Offline Cebuano Text to Speech in QnVoice for iOS with VoxCPM2

QnVoice
QnSub Software Solutions
Research preprint · November 2026

Abstract

QnVoice adds offline Cebuano text to speech to its iOS Speak feature by reusing a quantized VoxCPM2 runtime. Typed Cebuano goes directly to synthesis. English input can pass through a bundled English-to-Cebuano Marian model. We screen five synthetic-reference configurations on 12 authored phrases and compare the shared reference with the best new design on 16 separate phrases. BuzzASR character error is 3.32% for the shared reference and 5.63% for the designed reference in the second comparison. Word error is 18.89% and 32.22%, respectively. On an iPhone 15 Pro, the longest of six typed requests reaches the playback state in 4.67 s. The five later requests have a median of 3.54 s. The packaged voice profile pins the shared weights and includes the reference features, test phrases, and generated examples. Fluent listener ratings and coverage across devices remain open.

Keywords Cebuano, QnVoice, iOS, iPhone, text to speech, VoxCPM2, offline translation

1 Introduction

A text to speech feature for Cebuano should preserve the user's words, including everyday contractions and hyphenated forms. In QnVoice, the typed phrase “Maayong buntag” becomes speech with local model files. A separate input mode accepts English and translates it to Cebuano before synthesis. This distinction makes it possible to evaluate speech output with known Cebuano text and inspect machine translation separately.

QnVoice already ships a quantized VoxCPM2 runtime and a synthetic reference for Filipino [1, 2]. Sharing those resources could make Cebuano feasible without adding another multi-gigabyte checkpoint. Cebuano lies outside the 30 languages listed by the VoxCPM2 authors, so we test transfer directly on authored text. We also test designed references to see how conditioning affects the same Cebuano phrases.

The study measures transcription agreement, time from typed submission to app playback, and basic English translation behavior. These endpoints address different aspects of the experience. A single recognizer can surface possible omissions and substitutions, while a listener study can assess naturalness and pronunciation.

2 Related Work and Asset Selection

2.1 Cebuano speech resources

The Philippine Languages Database includes Cebuano in a multilingual collection designed for Philippine speech research [3]. Meta's Massively Multilingual Speech project provides a Cebuano VITS checkpoint with a noncommercial license [4]. The OmniVoice authors report Cebuano training coverage and release their pretrained weights under noncommercial terms [5]. These projects demonstrate multilingual coverage, while the distribution requirements for QnVoice favor the existing Apache-licensed VoxCPM2 weights.

Google's FLEURS corpus includes Cebuano read speech and distributes its data under CC BY 4.0 [6]. BuzzASR adapts a multilingual Whisper checkpoint to Cebuano using data that includes FLEURS [7]. We use BuzzASR as a desktop measurement tool. The production reference and all candidate designs are generated from written descriptions and project-authored text. BuzzASR runs only on the evaluation Mac.

2.2 Deployed model components

Table 1 Components and distribution terms

Component	Source	Terms and use
Speech generation	OpenBMB and ModelBest VoxCPM2 [1]	Apache 2.0 weights. QnVoice reuses its existing two GGUF files.
Reference	Original QnVoice synthetic reference [2]	Use terms require a paper citation from apps that use the reference.
Translation	Helsinki-NLP OPUS-MT English to Cebuano [8]	Apache 2.0. QnVoice exports the encoder and decoder to ONNX.
Runtime	llama.cpp-omni and ONNX Runtime [9, 10]	MIT. Native, on-device inference.

The app retains the existing speech checkpoint and its notices. The model card links to the exact shared-weight revision and provides the generated reference data. The two components can therefore be verified independently using their SHA-256 digests.

3 Text to Speech in QnVoice

3.1 Input and playback

In Speak, users select Cebuano and then choose Cebuano or English in the typed input selector. The choice controls whether the app sends text directly to synthesis or runs the translator first. The phrasebook supplies English entries to the translation path. In Record, the app obtains an English transcript from the installed iOS on-device recognizer. The original recording remains available for playback.

Table 2 Entry points for the Cebuano voice

Input	Text preparation	Output
Typed Cebuano	Preserve the typed string. Normalize speech punctuation and split it at sentence boundaries.	Cebuano speech and the original text.
Typed English or phrasebook entry	Translate with a reviewed phrase or the bundled Marian graphs.	Cebuano text and speech, with the original typed take available.
English recording	Recognize English on device, then translate the transcript.	Cebuano speech and translation, alongside the original recording.

The app stores the input language for typed Cebuano with the take. When a user edits that transcript, QnVoice invalidates its rendered voice variants and reads the corrected Cebuano words. It stages a render before adding it to the take's playback cache. Share and voice controls use the same rendered file.

3.2 Text boundaries and resource use

The frontend normalizes compatibility forms and quotation marks, turns long dashes into pauses, and keeps ordinary hyphens and apostrophes. A phrase may contain up to 400 characters. The synthesizer processes sentences in groups of up to 120 characters where word boundaries allow. The test case “Kumusta ka? Asa ta mangaon?” becomes two segments and uses a short gap between them.

The language and acoustic GGUF files occupy approximately 2.08 GB in total. Filipino and Cebuano share one serial runtime session and the same cached reference. The Cebuano addition ships about 137 MB of int8 translation files and tokenizer data. The speech runtime targets phones and tablets with at least 6 GB of physical memory and checks available process memory before loading.

4 Reference Design and Methods

4.1 Synthetic reference preparation

A reference waveform conditions VoxCPM2 on a speaker style through the AudioVAE encoder. The existing QnVoice reference was generated with a pinned four-bit MLX export [11]. We compare it with a two-stage Cebuano re-render and three independently designed references. For the new designs, we provide a written style description and a short Cebuano phrase. We resample to 16 kHz mono and match the reference length to 71,683 samples using padding or mild time adjustment.

Table 3 Reference alternatives in the screening

Reference	Generation method	Selection role
Shared	Original QnVoice synthetic Filipino reference. Design seed 1302.	Reuse comparison.
Re-render	Cebuano phrase generated while conditioning on the shared reference.	Transfer check.
Calm	Design seed 7410. Adult male, relaxed mid-low register.	Screened design.
Clear	Design seed 1676. Adult female, moderate pace.	Screened design.
Soft	Design seed 3681. Adult male, restrained pitch and clear vowels.	Lowest screening edit counts among the designs.

The published reference features contain 29 patches of four frames with 64 values per frame. The little-endian float32 file is 29,696 bytes. The engine can use these features directly with the pinned acoustic graph, or encode the provided waveform.

4.2 Matched desktop protocol

The screening uses 12 short prompts covering greetings, requests, directions, and everyday questions. After selecting the designed reference with the fewest edits, we compare it with the shared reference on 16 different phrases. Both variants use the same Mac runtime, text, six flow evaluations, guidance of 2, and seed 98000. The harness generates each prompt once with each reference. It measures native synthesis to the full waveform at 48 kHz.

BuzzASR uses the published Cebuano checkpoint, greedy decoding, a three-token repeated-sequence constraint, and a repetition penalty of 1.2. The scoring code case-folds the reference and hypothesis, strips diacritics and punctuation, and computes Levenshtein edits. Character scoring joins the tokens without spaces. Word scoring retains token boundaries.

5 Desktop Evaluation Results

Table 4 Reference screening on 12 Cebuano phrases

Reference	Character edits	CER	Word edits	WER
Shared	17 / 286	5.94%	20 / 62	32.26%
Cebuano re-render	23 / 286	8.04%	24 / 62	38.71%
Calm design	18 / 286	6.29%	18 / 62	29.03%
Clear design	18 / 286	6.29%	18 / 62	29.03%
Soft design	16 / 286	5.59%	17 / 62	27.42%

The soft design had 16 character edits and 17 word edits. The shared reference had 17 and 20 edits, respectively. This is a small screening difference, and we kept both conditions for the second comparison.

Table 5 Matched comparison on 16 separate prompts

Reference	Character edits	CER	Word edits	WER
Shared	13 / 391	3.32%	17 / 90	18.89%
Soft design	22 / 391	5.63%	29 / 90	32.22%

The shared reference recorded fewer edits on the separate set. We chose it for the app on that basis. A descriptive paired bootstrap using 10,000 prompt resamples gives an interval of −6.18 to +1.40 percentage points for the shared-minus-designed character error and −26.14 to 0.00 points for word error. These intervals describe uncertainty in the measured difference across this small set.

Word errors include spacing. The soft design's transcript for “Gipadala na nako ang mensahe” joined parts of the phrase and produced three word edits with zero character edits. In the other direction, the soft design matched “Tagpila ang usa ka kilo nga mangga?” exactly, while the shared reference had two character edits. The detailed record includes both transcripts [12].

In the 16-phrase desktop comparison, median synthesis time was 1.64 s with the shared reference and 1.73 s with the soft design. These values exclude session opening and measure the desktop synthesis call, with reference encoding inside the first call.

6 iPhone Evaluation

The QnVoice build 15 test ran on an iPhone 15 Pro with iOS 27.0. It submits six typed Cebuano requests in a fixed order and waits for the app player to report playback. The elapsed interval starts before the take is rendered and includes audio preparation and file publication. It ends at playback state. The test also verifies finite, audible samples and saves a clip for each request.

Table 6 Typed Cebuano playback on iPhone 15 Pro

Request	Cebuano input	Readiness	Audio duration
1	Maayong buntag.	3.14 s	1.28 s
2	Daghang salamat sa imong tabang.	3.54 s	2.24 s
3	Gusto ko og usa ka baso nga tubig.	3.73 s	2.40 s
4	Palihog, hinayi og sulti.	3.36 s	2.08 s
5	Kumusta ka? Asa ta mangaon?	4.67 s	2.60 s
6	Magkita ta ugma sa buntag.	3.36 s	1.92 s

First-request readiness was 3.14 s. The later median was 3.54 s, and the longest request reached playback in 4.67 s. The six requests are one sequence on one phone.

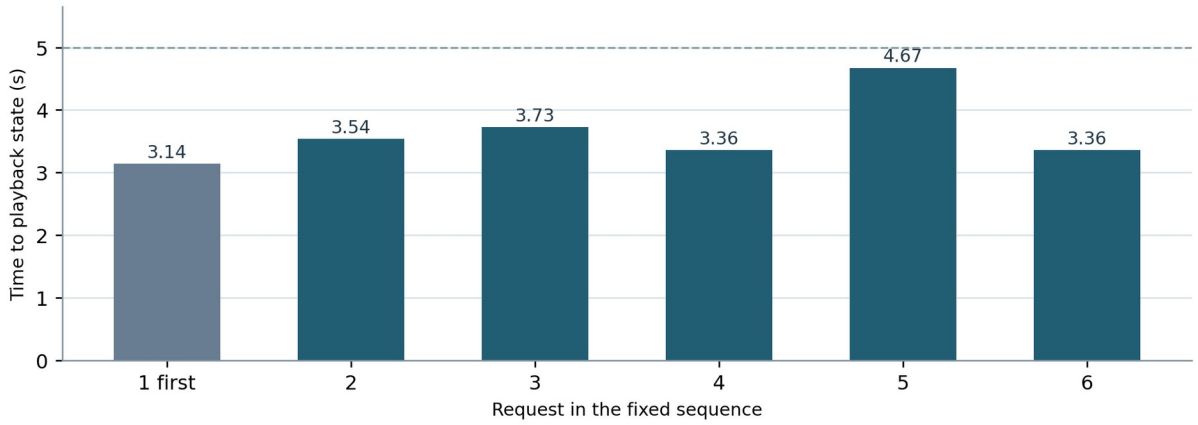


Figure 1 Time from typed submission to playback state. Requests 2–6 follow the first request in the same app process.

Five focused tests passed. They covered catalog and model presence, punctuation, typed Cebuano playback and transcript edits, typed English translation and playback, and the existing Urdu typed path. The edit test regenerated the Cebuano version and checked that the original audio bytes matched.

7 English to Cebuano Translation

7.1 Local Marian inference

The dedicated English-to-Cebuano Marian model uses SentencePiece tokens, an encoder, and a decoder with attention caching. QnVoice converts the pinned weights to ONNX and applies per-channel dynamic signed int8 quantization to supported operators. The decoder receives an integer cache flag from Swift and casts it to the graph's Boolean selector. Translation and speech run sequentially in the local app pipeline.

Table 7 Selected English requests and resulting text

English input	General model output	App handling
Good morning.	Maayong buntag.	Marian output.
I would like a glass of water.	Gusto kog basong tubig.	Marian output verified on iPhone.
How are you? Where shall we eat?	Unsaon man nimo? Asa man kami magkaon?	Kumusta ka? Asa ta mangaon?
Where is the nearest pharmacy?	Hain ang kinadul - ang botika?	Asa ang pinakaduol nga botika?
See you tomorrow morning.	Tan - awa ugma sa buntag.	Magkita ta ugma sa buntag.

For a small set of exact everyday requests, the app uses project-reviewed wording. This addresses observed literal translations such as “Tan - awa ugma sa buntag” for “See you tomorrow morning.” The review rules apply when all segments in a request match a reviewed phrase. Other text uses the general model.

7.2 What the translation check measured

A desktop smoke test ran the quantized graphs on the English glosses of the 12 screening prompts. The iPhone test verified the combined typed-English route for the glass-of-water request. It observed the displayed Cebuano text, generated audio, and the playback state. The examples cover a few everyday utterances. They provide evidence about routing and vocabulary, while translation quality across domains requires a larger parallel set and native-speaker assessment.

The English recording entry point relies on iOS availability and permission for local English recognition. The iPhone timing table covers typed Cebuano. The machine translation and recognition stages have separate latencies and error sources.

8 Discussion and Limitations

8.1 Reference selection and playback

The smallest transcription difference in screening did not persist on the separate prompts. The shared reference performed better on that set and costs the same model storage as the existing Filipino voice. This supports using it as the first Cebuano release profile. The release also includes the soft design so researchers can reproduce the comparison and listen to the differences.

Several word errors reflect alternative spellings or word boundaries, including “palihug” and “palihog.” Others resemble pronunciation or content errors, as when the recognizer substituted a different initial consonant. Transcripts alone cannot establish the cause. A blinded listening study could ask fluent Cebuano speakers to transcribe clips and rate pronunciation and naturalness separately using a documented protocol [13].

8.2 Coverage and measurement boundaries

The study uses short authored text, one hardware model for iPhone integration, and one recognizer. The 16-phrase comparison informed release selection, so an independent evaluation is needed to estimate the selected profile's accuracy on a new population of utterances. The tests give limited evidence about long passages, regional pronunciation, numbers, names, and Cebuano–English code switching.

Playback readiness varies with device memory, thermal state, phrase length, and whether the runtime is already loaded. The test run measures the player state on a single iPhone. It leaves first launch after a device reboot, peak memory, and timing across other devices unmeasured. The published audio examples let readers inspect representative outputs directly.

8.3 Data and reproducibility

QnVoice publishes the selected and research reference waveforms and encoded features, their SHA-256 digests, the pinned weights, and an evaluation record with prompts and transcripts. The references were generated from written descriptions and authored text. The benchmark examples contain synthesized speech. Inference in QnVoice uses bundled files. A desktop command first retrieves the shared weights when requested and can verify and reuse them locally.

9 Conclusion and References

A shared quantized VoxCPM2 runtime provides a practical offline Cebuano voice in QnVoice for iOS. The selected reference produced the lower automated error on the 16 matched phrases, and the tested iPhone completed all six typed requests within 4.68 s to playback state. These results support the tested deployment and identify questions for fluent-listener and device evaluation.

Cebuano model and evaluation files · Pinned shared GGUF weights · Filipino reference preprint

References

[1] Zhou, Y., et al. VoxCPM2 Technical Report. arXiv:2606.06928, 2026.

[2] Asis, O. P. Offline Filipino Text to Speech in QnVoice for iOS with Quantized VoxCPM2. Zenodo, 2026.

[3] Guevara, R. C. L., Cajote, R. D., Bayona, M. G. A. R., and Lucas, C. R. G. Philippine Languages Database. SIGUL, 2024.

[4] Pratap, V., et al. Massively Multilingual Speech Cebuano TTS checkpoint. Meta.

[5] Zhu, H., et al. OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models. arXiv:2604.00688, 2026.

[6] Conneau, A., et al. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. arXiv:2205.12446, 2022.

[7] Singh, S., Yadavalli, A., Arnett, C., and Warstadt, A. BuzzASR: A Swarm of 100+ Monolingual Speech Recognition Models. arXiv:2609.09554, 2026.

[8] Helsinki-NLP. OPUS-MT English to Cebuano. Pinned model card and weights.

[9] tc-mb. VoxCPM2 GGUF runtime. llama.cpp-omni, pinned revision.

[10] Microsoft. ONNX Runtime weight quantization documentation.

[11] MLX Community. VoxCPM2 four-bit model, pinned revision.

[12] QnVoice. Cebuano voice profile, evaluation record, and audio examples. Hugging Face, 2026.

[13] ITU-T. Method for subjective performance assessment of the quality of speech voice output devices. Recommendation P.85.

Appendix A Tested Phrases and Audio

The first six links open waveforms generated by the final iPhone test. The last three are desktop outputs from the separate comparison with the shared reference. The English glosses describe the Cebuano source phrases.

Table A1 Listening examples

Source	Cebuano input	English gloss	Audio
iPhone	Maayong buntag.	Good morning.	WAV
iPhone	Daghang salamat sa imong tabang.	Thank you very much for your help.	WAV
iPhone	Gusto ko og usa ka baso nga tubig.	I would like a glass of water.	WAV
iPhone	Palihog, hinayi og sulti.	Please speak slowly.	WAV
iPhone	Kumusta ka? Asa ta mangaon?	How are you? Where shall we eat?	WAV
iPhone	Magkita ta ugma sa buntag.	See you tomorrow morning.	WAV
Mac	Ayaw kalimot sa imong payong.	Do not forget your umbrella.	WAV
Mac	Tagpila ang usa ka kilo nga mangga?	How much is a kilo of mangoes?	WAV
Mac	Bugnaw ba ang tubig?	Is the water cold?	WAV

Run the Published Voice

Download the model profile and build a VoxCPM2 GGUF CLI using the pinned runtime. The first command downloads the exact shared weights and verifies their checksums. Subsequent synthesis requests use the local files.

```
python3 speak.py --fetch --verify
python3 speak.py --text 'Maayong buntag.' --output cebuano.wav
```

Offline model package · Full per-phrase evaluation record