Visuelle Wahrnehmung als Unterstützung zum Hörsinn: KI zum Lippenlesen auf Deutsch

Dinh Nam Pham[1,2] & Torsten Rahne[2]

[1] Technische Universität Berlin, Fakultät IV Elektrotechnik und Informatik., Berlin
[2] Universitätsmedizin Halle (Saale), Klinik und Poliklinik für Hals-Nasen-Ohrenheilkunde, Kopf- und Hals-Chirurgie, Halle (Saale)

dinh-nam.pham@campus.tu-berlin.de , torsten.rahne@uk-halle.de

Hintergrund

Sowohl Menschen mit Hörverlust als auch Normalhörende profitieren von den zusätzlichen visuellen Informationen aus den Lippenbewegungen des Sprechenden. Dieses sogenannte Lippenlesen ist jedoch fehleranfällig. So sind in der deutschen Sprache nur ungefähr 15% der Laute am Mundbild erkennbar. Bei Einbettung in den Kontext erhöht sich die Trefferquote einer Schätzung zufolge auf 50%. In den letzten Jahren wurde z.B. für die englische Sprache Künstliche Intelligenz (KI) mit Ansätzen von LipNet, einem Deep-Learning-Modell für das Lippenlesen auf Satzebene durch Vorhersagen von Zeichensequenzen, sowie von DeepMind, welches im Gegensatz zu LipNet, Phoneme statt Zeichen vorhersagt, verwendet. Die entwickelten Algorithmen zum Lippenlesen mit auf künstlichen neuronalen Netzwerken basierender KI verbessern die Wortkennung signifikant, stehen jedoch für die deutsche Sprache nicht zur Verfügung [1].

Abb. 1: Ausgewählte Bilder der Mundbewegung zum Wort „aber“.

Methoden

Datensatz

- Selektion von 1.806 Youtube-Videos (1280x720 Pixel, 25 FPS) mit jeweils nur einer deutsch sprechenden Person
- Unterteilung in Wortsegmente und Klassifikation mit VoskAPI (18 Klassen)

→ 38.391 Videos
→ 3 Teildatensätze mit unterschiedlichen Sprechenden

Datensatz		Original-Videos	Sprecher	Videos im Datensatz
A	Gesamt	250	6	3727
	Training			2973
	Validierung			754
B	Gesamt	1400	22	30684
	Training			24553
	Validierung			3060
C	Test	156	4	3950
	Gesamt			3950

Videobearbeitung

- Einheitliche Länge: 28 Frames
- Konvertierung in Farbräume RGB, Graustufen, HSV, LAB, XYZ und YcbCr
- Downsampling mit linearer Interpolation (Mund: 150x100 Pixel; Gesicht: 90x90 Pixel)

Modelle und Training

- Neuronale Netzwerke: 3D Convolutional Neural Network (Conv3D), Gated Recurrent Units (GRU) und GRUConv (Abb. 1), Details siehe [1].

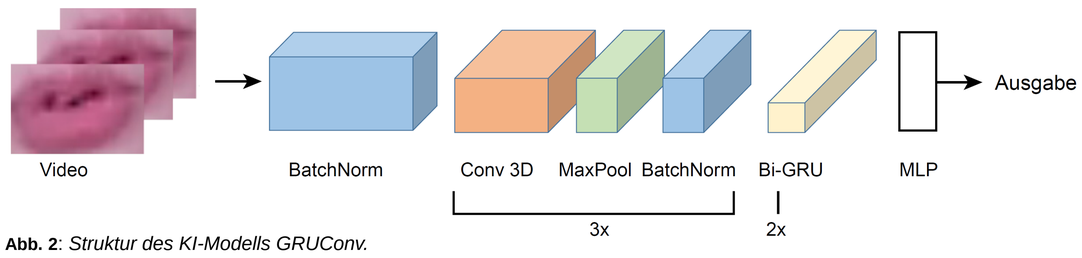

Abb. 2: *Struktur des KI-Modells GRUConv.*

Implementierung

- PyTorch Deep-Learning-Bibliothek
- 4 Nvidia Tesla-V100-SXM2-Grafikprozessoren mit 32 GB
- Modell-Parallelismus

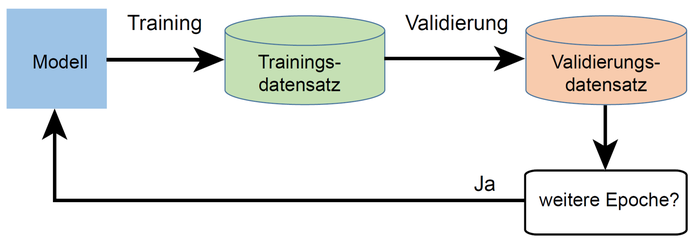

Abb. 3: *Schematische Darstellung der Trainingsstrategie des neuronalen Netzwerks.*

Ablauf

- Training der KI und Validierung mit getrennten Datensätzen
- Vergleich der KI-Modelle (Conv3D, GRU, GRUConv); Datensatz A
- Vergleich der Bildausschnitte (Gesicht, Mund); Datensatz A
- Vergleich der Farbräume (RGB, Graustufen, HSV, LAB, XYZ, YcbCr); Datensatz A

→ Bestimmung der Korrektklassifikationsrate jeweils innerhalb von 5000 Trainingsepochen

- Bestimmung des optimalen Modells, Training und Validierung mit vielen Instanzen (Datensatz B) und Testen mit unbekannten Sprechern (Datensatz C)

Ergebnisse

Vergleich der Bildausschnitte (mit Conv3D-Modell und RGB-Farbraum)

- Deutlich höhere Korrektklassifikationsrate bei Zuschnitt auf Mund (70%) als bei Zuschnitt auf das gesamte Sprechergesicht (34%)

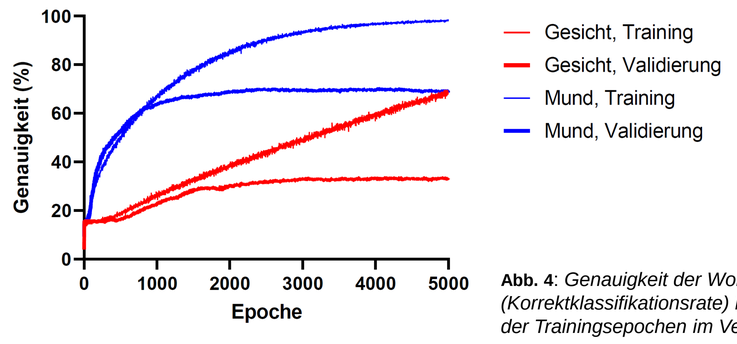

Abb. 4: *Genauigkeit der Worterkennung (Korrektklassifikationsrate) in Abhängigkeit von der Anzahl der Trainingsepochen im Vergleich der Bildausschnitte.*

Vergleich der Farbräume (Conv3D-Modell und Mundausschnitt)

- Korrektklassifikationsraten im Bereich von 69% (HSV) bis 73% (LAB)
- Nicht relevant verschieden

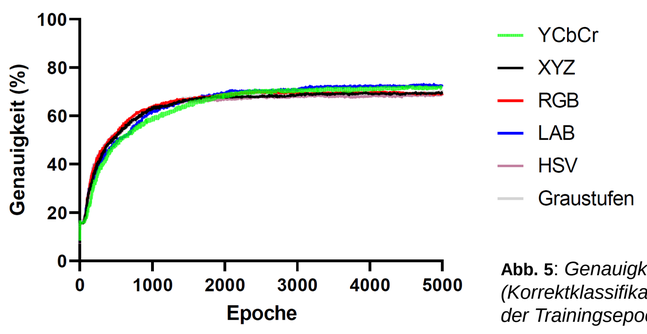

Abb. 5: *Genauigkeit der Worterkennung (Korrektklassifikationsrate) in Abhängigkeit von der Anzahl der Trainingsepochen im Vergleich der Farbräume.*

Vergleich der Modelle (Mundausschnitt und LAB-Farbraum)

- Conv3D: 73% maximale Validierungsgenauigkeit (4684 Epochen)
- GRU: 60% maximale Validierungsgenauigkeit (2466 Epochen)
- GRUConv: 78% maximale Validierungsgenauigkeit (4297 Epochen)

Training und Validierung (GRUConv, Mundausschnitt, LAB-Farbraum)

- 87% maximale Korrektklassifikationsrate bei bekannten Sprechenden
- 63% maximale Korrektklassifikationsrate bei unbekannten Sprechenden

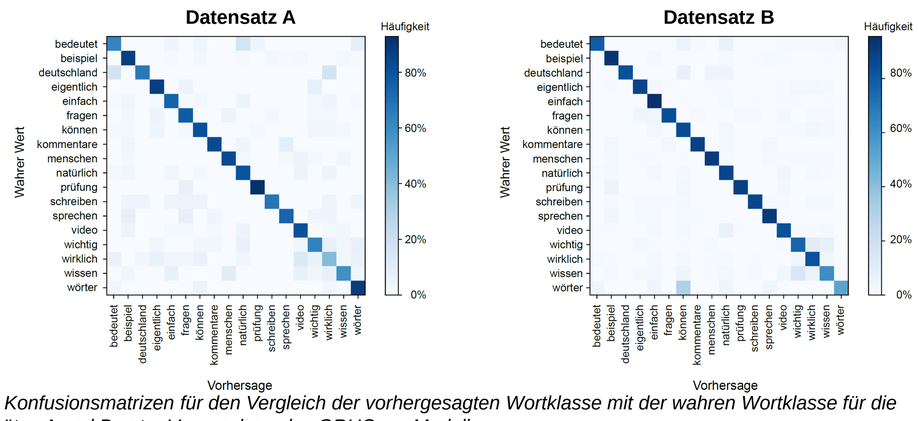

Abb. 6: *Konfusionsmatrizen für den Vergleich der vorhergesagten Wortklasse mit der wahren Wortklasse für die Datensätze A und B unter Verwendung des GRUConv-Modells.*

Literatur

[1] Pham & Rahne (2022) Entwicklung und Evaluation eines Deep-Learning-Algorithmus für die Worterkennung aus Lippenbewegungen für die deutsche Sprache. HNO 70, 456-465

Diskussion und Schlussfolgerung

Das erstmals für die deutsche Sprache entwickelte neuronale Netzwerk zum Lippenlesen zeigt eine sehr große, mit englischsprachigen Algorithmen vergleichbare Genauigkeit. Es funktioniert auch mit unbekannten Sprechenden und kann mit mehr Wortklassen generalisiert werden.