arXiv:2303.01639v1 [cs.SD] 3 Mar 2023

WESPER：面向基于 Whisper 的语音交互的零样本实时低语声到正常语音转换

Jun Rekimoto 东京大学 日本东京文京区本乡
7-3-1 索尼计算机科学实验室 日本京都府京都市下京区本通町 13-1 rekimoto@acm.org

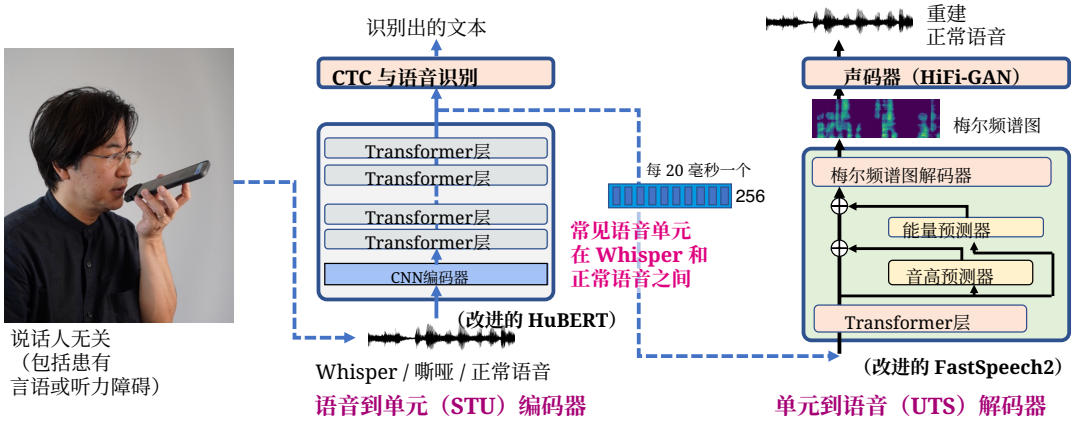

图 1：WESPER 是一种实时低语声到正常语音转换机制，由一个语音到单元（STU）编码器和一个单元到语音（UTS）解码器组成。该编码器利用自监督预训练，为低语和正常话语生成通用语音单元；解码器则根据这些语音单元还原语音。它可实时实现与用户无关的语音转换。

摘要

识别耳语语音并将其转换为正常语音，为语音交互创造了诸多可能性。由于耳语语音的声压显著低于正常语音，因此在公共场所中，它可以作为一种半静音语音交互方式使用，不会被他人听见。将耳语转换为正常语音，还能提升具有言语或听力障碍人群的语音质量。然而，传统的语音转换技术无法提供足够高的转换质量，或要求使用由耳语语音和正常语音话语配对组成的说话人依赖型数据集。为解决这些问题，我们提出了 WESPER——一种基于自监督学习的零样本、实时耳语到正常语音转换机制。WESPER 由一个语音到单元（STU）编码器和一个单元到语音（UTS）解码器，可从编码后的语音单元重建语音。与现有方法不同，这种转换不依赖特定用户，也不需要耳语声和正常语音的配对数据集。UTS 解码器可以根据语音单元，以任意目标说话人的声音重建语音，而且只需要目标说话人未标注的语音数据。我们确认，从耳语声转换得到的语音质量有所提升，同时保留了自然韵律。此外，我们还确认了所提出的方法在为言语或听力障碍人士进行语音重建方面的有效性。

CCS 概念

• **以人为中心的计算** → 基于声音的输入／输出；界面设计原型开发； 移动设备； • **计算方法** →**神经网络**。

关键词

语音交互、耳语声、耳语声转换、无声语音、人工智能、神经网络、自监督学习

ACM 参考格式：
Jun Rekimoto。2023。WESPER：面向基于 Whisper 的语音交互的零样本实时 Whisper 正常语音转换。在 *2023* 年计算机系统中的人为因素国际会议论文集（*CHI '23*）中，

在不收取费用的情况下，允许出于个人或课堂使用目的制作本作品的数字或纸质副本，或复制本作品的全部或部分内容，但前提是不得出于营利或商业利益进行复制或分发，并且副本必须附有本声明及第一页上的完整引文。对于本作品中由作者以外的其他人拥有的组件，其版权必须得到尊重。注明出处后可以进行摘要。若要以其他方式复制、再版、发布到服务器或重新分发至列表，则必须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 联系我们申请许可。*CHI '23*, *2023*年*4*月*23*日至*28*日，德国汉堡© 2023 版权所有归所有者/作者所有。出版权授权给 ACM。
ACM ISBN 978-1-4503-9421-5/23/04...$15.00
https://doi.org/10.1145/3544548.3580706

2023年4月23日至28日，德国汉堡。ACM，纽约，美国，13页。
https://doi.org/10.1145/3544548.3580706

1 引言

尽管语音交互系统已得到广泛应用，但通常不便于在人群中使用。在公共场所使用语音命令可能会让人觉得不合时宜，而且还存在机密信息泄露的风险。此外，在电话会议期间说话可能会让附近的人感到不自在，也可能损害谈话的保密性。

为解决这些问题，人们开发了多种无声语音输入技术 [3, 31, 32, 50, 51, 54]；然而，这些方法需要专用传感器，且语音识别准确率不高，仍停留在识别预定义命令的水平。此外，尚未实现将非条件无声语音转换为正常发声话语。

此外，无声语音还可用于采集言语或听力障碍者的话语。然而，由于上述限制，现有方法无法满足实用型无障碍辅助工具的需求。

与无声语音不同，我们关注耳语语音。耳语的声压足够低，能够确保保密性，而且耳语几乎等同于无声语音。耳语可以通过普通麦克风采集，无需任何特殊的传感器配置。言语障碍人士仍然可以耳语或用嘶哑嗓音说话，即使他们的发声器官可能受伤或被切除；因此，他们的声音仍可能被识别。

为了解决这些问题，我们提出了 WESPER，这是一种基于自监督学习的实时零样本耳语到正常语音转换方法。该方法旨在对耳语语音进行与说话人无关的转换，使其转为正常语音。无需针对每位用户进行训练，也不需要耳语与正常话语的配对数据集。所提出的架构由一个使用正常语音和耳语语音预训练的语音到单元（STU）编码器，以及一个根据语音单元生成目标语音的单元到语音（UTS）解码器组成（图 1）。通过使用（未配对的）耳语和正常语音进行预训练，STU 可以经过训练以缩小正常语音与耳语话语，并输出一个通用语音单元。

UTS 解码器可以使用特定说话人的语音数据进行训练（无需配套的文本标签）。如果使用的是某人的声音，则可以进行转换，将耳语声还原为该人的正常语音，甚至转换为另一个人的声音。

由于编码器和解码器以非自回归方式运行，整个系统能够实时运行。因此，例如将其用于电话会议时，参会者可以用耳语声说话，语音随后会实时转换，其他人则可以听到以正常语音播放的内容。

图 2 展示了使用所提方法进行耳语到正常语音转换的梅尔频谱图示例。更多转换结果请见随附视频。

本研究的贡献可概括如下。

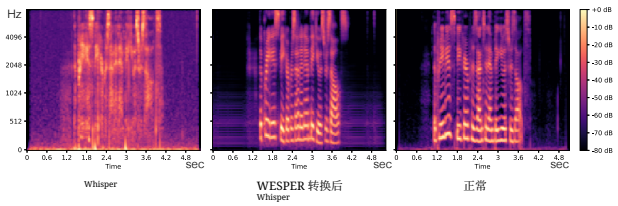

图 2：WESPER 转换结果：左：耳语语音；中：经 WESPER 转换的耳语语音；右：同一说话人的正常语音，且文本转写相同

- 我们提出了一种实时、说话人无关、无词汇限制的耳语转正常语音转换方法，仅使用非配对耳语语音和正常语音即可训练。
- 无需文本转写，即可通过特定说话人的语音样本学习目标语音。
- 我们通过实验确认，该方法提升了正常说话人以及构音障碍或听力障碍说话人的表现。

2 相关工作

2.1 无声语音与耳语语音研究

关于无声语音的各类研究开发了多种方法，可通过不同的传感器配置（包括唇读、肌电图（肌电描记术）和超声波）识别用户的无声话语或无声指令 [3, 31, 32, 50, 51, 54]。然而，这些系统需要专用数据集进行训练，因此增加了对专用传感器的需求，也阻碍了这些技术的广泛应用。因此，这些系统的词汇量有限（可用指令少于 100 条，通常约为 30 条）。尽管无声语音识别有时被认为是帮助发声障碍人士的一种可能途径，但由于词汇量有限，这些方法无法解读不受限制的自由语音。

耳语语音与无声语音具有一些相似特征，例如在公共场所使用时能保持社交可接受性。不过，由于可以使用普通麦克风，它的应用范围更广。

SilentVoice [15] 是一种语音交互技术，使用吸气语音，即在吸气时发出的语音。吸气语音的振幅较低，可视为无声语音的一种变体。它需要将麦克风式设备放置在非常靠近嘴部的位置，而且用户需要经过训练，才能正确地使用吸气语音说话。此外，还需要包含吸气语音及其文本转写的定制语料库。

此前已有针对耳语语音识别 [6, 11, 14, 18] 的研究。Alexa 智能音箱支持耳语模式 [23]。在此模式下，当用户对 Alexa 耳语时，Alexa 会以耳语方式回应 [10]。

DualVoice [48] 也被提出作为一种基于 wav2vec2.0 [2] 和 HuBERT [21]自监督语音识别系统，对耳语语音执行端到端识别的方法。他们还提出了一种交互技术，用于

模型	训练数据	每用户数据集	转换为正常语音	（Whisper）语音识别
无声语音 例如[3, 31, 32, 50, 51, 54]	用于.....的专用配对数据集 专用传感器	必需	both	YES （命令或字符）
Parotoron [4]	成对的耳语声与正常语音	必需	YES	YES （通过语音转换）
DualVoice [48]	带标签的未配对 耳语声与正常语音	必需	NO	YES
SilentVoice [15]	已标注 SilentVoice（自定义）	必需	NO	YES
CycleGAN-VC [30]	未配对、未标注 耳语语音和正常语音	不需要	YES	NO
MSpeC-Net [37]	配对的 耳语-正常语音	需要	YES	NO
AGAN-W2SC [16]	配对的 耳语-正常语音	必需	YES	NO
WESPER（我们的方法）	未配对且未标注 耳语声与正常语音	不需要	YES	YES （通过语音转换）

表1：无声与耳语语音研究

区分耳语声和正常话语。WESPER 可以与 DualVoice 结合，有选择地转换用户的耳语声。

2.2 语音转换

正常语音转换技术已被开发用于将一个有声话语转换为另一个 [19, 29, 30, 43, 46]；然而，将这些方法应用于耳语到正常语音转换时，其转换质量仍不尽如人意。

近年来，研究人员对多种基于机器学习的耳语到正常语音转换技术进行了研究 [41]。为了比较这些技术，必须考虑转换质量以及用于训练的数据集所需具备的特性。准备一个成对的耳语声与正常语音数据集（即包含同一话语的耳语版本和正常版本的数据集），或一个带标签的耳语语音数据集（即包含耳语话语及其对应文本标签的数据集），需要投入大量工作。相反，如果使用非成对且无标签的耳语语音数据集（即仅包含耳语话语、不含对应正常版本或标签的数据集），准备数据集所需的工作量就很小。

用于耳语到正常语音转换的注意力引导生成对抗网络（AGAN-W2SC）是一种基于生成对抗网络的耳语到正常语音转换 [16]。它将以梅尔频谱图表示的耳语声转换为对应正常语音的梅尔频谱图。该方法基于生成对抗网络，并使用注意力图进行转换。它需要成对耳语-正常语音数据集。MSpeC-Net [37] 是一种基于自编码器的多领域语音转换方法，支持耳语到正常语音的转换。它还需要成对耳语-正常语音数据集。

此外，耳语语音可以通过语音识别转换为文本，再使用文本转语音方法生成正常语音 [22]。然而，这种方法需要一个用于耳语语音识别的带标签的数据集，而且耳语话语中的韵律信息会丢失，因为中间文本表示无法传达这类信息。

Parrotron [4] 是一种语音转换系统，旨在改善发育不良说话人的语音。它基于一种符合文本转语音系统 Tacotron [58] 的编码器-解码器模型；不过，它与 Tacotron 的不同之处在于，输入和输出数据均采用梅尔频谱图格式。它还需要成对的源语音和目标语音，因此所需数据集的构建相对困难。

音素后验概率图（PPGs） [55]是从自动语音识别中获得的中间信息。由于音素后验概率图能够以说话人无关的方式表征所说内容的发音，因此可用于多对一说话人转换。据我们所知，尚未有研究探讨音素后验概率图对耳语声的影响。我们的方法同样使用中间语音单元，但不同于自动语音识别（自动语音识别）和音素后验概率图的做法，我们不需要基于文本的语料库。

相比之下，我们提出的系统只需要未配对的目标语音和源语音样本，无需提供相应的文本转录，因此数据集易于准备，且与目标语言无关。

表1总结了相关技术的特点。我们还展示了耳语转正常语音的转换示例，其中包括NMSE-DiscoGAN [53], MSpeC-Net [37], CycleGAN- VC [30], AGAN-W2SC [16], 以及 WESPER，示例见 https://wesperproj. github.io/。

2.3 语音的自监督表征学习

近年来，将在无标注语音数据上进行预训练与自监督表征学习相结合，并在带标签的语音数据上进行微调，受到了广泛关注。这些系统主要

用于语音识别应用，但也已被用于执行说话人、语言和情感识别 [44, 59]。

具体而言，隐藏单元BERT（hu-BERT）的预训练方法 [21] 与自然语言处理中使用的掩码语言模型——双向编码器表示变换器（BERT）[12] 的方法类似。该方法旨在遮盖部分输入，并根据输入的其余部分估计相应的表达特征。通过这种预训练，模型能够学习输入数据的声学属性以及语音的特征。

作为自动语音识别（ASR）使用时，预训练后仅使用少量带有文本转录的音频数据进行微调。系统添加了一个投影层和一个连接时序分类（CTC）层 [17]，用于从音频波形生成文本转录。

据 [2, 59],报道，自监督自动语音识别仅使用少量带标签的语音数据进行微调，便达到了与传统最先进的自动语音识别系统相当的语音识别准确率。因此，该架构可能适用于在耳语语音语料库有限的情况下进行耳语声识别。

WESPER 的独特之处在于，它通过预训练缩小编码耳语话语与正常话语的潜在向量之间的差异。

2.4 无文本自然语言处理

近年来，人们开发了无文本语音处理和语音转换方法。通过自监督学习，这些系统从未附带文本转录的语音数据中提取潜在表示。Textless-NLP[34, 35] 和 AudioLM [5] 在语音处理系统中不使用文本转录或音素符号，而是使用通过自监督学习构建的离散单元。软离散单元是无文本语音处理的另一种方法 [57]。

我们提出的方法同样使用通过自监督学习获得的非离散向量作为潜在语音表征，并且不会显式使用文本转录或音素符号。我们的研究独具特色，因为我们证明，通过自监督学习，耳语语音和正常语音可以由相似的语音单元表示。此外，我们的方法还旨在后续阶段与语音生成系统无缝协作。

3 WESPER 语音转换模型

WESPER 由 STU 编码器和 UTS 解码器组成。STU 将耳语或正常语音转换为通用语音单元。UTS 将通用语音单元转换为梅尔频谱图，声码器可据此重建语音。WESPER 的特点是，同一话语的通用语音单元可以相似，尽管 STU 仅使用（未配对的）耳语语音和正常语音进行预训练，且未使用耳语话语与正常话语的配对集合进行训练。下文将介绍各部分的详情以及模型的训练方法。

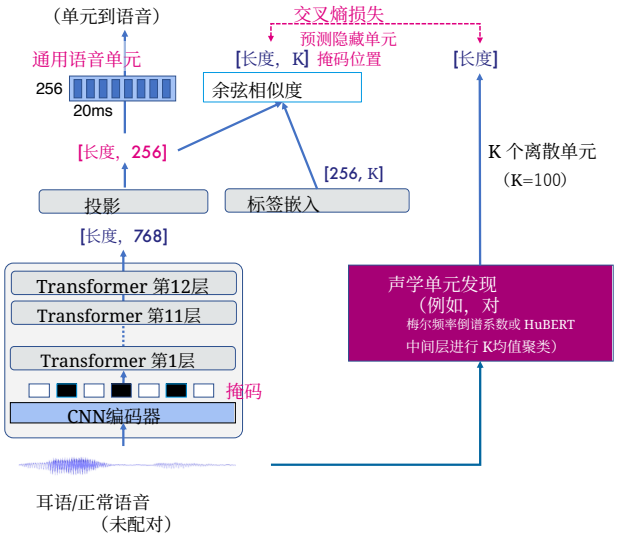

图3：STU预训练概览。预训练使用未配对的耳语语音或正常语音。通过根据掩码输入估计离散单元来训练 Transformer 层。Transformer层后的投影层会生成 256 维向量（每 20 毫秒一个），这些向量用作通用语音单元。

3.1 语音到单元（STU）编码器

STU 编码器以音频波形为输入，并输出语音单元。它基于 HuBERT [21], 这一用于语音的自监督神经网络。HuBERT 以类似 BERT [12] 的方式对大量未标注语音进行预训练；它学习从部分屏蔽的语音特征中恢复相关部分，从而习得语音语言模型。

我们的目标是生成在耳语语音和正常语音之间尽可能一致的语音单元。为此，我们使用耳语和正常语音话语的混合数据对 STU 进行预训练；正常英语语音来自包含许多说话人的 Librispeech 960 小时数据集 [42]。Librispeech 语音数据使用基于 LPC 的音频转换工具 [60] 机械转换为耳语声。此外，我们还使用了 wTIMIT 正常语音和耳语语音数据集 [36]，其语音时长为 58 小时。

图3详细展示了STU的预训练过程。预训练仅使用未配对的耳语语音或正常语音。通过从掩码输入中估计离散单元来训练Transformer层。与HuBERT类似，在第一阶段，首先对输入语音数据进行K 均值聚类，以生成离散目标单元；随后，再对Transformer 中间层的输出进行K 均值聚类。在我们的实验中，我们使用了100个离散单元。Transformer层之后的投影层会生成一系列256维向量（每20毫秒一个），这些向量被用作通用语音单元。

在 STU 中，CNN 特征提取器前设置了 12 个 Transformer层（图1）。预训练后，通过比较各层的输出，我们确认：（1）差异

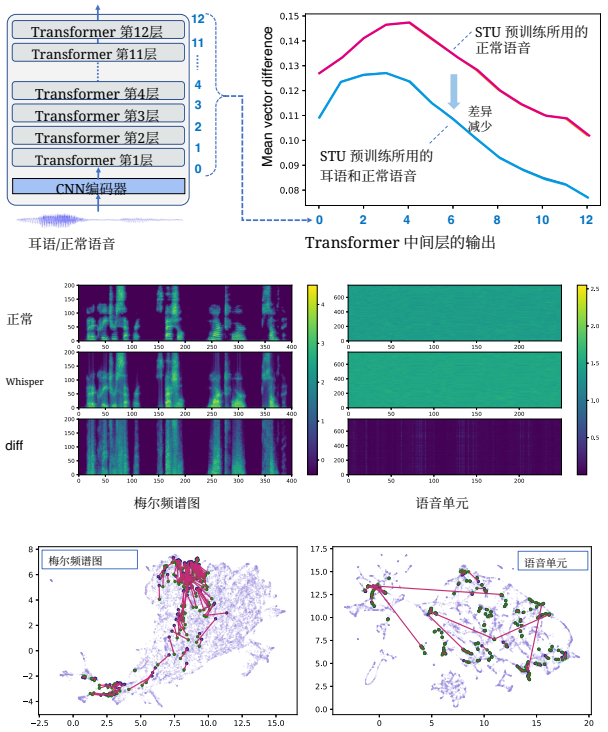

图4：耳语语音与正常语音之间差异的比较：使用正常语音和经过耳语转换的语音进行比较。（上）同时使用正常语音和耳语语音进行预训练的 STU，比仅使用正常语音进行预训练的 STU 产生的差异更少。随着 Transformer 层加深，两者之间的差异逐渐减小。（中）梅尔频谱图中存在差异，而这种差异随着语音单元的使用而减小。（下）UMAP [39] 可视化了梅尔频谱图空间和语音单元空间中耳语语音与正常语音之间的差异。

耳语和正常语音之间的语音单元值差异会随层深度增加而减小；（2）与仅使用正常语音话语进行预训练相比，使用耳语和正常语音话语共同对 STU 进行预训练时，差异会减小（图4）。图4（底部）还展示了耳语和正常语音在梅尔频谱图空间及通用语音单元空间中的 UMAP [39] 可视化。两段语音中对应的时段以线条相连。如图所示，在通用语音单元空间中，差异有所减小（少数距离较长的情况除外）。

尽管耳语和正常语音的语音特征值不同，我们认为，自监督预训练能够解决这一问题，使语言立场相似的话语以相似的单元表示。我们推测，我们的学习方法可以从正常和耳语话语中提取共同的发音，这类似于自监督预训练方法能够从自动语音识别系统中不同说话人的话语里提取共同发音。

此外，我们可以通过添加一个 CTC层 [17]（如 [48]所示），使用（附带转录文本的）wTIMIT 和 Librispeech 数据集对 STU 进行微调；因此，STU 可以用作自动语音识别系统，以识别耳语和正常语音。

3.2 UTS 解码器

UTS 解码器以 STU 生成的语音单元为输入，并重建目标（正常）语音。它基于非自回归文本到语音（TTS）系统 FastSpeech2 [49]。原版 FastSpeech2 包含一个嵌入层，用于接收文本或音素标记并将其转换为一系列向量嵌入；而 UTS 直接以语音单元作为输入，因此可以省去这一层。原版 FastSpeech2 还包含一个时长估计器，用于估计每个音素标记的时长，以及一个长度调节器，根据估计的时长调整内部向量的数量。由于 STU 以恒定速率生成语音单元，因此这些组件也可以省去。

训练原版 FastSpeech2 时，必须借助 Montreal Forced Aligner [38] 等外部工具，提供语料库中每个音素的真实时长。受此限制，FastSpeech2 的训练是语言相关的。相较之下，UTS 无需进行时长估计，因此可以做到语言无关。

UTS 的输出与 FastSpeech2 一样，是梅尔频谱图。声码器（HiFi-GAN [33]）将其转换为实际的语音波形。

在常规文本转语音系统中，训练需要目标语音及其对应的文本标签。相比之下，所提出的 UTS 仅需要目标语音，无需文本标签。目标语音通过 STU 处理，以获得与语音波形相关联的语音单元序列，并用于训练 UTS（图 6）。在我们的实验中，UTS 使用来自 LJSpeech 单个说话人的语音数据 [27]，以及从旁白数据中获取的其他说话人数据进行训练。

4 系统配置

我们使用 PyTorch 框架设计了 WESPER 模型。STU 基于 HuBERT，UTS 基于由 [8]修改实现的 FastSpeech2PyTorch 实现。我们使用 Librispeech 和 wTIMIT（两者均包含正常语音和耳语语音）进行预训练。使用双 NVIDIA R6000，预训练耗时 48 小时。每个目标语音的 UTS 训练耗时 26 小时。

在单块 NVIDIA R6000 上，执行转换所需的总处理时间约为实际语音时长的 1/20；在 Apple M1 Max CPU 上则约为 1/10。随附视频展示了转换的实际效果。

我们设计了两种类型的接口。第一种采用按键通话模式，用户在按住按钮的同时，先以耳语或正常语音说话。随后，松开按钮时，这段时间内录制的语音波形会被发送至语音转换神经网络，转换结果会立即播放出来。另一种接口会检测输入中的非音频时段，并在无需用户进行额外操作的情况下，自动转换语音片段。

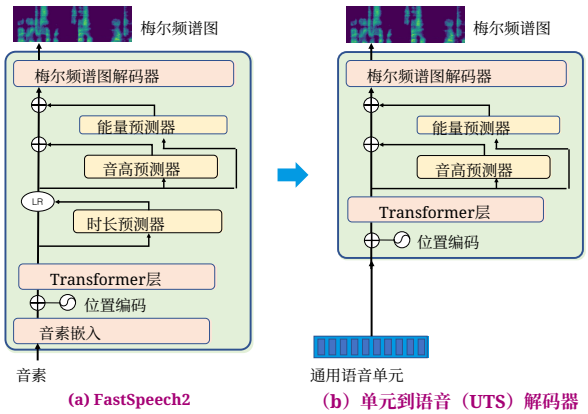

图 5：FastSpeech2 [49]和 UTS 解码器的比较：FastSpeech2 需要预测每个音素的时长，而 UTS 接受时长相同的通用语音单元。这样便无需使用时长预测器和长度调节器（图中的 LR）。由于通用语音单元不由离散令牌组成，因此也可以省略音素嵌入。

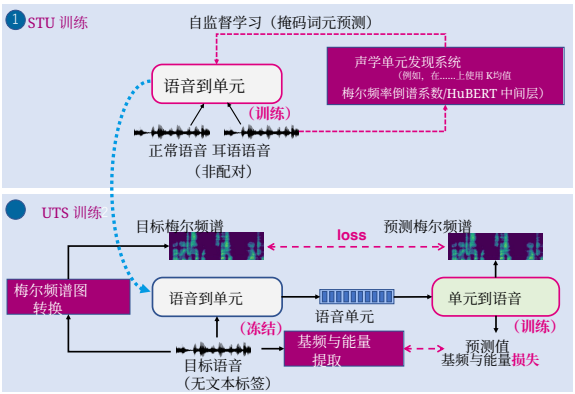

图 6：STU 训练（1）和 UTS 训练（2）：使用冻结的 STU，UTS 仅利用目标语音的波形数据学习将语音单元解码为目标语音。无需带文本标注的数据集。值得注意的是，WESPER 并未在带标注语料库上进行训练，而只是使用正常语音和耳语语音进行预训练。

图 7 展示了当前 WESPER 的语音输入配置。我们使用了 (a) 头戴式耳机、(b) 配备四个阵列式微机电系统（MEMS）麦克风的定向麦克风 [52]，以及 (c) 带有防喷罩的手机麦克风，以避免耳语时产生爆破音，并使用隔音材料降低环境噪声。

5 评估

WESPER 机制支持不受输入说话人影响的耳语转正常语音。在此，我们从三个方面评估转换

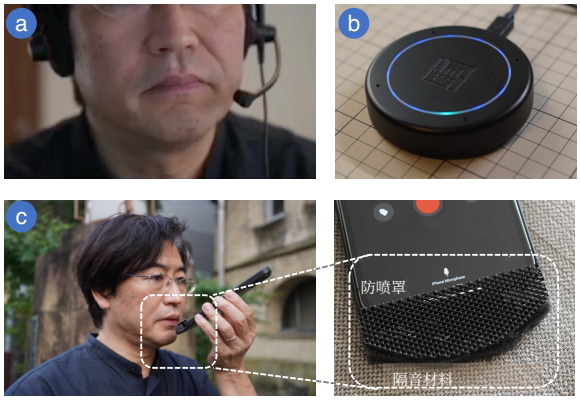

图 7：WESPER 语音输入设备：(a) 头戴式耳机、(b) 阵列式（定向）麦克风，以及 (c) 配备防喷罩和隔音材料的手机麦克风。

质量，涵盖耳语转正常语音，以及为有言语和听力障碍的人重建语音。

5.1 耳语语音到正常语音转换的质量

为了评估转换语音的质量，我们通过 Prolific 众包平台在线招募了 50 名性别均衡的参与者 [25]，，他们均已年满 18 岁且英语流利。每位参与者通过网页界面聆听四组正常语音、耳语语音和由 WESPER 转换的耳语语音（共 12 种声音），并使用 5 分制平均意见分（MOS）及其他问卷对这些话语进行评分（所附视频中包含声音示例）。

我们使用经 LJSpeech 训练的 WESPER 语音作为目标语音，并为所有语音使用相同的转写句子，以避免因句子内容不同而产生印象差异。

结果如图 8 所示。图 8(a) 展示了 MOS 评估结果。经 WESPER 转换的语音得分介于正常语音和耳语语音之间。结果证实，WESPER 转换提升了原始耳语语音的平均意见分（$p < 0.01$ 通过成对 t 检验，Cohen's d 效应量为 $= 0.54$）。图 8(b) 展示了对“这段语音听起来是嘶哑还是正常？”这一问题的回答；经 WESPER 转换的语音有明显改善（$p < 0.01$）。图 8(c) 展示了对“这段语音是否采用一致的发音、标准的语调和韵律？”这一问题的回答。在这方面，WESPER 与耳语语音的结果几乎相同。值得注意的是，WESPER 没有影响韵律的自然度。考虑到经 WESPER 转换的语音是由单元转语音模块和声码器生成的，这一结果表明，WESPER 能够保留原始语音的自然韵律。

多重刺激隐藏参考与锚定（*MUSHRA*）评估。除了 MOS 测试外，我们还使用多重刺激隐藏参考与锚定（MUSHRA）评估了语音质量。MUSHRA 是一种根据 ITU-R 建议书 BS.1534-3 [56] 定义的音频感知质量评估方法。MUSHRA 使用锚定音频和其他音频，并要求评估者

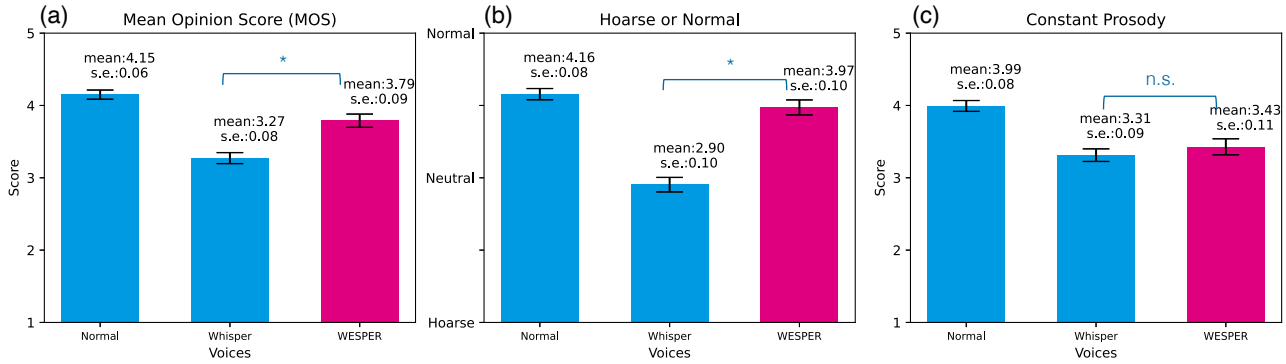

图 8：耳语转正常语音的质量：(a) 正常语音、耳语声及经 WESPER 转换的耳语语音的平均意见得分（MOS）。(b) 沙哑正常语音评分，以及 (c) 自然韵律评分（标准误：标准误，⋆：通过 t 检验得出 $p < 0.01$，n.s.：不显著）

模型	训练（微调）	test	词错误率（%）	字符错误率（%）	BLEU
谷歌	（由谷歌训练）	wTIMIT（N）	11.55	4.66	0.76
		wTIMIT（W）	44.70	28.38	0.34
		WESPER[wTIMIT（W）]	**26.68**	**12.70**	**0.52**
HuBERT base	Librispeech	wTIMIT（N）	21.06	8.17	0.54
	Librispeech	wTIMIT（W）	33.06	15.45	0.38
	Librispeech +wTIMIT（N,W）	wTIMIT（W）	**13.75**	**5.47**	**0.70**

wTIMIT（N）：wTIMIT [36] 正常语音　　wTIMIT（W）：wTIMIT 耳语语音
谷歌：Google Cloud Speech-to-Text [24]　　WESPER[•]：WESPER 转换后 •
HuBERT：使用 Librispeech [42] + wTIMIT(N,W) 预训练的 HuBERT 基础模型

表2：Google Cloud Speech-to-Text [24] 作为参考自动语音识别系统时的耳语语音识别准确率：在正常语音自动语音识别中，耳语声的识别率并不高；然而，使用 WESPER 将耳语声转换为正常语音后再进行识别，识别率有所提升。值得注意的是，WESPER 未使用标注数据进行训练，而是基于（未标注且未配对的）正常语音和耳语声进行预训练。

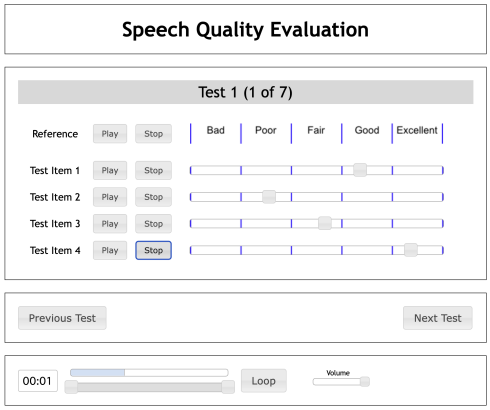

图 9：MUSHRA 评估网页界面示例。

将音频与作为参考的锚定音频进行比较，并为其打分（0 至 100 分）。每次评分时，参与者都可以随意多次播放每种语音。由于包含锚定音频和隐藏参考音频，MUSHRA 被认为比平均意见分更可靠。

我们通过 Prolific 众包平台，在互联网上招募了 50 名年龄超过 18 岁、性别均衡且讲英语的参与者 [25]。我们使用基于 JavaScript 的 MUSHRA 测试工具 [28]进行在线评估（图 9；该系统可从 https: //github.com/rkmt/mushraJS_prolific 获取）。我们使用了与之前 MOS 测试中相同的耳语声-正常语音样本。参与者的回答通过互联网收集。

结果如图 10 所示。该结果再次验证了 MOS 评估的结果（$p < 0.01$通过成对 t 检验，效应量 = 0.57）。

根据这些评估结果，我们可以得出以下结论：

- WESPER 可以将耳语语音转换为正常语音。
- 经 WESPER 转换的语音数据，其平均意见分高于耳语源语音。

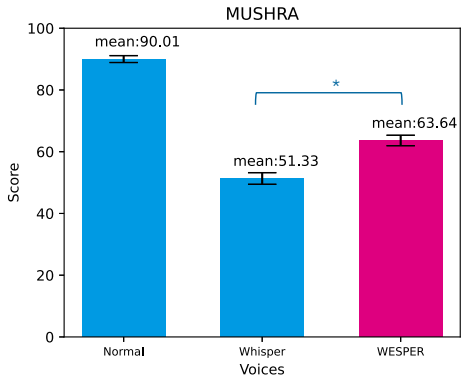

图 10：通过 MUSHRA（多重刺激隐藏参考与锚定）评估耳语声与 WESPER 转换后语音的语音质量。（*: $p < 0.01$，t 检验）

- WESPER 保留了源耳语语音的自然韵律。

5.2 语音识别准确率

使用 WESPER 作为语音识别器有两种可能的方式：使用 WESPER 进行语音识别，以及使用其他语音识别器识别经 WESPER 转换的语音。

在前一种方法中，使用耳语语音和正常语音预训练的 WESPER 模型（基于 HuBERT）会使用耳语语料库进行微调，并从耳语声中推断文本。在后一种情况下，经 WESPER 转换的耳语语音可以控制任何其他语音设备。如果耳语语音能够转换为正常语音，那么现有的语音设备便可立即使用，无需针对耳语语音对其进行修改。

我们以现有的语音识别器（谷歌云语音转文本 [24]）作为参考，测量了正常语音、耳语语音以及经 WESPER 转换的耳语语音的语音识别准确率。

测量采用了 wTIMIT语料库。wTIMIT 是一个符合 TIMIT 标准的转写语料库，其中包含带有标签的正常语音和耳语语音。该语料库以正常语音和耳语语音的识别准确率为基准，评估了经 WESPER 转换的耳语语音的识别准确率。

结果以词错误率（WER）、字符错误率（CER）以及双语评估替补（BLEU）的形式汇总于表2中。如表中所示，直接识别耳语语音时，识别准确率并不高（WER=44.70%），但使用 WESPER 进行转换后，准确率有所提高（WER=26.68%）。

使用 wTIMIT（W）进行测试时，在 Librispeech 和 wTIMIT（N、W）上预训练的 HuBERT 基础版表现优于 Google Cloud Speech-to-Text（HuBERT：词错误率=33.06%，谷歌：词错误率=44.70%）。据推测，使用耳语声和正常语音的混合语料进行预训练（但不使用耳语声进行微调）可能有助于提高自动语音识别的准确率。

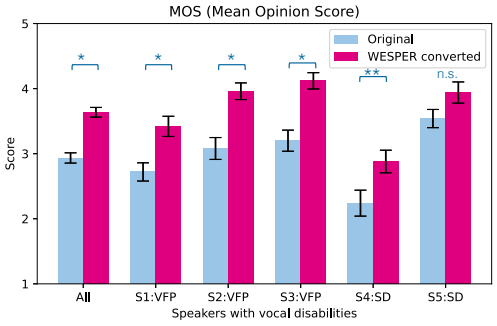

图11：按五点式MOS对言语障碍患者的语音质量评估排名：（*： $p < 0.01$，： $p < 0.05$，S1-S5：说话人，VFP：声带息肉，SD：痉挛性发声障碍）**

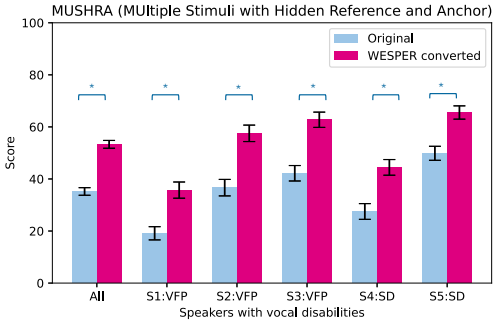

图 12：按 MUSHRA 排名的发声障碍人士语音质量评估。(*: $p < 0.01$，说话者1-S5：说话人，VFP：声带息肉，SD：痉挛性发声障碍)

值得注意的是，WESPER 并未使用带标注语料库进行训练，而是使用耳语声和正常语音进行预训练，从而提高了语音识别准确率。

因此，通过 WESPER 转换进行语音识别，无需为特定或未指定说话人准备耳语语音语料库。

5.3 针对言语障碍人士的语音重建评估

WESPER 的一个重要目标是重建言语障碍或听力障碍人士的非典型语音，使不熟悉其个人言语模式的人也能更容易听懂他们的讲话。

发声障碍是指不自主地产生嘶哑、气息声或紧绷的声音，或音量或基频偏低。其成因多种多样，包括痉挛、声道息肉及其他原因。如果因喉癌或其他原因切除了声带，发声会变得极其困难。对于听力障碍者，即使声带未受影响，控制声带也会变得困难，从而导致发声障碍。电子喉用于让喉部机械振动，为

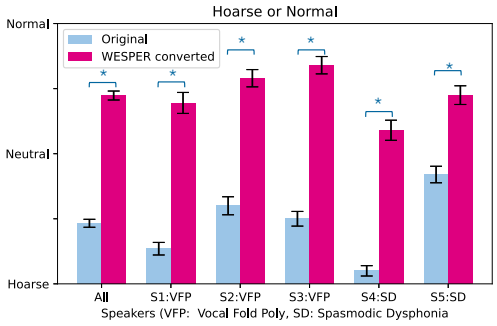

图 13：言语障碍者话语的嘶哑—正常评估（∗：$p < 0.01$）。

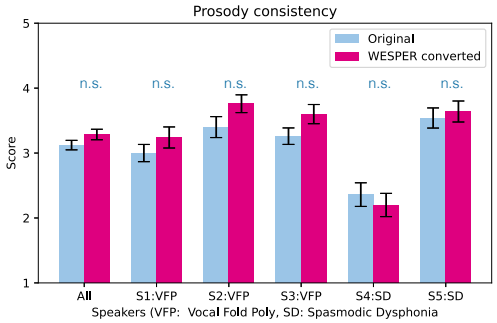

图 14：言语障碍者话语的言语韵律一致性。（n.s.：无显著差异）

声带损伤者发声，但其产生的声音是人工的，音高转换具有确定性，且与正常发声存在很大差异。发声障碍造成的沟通障碍是一个严重问题，通过语音转换技术消除这一障碍应具有重大的社会价值。

为考察 WESPER 语音转换所带来的改善效果，我们评估了两类言语障碍者的言语发声，具体如下。

声带息肉（下文简称 VFPs）：声带息肉是喉部最常见的良性病变之一，会影响发声质量。

痉挛性发声障碍（以下简称 SD）：这也称为喉肌张力障碍。SD 是另一种常见的神经系统疾病，会影响声音和言语。这是一种终身疾病，会导致发声肌肉痉挛。

我们使用萨尔布吕肯语音数据库（SVD）语料库 [1, 45]，从患有 VFP 和 SD 的人群中抽取话语。SVD 是一个常用的言语障碍患者语音语料库，其中包含元音发声录音，以及一句德语参考句“早上好，您好吗？”（“早上好，您好吗？”）的录音。该句用于评估。

评估通过 Prolific [25] 招募了 50 名参与者。参与者的性别比例均衡，且年龄均在 18 岁以上。这些参与者除英语外，他们也能流利使用德语，因为所用的参考句子是德语。

结果如图 11、12、13 和 14 所示。图 11 展示了平均意见分（MOS）方面的结果。对于 SD 和 VFP，WESPER 转换语音的 MOS 评分（$p < 0.01$，效应量 =0.66）和 MUSHRA 评分（$p < 0.01$，效应量 =0.79）均较高。图 13 展示了对问题“这个声音是嘶哑还是正常？”的回答；与原始 VFP 和 SD 语音相比，WESPER 转换语音有明显改善（$p < 0.01$）。图 14 展示了对问题“该声音的发音是否连贯、语调是否标准、韵律是否自然？”的回答。在这方面，WESPER 转换语音与原始语音的评分几乎相同，不过 WESPER 转换语音的评分略高。可以推断，WESPER 没有影响原始语音的韵律。

根据这些评估，我们可以得出以下结论。

- VFP 和痉挛性发声障碍（SD）患者的 WESPER转换语音质量更好。这表明，对于不熟悉他们各自言语模式的人而言，WESPER 可以提高这些患者语音的可懂度。
- 同样，WESPER 可以提高原始 VFP 和痉挛性发声障碍（SD）语音的自然度。
- WESPER 还可以保留源语音的自然韵律。

值得注意的是，这项测试使用德语句子进行。尽管 WESPER 仅使用英语语音进行预训练，并未使用德语语音，但源语音的韵律得以保留，且其平均意见分有所提高。因此，这一结果或许表明 WESPER 模型具有语言无关能力。随附视频展示了日语耳语转正常语音的示例。由于作为 STU 基础模型的 wav2vec 2.0 也展现出语言无关的预训练表现，[9]，WESPER 的语言无关能力或许是未来值得评估的一项特性。

5.4 听力障碍人士的语音重建评估

最后，我们评估了 WES-PER 语音重建对听力障碍人士的影响。听力损失人士听不到自己的语音，也听不到他人的语音；因此，他们往往难以做到让普通说话人轻松听懂自己说的话。然而，由于他们的发声器官正常，其语音特征与言语障碍人士不同。

我们使用了“用于声学和语音产生研究的聋人语音语料库”[40]。语料库中包含五位说话人的话语（一位听力正常，另外四位有听力障碍）。我们通过 Prolific [25],招募了 50 名评估参与者，他们的性别比例均衡，均能流利使用英语，且年龄均在 18 岁以上。我们请参与者从平均意见分、嘶哑度和自然韵律三个方面，对每条话语进行 5 分制评分。

结果见图 15、16、17 和 18。这些结果表明，MOS、MUSHRA 及其他评分

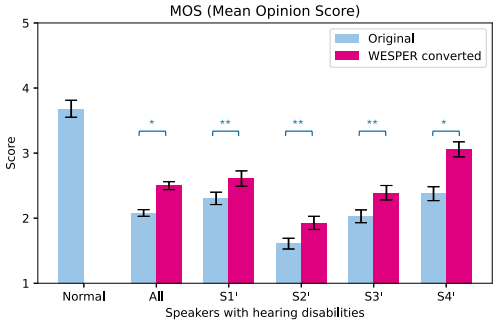

图15：听力障碍人士的语音质量评估：语音质量采用五点式MOS进行评定。（S1'-S4'：说话人，*：$p < 0.01$，**：$p < 0.05$）

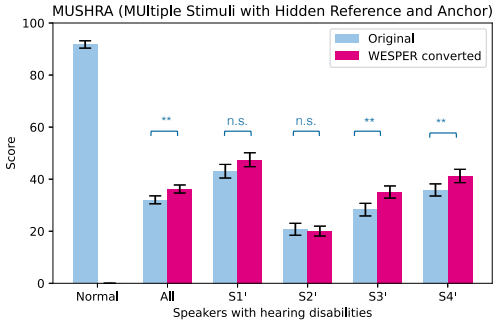

图16：听力障碍人士的语音质量评估：语音质量采用MUSHRA进行评定。（S1'-S4'：说话人，**：$p < 0.05$）

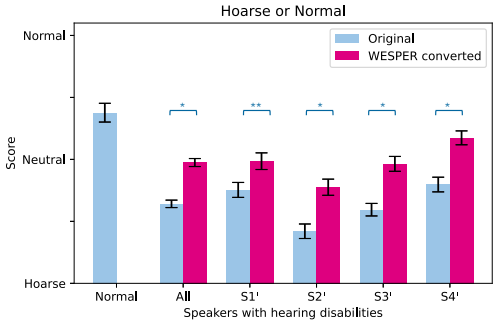

图17：听力障碍人士话语的嘶哑—正常评估。（*：$p < 0.01$，**：$p < 0.05$）

经 WESPER 转换的语音评分更高（$p < 0.05$，MOS 的效应量为 $= 0.45$；$p < 0.05$，MUSHRA 的效应量为 $= 0.21$），但提升幅度低于嗓音障碍说话人的语音。此外，与听力正常者的语音相比，听力受损者的语音韵律评分显著更低。因此，这些结果表明，听力损失者在说话时可能难以控制韵律。

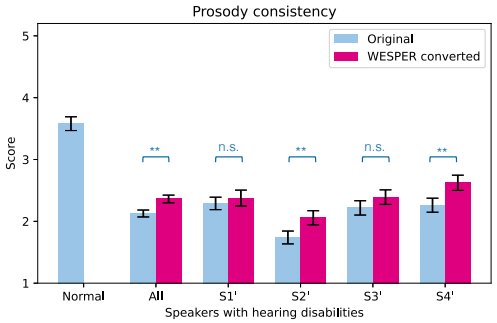

图18：听力障碍人士的语音自然韵律一致性。（**：$p < 0.05$）

所有实验的结果汇总于表 3。

6 讨论

*WESPER*与用户相关微调的结合。评估实验结果表明，听力障碍者语音的改善程度低于言语障碍者语音的改善程度。因此，我们推测前者的韵律变化较为显著。尽管本研究的主要目的是仅通过预训练实现说话人无关语音转换，但我们计划开展更多实验，以探究对每位说话人进行小规模微调能否提升转换性能。

即便如此，我们仍认为语音对是必要的，但不需要文本转写。与自动语音识别微调类似，可以使用经 STU 转换为共同语音单元的耳语和嘶哑语音话语对以及正常语音，来代替文本转写。

适用于耳语语音的音频输入设备。正如本研究所示，耳语或嘶哑嗓音可以转换为正常嗓音。实际应用中，选择合适的音频输入设备至关重要。目前，我们正在使用普通耳机和定向阵列麦克风（为智能音箱设计）测试所提出的方法，并已取得良好效果。对于可穿戴设备，可使用检测皮肤振动的非听觉发声（NAM）麦克风来采集不可听语音 [20]。结合 [7, 61] 等降噪技术也是未来的一项重要研究方向。

Philips 和 Dyson 也在开发能够提供动力呼吸通气的口罩，以防护空气污染和传染病 [13, 26]。在此类口罩内放置麦克风，就能拾取耳语声，其效果几乎等同于无声语音。

人机智能融合。本研究探讨将未指定说话人的耳语语音转换为正常语音的机器学习技术。然而在实践中，我们发现，用户注意到一些相似的耳语话语很容易转换为正常语音，而另一些则很难转换。他们尝试依靠机器学习技术来让用户

说话者类型		MOS	MUSHRA	Q1	Q2
正常		4.15	90.01	4.16	3.99
Whisper		3.27	51.33	2.90	3.31
WESPER[Whisper]		3.79	63.64	3.97	3.43
言语障碍说话者					
VFP	S1	2.72	19.13	1.54	3.00
	WESPER[S1]	3.42	35.72	3.78	3.24
	S2	3.08	36.66	2.20	3.40
	WESPER[S2]	3.96	57.53	4.16	3.76
	S3	3.20	42.19	2.00	3.26
	WESPER[S3]	4.12	62.74	4.36	3.60
SD	S4	2.24	27.51	1.20	2.36
	WESPER[S4]	2.88	44.45	3.36	2.20
	S5	3.54	49.86	2.68	3.54
	WESPER[S4]	3.94	65.53	3.90	3.64
All		2.93	35.20	1.93	3.12
WESPER[所有]		3.64	53.30	3.89	3.29
听力障碍说话者					
正常		3.68	91.75	3.76	3.75
S1'		2.30	43.05	2.29	2.51
WESPER[说话者1']		2.61	47.47	2.38	2.97
S2'		1.61	20.76	1.74	1.84
WESPER[S2']		1.93	20.05	2.06	2.55
S3'		2.03	28.30	2.19	2.22
WESPER[S3']		2.39	35.07	2.39	2.93
S4'		2.38	35.86	2.26	2.59
WESPER[S4']		3.06	41.23	2.62	3.35
All		2.08	32.07	2.28	2.12
WESPER[全部]		2.50	36.21	2.94	2.36

平均意见分 MUSHRA 多重刺激隐藏参考与锚定 问题1 "这个声音是嘶哑还是正常？"问题2 "声音是否使用一致的发音、标准语调和韵律？" WESPER[•] WESPER 转换了 •VFP 声带息肉说话人 SD 痉挛性发声障碍说话人 S1-S5、S1'-S4' 说话者编号

表3：语音转换质量评估汇总（MUSHRA：100分制，其他：5分制）

一方。这种互动表明，机器学习不仅单方面扩展人类能力，而且通过人类一方的学习也能实现进一步的协同效应。这可以被视为人机智能融合 [47] 通过发声实现的一个实际示例。

7 结论

在本研究中，我们提出了 WESPER，作为一种将耳语语音实时转换为正常语音的机制。我们证实，即使耳语语音和正常语音的声学特征不同，也可以通过自监督学习获得一种通用语音单元。只需使用任意目标说话人的语音数据，无需文本标签，即可学习语音再现。无需针对每位用户分别训练，也不需要耳语语音与正常语音的平行数据。WESPER 可以根据语音单元重建任意目标说话人的话语，且只需要目标说话人的无标注语音数据。我们证实，转换语音的质量有所提升，同时保留了语音的韵律。此外，我们还报告了对言语障碍者和听力障碍者语音话语重建结果的评估。

致谢

我们感谢审稿人提出的宝贵意见。本研究得到JST Moonshot研发计划资助（项目编号：JP-MJMS2012）、JST CREST资助（项目编号：JPMJCR17A3）、NICT日本委托研究项目，以及东京大学人类增强研究计划的支持。

参考文献

[1] A. Al-Nasheri、G. Muhammad、M. Alsulaiman 和 Z. Ali。2017 年。利用相关函数研究不同频率区域的语音病理检测与分类。《语音学杂志》(2017)。https://doi.org/10.1016/j. jvoice. 2016 01 014 . . Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael [2] Auli。2020 年。wav2vec 2.0：一种用于语音表征自监督学习的框架。*arXiv [cs.CL]* (2020 年 6 月)。[3] Abdelkareem Bedri、Himanshu Sahni、Pavleen Thukral、Thad Starner、David Byrd、Peter Presti、Gabriel Reyes、Maysam Ghovanloo 和 Zehua Guo。2015 年。迈向消费级可穿戴设备的无声语音控制。*Computer* 48, 10 (2015), 54–62。https://doi.org/ 10 1109 2015 310 . /MC. . Fadi Biadsy、Ron J. Weiss、Pedro J. Moreno、[4] Dimitri Kanevsky 和 Ye Jia。2019 年。Parrotron：一种端到端语音到语音转换模型及其在听力障碍语音和语音分离中的应用。https: //doi.org/10.48550/ARXIV.1904.04169[5] Zalán Borsos、Raphaël Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matt Sharifi、Olivier Teboul、David Grangier、Marco Tagliasacchi 和 Neil Zeghidour。2022 年。AudioLM：一种音频生成的语言建模方法。https://doi.org/ 10 48550 2209 03143 . /ARXIV. . Heng-Jui Chang、[6] Alexander H Liu、Hung-Yi Lee 和 Lin-Shan Lee。2020 年。采用频率加权方法和伪耳语预训练的端到端耳语语音识别。(2020 年 5 月 2020 2005 01972 [ ])。arXiv: . cs.CL Ishan [7] Chatterjee、Maruchi Kim、Vivek Jayaram、Shyamnath Gollakota、Ira Kemelmacher、Shwetak Patel 和 Steven M. Seitz。2022 年。ClearBuds。发表于第*20*届年度国际移动系统、应用与服务会议论文集。ACM。https://doi.org/ 10 1145 3498361 3538933 . / . Chung-Ming Chien、Jheng-Hao [8] Lin、Chien-yu Huang、Po-chun Hsu 和 Hung-yi Lee。2021 年。研究如何将预训练和可学习的说话人表征融入多说话人多风格文本转语音。发表于*ICASSP 2021 - 2021*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。8588–8592。https://doi.org/10.1109/ICASSP39728.2021.9413880[9] Alexis Conneau、Alexei Baevski、Ronan Collobert、Abdelrahman Mohamed 和 Michael Auli。2020 年。用于语音识别的无监督跨语言表征学习。https://doi.org/ 10 48550 2006 13979 . /ARXIV. . Marius Cotescu、[10] Thomas Drugman、Goeric Huybrechts、Jaime Lorenzo-Trueba 和 Alexis Moinet。2019 年。耳语语音合成的语音转换。(2019 年 12 月 2019 1912 05289 [ ])。arXiv: . cs.SD B. Denby、T. Schultz、K. Honda、T. Hueber、J. M. Gilbert 和 J. S. Brumberg。2010 年。[11] 无声语音接口。*Speech Commun.* 52, 4 (2010 年 4 月), 270–287。https://doi.org/ 10 1016 2009 08 002 . /j.specom. . . Jacob Devlin、Ming-Wei Chang、[12] Kenton Lee 和 Kristina Toutanova。2018 年。BERT：用于语言理解的深度双向变换器预训练。https://doi.org/ 10 48550 1810 04805 . /ARXIV. . Dyson。 . dyson zone：[13] 2022 配备主动降噪功能的空气净化耳机。https://www.dyson.co.uk/en. [14] Joo Freitas、Antnio Teixeira、Miguel Sales Dias 和 Samuel Silva。2016 年。无声语音接口简介(第1版)。Springer Publishing Company, Incorporated。[15] Masaaki Fukumoto。2018 年。SilentVoice：通过吸气语音实现不易察觉的语音输入。发表于第*31*届*ACM*用户界面软件与技术年度研讨会论文集(德国柏林) *(UIST '18)*。计算机协会，纽约州纽约市，美国， 237 246 10 1145 3242587 3242603 – 。[16] https://doi.org/ . / . Teng Gao、Jian Zhou、Huabin Wang、Liang Tao 和 Hon Keung Kwan。 . 2021 用于耳语到正常语音转换的注意力引导生成对抗网络

转换。https://doi.org/ 10 48550 2111 01342 . / [17] ARXIV. . Alex Graves、Santiago Fernández、Faustino Gomez 和 Jürgen Schmidhuber。2006 年。连接时序分类：利用循环神经网络标注未分段序列数据。发表于第*23*届国际机器学习会议(美国宾夕法尼亚州匹兹堡) *(ICML '06)*。计算机协会，纽约州纽约市，美国，369–376。https: //doi.org/ 10 1145 1143844 1143891 . / . [18] Dorde T. Grozdic 和 Slobodan T. Jovicic。2017 年。使用深度去噪自编码器和逆滤波进行耳语语音识别。*IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 25, 12 (2017 年 12 月), 2313–2322。https://doi.org/10. 1109/TASLP.2017.2738559[19] Tomoki Hayashi、Wen-Chin Huang、Kazuhiro Kobayashi 和 Tomoki Toda。2021 年。非自回归序列到序列语音转换。https: //doi.org/10.48550/ARXIV.2104.06793[20] Panikos Heracleous、Yoshitaka Nakajima、Hiroshi Saruwatari 和 Kiyohiro Shikano。2005 年。一种用于语音识别以保护隐私的组织传导式声学传感器。发表于*2005*年智能物体与环境智能联合会议论文集：创新情境感知服务: 用途与技术(法国格勒诺布尔) *(sOc-EUSAI '05)*。计算机协会，纽约州纽约市，美国，93 97 10 1145 1107548 1107577 – 。https://doi.org/ . / . [21] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。2021 年。HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习。*IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (2021 年 1 月), 3451–3460。https://doi.org/ 10 1109 2021 3122291 . /TASLP. . [22] Wen-Chin Huang、Tomoki Hayashi、Shinji Watanabe 和 Tomoki Toda。2020 年。2020年语音转换挑战赛的序列到序列基线：级联自动语音识别与文本转语音。https://doi.org/ 10 48550 2010 02434 . / [23] ARXIV. . Amazon.com Inc. . Alexa 如何变得越来越智能。2018。https://www.aboutamazon.com/devices/how-alexa-keeps-getting-smarter [24] Google Inc. 2020 年。Google Cloud Speech-to-Text。https://cloud.google.com/speech- to-text.[25] Prolific inc. 2014 年。Prolific。https://www.prolific.co [26] Philips Inc. 2021 年。Fresh Air Mask Series 6000。https://www.philips.com.sg/c- p/ACM066_01/fresh-air-mask-series-6000。[27] Keith Ito 和 Linda Johnson。2017 年。LJ语音数据集。https://keithito.com/LJ- Speech-Dataset/. [28] jfsantos。2019 年。mushraJS。https://github.com/jfsantos/mushraJS [29] Hirokazu Kameoka、Takuhiro Kaneko、Kou Tanaka 和 Nobukatsu Hojo。2018 年。StarGAN-VC：采用星形生成对抗网络的非平行多对多语音转换。https://doi.org/ 10 48550 1806 02169 . /ARXIV. . Takuhiro Kaneko 和 Hirokazu Kameoka。. 利用循环一致对抗网络实现无需平行数据的语音转换。https://doi.org/10.48550/ ARXIV.1711.11293[31] Arnav Kapur、Shreyas Kapur 和 Pattie Maes。2018 年。AlterEgo：个性化可穿戴无声语音接口。发表于第*23*届国际智能用户界面会议(日本东京) *(IUI '18)*。ACM，43–53。https://doi.org/10.1145/ 3172944.3172977 [32] Naoki Kimura、Michinari Kono 和 Jun Rekimoto。2019 年。SottoVoce：一种基于超声成像并采用深度神经网络的无声语音交互。发表于*2019*年*CHI*计算系统中的人为因素会议论文集(英国苏格兰格拉斯哥) *(CHI '19)*。计算机协会，纽约州纽约市，美国，1 11 10 1145 3290605 3300376 – 。https://doi.org/ . / . [33] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。. HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。https: //doi.org/ 10 48550 2010 05646 . /ARXIV. . Felix Kreuk、Adam Polyak、Jade Copet、Eugene Kharitonov、Tu-Anh Nguyen、Morgane Rivière、Wei-Ning Hsu、Abdelrahman Mohamed、Emmanuel Dupoux 和 Yossi Adi。2021 年。使用离散与分解表征进行无文本语音情感转换。https://doi.org/ 10 48550 2111 07402 . /ARXIV. . [35] Kushal Lakhotia、Evgeny Kharitonov、Wei-Ning Hsu、Yossi Adi、Adam Polyak、Benjamin Bolte、Tu-Anh Nguyen、Jade Copet、Alexei Baevski、Adelrahman Mohamed 和 Emmanuel Dupoux。2021 年。基于原始音频的生成式口语语言建模。https://doi.org/ 10 48550 2102 01192 . /ARXIV. . [36] Boon Pang Lim。2010。耳语与非耳语语音之间的计算差异，博士论文，伊利诺伊大学厄巴纳-香槟分校。[37] H. Malaviya、J. Shah、M. Patel、J. Munshi 和 H. A. Patil。2020 年。MSpec-Net：多领域语音转换网络。发表于*ICASSP 2020 - 2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。7764–7768。[38] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。2017 年。Montreal Forced Aligner：使用 Kaldi 进行可训练的文本-语音对齐。498 502 10 21437 2017 1386 – 。https://doi.org/ . / [39] Interspeech. - Leland McInnes、John Healy 和 James Melville。. UMAP：用于降维的统一流形逼近与投影。2018。https://doi.org/10. 48550/ARXIV.1802.03426[40] Lisa Lucks Mendel、Sungmin Lee、Monique Pousson、Chhayakanta Patro、Skylar McSorley、Bonny Banerjee、Shamima Najnin 和 Masoumeh Heidari Kapourchali。2017 年。用于声学和语音产生研究的聋人语料库。*The* 《美国声学学会期刊》142, (1) (2017), EL102. https://doi. org/ 10 1121 1 4994288 . / . [41] Marco A. Oliveira. 2022. 耳语到正常语音转换的机器学习方法：综述。波尔图大学工程学期刊8, 2 (2022), 202–212. https://doi.org/ 10 24840 2183 6493_008 002_0016 . / - . [42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech：一个基于公有领域有声书的自动语音识别语料库。发表于*2015 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。5206 5210 10 1109 2015 7178964 – . https://doi.org/ . /ICASSP. . [43] Santiago Pascual, Antonio Bonafonte, Joan Serrà, and Jose A. Gonzalez. 2018. 使用生成对抗网络进行耳语到有声无喉语音转换。https://doi.org/ 10 48550 1808 10687 . /ARXIV. . Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. 使用 Wav2vec 2.0嵌入进行语音情感识别。(2021年4月)。arXiv: 2104 03502 [ ] . cs.SD Manfred Putzer and William J. Barry. 2016. 萨尔布吕肯语音数据库。https: //stimmdb.coli.uni-saarland.de[46] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa- Johnson. 2019. AutoVC：仅使用自编码器损失实现零样本语音风格迁移。发表于第*36*届国际机器学习大会论文集（机器学习研究会论文集，第*97*卷），Kamalika Chaudhuri 和 Ruslan Salakhutdinov（编）。PMLR，5210–5219。https://proceedings.mlr.press/v97/qian19c.html [47] Jun Rekimoto. 2019. 赛博人：人机智能融合时代。https://doi.org/ 10 48550 1911 02637 . /arXiv. . Jun Rekimoto. . DualVoice：可区分正常语音和耳语声输入的语音交互。2022。发表于第*35*届*ACM*用户界面软件与技术年度研讨会论文集（美国俄勒冈州本德）*(UIST '22)*。计算机协会，纽约，美国，文章编号91，共10页。https://doi.org/ 10 1145 3526113 3545685 . / . Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. FastSpeech 2：快速、高质量的端到端文本转语音。https://doi.org/10.48550/ARXIV.2006.04558[50] Gabriel Reyes, Dingtian Zhang, Sarthak Ghosh, Pratik Shah, Jason Wu, Aman Parnami, Bailey Bercik, Thad Starner, Gregory D. Abowd, and W. Keith Edwards. 2016. Whoosh：用于智能手表的低成本、免手持、快速非语音声学输入。发表于*2016*年*ACM*可穿戴计算国际研讨会论文集（*ISWS'16*）。ACM，120–127。https://doi.org/10.1145/2971763. 2971765[51] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. 舌耳接口：用于无声语音识别的可穿戴系统。发表于*2014*年*ACM*可穿戴计算国际研讨会论文集（美国华盛顿州西雅图）*(ISWC '14)* 47 54 10 1145 2634317 2634322 。ACM，–。https://doi.org/ . / . [52] Seeed Studio. . ReSpeaker USB 麦克风阵列。2019。https://wiki.seeedstudio.com/ReSpeaker-USB-Mic-Array/ [53] Nirmesh Shah, Mihir Parmar, Neil Shah, and Hemant Patil. 2018. 用于跨域耳语转语音的全新 MMSE DiscoGAN。机器学习。发表于语音与语言处理（*MLSLP*）研讨会。[54] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact：利用无声语音指令改善移动设备交互。发表于第*31*届*ACM*用户界面软件与技术年度研讨会论文集（德国柏林）*(UIST '18)*。计算机协会，纽约，美国，581 593 10 1145 3242587 3242599 –。https://doi.org/ . / . [55] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. 2016. 无需平行数据训练的多对一语音转换音素后验概率图。发表于*2016 IEEE*多媒体与博览会国际会议（*ICME*）。1–6。https: //doi.org/ 10 1109 2016 7552917 . /ICME. . 国际电信联盟。2013. BS.1534：音频系统中间质量等级主观评估方法。https: //www.itu.int/rec/R-REC-BS.1534/en[57] Benjamin van Niekerk, Marc-Andre Carbonneau, Julian Zaidi, Matthew Baas, Hugo Seute, and Herman Kamper. 2022. 用于改进语音转换的离散和软语音单元比较。发表于*ICASSP 2022 - 2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE。https://doi.org/ 10 1109 43922 2022 9746484 . /icassp . . Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron：一个完全端到端的文本转语音合成模型。*CoRR* abs/1703.10135 (2017)。arXiv: 1703 10135 1703 10135 . http://arxiv.org/abs/ . Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. 将 Wav2vec2.0 应用于多种低资源语言的语音识别。(2020年12月 2020 2012 12121 [ ])。arXiv: . cs.CL zeta chicken. . toWhisper. 2017。https://github.com/zeta-chicken/toWhisper [61] Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2022. 鲁棒Data2vec：抗噪语音表征学习

通过结合回归和改进的对比学习实现自动语音识别。https://doi.org/10.48550/ARXIV.2210.15324