Kontinuierliche Schätzung von Sprechgeschwindigkeit mit einem Rekurrenten Neuronalen Netzwerk

Benjamin Weiss, Thilo Michael, Stefan Hillmann
TU Berlin
vorname.nachname@tu-berlin.de

Sprechgeschwindigkeit zählt mit der Intensität und der Grundfrequenz zu den drei primären Prosodien [1]. Alle drei spielen eine zentrale Rolle bei der Analyse und Synthese mündlicher Kommunikation (z.B. Sprecherwechsel, Emphase, Informationsstruktur), wie auch der Paralinguistik (Ausdruck und Zuschreibung von Sprecherzuständen und -merkmalen). Von diesen drei primären Prosodien stellt Sprechgeschwindigkeit derzeit auch die größte Herausforderung an die Erhebung aus einem Sprachsignal. Während akkurate Algorithmen zur Messung von Intensität und zur Schätzung der Grundfrequenz existieren, sind für Sprechgeschwindigkeit keine Methoden mit vergleichbarer Güte vorhanden. Dies liegt vor allem daran, dass Sprechgeschwindigkeit akustisch über die Anzahl linguistischer Einheiten in einem Zeitintervall erfasst werden (bspw. Silben pro Sekunde in germanischen Sprachen), und damit eine direkte akustische Definition fehlt. Der oben genannte Anwendungsbedarf hat zu Ansätzen geführt, die keine automatische Spracherkennung verwenden, sondern direkt aus dem Signal Silbenkerne schätzen, aus denen die Sprechgeschwindigkeit abgeleitet wird. Bei dem wohl aktuell verbreitetsten Verfahren handelt es sich um ein Praat-Script [2], seit dem Jahr 2010 in einer modifizierten Form [3], das über den Intensitätszeitverlauf und Stimmhaftigkeit die Silbenkerne und Pausen schätzt, um direkt Sprechgeschwindigkeit (Silbenanzahl pro Zeitintervall) und Artikulationsrate (Silbenanzahl pro Artikulationszeit) – gemittelt über die Aufnahme – bereitzustellen. Es wird derzeit (Stand 24.11.17) 262 Mal bei Google Scholar zitiert, weist aber auch bekannte Mängel bei der Erkennung unbetonter Silben auf. Der eigene Beitrag stellt einen Ansatz mit rekurrenten neuronalen Netzen und Long Short-Term Memory Zellen vor (32 MFCCs, deltas als Eingabe). Als Datenbasis wurde das Kielkorpus [4] genutzt, für das ein perzeptives und kontinuierliches Maß lokaler Sprechgeschwindigkeit [5] vorliegt [6]. Neben der Erhöhung der Schätzerleistung durch neuronale Netze (gemittelt pro Datei von r=.41 auf r>.70 für 30% der Datenbasis) ergeben sich weitere Vorteile durch das neue Verfahren:

1. Schnelle, kontinuierliche Schätzung (die Performanz in den frühen Signalphasen wurde jedoch noch nicht evaluiert).
    a. notwendig für die Umsetzung von akustisch-prosodischem Entrainment in Dialogsystemen
    b. direkt nutzbar für stark variierende Aufnahmedauern
2. Ein generischer Ansatz, der unabhängig von der Sprachfamilie genutzt werden kann.
3. Unabhängigkeit von automatischer Spracherkennung
4. Die Zielgröße ist kontinuierlich und kann somit auch Tempovariabilität für Material von wenigen Silben erfassen.

Der Vollbeitrag beschreibt im Detail die Datenaufteilung, Schätzer-Modellierung und -evaluierung, sowie den Vergleich der beiden Ansätze und diskutiert die Ergebnisse.

[1] Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., Neuschaefer-Rube, C. (2017): “Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis”, *Computer Speech and Language 41,* 116—127.
[2] de Jong, N. and Wempe, T. (2009): “Praat script to detect syllable nuclei and measure speech rate automatically”, *Behavior Research Methods 41(2),* 385—390.
[3] Quené, H., Persoon, I., de Jong, N. (2010): “Praat Script Syllable Nuclei v2”, url: https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2
[4] 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.
[5] Pfitzinger, H.R. (1999): “Local speech rate perception in German speech”, *Proc. of the 14th ICPhS*, San Francisco, 893–896.
[6] Weiss, B. (2008): “Sprechtempoabhängige Aussprachevariationen“, *Doktorarbeit*, Humboldt Universität zu Berlin.