基于预训练模型与序列建模的音素分割方法

杨尚龙[1,2]，余正涛[*1,2]，王文君[1,2]，董凌[1,2]，高盛祥[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1054179997@qq.com, ztyu@hotmail.com, 175360805@qq.com,
46761956@qq.com, gaoshengxiang.yn@foxmail.com

摘要

音素分割作为语音处理领域内的一个重要任务，对于关键词识别、自动语音识别等应用具有至关重要的意义。传统方法往往独立预测每一帧音频是否为音素边界，忽视了音素边界与整个音频序列以及相邻帧之间的内在联系，从而影响了分割的准确性和连贯性。本文提出一种基于预训练模型与序列建模的音素分割方法，在HuBERT模型提取声学特征的基础上，结合BiLSTM捕捉长期依赖，再用CRF优化序列，提升了音素边界检测的性能。在TIMIT和Buckeye数据集上的实验表明，本文方法优于现有技术，证明了序列建模在音素分割任务中的有效性。

关键词： 音素分割；自监督预训练；条件随机场；序列建模

Phoneme Segmentation Based on Pre-trained Model and Sequence Modeling

Shanglong Yang[1,2], Zhengtao Yu[*1,2], Wenjun Wang[1,2], Ling Dong[1,2], Shengxiang Gao[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
1054179997@qq.com, ztyu@hotmail.com, 175360805@qq.com,
46761956@qq.com, gaoshengxiang.yn@foxmail.com

Abstract

Phoneme segmentation is a crucial task in the field of speech processing, essential for applications such as keyword spotting and automatic speech recognition. Traditional methods typically predict phoneme boundaries independently for each audio frame, overlooking the intrinsic relationships between the phoneme boundaries, the entire audio sequence, and adjacent frames. This oversight can compromise the accuracy and coherence of segmentation. We proposes a phoneme segmentation method that leverages pre-trained models and sequence modeling. Based on acoustic features extracted by the HuBERT model, our method utilizes a BiLSTM to capture long-term dependencies and a CRF to optimize the sequence, thereby enhancing the accuracy and coherence of phoneme boundary detection. Experiments conducted on the TIMIT and Buckeye datasets demonstrate that our method outperforms existing techniques, validating the effectiveness of sequence modeling in phoneme segmentation tasks.

Keywords: Phoneme Segmentation , Self-supervised Pre-training , Conditional Random Fields , Sequence Modeling

*余正涛（通信作者）：ztyu@hotmail.com
基金项目：国家自然科学基金（U21B2027, 62376111, U23A20388, ）；云南省重点研发计划（202303AP140008, 202302AD080003, 202401BC070021）；云南省科技人才与平台计划（202105AC160018）

1 引言

音素分割，亦称音素边界检测，是在语音信号中确定离散音素单元间时间边界的关键任务，该任务在关键词识别(Keshet et al., 2009)、语音识别(Rybach et al., 2009)、语音合成(Lee et al., 2021)等语音处理领域发挥着至关重要的作用，为这些应用提供了基础的信息框架。

音素分割可以根据是否使用监督信息，分为有监督和无监督两大类。在无监督场景中，音素边界的推断仅依赖于语音信号本身(Kreuk et al., 2020; Cuervo et al., 2022)。而有监督学习则利用已知的音素边界位置作为训练标签，进一步可分为文本相关监督和文本无关监督。文本相关监督需要额外的音素标签作为输入(McAuliffe et al., 2017; Lin and Wang, 2022)，并要求推断的边界位置与输入标签严格对齐，但这种强制对齐的方法在应用时也需要音素标签，限制了其在新语种上的适用性。

本文聚焦文本无关的监督学习，即仅使用音素边界作为监督信息，不涉及音素标签。以往的研究主要将此任务视为二分类问题，将音频帧分类为边界或非边界。这些传统方法通常条件独立地对每一帧进行预测(Franke et al., 2016; Zhu et al., 2022; Kim and Choi, 2023; Strgar and Harwath, 2023)，忽略了音频信号中帧与帧之间的时间依赖性。鉴于音频信号的连续特性，一帧音频是否为边界并非孤立事件，而是与其周围帧以及整个音频序列紧密相关，可见现有方法存在明显的局限性。

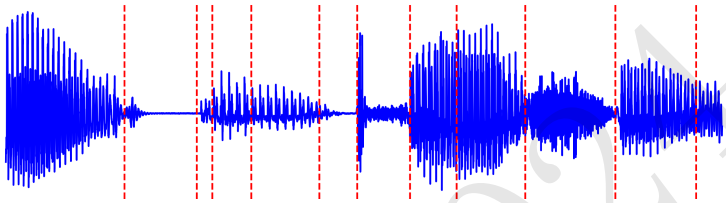


图1. 语音信号中的音素边界

受到序列标注任务的启发(Huang et al., 2015; Dalai et al., 2023; Zhou et al., 2019)，本文采用序列建模策略，提出了一种简单有效的音素分割方法。具体而言，首先使用预训练的HuBERT模型(Hsu et al., 2021)提取每一帧的音频表征，以利用无监督预训练过程中获得的丰富语音先验知识。然后，通过双向长短期记忆（BiLSTM）层(Graves and Graves, 2012)捕获音频序列中的长期依赖关系，进一步通过线性层将上下文表征转化为发射得分（emission scores），并利用条件随机场（CRF）层(Lafferty et al., 2001)对整个音频序列进行全局优化，以实现更加精确和连贯的音素边界判定。

本文主要贡献如下：（1）提出一种基于语音预训练模型与序列建模的音素分割新方法，强调了从整体序列角度进行建模的重要性；（2）在TIMIT(Garofolo, 1993)和Buckeye(Pitt et al., 2005)数据集上的实验证明，在不依赖额外音素标签信息的条件下，本文所提方法优于现有的音素分割方法；（3）本文方法具有良好的跨语言迁移能力，在训练未见的语言上直接推理也能取得较好的性能。

2 相关工作

2.1 音素分割

音素边界检测在多种实验设置下有着广泛的研究，具体可分为无监督、文本相关监督以及文本无关监督方法。

在无监督设置中，研究主要依赖于语音信号本身，不利用任何额外的音素或边界标签。传统方法侧重于利用信号处理技术，通过分析信号的频谱变化来探测音素边界(Hoang and Wang, 2015)。(Michel et al., 2016)利用循环神经网络（RNN）来预测下一帧音频，通过峰值检测算法发现可能的音素转变区域。(Kreuk et al., 2020)通过噪声对比估计技术来区分相邻帧和随机干扰帧，并应用峰值检测算法来定位音素边界。此外，对比预测编码（CPC）及其变体(Chorowski et al., 2021; Cuervo et al., 2022)在预测潜在向量时，尝试对未来多个潜在向量进行匹配，并使用对比损失进行模型训练，为无监督音素分割提供了新的思路。

在监督设置下，文本相关监督方法此前得到广泛研究。这类方法一般采用隐马尔可夫模型（HMM）或利用手工设计的特征进行结构化预测(Keshet et al., 2005; McAuliffe et al., 2017)。

虽然文本相关方法在一定程度上可以提升音素边界的检测准确性，但其性能在很大程度上受限于对音素信息的需求，在没有可靠音素标签的新语言资源中存在较大的局限。

文本无关监督则为本文所关注的重点。在该设置下，只需要音素边界作为监督信息而无需提供具体音素标签。(Franke et al., 2016)使用LSTM来捕捉音频的上下文依赖关系，并通过交叉熵损失进行优化。(Kreuk et al., 2020)采用可学习的分段特征识别可候选的音素边界，通过动态规划算法来最小化结构化损失，还证明了来自音素标签的额外监督信号可以改善边界检测。(Zhu et al., 2022)提出了一种半监督方法，通过对比学习以及前向和损失直接学习语音到音素的对齐。(Kim and Choi, 2023)采用自回归架构来利用之前的边界预测信息，通过提供有关先前帧是否被分类为边界的附加信息来防止不必要的边界重复。(Strgar and Harwath, 2023)将迁移学习应用于音素分割任务，证明了自监督预训练模型的表征能力在音素分割任务上的有效性。

2.2 自监督预训练

自监督学习为一种无需标注数据的语音表示学习方法，通过自监督预训练从大量语音数据中学习有用信息，能够适应各类下游任务。自监督预训练模型近年来在语音领域取得显著成功，通过微调或作为特征提取器的方式在各项下游任务上取得先进的性能。

目前最广泛使用的预训练模型为wav2vec 2.0(Baevski et al., 2020)与HuBERT(Hsu et al., 2021)。wav2vec 2.0结合对比学习方法和掩码技术，旨在最大化上下文表征和本地化表征之间的相似性。在生成局部表征时，模型通过随机掩码学习使用未被掩盖的表征预测被掩盖部分，这样的设计使模型更关注输入波形的上下文信息，而不仅依赖局部表征，从而极大提高了泛化能力。HuBERT训练过程类似于声学模型和语言模型的联合学习：首先，模型学习从未被掩码的时间步提取连续潜在表征，这些表征被映射到离散单元，类似于传统的基于帧的声学建模。其次，HuBERT学习捕捉长期的时序依赖关系，以对掩码时间步做出准确预测。近期的研究证明，预训练模型编码的信息不仅包括传统方法所提取的帧级特征，还涵盖音素级、词级，甚至是语义层级的特征(Pasad et al., 2023)，对各种语音任务的研究具有深远的意义。

3 方法

本文所提出的模型架构包括HuBERT层，BiLSTM层和CRF层。在训练阶段，音频首先通过预训练的HuBERT模型，以获得声学表征序列。随后，BiLSTM层被用于进一步捕获音频序列中的长期依赖关系。其输出的上下文表征再通过线性层处理，输出每一帧对应各类标签的发射得分，这些得分表示了当前帧处于边界或非边界状态的概率，并用于决定最终的分类标签。CRF层则以全局方式对音频序列进行建模，综合利用每一帧的发射得分和标签间的转移概率，以模拟所有可能的标签序列，整个训练过程对HuBERT进行微调。在解码阶段，采用维特比算法来解码，得到最优的标签序列。

BiLSTM能够在每个时间点捕获过去和未来的信息，从而进一步增强预训练模型提取的特征，为每个时间步生成一个更全面的上下文依赖表征，在特征层面可以为CRF层提供更准确的发射得分；CRF层计算相邻标签转换的概率（例如从“边界”到“非边界”），并结合BiLSTM层输出的发射得分来最大化整体标签序列的概率。二者的结合能够利用BiLSTM捕获长距离依赖的能力和CRF全局序列优化的能力，这样我们的模型不仅能感知单帧的局部特征，还能理解整个音频序列的时序依赖性，从而提升音素分割的准确性和一致性。完整的模型架构如图2所示。

3.1 声学特征提取

在音素分割任务中，输入为原始音频序列$\bar{x} = (x_1, ..., x_T)$，序列的长度$T$与音频的总时长相关，原始序列通过特征提取转换为$\bar{z} = (z_1, ..., z_K)$这样的声学特征向量序列。每个输入语音序列都对应一个时间戳标签序列$\bar{y} = (y_1, ..., y_K)$，其中时间戳标签序列长度与声学特征向量序列长度$K$取决于预处理过程与特征提取方法，每个$y_i$是一个二元分类标签，用以标识音素边界或非边界。

传统的特征提取方法，如梅尔频率倒谱系数（MFCC），在捕获丰富的语音表征方面存在局限，特别是在提取音素级特征时。近年来，自监督学习在语音处理领域取得显著成果，在多个下游任务中展现了巨大潜力，通过对预训练模型进行微调或应用迁移学习至特定任务，已在众多语音处理任务上实现了前所未有的性能。

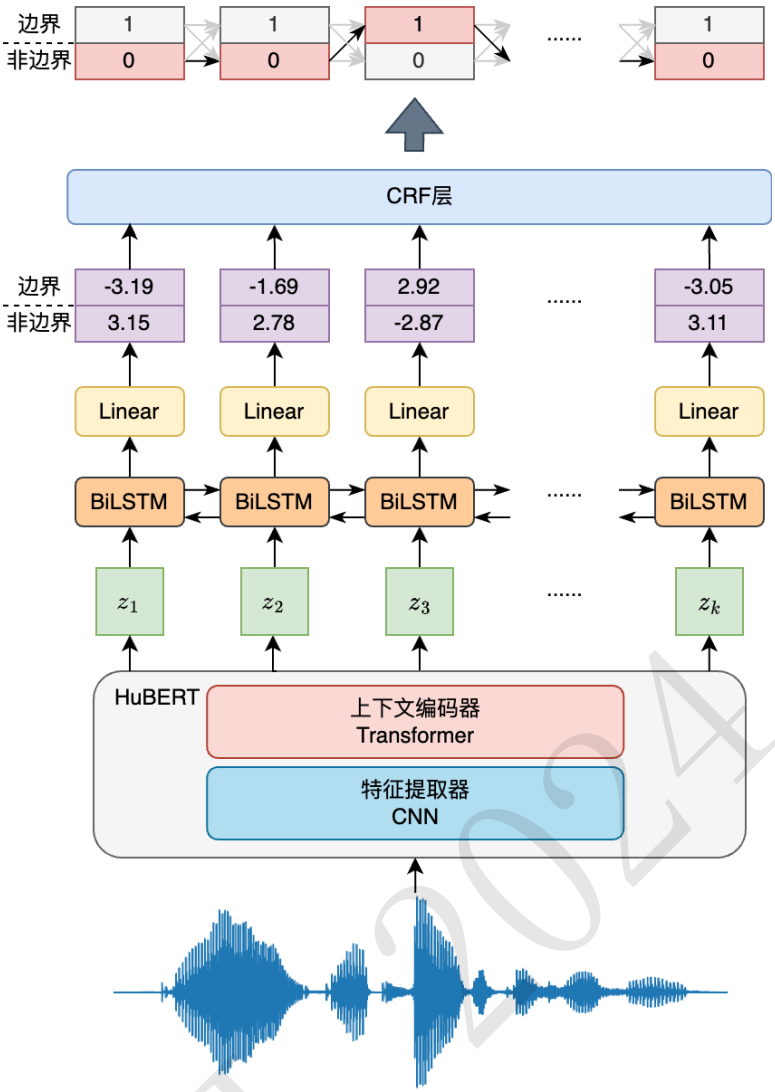


图2. 基于预训练模型与序列建模的音素分割架构

本文选择使用预训练的HuBERT模型作为声学特征的编码器。特征提取函数可以表示为$f \circ g$。其中f是一个卷积特征提取器，负责处理原始波形输入，并输出潜在的语音表示时间序列，这一序列在HuBERT的端到端训练过程中被学习；g是一个上下文编码器，它通过学习到的注意力掩码来从每个语音表征合成上下文感知的表征。函数f和g由HuBERT初始化，并在训练过程中接受梯度更新，从而优化特征表示，特征表示过程如式（1）所示。

$$\bar{z} = f \circ g(\bar{x}) \tag{1}$$

3.2 上下文信息建模

为了获得更丰富的上下文表征，本文采用双向长短期记忆网络（BiLSTM）进一步处理语音表征序列。长短期记忆网络（LSTM）是循环神经网络（RNN）的一种改良形式，设计用来学习序列数据中的长期依赖关系，有效地解决了梯度消失和梯度爆炸的问题。与RNN的主要区别在于，LSTM的隐藏层更新机制由专门设计的记忆单元替代，这些记忆单元负责控制信息的遗忘比例并将其传递到下一时间步。

在音素分割任务中，我们认为当前帧的预测不仅受到前序帧的影响，同样也与后续帧的信息密切相关。因此，我们引入BiLSTM来分别处理向前和向后输入序列，构建两种独立的隐藏状态，从而捕捉过去和未来的上下文信息。具体而言，可以将单向LSTM的隐藏状态更新过程表示为$\overrightarrow{h_t} = lstm(\overrightarrow{h_{t-1}}, e_t)$，其中$h_t$表示特定时间点的隐藏状态，$e_t$为当前时间步的输入。对于BiLSTM来说，我们得到两个方向的上下文表示，分别为：

$$\overrightarrow{h_t} = lstm(\overrightarrow{h_{t-1}}, e_t) \tag{2}$$

$$\overleftarrow{h_t} = lstm(\overleftarrow{h_{t-1}}, e_t) \tag{3}$$

其中$\overrightarrow{h}_t$和$\overleftarrow{h}_t$代表了从左到右（前向）和从右到左（后向）学习到的上下文信息。通过融合这两种方向的上下文表示，我们能够为每一帧音频构建更为全面的上下文表征。

$$h_t = \overrightarrow{h_t} + \overleftarrow{h_t} \tag{4}$$

3.3 标签序列解码

在音素分割任务中，利用相邻标签之间的依赖性或相关性对于减少独立预测每一帧音频可能产生的错误至关重要。例如，音素边界之后不太可能紧接着又出现一个边界，因此，利用相邻标签的相关性解码给定输入序列，可以确保生成的标签序列准确且合理。条件随机场（CRF）是一种在序列标注任务中广泛应用的方法，能够建模观测序列与状态序列之间的条件概率，在音素分割任务中，CRF在每个时间点的预测不仅考虑当前观测到的时间步，同时根据前后状态的依赖关系来优化整个序列的预测决策，它通过捕获序列中标签间的相互关系，优于单纯使用softmax函数的方法。我们将CRF用作标签序列解码器，以捕捉标签间的关联并对整个标签序列进行联合建模，而非单独预测每个标签。对于标签序列$\bar{y}$，我们定义第i个标签的得分为以下形式：

$$S_{y_i} = A_{y_{i-1},y_i} + P_{y_i} \tag{5}$$

这里的A是一个转移矩阵，其中$A_{i,j}$表示从标签i到标签j的转移得分，P是发射矩阵，由BiLSTM层的发射得分决定，因此，任何一个可能的标签序列的输出概率可以表示为：

$$p(Y_i \mid X) = \frac{exp(S(X, Y_i))}{\sum_{j=1}^{K} exp(S(X, Y_j))} \tag{6}$$

我们通过最大化对数似然函数来训练CRF，其损失函数的计算如下：

$$L_{CRF} = -\sum_{i=1}^{K} exp(S(y_i|x_i)) \tag{7}$$

通过这种方式，模型不仅能够有效地学习标签间的相互依赖性，还能够确保预测出的标签序列在整体上更加准确和连贯。

4 实验设置与结果分析

4.1 实验设置

4.1.1 数据集

本文使用TIMIT和Buckeye语音语料库对模型进行训练和评估。对于TIMIT数据集，我们按照标准流程将数据划分为训练集和测试集，并仿照先前的研究(Strgar and Harwath, 2023)，使用训练数据的10%作为验证集。在处理Buckeye数据集时，我们根据之前的工作设定了训练集、验证集和测试集的比例分别为80%、10%和10%(Kreuk et al., 2020; Strgar and Harwath, 2023)。此外，为了提高实验的准确性和可靠性，我们将较长的录音文件在非语音噪声和静音期间切割为较短的连续语音片段，确保每个片段的起始和结束部分的非语音时间不超过20毫秒。这样的处理旨在优化数据输入，从而为音素分割任务提供更加稳定和准确的实验条件。

我们还在老挝语上进行了跨语言实验，在不进行额外训练的前提下直接使用老挝语测试集测试模型性能。测试集由老挝本土人员人工标注，为了保持实验的一致性，我们将数据标签转化为TIMIT和Buckeye相同的格式，所使用的公共数据集和老挝语测试集的具体信息如表1所示。

	TIMIT	Buckeye	老挝语测试集
时长（小时）	3.45	7.69	1.89
句子数（条）	4009	10264	2050

表1. 数据集信息

4.1.2 数据预处理

为了让音素标签与经过HuBERT预处理的音频数据之间保持时间上的一致性，我们对数据集的标签格式进行了必要的转换。原始数据集以采样点为单位，记录了每个音素的开始和结束时间，其中采样率为16kHz。目标是将这些时间戳标签转换成二进制的音素边界标签，生成的标签以20ms为单位，标识该帧是否为音素边界（边界标记为1，非边界标记为0）。

为实现这一转换，我们采取了以下步骤。首先，模拟HuBERT处理流程中的7层卷积操作，确定卷积后音频的长度，并计算原始音频长度与卷积处理后音频长度之间的比例因子。然后，利用这个比例因子，将原始音素时间标注转换为与处理后音频长度相匹配的新时间帧格式。通过这种方法，我们确保了音素标签与HuBERT处理后的音频数据在时间上的对齐，为后续的训练和评价提供了准确的标签信息。

4.1.3 模型配置与评价指标

本文实验构建在fairseq(Ott et al., 2019)框架上，选择Librispeech(Panayotov et al., 2015)960小时数据集预训练的 HuBERT[1]作为基础检查点进行加载。在模型配置方面，我们设置BiLSTM的隐藏层维度为768，层数为2，对HuBERT层的微调学习率定为1e-4，而CRF层的学习率则设为1e-2。优化过程采用Adam优化器进行，其中β_1为0.9，β_2为0.999。所有的训练过程均在1张NVIDIA GeForce RTX 3090 GPU上进行，以16为批量大小，并设置默认训练周期为30。

评价指标方面，本文实验延续先前研究的做法，采用准确率、召回率、F1值以及R-value作为主要评价标准(Räsänen et al., 2009)，设置20ms为误差窗口。为了限制在一个误差窗口内，多个预测边界与单一真实边界匹配并重复计算贡献的情况，我们还采纳了(Strgar and Harwath, 2023)提出的更为严格的评价指标计算方法，即在多个预测边界与真实边界匹配的场景中，仅能有一个预测被视为正确并计数。考虑到R-value能够平衡召回率与过度分割的关系，其被认为是评估分割任务最为科学的指标。在模型训练期间，每处理50个批次的数据则在验证集上计算评价指标，根据严格指标下的最佳R-value保存最佳模型检查点，并用于最终测试。如果连续50次在验证集上的评估未能更新最佳模型将提前终止训练。

4.2 实验结果

4.2.1 与现有音素分割方法的对比实验

为了验证本文方法的有效性，我们在TIMIT和Buckeye数据集上与现有的音素分割方法进行了对比。选取了以下几种方法作为基线模型，具体介绍如下：

（1）segfeat：利用可学习的分段特征来识别潜在的音素边界，并通过动态规划算法最小化结构化损失，结果取自(Kreuk et al., 2020)。

（2）superseg：采用自回归架构，预测时考虑前一帧的边界预测信息，并利用先前帧是否被分类为边界的附加信息来限制边界的重复预测，结果取自(Kim and Choi, 2023)。

（3）wav2vec2.0 finetune：通过二元交叉熵损失微调预训练的wav2vec 2.0进行音素分割，目前为这两个数据集上表现最佳的模型，结果取自(Strgar and Harwath, 2023)。

如表2所示，本文所提方法在所有评价指标上优于现有方法，实现了最优性能，证明了序列建模在音素分割任务上的有效性。segfeat和superseg两种方法分别依赖于MFCC和对数梅尔频谱图进行特征提取，尽管这两种方法均采取了额外策略以约束预测结果，但传统特征提取方法在捕获音素级别的细节信息及其上下文信息方面存在限制，导致了性能的局限。wav2vec 2.0 finetune方法首次将自监督预训练模型应用于音素分割任务，并取得了一定成就，但其微调策略主要基于对独立音频帧的预测与优化，未能充分考虑音频序列的时序依赖性和标签间的转移关

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert

系，从而限制了模型的整体性能。本文方法在严格评价指标（*）的性能提升更加明显，充分体现了从序列角度优化并预测的合理性。

数据集	方法	标准指标				严格指标			
		准确率	召回率	F1	R-val	准确率*	召回率*	F1*	R-val*
TIMIT	segfeat	94.03	90.46	92.22	92.79	-	-	-	-
	superseg	95.63	94.77	95.29	95.82	-	-	-	-
	W2V2 finetune	96.90	96.30	96.60	97.04	94.35	93.91	94.13	94.96
	本文方法	**97.47**	**97.51**	**97.49**	**97.86**	**95.77**	**95.18**	**95.48**	**96.08**
Buckeye	segfeat	85.40	89.12	87.23	88.76	-	-	-	-
	superseg	89.92	89.94	89.93	91.40	-	-	-	-
	W2V2 finetune	94.01	93.08	93.54	94.41	90.56	90.28	90.42	91.81
	本文方法	**94.98**	**95.20**	**95.09**	**95.82**	**92.73**	**92.98**	**92.85**	**93.91**

表2. 在TIMIT和Buckeye数据集上与其它方法的比较，*代表严格指标，-代表未提供分数

4.2.2 声学上下文特征序列对音素分割性能的影响

CRF的解码结果依赖于其输入的发射得分，而不同的声学上下文特征序列所产生的得分也有所不同，为了评估不同特征序列对音素分割任务性能的影响，我们设置了多组实验来探究最适合CRF层的输入嵌入，具体实验设置如下：

（1）MFCC+BiLSTM：使用标准的MFCC特征提取方法，结合BiLSTM捕获音频的上下文依赖性。

（2）Whisper+BiLSTM：使用在大量数据上进行弱监督训练的Whisper(Radford et al., 2023)编码器提取特征，再结合BiLSTM，此处选择预训练的base版本的Whisper编码器[2]。

（3）HuBERT+BiLSTM：采用预训练的HuBERT模型作为特征提取器，利用其最后一层的输出作为音频表征，并结合BiLSTM以捕捉序列的上下文信息。在此配置下，HuBERT在整个训练过程中保持冻结状态。

（4）wav2vec 2.0+BiLSTM finetune：使用预训练的wav2vec 2.0替代HuBERT，并对其进行微调，为了公平，同样选择基于Librispeech 960小时进行训练的预训练检查点[3]。

（5）HuBERT finetune：直接使用HuBERT输出的表征计算发射得分，训练过程中对HuBERT进行微调，无BiLSTM层。

（6）HuBERT+LSTM finetune：将BiLSTM替换为单向LSTM，并对HuBERT进行微调。

（7）HuBERT+Transformer finetune：使用基于自注意力机制的Transformer编码器代替BiLSTM。Transformer在处理序列数据时捕捉长距离依赖，同时通过多头注意力机制并行处理不同表示子空间中的信息，设置隐藏层维度为768，多头注意力头数为8，训练过程中对HuBERT进行微调。

（8）HuBERT+BiLSTM finetune：本文采用的方法，结合HuBERT和BiLSTM，并对HuBERT进行微调。

实验结果如表3所示，比较（1）、（2）与其他结果，发现无论是否进行微调，使用自监督预训练模型（HuBERT或wav2vec 2.0）均优于传统的MFCC特征提取方法，这证实了自监督模型在捕获音素级信息及其上下文关系方面的优势，同时我们发现使用Whisper编码器提取特征在音素分割任务上表现不佳，这或许是因为Whisper主要适用于语音识别、语音翻译等任务，在音素粒度的表征较差。对比（3）、（8）可发现，对HuBERT进行微调能够明显提高其表征能力，比冻结预训练模型的方法能取得更佳的效果。实验（4）和（8）的比较可以证明在同样的条件下HuBERT比wav2vec 2.0在本任务上的效果略好，这和(Strgar and Harwath, 2023)的结论不同，这一发现可能归因于HuBERT和wav2vec 2.0不同的预训练策略所带来的上下文表征能力差异。相比于wav2vec 2.0通过计算量化器得到的码本与Transformer编码器输出的隐层向量之间的对比损失来优化模型，HuBERT通过离线聚类和掩蔽预测过程的交替进行，可以获得更优

[2]https://github.com/openai/whisper
[3]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

的上下文表征能力。通过CRF损失以全局序列建模并微调HuBERT更能激发其表征效果，使其在本任务上性能超越了wav2vec 2.0。（5）和（8）的对比实验证明了BiLSTM优化序列上下文依赖关系的有效性，进一步对比（6）和（8）验证了BiLSTM比单向LSTM取得更好的效果，证明了音素分割任务受到前后序列的双重影响。（7）和（8）的对比发现Transformer编码器在本任务上的结果略微低于BiLSTM，考虑到本任务的音频序列较短，同时相邻帧的依赖关系比全局的依赖关系更加重要，这也导致了Transformer编码器无法发挥其捕捉长距离依赖关系的优势。

嵌入	F1	R-val	F1*	R-val*
(1)	94.23	95.03	90.22	91.64
(2)	52.50	56.71	48.50	55.22
(3)	97.11	97.47	94.33	95.04
(4)	97.39	97.78	95.19	95.87
(5)	97.36	97.75	95.18	95.86
(6)	97.41	97.78	95.35	95.80
(7)	97.32	97.72	95.10	95.78
(8)	**97.49**	**97.86**	**95.48**	**96.08**

表3. 不同特征输入序列在TIMIT数据集上的实验结果比较，*代表严格指标

4.2.3 损失函数对音素分割性能的影响

为了证明序列建模方法相比于独立预测方法的优势，我们设计了不同的损失函数作为优化目标，在对HuBERT或wav2vec 2.0进行微调并加上BiLSTM框架的基础上，我们探索了以下三种损失函数对实验性能的影响：

（1）交叉熵损失（BCE loss）：通过独立比较每一帧音频与其真实标签间的差异，进行损失计算，其具体计算公式如式（8）。

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{8}$$

（2）焦点损失（focal loss）：对标准的交叉熵损失进行改进，引入了一个系数因子，允许根据预测的准确度自适应调整每个样本对总损失的贡献度，具体计算公式如式（9）。其中α是权衡因子，用于减少类别不平衡的影响，而γ是一个调节参数，用于调整易分类样本的损失贡献。我们设定α为0.6，γ为2，以提高对难以分类的音素边界样本对损失的影响。

$$L_{focal} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i)] \tag{9}$$

（3）CRF损失（CRF loss）：本文采用的方法，通过从整个序列的视角对模型进行建模和优化，其具体计算公式如式（7）。

实验结果如表4所示，焦点损失提升了对难分类样本以及音素边界样本在损失中的惩罚，我们实验中测试了不同的α与γ值，但发现无法产生明显的提升，说明了单独预测的方法在根本上存在局限性。采用CRF损失函数从序列整体进行建模与优化，相较于单个样本的独立预测方法，在所有评价指标上均获得了显著提升。这不仅证明了CRF损失函数在处理音素分割任务时的有效性，也突显了考虑音频序列时序依赖性的重要性。与单独对每一帧进行评估的交叉熵损失和尝试通过调整损失权重来提高难分样本影响的焦点损失相比，CRF损失通过全局优化整个序列的标注，有效捕获了音素边界之间的相互依赖关系，显著优化了模型对于音素分割的理解和预测能力。进一步比较HuBERT与wav2vec 2.0两种预训练模型在不同损失下的结果，我们发现HuBERT在损失函数为交叉熵损失和焦点损失时均落后于wav2vec 2.0，但采用CRF损失后指标的提升更明显，能够达到最佳的性能，说明了HuBERT通过CRF损失进行序列建模能够高度激发其上下文表征能力。这些发现强调了音素分割序列任务的本质，并且采用合适的损失函数对于提升模型性能具有关键作用。

嵌入	损失函数	F1	R-val	F1*	R-val*
W2V2 + BiLSTM	BCE loss	96.67	97.16	93.87	94.77
	focal loss	96.74	97.43	93.89	94.92
	CRF loss	97.39	97.78	95.19	95.87
HuBERT + BiLSTM	BCE loss	96.53	96.96	93.70	94.57
	focal loss	96.70	97.10	93.68	94.61
	CRF loss	**97.49**	**97.86**	**95.48**	**96.08**

表4. 不同损失在TIMIT数据集上的实验结果比较，*代表严格指标

此外，音素分割可能会应用于一些实时场景，为了验证本文方法引入复杂的解码方式后的实时性，我们进一步比较了不同损失函数的解码效率。具体而言，我们在TIMIT测试集上比较了不同方法的RTF值（RTF=解码时间/音频持续时间），各组实验均采用微调HuBERT模型并结合BiLSTM，只有损失函数不同，结果如表5所示，我们的模型在解码时间上确实会多于独立预测的方法，但是RTF值也接近于1，表示能够在一段音频的持续时间结束时完成解码，在实时应用的场景下并不会造成明显的效率降低。

损失函数	RTF
BCE loss	0.761
focal loss	0.776
CRF loss	1.093

表5. 不同损失在TIMIT数据集上的解码效率比较

4.2.4 跨语言音素分割实验

本文提出的音素分割方法在训练阶段没有使用任何额外的音素标签作为监督信息，同样，在推理阶段也完全摒弃了对音素词典的依赖。这种独立性不仅简化了模型的训练和应用过程，而且理论上模型具备跨语言迁移的强大潜力。

为了深入探索并验证本文方法的跨语言应用能力，我们在老挝语测试集上进行了测试。具体来说我们在英语数据集TIMIT上训练模型，然后将训练好的模型直接应用于老挝语的测试。值得注意的是，老挝语属于声调语言，其声调的多样性高于英语，在语言学特性上和英语有着显著差异，这对跨语言迁移产生了严峻的考验。

在没有对模型进行任何针对老挝语的微调操作的前提下，我们对比了本文方法与现有最佳模型在老挝语上的表现。实验结果如表6所示：由于在训练中未曾接触老挝语语料，模型在老挝语的测试中性能确实有所下降，但本文提出的模型相比于现有模型，仍展现出了更为卓越的性能，达到了71.07的R-value。这一结果表明了本文方法在处理未知语言时，仍然能够保持一定的鲁棒性和适应性。

方法	准确率	召回率	F1	R-val
W2V2 finetune	64.65	70.99	67.67	70.97
本文方法	**66.89**	**82.33**	**73.81**	**71.07**

表6. 在老挝语测试集上的实验结果比较

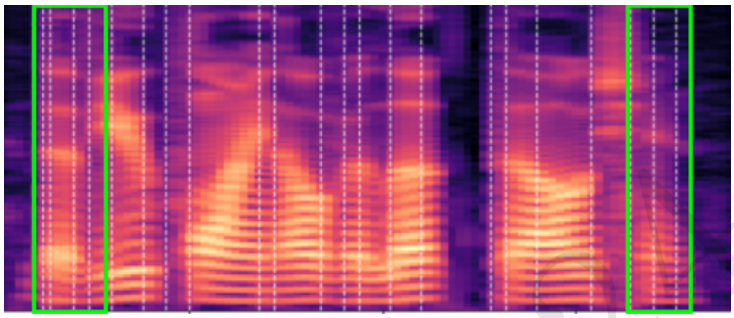
对于跨语种泛化能力的来源，我们认为传统的独立预测的方法只能学习到音频帧与音素边界的对应关系，即模型只能掌握如何根据当前帧的特征判断是否为音素边界，而我们提出的序列建模的方法不仅能够学习到这一对应关系，还能学习到相邻标签之间的依赖以及整个序列的长期依赖。对于不同的语言，音素类别会存在较大差异，传统方法在面对新的语言时可能会遇到很多训练中未见过的特征，从而导致性能急剧下降。我们提出的模型虽然也会面临这一问

题，但我们的模型相比传统方法学习到更多的序列间的依赖关系，并且对于不同语言，相邻标签之间的依赖与整体序列的依赖上的差异会相比音素特征的差异更小，因此本文方法在跨语言实验上展现了更好的能力。

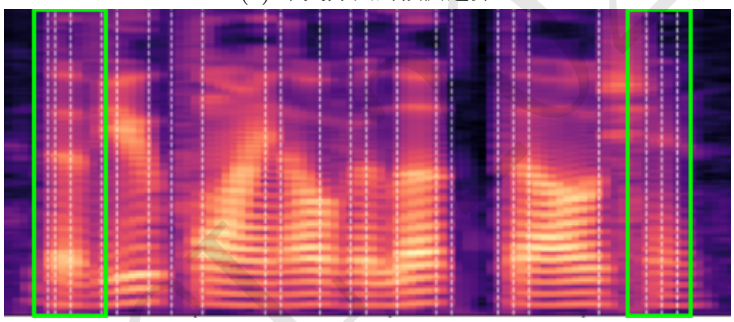
我们采用的预训练HuBERT模型仅在英文数据上进行了预训练，可能无法准确捕捉到老挝语的特定声学特征。然而，与基于独立预测的模型相比，本文方法通过序列建模，能够更有效地利用音频序列中的时序依赖信息。正是由于对音频序列时序特性的深度挖掘和利用，使得本文方法在跨语言迁移的应用上相比现有技术取得了更优的性能表现。

4.2.5 分割实例分析

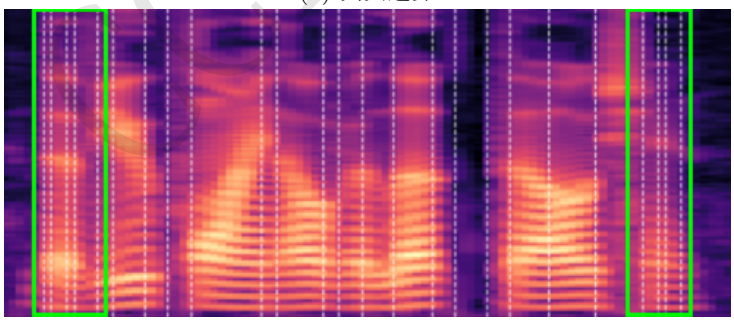
为了更加直观地分析本文方法的效果，我们选取了TIMIT数据集中的单条音频进行分割效果的展示。图3展示了一段特定音频的原始频谱图以及其中的音素边界，边界处为图中的白色虚线。其中（a）为本文方法的预测边界，（b）为真实标签中的边界，(c)为(Strgar and Harwath, 2023)中的最佳模型的预测边界。



(a) 本文方法的预测边界



(b) 真实边界



(c) 对比方法的预测边界

图3. 音素分割实例

通过对比分析这三个边界实例，可以明显观察到，在图3中我们标注出来的部分，（c）相较于真实边界，产生了若干不必要的预测边界。这种现象表明，尽管现有的方法在评价指标上表现出色，但它在预测时显著缺乏对音频序列时序信息的深入理解和准确捕捉。相反，本文提出的方法与真实的边界结果在大多数情况下保持了高度一致，这一比对结果不仅直观地证实了本文方法的有效性，也突显了它在理解音频序列的时序信息方面的明显优势。

此外，通过之前的实验验证，我们已经明确地证明了本文提出的方法在应用严格的评价指标时，能够取得相比其他现有方法更加显著的性能提升。特别是在解决音素边界的冗余预测问

题上，序列建模方法展现了其独特的有效性。这种成效归功于本文方法对音频数据内在时序依赖性的深刻理解和准确建模，确保了在保持预测精度的同时，有效避免了不必要的预测冗余。

5 结论

针对现有音素分割方法在捕获音频序列的时序依赖性和标签间关系方面的局限性，本文提出了一种融合语音自监督预训练模型和序列建模的音素分割策略。通过结合预训练的HuBERT模型和BiLSTM层，以及在序列层面优化的CRF损失函数，本文方法不仅优化了音素边界的识别准确性，还增强了模型对音频数据上下文信息的捕捉能力。实验结果表明，相较于传统的独立预测框架，本文方法在TIMIT和Buckeye数据集上均展现出了更优的性能，证明了序列建模对于提高音素分割任务准确性的重要作用。未来的工作将继续探索自监督学习和序列建模在音素分割及其他语音分割任务中的应用。此外，考虑到自监督预训练模型的强大表征能力，未来的研究还将探讨这类模型在跨语言和跨任务场景下的迁移学习能力，以期达到更广泛的应用和更优的性能。

参考文献

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518.

Alex Graves and Alex Graves. 2012. Long short-term memory. In *Supervised sequence labelling with recurrent neural networks*, pages 37–45. Springer.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Binghuai Lin and Liyuan Wang. 2022. Learning acoustic frame labeling for phoneme segmentation with regularized attention mechanism. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7882–7886.

Dac-Thang Hoang and Hsiao-Chuan Wang. 2015. Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America*, 137(2):797–805.

David Rybach, Christian Gollan, Ralf Schluter, and Hermann Ney. 2009. Audio segmentation for speech recognition using segment features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4197–4200.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. *arXiv preprint arXiv:2007.13465*.

Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi. 2020. Phoneme boundary detection using learnable segmental features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8089–8093.

Hyeongju Kim and Hyeong-Seok Choi. 2023. Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jan Chorowski, Grzegorz Ciesielski, Jarosław Dzikowski, Adrian Łańcucki, Ricard Marxer, Mateusz Opala, Piotr Pusz, Paweł Rychlikowski, and Michał Stypułkowski. 2021. Aligned contrastive predictive coding. *arXiv preprint arXiv:2104.11946*.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8167–8171.

Joerg Franke, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker, and Alex Waibel. 2016. Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium*, pages 1–5.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Xi Peng, Yang Xiao, and Zhiguo Cao. 2019. Roseq: Robust sequence labeling. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2304–2314.

John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, volume 1, number 2, page 3. Williamstown, MA.

John S Garofolo. 1993. TIMIT acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Joseph Keshet, David Grangier, and Samy Bengio. 2009. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329.

Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Dan Chazan. 2005. Phoneme alignment based on discriminative learning. In *INTERSPEECH*, pages 2961–2964.

Luke Strgar and David Harwath. 2023. Phoneme segmentation using self-supervised speech models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1067–1073.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. *NAACL HLT 2019*, page 48.

Okko Johannes Räsanen, Unto Kalervo Laine, and Toomas Altosaar. 2009. An improved speech segmentation quality measure: the r-value. In *Tenth Annual Conference of the International Speech Communication Association*.

Okko Räsanen. 2007. Speech segmentation and clustering methods for a new speech recognition architecture. *Helsinki University of Technology*.

Paul Michel, Okko Räsanen, Roland Thiolliere, and Emmanuel Dupoux. 2016. Blind phoneme segmentation with temporal prediction errors. *arXiv preprint arXiv:1608.00508*.

Sang-Hoon Lee, Ji-Hoon Kim, Hyunseung Chung, and Seong-Whan Lee. 2021. Voicemixer: Adversarial voice style mixup. *Advances in Neural Information Processing Systems*, 34:294–308.

Santiago Cuervo, Adrian Lancucki, Ricard Marxer, Paweł Rychlikowski, and Jan K Chorowski. 2022. Variable-rate hierarchical cpc leads to acoustic unit discovery in speech. *Advances in Neural Information Processing Systems*, 35:34995–35006.

Tusarkanta Dalai, Tapas Kumar Mishra, and Pankaj K Sa. 2023. Part-of-speech tagging of Odia language using statistical and deep learning based approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.