AI 加速器到專用處理器，語音應用更直觀

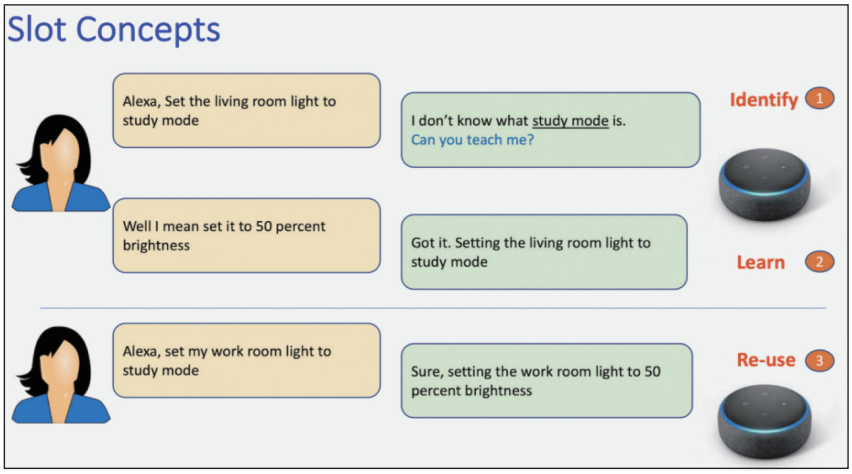
■文：任芷萍

語音辨識正在成為消費者語音助理的重要元素。亞馬遜(Amazon)在2017年為Alexa用戶創建語音配置文件，此後一直在增強該功能；去年夏天，語音助理開始深入用戶聯繫資訊來個性化與Alexa的交互。與此同時，Google也徹底修改Google Assistant的語音匹配功能的設置程序，增加步驟以提高安全性，並在個性化響應時使語音助理更加靈活。去年9月，亞馬遜再推「可教學AI」(Teachable AI)功能，若遇到Alexa不理解的語音命令，用戶可直接透過語音向Alexa下達指示、即時傳達定義，無需手動設置偏好或改寫Alexa邏輯規則。

語音助理變聰明、且更個性化！

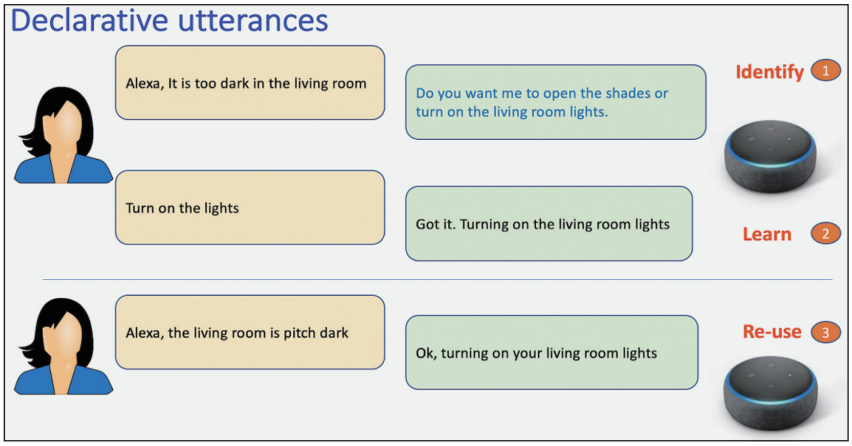
第一階段將專門用於照明和恆溫器等智能設備，但最終將包括其他類型的命令，其工作原理類似為Alexa Routines設置關鍵字，而非死記硬背的觸發器。Alexa首席科學家Rohit Prasad對此做了演示：「Rohit的閱讀模式」，一開始盲然不解的Alexa會主動詢問，在得知定義後將燈光亮度降低至40%。借助交互式教學，Alexa會立即學習這些定義和相關操作，並僅將它們儲存供特定用戶使用。對話管理(dialogue management)模型會檢查用戶的問題及答案是否

圖1：Alexa能與用戶即時互動式教學，將這些概念推廣到新的上、下文中，並將其與使用者帳戶連結



資料來源：https://www.amazon.science/blog/new-alexa-features-interactive-teaching-by-customers

圖2：交互式教學可讓Alexa學習「間接表述」



資料來源：https://www.amazon.science/blog/new-alexa-features-interactive-teaching-by-customers

在已知範圍內，例如，Alexa 會詢問用戶口中的「學習模式」是什麼意思？

若用戶回答：設置為良好的閱讀亮度水平。模型因無法理解、在每次嘗試定義失敗後，對話管理器會降低後續問題的複雜性。若概念提取模型在幾經詢問後仍無法獲知「學習模式」定義，對話管理器可能會直接追問：能為我提供亮度或顏色數值嗎？最後，「陳述推論」(declarative-reasoning) 模型會預測與用戶陳述話語對應的動作，還可在決定儲存所選動作以備將來重用之前，就其上、下文驗證所選動作的語義是否適當；成功後，先前學習的概念可沿用至相關文本，例如，客廳「學習模式」意味將燈光設置為 50% 亮度時，辦公室也將採用相同概念。

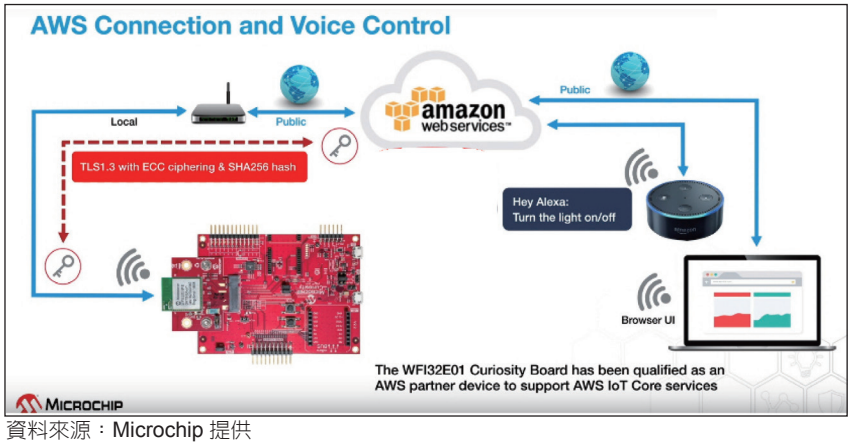
MCU 供應商群起支援 Amazon Voice Service

Alexa 甚至可被授予人類如何將陳述式語句視為變相命令，例如，告訴 Alexa 房間太暗了，它會詢問用戶是否要打開燈或窗簾、然後依定義動作。除了自動涵蓋所教概念外，可教學 AI 還允許用戶明確指示 Alexa 忘記最近或所有學到的概念。隨著 Alexa 語音服務 (Amazon Voice Service, AVS) 越來越強大，不少微控制器 (MCU) 供應商群起響應。意法半導體 (ST) STM32 系列 MCU 皆已整合 Alexa 語音使用者介面軟體，日前再推亞馬遜認證的智慧連網裝置參考設計套件，開發者可利用 AWS IoT Core 平台 AVS 功能，在簡易 MCU 研發內建 Alexa 之產品。

微芯科技 (Microchip) 推出首款為雲端身份驗證預先配置和設置的 Trust&Go Wi-Fi 32 位元 MCU 模組——WFI32E01PC，符合 Wi-Fi 聯盟 (WFA) 規範，並獲得美國聯邦傳播委員會 (FCC)、加拿大工業部 (IC) 和歐洲無線電設備指令 (RED) 三大世界級監管機構的全面認證，同時與 PIC32MZW1 Curiosity 開發板相容 (已通過 AWS IoT Core 平台認證並被列入 AWS 合作夥伴設備目錄)，可使用 AVS 與板載感測器互動。AWS IoT Core 認證平台包括程式碼範例、WLAN 軟體以及可在 MPLAB Harmony v3 找到的網路通訊協定堆疊。

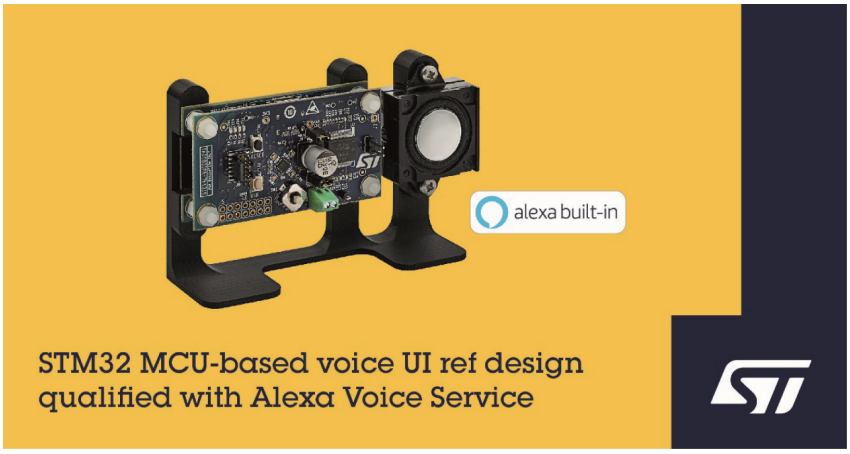
STM32 MCU 使用者可自訂和擴充系統設計、增加強化功能，例如：第二個喚醒關鍵字、附加的當地語系化命令、聲控圖形顯示。為進一步簡化原型設計和產品研發，參考設計硬體包括一個作為獨立模組的音訊子板，內含一個 ST FDA903D 音訊轉碼器、使用者 LED 和按鈕，以及兩個間隔 36mm 的 MP23DB01HP MEMS

圖 3：集成 PIC32MZ-W1 Wi-Fi SoC 及可選預配置 Trust & GO 安全元件的 WFI32E01 Wi-Fi MCU 模組，可與 Amazon Alexa 連接做語音控制



資料來源：Microchip 提供

圖 4：ST 推出亞馬遜認證的參考設計，簡化 Alexa 內建智慧家庭裝置開發



資料來源：ST 提供

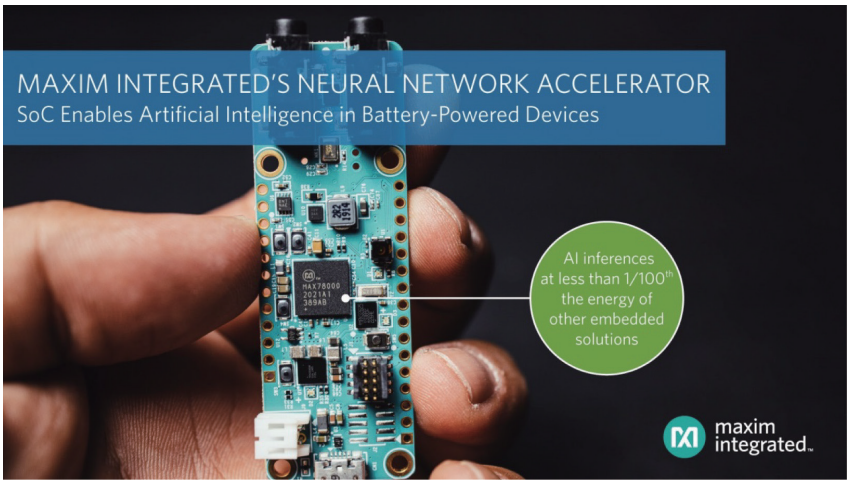
麥克風，適用於尺寸受限的產品，包括小至電源開關插頭。若需專用麥克風間距、聲學特性和使用者介面定義，模組化硬體還允許使用者自訂子板。即使環境吵雜，麥克風間隔小，音訊前端仍能提供出色的遠場語音偵測功能。

電池供電的邊緣設備，也能制訂複雜決策

顯然，語音助理越來越聰明，所肩負的任務越見繁複；於是，開啓了 AI 加速器、乃至專用處理器的採用風潮。美信 (Maxim) 去年底推出帶有神經網路加速器的 MAX78000 低功耗 MCU，支援電池供電的嵌入式物聯網 (IoT) 設備在邊緣透過快速、低功耗 AI 推論來制訂複雜決策；相較於軟體方案，採用 AI 技術的電池供電系統可大幅延長執行時間，且其成本僅是 FPGA 或 GPU 方案的零頭。MAX78000 核心是專用硬體，旨在最大程度降低卷積神經網路 (CNN) 的能耗和延遲，且運行時幾乎不需任何 MCU 介入，意味著操作的流暢度極高。

Maxim 表示，該硬體能量和時間僅用於實施 CNN 的數學運算，執行推論功耗不到 MCU 軟體運行功耗的 1%；若需將外部世界的採集資料高效輸入到 CNN 引擎，可整合 ARM Cortex-M4 或功耗更低的 RISC-V 內核。Syntiant 推出第二代 NDP120 神經決策處理器 (NDP)，亦強調電池供電設備的音訊和感測器應用；內嵌 Syntiant Core 2 靈活的低功耗深度神經網路推論引擎，以不到 1mW 的功耗同時運行多個應用程式，包括：迴聲消除、波束成形、噪聲抑制、語音增強、發言者辨識、關鍵字識別、多個喚醒詞、事件檢測和本地命令識別。

圖 5：Maxim 神經網路加速器，在電池供電設備中實現複雜的嵌入式決策



資料來源：Maxim 提供

NDP120 具有高度可配置的音訊前端，適用於遠距離語音濾波和迴聲消除，並支援紅外檢測、多軸加速度、傾斜、磁場和壓力等多模式感測器融合；Syntiant Core 2 是張量處理核心，每一層獨立控制參數、輸入和輸出張量，不影響編程簡便性。NDP120 的可編程數位訊號處理器 (DSP) 與高精度推論引擎結合，非常適合創建可在傳統演算法和機器學習 (ML) 之間運行的高性能語音命令應用程式。耐能智慧 (Kneron) AI SoC——KL720 NPU 版本，則強調能識別「整個詞典中的單詞」，不限特定喚醒單詞；另可重新配置的設計，KL720 NPU 可同時處理圖像和音訊。

加速器之後，專用晶片崛起

此外，亞馬遜第四代球形設計 Echo 與第三代 Echo Show 10 皆採用 AZ1 Neural Edge 處理器，專供邊緣設備的機器學習之用，可提供更自然的語音互動體驗並縮短數百毫秒的回應時間。AZ1 Neural Edge 處理器是聯發科技 (MediaTek) 專為「高階音訊處理和語音助理應用」之 MT8512 處理器的要員，可用於了解聲音方向來源，並決定何處、何時、何種速度調整攝影機；集成 2GHz 雙核 CPU，支援各種音訊處理外設及藍牙 5.0 ／ Wi-Fi 5 雙頻連接；配合高性能語音 DSP 使用，可快速、準確地檢測語音命令中的喚醒詞和關鍵詞，提升 Alexa 靈敏度。

當前的功能包括使用回饋搜索演算法來獲取用戶回饋並使用交互來糾正操作中的錯誤，並透過語音直接教導 Alexa，不必借道應用程式 (APP) 或網頁設置。順帶

圖 6：聯發科技 MT8512 專為高階音訊處理和語音助手應用而設計，內部集成亞馬遜 AZ1 Neural Edge 處理器，可在邊緣設備運行強大的推論引擎

資料來源：https://www.mediatek.com/blog/amazon-az1-neural-edge-processor-powered-by-mediatek

一提，帶顯示器的 Echo Show 10 使用具有電腦視覺 (CV) 的聲源定位 (SSL) 來識別視野中的物體和人，並辨識聲音發自何人？新一代 Echo Show 10 顯示器和攝影機可改變方向並對準空間中的揚聲器，在視訊通話實現更自然的交互，可一面走動、一面聊天或觀看視訊 (須事先儲存個人臉部和聲音特徵)。與此同時，雲端服務供應商 (CSP) 正往專用處理器晶片靠攏，以提升 AI 訓練、推論效能。

亞馬遜已宣佈未來將採用旗下 AWS Inferentia 處理器承擔 Alexa 大部分雲端處理，以加速文本到語音翻譯的大量機器學習。雖然市佔較小，但 Google Assistant 在理解自然語言和命令似更勝一籌，且 Google Nest Hub 智能顯示可控制和可視化管理整個智能家居，並識別家中所有成員、提供個性化資訊，其專用 TPU 貢獻不小。三星亦採用 Google Cloud 第三代 TPU 訓練自家 Bixby 語音助理 (在全球逾 1.6 億台設備上運行)，使用自動語音識別引擎將用戶語音命令轉換為文本，以減少 AI 訓練時間、縮小模型、降低單字錯誤率並提高運行速度。

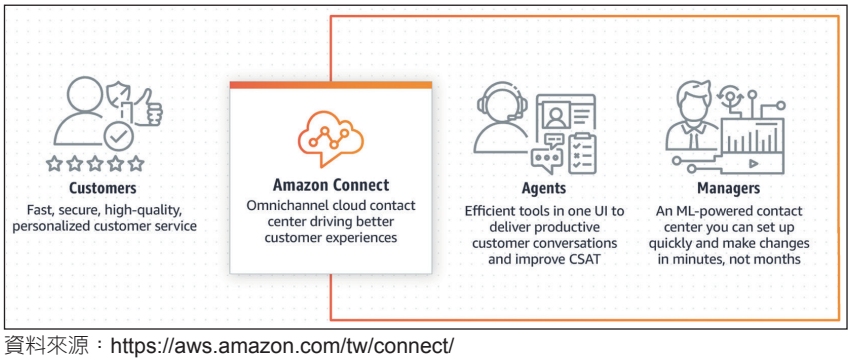
「以用戶爲中心」，語音辨識加速客服流程

根據 2020 年 eMarketer 的一項研究，美國有 38.5% 的人口使用語音助理連接智慧手機或其他小工具，且去年因為居家時間增多，成長率達兩位數。此一趨勢促使主要電商門戶網站開發聊天機器人或使用現有基於語音的集成來增加銷售；肺炎疫情爆發以來，制訂對話商務策略以彌補人際交流，已成商業新手段。亞馬遜宣佈將語音辨識技術集成到「虛擬聯絡中心」(客服中心) 平台，旗下 AWS 將使用 AI 來分析員工或客戶的聲音並悄悄確認其身份，作為與客戶交流的工具，並收集和分析有關這些對話的數據。

通常，企業依靠詢問生日、社會安全／身份證號碼或地址之類的辨識性問題來確認來電者身份；而上述語音 ID 目的是跳過通常很繁瑣的過程，並使得竊取呼叫中的身份更加困難。一旦使用者同意使用語音 ID，該軟體會使用幾秒鐘的通話來分解其語音生物特徵，以及音調、節奏之類的標籤元素，然後予以儲存並標記為個人的語音文件以備將來參考。當下次同一人來電並自報姓名，語音 ID 可提取聲紋並將其與當前語音做比對。若匹配無疑，會將呼叫轉移給某一客服人員，不必再確認身份；反之，則將經由標準篩選系統檢查。

一款專用的 SaaS (軟體即服務) 應用程式 Voice

圖 7：Amazon Connect 是易於使用的全通路雲端聯絡中心，採用全通路設計，為顧客和客服人員提供跨語音和聊天的無縫體驗



資料來源：https://aws.amazon.com/tw/connect/

Compass Journeys，利用 NLX Conversational AI 平台讓使用者無需與人交談、就能經由網頁、簡訊、電子郵件、手機、聊天軟體等，用「語音」下達指令。好處是：不必等待接線或複雜選單，且用戶可語音指導自訂控制呼叫速度，創建「以用戶為中心」的自助服務選項，亦有助服務商將 IT 服務台自動化、創建「旅程樣板庫」以滿足特定客戶需求。利用免費與 Voice Compass 服務集成的軟體開發套件 (SDK)，還可與網站、顧客關係管理 (CRM) 或第三方應用程式及 IoT 硬體設備整合。

圖 8：YoursTruly 開發目的是「利用原生音頻廣告體驗的巨大且快速增長的潛力」

資料來源：https://www.adtonos.com/news/its-yourstruly-offering-was-used-in-a-campaign-for-audi-in-london/

「交互式語音商務」新時代揭幕，風險隨之而至

使用加密聲波、可在任何設備離線非接觸支付的 ToneTag 公司，將音訊導入支付閘道器、推出「語音商務」；基於語音的支付解決方案利用聲波生成音頻 QR，作為交易媒介及支付資訊，支援 Amazon Pay、UPI 和信用卡等主要付款工具，使客戶能在離線商務獲得交互式、自定義和無縫體驗。用戶只需與他們的行動設備通話即可在咖啡館和快餐店訂購並預付帳款，到店後無需排隊就能取貨。這種「隨時隨地」的個性化離線零售，儼然是「交互式語音商務」新時代標誌。廣告商與內容製作商也正嘗試創建全新的交互式內容體驗。

英國一家線上廣告平台 AdTonos 擁有一項名為「YoursTruly」的技術，目的是利用「原生音頻」廣告體驗的巨大且快速增長的潛力。奧迪 (Audi) 交互式音頻廣告首次在商業廣播透過智能揚聲器發佈，以預訂奧迪汽車的試駕車或尋找經銷商。該廣告在廣告時段插入一個互動觸發器，聽眾可以語音命令對奧迪廣告做出反應，使語音助理參與並完成所需操作，再返回直播電台。英國廣播公司 (BBC) 一個名為「檢查室」的試點專案，是一個在亞馬遜 Alexa 運行的交互式科幻故事講述，讓用戶有效扮演其中一個角色，對某些受眾 (尤其是兒童) 頗具吸引力。

然而，如何針對自然語言搜索優化？如何創建針對答案引擎優化的內容？是兩大挑戰。AI 語音助理正在重塑消費者和企業與數位技術的交互方式；雖然，語音互動應用存在無限可能，也潛藏風險。一項甫在電腦協會嵌入式網路感測器系統會議 (SenSys 2020) 發表的研究揭露：即使沒有麥克風，家中掃地機器人等智能家電也可能被駭客竊聽家庭對話！透過遠程訪問光達 (LiDAR) 讀數對 Roborock 進行遠程竊聽攻擊。

光達可經由獲取屋中垃圾桶等特定物體的反射來捕獲聲音訊號，而這個反射物又會因為附近的聲源 (例如人們對話) 而振動。駭客可能會重新利用真空吸塵器的光達感測器來感知環境中的聲音訊號，從雲端遠程採集光達數據並使用深度學習處理原始訊號以提取音頻資訊。此一弱點恐會揭示電話會議的機密、信用卡訊息，或由正在播放的電視節目推估用戶偏好。魔鬼藏在細節裡，語音互動是最直觀的觸媒，伴隨而來的操作細膩度與資安風險防護卻不可不慎！ CTA