分析自监督语音模型中预训练语言、语音、声调和讲话者信息之间的关系

Michele Gubian[1], *Ioana Krehan*[1], *Oli Liu*[2], *James Kirby*[1], *Sharon Goldwater*[2]

[1]Institute for Phonetics and Speech Processing, Ludwig Maximilian University of Munich, Germany

[2] School of Informatics, University of Edinburgh, UK

m.gubian@phonetik.uni-muenchen.de, Ioana.Krehan@campus.lmu.de, Oli.Liu@ed.ac.uk, j.kirby@phonetik.uni-muenchen.de, sgwater@inf.ed.ac.uk

Abstract

对自监督语音模型的分析已经开始揭示出这些模型在哪里以及如何表示不同类型的信息。然而，几乎所有的分析都集中在英语上。在此，我们研究了在四种不同语言上训练的 wav2vec2 模型如何编码与语言匹配和不匹配的语音。我们使用探测分类器和几何分析来研究音素、词汇音调和说话人信息是如何表示的。我们表明，对于所有的预训练和测试语言，编码音素、音调和说话人的子空间在很大程度上是正交的，并且逐层探测准确率的模式是相似的，在后期层次中，对匹配语言的音素和音调（但不是说话人）探测有相对较小的优势。我们的研究结果表明，wav2vec2 学习到的表示结构在很大程度上与预训练期间使用的语音材料无关。

Index Terms: self-supervised speech models, interpretability, probing, representational geometry, tone languages, wav2vec

1. 介绍

自监督语音模型的成功激发了人们对理解这些模型在预训练期间学习到的表示性质的兴趣（概述请参见 [1] 的第五部分）。先前的研究者已经解决了诸如在不同层次中可以获得哪些类型的信息 [2–9] 以及这些信息是如何结构化的 [10, 11] 等问题。然而，该领域的大多数工作都是研究来自英语预训练模型的表示，并使用英语数据进行分析。因此，目前尚不清楚这些发现能在多大程度上推广到其他语言预训练的模型，或者当预训练和分析语言不匹配时——模拟跨语言微调场景开始时的情况。[1]

在本研究中，我们分析了预训练于四种不同语言（英语、法语、普通话和越南语）上的 wav2vec 2.0（以下简称为 w2v2）模型 [14]，并使用同样的四种语言加上泰语进行测试。我们分析每个模型对匹配语言数据（例如，我们使用普通话测试数据分析预训练于普通话的模型）和跨语言数据（例如，我们使用不同语言的测试数据分析普通话模型）的表现。我们使用了两种方法：

分层探测分析。受之前分层探测研究 [2–9] 的启发，我们训练了三种线性分类器（用于音素、声调和说话者），并研究了每个模型不同层的探测准确性。我们对声调的分层分析在很大程度上重现了 [3, 5] 的结果，但对说话者和音素的分析扩展了之前的工作。据我们所知，之前对自监督模型中说话者信息的分层分析只探讨了基于英语预训练模型 [7, 11]，而跨语言的音素探测结果仅限于单一语言对 [15] 或仅在英语上进行测试 [16]。我们的探测结果系统地涵盖了更广泛的条件（预训练语言、测试语言和探测类型），允许更普遍的结论。特别是，我们发现：

- 对于测试的所有模型和语言，音素和声调探测结果遵循与早期英语音素研究中观察到的相同的一般模式：对这些语言类别的探测准确率在模型的早期层中增加，在中间层达到顶峰，然后在后期层中再次下降。
- 跨语言和匹配语言的音素/声调探测之间的性能差异仅在中间层显现，这表明这是语言特定表示出现的地方。
- 然而，在大多数情况下，这种差异相对较小，这有些令人惊讶，因为这些语言具有相当不同的语音和音系特征。然而，这确实解释了为何跨语言微调可以有效。

几何分析。这种分析基于 [10, 11] 的研究，他表明说话者和语音信息在各种自监督模型中（均在英语上预训练）被捕获在几乎正交的子空间中。他们假设 [10] 这种正交性源于这两种信息类型的统计独立性。如果是这样，那么语音的其他独立变化特性，例如词汇声调，也应该被正交编码。[2] 为了检验这一假设，我们使用 [11] 的正交性度量来分析声调子空间的几何结构，以及它与音素和话者子空间的关系，跨越我们所有的预训练和分析语言——再次提供了比以往工作更为广泛的分析。

这一分析表明：

- 先前关于说话者和音素子空间 [10, 11] 正交性的发现不仅适用于训练在其他语言上的模型，还适用于跨语言测试场景。
- 正如假设的那样，其他的子空间对（说话者和声调，音素和声调）也基本上是正交的，尽管程度略低于说话者和音素的子空间。
- 在大多数情况下，根据预训练语言或测试语言来看，正交性的程度没有明显的区别或模式。一个例外是音素和声调子空间，其正交性程度在三种声调测试语言中有所不同，并与每种语言中音素和声调之间的统计（不）独立性程度相关联。

总而言之，我们发现，语言特异性的语言信息编码在 w2v2 模型中是轻微的，但在中间层中始终出现，而说话者信息在不同模型中没有一致编码。我们还进一步支持了这样一种假设：这些模型隐式地将统计上独立的信息来源解缠为正交子空间。

2. 相关工作

尽管有许多研究分析了自监督语音表示（例如 [2–9, 26–28]），但只有少数研究审查了非英语模型或数据。其中一些采用了与我们不同的方法，例如，与人类感知数据对比模型 [29] 或使用可视化 [30]。在使用探测分类器的研究中，我们仅知道有一项研究考虑了说话者信息，该研究在几个语言上进行了测试，但只使用了一个（非 w2v2）英语预训练模型 [7]。几个探测研究审查了非英语模型，但都在有

[1]现在，微调通常从多语言预训练模型开始。虽然我们在此并未探讨多语言模型，但我们的发现为将来与这些模型进行对比提供了很好的基础，特别是因为像 XLSR [12] 和 MMS [13] 这样的流行模型是基于我们在此研究的 wav2vec 2.0 架构。

[2]词汇声调主要通过音调的变化来表达，但必须大致独立于个体 *说话者* 的标准音域，否则无法表达与说话人无关的信息。将声调作为一种单独“层次”的信息的典型语言学分析 [17] 也意味着声调和 *部分* 相对独立，正如我们在 5.1 节中进一步讨论的那样。

Table 1: 我们研究中使用的数据和模型。所有语料库都是朗读语音，除了越南语模型的预训练数据，它还包含 *YouTube* 数据。评估发音者、音位和声调表示为分析选择的类别数量。

Language	**Tonal?**	**Pretraining**			**Evaluation**				
		Model	Data (hrs)	Spkrs	Data source	Aligner	Spkrs	Phones	Tones
English	No	[14][a]	960	2120	LibriSpeech dev-clean [18]	MAUS	40	45	-
French	No	[19][a]	7,600	2338+	Vibravox [20][b]	MFA	40	39	-
Mandarin	Yes	[21][a]	1,000	1991	THCHS-30 [22]	MFA	40	51	4
Vietnamese	Yes	[23][a]	13,000	?	VIVOS [24]	MAUS	40	46	6
Thai	Yes	-	-	-	Global TIMIT Thai [25]	MAUS	40	38	5

[a] All pre-trained models are from `https://huggingface.co`. We used `wav2vec2-base` (English); `wav2vec2-FR-7K-base` (French); `mandarin-wav2vec2` (Mandarin); and `wav2vec2-base-vi` (Vietnamese).
[b] `https://huggingface.co/datasets/Cnam-LMSSC/vibravox`; speech_clean, headset_microphone subset

限的情景范围内。具体而言，[31] 分析了在法语和英语上训练的 LSTM 模型，但仅在匹配语言条件下；[15] 测试了 w2v2 模型在匹配和不匹配条件下，但仅适用于单一语言对（印地语-英语）；而 [16, in Supplementary Information] 报告了在荷兰语、普通话和法语上预训练的 w2v2 模型的结果，但仅测试英语。此外，还有多语言预训练模型的音素探测结果，但没有按语言进行详细分析 [32,33] 。最后，两项研究探测了在音调和非音调语言上训练和分析的 w2v2 模型中的词汇音调表示 [4,5] 。我们的一些探测分析重复了 [4,16] 的部分内容，但总体而言，我们报告的条件范围比过去的工作更为广泛。

关于几何分析，这些分析在文本模型中（例如，[34–38] ）已经很成熟，但在语音模型中则不然。我们只知道几个例子，其中大多数是研究其他类型的模型（例如，声学词嵌入 [39] 或端到端自动语音识别系统 [40] ）。我们在这里基于上述 [10,11] 的工作，他（像我们一样）研究了自监督模型，但他的研究仅限于英语。

3. 数据和模型

我们研究中使用的材料总结在表格 1 中。我们使用 12 层的 w2v2 模型 [14] ，这些模型在两种声调语言（普通话、越南语）和两种非声调语言（英语、法语）上进行了预训练，并加上一个等效的未训练（随机初始化）模型作为基线。我们选择这些模型是为了最大化模型之间的可比性——这些模型都有相同的架构和层数——以及与 [4] 之前的声调探测结果进行比较，他们使用了相同的模型。

我们的分析使用来自相同四种语言外加泰语的评估数据，对于泰语我们无法获得预训练模型。在所有情况下，评估数据与预训练数据是不同的。探测和正交性分析都要求每个语音帧带有讲话者、声调或音素的标签。对于讲话者，我们使用语料库的元数据。对于声调和音素，我们使用 MAUS [41] 或 MFA [42] 来获得强制对齐。我们排除了非语音声音和一些不在所有讲话者中出现的罕见音素。Table 1 列出了每种语言的音素和声调类别数量。[3]

对于每个模型和每个测试语料库，我们提取了来自所有 Transformer 层（从 1 到 12 编号）以及卷积特征提取器输出（层 0）的表示。所有表示都是 768 维的实数向量，对应于单个 20 毫秒的语音输入帧。这些表示及其对应的说话人、语调和音素标签被用于下述的所有分析中。

[3] 普通话有四个声调对比，还有第五个“轻声”，主要出现在非重读音节中 [43] 。根据 [4] ，我们忽略轻声，仅考虑每个音节的词汇声调，也就是说，我们不考虑连读变调。有时候，越南语被分析为有八个声调，其中两个仅限于有阻音音尾的音节，但在正字法中只有六个声调被区分。泰语和越南语都没有像普通话中那样的连读变调。这三种语言在基因上没有联系。

4. 分析 1：探测分类器

在该分析中，我们训练线性分类器来预测从不同模型或层中提取的表征中的音素、音调或说话者标签，以量化这些信息类型在线性编码中的程度。

如同在 [4, 15, 26, 31] 中的一样，电话和音调分类器的输入是通过对同一标记段（电话或音调）沿持续时间的表示进行平均获得的。说话者分类器的输入是通过采用对应的电话数据集并将电话标签替换为适当的说话者标签来创建的。

分类器的训练集和测试集通过随机采样（输入，标签）对创建。为了简单起见，我们没有为训练集和测试集使用不同的说话人集，部分原因是 [11] 报告说，说话人相关和说话人无关的电话探测准确率非常相似，并且有很强的相关性。训练集的大小固定在大约 25k 对，这对于所有类型来说都远超训练准确度饱和。测试集的大小固定在 10k 对，这可以提供约为 ±1% 的分类准确度的 95 % 置信区间。探测分类器是线性的，全连接神经网络，使用交叉熵损失实现多项逻辑回归，使用 Adam 优化器，学习率为 10^{-3} 。训练运行了五个周期。测试在最后一个训练周期后的分类器状态上进行。

4.1. 结果

Figure 1 显示了对两个语言类别（音素和声调）探测分类器的结果。在几乎所有情况下，音素探测准确性都高于声调探测准确性，尽管声调类别的数量要少得多。这表明在模型中，声调比音素更难以获取（线性可分）。从质量上看，这两个类别显示出类似的模式：准确性在初始时上升，在中间层达到峰值，并在最后层下降。这与之前对英语 [2,6,8,15] 和印地语 [15] 的 w2v2 模型的语音分析以及关于声调的 [4,5] 结果是一致的。

我们还观察到当预训练语言和测试语言匹配时，会有一致的优势。虽然这一效应之前已在多种语言上的声调探测 [4] 和印地语/英语上的音素探测 [15] 中被发现，我们的结果表明这一效应在多种语言和语言类别中都很普遍。此外，我们更全面的结果强调，匹配语言的优势似乎始终出现在第 4-6 层左右，匹配语言探测器的准确性继续提高，而不匹配的探测器则趋于稳定或开始下降。这表明，特定语言的表示仅在模型的中间层开始出现。最后，我们注意到，对于越南语，匹配语言的优势最为显著，这可能与其更加庞大和多样化的预训练数据有关。需要进一步的工作来完全拆解语言、数据量和语音风格的影响。

讲者 对于说话人探测（Figure 2 ），结果在不同的模型和测试语言中大致一致，但与音素和音调不同，我们没有发现匹配语言优势的证据，且整体结果更加杂乱。分层趋势也不同，在早期层（大约 2-4 层）中精确度最高，而在 5-9 层中精确度较低，在这些层中音素和音调探测表现

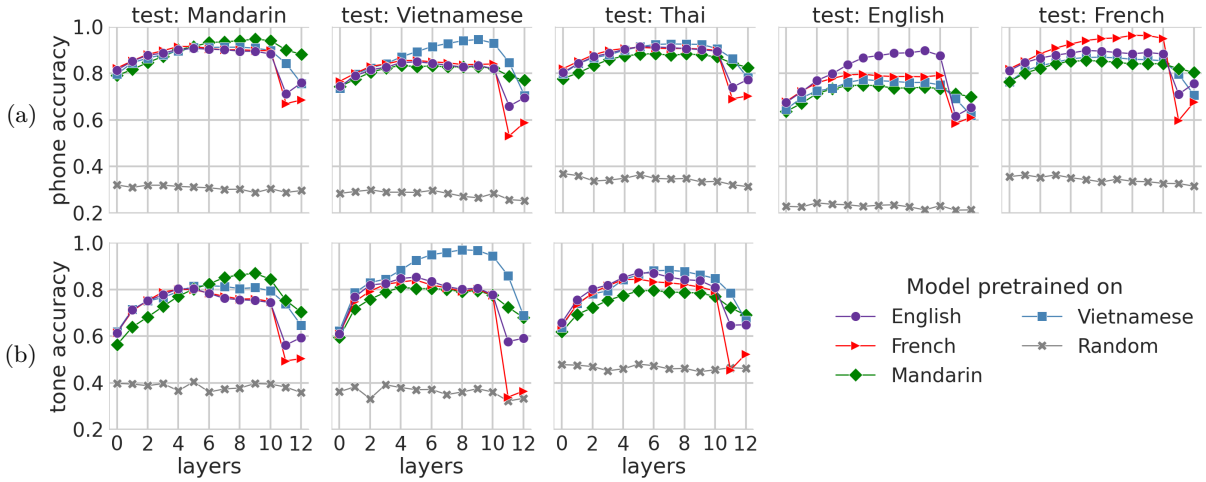

Figure 1: 各层探测分类器在五种不同测试语言上对 *(a)* 音素和 *(b)* 声调的准确率。每个图显示了四种不同的预训练模型和一个未经训练（随机）模型在单一测试语言上的结果。

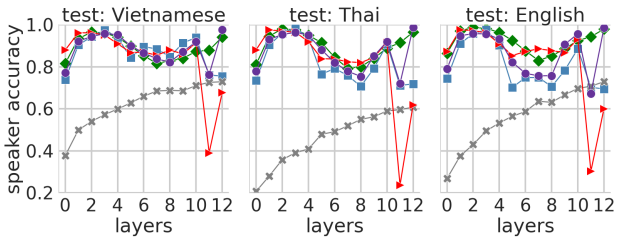

Figure 2: 在三种测试语言上的说话者检测准确率，对于与图 *1* 中相同的五种模型（使用相同的键）。其他两个测试语言（未显示）显示了类似的模式。

出最强的语言特定优势。说话人探测在第 11 层特别不稳定，此前在英语模型中已显示过异常行为 [2,4,5] 。我们发现法语模型中有更大的异常，在 11-12 层中甚至比随机模型还要差。（此异常在图 1 和 3 中也可见，并在附录 B 中有进一步分析）。

5. 分析 2：表现几何

我们遵循 [10,11] 中的方法来研究信息是如何在几何上表示的。该分析包括两个步骤：识别对应于音素、声调和说话者的子空间；以及测量每对子空间之间的正交性。

我们如 [10] 中所示识别每个子空间。首先，我们计算每个类别标签的质心：例如，对于语音子空间，我们对每个语音类别相关联的所有嵌入进行平均。这产生一个大小为 $N_c \times 768$ 的矩阵，其中 N_c 是语音类别的数量。通过对该矩阵进行主成分分析（PCA），我们获得跨越语音子空间的主成分（PCs）。我们以类似的方式获得说话者和语调子空间的主成分。

我们使用累计残差方差（CRV）[11] 来测量这些子空间之间的正交性。考虑两个由矩阵 X 和 Y 的主成分定义的不同子空间 R^n 。去除 Y 后的 X 的 CRV，记作 $X \setminus Y$ ，用于衡量在投影去除（“折叠”）Y 之后，子空间 X 定义的方差剩余多少。CRV 解释了 X 的给定主成分与 Y 中多种方向对齐的程度，以及与解释更多（更少）方差的主成分的对齐，也意味着 Y 与 X 的对齐程度更多（更少）。值为 1 表示 X 和 Y 正交（去除 X 的主成分对 Y 没有影响）；较低的值表明两个空间的主成分之间的对齐程度更高；而值接近 0 表示大部分的方差在 X 的第一个主成分中，并

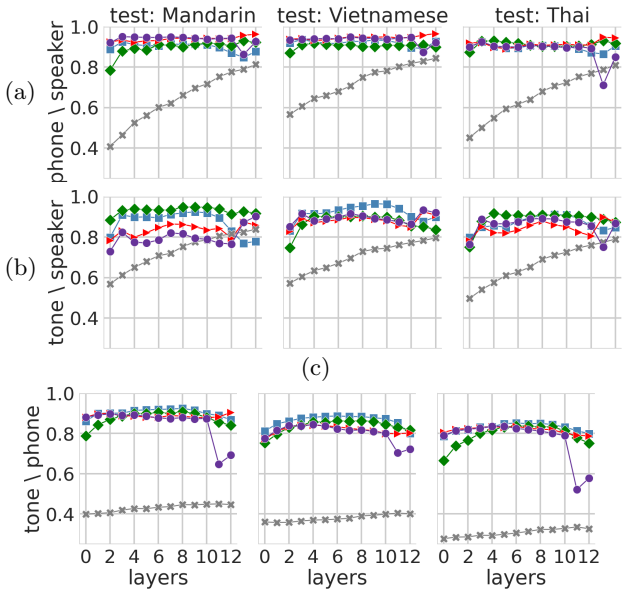

Figure 3: *CRV* 正交测量：*(a) Phone* 说话者（由于篇幅原因仅显示音调测试语言），*(b) Tone* 说话者，以及 *(c) Tone* 电话。附加 *CRV* 测量（*Speaker* 电话，*Speaker* 音调，和 *Phone* 音调）的结果可在附录 *A* 中找到，并显示出质上相似的模式。

且该主成分也与 Y 的第一个主成分紧密对齐。关于 CRV 的更精确和详细的描述，请参见 [11] 。对于电话和说话者子空间，在计算 CRV 时我们使用了前 35 个主成分。

5.1. 结果

Figure 3 报告了 CRV 测量结果。我们计算了所有组合 X\ Y 的 CRV，其中 X 和 Y 是 { 说话人、电话、声调 } 的不同值。涉及声调的 CRV 仅计算了三种声调语言。

首先看涉及音素和说话人的结果（图 3a ），我们发现之前在英语数据上使用英语预训练（或未训练）模型 [11] 找到的结果模式大致扩展到了其他预训练模型。特别是，音素和说话人子空间基本上是正交的，跨层之间只有小的变

Table 2: 每种语言在音节中的位置音调和音素之间的调整互信息（*AMI*）。*AMI* 的范围是从 *0*（统计独立）到 *1*（没有独立性）。

Language	Onset	Nucleus	Coda
Mandarin	0.048	0.020	0.002
Vietnamese	0.066	0.047	0.134
Thai	0.118	0.063	0.179

化（除了在英语模型第 11 层中的异常，已在第 4.1 节中提到）。与探测结果不同，我们没有看到任何明显证据表明预训练语言、测试语言和正交度之间存在明显关系。

带有声调的结果显示在图 3b 至 3c 中。图 3b 表明，去除说话者子空间后，声调子空间中保留了许多方差，尽管 CRV 不如 Phone\ 说话者 那么高（即，更低的正交性），而且与随机初始化模型相比，差异较小。这表明模型没有像音素和说话者信息那样（线性地）分离声调和说话者信息。此外，这里预训练语言的选择确实很重要，当（声调）预训练语言与测试语言匹配时，可观察到最高的 Tone\ 说话者 值。

对于 Tone\ 电话 结果（图 3c ），我们再次看到总体上相对较高的值，尤其是与随机基线相比。对于在声调语言上进行预训练的模型，正交性似乎略高，但最大影响是由于测试语言：在普通话上测试的模型得分高于在越南语上测试的模型，而后者得分又高于在泰语上测试的模型。如果我们关于统计独立性与表示空间中的正交性之间关系的假设成立，这些结果表明音调与音素之间的任何系统性相关性在泰语中最高，在普通话中最低，而越南语则位于中间的某个位置。

虽然一般分析认为声调在音节层面上运作且与宿主音节的段结构无关 [17] ，但它们之间通常存在一定程度的统计依赖关系，以音位配置约束的形式呈现，这种约束可以从弱到绝对 [44] 。为了检验我们研究中的三种声调语言的这种统计关系，我们使用从测试语料中提取的计数为每种语言计算了声调-音素共现。计数被分别收集用于音节的起始、核心和尾码，然后用来计算每种语言和音节成分的声调与音素之间的调整互信息 [45]（Table 2 ）。结果与我们的假设一致：声调和音素之间有高度的独立性，但也存在一致的排序，即普通话 < 、越南语 < 、泰语的顺序——与 Figure 3c 中的 CRV 结果的排序相同。虽然这还远未定论，但它表明表示的几何结构可能更多地依赖于嵌入语言的音位配置，而不是预训练时使用的语言材料的音位配置。

6. 结论

在本研究中，我们采用探测分类器和几何分析方法来检验自监督 w2v2 模型如何表征关于音素、声调和说话者的信息。与其他研究一致，我们发现我们的音素和声调探测器的准确性在网络中间层达到峰值。我们对几种语言的更广泛分析，包括跨语言测试，表明在第 5 到 9 层中，匹配语言探测对音素和声调的优势稳定出现，但相对较小。另一方面，说话者探测的准确性则表现出较大的变动性，没有表现出任何匹配语言的优势，并且在早期层（有时在最后一层）中达到峰值。我们对表征几何的分析表明，编码音素、声调和说话者的子空间大体上是正交的，并且正交性受预训练语言的影响较小，这可能反映了对音位组合约束的敏感性。总体而言，我们的研究结果表明，w2v2 架构学习以一种大体上独立于语言的方式编码语言和非语言信息，尽管在中到后期层中观察到一些段和超段特征的语言特定性。

7. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe et al. , "Self-supervised speech representation learning: A review," IEEE Journal of Selected Topics in Signal Processing , vol. 16, 2022.

[2] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in Proc. of ASRU , 2021.

[3] G. Shen, A. Alishahi, A. Bisazza, and G. Chrupała, "Wave to Syntax: Probing spoken language models for syntax," in Proc. of Interspeech , 2023.

[4] G. Shen, M. Watkins, A. Alishahi, A. Bisazza, and G. Chrupała, "Encoding of lexical tone in self-supervised models of spoken language," in Proc. of NAACL , 2024.

[5] A. de la Fuente and D. Jurafsky, "A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models," in Proc. of Interspeech , 2024.

[6] M. Yang, R. C. M. C. Shekar, O. Kang, and J. H. L. Hansen, "What can an accent identifier learn? Probing phonetic and prosodic information in a wav2vec2-based accent identification model," in Proc. of Interspeech , 2023.

[7] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? A layer-wise and neuron-level analysis," Computer Speech & Language , vol. 83, 2024.

[8] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in Proc. of ICASSP , 2023.

[9] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu, "What Do Self-Supervised Speech Models Know About Words?" Transactions of the Association for Computational Linguistics , vol. 12, 2024.

[10] O. D. Liu, H. Tang, and S. Goldwater, "Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces," in Proc. of Interspeech , 2023.

[11] M. Mohamed, O. D. Liu, H. Tang, and S. Goldwater, "Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations," in Proc. of Interspeech , 2024.

[12] A. Conneau, A. Baevski, R. Collobert, A. rahman Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in Proc. of Interspeech , 2021.

[13] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," Journal of Machine Learning Research , 2024.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. of NeurIPS , 2020.

[15] J. Rodriguez, K. Sreepada, R. L. Famularo, S. Goldwater, and N. Feldman, "Self-supervised speech representations display some human-like cross-linguistic perceptual abilities," in Proc. of NAACL , 2024.

[16] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning," Proc. of NeurIPS , vol. 35, 2022.

[17] J. A. Goldsmith, "Autosegmental phonology," PhD dissertation, Massachusetts Institute of Technology, 1976.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in Proc. of ICASSP , 2015.

[19] T. Parcollet, H. Nguyen, S. Evain, M. Z. Boito, A. Pupier, S. Mdhaffar, H. Le, S. Alisamir, N. Tomashenko, M. Dinarelli et al. , "LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech," Computer Speech & Language , 2024.

[20] J. Hauret, M. Olivier, T. Joubaud, C. Langrenne, S. Poirée, V. Zimpfer, and É. Bavu, "Vibravox: A Dataset of French Speech Captured with Body-conduction Audio Sensors," arXiv , 2024.

[21] K.-H. Lu and K.-Y. Chen, "A context-aware knowledge transferring strategy for CTC-based ASR," in Proc. of SLT , 2023.

[22] D. Wang and X. Zhang, "THCHS-30: A free Chinese speech corpus," arXiv , 2015.

[23] T. B. Nguyen, "Vietnamese end-to-end speech recognition using wav2vec 2.0," 2021. [Online]. Available: https://github.com/vietai/ASR

[24] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese speech recognition system," in Proc. of WLSI/OIAF4HLT , 2016.

[25] M. Liberman, J. Yuan, C. Cieri, and J. Wright, "Global TIMIT Thai," 2023.

[26] K. Martin, J. Gauthier, C. Breiss, and R. Levy, "Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration," in Proc. of Interspeech , 2023.

[27] K. Choi, A. Pasad, T. Nakamura, S. Fukayama, K. Livescu, and S. Watanabe, "Self-Supervised Speech Representations are More Phonetic than Semantic," in Proc. of Interspeech , 2024.

[28] D. Wells, H. Tang, and K. Richmond, "Phonetic analysis of self-supervised representations of English speech," in Proc. of Interspeech , 2022.

[29] J. Millet and E. Dunbar, "Do self-supervised speech models develop human-like perception biases?" in Proc. of ACL , 2022.

[30] J. Linke, M. Kádár, G. Dosinszky, P. Mihajlik, G. Kubin, and B. Schuppler, "What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers," in Proc. of Interspeech , 2023.

[31] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux, and G. Wisniewski, "Probing phoneme, language and speaker information in unsupervised speech representations," in Proc. of Interspeech , 2022.

[32] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in Proc. of NeurIPS , 2021.

[33] H. Xue, Q. Shao, K. Huang, P. Chen, J. Liu, and L. Xie, "SSHR: leveraging self-supervised hierarchical representations for multilingual automatic speech recognition," in Proc. of ICME , 2024.

[34] D. Mimno and L. Thompson, "The strange geometry of skip-gram with negative sampling," in Proc. of EMNLP , 2017.

[35] X. Cai, J. Huang, Y. Bian, and K. Church, "Isotropy in the Contextual Embedding Space: Clusters and Manifolds," in Proc. of ICLR , 2020.

[36] E. Hernandez and J. Andreas, "The Low-Dimensional Linear Geometry of Contextualized Word Representations," in Proc. of CoNLL , 2021.

[37] T. Chang, Z. Tu, and B. Bergen, "The Geometry of Multilingual Language Model Representations," in Proc. of EMNLP , 2022.

[38] K. Park, Y. J. Choe, and V. Veitch, "The Linear Representation Hypothesis and the Geometry of Large Language Models," in Proc. of Causal Representation Learning Workshop at NeurIPS , 2023.

[39] B. Abdullah and D. Klakow, "Analyzing the Representational Geometry of Acoustic Word Embeddings," in Proc. of BlackboxNLP Workshop at ACL , 2022.

[40] C. Stephenson, J. Feather, S. Padhy, O. Elibol, H. Tang, J. McDermott, and S. Chung, "Untangling in invariant speech recognition," Proc. in NeurIPS , 2019.

[41] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," Computer Speech & Language , 2017.

[42] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Proc. of Interspeech , 2017.

[43] J. Cao, "On neutral-tone syllables in Mandarin Chinese," Canadian Acoustics , 1992.

[44] J. Kirby, "Incorporating tone in the calculation of phonotactic probability," in Proc. of SIGMORPHON Workshop at ACL , 2021.

[45] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," Journal of Machine Learning Research , vol. 11, 2010.

Appendices

A. 完整的图集

Figure 4 reports all results of probing classifiers. Figure 5 reports all CRV measurements.

B. 第 11 层的异常

In Figure 4, a severe drop in probe accuracy values at layer 11 for the representations from the English- and French-trained models are apparent. CRV values in Figure 5 partly exhibit similar trends for the same models. The anomaly for the English checkpoint was also reported by [5]. In order to get closer to the origin of such anomaly, we analyzed the magnitude of the representation vectors as follows. For each layer, model and test data, mean and standard deviation of the magnitude of the representation vectors were computed. For simplicity, these computations were conducted on the set of per phone (or per speaker) aggregations, as defined in the main text. Formally, given an aggregate matrix X of dimensions $N_C \times d$, where each row x_i is a d -dimensional vector ($d = 768$) representing the mean of a phone (speaker) class, the following three scalar quantities were computed:

$$\begin{aligned} \mu_{|x|} &= \frac{1}{N_C} \sum_{i=1}^{N_C} |x_i| \\ \sigma_{|x|} &= \sqrt{\frac{1}{N_C - 1} \sum_{i=1}^{N_C} (|x_i| - \mu_{|x|})^2} \\ |\bar{x}|\,,\ \bar{x} &= \frac{1}{N_C} \sum_{i=1}^{N_C} x_i, \end{aligned} \tag{1}$$

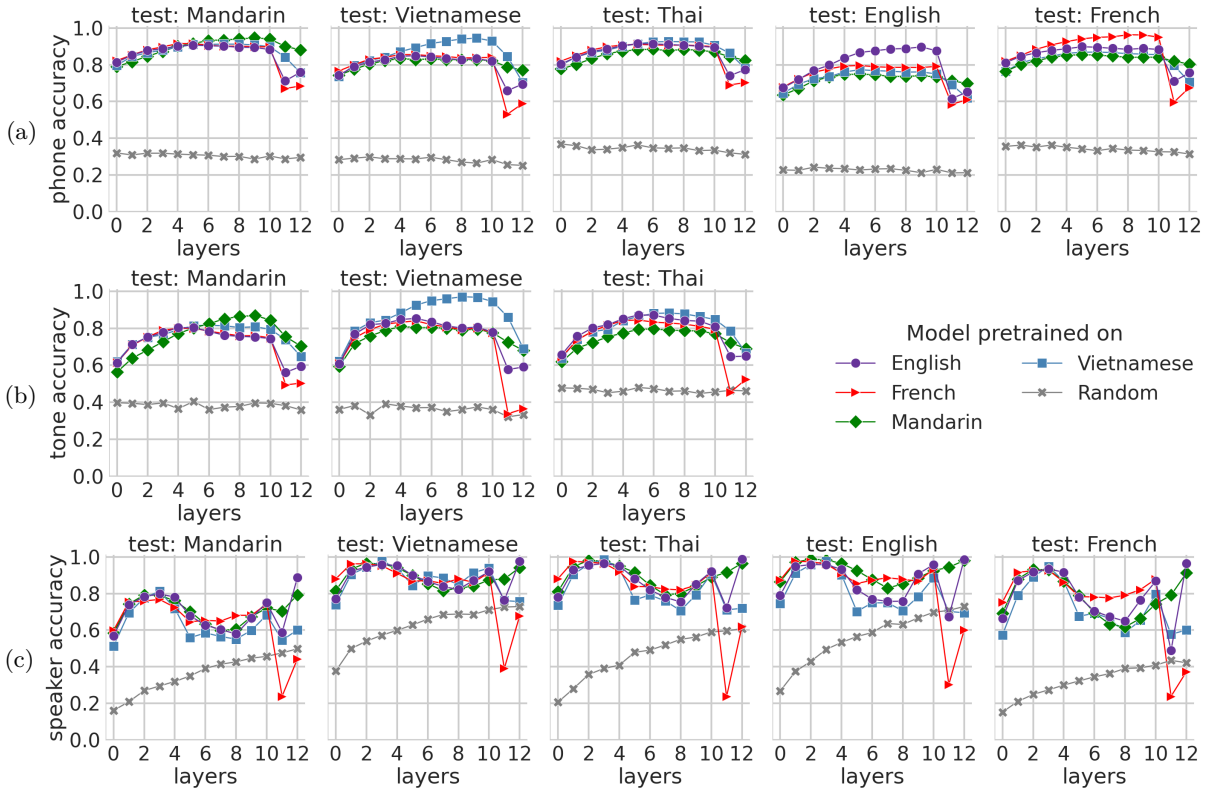

Figure 4: 层级探测分类器在五种不同测试语言中针对 *(a)* 音素，*(b)* 声调和 *(c)* 说话者的准确性。每个图显示了四种不同预训练模型和一个未训练（随机）模型在单一测试语言上的结果。**注释**：在普通话测试中，模型的整体较低的说话者探测准确性可能是由于标注上的一个问题。*THCHS-30* 语料库中的文件命名约定表明有 *60* 个说话者，而参考文献中声明有 *40* 个说话者（且探测结果的混淆矩阵表明这一数字更接近真实数量）。由于我们的说话者标注基于文件名编码，因此可能有些标签组指的是同一个说话者，导致性能的（表面）下降。我们认为这个问题不会实质性地影响本文中的其他结果（特别是正交性测量指标），因为拆分一位说话者的数据仅仅会为该说话者创建多个中心点。由于这些中心点彼此非常接近，它们不应该对说话者中心点的 *PCA* 捕获到的额外方差贡献太多，因此仅会影响最不重要的 *PCA* 成分。

where $\mu_{|x|}$ and $\sigma_{|x|}$ are the mean and standard deviation of the magnitudes of the representations, $|\bar{x}|$ is the magnitude of the mean of the representations. These statistics computed on the phone aggregate matrices are displayed in Figure 6; very similar results were obtained from speaker aggregates.

The first and the third row of plots report values for $\mu_{|x|}$ and $|\bar{x}|$, respectively. The fact that these quantities are very similar tells us that at layer 11, as well as in the other layers, representations are more likely to be all in a cloud away from the origin, rather than being on a shell surrounding the origin, which would have resulted in $|\bar{x}| \ll \mu_{|x|}$. A trace of an anomaly appears at layer 11 for the representations from the French-trained model, where the average distance from the origin remains constant but its standard deviation decreases dramatically. This fact may or may not be related to the drop of classification accuracy across the board at layer 11 for those representations visible in Figure 4. A similar phenomenon, though much less pronounced, is also visible for the representations from the Mandarin-trained model, while no major deviation at layer 11 appear for the English-trained model.

While we have no explanation for these anomalies at present, we can at least attest that these are neither English-specific nor strongly dependent on the matching between training and test language. Considering the results reported in Figure 2 of [11], we are inclined to believe that these anomalies are architecture-specific, as layers 11 and 12 of wav2vec2 stand out in some of the results, compared to all other architectures.

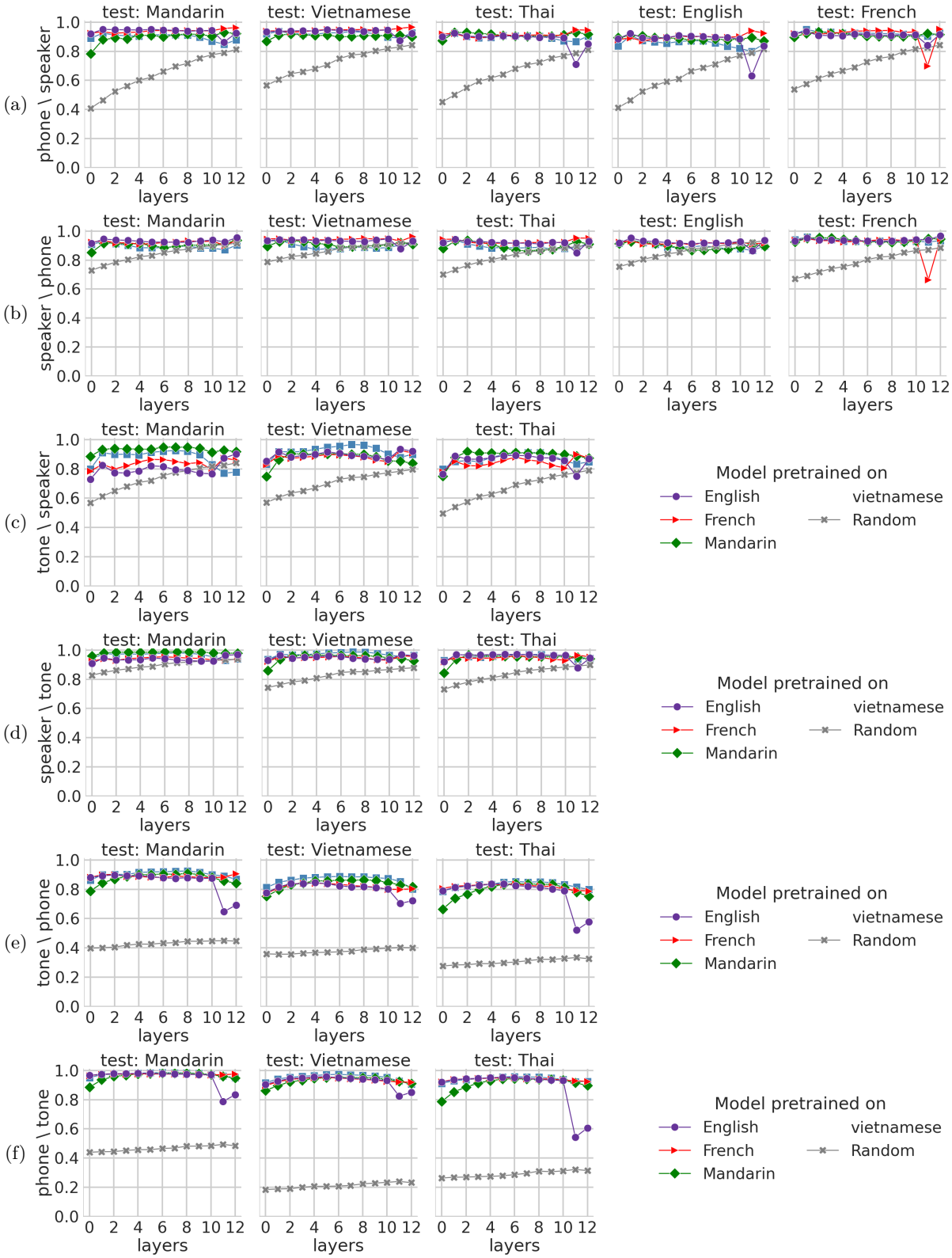

Figure 5: *CRV* 正交度量为 *(a) Phone*\ 说话者，*(b) Speaker*\ 电话，*(c) Tone*\ 说话者，*(d) Speaker*\ 语调，*(e) Tone*\ 电话，*(f) Phone*\ 语气。

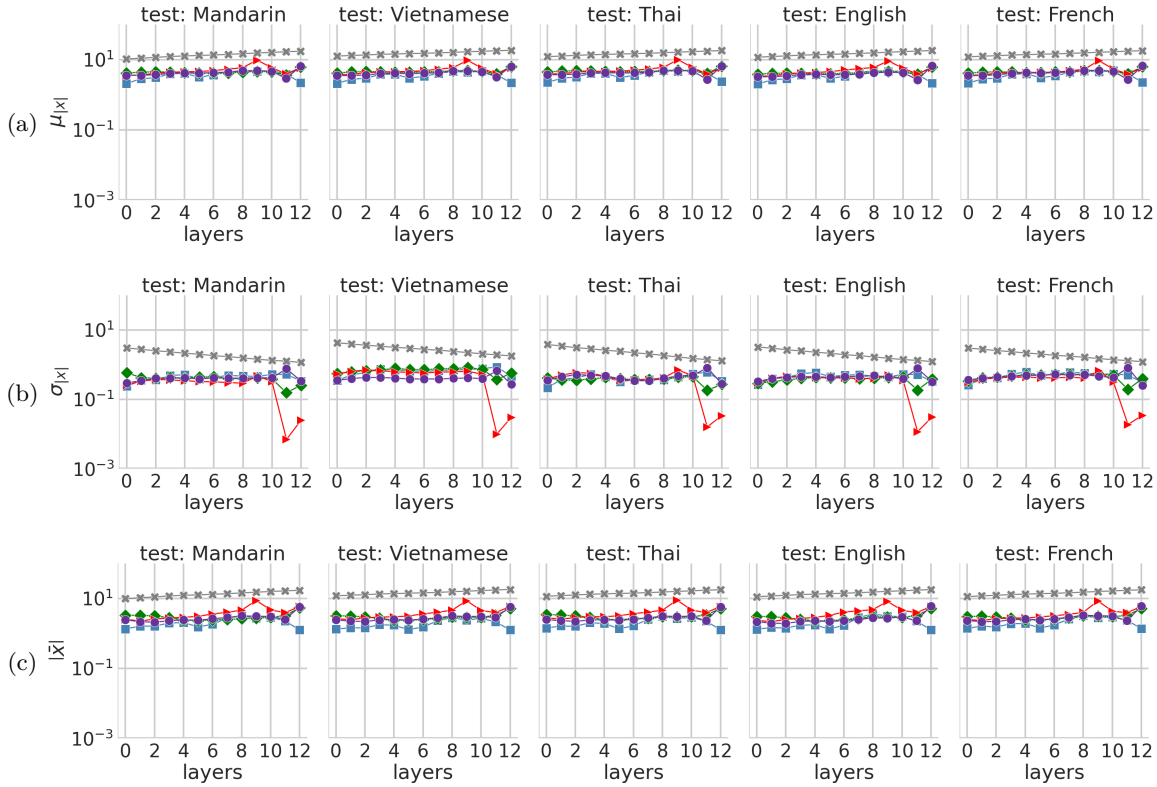

Figure 6: 在对数刻度上，方程 (1) 中定义的量：$\mu_{|x|}$ (a)，$\sigma_{|x|}$ (b)，$|\bar{x}|$ (c)。颜色编码与 *Figure 1* 和 *3* 中相同。