT	0.39 ± 0.04	0.29 ± 0.05	0.39 ± 0.04	0.54 ± 0.03	0.37 ± 0.05	0.54 ± 0.03
AISHELL-1	0.39 ± 0.04	0.37 ± 0.04	0.36 ± 0.04	0.54 ± 0.04	0.42 ± 0.04	0.47 ± 0.04
Wenet 有声读物	0.32 ± 0.05	0.27 ± 0.03	0.32 ± 0.05	0.54 ± 0.05	0.27 ± 0.06	0.44 ± 0.05
Wenet 对话	0.3 ± 0.05	0.25 ± 0.05	0.28 ± 0.05	0.45 ± 0.05	0.29 ± 0.04	0.39 ± 0.05
TIMIT	0.36 ± 0.07	0.36 ± 0.07	0.27 ± 0.06	0.58 ± 0.07	0.58 ± 0.07	0.33 ± 0.07
Giga有声书	0.32 ± 0.04	0.33 ± 0.04	0.22 ± 0.04	0.48 ± 0.04	0.43 ± 0.04	0.27 ± 0.06
TED-LIUM	0.28 ± 0.05	0.28 ± 0.05	0.19 ± 0.05	0.44 ± 0.05	0.3 ± 0.04	0.21 ± 0.05
Giga-interview	0.24 ± 0.05	0.25 ± 0.04	0.17 ± 0.04	0.43 ± 0.04	0.33 ± 0.04	0.22 ± 0.04

表3描述音节速率或调制速率的测量值（M ±SD，赫兹）。

语料库	**音节 mode**	**音节速率**	**发音速率**	**宽式 1**	**宽式 2**	**窄式 1**	**窄 2**
Chinese-TIMT	4.64 ± 0.39	4.96 ± 0.45	5.12 ± 0.45	4.30 ± 0.53	4.32 ± 0.83	5.25 ± 0.72	4.71 ± 0.75
AISHELL-1	4.18 ± 0.47	4.04 ± 0.47	4.25 ± 0.44	3.89 ± 0.41	4.21 ± 0.63	4.66 ± 0.61	4.18 ± 0.73
Wenet 有声读物	4.43 ± 0.37	4.50 ± 0.46	5.22 ± 0.39	4.28 ± 0.33	4.12 ± 0.84	5.70 ± 0.49	5.07 ± 0.98
Wenet 对话	5.27 ± 0.64	5.23 ± 0.68	6.04 ± 0.68	4.87 ± 0.54	4.77 ± 1.14	6.29 ± 0.73	5.85 ± 1.20
TIMIT	4.31 ± 0.54	4.67 ± 0.51	4.89 ± 0.50	3.61 ± 0.55	3.37 ± 0.77	4.57 ± 0.78	4.46 ± 1.26
Giga有声书	3.51 ± 0.54	3.99 ± 0.48	4.30 ± 0.49	3.88 ± 0.44	3.30 ± 0.59	5.16 ± 0.56	4.65 ± 0.80
TED-LIUM	4.26 ± 0.55	4.44 ± 0.48	5.16 ± 0.52	4.19 ± 0.53	3.86 ± 0.75	5.24 ± 0.56	4.99 ± 0.88
Giga-interview	4.10 ± 0.43	4.44 ± 0.42	4.95 ± 0.42	3.92 ± 0.47	3.72 ± 0.80	5.25 ± 0.54	4.88 ± 0.98
CV-Thai	/	3.53 ± 0.29	3.57 ± 0.29	4.32 ± 0.52	4.38 ± 0.59	4.93 ± 0.60	4.38 ± 0.83
CV-法语	/	4.34 ± 0.56	5.54 ± 0.60	5.06 ± 0.55	4.96 ± 0.96	5.50 ± 0.66	5.26 ± 0.95
CV-西班牙语	/	5.45 ± 0.57	5.59 ± 0.57	4.88 ± 0.61	4.70 ± 0.91	5.44 ± 0.77	5.07 ± 0.99
CV-波兰语	/	5.21 ± 0.51	5.39 ± 0.49	4.96 ± 0.56	4.85 ± 0.96	5.67 ± 0.69	5.35 ± 1.01
CV-俄语	/	6.14 ± 0.60	6.25 ± 0.59	4.87 ± 0.49	4.89 ± 0.81	5.40 ± 0.63	5.14 ± 1.04
CV-German	/	5.07 ± 0.54	5.12 ± 0.54	4.73 ± 0.55	4.32 ± 0.90	5.29 ± 0.60	5.00 ± 0.94