监督式声学嵌入及其跨语言可迁移性

Sreepratha Ram阿
联酋大学 sree_
ram@uaeu.ac.ae

Hanan AldarmakiMBZUAI
hanan.aldarmaki@mbzuai.ac.ae

arXiv:2301.01020v1 [cs.CL] 3 Jan 2023

摘要

在语音识别中，必须对输入信号的语音内容进行建模，同时剔除说话人变化、噪声等无关因素；而在低资源环境下，这一目标颇具挑战。人们提出了自监督预训练，以改进有监督和无监督语音识别，其中包括帧级特征表示，以及用于变长片段的声学词嵌入（AWE）。然而，自监督模型仅凭自身无法完美分离语言内容，因为它们的训练目标是优化间接目标函数。本研究将不同的预训练自监督特征作为声学词嵌入模型的输入进行实验，并表明这些特征在监督式框架下效果最佳。在英语上训练的模型无需适配即可迁移到其他语言，其表现优于仅在目标语言上训练的自监督模型。

关键词— 无监督自动语音识别、迁移学习、声学词嵌入

1 引言

随着大量标注数据和计算能力的 उपलब्ध，监督式语音识别系统变得更加稳健、准确（Gulati等人，2020年；Baevski等人，2020b），人们如今更加关注训练数据有限或完全缺乏的低资源语言（Aldarmaki等人，2022）。利用未标注语音进行无监督预训练，可提升有监督模型和无监督模型的性能；例如，研究表明，在多种语言的大量未标注语音上预训练的语音表示，能够提升低资源语言的自动语音识别性能（Kawakami等人，2020年；Conneau等人，2020年）。

虽然大多数有监督自动语音识别模型都在音素层面运行，但词级分段式自动语音识别通过对变长片段进行建模并将其嵌入为固定维度向量（Abdel-Hamid等人，2013；He和Fosler-Lussier，2015），也已得到探索，并取得了相对成功。类似地，声学词嵌入（AWEs）被提出作为一种在低资源环境中高效比较变长语音片段的方法（Peng等人，2020年；Kamper等人，2020年）。与书面词语不同，口语词语天然包含说话人差异和语音变异，因此在没有监督的情况下，更难在潜在空间中对其建模。自监督预训练和跨语言迁移是两种可行的方法，可使无监督模型更能适应输入信号中的非语言变化。

在本研究中，我们考察了声学词嵌入模型的自监督训练与有监督训练在零样本跨语言迁移中的表现。我们尝试了不同类型的声学特征，并分别评估它们的性能以及它们在声学词嵌入模型中的表现。我们发现，预训练声学特征在一定程度上提升了自监督声学词嵌入模型的性能；而使用少量来自另一种语言的标注数据，以有监督方式训练声学词嵌入模型，则能取得更大的提升。在不同语言中都持续观察到这种零样本跨语言迁移，尤其是在使用预训练特征表示时。我们的结果表明，与纯粹的自监督训练相比，采用零样本跨语言迁移的有监督训练，是低资源语音模型更有效的方法[1]。

2 背景与相关工作

口语通常使用时长为 10 至 30 毫秒的短定长帧进行建模，这

[1]我们提供 Python 训练和评估脚本，用于复现我们的实验：https://github.com/h-aldarmaki/acoustic_嵌入

会产生变长词片段。动态时间规整（DTW）是一种早期技术，它利用动态规划寻找最优的逐帧对齐，从而比较变长片段。动态时间规整效率较低，因此人们转而将变长片段嵌入为固定大小的向量，以便使用余弦距离或欧氏距离等更高效的度量进行比较（Levin等人，2013）。已有研究提出了不同类型的声学词嵌入（AWE）。由于这些技术通常面向低资源语言，因此一般以自监督方式训练，最常见的方法是使用带有重构损失的自编码器网络（Chung 等人，2016；Holzenberger等人，2018年）。与直接使用动态时间规整进行比较相比，这些 AWE 通常能取得相近或略优的性能，同时效率高得多（Holzenberger等人，2018）。Peng等人（2020年）介绍了一种使用对应自编码器的替代训练策略，该策略依赖通过无监督口语术语发现提取的词对；此外，采用对比学习和多语言适配还能进一步提升效果（Jacobs等人，2021年）。

上述模型使用静态声学特征（例如梅尔频率倒谱系数）作为输入。van Staden 和 Kamper (2021年)表明，使用 CPC（van den Oord 等人，2018年）等预训练特征能够提升无监督声学词嵌入模型的性能。已有研究反复表明，预训练特征能够提升有监督下游任务的性能（Yang等人，2021年）。此外，研究还表明，预训练特征能够跨语言迁移。例如，Riviere等人(2020年)介绍了 CPC 的一个修改版本（MCPC），并证明在英语上对这些特征进行预训练，能够提高其他语言的音素分类准确率。其他类型的预训练特征，例如 wav2vec 2.0（Baevski等人，2020a），已被证明能够提升有监督和无监督自动语音识别性能（Baevski等人，2021年），而对这些特征进行多语言训练（即 XLSR-53）与单语言预训练相比，能够提升多种语言的性能（Con- neau等人，2020年）。

3 目标与方法

本研究旨在探究预训练声学特征用作声学词嵌入模型输入时的有效性和可迁移性。为此，我们比较了直接在目标语言上训练的自监督声学词嵌入，以及在另一种源语言上训练的有监督声学词嵌入的零样本跨语言迁移。据我们所知，预训练特征与声学词嵌入模型的结合尚未得到充分研究；大多数声学词嵌入模型使用梅尔频率倒谱系数等标准声学特征进行训练，而自监督特征通常是在针对目标语言微调的有监督模型中进行评估。此外，有监督声学词嵌入的零样本跨语言迁移并非该领域先前研究的重点，这些研究主要关注如何改进自监督声学词嵌入。

为进行此次评估，我们为嵌入模型采用了相对简单的架构，并根据英语自监督声学词嵌入的初步验证结果固定了超参数[2]。我们不再对自监督模型或有监督模型进行进一步调优。我们使用英语作为源语言，并在另外四种语言上评估零样本迁移：法语、德语、西班牙语和阿拉伯语，其中阿拉伯语包含更多变异和噪声，因此被用作挑战集。除通过强制对齐获得的词边界外，目标语言均未使用标注数据。我们主要使用最小对立对 ABX 错误率来衡量语音辨别能力和说话人不变性。此外，我们还对嵌入后的词进行聚类，并测量同一词的不同出现形式最终被归入同一簇的频率。

4 实验设置

4.1 模型架构

我们的声学词嵌入模型由一个多层双向 LSTM 编码器和一个单向 LSTM 解码器组成，与 Chung 等人(2016) 以及 (Holzenberger等人，2018年) 的模型类似。编码器接收一个由 T 声学特征构成的序列，用以表示一个口语词。编码器最后一个隐藏层的前向和后向状态拼接起来，作为给定词的嵌入，记作 $\mathbf{h}^T$。解码器以 $\mathbf{h}^T$ 和前一时间步的输出为条件，每次生成一个时间步的目标序列，与 Chung 和

[2]我们观察到，自监督模型对架构和超参数的选择非常敏感，因此我们针对自监督模型固定了这些设置。正如后文各节所示，我们使用有监督模型仍然取得了更好的结果，这表明有监督模型更加稳健，也更易于优化，同时效果更佳。

Glass(2018年)。在自监督设置中，目标序列与输入序列相同，因此模型以均方误差损失训练为自编码器。在监督设置中，目标是表示输入词的一串音素，模型通过最小化负对数似然进行训练。大多数模型采用具有 100 个隐藏单元的 2 层网络，生成的嵌入大小为 200。我们还在输入特征上使用了概率为 0.3 的随机失活，与 Chung 等人(2016) 使用的去噪网络类似。有关参数和训练过程的更多细节，请参阅附录。

4.2 特征提取

为便于复现，我们使用了 s3prl 工具包[3]提取所有特征。我们使用了预训练的 s3prl 上游模型；在众多可用的预训练自监督语音表示中，根据基于 DTW 的 ABX 分数，选择了改进的 CPC、Wav2Vec2 和 XLSR-53[4]。除 XLSR 外，所有预训练模型都只使用英语数据进行预训练。XLSR-53 则使用来自 53 种语言的未标注语音进行预训练，其中包括我们实验中的所有目标语言。正如其他研究者（Bartelds 等人，2022）所观察到的，从基于 Transformer 的模型中提取的特征，其性能在很大程度上取决于所选的层；我们对改进的 CPC 使用最后一个隐藏层，对 Wav2Vec2 使用倒数第二个隐藏层，对 XLSR-53 使用中间隐藏层（第 12 层）。对所有层取平均也能取得不错的结果，但上述选择带来了最佳性能。对于 MFCC 特征，我们同样使用了 s3prl 的实现，其中包括 13 个静态特征，以及动态差分系数和二阶差分系数。

4.3 数据

我们使用了英语（en）、法语（fr）、德语（de）和西班牙语（es）的 Librispeech（Panayotov 等人，2015）和多语言 Librispeech（Pratap 等人，2020年）数据集。我们使用开发集进行训练，并使用测试集进行评估（英语使用 dev-clean 和 test-clean）。我们通过强制对齐自动获取词边界。对于阿拉伯语（ar），我们使用了 MGB2 的开发集和测试集（Ali 等人，2016）。该数据集预计更具挑战性，因为其中包含多种方言以及各种噪声条件。有关数据集和词对齐过程的更多详情，请参阅附录。

4.4 评估方案

我们按照 Schatz 等人 (2013) 中的描述构建了最小对立对 ABX 任务。ABX 任务通常用于衡量零资源设置下的音素辨别，由两个存在最小对立（例如仅一个音素不同）的片段 A 和 B，以及第三个与 A 或 B 匹配的片段 X 组成。我们使用动态时间规整或余弦距离等距离度量来寻找最接近的匹配。我们采用了此任务的两种变体：说话人内 ABX，即三个词均由同一说话人说出；以及跨说话人 ABX，即 X 由不同的说话人说出。我们从每个测试集中自动提取词语；通过寻找长度相同[5] 且莱文斯坦编辑距离为 1 或 2 的词对来选取 A 和 B，这大致对应于大多数情况下相差一个或两个音素。对于阿拉伯语，数据集没有说话人 ID，因此三个词可能来自不同的说话人。此外，由于该数据集中录音质量较低且存在噪声，其词对齐质量远低于其他语言，因此自动处理过程产生了许多无效片段。为获得更可靠的阿拉伯语测试集，我们人工检查了提取词语的有效性，并保留了 954 对经过验证的词对用于评估。

我们还使用聚类进行补充评估。我们使用 K-Means 对嵌入进行聚类，其中 K 为测试集中不同词语的数量。我们将聚类准确率计算为与其簇标签相匹配的词语所占的百分比；簇标签即每个簇中占多数的片段对应的词语 ID。这样我们便能衡量相同词语的嵌入是否足够相似，从而被聚到一起。

5 结果

表1 展示了直接使用输入特征（以动态时间规整作为距离度量）时的 ABX 错误率，以及在各语言上训练的自监督声学词嵌入，

[3]https://github.com/s3prl/s3prl[4]我们也使用了其他特征进行实验，例如 APC、VQ- APC 和 VQ-Wav2Vec，所得性能与 MCPC 和 Wav2Vec2 相近或更差。为简洁起见，我们省略了这些结果。

[5]由于自动词对齐在边界附近往往不准确，我们只使用至少包含五个字符的词语。

	en		fr		de		es		ar
	语言内	跨语言	语言内	跨语言	语言内	跨语言	在.....之	跨越	跨越
使用动态时间规整									
MFCC	9.98	19.85	12.59	24.82	11.46	25.03	11.83	25.27	40.98
wav2vec	8.51	11.15	9.95	15.61	9.01	15.08	10.13	15.46	37.42
MCPC	7.80	11.74	9.96	17.09	9.22	15.85	11.14	18.03	38.99
XLSR-53	9.45	13.72	10.97	16.77	10.89	15.76	14.31	19.31	40.15
各语言的自监督声学词嵌入									
MFCC	12.30	19.12	16.63	25.06	16.71	25.99	16.58	25.21	43.92
wav2vec	6.63	9.27	9.94	13.19	10.56	15.09	12.25	15.23	38.16
MCPC	7.66	9.53	11.64	16.19	10.24	15.30	13.25	16.29	41.09
XLSR-53	10.61	12.19	13.72	16.19	12.59	15.73	17.10	20.14	37.00
基于英语训练的有监督声学词嵌入									
MFCC	3.83	4.57	10.77	15.32	9.16	13.06	11.49	16.56	38.15
wav2vec	1.38	1.14	6.59	9.44	4.98	7.32	7.32	10.12	34.80
MCPC	2.49	2.51	8.13	12.23	6.66	10.68	9.89	13.99	39.20
XLSR-53	**0.93**	**0.79**	**4.12**	**5.71**	**1.92**	**2.83**	**5.05**	**6.21**	**31.76**

表 1：各语言的说话人内及跨说话人 ABX 错误率（%）

	en	fr	de	es	ar
各语言的自监督声学词嵌入					
MFCC	47.3	48.4	45.0	54.3	30.9
wav2vec	66.1	59.9	59.5	67.5	35.9
MCPC	57.2	53.6	53.9	60.4	33.5
XLSR-53	52.0	52.2	54.9	55.0	31.0
在英语上训练的有监督声学词嵌入					
MFCC	68.5	52.7	56.7	61.1	33.4
wav2vec	82.3	64.8	69.3	71.1	38.8
MCPC	74.5	56.4	62.7	66.2	35.6
XLSR-53	**84.3**	**69.1**	**78.1**	**75.8**	**41.8**

表 2：K 均值聚类准确率（%）

以及在英语上训练的有监督声学词嵌入。后两种设置使用余弦相似度作为度量。与先前的结果（Riviere等人，2020年）一致，我们确实观察到，改进的CPC和Wav2Vec2等预训练声学特征尽管仅使用英语未标注语音进行训练，仍能很好地迁移到其他语言。这些预训练特征在所有语言中始终优于 MFCC 特征，尤其是在跨说话人评估中。不出所料，英语的 ABX 得分总体最佳，这仅仅是因为所用的预训练特征均在英语上训练。自监督声学词嵌入模型的结果不一，但总体上与动态时间规整的表现处于同一水平，这也与此前发表的结果相符（Holzenberger等人，2018年）。

通过有监督训练，我们看到所有语言的错误率都显著降低。正如预期，英语测试集的错误率最低。更值得注意的是，使用 XLSR 特征时错误率降幅最大。有趣的是，与其他特征相比，XLSR 特征在自监督设置中的表现并不突出；仅在英语上训练的 Wav2Vec2 和 MCPC，在自监督框架下对所有测试语言都取得了更好的结果。使用这些跨语言特征的优势仅在监督和迁移学习设置下显现，它们始终优于其他所有特征。对于阿拉伯语，由于数据集的特性，其整体错误率较高，但我们仍观察到迁移学习设置下的错误率最低。

最后，我们从表格2 中可以看到，聚类准确率结果与 ABX 结果一致：在英语上训练的有监督模型，其准确率始终高于在目标语言上训练的自监督模型。

6 结论

我们的结果表明，与在目标语言中进行自监督训练相比，声学词嵌入的零样本迁移学习效果更佳。这对于低资源语言尤其有用，因为这类语言可能没有可用于有监督或自监督

训练的数据。这种迁移的机制主要是减少说话人差异，而通过有监督训练实现这一点要容易得多。此外，有监督训练能充分利用预训练特征，使错误率进一步降低，且降幅远高于自监督设置下观察到的降幅。噪声的存在自然会导致错误率上升；还需要进一步研究，以类似方式证明噪声鲁棒性的可迁移性。

致谢

本研究得到了阿联酋大学31T139号资助，并部分由阿联酋移动性研究中心通过UAEU-ZU联合研究基金G00003715（基金编号：12T034）提供资助。

参考文献

Ossama Abdel-Hamid、Li Deng、Dong Yu 和 Hui Jiang。2013。用于语音识别的深度分段神经网络。

Hanan Aldarmaki、Asad Ullah、Sreepratha Ram 和 Nazar Zaki。2022。无监督自动语音识别：综述。语音通信。

Ahmed Ali、Peter Bell、James Glass、Yacine Messaoui、Hamdy Mubarak、Steve Renals 和 Yifan Zhang。2016。mgb-2挑战赛：阿拉伯语多方言广播媒体识别。载于*2016 IEEE*口语语言技术研讨会（*SLT*），第279–284页。IEEE。

Alexei Baevski、Wei-Ning Hsu、Alexis Conneau 和 Michael Auli。2021年。无监督语音识别。*arXiv*预印本 *arXiv:2105.11084*。

Alexei Baevski、Henry Zhou、Abdel rahman Mohamed 和 Michael Auli。2020a。wav2vec 2.0：用于语音表征自监督学习的框架。*ArXiv*，abs/2006.11477。

Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020b。wav2vec 2.0：用于语音表征自监督学习的框架。神经信息处理系统进展，33：12449–12460。

Martijn Bartelds、Wietse de Vries、Faraz Sanal、Caitlin Richter、Mark Liberman 和 Martijn Wieling。2022。用于语音变异建模的神经表征。*JournalofPhonetics*，92：101137。

Mathieu Bernard 和 Hadrien Titeux。2021年。Phonem-izer：使用 Python 将文本转写为多种语言的音素。《开放源代码软件期刊》，6(68)：3958。

Yu-An Chung 和 James Glass。2018年。Speech2vec：一种用于从语音中学习词嵌入的序列到序列框架。*Interspeech 2018*，第 811–815 页。Yu-An Chung、Chao-Chung Wu、Chia-Hao Shen、Hung-Yi Lee 和 Lin-Shan Lee。2016。Audio word2vec：使用序列到序列自动编码器进行音频片段表征的无监督学习。*Interspeech 2016*，第 765–769 页。

Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 MichaelAuli。2020年。用于语音识别的无监督跨语言表征学习。*arXiv*预印本 *arXiv:2006.13979*。

Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu 等。2020年。Conformer：卷积增强的 Transformer 语音识别模型。*Interspeech 2020* 会议论文集，第 5036–5040 页。

Yanzhang He 和 Eric Fosler-Lussier。2015年。将分段条件随机场与深度神经网络声学模型结合，用于首次通过词语识别。在 国际语音通信协会第十六届年度会议。

Nils Holzenberger、Mingxing Du、Julien Karadayi、Rachid Riad 和 Emmanuel Dupoux。2018年。学习词嵌入：用于变长语音片段的定长表征的无监督方法。*Interspeech 2018* 会议论文集，第 2683–2687 页。

Christiaan Jacobs、Yevgen Matusevych 和 Herman Kamper。2021年。利用自监督对比学习和多语言适应，为零资源语言构建声学词嵌入。在 *2021 IEEE*口语语言技术研讨会 *(SLT)*，第 919–926 页。IEEE。

Herman Kamper、Yevgen Matusevych 和 Sharon Goldwater。2020年。用于处理零资源语言的多语言声学词嵌入模型。在*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第 6414–6418 页。IEEE。

KazuyaKawakami、Luyu Wang、Chris Dyer、Phil Blunsom 和 Aaron van den Oord。2020年。学习稳健的多语言语音表示。在计算语言学协会论文集：*EMNLP 2020*，第 1182–1192 页。

Keith Levin、KatharineHenry、Aren Jansen 和 Karen Livescu。2013年。低资源环境下变长片段的固定维度声学嵌入。载于*2013 IEEE*自动语音识别与理解研讨会，第410–415页。IEEE。

Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。2017年。蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本-语音对齐。*Interspeech 2017*论文集，第498–502页。

Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。2015年。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。载于*2015 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第5206–5210 页。IEEE。

Puyuan Peng、Herman Kamper 和 Karen Livescu。2020年。一种用于无监督声学词嵌入的对应变分自编码器。神经信息处理系统进展。

Vineel Pratap、Qiantong Xu、Anuroop Sriram、Gabriel Synnaeve 和 Ronan Collobert。2020年。MLS：用于语音研究的大规模多语言数据集。

Morgane Riviere、Armand Joulin、Pierre-Emmanuel Mazaré 和 Emmanuel Dupoux。2020年。无监督预训练能够很好地迁移到不同语言。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第7414–7418页。IEEE。

Thomas Schatz、Vijayaditya Peddinti、Francis Bach、Aren Jansen、Hynek Hermansky 和 Emmanuel Dupoux。2013年。使用最小对ABX任务评估语音特征：经典MFC/PLP流程分析。载于*INTERSPEECH 2013*：国际语音通信协会第*14*届年度会议，第1–5页。

Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。2018年。利用对比预测编码进行表征学习。*arXiv*电子预印本，第arXiv–1807页。

Lisa van Staden 和 Herman Kamper。2021年。自监督语音表征作为无监督声学词嵌入输入特征的比较。载于*2021*年 *IEEE*口语语言技术研讨会（*SLT*），第927–934页。IEEE。

Shu-wen Yang、Po-Han Chi、Yung-Sung Chuang、Cheng-I Jeff Lai、Kushal Lakhotia、Yist Y Lin、Andy T Liu、Jiatong Shi、Xuankai Chang、Guan-Ting Lin 等。2021年。SUPERB：语音处理通用性能基准。*arXiv*预印本 *arXiv:2105.01051*。

A 附录

A.1 数据集详情

A.2 模型架构与超参数

第4.1 节所述的架构模仿了其他声学词嵌入模型（Chung 等人，2016；Chung和Glass，2018年；Holzenberger等人，2018），并作了少许调整。我们发现，这种特定配置在不同声学特征上表现最佳，而其他选择则结果参差不齐。例如，与长短期记忆网络相比，使用门控循环单元配合预训练特征时效果不错，但用于梅尔频率倒谱系数时表现较差。采用位置编码而非先前输出的解码过程（Holzenberger等人（2018年））也导致性能较差。我们还发现，将教师强制作为解码器的输入，而不是使用模型先前的输出，会损害性能。最后，对于大多数自监督模型而言，使用两层网络对于取得与动态时间规整相当的结果至关重要。唯一的例外是采用 XLSR 特征的自监督模型：使用两层网络时，训练不稳定。我们发现，使用单层网络并略微增大嵌入维度，效果要好得多。总体而言，增大嵌入维度在一定程度上能提升性能，但超过我们选定的数值后，提升幅度会变小；此外，较小的维度在计算效率方面更有优势。由于我们的研究基于低资源场景这一前提（在这种场景下可能没有验证数据），因此未针对目标语言进行任何超参数调优。

数据集	test	dev
英语	52,576	54,402
法语	90,958	83,560
德语	121,713	122,903
西班牙语	88,417	87,417
阿拉伯语	62,745	57,532

表 3：每个数据集中的单词总数

表 4 展示了每个模型的参数数量。由于解码器仅用于训练，之后可以将其丢弃，因此我们只展示编码器参数的数量。

A.3 训练细节

有监督模型采用 NLL 损失进行训练，训练目标是使用 Phonemizer 软件包生成的音素序列 [6] (Bernard 和 Titeux，2021年）。这一选择起初似乎更合理，但我们发现，使用字符序列代替音素序列效果同样好。

该模型使用 PyTorch 实现，并

[6] https://github.com/bootphon/phonemizer

模型	输入	隐藏层	参数数量
自监督			
MFCC	39	100	354,400
wav2vec	768	100	937,600
MCPC	256	100	528,000
XLSR-53	1024	250	1,411,200
有监督			
MFCC	39	100	354,400
Wav2Vec	768	100	937,600
MCPC	256	100	528,000
XLSR-53	1024	100	1,142,400

表 4：各模型的输入大小、隐藏层大小及编码器参数总数。

在 AWS p2.xlarge 实例提供的 NVIDIA K80 GPU 上进行训练。对于优化，我们发现 Adam 优化器适用于除梅尔频率倒谱系数之外的所有特征；对于后者，采用循环或阶梯式学习率调度的随机梯度下降更加稳定。

表 3 显示了各数据集中词语的数量。词对齐是使用 The Montreal Forced Aligner 进行强制对齐获得的[7] （McAuliffe 等人，2017年），适用于英语、德语、法语和西班牙语。蒙特利尔对齐器使用自动语音识别引擎；由于这些数据集相对干净，对齐结果通常较为准确。对于阿拉伯语，最佳选择是 aeneas 工具包[8]，，它依靠文本转语音引擎将合成语词与实际音频片段对齐。我们使用 Amazon Polly 文本转语音以获得更高质量，但总体而言，对齐结果不如其他数据集准确。我们认为这是由于录音质量较低、存在噪声以及口音差异较大所致。片段标注不准确可能是聚类准确率较低的部分原因。针对阿拉伯语数据集进行 ABX 评估时，我们手动筛选了边界相对准确的片段；所选片段对中仍包含噪声程度较高的条件，例如背景音乐和干扰语音。

[7] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner [8] www.readbeyond.it/aeneas