四縣腔(含南四縣）語音辨識語料 396 時

(HAT-ASR-Sixian-Reading)

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 語料庫內容

語料庫總計 208 位發音員，總時長約 396 小時。

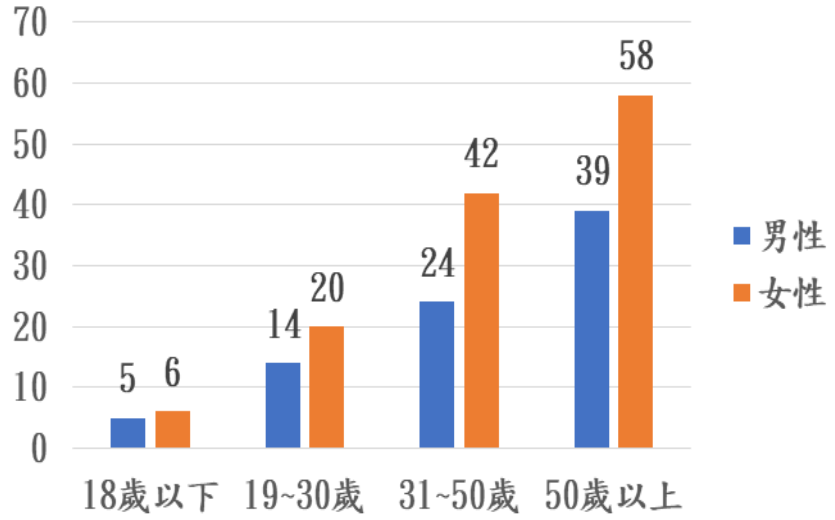
三、處理說明

資料庫為客語朗讀語料（reading speech)，是以原生客語文本，收集來自臺灣各地的四縣腔(含南四縣)語音，錄製環境為安靜密閉的空間，並同時以 8 支麥克風進行錄製完成。錄製的客語語音，經由客語老師兩次人工校正文本後，整理成可供語音辨認技術研究與開發使用之語音語料庫。

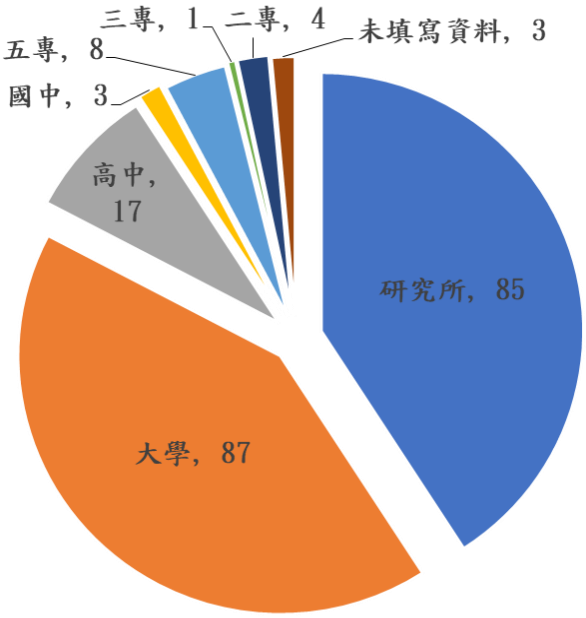
四、音檔格式

副檔名	Channels	Sample Rate	Precision	Sample Encoding
wav	8	16000Hz	24-bit	24-bit PCM

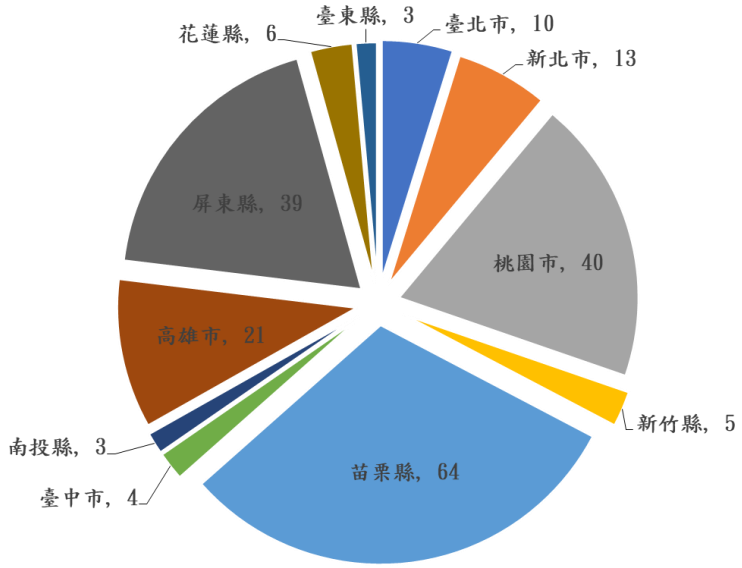
五、語者資訊



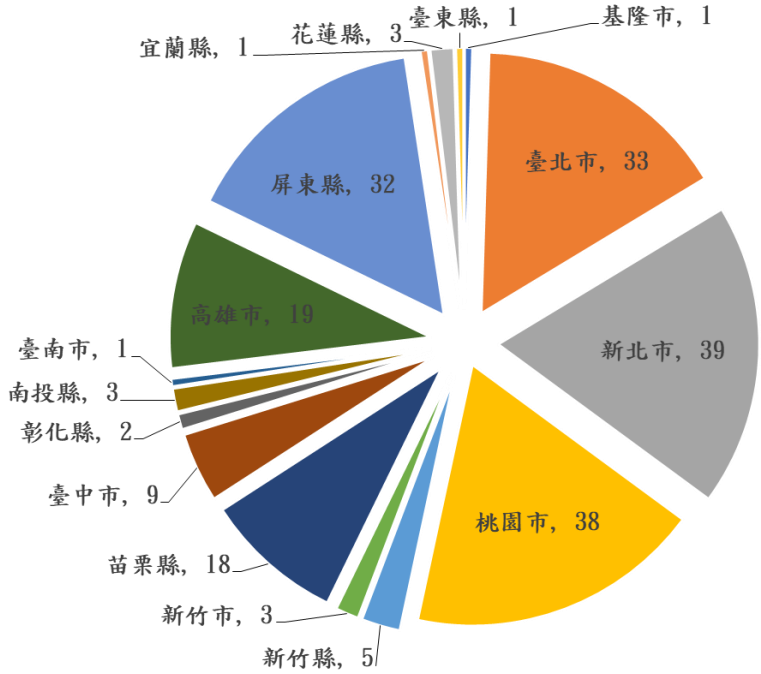
四縣腔(含南四縣)發音員性別及年齡(208 人)



四縣腔(含南四縣)發音員教育程度(208 人)



四縣腔(含南四縣)發音員 18 歲前居住地(208 人)



四縣腔(含南四縣)發音員現居地(208 人)

海陸腔語音辨識語料 300 小時

HAT-ASR-Hailu-Reading

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 語料庫內容

語料庫總計 151 位發音員，總時長約 300 小時。

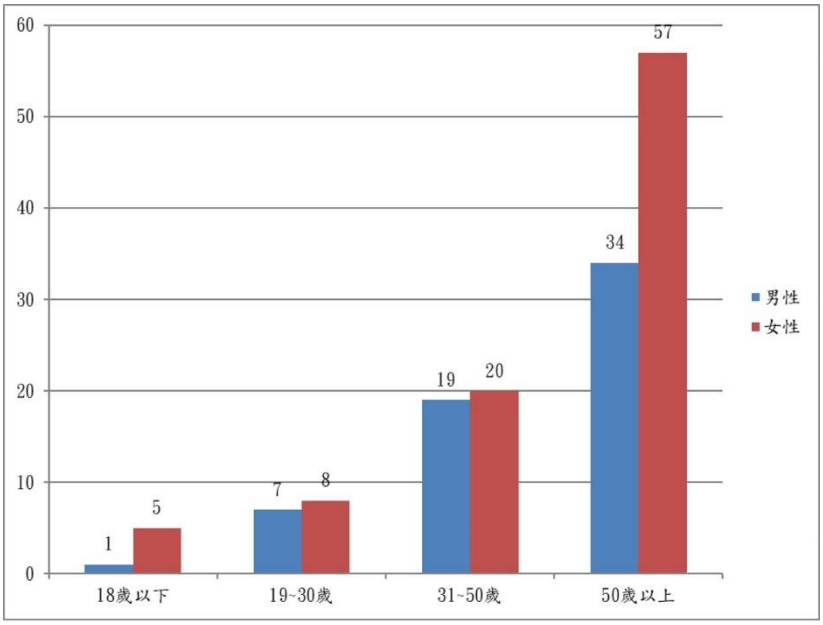
三、處理說明

資料庫為客語朗讀語料（reading speech），是以原生客語文本，收集來自臺灣各地的海陸腔語音，錄製環境為安靜密閉的空間，並同時以 8 支麥克風進行錄製完成。錄製的客語語音，經由客語老師兩次人工校正文本後，整理成可供語音辨認技術研究與開發使用之語音語料庫。

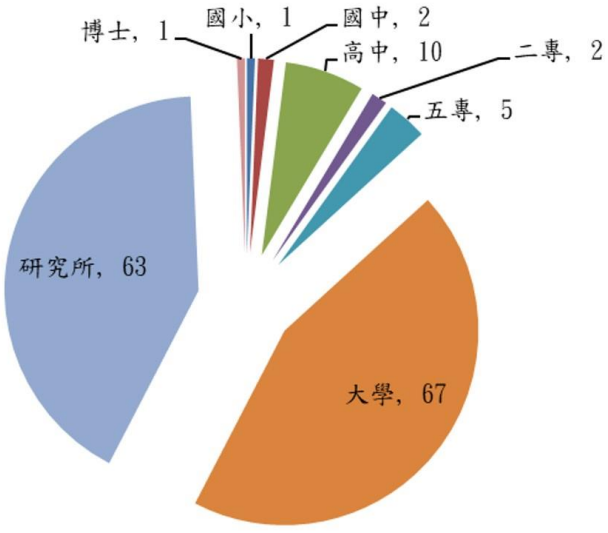
四、音檔格式

副檔名	Channels	Sample Rate	Precision	Sample Encoding
wav	8	16000Hz	24-bit	24-bit PCM

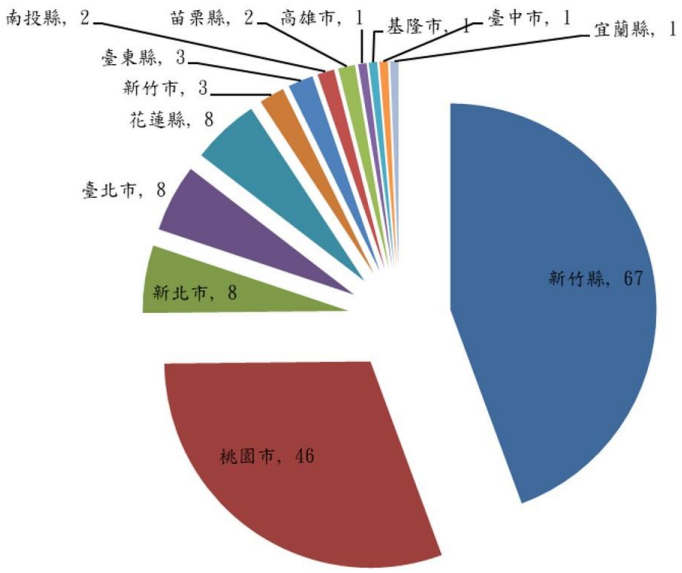
五、語者資訊



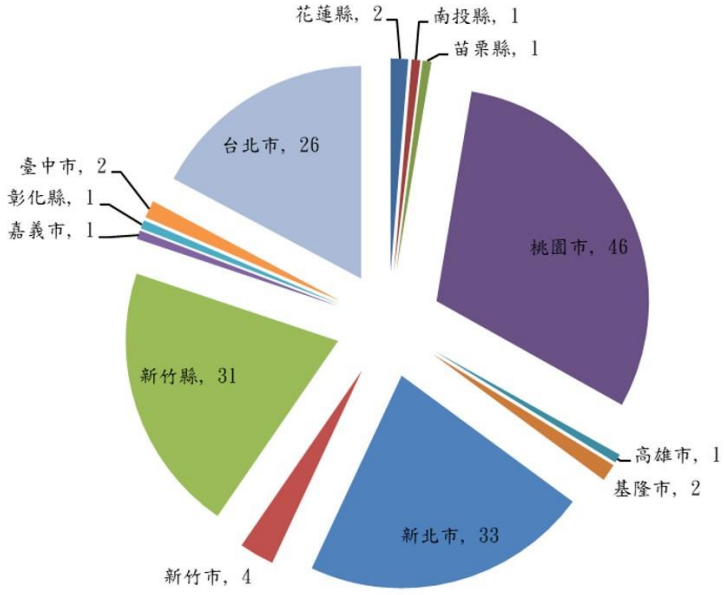
海陸腔發音員性別及年齡（151人）



海陸腔發音員教育程度(151人)



海陸腔發音員18歲前居住地(151人)



海陸腔發音員現居地(151人)

四縣腔語音合成語料 60 小時

HAT-TTS-Sixian

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 語料庫內容

語料庫為男性及女性各 1 位發音員，總時長約 60 小時。

三、處理說明

本語音資料庫為客語朗讀語料（reading speech），是以原生客語文本，男女性各 1 位發音員，於無噪音影響之專業錄音室內錄製，每人錄製約 30 小時。主要目的是要建置客語語音合成語料庫，作為研發人工智慧之基礎建設，尤其是針對基於深度學習之語音合成研究資料之要求。

四、音檔格式

副檔名	Channels	Sample Rate	Precision	Sample Encoding
wav	1	48000Hz	25-bit	32-bit PCM

五、語者資訊

性別	年齡	教育程度	18 歲前居住地	現居地
男性	53	大學	苗栗縣造橋鄉	新北市三重區
女性	64	研究所	新北市中和區	新北市永和區

海陸腔語音合成語料 60 小時

HAT-TTS-Hailu

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 語料庫內容

語料庫為男性及女性各 1 位發音員，總時長約 60 小時。

三、處理說明

本語音資料庫為客語朗讀語料（reading speech），是以原生客語文本，男女性各 1 位發音員，於無噪音影響之專業錄音室內錄製，每人錄製約 30 小時。主要目的是要建置客語語音合成語料庫，作為研發人工智慧之基礎建設，尤其是針對基於深度學習之語音合成研究資料之要求。

四、音檔格式

副檔名	Channels	Sample Rate	Precision	Sample Encoding
wav	1	48000Hz	25-bit	32-bit PCM

五、語者資訊

性別	年齡	教育程度	18 歲前居住地	現居地
男性	56	研究所	新竹縣竹東鎮	新竹縣竹東鎮
女性	54	大學	新竹縣竹東鎮	台北市大安區

四縣腔媒體語料 280 小時

HAT-ASR-Sixian-Broadcast

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 語料庫內容

語料庫為收錄客家之電視台及廣播電台，以生活、文化、新聞為主題之四縣腔音檔，總時長共計 280 小時。

三、處理說明

收錄音檔為電視台及廣播電台於無噪音影響之專業錄音室錄製，刪除音檔中空白及背景音樂部分，續以校正完成之媒體音檔及逐字稿轉換至客語語音資料庫格式，產出具metadata之標準化電子語音語料庫格式，音檔標註 json 檔(Unicode 編碼，儲存 metadata)，此口語語音資料，可以使人工智慧開發團隊或廠商，開發出更貼近生活用語之語音辨識系統。

四縣腔客華平行辭庫 50,000 詞

HAT-Lexicon-Sixian

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 平行辭庫內容

四縣腔客華平行辭庫實際數量為 50,000 詞，內容包含客語漢字及四縣腔拼音。辭庫收錄教育部國語辭典部分用詞、客語能力中高級認證、臺灣街道名及知名地標、常見生活詞彙(單位、日期、時間)、新聞常見用詞等等。

三、 處理說明

將華語辭典經由客語老師對譯為客語，包含客語漢字及四縣腔拼音，建立人工智慧使用之華客語平行辭庫。

海陸腔客華平行辭庫 50,000 詞

HAT-Lexicon-Hailu

一、 簡介

為了讓客語在 AI 時代綻放新光彩，客家委員會致力於打造一個創新的里程碑—「臺灣客語語音資料庫」。該計畫旨在收集和整合客語的龐大語音數據，不僅保存了珍貴的客語語音文化，更是推動語音 AI 技術和產業發展的強力動力。語料庫的目標是，讓客語與人工智慧的浪潮同步，開啟人機對話的全新篇章；在智慧科技的領域中發揮關鍵作用，並堅實其在 AI 應用世界中的基石。

二、 平行辭庫內容

海陸腔客華平行辭庫實際數量為 50,000 詞，內容包含客語漢字及海陸腔拼音。辭庫收錄教育部國語辭典部分用詞、客語能力中高級認證、臺灣街道名及知名地標、常見生活詞彙(單位、日期、時間)、新聞常見用詞等等。

三、 處理說明

將華語辭典對譯為客語，包含客語漢字及海陸腔拼音，建立人工智慧使用之華客語平行辭庫。