# Title：

Robust Speech Recognition via Large-Scale Weak Supervision

稳健的语音识别通过大规模的弱监督

# Abstract：

研究在网络上收集到的语音训练的模型的效果，在把训练数据集拓展到68万小时，对语言多任务训练时，结果非常好：在标准的测试任务上，它和在这些任务的训练数据集上充分训练好的模型相比，结果差不多，但是Whisper是zero-shot的训练（训练好的语音模型，部署到一个地方的时候就算是应该能直接用，而不是说需要额外收集一点数据，再做一次微调）。跟人类比，这个模型的精度和稳健度都差不多。

# Conclusion

Whisper指出，最近的语音识别领域，对把整个弱监督的预训练做大的事情认识不够，有了这个以后以后就不用做自监督或者自训练了。好处是在大的训练集上训练好就行了，不用做微调。

# Introduction：

在语音识别领域，最近的方法是无监督的预训练（比如Wav2Vec 2.0），核心思想是采集大的语音数据而不需要标号（采集这种数据比较容易，不需要太多的清理和不容易有版权问题）。大概意思是在没有标号的数据集上用自训练或者对比学习的方法训练一个比较好的encoder，能对语音数据做一个比较好的编码。因为这种方式收集数据比较简单，2021年Google的作者把这个数据集做到了1million hours的大小，远远超过了在学术界使用的标准数据集。然后使用在大的数据集上训练好的编码器用标准数据集微调，表现比只用标准数据集训练的模型要好（因为encoder已经足够好了，所以只用很少的数据做微调就行了）。

作者提出了一个问题：

这些预训练好的语音的编码器能够学习到高质量的语音的表示，但是他们只是用纯没有监督的方式训练好的，所以没有一个很好的解码器，所以要用的话还是要用带标号的数据进行微调，微调是一个复杂的过程，需要熟练工才能干，不用微调那就更好了。

为什么不用GPT解码器？（GPT是输入一段文字，预测下一个词是什么样子，一旦训练好可以直接拿来用，不需要微调）解码器自带预测功能，任何任务构造一个问答的模式（prompt），就能用GPT预测答案。

语音信号是一段声波，丢进GPT后它只能预测下一个微秒的声波的形状，但是预测声波对我们无用，因为对语音来讲，需要把词预测出来，语音信号转化成文本信号，但是没有标号的时候，只能预测声波，还是需要一个声波转文字的功能（需要标号）。这就是为什么做无监督训练好的任务需要微调，是因为在这个地方需要一个格式转换。

作者还提出：

在微调的时候，很容易对特定数据过拟合，降低泛化性。在ImageNet上训练的模型能够做到比人类好，但是在其他任务上不见得做的比人类好。

核心思想：

无监督的预训练确实把编码器改进了很多，但是却没有相对应的高质量的解码器会是一个巨大的缺陷，真正语音识别的目标是“out of box”，好的语音识别系统不应该是部署到新的环境就要重新训练一个解码器。

有监督学习的数据集和无监督学习的数据集量差的比较远，之前工作最多5k hours和1million hours差得比较远。作者觉得放松标号要求能达到更多的数据（在数量和质量之间做权衡）。于是作者把弱监督学习的数据拓展到了68万小时，多样性也保证了

# Approach

## Data Processing

跟之前做语音系统不一样的是直接预测原始文本（无标准化，纯粹靠seq2seq够强大），直接输出需要的文本，而不用做文本反向归一化的过程（单词变小写，去点标点符号，展开缩写形式等）。

数据集是在网上爬下来的，多样性很足但是文本质量的多样性不好，最好把质量不好的标号去除，为了做这个事情，他们做了一个自动过滤器：

filter：检测机器生成的文本并去除（机器生成的大都基于一些规则，有办法检测到）；然后是把声音的语言和文本的语言检测出来，如果文本是英语就行，那就把不匹配的删掉，它有任何语言到英语的翻译功能，只要文本是英语就行，还有去重。

然后把音频文件分成30s的区间，匹配上对应的文字就成了训练的样本。训练完成后又识别出错误的样本，再做一次过滤。最后就是把网上搜集的数据和用来测试的数据进行去重。

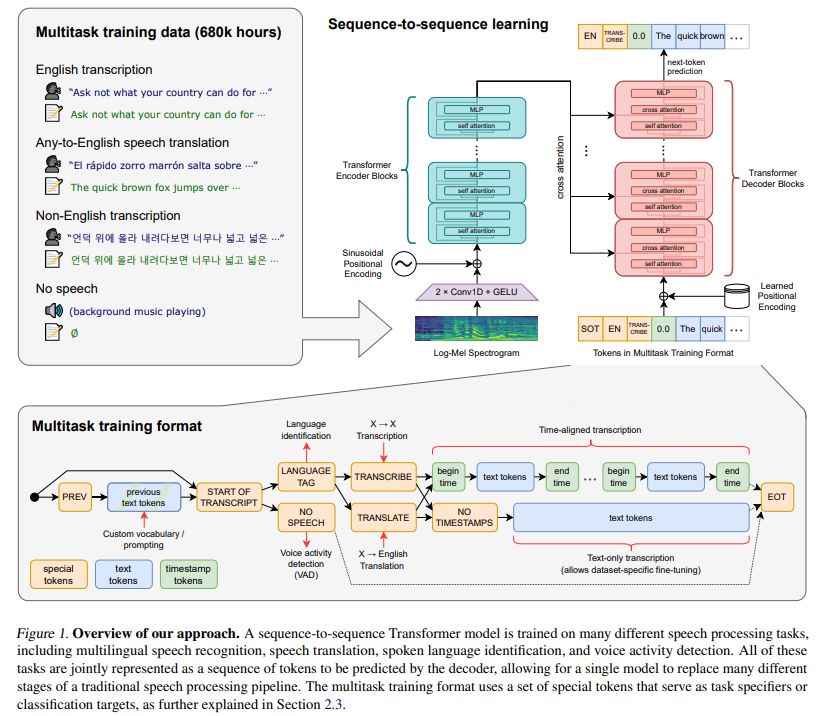
## Model

用了简单的off the shelf模型（原始代encoder-decoder的Transformer）。

声音如何输入？

音频采样到16000HZ，然后把它变成80通道的log-magnitude Mel spectrogram（用傅里叶变换把声音转化成频谱图，然后mel的意思是把低频做的好一点（因为人对低频声音敏感），然后把高频放在一起），每一个时间点抽取了80个维度的特征。每一个时间窗口使用的是25ms（大约600个点），做傅里叶之后通过mel做到80通道里面，长为30s的云信号最后变成了3000个数据点，每个数据点上的维度是80。最后把值变为-1到+1之间。

## Multitask Format



输入长度是3000，特征维度是80，首先进入到两个一维的卷积层里面，输出为80\*1500，然后输入Transformer的编码器解码器

有四个任务：english to english；any to english；其它语言的翻译；给一段音乐判定没人说话

通过prompt的结构区分这四个任务

虽然语音模型识别模型主要干的事情是给一个特定的语音片段把里面的词识别出来，但是对大部分的语言识别系统来讲还要干其他事情，比如是不是有人在说话（VAD：voice activity detection），是谁在说话，以及识别出来的文字最后要把它做成一个符合书写习惯的文字（大小写、标点符号），通常来讲这些事情是额外设计的，在本文作者希望通过一个单模型把所有的事情全部做好，对一个输入的语音可以做转录，翻译，VAD，时间戳，语言检测等等。

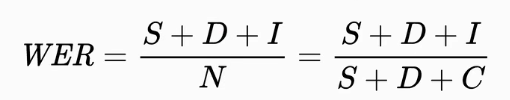
单模型做多个任务在实际中有很多问题：比如VAD，检测是不是有人说话，这样的模型通常来讲需要它运行的快一点（需要一个一直记录声音的麦克风，有个很小的模型一直在跑，功耗不能很高，一旦喊他，他就触发大的模型跑任务），如果在大的模型例做这个事情的话开销就大了；第二是每个任务的能力参差不齐，比一定能达到都很好的水平

# Experiments

## Zero-shot Evaluation

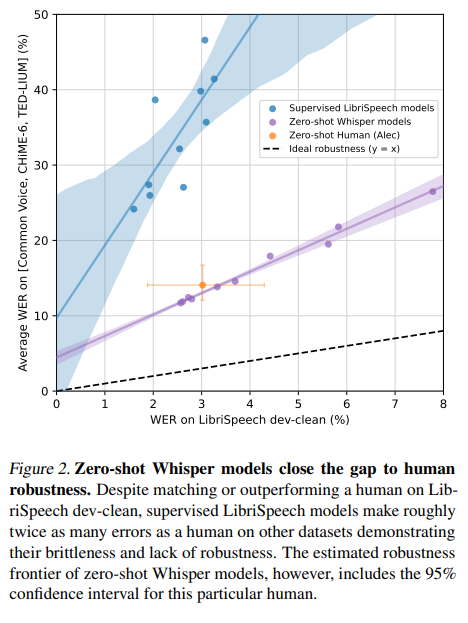
## Evaluation Metrics

word error rate (WER)

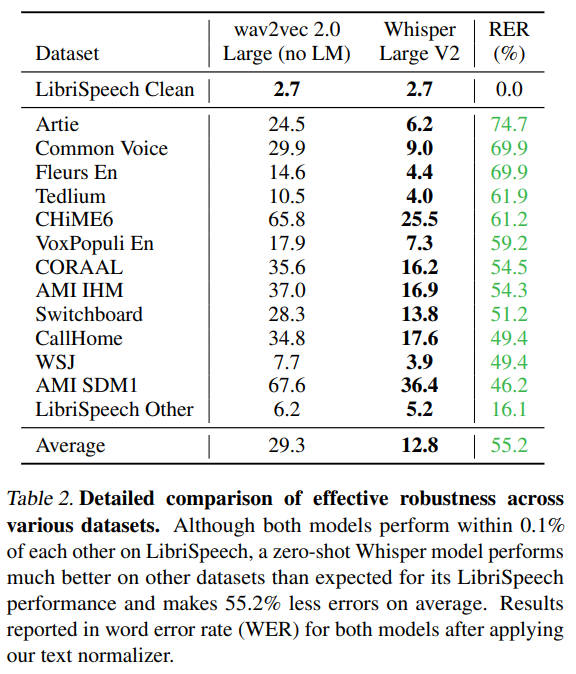


S为替换，D为删除，I是插入

## English Speech Recognition



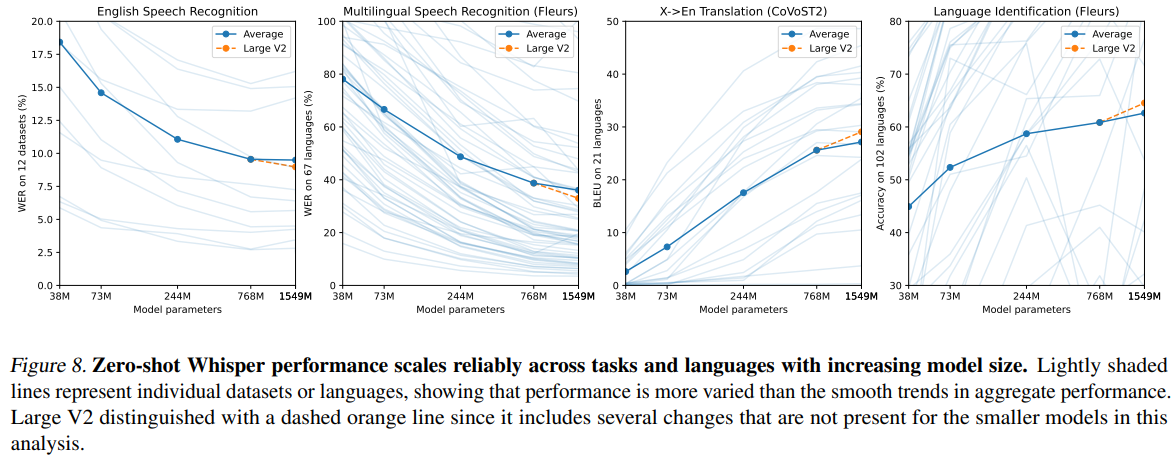
左图表现了Whisper的泛化能力比较好



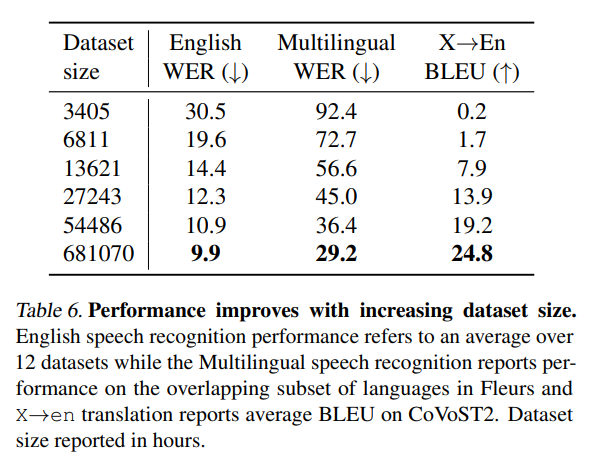
在一系列数据集上的错误率

# Analysis and Ablations

## Model Scaling

但模型变大的时候，性能还是有网上提升的，但是有饱和的迹象

## Dataset Scaling



# Limitations and Future Work

## Improved decoding strategies.

当模型变大的时候对于词的选择更好一些，可以根据上下文找出最好的词，但是对于比较长的语音转录还是有很多噪音，debug比较难