同行专家业内评价意见书编号：20240854203

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名：刘华岱

学号：22160146

申报工程师职称专业类别（领域）：电子信息

浙江工程师学院（浙江大学工程师学院）制

2024年03月28日

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

一． 对本专业基础理论知识和专业技术知识的掌握情况

1. 基础理论知识掌握情况：

1. 机器学习和深度学习：

我在人工智能领域的学习中掌握了机器学习和深度学习的基础理论知识，包括常见的神经网络结构、梯度下降优化算法、正则化技术等。我能够理解并应用常见的深度学习框架如TensorFlow和PyTorch进行模型搭建和训练。

2. 多模态理解：

我对多模态数据（如图像、文本、音频等）的理解和处理有深入的了解。我熟悉多模态数据融合的方法，包括融合层级特征和跨模态注意力机制等。我在相关研究中探索了如何有效地结合不同模态的信息以提高任务性能。

3. 自然语言处理：

作为人工智能专业的学生，我对自然语言处理领域有广泛的了解。我熟悉词嵌入技术、序列模型（如循环神经网络和Transformer）、语言生成模型等。我在学术和实践项目中探索了文本生成、情感分析、命名实体识别等任务。

2. 专业技术知识掌握情况：

1. 语音合成：

我的研究方向之一是语音合成。我深入研究了端到端的语音合成模型，包括基于深度神经网络的WaveNet和基于变分自动编码器的声码器。我能够设计和实现语音合成系统，并优化模型以提高语音质量和自然度。

2. 模型评估与调优：

我具备对人工智能模型进行评估和调优的技能。我能够使用标准评估指标（如BLEU、WER等）来评估模型的性能，并通过超参数调整、数据增强等技术来优化模型性能。我也了解模型的泛化能力和鲁棒性的重要性，在实践中注重模型的可解释性和可维护性。

3. 项目管理与团队合作：

我具备良好的项目管理和团队合作能力。我曾在实验室和实习公司均担任了算法负责人，负责项目规划、进度控制和团队协作。我能够有效沟通和协调团队成员，确保项目按时高质量完成。

二． 工程实践的经历

1. 结构化查询语言（SQL）的生成（时长：1年）：

（1）项目描述：

在这段实践经历中，我负责开发一个系统，旨在通过语音问题结合数据库生成SQL语句，以便用户能够通过口头提问直接获取数据库中的信息。

（2）工作内容：

我与团队合作，首先标注了多口音和多语者的语音SQL数据集，以用于训练模型。随后，我们面临了多口音和多语者的挑战，为了解决这一问题，我提出了梯度反转和语音重编程等技术，以帮助模型学习声学信息无关的语音表征。

（3）成果与贡献：

在这一项目中，我们成功开发了一个能够从语音问题中生成SQL语句的系统，并且通过实验验证了提出的梯度反转和语音重编程等方法的有效性，为实现语音与数据库的无缝交互提供了重要技术支持，并转化成了论文投稿于ACL会议。

2. 金融直播拆条（时长：1年）：

（1）项目描述：

在这段实践经历中，我参与了一个项目，旨在根据给定的多模态直播视频提取直播的热点片段，以便用户能够快速获取所关注领域的最新信息。

（2）工作内容：

在项目中，我们首先标注了首个金融直播数据集，这为模型训练提供了必要的数据支持。然后，面临多模态建模和长序列建模的挑战，我提出了新的方法来解决这些问题，包括多模态特征融合和长序列建模技术。

（3）成果与贡献：

在这一项目中，我们成功开发了一个能够从多模态直播视频中提取热点片段的系统，并且提出的多模态建模和长序列建模方法在实践中取得了显著效果，为金融直播内容的智能处理提供了新的解决方案，为公司的短视频生产和热点提取业务带来了提升，节省了成本。

三． 在实际工作中综合运用所学知识解决复杂工程问题的案例

在金融直播拆条项目中，我面临了多个挑战，包括如何生成多个兴趣点以提升用户点击和停留，以及如何生产高质量的短视频并优化产量、时效和成本。通过综合运用所学知识，我成功解决了这些复杂工程问题，取得了显著的成果。

1. 生成多个兴趣点以提升用户点击和停留：

利用多模态建模技术，结合视频内容、主播语音以及转录文本等信息，识别并生成多个兴趣点，以确保直播拆条的吸引力和多样性。

通过对用户行为数据进行分析和挖掘，我们不断优化兴趣点的生成算法，以提升用户点击率和停留时长。

2. 生产高质量短视频并优化产量、时效和成本：

利用多模态建模和长序列建模技术，对视频内容进行内容理解和情感分析，以确保生成的短视频质量高且符合用户兴趣。

通过自动化流程和智能调度算法，优化短视频生产流程，提高生产效率并降低成本。

在以上工程实践中，我充分综合运用了人工智能、多模态理解、语音合成等领域的知识和技术，成功解决了复杂的工程问题，并取得了显著的成果。通过我们的努力，金融直播拆条项目的使用率和用户体验得到了显著提升，使得直播打点每天有超过70%的机构使用并且直播总体停留时长增加了23%。同时，我们也为短视频产量和用户规模的增长做出了重要贡献，质量、产量、时效和成本均得到了优化，破局短视频发展，帮助短视频产量从800+增加到5000+，短视频整体每月活跃用户数从200万增加到1100万。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】

1.
公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
ViT-TTS: Visual Text-to-Speech with Scalable Diffusion Transformer	会议论文	2023年12月10日	EMNLP		
TranSpeech: Speech-to-Speech Translation With Bilateral Perturbation	会议论文	2023年04月06日	ICLR		
ProDiff: Progressive Fast Diffusion Model For High-Quality Text-to-Speech	会议论文	2022年10月12日	ACM MM		

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

金融直播拆条生成多个兴趣点，提升用户点击和停留，使得直播打点每天有超70%机构使用并且直播总体停留时长增强23%，同时直播拆条生产短视频，质量产量时效成本并优，破局短视频发展，帮助短视频产量从800+到5000+，短视频整体每月活跃用户数从200万增加到1100万。

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 81 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年（要求1年及以上） 考核成绩： 83 分（要求80分及以上）
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：** 刘笑伐	

2216016

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀 ☐良好 ☐合格 ☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：　　年3月29日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过 ☐不通过（具体原因：　　） 工程师学院教学管理办**公室**审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22160146	姓名：刘华岱	性别：男	学院：工程师学院	专业：电子信息	学制：2.5年
毕业时最低应获：24.0学分	已获得：24.0学分			入学年月：2021-09	毕业年月：2024-03
学位证书号：1033532024602171		毕业证书号：103351202402600397			授予学位：电子信息硕士

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2021-2022学年秋季学期	人工智能算法与系统		2.0	88	专业学位课	2021-2022学年夏季学期	药品创制工程实例		2.0	87	专业学位课
2021-2022学年冬季学期	新药发现理论与实践		2.0	86	专业学位课	2021-2022学年夏季学期	自然辩证法概论		1.0	77	公共学位课
2021-2022学年秋冬季学期	中国特色社会主义理论与实践研究		2.0	87	公共学位课	2021-2022学年夏季学期	批判性思维与科学研究		1.0	66	专业选修课
2021-2022学年秋冬季学期	数据分析的概率统计基础		3.0	95	专业选修课	2021-2022学年夏季学期	工程伦理		2.0	81	公共学位课
2021-2022学年秋冬季学期	机器学习		3.0	76	专业选修课	2023-2024学年夏季学期	研究生英语基础技能		1.0	免修	公共学位课
2021-2022学年秋冬季学期	研究生论文写作指导		1.0	66	专业学位课	2023-2024学年夏季学期	研究生英语应用能力提升		2.0	免修	公共学位课
2021-2022学年春季学期	生物智能与算法		2.0	84	专业选修课						

说明：1.研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2.备注中“*”表示重修课程。

学院成绩校核章：

成绩校核人：张梦依

打印日期：2024-04-02

1. 论文地址：aclanthology.org/2023.emnlp-main.990.pdf

ViT-TTS: Visual Text-to-Speech with Scalable Diffusion Transformer

Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, Zhou Zhao

Abstract

Text-to-speech(TTS) has undergone remarkable improvements in performance, particularly with the advent of Denoising Diffusion Probabilistic Models (DDPMs). However, the perceived quality of audio depends not solely on its content, pitch, rhythm, and energy, but also on the physical environment.In this work, we propose ViT-TTS, the first visual TTS model with scalable diffusion transformers. ViT-TTS complement the phoneme sequence with the visual information to generate high-perceived audio, opening up new avenues for practical applications of AR and VR to allow a more immersive and realistic audio experience. To mitigate the data scarcity in learning visual acoustic information, we 1) introduce a self-supervised learning framework to enhance both the visual-text encoder and denoiser decoder; 2) leverage the diffusion transformer scalable in terms of parameters and capacity to learn visual scene information. Experimental results demonstrate that ViT-TTS achieves new state-of-the-art results, outperforming cascaded systems and other baselines regardless of the visibility of the scene. With low-resource data (1h, 2h, 5h), ViT-TTS achieves comparative results with rich-resource baselines.

PDF

Cite

Search

Anthology ID: 2023.emnlp-main.990
Volume: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing
Month: December
Year: 2023
Address: Singapore
Editors: Houda Bouamor, Juan Pino, Kalika Bali
Venue: EMNLP
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 15957–15969

ViT-TTS: Visual Text-to-Speech with Scalable Diffusion Transformer

Huadai Liu[1]*, **Rongjie Huang**[1]*, **Xuan Lin**[2]*, **Wenqiang Xu**[2], **Maozong Zheng**[2], **Hong Chen**[2], **Jinzheng He**[1], **Zhou Zhao**[1]†
Zhejiang University[1], Ant Group[2]
{liuhuadai,rongjiehuang,jinzhenghe,zhaozhou}@zju.edu.cn
{daxuan.lx,yugong.xwq,zhengmaozong.zmz,wuyi.ch}@antgroup.com

Abstract

Text-to-speech(TTS) has undergone remarkable improvements in performance, particularly with the advent of Denoising Diffusion Probabilistic Models (DDPMs). However, the perceived quality of audio depends not solely on its content, pitch, rhythm, and energy, but also on the physical environment. In this work, we propose ViT-TTS, the first visual TTS model with scalable diffusion transformers. ViT-TTS complement the phoneme sequence with the visual information to generate high-perceived audio, opening up new avenues for practical applications of AR and VR to allow a more immersive and realistic audio experience. To mitigate the data scarcity in learning visual acoustic information, we 1) introduce a self-supervised learning framework to enhance both the visual-text encoder and denoiser decoder; 2) leverage the diffusion transformer scalable in terms of parameters and capacity to learn visual scene information. Experimental results demonstrate that ViT-TTS achieves new state-of-the-art results, outperforming cascaded systems and other baselines regardless of the visibility of the scene. With low-resource data (1h, 2h, 5h), ViT-TTS achieves comparative results with rich-resource baselines. [1] [2]

1 Introduction

Text-to-speech (TTS) (Ren et al., 2019; Huang et al., 2022a,b) aims to synthesize audios that is consistent with the reference samples in terms of semantic meaning, timbre, emotions, and melody, and has shown remarkable advancements with the advent of Denoising Diffusion Probabilistic Models (DDPMs). However, the perceived audio quality is not solely determined by these aspects, as it is also influenced by the surrounding physical environment. For instance, a room with hard surfaces like concrete or glass reflects sound waves, whereas a room with soft surfaces such as carpets or curtains absorbs them. This variance can drastically impact the clarity and quality of the sound we hear.

To ensure an authentic and captivating experience, it is imperative to accurately model the acoustics of a room, particularly in virtual reality (VR) and augmented reality (AR) applications. Recent years have seen a surge in significant research (Li et al., 2022; Radford et al., 2021; Li et al., 2023; Huang et al., [illegible] addressing the language-visual modeling problem. For instance, Li et al. (2022) have proposed a unified video-language pre-training framework for learning robust representation, while Radford et al. (2021) have focused on large-scale image-text pairs pre-training via contrastive learning. Visual TTS open-ups numerous practical applications, including dubbing archival films, providing a more immersive and realistic experience in virtual and augmented reality, or adding appropriate sound effects to games.

Despite the benefits of language-visual approaches, training visual TTS models typically requires a large amount of training data, while there are very few resources providing parallel text-visual-audio data due to the heavy workload. Besides, creating a sound experience that matches the visual content remains challenging when developing AR/VR applications, as it is still unclear how various regions of the image contribute to reverberation and how to incorporate the visual modality as auxiliary information in TTS.

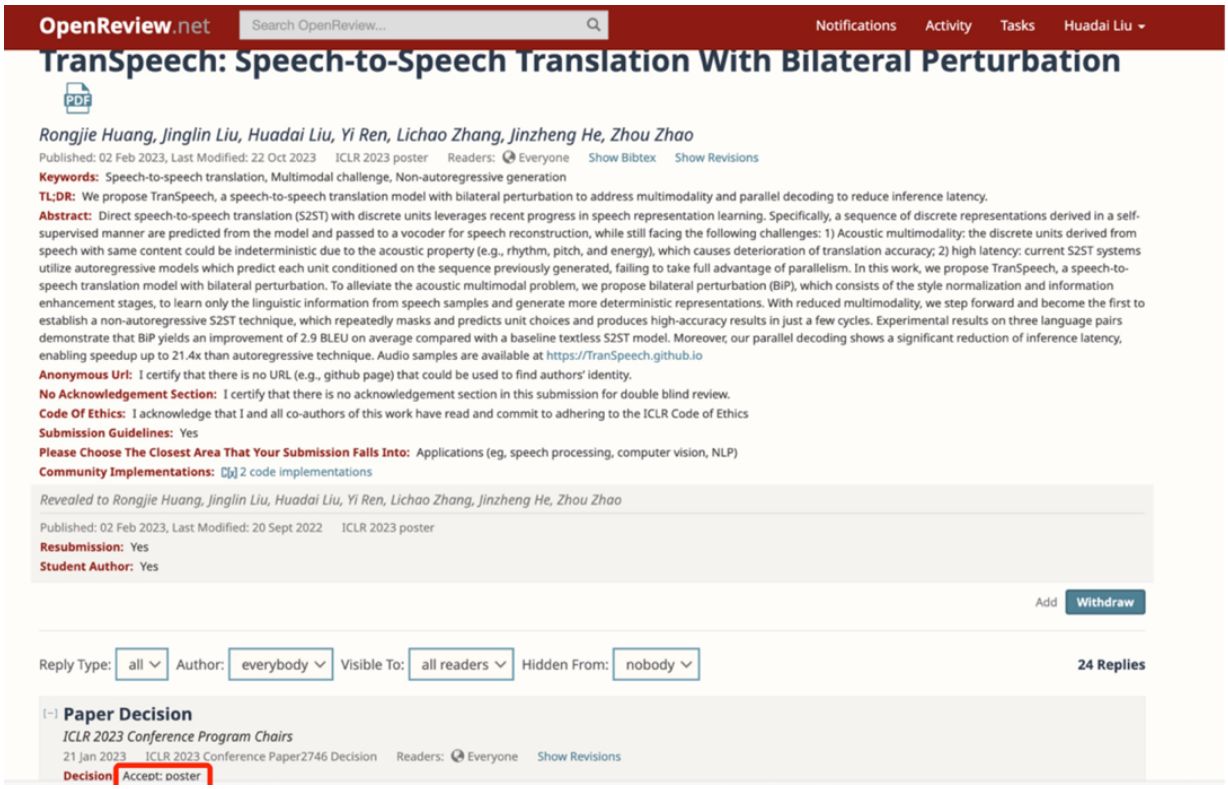
In this work, we formulate the task of visual TTS to generate audio with reverberation effects in target scenarios given a text and environmental image, introducing ViT-TTS to address the issues of data scarcity and room acoustic modeling. To enhance visual-acoustic matching, we 1) propose the visual-

*Equal contributions
†Corresponding author
[1]Audio samples are available at https://ViT-TTS.github.io/.
[2]Code is available at https://github.com/liuhuadai/ViT-TTS/

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15957–15969
December 6-10, 2023

导师签名：

2. 论文地址：openreview.net/pdf?id=UVAmFAtC5ye



TRANSPEECH: SPEECH-TO-SPEECH TRANSLATION WITH BILATERAL PERTURBATION

Rongjie Huang[1]*, **Jinglin Liu**[1]*, **Huadai Liu**[1]*, **Yi Ren**[2], **Lichao Zhang**[1], **Jinzheng He**[1], **Zhou Zhao**[1]†

[1]Zhejiang University
{rongjiehuang, jinglinliu, huadailiu, zhaozhou}@zju.edu.cn

[2]ByteDance
ren.yi@bytedance.com

ABSTRACT

Direct speech-to-speech translation (S2ST) with discrete units leverages recent progress in speech representation learning. Specifically, a sequence of discrete representations derived in a self-supervised manner are predicted from the model and passed to a vocoder for speech reconstruction, while still facing the following challenges: 1) Acoustic multimodality: the discrete units derived from speech with same content could be indeterministic due to the acoustic property (e.g., rhythm, pitch, and energy), which causes deterioration of translation accuracy; 2) high latency: current S2ST systems utilize autoregressive models which predict each unit conditioned on the sequence previously generated, failing to take full advantage of parallelism. In this work, we propose TranSpeech, a speech-to-speech translation model with bilateral perturbation. To alleviate the acoustic multimodal problem, we propose bilateral perturbation (BiP), which consists of the style normalization and information enhancement stages, to learn only the linguistic information from speech samples and generate more deterministic representations. With reduced multimodality, we step forward and become the first to establish a non-autoregressive S2ST technique, which repeatedly masks and predicts unit choices and produces high-accuracy results in just a few cycles. Experimental results on three language pairs demonstrate that BiP yields an improvement of 2.9 BLEU on average compared with a baseline textless S2ST model. Moreover, our parallel decoding shows a significant reduction of inference latency, enabling speedup up to 21.4x than autoregressive technique. [1]

1 INTRODUCTION

Speech-to-speech translation (S2ST) aims at converting speech from one language into speech in another, significantly breaking down communication barriers between people not sharing a common language. Among the conventional method (Lavie et al., 1997; Nakamura et al., 2006; Wahlster, 2013), the cascaded system of automatic speech recognition (ASR), machine translation (MT), or speech-to-text translation (S2T) followed by text-to-speech synthesis (TTS) have demonstrated reasonable results yet suffering from expensive computational costs. Compared to these cascaded systems, recently proposed direct S2ST literature (Jia et al., 2019; Zhang et al., 2020; Jia et al., 2021; Lee et al., 2021a;b) demonstrate the benefits of lower latencies as fewer decoding stages are needed.

Among them, Lee et al. (2021a;b) leverage recent progress on self-supervised discrete units learned from unlabeled speech for building textless S2ST systems, further supporting translation between unwritten languages. As illustrated in Figure 1(a), the unit-based textless S2ST system consists of

*Equal Contribution
†Corresponding Author
[1]Audio samples are available at https://TranSpeech.github.io/.

导师签名：

3. 论文地址：aclanthology.org/2023.acl-long.479.pdf

AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation

Rongjie Huang, Huadai Liu, Xize Cheng, Yi Ren, Linjun Li, Zhenhui Ye, Jinzheng He, Lichao Zhang, Jinglin Liu, Xiang Yin, Zhou Zhao

Abstract

Direct speech-to-speech translation (S2ST) aims to convert speech from one language into another, and has demonstrated significant progress to date. Despite the recent success, current S2ST models still suffer from distinct degradation in noisy environments and fail to translate visual speech (i.e., the movement of lips and teeth). In this work, we present AV-TranSpeech, the first audio-visual speech-to-speech (AV-S2ST) translation model without relying on intermediate text. AV-TranSpeech complements the audio stream with visual information to promote system robustness and opens up a host of practical applications: dictation or dubbing archival films. To mitigate the data scarcity with limited parallel AV-S2ST data, we 1) explore self-supervised pre-training with unlabeled audio-visual data to learn contextual representation, and 2) introduce cross-modal distillation with S2ST models trained on the audio-only corpus to further reduce the requirements of visual data. Experimental results on two language pairs demonstrate that AV-TranSpeech outperforms audio-only models under all settings regardless of the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yields an improvement of 7.6 BLEU on average compared with baselines. Audio samples are available at https://AV-TranSpeech.github.io/.

PDF

Cite

Search

Anthology ID: 2023.acl-long.479
Volume: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Month: July
Year: 2023
Address: Toronto, Canada
Editors: Anna Rogers, Jordan Boyd-Graber, Naoaki Okazaki
Venue: ACL
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 8590–8604

AV-TranSpeech: Audio-Visual Robust Speech-to-Speech Translation

Rongjie Huang[1]*, **Huadai Liu**[1]*, **Xize Cheng**[1]*, **Yi Ren**[2], **Linjun Li**[1], **Zhenhui Ye**[1], **Jinzheng He**[1], **Lichao Zhang**[1], **Jinglin Liu**[2], **Xiang Yin**[2], **Zhou Zhao**[1]†
Zhejiang University[1], ByteDance[2]
{rongjiehuang,liuhuadai,zhaozhou}@zju.edu.cn
{ren.yi,liu.jinglin,yinxiang.stephen}@bytedance.com

Abstract

Direct speech-to-speech translation (S2ST) aims to convert speech from one language into another, and has demonstrated significant progress to date. Despite the recent success, current S2ST models still suffer from distinct degradation in noisy environments and fail to translate visual speech (i.e., the movement of lips and teeth). In this work, we present AV-TranSpeech, the first audio-visual speech-to-speech (AV-S2ST) translation model without relying on intermediate text. AV-TranSpeech complements the audio stream with visual information to promote system robustness and opens up a host of practical applications: dictation or dubbing archival films. To mitigate the data scarcity with limited parallel AV-S2ST data, we 1) explore self-supervised pre-training with unlabeled audio-visual data to learn contextual representation, and 2) introduce cross-modal distillation with S2ST models trained on the audio-only corpus to further reduce the requirements of visual data. Experimental results on two language pairs demonstrate that AV-TranSpeech outperforms audio-only models under all settings regardless of the type of noise. With low-resource audio-visual data (10h, 30h), cross-modal distillation yields an improvement of 7.6 BLEU on average compared with baselines.[1]

1 Introduction

Speech-to-speech translation (S2ST) models (Tjandra et al., 2019; Zhang et al., 2020; Jia et al., 2021) relying on speech data have achieved high performance and significantly broken down communication barriers between people not sharing a common language, which attracts broad interest in the machine learning community (Huang et al., 2022c; Huang et al.). Among them, direct systems (Lee et al., 2021a,b; Huang et al., 2022d) leverage recent progress on self-supervised discrete units learned from unlabeled speech for building textless S2ST, further supporting translation between unwritten languages (Chen et al., 2022).

As speech production (Huang et al., 2023; Lam et al., 2021; Huang et al., 2022b) is accompanied by the movement of lips and teeth, it can be visually interpreted to understand speech. In recent years, significant research (Shi et al., 2022a; Prajwal et al., 2022) has introduced joint modeling of spoken language and vision: Shi et al. (2022b) investigate to learn lip-based audio-visual speaker embeddings, where the speaker's mouth area is used alongside speech as inputs. Chern et al. (2022) focus on audio-visual speech enhancement and separation which better integrates visual information. Despite their success, it is unclear how lip can contribute to audio-based S2ST, and how to incorporate visual modality as auxiliary information in S2ST. A visual translator may open up a host of practical applications: improving speech translation in noisy environments, enabling dictation, or dubbing archival films.

Despite the benefits of audio-visual approaches, training direct speech translation models without relying on intermediate text typically requires a large amount of training data, while there are very few resources providing parallel audio-visual speech due to the heavy workload. To mitigate the data scarcity, researchers have leveraged multitask learning (Lee et al., 2021a), data augmentation (Popuri et al., 2022), and weekly-supervised data with synthesized speech (Jia et al., 2022a) in audio S2ST.

In this work, we propose AV-TranSpeech, introducing the first AV-S2ST system without using text. As illustrated in Figure 1, our textless AV-TranSpeech inherits speech-to-unit translation (S2UT) framework (Lee et al., 2021b; Huang et al., 2022d), which consists of an audio-visual speech-to-unit translation (AV-S2UT) model followed by

*Equal contributions
†Corresponding author
[1]Audio samples are available at https://AV-TranSpeech.github.io/.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8590–8604
July 9-14, 2023

导师签名：