文本与语音预训练模型之间的自适应知识蒸馏

Jinjie Ni[1, 2, *], *Yukun Ma*[1], *Wen Wang*[1], *Qian Chen*[1], *Dianwen Ng*[1,2], *Han Lei*[2], *Trung Hieu Nguyen*[1], *Chong Zhang*[1], *Bin Ma*[1], *Erik Cambria*[2]

[1]阿里巴巴集团[2]南洋理工大学计算机科学与工程学院

arXiv:2303.03600v1 [cs.CL] 7 Mar 2023

摘要

在海量语音语料库上进行训练，促成了近期众多自监督语音模型的成功。通过知识蒸馏，这些模型也可以受益于在丰富文本资源上预训练的语言模型所编码的知识。然而，由于文本与语音嵌入空间之间存在模态差异，蒸馏过程颇具挑战。本文研究基于度量的蒸馏方法，旨在仅使用少量数据、无需修改模型结构，便能对齐文本与语音的嵌入空间。鉴于现有文献忽略了文本与语音之间的语义与粒度差距，从而影响蒸馏效果，我们提出先验引导的**自**适应**知**识**蒸**馏（PAD），通过自适应利用不同粒度的文本/语音单元和先验分布，在文本与语音预训练模型之间实现更好的全局与局部对齐。我们在三个口语语言理解基准上进行评估，结果表明，与其他基于度量的蒸馏方法相比，PAD在迁移语言学知识方面更有效。

索引词— 知识蒸馏、预训练模型、口语语言理解

1. 引言

近年来，预训练模型在自然语言 [1, 2, 3, 4, 5] 和语音 [6, 7, 8]等多个领域取得了巨大成功。预训练模型通过掩码语言建模（MLM）、下一句预测等自监督学习任务，将输入编码为高层特征，并显著提升了下游任务 [1, 2, 9]的性能。口语语言理解（SLU）是语音领域的一系列下游任务，旨在训练模型理解语音输入的内容。目前最先进的 SLU 模型大多通过在语音预训练模型上进行微调 [10]来实现。

尽管预训练语音编码器在保留基于内容的信息方面表现强大，但由于缺乏符号表示，它在编码语言信息方面仍有不足。为弥补语音预训练的不足，语音编码器可以从在无标注文本上预训练的语言模型中迁移知识，从而获益。然而，语音与文本的预训练阶段所获得的知识之间存在天然差距，这是由模态偏差和预训练方法的差异造成的 [11]。这促使我们研究如何将文本预训练模型中的知识蒸馏到语音预训练模型中，以提升它们在理解任务上的表现。

我们将现有利用文本知识进行端到端口语语言理解的工作分为两类：单塔和双塔。单塔方法通过共享编码器对单模态输入进行编码，并且在推理阶段可以灵活地接收单模态或多模态输入。双塔方法使用模态专用编码器分别学习语音与文本，并依靠额外的层/码本或对齐损失来对齐多模态嵌入空间。

本研究属于双塔类别，通过设计度量目标 [17, 18], 来对齐文本与语音嵌入空间，我们将其称为基于度量的蒸馏。与单塔方法相比，基于度量的蒸馏只需少量数据即可蒸馏文本知识 [18] ，并具有明确的对齐目标。它不会引入任何额外参数，从而确保模型的效率和可分离性。然而，由于语音与文本之间存在较大的语义和粒度差距，设计适用于基于度量的蒸馏的度量方法颇具挑战。以往工作 [17, 18]采用简单的度量（L2 距离和余弦相似度）将文本知识蒸馏到语音中，却忽视了两者之间的语义和粒度差异，从而降低了蒸馏效果。为此，我们提出先验引导的**P**自适应**A**知识**D**蒸馏（PAD），在自适应粒度上对齐文本与语音表征，并利用反映语义重要性的先验知识指导对齐过程。

*在 Speech Lab、阿里巴巴集团新加坡分部实习期间完成的工作。

*获得新加坡科技研究局（A*STAR）AME 项目资助计划（项目 # A18A2b0046）的支持。

我们的**贡献**如下：

• 我们提出采用基于注意力的重要性先验（ASP），以促进从文本向语音迁移**语义**知识。

• 我们提出基于锚点的自适应跨度聚合算法（AASA），以缩小对齐的模态**粒度**差距。

• 据我们所知，我们是首个研究基于度量的语音文本蒸馏可行性的工作，并评估了原始全局和局部对齐之外的多种不同对齐策略。在三个口语语言理解基准上的结果验证了我们的假设和主张。

2. 方法论

2.1. 预备知识：全局与局部对齐

基于度量的蒸馏通过对齐语音和文本预训练模型的表示空间来蒸馏文本知识。通常有两种对齐策略：全局对齐和局部对齐。全局级对齐缩小序列级表示之间的差距，而局部级对齐则使局部单元表示相似。完成对齐训练后，语音表示空间将更接近文本表示空间，从而实现文本知识的迁移。给定一对语音和文本数据，语音和文本预训练模型的输出分别表示为$\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_n]$和$\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_n]$，其中$\mathbf{s}_m \in \mathbb{R}^{d_s}$和$\mathbf{t}_m \in \mathbb{R}^{d_t}$。 d_s和d_t分别是语音模块和文本模块的模型维度。

全局级对齐。 可以从预训练模型的输出表示中提取序列级表示。我们将语音和文本模型的全局级输出表示分别记为$\hat{}\mathbf{s} \in \mathbb{R}^{d_s}$和$\hat{}\mathbf{t} \in \mathbb{R}^{d_t}$，它们是经过下采样的嵌入，用于表示句子，例如 BERT 中的 cls 嵌入或平均嵌入。通过计算两个序列表示的 L1距离来进行对齐：

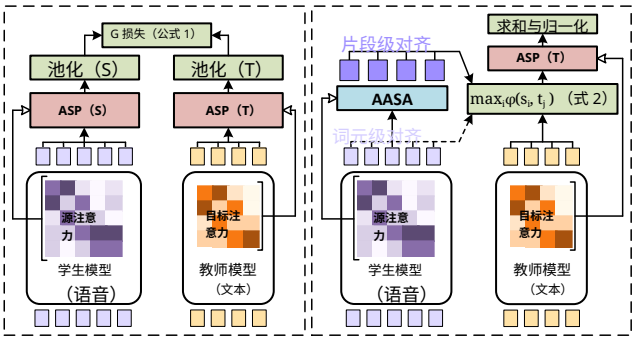
$$\mathcal{L}_{Glob} = ||\hat{\mathbf{s}} - \hat{\mathbf{t}}||_1 \tag{1}$$

在全局级对齐训练期间，文本模块被冻结，仅更新语音模块。

局部级对齐。 还可以在局部进行更精细的对齐。由于语音与文本词元之间不存在一对一映射，因此不能像公式 1 那样使用 L1距离。语音模块和文本模块的局部词元单元通过最大化针对所有文本单元的最大文本—语音相似度之和来进行对齐：

$$\mathcal{L}_{Loc} = -\frac{1}{n}\sum_{j=0}^{n} max_i \phi(\mathbf{s}_i, \mathbf{t}_j) \tag{2}$$

其中，ϕ 表示相似度度量，例如余弦相似度。与全局级对齐相同，文本模块
固定不变，仅更新语音模块。



（a）全局对齐　　（b）局部对齐

图 1：全局和局部 PAD。全局和局部对齐（第 2.1 节）均由**ASP** （第 2.2 节）提供信息，以缩小语义差距。**AASA** 自适应地重新组织语音序列，以缩小粒度差距。

2.2. 基于注意力的重要性先验

第 2.1 节中引入的全局与局部对齐将序列中的词元一视同仁。然而，语音与文本的模态偏差会导致某些对齐没有意义，例如，在文本中无法找到声学序列中的空白和噪声信号，尝试将这些信号与文本词元对齐可能无助于提升模型的理解能力。因此，我们认为，声学特征与其文本对应项之间的对齐应仅关注语义相关性最高的词元或帧，例如，非空白语音词元或文本序列中的关键词。由此，我们需要恰当地评估词元（或帧）的显著性。 [18] 使用逆文档频率（idf）对频繁出现的词元-帧对的显著性进行重新加权。然而，idf 存在两个问题：(a) 对于每个数据集而言它是固定的，而词元的显著性会随上下文不同而变化；(b) 它不适用于取连续值的语音特征。我们提出了基于注意力的显著性先验（ASP），从预训练 Transformer 模型的自注意力图中提取 $\mathcal{P}_{sig} \in \mathbb{R}^n$ 输出隐藏序列$\mathbf{H} = [\mathbf{h}_1, .., \mathbf{h}_n]$的显著性分布[3] （图 1）：

$$\mathcal{P}_{sig}(\mathbf{H}) = \frac{1}{L_0}\sum_{l=1}^{L_0} \frac{1}{\mathbf{e}A(\mathbf{H})^l\mathbf{e}^T} \sum_{m=1}^{n} A(\mathbf{H})^l_m \tag{3}$$

$\mathbf{h}_m$ 的重要性分数是对关注$\mathbf{h}_m$的单元的注意力权重取归一化均值得到的。 $A(\mathbf{H})^l = [A(\mathbf{H})^l_1, ..., A(\mathbf{H})^l_n]$ 表示 l层的自注意力图；$A(\mathbf{H})^l_m \in \{a \in \mathbb{R} | 0 < a < 1\}^n$ 表示关注$\mathbf{H}$的第 m个单元的各单元的注意力权重，即 $A(\mathbf{H})^l$的第 m 行； L_0 表示变换器模型的层数；$\mathbf{e} = \{1\}^n$。先验分布取决于上下文，我们将展示语音与

[3]我们发现，在某些情况下，最后一层注意力图能够更准确地估计 ASP，从而带来更好的性能。

算法1基于锚点的自适应跨度聚合

过程AASA ($\mathbf{S}$ ξ) ▷输入：$\mathbf{S} \in \mathbb{R}^{n \times d_s}$

初始化：锚点 $\mathbf{\Gamma} \leftarrow [id_{e0}]$；跨度池 $\mathbf{S} \leftarrow [\,]$。

$P'_{sig}(\mathbf{S})$，$Index_{sorted} \leftarrow$ 降序排序 $P_{sig}(\mathbf{S})$。

▷根据重要性分数（公式 3）选择锚点。

对于 id 属于 $Index_{sorted}$，**执行**

  如果 $\forall id_e \in \mathbf{\Gamma}$，$|id - id_e| > \xi/2$ **则** ▷控制密度。

    $span_m \leftarrow \{\mathbf{s}_{id-s_m} \cdots, \mathbf{s}^{id+s_m}\}$ 对于 $m \in 1, ..., c$

    ▷ $s_m \in \{\xi/2, \xi, 2\xi, ...\}$；$\xi$表示基准尺度。

    $\tilde{\mathbf{s}}_{id} \leftarrow \{(pooling(span_1) ..., pooling(span_c)\}$

    将 id 加到 $\mathbf{\Gamma}$；将 $\tilde{\mathbf{s}}_{id}$ 加到 $\tilde{\mathbf{S}}$.

返回 $\tilde{\mathbf{S}}$ ▷输出：$\tilde{S} \in \mathbb{R}^{k \times c \times d_s}$

文本模块如何缩小模态语义差距，并有助于获得更好的对齐结果。

对于全局对齐，在对语音和文本模块的输出序列应用显著性先验后，我们得到 $\mathbf{S}'$ 和 $\mathbf{T}'$：$\mathbf{S}' = \mathbf{S} \odot \mathcal{P}_{sig}(\mathbf{S})$、$\mathbf{T}' = \mathbf{T} \odot \mathcal{P}_{sig}(\mathbf{T})$。对于局部对齐，根据公式 2，我们得到关于文本输出 $\mathbf{T}$ 的相似度序列 $\mathbf{\Phi} = [\varphi_1, ..., \varphi_n]$，并得到受先验约束的 $\mathbf{\Phi}'$：$\mathbf{\Phi}' = \mathbf{\Phi} \odot \mathcal{P}_{sig}(\mathbf{\Phi})$。对于 PAD，我们在公式 1 和公式 2 中用其受先验约束的序列替换原始序列。

2.3. 基于锚点的自适应跨度聚合

实际上，一个文本标记可能对应多个语音标记，因为一个单词由多个音素组成。然而，第 2.1 节定义的局部对齐将每个文本标记与对应的语音标记进行对齐，这在粒度上并不一致。之所以能够进行这种对齐，是因为注意力表示具有上下文信息——每个语音嵌入都可能表示上下文化音素。然而，对于 Transformer 而言，注意力分数仍然高度局部化 [19, 20], ，这意味着嵌入往往只表示其相邻的近邻。因此，语音和文本标记的表示之间仍然存在粒度差距，直接对齐它们并不是最优选择。

我们提出自适应地生成局部跨度，其中每个跨度由若干语音嵌入聚合而成，以缩小粒度差距。对齐过程与公式 2 相同，唯一不同之处在于：原始语音序列被基于锚点的自适应跨度聚合（AASA）过程（算法 1）生成的局部跨度池所替代。我们将语音序列$\mathbf{S}$ 的局部跨度池表示为 $\tilde{\mathbf{S}} = [\tilde{\mathbf{S}}_1, ..., \tilde{\mathbf{S}}_k]$，其中 $\tilde{\mathbf{S}}_m \in \mathbb{R}^{c \times d_s}$。每个跨度池 $\tilde{\mathbf{S}}_m$ 由 c 个不同尺度的跨度组成，而跨度池的绝对位置由锚点决定。与局部级对齐（公式 2）一致，每个文本标记都与$\mathbf{S}$中的某个跨度 $\tilde{\mathbf{s}}_m$ 对齐，从而最大化所有文本标记对应的最大文本—跨度相似度之和。

表 1：IC、ER 和 SF 的准确率与 F1分数结果。我们的 PAD 优于所有对齐基线。

模型	对齐变体	意图分类 准确率 ↑	情绪识别 准确率 ↑	Slot 填充 F1 ↑
wav2vec 2.0 Base [7]	-	94.40	63.02	87.94
全局 - 分类 [18]	G	97.39	64.19	86.30
局部 - 逆文档频率 [18]	T	94.91	58.49	82.04
全局 - 平均	G	97.30	64.65	86.00
TLocal	T	95.90	59.18	82.27
TLocal - CW	T	96.81	60.27	84.88
TLocal - CG	T	96.76	61.97	84.89
TLocal - 或	T	90.30	59.32	77.36
SLocal - 或	S	97.57	63.87	85.12
PAD-Glob	G	97.68	**64.91**	87.67
PAD-TLocal	T	95.90	62.46	**88.40**
PAD-SLocal	S	**97.86**	62.74	87.19

3. 结果与分析

实验设置。 我们使用 Hugging Face[4] 提供的 *wav2vec2-base* 和*bert-base-uncased* 检查点作为语音与文本模块。我们为两个模型均选择 768 维，并将 BERT 词汇表大小缩减至 5000。按照 [18], 的设置，对齐训练使用从 Librispeech 960 [21]数据集的*train-clean-360*子集随机选取的 10 小时转录数据。对于所有下游任务的评估，我们遵循 SUPERB [10] 基准，并使用 S3PRL[5] 以实现公平比较。我们分别在 Fluent Speech Commands、IEMOCAP 和 Audio SNIPS 数据集上评估三个口语语言理解任务：意图检测（IC）、情感识别（ER）和槽位填充（SF）。

SUPERB 冻结上游预训练模型，并将所有其他设置控制为一致。 [18]对整个上游模型进行微调，并与基线方法进行比较，但没有说明是否将各项设置控制为一致。因此，他们的对齐效果不易比较，而我们基于 SUPERB 的结果更公平，也更便于后续研究复现。我们重新实现了他们的对齐方法以进行比较。

与各种对齐方法的结果。我们报告了表1中三个层级的 PAD 与基线方法的对比结果。G、T 和 S 分别表示全局级、词元级局部和跨度级局部对齐。对于全局级对齐（公式1），我们比较了两种句子嵌入变体：cls嵌入（**Glob-cls**） [18] 和平均池化（**Glob-avr**）。对于词元级局部对齐，我们比较了五种变体：公式1中的标准词元级对齐（**TLocal**）；使用逆文档频率忽略高频词的词元级对齐（**TLocal-idf**） [18]; 首先使用相同的10小时数据，通过 CTC 损失训练语音词元识别器，然后使用文本词汇表上的预测概率进行加权（**TLocal-CW**，通过将对齐约束为 *ASP*）或

[4]https://huggingface.co[5] https://github.com/s3prl/s3prl

表2：逐一移除特征，对意图分类、情感识别和槽位填充进行消融研究。

对齐变体	对齐变体	意图分类	情绪识别	Slot 填充
		准确率 ↑	准确率 ↑	F1 ↑
PAD-Glob	G	**97.68**	**64.91**	**87.67**
- 不含 s 先验	G	97.55	64.39	87.17
- 不含 t 先验	G	97.34	64.14	87.59
- 两者均不含	G	97.30	63.97	86.29
PAD-TLocal	T	**95.90**	**62.46**	**88.40**
- 不含先验	T	95.90	59.18	82.27
PAD-SLocal	S	**97.86**	**62.74**	**87.19**
- 不含先验	S	97.52	60.37	83.04
- 不含跨度池	S	97.86	62.72	87.06
- 不含锚点	S	96.73	60.88	86.10

引导对齐（**TLocal-CG**，通过最大概率实现语音-文本映射，并根据该映射而非公式*2*中的最大相似度进行对齐）；**TLocal-OR**将公式2中的相似度分布视为 CTC 计算中的概率分布，并将对齐损失计算为 CTC 损失，以获得基于最大期望的对齐。对于跨度级局部对齐，我们将其与跨度级有序对齐（**SLocal-OR**）进行比较；除粒度不同外，后者与**TLocal-OR**相同。

表1中的结果表明，我们的三种对齐方法在下游任务中的表现均优于所有基线方法。不过，全局对齐与局部对齐在不同任务上的表现存在权衡。采用全局对齐的模型在分类任务（意图分类和情感识别）中表现更好，而采用局部对齐的模型在序列生成任务（槽位填充）中表现更佳。分类任务需要更好的全局级语义，而序列生成任务则需要更精细的词元级语义。与词元级对齐相比，跨度级对齐在这两类任务上都取得了良好表现，这说明 AASP 能够缩小语音与文本之间的粒度差异，效果显著。

消融研究。我们通过消融 PAD 的各项特征，评估所做不同选择的影响（表2）。对于 PAD-Glob，我们逐一移除语音先验和文本先验，性能随之持续下降，验证了它们的重要性。同样，我们也验证了显著性先验对 PAD-TLocal 和 PAD-SLocal 的有效性。此外，对于 PAD-SLocal，我们发现，无论是 (a) 移除跨度池并使用固定跨度，还是 (b) 使用等步长跨度而非基于锚点的跨度，都会降低性能。

联合对齐分析。我们还评估了所有同时执行多种对齐的联合对齐组合（表3）。我们发现，它们的表现都不如单独对齐的组合结果。此外，我们观察到 PAD-G&S 和 PAD-T&S ＞ PAD-all joint ＞ PAD-G&T 的表现，这与我们的假设一致：全局级对齐与词元级对齐

表 3：联合对齐组合分析。

对齐变体	对齐变体	意图分类	情绪识别	Slot 填充
		准确率 ↑	准确率 ↑	F1 ↑
PAD-Glob	G	97.68	**64.91**	87.67
PAD-TLocal	T	95.90	62.46	**88.40**
PAD-SLocal	S	**97.86**	62.74	87.19
PAD-G&T	G+T	68.63	61.42	77.98
PAD-G&S	G+S	97.86	62.93	87.25
PAD-T&S	T+S	97.68	62.91	86.49
PAD-all joint	G+T+S	97.50	62.40	85.47

由于粒度差距，在本质上可能并不兼容；而得益于 AASA 的自适应聚合（公式1），这两者的粒度都接近跨度级对齐。

对齐损失分析。除了下游任务准确率之外，对齐前后对齐损失的变化是比较对齐效果的一个粗略指标，因为它反映了嵌入空间的同质化程度。在对所有条目进行适当缩放后，我们从训练轮次 1 开始绘制了全局对齐训练和局部对齐训练的验证损失曲线，如图2所示。图2(a)显示，与原始版本相比，受先验约束的三种对齐方式表现出更高的损失降低，这与上述结果一致，说明了语音先验和文本先验的有效性。在图2(b)中，受先验约束或利用自适应跨度的局部对齐表现出更高的损失降低，这同样与显著性先验和自适应跨度在局部级对齐实验中的有效性一致。

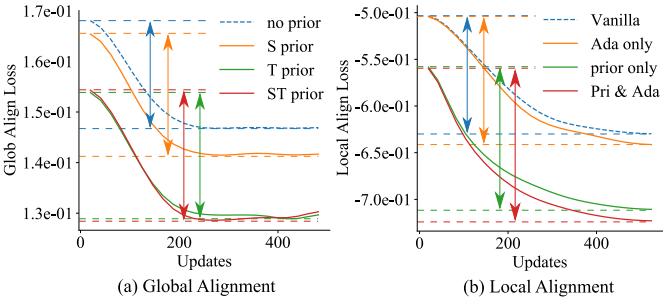


图 2：开发集上的全局和局部级对齐损失。S 和 T 分别表示语音和文本。

4. 结论

我们提出了一种利用自适应跨度和存储的显著性先验分布的方法，以在全局与局部级别上实现文本和语音预训练模型之间更好的对齐。我们还观察到，全局与局部对齐之间始终存在权衡，因为二者在本质上可能相互矛盾。未来，我们计划研究如何缓解这种权衡。

5. 参考文献

[1] JacobDevlin 等，“BERT：用于语言理解的深度双向 Transformer 的预训练”，载于 *NAACL-HLT 2019* 会议论文集，美国明尼苏达州明尼阿波利斯，*2019* 年 *6* 月 *2–7* 日，第 *1* 卷。2019 年，第 4171–4186 页，计算语言学协会。

[2] Zhenzhong Lan 等人，“ALBERT：一种用于语言表征自监督学习的轻量级 BERT”，载于第八届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。2020年，OpenReview.net。

[3] Yinhan Liu 等人，“RoBERTa：一种稳健优化的 BERT 预训练方法”，2019年。

[4] Colin Raffel 等人，“探索使用统一文本到文本变换器进行迁移学习的极限”，《机器学习研究杂志》，第21卷，第140:1–140:67页，2020年。

[5] Alec Radford，“语言模型是无监督的多任务学习器”，2019年。

[6] Steffen Schneider 等人，“wav2vec：用于语音识别的无监督预训练”，载于*Interspeech 2019*，奥地利格拉茨，*2019*年*9*月*15*日至*19*日，Gernot Kubin 和 Zdravko Kacic 编。2019年，第3465–3469页，ISCA。

[7] Alexei Baevski 等人，“wav2vec 2.0：一种用于语音表征自监督学习的框架”，发表于*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。

[8] Wei-Ning Hsu 等人，“HuBERT：通过对隐藏单元进行掩码预测的自监督语音表征学习”，*IEEE ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021 年。

[9]Mauajama Firdaus、Asif Ekbal 和 Erik Cambria，“对话系统中的多任务多语言意图检测与槽位填充”，信息融合，第 91 卷，第 299–315 页，2023 年。

[10]Shu-Wen Yang 等人，“SUPERB：语音处理通用性能基准”，发表于*Interspeech 2021*，捷克布尔诺，*2021*年*8*月*30*日至*9*月*3*日，Hynek Hermansky、Honza Cernock´y、Luk´as Burget、Lori Lamel、Odette Scharenborg 和 Petr Motl´ıcek 编，2021 年，第 1194–1198 页，ISCA。

[11] Junyi Ao 等人，“Speecht5：面向口语语言处理的统一模态编码器-解码器预训练”，载于*ACL 2022*，爱尔兰都柏林，*2022*年*5*月*22*日至*27*日，Smaranda Muresan、Preslav Nakov 和 Aline Villavicencio 编，2022 年，第 5723–5738 页，计算语言学协会。

[12] Sachidananda 等人，“Calm：对比式对齐的音频-语言多速率多模态表示”，*arXiv* 预印本 *arXiv:2202.03587*，2022 年。[13] Ankur 等人、Bapna，“Slam：通过语音-文本联合预训练实现语音与语言建模的统一编码器”，*arXiv* 预印本 *arXiv:2110.10329*，2021 年。[14] Ankur 等人、Bapna，“mslam：大规模多语言联合预训练语音与文本”，*arXiv* 预印本 *arXiv:2202.01374*，2022 年。

[15] Bhuvan 等人 Agrawal，“Tie your embeddings down：用于端到端口语理解的跨模态潜在空间”，收录于 *ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022 年，第 7157–7161 页。

[16] Alexander 等人 Liu，“跨模态离散表示学习”，*arXiv* 预印本 *arXiv:2106.05438*，2021 年。

[17] Pavel Denisov 和 Ngoc Thang Vu，“通过跨模态教师-学生学习，利用预训练语义语音嵌入实现端到端口语理解”，收录于*Interspeech 2020*，国际语音通信协会第 *21* 届年会，线上活动，中国上海，*2020* 年 *10* 月 *25—29* 日。2020 年，第 881–885 页，ISCA。

[18] Yu-An Chung 等人，“SPLAT：面向口语语言理解的语音-语言联合预训练”，载于*NAACL-HLT 2021*，线上举行，*2021*年*6*月*6*日至*11*日。2021年，第1897–1907页，计算语言学协会。

[19] Liunian Harold Li 等人，“配备视觉能力的 BERT 会关注什么？”，载于第*58*届计算语言学协会年会论文集，*ACL 2020*，线上举行，*2020*年*7*月*5*日至*10*日，Dan Jurafsky、Joyce Chai、Natalie Schluter 和 Joel R. Tetreault 编。2020年，第5265–5275页，计算语言学协会。

[20] Zhanghao Wu、Zhijian Liu、Ji Lin、Yujun Lin 和 Song Han，“采用长短程注意力的轻量级 Transformer”，载于第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。2020年，OpenReview.net。

[21] Vassil Panayotov 等人，“Librispeech：基于公有领域有声读物的自动语音识别语料库”，载于 *2015*年 *IEEE* 国际声学、语音与信号处理会议，*ICASSP 2015*，南布里斯班，昆士兰州，澳大利亚，*2015*年*4*月*19*日至*24*日。2015年，第 5206–5210 页，IEEE。