Unlocking Insights from Audio

Foundations of Speech Processing and Advanced Language Models

Warit Sirichotedumrong (Boom)

Research Scientist, SCB 10X R&D

SCB 10X R&D

A team of experienced AI professionals, specializing in Thai Natural Language Processing (NLP)

Driving Impactful Research and Development in the field of Thai NLP

Developing Open-Source AI Models, Datasets, and Tools

Exploring Real-World Use Cases and Applications in the Thai Market

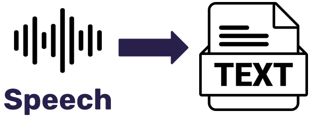
Fostering a Robust Thai NLP Ecosystem Through Collaboration & Community Building

Agenda

- Understanding How We Process Sound & Language
- Key Concepts in Speech Processing
- From Traditional Methods to the Power of AI
- A Deep Dive into “Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models”
- Applying Research to Complex Audio Challenges
- Key Takeaways & Q&A

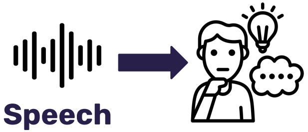
Turning Spoken Language to Actionable Information

Speech Recognition



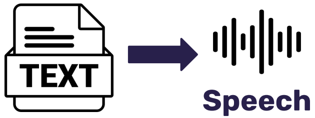
"Human voice into digital understanding"

Speech Understanding



"Interpreting meaning, intent, and emotion from the human voice"

Speech Synthesis



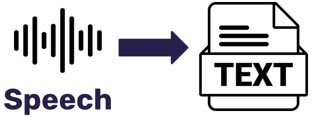
"Converting text into natural-sounding speech"

Other Areas

- Speaker Verification
- Speaker Diarization
- Emotion Detection

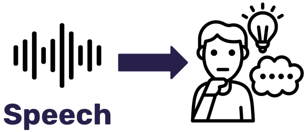
Turning Spoken Language to Actionable Information

Speech Recognition



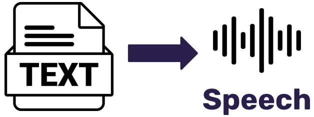
“Human voice into digital understanding”

Speech Understanding



“Interpreting meaning, Intent, and emotion from the human voice”

Speech Synthesis

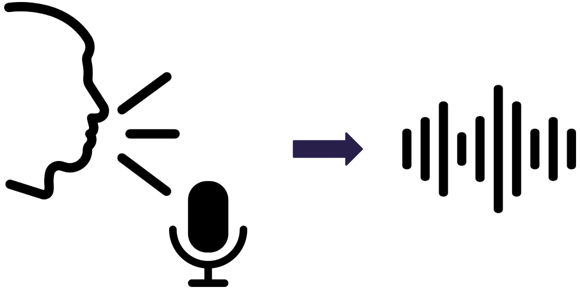


“Converting text into natural-sounding speech”

Other Areas

- Speaker Verification
- Speaker Diarization
- Emotion Detection

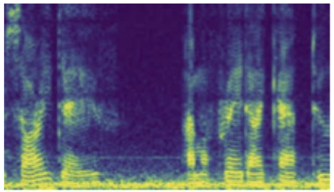
Key Concept in Speech Processing



Audio Capture (Microphone, Digital Signals)

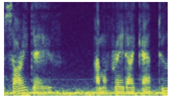
Speech Processing
- Noise Reduction
- Echo Cancellation

Feature Extraction
- Spectrogram
- Mel Frequency Cepstral Coefficients (MFCCs)



Key Concept in Speech Processing

Acoustic Features → Phonetic Units



[/s/ /à/ /w/ … /t/ /d/ /iː/]

"Maps acoustic features to phonetic units (the basic sounds of a language)"

Acoustic Modeling

→ Predict Word Sequence Probability

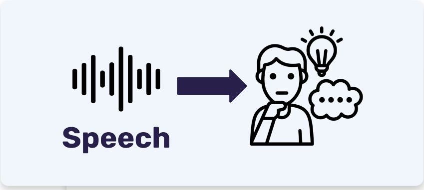
ฉันรักเธอ chăn rák ter ✅

ฉันหลักเธอ chăn làk ter ❌

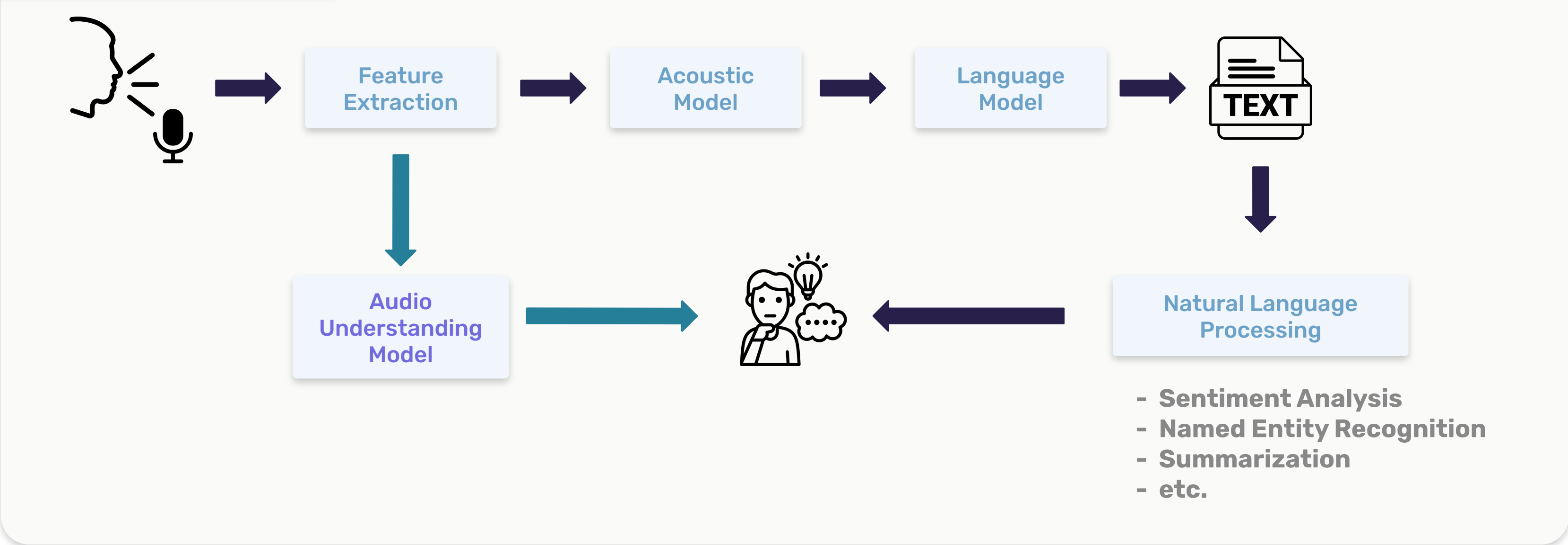
"Predict the likelihood of word sequences"

Language Modeling

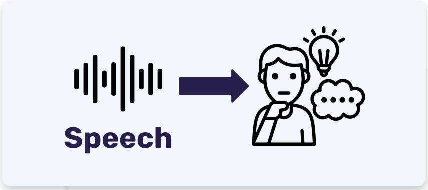
Natural Language Understanding



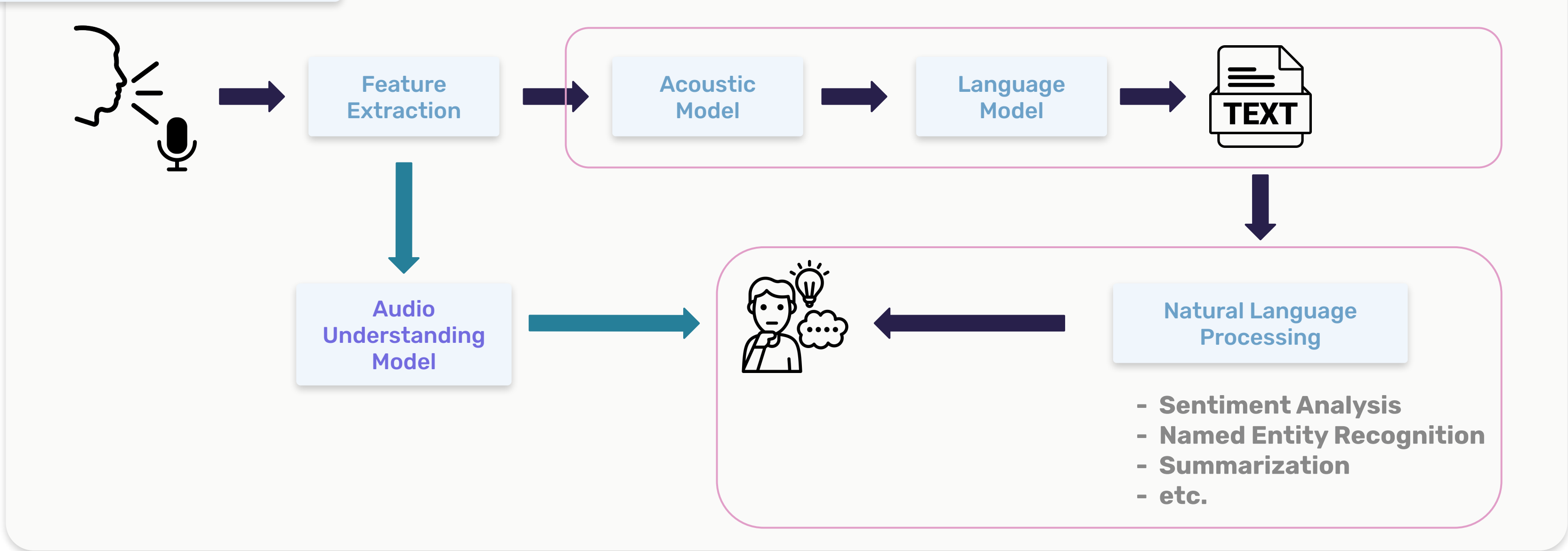
“NLU from speech: analyze the text after ASR, or understand directly from the audio.”



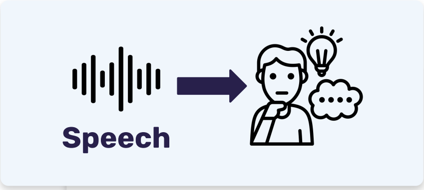
Natural Language Understanding



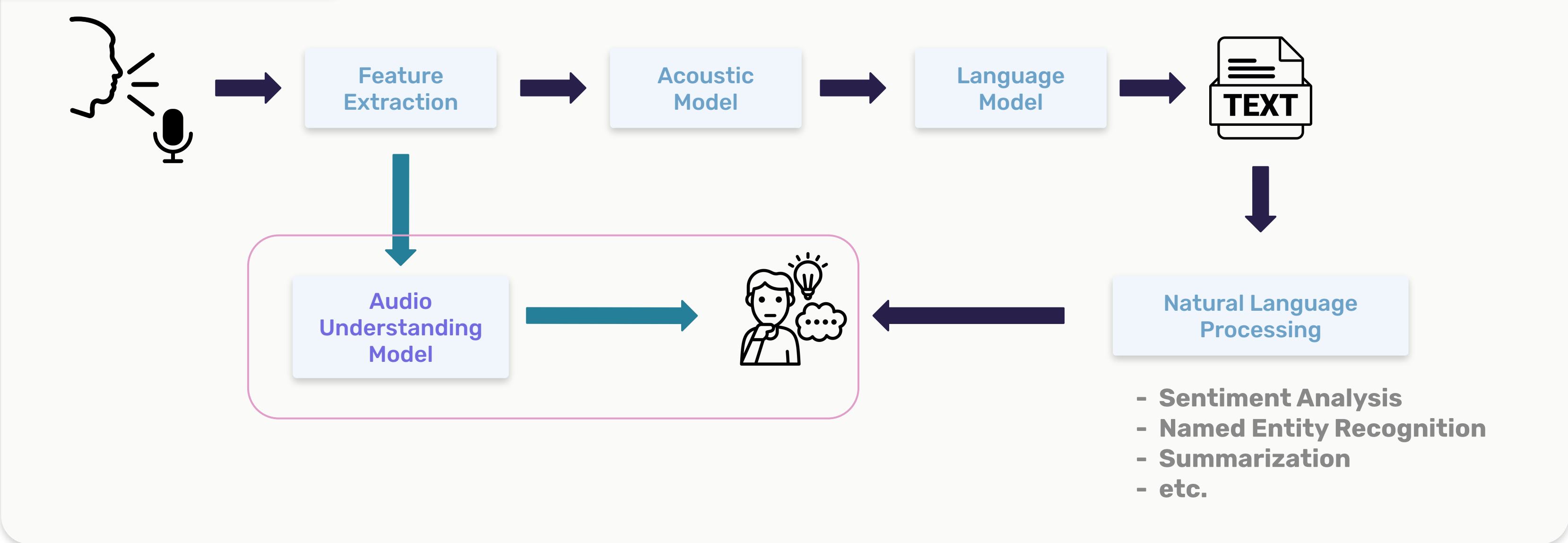
"NLU from speech: analyze the text after ASR, or understand directly from the audio."



Natural Language Understanding

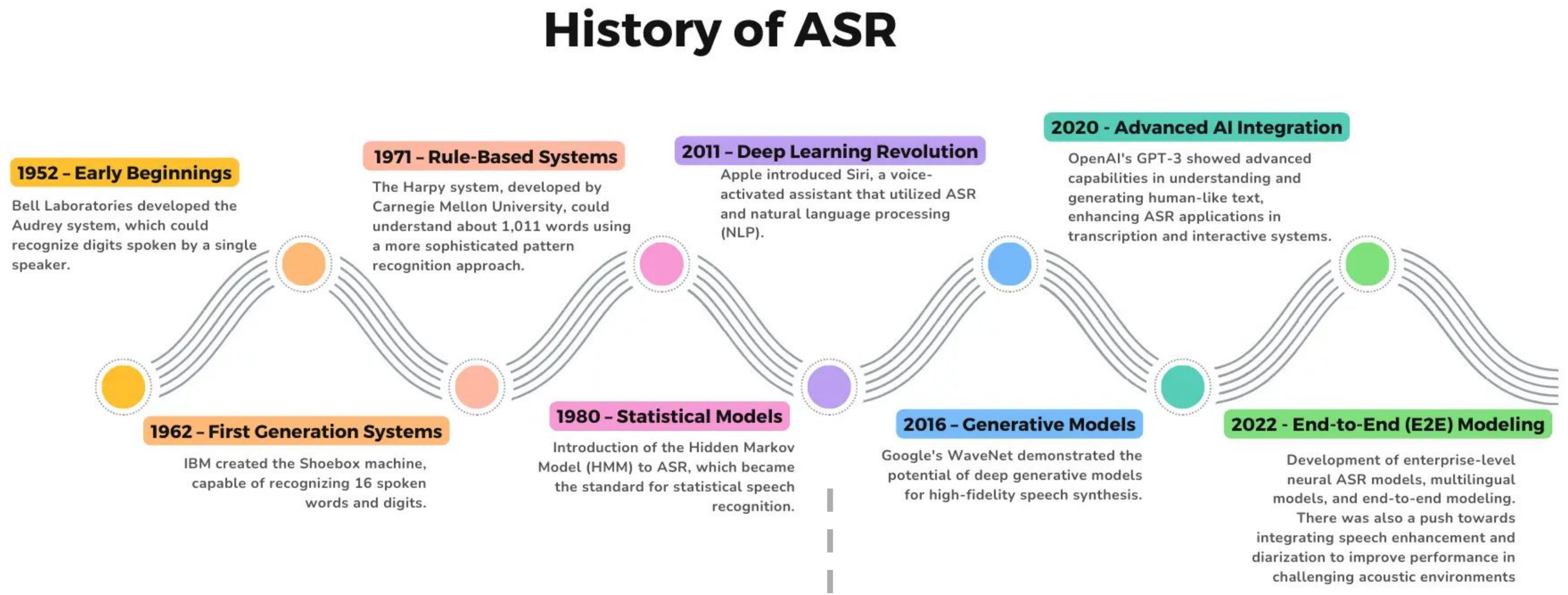


"NLU from speech: analyze the text after ASR, or understand directly from the audio."



From Traditional Methods to AI & LLMs

The AI Revolution in Understanding Speech



Statistical Era

(Required significant domain expertise and handcraft features)

Deep Learning Era

(End-to-End ASR: DNNs, CNNs, RNNs, Transformers)

Credit: https://www.labellerr.com/blog/asr-models-enhacement-with-llm/

DEEP DIVE INTO

“Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models”

Typhoon Audio Language Model
(To be published in Interspeech 2025)

TYPHOON | SCB 10X

- **Address a critical gap: making advanced audio language models (ALMs) work for less-resourced languages/domains**
- **Focuses on combining audio understanding with instruction following**
- **Provide methodology for building and training the model**
- **Link:** https://arxiv.org/abs/2409.10999

Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models

Potsawee Manakul[1], Guangzhi Sun[2]
Warit Sirichotedumrong[1], Kasima Tharnpipitchai[1], Kunat Pipatanakul[1]
[1]*SCB 10X* [2]*Univeristy of Cambridge*
{potsawee, warit, kasima, kunat}@scb10x.com, gs534@cam.ac.uk

***Abstract*—Audio language models can understand audio inputs and perform a range of audio-related tasks based on instructions, such as speech recognition and audio captioning, where the instructions are usually textual prompts. Audio language models are mostly initialized from pre-trained audio encoders and large language models (LLMs). Although these pre-trained components were developed to support multiple languages, audio-language models are trained predominantly on English data, which may limit their usability to only English instructions or English speech inputs. First, this paper examines the performance of existing audio language models in an underserved language using *Thai* as an example. This paper demonstrates that, despite being built on multilingual backbones, audio language models do not exhibit cross-lingual emergent abilities to low-resource languages. Second, this paper studies data mixture for developing audio language models that are optimized for a target language as well as English. In addition. this paper integrates audio comprehension and speech instruction-following capabilities into a single unified model. Our experiments provide insights into data mixture for enhancing instruction-following capabilities in both a low-resource language and English. Our model, *Typhoon-Audio*, outperforms existing open-source audio language models by a considerable margin, and it is comparable to state-of-the-art Gemini-1.5-Pro in both English and Thai languages.**

***Index Terms*—audio language model, large language model, instruction following, low-resource language**

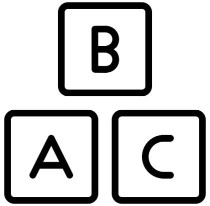
Audio language models typically comprise three key components: an audio encoder backbone, an LLM backbone, and an adapter module, as outlined in Section II. Despite leveraging multilingual backbones, most models are primarily trained on: (1) English data, and (2) only audio content understanding tasks, as seen in models like Qwen-Audio [1] and SALMONN [2], or only speech instruction understanding, such as AudioChatLlama [6]. Addressing these limitations, this work focuses on two goals. First, we examine model performance in a low-resource language using Thai as a case study, and we provide a recipe to enhance the low-resource language ability while retaining the English performance. Second, we integrate *improved* audio-understanding and speech instruction understanding capabilities into one unified model.

II. Related Work

Audio Language Models: *SALMONN* integrates three primary components: an LLM based on Vicuna [8], a speech encoder based on the encoder of Whisper-large-v2 [9], and BEATS [10] for audio events. The representations from Whisper and BEATS are concatenated and passed through an adapter (connection module based on Q-Former) to obtain

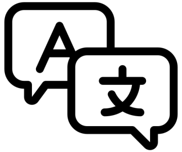
Core Problems

English Centric



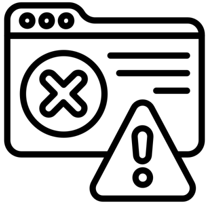
"Most open-source audio LMs are English Centric."

Low-resource Language Capability



"Without targeted training, ALMs built on multilingual backbones often fail to perform well on low-resource language"

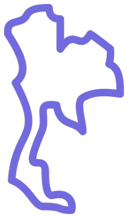
Limitation



"Many ALMs focus on either audio understanding or speech instruction following (Speech IF)"

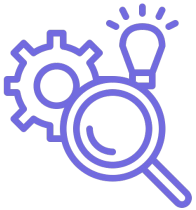
Goals

Performance Improvement in low-resource language (Thai)



"Improve performance in Thai while maintaining its English proficiency."

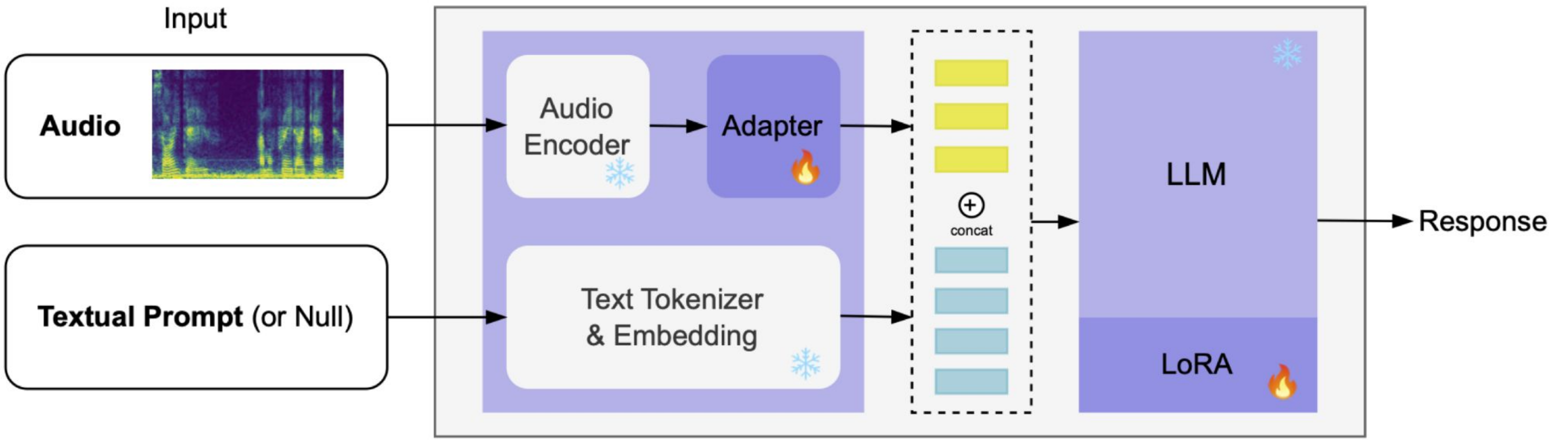
Both Speech IF and Audio Understanding



"Integrate improved audio comprehension and speech IF capability into a single model."

Model Architecture

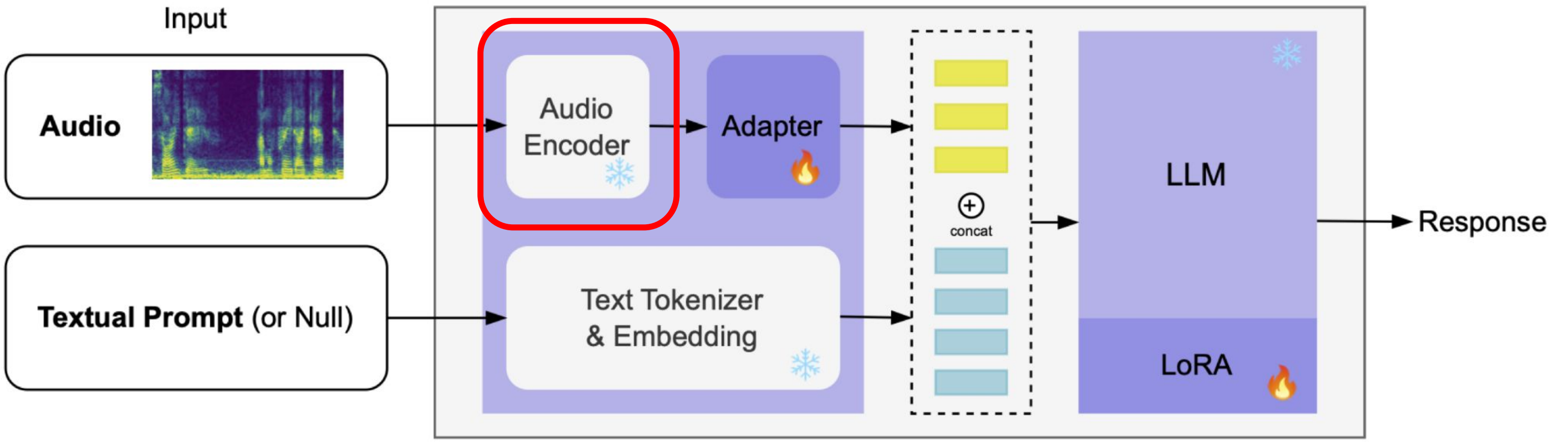
Input: Audio & Text Prompt
Output: Text Response



Model Architecture

Audio Encoder Backbone (Transforming raw audio input into rich numerical embeddings):

1. Processing speech inputs (biodatlab/Whisper-th-large-v3-combined)
2. BEATs: For encoding general audio events (music and environmental sound)

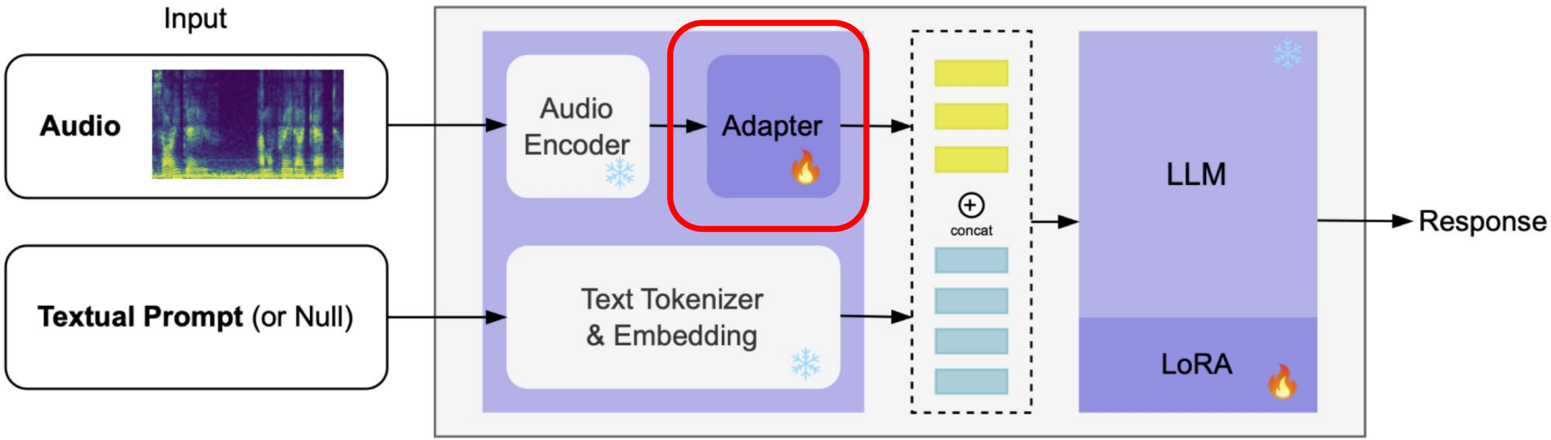


"Whisper-th-large-v3-combined specialized in turning speech spectrograms into advanced Thai language representation."

Model Architecture

Adapter Module (Q-Former)

Takes the audio representation from encoders and maps them into a sequence of embeddings that are "understandable" by the LLM.

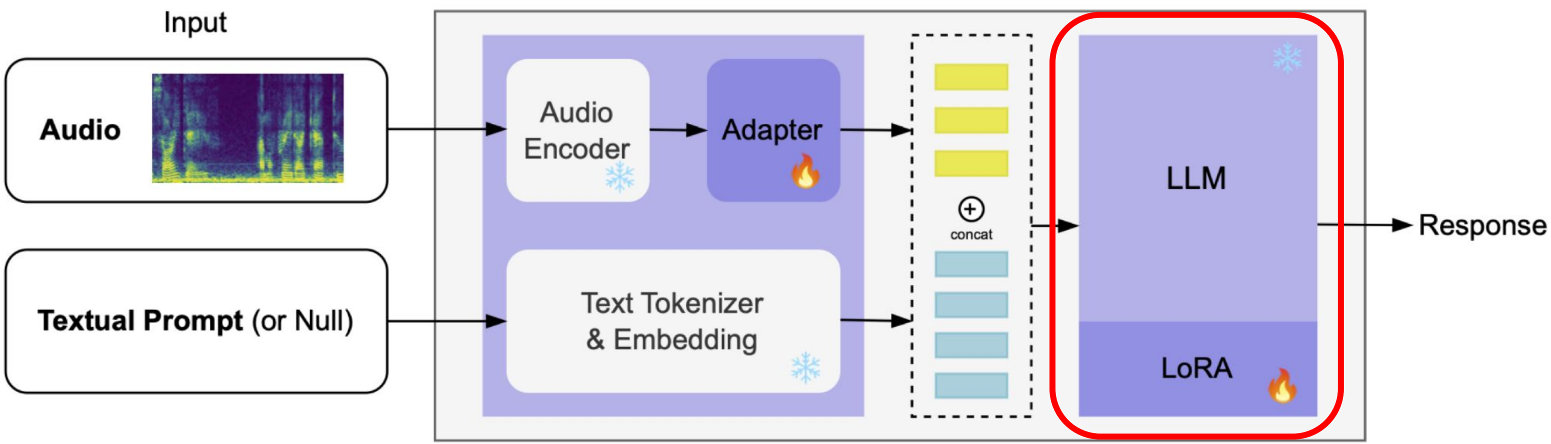


"This is to map the audio representations into the same semantic space as text embeddings."

Model Architecture

LLM Backbone

- Processes the combined audio features (via adapter) and any textual prompt to generate the textual response, follow instructions, or perform understanding task.



"Typhoon-1.5-8B-Instruct, a Llama3-based model, has been further pre-trained on a mix of English and Thai text, and then instruction fine-tuned."

Model Training Strategies

Pre-training the Adapter

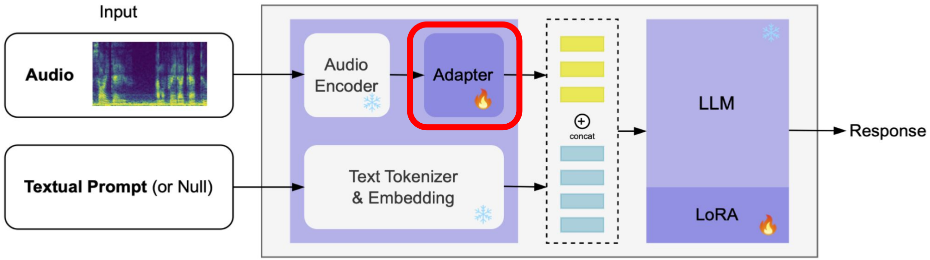
To align the audio representation from encoder with the textual representation space of the LLM.

Supervised Fine-Tuning (SFT)

To enhance the model instruction-following capability across a diverse range of tasks and in both English and Thai

Phase 1 - Pre-training the Adapter

Goal: To align audio representation with the textual representation space of the LLM



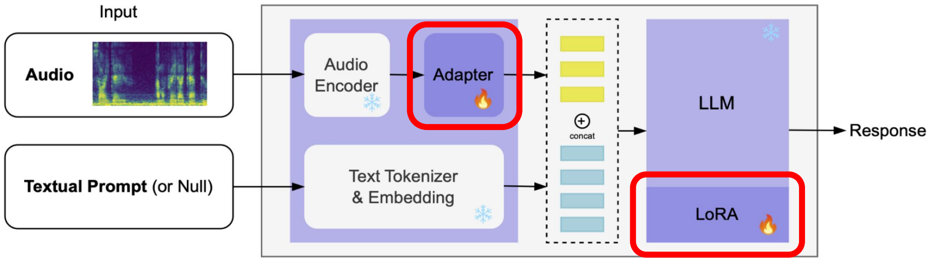
"Only the Adapter module. The pre-trained audio encoders and the LLM are kept frozen."

TABLE I
PRE-TRAINING DATA – 1.82M EXAMPLES IN TOTAL

Dataset	Task	Lang	#Examples
LibriSpeech [20]	ASR	En	281K
GigaSpeech-M [21]	ASR	En	900K
CommonVoice-Th [22]	ASR	Th	436K
Fleurs-Th [23]	ASR	Th	7.8K
Vulcan+Elderly+Gowajee	ASR	Th	65.1K
AudioCaps [24]	Audio Caption	En+Th	48.3K+48.3K
Clotho [25]	Audio Caption	En+Th	19.2K+19.2K

Phase 2 - Supervised Fine-Tuning (SFT)

Goal: To enhance the model instruction-following capability across a diverse range of tasks and in both English and Thai



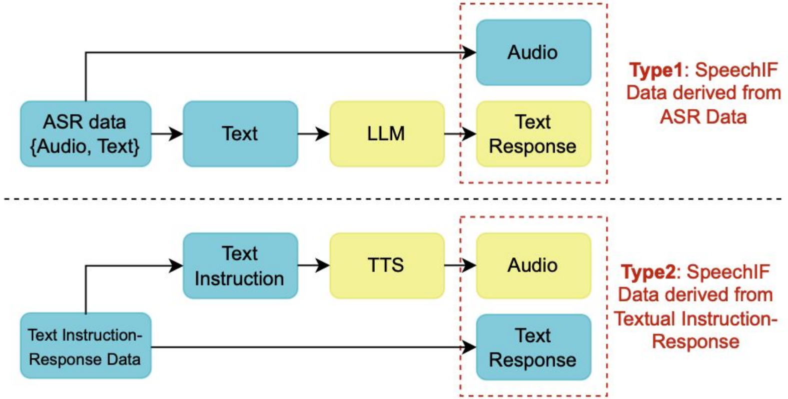
"Train both the Adapter and the LLM (LoRA). The audio backbone remains frozen."

TABLE II
SFT DATA OF TYPHOON-AUDIO – 640K EXAMPLES IN TOTAL

Dataset	Task	New	#Examples
QA pairs taken from SALMONN used in `SFT-v1`, `SFT-v2`, `SFT-v3`			
LibriSpeech [20]	QA (Speech-En)	✗	40.0K
AudioCaps [24]	QA (Audio)	✗	30.0K
QA pairs taken from LTU-AS used in `SFT-v1`, `SFT-v2`, `SFT-v3`			
LibriTTS [26]	QA (Speech-En)	✗	21.1K
IEMOCAP [27]	QA (Speech-En)	✗	4.3K
FSD50K [28]	QA (Audio)	✗	11.5K
AudioSet [29]	QA (Audio-Speech)	✗	20.0K
AS20k [30]	QA (Audio-Speech)	✗	12.0K
ASR, Translation, Audio Caption, QA used in `SFT-v2`, `SFT-v3`			
LibriSpeech [20]	ASR (En)	✗	32.0K
CommonVoice-Th [22]	ASR (Th)	✗	52.0K
SelfInstruct-Th	ASR (Th)	✓	18.9K
AudioCaps(Gemini)	Audio Caption	✓	48.3K
Covost2 [31]	Translate (X2Th)	✗	30.0K
CommonVoice-Th [22]	Translate (Th2X)	✗	7.3K
VISTEC-SER [32]	QA (Emotion & Gender)	✓	18.0K
Yodas2-30S [33]	QA (Speech-Th)	✓	90.0K
Speech Instruction Following used in `SFT-v3`			
GigaSpeech [21]	SpeechIF-Type1 (En)	✓	20.0K
CommonVoice-Th [22]	SpeechIF-Type1 (Th)	✓	120.5K
jan-hq-instruction-v1 [34]	SpeechIF-Type2 (En)	✗	20.0K
Airoboros-Th	SpeechIF-Type2 (Th)	✓	5.7K
Alpaca-Th	SpeechIF-Type2 (Th)	✓	20.0K
SelfInstruct-Th	SpeechIF-Type2 (Th)	✓	18.9K

Data Generation for Speech IF

"Speech Instruction Following (SpeechIF) requires models to listen to spoken instructions and directly response."



- Lacking of existing data for SpeechIF
- There are two types:
  a. Type 1: Derive from ASR data
  b. Type 2: Synthesized from Textual Instruction-Response data

Evaluation: Tasks & Metrics

Task	Direction	Metric
Automatic Speech Recognition (ASR)	⇩	Word Error Rate (WER) on English and Thai
Translation	⇧	BLEU score for Thai-to-English and English/Other-to-Thai
Gender Classification	⇧	Accuracy on English and Thai
Spoken QA	⇧	F1 score on English and Thai
SpeechIF (Judge)	⇧	Human/GPT4o judged score (1-10 scale) for English and Thai
ComplexIF (Judge)	⇧	Multi-step instructions (English only). Judged on Quality & Format

Key Results & Findings

TABLE IV
AUDIO LM EVALUATION IN ENGLISH AND THAI. ASR: EN=LIBRISPEECH-OTHER, TH=COMMONVOICE-17; TRANSLATION: TH-TO-EN=COMMONVOICE-17, EN/X-TO-TH=COVOST2 WHERE REFERENCE TEXTS ARE DERIVED FROM TRANSLATION; GENDER CLASSIFICATION: EN & TH = FLEURS; SPOKENQA: EN=SPOKENSQUAD [37], TH=COMMONVOICE-17 WHERE QA ARE GENERATED FROM REFERENCES USING GPT-4O; SPEECHIF: EN=ALPCAEVAL-TTS, TH=SELFINSTRUCT-TTS. COMPLEXIF: MIXTURE OF 5 OTHER TASKS IN ENGLISH. SPEECHIF/COMPLEXIF $\in$ [1,10]

Model	Size	ASR (%WER↓)		Translation (%BLEU↑)			Gender (%Acc↑)		SpokenQA (%F1↑)		SpeechIF (Judge↑)		ComplexIF (Judge↑)		
		En	Th	Th2En	En2Th	X2Th	En	Th	En	Th	En	Th	Qual	Format	Avg.
Qwen-Audio [1]	7B	6.94	95.12	0.00	2.48	0.29	37.09	67.97	25.34	0.00	1.07	1.03	3.13	1.68	2.41
SALMONN [2]	13B	**5.79**	98.07	14.97	0.07	0.10	95.69	93.26	52.92	2.95	2.47	1.18	4.10	5.09	4.60
DiVA [36]	8B	30.28	65.21	7.97	9.82	5.31	47.30	50.12	44.52	15.13	**6.81**	2.68	6.33	7.83	7.08
Gemini-1.5-Pro	-	5.98	**13.56**	22.54	**20.69**	**13.52**	90.73	81.32	**74.09**	62.10	3.24	3.93	**7.25**	**8.99**	**8.12**
Typhoon-Audio	8B	8.72	14.17	**24.14**	17.52	10.67	**98.76**	**93.74**	48.83	**64.60**	5.62	**6.11**	6.34	8.73	7.54

- **Typhoon-Audio significantly outperformed other open-source models on Thai ASR, Thai SpokenQA, and Thai SpeechIF. It was often comparable to Gemini-1.5-Pro for Thai tasks.**
- **Effective Instruction Following: Excelled in SpeechIF for both English and Thai, outperforming Gemini-1.5-Pro in their evaluation. Also strong on ComplexIF.**

Key Results & Findings

TABLE V
SFT RESULTS ON THAI TASKS AND ENGLISH COMPLEXIF. *ASR IS EVAL ON SUBSET-1K OF CV17. †AVERAGE OF QUAL AND FORMAT

Experiment	#Ex	ASR*↓	Th2En↑	SpQA↑	SpIF↑	CxIF†↑
Pre-trained	-	13.52	0.00	28.33	1.12	1.41
SFT-v1: 100% En-Prompt	600K	80.86	6.01	36.88	1.48	6.35
SFT-v2: 10% Th-Prompt	200K	16.80	0.00	35.26	3.72	5.08
+ QA	220K	16.93	0.02	46.82	4.29	5.33
+ QA + Trns	240K	18.33	21.53	44.93	4.25	5.97
+ 2*QA + Trns + MCQ	300K	19.84	22.04	61.63	4.60	6.31
SFT-v3: scaled-v2+SpIF	620K	19.07	23.77	62.79	6.32	6.45
+ ASR (SelfInstruct-Th)	640K	16.89	24.14	64.60	6.11	7.54

"A single model can achieve strong performance across diverse audio tasks and multiple languages if trained with the right data mixture."

Applying Paper's Insights to Broader Audio Challenges

Generalizing the "Low-Resource" Approach

1. **Unified Audio-LLM Architecture:** The idea of an audio encoder + adapter + LLM is a powerful template for building systems that need to reason about audio content.

2. **Instruction Following as a Paradigm:** Frame your audio analysis task as an "instruction" to the model.
   - Instead of building separate classifiers, ask the model: "Does this audio segment contain evidence of X?", "Summarize the main points discussed by speaker Y in this meeting audio."

3. **Strategic Data Augmentation & Synthesis:** If you lack real data for your specific niche:
   - Like SpeechIF Type 2: Use LLMs + TTS to synthesize spoken instructions/queries relevant to your domain.
   - Like AudioCaps-Gemini: Use LLMs to augment existing descriptions or create detailed scenarios.

4. **Curated Fine-Tuning Data:** Even a small, high-quality SFT dataset focused on your target domain can significantly boost performance of a pre-trained foundation model.

Data & Prompting Strategies for Your Own Audio Challenges

Fine-Tuning Open-source LLMs

- Specific questions that your model should handle?
- Generate the data based on what you need
- Translate/adapt existing textual instruction dataset

Using LLM APIs

- Prompt Engineering
- Few-Shot Examples (2-5 examples)
- Chain-of-Thought Prompting for complex reasoning
- Provide relevant knowledge within the prompt

Key Takeaways

- **Audio Understanding is Evolving Fast**
  - Moving from basic transcription to deep, instruction-based interaction with audio.
- **Audio LLMs & Smart Data are Driving Progress**
  - Integrated models and strategic data use are key to powerful audio analysis.
- **The Future: More Capable, Multimodal Audio Models**
  - Expect versatile models that handle audio, text, and video together.
- **Multiple Paths to Solutions:**
  - Direct Audio LLMs are promising, but pipeline methods (ASR -> Text -> LLM) are also effective options.

TYPHOON | SCB 10X

Be Part of the Open-Source LLM Revolution!

Connect & Collaborate

- Join our Discord community
- Chat with us today in person!

Build & Experiment

- Access our models in Hugging Face
- Create your own LLM application

Start Here: opentyphoon.ai

Do Your Best Here & Get Ready for The Upcoming Hack Opportunities

Regional Hackathon

- Southeast Asia AI Hackathon by AI Singapore and Country Partner (SCB 10X is a main Thailand partner)
- Using regional/local open-source LLMS such as Typhoon

Typhoon Community Events and Hackathon

- Q3-Q4 community events
- Typhoon Hackathon by early 2025