The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 197-212

基於半監督式學習之廣播節目語音逐字稿自動轉寫系統

Automatic Transcription of Broadcast Radio Speech Based on Quality Estimation-Guided Semi-Supervised Training

王星月 Sing-Yue Wang, 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
u00157104@gmail.com, asmayday24@gmail.com, yfliao@ntut.edu.tw

摘要

廣播節目製作時通常只有收錄語音訊號，沒有保留相對應的節目內容詮釋資料（metadata），導致節目播出後，很難檢索節目內容，或是加以組織再利用。針對此問題，常用的方法是以語音辨認器，自動轉寫廣播節目內容，產生語音逐字稿，但是目前缺乏已標記好的廣播語音語料庫，因此無法訓練出適合轉寫廣播節目的語音辨識器。所以在本論文中，我們探討如何同時使用語音訊號特徵參數、辨認器辨認結果與語言模型參數，訓練一語音品質估算(Quality Estimation，QE)器，取代傳統只依賴語音辨認器的信心值估算（Confidence Measure），從源源不絕，但未標記的大量廣播語料中，挑選適合訓練語音辨認器的語料，進行半監督式聲學模型訓練，以提升轉寫廣播語料逐字稿的效能。實驗中以一不佳錄音品質 NER-set1 與一優良 NER-set2 之廣播節目測試語料集，測試種子語音辨認器與經半監督式訓練後，新的語音辨認器轉寫語音逐字稿的效能。實驗結果顯示經半監督式訓練後，新的語音辨認器可以把 NER-set1 與 NER-set2 的字元辨認錯誤率（CER）從原始種子模型的 25%與 14.24%，壓低至 23.61%與 13.24%。此外，若進一步改用進階語言模型，更可將 CER 再改善至 23.25%與 12.63%。

關鍵詞：半監督式學習、品質估算、信心度評估、語音辨認系統

1. 簡介

廣播節目的語音資料源源不絕，但因人力、資源等因素，廣播節目製作完成後，通常只有保留最後要播出的語音訊號，沒有將錄製過程中的用到的相關資料，整理成後設資料（metadata）。導致節目播出後，很難檢再檢索節目內容，或是加以組織再利用。因此我們希望能夠轉寫廣播節目產生語音逐字稿，以便將廣播節目組織成有聲書，讓這些大量的語音資料可以有更多的加值運用。除了可以讓聽眾能夠容易地以文字檢索的方式，去找到最關鍵的講述內容部分，尤其是名人在節目中所說的故事、想法思維、新知等等，也可以利用逐字稿，將廣播節目的語音轉成字幕檔，變成多媒體視訊檔案，讓聾胞也可以從中獲知廣播節目內容，或是當做第二語言學習用的語音範例。

要能達到將廣播節目自動轉寫成語音逐字稿這個目的，通常需要先擁有一個適合辨認廣播節目語音的大詞彙語音辨認(Large Vocabulary Continuous Speech Recognition, LVCSR)器，但是因為廣播節目錄製時，通常不會先給主持人與來賓講稿，尤其是對話性質的節目，問答之間常讓來賓自由發揮，因此廣播節目中的語音通常為較隨興的口語，具有強烈的自發性語音（spontaneous speech）特性。

但是，目前非常缺乏高效能的自發性語音辨認器。這是因為若要提高自發性語音辨認器的效能，需要直接以大量的自發性語音語料與口語文字語料來訓練辨認器中的聲學模型與語言模型。但這兩種語料，通常很難取得。尤其是標註好逐字稿的自發性語音語料庫，因其需耗費大量人工、時間、金錢成本才能完成，因此通常有公開發行的自發性語音語料庫都很小，只適合進行語言學分析，探討語言現象。而若要建立基於深度學習的高效能語音模型，就需要很大的數據量，通常需要數百小時，或是數千小時標註好的自發性語料才能達成。因此如何獲得足夠的有標記自發性語音語料，是目前急需解決的一大問題。

針對此問題，一般的做法是以半監督式學習[1][2]方式解決，例如圖 1 所示的架構。方法是先利用較易取得、有正確標註的讀稿語料（reading speech），建立一種子語音辨認器，再用此種子語音辨認器，自動對大量未經人工標註的廣播電臺節目語音語料，進

行逐字稿轉寫。接著把自動產生的逐字稿加入訓練語料庫，重新訓練新的語音辨認模型，以改善辨認自發性語音的效能。

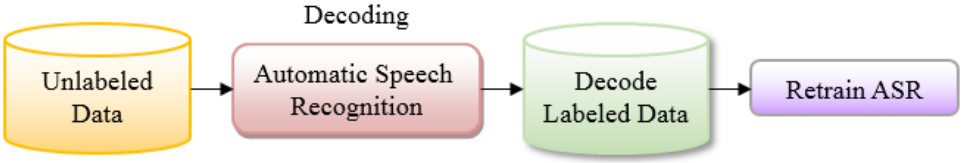

圖 1 利用未標記語料訓練語音辨認器架構

其中，因為自動轉寫出的逐字稿，通常會有錯誤，無法完全信任。因此傳統上會再利用如圖 2 的架構，增加一個信心值估算（Confidence Measure）[3]，計算每段逐字稿的辨認信心值，只挑選較可靠的轉寫結果，加入訓練語料庫。

圖 2 傳統半監督式學習語音辨認器訓練架構

逐字稿的信心值估算，通常是依賴種子辨認器的解碼輸出。然而因為種子辨認器一般是用讀稿語料建立，與自發性語音會有說話模式（speaking style）不匹配的問題，因此算出來的信心值估算不見得可靠。所以在本論文中，將改以同時利用多種語音品質線索，包括語音訊號本身的特徵參數，逐字稿文字內容特徵參數，與多種訊號與文字內容混合參數，建立一語音品質估算器（Quality Estimation，QE），直接預測未標記語料自動轉寫逐字稿的辨識字元錯誤率（CER），並且只挑選辨認錯誤率較低的轉寫結果，加入訓練語料庫。我們所提出的架構如圖 3 所示。

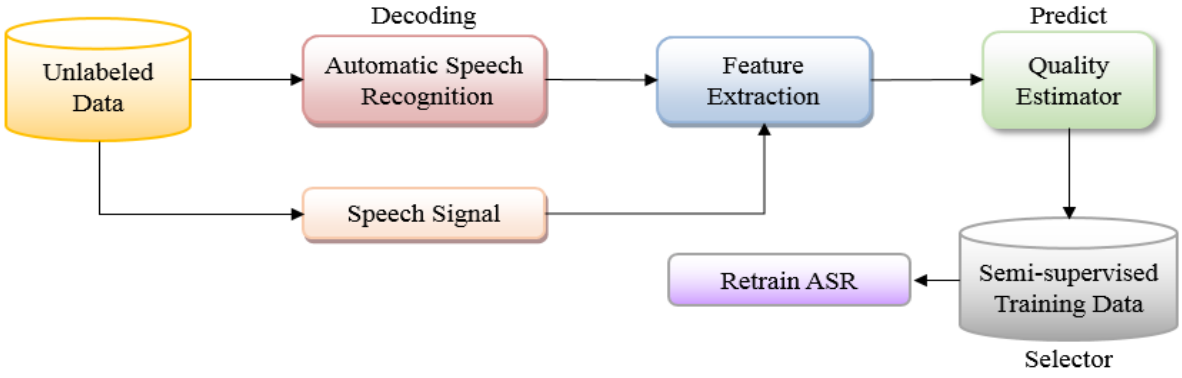

圖 3 以預測錯誤率挑選半監督式學習訓練語料架構

2. 基於語音品質估算之半監督式語音辨認模型

以下說明本論文提出的半監督式語音辨認器中的各模組，包括（1）種子語音辨認模型與（2）半監督式學習訓練的作法。

2.1. 種子語音辨認模型

本論文利用 Kaldi speech recognition toolkit[4]環境，建立種子語音辨認系統，包括語音特徵參數擷取、聲學模型訓練、語言模型訓練。最後以加權有限狀態轉換機[5]結合語言模型與詞典，對教育電臺廣播節目音檔進行轉寫逐字稿，整體架構如圖 4 所示。以下詳細說明各模組。

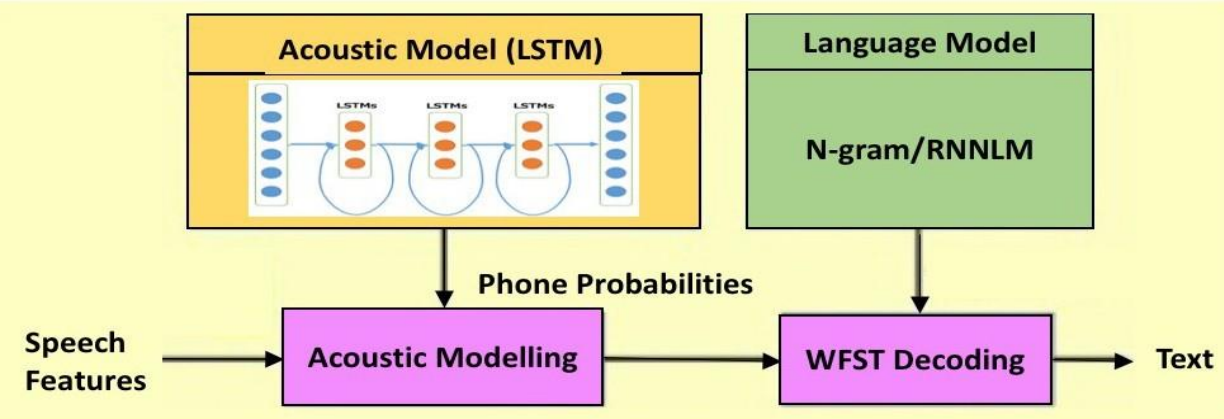

圖 4 種子語音辨認系統架構圖

2.1.1. 聲學模型

LSTM 網路是一種特殊的 RNN 結構[6]，可以記憶較長的時間資訊。其中所有有關訊息傳遞的運作都決定於門(gates)，而這些 gates 依據接收到的信號，計算激發強度來決定訊息是否通過或是被移除。LSTM 的結構能夠用來防止長距離依賴問題，也就是可以解決梯度消失的問題，在此論文中，我們使用我們所擁有的多個語料庫，包含中文、英文、中英夾雜語料，共約 400 小時，來訓練遞迴式類神經網路聲學模型。

2.1.2. 語言模型

語言模型最主要的目標為使用詞序列中先前出現的詞來預測現在最有可能用到的詞。較常被使用的語言模型為 n 連語言模型(n-gram language model)，其統計方式為計算詞與詞之間連接的可能性以挑選可能的字詞。目前進階的作法則是，使用遞迴式類神

經網路語言模型(Recurrent Neural Network Language Model, RNNLM) [7] ，在本論文中我們即使用 RNNLM 去增加語言模型的整體效能，以改善辨認系統的辨認率。

2.2. 半監督式學習訓練

廣播電臺的語音資料非常龐大，但未標記的語料無法拿去做聲學模型訓練。因此我們提出一新的半監督式學習方法，其包含訓練 QE 模型與挑選語料兩部分，整體架構如圖 5 與圖 6 所示。

主要做法是先利用種子語音辨認系統自動轉寫未標記的廣播語料，再使用訓練好的 QE 錯誤率模型，預測其辨認錯誤率，挑選錯誤率較低的語料當作半監督式學習訓練語料，加入種子模型訓練語料重新訓練聲學模型。我們的半監督式學習與傳統方法最大不同的地方，在於增加一個新的 QE 模型，以取代傳統的 CM 語料挑選方法，其中 QE 模型訓練，是依據圖 5 所示的架構，基於監督式學習訓練而成。一方面利用種子語音辨認系統將現有已標記的語料，自動轉寫出逐字稿，並利用已有的人工標記，計算出逐字稿的錯誤率。一方面從自動轉寫的逐字稿擷取文字相關特徵參數、訊號相關參數，以及利用種子語音辨認系統產生的切割時間，與對應的音檔訊號，提取混合特徵參數。再以實際錯誤率為目標，訓練出一個 QE 錯誤率預測模型。

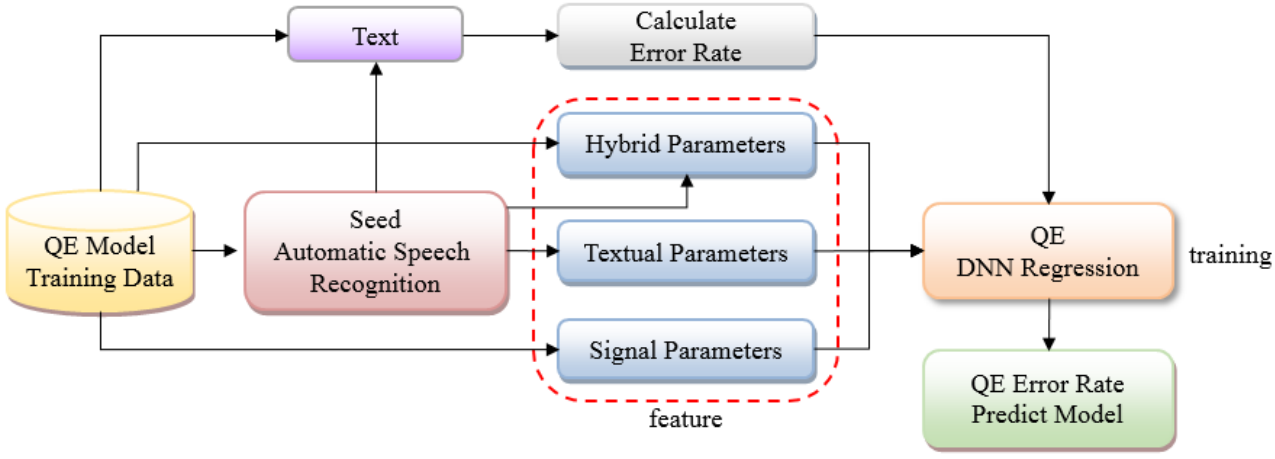

圖 5 QE 模型訓練架構

待 QE 模型訓練好後，就可以依圖 6 所示的架構，利用 QE 模型，預測訓練語料的辨認錯誤率，以從源源不絕，但未標記的大量廣播語料中，挑選適合訓練語音辨認器的語料，進行半監督式聲學模型訓練。

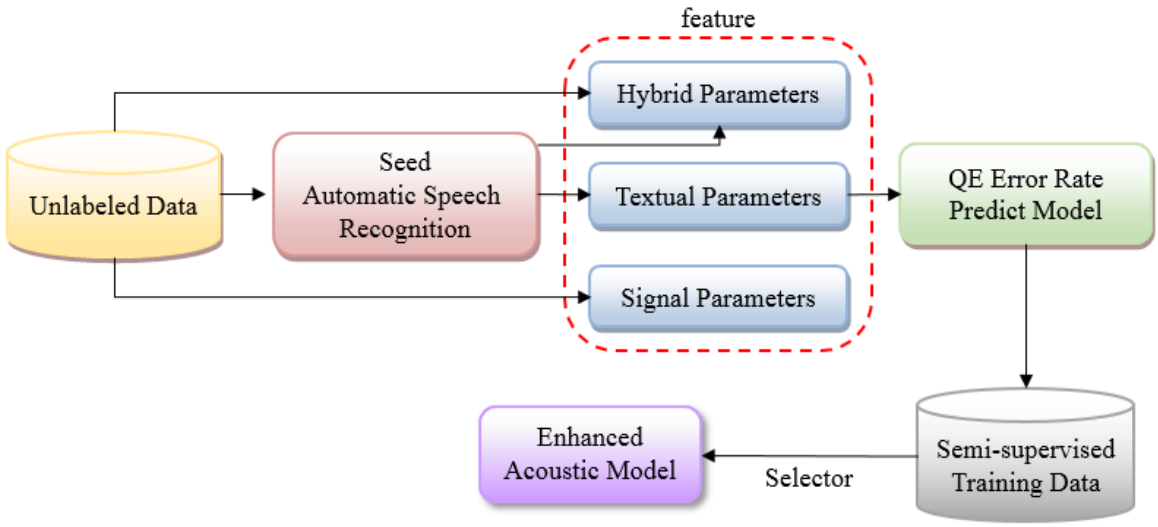

圖 6 利用 QE 名行挑選半監督式學習訓練語料架構

2.2.1. QE 訓練參數擷取

本論文中，我們共利用 93 個參數來訓練 QE 模型[8]，如表 1 所示，包含三類特徵參數。其中 16 個擷取自語音訊號參數、57 個擷取自語言模型參數、20 個擷取自語音訊號與辨認結果的混合參數。

表 1 QE 模型訓練參數

Signal(16)	Total segment duration (sec), Mean/Min/Max raw energy (dB), mean MFCC(12).
Hybrid(20)	SNR, Mean/Min/Max word energy, Mean/Min/Max noise energy, max word - min noise energy, No. of silences, ratio of silences and words, words per second, silences per second, total duration of words, total duration of silences, mean duration of words, mean duration of silences, ratio of total duration silences and total duration words, Std of word duration, Std of silence duration, total duration words - total duration silences.
Textual(57)	Mean of the probability of each word, Sum of log probability of each word, Perplexity in a sentence, probability of each phoneme.

2.2.2. QE 模型訓練

本實驗中利用深層類神經網路方法來製做 QE 模型。深層類神經網路是一種具備至少一個隱藏層的神經網路。其可以透過隱藏層層數的增加，提供更複雜的非線性處理能力，因而能提高模型的能力。

其中深層類神經網路的架構包含多個節點或神經元的多層次，架構如圖 7 所示，其隱藏層間的神經元互不連結，每個神經元使用 Relu 激活函數，用來解決更新權重值時的梯度消失問題，而訓練時使用的成本函數，為依據人工標記算出的真正錯誤率，與 QE 預測的錯誤率間的均方差(Mean Squared Error, MSE)。

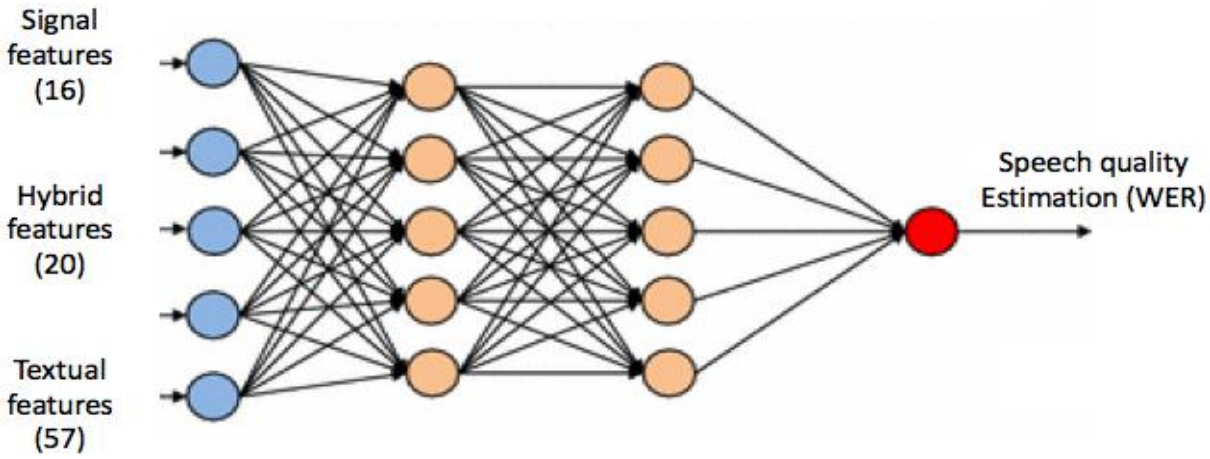

圖 7 DNN-based QE 架構

3. 實驗設定

3.1. 語料庫介紹

在本論文中，我們先使用多個已有的中文、英文及中英夾雜的語料庫，訓練種子語音辨認系統。再利用此種子語音辨認系統，對大量未標記的廣播語料，進行自動轉寫逐字稿，並分別對自動轉寫出的逐字稿做預測錯誤率與信心度評估(Confidence Measure)，挑選適合的語料，加入種子系統的訓練語料重新訓練聲學模型，以比較 QE 方法與 CM 方法的效能。

3.1.1. 種子辨認器訓練語料及測試語料

種子語音辨認系統的訓練語料如表 2 所示，包含 5 個語料庫，共約 400 小時。測試語料如表 3 所示，包含 7 組測試語料。尤其是在測試語料部分，增加了從教育廣播電臺節目挑出錄音品質較差，辨認率較不好的 NER-set1（技職最前線），以及錄音品質較佳，辨認率較好的 NER-set2（國際教育心動線），以測試半監督式訓練對不同廣播節目的轉寫效能。

表 2 種子系統訓練語料

訓練語料	時數	語者數	語句數
TCC300 (all)	26.4	300	27,375
MATBN (train)	127.3	5,207	29,549
OC16-CE80 (train)	63.8	1,163	58,132
SEAME	95.1	138	94,034
Librispeech (train-clean100hr)	100.6	251	28,539
Total	413.2	7,059	237,629

表 3 半監督式訓練系統測試語料

測試語料	時數	語者數	語句數
NER-set1	1.75	35	438
NER-set2	3.23	23	640
MATBN (test)	3.06	273	729
OC16-CE80 (test)	7.93	142	7,099
SEAME	13.70	18	12,104
Librispeech (test-other)	5.10	33	2,939
Librispeech (test-clean)	5.40	40	2,620
Total	40.17	564	26,569

3.1.2. QE 模型訓練語料

在此利用人工先標記部分教育電臺語料庫，來進行 QE 錯誤率預測模型訓練。其中包含八個不同節目廣播語料，總計約 65 小時、10526 語句數。詳細資料如表 4 所示。

表 4 QE 錯誤率預測模型訓練語料

訓練語料	時數	語句數
創設市集 On-Air	10.73	3,000
多愛自己一點點	6.77	720
兒童新聞	0.86	166
國際教育心動線	3.23	640
技職最前線	1.75	438
科學 So Easy	1.84	208
雙語新聞	34.49	4,042
文教新聞	5.46	1,312
Total	65.13	10,526

3.2. 中英夾雜語音辨認器

在華語國家由於被國際語言英文的影響狀況下時常在講話時會穿插一些英文字詞，而單語言語音辨認器顯然無法正常辨識多語言夾雜的說話語流，所以訓練一個多語言語音辨認器會是一個符合現代人在自然說話語流的說話形式，因此我們將中、英文利用 X-SAMPA 音素編碼規則產生音素，並且使用音素共享與中英混合模型來完成我們的中英夾雜語音辨認器。

3.2.1. 音素共享

在不同語言之間存在著相近音的現象，既然是相似的音就不需要因為語言的不同而分開訓練，而是將其相近的發聲音素合併建模訓練，在此實驗中所有音素編碼規則皆使用 X-SAMPA，中文音素共有 131 個(包含語調)，英文音素共有 69 個(包含重音)，其中選取了中、英音素在 X-SAMPA 用於相同符號表示的音素，去掉原本區別中、英音素的標籤，其中共享了 10 個子音音素如表 5 所示，最後，我們所使用的 X-SAMPA 共有 190 個音素。

表 5 音素共享表

共享(子音)音素	註音
j、w、t、s、p、n、m、l、k、f	ㄧ、ㄨ、ㄉ、ㄙ、ㄆ、ㄋ、ㄇ、ㄌ、ㄍ、ㄈ

3.2.2. 中英混合字典

我們設定了 X-SAMPA 有 190 個音素與整理完訓練文本後，需要整合成一個中英混合字典，從訓練文本中挑出所有不重複的單詞，並且按照 X-SAMPA 標記出音素，我們的中英混合字典最終有 455,715 個字詞。

3.2.3. 中英混合語言模型

表 6 則為種子系統語言模型的訓練文本，在此實驗我們集結諸多語料庫的文本進行訓練。但為了避免 inside test 情形發生，我們只抽取語料庫的訓練語料之文本做為訓練資料，並且使用 4-gram 與 RNNLM 來建構語言模型。

表 6 種子系統語言模型訓練文本

訓練文本	語句數	字詞數
TCC300 (all)	27,375	186,369
MATBN (train)	29,549	1,264,625
OC16-CE80 (train)	58,132	509,657
SEAME	94,034	1,200,121
Librispeech (train-960)	28,539	9,403,555
Giga Word	500,000	9,899,664
Total	737,629	22,463,991

4. 實驗結果與分析

4.1. 實驗一-種子語音辨認系統效能

我們先測試種子語音辨認系統的辨認效能，尤其是對從教育廣播電臺節目中挑出，辨認率較差的 NER-set1（技職最前線），以及辨認率較好的 NER-set2（國際教育心動線），分別做測試，以此做為 Baseline 系統的效能參考值。

表 6 為種子辨認器辨認效能實驗結果。從表 7 可以看到，雖然都是來自教育廣播電臺的語料，但是因為節目錄音品質不同，主持人及來賓的口語不同，所談論的話題不同，在整體的辨認上這兩個測試語料的錯誤率差距相當大（相差約 10%）。

表 7 種子語音辨認系統辨認率

測試語料	種子模型
NER-set1	25.00
NER-set2	14.24
MATBN (test)	13.18
OC16-CE80 (test)	16.30
SEAME	36.32
Librispeech (test-other)	18.17
Librispeech (test-clean)	5.00

4.2. 實驗二- QE 模型訓練結果

此實驗中將擷取的 93 個特徵參數，利用三種不同迴歸訓練架構，包含（1）支援向量回歸（Support Vector Regression, SVR）[9]、（2）極端隨機樹（Extremely randomized

trees, Extra-Tree）[10]與（3）DNN，分別訓練三種 QE 錯誤率預測模型，進行錯誤率預測效能比較。其中，DNN 的訓練參數設定為 learning rate= 0.001、epochs=100、batch size=500、dropout=0.1，DNN 的層數則嘗試使用 1~3 層隱藏層。

因訓練語料較少，為公平比較三種方式，我們利用 Cross-validation 訓練與測試方式，在每一次的訓練將其中一個節目當作測試資料，其餘七個節目當作訓練資料，總共進行八次訓練與測試，最後再將八次測試結果所計算的 MAE、MSE 加總平均。

首先，由於 SVR 與 Extra-Trees 是屬於淺層分析，在 DNN 部分先只使用一層隱藏層與 SVR 和 Extra-Trees 比較。實驗結果如表 8 所示，可以看到使用 DNN 架構訓練出來的預測模型，其錯誤率誤差最小。

表 8 QE 錯誤率預測模型三種架構訓練結果比較

廣播節目	1 layer DNN		SVR		Extra-Trees	
	MAE	MSE	MAE	MSE	MAE	MSE
多愛自己一點點	0.0819	0.0101	0.0769	0.0110	0.0893	0.0194
創設市集 On-Air	0.1134	0.0201	0.1057	0.0230	0.0991	0.0196
兒童新聞	0.0938	0.0141	0.1006	0.0157	0.0979	0.0147
國際教育心動線	0.0742	0.0100	0.0875	0.0129	0.0930	0.0138
技職最前線	0.0898	0.0128	0.0878	0.0138	0.0944	0.0153
科學 So Easy	0.0676	0.0069	0.0793	0.0098	0.0688	0.0096
雙語新聞	0.0989	0.0163	0.1103	0.0220	0.1078	0.0219
文教新聞	0.0966	0.0133	0.0932	0.0141	0.0965	0.0146
Average	0.0895	0.0129	0.0927	0.0153	0.0934	0.0161

然後，我們再訓練多層 DNN，看多層 DNN 是否可以進一步提升預測效果。從實驗結果中使用 2 layer DNN 會有較低的預測誤差，總結如下表 9 所示。因此，在以下的非監督式學習實驗中，皆使用兩層的 DNN 模型做 QE 預測。

表 9 QE 錯誤率預測模型測試結果

廣播節目	**1 layer DNN**		**2 layer DNN**		**3 layer DNN**	
	MAE	MSE	MAE	MSE	MAE	MSE
Average	0.0894	0.0164	**0.0855**	**0.0130**	0.0865	0.0134

4.3. 實驗三-基於 CM 及 QE 之半監督式訓練效能比較

為測試 QE 與 CM 兩種不同語料挑選方法，我們從教育電臺廣播語料中，先取出 16

個不同節目，總計約為 377 小時的未標記語料。經過信心度評估 CM 及品質估算 QE，分為 QE1，QE2，CM1 與 CM2 四種挑選機制。CM1 挑出約 210 小時（此部分因語料時數不同，僅為參考用），而 QE1，QE2 與 CM2 各挑選出約莫 38 小時的半監督式學習訓練語料，如表 10。

表 10 半監督式學習訓練語料挑選

使用語料時數	Total	CM1	CM2	QE1	QE2
廣播節目語料(hour)	377.58	209.25	38.28	38.44	38.21

其中，CM1 是以每一字詞的信心度評估，依照較高的轉寫信心程度($score \geq 0.9$)，以詞為單位做挑選。CM2 則是先計算每一句中所有字詞信心度評估，再以句為單位取平均($score \geq 0.9$)，進行挑選。QE1 使用 93 個特徵參數，預測出錯誤率，挑選錯誤率($wer < 0.3$)較低的語句。最後，QE2 是將 CM 值再加入原有的 93 個特徵參數，訓練出新的 QE 錯誤率預測模型，一樣從錯誤率較低($wer < 0.3$)的語句開始挑選。

依挑選結果，我們各自將所挑選出的語料加入原先的訓練語料，重新訓練四個聲學模型以測試四種挑選半監督式學習訓練語料方法的效能。

測試結果如表 11 所示，可以發現使用 QE 或是 CM 值挑選語料，所訓練出來的聲學模型，在不同測試語料的語音辨認上都能降低整體的錯誤率，並且用 QE1 或是 CM2 挑選訓練語料，訓練出的聲學模型，在平均語音辨認錯誤率上都有較好的表現。

表 11 基於半監督式學習聲學模型訓練結果

測試語料	種子模型	CM1(610h)	CM2(438h)	QE1(438h)	QE2(438h)
NER-set1	25.00	24.66	23.88	24.04	23.88
NER-set2	14.24	13.85	13.26	13.11	13.35
MATBN (test)	13.18	13.19	13.00	13.00	13.24
OC16-CE80 (test)	16.30	15.96	16.10	16.08	15.95
SEAME	36.32	36.70	35.85	35.96	36.13
Librispeech (test-other)	18.17	18.00	17.83	17.87	18.15
Librispeech (test-clean)	5.00	5.18	5.02	4.96	5.2
Average1 (with SEAME)	20.53	20.57	20.19	20.22	20.39
Average2 (without SEAME)	13.21	13.08	12.93	12.92	13.09

而若只針對教育電臺測試語料來看半監督式學習的效能，實驗結果顯示四種方法的相對辨認改善率如表 12 所示。其中錄音品質較差的 NER-set1 （技職最前線)的最佳相對改善率來到 4.48%（QE2 與 CM2)，錄音品質較好的 NER-set2（國際教育心動線）的最佳相對改善率來到 7.94%（QE1)。

表 12 廣播語料測試結果_相對改善率

辨認模型	CER in %		相對改善率	
	NER-set1	NER-set2	NER-set1	NER-set2
種子模型	25.00	14.24	-	-
CM1(610h)	24.66	13.85	1.36%	2.74%
CM2(438h)	23.88	13.26	4.48%	6.88%
QE1(438h)	24.04	13.11	3.84%	7.94%
QE2(438h)	23.88	13.35	4.48%	6.25%

4.4. 實驗四-挑選語料量與效能比較

以下實驗針對 QE1 錯誤率預測模型所挑出的語料，以漸進的方式加入半監督式學習訓練語料，重新訓練聲學模型，測試挑選語料量與效能的影響。語料挑選順序為依預測錯誤率從低到高，挑出四組，各有 38 小時，50 小時、50 小時與 60 小時。因此訓練語料總時數分變成為 438 小時、488 小時、538 小時、598 小時四組。

挑選語料量與效能實驗結果如表 13 所示。另外，圖 8 為只針對教育電臺測試語料來看半監督式學習的效能改善曲線，實驗結果顯示對 NER-set1，隨著加入的訓練語料的時數增加，所訓練出來的語音辨認器，的確有更好的辨認效能。

表 13 不同語料量對半監督式學習型訓練結果的影響

測試語料	種子模型	QE(438h)	QE(488h)	QE(538h)	QE(598h)
NER-set1	25.00	24.04	24.00	23.86	23.61
NER-set2	14.24	13.11	13.23	12.96	13.24
MATBN (test)	13.18	13.00	13.04	13.12	13.28
OC16-CE80 (test)	16.30	16.08	16.08	16.11	16.40
SEAME	36.32	35.96	36.15	36.45	36.84
Librispeech (test-other)	18.17	17.87	18.01	18.59	18.51
Librispeech (test-clean)	5.00	4.96	5.22	5.19	5.31
Average	18.32	17.86	17.96	18.04	18.17

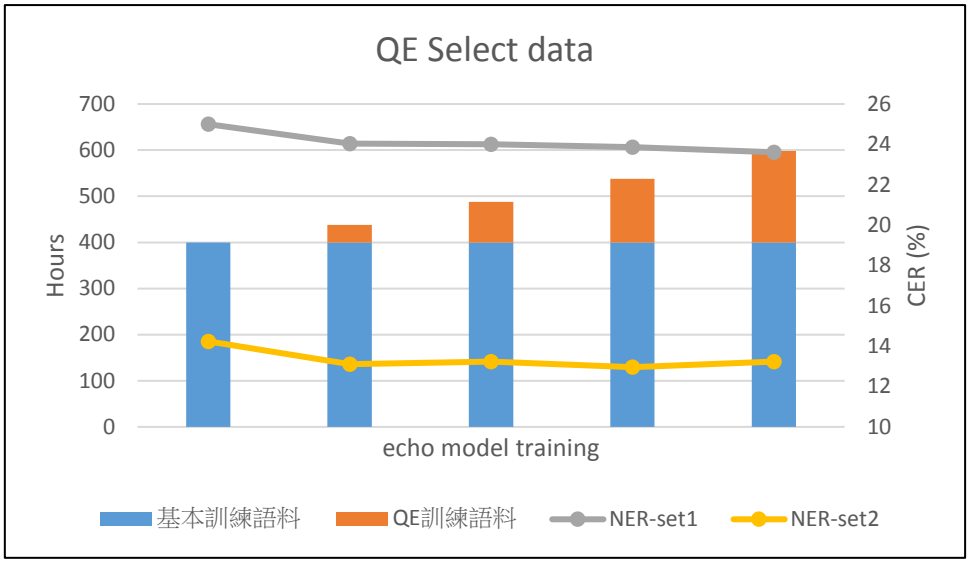

圖 8 半監督式學習聲學模型訓練結果(改善曲線)

4.5. 實驗五-語言模型改善

在上一節的實驗中，針對聲學模型的訓練上從原本的種子訓練語料 400 小時增加了 198 小時的半監督式學習訓練語料重新訓練，改善程度已經趨近於收斂狀態。而在實驗五，我們將加入更豐沛的訓練文本，來訓練我們的語言模型。主要是在種子系統語言模型訓練文本中，增加了更完整的訓練文本 Giga Word2。其統計資料如表 14 所示。

表 14 語言模型訓練文本

訓練文本	語句數	字詞數
TCC300 (all)	27,375	186,369
MATBN (train)	29,549	1,264,625
OC16-CE80 (train)	58,132	509,657
SEAME	94,034	1,200,121
Librispeech (train-960)	28,539	9,403,555
Giga Word	500,000	9,899,664
Giga Word2	16,500,000	441,889,701
Total	17,237,629	464,353,692

實驗的結果如表 15 所示，可以看到語言模型建模的能力對語音識別的結果有一定的影響，讓辨認率較差的 NER-set1(技職最前線)相對改善率來到 7%，辨認率較好的 NER-set2(國際教育心動線)相對改善率來到 11.3 %。

表 15 加入完整 Giga Word 語料後之廣播語料辨認錯誤相對改善率

辨認模型	CER in %		相對改善率	
	NER-set1	NER-set2	NER-set1	NER-set2
種子模型	25.00	14.24	-	-
種子模型+LM2	24.54	13.27	1.84%	6.81%
QE1(598h)	23.61	13.24	5.56%	7.02%
QE1(598h)+LM2	23.25	12.63	7%	11.3%

5. 結論

我們用 QE 模型挑選半監督式學習語料，對辨認率較差的 NER-set1 字元錯誤率（CER）來到 23.61%，辨認率較好的 NER-set2 字元錯誤率（CER）其辨認率來到 13.24%。若增加了更完整的語言模型訓練文本 Giga Word2，能讓 NER-set1 的最佳辨認率來到 23.25%，NER-set2 的最佳辨認率來到 12.63%。

最後，整體效能改善總結如圖 9 所示，其顯示使用 QE 模型來挑選半監督式學習訓練語料，重新訓練聲學模型，確實能有效提升聲學模型之效能。

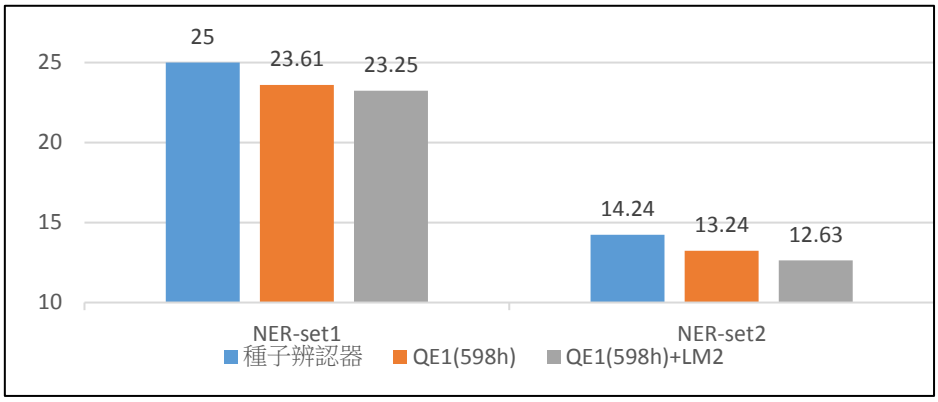

圖 9 廣播語料測試結果改善率

參考文獻

[1] Wessel, F. and Ney, H., “Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 1, 2005, pp. 257-265.

[2] Chen, B., Kuo, J.W., Tsai, W.H., “Lightly Supervised and Data-Driven Approaches to

Mandarin Broadcast News Transcription," International Journal of Computational Linguistics & Chinese Language Processing, vol. 10, no. 1, 2005, pp. 1-18.

[3] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Communication, vol. 45, no. 4, pp. 455 – 470, 2005.

[4] 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.

[5] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiat, S. Kombrink, ´ P. Motl´ıcek, Y. Qian, N. T. Vu, K. Riedhammer, and ˇ K. Vesely, "Generating exact lattices in the WFST ´ framework," 2011, submitted to ICASSP 2012.

[6] Understanding LSTM Networks http://colah.github.io/posts/2015-08-Understanding-LSTMs/ , 2016, July.

[7] Mikolov Tomáš, Kombrink Stefan, Deoras Anoop, Burget Lukáš, Černocký Jan: RNNLM - Recurrent Neural Network Language Modeling Toolkit, In: ASRU 2011 Demo Session.

[8] Negri, M., Turchi, M., Falavigna, D., C. de Souza, J.G., 2014. Quality Estimation for Automatic Speech Recognition, in: Proc. of COLING, Dublin, Ireland. pp. 1813–1823.

[9] 0Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004): 199-222.

[10] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees.Machine Learning, 63(1), 3-42.

[11] Raj Nath Patel, Sasikumar M. 2016. Translation Quality Estimation using Recurrent Neural Network. In Proceedings of the First Conference on Machine Translation, pages 819-824, Berlin, Germany.