The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 4-6

融合多任務學習類神經網路聲學模型訓練於會議語音辨識之研究
Leveraging Multi-task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition

楊明翰 Ming-Han Yang, 許曜麒 Yao-Chi Hsu, 洪孝宗 Hsiao-Tsung Hung, 陳映文 Ying-Wen Chen, 陳柏琳 Berlin Chen
國立台灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{mh_yang, ychsu, alexhung, cliffchen, berlin}@ntnu.edu.tw

陳冠宇 Kuan-Yu Chen
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
kychen@iis.sinica.edu.tw

摘要

語音長久以來一直是人跟人之間最自然的溝通方式；它在未來將是人與電腦等機器間溝通的一個不可或缺的重要工具。近六十年來，自動語音辨識的研究活動十分活躍，並且已取得了巨大的成功。在研究初期，語音辨識器只能在安靜的環境中識別一個單獨的詞彙。1980 年代，以高斯混合模型-隱藏式馬可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM)做為聲學模型使得語音辨識有能力進行大詞彙量連續語音識別[1]。由於 GMM-HMM 的架構易於訓練模型和進行聲學解碼，因此在近二十年來 GMM-HMM 是自動語音辨識系統的主流聲學模型，聲學模型的研究主要集中在以更好的模型結構與訓練演算法改良 GMM-HMM[1][2][3][4]。在過去的五年內，我們看見了深層學習架構和技術在語音領域的突破性的發展和卓越的成效[5][6][7]。深層類神經網路與其變體最終取代了高斯混合模型；時下的混合深層類神經網路-隱藏式馬可夫模型(hybrid deep neural networks-hidden Markov model, DNN-HMM)已成為大多數自動語音辨識系統的聲學模型[8][9][10]。雖然自動語音辨識技術已經是一項成熟的技術，但是在實際應用上仍有許多問題需要被解決。例如使用智慧型手機錄音時往往離手機麥克風較遠，錄音品質容易受環境影響。此外，現今語音辨識領域也面臨著海量詞彙、自由不受

限的任務、吵雜的遠距離語音、自發性的口語及語言混雜情景的挑戰[11]。而會議語音辨識正涵蓋了上述大部分的困境與挑戰，是一個相當困難的語音辨識任務。因此，本論文以會議語音辨識的發展為研究動機，旨在探索如何融合多任務學習(multi-task learning, MTL)技術於聲學模型之參數估測，藉以改善會議語音辨識(meeting speech recognition)之準確性。我們的貢獻主要有三點：(1)我們進行了實證研究以充分利用各種輔助任務來加強多任務學習在會議語音辨識的表現。此外，我們還研究多任務與不同聲學模型像是深層類神經網路(deep neural networks, DNN)聲學模型及摺積神經網路(convolutional neural networks, CNN)結合的協同效應，期望增加聲學模型建模之一般化能力(generalization capability)。(2)由於訓練多任務聲學模型的過程中，調整不同輔助任務之貢獻(權重)的方式並不是最佳的，因此我們提出了重新調適法，以減輕這個問題。我們基於在台灣所收錄的華語會議語料庫(Mandarin meeting recording corpus, MMRC)建立了一系列的實驗。與數種現有的基礎實驗相比，實驗結果揭示了我們所提出的方法之有效性。

關鍵詞：多任務學習，深層學習，類神經網路，會議語音辨識。

致謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫(104-2911-I-003-301)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3 和 MOST 105-2221-E-003-018-MY3)之經費支持，謹此致謝。

參考文獻

[1] M. N. Stuttle, *A Gaussian Mixture Model Spectral Representation for Speech Recognition training for large vocabulary speech recognition*, Ph.D. dissertation, University of Cambridge, 2003.

[2] V. Valtchev, J. J. Odell, P. C. Woodland, and S. Young, “Lattice-based discriminative training for large vocabulary speech recognition,” in *ICASSP*, 1996.

[3] P. C. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition,” *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[4] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. dissertation, University of Cambridge, 2004.

[5] O. A. Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[6] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, “Very deep multilingual convolutional neural networks for LVCSR,” in *Proc. ICASSP*, 2016.
[7] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, 2013.
[8] J. Li, A. Mohamed, G. Zweig, and Y. Gong, “Exploring multidimensional LSTMs for large vocabulary ASR,” in *Proc. ICASSP*, 2016.
[9] A. R. Mohamed, F. Seide, D. Yu, J. Droppo, A. Stolcke, G. Zweig and G. Penn, “Deep bi-directional recurrent networks over spectral windows,” in *Proc. ASRU*, 2015.
[10] T. Sainath, O. Vinyals, A. Senior and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks, ” in *Proc. ICASSP*, 2015.
[11] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2014.