LIPSODY：唇部 T 具有增强韵律一致性的语音合成 CY

Jaejun Lee⋆ *Yoori Oh*⋆ *Kyogu Lee*⋆†‡

⋆ 首尔大学音乐与音频研究组（MARG）、智能与信息系

† 首尔大学人工智能交叉学科项目
‡ 大韩民国首尔大学人工智能研究院

arXiv:2602.01908v1 [cs.SD] 2 Feb 2026

摘要

唇语到语音合成旨在通过从唇部运动中重建语言内容，直接根据无声面部视频生成语音音频，在音频信号缺失或劣化的情况下具有重要应用价值。尽管 LipVoicer 等近期基于扩散的模型在重建语言内容方面表现出色，但它们往往缺乏韵律一致性。本文提出 LipSody，一种增强韵律一致性的唇语到语音框架。LipSody 引入了一种韵律引导策略，利用三种互补线索：从面部图像中提取的说话人身份、根据唇部运动推断的语言内容，以及从面部视频中推断的情感语境。实验结果表明，与以往方法相比，LipSody 显著改善了韵律相关指标，包括全局和局部音高偏差、能量一致性以及说话人相似度。

索引词— 唇语到语音合成、视频到语音、多模态语音合成、韵律

1. 引言

唇语到语音合成旨在从无声视频输入中重建或生成语音音频，大多数研究着重于通过分析说话人嘴唇可见的发音运动来恢复原始语音内容。这项跨模态任务在语音信号缺失或劣化的场景中尤其有价值，例如无声视频片段、听障沟通辅助，或作为多模态语音合成中的补充模态。

得益于深度学习的发展，随着能够将面部视频帧直接映射到声学特征的端到端框架不断发展，唇到语音合成取得了长足进展。然而，现有大多数系统主要关注语言内容的重建，其进展通常通过词错误率（WER）等可懂度指标进行衡量。然而，鉴于面部视频中包含丰富且多样的视觉线索，同样重要的是生成能够传达说话人合理嗓音特征的语音。除了简单地实现唇部运动同步之外，要实现自然且个性化的语音合成，还需要具备与说话人的面部外观和情感表达相一致的逼真韵律与说话人身份。

受这些观察的启发，我们提出了 LipSody，这是一种新颖的唇语到语音框架，旨在增强说话人相关韵律的一致性。与以往主要关注可懂度的研究不同，LipSody 利用视觉输入，不仅获取发音内容，还推断韵律线索和说话人特定的声音特征。在先进的 LipVoicer [1]，基础上，我们的模型通过以三个完全从面部视频中提取的互补嵌入为条件，显式估计韵律特征——即音高和能量：基于唇部运动的语言内容、基于面部图像的说话人身份，以及基于人脸视频的情感表达。通过这一设计，LipSody 生成的语音不仅可懂，而且富有表现力且个性化，仅凭视觉输入便能提升韵律一致性和说话人相似度。

我们的贡献总结如下：

- 我们提出了 LipSody，这是一种基于扩散的唇到语音框架，在保持最先进可懂度的同时，显著提升了韵律一致性。
- 我们提出了一种新颖的纯视觉韵律估计方法，该方法利用三种互补线索：基于唇部运动的语言内容、基于人脸图像的说话人身份，以及基于人脸视频的情感表达。
- 我们使用多样化的韵律相关指标进行了广泛评估，证明了我们方法的有效性，并揭示了韵律一致性与语音可懂度之间的显著相关性。

演示可在以下地址获取：https://jaejunL.github.io/ LipSody_Demo/。

2. 相关工作

2.1. 唇语到语音

唇语到语音合成旨在从静音视频帧重建语音音频。早期工作训练CNN-RNN架构，直接从唇部运动序列预测语音特征 [2]，，主要通过词错误率（WER）评估性能。

近期研究致力于通过基于GAN的细化 [3]，、自监督单元表示 [4, 5], 以及语音分解方法 [6]等技术，提升生成语音的可懂度与自然度。值得注意的是，LipVoicer [1]等近期工作利用基于扩散模型的语音生成 [7]，在语音可懂度方面取得了出色结果。其方法同时利用分类器引导（CG） [8] 和无分类器引导（CFG） [9] 来提升条件生成质量。具体而言，该模型使用预训练唇读模型 [10] 生成伪真实文本描述，并使用预训练自动语音识别（ASR）模型 [11]作为引导分类器。

然而，大多数唇动到语音模型仍难以捕捉音高和能量变化等细粒度韵律特征，导致合成的语音不够自然，也缺乏个性化。

本文已被 ICASSP 2026 接收。

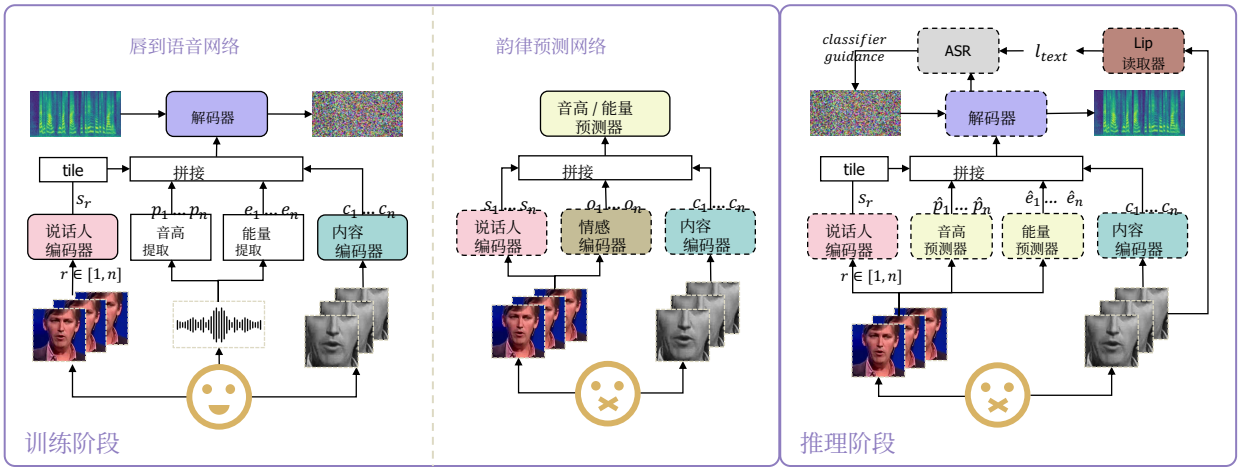


图 1。LipSody 概览：所提出的基于扩散模型的唇语转语音框架，具有增强的基频一致性。训练期间，使用真实基频和能量值，为唇到语音网络提供韵律相关监督。推理时，训练一个独立网络，根据基于唇部运动的语言内容、基于人脸图像的说话人身份以及基于人脸视频的情感表达来预测这些韵律特征。

3. 方法

我们提出 LipSody，这是一种专为增强韵律一致性而设计的唇语到语音框架，旨在生成音高和能量模式与说话人的视觉表情更加准确匹配的语音。

3.1. 基于扩散模型的唇到语音网络

我们的模型基于最先进的基于扩散模型的唇语到语音框架 LipVoicer [1], 构建，该框架已显著提升可懂度，尤其降低了词错误率（WER）。LipVoicer 使用条件去噪扩散概率模型（DDPM） [12], 生成梅尔频谱图，其中条件生成由 CFG [9]引导。条件信号仅来源于视频输入，包含两个部分：用于提取说话人身份特征（$\boldsymbol{s}$）的全脸图像（$\mathcal{S}$），以及用于提取语言内容特征（$\boldsymbol{c}$）的以嘴唇为中心裁剪的视频序列（$\mathcal{C}$）。对于 $\boldsymbol{s}$，从 n 帧中随机选取单帧；对于 $\boldsymbol{c}$，则使用全部 n 帧。推理期间，为进一步提高生成语音的可懂度，LipVoicer 使用 CG [8]。具体而言，它使用预训练自动语音识别模型计算引导信号：$\nabla_{\mathbf{x_t}} \log p(l \mid \mathbf{x}_t)$，其中 $\mathbf{x}_t$ 是扩散步 t处的中间梅尔频谱图，而 l 是预训练唇读模型预测的文本标签。最后，使用预训练声码器将生成的梅尔频谱图转换为波形。

3.2. 使用 LipSody 增强韵律一致性

受 LipVoicer 启发，我们提出了 LipSody——一种增强韵律一致性的基于扩散模型的唇语转语音框架。图 1 展示了整体训练与推理流程。与 LipVoicer 类似，LipSody 使用带 CFG 的基于 DDPM 的架构进行训练。然而，除了利用 $\boldsymbol{s}$ 和 $\boldsymbol{c}$之外，我们还通过注入逐帧基频（$\boldsymbol{p}$）和能量（$\boldsymbol{e}$）值，对韵律特征进行显式建模。这些值在训练期间从配对的真实语音中提取，用作预言机监督信号。

为获得梅尔频谱图，通常会将原始波形 y 除以每个片段中的最大绝对值，以此对其进行归一化。然而，我们假设每位说话人都具有独特的能量分布，而按片段进行归一化会抹去这些说话人特有特征。由于 LipSody 的目标是增强韵律一致性，我们改为按说话人进行归一化。形式上，说话人 i 的第 j个片段的归一化波形定义为 $y^{ij}_{\mathbf{norm}} = y^{ij}/\max(\forall y^{i})$，其中 y^{ij}表示属于说话人 i的第 j个片段的原始波形信号，最大值取自该说话人的所有片段。

在推理过程中，我们保留使用唇读器预测文本的分类器引导策略，以维持 LipVoicer 较强的可懂度。结合我们提出的韵律建模，在每个扩散步计算得到的噪声$\boldsymbol{\epsilon}$ 如下：

$$\begin{aligned}\boldsymbol{\epsilon} = (1+w_1)\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \boldsymbol{s}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{e}}, \boldsymbol{c}) - w_1\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) \\ -w_2\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x_t}} \log p(l \mid \mathbf{x}_t),\end{aligned} \tag{1}$$

其中，$\hat{}\boldsymbol{p}$ 和$\hat{}\boldsymbol{e}$ 分别表示预测基频和能量特征，而 w_1 和 w_2 控制 CFG 和分类器引导的强度。术语 $\bar{\alpha} = \prod_{s=1}^{t} \alpha_s$ 是噪声调度系数的累积乘积，其中 $a_t = 1 - \beta_t$ 和 β_t 是方差调度。与 LipVoicer 一样，我们在应用公式 1 时也采用梯度归一化。

3.3. 韵律预测

推理过程中无法获取 Oracle $\boldsymbol{p}$ and $\boldsymbol{e}$ 值。因此，我们分别训练韵律预测网络，以根据视频衍生特征估计各个值。我们不仅依赖基于人脸图像的说话人身份（$\boldsymbol{s}$）和基于唇部运动的语言内容（$\boldsymbol{c}$），还通过融入视频模态中捕捉到的丰富情感线索，进一步提升韵律估计效果。具体而言，我们使用预训练的情感编码器，生成情感嵌入 $\boldsymbol{o}$。在唇动到语音网络的训练过程中，仅使用随机选取的单帧来提取 $\boldsymbol{s}$，这意味着 $\boldsymbol{s}$ 携带的说话人信息相对粗粒度。我们推测，覆盖更广时间范围的情感嵌入能够捕捉细粒度信息，从而进一步优化逐帧

表1. 传统指标的评估结果

	词错误率↓	短时客观可懂度↑	DNSMOS ↑	LSE-D ↓	LSE-C ↑
$\text{LipVoicer}_{\text{official}}$	0.219	0.93	3.19	8.166	6.308
$\text{LipVoicer}_{\text{recon}}$	0.221	0.93	3.17	8.516	6.400
LipSody（我们的方法）	0.225	0.92	3.18	8.435	6.404

表2。韵律相关指标的评估结果

	GF_0 ↓	LF_0 ↓	EC ↓	Resem ↑	Resem_{tv} ↑
$\text{LipVoicer}_{\text{recon}}$	28.87	45.26	1.2667	0.5974	0.6578
LipSody（我们的方法）	**25.15**	**41.06**	**0.9141**	**0.6040**	**0.6618**

韵律预测。为估计韵律特征，我们将 $\boldsymbol{s}$、$\boldsymbol{o}$ 和 $\boldsymbol{c}$ 拼接后，将合并后的表示输入韵律预测网络。韵律预测网络中的说话人编码器和内容编码器均由基础 LipSody唇语到语音模型预训练；在此阶段，只有音高和能量预测器（图 1）保持可训练状态。每个预测器均使用均方误差（MSE）损失独立优化。

3.4. 模型架构

内容编码器：与 [10] 类似的三维卷积网络
说话人编码器：ResNet-18 [13] ，去掉最后两个全连接层。

解码器：基于 DiffWave 的解码器，由堆叠的扩张残差块组成，具有门控激活，并以梅尔频谱图和韵律特征 [7] 为条件。

基频提取：FastContext-base PitchEstimator (FCPE)，其归一化方式与 SoVITS[2] 类似

能量提取：对数梅尔频谱图沿频率轴求平均。

韵律预测器：由自注意力层和前馈层组成。

情感编码器：基于 EmoCLIP 的编码器，利用对比式文本–图像预训练 [14]。

自动语音识别：音视频 Conformer 网络 [11], , 但仅使用音频输入。

唇读器：基于 Conformer 的唇读模型架构 [15] **声码器**：HiFi-GAN 架构 [16]。

4. 实验

4.1. 数据集

我们使用了 LRS3 数据集 [17], , 其中包含 5,502 段 TED 和 TEDx 视频，总计超过 430 小时的语音。每段视频均经过预处理，包含一块分辨率为 224×224 像素的裁剪面部区域；视频以每秒 25 帧的速率采样，音频则以单声道 16kHz 波形录制。训练时，我们遵循预先定义的 预训练 和训练 划分，并采用测试 划分进行评估，使用 LipVoicer [1] 中采用的未见说话人协议。口部区域裁剪遵循 [18] 中的方法。我们提取以嘴部为中心的 96×96 边界框作为感兴趣区域。视频图像块会转换为灰度，并使用训练集的全局均值和方差进行归一化。

4.2. 对比系统

由于我们的首要目标是在保持较高可懂度的同时增强韵律，因此我们排除了近期一些报告了相对较差词错误率的研究（ [4], 29中为 33.9%， [5], 中为 8%，[6] 中为 28.5% ）。相反，我们将比较重点放在最先进的框架 LipVoicer [1]（词错误率为 21.9%）上，并针对其两个变体评估我们的方法。

[1]https://github.com/CNChTu/FCPE [2] https://github.com/svc-develop-team/so-vits-svc

第一种是原论文报告的官方模型，具体而言是采用 HiFi-GAN [16] 声码器的版本，记为 $\text{LipVoicer}_{\text{official}}$。第二种是我们使用公开代码库并采用默认训练设置重新实现的模型，记为 $\text{LipVoicer}_{\text{recon}}$。我们提出的模型融入了显式韵律建模，称为 LipSody。

4.3. 实现细节

对于预处理和基础唇动到语音模型，我们遵循官方LipVoicer实现 [1], , 包括采用

$T = 400$ $\beta_1 =$ 去噪步数、0.0001 $\beta_T = 0.02$ 方差调度从到的扩散设置。韵律预测器和情感编码器分别基于SoVITS和EmoCLIP [14], 的公开实现构建。训练按照原始协议进行100K步，批大小为256。关于CFG和分类器引导，权重 w_1 和 w_2 分别固定为2和1.5，遵循 [1]的原始配置。

4.4. 指标

我们使用客观指标和主观指标来评估模型。客观评估包括两类：（1）唇动到语音文献中常用于评估可懂度和质量的常规指标，以及（2）我们引入的韵律相关指标，用于评估模型捕捉韵律信息的能力，例如音高和能量一致性。对于主观评估，我们使用Amazon Mechanical Turk（MTurk）进行了自然度测试和ABX测试。在ABX测试中，评定者判断哪一个合成语音样本的韵律与真实值更为相似。

4.4.1. 客观指标（传统指标）

词错误率：使用自动语音识别模型 [15] 计算，以评估生成语音的可懂度。

短时客观可懂度：使用 STOI- Net [19]估算的非侵入式可懂度指标。

DNSMOS：用于评估感知语音质量 [20]的非侵入式神经网络平均意见分预测器。
LSE：使用 SyncNet [21], 测量唇音同步误差，包含两个组成部分：LSE-D（唇部与音频之间的距离）和 LSE-C（同步置信度）。

4.4.2. 客观指标（韵律相关）

$\textbf{GF}_0$（全局 F_0）：合成平均音高与说话人全局 F_0（所有片段的平均值） [22], 之间的 L1损失，用于评估说话人特定音高统计。

$\textbf{LF}_0$（局部 F_0）：合成语音与真实语音的逐帧音高轮廓之间的 L1损失 [23], , 用于评估细粒度、内容对齐的音高变化。

EC（能量一致性）：对数梅尔频谱图中逐帧能量值（在频率维度上取平均）之间的 L2损失。
Resem：使用 Resemblyzer [3] 计算合成音频与参考音频的说话人嵌入之间的余弦相似度。

[3]https://github.com/resemble-ai/Resemblyzer

表3. 主观指标的评估结果

	自然度 ↑	ABX 测试(%) ↑
真实语音	4.02	-
$\text{LipVoicer}_{\text{recon}}$	3.36	45.78
LipSody（我们的方法）	3.47	**54.22**

表4。根据韵律信息设置的结果

	词错误率 ↓	GF_0 ↓	LF_0 ↓	EC ↓
$\text{LipVoicer}_{\text{recon}}$	0.221	28.87	45.26	1.2667
LipSody（我们的方法）	0.225	25.15	41.06	0.9141
w/o 情感 $\boldsymbol{o}$	0.220	27.11	41.26	0.9402
w 预言 $\boldsymbol{p,e}$	**0.210**	**16.98**	**31.11**	**0.4620**

Resem$_{tv}$：时变Resem，采用窗口余弦相似度（2秒窗口，1秒步长），用于评估动态说话人身份。

4.4.3. 主观指标

自然度：使用 5 分制平均意见分（MOS）评估合成音频的感知质量（从完全不自然到完全自然）。每项任务分配给 100 名受试者，每人奖励 0.5 美元。

ABX测试：测量两个模型之间的主观偏好，参与者需判断哪个合成样本的韵律与配对的真实音频更为一致。参与者被要求仅关注音高、能量和语音风格等韵律方面，而忽略发音错误或音频质量。每项任务分发给100名受试者，报酬为2美元。

5. 结果

5.1. 使用客观传统指标进行评估

常规指标评估的结果见表1。首先，$\text{LipVoicer}_{\text{official}}$ 与 $\text{LipVoicer}_{\text{recon}}$ 的比较表明，重建模型在所有指标上均达到了与官方实现相当的性能。尽管$\text{LipVoicer}_{\text{recon}}$ 的词错误率和LSE略差，但差异很小。在将LipSody与$\text{LipVoicer}_{\text{recon}}$进行比较时，我们在任何传统指标上均未观察到统计学显著差异（配对t检验中$p > 0.05$ ）。这些结果表明，LipSody保持了较高的可懂度和感知质量，验证了我们的韵律增强策略不会损害基线性能。

5.2. 使用客观韵律相关指标进行评估

表2展示了韵律相关指标的结果，这些指标直接评估我们所提出方法的主要目标。在全部五项韵律指标上，LipSody均持续优于 LipVoicer（$p < 0.01$ 在配对t检验中）。GF_0 和LF_0 的改进表明，LipSody能够更好地捕捉说话人特定的声音特征，并且仅从视觉输入中生成更自然、与内容一致的音高轮廓。EC指标上的提升进一步支持了这一发现，表明模型对与韵律相关的逐帧语音动态建模能力有所增强。最后，Resem和Resem_{tv}的结果表明，LipSody不仅能够在粗粒度的片段级别保留说话人身份，还能捕捉其随时间变化的波动，从而实现更忠实、更具表现力的语音合成。

需要注意的是，EC 是从梅尔频谱图中以非参数方式计算得出的，而 GF_0 和 LF_0 则使用参数化音高提取器 FCPE 进行估计。为确保我们的发现并非依赖于这一特定估计器，我们还使用了若干替代参数化提取器 [24, 25]来计算 GF_0 和 LF_0 。在所有方案中，LipSody 的表现始终显著优于 LipVoicer。受篇幅限制，此处省略详细结果。

5.3. 主观指标评估

表3展示了自然度和 ABX 测试的结果。在自然度方面，LipSody 的得分略高于 LipVoicer，但差异在统计上并不显著（$p > 0.05$配对t检验）。这一结果符合预期，因为我们的主要贡献在于提升韵律，同时保持可懂度；因此，自然度达到与基准模型相当的水平是合理的。在衡量韵律与真实值一致性的ABX偏好测试中，LipSody 的表现显著优于 LipVoicer。我们进一步使用单样本t检验，检验 ABX 准确率是否高于 0.5 的随机水平，结果发现，我们的模型取得了显著优于随机水平的表现（$p < 0.05$）。这些主观评估结果进一步证明了我们的方法在保持可懂度的同时提升韵律一致性的有效性。

5.4. 韵律信息方面的性能

表4比较了LipSody框架三种变体的性能。第一种变体对应于表1和表2中使用的完整LipSody配置。第二种变体记为“w/o 情感 $\boldsymbol{o}$”，从韵律预测网络中移除了情感编码器（图 1 ）。在该设置下，韵律预测器仅接收说话人嵌入 $\boldsymbol{s}$ 和内容嵌入 $\boldsymbol{c}$，而不包含情感嵌入 $\boldsymbol{o}$。与完整的LipSody模型相比，其韵律相关指标出现了统计显著的下降，包括GF_0, LF_0, 和EC（配对t检验中 $p < 0.01$ ）。这表明，纳入从面部图像序列中提取的情感相关特征，对准确的韵律估计具有实质性贡献。第三种变体记为“w 理想 $\boldsymbol{p,e}$”，在推理过程中使用真实基频 $\boldsymbol{p}$ 和能量 $\boldsymbol{e}$ 。正如预期，该配置在韵律相关指标上带来了大幅改善，因为它作为上界参考，此时真实的韵律特征是可用的。有趣的是，该设置下词错误率也有明显改善。这一结果表明，更准确的韵律估计不仅提升了韵律一致性，也有助于提高语音可懂度。

6. 结论

我们提出了 LipSody，这是一种新颖的唇语到语音合成框架，通过将视觉驱动的音高和能量建模整合到基于扩散的生成流程中，增强了韵律一致性。与主要强调可懂度的以往方法不同，LipSody 明确利用视觉线索进行韵律建模，包括从面部图像中提取说话人身份、从唇部运动中提取语言内容，以及从面部视频中提取情感上下文。

我们的方法在保持较高可懂度的同时，显著提升了与韵律相关的指标。此外，我们发现，增强韵律建模也有助于提高可懂度，这凸显了表现力传达与内容清晰度之间的紧密联系。

7. 致谢

本研究获韩国国家研究基金会（NRF）通过教育部资助的基础科学研究计划支持 [编号：RS-2025-25398143，50%], 信息通信技术规划与评估院（IITP）获韩国政府（科学技术信息通信部）资助的项目 [编号：RS2022-II220641，45%], 和 [编号：RS-2021-II211343，5%]

8. 参考文献

[1] Yochai Yemini、Aviv Shamsian、Lior Bracha、Sharon Gan- not 和 Ethan Fetaya，《LipVoicer：在唇读的引导下从无声视频生成语音》，*arXiv* 预印本 *arXiv:2306.03258*，2023年。[2] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman，《深度音视频语音识别》，*IEEE* 模式分析与机器智能汇刊，第44卷，第12期，第8717–8727页，2018年。[3]Sindhu B Hegde、KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri 和 CV Jawahar，《面向真实环境中任意说话人的唇语到语音合成》，载于第*30*届*ACM*国际多媒体会议论文集，2022年，第6250–6258页。[4] Wei-Ning Hsu、Tal Remez、Bowen Shi、Jacob Donley 和 Yossi Adi，《Revise：利用视觉输入进行自监督语音再合成，以实现通用且泛化的语音再生成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第18795–18805页。[5] Jeongsoo Choi、Minsu Kim 和 Yong Man Ro，《利用语音单元实现清晰易懂的唇语到语音合成》，*arXiv* 预印本 *arXiv:2305.19603*，2023年。[6] Jeongsoo Choi、Ji-Hoon Kim、Jinyu Li、Joon Son Chung 和 Shujie Liu，《V2sflow：结合语音分解与整流流的视频到语音生成》，载于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2025年，第1–5页。[7] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro，《DiffWave：一种多用途音频合成扩散模型》，*arXiv* 预印本 *arXiv:2009.09761*，2020 年。[8] Prafulla Dhariwal 和 Alexander Nichol，《扩散模型在图像合成方面胜过生成对抗网络》，神经信息处理系统进展，第34卷，第8780–8794页，2021年。[9] Jonathan Ho 和 Tim Salimans，《无分类器扩散引导》，*arXiv* 预印本 *arXiv:2207.12598*，2022年。[10] Pingchuan Ma、Brais Martinez、Stavros Petridis 和 Maja Pan- tic，《迈向实用的唇读：蒸馏且高效的模型》，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第7608–7612页。

[11] Maxime Burchi 和 Radu Timofte，“用于鲁棒语音识别的音视高效 Conformer”，发表于 *IEEE/CVF* 冬季计算机视觉应用会议论文集，2023年，第2258–2267页。

[12] Jonathan Ho、Ajay Jain 和 Pieter Abbeel，“去噪扩散概率模型”，神经信息处理系统进展，第33卷，第6840–6851页，2020年。

[13] 何恺明、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，2016年，第770–778页。[14] Niki Maria Foteinopoulou 和 Ioannis Patras，“EmoCLIP：一种用于零样本视频面部表情识别的视觉语言方法”，载于*2024*年第*18*届*IEEE*自动人脸与手势识别国际会议（*FG*）。IEEE，2024年，第1–10页。[15] Pingchuan Ma、Alexandros Haliassos、Adriana Fernandez-Lopez、陈宏烈、Stavros Petridis 和 Maja Pantic，“Auto-avsr：使用自动标签的音视语音识别”，载于*ICASSP 2023-2023 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2023年，第1–5页。[16] 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”，神经信息处理系统进展，第33卷，第17022–17033页，2020年。[17] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman，“Lrs3-ted：用于视觉语音识别的大规模数据集”，*arXiv*预印本 *arXiv:1809.00496*，2018年。[18] Pingchuan Ma、Stavros Petridis 和 Maja Pantic，“真实环境下多种语言的视觉语音识别”，《自然·机器智能》，第4卷，第11期，第930–939页，2022年。[19]Ryandhimas E Zezario、Szu-Wei Fu、Chiou-Shann Fuh、Yu Tsao 和 Hsin-Min Wang，“STOI-Net：一种基于深度学习的非侵入式语音可懂度评估模型”，载于*2020*年亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）。IEEE，2020年，第482–486页。[20] Chandan KA Reddy、Vishak Gopal 和 Ross Cutler，“Dnsmos p. 835：一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 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第886–890页。[21] Joon Son Chung 和 Andrew Zisserman，“Out of time：真实环境中的自动唇形同步”，载于计算机视觉——*ACCV 2016*研讨会：*2016*年*11*月*20*日至*24*日在台北举行的*ACCV 2016*国际研讨会，修订精选论文，第二部分第*13*卷。施普林格，2017年，第251–263页。[22] Jaejun Lee、Yoori Oh、Injune Hwang 和 Kyogu Lee，“Hear your face：基于面部的语音转换与基频估计”，载于*Interspeech 2024*论文集，2024年，第4378–4382页。[23] Jaejun Lee、Yoori Oh 和 Kyogu Lee，“Speaking without sound：仅使用面部输入的多说话人无声语音发声”，载于*ICASSP 2025-2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2025年，第1–5页。[24] 魏浩杰、曹雪珂、单唐鹏和陈月国，“Rmvpe：一种用于复调音乐人声基频估计的鲁棒模型”，*arXiv*预印本 *arXiv:2306.15412*，2023年。[25] Jong Wook Kim、Justin Salamon、Peter Li 和 Juan Pablo Bello，“Crepe：用于基频估计的卷积表征”，载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第161–165页。