課程名稱：人工智慧

學期報告名稱：人工智慧相關應用─以google duplex為研究

授課教授：陳鍾誠

系級：電四乙

學號：110310232

姓名：官振鵬

1. 大綱

拆解在2018年Google I/O上所展示的Google Duplex，分成ASR、TFX、WaveNet然後Tacotron2最後Google Assistant，接下來介紹Duplex的工作，然後再做探討。希望透過了解Google Duplex的研究，如果能從中得到一些而外的想法，應該能改變創造的方向與想像。

1. 研究動機與目的

每年的GOOGLE I/O中，GOOGLE都會發布或展示新的軟硬體和相關的應用，每次都讓人驚呼、為之讚嘆，而這次的2018年展出中，讓大家最驚訝的莫過於就是當Google Assistant幫用戶，也就是使用者去訂餐廳的位子時，並不是網路的網頁預約，也不是轉介真人客服，而是使用人工智慧的Google助理打電話給餐廳進行預約，那和人維妙維肖的電話溝通，讓在場和觀看線上直播的各位都看呆了。同時網路也出現不少相關的討論與新聞，有的在讚嘆技術的進步，也有的擔心發展的同時所帶出的危險。於是，我相信只有深入了解，才能釐清所有的擔憂。

1. 文獻探討
2. ASR與STT

ASR全名Automatic Speech Recognition，或被稱之為STT（Speech-to-Text），這邊說的將會指Google語音辨識服務，是自然語言處理的領域之一（NLP：natural language processing）。

使用上已經相當廣泛，無論是在Google翻譯裡面，到Youtube中的自動偵測的字幕，都是使用ASR的應用。

在這我們引用(12)中的內容，英文原文來自tensorflow.org，透過TensorFlow（用於Google許多產品）來製作簡單的語音辨識。但由於我自己並不會使用TensorFlow，所以這邊只做粗略介紹，這個教學來自Google自己，需要透過Speech Commands資料庫，也就是指令語言的數據集合，在這裡說明一下這裡面的內容是透過CC BY授權，大小大於1GB，而不是竊取我們的通話之類的。

這資料庫中是有六萬五千個範例，而將這些範例丟進訓練模型中，在經過幾百次的步驟後，可以看出系統對無聲輸入的反應，如下圖3-1，在第一行是無聲的剪輯，而判斷0便是無聲，其他行分別是：未知單字，“yes”、“no”、“up”、“down”、“left”、“right”、“on”、“off”、“stop”或“go”。完成後，這個模型可以會嘗試分辨一秒的聲音為其中一個指令。

|  |
| --- |
| I0730 16:57:38.073667 55030 train.py:243] Confusion Matrix:  [[258 0 0 0 0 0 0 0 0 0 0 0]  [ 7 6 26 94 7 49 1 15 40 2 0 11]  [ 10 1 107 80 13 22 0 13 10 1 0 4]  [ 1 3 16 163 6 48 0 5 10 1 0 17]  [ 15 1 17 114 55 13 0 9 22 5 0 9]  [ 1 1 6 97 3 87 1 12 46 0 0 10]  [ 8 6 86 84 13 24 1 9 9 1 0 6]  [ 9 3 32 112 9 26 1 36 19 0 0 9]  [ 8 2 12 94 9 52 0 6 72 0 0 2]  [ 16 1 39 74 29 42 0 6 37 9 0 3]  [ 15 6 17 71 50 37 0 6 32 2 1 9]  [ 11 1 6 151 5 42 0 8 16 0 0 20]]  **圖3-1 經過400步驟後的紀錄-混淆矩陣** |

而最終，經過幾個小時的練習後，完成一萬八千步後，最終印出的混淆矩陣有近九成的準確度。

1. TFX

TFX: A TensorFlow-Based Production-Scale Machine Learning Platform，是指建立於TF上的包含機器學習整個生態的完整系統，不僅僅是提供模型訓練等，還有數據驗證與發佈相關的功能。

其整體設計原則如下四點：

* 建構可服務於多個學習任務的單一平台
* 支援持續訓練和服務
* 人工干預
* 可靠性和穩定性

相關的流程：數據分析→數據轉換→數據驗證，再來便是模型訓練→模型評估和驗證。

* 定義「好」模型

這個模型的需求是準確的答案還是穩定的回答，或是不佔用太多資源，根據不同需求，都會需要不同的設定。

* 驗證的敏感度

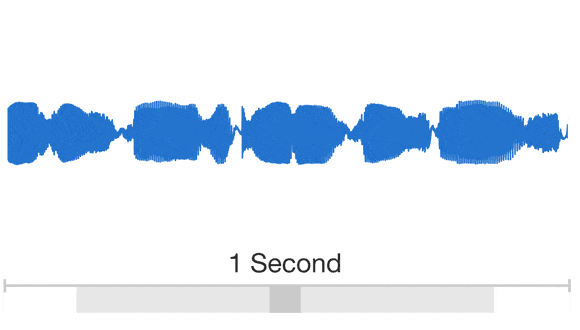
過於敏感只要稍有波動就會警報，相反的，如果過於不敏感，就會錯過漏掉問題，專欄的作者建議，看專案的重要性，業務的話就要敏感些。

* 分片驗證

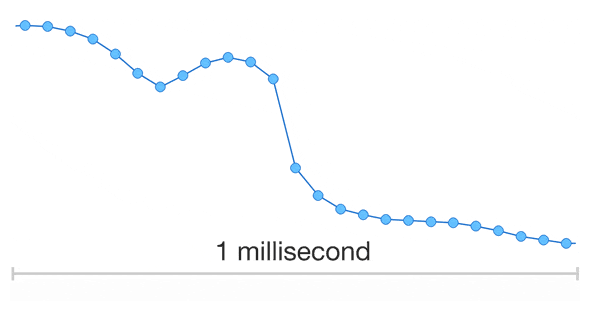
除了對模型做整體的驗證，有時需要對部分的數據做分片驗證。

1. WaveNet與TTS

要是說ASR的反向，那就是TTS了，但在GOOGLE中有個重要的部分，就是WaveNet，來自於有名的DeepMind。更加由技術性而言，是來自於2016年的論文。

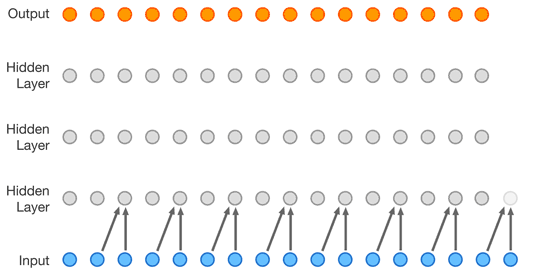


**圖3-2 放大前的聲音（5）**



**圖3-2 放大後的聲音節點（5）**

使用前饋神經網絡如同CNN：Convolutional Neural Network、卷積神經網絡，來訓練這個模型，去預測每個資料點的下個自然點，非傳統的TTS系統採用的拼接方法，就能創造更加自然的輸出效果。其最大的問題便是，使用這種回饋式的神經網絡，花太多時間了。於是乎Google使用運行於雲端的TPU來做處理，在模型優化後擁有1000倍的效率提升，並且只要花50毫秒就能產生一秒的聲音。

****

為了知道新的語音模型有多好，會反映在使用者的平均意見分數量表中，而在滿分五分的量表中，WaveNet得到4.1分的高分。而由於它的模型生成是使用訓練神經網絡的方式，不是使用拼接的方式，所以不用這麼多的錄製音頻，就能做出高音質的模型，結果就是其製造新的模型非常快速。

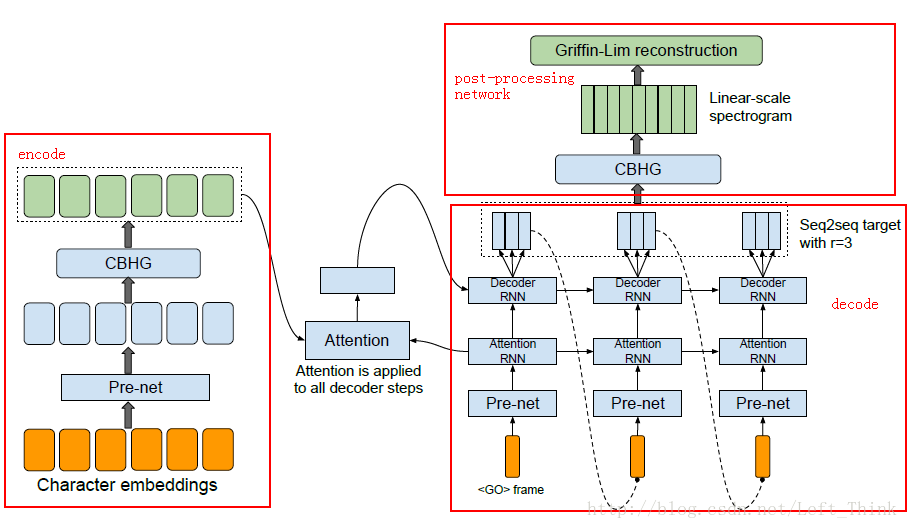
1. Tacotron2

這是一種端到端的方式，輸入文本後就直接輸出語音，使用端到端的好處如下：

1.減少特徵的工作工程，只需要輸入文本，其他的特徵由模型本身自行學習。

2.添加各種條件，如不同語言的重音節，人說話的語調和感情。

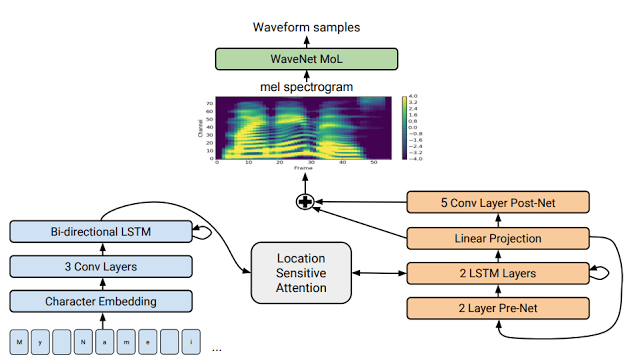
3.避免掉多個子模型所造成的累積傳遞誤差。



**圖3-4 tecotron結構（6)**

透過上圖3-4能看到，tecotron的結構是能分成不同的部分，所以一方面它是一個生成式模型，如WaveNet和DeepVoice；另一方面，它是一個端到端模型，如：Wang、Char2Wav，不過端到端模型最終還是需要語音合成，也就是說Tacotron是透過輸入的本文，經由轉換最後輸出一聲學參數，再進行合成語音，詳細在(7)。

而二代的Tecotron其實一點也不意外，便是將合成語音的部分，由WaveNet來完成，可以解決其速度問題，並且得到更高的擬真(8)。



**圖3-5 tecotron2結構（8)**

1. Google Assistant



**圖3-6 google assistant的標誌與標語（9)**

在2016年的5月，GOOGLE發表了個人助理。與GoogleNow的功能十分相像，不過新的Google助理透過自然語言處理和雙向對話。

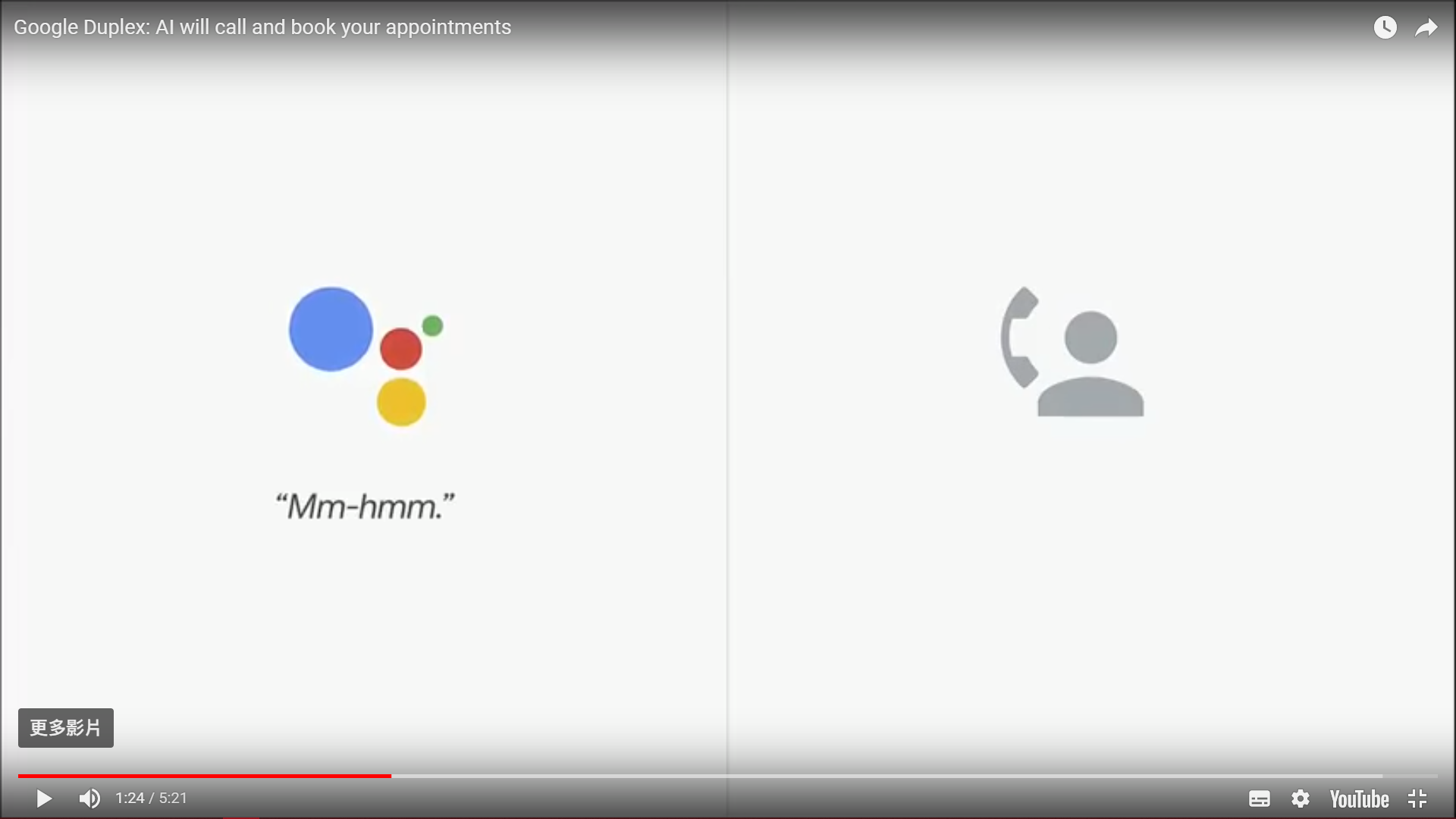
更加簡單的說法便是GOOGLE版本的SIRI吧，在人工智慧的蓬勃發展下，行動助理一直都是大家所期望能擁有的功能，而多虧了ASR的健全，SIRI對語言的準確度也大幅度上升，加上雲端的計算，為人們的生活方便不少。

而Google Assistant是類似的應用，但其強調使用更好的回應聲音，使用WaveNet的語音更加自然，而加上更強的人工智慧運算晶片，能進行持續的對話，並更加能理解不同規則的文法，不再只適用於例句的回答等。

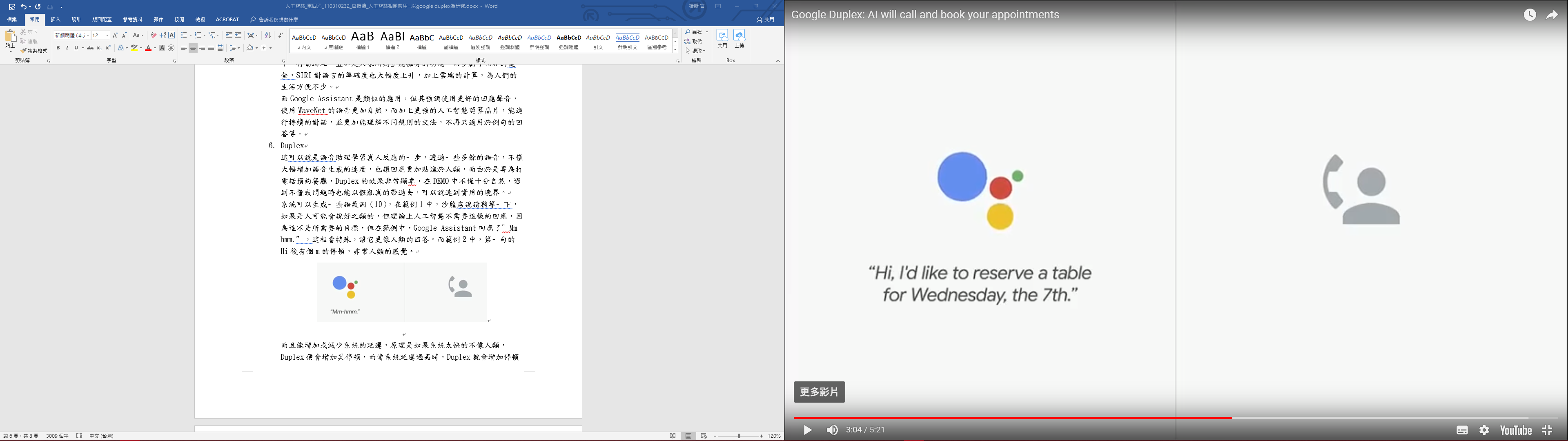
1. Duplex

這可以說是語音助理學習真人反應的一步，透過一些多餘的語音，不僅大幅增加語音生成的速度，也讓回應更加貼進於人類，而由於是專為打電話預約餐廳，Duplex的效果非常顯卓，在DEMO中不僅十分自然，遇到不懂或問題時也能以假亂真的帶過去，可以說達到實用的境界。

系統可以生成一些語氣詞（10），在範例1中，沙龍店說請稍等一下，如果是人可能會說好之類的，但理論上人工智慧不需要這樣的回應，因為這不是所需要的目標，但在範例中，Google Assistant回應了”Mm-hmm.”，這相當特殊，讓它更像人類的回答。而範例2中，第一句的Hi後有個m的停頓，非常人類的感覺。



**圖3-7 google assistant的範例一截圖（10)**



**圖3-8 google assistant的範例二節圖（10)**

而且能增加或減少系統的延遲，原理是如果系統太快的不像人類，Duplex便會增加其停頓，而當系統延遲過高時，Duplex就會增加停頓詞，更加極端的時候，更能直接使用快速逼近模型。

雖然聽起來不可思議，好像也相當有用，這還是讓許多人十分反彈，有些人對於這樣欺騙人類的模式敬謝不敏，也有的害怕被往犯罪之類的用法。

1. 探討

在經過這麼多資訊後，對GOOGLE的個人助理有更進一步的了解。當時看到I/O 2018時，我真的非常的驚訝，但心裡其實也有些疑問以及不安感，我想對於更不了解資訊相關的工作和領域的人想必會更加害怕吧，而雖然我算是電子科系，高中時也能算是資訊類組的高職生，但談到人工智慧與深度學習，也只能說出隻字片語，要想了解的更深，但卻受限於專業，我想借鏡其他的高人是最快的了。

GOOGLE行動助理主要是建立於發展日漸健全的TTS和SST，也就是說程式必須學會聽懂我們的話，也必須能使用語言回應我們，跳脫以前的文字交談，但說到底其實是相同的基礎，所以重點便在TTS和SST上面了。而科技不會有停下的一天，即便現在的SIRI小姐或GOOGLE小姐的聲音並不擬真，但確實的可以明白其想表達的意思，在我的觀點上，這已經是可以使用的功能，但科技不會就此止步，於是Tacotron和WaveNet便誕生了，這也是相當突破的發展，使用神經網絡來訓練模型，可以比之前建構更快、更擬真。我個人是非常歡喜的看待到來，更好聽的GOOGLE小姐以及更加完善的翻譯，甚至是在開發APP上也十分有幫助。

但最後還是必須說到Google Duplex這個惹人爭議的話題，我相信科技進步神速，同時水能載舟已能覆舟，而Duplex在我眼裡看起來像走火入魔或巴洛克藝術，在已經能完成的WaveNet後，加上看似多餘的停頓，是增加了擬真感，但我們真的需要嗎？需要多餘的停頓與贅詞嗎？我無法給出答案，我希望我的手機能幫我預定餐廳或行程，也希望在對談中更擬人，甚至有和人對話的感覺；但同時如果我在餐廳上班，說不定會直接掛掉它（Google表示會主動先表示自己是GOOGLE個人助理)。

從前，我們操弄程式，人們嘲笑SIRI，嘲笑GOOGLE有時講錯的發音，同時它們幫助我們工作、休閒，但從現在開始，我們也有可能被操弄，我們所收到的簡訊可能來自人工智慧，我們收到的電話可能不再是語音留言，而是人工智慧，更加驚人的是，我們工作的對象也有可能透過人工智慧和我交談，到頭來我是和他還是它工作呢？

我知道這問題終究會到來，當人分不清楚兩者的差異時，我們要怎麼防止人類的權利受損呢？

1. 資料來源與致謝
   1. http://coolmandiary.blogspot.com/2017/06/virtual-assistant-ttsstt.html
   2. https://zh.wikipedia.org/wiki/自然語言處理
   3. https://zhuanlan.zhihu.com/p/31041536
   4. https://buzzorange.com/techorange/2017/10/11/deepmind-wavenet-1000-times-faster/
   5. https://deepmind.com/blog/wavenet-generative-model-raw-audio/
   6. https://blog.csdn.net/Left\_Think/article/details/74905928
   7. https://gist.github.com/candlewill/aeaca14975d8ba4417f6a99ea91f196b
   8. https://www.inside.com.tw/2018/01/01/tacotron-2
   9. https://assistant.google.com/#?modal\_active=none
   10. https://www.leiphone.com/news/201805/99jY3lvPZ9Oh2m7j.html
   11. https://www.youtube.com/watch?v=ogfYd705cRs
   12. https://cloud.tencent.com/developer/section/1475685