台灣聽力語言學會電子學報
The Speech-Language-Hearing Association， Taiwan

- 主題文章：助聽器的頻道數多寡，在未來不應是決定助聽器性能好壞的主因
- 撰 稿 者：陳柏儒

主編的話

未來的輔具會呈現什麼樣貌呢，科技與數位演算不斷推進下，本期的電子報在理事長的媒合下，編輯群首次嘗試突破專業藩籬，邀請{洞見未來}科技的陳執行長柏儒來撰稿。

由於角度及專業領域的不同，在撰寫和後續校閱的過程中，因編輯群與筆者有多次的討論，也間接導致此次出刊的一再延宕，再次向讀者們致上萬分歉意。

為降低稿件潤飾後造成的語意更動的可能，本期特於文章下方，保留筆者與團隊間有想法差異的 Q&A 內容，供讀者參閱；讓讀者藉此機會從數位工程的角度來了解助聽器這門科學。

當所學的愈多，便會發現自己在醫學前顯得愈為渺小；希冀未來某天這些思想火花，能成為熊熊烈焰，照亮我們尚在漆黑中摸索的領域。

主編攜編輯群　敬上

主題文章

助聽器的頻道數多寡，

在未來不應是決定助聽器性能好壞的主因

陳柏儒

洞見未來科技執行長

30 年助聽器使用者，因自身的需求對聽力領域的技術熱愛，過去曾待在知名的半導體公司設計晶片。現為 RelaJet 洞見未來科技創辦人，全力發展聽力相關產品。洞見未來科技是全球少數幾個助聽器晶片提供方，同時為 ARM 以及美國高通的聲音技術合作夥伴，2021 年發展自有品牌 Otoadd，期待更廣泛地解決人們聽力問題 。

前期調查

調查過去 5 年間的助聽器購買者，許多人既定印象是門市人員總會告知助聽器頻道數越多價格越貴。而門市人員說法多種，我們整理出以下兩點經常是使用者聽到的推銷方式：

1. 頻道數越多，聲音細節越好，人聲更易聽清楚。
2. 學習英文需要更廣的頻域，因此頻道數越多，音域更廣。

其實，以上兩種說明方式只有對一半，讓我們以聲學的角度來做個說明。面對使用者做聽力測驗的時候，坊間的測驗器材會透過 250Hz、500Hz、1000Hz、2000Hz、4000Hz、8000Hz 這幾個頻段去進行分析使用者的聽力狀況。大多數的使用者是高頻聽損，為了讓助聽器可播放到 8000Hz 頻率的聲音，最少的要求必須以每秒 16000 次的採樣率（Sample rate） ，採樣率越高代表可聽見的最高頻率為（採樣率/2）的頻率，理論上人耳聽見頻率為 20-20000Hz 之間因此也不需要過高的採樣率即能涵蓋可聽頻率；而採樣率越高，也代表著每秒採樣的次數越多越密，採樣到的波形越完整 ；在相同時間內採樣數越高，帶給演算法的「可分頻」的空間越多（也就是坊間稱的頻道數）。[註 1]

了解聲學的因果關係後，回答上述銷售的說明方式一及方式二的時候，確實頻道數越多，其背後代表採樣率可能會越高，採樣的波形相對越完整聲音細節越好，但是人聲或者是英文是否能因此聽得「更」清楚？這點並非直接關係，因為在聲學領域中喇叭單體的限制（普遍 2000 Hz 以後頻響曲線下降）因此演算法處理人聲頻道的範圍大約會著重在能量較強的 200Hz-1000Hz 之間，就算是要提升英文氣音理解音

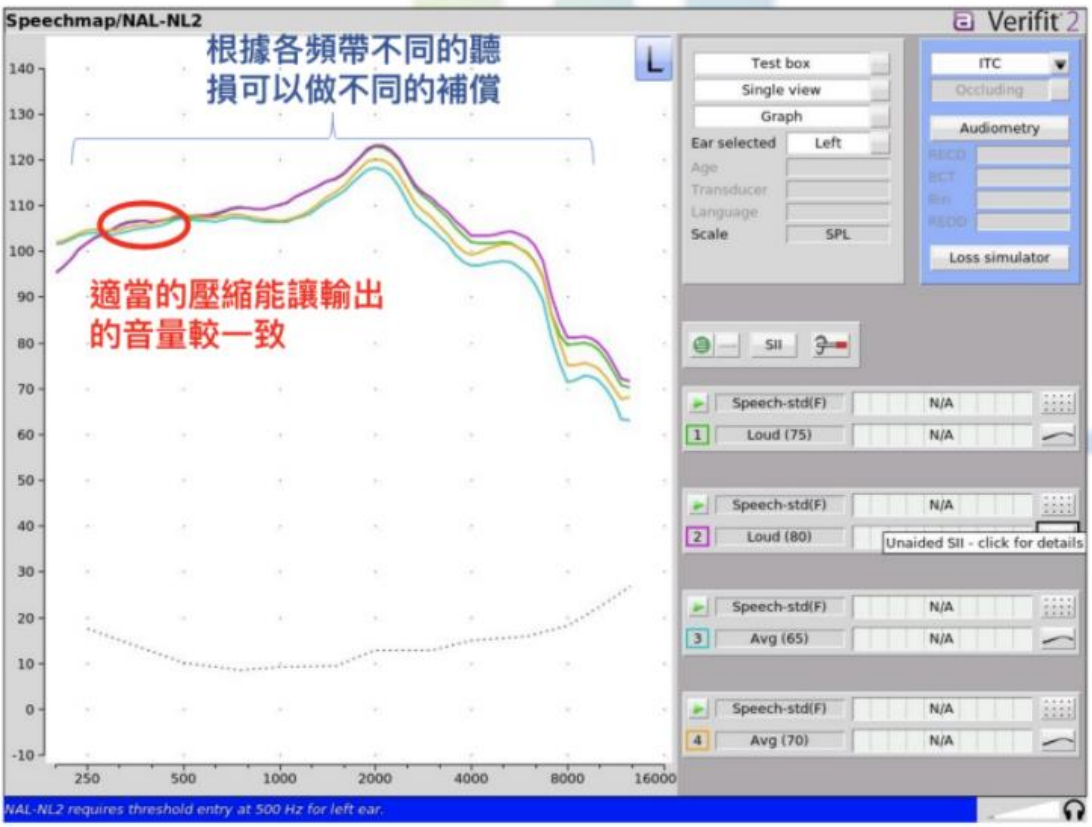
域，例如 [SH] 音大多強化 2000 Hz 的部分較多，至於 2000Hz 以上可以依照使用者聽損曲線是否超出單體的限制做微調（例如 [S] 音常發生在 4000-6000 Hz 可以依照單體限制微調）。整體來說這樣的頻率所對應不需要很高的採樣率就能夠聽到。而日常生活對話中，真正會造成干擾或聽不清楚主因是生活中有太多噪音跟人聲頻率近似，因此各家助聽器廠衍生出各種解決方式。

對於人聲處理的演進與未來，就是本篇欲詳細介紹之所在。

註 1 實務經驗來看坊間強調「多頻道」的助聽器演算法做頻率基底轉換幾乎是以 FFT 為主（演算法之中也有一些用 filter 作為分頻方法或者是直接波形圖做 DCT 之類做分頻，但這類方式在多頻道的時候需要龐大運算資源，用此兩個方法在有限資源內達到高頻道數幾乎是不可能的事情）若採樣數太低，加上助聽器限制在極短延遲的時間內，使用 FFT 轉換太多個頻段再 IFFT 回來精度會掉超多，因此沒有廠商會在低採樣率的時候開太多頻道。因此在實務經驗裡，頻道數越多採樣數越高是強相關的。

助聽器若只靠高頻道數或頻率演算已經是過時作法

面對使用者的反饋，常會聽到聽障者普遍認為在人聲鼎沸（例如酒吧，星巴克）等等的場域會聽不清楚。為了這個問題，我們內部實驗室做了各種比對找原因。我們來看一下目前主流的助聽器頻率響應曲線：



圖一

上圖曲線的部分，可以看出 g65/g70/g75/g80 這三條線相當緊密， 這樣曲線的調音邏輯若以白話文來解釋：把 g65/70 跟 75/80 更緊密是為了要讓小聲的聲音放大，讓使用者聽起來較小聲的聲音會更像在耳邊的聽感，在耳邊的聽感相對的語音理解能力也較好。

這樣的設計在安靜環境或者是環境噪音不那麼吵的情況下很有用。但是在吵雜環境的時，就會是一個巨大災難，這樣的調音邏輯會讓小聲的噪音放很大。另外，若人聲跟噪音剛好在近似的頻帶或者是能量一致時也皆會一起放大，進而影響人對語音的理解。

我自己身為一個使用者，有一個形容最貼切的：很明顯感受到在這樣環境下的聽感是一種非常「平面」的聽感。不論靠目標說話者多近，噪音因為被放更大，大到跟說話者音量一樣的時候聽起來就跟說話者的聲音是混在一起的導致聽感混淆，無法區分出聲音的立體感，造成聽不清楚。

當前各家執著在頻道數或頻率數也是無可厚非的事情，因為數位助聽器晶片平台因要兼顧短延遲以及省電的需求。剛好對頻率上的處理並非需要龐大的計算能力，若當晶片平台所能提供的算力很有限，各家技術也只能受限在目前算力瓶頸。

以聲學辨識角度來看，若只有靠頻率去做分析，面臨到日常生活中很多聲音的頻率/能量相當接近，很難抓出何者為對使用者有意義的聲音並適當放大，幫助使用者有更好的聽覺體驗。剛好這幾年深度學習在影像或者是聲音辨識已經相當成熟，但是龐大的深度學習的模型，難以運行在坊間助聽器使用的低運算力平台，若要追求更好的體驗，這方面的發展可想見在未來助聽器市場中會是兵家必爭之地。

將聲音圖像化進行 AI 辨識，將是未來助聽器的主流作法

針對聽障者聽不清楚的人聲鼎沸場域，使用頻率的做法無法有效去區分出各人的聲音，因為大多數的時候每個人講話頻率跟能量是相當接近的。過去這個問題，我們團隊洞見未來科技試驗過各種深度學習的方法。其中，把聲音的時間跟空間的關係變成一幅畫，進行影像辨識的方式找出關聯性，這是目前最有效率的方法。各位可以想像，當機器看一張照片抓出裡面的人臉時，對於機器來說就是先找到裡面的人臉輪廓，然後把人臉輪廓取出進行辨識身份。聲音處理的方式也是一樣，目前主流廠商的做法是設計一個已學習多種噪音輪廓的模型去做聲音處理，在吵雜環境時，因為模型已有學習過類似場景，因此可以把噪音的圖騰刪除，變成只留下人的聲音，增加聽障者的聽感。

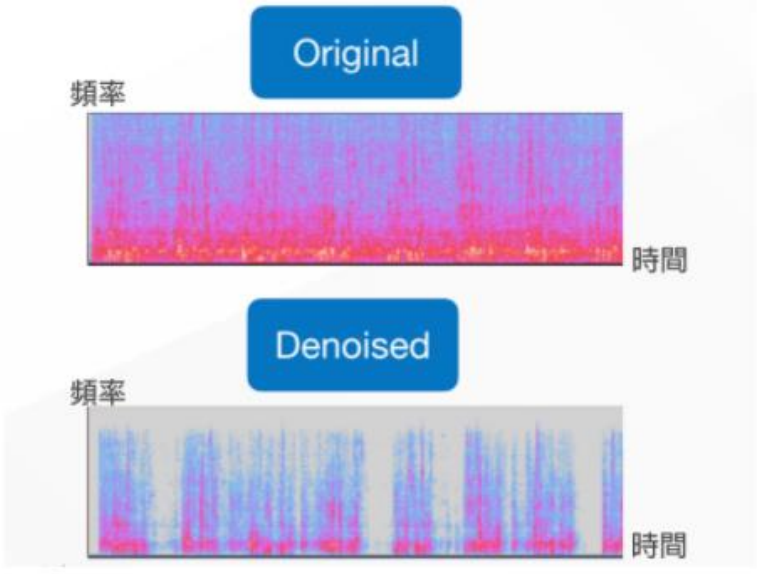
不過，這樣的深度學習的做法前提是：要有足夠的噪音場景的資料做訓練。實

際上，日常生活中的噪音非常多元，有些噪音中帶的人聲也是對使用者有意義的，例如像是電視，電視機發出的主播聲也是對聽障者有意義的，我們發現到早期的電視喇叭在低頻的播放能力稍弱，人聲聽起來會稍微尖銳，在不那麼足夠資料量訓練下來的模型會很容易把這樣的人聲當成是噪音直接濾掉。

因此，我們團隊反其道而行，專注在訓練如何更有效率追蹤聲紋，三年間錄製各國超過百萬小時聲紋資料，且聲紋的精準度已有在金融機構進行產品量產，以前述機器從照片中找人臉識別的例子為例，人臉的輪廓對應我們的想法就是聲紋，進行聲紋的分析並保留並刪去不必要的噪音後，讓使用者在這個場景之下更能聽清楚眼前這位說話者的聲音。

以下是我們在標準化實驗室錄製下的範例：

Noise Types	Noise Reduction Level
Pub_Noise	-32dB



「Original」下方的時頻圖是原始的音檔，背景噪音是模擬酒吧人聲鼎沸的時候的聲音（時頻圖中，顏色越接近粉紅色或白色代表能量越大，藍色或灰色代表能量越小）主要說話者在使用者正前方約 1m 左右說話（測試以 SNR = 5 為基準），經過我們聲紋追蹤並把非聲紋的部分（也就是噪音）移除後，下方的時頻圖可顯然後出講話者的聲音輪廓，幫助使用者在這樣的環境下能夠更專注聆聽主要說話者的聲音，而數據反映在降噪值約 32dB ，這是一個非常有效率解決吵雜場景聽覺痛點的方法。

人聲追蹤讓口罩人聲還原變得更可行

2020 年疫情爆發時，各國因為疫情關係大家都戴口罩，對話時會聽不清楚人

聲，我身為使用者也發現這個問題，針對這樣的情境做了許多研究並實驗。最終推出基於聲紋追蹤的方式去做口罩人聲的還原，這樣的做法有別於目前市面上已知的助聽器產品使用的口罩模式，當前口罩模式幾乎都是透過調高高頻的方式去補足因口罩阻礙聲音傳遞帶來的高頻減損，但是聲紋的形成是連貫性的，若只有增強高頻會讓人聲變形反而認不出來。身為使用者的我很能理解以語意理解的角度會希望對方的人聲仍然是自己所熟悉的人聲才能維持語意理解的水準。我們的作法是透過聲紋的追蹤，並且經過大數據訓練出的模型即時比對正常聲音的能量「應該」會是怎樣的樣子去做音量還原，去幫助使用者維持理解能力。

既然我們證明了深度學習的方式克服了許多問題，那是不是以後面對任何問題都永遠靠 AI 來解決？

凡事靠 AI 就對了嗎？ 那可未必。

就我們的觀點來看，過去助聽器發展了接近一個世紀，過去的作法以及上市後得到大量使用者反饋不斷地精進產品也許解決了 50 %的痛點，深度學習的方式「只」可能解決了另外 30 %的痛點，但是剩下的 20%我們認為更要從改善使用者體驗去下手才有辦法長足進步。為什麼深度學習的方式「只」可能解決 30%痛點呢？各位要知道的是，深度學習跑出來的結果都是機率，它的精確度取決於模型在訓練時的方式以及資料量的多寡。而模型訓練的成本是一個等比曲線的成長過程，換句話說，訓練到 90%準確率可能訓練只要花 10 小時，但是從 90~100%的過程可能就是要 100 小時以上的訓練時間倍數成長。且大多數的時候模型精準度跟模型大小強相關，模型越大所需的平台算力要更好以及要更省電，以目前的半導體技術現況，耳機類（包含消費型耳機或者是助聽器）的製程要做到 22 nm 以下才有辦法讓深度學習的模型有更長足的進步以及兼顧省電性能（篇幅有限日後有機會在說明原因），目前坊間市售的五大集團助聽器最好的製程到 28nm，就我們得知業界消息，包含我們團隊以及台灣多家 IC 設計公司正在進行 22nm 更先進製程的助聽器晶片試量產，相信 2022 年會有更豐富的產品給使用者做選擇。

即便是這幾年能解的技術問題，接下來的 20%使用者體驗是非常重要的，過去的助聽器對於使用者而言是單一化的產品，也就是助聽器提供怎樣的性能使用者就使用怎樣的性能，隨著科技的演進，可以想像未來助聽器慢慢會走向一個多元化工具的產品，讓使用者在聽不清楚的狀況之下也能夠善用這個「工具」輔助他聽得清晰。舉例來說，當使用者在更極端的環境下使用深度學習版本的模型降噪仍然聽不清楚時，是不是可以讓使用者只要做一點小改變，像是喬個位子能正面面對說話者或者是更靠近說話者就能夠讓模型更有效率處理，將這個過程變得更直觀，更方便使用將是助聽器製造商未來的課題。

最後，期待未來科技的演進，造福的都是廣大的聽障同胞。

QA 問題集

Q（編輯群）：您說「在相同時間內採樣數越高，帶給演算法的「可分頻」的空間越多（也就是坊間稱的頻道數），採樣數應和可處理的頻寬有關，兩者間成正比…」，就我們瞭解，採樣數雖和頻寬有所相關，但並不一定能對應頻道數目的切分，頻道數的切分是否應與分析的音框長度有更直接關係？

A（筆　者）：採樣數跟頻寬有關，也跟能做後續頻道數切分有關，因為坊間助聽器演算法的基底幾乎都是 FFT（當然有一些用 filter 分頻或者是直接波形圖做 DCT 之類，但這種要龐大運算資源，如果說一個號稱 32 個頻道的助聽器用這兩個方法去做根本不可能），採樣數太低，給 FFT 轉換太多個頻段再 IFFT 回來精度會掉超多，因此沒有廠商會在低採樣率的時候分一堆頻道。因此我才會說採樣數越高頻道數切分是相關的。

Q（編輯群）：您在文章內提到，「人聲頻道的範圍大約會著重在能量較強的 200 Hz-1000Hz 之間 ....」，不曉得此處所指的「人聲」是指人聲當中的哪些信息？ 因為就語音聲學分析的角度而言，語音信息特徵對應的頻寬應不僅落在 200 Hz-1000Hz 間，許多跟清晰度相關的子音訊息都是落在較高頻的位置；就如同雅文基金會 2020.02.20 製作的華語語音分布圖，華語語音的分布位置大致落在 250-8000Hz，尤其子音的部份大多落在 4000Hz-8000Hz 的高頻區域；又英文的語音分布也多集中在 500-4KHz（參考 Speech banana / string bean 或 Count the Dots（Killion& Mueller，2010）；與您內文所提到的資訊有所差異，不知道這部份的參考來源為何呢？

若您文章內所指的 200-1000 是指主要能量帶，那是否此敘述是與「基頻」、「母音」等較有關。另，以傳統的市話網路為例，其頻寬也有約 4kHz 左右，才能讓我們聽得清楚語音訊號，而非單純的 200-1000Hz；抑或您是指稱「聲紋追蹤的演算」只需要參考此段頻寬範圍，即可達到不錯的成果呢？

A（筆　者）：關於這點我先講一下我們研發（真實做產品）角度跟您提問的角度出發點本質不同。您提到的雅文華語語音分布圖在錄製的時候是以麥克風對真人（或者是其他錄音室）錄製，錄製到的聲音理論上是真實聲音的頻響圖沒錯，但是理論也是要給人聽的，我的角度是從發展一個助聽器產品面向，考量的是除了理論值之外，同時也是考量現行助聽器製作上面對這樣的問題之瓶頸，而助聽器專用的喇叭單體的性能就正是此瓶頸，以目前世界最大助聽器喇叭供應商婁氏動鐵產品為例，

幾乎所有動鐵單體皆在 2K 之後的頻率的響應會有將近 20db 以上落差，以 Knowles RVA-90080（Knowles RVA-90080 此單體很廣泛應用在重度助聽器之中）產品頻響圖來說，可見下圖做參考。

臨床個案如果其聽損圖需補償範圍在此單體極限頻響之下的範圍可以透過雅文的理論圖做補強，但是我們以產品角度出發，就必須要考量到全部患者可能的情境，特別是聽損圖界在此頻響的邊緣的時候要如何解決，基本上這時候還要對 2K 以上的頻響滿足做到雅文的理論值這樣的患者的需求受限於單體頻響，幾乎是不可能的事情。

在聲學產品設計領域來講通常我們會說人聲「大約介在 200Hz-1000Hz」主要原因是在講 200Hz-1000Hz 能量最強，並非是指整個人聲都在此頻率，而 1500-2000Hz 的英文強化的部分主要是考量上一段落講到動鐵單體的頻響極限大約在 2000Hz 最好，以後不是不行但若遇到像我們所說的情境（聽損圖屆在此頻響的邊緣）頻響落差會更大，因此我們在這個議題上以產品角度出發會以 2000Hz 為一個分水嶺。

還有，雅文那張圖表主要是來表達語言發音落在的頻率，他錄製的場景幾乎都是在很安靜室內或者是 SNR 極大的場景，但是如果考量現實生活的對話，現實生活中的各種聲音皆有所謂共振現象，在高頻率很有機率發生互抵現象，因此在助聽器麥克風收音的時候是否還能是保持高頻的完整性，並且經過演算法處理解決使用者需求，並非只有光靠「補償到雅文的理論值」這麼簡單。

Q（編輯群）：關於聲電圖（圖一）的部分，由於聽力師在調整助聽器時，其內部擴音設定調整會與使用者聽損程度有關，亦與聽力師的「主觀選擇」有關，因此您提供的此張圖片代表「該隻助聽器是這樣調整」，應無法代表「市售助聽器皆是如此的設計理念」，又或者「市售主流品牌助聽器頻率響應曲線」為此，不是嗎？

A（筆　者）：修改為目前內容

Q(編輯群)：以深度學習來進行降噪處理；在評估降噪效果時，如果是 mismatch testing condition 要達到 32dB 應該十分不容易的，能否請您就降噪測試的條件等相關資訊向我們多做說明呢？

A（筆　者）：測試條件原文有提到「主要說話者在使用者正前方約 1m 說話」，我們在多加入耳時的收音 SNR=5 的條件，之前只寫到前者是因為比較方便

讀者好理解大概就是生活中場景的距離。 有可能有專家會問為何不 SNR=0，因為 SNR=0 正常人要相對少數才能聽清楚，SNR=5-10 的範圍涵蓋率就高出很多。

Q（編輯群）：若是以大量資料來訓練，一般來說模型的維度都不低，也意味著系統的 group delay 可能偏高。要如何兼顧訓練成效和較低的延遲，或許也是值得探討的方向。

A（筆　者）：是的，最主要關鍵還是晶片能處理的能力，速度 / 省電 / 短延遲 ，我們有在籌備相關的文章，但目前比較會希望跟著晶片產品推出來之後發布。

編　輯

發行單位：台灣聽力語言學會　　發行日期：2021.12.01
發 行 人：葉文英　　聽語學報：第 100 期
主　　編：張晏銘　　執行編輯：鄭庭語
編輯顧問：曾進興　　助理編輯：潘沐萱
網　　址：www.slh.org.tw