語音處理

搭起人機界面的橋梁

簡仁宗

在科幻電影裏，常見的一個情節是主角在太空船中與船上的電腦系統相互對話。人們藉由語音直接與電腦進行溝通，或向電腦下達命令及詢問問題，而電腦也以語音進行回應，這樣的過程是不是讓你感到十分好奇呢？究竟其中的奧妙在哪裡？為什麼電腦可以聽得懂你說的話，更可以像我們一樣用語音回應呢？

在不久的未來，人與電腦將可以用語音互動

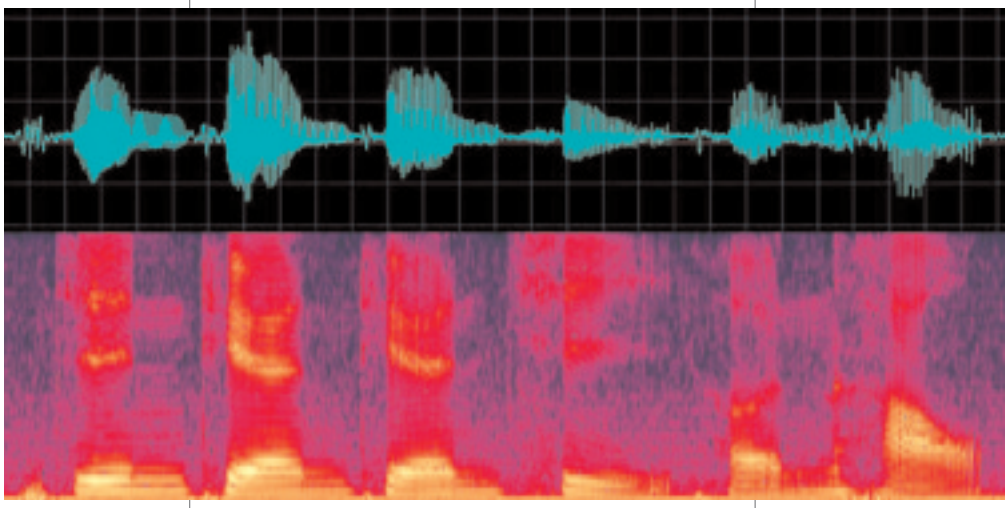
語音是人與人之間最自然的溝通方式，透過語音，人們無須借助額外的工具便可以彼此交換各種資訊。雖然我們從眼睛所接收到的刺激，比從耳朵所接受到的要來得多，但是不可否認的，要與他人進行溝通，使用語音絕對比只用眼睛更有效率。

近幾十年來，由於資訊科技的蓬勃發展，在日常生活中電腦幾乎已經成爲不可或缺的一部分，爲了讓人類也能藉由語音對電腦下指令、向電腦詢問問題，甚至讓電腦能自然地與人們交談，在各國學者們不斷地努力下，早已開啓了語音處理的研究領域，期望早日實現人機自然交談的夢想。

語音的本質有哪些？

語音是一種聲波，它本身帶有語言學上的相關資訊，即我們說出的這段語音所對應的文字訊息。除此之外，語者的聲調高低、抑揚頓挫，

以及語者說話當時的情緒，也可以從語音中得知。因此，使用語音交換資訊，在我們的日常生活中占有相當重要的地位。讀者們可以回想一下和別人講話時的情景，在高興時和悲傷時所講的話，語氣是不是不一樣呢？在和好朋友哈啦時和與喜歡的女生講話時，語氣是不是也不一樣呢？在回答老師所提的問題時和主動去教導同學時，語氣是不是也不相同呢？這些都是語音本身具有的資訊，而我們人類可以從中得知對方於交談中不斷變換的語意、語氣及情緒。至於電腦呢？就目前而言，還沒有辦法達到這樣的境界。



「今天天氣很好」的語音波形與其對應的頻譜圖。上半是語音錄製到電腦中直接根據時間軸所繪出的波形圖，我們可以觀察到不同的字詞在波形上有不同的表現；下半的頻譜圖則是將語音以不同於波形的另一種形式表達出來。

基本上，可以從兩個方面來觀察一段語音訊號的特性。其一是語音波形，這是經由音效卡將語音訊號轉錄爲數位資料後，以時間爲橫軸，將每一個時間點所對應的能量繪出的圖形。另一則是頻譜圖，如果我們視語音爲一種密碼，那麼頻譜圖就是提供一種用來將語音歸屬到不同密碼類別的最佳方式。

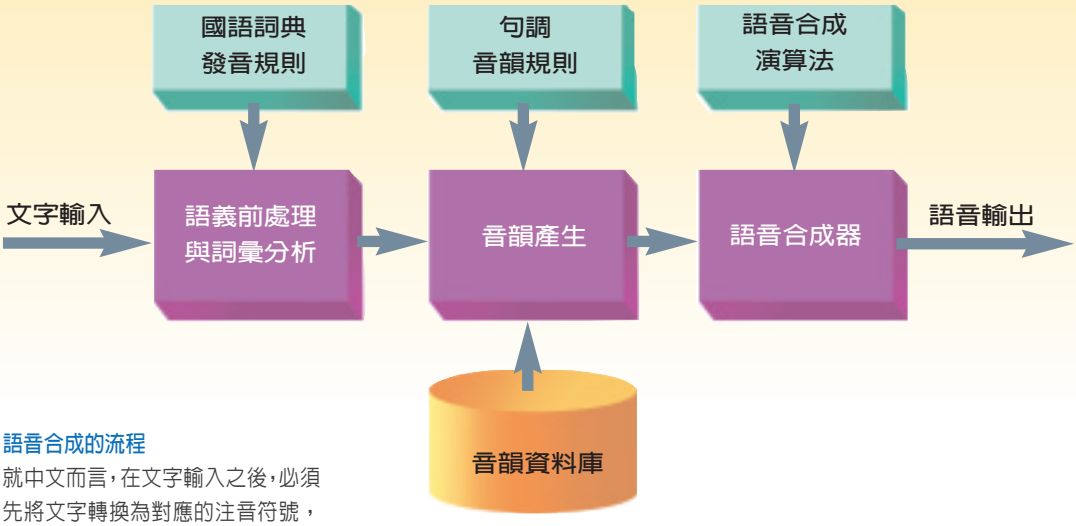
頻譜圖的產生是一種翻譯的過程，它將我們耳朵所處理的語音及其他聲音轉換成可見的影像，供我們進一步觀察。在很早以前，人類就已經有能力將聽覺資訊轉換爲另一種有意義的語言。然而，這樣的轉換過程我們還沒有辦法去分析它，只知道這樣的過程無須透過人類的意識，而是直接由神經元來完成轉換。

利用人工合成語音，讓機器可以與人對話

語音合成或稱爲語音生成，是一項取代傳統上人們只能由電腦螢幕上讀取電腦回應訊息的重要技術。在早期，爲了讓電腦能藉由語音回報使用者所下指令的結果，必須先行錄製各種不同的語音音檔儲存在電腦上，當電腦需要進行回應的時候，再從中找尋對應的音檔播出。這種打破傳統上只能用眼睛確認結果的方式，顯然是一種高效率也更直接的技術。

更有效率的方法是把所有基本的音節，像是中文裡的所有單字詞都錄製下來，需要輸出的時候，將這些單字詞前後連接起來經由喇叭播出。但是這種方式的缺點是字與字連接的地方聽起來會有不連貫的感覺，需要進一步將連接處進行平滑化處理。

最後一類是將所收集到的語音，轉換成具代表性的參數儲存起來，需要語音輸出的時候，將這些參數經過逆轉換，找出對應的語音訊號，並加上有關時間長度與聲調的資料修正後，再予以合成播放出來。這樣的方法應該是適用性最廣的，但是播出來的語音品質往往是最差的。所以，近來比較廣爲使用的合成技術，仍然是透過大量收集的語音資料，予以連接合成爲所需要的語音再播出。



語音合成的流程
就中文而言，在文字輸入之後，必須先將文字轉換為對應的注音符號，在此要參照現有中文字詞的發音規則考慮四聲的轉調，接下來透過對句調的分析，才能掌握語音輸出時，句子整體上音調的抑揚頓挫，讓輸出的語音顯得更自然，最後再經由語音合成器將語音合成並輸出。

語音合成是一種以人工方式產生語音，供各種場合使用的技術，它可以免除必須事先把語音錄好再播放的麻煩，同時也可以省下相當可觀的金錢與時間。語音合成的技術使得機器可以透過語音，傳達它對於使用者所下命令的回應訊息，拉近了人機的距離。

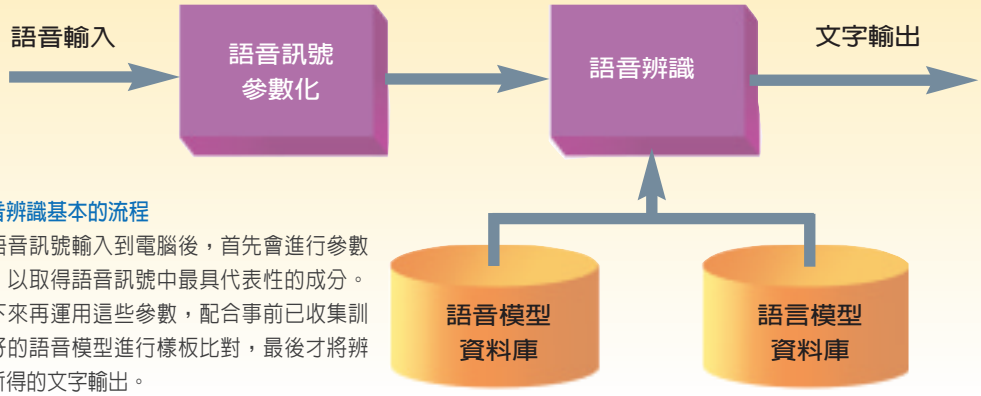
運用語音回應來傳達訊息的技術，在訊息傳播上有許多優點：任何人都可以輕易地了解訊息的內容，而無須經由訓練或過於高度專注的要求。訊息可以完全地被接受到，即使是當聽眾正在做其他的事情時，比如在走路、開車或是注視著某樣東西時。可以透過一般的電話網路，輕易地得知遠端的訊息內容。這種訊息傳遞方式，實質上是不需要紙張的，可以達到節省資源的效果。

不過，上述最後一項也意味著以聲音來傳遞的資訊內容，難以完整地複製下來。也就是說，合成語音並不適用於傳遞大量且內容複雜的資訊給一般大眾。但是語音合成在句子簡短、意義單純的對話內容上，倒是有相當程度的應用空間。

機器聽得懂人類語言的關鍵——語音辨識

語音辨識是讓電腦或是電子設備從我們所發出的語音中找出所含的文字意義，可視爲一種由語音轉換爲文字的過程。因爲這些文字意義是語音要傳遞的最基本資料，所以語音辨識技術的優劣，就取決於它是否能從接收到的語音訊號中擷取相對應的文字內容！

文獻中有關語音辨識的研究論文最早出現於一九五二年，內容是描述貝爾實驗室的一套語音數字辨識器。此後語音辨識方面的研究才逐漸熱絡起來，雖然近年來已建立許多語音辨識系統，但是它們的應用到目前爲止仍相當有限。即便如此，透過語音辨識與語音合成技術的整合，人類與機器進行對談的夢想將可以很快地實現。這樣的整合是相當有效率的，因爲我們可以透過語音對機器下達指令，機器的回覆可以透過語音合成直接聽到，如果有錯誤的話，我們也可以很方便



語音辨識基本的流程

在語音訊號輸入到電腦後，首先會進行參數化，以取得語音訊號中最具代表性的成分。接下來再運用這些參數，配合事前已收集訓練好的語音模型進行樣板比對，最後才將辨識所得的文字輸出。

地直接再次用語音更正。

與語音合成一樣，語音輸入也有幾項特點：語音輸入對大多數人類來說，是最簡單而又直接的方式，因爲它不須假借如打字鍵盤或是紙筆等其他工具的輔助。語音輸入的速度遠比打字或是寫字快，所以輸入會更有效率。即使是正在路上行走或是在做其他的事情，只要我們的嘴巴有空，就可以進行語音的輸入，這是使用其他方法無法辦到的。只要使用小小的麥克風或是電話即可輸入，省去了要借助其他裝置的麻煩。藉由電話線路的傳輸，更可以對遠端的機器下達指令。

然而，語音辨識並非沒有缺點。和語音合成一樣，所有輸入的或是進行中的對話都無法列印出來。另外，當我們在吵雜的環境中使用語音輸入時，辨識器需要先行去除雜訊，才能讓語音辨識的正確性維持在一定水準之上。

一般而言，語音辨識大致包括前半部的語音訊號參數化的程序與後半部的語音辨識兩大部分。前半部的參數化程序，主要的目的在於從原本的語音訊號中，擷取出最具代表性的語音特徵。就長時間來觀察，雖然語音訊號是變化多端的，然而將時間縮短到一定長度時，它則呈現出較爲穩定的狀態，擁有較固定的語音特性。所以，我們常以約32毫秒的時間長度爲單

位，稱爲一個語音音框，做爲擷取語音特徵的基本單位。在擷取出語音參數後，接著便進行語音辨識的步驟。

簡言之，語音辨識是一種樣本比對的程序，利用事先收集到的大量語音資料，經由同樣的參數擷取過程，統計出辨識所需的模型。進行辨識時，將所要辨識的語音參數序列與事先收集的語音模型進行比對，相似度最高者，便是這一句語音所對應的文字。

要求機器能善解人意——語意理解

假設一套語音辨識系統可以把一句語音翻譯出對應的文字，雖然有語音辨識演算法的協助，機器還是不知道我們說那句話的意思、想法是什麼？換句話說，機器並不知道那句話的「語意」是什麼。比如我六月十日早上八點要搭火車去台北，我們可以很清楚地知道說這句話的人，在六月十日那一天的早上八點要搭火車去台北。對我們來說，了解某個人說某些話的意思是輕而易舉的事，然而對電腦而言，事情並非如此。

簡單地說，語意的理解就是要知道人、事、時、地、物這五項的內容。只要能夠把這幾個項目填寫清楚，再搭配眼前的時間與空間資料，就可以初步理解一句話的意涵。就前述例子而言，「人」這一項可以填入「我」，指的是說話者；「事」就填入「搭火車去台北」；「時」就填入「六月十日早上八點」。將一句話拆解成五個要項進行分析，即可讓電腦具備簡單的語意理解能力。

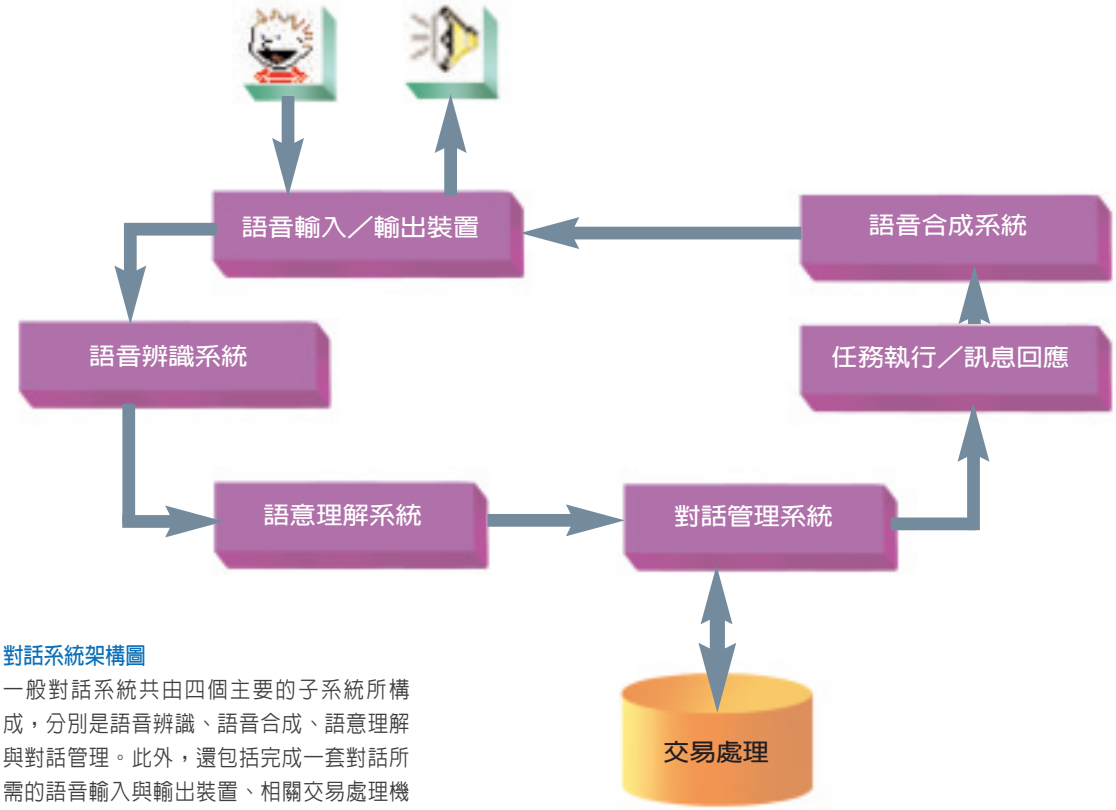
要讓電腦完全理解我們所說的每一句話的意思，迄今爲止仍是一件不容易做到的事。幸而，在一個對話系統中，由於系統是針對某種特定的應用領域而開發的，對於使用者語音的理解，在設計上可以將範圍縮小到一個固定的應用範圍內。對於超出系統應用範圍的文句，系統可採用無法理解的訊息來回應使用者，提示使用者改用其他語句重新輸入。

語言表達的功能——對話系統

顧名思義，這是一套人與機器之間進行對話的系統。透過它，我們可以跟電腦進行交談，使電腦理解我們說的話，或是命令電腦去做一些我們想要它去執行的工作。因此，這是一項非常重要的人機界面技術。

一般而言，對話系統包含語音辨識、語意理解、對話管理及語音合成四個子系統。首先，使用者輸入的語音經由語音辨識系統辨識出對應的文字，這些文字再送入語意理解系統分析出該句的語意之後，即由對話管理系統判斷系統應該執行的動作及適當的回應，回應的訊息經由語音合成系統做成語音輸出以回覆使用者，並等待使用者下一回合的輸入。

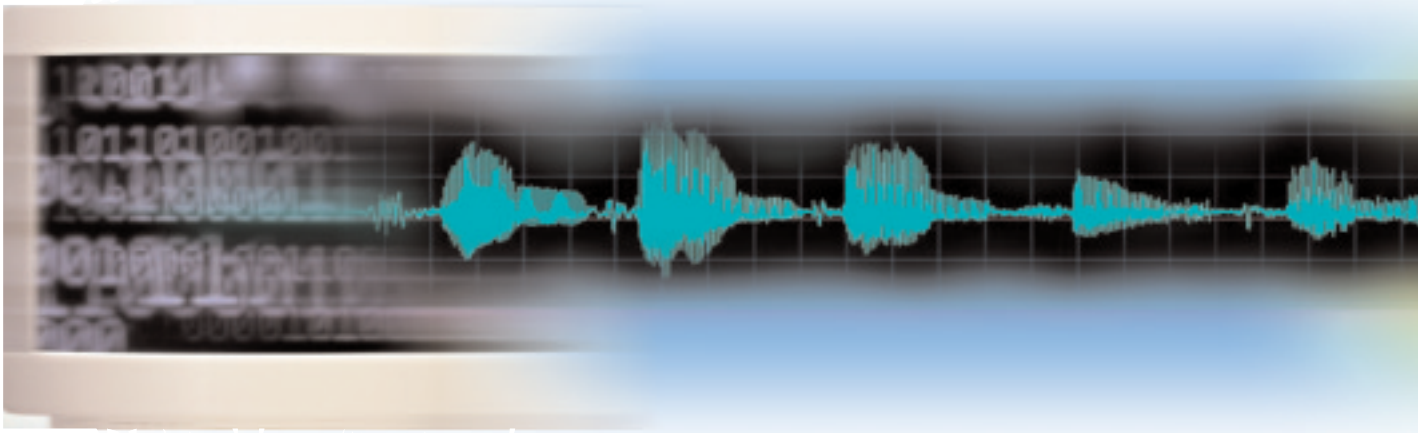
根據不同的系統對話主導方式，可以將對話系統概分爲兩大類。一是由使用者主導的對話方式，它是由使用者提出要求，讓電腦提供或協助查詢資訊，這種對話方式對一般使用者而言比較自然，但在系統實作上的困難度較高。



對話系統架構圖
一般對話系統共由四個主要的子系統所構成，分別是語音辨識、語音合成、語意理解與對話管理。此外，還包括完成一套對話所需的語音輸入與輸出裝置、相關交易處理機制及訊息回應與動作執行相關工作模組。

另一種則是由電腦主導的對話方式，這是目前最常見的語音對話系統架構。對話是由電腦以固定的開場白揭開序幕，提供某些固定的服務項目讓使用者選擇，於使用者回應且經電腦確認後，再由電腦回應使用者所要求的服務，或提出另一組子服務項目，由電腦提出選項，再由使用者進行回應。這是最常見的對話方式，雖然不甚人性化，但只要在架構系統時，將所有可能出現的對話模式加以分析，實作上較使用者主導的對話方式簡易許多，系統出錯的機會也較少。

由於這兩大類對話形態各具優缺點，於是有人將這兩類對話方式加以整合，提出另一種可以交互主導對話方式的系統架構，亦即除了電腦可以提出選項讓使用者選擇外，使用者也可以提出額外的選項讓電腦回應。舉例來說，使用者可以詢問系統還有什麼其他的服務項目？這個時候，對話系統就應該針對使用者的詢問，提出其他的服務項目。



當這類由使用者所提出的問題或選項出現時，由於在系統預定的對話流程中可能沒有涵蓋到，此時對話系統中的語意理解系統及對話管理系統，便要發揮應有的語意理解及適當的對話回應能力。因爲要考慮到設計時沒有收集到的對話內容，系統設計上的難度更高，但是對於人機界面的親和度會有較好的評價。

目前常見的對話系統主要應用在資料查詢上。航空運輸資訊服務系統，是由美國國防部高等研究計畫局資助的研究計畫所開發的。它主要是透過電話使用語音查詢飛機航班時間等相關資訊，及進行旅行的時程安排，是最常用來做爲對話系統效能評估及系統展示的一套系統。另一套系統是丘比特氣象查詢系統，由麻省理工學院發展出來的，它主要也是透過電話使用語音查詢與氣象相關的資訊，像是世界各大城市某一天或某幾天的氣象報告及回答關於天氣現象的一些問題，如氣溫、風速、降雨量、氣壓、溼度、日出時間等。雖然主要提供的氣象資訊是以英文爲主，但是它也可以具備轉換語言的能力，也就是將這些氣象資訊翻譯成其他不同語言輸出。

國內在對話系統的應用方面，也都是著重在查詢系統上，研究單位則包括台大、成大、工研院等。所開發出的相關系統有工研院前瞻技術研究中心所開發的伊喂得中文電話語音對話系統，它的應用範圍廣泛，主要有旅遊資訊查詢、網際網路線上資訊查詢、電話語音查詢、個人語音秘書、金融資訊查詢、電子銀行、語音電子商務、醫院掛號等。其他包括電話總機系統，是透過電話線路查詢某個人名的分機號碼；醫院掛號系統，是透過電話提供線上預約掛號的功能，查詢各科醫師的門診時間及目前已預約的門診人數；全球衛星定位行車資訊查詢系統，提供行車時所需的路況、餐飲、風景名勝、街道地圖，甚至當日股票交易的相關資料。

其中，車用對話系統是相當重要的一個研究項目。因爲在開車時，我們的眼睛、手腳忙於控制汽車，無法分心到其他動作上。如果可以使用語音來查詢資料，將是相當安全且便利的方法。

未來展望

經由上面的介紹，我們可以了解語音處理在人機界面所占的重要性，在國內外研究學者們不斷的努力下，人類與機器直接用口語溝通的夢想，得以在各個不同的應用環境中實現。然而，目前的一些應用，由於語音處理技術未臻成熟，尚有改善的空間。現有的系統也只能限定在某些特定的應用環境中，如前面提到的語音訂票系統、運輸工具航班查詢系統及節慶時間查詢系統等，以及一些少量且固定詞彙內容的對話系統上。我們相信只要持續地進行這方面的研究，投入更多的研究人力，將目前既有的研究成果加以改進，一定可以將人機界面建構得更好，在操作上變得更自然，人機間的互動可以更流暢，使得機器給人的印象也可以更人性化。 □

簡仁宗
成功大學資訊工程學系