于佳宁联想研究院
实习总结

技术辅导：刘博

实习目标

为解决数字虚拟人在说中文时口型表现不自然的问题，构建中文数据集，测试Wav2Lip模型在中文数据集上的性能指标并训练。

- 数据集构建
- 数据集的预处理
- 客观指标的测试
- 模型训练

问题

现存问题：在英文数据集LRS2上训练的Wav2Lip模型说中文表现不自然。主观表现为口型变化过快，口型与音频之间不够匹配。

可能的原因：英文音素数量大于中文（英文48个音素，中文32个音素），并且包含了更多的辅音音素，这些辅音需要较快的口型动作来产生。其次，相关研究也表明英文语速通常比中文快。

解决方法：尝试用中文数据集训练Wav2Lip模型。

用英文数据集训练的wav2lip模型生成的视频效果

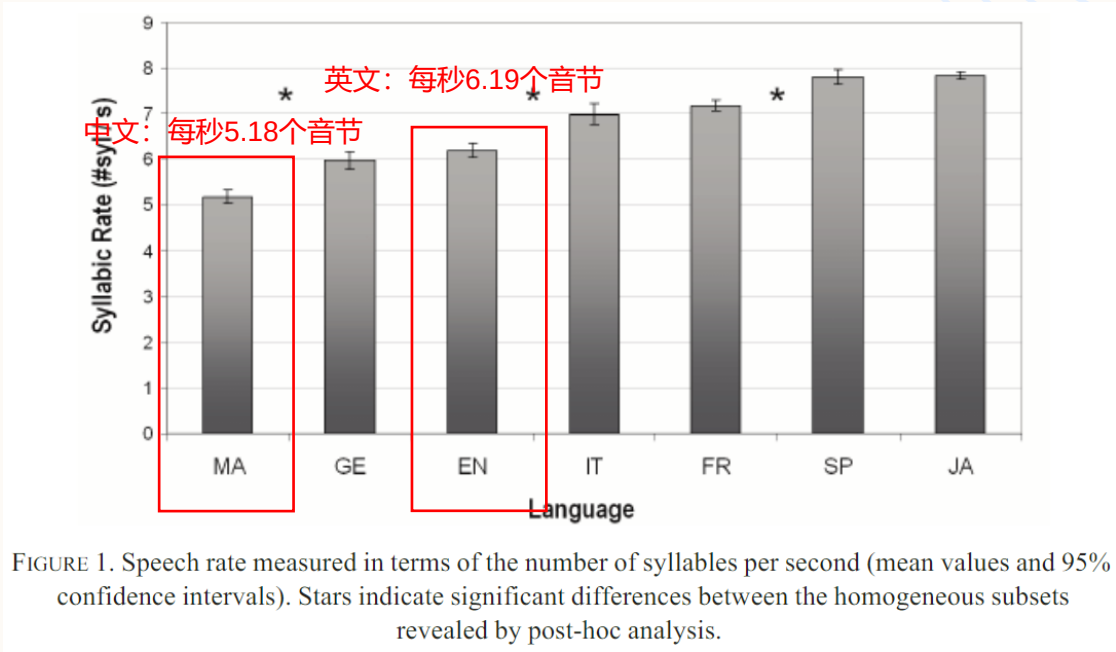


FIGURE 1. Speech rate measured in terms of the number of syllables per second (mean values and 95% confidence intervals). Stars indicate significant differences between the homogeneous subsets revealed by post-hoc analysis.

根据每秒音节数衡量的几种语言的语速对比

跨语言视角的语音信息率研究 里昂大学
Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. Language, 87(3), 539-558.

中文数据集构建方案

人工构建：优点是数据质量可控，缺点是周期长、成本高。

公开数据集：优点是成本低，问题是中文的高质量数据集数量少。

人工构建中文数据集

要求：

1. **说话视频**：有真人露脸说话
2. **面部清晰**：脸部无大面积遮挡
3. **多样性**：包含尽量多不同年龄段、性别、场景的人物
4. **分辨率**：视频分辨率≥720P
5. **帧率**：≥25FPS

流程：

1. **SyncNet 检测**：音视频同步性检测
    1. AV Offset: 音视频偏差大小，乘以40ms代表偏差的时间
    2. Min Dist: 音视频同源度，应该在7左右，否则不同源
    3. Confidence: 置信度，表示视频是否同源，应该在7左右
2. **ffmpeg 矫正**：比如延迟音频，插入空白画面等
3. **视频切分**：将视频切分为10秒左右的片段

收集数量：

访谈类13个，授课类8个，新闻类11个

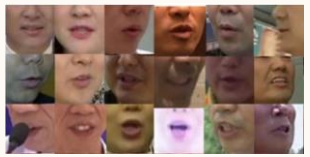
整理并收集公开的中文数据集

现有中文数据集主要分为两类：

1. **In the lab 数据集**：多为人工录制，背景单一，无法用到生产中
    - **CAVSR1.0**：人工录制，20人读78个汉字，样本数量少

    - **HIT Bi-CAV**：人工录制，10人基于96个音读出的200个常用汉语句子
    - **CMLR**：说话人只有11个，数据全部收集于新闻联播

2. **In the wild 数据集**：多为视频网站下载，说话者来自于多种真实环境，通用性更强
    - **LRW-1000**：无完整人脸，词汇级数据，非完整句子



    - **CN-CVS （2023.6）**

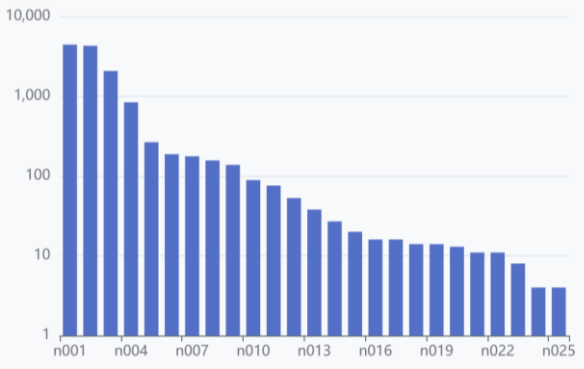
清华大学语音和语言技术中心
ICASSP 2023
国际声学、语音与信号处理会议

CN-CVS 数据集构成

新闻类

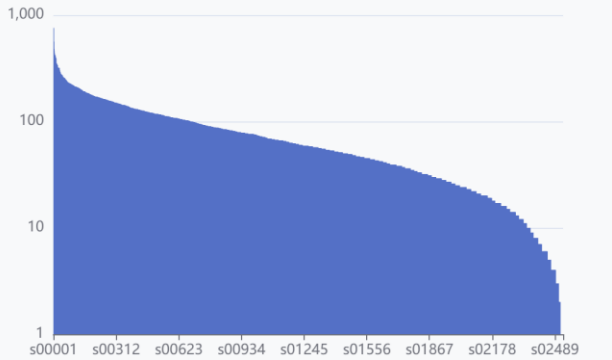
演讲类

新闻类每个说话人的语句数量



新闻类共有 28 个说话人和 13016 个语句，总时长近 35 小时，平均时长8.5秒

演讲类每个说话人的语句数量



演讲类共有来自 2529 个发言人的 193,245 个语句，总时长为 273 小时，平均时长3.7秒

数据整体和处理流程

2500+ 说话人
200000+ 语句
300+ 小时

片段平均长度为 6 秒。说话内容丰富，涉及新闻播报、演讲访谈等多个领域。包含来自不同职业和年龄段的说话人，且光照条件、视频角度也不同，保证了数据的多样性。

数据处理流程：

1. **镜头变化检测**：为避免在单个样本中出现镜头变化，使用ffmpeg检测镜头变化，并在检测到的位置剪切视频成片段
2. **人脸检测**：删除没有人脸或多个人脸的片段
3. **片段分割**：使用pydub工具检测静音帧，并根据检测结果将片段分割成短的小段。每个小段大致对应一个句子。
4. **同步性检测**：确保视频和音频同步。使用了SyncNet模型进行同步检测，并删除同步度较低的片段。

数据处理流程符合人工构建的要求。

数据来源：新闻30分，一席演讲，一刻TALKS等

CN-CVS 数据预处理

1. **提取视频帧**：使用ffmpeg从视频中提取帧并将其保存为JPG图像

   ‘ffmpeg -y -i {} -qscale:v 1 -qmin 1 -qmax 1 -vsync 0 {}%d.jpg’

   控制图像质量　　不进行视频同步 用于从帧率不固定的视频中提取帧

2. **提取人脸部分**：基于FAN-Face算法检测视频中的人脸框，从原始图像中提取对应的人脸图像并保存

3. **整合格式**：按照Wav2lip模型要求的数据格式，重新整理CN-CVS数据集

4. **生成图像总数量**：27713052 张（LRS2数据集261174张）

*.mp4
原视频

1.jpg　2.jpg　3.jpg　4.jpg
提取视频帧

1.jpg　2.jpg　3.jpg　4.jpg
提取人脸

```
cn-cvs
├── news
│       ├──n001-n028
│       │    ├── video
│       │    ├── audio
│       │    ├── txt
│       │    ├── landmark
│       │    ├── roi
│       │    └── n001.json
├── speech
│       ├── s00001-s02529
│       │    ├── video
│       │    ├── audio
│       │    └── ...
```

CN-CVS数据集原格式

```
lrs2_preprocessed
├── list of folders
│       ├── Folders with numbered video
│       │    ├── *.jpg
│       │    └── audio.wav
```

wav2lip模型使用的LRS2数据集格式

```
cn_cvs_processed
├── list of folders for each speaker n001-s02529
│       ├── videos for each speaker
│       │    ├── *.jpg
│       │    └── audio.wav
```

CN-CVS数据集处理后格式

测试

完整训练集：153323 (87%) 个，**验证集**：11253 (6%) 个，**测试集**11936 (7%) 个
完整测试集过程预计需要~100h，考虑时间因素，选用**小规模测试集**测试：3396个(耗时30h)

测试步骤：

1. **生成随机样本索引**：准备一个filelist.txt文件，将测试集的视频名称随机分成两列，第一列是在合成视频中输入的音频，第二列为视频，匹配过程中需要使用OpenCV库读取视频时长，并保证第二列视频的时长要大于第一列视频的时长。
2. **生成合成的视频**：使用Wav2Lip模型中gen_videos_from_filelist.py生成视频。生成的视频为原始视频的人物，但口型会变为新匹配的音频所对应的口型。
3. **测试**：使用合成视频和真实视频测试LSE-D，LSE-D，FID等指标。

原始视频

合成视频

测试指标：

使用**SyncNet模型**，用于衡量人的嘴唇动作和相应的语音之间的同步性

1. **LSE-C**：置信度，数值越高代表音频和视频越匹配
2. **LSE-D**：唇部表现和音频之间的距离的误差，数值越低代表越高的音频视频匹配度

PyTorch-FID库

3. **FID**：用于评估生成的图像质量。使用Inception网络对真实视频的图像和生成视频的图像进行特征提取，得到它们在特征空间中的分布。计算两个分布之间的Fréchet距离，衡量相似性。FID值越低，表示生成模型与真实数据之间的分布越接近，生成模型的质量越好。

	LSE-C	LSE-D	FID
LRS2	7.781	6.469	4.446
CN-CVS	3.591	8.258	3.270
指标差异	-2.878	-0.477	+1.176
CN-CVS news	5.411	7.598	2.607
CN-CVS speech	3.269	8.375	3.470

英文训练的Wav2Lip在中英文数据集上的指标对比

进展与成果

1. **数据集**：完成了人工构建中文数据集的标准流程，以及公开数据集的数据构建和预处理。
2. **模型训练**：Wav2Lip第一阶段的模型训练完成，同步损失指标达到了可用的范围。
3. **当前应用**：数据集用于会议分身模型训练，对数字虚拟人的主观视觉效果有一定程度的提升。
4. **后续应用**：可以使用纯中文数据集替换现在的英文数据集。

Wav2Lip是一个两阶段模型：

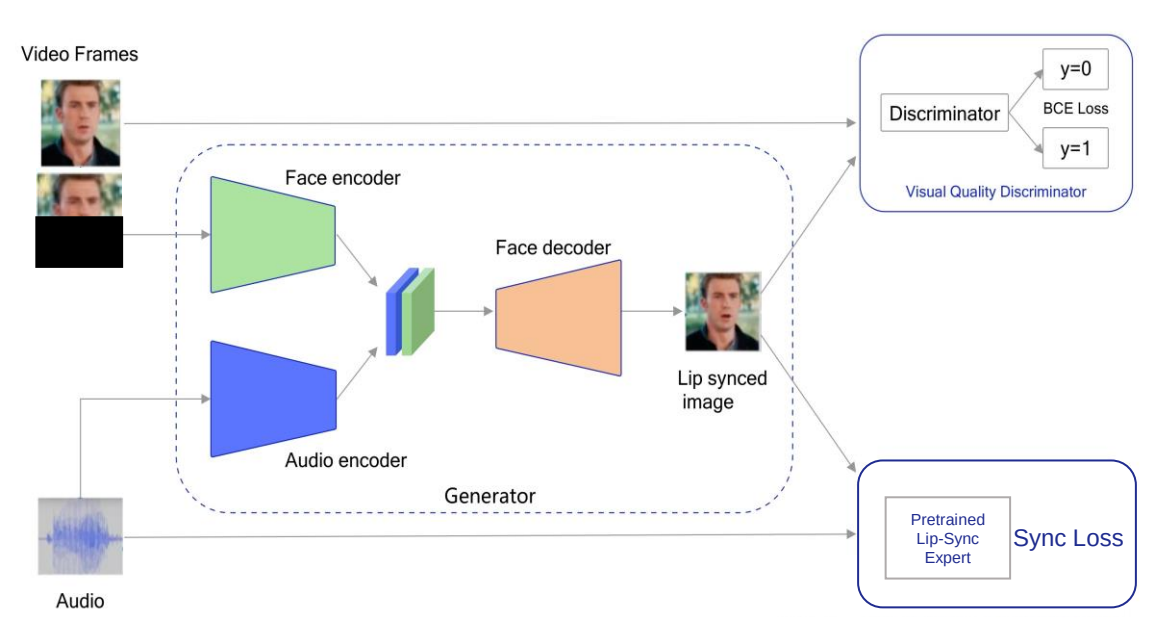
第一阶段：
用视频帧和音频训练一个音频和嘴型同步鉴别器SyncNet， 用于评估生成口型与音频的一致性和真实性。

第二阶段：
训练一个生成器Wav2Lip，生成口型。第二阶段的训练需要在第一阶段完成后开始。

训练过程：生成器模型的输入包含两部分（视频帧和音频），输出获得唇形和音频同步的图像帧。把原始视频帧和生成图像帧输入到视觉质量判别器中，表示是真实的还是生成的图片，进而提高图像质量。把生成图像帧和音频输入到预先训练好的唇形同步鉴别器中，判断唇形是否生成的精准。

训练指标：鉴别器在验证集的损失应降至约0.20-0.40，生成器在验证集的同步损失应降至约0.015-0.035。

目前进度：鉴别器的同步损失从0.79下降到了0.35（训练耗时约8天），达到可用范围，还需要进行生成器的训练。



实习总结

技术层面：

1. 学习了多种数据的处理方法和相关工具的使用
2. 操作系统相关技能
3. 加深了对深度学习模型的了解和应用

工作方法层面：

1. 对问题的及时沟通和反馈
2. 注意细节
3. 相关资料、学术成果的查找和阅读

个人感悟：

1. 代码能力和debug能力需要提升
2. 平时需要更多了解前沿的技术，多读相关论文