arXiv:2304.05574v1 [eess.AS] 12 Apr 2023

通过伪目标生成和域对抗训练，从静默舌部与唇部发音重建语音

Rui-Chen Zheng、*Yang Ai**、*Zhen-Hua Ling*

中国科学技术大学语音与语言信息处理国家工程研究中心，中国合肥

zhengruichen@mail.ustc.edu.cn，{yangai, zhling}@ustc.edu.cn

摘要

本文研究在静默说话模式下，从超声舌部图像和光学唇部视频重建语音的任务。在这种模式下，人们只活动口腔内外的发音器官，而不发出声音。该任务属于发音到声学转换的范畴，也可称为静默语音接口。我们提出一种结合伪目标生成、域对抗训练和迭代训练策略的方法，以提升从静默舌部与唇部发音中恢复语音的可懂度和自然度。实验表明，与基线 TaLNet 模型相比，我们提出的方法显著提升了静默说话模式下重建语音的可懂度和自然度。使用自动语音识别（ASR）模型评估可懂度时，我们提出的方法的词错误率（WER）比基线降低了 15% 以上。此外，在发声发音模式下重建语音的可懂度方面，我们提出的方法也优于基线，词错误率降低了约 10%。

索引词— 发音到声学转换，静音语音接口，伪目标，域对抗训练

1. 引言

语音产生过程涉及一系列发声器官的协调，例如舌、下颌、软腭和嘴唇 [1]。因此，发音特征与声学特征在本质上是相互关联的 [2]。发音到声学转换是一项源于上述理论的任务，旨在直接根据发音输入合成声学特征 [3, 4]。它也可被称为无声语音接口（SSI）；该接口依赖说话人在语音产生过程中生成的非声学信号，使其能够进行交流，尽管常规的言语交流无法实现 [5–7]。SSI 具有极其重要的研究意义，因为它可以帮助发声障碍用户恢复语音交流，并在无法或不宜使用语音时协助交流。

人们在不同场景下会采用不同的说话模式。在大多数情况下，说话人采用标准的发声说话模式，这意味着他们的喉部和肺部能够正常发挥作用。在过去几年中，针对发声说话模式下的 SSI 已开展了大量研究。早期的一项研究采用基于隐马尔可夫模型（HMM）的统计模型和单元选择算法，来预测与发声舌部和唇部发音的特征相对应的声学特征 [8]。随着深度学习的快速发展，研究人员开发了深度神经网络（DNN）和卷积神经网络（CNN），基于发声舌部或唇部运动记录进行语音重建 [9–11]。此外，还有一项研究使用基于双向长短期记忆（BLSTM）的循环神经网络（RNN），将记录的电磁矢状面构音仪（EMA）测量值转换为频谱特征和激励特征，并最终恢复语音波形 [12]。最近，有研究提出了一种名为 TaLNet 的编码器-解码器架构模型，该模型结合发声舌部超声和唇部视频重建语音，并利用从文本转语音（TTS）模型中迁移学习获得了令人瞩目的效果 [13]。

在某些需要保持安静的情况下，或对于一些喉切除患者，发声者倾向于采用静默说话模式，这意味着在发音过程中，说话人仅激活其口腔和鼻腔发音器官，但抑制喉部活动，因此不会产生声音输出。以往的研究很少关注在静默说话模式下重建语音的性能。从无声发音中重建语音面临以下两个挑战。首先，大量研究已证明，发声发音与无声发音之间存在差异。静默说话模式下缺乏口内压力，可能导致说话人产生不完整或减弱的发音动作 [14]。无声发音过程中的发音动作持续时间更长 [13–15]，，并表现出峰值速度降低以及发音子动作数量增加的现象 [16]。此外，在无声模式下产生的音素表现出不太明显的舌部运动模式，其特征是发音辨别度的空间范围缩小 [16, 17]。这些发现表明，无声发音和发声发音属于两个不同的域。因此，在发声数据上训练的模型无法直接应用于无声模式。其次，静默说话模式不会产生语音信号。因此，以无声发音为输入的模型无法采用与发声模式中相同的监督范式进行训练。

针对上述从无声舌部和唇部发音进行语音重建所面临的挑战，本文提出以下方法。（1）针对静默说话模式下训练时缺少对应自然语音输出的问题，我们采用动态时间规整（DTW） [18]为未标注的无声发音数据生成伪目标。（2）由于静默说话模式下的发音状态比发声模式下更不确定，我们结合发声数据与静默数据，并引入域对抗训练，以学习在发声域和静默域中均保持不变的鲁棒表征。（3）我们设计了一种迭代训练策略，反复执行前两步，以进一步

* 通讯作者。本工作部分得到了中央高校基本科研业务费的资助。

提升模型性能。在舌部与唇部（TaL）数据集 [19] 上的实验结果表明，所提出的方法有效提高了静默说话模式下重建语音的可懂度和自然度，同时也在发声模式下取得了一定程度的提升。

2. 相关工作

我们提出的方法基于 TaLNet [13], , 这是 TaL 数据集 [19]上以发声舌部和唇部发音为输入进行语音重建的最先进模型。它采用编码器—解码器架构。TaLNet 的编码器首先将输入的舌部图像和唇部视频编码为发音表征，随后由解码器将其解码为声学特征。最后，将声学特征输入经过充分训练的神经声码器，以生成最终的语音波形。

TaLNet 的编码器包含两个并行的子编码器，分别用于处理超声舌部图像和光学唇部视频。具体而言，每个子编码器均由堆叠的 3D CNN 组成。子编码器会为每个输入帧生成一个视觉向量。最后，将分别从舌部和唇部编码得到的向量通过拼接和线性投影进行融合，从而生成发音表征。

TaLNet 的解码器由基于 Tacotron2 的[20] TTS 声学模型迁移而来，不同之处在于采用强制注意力机制取代了标准的软注意力机制。每一帧声学特征均以自回归方式预测。自回归输入首先经过预处理网络，然后送入一个双层长短期记忆（LSTM）网络。初始解码器输出之后会经过一个基于 CNN 的后处理网络，以改进声学特征预测。

为了训练 TaLNet，首先在多说话人 TTS 语料库上构建一个多说话人 Tacotron2 模型。随后，将其解码器迁移为 TaLNet 解码器，同时用强制注意力替代软注意力。接着，将迁移后的解码器与 TaLNet 的编码器联合训练。关于 TaL Net 的更多详情，建议读者参阅其原始论文 [13]。

3. 提出的方法

本文提出采用伪目标生成和域对抗训练，以应对静默说话模式下 SSI 任务面临的两大主要挑战。此外，还设计了一种迭代训练策略，以进一步提升模型性能。在我们的

在所提出的模型中，编码器和解码器的结构均与 TaLNet [13] 中的一致。详情如图 1 所示，并将在本节中介绍。

3.1. 伪目标生成

在静默说话模式下，由于声带不振动，不会产生语音信号，因此难以应用传统的监督式序列到序列范式。为解决这一问题，我们提出为无声发音生成伪声学目标，以便采用与发声模式相同的监督学习范式。为无标签目标域生成伪目标的一般方法，是直接利用模型在目标域上的预测结果。然而，在源域上训练的模型通常无法在目标域上取得良好效果，因而生成的伪目标质量较差。因此，我们提出一种新的伪目标生成方法，利用 TaL 数据集 [19]中两种模式下具有相同语言内容的平行录音。

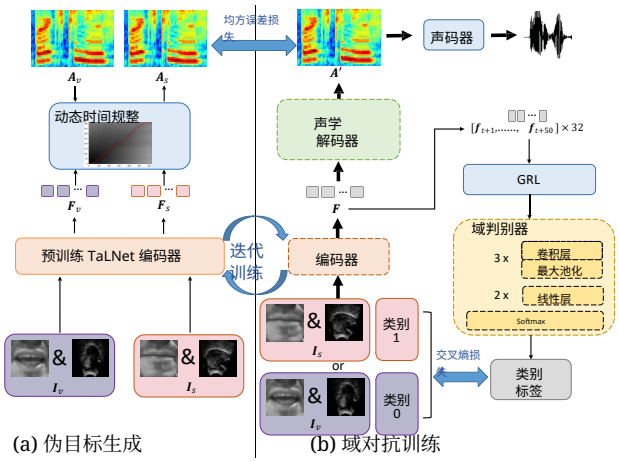


图 1。所提方法的细节。虚线框表示该部分的参数在训练过程中会更新。推理过程以粗黑箭头表示。图中出现的符号将在第 3.1 节中定义。

如图 1(a)所示，首先使用预训练 TaLNet 编码器对静默输入 $\boldsymbol{I}_s$ 及其对应的、具有相同语言内容的发声输入$\boldsymbol{I}_v$ 进行编码，得到输出 $\boldsymbol{F}_s$ 和 $\boldsymbol{F}_v$。通常，静默输入 $\boldsymbol{I}_s$ 的持续时间 T' 长于发声输入 $\boldsymbol{I}_v$ 的持续时间 T，因为无声发音过程中的发音动作通常比发声发音过程中的发音动作持续更久。随后，我们提出使用动态时间规整（DTW）[18]，将源发声域中的编码器输出 $\boldsymbol{F}_v$ 与目标静默域中的 $\boldsymbol{F}_s$ 进行对齐。由此得到用于对齐声学特征的对齐路径。之后，沿着 DTW 得到的对齐路径，对发声发音数据的自然声学特征 $\boldsymbol{A}_v$ 进行时间对齐，从而在目标静默域上生成伪声学特征 $\boldsymbol{A}_s$ 。为 $\boldsymbol{A}_s$ 静默输入获取伪声学特征后，即可遵循 TaLNet 在发声模式下的监督训练范式，在目标静默域上训练模型。

在实际应用中，动态时间规整算法有时会生成不恰当的对齐，导致部分训练样本的伪目标质量较差。为解决这一问题，在我们的实现中，仅选择具有可靠伪目标的样本进行监督训练。动态时间规整对齐的代价被用作筛选可靠样本的标准。令 D_s 表示所有无声样本的集合。然后， $D_s^\epsilon \subset D_s$ 表示动态时间规整代价低于特定阈值 ϵ的可靠无声样本集合。因此，使用伪目标进行模型训练的重构损失，即图 1 中的均方误差损失，可写为

$$L_{Recon,s} = \mathbb{E}_{D_s^\epsilon} ||\boldsymbol{A}_s - \boldsymbol{A}'_s||_2^2, \tag{1}$$

其中， $\boldsymbol{A}'_s$ 表示模型根据无声发音动作预测的声学特征。

3.2. 域对抗训练

鉴于采集发音数据存在困难，尤其是在无声发音场景中，我们的方法对无声发音和发声发音数据进行联合训练。对于发声发音数据，重构损失可写为

$$L_{Recon,v} = \mathbb{E}_{D_v} ||\boldsymbol{A}_v - \boldsymbol{A}'_v||_2^2, \tag{2}$$

where $\boldsymbol{A}'_v$ 表示由 v 预测得到的声学特征 ocalized

发音动作，而 D_v 表示所有有声训练样本的集合。此外，正如前文所述，有声发音与无声发音之间存在域差异。为解决这一问题，模型引入域对抗训练，促使编码器学习在无声域和有声域中均保持不变、更加稳健的发音动作表征。其详细结构如图 1(b) 所示。具体而言，在编码器之后接入一个域判别器，以编码器的输出$\boldsymbol{F}$ 为输入，判断其来自无声域还是有声域。由于发音动作表征 $\boldsymbol{F}$的长度各不相同，编码器的输出无法直接输入域判别器。因此，从编码器的输出中随机选取 32 个长度为 50 帧的片段，将它们拼接起来，再送入域判别器进行分类。域判别器由多层 CNN、最大池化层和线性层堆叠而成。最后，使用 Softmax 函数生成分类概率。

根据对抗训练的思想， [21], 通过最小化判别器的交叉熵（CE）损失 L_D 来训练判别器网络，以增强其判别能力；与此同时，编码器则需要最大化 L_D ，从而生成更具迷惑性的表征来欺骗判别器。因此，用于训练编码器和声学解码器的总体损失函数可写为

$$L = L_{Recon,s} + L_{Recon,v} - \lambda L_D, \tag{3}$$

其中，超参数 λ 控制域对抗训练的强度。在编码器与域判别器之间插入梯度反转层（GRL） [22] ，以实现式（3）中 L_D 的相反数。

3.3. 迭代训练策略

我们提出的方法所生成的伪目标质量，与编码器提供的发音动作表征是否恰当高度相关。在模型训练初期，使用预训练 TaLNet 编码器进行动态时间规整对齐和伪目标生成。然而，原始 TaLNet 模型仅使用发声数据训练，因此可能无法为静默数据提供恰当的发音动作表征。为解决这一问题，我们设计了一种迭代训练策略。具体而言，训练若干轮后，使用经第 3.2 节域对抗训练更新的编码器再次生成伪目标。该过程可以迭代重复，以逐步提高生成的伪目标质量。由于域对抗训练可能改变发音动作表征的尺度以及动态时间规整对齐的代价，因此也会相应更新第 3.1 节中用于筛选可靠静默训练样本的阈值 ϵ 。

4. 实验

4.1. 数据集

舌部和唇部（TaL）数据集 [19] 被用于我们的实验。该数据集包含来自81名英语母语说话人的同步音频、超声舌部图像，以及发声和无声说话模式下的唇部视频。语料库中无声说话模式下有1212条话语，每条话语都有一条具有相同语言内容的对应发声话语。由于每位说话人在无声模式下只有约15条话语，因此从每位说话人中随机选取2条话语，分别构成验证集和测试集，其余话语用于训练集。对于域对抗训练，我们没有使用全部数据集中的发声语句为 [13], ，但仅选择与静默发音数据具有相同语言内容的语句，并将其合并到训练集。我们还构建了一个发声测试集，其中包括每位说话人的一条发声语句，且这些语句与预训练 TaLNet 和我们提出的方法所使用的发声训练集不重叠。

4.2. 实验设置

我们采用先前工作中构建的 TaLNet 模型 [13] 作为实验的基线模型，该模型在 TaL 数据集中的 11487 条发声语句上进行了训练。考虑到每位说话人的静音语句数量有限，我们以说话人无关的方式构建了所提出的模型，未使用说话人相关数据进行进一步微调。为公平比较，TaLNet 模型也采用说话人无关的设置，未进行说话人相关微调。

在迭代训练的第一轮中，我们使用预训练 TaLNet 模型的编码器生成伪目标。对于域对抗训练，为了抑制训练早期阶段来自域判别器的噪声信号，我们按照以下计划将式（3）中的 λ 从 0 逐渐增大 [22]

$$\lambda = \frac{2}{1 + \exp\{-2 \cdot \frac{current\ epoch}{total\ epochs}\}} - 1. \tag{4}$$

在最初的 50 个训练轮次中，编码器参数每 5 个批次更新一次，而域判别器参数则每个批次更新一次。在剩余的训练轮次中，编码器与域判别器的更新频率互换。在我们的实现中，使用训练充分的 Parallel WaveGAN (PWG) [23]从预测的梅尔频谱图重建语音波形。

在我们的实现中，进行了三轮迭代训练。根据所有训练样本的 DTW 成本均值，我们采用启发式方法将用于筛选可靠训练样本的 DTW 成本阈值 ϵ 设为：第一轮为 40，后续两轮为 49。第一轮中，有 18 位说话人的所有无声语句的 DTW 成本均高于该阈值。我们怀疑这些说话人的无声发音可能不可靠，因此在后续实验中，将他们从所有无声模式模型的测试集中排除。

4.3. 实验结果

我们提出的方法与基线 TaLNet 模型在静默（S）和发声（V）两种说话模式下进行了比较。[1] 客观评价采用梅尔倒谱失真（MCD）、短时客观可懂度（STOI）以及语音识别引擎给出的词错误率（WER）作为指标。由于静默语句没有真实语音，因此将测试语句生成的伪声学特征输入 PWG 声码器生成波形，并在计算 MCD 和 STOI 时，将这些波形用作测试静默语句的参考语音。为进一步评估重建语音的可懂度，我们使用科大讯飞语音识别 API[2] 测量不同模型的词错误率。作为参考，测试集中所有发声自然录音的平均词错误率为 4.110%，上述测试静默语句参考语音的平均词错误率为 10.246%。我们还分别针对两种模式开展了两组主观听音测试，以评估重建语音的自然度平均意见分（MOS）。

[1]我们的演示可在此处查看：https://zhengrachel.github.io/ImprovedTaLNet-demo/.[2]https://www.xfyun.cn/services/lfasr

表 1。静默（S）和发声说话模式（V）下恢复语音的客观与主观评估结果。GT 表示发声说话模式下经声码器重新合成的自然语音。最佳结果以粗体突出显示。所有结果均为测试集上的平均值。 ± 表示 95% 置信区间。

Mode	方法	MCD/dB	STOI	词错误率/%	MOS
S	TaLNet	4.423	0.432	59.960	2.990±0.123
	Ours	**3.935**	**0.517**	**43.114**	**3.330±0.120**
V	TaLNet	3.421	0.666	26.890	3.421±0.122
	Ours	**3.382**	**0.684**	**17.309**	3.672±0.109
	GT	-	-	-	**4.161±0.096**

表 2. 消融研究中所提出方法的客观评价结果。此处，ITS 和 DAT 分别代表“迭代训练策略”和“域对抗训练”。最佳结果以粗体突出显示。

Mode	方法	MCD/分贝	STOI	词错误率/%
S	Ours	**3.935**	**0.517**	**43.114**
	不含 ITS	3.971	0.512	46.016
	不含 DAT	4.009	0.51	51.199
V	Ours	**3.382**	**0.684**	**17.309**
	不含 ITS	3.392	0.683	18.122
	不使用域对抗训练	3.434	0.670	25.032

每次测试中，我们通过亚马逊 Mechanical Turk[3] 招募了 21 名英语母语者，请他们为所听到的每条语句打分，采用 5 分制（1—很差，2—较差，3—一般，4—良好，5—优秀）。随机选取每种模式下由每个系统生成的 15 条语句进行 MOS 评估。

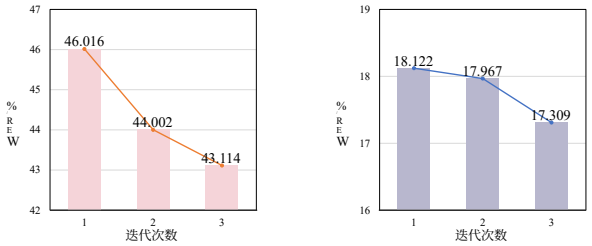
无声模式下的评估结果如表1前两行所示。可以看出，所提出的方法在所有客观和主观指标上均优于基线 TaLNet 模型，尤其是词错误率降低了15%，MOS显著提高了0.34（$p = 2.11 \times 10^{-6}$ 配对 t 检验），这表明语音的可懂度和自然度更高。

发声模式下的评估结果如表1最后三行所示。可以看出，所提出的方法显著提升了 TaLNet 利用发声发音数据恢复语音的自然度和可懂度。这可能是因为 DTW 生成的伪标签可视为一种数据增强，能够提供更多有用的训练数据。此外，域判别器使编码器专注于提取与域无关的发音表征，这也可能提升模型处理未见过的发声发音数据时的泛化能力。

4.4. 分析

为考察我们提出的方法中各个部分的有效性，
我们进行了一些消融研究。具体而言，我们比较了所提出的方法、未采用迭代训练策略的所提方法（“*w/o ITS*”），以及未采用域对抗训练的所提方法（“*w/o DAT*”）在静默和发声发音数据上的性能。对于“*w/o ITS*”，伪目标生成和域对抗模型训练过程仅进行一次。对于“*w/o DAT*”，公式（3）退化为$L_{Recon,\ s}$且不使用域判别器。如表2所示，在静默和发声说话模式下，“*w/oITS*”和“*w/oDAT*”在所有客观指标上的表现均不如我们提出的方法，这

[3]https://www.mturk.com/



(a) 静默说话模式　　(b) 发声说话模式

图 2。我们提出的方法在不同训练迭代次数下的测试集词错误率。

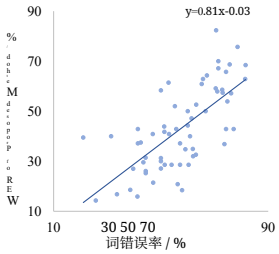


图 3。采用 TaLNet 和我们提出的方法恢复语音后，对词错误率进行拟合得到的 LSE 结果。系数大于 0 表明两种方法之间呈正相关。

证明了我们提出的迭代训练策略和域对抗训练的有效性。与表1中TaLNet的结果相比，可以看出，两种消融模型在所有客观指标上的表现仍优于TaLNet。此外，图2还展示了所提出方法在不同训练迭代次数下的测试集词错误率（WER），以说明迭代模型训练的有效性。

为研究不同说话人之间的性能差异，我们采用最小二乘估计（LSE）对使用 TaLNet 和我们提出的方法还原语音所得的逐说话人词错误率进行拟合。结果如图 3 所示，呈现出显著的正相关关系：也就是说，如果某位说话人在静默说话模式下使用 TaLNet 重构语音时词错误率较低或较高，那么使用我们提出的方法时，其词错误率也往往较低或较高。与 TaLNet [13]，的原论文结论相同，我们也观察到不同说话人的重构语音词错误率存在显著差异。原因可能是，部分说话人在缺少音频反馈时发音不准确，因此难以通过我们提出的方法提高其重构语音的可懂度。

5. 结论

本文提出采用伪目标生成和域对抗训练，以应对静默发音语音重建中的两大主要挑战。此外，还设计了一种迭代训练策略，以进一步提升性能。客观和主观实验结果均证实了所提方法的有效性。然而，静默模式与发声模式下发音到声学转换的性能之间仍存在明显差距。未来的工作将通过考虑两种模式之间固有的发音差异，改进模型框架。

6. 参考文献

[1] Peter B Denes, Peter Denes, and Elliot Pinson, The speech chain, Macmillan, 1993.[2]Zheng-Chen Liu, Zhen-Hua Ling, and Li-Rong Dai, “使用神经网络级联预测频谱和激励特征的发音到声学转换。,” in Proc. Interspeech, 2016, pp. 1502–1506.[3] Christopher T Kello and David C Plaut, “在发音参数录音上训练的发音—声学正向映射神经网络模型,” The Journal of the Acoustical Society of America, vol. 116, 2364, 2004 4, 2354no. pp. – .[4] Tam´as G´abor Csap´o, Csaba Zaink´o, L´aszl´o T´oth, G´abor Gosz- t olya, and Alexandra Mark´o, “基于超声的发音到声学映射与 WaveGlow 语音合成 ,” 2020, pp. 2727–2731.[5] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg, “静默语音接口,” Speech Communication, vol. 52, 287, 2010 270pp. – .[6] Tanja Schultz, Michael Wand, Thomas Hueber, Dean J Krusienski, Christian Herff, and Jonathan S Brumberg, “基于生物信号的口语交流：综述,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2257–2271, 2017.[7]Jose A Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M Mart´ın Do˜nas, Jos´e L P´erez-C´ordoba, and Angel M Gomez, “用于语音恢复的静默语音接口：综述,” IEEE access, vol. 8, pp. 177995–178021, 2020.[8]Thomas Hueber, Elie-Laurent Benaroya, G´erard Chollet, Bruce Denby, G´erard Dreyfus, and Maureen Stone, “由舌部和唇部超声及光学图像驱动的静默语音接口的开发,” Speech Communication, vol. 52, pp. 288–300, 2010.[9] Tam´as G´abor Csap´o, Tam´as Gr´osz, G´abor Gosztolya, L´aszl´o T´oth, and Alexandra Mark´o, “基于深度神经网络的超声到语音转换，用于静默语音接口,” i n Proc. Interspeech, 2017, pp. 3672–3676.[10] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani, “Lip2audspec：从静默唇部动作视频重建语音,” in Proc. ICASSP, 2018, pp. 2516–2520.[11] Yaman Ku mar, Rohit Jain, Khwaja Mohd Salik, Rajiv Ratn Shah, Yifang Yin, and Roger Zimmermann, “Lipper：利用多视角唇读合成语音,” in Proc. AAAI, 2019, vol. 33, pp. 2588–2595.[12]Zheng-Chen Liu, Zhen-Hua Ling, and Li-Rong Dai, “使用增强输入表征的 BLSTM-RNN 发音到声学转换,” Speech Communica- tion, vol. 99, pp. 161–172, 2018.[13] Jing-Xuan Zhang, Korin Richmond, Zhen-Hua Ling, and Lirong Dai, “TaLNet：通过从文本转语音合成迁移学习，实现从舌部和唇部发音重建语音,” in Proc. AAAI, 2021, pp. 14402–14410.[14] Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, and Steve Renals, “基于超声和视频的静默与常规多说话人语音识别,” in Proc. Interspeech, 2021, pp. 641–645.

[15] Christopher Dromey 和 Katherine M Black，“喉部活动对发音的影响”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 25 卷，第 2272–2280 页，2017 年。[16] Kristin J Teplansky、Brian Y Tsang 和 Jun Wang，“有声、耳语和无声元音产生中的舌部与唇部运动模式”，载于国际语音科学大会论文集，2019 年，第 1–5 页。[17] Kristin J Teplansky、Alan Wisler、Beiming Cao、Wendy Liang、Chad W Whited、Ted Mau 和 Jun Wang，“无喉语音中的舌部与唇部运动模式”，载于*Interspeech* 论文集，2020 年，第 4576–4580 页。[18] Meinard M¨uller，“动态时间规整”，音乐与运动的信息检索，第 69–84 页，2007 年。[19] Manuel Sam Ribeiro、Jennifer Sanger、Jing-Xuan Zhang、Aciel Eshky、Alan Wrench、Korin Richmond 和 Steve Renals，“TaL：一个同步的多说话人超声波舌部成像、音频和唇部视频语料库”，载于*SLT* 论文集，2021 年，第 1109–1116 页。[20] Jonathan Shen、Ruoming Pang、Ron J Weiss、Mike Schuster、Navdeep Jaitly、Zongheng Yang、Zhifeng Chen、Yu Zhang、Yuxuan Wang、Rj Skerrv-Ryan 等，“通过以梅尔频谱预测为条件的 WaveNet 实现自然语音 TTS 合成”，载于 *ICASSP* 论文集，2018 年，第 4779–4783 页。[21] Antonia Creswell、Tom White、Vincent Dumoulin、Kai Arulkumaran、Biswa Sengupta 和 Anil A Bharath，“生成对抗网络：综述”，*IEEE* 信号处理杂志，第 35 卷，第 1 期，第 53–65 页，2018 年。[22] Yaroslav Ganin 和 Victor Lempitsky，“通过反向传播实现无监督域适应”，载于*ICML* 论文集，2015 年，第 1180–1189 页。[23]Ryuichi Yamamoto、Eunwoo Song 和 Jae-Min Kim，“Parallel WaveGAN：一种基于生成对抗网络和多分辨率频谱的快速波形生成模型”，载于*ICASSP* 论文集，2020 年，第 6199–6203 页。