Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 63-72

基於鑑別式自編碼解碼器之錄音回放攻擊偵測系統

A Replay Spoofing Detection System Based on Discriminative Autoencoders

吳家隆*、許祥平*、呂滮鼎+、曹昱+、李鴻欣#、王新民#

Chia-Lung Wu, Hsiang-Ping Hsu, Yu-Ding Lu, Yu Tsao,

Hung-Shin Lee and Hsin-Min Wang

摘要

在此論文中，我們提出了一個基於鑑別式自編碼解碼器的神經網路模型，對語者辨識系統的錄音回放攻擊進行自動偵測，也就是判斷語者辨識系統所收到的音訊內容是屬於真實的人聲或是由錄音機所回放出來的人聲。在語者辨識領域中，以人為的聲音造假對語者辨識系統進行的攻擊稱之為欺騙攻擊(Spoofing Attack)。有鑑於深度類神經網路模型已被廣泛應用在語音處理相關問題，我們期望能夠應用相關模型在此類問題上。在所提出的鑑別式自編碼解碼器模型中，我們利用模型的中間層來達到特徵抽取的目的，並且提出新的損失函數，使得中間層的特徵將依照資料的標記結果做分群，因此新的特徵將具有能鑑別真偽人聲的資訊，最後再利用餘弦相似度來計算所抽取的特徵與真實的人聲相近與否，得到偵測的結果。我們採用 2017 Automatic Speaker Verification Spoofing and Countermeasures Challenge(ASVspoof-2017)所提供的資料庫進行

* 法務部調查局
Investigation Bureau, Ministry of Justice
E-mail: m38025@mjib.gov.tw
+ 中央研究院資訊創新科技研究中心
Research Center for Information Technology Innovation, Academia Sinica
E-mail: jolu.citi@gmail.com; yu.tsao @iis.sincia.edu.tw
The author for correspondence is Yu Tsao.
中央研究院資訊科學研究所
Institute of Information Science, Academia Sinica
E-mail: hungshinlee@gmail.com

測試，所提出的系統在開發數據集上得到了很好的成效，與官方所提供的測試方法相比，其準確度約有 42 %的相對進步幅度。

關鍵字：語者辨識，語者辨識攻擊，回放攻擊偵測，鑑別式自編碼解碼器，深度類神經網路

Abstract

In this paper, we propose a discriminative autoencoder (DcAE) neural network model to the replay spoofing detection task, where the system has to tell whether the given utterance comes directly from the mouth of a speaker or indirectly through a playback. The proposed DcAE model focuses on the midmost (code) layer, where a speech utterance is factorized into distinct components with respect to its true label (genuine or spoofed) and meta data (speaker, playback, and recording devices, etc.). Moreover, the concept of modified hinge loss is introduced to formulate the cost function of the DcAE model, which ensures that the utterances with the same speech type or meta information will share similar identity codes (i-codes) and higher similarity score computed by their i-codes. Tested on the development set provided by ASVspoof 2017, our system achieved a much better result, up to 42% relative improvement in the equal error rate (EER) over the official baseline based on the standard GMM classifier.

Keywords: Speaker Verification, Speakser Verification Attack, Spoofing Attack, Discriminative Autoencoder, Deep Neural Network.

1. 緒論 (Introduction)

在近幾年內，自動語者辨識的準確度已經有了顯著的提升，自動語者辨識系統也已經廣泛地應用在日常生活中，像是行動裝置或是個人語音裝置的登入系統。然而，由於語音合成(Speech Synthesis)、語音轉換(Voice Conversion)、文字轉語音(Text-to-speech)及錄音回放等技術的進步(Abe, Nakamura, Shikano & Kuwabara, 1990) (Chen, Ling, Liu & Dai, 2014) (Van Santen, Sproat, Olive & Hirschberg, 2013) (Ze, Senior & Schuster, 2013)，電腦越來越能夠模仿人類所發出來的聲音，因此，這些技術確實造成了自動語者辨識系統的潛在危機。

在 ASVspoof-2015 比賽中，其主要的任務目標為訓練一個系統來分辨真實人聲(Genuine)及由語音合成器產生的合成人聲(Spoofing) (Wu *et al.*, 2015)；故這次比賽所提供的數據庫中，主要組成為文字轉語音和語音合成的語料，並為與文本無關(Text independent)的內容，但並不包括由錄音機所播放出的回放數據。為了能夠訓練出更好的系統來面對更多的語者辨識攻擊(Alegre, Amehraye & Evansdoi, 2013) (Alam, Kenny, Bhattacharya & Stafylakis, 2015) (Xiao *et al.*, 2015) (Villalba, Miguel, Ortega & Lleida,

2015)，ASVspoof-2017 所提供的資料集就包含了由多個不同的播放設備所播出來的人聲錄音回放，參賽單位需要面對並解決新的問題。ASVspoof-2017 所提供的基本系統架構(Kinnunen *et al*., 2017) 以常數 Q 倒頻譜係數(Constant Q Cepstral Coefficients, CQCC)為特徵抽取的參考方法，並且利用了二類的高斯混合模型(Gaussian Mixture Models, GMM)作為分類器。此方法對語音合成及文字轉語音這類的攻擊有不錯的成效，故將此方法沿用至回放攻擊的偵側。

近年來，我們可以看到深度學習方法被廣泛應用在語者辨識系統中，並且取得了相當不錯的成績(Yamada, Wang & Kai, 2013) (Sarkar, Do, Le & Barras, 2014) (Lei, Scheffer, Ferrer & McLaren, 2014) (Kenny, Gupta, Stafylakis, Ouellet & Alam, 2014) (Variani, Lei, McDermott, Lopez Moreno & Gonzalez-Dominguez, 2014) (Lin, Mak & Chien, 2017)，但回放攻擊偵測系統則是一個尚未解決的問題。在這篇論文中，我們提出了一個基於深度自編碼解碼器(Chen, Sun, Rudnicky & Gershmandoi, 2016) (Bone, Lee & Narayanan, 2014) (Huang, Wu, Su & Fu, 2017) (Chung, Wu, Shen, Lee & Lee, 2016) (Richardson, Reynolds & Dehakdoi, 2015)的架構來作為偵測系統的基礎，稱之為鑑別式自編碼解碼器(Discriminative Autoencoders, DcAE) (Lee *et al*., 2017) (Yang *et al*., 2017)，我們嘗試使用自編碼解碼器來替換傳統分類方法，並且設計目標函數來偵測語者辨識攻擊。自編碼解碼器為一種對稱的神經網路架構，在此架構中分成編碼層以及解碼層，目標是希望輸出端能夠經由訓練重建輸入端的資料，為了最小化重建誤差，並且達到偵測語者辨識攻擊。我們也提出了一個新的損失函數，稱之為合頁損失(Hinge Loss)，來訓練這個鑑別式自編碼解碼器。根據這個損失函數，每一個經由音訊所輸入的音框將會編碼成相同的 identity codes(i-codes)，而這個 i-codes 就會具有辨別種類的能力，因此所有相同類別的 i-codes 則會表示成近似的特徵，使得 i-codes 可以利用簡易的分類法便可以輕易地分別數據。此外，在合頁損失的部分我們加入了可調動的邊界函數，此邊界函數可以限制自編碼解碼器所更新的權重，因此每次的更新將可以注重在最令模型困惑的數據上，提高模型泛化(Generalization)的能力。

本論文的主要貢獻包括：第一，i-codes 是一個全新的特徵表示法，並且此特徵能夠有效地偵測語者辨識攻擊，並達到有效分類的效果。第二，在鑑別式自編碼解碼器中，中間層具有聚集相同類型語句的能力，此方法為一個新穎且有效的類神經網路架構。第三，在多任務的鑑別式自編碼解碼器中，我們能夠結合額外的資料來改善我們所提出的系統架構，並且提高泛化的能力。

本論文的後續安排如下：第二節簡單介紹了 ASV-spoof 2017 Challenge；第三節敘述了我們所提出的兩個基於自編碼解碼器的架構；第四節介紹實驗語料與設定以及評估的方法跟結果，最後一節則是結論與未來研究方向。

2. 任務描述 (Task Description)

ASVspoof 是一個設定為偵測語者辨識攻擊的比賽，而在 2017 年，則注重在解決錄音回放攻擊，其中所提供的語料庫是來自於 RedDots 的語料庫跟此語料的錄音回放版本。在

這次比賽所提供的訓練數據中，包含了十位男性的語者，其中分別製作成 1508 句的真實語料和 1508 句的回放錄音語料，並且在錄音回放的設備中，利用了六種錄音設備以及三種播放設備。在開發數據中，包含了八個不同的語者和 760 筆真實語料以及 950 筆回放錄音語料，而在錄音回放設備中，則是有十種錄音及播放設備的組合。

在 ASVspoof-2017 中，主辦方所提供的評估標準為無門檻值的相等錯誤率(Equal Error Rate, EER)，意即在計算偵測系統時，在錯誤接收率(False Acceptance Rate, FAR)以及錯誤拒絕率(False Rejection Rate, FRR)相等時所得到的錯誤率。愈好的偵測系統，就會有愈低的相等錯誤率。

3. 系統描述 (System Description)

3.1 特徵抽取 (Feature Extraction)

根據 ASVspoof-2017 所提供的官方文件，常數 Q 倒頻譜係數對語者辨識攻擊偵測有良好的效果(Todisco, Delgado & Evans, 2016)，因此我們使用了常數 Q 倒頻譜係數作為特徵抽取的方法。常數 Q 倒頻譜係數為一個基於常數 Q 轉換(constant Q transform, CQT)的特徵表示法，這個方法最初使用在音樂訊號處理，並且在抽取的過程確保了頻譜上所有頻域的資訊都能夠被有效的保留下來，其計算過程可用下式表達：

$$\mathrm{CQCC(p)} = \sum_{l=1}^{L} \log|X^{CQ}(l)|^2 \cos\left[\frac{p\left(l-\frac{1}{2}\right)\pi}{L}\right] \tag{1}$$

其中 X^{CQ} 為經過 CQT 轉換輸入訊號後所得到的值，而 $p = 0, 1, \cdots, L\text{-}1$ ，l 則是取樣的頻帶(Frequency Bin)。

3.2 鑑別式自編碼解碼器 (Descriminative Autoencoder)

在自編碼解碼器中，主要的架構分為兩部分，編碼器 f 以及解碼器 g，在編碼器部分，自編碼解碼器會在隱藏層中建立出特別的特徵表示法 (Goodfellow, Bengio & Courville, 2016)，並且使得解碼器能夠利用這樣的特徵達到重建輸入資訊的效果，也就是 $X \xrightarrow{f} H \xrightarrow{g} X$，其中 H 為資料經過隱藏層所產生出來特徵表示法，為了訓練自編碼解碼器中的參數，來達到重建的效果，則需要利用重建誤差來更新參數，重建誤差的表示如下：

$$F_r(X) = \frac{1}{|X|}\sum_{x\in X} ||y\text{-}x||_2^2 \tag{2}$$

其中 $y = g(f(x))$ 是重建出來的結果，$\|\cdot\|_2$ 則是 2-範數(2-norm)，$|X|$ 則是樣本的數量，由於此模型會限制並且強迫輸入值可以被複製到輸出，因此可以從輸入資料樣本中學到有用的資料特性。

根據 (Krizhevsky, Sutskever & Hinton, 2012)，我們可以假設 H 中包含了許多有價值的特徵，並且也包含了代表回放聲音的特徵，H_g 為真實錄音語料的集合，H_s 則為錄音回

放語料的集合，因此我們提出了兩個特別的合頁損失來使得隱藏層中的特徵更加具有代表性；這兩個合頁損失適用於更新隱藏層，其表示如下：

$$F_P(H) = \frac{1}{H}\sum_{h_i,h_j\in H_g|h_i,h_j\in H_s} f\left(M_p - < h_i, h_j >\right)^2 \quad (3)$$

$$F_n(H) = \frac{1}{H}\sum_{h_i,h_j\in H_g|h_i,h_j\in H_s} f\left(< h_i, h_j > - M_n\right)^2 \quad (4)$$

其中 F_p 為正合頁損失，F_n 為負合頁損失。在經過合頁損失訓練的隱藏層中，將會提供一個獨立的子空間來分離真實說話人聲或是錄音回放人聲，使我們能夠達到分離兩種類別的目標；因為我們專注在解決資料的複雜性，合頁損失能夠有效地面對這類型的問題。舉例來說，式(3)中，相同類別之間的內積若是小於邊界 M，代表此資料配對結果相似度高但不具有代表性，將會更新這個模型中神經元的權重，使同一類別之間內積能夠提高，反之亦然，使得鑑別式自編碼解碼器能夠更新權重達到想達到的目的，進而分辨出輸入的資料為真實人聲與否，在式(3)及式(4)中 f 則為 ReLU (Luong, Le, Sutskever, Vinyals & Kaiser, 2015)，這個非線性轉換方程式能夠有效限制更新權重與否。最後，結合式(2)、式(3)及式(4)，在訓練過程中，模型的目標函數則如下所表示：

$$\alpha\left(F_r(X)\right) + \beta\left(F_p(H) + F_n(H)\right) \quad (5)$$

其中 α 控制重建誤差所佔的權重，β 則控制了隱藏層在此模型中佔的重要性。

在論文中，我們將鑑別式自編碼解碼器分成兩種不同的架構，如圖 1 所示。

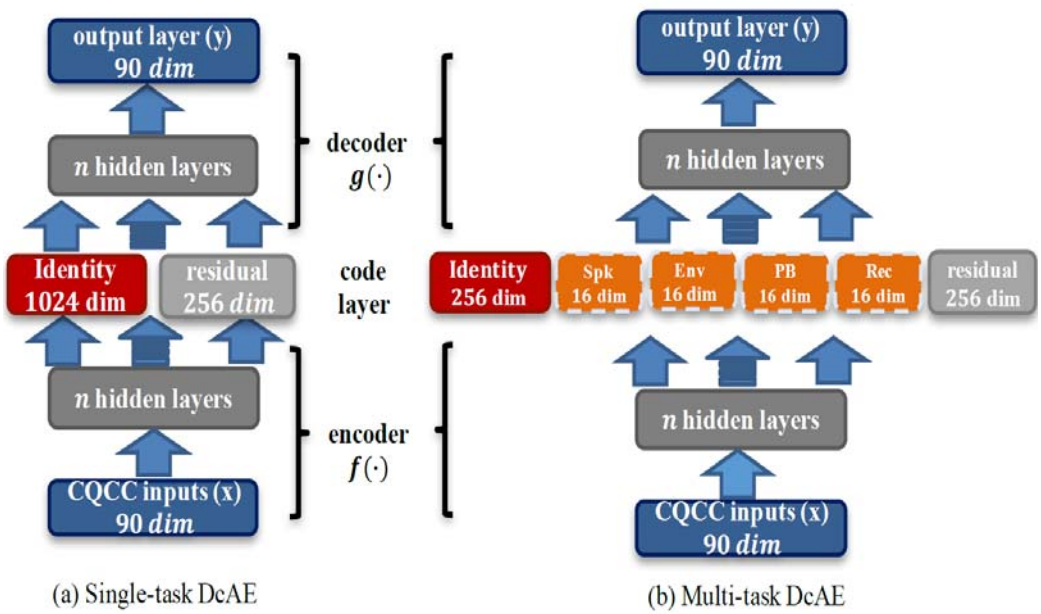

圖 1. 鑑別式自編碼解碼器示意圖。

[Figure 1. The two kinds of architecture in DcAE]

這兩個架構中，都是將輸入 *x* 經由模型中的 i-codes 轉換成 *y*，然而隱藏層的設計卻截然不同；圖左所表示的為單任務鑑別式自編碼解碼器，在此模型中，中間層被分成兩個部分，一個部分我們稱之為鑑別層，另一個則為剩餘層，在鑑別層中，我們加入了設計好的合頁損失，使得經由鑑別層所產生出來的特徵具有分開類別且避免同類別資料發散的能力，使得更能被辨別，而剩餘層則是希望能夠使解碼的部分較容易達到重建，另一方面，在多任務的鑑別式自編碼解碼器中，則是模仿多任務的類神經網路模型 (Liu *et al*., 2015) (Glorot & Bengio, 2010)，多任務模型具有利用額外的訓練資訊，使得模型泛化能力上升以及有效提升預測準確度，而在此，我們將額外資訊分成語者，錄音環境，錄音設備，以及回放設備，這些資訊皆能從 ASVspoof-2017 的官方資料中獲得，在多任務模型中，除了剩餘層外，我們都加入了合頁損失來讓每層都能夠跟著目標函數更新，提供更多的額外資訊，幫助主要目標來分類真實人聲或是回放人聲。此外，在編碼層及解碼層中則加入了隱藏層，使得此模型具有更多的參數，達到所謂深度學習的效果。

4. 實驗設定以及實驗結果 (Experiment Setting and Result)

在這一節，我們將會一一介紹實驗的設定以及所達到的結果，並且能夠加以討論，首先，在此篇論文中，我們會利用高斯混合模型以及類神經網路模型當作比較的標準，對於高斯混合模型，ASVspoof-2017 官方所提供的系統架構如下：利用常數 Q 倒頻譜係數抽取語料的特徵，並且利用此特徵訓練一個高斯混合模型，使得此模型具有分類的能力；另外，在類神經網路架構中，我們使用了三層隱藏層，並且每層有著 1024 個神經元，相同地，也利用了常數 Q 倒頻譜係數作為抽取語料的特徵表示法，此類神經網路的輸出則是用二元分類來判斷是否為真實人聲，另外，由於我們有提出多任務的鑑別式自編碼解碼器，於是我們也訓練了一個多任務類神經網路來當作比較的標準。在常數 Q 倒頻譜係數的設定中，我們抽取了 90 維的特徵向量作為所有系統的輸入向量，並且抽取完特徵後，我們使其標準正規化，並且使用訓練數據當作標準特徵來轉換開發數據。

在鑑別式自編碼解碼器中，我們將架構中，最中間的層分成了鑑別層以及剩餘層，在鑑別層中，常數 Q 倒頻譜係數將會被編碼為 1024 維的 i-codes，這裡所產生的 i-codes 將會具有分類是否為回放攻擊的能力，而特徵在剩餘層中則被編碼為 256 維；在合頁損失中，我們將正向邊界設定為 10，負向邊界設定為-10。在自編碼解碼器中，所有權重的初始值則是利用 Glorot Uniform 作為初始化設定，在所有的神經元中，我們則是選擇 tanh 作為激發函數(Activation Function)，除了最後一層用來還原數據則是利用選用了 linear 作為激發函數，另外，梯度下降的最佳化演算法則選用 Adaptive Moment Estimation(Adam) (Kingma & Ba, 2014)，使其能快速且有效的達到最佳化的目標。在多任務的鑑別式自編碼解碼器中，我們增加了中間層的數量，使得輸入特徵被額外編碼為其他的資訊，在這些額外的編碼層中，則設定為 16 個神經元，最後，我們則是利用向量內積來計算 i-codes 之間的相似度，因此可以利用此結果作為分類的依據，進而計算相等錯誤率。

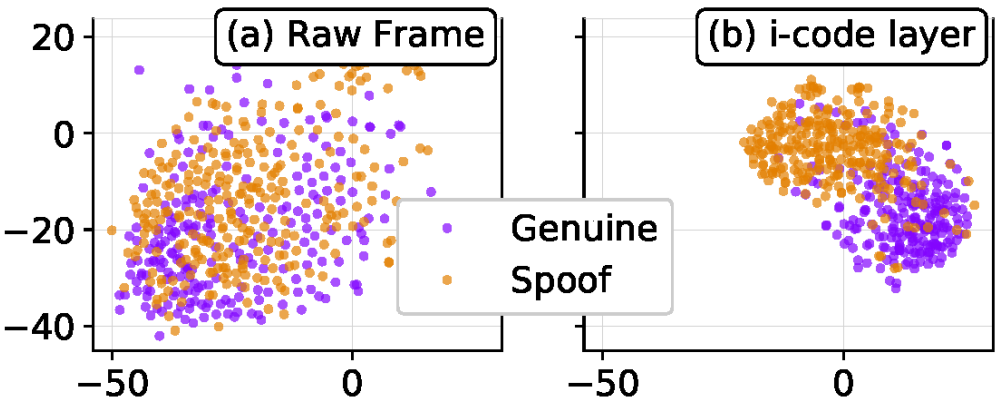

圖 2. 利用 t-SNE 表示在單任務鑑別式自編碼解碼器中，i-codes 的分佈情形。
[Figure 2. The result of single-task DcAE that t-SNE maps high dimension i-codes into 2 dimension.]

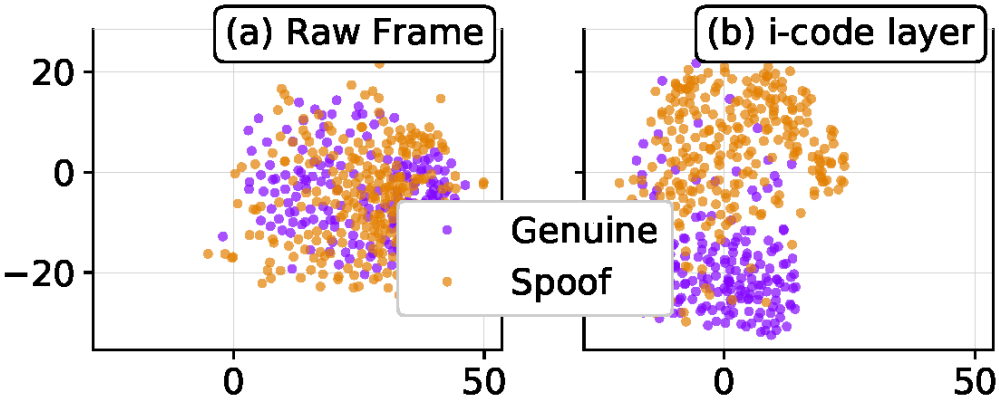

圖 3. 利用 t-SNE 表示在多任務鑑別式自編碼解碼器中，i-codes 的分佈情形。
[Figure 3. The result of multi-task DcAE that t-SNE maps high dimension i-codes into 2 dimension.]

表 1. 相等錯誤率在開發數據集上的結果。
[Table 1. Summary of development result in ASVspoof Task]

Method	EER(%)
Baseline	10.35
DNN	8.18
Multi-task DNN	7.6
Single-task DcAE	6.43
Multi-task DcAE	**5.99**

由表 1 可得知，由高斯混合模型作為 Baseline 方法的效果遠差於鑑別式自編碼解碼器所達到的效果，另外對於一般的深度類神經網路來說，我們的模型也能有較好得效果，在視覺化呈現的部分，如圖 2、圖 3，我們利用了 t-Distributed Stochastic Neighbor Embedding(t-SNE) (van der Maaten & Hinton, 2008) 來表示，由此圖可發現，經由鑑別層所產生出來的 i-codes 確實達到了使輸入數據依照目標分離的效果，由此可見，經由簡單的向量內積來計算即可快速的分辨輸入語料是否為真實人聲。

5. 結論 (Conclusion)

在這篇論文中，我們提出了一個全新的鑑別式自編碼解碼器來參與這次的 ASVspoof-2017，在這個新的架構中，我們加入了新設計的目標函數，使得我們可以利用新的特徵表示法 i-codes 來達到辨別是否為真實人聲的目標，另外，我們同時利用了多任務模型來增強預測的結果，最後，相比於高斯混合模型以及深度類神經網路，鑑別式自編碼解碼器更能夠使資料具有辨別的價值。在未來，我們希望能夠加以延伸發展此模型，以建立一個更泛用的系統架構。

致謝 (Acknowledgement)

特別感謝由法務部調查局所提供的贊助來完成此篇論文研究(Grant number: 106-1301-05-04-01-3)。

參考文獻 References

Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1990). Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, *11*(2), 71-76.

Alam, M. J., Kenny, P., Bhattacharya, G. & Stafylakis, T. (2015). Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. In *Proceedings of Interspeech 2015*, 2072-2076.

Alegre, F., Amehraye, A. & Evansdoi, N. (2013). Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2013.6638222

Bone, D., Lee, C.-C. & Narayanan, S. (2014). Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *IEEE Transactions on Affective Computing*, *5*(2), 201-213. doi: 10.1109/TAFFC.2014.2326393

Chen, L.-H., Ling, Z.-H., Liu, L.-J. & Dai, L.-R. (2014). Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech and Language Processing(TASLP)*, *22*(12), 1859-1872. doi: 10.1109/TASLP.2014.2353991

Chen, Y.-N., Sun, M., Rudnicky, A. I. & Gershmandoi, A. (2016). Unsupervised user intent modeling by feature-enriched matrix factorization. In *Proceedings of 2016 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6150-6154. doi:10.1109/ICASSP.2016.7472859

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y. & Lee, L.-S. (2016). Audio Word2Vec: Unsupervised Learning of Audio Segment Representations Using Sequence-to-Sequence Autoencoder. In *Proceedings of Interspeech 2016*. doi:10.21437/Interspeech.2016-82

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics*, *9*, 249-256.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT press.

Huang, K.-Y., Wu, C.-H., Su, M.-H. & Fu, H.-C. (2017). Mood detection from daily conversational speech using denoising autoencoder and LSTM. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5125-5129. doi:10.1109/ICASSP.2017.7953133

Kenny, P., Gupta, V., Stafylakis, T., Ouellet, P. & Alam, J. (2014). Deep neural networks for extracting Baum-Welch statistics for speaker recognition. In *Proceedings of Odyssey 2014*, 293-298.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representation*. Retrieved from https://sarXiv preprint arXiv:1412.6980

Kinnunen, T., Evans, N., Yamagishi, J., Lee, K. A., Sahidullah, Md., Todisco, M. & Delgado, H. (2017). ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training, 10*, 1508.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in neural information processing systems*, 1106-1114.

Lee, H.-S., Lu, Y.-D., Hsu, C.-C., Tsao, Y., Wang, H.-M. & Jeng, S.-K. (2017). Discriminative autoencoders for speaker verification. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5375-5379. doi:10.1109/ICASSP.2017.7953183

Lei, Y., Scheffer, N., Ferrer, L. & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2014.6853887

Lin, W.-w., Mak, M.-W. & Chien. J.-Z. (2017). Fast scoring for PLDA with uncertainty propagation via i-vector grouping. *Computer Speech & Language*, *45*, 503-515. doi:10.1016/j.csl.2017.02.009

Liu, X., Gao, J., He, X., Deng, L., Duh, K. & Wang, Y.-Y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *proceedings of HLT-NAACL 2015*.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O. & Kaiser, L. (2015). Multi-task sequence to sequence learning. In *Proceedings of ICLR 2016*. Retrived from https://arXiv preprint arXiv:1511.06114

Richardson, F., Reynolds, D. & Dehakdoi, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671-1675. doi:10.1109/LSP.2015.2420092

Sarkar, A. K., Do, C.-T., Le, V.-B. & Barras, C. (2014). Combination of cepstral and phonetically discriminative features for speaker verification. *IEEE Signal Processing Letters*, *21*(9), 1040-1044. doi: 10.1109/LSP.2014.2323432

Todisco, M., Delgado, H. & Evans, N. (2016). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Proceedings of Odyssey 2016*. doi: 10.21437/Odyssey.2016-41

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Van Santen, J. P. H., Sproat, R., Olive, J. & Hirschberg, J. (2013). *Progress in speech synthesis*. New York, NY: Springer Science & Business Media.

Variani, E., Lei, X., McDermott, E., Lopez Moreno, I. & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4080-4084. doi: 10.1109/ICASSP.2014.6854363

Villalba, J., Miguel, A., Ortega, A. & Lleida, E. (2015). Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In *Proceedings of Interspeech 2015*, 2067-2071.

Wu, Z., Kinnunen, T., Evans, N. W. D., Yamagishi, J., Hanilci, C., Sahidullah, M. & Sizov, A. (2015). ASVspoof 2015 - the first automatic speaker verification spoofing and countermeasures challenge. In *Proceedings of Interspeech 2015*.

Xiao, X., Tian, X., Du, S., Xu, H., Siong, C. E. & Li, H. (2015). Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In *Proceedings of Interspeech 2015*.

Yamada, T., Wang, L. & Kai, A. (2013). Improvement of distant-talking speaker identification using bottleneck features of DNN. In *Proceedings of Interspeech 2013*.

Yang, M.-H., Lee, H.-S., Lu, Y.-D., Chen, K.-Y., Tsao, Y., Chen, B. & Wang, H.-m. (2017). Discriminative autoencoders for acoustic modeling. In *Proceedings of Interspeech 2017*.

Ze, H., Senior, A. & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2013.6639215