◎ 蔡玉秀、廖元甫

語音辨識與資料探勘—在廣播媒體的運用

語音辨識與資料探勘技術導入廣播媒體後，
除了期望提高工作效率、創新加值服務，
讓廣播節目的內容更易搜尋與運用外，
把廣播語音資料加以處理
可創造教育電台典藏節目的新價值。

因應廣播節目的戰場已從空中逐步轉移至網路上，必須配合網路傳播的特性，找到創新的服務模式。

廣播媒體的挑戰

隨著網路科技的普及與影音媒體科技的快速發展，民眾持續加重其對網路媒體的依賴，閱聽收視習慣逐漸改變，網路電台及行動載具成為廣播電台最貼近民眾的渠道。如何突破現有實體廣播頻道框架，開創跨平台多通路媒體營運模式，成為傳統廣播媒體發展不可忽視的課題與挑戰。

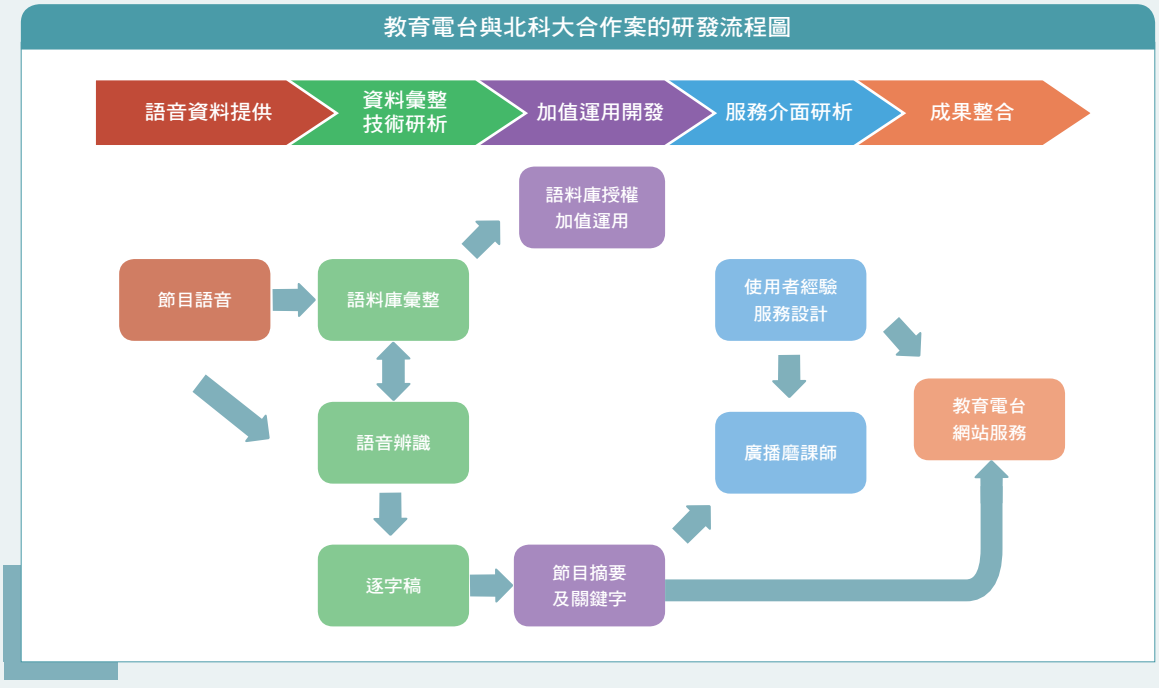
因應廣播節目的戰場已從空中逐步轉移至網路上，更要配合網路傳播的特性，不斷在節目內容、加值服務、行銷推廣、聽友經營等面向找到創新的服務模式。為了掌握使用者的喜好，需要分析其收聽行為，例如，各別使用者偏好什麼類型的節目，所收聽過的各單集節目有何共通點等。然而要分析這些資訊，需要有完整的節目文本資料才能進行精確的分析。而廣播節目都是以語音方式存在於資訊系統中，並非文字資料，造成分析上的困難。

另一方面，為達成節目內容在線上及不同網路平台上的多角化經營策略，節目製播人員的工作也隨著時代趨勢從單一的語音節目製播轉為節目製播、音檔處理、資料登錄操作、影音直播、社群經營、聽友互動等多元的工作內容。因此需要重新思考，在節目製播人員的工作流程中，是否可運用資訊技術減輕其不堪負荷的工作壓力，使其仍有餘力製作優質的節目，達成良性的工作循環。

跨域合作

為面對前述廣播媒體的挑戰，教育廣播電台（以下簡稱教育電台）與臺北科技大學（以下簡稱北科大）在教育部數位人文與科技部智慧博物館計畫經費的支持下，從 105 年 9 月起嘗試透過跨域合作方式，結合北科大電子工程系、資訊工程系、應用英文系、互動設計系等跨領域的人才進行以下的合作事宜。

教育電台與北科大合作案的研發流程圖



研析透過語音辨識技術把廣播節目語音轉為節目逐字稿，俾利廣播節目內容可用文字方式查詢、傳播、再利用，並運用於加值服務。

以資料探勘技術分析節目逐字稿，自動產製節目摘要、關鍵字。一方面可節省人工逐筆登錄的人力，一方面可避免因節目內容管理人員登錄習慣不同所造成資料的良莠不齊，同時相關資料可做為節目間關聯性的分析基礎。

把廣播節目語音透過轉逐字稿、人工校正、標注等過程產製為語料庫，提供語音辨識技術、語言學家語言教學時運用，期能建構國內主要的中文語料庫，為我國語音辨識及人工智慧、華語教學領域奠定厚實基礎，並創造數位典藏廣播節目的新價值。

以使用者經驗服務設計的觀點及科學化量測儀器，分析教育電台現行網站服務的優缺點，做為未來的改善方向。同時把合作案的成果建構為「廣播磨課師」，並呈現為廣播節目學習資源，供大眾學習運用。

當廣播媒體遇上語音辨識科技

在這次合作案中引進語音、語言及聲學處理技術，尤其是機器學習中的深層類神經網路技術，自動處理教育電台數位典藏的大量廣播節目語音資料，把它轉成節目的文字逐字稿，以便針對節目內容作全文檢索。

處理程序主要分為 2 個階段，包括：

語音與非語音段落偵測—把節目語音資料透過語音與非語音（如音樂）段落的偵測作音檔切割。主要是對節目進行中發生的所有事件，例如談話、音樂、過場或廣告等，利用深層類神經網路標注，然後把語音段落從節目中切割出來。透過音訊事件偵測系統區分語音、音樂和其他（笑聲、特效聲），當系統偵測到語音時會把它切割，進而區分成若干不同種類的音檔。

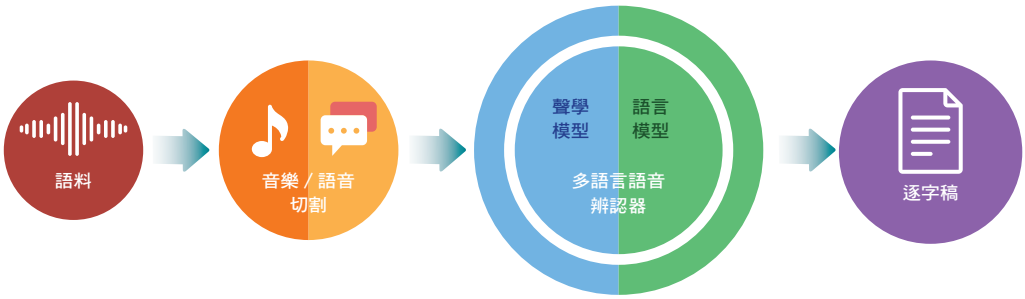
語音段落逐字稿的產生—把切割好的單純的語音檔案透過多語言語音辨認器辨識後產出逐字稿。針對上階段偵測到的每一語音段落，透過語言語音辨認器進行聽寫辨認，以產生節目的文字逐字稿。語言語音辨認器主要是透過聲學模型及語言模型的訓練，達成辨識語音的功能。

經過上述處理後，最後可產生帶有時間標記的語音內容文字。這結果可直接用於全文檢索，同時可產製成為字幕檔，也可作為節目同步字幕運用。

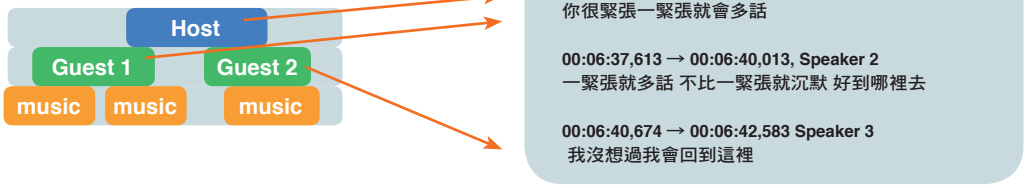
這項合作案共完成了 2,061 小時的節目語音檔彙整、1,922 小時逐字稿的產製及 380 小時人工逐字稿校正作業，也開發線上自動語音轉寫系統，於網頁直接上傳音檔，就可辨認並產生逐字稿。網頁操作介面產出逐字稿後，也可下載字幕檔格式儲存。

透過語音辨識技術把廣播節目語音轉為節目逐字稿，俾利廣播節目內容可用文字方式查詢、傳播、再利用，並運用於加值服務。

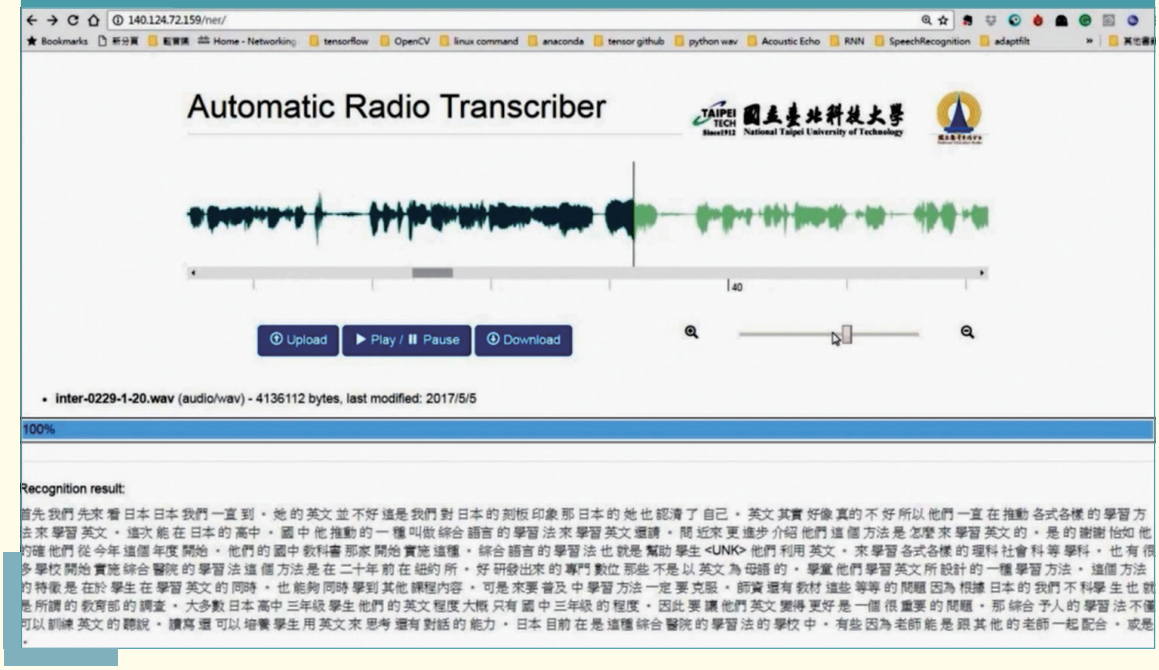
語音轉逐字稿的概念

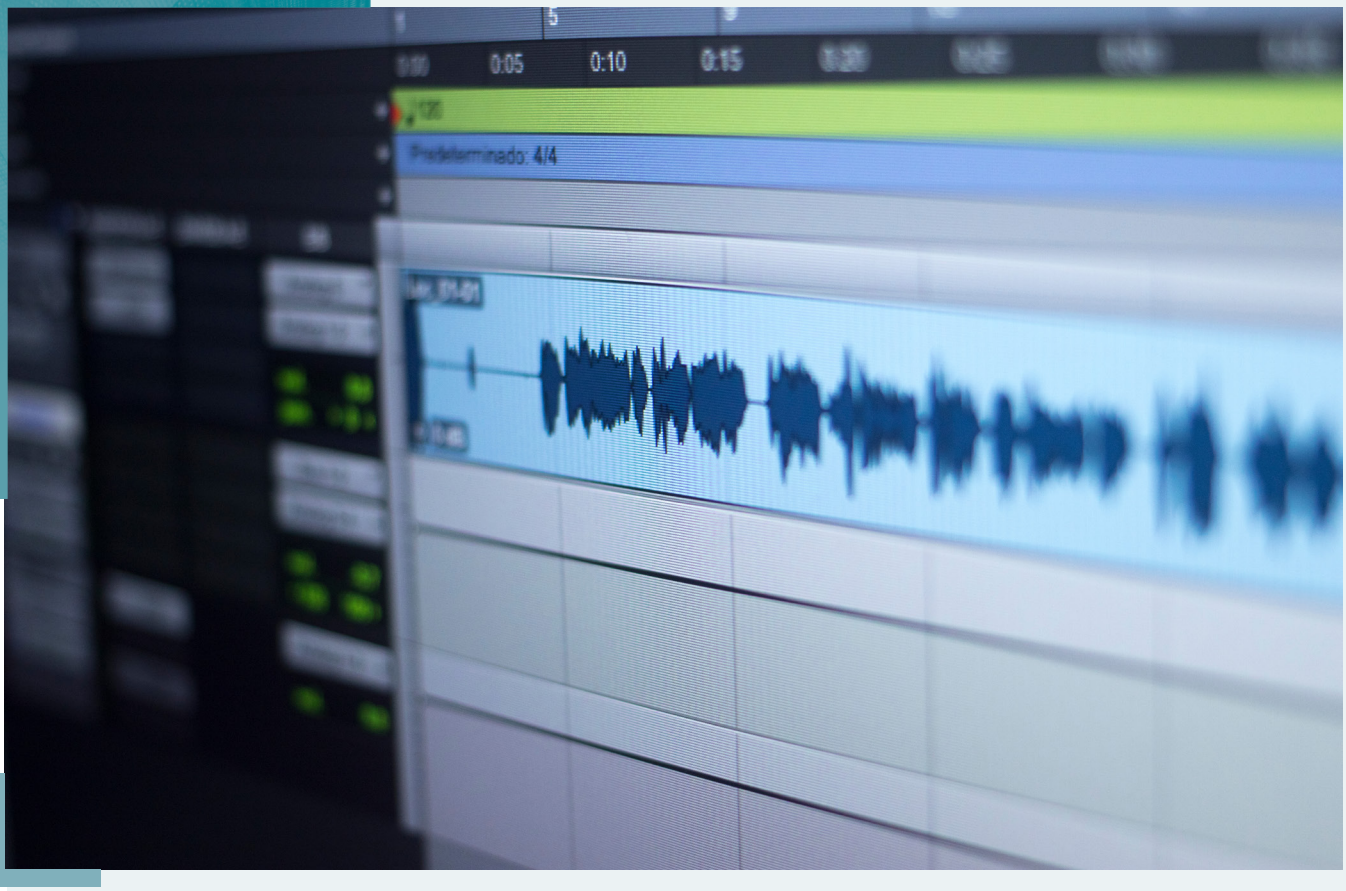


廣播節目語音文字逐字稿產出的示意圖



online 自動語音轉寫系統操作的介面



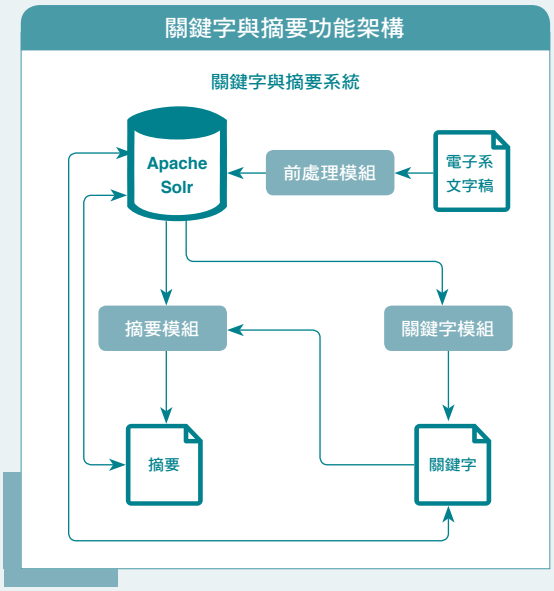


透過語音辨識技術把廣播節目語音轉為節目逐字稿，以利廣播節目內容可用文字方式查詢、傳播、再利用。（圖片來源：種子發）

廣播節目內容的資料探勘

完成節目逐字稿產製後，再透過資料探勘技術，研發產製逐字稿的關鍵字及摘要的功能，並把節目管理、逐字稿匯入、摘要與關鍵字產出等功能，開發成互動學習雛型平台「廣播磨課師」，以具體實踐對未來互動學習服務模式的雛型。在這系統中，透過資料探勘技術，自節目逐字稿中擷取關鍵字與文件摘要，說明如下。

前處理模組—由於匯入的逐字稿是以時間搭配字詞形成，因此針對產出的原始逐字稿進行前處理，根據時間資訊、空白進行過濾，加入標點符號形成句子，以便形成文章型式的處理後逐字稿。



關鍵字模組—把前處理模組完成的文章型式的文字稿，透過 Jieba 斷詞搭配 TF-IDF 方式取得各字詞的分數，把最高分 10 個字詞作為關鍵字，保存至 Apache Solr，輸入是文字稿，輸出是關鍵字。

摘要模組—把使用前處理完成的文章型式文字稿，透過標點符號斷成多個句子，再使用 Jieba 斷詞搭配 TF-IDF 取得每句分數，以分數高低作為依據產生摘要，保存至 Apache Solr，輸入是文字稿，輸出是摘要。

經過前述 3 個模組，當匯入逐字稿後，就可取得文章型式的文字稿、關鍵字及摘要。

語料庫產製與運用

在這次合作案中，另一個重要的成果是在把廣播節目轉文字的過程中，同時產製了語料庫。語料庫是藉由語言學的角度處理語音內容，並逐一進行標注而成的資料庫，可應用在詞典或文法書編纂、詞彙搭配研究、語言教學、自然語言處理（如機器翻譯）、翻譯研究等各種領域。

另外，語料庫是語音辨識科技的基礎，因發展語音合成及辨識科技的終極目標是合成出像真人發音一樣自然、順暢的語流，以及辨識不需特意清楚咬字的言語，也就是自發性的語言。因此，所有語音合成及辨識的語音模型，最後都必須用自發性的語料來訓練、測試它們的成效。

而人工智慧服務中最重要的人機介面技術，就是如何讓機器聽懂人類的語音，理解語意與溝通的意圖，以提供更貼近人性需求的服務內容。透過整理教育電台豐富的數位典藏節目音檔，經過語音辨識、校正、標注等過程，最後整理成訓練語音辨認器所需的格式，建立成為廣播節目加值語料庫，整理出來的語料共有 17 個節目，35,186 個檔案。

而自 2017 年 8 月起，適逢科技部為促進國內人工智慧技術的發展，期望透過「科技大擂台與 AI 對話」競賽，鼓勵尖端技術團隊競相投入 AI 關鍵技術研發。由於這個競賽需要相關中文語料庫，以提供參賽隊伍進行語音辨識模型訓練，以及測試競賽隊伍語音辨識的正確率，因此當教育電台與北科大得知這消息後，就與科技部及「科技大擂台與 AI 對話」競賽委辦單位國家實驗研究院接洽。經多次洽談後，教育電台與國家實驗研究院簽署語料庫授權協議書，在這競賽活動中運用教育電台語料庫，目前提供比賽的語料庫約為 443 小時。

合作成果的加值運用

為了發揮合作案成果，並提升教育電台數位典藏節目的加值應用，希望結合合作案中語料庫彙整、語音轉逐字稿產製、語音文件摘要與關鍵字分析、廣播磨課師系統開發、使用者服務設計等研發成果，為教育電台未來在節目內容的呈現與服務上奠定厚實的基礎，也提供廣播媒體界參考運用。未來相關成果預期可發展的加值運用方向如下。

導入自動化產製節目語音轉逐字稿、節目內容摘要及關鍵字等機制於教育電台節目管理流程中，藉此建立節目知識網，串聯各節目單元間的關係；應用節目知識網於服務平台上，提供更完整的節目資訊及延伸學習或節目搜尋運用。

藉由自動分析逐字稿、關鍵字等建立各節目單元間的關係，並記錄使用者線上

Corpus	abbreviation	Source	Hours	Remark
Mandarin Chinese Broadcast News corpus	MATBN	PTS	198.0	story and speaker boundaries
NER Phonetic Annotation corpus Vol. 1	NER-PHA-Vol1	NER	6.5	phone, syllable, speaker and code-switching
NER Manual Transcription corpus Vol.1	NER-Trs-Vol1	NER	126.6	manual, word sequences
NER Automatic Transcription corpus Vol.1	NER-Auto-Vol1	NER	309.6	auto., word sequences with recognition error rate prediction (QE) and confidence measure (CM)
PTS Manual Subtitlig corpus Vol.1	PTS-MSub-Vol1	PTS	264.0	manual subtitling with time-code
Total			879.0	exclude NER-PHA-Vol1

- PTS: Taiwan Public Television Service
- NER: National Education Radio

提供給科技大擂台比賽用的教育電台語料時數

收聽行為，運用資料探勘與分析技術深入分析使用者行為，以掌握使用者喜好進行精準而專屬個人的節目推薦服務。

持續彙整語音語料庫，並提升合作案的語音辨識技術。另發行語音語料庫，提供產官學研開發使用，並運用教育電台多種語言教學節目內容，未來可納入閩南語、客語、原住民語或其他語言的語料，發揮數位典藏節目的附加價值。

視語音辨識率的提升進度及人力狀況，可製作有同步逐字稿的廣播節目內容，除可以讓一般人士經由搜尋引擎直接收聽重點節目內容外，讓聽障人士或一般人在吵雜環境中也能直接閱讀節目內容。

結合教育電台多種語言教學節目內容，以及與北科大合作案的相關技術，研發語言教學相關功能，提供深入與精緻的語言學習服務。

持續進行使用者服務設計分析，激發未來創新服務模式並掌握媒體服務趨勢。同時整合教育電台相關系統與研發案成果，營造更好的工作流程與加值服務。

雖然教育廣播電台與北科大的合作案配合數位人文計畫期程於 2018 年 8 月底告一段落，但合作案成果對國內語音辨識技術的產業發展、電台未來工作流程與加值應用都有可預期的重大價值，預計以下列方式持續合作。

持續釋出彙整後的相關語料庫，讓國內學術界、產業界運用語料庫精進語音辨識技術，或運用於語言教學相關領域中，期能逐漸建構國內最大語料庫，解決國內語料庫缺乏的問題、為人工智慧領域奠基，並爭取後續相關計畫經費或透過語料庫授權所得，挹注於語料庫產製所需，建構永續營運模式。

委外開發「廣播節目知識系統」，結合雙方計畫成果、研發技術並導入產業界研發能量，期能建立產學研合作模式，持續運用計畫成果，進行語音辨識及資料探勘的技術交流，以建立語音轉文字稿、產製逐字稿、摘要、關鍵字等全自動處理流程，並開發後續加值服務。

在教育電台與北科大進行跨域合作的過程中，找到語音辨識與資料探勘技術導入廣播媒體後，未來發展的各種可能性。除了期望能提高工作效率、創新加值服務，讓廣播節目的內容更易搜尋與運用外，透過把廣播語音資料加以處理，可成為語音辨識、人工智慧等領域重要的資源與基礎，創造教育電台典藏節目的新價值。期待未來能具體把合作成果實踐為線上創新的加值服務，以嘉惠廣大的閱聽眾、貼近民眾的需求，為廣播媒體開創一條新的道路。

蔡玉秀
教育廣播電台

廖元甫
臺北科技大學電子工程系