iApp Technology · Technical Whitepaper WP-2026-17

English Speech-to-Text: Word Error Rate on LibriSpeech and FLEURS

Service 2.10.4 · Measured 28–29 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. The iApp English Speech-to-Text API transcribes English recordings in two tiers, Base and Pro. On the full LibriSpeech test-clean set — 2,620 read utterances — the Base tier has a word error rate of **1.93%** (95% CI 1.75–2.13), so 98.07% of words are right; on the harder test-other set 3.63% (CI 3.41–3.86), and on 647 FLEURS sentences read by non-professional speakers 5.62% (CI 4.95–6.33). The Pro tier adds a context-aware correction step. On the audiobook speech of LibriSpeech it made no measurable difference (+0.13 points on test-clean, CI −0.05 to +0.35; +0.03 on test-other); on FLEURS, with more names and technical words, it changed the WER by −0.45 points (CI −0.78 to −0.14). The Base tier processes an utterance in a median 0.53 s; through the public API it answers a 5.2-second recording in a median 1.17 s end to end and a 22.7-second one in 1.70 s. Until 29 October 2026 the public Base endpoint also ran the Pro step (Section 6).

1 Introduction

The English Speech-to-Text API (`POST /v3/store/speech/speech-to-text/base/en` and `/pro/en`) turns a recording of English 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. This paper reports how many words each tier gets right on public English test sets, so that anyone can repeat the measurement.

2 Test Data

LibriSpeech test-clean and test-other
The two standard evaluation sets of the LibriSpeech corpus [1]: 2,620 and 2,939 utterances of read audiobook speech, 16 kHz, from speakers not in its training portion. *test-other* holds the speakers whose speech was harder for a reference recogniser. Licence CC BY 4.0.

FLEURS English test
647 recordings of Wikipedia sentences read by non-professional speakers on their own equipment [2]. Licence CC BY 4.0.

LibriSpeech is drawn from public-domain audiobooks that are widely available on the internet. We cannot exclude that some of these recordings occur in the very large training corpus of the recognition model. FLEURS is included as a second, more recent read-speech set; the conclusions below hold on both.

3 Method

Paths. Every utterance was sent, as its original file, to the production speech service — the same instance that answers `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 200 test-clean, 200 test-other and 100 FLEURS utterances, 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, to time it end to end: Pro on 28 October, Base on 29 October, after the fix described in Section 6. The main runs used service version 2.10.3; version 2.10.4 changes only which tier the public Base endpoint runs, not the recogniser or the Pro step.

Metric. Word error rate (WER) is the number of substituted, deleted and inserted words divided by the number of reference words, summed over a set. Before scoring, the reference and the service's text (its segments joined) pass through the same standard English normaliser: lower case; punctuation removed; hyphenated words split; British spellings mapped to American; contractions and titles expanded (*it's* → *it is*, *Mr* → *mister*); numbers brought to one written form; hesitation words such as *uh* removed. A request that failed, or returned no text, counts as deleting every word.

Statistics. 95% confidence intervals are percentile bootstrap intervals over utterances (2,000 resamples, seed 1337). Base and Pro are compared on the same utterances, and the interval of the difference is paired.

4 Results

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

	Full set, Base		Seeded sample			
Set	Utt.	WER [95% CI]	Utt.	Base	Pro	Δ [95% CI]
LibriSpeech test-clean	2,620	1.93 [1.75, 2.13]	200	1.54	1.67	+0.13 [−0.05, +0.35]
LibriSpeech test-other	2,939	3.63 [3.41, 3.86]	200	4.09	4.12	+0.03 [−0.30, +0.36]
FLEURS English test	647	5.62 [4.95, 6.33]	100	6.08	5.64	−0.45 [−0.78, −0.14]

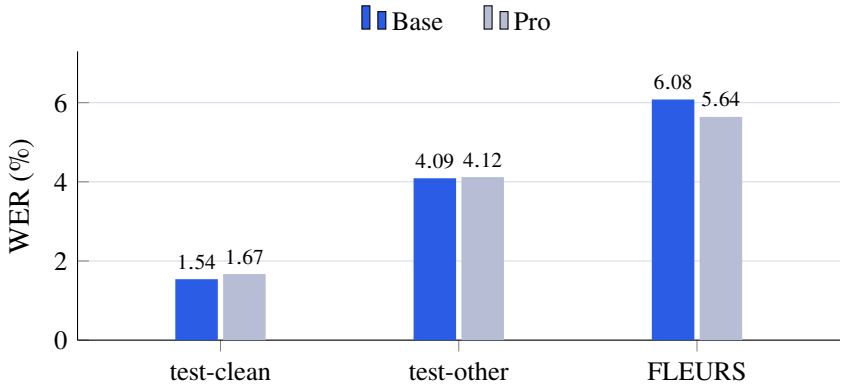


Figure 1: Word error rate, Base and Pro on the same seeded samples (Table 1, right-hand columns).

Table 2: Base tier, full sets: error types (% of reference words), utterances transcribed without a single word error, and failures.

Set	Substitutions	Insertions	Deletions	Error-free utt. (%)	Failed / empty
LibriSpeech test-clean	1.37	0.22	0.35	77.0	0 / 0
LibriSpeech test-other	2.74	0.40	0.49	64.2	0 / 0
FLEURS English	2.62	2.23	0.77	50.5	0 / 0

The error is uneven across speakers. Per speaker, Base WER on test-clean ranges from 0.54% to 5.00% (median 1.82%, 40 speakers); on test-other from 0.93% to 8.38% (median 3.34%, 33 speakers).

5 Discussion

Base is accurate on clean read English. On test-clean, 77.0% of utterances come back without a single word error; most remaining errors are substitutions of one word for a similar-sounding one. test-other, with less clear speakers, has a higher error rate, and FLEURS, recorded by non-professional readers on their own equipment, is harder again.

FLEURS insertions are partly a reference artefact. Insertions make up 2.23 of the 5.62 points on FLEURS, against 0.22 on test-clean. In the utterances with the most insertions, the extra words are the rest of the original Wikipedia sentence, which the speaker read aloud but the reference transcript leaves out. The FLEURS WER therefore overstates the service's error somewhat.

Pro helps with names and vocabulary, not with audiobooks. On the LibriSpeech samples the paired difference between Pro and Base is within a fraction of a point and its confidence interval includes zero (Table 1). The correction step fixed some words and introduced others: on test-clean it lowered the error on 3 utterances and raised it on 5; on test-other 7 and 7. The errors it introduced are rewordings towards modern, grammatical English: *thee* written as *you*, *says* as *said*, the audiobooks' old spelling *to morrow* as *tomorrow*, an added pronoun, a dropped phrase and a misspelt name. On FLEURS the balance is positive (10 utterances better, 1 worse): the fixes are place and organisation names and words that only the context decides, such as *riders* for *writers* and *banknotes* for *bake notes*. Read speech is close to the recogniser's best case, so there is little for a correction step to repair; this test does not tell how Pro performs on the conversational, noisy or jargon-heavy recordings it is meant for.

The claim the evidence supports: on clear read English, the Base tier transcribes about 98.07% of words correctly (LibriSpeech test-clean) and 96.37% on harder read speech (test-other). On read speech, Pro's gain is limited to text with uncommon names and vocabulary.

6 Production Performance

Service speed. Over the 6,206 Base requests of this benchmark, the service reported a median processing time of 0.53 s per utterance (90th percentile 0.82 s) for a median recording of 5.9 s; processing took 0.080 seconds per second of audio over all 12.5 hours. The Pro step adds the correction: median 4.57 s per utterance (90th percentile 5.34 s). The correction ran on 499 of 500 Pro requests; on 1 the service returned the recogniser's own text, as it is designed to when the correction is unavailable.

Public API. From a client in Bangkok, one request at a time, the Base endpoint answered the LibriSpeech sample (median recording 5.2 s) in a median 1.17 s end to end through `api.iapp.co.th` (90th percentile 2.11 s, 100 requests) and test-clean utterances of 20 s or more (median 22.7 s) in 1.70 s (90th percentile 2.33 s, 50 requests). The Pro endpoint took a median 4.32 s on short utterances (90th percentile 4.89 s; 60 requests). None of the Base answers used the Pro step; 100 of the 100 short and 50 of the 50 long ones were identical to the recogniser's answers in the main run, so the endpoint's WER equals the recogniser's on both samples (2.27% against 2.27%; 1.90% against 1.90%). Failed requests: 0 on Base and 0 on Pro.

The Base endpoint ran the Pro step until 29 October 2026. In a first run on 28 October, 100 of 100 Base responses had `asr_pro_is_used` set to `true`: the English Base endpoint returned Pro's text, at Pro's speed (median 4.46 s end to end), at the Base price. Service version 2.10.4, deployed on 29 October 2026, runs the Base endpoint without the Pro step; the Base times above were measured after it. The Base figures elsewhere in this paper are those of the recogniser alone, which is what the Base tier is.

7 Limitations

- **Read speech only.** All three sets are single speakers reading text aloud. Conversations, overlapping speakers, telephone audio, accents under-represented in these sets and domain vocabulary (medical, legal) were not measured, and the WER on such audio will be higher.
- **Possible exposure.** LibriSpeech audio may occur in the recognition model's training data (Section 2). FLEURS is less likely to, and agrees with the LibriSpeech picture.
- **Pro on samples.** Pro figures rest on 200, 200 and 100 utterances; their intervals are correspondingly wider. The correction step is not deterministic, so a repeat run can differ by a few words.
- **Short utterances.** Test utterances last seconds to about half a minute. Speaker labels and time stamps, which the service also returns, were not scored.

8 Reproducibility

All test data are public. The benchmark script, the scoring script (with the exact normaliser) and the raw response for every utterance are available from iApp Technology on request, sale@iapp.co.th.

References

[1] V. Panayotov, G. Chen, D. Povey and S. Khudanpur. *LibriSpeech: an ASR corpus based on public domain audio books*. ICASSP 2015. https://www.openslr.org/12

[2] 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

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