arXiv:2302.01984v1 [cs.CL] 3 Feb 2023

PSST！基于变换器的韵律语音分割

Nathan Roll[1], *CalbertGraham*[2],*Simon Todd*[11]加利福尼亚大学圣塔芭芭拉分校，美国[2]剑桥大学，英国nroll@ucsb.edu, crg29.cam.ac.edu, sjtodd@ucsb.edu

摘要

自注意力机制使 Transformer 在许多语音转文本（STT）任务中达到超越人类的表现，但自动韵律分段这一难题仍未解决。本文对 Whisper 这一预训练语音转文本模型进行微调，通过重新利用低频词元来标注语调单位（IU）边界。我们的方法准确率达到 95.8%，优于以往方法，且无需大规模标注数据或企业级计算资源。我们还通过应用一系列滤波器来削弱输入信号，发现使用截止频率为 3.2 千赫的低通滤波器，能够提升模型在样本外和分布外场景中的分割性能。我们发布了该模型[1]，既可用作转录工具，也可作为进一步改进韵律分段的基线。

索引词： 语调单元、自动语音识别、Whisper、Transformer、语音分段、边界检测

1. 引言

尽管语音以连续流的形式发出，并伴有协同发音效应，使音位的实现受到相邻音位的影响 [28, 41]，听者仍能够将语音感知为由不同声音构成的连贯序列。自动语音识别（ASR）是指机器将音频转换为一系列词语的过程。人类识别语音模式的能力是交际过程的基础。自动语音识别需要将词语、音节或音素等语言类别自动映射到相应的声学信号。自动语音识别有许多重要应用（例如虚拟助手和聊天机器人、语音命令与听写、实时字幕和转录文本）。然而，尽管技术取得了显著进步，机器语音分段仍然是语音处理领域最具挑战性的问题之一。造成这一困难的主要原因是语音存在多种变化来源，包括说话人特征（例如年龄、性别、声道长度）、环境因素（例如麦克风、房间声学）、说话风格（例如即兴表达与预先规划），等等。

传统上，语音的语言学分析侧重于对音段的分析。然而，越来越多的研究关注韵律：话语的自主音段结构，它编码了有关突显和短语组织的信息 [25, 21]。语音中的韵律现象包括相互关联、彼此作用的多种现象，例如语调、重音、节奏和短语划分 [1]。在英语中，短语组织将词语组合成若干语块，供听者和说话人用来处理话语。一个边界将这些语块逐一划分开来，这对于提高语音可懂度十分重要 [9, 32]。这有助于听者正确辨别话语的句法结构并推断其含义 [36, 42, 2, 10, 40]。本研究旨在开发一种适用于美式英语的自动语调单位切分和边界检测工具。

韵律线索在美式英语口语话语处理过程中发挥着重要作用 [39, 16]。文本转语音（TTS）系统的自然度部分取决于能否在话语中适当的位置生成停顿，也就是说话者自然会停顿的位置。一些研究人员利用时长线索和停顿来检测韵律短语边界 [44, 31], ，另一些则在自动边界检测中使用多种韵律线索[22, 24]。然而，由于以下若干原因，我们尚无法在自动语音识别技术中充分解释韵律：（1）韵律实现存在显著的说话者差异和语境变异；（2）韵律结构与话语中其他组织层级之间的关系极其复杂；（3）难以区分旨在标示边界的停顿与由无意失误造成的停顿。人类在识别语调单位（亦称语调短语或韵律单位）时，会通过有意识或无意识地识别各种往往细微的线索来克服这些挑战。这些线索包括语调轮廓的整体统一性、单位起始处的音高重置和弱起、单位末尾的延迟、单位边界处的停顿和换气，以及核重音 [13]。

语调单位识别的自动化方法差异很大，从简单的基于规则的算法 [5]到复杂的监督式机器学习模型 [35]。我们的方法在理论和架构上都具有创新性。我们认为，韵律并非一个独立的问题，而是与句法密切耦合的。从计算角度来看，Transformer模型在常规语音转文本任务中的成功，为端到端韵律转写应用提供了自然的起点；这类任务体现了句法与韵律接口的一半 [4], 。

本文探讨能否通过一个小型、高质量数据集提供的显式监督，“教会”预训练的基于Transformer的语音转文本模型将语音切分为语调单位。我们聚焦于以下主要研究目标：

1. 将针对自动语音识别优化的 Transformer模型改造为能够可靠地执行 IU 边界检测。2. 通过在缩减版微调集上重复实验，探究此类模型中声音与句法各自的作用。3. 通过谐波频率滤波以及对分布外语音数据的评估，检验韵律边界预测的稳健性。

[1]https://github.com/Nathan-Roll1/PSST

2. 方法

2.1. 数据

训练时，我们使用圣巴巴拉口语美式英语语料库（SBCSAE），因为该语料库参与者范围广泛，且转录文本质量高。语料库包含 60 段对话（210 位参与者）的自然话语及带有韵律标注的转录文本，总时长为 ~20 小时。说话人的年龄从 11 岁到 101 岁不等，自我认同为亚裔美国人、黑人／非裔美国人、拉丁裔／奇卡诺裔、西班牙裔、日裔、美国原住民、白人、双种族或其他族裔。他们来自美国 30 个州，受教育程度从小学到各类研究生学位不等。语料库中的性别比例大致均衡，55% 的说话人自我认同为女性，44% 自我认同为男性。其余说话人的性别数据不可用。SBCSAE 转录文本由多名受训的审查员完成，存在的差异由专家解决。所有个人身份信息及其他敏感信息（由语料库创建者认定）均使用 400 赫兹低通滤波器进行屏蔽，并在相关区域前后 45 毫秒内逐渐淡出 [12]。

我们的数据集版本包含 60 个单声道 .wav 文件，采样率为 22,050 赫兹。每个音频文件都附有一份 .cha 格式的文本转录稿，其中 IU 级时间戳的精度为 0.1 秒。

微调之前，我们从每份转录稿附带的 .cha 文件中提取 IU 级时间戳。在 60 份转录稿中，前五份（ ~2 小时）的有效片段被划分到测试集，其余部分则分配到训练集/验证集。若片段不存在重叠且彼此连贯，则视为有效片段。此外，鉴于输入长度固定为 30 秒，其他有效片段可能会被拆分为多个部分，每部分最多包含十个连续单元。使用前会删除代表其他语音现象的杂项词元，即呼吸声（吸气/呼气）和笑声。不过，填充停顿和言语不流畅现象（“um”“uh”等）会予以保留。

提取出的时间戳与输入音频匹配，并从 22,050 赫兹重采样至 16,000 赫兹。每个片段都会生成对数梅尔频谱图，使用 80 个通道、25 毫秒窗口和 10 毫秒步长。随后将输入矩阵重新缩放至 $([0, 1])$，并填充至 30 秒 [27]。

我们会人工检查现有词元词典，以找出最终转录中不需要且出现频率较低的词元，从而尽量减少对输出的干扰。对于 IU 边界，我们选择用五个连续感叹号表示的词元。

2.2. 模型

Transformer 是一种神经网络架构，于2017 [38]提出。与循环神经网络（RNN）或卷积神经网络（CNN）不同，它不对输入之间的关系施加任何结构约束，无论是在时间维度还是空间维度上。相反，它将位置信息编码到输入本身，从而通过并行化获得计算优势，并借助自注意力和交叉注意力提升性能 [37]。Transformer 在生物学、化学等多个领域取得了最先进的成果，但其最成功的应用领域是自然语言处理（NLP）任务 [18]。Whisper [27], 是这类 Transformer 模型之一，在多种语音处理用例中都取得了具有竞争力的表现，包括语音转文本（STT）合成。

我们的基于变换器的韵律语音分割（PSST）模型，是在 Whisper 最大的英语专用版本上微调而成，该版本拥有 7.64 亿个参数，大小为 3.06 GB。与 Whisper 初始训练周期相比，两项超参数调整为批量大小（从 256 调至 32）和梯度步数（从 1 调至 2），有效批量大小为 64。这些调整纯粹是出于计算资源限制。两个卷积层和一个高斯误差线性单元（GELU）激活函数，将 SBCSAE 输入音频的对数梅尔频谱图转换为线性向量，并与一个正弦位置编码向量相结合。该向量序列随后通过一系列编码器和解码器模块，每个模块均由注意力机制和多层感知机（MLP）组件构成。微调采用监督式方式进行，以人工生成的转录文本作为真实标注。

我们还实例化了模型的第二个版本（PSST-Acoustic），该版本使用经过语法屏蔽的 SBCSAE 进行训练（其他方面则完全相同）。所有文本词元都会替换为一个通用词元，并保留边界作为分隔符。这样做是为了确定 PSST 模型是否依赖句法。

为了检验 PSST 生成的边界是否仅仅源于词汇／句法概率，我们训练了一个仅使用文本的模型版本。我们不将音频转换为词元，而是让模型在 Whisper 生成的词元之间标注边界。我们的词汇分割模型以 GPT-NEO 的 12 亿参数（5.36 GB）分布版本启动 [6]，并使用来自 GPT-2 [27] 的分词器。训练使用 SCBSAE 预处理后的文本版本，数据划分与 PSST 和 PSST-Acoustic 所用的划分相同。基于 [27],的 PSST 基本架构如图1所示。

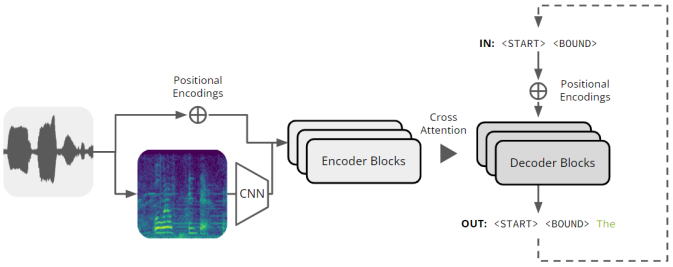

图1： *PSST* 架构

2.3. 评估

尽管大多数自动语音识别系统根据词错误率（WER）评估模型性能， [43], 但由于我们的任务涉及边界预测，因此模型采用分割指标进行评估。与我们用于构建模型架构的语音转文本方法不同，许多韵律任务的“真实标注”存在显著歧义 [23]。例如，语调短语边界的标注者间一致性为 93.4% [26]。

在此改用有意义的分割：若提出的边界在词级分隔上与样本外的专家转录文本一致，则视为准确。假阳性（过度分割错误）和假阴性（欠分割错误）的出现频率通常低于真阴性；真阴性出现在其余词与词之间的非边界划分处。因此，我们仅将准确率用作比较不同研究的参考指标。F1分数是精确率与召回率的调和平均值，因而更适合作为衡量有意义的分割性能的标准指标。词汇差

异（即生成的音频片段与专家转录片段之间的差异）通过基于 Transformer 的强制对齐技术 [45] 进行处理。自动语音识别和强制对齐不准确所引发的级联错误，尤其是个人标识符过滤导致的错误，可通过采用 20 毫秒的对齐窗口得到部分缓解。

2.4. 训练

出于内存管理的考虑，训练集拆分采用流式方式加载。所有训练均在一块配备 32 GB 显存的 NVIDIA V100 Tensor Core GPU 上完成。

PSST 和 PSST-acoustic 训练 400 步（即完整遍历训练数据两次）。为避免过早过拟合，前 50 步的学习率较低，随后该超参数线性增加，直至达到 10^{-5} 后保持不变。此阶段约需 2 小时 20 分钟。

词汇模型在与先前模型相同的硬件上进行微调，批量大小为 8，预热 100 步，权重衰减率为 10%。微调共进行两个周期，耗时不到 30 分钟。

2.5. 推理

PSST 和 PSST-acoustic 可在中央处理器上运行，但即使使用消费级 GPU，也能显著加快处理速度。平均而言，使用 NVIDIA T4 GPU 时，每个输入的推理仅需 4 秒（最长可达 30 秒）；而使用我们的中央处理器通常需要一分钟以上。降频折叠、重采样和特征生成所需的计算能力相对较少。借助 GPU 加速，PSST-lexical 的推理每个分块（1–10 个语调单位）仅需 1.2 秒。

2.6. 信号缩减实验

我们使用 PSST 和 PSST-acoustic 模型，对测试集中的音频应用一系列低通和高通滤波器。所选频率为 200 Hz、400 Hz、800 Hz、1.6 kHz 和 3.2 kHz，大致涵盖了 F0–F3频段，如 [34, 19] 中所述。我们针对截止频率分别设为上述各频率的低通和高通巴特沃斯滤波器计算评估指标 [7]。与 [14], 中的发现一致，我们观察到，较高的频段（3 kHz 以上）对语音可感知性的影响逐渐减弱。

3. 结果

3.1. 性能

我们的方法在 F1分数和总体准确率方面均达到了最先进的表现，甚至优于一些配有人工作标注正字法的方法。不过，片段定义、输入特征（句法和/或声学）、语料库内容（说话人数、是否为脚本化语音等）以及模型类型各不相同，因此难以进行比较。表 1 总结了以英语为对象的既有分段方法的表现。

我们还比较了IU长度的分布，发现PSST生成的IU密度与人工转录的IU密度在样本外具有相似性。这些分布如图2所示。

3.2. 在IViE语料库上的表现

我们在英语语调变异（IViE）语料库 [15] 上测试 PSST。与 SBCSAE 不同，IViE 聚焦于不列颠群岛城市地区的英语方言，并采用

表1： 分割性能

方法	*F1*	准确率
PSST	**0.87**	**0.96**
[29]	0.81	0.93
[30]	0.77	0.89
Whisper [27] +词汇	0.77	0.93
PSST-Acoustic	0.71	0.87
[17]	0.70	0.83
[5]	0.66	0.86
[20]	0.63	0.87
Whisper [27]	0.48	0.85

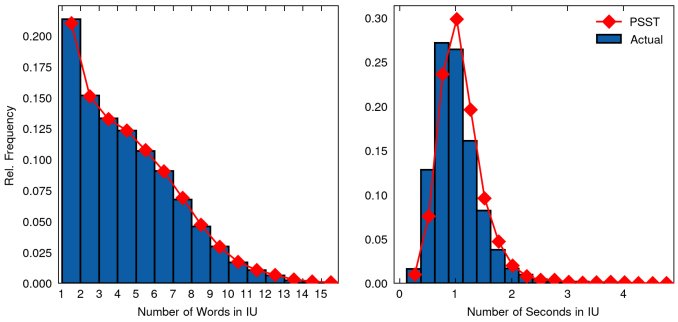

图 2： *IU*长度分布

一种独特的语调短语方法进行转写。IViE 标注系统改编自 ToBI 框架 [33, 3]。尽管如此，我们发现模型在这种分布外环境中仍表现稳健，但性能有所下降。表 2 总结了模型在 IViE 语料库上的表现。

表 2： *IViE* 语料库

方法	*F1*	准确率
PSST	**0.73**	**0.93**
Whisper [27] +词汇	0.56	0.89
Whisper [27]	0.35	0.87
PSST-Acoustic	0.00	0.82

3.3. 失败案例

失败案例大致分为两类：由自动语音识别导致的不准确性和韵律不准确性。我们模型中的语音转文本部分会生成一些专家转录中未包含的词元，也会漏掉另一些词元。我们发现，这在处理几乎听不见的填充停顿时尤其成问题。鉴于此类词元往往与边界相关，这些情况会拉低PSST的整体性能。听过相关音频后，我们注意到这些发声是否存在并不明确，而且即便它们确实存在，也不确定是否足以构成额外的语调单位。

另一类由自动语音识别导致的不准确性，是输出的词元较长或较短，无法在20毫秒的对齐窗口内完成对齐。扩大对齐窗口可以减少此类情况，但代价是可能会将不准确的边界误判为正确。

在词汇转录正确或接近正确的情况下 , IU

分段错误更明显地可归因于韵律因素。我们发现，这些错误往往伴随着更具主观性的示例。

3.4. 滤波器

应用 3.2 kHz 高通滤波器后，我们发现分割性能略有提升（∼0.1%）；而使用其他所有频率和滤波器时，性能均有所下降。如图 3 所示，滤波器越极端，整体性能下降越明显。

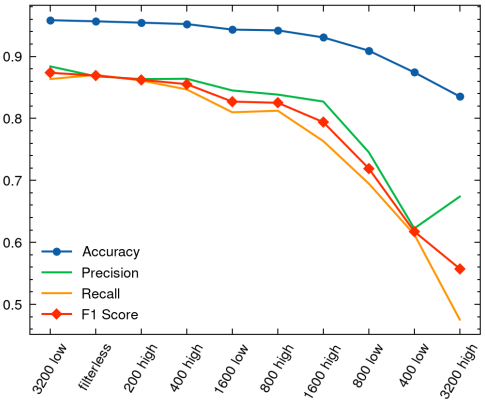

图 3： 滤波器频率与类型

尽管我们评估的是分割指标，但语音转文本准确率会产生内生效应。正字法完全一致的片段，其语调单位边界一致性也很高。因此，很难确定更居中的谐波频率捕捉到的是显著的韵律信息，还是仅仅捕捉到了正字法线索。

4. 讨论与结论

本研究旨在实现三个研究目标。关于目标 1，我们成功地改造了 Whisper [27]，使其能够将自发语音分割为语调单位。在此前未见过的示例上，我们取得了 0.87 的 F1分数，高于任何传统方法。Whisper 最初仅针对从音频中辨识词语这一简单目标进行训练，但我们能够借助少样本学习成功改造它，这一事实为其他依赖较小的数据集的自然语言处理研究带来了很大希望。

关于目标 2，我们考察了两个“同源”模型的表现：它们分别在 SBCSAE 的词汇版本和受正字法混淆的版本（PSST-acoustic）上进行微调。完整的 PSST 模型表现显著优于词汇模型和声学模型，后两者的 F1分数分别为 0.77 和 0.71。这些结果证实，韵律和句法都在边界判定中发挥作用。

关于目标 3，结果表明，模型在分布内和分布外数据集上的表现差异显著。在分布外的 IViE 数据集上应用时，PSST 模型成功预测了语调边界，F1分数为 0.731。不同口音之间的语调变异可能很大，例如不列颠群岛的口音，正如 Grabe 等，2001 所报告的那样。因此，美式和英式英语方言之间很可能存在更显著的差异 [11]。令人非常欣喜的是，该模型在 SBCSAE 数据集上训练，能够以这样的准确率预测其他英语方言中的边界。然而，我们认为，若要达到最佳表现，微调集应包含多种英语变体，包括母语影响各异的变体。同样，在不同的谐波滤波环境中也发现了表现差异；应用低于 800 Hz 和高于 1600 Hz 的掩蔽后，表现显著下降。200–1600 Hz 的范围大致对应英语中的第一共振峰和第二共振峰，其中包含了韵律分段任务最有用的信息 [8]。鉴于基频在语调中的突显作用，这一结果出乎意料。

综上所述，基于这项研究，我们推断，文本预测和韵律边界识别并非彼此独立的挑战，而只是统一语音处理目标的组成部分。只需以基于 Transformer 的模型能够处理的方式重新标记韵律特征，就能释放出模型识别语调单位的潜在能力。总体而言，我们的结果表明，鉴于这些语音转文本模型在少样本场景中的成功，它们会隐式地考虑韵律。此外，在受到中度基于频率的信号篡改，甚至完全基频掩蔽的情况下，分割性能依然稳健，这进一步支持了韵律—句法互动处于高性能自动语音识别模型“核心”的观点。

未来的工作可以考虑采用我们的自动语音识别重新分词过程来检测其他语音现象，例如韵律重音、嗓音质量变化，甚至环境语境。

5. 致谢

我们感谢 John DuBois 博士提供《SBC-SAE》副本，并感谢 Tirza Biron 博士协助确定稳健的性能指标。本研究由加利福尼亚大学圣塔芭芭拉分校的 URCA 资助资助。

6. 参考文献

[1] Amalia Arvaniti。《韵律语音学》。载于《牛津语言学研究百科全书》。牛津大学出版社，2020年7月。[2] CM Beach。句法结构歧义处韵律模式的解读：线索权衡关系的证据。《记忆与语言》，1991年。[3] Mary E Beckman 和 Gayle Ayers Elam。ToBI 标注指南（第3版，1997年3月）。俄亥俄州立大学研究基金会，1997年。[4] Ryan Bennett 和 Emily Elfner。句法—韵律界面。《语言学年度评论》，5(1):151–171，2019年。电子版：https://doi.org/10.1146/annurev-linguistics-011718-012503。[5] Tirza Biron、Daniel Baum、Dominik Freche、Nadav Matalon、Netanel Ehrmann、Eyal Weinreb、David Biron 和 Elisha Moses。自发语音中韵律边界的自动检测。《公共科学图书馆·综合》，16(5):e0250969，2021年。出版方：美国加利福尼亚州旧金山公共科学图书馆。[6] Sid Black、Gao Leo、Phil Wang、Connor Leahy 和 Stella Biderman。GPT-Neo：基于 Mesh-Tensorflow 的大规模自回归语言建模，2021年3月。[7] Stephen Butterworth。论滤波放大器理论。《无线工程师》，7(6):536–541，1930年。[8] J. C. Catford。《语音学实用导论》。载于《语音学实用导论》，第161页。克拉伦登出版社；牛津大学出版社，牛津 [英格兰]：纽约，1988年。[9] William E. Cooper 和 John M. Sorensen。《句子产出中的基频》。施普林格，纽约州纽约市，1981年。[10] D Crystal。韵律发展。载于《第一语言发展研究》，第174–197页。剑桥大学出版社，纽约州纽约市，1986年。

[11] Alex DiChristofano、Henry Shuster、Shefali Chandra 和 Neal Patwari。自动语音识别中不同口音间的性能差异，2022 年 8月。arXiv:2208.01157 [cs]。[12] John W. Du Bois、Wallace L. Chafe、Charles Meyer、Sandra A. Thompson、Robert Englebretson 和 Nii Martey。圣巴巴拉美式英语口语语料库，第1–4部分。费城：语言数据联盟，2000年。[13] John W. Du Bois、Susanna Cumming、Stephen Schuetze-Coburn 和 Danae Paolino。话语转写。《圣巴巴拉语言学论文集》，第4卷，1992年。[14] Emmanuel Ferragne、C´edric Gendrot 和 Thomas Pellegrini。迈向深度学习应用于语音比较时的语音学可解释性。载于国际语音科学大会，ISBN–978，第2019年。[15] Esther Grabe、B Post 和 F Nolan。英语语调变异建模。IViE系统。《韵律学*2000*会议论文集》，第51–57页，2001年。[16] Julia Hirschberg。交际与韵律：韵律的功能方面。《语音交际》，36(1-2):31–43，2002年。出版方：爱思唯尔。[17] Julia Hirschberg 和 Christine H. Nakatani。话题分段的声学指标。载于第五届国际口语处理会议（*ICSLP 1998*），论文编号0976–0。国际语音通信协会，1998年11月。[18] Katikapalli Subramanyam Kalyan、Ajit Rajasekharan 和 Sivanesan Sangeetha。Ammus：自然语言处理中基于Transformer的预训练模型综述。*arXiv*预印本 *arXiv:2108.05542*，2021年。[19] Raymond D. Kent 和 Houri K. Vorperian。元音共振峰频率和带宽的静态测量：综述。《言语沟通障碍杂志》，74:74–97，2018年。[20] Ondrej Klejch、Peter Bell 和 Steve Renals。采用声学和词汇方法对多类型广播节目进行带标点转写。载于*2016*年*IEEE*口语技术研讨会（*SLT*），第433–440页，美国加利福尼亚州圣迭戈，2016年12月。电气电子工程师学会。[21] D. Robert Ladd。语调音系学。剑桥大学出版社，第2版，2008年12月。[22] Shyamal Kr Das Mandal、A. K. Datta 和 B Gupta。利用超音段特征检测标准孟加拉口语（SCB）连续语音信号中的词边界。2003年。[23] Russell Moore、Andrew Caines、Calbert Graham 和 Paula Buttery。口语学习者英语中的语音单位自动划分。载于*COLING 2016*论文集：第二十六届国际计算语言学会议技术论文，第782–793页，日本大阪，2016年12月。COLING 2016组委会。[24] Benno Peters。德语自发语音中语音短语边界的多重线索。第1795–1798页，2003年。[25] J. Pierrehumbert。韵律与语调。载于《麻省理工学院认知科学百科全书》（编者：*Wilson RA* 和 *Keil F*），第479–482页。麻省理工学院出版社，马萨诸塞州剑桥，1999年。[26] John F. Pitrelli、Mary E. Beckman 和 Julia Hirschberg。评估ToBI框架中韵律转写标注的可靠性。第三届国际口语处理会议（*ICSLP 1994*），第123–126页，1994年9月。会议名称：第三届国际口语处理会议（ICSLP 1994）；出版方：国际语音通信协会。[27] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现稳健语音识别。*arXiv*预印本 *arXiv:2212 04356*，2022。。[28] Peter Roach、Helen Roach、Andrea Dew 和 Paul Rowlands。语音的语音学分析、自动分段与标注。《国际语音学会会刊》，20(1):15–21，1990年7月。出版方：剑桥大学出版社。[29] Andrew Rosenberg。韵律事件的自动检测与分类。哥伦比亚大学，2009年。

[30] Andrew Rosenberg。使用量化轮廓建模对韵律事件进行分类。载于人类语言技术：北美计算语言学协会分会*2010*年年会，第721–724页，美国加利福尼亚州洛杉矶，2010年6月。计算语言学协会。[31] Ariel Salomon、Carol Y. Espy-Wilson 和 Om Deshmukh。语音地标检测：时间信息的运用。《声学学会杂志》，115:1296–1305，2004年。[32] Elisabeth Selkirk。音系学与句法：声音与结构之间的关系。麻省理工学院出版社，马萨诸塞州剑桥，1984年。[33] Kim EA Silverman、Mary E Beckman、John F Pitrelli、Mari Ostendorf、Colin W Wightman、Patti Price、Janet B Pierrehumbert 和 Julia Hirschberg。Tobi：英语韵律标注标准。载于*ICSLP*，第2卷，第867–870页，1992年。[34] Sujeet Kumar Sinha 和 Vijayalakshmi Basavaraj。语音诱发听觉脑干反应：研究语音脑干编码的新工具。印度耳鼻喉科与头颈外科杂志：印度耳鼻喉科医师协会官方出版物，62:395–9，2010年10月。[35] Sabrina Stehwien 和 Ngoc Thang Vu。使用带上下文信息的卷积神经网络识别韵律事件，2017年6月。arXiv:1706.00741 [cs].[36] Lynn A. Streeter。短语边界感知的声学决定因素。《美国声学学会杂志》，64:1582–1592，1978年。[37] Yi Tay、Mostafa Dehghani、Dara Bahri 和 Donald Metzler。高效 Transformer：综述。《*ACM*计算调查》，55(6):109:1–109:28，2022年12月。[38] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。《Attention Is All You Need》，2017年12月。arXiv:1706.03762 [cs].[39] Jennifer J. Venditti 和 Julia Hirschberg。语调与话语处理。载于国际语音科学大会论文集，第315–318页。Citeseer，2003年。[40] Paul Warren。韵律与句法分析：导论。《语言与认知过程》，11:1–16，1996年。[41] Laurence White。语音分段。《牛津心理语言学手册》，第1页，2018年。出版社：牛津大学出版社。[42] Arthur Wingfield、Linda Lombardi 和 Scott Sokol。韵律特征与加速语音的可懂度：句法分段与周期性分段。《言语与听力研究杂志》，27:128–134，1984年。[43] J. P. Woodward 和 J.T. Nelson。语音识别性能的信息论度量。语音输入/输出技术标准化研讨会，海军航空发展中心，沃明斯特，宾夕法尼亚州，1982年。[44] Li-Chiung Yang。时长和停顿作为语音中短语与边界标记的指标。载于第*15*届国际语音科学大会论文集。巴塞罗那，第1791–1794页，2003年。[45] Jian Zhu、Cong Zhang 和 David Jurgens。无需文本的音素到音频对齐：一种半监督方法，2022年2月。arXiv:2110.03876 [cs, eess]。