arXiv:2303.00502v1 [cs.SD] 1 Mar 2023

关于唇到语音合成的视听同步

Zhe Niu 和*Brian Mak*香港科技大学计算机科学与工程系{zniu,mak}@cse.ust.hk

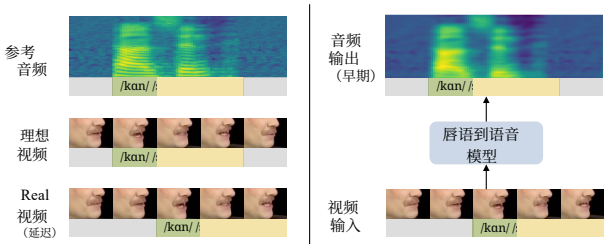
摘要

大多数唇语到语音（*LTS*）合成模型都基于这样的假设进行训练和评估：数据集中的视听配对数据完全同步。在本研究中，我们发现，常用的视听数据集（如 *GRID*、*TCD-TIMIT* 和 *Lip2Wav*）可能存在数据不同步问题。使用此类数据集进行唇到语音训练，可能进一步导致模型不同步问题，即生成语音与输入视频不同步。为解决这些异步问题，我们提出了一种带有自动同步机制（*ASM*）的同步唇语转语音（*SLTS*）模型，用于纠正数据不同步并惩罚模型不同步。我们还展示了在异步测试数据下，*LTS* 常用评估指标的局限性，并在对时间对齐敏感的指标之前引入音频对齐前端，以实现更好的评估。我们在传统指标和时间对齐指标上，将我们的方法与当前最先进的方法进行比较，以展示同步训练的优势。

1. 引言

唇语到语音（LTS）任务是根据无声视频中的唇部动作重建说话人语音。随着深度学习的发展，人们提出了许多数据驱动的深度网络模型来解决 LTS 任务。

在训练 LTS 模型时，通常会作出一个假设：对应视频和音频数据之间的时间偏移是一个较小的常数，或者为零。换句话说，同一段语音的视频和音频在时间上基本同步。然而，在使用唇同步模型 SyncNet 分析同步误差后 [4], 我们发现，常用于训练和评估 LTS 模型的音视频数据集中，音频与视频之间存在不同的时间偏移。有些数据集（例如 GRID） [5] 和 TCD-TIMIT [8] 的偏移量小于±1 一帧视频，但其他数据集（例如 Lip2Wav） [21], 可能



(a) 数据异步问题。 (b) 模型异步问题。

图1. 音视频不同步问题示意图。说话人正在说*constant*这个词。

存在数帧视频大小的较大偏移量。此外，不当的数据预处理也可能引入较大的时间偏移（例如使用 FFmpeg [26] 将视频文件分割成较小片段[1]）。我们将此称为数据不同步问题（见图1a），因为同步误差源于外部数据集，而非 LTS 模型本身。

尽管同步误差大多数时候肉眼几乎不可见，但它们会对 LTS 模型的优化产生不可忽视的影响。LTS 模型的训练通常采用时间维度学习目标（例如，音频梅尔频谱图之间的均方误差），而这类目标对时间偏移十分敏感。数据集中视频与音频之间的错位可能会误导模型，使其产生不同步的输出，从而导致模型异步问题 （见图1b）。此外，非恒定的时间偏移会导致训练不稳定性，使模型难以在大规模数据集上收敛。

在评估阶段，测试数据集中的视听异步也会使客观评估变得困难。常用的客观语音可懂度指标，如短时客观可懂度 [25] 和扩展短时客观可懂度 [12], , 要求参考音频与测试音频在时间上完全对齐，才能得出准确反映听力测试结果的分数。当模型异步和数据不同步同时

[1]Lip2Wav数据集的预处理流程中就存在这种分段：https://github.com/Rudrabha/Lip2Wav/blob/master/download_speaker.sh。

存在时，如果未妥善处理，同一段语音的音频与视频之间未对齐就可能导致评估结果不准确。

在本工作中，我们旨在解决训练和评估阶段中的异步问题。针对训练，我们引入了同步唇到语音（SLTS）架构，该架构包含一种自动同步机制（ASM），以确保训练阶段的模型与数据同步。此外，我们在评估期间为常用指标提出了一种侵入式时间对齐前端。所提出的前端将同步误差与传统评估解耦，即使测试集存在数据不同步，也能确保评分可靠。

在实验部分，我们在常用音视频数据集上进行了大量实验，以展示所提出自动同步机制的有效性。结果表明，所提出的同步方法既能处理人眼可见的长期异步（例如，超过一个视频帧的偏移量），也能处理细微的同步误差（例如，单帧或子帧偏移量）。SLTS还在多项客观指标上优于现有最先进水平模型，并在主观听测中取得了较高分数。

2. 相关工作

2.1. 唇语到语音模型中的同步

唇语到语音模型通常由能够提供较大时间感受野的组件构成，例如 3D 卷积堆叠 [21], 长短期记忆网络或门控循环单元 [1, 19, 21, 29], 位置敏感注意力 [10,21], 以及自注意力层 [14,28]。较大的感受野可能使模型生成存在偏移量的音频。Kim 等人 [14]指出，一些现有的 LTS 模型并未显式处理局部视觉特征，因此可能根据输入视频生成不同步的语音。他们提出在训练过程中加入额外的同步损失，以解决模型不同步问题。然而，他们的工作只考虑了模型不同步，而没有考虑数据不同步。我们的工作同时考虑这两种不同步问题，并提出相应的解决方案。

2.2. 唇音同步模型

唇音同步任务旨在预测视听偏移，以纠正唇音同步错误。现有研究（如 [4, 15],）假设音视频训练数据已同步，并设计不同的负样本对，通过对比损失训练模型。Chung等人 [4] 通过随机移动音频生成负（不同步）视听配对数据，并采用孪生网络 [3]中的对比损失来训练其网络。Kim 等人 [15] 则采用基于 softmax 的对比损失，并将不同时间步的视听特征视为负样本对。我们提出的数据同步模块（DSM）也可用于唇音同步。

与现有唇同步模型相比，DSM 不假设训练数据已同步。它处理一组候选配对，并在唇到语音学习目标（例如，梅尔频谱图之间的均方误差损失）的驱动下，以无监督方式发现正样本对和负样本对。

2.3. 端到端唇语到语音模型

唇语到语音模型通常并非为端到端生成波形而设计，因为人们通常希望采用更紧凑的声学表示（例如梅尔频谱图）来降低任务难度。随后，声码器会将这些紧凑的声学表示转换为音频波形。声码器可以是基于算法的，例如 [14,21,29], 中使用的 Griffin-Lim；也可以是单独训练的神经声码器，如 [10,13,18] 所示。近年来，直接生成音频波形的端到端 LTS 模型 [19,28] 越来越受到关注，因为与基于算法的声码器相比，这类模型生成的语音质量更高，且无需单独训练神经声码器。在这项工作中，我们还通过将 UnivNet 声码器 [11]与所提出的模型联合训练，研究了端到端建模。

3. 同步唇语到语音合成

我们首先在第3.1节中阐述数据和模型不同步问题，然后在第3.2节中介绍所提出的同步唇语转语音（SLTS）模型的整体架构。我们将在第3.3节中详细介绍我们的关键贡献——自动同步机制（ASM）及其两个组成部分：数据同步模块（DSM）和自同步模块（SSM）。

3.1. 问题表述

为简化起见，我们首先将无声唇部视频$x(t) \in \mathbb{R}^{H\times W\times 3}$ 和对应的音频 $y(t) \in \mathbb{R}$ 视为时间 $t \in \mathbb{R}$ 的连续函数。唇语转语音任务面临两类异步问题：数据异步问题和模型不同步问题。

理想情况下，唇部视频 $x(t)$ 应与音频 $y(t)$ 同步，即时间偏移为零。然而，在真实世界数据集中，唇部视频相对于音频可能存在大小不固定的 o_d 秒时间偏移（当 $o_d > 0$时，视频滞后于音频），从而导致视频不同步：$x^{o_d}(t) = x(t - o_d)$。我们将这种 o_d秒的数据异步问题称为图1a所示的问题。

另一方面，当用于训练模型的数据不同步时，LTS 模型会利用时间上下文，因此可能会在重构音频中引入自身的时间偏移量。给定视频 $x(t)$，LTS 模型可能会重构出音频 $\hat{y}^{o_m}(t) = \hat{y}(t - o_m)$，其与理想同步重构结果相差 o_m (秒)，

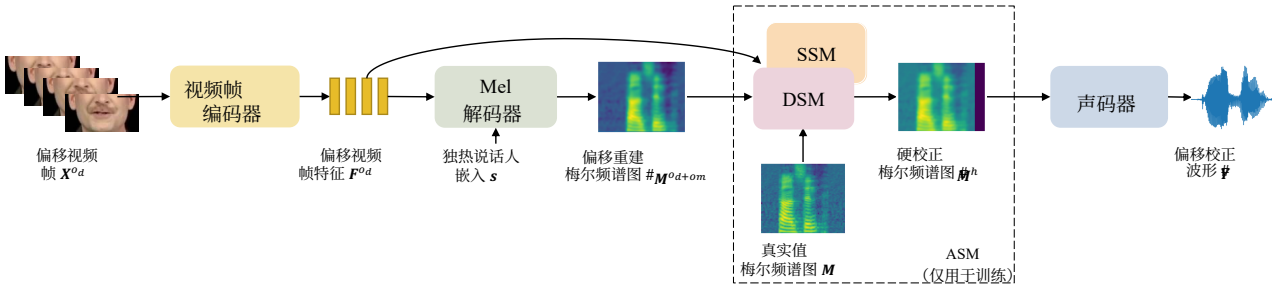


图 2. 所提出的 SLTS 架构概述。其中， o_d 由数据不同步导致， o_m 由模型不同步导致；两者均以秒为单位测量。这两种不同步问题分别由 DSM 和 SSM 处理。

时间 $\hat{y}(t)$。我们将其称为 o_m 秒模型不同步，如图1b所示。

通常，如果不存在数据不同步，也可能不会出现模型不同步，因为同步重建的音频应当优于其他不同步方案。另一方面，数据不同步会导致模型不同步，尤其是在不同样本之间的音视频偏移量各不相同的情况下。

3.2. 架构概述

在深入探讨不同步问题的解决方案之前，我们先介绍 LTS 模型的架构，如图 2 所示。实际应用中，视频和音频数据都是时间上的离散信号。静音 RGB 视频数据表示为 $\boldsymbol{X} \in \mathbb{R}^{T_v \times H \times W \times 3}$，其中 T_v、H、W分别表示视频帧数、帧高和帧宽，而 3 表示颜色通道数。包含 T_a 个采样点的单通道音频表示为$\boldsymbol{Y} \in \mathbb{R}^{T_a \times 1}$。在我们的工作中，视频数据的频率根据数据集不同为 25 或 30 Hz，音频频率则固定为 16kHz。

所提出的同步唇语转语音（SLTS）模型旨在从给定的无声视频 $\boldsymbol{X}^{o_d}$ 中重建偏移校正音频 $\boldsymbol{Y}$。该视频在训练期间存在 o_d 秒的偏移，以匹配学习目标（真实值）音频 $\boldsymbol{Y}$。在推理过程中，模型可以选择使用 ASM 生成与偏移视频 $\boldsymbol{X}^{o_d}$ 对齐的音频 $\boldsymbol{Y}^{o_d}$，也可以生成与参考音频 $\boldsymbol{Y}$ 对齐的经 ASM 校正的音频 $\boldsymbol{Y}$。后者主要用于需要音频对齐的评估。

SLTS 由视频帧编码器、解码器、两个同步模块（即 DSM 和 SSM）以及声码器组成。帧编码器基于 ResNet18 [9]，为 D_f每个单独的视频帧生成维特征 $\boldsymbol{F} \in \mathbb{R}^{T_v \times D_f}$。解码器由 Conformer [7]和基于一维卷积的后处理网络组成。25 Hz 的帧特征 $\boldsymbol{F}$ 首先与说话人嵌入进行拼接，然后输入 Conformer，以基于局部和全局上下文生成紧凑的声学表征。随后，紧凑的声学表征经过线性上采样至 100 Hz，并输入后处理网络，以生成 100 $\hat{}$ Hz 的梅尔频谱图 $\boldsymbol{M}$。

解码器之后是 ASM，它由两个模块组成：DSM 和 SSM，这也是本文的主要贡献。这两个模块分别在训练过程中学习并校正数据异步和模型不同步。DSM 以 25 Hz 的视频帧特征 $\boldsymbol{F}$、真实梅尔频谱图 $\boldsymbol{M}$ 和重建梅尔频谱图 $\boldsymbol{M}$ 作为输入，估计时间偏移，以校正音视频数据中的异步。另一方面，SSM 基于视频帧特征 $\boldsymbol{F}$ 和重建梅尔频谱图 $\boldsymbol{M}$ 生成自同步损失，以惩罚模型不同步。

最后，采用基于 UnivNet 的 [11]声码器生成音频波形。由于声码器通常使用时长短于 1.0 秒的音频片段进行训练，因此我们对经过偏移量校正的梅尔频谱图与参考音频波形进行 0.6 秒的随机分段。声码器与整个系统一同训练，使用多分辨率 STFT 损失和可微分的 STOI 损失。我们还提供了采用多分辨率频谱图判别器（MRSD）和多周期波形判别器（MPWD）的选项，以进一步提升主观语音质量，但客观评估得分会有所降低。

3.3. 自动同步机制

自动同步机制由两个组件构成：DSM 和 SSM，二者都依赖时间偏移预测器。我们先介绍时间偏移预测器，再介绍 DSM 和 SSM 的具体细节。

3.3.1 视听时间偏移预测器

时间偏移预测器（如图 3 所示）会为音视频偏移生成分类分布。时间偏移分类分布中的数值范围可以手动设置，通常为 150–300 毫秒。不同于 SyncNet [4]预测以视频帧数量表示的同步误差（在我们的工作中为 25 Hz），本文提出的偏移预测器

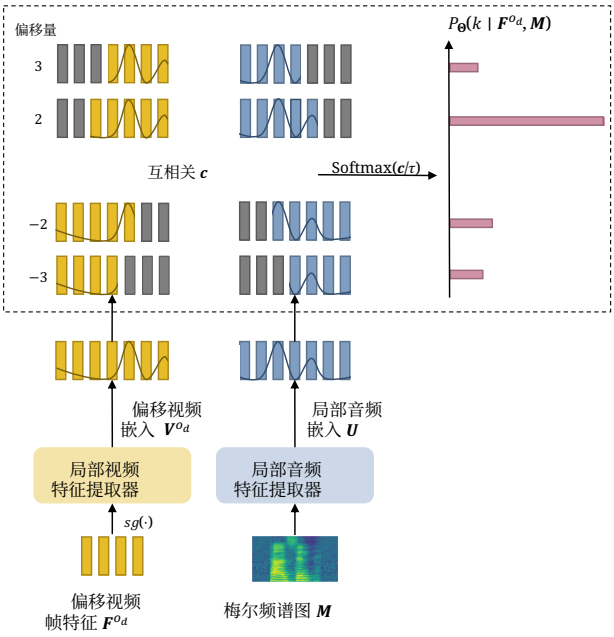


图 3. 音视频偏移预测器。

则预测以梅尔频谱帧数量表示的偏移量（在我们的工作中为 100 Hz），以实现更精确的同步。

偏移预测器分别包含用于视频和音频的两个局部特征提取器。每个特征提取器包含两个 Conv1D-BN-GELU 块、一个全连接层和一个 L2 归一化操作，用于生成归一化的局部嵌入。只有第一个 Conv1D 的卷积核大小为 3，另一个的卷积核大小为 1。我们有意限制感受野，以保留嵌入的时间精度，同时适度利用时间上下文来提高特征可区分性。视频特征提取器还包含一个额外的重采样操作，在第一个 Conv1D 之前将输入的视频特征从 25 Hz 线性上采样至 100 Hz，以匹配梅尔频谱图的采样率。

获得局部视频嵌入序列 $\boldsymbol{V}^{o_d} = (\boldsymbol{v}_0^{o_d}、\ .....、\ \boldsymbol{v}_{T_m-1}^{o_d})$ 以及局部音频嵌入序列 $\boldsymbol{U} = (\boldsymbol{u}_0、\ .....、\ \boldsymbol{u}_{T_m-1})$ 后，将在同步半径 $K \in \mathbb{N}^+$（一个超参数）范围内的样本之间计算这两个嵌入序列的互相关 $\boldsymbol{c} = (c_{-K}、\ .....、\ c_K)$：

$$c_k = \sum_{i=\max(k,0)}^{\min(k,0)+T_m-1} \langle \boldsymbol{v}_i^{o_d}, \boldsymbol{u}_{i-k} \rangle. \tag{1}$$

随后，使用经过人工调节温度参数 τ 的 softmax 函数对互相关进行归一化，以生成偏移分布：

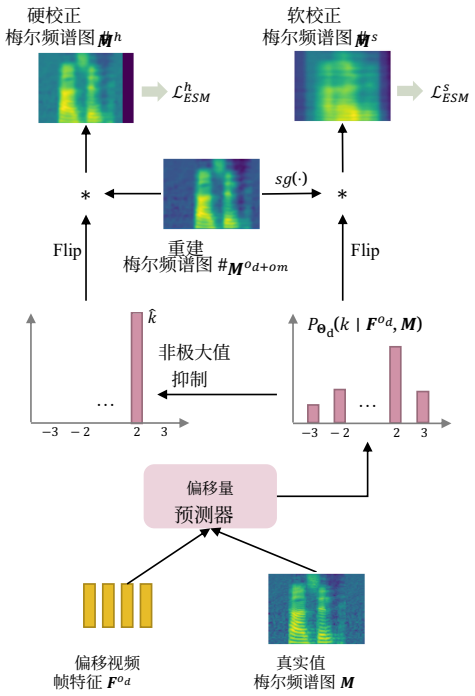


图 4. 数据同步模块。

$$P_\Theta(k|\boldsymbol{F}^{o_d}, \boldsymbol{M}) = \frac{\exp(c_k/\tau)}{\sum_{i=-K}^{K} \exp(c_i/\tau)}, \tag{2}$$

其中，Θ 是偏移预测器的参数（即，局部提取器的参数）。

3.3.2 数据同步模块

如图 4所示，数据同步模块（DSM）由一个偏移预测器组成，该预测器首先基于视频特征 $P_{\Theta_D}(k \mid \boldsymbol{F}^{o_d}, \boldsymbol{M})$ 和真实梅尔频谱图 $\boldsymbol{F}^{o_d}$，并使用一组 DSM 模型参数 $\boldsymbol{M}$，生成视听偏移的分类分布 Θ_D。

生成的偏移分布沿时间轴翻转，以获得校正卷积核，该卷积核用于校正偏移梅尔频谱图。通过将重建梅尔频谱图 $\hat{\boldsymbol{M}}^s = (\hat{\boldsymbol{m}}_0^s, \ ....., \ \hat{\boldsymbol{m}}_{T_m-1}^s)$ 与该卷积核进行卷积，可得到软校正后的梅尔频谱图 $\hat{\boldsymbol{M}}^{o_d+o_m} = (\hat{\boldsymbol{m}}_0^{o_d+o_m}, \ ....., \ \hat{\boldsymbol{m}}_{T_m-1}^{o_d+o_m})$：

$$\hat{\boldsymbol{m}}_i^s = \sum_{k=\max(-K,i-T+1)}^{\min(K,i)} P_{\Theta_e}(-k \mid \boldsymbol{F}^{o_d}, \boldsymbol{M}) sg(\hat{\boldsymbol{m}}_{i-k}^{o_d+o_m}), \tag{3}$$

其中 $sg(\cdot)$ 是梯度停止操作。然后，在真实梅尔频谱与软校正梅尔频谱图之间计算软 DSM 损失：

图与软校正梅尔频谱图之间：

$$\mathcal{L}^{s}_{DSM}(\boldsymbol{M}, \hat{\boldsymbol{M}}^{s}) := \|\boldsymbol{M} - \hat{\boldsymbol{M}}^{s}\|_2^2. \tag{4}$$

生成软校正梅尔频谱图时，对重建梅尔频谱图执行梯度停止操作至关重要。软校正梅尔频谱图由大量偏移提议组合而成，其中可能包含一些错误的提议。这些错误的提议可能导致错误的梯度反向传播至解码器，迫使其同时学习多个错误目标，从而造成收敛问题。

为了确保解码器仅从最可能的偏移提议中学习，除软校正梅尔频谱图外，还通过使用另一个校正核对重建梅尔频谱图进行卷积，计算出硬校正梅尔频谱图 $\hat{\boldsymbol{M}}^{h} = (\hat{\boldsymbol{m}}^{h}_{0}, \ldots\ldots \hat{\boldsymbol{m}}^{h}_{T-1})$；该校正核会抑制可能性较低的偏移量，结果如下：

$$\hat{\boldsymbol{m}}^{h}_{i} = \begin{cases} \hat{\boldsymbol{m}}^{o_d+o_m}_{i-\hat{k}}, & i \geq \hat{k} \\ \mathbf{0}, & i < \hat{k} \end{cases}, \tag{5}$$

其中 $k = \arg\max P(k \mid \boldsymbol{F}^{o_d}, \boldsymbol{M})$，越界的帧 $i < k$ 设为零，并在损失计算中排除。与软 DSM 损失类似，我们对硬校正梅尔频谱图采用均方误差损失：

$$\mathcal{L}^{h}_{DSM}(\boldsymbol{M}, \hat{\boldsymbol{M}}^{h}) := \|\boldsymbol{M} - \hat{\boldsymbol{M}}^{h}\|_2^2. \tag{6}$$

理想情况下，DSM 训练至收敛后，在重建梅尔频谱图上平移 $-k$ 帧即可校正 o_d秒的数据不同步。

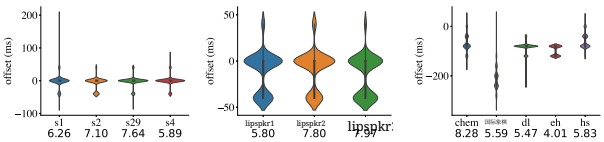
3.3.3 自同步模块

除了 o_d秒的数据不同步外，由于音频解码器的感受野较大，还可能存在 o_m秒的模型不同步。我们引入了一个自同步模块（SSM），用于尽可能减少模型不同步。

与 DSM 类似，SSM 也包含一个独立的偏移预测器，其参数由 Θ_S 确定，用于生成偏移分布 $P_{\Theta_S}(k \mid \boldsymbol{F}^{o_d}, \hat{\boldsymbol{M}}^{o_d+o_m})$。与 DSM 不同，SSM 旨在通过最小化以下 SSM 损失，减小视频特征$\boldsymbol{F}^{o_d}$与重建梅尔频谱图 $\hat{\boldsymbol{M}}^{o_d+o_m}$ 之间的偏移量：

$$\mathcal{L}_{SSM}(\boldsymbol{F}^{o_d}, \hat{\boldsymbol{M}}^{o_d+o_m}) := -\log P_{\Theta_S}(k = 0 \mid \boldsymbol{F}^{o_d}, \hat{\boldsymbol{M}}^{o_d+o_m}). \tag{7}$$

从经验上看，我们发现 SSM 提高了 DSM 训练的稳定性。没有 SSM 时，DSM 有时无法学到有效的偏移量，预测的偏移量会坍缩为一个常数。我们推测，这可能是由于噪声梯度从原本不受约束的软 DSM 损失中传播所致。



(a) GRID-4S。 (b) TCD-TIMIT-LS。 (c) Lip2Wav。

图 5。SyncNet 生成的偏移量。说话人名称下方的数字是 SyncNet 生成的置信度分数。分数越高，表示 SyncNet 对其输出的置信度越高。仅统计置信度大于 3．0 的偏移量。

数据集	说话人	训练 / 验证 / 测试样本	训练 / 验证 / 测试时长
GRID-4S [5]	4	3,600 / 200 / 200	2.98 / 0.17 / 0.17
TCD-TIMIT-LS [8]	3	1,017 / 57 / 57	1.64 / 0.09 / 0.09
Lip2Wav [21]	5	15,894 / 376 / 487	115.16 / 2.48 / 3.37

表 1. 我们实验中使用的数据集划分统计。同一数据集中的所有说话人均出现在训练、验证和测试划分中。

4. 数据集、指标与训练细节

4.1. 数据集概述

GRID-4S 是 GRID 视听语料库的一个四说话人子集 [5]。该子集包含两名男性说话者（*s1*、*s2*）和两名女性说话者（*s4*、*s29*），常用于文献中 [14,21] 评估唇语到语音模型。该语料库是在实验室条件下录制的，词汇量较小，并采用人工语法。

TCD-TIMIT-LS [8]是另一个在实验室条件下采集的视听语料库，使用真实的英语句子，词汇量更大。原始的 TCD-TIMIT 数据集由三位专业训练的唇语说话者和 59 位正常说话的志愿者录制。遵循相关文献 [14,21],，我们仅采用三位专业训练的唇语说话者的数据。

Lip2Wav [21] 是一个从 YouTube 讲座视频中收集的大规模视听数据集。该数据集包含五位不同的说话人，他们都用于我们的实验。

4.1.1 数据准备

对于 GRID-4S 和 TCD-TIMIT-LS 数据集，我们遵循惯例 [17,21,27]，从每位说话人的样本中随机抽取 90% 用于训练，5% 用于验证，另取 5% 用于测试。对于 Lip2 Wav，我们采用官方的数据划分[2]。我们采用 S^{3}FD [30] 人脸检测器，获取这三个数据集视频中的面部区域。在进行人脸检测之前，Lip2Wav 数据集中的长视频会被切分为时长最长为 30 秒的片段，遵

[2]Lip2Wav 官方数据划分：https://github.com/Rudrabha/Lip2Wav/tree/master/Dataset。

数据集	模型	短时客观可懂度 ↑	扩展短时客观可懂度 ↑	感知语音质量评估 ↑	梅尔倒谱失真 ↓	a-短时客观可懂度 ↑	a-扩展短时客观可懂度 ↑	a-感知语音质量评估 ↑	a-梅尔倒谱失真 ↓	偏移量-R^2 (F) ↑	偏移量-R^2 (SN) ↑	w-词错误率（%）↓	k-词错误率（%）↓
GRID-4S	VCA-GAN	0.688	0.500	1.917	29.720	0.732	0.552	1.910	28.437	-0.002†	-0.030†	23.67	8.50
	不含 ASM 的 SLTS	0.698	0.519	1.906	27.438	0.753	0.582	1.903	25.684	-0.001†	-0.030†	15.33	4.92
	SLTS	**0.703**	**0.525**	**1.932**	**27.327**	**0.761**	**0.592**	**1.933**	**25.404**	**0.862**	**0.383**	**12.83**	**2.92**
TCD-TIMIT-LS	VCA-GAN	0.577	0.398	1.373	33.450	0.593	0.412	1.376	33.175	-0.038†	-0.164†	79.96	-
	不含 ASM 的 SLTS	**0.622**	**0.460**	**1.480**	**30.334**	0.650	0.496	**1.482**	29.667	-0.585 †	-0.164†	50.40	-
	SLTS	0.606	0.445	**1.480**	30.818	**0.664**	**0.511**	1.480	**29.430**	**0.796**	**0.525**	**38.06**	-
Lip2Wav 化学	VCA-GAN	0.543	0.364	1.363	37.827	0.659	0.477	1.365	34.600	-0.000†	-2.725†	48.20	-
	不含 ASM 的 SLTS	**0.603**	**0.445**	1.478	**34.104**	0.736	0.578	1.481	30.291	-0.005 †	-2.725†	33.03	-
	SLTS	0.215	0.049	**1.520**	49.481	**0.760**	**0.616**	**1.515**	**29.130**	**0.982**	**0.704**	**24.69**	-

表 2. VCA-GAN、 [14], 训练期间不含 ASM 的 SLTS，以及 SLTS 之间的比较。 a-：采用时间对齐前端的指标。†：使用虚拟偏移预测器计算的结果（即，始终预测 0）。默认情况下，用于计算词错误率的参考文本来自数据集；Lip2Wav 除外，其参考文本通过对参考语音使用 Whisper 自动语音识别获得。

循官方的 Lip2Wav 预处理流程。预处理后的数据集统计信息见表1。

4.1.2 异步性分析

为研究数据集中的异步现象，我们使用预训练的 SyncNet[3] [4] 来估计三个视听数据集中的异步程度。SyncNet 以 25 帧/秒视频和 16 千赫音频作为输入，并以 40 毫秒的分辨率生成音视频偏移量。SyncNet 结果的统计数据如图 5所示。除少数离群值外，GRID-4S 和 TCD-TIMIT- LS 数据样本的音视频偏移量集中在 0 毫秒附近，部分样本存在约一个视频帧的轻微不同步（即，± 40 毫秒）。在 Lip2Wav数据集中，国际象棋 说话人的偏移量集中在 -200∼- 250 毫秒附近，而其他说话人的偏移量集中在 -80 毫秒附近。Lip2Wav 数据的音频滞后主要由数据预处理过程中的视频分割造成，国际象棋 说话人的数据除外，因为其原始视频通常超前。

4.2. 评估指标概述

感知语音质量评估 [23]：评估生成语音相较于干净参考语音的感知质量。我们遵循 [14, 21] 报告感知语音质量评估的窄带 MOS-LQO 分数。

短时客观可懂度 [25]**和扩展短时客观可懂度** [12]：根据生成语音与干净语音短时包络之间的相关性，预测可懂度听力测试的结果。这两种指标都假定音频在时间上对齐。

梅尔倒谱失真：是另一种对齐敏感指标，用于衡量从生成音频和参考音频中提取的两组梅尔倒谱序列之间的差异。

词错误率：统计生成音频转写中的词错误数。我们使用 Whisper [22]的中型版本获取转写，并将据此计算出的词错误率记作w-词错误率。GRID-4S 语句由具有受限词汇量的人工语法生成。然而，Whisper 是一种通用的大词汇量语音识别器。它会在 GRID-4S 上产生许多同音词（例如，“red” → “read”，“blue” → “blew”），这些都会被计为错误，从而导致评估不准确。因此，我们使用 GRID-4S 训练集训练了一个专用的 Kaldi 自动语音识别模型 [20]，用于识别生成的 GRID-4S 音频，并将由此得到的词错误率记作k-词错误率。

偏移量-R^2：表示 DSM 产生的偏移量与另一种方法（例如度量前端，见第4.3节）产生的偏移量之间的决定系数。如果另一种方法分别为度量前端和 SyncNet，则记作偏移量-R^2(F) 和偏移量-R^2(SN)。

4.3. 时间对齐度量前端

当两个输入音频未进行时间对齐时，对齐敏感指标（如 STOI、ESTOI 和 MCD）可能会得出不准确的分数（详情见第 5.2 节）。为解决这一问题，我们提出了一种时间对齐前端：首先，分别对生成音频和参考音频计算梅尔频谱图，窗口大小为 640，帧移为 160；随后沿通道维度对其进行归一化。接着，通过将生成音频从 -300 毫秒偏移至300 毫秒，每隔 10 毫秒生成一个偏移量，共得到 61 个对齐候选。选取使两个归一化梅尔频谱图之间的均方误差最小的偏移量，并在评分前据此校正生成音频。最优偏移量的相反数称为前端偏移量，记作 o_f。

4.4. 实现与训练细节

我们通过随机分块将视频片段长度限制为最长 3 秒，并采用批大小 32 来训练我们的 SLTS模型。采用 Adam优化器 [16] ，结合线性预热和余弦退火学习率，其中预热步数为 1k，最大学习率为 5×10^{-4}。对于 GRID 和 TCD-TIMIT 模型，我们选择 conformer (S)；对于 Lip2Wav 模型，则选择 conformer (M)。所有 SLTS模型最多训练 50k 次迭代（在 RTX 2080-Ti 上每次约需 1 天）。为了将模型装入显存，我们采用

[3]此处可获取 SyncNet 的实现和模型检查点：https://github.com/joonson/syncnet_python。

偏移量	短时客观可懂度 ↑	扩展短时客观可懂度 ↑	梅尔倒谱失真 ↓
0 ms	1.000	1.000	0.000
4 ms	0.916	0.869	11.651
8 ms	0.770	0.708	15.605
12 毫秒	0.660	0.594	18.712

表3. 偏移量对对齐敏感指标的影响。结果基于同一音频（来自 GRID-4S 的 bbaf2n）的两个副本计算，其中一个副本经过偏移，以模拟异步。

方法	α-短时客观可懂度 ↑	α-扩展短时客观可懂度 ↑	α-感知语音质量评估 ↑	α-梅尔倒谱失真 ↓	α-词错误率 (%)	平均意见分 (I) ↓↑	平均意见分 (N) ↑
VCA-GAN	0.659	0.477	1.365	34.600	48.20	3.250 ± 0.225	2.042 ± 0.179
SLTS	**0.760**	**0.616**	**1.515**	**29.130**	**24.69**	3.633 ± 0.228	1.858 ± 0.171
SLTS含有判别器	0.738	0.583	1.405	31.856	26.55	$\mathbf{4.483 \pm 0.139}$	$\mathbf{4.267 \pm 0.153}$
真实语音	1.000	1.000	4.549	0.000	0.00	4.808 ± 0.100	4.975 ± 0.028

表 4. Lip2Wav 化学结果。含有判别器：使用判别器训练以生成音频波形。平均意见得分及其 95% 置信区间根据 t 分布计算。

帧编码器上的梯度检查点 [2] 以减少显存占用。作为比较，我们还训练了最先进水平的 VCA-GAN 模型 [14]，最多进行 70k 次迭代，使用 Adam优化器和固定学习率 1×10^{-4}。由于该模型占用更大的内存，我们采用了更小的批大小 24。

5. 实验结果与讨论

除非另有说明，所报告的结果均来自训练过程中在验证集上获得最高时间对齐短时客观可懂度分数的模型。对于 GRID-4S 和 TCD-TIMIT-LS，每训练 1k 次迭代后计算一次时间对齐短时客观可懂度；对于 Lip2Wav，则每训练 5k 次迭代后计算一次。默认情况下，结果是在原始测试集上计算的，且未额外应用唇形同步方法。

5.1. 同步训练的有效性

无论数据集中异步问题的严重程度如何，按照时间对齐指标评估，SLTS 模型的得分均高于未进行同步的竞争模型（即，VCA-GAN 和不含 ASM 的 SLTS）（见表 2）。结果表明，经过恰当评估后，同步训练有助于提升重构音频的语音可懂度、感知质量和梅尔倒谱相似度。此外，根据词错误率（WER）测量，经过同步训练后，重构音频的内容准确性也有所提高。与 GRID-4S 和 TCD-TIMIT-LS 数据集相比，存在更严重数据异步问题的 Lip2Wav 化学数据集取得了更显著的性能提升，尤其是在可懂度和内容准确性方面。这表明，数据集中的严重异步问题不仅会导致生成音频不同步，还会影响生成语音的质量。

5.2. 非对齐指标的局限性

虽然使用 ASM 训练的模型在对齐不敏感指标（例如，感知语音质量评估和词错误率）以及采用对齐前端的指标上，表现出更好的可懂度、感知质量和内容正确性，但在常规短时客观可懂度、扩展短时客观可懂度和梅尔倒谱失真指标上，得分未必始终更高。对于这些指标，我们注意到，测试音频与参考音频之间的轻微偏移可能造成很大的负面影响（见表3）。问题在于，我们的 SLTS 模型经过训练，旨在校正训练数据中的数据不同步；但在测试时，如果参考测试音频与测试视频不同步，那么我们的 SLTS 模型根据测试视频重建出的音频（或许具有完美的零偏移量）也会与参考测试音频不同步。因此，SLTS 模型在短时客观可懂度、扩展短时客观可懂度和梅尔倒谱失真指标上的得分会更低。这体现了对齐敏感指标的局限性。如果没有适当对齐，即使模型表现更好，也可能得到较低的分数。

5.3. 数据同步模块的准确性

由于测试集中没有可用的视听偏移真实标注，我们通过比较不同方法产生的偏移量来评估 DSM 的准确性。R^2 DSM 预测的偏移量与度量前端或 SyncNet 之间的分数也列于表2中。较高的 R^2 DSM 预测的偏移量（即，$\hat{o}_d$）与度量前端（即，$o_f = o_d + o_m$，$o_m \approx 0$）之间的分数表明，DSM 能够准确预测数据不同步情况。o_d 另一方面，与假设所有偏移量均为 0 的虚拟偏移预测器相比，DSM 产生的偏移量能够解释 SyncNet 偏移量中更多的方差。在数据异步问题更严重的数据集上（例如，Lip2Wav 化学），由于视听偏移的总方差较高，R^2 分数会更加显著。

5.4. 判别器对声码器的影响

为了展示采用判别器训练的模型在音频生成方面的优越性（即，MRSD 和 MPWD），我们邀请 12 名志愿者对从 Lip2Wav化学 测试集中随机抽取的 10 个样本进行平均意见分（MOS）测试。评估可懂度和自然度。每位志愿者对这 10 个样本的四个版本（即，VCA-GAN、SLTS、SLTS w/ dis 和真实语音）进行评分。表 4的结果显示，采用判别器后，可懂度和自然度的 MOS 均显著提高。然而，我们注意到，采用判别器后，训练样本和测试样本的客观得分均有所降低。例如，加入判别器后，α-STOI 得分在 Lip2Wav 化学测试集上从 0.760 降至 0.738，在包含 200 个样本的训练子集上从 0.855 降至 0.825。我们推测，较低的客观得分是由其非侵入式特性所致。

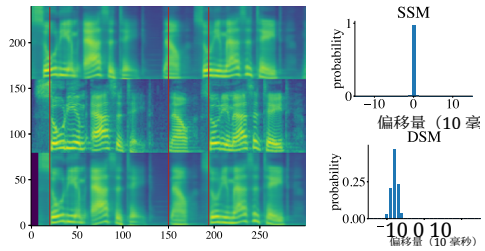


图 6. Lip2Wav 化学中的一个示例，展示了 ASM 的工作原理。左侧从上到下依次显示重建梅尔频谱图、真实梅尔频谱图和硬校正梅尔频谱图。

GAN 训练。判别器会促使生成的音频符合真实音频的分布，而非与相应的目标音频相匹配，因此更难满足均方误差、多分辨率短时傅里叶变换和短时客观可懂度等侵入式学习目标，导致侵入式指标得分较低。

5.5. 定性研究

为了展示 ASM 的工作原理，我们在图中给出了一个真实的训练示例。6在此示例中，重建音频比参考梅尔频谱图早 80 毫秒。DSM 将大部分概率质量分配给约 -80 毫秒的偏移量。将重建梅尔频谱图与硬校正核进行卷积后，所得梅尔频谱图便与真实梅尔频谱图精确对齐，从而能够更准确地计算重建梅尔频谱图与参考梅尔频谱图之间的损失。

5.6. 与其他最先进结果的比较

表 5 至 7将我们的结果与现有工作中报告的最先进结果进行了比较。由于 SLTS 模型生成的音频与输入视频同步，因此当测试集中的数据不同步情况严重时，使用常规指标评估可能会得分较低（例如，化学，如表 2所示）。为进行合理比较，我们报告了经相应 SLTS 模型的 DSM 进行唇音同步后的测试集结果。与其他最先进方法相比，SLTS 的结果相当或更优。在 GRID-4S 上，SLTS 的短时客观可懂度和梅尔倒谱失真指标表现最佳，且除工作 [24] 在 TCD-TIMIT-LS 上的感知语音质量评估外，在所有指标上均优于其他方法。对于 Lip2Wav 中的大多数说话人（即，化学、国际象棋和高中），SLTS 的可懂度显著更高，感知质量也相当（或更优）。在 深度学习和增强英语上，SLTS 的表现与最先进工作 [10]相当或略逊。我们注意到，来自 深度学习和增强英语的视频中，嘴部区域相对较小，因此难以识别视素。这与 SyncNet 的结果（图 5c）一致，该结果也显示其在深度学习和增强英语上的表现置信度较低。

方法	短时客观可懂度 ↑	扩展短时客观可懂度 ↑	感知语音质量评估 ↑	梅尔倒谱失真 ↓
E2E-V2AResNet [24]	0.627	-	2.030	27.790
Yadav等。 [29]	0.724	0.540	1.932	-
VCA-GAN [14]	0.724	**0.609**	**2.008**	-
Lip2Wav [21]	0.731	0.535	1.722	-
Kim等。 [10, 13]	0.738	0.579	1.984	-
SLTS	**0.757**	0.588	1.931	**25.491**

表5. GRID-4S数据集上的最先进结果对比。

方法	短时客观可懂度 ↑	扩展短时客观可懂度 ↑	感知语音质量评估 ↑	梅尔倒谱失真 ↓
E2E-V2AResNet [24]	0.472	-	**1.540**	36.190
Ephrat 等人. [6]	0.487	0.310	1.231	-
基于GAN [27]	0.511	0.321	1.218	-
Lip2Wav [21]	0.558	0.365	1.350	-
VCA-GAN [14]	0.584	0.401	1.425	-
SLTS	**0.661**	**0.507**	1.474	**29.689**

表6. 与 TCD-TIMIT-LS 上最先进结果的比较。

说话人	方法	短时客观可懂度 ↑	扩展短时客观可懂度 ↑	感知语音质量评估 ↑
chem	Lip2Wav [21]	0.416	0.284	1.300
	Hong等. [10]	0.566	0.429	**1.529**
	SLTS	**0.757**	**0.612**	1.514
国际象棋	Lip2Wav [21]	0.418	0.290	1.400
	Hong等. [10]	0.506	0.334	1.503
	SLTS	**0.680**	**0.451**	**1.604**
dl	Lip2Wav [21]	0.282	0.183	**1.671**
	Hong等人。 [10]	**0.576**	**0.402**	1.612
	SLTS	0.565	0.320	1.513
hs	Lip2Wav [21]	0.446	0.311	1.290
	Hong等人。 [10]	0.504	0.337	1.366
	SLTS	**0.590**	**0.394**	**1.402**
eh	Lip2Wav [21]	0.369	0.220	1.367
	Hong等人。 [10]	0.463	**0.304**	1.362
	SLTS	**0.482**	0.268	**1.428**

表 7. 最先进水平的工作与所提出模型在 Lip2Wav数据集 上的比较。与其他数据集不同，我们遵循惯例，为每位说话人训练说话人专属模型。

6. 结论

在这项工作中，我们发现唇语转语音合成训练过程中会出现两种不同步：数据不同步和模型不同步。为解决这些不同步问题，我们提出了一种同步唇到语音模型（SLTS）。在训练过程中，SLTS通过DSM主动学习视听时间偏移量，以纠正数据不同步。我们还通过SSM确保模型同步。此外，我们引入了一个时间对齐前端，将同步评估和音频质量评估与传统的时间对齐敏感指标（如短时客观可懂度、扩展短时客观可懂度和梅尔倒谱失真）分离开来。我们使用这些新指标进行了大量实验，以展示所提模型的优势。与现有最先进的工作相比，我们的方法在多项任务中取得了相当或更优的结果。

参考文献

[1] Hassan Akbari、Himani Arora、Liangliang Cao和Nima Mesgarani。Lip2AudSpec：从静音唇部运动视频重建语音。载于2018年IEEE声学、语音与信号处理国际会议（ICASSP），第2516–2520页。IEEE，2018年。2[2] Nanxin Chen、Yu Zhang、Heiga Zen、Ron J Weiss、Mohammad Norouzi和William Chan。Wavegrad：用于波形生成的梯度估计。载于国际学习表征会议，2021年。7[3] Sumit Chopra、Raia Hadsell和Yann LeCun。以判别方式学习相似性度量，并将其应用于人脸验证。载于2005年IEEE计算机学会计算机视觉与模式识别会议（CVPR'05），第1卷，第539–546页。IEEE，2005年。2[4] J. S. Chung和A. Zisserman。时机不对：野外自动唇音同步。载于多视角读唇研讨会，ACCV，2016年。1、2、3、6[5] Martin Cooke、Jon Barker、Stuart Cunningham和Xu Shao。用于语音感知和自动语音识别的视听语料库。《美国声学学会期刊》，120(5)：2421–2424，2006年。1、5[6] Ariel Ephrat、Tavi Halperin和Shmuel Peleg。从无声视频重建语音。载于IEEE国际计算机视觉会议研讨会论文集，第455–462页，2017年。8[7] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu等。Conformer：用于语音识别的卷积增强型Transformer。arXiv预印本 arXiv:2005.08100，2020年。3[8] Naomi Harte和Eoin Gillen。TCD-TIMIT：连续语音视听语料库。《IEEE多媒体汇刊》，17(5)：603–615，2015年。1、5[9] Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议论文集，第770–778页，2016年。3[10] Joanna Hong、Minsu Kim、Se Jin Park和Yong Man Ro。通过视觉语音记忆，利用回忆式声音进行语音重建。《IEEE/ACM音频、语音与语言处理汇刊》，29：3654–3667，2021年。2、8[11] Won Jang、Dan Lim、Jaesam Yoon、Bongwan Kim和Jun-tae Kim。UnivNet：一种采用多分辨率频谱图判别器的高保真波形生成神经声码器。arXiv预印本 arXiv:2106.07889，2021年。2、3[12] Jesper Jensen和Cees H Taal。一种预测调制噪声掩蔽器掩蔽下语音可懂度的算法。《IEEE/ACM音频、语音与语言处理汇刊》，24(11)：2009–2022，2016年。1、6[13] Minsu Kim、Joanna Hong、Se Jin Park和Yong Man Ro。通过记忆实现多模态关联桥接：从人脸视频中回忆语音声音。载于IEEE/CVF计算机视觉国际会议论文集，第296–306页，2021年。2、8[14] Minsu Kim、Joanna Hong和Yong Man Ro。基于视觉上下文注意力GAN的唇语转语音合成。《Advances 载于神经信息处理系统，34 2770，2021 2758 2、5、6、7、8：–。[15] You Jin Kim、Hee Soo Heo、Soo-Whan Chung 和 Bong-Jin Lee。基于模式分类的端到端唇部同步。载于 2021 IEEE 口语语言技术研讨会（SLT），第 598 605 2021 2 – 页。IEEE，。[16] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412 6980，2014 6。。[17] Daniel Michelsanti、Olga Slizovskaia、Gloria Haro、Emilia G´ omez、Zheng-Hua Tan 和 Jesper Jensen。基于声码器的无声视频语音合成，2020。5[18] Rodrigo Mira、Alexandros Haliassos、Stavros Petridis、Bj¨orn W Schuller 和 Maja Pantic。SVTS：可扩展的视频到语音合成。arXiv 预印本 arXiv:2205.02058，2022。2[19] Rodrigo Mira、Konstantinos Vougioukas、Pingchuan Ma、Stavros Petridis、Bj¨orn W Schuller 和 Maja Pantic。使用生成对抗网络的端到端视频到语音合成。IEEE《控制论汇刊》，2022。2[20] Daniel Povey、Arnab Ghoshal、Gilles Boulianne、Lukas Burget、Ondrej Glembek、Nagendra Goel、Mirko Hannemann、Petr Motlicek、Yanmin Qian、Petr Schwarz、Jan Silovsky、Georg Stemmer 和 Karel Vesely。Kaldi 语音识别工具包。载于 IEEE 2011 自动语音识别与理解研讨会。IEEE 信号处理学会，2011 年 12 月。IEEE 目录编号：CFP11SRW-USB。6[21] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar。学习个体说话风格以实现准确的唇语转语音合成。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 13796–13805 页，2020。1、2、5、6、8[22] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。技术报告，Tech. Rep.，技术报告，OpenAI，2022。6[23] Antony W Rix、John G Beerends、Michael P Hollier 和 Andries P Hekstra。语音质量的感知评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法。载于 2001 IEEE 声学、语音与信号处理国际会议论文集（目录号 01CH37221），第 2 卷，第 749–752 页。IEEE，2001。6[24] Nasir Saleem、Jiechao Gao、Muhammad Irfan、Elena Verdu 和 Javier Parra Fuente。E2E-V2SResNet：用于端到端视频驱动语音合成的深度残差卷积神经网络。《图像与视觉计算》，119 104389，2022 8：。[25] Cees H Taal、Richard C Hendriks、Richard Heusdens 和 Jesper Jensen。用于时频加权噪声语音的短时客观可懂度指标。载于 2010 IEEE 声学、语音与信号处理国际会议，第 4214–4217 页。IEEE，2010。1、6[26] Suramya Tomar。使用 FFmpeg 转换视频格式。《Linux期刊》，2006(146):10，2006。1[27] Konstantinos Vougioukas、Pingchuan Ma、Stavros Petridis 和 Maja Pantic。视频驱动的语音重建

使用生成对抗网络。*arXiv* 预印本 *arXiv:1906.06301*，2019 年。5，8[28] Yongqi Wang 和 Zhou Zhao。FastLTS：非自回归端到端无约束唇语到语音合成。*arXiv* 预印本 *arXiv:2207.03800*，2022 年。2[29] Ravindra Yadav、Ashish Sardana、Vinay P Namboodiri 和 Rajesh M Hegde。使用变分自编码器从无声视频中进行语音预测。载于 *ICASSP 2021-2021 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），第 7048–7052 页。IEEE，2021 年。2，8[30] Shifeng Zhang、Xiangyu Zhu、Zhen Lei、Hailin Shi、Xiaobo Wang 和 Stan Z Li。S^3FD：单次尺度不变人脸检测器。载于 *IEEE* 国际计算机视觉会议论文集，第 192–201 页，2017 年。5