“听音辨脸”的超能力，你想拥有吗？

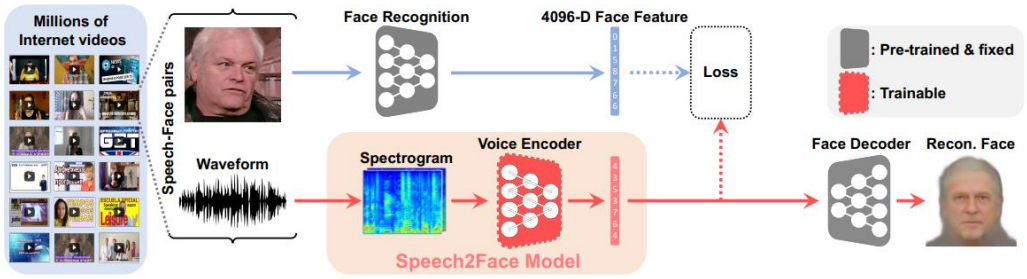
论文：Speech2Face: Learning the Face Behind a Voice（CVPR 2019,MIT）
项目地址：https://speech2face.github.io/

我们可以从一个人的说话方式推断出多少？在本文中,研究人员研究了从讲话人的简短录音中重建该人的面部图像的任务。他们设计并训练了一个深层的神经网络，使用来自 Internet / Youtube 的数百万人的自然视频来执行此任务。在训练过程中，模型学习视听和面部表情的相关性，从而使其产生可捕捉说话者各种身体属性（例如年龄，性别和种族）的图像。这是通过利用互联网视频中人脸和语音的自然共现以自我监督的方式完成的，而无需明确地对属性建模。直接从音频获得的重构揭示了脸部和声音之间的相关性。研究人员评估并以数字方式量化从音频中重建 Speech2Face 的方式如何以及以何种方式类似于扬声器的真实面部图像。

Speech2Face 模型：

自然面部图像中的面部表情，头部姿势，遮挡和照明条件的巨大差异，使 Speech2Face 模型的设计和训练变得不那么重要。例如,从输入语音退回到图像像素的直接方法不起作用；这样的模型必须学会排除数据中许多不相关的变化，并隐式提取人脸的有意义的内部表示，这本身就是一项艰巨的任务。

为了避开这些挑战，研究人员训练模型以回归到人脸的低维中间表示。利用 VGG-Face 模型（在大型人脸数据集上预先训练的人脸识别模型），并从网络的倒数第二层提取人脸特征。这些面部特征显示为包含足够的信息以重建相应的面部图像，并且具有一定的鲁棒性，模型整体框架如下。



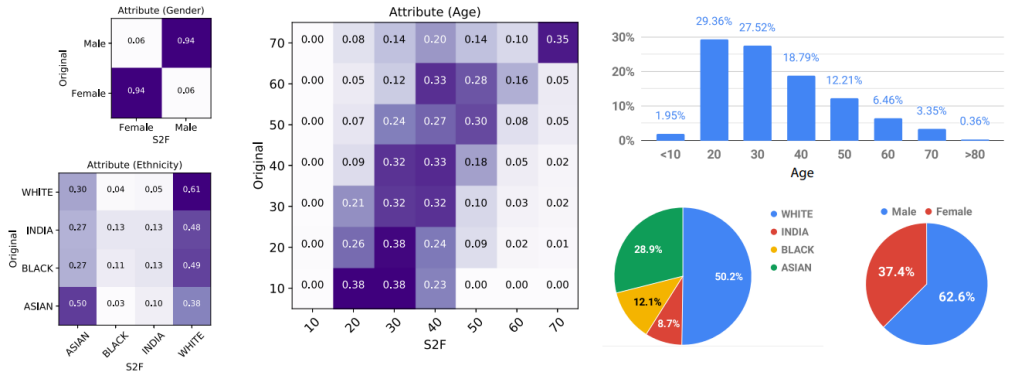
给这个网络输入一个复杂的声谱图，它将会输出 4096-D 面部特征，然后使用预训练的面部解码器将其还原成面部的标准图像。训练模块在图中用橙色部分标记。在训练过程中，Speech2Face 模型不会直接用人脸图像与原始图像进行对比，而是与原始图像的 4096-D 面部特征对比，省略了恢复面部图像的步骤。在训练完成后，模型在推理过程中才会使用面部解码器恢复人脸图像。训练过程使用的是 AVSpeech 数据集,它包含几百万个 YouTube 视频，超过 10 万个人物的语音-面部数据。在具体细节上，研究使用的中每个视频片段开头最多 6 秒钟的音频，并从中裁剪出人脸面部趋于，调整到 224×224 像素。

Speech2Face 管道包括两个主要组件：1）语音编码器，语音编码器模块是一个 CNN，它以语音的复杂声谱图作为输入，并预测将与相关联的脸部相对应的低维脸部特征；2）面部解码器，面部解码器的输入为低维面部特征，并以标准形式（正面和中性表情）产生面部图像。在训练过程中，人脸解码器是固定的，只训练预测人脸特征的语音编码器。语音编码器是作者自己设计和训练的模型，而面部解码器使用的是前人提出的模型。将实验结果更进一

步，Speech2Face 还能用于人脸检索。

研究结果：

Speech2Face 能较好地识别出性别，对白种人和亚洲人也能较好地分辨出来，另外对 30-40 岁和 70 岁的年龄段声音命中率稍微高一些。Speech2Face 似乎倾向将 30 岁以下的说话者年龄猜大，将 40-70 岁的说话者年龄猜小。除了比较基础的性别、年龄和种族，该模型甚至能猜中一些面部特征，比如说鼻子的结构、嘴唇的厚度和形状、咬合情况，以及大概的面部骨架。基本上输入的语音时间越长，预测的准确度会越高。但是该项研究的目的不是为了准确地还原说话者的模样，主要是为了研究语音跟相貌之间的相关性。在人口属性评估方面研究人员使用了 Face++，他们通过在原始图像和 Speech2Face 重建图像上运行 Face++分类器，评估并比较了年龄，性别和种族。此外，研究人员也从颅面属性（获取面部的比率和距离），特征相似度（直接测量预测特征与从说话者原始面部图像获得的真实特征之间的余弦距离）等方面进行比较。

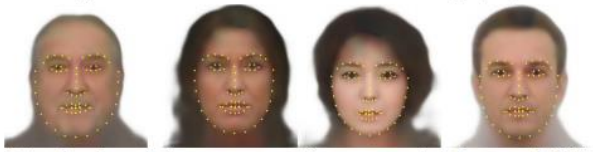


(a) Confusion matrices for the attributes

(b) AVSpeech dataset statistics

(a) 人口属性评估

(a) Landmarks marked on reconstructions from image (F2F)



(b) Landmarks marked on our corresponding reconstructions from speech (S2F)

Face measurement	Correlation	p-value
Upper lip height	0.16	$p < 0.001$
Lateral upper lip heights	0.26	$p < 0.001$
Jaw width	0.11	$p < 0.001$
Nose height	0.14	$p < 0.001$
Nose width	0.35	$p < 0.001$
Labio oral region	0.17	$p < 0.001$
Mandibular idx	0.20	$p < 0.001$
Intercanthal idx	0.21	$p < 0.001$
Nasal index	0.38	$p < 0.001$
Vermilion height idx	0.29	$p < 0.001$
Mouth face with idx	0.20	$p < 0.001$
Nose area	0.28	$p < 0.001$
Random baseline	0.02	–

(c) Pearson correlation coefficient

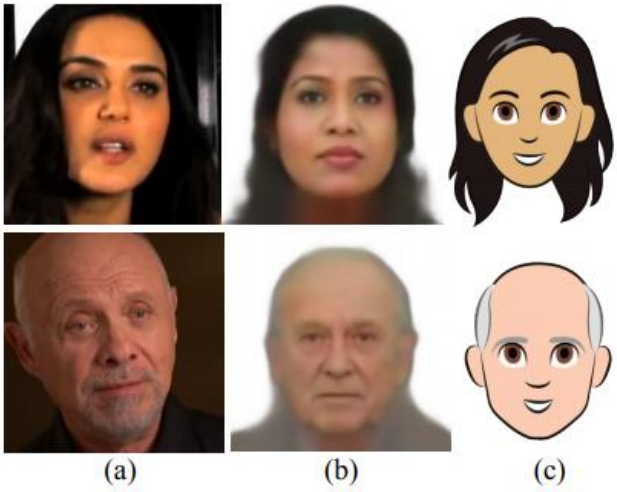
(b) 颅面属性

Length	cos (deg)	L_2	L_1
3 seconds	48.43 ± 6.01	0.19 ± 0.03	9.81 ± 1.74
6 seconds	45.75 ± 5.09	0.18 ± 0.02	9.42 ± 1.54

(c)特征相似度

研究应用：

如下图所示,研究人员从语音中重建的面部图像可用于从语音中生成说话者的个性化卡通形象。研究人员使用 Gboard(Android 手机上可用的键盘应用程序)，它还能够分析自拍图像以产生卡通版的脸。可以看出，Speech2Face 的重构能够很好地捕获面部特征，以使应用程序正常工作。



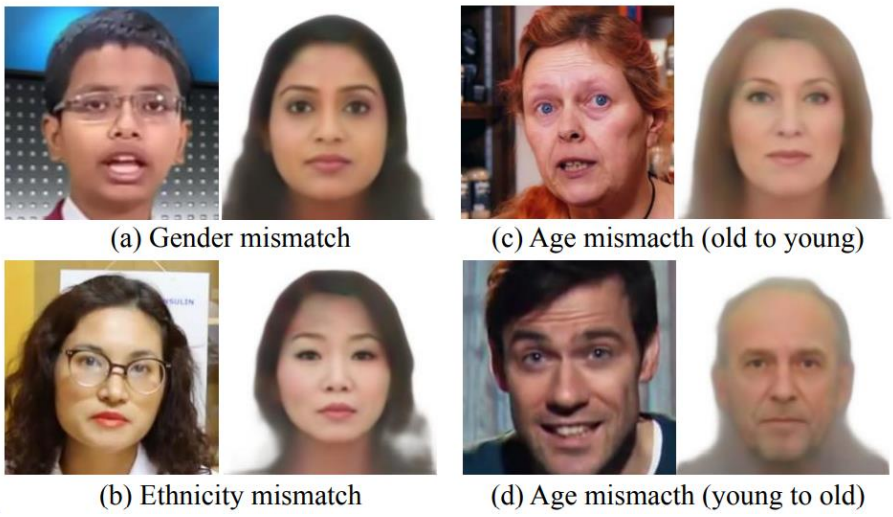
研究不足：

若根据语言来预测种族，那么一个人说不同的语言会导致不同的预测结果。研究人员让一个亚洲男性分别说英语和汉语，结果分别得到了 2 张不同的面孔。但是，模型有时候也能正确预测结果，比如让一个亚洲小女孩说英文，虽然恢复出的图像和本人有差距，但仍可以看出黄种人的面部特征。通常，观察到混合的行为，需要更彻底的检查以确定模型在多大程度上依赖语言。

An Asian male speaking in English (left) & Chinese (right)

An Asian girl speaking in English

除此以外，在其他的一些情况上，模型也会出错，比如：变声期之前的儿童，会导致模型误判性别发生错误；口音与种族特征不匹配；将老人识别为年轻人，或者是年轻人识别为老人。研究人员指出，Speech2Face 的局限性，部分原因来自数据集里的说话者本身种族多样性不够丰富，这也导致了它辨认黑种人声音的能力比较弱。



(a) Gender mismatch (c) Age mismacth (old to young)

(b) Ethnicity mismatch (d) Age mismacth (young to old)

麻省理工学院的研究人员在该项目的 GitHub 页面提出警告，承认该技术引发了关于隐私和歧视的问题。虽然这是纯粹的学术调查，但研究人员认为由于面部信息的潜在敏感性，在文章中明确讨论一套道德考虑因素很重要,对此进行任何进一步调查或实际使用都将会仔细测试，以确保训练数据能够代表预期的用户人群。

发布链接：https://mp.weixin.qq.com/s/2Otst8vOksYL8MNEafQDRw