

Building transcribed speech corpora quickly and cheaply for many languages

By Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu,
Pedro J. Moreno, Mike LeBeau / September, 2010

reporter : 許妙鸞
Professor : 陳嘉平

Abstract

- We present a system for quickly and cheaply building transcribed speech corpora containing utterances from many speakers in a variety of acoustic conditions.
- The system consists of a client application running on an Android mobile device with an intermittent Internet connection to a server.
- The client application collects demographic information about the speaker, fetches textual prompts from the server for the speaker to read , records the speaker's voice, and uploads the audio and associated metadata to the server.
- The system has so far been used to collect over 3000 hours of transcribed audio in 17 languages around the world.
- Index Terms: speech corpora, speech recognition, internationalization

Introduction

- 語音語料庫是系統用來建造聲學模型的生命線，但是錄音和轉錄語音語料庫是相當耗時且昂貴
- 在訓練聲學模型時，研究者往往依賴於著名的語料庫。例如:Switchboard

已存在的語料庫的缺點

- 透過telephony channel錄製的語音品質較低
- 缺乏多樣化的單字
- 有些價位昂貴或者需要授權
- 和現實生活使用的條件不符合
- 各種世界語言的轉錄語料庫不夠大

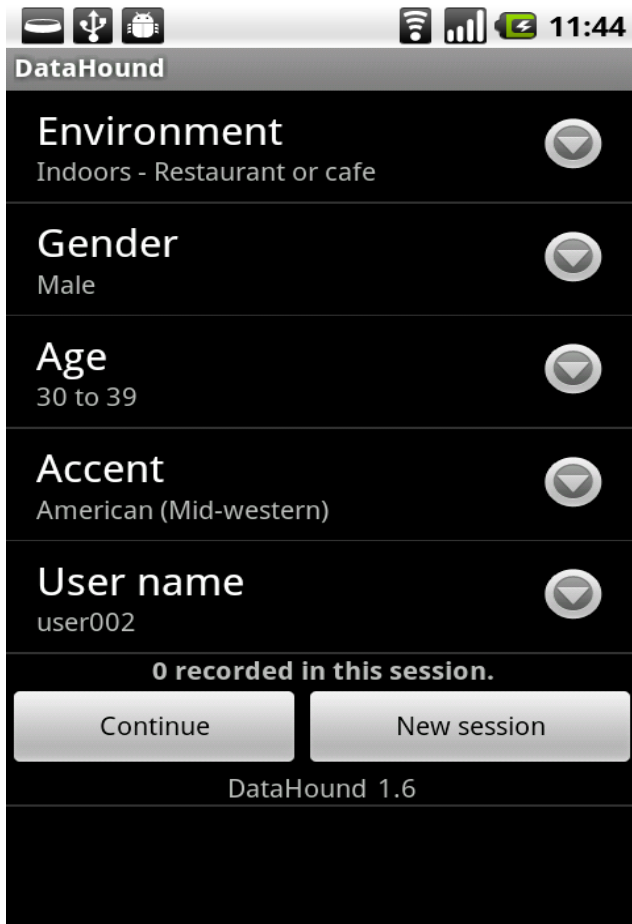
系統的優點

- 錄製語音語料庫快速且容易
- 可以選擇錄音的環境，錄製出來的品質比較高
- 可以透過網路，將語音和相關資料集中存在一個地方(**server**)，以便日後使用
- 可以同時做資料的收集和保持追蹤相關的資料

系統架構

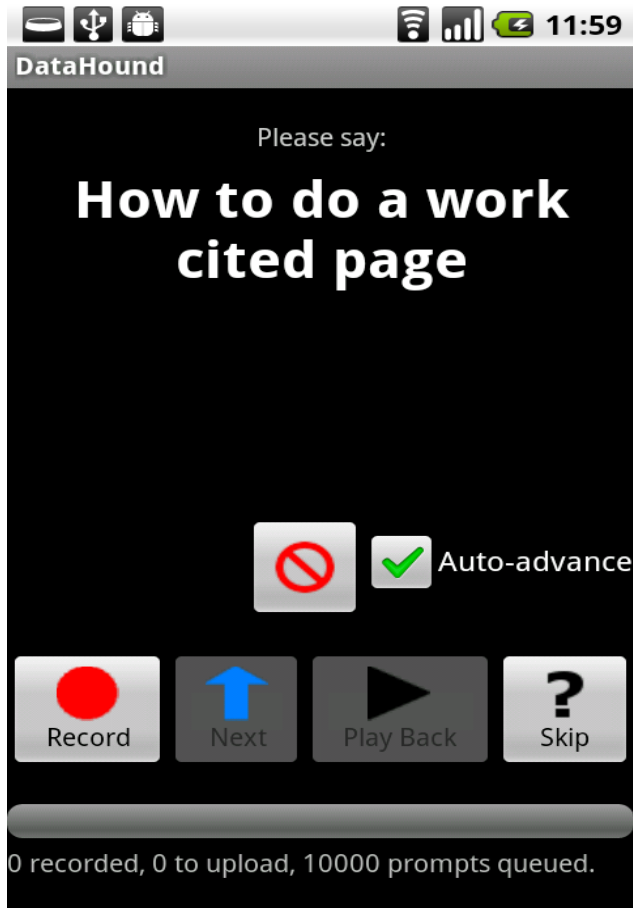
- Client/Server架構
- Client 使用 google Android平台的行動裝置
- Server 處理從Client發出的HTTP請求及儲存錄製好的語音和相關資料
- Client不需要一直與Server保持連線，只要在上傳錄製的語音 和 下載提示詞時連線即可



User Interface-Initial screen



- 聲音環境(室內，室外，在汽車上，背景有噪音)
- 性別
- 年齡:以十年為一個群組
- 口音(例如:美國)
- 使用者名稱:用來辨別session
- 應用程式自動標示
 - 日期和時間
 - 手機的硬體版本和Android OS版本
 - 電話的IMEI數字(國際行動裝置識別)
 - 地理位置(用GPS定位)

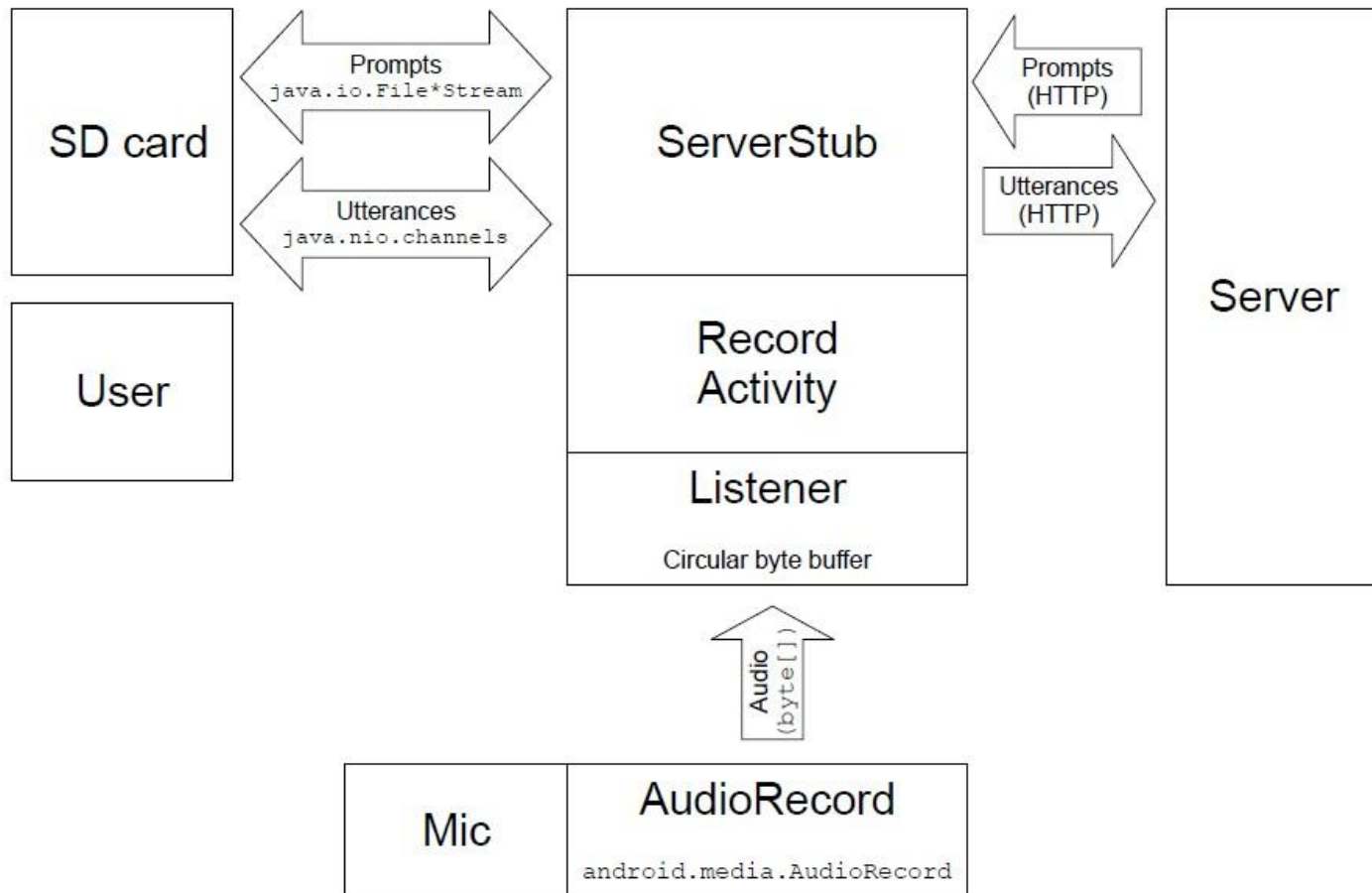
User Interface-Recording screen



- How to do a work... : 錄音的提示字
- Record : 錄音，再按一次結束錄音
- Next : 繼續錄下一個提示字
- Skip : 跳過自己不想要的提示字
-  : 使用者覺得受到提示字的冒犯
-  : 自動進入下一個提示字和錄音

Client會記錄有多少語音被錄製，當到達需要的數字，會通知使用者，並結束這個session

Client Implementation



Server Application

- 提供Client 要錄音的提示字
- 透過HTTP interface，接收和聚集錄好的語音及使用者的相關資料
- 如果Client 需要更多的提示字，可以透過HTTP POST 向Server 提出請求
- 使用認證，去防止未授權的使用者存取提示字或上傳錄音
- 當Server收到全部的語音和使用者的相關資料，會通知Client在SD卡上的語音可以安全移除

提示字的準備

- Server要先產生一個很大的提示字列表，讓speaker去讀
- 提示字的內容是google查詢常用的關鍵字或句子
- 使用查詢的關鍵字或句子的原因：
 - 使用這些語料庫去訓練google語音搜尋所使用的“聲學模型”
 - 我們希望打字和口說的網頁搜尋是類似的
 - 現代的關鍵字查詢都包含一些不在標準詞彙中的流行語，例如:pokemon,openvpn

語音錄製

- 在手機可以錄音之前，要先設定語言和環境
- 在沒有網路的環境下，要先下載大量的提示字到SD卡上
- 錄製一個500個語音的session:
 - 平均要花30分鐘
 - 改成auto-advance模式，只要花20分鐘
 - 一天可能錄製8000個語音

人群的來源

- 這套系統很容易使用且不貴，因此非相關技術人員也可以收集語音資料
- 在各個地方，聘請大學生去收集資料
- 這些學生會設定一個目標人數、地方分布、錄音環境分布

語料庫的特徵

- 語音簡短，只有幾個字
- 每個語音都有註釋和speaker相關的資料
- 要給speaker讀的提示字同樣有註釋相關資料

Speaker Errors

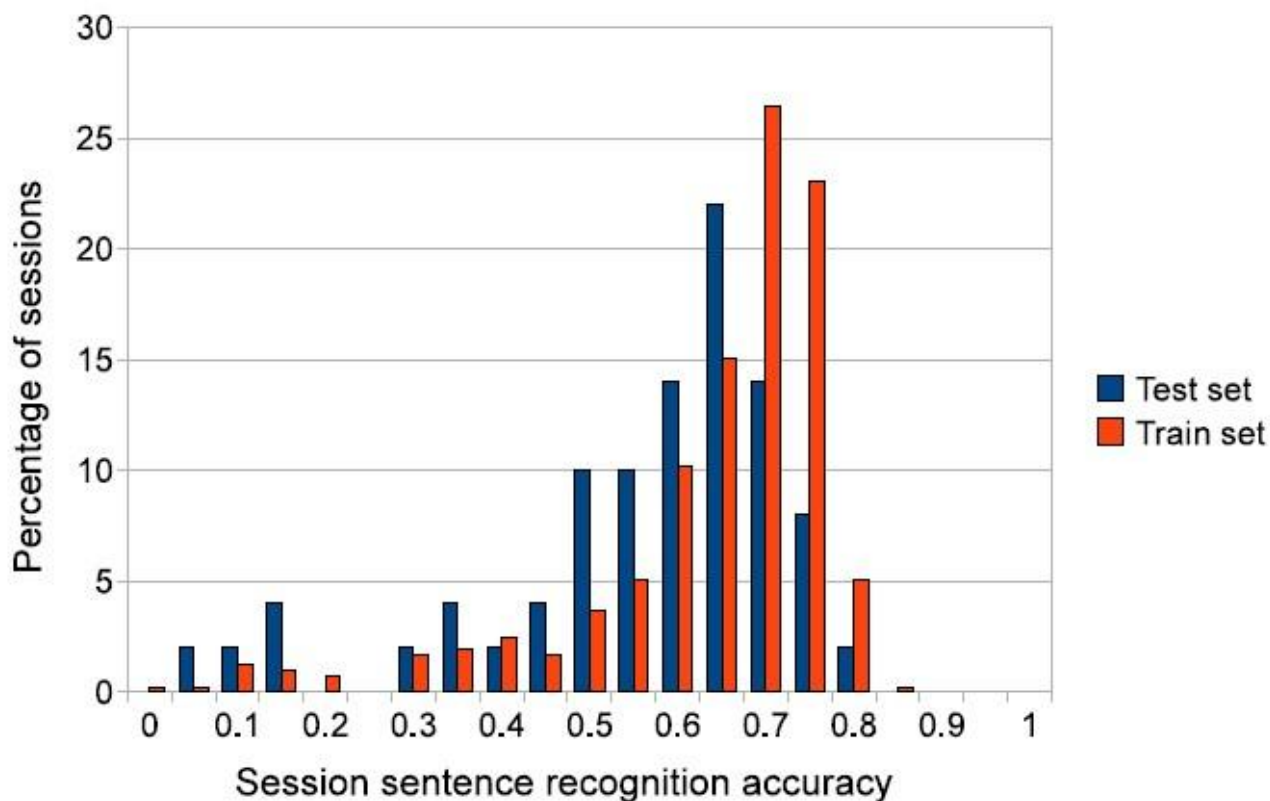
Error category	Rate	German example
Misread	3.5%	“fasanerie” → “fanaserie”
Side-speech/noise	2.5%	Coughing, extra commentary “scheffler” → “quatsch scheffler”
Restart	2.0%	“fm09” → “f m oh-f m null neun”
Truncation	1.5%	“kinoprogram” → “-gram”
Empty	0.5%	“konstanz” → “”

本例子是以德國人錄製語料庫的發聲錯誤為例，錯誤率總和是10%

評估辨識的效能

- 把新的語料庫分成兩個部分
 - training data : 80%
 - testing data : 20%
- 確認沒有speaker同時進行training和testing

每個session的辨識正確率



每一個句子在training的辨識正確率 和 test data

辨識不正確的單字

- **Speaker**錄製的提示字包含很多流行語是不存在於標準的詞彙中
- 辨識者利用文轉音的規則去產生發音，但不是每次都是正確的，例如: **pokemon** , **openvpn**
- 沒有工具可以自動校正不正確的發音，但可以半自動的利用在訓練的資料上找出辨識的最大可能性
- **MMI(Maximum Mutual Information)**可以解決這個問題

結語

- 這套系統是一套對於建立各種語言的語料庫方面，相當有效率