arXiv:2603.12642v1 [eess.AS] 13 Mar 2026

# 自监督语音模型通过位置相关正交子空间编码语音上下文

*Kwanghee Choi*[1], *Eunjung Yeo*[1], *Cheol Jun Cho*[2], *David R. Mortensen*[3], *David Harwath*[1]

[1] 德克萨斯大学奥斯汀分校 [2] 加州大学伯克利分校 [3] 卡内基梅隆大学{kwanghee,harwath}@utexas.edu

## 摘要

基于 Transformer 的自监督语音模型（S3Ms）通常被描述为具有上下文化特征，但其具体含义仍不清楚。本文聚焦于单个帧级 S3M 表示如何编码音段及其周围上下文。先前研究表明，S3Ms 以组合方式表示音段；例如，浊音性、双唇性和鼻音性等音系向量会叠加在 [m] 的 S3M 表示中。我们拓展这一观点，提出来自一系列相邻音段的音系信息也会以组合方式编码在单个帧中，使对应于前一个、当前和下一个音段的向量叠加在单个帧级表示中。我们表明，这种结构具有多种属性，包括不同相对位置之间的正交性，以及隐式语音边界的显现。总之，我们的发现加深了对上下文依赖型 S3M 表示的理解。[1]

**索引词**：自监督语音模型、可解释性、音系特征、语音上下文、上下文化

## 1. 引言

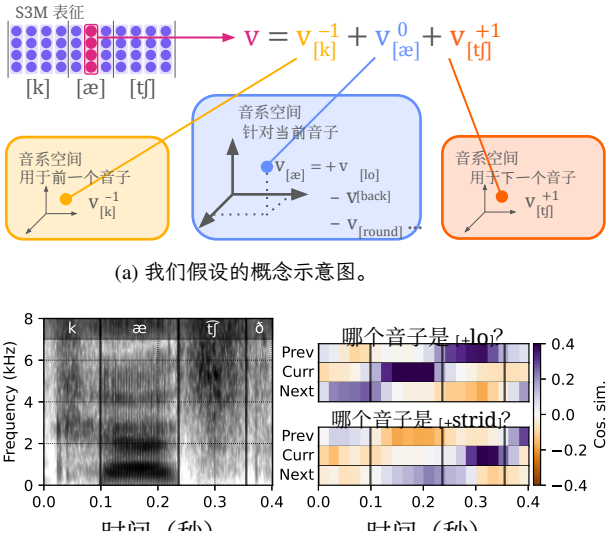
自监督语音模型（S3Ms）已成为一种核心范式，可从无标注音频中学习通用语音表征，并在广泛的下游任务中取得出色表现 [1–3]。这类模型的神经网络架构由 wav2vec [4] 首次提出，随后被后续的 S3Ms 广泛采用 [1–3]。该架构由两个主要组件构成：一个卷积网络，将原始波形转换为一系列潜在表示；以及一个基于 Transformer的上下文网络，能够整合跨时间的信息。

尽管这种架构已被证明非常有效，其内部机制仍未得到充分理解。先前的研究已确定 S3Ms 的不同层中存在哪些声学和语言信息 [5, 6]。然而，尽管涵盖了语言层级的多个层次，目前仍不清楚上下文网络如何在表征空间中组织这些信息。这一研究空白促使人们日益关注

旨在刻画 S3M 学习到的表示的内部结构的一系列研究 [7–14]。

[5] 已表明，S3M 的卷积网络主要捕捉局部声学信息。与声谱图类似，有限的感受野确保了时间局部性，使每一帧都与输入信号的特定区域对齐。此外， [9]指出，与各维度对应固定频带的声谱图不同，卷积表征空间是分布式且相对的。在这个空间中，声学

[1]所有实验代码均可在 https://github.com/juice500ml/phonetic-arithmetic 获取。




(a) 我们假设的概念示意图。

(b) 我们假设的实证展示。

图 1： *(a)* 每个帧级*S3M*表示都包含来自位置敏感的正交子空间的音系向量。*(b)* 帧级表示与前一个、当前及下一个音段子空间中的音系向量（低、刺耳音）之间的余弦相似度。相对于语音边界，每一帧都会编码来自当前音段及相邻音段的音系信息。

相似度体现为表征之间的余弦相似度，而非沿着固定轴体现。这些发现表明，尽管卷积网络能够提供局部声学信息，但其本身无法解释 S3M 如何整合局部声学范围之外的信息。

尽管关于卷积网络的图景相对清晰，但基于Transformer的上下文网络所起的作用仍未得到充分理解。实证结果表明，上下文网络至关重要。与卷积表示相比，较后层Transformer层的表示在音段和词身份等语音探测任务中始终表现更佳 [6], ；而结合Transformer各层的信息通常还能进一步提升性能 [15]。此外，扩大上下文网络的规模通常会带来显著提升，大型模型的表现始终优于基础模型 [1–3]。这些趋势表明，上下文网络的作用不只是对局部声学信息进行简单改进。

另一条线索来自离散化 S3M 表征在各种语音应用中的有效性

[16], 其中，对 S3M 表征进行 k均值聚类，会得到与语音类别相对应的离散单元，尽管这一过程没有监督 [1, 2, 14, 17]。此外，这些单元往往按照语音类别组织。例如，属于同一类别的音段之间的相似度高于属于不同类别的音段之间（例如，元音与辅音） [13, 18]。最后，在注意力图中也观察到了语音边界和音节边界 [19], ，这使得 S3Ms 可用于无监督语音分割 [20–24]。这些观察结果表明，上下文网络编码的是具有语言学意义的区别，而不只是反映细粒度的声学相似度。

为了阐释这类具有语言结构的表示是如何产生的，近期研究提出，S3M会编码音系向量 [11]。根据这一观点，单个音段由音系向量的线性组合表示，从而形成组合表示空间。例如， [æ] 的S3M表示包含音系向量的线性组合，其中包括低、前和不圆唇向量，并支持向量运算。尽管这一框架为上述音系能力提供了连贯的解释，但它只考察彼此孤立的单个音段，并未探究Transformer如何捕捉来自相邻音段的语音上下文。

这促使我们通过提出以下假设来扩展其框架：每个帧级*S3M*表示编码的不仅是单个音段，还包括音段序列（图 1）。这一扩展旨在阐明S3M如何编码上下文信息。具体而言，我们提出，每个上下文化的帧级表示都包含音系向量的线性组合，这些向量不仅对应于与该帧对齐的音段，也对应于其相邻音段。基于这一假设，我们提出以下四项预测。

**预测1(§ 3.1)**。单个帧级表示将呈现组合音系结构。我们评估表征空间中的向量运算是否反映音系类比例如，将浊音向量 [d] − [t] + [p] ≈ [b] 加到清塞音上，能否近似得到浊塞音。虽然 [11] 依赖均值池化表示，但我们评估此类音系类比是否也适用于单个帧表示。

**预测2（§ 3.2）**。单个帧级表示还将编码相邻音段的音系信息。为此，我们将上述音系类比测试扩展到相邻音段，考察单个帧表示是否支持针对前一个或下一个音段的音系类比。具体而言，我们测试表征空间中是否存在能够编码上下文音系属性的向量，例如“前一个音段是否为浊音？”或“下一个音段是否为高元音？”，以及能否通过对单帧表示进行线性运算来恢复此类属性。

**预测3（§ 3.3）**。位置依赖子空间彼此正交，使模型能够区分哪些信息对应于各个相对位置，例如前一个、当前或下一个音段。例如，如果当前音段和前一个音段都是浊音，那么仅仅将共享的音系向量相加所得的表示，无法区分浊音性源自当前音段还是前一个音段。为解决这种冲突，情境化表示应以位置依赖的方式编码音系信息，使与不同相对位置相关联的音系向量分别处于不同的子空间中。因此，我们检验与不同相对位置相关联的音系向量是否在单帧表示中以正交方式编码。

**预测4（§ 3.4）**。由于底层存在位置相关的子空间，帧级表示会随时间发生变化，从而隐含地标示出语音边界。为了区分前一个、当前和下一个音段，模型可能会隐式推断语音边界，并据此对其帧级表示进行分段。另一种可能是，位置相关的音系信息忽略语音边界，而是在固定的对称窗口 [8, 25]上运作。为了区分这两种可能性，我们考察帧级表示与位置相关的音系子空间之间的余弦相似度如何随帧变化，尤其是在语音边界附近。

总之，我们旨在阐明 S3Ms 如何编码语音上下文，即音段序列，而非单个音段。我们假设，帧级 S3M 表示以组合性和位置敏感的方式编码音系信息，并对以下预测进行检验：

- **帧级组合性**：单个帧表示可以表示为描述与该帧对齐的音段的音系向量的线性组合（§ 3.1）。
- **上下文音系向量**：单个帧级表示编码了关于相邻音段的音系信息（§ 3.2）。
- **位置正交性**：与相对位置（前一个、当前和下一个音段）相关的音系向量位于彼此正交的子空间中（§ 3.3）。
- **语音分割**：帧级表示中的位置相关音系子空间由语音边界决定（§3.4）。

## 2. 设置

### 2.1. 自监督语音模型（S3Ms）

我们分析了三种广泛使用的、以英语训练的单语 S3M 的表示：wav2vec 2.0 [1], HuBERT [2],和 WavLM [3], ，采用公开发布的大型检查点。[2] wav2vec 2.0 和 HuBERT 主要在有声读物上进行预训练，而 WavLM 使用了朗读语音和自发语音。三种模型都采用相同的架构，由卷积特征编码器和后接的基于 Transformer 的上下文网络组成。

遵循 [11], ，我们使用谱表示、对数梅尔频谱图（梅尔频谱）和梅尔频率倒谱系数作为基线。我们使用 librosa 实现 [26], ，窗口长度为 2048，跳跃长度为 512；梅尔频率倒谱系数不使用一阶和二阶差分系数。遵循 [5,6], ，我们对每个音素片段内的帧进行均值池化，以提取音子级表示，并将相同的池化过程应用于谱表示和 S3M 表示，作为第 3.1 节和第 3.2 节的基线。

### 2.2. 通过音系类比分析 S3M

**音系类比。** 基于音系类比的分析 [11, 27, 28] 用于检验音系关系是否对应于表征空间中的线性运算。给定一组构成一致音系对立的四个音段，该分析会评估两个音段之间的差异（例如， [b] − [d]）能否加到另一个音段上（例如，

[2]我们使用了 facebook/wav2vec2-large-lv60、facebook/hubert-large-ll60k 和 microsoft/wavlm-large 这几个检查点。

[t]），以近似预期目标（例如，[p]），从而使共同的音系差异通过向量减法体现出来。音系类比使用 PanPhon 自动构建，[29] 而音系对立可能涉及多个 PanPhon 特征差异。例如，[b] 和 [t] 的差异在于 [浊音]，[冠音]，[弥散]，和 [唇音]。

**成功率**。为了量化此类音系类比是否成立，我们采用 [11] 的成功率指标。对于每个类比，我们通过向量运算来近似目标音素的表示（例如，使用 [b] − [d] + [t] 近似 [p]）。随后，我们将目标表示与近似表示之间的余弦相似度（例如，cos([p], 近似的 [p]）与上下基线进行比较。上基线由不同话语中相同音子之间的相似度给出（例如，来自不同话语的 cos([p], [p] ）。下基线则由不同音子之间的相似度给出（例如，cos([p], 任意非 [p] 音子）。若平均余弦相似度介于这两个基线之间，则认为该类比成功；成功率定义为满足这一排序关系的类比所占比例。对于 §§ 3.1 和 3.2 中的实验，我们采用与 [11] 相同的超参数：有放回地随机抽取 1,000 个样本以估计平均余弦相似度，并基于 10 次独立重复计算 99% 置信区间。

### 2.3. 数据集

我们在 TIMIT [30]和 VoxAngeles[31], 这两个经过语音标注和人工分段的语音数据集上开展实验。TIMIT 包含句子级英语朗读语音，而 VoxAngeles 则包含来自 95 种 S3M 训练期间未见语言的词级朗读语音。按照 [11], 的做法，我们排除没有 PanPhon映射的音段，使音段清单从 47 个缩减为 47 个，再缩减为 44 个（TIMIT），并从 567 个缩减为 468 个（VoxAngeles）。这些筛选后的集合用于 §§ 3.3、3.4 和 4.1。此外，为确保成功率计算可靠，我们进一步排除出现次数少于 50 次的音段，最终 TIMIT 有 43 个音段，VoxAngeles 有 57 个音段。由此分别得到 236 个和 468 个有效四元组，用于 §§ 3.1 和 3.2。除非另有说明，我们使用 TIMIT 测试划分和完整的 VoxAngeles 数据集。

## 3. 实验

### 3.1. 帧级组合性

**设置**。为了评估帧级是否存在音系组合性，我们比较均值池化 [5, 6, 11]
到中心池化 [10, 12, 20]。一个音素片段通常跨越多个帧。均值池化在整个片段上聚合帧级表示，而中心池化仅使用时间上的中心帧。我们评估与均值池化相比，单个中心帧是否也能捕获音系信息。

**结果**。如图2所示，在不同数据集和 S3M 中，中心池化表示的成功率与均值池化表示相当，且往往更高；在某些层中，两者的表现都显著优于频谱基线。这一结果表明，音系组合性并不依赖于对音子片段进行时间平均，即均值池化。相反，单个帧级表示就能支持线性音系类比，这与我们关于帧级音系组合性的假设一致。

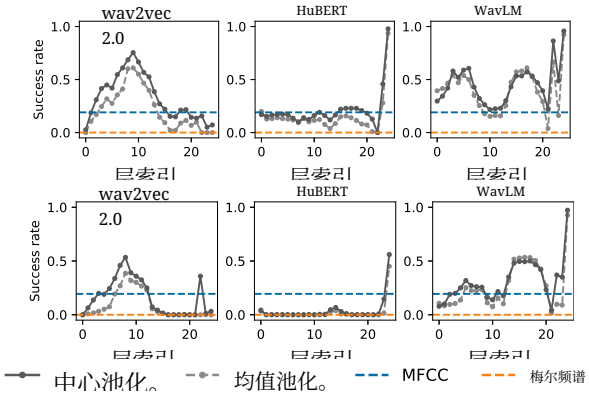



图2：比较*TIMIT*（上）和*VoxAngeles*（下）上均值池化与中心池化的音系类比成功率。中心池化与均值池化相当，且往往优于均值池化，表明音系组合性存在于单个帧表示层面。

### 3.2. 上下文音系向量

*3.2.1.* 中心帧编码了多少邻近上下文?

**设置**。对先前分析的自然延伸是询问单个帧级表示中编码了多少相邻音段的音系信息。具体而言，我们研究中心池化表示是否不仅支持关于当前音子的音系类比，也支持关于相邻音段的音系类比。

考虑一个音段序列 $[p^{-2}, p^{-1}, p^{0}, p^{+1}, p^{+2}]$。对于每个序列，我们从中心音素 $p^0$ 提取单个中心池化表示，并用该表示探测与每个相对位置相关的音系信息。这使我们能够检验 $p^0$的中心帧是否编码了对应于$p^0$之外音段的音系向量。例如，为了检验来自$p^0$ 的中心表示是否包含前一个音段 $p^{-1}$的浊音向量，我们比较 $p^{-1}$ 为浊音或清音音素的语境中的表示。

**结果**。如图3所示，中心池化的 S3M 表示不仅支持对当前音子进行音系类比，也支持对其紧邻的音段 $p^{-1}$、$p^0$ 和 $p^{+1}$ 进行音系类比。具体而言，S3Ms 的较后层尤其呈现出非零成功率，这表明，来自中心音素的单个帧级表示编码了相邻音段的音系信息。

*3.2.2.* 音系类比的有效窗口大小

**设置**。为进一步确定中心音素 $p^0$ 的音系信息编码位置，我们评估从音段 $p^{-2}$ 到 $p^{+2}$ 内不同时间位置的帧中提取的表示。给定与长度为 $n$ 帧的音素片段对齐的帧级表示$\mathbf{r}[1], \mathbf{r}[2], \ldots\ldots, \mathbf{r}[n]$（该片段可以来自 $p^{-2}$ 到 $p^{+2}$ 的任意音段），我们执行随机池化：

$$\mathbf{r}^{\text{random}} = \mathbf{r}[i] \text{ where } i \sim \text{Unif}\{1, \ldots, n\}. \quad (1)$$

我们按每个音素片段内的归一化位置 $i/n$ 对表示进行分箱，并汇总各箱的结果。对于 VoxAngeles，与 TIMIT 不同，按相对位置分箱会使可用的音系四元组数量减少

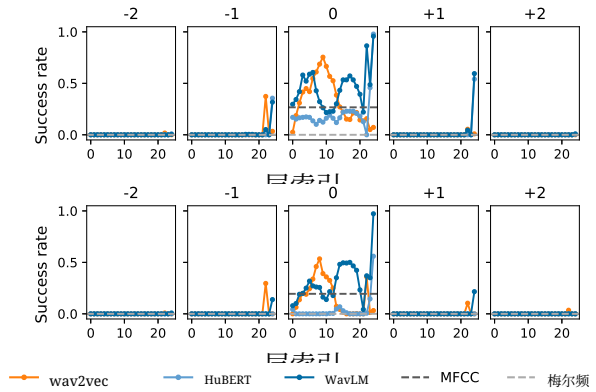



图3：在 *TIMIT*（上）和 *VoxAngeles*（下）上，探测编码于 $p^0$ 单个帧级表示中的上下文信息时的音系类比成功率。中心池化的 *S3M* 表征支持当前音子位置（0）及其相邻位置（±1）的音系类比。

从 468 个降至 128 个。根据 § 3.2.1，我们选择在 $p^0$上表现最好的 S3M 层：HuBERT 和 WavLM 的第 24 层，以及 wav2vec 2.0 的第 9 层。

**结果**。图4展示了帧级表示能够支持中心音素 $p^0$音系类比的有效窗口。在所评估的表示中，WavLM 在各数据集上呈现出最清晰的梯形模式。对于位置 0，成功率始终保持高；对于紧邻位置 ±1 则下降，而对于更远的位置 ±2 则趋近于零。

HuBERT 在 TIMIT 上呈现出类似模式，但这一行为在 VoxAngeles 上有所减弱。相比之下，wav2vec 2.0 在两个数据集上的表现都有限。最后，正如预期，频谱基线仅在位置 0 有效。

从协同发音的角度来看，S3M 表征可能会整合来自相邻音段的音系信息，以支持对中心音素的重建；§ 4.2 将对此进行详细讨论。我们还观察到，成功率似乎会随语音边界而变化，我们将在 § 3.4 中对此进行更直接的分析。

### 3.3. 位置正交性

**设置**。§ 3.2 中的结果表明，当前音子 $p^0$ 中的单帧级表示可以编码有关相邻音段 $p^{\pm 1}$的音系信息。这会带来一种歧义：如果当前音段和相邻音段的音系特征相同，该表示就必须区分它们的位置来源。我们假设，与不同相对位置相关的音系信息编码在彼此正交的子空间中。

为检验这一假设，我们遵循 [11]，并通过均值差提取音系向量。例如，与当前音子相关的浊音向量计算如下：

$$\mathbf{v}^0_{\text{voicing}} = \mathbb{E}_{p^0 \text{ is voiced}}[\mathbf{v}] - \mathbb{E}_{p^0 \text{ is unvoiced}}[\mathbf{v}], \qquad (2)$$

其中$\mathbf{v}$ 表示通过中心池化获得的帧级表示。我们不局限于最小对，而是对所有具有或不具有该特征的音段的表示取平均

目标音系特征。类似地，我们提取一个浊音性

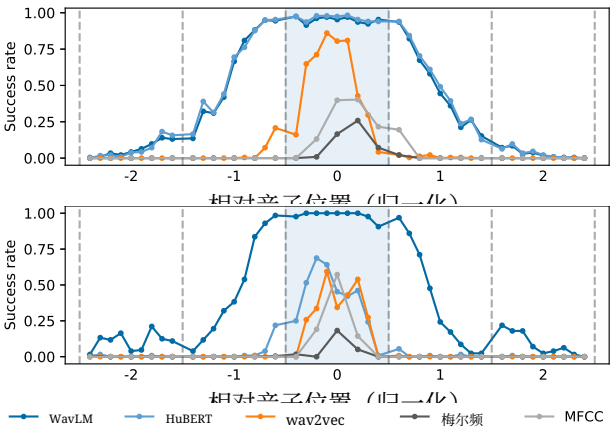



图4：在 *TIMIT*（上）和 *VoxAngeles*（下）上，中心音子的音系类比成功率$p^0$ 与相对位置的关系。*WavLM* 展现出最宽的窗口，在中心音子位置 0 处具有较高的成功率，随着 ±1 的增加而降低，并在 ±2 处接近于零。与 *S3Ms* 不同，谱表示仅在中心附近显示出非零的成功率。

向量，用于表示前一个音段 $p^{-1}$:

$$\mathbf{v}^{-1}_{\text{voicing}} = \mathbb{E}_{p^{-1} \text{ is voiced}}[\mathbf{v}] - \mathbb{E}_{p^{-1} \text{ is unvoiced}}[\mathbf{v}]. \qquad (3)$$

如果位置正交性成立，这些向量应当近似正交，即 $\cos(\mathbf{v}^{0}_{\text{voi.}}, \mathbf{v}^{-1}{}_{\text{voi.}}) \approx 0$，从而使单个帧级表示能够编码来自多个位置的音系信息而不会产生混淆。

在 [11], 之后，我们从四个元音特征（高、低、后、圆唇）和四个辅音特征（鼻音、响音、刺耳音、浊音性）中计算八个音系向量。对于元音特征，仅使用元音提取音系向量；对于辅音特征，仅使用辅音提取向量。所有分析均使用 WavLM-Large 的最终层，该模型在第 3.2 节中始终取得最强的性能。我们使用数据集的训练集划分来计算音系向量。对于 VoxAngeles，训练集划分是通过从数据集中随机选择部分语言创建的。

**结果**。我们首先考察从当前音子中提取的音系向量之间的余弦相似度，这些向量基于中心池化的帧级表示（图 5）。所得相似度结构与 [11] 中报告的结果高度吻合，后者采用了均值池化。值得注意的是，该结构与语言学直觉十分一致。例如，高和低 等相反特征呈现出强负相似度，而鼻音、响音和浊音性等相关特征则呈现正相似度。此外，元音和辅音特征近似正交。

接下来，我们比较不同相对位置上的音系向量（图 6）。与同一位置的向量相比，来自不同相对位置的音系向量彼此之间的余弦相似度显著更低，表明不同位置之间近似正交。此外，不同位置上的音系特征间相对相似度结构似乎得以保留（$-2$，$-1$，$0$，$+1$，$+2$），这表明每个位置都实例化了一个相似的音系子空间。总之，这些发现支持我们的假设：S3Ms 在依赖位置的正交子空间中编码音系信息。

有趣的是，位置结构不仅延伸至紧邻音段（±1），也延伸至距离更远的音段（±2），尽管

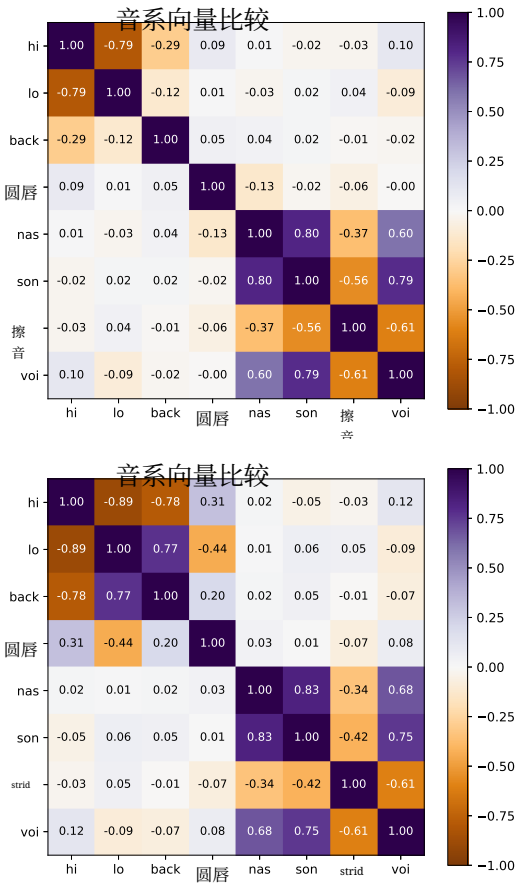



图 5： *TIMIT*（上）和 *VoxAngeles*（下）上从帧级 *S3M* 表示中提取的音系向量之间的余弦相似度。其结构与 *[11]*, 相呼应：相反特征之间的相似度为负，相关特征之间的相似度为正。

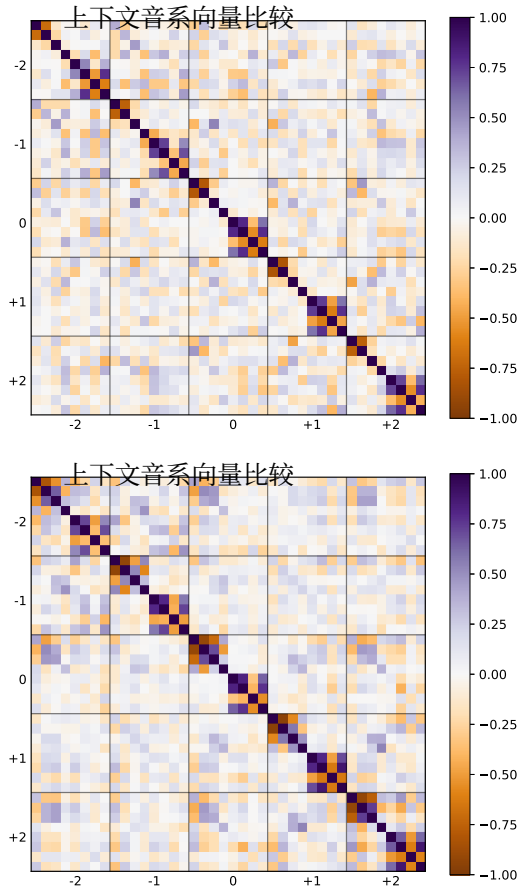



图 6： *TIMIT*（上）和 *VoxAngeles*（下）中与不同相对音素位置（$-2$ 至 $+2$）相关联的音系向量之间的余弦相似度。与图 *5* 相比，各位置内部保留了相对相似度结构。此外，不同位置的向量之间的相似度显著低于相同位置的向量，表明不同位置近似正交。

单个帧级表示的成功率几乎为零（§ 3.2）。这种表面上的矛盾表明，成功率分析可能低估了上下文信息的存在，因为单个帧表征可能会纠缠多个音系成分，导致在单帧级别应用音系算术时只能得到含噪的近似结果。与这一解释一致，图 7 显示，音系向量的范数会随着与中心音素的距离增加而减小，符合 $0 > \pm 1 > \pm 2$。

最后，我们考察这种位置结构是否仅存在于 WavLM 的最后一层，还是也贯穿不同模型和层（图 8）。我们比较了同位置配对和跨位置 配对的平均绝对余弦相似度。前者由来自同一自然类（元音或辅音）且相对位置相同的音系向量构成（$-2$，$-1$，$0$，$+1$，$+2$）；后者则包括其余所有组合。在所有模型和层中，同位置相似度始终高于跨位置相似度，这表明位置正交性在各层中大致得以保持。

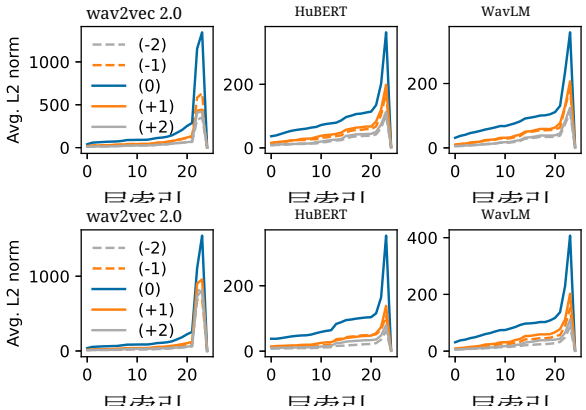



图 7： *TIMIT*（上）和 *VoxAngeles*（下）上各层中八个音系向量在各自相对位置上的平均 *L2* 范数。我们观察到，向量的平均大小随着其与中心音素距离的增加而减小。

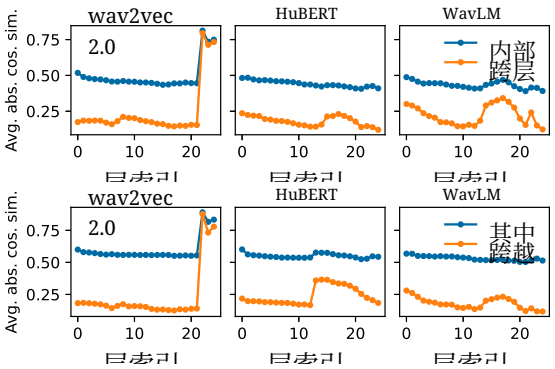



图 8：观察各层的正交程度。我们观察了 *TIMIT*（上）和 *VoxAngeles*（下）中子空间内和跨子空间余弦相似度绝对值的平均值。此处，我们将“同一子空间内”定义为属于相同自然类（元音或辅音）且处于相同位置（$-2$、$-1$、$0$、$+1$ 或 $+2$）。

### 3.4. 语音分割

**设置。** § 3.3 中观察到的位置正交性表明，S3M 表征应对语音边界敏感，以便选择恰当的位置相关音系子空间。如果音系信息是相对于前一个、当前和下一个音段编码的，那么模型应能隐式追踪语音边界。

为检验这一假设，我们分析语音边界附近的帧级表示。对于每个边界，我们提取十一个帧级表示：边界前的五帧（$\mathbf{r}[0],\ldots,\mathbf{r}[4]$）、边界帧本身（$\mathbf{r}[5]$）以及边界后的五帧（$\mathbf{r}[6],\ldots,\mathbf{r}[10]$）。对于每个帧级表示，我们计算其与前一个（$\mathbf{v}^{-1}$）、当前（$\mathbf{v}^{0}$）和下一个音段（$\mathbf{v}^{+1}$）所对应的音系向量之间的余弦相似度。我们考虑两类边界：（i）边界起始点，即 $p^{-1}$ 与 $p^{0}$ 之间的边界；以及（ii）边界终止，即 $p^{0}$ 与 $p^{+1}$ 之间的边界。

例如，考虑一个边界起始点，其中前一个音段是清音（[-voi]），当前音段是浊音（[+voi]）。如果模型以对边界敏感的方式编码音系信息，那么与当前音段相关联的浊音向量 $\mathbf{v}^{0}_{\text{voi.}}$，应当比起边界前帧（$\mathbf{r}[0],\ldots,\mathbf{r}[4]$）更好地与边界后帧（$\mathbf{r}[6],\ldots,\mathbf{r}[10]$）对齐。换言之，我们预期 $\cos(\mathbf{v}^{0}_{\text{voi.}},\mathbf{r}[i])$（其反映了当前音段 $p^{0}$ 是否为浊音）在边界后帧中会高于边界前帧。

在同一个例子中，考虑边界起始处与下一个音段相关联的浊音向量 $\mathbf{v}^{+1}_{\text{voi.}}$。由于 $p^{0}$，即，$p^{-1}$ 的“下一个”音段是 [+voi]，，我们预期对于边界前帧（属于 $p^{-1}$），$\cos(\mathbf{v}^{+1}_{\text{voi.}},\mathbf{r}[i])$ 的值会高于边界后帧（属于 $p^{0}$）。换言之，$\cos(\mathbf{v}^{+1}_{\text{voi.}},\mathbf{r}[i])$ 量化了“下一个”音段为浊音的证据。类似地，对于边界结束处与前一个音段相关联的浊音向量 $\mathbf{v}^{-1}_{\text{voi.}}$，$p^{0}$，即，$p^{+1}$ 的前一个音段是 [+voi]。因此，

$\cos(\mathbf{v}^{-1}_{\text{voi.}},\mathbf{r}[i])$ 对于边界后帧（即 $p^{+1}$ 的那些帧）应高于边界前帧（即 $p^{0}$ 的那些帧）

我们使用 WavLM 最后一层的表征，在 TIMIT训练集上进行此分析。我们汇总语音边界处相关音系特征发生变化的情况（例如，对于边界起始点，[-voi] 到 [+voi]；对于边界终止点，[+voi]到 [-voi]——

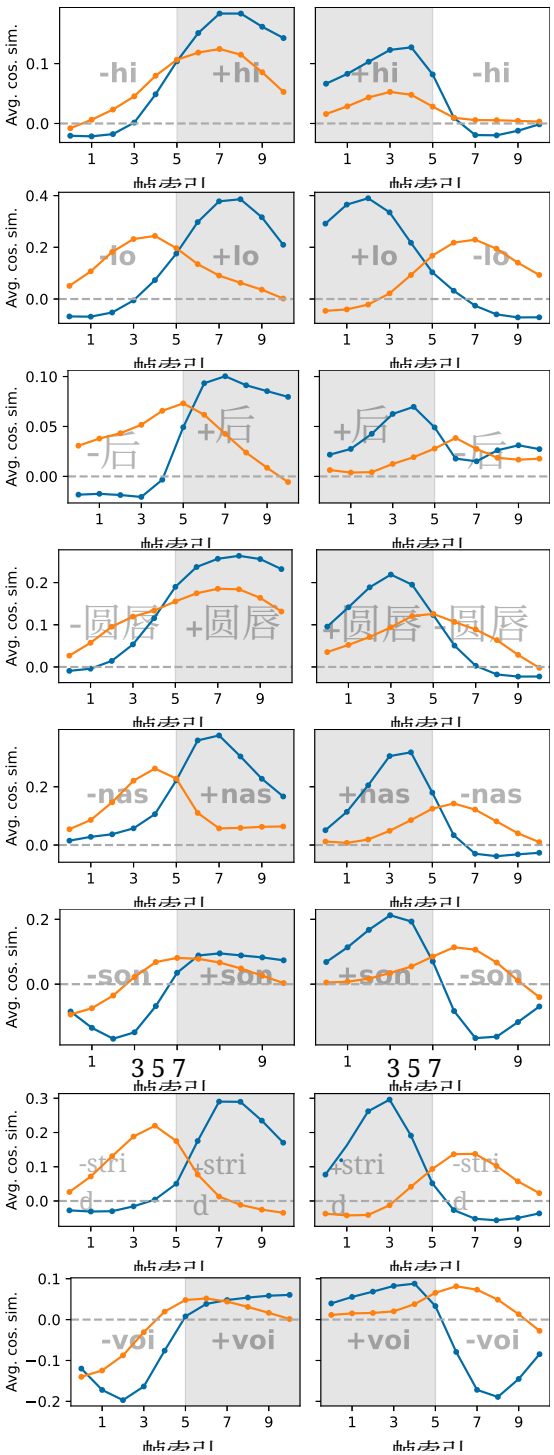



图9：在*TIMIT*上进行的语音分段实验，其中帧索引*5*表示*TIMIT*的语音边界。语音边界附近帧级 *S3M* 表示与位置依赖的音系向量之间的余弦相似度。例如，对于清音 *[-voi]* 与浊音 *[+voi]*，之间的边界，左图（边界起始点）比较当前音子为浊音的证据（$\cos(\mathbf{v}^{0},\mathbf{r}[i])$；蓝色）与后续音子为浊音的证据（$\cos(\mathbf{v}^{+1},\mathbf{r}[i])$；橙色），而右图（边界终止）比较当前音子为浊音的证据（$\cos(\mathbf{v}^{0},\mathbf{r}[i])$；蓝色）与前一音子为浊音的证据（$\cos(\mathbf{v}^{-1},\mathbf{r}[i])$；橙色）。

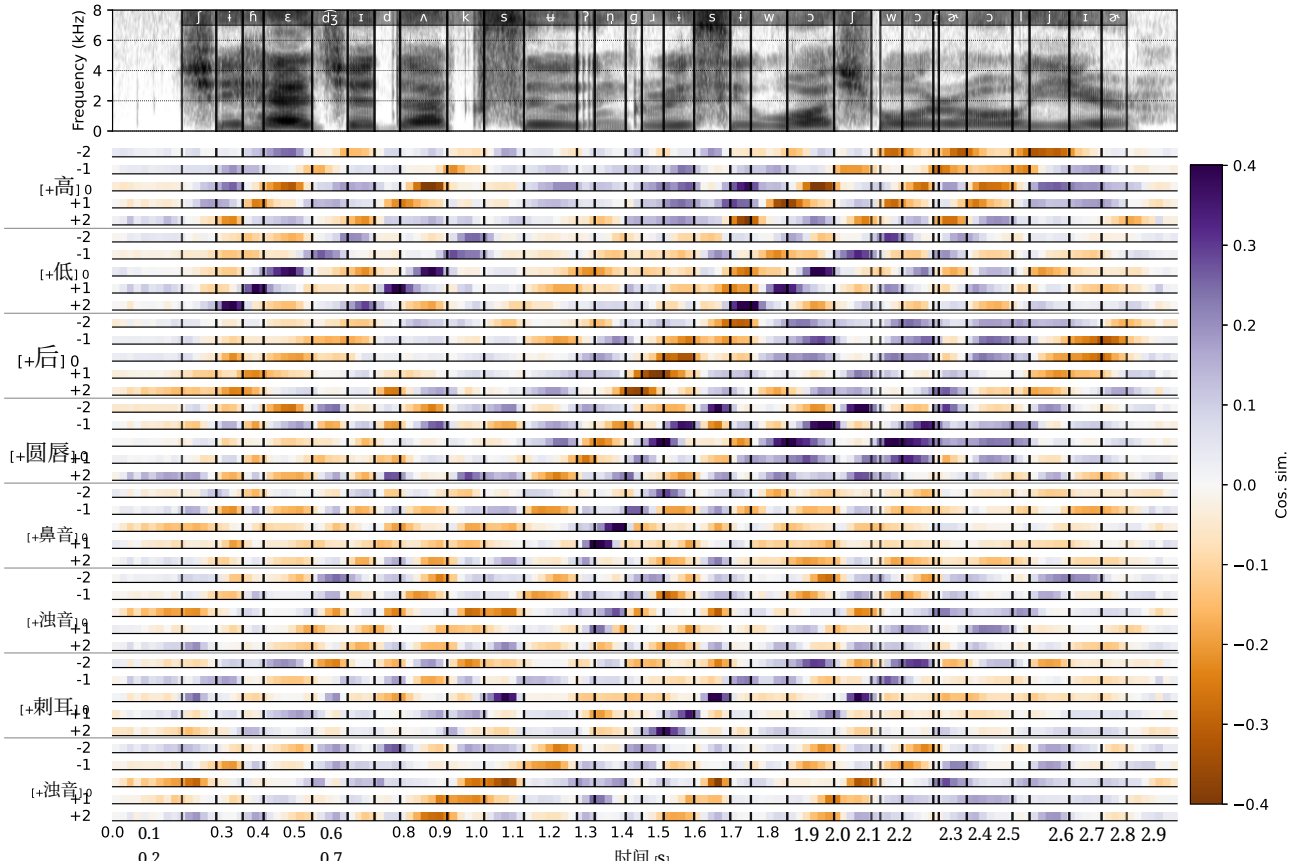



图10： 在一段完整的 *TIMIT* 话语中，相对音素位置（$-2$至 $+2$）对应的帧级 *S3M* 表示与音系向量之间的余弦相似度。阶梯状过渡通常与语音边界对齐，展示了特定位置的正交音系子空间。

点）。对于元音特征（高、低、后、圆唇）的音系向量，我们选择 $p^0$ 作为元音。类似地，对于辅音特征（鼻音、响音、刺耳音、浊音性），我们选择 $p^0$ 作为辅音。所报告的曲线显示了所有此类边界实例的余弦相似度均值。音系向量取自 § 3.3，使用 TIMIT训练集计算得出。

**结果**。 如图9所示，对于 TIMIT 的边界起始点（左）和结束点（右），两个余弦相似度相交处对应的帧均与标注的语音边界重合。这些观察表明，S3Ms 大致按照语音边界对帧级表示进行分段，其位置依赖子空间是基于音素段而非固定窗口构建的。

# 4. 讨论

## 4.1. 音系向量的定性示例

图1展示了与单个单词（“catch”）对应的短片段中，帧级 S3M 表示与音系向量之间的余弦相似度；现在，我们将这一可视化扩展到一段完整的 TIMIT 语句（“She had your dark suit in greasy wash water all year.”）。具体而言，图10展示了与相对音素位置 $-2$ to $+2$ ($p^{-2}$, $p^{-1}$, $p^0$, $p^{+1}$ $p^{+2}$ ,)相关的音系向量的余弦相似度。虽然我们展示了第3.3节和第3.4节中使用的八种音系特征，但同样的可视化也适用于 PanPhon 提供的其他音系特征。

在整段语句中，我们观察到清晰的阶梯状结构，与语音边界相吻合，而非固定窗口。例如，对于元音 [ɛ]，这一模式在 ［+低］音系向量中尤为明显，而在 ［+高］向量中则呈现相反的形式，这与二者的音系对立相符。尽管某些片段中的变化较弱或不够规律，但整体模式仍然清晰。

此外，阶梯状结构横跨相对位置 $-2$ 至 $+2$，这与 § 3.3 一致。该可视化进一步表明，§ 3.2 中基于成功率的分析低估了这一效应。

## 4.2. 逐层掩码填充行为

**设置**。 S3Ms 的训练目标是根据周围上下文 [1–4], 重建输入中的掩码部分，这促使表示编码有关相邻语音的信息。鉴于语音的协同发音特性，这一训练目标表明，应利用相邻音段的音系特性，即也就是说，周围语音环境 [12],, 以支持准确的掩码填充。

为考察这种基于上下文的预测行为如何随层变化，我们分析了掩码填充行为的逐层差异。具体而言，我们计算掩码部分中从原始信号提取的表示与从掩码信号提取的表示之间的余弦相似度。

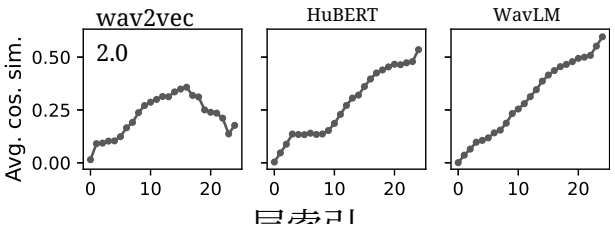



图 11： *TIMIT* 上的逐层掩码填充行为。余弦相似度越高，表示根据周围上下文恢复掩码表示的准确度越高。

相似度越高，表明从周围上下文中恢复掩码表示的准确度越高。

我们使用 TIMIT 测试划分，并采用每个模型的默认掩码配置（掩码长度为 10 帧，约 205 毫秒）。由于已知 Transformer 表示具有各向异性 [32], ，我们在计算余弦相似度之前进行 ZCA 白化 [33]。白化统计量根据从训练集中随机抽取的 100 个话语进行估计。

**结果**。 如图 11 所示，HuBERT 和 WavLM 从原始信号与掩码信号中提取的表示之间的余弦相似度，在所有层中持续上升。相比之下，wav2vec 2.0 在中间层达到峰值，之后则有所下降。我们推测，这些差异可能与 § 3.1 和 § 3.2 中观察到的逐层模式有关，也可能与先前研究的发现有关 [5, 6], 或许反映了 S3Ms 之间训练策略的差异 [34]。综合来看，对 HuBERT 和 WavLM 这一行为的一种可能解释是，所有层在很大程度上都由掩码填充目标驱动。由于掩码方式具有对称性，给定位置的表示可能会编码有助于预测当前音子之前或之后可能出现哪些音段的信息。此外，由于协同发音，准确地进行掩码填充可能需要有关相邻音段的信息。我们将更具体的探究留待未来工作。

### 4.3. 与先前研究观察结果的联系

**可解释性**。 先前关于 S3M 可解释性的研究主要关注逐层表示编码了什么 信息 [5, 6], ，例如频谱信息 [5, 9], 、语音信息 [10, 35, 36], 、音位信息 [36], 、发音信息 [37, 38], 、音节信息 [21–23], 、词汇信息 [10, 20, 39], 、句法信息 [40], 和语义信息[10,20]。相比之下，我们的研究关注如何 在 S3M 表征中组织和构建语音信息。

**音系向量**。 [11] 提出，S3Ms 通过音系向量的线性组合编码音系信息。他们对若干实证观察结果给出了有原则的解释，包括为何 S3Ms 以相对相似性的形式编码语音信息 [9, 10], 、为何从 S3Ms 得出的离散单元往往与语音单元表现相似 [1, 2, 14, 16], ，以及为何 S3M 表征中的相似性结构往往反映自然类 [13, 18]。我们扩展了这一框架，并进一步展示了位置依赖的正交音系子空间的存在。这一视角统一解释了先前的观察结果：词级 S3M 相似性与词语之间的语音相似性密切吻合 [10, 20], ，而且 S3Ms 编码了有关周围语音环境（即音位变体）的信息 [12]。

**上下文相关三音素隐马尔可夫模型**。在深度神经网络出现之前，自动语音识别系统通常会对协同发音进行建模，采用上下文相关模型 [41,42]。我们的结果表明，S3Ms 可能会通过位置依赖的音系向量重新发现类似的上下文结构。

**有效上下文窗口**。 § 3.2 中的结果与先前关于 S3M 有效时间上下文大小的分析相关 [8, 25]。以往研究通常将上下文化描述为固定的时间窗口，而我们的结果表明，S3M 表示是围绕语音边界组织的，同时位置依赖子空间会根据其与中心音素的接近程度进行加权（图 7）。

**无监督分割**。 § 3.4 中的语音分割分析为无需监督即可产生类似词和音节边界的机制提供了新的视角 [19–24]。通过证明 S3M 表示隐式编码了与语音边界对齐的位置依赖结构，我们确定了一种可用于无监督音子级和音节级分割的具体信号。

### 4.4. 对未来工作的启示

**可解释性**。 我们的发现与 [7,43,44], 密切相关，该研究指出，S3Ms 中的说话人信息编码于正交子空间中，并且可以用线性向量进行表征。结合我们的结果，这表明 S3Ms 中存在一种更广泛的表征原则：不同来源的信息（例如，音系、说话人或声学信息）可能被组织到结构化的解耦子空间中。未来研究的一个重要方向是刻画这些信息如何在表征空间中组织和相互作用。

**离散语音单元**。 我们的结果表明，从 S3M 中提取的离散单元 [16] 与神经音频编解码器 [45]学习到的声学词元之间可能存在差异。众所周知，S3M 中的离散单元天然具有更高的结构连贯性 [46] ，且自动语音识别性能 [16] 优于声学词元。我们假设，这种差异源于对语音上下文的编码。一个有前景的方向是开发新的架构和训练目标，以解耦与上下文无关和依赖上下文的表示 [47–49]。此外，§ 3.3 表明，多个相对位置的音系结构相似，这也带来了设计更具结构性的神经网络架构的可能性，此类架构可显式建模这一特性。

**可解释语音表征**。 § 4.1 中观察到的音系结构与语音和音系后验图 [50–52] 以及发音反演 [53]自然相关。这些联系表明，类似图10的表示可作为可解释且具有语言学基础的中间表征，应用于声码器和文本转语音模型等领域。

## 5. 结论

在本研究中，我们探讨了 S3M 如何在单个帧表征层面进行上下文建模。通过一系列分析，我们发现 S3M 以组合方式编码音素序列的音系信息。单个帧通过音系向量反映当前音段及相邻音段的特性。此外，我们发现这些向量位于相对位置相关的正交子空间中。最后，我们证明，这种结构会产生涌现的语音边界。综合而言，这些发现有助于统一理解 S3M 的基于Transformer的上下文网络如何编码语音上下文。

## 6. 致谢

作者谨此感谢 Stephen McIntosh 和 Ian Shih 提供的宝贵反馈。

## 7. 生成式人工智能使用声明

本文撰写过程中使用了生成式人工智能工具，主要用于代码自动补全，以及对措辞和语法进行少量编辑。所有科学内容和代码实现均由作者独立开发、验证并最终定稿。

## 8. 参考文献

[1] A. Baevski、Y. Zhou、A. Mohamed 与 M. Auli，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，载于 NeurIPS，2020。[2] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai 等，“HuBERT：通过隐单元掩码预测进行自监督语音表征学习”，IEEE/ACM TASLP，第29卷，2021。[3] S. Chen、C. Wang、Z. Chen 等，“WavLM：面向全栈语音处理的大规模自监督预训练”，IEEE 信号处理选题期刊，第16卷，2022。[4] S. Schneider、A. Baevski、R. Collobert 与 M. Auli，“wav2vec：面向语音识别的无监督预训练”，载于 Interspeech，2019。[5] A. Pasad、J.-C. Chou 与 K. Livescu，“自监督语音表征模型的逐层分析”，载于 ASRU，2021。[6] A. Pasad、B. Shi 与 K. Livescu，“自监督语音模型的比较性逐层分析”，载于 ICASSP，2023。[7] O. D. Liu、H. Tang 与 S. Goldwater，“自监督预测编码模型在正交子空间中编码说话人与语音信息”，载于 Interspeech，2023。[8] Y. Meng、S. Goldwater 与 H. Tang，“神经语音模型中的有效上下文”，载于 Interspeech，2025。[9] K. Choi 与 E. Yeo，“打开 wav2vec 特征编码器的黑箱”，arXiv 预印本 arXiv:2210.15386，2022。[10] K. Choi、A. Pasad、T. Nakamura 等，“自监督语音表征更偏语音而非语义”，载于 Interspeech，2024。[11] K. Choi、E. Yeo、C. J. Cho 等，“[b]=[d]-[t]+[p]：自监督语音模型发现音系向量运算”，arXiv 预印本 arXiv:2602.18899，2026。[12] K. Choi、E. Yeo、K. Chang 等，“利用自监督语音模型中的音位变体进行非典型发音评估”，载于 NAACL，2025。[13] A. Sicherman 与 Y. Adi，“面向口语语言建模分析离散自监督语音表征”，载于 ICASSP，2023。[14] D. Wells、H. Tang 与 K. Richmond，“英语语音自监督表征的语音分析”，载于 Interspeech，2022。[15] Y.-J. Shih 与 D. Harwath，“自监督语音模型的界面设计”，载于 Interspeech，2024。[16] X. Chang、B. Yan、K. Choi 等，“利用离散语音单元探索语音识别、翻译与理解：一项比较研究”，载于 ICASSP，2024。[17] M. de Heer Kloots 与 W. Zuidema，“神经语音模型中的类人语言偏差：Wav2Vec2.0 中的语音分类与音系制约”，载于 Interspeech，2024。[18] B. M. Abdullah、M. M. Shaik、B. M¨obius 与 D. Klakow，“自监督离散语音表征的信息论分析”，载于 Interspeech，2023。

[19] Y. Li、G. K. Anumanchipalli 与 A. o. Mohamed，“利用面向语音的深度神经网络剖析人类听觉通路中的神经计算”，《自然·神经科学》，第26卷，2023。[20] A. Pasad、C.-M. Chien、S. Settle 与 K. Livescu，“自监督语音模型对词语了解什么？”，《计算语言学会汇刊》，第12卷，2024。[21] A. Baade、P. Peng 与 D. Harwath，“SyllableLM：为语音语言模型学习粗粒度语义单元”，载于 ICLR，2025。[22] C. J. Cho、N. Lee、A. Gupta 等，“Sylber：由原始音频得到的语音音节嵌入表示”，载于 ICLR，2025。[23] C. J. Cho、N. Lee、A. W. Black 与 G. K. Anumanchipalli，“Sylber 2.0：一种通用音节嵌入”，arXiv 预印本 arXiv:2601.22306，2026。[24] N. Visser、S. Malan、D. Slabbert 与 H. Kamper，“ZeroSyl：面向口语语言建模的简单零资源音节切分”，arXiv 预印本 arXiv:2210.15386，2026。[25] K. Choi、M. Someki、E. Strubell 与 S. Watanabe，“设备端流式离散语音单元”，载于 Interspeech，2025。[26] B. McFee、C. Raffel、D. Liang 等，“librosa：用 Python 进行音频与音乐信号分析。”SciPy，2015。[27] R. Chaabouni、E. Dunbar、N. Zeghidour 与 E. Dupoux，“学习弱监督多模态音素嵌入”，载于 Interspeech，2017。[28] V. Zouhar、K. Chang、C. Cui 等，“PWESuite：语音学词嵌入及其所促进的任务”，载于语言资源与评估会议暨国际计算语言学会议，2024。[29] D. R. Mortensen、P. Littell、A. Bharadwaj 等，“Panphon：将国际音标音段映射为发音特征向量的资源”，载于国际计算语言学会议，2016。[30] J. S. Garofolo、L. F. Lamel、W. M. Fisher 等，“DARPA TIMIT：声学语音连续语音语料库光盘，美国国家标准与技术研究院语音光盘 1-1.1”，1993。[31] E. Chodroff、B. Paˇzon、A. Baker 与 S. Moran，“加州大学洛杉矶分校语音学实验室档案的语音分割”，载于语言资源与评估会议暨国际计算语言学会议，2024。[32] K. Ethayarajh，“上下文化的词表示有多依赖上下文? 比较 BERT、ELMo 与 GPT-2 嵌入的几何结构”，载于自然语言处理实证方法会议，2019。[33] A. Krizhevsky，“从微型图像中学习多层特征”，硕士学位论文，多伦多大学，2009。[34] R. Huo 与 E. Dunbar，“是迭代细化而非训练目标使 HuBERT 的行为不同于 wav2vec 2.0”，载于 Interspeech，2025。[35] K. Choi、J.-w. Jung 与 S. Watanabe，“通过互信息的变分界理解探针行为”，载于 ICASSP，2024。[36] K. Martin、J. Gauthier、C. Breiss 与 R. Levy，“探测自监督语音模型中的语音与音位信息：一项关于送气的案例研究”，载于 Interspeech，2023。[37] C. J. Cho、P. Wu、A. Mohamed 与 G. K. Anumanchipalli，“语音自监督学习中声道发音的证据”，载于 ICASSP，2023。[38] C. J. Cho、A. Mohamed、A. W. Black 与 G. K. Anumanchipalli，“语音自监督模型推断通用发音运动学”，载于 ICASSP，2024。[39] P. Peng 与 D. Harwath，“视觉锚定的自监督语音模型中的词语发现”，载于 Interspeech，2022。[40] G. Shen、A. Alishahi、A. Bisazza 与 G. Chrupała，“从声波到句法：探测口语语言模型中的句法”，载于 Interspeech，2023。[41] R. Schwartz、Y. Chow、O. Kimball 等，“依赖上下文的建模用于连续语音的声学语音识别

,” 在*ICASSP* , 1985。

[42] S. J. Young、J. J. Odell 和 P. C. Woodland，“用于高精度建模的基于树的状态绑定”，载于人类语言技术：在新泽西州普兰斯伯勒举行的研讨会论文集，1994 年。[43] C.-L. Feng、P.-c. Hsu 和 H.-y. Lee，“沉默胜于言语：利用沉默存储说话人信息的自监督模型”，*arXiv* 预印本 *arXiv:2205.03759*，2022 年。[44] H. Kamper、B. van Niekerk、J. Za¨ıdi 和 M.-A. Carbonneau，“Linearvc：从语音转换视角看自监督特征的线性变换”，载于 *Interspeech*，2025 年。[45] P. Mousavi、G. Maimon、A. Moumen 等，“离散音频词元：不仅仅是一篇综述！”*TMLR*，2025 年。[46] Z. Borsos、R. Marinier、D. Vincent等，“AudioLM：一种用于音频生成的语言建模方法”，*IEEE/ACM TASLP*，第 31 卷，2023 年。[47] A. D´efossez、L. Mazar´e、M. Orsini 等，“Moshi：用于实时对话的语音文本基础模型”，*arXiv* 预印本 *arXiv:2410.00037*，2024 年。[48] X. Zhang、D. Zhang、S. Li 等，“Speechtokenizer：用于语音语言模型的统一语音分词器”，载于*ICLR*，2024 年。[49] Z. Huang、S. McIntosh、D. Saito 和 N. Minematsu，“Kanade：用于口语语言建模的简单解耦分词器”，*arXiv* 预印本 *arXiv:2602.00594*，2026 年。[50] T. J. Hazen、W. Shen 和 C. White，“使用音素后验图模板的基于示例查询的口语术语检测”，载于*ASRU*，2009 年。[51] M. Morrison、C. Churchwell、N. Pruyne 和 B. Pardo，“细粒度且可解释的神经语音编辑”，载于*Interspeech*，2024 年。[52] M. Cernak、B. Potard 和 P. N. Garner，“使用人工神经网络进行音系声码器编码”，载于*ICASSP*，2015 年。[53] C. J. Cho、P. Wu、T. S. Prabhune 等，“通过声道运动学对语音进行编码”，*IEEE* 信号处理选题期刊，第 18 卷，2024 年。