KONTINUIERLICHE SCHÄTZUNG VON SPRECHGESCHWINDIGKEIT MIT EINEM REKURRENTEN NEURONALEN NETZWERK

Benjamin Weiss, Thilo Michael, Stefan Hillmann

Quality and Usability Lab, Technische Universität Berlin
vorname.nachname@tu-berlin.de

Kurzfassung: Um händische Segmentation bei der Analyse von Tempo aus Sprachaufnahmen zu vermeiden, können automatische Spracherkenner oder Verfahren wie das Praat-Skript von de Jong und Wempe 2009 verwendet werden. Als Ersatz wurde ein Verfahren für die kontinuierliche Schätzung von Sprechtempo entwickelt, das nicht wie automatische Spracherkenner auf ein Sprachmodell zurückgreifen muss, und dennoch praktikabler als das etablierte Praat-Skript ist, da es die Sprechzeit präziser berücksichtigt. Unser Ansatz nutzt ein rekurrentes neuronales Netz mit LSTM Zellen und wurde mit ca. 70% des Kiel Korpus trainiert. Das resultierende Modell schätzt die verbleibenden 30% Testdaten mit r=0,9.

1 Einführung

Sprechgeschwindigkeit zählt mit Intensität und 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 die Sprechgeschwindigkeit derzeit auch die größte Herausforderung an die Erhebung aus einem Sprachsignal. Während genaue 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 wird (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-Skript [4], seit dem Jahr 2010 in einer modifizierten Form [9], das über den Intensitätsverlauf und Stimmhaftigkeit die Silbenkerne und Pausen schätzt, um direkt Sprechgeschwindigkeit (Silbenanzahl pro Zeitintervall) und Artikulationsrate (Silbenanzahl pro Artikulationszeit) – gemittelt über die Aufnahme – bereitzustellen. Dieses Verfahren wird als Referenz verwendet, um einen aktuellen Ansatz, der auf künstliche rekurrente neuronale Netze (RNN) basiert, zu entwickeln und zu evaluieren. Auf die Alternative, einen Spracherkenner zu nutzen, um Silben und Lautdauern zu extrahieren, wurde verzichtet, um ein einfaches, anpassbares und konzeptuell schnelles Verfahren zu erreichen. Ein abschließender Vergleich der Laufzeiten zwischen dem hier vorgestellten Verfahren und Ansätzen mit automatischer Spracherkennung steht noch aus.

2 Automatische Schätzung von Sprechgeschwindigkeit

Das Praat-Skript von De Jong und Wempe [4] wurde für das Niederländische entwickelt. Es schätzt zuerst Stille und berechnet danach die Intensität mit einem Zeitfenster von 64 ms. Als lokale Intensitätsmaxima gelten solche mit mehr als 0 dB, bzw. 2 dB über dem Median, jedoch nur solche, die eine bestimmten vorherigen Intensitätsabfall nachfolgen. Abschließend werden alle potentiellen Silbenkerne entfernt, die stimmlos sind, um intensitätsreiche Konsonanten auszuschließen. Die Sprechrate und Artikulationsrate werden dann als Anzahl von Silbenkernen über die Äußerungsdauer bzw. Artikulationsdauer ermittelt.

Zur Evaluierung dieses Ansatzes wurde von 258 Personen ca. 46 Stunden Material aufgezeichnet, jeweils 8 Aufgaben pro Sprecher, und davon 50 Aufzeichnungen (75 min) ausgewählt und händisch auf Silbenebene annotiert. Talkspurts, also Sprechphasen, wurden anhand von Pausenzeiten ab 0,4 sec definiert. Die minimale Spurt-Dauer wurde auf 5 sec gesetzt. Die Korrelation zwischen der Sprechgeschwindigkeit aus annotierten und automatisch geschätzten Silbenkernen beträgt r=,71 für 441 *spurts*, und r=,88 für die 50 Sprecher, also gemittelt für je eine komplette Aufnahme.

Ein zweiter Datensatz besteht aus 125 min semi-spontaner Sprache von acht Sprechern (4m, 4w). Hier beträgt die Korrelation r=,77 für 1171 Spurts und r=,80 gemittelt pro Aufnahme (3 Aufgaben * 8 Sprecher).

Dieses Verfahren wird derzeit (Stand 24.11.2017) 262 Mal bei Google Scholar zitiert, weist aber auch bekannte Mängel bei der Erkennung unbetonter Silben auf. Die aus der Evaluierung hervorgehende hohe Übereinstimmung mit hand-annotierten Daten ist vermutlich der Grund, dass dieses Verfahren ohne Überprüfung für andere Daten und Sprachen verwendet wird. Als Beispiele seien hier die Forschungsthemen akustisch-prosodisches Entrainment [5,6], Sprecherattributionen und- externalisierungen [2,10], Soziale Signale [3], sowie Turn-Taking [7] genannt.

Eigene Anwendungen zeigten jedoch durchaus ein Abweichen von händisch erfasster Artikulationsrate. Für bestimmte Anwendungen mag solch ein systematisches Abweichen zu händisch annotierten Daten vernachlässigbar sein, solange dieses Abweichen als vergleichbar für alle Sprachsignale angenommen werden kann. Spezielle Anwendungen, wie etwa der Vergleich mit Daten aus anderen Quellen oder die Steuerung von Sprachsynthesen, verlangen jedoch eine hohe Genauigkeit. Zudem lässt der Erfolg neuronaler Netze in der Sprachtechnologie in den letzten Jahren das Verfahren von de Jong und Wempe als potentiell veraltet erscheinen.

Deshalb wird anhand eines bestehenden deutschen Korpus ein RNN trainiert und mit dem etablierten Verfahren verglichen. Neben einer potentiellen Steigerung der Genauigkeit wurde dieser Ansatz gewählt, da er weitere Vorteile verspricht:

1. Schnelle, kontinuierliche Schätzung von Sprechgeschwindigkeit
    a. notwendig für die Umsetzung von akustisch-prosodischem Entrainment in Dialogsystemen
    b. direkt nutzbar auch für kurze Aufnahmedauern
    c. Tempovariabilität kann erfasst werden
2. Verallgemeinerbarkeit
    a. unabhängig von der Sprachfamilie nutzbar
    b. unabhängig von automatischer Spracherkennung

3 Datenmaterial

Aus dem Kielkorpus wurde der Teil semi-spontaner Dialoge zur Terminabsprache [11] verwendet. Von insgesamt 32 Sprechern (14 Frauen, 18 Männer) des Standarddeutschen mit meist norddeutscher Ausprägung stehen Aufnahmen von etwa 3,75 Stunden zur Verfügung. Aufgrund der Experimentalsituation sind für alle Sprecher getrennte Monodateien (16 kHz, 16 bit) vorhanden. Aus den vorhanden segmentalen Transkriptionen sind reziproke Laut- und Silbendauern in zwei Signale überführt worden. Die „perzeptuelle lokale Sprechraten“ (PLSR) [8] ist die lineare Kombination aus den beiden Signalen, die über ein 625 ms Hann-Fenster geglättet wurden. Das Ergebnis ist nicht nur lokal und kontinuierlich, und erlaubt damit die Analyse von Tempovariabilität auch bei kurzen Äußerungen, sondern korreliert stärker mit Daten aus Wahrnehmungsexperimenten im Deutschen als die Silbenrate allein, da es bspw. die Silbenkomplexität berücksichtigt. Die verwendete lineare Kombination der beiden Raten beträgt:

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

Auf das Korpus mit der errechneten PLSR [12] konnte hier zurückgegriffen werden. Für das Training des RNN wurden 70% (bezogen auf die Dateianzahl) der Daten verwendet, um 30% für die Evaluierung zu behalten. Dabei wurden die Sprecher strikt aufgeteilt, d.h. das Material desselben Sprechers wurde entweder ausschließlich im Trainigsset oder ausschließlich im Testset verwendet. Alle Sprecher wurden fortlaufend von "G364" inklusive dem Evaluierungsset zugeordnet. So werden 1281 Dateien zum Training, und 555 Dateien für die Evaluierung verwendet.

4 Modellansatz und Training

Rekurrente neuronale Netze (RNN) nutzen die Ausgaben einer Schicht von Zellen als Eingabe für dieser oder früherer Schichten, um Abhängigkeiten zwischen verschiedenen Eingaben zu berücksichtigen. Damit eignen sie sich besonders für Zeitreihen, wie sie mit der PLSR vorliegen. Die direkte Rückkopplung mit Long Short-Term Memory (LSTM) Zellen sind hierbei Stand der Technik, da LSTM auch die Modellierung über längere Zeitfenster ermöglichen. In wie weit eine längere Backpropagation über die Zeit notwendig ist, wird im Vergleich zu kurzen Zeitfenstern überprüft.

Ein erstes RNN mit 52 LSTM Zellen wurde mit Keras[1] (Tensorflow als Backend) definiert, das mit 26 MFCCs und deren Ableitung (25 ms Fenstergröße, 10 ms Schrittweite) einen Eingabevektor von 52 Datenpunkten erwartet. Die Zielgröße PLSR wurde kontinuierlich modelliert. Beide Datenarten wurden für die Modellbildung normalisiert. Das Modell wurden mit einer Länge von 30 Eingabevektoren (*minibatch*) und einer Schrittweite von 3 trainiert. Die Schrittweite konnte hier aufgrund einer älteren Hardware nicht auf 1 gesetzt werden, da mit der derzeitigen Implementierung der RAM von 8GB nicht ausreichte. Dies wird jedoch in einer zukünftigen Version optimiert. Pro *minibatch,* die alle Daten einer Zeitreihe darstellen, wurde sowohl ein einzelner normalisierter PLSR-Wert, wie auch 30 Datenpunkte getestet. Dem RNN schließt sich eine Ausgabeebene mit 52 Neuronen und linearer Aktivierung an, die die Ausgabewerte aus den LSTM Zellen aggregiert. Als *loss*-Funktion wird der mittlere Quadratfehler (MSE) und als Optimierer der Lernrate *adam* mit 0,01 als Startwert für die Lernrate verwendet. Die batch-size beträgt 5000.

5 Evaluierung

Das RNN wird für dieselben Trainingsdaten mit der Schätzung durch die Methode von de Jong & Wempe verglichen. Für das Praat-Skript wurde die geschätzte Silbenrate pro Aufnahme mit den gemittelten Werten der PLSR aus der Datenbank korreliert. Dabei erwiesen sich die folgenden Einstellungen am besten, bezogen auf Pearson's Korrelationskoeffizient *r* und der Wurzel der mittleren quadratischen Abweichung *RMSE*: Silbenrate anstatt Artikulationsrate, silence threshold=25dB und minimum dips between peaks=3 (die minimum pause duration wurde nicht gestestet und blieb bei der Voreinstellung von 0,3 s). Die Korrelation beträgt r=,44 (RMSE=23,61), siehe auch Abbildung 1. Bei der Entfernung der acht Äußerungen ohne erkannte Silbenkerne verändert sich die Korrelation auf r=,43 (RMSE=22,05).

Für die Evaluierung des RNN wurden nur wenige Parametereinstellungen getestet. Das RNN mit many-to-one Struktur und 30 Samples Zeitfenster erreicht hierbei mit einer Korrelation von r=,9 (RMSE=11.29) eine deutlich bessere Vorhersage, wie in Abbildung 2 zu sehen ist. Dabei wird der Zeitverlauf nicht gleichmäßig gut modelliert (Abbildung 3).

Im direkten Vergleich zur many-to-many Version mit 30 PLSR Werten steigt die Korrelation nur wenig (r=,92, RMSE=10,58). Dagegen hat eine Verringerung der berücksichtigten Zeitspanne stärkere Konsequenzen, da bei einer Reduzierung auf 10 Samples die Korrelation auf

[1] https://*keras.io*

r=,86 (RMSE=17,10) und bei einem Sample auf r=,75 (RMSE=24,40) sinkt, während eine Erhöhung auf 40 Samples mit r=,91, RMSE=12,16 zu keiner Verbesserung führt. Während hier also bereits gute Parametereinstellungen für die gegebenen Daten gefunden wurden, ist die Nutzung von 52 LSTM-Zellen vermutlich noch übertrieben, da auch lediglich 10 Zellen mit nachfolgender *hidden layer* mit 25 einfachen Neuronen zu r=,89 (RMSE=12,35) und bei 10 Zellen zu r=,87 und RMSE=13,15 führt.

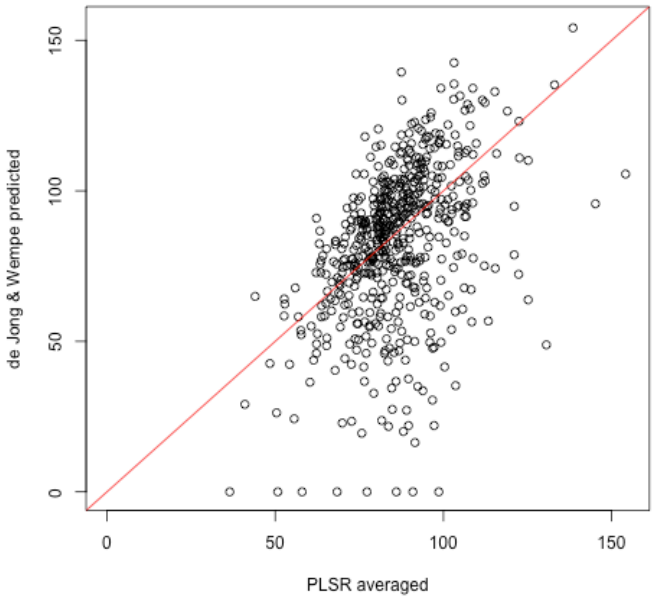

Abbildung 1 –PLSR-Werte gemittelt pro Aufnahme und die vorhergesagten Werte nach De Jong & Wempe

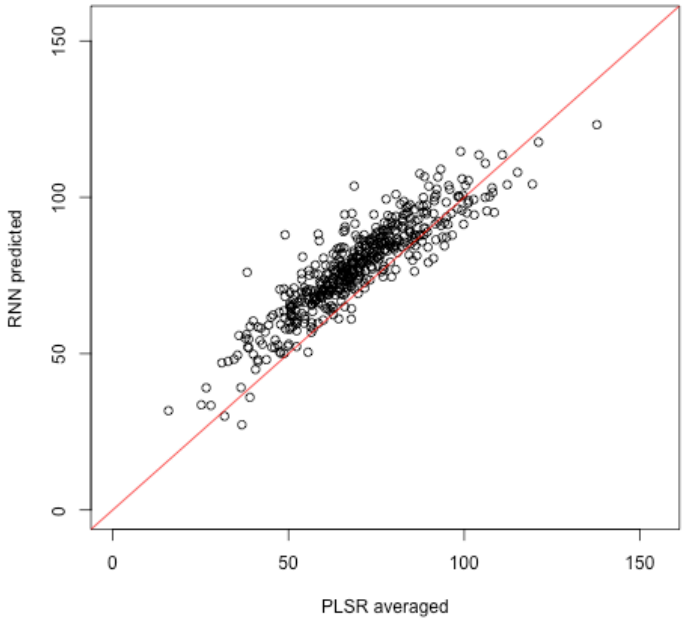

Abbildung 2 – PLSR-Werte gemittelt pro Aufnahme und die vorhergesagten Werte durch das RNN

6 Diskussion und Ausblick

Die mäßige Güte des Praat-Skriptes bestätigt frühere, anekdotische Erfahrungen bei der automatischen Schätzung von Sprechtempo für die Modellierung von Sympathie in Stimmen mit akustischen Parametern [13]. Da die PLSR hier ein kontinuierliches, gefenstertes Maß ist, das zudem noch über die Aufnahmedauer gemittelt wurde, stellt sich natürlich die Frage, ob das Praat-Skript die ursprüngliche Aufgabe der Silbendetektion besser bewältigt. Hier zeigt sich die eigentliche Stärke des Praat-Skriptes, dessen erkannte Silbenanzahl pro Datei mit den tatsächlichen Silben aus der Annotation mit r=,995 (RMSE=5,37) fast perfekt korreliert. Anscheinend liegt das Problem bei der Schätzung der Sprechzeit bzw. Stille. Ein Vergleich der Silbendetektion mit dem neuen Verfahren steht noch aus.

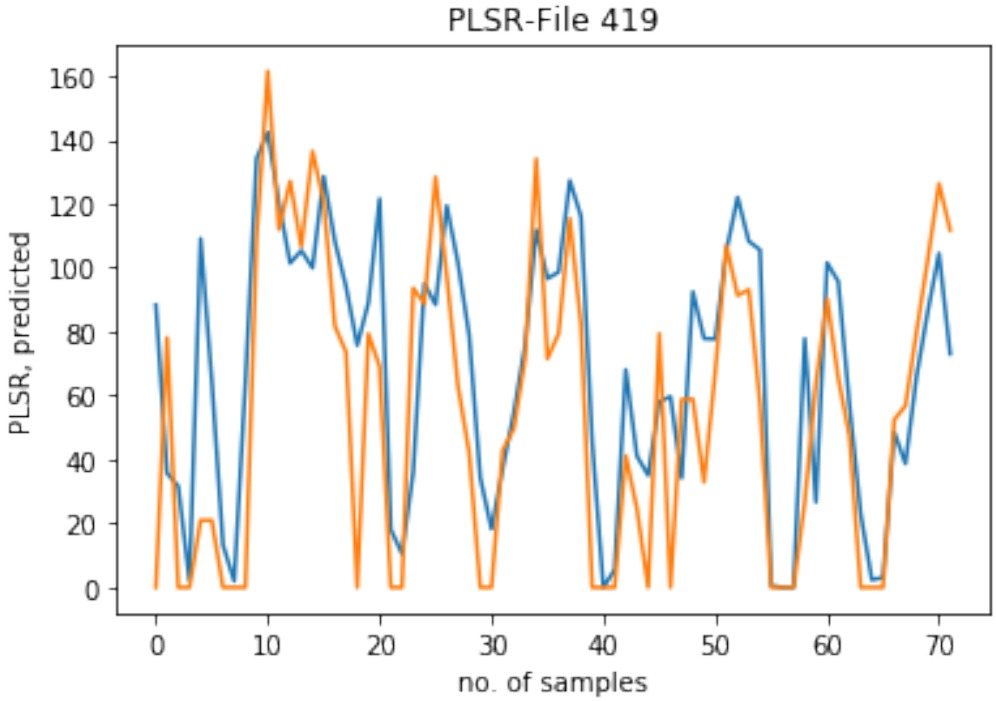

Abbildung 3 – Zufällige Aufnahme (Nr. 419) zeigt tatsächliche PLSR -Werte über die Dauer (gelb) und die vorhergesagten Werte des RNN (blau)

Das RNN-basierte Modell ist natürlich auf die Domäne und Sprache trainiert, so dass weitere Datenbanken zur Validierung und Generalisierung neben gelesener auch Spontansprache beinhalten müssen. Eventuell können auch Sprachsynthesesysteme für die Materialgenerierung verwendet werden.

Neben dieser Verbesserung des Schätzers durch das neue Verfahren – ausschließlich für das Sprechtempo, nicht die Silben, – soll der neue Ansatz auch einige weitere praktische Vorteile bringen. So soll mit der kontinuierlichen Schätzung von Sprechtempo ermöglicht werden, auch Schätzwerte für Zeitfenster einer oder unter einer Silbendauer zu ermöglichen, um für inkrementelle Dialogsysteme nutzbar zu sein. Zudem kann damit die Tempodynamik für paralinguistische Fragestellungen analysiert werden.

Das vorliegende Modell kann bereits direkt für die (para-)linguistische Analyse von Sprachdaten genutzt werden kann, für die keine Segmentierung vorhanden sind. Allerdings steht ein Vergleich mit der Sprechtemposchätzung über automatische Spracherkennung noch aus. Demgegenüber muss für die Anwendung in Sprachdialogsystemen, die Sprechtempoinformationen

für bspw. akustisch-prosodische Entrainment nutzen sollen, die Güte des Schätzers noch speziell für Äußerungsanfänge und kurze Äußerungen überprüft werden, und die Latenz gemessen und ggf. verringert werden.

7 Fazit

Der Ansatz, einen kontinuierlichen Sprechtemposchätzer mit RNN zu erstellen, war erfolgreich und verspricht für die kontinuierliche Schätzung von Sprechtempo eine höhere Genauigkeit als ein bisheriges, etabliertes Verfahren. Jedoch sind noch weitere Validierungen durchzuführen, bevor die Einsetzbarkeit gewährleistet werden kann. Prinzipiell ist dieser Ansatz, wie viele vergleichbare akustische Modelle, sprachunabhängig, da im Gegensatz zu Ansätzen mit automatischer Spracherkennung, für die ein Sprachmodell bzw. eine sprachannotierte Datenbank erforderlich ist, eine Datenbank ausreicht, die lediglich auf der entsprechend relevanten Ebene (Silbe, Phon, Mora) segmentiert ist, um ein ähnliches Modell zu trainieren. Hier muss kein perzeptives Modell, wie das der PLSR zugrundeliegende, angenommen werden.

8 References

[1] Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., Neuschaefer-Rube, C.: "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis", *Computer Speech and Language 41,* 2017, 116—127.

[2] Coutinho, E. and Dibben, N.: "Psychoacoustic cues to emotion in speech prosody and music", *Cognition and Emotion 27,* 2013, 658—684.

[3] 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.

[4] 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

[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] 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.

[7] ter Maat, M., Truong, K., Heylen, D.: "How Agents'Turn-Taking Strategies Influence Impressions and Response Behaviors", *Presence 20,* 2011, 412—430.

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

[9] 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

[10] 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.

[11] 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.

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

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