國立台北大學資訊工程學系專題報告

AI 判別與全自動影片剪輯

1.摘要

近年來，教學影片與自媒體影片爆發式增長，然而大量影片的錄製，其中品質卻良莠不齊，包括面對鏡頭講話結巴，錄製的影片講話斷斷續續或是口齒不清，這些狀況包括重複、停頓等問題，因此需要剪輯這些影片片段。傳統剪輯方式為人工進行分割再刪除，此過程極耗時與所費不貲。有鑑於此，我們利用深度學習中語音模型 Whisper 與大語言模型 BERT 開發一款自動剪輯軟體來濃縮影片長度並刪除冗餘的片段。該軟體先使用 Whisper 模型，將影片說話音訊轉換成的逐字稿，並將逐字稿的每個字對齊影片的時間軸，接下來會同時標出空白或停頓影片片段並刪減，再透過相位聲碼器疊合聲音，將發音過長的影片壓縮，防止聲音頻率變化，更進一步地，我們利用 BERT 模型開發能自動刪減冗詞的演算法，它首先將逐字稿經過微調訓練後的大語言模型進行移除逐字稿中的冗詞贅字，最後再將剪輯後的逐字稿對應到原始影片的剪輯，進而輸出對應的剪輯影片，達到精簡影片長度的目的。

2.簡介

現在有很多教授上課會錄影，我們看影片的時候會遇到一些問題，教授講話講太慢，講話口齒不清，講話有發語詞或是冗詞贅字，例如: 呃...這個、啊啊、然後、就是就是，其實其實，對等等。思考的時間停頓太久。這些影片片段，跳過怕會錯過內容，等待又會浪費時

間，一堂課上課 50 分鐘，有 5~10 分鐘的時間被消磨，如果能夠偵測並刪除這 10 分鐘的無效教學影片內容，上課時間就可以壓縮到 40 分鐘，影片再放上字幕，用兩倍速觀看，能夠在 20 分鐘內看完，一門三小時的課有機會在一小時內上完，如此可以大大地增加學習效率。另外，傳統的人力剪輯方式的問題，必須用人耳去辨別聲音斷點，才能將不需要的影片剪除，繁冗的工作需要耗費時間跟精力。一般的教授或創作者沒有這麼多精力與經費去做細微調整與精簡。

我們的研究計畫預計設計一個基於深度學習語音識別技術的自動化剪輯工具系統程式。如圖 1 所示，該系統會將影片的語音對齊中文字的長度，影片聲音精準切割到微秒，把編輯影片像編輯 word 檔案一樣輕鬆，將想要的部分直接從文字刪除，直接對應影片的片段。該技術可以提高編輯影片的效率和內容品質。我們的研究專注於去除冗言贅句，針對中文文法進行優化處理，以確保剪輯內容的準確性和流暢性。對於忙碌的教育工作者、自媒體，甚至是日常生活中需要快速處理大量語音信息的普通用戶來說，這樣的工具將大大提高他們處理信息的效率，節省時間。同時可以拉近大眾加入自媒體與教育業的距離，降低進入自媒體的門檻與時間成本。

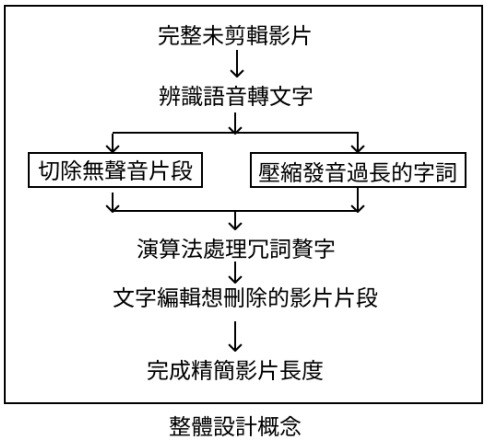


圖 1：本計畫開發系統整體設計概念

3.專題進行方式

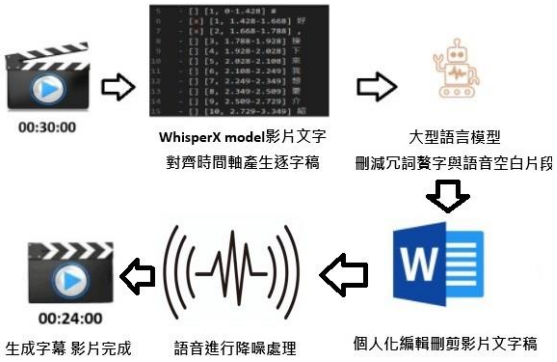


圖 2 :本專題流程圖

我們的專題為以下的流程

1. 使用 WhisperX 轉換影片：
   將影片轉換成逐字稿，並與時間軸對齊。
2. 自動化刪減冗詞贅字：
   使用大型語言模型 BERT，對逐字稿進行自動化處理，刪減冗詞贅字。
3. 個人化編輯影片文字稿：
   對逐字稿進行個人化編輯，確保內容精確且符合需求。
4. 語音降噪：
   對影片進行語音降噪處理，提升音質。
5. 生成字幕：
   根據編輯後的文字稿生成字幕，完成影片製作。

3.1 Whisper X

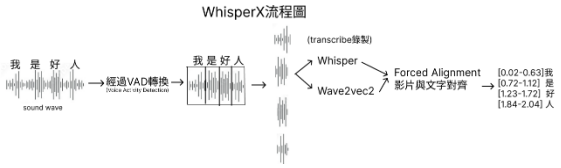


圖 3: 音訊轉文字時間軸架構圖（whisperx 流程圖）

如圖 2 所示，WhisperX (Bain et al. 2023.)[1] 使用 wav2vec2 對齊的準確字級時間戳，時間戳記是單字級別，Phoneme-Based ASR 經過微調的模型，識別區分一個單字和另一個單字的最小語音單位，可以到微秒等級，我們可以勾選或剃除想要的中文字(如圖 5)，對應到影片的內容，經過我們測試對多語言都有不錯的表現，但應用在中文時或許是因為訓練資料來源多為百度

[1] https://arxiv.org/abs/2303.00747

等論壇，套用台灣口音的影片結果並不理想，因此在使用時我們在 wav2vec2 階段使用了不同的模型。

```
- [] [1, 0-1.428] #
- [x] [1, 1.428-1.668] 好
- [x] [2, 1.668-1.788] ，
- [] [3, 1.788-1.928] 接
- [] [4, 1.928-2.028] 下
- [] [5, 2.028-2.108] 來
- [] [6, 2.108-2.249] 我
- [] [7, 2.249-2.349] 想
- [] [8, 2.349-2.509] 要
- [] [9, 2.509-2.729] 介
- [] [10, 2.729-3.349] 紹
- [x] [11, 3.349-3.569] 在
- [x] [12, 3.569-3.689] 介
- [x] [13, 3.689-3.81] 紹
- [x] [14, 3.81-3.93] 物
- [x] [15, 3.93-4.09] 理
- [x] [16, 4.09-4.25] 之
- [x] [17, 4.25-4.41] 前
```

圖 5: 經 DTW 算法後的路徑圖

3.2 WhisperX 模型架構

3.2.1 影片的語音轉文字

使用工具：使用 Whisper large-v2，這是一種自動語音識別(ASR)模型。

詳細過程：此模型將影片中的語音內容轉譯為中文文字。它生成一系列的文字標記（token），這些標記代表語音中的詞彙，但是它有個缺點是沒辦法獲取準確的文字時間戳記。

3.2.2 取得每個文字對應的語音時間軸

透過 wav2vec2 模型，他可以辨別音訊的細微變化，生成音速序列，舉例來說英文字母的 tap ，該模型可以辨別 p 的起始時間以及結束時間，也就是音速序列，將 whisper 的文字序列以及 Wav2vec2 生成的音素序列結果做 forced alignment 可以得到每個文字的時間戳記。

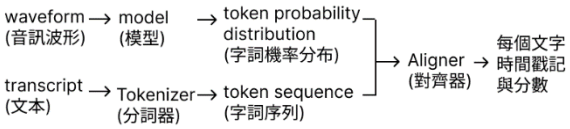


圖 4：強制對齊流程圖

其中(圖 4)的 Aligner 主要基於 Forced Alignment with Wav2Vec2

的 Pytorch 程式[2]，使用微調後的 wav2vec2 模型[3]，達成 token probability distribution。根據 Dynamic Time Warp 算法，找出最有可能的路徑，該路徑代表每個時間戳記最有可能的 token，如圖 5 所示。

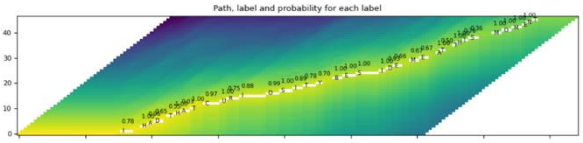


圖 5: 經 DTW 算法後的路徑圖

3.3 自動化處理贅字

有了每個文字的時間戳記，我們即可快速刪減影片內容。

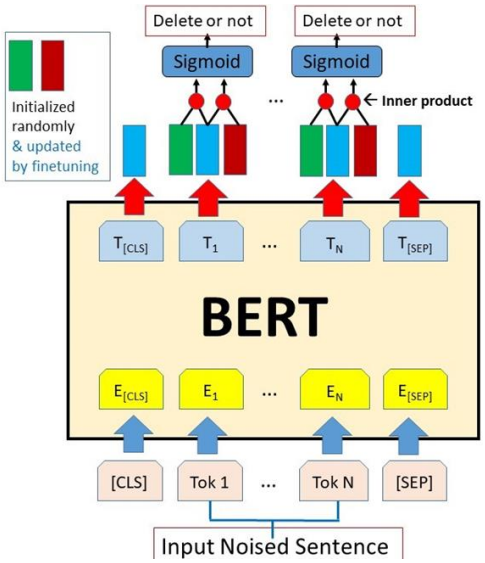
為了自動化刪減贅字已達成節省大量人力刪減時間的目標，我們採用了 fine tuning bert 的方法實作。我們的訓練資料有 5000 個贅句以及 2000 個非贅句。模型架構主要基於 bert-chinese，再加上 dropout layer 避免過度擬合，最後加上一個二元分類器即可達成自動化刪減贅字目標，bert 架構如圖 6 所示。

為什麼我們選用 bert?因為 bert 是 NLP 領域廣泛使用的模型。架構是採用 transformer 的 encoder。可以正確理解句意，以達成刪減贅字的目標。

為什麼不用 chatGPT?由於 chatGPT 會曲解句子的原意、新增或修改句子中的內容，無法達成只刪減冗詞贅字的目標。

[2] https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html

[3] https://huggingface.co/ydshieh/wav2vec2-large-xlsr-53-chinese-zh-cn-gpt



圖(6)bert 模型架構圖

bert 模型經過微調訓練。輸入句子，bert 可以辨別句子中的贅字與非贅字，例如:我向前走到前面去。模型辨別出贅字為向前以及面，輸出句子為:我走到前去。

由此可知模型有理解句子中的內容而非一味刪除重複的字詞。訓練成果如圖（7）所示。

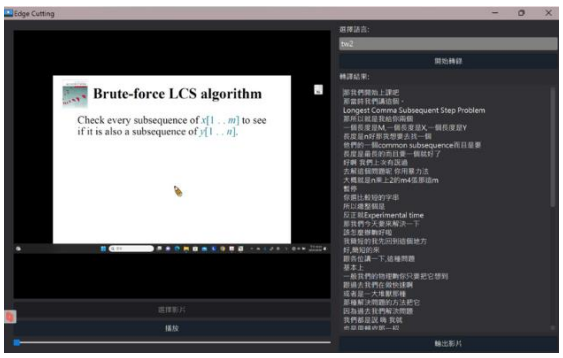
```
PS C:\Users\s4110\112project_combine\bert> python test.py
['我', '向', '前', '走', '到', '前', '面', '去']
[CLS]: 0
我: 0
向: 1
前: 1
走: 0
到: 0
前: 0
面: 1
去: 0
[SEP]: 0
PS C:\Users\s4110\112project_combine\bert>
```

0: 非贅字　1:贅字

圖 7: 刪減贅字後的結果圖

3.4 編輯逐字稿對應的影片內容

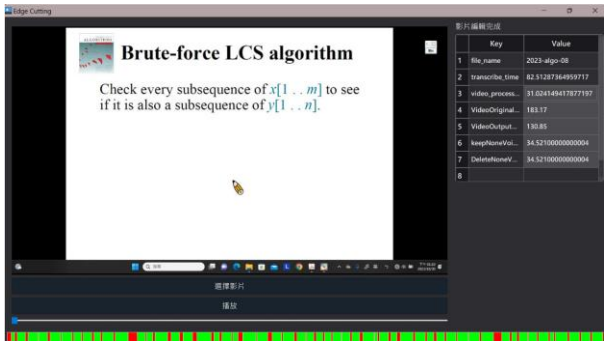
經過 bert 自動化篩檢文字內容後，若有不夠滿意的片段，我們可以透過文字編輯器來刪減想要去除的字詞，並直接對應影片中的片段。編輯影片可以如同編輯文字檔一樣簡單迅速，只需要按 backspace，文字刪除，影片也一併刪除。代替使用剪輯軟體，手動分割影片並刪除的步驟。我們的使用介面如下圖(8)。



圖(8)個人化影片編輯介面

4. 主要成果與評估

使用此軟體，快速辨別影片中的瑕疵片段，如下圖，綠色的片段為保留的影片，紅色為剪除的瑕疵或空白片段。



圖(9)

影片濃縮的多寡與講者本身的講話習慣密切相關。本軟體以演算法教授上課影片為例，以下為演算法 3 分鐘的影片資料分析。

	Key	Value
1	file_name	2023-algo-08
2	transcribe_time	82.51287364959717
3	video_process...	31.024149417877197
4	VideoOriginal...	183.17
5	VideoOutput...	130.85
6	keepNoneVoi...	34.52100000000004
7	DeleteNoneV...	34.52100000000004
8		

圖(10)影片內容資料分析

如上圖(10)，影片長度為 183 秒，剪輯完成後長度為 131 秒，刪減長度為 52 秒，濃縮比例為 71%。其中，刪減空白長度 35 秒，刪減錯誤字詞內容為 18 秒。

5. 結語與展望

傳統的影片剪輯方式分為粗剪與定剪，粗剪的方法為將原始影片剪除瑕疵片段與冗詞贅字，整理成可以使用的素材。定剪為將素材合併與加工，完成一部影片。粗剪的人力成本，時薪至少為 200 元起跳，定剪一部 10 分鐘影片則高達 3000 至 5000 元起跳。

目前市面的影片剪輯，可以做到一段影片對應的文字與時間軸，然而無法精確處理一個字的時間切割。我們的客群專攻在粗剪的市場上，替代繁冗的工

作，將低粗剪的成本。

我們的設計可以幫助自媒體或是製作教學影片，處理細微的影片瑕疵，例如刪除結巴的句子，剪去常見且冗餘的口頭禪，移除錄製影片中間產生的停頓或空白片段。預期影片長度縮減 10%~30%，提高繁冗的影片編輯效率，剪輯影片如同編輯 word 檔一樣直觀簡易。一門 50 小時的影片內容可以快速剪輯並縮短影片內容，用低成的方式製作影片。我們希望將這項技術推廣到數位化教學與自媒體上。同時，這項創新可以使廣大學生受益，對於沒有時間剪片的老師更是一項福音。

6. 銘謝

感謝指導教授提供我們所需的設備，並在我們遇到問題時，能在會議中與老師進行詳細的討論，研究各種解決方法，這對我們的學習和研究進展非常有幫助。您的指導和支持對我們來說是無價的，我們非常感激。

7. 參考文獻

[1]Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. “MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction. In Proceedings of NAACL-HLT"

[2]Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever "Robust Speech Recognition via Large-Scale Weak Supervision"

[3] Max Bain, Jaesung Huh, Tengda Han, Andrew Zisserman "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio"

[4] Yue Zhang1, Haochen Jiang1, Zuyi Bao2, Bo Zhang2, Chen Li2, Zhenghua Li1 "Mining Error Templates for Grammatical Error Correction"

[5] Sagar Ailani, Ashwini Dalvi, Irfan Siddavatam "Grammatical Error Correction (GEC): Research Approaches till now"

[6] mli/autocut:

https://github.com/mli/autocut

[7]

https://github.com/destwang/CTCR

esources#datasets