P1：尊敬的各位评委老师，上午好，我叫黄子恒，控制科学与工程专业，论文题目是面向车载嵌入式设备的智能语音对话方法研究。

P2：我将从研究背景及意义、主要工作和总结与展望这3个部分展开介绍。

P3：首先是研究背景及意义。

P4：车载语音交互已经成为汽车的标准化配置之一。自动语音识别和自然语言理解在深度学习的驱动下取得了重大进展。然而高精度的深度学习模型需要庞大的计算资源，当前主流车载语音对话系统采用的“云—端”方式存在着一定的数据安全隐患。为解决这一问题，论文重点研究针对ASR和NLU的技术创新应用方法，探索以离线方式运行的车载智能语音对话平台搭建，在本地实现数据安全、自然实时的语音对话。

P5：在ASR方面，现阶段主要基于注意力机制搭建序列到序列模型，其主要代表是Transformer模型，但存在参数量庞大、计算复杂和难以部署等缺点；在NLU方面，现阶段主要使用基于显式联合建模的方法，但该类模型没有高效的交互信息融合手段，导致模型交互能力不足，虽然可以引入预训练模型来补充丰富的语义特征，但是这会导致参数量剧增。

P6：针对上述问题，论文主要研究基于残差分组线性变换解码器的ASR模型和基于标签感知图交互的NLU模型，并搭建面向车载嵌入式设备的本地智能语音对话系统。

P7：第3章的主要内容为基于残差分组线性变换解码器的自动语音识别。

P8：Transformer模型在ASR领域表现亮眼，但该类模型参数量庞大。针对这一问题，提出了基于残差分组线性变换的解码器结构，该结构关键模块为“钻石”型缩放单元，其基本计算步骤如图（a）所示，首先通过切分混合函数对输入进行处理以保证学习效率，再使用分组线性变换进行前向传播。和普通全连接网络不同，分组线性变换允许分组的输入神经元和输出神经元进行局部连接，计算效率更高，参数量更少。

（左侧为编码器，对输入语音帧序列进行编码，使用注意力机制提取深层次特征并获取语义信息；右侧为解码器，根据历史文本对编码器输出进行解码，产生当前时刻文本。）

P9：基于“钻石”型缩放单元提出轻量级解码器块，它是通过在原始Transformer解码器块上进行两个改进得到的：1、在起始位置使用逐块缩放策略嵌入对应深度、宽度不均匀的缩放单元进行维度变换；2、使用轻量级前馈网络层代替普通前馈网络层。更进一步，用若干个轻量级解码器块堆叠形成解码器，得到改进的、轻量级Transformer模型。另外，使用卷积模块对语音特征进行时间维度上的压缩，减少其中噪点信息造成的影响。

（在对维度变化的倍数相同的情况下，轻量级前馈网络层的参数量是普通前馈网络层的1/r2。）

P10：这里是第3章使用的实验环境、数据集和评价指标。在数据集方面，使用了在ASR领域内应用较为广泛的AISHELL-1中文数据集和TED-LIUM2英文数据集。

P11：将改进的模型与近年来一些性能表现优秀的同类模型进行对比。在AISHELL-1上，与HA-Transformer相比，改进模型在测试集上的CER上升了0.71%，RTF值上升了0.015，这些在实际应用中影响较小，但参数量和计算量分别相对下降了48.35%和77.16%，同时GPU内存占用率也下降了24.94%，引入语言模型后测试集上的CER可进一步降低至6.38%；在TED-LIUM2上的对比实验结果与在AISHELL-1上类似。

P12：为了更加直观地看出所提方法的效果，将上述结果进行可视化处理。可以看出，改进模型的对应点在三幅图上都位于最左侧，除了在AISHELL-1上的CER略高于HA-Transformer，在参数量、计算量和GPU内存使用率等方面都达到了最优。

P13：围绕解码器类型、注意力表示子空间个数和解码器中FNL类型设计消融实验。可以看出，Base模型的对应点在三幅图上都位于最左侧，表明引入轻量级解码器，可以实现模型参数量、计算复杂度和GPU内存使用率的显著降低，而只牺牲小部分的识别性能，达到识别性能和模型轻量化的平衡状态。

（此外，引入轻量级解码器时，需要考虑对于其中注意力表示子空间的个数选取，因为这会在一定程度上影响模型的识别性能。采用单一表示子空间时，模型的轻量化程度几乎没有变化，但在识别性能上表现较差。 ）

P14：第4章的主要内容为基于标签感知图交互的自然语言理解。

P15：针对基于显式联合建模的自然语言理解模型交互能力不足的问题，提出了一种基于标签感知的图交互模型，主要由标签映射模块和全局图交互模块构成。在标签映射模块中，使用Bi-LSTM和注意力机制对标签数据进行编码得到一组基向量r并以此构建标签空间，对于一个给定的原始话语向量ru，可将其映射到空间上获取最佳线性逼近向量ru‘，以融合意图标签特征；在全局图交互模块中，通过构造图注意力网络连接意图节点和槽位节点，使用逐层特征融合实现意图与槽位之间的语句级交互，以实现全局优化。

P16：在经典显式联合建模方法的基础上，在意图解码器前嵌入标签映射模块捕获原始意图标签特征，在槽位解码器前嵌入全局图交互模块对语句级别的意图—槽位交互过程进行建模，得到改进的、基于标签感知的图交互模型。

P17：这里是第4章使用的实验环境、数据集和评价指标。实验环境与第3章一致。在数据集方面，使用了在NLU领域内应用较为广泛的多意图数据集MixATIS和MixSnips。

P18：将改进的模型与近年来一些性能表现优秀的同类模型进行对比，可以看出，在MixATIS上，与对比组中预测效果最好的SDJN相比，改进模型的整体准确率、意图准确率和槽位F1值分别上升了5.3%、0.7%和0.1%。在MixSnips上的结果与在MixATIS上类似。

P19：为了更加直观地看出所提方法的效果，论文将上述结果进行可视化处理。可以看出，改进模型的对应点在两幅图上都位于最高点，除了在Latency值上略高于AGIF，在整体准确率和GPU内存使用率两方面都达到了最优。

P20：围绕标签映射模块和全局图交互模块设计消融实验。可以看出，在IAcc方面，w/o LIM相较于Base模型的下降程度大于w/o GGIM，说明标签映射模块可以获取原始话语和标签语义之间的相关性，以辅助意图检测子任务；在F1值方面，w/o GGIM相较于Base模型的下降程度大于w/o LIM，说明全局图交互模块可以对语句级别的意图—槽位交互过程进行建模，并通过全局优化辅助槽位填充子任务，从而显著提高模型性能。

P21：第5章的主要内容为面向车载嵌入式设备的本地智能语音对话系统。

P22：针对“云—端”方式运行的车载智能语音对话存在数据安全隐患的问题，搭建了面向车载嵌入式设备的本地智能语音对话系统。首先对TX2进行运行环境搭建；然后根据实际应用场景开发硬件、软件平台收集专属驾驶数据，得到包含近52小时、58000条音频的驾驶数据集。

P23：将第3章和第4章提出的模型在专属驾驶数据集上进行训练，ASR模型和NLU模型的效果均优于其他对比方法。模型训练完成后，进行对话系统集成并移植至TX2上，观察其运行时的性能表现。GPU的最大占用率接近85%，在等待音频输入时约为55%；CPU的最大占用率为74.6%；内存占用为60%。上述结果表明，对话系统能在TX2上稳定运行。

P24：围绕TX2搭建全套硬件平台，相关器件主要分布于驾驶舱和后备箱中并用USB连接。TX2接收语音指令进行处理并传递给ARHUD，同时播放语音给予驾驶员反馈。

P25：整套系统已在实验用车上完成搭建。由利龙研究院测试部门主导，对系统进行内部验收测试。对话通过率为97%，对话系统平均响应时间为0.33s，ARHUD平均响应时间为1.87s，论文提出的面向车载嵌入式设备的智能语音对话系统已符合预期的设计需求。

P26：最后是对于论文工作内容的总结和对未来研究方向的展望。

P27：论文的主要研究成果可以总结为三点，提出了一种基于残差分组线性变换的解码器结构，提出了一种基于标签感知的图交互模型，搭建了面向车载嵌入式设备的本地智能语音对话系统。然而论文仍有一些不足之处，后续研究工作可以从两个方面展开，对于流式自动语音识别的支持和对于非自回归解码的支持。

P28：最后是攻读硕士学位期间的研究成果。

P29：恳请各位专家批评指正，谢谢！