WhispEar：通过伪平行耳语生成扩展耳语语音转换的双向框架

ZihaoFang [1]*, *Yingda Shen* [1,*], *Zifan Gua* [1], *Tongtong Song*[2], *Zhenyi Liu*[2], *Zhizheng Wu* [1]

[1] 香港中文大学（深圳）数据科学学院，中国深圳[2] 荣耀终端有限公司，中国

zihaofang@link.cuhk.edu.cn, yingdashen@link.cuhk.edu.cn

arXiv:2603.08046v1 [cs.SD] 9 Mar 2026

摘要

耳语语音缺乏声带振动和基频，导致声学线索退化，使耳语转正常语音（W2N）转换面临挑战，尤其是在平行数据有限的情况下。我们提出 WhispEar，这是一种基于统一语义表示的双向框架，能够捕捉耳语语音与正常语音共享的说话模式不变信息。该框架包含 W2N 和正常语音转耳语（N2W）模型。值得注意的是，N2W 模型能够利用大量正常语音进行零样本伪平行耳语语音生成，从而为 W2N 训练提供可扩展的数据增强。增加生成数据量能够持续提升性能。我们还发布了迄今规模最大的中英双语耳语语音—正常语音平行语料库。实验表明，WhispEar 的表现优于强基线方法，并能显著受益于可扩展的伪平行数据。

索引词：耳语语音转换、伪平行数据生成、可扩展训练

1. 引言

耳语转正常语音旨在从耳语语音中重建自然且可懂的语音，这对隐私通信和语音恢复至关重要。由于缺乏声带振动和周期性激励，耳语语音存在基频缺失以及声学特征退化[1], , 给生成具有自然韵律和一致音色的语音带来了巨大挑战。尽管现有研究推动了耳语转正常语音的发展，但仍面临严重局限。一方面，这些方法严重依赖有限的平行耳语—正常语音数据[2, 3], , 而传统的基于数字信号处理的伪耳语数据[4, 5] 与真实耳语语音存在分布差异，性能提升有限。基于对抗学习的方法还进一步受到训练不稳定性的影响[6, 7]。另一方面，大多数方法难以保持说话人音色和自然韵律[8, 9], , 导致生成质量不理想、说话人相似度较低。

基于内容解耦的模型在文本转语音和语音转换领域取得了显著成功[10,11], , 这得益于它们在保持音色和韵律的同时生成高质量语音的强大能力。受此启发，近期的耳语转正常语音方法（如 WESPER[8] 和 DistillW2N[12]）引入了 HuBERT[13]等语义表征，以提升转换性能。在这项工作中，我们观察到，尽管耳语和正常语音在声学特征上有所不同，但它们共享相同的高层次语言和语义信息。基于此，耳语转正常语音和正常语音转耳语这两项任务都可以视为从相同语义内容生成不同风格语音的生成任务，这也从逻辑上支持了这两种发声模式之间的双向转换。

为解决数据稀缺问题并进一步提升耳语转正常语音性能，我们提出了一种基于语义特征的双向转换框架 **WhispEar**。借助其正常语音转耳语模块，我们可以从大规模正常语音数据集中合成高质量伪耳语数据[14]。此外，我们招募志愿者采集真实耳语语音数据。通过结合真实录制数据与高质量伪数据，我们构建了迄今该领域规模最大的双语耳语-正常语音数据集wEar。实验结果表明，我们的方法有效缓解了数据限制，并在语音自然度、可懂度、韵律恢复和说话人音色相似度方面显著优于基线模型。

本文的主要贡献如下：

- 我们提出了**WhispEar**，一种基于统一语义表示的双向耳语语音转换框架。
- 我们提出一种**基于零样本正常语音转耳语合成的伪平行耳语生成策略**，无需额外录音，即可利用丰富的正常语音大规模扩充数据。
- 我们通过逐步增加生成的伪平行数据，开展**系统的规模扩展研究**，结果表明性能持续提升，验证了以数据为中心的扩展方法在耳语语音转换中的有效性。
- 我们发布迄今规模最大的**中文和英语双语耳语—正常语音平行语料库**，其中包含录制和生成的数据，为未来研究提供了宝贵的基准。

2. 方法

WhispEar 分三个连续阶段进行训练：(1) 语义分词器蒸馏，(2) 共享流匹配声学模型训练，以及 (3) 利用可扩展的伪平行数据生成进行统一分词器训练。

2.1. 阶段1：语义分词器蒸馏

我们首先从大型自动语音识别编码器中蒸馏出一个轻量级语义分词器。给定波形 x，教师模型 f_{large}会生成语义嵌入。我们通过以下方式训练紧凑型学生模型 f_ϕ，使其模仿教师模型：

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_x \left\| f_\phi(x) - f_{\text{large}}(x) \right\|_2^2 . \quad (1)$$

*这些作者贡献相同。演示页面可通过 https://whispear-demo.github.io/ 访问

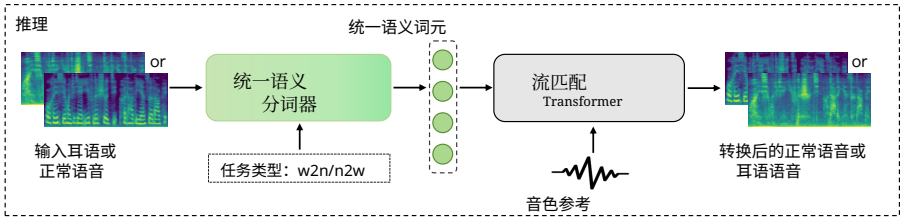

图 1： *WhispEar* 的双向转换推理流程 .

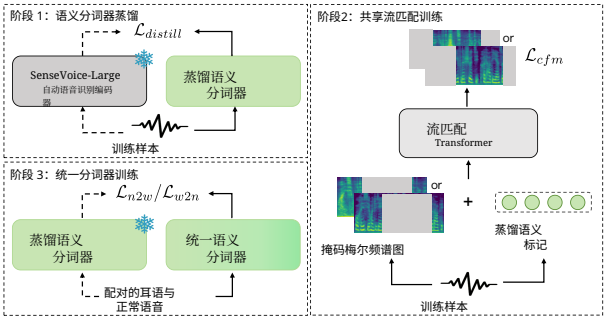

图 2： *WhispEar* 的训练流程。

学生分词器由多个堆叠的 Transformer 块组成，每个块包含基于 RoPE 的自注意力、一个 FSMN 块和一个前馈网络（FFN）。FSMN（前馈序列记忆网络）是一种轻量级时间建模模块，它利用卷积记忆机制捕捉局部序列依赖关系，无需循环连接，并能高效地进行上下文平滑。

训练数据包含耳语和正常语音，以促使模型学习说话模式不变表示。得到的嵌入使用有限标量量化（FSQ）[15] 进行量化，以获得离散语义标记。这些标记用于训练阶段 2 的声学模型。

2.2. 阶段 2：共享流匹配声学模型

我们训练一个条件流匹配 Transformer，根据离散语义标记生成梅尔频谱图。耳语转正常语音（w2n）和正常语音转耳语（n2w）共用同一个流匹配 Transformer 和同一个声码器。方向标记 $d \in \{\text{w2n}, \text{n2w}\}$ 用于指定转换方向。

重要的是，声学模型是在蒸馏分词器 f_ϕ生成的标记上训练的，而非后文介绍的统一分词器。这使声学建模与语义对齐相互独立。

与 CosyVoice2 一样，模型将语义标记$\mathbf{Z}$ 和掩码梅尔频谱图作为输入。掩码区域以高斯噪声初始化，模型沿最优传输概率路径预测掩码部分的速度场。

训练目标为

$$\mathcal{L}_{\text{cfm}} = \mathbb{E}\,\|(\mathbf{Y}_1 - \mathbf{Y}_0) - \nu_t(\mathbf{Y}_t \mid \mathbf{Z}, d)\|_1 . \tag{2}$$

其中损失仅在掩码区域上计算。模型以 CosyVoice2 为初始化，并在混合耳语与正常语音的数据上进行微调。

2.3. 阶段 3：统一分词器训练与可扩展扩充

固定声学模型后，我们训练两个统一语义分词器： f_{n2w} 和 f_{w2no} 这些模型将一种模态的语音映射到另一种模态的目标语义标记空间。

设 $\mathcal{T}(\cdot)$ 表示冻结的蒸馏分词器。对于配对耳语与正常语音 $(x_w,\ x_n)$，我们提取$\mathbf{z}_w = \mathcal{T}(x_w)$ 和$\mathbf{z}_n = \mathcal{T}(x_n)$。

训练目标如下：

$$\mathcal{L}_{\text{w2n}} = \mathbb{E}_{(x_w,x_n)\sim\mathcal{P}}\Big[\,\|f_{\text{w2n}}(x_w) - \mathbf{z}_n\|^2 + \lambda_n\,\|f_{\text{w2n}}(x_n) - \mathbf{z}_n\|^2\Big]. \tag{3}$$

$$\mathcal{L}_{\text{n2w}} = \mathbb{E}_{(x_w,x_n)\sim\mathcal{P}}\Big[\,\|f_{\text{n2w}}(x_n) - \mathbf{z}_w\|^2 + \lambda_w\,\|f_{\text{n2w}}(x_w) - \mathbf{z}_w\|^2\Big]. \tag{4}$$

步骤 1：训练较简单的 N2W 方向：从经验来看，正常语音转耳语更容易，仅用有限的平行数据就能达到良好的质量。我们首先仅使用真实配对数据训练 f_{n2w}。

步骤 2：生成伪平行数据：利用丰富的正常语音语料库和训练好的 N2W 流程，我们将真实正常语音 x_n合成为耳语语音 $\tilde{x}_w$。由此大规模生成完全对齐的伪配对数据（$\tilde{x}_w,\ x_n$）。

步骤 3：利用扩充后的数据训练较难的 W2N 方向：最后，我们使用真实配对数据和大规模伪平行数据训练 f_{w2n}。扩充后的数据集显著提升了耳语语音转正常语音性能，并支持开展系统性的扩展实验。

推理时，输入话语首先通过相应的统一分词器映射为目标语义词元，随后由共享的流匹配变换器和声码器生成输出波形。

3. 数据集

为解决大规模双语耳语-常规语音语料库稀缺的问题，我们构建了涵盖中文和英语的平行数据集**wEar**。

3.1. 数据收集

wEar（真实）语料库录制于空教室和隔音室，共有 146 位说话人，分为平行组和纯耳语组。每位说话人录制 20–40 分钟的数字和不同情绪的语句。录音使用三部手机以 48kHz/16-bit 单声道格式采集，背景噪声低于 30dB；说话人将手机放在耳边，以模拟电话通话。数据集按 91:6:3 的比例划分为训练集、验证集和测试集，且各集合之间没有说话人重叠。

表1： 从语言覆盖范围、总时长、平行语音对数量和说话人数等方面比较耳语语音转换数据集。

数据集	语言	时长（小时）	平行语音对	说话人
其他				
wTIMIT[16]	EN	26	19k	48
CHAINs[17]	EN	3	1k	36
iWhisper[18]	CN	15	–	80
wSPIRE[19]	EN	18	–	88
AISHELL6[20]	CN	30	20k	167
耳语40[21]	CN	6	1k	36
Ours				
wEar（真实）	CN	18	4k	146
wEar（伪）	英语，中文	3026	600k	230
wEar（总计）	英语，中文	**3044**	**604k**	**230**

3.2. 数据处理

波形经过重采样和归一化后，由 HuBERT ASR X-Large [13] 处理以获取表征和频谱图。经过对数 softmax 和 CTC 贪心解码后，强制对齐确定帧边界，随后进行词合并和静音裁剪。FastDTW [22] 计算对齐路径，从而生成近似一对一的帧映射。随后对齐的频谱图帧会被反转，以重建时域信号，将语义级对齐与帧级精度相结合。

3.3. 数据构建

为扩展语料库，我们使用提出的 WhispEar 模型生成大规模伪平行数据。对于每位说话人，高质量语音片段被用作音色提示，对从 Emilia 采样的话语进行语音转换 [14], ，从而生成音色匹配语音 A_1。正常语音转耳语模块（n2w）随后生成耳语语音 A_2，形成合成对（A_1, A_2）。该流程产出超过 3,000 小时的伪平行数据，大幅扩充了现有耳语–正常语音语料库，并提升了双向转换效果。表1总结了 wEar 数据集。

4. 实验

我们开展实验以回答以下评估问题：**EQ1：**与现有最先进方法相比，WhispEar 系统在耳语转正常语音（W2N）任务上的表现如何？**EQ2：**由 WhispEar 自身生成的伪耳语数据，对于训练高质量耳语转正常语音系统有多有效？**EQ3：**扩大伪耳语数据规模能否进一步提升耳语转正常语音系统的性能，以及在真实对齐数据上进行监督微调起到什么作用？

4.1. 实验设置

4.1.1. 训练数据

训练过程利用真实对齐数据和由 WhispEar-n2w 模型生成的伪耳语数据。真实对齐数据包括 wTIMIT 训练集（英语）和 wEar 训练集（中文），这些数据经对齐流程处理后，提供真实耳语语音及其对应的正常语音（见第 3 节）。为扩大训练数据规模并提升模型泛化能力，我们通过 WhispEar-n2w 生成大规模伪平行数据，涵盖 wTIMIT 的伪版本、wEar 的伪版本以及一个包含 3,000 小时数据的 Emilia 伪子集。所有数据集的统计信息汇总于表1。

4.1.2. 训练细节

我们训练了两个模型：**WhispEar** 和 **WhispEar-Scaled**。二者共享阶段1和阶段 2，仅阶段 3 不同。

阶段1：蒸馏。使用混合耳语与正常语音，从大型自动语音识别编码器中蒸馏出学生语义分词器。无需配对数据。模型使用 Adam 优化，学习率为 1×10^{-4}。

阶段 2：流匹配训练。使用蒸馏分词器生成的词元训练声学模型。训练设置遵循 CosyVoice2，但我们使用 Adam 和固定学习率 1×10^{-5} 。训练时使用混合耳语与正常语音，无需对齐。

阶段 3：统一分词器训练。

WhispEar：使用来自 wTIMIT 和 wEar 的对齐数据和伪平行数据训练统一分词器。伪配对数据通过正常语音到耳语流程生成。数据总规模约为 80 小时。

WhispEar-Scaled：我们进一步纳入由正常语音语料生成的大规模伪平行数据（约 3000 小时），用于训练耳语转正常语音分词器。

4.1.3. 评估指标

我们从四个方面评估性能：质量、可懂度、韵律和说话人相似度。质量通过 UTMOS [23], DNSMOS [24], 和 NISQA [25]进行衡量。使用 Whisper-large-v3 [26], 通过词错误率（英语）和字符错误率（中文）评估可懂度。韵律通过与目标正常语音的基频皮尔逊相关系数（F0 CoRR）进行评估；说话人相似度则计算由 wavlm-base-plus-sv [27]提取的说话人嵌入之间的余弦相似度。

4.1.4. 基线模型

我们将四种具有代表性的耳语转正常语音系统进行比较：WESPER [8], DistillW2N [12], CosyVoice2 [10], 和 MaskCycleGAN [7]。前三种基线方法使用其公开发布的检查点进行评估，而 MaskCycleGAN 则使用官方实现，在 wTIMIT 和 wEar 上重新训练。

4.2. WhispEar 的性能（EQ1）

表 2 汇报了主要结果，包括**WhispEar** 和 **WhispEar-Scaled**。

在英语基准测试中，WESPER 的音频质量尚可，但由于采用单说话人设计，说话人相似度较低。DistillW2N 通过知识蒸馏缩小了模型规模，但可懂度有所下降。CosyVoice2 针对正常语音进行了优化，在质量和相似度方面表现出色；然而，它在耳语转正常语音中的韵律恢复能力较弱，这一点体现在较低的基频相关性上。

WhispEar 在未使用大规模扩展数据进行训练的情况下（约 80 小时），在多个指标上已达到与最先进系统相当或更优的表现。加入大规模伪平行数据（约 3000 小时）后，**WhispEar-Scaled**在说话人相似度、可懂度、自然度和韵律方面均进一步稳定提升，整体表现明显优于其他系统。

在 wEar 测试集（中文）上，WESPER 和 DistillW2N（仅用英语训练）未能有效泛化，导致可懂度严重下降（字符错误率 $> 80\%$）。相比之下，WhispEar 保持了出色的跨语言性能，

表 2：比　　较基线方法在 *WTIMIT*（英语）和 *wEar*（中　　文）测试集

模型	WTIMIT（英语）						wEar（中文）					
	SIM	WER	UTMOS	DNSMOS	NISQA	F0 CoRR 相似度		CER	UTMOS	DNSMOS	NISQA	F0 CoRR
耳语语音	0.333	12.36%	1.46	1.81	1.66	0.029	0.550	18.09%	1.41	1.84	1.59	0.016
基线方法												
WESPER[8]	0.064	42.01%	3.63	**3.91**	4.37	0.373	0.004	141.36%	**3.64**	**4.05**	**4.67**	0.254
DistillW2N[12]	0.189	50.99%	1.97	3.61	2.25	0.113	0.216	128.06%	1.69	3.43	1.84	0.125
MaskCycleGAN[7]	0.237	48.53%	1.97	3.24	3.32	0.023	0.322	49.06%	1.67	3.32	2.98	0.023
CosyVoice2[10]	0.409	36.69%	2.51	2.54	3.59	0.052	0.622	29.29%	2.11	2.59	3.20	0.062
Ours												
WhispEar	0.554	30.74%	3.45	3.70	4.15	0.485	0.702	32.50%	2.59	3.51	3.78	0.259
WhispEar-Scaled	**0.577**	**22.44%**	**3.75**	3.76	**4.38**	**0.513**	**0.750**	**14.93%**	3.16	3.94	4.32	**0.396**

表3：　仅使用 *wTIMIT*，对不同的耳语数据构建策略进行消融研究，包括原始配对数据、*DSP*生成的耳语、对齐的真实配对数据、*WhispEar*生成的伪配对数据及其组合。

配置	SIM	WER	UTMOS	F0 CoRR
RAW	0.031	102.934%	1.326	0.009
传统				
DSP[5]	0.429	54.814%	2.446	0.208
提出的方法				
对齐（A）	0.416	44.594%	1.598	0.309
伪耳语（P）	0.454	49.179%	2.595	0.271
A +P	**0.513**	**34.871%**	**3.064**	**0.357**

而 WhispEar-Scaled 则进一步提升，展现了多语言和可扩展伪数据训练的有效性。

4.3. 伪耳语数据的有效性（EQ2）

表3展示了不同数据配置下的消融研究。所有训练数据均仅来自原始wTIMIT语料库；未引入任何外部数据。各配置的差异仅在于耳语语音的构建与处理方式：

•原始：直接使用wTIMIT中的原始正常语音录音和耳语录音，不做任何对齐，仅将配对样本填充至相同长度。严重的时间失配导致性能最差。

• DSP：使用wTIMIT中的原始正常语音录音，而耳语语音则通过基于传统DSP的正常语音转耳语方法[5]由这些正常信号生成。尽管时间上已匹配，信号处理转换仍引入了明显的质量差距。

• 对齐（A）：使用来自 wTIMIT 的原始正常语音和耳语录音，但采用提出的对齐流程，获取时间对齐的平行语对。

•伪耳语（P）：使用原始正常语音录音，来源于 wTIMIT，其中耳语语音由 WhispEar N2W 模型生成，以构成伪平行配对。

• A + P: 结合对齐的真实配对数据（A）和 WhispEar 生成的伪配对数据（P）。

结果表明，直接使用原始数据会严重降低性能。传统的基于DSP的生成方法相较于RAW有所改善，但仍然受到限制。相比之下，对齐（A）和由WhispEar生成的伪数据（P）均带来了显著提升，而二者的组合（A+P）在所有指标上都取得了最佳结果。这些发现表明，高质量的时间对齐和基于模型的伪耳语生成都是这对于缓解配对数据稀缺至关重要。

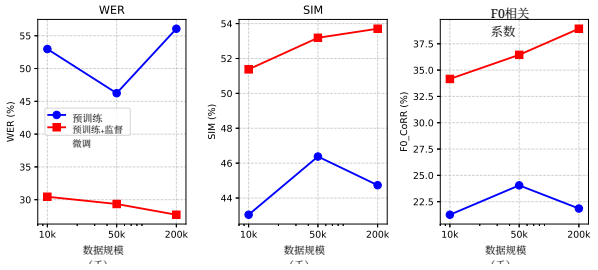

图3：　使用伪平行数据进行扩展实验。蓝色曲线表示仅使用伪数据预训练的模型，红色曲线表示预训练后进一步使用对齐的真实数据微调的模型。

4.4. 扩规模实验（EQ3）

我们研究了扩展由N2W流程生成的伪平行数据所产生的影响。首先，使用不同数量的伪数据（10k、50k和200k对）对统一分词器进行预训练，然后在两种设置下进行评估：

- **预训练：** 伪数据预训练后的直接评估。
- **预训练 + 监督微调：** 使用 EQ2 中的对齐真实数据进行额外微调。

图3显示，在仅预训练设置下，扩大伪数据规模带来的收益有限。然而，若随后使用真实对齐数据进行微调，更大规模的预训练在所有指标上都能持续提升性能。尤其是，200k 模型在微调后取得了最佳结果。

这些结果表明，大规模伪数据预训练能够提供更好的初始化，而少量真实对齐数据对于使模型有效适应耳语转正常语音任务是必不可少的。

5. 结论

我们提出 WhispEar，这是一种基于统一语义表示的双向耳语语音转换框架。通过结合共享声学建模与可扩展的伪平行数据生成，WhispEar 能够在配对数据有限的情况下实现稳定的耳语到正常语音转换，并且随着数据规模扩大，性能显著提升。我们还发布了 wEar，这是迄今规模最大的双语耳语—正常语音平行语料库。

未来的工作将聚焦于提升系统在噪声环境下的鲁棒性、拓展多语言泛化能力，并进一步提高效率，以便实际部署。

6. 生成式人工智能使用披露

在准备本研究期间，作者使用了 Gemini 来改进语言表达。使用该工具后，作者根据需要审阅并编辑了内容，并对出版物内容承担全部责任。

7. 参考文献

[1] V. C. Tartter, 《耳语中有什么？》, 《美国声学学会杂志》, 第 86 卷, 第 5 期, 第 1678–1683 页, 1989 年。[2] H. Lian、Y. Hu、W. Yu、J. Zhou 和 W. Zheng, 《利用结合听觉注意力的序列到序列映射模型进行耳语到正常语音转换》, IEEE Access, 第 7 卷, 第 130 495–130 504 页, 2019 年。 [在线]。网址: https://doi.org/10. 1109/ACCESS.2019.2940700[3] T. Gao、Q. Pan、J. Zhou、H. Wang、L. Tao 和 H. K. Kwan, 《一种用于耳语到正常语音转换的新型注意力引导生成对抗网络》, 《认知计算》, 第 15 卷, 第 2 期, 第 778–792 页, 2023 年。 [在线]。网址: https://doi.org/10.1007/s12559-023-10108-9[4] Z. Lin、T. Patel 和 O. Scharenborg, 《利用基于伪耳语语音的数据增强提升耳语语音识别性能》, 载于 IEEE自动语音识别与理解研讨会（ASRU 2023）, 台北，台湾, 2023 年 12 月 16–20 日。IEEE, 2023 年, 第 1–8 页。 [在线]。网址: https://doi.org/10.1109/ASRU57964.2023.10389801 [5] zeta, 《towhisper》, https://github.com/zeta-chicken/toWhisper。[6] M. Parmar、S. Doshi、N. J. Shah、M. Patel 和 H. A. Patil, 《跨域架构在耳语到正常语音转换中的有效性》, 载于第 27 届欧洲信号处理会议（EUSIPCO 2019）, 西班牙拉科鲁尼亚, 2019 年 9 月 2–6 日。IEEE, 2019 年, 第 1–5 页。 [在线]。网址: https: // doi.org/10.23919/EUSIPCO.2019.8902961[7] D. Wagner、I. Baumann 和 T. Bocklet, 《利用掩码循环一致生成对抗网络实现耳语语音的无声码器非平行转换》, 载于《文本、语音与对话——第 28 届国际会议（TSD 2025）论文集》, 埃朗根, 德国, 2025 年 8 月 25–28 日, 第一部分, 丛书《计算机科学讲义》, K. Ekstein 、 M. Konop´ık、O. Praz´ak 和 F. P´artl 编, 第 16029 卷。施普林格, 2025 年, 第 235–246 页。 [在线]。网址: https: // doi.org/10.1007/978-3-032-02548-7 20[8] J. Rekimoto, 《WESPER：面向基于耳语的语音交互的零样本实时耳语到正常语音转换》, 载于 2023 年 CHI计算系统中的人为因素会议论文集, CHI *2023, 23 28,* 汉堡, 德国, 4 月, A. Schmidt、K. Vaananen、T. Goyal、P. O. Kristensson、A. Peters、S. Mueller、J. R. Williamson 和 M. L. Wilson 编。美国计算机协会, 2023 年, 第 700:1–700:12 页。 [在线]。网址: https://doi.org/10.1145/3544548.3580706[9] F. Biadsy、R. J. Weiss、P. J. Moreno、D. Kanvesky 和 Y. Jia, 《Parrotron：一种端到端语音到语音转换模型及其在听障语音和语音分离中的应用》, 载于国际语音通信协会第 20 届年会（Interspeech 2019）, 格拉茨, 奥地利, 2019 年 9 月 15–19 日, G. Kubin 和 Z. Kacic 编。国际语音通信协会, 2019 年, 第 4115–4119 页。 [在线]。网址: https: // doi.org/10.21437/Interspeech.2019-1789[10] Z. Du、Y. Wang、Q. Chen、X. Shi、X. Lv、T. Zhao、Z. Gao、Y. Yang、C. Gao、H. Wang 等, 《CosyVoice 2：利用大语言模型实现可扩展的流式语音合成》, arXiv 预印本 arXiv:2412.10117, 2024 年。[11] X. Zhang、X. Zhang、K. Peng、Z. Tang、V. Manohar、Y. Liu、J. Hwang、D. Li、Y. Wang、J. Chan、Y. Huang、Z. Wu 和 M. Ma, 《Vevo：通过自监督解耦实现可控零样本语音模仿》, 载于第十三届国际学习表征会议（ICLR 2025）, 新加坡, 2025 年 4 月 24–28 日。OpenReview.net, 2025 年。 [在线]。网址: https://openreview.net/forum?id=anQDiQZhDP [12] T. Tan, H. Ruan, X. Chen, K. Chen, Z. Lin, and J. Lu, “Distillw2n：一种通过蒸馏自监督特征实现的轻量级单样本耳语到正常语音转换模型”, 载于 *2025*年美国电气与电子工程师协会声学、语音与信号处理国际会议（*ICASSP 2025*）, 印度海得拉巴, *2025*年*4*月*6*日至*11*日。IEEE, 2025, 第1–5页。 [在线]。链接: https://doi.org/10.1109/ICASSP49660.2025.10888480[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习”, *IEEE/ACM*音频、语音与语言处理汇刊, 第29卷, 第3451–3460页, 2021年。[14] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, “Emilia：面向大规模语音生成的广泛、多语言且多样化的语音数据集”, 载于 *IEEE*口语语言技术研讨会（*SLT 2024*）, 澳门, *2024*年*12*月*2*日至*5*日。IEEE, 2024, 第885–890页。 [在线]。链接: https://doi.org/10.1109/SLT61566.2024.10832365[15] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, “有限标量量化：让VQ-VAE化繁为简”, 载于 第十二届国际学习表征会议（*ICLR 2024*）, 奥地利维也纳, *2024*年*5*月*7*日至*11*日。OpenReview.net, 2024年。 [在线]。链接: https://openreview.net/forum?id=8ishA3LxN8[16] B. P. Lim, “耳语语音与非耳语语音之间的计算差异”, 博士学位论文, 伊利诺伊大学厄巴纳-香槟分校, 美国伊利诺伊州厄巴纳-香槟, 2011年5月, 电气与计算机工程博士论文; 获取地址: https://hdl.handle.net/2142/24283。[17] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “CHAINs语音语料库：刻画个体说话人”, 载于 *SPECOM*会议论文集, 2006年, 第1–6页。[18] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. F. Chen, and B. Ma, “面向语音技术应用的耳语普通话语料库”, 载于 *Interspeech 2014*会议论文集, 2014年, 第1598–1602页。[19] B. Singhal, A. R. Naini, and P. K. Ghosh, “wSPIRE：中性语音与耳语语音的并行多设备语料库”, 载于*2021*年第*24*届东方*COCOSDA*国际委员会语音数据库及评测技术协调与标准化会议（*O-COCOSDA*）。IEEE, 2021年, 第146–151页。[20] C. Li, F. Su, J. Liu, H. Bu, Y. Wan, H. Suo, and M. Li, “AISHELL6-Whisper：一个带有语音识别基线方法的中文普通话音视频耳语语音数据集”, *arXiv*预印本 *arXiv:2509.23833*, 2025年。[21] J. Yang and R. Zhou, “Whisper40：一个包含同文本中性语音的多人中文耳语说话人识别数据集”, *Information*, 第15卷, 第4期, 第184页, 2024年。[22] S. Salvador and P. Chan, “实现线性时间和空间内精确的动态时间规整”, 智能数据分析, 第11卷, 第5期, 第561–580页, 2007年。[23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “UTMOS：东京大学Sarulab的VoiceMOS挑战赛2022系统”, *arXiv*预印本 *arXiv:2204.02152*, 2022年。[24] C. K. Reddy, V. Gopal, and R. Cutler, “DNSMOS：一种用于评估噪声抑制器的非侵入式感知客观语音质量指标”, 载于*ICASSP*, 2021年。[25] G. Mittag, B. Naderi, A. Chehadi, and S. M¨oller, “NISQA：一种利用众包数据集进行多维语音质量预测的深度CNN自注意力模型”, 载于*Interspeech 2021*会议论文集, 2021年, 2127 2131pp. – 。[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “通过大规模弱监督实现稳健的语音识别”, 2022年。 [在线]。链接: https: // arxiv.org/abs/2212.04356

[27] 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, and F. Wei, “WavLM：面向全栈语音处理的大规模自监督预训练”，*IEEE*信号处理选题期刊，第16卷，第6期，第1505–1518页，2022年。 [在线]。链接：https://doi.org/10.1109/JSTSP.2022.3188113