The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 21-36

基於卷積類神經網路之廣播節目音訊事件偵測系統

Automatic Audio Event Detection of Broadcast Radio Programs Based on Convolution Neural Networks

陳智偉 Jhih-wei Chen, 許吳華 Wu-Hua Hsu, 廖元甫 Yuan-Fu Liao

國立臺北科技大學電子工程系

Department of Electronic Engineering, National Taipei University of Technology

t104368109@gmail.com, asmayday24@gmail.com, yfliao@ntut.edu.tw

摘要

廣播電臺節目中通常包含語音，音樂與其他音訊事件（如笑聲或特效聲）。若能偵測並切割這些音訊事件，就能進一步對廣播節目進行加值運用。例如，轉寫語音片段的逐字稿，或是辨認音樂片段的歌名與曲名，以利檢索。針對此問題，在本論文中，我們首先設計，並以人工標註出一廣播節目音訊事件資料庫，再利用 Convolutional Neural Network (CNN)自動擷取有效的特徵音訊參數，對廣播電臺的音檔做音訊事件偵測與切割，最後轉成具時間資訊的音訊事件標註檔。實驗方面我們從教育電臺節目中，選出新聞類與不同性質的談話類節目共 14 個，經人工標注後，獲得總長度共約 60 小時的音檔，並用來訓練與測試 CNN 和傳統 Gaussian Mixture Model（GMM）的效能。實驗結果顯示以 CNN 直接搭配頻譜參數，在偵測語音與非語音，音樂與非音樂或其它與非其它音訊事件等的錯誤率（equal error rates，EER），分別為 2.27%、12.52%與 9.51%，皆低於傳統以 GMM 搭配 Mel-Frequency Cepstral Coefficients（MFCCs）的 3.65%、15.68%與 13.25%。

關鍵詞：廣播節目資料庫、音訊事件偵測、卷積類神經網路。

1. 簡介

過去在網路（尤其是行動網路）還不夠普及的時候，用收音機收聽廣播電臺節目是人們主要的資訊與娛樂來源。但現今網路已經非常發達，許多人皆改成直接用手機觀賞電視與電影等視頻類節目。因此大部分的廣播電臺也積極轉型因應，除設立網路廣播電臺，線上即時（online）直播節目內容，吸引年輕聽眾收聽外。更試著把播過的節目內容，離線（offline）放在網路上，建立電臺的廣播節目典藏庫（archive），供聽眾自由安排時間收聽任何節目，以獲取更多元的聽眾來源。

但是，因傳統上廣播節目是聽覺媒體，且具有傳播速度快，時效性強等特性。目前多數電臺的廣播節目典藏，大都只有保存節目音檔本身，少有付加與節目內容相關的後設資料（metadata）以利查詢。因此聽眾若想要查詢與檢索歷史節目時，就相當不容易，尤其是常常無法直接找到聽眾最有興趣的某節目中的某段重要的內容。

因此本論文嘗試偵測與切割廣播節目中的多種音訊事件[1][2]，因為，廣播電臺節目中通常包含語音，音樂與其他音訊事件（如笑聲或特效聲）。若能偵測並切割這些音訊事件[3]，就能進一步組織廣播節目內容進行進一步加值。尤其是若能自動轉寫每一語音片段的逐字稿，擷取出關鍵字與摘要，或是自動辨認出每一音樂片段的歌名或曲名，就能讓聽眾直接以文字進行全文檢索，找到相關節目內容，或是以哼唱方式找到想聽的音樂歌曲段落。

針對此問題，本論文的整體處理程序如圖 1 所示，主要的想法包含（1）建立廣播節目音訊事件資料庫，如圖 1（a）;（2）訓練音訊事件模型，如圖 1（b）;（3）進行自動廣播節目音訊事件偵測與標記，如圖 1（c）。

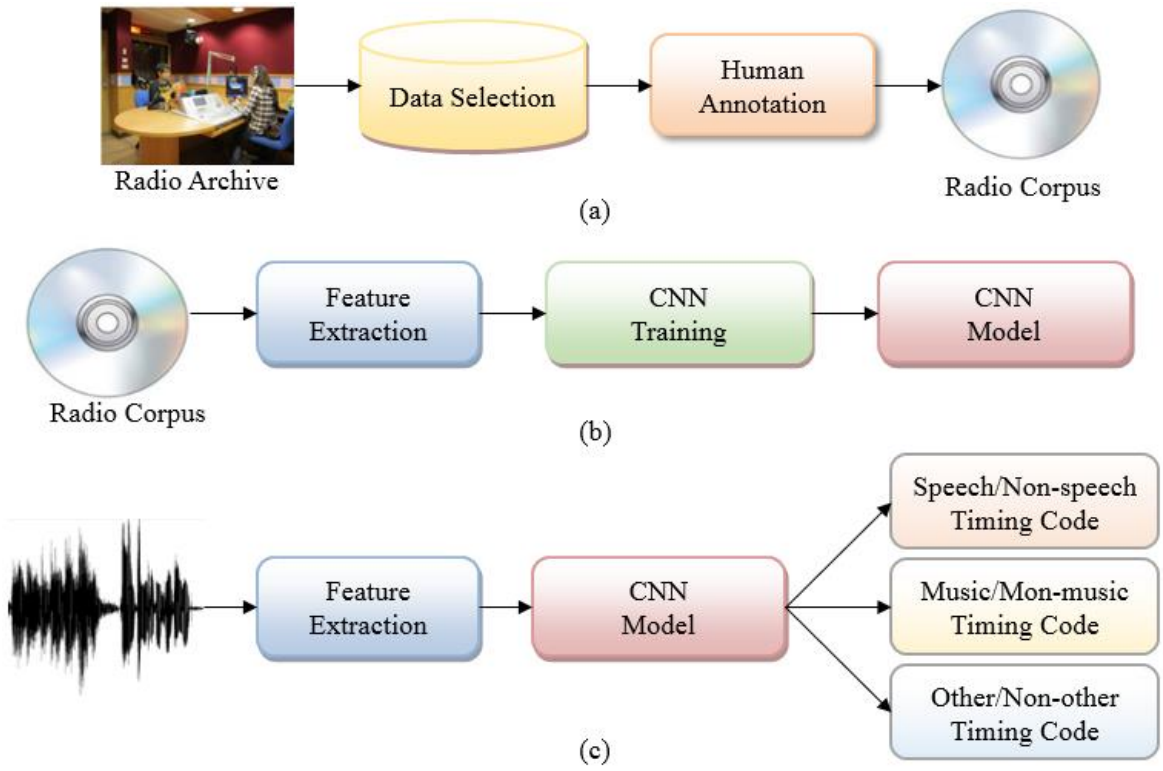

圖 1 音訊事件偵測系統架構圖

其中廣播節目音訊事件資料庫建立，將蒐集並對電臺節目做分類，再以人工進行標註，找出語音，音樂與其他音訊事件的起始與結束時間。音訊事件模型訓練，則是利用人工切割出之不同音訊事件的樣本與答案集，分別建立語音與非語音，音樂與非音樂，其他與非其他等音訊事件的模型。最後自動廣播節目音訊事件偵測與標記，就是利用所訓練出之三種音訊事件模型，偵測輸入之廣播節目中的各種音訊事件，並將其轉成與節目音檔相對應，具時間資訊的音訊事件標注檔。

本論文將使用 Convolutional Neural Network (CNN)[4]架構來完成我們的音訊偵測切割系統，一般音訊事件偵測系統競賽都是以傳統 GMM+MFCCs 架構來訓練，如 DCASE2016 Challenge[5]，我們將使用 CNN 架構與典型的 i-vector 系統 Kaldi speech recognition toolkit[6]中 Bn_music_speech 提供的一個 GMM 基礎作法做比較，主要是因為 CNN 具有以下特性：(1) 對音訊事件在輸入參數序列中的位置，具有時間與頻帶上的平移不變性，可以容忍音訊事件在時間與頻譜上的變異、(2) 能自我訓練如何擷取最佳化的音訊事件特徵參數，因此可以避免需專業知識，才能設計出適合的音訊參數的參數工程（feature engineering）問題。而能直接輸入頻譜參數，讓 CNN 自動去探索，除

了傳統 MFCCs 外，還有哪些特徵參數對音訊事件偵測效能最好。

2. 廣播節目音訊事件資料庫

2.1. 廣播節目資料搜集與設計

我們首先將廣播節目類型分為純語音、語音+音樂、語音+較多音樂三大類[5]，每一類節目各挑選多集節目。表 1 為廣播節目類型分類，與挑選出的節目與其長度。

表 1 廣播語料庫統計資料

類型	節目名稱	集數	挑選時長 (minute)	類型	節目名稱	集數	挑選時長 (minute)
純語音	創青宅急便	10	384	語音+音樂	創設市集 On-Air	10	267
	自然有意思	8	157		教育開講	10	261
	科學 SoEasy	10	261		今天不補習	10	298
	特別的愛	10	270		兒童新聞	10	98
	多愛自己一點點	10	289		文教新聞	10	68
	國際教育心動線	10	210	語音+較多音樂	從心歸零	10	540
	青年故事館	10	286		技職最前線	10	213

各挑選多集節目原因在於，教育廣播電臺的節目相當多元，音檔可能包含說話聲、音樂和特效等等。為了能夠讓所有情況都能夠收集到，因此我們需要拿多種不同類型的節目進行挑選，才能涵蓋所有可能情況。

2.2. 音訊事件人工標記規範

我們考慮在一段音檔裡，會有多種不同類型的音訊事件，且發生期間可能會重疊，或是多種事件一併發生的狀況。因此將標注準則設為一類一軌，各自獨立標注（如圖 2 之規範示意圖），以建立音訊事件資料庫。

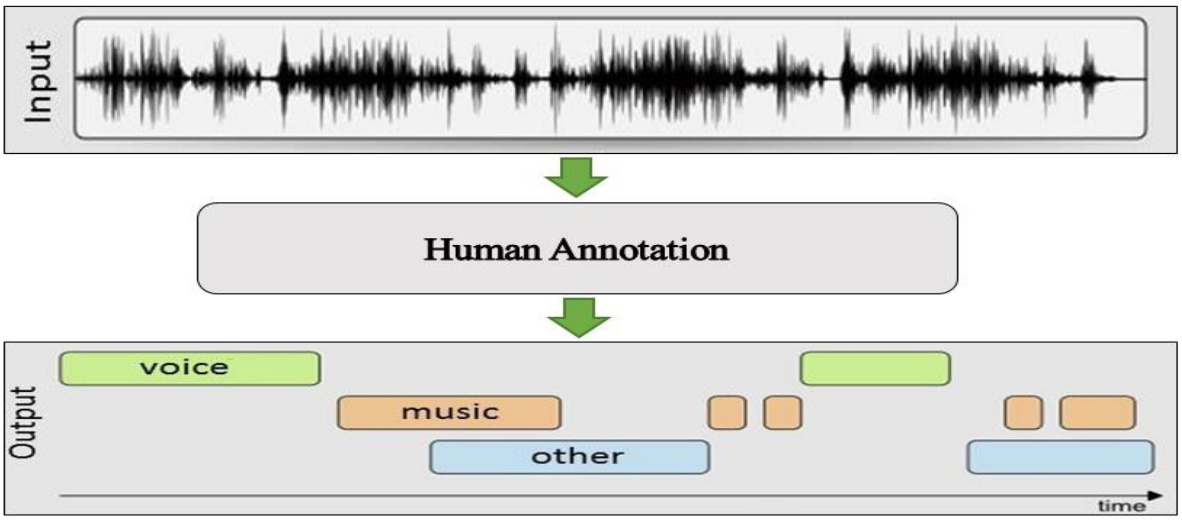

圖 2 音訊事件人工標誌的示意圖

標注程序，則是利用 Praat 軟體，先建立三軌標註面板，再依照下列規則標記。(1) 語音部分：只要音檔中有人講話的部分，且聽得出講話內容，皆標記成為語音（如：主持人或來賓講話或是 Call in 的民眾)。(2) 音樂部分：只要音檔有音樂，且聽得出音樂內容的話，皆標為音樂（如：高中生合唱團演唱、樂團表演以及流行音樂...等)。(3) 其它部分：則是在音檔內容出現笑聲或特效聲，皆標示為其它的部分。圖 3 為實際人工標註結果的範例。

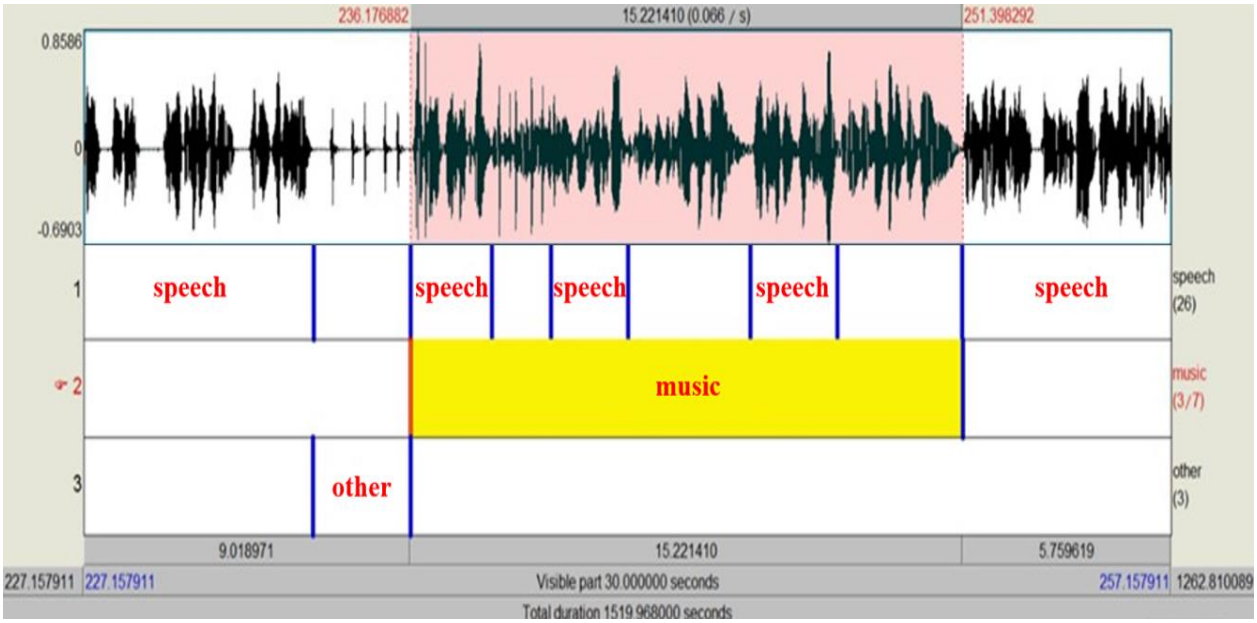

圖 3 Praat 人工標記音訊事件結果範例圖

2.3. 標記結果統計

表 2 則是進行人工標記後，各種標記出的音訊事件的時間長度統計資料。其中，語音事件最多，總數約為 3000 分鐘，音樂事件次之，約有 700 分鐘，其他事件最少，只有約 25 分鐘。因此資料分佈相當不平衡。

表 2 人工標記音訊分類時間總表(minute)

Program	speech	non-speech	music	non-music	other	non-other
創設市集 On-Air	222	45	84	180	0.86	264
教育開講	234	27	50	17	0.66	264
今天不補習	246	52	63	6	2	294
兒童新聞	92	6	7	92	0.4	99
技職最前線	186	27	72	114	3	150
從心歸零	372	168	52	342	7	354
特別的愛	270	0	0	300	5	294
創青宅急便	384	0	0	384	1	384
自然有意思	156	1	1	138	1	156
科學 SoEasy	246	15	60	198	0.78	204
文教新聞	61	7	9	60	0.66	68
多愛自己一點點	204	85	234	54	0.86	224
國際教育心動線	210	0	0	210	0.33	210
青年故事館	252	34	55	206	0.72	212
Total	3135	467	687	2301	25	3177

3. 基於 CNN 之音訊事件偵測系統

3.1. 特徵參數選擇與音訊模型訓練

在音訊事件偵測特徵參數部分，傳統上普遍使用的參數為 MFCCs。但其實對音訊事件偵測，MFCCs 不見得是最佳的。尤其是對音樂與其他事件，還有很多不同的參數被提出來。

由於，我們不能確定哪一種特徵參數，能夠有最好的音訊事件偵測效果，所以我們求取並測試各種不同參數的音訊事件偵測效能。因此在本論文中，將嘗試如圖 4 所示的架構，尤其是測試使用 Spectrum（Specgram），Mel-spectrum (Mels)與 MFCCs，分別進行語音與非語音，音樂與非音樂，其他與非其他音訊事件偵測模型。

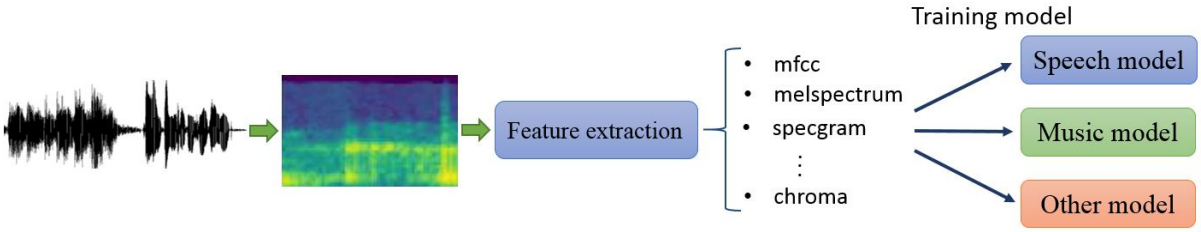

圖 4 選用不同的特徵參數示意圖

3.2. 卷積類神經網路模型架構

傳統類神經網路模型，都是先以人為方式，設計好要使用的特徵參數，然後直接採用全連接的 deep neural networks（DNNs）訓練模型。但由於我們不能確定哪一種特徵參數，能夠有最好的音訊事件偵測效果，所以本論文改使用卷積神經網絡（Convolutional Neural Networks, CNNs）架構來做音訊偵測切割系統，其架構如圖 5。此外，由於不同音訊事件可能會同時發生，所以每種音訊事件(speech、music and other)，都需要獨立建立一個模型，然後個別運作做偵測。

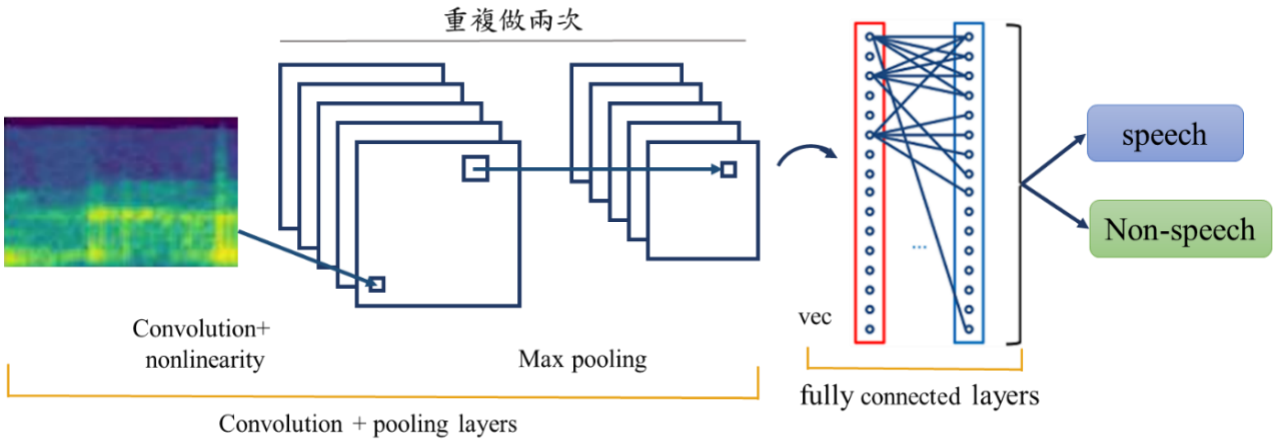

圖 5 CNN 音訊事件偵測系統

在此 CNN 網路架構中，有三個主要的神經層，包括（1）卷積層，（2）池化層與（3）全連接層。其中卷積層與池化層可重複數次。使用 CNN 的好處是，CNN 能利用卷積層，自動學習如何求取最佳參數，與利用池化層，容忍目標事件在頻譜上的位置變異。所以我們可以自由嘗試許多不同的特徵參數，尤其是可以不經過人為設計，直接輸入頻譜參數，訓練 CNN 模型。以下說明卷各層的運作方式。

3.2.1. 卷積層

卷積層可包含許多卷積核，其運作如下圖 6 所示（以二維輸入參數為例），主要是

每個卷積核會通過一滑動窗口（sliding window），掃描上一級輸入的參數，逐步計算其與卷積核的內積，輸出一卷積特徵參數圖（feature map）。因此，一個卷積核就相當於一個多維度匹配濾波器，但其權重可經由訓練自動最佳化。

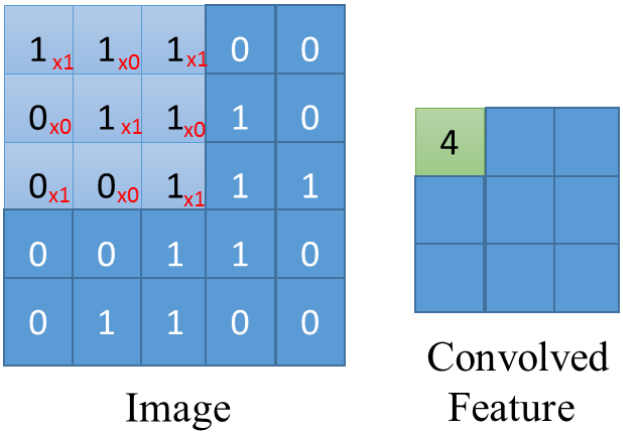

圖 6 CNN 卷積層運作方式

3.2.2. 池化層

池化層（Pooling）[6]接在卷積層之後，其運作如圖 7 所示。主要是將 feature map 劃分成數個區域，並以區域為單位，在每一區域以類似投票方式，只選取此區域中較強的卷積值做輸出，並丟掉較微弱的卷積值。此運作除可降低數據量、減小過擬合，最重要的是，可以容忍目標事件在頻譜上的位置變異。

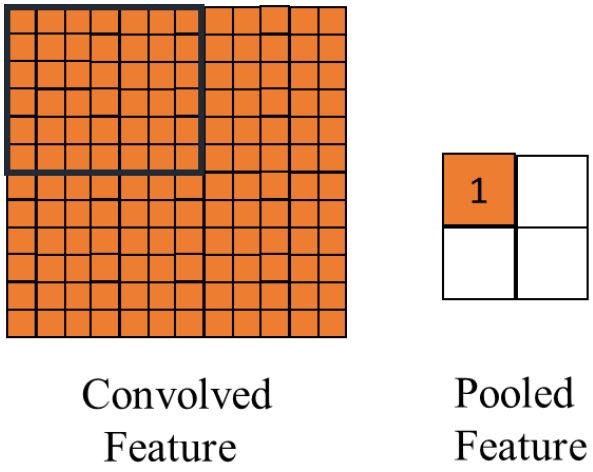

圖 7 池化層

3.2.3. 全連接層

最後則是利用全連接層辨認目標與非目標事件。全連接層的架構就是一個傳統的多層式類神經網路（通常只用兩層），但最後是以 softmax 激發函數，計算目標事件與非目標事件發生的機率值，用以檢測是否有目標音訊事件發生。

4. 音訊事件偵測實驗

4.1. 訓練與測試語料

本實驗將教育電臺節目分成純語音、語音+音樂、語音+較多音樂三種類型，共擷取 14 個節目（長度共約 60 小時），經人工標注後，將其分成訓練用與測試用兩組，用來比較 CNN 的效能。語料內容概況如表 3、表 4 所示。此外為了平衡不同事件的樣本數量，我們在訓練語料組，額外再加上 MUSAN[7]語料，以增加音樂與其他類別事件樣本的數量。

表 3 教育電臺音訊事件訓練語料 (minute)

Program	speech	non-speech	music	non-music	other	non-other
創設市集 On-Air	222	45	84	180	0.86	264
教育開講	234	27	50	17	0.66	264
今天不補習	246	52	63	6	2	294
兒童新聞	92	6	7	92	0.4	99
技職最前線	186	27	72	114	3	150
從心歸零	372	168	52	342	7	354
特別的愛	270	0	0	300	5	294
創青宅急便	384	0	0	384	1	384
MUSAN_Data	390	816	846	0	408	0
Total	2396	1141	1174	1435	427.92	2103

表 4 教育電臺音訊事件測試語料(minute)

Program	speech	non-speech	music	non-music	other	non-other
自然有意思	156	1	1	138	1	156
科學 SoEasy	246	15	60	198	0.78	204
文教新聞	61	7	9	60	0.66	68
多愛自己一點點	204	85	234	54	0.86	224
國際教育心動線	210	0	0	210	0.33	210

青年故事館	252	34	55	206	0.72	212
Total	1129	142	359	866	4.35	1074

4.2. CNN 設定

主要是需要考慮使用的特徵參數、CNN 網路結構大小與選擇適當的優化器。

4.2.1. 特徵參數設定：

在特徵參數方面，我們取了三種參數進行比較，包括（1）MFCCs，（2）Mel-spectrum (Mels)與（3）Raw Spectrum (Specgram)。

4.2.2. 網路配置：

在 batch size 方面，我們測試了 16、32、64、128 筆樣本四種變化，最後將 batch size 設成 64 筆。我們也在參數輸入層加入 dropout，嘗試過丟棄 2%、5%與 10%輸入參數等變化，最後設定為 2%，在卷積層的 dropout 則是設定為 25%。最後在層數方面，我們測試了 2 layers、3 layers、4 layers 與 5 layers，最後皆設成 4 layers。

4.2.3. 選擇優化器：

實驗中嘗試兩種優化器來做測試，一個是 SGD，另一個是 Adadelta。SGD 是指 gradient descent，是最常見的優化方法，但其不會自動調整學習率，須自行嘗試。而 Adadelta 則會對學習率進行自適應約束，使用起來必較方便。因此最後我們選用 Adadelta 優化器。

4.3. 實驗結果

以下分別進行三個實驗，包括，（1）比較 GMM 與 CNN 模型的效能，（2）比較不同參數的影響與（3）求取檢測錯誤權衡曲線（detection error tradeoff curves，DETs）以計算 equal error rate（EER），分別說明如下：

4.3.1. 實驗一-基於 MFCCs 特徵參數之 GMM 與 CNN 效能比較

我們先提取 MFCCs 當做我們的特徵參數，訓練 CNNs 與傳統 GMMs 系統做比較，從表 5 的實驗結果可知，CNN 系統在音訊事件偵測的效果較佳。

表 5 語音/音樂/其它事件偵測準確度

Accuracy(%)	Speech	Music	Other
GMM	97.12	94.88	94.15
CNN	98.46	96.43	97.47

4.3.2. 實驗二-基於不同參數之 CNN 效能測試

從上一個實驗，我們已經驗證了，在音訊事件模型裡，CNN 架構會比 GMM 架構來得有效。在此，我們進一步測試不同參數的效能。實驗結果如表 6 所示，可以看出直接輸入 raw spectrum，就可以得到最佳的辨認結果。

表 6 不同特徵參數的音訊事件偵測準確度

Accuracy(%)	MFCCs	Mels	Specgram
Speech	97.96	97.87	98.46
Music	95.23	96.43	96.28
Other	96.53	96.47	96.22
Average	96.57	96.92	96.99

4.3.3. 實驗三-EER 結果

圖 8 至圖 13 分別為所訓練好的語音、音樂、其他音訊事件偵測器，對訓練資料和對測試資料計算檢測錯誤權衡曲線的結果。其中 y 軸為錯誤拒絕率，x 軸為錯誤接受率。由表 7EER 的結果中，可知語音事件偵測的 EER 最低，而音樂與其他事件的 EER 都較高。這可能是音樂與其他事件的變化較多，在訓練語料中的樣本涵蓋率，還是比較不足。

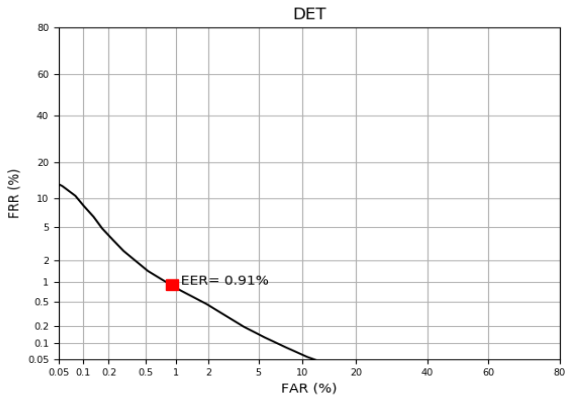

圖 8 語音訓練資料結果

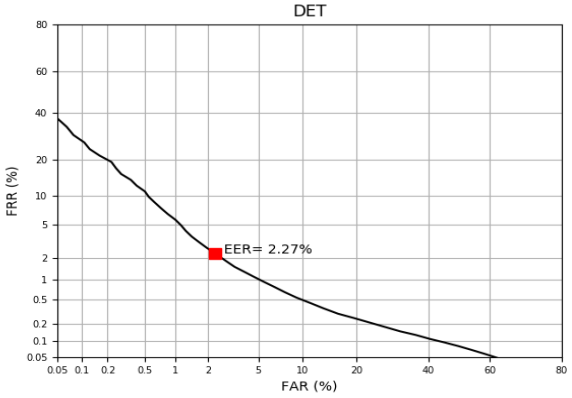

圖 9 語音測試資料結果

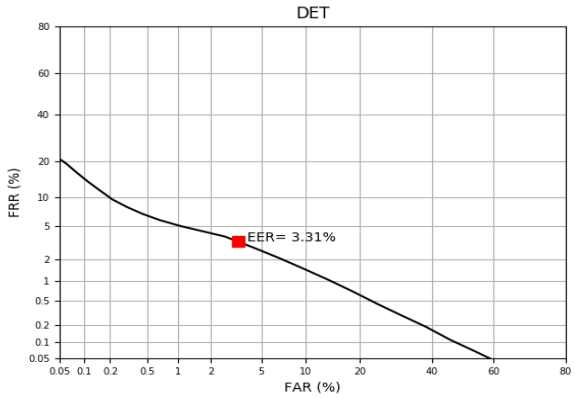

圖 10 音樂訓練資料結果

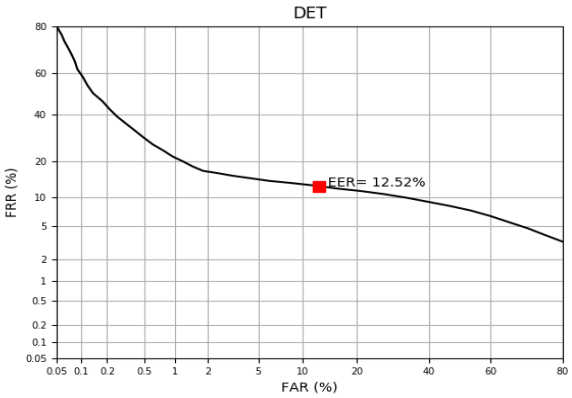

圖 11 音樂測試資料結果

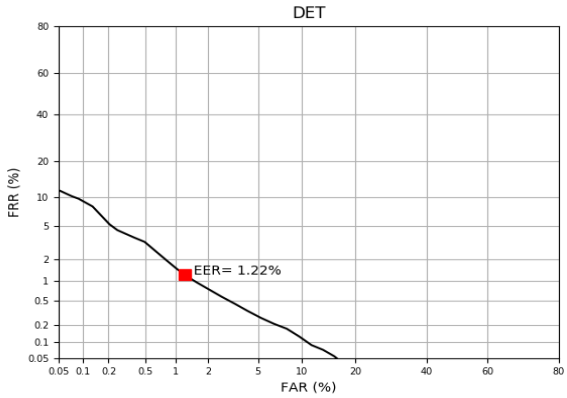

圖 12 其它訓練資料結果

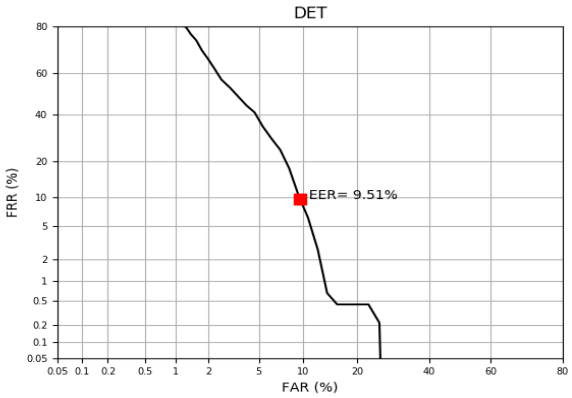

圖 13 其它測試資料結果

表 7 EER 結果

EER(%)	Train	Test
Speech	0.91	2.27
Music	3.31	12.52
Other	1.22	9.51

4.4. 實驗分析與討論

4.4.1. 語音事件偵測

圖 14 是經語音事件偵測器處理過的結果範例圖。圖中第一層為原始音檔波形、第二層為頻譜、第三層為只保留偵測出的語音事件部分的音檔波形、第四層為其頻譜、最下層則是人工所標記的音訊事件標準答案。從圖中可知道語音與音樂頻譜特性相當不同，因此非語音部分明顯可以被正確偵測並拿掉，只保留語音的部分。

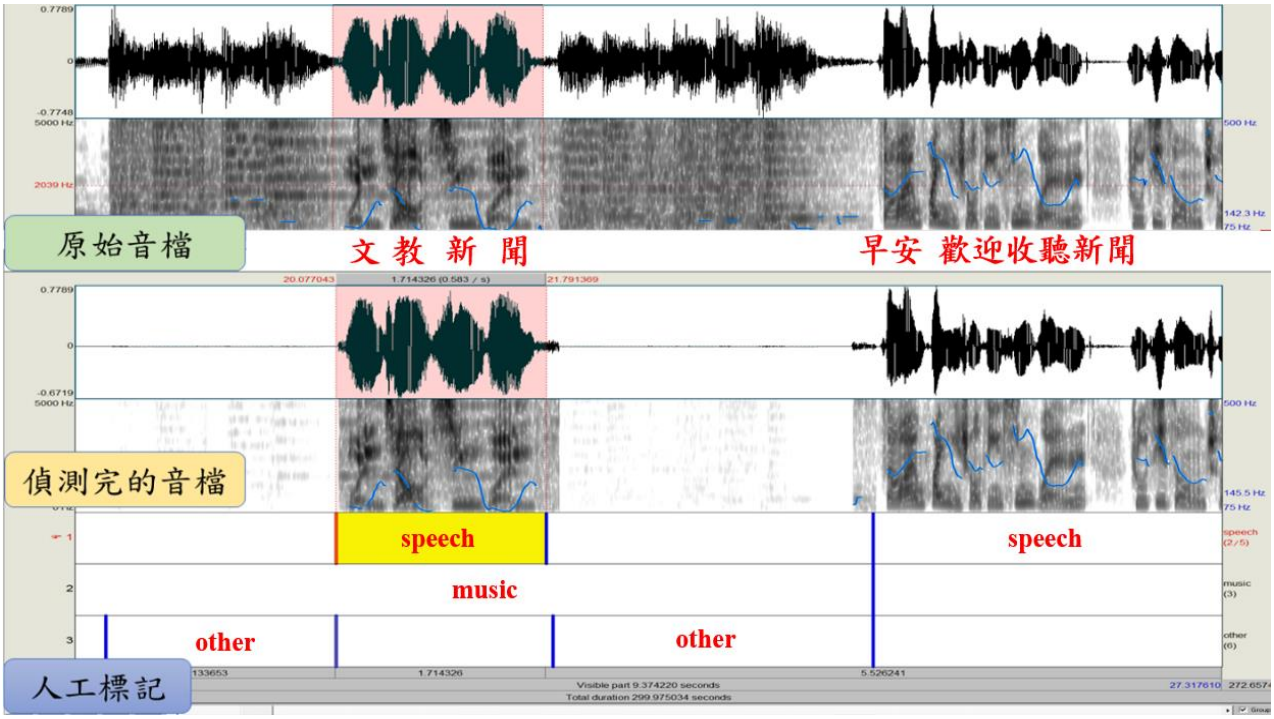

圖 14 語音偵測結果

4.4.2. 音樂事件偵測

從圖 15 可以看到經音樂事件偵測器處理過的結果範例圖。從圖中可知道非音樂部分可以明顯被拿掉，只保留音樂事件的段落。不過比較圖 14 與圖 15，可知在邊界的地方較容易發生錯誤，尤其語音容易被判別成音樂。

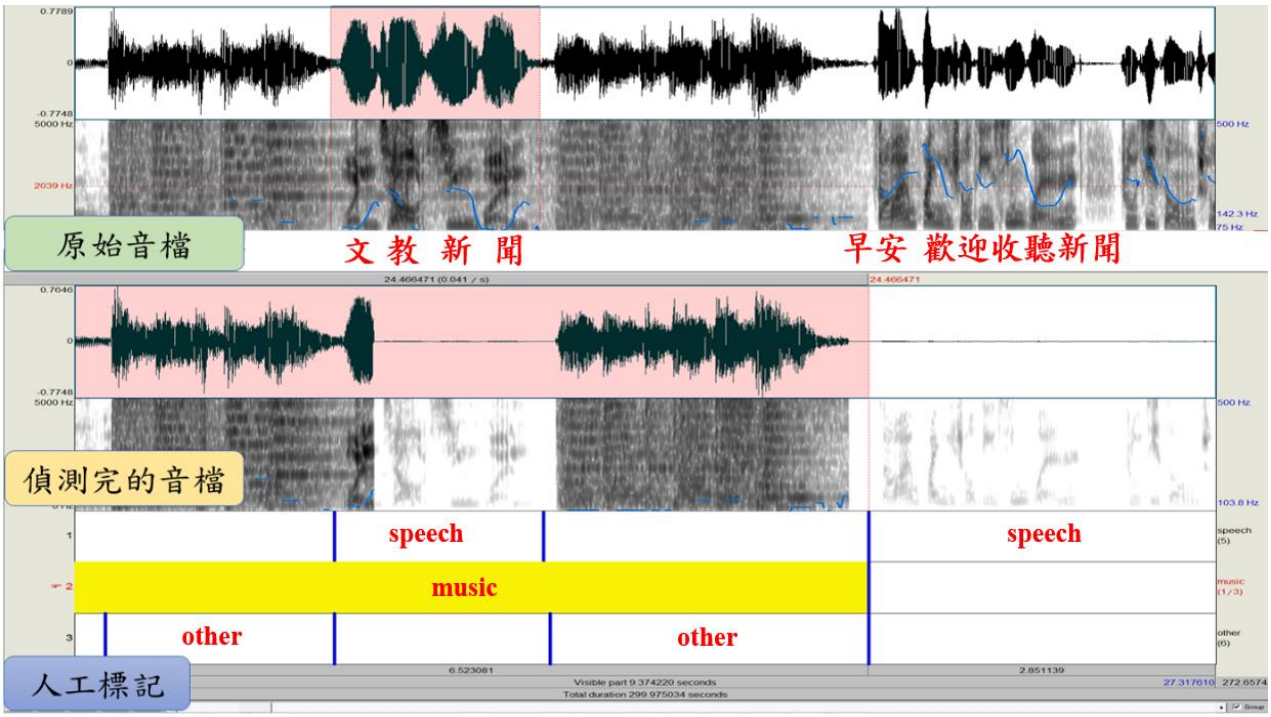

圖 15 音樂偵測結果

4.4.3. 其它(笑聲、特效聲) 事件偵測

圖 16 為經其他事件偵測器處理過的結果。範例圖中可以看到系統的確能夠正確偵測出笑聲事件。

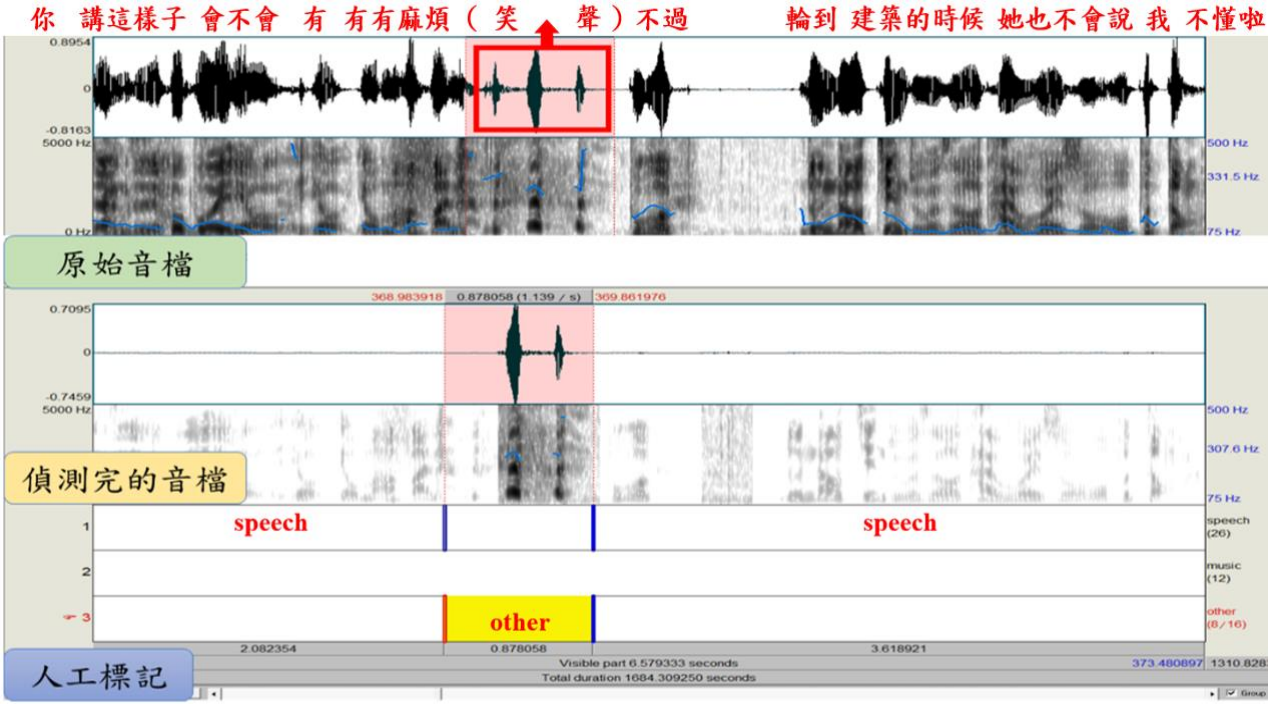

圖 16 其它偵測結果

5. 自動音訊事件標記應用

本論文最後使用前面訓練出來的三種音訊事件偵測器，將偵測到的音訊事件的起始與結束時間找出來。在合併這些資訊做出如圖 17，類似影片字幕具有 timing code 格式的檔案。

```
語音/音樂/其它切割：
00:00:00.660 - 00:00:20.660 -- music
00:00:19.250 - 00:00:20.070 -- other
00:00:20.100 - 00:00:22.200 -- speech
00:00:22.900 - 00:00:24.100 -- other
00:00:21.900 - 00:00:24.600 -- music
00:00:24.300 - 00:04:42.720 -- speech
00:04:42.300 - 00:04:59.110 -- music
```

圖 17 語音/音樂/其它切割結果標註檔格式

最後，我們即可以利用這個檔案，進一步組織並且對廣播節目內容進行做加值運用。尤其是自動轉寫每一語音片段的逐字稿，擷取出關鍵字與摘要，或是自動辨認出每一音樂片段的歌名或曲名。讓聽眾直接以文字進行全文檢索，找到相關節目內容，或是以哼唱方式找到想聽的音樂歌曲段落。

6. 結論

本論文先建立人工標記之廣播節目音訊事件資料庫，再使用 CNN 實作音訊偵測切割系統，並直接使用頻譜，避免參數設計工程問題。整體實驗結果顯示，如表 8 所示，以 CNN 直接搭配頻譜參數，在偵測語音與非語音，音樂與非音樂或其它與非其它音訊事件等的錯誤率 EER，分別為 2.27%、12.52%與 9.51%，皆低於傳統以 GMM 搭配 Mel-Frequency Cepstral Coefficients（MFCCs）的 3.65%、15.68%與 13.25%。因此本論文提出之 CNN 音訊切割架構確實可增強效果，許多文獻中也使用另一種 ivector 求參數，將來會考慮改用 ivector 求參數，跟 AlexNet, VGG, ResNet 等，做近一步實驗比較。

表 8 GMM 和 CNN 系統之正確率與 EER

	GMM		CNN	
Audio event	Accuracy(%)	EER(%)	Accuracy(%)	EER(%)
Speech	97.12	3.65	98.46	2.27
Music	94.88	15.68	96.43	12.52
Other	94.15	13.25	97.47	9.51
Average	95.38	10.86	97.45	8.1

參考文獻

[1] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," in In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016), 2016.

[2] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Contextdependent sound event detection,” in EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no.

323 1, 2013, p. 1.

[3] Yash Malviya, Shiv Kaul, Kushaagra Goyal, "Music Speech Discrimination",in CS 229 Machine Learning Final Projects, Autumn 2016.

[4] Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. In International Conference on Pattern Recognition (ICPR 2012), 2012.

[5] http://www.cs.tut.fi/sgn/arg/dcase2016/challenge

[6] D. Povey, A. Ghosal, G. Boulianne, L. Burgat, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, YM Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, “The Kaldi Speech Recognition Toolkit,” in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Big Island, Hawaii, 2011.

[7] Diego Castan, David Tavarez, Paula Lopez-Otero, Javier Franco-Pedroso, Hector Delgado,Eva Navas, Laura Docio-Fernandez, Daniel Ramos, Javier Serrano, Alfonso Ortega and Eduardo Lleida, " audio segmentation and classification in broadcast news domains",in EURASIP Journal on Audio Speech and Music Processing · December 2015.

[8] D. Scherer, A. Muller, and S. Behnke. " Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition ". In ICANN. 2010.

[9] MUSAN: A Music, Speech, and Noise Corpus,https://scirate.com/arxiv/1510.08484.