

StepAudio 3: A Voice Model That Thinks While It Speaks, and a Music Model That Plans the Song First

Kabui, Charles

2026-09-22

[Read at ToKnow.ai](#)



The graphic features a dark blue background with a subtle grid and a faint line-art illustration of a person's head and shoulders in the upper right. The main title is split into two lines: 'StepAudio 3: A Voice Model That Thinks While It Speaks' in white and 'A Music Model That Plans First' in a light blue/green. Below the title, a subtitle reads 'Two StepFun audio models that plan before they produce'. Three key metrics are highlighted in separate boxes: a red box for '98.9 Full-duplex speech score reported by StepFun', a yellow box for '1092 Elo on the independent music leaderboard, fourth place', and a blue box for '5:30 Longest single song the music model renders'. A horizontal line separates these from the text 'One model reasons while it talks; the other writes the arrangement first'. At the bottom left is the date 'September 22, 2026' and at the bottom right is the 'ToKnow.ai' logo.

StepAudio 3: A Voice Model That Thinks While It Speaks A Music Model That Plans First

Two StepFun audio models that plan before they produce

98.9 Full-duplex speech score reported by StepFun	1092 Elo on the independent music leaderboard, fourth place	5:30 Longest single song the music model renders
---	---	--

One model reasons while it talks; the other writes the arrangement first

September 22, 2026

ToKnow.ai

StepFun's StepAudio 3 family is five AI audio models covering speech, sound effects and music. Its two reports, from September 11 and 12, share one idea: settle the hard part before you hear the output. StepAudio 3 Realtime is a full-duplex speech model, meaning you and it can talk

at once, built on a listen, converse, think and act loop. Think-While-Speaking runs private reasoning alongside speech, so it starts answering while still working a question out, then folds in the tool results. StepAudio 3 Music works the other way: it writes the arrangement out first, in a text shorthand called ABC notation, so harmony and song structure are decided rather than improvised, then renders stereo songs up to 5 minutes 30 seconds long.

That removes the usual tradeoff between a voice model that reasons and one that answers fast. StepFun reports 98.9 on Artificial Analysis' full-duplex bench and 90.6 on MMSU, an audio-understanding test, but the model is absent from that site's main speech leaderboards, so the numbers are hard to check. Its own paper notes gaps in multi-turn constraint following, and on tau-Voice, a customer-service benchmark, it scores 56.0%, just behind Grok Voice Think Fast 2.0's 56.5%. Music is easier to verify: it places fourth on the independent Artificial Analysis music leaderboard, Elo 1092 from 2,043 blind votes, ahead of Suno V5 and Google's Lyria 3 Pro.

Neither ships open weights, and StepFun lists both as free previews with paid versions to come. In both cases the gain comes from an inspectable middle step, a written plan or a private reasoning track, not from one network doing both at once.

Read More: [Google's Lyria 3 generates full songs and lets you steer them while they play](#)

Sources:

- [StepAudio 3 Realtime Technical Report \(arXiv\)](#)
- [StepAudio 3 Music Technical Report \(arXiv\)](#)
- [StepFun documentation: StepAudio 3 Realtime models and preview access](#)
- [StepAudio 3 Music demos and capability walkthrough \(StepFun\)](#)
- [Artificial Analysis music arena vocals leaderboard, blind votes](#)

***Disclaimer:** For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)*