

聽聲音寫字的機器人

本課講 2015 年的一篇文章。百度硅谷 AI 實驗室的 Amodei 和同事們，用一個大網絡把聽到的聲音直接寫成文字。同一套辦法，英語和中文都能學。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、白紙、手機錄音（可選）、4 位同伴

1 熱身：聽寫裡的小陷阱

請同伴把下面的拼音讀給你聽。你只聽，不看字，把聽到的寫成漢字。

wǒ míngtiān kǎo shì

zhè jiàn shì hěn zhòngyào

tā shì wǒ tóngxué _____

wǒ zài shì yí cì _____

4 句裡都有「shì」，卻是 3 個不同的字：試、事、是。光聽一個音，分不出是哪個字，得聽完整句話。手機能把你說的話打成字，這個本事叫**語音識別**，就是把聽到的聲音變成文字。

2 老辦法：一條長長的流水線

以前，語音識別像工廠裡的**流水線**。一段聲音要經過好幾個工位，每個工位只做一件事：

1

從聲音裡挑出有用的
特點

2

聽出是哪些發音，比
如「sh」「i」

3

查發音詞典，把發音
拼成詞

4

挑出最通順的一句話

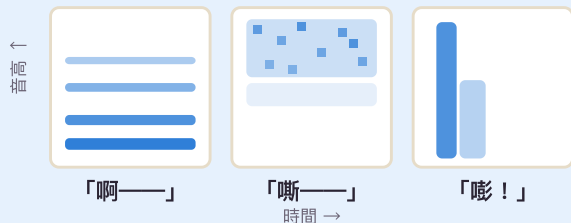
每個工位都要請**專家**來設計和調整。麻煩有 3 個：

- 換一種語言，發音表和詞典都要重新做。
- 中文更難：有聲調，「媽」「馬」「罵」只差一個調；詞和詞之間也沒有空格。
- 前面的工位一齣錯，後面只能將錯就錯。

3 新辦法：一個大網絡包辦到底

百度的團隊想：能不能不要流水線，讓一個大網絡從頭做到尾？聲音從一頭進去，文字從另一頭出來，中間讓網絡自己練。這種辦法叫**端到端**。中間不再需要專家寫的發音表和詞典。

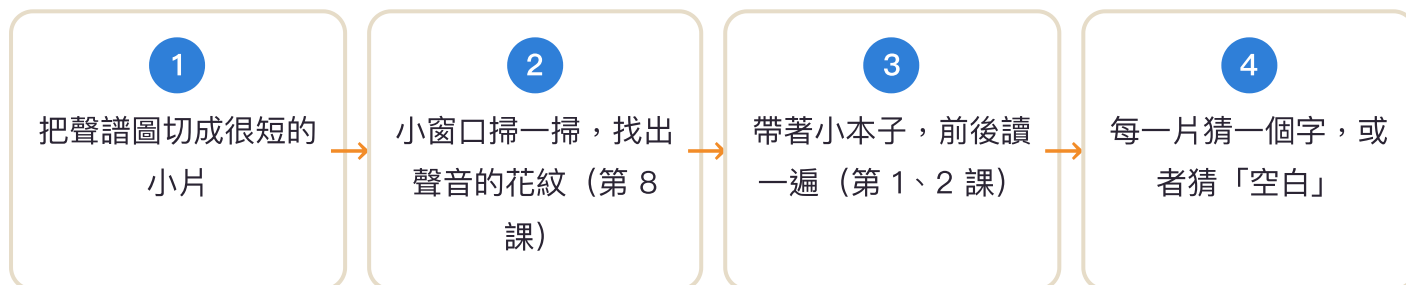
網絡只會看數字，所以先要把聲音變成一張圖。這張圖叫**聲譜圖**（也叫**頻譜圖**），你可以把它當成**聲音的畫**：



- 從左往右是時間，從下往上是音從低到高。
 - 顏色越深，那個音越響。
 - 「啊」拉長了，是幾條橫紋。「嘶」像一片細沙，在高處。「嘸」很短，是一道豎條。
- 示意圖，不是真的錄音。

4 網絡裡面：小窗口加小本子

網絡像切面包一樣，把聲譜圖切成**很短的小片**，再一層一層地處理：



第 2 步是**卷積**，就是第 8 課拿小窗口一格一格掃的辦法，這次掃的是聲音的畫。第 3 步是**循環網絡**，就是第 1 課帶著小本子邊讀邊記的程序。它正著讀一遍，還倒著讀一遍，每一片就知道前後在說什麼。最深的網絡有 11 層。學英語時，它寫出字母；學中文時，它**直接寫出漢字**，不用先寫拼音。

5 對不齊怎麼辦：空白和拖長音

說「媽媽」時，聲音會被切成很多片，字卻只有 2 個。**哪幾片對應第一個「媽」**？沒人標出來，一片一片去標也太累了。

科學家用了一個聰明的規矩，叫 **CTC**。每一片可以寫一個字，也可以寫「空白」。然後按兩步合併：



- ① 挨在一起的相同字，合成 1 個 ② 刪掉所有空白

結果：媽媽

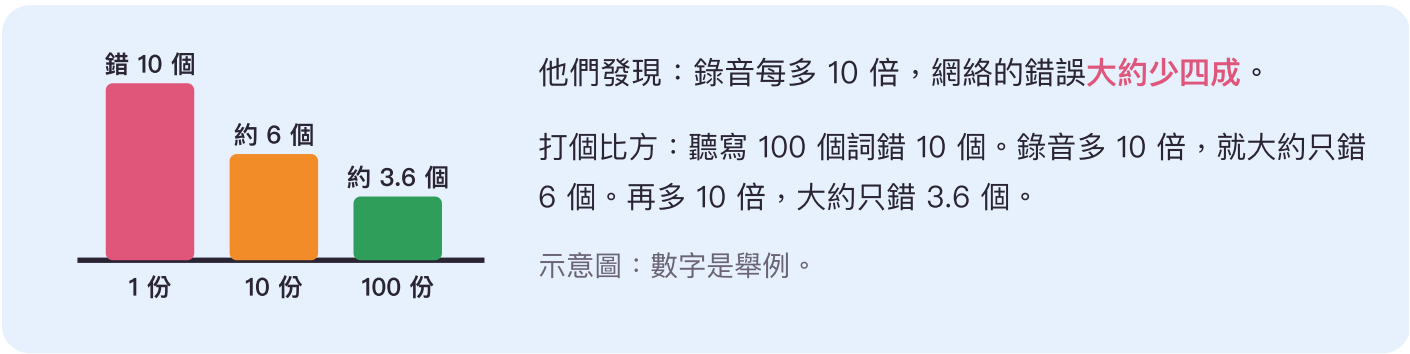
示意圖：「—」表示空白片。真的網絡切出的小片比這多得多。

拖長音：一個字說得慢，可以佔好幾片。**空白**可以填停頓，還能隔開兩個一樣的字。沒有空白，「媽媽媽媽媽」會合成 1 個「媽」。

這樣，練習時只要告訴網絡「這段是‘媽媽’」。能合併成「媽媽」的排法有很多種，CTC 都算對。哪一片寫什麼，網絡自己慢慢學會。

6 最重要的發現：越多越好

團隊準備了大約 1.2 萬小時的英語錄音和大約 9400 小時的普通話錄音。一個人不吃不睡地聽，每種都要聽 1 年多。他們問：**錄音越多，網絡會進步多少？**



錄音太多，團隊就讓**很多塊顯卡一起練**（顯卡見第 8 課）。每塊顯卡聽不同的錄音，各自想「旋鈕」（第 4 課）該怎麼調，再合起來一起調。就像全班分著改一大摞作業。

它和人比，**誰聽得準？**誠實地說一說：

聽什麼	結果
清楚的朗讀（念報紙、唸書的錄音）	比網上請來的聽寫員錯得少
中文的簡短語音搜索	比普通的中文母語聽寫員錯得少
吵鬧的地方、帶口音的話	還是比人差

「越多越好」非常重要。第 25 課會專門講它，叫「規模定律」。

7 這和今天有什麼關係

手機語音打字、視頻自動字幕，走的都是這條路：**一個大網絡聽很多錄音，直接把聲音寫成字**。聲譜圖和 CTC 今天還很常用。中間的循環網絡，後來大多換成了第 13 課「抬頭看」的注意力網絡。2022 年的 Whisper 聽了大約 68 萬小時錄音，能聽懂將近 100 種語言。

第一作者 Dario Amodei 後來創辦了 Anthropic 公司，就是做聊天 AI Claude 的公司。也要誠實地說：今天的語音識別遇到吵鬧和口音，還是會出錯。

8 動手：聲音的畫

玩法（2 個人，約 15 分鐘）

1. A 依次發出「啊——」「嘶——」「嘸！」。B 輕輕摸著 A 的喉嚨，感覺震不震。
2. 有手機：用錄音 App 錄下來，看屏幕上的波浪（這叫波形圖，和聲譜圖不一樣）。沒有手機：B 閉眼聽，在紙上畫出「聽起來的樣子」。然後換人再做一遍。

聲音	喉嚨震不震	長還是短	畫出來的樣子
啊——			
嘶——			
嘸！			

你會發現：「啊」喉嚨會震，「嘶」不震，「嘸」一下就沒了。每個聲音都有自己的樣子，網絡就是看這樣的「畫」學會分辨的。

9 動手：流水線和一個人包辦

玩法（5 個人，約 15 分鐘）

唸的人（1 人）：每次小聲念一句。① 這件事我試過了。② 他是我們班的班長。③ 期中考試明天開始。

流水線（3 人）：1 號只聽，寫成拼音，傳給 2 號。2 號只看拼音，一個音一個音寫成漢字，傳給 3 號。3 號只看漢字，把句子改通順。

一個人包辦（1 人）：直接聽，直接寫漢字。

句子	流水線最後寫成	錯幾個字	一個人寫成	錯幾個字
①				
②				
③				

你會發現：流水線上，前面一錯，後面跟著錯。2 號聽不到整句話，「shì」常常寫錯。一個人包辦，能用上所有線索。

10 詞語小卡片

語音識別

把聽到的聲音變成文字。手機的語音輸入就是它。

端到端

一個網絡從頭做到尾：聲音進去，文字出來，中間不用專家設計。

聲譜圖

聲音的畫。從左到右是時間，從下到上是音的高低，顏色越深越響。

CTC

對齊的規矩。每片寫一個字或空白，相同的挨著就合併，再刪掉空白。

11 想一想

1. 老辦法換一種語言，為什麼要重新請專家？端到端的辦法為什麼不用？

2. 用 CTC 的兩步合併下面的小片，結果是什麼？（「_」是空白）

好好 _ 好 _ _ 學學學習 _ 結果：_____

3. 語音輸入遇到吵鬧和口音，容易寫錯。你會給網絡多聽什麼樣的錄音？

一句話記住：給一個大網絡聽足夠多的錄音，它就能直接把聲音寫成字，英語、中文都行。

給家長和老師：本課講的是 Dario Amodei 等人（百度硅谷 AI 實驗室）2015 年的論文 [Deep Speech 2: End-to-End Speech Recognition in English and Mandarin](#) (ICML 2016)：頻譜圖經 1—3 層卷積、1—7 層雙向 RNN/GRU 和全連接層（最深 11 層）逐幀輸出字符（普通話直接輸出漢字），CTC 訓練，解碼仍結合 n-gram 語言模型；英語 11,940 小時、普通話 9,400 小時，數據每增 10 倍 WER 相對降約 40%；多 GPU 同步數據並行。熱身句、「啊/嘶/嘔」聲譜圖、「媽媽」切片圖、錯誤柱狀圖（10→6→3.6）和活動句子都是示意，不是原文內容。「聽 1 年多」按 $11,940 \div 24 \approx 497$ 天、 $9,400 \div 24 \approx 392$ 天推算。第 2 題答案：好好學習。想深入可看本站的 [第 22 天中文導讀](#)，代碼見開源項目 [30u30 第 22 天目錄](#)（英文）。