

听声音写字的机器人

本课讲 2015 年的一篇文章。百度硅谷 AI 实验室的 Amodei 和同事们，用一个大网络把听到的声音直接写成文字。同一套办法，英语和中文都能学。

适合 10 岁左右

约 40 分钟

需要：铅笔、白纸、手机录音（可选）、4 位同伴

1 热身：听写里的小陷阱

请同伴把下面的拼音读给你听。你只听，不看字，把听到的写成汉字。

wǒ míngtiān kǎo shì

zhè jiàn shì hěn zhòngyào

tā shì wǒ tóngxué _____ wǒ zài shì yí cì _____

4 句里都有“shì”，却是 3 个不同的字：试、事、是。光听一个音，分不出是哪个字，得听完整句话。手机能把你说说的话打成字，这个本事叫**语音识别**，就是把听到的声音变成文字。

2 老办法：一条长长的流水线

以前，语音识别像工厂里的**流水线**。一段声音要经过好几个工位，每个工位只做一件事：

1

从声音里挑出有用的
特点

2

听出是哪些发音，比
如“sh”“i”

3

查发音词典，把发音
拼成词

4

挑出最通顺的一句话

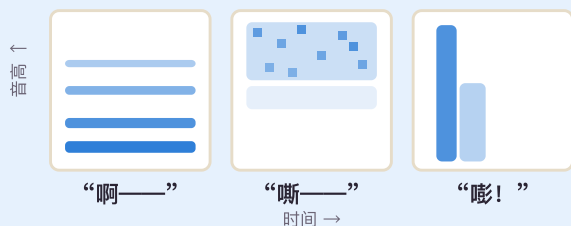
每个工位都要请**专家**来设计和调整。麻烦有 3 个：

- 换一种语言，发音表和词典都要重新做。
- 中文更难：有声调，“妈”“马”“骂”只差一个调；词和词之间也没有空格。
- 前面的工位一出错，后面只能将错就错。

3 新办法：一个大网络包办到底

百度的团队想：能不能不要流水线，**让一个大网络从头做到尾**？声音从一头进去，文字从另一头出来，中间让网络自己练。这种办法叫**端到端**。中间不再需要专家写的发音表和词典。

网络只会看数字，所以先要把声音变成一张图。这张图叫**声谱图**（也叫频谱图），你可以把它当成**声音的画**：

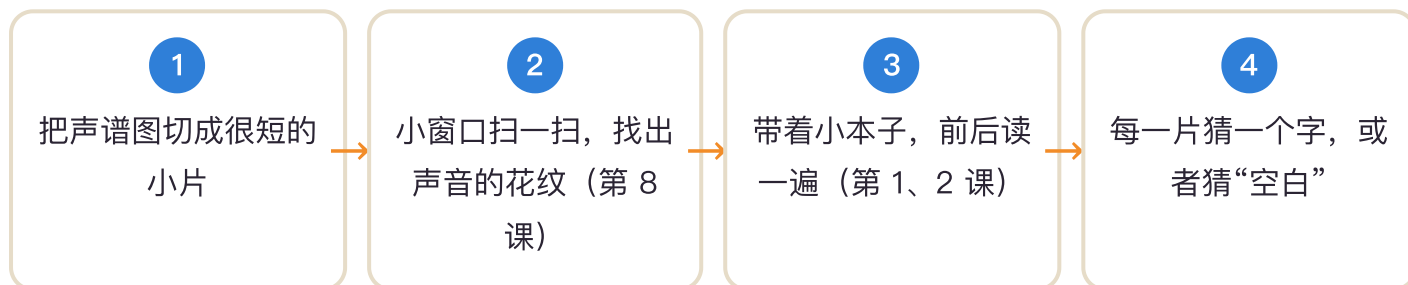


- 从左往右是时间，从下往上是音从低到高。
- 颜色越深，那个音越响。
- “啊”拉长了，是几条横纹。“嘶”像一片细沙，在高处。“膨”很短，是一道竖条。

示意图，不是真的录音。

4 网络里面：小窗口加小本子

网络像切面包一样，把声谱图切成**很短的小片**，再一层一层地处理：



第 2 步是**卷积**，就是第 8 课拿小窗口一格一格扫的办法，这次扫的是声音的画。第 3 步是**循环网络**，就是第 1 课带着小本子边读边记的程序。它正着读一遍，还倒着读一遍，每一片就知道前后在说什么。最深的网络有 11 层。学英语时，它写出字母；学中文时，它**直接写出汉字**，不用先写拼音。

5 对不齐怎么办：空白和拖长音

说“妈妈”时，声音会被切成很多片，字却只有 2 个。**哪几片对应第一个“妈”**？没人标出来，一片一片去标也太累了。

科学家用了一个聪明的规矩，叫 **CTC**。每一片可以写一个字，也可以写“空白”。然后按两步合并：



- ① 挨在一起的相同字，合成 1 个 ② 删掉所有空白

结果：妈妈

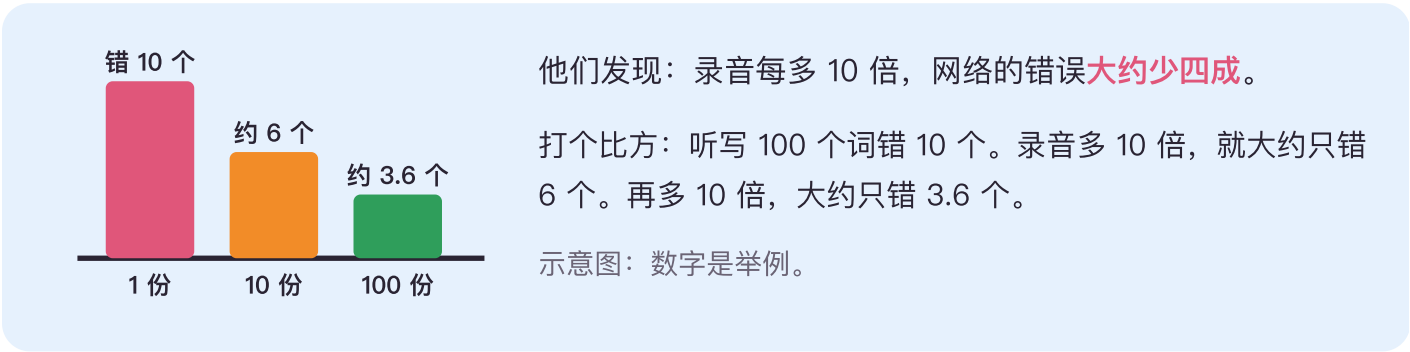
示意图：“_”表示空白片。真的网络切出的小片比这多得多。

拖长音：一个字说得慢，可以占好几片。**空白**可以填停顿，还能隔开两个一样的字。没有空白，“妈妈妈妈妈妈”会合成 1 个“妈”。

这样，练习时只要告诉网络“这段是‘妈妈’”。能合并成“妈妈”的排法有很多种，CTC 都算对。哪一片写什么，网络自己慢慢学会。

6 最重要的发现：越多越好

团队准备了大约 1.2 万小时的英语录音和大约 9400 小时的普通话录音。一个人不吃不睡地听，每种都要听 1 年多。他们问：**录音越多，网络会进步多少？**



录音太多，团队就让**很多块显卡一起练**（显卡见第 8 课）。每块显卡听不同的录音，各自想“旋钮”（第 4 课）该怎么调，再合起来一起调。就像全班分着改一大摞作业。

它和人比，**谁听得准？**诚实地说一说：

听什么	结果
清楚的朗读（念报纸、念书的录音）	比网上请来的听写员错得少
中文的简短语音搜索	比普通的中文母语听写员错得少
吵闹的地方、带口音的话	还是比人差

“越多越好”非常重要。第 25 课会专门讲它，叫“规模定律”。

7 这和今天有什么关系

手机语音打字、视频自动字幕，走的都是这条路：**一个大网络听很多录音，直接把声音写成字**。声谱图和 CTC 今天还很常用。中间的循环网络，后来大多换成了第 13 课“抬头看”的注意力网络。2022 年的 Whisper 听了大约 68 万小时录音，能听懂将近 100 种语言。

第一作者 Dario Amodei 后来创办了 Anthropic 公司，就是做聊天 AI Claude 的公司。也要诚实地说：今天的语音识别遇到吵闹和口音，还是会出错。

8 动手：声音的画

玩法（2 个人，约 15 分钟）

1. A 依次发出“啊——”“嘶——”“嘭！”。B 轻轻摸着 A 的喉咙，感觉震不震。
2. 有手机：用录音 App 录下来，看屏幕上的波浪（这叫波形图，和声谱图不一样）。没有手机：B 闭眼听，在纸上画出“听起来的样子”。然后换人再做一遍。

声音	喉咙震不震	长还是短	画出来的样子
啊——			
嘶——			
嘭！			

你会发现：“啊”喉咙会震，“嘶”不震，“嘭”一下就没了。每个声音都有自己的样子，网络就是看这样的“画”学会分辨的。

9 动手：流水线 and 一个人包办

玩法（5 个人，约 15 分钟）

念的人（1 人）：每次小声念一句。① 这件事我试过了。② 他是我们班的班长。③ 期中考试明天开始。

流水线（3 人）：1 号只听，写成拼音，传给 2 号。2 号只看拼音，一个音一个音写成汉字，传给 3 号。3 号只看汉字，把句子改通顺。

一个人包办（1 人）：直接听，直接写汉字。

句子	流水线最后写成	错几个字	一个人写成	错几个字
①				
②				
③				

你会发现：流水线上，前面一错，后面跟着错。2 号听不到整句话，“shì”常常写错。一个人包办，能用上所有线索。

10 词语小卡片

语音识别

把听到的声音变成文字。手机的语音输入就是它。

端到端

一个网络从头做到尾：声音进去，文字出来，中间不用专家设计。

声谱图

声音的画。从左到右是时间，从下到上是音的高低，颜色越深越响。

CTC

对齐的规矩。每片写一个字或空白，相同的挨着就合并，再删掉空白。

11 想一想

1. 老办法换一种语言，为什么要重新请专家？端到端的办法为什么不用？

2. 用 CTC 的两步合并下面的小片，结果是什么？（“_”是空白）

好好 _ 好 _ _ 学学学习 _ 结果：_____

3. 语音输入遇到吵闹和口音，容易写错。你会给网络多听什么样的录音？

一句话记住：给一个大网络听足够多的录音，它就能直接把声音写成字，英语、中文都行。

给家长和老师：本课讲的是 Dario Amodei 等人（百度硅谷 AI 实验室）2015 年的论文 [Deep Speech 2: End-to-End Speech Recognition in English and Mandarin](#) (ICML 2016)：频谱图经 1—3 层卷积、1—7 层双向 RNN/GRU 和全连接层（最深 11 层）逐帧输出字符（普通话直接输出汉字），CTC 训练，解码仍结合 n-gram 语言模型；英语 11,940 小时、普通话 9,400 小时，数据每增 10 倍 WER 相对降约 40%；多 GPU 同步数据并行。热身句、“啊/嘶/嘭”声谱图、“妈妈”切片图、错误柱状图（10→6→3.6）和活动句子都是示意，不是原文内容。“听 1 年多”按 $11,940 \div 24 \approx 497$ 天、 $9,400 \div 24 \approx 392$ 天推算。第 2 题答案：好好学习。想深入可看本站的 [第 22 天中文导读](#)，代码见开源项目 [30u30 第 22 天目录](#)（英文）。