iApp Technology · Technical Whitepaper WP-2026-28

ChindaTTS Text-to-Speech: Intelligibility Measured by Round-Trip Recognition

Service version 1.0.2 · Measured 28–29 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. ChindaTTS turns Thai text into speech in built-in voices or in a voice cloned from a short clip. Without a panel of listeners we measured intelligibility: 60 FLEURS and 60 Common Voice Thai sentences were spoken by each voice and transcribed back by the iApp Thai Speech-to-Text service. The default voice, Kaitom, had a round-trip character error rate of **4.93%** (95% CI 3.89–6.02); the three built-in voices together 5.40% and 10 cloned library voices 5.06%, from 4.27% to 5.73% per voice. On the Common Voice sentences, the built-in voices scored 4.72% and the library voices 3.13%, against 4.60% for the volunteers who recorded the same sentences: to the recogniser, the synthetic speech is as intelligible as human read speech. The eight speaking styles of the default voice scored from 4.20% to 5.98% on 30 sentences (neutral 4.59%), all within each other's confidence intervals. A built-in voice produces a second of audio in 0.45 s; cloning takes 1.07 s, longer than the audio itself. Through the public API, from a client in Bangkok, a built-in voice answered in a median 1.47 s and a cloned voice in 4.38 s. Naturalness and prosody were not measured.

1 Introduction

The ChindaTTS API (`POST /v3/store/audio/tts`) speaks Thai text, or Thai mixed with English, in one of three built-in voices and eight speaking styles, and returns a 48 kHz WAV or MP3 file. A second endpoint, `/v3/store/audio/tts/clone`, speaks in the voice of an uploaded clip of 10 to 20 seconds; iApp publishes ten such reference clips, a voice library from a child to a grandparent, that customers can clone.

A buyer wants to know whether listeners will understand the words. The usual measure of quality, a mean opinion score from human listeners, needs a listening panel, which this paper did not have. It measures instead whether a speech recogniser can recover the text from the audio. This is a standard proxy for intelligibility, and it can be repeated by anyone.

2 What This Method Shows, and What It Does Not

What it shows. Round-trip recognition measures whether the words in the audio are the words in the text: whether syllables are dropped, added or garbled, whether tones and vowel lengths are clear enough for the words to be recognised, and whether numbers and abbreviations are read out. A voice whose round-trip error is close to that of human speakers reading the same sentences is, for a recogniser, as intelligible as they are.

What it does not show. It does not measure naturalness, rhythm and intonation (prosody), whether the speaking styles sound as intended, how pleasant a voice is, or how closely a cloned voice resembles its reference speaker. A recogniser also forgives some errors a listener would notice, such as a wrong tone in a word it can complete from context, and it can itself mishear clear speech. Human listening tests are needed for those questions. Finally, the service checks each take with a speech recogniser of its own before returning it; the judge used here is a different recogniser, but round-trip error is partly a property the service already screens for.

© 2026 iApp Technology Co., Ltd. · iapp.co.th · WP-2026-28

3 Test Data

Sentences
60 sentences of the FLEURS Thai test set [1] (Wikipedia text: long, formal, with names, numbers and some English words; CC BY 4.0) and 60 sentences of the Common Voice 17.0 Thai test set [2] (short everyday sentences; CC0), each drawn at random with seed 1337. The Common Voice sentences are restricted to those of 20 characters or more that do not occur in the earlier Common Voice release on which the recogniser was trained, so that the recogniser has no advantage on them. Each sentence was sent exactly as published, punctuation included.

Voices
The three built-in voices in the neutral style: Kaitom (the default), Kaimook and Kaidang; and all 10 library voices, each cloned from its published reference clip through the cloning endpoint, exactly as a customer would.

Speaking styles
Kaitom in each of the eight styles, on 30 of the sentences: the first 15 of each test set in the seeded order.

Human reference
All 240 human recordings of the same 120 sentences in the two test sets.

4 Method

Synthesis. Each sentence was sent to the production TTS service — the same instance that answers `api.iapp.co.th` — two requests at a time: 1,770 requests in all, in two sessions. The first, on 28 October 2026, covered the built-in voices and eight library voices (1,350 requests); the second, on 29 October, the grandmother and grandfather voices and the speaking styles (420 requests). Between the sessions the service received a stability update. It changes how the service handles a request it cannot complete, so that the failure stays with that request; it does not change how speech is generated. To check this, the default voice was run again on the 30 style sentences in the second session (Section 5.2).

Recognition. Every returned file, and every human recording, was transcribed by the production iApp Thai Speech-to-Text service, Base tier. Its accuracy on human speech is reported in a companion paper [3].

Metric. Round-trip character error rate (CER): character substitutions, deletions and insertions between the text sent and the text recognised, divided by the characters sent, summed over the sentences of a voice. Both texts are normalised as in the companion paper: Unicode NFC; Thai digits to Arabic digits; punctuation and spaces removed; Latin letters lower-cased. Numbers are not spelled out. A synthesis request that failed counts as losing every character. For the human reference, each sentence's errors are averaged over its recordings. 95% confidence intervals are percentile bootstrap intervals over sentences (2,000 resamples, seed 1337).

5 Results

Table 1: Round-trip CER (%) by voice, neutral style, lower is better; 60 FLEURS and 60 Common Voice (CV) sentences. Insertions and deletions as % of characters sent; speaking rate in characters per second of audio. Human recordings: 240 recordings of the same sentences, each sentence's errors averaged over its recordings.

Voice	Sent.	CER [95% CI]	FLEURS	CV	Ins.	Del.	Chars/s
Kaitom (default)	120	**4.93** [3.89, 6.02]	4.92	4.96	1.60	1.06	11.3
Kaimook	120	**5.73** [4.33, 7.12]	6.66	2.64	1.76	1.43	11.0
Kaidang	120	**5.55** [4.21, 7.03]	5.24	6.56	1.67	1.59	11.5
Girl, about 8	120	**5.08** [3.98, 6.24]	5.30	4.37	1.44	1.33	11.7
Boy, about 7	120	**5.60** [4.22, 7.14]	6.43	2.87	2.14	1.09	11.4
Teenage girl, about 16	120	**4.27** [3.19, 5.55]	4.88	2.23	1.32	0.94	12.1
Teenage boy, about 16	120	**5.36** [4.05, 6.89]	6.29	2.28	1.89	1.34	11.4
Woman, about 30	120	**5.48** [4.09, 7.14]	6.27	2.87	1.58	1.58	12.1
Man, about 30	120	**4.78** [3.83, 5.81]	4.96	4.19	1.06	1.26	13.2
Woman, about 55	120	**5.60** [4.40, 6.98]	6.10	3.96	1.77	1.55	12.2
Man, about 55	120	**4.80** [3.95, 5.77]	5.32	3.05	1.25	1.43	12.3
Grandmother, about 80	120	**5.07** [3.82, 6.58]	5.68	3.05	1.75	1.20	11.2
Grandfather, about 80	120	**4.56** [3.35, 5.94]	5.21	2.41	1.02	1.35	12.1
Human recordings	120	**8.31** [6.58, 10.29]	9.43	4.60			

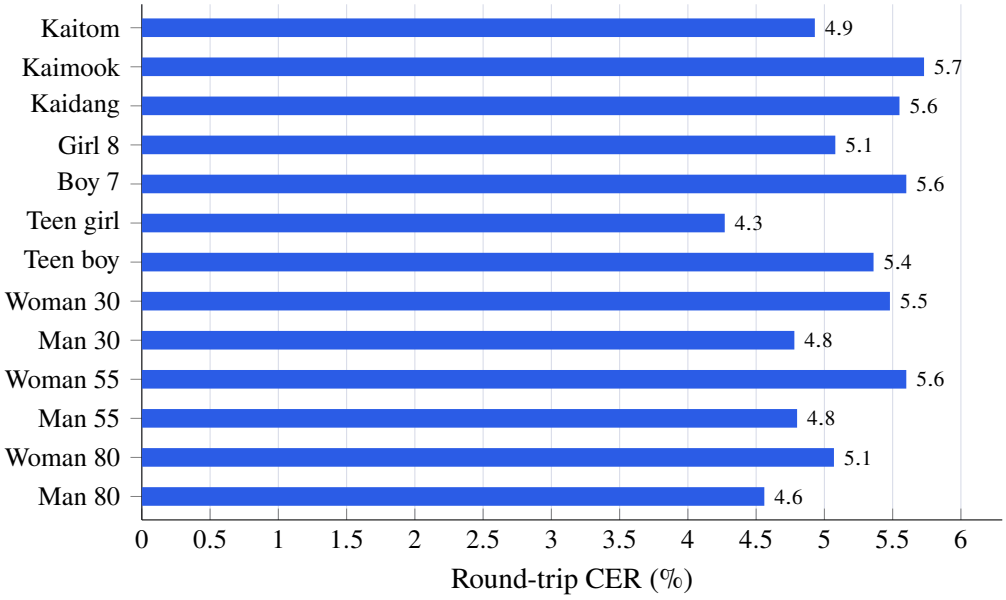


Figure 1: Round-trip CER by voice, neutral style, all 60 plus 60 sentences (Table 1).

All 1,770 synthesis requests returned audio, and every file was transcribed.

Voices. The differences between voices are small against their confidence intervals: all 13 lie between 4.27% and 5.73%. The cloned library voices, from a girl of about eight to a grandfather of about 80, are as intelligible as the built-in voices.

Long and short sentences. On the long FLEURS sentences, which carry names, numbers and abbreviations, the error is similar for every voice. On the short Common Voice sentences the built-in voices make more insertions (1.93% of characters) than the library voices (0.72%) or the human speakers (0.73%): in the sentences with the most insertions, the recogniser heard extra syllables after the end of the sentence. We did not listen to these files, so we cannot say how much of this is in the audio and how much is the recogniser's.

Numbers and abbreviations. The text was sent as published, so *70 km*, written in digits and a Thai abbreviation, was correctly read out in full words; the recogniser then wrote it in words and the metric counts the difference as errors. The same happens to the human speakers.

The human reference. On the Common Voice sentences the human reference is clean. On FLEURS it is not: the speakers read whole sentences of which the published reference covers only part, so the recogniser's transcript of the human recordings has many more insertions (4.05% of characters) than that of the synthetic speech (1.60%), which reads exactly the reference text. The FLEURS human figure, 9.43%, therefore overstates human error, and the comparison in this paper rests on the Common Voice sentences.

5.1 Speaking Styles

Table 2: Round-trip CER (%) of the default voice, Kaitom, by speaking style, on the same 30 sentences (15 FLEURS, 15 Common Voice); speaking rate in characters per second of audio. The neutral row is the neutral-style take of the same sentences from Table 1.

Style	Sent.	CER [95% CI]	Chars/s
Neutral	30	4.59 [2.75, 7.08]	11.5
Friendly	30	5.37 [3.45, 7.29]	11.9
Cheerful	30	4.59 [3.13, 6.59]	11.9
Calm	30	4.81 [3.32, 6.43]	11.2
Serious	30	5.37 [3.83, 6.92]	12.1
Sad	30	5.98 [3.06, 10.25]	11.1
Excited	30	5.11 [3.34, 7.45]	12.8
Empathetic	30	4.20 [2.79, 5.62]	12.3

No style makes the speech measurably harder to recognise: all eight lie between 4.20% (empathetic) and 5.98% (sad), within each other's confidence intervals, which are wide on 30 sentences. The speaking rate moves with the style, slowest for *sad* and fastest for *excited*; whether each style sounds as intended is a question for listeners (Section 2).

5.2 Control Repeat

The default voice was run again on the 30 style sentences in the second session, after the service update. Its round-trip CER was 3.81% [2.68, 4.93] against 4.59% [2.75, 7.08] for the takes of the same sentences on 28 October, at 11.8 and 11.5 characters per second. The difference is within the variation between two takes of the same text: the update did not change the speech. The figures in Table 1 for the default voice are those of 28 October.

6 Speed

Times are measured at the client, two requests at a time, while the service also served customers. In the first session the client's connection to the data centre added about 0.2–0.3 s to each request; the requests of the second session (the last two library voices) were sent from inside the data centre.

Table 3: Synthesis time per request and real-time factor (synthesis time per second of audio produced).

Voices	Median (s)	90th pct (s)	Median audio (s)	Real-time factor
Built-in	2.13	7.59	5.3	0.45
Library (cloned)	6.55	10.44	4.7	1.07

A built-in voice took a median 1.35 s for a Common Voice sentence (median 34 characters) and 3.27 s for a FLEURS sentence (median 124 characters). Cloning processes the reference clip with every request, and took a median 6.14 s and 7.30 s for the same sentences: an application that speaks in a cloned voice should expect several seconds per sentence and send long text in parallel parts.

6.1 Through the Public API

A seeded sample of 60 sentences (10 for each built-in voice, 3 for each library voice) was sent to `api.iapp.co.th` one request at a time from a client in Bangkok, through the same gateway, authentication and billing as a customer's request. All returned 48,000 Hz audio (0 failures).

Table 4: End-to-end time through the public API, client in Bangkok, one request at a time; real-time factor = time per second of audio returned.

Endpoint	Requests	Median (s)	90th pct (s)	Real-time factor
`/v3/store/audio/tts` (built-in)	30	1.47	7.01	0.61
`/v3/store/audio/tts/clone` (library)	30	4.38	8.06	0.81

These times are shorter than those of Table 3 because the requests were sent one at a time rather than two. Transcribed back, the audio returned by the API had a round-trip CER of 3.51% against 4.58% for the takes of the same sentences in Table 1: two independent takes of each sentence, within the variation shown in Section 5.2. The 60 requests, 4,119 characters in all, cost 1.46 IC at 0.02 IC per 100 characters, rounded up per request.

7 Limitations

- **Not a listening test.** See Section 2: naturalness, prosody, pronunciation that a listener would notice but a recogniser tolerates, and how much a cloned voice resembles its reference were not measured.
- **The judge is one recogniser.** Its own errors are part of every figure; the human reference shows their size. A different recogniser would give different absolute numbers.
- **Sentences.** 60 plus 60 sentences in Thai, mostly without English. Mixed Thai-English text, long passages and domain vocabulary were measured only as far as these sentences contain them.
- **Styles on one voice.** The speaking styles were measured on the default voice only, on 30 sentences; their confidence intervals are wide (Section 5.1).
- **Two sessions.** The measurements span two days and a service update between them; the control repeat found no change in the speech (Section 5.2).
- **One take per sentence.** The service generates speech stochastically; a repeat run produces different audio and slightly different figures. Two takes of the same 30 sentences scored 4.59% and 3.81% (Section 5.2).

8 Reproducibility

All sentences and human recordings are public, and so are the ten library reference clips (`https://iapp.co.th/docs/speech/text-to-speech/text-to-speech-v3`). The benchmark and scoring scripts and the raw results are available from iApp Technology on request, sale@iapp.co.th.

References

[1] 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

[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] iApp Technology. *Thai Speech-to-Text: Character Error Rate on Common Voice, FLEURS and TVSpeech*. Technical Whitepaper, 2026.

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