{% include header.html %}

{% include ai-share.html %}

🎙️ 語音識別 / 合成平台價格比較

~2025/04

名稱	功能	網址	說明
Whisper (開源)	語音識別、翻譯	每分鐘150字 × 10分鐘 = 1500字	
Fish Audio	語音識別、語音合成	TTS：英文 $0.0225，中文 $0.0675；ASR：30分鐘 = $0.18	
Deepgram	語音識別	TTS：英文 $0.02025，中文 $0.06075；ASR：30分鐘 = $0.147	
Microsoft Azure	語音合成	TTS：英文 $0.036，中文 $0.108；ASR：即時轉錄 $1/小時，超額 $0.8/小時	
Amazon Polly	語音合成	TTS：英文 $0.024，中文 $0.072	
Google WaveNet	語音合成	TTS：英文 $0.024，中文 $0.072	
Google Vertex AI	大型語言模型	Gemini/Claude 定價頁	
Google Cloud VM	虛擬機器	VM 執行個體定價頁面	

語音處理 (Speech Processing)

Speech-Processing

- 2025-05-14：ten-turn-detection；ten-vad
- 2025-01-19：小米語音首席科學家Daniel Povey：語音辨識捲完了，下一個機會在哪裡？
- ASR/TTS 開發避坑指南：語音辨識與合成的常見挑戰與對策；探討 ASR 和 TTS 技術應用中的問題，強調數據質量的重要性
- 那些語音處理踩的坑；分享語音處理領域的實務經驗，強調資料品質對模型效果的影響
- 音視頻開發基礎入門｜聲音的採集與量化、音頻數字信號質量、音頻碼率
- 一文總覽萬字語音合成系列基礎與論文總結
- Mozilla Common Voice Datasets - zhTW
- 語音識別資料匯總：常見庫和特徵對比
- 語音合成,語音辨識常見資料集

中文語音識別 (Chinese Speech Recognition)

Chinese-Speech-Recognition

https://www.twman.org/AI/ASR

通過語音信號處理和模式識別讓機器自動識別和理解人類的口述

- 2025-07-16：Voxtral Small 1.0 (24B) - 2507；Voxtral-Mini-3B-250；Mistral首個開源語音模型來了！全面碾壓Whisper，多項測試超越GPT-4o mini
- 2025-07-02：OpusLM：全開源！ CMU 發布OpusLM：統一語音辨識、合成、文字理解的大模型
- 2025-06-06：speakr；開源的轉錄音訊記錄工具，更夠設定音訊轉錄語言和AI 生成內容
- 2025-05-06：VITA-Audio；VITA-Audio：快速交錯跨模態令牌生成，用於高效的大型語音語言模型
- 2025-04-28：FireRedASR：AI語音助理語音轉文字FireRedASR轉API；如何使用
- 2025-04-02：Dolphin；Dolphin: A Large-Scale Automatic Speech Recognition Model for Eastern Languages
- 2024/07/03：SenseVoice：阿里開源語音大模型：語音辨識效果與表現強於Whisper，還能偵測掌聲、笑聲、咳嗽等！
- 2024/05/01：使用Hugging Face 推理終端建立強大的「語音辨識+ 說話者分割+ 投機解碼」工作流程

- **Whisper**
  - Whisper: openAI開源準確率最高的通用語言語音識別
  - Robust Speech Recognition via Large-Scale Weak Supervision
  - Introducing Whisper
  - Whisper-Finetune
  - CrisperWhisper：CrisperWhisper：精確的逐字語音轉錄工具
  - WhisperLive：免費的即時語音轉文字工具：Whisper Live，精準高效，支援多語言
  - distil-whisper：語音辨識的未來已來-探索Distil-Whisper，輕量級AI的強大力量
  - whisper-speculative-decoding：使用推測解碼使Whisper實現2 倍的推理加速
  - insanely-fast-whisper；Insanely Fast Whisper：超快速的Whisper語音辨識腳本
  - Whisper-Finetune：微調Whisper語音辨識模型與加速推理
  - fine-tune-whisper：使用Transformers 為多語種語音識別任務微調Whisper 模型
  - WhisperX
  - Faster-Whisper對影片進行雙語字幕轉錄實踐(Python3.10)
  - Whisper斷句不夠好？用AI LLM和結構化資料打造完美字幕
- **FunASR**
  - FunASR: A Fundamental End-to-End Speech Recognition Toolkit
  - 阿里達摩院開源大型端到端語音識別工具包FunASR
  - 達摩院FunASR離線文件轉寫SDK發布
- **WeNet**
  - 58同城：WeNet端到端語音識別大規模落地方案
  - WeNet: Production First and Production Ready End-to-End Speech Recognition Toolkit
- **PaddleSpeech**
- **Speech Brain**：A PyTorch-based Speech Toolkit
- **Kaldi 2**：FSA/FST algorithms, differentiable, with PyTorch compatibility.
  - Next-gen-Kaldi 近期進展

▶ 2020/03-2021/01 開發心得

中文語者(聲紋)識別 (Chinese Speaker Recognition)

Chinese-Speaker-Recognition

https://www.twman.org/AI/ASR/SpeakerRecognition

找到描述特定對象的聲紋特徵，通過聲音判別說話人身份的技術；借助不同人的聲音，在語譜圖的分佈情況不同這一特徵，去對比兩個人的聲音，來判斷是否同人。

相關論文

- Wespeaker: A Research and Production oriented Speaker Embedding Learning Toolkit
- SincNet：Speaker Recognition from Raw Waveform with SincNet

相關連結

- Wespeaker v1.2.0 發布：新增SSL Recipe，NIST SRE 數據集支持, PLDA 及自適應代碼等
- ASV-Subtools聲紋識別實戰
- ICASSP 2023說話人識別方向論文合集（一）
- 聲紋識別原理
- 深度學習在聲紋識別中的應用
- 相關聲紋識別介紹匯整
- 提高聲紋辨識正確率 更添防疫新利器
- CN-Celeb-AV: 多場景視聽多模態數據集發布

▶ 2020/03/08-2020/08/29 開發心得

技術指標： 錯誤拒絕率(False Rejection Rate, FRR)：同類的兩人被系統判別為不同類。FRR為誤判案例在所有同類匹配案例中的比例 錯誤接受率(False Acceptance Rate, FAR)：不同類的兩人被系統判為同類。FAR為接受案例在所有異類匹配案例中的比例 等錯誤率(Equal Error Rate, EER)：調整threshold，當FRR=FAR時，FRR和FAR的數值稱為等錯誤率 準確率(Accuracy，ACC)：ACC=1-min(FAR+FRR)

速度： Real Time Factor 實時比:衡量提取時間跟音頻時長的關係，ex:1秒可以處理80s的音頻，實時比=1:80；驗證比對速度：平均每秒能進行的聲紋比對次數 ROC曲線：描述FAR和FRR間變化的曲線，X軸為FAR,Y軸為FRR。 閥值：當分數超過閥值才做出接受決定。

中文語音增強(去噪) (Chinese Speech Enhancement)

Chinese-Speech-Enhancement

https://www.twman.org/AI/ASR/SpeechEnhancement

https://huggingface.co/spaces/DeepLearning101/Speech-Quality-Inspection_Meta-Denoiser

找到描述特定聲音特徵，並將其去除以提高質量；從含雜訊的語音信號中提取出純淨語音的過程

相關論文

- Real Time Speech Enhancement in the Waveform Domain

相關連結

- 2024-12-07：ClearVoice: Speech Enhancement
  - 一站式語音處理工具包ClearerVoice-Studio，輕鬆實現語音降噪、分離、說話人提取
  - 阿里巴巴開源超強語音處理神器，語音分離、音訊視訊說話者擷取等功能一站式解決。
  - HuggingFace Space Demo
  - https://github.com/facebookresearch/denoiser
  - https://www.youtube.com/watch?v=77cm_MVtLfk

▶ 2020/08/30-2021/01/25 開發心得

中文語者分離(分割) (Chinese Speaker Separation)

Chinese-Speaker-Separation

https://www.twman.org/AI/ASR/SpeechSeparation

https://huggingface.co/spaces/DeepLearning101/Speech-Separation

從多個聲音信號中提取出目標信號；多個說話人情況的語音辨識問題，比如雞尾酒會上很多人講話

相關論文

- Stabilizing Label Assignment for Speech Separation by Self-supervised Pre-training
  - https://github.com/SungFeng-Huang/SSL-pretraining-separation
- Self-supervised Pre-training Reduces Label Permutation Instability of Speech Separation
  - https://github.com/SungFeng-Huang/SSL-pretraining-separation
- Sudo rm -rf: Efficient Networks for Universal Audio Source Separation
  - https://github.com/asteroid-team/asteroid/blob/master/asteroid/models/sudormrf.py
  - https://github.com/etzinis/sudo_rm_rf
- Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation
- Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation
  - https://github.com/JusperLee/Dual-path-RNN-Pytorch
  - Dual-path RNN for Speech Separation 閱讀筆記

相關連結

- 2025-06-03：SoloSpeech；高品質語音處理模型，一鍵提取指定說話者音訊並提升提取音訊清晰度和品質
- 2024-12-07：ClearVoice: Speech Enhancement：阿里巴巴開源超強語音處理神器，語音分離、音訊視訊說話者擷取等功能一站式解決。，HuggingFace Space Demo
- ICASSP2023論文代碼開源｜TOLD能對混疊語音建模的說話人日誌框架
- ICASSP 2023論文模型開源｜語音分離Mossformer

▶ 2020/08/30-2021/01/25 開發心得

中間也意外發現了Google brain 的 wavesplit，在有噪音及兩個人同時講話情形下，感覺效果還不差，但沒找到相關的code，未能進一步驗證或是嘗試更改數據集。還有又是那位有一起用餐之緣的深度學習大神 Yann LeCun 繼發文介紹 完去噪後，又發文介紹了語音分離；後來還有像是最早應用在NLP的Transformer等Dual-path RNN (DP-RNN) 或 DPT-NET (Dual-path transformer) 等應用在語音增強/分割，另外VoiceFilter、TasNet 跟 Conv-TasNet還有sudo-rm等等也是語音分割相關，當然更不能錯過臺大電機李宏毅老師一篇SSL-pretraining-separation的論文 (務必看完臺大電機李宏毅老師的影片)，最後也是多虧李老師及第一作者黃同學的解惑，然後小夥伴們才又更深入的確認並且解決問題。 這裡做數據時相對簡單一點，直接打散混合，再從中隨機挑選兩個人，然後分別挑出語音做混合，若長度不同，選擇短者為參考，將長者切到與短者相同，兩兩完全重疊或者兩兩互不完全重疊等都對效果有不小的影響；同時也研究了Data Parallel 跟 Distributed Data Parallel 的差異，但是如何才能在 CPU 上跑得又快又準才是落地的關鍵

中文語音合成 (Chinese Speech Synthesis)

Chinese-Speech-Synthesis

相關連結

- OpenAI-Edge-TTS：開源的免費文字轉語音API接口，託管在GitHub上
- fish-speech
  - 性能超過F5、CosySense，一文帶你理論+實操多種語言克隆效果
  - fish-speech-gui；Github；Document
- 2025-08-08：KittenTTS；超迷你 TTS 模型（小於 25 MB）
- 2025-07-30：Microsoft DragonV2.1
- 2025-07-25：Higgs Audio V2
  - Github
  - HF Space Playground
  - Boson AI Playground
  - 李沐B站更新了！教你手搓語音大模型，程式碼全開源還能在線上試玩
  - 2025-08-03：Higgs audio學會了越南語
  - https://mp.weixin.qq.com/s/k7PMihbveN8XUS-bvQOOsw
- 2025-07-23：FreeAudio；AI音效90秒長時可控生成！「狼嚎2秒，蟋蟀鳴8秒」精準搞定
- 2025-07-14：MOSS-TTSD；邱錫鵬團隊開源MOSS-TTSD！百萬小時音頻訓練
- 2025-06-05：OpenAudio S1；重新定義文字轉語音，釋放聲音的無限潛能；全球唯一高可控多語言TTS-OpenAudio-S1，並開源mini版
- 2025-03-30：MegaTTS3：字節跳動MegaTTS3 開源：0.45B 參數實現高品質中英雙語TTS 與語音克隆

- 2025-03-21：Orpheus TTS
  - Space@HuggingFace
  - 一款剛剛開源的TTS語音模型！ 25ms超低延遲支援即時對話
- 2025-03-15：CSM-Conversational Speech Generation Model
  - AI語音合成新標竿！開源10小時斬獲8K Star！ 1B參數實現電影級人聲!
  - 驅動「超真人」虛擬助理Maya的即時語音對話模型CSM-1b開源！
- 2025-03-02：Spark-TTS：基於單流解耦語音令牌的高效能文字轉語音模型
- 2025-03-01：Step-Audio：[ComfyUI]最新聲音複製技術，配合數位人無敵了
- 2024/11/30：MockingBird：MockingBird 開源語音克隆神器，5 秒速「復刻」 聲音，摘得35.4k 星閃耀佳績！
- 2024-11-02：Kokoro-TTS：Kokoro TTS：文字轉語音AI
- 2024/10/15：阿里最強TTS，STT工具CosyVoice，SenseVoice，在ComfyUI中的使用，讓你實現流暢對話
- 2024/10/15：F5-TTS：上海交大開源超逼真聲音克隆TTS，15秒克隆聲音
- 2024/09/09：Parler-TTS：Hugging Face 開源TTS 模型！一行指令即可安裝！
- 2024/06/09：ChatTTS：開源文字轉語音模型本地部署、API使用和建立WebUI介面
- 2024/08/6：VALL-E X語音模型，手把手實操語音轉文字和聲音克隆
- MeloTTS：https://github.com/myshell-ai/MeloTTS
  - 多語言即時文字轉語音的高品質工具！無GPU也可靈活使用！
- GPT-SoVITS：Github 獲得 35.2k star的開源聲音克隆項目，1分鐘語音訓練TTS模型
- Deepgram