arXiv:2303.03144v1 [cs.MM] 6 Mar 2023

IPA-CLIP：将语音先验融入视觉与语言预训练

Chihaya Matsuhira
m
atsuhirac@cs.is.i.nagoya-u.ac.jp
名古屋大学 日本爱知县

Marc A. Kastner
m
kastner@i.kyoto-u.ac.jp
京都大学 日本京都

Takahiro Komamizu
taka-coma@acm.org 名
古屋大学 日本爱知县

Takatsugu Hirayama
t-hirayama@uhe.ac.jp 人类环
境大学 日本爱知县

Keisuke Doman
kdoman@sist.chukyo-u.ac.jp
中京大学 日本爱知县

Yasutomo Kawanishi
yasutomo.kawanishi@riken.jp
理化学研究所 日本京都

Ichiro Ide
ide
@i.nagoya-u.ac.jp 名古
屋大学 日本爱知县

摘要

近年来，大规模视觉与语言（V&L）预训练已成为许多多媒体系统的标准骨干。即便在未曾见过的情境中，这类模型也展现出卓越的性能，但其表现往往不符合人类的直觉。尤其是，它们通常不会考虑输入的发音，而人类会利用发音来理解语言，特别是遇到未知词时。因此，本文将语音先验引入对比式语言-图像预训练（CLIP）这一视觉与语言预训练模型，使其考虑不同发音输入之间的发音相似性。为此，我们首先提出一种音素嵌入，将国际音标（IPA）音标表所提供的音位关系作为语音先验加以利用。接着，通过蒸馏冻结的 CLIP 文本编码器，我们训练了一个采用基于IPA的嵌入的发音编码器。所提方法命名为 IPA-CLIP，由该发音编码器和原始 CLIP 编码器（图像和文本）组成。定量评估表明，使用所提音素嵌入时，嵌入空间中的音素分布能更准确地体现语音关系。此外，在一些多模态检索任务中，我们确认所提发音编码器能够提升文本编码器的性能，并且相较于文本编码器，发音编码器能以更符合语音特征的方式处理无意义词。最后，定性评估验证了发音编码器与人类对发音相似性的感知之间存在相关性。

特此授予许可：任何人均可免费为个人或课堂使用而制作本作品的全部或部分数字或纸质副本，但前提是不得出于营利或商业利益制作或分发副本，且副本首页必须附有此声明及完整引文。对于本作品中由非 ACM 所有的组件，必须尊重其版权。注明出处后可以进行摘要。除此之外，如需复制、重新出版、发布到服务器或重新分发到列表，必须事先获得具体许可和／或支付费用。请通过 permissions@acm.org 申请许可。会议缩写“*XX*”，*2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2023 Association for Computing Machinery。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX

CCS 概念

• **信息系统** → **语言模型**；相似性度量； • **计算方法** → 图像表示；**音系学／形态学**； • **应用计算** → 心理学。

关键词

发音、语音学、音系学、视觉与语言、蒸馏、多模态理解

ACM 参考文献格式：
Chihaya Matsuhira、Marc A. Kastner、Takahiro Komamizu、Takatsugu Hirayama、Keisuke Doman、Yasutomo Kawanishi 和 Ichiro Ide。2023年。IPA-CLIP：将语音先验融入视觉与语言预训练。在请从权利确认电子邮件中输入正确的会议标题（会议缩写“*XX*”）。ACM，纽约州纽约，美国，11页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

基于大规模图文数据集的视觉与语言（V&L）预训练，作为多媒体系统的基础模型，正受到越来越多的关注。对比式语言-图像预训练（CLIP） [18] 是此类视觉与语言预训练模型之一，由共享双模态嵌入空间的图像编码器和文本编码器组成。与以往模型相比，它使用了大一个数量级的训练数据，这保证了其在图像分类和检索 [18], 目标检测 [22], 图像生成 [4], 以及图像描述生成 [8]等各种应用中的有效性。此外，即使在训练集中未出现过的场景中，它也能表现良好。

然而，在许多情况下，这类模型的行为方式并不符合人类直觉。原因之一是，它们不考虑输入的语音特征，即词语如何发音；而人类会有意识或无意识地利用这些特征，凭直觉表达词语的含义。例如，当一位英语使用者在对话中使用“*Lump*”一词时，可能会联想到某种**沉重而圆润**的东西，这与其他发音相似的词语“*Bump*”“*Slump*”和“*Plump*”相似。

	双唇音		唇齿音		齿音		齿龈音		后置齿龈音		硬腭音		软腭音		声门音	
鼻音		m		ɱ				n				ɲ		ŋ		
塞音	p	b					t	d			c	ɟ	k	g	ʔ	
塞擦音							ts	dz	ʧ	ʤ	tɕ	dʑ				
擦音							s	z	ʃ	ʒ	ɕ	ʑ				
非塞擦音	ɸ	β	f	v	θ	ð					ç	ʝ	x	ɣ	h	ɦ
近音				ʋ				ɹ				j		ɰ		
边近音								l				ʎ		ʟ		

每列左侧的符号表示清辅音，右侧的符号表示浊辅音

(a) 辅音（肺部气流音）

	前		近前		中央		近back		Back	
近	i	ʏ			ɨ	ʉ			ɯ	u
次闭			ɪ	ʏ				ʊ		
半闭	e	ø			ɘ	ɵ			ɤ	o
Mid					ə					
半开	ɛ	œ			ɜ	ɞ			ʌ	ɔ
次开 æ					ɐ					
Open	a	ɶ			ä				ɑ	ɒ

每列左侧的符号表示不圆唇元音，
右侧的符号表示圆唇元音

(b) 元音

图1：本文使用的肺部气流辅音和元音国际音标表 [11] 。该表将自然语言中出现的几乎所有音素按照其语音关系联系起来。本文使用的英语音素以红色标出。

人类也会利用这类语音知识来处理口语 [10], ，尤其是在听到未知词或无意义词（简称非词）时。人类倾向于通过将这类非词与发音相似的词联系起来，想象其含义。例如，当英语使用者听到非词“*Britch*”的发音时，他们可能会将其与“*Bridge*”一词联系起来，因此“*Britch*”的含义可能会被理解为与桥梁有关的事物。与此同时，另一个非词“*Br-ish*”（与“*Fish*”押韵）由于与“*Bridge*”的语音相似性较低，可能较少被如此理解。若不了解语音关系，传统模型和系统便无法考虑这种对应关系。

因此，本文旨在将语音先验引入视觉与语言预训练模型，使其考虑不同发音输入之间的语音相似性。通过这种方式，我们还希望使模型能够处理包含非词的输入，这些非词的发音与某些现有词语略有不同。这将使模型对这类输入的处理更符合人类的预期。

将语音知识引入预训练模型的一种可行方法，是更改文本编码器的分词器，并使用图文对重新训练或微调整个模型。例如，在 CLIP 模型中，采用字节对编码 [21] 对文本输入进行分词并生成嵌入。然而，这种方法无法按上述方式处理非词。其他现有的嵌入方式，例如字符级嵌入 [6, 14], ，主要在应对拼写错误方面更为稳健，可能会更有效。但这类嵌入仍然不够理想，因为它们同样忽略了语音相似性。其障碍之一在于，这些方法的语言输入以字素书写，而字素不一定对应音位。此外，重新训练或微调原始预训练模型也可能存在缺点，因为这需要大量图文对和巨大的计算成本。

为解决这些问题，我们首先将语音先验融入通用的音位级嵌入。我们利用国际音标表 [11], （如图 1 所示），提出一种基于国际音标的音素嵌入。该音标表连接了几乎所有可能出现在任何自然语言中的音素，其中的邻近程度表示语音相似性。所提方法将国际音标表中的这些关系纳入嵌入空间的构建。虽然语音先验以及由此得到的所提音素嵌入应适用于任何语言，但本文其余部分所述的所提方法主要是针对英语设计、实现和评估的。

本文其余部分所述的所提方法主要是针对英语设计、实现和评估的。

接下来，我们采用蒸馏方法扩展视觉与语言预训练模型，使其能够接受以音素书写的语言输入。具体而言，我们按照图 2 所示实现 IPA-CLIP，这是一个扩展 CLIP 的模型。IPA-CLIP 由三个编码器组成：原有的 CLIP 图像编码器和文本编码器，以及一个新训练的发音编码器。发音编码器以由国际音标符号书写的音素数组作为输入（例如，/ə ˈfoʊˌtoʊ əv ə kæt./ 对应“ 一张猫的照片。”）。这使编码器中的基于国际音标的音素嵌入能够依据语音知识处理每个音素。为训练该编码器，我们冻结原始 CLIP 文本编码器，并使用若干文本-发音对对其进行蒸馏。这种蒸馏降低了获取发音编码器的成本，同时使其与原始 CLIP 的双模态联合嵌入空间共享特征空间。此外，共享嵌入空间还能拓宽 CLIP 的应用范围。由于发音编码器是通过蒸馏训练的，任何依赖 CLIP 文本编码器的系统，只需替换这两个编码器，也能接受缺乏正字法的语言的发音输入。

本文的贡献可总结如下：(1) 我们提出了一种基于国际音标的音素嵌入，将国际音标表中的语音相似性融入其音素嵌入空间；(2) 我们实现了 IPA-CLIP，这是一种扩展 CLIP [18], （视觉与语言预训练模型之一）的模型，利用基于国际音标的音素嵌入接受任意发音输入；(3) 我们确认，基于国际音标的音素嵌入的输出与国际音标表中的语音关系一致；(4) 我们将 IPA-CLIP 应用于若干检索任务，讨论其相较于原始 CLIP 的性能，以及它将未知词与发音相似的已知词联系起来的能力；(5) 我们评估 IPA-CLIP 的嵌入空间与人类对发音相似性的感知之间的相关性。

第 2 节首先介绍语音学所提供的语音关系，然后在第 3 节介绍相关研究。第 4 节详细说明所提出的基于国际音标的音素嵌入和 IPA-CLIP。我们在第 5 节和第 6 节报告实验并讨论结果，最后在第 7 节总结本文。

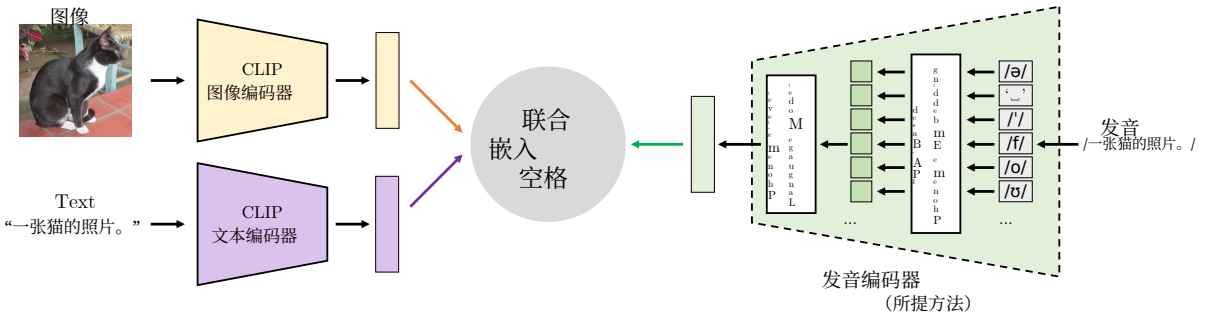


图2：所提方法 IPA-CLIP 模型概览

2 语音学中的音素属性

用于描述单词发音的字母称为语音符号。国际音标（International Phonetic Alphabet，IPA）转写是描述发音最常用的字母系统之一。它为每个音位分配一个独特的符号，同时也为音位以外的语音成分（如重音和音节边界）提供符号。国际音标表 [11], 如图 1 所示，列出了所有国际音标符号，并标示了所有可能出现在任何自然语言中的音位之间的关系。在表中，每个音位都由多个语音属性来描述。辅音有三个属性：浊音、发音部位和发音方式。元音也有三个属性：高度、舌位前后和圆唇度。例如，根据该表，清软腭塞音 /k/（如“*Coat*”中的音）具有“清音”（浊音）、“软腭音”（发音部位）和“塞音”（发音方式）这几项辅音属性；半闭后圆唇元音 /o/（如“*Coat*”中的音）则具有“半闭”（高度）、“后”（舌位前后）和“圆唇”（圆唇度）这几项元音属性。有些辅音（如浊唇软腭近音 /w/）具有多个发音部位。在这种情况下， /w/ 由四项辅音属性（“浊音”“唇音”“软腭音”和“近音”）来描述。

本文结合这些语音属性，构建所提基于国际音标的音素嵌入的嵌入空间。

3 相关工作

第3.1节介绍了若干先前研究，这些研究利用此类语音信息，通过计算方法构建语音空间。接下来，第3.2节介绍了几种尝试将 CLIP 扩展到其他模态的方法。

3.1 语音学的计算方法

已有若干研究将语音知识（如国际音标表）融入以字母或音标书写的词语之间语音相似性的计算中 [1, 10, 17, 25]。Vitz和Winkler [25] 提出了预测音位距离，这是一种根据类似编辑距离的方法计算两个词语发音之间差异度的指标。虽然该方法旨在预测发音之间的差异，却没有考虑语音关系。为解决这一问题，Hahn和Bailey [10] 将音位的语音特征纳入编辑距离度量。他们的方法将以下两种音位之间的距离视为较近：具有某些共同属性的两个音位（例如，清**软腭塞音** /k/ 和浊**软腭塞音** /g/）；而共同属性较少的其他音位对则被视为距离较远（例如，清软腭塞音 /k/ 和浊双唇鼻音 /m/）。他们将这种距离度量与人类感知进行比较，证明了其有效性，尤其适用于单音节词。Parrish [17] 提出了一种用于诗歌应用的二元语法模型。该模型根据与国际音标表所列特征相似的语音特征构建。作者提取了英语单词中可能出现的949种独特语音特征二元组合，并据此建立直方图。词语之间的语音相似性通过其直方图的余弦相似度来衡量。Bay 等人 [1] 利用国际音标表的结构计算语音相似性，并将其称为文本转换的押韵分数。他们将第2节所述的三种辅音属性都视为类别属性，并将元音属性中的高度 和舌位前后 视为连续属性。具体而言，为计算辅音之间的相似性，他们会检查两个辅音共有的属性，并将这些共有属性对应的克罗内克δ按经验权重加权后求和。对于元音，他们在二维笛卡尔平面上手动重建元音图（忽略圆唇度），并测量元音之间的欧几里得距离。

近期的自然语言处理（NLP）技术也能获得神经音素嵌入，无需显式先验和监督即可反映语音相似性 [2, 13]。Kolachina 和 Mag- yar [13] 评估了 Word2vec [15, 16] 能否学习音素之间的语音关系。他们使用人工语言语料库以及英语语料库训练模型，并得出结论：Word2vec 嵌入能够很好地捕捉此类语料库中的语音关系。Boldsen等人 [2] 采用长短期记忆网络（LSTM）和 Transformer，对多个语言中的字符嵌入进行了类似分析；他们通过训练这些模型完成下一个字符预测任务。所得字符嵌入与从英语音素分类中得出的声音表征呈现出很强的相关性，这为先前研究的结果提供了支持 [13]。

为了构建本文中的基于国际音标的音素嵌入，我们沿用 Bay 等人提出的国际音标表用法。 [1] 我们将两种元音属性视为连续型属性，而将其他属性视为类别型属性。我们还将基于国际音标的嵌入与一种通过训练获得、未使用此类先验的神经音素嵌入进行比较，并将后者作为基线。

我们的目标与他们的主要区别在于，我们的语音空间不仅考虑语音相似度，还考虑语义相似度。因此，例如，两个含义截然不同的单词“比特”和“皮特”在我们的嵌入空间中不一定会成为彼此的近邻，而在上述方法中则会。

3.2 CLIP 对其他类型数据的扩展

由于 CLIP 在各种多模态任务中表现出色，许多方法被提出用于将其扩展到其他模态。对于音频模态，一些研究 [7, 9, 26] 使用多模态数据集，在原有图像和文本编码器之外训练一个新的音频编码器。Guzhov 等人 [9] 利用多个单模态和多模态数据集，同时为每种模态训练一个编码器。Wu 等人 [26] 冻结并蒸馏 CLIP图像编码器，仅训练一个额外的音频编码器，并使用音视数据集。这样，音频编码器不仅能与图像编码器兼容，也能与文本编码器兼容。Elizalde 等人 [7] 使用文本-音频数据集，从头开始训练音频和文本编码器。所有这些方法都采用对比学习进行训练，这种方法源自原始 CLIP [18]。

在图像和文本模态中，Carlsson 等人 [3] 扩展了 CLIP 文本编码器。该编码器主要使用英语词汇进行训练，扩展后能够处理多种语言。他们首先准备一组英语句子，然后将其机器翻译成多种语言，以获得一个多语言数据集。与 Wu 等人 [26], 类似，他们利用该数据集，通过冻结并蒸馏 CLIP 文本编码器，使用多语言句对训练另一个多语言文本编码器。这确保机器翻译句子的输出与原始英语句子的输出相似。在训练过程中，他们没有使用对比损失，而是优化均方误差（MSE）损失，以蒸馏 CLIP 文本编码器。

本文将 CLIP 扩展至发音输入，提出 IPA-CLIP。为此，我们采用与 Carlsson 等人 [3] 提出的方法类似的蒸馏方法，以降低训练可兼容 CLIP 图像编码器的发音编码器的成本。我们自动将英语句子转换为音标转写，然后使用文本-发音对替代多语言文本对。

4 IPA-CLIP：CLIP 的语音嵌入蒸馏

本文提出 IPA-CLIP，这是 CLIP 面向发音输入的扩展。图 2 展示了 IPA-CLIP 的整体结构。它由三个编码器组成：CLIP图像编码器、CLIP文本编码器，以及一个新的发音编码器；三者共享同一个多模态嵌入空间。

在第4.1节中，我们首先介绍如何将国际音标表上的语音先验嵌入音位嵌入空间。第4.2节随后解释 IPA-CLIP 的蒸馏过程，该过程使用基于IPA的嵌入训练新的发音编码器。第4.3节则介绍所提方法的实现细节。

4.1 基于国际音标的音素嵌入

我们提出了一种音素嵌入，它考虑了国际音标表中各音素之间的语音关系。该音素嵌入层通过替换任何语言模型（包括 BERT）的词嵌入层来实现。由于该层基于国际音标表，因此输入其中的发音理论上具有普适性，不专属于任何语言。

如第 2 节所述，国际音标表为每个音位标注三种辅音属性：浊音、发音部位和发音方式；以及三种元音属性：高度、舌位前后和圆唇度。受运用这些属性的相关文献启发 [1], , 我们将两种元音属性——高度和舌位前后——视为连续属性。我们考虑不同音位在这些属性上的差异程度。例如，**闭**前不圆唇元音 /i/ 应视为比**半闭**前不圆唇元音 /e/ 与**开**前不圆唇元音 /a/更相似。相反，按照 Bay 等人 [1], 的做法，我们将其他四种属性视为类别属性。我们忽略国际音标表中音位的排列顺序，只关注两个音位是否具有任何共同属性。浊双唇鼻音 /m/与清软腭塞音 /k/ 之间的差异，被视为等同于 /m/ 清齿龈塞音 /t/ 之间的差异，尽管 /t/ 在表中的位置比 /m/更近。

如图 3a 所示，所提方法将音素嵌入$\mathbf{p}$计算为线性组合 $\sum_i x_i \mathbf{w}_i$，其中 x_i是幅度，$\mathbf{w}_i$是第 i 个属性的特征向量。具体而言，对于每个音素，我们计算 N 维稀疏幅值向量$\mathbf{x}$ 的转置与 $N \times D$特征矩阵W的乘积，记作

$$\mathbf{p} = \mathbf{x}^\top W = \sum_{i=1}^{N} x_i \mathbf{w}_i = x_1 \mathbf{w}_1 + x_2 \mathbf{w}_2 + \cdots + x_N \mathbf{w}_N. \qquad (1)$$

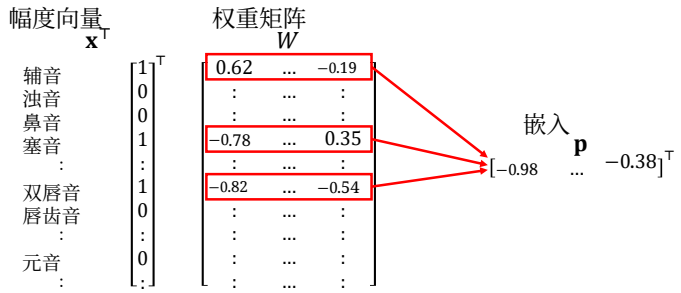
表 1 展示了部分音素对应的 N属性以及向量$\mathbf{x}$ 中的幅度示例。如表所示，$\mathbf{x}$ 还包含音素以外字母的属性，例如重音、空格、逗号和感叹号。尽管这些字母不是音素，所提方法仍会将它们投影到相同的音位嵌入空间中。这样做旨在确保发音编码器的输入具有与 CLIP 文本编码器同等的灵活性。因此，发音编码器能够区分同音文本，例如“everyday”和“every day”，以及“a cat”和“a cat!”。

这种基于国际音标的音素嵌入用于第4.2节所述发音编码器的音素嵌入层，以处理发音输入。

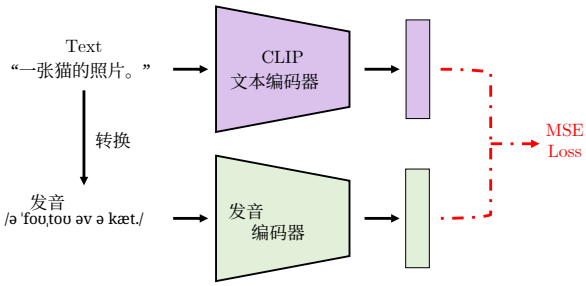
通过 DistillingCLIPTextEncoder 进行训练

IPA-CLIP 的发音编码器训练基于 Carlsson 等人 [3] 和 Wu 等人 [26]提出的蒸馏方法。尽管本文的实现仅针对英语，但如果具备相应资源，该蒸馏方法本身也可以应用于其他语言。

蒸馏过程如图3b所示。首先，为了准备句子-发音对数据集，我们使用给定的词语到发音词典，将每个句子转换为其发音。具体而言，通过查阅词典，我们将所有



(a) 所提的基于国际音标的音素嵌入。该图展示了清双唇塞音 /p/的音素嵌入计算过程。



(b) 蒸馏过程。

图3：IPA-CLIP发音编码器构建过程的详细示意图。

表1：幅度向量x各维度所表示属性的示例。

属性	类别	范围	示例x				
			/p/	/v/	/e/	/ʊ/	‘,’
辅音	辅音	$x_i \in \{0,1\}$	1	1	0	0	0
浊音		$x_i \in \{0,1\}$	0	1	0	0	0
发音方式 1：鼻音		$x_i \in \{0,1\}$	0	0	0	0	0
发音方式 2：塞音		$x_i \in \{0,1\}$	1	0	0	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮
发音部位 1：双唇音		$x_i \in \{0,1\}$	1	0	0	0	0
发音部位 2：唇齿音		$x_i \in \{0,1\}$	0	1	0	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮
元音	元音	$x_i \in \{0,1\}$	0	0	1	1	0
高度		$0 \le x_i \le 1$	0	0	$\frac{2}{6}$	$\frac{1}{6}$	0
舌位前后		$0 \le x_i \le 1$	0	0	0	$\frac{3}{4}$	0
圆唇度		$x_i \in \{0,1\}$	0	0	0	1	0
主重音 /ˈ/	其他	$x_i \in \{0,1\}$	0	0	0	0	0
次重音 /ˌ/		$x_i \in \{0,1\}$	0	0	0	0	0
字符‘ ’：空格		$x_i \in \{0,1\}$	0	0	0	0	0
字符“,”：逗号		$x_i \in \{0,1\}$	0	0	0	0	1
字符“!”：感叹号		$x_i \in \{0,1\}$	0	0	0	0	0
⋮		⋮	⋮	⋮	⋮	⋮	⋮

句子中的词语替换为其发音。例如，在此转换中，句子“a photo of a cat.”会被转换为其发音 /ə ˈfoʊˌtoʊ əv ə kæt./。如第4.1节所述，我们忽略拉丁字母中的情况，且不会排除拉丁字母以外的字母。

借助该数据集，我们将发音编码器与 CLIP文本编码器一同训练。在训练过程中，文本编码器的权重保持冻结。给定一个句子-发音对，我们训练发音编码器，使其输出与文本编码器计算得到的句子嵌入完全相同的发音嵌入。训练目标采用均方误差损失，而非训练 CLIP 时使用的对比损失 [18]。我们选择均方误差损失，是因为已知它在蒸馏任务中的效果优于对比损失 [3]。此外，使用对比损失（即余弦相似度）进行训练，会得到角度相似但幅度不同的嵌入。由于现有的若干系统可能需要 CLIP嵌入的精确幅度，这样的训练目标可能会限制其应用范围。

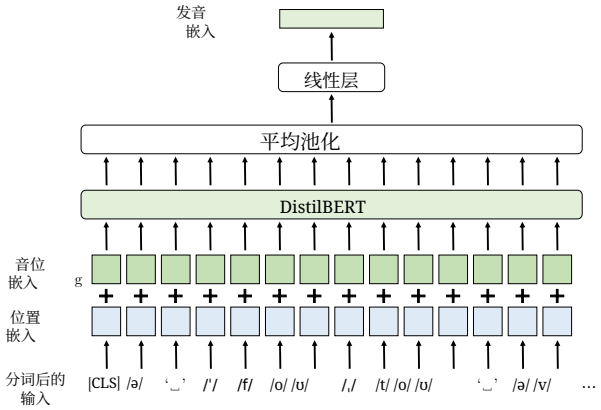


图4：IPA-CLIP中所使用的发音编码器示意图。IPA-CLIP在其音素嵌入层中采用了所提方法基于国际音标的音素嵌入。

4.3 实现

作为发音编码器的架构，我们采用 DistilBERT [20],，这是 BERT 的轻量高效版本。图 4 展示了其实现方式。我们将其词嵌入替换为所提出的基于国际音标的音素嵌入，并添加一个额外的线性层，使其输出维度与 CLIP 编码器的输出维度相匹配，但不修改 DistilBERT 的任何其他部分。发音编码器从头开始训练。

为了提炼主要使用英语数据训练的原始 CLIP 模型，我们使用了 Carls- son 等人 [3] 编制的英语句子列表。该列表混合了取自多个图像描述数据集的句子，这些句子可能与视觉领域有着紧密联系。此外，为了扩大词汇量，我们使用面向拼写检查器的词表（SCOWL）1——一个包含 102,305 个单词的英语词表——准备了仅由一个单词组成的句子。我们使用 Pyt hon 软件包 eng-to-ipa[2]，将这些句子转换为以国际音标书写的发音。该软件包使用卡内基梅隆大学（CMU）发音词典，许多先前的工作 [1, 13, 17] 也使用了该词典。我们移除

[1] http://wordlist.aspell.net/（访问日期：2023年1月19日）[2] https://pypi.org/project/eng-to-ipa/（访问日期：2023年1月19日）

包含软件包未提供其发音的单词的句子。最终得到的训练数据总计包含 1,168,451 个句子。遵循先前工作的实现方式 [3]，我们将验证集划分的大小固定为 1,000 个句子，从而得到包含 1,167,451 个句子的训练集和包含 1,000 个句子的验证集。

我们在两个不同的预训练 CLIP 模型上进行了测试：OpenAI 提供的模型卡片 [3]中列出的 ViT-B/32 和 ViT-L/14。ViT-B/32 是最简单、最轻量的模型，而 ViT-L/14 则是模型卡片中最新发布、性能最强的模型。这些模型都使用 Transformer 作为图像和文本编码器。我们以 5×10^{-5}的学习率、32 的批量大小和 Adam优化器 [12], 训练发音编码器，最多训练 50 个周期。

5 定量评估

本节以定量方式评估所提的基于 IPA 的音素嵌入和 IPA-CLIP。借助语音学先验知识，IPA-CLIP 通过蒸馏学习音素嵌入和发音嵌入。

在这些实验中，我们将 IPA-CLIP 的性能与一种基线方法进行比较。基线方法采用发音编码器，使用普通的字符级（音素级）嵌入层，而非基于 IPA 的嵌入层。因此，其得到的嵌入空间没有考虑国际音标表中的语音关系，而是仅从训练数据中的音素共现中隐式地学习这些关系。此外，我们还探究 IPA-CLIP 的特征矩阵W权重应设为可训练，还是冻结为随机初始化值。如果冻结，W所负责的映射就会变成随机映射；如果设为可训练，则该映射会反映从音素共现中学到的语音关系，这既可能有益，也可能有害。在以下各节中，我们分别将这两种情况称为*所提方法（可训练）*和 *所提方法（冻结）*。在这两种情况下，DistilBERT 和额外的线性层的权重始终都会参与训练。

请注意，为了公平比较所提方法 IPA-CLIP 与基线，以下各节所述实验仅衡量在英语中出现的音位的性能。

5.1 音位嵌入空间实验

本节实验比较所提方法中的基于IPA的嵌入（所提方法）与传统的音位级嵌入（基线方法）。我们使用不同的指标衡量所学习的音位嵌入空间的以下特征：（1）辅音簇和元音簇的分布有多大区别；（2）辅音布局如何表示辅音之间的语音关系；（3）元音布局如何表示元音之间的语音关系。

辅音和元音之间的明确区分，使 IPA-CLIP 更容易区分这两类音位。此外，音位布局与国际音标表相一致，这意味着 IPA-CLIP 可以根据语音相似性计算音位之间的距离。我们对辅音和元音使用不同的度量指标，因为正如第4.1节所述，辅音属性是类别型的，而某些元音属性是连续型的。

5.1.1 **辅音和元音分布的区分度。**为了衡量音位嵌入空间中辅音和元音的区分度，我们计算嵌入空间中辅音簇与元音簇之间的轮廓系数 [19]。给定 C_c（分别为 C_v）是辅音（元音）的集合，c是C_c中的一个元素，且 x_c是 c的嵌入向量，辅音 c 的系数s_c 计算如下：

$$s_c = \frac{b_c - a_c}{\max(a_c, b_c)}, \tag{2}$$

其中

$$a_c = \frac{1}{|C_c| - 1} \sum_{\hat{c} \in C_c, c \neq \hat{c}} d(x_c, x_{\hat{c}}),\ b_c = \frac{1}{|C_v|} \sum_{v \in C_v} d(x_c, x_v), \tag{3}$$

且 d 表示欧几里得距离。随后，通过对所有辅音之间的系数取平均值，计算辅音簇 $^{s}C_c$ 的轮廓系数。元音簇 $^{s}C_v$ 的系数也以相同方式获得，只需在公式中交换 c 和 v。

$^{s}C_c$（分别为 $^{s}C_v$）的取值范围为 $[-1, 1]$，其中较高的值表示辅音（元音）簇与元音（辅音）簇之间具有明显的区分度。

5.1.2 **辅音分布与语音学的一致性。**为了衡量辅音在语音空间中的分布与国际音标表中的辅音分类之间的一致性，我们计算每个辅音属性的平均精度均值（mAP）。它是平均精度分数（AP）的平均值，使用 Precision@k 和 Recall@k 分数形式化表示为

$$\mathrm{AP} = \sum_{k=1}^{|C_c|} \mathrm{Precision@}k\{\mathrm{Recall@}(k) - \mathrm{Recall@}(k-1)\}. \tag{4}$$

随后，我们通过测量语音空间中的欧几里得距离，对检索到的辅音进行排序。在整个计算过程中，我们将具有所关注属性的辅音视为相关项。例如，当从浊音属性的角度评估浊双唇塞音 /b/ 时，其相关辅音集合 R 就变为包含所有浊辅音的集合，例如 /d/、/m/ /g/ /b/ 和。如果该辅音的检索排序依次为[/p/,/d/,/t/,/m/,/g/,.....]，则 /b/ 的 AP 分数 $\mathrm{AP_b}$ 为

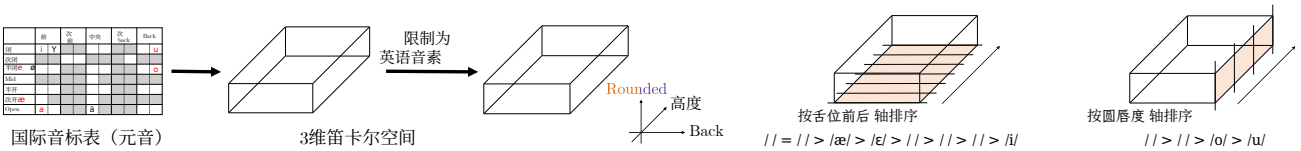
$$\begin{aligned} \mathrm{AP_b} &= \frac{0}{1}\left(\frac{0}{|R|} - \frac{0}{|R|}\right) + \frac{1}{2}\left(\frac{1}{|R|} - \frac{0}{|R|}\right) + \frac{1}{3}\left(\frac{1}{|R|} - \frac{1}{|R|}\right) \\ &\quad + \frac{2}{4}\left(\frac{2}{|R|} - \frac{1}{|R|}\right) + \frac{3}{5}\left(\frac{3}{|R|} - \frac{2}{|R|}\right) + \cdots \\ &= \frac{1}{|R|}\left(\frac{1}{2} + \frac{2}{4} + \frac{3}{5} + \cdots\right). \end{aligned} \tag{5}$$

最后，每个辅音属性的平均精度均值是通过对所有辅音的 AP 分数取平均计算得出的。因此，较高的平均精度均值意味着，就目标属性而言，语音相似的音位在音位空间中彼此距离较近。

[3]https://github.com/openai/CLIP/blob/main/model-card.md（访问日期：2023年1月19日）

表 2：音位嵌入空间的定量评估。轮廓系数（Silhouette）、平均精度均值（mAP）和斯皮尔曼秩相关（Rank Corr.）分别表示辅音与元音之间的区分度、辅音分布与语音学的一致性，以及元音分布与语音学的相关性。

		轮廓系数 ↑		平均精度均值 ↑（辅音）			秩相关系数↑（元音）			损失 ↓
Base	方法	s_{C_c}	s_{C_v}	浊音	发音部位	发音方式	高度	Back	圆唇	MSE
ViT-B32	基线	−0.006	0.049	0.585	0.433	0.394	0.524	0.574	0.796	**0.0084**
	所提方法（可训练）	0.036	0.193	**0.753**	0.763	0.717	**0.913**	0.666	0.882	0.0096
	所提方法（冻结）	**0.252**	**0.558**	0.735	**0.812**	**0.845**	0.889	**0.688**	**0.925**	0.0092
ViT-L14	基线	−0.014	0.054	0.589	0.421	0.342	0.541	0.592	0.753	**0.027**
	所提方法（可训练）	−0.036	0.217	0.705	0.767	0.642	**0.891**	0.680	**0.925**	0.028
	所提方法（冻结）	**0.233**	**0.568**	**0.735**	**0.810**	**0.837**	0.890	**0.685**	**0.925**	0.028
	上界	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	–



(a) 将国际音标表转换到三维笛卡尔空间。

(b) 计算 /ɑ/ 在高度上的真实排名轴。

图 5：（a）将国际音标表转换为三维笛卡尔空间，以及（b）创建真实值元音排名以测量秩相关分数的核心思路。

5.1.3 元音分布与语音学的相关性。本节介绍如何衡量语音空间中的元音分布与国际音标表中的元音顺序之间的相关性。

核心思路如图 5 所示。首先，受 Bay 等人 [1],的启发，我们将每个元音映射到能够复现国际音标表的三维笛卡尔空间中。笛卡尔空间的三个坐标轴分别表示元音的高度、舌位前后和圆唇度属性。接着，对于三个坐标轴／属性中的每一个，我们计算元音在该笛卡尔空间中的分布与其在语音空间中的分布之间的斯皮尔曼秩相关。对于每项属性，我们通过对共享另外两项属性之一的元音进行排序，创建两组真实值排名。例如，如图 5 所示，在评估高度属性上的元音 /ɑ/ 时，我们计算以下两组排名：（1）后元音中的排名：/ɑ/ > /ɔ/ > /o/ > /u/；（2）不圆唇元音中的排名：/ɑ/ > /a/ > /æ/ > /ɛ/ > /ə/ > /e/ > /ɪ/ > /i/。

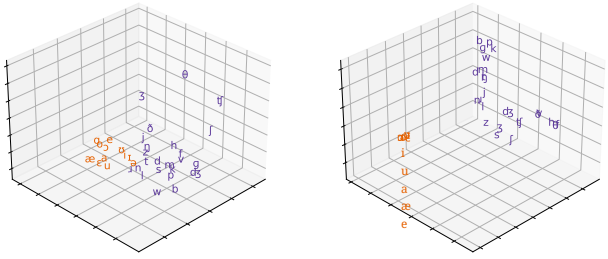
准备好这些真实排名后，我们计算目标元音的两种排名分别与在语音空间中检索到的相似元音排名之间的两个秩相关得分。检索依据欧几里得距离进行。最后，对于每种元音属性，我们计算所有元音的平均值，作为语音空间上的秩相关指标。因此，较高的秩相关指标意味着元音在音位空间中的分布能够很好地体现元音之间的语音关系。

5.1.4 结果与讨论。 该实验的结果如表 2 所示。表中还展示了第 50 个 epoch 时的验证损失。总体而言，所提方法（冻结）在几乎所有指标上表现最佳。所提方法（可训练）的表现也不相上下，轮廓系数除外。除了验证损失略好之外，基线方法的表现都不如所提方法。

与其他方法相比，所提方法（冻结）的一大优势在于轮廓系数。较高的轮廓系数意味着辅音簇和元音簇的分布彼此区分明显，因此在音位嵌入空间中的重叠较少。由于基线和所提方法（可训练）在训练权重矩阵的过程中轮廓系数有所下降，这一比较表明，从句子中的音位共现中学习到的嵌入无法清晰地区分辅音和元音。这与此类神经嵌入已知能够很好地表示语音关系这一事实相矛盾 [2, 13]。不过，由于基线在平均精度均值和秩相关指标上的表现也尚可，我们还可以确认，即使没有任何明确的先验，神经嵌入也能学习辅音内部及元音内部的关系。

此外，与基线相比，两种所提方法的平均精度均值和秩相关均有所提升。这表明，所提基于国际音标的嵌入在区分辅音内部和元音内部方面效果良好。相反，验证损失下降速度略有减缓，这表明引入的先验对模型施加了额外限制，从而影响了蒸馏过程中的优化。

与上界的比较表明，辅音簇 s_c 的轮廓系数并不高，即使采用所提方法



(a) 基线 **(b) 所提方法（冻结）**

图6：音位嵌入空间的可视化。辅音和元音分别以蓝色和黄色显示，以比较它们的分布。

（冻结）方法也是如此。造成这一现象的一个可能原因是，无论两个簇的区分程度如何，辅音的分布范围都比元音更广。为验证这一点，我们将基线方法和所提方法（冻结）的音位空间可视化。图6展示了所有辅音和元音在经过主成分分析（PCA）降维后的三维空间中的散点图。正如预期，在这两种方法中，辅音簇的分布范围都比元音簇更广，这可能会对轮廓系数指标产生很大影响。尽管如此，该图仍表明，就两个簇的分离程度及其与国际音标表的一致性而言，所提方法的空间排列比基线方法更合理。

5.2 发音空间上的实验

我们进一步开展实验，探讨 IPA-CLIP 在多模态任务中的适用性。我们测量其在以下三项检索任务中的表现：(1) 根据现有词语的发音进行基于检索的图像分类；(2) 根据非词的发音检索图像；(3) 根据非词的发音检索文本。请注意，这里的非词指英语中不存在、但听起来与某些现有词语相似的词。

通过测量模型在这三项任务中的表现，我们可以分别评估发音编码器与以下编码器的一致性：(1) 处理现有词语时的图像编码器；(2) 处理非词时的图像编码器；(3) 处理非词时的文本编码器。

在所有这些实验中，我们使用 ImageNet [5]验证数据集作为图文对的来源。该数据集为 1,000 个预定义类别中的每个类别提供 50 张图像。我们按照第 4.3 节所述的方法，将每个类别标签转换为其发音。移除无法将标签转换为发音的类别后，我们最终得到 912 个类别，每个类别各有 50 张图像。请注意，我们使用的类别标签与 CLIP [18], 作者采用的标签一致，这些标签不同于 ImageNet 提供的类别标签。作者修改了标签，以减少原始 ImageNet 标签在其基于检索的图像分类任务中的歧义。

此外，以下实验仅使用基于 CLIP ViT-L14 训练的模型，因为第5.1节中的实验表明，基础模型的选择不会带来显著差异。另外，我们不会将所提方法与 CLIP [18] 以外的模型进行性能比较。这是因为第3.1节中介绍的所有方法都不适用于多模态检索任务，而第3 2 [3, 节中26]介绍的其他基于CLIP的模型在这些任务上的表现均不如 CLIP。

5.2.1 基于发音的图像分类。我们在一项基于检索的图像分类任务上进行实验，该任务与先前一项研究中的任务类似 [18]。在此，IPA-CLIP 通过衡量图像嵌入与类别标签嵌入之间的余弦相似度来对图像进行分类，例如采用

/ə ˈfoʊˌtoʊ əv

< < >

CLASS>/ (“一张 CLASS 的照片”) 这种形式。举例来说，给定一张图像以及两个类别标签“狗”和“猫”，IPA-CLIP 首先计算图像的嵌入，以及这两个发音的嵌入 /ə ˈfoʊˌtoʊ əv dɔg/ 和 /ə ˈfoʊˌtoʊ əv kæt/。随后，它衡量图像嵌入与每个发音嵌入之间的余弦相似度。通过找出相似度最高的图像—标签对，它确定图像的类别标签。

在本实验中，我们还使用现有的 Python 软件包 [23]，根据类别标签的词频（以 Zipf 量表衡量，我们称之为 *Zipf*频率）筛选类别。这样做是为了观察罕见的类别标签如何影响分类结果；这些标签在 IPA-CLIP 的蒸馏过程中从未出现或很少出现。

我们将准确率作为衡量指标。我们将所提方法的性能与 CLIP [18], 进行比较，后者使用文本编码器根据文本标签对图像进行分类。为了了解 IPA-CLIP 发音编码器与 CLIP文本编码器在功能上的差异，我们还尝试通过在联合嵌入空间中取二者嵌入的平均值来融合它们。我们将其称为“所提方法（冻结）***+ CLIP***”。

5.2.2 非词到图像检索。 为了评估 IPA-CLIP 对具有某些相似发音现有词的无意义词的鲁棒性，我们通过对 ImageNet 的类别标签进行轻微修改，构建了一组非词。

首先，我们仅关注 Zipf频率达到 3 或更高的标签（297 个类别）。然后，对于以单个辅音开头的标签（其中有 216 个类别符合条件），我们用其他可能的辅音替换初始辅音（例如，从 /dɛsk/: “Desk” 变为 /zɛsk/、/nɛsk/ 等），从而构建一组与原词发音相似的非词。接下来，我们移除其中恰好存在于英语词汇中的生成词。为进行检查，我们使用 SCOWL 词表和 CMU 词典[4]。这一过程产生了 3,530 个非词，它们分别源自这 216 个类别中的某一个。

在此准备过程中，我们还通过自动将每个音位转换为其拼写形式（/zɛsk/ 对应“Zesk”），来准备文本等价形式，以便评估基于文本的原始 CLIP。表 3 列出了替换中使用的所有辅音及其对应的拼写形式。如表所示，候选辅音选自所有出现在英语单词词首的辅音，但 /ð/ 除外，因为它拼写后会与 /θ/ 相同。

[4]https://github.com/menelik3/cmudict-ipa/（访问于2023年1月19日）

表3：用于根据 Ima- geNet 的类别标签生成非词的候选辅音及对应拼写形式。

辅音	/s/	/n/	/f/	/l/	/z/	/b/	/ɹ/	/p/	/g/	/k/
拼写		‘s’	‘n’	‘f’	‘l’	‘z’	‘b’	‘r’	‘p’	‘g’	‘k’ or ‘c’
辅音	/d/	/m/	/θ/	/t/	/ʧ/	/j/	/h/	/v/	/ʃ/	/ʧ/	/w/
拼写		‘d’	‘m’	‘th’	‘t’	‘j’	‘y’	‘h’	‘v’	‘sh’	‘ch’	‘w’

表4：在1000类ImageNet [5] 数据集上的图像分类准确率。我们仅使用可转换为其发音的类别，并使用Zipf频率筛除频率较低且罕见的标签名称。

Zipf频率	≥ 0.0	≥ 1.5	≥ 3.0	≥ 4.5
类别数量	912	492	297	29
基线	0.600	0.696	0.777	0.877
所提方法（可训练）	0.581	0.686	0.769	0.886
所提方法（冻结）	0.590	0.683	0.764	0.885
CLIP [18]	**0.712**	0.751	0.765	0.891
所提方法（冻结） + CLIP [18]	0.705	**0.765**	**0.799**	**0.897**

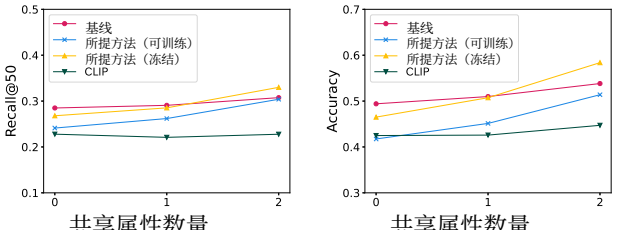
使用这些非词，我们执行非词到图像检索任务。给定一个非词，目标是检索出该非词所源自类别中的图像。例如，给定非词 /zɛsk/，我们衡量从 /ə ˈfoʊˌtoʊ əv zɛsk/ 的发音嵌入中能检索出“Desk”类别 50 张图像中的多少张。

我们使用 Recall@50 分数作为评估指标。我们根据非词与其原始词在语音上的相似程度对评估结果进行细分。具体而言，我们观察非词的首个辅音与其原始词首个辅音之间共享属性的数量。这样做旨在评估这些方法是否捕捉到了辅音之间的语音相似性。如果无论共享属性的数量多少，得分始终相近，则说明该方法没有考虑语音相似性。相反，如果得分与共享属性的数量相关，则可以认为该方法会基于语音相似性，将非词与发音相似的词关联起来。

5.2.3 非词到文本检索。非词到文本检索的流程与第 5.2.2 节所述流程十分相似，但本实验的检索目标是文本而非图像。我们使用第 5.2.2 节中为 216 个类别准备的 3,530 个非词。在本实验中，模型会检索非词所属类别的文本。例如，给定非词

/zɛsk/，我们评估这些方法能否从 216 个类别的文本嵌入中检索出句子“A photo of desk”的文本嵌入。我们以准确率作为衡量指标。

5.2.4 结果与讨论。首先，图像分类的结果如表4所示。表中显示，所提方法的性能受到类别标签出现频率（或稀有程度）的显著影响。如表格左侧所示，当类别中包含稀有词时，这些方法的表现远逊于 CLIP [18]。



（a）非词到图像 **（b）非词到文本**

图 7：（a）从以语音学符号（基线和所提方法）或拉丁字母（CLIP）书写的非词中进行图像检索，以及（b）文本检索的结果。与语音相似性的高度一致性使共享属性的数量（x 轴）与两项指标（y 轴）均呈现较强相关性。

这是因为这些模型作为学生模型，在蒸馏过程中很少接触这些稀有词。相比之下，当评估中不再包含稀有词时，它们的表现便与 CLIP 不相上下。

最有意思的是“所提方法（冻结） *+CLIP*”。尽管仅采用了简单的双模态融合策略，它在几乎所有设置下都取得了最佳表现。这表明，将发音模态引入现有视觉与语言预训练模型是有效的。对 297 类混淆矩阵的分析揭示了各编码器的优势与不足，但为节省篇幅，我们不展示具体矩阵。发音编码器对发音差异更敏感，而文本编码器更能应对含义上的差异。例如，所提方法（冻结） 将“刨底刨”误分类为“搭扣”，因为二者听起来相似。相比之下，CLIP 将“蛞蝓”误分类为“蜗牛”，因为二者的含义十分相近。由于“所提方法（冻结） *+ CLIP*”正确分类了这两种情况，对两个编码器的嵌入取平均值或可弥补各自的不足。

接下来，图 7 分别展示了非词到图像和非词到文本检索任务的结果。两个独立任务中的结果趋势相似：共享属性数量为 0 或 1 时，基线方法的检索效果最佳；而共享属性数量为 2 时，所提方法（冻结）表现最佳。这表明，所提方法（冻结）可能仅在非词与原词语音相似时，才会将两者关联起来。因此，这也证实了在发音嵌入空间中，所提方法（冻结）比其他方法更准确地考虑了辅音之间的语音相似性。

我们还观察到，在这些以非词为中心的任务中，所提方法始终优于 CLIP。当共享属性数量为 2 时，所提方法（冻结）使 Recall@50 提高了 0.112 分，准确率提高了 0.137 分。这一事实验证了所提出的发音模态使 CLIP 能够更好地应对非词。

为进一步分析 CLIP 与 IPA-CLIP 之间的差异，我们将词语和非词在二者共享的嵌入空间中的分布可视化。图8所示的散点图展示了现有词语以及听起来与“*Bridge*”相似的非词的嵌入，这些嵌入由 CLIP 或 IPA-CLIP 计算得出。图中显示，

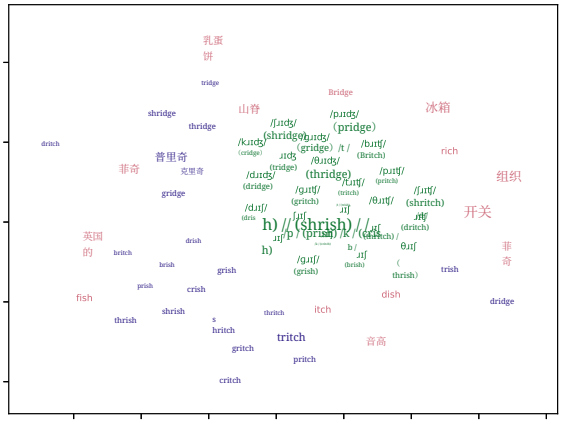


图8：非词的文本嵌入和发音嵌入的 t-SNE [24] 可视化。现有单词（红色）和非词（紫色）的文本嵌入由CLIP计算，而非词（绿色）的发音嵌入由IPA-CLIP计算。

IPA-CLIP 将“*Pridge*”（/pɹɪʤ/）和“*Britch*”（/bɹɪʧ/）等非词置于与发音相似的现有词“*Bridge*”（/bɹɪʤ/）非常接近的位置。相比之下，CLIP 并未将任何非词（甚至“*Pridge*”）置于“*Bridge*”附近。这支持了 IPA-CLIP 会考虑词语语音相似性的结论。

6 定性评估

我们还评估了所提方法在多大程度上使CLIP嵌入空间与人类对发音相似性的实际感知相一致。该评估采用了Vitz和Winkler [25]收集的发音相似度排名。在他们的四项试验中，美式英语使用者对给定目标词与其25个比较词中每个词之间纯粹的声音相似性进行了评分。他们的四项实验仅在目标词——分别为“坐”、“植物”、“奇迹”和“关系”——及其比较词方面有所不同。在前三项实验中，比较词由有效英语单词和无意义英语单词混合组成，它们具有与目标词相似的音节结构。在最后一项实验中，取消了对音节结构的这一限制。例如，目标词“坐”的比较词包括“*Pit*”、“*Sat*”和“*Tass*”，而目标词“关系”的比较词包括“*Belation*”、“*Fascinating*”和“*Get*”。

在我们的评估中，给定一个目标词，我们首先计算该词与每个对比词在各方法发音空间中的余弦相似度，并据此生成相似度排名。随后，我们将其秩相关与真实值进行比较，以此作为评估指标。数值越高，表示嵌入空间越符合人类对发音相似性的感知。

结果如表 5 所示。所有基于发音的方法都优于传统的基于文本的 CLIP，这验证了语音先验会促使发音相似的词彼此更加接近。在基于发音的方法中，当目标词是“坐”。这是因为其比较词的音节较短，因此比其他目标词产生了更多可能的相似发音词。因此，这些词的语音知识可能干扰了此类短音节词的嵌入计算，这可能是所提方法的一个缺点。尽管如此，由于此评估仅涵盖这四种特定情况，结果无法说明哪种基于发音的方法总体上效果最佳。

表 5：发音嵌入空间的定性评估。分数表示各方法嵌入空间中测得的词语相似度与人类评定的真实值之间的秩相关。

目标词	Sit	植物	奇迹	关系
基线	0.535	0.397	**0.693**	0.442
所提方法（可训练）	**0.642**	**0.549**	0.526	0.485
所提方法（冻结）	0.385	0.420	0.640	**0.504**
CLIP [18]	0.353	0.402	0.585	0.304

7 结论

我们提出了一种基于国际音标的音素嵌入，将国际音标（IPA）图表中的语音关系融入字符/音素级嵌入。我们还提出了 IPA-CLIP，它通过采用基于国际音标的音素嵌入，扩展视觉与语言（V&L）预训练模型 CLIP，使其能够接受任意发音输入。这使模型即使在输入中包含无意义词（非词）时也能进行处理。所提出的 IPA-CLIP 模型通过蒸馏 CLIP ViT-B/32 和 ViT-L/14 模型的文本编码器进行训练。评估结果表明，与不含语音先验的神经嵌入相比，采用基于国际音标的音素嵌入是有效的。我们还评估了 IPA-CLIP 在一些多模态检索任务上的表现，结果表明 IPA-CLIP 不仅性能相当，还有望超越原始的基于文本的 CLIP。在输入非词的设置下，IPA-CLIP 的表现始终优于 CLIP，验证了 IPA-CLIP 对非词的鲁棒性。进一步评估证实了 IPA-CLIP 的发音嵌入空间与人类对发音相似性的感知之间存在相关性。这些结果证明，IPA-CLIP 有助于增强多媒体系统的语言理解能力。

未来的工作需要进一步分析所提方法在哪些条件下优于基于文本的方法。此外，我们计划创建一个基于所提方法的系统，用于模拟人类对非词的感知。此外，我们还希望将 IPA-CLIP 扩展到多种语言，以量化人类感知中的跨语言差异。对此，我们认识到的一个局限是，所提基于IPA的嵌入忽略了发音的听觉特征，例如音素、声调和韵律之间的听觉相似性。为了解决这一局限，一种可能性是创建一个扩展至音频输入的 CLIP，类似于以往研究中已经实现的方案 [9, 26], ，并将其与所提 IPA-CLIP 结合。

致谢

本研究部分得到了微软研究院 CORE16 计划和 JSPS 科研经费补助（22H03612）的支持。本研究是名古屋大学与国立信息学研究所之间联合研究项目的成果。

第一作者谨借此机会，感谢由名古屋大学和日本科学技术振兴机构资助的“名古屋大学跨学科前沿奖学金”，该奖学金隶属于“旨在创建科学技术创新的大学奖学金设立项目”，项目编号为JPMJFS2120。

参考文献

[1] Benjamin Bay、Paul Bodily 和 Dan Ventura。2017。通过约束和词嵌入进行文本转换。载于第*8*届国际计算创造力会议论文集。佐治亚理工学院，美国佐治亚州亚特兰大，49–56。[2] Sidsel Boldsen、Manex Agirrezabal 和 Nora Hollenstein。2022。利用感知表征解读字符嵌入：形状、声音和颜色的案例。载于第*60*届计算语言学协会年会论文集（第*1*卷：长论文）。ACL，爱尔兰都柏林，6819–6836。https://doi.org/10.18653/v1/2022.acl-l ong.[3] Fredrik Carlsson、Philipp Eisen、Faton Rekathati 和 Magnus Sahlgren。2022。跨语言和多语言CLIP。载于第*13*届语言资源与评估会议论文集。ELRA，法国罗讷河口省马赛，6848–6854。[4] Katherine Crowson、Stella Biderman、Daniel Kornis、Dashiell Stander、Eric Halla- han、Louis Castricato 和 Edward Raff。2022。VQGAN-CLIP：利用自然语言引导进行开放领域图像生成与编辑。计算研究文库，arXiv预印本，arXiv: 2204 08583 10 48550 2204 08583 。[5] 2009 https://doi.org/ . /arXiv. . Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。. ImageNet：大规模分层图像数据库。载于*2009*年*IEEE*计算机学会计算机视觉与模式识别会议论文集。IEEE，美国佛罗里达州迈阿密，248–255。https://doi.org/10.1109/CVPR.2009.5206848[6] Hicham El Boukkouri、Olivier Ferret、Thomas Lavergne、Hiroshi Noji、Pierre Zweigenbaum 和 Jun’ichi Tsujii。2020。CharacterBERT：为来自字符的词级开放词汇表征调和ELMo与BERT。载于第*28*届国际计算语言学会议论文集。ACL，西班牙加泰罗尼亚巴塞罗那，6903–6915。https://doi.org/ 10 18653 1 2020 609 . /v / .coling-main. Benjamin Elizalde、[7] Soham Deshmukh、Mahmoud Al Ismail 和 Huaming Wang。2022。CLAP：从自然语言监督中学习音频概念。计算研究文库，arXiv预印本，arXiv:2206.04769。https://doi.org/10.48550/arxiv. 2206.04769[8] Federico Galatolo、Mario Cimino 和 Gigliola Vaglini。2021。通过CLIP引导的生成式潜在空间搜索从图像说明文字生成图像，反之亦然。载于国际图像处理与视觉工程会议论文集。SciTePress，捷克布拉格，166–174。https://doi.org/ 10 5220 0010503701660174 . / Andrey Guzhov、Federico [9] Raue、Jörn Hees 和 Andreas Dengel。2022。Audio- CLIP：将CLIP扩展到图像、文本和音频。载于*2022*年*IEEE*国际声学、语音与信号处理会议论文集。IEEE，新加坡，976–980。https://doi.org/10.1109/ ICASSP43922.2022.9747631[10] Ulrike Hahn 和 Todd M. Bailey。2005。是什么使词语听起来相似？认知 97, 3 2005 227 267 10 1016 2004 09 006 ( ), –。[11] 1999 https://doi.org/ . /j.cognition. . . 国际语音学会。. 国际语音学会手册：国际音标使用指南。剑桥大学出版社，英国英格兰剑桥。[12] Diederik P. Kingma 和 Jimmy Ba。2015。Ada m：一种随机优化方法。载于第*3*届国际学习表征会议论文集。ICLR，美国加利福尼亚州圣迭戈，15页。[13] Sudheer Kolachina 和 Lilla Magyar。2019。音素嵌入能学到哪些关于音系学的知识？载于第*16*届语音学、音系学和形态学计算研究研讨会论文集。ACL，意大利佛罗伦萨， 160 169 10 18653 1 19 4219 –。[14] https://doi.org/ . /v /W - Wentao Ma、Yiming Cui、Chenglei Si、Ting Liu、Shijin Wang 和 Guoping Hu。2020。CharBERT：字符感知预训练语言模型。载于第*28*届国际计算语言学会议论文集。ACL，西班牙加泰罗尼亚巴塞罗那，39–50。https: //doi.org/ 10 18653 1 2020 4 . /v / .coling-main. Tomas Mikolov、Kai Chen、[15] Greg Corrado 和 Jeffrey Dean。2013。向量空间中词表征的高效估计。计算研究文库，arXiv预印本，arXiv: 1301 3781 10 48550 1301 3781 。[16] https://doi.org/ . /arXiv. . Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。2013。词和短语的分布式表征及其组合性。载于神经信息处理系统进展，第26卷。Curran Associates，美国内华达州太浩湖，3111–3119。[17] Allison Parrish。2017。使用语音特征的诗歌声音相似度向量。载于第*13*届*AAAI*人工智能与互动数字娱乐会议论文集，第13卷，第2期。AAI Press，美国犹他州斯诺伯德，99–106。https://doi.org/10.1609/aiide.v13i2.12971

[18] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021。从自然语言监督中学习可迁移视觉模型。载于第*38*届国际机器学习会议论文集，机器学习研究论文集，第139卷。ML Research Press，在线，8748–8763。[19] Peter J. Rousseeuw。1987。轮廓系数：用于解释和验证聚类分析的图形辅助工具。计算与应用数学杂志20（1987），53–65。https://doi.org/ 10 1016 0377 0427 87 90125 7 . / - ( ) - Victor Sanh、Lysandre Debut、[20] Julien Chaumond 和 Thomas Wolf。2019。Distil- BERT，BERT的蒸馏版本：更小、更快、更便宜、更轻量。计算研究文库，arXiv预印本，arXiv:1910.01108。https://doi.org/10.48550/arXiv. 1910.01108[21] Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016。使用子词单元进行稀有词的神经机器翻译。载于第*54*届计算语言学协会年会论文集，第1卷。ACL，德国柏林，1715–1725。https://doi.org/10. 18653/v1/P16-1162[22] Hengcan Shi、Munawar Hayat、Yicheng Wu 和 Jianfei Cai。2022。ProposalCLIP：通过利用CLIP线索进行无监督开放类别目标提议生成。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议论文集。IEEE，美国路易斯安那州新奥尔良，9611–9620。[23] Robyn Speer、Joshua Chin、Andrew Lin、Sara Jewett 和 Lance Nathan。2018。LuminosoInsight/wordfreq：v 2 2 10 5281 1443582 。https://doi.org/ . [24] 2008 /zenodo. Laurens van der Maaten 和 Geoffrey Hinton。. 使用 t-SNE 可视化数据。机器学习研究杂志9，86（2008），2579–2605。[25] Paul C. Vitz 和 Brenda Spiegel Winkler。1973。预测英语单词被判断的“声音相似性”。言语学习与言语行为杂志12，4（1973），373–388。https://doi.org/ 10 1016 0022 5371 73 80016 7 . /S - ( ) - [26] Ho-Hsiang Wu、Prem Seetharaman、Kundan Kumar 和 Juan Pablo Bello。2022。Wav2CLIP：从CLIP学习鲁棒的音频表示。载于*2022*年*IEEE*国际声学、语音与信号处理会议论文集。IEEE，新加坡，4563–4567。https: // doi.org/10.1109/ICASSP43922.2022.9747669