ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 101－110
DOI:10. 16337/j. 1004-9037. 2023. 01. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于预训练与音素字节对编码的越南语识别

沈之杰，郭　武

(中国科学技术大学电子工程与信息科学系,合肥 230027)

摘　要: 基于无监督预训练技术的wav2vec 2.0在许多低资源语种上获得了良好的性能,成为研究的热点。本文在预训练模型的基础上进行越南语连续语音识别。将语音学信息引入到基于链接时序分类代价函数(Connectionist temporal classification,CTC)的声学建模中,选取音素与含位置信息的音素作为基础单元。为了平衡建模单元数目以及模型的精细程度,采用字节对编码(Byte-pair encoding,BPE)算法生成音素子词,将上下文信息结合到声学建模过程。实验在美国NIST的BABEL任务低资源的越南语开发集上进行,所提算法相对wav2vec 2.0基线系统有明显改进,识别词错误率由37.3%降低到29.4%。

关键词: 低资源语音识别;建模单元;字节对编码;音素子词;预训练;越南语识别

中图分类号: TN912.34　　**文献标志码:** A

Vietnamese Speech Recognition Based on Pre-training and Phone-Based Byte-Pair Encoding

SHEN Zhijie, GUO Wu

(Department of Electronic Engineering & Information Science, University of Science and Technology of China, Hefei 230027, China)

Abstract: Based on the unsupervised pre-training technology, wav2vec 2.0 has become a research hotspot for the state of the art performance in many low-resource languages. In this paper, the Vietnamese continuous speech recognition is carried out on the basis of the pre-trained model. The phonetics information is integrated into the connectionist temporal classification (CTC) loss function based acoustic modeling, and the phones and the position dependent phones are selected as the basic modeling units. To balance the number of modeling units and the refinement of the model, a byte-pair encoding (BPE) algorithm is used to generate phone based subwords, and the contextual information is integrated into the acoustic modeling process. Experiments are carried out on the low-resource Vietnamese development set of NIST's BABEL task, and the proposed algorithm significantly improves the wav2vec 2.0 baseline system. The word error rate is reduced from 37.3% to 29.4%.

Key words: low-resource speech recognition; modeling unit; byte-pair encoding; phone based subword; pre-training; Vietnamese speech recognition

---

基金项目: 国家自然科学基金(U1836219)。
收稿日期: 2021-07-27;**修订日期:** 2021-12-27

引　言

近年来基于端到端的语音识别声学建模取得了显著进展，这类模型通常基于序列到序列（Sequence-to-sequence，S2S）框架或链接时序分类（Connectionist temporal classification，CTC）代价函数。与传统的基于隐马尔可夫模型（Hidden Markov model，HMM）的框架相比，基于端到端的模型能达到更高的识别率。然而基于端到端的模型一般需要数百甚至数千小时的有标注音频数据，目的是减轻模型训练的过拟合问题并取得较好的自动语音识别（Automatic speech recognition，ASR）性能。对于英语和汉语等大语种来说，通常数据量的需求可以得到满足。然而世界上大多数其他语言的音频资源相当有限，研究者正越来越将注意力转移到许多低资源语种的连续语音识别上[1]。针对训练语料不足的问题，可以采用预训练技术[2]。在预训练模型的基础上再进行有监督训练，也称为有监督迁移学习[3]，可将从其他任务学到的知识应用于目标任务[4]。Dalmia等[5]将300 h的英语语料纳入训练数据，在蒙古语等4类小语种上实现了识别率的提升。Cho等[6]使用混合注意力/CTC网络，用10个低资源语种预训练模型，并迁移到其他4个低资源语种。另一个解决思路是使用无标注的数据对端到端模型进行无监督预训练。例如Jiang等[7]通过掩盖预测编码（Masked predictive coding，MPC）对Transformer声学编码器进行了预训练，并且该模型在较高的ASR基线系统之上得到了进一步的提升。Conneau等[8]基于wav2vec 2.0在53个语种上进行预训练，提出了XLSR-53，并在美国NIST的BABEL语料库的等15个语种上进行了微调。

近年来，许多端到端系统使用基于书写字符的子词作为声学建模单元[9-10]，例如使用字节对编码（Byte-pair encoding，BPE）算法或WordPiece算法来构建子词。这类方法的优点在于其简单性，此外得到的单元不仅可用于集外词（Out of vocabulary，OOV）识别，而且还避免了专业领域知识（如发音词典和基于决策树的状态绑定所需的问题集）与额外的处理步骤（如HMM模型中的音素状态建模）。然而相较于书写字符，语音识别中发音词典的音素与声学建模的联系更加紧密，并且对于本文所研究的越南语，它具有6种声调，但其书写系统却是全拉丁字符化的。拉丁字符几乎无法体现声调属性，这可能导致系统对多音字无法进行准确识别。另外，即使对于无调的英语，Zweig等[11]的研究表明，在Switchboard语料库上，基于字符的CTC系统在识别率上仍然不及基于音素的CTC系统。针对以上问题，Xu等[12]使用发音词典和对齐器找出子词单元和音素单元的对应性，并利用该信息指导子词分割的过程。其他研究[13-14]则尝试将单词分解为不同形式的建模单元序列（如字符序列或音素序列），对不同的单元序列分别构建子词单元，在关键词检索（Keyword search，KWS）或语音识别任务上均取得了性能提升。其中He等[13]研究表明，在低资源场景下，基于音素的子词单元能够有效提高系统处理OOV问题的能力。

本文重点研究越南语声学建模问题，在IARPA BABEL语料库上进行实验。该语料库包含共约80 h的训练集和开发集，数据量十分稀缺，因此本文使用预训练的wav2vec 2.0模型进行初始化，并探索了适合越南语的建模单元。具体而言，使用发音词典将标注文本中的单词序列转换为音素序列，并采用字节对编码算法逐步合并出现频率较高的音素对，最终得到基于音素的子词单元。与传统的基于HMM的ASR系统所使用的上下文依赖的音素类似，音素子词能捕获相邻的音素之间的关联性。此外，音素子词也能够平衡建模单元集合的大小与解码序列长度，有利于提升端到端系统性能和效率。

本文使用的算法在BABEL越南语的开发集上进行评测，测试结果表明，基于音素BPE构建声学建模单元的系统相对采用其他建模单元的系统有明显的性能提升，相对于采用单字符的基线系统，识别错误率相对下降约21.2%。此外，本文系统优于近期其他学者提出的越南语识别系统。需要强调的是，

虽然本文将发音词典用于生成子词和解码,但是避免了传统方法的决策树聚类等步骤。另一方面,发音词典资源并不难获得,即便是难以获得的发音,也可由字音转换等技术得到。因此,本文方法仍保持了端到端系统的简洁性。

1 基于预训练模型和CTC的ASR框架

越南语的数据量十分有限,因此本文将预训练的wav2vec 2.0作为初始化模型,通过CTC代价函数完成有监督训练完成越南语ASR任务。wav2vec 2.0是由Facebook提出的语音表征自监督学习框架[15],如图1所示。wav2vec 2.0的训练包括2个步骤:预训练与微调。在预训练期间,该模型采用对比损失函数学习音频的上下文表征。在微调阶段,wav2vec 2.0可用于ASR声学建模,本文在它的顶部增加了分类器网络,该网络结合CTC技术,输出越南语识别结果[8]。

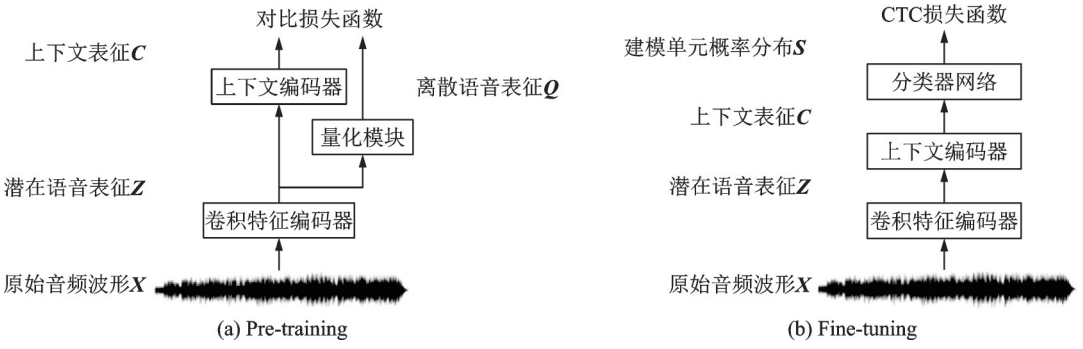

图1 wav2vec 2.0模型框架

Fig.1 Framework of the wav2vec 2.0 model

wav2vec 2.0包含一个多层卷积特征编码器$f: \boldsymbol{X} \mapsto \boldsymbol{Z}$,一个上下文编码器$g: \boldsymbol{Z} \mapsto \boldsymbol{C}$和一个量化模块$h: \boldsymbol{Z} \mapsto \boldsymbol{Q}$,其中$\boldsymbol{X}$、$\boldsymbol{Z}$、$\boldsymbol{C}$、$\boldsymbol{Q}$分别表示输入音频、潜在语音表征、上下文表征以及离散语音表征。wav2vec 2.0的输入是原始音频,原始音频首先经过多层卷积特征编码器,被编码为潜在语音表征$\boldsymbol{Z}=(z_1, \cdots, z_T)$,其中$T$为编码后的帧数。接着,潜在语音表征$\boldsymbol{Z}$被送入上下文编码器,得到上下文表征$\boldsymbol{C}=(c_1, \cdots, c_T)$。上下文表征包含了整个序列的信息。量化模块仅在预训练阶段使用,它将潜在语音表征$\boldsymbol{Z}$进行离散化处理,得到离散语音表征$\boldsymbol{Q}=(q_1, \cdots, q_T)$,离散语音表征$\boldsymbol{Q}$用于在自监督损失函数中表示目标。

1.1 特征编码器

特征编码器f(图2)由若干结构相同的子模块级联而成,其中N_f指特征编码器中级联结构相同的子模块个数。每个子模块包含一个一维时间卷积(Conv1d)、层归一化(LayerNorm)和高斯误差线性单元(Gaussian error linear unit,GELU)激活函数[16],计算过程为

$$\mathrm{Block}(\boldsymbol{X})=\mathrm{GELU}\left(\mathrm{LayerNorm}\left(\mathrm{Conv1d}(\boldsymbol{X})\right)\right) \tag{1}$$

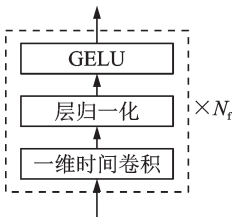

图2 特征编码器框架

Fig.2 Framework of the feature encoder

1.2 上下文编码器

上下文编码器g的输入是潜在语音表征$\boldsymbol{Z}$,输出是上下文表征$\boldsymbol{C}=(c_1, \cdots, c_T)$。上下文编码器的底部是卷积位置编码层,上方堆叠了若干个结构相同的Transformer层,其结构如图3所示,其中N_g指上下文编码器上方堆

叠的结构相同的Transformer层的个数，关键公式为

$$\text{Attention}(\boldsymbol{R},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{R}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D}}\right) \tag{2}$$

$$\text{MultiHead}(\boldsymbol{R},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{HEAD}_1,\cdots,\text{HEAD}_h)\boldsymbol{W}^{O} \tag{3}$$

$$\text{HEAD}_i=\text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}\right) \tag{4}$$

式中：$\boldsymbol{R}\in\mathrm{R}^{T\times D}$、$\boldsymbol{K}\in\mathrm{R}^{T\times D}$、$\boldsymbol{V}\in\mathrm{R}^{T\times D}$分别为查询、键与值，且三者相等；$D$为表征维度；$h$为头数；$\boldsymbol{W}^{O}\in\mathrm{R}^{hD'\times D}$为输出投影矩阵；$\boldsymbol{W}_i^{Q}\in\mathrm{R}^{D\times D'}$、$\boldsymbol{W}_i^{K}\in\mathrm{R}^{D\times D'}$、$\boldsymbol{W}_i^{V}\in\mathrm{R}^{D\times D'}$为第$i$个头的投影矩阵；$D'$为内部维度；$i=1,2,\cdots,h$。

1.3 微调

本文在已通过预训练的wav2vec 2.0模型上进行微调，如图1(b)所示。具体而言，去除了量化模块，并在上下文编码器的顶部增加了分类器网络h：$\boldsymbol{C}\mapsto\boldsymbol{S}$，将上下文表征映射为建模单元的概率分布[8]。该网络包含一个随机初始化的全连接层与归一化指数激活函数(Softmax)，表达式为

$$\boldsymbol{S}=\text{Softmax}(\boldsymbol{C}\boldsymbol{W}+\boldsymbol{b}) \tag{5}$$

式中：$\boldsymbol{C}\in\mathrm{R}^{T\times D}$为上下文表征；$\boldsymbol{W}$为权重矩阵；$\boldsymbol{b}$为偏置向量；$\boldsymbol{S}\in\mathrm{R}^{T\times|L'|}$为每一帧在扩展的建模单元$L'=L\cup\varnothing$上的概率分布；$L$为原始的建模单元集合；$\varnothing$为空白符号。

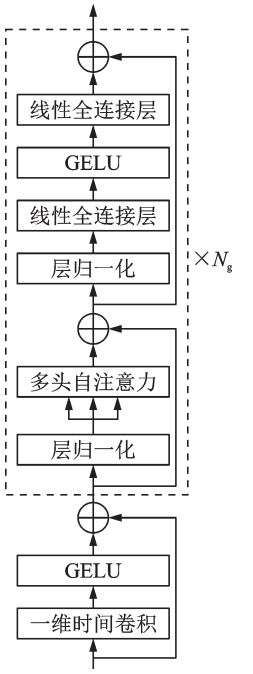

图3 上下文编码器框架

Fig.3 Framework of contextualized encoder

本文针对低资源越南语，对3种类型的建模单元展开研究，分别是字符、单音素与含位置信息的音素，并结合BPE生成相应的子词单元，具体内容将在第2部分介绍。最后本文使用CTC算法[17]计算代价函数。

1.4 解码

由于本文的研究重点为ASR声学建模，因此将模型在越南语微调之后，使用常规的浅融合策略[18]对ASR系统进行解码，即将声学模型得分与语言模型得分进行线性组合，并使用束搜索算法，即

$$\hat{\boldsymbol{Y}}=\arg\max_{\boldsymbol{Y}}\left(\log P(\boldsymbol{Y}|\boldsymbol{X})+\gamma\log P(\boldsymbol{Y})\right) \tag{6}$$

式中：$\hat{\boldsymbol{Y}}$为解码得到的建模单元序列；γ是介于0与2的可调超参数。在本文中，$\hat{\boldsymbol{Y}}$可能为字符序列、单音素序列、含位置信息的音素或它们对应的子词序列。$P(\boldsymbol{Y}|\boldsymbol{X})$由声学模型计算得到，而$P(\boldsymbol{Y})$由$N$元语法($N$-gram)语言模型计算得到。系统解码框架如图4所示。

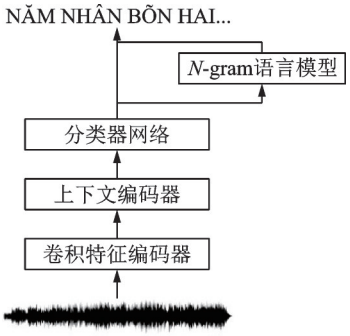

图4 系统解码框架

Fig.4 Decoding framework of the system

2 采用BPE的建模单元构建

越南语的书写系统是拉丁化的，端到端建模中最常见的是采用字符和词作为建模单元。在训练语料不足情况下，由于会出现过拟合，一般不采用词作为建模单元。另一方面，既然越南语的发音词典是很容易获得的，本文选择3种类型的基础单元：字符、单音素与含位置信息的音素，并利用BPE算法分别将这些单元构造为子词。表1展示了同一句文本在不同建模单元下的形式作为对比。

2.1 字符

越南语每个单词由多个字符(即字母)构成，字符有大小写之分。考虑到字符的大小写只取决于语

表1 不同建模单元的标注文本示例

Table 1 Transcription examples of different modeling units

建模单元	标注文本示例
原始文本	KHÔNG HỐI HẬN HÈ
字符	K H Ô N G | H Ố I | H Ậ N | H È |
单音素	x_1 o_1 N_1 | h_2 oI_2 | h_6 @_6 N_6 | h_3 E_3 |
含位置信息的音素	x_1_B o_1_I N_1_E | h_2_B oI_2_E | h_6_B @_6_I N_6_E | h_3_B E_3_E |
字符子词	KHÔNG H@@ ỐI H@@ ẬN HÈ
单音素子词	x_1o_1N_1 h_2@@ oI_2 h_6@@ @_6N_6 h_3E_3
含位置信息的音素子词	x_1_Bo_1_IN_1_E h_2_B@@ oI_2_E h_6_B@@ @_6_IN_6_E h_3_BE_3_E

法,而与发音无关,本文将越南语的文本全部转为大写字符,共计93个。为了显式地学习文本中单词的分割,以提升系统词识别率,本文对音频标注文本进行了预处理操作,在前一个单词的末尾字符与后一个单词的首字符之间插入了“空格”符号(“|”),并将该符号加入建模单元集合中。

2.2 单音素与含位置信息的音素

与大语种类似,越南语也有完成字到音素对应关系的发音词典。越南语属于有调语言,共6种声调,发音的音素信息与书写的字符之间不存在严格的对应关系。在语音识别领域,单音素单元天生就是声学建模的基础单元。利用数据集已有的发音词典文件,将以词为单元的音频标注文本修改为以音素为单元的序列串。

更进一步地,即便是同一个音素,它在一个词的不同位置也可能对应不同的发音,本文假设更精细的音素建模单元有助于神经网络学习输入与输出之间的映射关系,因此本文将音素细分为4类:单个词、词头、词中与词尾,分别以后缀“_S”“_B”“_I”与“_E”进行区分。单音素共356个,含位置信息的音素共693个。值得注意的是,训练集中的许多单音素在4个类别中不一定都存在样本,例如,N_1这个音素只在词头、词中与词尾出现,不存在单个词的情况。因此,含位置信息的音素总数并非恰好为单音素总数的4倍。与2.1节的做法一致,对于单音素与含位置信息的音素,本文也在单词的边界处插入了“空格符”。

2.3 基于BPE算法的子词分解

以单个字符或音素作为建模单元均忽视了建模单元之间的上下文之间相关性,因此本文使用BPE算法构造子词单元。BPE算法是Gage等[18]提出的数据压缩算法。该算法迭代地执行,每次迭代时将最频繁出现的字节对替换为一个新的独立字节。

将BPE算法用于生成子词的方法如下:首先进行算法的初始化,建模单元集合构建为初始的符号(如越南语的字符)集,并将标注文本中的每个词表示为符号序列。算法在迭代的每一轮中,统计所有的符号对(即相邻的符号)在标注文本中出现的次数,将出现频率最高的符号对替换为一个新的符号。BPE算法需要预先设定迭代次数,用于算法的终止。最终,BPE算法将生成介于字符与单词之间的子词,并且子词的粒度和子词集合的大小仅取决于算法的迭代次数。例如,如果建模单元集合中的两个单元分别是“L”和“O”,并且它们在标注文本中相邻出现的频率很高,那么通过应用BPE操作,可以将它们合并到一个新的子词单元“LO”中。

在获得子词单元之后,需要对音频标注文本进行切分,将单词序列转换为子词序列。具体而言,对

标注文本中的每个词，遍历所有的子词单元，找到构成该单词的所有子词单元，从而将单词分解为子词。另外，由于当BPE操作数较大时，许多生成的BPE子词即为单词本身（如表1中的KHÔNG与HÈ）。针对此现象，本文对于子词单元不显式地学习单词的分割，即不再像2.1节与2.2节那样插入“空格”符号。因此，在子词中加入位置信息，对于那些处于原本单词开头或中间位置的子词，本文在其末尾添加特殊标识符（@@）。

2.4 基于BPE的越南语音素子词

考虑到字符、音素或者含位置信息的音素3种基础单元，本文将BPE算法应用到这3种单元，都可以获得不同的建模单元。许多端到端的ASR系统已经将BPE算法运用于音频标注文本，将单词序列分解为字符子词序列，并将这类字符子词作为声学模型建模单元。这种做法保持了端到端属性，无需外部发音词典等语言知识，同时平衡了建模单元集合的大小与解码序列的长度。

然而，对于越南语来说，以字符子词作为建模单元，依然存在着2.2节描述的发音与标注文本的不匹配问题。于是，本文研究了基于音素和含位置信息的音素的子词单元（统称为音素子词）。具体做法是，首先利用越南语的发音词典文件，该词典包含2列，第1列为单词，第2列为此单词对应的音素序列。对标注文本中的每个单词，均遍历发音词典，在找到匹配的词后，将单词替换为对应的音素序列。使用2.3节描述的BPE算法构造音素子词，并将音素串序列分解为音素子词序列。单音素子词的构造过程如图5所示。音素子词的优势主要有以下几点：

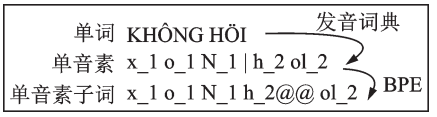

图5 单音素子词构造过程

Fig.5 Construction process of monophone subword

（1）继承了字符子词的多数优点。由于音素子词也是由BPE算法生成的，该算法能有效控制建模单元集合的大小。并且，对于OOV，只要它可以分解为集内的音素子词，系统也能进行有效识别。

（2）稳定性较好。每一个音素子词由1个或多个单音素构成，纳入了前后发音单位之间的联系，因此音素子词的发音相对稳定，不太容易存在单音素那样由前后音素的不同而导致发音不同的问题。

（3）保持了端到端系统的简洁性。传统的隐马尔科夫模型常常采用三音子作为建模单元，音素子词与这类单元均考虑了音素的上下文信息。但是，三音子需要进行状态绑定与决策树聚类等操作，并且需要得到强制对齐的标注。相比之下，音素子词的构建相对简洁，只需在文本层面上利用发音词典与BPE算法。

为了与音素子词进行对比，本文同样使用BPE算法构建字符子词。

3 实验配置及结果分析

3.1 实验数据集与评价指标

本文在BABEL语料库上进行了实验。该语料库收集自IARPA BABEL项目，主要包含亚洲和非洲等多个低资源语种的电话对话语音、脚本录音和远场录音。本文使用BABEL语料库所包含的越南语，该语言是NIST 2013语音识别及关键词识别评测的比赛语种，数据库包括共约80 h的训练集和开发集，以及约10 h的评估集。由于评估集没有标注，本文将BABEL的开发集作为本文的测试集，BABEL的训练集随机的10%作为本文的开发集。语音数据的格式为8 bit、8 kHz采样率、A律。本文将所有音频都上采样到16 kHz，以满足wav2vec 2.0的输入要求，以fairseq[19]作为实验平台。本文主要采用词错误率（Word error rate，WER）作为系统评价指标。

3.2 基于wav2vec 2.0的ASR系统

3.2.1 声学模型

声学模型的输入为16 kHz的wav音频文件，输出为相应的建模单元，以及“空格符”与“空白符”，输

出节点数目如表2所示。wav2vec 2.0网络结构中,特征编码器包含7个结构相同的子块,每个子块内的时间卷积层的通道数为512,步长分别为(5,2,2,2,2,2,2),卷积核尺寸分别为(10,3,3,3,3,2,2)。上下文编码器包含24个Transformer层,表征维度D为1 024,内部维度D'为4 096,注意力头数h为16,Dropout概率为0.1。用于建模相对位置嵌入的卷积层的卷积核大小为128,组数为16。采用Adam优化器训练模型,学习率分3个阶段调整,在前10%个更新进行预热,从3e−7线性增加至3e−5,在接下来的40%个更新保持学习率恒定,在剩余的50%个更新使学习率线性降低至1.5e−6。在前10 000个更新,只训练网络顶部的分类器网络,在10 000个更新之后才对上下文编码器进行训练。特征编码器f的参数始终固定。每块显卡的一个批次包含1.28 m输入样本点数,在16 kHz的采样率下相当于80 s的音频。共使用4张TI TAN RTX显卡进行训练。

表2 不同建模单元对应的输出节点数目

Table 2 The number of output nodes corresponding to different modeling units

项目	建模单元	空格符	空白符	输出节点
字符	93	1	1	95
单音素	357	1	1	358
含位置信息的音素	694	1	1	696

3.2.2 语言模型与解码

本文采用KenLM[21]实现3-gram语言模型,该模型由训练集的标注构建而来,并且基于IARPA BABEL语料库自身提供的词汇表。具体而言,将标注文本按表1的方式,转换为不同形式的建模单元序列,并分别构建语言模型。语音解码时,式(6)中浅融合参数γ设为1.0。训练越南语ASR系统流程如算法1所示。

算法1: 训练越南语ASR系统

Input: 原始音频与标注文本对$(\boldsymbol{X}, \boldsymbol{Y})$,特征编码器$f$,上下文编码器$g$,分类器网络$h$

将原始音频进行升采样,按2.4节描述步骤将标注文本的单词序列转换为对应的建模单元(字符、音素或子词单元)序列

for $i \in \{1, \cdots, M\}$

for 一批训练样本$\{\boldsymbol{X}_k, \boldsymbol{Y}_k\}_1^N$ do

$\hat{\boldsymbol{Y}}_k = h\left(g\left(f(\boldsymbol{X}_k)\right)\right)$

计算CTC代价函数$J_{\mathrm{CTC}}\left(\boldsymbol{Y}_k, \hat{\boldsymbol{Y}}_k\right)$

end

更新网络g与h,最小化CTC代价函数$J = \sum_{k-1}^{N} J_{\mathrm{CTC}}\left(\boldsymbol{Y}_k, \hat{\boldsymbol{Y}}_k\right)$,即$\boldsymbol{\theta} = \mathrm{Optim}(\boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} J)$

end

Return 越南语微调后的网络g与h

3.3 实验结果

3.3.1 不同基础建模单元实验

本文在wav2vec 2.0模型基础上,分别将越南语的字符、单音素与含位置信息的音素作为CTC模型建模单元。其中,字符单元为本文的基线系统。其他系统在声学模型结构上,除了分类器网络拥有不同的输出节点之外,其他都与3.2.1节描述的基线系统一致。词错误率结果如表3所示。从表3可以看出,单音素和含位置信息音素的单元错误率在词错误率上的降低十分显著,分别相对降低约11.8%和

11.0%。语音识别本质上是一种把发音映射到文字的过程，采用发音词典得到的音素信息即使对于CTC这种端到端的建模方式仍然是有效的，能够比仅仅采用字母的系统有明显的提升。另一方面，简单地通过位置信息扩充建模单元，不仅对性能没有帮助，反而有稍微的下降(WER由32.9%降为33.2%)。

表3 不同建模单元的词错误率结果

Table 3 Results of WER under different modeling units

建模单元	WER/%
字符(基线)	37.3
单音素	32.9
含位置信息的音素	33.2

进一步地，本文对3种建模单元分别使用BPE算法得到相应的子词单元，BPE操作数分别为1 000与1 500。词错误率结果如表4所示，表中第2列也就是表3的结果，代表未使用BPE算法的词错误率。对比表格的第2列与第3列结果可见，BPE算法的引入能够大幅降低词错误率，并且对字符建模单元最明显，相对提升16.9%左右。纵向对比表格的结果可知，在相同BPE操作数的前提下，音素子词的词错误率显著低于字符子词，并且基于单音素的识别率高于含位置信息的音素。相比字符为建模单元的基线系统(37.3%的WER)，表中最好的音素子词系统(29.4%的WER)取得了21.2%的相对提升。另外，本文对语音识别系统的另一个重要性能指标——实时率(Real time factor，RTF)进行了对比，采用单张TITAN RTX显卡进行测试，结果如表4所示。从表4可以看出，使用BPE子词作为建模单元将大大增加解码所需时间，损害系统的实时性。比如BPE1500的RTF比单字符的RTF相对增加约60.1%。此外，基于音素单元的系统实时性逊于基于字符的系统。综合表4中WER结果与RTF结果不难发现，本文的改进系统在取得更低错误率的同时损害了解码实时性，更适合对解码时延不敏感的场景。

表4 不同子词单元的词错误率与实时率结果

Table 4 Results of WER and RTF under different sub-word units

建模单元	WER/%			RTF		
	未引入BPE	BPE操作数为1 000	BPE操作数为1 500	未引入BPE	BPE操作数为1 000	BPE操作数为1 500
字符	37.3	31.0	30.4	0.011 48	0.018 43	0.018 38
单音素	32.9	29.9	29.4	0.012 98	0.019 82	0.019 93
含位置信息的音素	33.2	30.4	29.9	0.012 14	0.019 99	0.020 34

3.3.2 对比其他系统

最后对本文识别系统与其他研究者近期提出的越南语识别系统进行了性能对比，所选系统均基于BABEL语料库进行搭建，并且均在BABEL开发集上进行测试，如表5所示。其中，Conneau等[8]使用与本文相同的预训练模型进行微调，与之相比，本文的ASR系统在字符错误率(Character error rate，CER)上相对提升约6.0%。而Yi等[22]采用语言对抗性迁移学习的策略，对比该方法，本文在WER上取得了约35.7%的相对性能提升。

表5 对比其他越南语识别系统

Table 5 Comparison with other Vietnamese recognition systems

声学模型	语言模型	CER/%	WER/%
XLSR-53(Conneau等[8])	4-gram	21.8	
SHL-Model(Yi等[22])	3-gram		45.7
本文模型	3-gram	20.5	29.4

4 结束语

本文研究了基于wav2vec 2.0的语音识别技术,并在低资源越南语数据集上搭建了完整的语音识别系统。本文通过将声学信息纳入模型,大幅提升了模型的识别率。以单音素和含位置信息的音素作为输出节点系统的词错误率都显著低于以字符作为输出节点的系统,相对提升分别达到11.8%和11.0%。此外,本文基于BPE算法构造音素子词单元,考虑前后音素的影响,相比基线系统获得了进一步的提升。下一步的研究方向是如何将不同建模单元的信息进行更有效的融合。

参考文献:

[1] 刘加, 张卫强.低资源语音识别若干关键技术研究进展[J].数据采集与处理, 2017, 32(2): 205-220.
LIU Jia, ZHANG Weiqiang. Research progress on key technologies of low resource speech recognition[J]. Journal of Data Acqusition and Processing, 2017, 32(2): 205-220.

[2] SALAKHUTDINOV R, HINTON G. A better way to pretrain deep Boltzmann machines[C]//Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates, 2012: 2447-2455.

[3] CHUNG Y, LEE H, GLASS J. Supervised and unsupervised transfer learning for question answering[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana: Association for Computational Linguistics, 2018: 1585-1594.

[4] ZHOU Shiyu, XU Shuang, XU Bo. Multilingual end-to-end speech recognition with a single transformer on low-resource languages[EB/OL]. (2018-06-12)[2021-07-27]. https://arxiv.org/pdf/1806.05059.

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

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

[7] JIANG Dongwei, LEI Xixoaning, LI Wubo, et al. Improving transformer-based speech recognition using unsupervised pre_training[EB/OL].(2019-10-22)[2021-07-27]. https://arxiv.org/pdf/1910.09932.

[8] CONNEAU A, BAEVSKI A, COLLOBERT R, et al. Unsupervised cross-lingual representation learning for speech recognition[EB/OL].(2020-06-24)[2021-07-27]. https://arxiv.org/pdf/2006.13979.

[9] LIU Hairong, ZHU Zhenyao, LI Xiangang, et al. Gram-CTC: Automatic unit selection and target decomposition for sequence labelling[C]// Proceedings of the 34th International Conference on Machine Learning. Cambridge, MA, USA: PMLR, 2017: 2188-2197.

[10] ZEYER A, IRIE K, SCHLÜTER R, et al. Improved training of end-to-end attention models for speech recognition[EB/OL]. (2018-05-08)[2021-07-27]. https://arxiv.org/pdf/1805.03294.

[11] ZWEIG G, YU C, DROPPO J, et al. Advances in all-neural speech recognition[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017: 4805-4809.

[12] XU Hainan, DING Shuoyang, WATANABE S. Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling[EB/OL].(2018-11-10)[2021-07-27]. https://arxiv.org/pdf/1811.04284v2.

[13] HE Yanzhang, BAUMANN P, FANG Hao, et al. Using pronunciation-based morphological subword units to improve OOV handling in keyword search[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(1): 79-92.

[14] WANG Weiran, WANG Guangsen, BHATNAGAR A, et al. An investigation of phone-based subword units for end-to-end speech recognition[EB/OL]. (2020-04-08)[2021-07-27]. https://arxiv.org/pdf/2004.04290.

[15] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[EB/OL]. (2020-06-20)[2021-07-27]. https://arxiv.org/pdf/2006.11477.

[16] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[EB/OL]. (2016-06-27)[2021-07-27]. https://arxiv.org/

pdf/1606.08415v4.

[17] GRAVES A, FERNANDEZ 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. Pittsburgh, USA: Association for Computing Machinery, 2006: 369-376.

[18] CHO J, WATANABE S, HORI T, et al. Language model integration based on memory control for sequence to sequence speech recognition[C]// Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 6191-6195.

[19] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. (2015-08-31) [2021-07-27]. https://arxiv.org/pdf/1508.07909v5.

[20] OTT M, EDUNOV S, BAEVSKI A, et al. fairseq: A fast, extensible toolkit for sequence modeling[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 48-53.

[21] HEAFIELD K. KenLM: Faster and smaller language model queries[C]//Proceedings of the Sixth Workshop on Statistical Machine Translation. Edinburgh, Scotland: Association for Computational Linguistics, 2011: 187-197.

[22] YI Jiangyan, TAO Jianhua, WEN Zhengqi, et al. Language-adversarial transfer learning for low-resource speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 621-630.

作者简介:

沈之杰(1997-),男,硕士研究生,研究方向:语音识别,E-mail:zjshen20@mail.ustc.edu.cn。

郭武(1973-),**通信作者**,男,博士,副教授,研究方向:语音、语言信号处理,E-mail:guowu@ustc.edu.cn。

(编辑:张黄群)