VERGLEICH VERSCHIEDENER MACHINE-LEARNING ANSÄTZE ZUR KONTINUIERLICHEN SCHÄTZUNG VON PERZEPTIVEM SPRECHTEMPO

Benjamin Weiss[1], *Thilo Michael*[2], *Uwe Reichel*[1], *Oliver Pauly*[1]

[1]*audEERING GmbH,* [2]*Technische Universität Berlin*
bweiss@audeering.com

Kurzfassung: Perzeptives, lokales Sprechtempo im Deutschen ist nach Pfitzinger (1999) ein geglättetes kontinuierliches Signal, das aus einer Kombination von Silben- und Lautrate generiert wird und für die menschliche Wahrnehmung von Sprechtempo im Deutschen validiert ist. Ein bestehender akustischer Schätzer auf Basis von LSTMs mit MFCCs als Eingangsparametern wurde auf der ESSV 2018 vorgestellt. Dieser Ansatz wird nun mit moderneren Ansätzen verglichen. Zum einen werden komplexere neuronale Netzarchitekturen verwendet, die CNN und LSTM kombinieren. Zum anderen werden unterschiedliche Eingangsdaten getestet, indem zusätzlich zu und anstatt MFCCs die Einhüllende des Zeitsignals verwendet wird. Für die abschließende Evaluierung wird mit PhonDat 1 ein zusätzlicher Datensatz mit unterschiedlichem linguistischem Material herangezogen. Einfache rekurrente Netze sind hierbei CNNs etwas überlegen. Eine einfache Kombination von CNN und LSTM führt nicht zu einer Verbesserung. Zudem zeigt sich, dass komplexere CNN Architekturen MFCCs als Merkmale überflüssig machen können.

1 Einleitung

Sprechtempo gilt neben der Tonhöhe und Lautstärke zu den drei grundlegenden Suprasegmentalia [4]. Dabei wird üblicherweise zwischen der sogenannten Sprechrate als Silben pro Zeitdauer und Artikulationsrate als Silben pro Artikulationsdauer, also ohne Sprechpausen, unterschieden [14].

In [16] wurde ein einfaches akustisches Vorhersagemodell für kontinuierliche, perzeptive Sprechrate vorgestellt. Neben einer verbesserten Vorhersage gegenüber eines durch Experten erstellten, regelbasierten Ansatzes der Silbenkerndetektion, soll ein solches Modell erlauben, kontinuierlich Schätzwerte zum Sprechtempo, und auch deren Streumaße für kurze Äußerungen, auszugeben. Idealerweise funktioniert ein solcher Schätzer in Echtzeit und bedarf keiner linguistischen Komponente zur Silbenerkennung, benötigt also keine Sprach- oder Phonemerkennung. Zu den Anwendungsgebieten zählt die paralinguistische Analyse von Sprecherzuständen und -eigenschaften [1][12][15], aber auch die Gesprächsanalyse, wie etwa die Vorhersage von Äußerungsenden und die damit verbundene Erweiterung von Sprachdialogsystemen im Bereich Sprecherwechsel und akustisch-prosodischen Entrainment [2][5][8][9].

Für germanische Sprachen sind verschiedene automatische Verfahren etabliert. Neben der Spracherkennung und -segmentierung gibt es auch rein akustische Ansätze zur Silben(kern)detektion, von denen ein Praat-Skript [3], im Jahr 2010 modifiziert [11], das bekannteste und verbreitetste sein dürfte. Hierbei werden mit einfachen Regeln über den Intensitätsverlauf und Stimmhaftigkeit Silbenkerne und Pausen geschätzt. In beiden Ansätzen wird über den Kehrwert der Silbendauern oder Silbenkerndauern, Sprechgeschwindigkeit bzw. Artikulationsrate (unter Ausschluss von Sprechpausen) – gemittelt über eine Aufnahme – ermittelt. Dabei zeigt das Praat-Skript eine sehr gute Erkennung von Silbenkernen, obwohl erfahrungsgemäß unbetonte Silben und Sprechpausen übersehen werden können.

Beide Ansätze erlauben die Vorhersage von Sprechrate jedoch nur im Anschluss der Silbendetektion und sind damit prinzipiell zwei-schrittig und langsam, auch bei Fensterung von längeren Äußerungen in kurze Abschnitte von nur wenigen Silben. Im Gegensatz zur Silbenkern- oder Silbensegmentdetektion verspricht die Verwendung einer kontinuierlichen Darstellung von Sprechrate als Zielgröße eine schnellere Vorhersage. Dies konnte auf Basis von MFCC-basierten Eingangsmerkmalen mit einem einfachen rekurrenten neuralen Netz (RNN) signifikant und

mit deutlich kürzeren Signalfenstern, nämlich 300ms, als Modell umgesetzt werden [16]. Dieses Modell zeigt eine bessere Vorhersage als die in dem Praat-Skript formulierten Regeln von De Jong & Wempe [11].

Der aktuelle Ansatz mit neuronalen Netzen soll im Folgenden systematisch erweitert und evaluiert werden. Dazu werden insbesondere Convolutional Neural Networks (CNNs) mit dem bisherigen Recurrent Neural Networks (RNN) verglichen und zusätzlich zu MFCCs auch als akustische Eingangsmerkmal die Einhüllende verwendet.

2 Vorarbeiten

Als Daten für einen ersten Schätzer wurden Mel-skalierte cepstrale Koeffizienten (MFCCs) zur Repräsentation des Sprachsignals verwendet. Diese wurden aus den semi-spontanen Terminabsprachen des Kielkorpus extrahiert [13]. Zur Evaluierung wurden diese Aufnahmen in ein Trainings- und ein Validierungsset geteilt (vgl. Abschnitt 3). Im Vergleich der beiden Verfahren, dem regelbasierten Ansatz und dem RNN, wurden die geschätzten Sprechraten über gesamte Äußerungen gemittelt. Die Evaluierung fand anhand des Pearson's Korrelationskoeffizienten *r* und der Wurzel der mittleren quadratischen Abweichung *RMSE* statt. Die Korrelation beträgt r=,44 (RMSE=23,61, siehe auch Tabelle 1). Bei der Entfernung der acht Äußerungen ohne erkannte Silbenkerne durch das Skript verändert sich die Korrelation auf r=0,43 (RMSE=22,05). Details zu den verwendeten Daten finden sich in Abschnitt 3.

Ziel der hier vorgestellten Arbeit ist ein Vergleich verschiedener aktueller neuronaler Netze für die Verbesserung des Schätzers. Dafür werden verschiedene Ansätze, Netzgrößen und Eingangsmerkmale miteinander verglichen. Im Vergleich zu [16] wurden folgende Änderungen umgesetzt:

- Die batches der Trainingsdaten werden nicht mit einzelnen Aufnahmen, also Dateien mit variablen Längen) gleichgesetzt, sondern konstant auf die batch-size von 5000 gesetzt. Daraus folgt die Festlegung der batches auf 170, um bei zufälliger Auswahl ungefähr auf die Größe der Trainingsdaten zu kommen.
- Stille bzw. Tempowerte von 0 werden nicht ausgeschlossen, um in konkreten Anwendungsfällen von Voice-Activity-Detection unabhängig zu sein.
- Alle Daten werden Anhand des Mittelwertes und der Standardabweichung der Trainingsdaten normalisiert.

Durch diese Maßnahmen sind die Ergebnisse vom letzten Jahr nicht direkt vergleichbar, weswegen auch der regelbasierte Ansatz hier wiederholt wurde, mit den Einstellungen *silence threshold*=25dB, *minimum dips between peaks*=3 und *minimum pause duration*=0.3s. Für das auf Grundfrequenz und Intensität beruhende Skript ergibt sich eine Korrelation von r=,71 (RMSE=0,33).

3 Material

Von der verwendeten Datenbank mit insgesamt 32 Sprechern (14 Frauen, 18 Männer), mit Aufnahmen von etwa 3,75 Stunden, werden etwa 70% der Aufnahmen für das Training und der Rest, nämlich ab Dialog "G364" für die Validierung verwendet. Als echtes Testset werden mit den vorgelesenen kurzen Geschichten aus Phondat 1 (der „Buttergeschichte" und „Nordwind und Sonne" andere Sprecher und eine andere Gattung genutzt [7].

Die „perzeptuelle lokale Sprechraten" (PLSR) [10] ist die lineare Kombination aus der Silben-, aber auch Phonrate. Aus beiden Ebenen der Transkription werden separat Kehrwerte der jeweiligen Segmentdauern in zwei separate Signale überführt, die über ein 625 ms Hann-Fenster geglättet werden. Das Ergebnis ist lokal und kontinuierlich, und wurde mit Perzeptionsexperimenten auf Ihre Validität für das Deutsche überprüft. Die beiden Signale wurden auf folgende Weise miteinander kombiniert:

$$(1)\quad PLSR = 8{,}14 * Silbenrate + 3{,}31 * Phonrate + 6{,}07$$

Alle Aufzeichnungen mit einer Gesamtdauer unter 1.5 s wurden ausgeschlossen, so dass von ursprünglich 1281 Dateien zum Training und 555 Dateien für die Evaluierung noch 1115 und 475 verwendet werden.

Für das Testset wurde PLSR anhand automatisiert segmentierter Annotationen berechnet. Die Konversion des an der TU Berlin vorhandenen Korpus in das EMU-Datenbankformat schlug aufgrund überlappender Segmente für einige Äußerungen fehlt, so dass 60 Dateien von 58 Sprechern (30 Männern, 28 Frauen) mit insgesamt 71.9 Min. Dauer genutzt werden. Die Automatisierung der Segmentierung hat auch zur Folge, dass sich die PLSR Werte systematisch unterscheiden müssten, da bspw. Glottalverschlüsse deutlich länger zu sein scheinen und da Verschluss- und Öffnungsphase nicht, wie in den Kielkorpus-Daten [15] zu einem einzigen Plosivsegment zusammengefasst wurden. Diese strukturellen Unterschiede wurden bewusst nicht verändert, da sie sich komplett nur mit erheblichem manuellem Aufwand angleichen ließen und somit ein realistisches Testszenario für das Trainieren und Evaluieren mit verschiedenen Datenbanken darstellen.

4 Modellansätze und akustische Merkmale

Wie bei dem Ansatz 2017 werden als Eingangsdaten 30 Werte pro (MFCC-)Merkmal als zeitlicher Kontext von 300 ms verwendet. Für diese Daten wird der späteste vorliegende PLSR-Wert dieses Zeitfensters als Zielwert genutzt. Dieses überlappende Fenster mit Eingangsgrößen wird nun schrittweise, und im Gegensatz zu [16], mit jedem Wert, also alle 10 ms anstatt alle 30 ms, verschoben, da eine ressourcenschonendere Arbeitsweise umgesetzt wurde. Die Anzahl von Epochen bleibt mit 30 identisch. Anstatt alle „batches“ für eine Epoche zu verwenden, wurde jetzt diese Anzahl auf 170 festgelegt und die jeweiligen „batches“ zufällig ausgewählt. Außerdem wurden einige wenige PLSR-Daten bereinigt, um auch alle Häsitationen vollständig zu berücksichtigen. Dies war in der früheren Version nicht für alle Token der Fall.

Als direkter Vergleich verschiedener Ansätze werden rekurrente Netze mit LSTM Zellen (LSTM), sowie Convolutional Neural Networks (CNN) mit den zuletzt verwendeten MFCC-Merkmalen trainiert und evaluiert. Die 26 MFCC mit ihren Ableitungen (δ) wurden alle 10 ms mit einer Fensterlänge von 25 ms erhoben. Als weitere Testbedingung wird die zusätzliche Verwendung der Einhüllenden getestet, die als Repräsentation Intensitätskontur einen inhaltlich starken Bezug zum Silbenkerninformation hat.

Insgesamt werden neun Bedingungen für diese beiden Netzarten getestet: jeweils alle 26 MFCCs, 13 MFCCs, jeweils mit und ohne δ, und jeweils mit und ohne Einhüllende (Env), sowie nur die Einhüllende, vgl. Tabelle 1.

Als ein derzeit verbreiteter Ansatz für Sprachdaten werden CNN-Schichten mit (einer) späteren RNN-Schicht verbunden. Auch dieser Ansatz wird hier rudimentär, also mit jeweils einer Schicht und Standardwerten, überprüft.

Abschließend wird getestet, ob ein einfacher, aber aktueller Ansatz der direkten Nutzung von eindimensionalen Zeitsignalen, hier die Einhüllende, vergleichbare Ergebnisse liefert. Dieser Ansatz wurde bspw. für die Grundfrequenzdetektion verwendet, indem Fenster des Audiosignals direkt in mehreren 1D-CNN Schichten mit abschließendem FF-Schicht verarbeitet werden [6].

Tabelle 1 - Liste der verglichenen Modelle zur Tempo-Schätzung (Werte gemittelt über Aufnahmen)

Architektur	Merkmale	Validierung	Test
LSTM(32) + FC(1)	53: 26MFCC, 26δ, Env	r=,79 (RMSE=0,46)	r=,49 (RMSE=0,30)
	52: 26MFCC, 26δ	r=,86 (RMSE=0,33)	r=,52 (RMSE=0,41)
	27: 26 MFCC, Env	r=,76 (RMSE=0,46)	r=,35 (RMSE=0,36)
	26: 26 MFCC	r=,81 (RMSE=0,42)	r=,53 (RMSE=0,31)
	27: 13MFCC, 13δ, Env	r=,83 (RMSE=0,38)	r=,53 (RMSE=0,32)
	26: 13MFCC, 13δ	r=,88 (RMSE=0,34)	r=,60 (RMSE=0,34)
	14: 13MFCC, Env	r=,78 (RMSE=0,34)	r=,37 (RMSE=0,41)
	13: 13MFCC	r=,82 (RMSE=0,35)	r=,52 (RMSE=0,34)
	1: Env	r=,76 (RMSE=0,44)	r=,29 (RMSE=0,32)
CNN(32) + MaxPooling2D + FC(1)	53: 26MFCC, 26δ, Env	r=,65 (RMSE=0,34)	r=,33 (RMSE=0,53)
	52: 26MFCC, 26δ	r=,56 (RMSE=0,55)	r=,42 (RMSE=0,31)
	27: 26 MFCC, Env	r=,42 (RMSE=0,64)	r=,21 (RMSE=0,38)
	26: 26 MFCC	r=,60 (RMSE=0,94)	r=,33 (RMSE=0,53)
	27: 13MFCC, 13δ, Env	r=,60 (RMSE=0,73)	r=,32 (RMSE=0,58)
	26: 13MFCC, 13δ	r=,59 (RMSE=0,33)	r=,30 (RMSE=0,49)
	14: 13MFCC, Env	r=,50 (RMSE=0,53)	r=,40 (RMSE=0,29)
	13: 13MFCC	r=,56 (RMSE=0,79)	r=,48 (RMSE=0,42)
	1: Env	r=,59 (RMSE=0,38)	r=,27 (RMSE=0,63)
CNN(32)+LSTM(1)+FC(1)	Bestes Set (13MFCC 13δ)	r=,78 (RMSE=0,33)	r=,54 (RMSE=0,33)
CREPE-modifiziert CNN(128,16,32,64) + D(1 linear)	Einhüllende	r=,86 (RMSE=0,38)	r=,63 (RMSE=0,36)

4.1 Recurrent Neural Networks

Das ursprüngliche Modell basiert auf einer sehr einfachen LSTM-RNN Architektur mit nur einem RNN Layer, dessen Ausgaben der LSTM Zellen mit einer abschließenden Fully-Connected (FC) Layer aggregiert werden. Dabei sind fast ausschließlich Standardeinstellungen verwendet worden[1], bspw. “tanh“ Aktivierungen für die LSTM-Zellenausgaben,

1 https://keras.io, mit Tensorflow als backend

“lineare“ Aktivierungen für die FF-Schicht, ein Startwert für die Lernrate von 0,01 mit „adam“ Optimierer, sowie der mittlere Quadratfehler als „loss“-Funktion. Es wurde auch auf Regularisierungsmethoden verzichtet. Lediglich die Initialisierung der FF-Schicht wurde von „glorot_uniform“ auf „normal“ umgestellt. Aufgrund der in Abschnitt 2 dargestellten Veränderungen wurde dieser Ansatz mit unterschiedlichen Parametern neu trainiert und den neuen Testdaten evaluiert. Diese Architektur wird nun mit ebenso rudimentären CNNs vergleichen werden.

4.2 Convolutional Neural Networks

Da CNNs eine hohe Anzahl neuer Merkmale erzeugen, wird eine weitere Schicht zur Aggregierung („MaxPooling2D“) vor einer abschließenden FF-Schicht eingefügt. Die Anzahl der Filter wurde variiert, während die „kernel-size“ auf (3, 3) und die „strides“ auf jeweils 1 gesetzt wurde. Als Aktivierungsfunktion wurde „relu“ gewählt. Die genauen Architekturen sind in Tabelle 1 angegeben.

4.3 Andere Ansätze

Die CREPE genannte Architektur nutzt direkt das Sprachsignals zur Grundfrequenzdetektion. Sie verwendet 1024 Samples bei 16kHz Sprachsignalen als Eingangsdaten und eine Schrittweite von 10ms. Zum direkten Vergleich mit den CNN und RNN wurde auf Regularisierung, in diesem Fall batch-Normalisierung und drop-out Layer, verzichtet (vgl. Tabelle 1). Außerdem wird anstatt des rohen Sprachsignals die Einhüllende als Eingangsdatum verwendet, allerdings aufgrund der langsam variierenden Daten mit reduzierten Schichten und Parametern (width=30,4,4,4; strides=1, pool-size=2). Die für Grundfrequenz optimierte Fensterlänge von 64ms erscheint für die Tempoerkennung zu gering, weshalb eine Fenstergröße von 300 ms, wie bei den anderen Modellen, getestet wird. Dazu wurde die Ausgangsschicht mit 360 Zellen für Wertintervalle durch ein einzelnes Ausgangsneuron für den Gesamtwert ersetzt.

5 Ergebnisse und Diskussion

Die in Tabelle dargestellten Ergebnisse zeigen jeweils das Ergebnis mit dem höchsten Pearson‘s r aus zwei Trainingsläufen für die Validierungsdaten und dem dazugehörigen Ergebnis für das Testset. Dabei zeigt sich kein deutlicher Gewinn durch mehr Merkmale. Insbesondere die Einhüllende verbessert nicht die Modelle, weder bei LSTMs, noch bei CNNs. Als einziges Merkmal zeigt es aber bereits gute Ergebnisse auf den Validierungsdaten, jedoch nicht für die Testdaten. Interessanterweise sind die Validierungsergebnisse für CNNs und LSTM durchweg besser für die per Aufnahme gemittelten Daten, während die Testdaten (hier nicht abgebildet) systematisch besser (für zahlreiche Bedingungen r>.60) für die ungemittelten Werte sind. Insgesamt erscheint der Ansatz mit LSTMs vielversprechender als CNNs, während eine Kombination nicht besser als ein vergleichbares LSTM wird.

Nicht überraschend zeigt sich die komplexere Architektur als überlegen. Selbst ein deutlich vereinfachtes CREPE-inspiriertes Modell mit 4 CNN Schichten für die Einhüllende erreicht vergleichbare Ergebnisse wie das beste der grundlegenden Modelle (LSTM mit 13 MFCCs und deltas). Dieses Ergebnis motiviert zur zur direkten Nutzung des Sprachsignals ohne Vorverarbeitung.

6 Fazit

Die hier gezeigten Ansätze zeigen brauchbare Ergebnisse für verschiedene Merkmalsbündel und Modell-Architekturen. Mit dem zusätzlichen Datensatz konnte prinzipiell die Generalisierbarkeit belegt werden. In einer abschließenden Arbeit muss nun ein Modell ausgewählt und trainiert werden, das bezüglich Echtzeit- und Anwendungsfähigkeit optimiert ist. Die positiven Ergebnisse für die (nicht gemittelten) Einzelwerte der Testdaten sind für den Anwendungsfall der Nutzung für Sprachdialogsysteme mit kurzen Äußerungen, bspw. für die Schätzung von Tempovariabilität, sehr vielversprechend.

Literatur

[1] COUTINHO, E. AND DIBBEN, N.: "Psychoacoustic cues to emotion in speech prosody and music", *Cognition and Emotion 27,* 2013, 658—684.

[2] DAMIAN, I., TAN, C.S.S, BAUR, T., SCHÖNING, J., LUYTEN, C., ANDRÉ, E.: „Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques", *Proc. ACM Conference on Human Factors in Computing Systems*, 2015, 565—574.

[3] DE JONG, N. AND WEMPE, T.: "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior Research Methods 41(2),* 2009, 385—390.

[4] LEHISTE, I.: *Suprasegmentals*. Cambridge: MIT Press, 1970.

[5] DE LOOZE, C., SCHERER, S., VAUGHAN, B., CAMPBELL, N.: "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction", *Speech Communication 58,* 2014, 11—34.

[6] KIM, J.W., SALAMON, J., LI, P., BELLO, J.P.: Crepe: A Convolutional Representation for Pitch Estimation. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[7] KOHLER, K. J., PÄTZOLD, M., SIMPSON, A. P.: "*From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech", Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel* (AIPUK) 29, 1995.

[8] LUBOLD, N. AND PON-BARRY, H.: "Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues", *Proc. ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014 5—12.

[9] TER MAAT, M., TRUONG, K., HEYLEN, D.: "How Agents'Turn-Taking Strategies Influence Impressions and Response Behaviors", *Presence 20,* 2011, 412—430.

[10] PFITZINGER, H.R.: "Local speech rate perception in German speech", *Proc. of the 14th ICPhS*, 1999, 893–896.

[11] QUENÉ, H., PERSOON, I., DE JONG, N.: "Praat Script Syllable Nuclei v2", 2010, url: https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2

[12] SCHERER, S., WEIBEL, N., MORENCY, L.-P, OVIATT, S.: "Multimodal prediction of expertise and leadership in learning groups", *Proc. 1st International Workshop on Multimodal Learning Analytics*, 2012, paper 1, 1—8.

[13] SIMPSON, A.P. (1998): „Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung", *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)*, Band 33.

[14] WEISS, B.: "Sprechtempoabhängige Aussprachevariationen", *Doktorarbeit*, Humboldt Universität zu Berlin, 2008.

[15] Weiss, B.: „Akustische Korrelate von Sympathieurteilen bei Hörern gleichen Geschlechts". In: *26. Konferenz Elektronische Sprachsignalverarbeitung* (ESSV), Eichstädt. Vol. 78. Studientexte zur Sprachkommunikation. Dresden: TUDpress, 2015, pp. 165–171.

[16] WEISS, B., MICHAEL, T., HILLMANN, S.: "Kontinuierliche Schätzung von Sprechgeschwindigkeit mit einem Rekurrenten Neuronalen Netzwerk". In: *29. Konferenz Elektronische Sprachsignalverarbeitung* (ESSV), Ulm. *Studientexte zur Sprachkommunikation*. Dresden: TUDpress, 2018, pp. 186–191.