



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

Service version 2.10.3 · Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. The iApp Thai Speech-to-Text API transcribes Thai recordings in two tiers, Base and Pro. On the full Common Voice 17.0 Thai test set — 11,042 clips, the set our documentation cites — the Base tier has a character error rate of **5.21%** (95% CI 5.02–5.41), so 94.79% of characters are right; the documentation states 5.87%, and with minimal text normalisation we measure 5.52%. On the 9,856 clips whose sentences are not in the recogniser’s training data, 5.27%. Formal Wikipedia sentences (FLEURS) and natural speech from video with dense jargon (TVSpeech) are harder: 9.93% and 18.50%. Pro’s correction step lowered the error on all three samples, by more than chance alone would explain only on TVSpeech (–1.53 points, CI –2.42 to –0.47); on Common Voice it reached 5.11%, not the 2.19% our documentation states. Base processes a typical clip in a median 0.51 s.

1 Introduction

The Thai Speech-to-Text API (POST `/v3/store/speech/speech-to-text/base` and `/pro`) turns a recording of Thai speech into text, one time-stamped segment per speaker turn. Base costs 1 IC and Pro 2 IC per started minute of audio. Base is the speech recogniser alone; Pro passes the recogniser’s output, part by part, through a second step that re-reads each part of the recording in context and revises the transcript.

Our documentation states that Base gets 5.87% of characters wrong and Pro 2.19% on the 11,042 clips of the Common Voice 17.0 Thai test set [1]. Those figures were measured in April 2024, and the method of text normalisation was not recorded. This paper repeats the measurement on the current production service with a stated method, adds two more public Thai test sets, and checks how much of the Common Voice test set the recogniser could have seen in training.

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). This is the set our documentation cites.

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, from Thai video 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, after the recogniser was trained.

What the recogniser was trained on. Its training data includes the training and validation splits of an earlier release, Common Voice 11.0 Thai, the training split of FLEURS Thai, and further public Thai corpora; we therefore did not use those corpora, such as Gowajee and Thai Elderly Speech, as test sets. Common Voice re-draws its splits at each release, so the 17.0 test split is not guaranteed to be disjoint from

the 11.0 training data. We matched sentences: 1,186 of the 11,042 test clips (10.7%) carry a sentence that occurs in the 11.0 training or validation data. Section 4.1 reports them separately. The FLEURS test sentences do not occur in its training or development splits.

3 Method

Paths. 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. Base was requested with the Pro step switched off, which is what the Base tier is; Pro with it switched on. The Pro step was run on seeded samples (seed 1337) of 300 Common Voice, 120 FLEURS and 80 TVSpeech clips, because it calls a rate-limited correction step shared with customers; we paced it to stay well inside that limit. Seeded samples were also sent through the public API, exactly as the documentation shows (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 the reference and in digits by the service counts as errors. As secondary measures we report word error rate (WER) after segmenting both texts into words 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 over clips (2,000 resamples, seed 1337). Base and Pro are compared on the same clips, and the interval of the difference is paired.

4 Results

Table 1: Character error rate (%), lower is better, with 95% CIs. Base on the full sets (WER secondary); Base and Pro on the same seeded sample. Δ is Pro minus Base, in points.

Set	Full set, Base			Seeded sample			
	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 [17.71, 19.30]	29.02	80	19.21	17.69	−1.53 [−2.42, −0.47]

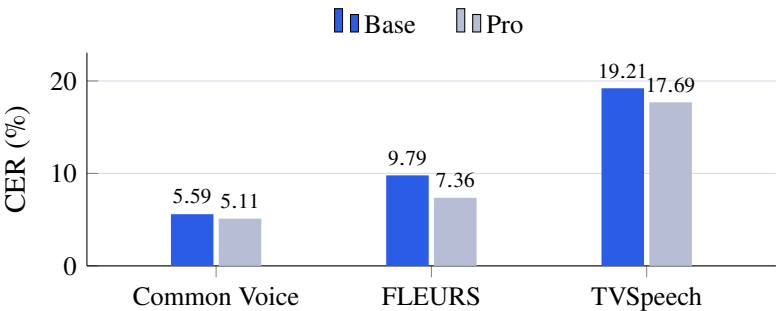


Figure 1: Character error rate, Base and Pro on the same seeded samples (Table 1, right-hand columns).

Common Voice. 57.2% of clips come back without a single character error. With numbers written as words in both texts the CER is 5.18%, so number formatting plays almost no part on this set. In 37 clips, mostly short sentences, the service returned no text at all; they count as fully deleted and account for a small part of the error.

FLEURS and TVSpeech. Both sets have long sentences with names, numbers and English terms. The service writes numbers in digits where the references spell them out (285 of the 12,633 Base answers contain digits that the reference does not); with every number written as words in both texts, the CER is 8.63% on FLEURS and 16.40% on TVSpeech. English terms are a second source of disagreement that is a matter of convention: TVSpeech writes *think tank* in Latin script, and the service in Thai transliteration. The FLEURS references, moreover, cover only part of what some speakers read, which adds insertions. Of the TVSpeech clips, with a median length of 27.2 s, only 0.9% are error-free, but the per-character error is what a customer transcribing a video would see.

4.1 Held-out check

Table 2: Base CER (%) on Common Voice clips by whether the recogniser could have seen the sentence.

Clips	Number	CER [95% CI]
17.0 test, sentence not in the 11.0 training data	9,856	5.27 [5.05, 5.48]
17.0 test, sentence in the 11.0 training data	1,186	4.66 [4.17, 5.17]
17.0 train, seeded sample	1,000	8.53 [7.53, 9.56]

The recogniser could have seen 1,186 test sentences. On those clips its CER is 4.66%, against 5.27% on the others: somewhat lower; without them the overall figure would be 5.27% rather than 5.21%. The training clips themselves show no sign of memorisation: on a seeded sample of 1,000 clips from the 17.0 training split, almost all of whose sentences were in the training data, the CER is 8.53%, higher than on the test set. The training split holds more poor recordings — truncated clips, and clips with no clear speech, on which the recogniser occasionally returned a fluent sentence that was never said — so it is a weak comparison, but it is not the pattern of a model that has memorised its data.

4.2 Speakers

Table 3: Base CER (%) on the held-out Common Voice clips (sentence not in the 11.0 training data) by the speaker’s self-reported gender and age; groups of 50 clips or more.

Group	Clips	CER
Male	352	5.85
Female	256	4.47
Teens	117	6.14
Twenties	321	4.88
Thirties	128	5.90
Not stated (gender)	9,248	5.27

5 Comparison with the Published Figures

Table 4: The documentation’s figures (April 2024) and this measurement (28 October 2026), Common Voice 17.0 Thai test.

Measure	Documentation	This paper	Note
Base CER	5.87	5.21	normalised as in Section 3; 5.52 with minimal normalisation
Base WER	14.52	15.29	word segmentation differs between the two
Pro CER	2.19	5.11	on a seeded sample of 300 clips

The Base figure is reproduced: the current service is as accurate as, or slightly more accurate than, the documentation states. The Pro figure is not. On the seeded sample, Pro’s CER was 5.11%, close to Base’s 5.59% on the same clips (difference -0.48 points, CI -1.09 to $+0.20$). We could not establish how the 2024 Pro figure was obtained and could not reproduce it with the current service; the figures in this paper supersede it.

6 Discussion

Base is accurate on clear read Thai and less so on natural speech. The error rises from 5.21% on Common Voice to 18.50% on video speech with jargon, of which a part is number and transliteration conventions rather than mishearing (16.40% with numbers compared as words).

Pro helps most where Base struggles. The correction step lowered the error on 70 FLEURS clips and raised it on 9; on TVSpeech 48 and 12; on Common Voice 45 and 20. Its fixes are words that only the context decides; its errors are fluent rewordings: an added or dropped connective, a dropped word, an English loanword written in Latin script. Because it produces fluent text, a Pro error is harder to spot than a Base error; the segment score that the service returns (`avg_logprob`) describes the audio and helps find the passages to check.

The claim the evidence supports: on clearly read Thai, the Base tier transcribes about 94.79% of characters correctly; Pro lowers the error further on harder, jargon-rich speech: -1.53 points on TVSpeech (CI -2.42 to -0.47) and -2.43 on FLEURS, where the interval (-4.66 to $+0.17$) still includes zero.

7 Production Performance

Service speed. Over the 12,633 Base requests of this benchmark, the service reported a median processing time of 0.51 s per clip (90th percentile 1.20 s) and 1.43 s for the TVSpeech clips; processing took 0.103 seconds per second of audio over all 22.2 hours. With the correction step: median 4.84 s (90th percentile 5.93 s). The correction ran on 498 of 500 Pro requests; on the others the service returned the recogniser’s own text, as designed.

Public API. From a client in Bangkok, one request at a time, the Base endpoint answered Common Voice clips (median 4.4 s of audio) in a median 0.68 s end to end (90th percentile 0.90 s, 100 requests) and TVSpeech clips (median 26.1 s) in 2.13 s (90th percentile 2.44 s). The Pro endpoint took a median 4.52 s on Common Voice clips and 6.01 s on TVSpeech clips. All 150 of the 150 Base answers were identical to the production service’s answers in the main run, and none used the Pro step; failed requests: 0 on Base and 0 on Pro.

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.** Part of the Common Voice test set shares sentences with the recogniser’s training data; Table 2 separates it. FLEURS is in the training domain (its training split was used), though not its test sentences.
- **Pro on samples.** Pro figures rest on 300, 120 and 80 clips. The correction step is not deterministic, so a repeat run can differ by a few characters.
- **Not scored.** Speaker labels and time stamps, which the service also returns.

9 Reproducibility

All test data are public. The clip lists, 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. *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