語音語言模型最新進展：一份綜述

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King, Fellow, IEEE

摘要

大型語言模型（LLMs）近期獲得了廣泛關注，主要歸因於其在基於文字互動方面的能力。然而，自然的人機互動往往依賴於語音，這使得我們必須轉向基於語音的模型。實現這一目標的一個直接方法是採用「自動語音辨識（ASR）+ LLM + 文字轉語音（TTS）」的管線，其中輸入語音被轉錄為文字，由 LLM 處理，然後再轉換回語音。儘管這種方法直接，但它存在固有的限制，例如模態轉換過程中的資訊損失、複雜管線導致的顯著延遲，以及三個階段中的錯誤累積。為了解決這些問題，語音語言模型（SpeechLMs）——無需從文字轉換即可生成語音的端到端模型——已成為一個有前景的替代方案。本綜述論文首次全面概述了建構 SpeechLMs 的最新方法，詳細闡述了其架構的關鍵組成部分以及對其發展至關重要的各種訓練方案。此外，我們系統地調查了 SpeechLMs 的各種能力，對其評估指標進行了分類，並討論了這個快速發展領域中的挑戰和未來研究方向。 [1]

索引詞彙—語音語言模型、語音互動、大型語言模型。

I. 介紹

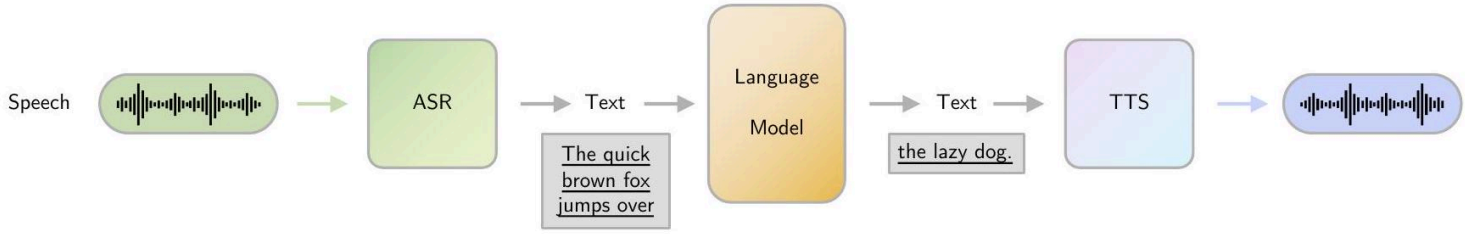
大型語言模型（LLMs）在生成文本和執行各種自然語言處理任務方面展現了卓越的能力 [1]-[3]，成為人工智慧驅動的語言理解和生成強大基礎模型。它們的成功也激發了在其他各個領域的眾多應用，然而，僅依賴基於文本的模態存在顯著的局限性。這導致了基於語音的生成模型的發展，這些模型允許更自然、直觀地與人類互動。語音的納入不僅促進了即時語音互動，還透過結合文本和語音資訊豐富了溝通 [4], [5]。

鑑於文本和語音之間存在廣泛的互資訊，修改現有的 LLMs 以實現語音互動能力是很自然的。一種直接的方法是採用「自動語音辨識（ASR）+ LLM + 文字轉語音（TTS）」框架（圖 1a）[6], [7]。在此設定中，

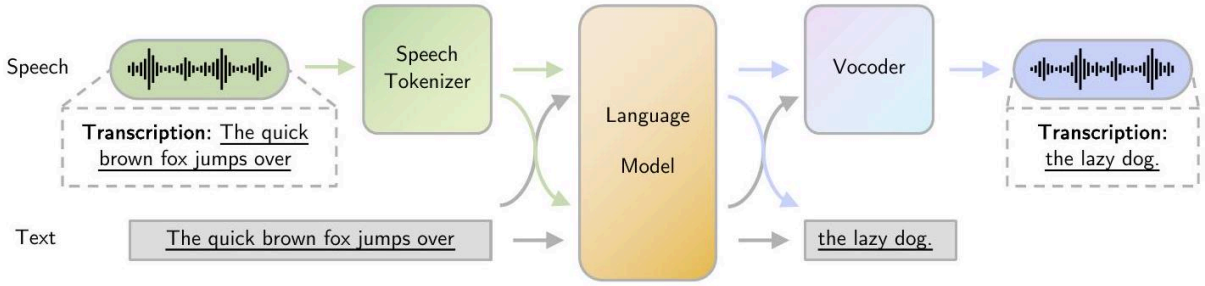
使用者的語音輸入首先由 ASR 模組處理，將其轉換為文字。然後，LLM 根據此轉錄生成文字回應。最後，TTS 模組將文字回應轉換回語音，並播放給使用者。然而，這種簡單的解決方案主要存在以下三個問題。

資訊遺失。語音訊號不僅包含語義資訊（即語音的意義），還包含副語言資訊（例如音高、音色、音調等）。將純文字的 LLM 置於中間會導致輸入語音中的副語言資訊完全遺失 [8]、[9]。辨識副語言特徵可以實現更具吸引力且身臨其境的互動，因為模型可以回應更豐富的上下文。此外，它還能讓模型在某些情況下更準確地解讀使用者的意圖，因為語音的意義會因語氣而異。
顯著延遲。ASR、LLM 和 TTS 的循序操作會導致相當大的延遲，因為這些模組的結構和管線本質上就很複雜 [9]-[11]。例如，ASR 通常包含一個額外的文字生成器 [12]、[13]，而 TTS 通常依賴於文字分詞器，這兩者都會增加計算需求。此外，實施進階解碼技術，例如用於 ASR 文字解碼的波束搜尋，可能會進一步導致延遲 [12]、[13]。因此，開發端到端的語音生成模型可以顯著減少延遲。
累積錯誤。這種分階段的方法很容易導致整個管線中累積錯誤，尤其是在 ASR-LLM 階段 [14]、[15]。具體來說，ASR 模組中將語音轉換為文字時發生的轉錄錯誤可能會對 LLM 的語言生成效能產生負面影響。此外，如果 LLM 生成無法合成的文字，TTS 效能可能會顯著下降。因此，開發統一架構在減少累積錯誤方面扮演著關鍵角色。

樸素的 ASR + LLM + TTS 框架的限制促使了語音語言模型（SpeechLMs，圖 1b）的發展。與樸素框架不同，SpeechLMs 直接將語音波形編碼為標記或表示，從音訊中捕捉基本特徵和資訊（第 III-A 節）。儘管單個語音標記可能不帶有詞級語義，但它們捕捉了語音語句的語義資訊。



(a) Illustration of the “ASR + LLM + TTS” framework.



(b) Illustration of the architecture of a SpeechLM.

圖 1：圖 1。「ASR + LLM + TTS」框架和 SpeechLM 的架構。我們強調，對於 SpeechLM，相同的內容可以跨語音和文字兩種模態使用，這表示任何輸入模態都可以產生相同結果的任何輸出模態。圖中刻意重複輸入/輸出內容以突顯此點。

表 1：表一
本調查中使用的符號。

符號	描述
a	語音音訊波形
â	模型重建的語音音訊波形
t	一段文字
M	包含語音和/或文字的多模態序列
θ	模型參數
$f_E(\cdot)$	語音編碼器
$d(\cdot)$	語音量化器
LM	語言模型
v	編碼語音表示
S	語音符記
m	多模態符記（語音和/或文字）
V	語言模型的詞彙表
E	語言模型的嵌入矩陣
解碼器	Transformer 解碼器區塊
語音 (Vo)	聲碼器 (Vocoder)
G	生成對抗網路中的生成器 (Generator in a Generative Adversarial Network)
D	生成對抗網路中的判別器 (Discriminator in a Generative Adversarial Network)
毫秒	語音音訊波形的梅爾頻譜圖
F_0	基頻

並保留有價值的副語言資訊，從而防止資訊遺失。語音語言模型（SpeechLMs）隨後以自迴歸方式對這些標記進行建模，不單純依賴文字輸入，這使得它們能夠利用額外的副語言資訊來生成更具表現力和細微差別的語音（第三節-B）。最後，生成的標記會被合成回語音（第三節-C）。這種整合方法消除了將三個獨立模組串聯起來的需求，顯著降低了延遲。此外，透過直接處理編碼的語音標記，語音語言模型有效地減輕了累積錯誤，因為它們的訓練與語音編碼整合在一起，而 LLMs（語言建模）的訓練在樸素框架中與 ASR（語音辨識）模組完全獨立。

語音語言模型（SpeechLM）有能力
超越簡單的對話任務，處理更複雜和多樣化的應用。首先，語音語言模型可以捕捉說話者特定的資訊和情感細微差別（圖2），這使得它們能夠在對話中區分不同的說話者，並理解和生成帶有特定情感語氣的語音。這些進步對於個人化助理、情感感知系統以及更細緻的人機互動情境等領域的應用至關重要。其次，語音語言模型可以設計成實現即時語音互動，模型可以被人類打斷，或者在使用者仍

在說話時選擇發言，更自然地模仿人類對話的動態。此外，由於語音語言模型直接在語音資料上進行訓練，它們有潛力促進稀有語言的溝通，在這些語言中，口語內容比書面材料更為普遍。

貢獻。在本調查中，我們首次全面概述了建構語音語言模型（SpeechLM）的最新努力。我們探討了構成其架構的各種組件（第三節）以及其開發所涉及的訓練方法（第四節）。我們旨在透過從上述角度分析這些模型，闡明該領域的現狀。此外，我們調查了語音語言模型的下游應用（第五節），對評估語音語言模型的指標進行了分類（第六節），討論了這個快速發展領域中遇到的挑戰，並概述了可能推動語音語言模型技術進一步發展的有前景的未來研究方向（第七節）。我們的貢獻總結如下：

我們發表了語音語言模型領域的第一份調查報告。
我們提出了一種新穎的分類法（圖 4），用於從底層組件和訓練方法對語音語言模型（SpeechLM）進行分類。
我們為語音語言模型（SpeechLM）的評估方法提出了一種新穎的分類系統。

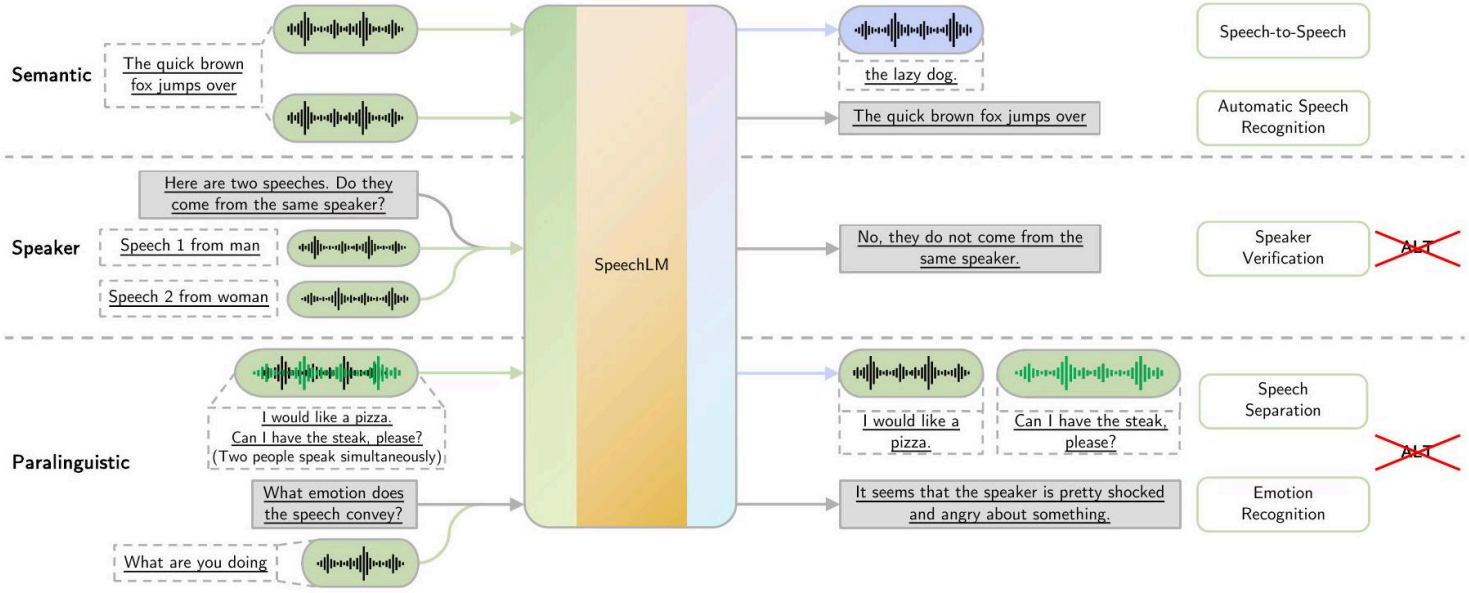


圖 2：圖 2. 語音語言模型（SpeechLM）的應用。我們使用 ALT 代表「ASR + LLM + TTS」框架。

我們在建構語音語言模型（SpeechLM）時發現了幾個挑戰。

與其他調查的連結。有幾項調查專注於傳統語音和音訊技術，例如口語理解（SLU）[16]、音訊和語音自監督學習（SSL）[17]、[18]，以及語音與其他模態的整合[19]、[20]。隨著 LLMs 的快速發展，一些研究回顧了單模態[21]、[22]和多模態[23]-[25]的 LLMs。此外，還有一些調查探討了音訊模態與 LLMs 之間的重疊。Latif 等人[26]檢視了音訊處理中的 LLMs，Peng 等人[27]回顧了 SLU 領域中的 SpeechLLMs，而 Ji 等人[28]則專注於包含語音、聲音和音樂的口語對話系統。

二、問題闡述

在本節中，我們將提供語音語言模型（Speech Language Models）的正式定義。語音語言模型（SpeechLM）是一種自迴歸基礎模型，它能端到端地處理和生成語音，並利用上下文理解來生成連貫的序列。這種能力使其能夠透過基於語音的互動來執行各種任務。儘管 SpeechLM 需要執行端到端的語音互動，但它們也可以整合文字，從而實現跨模態功能，例如語音輸入文字輸出（speech-in-text-out）反之亦然。我們注意到，SpeechLM 的概念與傳統的基於文字的語言模型（例如 LLM）形成對比，在傳統模型中，模型內部處理的唯一模態是文字。因此，為了避免混淆，在本調查中，我們將這些基於文字的語言模型稱為 TextLM。

我們提供了一個統一的框架，其中語音語言模型（SpeechLM）可以處理和生成語音資料、文字資料，甚至交錯的語音和文字資料。具體來說，語音音訊波形 $\mathbf{a} = (a_1, a_2, \ldots, a_Q)$ 由長度為 Q 的音訊樣本序列 $a_i \in \mathbb{R}$ 組成，其中 $1 \leq q \leq Q$ 。同樣地，文字片段 $\mathbf{t} = (t_1, t_2, \ldots, t_K)$ 由長度為 K 的文字標記序列 t_j （詞、子詞、字元等）組成。令 $\mathbf{M} = (M_1, M_2, \ldots, M_N)$ 表示長度為 N 的多模態序列，其中每個元素為 $M_i \in \{a_i, t_j\}$ 。我們將 $\mathbf{M}^{\text{in}} = \left(M_1^{\text{in}}, M_2^{\text{in}}, \ldots, M_{N_{\text{in}}}^{\text{in}}\right)$ 定義為輸入多模態序列，將 $\mathbf{M}^{\text{out}} = \left(M_1^{\text{out}}, M_2^{\text{out}}, \ldots, M_{N_{\text{out}}}^{\text{out}}\right)$ 定義為輸出多模態序列，其中 $N_{\text{in}} \geq 0$ 和 $N_{\text{out}} \geq 0$ 。那麼，由 θ 參數化的 SpeechLM 可以表示為：

$$\mathbf{M}^{\text{out}} = \text{SpeechLM}\left(\mathbf{M}^{\text{in}}; \theta\right).$$

三、SpeechLM 中的組件

SpeechLM 內部主要有三個組件，分別是語音分詞器、語言模型和語音合成器（聲碼器），如圖 1 所示。這種三階段設計模式的根本原因在於使用語言模型架構（例如，僅解碼器 Transformer）以音訊波形的形式自迴歸地建模語音。由於語言模型的輸入和輸出都是標記，因此需要將額外的模組連接到語言模型以處理輸入/輸出格式。具體來說，語音分詞器首先將連續的音訊波形轉換為標記或表示，作為語言模型的輸入，然後語言模型根據輸入語音標記執行下一個標記預測。最後，聲碼器將語言模型輸出的標記轉換回音訊波形。我們在此強調，我們的重點在於這三個組件如何組合形成 SpeechLM，而不是對每個組件進行全面概述。因此，對於語音分詞器和聲碼器，我們主要總結現有 SpeechLM 中使用的方法。表二總結了各種 SpeechLM 論文中這三個組件的流行選擇。

A. 語音分詞器

語音分詞器是語音語言模型（SpeechLM）中的第一個組件，它將連續的音訊訊號（波形）編碼為詞元。語音分詞器旨在捕捉音訊的基本特徵，

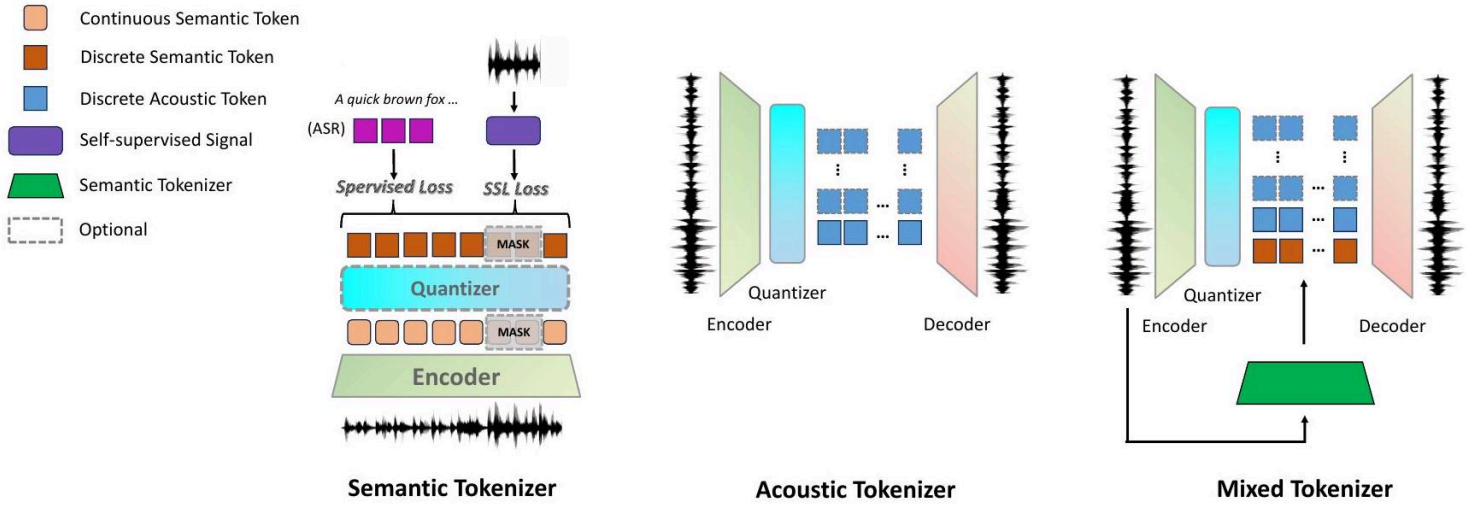


圖3：圖3。三種語音分詞器的示意圖。

同時降低其維度，並允許語言模型有效處理音訊輸入以進行自迴歸生成。語音分詞器透過逐段編碼音訊來運作，產生兩種可能的詞元類型（特徵）：離散詞元和連續詞元。離散詞元（第 IV-A1 節）使用特定索引來表示每個語音片段，而連續詞元（第 IV-A2 節）則使用嵌入來表示片段 [2] 。這兩種詞元類型都可以用作語言模型在自迴歸建模中的輸入。在本節中，我們根據語音分詞器在建模原始音訊不同方面的側重，將其分為三類。圖 3 說明了這三種語音分詞器。

語義理解目標：以語義理解為目標設計的語音分詞器旨在將語音波形轉換為能準確捕捉語音內容和意義的詞元。這些分詞器專注於從波形中提取語義特徵，這有助於提高語音辨識（ASR）等任務的效能。

語義理解語音分詞器通常包含語音編碼器和量化器。語音編碼器將波形中的基本資訊編碼為連續嵌入。然後，通常會整合一個量化器，將連續嵌入轉換為離散索引。令 $f_E(\cdot)$ 表示由 θ_{f_E} 參數化的語音編碼器，我們有 $\mathbf{v} = f_E(\mathbf{a}; \theta_{f_E})$ ，其中 $\mathbf{v} = (v_1, v_2, \ldots, v_P)$ 代表編碼後的嵌入。由於 $\mathbf{v}$ 仍然是連續的，因此使用量化器 $d(\cdot)$ 將嵌入離散化。根據不同的設計選擇，語音標記 $\mathbf{s} = (s_1, s_2, \ldots, s_P)$ 可以從 $\mathbf{a}$ 或 $\mathbf{v}$ 導出。因此，我們有 $\mathbf{s} = d(\mathbf{v}; \theta_d)$ 或 $\mathbf{s} = d(\mathbf{a}; \theta_d)$ 用於離散標記，以及 $\mathbf{s} = \mathbf{v}$ 用於連續標記。之後， $\mathbf{s}$ 可以用作訓練語音分詞器的目標標籤（例如遮蔽 $\mathbf{a}_{\text{mask}} \subset \mathbf{a}$ 並重建其對應的標籤 $\mathbf{s}_{\text{mask}} \subset \mathbf{s}$ [33]），或用於訓練後續的語言模型。

關鍵的設計選擇在於如何有效地將語音編碼和/或量化為標記。Wav2vec 2.0 [30] 使用卷積編碼器，然後是乘積量化

模組 [74] 將連續波形離散化。然後，對一部分量化表示進行遮蔽，並使用對比損失進行建模。W2v-BERT [32] 建立在 wav2vec 2.0 的基礎上，並提出除了對比損失之外，還使用遮蔽語言建模 (MLM) 損失 [75]。類似地，HuBERT [33] 使用 k-means 演算法將語音語句聚類成多個隱藏單元，然後執行 MLM 以從遮蔽的語音語句中預測目標隱藏單元。為了更好地對齊文本和語音模態的表示，Google USM [36] 在第二個預訓練階段利用文本注入損失 [76] 來提高下游任務的性能和穩健性。WavLM [37] 在預訓練期間增加了語音去噪目標。雖然大多數語音分詞器研究都集中在語義相關任務，例如 ASR 和 TTS，但 WavLM 表明語音去噪可以提高非語義任務的性能，例如說話者驗證和語音分離。第五節列出了所有下游任務。
2) 聲學生成目標：具有聲學生成目標的語音分詞器專注於捕捉生成高品質語音波形所需的聲學特徵。這些分詞器優先保留基本的聲學特性而非語義內容，使其適用於語音（重新）合成任務。

為了生成高品質的語音波形，聲學生成語音分詞器採用語音合成或語音重建的目標。為此，其架構通常包含一個編碼器、一個量化器和一個解碼器。如同之前所述，編碼器 $f_E(\cdot)$ 和量化器 $d(\cdot)$ 將原始波形轉換為語音標記。之後，解碼器 $f_D(\cdot)$ 將這些語音標記重建回語音波形。此過程由 $\hat{\mathbf{a}} = f_D(\mathbf{s}; \theta_{f_E})$ 表示，其中 $\hat{\mathbf{a}}$ 是生成的或重建的波形。

神經音訊編解碼器非常適合且主要用作聲學生成語音分詞器 [35]、[49]。這些編解碼器利用深度神經網路的先進建模能力，將音訊波形壓縮成緊湊的表示形式，通常以離散標記的形式呈現。透過編碼器-量化器-解碼器架構，編碼器將音訊壓縮成潛在表示，

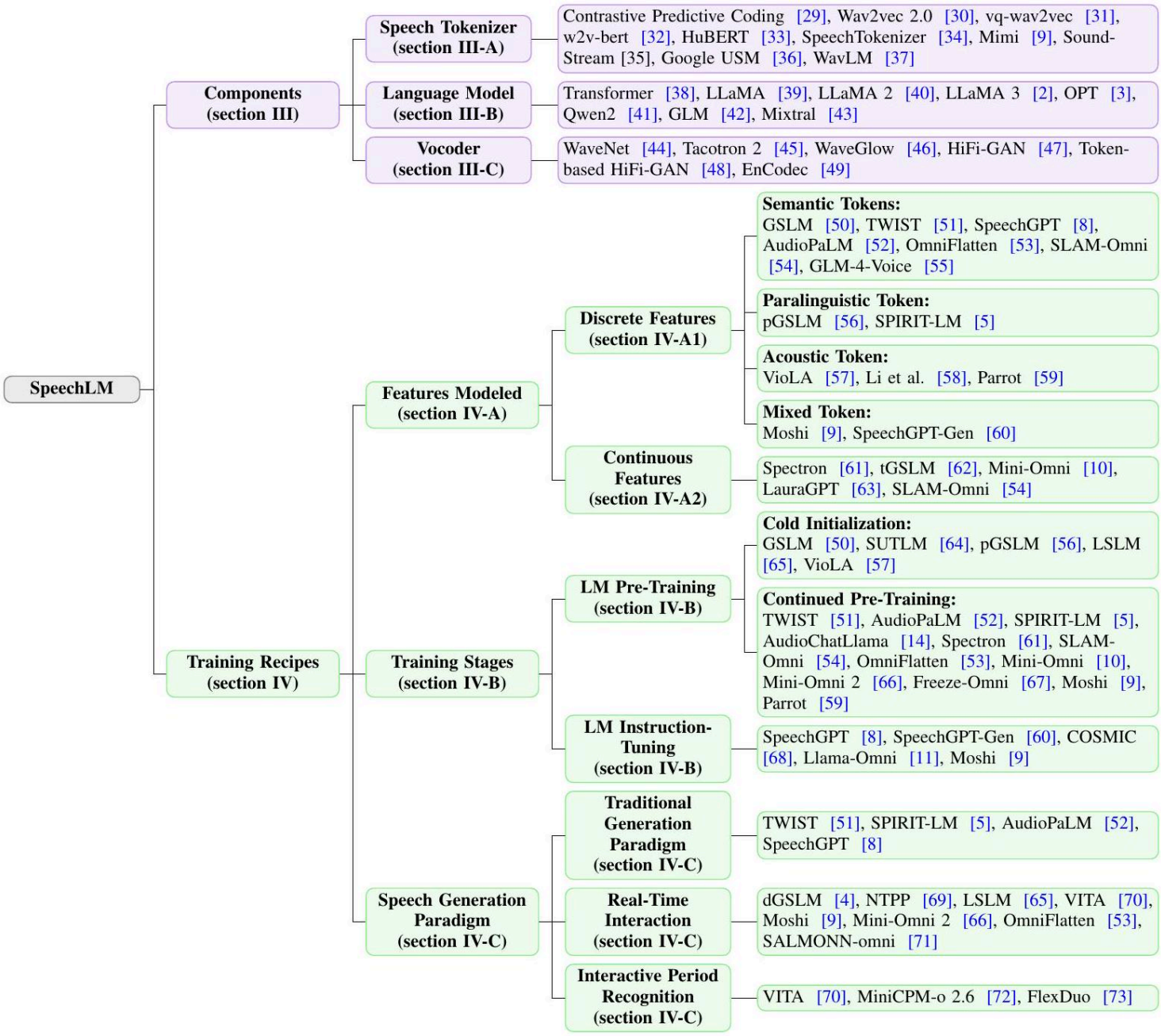


圖 4：圖 4. 語音語言模型分類。

量化器將這些表示離散化（通常透過向量量化 (VQ) [77] 或殘差向量量化 (RVQ) [35]），解碼器則將離散標記重建回音訊波形。因此，編碼器和/或量化器被用作聲學語音分詞器。

3) 混合目標：具有混合目標的語音分詞器旨在平衡語義理解和聲學生成。目標是利用這兩種類型分詞器的優勢。目前，這些分詞器的發展仍處於早期階段。大多數現有的混合語音分詞器主要採用聲學生成語音分詞器的架構，並專注於將語義分詞器的資訊提取到聲學分詞器中。SpeechTokenizer [34] 利用 RVQ-GAN [35], [49] 架構，將 HuBERT [33] 的語義資訊提取到 RVQ 的第一層。受 SpeechTokenizer 啟發，Mimi [9] 採用單一 VQ 從 WavLM [37] 中提取資訊，並結合另一個 RVQ 模組來學習聲學資訊。

在接下來的段落中，我們將介紹三種代表性語音分詞器的符號，每種都代表一個不同的類別。我們將 HuBERT 視為語義目標分詞器，
Encodec 視為聲學目標分詞器，而 SpeechTokenizer 則視為混合目標分詞器。

HuBERT。作為代表性的語義目標分詞器，HuBERT [33] 採用特徵編碼器 f_E 將原始音訊波形 a 轉換為連續嵌入 v，即 $f_E\left(\mathbf{a};\theta_{f_E}\right)=\mathbf{v}$ 。然後，這些嵌入透過 MFCC 特徵的 $\mathbf{k}$ -means 聚類量化為離散語音標記 $\mathbf{s}$ ，表示為 $d\left(\mathrm{MFCC}(\mathbf{a});\theta_d\right)=\mathbf{s}$ 。該模型以遮罩預測目標進行訓練，旨在最大化遮罩位置處正確標記的可能性：

$$\mathcal{L}(\theta)=\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\left[\sum_{i\in\mathcal{M}}-\log p\left(s_i\mid\mathbf{v}_{\setminus\mathcal{M}};\theta\right)\right],$$

其中 $\mathcal{M}$ 表示被遮蔽的索引。HuBERT 進一步疊代地精煉其語音標記，在每個步驟中更新編碼器和離散器參數，如下所示：

$$\mathbf{s}^{(n+1)}=d\left(f_E\left(\mathbf{a};\theta_{f_E}^{(n)}\right);\theta_d^{(n)}\right).$$

這種疊代過程使得學習越來越有意義的語音表徵成為可能。

Encodec。作為一種具代表性的聲學客觀分詞器，Encodec [49] 採用了帶有殘差向量量化 (RVQ) 的卷積編碼器-解碼器架構。編碼器 f_E 將原始音訊波形 a 映射到連續嵌入 $\mathbf{v}$ ，即 $\mathbf{v} = f_E\left(\mathbf{a}; \theta_{f_E}\right)$ 。然後使用多階段 RVQ 對這些嵌入進行離散化，其中每個階段 r 量化前一階段的殘差：

$$\mathbf{s} = d\left(\mathbf{v}; \theta_d\right) = \big(d_1\left(\mathbf{v}; \theta_{d_1}\right), d_2\left(\mathbf{v} - \hat{\mathbf{v}}_1; \theta_{d_2}\right), \\ \ldots, d_R\left(\mathbf{v} - \textstyle\sum_{r=1}^{R-1} \hat{\mathbf{v}}_r; \theta_{d_R}\right)\big),$$

其中 $\hat{\mathbf{v}}_r$ 表示在階段 r 處的量化嵌入。解碼器 f_D 從量化標記 $\hat{\mathbf{a}} = f_D\left(\mathbf{s}; \theta_{f_D}\right)$ 重建音訊波形。這種設計使 Encodec 能夠產生離散的聲學標記，這些標記保留了高傳真音訊資訊，適用於下游建模。

SpeechTokenizer。作為混合目標分詞器（mixed objective tokenizer）的代表，SpeechTokenizer [34] 透過利用 HuBERT 和殘差向量量化（RVQ）機制，結合了語義和聲學目標。編碼器 f_E 首先將輸入音訊波形 a 轉換為連續嵌入 $\mathbf{v}$ ，即 $\mathbf{v} = f_E\left(\mathbf{a}; \theta_{f_E}\right)$ 。離散化是透過多階段 RVQ 進行的。離散化過程使用多階段 RVQ，其操作方式與 Encodec 類似，不同之處在於第一個 RVQ 階段提取來自 HuBERT 的詞元，而後續階段則量化殘差。這種混合方法使 SpeechTokenizer 能夠捕捉高層語義和低層聲學資訊，以實現穩健的語音表示學習。

B. 語言模型

由於 TextLM [1]、[2]、[120] 的成功，大多數 SpeechLM 都遵循其架構。它們主要採用 Transformer [38] 或僅解碼器架構（例如 OPT [3]、LLaMA [39]）以自迴歸方式生成語音。為了正式定義它，給定 $|V_t|$ 作為詞彙量大小，以及 h 作為隱藏維度，一個典型的基於文本的僅解碼器 Transformer 語言模型由一個嵌入矩陣 $E_t \in \mathbb{R}^{|V_t| \times h}$ 、一系列 L 個 Transformer 解碼器區塊 $\mathbf{De} = \{De_1, De_2, \ldots, De_L\}$ 和一個輸出嵌入矩陣 $E_t' \in \mathbb{R}^{h \times |V_t|}$ 組成。因此，語言模型（LM）可以表示為

$$\mathbf{t}^{\text{out}} \sim \text{LM}\left(\mathbf{t}^{\text{in}}, \left(E_t, \mathbf{De}, E_t'\right)\right).$$

為了使語言模型適應生成語音，原始的文本分詞器被替換為第 III-A 節所示的語音分詞器。當使用離散詞元時， $E_t \in \mathbb{R}^{|V_t| \times h}$ 變為語音嵌入矩陣 $E_s \in \mathbb{R}^{|V_s| \times h}$ ，其中 $|V_s|$ 表示語音分詞器的詞彙量大小。輸出嵌入矩陣也從 $E_t' \in \mathbb{R}^{h \times |V_t|}$ 變為 $E_s' \in \mathbb{R}^{h \times |V_s|}$ 。因此，SpeechLM 中的語言模型表示為

$$\mathbf{s}^{\text{out}} \sim \text{LM}\left(\mathbf{s}^{\text{in}}, \left(E_s, \mathbf{De}, E_s'\right)\right).$$

由於語音語言模型（SpeechLMs）的語言模型架構是借鑒自文字語言模型（TextLMs），因此所產生的模型可以同時對文字和語音兩種模態進行建模 [5], [8]。為實現此目的，一種直接且最常採用的方法是擴展原始文字語言模型的詞彙表，以納入文字和語音符元。具體來說，語音嵌入矩陣通常會附加到文字嵌入矩陣的末端，從而形成一個更大的嵌入矩陣 $E_m \in \mathbb{R}^{(|V_t|+|V_s|) \times h}$ 。令 $\mathbf{m}$ 為一個包含語音和文字符元的符元序列，則所產生的語言模型變為

$$\mathbf{m}^{\text{out}} \sim \text{LM}\left(\mathbf{m}^{\text{in}}, \left(E_j, \mathbf{De}, E_j'\right)\right).$$

透過這種方式，模型可以在單一序列中同時生成文字和語音，從而實現更多樣化的應用（請參閱第五節）。相對地，當使用連續符元進行建模時，從語音符元器衍生的嵌入會直接輸入到語言模型中。在這種情況下，語言模型的架構保持不變。

C. 符元轉語音合成器（聲碼器）

在語言模型組件自動迴歸生成語音標記後，會利用一個語音標記轉語音模組（通常稱為聲碼器）將所有語音標記合成回語音波形。這個過程涉及將生成的語音標記所代表的語言和副語言資訊轉換為可聽見的音訊波形。這可以看作是語音標記器的逆向過程，因此可以表示為

$$\mathbf{a} = Vo\left(\mathbf{s}; \theta_{V_o}\right),$$

其中 Vo 是由 θ_{V_o} 參數化的聲碼器模型。
SpeechLM 聲碼器的管線可能因底層聲碼器模型而異。主要有兩種管線：直接合成和輸入增強合成。直接合成是指聲碼器直接將語言模型生成的語音標記轉換為音訊波形的管線。例如，Polyak 等人 [48] 採用 HiFi-GAN [47] 架構並將語音標記作為輸入。相較之下，輸入增強合成則採用額外的模組，在將語音標記輸入聲碼器之前，將其轉換為連續的潛在表示 [121]、[122]。使用此管線的主要原因是聲碼器通常需要中間音訊表示，例如梅爾頻譜圖 [47]、[80]、[123]，作為輸入。例如，CosyVoice [88] 引入了條件流匹配 (CFM) 模型將語音標記轉換為梅爾頻譜圖，然後利用 HiFi-GAN 合成最終波形。比較這兩種管線，直接合成通常比輸入增強合成更簡單、更快。然而，管線的選擇取決於所使用的輸入語音標記類型。來自聲學生成標記器的語音標記包含足夠的聲學資訊，使其適合直接合成。相反，來自語義理解標記器的語音標記提供豐富的語義資訊，但缺乏精細的聲學細節，尤其是在較高頻率。因此，這些語音標記最好在合成最終語音之前增強為富含聲學資訊的表示，例如梅爾頻譜圖。

表 2：表二

熱門語音大型語言模型中語音分詞器、語言模型和聲碼器的架構選擇摘要。「-」表示不存在或未指明，「*」表示該架構主要基於書面文本，「A, B」表示作者同時實驗了「A」和「B」作為組件，而「A + B」表示「A」和「B」結合起來作為組件。

方法	語音分詞器	語言模型	聲碼器
Kimi-Audio [78]	Whisper 編碼器 [12] + 線性投影器	Qwen2.5 [79]	BigVGAN [80]
Qwen2.5-Omni [81]	Whisper	Qwen2.5	Talker + Codec Decoder [81]
Minmo [82]	SenseVoice [83]	Qwen2.5	CosyVoice 2 [84]
Lyra [85]	Whisper [12]	Qwen2-VL [86]	HuBERT + HiFi-GAN
Flow-Omni [87]	Whisper 編碼器 + 線性投影器	Qwen2 [41]	流匹配（Transformer + MLP）+ HiFi-GAN
SLAM-Omni [54]	Whisper 編碼器 + 線性投影器	Qwen2 [41]	-
OmniFlatten [53]	CosyVoice 編碼器 [88]	Qwen2	CosyVoice 解碼器 [88]
SyncLLM [89]	HuBERT [33]	LLaMA-3 [2]	HiFi-GAN [47], [48]
EMOVA [90]	SPIRAL [91]	LLaMA-3	VITS [92]
Freeze-Omni [67]	Transformer [38]	Qwen2	TiCodec [93]
IntrinsicVoice [94]	HuBERT	Qwen2	HiFi-GAN
Mini-Omni2 [66]	Whisper	Qwen2	Mini-Omni [10]
SALMONN-omni [71]	Mamba Streaming 編碼器 [95]	-	VoiceCraft [96] + Codec 解碼器
Zeng 等人 [97]	Whisper + VQ	GLM [42]	CosyVoice
NTPP [69]	VQ-VAE	LLaMA-3、Mistral、Gemma 2	HiFi-GAN
GPST [98]	EnCodec [49]	Transformer	Codec 解碼器
GLM-4-Voice [55]	Whisper + VQ [9]	GLM-4-9B-Base [42]	CosyVoice
Moshi [9]	Mimi [9]	Transformer*	咪咪
VITA [70]	CNN + Transformer + MLP [70]	Mixtral [43]	文字轉語音工具包 [70]
LSLM [65]	vq-wav2vec [31]	僅解碼器 Transformer	UniVATS [99]
SpiRit-LM [5]	HuBERT, VQ-VAE [77], speechprop	LLaMA-2 [40]	HiFi-GAN
TWIST [51]	HuBERT	OPT [3], LLaMA [39]	HiFi-GAN
PSLM [100]	HuBERT	NekoMata [101]	HiFi-GAN
VOXTLM [102]	HuBERT	OPT [3]	HiFi-GAN
Voicebox [103]	EnCodec	Transformer* [38]	HiFi-GAN
Park 等人 [104]	AV-HuBERT [105]	OPT	HiFi-GAN
USDM [106]	XLS-R [107]	Mistral	Voicebox [108]
VioLA [57]	EnCodec	Transformer*	Codec 解碼器 [49]
FunAudioLLM [83]	SAN-M [109]	Transformer*	HiFTNet [110]
SpeechGPT-Gen [60]	SpeechTokenizer [34]	LLaMA-2	SpeechTokenizer 解碼器 [34]
ICoT [?]	SpeechTokenizer	LLaMA-2	SoundStorm
AnyGPT [111]	SpeechTokenizer	LLaMA-2	SoundStorm
LauraGPT [63]	Conformer*	Qwen [112]	Transformer + Codec Decoder
Spectron [61]	Conformer*	PaLM 2* [113]	WaveFit [114]
AudioLM [115]	w2v-BERT [32]	僅解碼器 Transformer*	SoundStream* [35]
UniAudio [116]	EnCodec、Hifi-codec [117]、Improved RVQGAN [118]	Transformer*	Codec 解碼器
Llama-Omni [11]	Whisper	LLaMA-3.1	HiFi-GAN
Mini-Omni [10]	Whisper + ASR 轉接器 [10]	Qwen2	TTS 轉接器 [10]
tGSLM [62]	語音分割 + SSE [119] + 詞彙嵌入器	Transformer*	Tacotron-2 + Waveglow [45], [46]
SpeechGPT [8]	HuBERT	LLaMA	HiFi-GAN
dGSLM [4]	HuBERT	Dialogue Transformer [4]	HiFi-GAN
SUTLM [64]	HuBERT	Transformer*	-
pGSLM [56]	Hubert	MS-TLM [56]	HiFi-GAN
GSLM [50]	HuBERT、CPC [29]、Wav2vec 2.0 [30]	Transformer*	Tacotron-2 + Waveglow

聲碼器可依其架構選擇進行分類。在以下章節中，我們將概述在語音語言模型開發中主要採用的聲碼器。

基於 GAN 的聲碼器：生成對抗網路（GAN）是聲碼器最常採用的架構 [47]、[48]、[80]、[123]、[124]。它以在語音合成任務中快速且高傳真地生成而聞名。GAN 的架構包括一個生成器和一個判別器。具體來說，生成器從隨機雜訊或輸入特徵中創建逼真的音訊波形，而判別器則根據真實音訊樣本評估生成音訊的真實性。

為了利用 GAN 合成高傳真語音，設計了各種訓練目標，側重於不同的方面。首先，GAN 損失被用作生成器和判別器操作的基本目標。具體來說，生成器（ G ）和判別器（ D ）的 GAN 損失通常選擇使用最小平方損失函數。生成器（ $\mathcal{L}_{\text{GAN}}(G;D)$ ）和判別器（ $\mathcal{L}_{\text{GAN}}(D;G)$ ）的 GAN 損失為

$$\mathcal{L}_{\text{GAN}}(G;D) = \mathbb{E}_{ms}\left[(D(G(ms))-1)^2\right]$$

和

$$\mathcal{L}_{\text{GAN}}(D;G) = \mathbb{E}_{(x,ms)}\left[(D(x)-1)^2 + (D(G(ms)))^2\right]$$

分別。在這些損失函數中， x 代表真實音訊， ms 代表其梅爾頻譜。其次，大多數基於 GAN 的聲碼器從梅爾頻譜合成語音波形，因此提出了梅爾頻譜損失，以對齊生成器合成的梅爾頻譜和從真實波形轉換而來的梅爾頻譜，以提高生成語音的傳真度。梅爾頻譜損失（ $\mathcal{L}_{\text{Mel}}(G)$ ）透過最小化上述兩種梅爾頻譜之間的 L1 距離來運作。其公式如下所示：

$$\mathcal{L}_{\text{Mel}}(G) = \mathbb{E}_{(x,ms)}\left[\|\phi(x) - \phi(G(ms))\|_1\right],$$

其中 $\phi(\cdot)$ 是將波形轉換為對應梅爾頻譜圖的函數。第三，為了進一步提升生成保真度，提出了特徵匹配損失 ($\mathcal{L}_{FM}(G;D)$)，用於對齊判別器編碼的特徵。
真實樣本與生成樣本之間的 L1 距離，其公式如下：

$$\mathcal{L}_{FM}(G;D) = \mathbb{E}_{(x,ms)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(x) - D^i(G(ms))\right\|_1\right],$$

其中 $D^i(\cdot)$ 和 N_i 分別表示鑑別器第 i 層的特徵和特徵數量。

在架構選擇方面，基於 GAN 的聲碼器著重於注入歸納偏置（inductive biases）以生成音訊波形。MelGAN [123] 在生成器中加入了帶有擴張（dilations）的殘差區塊，以模擬音訊時間步長之間的長程相關性，並提出了一種多尺度架構用於鑑別器，以模擬音訊的不同頻率範圍。基於多尺度鑑別器的概念，HiFi-GAN [47] 提出了一種多週期鑑別器，以模擬音訊波形中多樣的週期性模式。為了保留高頻內容，Fre-GAN [124] 採用離散小波轉換（Discrete Wavelet Transform, DWT）進行降採樣，並學習多個頻帶上的頻譜分佈。與傳統方法如平均池化（Average Pooling, AP）不同，DWT 能有效地將訊號分解為低頻和高頻子頻帶。BigVGAN [80] 引入了一種稱為蛇形函數（snake function）的週期性活化函數，並結合抗鋸齒表示（anti-aliased representation），以減少合成音訊中的高頻偽影。
2) 基於 GAN 的神經音訊編解碼器：鑑於許多神經音訊編解碼器採用 GAN 架構，它們可以在基於 GAN 的聲碼器背景下進行有效討論。與語音標記器不同，編解碼器中的解碼器被用作聲碼器[35]、[49]。Polyak 等人[48]利用 HiFi-GAN [47]作為聲碼器骨幹，並提出將聲碼器的輸入特徵分解為不同的屬性[48]，其中包括語義標記、音高標記和說話者嵌入。這種設計選擇使編解碼器能夠更好地執行音高和說話者相關任務，例如語音轉換和 F_0 操作。
3) 其他類型的聲碼器：聲碼器的種類不限於前面提到的那些，因為那些是語音語言模型（SpeechLM）中常用的。本節簡要概述了其他潛在的聲碼器類型，這些類型很少被探索作為語音語言模型中的組件。

純訊號處理聲碼器。純訊號處理聲碼器是傳統方法，依賴確定性演算法而非深度學習模型來合成語音[125]、[126]。然而，這類聲碼器在合成音訊中引入了明顯的偽影，因此很少使用。

自迴歸聲碼器。自迴歸聲碼器一次生成一個音訊波形樣本，每個樣本都以先前生成的樣本為條件[44]。這種方法由於其序列性質以及捕捉音訊訊號中複雜時間依賴關係的能力，可以實現高品質的音訊合成。然而，序列生成過程可能計算成本高且耗時，使得自迴歸模型與基於 GAN 的聲碼器等並行化方法相比效率較低。

基於流的聲碼器。基於流的聲碼器旨在建立一系列可逆轉換，將簡單的離散訊號映射為

表3：表三
語音語言模型預訓練和指令調優階段常用的資料集摘要。*表示該資料集是使用文字轉語音（TTS）合成的文字資料集的語音版本。S2ST 和 S2TT 分別代表語音到語音翻譯和語音到文字翻譯。

資料集	類型	階段	時數	年份
LibriSpeech [131]	ASR	預訓練	1k	2015
多語種 LibriSpeech [132]	ASR	預訓練	50.5k	2020
LibriLight [133]	ASR	預訓練	6萬	2019
People 資料集 [134]	ASR	預訓練	3萬	2021
VoxPopuli [135]	ASR	預訓練	1.6 k	2021
Gigaspeech [136]	ASR	預訓練	40k	2021
Common Voice [137]	ASR	預訓練	2.5k	2019
VCTK [138]	ASR	預訓練	0.3k	2017
WenetSpeech [139]	ASR	預訓練	22k	2022
LibriTTS [140]	TTS	預訓練	0.6k	2019
CoVoST2 [141]	S2TT	預訓練	2.8k	2020
CVSS [142]	S2ST	預訓練	1.9k	2022
VoxCeleb [143]	說話者辨識	預訓練	0.4 k	2017
VoxCeleb2 [144]	說話者辨識	預訓練	2.4 k	2018
Spotify 播客 [145]	播客	預訓練	47k	2020
費雪 [146]	電話交談	預訓練	2千	2004
SpeechInstruct* [8]	指令遵循	指令微調	-	2023
InstructS2S-200K* [11]	遵循指令	指令微調	-	2024
VoiceAssistant-400K* [10]	遵循指令	指令微調	-	2024

分佈，例如高斯分佈，轉換為音訊樣本的複雜分佈。這種機制可以實現高效的取樣和密度評估，使模型能夠並行而非依序地合成音訊，從而顯著提升速度和品質 [46]。與基於 GAN 的聲碼器相比，基於 Flow 的聲碼器通常需要更多的參數和記憶體來訓練模型，這阻礙了它們的有效利用 [123]。

基於 VAE 的聲碼器。變分自編碼器 (VAE) 是強大的生成模型，可學習將輸入資料編碼為壓縮的潛在空間，同時允許重建原始資料 [77]、[127]。然而，VAE 很少被探索作為聲碼器的底層架構。

基於擴散的聲碼器。擴散模型近年來已成為一類強大的生成模型，可用於高傳真語音合成。它們透過逐步向輸入資料（例如音訊波形）添加雜訊來創建一系列越來越嘈雜的表示，然後學習反轉此過程以生成新樣本 [128][130]。例如，DiffWave [128] 使用去噪擴散機率模型 (DDPM) 來合成音訊。

在以下段落中，我們將介紹 HiFiGAN [47] 的符號表示法，因為它是語音語言模型 (SpeechLM) 中最常用的聲碼器。HiFiGAN 使用生成器-判別器框架，從梅爾頻譜圖或語音標記合成高傳真音訊波形。生成器 $G\left(\mathbf{s};\theta_G\right)$ 將語音標記序列 $\mathbf{s}$ 映射到輸出音訊波形 $\mathbf{a}$ ，即：

$$\mathbf{a}=Vo\left(\mathbf{s};\theta_{Vo}\right)=G\left(\mathbf{s};\theta_G\right),$$

其中 Vo 表示聲碼器函數， $\theta_{Vo}=\theta_G$ 是其參數。HiFi-GAN 採用多週期和多尺度判別器 $D_{MPD}\left(\mathbf{a};\theta_{MPD}\right)$ 和 $D_{MSD}\left(\mathbf{a};\theta_{MSD}\right)$ ，在對抗訓練期間區分真實音訊和生成音訊。在推論時，僅使用生成器 G 來有效重建語音波形。

IV. 訓練方法

在本節中，我們將對近期語音語言模型論文中常用的訓練方法進行分類和總結。這
包含對語音語言模型（SpeechLM）中建模的特徵類型、各個訓練階段及其所採用的技術，以及生成語音的不同範式進行概述。

A. 建模的特徵

建模的特徵指的是語音分詞器輸出的特徵或符記類型，以及語音語言模型（SpeechLM）中語言模型組件所建模的內容。這些特徵在決定語音語言模型的能力和性能方面扮演著關鍵角色。不同的特徵從不同方面對語音波形進行建模。在本節中，我們將總結語音語言模型中常用的特徵，並著重探討不同特徵如何影響語音語言模型的性能。根據最新的發展，我們可以將語音語言模型所建模的特徵分為兩大主要類型：離散特徵和連續特徵。

離散特徵：離散特徵（或離散符記）指的是語音訊號的量化表示，可以表示為獨立、可計數的單位或符記。這些特徵通常透過各種編碼和量化過程從語音訊號中提取，產生一組有限的可能值。離散特徵是語音語言模型最常使用的特徵，因為它們可以表示為符記，並能以與文字語言模型（TextLM）中的文字符記完全相同的方式進行建模。

大多數語音語言模型（SpeechLMs）僅使用語義標記（由語義理解標記器生成，第 III-A1 節）來表示語音，因為語義資訊在口語交流中扮演著最關鍵的角色。GSLM [50] 是第一個語音語言模型，它比較了三種標記器，包括對比預測編碼（Contrastive Predictive Coding, CPC）[29]、wav2vec 2.0 [30] 和 HuBERT [33]。它得出結論，HuBERT 在語音重合成和語音生成等各種任務上表現最佳。大量研究遵循此設定，並使用 HuBERT 作為語音標記器 [5]、[8]、[51]。AudioPaLM [52] 實驗了 w2v-bert [32]、USM-v1 [36] 和 USM-v2 [52]（USMv1 的修改版本）之間的選擇，並得出結論，USM-v2 是在 ASR 和語音翻譯（ST）任務上表現最佳的語音標記器。

儘管語義標記因其對語音波形中上下文資訊的建模而擅長生成具有語義意義的語音，但研究人員發現，僅基於語義標記生成的語音缺乏表達性資訊，例如語調和不同的音高或音色 [5]、[147]。為了克服這個限制，副語言標記可以整合到建模過程中，以捕捉語音中的表達性資訊。pGSLM [56] 提出使用基頻（F0）和單位持續時間作為韻律特徵來補充 HuBERT 語義標記，並訓練一個多流變壓器語言模型來分別預測語義標記、音高（F0）和單位持續時間。類似地，SPIRIT-LM [5] 用音高和風格標記 [148] 補充了 HuBERT 語義標記。這種額外聲學標記的整合使語音語言模型能夠更有效地捕捉表達元素，而不會顯著損害語義理解 [5]。

另一種類型是聲學標記，這些標記旨在捕捉基本的聲學特徵以重建高傳真語音，主要從神經音訊編解碼器模型中獲得（第 III-A2 節）。一些研究直接在語言模型中建模編解碼器標記，這通常被視為編解碼器語言模型（CodecLMs）。例如，Viola [57] 訓練了一個能夠執行 ASR、TTS 和機器翻譯的 CodecLM。NTPP [69] 則在 VQ-VAE [77] 標記上進行訓練，以建模雙通道口語對話資料。

討論。不同類型的標記（token）會以不同方式影響語音語言模型（SpeechLM）的語音品質，這通常會導致權衡取捨 [115]。例如，語義標記雖然與文本高度對齊，擅長生成語義連貫的語音，但生成的語音往往缺乏聲學細節，例如高頻資訊。恢復和增強這些細節通常需要後處理，例如擴散模型（diffusion model），這會顯著增加模型的延遲。相反地，聲學標記可以促進高傳真音訊的生成，但卻常在內容生成方面出現不準確的問題 [34]。研究人員嘗試了兩種方法來平衡這些權衡。第一種方法是將語義標記和聲學標記組合成一個序列。AudioLM [115] 提出了一種分層建模方案，首先從 w2v-bert [32] 建模語義標記，然後使用這些標記來預測來自 SoundStream [35] 的聲學標記，最終生成語音。然而，這種方法會增加序列長度，進而增加建模複雜度。第二種策略是利用混合標記（第三章 A3 節）來共同建模語義和聲學資訊，這在 Moshi [9] 和 SpeechGPT-Gen [60] 中顯示出有前景的結果。
2) 連續特徵：連續特徵（或連續標記）與離散特徵不同，它們是未量化的、實數值的語音訊號表示，存在於連續尺度上（連續標記）。連續特徵可以包括頻譜表示，例如梅爾頻譜圖（mel-spectrograms），或從神經網路中提取的潛在表示。Spectron [61] 透過逐幀預測頻譜圖來執行語音延續。Mini-Omni [10] 和 SLAM-Omni [54] 從凍結的 Whisper 編碼器中提取中間表示作為 SpeechLM 的輸入，而 LauraGPT [63] 則採用與語言模型一同訓練的音訊編碼器，從輸入語音中導出潛在表示。連續特徵可以捕捉語音中可能在離散化過程中丟失的細微、微妙的方面。然而，使用這些特徵通常需要修改語言模型的現成訓練流程，因為傳統的基於文本的模型是為處理離散單元而構建的。此外，與離散特徵相比，連續特徵需要更大的儲存容量。

B. 訓練階段

訓練語音語言模型（SpeechLM）涉及訓練三個主要組件：語音分詞器、語言模型和聲碼器。與文本語言模型（TextLM）類似，訓練 SpeechLM 的關鍵在於有效建模語音延續，這主要是語言模型的職責。語音分詞器和聲碼器通常依賴於既定方法，並且

表四：表四
四種不同的語音和文字符號建模方法。

建模方法	範例	說明
僅語音	[語音] S12 S34 S33 ... S11 S59	僅提供語音序列。
純文字	[文字] 敏捷的棕色狐狸躍過懶惰的狗。	僅提供文字序列。
語音-文字串接	[語音] S12 S34 S33 ... S11 S59 [文字] 一隻敏捷的棕色狐狸跳過一隻懶惰的狗。	語音序列和文字序列被串聯在一起。
交替的語音-文字	[語音] S12 S34 S33 [文字] 棕色狐狸跳過一隻懶惰的 [語音] S11 S59	該序列會與語音和文字符記交錯。

是使用針對每種語音語言模型方法特有的不同訓練資料集進行訓練的。因此，本節將回顧用於訓練語言模型組件的主要技術。繼文字語言模型之後，我們將語音語言模型的訓練過程分為三個階段，包括預訓練、指令微調和後對齊。

語言模型預訓練：語音語言模型中的語言模型預訓練是一個關鍵階段，它顯著影響模型生成連貫且與上下文相關的語音的能力。此階段通常涉及訓練語言模型，使其能夠在大量的語音符記語料庫上自動迴歸地預測下一個符記。此階段的主要目標是學習語音資料中固有的統計模式和依賴關係，使模型能夠根據前面的上下文預測序列中的下一個符記。

訓練資料。語音語言模型的預訓練主要利用大規模的開源語音資料。常用的資料集包括用於自動語音辨識 (ASR) [131]、[133]-[135]、文字轉語音 (TTS) [140]、語音翻譯 (ST) [135]、[142]、播客 [145] 和對話 [146] 的資料集。表 III 包含預訓練階段中使用的熱門資料集。有些資料集僅包含語音資料，而另一些則同時包含語音和相應的文字轉錄。包含文字轉錄可以透過讓模型學習口語及其書面形式之間的關係來增強模型的表示，這將在稍後討論。

冷啟動。有些語音語言模型（SpeechLM）在預訓練階段使用冷啟動，此時模型參數會隨機初始化。開創性的語音語言模型 GSLM [50] 從頭開始訓練一個 Transformer [38] 作為語言模型。這項研究展示了 SpeechLM 管線的有效性，並比較了各種語音分詞器選項

的效能。他們發現 HuBERT [33] 在理解語音內容和生成自然語音方面優於 CPC [29] 和 wav2vec 2.0 [30]。SUTLM [64] 也使用 Transformer 作為語言模型。他們透過比較四種不同的建模方法：僅語音、僅文本、語音-文本串聯以及交替（穿插）語音-文本，研究了共同建模語音和文本分詞的關鍵問題。他們發現交替語音-文本的設定在跨模態評估中表現最佳。表 IV 說明了這四種建模方法。

有些研究利用與標準 Transformer 不同的架構。當架構與標準 Transformer 或 TextLM 差異過大時，這些模型通常會從頭開始訓練。例如，pGSLM [56] 提出了一種多流 Transformer 語言模型（MS-TLM）
它接收多個輸入流並預測多個輸出流，以同時生成語音單元、持續時間和音高嵌入。dGSLM [4] 引入了對話 Transformer 語言模型（DLM），以共同建模來自兩位說話者的對話語音資料。為了使 SpeechLM 在說話時具備聆聽能力，LSLM [65] 建議將串流式自我監督學習（SSL）編碼器連接到基於自迴歸分詞的 TTS 模型。

持續預訓練。與冷啟動相反，持續預訓練涉及使用來自 TextLM 的預訓練權重初始化語言模型，然後使其適應處理語音分詞。這種方法利用了 TextLM 中嵌入的語言知識，從而實現了更有效率和更有效的 SpeechLM 訓練。Hassid 等人 [51] 發現，從文本預訓練語言模型（OPT [3] 和 LLaMA [39]）開始，可以提高模型的收斂速度並顯著改善其語音理解能力。他們還證明，雖然從文本預訓練檢查點進行訓練優於冷啟動，但從圖像預訓練檢查點進行訓練的結果比冷啟動差。這表明並非所有預訓練檢查點都同樣有效。此外，AudioPaLM [52] 使用 PaLM 和 PaLM-2 [149]、[150] 訓練 SpeechLM，表明 SpeechLM 受益於預訓練檢查點尺寸的增加和更大的訓練資料集。

透過對齊文字和語音模態表示，可以進一步提升語音語言模型（SpeechLM）的效能。有些研究將文字和語音表示對齊在單一序列中。SPIRIT-LM [5] 發現，在文字語言模型（TextLM）檢查點上，使用交錯的文字和語音標記進行持續預訓練，可以顯著提升模型在語音理解和生成方面的效能。此外，他們的視覺化結果顯示，與未採用此方法的模型相比，使用交錯標記序列訓練的模型，其文字和語音特徵之間的相似度顯著更高。Spectron [61] 透過共同監督多個目標來解決文字-語音表示對齊問題。具體來說，輸入的語音提示首先被轉錄成文字標記，然後模型預測文字標記回應。最後，文字回應被合成為輸出語音。SpeechGPT [8] 也採用了這個概念，但將其應用於指令微調階段。其他一些方法則執行多序列表示對齊。這種方法同時生成文字序列和語音序列。例如，Llama-Omni 使用 LLM 輸出隱藏狀態來解碼文字標記並同時生成離散語音
標記。Mini-Omni [10] 平行生成一個文字標記序列和七個聲學標記序列，所有這些序列都在句子層級對齊。同樣地，Moshi [9] 平行生成一個文字標記序列、一個語義標記序列和七個聲學標記序列，這些序列在詞彙層級對齊。

討論。對齊文字和語音表示的主要目標是利用基於文字模型的優勢來增強基於語音的模型。研究人員發現，訓練語音語言模型（SpeechLM）比訓練文字語言模型（TextLM）更具挑戰性。這種困難源於文字是知識的濃縮形式，而語音則要求模型獨立學習口語規則。對齊文字和語音表示已證明有效，但其中涉及各種權衡。首先，文字主要傳達語義資訊，這可以提高 SpeechLM 的語義建模能力，但在對齊過程中可能會損害其捕捉語氣和情感等副語言特徵的能力。其次，對齊模型有兩種主要的推論方法：文字存在（text-present）和文字獨立（text-independent）。文字存在推論同時解碼文字和語音，這可能會增加延遲，但能增強 SpeechLM 的推理能力 [10] 並減少可能的幻覺 [9]。相反地，文字獨立推論效率更高，但可能缺乏穩定性。此外，是否納入文字模態以增強 SpeechLM 效能仍然是一個懸而未決的問題，特別是考慮到人類通常在掌握書面語言之前就已習得口語技能。
2) 語言模型指令微調：指令微調是指對語音語言模型（SpeechLMs）進行微調的過程，使其能遵循特定指令來執行各種任務。此階段對於增強預訓練模型的泛化能力，並使其更能適應多樣化應用至關重要。因此，重點在於建立有效的指令遵循資料集。

目前已提出數種方法來建構語音語言模型的指令遵循資料集。SpeechGPT [8] 和 SpeechGPT-Gen [60] 提出兩階段的指令微調，包括跨模態指令微調和模態鏈指令微調。在第一階段，指令資料是根據自動語音辨識（ASR）資料集生成的，方法是將指令附加到配對的 ASR 資料中，要求模型將語音轉換為文字。同樣地，配對資料也用於建立執行文字轉語音（TTS）的指令資料。在第二階段，他們透過使用 TTS 轉換基於文字的指令遵循資料集，來建構語音輸入-語音輸出的資料集。Llama-Omni [11] 也透過合成基於文字的資料集來建立指令遵循資料，並遵循特定限制。首先，他們將輸入的文字提示轉換為模仿自然語音模式的格式。接著，他們捨棄原始的文字回應，並使用文字語言模型（TextLM）來生成對轉換後提示的答案，確保這些回應也遵循自然語音模式。最後，他們使用 TTS 合成提示/回應對。COSMIC [68] 透過要求 GPT-3.5 根據英文 TED 演講的轉錄稿生成問答對，來建構語音問答資料。他們展示了在他們提出的
語音問答資料集上訓練的模型，可以透過情境學習泛化到未見過的任務，例如語音到文字翻譯。
3) 語言模型後對齊：後對齊是精煉語言模型行為以符合人類偏好的關鍵過程，確保其輸出既安全又可靠。此階段通常被視為語言模型訓練的最後階段。它經常採用諸如人類回饋強化學習（RLHF）等技術，特別是近端策略最佳化（PPO）[151] 和直接偏好最佳化（DPO）[152] 等方法。

語音語言模型（SpeechLMs）的後對齊（Post-alignment）著重於解決語音互動流程中固有的獨特挑戰。Align-SLM [153] 指出，SpeechLMs 在接收相同提示時，經常產生不一致的語義內容。它透過使用文字語言模型（TextLM）來選擇 SpeechLMs 經由自動語音辨識（ASR）轉錄後的偏好回應，然後使用 DPO 對這些偏好進行對齊來解決這個問題。另一方面，SpeechAlign [154] 則專注於 SpeechLMs 的聲學品質。它觀察到「黃金」語音標記與語言模型生成的標記之間的差異，導致生成的語音聲學品質不佳，因為聲碼器在推論期間是從生成的標記合成語音。為了緩解這個問題，它採用各種最佳化技術，將語言模型的輸出與「黃金」標記的分佈對齊。儘管其重要性，SpeechLMs 的後對齊仍未被充分探索。後對齊的一個關鍵應用是減輕生成模型相關的安全風險。因此，未來的研究應優先識別和解決 SpeechLMs 帶來的獨特安全挑戰（參見第七節-D）。

C. 語音互動範式

前面章節涵蓋的大多數方法都遵循 SpeechLMs 的傳統生成範式，即接收預定義的輸入序列並生成完整的響應。然而，這種方法並不能反映語音互動的自然流程。例如，在對話中，一個人可能會打斷另一個人，從聽轉為說。此外，如果對方正在與其他人交談，一個人可能會選擇不回應。基於這些觀察，我們為 SpeechLMs 識別了進階語音互動技能的兩個關鍵面向：即時互動和互動期辨識。

SpeechLMs 的即時互動涉及對來自兩個或更多人的對話資料進行進階處理，並且可以透過幾個漸進階段來理解。初始階段是採用串流分詞器和聲碼器，這消除了語言模型在處理前等待完整語音編碼的需要。這種架構能夠對使用者查詢提供即時、低延遲的響應，標誌著相對於傳統互動範式的顯著改進。儘管如此，雖然這種串流方法支援基本的即時互動，但它仍然不足以捕捉自然對話中觀察到的更複雜的互動模式。下一個前沿是全雙工建模，它允許 SpeechLMs 支援
同步雙向通訊——特別是處理由使用者或模型發起的打斷的能力。它主要包含兩個功能：1) 使用者打斷，模型在對話中可以被打斷並適當地回應新的指令；2) 同步回應，使模型能夠同時處理輸入和產生輸出。實現這點需要對使用者和模型的音訊流進行聯合建模。dGSLM [4] 為雙人對話中的每個參與者使用一個獨立的 Transformer，並透過交叉注意力層捕捉說話者之間的互動。然而，大多數方法都依賴單一語言模型。NTPP [69] 採用「下一個詞元對預測」方法，使用僅解碼器 Transformer 來預測兩個通道的詞元。Moshi [9] 將使用者輸入和模型回應通道的資料串聯起來，使用 RQ-Transformer 一起處理資料。LSLM [65] 專注於使用僅解碼器 Transformer 建模單一說話者的語音，整合串流 SSL 編碼器以融合聽覺和說話通道的嵌入。

互動期間辨識 (IPR) 指的是辨識使用者是否正在與其互動的能力。語音語言模型 (SpeechLM) 應在互動期間提供回應，並在非互動期間保持靜默。IPR 對於創造自然的對話流程至關重要，讓模型避免不必要的打斷。這對於一小群使用者正在討論的情況至關重要，因為 SpeechLM 需要辨別何時加入以及何時保持靜默。此外，模型學習何時忽略使用者未對其說話時的指令也很重要。實現 IPR 的一種方法是透過語音活動偵測 (VAD) 模組。MiniCPM-o 2.6 [72] 整合了一個 VAD 模組，以確保模型僅在輸入音訊超過預設的 VAD 閾值時才回應。低於此閾值的輸入被視為噪音並被忽略。VITA [70] 採用不同的方法，透過訓練 SpeechLM 來區分查詢語音和非查詢音訊。當偵測到非查詢音訊時，模型會學習輸出序列結束詞元以終止其回應。

V. 下游應用

與傳統的語音系統（如 ASR 和 TTS）通常專注於特定任務不同，語音語言模型（SpeechLM）作為生成式基礎模型運作。它們可以透過遵循各種指令來處理多樣化的純語音、純文字和多模態任務。在本節中，我們將探討 SpeechLM 的主要下游應用。這裡討論的任務主要包括傳統的語音相關任務，以及一些 SpeechLM 獨有的任務。與僅生成包含語義資訊的文字的文字語言模型（TextLM）不同，SpeechLM 可以同時建模語義和副語言資訊，例如音高和音色，這使得它們成為更強大的模型。因此，我們將 SpeechLM 的下游應用分為三大類：語義相關應用、說話者相關應用和副語言應用。表 V 提供了每個下游任務的範例。

A. 語義相關應用

語義相關應用涵蓋了促進人機之間有意義互動的關鍵任務。這些應用要求 SpeechLM 理解輸入的語義，並生成不僅與上下文相關，而且邏輯連貫的回應。SpeechLM 的主要語義相關應用如下。

語音對話。語音對話是 SpeechLM 最自然的應用。語音對話系統旨在促進人機之間以語音形式進行自然對話。它們可以讓使用者參與互動式交流，根據對話的上下文理解並生成回應。與 TextLM 不同，SpeechLM 能夠直接以語音與人類進行對話，這是一種更自然的溝通方式。請注意，SpeechLM 不僅可以執行純語音對話，還可以執行跨模態對話，例如將文字作為輸入並以語音格式回應。

語音翻譯。語音翻譯（ST）是將一種口語轉換為另一種口語的過程。與口語對話類似，語音語言模型（SpeechLMs）可以在單模態和跨模態設定中執行語音翻譯。具體來說，語音翻譯任務的輸入和輸出可以是文字或語音格式。

自動語音辨識。自動語音辨識（ASR）使系統能夠將口語轉換為文字。ASR 的輸入是語音波形，系統輸出文字形式的轉錄。對於語音語言模型，輸入將是語音波形和指示的組合，以告知模型對給定的語音執行 ASR。

關鍵字偵測。關鍵字偵測可以被視為一種特殊類型的 ASR，其主要目標是在連續語音中識別特定單詞或短語。傳統的 ASR 系統旨在將整個口語語句轉錄為文字，而關鍵字偵測則專注於識別和提取連續語音中預定義的關鍵字或短語。關鍵字偵測的主要應用是在智慧家庭設備中建立語音啟動助理。這些設備在特定關鍵字被觸發時啟動。因此，儘管語音語言模型能夠偵測和理解的不僅僅是幾個單詞，但關鍵字偵測可以用來有效地觸發語音語言模型以回應使用者輸入。

文字轉語音合成。文字轉語音合成（TTS）使系統能夠將書面文字合成為口語。與 ASR 相反，TTS 接收文字作為輸入並輸出轉換後的語音波形。同樣，語音語言模型的輸入將是待合成文字和指令的組合，輸出則是合成語音。

意圖分類。意圖分類是一項關鍵任務，旨在識別使用者輸入語音背後的潛在意圖。人工智慧系統隨後可以根據識別出的使用者意圖執行某些動作（例如，預訂航班）。意圖分類在虛擬助理、客戶服務機器人以及互動式語音應答系統等應用中尤為重要。為了執行意圖分類，更

表 5：表五
SpeechLM 各種功能的範例。

任務	輸入	輸出
語義相關應用		
語音對話	語音（我現在好生氣！）	語音（很抱歉聽到這個消息。發生了什麼事？）
語音翻譯	語音（早安，我的朋友！）請將這段語音翻譯成法語。	語音(Bonjour mon ami !)
自動語音辨識	語音(A river is on your left side.) 請對語音執行自動語音辨識	轉錄的語音是：A river is on your left side.
關鍵字偵測	語音（你今天過得如何？（對著別人說）嘿，Siri。）請偵測給定的語音是否包含任何預先定義的關鍵字。	偵測到的關鍵字：「嘿，Siri。」
文字轉語音合成	請將以下文字合成語音：「你的左邊有一條河。」	好的。這是合成語音：語音（你的左邊有一條河。）
意圖分類	語音（你能幫我訂機票嗎？）請偵測給定語音的意圖。	偵測到意圖：「預訂航班」。
槽位填寫	語音（我想在6月5日從紐約飛往舊金山。）請根據給定的語音填寫以下槽位：出發城市、目的地城市和日期	好的。以下是已填寫的槽位。出發城市：紐約。目的地城市：舊金山。日期：6月5日。
以範例語音詞彙檢索	語音（你能檢索包含「快樂」這個詞的音訊文件嗎？）[一批音訊文件]	好的。我已找到文件2和文件4包含「快樂」這個詞。
與說話者相關的應用		
說話者辨識	以下語音來自露西：「[露西的語音]」。以下語音來自鮑伯：「鮑伯的語音」。問題：以下語音「[未知語音]」的說話者是誰？	以下語音：「[未知語音]」的說話者是鮑伯。
說話者驗證	這裡有兩段語音。語音 1：[語音 1]。語音 2：[語音 2]。它們是來自同一個說話者嗎？	不，它們不是來自同一個說話者。
說話者辨識	[會議的音訊檔案] 請提供說話者辨識結果，格式如下：[開始時間] [結束時間]：說話者 [說話者]。	好的。這是說話者辨識結果：00:00-00:15：Alice，00:16-00:40：Bob，00:41-01:00：Carol，01:01-01:20：Alice，01:21-01:45：Bob 和 Carol（語音重疊），01:46-02:00：Carol。
語音條件式語音生成	你能用機器人聲音對我說「Hello world!」嗎？	好的。語音（[用機器人聲音] Hello world!）
語音伴隨應用		
情緒辨識	語音（你在做什麼）語音傳達了什麼情緒？	聽起來說話者對某件事感到非常震驚和憤怒。
語音分離	語音（[兩段重疊的語音]）你能將這兩段語音分開嗎？	當然。這是語音分離的結果。語音1：語音（我應該選這個嗎？）。語音2：語音（那個看起來不錯。）
副語言增強生成	語音（請生成一段聽起來很開心的語音...	語音（[心情愉悅] 我的朋友剛才給了我一顆糖果！）

對於語音語言模型（SpeechLMs）來說，接收語音輸入並將結果分類為文字是更自然的做法，因為以文字形式解析和分類意圖分類結果比以語音形式更容易。

槽位填充。槽位填充是口語理解中的一項重要任務，它涉及從使用者輸入中識別和提取特定資訊片段，並將其歸入預定義的類別，例如意圖、實體和參數，這些對於完成任務至關重要。例如，槽位填充會將「我想在 6 月 5 日從紐約飛往舊金山」這句話提取成不同的槽位，例如「出發城市」（紐約）、「目的地城市」（舊金山）和「日期」（6 月 5 日）。與意圖分類類似，對於語音語言模型來說，接收語音輸入並以文字形式提取這些片段是更自然的做法。

以範例語音詞彙偵測進行查詢。另一項語音詞彙偵測任務是以範例語音詞彙偵測（QbE-STD），它允許使用者透過提供所需詞彙的範例，在較大的音訊串流中識別特定的語音詞彙或短語。與依賴預定義關鍵字列表的傳統關鍵字偵測方法不同，QbE-STD 利用基於範例查詢的靈活性，讓使用者能夠透過
音訊樣本指定其搜尋詞彙。

B. 說話者相關應用

說話者相關應用是指涉及處理與說話者身份相關資訊的任務。它可能涉及分類任務，例如根據說話者獨特的聲音特徵識別、驗證和區分個別說話者，以及生成任務，例如維持或修改給定語音的音色。雖然我們承認聲音特徵可以被視為副語言資訊，但我們認為說話者相關應用是獨特的，因為它們使語音語言模型（SpeechLMs）能夠在複雜的場景中運作，例如參與多說話者對話。在本節中，我們將探討語音語言模型常見的說話者相關應用。

說話者辨識。說話者辨識是根據一個人的聲音特徵來識別其身份的過程。它是一個將給定語音作為輸入的多類別分類任務。語音語言模型（SpeechLMs）可以透過接收輸入語音並以文字或語音格式輸出分類結果來執行此任務。此外，語音語言模型還可以隱含地識別不同的說話者。具體來說，它可以同時與多個說話者聊天，區分不同說話者的話語並適當地回應每個說話者。

說話者驗證。說話者驗證涉及判斷一對語音的說話者是否匹配。與說話者辨識（一個多類別分類過程）不同，說話者驗證是一個二元分類過程。

說話者分離。說話者分離是根據說話者的身份將音訊流分割成片段的過程。它預測每個時間戳「誰在何時說話」[155]。將說話者分離整合到語音語言模型中的一種自然方式是讓模型生成每個音訊片段的轉錄稿以及說話者的識別。

語音條件式語音生成。語音條件式語音生成涉及根據特定說話者的聲音特徵來合成語音。這可能包括語音複製和語音轉換。語音複製利用說話者的語音樣本作為參考，使模型能夠在從輸入文字生成語音時重現說話者的音色。另一方面，語音轉換則修改現有的語音訊號，使其聽起來像是由不同的說話者產生，同時保留原始內容。此外，除了提供目標聲音特徵外，語音語言模型（SpeechLMs）也應該能夠根據各種語音或文字指令調整其輸出音色。

C. 副語言應用

副語言（Paralinguistics）指的是伴隨口語的非語言溝通元素。它包含各種聲音屬性，這些屬性傳達的意義超越了實際說出的詞語。這些元素可以顯著影響訊息的解釋和理解方式。副語言的關鍵元素包括音高、音色、音量、語速、停頓等。由於副語言元素的組合方式不同，可以產生帶有不同情感的語音，因此我們也將情感相關任務納入副語言應用。

情感辨識。情感辨識任務涉及識別並將給定語音所帶有的情感歸類到預定義的類別中。與說話者辨識類似，語音語言模型（SpeechLMs）不僅能夠直接執行此任務，還能透過使用者的語音查詢隱式辨識其情感並做出相應的回應。

語音分離。語音分離是指從混合的聲音中分離出個別語音訊號的過程，例如當多個說話者同時說話時。在分離輸入語音時，語音語言模型（SpeechLMs）不僅能輸出語音中每個人的內容，還能以文字格式（即轉錄）輸出。

副語言增強生成。副語言增強生成是指指示語音語言模型（SpeechLMs）產生具有特定副語言特徵的語音的過程。使用者可以在提示中定義這些特徵，讓模型生成語音，而這些語音

表6：表六
語音語言模型（SpeechLMs）評估常用基準的摘要。i/O、 A 和 T 分別代表輸入/輸出模式、音訊和文字。

名稱	評估類型	# 任務	音訊類型	輸入/輸出
ABX [156]-[158]	表徵	1	語音	$A \rightarrow -$
sWUGGY [158]	語言學	1	語音	$A \rightarrow -$
sBLIMP [158]	語言學	1	語音	$A \rightarrow -$
sStoryCloze [51]	語言學	1	語音	$A/T \rightarrow -$
STSP [5]	副語言	1	語音	$A/T \rightarrow A/T$
MMAU [159]	下游	27	語音、聲音、音樂	$A \rightarrow T$
Audiobench [160]	下游	8	語音、聲音	$A \rightarrow T$
AIR-Bench [161]	下游	20	語音、聲音、音樂	$A \rightarrow T$
SD-Eval [162]	下游	4	語音	$A \rightarrow T$
SUPERB [163]	下游	10	語音	$A \rightarrow T$
VoxDialogue [164]	下游	12	語音、聲音、音樂	$A \rightarrow T$
Dynamic-SUPERB [163]	下游	180	語音、聲音、音樂	$A \rightarrow T$
SALMON [165]	下游	8	語音	$A \rightarrow -$
VoiceBench [166]	下游	8	語音	$A \rightarrow T$
VoxEval [167]	下游	56	語音	$A \rightarrow A$

以符合其規範。旁語言增強生成（paralinguistics-enhanced generation）的例子包括合成特定風格的語音、快速說話，甚至是唱歌。這種能力使語音語言模型（SpeechLMs）有別於文字語言模型（TextLMs），並促進了與 AI 模型之間更具吸引力和互動性的溝通形式。

六、評估

與文字語言模型（TextLM）類似，語音語言模型（SpeechLM）具備廣泛的能力，這使得比較不同的語音語言模型變得具有挑戰性。因此，從多個角度評估語音語言模型以確定其有效性至關重要。在本節中，我們將回顧評估語音語言模型常用的方法和基準（表六）。我們將這些評估方法分為自動評估和人工評估，每種方法都包含不同的評估面向。

A. 自動（客觀）評估

自動評估方法對於提供語音語言模型的快速且一致的評估至關重要。這些方法通常依賴於無需人工干預即可計算的量化指標。以下，我們概述了一些最常用的自動評估技術。

表徵評估。表徵（嵌入）是語音語言模型（和文字語言模型）中的關鍵組成部分。它指的是輸入資料（例如語音或文字）如何轉換為模型可以理解和處理的格式。有效的表徵為模型理解詞彙、語法和上下文資訊奠定了堅實的基礎，這些資訊對於生成連貫且與上下文相關的輸出至關重要。

在語音語言模型（SpeechLMs）的背景下，表徵評估著重於模型將語音特徵編碼成有意義向量的效能。GSLM [50] 使用說話者間 ABX 分數來衡量嵌入的相似性，它量化了語音類別的分離程度。具體來說，它透過比較三個聲音樣本來運作：其中兩個來自同一類別 (A)，一個來自不同類別 (B)。此測試衡量系統正確識別來自類別 A 的兩個聲音比來自 A 的一個聲音與來自 B 的一個聲音更相似的頻率。另一種評估表徵的方法是透過語音再合成 [50]。

具體來說，輸入語音會被編碼成語音標記，然後再合成回語音。接著，可以計算輸入語音和再合成語音的自動語音辨識（ASR）結果的詞錯誤率（WER）或字元錯誤率（CER）。這衡量了將輸入語音離散化為語音標記所造成的信息損失，從而評估了潛在表徵的穩健性。

語言學評估。語言學評估，包括詞彙、句法和語義評估方法，旨在評估模型生成和理解構詞、造句和有意義內容規則的能力。這些評估著重於詞彙選擇的正確性和適當性、輸出內容的語法結構，以及生成內容的連貫性和相關性。在基準資料集方面，sWUGGY [158] 在詞彙層面進行評估，判斷模型是否能區分真實詞彙與（真實、非真實）詞彙對。sBLIMP [158] 在句法層面進行評估，判斷模型是否能從（合乎語法、不合語法）句子對中識別出合乎語法的句子。Spoken StoryCloze [51] 透過評估模型從一對結局選項中選擇故事真實結局的能力，來評估語義理解。所有評估都是透過比較模型對資料對的負對數似然值來進行的。

副語言評估。與語言評估不同，副語言評估著重於伴隨語音的非語言溝通層面。有些研究選擇將副語言標記與語義標記一同使用，以增強語音語言模型（SpeechLMs）的副語言能力 [5]、[56]，因此一種方法是評估副語言標記。pGSLM [56] 測量語音標記的正確性、一致性和表達性。正確性透過計算 20 個生成樣本的語音標記與參考語音標記之間的最小平均絕對誤差（min-MAE），來評估模型生成準確語音輪廓的能力；一致性則透過提示語音標記的平均值與其生成延續語音標記之間的皮爾遜相關係數來評估；表達性則透過生成語音標記值的標準差來測量，並期望其與真實情況的變異性相符。我們注意到，相同的指標也可以應用於其他副語言標記。SPIRITLM [5] 則提出在感知層面進行測量，而非從標記層面評估。他們引入了一個語音-文本情感保留基準（STSP），要求模型生成一個能保留提示情感的文本或語音序列標記。情感分類器用於評估生成語音中的情感。值得注意的是，儘管他們僅將保留方法應用於情感，但這個概念可以推廣到其他副語言特徵，例如音色或韻律。

生成品質與多樣性。品質與多樣性是模型生成的兩個關鍵面向。通常，在不同溫度下取樣模型回應時，這些維度之間存在權衡，因此 GSLM [50] 建議使用曲線下面積（AUC）與各種溫度值。具體來說，困惑度與 VERT 的 AUC 被用於評估這些因素，其中 VERT 代表生成語音中至少重複一次的 k-gram 比率的幾何平均值。此外，ChatGPT 分數可用於評估生成語音的品質。在此過程中，生成的語音會使用最先進的 ASR 模型進行轉錄，然後發送給 ChatGPT 進行品質（和多樣性）評估。

即時互動評估。即時互動評估涉及評估語音語言模型（SpeechLMs）即時互動的能力，這對於促進串流或全雙工互動的模型至關重要。目前的研究重點是評估即時生成語音的自然度和實用性。dGSLM [4] 透過引入不同的輪流發言事件，例如語音片段（Inter-Pausal Unit, IPU）、語音內的停頓（pause）、語音間的停頓（gaps）以及重疊語音（overlap），來檢視兩位說話者之間對話的自然度。如果這些輪流發言事件的統計數據與人類對話中的統計數據非常相似，則生成的語音被認為更自然。另一種方法涉及語音延續，如果語音提示的輪流發言統計數據與隨後延續的統計數據緊密對齊，則生成的語音被認為更自然。最近，Full Duplex Bench [168] 作為一個基準，用於評估全雙工語音語言模型中各種輪流發言能力，而 Talking Turns [169] 則透過訓練一個神經網路模型來預測全雙工語音語言模型輸出的事件，從而評估輪流發言事件。此外，評估語音語言模型作為 AI 助理在即時互動中的實用性也至關重要。NTPP 模型 [69] 建議使用反思性停頓和打斷來評估實用性，其中反思性停頓評估語音語言模型在使用者說話時保持沉默的能力，而打斷則衡量語音語言模型在被打斷時停止說話的能力。

下游評估。下游評估指的是評估語音語言模型執行特定任務的能力，例如自動語音辨識（ASR）、文字轉語音（TTS）、說話者辨識等。評估可以在預訓練模型上進行，方法是在提示開頭添加少量範例，或在經過指令微調的模型上直接指示它們執行任務。有幾個基準彙編了廣泛的下游任務，以提供對語音語言模型的全面評估。SUPERB [155] 包含了各種語音理解任務。SD-Eval [162] 使用情感、年齡、環境和年齡分類任務來評估語音語言模型的副語言理解能力。SALMON [165] 測試語音語言模型生成具有一致副語言和環境特徵語音的能力。Voicebench [166] 評估語音語言模型的通用能力。Dynamic-SUPERB [163]、MMAU [159]、AirBench [161] 和 AudioBench [160] 則超越了傳統語音任務，還包含了聲音和/或音樂相關的挑戰。

儘管這些基準提供了音訊相關任務的全面覆蓋，但它們主要要求模型以文字回應，這為端到端語音互動評估製造了障礙。為了解決這個限制，VoxEval [167] 專注於基準測試語音語言模型的知識理解能力，提供問答對以全面評估。
涵蓋語音格式的綜合主題，並提供專為語音輸出量身打造的評估流程。此外，它還提出了在各種輸入音訊條件下的問題，以評估模型的穩健性，並開創了口語數學推理評估。

B. 人工（主觀）評估。

人工評估在評估語音語言模型（SpeechLM）的效能方面扮演著關鍵角色，因為語音最終是設計給人類聽覺和感知的。這類評估依賴人類判斷來評估 SpeechLM 生成輸出的品質。以下，我們概述了幾種常用的人工評估方法。

平均意見分數。平均意見分數（MOS）是語音評估領域中廣泛使用的指標，它量化了人類聽眾判斷的語音輸出感知品質。通常，一群評估者會聽取 SpeechLM 生成的一系列音訊樣本，並根據預定義的量表（通常從 1 分（品質差）到 5 分（品質優異））對每個樣本進行評分。

MOS 是透過將所有評估者對每個音訊樣本給予的分數取平均值來計算的，提供一個單一分數，反映人類感知到的整體品質。MOS 的變體側重於語音品質的不同方面，包括 MMOS、PMOS 和 SMOS [56]、[60]。它們分別評估給定語音的自然度、語調和音色相似度。

通常，評估自然度或音色相似度涉及收集人類意見。然而，由於招募參與者和收集其評估的挑戰，這個過程可能會很複雜。因此，研究人員經常轉向基於機器的評估。他們通常採用專門為這些任務訓練的神經網路模型。例如，自然度預測模型 [170] 可以評估生成輸出的自然度，而說話者識別模型可以評估音色相似度。

七、挑戰與未來方向

儘管語音語言模型（SpeechLMs）已展現出令人印象深刻的能力，但該領域的研究仍有待深入探索。在本節中，我們將探討語音語言模型研究中的挑戰、未解決的問題以及未來研究的可能方向。

A. 了解不同的組件選擇

目前關於語音語言模型（SpeechLMs）的研究涵蓋了語音分詞器、語言模型和聲碼器等關鍵組件，每個組件都提供了多樣化的選項。儘管有些研究比較了各種組件選擇——主要集中在語音分詞器上——但這些比較往往範圍和深度有限 [50], [52]。因此，在理解不同組件選擇的優缺點方面仍存在顯著的空白。因此，旨在全面比較這些選擇的研究至關重要。這類調查將產生寶貴的見解，並為開發語音語言模型時選擇更高效的組件提供指導。

B. 端到端訓練

儘管語音語言模型可以直接生成語音而無需依賴文字訊號，但有些研究將這三個組件分開訓練。這種獨立最佳化可能會阻礙模型的整體潛力。因此，值得研究是否可以採用端到端的方式進行訓練，讓梯度從聲碼器的輸出反向傳播到分詞器的輸入。透過探索這種完全端到端的方法，我們有可能使語音語言模型產生更連貫、更具上下文相關性且更高傳真度的語音輸出。

C. 即時語音生成

在語音語言模型（SpeechLM）中，實現即時語音生成至關重要，因為它能促進與人類互動的更具互動性的方式。然而，第三節所述的大多數常用方法在輸入和輸出語音生成之間仍然會產生明顯的延遲。這種延遲的發生是因為典型的聲碼器必須等待語言模型生成整個輸出詞元序列後才能運作，這使其成為推論管線中最耗時的過程。改善延遲的一種潛在解決方案是開發可串流的管線，允許語音輸入和輸出以區塊方式處理和生成。另一種選擇可能涉及語音語言模型自主生成波形音訊樣本。總體而言，即時語音生成這個領域仍未被充分探索，需要進一步研究。

D. 語音語言模型中的安全風險

安全是機器學習領域中一個高度重要的議題，尤其是在大型生成式人工智慧模型方面。儘管在文字語言模型（TextLM）的安全問題上已有廣泛研究，但語音語言模型（SpeechLM）中的安全問題尚未得到徹底調查。語音語言模型中的安全挑戰與文字語言模型相比，既有相似之處，也有獨特之處，正如 OpenAI 最近關於 GPT-4o 語音模型安全問題的報告 [171] 所強調的。因此，未來的研究探索語音語言模型中的安全漏洞並開發更安全的語音語言模型至關重要。

語音語言模型（SpeechLM）安全問題的主要考量包括但不限於毒性與隱私。毒性指的是語音語言模型所生成內容的有害性質。例如，這些模型可能會產生語義上危險的內容，例如製造爆炸物的說明。此外，它們也可能生成聲學上不適當的內容，例如色情語音 [171]，這帶來了獨特的挑戰。隱私則涉及語音輸入經語音語言模型處理後，洩露個人資訊的風險。例如，模型可能會根據語義內容或輸入的聲學特徵推斷說話者的身份。更令人擔憂的是，模型可能會根據不足的（例如聲學）資訊，對說話者做出帶有偏見的推斷，例如他們的種族或宗教信仰 [171]。

E. 稀有語言的表現

語音語言模型直接對語音資料進行建模，這使得它們比文字語言模型更能有效處理「低資源」語言。「低資源」語言是指缺乏大量文字資料的語言，這使得文字語言模型難以有效建模。相較之下，語音語言模型透過對這些「低資源」語言的語音資料進行建模，提供了更好的解決方案，因為這些語言通常擁有比文字更多的音訊資料 [50]。因此，未來的研究可以專注於訓練「低資源」語言或方言的語音語言模型，以擴展其能力。

VIII. 結論

本調查全面概述了語音語言模型（SpeechLMs）的最新進展。我們首先探討了將自動語音辨識（ASR）、大型語言模型（LLMs）和文字轉語音（TTS）系統結合用於語音互動的簡單框架所存在的限制。接著，我們強調了 SpeechLMs 所提供的關鍵優勢。隨後，我們探討了 SpeechLMs 的架構，詳細介紹了所涉及的組件及其訓練方法。我們還討論了它們在各種下游應用中的能力以及不同的評估方法。最後，我們指出了開發 SpeechLMs 的主要挑戰，並概述了未來研究的潛在方向。我們希望這項調查能闡明該領域，並協助研究社群建立更強大的語音語言模型。

參考文獻

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
[2] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.
[3] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin et al., "Opt: Open pre-trained transformer language models," arXiv preprint arXiv:2205.01068, 2022.
[4] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed et al., 「生成式口語對話語言模型」(Generative spoken dialogue language modeling)，收錄於《計算語言學學會彙刊》(Transactions of the Association for Computational Linguistics)，第 11 卷，頁 250-266，2023 年。
[5] T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, P.-A. Duquenne, R. Algayres, R. Mavlyutov, I. Gat et al., 「Spirit-lm：交錯式口語與書面語言模型」(Spirit-lm: Interleaved spoken and written language model)，arXiv 預印本 arXiv:2402.05755，2024 年。
[6] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu et al., 「Audiogpt：理解與生成語音、音樂、聲音和說話頭像」(Audiogpt: Understanding and generating speech, music, sound, and talking head)，收錄於《AAAI 人工智慧會議論文集》(Proceedings of the AAAI Conference on Artificial Intelligence)，第 38 卷，第 21 期，2024 年，頁 23 802-23804。
[7] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, 「Hugginggpt：利用 chatgpt 及其在 hugging face 中的朋友解決 AI 任務」(Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face)，收錄於《神經資訊處理系統進展》(Advances in Neural Information Processing Systems)，第 36 卷，2024 年。
[8] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, 「SpeechGPT: 賦予大型語言模型內在的跨模態對話能力，」收錄於計算語言學學會論文集：EMNLP 2023，H. Bouamor, J. Pino, and K. Bali 編。新加坡：計算語言學學會，2023 年 12 月，頁 15757-15773。[線上]。網址：https://aclanthology.org/2023.findings-emnlp。1055
[9] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, 「Moshi: 一種用於即時對話的語音-文本基礎模型，」Kyutai，技術報告，2024 年 9 月。[線上]。網址：http://kyutai.org/Moshi.pdf
[10] Z. Xie and C. Wu, 「Mini-Omni: 語言模型在串流中思考時能聽、能說，」arXiv 預印本 arXiv:2408.16725, 2024。
[11] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, 「Llamaomni: 大型語言模型的無縫語音互動，」arXiv 預印本 arXiv:2409.06666, 2024。
[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, 「透過大規模弱監督實現穩健語音辨識」(Robust speech recognition via large-scale weak supervision)，收錄於《國際機器學習會議》(International Conference on Machine Learning)。PMLR，2023 年，頁 28492-28518。
[13] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, 「用於聯合自動語音辨識和多語言語音翻譯的雙解碼器 Transformer」(Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation)，收錄於 D. Scott, N. Bel, and C. Zong 編，《第 28 屆國際計算語言學會議論文集》(Proceedings of the 28th International Conference on Computational Linguistics)。西班牙巴塞隆納（線上）：國際計算語言學委員會，2020 年 12 月，頁 3520-3533。[線上]。網址：https://aclanthology.org/2020.coling-main。314
[14] Y. Fathullah, C. Wu, E. Lakomkin, K. Li, J. Jia, Y. Shangguan, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer, 「AudioChatLLaMA：邁向 LLMs 的通用語音能力」(Audiochatllama: Towards general-purpose speech abilities for LLMs)，收錄於《北美計算語言學協會 2024 年會議論文集：人類語言技術》(Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies) (第 1 卷：長篇論文)，2024 年，頁 5522-5532。
[15] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, 「SALMONN：邁向大型語言模型的通用聽覺能力」(SALMONN: Towards Generic Hearing Abilities for Large Language Models)，收錄於《第 12 屆國際學習表徵會議》(Proceedings of the 12th International Conference on Learning Representations (ICLR))，2024 年。[線上]。網址：https://openreview.net/forum?id=14rn7HpKVk
[16] L. Qin, T. Xie, W. Che, and T. Liu, 「A survey on spoken language understanding: Recent advances and new frontiers」, 發表於第三十屆國際人工智慧聯合會議論文集，IJCAI-21，Z.-H. Zhou 編。國際人工智慧聯合會議組織，2021 年 8 月，第 4577-4584 頁，調查軌。[線上]。網址：https://doi.org/10.24963/ijcai.2021/622

[17] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, 「Audio self-supervised learning: A survey」, Patterns, 第 3 卷，第 12 期，2022 年。
[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe et al., 「Self-supervised speech representation learning: A review」, IEEE Journal of Selected Topics in Signal Processing, 第 16 卷，第 6 期，第 1179-1210 頁，2022 年。
[19] G. Chrupała, 「Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques」, Journal of Artificial Intelligence Research, 第 73 卷，第 673-707 頁，2022 年。
[20] C. Sheng, G. Kuang, L. Bai, C. Hou, Y. Guo, X. Xu, M. Pietikäinen, and L. Liu, "Deep learning for visual speech analysis: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 9, p. 6001-6022, Mar. 2024. [Online]. Available: https://doi.org/10.1109/TPAMI.2024. 3376710
[21] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.
[22] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," arXiv preprint arXiv:2402.06196, 2024.
[23] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," National Science Review, p. nwae403, 11 2024. [Online]. Available: https://doi.org/10.1093/nsr/ nwae403
[24] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 2247-2256.
[25] T. Bai, H. Liang, B. Wan, Y. Xu, X. Li, S. Li, L. Yang, B. Li, Y. Wang, B. Cui et al., "A survey of multimodal large language model from a data-centric perspective," arXiv preprint arXiv:2405.16640, 2024.
[26] S. Latif, M. Shoukat, F. Shamshad, M. Usama, Y. Ren, H. Cuayáhuitl, W. Wang, X. Zhang, R. Togneri, E. Cambria et al., "Sparks of large audio models: A survey and outlook," arXiv preprint arXiv:2308.12792, 2023.
[27] J. Peng, Y. Wang, Y. Xi, X. Li, and K. Yu, "A survey on speech large language models," arXiv preprint arXiv:2410.18908, 2024.
[28] S. Ji, Y. Chen, M. Fang, J. Zuo, J. Lu, H. Wang, Z. Jiang, L. Zhou, S. Liu, X. Cheng et al., 「Wavchat: 語音對話模型綜述，」 arXiv 預印本 arXiv:2411.13577, 2024。
[29] A. v. d. Oord, Y. Li, and O. Vinyals, 「使用對比預測編碼的表徵學習，」 arXiv 預印本 arXiv:1807.03748, 2018。
[30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, 「wav2vec 2.0: 語音表徵自監督學習框架，」 Advances in neural information processing systems, 第 33 卷，第 12449-12460 頁，2020 年。
[31] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020. [Online]. Available: https://openreview.net/forum?id=rylwJxrYDS
[32] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2021, pp. 244-250.
[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 3451-3460, 2021.
[34] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," in Proceedings of the 12 th International Conference on Learning Representations (ICLR), 2024. [Online]. Available: https://openreview. net/forum?id=AF9Q8Vip84
[35] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, 「Soundstream: 一種端到端神經音訊編解碼器，」 IEEE/ACM 音訊、語音與語言處理學報，第 30 卷，第 495-507 頁，2021 年。
[36] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang et al., 「Google USM：將自動語音辨識擴展至 100 種以上語言，」 arXiv 預印本 arXiv:2303.01037，2023 年。
[37] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., 「WavLM：用於全堆疊語音處理的大規模自我監督預訓練，」 IEEE 訊號處理精選主題期刊，第 16 卷，第 6 期，第 1505-1518 頁，2022 年。
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, 「Attention is All You Need，」 收錄於《神經資訊處理系統進展》，I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett 編，第 30 卷。Curran Associates, Inc.，2017 年。[線上]。網址：https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al., 「Llama: 開放且高效的基礎語言模型，」 arXiv preprint arXiv:2302.13971, 2023。
[40] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., 「Llama 2: 開放式基礎與微調聊天模型，」 arXiv preprint arXiv:2307.09288, 2023。
[41] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang et al., 「Qwen2 技術報告，」 arXiv preprint arXiv:2407.10671, 2024。
[42] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao et al., 「Chatglm: 從 glm-130b 到 glm-4 全工具的大型語言模型家族，」 arXiv preprint arXiv:2406.12793, 2024。
[43] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, 「Mixtral of Experts」，2024 年 1 月。
[44] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, 「Wavenet: A generative model for raw audio」，收錄於《Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)》，2016 年，第 125 頁。
[45] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan et al., 「Natural tts

synthesis by conditioning wavenet on mel spectrogram predictions」，收錄於《2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)》。IEEE，2018 年，第 4779-4783 頁。
[46] R. Prenger, R. Valle, and B. Catanzaro, 「Waveglow: A flow-based generative network for speech synthesis」，收錄於《ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)》。IEEE，2019 年，第 3617-3621 頁。
[47] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems, vol. 33, pp. 17022-17033, 2020.
[48] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in Proc. Interspeech 2021, 2021, pp. 3615-3619.
[49] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," Transactions on Machine Learning Research, 2023.
[50] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed et al., "On generative spoken language modeling from raw audio," Transactions of the Association for Computational Linguistics, vol. 9, pp. 1336-1354, 2021.
[51] M. Hassid, T. Remez, T. A. Nguyen, I. Gat, A. Conneau, F. Kreuk, J. Copet, A. Defossez, G. Synnaeve, E. Dupoux et al., "Textually pretrained speech language models," Advances in Neural Information Processing Systems, vol. 36, 2024.
[52] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov et al., "Audiopalm: A large language model that can speak and listen," arXiv preprint arXiv:2306.12925, 2023.
[53] Q. Zhang, L. Cheng, C. Deng, Q. Chen, W. Wang, S. Zheng, J. Liu, H. Yu, and C. Tan, "Omniflatten: An end-to-end gpt model for seamless voice conversation," arXiv preprint arXiv:2410.17799, 2024.
[54] W. Chen, Z. Ma, R. Yan, Y. Liang, X. Li, R. Xu, Z. Niu, Y. Zhu, Y. Yang, Z. Liu et al., "Slam-omni: Timbre-controllable voice interaction system with single-stage training," arXiv preprint arXiv:2412.15649, 2024.
[55] A. Zeng, Z. Du, M. Liu, K. Wang, S. Jiang, L. Zhao, Y. Dong, and J. Tang, 「Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot」, arXiv preprint arXiv:2412.02612, 2024.
[56] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. A. Nguyen, M. Riviere, A. Mohamed, E. Dupoux, and W.-N. Hsu, 「Text-free prosody-aware generative spoken language modeling」, 載於《Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)》, S. Muresan, P. Nakov, and A. Villavicencio 編. 愛爾蘭都柏林: Association for Computational Linguistics, 2022 年 5 月, 頁 8666-8681. [線上]. 可取用: https://aclanthology.org/2022.acl-long. 593
[57] T. Wang, L. Zhou, Z. Zhang, Y. Wu, S. Liu, Y. Gaur, Z. Chen, J. Li, and F. Wei, 「Viola: Conditional language models for speech recognition, synthesis, and translation」, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
[58] J. Li, D. Wang, X. Wang, Y. Qian, L. Zhou, S. Liu, M. Yousefi, C. Li, C.-H. Tsai, Z. Xiao et al., 「Investigating neural audio codecs for speech language model-based speech generation」, arXiv preprint arXiv:2409.04016, 2024.
[59] Z. Meng, Q. Wang, W. Cui, Y. Zhang, B. Wu, I. King, L. Chen, and P. Zhao, "Parrot: Autoregressive spoken dialogue language modeling with decoder-only transformers," in Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation, 2024.
[60] D. Zhang, X. Zhang, J. Zhan, S. Li, Y. Zhou, and X. Qiu, "Speechgptgen: Scaling chain-of-information speech generation," arXiv preprint arXiv:2401.13527, 2024.
[61] E. Nachmani, A. Levkovitch, R. Hirsch, J. Salazar, C. Asawaroengchai, S. Mariooryad, E. Rivlin, R. Skerry-Ryan, and M. T. Ramanovich, "Spoken question answering and speech continuation using spectrogram-powered 11m," in Proceedings of the 12th International Conference on Learning Representations (ICLR), 2024. [Online]. Available: https://openreview.net/forum?id=izrOLJov5y
[62] R. Algayres, Y. Adi, T. Nguyen, J. Copet, G. Synnaeve, B. Sagot, and E. Dupoux, "Generative spoken language model based on continuous word-sized audio tokens," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3008-3028. [Online]. Available: https://aclanthology.org/2023.emnlp-main. 182
[63] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng, C. Zhou, Z. Yan, and S. Zhang, "LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT," https://arxiv.org/abs/2310.04673v4, Oct. 2023.
[64] J.-C. Chou, C.-M. Chien, W.-N. Hsu, K. Livescu, A. Babu, A. Conneau, A. Baevski, and M. Auli, "Toward joint language modeling for speech units and text," in Findings of the Association for Computational Linguistics: EMNLP 2023, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6582-6593. [Online]. Available: https://aclanthology.org/2023.findings-emnlp. 438
[65] Z. Ma, Y. Song, C. Du, J. Cong, Z. Chen, Y. Wang, Y. Wang, and X. Chen, "Language model can listen while speaking," arXiv preprint arXiv:2408.02622, 2024.
[66] Z. Xie and C. Wu, "Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities," arXiv preprint arXiv:2410.11190, 2024.
[67] X. Wang, Y. Li, C. Fu, L. Xie, K. Li, X. Sun, and L. Ma, 「Freezeomni: 一種具有凍結 1LLM 的智慧低延遲語音對語音對話模型」，arXiv 預印本 arXiv:2411.00774，2024。
[68] J. Pan, J. Wu, Y. Gaur, S. Sivasankaran, Z. Chen, S. Liu, and J. Li, 「Cosmic: 語音情境學習的資料高效指令微調」，arXiv 預印本 arXiv:2311.02248，2023。
[69] Q. Wang, Z. Meng, W. Cui, Y. Zhang, P. Wu, B. Wu, I. King, L. Chen, and P. Zhao, 「NTPP: 透過下一個詞元對預測實現雙通道口語對話的生成式語音語言建模」，arXiv 預印本 arXiv:2506.00975，2025。
[70] C. Fu, H. Lin, Z. Long, Y. Shen, M. Zhao, Y. Zhang, X. Wang, D. Yin, L. Ma, X. Zheng et al., 「VITA: 邁向開源互動式全方位多模態 LLM」，arXiv 預印本 arXiv:2408.05211，2024。
[71] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, "Salmonn-omni: 一種用於全雙工語音理解與生成的無編解碼器 LLM，" arXiv 預印本 arXiv:2411.18138, 2024。
[72] OpenBMB, "Minicpm-o 2.6: 一款適用於手機視覺、語音和多模態直播的 GPT-4.0 級 MLLM，" https://huggingface.co/openbmb/MiniCPM-o-2_6-int4, 2024。

[73] B. Liao, Y. Xu, J. Ou, K. Yang, W. Jian, P. Wan, and D. Zhang, “Flexduo: 一種可插拔系統，用於在語音對話系統中實現全雙工功能，” arXiv 預印本 arXiv:2502.13472, 2025。
[74] H. Jegou, M. Douze, and C. Schmid, “用於最近鄰搜尋的乘積量化，” IEEE transactions on pattern analysis and machine intelligence, 卷 33, 期 1, 頁 117-128, 2010。
[75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 「BERT: Pretraining of deep bidirectional transformers for language understanding，」收錄於《Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)》，J. Burstein, C. Doran, and T. Solorio 編。明尼蘇達州明尼亞波利斯：Association for Computational Linguistics，2019 年 6 月，頁 4171-4186。[線上]。網址：https://aclanthology.org/N19-1423
[76] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. J. Moreno, A. Bapna, and H. Zen, 「Maestro: Matched speech text representations through modality matching，」收錄於《Proceedings of Interspeech》，2022 年，頁 4093-4097。
[77] A. Van Den Oord, O. Vinyals et al., 「Neural discrete representation learning，」《Advances in neural information processing systems》，第 30 卷，2017 年。
[78] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang et al., 「Kimi-audio technical report，」arXiv preprint arXiv:2504.18425，2025 年。
[79] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu et al., 「Qwen2.5 技術報告」，arXiv 預印本 arXiv:2412.15115，2024 年，第 2 版，2025 年 1 月 3 日修訂。[線上]。網址：https://arxiv.org/abs/2412.15115
[80] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, 「Bigvgan: 一種具備大規模訓練的通用神經聲碼器」，收錄於《第 11 屆國際學習表徵會議 (ICLR) 論文集》，2023 年。[線上]。網址：https://openreview.net/forum?id= iTtGCMDEzS_
[81] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., 「Qwen2.5-omni 技術報告」，arXiv 預印本 arXiv:2503.20215，2025 年。
[82] Q. Chen, Y. Chen, Y. Chen, M. Chen, Y. Chen, C. Deng, Z. Du, R. Gao, C. Gao, Z. Gao et al., 「Minmo: 一種用於無縫語音互動的多模態大型語言模型」，arXiv 預印本 arXiv:2501.06282，2025 年。
[83] T. SpeechTeam, “FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs,” arXiv preprint arXiv:2407.04051, 2024.
[84] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang et al., “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” arXiv preprint arXiv:2412.10117, 2024.
[85] Z. Zhong, C. Wang, Y. Liu, S. Yang, L. Tang, Y. Zhang, J. Li, T. Qu, Y. Li, Y. Chen et al., “Lyra: An efficient and speech-centric framework for omni-cognition,” arXiv preprint arXiv:2412.09501, 2024.
[86] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” arXiv preprint arXiv:2409.12191, 2024. [Online]. Available: https://arxiv.org/abs/2409.12191
[87] Z. Yuan, Y. Liu, S. Liu, and S. Zhao, “Continuous speech tokens makes llms robust multi-modality learners,” arXiv preprint arXiv:2412.04917, 2024.
[88] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma et al., “Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” arXiv preprint arXiv:2407.05407, 2024.
[89] B. Veluri, B. N. Peloquin, B. Yu, H. Gong, and S. Gollakota, “Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents,” in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 21 390-21402. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1192/
[90] K. Chen, Y. Gou, R. Huang, Z. Liu, D. Tan, J. Xu, C. Wang, Y. Zhu, Y. Zeng, K. Yang et al., “Emova: Empowering language models to see, hear and speak with vivid emotions,” arXiv preprint arXiv:2409.18042, 2024.
[91] W. Huang, Z. Zhang, Y. T. Yeung, X. Jiang, and Q. Liu, “Spiral: Selfsupervised perturbation-invariant representation learning for speech pre-training,” arXiv preprint arXiv:2201.10207, 2022.
[92] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in International Conference on Machine Learning. PMLR, 2021, pp. 5530-5540.
[93] Y. Ren, T. Wang, J. Yi, L. Xu, J. Tao, C. Y. Zhang, and J. Zhou, “Fewertoken neural speech codec with time-invariant codes,” in ICASSP 20242024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 12737-12741.
[94] X. Zhang, X. Lyu, Z. Du, Q. Chen, D. Zhang, H. Hu, C. Tan, T. Zhao, Y. Wang, B. Zhang et al., “Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities,” arXiv preprint arXiv:2410.08035, 2024.
[95] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in First Conference on Language Modeling, 2024. [Online]. Available: https://openreview.net/forum?id= tEYskw1VY2
[96] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, “VoiceCraft: Zero-shot speech editing and text-to-speech in the wild,” in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12442-12462. [Online]. Available: https://aclanthology.org/2024.acl-long.673/
[97] A. Zeng, Z. Du, M. Liu, L. Zhang, S. Jiang, Y. Dong, and J. Tang, “Scaling speech-text pre-training with synthetic interleaved data,” arXiv preprint arXiv:2411.17607, 2024.
[98] Y. Zhu, D. Su, L. He, L. Xu, and D. Yu, “Generative pre-trained speech language model with efficient hierarchical transformer,” in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1764-1775. [Online]. Available: https://aclanthology.org/2024.acl-long. 97
[99] C. Du, Y. Guo, F. Shen, Z. Liu, Z. Liang, X. Chen, S. Wang, H. Zhang, and K. Yu, “Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, 2024, pp. 1792417932.

[100] K. Mitsui, K. Mitsuda, T. Wakatsuki, Y. Hono, and K. Sawada, “Pslm: Parallel generation of text and speech with llms for low-latency spoken dialogue systems,” arXiv preprint arXiv:2406.12428, 2024.
[101] K. Sawada, T. Zhao, M. Shing, K. Mitsui, A. Kaga, Y. Hono, T. Wakatsuki, and K. Mitsuda, “Release of pre-trained models for the japanese language,” arXiv preprint arXiv:2404.01657, 2024.
[102] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, “Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks,” in ICASSP 2024-

2024 IEEE 國際聲學、語音與訊號處理會議 (ICASSP)。IEEE，2024 年，頁 13326-13330。
[103] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, and J. Mahadeokar, 「Voicebox：大規模文字引導多語通用語音生成，」神經資訊處理系統進展，第 36 卷，2024 年。
[104] S. J. Park, C. W. Kim, H. Rha, M. Kim, J. Hong, J. H. Yeo, and Y. M. Ro, 「Let’s go real talk：面對面交談的口語對話模型，」arXiv 預印本 arXiv:2406.07867，2024 年。
[105] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, 「透過遮罩多模態群集預測學習視聽語音表徵，」arXiv 預印本 arXiv:2201.02184，2022 年。
[106] H. Kim, S. Seo, K. Jeong, O. Kwon, S. Kim, J. Kim, J. Lee, E. Song, M. Oh, J.-W. Ha et al., 「Paralinguistics-aware speech-empowered large language models for natural conversation，」 arXiv preprint arXiv:2402.05706, 2024。
[107] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino et al., 「Xls-r: Self-supervised cross-lingual speech representation learning at scale，」 arXiv preprint arXiv:2111.09296, 2021。
[108] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar et al., 「Voicebox: Text-guided multilingual universal speech generation at scale，」 Advances in neural information processing systems, vol. 36, pp. 14005-14034, 2023。
[109] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, 「San-m: Memory equipped self-attention for end-to-end speech recognition，」 arXiv preprint arXiv:2006.01713, 2020。
[110] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, “Hiftnet: A fast highquality neural vocoder with harmonic-plus-noise filter and inverse short time fourier transform,” arXiv preprint arXiv:2309.09493, 2023.
[111] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li et al., “Anygpt: Unified multimodal 1 lm with discrete sequence modeling,” arXiv preprint arXiv:2402.12226, 2024.
[112] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang et al., “Qwen technical report,” arXiv preprint arXiv:2309.16609, 2023.
[113] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen et al., “Palm 2 technical report,” arXiv preprint arXiv:2305.10403, 2023.
[114] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, “Wavefit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration,” in 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023, pp. 884-891.
[115] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi et al., “Audiolm: a language modeling approach to audio generation,” IEEE/ACM transactions on audio, speech, and language processing, vol. 31, pp. 2523-2533, 2023.
[116] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, and X. Wu, “Uniaudio: An audio foundation model toward universal audio generation,” arXiv preprint arXiv:2310.00704, 2023.
[117] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, “Hifi-codec: Group-residual vector quantization for high fidelity audio codec,” arXiv preprint arXiv:2305.02765, 2023.
[118] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, “High-fidelity audio compression with improved RVQGAN,” in Thirtyseventh Conference on Neural Information Processing Systems, 2023. [Online]. Available: https://openreview.net/forum?id=qjn11QUnFA
[119] R. Algayres, A. Nabli, B. Sagot, and E. Dupoux, “Speech sequence embeddings using nearest neighbors contrastive learning,” in Interspeech 2022, 2022, pp. 2123-2127.
[120] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth et al., “Gemini: a family of highly capable multimodal models,” arXiv preprint arXiv:2312.11805, 2023.
[121] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao et al., “Seed-tts: A family of high-quality versatile speech generation models,” arXiv preprint arXiv:2406.02430, 2024.
[122] J. Betker, “Better speech synthesis through scaling,” arXiv preprint arXiv:2305.07243, 2023.
[123] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” Advances in neural information processing systems, vol. 32, 2019.
[124] J.-H. Kim, S.-H. Lee, J.-H. Lee, and S.-W. Lee, “Fre-gan: Adversarial frequency-consistent audio synthesis,” in Proceedings of Interspeech, 2021, pp. 2197-2201.
[125] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 2, pp. 236-243, 1984.
[126] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 18771884, 2016.
[127] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, “Refined wavenet vocoder for variational autoencoder based voice conversion,” in 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019, pp. 1-5.
[128] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” in Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J
[129] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “Wavegrad: Estimating gradients for waveform generation,” in Proceedings of the 9th International Conference on Learning Representations (ICLR), 2021. [Online]. Available: https://openreview. net/forum?id=NsMLjcFaO8O
[130] S.-g. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu, 「Priorgrad: Improving

conditional denoising diffusion models with data-dependent adaptive prior，」收錄於第十屆國際學習表徵會議（ICLR）論文集，2022 年。[線上]。網址：https: //openreview.net/forum?id=_BNiN4IjC5
[131] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 「Librispeech: an asr corpus based on public domain audio books，」收錄於 2015 年 IEEE 國際聲學、語音與訊號處理會議（ICASSP）。IEEE，2015 年，頁 5206-5210。
[132] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, 「Mls: A large-scale multilingual dataset for speech research，」收錄於 Interspeech 2020，2020 年，頁 2757-2761。
[133] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen et al., 「Libri-light: A benchmark for asr with limited or no supervision，」收錄於 ICASSP 20202020 IEEE 國際聲學、語音與訊號處理會議（ICASSP）。IEEE，2020 年，頁 7669-7673。
[134] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," in Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS), 2021.
[135] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A largescale multilingual speech corpus for representation learning, semisupervised learning and interpretation," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 993-1003. [Online]. Available: https: //aclanthology.org/2021.acl-long. 80
[136] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 6. International Speech Communication Association, 2021, pp. 4376-4380.
[137] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in Proceedings of the Twelfth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4218-4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520
[138] C. Veaux, J. Yamagishi, and S. King, 「The voice bank corpus: Design, collection and data analysis of a large regional accent speech database，」收錄於 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)。IEEE，2013，頁 1-4。
[139] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng et al., 「Wenetspeech: A 10000+ 小時多領域中文語料庫用於語音辨識，」收錄於 ICASSP 2022-2022 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP)。IEEE，2022，頁 6182-6186。
[140] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, 「Libritts: A corpus derived from librispeech for text-to-speech，」收錄於 Proc. Interspeech 2019，2019，頁 1526-1530。[線上]。網址：http://www.openslr.org/60/
[141] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," in Proc. Interspeech 2021, 2021, pp. 22472251.
[142] Y. Jia, M. Tadmor Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6691-6703. [Online]. Available: https://aclanthology.org/2022.lrec-1.720
[143] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612, 2017.
[144] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in Interspeech 2018, 2018, pp. 1086-1090.
[145] A. Clifton, A. Pappu, S. Reddy, Y. Yu, J. Karlgren, B. Carterette, and R. Jones, 「The spotify podcast dataset，」 arXiv preprint arXiv:2004.04270, 2020。
[146] C. Cieri, D. Miller, and K. Walker, 「The fisher corpus: A resource for the next generations of speech-to-text。」 in *LREC*, vol. 4, 2004, pp. 69-71。
[147] T. A. Nguyen, W.-N. Hsu, A. d'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid et al., 「Expresso: A benchmark and analysis of discrete expressive speech resynthesis，」 in Proceedings of INTERSPEECH 2023, 2023, pp. 4823-4827。
[148] P.-A. Duquenne, K. Heffernan, A. Mourachko, B. Sagot, and H. Schwenk, 「Sonar expressive: Zero-shot expressive speech-to-speech translation，」 Meta AI Research, 2023。
[149] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, and S. Gehrmann, 「Palm: Scaling language modeling with pathways」, Journal of Machine Learning Research, vol. 24, no. 240, pp. 1-113, 2023.
[150] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen et al., 「Palm 2 technical report」, arXiv preprint arXiv:2305.10403, 2023.
[151] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, 「Proximal policy optimization algorithms」, arXiv preprint arXiv:1707.06347, 2017.
[152] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, 「Direct preference optimization: Your language model is secretly a reward model」, 發表於 Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.
[153] G.-T. Lin, P. G. Shivakumar, A. Gourav, Y. Gu, A. Gandhe, H.y. Lee, and I. Bulyko, "Align-slm: Textless spoken language models with reinforcement learning from ai feedback," arXiv preprint arXiv:2411.01834, 2024.

[154] D. Zhang, Z. Li, S. Li, X. Zhang, P. Wang, Y. Zhou, and X. Qiu, “Speechalign: Aligning speech generation to human preferences,” in Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), December 2024, poster presentation on December 11, 2024, from 11 a.m. to 2 p.m. PST.
[155] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin et al., “Superb: Speech processing universal performance benchmark,” in Proc. Interspeech, 2021, pp. 1194-1198.
[156] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, “The zero resource speech challenge 2015: Proposed approaches and results,” Procedia Computer Science, vol. 81, pp. 67-72, 2016, sLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S187705091630045X
[157] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux, 「The zero resource speech challenge 2019: Tts without t，」收錄於 Interspeech 2019，2019 年，頁 1088-1092。
[158] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, 「The zero resource speech
benchmark 2021: Metrics and baselines for unsupervised spoken language modeling，」收錄於 Proceedings of the Workshop on Self-Supervised Learning for Speech and Audio Processing. NeurIPS，2020 年。
[159] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, 「Mmau: A massive multitask audio understanding and reasoning benchmark，」arXiv 預印本 arXiv:2410.19168，2024 年。
[160] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, 「Audiobench: A universal benchmark for audio large language models,」 arXiv preprint arXiv:2406.16020, 2024.
[161] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou et al., 「Air-bench: Benchmarking large audio-language models via generative comprehension,」 arXiv preprint arXiv:2402.07729, 2024.
[162] J. Ao, Y. Wang, X. Tian, D. Chen, J. Zhang, L. Lu, Y. Wang, H. Li, and Z. Wu, 「Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words,」 arXiv preprint arXiv:2406.13340, 2024.
[163] C.-y. Huang, W.-C. Chen, S.-w. Yang, A. T. Liu, C.-A. Li, Y.-X. Lin, W.-C. Tseng, A. Diwan, Y.-J. Shih, J. Shi et al., 「Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks,」 arXiv preprint arXiv:2411.05361, 2024.
[164] X. Cheng, R. Hu, X. Yang, J. Lu, D. Fu, Z. Wang, S. Ji, R. Huang, B. Zhang, T. Jin et al., “Voxdialogue: Can spoken dialogue systems understand information beyond words?” in The Thirteenth International Conference on Learning Representations, 2025.
[165] G. Maimon, A. Roth, and Y. Adi, “A suite for acoustic language model evaluation,” arXiv preprint arXiv:2409.07437, 2024.
[166] Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, “Voicebench: Benchmarking llm-based voice assistants,” CoRR, vol. abs/2410.17196, 2024. [Online]. Available: https://doi.org/10.48550/arXiv. 2410.17196
[167] W. Cui, X. Jiao, Z. Meng, and I. King, “Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models,” arXiv preprint arXiv:2501.04962, 2025.
[168] G.-T. Lin, J. Lian, T. Li, Q. Wang, G. Anumanchipalli, A. H. Liu, and H.-y. Lee, “Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities,” arXiv preprint arXiv:2503.04721, 2025.
[169] S. Arora, Z. Lu, C.-C. Chiu, R. Pang, and S. Watanabe, “Talking turns: Benchmarking audio foundation models on turn-taking dynamics,” arXiv preprint arXiv:2503.01174, 2025.
[170] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in Proceedings of Interspeech 2021, 2021, pp. 2127-2131.
[171] OpenAI, “Gpt-4o system card,” 2024, online; Accessed on 6-September-2024. [Online]. Available: https://openai.com/index/gpt-4o-system-card/

溫謙·崔 (wenqian.cui@link.cuhk.edu.hk)、余典之和金國慶 (king@cse.cuhk.edu.hk) 任職於香港中文大學計算機科學與工程學系，中國香港。焦曉琪和張廣彥任職於中國深圳光速工作室。孟子喬任職於新加坡國立大學，新加坡。王啟超任職於中國深圳騰訊。郭怡文為獨立研究員。

[1] Github: https://github.com/dreamtheater123/Awesome-SpeechLM-Survey

[2] 為求簡潔，我們將這兩種表示方式統一稱為「符元」(tokens)