iApp Technology · Technical Whitepaper WP-2026-18

Chinese Speech-to-Text: Character Error Rate on AISHELL-1 and FLEURS

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

Executive summary. The iApp Chinese Speech-to-Text API transcribes Mandarin recordings in two tiers, Base and Pro. On the full AISHELL-1 test set — 7,176 read utterances by 20 speakers held out from the corpus's training portion — the Base tier has a character error rate of **2.87%** (95% CI 2.74–3.01), so 97.13% of characters are right, and 74.7% of utterances come back without a single error. On 945 FLEURS sentences its CER is 10.58%, but most of that gap comes from reference text the speakers do not say as written (bracketed English glosses of names, numbers in digits): on the 654 FLEURS sentences written entirely in Chinese characters it is 3.43%. The Pro tier's correction step helps on the harder FLEURS sentences (−0.55 points on the all-Chinese ones, CI −0.93 to −0.22) and hurts on AISHELL-1 (+1.16 points, CI +0.14 to +2.72), where it also wrote numbers as digits, returned traditional characters in 2 answers and translated 1 utterance into English. Base processes an utterance in a median 0.51 s; through the public API it answers a 4.6-second recording in a median 1.25 s end to end and a 21.5-second one in 4.15 s, mostly upload time (Section 6). Until 29 October 2026 the public Base endpoint also ran the Pro step.

1 Introduction

The Chinese Speech-to-Text API (`POST /v3/store/speech/speech-to-text/base/zh` and `/pro/zh`) turns a recording of Mandarin Chinese into text in simplified characters, 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 characters each tier gets right on two public Mandarin test sets.

2 Test Data

AISHELL-1 test
: The evaluation set of the AISHELL-1 corpus [1]: 7,176 utterances from 20 speakers who are not in its training portion, reading news-domain sentences (finance, technology, sports, entertainment) on a high-fidelity microphone, 16 kHz. Licence Apache 2.0. We checked every transcript against the corpus's official transcript file before scoring.

FLEURS Mandarin test
: 945 recordings of Wikipedia sentences in Mandarin, read by non-professional speakers on their own equipment [2]. Licence CC BY 4.0. The sentences contain foreign names in Latin script, numbers and punctuation, which makes this set harder.

AISHELL-1 is among the most widely used Mandarin training corpora, so its training portion may be in the recognition model's training data. Its test speakers and sentences are held out by design. Section 4.1 checks that the test set behaves as unseen data.

© 2026 iApp Technology Co., Ltd. · iapp.co.th · WP-2026-18

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 300 AISHELL-1 and 150 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, on 28 October and, for Base, again 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. Character error rate (CER) is the number of substituted, deleted and inserted characters divided by the number of reference characters, summed over a set. Before scoring, the reference and the service's text pass through the same normalisation: full-width letters and digits to half-width; traditional characters to simplified; punctuation, symbols and spaces removed; Latin letters lower-cased. Each Chinese character, Latin letter and digit counts as one character. Numbers are not converted between Chinese numerals and digits. A request that failed, or returned no text, counts as deleting every character.

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: Character 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.	CER [95% CI]	Utt.	Base	Pro	Δ [95% CI]
AISHELL-1 test	7,176	2.87 [2.74, 3.01]	300	2.87	4.04	+1.16 [+0.14, +2.72]
FLEURS Mandarin test	945	10.58 [9.56, 11.56]	150	12.51	11.10	−1.41 [−2.46, −0.59]

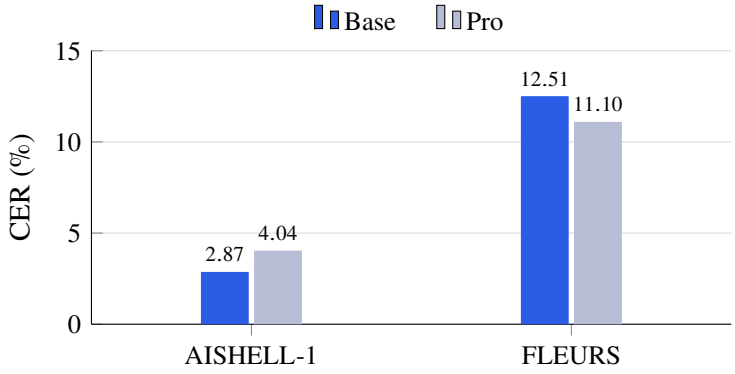


Figure 1: Character error rate, Base and Pro on the same seeded samples (Table 1, right-hand columns).

Table 2: Base tier, full sets: error types (% of reference characters), error-free utterances, and failures.

Set	Substitutions	Insertions	Deletions	Error-free utt. (%)	Failed / empty
AISHELL-1 test	2.69	0.08	0.10	74.7	0 / 0
FLEURS Mandarin	4.92	1.02	4.64	33.8	0 / 0

Substitutions make up nearly all of Base's error on AISHELL-1: mostly homophones and near-homophones, often in names of people, places and companies, which Mandarin speech does not disambiguate without context (for example the city of Shiyan written as a homophone that means "table salt"). Per speaker, Base CER ranges from 1.81% to 4.89% (median 2.67%). Base never answered in Arabic digits on either set: it writes numbers in Chinese numerals, as AISHELL-1 does.

On FLEURS, deletions are 4.64 of the 10.58 points. The FLEURS reference text gives many foreign names twice, in Chinese and again in brackets in Latin script, and writes numbers as digits; the speakers read the Chinese and say the numbers as words. The service transcribes what is said, so the bracketed names count as deleted and the numbers as substituted. On the 654 FLEURS sentences (69.2%) whose reference is written entirely in Chinese characters, Base CER is 3.43% (CI 3.11–3.77), in line with AISHELL-1.

Table 3: Form of the answer on the 300 AISHELL-1 utterances of the Pro sample: number of utterances.

	Base	Pro
Numbers written in Arabic digits	0	3
Traditional characters in the answer	0	2
Answer translated into English	0	1

4.1 Held-out check

A model that had memorised AISHELL-1 would transcribe the corpus's training speakers far better than its test speakers. On a seeded sample of 500 utterances by two training speakers, Base CER was 2.14%, against 2.87% on the test set. The two training speakers fall inside the range of the 20 test speakers (1.81% to 4.89%), so the check shows no sign that the recogniser memorised the corpus. Two speakers cannot rule out that its training data included the AISHELL-1 training portion, as is common for Mandarin recognisers; the test speakers are held out by the corpus design.

5 Discussion

Base is accurate on read Mandarin. 97.13% of characters are right on AISHELL-1, with no failed requests and no empty answers, and the error is similar on the all-Chinese FLEURS sentences, which come from a different domain (Wikipedia) and different recording conditions.

Pro helps on the harder text and hurts on the easier. On the FLEURS sentences, with more foreign names and technical vocabulary, the correction step lowered the error on 23 utterances and raised it on 4. On AISHELL-1 it lowered the error on 8 and raised it on 16. Part of that is a change of form rather than of words: it wrote numbers as Arabic digits, which the metric counts as errors; without the 3 utterances where either tier answered in digits, Pro's difference is +0.65 points (CI −0.12 to +1.96). The rest is the correction step's own errors: it removed words the speaker repeated, replaced a place name with a better-known one, returned traditional characters in 2 answers, and in 1 case returned an English translation instead of the Chinese transcript (Table 3). An application that needs simplified characters only should convert Pro's text, and should check answers that contain no Chinese characters.

The claim the evidence supports: on clearly read Mandarin, the Base tier transcribes about 97.13% of

characters correctly. Pro is the better choice for text with uncommon names and vocabulary, and on plain read news text Base is as accurate or more.

6 Production Performance

Service speed. Over the 8,121 Base requests of this benchmark, the service reported a median processing time of 0.51 s per utterance (90th percentile 0.68 s); processing took 0.095 seconds per second of audio over all 13.1 hours. With the correction step: median 4.55 s (90th percentile 5.20 s). The correction ran on 449 of 450 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 AISHELL-1 utterances (median recording 4.6 s) in a median 1.25 s end to end (90th percentile 2.57 s, 100 requests) and the FLEURS clips of 20 s or more (median 21.5 s) in 4.15 s (90th percentile 7.68 s, 39 requests). The service itself reported a median 0.42 s and 0.84 s of processing; most of the long clips' remaining time is the upload, as FLEURS files are 32-bit WAV (median 1.4 MB, twice the bytes of the same audio in 16-bit). None of the Base answers used the Pro step; 100 of the 100 short and 39 of the 39 long ones were identical to the recogniser's answers in the main run (CER 3.12% against 3.12%; 13.05% against 13.05%). Failed requests: 0. The Pro endpoint was not timed through the public API: the first run, on 28 October, stopped at a server error (a 502 returned by the gateway while the speech service itself was answering normally) before it reached the Pro stage.

The Base endpoint ran the Pro step until 29 October 2026. In that first run, 48 of 48 Base responses had `asr_pro_is_used` set to `true`: the Base endpoint returned Pro's text at Pro's speed (median 4.61 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.** Both sets are single speakers reading text aloud. Conversations, overlapping speakers, telephone audio, regional accents and Cantonese were not measured.
- **Numbers.** The metric does not convert between Chinese numerals and digits; an application that needs one form should normalise the service's text itself.
- **Pro on samples.** Pro figures rest on 300 and 150 utterances. 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.

8 Reproducibility

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

References

[1] H. Bu, J. Du, X. Na, B. Wu and H. Zheng. *AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline*. O-COCOSDA 2017. https://www.openslr.org/33

[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