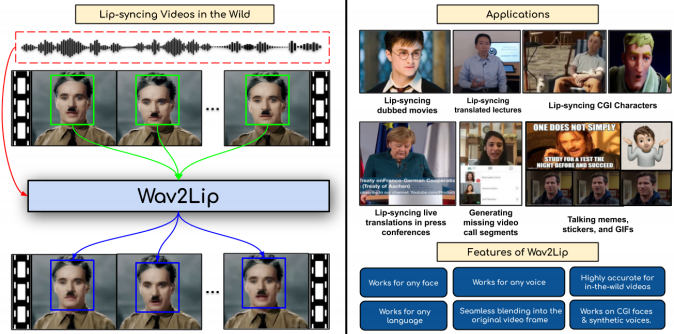
**A Lip Sync Expert Is All You Need for Speech to Lip Generation**

**In The Wild**





**图1：我们的新型Wav2Lip模型在动态、不受约束的说话人脸视频中产生了显著更准确的唇同步。定量指标表明，我们生成的视频中的口型几乎和真实的视频一样好。因此，我们相信我们的模型可以实现广泛的现实应用，以前的说话者独立的语音方法需要克服一些困难[17,18]才能产生令人满意的结果。**

**摘要**

在这项工作中，我们研究了用任意身份的语音视频来匹配目标语音片段的问题。目前的工作擅长在训练阶段看到的特定人的静态图像或视频上产生准确的嘴唇动作。

\*两位作者对这项研究的贡献相同。

允许为个人或课堂使用全部或部分作品的数字或硬拷贝，但副本不是为利润或商业利益而制作或分发，且副本载有本通知和完整引用。必须尊重作者以外的作品的版权。给予原文适当的署名或认可的情况下摘录文章的部分内容是被允许的。若要以其他方式复制或重新发布，或在服务器上发布，或分发到邮件列表，需要事先获得特定的许可和/或支付相关费用。从permissions@acm.org中请求权限。

2020年10月10-12-16日，美国华盛顿州西雅图

©2020版权由所有者/作者所有。授权给ACM的出版权利。

ACM ISBN 978-1-4503-7988-5/20/10。..$15.00

https://doi.org/10.1145/3394171.3413532

然而，在动态、不受约束的人脸视频中，它们不能准确地变化任意身份的嘴唇动作，导致视频的大部分与新音频不同步。我们确定了与此相关的关键原因，并通过从一个强大的对口型鉴别器来解决它们。接下来，我们提出了新的、严格的评估基准和度量标准来准确地测量无约束视频中的嘴唇同步。对我们具有挑战性的基准进行的广泛定量评估表明，我们的Wav2Lip模型生成的视频的对口型精度几乎和真实同步的视频一样好。我们在网站上提供了一个演示视频，清楚地展示了我们的Wav2Lip模型和评估基准的实质性影响: cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in-the-wild.这些代码和模型将在这里发布： github.com/Rudrabha/Wav2Lip。你也可以在这个链接上试用交互式演示：: bhaasha.iiit.ac.in/lipsync.

**CCS概念**

**.Computing methodologies → Computer vision; Learning**

**from critiques; Phonology/morphology.**

**关键字**

对口型，视频生成，说话人脸生成

**ACM参考格式：**

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C V Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to

Lip Generation In The Wild。收录于第28届ACM国际多媒体会议论文集

（MM '20）2020年10月12日至16日，美国华盛顿州西雅图举办。美国纽约

市，美国计算机协会（ACM），共10页https://doi.org/10.1145/3394171.3413532

**1 介绍**

随着视听内容消费的指数级增长[[21](#_bookmark4)]，快速创建视频内容已经成为一个典型的需求。与此同时，使这些视频能够以不同语言呈现也是一个关键挑战。例如，一个深度学习系列讲座、一部著名电影，或一个对全国的公开演讲，如果翻译成理想的目标语言，可以提供给数百万观众。翻译这类面部语音视频或创建新视频的一个关键方面是纠正口型，以匹配提供的目标语音。因此，匹配给定的输入音频流的对口型面部视频在研究界受到了相当大的关注[[6](#_bookmark5),[13](#_bookmark6),[17](#_bookmark1),[18](#_bookmark2),[23](#_bookmark7)]。

在这一领域，最初使用深度学习的工作是通过几个小时的单一发言人数据学习了从语音表示到唇部标记的映射。最近的作品[[13](#_bookmark6),[23](#_bookmark7)]在这一方向中，直接从语音表示中生成图像，并显示出他们经过训练的特定演讲者的特殊生成质量。然而，许多实际应用程序都需要能够很容易地用于通用身份和语音输入的模型。这导致了说话模型的产生需要[[17](#_bookmark1),[18](#_bookmark2)] 经过数千种身份和声音的训练。它们可以在任意音频，任意身份的单一静态图像上产生精确的嘴唇运动，包括由文本到语音系统生成的合成语音的图像[[18](#_bookmark2)]然而，要应用于翻译讲座/电视等方面，这些模型需要能够变化这些动态且无约束视频中广泛的嘴唇形状，而不仅仅是在静态图像上。

我们的工作建立在后一类独立于说话者的分类上，这些作品能够同步任何身份和声音的面部视频。我们发现，这些适合静态图像的模型不能准确地变化无约束视频内容中的大量嘴唇形状，导致生成的视频的很大部分与新的目标音频不同步。观看者可以识别出只有≈0.05-0.1秒那么小的不同步视频片段[9]。因此，将一个真实世界的视频对口型转换成一个较为真实，能令人信服的全新说话视频，是相当具有挑战性的。此外，我们独立于说话者，而没有任何额外的特定于说话者的数据开销的目标使我们的任务更加艰巨。现实世界的视频包含快速的姿势、比例和照明变化，生成的面部结果也必须无缝地融合到原始目标视频中。

我们首先查看现有的独立于说话人的方法。我们发现，这些模型并没有充分惩罚错误的唇形形状，无论是由于只使用重建损失还是弱唇同步鉴别器。我们采用了一个强大的唇同步鉴别器，可以强制发电机持续地产生准确，真实的嘴唇运动。接下来，我们重新检查当前的评估方案，并从三个标准测试集中设计出新的、严格的评估基准。我们还提出了使用SyncNet [[9](#_bookmark10)]来精确评估无约束视频中的对口型。我们还收集并发布了ReSyncED，这是一组具有挑战性的真实世界视频，可以作为模型在实践中的基准。我们进行了广泛的定量和主观的人类评估，并在所有的基准测试中都大大优于以前的方法。我们的主要贡献/索赔要求如下：

我们提出了一种新的唇同步网络，Wav2Lip，它明显比以前的工作更精确的唇同步任意语音的面部视频。

我们提出了一个新的评估框架，包括新的基准和度量，以使一个公正的判断在无约束视频的唇同步。

我们收集并发布了ReSyncED，一个真实的对口型评估数据集，用来在野外完全看不见的视频上对对口型模型的性能进行基准测试。

Wav2Lip是第一个生成视频的假唱精度与真实的同步视频。人工评估表明，Wav2Lip生成的视频比现有的方法和超过90%的时间的非同步版本更受青睐。

演示视频可以在我们的网站找到[website1](#_bookmark11)通过几个定性的例子，清楚地说明了我们的模型的影响。我们还将在网站上发布一个交互式演示，允许用户使用他们选择的音频和视频样本来试用该模型。本文的其余部分组织如下：以下部分[2](#_bookmark12)调查了从言语到唇代等领域的最新发展[3](#_bookmark13)讨论现有工作的问题，并描述我们提出的方法来减轻它们，部分[4](#_bookmark14)提出了一个新的、可靠的评价框架。我们描述了各种潜力

应用和解决部分中的一些伦理问题[5](#_bookmark15)并在章节中结束[6](#_bookmark16).

**2相关工作**

**2.1从言语中产生有限的说话人脸**

我们首先回顾了关于谈论人脸生成的工作，这些工作要么受到它们可以产生的身份范围的限制，要么受到它们被限制的词汇表范围的限制。最近的一些作品实现了真实生成的面部对话视频[[19](#_bookmark8),[22](#_bookmark9)奥巴马的视频。他们学习输入音频之间的映射

1[cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-言语对唇的一代](cvit.iiit.ac.in/research/projects/cvit-projects/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in-the-wild)

以及相应的嘴唇标志。由于他们只接受过一个特定的说话者的训练，他们无法合成新的身份或声音。它们还需要一个特定扬声器的大量数据，通常需要几个小时。最近的一项研究表明，关于这一点的研究成果[[13](#_bookmark6)]建议通过在演讲中添加或删除短语来无缝编辑单个演讲者的视频。他们仍然需要每个说话者一个小时的数据来完成这项任务。最近，另一项工作[[23](#_bookmark7)]试图通过使用两阶段的方法来最小化数据开销，他们首先学习独立于说话人的特征，然后学习使用≈5分钟数据的渲染映射。然而，他们在一个小得多的语料库上训练说话者独立网络，并且在需要为每个目标说话者生成干净的训练数据方面还有额外的开销。现有作品的另一个限制是在词汇量方面。几件作品[5](#_bookmark17),[26](#_bookmark18),[28](#_bookmark19)]在数据集上的训练与一个有限的单词集，如GRID [[10](#_bookmark20)]（56个单词），TIMIT [[14](#_bookmark21)]和LRW [[8](#_bookmark22)]（1000个单词）极大地阻碍了一个模型学习真实视频中音素-视觉素映射的巨大多样性[[18](#_bookmark2)].我们的工作重点是对口型、不受约束的会说话的面部视频，以匹配任何目标语音，而不受身份、声音或词汇的限制。

**2.2从言语中产生的不受约束的会说话的人脸**

尽管关于语音驱动的人脸生成的作品数量在增加，但令人惊讶的是，很少有作品被设计成任意身份、声音和语言的假型视频。他们没有接受过一小组身份或少量词汇表的训练。这允许他们在测试时对任何语音进行对口型随机身份。据我们所知，只有两部这样杰出的作品[[17](#_bookmark1),[18](#_bookmark2)]存在于当前的文献中。请注意，[17](#_bookmark1)]是[的一个扩展版本[7](#_bookmark24)].这两项工作[17](#_bookmark1),[18](#_bookmark2)]表示野外唇同步学习任务如下：给定短语音段S和随机参考人脸图像R，网络的任务是生成唇同步版本Lд与音频相匹配的输入面。此外，LipGAN模型还输入下半部分掩蔽的目标面作为姿态先验。这是至关重要的，因为它允许生成的人脸作物被无缝地粘贴回原始视频中，而不需要进行进一步的后处理。它还训练一个鉴别器与生成器一起来区分不同步或不同步的音频-视频对。然而，这两种作品都有一个显著的局限性：它们在任意身份的静态图像上工作得很好，但当它们在野外尝试假唱不受约束的视频时，会产生不准确的嘴唇生成。与在LipGAN中使用的GAN设置相比，[[18](#_bookmark2)]，我们使用一个预先训练的，精确的，唇同步鉴别器，没有用生成器进一步训练。我们观察到，这是一个重要的设计选择，以实现更好的对口型结果。

**3个准确的语音驱动的假唱视频在野外**

我们的核心架构可以总结为“通过从一个训练有素的对口型专家那里学习来生成准确的对口型”。为了理解这种设计选择，我们首先确定了现有架构的两个关键原因[2.2](#_bookmark23))在野外产生不准确的假唱视频。我们认为，损失的函数，即

在两个现有工程中使用的L1重建损失[[17](#_bookmark1),[18](#_bookmark2)]和LipGAN中的鉴别器损失[[18](#_bookmark2)]不足以惩罚不准确的假唱生成。

**3.1像素级重建损失是唇同步的弱判断**

计算整个图像的人脸重建损失，以确保正确的姿态生成，保留身份，甚至在人脸周围的背景。唇部区域对应的区域不到总重建损失的4%（基于空间范围），因此在网络开始进行细粒度嘴唇形状校正之前，首先对周围的大量图像重建进行优化。这进一步支持了一个事实，即网络开始变形的嘴唇只有在大约一半(≈11th通过其训练过程(≈20个时代[[18](#_bookmark2)]).因此，有一个额外的鉴别器来判断假型是至关重要的，就像在LipGAN中所做的那样[[18](#_bookmark2)].但是，在LipGAN中使用的鉴别器有多强大呢？

**3.2一个弱唇同步鉴别器**

我们发现，在LRS2测试集上检测非同步时，LipGAN的唇同步唇对的准确率仅为56%左右。为了进行比较，我们将在这项工作中使用的专家鉴别器在同一测试集上的准确率为91%。我们假设了造成这种差异的两个主要原因。首先，LipGAN的鉴别器使用a

单帧检查是否有唇同步。在表中[3](#_bookmark26)，我们展示了一个小的

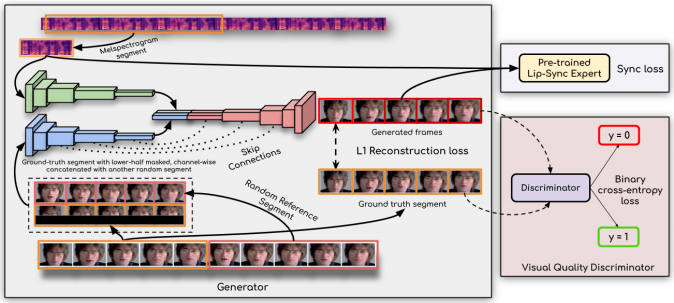
时间上下文在检测对口型时非常有用。其次，在训练过程中生成的图像由于大尺度和姿态变化而包含大量的伪影。我们认为，在GAN设置的基础上训练这些噪声生成的图像，就像在LipGAN中所做的那样，导致鉴别器聚焦于视觉伪影，而不是音频-唇对应。这导致了

不同步检测精度的大幅下降(表[3](#_bookmark26)).我们论证并表明，从实际的视频帧中捕获的“真实的”、准确的假唱概念可以用来准确地区分和加强生成的图像中的假唱。

**.33你只需要一个唇同步专家**

基于上述两个发现，我们建议使用一个预先训练好的专家对口型鉴别器，以准确地检测真实视频中的同步。此外，它不应该像在LipGAN中那样对生成的帧进行进一步的微调。其中一个这样的网络，已被用于纠正对口型错误，以创建大型对口型数据集[[1](#_bookmark28),[3](#_bookmark29)]是SyncNet [[9](#_bookmark10)模型我们建议调整和训练一个改进版本的SyncNet [[9](#_bookmark10)为我们的任务。

3.3.1SyncNet概述。SyncNet [[9](#_bookmark10)]输入一个窗口V，该窗口V为电视机连续的面部帧（仅限下半部分）和大小的语音段S谢谢×D，其中电视机和谢谢分别是视频和音频的时间步长。它被训练为通过随机采样一个音频窗口来区分音频和视频之间的同步谢谢×D，要么与视频对齐（不同步），要么来自不同的时间步长（不同步）。它包含一个面编码器和一个音频编码器，两者都由一堆二维卷积组成。计算由这些编码器生成的嵌入之间的L2距离，并使用最大边际损失训练模型，以最小化（或最大化）同步（或未同步）对之间的距离。



**图2：我们的方法通过向“已经训练有素的对口型专家”学习来生成准确的对口型。不像以前的工作，只使用一个重建损失[[17](#_bookmark1)或者在GAN设置中训练一个鉴别器[18](#_bookmark2)]，我们使用了一个预先训练过的鉴别器，它在检测对口型错误方面已经相当准确了。我们表明，对产生噪声的人脸进行进一步的微调，阻碍了鉴别器测量唇同步的能力，从而也影响了生成的嘴唇形状。此外，我们还采用了一个视觉质量鉴别器来提高视觉质量和同步精度。**

3.3.2，我们的专家对口型鉴别器。我们对SyncNet进行了以下更改[[9](#_bookmark10)来训练一个适合我们的唇型生成任务的专家唇型识别器。首先，我们不像原始模型那样提供按通道连接的灰度图像，而是提供彩色图像。其次，我们的模型明显更深，有剩余的跳跃连接[[15](#_bookmark31)].第三，受到这种公共实施的启发[2](#_bookmark32)，我们使用了一个不同的损失函数：余弦相似度和二值交叉熵损失。也就是说，我们计算relu激活的视频和语音嵌入v之间的点积，得到每个样本的[0,1]之间的一个值，表示输入的音频-视频对同步的概率：

|  |  |  |
| --- | --- | --- |
| P同步= | v · s | (1) |
| max(∥v ∥2· ∥s∥2, e) |

我们在LRS2列车上训练我们的专家对口型鉴别器

分割（≈29小时），批量大小为64，Tv=5帧使用Adam优化器[[12](#_bookmark34)]，初始学习速率为1e−3.我们的专家对口型鉴别器在LRS2测试中的准确率约为91%

而在LipGAN中使用的鉴别器在同一测试集上的准确率仅为56%。

**3.4通过向唇同步专家学习，产生准确的唇同步**

现在我们有了一个精确的唇同步鉴别器，我们现在可以使用它来惩罚生成器(图[2](#_bookmark30))，表示在训练期间进行不准确的生成。我们从描述生成器架构开始。

2 github.com/joonson/syncnet\_trainer

3.4.1.发电机体系结构的细节。我们使用了与LipGAN所使用的类似的生成器架构[[18](#_bookmark2)].我们的关键贡献在于用专家鉴别器来训练它。生成器G包含三个块：(i)身份编码器，（ii）语音编码器，和（iii）人脸解码器。身份编码器是一堆残差卷积层，编码一个随机参考系R，与沿通道轴的姿态先验P（下半部分掩蔽的目标面）连接。语音编码器也是2个卷积的堆栈，用于对输入的语音段S进行编码，然后与人脸表示相连接。解码器也是一堆卷积层，以及用于上采样的转置卷积。生成器被训练为最小化生成的帧L之间的L1重建损失д和真实的框架lg:

N

L反演=**x**||Lд− LG||1(2)

=1

i

因此，该生成器类似于之前的工作，一个独立生成每一帧的2D-CNN编码-解码器网络。

然后我们如何使用我们预先训练过的专家对口型鉴别器，这需要一个时间窗口电视机=5帧作为输入吗？

3.4.2惩罚不准确的嘴唇生成。在培训期间，作为专家鉴别器在分科接受培训[3.3](#_bookmark27)过程电视机=5连续帧，我们还需要生成器G来生成所有的电视机=5帧。我们为参考系采样一个随机连续窗口，以确保尽可能多的姿态时间一致性等。穿过电视机窗户当我们的生成器独立地处理每一帧时，我们会堆叠时间步长

批量维度，同时提供参考帧，以获得一个输入形状(N·电视机，H、W、3)，其中N、H、W分别为批的大小、高度和宽度。当将生成的帧提供给专家鉴别器时，时间步长沿着通道维数连接起来，就像在训练鉴别器时所做的那样。向专家识别器得到的结果输入形状为（N、H\2、W、3·Tv），其中只有生成的人脸的下半部分用于识别。生成器也被训练以最小化“专家同步损失”E同步专家鉴别员：

N

E同步=**x**−日志(P yncis) (3)

=1

i

其中Pi同步是根据公式计算的吗[1](#_bookmark33).注意，

专家鉴别器的权重在训练过程中保持冻结

的发电机。这种强烈的歧视纯粹是基于

从真实视频中学到的对口型概念迫使发电机

也可以实现现实的对口型，以尽量减少对口型的损失E同步.

**3.5生成逼真的面孔**

在我们的实验中，我们观察到使用一个强唇同步鉴别器迫使发生器产生精确的嘴唇形状。然而，它有时会导致变形的区域稍微模糊或包含轻微的伪影。为了减轻这种质量上的微小损失，我们在GAN设置中训练了一个简单的视觉质量鉴别器。因此，我们有两个鉴别器，一个用于同步精度，另一个用于更好的视觉质量。对口型鉴别器没有在GAN设置中进行训练

解释在[3.2](#_bookmark25).另一方面，由于视觉质量鉴别器不对唇同步进行任何检查，只惩罚不现实的人脸生成，所以它是在生成的人脸上进行训练的。

鉴别器D由一堆卷积块组成。每个块由一个卷积层组成，然后是一个泄漏的ReLU激活[[20](#_bookmark38)].鉴别器被训练以最大化

目标函数Ldisc（方程[5](#_bookmark39)):

Lдen=Ex∼Lд[loд(1 − D (x)] ([4)](#_bookmark41)

Ldisc=Ex∼LG[loд(D (x))] + Lдen([5)](#_bookmark36)

其中Lд对应于来自生成器G的图像，和lg对应于真实的图像。

发电机最小化方程[6](#_bookmark42)，为重建损失的加权和(式[2](#_bookmark35)）、同步损失（式[3](#_bookmark37))和对抗性损失Lдen（方程[4](#_bookmark40)):

L总数= (1 −sw − sд)· L反演+sw· E同步+ sд· Lдen(6)

在哪里sw同步惩罚权重为sд是对抗性损失，在我们所有的实验中，经验设置为0.03和0.07。因此，我们的完整的网络是优化为优越的同步精度和质量使用两个不相交的鉴别器。 我们只在LRS2列车集上训练我们的模型[[1](#_bookmark28)]，批量大小为80。我们使用Adam优化器[[12](#_bookmark34)]，初始学习率为1e−4和贝塔斯β1= 0.5, β2=为0.999，同时用于发生器和视觉质量鉴别器D。注意，唇同步鉴别器没有进一步微调，因此它的权重被冻结。我们通过解释来总结对我们所提出的架构的描述

它在对真实视频的推理过程中是如何工作的。类似于LipGAN [[18](#_bookmark2)，该模型逐帧生成一个会说话的人脸视频。每个时间步长的视觉输入是当前的面部作物（从源帧），与相同的当前面部作物连接，下半部分掩蔽用作姿态优先。因此，在推理过程中，模型不需要改变姿态，从而显著地减少了伪影。相应的音频段也作为语音子网络的输入，网络生成输入面裁剪，但嘴区域变形。

我们所有的代码和模型都将公开发布。我们现在将定量地根据以前的模型来评估我们的新方法。

**4定量评价**

尽管只在LRS2训练集上进行了训练，但我们在3个不同的数据集上评估了我们的模型。但在这样做之前，我们重新研究了之前工作中遵循的当前评估框架，以及为什么它远不是评估这个领域工作的理想方法。

**4.1重新思考野外语音驱动唇同步的评估框架**

目前的与说话者无关的语言表达评估框架对模型的判断与假唱真实视频时的使用方式不同。具体地说，不是提供当前帧作为参考（如前一节所述），而是选择视频中的随机帧作为参考，以在评估期间不泄漏正确的嘴唇信息。我们强烈认为，前一段中的评估框架对于评估对口型的质量和准确性并不理想。在对上述评价系统的仔细研究中，我们观察到一些关键的限制，我们将在下面讨论。

4.1.1并不能反映现实世界的用法。如前所述，在测试时的生成过程中，模型不能改变姿态，因为生成的面需要无缝地粘贴到框架中。然而，当前的评估框架在输入中输入随机的参考系，因此要求网络改变姿态。因此，上述系统并不评估该模型在现实世界中如何使用。

4.1.2评价不一致。由于参考系是随机选择的，这意味着测试数据在不同的工作中是不一致的。这将导致不公平的比较，并阻碍结果的可重复性。

4.1.3不支持检查时间一致性。由于参考帧在每个时间步长都是随机选择的，所以当帧在随机姿态和尺度上生成时，时间一致性已经丢失。目前的框架不能支持一个新的指标或未来的方法来研究这个问题的时间一致性方面。

4..41.当前的指标并不是特定于对口型的。现有的指标，如SSIM [[27](#_bookmark43)]和PSNR，用于评估整体图像质量，而不是细粒度的唇同步错误。虽然LMD [[4](#_bookmark44)]关注嘴唇区域，我们发现嘴唇标记在生成的脸上可能相当不准确。因此，需要一个专门为测量对口型误差而设计的度量标准。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LRW [[8](#_bookmark22)] | | | LRS2 [[1](#_bookmark28)] | | |  | LRS3 [[3](#_bookmark29)] | | |
| **方法** | **LSE-D t** | **LSE-C ↑** | **FID t** | **LSE-D t** | **LSE-C ↑** | **FID t** | | **LSE-D t** | **LSE-C ↑** | **FID t** |
| Speech2Vid [[17](#_bookmark1)]  LipGAN [[18](#_bookmark2)]  **Wav2唇（我们的）Wav2唇+口（我们的）** | 13.14  10.05  **6.512**  6.774 | 1.762  3.350  **7.490**  7.263 | 11.15  2.833  3.189  **2.475** | 14.23  10.33  **6.386**  6.469 | 1.587  3.199  **7.789**  7.781 | 12.32 4 4.861.887  **4.446** |  | 13.97  10.65  **6.652**  6.986 | 1.681  3.193  **7.887**  7.574 | 11.91 4 4.732.844  **4.350** |
| 真实视频 | 7.012 | 6.931 | — | 6.736 | 7.838 | — |  | 6.956 | 7.592 | — |

**表1：我们提出了两个新的指标“唇同步误差-距离”（越低越好）和“唇同步误差-置信度”（越高越好），它们可以可靠地测量无约束视频中的唇同步精度。我们看到，使用Wav2Lip生成的视频的对口型精度几乎和真实的同步的视频一样好。请注意，我们只在LRS2上设置的火车上训练[[1](#_bookmark28)]，但我们轻松地概括了所有数据集，而没有进行任何进一步的微调。我们还报告了FID评分（越低越好），这清楚地表明，使用视觉质量鉴别器显著提高了质量。**

**4.2一个新的评估野外唇同步的基准和度量标准**

对随机帧进行采样进行评估的原因是，当前帧已经与语音同步，导致输入本身的唇形泄漏。而以前的工作并没有尝试采样不同的语音片段，而不是采样不同的帧，因为采样语音的地面-真实的嘴唇形状是不可用的。

4.2.1测量唇同步误差的度量。我们建议使用预先训练过的SyncNet [[9](#_bookmark10)]公开提供[3](#_bookmark46)测量生成的帧和随机选择的语音段之间的对口型误差。SyncNet在一个视频剪辑上的平均精度超过99% [[9](#_bookmark10)].因此，我们相信这可以是一个很好的自动评估方法，明确地测试准确的无约束视频nc。请注意，这不是我们在上面训练过的专家对口型鉴别器，而是钟和齐瑟曼发布的那个[[9](#_bookmark10)]，它是在一个不同的非公开数据集上进行训练的。使用SyncNet解决了现有评估框架的主要问题。我们不再需要采样随机的、时间上不相干的帧，而SyncNet在评估唇同步时也考虑了短期的时间一致性。因此，我们提出了使用SyncNet模型自动确定的两个新的指标。第一个是根据嘴唇和音频表示之间的距离计算出的平均误差度量，我们将其代号为“LSE-D”（“嘴唇同步误差-距离”）。较低的lsed表示较高的视听匹配，i。e., 说话和嘴唇的动作是同步的。第二个指标是平均置信度得分，我们将其代号为“LSE-C”(嘴唇同步错误

-信心）。置信度越高，视听相关性越好。较低的置信度得分表示视频中有几个部分的嘴唇运动完全不同步。进一步的细节可以在SyncNet的论文中找到[[9](#_bookmark10)].

4.2.2是评估野外唇同步的一致基准。现在我们有了一个自动的、可靠的度量，可以计算任何视频和音频对，我们可以在每个时间步长采样随机语音样本，而不是随机帧。因此，我们可以创建一个视频对的列表和一个伪随机选择的音频作为一个一致的测试集。我们创建了三个一致的基准测试集，每个都使用LRS2的测试集视频[[1](#_bookmark28)], LRW [[8](#_bookmark22)和LRS3 [[3](#_bookmark29)各自地对于每个视频vs，我们从另一个随机采样的视频中获取音频速度与长度的条件

3 github.com/joonson/syncnet\_python

他的演讲速度小于号对.我们使用LRS2创建了14K的音频-视频对。使用LRW测试集，我们创建了28K对，这个集合测量了额/近额视频的表现[[2](#_bookmark48)].

我们还使用LRS3测试集创建了14K对，这也将是轮廓视图中对口型的基准。完整的评估工具包将公开发布，以对野外对口型视频进行一致和可靠的基准测试。

**4.3比较在新的基准测试上的模型**

我们比较了前两种方法[[17](#_bookmark1),[18](#_bookmark2)使用LSE-D和LSE-C指标新创建的测试集。在推理过程中，我们现在为每个人提供相同的参考和姿态先验

时间步长，类似于之前在架构部分中描述的方式。LSE-D和LSE-C的平均得分见表[1](#_bookmark45)对于所有三个测试分割中的音频-视频对。此外，为了测量生成的人脸的质量，我们还报告了FrAlchet起始距离（FID）。我们的方法比以前的方法有很大的优势，表明强唇同步辨别的显著效果。我们还可以看到，在使用视觉质量鉴别器和对口型专家鉴别器后，质量的显著提高。然而，我们观察到在使用视觉质量描述后，同步精度略有下降

nator.因此，我们将发布这两个模型，因为它们在视觉质量和同步精度之间有轻微的权衡。

**4.4真实世界评价**

除了只对标准数据集进行评估外，我们新的评估框架和指标还允许我们评估现实世界中最有可能使用这些模型的视频。此外，考虑到人类对声唇同步的敏感性[[9](#_bookmark10)]，也有必要在人类评估者的帮助下评估我们的结果。因此，与之前关于扬声器独立对口型的工作相反，我们首次对网络上无约束的真实视频进行了定量和人工评估实验。因此，我们收集并公开发布了“重同步”一个“真实世界的评估数据集”，以主观和客观地测试假型作品的性能。

4.4.1曲线重新同步。我们所有的视频都是从YouTube上下载的。我们特别选择了三种类型的视频例子。第一种类型的“配音”包含音频自然不同步的视频，比如配音电影剪辑或公共地址被现场翻译成不同语言的视频(所以收件人的嘴唇

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **方法** | **视频类型** | **LSE-D t** | **LSE-C ↑** | **FID t** | | **同步Acc。** | **Visual质量。** | **整体经验** | **优选** |
| 未同步的Orig。视频语音2Vid[[17](#_bookmark1)] LipGAN [[18](#_bookmark2)]  **Wav2唇（我们的）Wav2唇+口（我们的）** | 被称为 | 12.63  14.76  10.61  **6.843**  7.318 | 0.896  1.121  2.857  **7.265**  6.851 | —  19.31 12.87 15.65  **11.84** |  | 0.21  1.14  2.98  **4.13**  4.08 | 4.81  0.93  3.91  3.87  **4.12** | 3.07  0.84  3.45  4.04  **4.13** | 3.15%  0.00%  2.35%  34.3%  **60.2%** |
| 没有嘴唇同步语音2Vid[[17](#_bookmark1)] LipGAN [[18](#_bookmark2)]  **Wav2唇（我们的）Wav2唇+口（我们的）** | 随机的 | 17.12  15.22  11.01  **6.691**  7.066 | 2.014  1.086  3.341  **8.220**  8.011 | —  19.98 14.60 14.47  **13.12** |  | 0.15  0.87  3.42  **4.24**  4.18 | 4.56  0.79  3.77  3.68  **4.05** | 2.98  0.73  3.57  4.01  **4.15** | 3.24%  0.00%  3.16%  29.1%  **64.5%** |
| 没有嘴唇同步语音2Vid[[17](#_bookmark1)] LipGAN [[18](#_bookmark2)]  **Wav2唇（我们的）Wav2唇+口（我们的）**  未翻译的视频 | 电报排字机 | 16.89  14.39  10.90  **6.659**  7.225  7.767 | 2.557  1.471  3.279  **8.126**  7.651  7.047 | —  17.96 11.91 12.77  **11.15**  — |  | 0.11  0.76  2.87  **3.98**  3.85  4.83 | 4.67  0.71  3.69  3.87  **4.13**  4.91 | 3.32  0.69  3.14  3.92  **4.05**  — | 8.32%  0.00%  1.64%  41.2%  **51.2%**  — |

**表2：使用我们新收集的ReSyncED基准测试进行的真实世界评估。我们使用定量指标和人工评估分数对三类真实视频进行评估。我们可以看到，在所有的情况下，Wav2Lip模型都可以产生高质量、准确的对口型视频。具体来说，这些指标表明，我们的对口型视频和真实的同步视频一样好。我们还注意到，人类的评估表明，当试图对口型同步TTS生成的语音时，有一个改进的空间。最后，值得注意的是，我们的唇型同步视频有超过90%的时间比现有的方法或实际的未同步视频更受欢迎。**

与翻译后的语音不同步)。第二种类型是

“随机”，我们有一个视频集合，我们创作

随机的视听对，类似于[4.2.2](#_bookmark47).第三种也是最后一种类型的视频，“TTS”，是专门为从文本到语音系统获得的合成语音的假唱性能进行基准测试的。这对于未来渴望自动翻译视频的作品至关重要(面对面翻译[[18](#_bookmark2)])或快速创建新的视频内容。我们手动转录文本，使用谷歌翻译（大约总共5种语言）和公开可用的文本到语音模型来为这类视频生成合成翻译语音。其任务是纠正原始视频中的嘴唇动作，以匹配这种合成语音。

4.4.2对重合成器的真实世界评估。我们首先评估生成的真实视频结果，使用我们的新的自动指标，“LSED”和“LSE-C”从SyncNet [[9](#_bookmark10)].人类评估，我们要求14评估者判断不同的同步版本的视频基于以下参数： (a)同步精度(b)视觉质量（评估视觉伪影的程度），(c)整体经验（评估视听内容的整体体验），和(d)偏好，观众选择视频的版本是最吸引人的观看。前三个参数的得分在1−5之间，而(d)是一个单个选择的投票，我们报告了由一个模型获得的选票的百分比。我们分别评估了这三类视频中的每一类，并在表中报告了我们的结果[2](#_bookmark49).一个值得注意的结果是，之前的工作[[17](#_bookmark1),[18](#_bookmark2)]，它产生了一些不同步的片段，比不同步的版本不太可取，因为后者仍然保持了良好的视觉质量。因此，我们的工作是第一个提供了一个显著的改进，比不同步的谈话人脸视频在野外。我们还在图中展示了一些定性的比较[3](#_bookmark50)其中包含了一些从ReSyncED测试集中生成的样本。

**4.5我们的专家鉴别器是其他选择中最好的吗？**

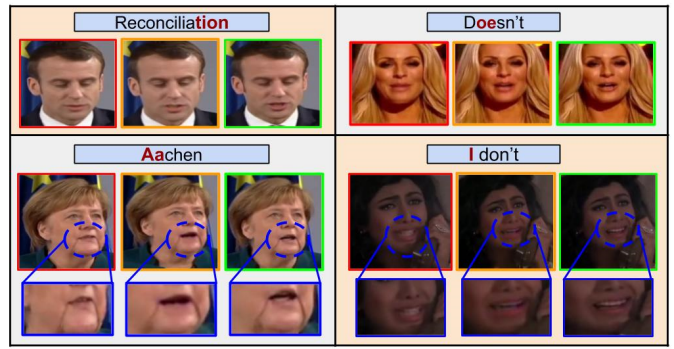
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 微调了吗？ | 不同步Acc。 | LSE-D | LSE-C |
| 电视机= 1 [[18](#_bookmark2)]我们的，电视机= 1 | ✓  × | 55.6%  79.3% | 10.33  8.583 | 3.19  4.845 |
| 我们的，电视机我们的=3，电视机= 3 | ✓  × | 72.3%  87.4% | 10.14  7.230 | 3.214  6.533 |
| 我们的，电视机**我们的=5，**电视机= 5 | ✓  × | 73.6%  **91.6%** | 9.953  **6.386** | 3.508  **7.789** |

**表3：一个更大的时间窗口允许更好的对口型识别。另一方面，在生成的面孔上训练唇同步识别器会降低其检测不同步的音频-唇对的能力。因此，使用这种鉴别器训练lipsync生成器会导致视频对口型不佳。**

我们的专家鉴别器使用电视机=5个视频帧来测量对口型误差。它也没有对中生成的面进行微调

一个GAN设置。我们在此消融术中证明了这两种设计选择

学习我们可以通过从LRS2测试集中随机采样非同步和非同步对来测试鉴别器的性能。我们改变了它的大小电视机通过=1、3、5，了解其对同步检测的影响。我们还微调/冻结的三种变体电视机同时训练Wav2Lip模型。因此，我们在表中总共得到了6个变体[3](#_bookmark26)由此我们可以清楚地做出两个观察结果。增加时间窗口的大小电视机始终提供更好的对口型辨别表现。更重要的是，我们看到，如果我们对生成的包含伪影的人脸进行微调的鉴别器，那么鉴别器就会失去其检测不同步的视听对的能力。我们认为，这是因为精细的鉴别器专注于生成过程中的视觉伪影



**图3：从我们提出的模型中生成的面孔（绿色和黄色轮廓）的例子。我们与目前的最佳方法进行了比较[[18](#_bookmark2)]（红色轮廓）。文本作为说明，表示在所示的框架中所说的话语。我们可以看到，我们的模型产生了准确的、自然的嘴唇形状。视觉质量鉴别器的加入也显著提高了视觉质量。我们强烈鼓励读者在我们的网站上查看演示视频。**

面孔的歧视，而不是细粒度的嘴唇对应。因此，它将真实的未同步的图像对归类为“不同步的”，因为这些真实的人脸图像不包含任何伪影。接下来，使用如此弱的鉴别器会导致我们的生成器的对口型惩罚不佳，导致对口型的说话面部视频不佳。

**5个应用程序和合理使用**

在一个我们的内容消费和社会社区-视听越来越视听，迫切需要大规模的视频翻译和创作。Wav2Lip可以在满足这些需求方面发挥重要作用，因为它在野外的视频是准确的。例如，通常用英语制作的在线讲座视频现在可以与其他本地的（自动）配音语音假唱

语言（表[2](#_bookmark49)，最后一个块)。我们也可以假唱配音电影，让它们看起来很愉快[2](#_bookmark49)，第一个块)。每天

在全球范围内，新闻发布会和公开演讲都有现场翻译，但收件人的嘴唇与翻译后的演讲不同步。我们的模型可以无缝地纠正这一点。自动动画CGI角色的嘴唇到语音演员的演讲，可以节省几个小时的手工努力，同时创建动画电影和丰富的对话游戏内容。我们在所有这些应用程序上演示我们的模型，并在我们网站上的演示视频中演示更多。

我们认为，讨论和促进合理使用日益有能力的假唱作品也是必要的。的广泛适用性

我们的模型对任何身份和声音都具有近乎现实的假唱能力，这引起了人们对潜在误用的担忧。因此，我们强烈建议，使用我们的代码和模型创建的任何结果必须明确地以合成的形式呈现。除了上面提到的强烈的积极影响之外，我们想要完全开源，我们的工作的意图是，它也可以同时鼓励努力[[11](#_bookmark51),[16](#_bookmark52),[24](#_bookmark53),[25](#_bookmark54)]来检测被操纵的视频内容及其被滥用。我们相信Wav2Lip可以使一些积极的应用，并鼓励富有成效的讨论以及关于合理使用合成内容的研究工作。

**6结论**

在这项工作中，我们提出了一种新的方法来生成准确的野外视频视频。我们已经强调了两个主要的原因，为什么目前的方法是不准确的，而对口型不受约束的谈话面部视频。基于此，我们认为一个预先训练过的、精确的唇型“专家”可以强制执行精确的、自然的嘴唇运动生成。在评估我们的模型之前，我们重新检查了当前的定量评估框架，并强调了几个主要问题。为了解决这些问题，我们提出了几个新的评估基准和度量标准，以及一个真实世界的评估集。我们相信，在这个新的框架下，我们可以可靠地判断未来的工作。我们的Wav2Lip模型在定量指标和人工评估方面都大大优于目前的方法。我们还在一个消融研究中调查了我们的鉴别器的设计选择背后的原因。我们鼓励读者在我们的网站上观看演示视频。我们相信我们的努力和想法

在这个问题中，可以导致新的方向，如综合表情和头部姿势以及准确的嘴唇运动。

**参考文献**

[1] T.Afouras, J.S.钟，A。高级，O。乙烯基和A。齐塞尔曼。2018.深

视听语音识别。在arXiv：1809.02108。

[2] T.Afouras, J.S.钟和A。齐塞尔曼。2018.对话：深度音频

视觉语音增强。在中间话。

[3]，琼儿子钟，安德鲁齐塞曼。2018.LRS3-TED：一个用于视觉语音识别的大型数据集。arXiv预印本arXiv：1809.00496（2018）。

陈[4]、李志恒、罗东、段志耀、徐辰良。2018.

嘴唇运动的一代，就在一瞥。在欧洲计算机视觉会议（ECCV）的论文集上。520–535.

陈[5]乐、马多克斯、段志耀、徐陈良。2019.分层跨模态的语音人脸生成。在IEEE关于计算机视觉和模式识别的会议的论文集上。7832–7841.

陈[6]、海地郑、马氏、段志耀、徐辰良。

2019.声音到视觉：层次跨模式说话人脸视频生成。在IEEE计算机学会计算机视觉和模式识别研讨会上。

[7]琼的儿子钟，阿米尔贾马鲁丁，和安德鲁齐塞尔曼。2017.你这么说？

arXiv预印本arXiv：1705.02966（2017）。

[8]琼的儿子钟和安德鲁·齐塞尔曼。2016.在野外读唇术。在亚洲

计算机视觉会议。施普林格，87-103。

[9]琼的儿子钟和安德鲁·齐塞尔曼。2016.超时：在野外自动对口型。在关于多视图唇部阅读的研讨会上，ACCV。

[10]马丁·库克，乔恩·巴克，斯图尔特·坎宁安，和徐绍。2006.一种用于语音感知和自动语音识别的视听语料库。《美国声学学会学报》120,5（2006），2421-2424。

[11]，布莱恩·多尔汉斯基，乔安娜·比顿，本·普弗拉姆，陆智国，拉斯·豪斯，梦林

王和克里斯蒂安·坎顿·费雷尔。2020.深度假检测挑战数据集。arXiv[:2006.07397](https://arxiv.org/abs/2006.07397)[cs.CV]

[12] John Duchi，Elad Hazan，和约兰的歌手。2011.在线学习和随机优化的自适应次梯度方法。机器学习研究杂志12,7（2011）。

[13] Ohad弗里德，阿尤什·特瓦里，迈克尔·佐尔霍弗，亚当·芬克尔斯坦，伊莱·谢克特-

男人，丹B戈德曼，凯尔热诺亚，泽宇金，基督教神堡，和阿格拉瓦拉。2019.基于文本的语音视频编辑。ACM交易

图形（TOG）38、4（2019年），1-14。

[14] Naomi Harte和约恩·吉伦。2015.TCD-TIMIT：一个视听语料库

连续话音IEEE多媒体学报17、5（2015），603-615。

何[15]凯明、张翔宇、任少卿、孙吉安。2016.用于图像识别的深度残差学习。在IEEE关于计算机视觉和模式识别的会议的论文集中。770–778.

徐志忠，庄义秀，李嘉妍。2020.深假图像

基于成对学习的检测。应用科学第10（2020年），370年。

[17]阿米尔贾马鲁丁，琼的儿子钟，和安德鲁齐塞曼。2019.你这么说？:

从音频中合成会说话的面孔。《国际计算机视觉杂志》127,11-12（2019），1767-1779。

[18] Prajwal，杰林，菲利普，阿比沙克，维奈南

布迪里和贾瓦哈尔。2019.朝向自动面对面的翻译。在第27届ACM多媒体国际会议会议记录上。ACM，1428-1436。

[19]里塞斯库马尔，何塞索特洛，昆丹库马尔，亚历山大德布雷比森，和约舒亚本吉奥。2017.奥巴马：逼真的对口型。arXiv预印本arXiv：1801.01442（2017）。

[20]安德鲁马斯，汉农和安德鲁吴。2013.整流器的非线性改善了神经网络的声学模型。在程序中。icml，卷。30.3.

[21] NPD.2016.52%的千禧一代智能手机用户使用他们的设备进行视频制作

打电话，根据NPD小组的说法。[..](https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use-their-device-for-video-calling-according-to-the-npd-group/)https://wwwnpdcom/wps/portal/npd/us/ [-2016年发布的52%](https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use-their-device-for-video-calling-according-to-the-npd-group/)news/pressofmillennialsmartphone [------到- - -](https://www.npd.com/wps/portal/npd/us/news/press-releases/2016/52-percent-of-millennial-smartphone-owners-use-their-device-for-video-calling-according-to-the-npd-group/)theirdeviceforvideocallingaccordingthenpd group/

[22]苏瓦角，史蒂文塞茨和艾拉凯梅尔马赫-什利泽曼。2017.

合成奥巴马：从音频中学习对口型。ACM图形交易（TOG）36、4（2017）、95。

[23] Justus·提伊斯，穆罕默德·埃尔加里布，阿尤什·特瓦里，克里斯蒂安·西奥巴特和马提亚斯

Nießner .2019.神经声音木偶法：音频驱动的面部再现。arXiv预印本arXiv：1912.05566（2019年）。

[24]鲁本·托洛萨纳，鲁本·维拉-罗德里格斯，朱利安·费雷斯，莫拉莱斯，和

哈维尔·奥尔特加加西亚。2020.《深度伪造与超越：对面部操纵和虚假检测的调查》。arXiv[:2001.00179](https://arxiv.org/abs/2001.00179)[cs.CV]

[25]埃莉诺特斯曼，玛丽莲乔治，塞卡马拉，和詹姆斯汤普金。2020.

走向不可信的社交视频验证，通过人脸几何形状的一致性来对抗深度造假。在IEEE/CVF计算机视觉和模式识别（CVPR）研讨会的论文集中。

[26]康斯坦丁诺斯维吉卡斯，斯塔夫罗斯佩特里迪斯，和马加恐慌。2019.现实的语言驱动面部动画与甘。《国际计算机视觉杂志》（2019年），第1-16页。

[28]杭州，刘玉、刘紫薇、罗平、王小刚。2018.会说话的脸

[27] Zhou Wang，Alan C Bovik，哈米德·谢赫，西蒙塞利等。2004.图像质量评估：从误差可见性到结构相似性。IEEE关于图像处理的事务13、4（2004），600-612。

通过横向分离的视听表示来生成。arXiv预印本arXiv：1807.07860（2018）。