Hier spricht der Klon

Zukunft Die perfekte Fälschung: Künstliche Intelligenz macht es möglich, die Stimme jedes beliebigen Menschen im Computer zu erschaffen – und damit böse Spiele zu treiben.

Wer nimmt als Erster einen neuen Song mit Michael Jackson auf? Wann lässt sich Elvis mit frischem Liedgut hören?

Bald sei es so weit, glaubt Jordi Janer, Klangforscher in Barcelona. Seine Firma Voctro Labs arbeitet bereits an der nötigen Technik. Sie erlaubt es, beliebige Stimmen im Computer zu klonen. Ein paar Stunden Tonaufnahmen vom Original genügen – und wenig später erwachen auf ewig Verstummte zu künstlichem Leben.

Schon Anfang vorigen Jahres verblüffte die Firma mit einem singenden Donald Trump. Ein kleines YouTube-Video zeigt, wie der damalige Präsidentschaftsbewerber eine Jazzballade vorträgt, schwungvoll intoniert und mit Mut zum Vibrato. Alles nur getrickst, aber recht überzeugend.

Heute ist die Technik einen großen Schritt weiter. Die neuesten Kunststimmen klingen so gut wie echt. Und sie singen, was auch immer man ihnen in den Mund legt: jeden Text, jede Melodie.

Darauf beruht eine der Geschäftsideen von Voctro Labs. „Die Werbewirtschaft“, sagt Janer, „könnte bekannte Songs auf ihre Produkte umschreiben.“ Nicht auszuschließen also, dass schon bald die Beatles „All You Need Is Coke“ anstimmen.

Janer hat seine kleine Firma zusammen mit drei Kollegen von der Universität gegründet. Sie gehören dort der Music Technology Group an, die für ihre Erfinderfreude bekannt ist. Eine Forscherin tüftelt zum Beispiel bereits an künstlichen Chören, bestehend aus je zwölf virtuellen Sängern – ein jeder braucht seine eigene Stimme, damit das Ensemble am Ende lebendig klingt. Eines Tages soll es Chorwerke selbstständig vom Blatt einspielen. Oder menschlichen Sängern zum häuslichen Üben dienen.

Die Firma Voctro Labs will ihre Technik auch für Gesprochenes einsetzen. Eine erste Anwendung ist das automatische Synchronisieren von Filmen. Die geklonte Stimme von Brad Pitt etwa könnte die Rollen des Schauspielers auf Deutsch oder auch Kurdisch nachsprechen, und es klänge in jeder Sprache nach dem echten Pitt.

Start-ups in aller Welt entwickeln derzeit solche künstlichen Sprecher. Besonders forsch geht die kanadische Firma Lyrebird zu Werk, Studenten der Universität von Montréal haben sie gegründet. Lyrebird kann angeblich schon aus einer einminütigen Stimmprobe, notfalls von YouTube heruntergeladen, einen brauchbaren Klon erzeugen. Auf der Homepage der Firma ist zu hören, wie Barack Obama, Hillary Clinton und Donald Trump („This is huge“) über ihre Software fachsimpeln. Der gefälschte Disput ist noch nicht perfekt, aber die Technik kommt allenthalben in großen Schritten voran.

Bislang war es langwierig und teuer, eine Computerstimme zu erschaffen. Ein Sprecher musste für viele Stunden ins Studio und dort Tausende Sätze aufnehmen – langsam und schnell, laut und leise, in verschiedener Betonung, als Frage, als Aufforderung. Diese Aufnahmen wurden in einzelne Lautabschnitte zerhackt. Daraus stückelte der Computer dann, je nach Text, neue Sätze zusammen (siehe Grafik).

Es bedurfte einiger Tricks, bis eine halbwegs erträgliche Satzmelodie herauskam. Trotzdem klang die Maschine oft blechern und geknödelt, kurz: wie ein Roboter.

Heute dagegen hören sich die besten Kunststimmen auf wundersame Weise lebendig an. Dahinter steckt künstliche Intelligenz.

Der Computer verkettet keine vorgefertigten Wortschnipsel mehr – er spricht einfach selbst. Wie das geht, lernt er anhand von Aufnahmen verschiedener Sprecher. Sogenannte neuronale Netze trainieren damit so lange, bis sie imstande sind, menschliche Laute zu erzeugen. Nach einiger Zeit gelingt das recht fließend, mit 16 000 Signalen und mehr pro Sekunde. Das genügt, um hie und da auch schon ein feines Gehör zu täuschen.

Allerdings hat so eine künstliche Stimme keinerlei Sprachverstand. Schaltet man sie ein, so brabbelt sie in unverständlichen Lauten vor sich hin. Aber sobald der Computer einen Text bekommt, kann er ihn manierlich sprechen.

Der letzte Durchbruch gelang im vergangenen September. Damals stellte die Google-Tochter DeepMind ein grundlegend neues Verfahren vor. Wenig später, im November, folgte der Softwarekonzern Adobe mit einer ähnlichen Technik namens Voco.

Zunächst hieß es noch, der Rechenaufwand für solche synthetischen Stimmen sei immens, er übersteige die Kräfte privater Computer bei Weitem. Inzwischen aber konnten pfiffige Unternehmen zeigen, dass es auch schnell und billig geht. Die Qualität leidet etwas, aber auch die Stimmen zweiter Wahl können mitunter schon beeindrucken.

Das Pariser Start-up CandyVoice zum Beispiel arbeitet an einer App für Smartphones. Der Kunde nimmt damit nur 160 Mustersätze auf, und auf den Servern von CandyVoice entsteht daraus ein Modell seiner Stimme.

Das geht, weil der Computer nicht jedes Mal von Grund auf neu zu sprechen lernt. Ist er einmal trainiert, verfügt er quasi über eine Allzweckstimme. Dann müssen nur noch die Eigenheiten der gewünschten Sprecher hinzugemischt werden.

Wer will, kann auf der Homepage von CandyVoice ein paar Worte ins Mikrofon sprechen – wenig später ist die eigene Stimme in vielerlei Verwandlung zu hören: als Frau, als Kind, als Greis. Wie am Mischpult lassen sich auf diese Weise auch neue Stimmen kombinieren.

Theoretisch könnte eine mittelbegabte Sängerin ihren Stimmklon eines Tages mit dem Schmelz eines Opernsoprans veredeln lassen. Auch Tempo und Gefühlsausdruck sind im Prinzip steuerbar. Darauf ist die kleine Firma Vivotext in Tel Aviv spezialisiert, die der Konzertpianist Gershon Silbert gegründet hat. Vivotext will nächstes Jahr eine App auf den Markt bringen, die mit Reglern für Fröhlichkeit oder Melancholie aufwartet.

Die Firma hat ihre Software an den Spielzeugkonzern Hasbro lizenziert. Gut möglich also, dass eines Tages interaktive Plüschtiere in den Kinderzimmern auftauchen, die mit den Stimmen der viel beschäftigten Eltern – oder des neuesten Teeniestars – Gutenachtgeschichten vorlesen.

Die gleiche Technik wäre für Kranke hilfreich, die ihre Sprechfähigkeit zu verlieren drohen. Sie könnten beizeiten mit künstlichem Ersatz vorsorgen.

Jux und Nutzwert sind hier nicht weit auseinander. Absehbar ist eine Hochkonjunktur der Telefonstreiche („Hier spricht Franz Beckenbauer“). Es drohen aber auch weniger witzige Attentate auf die Gutgläubigkeit. Was, wenn einmal im Internet ein Redemitschnitt von Angela Merkel auftaucht, in dem sie sich über den Propheten Mohammed lustig macht?

Die Täuschung wäre komplett, wenn es dazu noch ein Video gäbe: der gefälschte Text, vorgetragen von der Bundeskanzlerin mit sprachsynchronen Lippenbewegungen.

Video: So klingen geklonte Präsidentenstimmen

spiegel.de/sp222017stimme
oder in der App DER SPIEGEL

Erste Versuche in dieser Richtung gab es schon. Der Doktorand Justus Thies an der Uni Erlangen-Nürnberg zeigte an Videos von Donald Trump und George W. Bush, dass er die Kontrolle über deren Mimik übernehmen kann. Er schnitt Grimassen in eine Kamera, bewegte den Mund, und wie ferngesteuert zeigte sich das gleiche Mienenspiel auf den Gesichtern der Zielpersonen – Thies' Software übertrug alles in Echtzeit.

Bis heute galt: Einer Stimme glaubt man, dass sie echt ist. Sie kommt aus der Mitte des Körpers, sie ist Teil der Persönlichkeit – eigen und unverwechselbar. Nur ein paar Spezialtalente konnten anderer Leute Sprechweise nachmachen. Nun kann es bald jeder.

Die Stimme als Ausweis des Authentischen ist damit wohl erledigt. Dumm nur, dass sie mancherorts bereits als Passwortersatz erprobt wird. Die Firma Nuance bietet eine Stimmerkennung, mit der ein Callcenter Anrufer identifizieren kann. Die britische Großbank HSBC setzt ein ähnliches System beim Onlinebanking ein.

Die Hersteller versichern, mit forensischer Software lasse sich nach wie vor herausfinden, ob ein Mensch spricht oder ein neuronales Netz, das ihn nachäfft. Es gebe immer Spuren, die auf künstliche Herkunft hindeuteten.

Allerdings weiß auch die Gegenseite sich zu helfen. „Wir könnten ein zweites neuronales Netz darauf trainieren, ebensolche Spuren zu verwischen", sagt der kanadische KI-Pionier Yoshua Bengio. Sein Labor an der Uni von Montréal hat nicht nur die Grundlage für die Sprachsynthese des Start-ups Lyrebird geschaffen; hier entstand auch die Idee für eine neue Stufe maschinellen Lernens – nach Art eines Duells. „Wir sprechen von feindlichen Netzwerken", sagt Bengio. Das eine lauert auf verräterische Schwachstellen, das andere lernt, sie immer besser zu verbergen. So treiben die beiden Kontrahenten sich wechselseitig voran.

Bengio glaubt, dass weitere Fortschritte zu erwarten sind. Ein Computer, der Sprachlaute erzeugt, werde das irgendwann auch mit Bewegtbildern hinbekommen. Statt der Sprachaufnahmen gibt man den lernenden Algorithmen dann eben Videos zum Trainieren – bis sie ähnliche Aufnahmen selbst hervorbringen, im Extremfall täuschend echt. „Unseren Glauben an die Wahrhaftigkeit von Bildern und Tönen", sagt Bengio, „dürfte das ganz schön verstören."

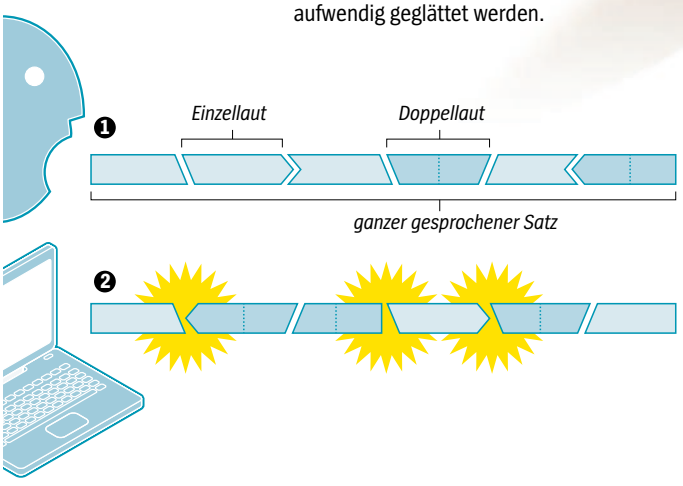
Manfred Dworschak
Mail: manfred.dworschak@spiegel.de

In den Mund gelegt

Funktionsweisen von **bisheriger Sprachsoftware . . .**

❶ Der Computer zerlegt Aufnahmen eines Sprechers mit Tausenden Sätzen typischerweise in Einzel- oder Doppellaute.

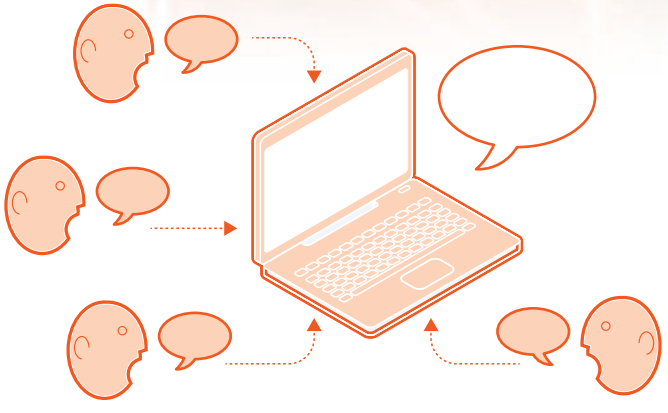
❷ Um eigene Sätze zu erzeugen, fügt die Software die Lautschnipsel nach Bedarf neu zusammen. An vielen Stellen passt dadurch die Sprachmelodie nicht mehr, es klingt eckig. Die Holperstellen müssen aufwendig geglättet werden.



. . . und **Programmen mit künstlicher Intelligenz**

❸ Programme mit künstlicher Intelligenz lernen anhand von Aufnahmen verschiedener Sprecher, selbst täuschend echte Sprachlaute zu erzeugen.

❹ Der Computer kann die Stimme einzelner Sprecher imitieren, indem er auf deren Eigenheiten trainiert wird. Mit solchen künstlichen Stimmen lassen sich auch Äußerungen fälschen, die so nie gefallen sind.



ALEXANDER HASSENSTEIN / BONGARTS / GETTY IMAGES

DER SPIEGEL