语音识别技术及多语言语音识别发展综述

课　　程：　语音交互

学　　院：　人工智能学院

班　　级：　1603 班

姓　　名：　孟令辉

学　　号：　201918014628045

2020 年　06 月 12 日

目录

语音识别技术及多语言语音识别发展综述

摘要

摘要：随着语音识别系统的发展，多种框架被相继提出，用来提升传统识别系统的速度和准确率。但在很长一段时间里，语音识别系统效果的提升很大程度受限于资源受限的语言，搭建语音识别系统受到很大限制。这其中的标注工作带来大量的工作量以及人力物力成本。而随着 ASR 系统的发展，语音识别从传统的 GMM-HMM 混合模型建模，到 CTC-DNN 框架，再到现如今非常流行的端到端框架 Attention 等机制都带来了系统性能或架构上的优化提升，并且已经实现了较好的性能。本文首先介绍了自动语音识别的发展历程以及其中用到的主要架构，随后在低资源下的多语言语音识别技术的发展概况，并对未来做出展望与规划。
关键词：语音识别系统、多语言语音识别、深度学习

引言

经过了多次技术的发展革新，语音识别技术已经经历了上个世纪的仅从声音的波形特征识别个别字符、利用模板匹配的方法识别特定说话人的语音、结合隐马尔可夫模型（HMM）的大语量连续语音识别系统、用 n-gram 的方式训练语言模型加入的识别系统、结合高斯混合模型（GMM）对语音观察概率序列建模并由 HMM 对语音时序建模进行识别等多种经典的语音识别框架。而随着 Hinton 等人的工作，将原先的感知机逐步摆脱离散函数的约束后并继续解决梯度消失、局部最优解等问题，全面掀起了深度学习的发展浪潮。结合深度神经网络（DNN）和 HMM 的语音识别框架[1]，将声学模型加入到网络当中，并利用统计的方法进行语音识别变得流行。在这过程中的 DNN 出现了很多种形式被应用其中，包括可以很好地学习历史信息的基于语音时序特性建模的循环神经网络（RNN）和长短期记忆网络（LSTM）[2]，之后再融合 HMM 完成语音识别。由于 RNN 等神经网络的结构特性，已经能够很好地学习历史信息作为辅助完成任务。

上述的神经网络部分的难度在于对输入时序信号帧级别的对齐，比如前馈神经网络中要通过左右拼帧的方式对时序信号长时相关性建模。由 Graves 等人在 2006 年提出的 Connectionist temporal classification（CTC）[3]的方式训练网络能够很好地解决之前网络的边界判断问题，其利用前后向算法自动学习语音特征的

模型边界，解决了标注序列与特征序列长度不同的问题，突破之前对齐操作费时长、精度低的瓶颈，这种模型可以直接与 LSTM 等网络结合适用于端到端语音识别系统。这种端到端的方式已经在工业界逐渐成熟，对多数的语音识别任务如大词汇量语音识别（LVSR）[4]、关键词识别（KWS）都有很好的表现，并成为全球各大语音实验室的主要研究方向，所以本次对基于这种框架下的语音识别系统的优化是具有十分积极的意义的。而加权有限状态转换器（WFST）[5]可以事先离线结合发音词典，或者利用 n-gram 训练的语言模型，生成对于计算机方便搜索和优化的搜索图[6]，对于解码网络是十分必要的。传统的语音识别框架一般是分别训练声学模型和语言模型作为识别效果的优化方法，利用 n-gram 方法生成的语言模型的 WFST 搜索图[7][8]，再利用前缀束搜索的方法逐帧解码，通过对构建搜索图和实现解码的方法的优化实现本次研究最终识别系统，提升语音识别系统性能。

语音识别发展历程

作为传递信息的媒介之一，语音在人们的生活中无处不在。它出现在人们的信息交流当中，有着多种多样的用途。从古时刀耕火种时代的各类语言就已经开始对人类生存发展起着至关重要的作用，再到现代社会中各类工作场所、休闲娱乐场所也都有着语音用来传递信息的重要应用。而随着当代的计算机技术、模式识别技术、信号处理技术的不断进步，语言学和社会学的不断发展，语音识别自然而然就成为了顺应时代的需求。

自动语音识别（ASR）系统的发展经历了漫长的过程，随着数学理论的发展、信号处理的成熟以及计算机技术突飞猛进的发展，出现了多种多样的技术，并且已经发展得越来越成熟。国内外众多学者积极投身与语音识别系统的开发及优化，并且由于编程语言的便利和计算机的算力的提升，ASR 系统的性能已经得到了大幅提升，甚至在一些常见的语音识别任务中超越人类的表现，这样的结果是十分令人激动的。虽然在 ASR 系统的发展过程中，研究者们提出了多种不同的算法来提升其识别的速率和准确率，但其大体框架依然是由后续描述的结构组成，只是各个模块的实现方式各不相同。

语音识别在实现的早期是在上个世纪五十年代美国 AT&T 的贝尔实验室，利用共振峰的特性用信号处理的方式，识别出有限的是个英文数字。之后的六七十年代出现了动态规划和线性预测分析两种技术，对语音识别的影响是巨大的，动态规划很好地将语音信号的特征提取问题解决，为之后的识别系统的发展打下基础。还有针对小词汇表的孤立词识别，利用模板匹配技术将孤立词作为模板。此

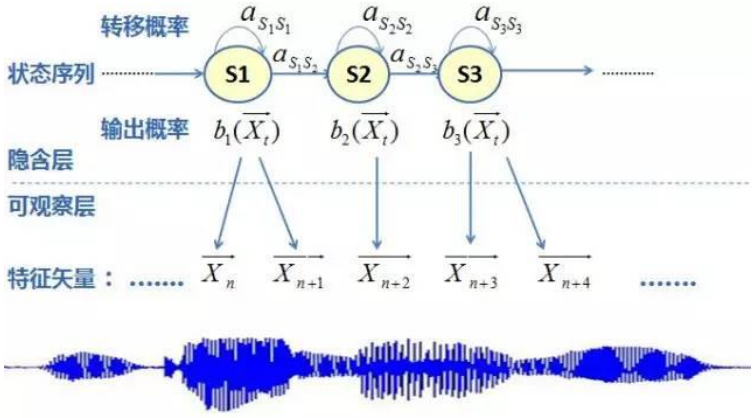
时的线性预测编码技术（Linear Predict Coding， LPC）被应用在语音识别中。随后的八十年代出现了对统治语音识别领域几十年的隐马尔可夫模型（HMM），其由贝尔实验室的 Rabiner 等研究人员提出，基于统计模型的方法对语音的时序进行建模。同时期由 n-gram 的方法，对识别系统加入语言模型的辅助，使系统结合语言学特征提升系统准确率。1988 年的非特定人连续语音识别系统 SPHINX 系统就是利用了结合语言模型和 HMM 的系统。20 世纪 90 年代到本世纪初，高斯混合模型（GMM）-HMM 框架成为主流技术，由 GMM 对语音的观察概率进行建模，配合 HMM 对状态序列建模，大幅提升了训练的速度，声学模型较小容易移植应用。再到 2006 年 Hinton 等人用预训练的方法缓解了局部最优的问题，将神经网络从最初的感知机真正变成了深度神经网络（DNN），使得深度学习变得十分火热。微软邓力和俞栋提出了 DNN-HMM，用 DNN 代替之前的 GMM 对观察序列建模输出概率。这种方法的优点是能够利用帧的上下文信息，学习到之前的历史信息对识别性能提升。在此之后 RNN、LSTM、CNN 都被广泛应用于语音识别当中。几年前由多伦多大学的 Alex Graves 又将 CTC（connectionist temporal classification）目标函数带入到语音识别当中，用来训练神经网络生成最后的概率分布，这种方法不必进行特意的对齐操作，通过前后向算法自动学习边界。现在结合 CTC 和解码网络的语音识别系统已经十分流行[9]，全球研究者们也都不断探索这种架构下的优化[10]，之后的加权有限状态转换器的引入，加入了语言模型提升识别的准确率和解码速度[11]。还有基于编码解码器（encoder-decoder）的神经网络，将输入输出在处理过程中以向量表示，这得益于 attention 机制的提出，其嵌入矩阵、残差连接和位置编码的机制将神经网络的梯度消失等问题进一步优化，又开辟了语音识别发展的新方向。

GMM-HMM 基本结构

马尔可夫链（Markov chain）是一种用来表示状态转移的链式结构，由俄罗斯数学家安德烈.马尔可夫得名，与确定性转移不同的是其状态的转移是具有随机性的，状态转移的边上有一定概率，这些状态构成的集合成为状态空间。当前状态出现的概率值与上一状态有关的马尔可夫模型成为一阶马尔可夫模型。每条代表状态转移的弧上的概率成为状态转移概率，这些概率构成状态转移矩阵，当一阶模型具备 N 个状态时，状态转移矩阵就有 N^2 个元素。一阶模型的性质如公式（2-1）：

$$P(X_i|X_{i-1} \dots X_1) = P(X_i|X_{i-1}) \quad \text{(2-1)}$$

式中 X_1、X_2、X_3….X_n 是随机变量。而隐马尔可夫模型是一种统计模型，用来描述一个含有未知参数的马尔可夫过程，该模型以马尔可夫的形式构建两种类别的状态转移图，一种是可观测状态，另一种为隐藏状态。利用观测到的状态来确定隐藏状态的序列，如自然语言处理中的词性标注：句子中的单词是可观察的，而词性是隐藏的状态，通过上下文信息找到最有可能的隐藏状态序列，得到该单词的词性。



DNN 和 HMM 框架构成图

这种模型被用来构建语音识别系统，配合高斯混合模型训练声学模型。模型需要通过观察序列 O 来得到最优序列组合 W，O 是语音中特征提取得到的参数，一般为 MFCC 倒谱系数，目的是求使得条件概率 P(O|W)最大的 W。HMM 在语音识别中的基本链式结构如图，针对语音当中不同的特征向量 HMM 会输出对应概率。对于 HMM 这种链式结构是通过状态之间的转移和状态时间的相关性来体现语音的时序线性序列特性的。图中 HMM 输出对应的每个特征矢量的概率是由下面的高斯混合模型（GMM）建模公式计算：

$$b(X) = \sum_{k=1}^{k} p(k)p(x|k) = \sum_{k=1}^{k} \pi_k N(X|u_k, \sigma_k^2) \text{.....(2-2)}$$

式中X——提取的特征矢量；k——是高斯个数；

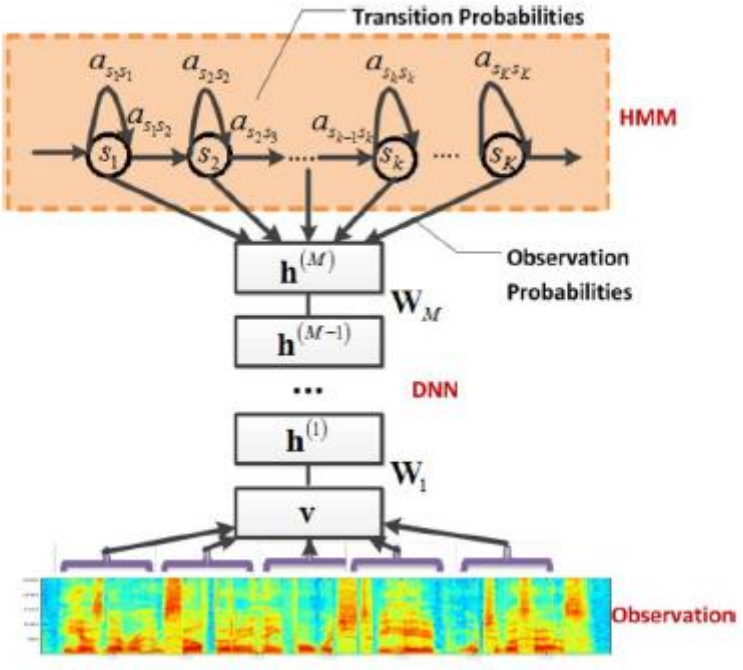
u——均值；

σ——方差

HMM 最基本的音素单元，实际使用是三音素的建模模型，这样可以使模型学习上下文信息，声学模型由从语音中提取的特征矢量表示，经模型训练的得到输出概率，通过解码得到最后的结果。GMM-HMM 构建过程复杂，且构建解码器的过程需要大量技巧，但模型解码速度快方便部署。

DNN-HMM 基本结构

在经历了 GMM-HMM 框架统治了二十余年后，深度学习被研究人员们引入语音识别，代替 GMM 的作用。实际仍然是通过 HMM 作为状态转移表示音素的时间序列特性。



DNN-HMM 基本结构

图中表示该模型的框架，由语音做特征提取输出的特征向量输入到 DNN 的输入层中作为训练样本经过隐层，最后通过 SoftMax 这种分类器作为输出，学习深层非线性特征变换。往往需要先用 GMM-HMM 训练得到输入特征标签和语音的对应关系，进行对齐工作。在这过程中深度神经网络作为判别模型给出的是后验概率，需要经过贝叶斯公式转换成 HMM 的输出的形式。通过最大似然的方式最优化 HMM 网络的参数，HMM 相当于生成模型，生成根据状态序列的概率和给定观测值之后的条件概率，而 DNN 相当于判别模型输入是特征向量，输出是对应各个状态的概率即观测概率。这个过程中的声学模型是根据语音的特征提取体现的，好的特征提取方法会给识别效果带来十分积极的作用。而语言模型仍然是由音素模型来体现。加入 DNN 避免了原来 GMM 无法表述复杂函数的缺点，根据训练样本和模型参数，得到最好的 HMM 网络参数，从而得到在 HMM 中由观测状态得到的最有可能出现的状态序列。

DNN+CTC 声学模型

这种模型与传统的带有 HMM 框架的方法不同，之前的模型训练时都需要将

训练数据的每一帧进行标注，对 DNN 的输出进行标号的操作，而且对语音需要按帧甚至按照音素裁剪或对齐。而 DNN+CTC 的方法的巧妙之处在于，按照 CTC 定义下的目标函数和 blank 概念的引入，加上前后向算法的结合使得模型可以自动学习语音特征的边界，解决了标注序列与特征序列长度不相等的问题，并且 blank 还可以吸收发音单元的混淆性，提升了系统的性能。下面将介绍本次研究框架下用到的神经网络 LSTM 和 CTC 的基本结构。应用于 RNN-T[12]该模型虽然缺少并行编码优点且解码速度慢，但联合 CTC 可有效辅助对齐效果。最新研究也应用了 CTC loss 的方式进行模型训练以辅助对齐[13][14]。

长短期记忆模型（LSTM）

循环神经网络（RNN）虽然是依据语音的时序性序列增强了长时建模的能力，但实现过程中仍然会出现深层次网络的梯度消失或梯度爆炸的问题，对于后面的节点，前面的节点感知力下降，使得深度没法加大。所以 LSTM 应运而生利用其自身的特殊结构，引入 Cell 解决了梯度的问题，同样学习序列的历史信息实现对网络的训练，生成最后的概率分布序列。LSTM 是一种学习长期依赖信息的网络也是改进后的 RNN。下面将从其结构和前后向公式的推导，以及权重更新方式等方面来介绍每层仅有一个 block 且每个 block 只有一个 cell 的 LSTM。

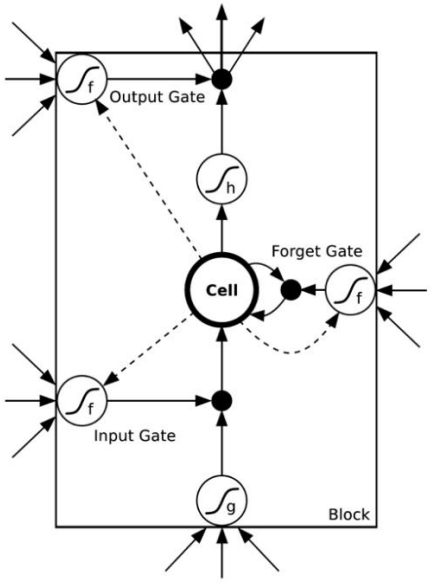
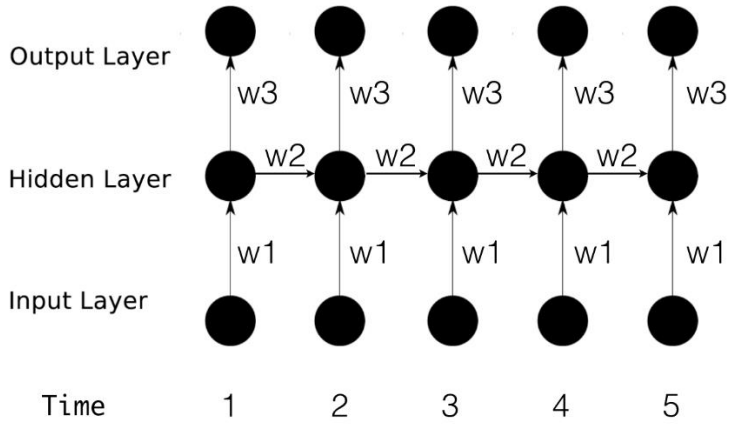


图 LSTM 中 Cell 的基本结构

为了解决感知力下降问题，LSTM 引入 Cell 的概念。LSTM 与传统 RNN 的大体结构是类似的，如图。但在每个时刻的隐层中存在多个 block，而每个 block 存在多个 Cell 来进行记忆存储。图中表示了 LSTM 最基本的单元 Cell，它不但受当前时刻的输出 Cell 的输出影响还受上一时刻的 Cell 的输出的影响。它包含

输入门、输出门和遗忘门，并且有着不同的门函数。



传统 RNN 示意图

针对各个门的参数变量表示，其前传公式如下：

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{c=1}^{C} w_{Cl} s_c^{t-1} \quad \text{(2-3)}$$

$$b_l^t = f(a_l^t) \quad \text{(2-4)}$$

式中a_l^t——输入门的输入；

b_l^t——输入门的输出。

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1} \quad \text{(2-5)}$$

$$b_\phi^t = f\left(a_\emptyset^t\right) \quad \text{(2-6)}$$

式中a_ϕ^t——遗忘门的输入；

b_ϕ^t——遗忘门的输出。

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t \quad \text{(2-7)}$$

$$s_c^t = b_\emptyset^t s_c^{t-1} + b_l^t g(a_c^t) \quad \text{(2-8)}$$

式中a_c^t——Cell 的输入；

s_c^t——Cell 的输出。

$$a_w^t = \sum_{i=1}^{I} w_{iw} x_i^t + \sum_{c=1}^{C} w_{cw} s_c^t \quad \text{(2-9)}$$

$$b_w^t = f(a_w^t) \quad \text{(2-10)}$$

式中a_w^t——输出门的输入；

b_w^t——输出门的输出。

$$b_c^t = b_w^t h(s_c^t) \quad \text{(2-11)}$$

式中b_c^t——最终的输出。

根据前传公式，LSTM 会利用求损失函数的梯度的方式对各个参数进行更新从而进行反向传播。损失函数如下，将输出真实值 z 与输入 x 求条件概率，用损

失函数对上述各参数分别求导，利用梯度下降的方法进行更新。

$$L(x,z) = -\ln P(z|x) \text{.....(2-12)}$$

求得损失函数$L(x,z)$对前向传播过程中的各参数的梯度，利用如下公式更新权重

$$\Delta w^n = m\Delta w^{n-1} - \alpha\frac{\partial L}{\partial w}n \text{.....(2-13)}$$

式中$m\Delta w^{n-1}$为上一次权重的更新，m∈[0, 1]

Connectionist Temporal Classification（CTC）

CTC 由于引入 blank 的标签项解决了标注序列与特征序列长度不相等的问题，直接与 LSTM 结合可以直接实现端到端的语音识别系统，无需标注输入输出序列之间的映射关系，避免了对齐操作。CTC 的出现颠覆了之前的 HMM 的统治地位。CTC 假设每个时刻输出的概率分布是相互独立的，这样在计算时就可以直接相乘。

当输入为长度为 T 的序列 x、输出为序列 y 时，在 t 时刻的第 k 个输出单元的概率为y_k^t，L 作为输出序列的集合，则输出序列为π时的概率有如下公式：

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t, \forall\pi \in L' \text{.....(2-14)}$$

式中L'——集合 L 加入 blank 后的集合。

设定多对一函数β，将不同输入做去重和去掉 blank 的操作，如β（a--bc）=β（abbc）=β（abc）。对于多条路径都对应于同一输出结果的概率表示为如下公式：

$$p(l|x) = \sum_{\pi\in B_{(l)}^{-1}} p(\pi|x) \text{.....(2-15)}$$

式中 l——给定的标签。

最终就是要得到使得上述公式概率最大的输出标签 l。通过选择 blank 为分割点，分别识别各个分区的 label 最后再将其串联得到最终结果，这样分割的方法有利于进行迭代过程逐步识别和解码。在训练网络阶段，定义目标函数的过程中，需要利用前后向算法，但为了满足原有的 blank 的存在还需要重新在每对 label 之间和序列的开始和末尾加入 blank，之后确定前后向变量表示条件概率。在定义变量时由于考虑到解码方法选择，当解码方式为前缀搜索时，需要区分末位标签是否为 blank 的情况。此处就相当于在训练 CTC 网络时加入了声学模型。为了防止溢出需要将前后向变量分别用求和项作为因子约束。最后表示当前输入情况下的输入的概率，公式（2-16）就是利用极大似然的方式定义最终的目标函数，训练网络参数。

$$O^{ML}(S,N_w) = -\sum_{(x,z)\in S} ln\big(p(z|x)\big) \quad (2\text{-}16)$$

式中 S——输入输出序列的集合；

N_w——整个神经网络，权重集合为 w。

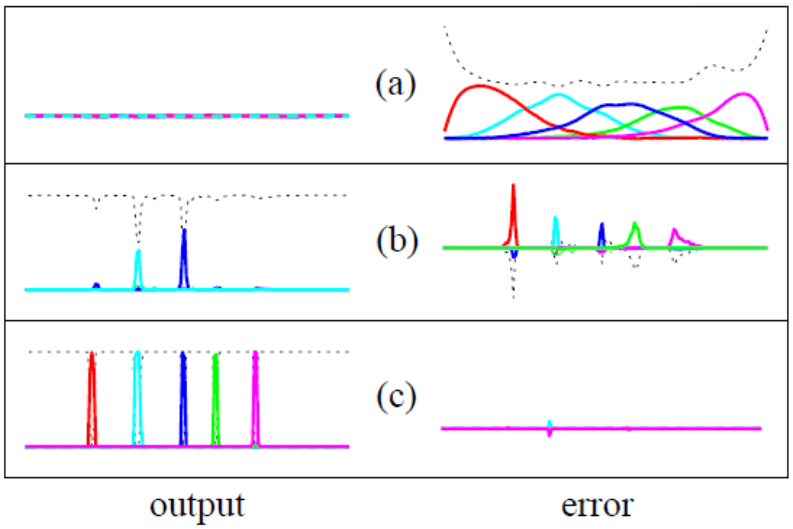
由于极大似然法特性，概率用 ln 函数表示，且最终训练目的是将目标函数的值尽可能地降低，得到最好的输出序列。若要用梯度更新的方式训练网络，则需计算目标函数的梯度以及其更新方式，这时要联合定义的前后向变量表示。由于最终要做分类的是网络的输出，即训练的输出标签y_k^t，所以需要计算目标函数对输出的梯度。式（2-17）为目标函数对输出的梯度表示，式（2-18）中表明了前后向变量替换条件概率在式（2-19）中的位置。

$$\frac{\partial O^{ML}(S,N_w)}{\partial y_k^t} = -\frac{\partial\, ln\big(p(z|x)\big)}{\partial y_k^t} \quad (2\text{-}17)$$

$$\frac{\partial p_{(l|x)}}{\partial y_k^t} = \frac{1}{y_k^{t^2}}\sum_s \alpha_t(s)\beta(s) \quad (2\text{-}18)$$

$$\frac{\partial\, ln\big(p(l|x)\big)}{\partial y_k^t} = \frac{1}{p(l|x)}\frac{\partial p(l|x)}{\partial y_k^t} \quad (2\text{-}19)$$

得到最后的梯度表示如公式（2-19），用来进行梯度下降的梯度更新，图 2.5 表示了网络训练的过程中输出和错误率的变化，最终得到训练好的网络。



CTC 网络训练进化过程

基于注意力机制的模型结构

模型包括三部分：编码器、解码器和 attention 模块。以下对这三部分分别进行详细介绍。

编码器模块（Encoder）:编码器采用深层神经网络将输入序列编码成一个隐层序列。该隐层序列是输入序列的抽象表达，方便与解码器协调运算。编码器

通常采用 Bi-RNN 或卷积神经网络。在语音识别中由于输入语音远大于输出文本序列，并且语音特征序列存在较大冗余，即编码器端一般会采用金字塔状的结构，即神经网络下面宽上面窄，一般采用 CNN 等网络分层进行逐步压缩，这样可以使编解码器对其更同意，下图中编码器端由两层 RNN 组成，并且第二层比第一层窄。编码器可以看成是对语音特征序列的声学模型构建，抽象表达为公式：

$$h = encoder(x)$$

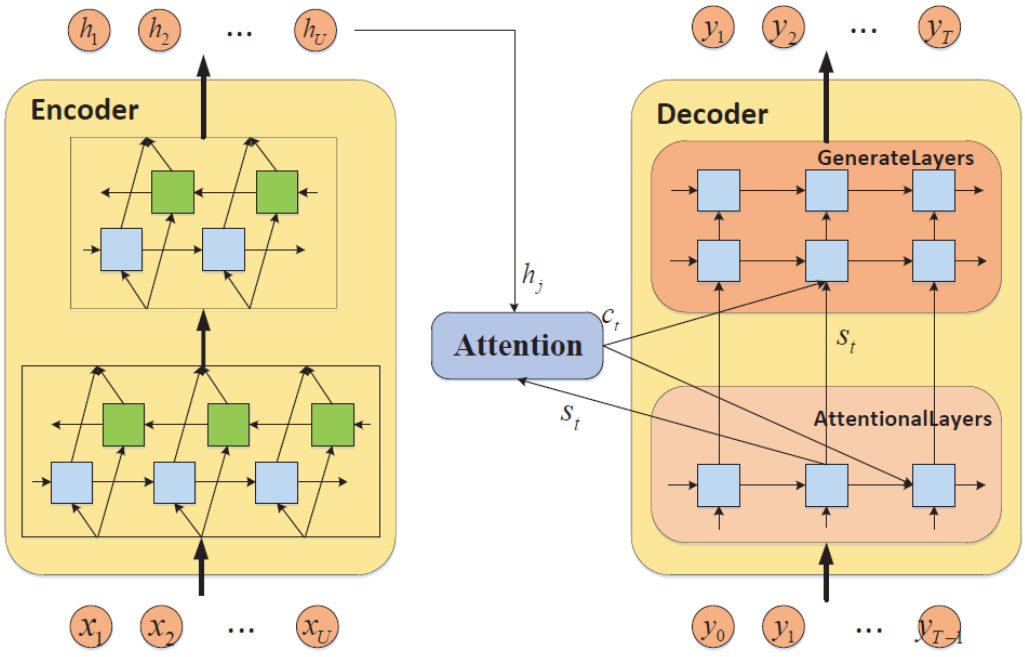
解码器模块（Decoder）：解码器采用自回归的方式逐步检测输出序列，在每一步预测时，解码器的输入来自两部分：一部分时上一步的预测结果 y_{t-1}；另一部分是通过扫描编码器的隐层序列，根据子注意力机制给出置信度得分，选择关注相应位置生成的注意力向量 c_t。解码器从功能上可以分为两个子模块：注意力模块和生成模块。注意力模块生成解码器端的隐状态 s_t，用于参与注意力模块的计算和生成模块的计算。生成模块将注意力模块生成的隐状态和注意力曾得到的注意力向量 c_t 作为输入预测当前步的输出 y_t。公式可以表示如下：

$$s_t = AttentionLayers(y_{t-1}, s_{t-1}, c_{t-1})$$

$$y_t = GenerateLayers(s_t, c_t)$$

解码器相当于在升学特征序列条件下的语言模型，他根据编码段的隐层蓄力通过自回归的方式扩展输出单元序列。

注意力模块（Attention）：注意力模块将解码器中注意力模块的隐状态和



隐层序列的每个元素打分，这个归一化后的得分可视为每个元素被解码器当前步

选择的概率，它表征了输入序列和输出序列的对齐关系。最后使用归一化得分对所有元素线性求和得到注意列向量。公式如下：

$$p_{t,j} = soft\max(score(s_t, h_j)),\ for\ \ j = 1, 2..., U$$

$$c_t = \sum_{j=1}^{U} p_{t,j} h_j$$

得分函数经常使用的是 scale dot-product 作为打分函数。2017 年 Chiu 等采用子词建模单元 WPM 在 12500 小时的英文数据集上取得了 sota 模型。现在用在语音识别上的基于注意力机制的端到端模型的建模单元进行过对比研究，包括字母、上下文无关音素、上下文相关因素以及子词建模单元此片段模型。实验结果显示在中文连续语音识别任务上采用汉字建模单元目前可以取得最好性能。基于 attention 机制的语音识别系统有较好的并行编码特性，但仍需要整句编码，不利于流式语音识别的发展。

本章小结

本章介绍了多种语音识别框架，它们均有自己的优势和不足，但都在实际应用中大放光彩，尤其是 GMM-HMM 框架流行多年。但在 CTC 这种将语音对齐操作去除的框架被提出后，与 LSTM 和解码器结合的语音识别框架有着极大的发展潜力，到如今已经发展的较为成熟。最近又部分基于端到端的语音识别，包括简化 transformer 模型[15]、融入卷积网络[16]等都取得了较好的效果。但在可解释性、流式识别、低资源等领域仍面临较大挑战。

多语言语音识别技术研究现状

发展历程概述

多语言语音识别技术是近几年发展起来的新的方向，也是语音识别中比较具有挑战性的工作。其主要包括两类问题，即不同语言同意系统识别整个序列的小语种语音识别问题，和语码转换问题即在源语言插入目标语言。由于自动语音识别系统在发展初期的很长一段时间都在考虑算力和单语种准确度提升的问题，近几年才陆续有研究人员针对低资源下的多语言语音识别做系统性的归纳与研究。最初的研究形式是语种识别（Language Identification，LID），早在 1973 年德州仪器公司就开始了这项研究。1980 年研究人员开始使用 Markov 模型进行语种识别技术的研究。随着研究的深入，出现了各种算法如基于声学特征和多项式决策函数的算法、VQ 算法等。一直到二十世纪九十年代中期，多语言语音识别的研究重点都是语种识别。

之后的研究重点转移到多语言连续语音识别。出现了 IARPA 提出的“巴别塔计划”。最初的实现方式是将语种识别技术和语音识别技术相结合进行多语言语音识别。先识别出语种，然后调用对应语言的识别引擎进行识别，分别使用各自的语言模型和声学模型。优点在于语种识别结果正确时识别性能较好，但缺点时对语种识别精度高度依赖。之后 Imseng 等提出基于隐式 LID 的多语言语音识别系统，将多个单语言语音识别系统直接并联后形成多语言语音识别系统，各自使用独立的训练语料、音子集和发音词典进行构建，但由于每条语音都需要经过所有的单语言语音识别器计算量较大，且对于语码转换问题无能为力。针对这个问题 Wu 等提出了一种语音自动切分和判断语种的方法，利用这种方法可以包含两种以上的语言的语音识别系统。

另一种更通用的方法是利用通用音子集的多语言语音识别方法。该方法将多种语言当成一种新的语言对待，它的音子集是采用通用音子集，声学模型和语言模型采用多种语言的数据共同训练而成，所以该模型不需要对语音进行雨中识别以及语种切分。声学模型针对非特定语言的，也称为非特定语言语音识别（Language-independent Speech Recognition）[17]。该方法的难点在于通用音子集的构建，一种方法是将多种语言的音子集合并，不同语言的音子集不进行共享；另一种方法是采用语音学的知识将每种语言映射到国际音标集 IPA 上作为全局的音子集清单[18][19][20]；第三种方法是基于数据驱动的音子聚类方法。MIT 的

计算机科学实验室，JHU 的语言与语音处理中心等都在研究。基于通用音子集的多语言语音识别技术是目前主流的多语言语音识别的研究路线。

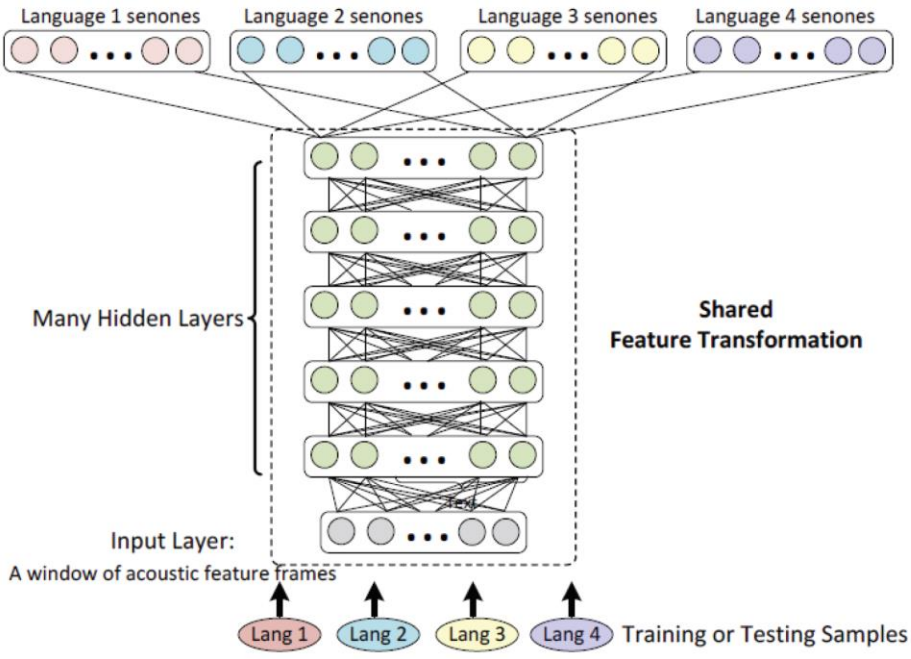
现阶段的端到端学习方法十分流行，2017 年文献[21][22]工作将语音识别中的端到端引入到多语言语音识别中。另外端到端模型可以采用字或者子词等与发音词典无关的建模单元，不需要复杂的通用音子集构建的相关流程，摆脱了发音词典的限制。

具体框架介绍

下面介绍三种主流的多语言语音识别框架：

共享隐层的多语言语音识别框架

2013 年微软的 Huang 等[23]提出共享隐层模型（Shared Hidden Layer Model, SHL-Model），它的输入层和隐层被所有语言所共享，但是输出层不进行共享，每种语言单独用用 softmax 层来估计绑定三音素状态的后验概率。文中采用前馈神经网络 DNN 作为共享隐层的网络结构，被称为共享隐层的多语言神经网络（Shared-Hidden-Layer Multilingual DNN, SHL-MDNN）。具体结构如下图。



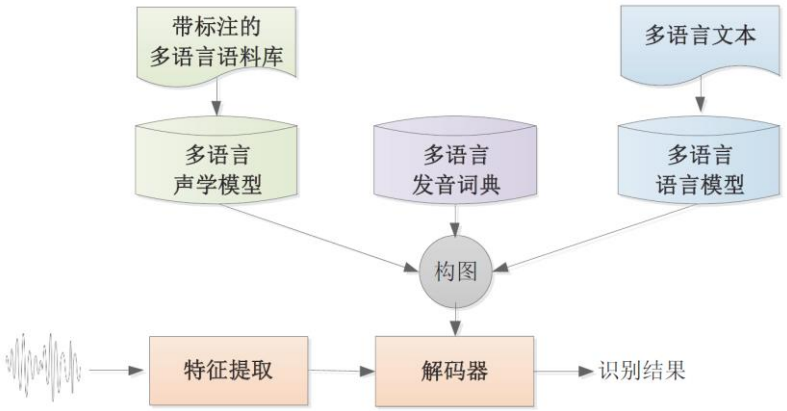
除输出层之外的网络可以堪称是一个对多语言共享的特征提取器。SHL-MDNN 机器训练过程可以看作是一种特殊的多任务学习[24]，等价于采用共享的特征表示来进行并行的多任务学习。相比于单语言的 DNN，SHL-MDNN 可以通

过寻找多语言任务的局部最优点，共享隐层在特征表达上更具有通用性，且多语言数据训练的共享隐层更加可靠，可以有效缓解训练数据不足导致的模型训练不充分和过拟合问题，并且有助于并行地学习特征有助于模型泛化，因为训练中包含了多个数据集的噪声等优点。使用时需要预先给定待识别的语音语种信息，这样准确地将语音送入到最终该语言对应的输出层解码。由此工作衍生出不同的多语言模型，如 SHL-MLSTM[25]、SHL-MTDNN-BLSTM[26]等。并且该模型在跨语言方面也取得了广泛的应用。Xu 等[27]将该方法与半监督学习相结合，从基于 DNN 的声学模型中迁移跨语言的知识到目标声学模型中，Li 等[28]采用该方法提取瓶颈特征进行跨语言知识迁移。

通用音子集的多语言语音识别框架

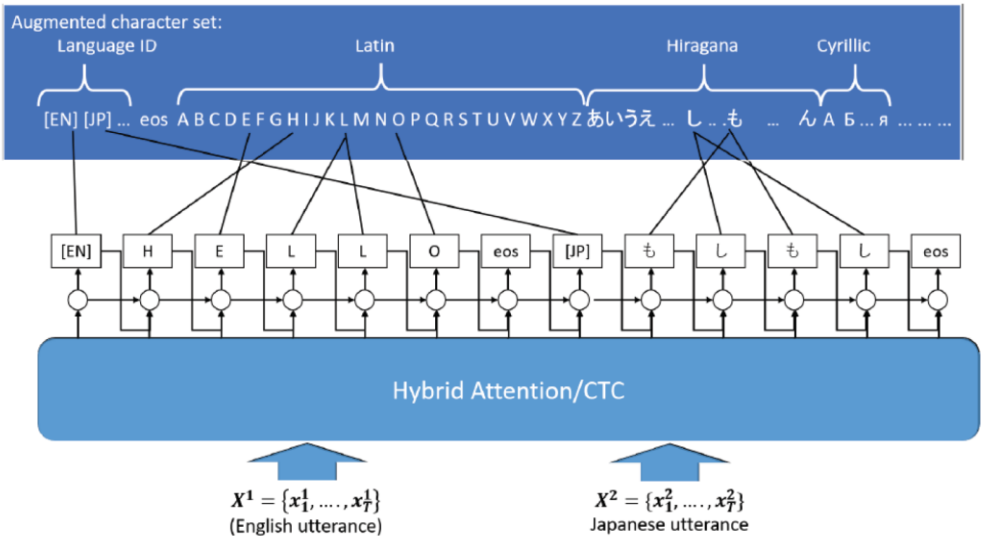
音子集的构建对于声学模型构建至关重要，其出发点是将多语言映射到统一的特征空间作为一种新的语言，它的音子集包含多种语言的通用音子集，声学模型和语言模型也都是用多种语言的数据共同训练得到的。该系统不会显示语种信息，而是直接语音识别，框架图如下。

构建音子集分为三种方法：一是直接将不同语言的音子集合并定义新的音子集，每种语言之间互相独立；二是基于知识驱动的方法，将不同语种映射达公共空间国际音标集 IPA[18]、SAMPA 和 Worldbet 等，优点是可以实现数据共享，并且通用音子集有比较明确的语音学定义，扩展性好；三是采用数据驱动的方法对多语种语言的音子集进行聚类生成通用的多语言音子集，分为自顶向下（决策树算法）和自底向上（根据相似性举例进行自动合并聚类，引入声学模型举例度量巴氏距离、马氏距离和 KL 距离等）两种算法。



基于端到端的多语言语音识别框架

这种模型时在端到端框架中产生的一种新的多语言语音识别框架。从整体框架上，基于端到端的多语言语音识别与特定语言的端到端语音识别几乎完全相同，不同之处在于 softmax 输出层将单语言的输出节点变成了多语言的输出节点。因此理论上前面介绍过的基于 CTC 的语音识别框架和基于注意力机制的语音识别框架都可以采用相同的方式迁移到多语言语音识别任务。2017 年 Watanabe 等[22]将语音识别中的端到端模型 CTC-Attention 应用到 10 种不同语言的多语言任务上取得了不错的效果。用字素作为建模单元，直接将多语言的字素合并后作为多语言模型的字素。为更好利用不同语言的语种信息，文中将语种标签增加到输出文本开头的方法进行训练。解码时模型先解码出语种标签 ，然后解码出后续文本，因此模型内部自动包含识别语种和语音识别部分。具体结构如下图



Toshniwal 等[21]采用基于注意力机制的 LAS 模型进行了类似的多语言语音识别的相关工作，并在 9 种不同的印度语言中进行验证，另外 Li 等[29]将 LAS 模型应用到 7 种英语方言任务上也取得了不错的效果。

相比共享隐层的多语言语音识别框架和通用音子集的多语言语音识别框架，基于端到端的多语言语音识别框架主要具有一下两点优势：首先可以采用音素等发音词典无关的的建模单元，因此不用以来发音词典，这样可以避免复杂的通用音子集构造的过程，这对于没有发音词典和发音词典获取比较困难的小语种非常适用；其次该模型将语音识别和语种识别统一符合端到端的特性，省去了语种识别和语种切分等预处理操作。

本章小结

本章对多语言语音识别任务的发展历程以及主要框架进行了介绍。多语

言语音识别系统逐渐从构建较为复杂的语种独立的专家知识到构建复杂的通用音子集，再到直接利用端到端语音识别框架，不断简化系统架构取得了比较良好的效果。最近也有探讨多语言的迁移学习的应用[30]，以及讨论多语种共享编码空间的工作[31]，都对未来多语言语音识别有重要的指导意义。但目前仍在声学编码的可解释性、多语言的数量和质量的量化分析上面临较大挑战。

未来展望

语音识别系统在端到端领域已经逐渐发展成熟，随着越来越多新的模型算法的加入，该领域仍面临可解释性低、流式识别和低资源等挑战。未来结合迁移学习和无监督学习以及更具可解释性的研究都是对该领域十分有意义的。而多语言语音识别的框架和算法以及可行性依据都在探索过程中，从传统建模加解码器再到现在学术界的主流端到端框架都在对多语言语音识别进行不断尝试，从多个维度和模块尝试融合多语种信息构建统一的多语言语音识别系统。未来主要会对端到端的多语言语音识别系统进行研究，探索迁移学习、多任务学习、强化学习以及生成对抗网络用在多语言语音识别系统构建上的尝试。

参考文献

[1] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1): 30-42.

[2] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 369–376.

[4] P. C.Woodland, J. J. Odell,V.Valtchev, and S. J.Young, “Large vocabulary continuous speech recognition using HTK,” in Proc. 1994 IEEE Int. Conf. Acoust., Speech, Signal Process., 1994, vol. 2, pp. 125–128.

[5] M. Mohri, F. Pereira, and M. Riley, “The design principles of a weighted finite-state transducer library,” Theoretical Comput. Sci., vol. 231, no. 1, pp. 17–32, 2000.

[6] M. Mohri, F. Pereira, and M. Riley, “Weighted finite-state transducers in speech recognition,” Comput. Speech Lang., vol. 16, no. 1, pp. 69–88, 2002.

[7] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of the 31st International Conference on Machine Learning (ICML-14). 2014: 1764-1772.

[8] Y. Miao, M. Gowayyed, and F. Metze, “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” arXiv:1507.08240, 2015.

[9] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, “An empirical exploration of CTC acoustic models,” in Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 2623–2627

[10] D. Nolden, R. Schl¨uter, and H. Ney, “Advanced search space pruning with acoustic look-ahead for WFST based LVCSR,” in Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process., 2013, pp. 6734–6738.

[11] T. Hori, C. Hori, Y. Minami, and A. Nakamura, “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1352–1365, May 2007.

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

[13] Zhang Q, Lu H, Sak H, et al. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7829-7833.

[14] Dong L, Xu B. CIF: Continuous integrate-and-fire for end-to-end speech recognition[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP). IEEE, 2020: 6079-6083.
[15] Luo H, Zhang S, Lei M, et al. Simplified Self-Attention for Transformer-based End-to-End Speech Recognition[J]. arXiv preprint arXiv:2005.10463, 2020.
[16] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[J]. arXiv preprint arXiv:2005.08100, 2020.
[17] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and T. Wang, "Towards language independent acoustic modeling," in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol. 2, pp. 1029–1032, 2000.
[18] D. Horga, "Handbook of the international phonetic association. a guide to the use of the international phonetic alphabetcambridge: Cambridge university press (1999), (204 stranice)," Govor, vol. 16, no. 2, pp. 181–188, 1999.
[19] J. L. Hieronymus, "Ascii phonetic symbols for the world"s languages: Worldbet,"1993.
[20] J. C. Wells, "Computer-coded phonemic notation of individual languages of the european community," Journal of the International Phonetic Association, vol. 19, no. 1, pp. 31–54, 1989.
[21] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. J. Moreno, E.Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model,"pp. 4904–4908, 2018.
[22] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition,"in Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE, pp. 265–271, IEEE, 2017.
[23] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7304–7308, IEEE, 2013.
[24] R. Caruana, "Multitask learning," Machine Learning - Special issue on inductive transfer archive, 1997.
[25] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," Proc. Interspeech 2017, pp. 704–708, 2017.
[26] S. Feng and T. Lee, "Improving cross-lingual knowledge transferability using multilingual tdnn-blstm with language-dependent pre-final layer," in Proc.

Interspeech 2018, pp. 2439–2443, 2018.

[27] H. Xu, H. Su, C. Ni, X. Xiao, H. Huang, E. S. Chng, and H. Li, “Semisupervised and cross-lingual knowledge transfer learnings for dnn hybrid acoustic models under low-resource conditions.,” in Interspeech 2016, pp. 1315–1319, 2016.

[28] J. Li, R. Zheng, B. Xu, et al., “Investigation of cross-lingual bottleneck features in hybrid asr systems.,” in INTERSPEECH, pp. 1395–1399, 2014.

[29] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, “Multi-dialect speech recognition with a single sequence-to-sequence model,” arXiv preprint arXiv:1712.01541, 2017.

[30] Kamper H, Matusevych Y, Goldwater S. Improved acoustic word embeddings for zero-resource languages using multilingual transfer[J]. arXiv preprint arXiv:2006.02295, 2020.

[31] Żelasko P, Moro-Velázquez L, Hasegawa-Johnson M, et al. That Sounds Familiar: an Analysis of Phonetic Representations Transfer Across Languages[J]. arXiv preprint arXiv:2005.08118, 2020.