# 利用门控语言专家和课程训练构建高准确率多语言自动语音识别

*Eric Sun*、*Jinyu Li*、*Yuxuan Hu*、*Yimeng Zhu*、*Long Zhou*、*Jian Xue*、*Peidong Wang*、*Linquan Liu*、*Shujie Liu*、*Edward Lin*、*Yifan Gong*

Microsoft

**摘要**

我们提出门控语言专家和课程训练方法，以增强多语言 Transformer 转导器模型，且在推理期间无需用户提供语言识别（LID）输入。我们的方法引入门控机制和语言识别损失，使 Transformer 专家能够学习特定语言信息。通过结合门控 Transformer 专家与共享 Transformer 层，我们构建多语言 Transformer 模块，并利用线性专家对联合网络进行有效正则化。课程训练方案利用语言识别引导门控专家提升各自语言的性能。在涉及英语和西班牙语的双语任务上的实验结果表明，与基线双语模型和单语模型相比，平均相对词错误率分别降低了 12.5% 和 7.3%，取得了显著提升。值得注意的是，我们的方法达到了与使用预言机语言识别进行训练和推理的上界模型相当的性能。将该方法扩展至三语、四语和五语模型后，我们观察到与双语模型类似的优势，凸显了该方法易于扩展至多种语言。

索引词—多语言自动语音识别，Transformer转导器，语言识别，专家

## 1. 引言

尽管端到端（E2E）模型在自动语音识别（ASR）领域取得了快速进展 [1-8], ，但鉴于全球超过 60% 的人口会说两种以上语言 [10]，对多语言自动语音识别模型的需求十分迫切。为开发端到端多语言模型，人们进行了大量尝试 [11-26], 。这些模型通过引入语言识别（LID）信息（采用独热向量或可学习嵌入向量），区分不同语言，因而能够达到与单语基线相当甚至更优的自动语音识别性能。然而，要构建在实际应用中表现与单语系统相近的流式多语言自动语音识别系统，关键在于模型推理时不要求用户输入任何 LID 信息。一种可能的解决方案是将 LID 推断为嵌入向量，并将其附加到输入特征上 [18, 19, 21]。然而，这种方法要么因语言识别预测不准确而导致提升有限，要么为了获得可靠的语言识别预测而引入额外延迟 [18, 19]。

本文介绍了一种带有辅助语言识别损失和线性专家的门控 Transformer，旨在提升多语言语音识别性能。所提出的门控 Transformer 专家能够构建紧凑模型，同时促进不同语言间语音信息的有效共享。线性专家用于调节联合网络输出，从而显著提升模型训练稳定性。此外，我们提出了一种课程训练策略，在训练过程中利用语言识别输入，增强专家对特定语言信息的学习。值得注意的是，我们的模型在推理过程中无需用户提供任何语言识别输入。在英语和西班牙语双语模型上的实验结果表明，与不使用语言识别输入的双语模型基线相比，词错误率（WER）相对降低了 12.5%；这一性能与使用预言机语言识别输入的双语模型相近。此外，我们的双语模型优于单语基线。我们的方法还成功扩展至三语、四语和五语模型，模型大小仅略有增加，并取得了与双语模型相近的效果。

## 2. 相关工作

专家的概念此前已应用于 [27, 28], 中的自动语音识别，也曾被用于应对双语代码切换的挑战，如[29, 30], 所示，其中专用编码器充当特定语言的专家。该方法利用门控函数组合不同专家的输出，但没有使用语言识别损失来规范这些输出。在[31], 中，采用基于 RNN-Transducer 且以语言识别输入为条件的专家进行多语言自动语音识别。我们的方法与 [31] 中所述方法的一个关键区别在于，后者所用的条件专家在训练和推理时均以预言机语言识别结果作为输入。相比之下，我们的方法无需用户在推理时提供 LID 信息。尽管 [31] 简要讨论了使用 LSTM门控取代基于语言识别的语言门控，但其训练过程并未纳入任何语言识别损失，因此无法学习

特定语言信息。在我们的设计中，我们采用与语言识别损失相关联的门控机制，并运用课程训练，使专家真正掌握依赖语言的信息，从而在推理时有效地区分不同语言。 [20]提出了一种可配置多语种模型，该模型只需训练一次，便可适配不同的语言组合。它在编码器和预测网络中均采用线性语言专家。

## 3. 采用门控语言专家的多语言 Transformer 转导器

### 3.1 Transformer转导器模型

神经转导器模型 [4] 由三个部分组成：声学编码器、标签预测网络和联合网络。神经转导器模型可以使用不同类型的模型作为编码器，例如 RNN-T 中的长短期记忆网络 [4]，以及 Transformer转导器（T-T）中的 Transformer [7, 8, 9, 17, 20, 21, 22]。在本研究中，我们使用 T-T 作为开发所用的骨干模型。编码器网络中的每个 Transformer 模块均由一个多头自注意力层和其后的一个前馈层构成。神经转导器模型的损失函数是给定输入声学特征 $\boldsymbol{x}$ 时输出目标标签 $\boldsymbol{y}$ 的负对数后验概率，其定义为

$$L_{rnnt} = -\log P(\boldsymbol{y}|\boldsymbol{x}) \quad (1)$$

该值通过 [4] 中的前向后向算法计算得出。

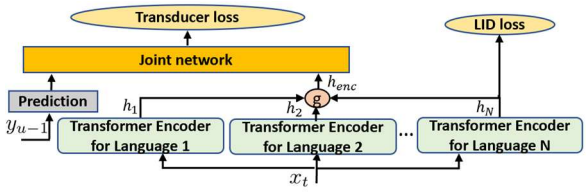



图 1：采用针对不同语言分别设置的 Transformer 编码器的多语 T-T 模型架构

### 3.2 编码器中的门控语言 Transformer 专家

编码器是 T-T 模型中最重要的组件。在多语言语音识别中，如果所有语言共用一个编码器，不同语言可能会相互混淆，从而影响模型性能，正如 [22] 中所讨论的那样。在本研究中，我们为每种语言配置了专属的 Transformer 编码器，如图 1 所示。不同编码器可以通过门控 $g$ 进行组合，该门控定义为

$$O = W_o(\tanh(\sum_{i=0}^{N} W_i\, h_i))) \quad (2)$$

$$g = Softmax(O) \quad (3)$$

其中，$\mathrm{h}_i$ 是语言 $i$ 的编码器嵌入，$W_i$ 是与每种语言相关联的线性矩阵，$N$ 是语言数量。整个编码器网络的嵌入输出为

$$h_{enc=} \sum_{i=0}^{N} g_i h_i \quad (4)$$

由于门控 $g$ 会组合来自不同语言的编码器嵌入，编码器网络本身并不知道应作为相应语言的 Transformer 编码器服务。为了使编码器网络学习各自对应的语言，提出了语言识别交叉熵（CE）损失，其定义如下

$$L_{lid} = CE(O) \quad (5)$$

因此，总损失定义为原始转导器损失与语言识别损失的组合，如下所示

$$L = L_{rnnt} + \lambda L_{lid} \quad (6)$$

其中，$\lambda$ 是用于调整这两种损失比例的权重。

上述设计的一个缺点是，每种语言都有自己的编码器，这使得模型难以随着语言数量的增加而扩展。此外，为不同语言设置彼此分离的编码器可能不是最优选择，因为不同语言之间仍然可以共享许多声学条件、说话人声音特征，甚至发音相似性。因此，我们针对基于 T-T 的多语言自动语音识别提出了另一种更有效且紧凑的编码器结构，如图 2 所示。不同于为每种语言构建完全分离的编码器，不同语言可以共享 Transformer 层，同时每种语言仍可拥有各自对应的 Transformer 专家；这些专家通过门控 $g$ 进行组合，如公式（2）和（3）所定义，而语言识别损失函数定义如下

$$L_{lid} = CE(\sum_{l=1}^{M} O_l) \quad (7)$$

其中，$O_l$ 是公式（2）中的 logit 输出，但来自不同的层 $l$。如图 2 虚线所示，语言相关专家、其门控以及共享 Transformer层可以构成一个多语言 Transformer 模块。在 T-T多语言模型中，可以包含多个多语言 Transformer 模块。此外，我们还可以在网络底部应用共享Transformer层，因为不同语言的输入滤波器组语音特征具有许多共同特征。借助这一新结构，网络可以在学习不同语言的语言学知识的同时，共享共同语音信息。此外，多语言模型的大小可以通过模块数量轻松控制，这有利于将多语模型扩展到更多语言。

### 3.3 用于联合网络的门控语言线性专家

T-T模型中的联合网络结合了声学信息和

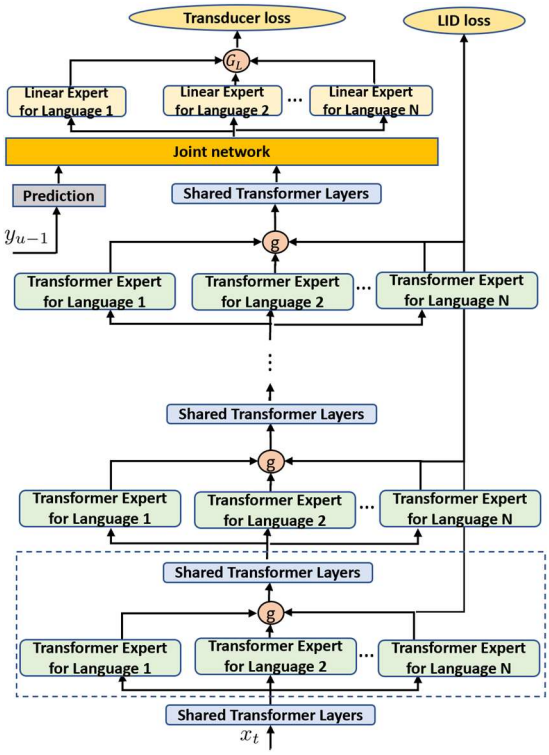



图 2：多语 T-T 模型的架构：底部为共享Transformer层，其上为多语言 Transformer 模块（虚线所示为一个模块），包括门控语言 Transformer 专家和共享层，以及用于联合网络的线性专家

来自编码器和预测网络的语言信息。受 [20]，启发，我们不在编码器和预测网络上分别添加特定语言的门控线性专家，而是直接将线性专家应用于联合网络的输出，如图 2 所示。定义 $W = \{w_1, w_2, \dots, w_N\}$ 为 $w_i$ ($i$ = 1 至 N) 的线性矩阵组合，其中 $w_i$ 是与某种语言对应的线性专家矩阵$i$。乘以这个语言门控线性矩阵后，新的联合网络输出定义为

$$h'_{joint} = h_{joint}\, W\, G_L \qquad (8)$$

其中，$\mathrm{h}_{joint}$ 是联合网络的输出，$G_L = \{g_{L_1}, g_{L_2}, \dots, g_{L_N}\}$是门控网络，即一种仿射 LID 投影，被称为“LID 门控”，由语言识别输入控制。为了获得更好的模型性能，我们在模型训练中确实利用了预言机语言识别；但在推理过程中，不需要用户输入任何 LID 信息，具体细节将在第3.4节中进一步说明。此外，还会对 $\mathrm{h}'_{joint}$ 应用层归一化，以稳定训练。

### 3.4 基于 LID 的课程训练策略

尽管应用 LID 损失来促使 Transformer 专家学习其对应语言的语音信息，但网络仍可能受到单词发音相似且书写字母相同的语言干扰，尤其是在训练初期，这可能导致模型性能下降。在 [29], , 我们预训练了一个明确利用 LID 信息的种子模型，以缓解这种语言混淆问题。在我们的方法中，在模型训练初期，将单个 LID 输入 LID 门控，如第 3.3 节所述，将其对应的值设为 1，从而引导 Transformer 专家学习各自的语言。随着训练的进行，我们还会将所有语言的多语言 LID 输入 LID 门控，同时仍保留一定比例的单个 LID 输入，以逐步实现从单个 LID 训练到多个 LID 训练的过渡，这称为针对 LID 输入的课程训练策略。在模型训练的最后阶段，只将所有语言对应的多个 LID 输入 LID 门控。以双语模型为例。在训练开始时，我们只针对不同的输入语言，将 LID 向量 [0, 1]或 [1, 0] 输入 LID 门控；在训练中期，我们以概率 p 将 1 热向量 [0,1], [1,0], 输入模型训练，以概率 1- p 将双热向量 [1,1] 输入模型训练。随着训练的进行，p 逐渐减小。在训练的最后阶段，p 降至 0，模型训练中只输入双热向量 [1,1]。于是，公式（2）进一步改进为

$$O_l = W_o(\tanh(\textstyle\sum_{i=0}^{N} W_{l_i}\, h_{l_i}\, g_{L_i}))) \qquad (9)$$

其中，$l$ 表示多语言 Transformer 模块的层数。在训练开始时，语言门控 $G_L$ 中与语言 $i$ 对应的唯一一个 $g_{L_i}$ 设为 1；在训练的最后阶段，$G_L$ 中的所有值均设为 1。（注：为使图示更简洁，我们在图 2 中省略了 Transformer 专家的 $G_L$ 。）推理时，系统会将所有元素值均为 1 的多热 LID 向量传递给 LID 门控，以进行多语言语音识别，用户无需输入任何特定的 LID 信息。这种训练策略也适用于模型训练中公式（8）所述联合网络的线性专家。

## 4. 实验设置

### 4.1 语言和数据

我们开发了多语 T-T 模型，支持最多五种语言：英语（EN）、西班牙语（ES）、德语（DE）、意大利语（IT）和法语（FR）。对于所有这些语言，训练数据和测试数据均经过转录和匿名化处理，并移除了个人身份信息。测试数据既包括从与训练数据相同分布中采样的域内数据，也包括与训练数据分布不同的域外数据。每种语言的训练数据和测试数据量汇总于表 1。

### 4.2 模型结构和训练配置

在我们的基线 T-T 模型中，编码器采用 18 个基础 Transformer 模块，每个模块包含 320 个隐藏节点、8 个注意力头和 2048 个前馈节点；预测网络采用 2 个 LSTM 层，嵌入层和隐藏层的维度均为 1024。基础 Transformer 模块也被用作多语言 Transformer 块中的 Transformer 专家，如图 2 所示，结构不作任何改变。我们使用 80 维对数梅尔滤波器组，窗口长度为 25 毫秒（ms），步长为 10ms。在模型训练和推理过程中，均将 LID 向量附加到输入特征上。我们采用两个卷积层，以获得采样率为 40ms 的特征。输入声学特征序列被分割成块，块大小为 4，且各块之间不重叠。此外，我们还使用左侧的 18 个块，以利用历史声学信息。一个有效的

表 1：各语言的训练和测试数据（小时）

| 语言 | 训练 | Test |
|---|---|---|
| EN | 23,035 | 208 |
| ES | 3,770 | 33 |
| DE | 2,893 | 38 |
| IT | 3,345 | 19 |
| FR | 3,176 | 33 |
| 总计 | 36,219 | 331 |

我们设计了掩码策略，如 [9] 所示，用于截断历史信息并允许有限的未来前瞻信息。学习率预热策略与 [32] 相同。每个训练小批次都包含来自所有语言的话语，并根据各自的训练数据分布进行抽样。我们分别训练 BPE模型，为每种语言生成词元列表，然后将这些词元列表合并为多语言模型的输出。对于单语、双语、三语、四语和五语模型，其输出词元数量分别为 4k、7k、10k、12k 和 14k。

## 5. 结果

### 5.1 英语和西班牙语双语模型

我们在第3节中使用英语和西班牙语双语模型，开始研究所提出的方法。表2中的基线英语和西班牙语双语模型称为 B2，它通过汇集两种语言的所有数据进行训练，且未向模型提供 LID 信息。B2 的模型大小为 80M。此外，我们还训练了单语模型作为另一种基线。表2还列出了不同模型结构的参数规模信息。

从表2可以看出，B2的平均词错误率为16.0%，高于单语模型的平均词错误率15.1%。此外，我们训练了一个双语模型B1，在训练和推理阶段均将预言机语言识别作为输入，以确定双语模型性能的上界。与B2相比，B1的相对词错误率降幅为13.1%。按照与[16]类似的方法，预言机语言识别以独热向量的形式附加到输入特征中。

表2：英语和西班牙语双语模型的词错误率（%）和参数数量（M）

| 模型 | 参数 | EN | ES | Avg |
|---|---|---|---|---|
| 单语 | 78*2 | 13.2 | 16.2 | 15.1 |
| B1 预言机语言识别 | 80 | 12.8 | 14.9 | 13.9 |
| B2 无语言识别基线 | 80 | 14.9 | 17.1 | 16.0 |
| B3 完全分离的编码器 | 133 | 13.6 | 16.0 | 14.8 |
| B4 6 个 Transformer 块 | 100 | 13.1 | 15.5 | 14.3 |
| B5 3 个 Transformer 块 | 90 | 13.2 | 15.8 | 14.5 |
| B6 + 联合线性专家 | 90.5 | 13.1 | 15.6 | 14.4 |
| B7 + 用于语言识别输入的课程学习策略 | 90.5 | 13.0 | 14.9 | 14.0 |

随后，我们训练了一个双语模型 B3，其编码器完全分离，如图 1 所示。为防止模型发散，门控机制基于两个编码器最后 3rd 层的嵌入，并将最后两层用作共享层。B3 的词错误率为 14.8%，与 B2 相比，相对词错误率降低了 7.5%。然而，编码器的模型大小几乎翻了一倍，参数总量也增加到 133M，因此难以扩展至更多语言。

如第 3.2 节所述，我们引入了模型 B4，并使用 6 个多语言 Transformer 块进行训练（图 2 中以虚线表示），以减小模型大小的增幅。每个多语言 Transformer 块由两个共享层组成，且在多语言 Transformer 块之前的底层不设置共享层。B4 不仅结构更紧凑，模型参数仅为 100M，还促进了不同语言之间语音信息和语言信息的共享。与 B2 和 B3 相比，B4 的相对词错误率分别降低了 10.6% 和 3.4%。为进一步减小模型大小，我们将多语言 Transformer 块的数量从 6 个减至 3 个，并在多语言 Transformer 块之前增加 9 个共享 Transformer 层，以训练模型 B5。与 B4 相比，该模型的相对词错误率仅差 1.4%，但模型参数减少了 10 M。我们尝试在不同位置添加 Transformer 块，结果表明，将其放置在靠近模型输出的位置对 B5 最为有效。

在 B5 的基础上，我们按照第 3.3 节提出的方法，采用联合线性专家训练了 B6 模型。这使性能略有提升，平均词错误率从 14.5% 降至 14.4%，模型参数仅增加了 0.5M。此外，B6 的训练方案稳定性显著更高，尤其是在构建多语言模型时纳入更多语言的情况下。

最后，我们对语言识别输入采用课程学习策略（CT），引导语言相关 Transformer 和线性

表 4： 三语、四语和五语模型的词错误率（%）及参数数量（M）

| 语言 | 参数 | EN | ES | DE | IT | FR | Avg |
|---|---|---|---|---|---|---|---|
| 单语 | 78*n (n=3,4,5) | 13.2 | 16.2 | 15.7 | 13.2 | 16.5 | 15.1 |
| T1 三语预言机语言识别 | 82 | 12.8 | 14.5 | 15.5 | - | - | 14.3 |
| T2 无语言识别的三语 | 82 | 14.9 | 17.2 | 16.0 | - | - | 16.0 |
| T3 门控专家三语 | 100.5 | 12.9 | 14.6 | 15.4 | - | - | 14.3 |
| Q1 四语预言机语言识别 | 84 | 12.9 | 14.6 | 15.2 | 12.0 | - | 13.7 |
| Q2 无语言识别的四语 | 84 | 15.0 | 17.6 | 16.2 | 14.7 | - | 15.9 |
| Q3 门控专家四语 | 110.5 | 13.0 | 14.8 | 15.4 | 12.2 | - | 13.9 |
| P1 五语预言式语言识别 | 86 | 12.9 | 14.5 | 15.3 | 12.0 | 15.2 | 14.0 |
| P2 无语言识别的五语 | 86 | 15.2 | 18.2 | 16.5 | 15.6 | 16.6 | 16.4 |
| P3 门控专家五语 | 120.5 | 13.2 | 14.8 | 15.5 | 12.1 | 15.5 | 14.2 |

专家学习各自对应的语言。由此得到模型 B7，与 B2基线和单语模型相比，其相对词错误率降幅分别达到 12.5% 和 7.3%。B7 的平均结果甚至与 B1 的上层模型相近（14.0% 对比 13.9%）。此外，我们增加了 B2 的参数数量，使其与 B7 相匹配；参数略微增加并未对词错误率产生显著影响。

我们还在西班牙语/英语语码转换测试集上评估了模型 B2 和 B7 的性能，其中英语词语作为实体名称纳入测试。表3中的结果表明，与 B2基线相比，B7 的相对词错误率降幅显著，达到 9.3%。这进一步证明了我们的方法在多语种自动语音识别方面的有效性。

表3：西班牙语/英语语码转换测试集结果

| 模型 | WER |
|---|---|
| B2 | 28.1% |
| B7 | 25.5% |

### 5.2 扩展至更多语言

在成功构建英语和西班牙语双语模型的基础上，我们进一步扩展方法，构建涵盖德语、意大利语和法语等更多语言的多语言模型。这些模型的结构与双语模型 B7 高度相似，但针对每种新增语言加入了专用 Transformer 和线性专家。具体而言，对于每种新增语言，我们在模型 B7 的 Transformer 块中加入了三个语言相关 Transformer 和一个线性专家，使模型参数量增加了 10M，如表 4 所示。为衡量多语言模型的性能，我们提供了单语模型作为参考。表 4 中列出了未使用任何 LID 信息的基线模型结果，以及采用预言机语言识别进行训练和推理的模型上界结果，以便与我们提出的模型进行比较。

随着纳入更多语言，基线模型的参数规模也略有增加，这是由于输出层中纳入了更多词元标签。同样，我们的门控专家模型始终表现出显著提升。与相应的不使用语言识别的三语、四语和五语基线模型相比，其词错误率平均分别降低了 10.6%、12.6% 和 13.4%。将我们的模型与使用预言机语言识别的上界模型进行比较后，我们发现两者的词错误率相近：三语模型为 14.3% 对 14.3%，四语模型为 13.7% 对 13.9%，五语模型为 14.0% 对 14.2%。

在考察各个语言时，我们观察到，对于未使用 LID 的西班牙语，随着所加入语言数量的增加，词错误率逐渐升高：三语、四语和五语模型分别为 17.2%、17.6% 和 18.2%。德语和意大利语也呈现出类似趋势。然而，我们的模型在向多语言模型中加入更多语言时，并未出现性能下降。其性能趋势保持稳定，与采用预言机 LID 信息进行训练和推理的模型相似。最后，我们基于门控专家的三语、四语和五语模型，其词错误率始终达到与单语模型相当或更优的水平。

## 6. 结论

本文提出了一种新颖方法，通过利用门控语言专家和辅助语言识别损失来提升多语 T-T 模型的性能，同时在模型推理期间无需用户提供任何语言识别输入。我们的方法是在编码器中构建多语言 Transformer 块，该模块由门控 Transformer 专家和共享层组成。该框架通过共享层促进通用语音和声学信息的共享，同时 Transformer 专家专注于学习语言相关知识。此外，我们将线性专家纳入联合网络输出，以有效正则化联合语音声学与词元标签信息，显著提高训练稳定性。为进一步增强特定语言专家的学习，我们引入了一种结合语言识别输入的课程训练策略。在英语和西班牙语双语

模型上的实验结果表明，与不含 LID 信息的双语模型基线和单语模型相比，词错误率相对降幅分别达到 12.5% 和 7.3%。此外，我们的方法实现了与使用预言机语言识别进行训练和推理的模型相近的模型性能。将我们的方法扩展至五种语言后，也得到了与双语模型中观察到的相似模式。

## 7. 参考文献


[1] J. Li，《端到端自动语音识别的最新进展》，APSIPA信号与信息处理汇刊，第11卷，第1期，2022年。

[2] S. Watanabe 等人，《用于端到端语音识别的混合连接时序分类/注意力架构》，IEEE信号处理精选主题期刊，第11卷，第8期（2017年）：1240–1253。

[3] W. Chan、N. Jaitly、Q. Le 和 O. Vinyals，《聆听、关注与拼写：一种用于大词汇量会话语音识别的神经网络》，载于 ICASSP 会议论文集，2016年，第4960–4964页。

[4] Alex Graves，《使用循环神经网络进行序列转导》，载于第29届 ICML 会议论文集，2012年。

[5] Y. He、T.N. Sainath、R. Prabhavalkar 等人，《面向移动设备的流式端到端语音识别》，载于 ICASSP 会议论文集，2019年。

[6] J. Li、R. Zhao、H. Hu 和 Y. Gong，“改进用于端到端语音识别的循环神经网络转导器建模”，载于 ASRU 论文集，2019年。

[7] Q. Zhang、H. Lu、H. Sak、A. Tripathi、E. McDermott、S. Koo 和 S. Kumar，“Transformer 转导器：一种采用 Transformer 编码器和 RNN-T 损失的可流式语音识别模型 ”，载于 ICASSP 2020 年论文集。

[8] C.-F. Yeh、J. Mahadeokar、K. Kalgaonkar、Y. Wang、D. Le、M. Jain、K. Schubert、C. Fuegen 和 M. L. Seltzer，“Transformer 转导器：采用自注意力的端到端语音识别”，arxiv.org/abs/1910.12977，2019年。

[9] X. Chen、Y. Wu、Z. Wang、S. Liu 和 J. Li，“在大规模数据集上开发用于语音识别的实时流式 Transformer 转导器”，载于 ICASSP 2021年论文集。

[10] “多语言人群”，http://ilanguages.org/bilingual.php，访问日期：2021年10月16日。

[11] Watanabe, S., Hori, T. 和 Hershey, J. R.，“用于联合语言识别与语音识别的语言无关端到端架构”，载于 ASRU 会议论文集，第 265-271 页，2017 年。

[12] S. Toshniwal, T. Sainath, R. Weiss, B. Li, P. Moreno, E. Weinstein 和 K. Rao，“采用单一端到端模型的多语言语音识别”，载于 ICASSP 会议论文集，2018年。

[13] B. Li、Y. Zhang、T. Sainath、Y. Wu 和 W. Chan，《字节就是一切：基于字节的端到端多语言语音识别与合成》，载于 ICASSP 论文集，2018年。[14] V. Pratap、A. Sriram、P. Tomasello、A. Hannun、V. Liptchinsky、Synnaeve 和 R. Collobert，《大规模多语种自动语音识别：50 种语言、1 个模型、10 亿参数》，载于 INTERSPEECH 论文集，2020年。[15] O. Adams、M. Wiesner、S. Watanabe 和 D. Yarowsky，《大规模多语种对抗式语音识别》，arXiv:1904.02210，2019年。[16] A. Kannan、A. Datta、T. N. Sainath、E. Weinstein、B. Ramabhadran、Y. Wu、A. Bapna、Z. Chen 和 S. Lee，《采用流式端到端模型的大规模多语言语音识别》，载于 INTERSPEECH 论文集，2019年。[17] Y. Zhu、P. Haghani、A. Tripathi、B. Ramabhadran、B. Farris、H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur, P. J. Moreno, Q. Zhang，“采用自注意力结构化参数化的多语言语音识别”，载于 INTERSPEECH 会议论文集，2020年。

[18] A. Waters、N. Gaur、P. Haghani、P. Moreno、Z. Qu，“在多语言端到端语音识别中利用语言识别”，载于 ASRU 会议论文集，2019年。

[19] S. Punjabi、H. Arsikere、Z. Raeesy、C. Chandak、N. Bhave、A. Bansal、M. Muller、S. Murillo、A. Rastrow、S. Garimella 等，“借助联合语言识别实现流式端到端双语自动语音识别系统”，arXiv 预印本 arXiv:2007.03900，2020年。

[20] L. Zhou、J. Li、E. Sun 和 S. Liu，“可配置多语言模型即可识别所有语言”，载于 ICASSP 2022年会议论文集。

[21] C. Zhang、B. Li、T.N. Sainath、T. Strohman、S. Mavandadi、S. Chang、P. Haghani，“借助联合语言识别实现流式端到端多语言语音识别”，载于 INTERSPEECH 2022年会议论文集。

[22] E. Sun, J. Li, Z. Meng, Y. Wu, J. Xue, S. Liu 和 Y. Gong，《通过减少语言混淆改进多语言 Transformer 转导器模型》，载于 INTERSPEECH 会议论文集，2021年。

[23] C. Zhang, B. Li, T.N. Sainath, T. Strohman, S. Mavandadi, S. Chang, P. Haghani，《结合联合语言识别的流式端到端多语言语音识别》，载于 INTERSPEECH 会议论文集，2022年。

[24] B. Houston 和 K. Kirchhoff，《探索用于多语言端到端语音识别的语言专用自注意力参数》，载于 SLT 会议论文集，2022年。

[25] C. Zhang, B. Li, T.N. Sainath, T. Strohman 和 S. Chang，《UML：用于多语言自动语音识别的通用单语输出层》，arxiv.org/abs/2302.11186，2023年。

[26] C. H. Yang、B. Li、Y. Zhang、N. Chen、R. Prabhavalkar 和 T.N. Sainath 和 T. Strohman，“从英语到更多语言：用于跨语言语音识别的参数高效模型重编程”，arxiv.org/abs/2301.07851，2023年。

[27] S. Papi、E. Trentin、R. Gretter、M. Matassoni 和 D. Falavigna，“用于自动语音评分的深度神经专家混合”，载于 INTERSPEECH 会议论文集，2020年，第3845–3849页。

[28] A. Jain、V. P. Singh 和 S. P. Rath，《使用专家混合的多口音声学模型进行语音识别》，载于 INTERSPEECH 会议论文集，2019年，第 779–783 页。

[29] Y. Lu、M. Huang、H. Li、J.、Y. Qian，《使用专家混合的普通话-英语语码转换语音识别双编码器Transformer网络》，载于 INTERSPEECH 会议论文集，2020年。

[30] S. Dalmia、Y. Liu、S. Ronanki 和 K. Kirchhoff，《用于语码转换语音识别的 Transformer-Transducers》，载于 ICASSP 会议论文集，2021年。

[31] N. Gaur、B. Farris、P. Haghani、I Leal、P. J. Moreno 等，《用于多语言语音识别的知情专家混合》，载于 ICASSP 会议论文集，2021年。

[32] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N Gomez、Ł. Kaiser 和 I. Polosukhin，《Attention Is All You Need》，载于《神经信息处理系统进展》，第 5998–6008 页，2017 年。