arXiv:2602.04217v1 [cs.SD] 4 Feb 2026

前端　　面向基于词元的语音识别的 KEN 增强

Takanori As　hihara、*Shota Horiguchi*、*Kohei Matsuura*、*Tsubasa Ochiai* 和 *Marc Delcroix*

NTT株式会社，日本

摘要

语音信号的离散化表示是各种语音应用中连续特征的高效替代方案，包括自动语音识别（ASR）和语音语言模型。然而，这些表示（例如从自监督学习（SSL）语音模型聚类输出中得到的语义令牌或语音令牌）容易受到环境噪声的影响，从而降低后端任务性能。在本文中，我们提出了一种前端系统，用于从含噪语音中估计干净语音令牌，并使用语义令牌在自动语音识别后端上对其进行评估。我们根据输入/输出域考虑了四类增强模型：波形到波形、令牌到令牌、连续自监督学习特征到令牌以及波形到词元。这些模型均独立于自动语音识别后端进行训练。在 CHiME-4数据集上的实验表明，波形到词元增强在各前端中取得了最佳性能。此外，它在大多数情况下都优于基于连续自监督学习特征的自动语音识别系统。

索引词— 离散语音表示、语义令牌、自动语音识别、语音增强、噪声鲁棒性

1. 引言

离散化语音表示近来已成为构建高效语音处理系统的一项重要研究方向。例如，可以对自监督学习（SSL）语音模型提取的特征应用 k-均值聚类，获得语义令牌或语音令牌。这些表示已成功应用于自动语音识别（ASR） [1–3] 和语音语言模型（SLM） [4]。生成的词元通常会经过序列缩短技术处理，例如去重 [5]和字节对编码（BPE） [6], , 从而提高数据传输、模型训练和解码的效率。

尽管取得了这些进展，使用语义令牌的自动语音识别系统（下文简称词元语音识别）的噪声鲁棒性尚未得到充分研究。在现实场景中，录音经常受到环境噪声干扰，导致自动语音识别性能严重下降。因此，提高自动语音识别的噪声鲁棒性是实际部署中的一个根本问题。为推动这一领域的发展，已举办了多项基准挑战赛 [7, 8]。在这些挑战赛中，使用连续值向量的自动语音识别系统（下文简称连续语音识别），例如采用对数梅尔滤波器组特征（FBANK）或自监督学习派生特征的系统，与语音增强（SE）前端结合后取得了最佳性能 [9, 10]。此外，一些研究报告称，将语音增强前端与自动语音识别目标联合微调还能进一步提升性能 [11]。

尽管已有大量研究探讨了连续语音自动语音识别的鲁棒性，但前端增强对鲁棒词元语音识别的潜力仍基本未得到探索。更具体地说，由于词元语音识别以离散表示作为输入，目前尚不清楚增强应在波形级进行，作为输入传递给自监督学习模型，还是直接在离散词元级进行。这一研究空白凸显了探究词元语音识别能否在噪声条件下兼具效率与鲁棒性的必要性。

本文介绍了多种前端增强策略，并探讨它们在提升词元自动语音识别后端噪声鲁棒性方面的效果。为进行系统比较，我们根据训练时所用的输入/输出表示对增强前端进行分类：波形到波形（W2W-E；传统语音增强）、令牌到令牌（T2T-E）、向量到令牌（V2T-E；其中向量是自监督学习模型输出的加权和特征），以及波形到词元增强（W2T-E）。由于增强前端的训练独立于自动语音识别后端，因此这种模块化设计使增强组件在替换或更新自动语音识别模型时仍可保持不变。

本研究的主要贡献如下：

- 我们首次对多种增强前端进行了全面、系统的评估，以提升离散单元自动语音识别系统的噪声鲁棒性，其中包括提出新颖的 V2T-E 和 W2T-E 方法。
- 在 CHiME-4 [7] 数据集上的实验表明，尽管 W2T-E 方法简单，却在所有前端中取得了最佳的词元自动语音识别性能，在大多数情况下甚至优于采用加权和自监督学习特征的连续语音识别系统。
- 我们表明，自动语音识别（ASR）准确率与离散单元级准确率（通过参考离散单元和增强后离散单元测得）并不总是相关，这说明该指标在预测识别准确率方面存在局限性。

我们希望这项工作能为未来基于词元的语音处理前端增强研究奠定基础。

2. 相关工作

已有多项研究提出了语义词元，旨在使其在不同类型的扰动下保持不变。例如，研究者以基于自监督学习特征训练的多层感知机（MLP）量化器取代了标准的 k-均值量化器 [12]。这种方法增强了数据增强不变性，并在一项以词元作为训练目标的无文本语音到语音翻译任务中取得了更好的性能 [5, 13], 。类似地，有研究提出了一种基于自监督学习特征训练的量化器，使其在各种失真下保留局部信息，同时解耦说话人身份等全局属性 [14]。该方法在零资源语音基准上取得了出色的结果 [15]。这两种方法均可归类为 V2T-E 方法。此外，研究者还探索了其他扰动不变离散单元标记器，例如用于提升对说话人变化的鲁棒性 [13, 16, 17]。在这些工作的基础上，我们探讨了噪声不变语义离散单元作为词元语音识别前端增强方案的适用性。此外，我们还针对不同的输入/输出域进行了全面评估，研究范围超越了以往研究所考虑的 V2T-E [12, 14]。这为输入/输出设计以及如何将增强策略融入离散单元自动语音识别流程提供了更全面的视角。

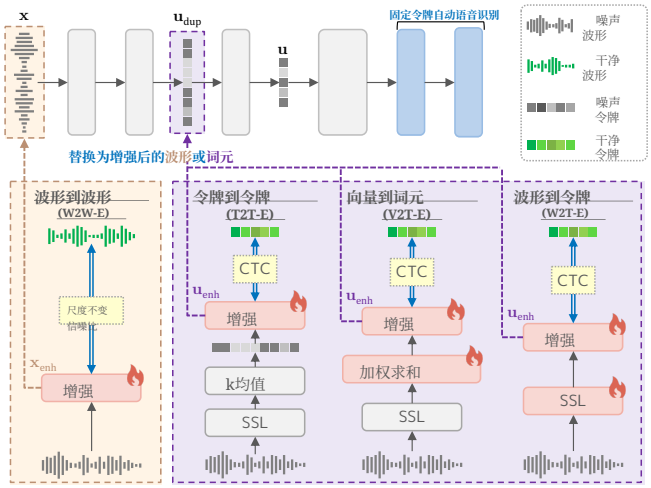


图1：词元自动语音识别后端（上方）以及用于提升后端噪声鲁棒性的四类增强前端（下方）的示意图。

3. 词元自动语音识别的前端增强

本文采用的语义词元自动语音识别后端 [1], 如图1上半部分所示。具体而言，语音信号 $\mathbf{x}$经过自监督学习特征提取后，通过分配 k-均值聚类标签转换为词元序列 $\mathbf{u}^{\text{dup}}$。序列 $\mathbf{u}^{\text{dup}}$ 经过去重得到$\mathbf{u}$，随后通过子词建模（例如字节对编码）进一步压缩。在词元自动语音识别训练中，模型的输入层包含一个可学习嵌入层，用于将离散词元映射到自动语音识别模型表示空间。

然而，在噪声录音中，词元往往会偏离其干净语音对应项，导致自动语音识别性能下降。为此，我们研究增强词元的策略，以实现具有噪声鲁棒性的词元语音识别。传统上，对于连续语音识别，会将语音增强系统（称为 W2W-E）作为前端，以将含噪信号转换为增强后的信号 $\mathbf{x}_{\text{enh}}$ （图1左下部分）。对于词元语音识别输入，$\mathbf{x}_{\text{enh}}$ 随后经过自监督学习特征提取和词元化，生成增强后离散单元。

为全面研究模块化前端，本研究除引入词元级增强方案外，还提出了波形级增强方案（即 W2W-E）。这些方案直接生成增强后的词元序列 $\mathbf{u}_{\text{enh}}$，可作为训练好的词元自动语音识别后端的输入，如图1右下部分所示。由于输出是离散的，训练目标可以采用交叉熵或连接主义时序分类（CTC）。根据初步实验，我们采用基于去重词元的 CTC 作为训练目标，因为其表现略优于基于重复词元的交叉熵。

为了系统地比较词元级增强前端，我们提出了三种方法，分别根据训练时的输入表示来定义：词元、连续向量或波形。与词元相比，后两种表示天然包含更丰富的信息，因此有望实现更高的准确率。然而，它们的维度更高，使学习更具挑战性，因此需要精心设计优化策略。

3.1. 词元到词元增强（T2T-E）

T2T-E 前端是一个简单的网络，用于学习从重复噪声词元到$\mathbf{u}_{\text{enh}}$的映射。该前端采用了输入端的可训练嵌入层，与词元语音识别类似。该映射网络的设计受到近期语音语言模型研究的启发，已有多项研究凸显了它们在基于词元的语音增强方面的有效性 [18, 19]。离散音频与语音基准的基线系统也采用了类似的架构 [20]。

3.2. 向量到词元增强（V2T-E）

V2T-E 前端的设计受到 [12], 的启发：该方法在从特定层（即用于 k-means 聚类的层）提取的自监督学习特征上训练多层感知机，并使用连接时序分类损失来构建扰动不变词元器。在多层感知机训练期间，自监督学习模型的参数保持固定。这个过程可以看作从 k-means 模型中蒸馏量化知识，同时融入噪声不变性。在我们的研究中，我们探讨了将其用作词元语音识别词元增强的适用性，并从两个方面扩展了这一策略：i）用功能更强大的模型替换词元器（即多层感知机）；ii）使用加权和特征。

对于词元化模型，除了原始的多层感知机外，我们还评估了 E-Branchformer [21] 和时序卷积网络（TCN） [22]。E-Branchformer 能增强局部聚合能力，且已有研究表明，它在自动语音识别任务中优于传统的 Transformer/Conformer 模型。[1] TCN 由堆叠的一维膨胀卷积块组成，并在 Conv-TasNet [23]中用作分离网络。

作为 V2T-E 的输入特征，我们采用自监督学习模型所有层的加权求和表示，并为各层设置可训练权重。该方法旨在捕捉自监督学习表示的层次性；这些表示将不同的语音特征（例如说话人身份和语言信息）分布在不同层中，这与 SUPERB 中的做法相同。 [24]请注意，先前的一项研究 [25] 也提出使用此类加权和特征训练词元估计模型。不过，该研究关注的是干净条件下的流式场景，而我们的目标是降噪和增强。

3.3. 波形到词元增强（W2T-E）

W2T-E 前端与 V2T-E 相似，但在自监督学习模型的使用方式上有所不同：该模型直接作为分词器进行训练，以噪声语音信号为输入来估计 $\mathbf{u}_{\text{enh}}$。我们在自监督学习模型之上添加单个线性层，并使用连接时序分类损失对整个网络进行微调，这与采用自监督学习模型初始化的自动语音识别系统中的常见做法相同 [26–28]。请注意，与 V2T-E 一样，我们也研究了用更强大的分类器替代线性层，但性能几乎没有变化。W2T-E 设计的训练成本在令牌级增强方法中最高，因为自监督学习模型本身也是系统训练的一部分。不过，由于它不会生成任何中间表示（即词元或连续向量），因此整体系统最简单，在所考虑的前端中推理成本也最低。

4. 实验设置

4.1. 评估指标和数据集

我们通过词错误率（WER）评估自动语音识别（ASR）性能，以此比较增强前端。连续语音识别使用波形级增强（即 W2W-E）进行评估，而词元语音识别则使用波形级和令牌级增强（即 T2T-E、V2T-E 和 W2T-E）进行评估。为了直接评估增强输出，我们还分别使用尺度不变信噪比（SI-SNR）和单元编辑距离（UED）来评估波形级增强和令牌级增强。此处，UED 定义为

[1]我们的初步实验也表明，令牌级增强任务中存在类似趋势。

去重参考序列（干净）与含噪/增强后的词元序列之间的Levenshtein距离，并以去重参考词元序列的长度 [12] 进行归一化。

所有实验均在 CHiME-4数据集 [7]上进行。我们采用 ESPnet 的配置，并使用相应的配方训练连续语音识别、词元语音识别和增强模型。[2] 我们仅报告单通道数据集上的性能。[3] 此外，我们使用评估子集 et05 simu 中的纯净语音报告自动语音识别系统的上界性能（下文简称为“干净”）。计算 UED 时，我们使用模拟数据集（即 {dt05,et05} simu）。为简洁起见，下文省略数据集名称中的“05”。

4.2. 模型与训练

后端语音识别系统： 我们训练了一个连续语音识别模型 作为基线，并训练了一个词元语音识别模型。对于这两种自动语音识别模型，我们采用联合CTC/注意力编码器-解码器模型（以下简称 AED）。这些模型由 12 个 E-Branchformer 编码器块和 6 个 Transformer解码器块组成，配置遵循 ESPnet 的 asr2 配方。我们还使用了具有相同编码器架构的纯CTC词元语音识别模型，以检验增强后离散单元能否在不同的自动语音识别系统之间复用。

连续语音识别模型使用 80 维 FBANK特征或 WavLM Large 的加权和特征 [28] 作为输入。词元语音识别以 2k 个字节对编码单元为输入，这些单元由 1k 个簇得出；这些簇是通过对 WavLM Large 第 21 层的隐藏表示进行 k-均值聚类获得的。词元语音识别的嵌入维度设为 512。对于编码器中的卷积下采样，基于 FBANK 的自动语音识别模型采用两层带 ReLU 的二维卷积（将序列长度缩短至四分之一）。其他自动语音识别模型则采用两层带 ReLU 的一维卷积（保持序列长度不变）。其他配置遵循 ESPnet 的默认设置。

为建立更强的基线，我们进一步微调了基于 WavLM 的连续语音识别模型（包括 WavLM），并使用训练好的 AED 的参数进行初始化。整个模型以 5e-5 的学习率进行联合微调，采用 reduce-on-plateau 调度器，并使用早停（耐心值为 5）。请注意，微调后的WavLM模型失去了下游无关属性，因为其输出特征不再能用于其他自动语音识别模型。

此外，为了与现有系统进行比较，我们使用根据 ESPnet 配置训练的 16 层 Transformer 语言模型对自动语音识别系统进行了解码。

前端增强系统： 对于 W2W-E，我们重点研究了两种广泛使用的语音增强模型：Conv-TasNet [23] 和 TF-GridNet [10]。构建这些模型时，我们遵循了 ESPnet 中提供的配置。

对于 T2T-E（第3.1节），映射网络采用了四个 E-Branchformer 块，除特征维度缩减至 256 外，其余设置与上文所述的 ASR编码器相同。对于嵌入层，我们根据初步实验 [4] 采用了与词元语音识别相同的配置（总计 9.20M 个参数）。

对于采用多层感知机的 V2T-E（第3.2节），我们使用了两个维度为 512 的线性层，并采用 LeakyReLU 激活函数，其架构与 [12] 中的架构相似（共 1.30M 个参数）。对于采用时序卷积网络的 V2T-E，其架构由三个时序卷积网络模块组成。每个模块包含八个一维卷积层，卷积核大小为 3，隐藏特征维度为 512，瓶颈维度为 128（共 3.95M 个参数）。采用 E-Branchformer 的 V2T-E 使用了与 T2T-E 映射网络相同的架构（共 9.08M 个参数）。请注意，我们报告的 V2T-E 结果仅采用所有自监督学习层的加权和特征，因为与词元化所用单层的特征相比，这些特征通常能提升性能。

对于 W2T-E（第 3.3 节），模型使用 Layer- Drop [29] 以 0.1 的比率进行训练；训练期间始终冻结卷积特征编码器，并在前 15k 步中冻结 Transformer编码器[5]；可训练参数数量为 311.74M。

所有词元级前端模型均训练了 30 个 epoch，批量大小为 16。在学习率调度方面，除采用多层感知机和时序卷积网络的 V2T-E 使用平台期衰减调度器外，其他模型均使用预热调度器，预热步数为 25k。优化器为 Adam，T2T-E 的学习率为 5e-3，采用多层感知机的 V2T-E 为 1e-4，采用时序卷积网络的 V2T-E 为 1e-3，采用 E-Branchformer 的 V2T-E 为 5e-3，W2T-E 为 1e-4。

表1：CHiME-4 [7] 的模拟和真实数据集上使用各种前端的评估结果，以词错误率（%）和 UED 表示。结果是在不使用语言模型的情况下获得的。

ID 系统	WER dt_ 模拟	WER dt_ 真实	WER et_ 模拟	WER et_ 真实	WER et_ 干净	UED dt_ 模拟	UED et_ 模拟
连续语音识别（AED）							
B1 FBANK	18.0	15.1	25.1	23.0	7.4	-	
B2 WavLM（加权求和）	8.1	6.0	11.0	6.8	**1.5**	-	
B3 +W2W-E（Conv-TasNet）	9.1	6.0	17.2	13.7	**1.5**	-	
B4 +W2W-E（TF-GridNet）	5.9	**3.8**	11.2	8.2	**1.5**	-	
词元语音识别（AED）							
A1 WavLM	17.3	12.6	18.6	13.5	3.4	63.6	65.7
A2 + W2W-E（Conv-TasNet）	12.9	9.5	21.6	19.3	3.4	46.3	55.3
A3 +W2W-E（TF-GridNet）	9.2	6.7	15.1	12.4	3.2	42.1	48.7
A4 +T2T-E	17.0	12.1	18.6	9.4	3.5	37.4	39.3
A5 +V2T-E（多层感知机）	11.5	8.8	14.7	10.0	3.5	34.6	37.1
A6 +V2T-E（时序卷积网络）	10.4	8.0	13.7	9.4	3.6	32.0	34.5
A7 +V2T-E（E-Branchformer）	9.8	7.7	13.6	8.9	3.3	30.8	33.5
A8 +W2T-E	**5.6**	4.5	**8.2**	**6.5**	3.4	**27.2**	**29.2**
词元语音识别（仅连接主义时序分类）							
C1 WavLM	21.9	16.2	23.6	17.3	5.1	63.6	65.7
C2 +W2T-E	6.6	6.1	9.9	8.2	4.6	**27.2**	**29.2**

5. 结果

5.1. 性能比较

表1总结了评估结果。对于基线连续语音识别模型，采用传统语音增强前端（B3、B4）在某些情况下改善了性能，但正如先前研究 [30]所示，语音增强系统引入的伪影可能会对识别产生不利影响。在 et simu 上，Conv-TasNet 和 TF-GridNet 分别取得了 11.38 dB 和 13.17 dB 的尺度不变信噪比。

对于词元语音识别，未进行增强的注意力编码器-解码器模型（A1）优于基于滤波器组特征的连续语音识别（B1），但不及基于 WavLM 的连续语音识别（B2），这与先前报告的结果 [1]一致。这一趋势表明自监督学习模型提供的预训练知识十分重要。通过比较 et simu 与干净条件下的词错误率，我们还确认，词元语音识别（A1）与连续语音识别系统（B1、B2）类似，对噪声敏感。

对于采用 W2W-E（A2、A3）的词元语音识别，Conv-TasNet 有时会增加词错误率，而 TF-GridNet 则始终能降低词错误率。这一结果表明，要达到一定的尺度不变信噪比，

[2] espnet/egs2/chime4/{asr1,asr2,enh1}[3]{dt05,et05} { simu,real}isolated 1ch track[4]我们还测试了使用 k-均值聚类的质心初始化嵌入层，但性能差异并不显著。

[5]这些冻结技术尤其改善了词错误率和 UED，而类似 SpecAugment 的掩蔽 [26–28] 则未能改善任何一项指标。

表2：与现有系统相比，单通道数据集上的词错误率（%）。所有系统均使用语言模型进行解码。现有系统的结果取自原论文 [11]。

ID	系统	dt_simu	dt_real	et_simu	et_real	干净
E1	IRIS [11]	3.2	**2.0**	6.1	3.9	-
B2	WavLM（加权求和）	5.4	3.7	8.0	4.4	**0.8**
B4	＋W2W-E（TF-GridNet）	4.1	2.3	8.2	5.5	**0.8**
B5	+联合微调 WavLM	**3.1**	**2.0**	**5.6**	**3.5**	1.0
A8	WavLM（词元）＋W2T-E	3.2	2.3	6.1	4.0	1.6

W2W-E 始终能提高词元语音识别的准确率。有趣的是，这与上文所述连续语音识别的趋势（B3、B4）相反。

对于 T2T-E (A4)，其前端的 UED 表现更好，但在大多数情况下，词错误率比基于 TF-GridNet 的 W2W-E (A3) 更差（详细分析请参见第 5.2 节）。

对于 V2T-E，即使使用基于多层感知机的模型 (A5)，其词错误率和 UED 表现也大多优于 T2T-E；采用更强大的架构还能进一步提升性能 (A6, A7)。值得注意的是，尽管 TCN (A6) 的可训练参数更少（3.95M 对比 9.08M），其性能仍与 E-Branchformer (A7) 相当，凸显了它的参数效率。

最后，W2T-E (A8) 在噪声条件下的词错误率和 UED 综合表现最佳。除 dt real 外，其词错误率优于基于 WavLM 的连续语音识别系统 (B2, B4)，同时由于输入序列更短，计算成本也更低。具体而言，与 WavLM 的输出帧相比，采用字节对编码的 W2T-E 将 et real 和 et simu 的平均词元序列长度分别缩短了 68.8% 和 68.2%。此外，包括 W2T-E 在内的增强前端在干净条件下均未出现明显性能下降，这表明它们可以可靠地替代基于 k-means 的分词器。值得注意的是，W2T-E 只需额外增加一个线性层，因此推理时的计算开销极小。这些结果表明，要充分发挥自监督学习模型的能力，仅仅添加一个固定的分词器模块（例如 k-means 词元化）是不够的。

此外，为了验证前端的模块化特性，我们使用纯CTC词元语音识别后端（C1、C2）评估了所提出的 W2T-E。C2 系统的结果证实，即使使用不同的后端，该系统仍取得了显著且稳定的改进。

为明确我们的方法在 CHiME-4 单通道数据集中的定位，我们将基于语言模型的结果与同样使用 WavLM Large 的现有系统 IRIS [11] 的结果进行了比较。

IRIS 是一个端到端优化的系统，它将 Conv-TasNet 与基于冻结的 WavLM 特征的注意力编码器-解码器模型相结合，达到了最先进的性能。如表2所示，采用 W2T-E（A8）的词元语音识别性能与 IRIS（E1）相当，同时受益于词元语音识别的高效性。值得注意的是，通过对 B2 继续进行联合自监督学习-自动语音识别微调得到的 B5，取得了整体最佳性能。这表明，将令牌级增强与词元语音识别联合优化，或许能进一步降低词错误率。不过，这超出了本文的讨论范围，因为此类增强不适用于任意离散单元自动语音识别系统。

5.2. UED 与词错误率之间的关系

根据表1中的 A3 和 A4 结果，我们通过逐话语计算这两项指标，对 UED 与词错误率之间缺乏相关性的问题进行了深入分析。图2根据增强后离散单元中 UED 和词错误率的变化，将话语分为四类；这些变化均以

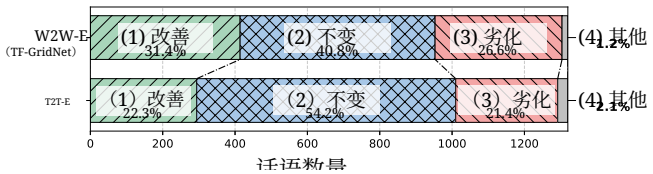


图 2：按 UED-词错误率变化组统计的语句数量（模拟 et simu）。

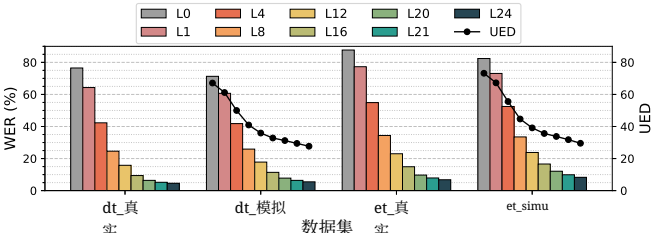


图 3：W2T-E 中词错误率（柱形）和 UED（黑线）随自监督学习模型深度的变化。L0 表示 WavLM Large 的卷积特征编码器输出。

原始 et simu 噪声词元为参照：（1）UED 和词错误率均改善；（2）UED 改善，而词错误率不变；（3）UED 改善，而词错误率劣化；（4）UED 不变或劣化（“其他”）。图2显示，尽管 T2T-E 的平均 UED 更好，但与基于 TF-GridNet 的 W2W-E 相比，其词错误率不变的话语数量显著更多。这一观察表明，词元语音识别对噪声引起的词元序列变化具有一定的鲁棒性，至少在 T2T-E 所涵盖的范围内如此。总之，UED 是评估词元化器抗噪鲁棒性的有用指标，但解读时应谨慎，因为基于词元的后端处理变化的固有能力也起着关键作用。

5.3. 自监督学习模型深度对W2T-E的影响

在消融实验中，我们进一步研究了改变W2T-E中自监督学习深度的影响。在训练W2T-E时，我们考察了将CTC线性层连接到中间WavLM层而非最终层是否能达到相当的准确率，因为分词最初基于第21层。若可行，这种设计将产生计算效率更高的分词器。为确保公平比较，本实验未应用LayerDrop，因为它不适用于零层条件，且其性能增益有限。图3展示了各数据集的词错误率和UED随W2T-E模型深度的变化。结果表明，减少层数并不可行，因为将层数从24减至21会导致et real和et simu上的词错误率分别相对恶化16.2%和19.3%。一种可能的解释是，除了编码音素或语言表征外，确保对噪声的鲁棒性可能也需要额外的层。

6. 结论

本文介绍了面向词元语音识别的前端增强，并探讨了噪声鲁棒性的有效性。实验结果表明，在噪声条件下，W2T-E 能显著提升词元语音识别的性能，优于基于连续向量的语音识别，同时只需额外添加一个轻量级分类器。未来，我们计划将评估扩展到其他后端系统，即语音识别之外的语音任务，尤其是借助语音语言模型进一步考察后端无关特性。

7. 参考文献

[1] X. Chang、B. Yan、K. Choi、J.-W. Jung、Y. Lu、S. Maiti、R. Sharma、J. Shi、J. Tian、S. Watanabe、Y. Fujita、T. Maekaku、P. Guo、Y.-F. Cheng、P. Denisov、K. Saijo 和 H.-H. Wang，《使用离散语音单元探索语音识别、翻译与理解：一项比较研究》，载于 *ICASSP*，2024年。[2] X. Chang、J. Shi、J. Tian、Y. Wu、Y. Tang、Y. Wu、S. Watanabe、Y. Adi、X. Chen 和 Q. Jin，《Interspeech 2024 离散单元语音处理挑战赛》，载于 *Interspeech*，2024年。[3] Y. Yang、F. Shen、C. Du、Z. Ma、K. Yu、D. Povey 和 X. Chen，《迈向通用语音离散词元：语音识别与语音合成案例研究》，载于 *ICASSP*，2024年。[4] S. Arora、K.-W. Chang、C.-M. Chien、Y. Peng、H. Wu、Y. Adi、E. Dupoux、H.-Y. Lee、K. Livescu 和 S. Watanabe，《口语语言模型概述：全面综述》，*arXiv:2504.08528*，2025年。[5] A. Lee、P.-J. Chen、C. Wang、J. Gu、S. Popuri、X. Ma、A. Polyak、Y. Adi、Q. He、Y. Tang、J. Pino 和 W.-N. Hsu，《基于离散单元的直接语音到语音翻译》，载于 *ACL*，2022年。[6] P. Gage，《一种新的数据压缩算法》，*C Users Journal*，1994年。[7] E. Vincent、S. Watanabe、A. A. Nugraha、J. Barker 和 R. Marxer，《鲁棒语音识别中的环境、麦克风与数据仿真失配分析》，计算机语音与语言，第46卷，第535–557页，2017年。[8] S. Cornell、M. S. Wiesner、S. Watanabe、D. Raj、X. Chang、P. Garcia、Y. Masuyama、Z.-Q. Wang、S. Squartini 和 S. Khudanpur，《CHiME-7 DASR 挑战赛：多样化场景下使用多台设备进行远场会议转写》，载于 *CHiME*，2023年。[9] C. Boeddecker、J. Heitkaemper、J. Schmalenstroeer、L. Drude、J. Heymann 和 R. Haeb-Umbach，《面向 CHiME-5 晚宴场景的前端处理》，载于*CHiME*，2018年。[10] Z.-Q. Wang、S. Cornell、S. Choi、Y. Lee、B.-Y. Kim 和 S. Watanabe，《TF-GridNet：融合全频带与子带建模的语音分离》，*TASLP*，第31卷，第3221–3236页，2023年。[11] X. Chang、T. Maekaku、Y. Fujita 和 S. Watanabe，《语音识别、语音增强与自监督学习表征的端到端集成》，载于 *Interspeech*，2022年。[12] I. Gat、F. Kreuk、T. Anh Nguyen、A. Lee、J. Copet、G. Synnaeve、E. Dupoux 和 Y. Adi，《生成式口语语言建模中的增强不变离散表征》，载于 *IWSLT*，2023年。[13] A. Lee、H. Gong、P.-A. Duquenne、H. Schwenk、P.-J. Chen、C. Wang、S. Popuri、Y. Adi、J. Pino、J. Gu 和 W.-N. Hsu，《真实数据上的无文本语音到语音翻译》，载于 *NAACL*，2022年。[14] S. Messica 和 Y. Adi，《NAST：面向语音语言模型的噪声感知语音分词》，载于 *Interspeech*，2024年。[15] T. A. Nguyen、M. d. Seyssel、P. Roz´e、M. Rivi`ere、E. Kharitonov、A. Baevski、E. Dunbar 和 E. Dupoux，《The 零资源语音基准 2021：无监督口语语言建模的指标与基线”，发表于*NeurIPSWorkshopSelf-Supervised Learning Speech Audio Processing*，2020。[16] K. Qian、Y. Zhang、H. Gao、J. Ni、C.-I. Lai、D. Cox、M. Hasegawa-Johnson 和 S. Chang，“ContentVec：通过分离说话人因素改善的自监督语音表征”，发表于*ICML*，2022。[17] H.-J. Chang、A. H. Liu 和 J. Glass，“通过说话人不变聚类进行自监督微调，以改善内容表征”，发表于*Interspeech*，2023。[18] Z. Wang、X. Zhu、Z. Zhang、Y. Lv、N. Jiang、G. Zhao 和 L. Xie，“SELM：使用离散词元和语言模型进行语音增强”，发表于*ICASSP*，2024。[19] B. Kang、X. Zhu、Z. Zhang、Z. Ye、M. Liu、Z. Wang、Y. Zhu、G. Ma、J. Chen、L. Xiao、C. Weng、W. Xue 和 L. Xie，“LLaSE-G1：提升基于 LLaMA 的语音增强的泛化能力”，发表于*ACL*，2025。[20] P. Mousavi、L. Della Libera、J. Duret、A. Ploujnikov、C. Sub- akan 和 M. Ravanelli，“DASB——离散音频与语音基准”，*arXiv:2406.14294*，2024。[21] K. Kim、F. Wu、Y. Peng、J. Pan、P. Sridhar、K. J. Han 和 S. Watanabe，“E-Branchformer：采用增强型融合的 Branchformer，用于语音识别”，发表于*SLT*，2023。[22] C. Lea、R. Vidal、A. Reiter 和 G. D. Hager，“时序卷积网络：动作分割的统一方法”，发表于*ECCV*，2016。[23] Y. Luo 和 N. Mesgarani，“Conv-TasNet：超越用于语音分离的理想时频幅度掩蔽”，*TASLP*，第 27 卷，第 8 期，第 1256–1266 页，2019。[24] S. w. Yang、P.-H. Chi、Y.-S. Chuang、C.-I. J. Lai、K. Lakhotia、Y. Y. Lin、A. T. Liu、J. Shi、X. Chang、G.-T. Lin、T.-H. Huang、W.-C. Tseng、K. t. Lee、D.-R. Liu、Z. Huang、S. Dong、S.-W. Li、S. Watanabe、A. Mohamed 和 H. y. Lee，“SUPERB：语音处理通用性能基准”，发表于*Interspeech*，2021。[25] K. Choi、M. Someki、E. Strubell 和 S. Watanabe，“设备端流式离散语音单元”，发表于*Interspeech*，2025。[26] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：自监督学习语音表征的框架”，发表于*NeurIPS*，2020。[27] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhut- dinov 和 A. Mohamed，“HuBERT：通过隐单元掩码预测进行自监督语音表征学习”，*TASLP*，第 29 卷，第 3451–3460 页，2021。[28] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao、J. Wu、L. Zhou、S. Ren、Y. Qian、Y. Qian、J. Wu、M. Zeng、X. Yu 和 F. Wei，“WavLM：面向全栈语音处理的大规模自监督预训练”，*JSTSP*，第 16 卷，第 6 期，第 1505–1518 页，2022。[29] G. Huang、Y. Sun、Z. Liu、D. Sedra 和 K. Q. Weinberger，“具有随机深度的深度网络”，发表于*ECCV*，2016。[30] T. Ochiai、K. Iwamoto、M. Delcroix、R. Ikeshita、H. Sato、S. Araki 和 S. Katagiri，“重新思考处理失真：分离语音增强错误对语音识别性能的影响”，*TASLP*，第 32 卷，第 3589–3602 页，2024。