Text-independent Speaker Verification using a Hybrid I-Vector/DNN Approach

結合 I-Vector 及深層神經網路之語者驗證系統

張雲帆 Yun-Fan Chang[1]，曹昱 Yu Tsao[1]，鄭少樺 Shao-Hua Cheng[2]，詹凱軒 Kai-Hsuan Chan[2]，廖嘉維 Chia-Wei Liao[2]，張文村 Wen-Tsung Chang[2]
[1] 中央研究院資訊科技創新研究中心
[2] 財團法人資訊工業策進會前瞻科技研究所
{she2113, yu.tsao}@citi.sinica.edu.tw,
{briancheng, kaihsuanchan, cliao, wtchang}@iii.org.tw

摘要

語者驗證的目的是以語音訊號來驗證特定語者的身份(Identity)，此項研究在近年的智慧生活環境已成為一個重要的研究議題。不論是在門禁系統，亦或是搜尋、偵測特定語者語音等，都被廣泛應用。語者驗證又分為文字特定模式(Test-dependent Mode)與文字不特定模式(Text-independent Mode) 兩類 [1]，前者的好處為已知較多語音資訊，可以大幅改善系統的驗證效能，但實際的應用限制較多，後者因為是隨機的語音訊號，資訊量較少，相對驗證效果不如前者，但也因為限制較少，應用層面相對較大。在本研究中，我們著重於文字不特定模式的語者驗證。

傳統的語者驗證系統是使用高斯混合模型的架構，其作法是訓練一套 Universal Background Model (UBM)高斯混合模型(Gaussian Mixture Model，GMM)，UBM-GMM。接著利用每一位語者的語音訊號，以及最大後驗概率法則(Maximum A Posteriori，MAP)對 UBM-GMM 作調整以得到每位語者專屬模型，接著再對測試語句利用 UBM-GMM 及 Speaker-specific GMM 分別計算似然值 [2, 3]。另外，還有將 GMM 抽取 Mean 串成 Supervector 再使用 Support Vector Machine(SVM)作辨識的方法 [4, 5]。

近年來在 NIST Speaker Recognition Evaluations(SRE)發展了一套 I-Vector 的特徵擷取方式，其特徵擷取包含以下三個步驟:1.對語音訊號作 MFCC 特徵擷取。2.利用 UBM-GMM 計算出每位語者的 Supervector。3.使用 Baum-Welch Statistics 計算出 I-Vector。過去的研究證實，I-Vector 搭配 SVM 分類器，能有效地完成語者識別 [6]。近日，深層神經網路(Deep Neural Network，DNN)已被廣泛地應用在各類型的分類問題 [7-10]。本論文提出使用 I-Vector 結合深層神經網路進行語者驗證。

本實驗所使用的資料為某談話性節目實際語音資料，目的為找出特定女主持人的語音片段。訓練語料為 177 句女性談話語句，目標訓練語句為某女主持人的 12 句語料，其長度均約 6 秒。經過 Voice Activity Detection(VAD)處理後，訓練語料切割成 1,921 句語句，目標訓練語句切割成 118 句語句。測試語料共 300 句，其中 30 句為目標語句，其餘 270 句為男女混合之語料，長度均約 3 秒。實驗設定部分，MFCC 特徵為 13 維展延成 39 維 MFCC，I-Vector 使用 256 個高斯混合數的 UBM-GMM，其維度為 64 維。在此篇論文的實驗結果顯示，增加維度不會明顯提升辨識結果，而相對會產生額外的運算

量。DNN 使用兩層隱藏層，神經單元均設為 150，其原因與上述相同。

在實驗中，對於語者驗證系統的評量，我們通常使用 Equal Error Rate(EER)做為評量標準。另外，我們還使用 Precision、Recall、F-measure 和 Accuracy 評量模型效能，我們將實驗結果整理於下表一。由實驗結果可知，我們提出的 I-Vector 搭配 DNN 系統在各種評量方式皆優於 I-Vector 搭配 SVM 系統。

表一、評估結果

	Precision	Recall	F-measure	Accuracy	EER
SVM	35%	67%	46%	84%	19.26%
DNN	70%	70%	70%	94%	12.22%

參考文獻

[1] H. Gish and M. Schmidt, “Text-independent speaker identification,” *IEEE, Signal Processing Magazine*, vol. 11, pp. 18-32, 1994.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *ELSEVIER, Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[3] D. A. Reynolds, “An overview of automatic speaker recognition technology,” *IEEE Transactions, Acoustics Speech and Signal Processing*, vol. 4, pp. 4072-4075, 2002.

[4] S. Fine, J. Navratil and R. A. Gopinath, “A hybrid GMM/SVM approach to speaker identification,” *IEEE Transactions, Acoustics Speech and Signal Processing*, vol. 1, pp. 417-420, 2001.

[5] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *ELSEVIER, Computer Speech & Language*, vol. 20, pp. 210-229, 2006.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions, Audio, speech ,and Language Processing*, vol. 19, pp. 788-798, 2011.

[7] H. Larochelle, Y. Benqio, J. Louradour, and P. Lamblin, “Exploring strategies for training deep neural networks,” *Machine Learning*, vol. 10, pp. 1-40, 2009.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” *IEEE, Signal Processing Magazine*, vol. 29, pp. 82-97, 2012.

[9] Y. Bengio, “Learning deep architectures for AI,” *Found. Trends Mach. Learn*, vol. 2, pp. 1-127, 2009.

[10] A. Mohamed, G. E. Dahl, and G. E. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions, Audio, speech ,and Language Processing*, 20, 14–22, 2013.