ZUR EVALUIERUNG VON INTONATIONSKONTUREN BEIM FREMDSPRACHENERWERB

Tristan Langenberg und Oliver Jokisch

Institut für Kommunikationstechnik, Hochschule für Telekommunikation Leipzig (HfTL)

tristan.langenberg@hftl.de, jokisch@hftl.de

Kurzfassung: Systeme für das Aussprachetraining basieren im Allgemeinen auf konventioneller Phonemerkennung und ermöglichen eine dezidierte Lernrückkopplung bezüglich segmenteller Merkmale - z. B. durch die Markierung von Artikulationsfehlern. Prosodische (suprasegmentelle) Parameterverläufe, z. B. f0-Konturen oder rhythmische Strukturen können unter guten akustischen Bedingungen zuverlässig bestimmt und visualisiert werden. Darüber hinaus gibt es etablierte, multilinguale Prosodiemodelle und Untersuchungen zur Variation von Modellparametern bei einem fremdsprachlichen Akzent. In vielfältigen Projekten werden sprachübergreifende Daten von Lernenden und Muttersprachlern erhoben. Dennoch halten prosodische Bewertungsverfahren nur langsam Einzug in die Sprachlernsysteme.

Der Artikel diskutiert Ansätze zur Bewertung von f0-Konturen auf Basis einer einfachen Abstandsanalyse zu Referenzkonturen, unter expliziter Berücksichtigung eines quantitativen Intonationsmodells der Zielsprache sowie mittels impliziter Information, z. B. aus der Hauptkomponentenanalyse (PCA) muttersprachlicher Referenzphrasen. In einer Pilotstudie zur Sprachstufeneinteilung von Deutschlernern wird die PCA-Methode angewendet. Dabei wird ausschließlich auf f0-Information zurückgegriffen. Die Klassifikationsergebnisse werden mit der Bewertung durch den Lehrer bzw. mit perzeptiven Testergebnissen von Muttersprachlern verglichen.

1 Einführung

Der Fokus auf segmentelle Aspekte bei der Ausspracheevaluation im computer assisted pronunciation tutoring (CAPT) ergibt sich aus der breiten Verfügbarkeit von phonembasierten Spracherkennungssystemen (z. B. HMM-Erkenner) und umfangreichen, muttersprachlichen Referenzdaten. Dabei werden die vorhandenen Spracherkenner für den spezifischen Einsatzzweck (Verifikation vorgegebener Phonemsequenzen und -qualitäten) lediglich adaptiert. Suprasegmentelle (prosodische) Merkmale spielen eine untergeordnete Rolle in der Spracherkennung und damit auch in CAPT-Systemen. Prosodische Aspekte fremdsprachlicher Einflüsse (second language, L2) werden in vielen Studien untersucht und auf verschiedenen Ebenen qualitativ/symbolisch, quantitativ oder perzeptiv modelliert. Für eine Evaluierung prosodischer Merkmale, u. a. der Intonation, stellt sich die Frage geeigneter Abstandsmaße zwischen realisierten Verläufen der L2-Lerner und Referenzbeispielen von Muttersprachlern (first language, L1).

Im folgenden Abschnitt gehen wir kurz auf die Intonationsanalyse (f0-Verlauf) sowie mögliche Abstandsmaße als Evaluierungskriterium ein. Die zugrunde liegenden Intonationsmodelle der Zielsprache sind entweder komplex und fehleranfällig (z. B. Analyse der Fujisakiparameter), oder die Analyse ist zu unspezifisch (vgl. z. B. Abstandsanalyse zu Referenzkonturen mittels RMSE). Der dritte Abschnitt diskutiert die Berücksichtigung impliziter Information aus einer f0-Hauptkomponentenanalyse (principal component analysis, PCA) muttersprachlicher Referenzphrasen. In einer Kurzstudie zur Sprachstufeneinteilung von Deutschlernern (Abschnitt 4)

wird die PCA-Methode angewendet und eine einfache Abstandsklassifikation auf Basis der PCA-Komponenten durchgeführt. Dabei dient uns die Sprachstufenklassifikation auf Basis der beschränkten f0-Information lediglich als Arbeitshypothese für Demonstrationszwecke.

2 Intonationsanalyse und Abstandbewertung

Aktuelle L2-Evaluierungsansätze folgen in der Regel dem gängigen Spracherkennungsparadigma, alle relevanten Sprachparameter in höher-dimensionalen Merkmalsvektorfolgen abzubilden und einen geeigneten Klassifikator mit annotiertem Sprachdatenmaterial zu trainieren. Dabei werden verschiedene phonetisch-akustische, spektrale und prosodische Merkmale, die sich in der Spracherkennung, Sprecheridentifikation oder auch Emotionserkennung etabliert haben, gemischt und Evaluierungsergebnisse erzielt, die gut mit perzeptiven Urteilen zur L2-Qualität korrelieren. In Hönig et al. [1] erreicht die beste Merkmalskombination mit $k = 64$ Komponenten eine Pearson-Korrelation von $\rho = 0,620$ zum Perzeptionskriterium *melody*.

Für ein spezifisches Feedback an den L2-Lerner ist es allerdings wünschenswert, einzelne Zielparameter wie den Intonationsverlauf getrennt zu beurteilen, was zur etablierten, quantitativen f0-Modellierung z. B. mit dem Fujisakimodel (Superposition von Akzent- und Phrasenkommandos) führt [2]. Die analysierte Kommandosequenz stellt eine Verbindung zwischen linguistischer Funktion (Akzentuierung und Phrasierung) und äußerer Form (realisierte f0-Kontur) her, weist jedoch je nach algorithmischer Konfiguration Mehrdeutigkeiten auf und ist bereits bei der L1-Analyse fehleranfällig. In der L2-Analyse wird das Problem durch die höhere Variation verstärkt. Tabelle 1 vergleicht die Ausprägung von Akzentkommados russischer und chinesischer Deutschlerner mit der muttersprachlichen Referenz. Im Sinne einer einfachen, robusten

Tabelle 1 - Mittlere Akzentkommandoamplituden und Kommandodauern aus [2].

Sprechergruppe		A_a	Dauer [ms]
L1 DE (Referenz)	μ	0,29	294
	σ	0,13	154
L2 DE (L1 RU)	μ	0,38	239
	σ	0,20	108
L2 DE (L1 CN)	μ	0,31	243
	σ	0,17	136

f0-Analyse greifen daher einige Bewertungsansätze auf Fehlermaße wie root mean square error (RMSE) zurück. Dieser Ansatz kann mit dem klassischen Lernprinzip des Nachsprechens assoziiert werden: Bei identischer Äußerung wird die vom Schüler erzeugte f0-Kontur mit der Referenzkontur des Lehrers verglichen, wobei vor der Berechnung des mittleren f0-Fehlers noch eine Zeitanpassung mittels dynamic time warping (DTW) durchgeführt wird. In [3] beträgt der RMSE für fünf männliche L2-Sprecher in einem baskischen Testkorpus im Mittel $17,1$ Hz – verglichen mit 14 Hz für einen männlichen L1-Referenzsprecher.

Die Nachteile der bisher beschriebenen f0-Evaluierungsansätze betreffen u. a.:

- die Überlagerung mit anderen Kriterien z. B. im Perzeptionstest,
- eine komplexe, fehleranfällige Berechnung der Modellparameter,
- die fehlende Funktion-Form-Relation.

Darüber hinaus wird die Evaluierung durch die perzeptive Toleranz bestimmter f0-Abweichungen beeinflusst – vgl. lineare Stilisierung im Intonationsmodell von Adriaens [4].
Bei kritischer Betrachtung der Vorarbeiten stellt sich die Frage, ob sich prinzipielle Intonationskomponenten finden lassen, die gewichtet zu der gesuchten Bewertungsfunktion (exemplarisch in unserer Sprachstufenklassifikation) beitragen. Diese Analyse kann quasi *rein mathematisch* und unabhängig vom konkreten Fachkontext als f0-Hauptkomponentenanalyse (principal component analysis, PCA) erfolgen. In anderen akustischen Anwendungsbereichen, z. B. bei der Schallquellenortung, liefert die PCA gute Ergebnisse.

3 Hauptkomponentenanalyse (PCA) für funktionale Daten

Die Hauptkomponentenanalyse identifiziert Korrelationen zwischen einzelnen Funktionen und überträgt deren Informationsgehalt auf eine geringere Anzahl komplexerer Funktionen, die sogenannten Hauptkomponenten (principal components, PC), wobei ein minimaler Informationsverlust angestrebt wird [5]. Die funktionalen Daten werden im Anschluss mit Hilfe der Hauptkomponentenpunkte kategorisiert.

3.1 Sprachdatenaufbereitung

Abbildung 1 veranschaulicht den Prozessablauf. Die Datenaufbereitung beinhaltet eine Zeitanpassung mittels DTW und die Erzeugung der f0-Kontur [6] sowie eine Interpolation und Glättung.

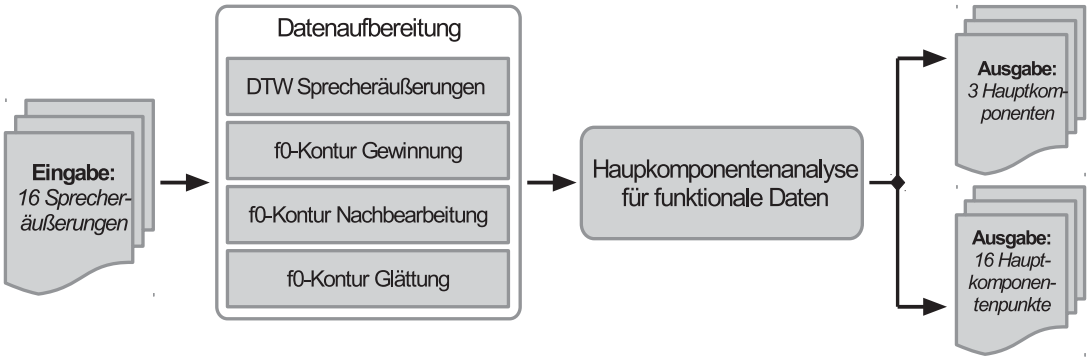

Abbildung 1 - PCA-Methode; Prozessablauf und Datenfluss zum Fallbeispiel in Abschnitt 4.

3.2 Methode

Die Suche von Hauptkomponenten für n funktionale Daten ermittelt die Eigenfunktionen ξ und deren Eigenwerte μ mit dem Ziel einer größtmöglichen Abweichung [7]. Die Hauptkomponentenanzahl p wird durch das Kaiser-Kriterium $p = max\{j|\mu_j > 1\}$ festgelegt [5].
Zunächst wird die Kovarianzfunktion

$$v(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(s) - \bar{x}_i(s)] \cdot [x_i(t) - \bar{x}_i(t)] \tag{1}$$

bestimmt. Die Gleichung

$$\mu = max_{\xi}\left\{\sum_{i=1}^{n}[\int \xi(t)x_i(t)\,dt]^2\right\} \tag{2}$$

definiert die Maximalwerte für μ. Aus den Gleichungen 1 und 2 wird ein Eigenwertproblem formuliert [8]:

$$\int v(s,t)\xi_j(t)\,dt = \mu_j\xi_j(s). \tag{3}$$

Das Eigenwertproblem wird in der linearen Algebra mit $\mathbf{V}\vec{\xi} = \mu\vec{\xi}$ beschrieben. Nach Umformung kann $(\mathbf{V} - \mu \cdot \mathbf{E}) \cdot \vec{\xi} = \Theta$ geschrieben werden. Alle Eigenfunktionen ξ und Eigenwerte μ werden dann unter der Bedingung $det(\mathbf{V} - \mu \cdot \mathbf{E}) \neq 0$ berechnet [9]. Daraus lassen sich gemäß [10] die Hauptkomponenten

$$PC_j(t) = \overline{x}(t) + \sqrt{\mu_j}\xi_j(t) \tag{4}$$

sowie gemäß [8] die Hauptkomponentenpunkte erzeugen

$$\mathbf{C}_{scr}(i,j) = \sum_{j=1}^{p}\sum_{i=1}^{n}\int \xi_j(t)[x_i(t) - \overline{x}(t)]\,dt. \tag{5}$$

4 Fallbeispiel zur L1/L2-Unterscheidung und Sprachstufeneinteilung

4.1 Testdaten

Die Teststichprobe wird in Tabelle 2 dargestellt und beinhaltet jeweils eine Äußerung pro Sprecher: *Nein sie kann es nicht* von acht deutschen Muttersprachlern (L1 DE) sowie acht Deutschlernern (L2 DE) mit der Muttersprache Russisch aus dem Euronounce/Veith-Korpus [11].

Tabelle 2 - Testdatensatz für die f0-Analyse und -Klassifikation.

Gruppe	Sprecherbeschreibung	Sprecheranzahl
$L1m_x$	L1 DE, männlich	4
$L1f_x$	L1 DE, weiblich	4
$L2m_x$	L2 DE (L1 RU), männlich	4
$L2f_x$	L2 DE (L1 RU), weiblich	4

4.2 RMSE-Analyse

Zum Vergleich mit der PCA-basierten Klassifikation analysieren wir den RMSE für alle L1/L1- und L1/L2-Kombinationen (realisierte f0-Kontur versus Referenz) entsprechend der Methode aus [3]. Die relativen Abweichungen (in %) werden in den Matrizen der Abbildung 2 zusammengefasst, wobei die f0-Werte der weiblichen und männlichen Sprecher auf ihre mittlere Grundfrequenz f0 normiert vorliegen. Bester Referenzsprecher bezüglich der RMSE-Analyse ist $L1m_2$ (kleinster Mittelwert von 13 %, Varianz von 64 %). Die mittleren Abweichungen betragen 15 % (L1/L1) bzw. 22 % (L1/L2).

4.3 PCA-basierte Klassifikation

Für die Klassifikation werden die Hauptkomponentenpunkte ermittelt und in Abbildung 3 für die drei Hauptkomponenten und Dimensionen dargestellt.

	$L1m_1$	$L1m_2$	$L1m_3$	$L1m_4$	$L1w_1$	$L1w_2$	$L1w_3$	$L1w_4$
$L1m_1$	0	20	23	29	50	5	4	5
$L1m_2$	17	0	2	7	25	13	14	13
$L1m_3$	19	2	0	4	22	15	16	15
$L1m_4$	22	6	4	0	17	18	19	19
$L1w_1$	31	18	16	13	0	28	28	28
$L1w_2$	4	13	16	20	38	0	1	0
$L1w_3$	3	14	16	21	39	1	0	0
$L1w_4$	4	13	16	20	38	0	0	0

	$L1m_1$	$L1m_2$	$L1m_3$	$L1m_4$	$L1w_1$	$L1w_2$	$L1w_3$	$L1w_4$
$L2m_1$	1	22	25	30	66	21	20	20
$L2m_2$	18	43	46	52	95	41	40	41
$L2m_3$	35	22	20	16	7	23	23	23
$L2m_4$	28	13	11	7	19	14	14	14
$L2w_1$	25	12	10	6	17	13	13	13
$L2w_2$	8	8	11	15	44	7	7	7
$L2w_3$	26	12	10	7	16	13	14	13
$L2w_4$	4	22	25	30	62	21	20	21

Abbildung 2 - Relative f0-Abweichungen (in %) – L1 vs. L1 (linke Matrix) und L1 vs. L2 (rechts).

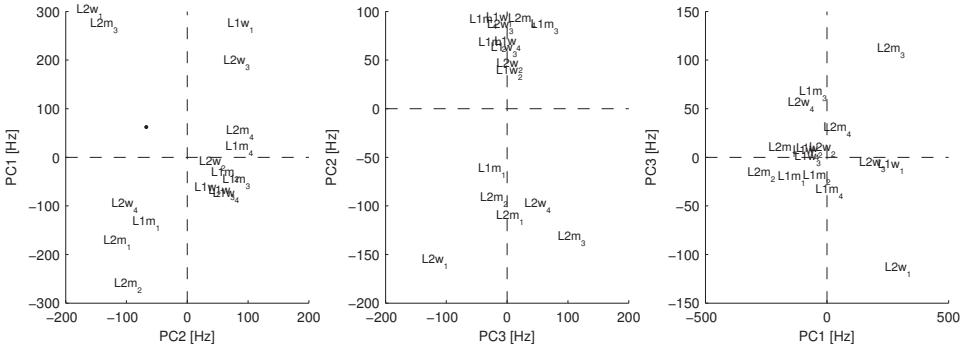

Abbildung 3 - Sprecher-Komponenten: PC1 vs. PC2 (links), PC2 vs. PC3 (Mitte), PC3 vs. PC1 (rechts).

4.3.1 L1/L2-Klassifikation

Über die Mittelpunkte m_{group} der vier Sprechergruppen $|g|$, mit $g = \{1,2,3,4\}$, wird die geometrische Distanz, mit dem Informationsgehalt der Eigenwerte $fac(j) = \frac{\mu(j)}{\sum_{i=1}^{p} \mu_i}$ gewichtet,

$$\vec{d}_{mdl}(i,g) = \sqrt{\sum_{j=1}^{p} fac_j \cdot (\mathbf{C}_{scr}(i,j) - m_{group}(g))^2} \tag{6}$$

gebildet. Anschließend werden die Standardabweichungen $\sigma_{group}(g)$ für die vier Punktwolken bestimmt. Der Indexwert $idx_{L1L2}(i)$, der die Klassenzugehörigkeit (L1 oder L2) angibt, wird wie dann folgt berechnet:

$$idx_{L1L2}(i) = \left| \sum_{g=1}^{|g|} \frac{\vec{d}_{mdl}(i,g)}{\sigma_{group}(g)} \cdot (-1)^{g-1} \right| \tag{7}$$

Der Wert wird abschließend auf die Skala 0..1 normiert, wobei 0,5 als Entscheidungsschwelle fungiert (L1: $idx_{L1L2} > 0,5$ bzw. L2: $idx_{L1L2} \leq 0,5$).

4.3.2 Sprachstufenklassifikation

Zunächst unterstellen wir bei allen $a = 8$ Muttersprachlern ein sehr gutes Sprachniveau (Stufe 1). Somit kann ein Sprachstufenreferenzwert ref_{level}, aus den geometrischen Abständen

$$\vec{d}_{geo}(i,ii) = \sqrt{\sum_{j=1}^{p} fac(j) \cdot (\mathbf{C}_{scr}(i,j) - \mathbf{C}_{scr}(ii,j))^2} \tag{8}$$

aller Hauptkomponentenpunkte der L1-Sprecher untereinander (mit $fac(j)$ gewichtet) ermittelt werden. Diese Abstände bilden gemeinsam eine Entfernungsmatrix E, deren Frobeniusnorm den Sprachstufenreferenzwert wiedergibt. Es gilt $ref_{level} = ||E(d_{geo})||_F$.
Die Sprachstufe spk_{level} eines Sprechers wird mittels eines Vergleichswerts cmp_{level} bestimmt. Dieser Wert wird aus der mittleren geometrischen (mit $fac(j)$ gewichteten) Abweichung eines Hauptkomponentenpunktes $C_{scr}(i,j)$ zu allen a Referenzpunkten (L1) berechnet:

$$cmp_{level}(i) = \sqrt{\frac{1}{a}\sum_{ii=1}^{a} d_{geo}(i,ii)^2}. \tag{9}$$

Der Sprachstufenwert des i-ten Sprechers ergibt sich dann wie folgt: $spk_{level}(i) = \frac{cmp_{level}(i)}{ref_{level}}$, wobei die Sprachstufe ebenfalls normiert wird – auf eine Skala 1 (sehr gut) .. 6 (schlecht).

4.4 Hörtest

Im Hörtest wurden fünf weibliche und zehn männliche deutsche Muttersprachler, Altersdurchschnitt 26,3 Jahre (Alterspanne von 20 bis 53), ohne Fachexpertise im Untersuchungsgebiet befragt. Die Probanden schätzten im Blindtest alle 16 Sprecher nach folgenden Kriterien ein:

- Geschlecht (männlich/weiblich),
- Deutsch als Muttersprache (ja/nein),
- Sprachstufe des Sprechers – sehr gut (1) bis schlecht (6).

Die maximale Einzelabweichung von Probanden-Einschätzungen beträgt drei Sprachstufen. Die über alle Sprecher gemittelte Standardabweichung der Einschätzung liegt bei 0,16 Stufen.

4.5 Ergebnisse

4.5.1 L1/L2-Unterscheidung

Abbildung 4 stellt die L1/L2-Sprecherzuordnung der verschiedenen Methoden bei gleicher Skalierung dar. Eine RMSE-basierte Klassifikation (dunkelgrau hinterlegt) mittels Referenzsprecher $L1m_2$ führt zu sechs korrekten und zwei undefinierten Entscheidungen (in etwa statistischer Erwartungswert). Die PCA-Methode (grau) weist hingegen 12 und der Hörtest (weiß) 14 korrekte Zuordnungen auf.

4.5.2 Zuordnung der Sprachstufe

Abbildung 5 zeigt die Sprachstufenzuordnung. Als Referenzwert dient die Einordnung eines professionellen Sprachlehrers (schwarz hinterlegt). Die PCA-Methode (grau) ordnet die Sprecher in neun Fällen ähnlich der Sprachlehrer-Einschätzung zu, weist aber fünf Abweichungen um bis zu zwei Stufen und zwei größere Abweichungen auf. Der Hörtest (weiß) führt zu 13 weitgehend korrekten Entscheidungen.

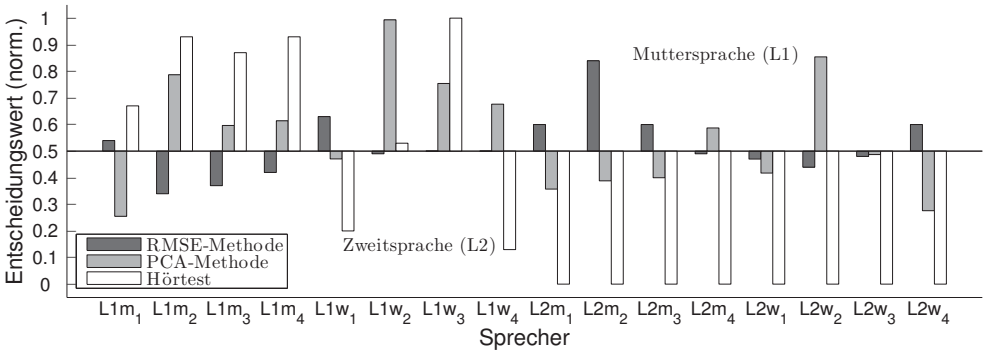

Abbildung 4 - Zuordnung: L1/L2-Sprecher des Deutschen.

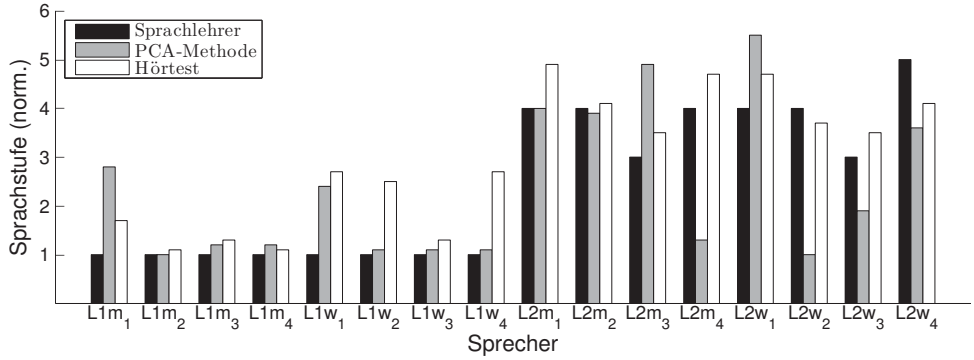

Abbildung 5 - Zuordnung der Sprachstufe (Deutsch).

5 Diskussion und Zusammenfassung

Die im zweiten Abschnitt diskutierten f0-Evaluierungsansätze sind im Kontext von Sprachlernsystemen eher unbefriedigend. Die im dritten Abschnitt vorgestellte und mathematisch motivierte Hauptkomponentenanalyse (PCA) funktionaler Daten in Anlehnung an Ramsay [7] und Sprachanalysen von Gubian (z. B. [10]) lässt sich hingegen gut auf unsere Testdaten anwenden.

Im dargestellten Fallbeispiel für 16 Sprecher wird nur die f0-Kontur jeweils einer Äußerung analysiert. Trotz der stark beschränkten Information besitzt die PCA-basierte Methode eine Performanz deutlich über einer konventionellen Abstandsanalyse mittels RMSE und in etwa in der Größenordnung des Hörtests, bei dem die Probanden auf zusätzliche – z. B. phonetische – Merkmale zurückgreifen können.

Bezüglich der L1/L2-Zuordnung beträgt die Trefferquote der PCA-Methode 75 % (versus RMSE-Analyse mit 43 % auf Zufallsniveau und dem Hörtest von 87 %).

Eine Sprachstufenzuordnung auf Basis der f0-Hauptkomponenten erfordert deutliche Verbesserungen. Der Hörtest zeigt hierbei ebenfalls Abweichungen zur Stufeneinordung des Sprachlehrers (auf Basis längerer Beobachtungen), was wiederholt auf die beschränkte Testdatenlage für die PCA-Methode sowie den Hörtest hinweist.

Dennoch liegen die Ergebnisse der kleinen Pilotstudie oberhalb unserer Erwartungen und lassen hoffen, dass die PCA-basierte Methode für weitere prosdische Analysen im Sprachlernkontext geeignet ist. Die vorgestellte Methode soll anhand zusätzlicher Evaluierungsmerkmale und mit einem größeren Datenumfang weiterentwickelt werden. Dabei sind bessere Klassifikatoren und eine Einteilung in Trainings-, Test- sowie Validierungsdaten notwendig.

6 Danksagung

Die L1/L2-Testdaten stammen aus dem Euronounce/Veith-Korpus und wurden von Rainer Jäckel, TU Dresden, aufgezeichnet. Wir möchten außerdem Michael Graf und Ines Rennert, Hochschule für Telekommunikation Leipzig, für ihre wertvollen Hinweise danken.

Literatur

[1] Hönig, F., Bocklet, T., Riedhammer, K., Batliner, A. and Nöth, E., "The automatic assessment of non-native Prosody: combining classical prosodic analysis with acoustic modelling," in proc. Interspeech, Portland, Oregon, 2011.

[2] Hilbert, A., Mixdorff, H., Ding, H., Pfitzinger, H. and Jokisch, O., "Prosodic analysis of German produced by Russian and Chinese learners," in Proc. 5th Intern. Conference on Speech Prosody, Chicago, Illinois, 2010.

[3] Odriozola, I., Jokisch, O., Hernaez, I. and Hoffmann, R., "A Pronunciation Tutoring System for Basque - First Development Steps," in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Cottbus, 2012.

[4] Adriaens, L. M. H., "Ein Modell deutscher Intonation", Dissertation, TU Eindhoven, 1991.

[5] Elle, O., "Einführung in die multivariante Statistik für Feldornithologen: Haupkomponentenanalyse, Diskriminanzanalyse und Clusteranalyse," Vogelwarte, Bd. 43(1), 19–38, 2005.

[6] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer" (version 5.3.05). Retrieved February 24, 2012 from https://www.praat.org.

[7] Ramsay, J. O. and Silverman, B. W., "Functional Data Analysis", Springer, New York, 1997.

[8] Ramsay, J. O., Hooker, G. and Graves, S., "Functional Data Analysis with R and MATLAB", Springer, 100–103, New York, 2009.

[9] Leupold, W., "Mathematik - ein Studienbuch für Ingenieure", Fachbuchverlag Leipzig, Bd. 1, 2. Auflage, Leipzig, 2004.

[10] Gubian, M., Boves, L. and Cangemi, F., "Joint analysis of f_0 and speech rate with Functional Data Analysis," in proc. ICASSP, 4972–4975, Florence, 2011.

[11] Jokisch, O., Jäckel, R., Rusko, M., Demenko, G., Cylwik, N., Ronzhin, A., Hirschfeld, D., Koloska, U., Hanisch, L., Hoffmann, R., "The EURONOUNCE project - An intelligent language tutoring system with multimodal feedback functions: Roadmap and specification," in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 116–123, Frankfurt, 2008.