“Gesprochene Wikipedia” und künstliche Stimmerzeugung *(Barrierearmut)*

Digitaler Themen Stammtisch

Präsentation: Thorsten Müller (*MrThorstenM*)
https://de.wikipedia.org/wiki/Thorsten_(Stimme)

Datum: 14. September 2021

Über mich & Motivation

- Mein Name ist Thorsten Müller
  https://twitter.com/ThorstenVoice / https://github.com/thorstenMueller/deep-learning-german-tts/

- Beruflich und Privat bin ich leidenschaftlich als Informatiker aktiv

- Privat gilt mein Interesse offenen Sprachsysthemen
  (*insbesondere Sprachausgabe*) und natürlich Wikipedia :-)

- Ich bin Sprecher und Stimmspender des freien deutschen Sprachdatensatzes “Thorsten”
  (*Keine Angst, was eine “Stimmspende” ist erkläre ich noch*)
  https://de.wikipedia.org/wiki/Thorsten_(Stimme)

“Meine Motivation ist die Bereitstellung einer kostenfreien, qualitativ hochwertigen, deutschen künstlichen Stimme zur Sprachsynthese, die offline erzeugt werden und jeder Zielgruppe kostenfrei und ohne lizenzrechtliche Einschränkungen zur Verfügung stehen soll.”

Was ist eine künstliche Stimme (*ohne Technikdetails*)

- Auch bekannt als Sprachsynthese oder Text-to-Speech/TTS
- Die Idee einer “sprechenden Maschine” gab es schon lange vor Computern
  (siehe Kempelens Sprechmaschine)
- In frühen Computerspielen konnten Figuren schon sprechende Piepslaute machen
  (es war viel Phantasie nötig, um darin Ansätze menschlicher Sprache zu hören)
- Sprechende Maschinen faszinierten auch in Filmen und Serien wie
  Star Trek, 2001 Odyssee im Weltraum, Knight Rider, … .
- Künstliche Sprachausgabe kann auch im Internet Browser eingebunden werden (*Barrierearmut*)
- Der Durchbruch kam durch Sprachassistenten
- Die Qualität der Stimme ist wesentlich für die Akzeptanz und die großen Technologieanbieter haben die Messlatte hoch gesetzt

Statisches TTS

- Aufgezeichneter Text, der situationsabhängig neu zusammengestellt wird (*vergleichbar einer Musik Playlist*).

  Bspw. Navi: In <300/200/100> Metern im Kreisverkehr an der <ersten/zweiten> Ausfahrt abfahren und dann <links/rechts> halten.

- Vorteile
  - Geringer Aufwand bei der Spracherzeugung und Aufzeichnung
  - Einzelne Wörter klingen natürlich, da sie von einem Mensch aufgenommen wurden

- Nachteile
  - Nur aufgenommen Texte können situationsabhängig zusammengestellt werden (*kann beispielsweise keine Orts- oder Straßennamen aussprechen*)
  - Die Betonung klingt durch zusammenfügen abgehackt und unnatürlich

Dynamisches TTS (Sprach “Dataset”)

- Auf Basis eines Sprach “Dataset” wird mit maschinellem Lernen (*technologische Grundlage von KI*) ein “Modell” einer Stimme trainiert

- Ein “Dataset” besteht aus Audio Aufnahmen (*einzelnen Sätzen*) und einer Textdatei mit dem gesprochenen Inhalt

- Dazu sind tausende Sätze eines einzelnen Sprechers notwendig (*hohe phonetische Abdeckung*)
  - Qualitativ hochwertige Aufnahmen ohne Rauschen und Kratzen
  - Konstante Sprechgeschwindigkeit
  - Saubere aber natürliche Betonung
  - Bitte “hochwertiges” Mikrofon und Raumsituation verwenden!
  - *Einige Kriterien/Empfehlungen mehr …*

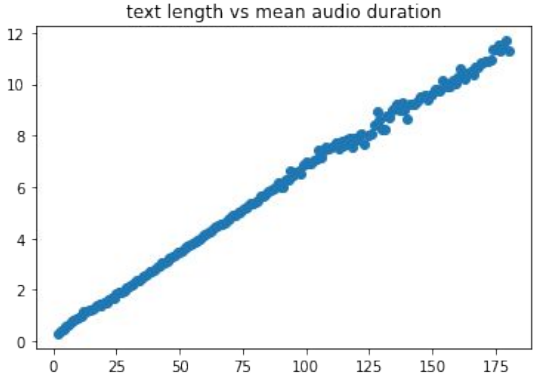
- Die Aufnahmen müssen lizenzrechtlich für das Training einer künstlichen Stimme zulässig sein
  Bitte nicht einfach die Tonspur des Lieblings Moderators, Podcasters oder Youtubers dafür verwenden

Dynamisches TTS (Dataset “Thorsten”)

Ich habe das Thema am Anfang massiv unterschätzt!

Beide Datasets wurden von mir aufgenommen und stehen unter der CC0 Lizenz zur Verfügung.
(https://github.com/thorstenMueller/deep-learning-german-tts/)

- **Neutrales Dataset**
  - Oktober 2019 - Juni 2020
  - **22.668 Aufnahmen** (entspricht über **23 Stunden reinem Audio**)
  - Mono und Samplerate von 22kHz
  - Konstante Sprechgeschwindigkeit von 14 Zeichen pro Sekunde
  - Textlänge: 2 - 180 Zeichen (Durchschnitt 52 Zeichen)
  - Die ersten sechs Monate (*leider*) mit schlechtem Headset



- **Emotionales Dataset**
  - Identische 300 Sätze in folgenden Emotionen: Neutral, Wütend, Angewidert, Fröhlich, Müde, Überrascht, Flüsternd, Betrunken

Das ist meine **Stimmspende** an die “Menschheit”
Details und Download unter https://github.com/thorstenMueller/deep-learning-german-tts/

Tips & Tricks zum Aufnehmen eines Datasets

- Gutes Mikrofon und Aufnahmesituation verwenden
- Text mit hoher phonetischer Abdeckung
- Zahlen und Abkürzungen bereinigen
- Ruhige und konstante Aufnahmesituation
- Immer gleichen Abstand zwischen Mund und Mikrofon
- Ventilatoren, Lüfter, etc. möglichst abschalten
- Neutral aber natürlich sprechen
- Nicht nuscheln oder Buchstaben verschlucken
- Stimme anpassen bei Satzzeichen

Tips & Tricks zum Aufnehmen eines Datasets

- Gleichmäßige Sprechgeschwindigkeit
- Keine Pause am Anfang und Ende der Aufnahme
- Die Aufnahmen kontrollieren auf die Qualität (*auf maximaler Lautstärke*)
- Maximal 30 Minuten am Stück lesen und nicht länger als 4 Stunden pro Tag
- Bei Erkältung keine Aufnahmen machen
- Fehlerfrei lesen, genau wie im Text geschrieben

Dynamisches TTS

- Nachteile
  - Benötigt sehr viele Aufnahmen eines Sprechers in sehr guter Qualität
  - Training ist aufwendig *(Dauer abhängig von verfügbarer Infrastruktur)*

- Vorteile
  - Meist natürlicher Sprachfluss
  - Kann auch Text synthetisieren, die vorher nicht aufgenommen wurden

Kleines Beispiel:

https://drive.google.com/file/d/17nArItdRYnSIU34wI2kUwr_DT7eoRdex/view?usp=sharing

Derzeitige Grenzen von TTS

- Moderne künstliche Spracherzeugung kommt menschlicher Stimme sehr nah und ist “fast” nicht zu unterscheiden. Menschliche Stimme ist aber noch vorne.

- TTS kann (*derzeit*) noch nicht gemäß Inhalt unterschiedlich emotional betonen, sondern klingt immer neutral.

- Texte müssen durch TTS oder im Vorfeld “vorbereitet” werden.
  - Bspw.: Nummern in Text umwandeln (1 = Eins), Datums- und Jahresformate konvertieren 15.09.2021, Abkürzungen ausschreiben (a.D.)

- Kann Schwierigkeiten mit der Betonung von Fremdwörtern haben.

Kann TTS die “gesprochene Wikipedia” unterstützen?

Ich hoffe und glaube **ja**:

- Artikel können deutlich effizienter maschinell “gesprochen” werden *(auch bei Artikeländerungen)*.
- Aktuell sind ca. 1.262 Artikel (***Respekt :-)***) von insgesamt 2.612.309 menschlich vertont. Dies entspricht ungefähr 0,05%.
- Selbst mit der “naiven” Annahme dass das Aufnehmen, Bearbeiten und Hochladen eines kompletten Artikels nur eine Stunde dauern würde, würde das folgendes bedeuten:
  - Noch 2.611.047 Artikel zu vertonen (= 2.611.047 Stunden)
  - Einer Person würde 24 Stunden am Tag lesen wären 108.793 Tage notwendig
  - Das bedeutet bei 24 Stunden, 7 Tage die Woche einsprechen wäre **eine Person 298 Jahre beschäftigt**!

Alle Wikipedia Artikel manuell einzusprechen ist ein “Faß ohne Boden”
(vergleichbar mit dem Versuch Youtube “durchzuschauen”).

Wie geht es weiter

- Ich nehme ein neues neutrales Dataset mit besserer Qualität und natürlicherem Sprachfluss auf.

- Ich trainiere “effizientere” und klanglich bessere Modelle.

- **Ich würde mich freuen mit meiner freien Stimme Wikipedia im Bereich der Barrierearmut unterstützen zu können.**

Falls jemand auch seine Stimme spenden möchte

https://commonvoice.mozilla.org/de freut sich immer über Unterstützung :-)

„Für mich sind alle Menschen gleich, unabhängig von Geschlecht, sexueller Orientierung, Religion, Hautfarbe oder Geokoordinaten der Geburt. Ich glaube an eine globale Welt, wo jeder überall willkommen ist und freies Wissen und Bildung kostenfrei für jeden zur Verfügung steht. Ich habe meine Stimme der Allgemeinheit gespendet, in der Hoffnung darauf, dass sie in diesem Sinne genutzt wird.“

(Thorsten Müller)