2025 TSMC IT
CareerHack

ONLINE WORKSHOP

建立多語言的即時語音翻譯
企業溝通無國界翻譯系統

徐梓軒 Ivan Hsu
資訊建構暨通訊服務處
ICSD / Infrastructure and Communication Services Div.

2025 TSMC IT CareerHack

Agenda

- Topic Introduction
- Background
- Goal
- Evaluation Criteria
- Dataset Description
- Notice
- Submit Results
- Q&A

Speaker

徐梓軒

EPCD / Enterprise and Public Cloud Department

IT 工程師

education

交大AI所

experience

台積電 IT 2 年

- EPCD 工程師

thhsuy@tsmc.com

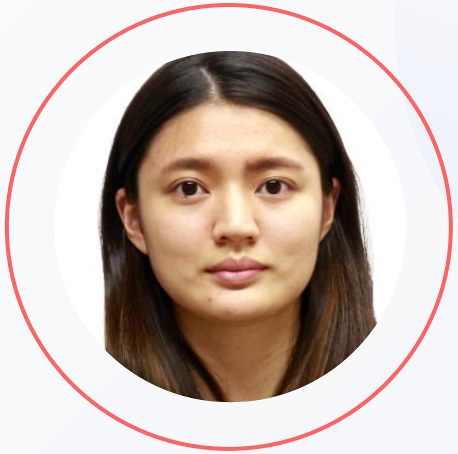
Meet Our Mentors

蘇予宣
Chris Su
IT Engineer

企業及公有雲建構部
EPCD
yxsuc@tsmc.com

陳德優
Luis Chen
Technical Manager

企業及公有雲建構部
EPCD
tychenx@tsmc.com



李毓簫
Daisy Li
IT Engineer

企業及公有雲建構部
EPCD
yhlizzi@tsmc.com

陳伯維
Martin Chen
IT Engineer

企業及公有雲建構部
EPCD
bwchent@tsmc.com

Background

- As new factories expand globally, **cross-language discussions** are often necessary for internal meetings, handovers, and client meetings.
- The abundance of **proper nouns** in the company hinders from understanding discussions and key points.
- Without **meeting minutes**, it is easy to forget the key points and decisions that have been discussed when there are many topics.

Three major system requirements

Real-Time Speech To Text Translation

ASR, Text Translation

Company-Specific Terminology Detector

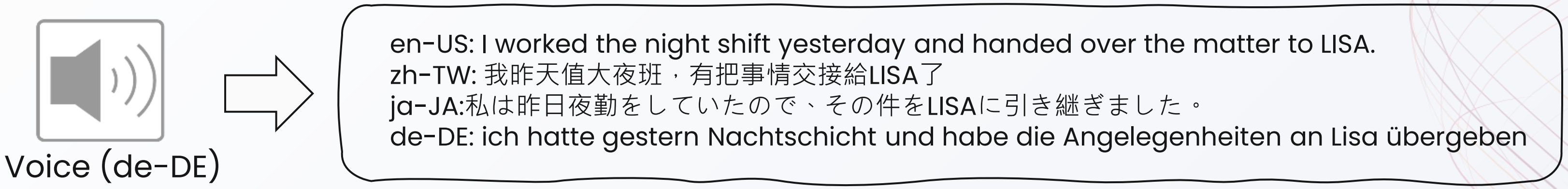
Key Word Detect, Specific Translate

Clerk Of Meetings

Search, Summarize

Task 1: Real-Time Speech To Text Translation

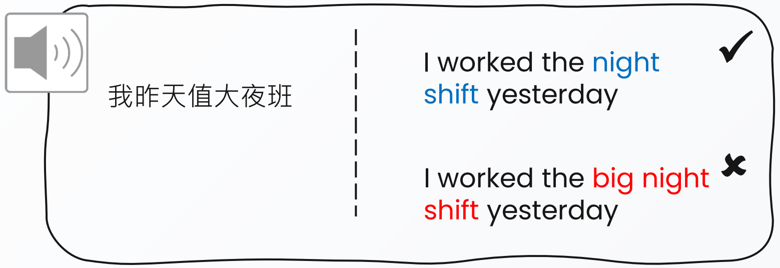
- **Objective: The system should provide instantaneous speech to text translation between multiple languages to facilitate seamless communication.**
- **Evaluation:**
  - Ability to handle multiple languages (Chinese, English, Japanese, German)
  - Obtain correctly translation text result
  - Real-time speech to text translation



en-US: I worked the night shift yesterday and handed over the matter to LISA.
zh-TW: 我昨天值大夜班，有把事情交接給LISA了
ja-JA:私は昨日夜勤をしていたので、その件をLISAに引き継ぎました。
de-DE: ich hatte gestern Nachtschicht und habe die Angelegenheiten an Lisa übergeben

Task 2: Company-Specific Terminology Detector

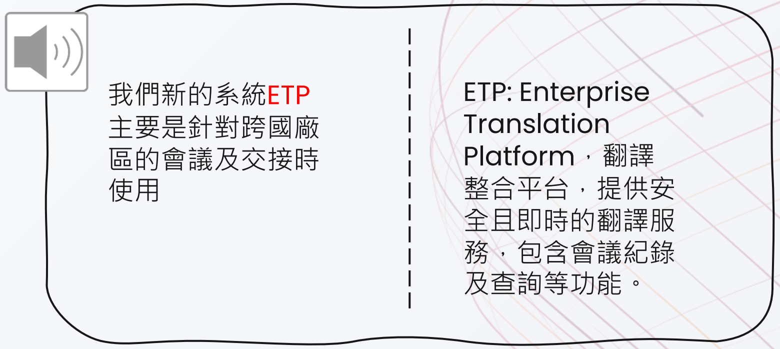
- **Objective: Integrate GenAI capabilities to understand and appropriately translate TSMC-specific jargon and technical terms.**
- **Evaluation:**
  - Detect technical terms and translate them accurately
  - Obtain correctly translation text result
  - Attached the description of professional terminology



我昨天值大夜班

I worked the night shift yesterday ✓

I worked the big night shift yesterday ✗

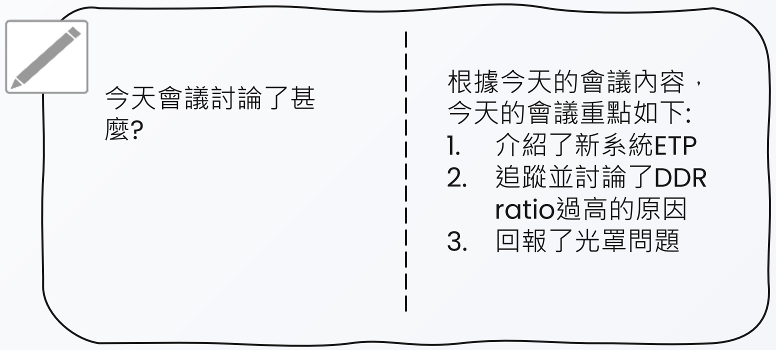


我們新的系統ETP主要是針對跨國廠區的會議及交接時使用

ETP: Enterprise Translation Platform，翻譯整合平台，提供安全且即時的翻譯服務，包含會議紀錄及查詢等功能。

Task 3: Clerk Of Meetings

- **Objective: Ability to record and transcribe meetings for future reference , system should enable users to search and summarize through the transcribed content after meeting.**
- **Evaluation:**
  - UI for retrieving historical transcriptions or performing transcription searches
  - Ease of use, including the clarity of the interface and search functionality



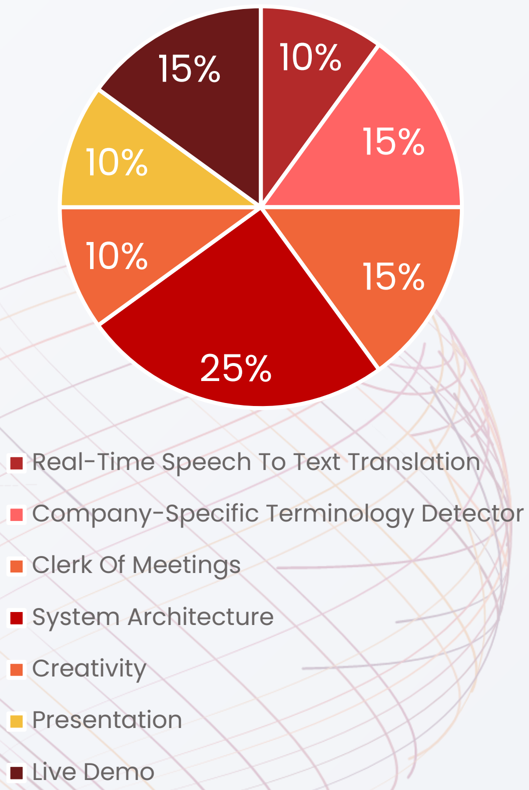
今天會議討論了甚麼?

根據今天的會議內容，今天的會議重點如下:

1. 介紹了新系統ETP
2. 追蹤並討論了DDR ratio過高的原因
3. 回報了光罩問題

Evaluation Criteria

Item	Evaluation Description	Usage Dataset	Score
Real-Time Speech To Text Translation	• Real-time translation function, can select specify language. • Output your Transcript, Chinese is the main language for scoring.	Testing dataset	10
Company-Specific Terminology Detector	• Detect and accurately translate proper nouns based on the provided dataset, 10~15 proper nouns includes in audio file. • List the proper nouns you detected. • Display the explanations of proper nouns on front-end.	Testing dataset	15
Clerk Of Meetings	• Store and organize meeting minutes, display with user input language. • The presentation results need to include the following items: meeting summary, assigned tasks.	Training dataset Testing dataset	15
System Architecture	• Including completion, design flow, UI or full system.	-	25
Creativity	• In addition to data processing and analysis, what other usage or scenario can be used?	-	10
Presentation	• The solution should be presented in a clear and concise manner, with appropriate documentation.	-	10
Live Demo	• Live demo your solution with demo audio, the audio file lasts about 1 minute. • Output your Transcript, Chinese is the main language for scoring.	Demo dataset	15



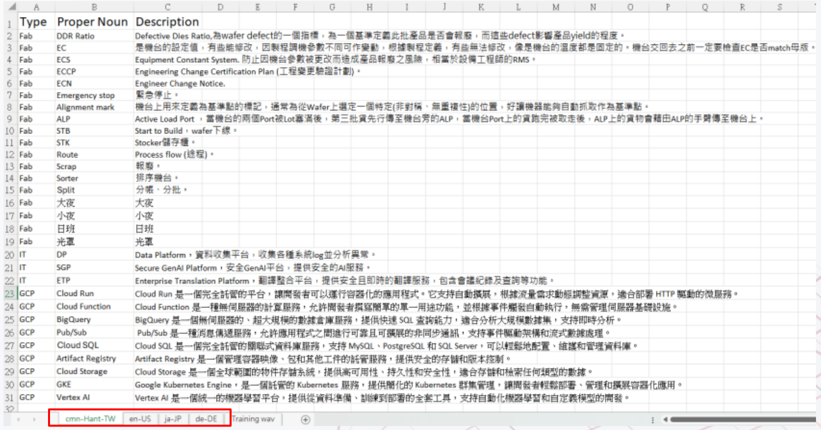
Dataset Description

- **Training.zip:**
  - Training.wav: Recording during the meeting, including Chinese, English, Japanese and German.
  - Knowledge Dataset.xlsx: Special terminology data set, including Chinese, English, Japanese and German.
- **Testing.zip:**
  - Testing.wav: Recording during the meeting, including Chinese, English, Japanese and German.

Knowledge Dataset.xlsx

Contains proper nouns and corresponding translations in four languages

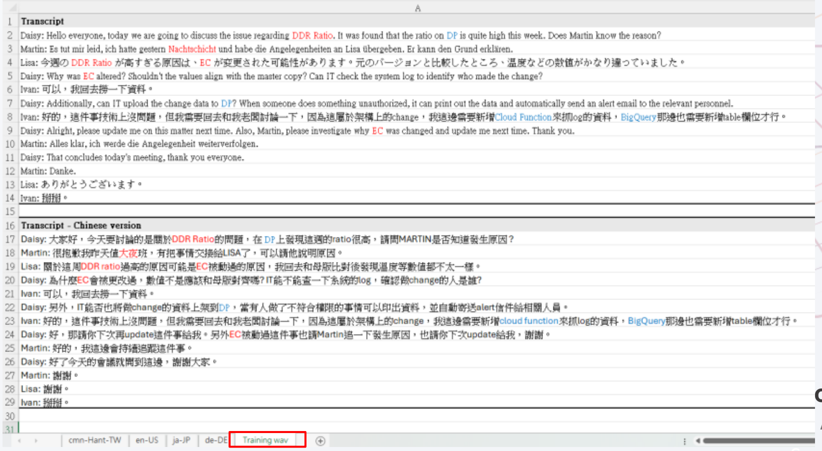
Sheet: cmn-Hant-TW, en-US, ja-JP, de-DE



Type	Proper Noun	Description
Fab	DDR Ratio	Defective Dies Ratio,為wafer defect的一個指標，為一個基準定義此批產品是否會報廢，而這些defect影響產品yield的程度。
Fab	EC	是機台的設定值，有些能修改，因製程調機參數不同可作變動，根據製程定義，有些無法修改，像是機台的溫度都是固定的。機台交回去之前一定要檢查EC是否match母版。
Fab	ECS	Equipment Constant System. 防止因機台參數被更改而造成產品報廢之風險，相當於設備工程師的RMS。
Fab	ECCP	Engineering Change Certification Plan (工程變更驗證計劃)。
Fab	ECN	Engineer Change Notice.
Fab	Emergency stop	緊急停止。
Fab	Alignment mark	機台上用來定義為基準點的標記，通常為從Wafer上選定一個特定(非對稱、無重複性)的位置，好讓機器能夠自動抓取作為基準點。
Fab	ALP	Active Load Port ，當機台的兩個Port被Lot塞滿後，第三批貨先行傳至機台旁的ALP，當機台Port上的貨跑完被取走後，ALP上的貨物會藉由ALP的手臂傳至機台上。
Fab	STB	Start to Build，wafer下線。
Fab	STK	Stocker儲存櫃。
Fab	Route	Process flow (途程)。
Fab	Scrap	報廢。
Fab	Sorter	排序機台。
Fab	Split	分帳、分批。
Fab	大夜	大夜
Fab	小夜	小夜
Fab	日班	日班
Fab	光罩	光罩
IT	DP	Data Platform，資料收集平台，收集各種系統log並分析異常。
IT	SGP	Secure GenAI Platform，安全GenAI平台，提供安全的AI服務。
IT	ETP	Enterprise Translation Platform，翻譯整合平台，提供安全且即時的翻譯服務，包含會議紀錄及查詢等功能。
GCP	Cloud Run	Cloud Run 是一個完全託管的平台，讓開發者可以運行容器化的應用程式。它支持自動擴展，根據流量需求動態調整資源，適合部署 HTTP 驅動的微服務。
GCP	Cloud Function	Cloud Function 是一種無伺服器的計算服務，允許開發者撰寫簡單的單一用途功能，並根據事件觸發自動執行，無需管理伺服器基礎設施。
GCP	BigQuery	BigQuery 是一個無伺服器的、超大規模的數據倉庫服務，提供快速 SQL 查詢能力，適合分析大規模數據集，支持即時分析。
GCP	Pub/Sub	Pub/Sub 是一種消息傳遞服務，允許應用程式之間進行可靠且可擴展的非同步通訊，支持事件驅動架構和流式數據處理。
GCP	Cloud SQL	Cloud SQL 是一個完全託管的關聯式資料庫服務，支持 MySQL、PostgreSQL 和 SQL Server，可以輕鬆地配置、維護和管理資料庫。
GCP	Artifact Registry	Artifact Registry 是一個管理容器映像、包和其他工件的託管服務，提供安全的存儲和版本控制。
GCP	Cloud Storage	Cloud Storage 是一個全球範圍的物件存儲系統，提供高可用性、持久性和安全性，適合存儲和檢索任何類型的數據。
GCP	GKE	Google Kubernetes Engine，是一個託管的 Kubernetes 服務，提供簡化的 Kubernetes 群集管理，讓開發者輕鬆部署、管理和擴展容器化應用。
GCP	Vertex AI	Vertex AI 是一個統一的機器學習平台，提供從資料準備、訓練到部署的全套工具，支持自動化機器學習和自定義模型的開發。

Contains Transcript and Chinese translation of Training.wav, Red is type fab, Blue is type IT and GCP

Sheet: Training wav



Transcript

Daisy: Hello everyone, today we are going to discuss the issue regarding DDR Ratio. It was found that the ratio on DP is quite high this week. Does Martin know the reason?

Martin: Es tut mir leid, ich hatte gestern Nachtschicht und habe die Angelegenheiten an Lisa übergeben. Er kann den Grund erklären.

Lisa: 今週の DDR Ratio が高すぎる原因は、EC が変更された可能性があります。元のバージョンと比較したところ、温度などの数值がかなり違っていました。

Daisy: Why was EC altered? Shouldn't the values align with the master copy? Can IT check the system log to identify who made the change?

Ivan: 可以，我回去撈一下資料。

Daisy: Additionally, can IT upload the change data to DP? When someone does something unauthorized, it can print out the data and automatically send an alert email to the relevant personnel.

Ivan: 好的，這件事技術上沒問題，但我需要回去和我老闆討論一下，因為這屬於架構上的change，我這邊需要新增Cloud Function來抓log的資料，BigQuery那邊也需要新增table欄位才行。

Daisy: Alright, please update me on this matter next time. Also, Martin, please investigate why EC was changed and update me next time. Thank you.

Martin: Alles klar, ich werde die Angelegenheit weiterverfolgen.

Daisy: That concludes today's meeting, thank you everyone.

Martin: Danke.

Lisa: ありがとうございます。

Ivan: 謝謝。

Transcript - Chinese version

Daisy: 大家好，今天要討論的是關於DDR Ratio的問題，在 DP上發現這週的ratio很高，請問MARTIN是否知道發生原因？

Martin: 很抱歉我昨天值大夜班，有把事情交接給LISA了，可以請他說明原因。

Lisa: 關於這周DDR ratio過高的原因可能是EC被動過的原因，我回去和母版比對後發現溫度等數值都不太一樣。

Daisy: 為什麼EC會被更改過，數值不是應該和母版對齊嗎? IT能不能查一下系統的log，確認做change的人是誰?

Ivan: 可以，我回去撈一下資料。

Daisy: 另外，IT能否也將做change的資料上架到DP，當有人做了不符合權限的事情可以印出資料，並自動寄送alert信件給相關人員。

Ivan: 好的，這件事技術上沒問題，但我需要回去和我老闆討論一下，因為這屬於架構上的change，我這邊需要新增cloud function來抓log的資料，BigQuery那邊也需要新增table欄位才行。

Daisy: 好，那請你下次再update這件事給我。另外EC被動過這件事也請Martin追一下發生原因，也請你下次update給我，謝謝。

Martin: 好的，我這邊會持續追蹤這件事。

Daisy: 好了今天的會議就開到這邊，謝謝大家。

Martin: 謝謝。

Lisa: 謝謝。

Ivan: 謝謝。

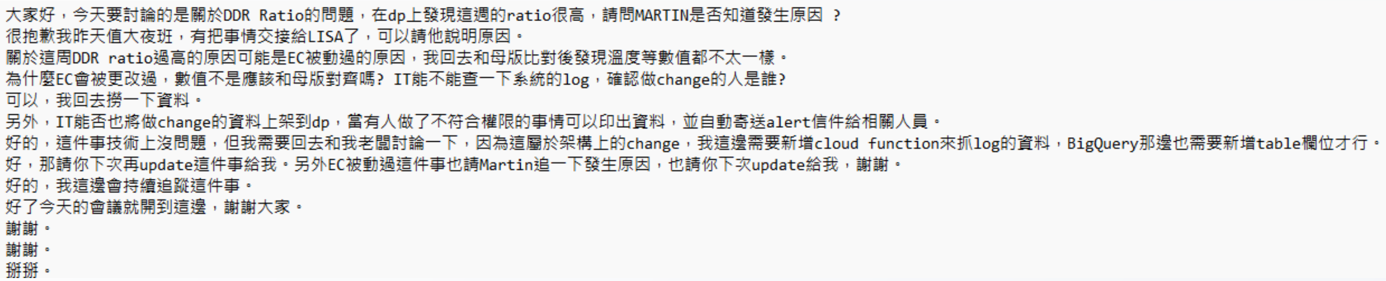
Notice

- **Remember turning off GCP environment once you don’t perform anything.**
- **Download dataset from “careerhack2025-icsd-resource-bucket” GCP bucket with gsutil command.**
  - gsutil cp gs://careerhack2025-icsd-resource-bucket/Training.zip gs://{your bucket name}/
- **Resource:**
  - Cloud Storage & Artifact Registry: Each team has 200 GB to store your data.
  - Compute Engine, Cloud Run, Cloud SQL, GKE: Each team has 60 hours to use.
  - 5 points will be deducted if GPU(GKE) exceeds one hour
- **Teamwork is important, which is also important in the workplace.**
- **Please send your question to careerhack@tsmc.com with title [企業溝通無國界翻譯系統] C* after meeting.**

Submit Results

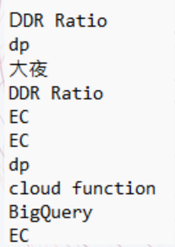
Transcript & Proper Nouns

- **Submit the transcript & proper nouns list you identified based on the testing wav, upload to your own bucket before 2/15 13:00**
- **List all detected proper nouns. If there are duplicates, don't list only one.**
- **Naming by [Cnumber_transcript] and [Cnumber_proper_nouns].**
  - C1_ transcript.txt
  - C2_ proper_nouns.txt



```
大家好，今天要討論的是關於DDR Ratio的問題，在dp上發現這週的ratio很高，請問MARTIN是否知道發生原因 ?
很抱歉我昨天值大夜班，有把事情交接給LISA了，可以請他說明原因。
關於這周DDR ratio過高的原因可能是EC被動過的原因，我回去和母版比對後發現溫度等數值都不太一樣。
為什麼EC會被更改過，數值不是應該和母版對齊嗎? IT能不能查一下系統的log，確認做change的人是誰?
可以，我回去撈一下資料。
另外，IT能否也將做change的資料上架到dp，當有人做了不符合權限的事情可以印出資料，並自動寄送alert信件給相關人員。
好的，這件事技術上沒問題，但我需要回去和我老闆討論一下，因為這屬於架構上的change，我這邊需要新增cloud function來抓log的資料，BigQuery那邊也需要新增table欄位才行。
好，那請你下次再update這件事給我。另外EC被動過這件事也請Martin追一下發生原因，也請你下次update給我，謝謝。
好的，我這邊會持續追蹤這件事。
好了今天的會議就開到這邊，謝謝大家。
謝謝。
謝謝。
掰掰。
```

C1_ transcript.txt



```
DDR Ratio
dp
大夜
DDR Ratio
EC
EC
dp
cloud function
BigQuery
EC
```

C2_ proper_nouns.txt

Submit Results

Demo ppt & Presentation

- Upload your **demo ppt to your own bucket before 2/15 13:45**.
- **Demo ppt naming by [Cnumber_vnumber].**
  - C1_v1.pptx
  - C2_v3.pptx
- **Presentation Timeline Description:**
  - Presentation: 15 mins (include live demo)
  - QA: 5 mins
  - Live demo: 1 min

Timeline

2/7 Workshop

1. Workshop
2. Provide cloud resources
3. Provide training dataset

2/14 Hack Day 1

1. Hack Day 1
2. Provide testing dataset

2/15 Hack Day 2

1. Hack Day 2
2. Submit the transcript and the proper nouns you identified based on the testing wav, upload to your own bucket before 13:00
3. Present your project
4. Upload your ppt before 13:45

Reference

GCP Vertex AI RAG Sample Code: Link

GCP Speech-to-text Sample Code: Link

Langchain with Cloud SQL: Link

Q&A

2025 TSMC IT
CareerHack

Thank You

SEE YOU NEXT PRESENTATION