计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900013

多语言语音识别声学模型建模方法最新进展

程高峰[1] 颜永红[1,2]
1 中国科学院声学研究所 北京 100190
2 中国科学院大学电子电气与通信工程学院 北京 100049
(chenggaofeng@hccl.ioa.ac.cn)

摘 要 随着多媒体信息和通信技术的快速发展,网络上的多语言语音数据日益增多。语音识别作为语音分析与处理的核心技术,如何快速地把中文和英文等少数多资源主要语言处理能力推广到更多的低资源语言,是当前识别技术迫切需要突破的瓶颈。文中试图总结声学模型建模领域的最新进展,探讨传统语音识别技术从单语言向多语言跨越过程中可能面临的困难。并在此基础之上,探索了最新的端到端语音识别技术在关键词检索系统构建上的作用,以进一步改善系统的整体效果。最后总结了如下最新研究进展:1)基于模型参数共享的多语言声学建模;2)基于语种分类信息的多语言声学建模;3)基于帧级别对齐的端到端关键词检索技术。

关键词:多语言;语音识别;声学模型
中图法分类号 TP391

Latest Development of Multilingual Speech Recognition Acoustic Model Modeling Methods

CHENG Gao-feng[1] and YAN Yong-hong[1,2]
1 Institute of Acoustics,Chinese Academy of Sciences,Beijing 100190,China
2 School of Electronic,Electrical and Communication Engineering,University of Chinese Academy of Sciences,Beijing 100049,China

Abstract With the rapid development of multimedia and communication technology,the amount of multilingual speech data on the Internet is increasing. Speech recognition technology is the core for media analysis and processing. How to quickly expand from a few major languages such as Chinese and English to more languages has become a prominent issue yet to be overcome in order to improve multilingual processing capabilities. This article summarizes the latest progress in the field of acoustic model modeling,and discusses breakthroughs needed by traditional speech recognition technology in the course of moving from single language to multi-languages. The latest end-to-end speech recognition technology was exploited to construct a keyword spotting system,and the system achieves favorable performance. The approach is detailed as follows:1)multi-lingual hierarchical and structured acoustic model modeling method;2)multilingual acoustic modeling based on language classification information;3)end-to-end keyword spotting based on frame-synchronous alignments.

Keywords Multilingual,Speech recognition,Acoustic model

1 引言

语音交互作为人类重要的交流方式,已成为智能交互领域的重要组成部分。由于地域及文化的差异,全球语言种类多达7000余种,其中20种语言的使用人口超过5000万。2013年9月和10月,中国先后提出共建“丝绸之路经济带”和“21世纪海上丝绸之路”(以下简称“一带一路”)的重大倡议,并得到了国际社会的高度关注。“一带一路”覆盖沿线国家和地区26个,涉及中亚、东南亚、南亚、西亚乃至欧洲部分区域,东牵亚太经济圈,西系欧洲经济圈。面对众多国家和地区范围的海量信息,多语言语音识别是及时分析和处理这些信息的一个重要技术手段。然而,目前主流语音识别技术打造一个工业级的语音识别系统需要几千甚至上万小时来标注语音。现阶段,构建一个新语种的语音识别系统大约需要2年时间,这无法满足实际应用的需求。

经过多年发展,当前国内语音信息处理技术基本上与国外主流研究机构并驾齐驱,中国科学院声学所、中国科学院自动化所、北京大学、清华大学、中国科学技术大学等多家单位都取得了丰硕成果。国内语音技术研发机构在国内、国际评测中都取得了优异的成绩。但时至今日,多语言语音识别技术的应用仍然困难重重。不同语言间言语生成、听觉感知和言语描述体系等方面的差异,给我国多语言语音信息处理技术带来了巨大挑战。

本文总结了声学研究所在解决多语言语音识别这一技术问题方面的研究进展:1)基于模型参数共享的多语言声学建模[1-2],即基于模型参数共享的多语言建模策略,该方法借助

到稿日期:2021-09-01 返修日期:2021-10-18
通信作者:颜永红(yanyonghong@hccl.ioa.ac.cn)

已经引入声学模型建模中的机器学习策略，从更高维度的向量空间中寻找不同发音特征的共性，以实现数据共享；2）基于语种分类信息的多语言声学建模[3-4]，语种分类是一个基于长时的分类任务，该方法利用神经网络长时与短时相结合，进行多任务深度神经网络框架的优化训练，同时利用语种信息辅助多语言声学模型进行自适应；3）基于帧级别对齐的端到端关键词检索技术，该技术为端到端语音识别提供了识别结果的帧级别对齐信息，显著提升了关键词的时间戳精度与置信度可靠性。

本文第2节介绍了基于模型参数共享的多语言声学建模工作；第3节描述了基于语种分类信息的多语言声学建模工作；第4节总结了基于帧级别对齐的端到端关键词检索技术方面的工作；最后总结全文。

2 基于模型参数共享的多语言声学建模

基于模型参数共享的多语言建模技术借助已经引入声学模型建模中的机器学习策略，从更高维度向量空间中寻找不同发音特征的共性，从而达到数据共享的目的。

基于模型参数共享的多语言声学建模技术一般采用多任务学习的深度神经网络结构，多个逻辑分类 softmax 单元对应多个参与训练的语言。在模型优化期间，深度神经网络隐含层在多个语言之间共享，而输出层（softmax 层）是语言相关的基于区间的 softmax（block-softmax）结构。共享的隐含层基于全部数据进行更新，各个独立的输出层则基于单语言数据进行更新。在基于隐马尔可夫模型[1-2]（HMM）或联结主义时间分类模型[3]（CTC）的多语言语音识别模型中，通常共享除输出层之外的所有隐含层[4-11]。在基于编-解码器模型的多语言语音识别模型中[12-13]，通常只共享编码器[14-15]。在大规模的多语言训练数据上，基于 CTC 或者编-解码器的语音识别模型比基于 HMM 的语音识别模型的精度更高，但是在低资源的多语言训练数据上（如 10 h 以内）HMM 模型更容易收敛。在对多语言进行声学建模时，若采用统一的建模单元，如国际音标（IPA）[5-7]、拉丁字母[4,16-19]、字节对编码（BPE）[20]、UTF-8 编码字符[21]等，或者直接将各个语种的字符集合并[22-23]，则多语言之间将共享全部的模型。低资源的多语言训练数据更适合使用 IPA 作为建模单元，因为 IPA 包含了语音学知识，所以有利于多语言语音识别模型在少量的训练数据上进行收敛。相反地，拉丁字母、BPE 和 UTF-8 编码字符缺少语音学知识，且多语言语音识别模型需要大量的多语言训练数据才能收敛。

基于模型参数共享的多语言声学建模通常会遇到数据不平衡及低资源语种的语音识别的精度依然较低的问题，可以在训练时增加低资源语种音频的采样概率[14-15]，或者通过数据增强的手段来增加低资源语种音频[4,11,16]，以确保各个语种训练数据具有相近的规模。也可以使用知识迁移，在多语言训练的语音识别模型的基础上，对目标低资源语种进行自适应，该方法通常能在目标语种上获得更高的精度[5,17]。因此，首先使用多语言训练数据，通过预训练得到多语言语音识别模型，然后使用目标语种训练数据微调模型参数。在预训练的过程中使用元学习的训练方法，可以在参数微调阶段获得更好的识别精度[8]。

根据实际使用经验可知，基于模型参数共享的多语言声学建模技术简单易用，能够稳定地提升语音识别精度。选取中文普通话、英语、韩语作为丰富资源多语言语种，选取日语作为低资源语言。另外，为了验证该方法在跨语言上的作用，添加广东话作为另一种低资源语言。建模过程分为两个步骤，即模型参数共享的多语言声学模型训练阶段和跨语言快速调整阶段。其中，训练数据在语音之间进行句子级别和帧级别的顺序被打乱，以最大限度地避免语言和数据的偏向性。选取实验平台 Kaldi[24] 作为语音识别的工具。实验结果表明，基于模型参数共享的多语言声学模型在混合中、英、韩数据上训练后，中文普通话的字错率为 42.4%，英语的词错率为37.7%，韩语的字错率为 49.4%。相比各自独立语种训练的基线系统，本文方法相对提升比例为 0.4%～2.5%。通过共享隐含层参数，多种语言的数据在建模过程中能互相辅助和促进。本文方法在日语测试集上的字错误率为 54.5%，比基线系统提升了 22.6%。基于模型参数共享的多语言声学建模对快速辅助新的低资源语言建模有提升效果。5 h 数据训练的广东话基线模型字错率为 73.8%，在基于模型参数共享的多语言声学模型下广东话的字错率为 59.6%，此基线系统提升了 19.2%，因此，基于模型参数共享的多语言声学建模可以有效缓解共享音素集覆盖情况差的目标语言识别精度不足的问题。

3 基于语种分类信息的多语言声学建模

将语种分类与多语言语音识别两个任务相结合的多任务深度神经网络框架是解决多语言语音识别问题的方案之一。多语言声学模型建模部分采用多语言隐含层参数共享、输出层参数独立的建模方式。语种分类是一个基于长时的分类任务，传统的神经网络分类大都是基于帧级别的短时分类任务，因此在语种分类部分，通常加入统计信息元组对段级别的信息进行统计，以提升语种分类任务的鲁棒性[25-30]。语种分类信息通常以如下两种方式与多语言声学建模结合。

（1）在输入层或者隐含层中拼接语种向量。语种向量包括热独（one-hot）向量[31]、嵌入（embedding）特征[32-33]、神经网络瓶颈（bottleneck）特征[34]。通过语种分类信息，多语言声学模型可以充分学习各个语种的表示，而不仅仅是学习单一语种的表示。由于基于热独向量的模型需要在训练和解码阶段事先知道语音的语种，因此使用场景受到了明显的限制；基于嵌入特征或者瓶颈特征的模型在解码阶段不需要事先知道语音的语种，因此使用范围更加广泛。但是嵌入特征或者瓶颈特征的质量将会决定多语言模型的识别精度，而声学环境和说话人对这类语种相关特征产生的影响尚无详细研究。

（2）在输出层预测语种。例如，在预测文本的句首或者句尾插入语种标签[22,35-36]，并由语音识别模型直接预测，或者由语种分类器预测语种标签，并与语音识别模型进行联合训练[32-33,37]。语种分类器不仅辅助选择多语言解码器的特定语言的输出，还提供语种分类信息辅助进行多语言声学模型的自适应训练[38-40]。多语言声学模型的自适应训练主要是在多语言声学模型的共享隐含层与特定输出层之间加入语言门

结构[41],该结构将共享隐含层的输出与语种分类信息相结合,以提升共享隐含层输出的语言特定性,进而提升参数共享的多语言声学模型的性能。此外,若测试语种不属于训练语种,通常结合对抗学习的方法,使得多语言语音识别模型学习语种无关的表示[42-43]。目前多语言语音识别在使用语种分类器时,尚未充分认识语种分类是一个基于长时的分类任务,也未能与语种识别领域的前沿研究充分结合。

为了验证语种分类信息对多语言语音识别的识别精度的影响,选取 IARPA BABEL[1] 语料集中的广东话、土耳其语和越南语 3 种语言进行相关的实验探索。其中,广东话有 78.9 h,土耳其语有 75.2 h,越南语有 78.5 h。基于共享隐含层的多任务神经网络来构建多语言神经网络模型。通过官方的 20 h 开发集数据来评估多语言语音识别系统的识别准确率和语种分类准确率,每种语言开发集语音的平均时长为:广东话 12.3 s,土耳其语 7.1 s,越南语 9.6 s。

选取实验平台 Kaldi 作为语音识别的工具。3 种语言的训练集的标注数据被用于构建相应的单语种语言模型,构建语言模型的实验平台为 SRILM[44] 开源工具包,构建声学模型的基本神经网络单元为时延神经网络。神经网络模型均基于交叉熵训练准则进行模型参数更新。前端声学特征为 40 维梅尔频率倒谱系数特征,该特征为 Kaldi 工具中标准的高分辨率梅尔频率倒谱系数特征,使用的语言模型为各语言训练集标注文本构建的 3 阶单语种语言模型。发音字典为 IARPA BABEL 数据集提供的基于 SAMPA[45] 音素标注的发音字典。在声学模型建模过程中,构建基于高斯混合模型-隐马尔可夫模型的声学模型,用于生成帧级别的三音素状态对齐信息。其中,三音素状态是通过决策树聚类算法生成的上下文相关的三音素状态,决策树分裂的问题集是由 K-means 聚类算法通过音素统计信息得到的。

基于共享隐含层的多任务神经网络的多语言声学模型的具体配置结构为:前 5 层时延神经网络[46-48] 结构为共享隐含层网络,之后隐含层为每一种语言配置了一层语种特定输出层。为避免模型规模对实验结果的影响,单语言声学模型的结构与多语言声学模型的结构保持一致,因此,单语言声学模型由 6 层时延神经网络结构组成,每层的节点数与多语言声学模型保持一致。在多语言语音内容和语种分类协同框架下,语种分类模块可以生成语种特征向量信息,以协助完成多语言声学模型语种的自适应训练。

相比单语言声学模型(3 种语言的平均错误率为 51.77%),直接将 3 种语言进行多语言声学模型的构建会直接影响各语言的声学模型建模的效果,使平均错误率升高至 52.68%。应用多语言声学模型语种自适应训练的方法,可以从一定程度上提升多语言声学模型的性能。改进的基于语言门控单元的语种自适应训练方法(缩写为“LGU”)可以在一定程度上提升多语言声学模型的性能(平均错误率为 49.97%)。基于共享隐含层的语种自适应训练方法(缩写为“SA”)和基于输入的语种自适应训练方法(缩写为“AI”)可以实现性能类似或更好的声学模型,平均错误率分别为 50.12%和 49.69%。

对于 LGU 和 SA,语言种类信息均与共享隐含层信息进行结合,从而提升共享隐含层输出信息的语种区分性。实验结果显示,在这两种将语言种类特征从共享隐含层之后输入到声学模型的方法中,LGA 可以提供更好的语种自适应训练效果。而 AI 则是将语言种类信息与原始频谱声学特征信息进行结合,从声学模型建模前端来提升模型整体的语种区分性。对于 SA 和 AI 这两种直接将语言种类信息输入到声学模型中的语种自适应训练方法来说,从声学模型的前端输入语言种类特征可以利用较多的参数进行声学模型的语种自适应训练。

4 基于帧级别对齐的端到端关键词检索

近年来,端到端的语音识别框架发展迅速,已经成为了与基于隐马尔可夫模型的语音识别并列的主流语音识别框架之一。然而,现有的两大类端到端语音识别框架,即基于注意力机制的编码器-解码器架构[13,49] 以及基于联结时间分类的神经网络架构[3,50],由于在模型训练原理上存在限制,自身无法为其识别结果提供准确的时间起止点和可靠的置信度,因此均难以被直接应用于基于文本查询的关键词检索任务。

基于示例查询(QBE)的关键词检索直接通过语音来查询多语言音频中是否包含关键词,适用于语种无关的关键词检索任务。通常借助多语言声学建模的方法,从语音中提取多语种瓶颈特征来训练 QBE 模型[51-53]。但目前 QBE 关键词检索的性能仍然非常低,不是主流的关键词检测方法,因此本文详细介绍了基于文本查询的关键词检索任务。

基于帧级别对齐的端到端关键词检索框架采用如下具体策略:首先使用联结时间分类/注意力联合端到端语音识别模型[54] 解码待测语句,将解码结果语句转换为音素序列,利用一个基于交叉熵准则训练的逐帧音素分类器输出的音素概率,使用一个基于动态规划的帧级别对齐算法为各个音素计算时间起止点和置信度,进而获得各个单词的时间起止点和帧平均音素后验概率置信度。最后在解码 N-最佳假设结果中匹配关键词,并利用时间起止点和置信度信息合并中被重复匹配的关键词,保留各个时间范围内互相冲突的相同关键词识别结果中置信度得分最高的一个,得到最终的检索结果。

在 ESPnet 平台[55] 上开展相关实验。基于帧级别对齐的端到端关键词检索系统使用 Transformer[56] 作为神经网络的基本结构。音素分类器与语音识别共享底层的 9 层 Transformer 编码器,另外各自独占较高的 3 层编码器。语音识别部分的解码器为 6 层。各个编码器与解码器层中多头注意力的维度为 320,头数为 4,前馈神经网络的维度为 2048。语音识别模型与音素分类器进行多任务学习联合训练,两者的损失函数按照 0.9 和 0.1 的比例进行线性插值。实验在越南语自由交谈数据集上进行,训练集和测试集各约 100 h 和 5 h。对于越南语,音素分类器的建模单元为 24 个辅音、11 个不带调的元音以及静音,共 36 个音素标签。语音识别的预测基本单元为由字节对编码[57-58] 算法生成的 1 000 个越南语子词。

[1] https://www.iarpa.gov/index.php/research-programs/babel

使用一个由 Kaldi 平台标准流程训练的三音子建模的高斯混合模型-隐马尔可夫模型语音识别系统，得到音素分类器训练所需的训练集语音逐帧音素标注。

实验结果显示，基于帧级别对齐的端到端关键词检索系统相比基于隐马尔可夫模型的基线系统，其关键词检索 $F1$ 值从 72.0%提高到了 77.6%，且精度和召回率均有提升。分析显示，音素分类器和帧级别对齐在不给语音识别性能造成负面影响的同时，能够为关键词检索提供更准确的时间起止点，提高系统的检索性能。此外，使用解码器置信度和对齐置信度进行插值可得到混合置信度，基于帧级别对齐的端到端关键词检索有效弥补了端到端语音识别预测标签置信度在高置信度区域不可靠的缺陷，因此，基于帧级别对齐的端到端关键词检索系统在各个置信度区间内都可以依靠置信度过滤并进一步调节关键词检索结果的精度和召回率。

结束语 本文总结了声学研究所在解决多语言语音识别技术问题方面的近期研究进展：1）基于模型参数共享的多语言声学建模缓解了共享音素集覆盖情况差的目标语言识别精度不足的问题；2）基于语种分类信息的多语言声学建模提升了多语言声学模型的性能；3）基于帧级别对齐的端到端关键词检索技术显著提升了关键词的时间起止点准确度和置信度可靠性，使端到端语音识别在关键词检索任务中得到了有效利用。未来，针对句内语码转换场景下的多语言语音识别，无监督学习框架下的多语言语音识别等问题可能会成为多语言语音识别领域下一阶段的研究重点。

参 考 文 献

[1] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[2] POVEY D, PEDDINTI V, GALVEZ D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI [C]// Interspeech. 2016: 2751-2755.

[3] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// Proceedings of the 23rd International Conference on Machine Learning. 2006: 369-376.

[4] LIU C, ZHANG Q, ZHANG X, et al. Multilingual graphemic hybrid ASR with massive data augmentation[J]. arXiv: 1909.06522, 2019.

[5] TONG S, GARNER P N, BOURLARD H. An investigation of multilingual ASR using end-to-end LF-MMI[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). IEEE, 2019: 6061-6065.

[6] TONG S, GARNER P N, BOURLARD H. Cross-lingual adaptation of a CTC-based multilingual acoustic model [J]. Speech Communication, 2018, 104: 39-46.

[7] TONG S, GARNER P N, BOURLARD H. Fast Language Adaptation Using Phonological Information [C]// INTERSPEECH. 2018: 2459-2463.

[8] HSU J Y, CHEN Y J, LEE H. Meta learning for end-to-end low-resource speech recognition[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7844-7848.

[9] DALMIA S, SANABRIA R, METZE F, et al. Sequence-based multi-lingual low resource speech recognition[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4909-4913.

[10] CHEN Y C, HSU J Y, LEE C K, et al. DARTS-ASR: Differentiable architecture search for multilingual speech recognition and adaptation[J]. arXiv: 2005.07029, 2020.

[11] THOMAS S, AUDHKHASI K, KINGSBURY B. Transliteration Based Data Augmentation for Training Multilingual ASR Acoustic Models in Low Resource Settings [C]// INTERSPEECH. 2020: 4736-4740.

[12] GRAVES A. Sequence transduction with recurrent neural networks[J]. arXiv: 1211.3711, 2012.

[13] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 4960-4964.

[14] PRATAP V, SRIRAM A, TOMASELLO P, et al. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters [J]. arXiv: 2007.03001, 2020.

[15] LI B, PANG R, SAINATH T N, et al. Scaling end-to-end models for large-scale multilingual asr[J]. arXiv: 2104.14830, 2021.

[16] DATTA A, RAMABHADRAN B, EMOND J, et al. Language agnostic multilingual modeling [C] // ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 8239-8243.

[17] KARAFIÁT M, BASKAR M K, WATANABE S, et al. Analysis of multilingual sequence-to-sequence speech recognition systems [J]. arXiv: 1811.03451, 2018.

[18] ADAMS O, WIESNER M, WATANABE S, et al. Massively multilingual adversarial speech recognition [J]. arXiv: 1904.02210, 2019.

[19] CHO J, BASKAR M K, LI R, et al. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling[C]// 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018: 521-527.

[20] ZHOU S, XU S, XU B. Multilingual end-to-end speech recognition with a single transformer on low-resource languages[J]. arXiv: 1806.05059, 2018.

[21] LI B, ZHANG Y, SAINATH T, et al. Bytes are all you need: end-to-end multilingual speech recognition and synthesis with bytes[C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 5621-5625.

[22] HOU W, DONG Y, ZHUANG B, et al. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning[C]// INTERSPEECH. 2020: 1037-1041.

[23] WATANABE S, HORI T, HERSHEY J R. Language independent end-to-end architecture for joint language identification and speech recognition[C]// 2017 IEEE Automatic Speech Recogni-

tion and Understanding Workshop (ASRU). IEEE, 2017: 265-271.

[24] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]// IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.

[25] CAI W, CAI Z, ZHANG X, et al. A novel learnable dictionary encoding layer for end-to-end language identification[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5189-5193.

[26] CAI W, CAI Z, LIU W, et al. Insights in-to-end learning scheme for language identification[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5209-5213.

[27] MIAO X, MCLOUGHLIN I. Lstm-tdnn with convolutional front-end for dialect identification in the 2019 multi-genre broadcast challenge[J]. arXiv: 1912.09003, 2019.

[28] MIAO X, MCLOUGHLIN I, YAN Y. A New Time-Frequency Attention Tensor Network for Language Identification[J]. Circuits, Systems, and Signal Processing, 2020, 39(5): 2744-2758.

[29] BEDYAKIN R, MIKHAYLOVSKIY N. Low-Resource Spoken Language Identification Using Self-Attentive Pooling and Deep 1D Time-Channel Separable Convolutions [J]. arXiv: 2106.00052, 2021.

[30] TJANDRA A, CHOUDHURY D G, ZHANG F, et al. Improved language identification through cross-lingual self-supervised learning[J]. arXiv: 2107.04082, 2021.

[31] KANNAN A, DATTA A, SAINATH T N, et al. Large-scale multilingual speech recognition with a streaming end-to-end model[C]// Proc. Interspeech 2019, 2019: 2130-2134.

[32] TOSHNIWAL S, SAINATH T N, WEISS R J, et al. Multilingual speech recognition with a single end-to-end model[C]// 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 4904-4908.

[33] PUNJABI S, ARSIKERE H, RAEESY Z, et al. Streaming end-to-end bilingual asr systems with joint language identification [J]. arXiv: 2007.03900, 2020.

[34] MIILLER M, STIIKER S, WAIBEL A. Multilingual adaptation of RNN based ASR systems[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5219-5223.

[35] SEKI H, WATANABE S, HORI T, et al. An end-to-end language-tracking speech recognizer for mixed-language speech [C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4919-4923.

[36] WATERS A, GAUR N, HAGHANI P, et al. Leveraging language id in multilingual end-to-end speech recognition [C]// 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019: 928-935.

[37] PUNJABI S, ARSIKERE H, RAEESY Z, et al. Joint ASR and language identification using RNN-T: An efficient approach to dynamic language switching[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 7218-7222.

[38] LIU D, WAN X, XU J, et al. Multilingual Speech Recognition Training and Adaptation with Language-Specific Gate Units [C]// 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2018: 86-90.

[39] LIU D, XU J, ZHANG P, et al. A unified system for multilingual speech recognition and language identification[J]. Speech Communication, 2021, 127: 17-28.

[40] LIU D, XU J, ZHANG P. End-to-End Multilingual Speech Recognition System with Language Supervision Training[J]. IEICE TRANSACTIONS on Information and Systems, 2020, 103(6): 1427-1430.

[41] KIM S, SELTZER M L. Towards language-universal end-to-end speech recognition[C]// Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2018: 4914-4918.

[42] YI J, TAO J, WEN Z, et al. Adversarial multilingual training for low-resource speech recognition[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 4899-4903.

[43] YI J, TAO J, WEN Z, et al. Language-adversarial transfer learning for low-resource speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 27(3): 621-630.

[44] STOLCKE A. Srilm—an extensible language modeling toolkit [C]// Proc. of the International Conference on Spoken Language Processing. 2002: 901-904.

[45] WELLS J. SAMPA computer readable phonetic alphabet[M]// Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter, 1997.

[46] HAMPSHIRE W. A novel objective function for improved phoneme recognition using time delay neural networks[C]// Proc. of the International 1989 Joint Conference on Neural Networks. 1989: 235-241.

[47] WAIBEL A, HANAZAWA T, HINTON G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(3): 328-339.

[48] HAMPSHIRE J B, WAIBEL A H. A novel objective function for improved phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Neural Networks, 1990, 1(2): 216-228.

[49] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]// Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. 2015: 577-585.

[50] LI J, YE G, DAS A, et al. Advancing acoustic-to-word CTC model[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5794-5798.

[51] YUAN Y, LEUNG C C, XIE L, et al. Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection[C]// 2017 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 5645-5649.
[52] RAM D, MICULICICH L, BOURLARD H. Multilingual bottleneck features for query by example spoken term detection[C]// 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019: 621-628.
[53] RAM D, MICULICICH L, BOURLARD H. Neural network based end-to-end query by example spoken term detection[J]. IEEE/ACM Transactionson Audio, Speech, and Language Processing, 2020, 28: 1416-1427.
[54] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.
[55] WATANABE S, HORI T, KARITA S, et al. Espnet: end-to-end speech processing toolkit[C]// Interspeech. 2018: 2207-2211.
[56] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in neural information processing systems. 2017: 5998-6008.
[57] GAGE P. A new algorithm for data compression[J]. C Users Journal, 1994, 12(2): 23-38.
[58] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[J]. arXiv: 1508.07909, 2015.

CHENG Gao-feng, born in 1990, Ph.D, assistant professor. His main research interests include speech recognition and deep learning.

YAN Yong-hong, born in 1967, Ph.D, professor. His main research interests include speech processing and recognition, language/speaker recognition, and human computer interface.

(责任编辑:李亚辉)