TTS 2026: OmniVoice, Qwen3-TTS, Chatterbox, Kokoro, NeuTTS, VoxCPM2 und Gepard 1.0 im Vergleich

Die Open-Source-Landschaft für Text-to-Speech (TTS) hat sich in den letzten 12 Monaten dramatisch verändert. Was früher nur mit teuren APIs wie ElevenLabs möglich war – Zero-Shot-Voice-Cloning, emotionale Kontrolle, Streaming in Echtzeit – gibt es heute frei verfügbar zum Download. In diesem Artikel werfen wir einen detaillierten Blick auf sieben der spannendsten aktuellen TTS-Modelle: **OmniVoice**, **Qwen3-TTS**, **Chatterbox**, **Kokoro**, **NeuTTS**, **VoxCPM2** und **Gepard 1.0**.

Kurzüberblick: Die Modelle auf einen Blick

Modell	Entwickler	Parameter	Lizenz	Kernstärke
OmniVoice	k2-fsa	– (Diffusion-LM-Architektur)	Apache 2.0	600+ Sprachen, breiteste Sprachabdeckung
Qwen3-TTS	Alibaba (Qwen-Team)	0.6B / 1.7B	Apache 2.0	Ultra-Low-Latency-Streaming, Instruction-Control
Chatterbox	Resemble AI	0.5B	MIT	Emotion-Exaggeration, schlägt ElevenLabs in Blindtests
Kokoro	hexgrad (Indie)	82M	Apache 2.0	Winziges, extrem effizientes Modell

Modell	Entwickler	Parameter	Lizenz	Kernstärke
NeuTTS (Air/Nano/2E)	Neuphonic	120–360M aktiv	Apache 2.0 / NeuTTS Open License	On-Device/Offline, läuft auf CPU/Raspberry Pi
VoxCPM2	OpenBMB	2B	Open Source	Tokenizer-freie Diffusion, 48kHz Studio-Qualität
Gepard 1.0	nineninesix	555M	Open Source	Streaming-first, ~50ms Time-to-First-Audio

---

1. OmniVoice – Der Sprachen-Champion

OmniVoice von k2-fsa (bekannt aus dem sherpa-onnx-Ökosystem) ist ein massiv multilinguales Zero-Shot-TTS-Modell und unterstützt beeindruckende **über 600 Sprachen** – das ist die derzeit breiteste Sprachabdeckung unter allen Zero-Shot-TTS-Systemen überhaupt.

Highlights:

- Basiert auf einer neuartigen **Diffusion-Language-Model-Architektur**
- Voice Cloning bereits mit 3–10 Sekunden Referenzaudio
- **Voice Design** über Attribute wie Geschlecht, Alter, Tonhöhe, Akzent/Dialekt (inkl. chinesischer Regionaldialekte wie Sichuanesisch) – ganz ohne Referenzaudio
- Feinsteuerung über Inline-Tags (`[laughter]`, `[sigh]`) und Aussprachekontrolle via Pinyin oder CMU-Phonemen
- Sehr schnelle Inferenz: RTF (Real-Time Factor) bis zu 0,025 – also **40× schneller als Echtzeit**
- Lizenz: Apache 2.0, auf GitHub bereits 8.500+ Stars

Einsatzgebiete: Ideal für Projekte, die tatsächlich globale Sprachabdeckung brauchen (Lokalisierung, Sprachassistenten für Nischensprachen), weniger für Ultra-Low-Latency-Dialoganwendungen.

---

2. Qwen3-TTS – Alibabas Streaming-Kraftpaket

Alibabas Qwen-Team hat mit **Qwen3-TTS** ein technisch besonders ausgereiftes Modell veröffentlicht, das in zwei Größen erscheint: **0.6B** und **1.7B** Parameter, jeweils in den Varianten Base (Voice Clone), CustomVoice (feste Premium-Stimmen) und VoiceDesign (Stimme per Text-Prompt erzeugen).

Highlights:

- Eigener Tokenizer **Qwen3-TTS-Tokenizer-12Hz** – erzielt in Benchmarks Bestwerte bei PESQ, STOI und UTMOS gegenüber Konkurrenten wie Mimi oder X-Codec2
- **Dual-Track-Streaming-Architektur**: Ein Modell für Streaming UND Nicht-Streaming, End-to-End-Latenz bis **97ms**
- Deckt 10 Hauptsprachen ab: Chinesisch, Englisch, Japanisch, Koreanisch, Deutsch, Französisch, Russisch, Portugiesisch, Spanisch, Italienisch
- In Benchmarks (Seed-TTS-Test) schlägt Qwen3-TTS-12Hz-1.7B-Base sogar CosyVoice3 und MiniMax-Speech bei der Wortfehlerrate (WER) im Englischen (1,24 vs. 1,45/1,65)
- Bei Sprecherähnlichkeit (SIM) und Instruction-Following liegt es klar vor ElevenLabs und GPT-4o-Audio
- Day-0-Support in vLLM-Omni für produktionsreife Bereitstellung
- Lizenz: Apache 2.0, 12.600+ GitHub-Stars

Einsatzgebiete: Sehr starke Allround-Wahl für mehrsprachige Sprachassistenten und Voice-Agents, besonders wenn Instruction-basierte Emotionssteuerung gewünscht ist.

---

3. Chatterbox – Der ElevenLabs-Herausforderer

Chatterbox von Resemble AI war einer der ersten Open-Source-Releases, der es 2025 schaffte, in Blindtests gegen ElevenLabs zu bestehen – und ist mittlerweile in Version **Multilingual V3** angekommen.

Highlights:

- 0,5B-Parameter-Modell mit **Llama-3-Backbone**
- Als erstes Open-Source-TTS-Modell mit **Emotion-Exaggeration-Control** (regelbare emotionale Übertreibung via `exaggeration`- und `cfg`-Parameter)
- Multilingual V3 deckt **23+ Sprachen** ab, inklusive dedizierter "Single Language Packs" für Chinesisch, Hindi und mehrere Spanisch-/Portugiesisch-Varianten
- Trainiert auf 0,5 Mio. Stunden Audiodaten
- Eingebautes **PerTh-Wasserzeichen** (überlebt MP3-Kompression) für Responsible-AI-Zwecke
- MIT-Lizenz – die freieste Lizenz unter allen hier verglichenen Modellen

- Über 2,5 Mio. Downloads/Monat auf Hugging Face

Einsatzgebiete: Sehr beliebt für kreative Inhalte (Memes, Videos, Games), Voice-Agents mit emotionalem Ausdruck – Referenz-Community-Favorit.

---

4. Kokoro – Klein, aber oho

Kokoro-82M ist das Gegenstück zu den riesigen Modellen: Mit nur **82 Millionen Parametern** liefert es laut Entwickler hexgrad Qualität, die mit deutlich größeren Modellen konkurrieren kann.

Highlights:

- Basiert architektonisch auf **StyleTTS 2**
- Extrem ressourcenschonend – läuft flüssig auf CPU, sogar im Browser (Kokoro.js via ONNX)
- Nutzt die eigene G2P-Bibliothek **misaki**
- Unterstützt mehrere Sprachvarianten (US-/UK-Englisch, Spanisch, Französisch, Hindi, Italienisch, Japanisch, Brasilianisches Portugiesisch, Mandarin)
- Apache-2.0-Lizenz, 8.100+ GitHub-Stars
- Kein Voice-Cloning-Feature – arbeitet mit vordefinierten Stimm-Presets (z. B. `af_heart`)

Einsatzgebiete: Perfekt für kosteneffiziente Massenproduktion von Sprache (Podcasts, Hörbuch-Vertonung, eingebettete Anwendungen), wo Server-Kosten und Latenz wichtiger sind als Voice-Cloning.

---

5. NeuTTS – On-Device-Champion

Die **NeuTTS**-Familie von Neuphonic (Air, Nano-Multilingual-Collection, 2E) ist explizit für den **offline, on-device Betrieb** konzipiert – bis hin zum Raspberry Pi.

Highlights:

- Verschiedene Backbone-Größen: NeuTTS-Air (~360M aktiv), NeuTTS-Nano (~120M aktiv), NeuTTS-2E (~125M aktiv, mit Emotionssteuerung)
- Eigener Audio-Codec **NeuCodec** (50Hz, Single-Codebook, hohe Qualität bei niedriger Bitrate)
- **GGUF-Quantisierungen** (Q4/Q8) für maximale Effizienz auf CPU
- Benchmark-Werte beeindruckend: Auf einem AMD Ryzen 9HX 370 erreicht NeuTTS-Nano 221 Tokens/Sekunde rein auf CPU
- Instant Voice Cloning mit nur 3 Sekunden Referenzaudio

- Unterstützt Englisch, Spanisch, Deutsch, Französisch (modellabhängig)
- Lizenzmix: NeuTTS-Air unter Apache 2.0, neuere Modelle (Nano, 2E) unter der "NeuTTS Open License 1.0"
- Eingebautes Perth-Wasserzeichen
- Auf GitHub bereits 6.200+ Stars, sehr aktive Entwicklung (fast täglich neue Commits)

Einsatzgebiete: Ideal für Privacy-first-Anwendungen, eingebettete Sprachassistenten, Spielzeuge oder Compliance-sensible Apps, bei denen keine Daten das Gerät verlassen dürfen.

---

6. VoxCPM2 – Tokenizer-freie Diffusion aus China

VoxCPM2 von OpenBMB (bekannt von den MiniCPM-Sprachmodellen) verfolgt einen architektonisch besonders interessanten Ansatz: Es ist **tokenizer-frei** und generiert kontinuierliche Sprachrepräsentationen direkt über ein End-to-End-Diffusion-Autoregressive-Modell.

Highlights:

- **2 Milliarden Parameter** – das größte Modell in diesem Vergleich
- Kein diskreter Audio-Tokenizer nötig – arbeitet direkt mit kontinuierlichen Repräsentationen, was Informationsverlust reduziert
- Unterstützt **30+ Sprachen**
- Ausgabe in **48kHz Studio-Qualität**
- Trainiert auf über 2 Millionen Stunden Sprachdaten
- Laut Community-Berichten erreicht VoxCPM2 im Similarity-Benchmark rund **85,4 %** bei englischer Sprache und gilt als eines der Modelle, die ElevenLabs bei der Sprecherähnlichkeit übertreffen
- State-of-the-art oder konkurrenzfähige Ergebnisse bei wichtigen Zero-Shot- und Controllable-TTS-Benchmarks
- Open Source über GitHub/Hugging Face

Einsatzgebiete: Für Anwendungen, die höchste Audioqualität (48kHz) und kontextbewusste, ausdrucksstarke Sprache brauchen – z. B. professionelle Synchronisation, Hörbücher, kreative Content-Produktion.

---

7. Gepard 1.0 – Der Sprint-Champion für Echtzeit-Dialoge

Gepard 1.0 (passenderweise nach dem schnellsten Landtier benannt) von **nineninesix** ist das jüngste und auf reine **Streaming-Performance** getrimmte Modell in diesem Vergleich.

Highlights:

- 555 Millionen Parameter, **Qwen3.5-Backbone**
- **Streaming-first**: Das Modell beginnt zu sprechen, sobald der erste Text-Chunk eintrifft – Audio wird frame-weise generiert
- Beeindruckende Performance-Kennzahlen: **~50ms Time-to-First-Audio (TTFA)** und **~20× Real-Time-Factor** auf einer einzelnen RTX 5090
- Explizit für **Real-Time-Dialogue-Synthese** konzipiert – also für Sprachassistenten und Konversations-KI, bei denen jede Millisekunde Latenz zählt
- Bereits Tag-0-Integrationsarbeit in vLLM-Omni in Vorbereitung
- Sehr frisch: erst vor wenigen Wochen open-sourced, aber schon große Aufmerksamkeit auf Reddit/HackerNews/X

Einsatzgebiete: Die klare Nummer 1 für latenzkritische Voice-Agent- und Conversational-AI-Anwendungen (z. B. Telefonie-Bots, Live-Übersetzung), wo eine Reaktionszeit im Millisekundenbereich entscheidend ist.

Direkter Vergleich nach Anwendungsfall

Anwendungsfall	Empfehlung
Maximale Sprachabdeckung (Nischensprachen)	**OmniVoice** (600+ Sprachen)
Bester Allround-Multilingual-Agent mit Instruction-Control	**Qwen3-TTS**
Emotionale/kreative Inhalte, MIT-Lizenz gewünscht	**Chatterbox**
Extrem günstige Massenproduktion (CPU/Browser)	**Kokoro**
Offline/On-Device/Privacy (Raspberry Pi, Mobile)	**NeuTTS**
Höchste Audioqualität (48kHz Studio-Sound)	**VoxCPM2**
Echtzeit-Dialog mit minimaler Latenz	**Gepard 1.0**

Gemeinsame Trends

Bei der Recherche fallen mehrere Muster auf, die die gesamte Open-Source-TTS-Szene 2026 prägen:

1. **LLM-Backbones sind Standard geworden**: Chatterbox nutzt Llama 3, NeuTTS und Gepard bauen auf Qwen-Backbones – TTS-Modelle werden zunehmend wie kleine Sprachmodelle behandelt.
2. **Voice Cloning mit wenigen Sekunden Audio** ist mittlerweile Basisfunktion, nicht mehr das Alleinstellungsmerkmal.
3. **Watermarking** (meist via Resemble AIs Perth) setzt sich als Responsible-AI-Standard durch – gleich mehrere Modelle (Chatterbox, NeuTTS) integrieren es standardmäßig.
4. **Streaming-Fähigkeit** wird zum entscheidenden Wettbewerbsfaktor – sowohl Qwen3-TTS als auch Gepard 1.0 werben explizit mit Latenzen unter 100ms.
5. **Apache 2.0/MIT dominieren** als Lizenzmodell, was kommerzielle Nutzung stark vereinfacht (Ausnahme: NeuTTS-Nano/2E mit eigener "Open License").

Fazit

Es gibt kein "bestes" TTS-Modell – die Wahl hängt vollständig vom Anwendungsfall ab. Wer **Reichweite** über möglichst viele Sprachen sucht, landet bei OmniVoice. Wer ein **produktionsreifes Allround-System** mit starker Instruction-Steuerung will, ist mit Qwen3-TTS gut bedient. Für **kreative, emotionale Sprachausgabe** unter freizügiger Lizenz bleibt Chatterbox der Community-Favorit. Wer **Kosten minimieren** will, greift zu Kokoro. **Datenschutz und Offline-Betrieb** sprechen für NeuTTS, **Studio-Qualität** für VoxCPM2 – und wer eine **blitzschnelle Sprachausgabe für Echtzeit-Konversationen** braucht, sollte sich Gepard 1.0 genau ansehen.

Die gute Nachricht: Alle sieben Modelle sind Open Source und lassen sich kostenlos selbst hosten – ein enormer Fortschritt gegenüber der Zeit, in der hochwertige TTS-Qualität ausschließlich hinter kostenpflichtigen APIs verschlossen war.

Revision #1
Created 2026-07-25 17:02:02 UTC by art10m
Updated 2026-07-25 17:03:45 UTC by art10m