分析自监督模型中的说话人信息改进无监督语音识别

摘要

语音表征的质量是无监督语音识别成功的关键，自监督模型中包含多种音频信息（例如：说话人、性别等），如果能去掉非语音信息对无监督语音识别是很有帮助的。在本文中，我们首先通过定性和定量分析自监督的每话语均值在很大程度上捕获了说话人信息。接下来，应用对特征进行归一化可以有效地去除说话者信息。使用 TIMIT 数据集分别在 wav2vec2.0 和 hubert 模型上进行实验，我们得出说话人归一化方法可以显著提高无监督语音识别的性能。

一、引言

在先前的工作中[1,2]使用数千小时有标注的语音数据在主流语言上语音识别性能取得了不错的效果，可见为更多语言构建语音识别系统的限制是需要大量有标签的数据。在低资源语音上，带标签语音数据的获取是很困难的，通常需要耗费大量的人力、物力和财力。

幸运的是，无监督学习在机器翻译任务中取得巨大成功。现在已经有一些[3-5]以完全无监督的方式训练语音识别系统的方法，我们遵循 wav2vec-U[5]的架构，通过从自监督语音表征模型【综述】中提取高级语音表征，然后用对抗生成网络[6]以无监督的方法进行训练，将语音转录为音素序列。

由于语音表征的质量是无监督语音识别成功的关键[5],在这项工作中我们首先对比了使用 wav2vec2.0[7]和 hubert[8]两种语音表征模型提取特征，用作无监督语音识别的效果，即使用不同的特征提取前端，从实验结果得出，使用同等参数量的预训练模型，hubert 在无监督语音识别任务中表现更佳。更进一步，我们分析了自监督语音表征模型中的说话人信息。我们定性地（第 3.1.1 节）和定量地（第 3.1.2 节）分析表明，在 wav2vec2.0 和 hubert 特征上的每话语均值捕获了的说话人信息。和[9]中不同的是，我们将一个简单的说话人归一化步骤添加到音频预处理阶段中，可以从语音高级表征中有效地去除说话人信息。然后，我们展示了使用说话人归一化后无监督语音识别的性能。

二、相关工作

使用原始语音进行训练的自监督学习在最近的研究中变得越来越流行，因为自监督模型在各种下游任务中取得了先进的结果，包括说话人识别、语音识别、意图分类等。最近比较流行的自监督表征模型有 wav2vec2.0、hubert 等。在这些自监督模型的基础上，已经有一些无监督方法进行 ASR：

1.wav2vec2.0

wav2vec2.0 模型的思想来源于 CPC 模型，结合了掩码和对比学习，它使用 InfoNCE 损失[10]使得上下文表征和量化的目标向量之间的距离最小。具体来说，将原始音频输入多层卷积特征编码器中，输出潜在语音表征。然后将他们送入上下文网络 Transformer[11]中去，从而捕获整个序列的信息。另一方面，特征编码器的输出通过量化模块生成自监督学习的目标。

2.Hubert

Hidden Unit BERT（HuBERT）模型没有依靠先进的表征学习模型来离散连续的语音输入，而是使用在 MFCC 特征上训练的经典 k-means 单元的有效性。此外，与 wav2vec2.0 类似 HuBERT 模型使用连续波形（或 log-mel 特征)作为特征编码器的输入，以确保自监督模型获得预训练和微调所需的完整输入语音信息。具体

来说，HuBERT 模型使用带掩码的连续语音特征来预测预先确定的类别分配。预测损失仅应用于被屏蔽的区域，迫使模型学习未屏蔽输入的良好高级表征，以正确推断被屏蔽输入的目标。

3.wav2vec-U

Wav2vec-U 或 wav2vec Unsupervised，利用 wav2vec2.0 的自我监督表示来嵌入语音音频，并使用简单的 k-means 聚类方法将音频转换为单元。使用对抗训练学习语音片段和音素之间的映射，显示了很好的效果。具体来说，他们通过自我监督的预训练模型 Wav2vec 2.0 来提取声学表征，该模型由一个可以将原始波形转换为潜在表征的卷积特征编码器和一个 Transformer[10]组成，用于通过自监督的目标学习上下文表征。接下来，他们对 Wav2vec 2.0 提取的表征进行聚类，应用 PCA 来降低维度，并对同一聚类中的表征进行合并以生成浓缩的片段表征。他们应用 GAN 将音段表征映射到音素预测中，初步的音素预测已经以很大的幅度超过了以前的工作。最后，为了进一步完善预测，他们采用了 HMM 的自训练，通过将 GAN 的结果作为伪标签或通过伪标签对另一个 Wav2vec 2.0 模型进行微调来重新训练 HMM。在随后的自训练后，错误率接近于有监督的 ASR 模型。

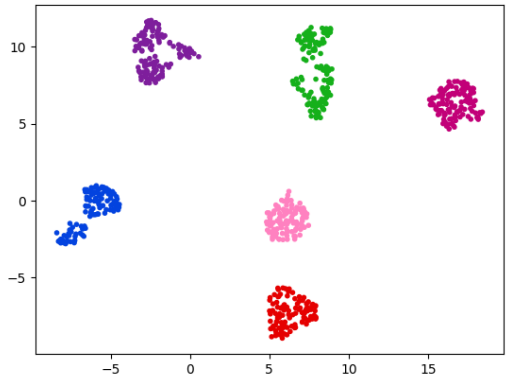
三．实验方法

这项工作的主要兴趣是研究不同预训练模型提取的特征对无监督 ASR 方法 Wav2vec-U 的影响。我们的工作重点是分析自监督模型中的说话人信息，并验证说话人归一化方法对无监督训练的影响，所以我们没有使用 HMM 自训练和微调等技巧进行完善。在生成器的输出，我们使用 4-gram 和 Viterbi 解码测量音素错误率（PER）。

3.1 自监督模型特征分析

3.1.1 定性分析

以前的工作[12,13]表明，自监督模型特征同时捕获语音和说话人信息。然而，目前尚不清楚表示如何重构这些信息。我们假设特征的每话语均值捕获了大量的说话人信息。这是合理的，假设说话者身份在一段话语中保持不变，而语音内容在较短的时间尺度内变化[14]。作为验证这一假设的第一步，我们使用 UMAP[15]探索 wav2vec2.0 和 hubert 两种特征。图 1(a)显示了从 LibriSpeech train-clean 集[16]中选择的六个说话者的 wav2vec 特征的每次话语平均值降维后的结果，图 1（b）显示了使用 hubert 特征使用相同降维方法和数据集的效果。两幅图都可以看出不同的说话人明显分开，表明句子的均值确实捕获了说话人信息。



（图上添加说话人 id）

图 1（a）wav2vec 特征可视化

待补充图

图 1（b）hubert 特征可视化

3.1.2 定量分析：

在本节中，我们定量验证自监督模型特征的每话语均值捕获说话人信息的大小。我们表明，简单地比较方法在说话人验证任务上表现良好。给定一组注册话语，说话人验证的目标是确定新话语是否属于特定说话人。为了设置任务，我们为 LibriSpeech dev-clean 集中的每个说话人随机选择了 4 个注册话语，保留其余部分用于测试。我们比较了三个系统的等错误率(EER)指标。

在这里的说话人验证任务中，我们的基线系统使用 MFCC 特征，该系统为自监督模型的性能提供一个下限。

第二个系统是基于 wav2vec 特征的方法，第三个系统是基于 Hubert 特征的方法。对于这两种模型我们采取相同的手段，在注册步骤中，我们提取 CPC 特征并计算每个话语的平均值。然后聚合这些方法以找到单个说话者嵌入。在测试时，我们使用余弦相似度打分将话语的特征均值与参考说话人嵌入进行比较。对于 EER 指标，我们设定距离以决定测试话语是否与给定说话者匹配。

从表一的结果可以看出，基于 wav2vec 和 hubert 的方法明显优于基线系统，这证明了在自监督模型中话语的均值确实可以捕获说话人信息。

表 1：基于 wav2vec、hubert 和 mfcc 系统的说话人验证结果

特征类型	EER(%)
Mfcc 的均值	15.8
Hubert 的均值	9.2
Wav2vec 的均值	6.3

基于上述观察，我们提出在无监督语音识别系统的流程中加入说话人归一化步骤。说话人归一化的简单定义如下：给定来自单个说话者的话语（或一组话语），我们通过减去均值并缩放到单位方差来从自监督模型特征中删除说话者信息。

$$f'(x) = \frac{f(x)-u}{\sigma} \quad (1)$$

在公式 1 中，f(x)为通过自监督模型提取的特征，u 为每句话语的均值，σ 为每句话语的标准差。通过这样简单的方法可以有效地从高级语音表征中去除说话人信息，从而减少说话人信息对下游任务的影响。

3.2 模型架构

我们的实验遵循 wav2vec-U 的实验核心架构，分为前端特征提取、特征预处理、无监督训练三个部分。具体的结构如图 2 所示，在前端特征提取部分分别使用 wav2vec 和 hubert 两种自监督模型，另外通过在特征预处理步骤中通过加入说话人归一化方法，从原始高级语音表征中删除说话人信息，从而减少无用的说话人信息对无监督语音识别的影响。

3.2.1 语音音频表征

如图 2 中所示，我们使用自监督模型嵌入未标记的语音音频以获得语音表征。具体来说，我们使用上下文 Transformer 网络的输出 c1,…,cT。每个 ct 表示以当前时间步 t 为中心，使用自注意力机制从整个话语计算的特征；随后的上下文表征之间有 20 毫秒的间隔。在这里我们分别使用了 wav2vec2.0 和 hubert 的 big 设置的预训练模型，其中上下文网络包含 24 个 Transformer 块，我们将块 *l*

在时间步 t 的输出表示为 c^{l}_{t}。从之前的经验中，Transformer 中前面块编码句法信息，而最上面的块编码语义信息。我们的任务是要找到从当前时间步长到音素的良好映射，所以我们在这里使用中间层第 15 层来提取特征。

3.2.2 语音表征分段

一旦嵌入了语音信号，我们就会识别出与有意义的单元相对应的片段，这些单元可以映射到音素。分割已被证明在先前的工作中至关重要[17]，因为输入表示中的正确边界可以更容易地预测输出序列。在无监督语音分割方面已经有很多先前的工作[18,19]，但这里我们只是使用了一种基于对 wav2vec2.0 语音表示 c1,...,cT 进行聚类的方法。第一步，我们收集未标记语音数据的所有语音表示，并执行 k-means 聚类以识别 K=128 个聚类。我们使用 FAISS 库在 GPU 上进行快速聚类[20]。接下来，每个 ct 都用相应的集群 ID $i_t \in \{1,..,K\}$进行表示，并且我们在集群 ID 改变时引入语音段边界。

一旦对语音音频表示进行分割，我们就会在训练集的 wav2vec2.0 输出的所有语音表示上计算 512 维 PCA。接下来，我们对特定片段的 PCA 表示进行均值池化，以获得片段的平均表示。PCA 仅保留最重要的功能，我们发现这是有效的。由于缺乏监督，段边界很嘈杂，因此我们发现对相邻段表示的平均池对也很有用，以提高鲁棒性。这导致对于给定的话语，生成语音片段表征序列 S=s1,...,sT。

3.2.3 说话人归一化

在前面我们已经分析了自监督模型中的说话人信息，如公式 1 所示可以通过一个简单的归一化方法从原始的高级语音表征中去除说话人信息。在这里我们对每个句子的分段表示执行说话人归一化方法，生成更加纯净的语音表征向量，然后将其送入 GAN 网络的生成器当中去。

3.2.4 文本数据预处理

类似于我们如何将未标记的语音音频数据分割成适合无监督学习的单元，我们对未标记的文本数据执行相同的操作。我们对文本数据应用两个预处理步骤：音素化和静音标记插入。

音素化。音素表征了将单词彼此区分开来的不同声音，例如，对于单词 cat 有三个音素对应于单词发音中的三个不同声音：K、AE、T。我们对文本数据进行音素化，因为我们发现学习语音音频和单词的不同声音之间的映射比学习音频和单词或字母之间的映射更容易。音素化将单词序列 Y 转换为音素序列 P=[p1,⋯,pM]，其中 pm∈O 和 O 是音素清单。我们使用现成工具 Phonemize 来完成这一步骤。

静音标记插入。通过应用无监督静音去除对未标记的语音音频数据进行预处理。然而，这个过程并不总是准确的，语音音频中仍然存在许多静音。为了解决这个问题，我们使无监督模型能够用音素静音标记（SIL）标记一些片段。我们通过在音素化的未标记文本数据中插入静音标记来解决这个问题。首先，我们在所有音素化的未标记文本句子的开头和结尾添加一个 SIL 标记。其次，我们在单词之间或与单词对应的音素组之间以一定的速率随机插入 SIL 标记。

3.2.5 无监督训练

本文使用生成对抗训练来训练无监督语音识别模型，该模型使用未标记的语音音频数据和未标记的音素化文本数据。生成对抗网络由生成器网络和判别器网络两部分组成，其中生成器产生的样本由鉴别器进行判断。对鉴别器进行训练，以分类样本是来自生成器还是来自真实数据分布。生成器的目标是产生鉴别器无法区分的样本。

具体而言，G 的输入为长度为 T 的片段表征 S=[s1,...,sT]，然后将其映射到长度为 M 的音素序列 G(S)=[p1,...,pM]。生成器预测每个片段在音素集 O 上的分布，并输出概率最高的音素。如果连续片段的 argmax 预测导致相同的音素，那么我们对这些片段之一进行采样，因此 M≤T。音素集 O 包括一个静音标签 SIL，以便能够在语音音频中标记静音。与随后的语言模型解码相联系。在反向传播中，我们通过在生成器输出处采样的片段进行反向传播。在无监督训练过程中，我们不修改片段表征 S。生成器是一个单层卷积神经网络。鉴别器将一系列序列作为输入，该序列表示来自真实数据分布 Pr 的音素化文本的 One-hot 向量，或来自生成器 G（S）的输出分布序列。每个输入向量都有|O|维来表示每个段的音素分布。鉴别器也是一个 CNN，它输出一个概率，指示样本来自数据分布的可能性。

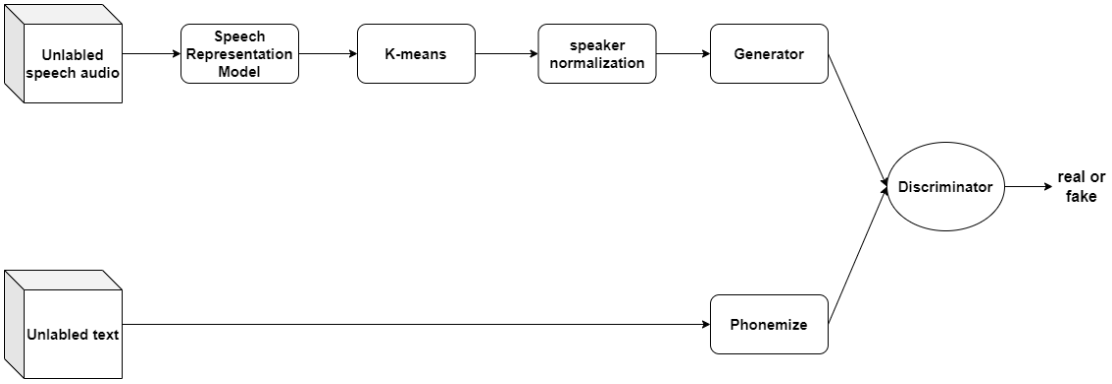


图 2：系统架构图

四、实验

1.实验数据

为了尽可能的模拟在低资源语言上仅有受限数据可用的情况，我们使用了 TIMIT 数据集，它包含大约五个小时的录音，带有时间对齐的语音转录本[21]。为了与之前的工作进行比较，我们考虑了两种设置：matched 设置使用来自同一组话语的文本和语音来训练模型，而 unmatched 设置确保未标记的文本数据不包含音频数据的转录。对于 matched 的设置，我们按照[4]等人所做的 TIMIT 的标准训练的 train dev test 切分。这是 3,696/400/192 个训练开发测试话语，其中仅包含 SX（紧凑）和 SI（多样化）句子。对于 unmatched 的设置，我们遵循[3]通过对完整数据集拆分的训练部分的 3000 条语音和 1,000 条转录进行训练。我们使用剩余的 620 个训练话语进行验证，并测试 1680 个句子进行测试。完整的数据集拆分包含 4,620 个训练语句和 1,680 个测试语句，以及额外的 SA（方言）句子。

2.实验一

我们考虑两种设置来与以前的工作进行比较：在匹配设置中，未标记的文本数据只是未标记的音频数据的转录，但未配对。在不匹配的设置中，未标记的文本数据不包含音频数据的转录，这是更现实的设置。我们测量标准 Kaldi 开发和测试集(core-devcore-test)以及稍大版本的测试集(all-test)的性能。我们报告了 wav2vec-U 在 TIMIT 语言建模数据上训练的 4-gram 语言模型的性能，表 2 的 2-5 行表明，使用 hubert 模型的性能更优。

3.实验二

在上述实验的基础上，我们在两种模型上添加说话人归一化步骤从中去除说话人信息。然后同样在 TIMIT 的 matched 和 unmatched 两种设置中进行了实验，结果如表 2 第 6-9 行所示，可以看出添加说话人归一化方法后，效果有明显提升。

表 2：在 TIMIT 数据集上 matched 和 unmatched 两种设置的实验结果

Model	LM	train	dev	test
Wav2vec_matched				
Wav2vec_unmatched				
Hubert_matched				
Hubert_unmatched				
Wav2vec_SN_matched				
Wav2vec_SN_matched				
Hubert_SN_matched				
Hubert_SN_unmatched				

五、结论与展望

我们分别利用三种不同的自监督语音表征模型提取特征，使用 hubert 自监督语音表征模型对无监督 ASR 具有更好的效果，同时应用说话人归一化方法可以降低无监督 ASR 的 WER。在未来，我们将探索从自监督语音表征模型中去除更多的非语音内容信息，那么将对无监督语音识别有很大帮助。

参考文献：

1. Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.
2. Han W, Zhang Z, Zhang Y, et al. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context[J]. arXiv preprint arXiv:2005.03191, 2020.
3. Liu D R, Chen K Y, Lee H, et al. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings[J]. arXiv preprint arXiv:1804.00316, 2018.
4. Yeh C K, Chen J, Yu C, et al. Unsupervised speech recognition via segmental empirical output distribution matching[J]. arXiv preprint arXiv:1812.09323, 2018.
5. Baevski A, Hsu W N, Conneau A, et al. Unsupervised speech recognition[J]. Advances in Neural Information Processing Systems, 2021, 34: 27826-27839.
6. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
7. Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.
8. Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.
9. van Niekerk B, Nortje L, Baas M, et al. Analyzing speaker information in self-supervised models to improve zero-resource speech processing[J]. arXiv preprint arXiv:2108.00917, 2021.
10. Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural

information processing systems, 2017, 30.
12. Liu A T, Li S W, Lee H. Tera: Self-supervised learning of transformer encoder representation for speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2351-2366.
13. Chung Y A, Zhang Y, Han W, et al. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021: 244-250.
14. Hsu W N, Tang H, Glass J. Unsupervised adaptation with interpretable disentangled representations for distant conversational speech recognition[J]. arXiv preprint arXiv:1806.04872, 2018.
15. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.
16. Panayotov V, Chen G, Povey D, et al. Librispeech: an asr corpus based on public domain audio books[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 5206-5210.
17. Chung Y A, Weng W H, Tong S, et al. Unsupervised cross-modal alignment of speech and text embedding spaces[J]. Advances in neural information processing systems, 2018, 31.
18. Kamper H, Jansen A, Goldwater S. A segmental framework for fully-unsupervised large-vocabulary speech recognition[J]. Computer Speech & Language, 2017, 46: 154-174.
19. Kreuk F, Keshet J, Adi Y. Self-supervised contrastive learning for unsupervised phoneme segmentation[J]. arXiv preprint arXiv:2007.13465, 2020.
20. Johnson J, Douze M, Jégou H. Billion-scale similarity search with gpus[J]. IEEE Transactions on Big Data, 2019, 7(3): 535-547.
21. Garofolo J S. Timit acoustic phonetic continuous speech corpus[J]. Linguistic Data Consortium, 1993, 1993.