

000 001 002 003 004 005 006 007 008 009 010 011 012 SEA-SPEECHBENCH: A LARGE-SCALE MULTITASK BENCHMARK FOR SPEECH UNDERSTANDING ACROSS SOUTHEAST ASIA

013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
Anonymous authors

Paper under double-blind review

ABSTRACT

The rapid advancement of audio and multimodal large language models has unlocked transformative speech understanding capabilities, yet evaluation frameworks remain predominantly English-centric, leaving Southeast Asian (SEA) languages critically underrepresented. We introduce SEA-SPEECHBENCH, the first large-scale multitask benchmark that evaluates speech understanding in 11 SEA languages through more than 97,000 samples and 597 hours of curated audio data. Our benchmark comprises 9 diverse tasks across 3 categories: speech processing (automatic speech recognition, speech translation, spoken question answering), paralinguistic analysis (emotion, gender, age, speaker recognition), and temporal understanding, a novel dimension featuring timestamped content queries and temporal localization within extended audio sequences up to 3 minutes. We implement multilingual prompting in both native SEA languages and English to reflect user interactions with audio-language models. Evaluation of leading open-source and proprietary systems reveals marked performance gaps. Across all models, performance remains underwhelming on temporal reasoning, emotion recognition, and speech translation, with most scores falling below 20. Prompting in low-resource languages such as Burmese, Lao, Tamil, and Khmer lag behind English by over 5%. Our findings expose critical model limitations and underscore the need for inclusive model development. We will release datasets and the evaluation framework upon paper publication to facilitate reproducible benchmarking.

1 INTRODUCTION

Recent advancement in audio large language models (AudioLLMs) has led to transformative applications in voice assistants, transcription, accessibility technologies, and multimodal reasoning (Wu et al., 2024; Gemini Team, 2025; Zhang et al., 2023). Despite these advances, research in speech understanding has been disproportionately concentrated on high-resource languages, particularly English and a small number of European or East Asian languages (Yang et al., 2021; Wang et al., 2021; Bu et al., 2017). While recent benchmarking efforts (Sakshi et al., 2024; Wang et al., 2024; Yang et al., 2024) have made significant strides in evaluating audio-language models across diverse tasks and modalities, they universally overlook Southeast Asian (SEA) languages, leaving an entire linguistic region underexplored despite representing over 650 million speakers worldwide.

Developing comprehensive benchmarks for SEA languages also presents unique technical challenges. The region’s speech landscape is characterized by extraordinary linguistic diversity, rich tonal and phonetic structures, and substantial resource disparities across languages: factors that create evaluation complexities absent from English-centric benchmarks. Many SEA languages operate in low-resource contexts with limited annotated data and sparse digital representation, making robust evaluation both methodologically challenging and critically important for equitable technological development. While recent initiatives, such as MERaLiON (MERaLiON Team, 2024) which targets Singapore’s multilingual context, Typhoon2-Audio (Pipatanakul et al., 2024) which focuses on Thai, and SeaLLMs-Audio (Liu et al., 2025) which extends capabilities to selected SEA languages, have begun to build general-purpose speech-language models for the SEA region, these efforts remain limited in both scope and linguistic coverage. Crucially, they lack comprehensive evaluation frameworks necessary to systematically assess capabilities across the full spectrum of

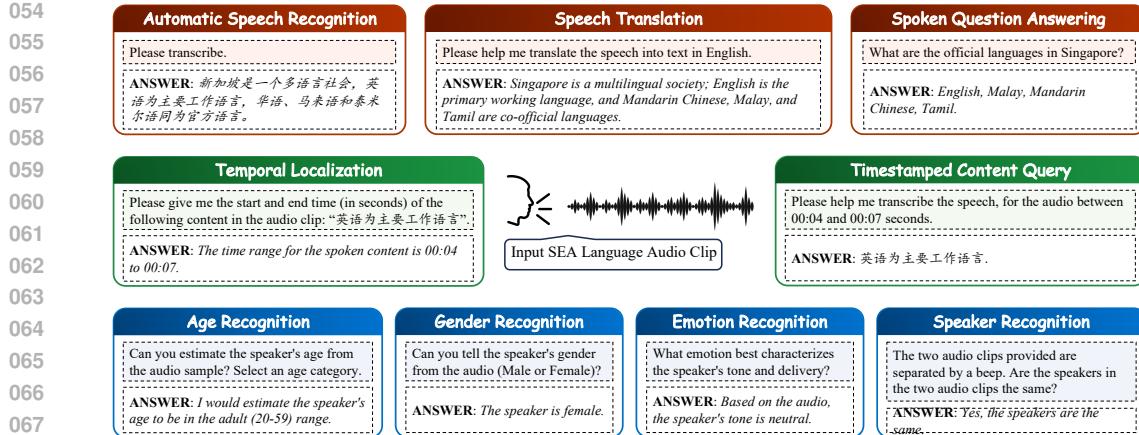


Figure 1: Overview of SEA-SPEECHBENCH task suite. The suite covers nine tasks in three categories. ■ **Speech processing**, ■ **Paralinguistic**, and ■ **Temporal understanding**.

Southeast Asian speech understanding tasks. Meanwhile, fragmented data collection efforts across research groups have yielded heterogeneous SEA datasets (Lovenia & et al., 2024; Pham et al., 2023; Bustamin et al., 2024; Magic Data Technology, 2025), but without unified frameworks for task definitions, prompting, and normalization, standardized comparison remains difficult.

In this work, we introduce the first ever comprehensive benchmark for speech understanding in Southeast Asian languages, designed to evaluate the capabilities of general-purpose speech-text LLMs. We focus on the official languages of Southeast Asian countries, as summarized in Table 1, to ensure broad regional coverage and practical relevance.

Table 1: Southeast Asian countries and their official language(s). Note: Several countries recognize additional regional or minority languages at sub-national levels; this table lists state-level official languages corresponding to the language codes used in our benchmark.

Country	Official language(s)	ISO 639-1 Code(s)
Singapore	English, Malay, Mandarin Chinese, Tamil	en, ms, zh, ta
Malaysia	Malay (Bahasa Melayu)	ms
Indonesia	Indonesian (Bahasa Indonesia)	id
Philippines	Filipino / Tagalog, English	tl, en
Thailand	Thai	th
Cambodia	Khmer (Cambodian)	km
Lao PDR	Lao	lo
Myanmar	Burmese (Myanmar)	my
Vietnam	Vietnamese	vi

Our benchmark encompasses 11 SEA languages with over 597 hours of audio data, curated from existing sources and synthesized into new tasks and datasets through systematic processing. The benchmark spans 9 diverse tasks across 3 broad categories: speech processing, paralinguistics, and temporal understanding. We introduce two temporal understanding tasks that evaluate models' capacity for temporal reasoning and localization within audio streams, addressing a previously unexplored dimension in audio LLM evaluation. These tasks test models' ability to navigate time-dependent information and extract content from specific temporal locations. The complete task suite is illustrated in Figure 1. To better reflect authentic usage scenarios, we evaluate each task using both English and native language text prompts.

The main contributions of this paper are threefold. First, we present SEA-SPEECHBENCH, the first ever large-scale multitask benchmark that systematically evaluates speech processing, paralinguistic analysis, and temporal understanding across Southeast Asian languages. It comprises a total of 99 evaluation sets with more than 97,000 audio samples. Second, we introduce temporal understanding tasks that assess models' ability to navigate and reason about time-dependent information in extended audio sequences. Finally, we conduct a comprehensive evaluation of both open-source

108 and proprietary models, offering critical insights into their strengths, limitations, and areas requiring
 109 further research.

112 2 RELATED WORKS

114 **Audio/Multimodal LLMs.** Recent language models have evolved to process spoken audio
 115 alongside text through diverse architectural approaches. Early work established key paradigms:
 116 alignment-based models (Tang et al., 2024; Zhang et al., 2023; Wu et al., 2023) connect speech en-
 117 decoders to LLMs via lightweight adaptors, while unified decoders (Rubenstein et al., 2023; Nguyen
 118 et al., 2024) share token spaces for joint speech-text modeling. Contemporary systems demon-
 119 strate varied innovations: Phi-4 (Microsoft, 2025) employs mixture-of-LoRAs for multimodality,
 120 Qwen2.5-Omni (Xu et al., 2025) advances temporal understanding through time-aligned position
 121 encoding, and Voxtral (Voxtral Team, 2025) handles extended recordings up to 40 minutes without
 122 separate ASR. Commercial systems like GPT-4o (OpenAI, 2024) and Gemini (Gemini Team, 2025)
 123 provide end-to-end audio understanding within unified frameworks. For Southeast Asian languages,
 124 MERA LiON (MERA LiON Team, 2024) targets Singapore’s multilingual context while SeaLLMs-
 125 Audio (Liu et al., 2025) extends capabilities to five major SEA languages. However, these efforts
 126 remain limited in scope, overlooking lower-resource languages and lacking comprehensive evalua-
 127 tion frameworks for the region’s full linguistic diversity.

128 **Audio/Multimodal LLM Benchmarks.** Audio-language evaluation has evolved from basic
 129 speech recognition to comprehensive multimodal assessment. While early efforts like SU-
 130 PERB (Huang et al., 2024) aggregated speech tasks universally, newer benchmarks (Wang et al.,
 131 2024; Yang et al., 2024) emphasize instruction-following across diverse audio types. Advanced
 132 benchmarks (Sakshi et al., 2024; Kumar et al., 2025) introduce complex reasoning with extended
 133 audio and multi-stream processing, while specialized evaluations target instruction-following (Gao
 134 et al., 2025) and domain-specific tasks (Ma et al., 2025). However, existing benchmarks remain pre-
 135 dominantly English-centric with minimal Southeast Asian language coverage, creating a significant
 136 evaluation gap for low-resource linguistic contexts. This bias fundamentally limits understanding
 137 of model performance across the global linguistic landscape where inclusive deployment is most
 138 critically needed.

139 3 SEA-SPEECHBENCH: TASK, DATASET AND EVALUATION SUITE

141 3.1 TASK SUITE

143 SEA-SPEECHBENCH comprises 9 core tasks across 3 categories: *speech processing*, *paralinguistic*
 144 *analysis*, and *temporal understanding*. All tasks require models to produce textual responses given
 145 an audio input and a text query.

146 *Speech processing* covers three fundamental capabilities: Automatic Speech Recognition (ASR),
 147 Speech Translation (ST) from Southeast Asian languages to English, and Spoken Question Answer-
 148 ing (SQA) based on SEA speech inputs.

149 *Paralinguistic analysis* examines vocal cues beyond linguistic content. It includes four tasks: Emo-
 150 tion Recognition (ER), which classifies emotional states; Gender Recognition (GR), which predicts
 151 gender from voice characteristics; Age Recognition (AgeR), which categorizes speakers as teens
 152 (10–19), adults (20–59), or seniors (60–100); and Speaker Recognition (SpkR), which determines
 153 whether two clips belong to the same speaker.

154 *Temporal understanding* introduces two novel tasks designed for extended audio, motivated by “skip
 155 to the content” and “what’s said at this time” use cases. Timestamped Content Query (TCQ) requires
 156 extracting content within a specified interval $[t_s, t_e]$, testing temporal grounding and localized re-
 157 trieval. Temporal Localization (TLoc) asks models to predict the exact time span $\hat{y} = [\hat{t}_s, \hat{t}_e]$ where
 158 queried information appears, evaluating boundary detection and alignment. Together, these tasks
 159 formalize time-referenced retrieval and support fine-grained navigation across recordings.

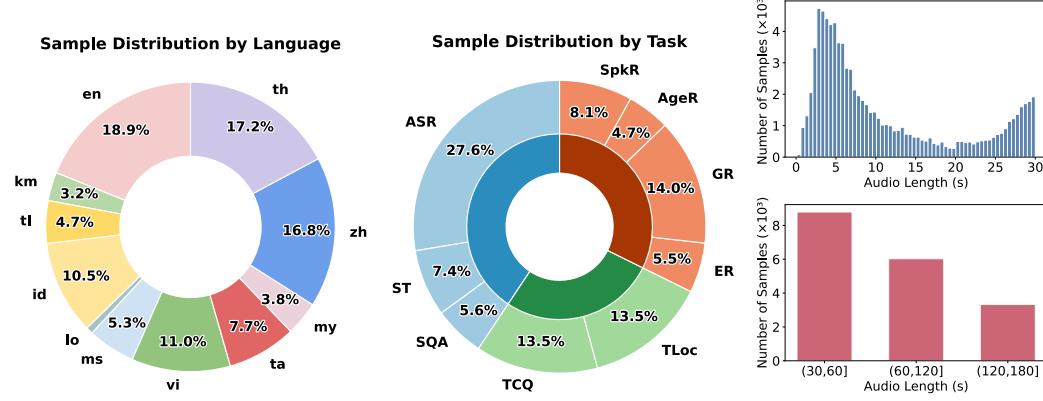
161 The first two categories use short clips (≤ 30 s) to align with current model input limits. As real-
 162 world audio applications increasingly involve longer recordings where users require temporal nav-

162
163 Table 2: Summary of curated datasets for SEA-SPEECHBENCH. For multilingual datasets, each
164 language-specific sub-dataset is counted separately in #Datasets.
165

Task	Languages	#Datasets	#Samples	Total L (h)	Min L (s)	Max L (s)
ASR	en tl id km lo ms my ta vi zh th	33	26,863	52.88	0.47	30.00
ST	en tl id km lo ms my vi th	9	7,189	26.46	3.06	30.00
SQA	en zh id th vi	7	5,462	40.57	20.00	30.00
ER	zh id th en ta	7	5,356	5.33	0.12	29.86
GR	zh id th en ta vi km my	16	13,599	22.02	0.12	29.90
AgeR	zh th en ta vi	5	4,608	6.55	0.58	20.78
SpkR	zh th en ta vi my	8	7,827	19.72	2.10	30.38
TCQ	zh en th id vi	7	13,145	211.98	20.00	180.00
TLoc	zh en th id vi	7	13,145	211.98	20.00	180.00
Total	-	99	97,194	597.49	-	-

180
181
182 igation, SEA-SPEECHBENCH introduces the temporal understanding tasks to assess model ability
183 to perform reasoning and localization over extended sequences of up to 3 minutes.
184

185 3.2 DATA CURATION



199 (a) Sample distribution by language. Colors match Table 2 language tags.
200

(b) Sample distribution by task.

(c) Audio duration distribution overview.

201 Figure 2: SEA-SPEECHBENCH composition: (a) language distribution by audio hours, (b) task
202 distribution by sample count, and (c) audio duration distribution. (Upper: short audios (≤ 30 s) with
203 fine-grained histogram. Lower: long audios (> 30 s) aggregated into three duration ranges.)
204

205 **Dataset Statistics:** As shown in Table 2, SEA-SPEECHBENCH is a comprehensive benchmark
206 comprising of over 97,000 samples across 9 tasks and 11 Southeast Asian languages.
207

208 Traditional speech processing tasks span the broadest linguistic coverage with standard 30-second
209 clips, while our novel temporal understanding tasks focus on 5 key languages with extended record-
210 ings up to 180 seconds. Detailed information for each source dataset is provided in Appendix A.

211 **Language and Task Distribution:** Figure 2 reveals the linguistic and task composition of our
212 benchmark. The language distribution (Figure 2a) demonstrates substantial coverage across South-
213 east Asian languages, with English (18.9%), Thai (17.2%) and Chinese (16.8%) representing the
214 largest segments, followed by Vietnamese (11.0%) and Indonesian (10.5%). We also include low-
215 resource languages such as Khmer and Lao to ensure representation of the region's full linguistic
spectrum. Figure 2b shows coverage across evaluation task families.

216 Audio Length Distribution: We cover both short and long audios to enable more comprehensive
 217 and realistic evaluation. As shown in Figure 2c, our benchmark encompasses a broad distribution
 218 of utterances spanning 0–30 seconds, with natural concentration around sentence-length segments
 219 for traditional tasks. For temporal understanding evaluation, we systematically curate extended
 220 recordings through strategic segmentation of long-form datasets, creating stratified duration bins of
 221 30–60s, 60–120s, and 120–180s, totaling over 10,000 samples for long audios.

222 We applied uniform data processing across all datasets and synthesized additional evaluation sets for
 223 tasks lacking suitable existing data. Processing and synthesis details are provided in Appendix B.

225 3.3 METRICS

226 Table 3 presents each task’s output format and primary evaluation metric. For ASR and TCQ, where
 227 the expected outputs are text transcripts, we employ script-dependent error rates: Character Error
 228 Rate (CER) for languages without explicit word boundaries (Chinese (zh), Thai (th), Khmer (km),
 229 Lao (lo), and Burmese (my)), and Word Error Rate (WER) for the other space-delimited languages.
 230 For ST, we use the BLEU score to measure n-gram overlap between generated translations and
 231 reference texts. For TLoc, given a actual time span $y = [t_s, t_e]$ and a prediction $\hat{y} = [\hat{t}_s, \hat{t}_e]$, define
 232 the intersection $I = [\min(\hat{t}_e, t_e) - \max(\hat{t}_s, t_s)]_+$ with $[x]_+ = \max(x, 0)$. Coverage and purity are
 233 $C = \frac{I}{t_e - t_s}$ and $P = \frac{I}{\hat{t}_e - \hat{t}_s}$, and the metric is $F1 = \frac{2CP}{C+P}$ for $C + P > 0$ (otherwise $F1 = 0$).

234 For the other tasks, due to the inherent variability of free-form outputs, we employ an LLM-based
 235 judge for consistent evaluation, adapted based on the methodology proposed in Wang et al. (2024).
 236 For SQA, the judge assigns a 0–5 quality score; we report this as a percentage via a linear scaling
 237 $s_{SQA} = 20 \times \text{score}$, $\text{score} \in \mathbb{Z}_{[0,5]}$. For paralinguistic tasks (ER, GR, AgeR, SpkR), the judge
 238 canonicalizes outputs and renders binary correctness, reported as accuracy in percent. Detailed
 239 judging prompts are provided in the Appendix F.

240 To enable aggregation across heterogeneous metrics, we transform error rates WER/CER into
 241 higher-is-better scores on the range $[0, 1]$ via the monotone reciprocal mapping: $s_{WER} = \frac{1}{1+WER}$,
 242 $s_{CER} = \frac{1}{1+CER}$. This transformation preserves relative ordering while ensuring well-defined
 243 scores even when excessive insertions yield $WER/CER > 1$. All remaining metrics in our
 244 evaluation framework are inherently higher-is-better, requiring no additional transformation. For
 245 presentation clarity, we report all metrics as percentages throughout our analysis. This aggregation
 246 facilitates direct comparison of overall model performance across tasks. Nonetheless, **aggregated**
 247 **scores should be viewed as a high-level summary**, and readers are encouraged to consult the individual
 248 metrics for detailed performance insights.

249 Table 3: Outputs and metrics by task. For ASR and TCQ, error-based metrics (WER/CER) are
 250 transformed into higher-is-better scores for comparability. [†] *Judge-provided metrics*: scores are
 251 produced by a LLM serving as an external judge.

Task	Expected Output	Metric(s)
ASR	Text transcript	s_{WER}/s_{CER}
ST	Text translation	BLEU
SQA	Short textual answer	Scaled judge score s_{SQA} [†]
AgeR	Age bin $\in \{\text{teens, adults, seniors}\}$	
ER	Emotion label	
GR	Gender label $\in \{\text{male, female}\}$	Judge-based classification Acc [†]
SpkR	Speaker identity match or mismatch	
TCQ	Text transcript at given time	s_{WER}/s_{CER}
TLoc	Time span $[t_s, t_e]$	$F1$ score

264 4 EXPERIMENTS AND RESULTS

266 4.1 EXPERIMENTAL SETTING

268 **269 Evaluated Models.** We conduct extensive evaluation across several state-of-the-art open-source au-
 270 dio/multimodal LLMs with multiple size variants ranging from 2B to 10B parameters: MERaLiON-

2 (MERaLiON Team, 2024), SeaLLMs-audio (Liu et al., 2025), Phi-4-multimodal-instruct (Microsoft, 2025), Qwen2-Audio-Instruct (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), Gemma-3n-it (Gemma Team, 2025), Voxtral (Voxtral Team, 2025), and Kimi-audio (KimiTeam, 2025). All models are evaluated using their official released checkpoints with recommended inference configurations to ensure fair comparison. To establish comprehensive performance baselines, we also evaluate two leading commercial models: Gemini 2.0 Flash (Gemini Team, 2025) and GPT-4o (OpenAI, 2024). For GPT-4o, we employ the specialized Whisper-based transcription API for ASR tasks and the general audio understanding model for all other tasks, ensuring optimal performance.

Model as Judge. We leverage Gemma-3 27B Instruct (Gemma Team, 2025) as our evaluation judge to assign quality scores and correctness decisions for model outputs, chosen for its demonstrated reliability in canonicalizing free-form text responses across Southeast Asian languages. This approach ensures scalable and standardized evaluation across the diverse range of audio understanding tasks in our benchmark.

4.2 MAIN RESULTS

Table 4 presents our results grouped by task families. To allow fair comparison across models with different maximum input audio length capacities, this table focuses on clips shorter than 30 seconds. Results on >30 s audios for temporal reasoning tasks are discussed separately in Section 4.3.1. Comprehensive per-language results are provided in the Appendix C.

Proprietary vs. Open-Source Models. Proprietary models maintain a performance lead across tasks. Gemini 2.5 Flash achieves the highest overall scores across task categories, reflecting advantages in computational scale and proprietary training resources. Among open-source models, MERaLiON-2-10B emerges as the strongest performer, demonstrating speech processing capabilities that closely match GPT-4o’s performance while achieving the second-best performance among open-source models on paralinguistic evaluations.

Table 4: Model performance across tasks under English (ENG) and SEA prompts. Category-average scores (%) are unweighted macro-averages over constituent tasks.

Model	Size	Speech Processing								Temporal Understanding							
		ASR		ST		SQA		Average		TCQ (30s)		TLoc (30s)		Average			
		ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA
Gemma-3n-it	2B	33.23	28.53	8.97	8.72	65.93	65.91	36.24	35.20	12.01	11.49	11.82	11.89	11.92	11.69		
Qwen2.5-omni	3B	59.70	47.31	7.60	6.27	67.50	56.64	44.94	40.1	10.27	14.34	30.49	25.45	20.38	19.90		
MERaLiON-2	3B	70.92	69.93	7.56	7.42	60.08	62.47	46.19	46.61	17.32	17.83	18.82	17.47	18.07	17.65		
Voxtral	3B	41.32	40.82	19.98	18.15	74.33	76.03	45.21	45.00	17.42	17.68	17.85	16.47	17.63	17.08		
Gemma-3n-it	4B	44.04	30.46	10.98	13.58	71.85	74.51	42.29	39.52	12.89	12.70	13.25	13.98	13.07	13.34		
Phi-4	5.6B	30.37	22.76	3.04	0.32	58.59	53.49	30.67	25.52	4.61	7.31	12.97	13.51	8.79	10.41		
SeaLLMs-Audio	7B	60.72	57.32	10.74	10.11	75.20	63.16	47.27	47.54	15.44	15.92	11.57	12.94	13.50	14.43		
Qwen2-Audio-it	7B	51.28	45.45	4.54	3.20	59.20	60.41	38.34	36.35	15.23	15.64	33.30	31.41	24.27	23.52		
Qwen2.5-omni	7B	44.68	47.96	7.91	8.06	61.50	66.03	38.03	40.69	15.41	14.30	35.74	33.08	25.57	23.69		
Kimi-Audio	7B	17.09	18.88	3.36	7.71	63.85	62.13	28.10	29.57	6.67	8.42	14.49	12.96	10.58	10.69		
MERaLiON-2	10B	78.74	78.74	17.75	19.52	76.03	80.43	57.51	59.56	16.34	17.26	22.22	22.05	19.28	19.66		
Gemini 2.5 Flash	-	87.72	87.72	16.86	18.89	84.51	87.02	63.03	64.54	29.33	28.79	11.66	13.59	20.49	21.19		
GPT-4o	-	81.79	81.79	21.24	21.39	74.08	75.25	59.04	59.48	16.50	15.56	26.47	25.43	21.49	20.49		

(a) Results on speech processing tasks (ASR, ST, SQA) and temporal understanding tasks (≤ 30 s) (TCQ, TLoc).

Model	Size	Paralinguistic									
		AgeR		ER		GR		SpkR		Average	
		ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA	ENG	SEA
Gemma-3n-it	2B	65.91	38.76	12.21	13.17	27.71	14.65	42.35	39.32	37.05	26.47
Qwen2.5-omni	3B	50.50	34.47	13.45	9.95	36.12	20.19	28.59	17.79	32.16	20.60
MERaLiON-2	3B	68.20	46.46	23.99	18.73	48.59	46.35	39.75	33.71	45.13	36.31
Voxtral	3B	75.71	52.61	10.62	5.35	27.65	9.89	42.90	36.09	39.22	25.99
Gemma-3n-it	4B	69.68	58.74	12.46	13.93	53.36	27.23	39.88	39.72	43.84	34.90
Phi-4	5.6B	43.45	41.98	20.87	9.20	44.21	32.60	23.27	27.48	32.95	27.82
SeaLLMs-Audio	7B	63.16	11.64	12.34	9.17	54.97	43.66	53.59	34.37	46.01	24.71
Qwen2-Audio-it	7B	20.21	23.13	24.47	19.36	92.25	66.31	49.85	36.56	46.70	36.34
Qwen2.5-omni	7B	20.15	23.13	16.33	10.45	50.46	34.58	14.63	11.44	25.39	19.90
Kimi-Audio	7B	47.91	48.42	36.86	42.55	90.97	79.23	54.16	48.16	57.48	54.59
MERaLiON-2	10B	66.24	57.31	18.73	20.34	59.93	46.28	53.25	46.12	49.54	42.51
Gemini 2.5 Flash	-	76.00	79.40	19.50	16.79	92.50	90.06	71.13	71.75	64.78	64.50
GPT-4o	-	76.60	61.20	17.00	19.50	46.13	31.63	10.50	8.38	37.56	30.18

(b) Results on paralinguistic tasks (AgeR, ER, GR, SpkR).

324 **Task Specialization.** Within specific task categories, several models exhibit particular strengths:
 325 Qwen2.5-omni-7B delivers the best open-source temporal localization performance, while Voxtral
 326 and SeaLLMs-Audio excel in spoken question answering tasks. Kimi-Audio demonstrates robust
 327 paralinguistic capabilities among open-source systems. Further details appear in Section 4.3.2.

328 **Failure Cases Analysis.** Temporal understanding tasks universally present the greatest challenge.
 329 Even leading commercial models achieve only approximately 20% scoring on TCQ and TLoc, sub-
 330 stantially lower than traditional speech processing performance. Details of this task will be discussed
 331 separately in Section 4.3.1. Inadequate ST performance stems from two systematic failure modes.
 332 First, task confusion: many models treat ST as ASR, returning source-language transcripts instead
 333 of English translations. Second, target-language mismatch: outputs appear in non-English or code-
 334 switched text despite explicit English requirements in prompts. These failures persist across our
 335 standardized templates and bilingual prompting conditions, indicating weaknesses in cross-lingual
 336 semantic grounding. For ER, we observe negative emotion clustering: anger, disgust, fear, and
 337 sadness are frequently conflated, while low-arousal negative emotions are often collapsed to neu-
 338 tral. This reflects the small inter-class separations and fuzzy boundaries within the negative emotion
 339 family, compounded by limited training data for some languages, lead to elevated confusions. A
 340 striking anomaly emerges in GPT-4o’s notably weak performance on speaker recognition (8.38-
 341 10.50%), contrasting sharply with its competitive results across other paralinguistic tasks. This
 342 underperformance will be discussed in Section 4.3.4.

343 **Robustness to Multilingual Prompts.** SEA language prompts consistently underperform relative
 344 to English prompts. The SEA-ENG gap is largest for paralinguistic tasks, plausibly because strict
 345 classification scoring penalizes parsing errors and refusals caused by weaker instruction-following
 346 under SEA prompts, compounded by non-Latin script tokenization issues. Gemini 2.5 Flash and
 347 Kimi-Audio prove most robust to prompt language variation, while SeaLLMs-Audio and Qwen2.5-
 348 omni-3B are the least robust. By contrast, best performing ASR models Gemini 2.5 Flash and
 349 MERaLiON-2-10B exhibit near-identical performance under both prompt languages, suggesting
 350 ASR task is heavily trained and largely prompt-agnostic. We provide more detailed analysis of
 351 cross-linguistic performance patterns and their underlying linguistic determinants in Section 4.3.3.

352 4.3 ANALYSIS AND INSIGHTS

354 4.3.1 TEMPORAL REASONING IN AUDIO LLMs

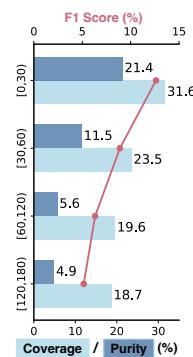
355 In this section, we provide detailed analysis of temporal understanding capabilities, stratifying per-
 356 formance across four duration bins: [0,30), [30,60), [60,120), [120,180) seconds as shown in Ta-
 357 ble 5a. Figure 5b presents coverage (C), purity (P), and F_1 of TLoc task across audio-duration
 358 ranges, using MERaLiON-2-10B as a case study.

360 Table 5: Temporal understanding performance by duration.

361 (a) TCQ and TLoc results (%) by duration, averaged across SEA and English
 362 prompts. A dash (–) indicates audio lengths for which the model is unable to perform
 363 inference.

364 (b) TLoc metrics for
 365 MERaLiON-2-10B.

366 Model	367 Size	368 TCQ				369 TLoc			
		370 0-30 s	371 30-60 s	372 60-120 s	373 120-180 s	374 0-30 s	375 30-60 s	376 60-120 s	377 120-180 s
SeaLLMs-Audio	7B	12.26	-	-	-	15.67	-	-	-
Qwen2-Audio-it	7B	15.44	-	-	-	32.35	-	-	-
Qwen2.5-omni	3B	11.97	10.82	-	-	27.97	17.02	-	-
Gemma-3n-it	2B	11.75	11.82	-	-	11.85	8.45	-	-
Gemma-3n-it	4B	12.79	12.04	-	-	13.61	8.95	-	-
Phi-4	5.6B	5.66	6.99	6.02	-	13.24	6.58	3.76	-
MERaLiON-2	3B	17.57	9.73	8.57	-	18.14	9.77	4.92	-
Qwen2.5-omni	7B	14.84	13.08	10.42	-	34.41	20.54	11.02	-
Kimi-Audio	7B	7.44	5.59	4.12	2.40	13.88	8.53	3.31	3.00
Voxtral	3B	17.55	12.19	8.31	5.30	17.16	9.61	3.51	2.33
MERaLiON-2	10B	16.79	11.27	6.52	5.43	22.14	12.63	6.50	5.25
Gemini 2.5 Flash	-	29.05	27.17	13.85	19.83	12.62	8.89	6.33	5.15
GPT-4o	-	16.01	12.12	11.62	10.08	25.95	18.15	8.90	5.44



378 **Metrics Analysis.** First, we observe systematic over-coverage in temporal understanding across
 379 models. In TCQ, models frequently produce content that extends beyond the queried time window.
 380 In TLoc, as demonstrated in Figure 5b, coverage consistently exceeds purity across all durations,
 381 which is a pattern we observe in most evaluated models. This asymmetry indicates weak temporal

boundary localization and alignment, reflecting a recall-seeking strategy that favors longer spans, and thus higher coverage, at the cost of precision. These motivate finer-grained temporal grounding, boundary-aware training objectives, and decoding constraints that penalize span over-coverage.

Constraints on Audio Length. As shown in Table 5a, only a select subset of models sustains inference availability across all duration bins. Among open models, Voxtral, Kimi-Audio, and MERA LiON-2-10B demonstrate consistent availability, while commercial models handle the full range. Meanwhile, both TCQ and TLoc exhibit performance degradation with increasing duration: errors accumulate over longer contexts, manifesting as boundary drift and truncation, exposing current architectural limits in context window, frame compression, and long-range memory.

These findings underscore that temporal grounding represents an unresolved challenge in current audio-language architectures, with this deficiency becoming critically pronounced in long-form audio contexts where existing architectures prove inadequate for practical deployment.

4.3.2 BEST-PERFORMING MODELS BY LANGUAGE AND TASK

To highlight model strengths across tasks and languages, we plot a winner map that marks, for each task-language pair, the best-performing **open-sourced** model in Figure 3. Each cell shows the top model and color-codes model identity. MERA LiON-2-10B establishes clear dominance in speech processing tasks, consistently achieving top performance across multiple languages including Indonesian, Vietnamese, and Filipino. Paralinguistic task leadership proves more distributed, with Kimi-Audio and Qwen2-Audio-Instruct alternately excelling in different linguistic contexts. Temporal understanding tasks reveal limited model coverage and inconsistent performance patterns, suggesting that model strengths are specialized rather than generalizable, with no single model demonstrating comprehensive temporal reasoning proficiency.

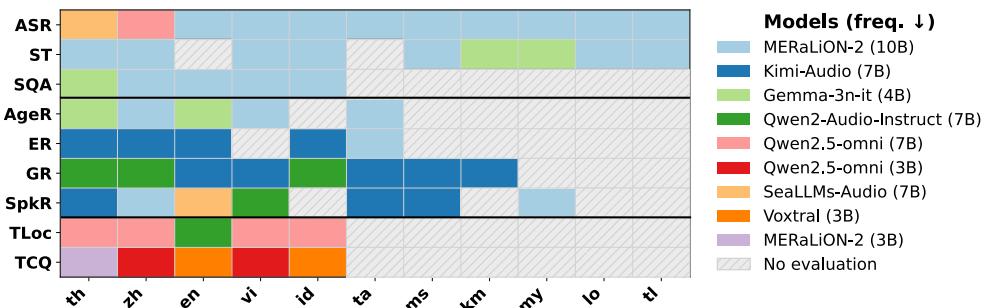


Figure 3: Winner map for open-sourced models across tasks and languages. Each cell marks the top model for a task–language pair; legend is ordered by overall win frequency.

4.3.3 EFFECT OF PROMPT LANGUAGE

We systematically investigate how prompt language choice: English versus native SEA language, affects model performance across our benchmark.

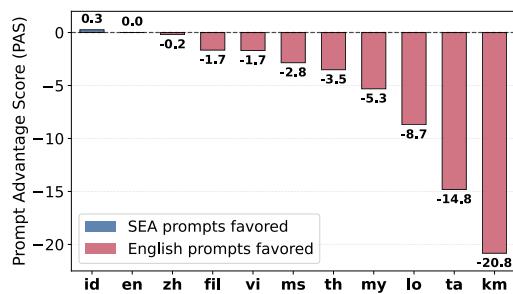
Figure 4 demonstrates a cross-linguistic hierarchy in prompt sensitivity. We define a Prompt Advantage Score (PAS) to quantify this effect, with its detailed formulation provided in Appendix C. Higher PAS values indicate stronger local language prompt advantage, while negative scores suggest English prompt superiority for that particular language. Indonesian (id) emerges as the sole language showing consistent local prompt advantage (+0.3), while English (en) and Chinese (zh) exhibit near-neutral behavior (0.0, -0.2). The remaining SEA languages display increasing English preference across two distinct clusters: moderate disadvantages for Filipino, Vietnamese, Malay, Thai, and Myanmar (ranging from -1.7 to -5.3), and severe English advantages for Lao, Tamil, and Khmer (-8.7 to -20.8). This variation correlates strongly with orthographic and computational factors that influence instruction parsing effectiveness:

Script Complexity and Tokenization. Non-Latin scripts create fundamental computational barriers. Languages like Thai, Lao, and Khmer lack clear word boundaries and employ complex grapheme clusters that disrupt standard tokenization processes. Abugida systems such as Myanmar

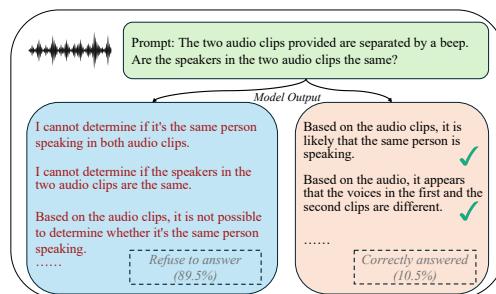
432 and Tamil further complicate parsing through character-level ambiguities. These structural challenges impair instruction processing, while English prompts leverage well-established tokenization patterns that avoid such complications.
 433
 434
 435

436 **Training Data Quality and Orthographic Consistency.** Corpus quality directly affects local
 437 prompt performance. Indonesian succeeds due to abundant Latin-script training data with consistent
 438 orthographic standards and imperative constructions, supporting reliable instruction following.
 439 In contrast, lower-resource languages suffer from limited, inconsistent training corpora marked by
 440 orthographic variations and dialectal diversity, making local prompts less reliable than standardized
 441 English alternatives.
 442

443 These results expose a critical deployment gap: when users issue prompts in Southeast Asian lan-
 444 guages, which is the natural interaction mode for regional populations, performance degrades sub-
 445 stantially compared to English-prompted evaluation. This asymmetry underscores the urgent need
 446 for multilingual instruction-tuning that aligns with authentic user interaction patterns throughout
 447 model development and evaluation.
 448



449
 450
 451
 452
 453
 454
 455
 456
 457 Figure 4: Cross-linguistic prompt sensitivity
 458 measured by Prompt Advantage Score (PAS).



459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470 Figure 5: Speaker recognition failure examples
 471 in GPT-4o responses.
 472
 473
 474

4.3.4 REFUSAL BEHAVIOR: CAUSES AND PREVALENCE

475 From Table 4, GPT-4o attains only 10.50% accuracy on speaker recognition. Inspecting the error
 476 breakdown shows that these “errors” are refusals rather than wrong predictions. As illustrated in
 477 Figure 5, GPT-4o refuses to answer 89.5% of the queries. When the model does answer for 10.5%
 478 of the queries, it is consistently correct (non-refusal accuracy = 100%). This pattern emerges on our
 479 self-constructed SpkR dataset, which is likely out-of-distribution for GPT-4o, suggesting limited
 480 task generalization. Concurrently, the model may adopt a conservative, uncertainty-aware strategy
 481 rather than making overconfident predictions. By contrast, Qwen2.5-Omni-7B also attains low SpkR
 482 performance with frequent refusals, but its accepted responses contain nontrivial mistakes, pointing
 483 to weaker calibration and label grounding rather than abstention alone.
 484
 485

5 CONCLUSION

486 We present SEA-SPEECHBENCH, the first comprehensive benchmark for evaluating speech under-
 487 standing across 11 Southeast Asian languages, comprising 97,000+ samples across 9 tasks in speech
 488 processing, paralinguistics, and temporal understanding. Our standardized framework enables re-
 489 producible, cross-linguistic comparisons through unified normalization and bilingual prompting.
 490

491 Evaluation of leading commercial and open-source systems exposes systematic weaknesses: per-
 492 formance collapses on long audio (temporal brittleness), English prompts consistently outperform
 493 native languages (linguistic inequity), and tasks such as temporal reasoning, emotion recognition,
 494 and speech translation remain far below usability thresholds. These findings underscore persistent
 495 scalability and generalization gaps. By surfacing these limitations, SEA-SPEECHBENCH seeks to
 496 establish a rigorous baseline for developing temporally robust, linguistically inclusive, and practi-
 497 cally deployable speech technologies for Southeast Asia’s diverse communities.
 498

486 REFERENCES
487

488 Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer,
489 Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice:
490 A massively-multilingual speech corpus, 2020. URL <https://arxiv.org/abs/1912.06670>.

491

492 Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin
493 speech corpus and a speech recognition baseline, 2017. URL <https://arxiv.org/abs/1709.05522>.

494

495 Anugrayani Bustamin, Andi M. Rizky, Elly Warni, Intan Sari Areni, and Indrabayu Indrabayu.
496 Indowavesentiment: Indonesian audio dataset for emotion classification. Mendeley Data, Version
497 1, 2024. URL <https://data.mendeley.com/datasets/j9ytfdz27/1>. CC BY
498 4.0 license; 300 audio files across five emotion categories.

499

500 Ananlada Chotimongkol, Kwanchiva Saykhum, Patcharika Chootrakool, Nattanun Thatphithakkul,
501 and Chai Wutiwiwatchai. Lotus-bn: A thai broadcast news corpus and its research applications.
502 In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, pp.
503 44–50, 2009. doi: 10.1109/ICSDA.2009.5278377.

504

505 Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv,
506 Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv
507 preprint arXiv:2407.10759*, 2024.

508

509 Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason
510 Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal repre-
511 sentations of speech. *arXiv preprint arXiv:2205.12446*, 2022. URL <https://arxiv.org/abs/2205.12446>.

512

513 Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. Ifeval-audio: Benchmark-
514 ing instruction-following capability in audio-based large language models. *arXiv preprint
515 arXiv:2505.16774*, 2025.

516

517 Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long
518 context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.

519

520 Gemma Team. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2504.18425>.

521

522 Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander
523 Gutkin, Isin Demirsahin, Cibu Johnny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsri-
524 sawat. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam,
525 Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language
526 Resources and Evaluation Conference (LREC)*, pp. 6494–6503, Marseille, France, May 2020.
527 European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/2020.lrec-1.800>.

528

529 Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu,
530 Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. Dynamic-superb: Towards
531 a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In
532 *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing
533 (ICASSP)*, pp. 12136–12140. IEEE, 2024.

534

535 Yerbolat Khassanov, Zhiping Zeng, Van Tung Pham, Haihua Xu, and Eng Siong Chng. Enriching
536 rare word representations in neural language models by embedding matrix augmentation. In
537 *Interspeech 2019*, pp. 3505–3509. ISCA, September 2019. doi: 10.21437/interspeech.2019-1858.
538 URL <http://dx.doi.org/10.21437/Interspeech.2019-1858>.

539

KimiTeam. Kimi-audio technical report, 2025. URL <https://arxiv.org/abs/2504.18425>.

540 Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand,
 541 Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A chal-
 542 lenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv*
 543 *preprint arXiv:2508.13992*, 2025.

544 Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel
 545 Whitenack. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream
 546 tasks, 2022. URL <https://arxiv.org/abs/2210.14712>

547 Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watan-
 548 abe. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech
 549 Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

550 Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and
 551 Wenxuan Zhang. Seallms-audio: Large audio-language models for southeast asia. <https://github.com/DAMO-NLP-SG/SeaLLMs-Audio>, 2025.

552 Holy Lovenia and et al. Seacrowd: A multilingual multimodal data hub and benchmark suite for
 553 southeast asian languages. *arXiv preprint arXiv: 2406.10118*, 2024.

554 Yinghao Ma, Siyou Li, Juntao Yu, Emmanouil Benetos, and Akira Maezawa. Cmi-bench:
 555 A comprehensive benchmark for evaluating music instruction following. *arXiv preprint
 556 arXiv:2506.12285*, 2025.

557 Magic Data Technology. Asr-sgpccsc. <https://magichub.com/datasets/singaporean-chinese-conversational-speech-corpus/>

558 Magic Data Technology. Asr-sfdusc: A scripted filipino daily-use speech corpus. MagicHub,
 559 2023. URL <https://magichub.com/datasets/filipino-scripted-speech-corpus-daily-use-sentence/>. 4.58 hours of transcribed Filipino scripted speech fo-
 560 cusing on daily use sentences; 4,073 utterances by 10 speakers. Licensed under CC BY-NC-ND
 561 4.0.

562 Magic Data Technology. Asr-malcsc: Malay conversational speech corpus. Online: MagicHub,
 563 2025. URL <https://magichub.com/datasets/malay-conversational-speech-corpus/>. 5 hours of transcribed Malay conversational speech, 10 conversations
 564 between 5 speaker pairs, licensed under CC BY-NC-ND 4.0.

565 MERaLiON Team. Meralion-audiollm: Bridging audio and language with large language models,
 566 2024. URL <https://arxiv.org/abs/2412.09818>

567 Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via
 568 mixture-of-loras, 2025. URL <https://arxiv.org/abs/2503.01743>

569 Myanmar Innovative Group. mig-burmese-audio-transcription. <https://huggingface.co/datasets/Ko-Yin-Maung/mig-burmese-audio-transcription>, 2025.

570 Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri,
 571 Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary
 572 Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. Spirit lm:
 573 Interleaved spoken and written language model, 2024. URL <https://arxiv.org/abs/2402.05755>

574 Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin,
 575 Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson, and Alexander Gutkin. Burmese Speech
 576 Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to
 577 Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference
 578 (LREC)*, May 2020.

579 OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>

580 Anh Pham, Khanh Linh Tran, Linh Nguyen, Thanh Duy Cao, Phuc Phan, and Duong A Nguyen.
 581 Bud500: A comprehensive vietnamese asr dataset, 2024. URL <https://github.com/quocanh34/Bud500>

594 Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. Vietnam-Celeb:
 595 a large-scale dataset for Vietnamese speaker recognition. In *Proc. INTERSPEECH 2023*, pp.
 596 1918–1922, 2023. doi: 10.21437/Interspeech.2023-1989.

597

598 Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotendumrong, Surapon
 599 Nonesung, Teetouch Jaknamon, Parinthipat Pengpun, Pittawat Taveekitworachai, Adisai Na-
 600 Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. Ty-
 601 phoon 2: A family of open text and multimodal thai large language models, 2024. URL
 602 <https://arxiv.org/abs/2412.13702>

603 Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos,
 604 Félix de Chaumont Quirly, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Han-
 605 nnah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk,
 606 Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Ve-
 607 limirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang,
 608 Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can
 609 speak and listen, 2023. URL <https://arxiv.org/abs/2306.12925>.

610 S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ra-
 611 mani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio
 612 understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

613

614 Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De
 615 Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open
 616 Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese.
 617 In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Lan-
 618 guages (SLTU)*, pp. 66–70, Gurugram, India, August 2018. URL <https://dx.doi.org/10.21437/SLTU.2018-14>.

619

620 Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and
 621 Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL
 622 <https://arxiv.org/abs/2310.13289>

623 Thanushs25. tamil-audio-emotion-classification. <https://huggingface.co/datasets/Thanushs25/tamil-audio-emotion-classification> 2024.

624

625 Jubeerathan Thevakumar, Luxshan Thavarasa, Thanikan Sivatheepan, Sajeet Kugarajah, and
 626 Uthayasanker Thayasilam. Emota: A tamil emotional speech dataset. In Kengatharaiyer
 627 Sarveswaran, Ashwini Vaidya, Bal Krishna Bal, Sana Shams, and Surendrabikram Thapa (eds.),
 628 *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiP-
 629 SAL 2025)*, pp. 193–201, Abu Dhabi, UAE, January 2025. International Committee on Compu-
 630 tational Linguistics. URL <https://aclanthology.org/2025.chipsal-1.19/>.

631

632 Voxtral Team. Voxtral, 2025. URL <https://arxiv.org/abs/2507.13264>.

633

634 Hoang Long Vu, Phuong Tuan Dat, Pham Thao Nhi, Nguyen Song Hao, and Nguyen Thi
 635 Thu Trang. Voxvietnam: a large-scale multi-genre dataset for vietnamese speaker recognition.
 636 In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Process-
 637 ing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890124.

638

639 Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi
 640 Aw, and Nancy F Chen. Audiobench: A universal benchmark for audio large language models.
 641 *arXiv preprint arXiv:2406.16020*, 2024.

642

643 Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary
 644 Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech
 645 corpus for representation learning, semi-supervised learning and interpretation, 2021. URL
 646 <https://arxiv.org/abs/2101.00390>

647

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu,
 648 Bo Ren, Linquan Liu, and Yu Wu. On decoder-only architecture for speech-to-text and large
 649 language model integration, 2023. URL <https://arxiv.org/abs/2307.03917>

648 Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multi-
 649 modal llm, 2024. URL <https://arxiv.org/abs/2309.05519>

650

651 Li Xinjian, Takamichi Shinnosuke, Saeki Takaaki, Chen William, Shiota Sayaka, and Watanabe
 652 Shinji. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech
 653 Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

654 Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang
 655 Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical
 656 report. *arXiv preprint arXiv:2503.20215*, 2025.

657

658 Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun
 659 Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via
 660 generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

661 Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhota, Yist Y Lin,
 662 Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing
 663 universal performance benchmark. *Interspeech 2021*, 2021.

664

665 Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu.
 666 Speechgpt: Empowering large language models with intrinsic cross-modal conversational abil-
 667 ities, 2023. URL <https://arxiv.org/abs/2305.11000>.

668

669 Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou
 670 Li. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint
 671 arXiv:2205.10237*, 2022.

672

673 Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases
 674 and esd. *Speech Communication*, 137:1–18, 2022.

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701