Exploring Web-Browser based Runtimes Engines for Creating Ubiquitous Speech Interfaces

Reporter:吳柏鋒
Professor:陳嘉平

摘要

This paper describes an investigation into current browser based runtimes including Adobe's Flash and Microsoft's Silverlight as platforms for delivering web based speech interfaces. The key difference here is the browser plugin is used to perform all the computation without any server side processing. The first application is an HMM based text-to-speech engine running in the Adobe Flash plugin. The second application is a WFST based large vocabulary speech recognition decoder written in C# running inside the Silverlight plugin.

摘要

- 簡介
- 瀏覽器內嵌語音合成系統
- 瀏覽器內嵌語音辨識系統
- 實驗

簡介

- 研究如何使用網頁瀏覽器內嵌技術來執行以語音介面為基礎的網頁

- 近年來提出的雲端技術，語音處理引擎是建立在server端，就因為透過分散式的方式作處理，所以相對需要消耗較大的處理器能源

簡介

- 在此論文中與以往最大不同之處在於所有的運算過程都不在server端執行
- 本論文希望可以使用以rich-client為基礎的技術，將所有運算部分建構在client端

瀏覽器內嵌語音合成

- 提出以HMM為基礎的文字轉語音(TTS)引擎其執行在Adobe Flash插件上

- 為了加速以語音合成為基礎之瀏覽器的發展，並不需要改寫整個系統的Action script，而是透過使用Alchemy 編譯器直接在flash元件中編譯HTS+Flite引擎

瀏覽器內嵌語音辨識

- 近年來有許多高效能的分散式語音辨識系統可以在不同平台上面使用，最常見的就是智慧型手機，另一種就是以網頁為基礎的辨識翻譯引擎

- 但這些主要的運算都還是在server端執行，缺點就是大型運算設備太消耗能源，且研究與管理維護設備需要很多時間

瀏覽器內嵌語音辨識

- 提出以WFST(Weighted finite state transducer)為基礎的大型字彙語音辨識解碼器其執行在Microsoft Silverlight插件上

- 即語音辨識的動作是在client端執行，如此可以降低辨識的延遲性且不需要使用任何server端的支援

T4解碼器

- 在辨識核心系統方面，使用以C#語言寫成T4解碼器其能夠在不同的.NET framework上運行

- T4解碼器使用Viterbi beam search演算法，可以同步搜尋相對應的麥克風輸入與訊號處理元件

T4解碼器架構圖

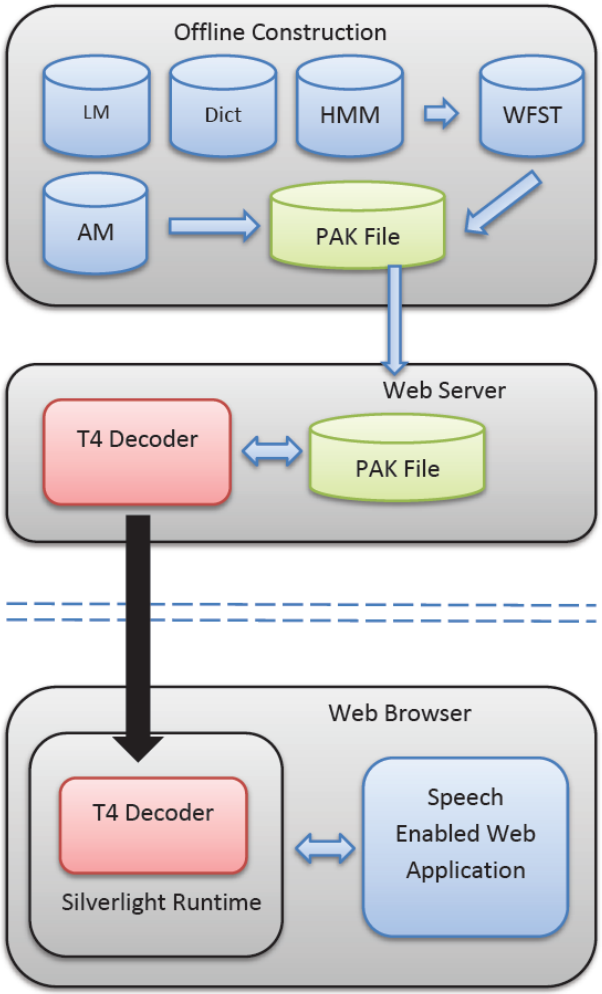

Figure 1: Diagram showing the model preparation and server-client architecture of the Silverlight decoder.

多核心的聲學評分

- 聲學評分在解碼乾淨語音與分析T4解碼器需要占用大量的CPU

- 在此提出一個有效加速的.NET threadPool 技術，用來計算平行處理的每frame中所需的states分數

實驗

- 使用Corpus of Spontaneous Japanese(CSJ)
  語料庫，以頭戴式麥克風錄音，錄音時間650小時，共700萬字，取樣頻率16KHz

- 測試集由10場演講所取得音段句子

實驗

- 將語音檔轉換成39維的特徵向量，為12維 MFCCs、log delta 和 log delta-delta，在加上 log energy
- 每個發聲模型是3個states的tri-phone HMM 模型
- 實驗中，word 正確性漸近81.34%