# Speech Recognition Is Not Solved 语音识别未解决

Posted on October 11, 2017  
2017年10月11日发布

Ever since Deep Learning hit the scene in speech recognition, word error rates have fallen dramatically. But despite articles you may have read, we still don’t have human-level speech recognition. Speech recognizers have many failure modes. Acknowledging these and taking steps towards solving them is critical to progress. It’s the only way to go from ASR which works for some people, most of the time to ASR which works for all people, all of the time.  
自从深度学习在语音识别领域大行其道以来，单词错误率急剧下降。但是，尽管你可能读过文章，我们仍然没有人类级别的语音识别。语音识别器有很多故障模式。承认这些并采取措施解决这些问题对取得进展至关重要。这是从对某些人有效的ASR到对所有人有效的ASR的唯一途径。

Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.  
提高交换机会话语音识别基准上的字误码率。测试集是2000年收集的。它由两个随机的英语母语者之间的40个电话对话组成。

Saying we’ve achieved human-level in conversational speech recognition based just on Switchboard results is like saying an autonomous car drives as well as a human after testing it in one town on a sunny day without traffic. The recent improvements on conversational speech are astounding. But, the claims about human-level performance are too broad. Below are a few of the areas that still need improvement.  
说我们已经在基于交换机结果的会话语音识别方面达到了人类水平，就像说在没有交通的晴天在一个城镇测试了自动驾驶汽车和人类一样。最近会话能力的提高令人震惊。但是，关于人的水平表现的说法过于宽泛。以下是一些仍需改进的方面。

## Accents and Noise 口音和噪音

One of the most visible deficiencies in speech recognition is dealing with accents and background noise. The straightforward reason is that most of the training data consists of American accented English with high signal-to-noise ratios. For example, the Switchboard conversational training and test sets only have native English speakers (mostly American) with little background noise.  
语音识别中最明显的缺陷之一是处理重音和背景噪声。最直接的原因是大部分的训练数据都是由美国口音的英语构成的，具有很高的信噪比。例如，总机会话训练和测试设备只有以英语为母语的人（大部分是美国人），背景噪音很小。

But, more training data likely won’t solve this problem on its own. There are a lot of languages many of which have a lot of dialects and accents. It’s not feasible to collect enough annotated data for all cases. Building a high quality speech recognizer just for American accented English needs upwards of 5 thousand hours of transcribed audio.  
但是，更多的训练数据可能无法单独解决这个问题。有很多语言，其中许多有很多方言和口音。为所有情况收集足够的注释数据是不可行的。为美国口音的英语建立一个高质量的语音识别器需要5000小时以上的转录音频。

Comparison of human transcribers to Baidu’s Deep Speech 2 model on various types of speech. Notice the humans are worse at transcribing the non-American accents. This is probably due to an American bias in the transcriber pool. I would expect transcribers native to a given region to have much lower error rates for that region’s accents.  
比较人类转录者与百度深度语音2模式对各类语音的影响注意，人类在转录非美国口音方面更差。这可能是由于美国在转录池的偏见。我希望某个地区的本地转录者对该地区口音的错误率要低得多。

With background noise, it’s not uncommon for the SNR in a moving car to be as low as -5dB. People don’t have much trouble understanding one another in these environments. Speech recognizers, on the other hand, degrade more rapidly with noise. In the figure above we see the gap between the human and the model error rates increase dramatically from the low SNR to the high SNR audio.  
在背景噪声下，运动车辆的信噪比低至-5dB并不少见。在这样的环境中，人们相互理解并没有太大的困难。另一方面，语音识别器在噪声的作用下退化得更快。在上图中，我们看到从低信噪比音频到高信噪比音频之间的人与模型错误率之间的差距显著增加。

## Semantic Errors 语义错误

Often the word error rate is not the actual objective in a speech recognition system. What we care about is the semantic error rate. That’s the fraction of utterances in which we misinterpret the meaning.  
在语音识别系统中，错误率往往不是实际的目标。我们关心的是语义错误率。这是我们曲解意思的话语的一部分。

An example of a semantic error is if someone said “let’s meet up Tuesday” but the speech recognizer predicted “let’s meet up today”. We can also have word errors without semantic errors. If the speech recognizer dropped the “up” and predicted “let’s meet Tuesday” the semantics of the utterance are unchanged.  
语义错误的一个例子是，如果有人说“让我们周二见面”，但是语音识别器预测“让我们今天见面”。我们也可以有没有语义错误的单词错误。如果语音识别器去掉“up”并预测“让我们星期二见面”，则该语句的语义不变。

We have to be careful when using the word error rate as a proxy. Let me give a worst-case example to show why. A WER of 5% roughly corresponds to 1 missed word for every 20. If each sentence has 20 words (about average for English), the sentence error rate could be as high as 100%. Hopefully the mistaken words don’t change the semantic meaning of the sentences. Otherwise the recognizer could misinterpret every sentence even with a 5% WER.  
当使用单词错误率作为代理时，我们必须小心。让我举一个最坏的例子来说明原因。5%的WER大致相当于每20个单词中有1个漏掉。如果每个句子有20个单词（英语大约是平均的），句子错误率可以高达100%。希望错误的单词不会改变句子的语义。否则，识别器可能会曲解每个句子，即使只有5%的答案。

When comparing models to humans, it’s important to check the nature of the mistakes and not just look at the WER as a conclusive number. In my own experience, human transcribers tend to make fewer and less drastic semantic errors than speech recognizers.  
在将模型与人类进行比较时，重要的是要检查错误的性质，而不是仅仅将WER看作一个决定性的数字。根据我自己的经验，人类转录者比语音识别者容易犯更少和更少的严重语义错误。

Researchers at Microsoft recently compared mistakes made by humans and their human-level speech recognizer. One discrepancy they found was that the model confuses “uh” with “uh huh” much more frequently than humans. The two terms have very different semantics: “uh” is just filler whereas “uh huh” is a backchannel acknowledgement. The model and humans also made a lot of the same types of mistakes.  
微软的研究人员最近比较了人类和他们的人类级语音识别器所犯的错误。他们发现的一个差异是，这个模型比人类更容易混淆“嗯”和“嗯哼”这两个术语有着非常不同的语义：“嗯”只是填充词，而“嗯哼”是一个反向确认。模型和人类也犯了很多相同类型的错误。

## Single-channel, Multi-speaker 单声道，多扬声器

The Switchboard conversational task is also easier because each speaker is recorded with a separate microphone. There’s no overlap of multiple speakers in the same audio stream. Humans on the other hand can understand multiple speakers sometimes talking at the same time.  
交换机会话任务也更容易，因为每个扬声器都用单独的麦克风录制。同一音频流中的多个扬声器没有重叠。另一方面，人类可以理解多个说话者有时同时说话。

A good conversational speech recognizer must be able to segment the audio based on who is speaking (diarisation). It should also be able to make sense of audio with overlapping speakers (source separation). This should be doable without needing a microphone close to the mouth of each speaker, so that conversational speech can work well in arbitrary locations.  
一个好的会话语音识别器必须能够根据说话的人来分割音频（分录）。它还应该能够理解与重叠扬声器（源分离）的音频。这应该是可行的，而不需要一个麦克风接近每一个发言者的嘴，这样，会话语音可以在任何地点工作良好。

## Domain Variation 域变化

Accents and background noise are just two factors a speech recognizer needs to be robust to. Here are a few more:  
口音和背景噪声只是语音识别器需要鲁棒性的两个因素。这里还有一些：

* Reverberation from varying the acoustic environment.  
  不同声环境产生的混响。
* Artefacts from the hardware.  
  五金制品。
* The codec used for the audio and compression artefacts.  
  用于音频和压缩人工制品的编解码器。
* The sample rate.  
  采样率。
* The age of the speaker.  
  演讲者的年龄。

Most people wouldn’t even notice the difference between an mp3 and a plain wav file. Before we claim human-level performance, speech recognizers need to be robust to these sources of variability as well.  
大多数人甚至不会注意到mp3和普通wav文件之间的区别。在我们声明人类级别的性能之前，语音识别器也需要对这些变异源具有鲁棒性。

## Context 上下文

You’ll notice the human-level error rate on benchmarks like Switchboard is actually quite high. If you were conversing with a friend and they misinterpreted 1 of every 20 words, you’d have a tough time communicating.  
你会注意到像Switchboard这样的基准测试的人为错误率实际上相当高。如果你和一个朋友聊天时，他们每20个字就曲解一个，你就很难交流了。

One reason for this is that the evaluation is done context-free. In real life we use many other cues to help us understand what someone is saying. Some examples of context that people use but speech recognizers don’t include:  
其中一个原因是评估是在上下文无关的情况下完成的。在现实生活中，我们使用许多其他的线索来帮助我们理解别人在说什么。一些人们使用但语音识别器不包括的上下文示例：

* The history of the conversation and the topic being discussed.  
  谈话的历史和正在讨论的话题。
* Visual cues of the person speaking including facial expressions and lip movement.  
  说话人的视觉暗示，包括面部表情和嘴唇运动。
* Prior knowledge about the person we are speaking with.  
  关于我们谈话对象的先验知识。

Currently, Android’s speech recognizer has knowledge of your contact list so it can recognize your friends’ names. The voice search in maps products uses geolocation to narrow down the possible points-of-interest you might be asking to navigate to.  
目前，Android的语音识别器知道你的联系人列表，因此它可以识别你朋友的名字。地图产品中的语音搜索使用地理位置来缩小可能需要导航到的兴趣点。

The accuracy of ASR systems definitely improves when incorporating this type of signal. But, we’ve just begun to scratch the surface on the type of context we can include and how it’s used.  
当加入这类信号时，ASR系统的精度明显提高。但是，我们刚刚开始对我们可以包含的上下文类型以及它的使用方式进行初步研究。

## Deployment 部署

The recent improvements in conversational speech are not deployable. When thinking about what makes a new speech algorithm deployable, it’s helpful to think in terms of latency and compute. The two are related, as algorithms which increase compute tend to increase latency. But for simplicity I’ll discuss each separately.  
最近在会话语音方面的改进是不可部署的。当考虑什么使新的语音算法可以部署时，从延迟和计算的角度考虑是有帮助的。两者是相关的，因为增加计算量的算法往往会增加延迟。但为了简单起见，我将分别讨论每一个问题。

Latency: With latency, I mean the time from when the user is done speaking to when the transcription is complete. Low latency is a common product constraint in ASR. It can significantly impact the user experience. Latency requirements in the tens of milliseconds aren’t uncommon for ASR systems. While this may sound extreme, remember that producing the transcript is usually the first step in a series of expensive computations. For example in voice search the actual web-scale search has to be done after the speech recognition.  
潜伏期：有潜伏期，我指的是从用户完成通话到转录完成的时间。低延迟是ASR中常见的产品约束。它会显著影响用户体验。数十毫秒的延迟要求在ASR系统中并不少见。虽然这听起来有点极端，但请记住，在一系列昂贵的计算中，生成文本通常是第一步。例如在语音搜索中，实际的网络规模搜索必须在语音识别之后进行。

Bidirectional recurrent layers are a good example of a latency killing improvement. All the recent state-of-the-art results in conversational speech use them. The problem is we can’t compute anything after the first bidirectional layer until the user is done speaking. So the latency scales with the length of the utterance.  
双向递归层是延迟消除改进的一个很好的例子。所有最新的最先进的会话结果都使用它们。问题是，在用户说完之前，我们无法计算第一个双向层之后的任何内容。所以潜伏期随着话语的长度而变化。

Left: With a forward only recurrence we can start computing the transcription immediately. Right: With a bidirectional recurrence we have to wait until all the speech arrives before beginning to compute the transcription.  
左：只有向前重复，我们可以立即开始计算转录。右图：在双向重复的情况下，我们必须等到所有的语音到达后才开始计算转录。

A good way to efficiently incorporate future information in speech recognition is still an open problem.  
如何在语音识别中有效地融合未来的信息仍然是一个有待解决的问题。

Compute: The amount of computational power needed to transcribe an utterance is an economic constraint. We have to consider the bang-for-buck of every accuracy improvement to a speech recognizer. If an improvement doesn’t meet an economical threshold, then it can’t be deployed.  
计算：转录话语所需的计算能力是一个经济限制。对于语音识别器的每一次精度改进，我们都必须考虑到一劳永逸。如果改进不符合经济阈值，则无法部署。

A classic example of a consistent improvement that never gets deployed is an ensemble. The 1% or 2% error reduction is rarely worth the 2-8x increase in compute. Modern RNN language models are also usually in this category since they are very expensive to use in a beam search; though I expect this will change in the future.  
一个从未部署的一致性改进的经典例子是集成。减少1%或2%的错误很少值得增加2-8倍的计算。现代RNN语言模型通常也属于这一类，因为它们在波束搜索中的使用非常昂贵；不过我希望将来会有所改变。

As a caveat, I’m not suggesting research which improves accuracy at great computational cost isn’t useful. We’ve seen the pattern of “first slow but accurate, then fast” work well before. The point is just that until an improvement is sufficiently fast, it’s not usable.  
作为一个警告，我并不是说以巨大的计算成本来提高精确度的研究是没有用的。我们早就看到过“先慢后准，后快”的模式。关键是，除非改进足够快，否则它是不可用的。

## The Next Five Years 未来五年

There are still many open and challenging problems in speech recognition. These include:  
在语音识别领域，还存在着许多开放性和挑战性的问题。其中包括：

* Broadening the capabilities to new domains, accents and far-field, low SNR speech.  
  扩展能力到新的领域，口音和远场，低信噪比语音。
* Incorporating more context into the recognition process.  
  在识别过程中加入更多的上下文。
* Diarisation and source-separation.  
  重氮化和源分离。
* Semantic error rates and innovative methods for evaluating recognizers.  
  语义错误率和评估识别器的创新方法。
* Super low-latency and efficient inference.  
  网络错误

I look forward to the next five years of progress on these and other fronts.  
网络错误

### Acknowledgements 网络错误

Thanks to [@mrhannun](https://twitter.com/mrhannun) for useful feedback and edits.  
网络错误

### Edit 网络错误

Hacker News .  
网络错误

### Footnotes 网络错误

1. Just ask anyone with a .   
   网络错误
2. These results are from . The accented speech comes from . The noise-free and noisy speech comes from the third challenge.   
   网络错误
3. [Stolcke and Droppo, 2017](https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/paper-revised2.pdf) [↩](https://awni.github.io/speech-recognition/?from=hackcv&hmsr=hackcv.com&utm_medium=hackcv.com&utm_source=hackcv.com#fnref:human_comparison)
4. See for an example of how to improve contact name recognition.   
   网络错误
5. See for an example of how to incorporate speaker location.   
   网络错误

语音识别未解决发布于2017年10月11日自深度学习语音识别出现以来，语言错误率急剧下降。但是，尽管您可能已阅读过文章，但我们仍然没有人类语音识别。语音识别器具有许多故障模式。承认这些并采取措施解决这些问题对于取得进展至关重要。这是从ASR出发的唯一途径，适用于某些人，大部分时间都是ASR，适用于所有人。 在Switchboard会话语音识别基准测试中，随着时间的推移，字错误率有所提高。该测试集于2000年收集。它包括两个随机母语为英语的人之间的40次电话交谈。 说我们在交换机结果的基础上实现了人类对话语音识别，就像在一个没有交通的晴天在一个城镇测试一辆自动驾驶汽车和人类一样。最近对话语言的改进令人震惊。但是，关于人类表现的说法过于宽泛。以下是一些仍需改进的领域。 口音和噪音语音识别中最明显的缺陷之一是处理重音1和背景噪音。直截了当的原因是大多数训练数据由具有高信噪比的美式口音英语组成。例如，Switchboard会话训练和测试集只有母语为英语的人（大多数是美国人），几乎没有背景噪音。 但是，更多的培训数据可能无法单独解决这个问题。有很多语言，其中很多都有很多方言和口音。为所有案例收集足够的注释数据是不可行的。为美国口音英语构建一个高质量的语音识别器需要超过5000小时的转录音频。 人类抄写员与百度深语言2模型在不同类型语音上的比较.2请注意，人类在抄写非美国口音方面更糟糕。这可能是由于美国人对抄录员库的偏见。我希望某个地区的抄写员对该地区的口音的错误率要低得多。 对于背景噪声，移动车辆的SNR低至-5dB并不罕见。人们在这些环境中彼此了解并不会有太多麻烦。另一方面，语音识别器随着噪声而更快地降级。在上图中，我们看到人类和模型错误率之间的差距从低SNR到高SNR音频急剧增加。 语义错误通常，单词错误率不是语音识别系统中的实际目标。我们关心的是语义错误率。这是我们误解其含义的话语的一小部分。 语义错误的一个例子是，如果有人说“让我们在星期二见面”，但语音识别器预测“让我们今天见面”。我们也可以出现没有语义错误的单词错误。如果语音识别器掉落“向上”并预测“让我们见周二”话语的语义不变。 使用单词错误率作为代理时，我们必须要小心。让我举一个最坏的例子来说明原因。 5％的WER大致对应于每20个1个遗漏的单词。如果每个句子有20个单词（大约是英语的平均值），则句子错误率可能高达100％。希望错误的词语不会改变句子的语义。否则识别器可能会误解每个句子，即使是5％的WER。 在将模型与人类进行比较时，重要的是要检查错误的性质，而不仅仅是将WER看作结论性数字。根据我自己的经验，人类抄写员倾向于比语音识别器产生更少且更少激烈的语义错误。 微软的研究人员最近比较了人类和他们的人类语音识别器所犯的错误。他们发现的一个差异是模型比人类更频繁地混淆“呃”和“嗯嗯”。这两个术语具有非常不同的语义：“呃”只是填充物，而“嗯嗯”是反向信道确认。模特和人类也犯了很多相同类型的错误。 单声道，多扬声器交换机会话任务也更容易，因为每个扬声器都使用单独的麦克风录制。同一音频流中的多个扬声器没有重叠。另一方面，人类可以理解多个发言者有时同时说话。 一个好的会话语音识别器必须能够根据讲话的人（diarisation）来分割音频。它还应该能够理解重叠扬声器的音频（源分离）。这应该是可行的，而不需要麦克风靠近每个扬声器的嘴，这样会话语音可以在任意位置很好地工作。 域变化口音和背景噪声只是语音识别器需要健壮的两个因素。这里还有一些：改变声学环境的混响。 来自硬件的人工制品。 用于音频和压缩工件的编解码器。 采样率。 发言者的年龄。 大多数人甚至都不会注意到mp3和普通wav文件之间的区别。在我们声称人类级别的表现之前，语音识别器也需要对这些可变性来源具有鲁棒性。 上下文您会注意到Switchboard等基准测试中的人为错误率实际上非常高。如果你和朋友交谈并且他们误解了每20个单词中的一个，你就很难沟通。 其中一个原因是评估是在无环境的情况下完成的。在现实生活中，我们使用许多其他线索来帮助我们理解某人所说的话。人们使用但语音识别器的上下文的一些示例不包括：会话的历史和正在讨论的主题。 说话的人的视觉线索包括面部表情和嘴唇运动。 关于我们与之交谈的人的先验知识。 目前，Android的语音识别器具有您的联系人列表的知识，因此它可以识别您朋友的名字.4地图产品中的语音搜索使用地理位置来缩小您可能要求导航到的可能的兴趣点。当结合这种类型的信号时，ASR系统肯定会得到改善。但是，我们刚刚开始研究我们可以包含的上下文类型以及如何使用它。 部署最近对话语音的改进是不可部署的。在考虑什么使新的语音算法可部署时，考虑延迟和计算是有帮助的。这两者是相关的，因为增加计算的算法往往会增加延迟。但为了简单起见，我将分别讨论每个问题。 延迟：延迟时间是指用户完成说话到转录完成时的时间。低延迟是ASR中常见的产品约束。它可以显着影响用户体验。对于ASR系统，数十毫秒的延迟要求并不少见。虽然这可能听起来很极端，但请记住，制作成绩单通常是一系列昂贵计算的第一步。例如，在语音搜索中，必须在语音识别之后进行实际的网络规模搜索。 双向复发层是延迟杀死改进的一个很好的例子。所有最近最先进的会话语音结果都使用它们。问题是我们无法在第一个双向层之后计算任何东西，直到用户完成说话。所以延迟随着话语的长度而变化。 左：只有前向复发，我们可以立即开始计算转录。右：由于双向复发，我们必须等到所有语音到达才开始计算转录。 将未来信息有效地融入语音识别的好方法仍然是一个悬而未决的问题。 计算：转录话语所需的计算能力是一种经济约束。我们必须考虑对语音识别器的每次准确性改进的降压。如果改进不符合经济阈值，则无法部署。 一个永远不会被部署的持续改进的典型例子是整体。 1％或2％的误差减少很少值得计算增加2-8倍。现代RNN语言模型通常也属于这一类，因为它们在波束搜索中使用起来非常昂贵。虽然我预计这将在未来发生变化。 作为一个警告，我并不是建议以高计算成本提高准确性的研究是没有用的。我们之前已经看到了“先缓慢但准确，然后快速”的模式。重点是，只要改进足够快，它就无法使用。 未来五年语音识别仍然存在许多开放和具有挑战性的问题。其中包括：扩展新域，重音和远场，低SNR语音的功能。 将更多背景纳入识别过程。 Diarisation和源分离。 语义错误率和评估识别器的创新方法。 超低延迟和高效推理。 我期待未来五年在这些和其他方面取得进展。  
网络错误