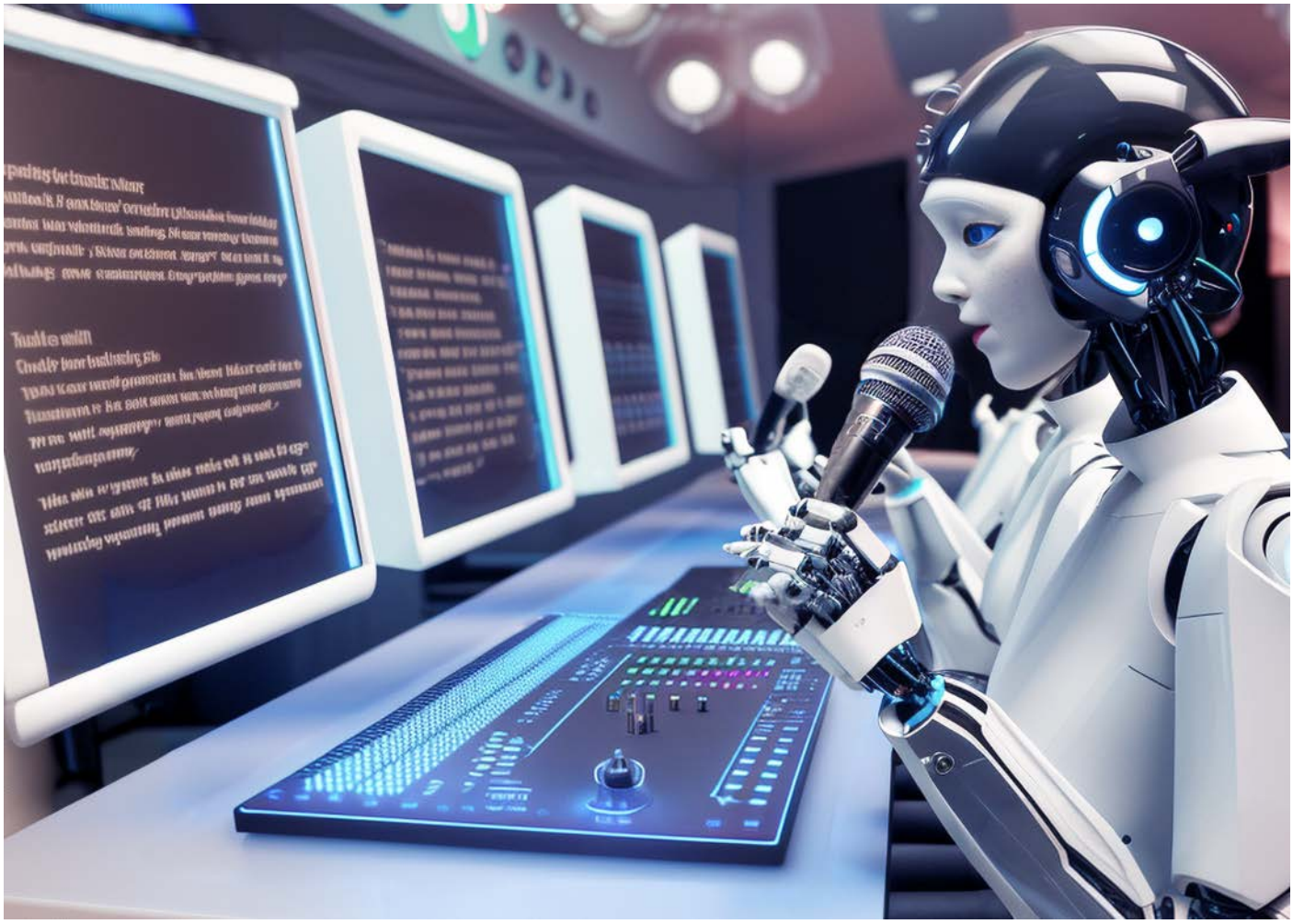


Bild: KI Stable Diffusion | Bearbeitung c't

Stimmprobe

Sieben Anbieter von KI-Stimmen für Text-to-Speech im Vergleich

Dank künstlicher Intelligenz klingt computergenerierte Sprache immer natürlicher. Inzwischen will sie sogar menschliche Sprecher ersetzen und diese klonen können. Wie gut das funktioniert, haben wir anhand von sieben TTS-Diensten untersucht.

Von Kai Schwirzke

Glaubt man den Anbietern, so sprechen künstliche Stimmen mittlerweile Texte auf Knopfdruck in wenigen Sekunden professionell ein. KI-basierte Text-to-Speech-Algorithmen sollen eine so realistische Sprachausgabe erreichen, dass man die maschinelle Herkunft nicht mehr erkennt. Aus dem kaum überschaubaren Angebot haben wir sieben interessante Dienste ausgewählt und getestet. Die Auswahl deckt einen Querschnitt der aktuellen Online-Angebote ab und reicht von günstigen Start-Ups wie ElevenLabs und Speecheasy über Anbieter mit hunderten verschiedener Stimmen wie Beepbooply und Uberduck, Spezialisten für Dialoge wie Coqui, Videovertonung wie Murf bis hin zu teureren Angeboten wie Revoicer, die mit besonders emotionalen Stimmen werben.

Dabei hat uns vor allem interessiert, welchen Mehrwert diese Systeme gegenüber der mittlerweile in jedem modernen Betriebssystem integrierten Sprachausgabe bieten. Denn unter Windows und macOS (unter Linux muss man nachinstallieren) lesen männliche und weibliche Stimmen Textdokumente, Webseiten oder E-Mails bereits in ordentlicher Qualität vor. Besonders gut gelingt dies in Eng-

lisch, Deutsch klingt oft holpriger. Um Menschen mit eingeschränktem Sehvermögen Inhalte zu vermitteln, reicht die Qualität der integrierten Stimmen jedoch allemal aus.

Vier der untersuchten Online-Dienste (Beepbooply, ElevenLabs, Murf und Revoicer) sprechen neben Englisch und anderen Sprachen auch Deutsch. Coqui, ElevenLabs, Murf und Uberduck können zudem Stimmen klonen. Alle Anbieter arbeiten browserbasiert. Mit Ausnahme von Revoicer können Sie alle Dienste kostenlos ausprobieren. Ein Download der Audiodaten mit den gesprochenen Texten ist oft erst nach Abschluss eines zahlungspflichtigen Abos möglich, dessen Einstiegspreise von 5 bis 30 US-Dollar pro Monat reichen. Nach einer Zahlung können Sie die Aufnahmen bei allen Anbietern herunterladen und fortan kommerziell auch nach Ende der Abozeit nutzen.

Alle Testkandidaten haben das gleiche Bedienkonzept: Man kopiert den zur Vertonung vorgesehenen Text in das Eingabefeld, setzt einige Parameter wie Sprache und Geschlecht und lässt die KI rechnen. Zur weiteren Verarbeitung laden Sie die heruntergeladene Datei in einen Audio- oder Videoeditor Ihrer Wahl. Coqui und Murf bieten zusätzliche Editoren für Dialoge an. Coqui ist zudem der einzige Anbieter ohne Abozwang.

Podcasts, Videovertonung, Spiele

KI-Stimmen kommen dort zum Einsatz, wo bisher menschliche Sprecher vor das Mikrofon traten, etwa bei Podcasts oder Video-Tutorials. Auch Spieleentwickler haben nicht immer das nötige Budget, um gute Sprecherinnen und Sprecher zu engagieren. Zudem will ordentliches Aufnahmeequipment bezahlt werden, und professionell klingende Sprachaufnahmen sind nicht in jedem Wohnzimmer möglich.

Wenn ein international tätiger Anbieter Video-Tutorials in mehreren Sprachen erstellen will, erhöht sich der Aufwand enorm. Nicht nur, dass der Text übersetzt werden muss. Für jede Version wird auch ein Sprecher benötigt, der die Zielsprache einigermaßen beherrscht. Das Zusammenspiel von KI-Übersetzern wie DeepL, deren Texte ein multilinguales TTS-System vertont, kann dabei viel Zeit und Geld sparen. Nicht zuletzt nimmt eine Aufnahmesession mit menschlichen Sprechern und deren Bearbeitung deutlich mehr Zeit in Anspruch als die eigentliche Sprechzeit. All das kostet Geld, das kommerzielle Anbieter gerne sparen. Und auch der Amateur, der aus Spaß gelegentlich Video-Tutorials auf YouTube stellt, fühlt sich vielleicht besser, wenn er eine knifflige Passage nicht ständig wiederholen muss, weil er sich regelmäßig verhaspelt.

Vorbereitungen und Nacharbeiten

Wer KI-generierte Stimmen zur Vertonung eines Videos einsetzen möchte, muss allerdings mit umfangreichen Vorbereitungen und Nacharbeiten rechnen. Während menschliche Kommentatoren bereits bei der Aufnahme darauf achten, bildunterstützend zu sprechen, also passend zum Bild und zum Originalton, müssen Sie die Sprachausgabe des KI-Kollegen in Ihrer Videoschnitt-App manuell nachbearbeiten. Denn die KI kennt weder die Länge noch den Inhalt oder den Schnitt Ihres Videos. Bei Murf funktioniert die Bearbeitung dank einer Zeitleiste etwas einfacher als bei den übrigen Anbietern.

Die zu sprechenden Texte sollten Sie in jedem Fall in einem Skript vorbereiten. Denn die Ausgabe der TTS-Stimmen wird stets nach der Anzahl der Buchstaben oder der gesprochenen Zeit berechnet. Jeden neuen Durchgang lassen sich die Anbieter bezahlen.

Damit nicht alles monoton klingt, bieten alle Dienste mindestens ein Dutzend verschiedene Sprecherinnen und Sprecher an, die teilweise sogar Dialekte intonieren. So gibt es gerade im englischsprachigen Raum ein umfangreiches Angebot vom breiten Südstaaten-Slang über gepflegtes Westküsten-Englisch bis zum indischen oder schottischen Zungenschlag.

Bei den meisten anderen europäischen Sprachen ist die Sprechervielfalt jedoch geringer. Dies gilt insbesondere für die deutschen Stimmen von Beepbooply, ElevenLabs, Murf und Revoicer. Allerdings bauen die Anbieter ihr Repertoire an internationalen Sprechern tagtäglich aus, sodass die Auswahl in einigen Wochen und Monaten schon deutlich besser aussehen könnte.

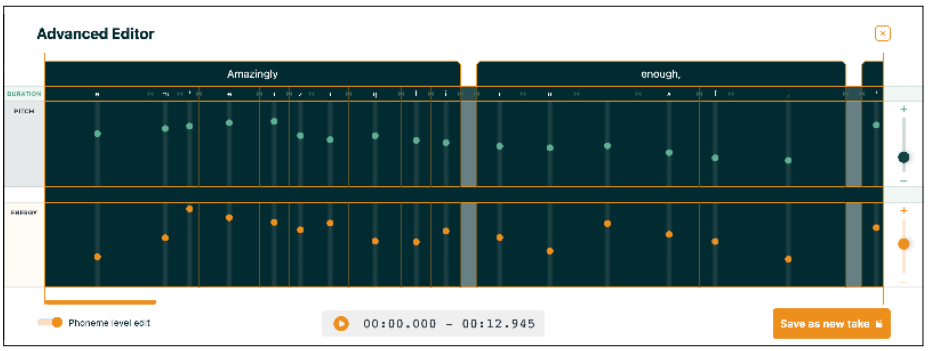
Authentizität

Um das Versprechen der KI-Dienste nach lebensnahen Stimmen einzulösen, müssen die Algorithmen die Sprache kontextabhängig modulieren, also je nach Inhalt Hebungen, Senkungen und Temposchwankungen einbauen. Das klappt bei allen Anbietern ganz gut, wenn auch nicht überragend. Die KIs erahnen nur ansatzweise, ob etwas lustig, traurig, spannend oder langweilig ist.

Das hat aber auch Vorteile: Zu viel Pathos nervt. Kulturen reagieren sehr unterschiedlich auf verbalisierte Emotionen. Wo in Norddeutschland oft ein schlichtes „Jo!“ genügt, um Ablehnung oder Zustimmung auszudrücken, mag man es in Italien lieber deftiger. Emotionale Kompetenz haben KI-Stimmen nur, wenn sie landestypisch trainiert wurden.

c't kompakt

- Aktuelle TTS-KIs geben Texte erstaunlich authentisch wieder, besonders in Englisch.
- Gute deutsche Stimmen sind selten zu finden.
- Manche Dienste erlauben eine manuelle Korrektur der Aussprache und Betonung.



Im Editor von Coqui korrigiert man die Aussprache bis hinunter zur Phonem-Ebene.

Dies funktioniert derzeit noch nicht länderübergreifend.

Immerhin reagieren alle von uns getesteten Anbieter auf Satzzeichen: Bei Komma, Punkt, Bindestrich oder Doppelpunkt fügen sie eine kleine Sprechpause ein. Das kann und sollte man nutzen, auch wenn Vater Duden sich wegen der zusätzlichen Kommata im Grabe umdreht. Mit wenig Aufwand klingt eine KI-Stimme so viel natürlicher. Auf die Länge der Pausen hat man aber meist keinen Einfluss. Die Engine wartet so lange, wie sie es für richtig hält. In Murf können Sie Aussprachefehler durch die Wahl verschiedener Lautschriftvarianten korrigieren. Wesentlich umfangreicher sind solche Eingriffe per SSML (Speech Synthesis Markup Language). Diese von Google erfundene und von einigen Diensten unterstützte Auszeichnungssprache ähnelt HTML oder XML. Sie erlaubt es, Ausführungsanweisungen in einen Text einzufügen. So sorgt beispielsweise `<prosody speed="slow" pitch="-1st">Aber gut aufpassen!</prosody>` dafür, dass der Text „Aber gut aufpassen“ langsamer und einen Halbton (st = semi tone) tiefer synthetisiert wird. `<break time=200ms>` fügt eine Sprechpause von 200 Millisekunden ein.

Kann ein Sprachmodell wie Speecheasy solche Anweisungen umsetzen, klingen KI-Sprecher deutlich natürlicher. Der Aufwand, einen längeren Text mit solchen Tags auszustatten, ist allerdings erheblich. Außerdem muss man – wie bei jeder Programmiersprache – mit mühseligem Debugging rechnen. Vergessene spitze Klammern oder einen fehlenden Schrägstrich quittieren die Dienste mit Fehlermeldungen.

Der Weg zur idealen KI-Stimme ist also mühsam. Zudem sind die Iterationen kostspielig. Denn jeder neue Durchlauf belastet das monatliche Kontingent der erlaubten Textmenge, auch wenn man ihn später gar nicht verwenden kann.

Geklonte Stimmen

Drei der getesteten Anbieter (Coqui, ElevenLabs und Uberduck) können Stimmen eines menschlichen Sprechers direkt im Browser-Interface klonen. Dazu lädt man üblicherweise eine kurze Audiodatei hoch, die den zu kopierenden Sprecher möglichst ohne Umgebungsgeräusche enthält. Nach einer kurzen Rechenpause steht dann die gewünschte Stimme zur Verfügung. Die Qualität der Klone reicht bei Coqui und ElevenLabs allerdings nicht an die eleganteren Betonungen und Sprachmelodien der vorproduzierten Stimmen heran.

Wer eine höhere Qualität benötigt, dem raten die Anbieter ein intensiveres Training. Das ist auf individuelle Anfragen möglich, bei denen nicht zuletzt auch die rechtlichen Modalitäten zu klären sind. Wie ein professionelles Training für eine Klonstimme aussehen kann, haben wir in [2] näher beleuchtet. Ein solches Sondertraining bietet auch Murf an, bislang aber nur für englische Stimmen. Im Rahmen dieses Tests haben wir auf solche Spezialtrainings verzichtet.

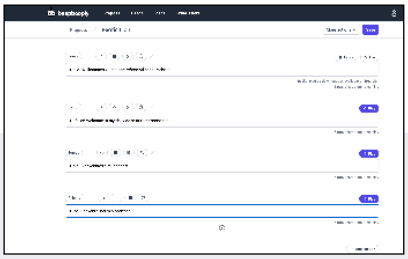
Im Hinblick auf die Performance und Verfügbarkeit der Dienste konnten wir keine signifikanten Unterschiede zwischen den Anbietern feststellen. Das Wandeln einer kurzen Textdatei in eine Audiodatei dauerte bei allen Diensten nur wenige Sekunden. Wer aber 5.000 Zeichen und mehr in Sprache umsetzen lässt, sitzt schon mal einige Minuten am Rechner, bis der Durchlauf fertig ist.

Rechtliches

Wie es um die Datensicherheit bestellt ist, bleibt bei allen Testkandidaten unklar. Aus den Angaben auf den Webseiten geht nicht hervor, ob, wie und wo sie die Daten speichern und wer die Texte eventuell zu Gesicht bekommt. Keiner der Anbieter klärt deutsche Kunden in deutscher Sprache über ihre Datenschutzrechte auf, wie es die DSGVO verlangt. Lediglich Coqui, ElevenLabs und Murf gehen auf Besonderheiten der DSGVO überhaupt ein. Speecheasy hat nicht einmal ein Impressum.

Gesunde Skepsis ist insbesondere bei der zweifellos faszinierenden Technik des Stimmenklonens angebracht. Denn mit dem digitalen Abbild der eigenen Stimme lässt sich einiges an Schaden anrichten. Zum Beispiel, wenn das Modell durch kompromittierte Zugangsdaten in falsche Hände gerät. So konnte ein amerikanischer IT-Journalist mit seiner geklonten Stimme das Telefonbanking austricksen und auf sein Konto zugreifen. Käme eine Fremde an seine Klonstimme, hätte sie ebenfalls Zugriff auf das Konto.

Ebenso könnten vermeintliche Scherzbolde gefälschte „Originalstimmen“ in Umlauf bringen, was nicht nur ärgerlich, sondern – schlimmer noch – rufschädigend oder gar existenzbedrohend werden kann. Es ist daher ratsam, sich genau zu überlegen, ob die Digitalisierung der eigenen Stimme wirklich sinnvoll ist.



Beepbooply

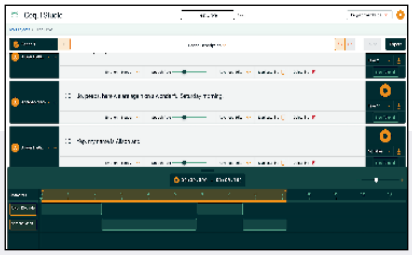
Der KI-Dienst mit dem drolligen Namen redet nicht lange um den heißen Brei herum: Die weit über 900 Stimmen aus aller Herren Länder stammen aus den Modellen von Amazon, Google und Microsoft. Wie so oft in diesem Testfeld gibt sich das Webinterface mehr als schlicht: In eine Textbox, Section genannt, fügt man seine zu sprechenden Inhalte ein, wählt Sprache und Stimme aus und klickt auf Play. Ein Projekt besteht aus beliebig vielen solcher Sektionen, die Sie individuell in Geschwindigkeit, Tonhöhe und Lautstärke verändern können. Diese Einstellungen gelten für den gesamten Text einer Sektion. Zwischen einzelnen Wörtern darf man zudem beliebig lange Pausen setzen.

Die Auswahl der passenden Stimme gelingt angesichts der schier unüberschaubaren Anzahl von Stimmen relativ gut. Hat man sich für ein Land (die Liste reicht von Afrikaans bis Zulu) und ein Geschlecht entschieden, kann man das nun gefilterte Angebot anhand eines Beispielsatzes kostenlos vorhören.

Qualitativ bietet Beepbooply eine große Bandbreite von „geht so“ bis „ziemlich gut“, was auch daran liegt, dass der Dienst zwischen Basic und Realistic Voices unterscheidet. Letztere kann man mit einem kostenlosen Zugang ausprobieren, aber nicht herunterladen. Ohne Abonnement beschränkt Beepbooply die monatliche Textmenge auf 10.000 Zeichen.

+ große Auswahl an Stimmen
+ darunter brauchbare deutsche
– schwankende Qualität der Stimmen

Preise: kostenlos 10.000 Zeichen pro Monat, monatlich ab 8 Euro



Coqui

Coqui richtet sich vor allem an Produzenten von Podcasts und Videos mit mehreren Sprechern. Die Entwickler sitzen zwar in Berlin, bieten aber bisher nur englische Stimmen an. Auch die Datenschutzerklärung ist auf Englisch verfasst.

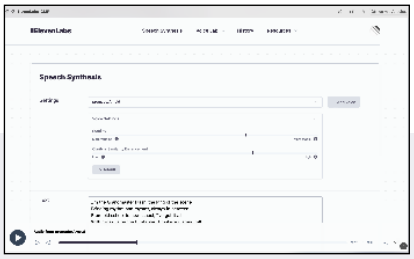
Die Benutzeroberfläche unterscheidet sich wohltuend von der Konkurrenz. Ein Projekt besteht aus mehreren Textzeilen, denen verschiedene Sprecher zugeordnet werden. Für jede Spur können mehrere Takes, also Variationen, erzeugt werden. Die KI spricht beispielsweise Take 1 neutral und Take 2 lebhaft. In einer zweiten Ansicht sieht man die einzelnen Sprecher und ihre Textzeilen sortiert auf einem Zeitlineal, ganz so wie bei einem Musikproduktionsprogramm. So erzeugt man im Handumdrehen ein Gespräch.

Mit komfortablen Werkzeugen verleiht man den Audioaufnahmen den letzten Schliff. Wer möchte, verändert im Advanced Editor die Länge und Lautstärke einzelner Wörter oder Intensitäten und Tonhöhen auf Phonemebene.

Reichen die angebotenen 30 englischen Stimmen nicht aus, kann man auch eigene Sprecher trainieren. Das Stimmen-Cloning kommt mit einem etwa 30 Sekunden langen Audiobeispiel aus. Die Ergebnisse sind erstaunlich, wenn auch nicht perfekt. Außerdem lassen sich zwei vorhandene Stimmen zu einer neuen verschmelzen oder Sie generieren eine Stimme per Prompt: „An older man with a British accent and a pleasing, deep voice."

- optimiert für Dialoge
- gute Klone
- keine deutschen Stimmen

Preise: 30 Minuten kostenlos, 4 Stunden für 20 US-Dollar



ElevenLabs

ElevenLabs haben wir bereits in [1] für künstliche Rapper vorgestellt. Das Webinterface ist denkbar einfach: Text eingeben, „Generieren" klicken, fertig. Zuvor wählt man noch eine von acht Sprachen und den Sprecher aus, neuerdings auch Deutsch. Außerdem gibt es die Parameter Stability und Clarity+-Enhancement. Beide dienen dazu, die Balance zwischen einer stabilen Sprechstimme und einer etwas menschlicheren Performance zu finden.

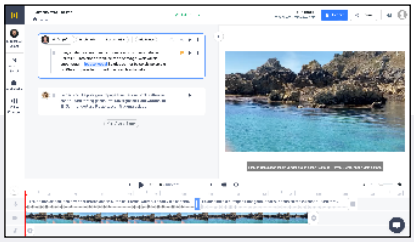
Bereits mit einem kostenlosen Konto hat man Zugang zum Voice Lab, in dem man per Zufallsgenerator neue Stimmen erzeugt und abspeichert. Der Nutzer gibt lediglich das Alter (jung, mittel, alt), das Geschlecht und die Art des englischen Dialekts vor (etwa amerikanisch, indisch oder britisch). Fünf solcher Stimmvariationen sind in der Basisversion möglich, für 5 US-Dollar monatlich bereits zehn.

In diesem Tarif, Starter genannt, darf man im Voice Lab auch auf das Instant Voice Cloning zugreifen. Dafür ist eine Stimmaufnahme zwischen einer und fünf Minuten notwendig. Das Ergebnis ähnelt dem Original, die Unterschiede bleiben aber deutlich erkennbar.

Die Stimmen von ElevenLabs haben uns gut gefallen. Die neue deutsche Spracherzeugung liefert auch mit den bereits vorhandenen Sprechern erstaunlich gute Ergebnisse. An der Sprachmelodie muss der Anbieter allerdings noch feilen.

- realistische englische Stimmen
- einfaches Klonen
- monotone deutsche Stimmen

Preise: kostenlos 10.000 Zeichen pro Monat, monatlich ab 5 US-Dollar



Murf

Murf listet im oberen Teil des Bildschirms einzelne Sprecherabschnitte auf. Diese ordnet die Software ähnlich wie bei Coqui auf einer Zeitleiste am unteren Rand des Bildschirms an. Zusätzlich zu den Sprechertexten können Sie Videos in Murf hochladen, um sie mit künstlichen Stimmen zu unterlegen. Auf Wunsch blendet Murf den hochgeladenen Text als Untertitel ein. Zur zusätzlichen Untermalung können Abonnenten auf über 8000 lizenzierte Musikstücke zugreifen.

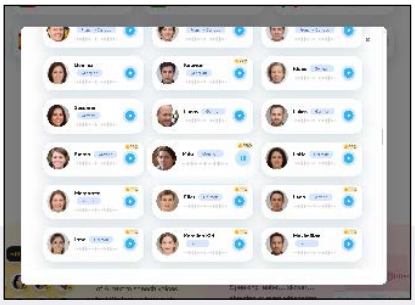
Die Stimmen von Murf gehören zu den ansprechendsten ihrer Art. Wenn die Standardbetonung nicht passt, lassen sich Aussprachefehler einzelner Wörter korrigieren. Dazu gibt man den IPA-Code (Internationales Phonetisches Alphabet) oder eine alternative Schreibweise ein. Sie können auch die Betonung ändern und Pausen einfügen. Das ist vor allem bei längeren Texten notwendig, aber zeitaufwändig und nicht immer erfolgreich.

Die Auswahl der deutschen Stimmen ist durchwachsen: Von den sieben Stimmen sprechen nur zwei halbwegs glaubwürdig, Stefan und Lena. Erstere intoniert mit deutlichem Akzent.

10 Minuten Sprachausgabe sind gratis. Wer mehr braucht, zahlt im günstigsten Tarif 29 US-Dollar pro Monat, darin sind 60 Stimmen inklusive Deutsch enthalten. Für 10 Dollar mehr erhält man Zugriff auf über 120 Stimmen in mehr als 20 Sprachen.

- viele gute Stimmen
- Videounterstützung
- kaum brauchbare deutsche Stimmen

Preise: kostenlos 10 Minuten pro Monat, monatlich ab 29 US-Dollar, Klone auf Anfrage



Revoicer

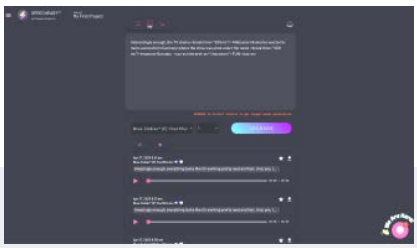
Revoicer wirbt damit, dass seine mehr als 80 Stimmen in vielen Sprachen emotionaler und damit glaubwürdiger klingen als die Stimmen der Konkurrenz. Die Gratisbeispiele schummeln jedoch: Sie kaschieren die mangelnde Stimmqualität mit dramatischer Musikuntermalung.

Das Konzept ähnelt dem der meisten anderen Kandidaten: Text in ein Feld einfügen, Sprecher auswählen und die KI rechnen lassen. Revoicer erlaubt zwar, die Sprechgeschwindigkeit zu variieren, aber die beworbene Funktion, in Textpassagen Emotionen einfügen zu können, war in unserem Testzeitraum noch nicht implementiert. Die überflüssige Musikuntermalung steht zahlenden Kunden ebenfalls zur Verfügung.

Die Qualität der Stimmen lässt sich leider erst nach Abschluss eines kostenpflichtigen Abonnements überprüfen. Bei der englischen Sprachausgabe liegt Revoicer im Mittelfeld. Schein und Sein zeigen sich bei den deutschen Stimmen. Sie stammeln maschinenhaft vor sich hin. Von Emotionen keine Spur, Betonungen werden falsch gesetzt. Weder die mangelhafte Sprachqualität noch das belanglose Gedudel der Hintergrundmusik können die hohen Abopreise rechtfertigen.

- unseriöse Werbung
- schlechte deutsche Stimmen
- keine kostenlose Probe

Preis: 27 bis 127 US-Dollar pro Monat



Speecheasy

Die vergleichsweise simpel aufgebaute Website von Speecheasy wirkt noch unfertig und enthält noch nicht einmal ein Impressum. Jedoch kann man bereits in der kostenlosen Testversion knapp hundert Stimmen ausprobieren. Neben Englisch sind auch Spanisch und Portugiesisch vorhanden. Deutsch und andere europäische Sprachen fehlen allerdings.

Besitzer eines Free-Accounts können bereits den erweiterten Texteditor mit SSML-Unterstützung nutzen und damit die Sprachausgabe manuell optimieren. Wer mehr als 20 Sekunden vertonen möchte, benötigt ein kostenpflichtiges Konto. Für 6 US-Dollar im Monat spricht Speakeasy jeweils bis zu 15 Minuten am Stück in beliebig vielen Aufnahmen.

Die Sprachqualität liegt im Mittelfeld der hier vorgestellten Dienste. Die Sprachauswahl ist zwar begrenzt und der Entwickler hält wichtige Informationen für die Nutzer, unter anderem zum Datenschutz, zurück. Als kostenloser Dienst kann das Angebot für kleinere Projekte aber durchaus nützlich sein.

- SSML-Unterstützung
- auch kostenlos gut nutzbar
- keine deutschen Stimmen

Preis: kostenlose 20-Sekunden-Clips, ab 6 US-Dollar pro Monat

Uberduck

Uberduck richtet sich mit über 5000 Stimmen an ein eher spaßorientiertes Publikum. Die Auswahl besteht aus unzähligen geklonten Schauspielern und Persönlichkeiten.

Unübersichtliche Drop-Down-Menüs machen die Auswahl der richtigen Stimme zur Qual. Eine Sortierung nach Sprache ist ebenso wenig möglich wie die Wahl zwischen einer männlichen und einer weiblichen Stimme. Die Resultate sind nicht beeinflussbar.

Uberduck kann auch Rap-Videos erzeugen. Dazu wählt man ein Playback aus dem Angebot und gibt in eine ausführliche Beschreibung an, worüber die Stimme rappen soll. Die musikalischen Ergebnisse sind indiskutabel schlecht.

Wer einfach nur ein wenig in der Sprachausgabewelt herumstöbern möchte, kann sich einen kostenlosen Account anlegen. Dann erfährt er, wie unbegabte KI-Rapper sich als MC versuchen oder es eben nicht klingt, wenn Rowan Atkinson in seinem typischen Englisch vor sich hin nuschelt.

Angesichts des unseriösen, zusammengewürfelten Angebots wirken die Bezahlmodelle überteuert. Bis zu 5 Minuten pro Monat sind kostenlos. Die kommerzielle Nutzung der verfügbaren, zumeist minderwertigen Stimmen kostet 10 US-Dollar im Monat. Für das Klonen weiterer Stimmen werden 100 Dollar pro Stimme und Monat fällig. Wer ernsthaft TTS nutzen will, sollte besser schnell das Weite suchen.

- viele Spaß-Stimmen
- unübersichtliche Auswahl
- schlechte und teure Klone

Preise: kostenlos 5 Minuten pro Monat, ab 10 US-Dollar / Klone ab 100 US-Dollar pro Monat

Zumal niemand weiß, wie verantwortungsvoll der Anbieter mit den Daten und Modellen umgeht.

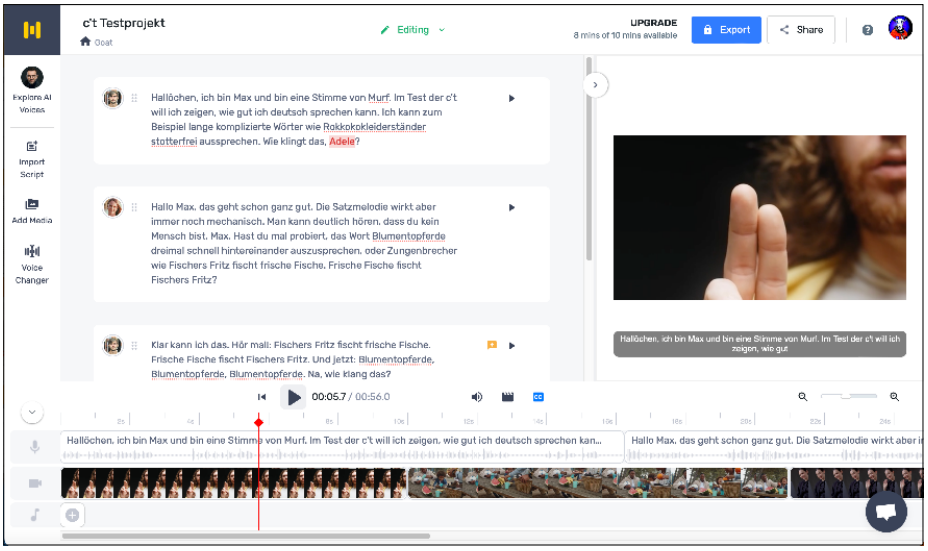
Fazit

Die Qualität der Sprachausgabe schwankt deutlich. Dies gilt insbesondere für die großen Pools von Beepbooply und Uberduck mit mehreren hundert Stimmen. Deshalb sollten Sie bei einem Dienst zunächst die kostenlosen Stimmen in Ruhe ausprobieren, ob sie für Ihren Einsatzzweck geeignet sind.

Die meisten Stimmen der KI-Dienste klingen im Vergleich zu den Stimmen von Windows und macOS realistischer. Um sich eine Webseite oder eine E-Mail vorlesen zu lassen, lohnen sie sich zwar nicht, wohl aber, um Video-Tutorials, Spiele und kurze Texte mit verteilten Sprechern vertonen zu lassen. Letzteres funktioniert mit Coqui und Murf sehr gut, wenn auch beim Erstgenannten bisher nur in Englisch. Coqui ist zudem der einzige Dienst ohne monatliche Abogebühren.

Wer eine deutsche Sprachausgabe benötigt, ist mit Beepbooply und Murf am besten bedient. Bei ElevenLabs kann die deutsche Sprachausgabe nicht mit den hervorragenden englischen Sprechern mithalten. Die deutschen Stimmen von Revoicer sind unbrauchbar und stottern mechanisch vor sich hin.

Wer Stimmen klonen möchte, kann dies relativ günstig bei Coqui und ElevenLabs ausprobieren. Auf Englisch funktioniert das am besten. Unsere Ergebnisse ähnelten zwar den Originalsprechern, kopierten sie aber nicht täuschend echt.



Im Editor von Murf passen Sie Dialoge auf einer Timeline an und unterlegen sie mit Musik und Videos – entweder eigene oder aus der Stock-Datenbank von Murf.

Dafür wäre ein aufwändigeres und teureres Training nötig.

Klar sollte sein, dass KI-Stimmen relativ unpersönlich sprechen, zumal man den Sprachduktus nur marginal und dann recht umständlich beeinflussen kann. Die schnellste und einfachste Methode sind noch zusätzliche Satzzeichen. Speecheasy erlaubt per SSML zwar feinere Eingriffe, die sind jedoch extrem aufwändig und verwandeln den Ausgangstext schnell in eine Codewüste.

Professionelle Sprecher lassen sich durch eine KI derzeit also noch nicht komplett ersetzen. Das gilt vor allem für längere Podcasts und Hörbücher sowie für Live-Sendungen, bei denen ein Sprecher spontan auf sein Gegenüber reagiert und emotionale Nuancen in seine Sprache einbaut. Für kürzere Clips, maschinelle Übersetzungen oder Avatare in Videospielen sind die KI-Stimmen schon heute günstige Alternativen, die menschliche Sprecher in diesen Bereichen bald verdrängen könnten. *(hag@ct.de)* c't

Literatur

[1] Kai Schwirzke, KI rappt wie Eminem, Wie man mit Chatbots und Sprachtools einen Rap-Song produziert, c't 7/2023, S. 148
[2] Dorothee Wiegand, Klon am Mikrofon, Was synthetische Stimmen leisten, c't 24/2022, S. 130

Online-Dienste für Text-to-Speech

Dienst	Beepbooply	Coqui	ElevenLabs	Murf	Revoicer	Speecheasy	Uberduck
URL	beepbooply.com	coqui.ai	elevenlabs.com	murf.ai	revoicer.com	speecheasyapp.com	uberduck.ai
Sprachen	81	1 (Englisch)	6	über 20	40	7	keine Angaben
Anzahl der Stimmen	über 900	33	9	über 120	80	92	über 5000
deutsche Stimmen	✓	–	✓	✓	✓	–	–
Stimmklone	–	✓	✓	Englisch auf Anfrage	–	–	✓
manuelle Bearbeitung	Tonhöhe, Geschwindigkeit, Pause	Dialog-Editor, Advanced Editor mit Phonem-Level-Anpassung	Stability, Clarity	Emphasis, Aussprache (IPA, alternative Schreibung)	Geschwindigkeit, Tonhöhe, Pause	SSML	–
Wertung und Preise							
Qualität deutsch / englisch	⊕ / ⊕	– / ⊕⊕	○/ ⊕⊕	⊕ / ⊕⊕	⊖ / ○	– / ○	– / ⊖⊖
Sprecherauswahl	⊕⊕	⊕	⊕	⊕⊕	⊕	○	⊖
Steuerung	⊕	⊕⊕	⊕	⊕⊕	○	○	⊖
kostenlose Version / Download erlaubt	10.000 Zeichen / –	30 Minuten / ✓	10.000 Zeichen / ✓	10 Minuten / –	– / –	je 20-Sekunden-Clips / ✓	5 Minuten / ✓
Abopreise pro Monat	ab 7 US-$ für 100.000 Zeichen	ab 20 € für 4 Stunden (kein Abo)	ab 5 US-$ für 30.000 Zeichen	ab 29 $ für 2 Stunden Audio	ab 27 US-$ für 60.000 Zeichen	ab 6 US-$ für 15-Minuten-Clips	ab 10 US-$ für 1 Stunde

✓ vorhanden – nicht vorhanden ⊕⊕ sehr gut ⊕ gut ○ befriedigend ⊖ schlecht ⊖⊖ sehr schlecht