iApp Technology · Technical Whitepaper WP-2026-16

Thai Speech-to-Text: Character Error Rate on Common Voice, FLEURS and TVSpeech

Revision 2: Chinda ASR Pro v2, release v2.2.6 · 9 November 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. Chinda ASR, iApp's Thai speech-to-text API, now answers with Pro v2. On the 3,539 clips of the Common Voice 17.0 Thai test set whose sentence and recording are in none of Pro v2's training data, its character error rate is **2.06%** (95% CI 1.89–2.23), so 97.94% of characters are right (97.95% before the spelling change of 9 November 2026); the Base tier gets 5.33% wrong on the same clips. On natural speech from video with dense jargon (TVSpeech), Pro v2 gets 7.28% wrong against Base's 18.50% (difference −11.22 points, CI −13.53 to −8.09); on formal Wikipedia sentences (FLEURS), 9.43% against 9.93%, a gain that is small but larger than chance (CI −0.96 to −0.05). Base and the first Pro tier, measured on 28 October 2026 and kept here as the baseline, end on 15 November 2026, when their endpoints start answering with Pro v2. Through the public API, Pro v2 answered a typical Common Voice clip in a median 0.68 s.

1 Introduction

Chinda ASR turns a recording of Thai speech into text, one time-stamped segment per speaker turn. Its first two tiers, until 14 November 2026, are Base (`/v3/store/speech/speech-to-text/base`, 1 IC per started minute), the speech recogniser alone, and Pro (`/pro`, 2 IC), which passes the recogniser's output through a second step that re-reads each part in context and revises it. Pro v2 (`/pro-v2`), available since 8 November 2026, is a new recogniser: it writes numbers as words, labels each speaker (`SPEAKER_00`, `SPEAKER_01`, . . .) and takes up to 8 hours of audio per request. It is free until 30 November 2026, then 0.003 IC per second of audio. From 15 November 2026 the Base and Pro endpoints answer with Pro v2.

The first edition of this paper (28 October 2026) measured Base on three public Thai test sets in full and Pro on samples, and found that the figures our documentation then stated for 2024 could be reproduced for Base but not for Pro. This revision adds Pro v2 on the same three sets in full, scored with the same method, and compares it with Base clip by clip.

2 Test Data

Common Voice 17.0 Thai, test split
11,042 short clips of crowd-sourced read speech, mostly everyday sentences, recorded by volunteers on their own devices [2]. Released into the public domain (CC0).

FLEURS Thai, test split
1,021 recordings of 349 Wikipedia sentences, read by non-professional speakers [3]. Licence CC BY 4.0. The sentences are longer and more formal, and contain names, numbers and English words.

TVSpeech
570 in-the-wild clips, 3.75 hours in total, cut from 18 videos of Thai channels on finance, technology and lifestyle, hand-transcribed and chosen for dense jargon and proper names [4]. Licence CC BY. Released as a test-only set in 2026.

What Pro v2 was trained on. Its training data includes other releases of Common Voice Thai. Because each release re-draws its splits, 7,503 of the 11,042 clips of the 17.0 test split (67.9%) carry a sentence or a recording that is in Pro v2's training data. Our headline figure is therefore measured on the other 3,539 clips; Section 4.1 reports both groups. No TVSpeech or FLEURS sentence or recording is in Pro v2's training data (checked against all of its training lists).

What Base was trained on. Base's recogniser was trained on the training and validation splits of Common Voice 11.0 Thai, the training split of FLEURS Thai and further public Thai corpora. 1,186 of the 17.0 test clips (10.7%) carry a sentence from the 11.0 training or validation data. The FLEURS test sentences are not in its training or development splits.

3 Method

Paths. Base and Pro: every clip was sent, as its original file, to the production speech service — the same instances that answer `api.iapp.co.th` — on 28 October 2026, one request at a time. Pro was run on seeded samples (seed 1337) of 300 Common Voice, 120 FLEURS and 80 TVSpeech clips, because its correction step is rate-limited and shared with customers. Pro v2: the Chinda ASR team decoded every clip of the three sets on 9 November 2026 with the production service's own code, model files and decoding settings, and applied the service's spelling step of release v2.2.6; these answers did not pass through the gateway. To check that they are what a customer receives, we sent seeded samples of the same clips through the public API (Section 7).

Metric. Character error rate (CER) is the number of substituted, deleted and inserted characters divided by the number of reference characters, summed over a set. Thai is written without spaces between words and transcribers place spaces inconsistently, so spaces are ignored. Before scoring, the reference and the service's text pass through the same normalisation: Unicode NFC; zero-width characters removed; the two-character spellings of sara ae and sara am replaced by the single characters; Thai digits mapped to Arabic digits; punctuation and symbols removed (the repetition mark is kept); Latin letters lower-cased. Numbers are not spelled out, so a number written in words in one text and in digits in the other counts as errors; we also report CER with every number written as words in both texts. Secondary measures: word error rate (WER) after segmenting both texts with the PyThaiNLP *newmm* dictionary segmenter, and CER with minimal normalisation (NFC and spaces removed only). A request that failed, or returned no text, counts as deleting every character.

Statistics. 95% confidence intervals are percentile bootstrap intervals (2,000 resamples, seed 1337) over clips, except on TVSpeech, where we resample its 18 source videos: clips cut from one video share a speaker, a topic and a recording, and one video holds 115 of the 570 clips. Tiers are compared on the same clips, and the interval of the difference is paired.

4 Results

Table 1: Character error rate (%), lower is better: Pro v2 (v2.2.6) and Base on the same clips, full sets. Δ is Pro v2 minus Base, in points, with paired 95% CIs; TVSpeech intervals resample its source videos.

Set	Clips	Base	Pro v2 [95% CI]	Δ [95% CI]
Common Voice 17.0 test, not in Pro v2 training	3,539	5.33	2.06 [1.89, 2.23]	−3.28 [−3.57, −2.99]
Common Voice 17.0 test, all clips	11,042	5.21	1.89 [1.80, 1.98]	−3.32 [−3.51, −3.14]
FLEURS test	1,021	9.93	9.43 [8.57, 10.37]	−0.50 [−0.96, −0.05]
TVSpeech	570	18.50	7.28 [6.21, 8.63]	−11.22 [−13.53, −8.09]

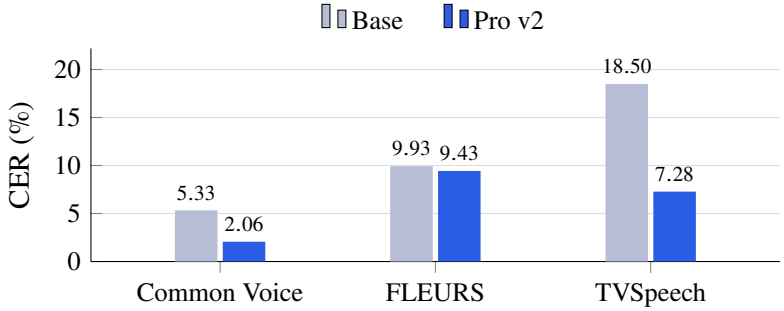


Figure 1: Character error rate, Base and Pro v2 on the same clips. Common Voice: the 3,539 clips not in Pro v2's training data.

Common Voice. On the 3,539 held-out clips Pro v2 gets 2.06% of characters wrong and Base 5.33%; Pro v2 was better on 1,228 of these clips and worse on 193. 76.9% of them come back without a single character error (Base: 55.4%). On all 11,042 clips, including those in its training data, Pro v2 measures 1.89% (98.11% correct). Pro v2 returned text for every clip; Base returned none for 37 short clips.

FLEURS. The references write most numbers in digits and Pro v2 writes them as words, which the main metric counts as errors. With every number written as words in both texts, Pro v2 measures 7.01% and Base 8.63%; the main metric's smaller gap (9.43% against 9.93%) is largely this convention. Part of every FLEURS figure is also the references covering only part of what some speakers read, which adds insertions for every tier.

TVSpeech. The largest gain is on natural speech: 7.28% against 18.50%, better on 505 of the 570 clips and worse on 57. With numbers compared as words, 7.21% against 16.40%. English terms remain a matter of convention: TVSpeech writes *think tank* in Latin script, and the service writes English words in Thai script.

Spelling. From release v2.2.6 (9 November 2026) Pro v2 writes words that are often spelled colloquially in their standard form (for example the question particle *mai* and the pronoun *khao*) and writes a repeated word out in full rather than with the repetition mark; a word the customer lists in the `vocabulary` field keeps the customer's spelling. A reference that uses the colloquial form counts the difference as errors. On the 3,539 held-out Common Voice clips the change moved accuracy from 97.95% to 97.94%, inside the confidence interval.

Base and the first Pro tier. Table 2 keeps the first edition's measurement. Pro's correction step lowered the error on all three samples, by more than chance alone would explain only on TVSpeech.

Table 2: The first edition (28 October 2026): Base on the full sets, and Base and Pro on the same seeded samples. CER (%), lower is better; Δ is Pro minus Base, in points.

	Full set, Base			Seeded sample			
Set	Clips	CER [95% CI]	WER	Clips	Base	Pro	Δ [95% CI]
Common Voice 17.0 test	11,042	5.21 [5.02, 5.41]	15.29	300	5.59	5.11	−0.48 [−1.09, +0.20]
FLEURS test	1,021	9.93 [9.17, 10.79]	25.76	120	9.79	7.36	−2.43 [−4.66, +0.17]
TVSpeech	570	18.50 [15.58, 21.44]	29.02	80	19.21	17.69	−1.53 [−2.54, −0.48]

4.1 Held-out check

Table 3: CER (%) on Common Voice clips by whether the tier could have seen them in training.

Tier	Clips	Number	CER [95% CI]
Pro v2	17.0 test, sentence and recording not in its training data	3,539	2.06 [1.89, 2.23]
Pro v2	17.0 test, sentence or recording in its training data	7,503	1.81 [1.70, 1.92]
Base	17.0 test, sentence not in the 11.0 training data	9,856	5.27 [5.05, 5.48]
Base	17.0 test, sentence in the 11.0 training data	1,186	4.66 [4.17, 5.17]
Base	17.0 train, seeded sample	1,000	8.53 [7.53, 9.56]

Pro v2 is slightly more accurate on the clips it could have seen (1.81% against 2.06%), so the held-out figure is the one we publish. The difference is small, which is not the pattern of a model that has memorised its data. Base shows the same pattern (4.66% against 5.27%), and on a seeded sample of 1,000 clips from the 17.0 training split it measures 8.53%, higher than on the test set; that split holds more poor recordings, so it is a weak comparison.

4.2 Speakers

Table 4: CER (%) on the 3,539 Common Voice clips not in Pro v2's training data, by the speaker's self-reported gender and age; groups of 50 clips or more.

Group	Clips	Base	Pro v2
Male	160	5.14	2.29
Female	120	5.02	1.77
Teens	65	5.65	2.40
Twenties	147	5.02	2.38
Thirties	53	5.19	0.70
Not stated (gender)	3,259	5.36	2.06

5 The 2024 Figures

Before the first edition, our documentation stated 5.87% for Base and 2.19% for Pro on the 11,042 Common Voice 17.0 Thai test clips, measured in April 2024 with an unrecorded normalisation. The first edition reproduced the Base figure (5.21%; 5.52% with minimal normalisation) but not the Pro figure: on its seeded sample Pro measured 5.11%, close to Base's 5.59% on the same clips. The documentation now cites this paper. Pro v2's 1.89% on all 11,042 clips is lower than the 2024 Pro figure, but 67.9% of those clips are in its training data; the figure to compare tiers by is the held-out 2.06%.

6 Discussion

Pro v2 is the more accurate tier on all three sets. The gain is largest where Base struggled: natural speech with jargon (−11.22 points on TVSpeech) and everyday read speech (−3.28 points on held-out Common Voice). On formal read sentences the gain under the main metric is small, and it widens once numbers are compared as words.

Its conventions differ from Base's. Pro v2 writes numbers as words, English words in Thai script, standard spellings of colloquial words, and repeated words in full. A customer who compares its text with transcripts in another convention will count those differences as errors, as this paper's main metric does.

The claim the evidence supports: on clearly read Thai that it was not trained on, Pro v2 transcribes 97.94% of characters correctly, against 94.67% for Base on the same clips; on natural speech from video, 92.72%, against 81.50%.

7 Production Performance

Pro v2 through the public API. On 9 November 2026, from a client in Bangkok, one request at a time, we sent 180 seeded clips to the `/pro-v2` endpoint exactly as the documentation shows. It answered Common Voice clips (median 4.4 s of audio) in a median 0.68 s end to end (90th percentile 1.11 s, 100 requests), TVSpeech clips (median 26.9 s) in 4.73 s (90th percentile 11.40 s) and FLEURS clips (median 12.1 s) in 2.77 s. 175 of the 180 answers were identical, after normalisation, to the full-set answers of Section 3. The other 5, all TVSpeech clips of 28 s or longer, differed by a word or two; on the 40 TVSpeech clips of the sample, the API's answers measured 7.25% against 7.15% for the full-set answers. Failed requests: 0. For long recordings, the Chinda ASR team measured an 85-minute MP3 at 9.3 minutes through the API [1].

Base and Pro (first edition). Over the 12,633 Base requests, the service reported a median processing time of 0.51 s per clip (90th percentile 1.20 s); with the correction step, median 4.84 s. Through the public API, Base answered Common Voice clips in a median 0.68 s end to end and TVSpeech clips in 2.13 s; Pro took 4.52 s and 6.01 s.

8 Limitations

- **Mostly read speech.** Common Voice and FLEURS are read sentences; only TVSpeech is natural speech from video. Telephone audio, meetings with overlapping speakers, regional dialects and noisy field recordings were not measured.
- **Training overlap.** 67.9% of the Common Voice test clips are in Pro v2's training data; Table 3 separates them, and the headline uses the others.
- **Pro v2's full sets were decoded outside the gateway,** with the service's own code. A service change after 9 November 2026 would not be reflected here.
- **TVSpeech has 18 source videos,** so its intervals are wide.
- **Not scored.** Speaker labels and time stamps, which the service also returns.

9 Reproducibility

All test data are public. The clip lists, including the 3,539 held-out Common Voice clips, the benchmark and scoring scripts, and the raw response for every clip are available from iApp Technology on request, sale@iapp.co.th.

References

[1] iApp Technology. *Chinda ASR Thai Speech-to-Text*, API documentation, section "Measured". https://iapp.co.th/docs/speech/speech-to-text/thai

[2] R. Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*. LREC 2020. Release 17.0, Thai. https://commonvoice.mozilla.org/th/datasets

[3] A. Conneau et al. *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*. IEEE SLT 2022. arXiv:2205.12446. https://arxiv.org/abs/2205.12446

[4] W. Sirichotedumrong et al. *Typhoon ASR Real-time: FastConformer-Transducer for Thai Automatic Speech Recognition*. arXiv:2601.13044, 2026. TVSpeech test set: https://huggingface.co/datasets/typhoon-ai/tvspeech

---

About iApp Technology. iApp Technology Co., Ltd. is a Thai artificial-intelligence company providing OCR, document understanding, speech, identity verification and large-language-model services for the Thai language.
Web: iapp.co.th · E-mail: sale@iapp.co.th · Tel: +66 2 124 4041