arXiv:2303.00978v1 [cs.CL] 2 Mar 2023

利用大型　　文本语料库进行端到端语音摘要

Kohei Matsuura、*Takanori Ashihara*、*Takafumi Moriya*、*To mohiro Tanaka*、*Atsunori Ogawa*、*Marc Delcroix*、*Ryo Masumura*

日本电信电话株式会社，日本

摘要

端到端语音摘要（E2E SSum）是一种直接从语音生成摘要句的技术。与结合自动语音识别（ASR）和文本摘要模型的级联方法相比，端到端方法更具优势，因为它能够减轻自动语音识别错误、纳入非言语信息，并简化整个系统。然而，由于难以收集大量配对数据（即语音和摘要），训练数据通常不足以训练出稳健的端到端语音摘要系统。本文提出两种新颖的方法，利用大量外部文本摘要数据进行端到端语音摘要训练。第一种方法是利用文本转语音（TTS）系统生成合成语音，并将其与文本摘要一起用于端到端语音摘要训练。第二种方法无需文本转语音，而是直接向端到端语音摘要模型输入音素序列，以替代合成语音。实验表明，我们提出的文本转语音方法和基于音素的方法在 How2 数据集上的多项指标均有所提升。尤其是，我们的最佳系统大幅超越了此前的最先进系统（即 METEOR 分数提高了 6 分以上）。据我们所知，这是首项将外部语言资源用于端到端语音摘要的研究。此外，我们还对 How2 数据集进行了详细分析，以验证所提出的端到端语音摘要系统的有效性。

索引词— 端到端语音摘要、合成数据增强、多模态数据增强、How2 数据集

1. 引言

随着深度学习的引入，自动语音识别（自动语音识别）技术取得了显著进展 [1]。尽管取得了成功，自动语音识别转写文本的可读性并不高，因为其中通常包含填充词、非流利现象或冗余表达 [2]。为了生成更易读、信息更丰富的文本，语音摘要（SSum）技术已在会议 [3]、教育视频 [4], 以及医患对话 [5]等领域受到广泛关注。

通常，语音摘要采用级联方法实现，该方法结合了自动语音识别和文本摘要（文本摘要）模型，如图 1(a) 所示。级联语音摘要取得了较高的性能，由高精度自动语音识别模型和在大量非配对文本数据上预训练的文本摘要模型组成的组合系统 [6,7]。尽管级联方法取得了成功，但它存在自动语音识别错误传播的问题，也缺乏非语言信息和声学信息 [8]。为解决这些问题， [9] 将自动语音识别的 N-best 假设输入文本摘要模型，以减轻自动语音识别错误的影响； [10] 则在输入文本中加入了声学特征。

近年来，也有人提出端到端语音摘要（E2E SSum）来解决上述问题。E2E SSum 系统类似于端到端自动语音识别（E2E ASR）系统，联合优化声学模型和语言模型，并生成抽象式摘要

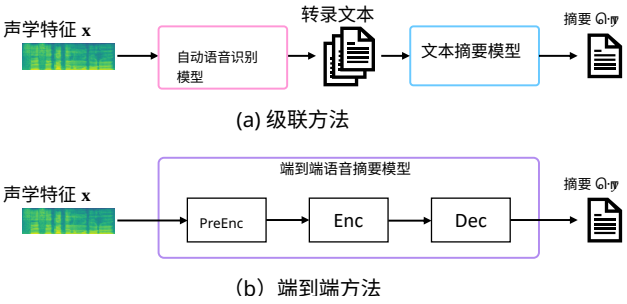

图 1. 语音摘要中（a）级联方法与（b）端到端（E2E）方法的比较。

如图1(b)所示，直接通过单个模型生成摘要。由于该方法不以文本作为中间输出，模型可以充分利用声学信息来生成摘要，并且可能不受自动语音识别错误的影响。 [11] 采用了这种方法，并结合受限自注意力机制，报告称其性能优于级联方法。

然而，端到端语音摘要模型需要大量成本高昂的语音-摘要对，而大多数语音摘要任务可用的配对数据有限。我们认为，如同其他端到端语音处理领域一样，丰富的外部语言资源有望提升端到端语音摘要模型的性能 [12–18]。

本文尝试利用文本摘要数据，即 Giga- word [19,20], ，使模型能够学习更好的语言表征。我们提出了一种基于文本到语音（TTS）的语音数据增强方法，通过将外部文本摘要数据集中的文本合成为语音，获得用于端到端语音摘要训练的伪语音与摘要配对数据。尽管这种方法应用起来相当简单，但合成伪语音需要很高的计算成本。为消除这些开销，我们还提出了一种受[16, 17]启发、通过添加音素预编码器来扩展模型的方法。当输入音素序列而非合成语音时，音素预编码器与语音预编码器分离。与基于 TTS 的数据增强相比，该方法所需的计算成本和存储空间都要低得多。

所有实验均在 How2 数据集 [21] 上进行，该数据集由简短的视频片段及其摘要组成。我们对语料库进行了深入分析，发现部分训练数据与评估数据非常相似，这可能会使结果产生偏差。因此，我们提出了一种评估方案，利用定量指标筛除这类数据。采用所提出的评估方案后，我们能够更有把握地评估摘要性能。

我们的主要贡献如下：

1. 我们最好的系统在 How2 摘要任务上达到了最先进的性能。

2. 我们证实，外部文本摘要数据结合文本转语音技术能够提升端到端语音摘要模型的性能。3. 我们提出了一种基于音素的数据增强方法，相较于基于文本转语音的方法，计算成本更低。4. 我们指出，对 How2 数据集的评估集进行筛选，对于更合理地评估模型的摘要能力至关重要。

2. 相关工作

以往研究已证明，利用外部文本数据对端到端自动语音识别和端到端语音翻译系统有效。对于端到端自动语音识别系统， [12] 在领域自适应场景中使用 Tacotron2 [22] 从外部文本数据合成声学特征，生成了额外的语音—转录文本数据。其他研究则利用 vq-wav2veq 提取的离散语音表征 [13], 、端到端自动语音识别编码器的潜在表征 [14], 以及音素序列 [16, 17]来使用文本数据。此外, [23]提出了一种名为“语音链”的方法，通过额外的非配对文本和语音联合优化文本转语音和自动语音识别模型。 [24] 通过引入一致性损失扩展了该方法。类似思路也被用于为端到端语音翻译系统生成训练数据 [18]。本研究受到这些成果的启发，尝试将数据增强方法应用于端到端语音摘要系统。

3. 端到端语音摘要模型

如图 1 所示，端到端语音摘要系统将自动语音识别和文本摘要模型整合为一个模型，而级联系统则通过转录文本处理输入语音。具体而言，本文采用基于 Transformer 的注意力编码器-解码器（AED）模型 [1] 。首先，预编码器 PreEnc 将声学特征$\mathbf{x}$ 嵌入为经过子采样的隐藏表征序列$\mathbf{h}^{\mathrm{pre}}$:

$$\mathbf{h}^{\mathrm{pre}} = \mathrm{PreEnc}(\mathbf{x}), \tag{1}$$

其中，PreEnc(·) 由二维卷积神经网络（CNN）组成。其次，编码器 Enc(·) 根据$\mathbf{h}^{\mathrm{pre}}$计算潜在表征$\mathbf{h}^{\mathrm{enc}}$:

$$\mathbf{h}^{\mathrm{enc}} = \mathrm{Enc}(\mathbf{h}^{\mathrm{pre}}, \mathbf{R}), \tag{2}$$

其中，Enc(·) 由 Conformer [25] 模块构成，可在自动语音识别任务中实现高准确率；$\mathbf{R}$ 表示相对位置嵌入。最后，解码器 Dec(·) 接收 $\mathbf{h}^{\mathrm{enc}}$ 和先前的输出 $\{\hat{y}_i\}_{i=1}^{l-1}$ 作为线索，以自回归方式推断每个词元:

$$\mathbf{h}_l^{\mathrm{dec}} = \mathrm{Dec}([\mathrm{Emb}(\hat{y}_i) + \mathrm{PE}_i]_{0\le i\le l-1}, \mathbf{h}^{\mathrm{enc}}), \tag{3}$$

$$\hat{\mathbf{y}}_l = \mathrm{softmax}(\mathbf{h}_l^{\mathrm{dec}}), \tag{4}$$

其中，softmax(·) 是 softmax 激活函数，PE_i 是第 i 个绝对位置编码 [26]。解码器 Dec(·) 由 Transformer 模块组成，而词元嵌入模块

Emb(·) 是一个可学习的线性层。整个摘要句子$\hat{y}$ 通过束搜索技术进行估计。在推理开始时，解码器接收特殊符号 <sos>，表示句子的开始，并持续输出标记，直到估计出特殊的结束标记 <eos> 。训练期间，端到端语音摘要模型通过 $\hat{\mathbf{y}}_l$ 与 y_l的独热向量之间的交叉熵损失进行优化。为了稳定模型训练，公式（3）中使用真实标记 y_i 代替$\hat{y}_i$。

除上述架构外，我们还对端到端语音摘要模型进行了以下两项修改。首先，我们将

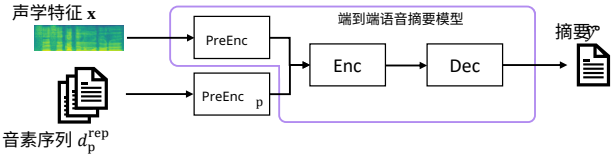

图. 2 带有音素预编码器 Pr 的端到端语音摘要模型 eEncP。

Conformer模块中的批归一化（BN）替换为层归一化（LN）。由于语音摘要系统的输入序列很长，批大小过小，训练期间难以学习到泛化能力良好的BN层统计量。其次，我们建议将式（3）中的绝对位置编码替换为可学习的位置嵌入（LPE） [27] , 具体如下:

$$\mathbf{h}_l^{\mathrm{dec}} = \mathrm{Dec}([\mathrm{Emb}(\hat{y}_i) + \mathrm{LPE}(i)]_{0\le i\le l-1}, \mathbf{h}^{\mathrm{enc}}), \tag{5}$$

其中，可学习的线性层 LPE(i) 对第 i 个标记的位置进行嵌入。鉴于 LPE 已成功应用于自然语言处理领域 [27, 28], , 我们预计 LPE 也能提升摘要任务的性能。

4. 提出的方法

4.1. 多阶段训练

我们的端到端语音摘要模型接收三个训练数据集：语音与转录文本的配对数据（$\mathbf{x}$ y_{asr}） $\in \mathcal{R}_{\mathrm{ASR}}$、语音与摘要的配对数据（$\mathbf{x}$ y） $\in \mathcal{R}_{\mathrm{Sum}}$，以及外部文本摘要数据（$d$， y） $\in \mathcal{R}_{\mathrm{Ext}}$。这里， d 是文本摘要的输入文本。为利用多种知识，我们采用了由多个阶段组成的训练策略，具体如下:

(i) 使用$\mathcal{R}_{\mathrm{ASR}}$ 中的配对数据，将注意力编码器-解码器模型训练为自动语音识别模型，使其在内部获得转录能力。(ii) 为构建端到端语音摘要模型，使用 $\mathcal{R}_{\mathrm{Sum}}$对上述模型进行额外微调。(iii) 使用由$\mathcal{R}_{\mathrm{Sum}}$ 和从 $\mathcal{R}_{\mathrm{Ext}}$衍生的人工数据组成的组合数据集，进一步微调该模型。

我们将阶段（ii）后的模型定义为基线模型，将阶段（iii）后的模型定义为通过文本转语音或基于音素的方法增强的模型。

4.2. 基于文本转语音和音素的数据增强

为了将文本摘要数据用于端到端语音摘要，我们提出了两种方法：基于文本转语音和基于音素的数据增强。对于基于文本转语音的方法，我们根据文本摘要数据集的输入文本 d 合成原始波形，并提取声学特征用于阶段 (iii)。

尽管这种基于文本转语音的方法可以直接用于扩充语音摘要数据，但合成语音需要很高的计算成本。因此，我们还提出了一种替代方法，即将音素序列输入端到端语音摘要模型。该方法的灵感来自 [16], 中的性能提升，该研究提出了一种用于端到端自动语音识别的多模态（即声学和文本）数据增强技术。具体而言，首先利用字素到音素模型，将词序列 d 自动转换为音素序列 d_p 。然后，根据单独计算的统计数据，通过重复 d_p 中的每个音素来建模音素时长。重复后的音素序列记为 d_p^{rep}。为了将 d_p^{rep}与 $\mathbf{x}$ 一同输入，我们还添加并优化了音素预编码器

表1。先前工作的结果（P-1）、 [11], 我们的基线（C-1、B-1、B-2）以及提出的模型（T-1、2）。

	ID	模型	增强数据	#参数	ROUGE-1、2、L（↑）	METEOR（↑）	BERTScore（↑）
基线	C-1	级联	-	201M+140M	60.9, 43.3, 55.4	30.3	92.6
	P-1	[11]		104M	60.9, 43.0, 55.9	28.8	91.0
	B-1	base		98M	65.3, 51.4, 62.1	32.5	93.0
	B-2	大型		203M	65.6, 50.9, 62.0	32.8	93.2
提出的	T-1	大型	TTS	203M	**68.4**,**54.1**, **65.0**	**34.9**	**93.8**
	T-2	大型	音素	203M(+7M)[1]	67.4, 53.2, 64.1	33.9	93.6

$\text{PreEnc}_{\text{P}}(\cdot)$，并在 PreEnc 之外同时作为

$$\mathbf{h}^{\text{pre}} = \begin{cases} \text{PreEnc}(\mathbf{x}) \\ \text{PreEnc}_{\text{p}}(d_p^{\text{rep}}). \end{cases} \tag{6}$$

如图 2 所示，两个输入 $\mathbf{x}$ 和 d_p^{rep}共享其他模块。这里，我们选择双向长短期记忆网络，后接一个层归一化层和用于 PreEnc_{P} 的子采样卷积神经网络。

5. 实验

5.1. 数据集

我们使用了 How2 数据集，该数据集包含约 2,000 小时的语音。由于其语音输入相对较短，且抽象式摘要的压缩率较高，因此适用于端到端语音摘要任务。训练集、验证集和评估集分别包含 68,336、1,600 和 2,127 对样本，其口语文档的平均时长分别为 84.7、76.0 和 98.7 秒。输入语音截取至最长 100 秒，由 40 维对数梅尔滤波器组输出和 3 维基频特征组成。根据 [21], , 摘要是从 YouTube 上的“description”标签自动收集的。

在数据增强方面，我们使用了 Gigaword语料库 [19, 20] 3 8 ，其中包含 . M 对新闻文章首句及其标题。该数据集的输入文本和摘要平均词数分别为 31.4 和 8.3。由于平均输入长度远短于其他文本摘要语料库（例如 CNN Daily Mail [29], ），Gigaword语料库的合成成本相对较低。尽管这一外部语料库的领域与 How2 数据集不匹配，但我们预计模型能够学习更多词汇和表达方式。

5.2. 训练与评估

模型架构采用了一个 12 层的 Conformer编码器和一个 6 层的 Trans- former 解码器。预编码器由 4 层卷积神经网络组成，子采样率为 4。我们评估了两种模型规模：基础版和大型模型，其嵌入维度分别为 512 和 768。两者的编码器均包含 2048 个前馈（FF）单元、8 个注意力头，卷积核大小为 31。基础版模型的解码器包含 2048 个 FF 单元和 4 个注意力头；大型模型的解码器则包含 3072 个 FF 单元和 12 个注意力头。

在第4.1节所述的训练阶段（i）中，我们使用Adam优化器 [30] 和Noam调度器[26] 训练端到端自动语音识别模型，学习率为 $2 \cdot 10^{-3}$，预热步数为40k，权重衰减率为 10^{-5}，批大小为512。基础版和大型模型在评估集上的词错误率（↓）分别为9.8%和9.5 %，束大小为16。

在训练阶段（ii），模型经过微调，成为端到端语音摘要模型，学习率为 10^{-4}，衰减因子为 0.5，批量大小为 30。

[1]自动语音识别模型在推理过程中不需要音素编码器（7M） e.

在训练阶段（iii），我们使用增强数据对模型进行了进一步微调。对于文本转语音模型，我们使用了基于 LJSpeech语料库 [2]训练的 VITS [31] 。由于 LJSpeech语料库只包含一位说话人的语音，合成语音不具有说话人多样性。这使我们能够仅评估额外语言学信息的有效性，而排除说话人多样性的增强所带来的影响。生成原始波形后，我们从波形中提取了 43 维声学特征，窗口大小为 25 ms、帧移为 10ms，并应用倒谱均值方差归一化（CMVN），使其与 How2语料库中的原始特征相匹配。在阶段（iii）训练模型时，我们采用的学习率为 10^{-4}，衰减因子为 0.5，解码器的学习率为 10^{-3}。每个批次中输入序列的最大总长度设为 300,000。每个批次仅包含真实样本或人工生成的样本。

对于基于音素的方法，我们使用g2pE工具包[3] 将词序列转换为音素序列。音素单元的数量为42。每次音素 p 出现时，其持续时间均按照正态分布 $\mathcal{N}$（μ_p，σ_p）确定。其中，p的均值 μ_p和标准差 σ_p使用TIMIT语料库 [32]进行估计。音素预编码器由3层双向长短期记忆网络组成，后接层归一化和4层卷积神经网络，子采样率为4。音素预编码器的学习率为 10^{-3}。其他配置与基于文本转语音的方法相同。在阶段（i）、（ii）和（iii）中，我们使用字节对编码，并将词汇量设为1,000。

评估模型性能时，我们选择验证准确率最高的检查点，并在所有条件下使用束宽为 4 的解码器进行解码。评估指标方面，我们选用了文本摘要任务中常用的 ROUGE [33], METEOR [34], 和 BERTScore [35] 分数。

我们还评估了传统的级联模型。自动语音识别模型与训练阶段 (i) 后得到的大型模型相同。文本摘要模型是一个基于 BART 的模型，使用 CNN/Daily Mail语料库[4] [29]和 How2 数据集进行了微调。使用 How2 数据集进行微调时，我们使用 Adam优化器训练模型 1M 步，学习率为 5×10^{-5} 并采用线性衰减，批大小为 8。所有实现、训练和评估均使用 ESPnet2[5] 。

5.3. 整个 How2 评估集上的结果

表1展示了增强模型（T-1、T-2）、先前最先进水平（P-1） [11], 级联模型（C-1）以及我们的基线系统（B-1、B-2）在How2评估集上的评估得分。与模型P-1相比，采用全注意力、层归一化和LPE的模型B-1显著提升了得分，达到了当前最先进水平。

[2] https://github.com/espnet/espnet/tree/master/ egs2/ljspeech/tts1# pretrained-models-1 [3] https://github.com/Kyubyong/g2p [4] https://huggingface.co/ainize/bart-base-cnn [5] https://github.com/espnet/espnet

表2。表中列出了 Model B-2、T-1 和 T-2 对评估集内 ID 为 zpoDukAvPPY 的样本生成的部分摘要，以及真实摘要（g.t.）。How2 训练集不包含加粗的单词 角斗士，而 Gigaword 训练集包含该词。

ID	摘要
B-2	学习在时尚专家的指导下绘制小牛皮鞋 在这段免费的时装设计视频中。
T-1	学习在这段免费的时尚视频中绘制角斗士凉鞋 由一位时尚设计研究生制作的插画视频。
T-2	学习绘制角斗士凉鞋的技巧 来自时尚专家的这段免费时尚设计视频中。
g.t.	在这段免费的时尚视频中学习绘制角斗士凉鞋 来自一名时尚设计研究生。

表 3。使用各种阈值 α 筛选后每个评估集所包含的样本数量。

α	0.5. . . 0 6	0 7	0 8	0 9	1 0（完整）
# 样本数	110	441	921	1336	1666	2127

如模型B-2所示，扩大模型规模带来了小幅提升；这一强大的模型也是所提出数据增强方法的基线。模型T-1的结果表明，基于TTS的数据增强方法进一步提升了METEOR、ROUGE-L和BERTScore得分，分别最多提高2.1、3.0和0.6分。此外，尽管成本较低，采用基于音素的方法训练的模型T-2也有所提升：METEOR、ROUGE-L和BERTScore得分分别提高了1.1、2.1和0.4分。尽管增强数据的领域与How2 数据集完全不同，但它帮助模型扩充了词汇。例如，如表2所示，模型B-2无法输出单词 角斗士，该词并未出现在How2 训练集中；而模型T-1和T-2似乎从Gigaword 训练集中学会了这个词。

5.4. 未过滤的 How2评估集分析与结果

虽然我们在第5.3节展示了系统的有效性，但我们发现，与级联模型相比，端到端语音摘要模型在若干评估样例上取得了异常高的分数。我们分析了这些样本，发现训练数据集中包含与评估集中的摘要相似或完全匹配的样本。例如，训练集中的 ID: 6D0hfwYIwm4 的摘要：

> 印度菜充满了有趣的香料和风味。了解如何烹制 印度土豆、胡萝卜和豌豆.....

与评估集中的 ID: 2NdV94T-JNc 的摘要非常相似：

> 印度菜充满了有趣的香料和风味。了解如何调味 印度土豆、胡萝卜和豌豆.....

与这些摘要相对应的音频数据通常来自同一视频的不同片段（例如，视频的第一部分和第二部分分别位于训练集和评估集中）。这可能是因为 How2 数据集将 YouTube 的“description”标签定义为摘要（如第 5.1 节所述），且训练集和评估集是随机划分的。在这种情况下，很难评估模型的泛化能力。例如，过拟合训练数据的模型可能会在与训练集重叠的测试样本上取得高分。

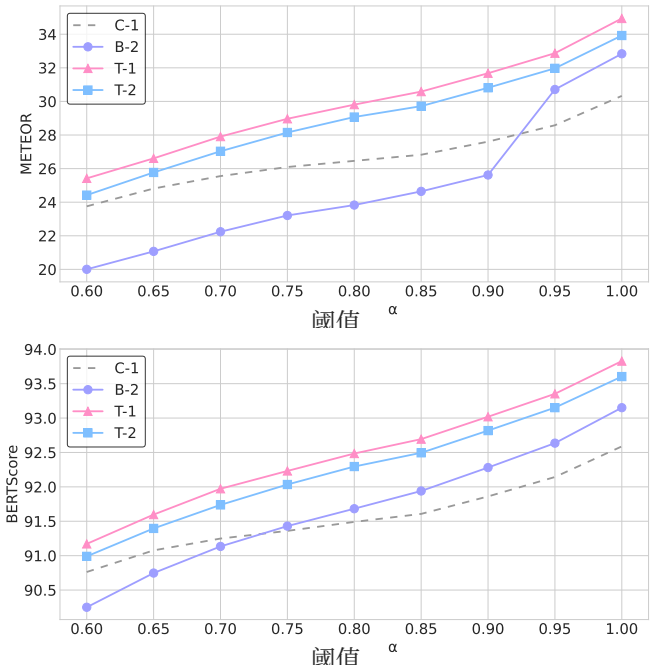

图 3。模型 B-2、T-1、T-2 和 C-1 的 METEOR（上方）和 BERTScore（下方）得分随筛选阈值 α 从0.6到1.0的变化。

因此，我们提出利用定量标准筛选评估集，剔除这类样本。首先，我们计算了评估集中每条摘要与整个 How2 数据集之间的 ROUGE-L 分数。我们将每条评估集摘要的最高分定义为其泄漏分数。例如，ID: 2NdV94T-JNc 的泄漏分数为 0.95。随后，我们从评估集中移除了泄漏分数高于阈值 α 的样本。表 3 展示了不同阈值下各筛选后评估集所包含的样本数量。

图 3 展示了模型 B-2、T-1、T-2 和 C-1 在不同阈值 α下的 METEOR 和 BERTScore 得分。尽管所有模型的得分总体上都会随着 α 降低而变差，但模型 B-2 的这一趋势最为明显。结果表明，基线端到端系统对训练数据过拟合，且容易受到陌生样本的影响。模型 C-1 对未见样本表现出较强的鲁棒性，得分下降也最慢，这可能得益于其 TSum 模型丰富的语言学知识。对于我们提出的方法，模型 T-1 和 T-2 在所有条件下都取得了最高准确率，似乎从外部 TSum 语料库中获得了语言学信息。

6. 结论与未来工作

本文研究了在 How2 数据集上使用外部文本摘要数据进行端到端语音摘要的数据增强方法的有效性。与强大的基线相比，文本转语音和基于音素的数据增强方法分别使 METEOR 分数提高了 2.1 分和 1.1 分。此外，我们还展示了在端到端语音摘要模型中利用文本摘要数据能够获得语言知识的明显趋势。

本文仅使用文本摘要数据作为端到端语音摘要模型的输入。作为未来研究的一部分，我们将尝试利用大型外部语言模型，为模型提供来自多种非配对文本数据的知识。

7. 参考文献

[1] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. Yalta, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, 《语音应用中 Transformer 与循环神经网络的比较研究》，载于自动语音识别与理解论文集，2019。[2] J. Goldman, S. Renals, S. Bird, F. De Jong, M. Federico, C. Fleischhauer, M. Kornbluh, L. Lamel, D. W. Oard, C. Stewart, and R. Wright, 《访问口语》，国际数字图书馆期刊，2005。[3] G. Murray, G. Carenini, and R. Ng, 《生成并验证会议对话摘要：一项用户研究》，载于国际自然语言生成会议论文集，2010。[4] R. Sharma, M. Mahrishi, S. Morwal, and G. Sharma, 《利用余弦相似度检测教育视频中的文本摘要索引点》，材料科学与工程，2021。[5] G. Finley, E. Edwards, A. Robinson, M. Brenndoerfer, N. Sadoughi, J. Fone, N. Axtmann, M. Miller, and D. Suendermann-Oeft, 《用于记录临床就诊的自动化医疗记录员》，载于北美计算语言学协会年会论文集，2018。[6] Y. Zhang, A. Ni, T. Yu, R. Zhang, C. Zhu, B. Deb, A. Celikyilmaz, A. H. Awadallah, and D. Radev, 《长对话摘要探索性研究：哪些方法有效，下一步是什么》，载于自然语言处理实证方法会议论文集，2021。[7] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. R. Radev, 《QMSum：用于基于查询的多领域会议摘要的新基准》，载于北美计算语言学协会人类语言技术年会论文集，2021。[8] M. ´A. T¨undik, V. Kasz´as, and G. Szasz´ak, 《评估自动语音识别错误传播造成的语义空间偏差及其对口语文档摘要的影响》，载于*INTERSPEECH*论文集，2019。[9] T. Kano, A. Ogawa, M. Delcroix, and S. Watanabe, 《基于注意力的语音摘要多假设融合》，载于自动语音识别与理解论文集，2021。[10] T.-E. Liu, S.-H. Liu, and B. Chen, 《口语文档分层神经摘要框架》，载于*ICASSP*论文集，2019。[11] R. Sharma, S. Palaskar, A. W. Black, and F. Metze, 《使用受限自注意力的端到端语音摘要》，载于*ICASSP*论文集，2022。[12] M. Mimura, S. Ueno, H. Inaguma, S. Sakai, and T. Kawahara, 《利用序列到序列语音合成增强声学到词语音识别》，载于口语语言技术研讨会论文集，2018。[13] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, 《通过离散表示利用文本转语音进行自动语音识别的数据增强》，载于自动语音识别与理解论文集，2021。[14] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, 《用于端到端自动语音识别的反向翻译式数据增强》，载于口语语言技术研讨会论文集，2018。[15] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, 《合成波形序列到序列数据以增强序列到序列语音识别的训练数据》，声学科学与技术，2021。[16] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, 《用于端到端自动语音识别的多模态数据增强》，*INTERSPEECH*论文集，2018。[17] R. Masumura、N. Makishima、M. Ihori、A. Takashima、T. Tanaka 和 S. Orihashi, 《基于音素到字素转换的端到端自动语音识别大规模预训练》，2020 年。[18] Y. Jia、M. Johnson、W. Macherey、R. J. Weiss、Y. Cao、C.-C. Chiu、N. Ari、S. Laurenzo 和 Y. Wu, 《利用弱监督数据改进端到端语音到文本翻译》，载于*ICASSP* 论文集，2019 年。[19] D. Graff、J. Kong、K. Chen 和 K. Maeda, 《English Giga-word》，语言数据联盟，费城，2003 年。[20] A. M. Rush、S. Chopra 和 J. Weston, 《用于抽象式句子摘要的神经注意力模型》，2015 年。[21] R. Sanabria、O. Caglayan、S. Palaskar、D. Elliott、L. Barrault、L. Specia 和 F. Metze, 《How2：用于多模态语言理解的大规模数据集》，载于*NeurIPS ViGIL* 研讨会论文集，2018 年。[22] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan、R. A. Saurous、Y. Agiomvrgiannakis 和 Y. Wu, 《通过基于梅尔频谱图预测对 WaveNet 进行条件控制，实现自然文本到语音合成》，载于*ICASSP* 论文集，2018 年。[23] A. Tjandra、S. Sakti 和 S. Nakamura, 《通过直通估计器实现语音链框架中的端到端反馈损失》，载于*ICASSP* 论文集，2019 年。[24] G. Wang、A. Rosenberg、Z. Chen、Y. Zhang、B. Ramabhad- ran、Y. Wu 和 P. Moreno, 《利用合成语音上的一致预测改进语音识别》，载于*ICASSP* 论文集，2020 年。[25] A. Gulati、J. Qin、C.-C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 和 R. Pang, 《Conformer：用于语音识别的卷积增强型Transformer》，载于*INTERSPEECH* 论文集，2020 年。[26] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin, 《注意力就是你所需要的一切》，载于*NIPS* 论文集。[27] J. Gehring、M. Auli、D. Grangier、D. Yarats 和 Y. N. Dauphin, 《卷积序列到序列学习》，载于*ICML* 论文集，2017 年。[28] M. Lewis、Y. Liu、N. Goyal、M. Ghazvininejad、A. Mohamed、O. Levy、V. Stoyanov 和 L. Zettlemoyer, 《BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练》，载于*ACL* 论文集，2020 年。[29] K. M. Hermann、T. Ko˘cisk´y、E. Grefenstette、L. Espeholt、W. Kay、M. Suleyman 和 P. Blunsom, 《教机器阅读和理解》，载于*NIPS* 论文集，2015 年。[30] D. P. Kingma 和 J. Ba, 《Adam：一种随机优化方法》，载于*ICLR* 论文集，2015 年。[31] J. Kim、J. Kong 和 J. Son, 《用于端到端文本到语音的对抗学习条件变分自编码器》，载于*ICML* 论文集，2021 年。[32] L. Lamel、R. Kassel 和 S. Seneff, 《语音数据库开发：声学语音学语料库的设计与分析》，载于*DARPA CSR* 论文集，1989 年。[33] C.-Y. Lin, 《ROUGE：摘要自动评估工具包》，载于*Text Summarization Branches Out*，2004 年。[34] M. Denkowski 和 A. Lavie, 《Meteor universal：适用于任意目标语言的语言特定翻译评估》，载于*EACL WMT* 论文集，2014 年。[35] T. Zhang、V. Kishore、F. Wu、K. Q. Weinberger 和 Y. Artzi, 《BERTScore：使用 BERT 评估文本生成》，载于*ICLR* 论文集，2020 年。