arXiv:2304.00649v1 [cs.CL] 2 Apr 2023

多语言词错误率估计：E-WER3

Shammur Absar Chowdhury，Ahmed Ali

卡塔尔计算研究所，哈马德·本·哈利法大学，多哈，卡塔尔

摘要

多语言自动语音识别系统的成功推动了许多语音驱动应用的发展。然而，由于这类系统在单语和多语场景下都依赖人工转录语音数据，衡量其性能仍是一项重大挑战。本文提出了一种新颖的多语言框架——eWER3，该框架联合利用声学和词汇表示进行训练，以估计词错误率。我们证明，eWER3 能够 *(i)* 在不使用自动语音识别系统任何内部状态的情况下预测词错误率，并*(ii)* 利用多语言共享潜在空间提升关系密切的语言的性能。结果表明，与以往的单语言词错误率估计方法（eWER2）相比，我们提出的多语模型使皮尔逊相关系数（PCC）绝对提高了 9%，且对预测词错误率与参考词错误率的整体估计效果更佳。

索引词— 多语言词错误率估计，端到端系统

1. 引言

近年来，单语和多语语音识别的性能显著提升，在成熟基准测试上的准确率已达到甚至超过人类水平 [1, 2]。随着这类技术取得成功，自动语音识别（ASR）系统已成为许多语音处理流程中的标准组件，应用于个人助理设备、广播媒体监测等语音驱动应用。然而，我们评估自动语音识别输出是否有用的方法基本上仍未改变。

词错误率（WER）是评估自动语音识别系统性能的标准指标。要可靠地估计词错误率，通常至少需要两小时的人工转录测试数据，这一过程既耗时又昂贵。语音驱动应用通常需要快速评估自动转录文本的质量，而传统的基于参考文本的指标无法满足这一需求。此外，即使对于离线应用，获取标准参考文本（尤其是在多语言场景下）来评估转写质量也并非总是可行。因此，我们需要开发无需人工付出、能够自动估计自动语音识别转录文本质量并处理多语言问题的技术 [3, 4]。

多项研究探索了词错误率的自动估计。这些研究使用大量提取的特征（可访问或不可访问自动语音识别系统的内部信息）来训练神经网络回归或分类模型 [5, 6, 7]。一些研究提出了一种新颖的神经网络零膨胀模型 [8], , 另一些则对预测中的不确定性进行建模[9], 以应对不同的挑战。然而，所有这些研究都是在单语环境中直接训练和测试网络。

在这项工作中，我们设计了一个单一的多语言端到端模型，只需输入原始音频和自动转录，即可估计来自不同（单语和多语）现成自动语音识别系统的词错误率，而无需访问自动语音识别系统的内部特征表示（这一概念见图 1）。为此，我们采用大型自监督预训练模型作为特征提取器，并利用现有的多语言语料库。

我们使用阿拉伯语、意大利语、西班牙语、英语和俄语 测试集评估结果。我们训练一个单语估计器，并将其与提出的多语模型进行比较，以展示后者在提升性能方面的有效性。我们的贡献如下：

- 设计首个不使用自动语音识别系统任何内部特征的多语言词错误率估计方法（黑盒）；
- 将我们的方法与以往的最先进结果进行比较（e-wer [6] 和 e-wer2 [7]）；
- 分析词错误率分布不均衡对估计器性能的影响，并提出一种新的采样技术。

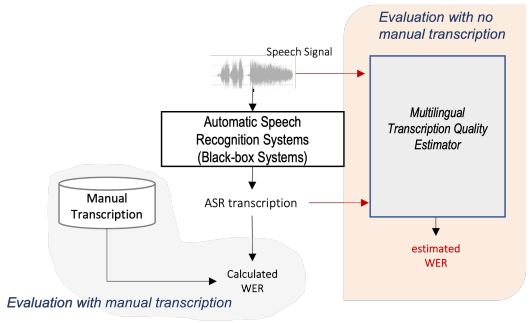


图 1：研究概念与所提框架概述。

2. 端到端多语词错误率估计器

图 2 展示了端到端系统架构概述，该架构旨在无需金标准参考转录即可估计语音识别词错误率。作为估计器的输入，我们首先输入原始音频及其从语音识别系统获得的自动转录。我们提取语音和词汇表示，并联合利用这些表示来训练多语回归模型。

声学表示： 我们使用 XLSR-53 提取具备音素感知能力的语音表示。XLSR-53 模型是 wav2vec 2.0 模型的一种多语变体，并经过跨语言音素识别任务微调 [14, 15]。在本研究中，我们移除输出层（语言模型头），仅使用其表示。我们将 XLSR-53 用作特征提取器，其中包含一个级联时序卷积网络，用于将原始音频 $X = \{x_1, x_2.., x_n\}$ 映射到潜在语音表征 $Z = \{z_1, z_2.., z_t\}$。这一潜在

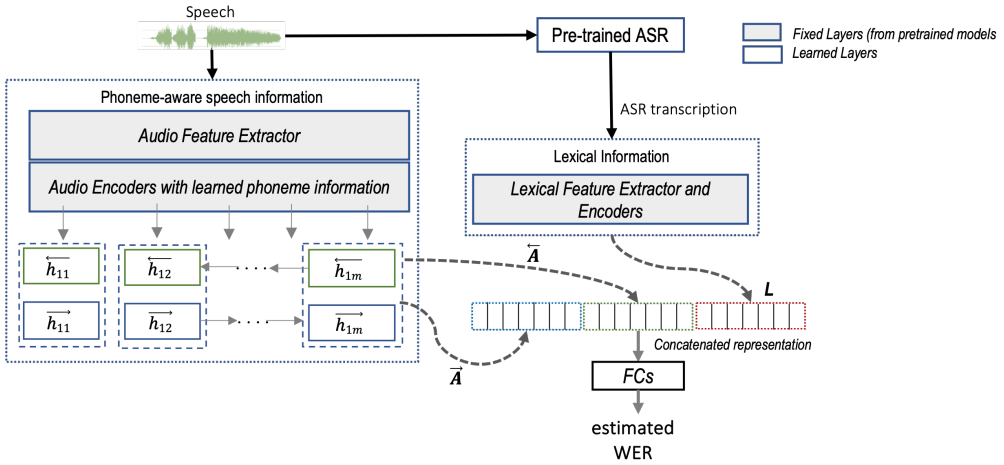


图 2：用于估计词错误率的端到端架构。A^*：来自双向长短期记忆网络的声学表示，L：词汇表示。

Lang	自动语音识别训练所用数据	架构	自动语音识别类型	估计器训练所用数据	估计器测试所用数据
英语	LibriSpeech（960小时）	Conformer	Mono	LibriSpeech +TEDLIUM3 开发集	TEDLIUM3 测试集
西班牙语	CommonVoice（CV-ES） [10]	Conformer	Mono	CV-ES 开发集	CV-ES 测试集
意大利语	CommonVoice（CV-IT）	Conformer	Mono	CV-IT 开发集	CV-IT 测试集
俄语	CommonVoice (CV-RU)	Conformer	Mono	CV-RU 开发集	CV-RU 测试集
阿拉伯语	QASR（阿拉伯语）+ LibriSpeech（英语）（200小时） [11]	Conformer	多语	MGB2 [12] +QASR 开发集	SUMMA [13]

表1：用于训练和/或测试所提出的估计器的自动语音识别系统，以及估计器的训练集和测试集的说明。*Lang.* 表示用于训练估计器的语言数据。

训练集（开发集 ǀ 测试集）	**T.D**（小时）	**A.D**（秒）	**A.U**（单词）	**#**
阿拉伯语	10.36（0.56 ǀ 2.94）	6.14（6.08 ǀ 7.52）	12.66（12.45 ǀ 13.96）	6077 (332 ǀ 1410)
英语	3.72 (0.31 ǀ 2.62)	5.83 (5.65 ǀ 8.18)	16.68 (16.04 ǀ 23.95)	2295 (196 ǀ 1151)
西班牙语	7.37 (0.84 ǀ 17.31)	6.05 (6.11 ǀ 6.12)	10.11 (10.17 ǀ 9.84)	4384 (496 ǀ 10179)
意大利语	4.94（0.53 ǀ 12.26）	5.94 (5.91 ǀ 6.13)	10.2 (10.13 ǀ 9.74)	2991 (322 ǀ 7200)
俄语	1.73 (0.17 ǀ 9.57)	5.94（5.83 ǀ 5.99）	9.86（9.72 ǀ 9.58）	1045（102 ǀ 5748）

表 2：训练集、开发集和测试集（开发集和测试集在括号内，以 ǀ 分隔）数据描述。*T.D*：数据集总时长，单位为小时；*A.D*：平均话语时长，单位为秒；*A.U*：平均话语长度，单位为词元；*#*：实例总数。

信息随后经过 24 个 Transformer [16] 块，模型维度为 1,024，包含 16 个注意力头，以捕获上下文表示，C（$g: Z \mapsto C$）。随后，我们将逐帧表示输入双向长短期记忆网络，并提取最后一步的表示（$\overleftarrow{A}, \overrightarrow{A}$）。

词汇表示： 同时，为了提取词汇嵌入，我们将自动语音识别转录文本输入 XLM-RoBERTa-Large 模型 [17], 。该模型使用 100 种不同语言进行预训练。预训练模型遵循与 BERT [16]，相同的架构，包含 24 层 Transformer模块、1,024 个隐藏状态和 16 个注意力头。该模型使用字节级BPE作为分词器，并输出整个输入句子对应的隐藏状态序列。为了获得最终的词汇表示（L），我们对各序列中的嵌入进行平均。

表示融合： 我们将声学模块和词汇模块的输出表示（$\overleftarrow{A} + \overrightarrow{A} + L$）拼接起来，并将其依次输入两个全连接层，再送入输出层，以完成回归任务。

3. 实验设置

3.1. 语音识别系统

为了训练估计器，我们选择了当前最先进的、基于 Conformer [18] 的端到端语音识别系统（见表1）。

对于西班牙语、意大利语和俄语自动语音识别系统，模型分别使用各自的 Common Voice 训练集进行训练。该模型包含 12 个编码器层和 6 个解码器层，每层均包含 2，048 个来自前馈网络的编码器/解码器单元、4 个注意力头，变换维度为 256，并使用 15 个卷积核。

对于英语自动语音识别系统，我们使用一个大型 Conformer 模型，其中包含 12 个编码器层和 6 个解码器层，具有 8 个注意力头、512 个变换维度和 31 个卷积核。该大型自动语音识别系统使用广为人知的 960 小时 LibriSpeech 数据进行训练。对于多语种阿拉伯语自动语音识别，我们采用类似的架构，[19] 使用阿拉伯语 QASR [11] 以及英语 LibriSpeech 数据进行训练。

3.2. 数据

数据准备： 我们使用表1中提到的数据集训练多语言估计器。首先将输入估计器的音频下采样至16千赫，以匹配自动语音识别的输入采样率。对于

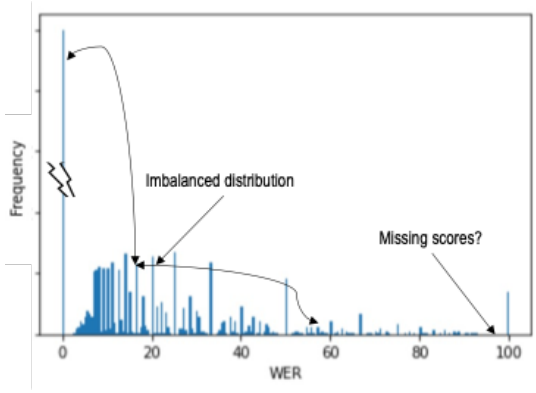


图 3：数据不平衡和缺失的目标值。

模型训练，我们选择时长不超过10秒的音频样本；这是根据整体时长分布的上尾部分确定的。

分布不平衡：鉴于当前端到端自动语音识别模型表现出色，词错误率往往呈现不平衡分布，即某些目标值的观测数量显著少于其他目标值。在这种情况下，训练集中的大多数词错误率都为“0”，使训练数据高度偏斜（见图 3）。此外，对于某些目标值，数据集中还存在缺失数据的倾向，因此任务更具挑战性。为处理训练集中大量的“0”词错误率分数，我们从每种语言中抽取了 n 个实例，且 $WER = 0$。我们根据落入接下来两个最常见分数组的实例总数来确定该值 n 。

数据划分：对于开发集，我们将训练数据集按目标词错误率划分为 10 个等距分箱，例如 $([0, 10), [10, 20) \cdots [90, 100])$。随后，从每个分箱中随机抽取 % 的实例，以创建验证集。最终（平衡后的）划分详情见表 2。 ≈ 10

估计器输出： 作为估计器的输出分数，需要注意的是，我们将估计器的目标值（WER）限制在 [0,1]（即 0 - 100%）的范围内。[1]

3.3. 词错误率估计器设计

模型参数：我们使用 Adam 优化器对端到端词错误率估计器进行 20 个训练轮次的训练，学习率为 $1e-3$，随机失活率为 0.1，并冻结预训练自监督模型的参数。在声学流水线中，我们使用一层双向长短期记忆网络模型；对于联合训练，我们选择两个全连接层（600、32 个神经元），并使用 ReLU 激活函数。至于损失函数，我们使用均方误差。训练单语和多语模型时，采用相同的架构和超参数，并使用平衡数据和自然分布数据。

3.4. 评估指标

鉴于分数分布不均（偏向较小的 WER 值），我们使用皮尔逊相关系数（PCC）作为主要评估指标。不过，我们也报告均方根误差（RMSE），以便与以往研究进行比较 [6, 7]。此外，为了有效估计完整测试集的 eWER3，我们报告加权词错误率：($eWER3 = \frac{\sum \widehat{WER_{utt}} * Dur_{(utt)}}{\sum^{n} Dur}$)，~~该指标使用话语级~~估计的 WER（$\widehat{WER_{utt}}$）及其对应的时长（$Dur$（$utt$））计算。

[1]如果 WER $> 100\%$，则将该值缩放为 100。

4. 结果与讨论

4.1. 单语比较

我们在单语（阿拉伯语）设置下对所提出的框架 eWER3 进行基准测试，并将其与此前的估计模型——eWER 和 eWER2 进行比较。表3所示结果表明，我们的模型优于 eWER 和 eWER2，其皮尔逊相关系数分别绝对提高了 3% 和 9%，均方根误差分别降低了 21% 和 6%。这一改进表明，即使不使用自动语音识别解码器提供的任何额外特征，我们的架构仍具有强大的估计能力。

此外，将单语模型（阿拉伯语和英语模型，见表4）应用于跨语言意大利语数据集进行测试时，两个模型的性能（无论是相关系数还是均方根误差）都大幅下降。然而，意大利语测试集从英语单语模型中获益更多：该模型的均方根误差为 0.19，而阿拉伯语模型的均方根误差为 0.32。这表明，在多语设置下，近距语言共享潜在空间可能具有优势。

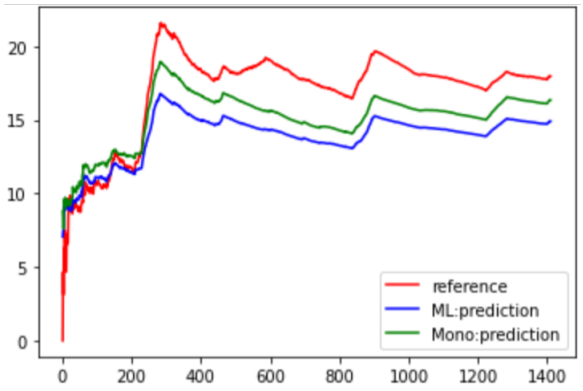


图4：阿拉伯语*SUMMA*测试集 *(a)* 中所有句子的累计词错误率随时间的变化。横轴为实例数，纵轴为聚合词错误率（%）。

4.2. 多语端到端估计器性能

表4显示，与单语言模型相比，多语模型的相关性和均方根误差相当。我们注意到，在多语设置下，英语测试集的皮尔逊相关系数提升了4%，这体现了此类多语估计器的额外优势。此外，对于所有语言测试集（阿拉伯语、英语、意大利语、西班牙语、俄语），除了逐话语层面的均方根误差更小且相关性显著之外，总体预测词错误率也与实际词错误率相差不大（$0-5$ 个百分点）。

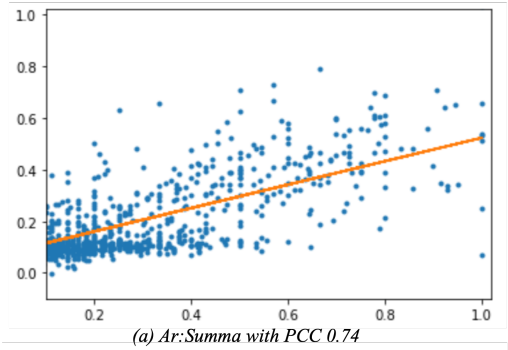
为简洁起见，我们在图4中展示了按句子汇总的阿拉伯语SUMMA测试集累积词错误率，并在图5中展示了阿拉伯语（获得的皮尔逊相关系数最高）和俄语（表现最差的皮尔逊相关系数）测试集对应的散点图。

Ar:SUMMA	**PCC**	**RMSE**	**输入至估计器**
eWER	0.66	0.35	词汇 + 字素 +解码器 + 数值 [6]
eWER2	0.72	0.20	MFCC+词汇 +语音 [7]
扩展词错误率（单声道）	**0.75**	**0.14**	原始音频，词汇转写

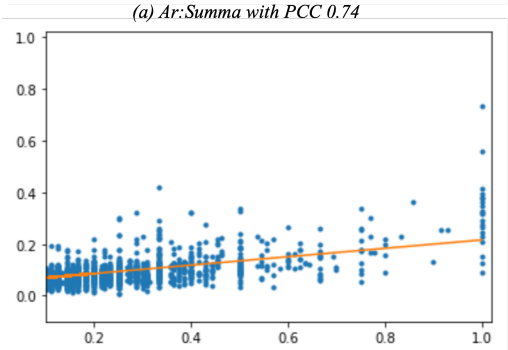
表3： *Ar:SUMMA* 测试集上的单语（阿拉伯语）转写质量估计器结果。

Sets	PCC	RMSE	eWER3	WER	#
单语估计模型 - 阿拉伯语					
Ar:SUMMA	0.75	0.14	16.0%	18.0%	1410
It:CV	0.45	0.32	41.0%	17.0%	7200
单语估计模型 - 英语					
En:TedL	0.62	0.14	7.0%	12.0%	1151
It:CV	0.49	0.19	10.0%	17.0%	7200
多语言估算器模型					
Ar:SUMMA	0.74	0.15	15.0%	18.0%	1410
It:CV	0.60	0.17	14.0%	17.0%	7200
西班牙语：CV	0.53	0.14	13.0%	11.0%	10179
En:TedL	0.66	0.14	8.0%	12.0%	1151
Ru:CV	0.51	0.12	6.0%	7.0%	5748

表4： 阿拉伯语（阿拉伯语）、英语（英语）、意大利语（意大利语）、西班牙语（西班牙语）和俄语（俄语）测试集上的单语和多语言词错误率估计器性能。



(a) Ar:Summa with PCC 0.74



(b) CV:RU with PCC 0.51

图5： 测试集散点图：皮尔逊相关系数最高为 *0.74*（阿拉伯语 *(a)*），最低为 *0.51*（俄语）。

4.3. 数据不平衡与采样的影响

我们分析了使用采样数据训练模型（采样模型： ψ）而非自然分布（自然模型：φ）的影响。就 ψ而言，我们注意到 φ 的相关系数略高——数值虽然较高，但均方根误差值更大，且聚合估计 eWER3 与 Oracle词错误率之间的差距也更大。例如，对于 ES:CV 测试集，$\varphi(PCC) = 0.58, \varphi(RMSE) = 0.15, \varphi(eWER3) = 7.0\%$，而 $\psi(PCC) = 0.53, \psi(RMSE) = 0.14, \psi(eWER3)$ 13.0 [2] %。

根据 φ 和 ψ 模型预测得到的密度曲线（图6）表明，在自然分布下，模型（φ）更擅长学习预测较低的词错误率，而非 ψ。不过，预测值被压缩到更低的范围（见两条曲线峰值的偏移），从而增大了均方根误差，以及整体预测 eWER3 与 Oracle词错误率之间的差距。这是本研究当前的一项潜在局限；未来可进一步尝试在此类多语言网络中使用零膨胀输出层 [20]。

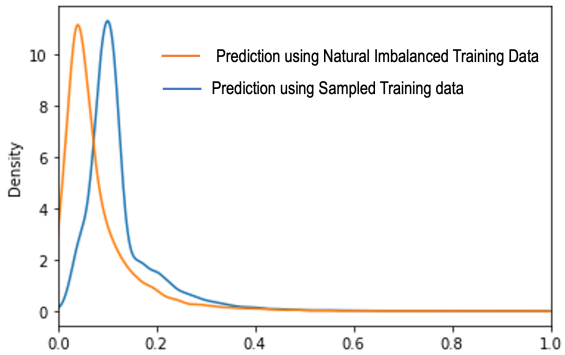


图 6： 采用采样分布（蓝线）和自然分布（橙线）训练集训练的多语模型，其估计词错误率的密度曲线，展示了不平衡数据标签的影响。*x* 轴表示词错误率。预测结果来自聚合的语内测试集。

5. 结论

在本研究中，我们提出了一种新颖的框架，无需人工转录即可估计多语言词错误率。我们提出的框架是一种联合声学词法模型，利用自监督学习范式。基于一小部分语言的实验结果表明，该模型能够有效预测测试集的话语级和总体词错误率。与单语模型相比，该多语言框架在远距语言（例如阿拉伯语）上的表现相当，同时提升了近距语言的表现（例如， En_{mono}: 0.62 皮尔逊相关系数 与 En_{multi}: 0.66 皮尔逊相关系数）。本研究可作为概念验证，说明如何利用表征模型为自动语音识别系统设计此类预测器。为确保计算可行性，我们将预训练模型用作特征提取器。未来，我们将致力于通过处理不平衡目标分布、改进神经网络架构并覆盖更多语言来提升性能。

[2]使用自然分布时，其他数据集的均方根误差和总体词错误率差异更大。

6. 参考文献

[1] Wayne Xiong、Jasha Droppo、Xuedong Huang、Frank Seide、Mike Seltzer、Andreas Stolcke、Dong Yu 和 Geoffrey Zweig，“实现对话语音识别中的人类水平表现”，*arXiv* 预印本 *arXiv:1610.05256*，2016年。[2] George Saon、Gakuto Kurata、Tom Sercu、Kartik Audhkhasi、Samuel Thomas、Dimitrios Dimitriadis、Xiaodong Cui、Bhu- vana Ramabhadran、Michael Picheny、Lynn-Li Lim、Bergul Roomi 和 Phil Hall，“人类与机器的英语对话电话语音识别”，*arXiv* 预印本 *arXiv:1703.02136*，2017年。[3] Matteo Negri、Marco Turchi、José GC de Souza 和 Daniele Falavigna，“自动语音识别的质量估计”，载于 *COLING 2014* 论文集：第*25*届国际计算语言学会议技术论文，2014年，第1813–1823页。[4] Julia Kiseleva、Kyle Williams、Jiepu Jiang、Ahmed Has- san Awadallah、Aidan C Crook、Imed Zitouni 和 Tasos Anas- tasakos，“理解用户对智能助手的满意度”，载于*2016*年 *ACM* 人类信息交互与检索会议论文集，2016年，第121–130页。[5] Shahab Jalalvand、Matteo Negri、Marco Turchi、José GC de Souza、Daniele Falavigna 和 Mohammed RH Qwaider，“Transcrater：一种用于自动语音识别质量估计的工具”，*AC*，2016年。[6] Ahmed Ali 和 Steve Renals，“语音识别的词错误率估计：e-WER”，载于*ACL*，2018年。[7] Ahmed Ali 和 Steve Renals，“无需自动语音识别输出的词错误率估计：e-wer2”，*Interspeech 2020* 论文集，第616–620页，2020年。[8] Kai Fan、Jiayi Wang、Bo Li、Boxing Chen 和 Niyu Ge，“用于自动语音识别系统的神经零膨胀质量估计模型”，*arXiv* 预印本 *arXiv:1910.01289*，2019年。[9] Apoorv Vyas、Pranay Dighe、Sibo Tong 和 Hervé Bourlard，“使用随机失活分析语音识别中的不确定性”，载于*ICASSP*，2019年。

[10] Rosana Ardila、Megan Branson、Kelly Davis、Michael Henretty、Michael Kohler、Josh Meyer、Reuben Morais、Lindsay Saunders、Francis M Tyers 和 Gregor Weber，《Common Voice：大规模多语语音语料库》，*arXiv* 预印本 *arXiv:1912.06670*，2019年。

[11] Hamdy Mubarak、Amir Hussein、Shammur Absar Chowdhury 和 Ahmed Ali，《QASR：QCRI 半岛电视台语音资源——大规模标注阿拉伯语语音语料库》，载于 *ACL/IJCNLP*（*1*），2021年。

[12] Sameer Khurana 和 Ahmed Ali，“用于阿拉伯语多方言广播媒体识别的 QCRI 高级转录系统（QATS）：MGB-2 挑战赛”，发表于*SLT*，2016年。[13] Walid Magdy、Ahmed Ali 和 Kareem Darwish，“面向时效性社交媒体的摘要工具”，发表于第*21*届 *ACM* 信息与知识管理国际会议论文集。ACM，2012年，第2695–2697页。[14] Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli，“用于语音识别的无监督跨语言表征学习”，*arXiv* 预印本 *arXiv:2006.13979*，2020年。

[15] Qiantong Xu、Alexei Baevski 和 Michael Auli，“简单有效的零样本跨语言音素识别”，*arXiv* 预印本 *arXiv:2109.11680*，2021年。[16] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，“BERT：用于语言理解的深度双向 Transformer 预训练”，*arXiv* 预印本 *arXiv:1810.04805*，2018年。[17] Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov，“大规模无监督跨语言表征学习”，*arXiv* 预印本 *arXiv:1911.02116*，2019年。[18] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu 等，“Conformer：用于语音识别的卷积增强型 Transformer”，*arXiv* 预印本 *arXiv:2005.08100*，2020年。[19] Shammur Absar Chowdhury、Amir Hussein、Ahmed Abdelali 和 Ahmed Ali，“迈向一统模型：阿拉伯语方言语码转换的多语言策略”，*Interspeech*，2021年。[20] Kai Fan、Bo Li、Jiayi Wang、Shiliang Zhang、Boxing Chen、Niyu Ge 和 Zhijie Yan，“用于自动语音识别系统的神经网络零膨胀质量估计模型”，发表于*Interspeech*，2020年。