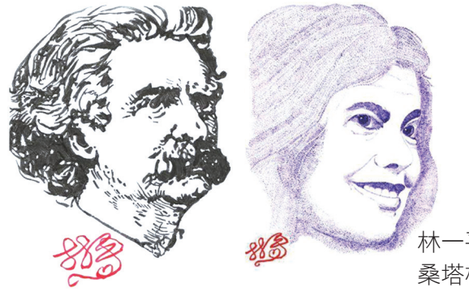
語音 AI 仿真的關鍵：停頓

文／林一平 講座教授



林一平手繪之馬克吐溫 (左) 與桑塔格 (右)。林一平提供。

最近開始流行基於語音的多媒體物聯網 (IoMT)，被大量用於語音到文本的翻譯和語音控制應用。對於此類應用，核心技術是自然語言處理。我的研究團隊發展一套語音談話的 IoT 應用開發平台，稱為 VoiceTalk，詳細闡述了基於語音的 IoMT 開發問題。我們提出了一種新的自然語言處理機制，進行自動語音辨識，藉此發展了不少有趣的互動應用。

利用語音來進行電器控制較為簡單，例如燈光控制，或冷氣控制，只要轉譯為指令即可。其商業化的產品也都極為成熟，例如 Google、亞馬遜 (Amazon) 及小米都有語音控制的產品。

而本文翻譯 (voice to text transcription) 這項科技的發展，其難度則遠高於語音控制，若無人文素養的加持，終將流於膚淺。個人淺見，最難之處之一，在於處理語句之間的停頓 (pause)。寫文章時，句子內部主語與謂語之間如需停頓、分開的地方，就用像一隻小蝌蚪的逗號來標明。因此在進行語音辨識，轉化為文字時，聲音的停頓處，就被翻譯成逗號。然而如何找出「停頓」轉化為逗號，頗有學問。

「停頓」的運用之妙，存乎一心。厲害的作家及演說家，都各自有妙招，呈現他們不同的體會。馬克吐溫 (Mark Twain) 這麼說：「正確的用詞可能很有效果，但沒有一個用詞如同在正確的時刻暫停那樣有效。」蘇珊．桑塔格 (Susan Sontag) 則承認：「無可避免的，沉默仍然是對話中的一種語言形式和元素。」尤其，沉默也是一種回答，可微妙的代表不同意義，例如默認。

談說中在何時停頓，意思可能完全不同。換言之，在一串文字中放逗號於不同位置，意思會有很大差距。二次世界大戰時的汪精衛政權，有一位女作家名叫蘇青。蘇青的成名作，僅僅將逗點移動一個位置。《禮記．禮運》寫著：「飲食男女，人之大欲存焉。」這位女作家將之改寫為「飲食男，女人之大欲存焉。」當時民風保守。她的創作大膽前衛，自我物化，一夕成名。遇到這種語帶雙關的讀法，停頓的判讀變得很重要，否則轉譯成文字時，差之毫釐，失之千里，就貽笑大方了。

詩人朗誦時，我們的 VoiceTalk 若進入「詞」的模式，會將朗誦的詩下標點成為一闋詞。例如千家詩中的七絕詩《清明即景》:「清明時節雨紛紛，路上行人欲斷魂。借問酒家何處有，牧童遙指杏花村。」經過人工智慧，將標點符號挪移一番，就變成一闋詞:「清明時節雨，紛紛路上行人；欲斷魂！借問酒家何處？有牧童遙指杏花村。」我們正在思索如何利用 VoiceTalk 改變莎士比亞作品中的「停頓」，將莎翁的雙關語化為「三」關語。

林一平
國立陽明交通大學資工系終身講座教授暨華邦電子講座

現為國立陽明交通大學資工系終身講座教授暨華邦電子講座，曾任科技部次長，為 ACM Fellow、IEEE Fellow、AAAS Fellow 及 IET Fellow。研究興趣為物聯網、行動計算及系統模擬，發展出一套物聯網系統 IoTtalk，廣泛應用於智慧農業、智慧教育、智慧校園等領域 / 場域。興趣多元，喜好藝術、繪畫、寫作，遨遊於科技與人文間自得其樂，著有 < 閃文集 >、< 大橋驟雨 >。

The Key to Voice AI Simulation: Pause

The emergence of Voice-Activated multimedia Internet of Things (IoMT) has been used extensively in voice to text transcription and voice control application. Natural language processing is the core technology for such applications. My research team has developed a voice over IoT platform, VoiceTalk, and elaborated on the developmental issues on the IoMT. Furthermore, we have proposed an innovative natural language processing mechanism for automatic speech recognition. Many interesting interactive applications were developed with this mechanism.

It is quite straightforward to use voice control on electrical appliances, such as lights or air conditioners, as long as the voice commands are translated into instructions. This has been widely developed into commercial products; for instance, leading vendors such as Google, Amazon and Xiaomi have launched their voice-enabled products.

The technological advancement of voice to text transcription is much more difficult than simple voice control. Without a deep understanding of the humanities and literacy, any attempts at advancement would merely be superficial and cosmetic. In my opinion, dealing with pauses between sentences is one of the most difficult jobs. When writing an essay, we may place a comma, a little tadpole-like critter, in a pause or separation between the subject and the predicate in the sentence. Therefore, when using speech recognition to convert speech into text, the pause of the sound is translated into a comma. How to gracefully detect "pause" and turn it into a comma is quite subtle.

Ingenuity in using “pause” arises from practical intelligence. Most influential speakers and writers have tactics to illustrate their understanding. Mark Twain put it this way, "The right word may be effective, but no word was ever as effective as a pause with right timing." Susan Sontag wrote that "silence remains, inescapably, a form of speech." In particular, silence subtly forms answers with different meanings, such as acquiescence.

Different timing of pauses in speech may deliver different messages. In other words, misuse of commas can cause confusion or even misunderstanding of the information. Su Qing was a female writer during the Wang Jingwei regime in World War II. Her most acclaimed work was merely moving a comma one Chinese character further in a sentence from the Book of Rites (Liyun). The Book of Rites (Liyun) states “the things which men greatly desire are comprehended in meat and drink and sexual pleasure.” And Su Qing rewrote the sentence as "the things which women greatly desire are comprehended in meat and drink and sexual pleasure with men.” Social mores were conservative at the time. Therefore, her radical creativity in pursuit of self-objectification attracted people’s attention so that she became an overnight sensation. While confronting such ambiguities in speech communication, it can’t be overemphasized to interpret pauses carefully. Otherwise, a tiny lapse can lead to a huge mistake and much embarrassment.

Our VoiceTalk software has a “ci” mode, in which a poet reciting their work can have their poem converted into iambic verse. For example, the well-known Chinese poem “Tomb Sweeping Day” in the Thousand Poems, is translated thus: "Tomb Sweeping Day sees drizzles running and flying, and hearts lost in gloom, mourners on paths crying. ‘Any tavern near and far?’ I ask a boy, who points to Almond Bloom Village beyond eyeing." After the process of artificial intelligence, the punctuation marks are relocated to turn the poem into an iambic verse, "It rains on Tomb Sweeping Day, Travelers along the road; Looking gloomy and miserable! I ask ‘any tavern near and far?’ A boy points to Almond Bloom Vill beyond eyeing." Currently, we are thinking about how to use VoiceTalk to alter the placement of "pauses" in Shakespeare's works to add an additional layer of nuance to the double-meaning of Shakespearean puns.

Dr. Jason Yi-Bing Lin

Lifetime Chair Professor of the Department of Computer Science at National Yang Ming Chiao Tung University and Winbond Chair Professor

Dr. Lin is currently a lifetime chair professor of the Department of Computer Science at National Yang Ming Chiao Tung University and Winbond chair professor. He is an ACM Fellow, IEEE Fellow, AAAS Fellow and IET Fellow. His research interests include Internet of Things, mobile computing, and system simulation. He has developed an Internet of Things system called IoTtalk, which is widely used in smart agriculture, smart education, smart campus, and other fields. He has a variety of interests, such as art, painting, and writing, as well as voyaging through science, technology, and humanities.