Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 38 (2001) 117–264

Phonetische Analyse der Sprechgeschwindigkeit

Hartmut R. Pfitzinger

Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstr. 3
D-80799 München

Abstract

Title: Phonetic Analysis of Speech Rate

A model for deriving perceptual local speech rate directly from the speech signal is developed based on experimental analysis of the relationship between speech acoustics and perception. Since local speech rate modifies acoustic cues (e.g. transitions and voice-onset time), phones, syllables, and even words, it is one of the most important components of prosody.

Our local speech rate estimation method is based on a linear combination of the local syllable rate and the local phone rate, since earlier investigations strongly suggest that neither the syllable rate nor the phone rate on its own represent the speech rate sufficiently.

To test this hypothesis training and evaluation data was needed. This we obtained by conducting two interactive perception experiments. In the first experiment 60 subjects were instructed to place 141 short speech stimuli along an interval scale, and were also asked to ensure that perceptual speech rate differences corresponded to distances on the interval scale. In the second experiment 30 subjects judged 100 stimuli. Additional perception experiments supported the reliability and validity of our method and data. The results enable us to calculate the linear correlation coefficients of local syllable rate or local phone rate with the mean perception results. Additionally we are able to approximate a linear combination of syllable rate and phone rate, thus obtaining a model for deriving perceptual local speech rate directly from the speech signal with sufficient accuracy for future phonetic research. In the literature, effects of F_0 level and F_0 movement on speech rate perception have been reported. Therefore we also included these cues.

Our results show 1) that the duration of speech stimuli has a strong influence on the perception of speech rate (a duration of 425 ms and less causes perceptual overshoot), 2) that the linear combination of local syllable rate and phone rate is well-correlated with perceptual local speech rate (r=0.912), 3) that the incorporation of F_0 measurements does not increase the accuracy of the model, and 4) that the perceptual local speech rate (PLSR) is a meaningful prosodic contour which now can be extracted from the speech signal with sufficient accuracy.

Diese Arbeit ist die überarbeitete Fassung meiner Dissertation, eingereicht im März 2001 bei der Philosophischen Fakultät für Sprach- und Literaturwissenschaften II der Ludwig-Maximilians-Universität München für das Fachgebiet „Phonetik und Sprachliche Kommunikation“.

Seit 2008 ist die gedruckte Version vergriffen und wird auch nicht mehr aufgelegt. Daher ist die vorliegende Version eine nahezu unveränderter Nachdruck, der auch online als pdf-Dokument zu finden ist unter: `http://www.ipds.uni-kiel.de/hpt/pub/f38_hp_1.pdf`

„[. . .] do not allow for deciding
whether the rate of phones or the rate of syllables
correspond to the speech rate. It is rather likely
that speech rate is a combination of both.“

1996 [169, S.426]

Vorwort

Als ich in einem Beitrag zur *International Conference on Speech Science and Technology (SST ’96)* in Adelaide mit oben zitierter These meine Absicht begründete, Perzeptionsexperimente zur Sprechgeschwindigkeit durchführen zu wollen, um deren Ergebnisse mit Phon- und Silbenraten zu korrelieren, war dies damals wirklich nur als ein „kurzes“ Intermezzo in einer Folge von Untersuchungen zur zeitlichen und spektralen Strukturierung lautsprachlicher Äußerungen geplant. Erst sehr viel später zeigte sich, daß die Ergebnisse so neu und zugleich vielversprechend waren, daß sie als Kern einer Dissertation bereits tragfähig waren. Dennoch wäre diese Promotionsschrift ohne die inspirierenden und motivierenden Gespräche mit Prof. Hans G. Tillmann nie begonnen worden, und sie wäre in absehbarer Zeit ohne seine ebenso konkreten Kürzungsvorschläge nicht beendet worden. Für die Förderung in den vergangenen Jahren danke ich ihm herzlich. Auch allen anderen, die sich in kritischer Weise mit Teilen der Arbeit auseinandergesetzt und mich zu vielen Verbesserungen veranlaßt haben, sei Dank gesagt: Olga Dioubina, Sebastian Heid, Phil Hoole, Michiko Inoue, Irene Jacobi, Christian Kroos, Kirsten Machelett und Uwe Reichel.

Inhalt

1

Einleitung

Im Mittelpunkt dieser Untersuchung steht die Entwicklung eines einfachen linearen Modells zur akkuraten Vorhersage der perzipierten lokalen Sprechgeschwindigkeit im Deutschen auf der Basis von automatisch extrahierbaren akustischen Merkmalen.

1.1 Übersicht

Ausgehend von einer knappen Zusammenfassung der einschlägigen Literatur werden zunächst zentrale Begriffe der Sprechgeschwindigkeit und zeitlichen Struktur von gesprochener Sprache geklärt. Daran anschließend liefern sechs Perzeptionsexperimente die notwendigen Fakten, um ein Sprechgeschwindigkeitsmodell entwickeln und testen zu können, das auf Markierungen der Mitten von Silbenkernen und Phongrenzensegmentationen aufbaut. Zusätzlich wird untersucht, inwiefern die Berücksichtigung der Grundfrequenz als ein weiterer akustischer Parameter zur Verbesserung des Modells beiträgt.

Abschließend zeigen ausgewählte Anwendungen in der Sprachanalyse und -synthese, daß die hier entwickelte Methode zur Messung der lokalen Sprechgeschwindigkeit nicht nur in der Praxis sehr hilfreich ist, sondern auch ganz neue Lösungswege für seit langem unbeantwortbare Fragen der phonetischen und auch phonologischen Forschung eröffnet.

Exemplarisch seien hier einerseits das Problem der Invarianz und Variabilität bei der zeitlichen Organisation von Sprachsegmenten und andererseits die Schwierigkeit eines signalphonetischen Nachweises von Sprechrhythmus und Isochronie genannt.

1.2 Zum fehlenden Konsens bei der Definition der Sprechgeschwindigkeit

In der bisherigen Forschung gehen die Ansichten darüber auseinander, was unter dem Begriff der *Sprech*geschwindigkeit zu verstehen sei. Die Geschwindigkeit eines Kraftfahrzeugs besagt, wieviele Kilometer pro Stunde zurückgelegt werden. Welche Ereignisse pro Zeiteinheit werden aber bei der Sprechgeschwindigkeit gemessen?

Zweifellos werden hohe Sprechgeschwindigkeiten vorwiegend von überdurchschnittlichen Wort-, Silben- und Phonraten begleitet und niedrige Sprechgeschwindigkeiten eher von unterdurchschnittlichen Raten. Diese oft beobachtete Kovariation führte in der phonetischen Forschung zu der weit verbreiteten Praxis, Sprechgeschwindigkeit durch eine beliebige dieser drei Raten zu kennzeichnen.

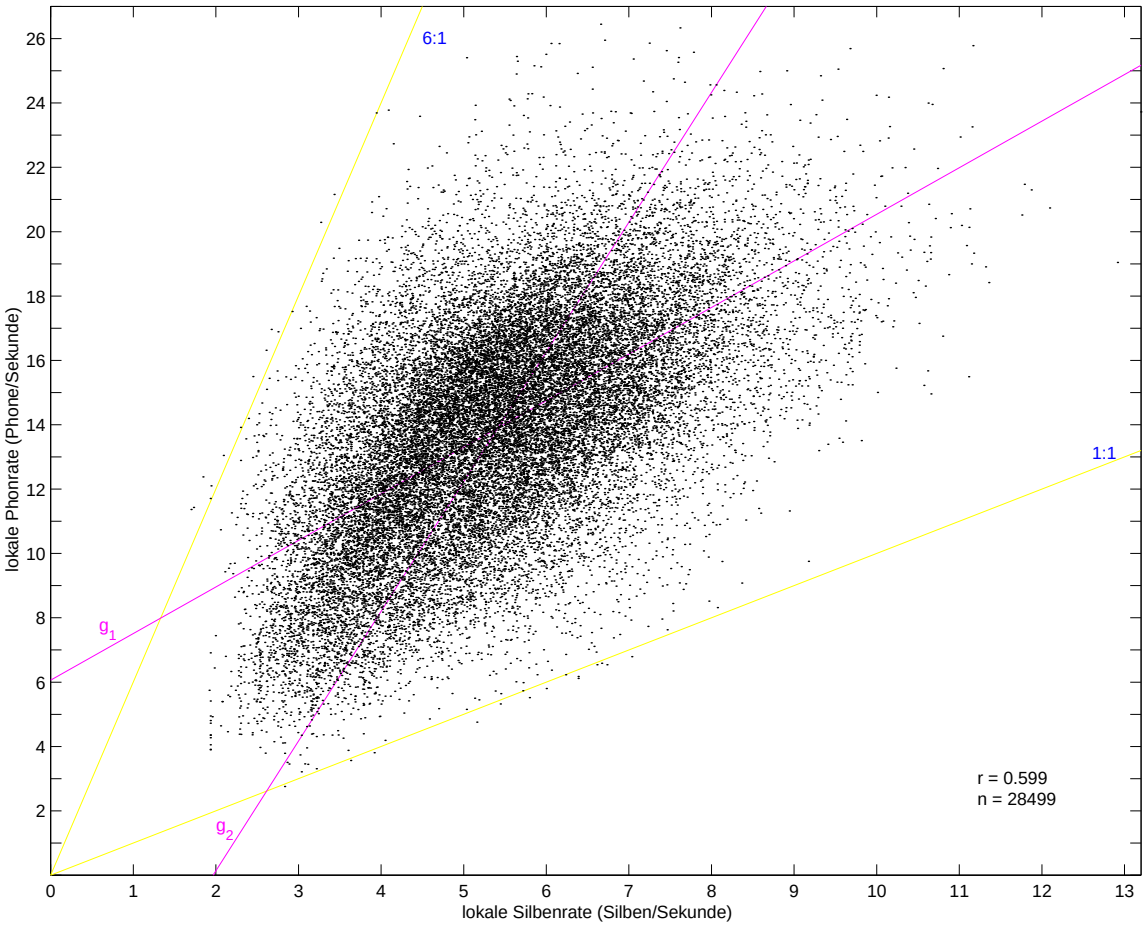

Abb. 1.1: Jeder der 28499 Punkte in diesem Streuungsdiagramm repräsentiert die Silben- und Phonrate eines 625 ms dauernden Äußerungsteils, die in 100 ms Schritten aus handsegmentierten Mitten von Silbenkernen bzw. Phongrenzen gewonnen wurden. Das zugrundeliegende Korpus umfaßt 16 Sprecher mit jeweils durchschnittlich 178.1 Sekunden Lesesprache (1024 der 3200 Sätze des *PhonDatII*-Korpus).

Im Deutschen korrelieren Silben- und Phonrate aber nur mäßig, wenn sie *lokal*, d.h. über vergleichsweise kurze Äußerungsausschitte von etwa 500 ms bis 800 ms Dauer ermittelt werden ($r \approx 0.6$ bei 625 ms Dauer, siehe Abb. 1.1). Sogar bei drei Sekunden langen Ausschnitten, die wohl nicht mehr lokal genannt werden können, bleibt der Korrelationskoeffizient unter 0.74 (Pfitzinger 1998 [170]).

Daraus folgt, daß beide Maße unterschiedliche Informationen widerspiegeln. Das ist nicht weiter verwunderlich: Jedes Wort setzt sich in anderer Weise aus Silben zusammen, deren Phonstrukturen wiederum unterschiedlich komplex sind. Dadurch ergibt sich ein von Wort zu Wort variierendes Verhältnis zwischen Silben- und Phonrate, wie folgende Wortbeispiele demonstrieren: Während das Wort *Banane* drei Silben und sechs Phone — also eine doppelt so hohe Phonrate — aufweist, hat das Wort *schimpfst* nur eine Silbe, aber etwa sieben Phone[1] und damit eine siebenmal höhere Phonrate.

Zudem weist das Verhältnis zwischen Silben- und Phonrate eine hohe Lokalität auf; seine Variationen sind bereits auf Silbenebene feststellbar und damit auch *innerhalb* von Wörtern, wie

[1] Affrikaten werden in dieser Untersuchung als zwei Phone gezählt. Wir reden in dieser Arbeit grundsätzlich von *Phonen* statt *Phonemen*, da die Sprachlaute der zugrundeliegenden Korpora aus guten Gründen abweichend vom Phoneminventar des Deutschen etikettiert sind (siehe hierzu Kap. 5.4 auf S. 166).

die beiden folgenden zweisilbigen Wörter zeigen: Beim Wort *Strafe* halbiert sich das Verhältnis von der ersten zur zweiten Silbe, während es sich beim Wort *Ankunft* mehr als verdoppelt.

Schon diese wenigen Wortbeispiele legen nahe, daß weder die Silben- noch die Phonrate als Sprechgeschwindigkeit bezeichnet werden sollte, solange keine einhellige Meinung darüber herrscht, wie Sprechgeschwindigkeit definiert werden soll.

1.3 Die Problematik der zeitlichen Struktur von Sprache

Gesprochene Sprache kann als ein akustischer Vorgang, der durch eine permanente spektrale Variation in der Zeit gekennzeichnet ist, aufgefaßt werden. Dabei folgt die Variation einer Vielzahl von Gesetzen, die den Sprachschall erst von jedem anderen zeitvariablen Schall differenzieren. Damit ein Schall sprachähnlich und verständlich wird, muß er weder von einem Mensch produziert werden, noch eine variierende Intonation aufweisen (siehe *Vocoder*-Sprache). Aber er muß die Gesetze der zeitlichen Strukturierung gesprochener Sprache einhalten: Zwar ist Sprache, die künstlich bis zu dreifach beschleunigt oder bis zu etwa zehnfach verlangsamt wurde, gerade noch verständlich, aber wenn die inneren zeitlichen Relationen verzerrt werden (etwa dadurch, daß einige Silben zu lang und gleichzeitig benachbarte Silben zu kurz werden), so ist das Ergebnis unverständliche Sprache, ohne daß sich dabei die Gesamtdauer der Äußerung geändert hätte.

Lokale Sprechgeschwindigkeit und zeitliche Struktur gesprochener Sprache stehen in so enger Beziehung zueinander, daß wir das eine nicht ohne das andere betrachten sollten. Demzufolge geht der theoretische Teil dieser Arbeit ausführlich auf beides ein. Hier sind zusätzlich zu den rein phonetischen Herangehensweisen insbesondere Ansätze aus der Phonologie, Psychologie, Psycholinguistik, Psychoakustik und Elektrotechnik nennenswert.

Wir werden mit unserem Überblick bei den sog. *phonetischen Quantitätsgesetzen* aus der Lautdauerforschung beginnen und die Bedeutung der Sprechgeschwindigkeit in den verschiedenen Forschungsrichtungen beleuchten. Dann werden wir auf die zeitliche Struktur von Sprache eingehen und anhand der jeweils einschlägigen Literatur die folgenden Feststellungen belegen:

- Die sog. *P-center*-Forschung hat zwar Modelle zur Bestimmung des Ereigniszeitpunkts einer Silbe geliefert und damit neue Möglichkeiten eröffnet. Aber weder konnte sie dazu beitragen, die Isochronie-Hypothese zu widerlegen oder zu bestätigen, noch verbesserte sie die Meßbarkeit von Sprechgeschwindigkeit, und ließ damit zwei wichtige Fragen weiterhin offen (siehe Kap. 3.1).
- Die gestaltpsychologische Betrachtung des Sprechrhythmus kann die perzipierte Struktur anhand der Gestaltgesetze im Nachhinein zwar besser erklären als rein physikalische Ansätze, aber ebenfalls nicht vorhersagen, da funktionierende Modelle fehlen (siehe Kap. 3.2).
- Der linguistische Ansatz, der sich in der Metrischen Phonologie manifestiert, betrachtet nur prototypische zeitliche Strukturen in Form von qualitativen und relativen Dauerinformationen, läßt aber die tatsächlich produzierten Zeitverhältnisse außer Acht (siehe Kap. 3.5).
- Die psychologische Erforschung der Zeitverarbeitung und ihre psycholinguistische Anwendung liefert immerhin die Antwort auf die Frage, warum künstlich zu stark verlangsamte bzw. beschleunigte Sprache nicht mehr verständlich ist: Die gedehnten sprachbildenden Einheiten überschreiten die *Präsenzzeit* und werden dann nicht mehr als zusammengehörig wahrgenommen. Bei zu starker Beschleunigung unterschreiten die Segmente die *Ordnungsschwelle*: Die Reihenfolge ist dann nicht mehr wahrnehmbar (siehe Kap. 3.7).

Keiner der bisherigen Ansätze bringt uns den entscheidenden Schritt weiter. Darüber hinaus müssen wir feststellen, daß auch artikulatorische Messungen als alleinige Referenz für Aussagen zum Sprechtempo oder zum Sprechrhythmus ungeeignet sind: Durch einerseits komplexe Koartikulationsphänomene und andererseits individuell unterschiedliche Produktionsstrategien sind direkte Beziehungen zwischen artikulatorischen Meßdaten und der lokalen Sprechgeschwindigkeit oder dem Sprechrhythmus ohne weitere Informationen nur äußerst schwierig aufzudecken.

In diesem Dilemma könnte der Ausweg, um die zeitliche Struktur gesprochener Sprache einer genauen empirischen Analyse zugänglich machen zu können, darin liegen, *zuerst* deren lokale Geschwindigkeit zu ermitteln. Aber auf die Frage nach einem Meßwert der Sprechgeschwindigkeit ist die bislang einzige Antwort immer noch die Anzahl sprachlicher Elemente pro Zeiteinheit, die wir jedoch im vorigen Kapitel als widersprüchlich und damit unbefriedigend entlarven mußten.

1.4 Der psychophysikalische Zugang

Silben- und Phonrate können heute mit automatischen Verfahren für phonetische Zwecke hinreichend genau gemessen werden (Pfitzinger 1996 [169]). Ein gewisser Fehleranteil ist allerdings momentan noch unvermeidbar, da kein Verfahren alle im Sprachsignal verborgenen Hinweise auf Silben und Phone, die der Mensch wahrnimmt, auswerten kann. Allein der Mensch verfügt über die Fähigkeit, Sprache zu *verstehen*, und ist damit in der Lage, die kognitiven *top-down*–Prozesse seiner Sprachwahrnehmung auch dafür zu nutzen, Silben und Phone richtig zu erkennen und die Fehler der Algorithmen aufzudecken.[2] Demnach gibt es bei der Silben- und Phonmarkierung nur eine sichere Referenz: die Perzeption des Menschen.[3]

Ausgehend von diesem Standpunkt wird es unmittelbar notwendig, anhand von Perzeptionsexperimenten zu untersuchen, ob der Mensch auch Sprechgeschwindigkeiten einschätzen kann und wie präzise und konsistent er dies vollführt. Daß dabei eine Rolle spielt, ob man Äußerungen in der eigenen Muttersprache oder einer unbekannten Sprache zu beurteilen hat, ist anzunehmen, wenn *top-down*–Sprachwahrnehmungsprozesse beteiligt sind. Die Sprechgeschwindigkeit wäre also abhängig von einer Interpretation der zugrundeliegenden lautsprachlichen Äußerung und nicht *per se* gegeben. Messungen der Anzahl von Silben oder Phonen liefern demnach eben nur *Silben-* oder *Phonraten*, nicht aber zwingend das, was allgemein mit *Sprechgeschwindigkeit* umschrieben und gemeint wird.

Wenn bei den angestrebten Perzeptionsexperimenten reliable und valide Ergebnisse entstehen, die Versuchspersonen also untereinander übereinstimmen und auch bei Wiederholungsexperimenten zu den gleichen Urteilen kommen, so muß dem eine psychophysikalische Relation zugrundeliegen, die die physikalische Realität der akustischen Sprachstimuli — und damit auch deren meßtechnische Repräsentationen — mit Beurteilungswerten für die perzipierte Sprechgeschwindigkeit in Beziehung bringt. Erst auf dieser Grundlage kann ein Modell entwickelt werden, das diese Beziehung möglichst adäquat beschreibt. Und erst dann kann auch gesagt werden, was genau Sprechgeschwindigkeit ist.

---

[2] Siehe zur automatischen Extraktion von Silbenkernen beispielsweise Pfitzinger, Burger & Heid 1996 [172] und von Phongrenzen etwa Verhasselt & Martens 1996 [228].

[3] Hier sei auf S. 140 in Kap. 2.5 und auf Kap. 4.5 ab S. 157 verwiesen, wo auf den grundsätzlich vorhandenen Unterschied zwischen den kanonisch gegebenen und den bei einer lautsprachlichen Äußerung tatsächlich realisierten Sprachsegmenten eingegangen wird.

ERSTER TEIL

DIE THEORETISCHEN GRUNDLAGEN

2

Forschungsüberblick zur Sprechgeschwindigkeit

Die wissenschaftliche Bedeutung der Sprechgeschwindigkeit hat im Laufe des 20. Jahrhunderts — in jüngerer Zeit auch aufgrund der Fortschritte bei der automatischen Spracherkennung — einen Wandel erlebt von einem (unter vielen) bei phonetischen Untersuchungen kontrollierbaren Einflußfaktor hin zu einem eigenständigen Forschungsgegenstand. Dennoch wird auch in den gegenwärtigen Publikationen nicht der Versuch unternommen, den Sprechgeschwindigkeitsbegriff sauber zu definieren, sondern er wird, wie schon in der Vergangenheit, weiterhin intuitiv verwendet. Die Bedeutung des Begriffs wird stillschweigend als selbstverständlich vorausgesetzt.

Tatsächlich wird dann aber statt der Sprechgeschwindigkeit typischerweise die Silben- oder Phonrate gemessen, die sich jedoch beide — wie wir im Laufe dieser Arbeit noch feststellen werden — zumindest im Deutschen von der eigentlichen Sprechgeschwindigkeit gravierend unterscheiden.

Man ist sich auch des lokalen Charakters der Sprechgeschwindigkeit als Prosodie nicht bewußt, auch wenn man ihre Bedeutung als einen der ausschlaggebenden Faktoren, welche gemeinsam „phonetische Variation“ hervorrufen, erkannt hat. Daher berücksichtigt man Sprechgeschwindigkeit meistens, indem man sie auf ein globales satzumfassendes Phänomen reduziert, das durch eine einzige Kennzahl für einen Sprecher bzw. für einen größeren Äußerungsabschnitt als hinreichend genau beschrieben angenommen wird.

Symptomatisch für diese Sichtweise ist beispielsweise das Buch *Computing Prosody* (herausgegeben von Sagisaka, Campbell & Higuchi 1997 [192]), in dessen Rahmen der Begriff der Prosodie geradezu synonym für Intonation gebraucht wird, welche dann auch Gegenstand der meisten dort dargestellten Untersuchungen ist. In der Arbeit von van Santen 1997 [223] wird zwar die Problematik der Berechnung von Segmentdauern in der Sprachsynthese thematisiert, jedoch ebenfalls nicht auf Sprechgeschwindigkeit eingegangen. Diese Feststellung trifft auch auf den Beitrag von Kato, Tsuzaki & Sagisaka 1997 [98] zu. Aber immerhin setzen sie sich in einer für unsere eigenen Untersuchungen relevanten Weise mit Lautdauerkompensation und Zeitstrukturwahrnehmung auseinander. Daher werden wir in Kap. 3.7 auf S. 151 auf diese Arbeit zurückkommen.

In den folgenden Abschnitten werden nun vorwiegend diejenigen wissenschaftlichen Veröffentlichungen knapp dargestellt, die über die hier angedeuteten vereinfachenden Ansichten hinausgehen und zu dem heutigen Verständnis des Begriffs der Sprechgeschwindigkeit wesentlich beigetragen haben.

2.1 Lautdauern in der frühen phonetischen Forschung

Bereits wenige Jahre nach der Entwicklung der ersten Verfahren zur Aufzeichnung von Visualisierungen gesprochener Sprache (Marey 1878 [133]), die erst die Voraussetzungen für die Entstehung der Experimentalphonetik schufen, wurden Lautdauern gemessen.[1]

In den frühen Bonner Arbeiten zu Silben- und Lautdauern im Spanischen bzw. Deutschen (Menzerath & de Oleza S. J. 1928 [136] bzw. Weitkus 1931 [234]), die auf dem Kymographen basieren, ist die wachsende wissenschaftliche Bedeutung der Sprechgeschwindigkeit dokumentiert. In der Arbeit von Menzerath & de Oleza S. J. wird sie nur beiläufig erwähnt [136, S.3]. Im Mittelpunkt ihrer Untersuchung stehen vielmehr die von ihnen formulierten „allgemeinen Quantitätsgesetze des Lautes, der Silbe und des Wortes“:

- „Die Durchschnittsdauer des Lautes im Wort wird kleiner, wenn die Lautzahl des Wortes steigt.“ [136, S.68]
- „Die Durchschnittsdauer der Silbe [...] nimmt ab, wenn die Silbenzahl des Wortes zunimmt.“[2] [136, S.71]
- „Die Wortdurchschnittsdauer nimmt zu mit steigender Lautzahl des Wortes.“ [136, S.73]
- „Die Wortdurchschnittsdauer nimmt zu mit steigender Silbenzahl der Wörter.“ [136, S.74]

Sie bezeichnen ihr Hauptergebnis, *„[...] daß allgemein ein Laut umso kürzer ist, je länger das Wort ist oder je mehr Silben das Wort hat, zu dem er gehört“*, als „erstes allgemeines phonetisches Quantitätsgesetz“ [136, S.91]. Dabei sind sie sich der Problematik des Begriffs *Gesetz* bewußt, empfanden diesen dennoch als angemessen.

Bemerkenswert ist noch, daß dieses Gesetz für mittlere Lautdauern nicht nur auf Wortebene gültig ist, sondern auch auf Silbenebene. Daher formulierten Menzerath & de Oleza S. J. ein diesen Sachverhalt beschreibendes „weiteres phonetisches Quantitätsgesetz“: *„[...] die lautreichere Silbe [ist] auch die relativ kürzere“* [136, S.91]. Damit meinen sie, daß eine lautreichere Silbe kürzer ist, als aufgrund der Lautanzahl bei konstanter mittlerer Lautdauer zu erwarten wäre. Daraus folgt unmittelbar, daß die durchschnittliche Dauer des einer lautreicheren Silbe angehörenden Lautes kleiner ist als bei einer lautärmeren Silbe. Bereits in dieser sehr frühen Arbeit wird mithin schon der Zusammenhang zwischen Silben- und Lautdauern analysiert und diskutiert.

Weitkus 1931 [234] spricht sein Korpus (die 40 sog. Wenker-Sätze[3]) im August 1927 viermal selbst. Anschließend teilt er es entsprechend der Satzdauer in drei Sprechgeschwindigkeitsklassen auf: langsam, normal und schnell [234, S.6]. Dauermessungen der Laute, getrennt nach den drei Klassen, offenbaren schließlich, daß Weitkus mit 5.4, 6.2 und 7.1 Lauten pro Sekunde gesprochen hat [234, S.25]. Er zählt zu seinen Hauptresultaten, daß sich die Laute in zwei Klassen aufteilen lassen entsprechend der Veränderung ihrer Dauer im Rahmen einer Sprechgeschwindigkeitszunahme [234, S.28]:

[1] Als Beispiele seien hier Wagner 1891, 1892 [230, 231], Rousselot 1891, 1897/1908 [189, 190] und Meyer 1903 [139] angeführt. Letzterer entdeckte und formulierte in seiner Untersuchung das sog. *Vokalquantitätsgesetz*. Es besagt, daß intrinsische Vokaldauern in allen von ihm zu der Zeit untersuchten Sprachen (Deutsch, Englisch und Ungarisch) mit zunehmender Zungenhöhe abnehmen. Zu den Anfängen der Experimentalphonetik siehe Tillmann 1994 [209].

[2] Für das Französische war dies bereits seit den Untersuchungen von Roudet 1910 [188] bekannt, der die Dauer der Silbe [pa] in den Wörtern *pâte* (270 ms), *pâté* (200 ms), *pâtisserie* (140 ms) und *pâtisserie St. Germain* (120 ms) maß.

[3] Die Sätze von Wenker sind im Anhang der Arbeit von Weitkus abgedruckt [234, S.I].

(1) Zu den Lauten, bei denen eine relative Verkürzung[4] auftritt, gehören alle Konsonanten mit Ausnahme der stimmlosen Frikative.

(2) Dagegen werden Vokale, Diphthonge und eben auch die stimmlosen Frikative relativ verlängert.

Betrachtet man dagegen die Dauern absolut, so werden mit zunehmender Sprechgeschwindigkeit erwartungsgemäß im Mittel alle Laute kürzer.

Es sollte nicht verschwiegen werden, daß die von Weitkus angegebene über 7480 Laute gemittelte Lautdauer von 161.6 ms [234, S.11] unerklärlicherweise mehr als doppelt so lang ist, wie für einen gelesenen deutschen Text erwartet werden kann.[5] Auch die von Menzerath & de Oleza S. J. über 7883 Laute berechnete mittlere Lautdauer des Spanischen erscheint mit 130.3 ms ungewöhnlich groß [136, S.36], wobei dieses Korpus allerdings aus isoliert gesprochenen Wörtern und nicht aus Sätzen besteht.[6] Kurios erscheint auch das Formulieren von *allgemeinen Quantitätsgesetzen* auf der Grundlage eines einzigen Sprechers (de Oleza S. J. sprach selbst das zugrundeliegende Korpus im März 1925) mit der Begründung: „[...] die kleine Zahl der Versuchspersonen wird durch die Häufung der Versuche selbst mehr als ausgewogen." [136, S.10]

Für vergleichende Untersuchungen zu Lautdauern stellte Heinitz bereits 1921 [78] die Methode der „relativen Lautdauer" vor. Sie sollte die durch individuelle Sprechgeschwindigkeiten hervorgerufenen Dauervariationen normalisieren, indem die Dauer jedes einzelnen Lautes einer Folge von Lauten durch deren mittlere Lautdauer dividiert wird. Überdurchschnittlich langen Lauten werden so *relative Lautdauern* größer als 1 zugeordnet, während unterdurchschnittlich kurze Laute Werte unter 1 erhalten.

Allerdings zeigte Hildebrandt 1961 [81] anhand einiger Beispiele, daß die Methode der „relativen Lautdauer" von Heinitz in Abhängigkeit des Lautkontexts derartig unterschiedliche Werte für denselben Laut liefern kann, daß Fehlinterpretationen unvermeidlich sind. Er führt daraufhin die „durative Funktion" ein, indem er bei der Mittelwertbildung die Dauer des jeweils betrachteten Lautes ausschließt. Dabei entging ihm allerdings, daß seine Methode denjenigen Fehler vergrößert, der bei Lauten mit etwa durchschnittlicher Dauer auftritt. Damit ist sein Verfahren in phonetischen Untersuchungen keinesfalls besser zu gebrauchen als die Heinitzsche Methode.

Von Essen entscheidet sich 1949 [229] für das Phon und gegen die Silbe als Grundlage der Sprechgeschwindigkeitsbestimmung. Seiner Argumentation folgend ist die Dauer einer Silbe *abhängig* von der Anzahl der ihr angehörenden Laute und damit zu variabel.[7] Seine „Lautungsfrequenz" ist einfach die Anzahl der Phone eines Lautkomplexes dividiert durch dessen Gesamtdauer in Sekunden und entspricht damit exakt der Phonrate.

Als Erster macht Hildebrandt 1963 [82] darauf aufmerksam, daß man zwischen dem „reflexiven" Sprechtempo, dem die *kanonisch gegebene* Lautzahl zugrunde liegt, und dem „effektiven"

[4] Unter *relativer Verkürzung* versteht Weitkus die prozentuale Abnahme des Daueranteils einer Lautklasse an der Gesamtdauer aller Laute. So haben in seiner Untersuchung etwa stimmhafte Plosive bei langsamem Sprechen einen Anteil von 6.2% an der Gesamtdauer aller Laute. Bei normalem Sprechen sind es noch 5.9% und bei schnellem Sprechen nur noch 5.67% [234, S.28].

[5] Beispielsweise beträgt die mittlere Lautdauer über 39612 Laute von 1024 Sätzen des *PhonDatII*-Korpus 71.9 ms.

[6] Es ist bekannt, daß der Kymograph nur bei sehr lautem Sprechen klare Abbildungen lieferte. Möglicherweise wurde zugleich auch erheblich langsamer gesprochen, wodurch die Dauern so ungewöhnlich lang geraten sein könnten.

[7] Von Essen erwähnte allerdings das hier auf S. 130 zitierte zusätzliche „phonetische Quantitätsgesetz" von Menzerath & de Oleza S. J. [136, S.91] nicht. Dabei kann es als Argument gegen das Phon interpretiert werden: Die Dauer eines Lautes hängt nämlich *ebenfalls* von der Lautanzahl der jeweiligen Silbe ab. Folglich existieren wechselseitige Abhängigkeiten zwischen Silben- und Phondauer.

Sprechtempo, das auf der *tatsächlich realisierten* Lautzahl basiert, unterscheiden muß.[8] Während beim langsamen und deutlichen Sprechen beide Tempi nahezu identisch sind, steigt mit zunehmender Sprechgeschwindigkeit deren Differenz, da Elisionen von Phonen häufiger auftreten und somit *tatsächlich* weniger Laute realisiert werden als *kanonisch gegeben* sind.[9]

Es sollte allerdings hervorgehoben werden, daß Hildebrandt den Begriff der „Übermittlungsgeschwindigkeit von Information“ bemüht, um ihn mit dem „reflexiven“ Sprechtempo gleichzusetzen. Da jedoch die Frage, ob und welche *Information* in einem Kommunikationsprozeß durch Silben oder Laute oder gar Wörter übertragen wird, weder trivial noch geklärt ist, kann über diesen Begriff ohne weiteres kein Zugang zur *wirklichen* Sprechgeschwindigkeit gefunden werden. Wir werden aber auf diese Thematik in Kap. 4.5 ab S. 157 noch einmal zurückkommen.

2.2 Prosodie und Lokalität der Sprechgeschwindigkeit

Peterson & Lehiste 1960 [166] — die in ihrer Publikation auch den Begriff der *intrinsischen Silbendauer* einführten — gingen in einer Untersuchung zum Einfluß von benachbarten Konsonanten auf die Dauer betonter Vokale und Diphthonge bereits ganz selbstverständlich davon aus, daß die Suprasegmentalia Intonation, Akzent und Sprechgeschwindigkeit für einen Teil der Dauervariabilität verantwortlich sind [166, S.698]. Den durch verschiedene Sprechgeschwindigkeiten hervorgerufenen Variabilitätsanteil untersuchten Peterson & Lehiste gesondert, indem sie die Dauern betonter Wörter in mehreren Trägersätzen, gesprochen in zwei verschiedenen Sprechgeschwindigkeiten, erfaßten. Nach ihren Ergebnissen wurden bei einer Halbierung der Dauer des Trägersatzes die Dauern der Zielwörter nur um den Faktor 1.5 kürzer. Demnach wirkt sich die Sprechgeschwindigkeit auf eine Äußerung lokal sehr unterschiedlich aus.

Tillmann entwickelte 1964 [207, S.125ff] seine Theorie des *Ausprägungskodes*, die er 1972 [208] weiter verfeinerte, indem er auch den Begriff des *Reduktionskodes* bzw. *Reduktionsverlaufs* einführte und sich damit auf die Prosodie des Reduktionsgrades bezog, die er als lokal unterschiedlich ausgeprägtes bzw. unterschiedlich reduziertes Silbenprofil über gegebene Sätze interpretiert wissen wollte. In seiner Arbeit werden auch erstmals die Interaktionen zwischen den drei Prosodien Reduktionsgrad, Sprechgeschwindigkeit[10] und Intonation thematisiert. Seiner Ansicht nach indiziert einerseits der Intonationsverlauf das Ausprägungsprofil und andererseits wird es durch die Sprechgeschwindigkeit beeinflußt.

Mit seiner Sichtweise der lokal variierenden und zugleich vielschichtig interagierenden Prosodien ist diese Arbeit den damals gängigen Theorien weit voraus, wenngleich Firth bereits 1948 [48, S.151] das Nebeneinanderbestehen von Systemen zusammenhängender prosodischer und phonematischer Kategorien vorschlägt.[11] Trotzdem wurden diese von Tillmann angedeuteten In-

[8] Wir werden uns dieser Nomenklatur nicht anschließen, sondern die 1973 von Wood [241] gewählten synonymen Begriffe „Brutto-“ und „Netto“-Sprechgeschwindigkeit verwenden. Diese werden in Kap. 2.5 auf S. 140 definiert.

[9] Kühnert 1996 [116, S.357ff] zeigt mit Hilfe von EMA- und EPG-Messungen, daß bei Äußerungen wie etwa „Das Blatt kam von der Eiche“ zwischen den Realisierungen [atka:] und [aka:] nahezu beliebig viele Abstufungen des Reduktionsgrades auftreten können, was die beobachteten /t/-Bewegungen der Zungenspitze belegen.

[10] Tillmann bezeichnete sie 1972 als *Redegeschwindigkeit* und postulierte, daß sie in Silben pro Zeiteinheit zu messen sei [208, S.263].

[11] Der Begriff der *Prosodien* (oder auch *prosodischen Merkmale*) wurde 1948 von Firth [48] im Rahmen eines formalen Ansatzes zur Beschreibung aller phonemübergreifenden Phänomene unter Verwendung mehrerer prosodischer Ebenen, die parallel zur Ebene der Lautsegmente verlaufen, eingeführt. Damit divergiert der von Firth leider nicht exakt definierte Begriff der *Prosodien* wesentlich von dem, was heute unter *Prosodie* verstanden wird. Nach Firth ist z.B. Betonung als ein prosodisches Merkmal von Wörtern und Phrasen zu sehen und z.B. Intonation als ein prosodisches Merkmal von Teilsätzen. Aber auch Wort- und Silbenstruktur sowie Vokal- und Konsonantenkombinationen fallen nach seiner Theorie unter den Begriff der *Prosodien*.

teraktionen bis heute, fast 30 Jahre später, noch immer nicht im Rahmen umfangreicher prosodischer Untersuchungen aufgegriffen.

Statt dessen galt es in den 70er Jahren lediglich als Selbstverständlichkeit, daß bei der Sprachproduktion erhebliche Sprechgeschwindigkeitsvariationen auftreten können. Damals wurden sie aber in mehreren größeren Untersuchungen allein auf den Einfluß der Häufigkeit und Dauervariation von Sprechpausen zurückgeführt. Diese Meinung äußerten bereits Menzerath & de Lacerda 1933 [135, S.56], obwohl Weitkus 1931 [234] schon klar gezeigt hatte, daß die Phondauern selbst mit zunehmender Sprechgeschwindigkeit abnehmen.

Grosjean & Collins untersuchten 1979 [69] Sprechpausen und -atmung bei sechs Probanden, die eine Kurzgeschichte mit 116 Wörtern in fünf verschiedenen Sprechgeschwindigkeiten lasen, indem zuerst eine normale Geschwindigkeit gewählt wurde, welcher dann nach der Methode der *magnitude production* der Sprechgeschwindigkeitswert 10 zugeordnet wurde, woraufhin die geschätzten Geschwindigkeiten 2.5, 5, 20 und 30 produziert werden sollten. Für jede Kurzgeschichte wurde die Wortrate pro Minute berechnet und in Bezug zur Pausendauer und -häufigkeit gesetzt. Es stellte sich heraus, daß mit zunehmender Sprechgeschwindigkeit sowohl die Pausenhäufigkeit als auch die mittlere Pausendauer abnimmt.

Die ohne Einbeziehen der Sprechpausen ermittelte Sprechgeschwindigkeit, die sog. *articulation rate*, wurde zu der Zeit beim Vergleich vieler Äußerungen jeweils nur eines Sprechers als verhältnismäßig invariant, aber von Sprecher zu Sprecher als deutlich variierend und damit als ein sprecherspezifisches Merkmal eingeschätzt (siehe etwa Goldman-Eisler 1968 [63]). Dieses Ergebnis wurde auch untermauert durch die u.a. von Lass 1970 [120] gefundene und von Lane & Grosjean 1973 [119] bestätigte reziproke Beziehung zwischen Pausendauern und perzipierter Sprechgeschwindigkeit.

Erst eine Untersuchung von Miller, Grosjean & Lomanto 1984 [142] zeigte schließlich, daß die damals berichtete geringe Variabilität der Sprechgeschwindigkeit aus dem Berechnen eines Mittelwerts über lange Sprachsignalabschnitte resultierte, was der vergleichsweise großen lokalen Variation der Sprechgeschwindigkeit nicht Rechnung trug. Als Levelt sogar noch 1989 [123, S.390] schrieb: „The speaker increases his rate mainly by cutting back on pausing“, war diese Ansicht also längst überholt.

Im Rahmen einer Untersuchung von Koopmans-van Beinum & van Donzel wurde 1996 [112] der Versuch unternommen, der lokalen Natur des Sprechtempos gerecht zu werden, indem für jeden durch Sprechpausen begrenzten Sprachsignalabschnitt jeweils eine eigene durchschnittliche Silbendauer berechnet wurde. Obgleich die Autoren eingangs postulieren: „[. . .] the speaker is alternately speeding up and slowing down his/her speech production“ [112, S.1724], führt dieses von ihnen eingeführte quasi-lokale Sprechtempomaß schließlich doch nur zu dem Ergebnis: „[. . .] accounting for variations in speaking rate [. . .] is a very complicated task.“ [112, S.1727]

Die wissenschaftliche Fundiertheit der Aussagen von Batliner, Kießling, Kompe, Niemann und Nöth 1997 [12] aufgrund von nur 322 von Hand aufgedeckten Sprechtempowechseln (*schnell*↔*langsam*) in einem sehr großen spontansprachlichen Korpus mit 149.643 Wörtern erscheint fragwürdig, wenngleich der Ansatz, Sprechgeschwindigkeit nicht absolut, sondern relativ zu erfassen und damit das Problem des Sprechtempomaßes zu umgehen, sehr vielversprechend ist. Immerhin erkennen die Wissenschaftler das große Potential lokaler Tempowechsel und bezeichnen sie schließlich als „[. . .] interesting topic by its own.“ [12, S.766] Ihr Hauptergebnis, daß die automatische Extraktion prosodischer Phrasengrenzen vom Einbeziehen der Sprechgeschwindigkeit profitiert, wurde auch von Kim & Oh 1996 [103] festgestellt.

Dankovičová mißt 1999 [33] die lokale Artikulationsrate auf der Basis von phonologischen Wörtern und kann zeigen, daß vom ersten zum letzten phonologischen Wort einer Intonationsphrase mit zwei, drei oder vier Wörtern sowohl im Englischen als auch im Tschechischen eine sukzessive Verlangsamung der lokalen Artikulationsrate, ein *Rallentando* auftritt.

Ansonsten tritt die Lokalität der Variation vergleichsweise selten in den Fokus phonetischer Forschung — vermutlich aus dem von Koopmans-van Beinum & van Donzel genannten Grund. Immerhin ist der Einfluß der Sprechgeschwindigkeit als globaler Faktor auf akustische Merkmale wie z.B. *voice-onset time* (VOT)[12], auf Konsonanten[13], Vokale[14], Diphthonge[15], Silben[16] und auf die prosodische Strukturierung von Äußerungen[17] Gegenstand vieler phonetischer Untersuchungen. Auch in der kognitiven Sprachwahrnehmung, der Sprechplanung sowie in der Soziophonetik findet die Sprechgeschwindigkeit mittlerweile zunehmende Beachtung[18] und wurde in jüngster Zeit herangezogen, um Sprechstil genauer zu charakterisieren.[19]

Schließlich soll noch einmal betont werden, daß die Dauern der phonetischen Einheiten Konsonanten, Vokale und Silben nur bis zu einem gewissen Grad durch die Sprechgeschwindigkeit moduliert werden. Wie eine sehr umfassende Studie von Kohler 1986 [106] belegt, sind für Phondauern auch die zugehörige phonologische Kategorie, die Position innerhalb einer Silbe und die innerhalb eines Fußes ausschlaggebend. Auf Silbendauern wie auch auf Phondauern wirken sich zusätzlich der Silbenkontext, die rhythmische Struktur und schließlich die Intonation aus.

2.3 Artikulation und Sprechgeschwindigkeit

Bei dem Großteil der phonetischen Untersuchungen zur Sprachproduktion diente Sprechgeschwindigkeit nur als ein kontrollierbarer Faktor (wie etwa auch Betonung), um gezielt Variationen der Auslenkung, Geschwindigkeit und Beschleunigung von Artikulatoren für die Suche nach Invarianzen bei den Bewegungsmustern zu produzieren. Allerdings spielte dabei die lokal unterschiedliche Ausprägung der Sprechgeschwindigkeit bisher keine Rolle.

So ließen Kuehn & Moll 1976 [114] kurze CV-Sequenzen (z.B. /papapa/) von fünf Probanden mit langsamer sowie schneller Sprechgeschwindigkeit produzieren. Dabei stellten sie individuelle Reduktionsstrategien fest: Bei schnellem Sprechen verkürzten zwar alle Sprecher die Transitionen um 41% bis 56%, aber während ein Sprecher die Artikulatorengeschwindigkeiten kaum veränderte und statt dessen die Artikulatorenauslenkungen um etwa 55% verringerte, erhöhte ein anderer Sprecher die Geschwindigkeiten um 35%, verringerte die Auslenkungen aber nur um etwa 15%. Die Ergebnisse der anderen drei Sprecher lagen zwischen diesen beiden Extremen, ohne daß auch nur zwei Sprecher eine gemeinsame Strategie erkennen ließen [114, S.314].

Tuller 1980 [215] sowie Tuller, Harris & Kelso 1981/82 [218, 219] untersuchten anhand von elektromyographischen Messungen, inwiefern sich die Variation von Betonung und Sprechge-

[12] Kessinger & Blumstein 1997/98 [101, 102]

[13] Crystal & House 1988 [28], Byrd & Tan 1996 [16], Miller & Baer 1983 [141], Miller, O'Rourke & Volaitis 1997 [143], Arvaniti 1999 [8] und Rallo Fabra 1999 [183]

[14] Gay 1978 [60], Crystal & House 1988 [29], Gottfried, Miller & Payton 1990 [66], Gopal 1990 [65] und van Son & Pols 1993 [224]

[15] Gay 1968 [59]

[16] Crystal & House 1990 [30], Cedergren & Perreault 1994 [22] und Kuwabara 1998 [117]

[17] Ladd & Campbell 1991 [118], Fougeron & Jun 1998 [52] und Trouvain & Grice 1999 [212]

[18] Smith, Sugarman & Long 1983 [200], Francis & Nusbaum 1996 [57], Fon 1999 [50] und Tennant 1999 [204]

[19] Nach Messungen von Hirschberg 2000 [83] weist Spontansprache eine niedrigere durchschnittliche Silbenrate als Lesesprache auf. Ob dies bei Spontansprache durch mehr Pausen oder eine geringere Sprechgeschwindigkeit verursacht wird, muß ihrer Ansicht nach noch untersucht werden.

schwindigkeit auf die Muskelaktivitäten bei der Sprachproduktion auswirkt.[20] Beide Parameter beeinflussen sehr wohl die Dauer der akustischen Silbe; es stellte sich allerdings heraus, daß alle fünf Sprecher der Untersuchung zwar vergleichbare artikulatorische Strategien zur Variation des Betonungsgrades verwenden, bei der Variation der Sprechgeschwindigkeit allerdings sehr stark divergieren. Offenbar ist das Sprechgeschwindigkeitsverhalten aufgrund ihres suprasegmentalen Charakters produktionsseitig in hohem Grad sprecherspezifisch.

Insbesondere durch die Untersuchung von Gay 1981 [61] wird die Komplexität der artikulatorischen Kontrolle der Sprechgeschwindigkeit deutlich. Zwar wird die Inadäquatheit eines eindimensionalen Modells zur artikulatorischen Planung gezeigt, jedoch kann kein alternatives Modell vorgestellt werden. Statt dessen beschränkt sich der Autor darauf, mit elektromyographischen Methoden die zeitliche Verschiebung und veränderliche Geschwindigkeit von artikulatorischen Bewegungen in Verbindung mit dem Überlappen von Bewegungen verschiedener Artikulatoren nachzuweisen. Seiner Ansicht nach bewirkt Sprechgeschwindigkeit keine einfachen, sondern komplexe Transformationen der Segmentdauern, der Amplituden und Geschwindigkeiten von Artikulationsbewegungen und der zeitlichen Überlappung benachbarter Laute. Diese Ansicht finden wir bereits 1974 bei Gay, Ushijima, Hirose & Cooper [62].

Dagegen entdecken Ostry & Munhall 1985 [164] auf der Suche nach konsistenten kinematischen und physiologischen Mustern durch ihre Ultraschall-Experimente, bei denen drei Probanden CV-Sequenzen in zwei Sprechgeschwindigkeiten produzierten, einfache Beziehungen zwischen der Bewegungsamplitude, der maximalen Geschwindigkeit und der Dauer der Öffnungsgeste des Zungendorsums. Aber auch hier treten individuelle Artikulationsstrategien auf.

Engstrand 1988 [46] behauptet auf der Grundlage von V-/p/-V–Stimuli, gesprochen von zwei Probanden mit jeweils zwei Sprechgeschwindigkeiten, daß bei größerer Sprechgeschwindigkeit artikulatorische Vokal- und Konsonantengesten in einem höheren Grad koproduziert werden.

Flege 1988 [49] kann bezüglich der Zungenhöhe bei drei Sprechern wieder nur uneinheitliche Artikulationsstrategien feststellen; Sonoda 1987 [201] belegt auf der Grundlage der Daten von zwei Versuchspersonen und zwei Sprechgeschwindigkeiten individuelle Artikulationsstrategien bei Kieferbewegungen und -geschwindigkeiten.

Wieneke, Janssen & Belderbos 1987 [239] untersuchen ebenfalls Kieferbewegungen, allerdings bei vier Probanden und mit vier Sprechgeschwindigkeiten. Bei ihnen zeigen sich signifikante Korrelationen zwischen der Dauer einer Artikulationsbewegung und ihrer Amplitude (im Mittel beträgt der Korrelationskoeffizient $r = 0.77$).

Shaiman, Adams & Kimelman 1995 [197] untersuchen den Phasenwinkel zwischen der Bewegung der Oberlippe nach unten und der Bewegung des Kiefers nach oben bei bilabialen Verschlußbildungen im Hinblick auf die von Harris, Tuller & Kelso 1986 [74] verbreitete Hypothese, es gäbe unabhängig von Variationen der Sprechgeschwindigkeit invariante Zeitverhältnisse zwischen verschiedenen Artikulatoren. Auch hier zeigt sich, daß der Phasenwinkel durchaus sprechgeschwindigkeitsabhängig ist; die Richtung der Veränderung ist allerdings sprecherspezifisch.

Byrd & Tan 1996 [16] gehen in ihrer Untersuchung der Frage nach, ob eine höhere Sprechgeschwindigkeit bei der Produktion eher durch Verkürzen von artikulatorischen Bewegungen (entweder durch eine höhere Geschwindigkeit des Artikulators oder eine geringere Auslenkung) oder durch stärkeres Überlappen verschiedener Bewegungen erreicht wird. Zusätzlich zur Erkenntnis,

[20] Harris, Tuller & Kelso 1986 [74] entwickeln aufbauend auf diesen Untersuchungen eine Theorie der temporalen Organisation des Sprechens auf der artikulatorischen Ebene, nach der weder absolute noch relative Zeit dargestellt oder kontrolliert werden muß, sondern die auf intrinsischem Timing aufbaut.

daß beide Effekte gleichzeitig wirken, ergibt sich, daß bei Konsonanten Artikulationsort und -art sowie die Position in der Silbe die Artikulationsstrategie des Sprechers bei der Sprechgeschwindigkeitserhöhung beeinflussen.

Adams, Weismer & Kent 1993 [2] messen mit *x-ray microbeam* Unterlippen- und Zungenspitzenbewegungen von fünf Probanden bei fünf Sprechgeschwindigkeiten zwischen *sehr schnell* und *sehr langsam*. Dabei stellen sie fest, daß sich der Geschwindigkeitsverlauf der Artikulatoren während einer einzelnen Geste von einer symmetrischen Form mit einem Gipfel bei schnellem Sprechtempo zu einem asymmetrischen mehrgipfeligen Verlauf bei langsamem Tempo verändert. Daraus schließen sie, daß Veränderungen des Sprechtempos mit Veränderungen der Strategie der *motor control* einhergehen. Insbesondere sollen sich langsame Bewegungen aus vielen Teilbewegungen zusammensetzen, die durch Rückkopplungsmechanismen beeinflußt werden, während schnelle Bewegungen ihrer Ansicht nach hauptsächlich vorprogrammiert zu sein scheinen.

Kühnert 1996 [116, S.252f] stellt anhand von EMA- und EPG-Messungen bei fünf Versuchspersonen und jeweils zwei Sprechgeschwindigkeiten fest, daß die durchschnittliche Silbendauer in keinem Verhältnis zur Häufigkeit von Reduktionen alveolarer Artikulationsbewegungen steht. Zusätzlich bezeichnet sie in einer Fußnote die Bewertung der Sprechgeschwindigkeit aufgrund von Silbensegmentationen als „heikel“, da die beim Sprechen typischen Laut- und Silbenverschmelzungen[21] die Meßgrundlage verzerren.

Edwards, Beckman & Fletcher 1991 [43] untersuchen artikulatorische Kinematik unter drei Bedingungen, von denen bekannt ist, daß sie Silbendauern verlängern: *final lengthening*[22], Akzent und Sprechgeschwindigkeit. Sie stellen in Experimenten mit zwei Probanden fest, daß eine Verlangsamung der Sprechgeschwindigkeit durch Herabsetzen der Steifheit[23] innerhalb einer artikulatorischen Geste realisiert wird, während Akzentuierung das Überlappen von Gesten verringert. Die Kinematik des *final lengthening* ist ihrer Meinung nach mit einer lokalen Verlangsamung des Sprechtempos zu vergleichen.

Zusammenfassend lassen sich bei den hier betrachteten Untersuchungen einerseits erhebliche Abweichungen zwischen den sprecherindividuellen Artikulationsstrategien feststellen, so daß allgemeingültige Aussagen über den Einfluß der Sprechgeschwindigkeit auf die Artikulation — nicht nur wegen der geringen Versuchspersonenzahlen — problematisch sind; andererseits wird die Möglichkeit von *lokal* ausgeprägten Variationen der Sprechgeschwindigkeit, mit Ausnahme der Untersuchung von Edwards, Beckman & Fletcher 1991 [43], nicht in Erwägung gezogen.

Im Hinblick auf die hier angestrebten Experimente soll zuletzt noch auf eine Untersuchung von Okadome, Kaburagi & Honda 1999 [161] hingewiesen werden, in der die Beziehung zwischen Artikulation und Akustik unter Berücksichtigung der Sprechgeschwindigkeit mit Hilfe eines linearen Modells hergestellt wird. Das Modell basiert auf elektromagnetischen Messungen mittels EMMA von Kiefer, Ober- und Unterlippe, Zunge, Velum und Larynx dreier männlicher Sprecher, von denen jeder 16 Sätze in drei verschiedenen Sprechgeschwindigkeiten produzierte. Zusätzlich wurden die Segmentdauern der resultierenden 144 Sätze bestimmt. Die Autoren stellen anhand von sechs weiteren Sätzen fest, daß ihr Modell in der Lage ist, artikulatorische Reduktion vorherzusagen. Doch leider wird auch hier nicht der suprasegmentale Charakter der Sprechgeschwindigkeit und die damit verbundene große lokale Variabilität berücksichtigt.

[21] So können die beiden Wörter *mit dem* [mɪt de:m] in einer Äußerung durchaus zu [mɪm] reduziert werden.

[22] Daß Silben und Laute zum Ende einer Äußerung oder Phrase hin typischerweise überdurchschnittliche Dauern aufweisen, wird entweder als *final lengthening*, *pre-final lengthening* oder auch als *pre-pausal lengthening* bezeichnet. Bereits 1899 konnte Grégoire [67] dieses Phänomen aufzeigen.

[23] Der Begriff *Steifheit* wird hier im Sinne der Federkonstante in einem Masse-Feder-Dämpfungs–System verstanden.

2.4 Sprechgeschwindigkeit in der automatischen Spracherkennung

Im Forschungsbereich der automatischen Spracherkennung setzte sich erst in den letzten Jahren die Einsicht durch, daß die große Variabilität der Sprechgeschwindigkeit eine der Ursachen für einen beträchtlichen Einbruch der Erkennungsleistung ist (Jones & Woodland 1993 [97], Osaka, Makino & Sone 1994 [162]). Insbesondere bei sehr schnellen oder sehr langsamen Äußerungen konnte die Erkennungsfehlerrate um 50% und mehr gegenüber einer durchschnittlichen Sprechgeschwindigkeit ansteigen (Siegler & Stern 1995 [199], Mirghafori, Fosler & Morgan 1995 [145], Verhasselt & Martens 1996 [228]).[24]

Im Bereich der Sprachsynthese wurde zwar früher begonnen, Einfluß und Bedeutung der Sprechgeschwindigkeit zu erforschen,[25] jedoch wurden von den Forschern aus dem Bereich der automatischen Spracherkennung diese Ergebnisse kaum aufgegriffen.

Um Spracherkennungssysteme an die jeweils vorliegende Sprechgeschwindigkeit adaptieren zu können und damit ein neues weites Forschungsfeld zur Verbesserung der Erkennungsleistung zu erschließen, entstand schnell das Bedürfnis nach automatischen Verfahren zur Schätzung der Sprechgeschwindigkeit.

Einen besonders vielversprechenden Ansatz lieferte Heid 1998 [77, S.296], als er „relative Dauerkonturen“ von Äußerungen darstellte, indem er die Abweichungen der gemessenen Phondauern (*tokens*) von zuvor anhand des *PhonDatII*-Korpus ermittelten und damit prototypischen Phondauern (*types*) für jedes einzelne Phon einer Äußerung berechnete. Die Grundidee dazu lieferten Campbell & Isard 1991 [21], als sie die in der Prüfstatistik übliche Abbildung einer Variablen auf die Standardnormalverteilung bei der Lautdauervorhersage in der Sprachsynthese verwendeten. Während Heid dieses Verfahren zur Analyse und Charakterisierung des *PhonDatII*-Korpus benutzte, könnte eine Durchschnittsbildung über die Abweichungen einer Menge benachbarter Phone die *relative lokale Sprechgeschwindigkeit* liefern: würde z.B. schneller gesprochen, so lägen die gemessenen Phondauern im Mittel unter denen der Prototypen, womit sich eine überdurchschnittliche Sprechgeschwindigkeit ergäbe. Eine ganz ähnliche Methode verwendete bereits Wightman 1992 [240], die er als „*mean normalized speech duration*“ τ bezeichnete.

Dieser Ansatz kam in vergleichbarer Weise auch bei Martínez, Tapias, Álvarez & León 1997 [134] zum Einsatz. Sie ermitteln die Sprechgeschwindigkeit, nachdem ein Spracherkennungsprozeß mit Viterbi-Alignment die Segmentgrenzen lieferte.

Ohno, Fujisaki & Taguchi 1997 [158] arbeiten in einer neueren Arbeit auch mit Prototypen zur Berechnung der relativen Sprechgeschwindigkeit. Allerdings kommen hier Prototypen auf Wortebene zum Einsatz.

Da die für derartige Sprechgeschwindigkeitsmessungen notwendige automatische Segmentation phonetischer Einheiten aufwendig und zugleich fehlerträchtig ist, wurden und werden viele Experimente durchgeführt, Repräsentationen der Sprechgeschwindigkeit noch vor dem eigentlichen Spracherkennungsprozeß aus dem Sprachsignal zu extrahieren und damit unabhängig von der Qualität der automatischen Segmentation zu werden.

Hier sei die „enrate“ (= „*energy rate*“) von Morgan, Fosler & Mirghafori 1997 [146] erwähnt, der die Annahme zugrunde liegt, daß die Modulationsgeschwindigkeit der Energiehüll-

[24] Diese Erkenntnisse waren auch mit ursächlich für das Wiederaufleben der Forschung zur automatischen Silbenextraktion (siehe etwa Reichl & Ruske 1993 [186] oder Pfitzinger, Burger & Heid 1996 [172]), auch wenn Mermelstein bereits 1975 [137] die Qualität seiner automatischen Extraktion als hinreichend gut einschätzte.

[25] Siehe hierzu z.B. Campbell 1988/90 [17, 18, 19], der zugunsten der Silbe gegen die linguistischen Einheiten des Wortes und des Phons als Meßgrundlage der Sprechgeschwindigkeit argumentiert.

kurve von Sprachsignalen die Sprechgeschwindigkeit repräsentiert. Zunächst erscheint diese Idee plausibel, da automatische Silbenextraktionsverfahren ebenfalls auf Amplitudenverläufen aufbauen. Der Korrelationskoeffizient mit der gemessenen Silbenrate beträgt allerdings nur $r = 0.42$. Aus anderen linguistischen Einheiten, wie z.B. Phonen oder Wörtern berechnete Geschwindigkeiten korrelierten die Forscher mit der „enrate“ bisher bedauerlicherweise nicht.

Statt dessen publizieren Morgan & Fosler-Lussier 1998 [147] eine interessante Weiterentwicklung der „enrate“, die sie als „mrate“ (= „*multiple rate estimator*“) bezeichnen. Hier wird die „enrate“ zusätzlich mit der Anzahl der Energiemaxima in vier kreuzkorrelierten Teilfrequenzbändern kombiniert. Auf diese Weise kommt es zu einer deutlichen Verbesserung des Korrelationskoeffizienten mit der gemessenen Silbenrate auf $r = 0.67$. Am Rand soll angemerkt werden, daß Fosler-Lussier & Morgan 1999 [51] in einer späteren Untersuchung das bereits gut bekannte Phänomen bestätigen können, daß Sprechgeschwindigkeit die Aussprache von häufigen Wörtern deutlich modifiziert.

Samudravijaya, Singh & Rao 1998 [193] entwickeln und prüfen drei verschiedene Methoden zur Sprechgeschwindigkeitsextraktion, die ebenfalls nicht auf linguistischen Einheiten basieren. Demzufolge nennen die Forscher sie „pre-recognition measures“ und korrelieren sie mit der nachträglich extrahierten Phonrate. In der Untersuchung nimmt mit steigender Phonrate die Dauer stationärer Sprachsignalpassagen ab und damit der Transitionsanteil zu ($r = 0.42$). Die Häufigkeit des Wechsels zwischen stimmhaften und stimmlosen Passagen steigt ($r = 0.21$), und die Variation der Signalamplitude fällt ($r = -0.19$).

Schließlich sei in dieser Kategorie noch ein weiterer Ansatz vorgestellt, der in den Arbeiten von Pfau 2000 [167] sowie Pfau, Faltlhauser & Ruske 2000 [168] verfolgt wird: Aufgrund von normierter modifizierter Lautheit und Nulldurchgangsrate werden die Silbenkerne extrahiert und damit wiederum die Silbengeschwindigkeit berechnet, die im nachfolgenden HMM-basierten Spracherkennungsprozeß zur Sprechgeschwindigkeitsnormalisierung der Zeitachse noch vor der Durchführung des Viterbi-Algorithmus verwendet wird.

Die hier vorgestellten Repräsentationen der Sprechgeschwindigkeit wurden beispielsweise verwendet, um zwischen verschiedenen Mengen von Hidden Markov Modellen auszuwählen, die für verschiedene Sprechgeschwindigkeiten trainiert wurden.

Ansätze dieser Art konnten die Erkennungsraten verbessern (Wu, Kingsbury, Morgan & Greenberg 1998 [243], Richardson, Hwang, Acero & Huang 1999 [187]), doch waren die in den verschiedenen Untersuchungen verwendeten Repräsentationen der Sprechgeschwindigkeit keineswegs vergleichbar. So wurden entweder Silben bzw. Phone pro Sekunde gemessen oder Modelle entwickelt, die Werte lieferten, welche jeweils einer dieser beiden Geschwindigkeiten ähnlich waren — und all diese Meßwerte wurden gleichermaßen als *Sprechgeschwindigkeit* bezeichnet.

In jüngster Zeit werden in der Spracherkennung nun auch Ansätze verfolgt, die nicht mehr darauf abzielen, explizit Sprechgeschwindigkeitsmessungen durchzuführen, sondern den gängigen HMM-Formalismus modifizieren, indem z.B. Zustandsübergänge zusätzlich mit Emissionswahrscheinlichkeiten für die Sprechgeschwindigkeit versehen werden (Tuerk & Young 1999 [214]) oder indem zusätzliche Freiheitsgrade auf der Zeitachse eingeführt werden, um den Sprechgeschwindigkeitsvariationen besser folgen können (Saul & Rahim 1999 [194], Faltlhauser, Pfau & Ruske 1999 [47]). Der Hauptvorteil dieser neuen Ansätze ist darin zu sehen, daß einerseits wieder Standardverfahren zur Signalvorverarbeitung eingesetzt und andererseits die modifizierten HMMs homogen trainiert werden können.

Am Rande sei bemerkt, daß Anderson, Liberman, Gillick, Foster & Hama 1999 [7] sogar den Weg gingen, die Sprecher auf das Spracherkennungssystem zu trainieren, statt wie üblich umgekehrt vorzugehen. Ihre Hoffnung, die Sprecher würden ihren Sprechstil und ihre Sprechgeschwindigkeit derart verändern, daß sich die Erkennungsraten erhöhen, erfüllte sich in einer geringen aber signifikanten Verbesserung.

2.5 Definitionen: globale, lokale und relative Sprechgeschwindigkeit

In diesem Abschnitt soll etwas mehr Klarheit in die Begriffsvielfalt gebracht werden, die sich um die Sprechgeschwindigkeit herum gebildet hat, und insbesondere auch beschrieben werden, wie die jeweiligen aus der Literatur bekannten und bereits etablierten „Sprechgeschwindigkeiten“ zu ermitteln sind.

Globale Sprechgeschwindigkeit wird berechnet, indem man die Anzahl phonetischer bzw. linguistischer Einheiten eines Redebeitrags[26] durch die akkumulierte Gesamtdauer der Einheiten teilt. Sie wird in Einheiten pro Sekunde angegeben, wobei als Einheiten typischerweise entweder Phone, Silben oder Wörter verwendet werden. Denkbar wären aber auch Moren, Füße oder Morphe. Zu beachten ist, daß sich aufgrund ihrer unterschiedlichen mittleren Dauern je nach gewählter phonetischer bzw. linguistischer Einheit ein anderer Wertebereich für die zugehörige Geschwindigkeit ergibt.

Lokale Sprechgeschwindigkeit ergibt sich ebenfalls aus der Anzahl phonetischer bzw. linguistischer Einheiten. Allerdings liegt nicht die gesamte Äußerung einem einzelnen Meßwert zugrunde, sondern man berechnet vom Beginn bis zum Ende der fraglichen Äußerung in gleichmäßigen Abständen von z.B. 20 ms je einen Meßwert, indem man mit Hilfe einer um den jeweiligen Meßpunkt zentrierten Fensterfunktion einen Signalausschnitt von z.B. 500 ms Dauer extrahiert und aus ihm dann die Einheiten pro Sekunde ermittelt. So ergibt sich in diesem Beispiel alle 20 ms ein lokaler Sprechgeschwindigkeitswert und damit im Ganzen eine synchron zum Signal verlaufende Sprechgeschwindigkeitskurve, die bei langsamen Äußerungsteilen einen niedrigeren und bei schnellen einen entsprechend höheren Wert aufweist.

Relative lokale Sprechgeschwindigkeit wurde von Ohno & Fujisaki 1995 [156] bzw. von Ohno, Fukumiya & Fujisaki 1996 [160] eingeführt. Sie kann nur zwischen Äußerungen mit identischem Wortlaut berechnet werden, weil ein *dynamic-time-warping*–Algorithmus (DTW) aus zwei Äußerungen diejenige *warping*-Kurve ermitteln muß, die die zeitliche Abfolge korrespondierender Laute beider Sprachsignale aufeinander abbildet. Die Steigung der Kurve wird dann durch ein 270 ms langes Dreieckfenster geglättet. Das Resultat ist eine Vergleichskurve, die den Wert 1 bei jenen Äußerungsabschnitten annimmt, die gleich schnell produziert wurden, und beispielsweise den Wert 0.5 bzw. 2, wenn ein Abschnitt des einen Signals halb bzw. doppelt so lange dauert wie der entsprechende Abschnitt des anderen Signals. Für die Wahrnehmung der Sprechgeschwindigkeitsverhältnisse sollte jedoch keine direkte Proportionalität zu den Meßwerten postuliert werden.

Oft ist auch die Rede von der *Artikulationsrate* (Miller, Grosjean und Lomanto 1984 [142], Crystal & House 1990 [30], Dankovičová 1997 [32]). Laver 1994 [121] unterscheidet zwischen „articulation rate“ und „speaking rate“ und bezeichnet mit ersterer das Tempo, mit dem eine Äußerung inklusive aller Häsitationen und gefüllten Pausen, aber ohne ungefüllte Pausen gesprochen wurde. Dagegen basiert die „speaking rate“ laut Laver auf längeren Äußerungen, z.B. auf vorgelesenen oder erzählten Geschichten, und bezieht zusätzlich ungefüllte Pausen in die Be-

[26] Hiermit ist wenigstens ein langer Satz, eher noch eine Menge von Sätzen gemeint.

rechnung ein, die etwa beim Luftholen zwischen Äußerungsteilen entstehen. Beide Raten werden in sprachlichen Einheiten pro Sekunde gemessen. Laver definiert hier offensichtlich *globale* Geschwindigkeiten, da es ihm um nur *eine* Maßzahl für einen ganzen Redebeitrag geht. Im folgenden Abschnitt werden wir noch einmal auf die Problematik der Häsitationen und Sprechpausen zurückkommen.

Brutto- und *Nettosprechgeschwindigkeit* spielen nach Woods Auffassung (1973 [241]) eine besondere Rolle: „In any investigation of tempo, it is necessary to distinguish between gross and net measures of rate." [241, S.11][27] Werden beim Zählen der sprachlichen Einheiten Elisionen nicht mitgezählt, sondern nur die tatsächlich im Sprachsignal realisierten Einheiten (einschließlich der insertierten Einheiten) einbezogen, so ergibt sich eine Nettosprechgeschwindigkeit. Von Bruttosprechgeschwindigkeit spricht Wood dann, wenn die Anzahl der Wörter, Silben oder Phone aus dem einer Äußerung zugrunde liegenden Textmaterial über die kanonische Form gewonnen wird und nicht aus dem realisierten Sprachsignal.

Zusammenfassend kann nun gesagt werden, daß bei Angaben zur Sprechgeschwindigkeit fünf Aspekte geklärt sein müssen: Wurde sie *erstens* global oder lokal, *zweitens* relativ oder absolut, *drittens* brutto oder netto, *viertens* mit oder ohne Einbeziehung ungefüllter Pausen berechnet und, *fünftens*, welche sprachlichen Einheiten lagen der Messung zugrunde?

2.6 Sprechgeschwindigkeit in der Wahrnehmung

Alle im vorausgegangenen Kapitel dargestellten Definitionen der Sprechgeschwindigkeit basieren auf Merkmalen, die entweder direkt der akustischen Domäne entstammen oder auf zusätzliche Informationen aus einem linguistischen Abstraktionsprozeß angewiesen sind. In Kap. 1.4 auf S. 126 stellten wir aber fest, daß zwischen Akustik und Perzeption der Sprechgeschwindigkeit eine psychophysikalische Relation erwartet werden kann. Somit ist nur auf der Grundlage von Perzeptionsexperimenten die Frage zu beantworten, welche akustischen Merkmale tatsächlich die Sprechgeschwindigkeit kennzeichnen.

Dieser Frage gingen in der Vergangenheit bereits einige Forscher nach. Während Osser & Peng 1964 [163] annahmen, daß die Phonemrate[28] der perzipierten Sprechgeschwindigkeit am ähnlichsten war, kamen Grosjean & Lane 1976 [70] sowie Grosjean & Lass 1977 [71] zu dem Ergebnis, daß einerseits Artikulationsrate und andererseits Anzahl und Dauer der Sprechpausen in einer dem Konzept der *trading relations*[29] entsprechenden Weise die Sprechgeschwindigkeitswahrnehmung maßgeblich bestimmen.

Im Gegensatz dazu bevorzugte Butcher 1981 [15], der in einer Untersuchung kurze Sprachstimuli beurteilen ließ, die Silbenrate als den besten Kandidaten zur Repräsentation der perzipierten Sprechgeschwindigkeit. Bedauerlicherweise nahmen an seinen Perzeptionsexperimenten nur zwei bzw. zehn Probanden teil, die zudem nur drei Urteile „*langsam*", „*normal*" und „*schnell*" bzw. fünf Urteile von „*sehr langsam*" bis „*sehr schnell*" vergeben durften.

[27] Wie bereits auf S. 131 in Kap. 2.1 dargestellt, legte auch Hildebrandt 1963 [82] — also schon zehn Jahre vor der Veröffentlichung von Wood — Wert auf diese Unterscheidung, verwendete allerdings andere Begriffe.

[28] Die *Phonemrate* unterscheidet sich von der aus dem vorigen Abschnitt herleitbaren *Bruttophonrate* dadurch, daß die Transkription des der Äußerung zugrundeliegenden Textes genau das *Phoneminventar* der jeweiligen Sprache nutzt, während eine Bruttophonrate z.B. Sproßvokale und -konsonanten einbeziehen könnte, obwohl ihr Auftreten vorhersagbar ist und sie daher in einer phonologischen Transkription keine Berücksichtigung finden würden.

[29] Mit *trading relations* bezeichnet man das synergetische Wirken unterschiedlicher Merkmale. Insbesondere kann ein überdurchschnittlich ausgeprägtes Merkmal durch ein anderes unterdurchschnittlich ausgeprägtes Merkmal in der Gesamtwirkung kompensiert werden.

Den Os 1985 [38] verwendete aus zwei Sprachen (Niederländisch und Italienisch) je neun verschieden schnell produzierte Stimuli, die sie jeweils unter drei Bedingungen (normal, monoton und reduziert auf die prosodische Information) einem Auditorium präsentierte. Weiterhin extrahierte sie deren Bruttosilbenraten, Nettosilbenraten und Nettophonraten, um herauszufinden, welche der Raten am besten mit der perzipierten Sprechgeschwindigkeit korreliert. Diese wiederum erhielt sie auf zwei verschiedene Weisen: Einerseits ermittelte sie auf der Basis von *paired-comparison*-Tests eine Rangfolge der Stimuli und andererseits ließ sie die Stimuli auf einer siebenstufigen Skala einschätzen, um daraus intervallskalierte Beurteilungen abzuleiten.

Ihren Resultaten zufolge eignen sich die Maße der Bruttosilbenrate („*linguistische Silben*“) sowie der Nettophonrate gut, um perzipierte Sprechgeschwindigkeit vorherzusagen ($r \approx 0.9$). Dagegen weist die Nettosilbenrate („*phonetische Silben*“) grundsätzlich niedrigere Korrelationskoeffizienten auf. Die Varianzaufklärung einer ANOVA ihrer Perzeptionsergebnisse zeigt, daß der Faktor „Stimulus“ immerhin 83.6% der beobachteten Varianz erklärt und damit hauptsächlich für die Variationen im Versuchspersonenverhalten verantwortlich ist.

Bemerkenswert ist außerdem auch die Hypothese von den Os, daß Italiener, die Niederländisch beurteilen und mit der Sprache nicht vertraut sind, ihre Urteile eher an der Phonrate als an der Silbenrate orientieren. Da den Os Sätze mit einer Dauer von etwa drei Sekunden als Stimuli auswählte, die jeweils nur durch einen Meßwert repräsentiert und in ihrer Gesamtheit eingeschätzt wurden, handelte es sich in ihrer Untersuchung um globale Raten.

Lass berichtete 1970 [120], daß beim Vorlesen die systematische Variation der Sprechpausen trotz konstant gehaltener Segmentdauern eine Änderung der Sprechgeschwindigkeitswahrnehmung hervorruft. Dies ergab sich anhand der Urteile von 78 Versuchspersonen, denen sechs Urteilsstufen zwischen „*sehr langsam*“ und „*sehr schnell*“ zur Auswahl standen, gilt allerdings nur für globale Sprechgeschwindigkeit, da die 31 beurteilten Stimuli jeweils etwa 30 Sekunden dauerten und eine Kurzgeschichte umfaßten.

An dieser Stelle muß im Hinblick auf unser eigenes Vorgehen in Kap. 7 festgehalten werden, daß sich Sprechpausen im lokalen Fall anders auswirken: Hört ein Proband ein kurzes aus einer Äußerung herausgeschnittenes Sprachsignalstück, das mit einer Sprechpause beginnt oder endet, kann er nur den entsprechend kürzeren Sprachteil bewerten (die Sprechpause unterscheidet sich ja nicht von der vorausgehenden bzw. nachfolgenden Stille). Eine Sprechpause würde also die vorgesehene Stimulusdauer unkontrollierbar verkürzen. Daraus ergibt sich für die anschließenden Perzeptionsexperimente, daß die akustischen Stimuli keine Sprechpausen enthalten sollten. Für die lokale akustische Analyse folgt dementsprechend, daß bei der Berechnung lokaler Geschwindigkeiten Sprechpausen ausgeschlossen werden sollten.[30]

Ventsov weist 1981 [226] darauf hin, daß er bereits 1976 an der Prägung des Begriffs des „Momentantempos“[31] beteiligt war. Hiermit sollte ein hypothetischer wahrnehmbarer Parameter bezeichnet werden, der der lokalen Sprechgeschwindigkeit folgt und insbesondere zur Detektion von Sprechgeschwindigkeitsänderungen — seien sie durch emphatischen Akzent oder durch *pre-final lengthening* hervorgerufen — dient und damit perzeptiv relevant ist.

Mit Hilfe von Perzeptionsexperimenten belegt Ventsov, daß bei synthetischen Stimuli mit acht bis zehn Vokalen, die durch kurze Pausen oder kurze Nasale getrennt präsentiert werden, Versuchspersonen in der Lage sind zu hören, ob die Silbengeschwindigkeit konstant ist oder sich

[30] Erst auf einer höheren und weniger lokalen Ebene sollten Sprechpausen dann Berücksichtigung finden.

[31] Er spricht von „running tempo“ oder „momentary tempo“. Da in der Systemtheorie die Begriffe Momentanfrequenz, -phase und -amplitude wohldefiniert sind, wollen wir den Begriff „Momentantempo“ möglichst vermeiden und statt dessen „lokales Sprechtempo“ oder „lokale Sprechgeschwindigkeit“ sagen.

innerhalb eines Stimulus verändert. Er kommt zu dem Schluß, daß die Dauer der geschlossenen Silbe ausschlaggebend für die Wahrnehmung der Sprechgeschwindigkeit und damit auch deren lokaler Variation ist.

Pompino-Marschall, Piroth, Tilk, Hoole & Tillmann 1982/84 [178, 179, 180] überprüfen diese Hypothese anhand von Stimuli, die auch auf komplexeren Silbenstrukturen basieren, und können Ventsov eine Überinterpretation seiner Ergebnisse nachweisen. Ihren Experimenten zufolge wird die lokale Silbengeschwindigkeit vielmehr durch „Koartikulationspunkte“ konstituiert. Pompino-Marschall et al. schreiben, „[. . .] daß der psychologische Moment des Silbenbeginns im ‘Koartikulationspunkt’ liegt, d.h. dem Punkt der gleichzeitigen Produktion des prävokalischen Konsonanten und des Vokals.“ [179, S.312][32]

Den Einfluß von F_0 auf die Wahrnehmung der Sprechgeschwindigkeit untersuchten erstmals Hoequist[33] 1983/84 [85, 87], den Os 1985 [38] und Kohler 1986 [107, 108], der den Effekt folgendermaßen zusammenfaßt: Bei sonst unveränderten Stimuli bewirkt eine *monotone*, aber höhere durchschnittliche Grundfrequenz auch die Wahrnehmung einer höheren Sprechgeschwindigkeit und eine tiefere Grundfrequenz entsprechend eine niedrigere Sprechgeschwindigkeit. In derselben Weise bewirkt eine *Bewegung* von F_0 in Richtung höherer Frequenzen eine höhere perzipierte Sprechgeschwindigkeit, während die Bewegung zu tieferen Frequenzen hin wiederum eine niedrigere perzipierte Sprechgeschwindigkeit auslöst.

Nach einer Reihe weiterer Untersuchungen zur Sprechtempowahrnehmung, die systematische Variationen der rhythmischen Struktur der verwendeten Stimuli auf der Fußebene[34] umfassen, kommt Hoequist 1986 [89] zu dem Schluß, daß segmentale Dauer niemals nur Informationen über das Sprechtempo liefert, sondern immer auch Informationen über Segmenttyp, Betonung und syntaktische Grenzen trägt.

Dies führt uns direkt zu einer weiteren Klasse von Untersuchungen, die sich mit dem Einfluß von Sprechgeschwindigkeit auf die Lautwahrnehmung beschäftigen und belegen, daß z.B. bei dem Phonem /w/ des Englischen die Dauer der Transitionen bei niedrigen Sprechgeschwindigkeiten länger sein muß als bei höheren Sprechgeschwindigkeiten, um eine Verwechslung mit dem Phonem /b/ zu vermeiden (Miller 1981 [140]).

Allgemeiner ausgedrückt sind Hörer zum einen sehr sensibel bezüglich durch unterschiedliche Sprechgeschwindigkeiten hervorgerufener Veränderungen in der akustischen Feinstruktur, und zum anderen verarbeiten sie segmental relevante akustische Eigenschaften in Bezug auf die jeweils vorliegende Sprechgeschwindigkeit (Summerfield 1981 [202]). Nach Nooteboom 1979 [151] ist hier die Sprechgeschwindigkeit des unmittelbar folgenden Kontextes als Referenz wirksam. Demzufolge muß ein Sprecher etwa auch die *voice-onset time*, die z.B. die Phonemkategorien /p/ und /b/ unterscheidet, adäquat an die jeweils gewählte lokale Sprechgeschwindigkeit anpassen (Miller & Volaitis 1989 [144]). Diese Klasse von Untersuchungen ruft geradezu zwingend nach einem allgemeingültigen Maß der Sprechgeschwindigkeit, so wie es in dieser Arbeit entwickelt werden soll.

[32] Ihren Begriff des „Koartikulationspunkts“ wollen sie im Sinne der Konzepte der *Koartikulation* und der *Steuerung* von Menzerath & de Lacerda 1933 [135] verstanden wissen.

[33] Hoequist prägte 1984 [88] den Begriff „Lokaltempo“ für die von ihm beobachteten Wahrnehmungsphänomene bezüglich der Sprechgeschwindigkeit auf Silbenebene.

[34] Siehe Fußnote 12 auf S. 146.

3

Zeitliche Struktur der Sprache

Den Eindruck einer Geschwindigkeit bekommt man trivialerweise nur anhand solcher *wahrnehmbarer Veränderungen*, die in der Zeitebene stattfinden.[1] Der Begriff der „wahrnehmbaren Veränderung“ kommt allerdings ohne eine genauere Erläuterung nicht aus, denn beispielsweise wäre das Drehen eines gänzlich einfarbigen und unstrukturierten Reifens trotz einer physikalisch nachweisbaren permanenten und gleichförmigen Veränderung in der Zeit nicht wahrnehmbar.

Die Rotation kann in ihrer physikalischen Realität immer als ein ablaufender *Vorgang*[2] bezeichnet werden. Wenn nun dieser Vorgang wahrnehmbar ist — etwa durch das Vorhandensein einer herausstechenden und damit für die Sinnesorgane faßbaren Strukturierung —, so wird die Abfolge der diese *faßbare Strukturierung* bildenden Elemente zu einer Abfolge von wahrnehmbaren *Ereignissen*[2].

Sind beispielsweise bei einem Reifen Speichen in 10-Grad–Winkeln angeordnet, so kann eine langsame gleichförmige Rotation als ein Vorgang wahrgenommen und zugleich jede Teilrotation um 10 Grad als Ereignis aufgefaßt werden, weil dann jedesmal der gleiche Augenblickszustand erreicht ist wie zu Beginn jeder Teilrotation. Ist die Rotation zu schnell, um der Bewegung der Speichen noch visuell folgen zu können, dann wäre etwa das Ventil des Reifens, das nur einmal in jeder vollen Umdrehung seinen Ausgangspunkt erreicht, ein potentieller Kandidat für ein Ereignis. Bei einer weiteren Steigerung der Rotationsgeschwindigkeit, wenn selbst das Ventil zu einer Kreisbahn verschwimmt, ist nur noch aus der Erfahrung heraus zu entscheiden, daß es sich um eine Rotation handelt und nicht etwa um einen Rotationskörper, der still steht. Tatsächlich entzöge sich ein derartiger Vorgang der Wahrnehmung, und die nicht mehr wahrnehmbare Strukturierung ließe auch keine sichere Identifikation von Ereignissen zu.

Offensichtlich ist einerseits eine wahrnehmbare Struktur erforderlich, anhand der die Veränderung in Form von Ereignissen für den Menschen erst faßbar wird, und andererseits hängt die Wahrnehmbarkeit einer Veränderung auch von der Geschwindigkeit ab, mit der Ereignisse aufeinander folgen.[3] Diesem Gedankengang folgend stellt sich zwangsläufig die Frage nach derjenigen zeitlichen Struktur von gesprochener Sprache, die eine Tempowahrnehmung ermöglicht.

Diese Frage führt geradezu unmittelbar zu einer Reihe von einschlägigen Begriffen aus der phonetischen Forschung (*P-centers, Rhythmus* und *Isochronie*), anhand derer die Problematik der

[1] So hat ein gedruckter Text (wie etwa der auf dieser Seite) keine Geschwindigkeit, obwohl von Buchstabe zu Buchstabe Veränderungen zu beobachten sind. Diese treten jedoch im Raum auf, nicht in der Zeit. Es kann sich allerdings eine Lesegeschwindigkeit ergeben oder die Geschwindigkeit eines Laufbands, das den Text nach und nach anzeigt.

[2] Die Begriffe *phonetisches Ereignis* und *phonetischer Vorgang* sowie deren Relation wurden von Tillmann 1980 [210] ausführlich thematisiert.

[3] Bereits Augustinus schrieb 397/398 [11], daß sich auch die subjektive Wahrnehmung der Zeit an der Wahrnehmung von Ereignissen festmacht.

zeitlichen Struktur von Sprache deutlich wird. Daher werden diese Begriffe nun in den anschließenden Abschnitten erläutert, wobei auch phonologische und psycholinguistische Ansätze einbezogen werden.

3.1 P-centers / Ereigniszeitpunkte

Perceptual centers (= *P-centers*) sind Wahrnehmungszeitpunkte auf Silbenebene, die aufgrund der Annahme unterstellt werden, daß Sprechen in eine Abfolge wiederkehrender *Ereignisse* strukturiert sein muß, um z.B. die Wahrnehmung von Rhythmus und auch Geschwindigkeit überhaupt ermöglichen zu können.

Marcus führte 1976 [132] den Begriff des *perceptual centers* ein und bezeichnete damit den *psychologischen Ereigniszeitpunkt* („*psychological moment of occurrence*“) eines Klanges. Nach Marcus ist eine Folge von einsilbigen Wörtern per Definition genau dann *regelmäßig*, wenn deren P-centers in regelmäßigen zeitlichen Abständen angeordnet sind und nicht etwa deren physikalisch meßbare Wortanfänge, die sich gravierend von den P-centers unterscheiden.

Obwohl bereits Allen 1970/72 [3, 4, 5], Rapp 1971 [185] und Huggins 1972 [91, 90] umfangreich erforscht hatten, inwiefern das P-center von der Silbenkomplexität abhängt, und auch schon erste Modelle entwickelten, wie man diesen Zeitpunkt im Sprachsignal vorhersagen kann, gerieten erst mit den vielzitierten Veröffentlichungen von Marcus 1976 [132], Morton, Marcus & Frankish 1976 [148] sowie Terhardt & Schütte 1976 [205] Untersuchungen zum P-center für längere Zeit in den Fokus wissenschaftlichen Interesses.[4]

Terhardt & Schütte untersuchten das Phänomen unabhängig von der erstgenannten Forschergruppe und auch bei nichtsprachlichen Schallereignissen. Sie bezeichneten es in ihrer eigenen, generell psychoakustisch ausgerichteten, deutschen Terminologie sehr treffend als *Ereigniszeitpunkt*.[5] Nach ihren Untersuchungen kann bei nichtsprachlichen Stimuli (Sinustonimpulse oder Impulse mit Weißem Rauschen) der Ereigniszeitpunkt gegenüber dem physikalisch meßbaren Geräuscheinsatz um bis zu 80 ms später auftreten.

Dieser Wert liegt weit über dem gerade wahrnehmbaren Verlassen der Isochronie eines Rhythmus, das bei Intervallen mit weniger als 200 ms Dauer unter 8 ms liegt und bei größeren Intervallen bis zu 1200 ms etwa 3% bis 6% des Intervalls beträgt (Friberg & Sundberg 1995 [58]). Bei Experimenten, in denen zwischen lediglich zwei aufeinanderfolgenden Intervallen in einem Bereich bis zu 2000 ms Dauer verglichen wird, ist die Wahrnehmungsschwelle mit etwa 10% deutlich größer (Woodrow 1951 [242]).

Im Rahmen von Untersuchungen zur zeitlichen Steuerung des artikulatorischen Verhaltens (Fowler 1977 [54]) wurde bereits sehr früh die Hypothese aufgestellt, daß P-centers durch das zugrundeliegende Artikulationsverhalten determiniert werden (Fowler 1979 [55]). So wurde in einer Untersuchung von Tuller & Fowler 1980 [216] perzeptive Isochronie, die bekanntermaßen nicht mit akustischer Isochronie einhergeht, auf der artikulatorischen Ebene mit Hilfe elektromyographischer Messungen der Lippenmuskulatur untersucht. Dabei stellte sich heraus, daß perzeptive Isochronie tatsächlich von artikulatorischer Isochronie begleitet wird. Hiermit wird auch ein Hinweis auf die enge Beziehung zwischen Produktion und Perzeption bei der rhythmischen Organisa-

[4] Siehe weiterhin zur P-center-Forschung auch Kohler 1981 [104], Köhlmann 1982/84 [110, 111], Pompino-Marschall 1989 [175], Janker 1989 [94] und Kühnert 1989 [115].

[5] Aufgrund der unterschiedlichen Termini nahmen sich die beiden Forschergruppen lange Zeit gegenseitig nicht wahr, obwohl der Begriff des *Ereigniszeitpunkts* der Übersetzung des von Marcus verwendeten Begriffs des *moment of occurrence* entspricht.

tion von Sprache gegeben. Einen weiteren Hinweis liefern die Experimente von Kohler, Schäfer, Thon & Timmermann 1981 [109], die eine enge Verbindung der rhythmischen Struktur in der Produktion mit der Geschwindigkeitswahrnehmung nahelegen.

Tuller & Fowler 1981 [217] führten weiterhin perzeptive Untersuchungen zum Einfluß des Amplitudenverlaufs von Logatomen auf die Wahrnehmung des P-centers durch. Diese ergaben, daß ein durch *infinite peak clipping* in eine nahezu rechteckig verlaufende Hüllkurve gezwungenes Sprachsignal keine andere P-center–Wahrnehmung hervorruft als das unveränderte Signal. Der maximale Anstieg der Amplitude, der sich durch diese Signalmanipulation immer an den Anfang einer Silbe verschob, konnte also nicht ausschlaggebend für die P-centers sein.

Janker 1995 [95] widersprach diesem Ergebnis. Er führte sechs Adjustierungsexperimente mit nur drei Probanden[6] und ein Mittastexperiment mit 30 Probanden durch, von denen er allerdings neun nicht in die Endergebnisse einbezog. Aufbauend auf diesen Untersuchungen vertritt er die Hypothese, daß P-centers in gesprochener Sprache besser mit Konsonant-Vokal–Übergängen korrelieren, als mit den Zeitpunkten, die durch Modelle von z.B. Marcus 1976 [132] oder auch Pompino-Marschall 1990 [176] entwickelt wurden.

Janker greift damit die ein Jahrhundert alten Überlegungen von Meyer 1898 [138, S.490f] auf, der diesem Übergang aufgrund der „besonderen Energieausgabe“ beim Wechsel der Artikulatorenbewegungsrichtung den Moment der „Wendung“ zuspricht. Auch durch Taktschlag-Experimente mit zwei Probanden konnte Meyer schon damals feststellen, daß sich der resultierende Taktschlag bei rezitierten Versen nahe „der Grenze zwischen anlautendem Konsonanten und Vokal“ befindet und damit ebenfalls jenen „Moment der Wendung“ trifft [138, S.127]. Seit Jankers Arbeiten ist es um P-centers in der Wissenschaft sehr viel ruhiger geworden.

3.2 Rhythmus

Der Begriff des Rhythmus war in der Vergangenheit Gegenstand zahlreicher wissenschaftlicher Abhandlungen,[7] vollständiges Einvernehmen hat sich indes nicht eingestellt. Dabei zeigten Demany, McKenzie und Vurpillot schon 1977 [37], daß Neugeborene bereits im Alter von 71 ± 12 Tagen rhythmische Strukturen unterscheiden können, komplexe Rhythmen aber erst später in ihrer Weiterentwicklung erkennen und dann auch wiedergeben können. Und gerade im Bereich musikalischer Rhythmen wird sehr schnell offenkundig, daß kulturelle Unterschiede das Rhythmusempfinden ganz wesentlich prägen.[8]

Demnach liegen für das Phänomen des Rhythmus Erklärungsversuche unter Wahrnehmungsaspekten geradezu auf der Hand, während rein physikalische Ansätze einstweilen als inadäquat eingestuft werden müssen. Insbesondere unter der Annahme gestaltpsychologischer Konzepte wird Rhythmuswahrnehmung plausibel.

Zu den wichtigsten Begründern der Gestaltpsychologie zählt Wertheimer, der in zahlreichen Perzeptionsexperimenten (1912/1923 [237, 238]) die Existenz und Gültigkeit der Gestaltgesetze *Einfachheit, Nähe, Ähnlichkeit, gute Fortsetzung, gemeinsames Schicksal, Symmetrie* und *restlo-*

[6] Janker selbst befindet sich unter den drei Probanden und auch die anderen beiden Probanden (Bernd Pompino-Marschall und Barbara Kühnert) hatten bereits zuvor auf dem Gebiet der P-centers intensiv geforscht.

[7] Siehe hierzu etwa Sachs 1953 [191], Fraisse 1982 [56], Handel 1989 [73], Couper-Kuhlen 1993 [25] und Auer, Couper-Kuhlen & Müller 1999 [9]. Rhythmus in der Schriftsprache, wie z.B. von Lösener 1999 [129] behandelt, ist nicht Gegenstand dieser Untersuchung.

[8] Für Europäer sind z.B. die komplexen Rhythmusstrukturen der indonesischen Gamelanmusik aus z.B. Java oder Bali nur schwer faßbar.

se Verarbeitung belegte. Es sollte nicht verschwiegen werden, daß die frühe Gestaltpsychologie hinsichtlich ihres eher beschreibenden als erklärenden Charakters und hinsichtlich konkurrierender Gestaltgesetze mit undefinierten Prioritäten kritisiert wird. Anhänger der Gestaltpsychologie versuchen mit Hilfe der genannten Gestaltgesetze, die Wahrnehmungsphänomene der Objekterkennung und -gruppierung zu erklären, und greifen dabei wieder die aristotelische These auf, daß „das Ganze mehr ist als die Summe seiner Teile“.[9]

Aus gestaltpsychologischer Sicht ist Rhythmus also als ein Wahrnehmungsinhalt aufzufassen. Und damit kann er *mehr sein* als die rein physikalisch gemessene Reizgrundlage und auch ihre auf die wahrnehmbaren akustischen Ereignisse abstrahierte und durch Anzahl und Zeitpunkte der *P-centers* bzw. *Ereigniszeitpunkte* vollständig beschriebene Form.

Die Anwesenheit von Rhythmus in der gesprochenen Sprache war unter Sprachwissenschaftlern zu keiner Zeit strittig (Cruttenden 1997 [27, S.13ff]). Dennoch scheiterten bis heute alle Untersuchungen zum Sprechrhythmus an dem Ziel, aufzudecken und schlüssig darzulegen, *wie* sich Rhythmus manifestiert, da er nicht offensichtlich vorliegt und bisher mit keiner meßtechnischen Methode nachgewiesen werden konnte.[10]

Sprechrhythmus scheint sich im zyklischen Wechsel zwischen den P-centers starker und schwacher Silben widerzuspiegeln. Üblicherweise wird angenommen, daß man die Abfolge der Silben jeder natürlichsprachlichen Äußerung *akzentzählender*[11] Sprachen in eine Fußstruktur[12] gliedern kann: Nach dieser Theorie folgen etwa im Deutschen einer starken Silbe entweder eine oder zwei schwache Silben (*Trochäus* oder *Daktylus*). Vennemann 1995 [225] geht davon aus, daß eine der Sprachgemeinschaft nicht von außen aufgedrängte, sondern in ihr über längere Zeiträume entstandene und damit natürliche poetische Metrik lediglich Sprachzüge stilisiert, die auch der Alltagssprache angehören. In der poetischen Metrik wird diese Fußstruktur schnell deutlich, was für Vennemann ein eindeutiges Indiz für deren Präsenz auch in der Alltagssprache ist. Unter Annahme dieser Fußstruktur wird dann leicht nachvollziehbar, daß Füße auch als *Takte* bezeichnet werden. Über diese ist wiederum aus der Musik bekannt, daß ihre innere Zeitstruktur gemeinsam mit ihrer sukzessiven Abfolge wiederum Rhythmus konstituiert.

Eine zeitliche Struktur sollte allerdings nur dann als *rhythmisch* bezeichnet werden, wenn ein zusätzliches Kriterium erfüllt ist: Rhythmus liegt vor, sobald man aufgrund des Wahrgenommenen *vorhersagen* kann, was folgt — wenn man also durch *Synchronisation* mit der Geschwindigkeit und rhythmischen Struktur eines Signals in die Lage versetzt wird, die Schläge des Rhythmus zu antizipieren (Fraisse 1982 [56, S.150]). Wird dieses Kriterium nicht erfüllt, so erscheint eine zeitliche Struktur demnach nicht rhythmisch.

Außerdem werden wiederholte akustische Ereignisse bei Intervallen oberhalb von ca. 1800 ms nicht mehr gruppiert und zu einem Rhythmus gehörig wahrgenommen. Dementsprechend sind Probanden dann auch nicht mehr in der Lage zu antizipieren, sondern statt dessen *reagieren* sie auf jedes der Ereignisse. Auch Intervalle unterhalb von etwa 120 ms — was mehr als 500 Schlägen pro Minute (BPM) bzw. mehr als 8.3 Hz entspricht — sind nicht mehr als Rhythmus bildend wahrnehmbar, sondern am ehesten noch als ein regelmäßig „ratterndes“ Geräusch.

---

[9] Neuere Darstellungen der Gesetze finden sich u.a. bei Bruhn, Oerter & Rösing 1993 [13, S.469ff], Goldstein 1997 [64, S.168ff] und Cook 1999 [24, S.33f].

[10] Daß der scheinbar offensichtliche Rhythmus selbst bei prosaischem Sprechen, skandierendem Verssprechen und Musik kaum systematisch greifbar ist, wird anhand der Versuche von Nord, Kruckenberg & Fant 1989 [155] deutlich.

[11] Dieser Begriff wird erst im folgenden Abschnitt in Fußnote 14 auf S. 147 erklärt.

[12] Der Begriff des *Fußes* stammt aus der traditionellen Versmetrik und beschreibt die kleinste rhythmische Einheit, die aus einer starken und einer oder mehreren schwachen Silben besteht. Die starke Silbe ist der Kopf des Fußes. In der literatur- und sprachwissenschaftlichen Sprachrhythmusforschung ist der Begriff des Fußes allgegenwärtig.

3.3 Isochronie

Über den Rhythmus hinausgehend unterstellt die Isochronie-Theorie der gesprochenen Sprache gleichbleibende Zeitintervalle zwischen bestimmten aufeinanderfolgenden linguistischen Einheiten. So sollen beispielsweise im Deutschen aufeinanderfolgende Füße *ungefähr* isochron sein. Selbst in dieser deutlich abgemilderten Formulierungsweise konnte sie auf der Grundlage von Aufzeichnungen natürlich gesprochener Sprache nur in den seltensten Fällen und auch dann nur über kurze Sprechabschnitte nachgewiesen werden (Lehiste 1977 [122]).[13] Diese ganz grundsätzliche Erkenntnis lieferten aber schon mehrere Untersuchungen im späten 19. Jahrhundert anhand von Messungen der Fußdauern in rezitierten Versen (vgl. Scripture 1902 [195, S.537ff]).

Auer, Couper-Kuhlen & Müller 1999 [9, S.51] führen als Beispiel zwei von ihnen transliterierte und segmentierte spontansprachliche Äußerungen auf, anhand derer sie Isochronie demonstrieren können. In diesen Fällen liegen die Dauerabweichungen aufeinanderfolgender Füße bei -3% bzw. -1% und damit weit unterhalb der Wahrnehmungsschwelle für Intervallunterschiede von etwa 10% (Woodrow 1951 [242]). Allerdings messen sie ansonsten bei den von ihnen als rhythmisch eingestuften Sprechabschnitten Intervallunterschiede von bis zu 35%, die sehr deutlich über dieser Wahrnehmungsschwelle liegen. Zwar gehen Auer et al. davon aus, daß bei perzeptiv isochronen Silben Intervallmessungen zwischen aufeinanderfolgenden Vokalanfängen bessere Ergebnisse liefern sollten als Messungen von einem Silbenbeginn zum nächsten, da im letzteren Fall einbezogene silbeninitiale Konsonanten-Cluster aufgrund ihrer unterschiedlichen Komplexität die Messungen verzerren würden. Dennoch haben sie keine anderen Messungen durchgeführt, denn sie sehen P-centers als noch zu wenig verstanden an [9, S.53].

Das Konzept der Isochronie wird trotz des bisherigen Scheiterns eines schlüssigen empirischen Nachweises nach wie vor als wichtig zur Unterscheidung zwischen den sog. *akzentzählenden*[14], *silbenzählenden*[15] und *morenzählenden*[16] Sprachen erachtet.

Pike 1945 [174] prägte die Begriffe *stress-timed* (akzentzählend) und *syllable-timed* (silbenzählend). Aus seiner Sicht ist Englisch akzentzählend, denn obwohl seiner Erfahrung nach im Englischen beide Rhythmustypen anzutreffen sind, kommen silbenzählende Passagen doch nur sehr selten vor. Lloyd James 1940 [128] fand deutlich vor Pike mit den Begriffen *morse-code rhythm* bzw. *machine-gun rhythm* nachvollziehbare Bilder für die beiden Rhythmustypen.[17]

Jones [96] hatte bereits in den frühsten Auflagen seines Buches *An Outline of English Phonetics* seit 1918 detailliert den Rhythmus im Englischen beschrieben. Dabei legte er großen Wert auf folgende Feststellung bezüglich der von ihm angegebenen Dauerverhältnisse: „[They] are not the lengths of the syllables but the lengths separating the ‘stress-points’ or ‘peaks of prominence’ of the syllables.“ [96, S.240]

---

[13] Z.B. kann Uldall 1978 [221] in keiner der zwei Aufzeichnungen von „*The North Wind and the Sun*“, die David Abercrombie in normalem bzw. sehr schnellem Sprechtempo gelesen hat, meßtechnisch Isochronie nachweisen, während sie allerdings noch 1971 [220] bei der alleinigen Betrachtung der mit normalem Sprechtempo gelesenen Version eine starke Tendenz zur Isochronie sieht.

[14] Hierzu zählen u.a. Englisch, Niederländisch und Deutsch. Aber auch vom Russischen und Arabischen wird behauptet, daß sie zu den akzentzählenden Sprachen gehören. Grundsätzlich nimmt man für diesen Sprachrhythmus an, daß die Zeitintervalle zwischen den Akzenten zur Isochronie tendieren (Abercrombie 1967 [1]).

[15] Als Beispiele seien das Französische, Spanische, Tschechische und Ungarische genannt. Bei diesen Sprachen sollen die Silbenintervalle Isochronie anstreben (Auer & Uhmann 1988 [10]).

[16] Bisher wurden nur wenige Sprachen als morenzählend eingestuft. Darunter sind Japanisch und auch Finnisch sowie Slovakisch (Trubetzkoy 1939 [213]). CV-Silben entsprechen in der Regel einer Mora. Weisen Silben aber Langvokale, mehrere Konsonanten oder Geminaten auf, werden sie mit zwei Moren gezählt (Warner & Arai 2001 [233]).

[17] Wir wollen diese belasteten Begriffe hier nur der Vollständigkeit halber erwähnen, aber nicht weiter verwenden.

Zweifellos hat er damit ohne große Umschweife die Zeitintervalle zwischen den P-centers als ausschlaggebend für den Rhythmus in der Sprache bezeichnet und auf diese Weise einen gerade heute wieder in Betracht gezogenen und oben erwähnten Ansatz (vgl. Auer, Couper-Kuhlen & Müller 1999 [9, S.53]) schon damals geliefert.

3.4 Messungen von sprachindividuellen Rhythmen

Daß das Isochronie-Konzept Sprachrhythmen hinreichend differenzierbar machen soll, wurde in der Vergangenheit schon oft bezweifelt. So zeigen Wenk & Wioland 1982 [236], daß die bisher als geradezu paradigmatisch silbenzählend geltende Sprache Französisch keineswegs isochrone Silben aufweist.

Auch Dauer widerspricht 1983 [34] der These, daß die akzentzählende Sprache Englisch zur Isochronie bei akzentuierten Silben tendiert, wohingegen die silbenzählende Sprache Spanisch eher isochrone Silben aufweisen soll. Ihren Untersuchungen zufolge finden sich keine signifikant unterschiedlichen Standardabweichungen zwischen Akzentintervallen im Englischen vs. Spanischen. Nach ihrer Ansicht spiegelt sich Rhythmus vielmehr in der Silbenstruktur, in Reduktionsprozessen und in der phonetischen Realisierung von Akzent wider. So sind betonte, nichtfinale, offene Silben im Spanischen nur 10% länger als unbetonte, im Englischen 60% und im Deutschen 50% (Delattre 1966 [36]).[18]

Entsprechend konstruiert Dauer 1987 [35] ein Modell zur Bestimmung des Rhythmustyps einer Sprache, das nur auf vergleichsweise einfach durchzuführende Messungen angewiesen ist und das Gesamtergebnis in Teilentscheidungen über An- („+“) oder Abwesenheit („−“) von einzelnen Komponenten zerlegt. Ist eine klare Entscheidung bezüglich einer Komponente nicht möglich, so wird eine „0“-Marke vergeben. Im einzelnen wird geprüft, ob *i)* betonte Vokale mindestens 50% länger sind als unbetonte, *ii)* viele unterschiedlich komplexe Silbenstrukturen auftreten, *iii)* Akzente mit dem Intonationsverlauf korrelieren, *iv)* Vokale in unakzentuierten Silben reduziert sind, *v)* Konsonanten ebenfalls in unakzentuierten Silben stärker reduziert werden und *vi)* ob der Akzent sich im Wort verschieben kann.

Dimitrova wendet 1998 [40] die Methode von Dauer an, um zu bestimmen, ob Bulgarisch akzent- oder silbenzählend ist. Da die einzelnen Komponenten des Modells von Dauer gleichviele „+“, „−“ und „0“ liefern, stuft Dimitrova den Sprachrhythmus des Bulgarischen als genau zwischen den beiden Klassen liegend ein.

Auch Ramus, Nespor & Mehler 1999 [184] unterlassen Messungen von Fußdauern und entwickeln eine neue Methode, die die Segmentation des Sprachsignals in lediglich vokalische vs. nicht-vokalische Bereiche erfordert. Dann wird der prozentuale vokalische Anteil am Gesamtsignal (%V), die Standardabweichung der Dauern vokalischer Bereiche (ΔV) und die der nichtvokalischen Bereiche (ΔC) bestimmt. Anhand dieser drei Größen sehen sie sich in der Lage, die Sprachrhythmen von vier (Englisch, Niederländisch, Polnisch und Japanisch) der acht[19] untersuchten Sprachen auseinanderzuhalten. Allerdings muß kritisch angemerkt werden, daß die zugrundeliegenden Meßgrößen nur Eigenschaften von Silbenstrukturen widerspiegeln, anhand derer sich zweifellos viele Sprachen unterscheiden lassen. Aber über den *Rhythmus* der jeweiligen Sprache kann mit dieser Methode genaugenommen nichts ausgesagt werden.

[18] Siehe auch Lisker 1974 [127], Allen 1975 [6], Huggins 1975 [92] und Kohler 1983 [105].

[19] Die anderen vier Sprachen (Französisch, Italienisch, Katalan und Spanisch) unterscheiden sich aufgrund der drei vorgeschlagenen Parameter nicht gravierend.

Deterding 2001 [39] verwendet zur Rhythmusmessung den sog. „variability index“ (*VI*). Dieser wurde von Low 1994 [130] eingeführt und entspricht dem Mittelwert über vorzeichenlose Silbendauerunterschiede jeweils zweier benachbarter Silben.[20] Zuvor führt Deterding eine Sprechgeschwindigkeitsnormalisierung durch, indem er alle Silbendauern durch die mittlere Silbendauer des jeweiligen Sprechers und der jeweiligen Äußerung dividiert. Weiterhin schließt er äußerungsfinale Silben aus. Ein kleinerer *VI* deutet auf ähnlichere Silbendauern bei sukzessiven Silben und damit auf einen stärker silbenzählend ausgeprägten Sprachrhythmus hin. Am Rande seien seine Ergebnisse zitiert, die zeigen, daß Singapur-Englisch über einen stärker silbenzählenden Charakter als das britische Englisch verfügt ($VI_{SE} = 0.448 < VI_{BE} = 0.565,\ p < 0.001$).

Abschließend sei betont, daß wir Isochronie für ein obligatorisches Merkmal von Rhythmus halten. Unter dieser Annahme verbietet die Abwesenheit von Isochronie den Gebrauch des Terminus „Rhythmus“. Wenn also in neueren Ansätzen wie z.B. dem von Ramus et al. 1999 [184] nur die Silbenstruktur einer Sprache, nicht aber der Grad an Isochronie berücksichtigt und beschrieben wird, dann sollte auf Aussagen über den Sprachrhythmus verzichtet werden.

3.5 Metrische Phonologie

Auch in der Phonologie gibt es Erklärungsansätze zur zeitlichen Struktur von Sprache. Hier gilt vor allem die Metrische Phonologie als wegweisend. Sie wird neben der Autosegmentellen Phonologie zur nichtlinearen Phonologie gezählt und trägt der Einsicht der Phonologen in den 70er Jahren Rechnung, daß der Akzent weder eine dem Segment inhärente Eigenschaft ist, noch als ein *absolutes* Merkmal beschrieben werden kann. Vielmehr drückt sich Akzent in der Ausgeprägtheit von Prominenzdifferenzen zwischen benachbarten Silben aus. Daher ist er auch nicht anhand des distinktiven Merkmals [$\pm$betont] adäquat und sinnvoll zu erfassen, was allerdings noch in der Generativen Phonologie wie z.B. bei Chomsky & Halle 1968 [23] versucht wurde. Ihren Ursprung hatte die Metrische Phonologie in einer Dissertation von Liberman 1975 [124] und wurde im Laufe weiterer Veröffentlichungen von Liberman & Prince 1977 [125], Halle & Vergnaud 1978 [72] und Hayes 1981/84 [75, 76] ausgebaut und weiterentwickelt.

Die Metrische Phonologie beschäftigt sich mit rhythmischen Phänomenen, dabei insbesondere mit Akzent, und sieht den Fuß als zentrale Einheit metrischer Sprachstrukturen. Die Prominenz eines Akzents wird nicht absolut angegeben, sondern immer in Relation zu anderen Silben, und kann entweder *stark* oder *schwach* sein. Metrische Strukturen linguistischer Formen werden im *metrischen Baum* analysiert und dargestellt, dessen Wurzel meist ein Wort ist, aber auch eine Phrase oder eine Äußerung sein kann. Den Knoten entsprechen Füße und darunter Silben; schließlich sind die Blätter die Laute jenes Wortes. Die zweite wichtige Darstellungsart von metrischen Sprachstrukturen ist das *metrische Gitter*, wobei allerdings verschiedene Autoren ihre ganz persönlichen Methoden der Gitterrepräsentation entwickelt haben. Hayes betont 1984 [76], daß für die Repräsentation von Akzent die Darstellung im metrischen Baum notwendig ist, während für die rhythmische Struktur die Analyse im metrischen Gitter besser geeignet ist.

Die Metrische Phonologie erhebt nicht den Anspruch, sprachsignalnah werden zu können; mehr noch, sie zieht nicht einmal die Möglichkeit in Erwägung. Vielmehr bewegt sie sich auf einem vergleichsweise hohen Abstraktionsgrad weit weg vom Sprachsignal und strebt dabei an, die rhythmische und die Akzentstruktur der Sprache von Wörtern bis Äußerungen auf einer rein symbolischen Ebene zu beschreiben und zu erklären. Daß dabei einerseits die temporale Realität

[20] Siehe auch Low, Grabe & Nolan 2000 [131].

gesprochener Sprache kaum vorhergesagt werden kann und andererseits auch für Messungen von Rhythmus bei gesprochener Sprache keinerlei verwertbare Informationen geliefert werden können, wird u.a. von Couper-Kuhlen 1993 [25] kritisch diskutiert. Offensichtlich berücksichtigt die Metrische Phonologie hauptsächlich morphosyntaktische Merkmale und läßt wesentliche — zugegebenermaßen aufwendig aus der orthographischen Form zu extrahierende — phonologische und prosodische Information (z.B. *pitch-accent*) unberücksichtigt. Deren Einbeziehen ist allerdings notwendig, um die tatsächliche prosodische Struktur realer Äußerungen vorherzusagen, und hätte möglicherweise einen praktischen Einsatz in der Sprachsignalanalyse und -synthese gestattet.

3.6 Spontanes Tempo

Auf der Suche nach der biologischen Uhr des Menschen, die auch als Taktgeber für rhythmische motorische Aktivitäten wie z.B. Nuckeln, Schaukeln oder Gehen, aber auch Atmen oder den Herzschlag angenommen wurde, ergab sich auch eine Reihe von Produktionsexperimenten. So sollten Probanden spontan einen Rhythmus erzeugen durch z.B. Tippen mit dem Finger auf eine Tischplatte, Schlagen mit der Handfläche auf den Oberschenkel, Schwingen mit dem Bein, u.ä. Dabei stellte sich heraus, daß die spontan gewählten Tempi von ein und derselben Versuchsperson bei verschiedenen Arten der Produktion sowie bei mehrfachen Wiederholungsexperimenten im Bereich von 0.75 bis 0.95 miteinander korrelierten, während die Tempi unterschiedlicher Probanden kaum korrelierten.

Dieses Tempo wird als *spontanes Tempo* oder auch *persönliches Tempo* bzw. *mentales Tempo* bezeichnet. Die Intervalle zwischen aufeinanderfolgenden Impulsen liegen je nach Proband in einem Bereich zwischen 200 ms und 1400 ms (300 BPM bis 43 BPM) mit einem die Verteilung am besten repräsentierenden Wert von 600 ms, welcher 100 BPM entspricht (Fraisse 1982 [56]).

Obwohl sich der Herzrhythmus mit einer mittleren Intervalldauer von etwa 800 ms in einer ähnlichen Region befindet, zeigt sich bei Versuchspersonen keine Korrelation mit ihrem spontanen Tempo. Auch erhöht sich das spontane Tempo nicht, wenn der Herzschlag beschleunigt wird. Demnach darf hier kein Zusammenhang unterstellt werden.

In Bezug auf Rhythmus bekommt das spontane Tempo aber eine besondere Bedeutung, da bei der Darbietung von polyrhythmischem akustischem Testmaterial eine Versuchsperson denjenigen Rhythmus mitklopft, der ihrem individuellen spontanen Tempo am nächsten kommt.

Das spontane Tempo muß übrigens vom sog. *bevorzugten Tempo* unterschieden werden. Um letzteres zu messen, wird in Perzeptionsexperimenten nach derjenigen Geschwindigkeit einer Ton- oder Lichtfolge gefragt, die weder als zu schnell noch als zu langsam empfunden wird. Auch hier ergibt sich eine Intervalldauer von etwa 600 ms. Sie korreliert allerdings bei Versuchspersonen kaum mit ihren jeweiligen spontanen Tempi ($r = 0.40$).

3.7 Psycholinguistische Modelle der Zeitverarbeitung

Im vorigen Abschnitt erfuhren wir, daß Probanden sowohl beim Produzieren als auch beim Perzipieren ein ganz individuelles Tempo wählen, das allerdings bei allen Versuchspersonen um eine mittlere Dauer von 600 ms verteilt ist. Nun stellt sich die Frage, ob wir es hier mit einer Dauer zu tun haben, die Aufschlüsse über die psychologische Zeitverarbeitung des Menschen zuläßt.

Pöppel hat 1978 [181] ein Modell der psychologischen Zeitverarbeitung entwickelt und 1979 [182] den Versuch unternommen, es auf gesprochene Sprache zu übertragen. Er geht davon aus,

daß wenigstens vier der aus seiner Sicht fünf für die menschliche Zeitwahrnehmung charakteristischen Phänomene auch in die Sprachanalyse involviert sind:

1. *Zeitliches Auflösungsvermögen*: Zwei Reize müssen einen zeitlichen Abstand von mindestens 2 ms (*Fusionsschwelle*) aufweisen, um nicht mehr als nur ein einziger Reiz wahrgenommen zu werden.
2. *Identifikation*: Erst oberhalb eines Intervalls von 20 ms (*Ordnungsschwelle*) sind Probanden in der Lage, die Reihenfolge der beiden Reize anzugeben. Aus der Tatsache, daß die Größe der Ordnungsschwelle unabhängig von der Modalität (auditiv, visuell oder taktil) ist, wird oft abgeleitet, daß ein zentraler zeitlicher Organisationsmechanismus involviert sein könnte.
3. *Sequenzierung*: Das Bilden und Reproduzieren von Ereignisketten basiert wiederum auf einem anderen temporalen Mechanismus. Dies zeigt sich bei Patienten mit Korsakowsyndrom, die zwar die wahrgenommenen Ereignisse wiedergeben, dabei aber nicht mehr die richtige Reihenfolge einhalten können.
4. *Zeitliche Integration*: Die einzelnen Elemente einer Ereigniskette werden in einer höheren syntaktisch-semantischen Struktur organisiert, wobei ihre rein serielle Anordnung an Bedeutung verliert. Das Zeitintervall dieser Ebene hat eine Dauer von etwa 2 bis 3 Sekunden und wird auch als „subjektives Jetzt“ oder „Präsenzzeit“ bezeichnet.

Kegel berichtete 1990 [99] über eigene im Rahmen eines DFG-Projekts (Kegel, Dames & Veit 1988 [100]) durchgeführte, umfangreiche psycholinguistische Experimente zur Wahrnehmung der zeitlichen Struktur von Sprache, die neue Erkenntnisse bezüglich Funktion und Entwicklung der Zeitverarbeitung — auch und gerade während des Spracherwerbs — ermöglichen sollten. Dabei kommt er zu einem Modell, das im Vergleich zum gerade dargestellten Modell von Pöppel nur drei Ebenen aufweist und sich allein auf die Sprachverarbeitung konzentriert.

Es differenziert zwischen erstens der *Ordungsebene* mit einer Fenstergröße zwischen 20 und 40 ms, auf der akustische Merkmale wie z.B. Formanttransitionen verarbeitet werden, zweitens der *Strukturierungsebene* mit einer Dauer zwischen 100 und 500 ms, die Silben verarbeitet, denen die akustischen Merkmale aus der Ordnungsebene zugeordnet sind, und drittens der *Integrationsebene* mit 2 bis 5 Sekunden Dauer, auf der prosodische Strukturen und auch emotionale Informationen erkannt und Sätzen oder Satzkonstituenten zugeordnet werden.

Um die Wahrnehmung von Zeitstrukturen in der Sprache zu analysieren, präsentierten Kato, Tsuzaki & Sagisaka 1997 [98] sechs Probanden 435 Stimuli, die auf 15 viersilbigen japanischen Wörtern basierten. Im einzelnen wurde die Dauer des Konsonanten oder/und Vokals der ersten, zweiten oder dritten Silbe um 15 bzw. 30 ms gestaucht oder gedehnt. Auf die Frage, wie akzeptabel die jeweilige zeitliche Manipulation war, zeigte sich, daß die Versuchspersonen die gemeinsame Dehnung oder Stauchung von zwei Segmenten als etwa doppelt so inakzeptabel beurteilen wie die nur eines Segments. Wurde dagegen der Konsonant um denjenigen Betrag verlängert, um den der benachbarte Vokal gekürzt wurde (oder umgekehrt), so ergab sich ungefähr die gleiche Akzeptanz wie bei der Manipulation nur eines Segments.[21]

Die Autoren belegen damit die Existenz einer perzeptiven Dauerkompensation zwischen benachbarten Lauten im Rahmen des silbischen Sprechrhythmus und schließen daraus, daß das Fenster, innerhalb dessen die zeitliche Struktur gesprochener Sprache kognitiv ausgewertet wird, wenigstens der Dauer einer Silbe entspricht.

[21] Vergleichbare Ergebnisse präsentierte Huggins bereits 1972 [91] für das Englische.

3.8 A-, B- und C-Prosodie

Tillmann prägte 1980 [210, S.39f, S.108ff] die Begriffe *A*-, *B*- und *C-Prosodie*.[22] Dabei wurde er inspiriert durch eine Feststellung von Scripture 1929 [196], daß ein Ton, dessen Tonhöhe oder Lautstärke sich sowohl *stetig* als auch *gleichmäßig* in eine Richtung ändert, als ein einziges Klangereignis wahrgenommen wird, „vorausgesetzt, daß die Geschwindigkeit der Änderung ein gewisses Maß nicht überschreitet." [196, S.25] Tillmann fügte die *zyklische Variation der Geschwindigkeit und Richtung der Änderung* hinzu, also etwa eine sinusförmige periodische Modulation der Tonhöhe bzw. Lautstärke eines Tones, denn dann „hängt das Wahrnehmungsereignis sehr wesentlich von dieser [Modulations-]Geschwindigkeit ab." [210, S.39]

Beispielsweise bietet eine heulende Sirene, deren Tonhöhe mit einer Geschwindigkeit von 1 Hz oder weniger moduliert wird, einen Tonhöhenverlauf, den die menschliche Wahrnehmung unmittelbar beobachten und verfolgen kann (Tillmann nennt dies den *A-Fall*). Das ändert sich bereits bei einer Verdopplung auf 2 Hz, also bei einer Änderung der Periodendauer von 1000 ms auf 500 ms: nun beginnen die Tonhöhenmaxima oder -minima, zu einer Folge von rhythmischen Schlägen zu werden. Der an sich immer noch kontinuierliche Tonhöhenverlauf wird zu schnell, als daß er unmittelbar wahrgenommen werden könnte; dennoch sind die rhythmischen Schläge zählbar und auch mitklopfbar (*B-Fall*). Bei weiterer beträchtlicher Erhöhung der Modulationsfrequenz auf etwa 10 Hz und mehr verliert sich die Empfindung einer Folge von rhythmischen Schlägen und geht über in die Wahrnehmung einer Triller-Struktur, bei der die einzelnen Elemente des Trillers weder zählbar noch mitzuklopfen sind (*C-Fall*).

Tillmann überträgt diese klar unterscheidbaren Fälle der Wahrnehmung auf die zeitliche Strukturierung von gesprochener Sprache, indem er den unmittelbar beobachtbaren Verlauf der Sprechmelodie *A-Prosodie* nennt, da er sich naturgemäß über mehrere Silben erstreckt und damit in seiner zeitlichen Ausdehnung mit dem oben beschriebenen *A-Fall* vergleichbar ist. Als *B-Prosodie* bezeichnet er die Variationen der Ausprägung und Akzentuierung von Silbe zu Silbe, die wesentlich schneller ablaufen als die A-Prosodie und damit den rhythmischen Charakter des *B-Falls* zeigen. Seiner Ansicht nach tritt der *C-Fall* nicht nur bei Vibranten ein, deren Geschwindigkeit aufgrund der durch Masse, Kraft und Federkonstante der artikulierenden Organe definierten ballistischen Eigenschaften determiniert wird; *alle* konsonantischen Bewegungen der einzelnen Artikulatoren liefern die *C-Prosodie* und machen die silbischen Einheiten differenzierbar [210, S.112].

Bemerkenswert ist, daß diese drei von Tillmann definierten Prosodien den drei von Kegel 1990 [99] experimentell bestimmten Verarbeitungsebenen ungefähr zugeordnet werden können und damit nicht nur durch das von Tillmann vorgeschlagene Modulationsexperiment unmittelbar erfahrbar und reproduzierbar sind, sondern darüber hinaus durch die Forschungsergebnisse von Kegel zusätzliche psycholinguistische Relevanz gewinnen.

Auch die aus der Psychoakustik (Zwicker & Feldtkeller 1967 [244]) bekannten Ergebnisse bezüglich des gerade wahrnehmbaren Amplituden- und Frequenzmodulationsgrads lassen sich zumindest hinsichtlich des Unterschieds zwischen *B*- und *C-Prosodie* interpretieren: Sie zeigen, daß unser Gehör in einem Bereich um 4 Hz besonders empfindlich für Modulationen ist. „Diesem Bereich folgt ein zweiter [. . .] in der Umgebung von 30 Hz [. . .]. [Hier] ist die Amplitudenschwankung nur als Rauigkeit des Tones zu erkennen." [244, S.167] Ganz offensichtlich geben Zwicker & Feldtkeller die Mitten der Bereiche an während wir oben die Grenzen angedeutet haben.

[22] Siehe zum Begriff der Prosodie Fußnote 11 auf S. 132.

4

Theoretische Vorüberlegungen

Die Darstellung der relevanten Forschungsliteratur in den vergangenen Kapiteln erlaubt nun folgende Feststellung: Bisher wurden in der phonetischen Forschung keine Perzeptionsexperimente zur Sprechgeschwindigkeitswahrnehmung durchgeführt, die über eine Zuordnungsaufgabe von speziell für diesen Zweck hergestellten Sprachstimuli zu jeweils einer aus maximal fünf (Butcher 1981 [15]) bis sieben (den Os 1985 [38]) Geschwindigkeitsklassen hinausgehen. Hier besteht ganz offensichtlich Bedarf nach genaueren Versuchsmethoden, die die wissenschaftliche Suche nach akustischen Korrelaten und damit einhergehend die Modellbildung zulassen und zudem möglichst auch noch psychophysikalisch fundierten Aufschluß über den gerade wahrnehmbaren Sprechgeschwindigkeitsunterschied (*JND*) zulassen.

Bevor wir also mit eigenen empirischen Untersuchungen beginnen, sollten wir im Rahmen dieses Kapitels einige Aspekte der zuvor dargestellten Forschungsliteratur hinsichtlich ihres Einflusses auf unsere ins Auge gefaßten Experimente herausarbeiten und diskutieren.

4.1 Hierarchisches Modell der zeitlichen Struktur akzentzählender Sprachen

Abb. 4.1 zeigt unseren auf den bisher gewonnenen Erkenntnissen basierenden Versuch der Visualisierung eines hierarchischen Modells, das das Ineinandergreifen der verschiedenen zeitlichen Strukturierungsebenen von akzentzählenden Sprachen repräsentieren soll. Selbstverständlich ergeben sich bei silben- oder morenzählenden Sprachen leicht modifizierte Modelle, da Isochronie dort nicht auf der Fußebene erwartet wird.

Je höher eine Strukturierungsebene in diesem Modell angesiedelt ist, desto entscheidender sind kognitive *top-down*–Prozesse für die Identifikation der diese Ebene konstituierenden Elemente. *Top-down*–Prozesse werden erst ermöglicht durch phonologisches, lexikalisches, morphologisches, syntaktisches, semantisches und pragmatisches Wissen und lassen sich unter dem Begriff der *sprachlichen Kompetenz des Hörers* subsummieren. Dagegen manifestieren sich niedrigere Ebenen an akustischen Merkmalen, deren automatische Extraktion signalnah ist und auf weniger Wissen zurückgreifen muß als bei höheren Ebenen.

Der Abstraktionsgrad nimmt mit höheren Ebenen zu, was sich u.a. auch darin widerspiegelt, daß ein zunehmendes Maß an Interpretationsaufwand notwendig wird, um die erwarteten Strukturen wiederzufinden, während sich diese in niedrigen Ebenen noch mit einfachsten Meßmethoden nachweisen lassen. Dies bedeutet jedoch nicht, daß in absehbarer Zeit nicht auch Meßinstrumente zur Bestimmung z.B. des Rhythmus zur Verfügung gestellt werden könnten.

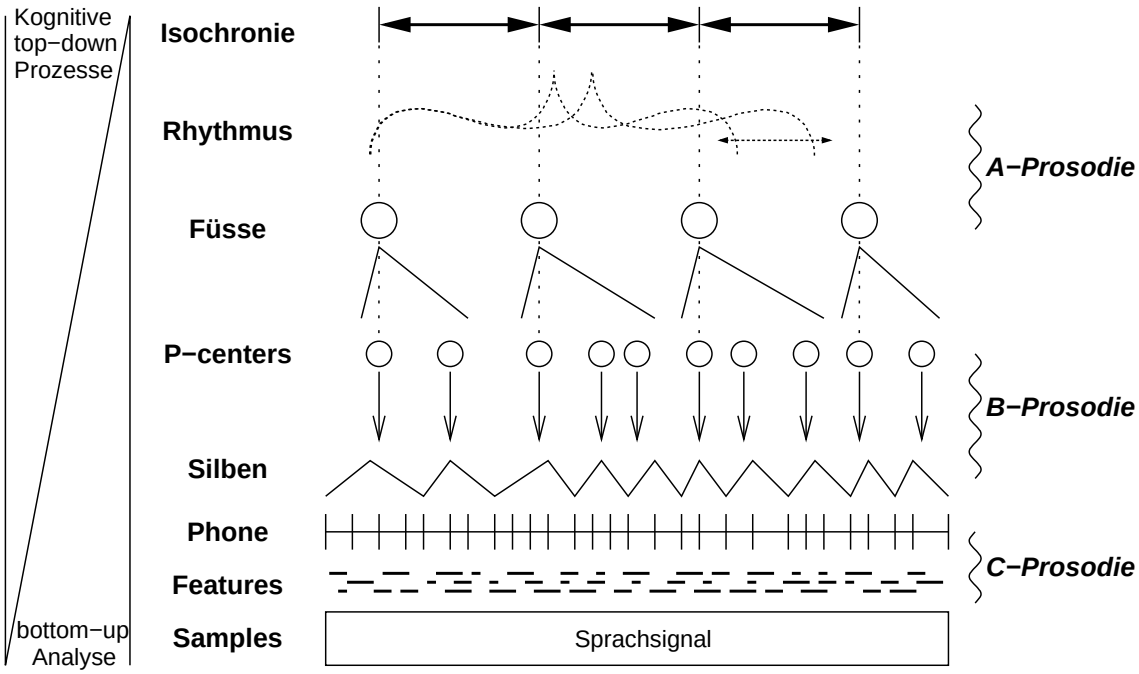

Abb. 4.1: Ein hierarchisches Modell der zeitlichen Struktur akzentzählender Sprachen, das auf der hier dargestellten Forschungsliteratur aufbaut. Sprechgeschwindigkeit spiegelt sich in allen Ebenen wider.

Weiterhin ist bei unteren Ebenen bekannt, wie groß die Zeitfenster sind, innerhalb derer Aussagen über deren Elemente eindeutig entscheidbar sind, während dies bei höheren Ebenen zunehmend unklarer wird. Insbesondere die Ebene des Rhythmus und die der Isochronie sind in ihrer minimalen zeitlichen Ausdehnung nicht klar definiert. (Dies wurde in Abb. 4.1 durch die variabel groß gezeichnete Klammer auf der Rhythmus-Ebene angedeutet.) Beide sind theoretisch auf mindestens zwei Füße angewiesen, könnten aber durchaus drei und mehr Füße umfassen.

Nun werden auch ganz selbstverständlich die Verbindungen zur A- und B-Prosodie nach Tillmann 1980 [210, S.39f, S.108ff] offensichtlich, weswegen sie im rechten Bereich von Abb. 4.1 angedeutet wurden. Gehen wir von dem Bereich der B-Prosodie aus: Silben werden in natürlicher Sprache typischerweise gerade so schnell gesprochen, daß ihre Geschwindigkeit es nicht mehr gestattet, den Verlauf an sich — etwa von einer Silbenmitte als Verlaufsmaximum über die Silbengrenze als Minimum zur nächsten Silbenmitte hin — unmittelbar zu beobachten, sondern statt dessen werden die rhythmischen Schläge, die P-centers wahrgenommen. Diese Schläge sind im gesprochenen Deutsch bekanntermaßen und sinnvollerweise nicht alle gleichermaßen ausgeprägt, sondern es gibt starke und schwache Silben, deren trochäisches oder daktylisches Alternieren die wesentlich langsamer ablaufende Fußstruktur erzeugt, die damit in den Bereich der A-Prosodie fällt und so unmittelbar beobachtbar wird. Welche der Silben oder welche Wörter einer lautsprachlichen Äußerung betont produziert wurden, ist also unmittelbar wahrzunehmen wie z.B. auch, welches Wort sich im Satzfokus befindet.

4.2 Wie „lokal“ ist Sprechgeschwindigkeit?

Sprechgeschwindigkeit ist ein prosodisches Merkmal, dessen Variation sich auf Silbenebene bereits ausprägen kann, das sich also möglicherweise schon von jeder beliebigen Silbe zur darauffolgenden verändern könnte und nicht erst an Wort-, Fuß- oder Intonationsphrasengrenzen. Dagegen

können sich komplementäre Dauervariationen zwischen benachbarten Phonen gegenseitig kompensieren, wie die in Kap. 3.7 auf S. 151 dargestellten Untersuchungen von Kato, Tsuzaki & Sagisaka 1997 [98] belegen, so daß Sprechtempo*variation* auf Lautebene und in den damit verbundenen kurzen Zeitfenstern noch nicht repräsentiert sein kann.

Ob zur Einschätzung der Sprechgeschwindigkeit die Wahrnehmung des Sprechrhythmus erforderlich ist, so daß der Silbenrhythmus auch in seiner A-prosodischen Ausprägung wahrnehmbar wird, ist an dieser Stelle aufgrund der bisher vorliegenden Forschungsliteratur nicht mit Sicherheit festzustellen. Dann wären in Perzeptionsexperimenten zur Sprechgeschwindigkeit Stimuli nötig, die mindestens zwei Füße enthalten.

Um nun bei durchschnittlichem Sprechtempo und einer damit verbundenen mittleren Silbendauer bei Lesesprache von etwa 185 ms[1] die Mindestanforderung zu gewährleisten, daß im Mittel mehr als drei Silben und damit auch etwas mehr als ein Fuß in einem Sprachstimulus auftreten, sollte die minimale Stimulusdauer für Perzeptionsexperimente zur Sprechgeschwindigkeit etwa 600 ms bis 700 ms nicht unterschreiten. Erst dann ist nämlich sichergestellt, daß wahrgenommen werden kann, wie sich die Silben in eine Fußstruktur eingliedern und welche betont und unbetont sind, wenngleich hier nur vermutet werden kann, daß diese Information in Wahrmehmungsexperimenten zum Sprechtempo nicht unterschlagen werden darf.

Das mittlere spontane Tempo liegt bei 600 ms, das bevorzugte Tempo ebenfalls. Bei Perzeptionsexperimenten zum Verstehen von normal und schnell gesprochenen Sprachstücken ermittelten Pickett & Pollack 1963 [173] eine Stimulusdauer von etwa 700 ms, über der 90% aller Wörter unabhängig von der Sprechgeschwindigkeit richtig erkannt werden. Bei kürzeren Stimuli dagegen werden signifikant weniger Wörter richtig erkannt.

Diese hier aufgezeigten Dauer-Koinzidenzen in einem Bereich zwischen 600 ms und 700 ms implizieren ohne weitere Information zwar keine Kausalitäten, geben aber diesem Intervallbereich eine besondere Bedeutung, die sich auch im Laufe der im Rahmen dieser Arbeit durchgeführten Perzeptionsexperimente als relevant herausstellen sollte.

4.3 Determinieren P-centers die Sprechgeschwindigkeit?

Aus der im vorigen Kapitel getroffenen Feststellung, daß sich Sprechgeschwindigkeit bereits auf Silbenebene ausprägen kann, läßt sich jedoch keinesfalls zwingend schließen, daß P-centers diejenigen Ereigniszeitpunkte sein müssen, die Sprechgeschwindigkeit allein und vollständig determinieren. Viel näher liegt die Annahme, daß die interne Struktur der jeweils beteiligten Silben ebenfalls aussagekräftig sein muß, da auch einsilbige Wörter, die naturgemäß nur ein P-center aufweisen, völlig problemlos in verschiedenen Sprechgeschwindigkeiten produziert und wahrgenommen werden können.

Zudem kann bei sehr langsamem Sprechen während einer Stimulusdauer von 600 ms durchaus nur eine Silbe[2] produziert worden sein, so daß dann der Fall eintritt, daß ein derartiger Stimulus nur ein P-center enthält. Sogar dann wird man die Sprechgeschwindigkeit einschätzen können, denn die ungewöhnlich lange Dauer der Silbe ist ja wahrnehmbar und auffällig.

Und auch die Rekonstruktions- und Antizipationsfähigkeiten des kognitiven Sprachwahrnehmungsprozesses könnten hier zusätzliche verwertbare Tempoinformationen generieren. Für die *Empfindung* der Sprechgeschwindigkeit sind also weder mindestens zwei P-centers noch

[1] Die durchschnittliche Silbenrate im *PhonDatII*-Korpus läßt sich aus Abb. 1.1 auf S. 124 abschätzen.

[2] Die längste im *PhonDatII*-Korpus gemessene Silbennukleusdistanz ohne Sprechpause beträgt 591.8 ms.

Silbenanfangs- und -endemarken erforderlich. Daß dagegen ein Algorithmus zur Geschwindigkeitsbestimmung, der auf physikalischen Dauermessungen linguistischer Einheiten basiert, wenigstens *ein Zeitintervall* (und damit *zwei Zeitmarken*) benötigt, liegt in der Natur dieser speziellen Meßmethode und erlaubt selbstverständlich keine Rückschlüsse auf die menschliche Geschwindigkeits- und Zeitverarbeitung.

P-centers alleine determinieren wahrgenommenes Sprechtempo also nicht vollständig, aber ihr Einfluß ist unbestreitbar und sie sind entscheidend bei der Rhythmus- und Isochroniewahrnehmung, wie in Kap. 3 ab S. 143 dargestellt wurde. Daher sollten sie bei perzeptiven Sprechgeschwindigkeitsuntersuchungen vorerst berücksichtigt werden.

Es muß noch festgehalten werden, daß aufgrund der Darstellungen zum P-center in Kap. 3.1 auf S. 144 klar geworden ist, daß die exakten P-centers ausgesprochen schwierig zu detektieren sind. Aber da die Mitten von Silbenkernen bessere Approximationen der P-centers darstellen als Silbengrenzen, kann hinreichend gerechtfertigt werden, daß in dieser Arbeit der Segmentation von Silbenkernmitten gegenüber Silbengrenzen — die zweifellos ebenfalls schwierig zu detektieren sind — der Vorzug gegeben wird.

4.4 Verdeckt lokale Sprechgeschwindigkeitsvariation Isochronie?

Lokale Sprechgeschwindigkeit kann bereits auf Silbenebene variieren, wie in Kap. 4.2 festgehalten wurde, und moduliert dann — damit zwangsläufig einhergehend — die intrinsischen Phon- und Silbendauern. Daher sollte es auch gar nicht verwundern, wenn Isochronie in gesprochener Sprache anhand von Dauermessungen eben nicht ohne weiteres nachzuweisen ist. Durch Untersuchungen anhand von musikalischen Rhythmen ist aber gut bekannt, daß Antizipation des Rhythmus auch bei Rallentandi/Ritardandi und Accelerandi/Stringendi[3] möglich ist, daß man sich also auch auf eine sich gleichmäßig verändernde Geschwindigkeit synchronisieren kann (Ehrlich 1958 [44]). Entsprechend könnte die Wahrnehmung von Sprechrhythmus auch bei variierender Sprechgeschwindigkeit möglich sein. Vor dem erneuten Versuch eines meßtechnischen Nachweises von Isochronie sollten demnach zumindest alle Sprechgeschwindigkeitsvariationen nivelliert werden, um die tatsächlichen intrinsischen Silbendauern aufzudecken und auf sie zurückgreifen zu können.

Laut einer Untersuchung von Donovan & Darwin 1979 [41] zeigt sich bei Wahrnehmungsexperimenten mit *nicht*-isochronen Sätzen, daß Probanden bei Matching- und Tapping-Aufgaben dennoch zur *perzipierten* Isochronie tendieren. Wir wollen unseren gewagten Erklärungsversuch hierfür nicht verschweigen: Rhythmen mit kürzeren und längeren Intervallen könnten dann quasiisochron wahrgenommen werden, wenn man annimmt, daß längere Intervalle mehr Information enthalten und somit eine längere kognitive Verarbeitung erfordern, was die subjektive Zeitwahrnehmung im Bereich der A-Prosodie in dem Sinne verfälschen könnte, daß *gut genutzte* Zeit (also die Zeit intensiver Informationsverarbeitung) schneller zu vergehen scheint als *gewartete* Zeit (die Zeit geringer Informationsdichte). Wenn dem so wäre, ließe sich Isochronie so lange nicht meßtechnisch nachweisen, bis die Beziehungen zwischen kognitiver Sprachverarbeitung und *sprachlicher Information* bzw. *Informationsdichte* hinreichend geklärt sind.

Erschwerend kommt hinzu, daß der Mensch, gestaltpsychologisch[4] betrachtet, dazu tendiert, wahrgenommene akustische Zeitstrukturen entsprechend der Gestaltgesetze zu gliedern. Auf diese

[3] Das allmähliche Langsamerwerden des musikalischen Tempos wird als *Rallentando* oder auch als *Ritardando* bezeichnet, wobei in der Fachliteratur Uneinigkeit bezüglich einer vollkommenen Synonymität beider Begriffe festgestellt werden kann. Dagegen ist ein *Accelerando* das Beschleunigen des Tempos und synonym zu *Stringendo*.

[4] Siehe zur Gestaltpsychologie Kap. 3.2 auf S. 145.

Weise interpretiert er möglicherweise isochrone und rhythmische Strukturen in die Sprache hinein, die physikalisch ohne weiteres nicht belegbar sind.

Die Hoffnung, durch eine Normalisierung der Sprechgeschwindigkeit diejenigen Zeitstrukturen aufdecken zu können, die maßgeblich für den Wahrnehmungseindruck der Isochronie verantwortlich sind, sollte demnach nicht allzu groß sein.

4.5 Informationsgehalt und Reduktion

In diesem Abschnitt wollen wir noch einmal auf den Unterschied zwischen Brutto- und Nettosprechgeschwindigkeit[5] eingehen und dabei auch versuchen, bisher ungeklärte Begriffe wie z.B. *sprachliche Information* und *geringe* bzw. *hohe Informationsdichte* anhand der Ergebnisse einer Untersuchung zu erläutern, die sich mit diesen Begriffen zwar gar nicht beschäftigte, aber aus der wir dennoch Aussagen über das mit den Begriffen Bezeichnete ableiten können.

In einer Untersuchung von Greisbach 1992 [68] lasen acht Probanden mit ihren normalen Sprechgeschwindigkeiten eine Kurzgeschichte, die insgesamt 196 kanonische Silben umfaßte, in durchschnittlich 41.3 Sekunden vor. Das entspricht 4.74 kanonischen Silben pro Sekunde. Als sie anschließend aufgefordert wurden, maximal schnell zu lesen, und nachdem sie eine zeitlang üben konnten, benötigten sie im Mittel nur noch 23.4 Sekunden und erreichten damit eine mittlere kanonische Silbenrate von 8.38 Silben/s. Der schnellste Sprecher benötigte sogar nur 18.7 Sekunden bei durchschnittlich 10.48 kanonischen Silben pro Sekunde.

Man könnte nun in einer ersten, naiven Näherung den Standpunkt einnehmen, daß der *Informationsgehalt* der vorgelesenen Geschichte unabhängig von der Sprechgeschwindigkeit gleich geblieben sein muß, aber daß sich dabei mit zunehmendem Vorlesetempo die *Informationsdichte* erhöht hat, da der gleiche *Informationsgehalt* in weniger Zeit übertragen wurde. Diesen Standpunkt vertritt Hildebrandt bereits 1963 [82], als er mit dem Begriff der „Übermittlungsgeschwindigkeit von Information“ die Anzahl der kanonisch gegebenen Laute pro Zeiteinheit gleichsetzte,[6] denn dadurch, daß kanonische Laute per se sprechgeschwindigkeitsunabhängig sind, muß dies in der Konsequenz dann auch für den *Informationsgehalt* gelten, so wie er von diesem Standpunkt aus zu deuten wäre.

Greisbach war an Reduktionsphänomen auf Lautebene interessiert, die seiner Meinung nach mit der Methode des maximal schnellen Vorlesens ähnlich ausgeprägt sind wie bei Spontansprache. Als Beispiel aus seiner Geschichte nannte er die Wörter *zwei Eier*, die bei normalem Sprechtempo durchaus kanonisch ausgesprochen wurden [tsvaɪʔaɪɐ] (Sprecher BA), bei sehr schnellem Lautlesen aber bis zu [tsvaɪɐ] (Sprecher MI) reduziert wurden, so daß eine ganze Silbe ausfiel. Phänomene dieser Art erwarten wir, da grundsätzlich mit wachsendem Sprechtempo auch zunehmend mehr Phone und Silben elidiert werden, und wir halten zunächst fest, daß ganz offensichtlich Nettophonraten und -silbenraten umso deutlicher unter den kanonisch basierten Bruttoraten liegen, je schneller gesprochen wird.[7]

Schließlich führte Greisbach einen Verständnistest mit zwei bis drei Sätzen jedes der drei schnellsten Sprecher aus dem Vorleseexperiment durch, an dem zehn Versuchspersonen teilnahmen. Diese waren trotz vielfachen Anhörens nicht in der Lage, die vorgelesenen Sätze vollständig zu rekonstruieren. Hiermit wird nun der obige, naive Standpunkt, daß der *Informationsgehalt* der

[5] Wir hatten diese Begriffe bereits auf S. 131 in Kap. 2.1 eingeführt und auf S. 140 in Kap. 2.5 definiert.

[6] Auf S. 131 in Kap. 2.1 wurde der Gedankengang von Hildebrandt bereits knapp skizziert.

[7] Auch der umgekehrte Fall, daß sich bei besonders langsamem Sprechen durch das Einfügen von z.B. Sproßvokalen und -konsonanten höhere Netto- als Bruttoraten ergeben, ist grundsätzlich denkbar.

vorgelesenen Geschichte unabhängig von der Sprechgeschwindigkeit sei und allein von den kanonischen Lauten abhänge, eindeutig widerlegt. Tatsächlich ist der übertragene Informationsgehalt nicht kanonisch gegeben und hat sich bei schnellem Sprechen definitiv verringert: die Sprecher elidierten nicht nur *redundante* Laute und Silben, sondern ganz offensichtlich auch zur Rekonstruktion der ursprünglichen Äußerungen *notwendige* Information, denn sonst wäre es nicht zu den Verständnisproblemen gekommen.

Gesprochene Sprache muß bis zu einem gewissen Grad redundant sein, damit sie dem Hörer in natürlichen Kommunikationssituationen mit Störgeräuschen die lückenlose Rekonstruktion der Information, auch bei teilweise maskierten *acoustic cues*, ermöglicht. So kann bereits der Wortkontext völlig ausreichen, um ein durch eine Störung vollständig maskiertes Wort trotzdem korrekt vorherzusagen. Die realisierten oder gar kanonisch gegebenen Laute müssen dazu nicht rekonstruiert werden; verwertbare Redundanz findet sich auch in der lexikalischen, syntaktischen, semantischen und pragmatischen Ebene. Wird allerdings die übertragene Redundanz und Information — etwa durch erhebliche Störgeräusche oder durch übermäßig reduzierende Aussprache — zu stark verringert, verstehen Hörer eine Äußerung merklich schlechter oder gar nicht mehr.

Wir gehen also ganz im Gegensatz zu Hildebrandt davon aus, daß sich der *Informationsgehalt* beim Sprechen nicht in den Brutto-, sondern in den Nettoraten widerspiegelt. Man sollte aber nicht dem Irrtum erliegen, daß Sprecher ihre Sprechgeschwindigkeit allein anhand der Häufigkeit von Elisionen regulieren würden. Schnell gesprochene Laute und Silben sind grundsätzlich kürzer als langsam gesprochene und tragen so ebenfalls zur schnelleren Aussprache bei.[8]

Aber selbst wenn die geringere Gesamtdauer der vorgelesenen Kurzgeschichte allein durch das Elidieren von Phonen und Silben bei ansonsten konstant gehaltenen Raten erreicht worden wäre, hätte sich trotzdem die *Informationsdichte* erhöht, da Sprecher typischerweise zuerst redundante Segmente elidieren und damit den Anteil informationstragender Segmente *pro Zeit* vergrößern.

Demnach erhöht schnelles Sprechen also grundsätzlich die Informationsdichte. Vermutlich ist dies ist auch der Grund, warum durch automatische Zeitkorrekturverfahren dauerkomprimiertes Sprechen *schwerer* zu verstehen ist als von einem Menschen genauso schnell produziertes Sprechen:[9] Der Mensch verringert durch Elisionen den Informationsgehalt, während Zeitkorrekturalgorithmen den Informationsgehalt unabhängig von der Zielgeschwindigkeit beibehalten und damit bei stärkerer Dauerkompression vermutlich eine für die kognitive Sprachverarbeitung des Menschen zu hohe Informationsdichte generieren.

Eine zusätzliche Ursache für die schlechtere Verständlichkeit erscheint ebenfalls naheliegend und soll hier der Vollständigkeit halber erwähnt werden: Durch die gängigen Zeitkorrekturverfahren werden auch diejenigen akustischen Merkmale (wie etwa Transitionen, Plosionen, Diphthongierungen, usw.) zeitlich verzerrt, deren Dauer selbst einer der *acoustic cues* ist.[10] Hier könnten unvorhersehbare Kategorienwechsel verursacht werden, die den Dekodierungsprozeß aufgrund widersprüchlicher Merkmale erschweren und auch Fehler verursachen.

---

[8] Dies sei im Vorgriff auf den experimentellen Teil dieser Arbeit, in dem sich u.a. herausstellen wird, daß sowohl die Nettophonrate als auch die Nettosilbenrate mit dem Sprechtempo korrelieren, bereits an dieser Stelle verraten.

[9] Siehe hierzu Foulke 1971 [53] und Covell, Withgott & Slaney 1998 [26]. Letztere zitieren Ergebnisse, nach denen natürliche Sprache noch mit bis zu 500 Wörtern pro Minute (wpm) verständlich ist, während dauerkomprimierte Sprache bereits bei maximal 270 wpm unverständlich wird. Sie entwickeln daher ein Kompressionsverfahren, das die Kompressionsphänomene schneller Sprache nachbildet: Pausen werden am stärksten komprimiert, betonte Vokale am geringsten, unbetonte Vokale werden durchschnittlich komprimiert und Konsonanten werden im Mittel stärker komprimiert als Vokale, aber immer in Abhängigkeit der benachbarten Vokale.

[10] So führt etwa eine *glide*-Phase von mehr als 120 ms zur Wahrnehmung eines Diphthongs, während bei weniger als 80 ms ein Hiat wahrgenommen wird (Peeters 1991 [165, S.193ff, S.200ff]).

Unserer Ansicht nach lassen sich zwischen Sprechgeschwindigkeit und Reduktionsgrad sicherlich Korrelationen nachweisen, und trotzdem sind beide bis zu einem gewissen Grad unabhängig voneinander: im Verhältnis zwischen den lokalen Netto- und Bruttoraten von Phonen und Silben spiegelt sich der Grad der lokalen Reduktion wider.

Festzuhalten ist, daß Sprechgeschwindigkeit *Informationsdichte* beeinflußt, und man kann durchaus den Standpunkt vertreten, daß die Steuerung der Sprechgeschwindigkeit einem Sprecher dazu dient, die Informationsdichte seiner Äußerung zu kontrollieren. Daß allerdings ausgerechnet das Ende einer Phrase (*pre-final lengthening*) typischerweise deutlich langsamer gesprochen wird, obwohl es aufgrund der vorangegangenen Äußerung oft am besten vorauszusagen und damit die größte Redundanz und den geringsten Informationsgehalt besitzt, zeigt, daß Sprechgeschwindigkeit nicht — oder zumindest nicht nur — dazu dient, die Informationsdichte möglichst *konstant* zu halten, also redundante Passagen schneller und wichtige, informationsreiche Passagen — wie etwa den Satzfokus — langsamer zu sprechen. Vielmehr dominiert in diesem Fall die für das reibungslose Funktionieren der menschlichen Kommunikation obligatorische Vorankündigung und Kennzeichnung von Äußerungsenden.

Im Sinne der von Shannon & Weaver 1949 [198] begründeten Informationstheorie läßt sich jede Art von Information mathematisch erfassen und beschreiben. Der *tatsächliche* Informationsgehalt einer lautsprachlichen Äußerung ergibt sich allerdings erst, wenn Art und Anzahl der möglichen sprachlichen Einheiten sowie deren Abhängigkeiten bekannt sind. Aber bereits auf der Ebene der Phonotaktik einer Sprache existieren unzählige informationsmindernde Abhängigkeiten, aus denen auch Redundanzen hervorgehen können. Dies gilt in zunehmendem Maße für alle weiteren Verarbeitungsebenen von der Syntax bis zur Pragmatik. Wie wenig über die akustische Kommunikation des Menschen bekannt ist, wird deutlich, wenn eine konkrete Aussage bezüglich des Grades an Redundanz einer ganz bestimmten sprachlichen Einheit gefordert wird. Ein weiteres Verfolgen dieses formal informationstheoretischen Ansatzes im Rahmen dieser weit fortgeschrittenen Fragestellungen erscheint beim jetzigen Stand der Forschung verfrüht.[11]

Die zukünftige Untersuchung des Sprechgeschwindigkeitsverhaltens beim *pre-final lengthening*, aber auch weiterer Divergenzen zwischen Sprechgeschwindigkeit und Informationsdichte könnte neue Erkenntnisse über das Funktionieren lautsprachlicher Kommunikation liefern, erfordert aber eine praktisch anwendbare Definition des Begriffs der *sprachlichen Information* und liegt daher außerhalb der in der vorliegenden Arbeit geplanten Untersuchungen.

4.6 Terminologie

Im Laufe des ersten Teils dieser Arbeit sind wir auf eine Reihe von Begriffen zur Bezeichnung des Sprechgeschwindigkeitsphänomens gestoßen: *running tempo, momentary tempo, Momentantempo, Lokaltempo, lokales Sprechtempo und lokale Sprechgeschwindigkeit*. Mit all diesen Begriffen ist grundsätzlich das lokale Sprechgeschwindigkeitsverhalten gemeint. Dagegen geht aus den ebenfalls gängigen Termini *articulation rate, speech rate, speaking rate, Tempo, Sprechtempo, Redegeschwindigkeit* und *Sprechgeschwindigkeit*, ohne die genaue Definition mitzuliefern, nicht eindeutig hervor, ob damit eine lokale oder globale Betrachtungsweise verbunden ist.

Um die Mannigfaltigkeit der Terminologie und damit verbundene Verwirrung nicht weiter zu unterstützen, werden wir uns im Deutschen auf die Begriffe *lokales Sprechtempo, lokale Sprech-*

[11] Wie man diesen Ansatz grundsätzlich in der phonetischen Forschung anwenden kann, hat Heid 1998 [77, S.216ff] anhand des *PhonDatII*-Korpus gezeigt.

rate oder *lokale Sprechgeschwindigkeit* und im Englischen auf den Begriff *local speech rate* beschränken und diese hier vollkommen synonym verwenden. Außerdem werden wir für die hier zu entwickelnde psychophysikalisch fundierte Sprechgeschwindigkeit einen eigenen Namen finden müssen.

ZWEITER TEIL

DAS EXPERIMENTELLE VORGEHEN

5

Stimuli und Sprachsignalkorpora

Für die folgenden Perzeptionsexperimente zur Sprechgeschwindigkeit kommen synthetische Stimuli nicht in Frage, da die von aktuellen Syntheseverfahren generierte zeitliche Struktur der Sprache immer noch mit Artefakten behaftet ist, deren Einfluß auf die Beurteilung von Sprechgeschwindigkeit nicht abschätzbar ist. Die Experimente müssen also mit natürlichsprachlichen Äußerungen durchgeführt werden. Somit bleibt nur die Wahl zwischen speziell für die hier angestrebten Experimente durchzuführenden Sprachaufnahmen und dem Rückgriff auf bereits existierende Sprachkorpora.

5.1 Korpusbasierte Wissenschaft

In der phonetischen Forschung werden vielfältige Methoden eingesetzt, um Sprachmaterial als Ausgangsbasis für empirische artikulatorische, akustische und perzeptive Untersuchungen zu gewinnen. Der höchste Grad an Kontrolle über das Material wird erreicht, indem systematisch variierte Logatomwörter, Silben, Vokale oder Konsonanten in gleichbleibende Trägersätze eingebettet werden. Auf diese Weise sind die fokussierten Phänomene sehr genau meßbar und statistisch zuverlässig zu analysieren. Allerdings wirft diese von jeder natürlichen Kommunikationssituation weit entfernte und damit artifizielle Aufnahmesituation kritische Fragen auf:

- Zeigen sich im Trägersatz Reduktionsphänomene, die allein durch sein häufiges Auftreten und Wiederholen hervorgerufen werden und sich auf die Zielwörter auswirken?
- Werden die variierenden Zielwörter oder -silben mit einem Kontrastakzent oder auf andere Weise übermäßig deutlich ausgesprochen?
- Inwiefern sind die Ergebnisse derartiger Studien generalisierbar?

Die Beantwortung zumindest der ersten beiden Fragen erübrigt sich, wenn Sprachmaterial untersucht wird, das während einer möglichst natürlichen Kommunikationssituation — beispielsweise während eines spontansprachlichen Dialogs — aufgezeichnet wurde. Aber auch in diesem Fall ist eine Reihe von wichtigen Fragen zu klären:

- Wie *natürlich* war die Aufnahmesituation? Führt bereits die bloße Anwesenheit von Aufzeichnungsgeräten während eines Kommunikationsprozesses grundsätzlich dazu, daß man als Ergebnis mit *Laborsprache* rechnen muß?
- Wie groß ist der Einfluß unerkannter Faktoren auf die Variation des fokussierten Phänomens?

- Ist das Sprachmaterial umfangreich genug, so daß sich das jeweils interessierende Phänomen hinreichend oft beobachten läßt, um statistisch reliable Aussagen treffen zu können? Ist also auf der Grundlage des Sprachmaterials der wissenschaftliche Anspruch an die Qualität einer Antwort erfüllbar?
- Ist das Korpus repräsentativ? Genauer gefragt: Lassen sich die Ergebnisse auf beliebige Sprecher und beliebige, nicht im Korpus enthaltene Äußerungen generalisieren?

Insbesondere bei Perzeptionsexperimenten ist es üblich, gezielt Manipulationen der Stimuli durchzuführen, so daß der interessierende Parameter in äquidistant variierten Schritten vorliegt. Dagegen wird beim korpusbasierten Vorgehen der Prozeß der Manipulation durch den der Suche abgelöst, die als Ziel ebenfalls ein möglichst gleichmäßig verteiltes Spektrum aller beobachtbaren Ausprägungen des interessierenden Faktors hat. Es muß als Vorteil gewertet werden, daß im zweiten Fall nur solche Stimuli auftreten können, die ein Mensch mit seinem Artikulationsapparat *tatsächlich hervorbringen* kann. Falls sicher gesagt werden kann, daß in der verwendeten Datenbank hinreichend unterschiedliche Ausprägungen vorliegen, so können Ergebnisse, die aufgrund korpusbasierter Stimuli erzielt werden, selbstverständlich statistische Reliabilität erreichen und bei genügend vielen Sprechern auch repräsentativ und generalisierbar sein.

Es ist anzunehmen, daß auf der Grundlage von äquidistant manipulierten Stimuli diese Ansprüche an Versuchsergebnisse schlechter zu erfüllen sind, als mit natürlichsprachlichen Stimuli aus Datenbanken. Denn schließlich werden die durch natürliches Sprechen evozierten Perzeptionsurteile zusätzlich durch bisher unerkannte oder unberücksichtigte und damit bei zufälliger Stimulusauswahl ebenfalls zufällig ausgeprägte Faktoren — wie z.B. Variationen in der Stimmqualität eines Sprechers — beeinflußt und lassen sich daher schlechter aus den kontrollierten Faktoren vorhersagen. Wenn also trotz unkontrollierter phonetischer Variation ein Modell mit guter Vorhersagequalität entwickelt werden könnte, so wäre es zuverlässiger, als ein allein auf der Grundlage von äquidistant manipulierten Stimuli entwickeltes Modell gleicher Vorhersagequalität.

Die hier angestrebten Untersuchungen sollen weitgehend sprechstilunabhängige Aussagen über die Prosodie der Sprechgeschwindigkeit erlauben. Deswegen müssen die Stimuli auf verschiedenen Sprechstilen basieren. Nun ist wenigstens für Spontan- vs. Lesesprache unterschiedliches prosodisches Verhalten dokumentiert (Hirschberg 2000 [83]), was das Einbeziehen dieser beiden Sprechstile nahelegt. Es ist jedoch keine Methode bekannt, um systematisch variierte spontansprachliche Äußerungen zu elizitieren. Also müssen die erforderlichen Sprachstimuli aus möglichst großen, bereits existierenden, handsegmentierten spontansprachlichen Korpora extrahiert werden. Um die Vergleichbarkeit aller Experimente hinsichtlich des Faktors *Stimulusherstellung* gewährleisten zu können, sollten dann auch die auf Lesesprache basierenden Stimuli aus großen Sprachkorpora stammen.

5.2 Entwicklungs- und Testkorpus

Eine Aufteilung der Stimuli einerseits in Entwicklungs- bzw. Trainingsdaten und andererseits in Test- bzw. Evaluationsdaten ist zur Entwicklung und Optimierung von tatsächlich anwendbaren Modellen unumgänglich, wenn eine Überadaption an die Trainingsdaten vermieden und damit die Generalisierungseigenschaft des entwickelten Verfahrens garantiert werden soll.

Beide Korpora sollten natürlich so groß wie möglich sein; typischerweise wird aber ein größeres Trainingskorpus verwendet, um eine Spezialisierung der verwendeten Algorithmen auf das

trainierte Material und damit das pure „Auswendiglernen“ während der Entwicklungsphase zu erschweren. Die Qualität der auf dem Trainingsmaterial entwickelten Verfahren wird erst *nach dem Abschluß* der Entwicklung anhand des Testmaterials evaluiert. Dieses sollte möglichst disjunkt mit dem Trainingskorpus sein, um die Generalisierungsfähigkeit des Verfahrens an während des Trainings ungesehenem Sprachmaterial testen zu können.

In dem Fall, daß ein sich mehrfach abwechselndes Trainieren und Evaluieren geplant sein sollte, wäre die Verwendung von *drei* Korpora unabdingbar, da sich die entwickelten Verfahren im Laufe des mehrfachen Wechselns zunehmend auch an die Testdaten adaptieren, wodurch die Objektivität solcher Tests abnimmt. Deswegen wäre hier ein zusätzliches eigenes Korpus für die abschließende Einschätzung der Qualität erforderlich.

Da wir jedoch nicht planen, nach der Evaluation wieder zurück in die Entwicklungsphase zu wechseln, kommen wir an dieser Stelle mit zwei disjunkten Korpora aus, wobei ein größerer Umfang für das Trainingskorpus angestrebt wird.

5.3 PhonDatII und VerbmobilI

Als Sprachsignalquelle für die Stimuli des ersten bis vierten Perzeptionsexperiments bot sich das *PhonDatII*-Korpus[1] an, da etwa ein Drittel der enthaltenen Äußerungen manuell segmentiert sowie mehrfach korrigiert und damit in höchster Qualität vorliegt. Es umfaßt 200 Sätze aus einem Szenario der Zugauskunft, die von je sechs weiblichen und zehn männlichen Sprechern aus Kiel, Bonn und München vorgelesen — und teilweise sogar annähernd geschauspielt — wurden. Das Korpus deckt damit nicht nur gelesene Sprache, sondern einen breiteren Bereich an Sprechstilen zwischen gelesener und schauspielerischer Sprache ab.

Alle Äußerungen wurden in speziellen reflektionsarmen Studioräumen mit dem auf Nierenrichtcharakteristik eingestellten hochqualitativen Kondensatormikrofon *Neumann U87*, dem Vorverstärker *John Hardy M1* und dem DAT-Rekorder *Sony PCM-2500* in 16-Bit-Auflösung aufgezeichnet und digital auf 16 kHz umabgetastet (Thon 1992 [206], Pompino-Marschall 1992 [177], Hess 1993 [80, S.330]). Die Sprachaufnahmen fanden im Frühjahr 1992 statt, nachdem auf der sog. Hamburger Datensitzung vom 11.12.1991 das aufzunehmende Sprachmaterial ausgewählt und beschlossen worden war ([177, S. 107]). Die manuelle Segmentation und Etikettierung von etwa 40.000 Phonen[2] in 1024 Sätzen (64 der von jedem der 16 Probanden gesprochenen 200 Sätze) wurde bis Dezember 1992 abgeschlossen.

Später wurden im Rahmen einer Untersuchung zur automatischen Silbenextraktion (Pfitzinger, Burger & Heid 1996 [172]) zusätzlich 15083 Mitten von allen Silbennuklei der 1024 Äußerungen durch drei Phonetiker von Hand markiert.

In Abb. 1.1 auf S. 124 haben wir bereits die lokalen Silben- und Phonraten von denjenigen händisch bearbeiteten 1024 Äußerungen graphisch dargestellt, die vollständig mit handsegmentierten Silbennuklei und Phongrenzen versehen sind. Dabei handelt es sich um etwa 40 Minuten gesprochener Sprache, was ca. einem Drittel des Gesamtmaterials entspricht. Für die übrigen 2176 Äußerungen liegen zwar manuell bestimmte Wortsegmentationen vor, sie finden in dieser Untersuchung aber keine Verwendung.

---

[1] *PhonDatII* wurde mit Mitteln des damaligen Bundesministeriums für Forschung und Technologie (heute BMBF) vom 1.1.1991 bis 31.12.1992 unter dem Kennzeichen DLR01IV103 gefördert.

[2] Als Resultat jüngster Korrekturen liegen momentan 39612 Phone vor.

Da es sich bei den Äußerungen aus dem *PhonDatII*-Korpus vorwiegend um Lesesprache handelt und in dieser Arbeit auch geprüft werden soll, ob die Ergebnisse auf im Trainingskorpus ungesehene Sprechstile verallgemeinerbar sind, lagen dem fünften und sechsten Experiment Stimuli aus dem spontansprachlichen Korpus *VerbmobilI* zugrunde. Dieses wurde in den vergangenen Jahren vielfach dokumentiert und war Grundlage zahlreicher Untersuchungen. In dem von Wahlster 2000 [232] zum *Verbmobil*-Projekt[3] herausgegebenen Buch findet sich eine sehr umfangreiche Darstellung. Daher darf an dieser Stelle auf die Wiederholung der vorliegenden und gut zugänglichen Beschreibungen verzichtet werden.

5.4 Segmentation und Etikettierung

Die Etikettierung gesprochener Sprache wird sinnvollerweise nicht mit Phonemen durchgeführt, sondern mit einem Lautinventar, das sich zwar an dem Phoneminventar der jeweiligen Sprache orientiert, aber nach phonetischen und auch praktischen Gesichtspunkten gezielt an die Erfordernisse beim Segmentieren angepaßt wurde. So sind etwa im Deutschen folgende Besonderheiten zu berücksichtigen:

1. Die vokalisierte Aussprachevariante des /r/, das [ɐ], tritt zwar grundsätzlich wortfinal und auch präkonsonantisch auf, so daß sie sich vorhersagen ließe, aber die phonetische Ähnlichkeit dieses Allophons mit den frikativischen Aussprachevarianten des /r/ ist so gering, daß sich hierfür ein zusätzliches Symbol empfiehlt.
2. Der Glottalverschluß, der aus phonologischer Sicht nicht gekennzeichnet werden muß, da er vor Vokalen in morphologisch gekennzeichneten Positionen auftritt, wird als eigenständiger Laut segmentiert, um bei fehlenden Morphemgrenzen dennoch Wörter wie z.B. *ver-reisen* und *ver-eisen* auseinanderhalten zu können.
3. Die beiden Allophone [ç] und [x], die einem Phonem zugeordnet werden können, sollten trotzdem als zwei verschiedene Segmente behandelt werden, da sie bei fehlenden Morphemgrenzen die notwendige Information liefern, um die Aussprache von Wörtern wie z.B. *Frauchen* vs. *rauchen* zu differenzieren.

Die manuellen Phonsegmentationen der beiden Korpora *PhonDatII* und *VerbmobilI* basieren prinzipiell auf dem gleichen Lautinventar, das in seiner Struktur und seinem Umfang nahezu dem IPA Symbolinventar entspricht, welches von der *International Phonetic Association* für das Deutsche vorgeschlagen wurde (IPA 1999 [93, S.86f]). Die wesentlichen Unterschiede liegen *i)* in der Verwendung von SAM-PA- statt IPA-Symbolen im Hinblick auf eine einfachere maschinelle Verarbeitung (Wells, Barry & Fourcin 1989 [235]) und *ii)* in der Zusammenfassung von Vokal-[ɐ]–Diphthongen zu jeweils einem Segment, die aus rein praktischen Erwägungen beschlossen wurde, nämlich um die zu erwartende große Anzahl an unsicheren Segmentgrenzen zwischen den beiden Vokalen zu vermeiden (van Dommelen 1992 [222, S.201]). Gerade auch durch diesen zweiten Punkt weicht das in den Tabellen 5.1, 5.2 und 5.3 dargestellte Etikettierungsinventar vom gemeinhin anerkannten Lautinventar des Deutschen erheblich ab.

Die ursprünglichen Segmentationskonventionen für *VerbmobilI* stimmen zwar nicht mit den wesentlich älteren *PhonDatII*-Konventionen überein. Aber dies betrifft hauptsächlich die Substitu-

[3] *Verbmobil* wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) im Zeitraum zwischen 1.1.1993 und 30.9.2000 in zwei Phasen *I* und *II* unter den Kennzeichen DLR01IV101 und DLR01IV102 gefördert.

	bilabial	labio-dental	alveolar	post-alveolar	palatal	velar	uvular	glottal
Plosiv	p b		t d			k g		Q
Nasal	m		n			N		
Frikativ		f v	s z	S Z	C	x	r	h
Approximant					j			
Lateral			l					

Tabelle 5.1: Die SAM-PA–Symbole für die deutschen Konsonanten.

	vorne	gerundet	hinten
geschlossen	i: I	y: Y	u: U
halb-geschlossen	e:	2:	o:
halb-offen	E: E	9	O
offen	a: a		

vorne	gerundet	hinten
i:6 I6	y:6 Y6	u:6 U6
e:6	2:6	o:6
E:6 E6	96	O6
a:6 a6		

Tabelle 5.2: Links: Die Monophthong-Symbole. Rechts: Die speziellen Symbole für die Vokal-[ɐ]–Diphthonge. Gespannte Vokale sind mit [:] gekennzeichnet.

aI	aU	OY

@	6

Tabelle 5.3: Links: Die drei deutschen Diphthonge. Rechts: Die beiden Zentralvokale.

tionsregeln, nach welchen die Segmentierer kanonisch vorgegebene Etiketten durch andere Etiketten, die die tatsächlich realisierten Aussprachevarianten genauer repräsentieren, ersetzen durften. Für unsere Vorgehensweise sind aber die gewählten phonetischen Symbole generell nur von geringer Relevanz; wir interessieren uns vielmehr für die zeitlichen Positionen der Etiketten.

Deswegen wurde hier, was die Zeitpositionen der Phongrenzen angeht, mit einer automatisch angepaßten und teilweise manuell nachkorrigierten Version der *VerbmobilI*-Segmentationen gearbeitet, die eine hinreichend hohe Konsistenz mit den *PhonDatII*-Daten sicherstellt.

5.5 Stimulusauswahl für die Perzeptionsexperimente 1 bis 4

Als Grundlage für das Entwicklungskorpus, das den späteren Perzeptionsexperimenten 1 bis 4 als Ausgangsbasis dienen sollte, wurden aus dem *PhonDatII*-Korpus 144 Stimuli derart zusammengestellt, daß von den 16 Sprechern je neun Signale gemäß Abb. 5.1 vorlagen. Dabei sollten nur Stimuli akzeptiert werden, deren Silben- und Phonrate und damit auch das Verhältnis zwischen ihnen im zeitlichen Verlauf eines Stimulus möglichst wenig variierten. Dies gelang jedoch nur begrenzt, denn die beiden Raten zeichnen sich gerade durch ihre große Variabilität aus.

Schwierig gestaltete sich auch die Auswahl der langsamsten Stimuli, denn viele niedrige Silben- und Phonraten treten unmittelbar vor Sprechpausen auf und zeigen das Phänomen des *pre-final lengthening*,[4] so daß bei Unachtsamkeit aus derartigen Sprechpassagen ausgewählte

[4] Das Phänomen des *pre-final lengthening* ist knapp in Fußnote 22 auf S. 136 beschrieben.

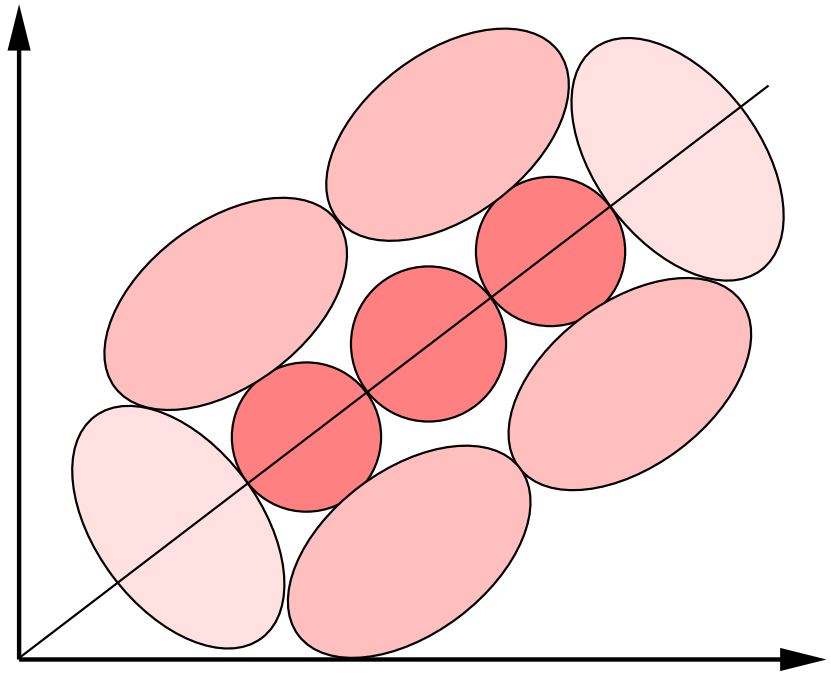

Abb. 5.1: Aus jedem dieser neun Bereiche der in Abb. 1.1 auf S. 124 dargestellten Verteilung zwischen Silben- und Phonrate wurde für jeden der 16 Sprecher des *PhonDatII*-Korpus ein Stimulus ausgewählt.

Sprachsignalabschnitte leicht bereits einen Teil der Sprechpause oder auch deutlich wahrnehmbare Sprechgeschwindigkeitsvariation enthalten konnten.

Auf der Basis dieser Auswahl wurden Sprechpassagen mit einer Dauer von zunächst 1400 ms automatisch aus dem Korpus geschnitten, um sie später auf die gewünschte Dauer zu kürzen, da erst die nachfolgenden Perzeptionsexperimente die optimale Stimulusdauer liefern sollten, von der jedoch erwartet wurde, daß sie nicht länger als 1400 ms ist.

Grundsätzlich werden durch die Schnitte, welche teilweise sogar mitten in Phonen durchgeführt wurden, Artefakte eingeführt; sie entstehen allerdings aufgrund der Koartikulation auch dann, wenn die Schnitte mit Silben-, Morph- oder Wortgrenzen synchronisiert sind. Versuchsweise wurden Stimuli mit der Absicht, artefaktfreie und unauffällige Schnitte zu erreichen, manuell segmentiert. Diese Stimuli begannen oder endeten dann jedoch meist an Satzgrenzen und waren damit in ihrer Bandbreite unterschiedlicher Sprechgeschwindigkeiten deutlich eingeschränkt und insofern weitaus weniger brauchbar als die automatisch geschnittenen Stimuli. Aus diesem Grund wurde der Methode des unbeaufsichtigten Schneidens schließlich der Vorzug gegeben.

Um die Schnittartefakte zu minimieren, wurden alle verwendeten Stimuli am Anfang und am Ende jeweils über 10 ms ein- bzw. ausgeblendet. Dieses Zeitintervall verhindert einerseits sicher jedes „Klick“-Geräusch und führt andererseits aufgrund seiner sehr kurzen Dauer nicht zu einem wahrnehmbaren An- und Abschwellen der Amplitude. Zusätzlich wurden alle Stimuli mit der gleichen Lautheit abgespielt.

Drei Stimuli, einer mit niedriger, einer mit durchschnittlicher und einer mit hoher Silben- und Phonrate, wurden als Ankerschalle für die kommenden Perzeptionsexperimente ausgewählt. Sie stammen aus den drei in Abb. 5.1 dunkel gezeichneten Bereichen, so daß später in den Perzeptionsexperimenten auch Stimuli mit noch höheren bzw. niedrigeren Silben- und Phonraten als der schnelle bzw. der langsame Ankerschall auftreten. Für die angestrebten Perzeptionsexperimente 1 bis 4 bleiben also 141 Stimuli.

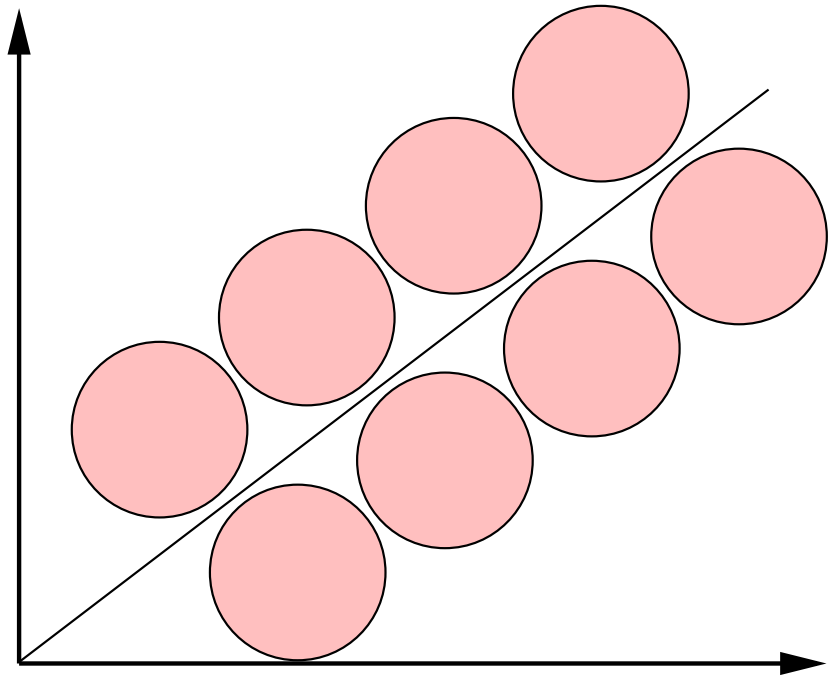

Abb. 5.2: Aus jedem dieser acht Bereiche der in Abb. 5.3 dargestellten Verteilung zwischen Silben- und Phonrate wurde für 12 Sprecher des *VerbmobilI*-Korpus jeweils ein Stimulus ausgewählt.

5.6 Stimulusauswahl für die Perzeptionsexperimente 5 und 6

Den Perzeptionsexperimenten 5 und 6 sollte das Testkorpus zugrunde liegen. Dieses durfte wesentlich kleiner als das Entwicklungskorpus sein, sollte aber trotzdem aus insgesamt immerhin 100 Stimuli bestehen. Die drei Ankerschalle von Experiment 4 wurden auch hier eingesetzt, um eine Vergleichbarkeit bei der Verwendung der Geschwindigkeitsskalen sicherzustellen.

Zusätzlich sollten vier der 141 im vorigen Kapitel ausgewählten Stimuli auch in den Experimenten 5 und 6 wiederverwendet werden, um später die Reliabilität der Urteile unter der Bedingung, daß sich die übrigen gleichzeitig präsentierten Stimuli bezüglich Anzahl, Zusammenstellung, Wortlaut, Sprechstil und zugrundeliegender Sprecher vollständig ändern, überprüfen zu können. Die Wahl fiel auf die Stimuli *n7*, *a3*, *c6* und *c8*, die hier in die Stimuli *96*, *97*, *98* und *99* umbenannt wurden.[5] Diese vier Stimuli wurden auf der Grundlage der Sprechgeschwindigkeitsbeurteilungen des Perzeptionsexperiments 4 ausgewählt, das erst in Kap. 7.6 ab S. 190 beschrieben ist. Auswahlkriterium war einerseits eine geringe Standardabweichung der Urteile und andererseits eine große Ähnlichkeit der mittleren Beurteilungen jeweils zweier Stimuli zum langsamen und zum schnellen Ankerschall.

Die übrigen 96 Stimuli des Evaluationskorpus wurden aus dem bereits in Kap. 5.3 beschriebenen spontansprachlichen *VerbmobilI*-Korpus ausgewählt, da dieses Korpus sich vom *PhonDatII*-Korpus erstens durch die Sprecher, zweitens durch die Domäne, der das Sprachmaterial entstammt, und drittens durch den Sprechstil unterscheidet, und damit in seinen entscheidenden Eigenschaften disjunkt mit dem Trainingskorpus ist.

In der Absicht, im Evaluationskorpus Stimuli entlang der Diagonale der Verteilung zwischen lokaler Silben- und Phonrate in Abb. 5.3 weitestgehend zu vermeiden und damit einen weiteren

[5] Die Oszillogramme, Sonagramme, Segmentierungen, Phon- und Silbengeschwindigkeiten und Grundfrequenzverläufe der Stimuli *c6* und *c8* sind in den Abb. 6.1 und 6.2 auf den S. 172 bzw. 173 dargestellt.

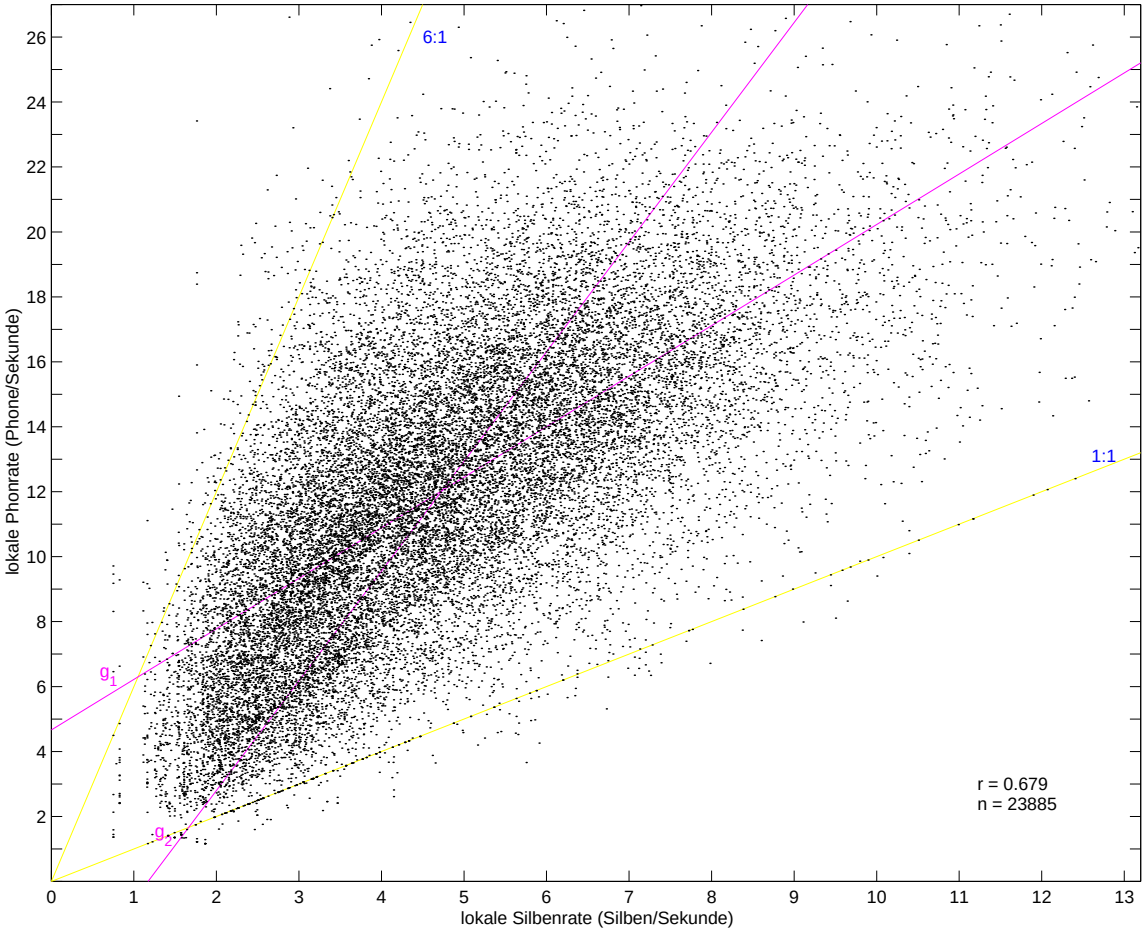

Abb. 5.3: Jeder der 23885 Punkte in diesem Streuungsdiagramm repräsentiert die Silben- und Phonrate eines 625 ms dauernden Äußerungsteils, die in 100 ms Schritten aus handsegmentierten Mitten von Silbenkernen bzw. Phongrenzen gewonnen wurden. Hier sind 10 Dialoge des *VerbmobilI*-Korpus mit 20 Sprechern und 372 Turns dargestellt, wobei ein Turn eine durchschnittliche Dauer von 6.4 Sekunden hat.

Schwierigkeitsgrad einzuführen, wurden diesmal — im Gegensatz zur vorigen Stimulusauswahl — möglichst keine Stimuli mit einem durchschnittlichen Phon-/Silbenratenverhältnis ausgewählt, sondern nur solche mit einem über- oder unterdurchschnittlichen Verhältnis. Dadurch ergaben sich die acht in Abb. 5.2 gekennzeichneten Bereiche, die von den in Abb. 5.1 dargestellten und den Perzeptionsexperimenten 1 bis 4 zugrundeliegenden Bereichen deutlich abweichen.

Da pro Sprecher für jeden der acht Bereiche nur ein Stimulus im Testkorpus vertreten sein sollte, wurden 12 Sprecher benötigt, um auf 96 Stimuli zu kommen. Der uns vorliegende handsegmentierte Ausschnitt aus dem *VerbmobilI*-Korpus bestand aus 10 Dialogen mit 20 Sprechern, von denen 12 aufgrund ihrer reichhaltigen Verteilung zwischen lokaler Silben- und Phonrate als Grundlage für die Auswahl der Stimuli selektiert wurden.[6] Schließlich wurde per Zufall für jeden der 12 Sprecher aus jedem der acht in Abb. 5.2 gekennzeichneten Bereiche ein Stimulus ausgewählt.

[6] Es handelt sich dabei um die sieben Sprecherinnen *m112a, m114b, m116b, m222a, m222b, m224a* und *m230a* sowie die fünf Sprecher *m113a, m113b, m114a, m115a* und *m117a*.

6

Akustische Untersuchungen der Stimuli

6.1 Vorüberlegungen

Als akustische Merkmale für die Sprechgeschwindigkeit kommen zunächst linguistisch relevante Einheiten in Frage, deren Dauern meßbar sind, also Phone, Silben, Füße, Morphe, Wörter, Akzente, Pausen und Häsitationen. Das Ziel dieser Untersuchung besteht allerdings darin, nur diejenigen akustischen Merkmale einzubeziehen, die durch bereits erforschte und bewährte Verfahren der digitalen Sprachsignalverarbeitung mit akzeptabler Fehlerrate extrahierbar sind.

Wort- und Morphgrenzen sind ohne das Wissen über die Sprache und die Erkennung der Wörter automatisch nicht zu markieren. Pausen und Häsitationen sind im Vergleich zu Wörtern sehr selten und kommen zudem in den Trainings- bzw. Teststimuli aus den Kap. 5.5 bzw. 5.6 ganz bewußt nicht vor. Akzente und Füße sind automatisch sehr schwierig zu extrahieren, während sich Silbenkerne und Phongrenzen bei Inkaufnahme einer gewissen Fehlerrate automatisch extrahieren lassen. Die vorliegenden akustischen Untersuchungen konzentrieren sich daher auf die automatisch extrahierbaren linguistischen Einheiten Silbe und Phon sowie zusätzlich auf die Grundfrequenz, deren Relevanz für die Wahrnehmung der Sprechgeschwindigkeit bereits früh dokumentiert wurde (Hoequist 1983/84 [85, 87]).

Unser Ziel besteht darin, die Prosodie der Sprechgeschwindigkeit als lokale Geschwindigkeit zu bestimmen und die globale als eine Art durchschnittliche Geschwindigkeit davon zu unterscheiden; im späteren Verlauf der Arbeit soll es mit Methoden der digitalen Sprachsignalverarbeitung sogar möglich werden, die lokalen Sprechgeschwindigkeitsvariationen durch inverse Verzerrungen der Zeitachse von Sprachsignalen vollständig aufzuheben und eine normalisierte Durchschnittsgeschwindigkeit zu erhalten, die durch einen nahezu konstanten Verlauf der Sprechgeschwindigkeit ohne relevante Minima oder Maxima gekennzeichnet ist. Dabei bleiben die mikroprosodischen Dauerstrukturen, die etwa in den intrinsischen Lautdauern kodiert sind, weitgehend erhalten. Ein „lokaler“ Ansatz ist also nicht mit einem „momentanen“ zu verwechseln, dessen Geschwindigkeitsprofil bereits auf der Lautebene variiert und dessen Normalisierung auch mikroprosodische Variation nivellieren würde.

6.2 Segmentation von Silbenkernen und Phongrenzen

Um in dieser Untersuchung solche Fehler, die jede automatische Segmentation zwangsläufig mit sich bringt, grundsätzlich auszuschließen, basiert die hier verwendete Silbenrate auf von drei Phonetikern manuell markierten Silbenkernen (Pfitzinger, Burger & Heid 1996 [172]) und die Phon-

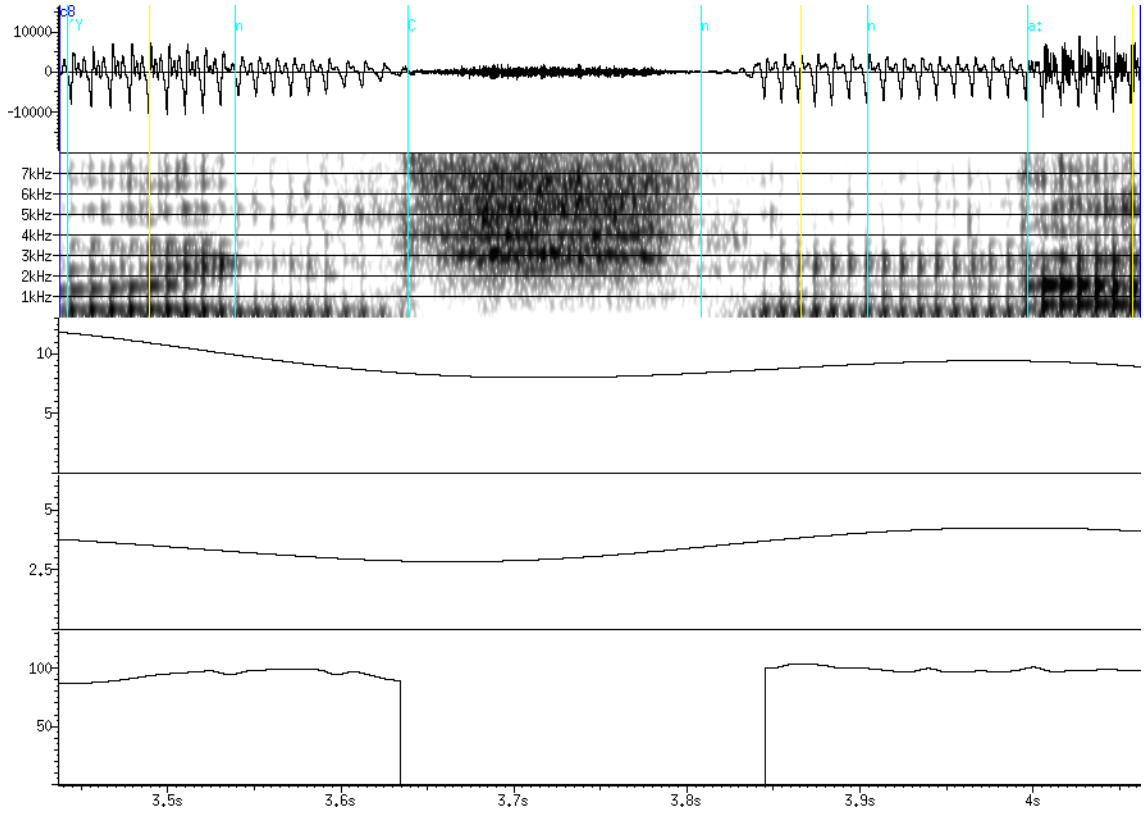

Abb. 6.1: Stimulus *c8* aus dem Signal *hgpd5340* (Lautfolge [YnCn na:] aus den Wörtern *München nach*). Oszillogramm und Sonagramm mit Phongrenzen- und Silbenkernmarken, lokaler Phongeschwindigkeitsverlauf (Phone/s), lokaler Silbengeschwindigkeitsverlauf (Silben/s) und Grundfrequenzverlauf (Hz).

rate auf den nach der *PhonDatII*-Konvention handsegmentierten und mehrfach evaluierten Phongrenzen (Pompino-Marschall 1992 [177]).

Aber auch bei manueller Segmentation treten immer wieder Zweifelsfälle auf, die von verschiedenen Phonetikern unterschiedlich beurteilt werden.[1] Die Fehlerrate gut ausgebildeter menschlicher Segmentierer ist jedoch beträchtlich geringer und wirkt sich zudem weniger gravierend aus als die von Automaten, welche insbesondere bei der Segmentation von Phonen gelegentlich zu schweren Fehlern tendieren, bei denen die zugeordneten Segmentgrenzen nichts mehr mit dem zugrundeliegenden Sprachsignal zu tun haben.[2] Gerade aus diesem Grund ist es bei der akustischen Analyse der ausgewählten Trainings- und Teststimuli wichtig, deren Segmentation von Hand vorzunehmen und so diese Fehlerquelle in der Entwicklungsphase auszuschließen.

Um einige Problemfälle bei der Segmentation an ganz konkreten Beispielen zu verdeutlichen, sollen an dieser Stelle einerseits diejenigen zwei Stimuli vorgestellt werden, deren erst später in Kap. 7.6 beschriebene Sprechgeschwindigkeitseinschätzungen von 60 Probanden die kleinsten Standardabweichungen und damit die größten Übereinstimmungen aufwiesen (siehe Abb. 6.1 und 6.2), und andererseits diejenigen zwei Stimuli, deren Standardabweichungen am größten waren (siehe Abb. 6.3 und 6.4).[3]

[1] Eisen, Tillmann & Draxler 1992 [45] sprechen in diesem Zusammenhang von klaren und unklaren Fällen.

[2] Im Extremfall wird ein Phon oder sogar eine Sprechpause in viele Lautsegmente unterteilt, während zwei andere Segmentgrenzen eine längere Folge von verschiedenen Phonen zu einem Segment zusammenfassen.

[3] Im Anhang ab S. 230 sind die Standardabweichungen und Histogramme aller Perzeptionsurteile zu allen Stimuli grafisch dargestellt.

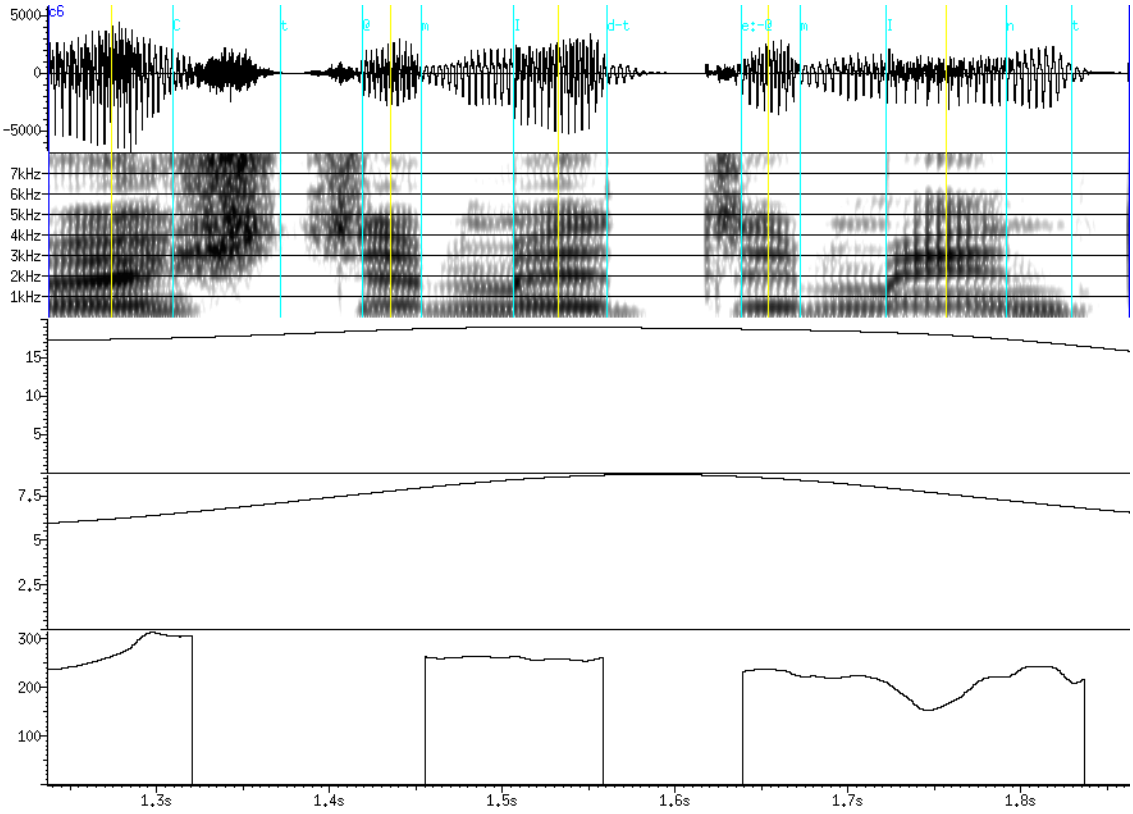

Abb. 6.2: Stimulus *c6* aus dem Signal *mknn5270* (Lautfolge [9Ct@ mIt@m Int] aus den Wörtern *möchte mit dem Intercity*). Oszillogramm und Sonagramm mit Phongrenzen- und Silbenkernmarken, lokaler Phongeschwindigkeitsverlauf (Phone/s), lokaler Silbengeschwindigkeitsverlauf (Silben/s) und Grundfrequenzverlauf (Hz).

Bei der Segmentation von Stimulus *c8* (Abb. 6.1: Lautfolge [YnCn na:] aus den Wörtern *München nach*) fällt auf, daß der Nasal bei ca. 3.9 s in zwei Segmente aufgeteilt wurde und damit im ersten Segment syllabischen Charakter zugewiesen bekam. Hier stimmen Silben- und Phonsegmentation überein, die beide unabhängig voneinander durchgeführt wurden. Entweder wurden beide Segmentierer durch die kanonische Form des Wortes *München* in ihren Urteilen beeinflußt, oder sie hörten den ersten Teil des Nasals tatsächlich syllabisch.

Man kann immerhin im Amplitudenverlauf des Oszillogramms einen Abfall vom ersten zum zweiten Teil des Nasals und eine überlange Gesamtdauer beobachten. Hätten sich die Segmentierer allerdings dazu entschlossen, beide Teile des Nasals als ein Segment zu werten, wären lokale Silben- und Phonrate deutlich niedriger ausgefallen. Wir werden später feststellen, daß unser Prädiktionsverfahren diesen Stimulus im Vergleich mit den Perzeptionsurteilen zu schnell einschätzt. Es hätte aufgrund der alternativen Segmentation eine bessere Schätzung abgegeben.

Abb. 6.2 zeigt den wesentlich schneller gesprochenen Stimulus *c6* (Lautfolge [9Ct@ mIt@m Int] aus den Wörtern *möchte mit dem Intercity*), der ebenfalls von den Probanden sehr einheitlich eingeschätzt wurde. Wie zu erwarten gibt es auch hier Zweifelsfälle: Zum einen hat der alveolare Plosiv bei ca. 1.6 s eine auffällig große Dauer, die zwei Plosiven entsprechen könnte. Da aber keine zwei Lösungsgeräusche auftreten, handelt es sich, phonetisch betrachtet, nur um ein Segment. Aus phonologischer Sicht wurden hier natürlich zwei Plosive mit gleichem Artikulationsort — der finale Plosiv von *mit* und der initiale von *dem* — zu einem verschmolzen, wobei allerdings die Einzeldauern beibehalten worden sein könnten.

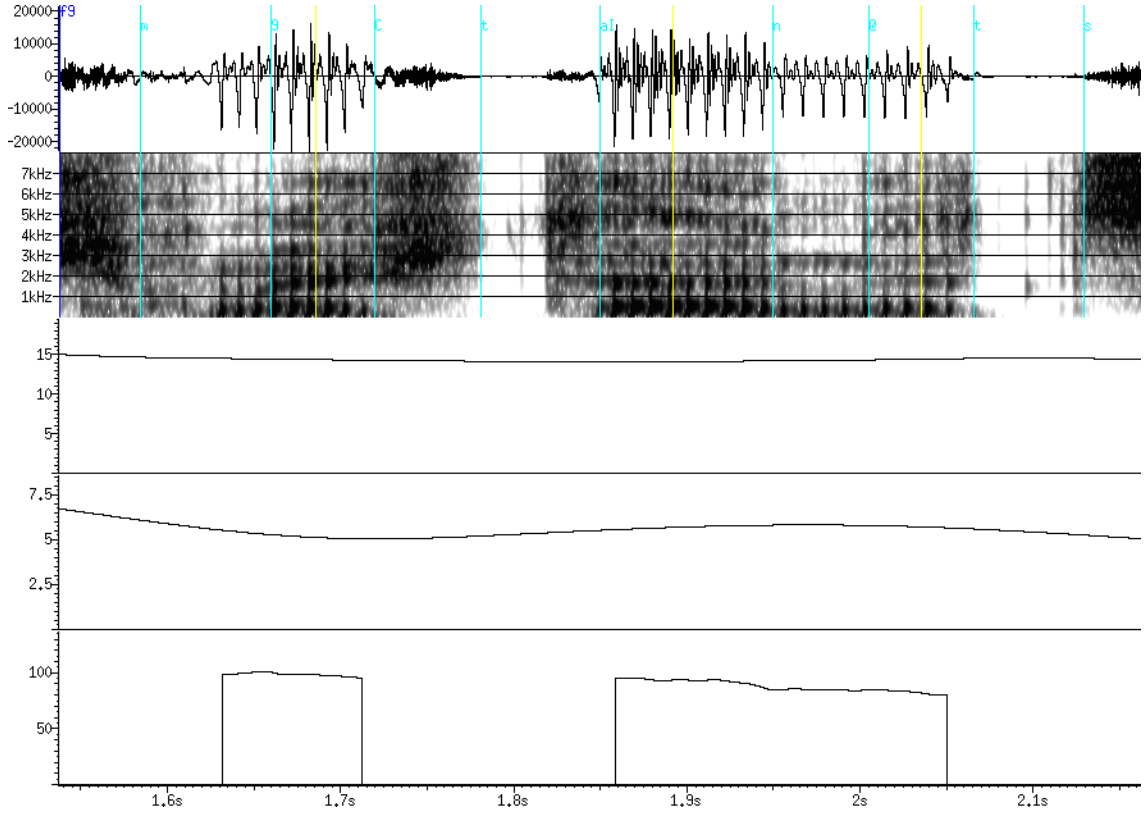

Abb. 6.3: Stimulus *f9* aus dem Signal *hgpd7190* (Lautfolge [C m9Ct aIn@ ts] aus den Wörtern *ich möchte eine Zugverbindung*). Oszillogramm und Sonagramm mit Phongrenzen- und Silbenkernmarken, lokaler Phongeschwindigkeitsverlauf (Phone/s), lokaler Silbengeschwindigkeitsverlauf (Silben/s) und Grundfrequenzverlauf (Hz).

Zum anderen zeigt der Grundfrequenzverlauf bei 1.75 s einen V-förmigen Einbruch, der den initialen Glottalverschluß von *Intercity* repräsentiert. Ob dies als Glottalverschluß und damit als ein eigenständiges produziertes Segment empfunden werden könnte, das zwar zu einer Laryngalisierung reduziert wurde, aber dennoch zeitkonsumierend und damit schneller gesprochen erscheint als eine vollständige Reduktion, kann hier nicht mit Sicherheit festgestellt werden.

In dieser Genauigkeit wollen wir die letzten beiden Beispiele in den Abb. 6.3 und 6.4 nicht mehr diskutieren, zumal wir uns auf einer sehr spekulativen Ebene bewegen. Und letztendlich können wir anhand eines Vergleichs der vier Abbildungen auch nicht eindeutig feststellen, warum Probanden bei den letzten beiden Stimuli erheblich uneinheitlichere Urteile abgegeben haben, während sie bei den ersten beiden Stimuli einen hohen Grad an Übereinstimmung erreichten. Somit bleibt zukünftigen Untersuchungen vorbehalten, diese interessante Frage wieder aufzugreifen.

6.3 Geschwindigkeitsmessung anhand von Zeitmarken

Eine konkrete Silben- bzw. Phongeschwindigkeit wird üblicherweise in *Silben/s* bzw. *Phonen/s* angegeben. Das Messen einer Geschwindigkeit auf der Basis von Intervallen zwischen Zeitpunkten, wie etwa Silbenkernmarken oder Phongrenzen, ist allerdings keineswegs so trivial, wie man zunächst annehmen könnte. Die damit verbundene Problematik soll hier anhand eines allgemein bekannten Beispiels aus dem Bereich der Mechanik verdeutlicht und eine Lösung für phonetische Fragestellungen geliefert werden.

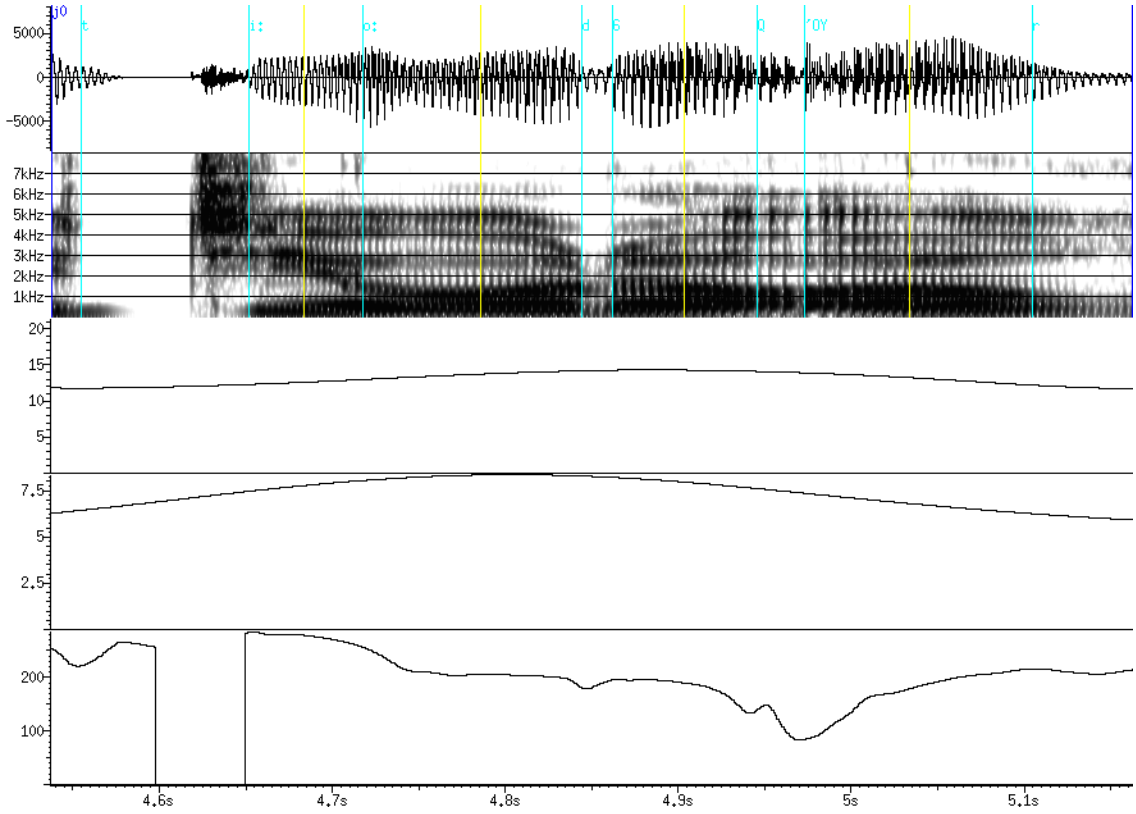

Abb. 6.4: Stimulus *j0* aus dem Signal *sscn7030* (Lautfolge [Iti: o:d6 QOYr] aus den Wörtern *Intercity oder Eurocity*). Oszillogramm und Sonagramm mit Phongrenzen- und Silbenkernmarken, lokaler Phongeschwindigkeitsverlauf (Phone/s), lokaler Silbengeschwindigkeitsverlauf (Silben/s) und Grundfrequenzverlauf (Hz).

Bereits in der Einleitung (S. 123) wurde auf die Vergleichbarkeit zwischen der Silben- bzw. Phongeschwindigkeit und der Geschwindigkeit eines Kraftfahrzeugs angespielt.

Wir nehmen einmal an, daß sich an einem Reifen eines Kraftfahrzeugs ein Impulsgeber befindet. Dieser soll bei jeder vollständigen Reifenumdrehung einen Impuls auslösen (siehe Abb. 6.5). Wird das Zeitsignal des Impulsgebers aufgezeichnet, so ergibt sich während der Fahrt eine Folge von Zeitmarken, deren Intervalle bei konstanter Geschwindigkeit des Fahrzeugs im Rahmen der Meßgenauigkeit gleich groß sind.

Aufgrund dieser Meßdaten läßt sich bereits leicht die Anzahl der Impulse innerhalb eines

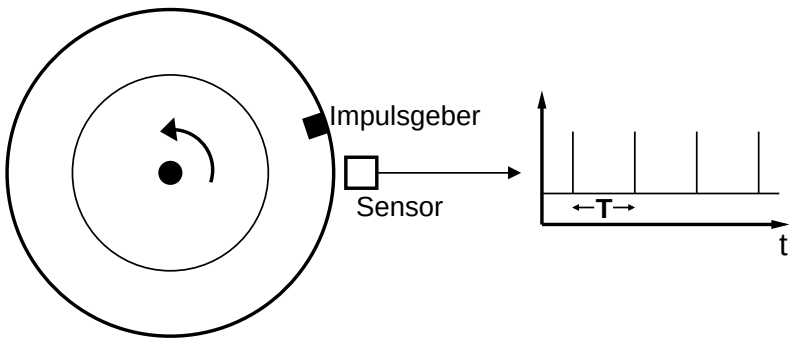

Abb. 6.5: Ein Impulsgeber an einem sich drehenden Rad liefert bei jeder vollständigen Umdrehung einen elektrischen Impuls. Aus den Zeitintervallen zwischen den Impulsen läßt sich die Geschwindigkeit ableiten.

konstanten Zeitintervalls, z.B. während einer Sekunde, zählen und mit dem Reifenumfang (gemessen z.B. in Metern) multiplizieren, um so die Geschwindigkeit des Kraftfahrzeugs in der physikalischen Einheit m/s zu erhalten, die sich ohne weiteres durch Multiplikation mit dem Faktor 3.6 in eine uns wesentlich geläufigere Geschwindigkeitsangabe, in km/h umrechnen ließe. Diese Messung könnte z.B. jede Sekunde wiederholt werden, um etwa die Geschwindigkeitsanzeige im Kraftfahrzeug regelmäßig zu aktualisieren.[4]

Auf der anderen Seite könnte man die Dauer des Zeitintervalls zwischen zwei aufeinanderfolgenden Impulsen messen. Auf diese Weise wäre bei der Vollendung jeder Reifenumdrehung ein neuer Geschwindigkeitswert möglich, indem der Reifenumfang durch die gemessene Dauer dividiert wird.

Im ersten Fall messen wir *Strecke pro Zeit*, im zweiten Fall *Zeit pro Strecke*. Sind diese beiden Meßmethoden äquivalent oder gibt es einen systematischen Unterschied?

Im Gegensatz zur ersten Meßmethode, die unabhängig von der Geschwindigkeit eine konstante Anzahl an Meßwerten pro Zeit liefert, würde die zweite Methode bei hohen Geschwindigkeiten viele Werte liefern, aber bei niedrigen Geschwindigkeiten wenige. Und während die erste Methode die Impulsanzahl nur schlicht zählen und nicht mit Nachkommastellen erfassen kann (da nicht bekannt ist, ob das Rad zwischen zwei Impulsen plötzlich beschleunigt oder verlangsamt wird) und damit einen beträchtlichen Meßfehler aufweist, ist die zweite Methode bezüglich jedes einzelnen Meßwertes exakt, aber die Zeitintervalle zwischen den Meßergebnissen sind unregelmäßig.

Übertragen wir diese Einsichten nun auf den Fall von Sprache: Bei Sprachsignalen führt die automatische sowie die manuelle Segmentation von Silben bzw. Phonen ebenfalls zu Signalen mit Folgen von Zeitmarken, die den Silbenkernen bzw. Phongrenzen entsprechen. Während jedoch bei einem fahrenden Kraftfahrzeug zwei unmittelbar aufeinanderfolgende Zeitintervalle zwischen den Meßimpulsen stark korreliert sind, da das Fahrzeug aufgrund seiner Masseträgheit nicht in beliebig kurzer Zeit die Geschwindigkeit ändern kann, sind Phondauern sehr stark von der jeweiligen Lautklasse, vom Betonungsgrad, von der Position im Satz und im Wort, von der Sprechgeschwindigkeit und von sprecherindividuellem Verhalten abhängig. Die Dauern benachbarter Phone sind demnach nur geringfügig korreliert. Dies darf auch für Silbendauern angenommen werden. Die Berechnung der *durchschnittlichen* Silben- und Phongeschwindigkeit über eine ganze Äußerung wird dadurch in keiner Weise erschwert. Problematisch gestaltet sich dagegen die Berechnung des lokalen Geschwindigkeitsverlaufs.

6.3.1 Eine erste Näherung

Eine erste Näherung der Lösung besteht darin, für die Dauer jedes Zeitintervalls zwischen zwei Impulsen — bzw. im Fall von Sprache zwischen zwei Segmentgrenzen — eine konstante Momentangeschwindigkeit anzunehmen, die dann wie folgt berechnet wird:

$$v_M(t) = \frac{1}{S_{i+1} - S_i}, \quad S_i \leq t < S_{i+1}, \tag{6.1}$$

wobei $v_M(t)$ die Momentangeschwindigkeit zum Zeitpunkt t ist, S_i der Anfangszeitpunkt des i-ten Segments und S_{i+1} der Endzeitpunkt des i-ten Segments sowie gleichzeitig der Anfangszeitpunkt des $i+1$-ten Segments.

[4] Auf diese Weise ergäbe sich zugleich eine *lokale* Geschwindigkeit, d.h. eine Geschwindigkeit an einem bestimmten Ort oder zu einem bestimmten Zeitpunkt, die sich schon im nächsten Augenblick durch Beschleunigen oder Bremsen wieder verändern könnte.

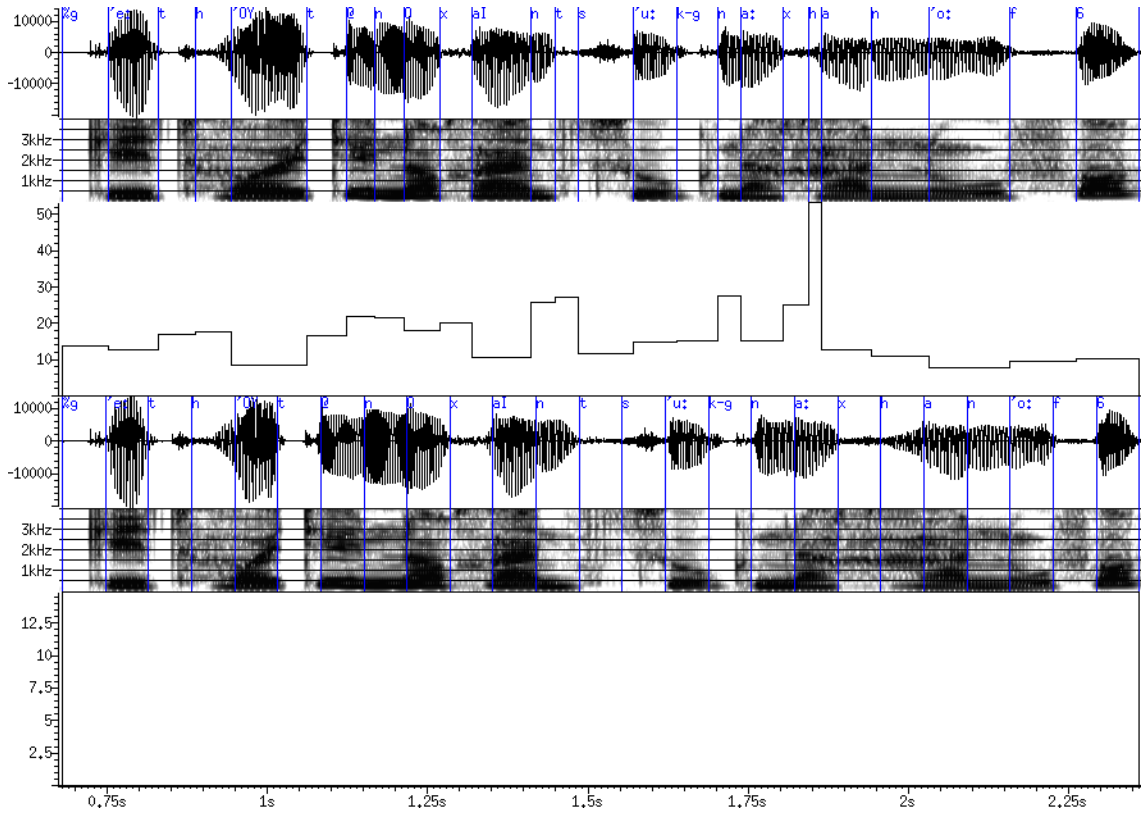

Abb. 6.6: Signal *awed5050* aus dem *PhonDatII*-Korpus: „*Geht heute noch ein Zug nach Hannover*“. Darunter: der durch die erste Näherung ermittelte *Momentan*-Phongeschwindigkeitsverlauf. Darunter: Das entsprechend dieses Verlaufs normalisierte Signal. Unten: Die lokale Phongeschwindigkeit des normalisierten Signals ist konstant 14.88 Phone/s.

In einer Umkehrung dieser Rechnung führt die Multiplikation einer Segmentdauer mit seiner Momentangeschwindigkeit trivialerweise zu einer Konstanten. Die Unterschiede zwischen den Segmentdauern werden also bei dieser Rechnung aufgehoben. Das Resultat dieser Rechnung angewendet auf ein Sprachsignal ist in Abb. 6.6 dargestellt. Zum einen wird hier anhand des Momentan-Phongeschwindigkeitsverlaufs deutlich, wie unterschiedlich die Dauern von Sprachlauten sein können, und zum anderen ist gut zu erkennen, daß die Anwendung der Rechnung tatsächlich zu einer Normalisierung der Segmentdauer führt: alle Phone haben nun dieselbe Dauer — in diesem Fall $67.2\text{ms} = \frac{1}{14.88\ \text{Phone/s}}$ —, da derart normalisiert wurde, daß das sich ergebende Signal die gleiche Gesamtdauer wie das Ursprungssignal haben sollte ($67.2\text{ms} \cdot 25\ \text{Phone} = 1.68\text{s}$ Gesamtdauer).

Diese Art der Normalisierung ist nicht in Einklang zu bringen mit unserer Auffassung von lokaler Sprechgeschwindigkeit als suprasegmentale Eigenschaft, da hier alle mikroprosodischen Dauervariationen gänzlich eliminiert werden. Hiermit sei belegt, daß ein Normalisierungsansatz unter Verwendung der tatsächlichen Momentangeschwindigkeit phonetisch uninteressant ist, da lediglich alle Lautdauern auf einen konstanten Wert abgebildet werden. Es muß also zur Ermittlung der lokalen Sprechgeschwindigkeit ein gleitender Durchschnitt über ein längeres Zeitintervall, das typischerweise mehrere Phone einbezieht, verwendet werden.

Bereits in Kap. 2.5 wurde das Prinzip der Berechnung lokaler Geschwindigkeitsverläufe aus linguistischen Einheiten erläutert. Allerdings verursacht dabei die Verwendung eines Rechteckfensters jedesmal dann, wenn während des schrittweisen Verschiebens des Fensters ein neues

Segment in das Fenster hinein- oder ein berücksichtigtes herausfällt, artefaktische Diskontinuitäten in der resultierenden Geschwindigkeitskurve. Diese können deutlich reduziert werden durch folgende Rechenvorschrift, die die lokale Geschwindigkeit zwischen den Zeitpunkten L und R angibt und die ich bereits 1996 [169] entwickelt hatte:

$$\text{Geschwindigkeit}_{LR} = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} + \frac{w_R-S_r}{S_{r+1}-S_r} + r - l - 1}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i}. \tag{6.2}$$

Hier sind S_i die Anfangszeitpunkte der jeweiligen Segmente und w_L bzw. w_R die linke bzw. rechte Endbegrenzung des Fensters. Insbesondere sind S_l das linke und S_r das rechte Segment, die beide jeweils nur teilweise ins Fenster fallen und daher anteilig gewichtet werden. Doch behält das Rechteckfenster trotz dieser anteiligen Gewichtung seine meßtechnisch nachteilige Eigenschaft, daß von der Fenstermitte weit entfernt liegende Segmente den das jeweilige Fenster repräsentierenden Geschwindigkeitswert ebenso stark beeinflussen wie in der Mitte liegende Segmente.[5]

6.3.2 Das verwendete Meßverfahren

Aus meßtechnischer Sicht bietet sich daher statt des Rechteck- eher das Hanning-Fenster an, da letzteres zentrale Segmente stärker und marginale geringer gewichtet und so einen stetigen Verlauf der resultierenden Kurve garantiert. Die einfachste Lösung für die Einführung des Hanning-Fensters besteht im Falten der Momentangeschwindigkeitsfunktion, wie sie in Formel (6.1) definiert wurde und z.B. in Abb. 6.6 zu sehen ist, mit dem gewünschten Hanning-Fenster und kann auch als Glätten interpretiert werden. Eine mathematisch äquivalente und rechnerisch erheblich weniger aufwendige Formel läßt sich durch Verwendung von Gewichtungen der beteiligten Segmente, die auf der Stammfunktion der gewünschten Fensterfunktion basieren, finden. So gilt für das in seiner Gesamtfläche auf 1 normalisierte Hanning-Fenster:

$$\int_a^b (1 - \cos x)\,\mathrm{d}x = \Big[x - \sin x\Big]_a^b = b - a + \sin a - \sin b. \tag{6.3}$$

Da nicht die Fläche eines Segments unter der Fensterfunktion, sondern die mittlere Höhe als Gewicht für das jeweilige Segment ausschlaggebend ist, ergibt sich aus (6.3) die gesuchte Gewichtungsfunktion H für ein Segment, das bei a beginnt und bei b endet, wie folgt:

$$H(a,b) = 1 + \frac{\sin a - \sin b}{b - a}, \quad 0 \le a, b \le 2\pi. \tag{6.4}$$

Nun müssen lediglich die Zeitpositionen von w_L, w_R, und S_i, die beim Aufruf von W in einer beliebigen Einheit (wie z.B. Samples, Sekunden oder Millisekunden) vorliegen können, für H auf den Bereich von 0 bis 2π abgebildet werden:

$$W(x,y) = H\left(2\pi\frac{x - w_L}{w_R - w_L}, 2\pi\frac{y - w_L}{w_R - w_L}\right). \tag{6.5}$$

Durch Verwendung der Gewichtungsfunktion W aus (6.5) in der bereits eingeführten Formel (6.2) ergibt sich schließlich der gewünschte Ausdruck:

$$\text{Geschwindigkeit}_{LR} = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} W(w_L, S_{l+1}) + \frac{w_R-S_r}{S_{r+1}-S_r} W(S_r, w_R) + \sum_{i=l+1}^{r-1} W(S_i, S_{i+1})}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i}. \tag{6.6}$$

[5] An dieser Stelle muß ungeklärt bleiben, wie stark die *wahrgenommene* Sprechgeschwindigkeit eines Sprachsignalausschnitts durch Randsegmente im Vergleich zu zentralen Segmenten bestimmt wird.

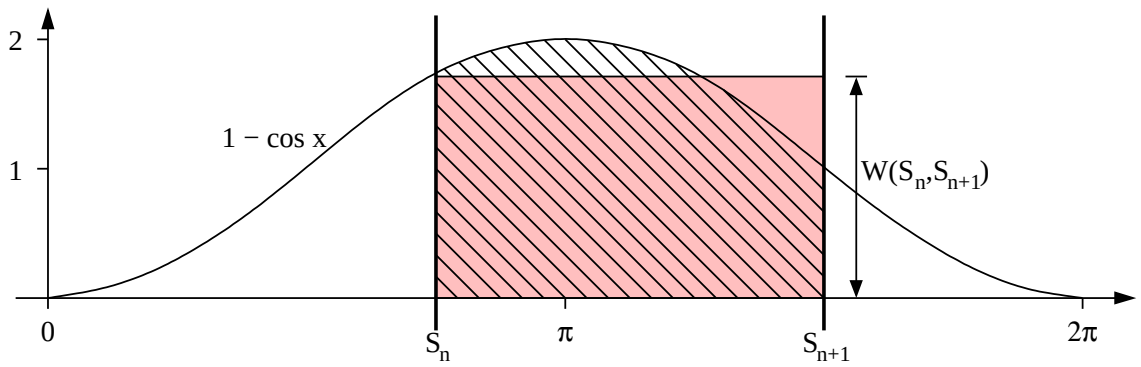

Abb. 6.7: Geometrische Interpretation der Gewichtungsfunktion $W(S_n, S_{n+1})$ aus Formel (6.5).

Zur Berechnung aller in dieser Arbeit verwendeten Silben- und Phonraten wurde das Verfahren (6.6) mit einem Hanning-Fenster von 625 ms Dauer angewendet. Diese kritische Dauer hat sich erst später aufgrund der Perzeptionsexperimente ergeben, soll aber der Vollständigkeit halber hier schon einmal genannt werden.

Es muß darauf hingewiesen werden, daß der Ausdruck im Nenner der Formel zwar in den meisten Fällen der Fensterlänge (in Samples) entspricht. Aber bei Beginn bzw. Ende einer Äußerung wird das Fenster am linken bzw. rechten Ende kürzer, da es links über das erste bzw. rechts über das letzte Segment der Äußerung hinausgeht. In diesen Fällen muß noch jeder Wert von W in (6.6) durch einen Korrekturwert ($W(S_1, w_R)$ bzw. $W(w_L, S_N)$) dividiert werden, da die Teilfläche und damit die mittlere Gewichtung des Teils der Fensterfunktion, der das Signal überdeckt, dann nicht mehr 1 ist. Dieses Vorgehen garantiert auch in den Randbereichen von Äußerungen einen stetigen und korrekten Verlauf der Geschwindigkeitskurve.

6.3.3 Detektion von Äußerungsabschnitten und Sprechpausen

Bei den hier durchgeführten Berechnungen von Silben- vs. Phongeschwindigkeiten treten praktische Unterschiede auf, die allerdings nicht von der Art der zugrundeliegenden linguistischen Einheiten abhängen, sondern davon, ob Anfang und Ende einer Einheit segmentiert wurden oder deren Zentrum. Beiden Fällen liegen nicht überlappende Segmentierungen zugrunde: das Ende eines Phonsegments ist bei den hier verwendeten Daten immer der Anfang des darauffolgenden Phonsegments und für jeden Silbenkern wurde nur *eine* Mittensegmentation durchgeführt.

Im Fall der Phongeschwindigkeit sind Beginn und Ende eines Bereiches, innerhalb dessen sukzessive lokale Geschwindigkeitswerte zu berechnen sind, *per se* gegeben: durch die Anfangszeitmarke des ersten Segments und die Endzeitmarke des letzten Segments eines bündig segmentierten Äußerungsabschnitts. Insbesondere sind Sprechpausen innerhalb einer Äußerung daran erkennbar, daß sie entweder keine Anfangs- und Endmarke besitzen oder daß sie segmentiert und mit einem eindeutigen Pausen-Label versehen sind. Letzteres muß dann dem Berechnungsverfahren mitgeteilt werden, damit dort keine lokalen Geschwindigkeitswerte geliefert werden.

Im zweiten Fall, in dem nur die Zentren von Einheiten, z.B. die Mitten der Silbenkerne, markiert wurden, kann weder eine Sprechpause, die sich zwischen zwei Marken befinden könnte, erkannt werden, noch können Anfangs- oder Endzeitpunkt einer Äußerung festgestellt werden, da jede lautsprachliche Äußerung immer schon vor dem Zentrum der ersten Einheit beginnt und auch nach dem Zentrum der letzten Einheit endet. Hier sind drei Lösungen zur Erkennung von Sprechpausen und Äußerungsabschnitten denkbar:

1. Das erste Lösungsverfahren arbeitet mit zwei einstellbaren Zeitkonstanten: die eine bestimmt, oberhalb welcher Dauer zwischen zwei Marken eine Sprechpause vermutet werden muß, weil derart lange Einheiten bei natürlichem Sprechen unüblich sind, die andere legt fest, wie weit — ausgehend von der ersten bzw. letzten Marke — in Richtung der erkannten Sprechpause extrapoliert werden soll, da näherungsweise angenommen werden kann, daß ein Äußerungsabschnitt etwa die halbe Dauer einer typischen Einheit früher beginnt als die erste Segmentmittenzeitmarke und entsprechend später endet als die letzte.
2. Äußerungsbeginn und -ende sind mit akzeptabler Fehlerrate automatisch durch spezielle Algorithmen detektierbar, welche bekannt sind aus dem Bereich der Grundfrequenzextraktion und dort bestimmen, wo stimmhaft, stimmlos oder gar nicht gesprochen wurde. Es ergibt sich allerdings das Problem der Differenzierung zwischen Verschlußphasen von Plosiven und Glottalverschlüssen auf der einen Seite und Sprechpausen auf der anderen Seite. Auch hier lassen Schwellwerte eine akzeptable Lösung zu.
3. Man kann diese Problematik des ungenauen Äußerungsbeginns und -endes bei der Geschwindigkeitsberechnung basierend auf Segmentmitten dadurch umgehen, daß man die tatsächlichen Anfänge und Enden von ununterbrochenen Äußerungsabschnitten aus einer anderen Segmentationsspur — etwa der Phonsegmentation — extrahiert, die diese Information enthält.

In den dieser Arbeit zugrundeliegenden Geschwindigkeitsberechnungen wurde das dritte Lösungsverfahren bevorzugt, da es lediglich von der Qualität der Segmentation abhängig ist und bei händischer Segmentation nahezu fehlerfreie Ergebnisse liefert.

6.4 Messung der Grundfrequenz

Zur Messung von Grundfrequenzverläufen stehen zahlreiche Algorithmen zur Verfügung, von denen die älteren von Hess 1983 [79] umfassend und detailliert dargestellt wurden. Bis auf Grenzfälle, wie etwa Glottalisierungen, sowie die jeweils ersten und letzten Glottisperioden von stimmhaften Äußerungsabschnitten liefern die meisten Verfahren Verläufe mit nur wenigen Hertz Abweichung untereinander. In dieser Arbeit fallen diese geringen Abweichungen nicht ins Gewicht; daher wird an dieser Stelle auf die Darstellung des verwendeten Standardverfahrens verzichtet.

6.5 Meßergebnisse: Silbenrate, Phonrate und Grundfrequenz

Die Ergebnisse der akustischen Analyse von Silben- und Phonraten der 141 Trainingsstimuli sind in Abb. 6.8 dargestellt (vgl. auch Abb. 1.1 auf S. 124 und Abb. 5.1 auf S. 168) und die der 96 Teststimuli in Abb. 6.9 (vgl. auch Abb. 5.2 auf S. 169 und Abb. 5.3 auf S. 170). Anhand eines Vergleichs der beiden Streuungsdiagramme wird unmittelbar deutlich, daß Silben- und Phonraten der spontansprachlichen Stimuli für die Perzeptionsexperimente 5 und 6 einen höheren Korrelationskoeffizient aufweisen als bei denjenigen Stimuli, die auf gelesener Sprache basieren.

Man könnte nun einerseits vermuten, daß dies eine Folge des frei bestimmbaren und damit vielleicht bezüglich der Silbenstruktur einfacheren spontansprachlichen Vokabulars ist, und andererseits, daß im Fall von Spontansprache eine Bestrebung eintritt, durch Elisionen und Insertationen von Lauten und Silben die Kovariation von Silben- und Phonraten stärker zu koppeln. Beides sind sehr vage Hypothesen, die in der zukünftigen Forschung der Klärung bedürfen. Es

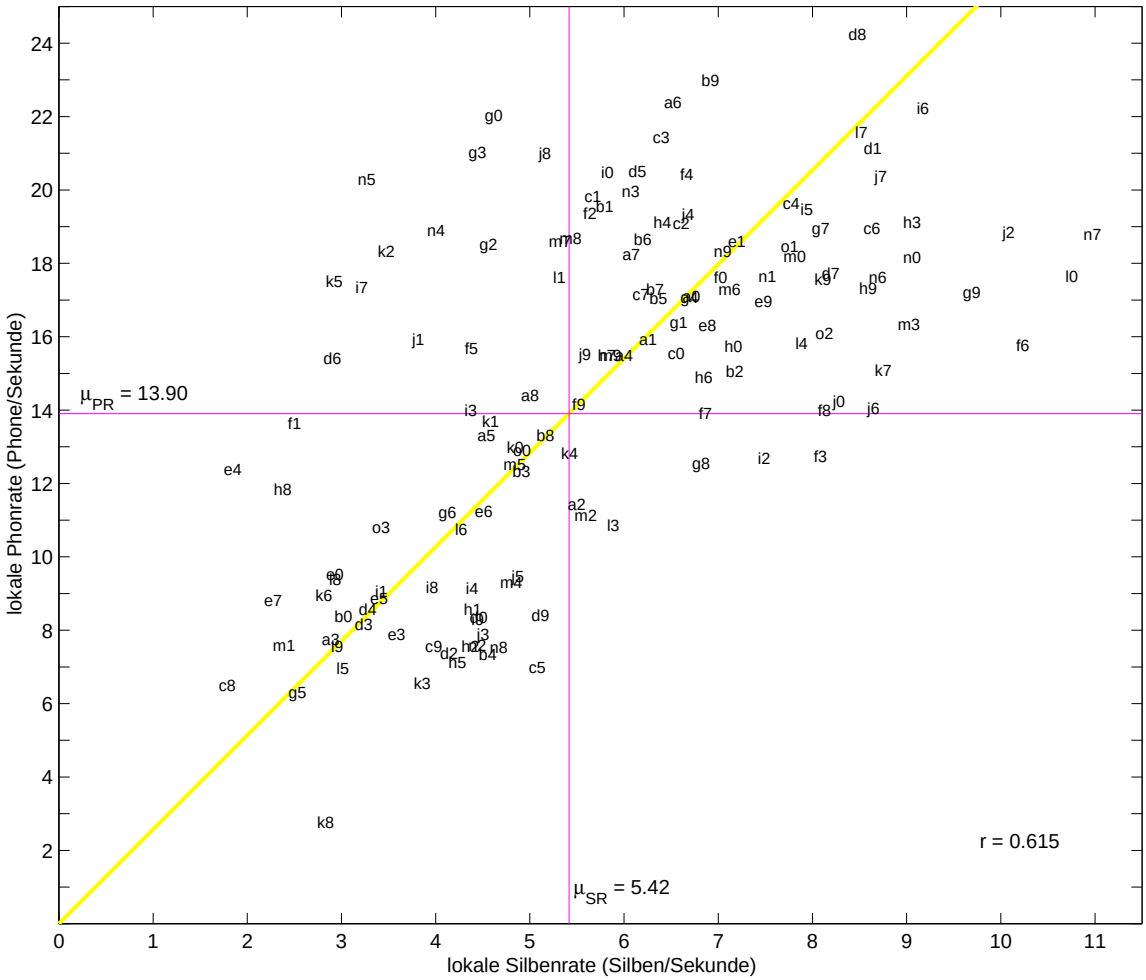

Abb. 6.8: Akustische Analyse der 141 Stimuli der Perzeptionsexperimente 1 bis 4 bezüglich ihrer lokalen Silben- und Phonraten. Die Steigung der Diagonale entspricht dem Verhältnis zwischen mittlerer Phon- und Silbenrate ($\mu_{\mathrm{PR}}/\mu_{\mathrm{SR}} = 2.57$, vgl. Abb. 1.1 auf S. 124 und Abb. 5.1 auf S. 168).

wird jedenfalls schon beim Vergleich der Rohdaten in Abb. 1.1 auf S. 124 und Abb. 5.3 auf S. 170 deutlich, daß die Korrelationskoeffizienten unterschiedlich sind.

Der Einfluß der lokalen Variation der Raten innerhalb der Stimuli auf die Sprechgeschwindigkeitsperzeption sollte ebenfalls erfaßbar sein. Deshalb wurden neben der Messung im Zentrum jedes Stimulus links und rechts im Abstand von 100 ms zum Zentrum zusätzlich die Silben- und Phonrate gemessen und anschließend die Standardabweichungen und Steigungen der drei Meßwerte für jeden Stimulus bestimmt.

Der Grundfrequenzverlauf innerhalb jedes Stimulus wurde ebenfalls extrahiert und auf Mittelwert, Standardabweichung sowie Steigung reduziert. Zusätzlich wurde die Indifferenzlage jedes der 16 Sprecher grob geschätzt, indem die mittlere F_0 über seine neun Stimuli berechnet wurde. Abschließend wurde die jeweilige Abweichung eines Stimulus vom zugehörigen mittleren F_0 berechnet, um feststellen zu können, ob die absolute Grundfrequenz oder eher die Abweichung von der Indifferenzlage perzeptiv relevant ist.

Weitere akustische Parameter, wie die von Morgan, Fosler & Mirghafori 1997 [146] vorgeschlagene Modulationsgeschwindigkeit der Energiehüllkurve[6] oder die von Samudravijaya, Singh

[6] Siehe zu weiteren Details bezüglich dieser Methoden S. 138.

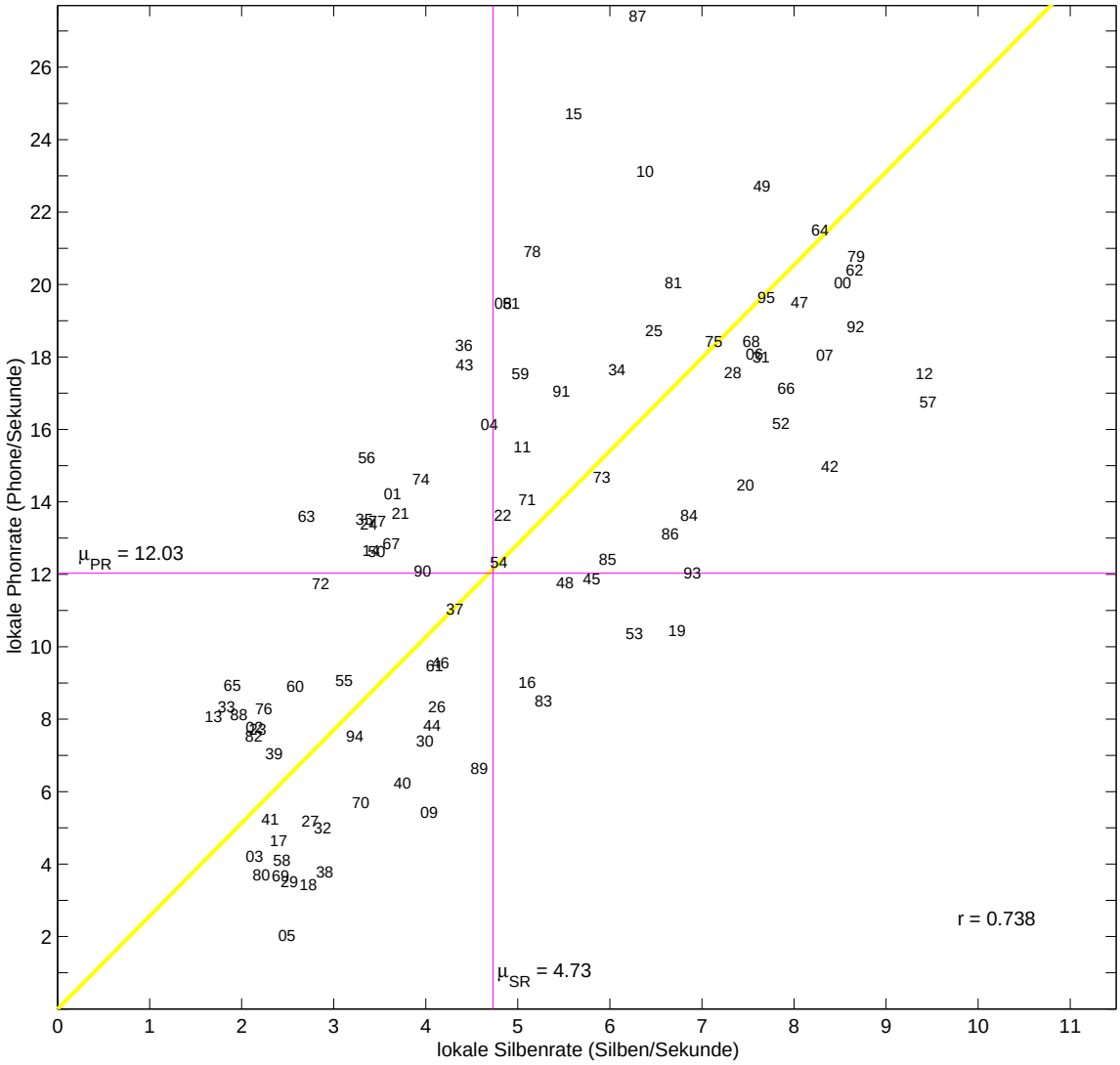

Abb. 6.9: Akustische Analyse der 96 Stimuli der Perzeptionsexperimente 5 und 6 bezüglich ihrer lokalen Silben- und Phonraten. Die Steigung der Diagonale entspricht dem Verhältnis zwischen mittlerer Phon- und Silbenrate ($\mu_{\mathrm{PR}}/\mu_{\mathrm{SR}} = 2.54$, vgl. Abb. 5.2 auf S. 169 und Abb. 5.3 auf S. 170).

& Rao 1998 [193] vorgeschlagenen drei Verfahren (die Dauern stationärer Sprachsignalpassagen, die Häufigkeit des Wechsels zwischen stimmhaften und stimmlosen Passagen und die Variation der Signalamplitude),[6] könnten an dieser Stelle neben den Resultaten eigener neu zu entwickelnder Verfahren zusätzlich extrahiert und in die Untersuchung einbezogen werden.

Aber einerseits würde sich an der grundsätzlichen Methodik der hier vorliegenden Arbeit durch die Hinzunahme dieser akustischen Informationen prinzipiell nichts ändern und andererseits erscheint eine Zweiteilung von akustischen Informationen in *i)* diejenigen, die auf phonetischen Einheiten basieren und Gegenstand dieser Untersuchung sind, und *ii)* diejenigen, die mit geringem phonetischem Wissen aus dem Sprachsignal extrahierbar sind und in einer zukünftigen Untersuchung betrachtet werden sollen, als angemessen.

Aber zumindest anhand von F_0, einem akustischen Merkmal, das zu letzterer Kategorie zu zählen ist, soll gezeigt werden, daß auch Vertreter dieser Kategorie mit rein akustischen Informationen leicht in unser noch zu entwickelndes Modell inkorporiert werden und dann unmittelbar zu Verbesserungen führen können.

7

Sechs Perzeptionsexperimente zur Sprechgeschwindigkeit

Die Idee, daß die Sprechgeschwindigkeit eine lokale Größe ist, die ihrerseits einer prosodischen Variation unterzogen wird, so daß sich zu jeder natürlich produzierten Äußerung eine ganz individuelle Sprechgeschwindigkeitskontur mit Maxima und Minima ergibt, die sich mit geeigneten Mitteln aus dem Sprachsignal extrahieren lassen müßte, ist uns in der Forschungsliteratur in dieser Form bisher noch nicht begegnet.

Um dieser Idee nachgehen zu können, haben wir die in den folgenden Abschnitten beschriebene Serie von Experimenten durchgeführt, deren primäres Gesamtziel darin besteht, zu den im vorigen Kapitel sorgfältig ausgewählten und akustisch analysierten Stimuli verläßliche Sprechgeschwindigkeitseinschätzungen zu erhalten. Die hier durchgeführten Perzeptionsexperimente gehen weit über die in Kap. 2 referierten Perzeptionsexperimente zur Sprechgeschwindigkeitsbeurteilung von Lass 1970 [120], Grosjean & Lass 1977 [71], Butcher 1981 [15], den Os 1985 [38] und Samudravijaya, Singh & Rao 1998 [193] hinaus.

Durch sechs aufeinander aufbauende Perzeptionsexperimente mit insgesamt 146 Versuchspersonensitzungen soll hier gezeigt werden, daß Probanden in der Lage sind, Sprachausschnitte bezüglich ihrer Sprechgeschwindigkeit konsistent zu beurteilen. Die ersten drei dieser sechs Experimente wurden dazu konzipiert, verschiedene Aspekte der Stimulusdauer bei der Wahrnehmung von Sprechgeschwindigkeit zu analysieren.

Ob und inwiefern die Stimulusdauer einen Einfluß auf die Wahrnehmung der Sprechgeschwindigkeit hat, spielt zwangsläufig eine zentrale Rolle bei Perzeptionsexperimenten zur Sprechgeschwindigkeit. Nur wenn das Perzeptionsergebnis nicht durch die Stimulusdauer bestimmt wird, darf sie in späteren Experimenten vernachlässigt werden. Anderenfalls muß ihre Einflußnahme sorgfältig erfaßt werden.

Im vierten Experiment besteht die Aufgabe von 60 Probanden darin, die Sprechgeschwindigkeiten von 141 Sprachstimuli zu schätzen, so daß für die anschließende Suche nach Relationen zwischen Akustik und Perzeption der Sprechgeschwindigkeit zusätzlich zur akustischen auch eine perzeptive Referenz zur Verfügung gestellt werden kann.

Am fünften Experiment nehmen 30 Versuchspersonen teil. Sie bestimmen die Sprechgeschwindigkeit von diesmal 100 Sprachstimuli, die auf Spontansprache basieren, um zu einem Testkorpus zu gelangen.

Schließlich müssen im sechsten Experiment zehn der 30 Probanden des fünften Experiments das selbe Experiment nach einem halben Jahr wiederholen, um feststellen zu können, wie hoch die Reliabilität bei dieser Art von Experimenten ist.

7.1 Versuchspersonen

Die Versuchspersonengruppen, die an den folgenden Experimenten teilnahmen, wurden jedesmal unterschiedlich zusammengesetzt, es sei denn, daß es bei Wiederholungsexperimenten darauf ankam, noch einmal dieselben Versuchspersonen heranzuziehen.

Im Durchschnitt setzte sich eine Versuchspersonengruppe zu 65% aus Studenten der Phonetik des zweiten bis sechsten Semesters, zu 25% aus Mitarbeitern des Instituts für Phonetik und Sprachliche Kommunikation in München und zu 10% aus fachfremden und teilweise auch älteren Menschen zusammen.

Manche der studentischen und fachfremden Versuchspersonen wurden für ihre Teilnahme am jeweiligen Experiment bezahlt. Einige Versuchspersonen nahmen an mehreren Experimenten teil; zwischen zwei Experimenten lagen in der Regel jedoch mehr als 6 Monate, so daß Lerneffekte ausgeschlossen werden können.

7.2 Experiment 1: Einfluß der Stimulusdauer auf den Schwierigkeitsgrad einer Sprechgeschwindigkeitsschätzung

Im Mittelpunkt des ersten Perzeptionsexperiments steht die Hypothese, daß die Dauer von Sprachausschnitten die Empfindung von Versuchspersonen, Sprechgeschwindigkeit leicht oder schwierig einschätzen zu können, nachweislich beeinflußt. Wir vermuten insbesondere, daß sehr kurze Stimuli als außergewöhnlich schwierig einschätzbar empfunden werden.

Um diese Hypothese überprüfen zu können, wurden Stimuli mit unterschiedlichen Dauern benötigt. Also wurde aus den in Kap. 5.5 vorgestellten 141 Stimuli per Zufall einer gestrichen, so daß 140 Stimuli vorlagen. Diese wurden derart in sieben Gruppen zu je 20 Stimuli aufgeteilt, daß in jeder Gruppe eine möglichst gleichmäßige Verteilung unterschiedlicher Stimuli gemäß Abb. 5.1 auf S. 168 erreicht wurde.

Die Stimuli der sieben Gruppen wurden nun auf sieben verschiedene Dauern (0.2 s, 0.4 s, 0.6 s, 0.8 s, 1.0 s, 1.2 s, 1.4 s) gekürzt, indem sie jeweils über 10 ms ein- und ausgeblendet wurden. Dann wurden sie über einen hochqualitativen Lautsprecher (*Pfleid FRS20R*) in einem akustisch unbehandelten, für Perzeptionsexperimente gut geeigneten Raum einem Auditorium von zwölf Probanden, die in ungefähr gleichem Abstand zum Lautsprecher saßen, in randomisierter Reihenfolge dargeboten. Die Versuchspersonen mußten ihre Urteile auf Antwortbögen eintragen. Sie wurden instruiert, jeden der 140 präsentierten Stimuli auf der fünfstufigen Skala, die in Abb. 7.1 dargestellt ist, gemäß der Frage zu bewerten, wie schwierig oder leicht ihnen eine Einschätzung der Sprechgeschwindigkeit fallen würde.

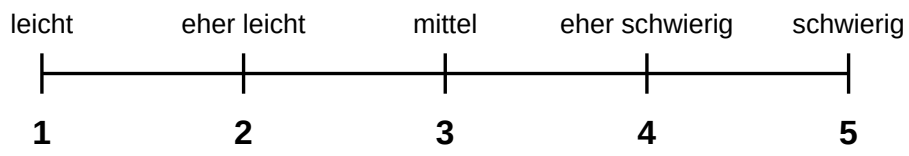

Abb. 7.1: Werte und Bedeutungen der Bewertungsskala des ersten Perzeptionsexperiments.

Dabei stellte sich heraus, daß die Probanden den kurzen Stimuli (200 und 400 ms) einen hohen Schwierigkeitsgrad zuordneten (siehe Abb. 7.2). Auch lange Stimuli mit 1400 ms erhielten einen

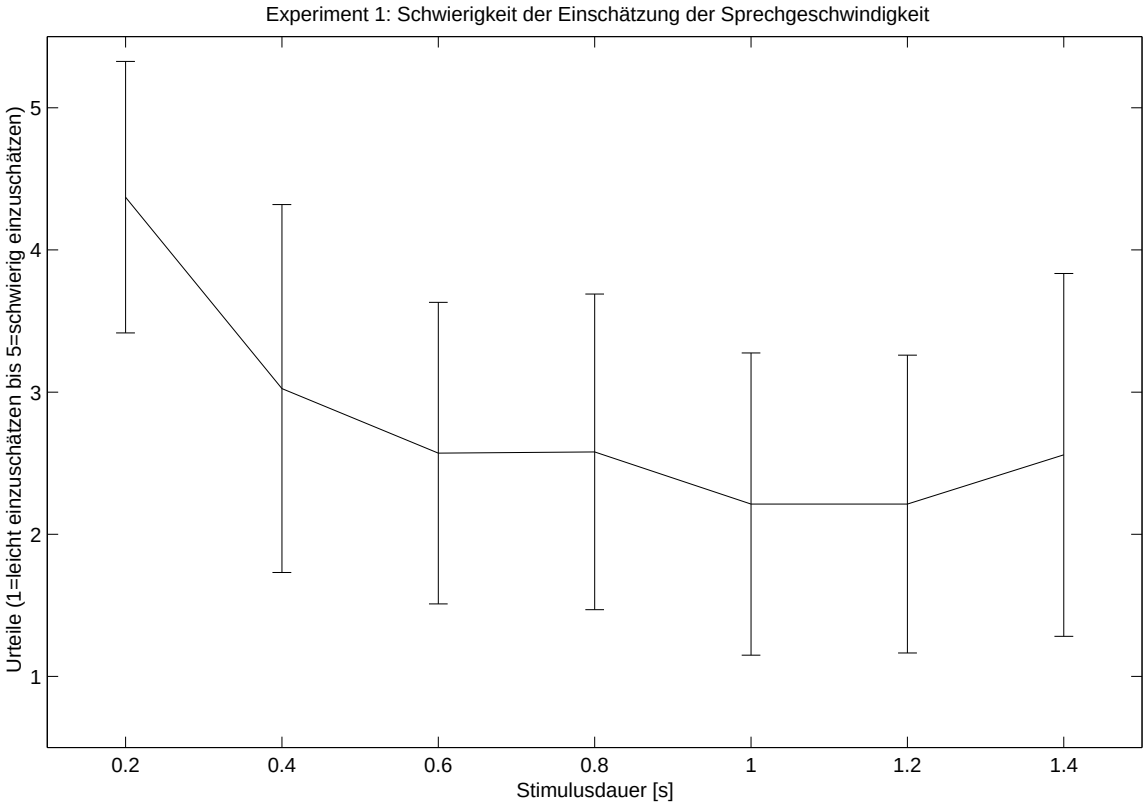

Abb. 7.2: Gesamtbeurteilungen der zwölf Probanden (Mittelwerte und Standardabweichungen über jeweils 240 Beurteilungen).

etwas höheren Schwierigkeitsgrad als Stimuli zwischen 600 ms und 1200 ms. Die Hypothese, eine gewisse Mindestdauer der Stimuli sei erforderlich, um eine Sprechgeschwindigkeitsbeurteilung zu ermöglichen, kann akzeptiert werden, da zumindest die 200-ms-Stimuli sehr oft mit dem höchsten Schwierigkeitsgrad bewertet wurden.

Eine nachträgliche Befragung der Versuchspersonen ergab, daß kurze Stimuli sehr wenige Anhaltspunkte zur Beurteilung lieferten. Dagegen würden lange Stimuli deswegen schwieriger zu beurteilen sein, weil sie oft deutliche Veränderungen der Sprechgeschwindigkeit enthielten.

Aufgrund dieses Ergebnisses sollten Stimuli für Wahrnehmungsexperimente zur Sprechgeschwindigkeit mit natürlichsprachlichen Äußerungen länger als 400 ms und kürzer als 1400 ms sein, um den Schwierigkeitsgrad bei der Beurteilung niedrig zu halten.

7.3 Experiment 2: Einfluß der Stimulusdauer auf die Wahrnehmbarkeit von Sprechgeschwindigkeitsvariationen

Aus der informellen Befragung der Teilnehmer des ersten Perzeptionsexperiments folgt unmittelbar die Hypothese für das zweite Perzeptionsexperiment: die Sprechgeschwindigkeit wird nur während kurzer Sprechpassagen als konstant empfunden. Ab einer gewissen Stimulusdauer können Probanden Sprechgeschwindigkeitsvariationen innerhalb der Stimuli wahrnehmen.

Um diese Hypothese überprüfen zu können, wurden Stimuli, Versuchsaufbau und Durchführung unverändert vom ersten Experiment übernommen. Diesmal nahmen 14 Probanden am Expe-

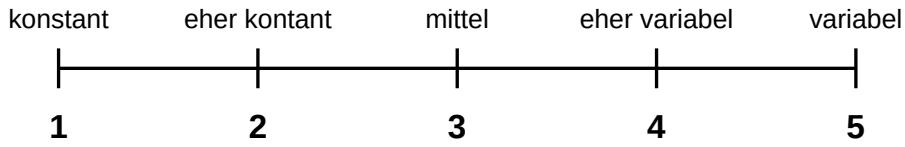

Abb. 7.3: Werte und Bedeutungen der Bewertungsskala des zweiten Perzeptionsexperiments.

riment teil, von denen sechs auch am ersten Experiment partizipierten. Sie wurden instruiert, auf der fünfstufigen Skala, die in Abb. 7.3 dargestellt ist, zu beurteilen, wie konstant bzw. variabel sie die Sprechgeschwindigkeit innerhalb des jeweils dargebotenen Stimulus empfinden.

Die Ergebnisse des zweiten Perzeptionsexperiments sind in Abb. 7.4 zusammenfassend dargestellt. Die Hypothese, daß nur kurze Sprachausschnitte die Empfindung einer konstanten Sprechgeschwindigkeit hervorrufen würden, kann akzeptiert werden, allerdings gilt dies nur bei 200 ms. Schon bei 400 ms Dauer beurteilen die Probanden die Sprechgeschwindigkeitsänderungen als deutlich ausgeprägter, und mit zunehmender Stimulusdauer werden die wahrgenommenen Variationen noch größer. Bei 1400 ms nimmt die Wahrnehmung von Sprechgeschwindigkeitsvariation im Vergleich zu Stimuli zwischen 800 ms und 1200 ms noch einmal deutlich zu.

Die Ergebnisse dieses Experiments legen nahe, daß Stimuli für Wahrnehmungsexperimente zur Sprechgeschwindigkeit mit natürlichsprachlichen Äußerungen möglichst kurze Dauern aufweisen sollten, um wahrnehmbare Variationen zu umgehen.

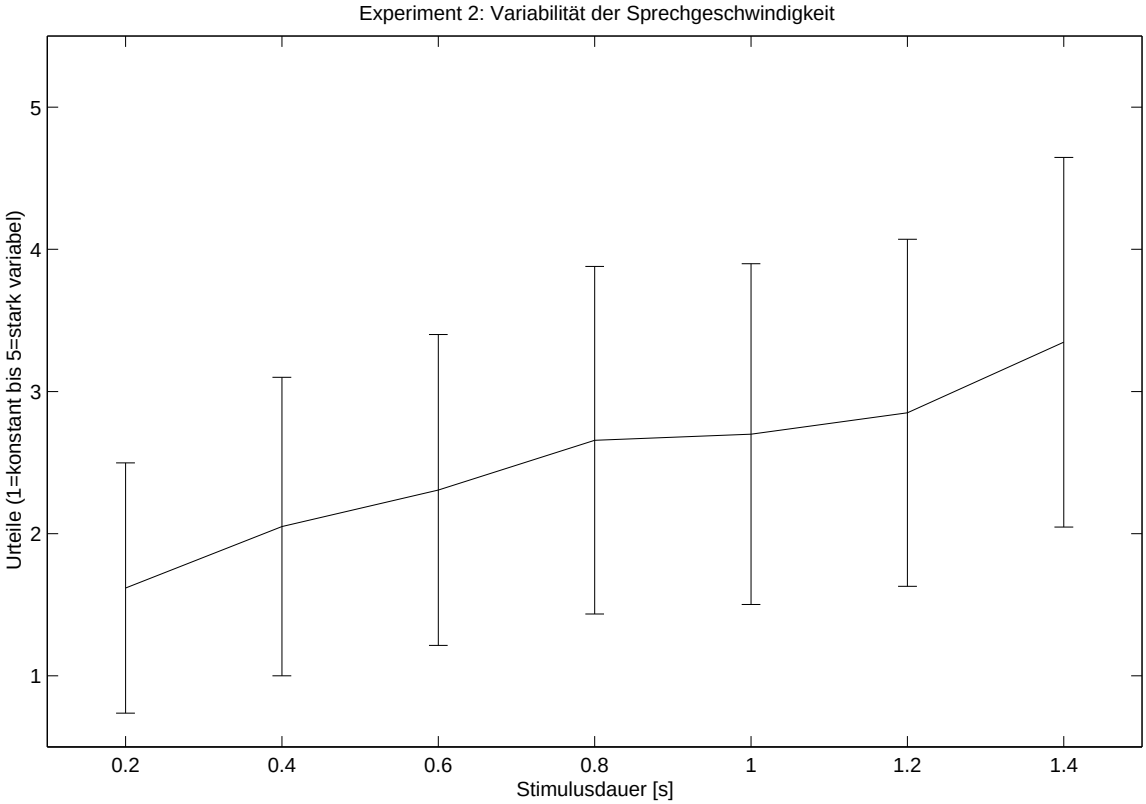

Abb. 7.4: Gesamtbeurteilungen der 14 Probanden (Mittelwerte und Standardabweichungen über jeweils 280 Beurteilungen).

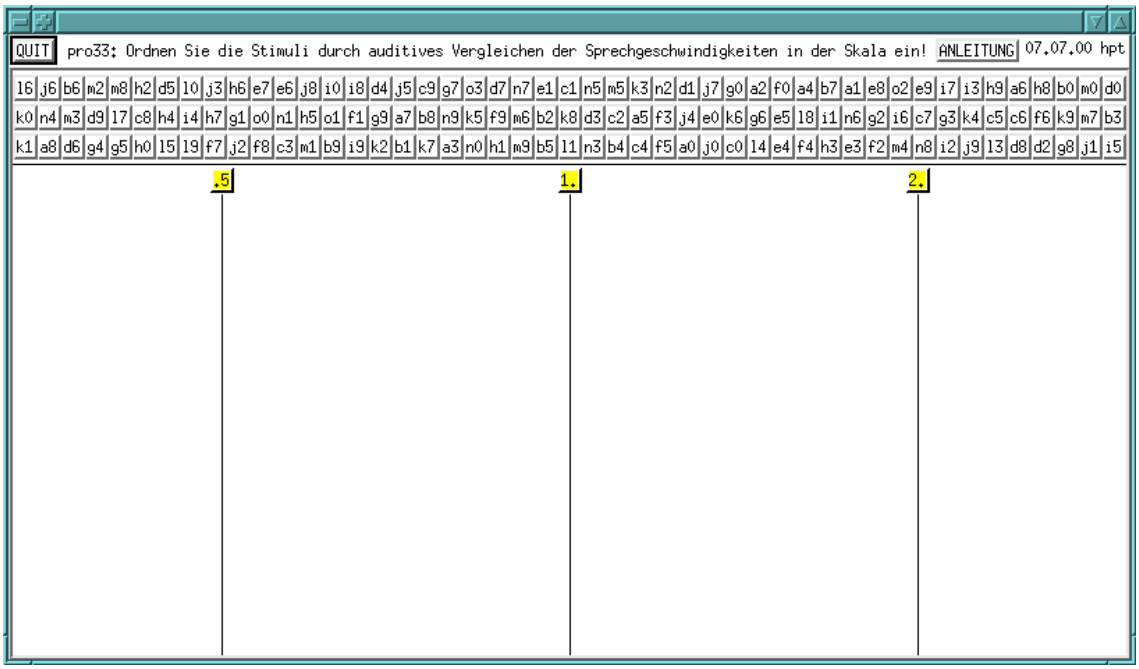

Abb. 7.5: Bedienungsoberfläche des computergestützten Perzeptionsexperiments vor Beginn des Experiments. Die zu beurteilenden Stimuli befinden sich noch im oberen Teil des Arbeitsbereiches (vgl. Abb. 7.6).

7.4 Ein neues computergestütztes Verfahren für Perzeptionsexperimente

Die beiden folgenden Perzeptionsexperimente werden mit einem neuen computergestützten Verfahren durchgeführt, das Identifikations- und Diskriminationstest miteinander vereint. Die Bedienungsoberfläche ist in Abb. 7.5 und 7.6 zu sehen. Die Probanden sitzen an Bildschirmen, tragen hochqualitative geschlossene Kopfhörer (*beyerdynamic DT770Pro*) und bedienen mit einer Maus ein speziell für die folgenden Untersuchungen entwickeltes Computerprogramm.

Die Aufgabe besteht darin, alle Stimuli, die als bewegliche Knöpfe symbolisiert sind und sich in ihrer Ausgangsposition im oberen Bereich der Bedienungsoberfläche befinden, anhand von perzeptiven Beurteilungen in eine Skala einzuordnen, die durch drei Ankerschalle horizontal aufgespannt wird. Die Höhe der Skala hat keine Bedeutung, sondern soll der Versuchsperson das Verschieben und Plazieren erleichtern (siehe Abb. 7.6). Beim Anklicken mit der Maus lassen die Ankerschalle einen langsamen, einen normalen und einen schnellen Sprachausschnitt erklingen, an denen sich die Versuchsperson orientieren soll.

Je langsamer gesprochen ein Stimulus empfunden wurde, umso weiter nach links sollte sein Knopf verschoben werden, und je schneller gesprochen wurde, umso weiter nach rechts. Die Beschriftungen der Knöpfe sind lediglich eine Hilfe für die Probanden, um bei Bedarf denselben Knopf wiederzufinden.

Die Versuchspersonen wurden zusätzlich instruiert[1], anfangs immer mit den Ankerschallen zu vergleichen und immer auch die Sprechgeschwindigkeitsabstände zu ihnen einzuschätzen. Sobald die ersten Stimuli in die Skala bewegt wurden, sollten auch diese immer wieder für Vergleiche herangezogen werden. Nach der Positionierung aller Stimuli sollte noch einmal jeder mit seinen

[1] Der vollständige Instruktionstext ließ sich während des Experiments jederzeit aufrufen. Er ist in Anhang A.1 auf S. 223 abgedruckt.

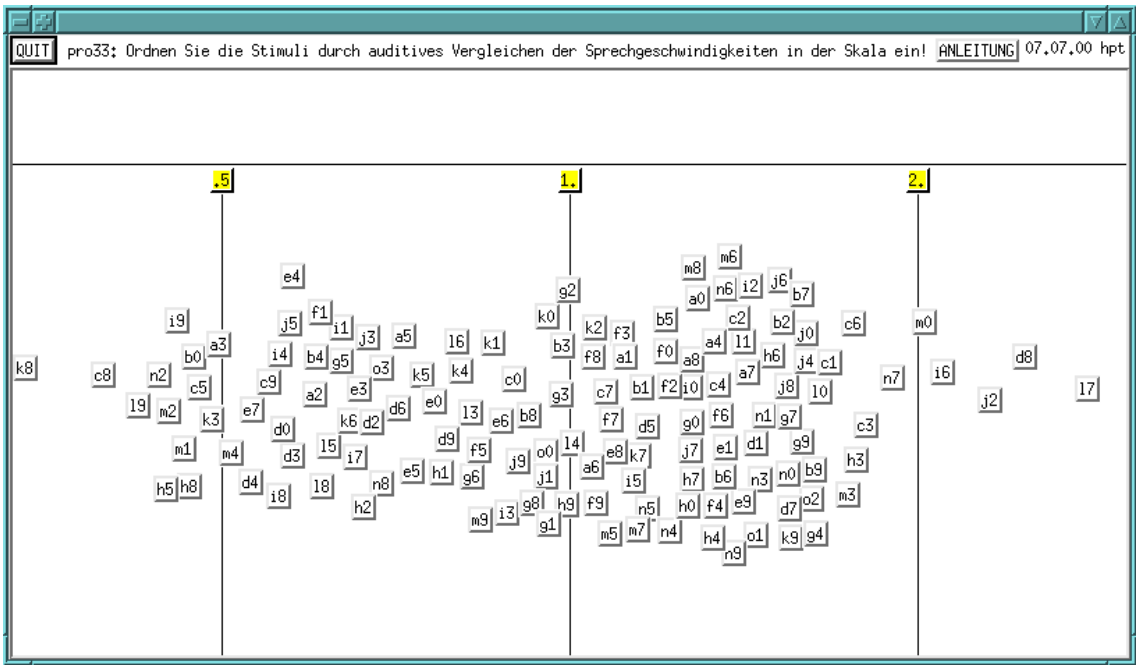

Abb. 7.6: Bedienungsoberfläche des computergestützten Perzeptionsexperiments. Hier bewegte Proband 33 (siehe Rohdaten auf S. 227, Abb. A.3) die Stimuli aus dem oberen Teil des Arbeitsbereiches in die Skala, die durch die drei Ankerschalle und drei senkrechte Linien aufgespannt wird (vgl. Abb. 7.5).

unmittelbaren Nachbarn verglichen werden und ggf. die Distanzen zwischen ihnen an den Grad ihres Sprechgeschwindigkeitsunterschieds angepaßt werden.

7.5 Experiment 3 (a und b): Einfluß der Stimulusdauer auf die perzipierte Sprechgeschwindigkeit

Nach den ersten beiden Experimenten ist nun bekannt und festzuhalten, daß kurze Sprachausschnitte unter 600 ms Dauer als schwierig einzuschätzen beurteilt werden, während eine größere Stimulusdauer zu einer besseren Wahrnehmbarkeit von Veränderungen der Sprechgeschwindigkeit innerhalb eines Stimulus führt. Demnach ist zu erwarten, daß die Stimulusdauer einen Einfluß auf die Einschätzung der Sprechgeschwindigkeit ausübt.

Durch zwei interaktive Perzeptionsexperimente, die im Abstand von sechs Monaten mit den gleichen zehn Probanden durchgeführt wurden, sollte nun quantitativ erfaßt werden, welchen Einfluß die Stimulusdauer auf die perzipierte Sprechgeschwindigkeit hat. Hierzu wurden 60 der in Kap. 5.5 beschriebenen Stimuli zufällig ausgewählt, wobei aber gewährleistet wurde, daß große und kleine Silben- und Phonraten gleich häufig auftraten.

Für den ersten Perzeptionstest (Experiment 3a) wurden alle 60 Stimuli auf eine Dauer von 625 ms gekürzt, indem am Anfang und am Ende der ursprünglichen Stimuli ein gleich langes Signalstück weggeschnitten wurde. Auch hier wurde über jeweils 10 ms ein- und ausgeblendet, um auffällige Schnitte zu vermeiden.

Für das zweite Experiment 3b wurden dieselben 60 Stimuli in fünf Gruppen zu je zwölf Stimuli eingeteilt, wobei jede Gruppe eine der Dauern 225, 425, 625, 825 und 1025 ms erhielt.

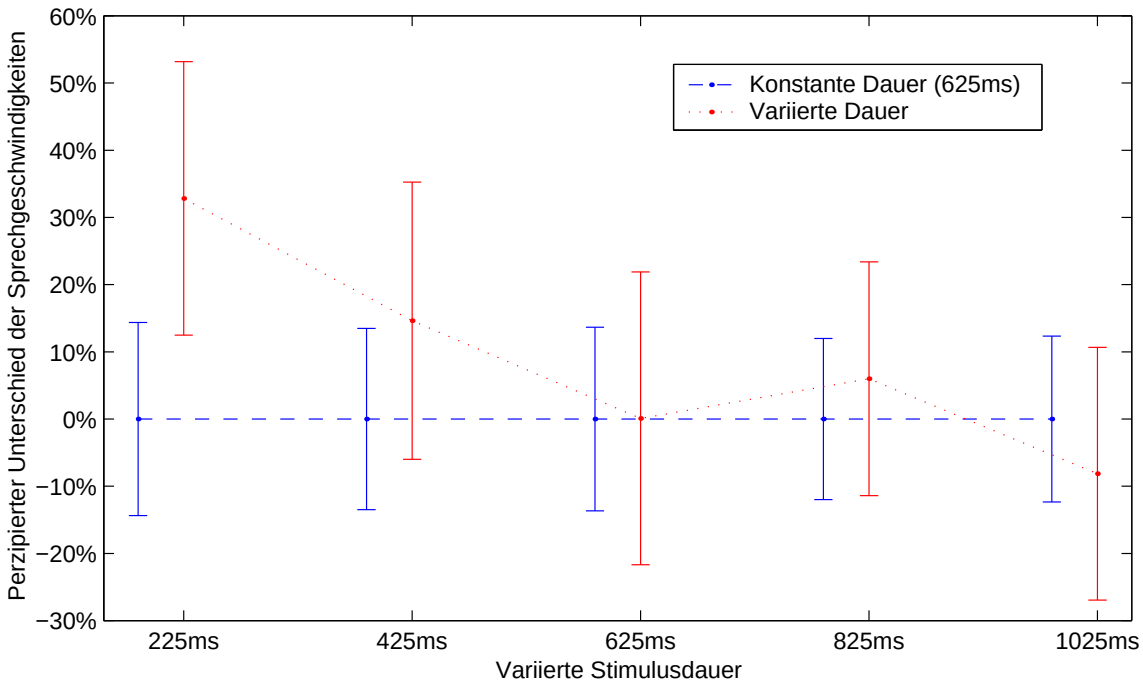

Abb. 7.7: Unterschiede in den perzipierten Sprechgeschwindigkeiten von Stimuli mit 625 ms Dauer und Stimuli mit variierter Dauer.

Die beiden Perzeptionsexperimente liefern abhängige Stichproben, die einen Vergleich der 625 ms dauernden Stimuli mit jeweils einer der fünf Gruppen zulassen. Insbesondere der Vergleich mit der Gruppe, die ebenfalls auf 625 ms gekürzt wurde, läßt einen Aufschluß über die Reliabilität des interaktiven Experiments zu.

Tabelle 7.1 sowie Abb. 7.7 zeigen, daß bei dem Experiment mit variabler Dauer die Varianzen aller fünf Gruppen durchweg größer sind als bei dem Experiment mit 625 ms Dauer. Die Versuchspersonen gaben an, daß es schwieriger sei, die Stimuli im Experiment mit unterschiedlichen Dauern zu vergleichen und einzuordnen.

Stimulusdauer	225 ms	425 ms	625 ms	825 ms	1025 ms
$\mu_2 - \mu_1$	32.8	14.6	0.102	5.99	-8.14
σ_1	14.4	13.5	13.7	12.0	12.3
σ_2	20.3	20.6	21.8	17.4	18.8
F	1.42	1.53	1.59	1.45	1.52
$F_{0.05;119/119} = 1.36$	$\star$	$\star$	$\star$	$\star$	$\star$
$\hat{t}$	14.4	6.50	0.043	3.11	3.96
f	214	205	200	211	205
α	0.001	0.001	0.1	0.005	0.001
$t_{f;\alpha}$	3.34	3.34	1.65	2.84	3.34
Signifikanzniveau	$\star\star\star$	$\star\star\star$	n.s.	$\star\star$	$\star\star\star$

Tabelle 7.1: Einfluß der Stimulusdauer auf die perzipierte Sprechgeschwindigkeit: Vergleich von 625 ms dauernden Stimuli mit den Stimuli variabler Dauer. (F-Test für Homogenität zweier Stichproben und t-Test für Stichproben mit inhomogenen Varianzen mit Korrektur der Freiheitsgrade nach Welch.)

Zu den Unterschieden bei den Mittelwerten gibt es drei bemerkenswerte Feststellungen, die in Abb. 7.7 besonders deutlich werden:

1. Die Kontrollgruppe mit 625 ms Stimulusdauer führt auch nach einem halben Jahr zu den gleichen Ergebnissen und bestätigt damit, daß die experimentelle Prozedur reliabel ist.
2. Kürzen der Stimuli auf 425 ms bzw. auf 225 ms führt dazu, daß die Versuchspersonen die Sprechgeschwindigkeit um 14.6% bzw. um 32.8% zu hoch einschätzen. Es zeigt sich also bei kürzeren Stimuli das aus anderen Domänen der Perzeptiven Phonetik bekannte Paradigma des *perceptual overshoot*.
3. Die Gruppe mit 825 ms dauernden Stimuli wird als 6% schneller empfunden, während die 1025-ms-Gruppe 8% langsamer erscheint. Mit anderen Worten, die Stimuli mit 625 ms Dauer werden gegenüber den 825-ms-Stimuli als langsamer bewertet, obwohl sie kürzer sind; offensichtlich greift das Paradigma des *perceptual overshoot* erst unterhalb von 625 ms. Das wechselnde Vorzeichen der Abweichung bei 825 ms vs. 1025 ms könnte dadurch entstehen, daß diese Stimuli zunehmend eine variierende Sprechgeschwindigkeit beinhalten, die zu unsystematisch abweichenden Urteilen führt.

Wir dürfen also festhalten, daß die Dauer der Stimuli einen gravierenden Einfluß auf die perzipierte Sprechgeschwindigkeit hat. Somit müssen also Perzeptionsexperimente mit variabler Stimulusdauer vorerst unterbleiben. Die Dauer käme sonst als weiterer die Wahrnehmung bestimmender Faktor zu den akustischen Merkmalen des Stimulus hinzu und müßte daher auch in den einfachsten Modellen berücksichtigt werden. Dann bestünde allerdings die Gefahr, daß allein die oben dargestellte einfache Beziehung des *perceptual overshoot*, sogar in einem Modell mit ungünstig gewählten akustischen Merkmalen, zu einer tendenziell stimmigen Vorhersage der perzipierten Sprechgeschwindigkeit führen könnte.

Durch die Wahl einer konstanten Stimulusdauer kann diese Variable bei dem hier zu entwickelnden Sprechgeschwindigkeitsmodell vorerst außer Acht bleiben. Da die akustischen Stimuli möglichst kurz sein sollten, um wenig variierende Sprechgeschwindigkeit zu enthalten, und bei 625 ms noch kein *perceptual overshoot* auftritt, wird diese Dauer für das folgende Perzeptionsexperiment gewählt.

7.6 Experiment 4: Einschätzung der lokalen Sprechgeschwindigkeit bei Lesesprache

Das Hauptziel des hier beschriebenen Experiments besteht darin, für die in Kap. 5.5 beschriebenen 141 Sprachstimuli perzeptive Beurteilungen zu erhalten. Diese können dann auf interindividuelle Varianz und Plausibilität geprüft werden. Aber vor allem stehen sie als Referenzdaten der Entwicklung eines Modells zur Verfügung, das in der Lage sein soll, eine Schätzung dieser Daten aus den akustischen Eigenschaften der Stimuli abzuleiten.

60 Versuchspersonen — vorwiegend Studentinnen und Studenten im Alter zwischen 25 und 35 Jahren, teilweise mit phonetischer Ausbildung — nahmen im Zeitraum von August 1998 bis Juni 1999 an dem Experiment teil. Sie wurden instruiert, unter Verwendung der in Kap. 7.4 vorgestellten computergestützten Versuchsmethode die 141 auf 625 ms geschnittenen Stimuli in die Sprechgeschwindigkeitsskala einzutragen.[2]

[2] Der vollständige Instruktionstext ist in Anhang A.1 auf S. 223 abgedruckt.

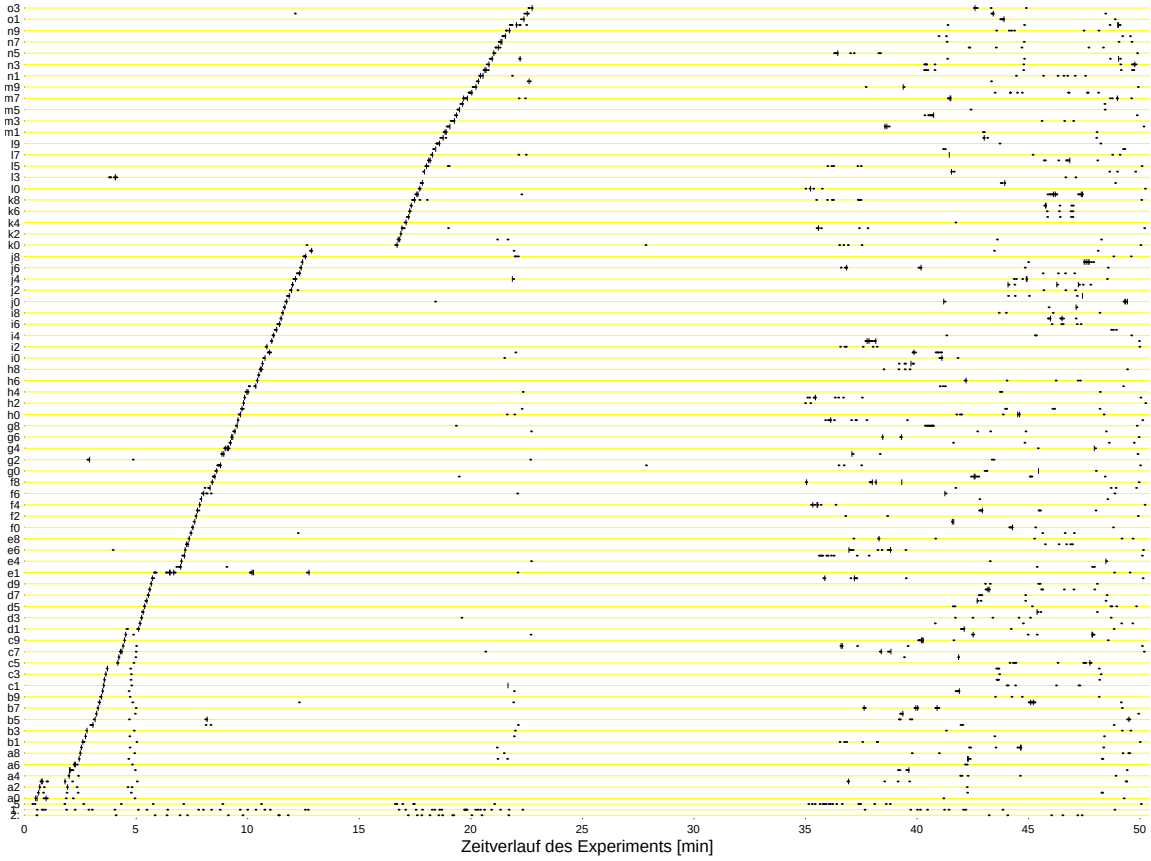

Abb. 7.8: Der Versuchsverlauf von Proband 38.

Da diese Versuchsmethode mit Hilfe einer Protokoll-Funktion für jede der 60 Versuchspersonen detailierte Informationen über jeden Versuchsdurchgang lieferte, seien hier wenigstens anhand der Daten von zwei ausgewählten Probanden typische Verhaltensweisen während des Versuchsablaufs dargestellt:

In Abb. 7.8 ist auf der X-Achse der zeitliche Ablauf des Experiments von Versuchsperson 38 zu betrachten, der etwa 50 Minuten dauerte. Auf der Y-Achse sind alle Stimuli von „*o3*“ bis „*a0*“ und die drei Ankerschalle „*.5*“, „*1.*“ und „*2.*“ aufgelistet. Diese Visualisierung ist vergleichbar mit einer Partitur-Lochkarte für z.B. Drehorgeln oder Player-Pianos. Es wird deutlich wie die Versuchsperson die Stimuli beginnend bei „*a0*“ bis „*a5*“ anhört[3], dabei immer wieder mit den Ankerschallen vergleicht, und auch verschiebt.[4]

Versuchsperson 38 hört in den ersten 22 Minuten alle Stimuli in der Reihenfolge ihrer bedeutungslosen Bezeichnungen an und sortiert vor, um dann eine mehr als zehnminütige Pause zu machen. Das Nachlassen der Leistungsfähigkeit vor der Pause deutet sich in Abb. 7.8 bereits durch die flachere Diagonale des letzten Stimulusblocks an, denn hier drückt der Winkel aus, wie lange VP 38 einen Stimulus bearbeitet, bevor sie zum alphanumerisch nächsten wechselt. Warum VP 38 die Stimuli ausgerechnet in dieser Reihenfolge bearbeitet, wird im Dunkeln bleiben müssen.

Dagegen beginnt Versuchsperson 11 (vgl. Abb. 7.9) gleich mit dem sehr gründlichen Abhören, Vergleichen und Einsortieren eines Stimulus nach dem anderen. Besonders auffällig ist, daß VP 11 während des gesamten Experiments sehr intensiven Gebrauch von den Ankerschallen

[3] Das Anhören eines Stimulus ist grafisch durch einen kleinen Punkt verdeutlicht.

[4] Das Verschieben eines Stimulus ist durch einen senkrechten Strich gekennzeichnet.

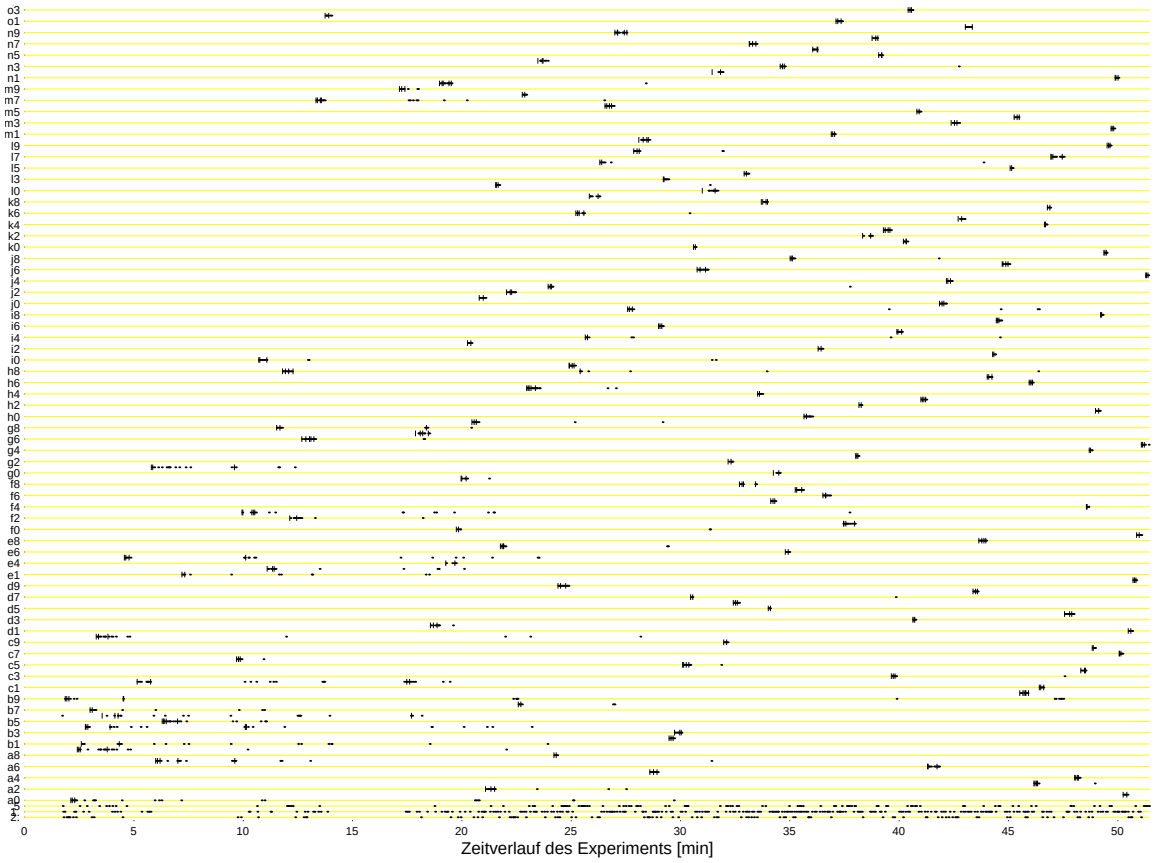

Abb. 7.9: Der Versuchsverlauf von Proband 11.

macht, wohingegen VP 38 nach ihrer Pause noch seltener als zuvor auf die Ankerschalle zugreift. Statt dessen hört sie in den letzten Minuten noch einmal alle Stimuli — teilweise auch mehrfach — an, verschiebt aber nur noch ganz selten. Offensichtlich hat sie sich diesbezüglich genau an die Instruktionen gehalten.

Über diese ganz direkte Interpretation der Versuchsprotokolle hinaus lassen sich bemerkenswerte Feststellungen machen, die mit einem herkömmlichen Versuchsaufbau unmöglich gewesen wären: So muß etwa hervorgehoben werden, daß Stimuli mit großen Standardabweichungen öfter verschoben und abgehört werden als solche mit geringeren Standardabweichungen.

Zusammenfassend läßt sich folgendes sagen: im Durchschnitt benötigte ein Proband für das Experiment 51 Minuten, verschob jeden Stimulus 3.5 mal und hörte ihn 10 mal ab. Außerdem griff er 39 mal auf den langsam gesprochenen Ankerschall zurück, 82 mal auf den normalen und 42 mal auf den schnell gesprochenen. Das Experiment bewertete er häufig als „anstrengend, aber machbar“.

Abb. 7.10 zeigt die Rohdaten[5] des Experiments (vgl. mit Abb. 7.6). Auf der Y-Achse sind alle Stimuli dargestellt und auf der X-Achse die Sprechgeschwindigkeitsskala. Da alle Probanden jeden Stimulus beurteilten, existieren für jeden Stimulus 60 Urteile, aus denen Mittelwerte (Kreuze in der Abbildung) und Standardabweichungen (graue waagerechte Balken) berechnet und in die Abbildung eingezeichnet wurden. Außerdem sind die Stimuli entlang der Y-Achse nach ihren Mittelwerten geordnet.

[5] In Anhang A.2 ab S. 224 sind die Korrelationen aller Probanden mit ihrem Gruppenmittelwert dargestellt und in Anhang A.3 ab S. 230 die Urteile zu jedem Stimulus als Histogramme.

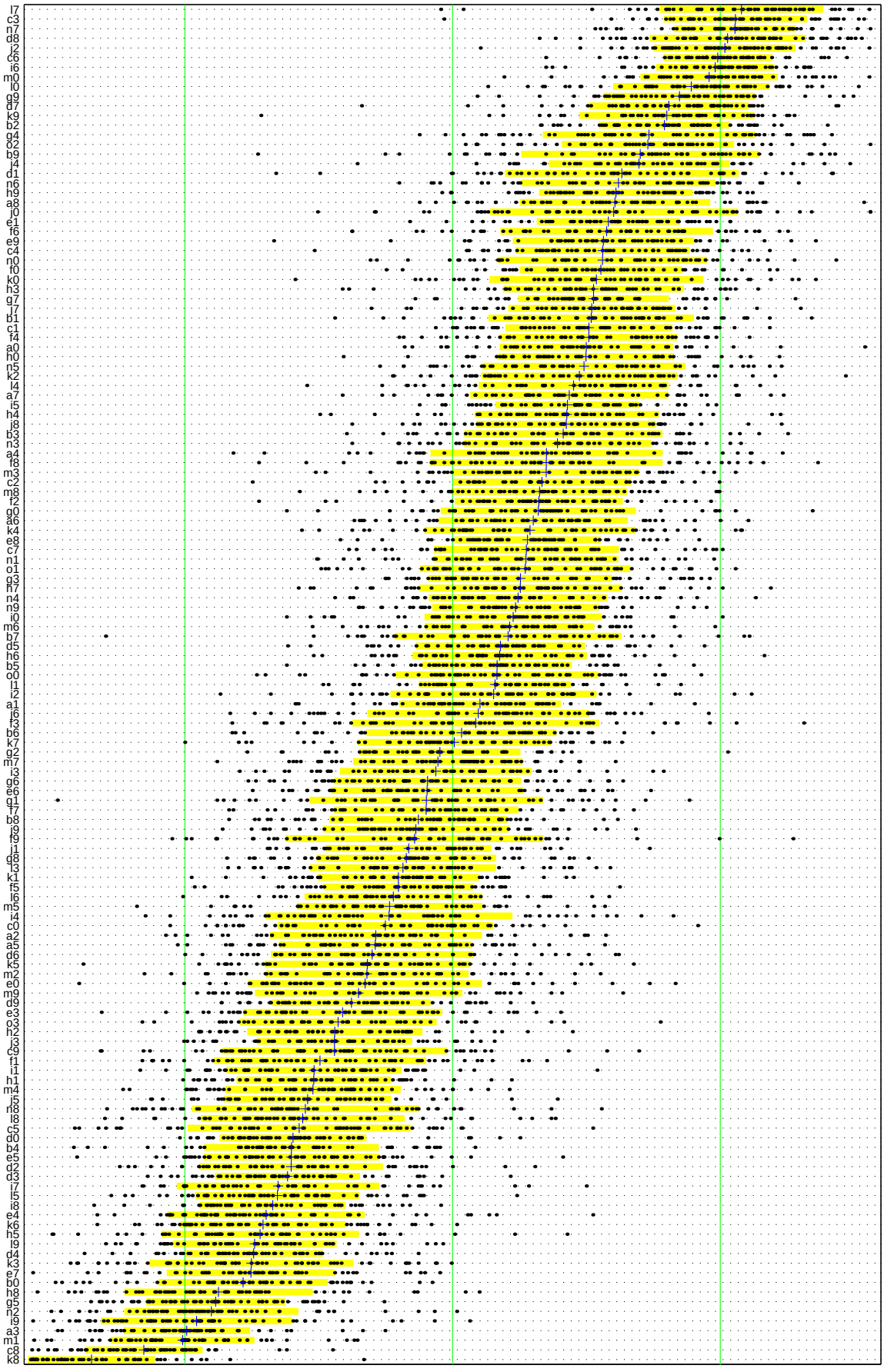

Abb. 7.10: Die Urteile aller 60 Probanden zu 141 Stimuli, die nach Mittelwerten sortiert wurden.

Die 60 Urteile zu jedem Stimulus folgen der Gauß-Verteilung und lassen sich daher vollständig durch Mittelwert und Standardabweichung beschreiben. Da auch frühere Auswertungen der Daten bereits mit weniger Versuchspersonen (Pfitzinger 1998/99 [170, 171]) Gauß-Verteilung offenbarten, wird eine weitere Erhöhung der Versuchspersonenzahl voraussichtlich nur noch eine — wenn auch gewünschte — Verringerung der Vertrauensbereiche nach sich ziehen.

Eine zweifaktorielle Varianzanalyse der Rohdaten mit einer Beobachtung pro Zelle führte zu hoch signifikanten Ergebnissen für den Faktor „Stimulus“ ($F(140, 8260) = 186.2, p < 0.001$) und den Faktor „Proband“ ($F(59, 8260) = 25.04, p < 0.001$). Beide Faktoren haben also bedeutenden Einfluß auf die Variation der Perzeptionsurteile. Eine Varianzaufklärung offenbarte allerdings, daß der Faktor „Stimulus“ 72.82% der Varianz erklärt, während der Faktor „Proband“ nur für 4.13% verantwortlich ist und der Rest der Varianz von 23.05% unerklärt bleibt. Hierbei könnte es sich um Wechselwirkungen handeln, also etwa individuell abweichende Bewertungsstrategien bezüglich ganz bestimmter Stimuli bei ansonsten vergleichbarem Verhalten. Diese Interaktionen lassen sich allerdings mit nur einer Beobachtung pro Zelle nicht nachweisen.

Die Ergebnisse dieser Varianzaufklärung lassen sich mit denen von den Os 1985 [38] vergleichen:[6] In ihrem zweiten Experiment erklärt die bei ihren neun verschieden schnell produzierten Stimuli auftretende Tempovariation 83.6% der Gesamtvarianz. Es ist allerdings nachvollziehbar und zu erwarten, daß eine Erhöhung der Stimuluszahl auf ein Vielfaches von Neun mehr Widersprüche in den Perzeptionsurteilen und damit niedrigere Korrelationen und eine geringere erklärte Varianz hervorruft. Die in unserem Experiment resultierende erklärte Varianz von fast drei Vierteln der Gesamtvarianz erscheint hinreichend groß.

Offensichtlich gibt es interindividuelle Unterschiede, doch sie haben nur einen Bruchteil des Einflusses des Faktors „Stimulus“ auf die Perzeptionsergebnisse. Folglich eignen sich die hier gewonnenen Daten sehr gut als Repräsentation der perzipierten Sprechgeschwindigkeit und liefern damit zugleich abgesicherte Informationen zu den Stimuli, die gemeinsam mit den Resultaten der in Kap. 6 erfolgten akustischen Untersuchungen zum Vergleichen herangezogen werden dürfen.

7.7 Experiment 5: Einschätzung der lokalen Sprechgeschwindigkeit bei Spontansprache

In diesem Experiment ging es um die Herstellung der Testdaten für die spätere Modellevaluation. Daher kamen hier die in Kap. 5.6 beschriebenen Stimuli zum Einsatz, die sich sowohl in den Sprechern als auch im gesprochenen Material und im Sprechstil vom vorangegangenen Experiment unterschieden.

30 Versuchspersonen — wie bereits in Experiment 4 vorwiegend Studentinnen und Studenten im Alter zwischen 25 und 35 Jahren und teilweise mit phonetischer Ausbildung — nahmen im Zeitraum von Juni bis November 2000 an diesem Experiment teil. Sie wurden instruiert, unter Verwendung der in Kap. 7.4 vorgestellten computergestützten Versuchsmethode die 100 auf 625 ms geschnittenen spontansprachlichen Stimuli in die Sprechgeschwindigkeitsskala einzutragen.[7]

Eine zweifaktorielle Varianzanalyse der in Abb. 7.11 dargestellten Rohdaten[8] mit einer Beob-

[6] Siehe zu weiteren Details bezüglich ihrer Experimente S. 140.

[7] Die Veränderungen des Instruktionstexts im Vergleich zu Experiment 4 sind in Anhang B auf S. 237 erwähnt.

[8] In Anhang B ab S. 237 sind die Korrelationen aller Probanden mit ihrem Gruppenmittelwert dargestellt und in Anhang B.2 ab S. 241 die Urteile zu jedem Stimulus als Histogramme.

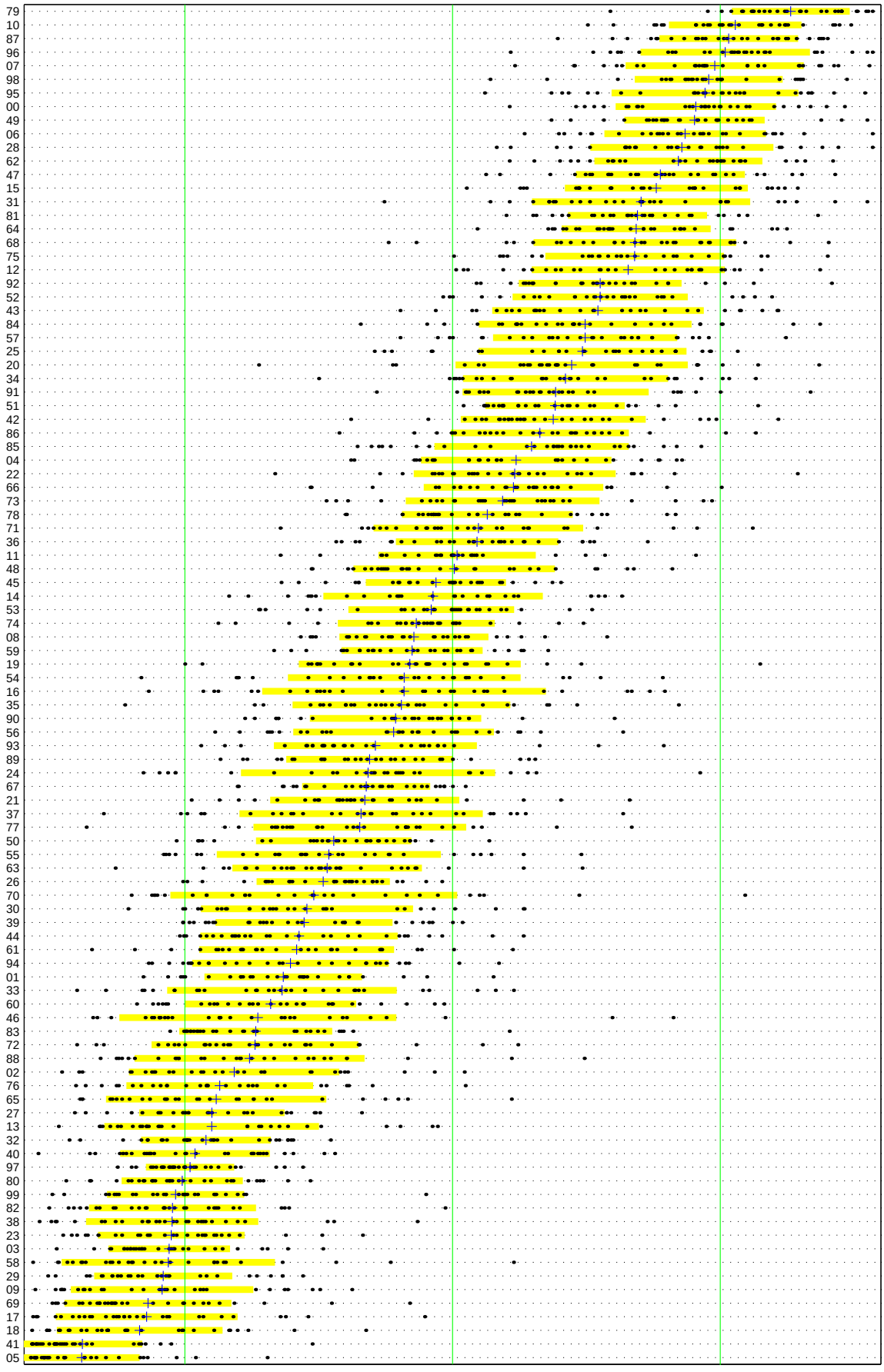

Abb. 7.11: Die Urteile aller 30 Probanden zu 100 Stimuli, die nach Mittelwerten sortiert wurden.

achtung pro Zelle führte auch bei diesem Experiment zu hoch signifikanten Ergebnissen für den Faktor „Stimulus“ ($F(99, 2871) = 165.9, p < 0.001$) und den Faktor „Proband“ ($F(29, 2871) = 35.45, p < 0.001$). Beide Faktoren haben also wieder bedeutenden Einfluß auf die Variation der Perzeptionsurteile. Die Varianzaufklärung zeigte, daß der Faktor „Stimulus“ 80.81% der Varianz erklärt, während der Faktor „Proband“ nur für 5.06% verantwortlich ist und der Rest der Varianz von 14.13% unerklärt bleibt.

Grundsätzlich gilt auch hier das bereits zu Experiment 4 Gesagte: *i)* Die unerklärte Varianz könnte durch divergierendes Verhalten verschiedener Versuchspersonen bei ganz bestimmten Stimuli hervorgerufen werden und *ii)* mit über vier Fünfteln durch den kontrollierten Faktor erklärter Varianz müssen diese Ergebnisse als ausgezeichnet eingestuft werden.

Möglicherweise hat im Vergleich zu Experiment 4 die durch den Faktor „Stimulus“ erklärte Varianz deswegen von 72.82% auf 80.81% zugelegt, weil *i)* es sich um weniger Stimuli hielt oder/und *ii)* die Stimuli leichter einzuschätzen sein könnten oder/und *iii)* zufällig mehr Versuchspersonen ausgewählt wurden, die eine ähnliche Bewertungsstrategie bei dieser Art der experimentellen Aufgabe anwenden, als im Experiment 4.

7.8 Experiment 6 (a und b): Reliabilität bei der Einschätzung der lokalen Sprechgeschwindigkeit

Im vierten und fünften Experiment wurden Einschätzungen der Sprechgeschwindigkeiten von insgesamt 238 verschiedenen Stimuli gesammelt. Aber ob man sich während der Entwicklung von Modellen zur Sprechgeschwindigkeitsbestimmung darauf verlassen darf, ist noch nicht hinreichend geprüft worden. Diese Aufgabe sollen die beiden Teile des sechsten Experiments erfüllen.

7.8.1 Teil a: Vergleich von vier Stimuli des vierten und fünften Experiments

Die 100 Stimuli des fünften Experiments umfaßten 96 Stimuli aus dem *VerbmobilI*-Korpus und zusätzlich vier Stimuli aus dem vierten Experiment, um testen zu können, ob die Art der Zusammenstellung der Stimuli in einem Experiment Einfluß auf die Beurteilung eines Stimulus hat. Es sollte insbesondere die Hypothese geprüft werden, daß die Positionierung eines Stimulus aufgrund des häufigen Vergleichens mit den benachbarten Stimuli durch dieselben beeinflußt wird. So würden die Ankerstimuli ihren Einfluß verlieren und sich die Beurteilung auf der Skala verschieben.

Wie bereits in Kap. 5.6 auf S. 169 ausgeführt wurde, handelt es sich bei den vier Stimuli um zwei langsame und zwei schnelle Stimuli, so daß sie zwei zweistufigen Faktoren zugeordnet werden können (*Sprechgeschwindigkeit* und *Stimulus*).

Mit Hilfe der statistischen GLM-Analyse (*General Linear Model*) wurde der Einfluß der zweistufigen Faktoren *Experiment*, *Stimulus* und *Sprechgeschwindigkeit* auf die Varianz der Beurteilungen der vier Stimuli getestet. Die Ergebnisse dieser Analyse sind in Tab. 7.2 aufgelistet. Es kann festgestellt werden, daß das Experiment und damit auch die Zusammenstellung der übrigen gleichzeitig dargebotenen Stimuli keinen signifikanten Einfluß auf die Beurteilung hat. Dagegen hat die Sprechgeschwindigkeit erwartungsgemäß hoch signifikanten Einfluß. Aber auch der vergleichsweise geringe Sprechgeschwindigkeitsunterschied zwischen dem ersten und zweiten Stimulus entpuppt sich als hoch signifikant. Interaktionen zwischen den verschiedenen unabhängigen Faktoren sind nicht signifikant.

Faktor	Freiheitsgrade	F	p	Signifikanz
Experiment	1	0.337	0.562	n.s.
Stimulus	1	10.502	0.001	$\star\star\star$
Rate	1	6057.522	<0.001	$\star\star\star$
Experiment$\times$Stimulus	1	1.079	0.300	n.s.
Experiment$\times$Rate	1	3.653	0.057	$(\star)$
Stimulus$\times$Rate	1	0.701	0.403	n.s.
Experiment$\times$Stimulus$\times$Rate	1	0.969	0.326	n.s.

Tabelle 7.2: Jeweils dieselben zwei langsamen und zwei schnellen Stimuli (Faktoren *Rate: langsam vs. schnell* und *Stimulus: erster vs. zweiter*) kamen in beiden Experimenten vor (Faktor *Experiment*). Während die vier Stimuli in den beiden Experimenten nicht unterschiedlich eingeschätzt wurden, führte sowohl die Sprechgeschwindigkeit wie auch der zweite gegenüber dem ersten Stimulus aus der gleichen Sprechgeschwindigkeitsregion zu hoch signifikant unterschiedlichen Beurteilungen.

Zusammenfassend läßt sich sagen, daß die statistische Auswertung anhand der vier ausgewählten Stimuli eindeutig belegt, daß der Wechsel der Stimulusanzahl und des Sprechstils der übrigen gleichzeitig dargebotenen Stimuli von einem Experiment zum anderen keinen signifikanten Einfluß auf die Positionierung der vier Stimuli hatte. Ursache dafür könnten die gleich gebliebenen Ankerschalle sein, die damit ihre Funktion, die Ausnutzung der Beurteilungsskala in vergleichbarer Weise zu ermöglichen, voll erfüllen.

7.8.2 Teil b: Wiederholung des fünften Experiments

Ziel des zweiten Teils des sechsten Experiments ist die Überprüfung der Reliabilität des fünften Experiments, um dadurch zugleich eine grundsätzliche Aussage über Experimente dieser Art zu erhalten.

Zur Teilnahme am Perzeptionsexperiment wurden zehn derjenigen 30 Probanden gebeten, die bereits am fünften Experiment teilgenommen hatten. Sie wurden aber nicht zufällig ausgewählt, sondern aufgrund der Kriterien einer möglichst hohen Korrelation und geringen mittleren Abweichung zum Gruppenmittelwert des fünften Experiments.[9] Zudem mußten dieselben Versuchspersonen über einen längeren Zeitraum verfügbar sein.

Durch diese Auswahlkriterien sollte zusätzlich getestet werden, ob für brauchbare Ergebnisse bei Perzeptionsexperimenten zur Sprechgeschwindigkeit bereits zehn gezielt selektierte Probanden ausreichen könnten. Dabei wurde allerdings in Kauf genommen, daß dieses Experiment eben wegen dieser Auswahlkriterien seine Validität verlieren könnte.

Der Versuchsaufbau entsprach in allen Einzelheiten dem fünften Experiment, fand allerdings für jede Versuchsperson etwa sechs Monate nach ihrer Sitzung im Experiment 5 statt, so daß Gedächtnis-Effekte der Probanden weitgehend ausgeschlossen werden konnten. Zu erwarten ist ohnehin, daß die Probanden sich nicht an die im vorangegangenen Experiment gewählten Positionen auch nur eines Stimulus erinnern konnten — zumal ihnen verschwiegen wurde, daß es sich um dieselben Stimuli des alten Experiments handelte. Aber es sollte auch sichergestellt werden, daß eine während des alten Experiments eventuell entwickelte Strategie in Vergessenheit gerät.

Zuerst wurde wieder eine zweifaktorielle Varianzanalyse der Rohdaten mit einer Beobachtung pro Zelle durchgeführt. Wie bei Experiment 4 und 5 lieferte auch Experiment 6 hohe Signi-

[9] Siehe zu den Einzelergebnissen der 30 Probanden von Experiment 5 Anhang B ab S. 237.

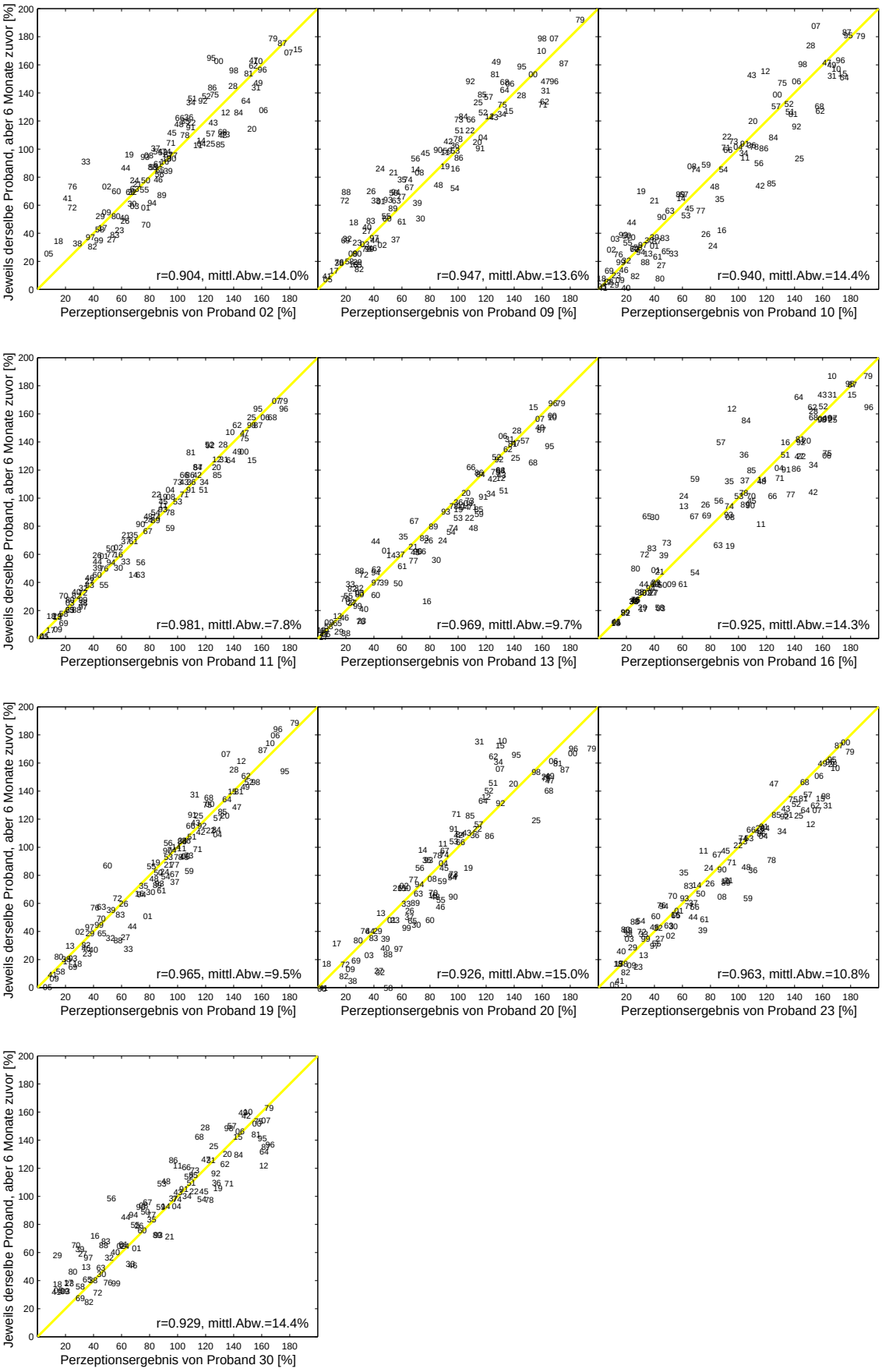

Abb. 7.12: Korrelationen der Ergebnisse von 10 Probanden mit ihren Ergebnissen 6 Monate zuvor.

fikanz der Faktoren „Stimulus“ ($F(99, 891) = 85.29, p < 0.001$) und „Proband“ ($F(9, 891) = 12.41, p < 0.001$). Beide Faktoren haben wieder bedeutenden Einfluß auf die Variation der Perzeptionsurteile. Die Varianzaufklärung zeigte hier, daß der Faktor „Stimulus“ 89.38% der Variation erklärt und damit gegenüber Experiment 4 und 5 noch weiter zugenommen hat, während der Faktor „Proband“ sogar nur noch für 1.18% verantwortlich ist, und nur eine Restvariation von 9.44% unerklärt bleibt.

Diese Ergebnisse sind insofern bemerkenswert, als nun mit großer Sicherheit festgestellt werden kann, welcher der in Experiment 5 genannten Erklärungsansätze für die Zunahme der aufgeklärten Varianz verantwortlich sein muß: Während nämlich sowohl die Stimuluszahl wie auch die Stimuli selbst gleich geblieben sind, hat die Versuchspersonenzahl abgenommen. Weniger Versuchspersonen können aber ebenso stark voneinander abweichen wie viele Versuchspersonen. Daher muß die Ursache in einer sehr ähnlichen Bewertungsstrategie dieser Probanden liegen.

Es muß nun auch in Erwägung gezogen werden, daß sich diese zehn Versuchspersonen trotz sehr ähnlicher Bewertung anders verhalten, als sich die Grundgesamtheit verhalten hätte, man also hoch reliable Ergebnisse erhält, die aber nicht valide sind. Daher soll im folgenden durch Gruppenvergleiche festgestellt werden, ob diese zehn Versuchspersonen sich anders verhalten als die 20 des vorigen Experiments.

Aber zunächst ist in Abb. 7.12 für jeden der 10 Probanden eine Korrelation seiner aktuellen Perzeptionsergebnisse mit denjenigen, die er sechs Monate zuvor geliefert hat, dargestellt. Bei der knappen Mehrheit der Versuchspersonen (*02, 09, 10, 19, 20* und *30*) sind die Korrelationskoeffizienten r und die mittleren Abweichungen mit sich selbst schlechter oder genauso gut wie ihre alten Perzeptionsergebnisse verglichen mit den Gruppenmittelwerten im fünften Experiment (vgl. mit Abb. B.1 bis Abb. B.3 ab S. 238). Auch verglichen mit den aus Experiment 6 stammenden Mittelwerten über die jeweils übrigen neun[10] Versuchspersonen (siehe Abb. C.1 auf S. 248) zeigen die Korrelationskoeffizienten r und mittleren Abweichungen mit sich selbst wieder bei sechs Versuchspersonen (*02, 09, 10, 20, 23* und *30*) nahezu gleich gute oder schlechtere Werte.

Die Probanden *02, 09, 10, 20* und *30* korrelieren also in beiden Experimenten 5 und 6 etwa genauso gut oder besser mit dem Gruppenmittelwert als mit sich selbst. Vermutlich streuen sie in den beiden Durchgängen des Experiments zu beiden Seiten der Mittelwerte, was sich erklären ließe durch eine geringere Strategiebildung, die Entwicklung einer anderen Strategie als im fünften Experiment oder schlechteres Einschätzungsvermögen. Den tatsächlichen Grund müssen wir zwar schuldig bleiben, aber zumindest können wir feststellen, daß eine hohe Korrelation mit sich selbst nicht zwangsläufig auch eine hohe Korrelation mit dem Gruppenmittelwert nach sich zieht.

Wir wollen nun die Gruppenvergleiche nachholen. In Abb. 7.13 sind drei Streuungsdiagramme dargestellt: im oberen wurden die Versuchspersonen des fünften Experiments aufgeteilt in die Gruppe der zehn Probanden, die auch das Experiment 6 durchführten, und in die andere Gruppe der übrigen 20 Versuchspersonen. Hier zeigt sich, daß das mittlere Ergebnis der zehn Probanden sehr gut mit dem der 20 Probanden korreliert.[11] Wenn verschiedene Gruppen von Probanden im

[10] Hier handelt es sich um einen *one-to-many*–Vergleich (mit *one-leave-out*), den wir als „konservativ“ bezeichnen möchten, weil er um bis zu etwa 10% schlechtere Korrelationskoeffizienten und mittlere Abweichungen liefern kann als der Vergleich mit allen zehn Probanden, denn bei nur zehn Probanden beträgt der Einfluß des Einzelnen auf die Gruppenmittelwerte maximal 10%. Bei 30 oder 60 Probanden wie in Experiment 4 und 5 ergeben sich dagegen keine gravierenden Unterschiede zwischen Einbeziehen oder Auschließen des jeweiligen Probanden aus dem Mittelwert.

[11] In einer erschöpfenden Durchführung dieser Aufteilung, in der permutativ zwei gleich große Gruppen gebildet werden, ergibt sich bei jeder beliebigen Kombination immer ein Korrelationskoeffizient $r > 0.98$ und eine mittlere Abweichung unter 8%.

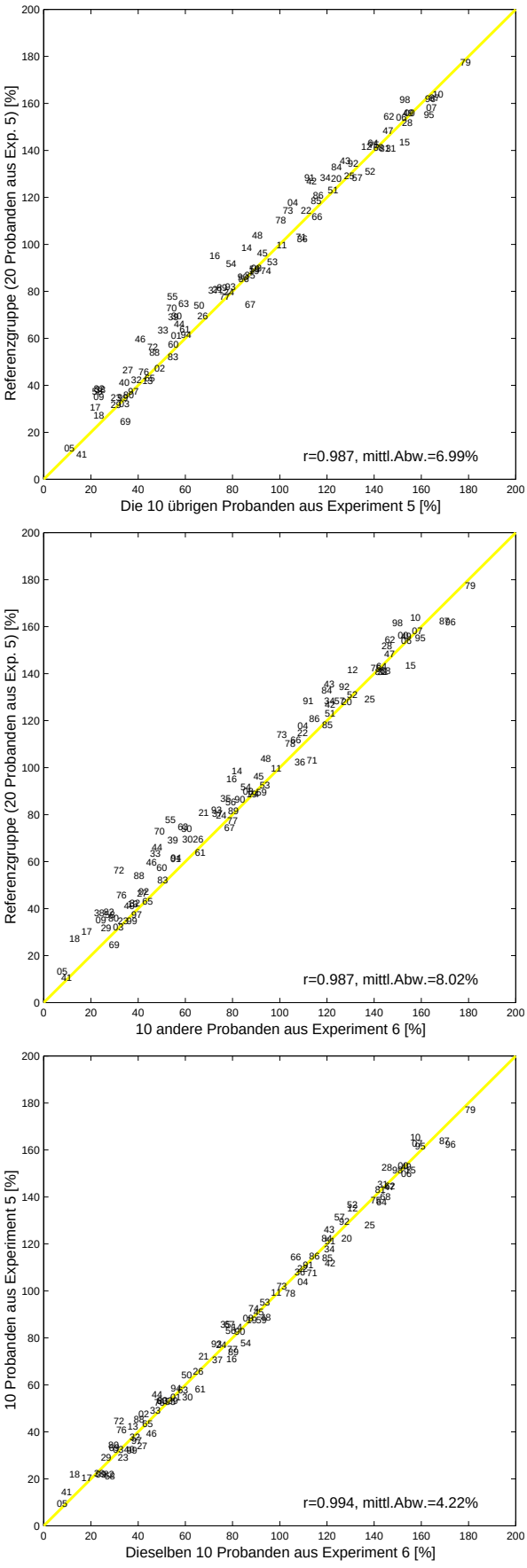

Abb. 7.13: Korrelationen zwischen Gruppen von Probanden aus Experiment 5 und 6.

selben Experiment zum annähernd selben Ergebnis kommen, dann muß man die Validität der Urteile akzeptieren.

Im mittleren Streuungsdiagramm von Abb. 7.13 werden die Gruppenmittelwerte von Experiment 6 mit der Gruppe der 20 Versuchspersonen aus Experiment 5 verglichen. Die Korrelation ist identisch, aber die mittlere Abweichung hat leicht abgenommen von etwa 7% im obigen Diagramm auf nunmehr 8%. Da keine der zehn Versuchspersonen seine Ergebnisse exakt wiederholen konnte, sind derartige Abweichung zu erwarten. Ein Lerneffekt hat sich aber offenbar nicht eingestellt, denn sonst hätten die neuen Ergebnisse besser sein müssen als die älteren.[12] Auch diese Ergebnisse unterstützen die Validität der Urteile.

Im unteren Diagramm von Abb. 7.13 werden schließlich die Gruppenmittelwerte derselben zehn Probanden aus Experiment 5 und 6 miteinander verglichen. Die extrem hohe Korrelation von $r = 0.994$ und die geringe mittlere Abweichung von 4.22% sprechen für sich: Die Reliabilität von Gruppen mit zehn und mehr Versuchspersonen ist ausgezeichnet.

Es stellt sich die Frage, ob die Unterschiede in den Ergebnissen im Abstand von sechs Monaten statistisch signifikant sind, ob also die Wiederholung des Experiments einen bedeutenden Effekt auf die Perzeptionsergebnisse hat. Hierzu wird eine zweifaktorielle Varianzanalyse mit Meßwiederholungen durchgeführt, die immer dann angezeigt ist, wenn es um die Erfassung von Veränderungen über die Zeit geht und wenn dieselben Versuchspersonen an den Wiederholungsexperimenten teilnehmen.

Als unabhängige Variablen wurden das Experiment und der Stimulus vorgegeben. Als Wiederholungen wurden die Perzeptionsergebnisse der zehn Probanden verwendet. Ohne die statistischen Ergebnisse hier im Detail auszubreiten, läßt sich aus der Analyse zusammenfassend festhalten, daß keiner der beiden Faktoren signifikant ist. Wie aufgrund des Gesamteindrucks anhand der Abb. 7.12 und 7.13 bereits vermutet werden konnte, findet sich kein Effekt aufgrund der Wiederholung des Experiments nach einem halben Jahr und damit eine letzte Unterstützung für die Reliabilität des Experiments zur Sprechgeschwindigkeitseinschätzung.

7.9 Zusammenfassung der Perzeptionsergebnisse

Die Reihe der durchgeführten Experimente hat uns in vier Schritten zum Ziel geführt: Im ersten Schritt wurde festgestellt, daß die Stimuli für Wahrnehmungsexperimente zur Sprechgeschwindigkeit länger als 400 ms und kürzer als 1400 ms sein sollten, um den Schwierigkeitsgrad bei der Beurteilung niedrig zu halten.

Im zweiten Schritt ergab sich, daß die Stimuli zugleich möglichst kurze Dauern aufweisen sollten, um die Wahrnehmbarkeit von Sprechgeschwindigkeitsvariationen zu verringern.

Im dritten Schritt wurde zum einen die Reliabilität von Sprechgeschwindigkeitsbeurteilungen nachgewiesen und zum anderen, daß Perzeptionsexperimente mit variabler Stimulusdauer aufgrund des gravierenden Einflusses auf die perzipierte Sprechgeschwindigkeit vorerst unterbleiben müssen.

Schließlich ergab sich im vierten Schritt, daß sich die hier gewonnenen Daten sehr gut als Repräsentation der perzipierten Sprechgeschwindigkeit eignen und somit als Vergleichsreferenz für akustische Untersuchungen zur Verfügung stehen.

[12] Hier und auch in Abb. 7.12 deutet sich schon an, daß drei Ankerstimuli möglicherweise zu wenig sind, so daß sich bei wiederholten Experimenten Verschiebungen und Rotationen der Daten ergeben können, wobei der Korrelationskoeffizient etwa gleich bleibt, aber sich die mittlere Abweichung verändert.

Das fünfte Experiment lieferte nach dem Muster des Experiments 4 zusätzliche Evaluationsdaten für die spätere Modellentwicklung. Und mit Hilfe der beiden Teile des sechsten Experiments konnte sowohl die völlig neue experimentelle Methode des interaktiven Perzeptionstests validiert werden, als auch die Reliabilität und Validität der Perzeptionsdaten gezeigt werden.

Damit ist die bisher fehlende Voraussetzung für die akustische Extensionalisierung der perzeptiv gegebenen Kategorie der lokalen Sprechgeschwindigkeit erfüllt.

Festzuhalten bleibt, daß Versuchspersonen zwar unterschiedliche Bewertungsstrategien bezüglich einiger Stimuli anwenden, aber daß Probanden mit hohen Korrelationswerten und niedrigen mittleren Abweichungen, wie schließlich Experiment 6 gezeigt hat, grundsätzlich den Gruppenmittelwert und mithin die Grundgesamtheit besser als andere Probanden repräsentieren, und nicht etwa in reliabler Weise stark abweichende Ergebnisse produzieren. Mit einer kleinen aber sorgfältig anhand von Perzeptionstests ausgewählten Gruppe von Versuchspersonen sollten also auch valide Perzeptionsdaten zu erreichen sein. Dieses Ergebnis kann in zukünftigen weiterführenden Untersuchungen berücksichtigt werden und den Aufwand gravierend verringern.

8

Relationen zwischen Akustik und Perzeption der Sprechgeschwindigkeit

8.1 Vorüberlegung

Unser Ziel ist die explizite Extensionalisierung der perzipierten lokalen Sprechgeschwindigkeit anhand ihrer meßbaren akustischen Korrelate. Die zu lösende Aufgabe besteht in der Entwicklung eines hinreichend genauen prädiktiven Modells.

Hierzu werden wir zunächst Korrelationen zwischen einzelnen akustischen Merkmalen und den Perzeptionsergebnissen analysieren. In einem zweiten Schritt vergleichen wir dann verschiedene Linearkombinationen von akustischen Merkmalen hinsichtlich ihres Korrelationskoeffizienten und ihrer mittleren Abweichung mit den Perzeptionsergebnissen.

Grundsätzlich gilt, daß die Genauigkeit von etwaigen Modellkoeffizienten unter den Ungenauigkeiten nicht nur der Perzeptionsergebnisse, sondern zusätzlich auch der akustischen Messungen zu leiden hat. Zwar besitzen die Perzeptionsergebnisse vergleichsweise hohe Stabilität, aber die akustischen Messungen sind stark abhängig von Segmentationsstrategie und -fehlern. Zudem führt ein Wechsel der Modelltopologie zu inkompatiblen Modellkoeffizienten. Um trotz all dieser Faktoren überhaupt Aussagen zur Stabilität der entwickelten Modelle machen zu können, sollen hier auch die Vertrauensbereiche der Modellkoeffizienten angegeben werden.

8.2 Untersuchung von Korrelationen

Die in Kap. 6.5 aus den 141 Trainingsstimuli extrahierten akustischen Merkmale der lokalen Silben- und Phonrate sowie der Grundfrequenz sollen nun auf ihre Beziehung zur perzipierten lokalen Sprechgeschwindigkeit überprüft werden. Wir verwenden hier die mittlere perzipierte lokale Sprechgeschwindigkeit, die für jeden Stimulus durch eine Zahl repräsentiert wird, nämlich den jeweiligen Mittelwert über alle 60 Probanden aus Experiment 4.[1] Die Mittelwerte sind zwar ohne plausiblen Nachweis vorerst nicht als Referenzwerte zu verstehen, aber sie stellen zumindest das Wahrnehmungsverhalten einer durchschnittlichen „künstlichen Versuchsperson“ dar.

Als Beispiel ist in Abb. 8.1 ein Streuungsdiagramm zwischen lokaler Silbenrate und mittlerer perzipierter lokaler Sprechgeschwindigkeit dargestellt. Offensichtlich sind hohe perzipierte Sprechgeschwindigkeiten vorwiegend von überdurchschnittlichen Silbenraten begleitet und niedrige Sprechgeschwindigkeiten von unterdurchschnittlichen Silbenraten. Nichtsdestotrotz spiegelt

[1] Siehe hierzu die Mittelwertmarkierungen der Stimuli in Abb. 7.10 oder deren Histogramme in Anhang A.3.

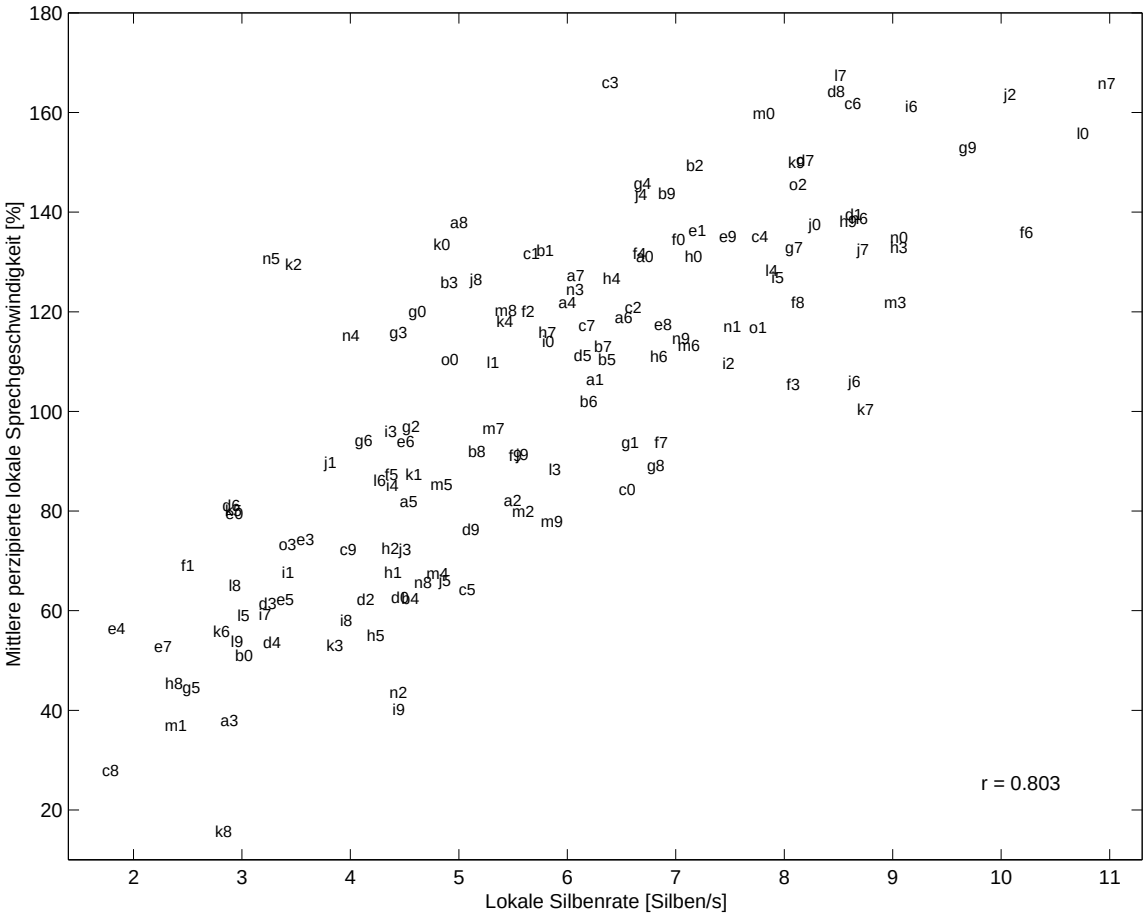

Abb. 8.1: Streuungsdiagramm zwischen gemessener lokaler Silbenrate der 141 Stimuli und der Sprechgeschwindigkeitsperzeption, die sich für jeden Stimulus durch Mitteln über die Resultate aller 60 Probanden von Experiment 4 ergibt.

der mäßige Korrelationskoeffizient $r = 0.803$ die nur bedingte Brauchbarkeit der Silbenrate als alleinigen Prädiktor für die perzipierte Sprechgeschwindigkeit wider. Übrigens korrelieren auch die individuellen Perzeptionsergebnisse der einzelnen Probanden selten besser mit der Silbenrate ($0.54 < r < 0.85$) als die der durchschnittlichen „künstlichen Versuchsperson“.

In Tab. 8.1 sind die Korrelationskoeffizienten aller akustischen Merkmale aus Kap. 6.5 mit der mittleren perzipierten lokalen Sprechgeschwindigkeit dargestellt. Für diese Darstellung wurden aus den akustischen Messungen zusätzliche Merkmale abgeleitet. So konnten die mittleren Silben- und Phonraten und deren Standardabweichungen und Steigungen ermittelt werden, indem nicht nur ein Wert pro Stimulus gemessen wurde, sondern drei: einer im Zentrum jedes Stimulus und je ein weiterer Wert im Abstand von 100 ms links und rechts des Zentrums.

Aus den in einer ersten Stufe mit einer Schrittweite von 5 ms extrahierten F_0-Rohmessungen wurden in einer zweiten Stufe ebenfalls Mittelwert, Standardabweichung und Steigung berechnet. Schließlich wurde grob die Indifferenzlage jedes Sprechers geschätzt, indem aus allen F_0-Messungen aller Stimuli jeweils eines Sprechers ein individueller Mittelwert berechnet wurde. Mit dessen Hilfe konnte für jeden Stimulus die Abweichung aus der approximierten Indifferenzlage ermittelt werden und damit ein grobes Maß für die Anspannung bei F_0.

Bei den in Tab. 8.1 dargestellten Ergebnissen ist die große Ähnlichkeit der Koeffizienten bei Silben- und Phonrate auffällig, die sich auch als statistisch nicht signifikant unterschiedlich her-

Akustische Messung	Korrelationskoeffizient r der akustischen Messung mit mittlerer perzipierter lokaler Sprechgeschwindigkeit
Silbenrate	0.803
Mittlere Silbenrate	0.812
Standardabweichung der Silbenrate	0.294
Steigung der Silbenrate	-0.115
Phonrate	0.836
Mittlere Phonrate	0.846
Standardabweichung der Phonrate	-0.183
Steigung der Phonrate	0.014
Mittlere F_0	0.253
Standardabweichung der F_0	0.041
Steigung der F_0	-0.130
Mittlere F_0 − mittlere F_0 des jeweiligen Sprechers	0.340

Tabelle 8.1: Verschiedene akustische Analysen (siehe Text) und die jeweiligen Korrelationskoeffizienten mit der perzipierten lokalen Sprechgeschwindigkeit.

ausstellen ($\hat{t}_{141} = 1.0669 < Z_{0.10} = 1.6449, n.s.$). Beide sind demnach gleichermaßen schlecht geeignet, die perzipierte Sprechgeschwindigkeit vorherzusagen. Auch F_0 erweist sich mit einem maximalen Koeffizient von $r = 0.34$, der bei dem akustischen Merkmal der Abweichung von F_0 aus der grob genäherten Indifferenzlage auftritt, als alleiniger Prädiktor gänzlich unbrauchbar.

8.3 Modelle zur Prädiktion der perzipierten Sprechgeschwindigkeit

Nachdem wir im vergangenen Abschnitt Korrelationen zwischen einzelnen akustischen Merkmalen und den Perzeptionsresultaten betrachtet haben und dabei letztendlich feststellen mußten, daß mit isolierten Merkmalen kein brauchbares Modell entstehen kann, drängt sich die Frage auf, ob und inwiefern *Kombinationen* von akustischen Merkmalen besser geeignet sind, die perzipierte lokale Sprechgeschwindigkeit vorherzusagen.

Mit Hilfe der multiplen linearen Regression läßt sich diese Frage klären. In der Praxis werden hierbei die akustischen Merkmale als unabhängige Variablen betrachtet und in einer Linearkombination durch Koeffizienten so gewichtet, daß die abhängige Variable — in unserem Fall die Perzeptionsresultate — mit einem möglichst geringen quadratischen Fehler vorhersagt wird.

Da eine Zunahme der Freiheitsgrade, die sich entweder durch Vergrößern der Anzahl von unabhängigen Variablen im Modell oder durch den Übergang zur Klasse der nichtlinearen Modelle ergibt, zwar immer eine Verbesserung der Vorhersagegenauigkeit während des Trainings bedeutet, aber auch eine Überadaption des Modells an die Trainingsdaten nach sich ziehen kann, das damit seine Generalisierungseigenschaften verliert, sollen *i)* nur lineare Modelle zum Einsatz kommen, *ii)* die Anzahl der Variablen möglichst klein gehalten werden und *iii)* eine mögliche Überadaption anhand eines Evaluationskorpus getestet werden.

In Pfitzinger 1998 [170] wendeten wir die multiple lineare Regression erstmals auf unsere Sprechgeschwindigkeitsdaten an. 1999 [171] verbesserten wir die Vorhersagegenauigkeit des Modells, indem wir einen dritten Koeffizient einführten. Dieses genauere Modell wird hier *Modell 1* genannt und noch einmal dargestellt, da sich aufgrund der gegenüber 1999 mehr als verdoppel-

ten Versuchspersonenzahl und aufgrund von Fehlerbereinigungen der akustischen Messungen die Modellkoeffizienten verändert und vor allem stabilisiert haben.

Das Modell 1 schätzt die perzipierte lokale Sprechgeschwindigkeit (PLSR) aus den akustischen Messungen durch

$$\widehat{\text{PLSR}} = s \cdot sr + p \cdot pr + c.$$

Das bedeutet, die lokale Silbenrate sr wird mit dem Koeffizient s multipliziert und die lokale Phonrate pr mit dem Koeffizient p. Zusätzlich gibt es noch einen weiteren Koeffizient c, der die Minimierung des Vorhersagefehlers durch eine Rotation der Streuung ermöglicht.

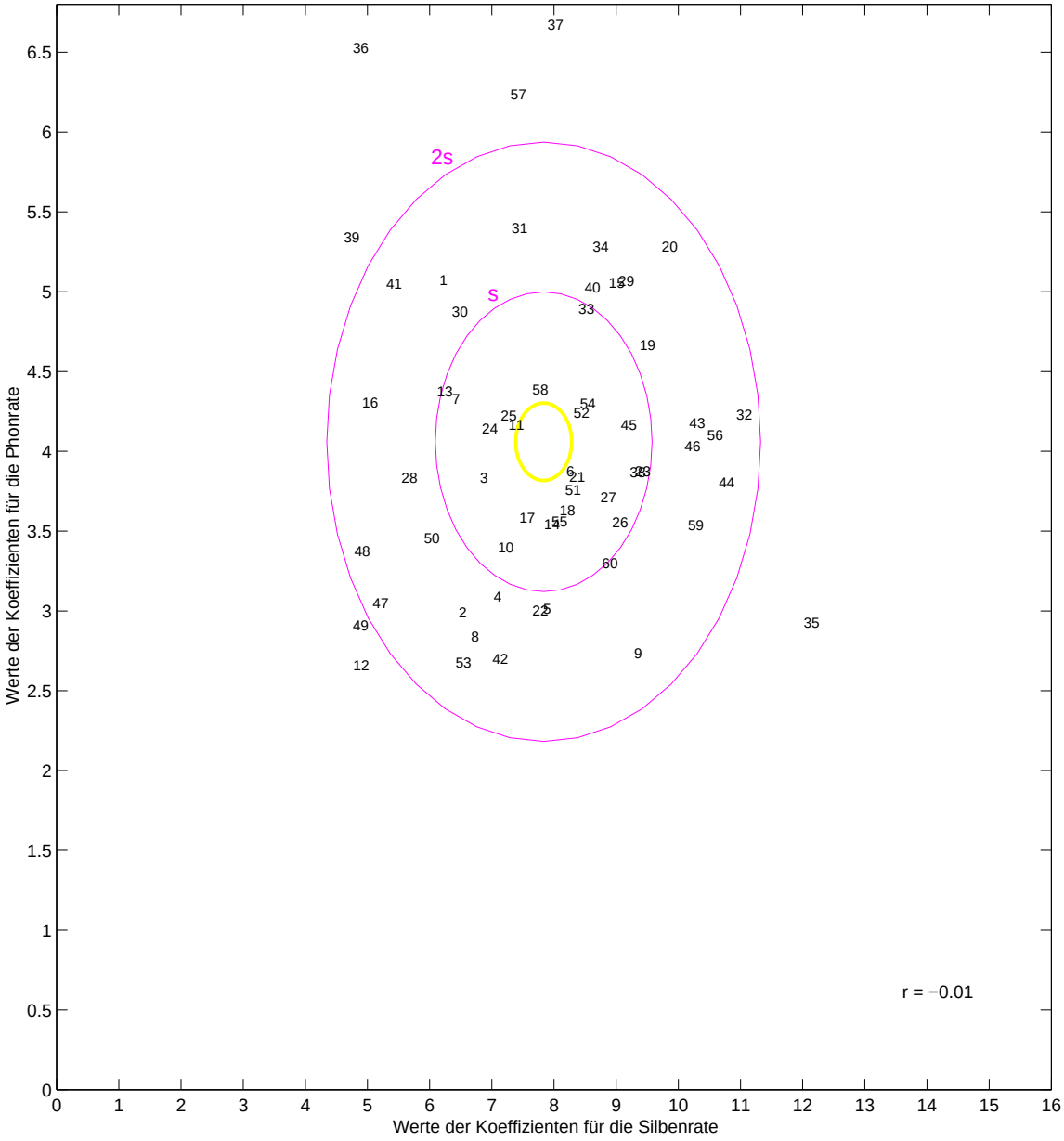

Abb. 8.2: Streuung der Koeffizienten s und p des Modells 1 für jeden der 60 Probanden. Die dunkel gedruckten Ellipsen zeigen die einfache und zweifache (s und 2s) Standardabweichung, die hell gedruckte Ellipse den 95%-Vertrauensbereich der Verteilung der Koeffizienten. X- und Y-Achsen wurden so skaliert, daß sie im Verhältnis 1:2.57, dem durchschnittlichen Verhältnis zwischen Silben- und Phonrate, stehen. Die größere y-Ausdehnung der Ellipsen spiegelt die größere Streuung des Phonratengewichts wider (vgl. Tab. 8.2), die Lage der Verteilung oberhalb der Winkelhalbierenden den größeren Einfluß der Phonrate.

Die unbekannten Koeffizienten s, p und c ergeben sich, indem die akustischen Messungen und die mittleren Perzeptionsergebnisse der 141 Stimuli zu dem folgenden überbestimmten linearen Gleichungssystem zusammengefaßt werden, und dieses nach der Methode der kleinsten Quadrate gelöst wird:

$$\begin{bmatrix} sr(1) & pr(1) & 1 \\ sr(2) & pr(2) & 1 \\ \vdots & \vdots & \vdots \\ sr(M) & pr(M) & 1 \end{bmatrix} \begin{bmatrix} s \\ p \\ c \end{bmatrix} = \begin{bmatrix} pz(1) \\ pz(2) \\ \vdots \\ pz(M) \end{bmatrix},$$

wobei M die Anzahl der Stimuli und pz die mittleren Perzeptionsresultate darstellen.

Für Abb. 8.2 wurden die Koeffizienten des Modells 1 bei jeder der 60 Versuchspersonen getrennt ermittelt. Man könnte nun vermuten, daß ein Proband, der sich bei der Beurteilung der Sprechgeschwindigkeit weniger an der Silbenrate orientiert, dafür mehr auf die Phonrate achtet und umgekehrt; doch tatsächlich sind beide Koeffizienten vollständig unabhängig, wie der Korrelationskoeffizient $r = -0.01$ offenbart.

An dieser Stelle nun können wir der *mittleren* perzipierten Sprechgeschwindigkeit im Gegensatz zum vorigen Abschnitt tatsächlich Referenzeigenschaft nachweisen, da das obige Gleichungssystem mit 141 Gleichungen äquivalent ist zu einem viel größeren Gleichungssystem mit $141 \cdot 60$ Gleichungen (eine für jedes einzelne Perzeptionsurteil jeder Versuchsperson). In diesem Fall repräsentiert die mittlere perzipierte Sprechgeschwindigkeit tatsächlich *alle* Perzeptionsergebnisse. Daher wird sie nun auch im folgenden als Referenz für die Bewertung der Residuen der Modelle herangezogen, die in den Abb. 8.3, 8.4 und 8.5 gezeigt sind.

Die Koeffizienten des Modells 1 und zweier Erweiterungen des Modells (*A* und *B*) sind in Tab. 8.2 mit den zugehörigen 95%-Vertrauensbereichen dargestellt. Zusätzlich wurden die Ver-

	Koeffizient für	Mittelwert	95%-Vertrauensbereich	
			absolut	relativ
Modell 1 (Abb. 8.3)	Silbenrate	7.83	±0.45	±5.8%
	Phonrate	4.06	±0.24	±6.0%
	Rotation	-1.41	±4.80	±4.8%
Modell A (Abb. 8.4)	Mittlere Silbenrate	8.21	±0.48	±5.9%
	Stdabw$_S$	2.23	±2.04	±91.4%
	Mittlere Phonrate	4.38	±0.27	±6.1%
	Stdabw$_P$	-0.20	±0.74	±372.1%
	Rotation	-8.66	±4.79	±4.8%
Modell B (Abb. 8.5)	Mittlere Silbenrate	7.72	±0.52	±6.7%
	Stdabw$_S$	3.18	±2.01	±63.2%
	Mittlere Phonrate	4.38	±0.27	±6.1%
	Stdabw$_P$	-1.38	±0.73	±52.4%
	Mittlere F_0	0.100	±0.019	±19.4%
	Stdabw$_{F_0}$	0.014	±0.034	±242.6%
	Rotation	-21.55	±5.05	±5.1%

Tabelle 8.2: Mittelwerte und 95%-Vertrauensbereiche der Koeffizienten der drei Sprechgeschwindigkeitsmodelle (ohne und mit Berücksichtigung von Standardabweichungen, Steigungen oder F_0) basierend auf den Perzeptionsergebnissen von 60 Versuchspersonen.

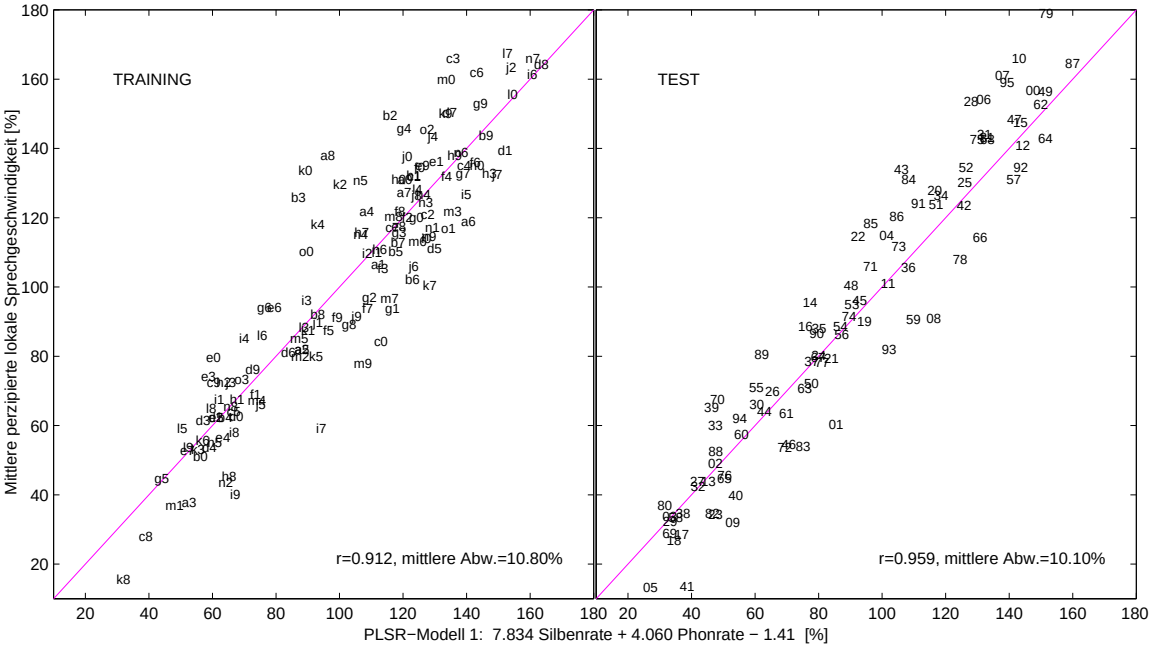

Abb. 8.3: Streuungsdiagramme zwischen der durch das einfache akustische Modell 1 geschätzten Sprechgeschwindigkeit und den Perzeptionsergebnissen.

trauensbereiche relativ zum Wertebereich des jeweiligen Koeffizienten angegeben. Man kann die 95%-Vertrauensbereiche auch so interpretieren, daß bei 95 von 100 zusätzlich durchgeführten gleichartigen Perzeptionsexperimenten die jedesmal neu ermittelten Koeffizienten um weniger als den in der vierten Spalte von Tab. 8.2 angegebenen Prozentsatz von den hier angegebenen Koeffizienten abweichen würden.

Während die meisten Vertrauensbereiche in Tab. 8.2 eine Größenordnung von ca. 6% des jeweiligen Koeffizienten besitzen, liegen zwei weit über 200%: 1) die Standardabweichung der mittleren Phonrate im Modell A und 2) die Standardabweichung der Grundfrequenz in Modell B. Derart große Vertrauensbereiche bedeuten, daß der Koeffizient keineswegs stabil ist. Ein Verzicht auf die zugehörigen Faktoren würde die Modelle kaum beeinträchtigen. Es ist aber zu bedenken, daß die Interaktionen zwischen den Faktoren unvorhersehbar sind: So ist die Standardabweichung der mittleren Phonrate, die im Modell A kaum Einfluß hat, im Modell B deutlich stabiler.

Die Abb. 8.3, 8.4 und 8.5 zeigen auf der linken Seite die Korrelation des jeweiligen Modells mit den Trainingsdaten und auf der rechten Seite mit den Evaluationsdaten. Hier ist ungewöhnlich, daß trotz sorgfältiger Stimulusauswahl und Wechsel der Sprecher sowie des Sprechstils von gelesener zur Spontansprache die Evaluationsdaten mit den Modellkoeffizienten, die allein aufgrund der Trainingsdaten gewonnen wurden, bei jedem Modell deutlich bessere Korrelationen und mittlere Abweichungen liefern als die Trainingsdaten. Vermutlich weisen die Trainingsstimuli eine größere Variation der Silbenstruktur auf und sind damit vielfältiger als die auf spontan produzierten Äußerungen basierenden Evaluationsstimuli. Wir hatten ja bereits in Kap. 6.5 darauf aufmerksam gemacht, daß schon die zugrundeliegenden Rohdaten beider Korpora unterschiedliche Korrelationskoeffizienten zwischen ihren lokalen Silben- und Phonraten aufweisen.[2]

Bei alleiniger Betrachtung der Trainingskorpora prädiziert Modell A die perzipierte Sprechgeschwindigkeit hoch signifikant besser als Modell 1 ($\hat{t}_{141} = 3.1853 > Z_{0.01} = 2.5758, \star\star$). Die Erweiterung von Modell A um die mittlere F_0 und deren Standardabweichung zum Modell B

[2] Vergleiche hierzu auch Abb. 1.1 auf S. 124 und Abb. 5.3 auf S. 170.

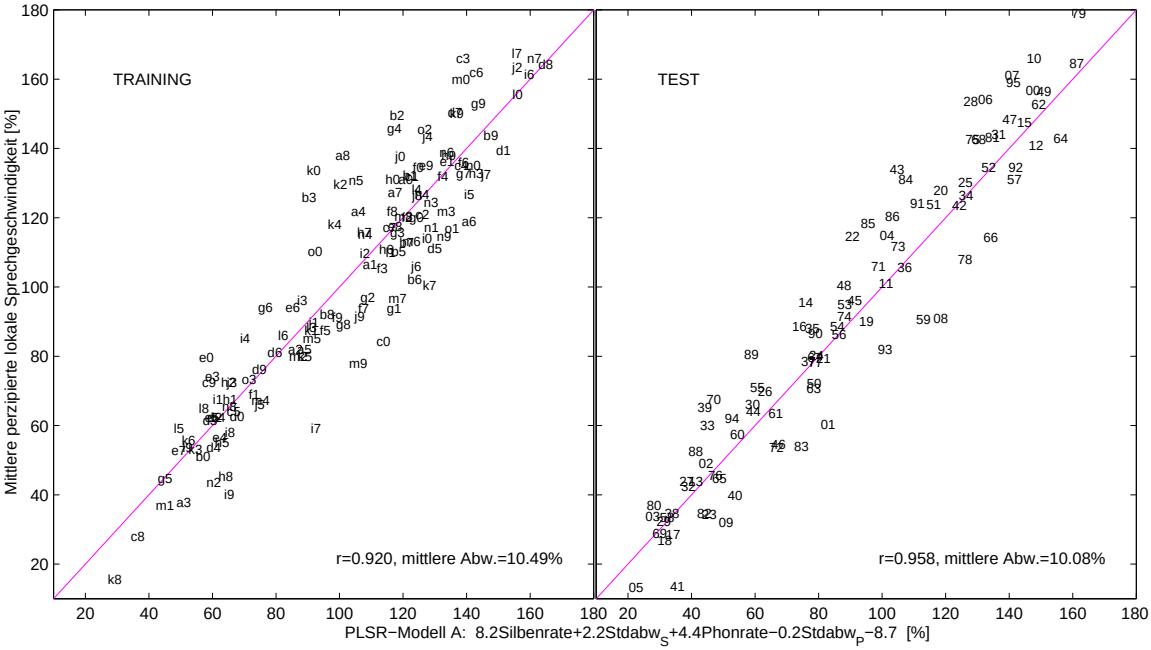

Abb. 8.4: Streuungsdiagramme zwischen der durch das akustische Modell A (mit Standardabweichungen und Steigungen, ohne F_0) geschätzten Sprechgeschwindigkeit und den Perzeptionsergebnissen.

zeigt eine zusätzliche signifikante Verbesserung ($\hat{t}_{141} = 2.1260 > Z_{0.05} = 1.96, \star$). Während in Pfitzinger 1999 [171] der Einfluß von F_0 aufgrund der geringeren Versuchspersonenzahl noch nicht signifikant war, scheint nun das Einbeziehen von F_0 zu einem etwas besseren Vorhersagemodell der perzipierten Sprechgeschwindigkeit zu führen.

Werden aber die Evaluationsergebnisse einbezogen, so finden sich zwischen den drei Modellen keine statistisch signifikant unterschiedlichen Korrelationskoeffizienten ($r_1 = 0.959$, $r_A = 0.958$, $r_B = 0.954$). Zwar werden die mittleren Abweichungen sowohl bei den Trainings-, wie auch bei den Evaluationsdaten mit zunehmender Modellkomplexität kleiner ($10.80 > 10.49 > 10.04$, $10.10 > 10.08 > 9.95$), aber eben nur in einer vernachlässigbaren Größenordnung.

Man muß wohl anhand dieser Ergebnisse zu dem Schluß kommen, daß das Einbeziehen weder von Standardabweichungen der Silben- und Phonraten, noch von F_0 eine höhere Genauigkeit liefert. Während die alleinige Betrachtung der Trainingsergebnisse ein anderes Fazit zuließe, offenbaren die Evaluationsergebnisse, daß die Signifikanz der Verbesserungen auf Überadaption an die Trainingsdaten zurückgeführt werden muß. Die Modelle sind demnach gleich gut.

8.4 PLSR — Perzipierte Lokale Sprechrate

Das gestellte Ziel ist erreicht. Nachdem es uns auf dem in Kap. 7 beschriebenen Weg gelungen ist, anhand des für BMBF-Projekte erstellten sehr umfangreichen Sprachmaterials der *PhonDatII*- und *VerbmobilI*-Korpora die lokale Sprechgeschwindigkeit als perzeptiv gegebene Kategorie zu etablieren, haben wir in diesem Kapitel anhand der akustisch meßbaren Korrelate Silbenrate, Phonrate und Grundfrequenz, deren Messungen in Kap. 6 beschrieben wurden, drei prädiktive Modelle (1, A und B) unterschiedlicher Komplexität entwickelt. Da diese dennoch nahezu identische Ergebnisse liefern, wie die Evaluationsergebnisse schließlich gezeigt haben, werden wir in zukünftigen Anwendungen das einfachste Modell präferieren: das Modell 1.

Die hier eingeführte perzeptive lokale Sprechrate — die PSLR — definiert ein prosodisches Maß, mit dessen Hilfe sich die gesuchten Konturverläufe der lokalen Sprechgeschwindigkeit aus den Sprachsignalen ohne weitere Perzeptionsexperimente herstellen lassen.

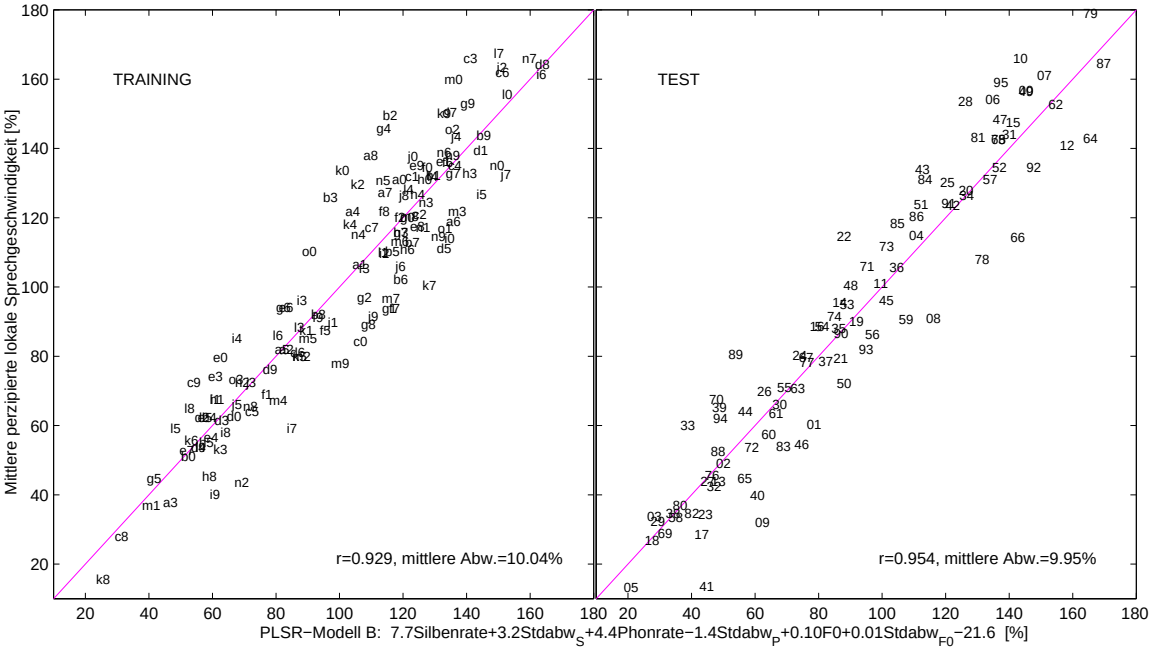

Abb. 8.5: Streuungsdiagramme zwischen der durch das akustische Modell B (mit Standardabweichungen, Steigungen und F_0) geschätzten Sprechgeschwindigkeit und den Perzeptionsergebnissen.

8.5 Zukünftige Weiterentwicklungen des Modells

Die im Hinblick auf umfangreiche Anwendungen ganz offensichtlich größte Schwachstelle, die alle hier eingeführten Modelle gemeinsam haben, ist die Notwendigkeit einer möglichst zuverlässigen Silbenkern- und Phonsegmentation des zu analysierenden Sprachsignals. Wir hatten aber schon in Kap. 6.5 auf S. 180ff betont, daß es nicht Ziel dieser Arbeit ist, alle denkbaren Möglichkeiten auszuschöpfen. Vielmehr stehen nun Perzeptionsdaten und Stimuli zur Verfügung, mit denen die Entwicklung eines neuen Modells möglich sein sollte, das keine Segmentationsdaten benötigt und damit anwendungsorientierter ist.

Aber auch hier gilt, daß die Anzahl der Modellkoeffizienten möglichst gering sein muß, um eine Überadaption zu vermeiden. Dies ist mit vollautomatischen Verfahren schwierig zu erreichen, da jeder beliebige 625 ms lange Signalabschnitt anhand von wenigen Parametern auf eine die Sprechgeschwindigkeit repräsentierende Zahl datenreduziert werden muß. So müßten etwa bei Verwendung von künstlichen neuronalen Netzen, die sich zunächst anbieten würden und z.B. auf 31 Frames mit jeweils 20 ms Dauer und 10 Spektralkoeffizienten zugreifen würden, bereits 310 Gewichte trainiert werden. Die Folge wäre ein unterbestimmtes System, das die Stimuli auswendig lernen und damit jede Generalisierungsfähigkeit verlieren würde.

Dennoch ist die sich abzeichnende Situation keineswegs hoffnungslos, denn mit Hilfe unserer bereits existierenden Modelle ließe sich ein großes Evaluationskorpus aus beliebigem handsegmentiertem Sprachmaterial berechnen, so daß alle 141 Trainings-, 96 Test- und 3 Ankerstimuli gemeinsam (insgesamt 240 Stimuli) in das Training des neuen Modells einfließen könnten.

Die Entwicklung dieses neuen Modells ist höchst wünschenswert; aber vor allem ist sie eine der wichtigsten Voraussetzungen, um in Zukunft das lokale Sprechgeschwindigkeitsverhalten anhand von großen Sprachkorpora untersuchen zu können.

9

Schlußbetrachtungen

Es ist nicht angestrebt und würde auch den Rahmen dieser Arbeit bei weitem sprengen, auf das umfangreiche vom Autor dieser Arbeit in den letzten Jahren entwickelte PHD-System[1] zur Entwicklung von automatischen Sprachsynthesemethoden einzugehen, das dem Konzept der *Synthese-durch-Analyse*[2] folgt und in dessen Entwicklungsumgebung auch die in der vorliegenden Untersuchung beschriebenen Verfahren zur Messung der Silben-, Phon- und Sprechgeschwindigkeit implementiert und verglichen wurden.

Die hier in der Arbeit neu eingeführte Meßmethode zur Ermittlung der perzipierten lokalen Sprechgeschwindigkeit bewährt sich im PHD-System bereits, und auch die in diesem Kapitel dargestellten Beispiele wurden mit Hilfe dieses Systems hergestellt. Da sie allerdings in der gedruckten Fassung der Arbeit nicht hörbar gemacht werden können, haben wir die zugehörigen Sprachsignale im WWW unter `http://www.phonetik.uni-muenchen.de/~hpt/sr` zum Abruf bereitgestellt.

Im folgenden sollen nun in knapper Form und anhand von ganz konkreten Sprachsignalen und deren Analysen gezeigt werden, wie die lokale Sprechgeschwindigkeit tatsächlich verläuft und wie sie für die zukünftige phonetische Forschung mit automatischen Methoden normalisiert und übertrieben werden kann, um Effekte in der zeitlichen Strukturierung gesprochener Sprache herauszuarbeiten. Schließlich folgt die Darstellung einer etwas komplexeren Anwendung, die dazu dient, das prosodische Verhalten eines Referenzsprechers auf beispielsweise das synthetisierte Signal eines automatischen Sprachsynthesesystems zu kopieren, ohne dabei die mikroprosodischen Strukturen des Zielsignals zu zerstören.

9.1 Zwei exemplifizierende Analysen

In den Abb. 9.1 und 9.2 sind die Analysen zweier exemplarischer Äußerungen dargestellt. In Abb. 9.1 wurde die Äußerung „*Eine Thermoskanne voll Kaffee ist im Rucksack verstaut*“ analysiert: Die perzipierte lokale Sprechgeschwindigkeitskontur zeigt ein deutlich ausgeprägtes Maximum, das während der Teiläußerung „*ist im*“ auftritt. Diese Wörter wurden mit etwa 130% der mittleren Sprechgeschwindigkeit[3] und damit vergleichsweise schnell ausgesprochen. Man kann

[1] *PHD* steht für *parametric high-definition* in Anlehnung an die von Campbell 1996 [20] für sein Syntheseverfahren verwendete Bezeichnung des *high-definition speech re-sequencing* Systems.

[2] Siehe hierzu Tillmann & Pfitzinger 2000 [211].

[3] Die mittlere Sprechgeschwindigkeit (=100%) wurde aus allen *PhonDatII*-Äußerungen der 16 Sprecher gewonnen. Sie ist nur etwa 10% schneller als der Mittelwert über die 12 Sprecher der spontansprachlichen Teststimuli. Also ist auch anzunehmen, daß sie den Mittelwert der Grundgesamtheit ebenfalls ausreichend repräsentiert.

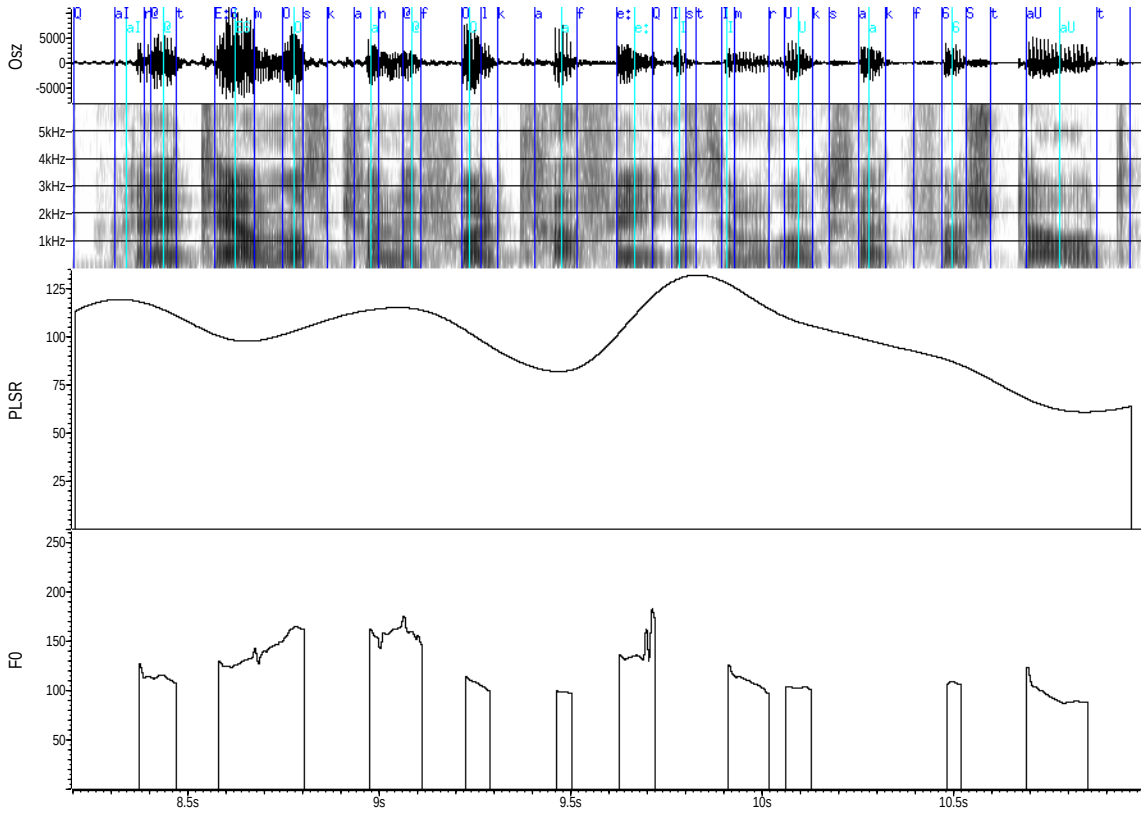

Abb. 9.1: *Oben:* Zeitsignal mit Sonagramm sowie manueller Phon- und Silbenkernsegmentation, *Mitte:* Verlauf der perzipierten lokalen Sprechrate, *unten:* Grundfrequenzverlauf.

diese Wörter als Funktionswörter klassifizieren, die laut der Hyper&Hypo-Theorie von Lindblom 1990 [126] sehr stark reduziert und somit auch schnell gesprochen werden können, wohingegen das unmittelbar vorausgehende Wort *Kaffee* als Träger des Satzfokus und damit eines ganz wesentlichen Teils der Information typischerweise deutlich und langsam ausgesprochen wird. Auch dies spiegelt sich in der Sprechgeschwindigkeitskurve wider und zwar im lokalen Minimum.

Schließlich wird anhand des Verlaufs der Sprechgeschwindigkeit die allmähliche Verlangsamung zum Satzende hin offensichtlich: Das aus der Literatur wohlbekannte *pre-final lengthening* (siehe Fußnote 22 auf S. 136) setzt in dieser Äußerung sehr früh ein und ist stark ausgeprägt. Dies wird vermutlich auch durch das Wort *Rucksack* mitverursacht, das wesentliche Informationen zum Verständnis des Satzes liefert und somit langsamer als die Funktionswörter produziert wird.

Der Grundfrequenzverlauf, auf den die heutige Forschung zur Prosodie ganz wesentlich — und teilweise sogar ausschließlich — zurückgreift, ist in Abb. 9.1 ebenfalls dargestellt. Beim Vergleichen der beiden prosodischen Konturen Sprechgeschwindigkeit und Intonation offenbart sich, daß sie nicht nur gänzlich unterschiedlich verlaufen, sondern auch, daß die sich in der Sprechgeschwindigkeitskontur widerspiegelnden linguistischen Eigenschaften der Äußerung nur schwer bis gar nicht im Grundfrequenzverlauf manifestiert sind. Dies wirft die Hypothese auf, daß ganz wesentliche linguistische Eigenschaften von Äußerungen eben nicht in der Intonation, sondern im lokalen Sprechgeschwindigkeitsverlauf kodiert und auch nur dort zu entdecken sind.

Im zweiten Beispiel in Abb. 9.2 ist die Äußerung „*Das ist der schnellste Weg zu den Kaligruben hinter den Hügeln*“ dargestellt. Auch hier werden die höchsten lokalen Sprechgeschwindigkeiten bei den Funktionswörtergruppen „*zu den*“ und „*hinter den*“ erreicht. Und auch *pre-final lengthening* läßt sich wieder beobachten, wenn auch nicht so ausgeprägt wie im vorigen Beispiel.

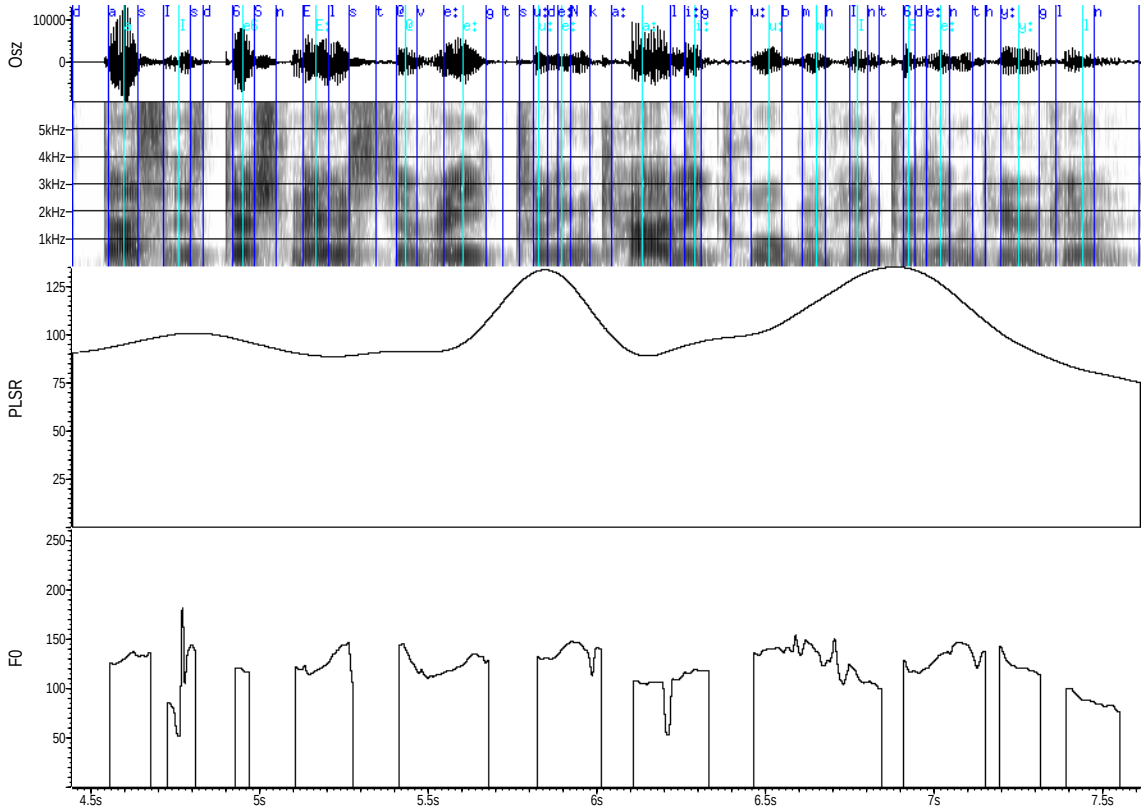

Abb. 9.2: *Oben:* Zeitsignal mit Sonagramm sowie manueller Phon- und Silbenkernsegmentation, *Mitte:* Verlauf der perzipierten lokalen Sprechrate, *unten:* Grundfrequenzverlauf.

Ob allerdings der Satzfokus auf dem Äußerungsteil „*schnellste Weg*“ oder „*Kaligruben*“ liegt, ist anhand unserer gerade erst erworbenen und recht unvollständigen Erfahrungen über den Verlauf von Sprechgeschwindigkeitskonturen ohne weiteres nicht eindeutig festzustellen. Hier tut sich ein ganz neues Forschungsfeld auf, das sicherlich in naher Zukunft umfassend bearbeitet wird.

9.2 Anwendungen in der phonetischen Forschung

Anhand der beiden Beispiele wurden die tatsächlich gemessenen Verläufe der perzipierten lokalen Sprechgeschwindigkeit demonstriert. Es ist aber durchaus möglich und auch sehr aufschlußreich, diese Konturen gezielt zu verändern und dann wieder auf das ursprüngliche Sprachsignal zu übertragen, so daß eine erneute Sprechgeschwindigkeitsanalyse des manipulierten Signals die gewünschte Kontur liefert. In den folgenden Abschnitten sollen hier insbesondere die Sprechgeschwindigkeitsnormalisierung und die Übertreibung dargestellt werden.

9.2.1 Normalisierung der Sprechgeschwindigkeit

Die gezielte Manipulation der gemessenen Sprechgeschwindigkeitskontur ist unproblematisch: So ist für die Normalisierung der Sprechgeschwindigkeit lediglich die Invertierung der Kontur an der 100%-Linie erforderlich. Dabei handelt es sich nicht um eine Spiegelung, sondern um eine Kehrwertbildung: Beispielsweise muß ein Signalabschnitt, der mit 130% der mittleren Sprechgeschwindigkeit wahrgenommen oder gemessen wurde, idealerweise auf $\frac{1}{130\%} = 76.9\%$ seiner lokalen Sprechgeschwindigkeit abgesenkt werden, um später mit der mittleren Sprechgeschwindigkeit von 100% wahrgenommen zu werden. Umgekehrt muß eine langsam gesprochene Passage

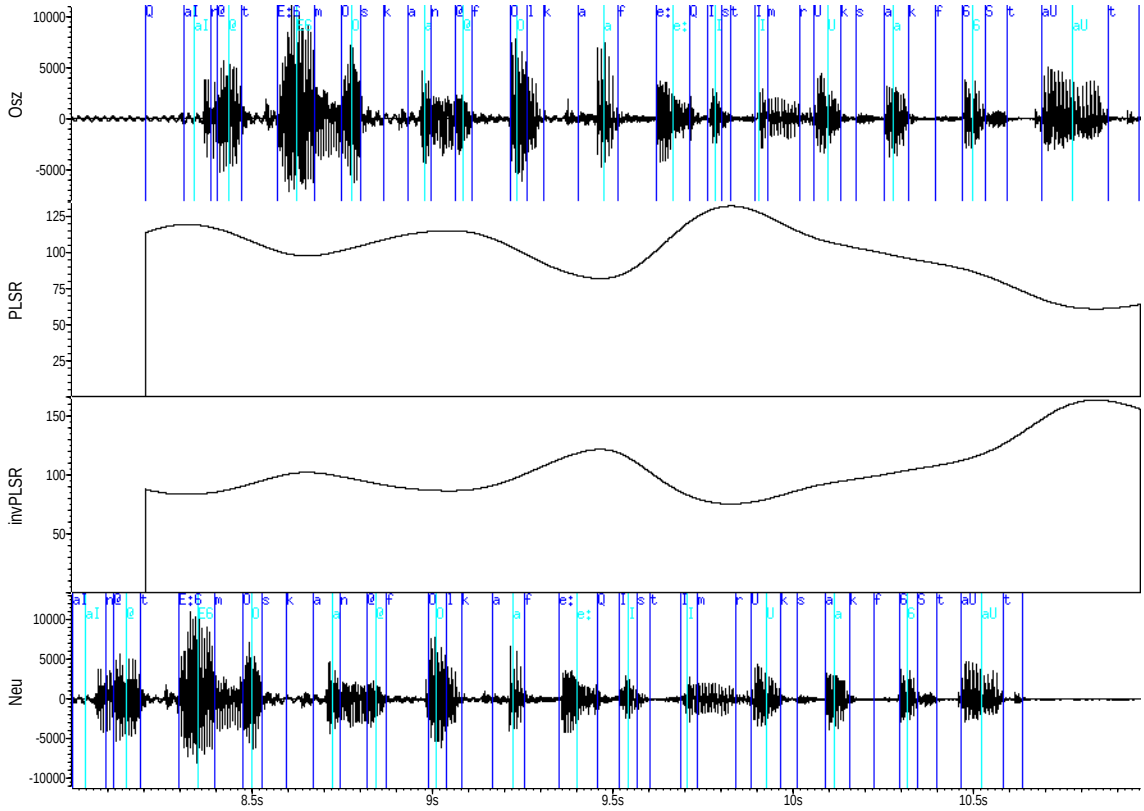

Abb. 9.3: *Oben:* Zeitsignal mit manueller Phon- und Silbenkernsegmentation, *darunter:* Verlauf der perzipierten lokalen Sprechrate, *darunter:* Verlauf der invertierten perzipierten lokalen Sprechrate, *unten:* sprechgeschwindigkeitsnormalisiertes Zeitsignal.

von z.B. 63% der mittleren Sprechgeschwindigkeit auf $\frac{1}{63\%} = 158.7\%$ beschleunigt werden, damit sie später ebenfalls mit 100% wahrgenommen wird.

Dies ist in Abb. 9.3 verdeutlicht: Wieder wurde die Äußerung „*Eine Thermoskanne voll Kaffee ist im Rucksack verstaut*“ verwendet. Der dritte Teil der Abbildung zeigt die invertierte Sprechgeschwindigkeitskontur. Das aufgrund dieser neuen Kontur resynthetisierte Sprachsignal ist im unteren Teil der Abbildung dargestellt. Die lokalen Veränderungen im zeitlichen Verlauf wurden mit Hilfe eines PSOLA-ähnlichen[4] Verfahrens im Rahmen des PHD-Systems durchgeführt.

Festzuhalten ist, daß die Manipulation der lokalen Sprechgeschwindigkeit die mikroprosodischen Segmentdauerverhältnisse, die durch Faktoren wie Kontext, Sprecher und Dialekt bestimmt werden, kaum beeinträchtigt. Dadurch verbessert sich die empfundene Natürlichkeit der manipulierten Äußerung auf das Niveau einer unmanipulierten Äußerung. Dennoch fällt die manipulierte Äußerung aus dem Rahmen wegen ihres (so die Aussagen mehrerer Versuchspersonen in einem informellen Akzeptanzexperiment:) „leiernden“ Sprechrhythmus und vor allem aufgrund des fehlenden *pre-final lengthening*. Dieses Fehlen ruft sogar den Eindruck einer finalen Sprechgeschwindigkeitserhöhung hervor, da Hörer vermutlich die sonst phrasenfinal übliche Dehnung kompensieren, was dann im Fall des Fehlens zu einem *perceptual overshoot*-Effekt führt.

9.2.2 Übertreibung der Sprechgeschwindigkeitsvariation

Schließlich soll noch in knapper Form auf den zur Normalisierung inversen Effekt eingegangen werden: Die Übertreibung oder auch Karikatur. Hierzu wird der extrahierte lokale Sprechge-

[4] PSOLA bedeutet *pitch synchronous overlap and add*. Siehe hierzu etwa Moulines & Charpentier 1990 [149].

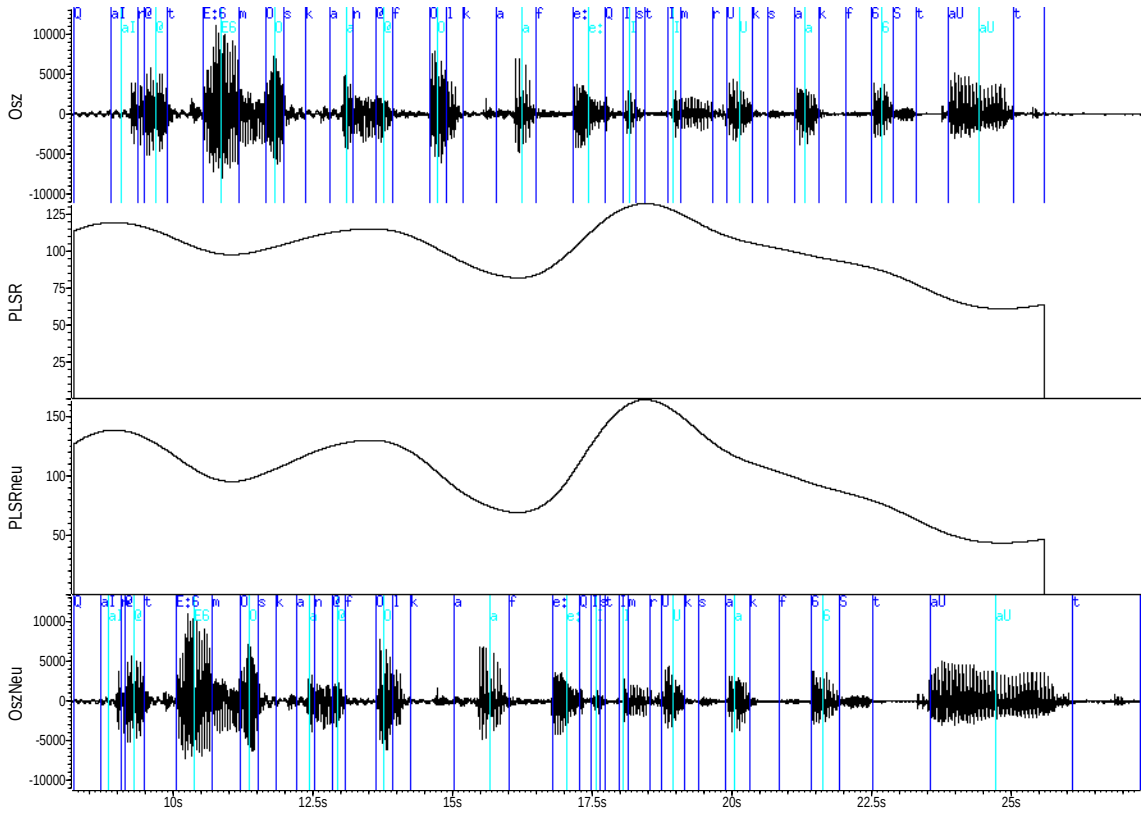

Abb. 9.4: *Oben:* Zeitsignal mit manueller Phon- und Silbenkernsegmentation, *darunter:* Verlauf der perzipierten lokalen Sprechrate, *darunter:* Verlauf der um 100% übertriebenen perzipierten lokalen Sprechrate, *unten:* Zeitsignal, das im lokalen Sprechgeschwindigkeitsverlauf übertrieben wurde.

schwindigkeitsverlauf mit einem gewünschten Faktor multipliziert, so daß in der resultierenden Kontur schnelle Passagen noch schneller und langsame Passagen entsprechend noch langsamer als die mittlere Sprechgeschwindigkeit verlaufen.

Ein Beispiel hierzu ist in Abb. 9.4 gezeigt. Wieder kam die Äußerung „*Eine Thermoskanne voll Kaffee ist im Rucksack verstaut*“ zum Einsatz. Wird die im dritten Teil der Abbildung in ihrer Modulationshöhe um 100% übertriebene Sprechgeschwindigkeitskontur wieder auf das Sprachsignal übertragen, so treten die lokalen Reduktions- und Dehnungsphänomene, die im ursprünglichen Signal wie beim üblichen Sprechen grundsätzlich eher unauffällig sind, sehr deutlich in Erscheinung. Insbesondere wird offensichtlich, wie sprecherspezifisch bestimmte Ausprägungen der lokalen Sprechgeschwindigkeit sind. Der Begriff der Karikatur ist hier durchaus angebracht, da alle Abweichungen von der mittleren Sprechgeschwindigkeit in ihrem Ausmaß übertrieben werden ähnlich einer gezeichneten Karikatur eines Gesichts, bei der auch alle Abweichungen vom Durchschnittsgesicht übertrieben dargestellt werden.

Am Rande sei angemerkt, daß die Ursache für die in Abb. 9.4 offensichtliche Zunahme der Gesamtdauer des neuen Signals gegenüber dem Ausgangssignal darin begründet liegt, daß der Sprecher die Äußerung im Mittel langsamer als 100% gesprochen hat. Dies muß in der Karikatur zwangsläufig zu einer weiteren Verlangsamung der Äußerung führen.

9.2.3 Kopieren prosodischer Sprechereigenschaften auf andere Sprecher

In Abb. 9.5 ist die synthetische Äußerung „*DBV-Präsident Sonnleitner wies die Kritik von Bundeskanzler Schröder am Kurs seines Verbands zurück*“ zusammen mit dem Sprachsignal eines

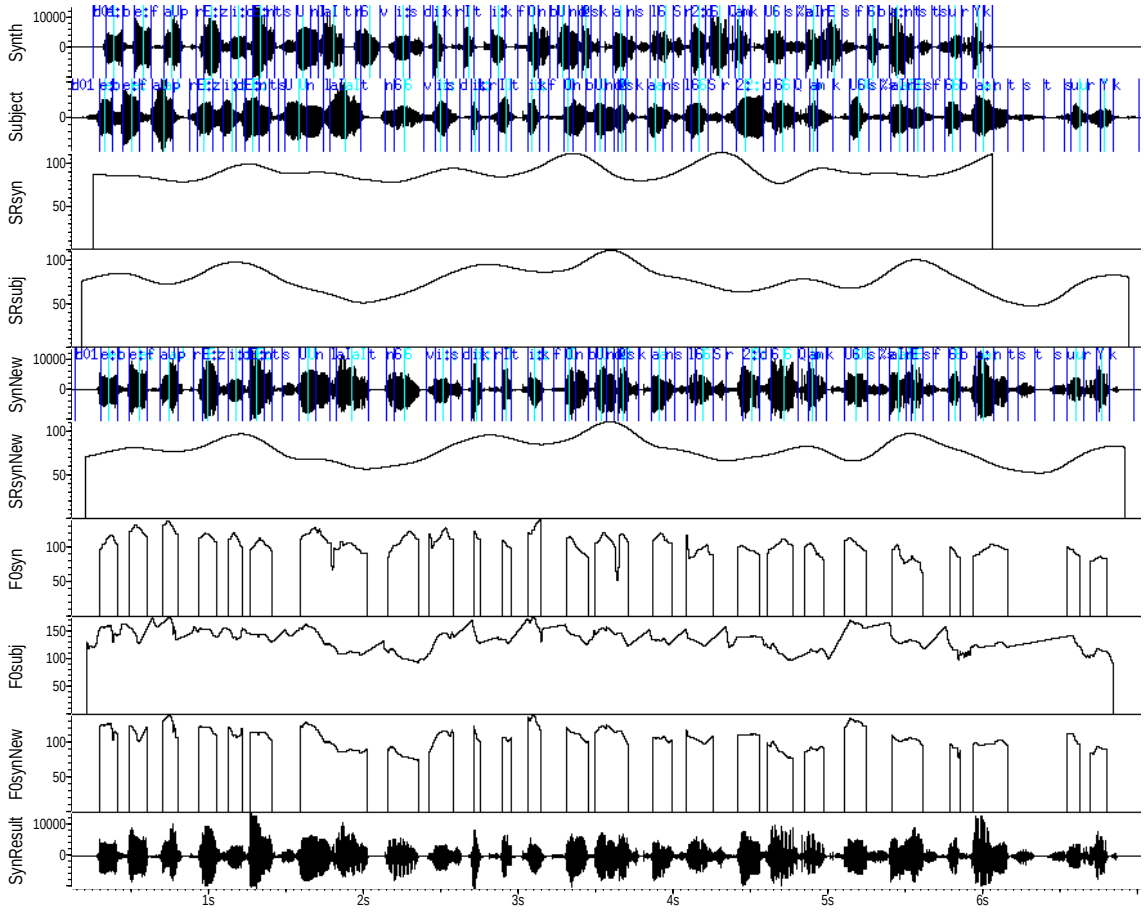

Abb. 9.5: Äußerung „*DBV-Präsident Sonnleitner wies die Kritik von Bundeskanzler Schröder am Kurs seines Verbands zurück*“. *Synth*: Synthetisches Signal. *Subject*: Von einem Probanden gesprochen. *SRsyn*: Sprechgeschwindigkeit der Synthese. *SRsubj*: Sprechgeschwindigkeit des Probanden. *SynNew*: Synthestisches Signal mit neuem Sprechgeschwindigkeitsverlauf des Probanden. *SRsynNew*: Zugehöriger im Nachhinein extrahierter Sprechgeschwindigkeitsverlauf. *F0syn*: Grundfrequenzverlauf des synthetischen Signals. *F0subj*: Extrapolierter Grundfrequenzverlauf des Probanden. *F0synNew*: Neuer Grundfrequenzverlauf des synthetischen Signals. *SynResult*: Synthetisches Signal mit der Prosodie des Probanden.

Probanden, der die gleiche Äußerung spricht, und verschiedenen Analysen und Manipulationen der beiden Ausgangssignale abgebildet. Für die hier dargestellten Arbeitsschritte ist unerheblich, welches Sprachsynthesesystem die Äußerung generiert hat. Wichtig ist lediglich, daß es deutliche Fehler im Timing und Grundfrequenzverlauf produzierte.

Bevor das Kopieren der Prosodie erklärt wird, sollte nicht unterschlagen werden, daß es uns mit Hilfe der im vorigen Abschnitt vorgestellten Methode der Übertreibung auch möglich gewesen wäre, den lokalen Sprechgeschwindigkeitsverlauf der synthetisch generierten Äußerung zu manipulieren. Beispielsweise werden durch Abhören einer um 100% in der Sprechgeschwindigkeitsmodulation überhöhten Variante des synthetischen Signals dessen Defekte im Timing auditiv hervorgehoben: So ist das satzfinal zu erwartende *pre-final lengthening* nicht auszumachen. Vielmehr tritt hier das Gegenteil ein: Zum Satzende hin steigt die lokale Sprechgeschwindigkeit des synthetischen Sprachsignals erheblich an. Dies zeigt auch das mit *SRsyn* bezeichnete dritte Teilbild von Abb. 9.5 sehr deutlich.

Desweiteren sollte der Name „*Schröder*“ in dieser Äußerung nicht mit maximaler Geschwindigkeit gesprochen werden, sondern mit deutlich unterdurchschnittlicher Geschwindigkeit, da er

stark betont realisiert werden sollte. Statt dessen könnte aber der Äußerungsteil „*wies die Kritik*“ deutlich schneller synthetisiert werden.

Durch diese Vorgehensweise ließen sich noch eine Reihe von echten Fehlern im Timing der Sprachsynthese aufdecken; wir wollen uns nun aber darauf konzentrieren, in mehreren Signalverarbeitungsschritten zuerst das Sprechgeschwindigkeitsverhalten der Versuchsperson auf das synthetische Sprachsignal zu kopieren und danach den Grundfrequenzverlauf, so daß sich ein synthetisches Sprachsignal mit dem Sprechgeschwindigkeits- und Grundfrequenzverlauf der Versuchsperson ergibt.

Bei den vorgenommenen Modifikationen sind viele kleine Zwischenschritte notwendig, deren Ergebnissignale aus Platzgründen nicht alle in Abb. 9.5 gezeigt werden können. Aber der detaillierte Ablauf des Verfahrens ist als Flußdiagramm in Abb. 9.6 dargestellt und zusätzlich als Unix-Shell-Programm in Anhang D ab S. 249 abgedruckt. Dort wird auch die große durch den Phonetiker kontrollierbare Parameteranzahl deutlich.

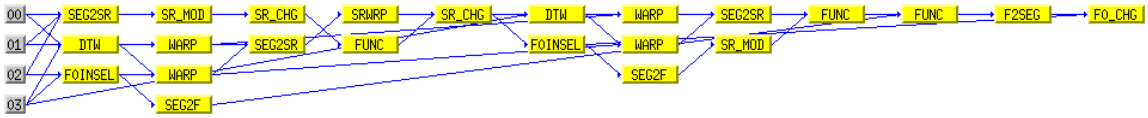

Abb. 9.6: Blockdiagramm des Algorithmus zur Erzeugung der in Abb. 9.5 dargestellten Signale. Links befinden sich die Eingangssignale: *00* = Synthesesignal, *01* = Phonsegmentation, *02* = Silbensegmentation und *03* = Probandensignal. Jeder Verbindungspfeil zwischen Modulen entspricht einem als Signal darstellbaren Zwischenergebnis (siehe hierzu auch Anhang D ab S. 249).

Die Abweichungen zwischen den beiden Sprechgeschwindigkeitsverläufen des Probanden und der manipulierten Synthese (vergleiche in Abb. 9.5 die Teilbilder *SRsubj* vs. *SRsynNew*) erklären sich durch den Einfluß der unterschiedlichen intrinsischen Segmentdauern beider Quellen, die nicht durch die zum Einsatz gekommenen Meß- und Manipulationsmethoden geändert werden können und sollen.

Trotz der vielen Verarbeitungsschritte bleiben — perzeptiv betrachtet — das Reduktionsverhalten, die Stimmqualität und die Sprechercharakteristik der synthetischen Stimme weitgehend erhalten; nur die makroprosodischen Merkmale der lokalen Sprechgeschwindigkeit und der Intonation werden einschneidend verändert.

Diese aufwendige Prozedur erlaubt nun auch tiefergehende Einblicke in die zugrundeliegende Synthesestruktur auf segmentaler Ebene und erste darauf basierende Verbesserungsvorschläge: so wird deutlich, daß die Lautdauern, die durch den Auswahlprozeß der konkatenativen Sprachsynthese zur Anwendung gekommen sind und durch unsere Sprechgeschwindigkeitsmanipulation nicht unabhängig von unmittelbar benachbarten Lauten verändert wurden, schon in ihrer ursprünglichen Form zu kurz (z.B. *Schröder*) oder zu lang (z.B. *Kritik*) waren. Man muß wohl festhalten und bei zukünftigen Weiterentwicklungen der Sprachsynthese berücksichtigen, daß die intrinsischen Dauern von Sprachlauten als Auswahlkriterium bei der Synthese ganz offensichtlich einen erheblichen Einfluß auf die spätere Synthesequalität ausüben.

In konkatenativen Sprachsynthesesystemen ergibt sich der lokale Verlauf der Sprechgeschwindigkeit bisher indirekt und als Folge aus den intrinsischen Dauern der zugrundeliegenden Sprachsegmente, die als Bausteine für eine konkrete zu realisierende Äußerung automatisch ausgewählt werden. Nicht zuletzt weil die lokale Sprechgeschwindigkeit sich bisher einer Messung entzog, waren die perzeptiven Auswirkungen gezielter Manipulationen, z.B. der Längung eines Wortes im Satzfokus, unvorhersehbar bzw. nur mit aufwendigen perzeptiven Experimenten abschätzbar.

Auf dem heutigen Stand der Forschung ist es trotz der Kenntnis von der lokalen Sprechgeschwindigkeit schwierig vorherzusagen, welche Passagen einer Äußerung schneller und welche langsamer zu produzieren sind. Zu erwarten ist, daß ähnliche Verfahren[5] wie diejenigen zur Vorhersage des Grundfrequenzverlaufs einer zu generierenden Äußerung auch genutzt werden können, um den lokalen Sprechgeschwindigkeitsverlauf vorherzusagen. Doch da erste bisher noch unveröffentlichte Studien zeigen, daß die lokale Sprechgeschwindigkeit nicht unmittelbar aus dem Grundfrequenzverlauf abgeleitet werden kann,[6] müssen an dieser Stelle erst noch weitere Ergebnisse aus der Grundlagenforschung abgewartet werden.

Zuletzt soll noch eine weiter auszuarbeitende Grundidee angedeutet werden: In dem hier beschriebenen Vorgehen könnte eine neue Methode der Evaluation von Sprachsynthese begründet werden, die darin bestünde, jeweils eine generierte Eigenschaft des Synthesesignals durch korrespondierendes menschliches Verhalten zu ersetzen, um dann in einem perzeptiven Vergleich den Grad der damit erzielten Veränderung in der Synthesequalität zu ermitteln.

9.3 Zusammenfassung der Ergebnisse

Unsere Untersuchungen liefern eine Reihe bemerkenswerter Ergebnisse: So ist zunächst festzustellen, daß die hier entwickelten linearen Modelle zur Prädiktion der perzipierten lokalen Sprechgeschwindigkeit aufgrund ihrer geringen mittleren Abweichung von 10.8% bei Lesesprache und 10.1% bei Spontansprache für die phonetische Forschung hinreichend genau sind.

Die Modelle besagen, daß perzipierte lokale Sprechgeschwindigkeit sowohl im gelesenen wie auch im spontan gesprochenen Deutsch hauptsächlich eine Kombination aus Silben- und Phonrate ist, während die anderen untersuchten Parameter wie z.B. Standardabweichungen und Steigungen von Silbenrate, Phonrate und F_0 keinen Einfluß haben. Insbesondere führt die Integration der Grundfrequenz nur zu einer scheinbaren, aber statistisch nicht signifikanten Verbesserung der Vorhersage. Aufgrund der hier ebenfalls durchgeführten Evaluation kann F_0 nicht zu den akustischen Merkmalen der perzipierten lokalen Sprechgeschwindigkeit hinzugezählt werden.

Viele weitere möglicherweise erfolgversprechende akustische Parameter blieben in dieser Arbeit bewußt unberücksichtigt, da hier in der Hauptsache gezeigt werden sollte, daß die lokale Sprechgeschwindigkeit als Prosodie meßbar und aussagekräftig ist und daher Einzug in die zukünftige Forschung zur Prosodie der Sprache halten sollte. Ziel war also nicht der Versuch der Berücksichtigung aller denkbaren Parameter und einer möglichst perfekten Vorhersage der perzipierten Sprechgeschwindigkeit. Der Einfluß weiterer akustischer Parameter auf die Wahrnehmung der Sprechgeschwindigkeit kann aber in Zukunft mit Hilfe der hier entwickelten Datenbasis und Methodik geprüft werden.

Hinter der Summenbildung in den verwendeten linearen Modellen zur Vorhersage der perzipierten Sprechgeschwindigkeit verbirgt sich das in der Perzeptiven Phonetik unter der Bezeichnung *trading relations* bekannte Paradigma, das sich hier wie folgt darstellt: Eine höhere Silbenrate kann durch eine gleichzeitig auftretende niedrigere Phonrate perzeptiv kompensiert werden, und auch umgekehrt.

[5] Diese Verfahren basieren auf z.B. künstlichen neuronalen Netzen, Hidden Markov Modellen oder auch syntaktisch-semantischen Analysen und Entscheidungsbäumen.

[6] Die Hypothese, daß eine steigende Grundfrequenz zur Markierung einer Betonung begleitet wird von fallender lokaler Sprechgeschwindigkeit, weil Betonung auch über längere Segment- und Silbendauern markiert wird, muß voraussichtlich verworfen werden.

Interessant ist nun, daß sich dieses perzeptive Verhalten zumindest im Fall von Silben- und Phonrate mit dem artikulatorischen Verhalten deckt: Einem der „phonetischen Quantitätsgesetze“ von Menzerath & de Oleza S. J. [136, S.91] folgend, das hier auf S. 130 zitiert wurde, hat ein Laut einer komplexeren Silbe im Durchschnitt eine kürzere Dauer als der einer einfacheren Silbe. Komplexe Silben sind aber gleichzeitig im Durchschnitt länger als einfache Silben. Aus diesen beiden Feststellungen folgt ganz unmittelbar, daß bei komplexen Silben die Silbenrate abnimmt, während die Phonrate zunimmt. Dagegen nimmt bei einfachen Silben die Silbenrate zu, während die Phonrate abnimmt. Dieser reziproke Zusammenhang kann als zwingendes zeitliches Strukturierungsprinzip aufgefaßt werden, dem die perzeptive Kompensation Rechnung trägt.

Den in dieser Arbeit vorgeschlagenen akustischen Modellen zur Vorhersage der perzipierten Sprechgeschwindigkeit liegen Netto- und keine Bruttoraten zugrunde, da Nettoraten aus dem Signal mit vergleichsweise einfachen Verfahren und einem akzeptablen Fehler extrahierbar sind, während Bruttoraten den vollständigen Spracherkennungsprozeß samt Graphem-Phonem-Konvertierung erfordern, um die kanonischen Formen zu erhalten, deren Geschwindigkeit erst dann gemessen werden kann. Zudem spiegeln kanonische Formen die artikulatorischen Strategien bei der Produktion schneller Sprache nicht wider, die eine ausreichende Verständlichkeit bei möglichst geringem Aufwand garantieren: Phoneme und sogar Silben werden assimiliert und auch elidiert, so daß die Nettophonrate deutlich unter der Bruttophonrate liegt. Welche Strategien in welcher Weise vom Sprecher angewendet werden, um höhere Sprechgeschwindigkeiten zu erreichen, ist Gegenstand der aktuellen Forschung (z.B. Kröger 1996 [113]) und noch nicht so gut verstanden, daß die tatsächlichen Realisierungen aus den kanonischen Formen vorhergesagt werden könnten. Die Interaktionen von Sprechgeschwindigkeit und Reduktionsgrad gehören nach wie vor zu den Herausforderungen in der phonetischen Forschung.

In unseren Untersuchungen wurde eine neue psychophysikalische Relation aufgedeckt und analysiert, die wir mit *PLSR* (*perceived local speech rate* oder *perzipierte lokale Sprechgeschwindigkeit*) bezeichnet haben. Sie tritt an die Stelle der bisherigen Maße für Sprechgeschwindigkeit, denn sie spiegelt wider, was vom Hörer als hohe oder niedrige Sprechgeschwindigkeit empfunden und bezeichnet wird. Beim jetzigen Stand der Forschung kann sie allerdings noch nicht als allgemeingültig betrachtet werden, sondern sie bezieht sich nur auf gelesenes und spontan gesprochenes Deutsch.

Mit dieser vorläufigen Einschränkung erlauben die hier entwickelten Modelle die automatische Extraktion einer bisher nicht meßbaren Prosodie aus Sprachsignalen, nämlich der Prosodie der Sprechgeschwindigkeit. Offenbar ist weder die Silben-, noch die Phonrate allein eine akzeptable Näherung für die PLSR. Darum sollten sie in Zukunft auch nicht mehr als Maß für die *Sprechgeschwindigkeit* herangezogen werden.

Zuletzt muß festgehalten werden, daß das Paradigma des *perceptual overshoot* auch für den Bereich der Sprechgeschwindigkeitswahrnehmung von Sprachsignalen mit Dauern unter 625 ms bestätigt werden kann: Die Sprechgeschwindigkeit wird von den Versuchspersonen umso höher eingeschätzt, je kürzer die zu beurteilende Signaldauer ist.

9.4 Ausblick

Bei den in Kap. 6 entwickelten Modellen tritt eine unerklärte Varianz auf, die darauf hindeutet, daß es noch akustische Merkmale gibt, die bisher nicht berücksichtigt wurden. Silben- und Phonrate allein können die unterschiedlichen Wort- und Silbenstrukturen des Deutschen nicht vollständig

repräsentieren. Für adäquatere Modelle sollten möglicherweise noch Morph-, Wort- oder auch Fußrate hinzugezogen und ergänzende akustische Merkmale (z.B. Stimmqualität, Lautstärke, ...) gesucht werden. Die Frage, ob Prominenzrelationen — eventuell in Verbindung mit rhythmischen Mustern — auf der Fußebene Einfluß auf die Sprechgeschwindigkeitswahrnehmung haben, muß an dieser Stelle beim jetzigen Stand der Forschung leider unbeantwortet bleiben.

Desweiteren sind Experimente mit anderen Sprechstilen wie z.B. mit skandierendem Versprechen, aber auch mit anderen Sprachen denkbar und naheliegend, wobei insbesondere weitere akzentzählende Sprachen (z.B. Englisch) untersucht werden sollen, dann aber auch silbenzählende (z.B. Französisch oder Spanisch) und morenzählende Sprachen (z.B. Japanisch).

Der Versuch eines instrumentalphonetischen Nachweises von Isochronie (Lehiste 1977 [122], Hoequist 1983 [84], Couper-Kuhlen 1993 [25]) in der Akustik einer Sprache scheitert unserer Ansicht nach nicht zuletzt daran, daß durch permanente Variationen der lokalen Sprechgeschwindigkeit immer verschiedene Segmentdauern hervorgerufen werden.[7] Wenn es mit Hilfe unserer Modelle möglich ist, die Sprechgeschwindigkeit vollständig zu normalisieren, ließe sich testen, ob dadurch mehr Hinweise auf Isochronie im Sprechrhythmus in Erscheinung treten. Dann könnte auch die Untersuchung der zeitlichen Organisation rhythmisch unterschiedlicher Sprachen zu neuen Erkenntnissen führen.[8]

9.5 Abschließende Bemerkung zur Frage von Diskontinuitäten

Aufgrund unserer im Laufe dieser Untersuchung gesammelten Erfahrungen sind wir mittlerweile fest davon überzeugt, daß bei lautsprachlichen Äußerungen immer wieder Passagen auftreten, in denen sich die Sprechgeschwindigkeit abrupt und diskontinuierlich ändert. Allerdings ist unser Verfahren ohne weitere Veränderungen systembedingt nicht in der Lage, diese zu detektieren, falls sie nicht in Verbindung mit Sprechpausen auftreten. Mit Hilfe des vergleichsweise großen Analysefensters von 625 ms werden zwar mikroprosodische Segmentdauervariationen von makroprosodischen Sprechgeschwindigkeitsvariationen getrennt, aber es wird eben zwingend auch immer ein kontinuierlich geglätteter Verlauf produziert. Solange die Zeitpunkte der Diskontinuitäten im Sprachsignal nicht mit automatischen Methoden detektiert werden können, muß der resultierende Fehler in Kauf genommen werden.

Wir vermuten zwar, daß diejenigen Diskontinuitäten, die nicht von einer kurzen Sprechpause begleitet werden, sehr selten sind, so daß die Anzahl dieser Fehler vergleichsweise gering bleibt. Aber diese spezielle Art von Diskontinuitäten könnte, falls sie existieren sollte, eine besondere kommunikative Funktion haben, die beim gegenwärtigen Stand unserer Methode leider verdeckt bleiben muß. Erst durch kommende Weiterentwicklungen des hier vorgestellten Meßverfahrens könnten die vermuteten Diskontinuitäten der experimentellen Untersuchung zugänglich gemacht werden.

Allerdings wird in den meisten Fällen bereits das jetzige Verfahren ausreichen, um zu neuen Erkenntnissen über die prosodische Strukturierung gesprochener Sprache zu gelangen.

---

[7] So, wie auch ein ständiger Wechsel zwischen *Rallentandi/Ritardandi* und *Accelerandi* in der Musik zu verschieden langen Taktdauern und damit zu einem verzerrten Rhythmus führt.

[8] Siehe hierzu Hoequist 1983 [86], Dimitrova 1998 [40], Cummins & Port 1998 [31].

ANHANG

A

Perzeptionsexperiment 4

A.1 Instruktion der Versuchspersonen

Beim Perzeptionsexperiment zur Einschätzung der Sprechgeschwindigkeit (Kap. 7.6) erhielten die Versuchspersonen folgenden Instruktionstext, den sie jederzeit während des Experiments im Menü abrufen konnten (siehe hierzu Abb. 7.6 auf S. 188):

> Im oberen Bereich des Fensters sehen Sie 141 kleine graue Knöpfe. Ihre Aufgabe besteht darin, diese Knöpfe in den unteren Bereich des Fensters entsprechend der subjektiv empfundenen Sprechgeschwindigkeit zu verschieben. Durch Anklicken eines Knopfes mit der linken Maustaste wird das entsprechende Sprachstück abgespielt. Um einen Knopf zu bewegen, muß die linke Maustaste gedrückt gehalten und die Maus dabei bewegt werden.
>
> Drei gelbe Knöpfe über den drei senkrechten Linien helfen beim Plazieren der einzuordnenden Knöpfe, indem sie als Anhaltspunkte und Vergleichsmöglichkeiten dienen und ein langsames, ein normales und ein schnelles Sprachstück abspielen. Diese drei Knöpfe spannen also eine Sprechgeschwindigkeitsskala auf, die von links nach rechts ansteigt. Je langsamer gesprochen wurde, umso weiter links sollte der einzuordnende Knopf plaziert werden, und je schneller gesprochen wurde, umso weiter rechts.
>
> Hilfreich beim Einschätzen der Sprechgeschwindigkeiten kann es sein, das Gehörte innerlich nachzusprechen und den eigenen Aufwand dabei zu beurteilen. Es kann auch helfen, sich beim Vergleichen zweier Sprachstücke das eine vom anderen der beiden Sprecher gesprochen vorzustellen.
>
> Anfangs sollten Sie immer mit den drei gelben Knöpfen vergleichen und vor allem die Geschwindigkeitsabstände zu diesen Knöpfen beurteilen: klingt z.B. das zu bewertende Sprachstück viel schneller als der zweite gelbe Knopf aber nur wenig langsamer als der dritte, so sollte der einzuordnende Knopf näher an der dritten senkrechten Linie plaziert werden als an der zweiten.
>
> Je mehr Knöpfe nun im unteren Teil des Fensters positioniert wurden, desto mehr Nachbarn entstehen, mit denen Sie dann auch vergleichen sollten. Gerade zu Beginn werden Sie sich immer wieder korrigieren und Positionierungen verfeinern. Deswegen ist es in dieser Phase wichtig, immer wieder die drei gelben Knöpfe heranzuziehen und den unteren Teil des Fensters gleichmäßig zu verdichten: nehmen wir also an, Sie würden durch Zufall immer wieder fast gleich schnelle Sprachstücke erwischen, dann sollten Sie diese zurückstellen und Sprachstücke aus anderen Geschwindigkeitsbereichen vorziehen, die bisher noch nicht besetzt wurden, um die Skala gleichmäßig aufzufüllen.
>
> Hier noch ein paar kurze Hinweise:
>
> - Es handelt sich um 141 Sprachstücke von verschiedenen Sprechern.

```
- Sie sollten zwischen 45 und 65 Minuten für das Experiment aufwenden.

- Manche Sprachstücke sind schlecht verständlich und trotzdem einschätzbar.

- Bitte nicht über den Kontext von Sprachstücken nachdenken oder diesen gar
  in die Bewertung einfließen lassen.

- Bewerten Sie nicht durch Zählen von Worten, Morphen, Silben, Lauten oder
  ähnlichem oder durch Einschätzen der Sprechtonhöhe; die Sprachstücke sind
  auf diesen Ebenen nicht vergleichbar.

- Es gibt sowohl langsamere Sprachstücke als den linken gelben Knopf als
  auch schnellere als den rechten.

- Die Beschriftung der Knöpfe ist lediglich eine Hilfe, um bei Bedarf
  denselben Knopf wiederzufinden.

- Das Fenster ist deswegen so hoch, um immer allen Knöpfen genügend Platz
  zu bieten.

- Nur mit dem "QUIT"-Knopf oben links können Sie jederzeit beenden. Bei
  einem erneuten Start sind dann alle Knöpfe da, wo Sie sie plaziert hatten.

                                                              Viel Erfolg!
```

A.2 Einzelergebnisse sortiert nach Probanden

In den nachfolgenden Abbildungen A.1 bis A.5 sind die Streuungsdiagramme zwischen den Perzeptionsergebnissen jeder einzelnen der 60 Versuchspersonen und den mittleren Perzeptionsergebnissen über alle 60 Versuchspersonen dargestellt. Auf den X-Achsen der Diagramme sind demnach die durch das Programm aus Kap. 7.4 erhaltenen Rohdaten abgetragen. Hierbei entsprechen 100% dem zweiten, in Abb. 7.6 mit „1.“ gekennzeichneten Ankerschall. Desweiteren ist in jedem Diagramm der Korrelationskoeffizient r sowie die prozentuale mittlere Abweichung der Stimuli jeder Versuchspersonen von den Gruppenmittelwerten (*mittl. Abw.*) angegeben.

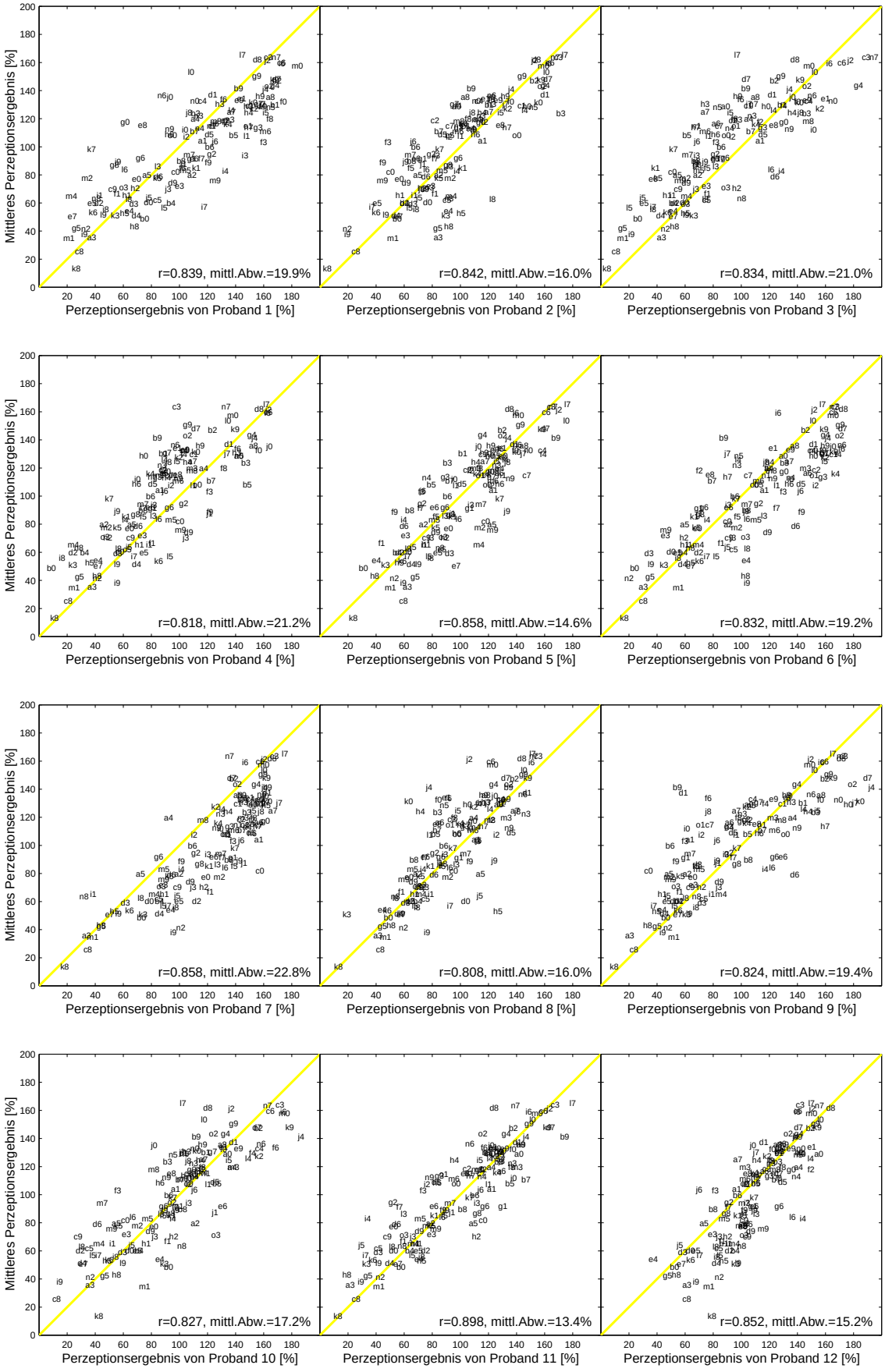

Abb. A.1: Einzelergebnisse der Probanden 1 bis 12 korreliert mit dem Gruppenergebnis.

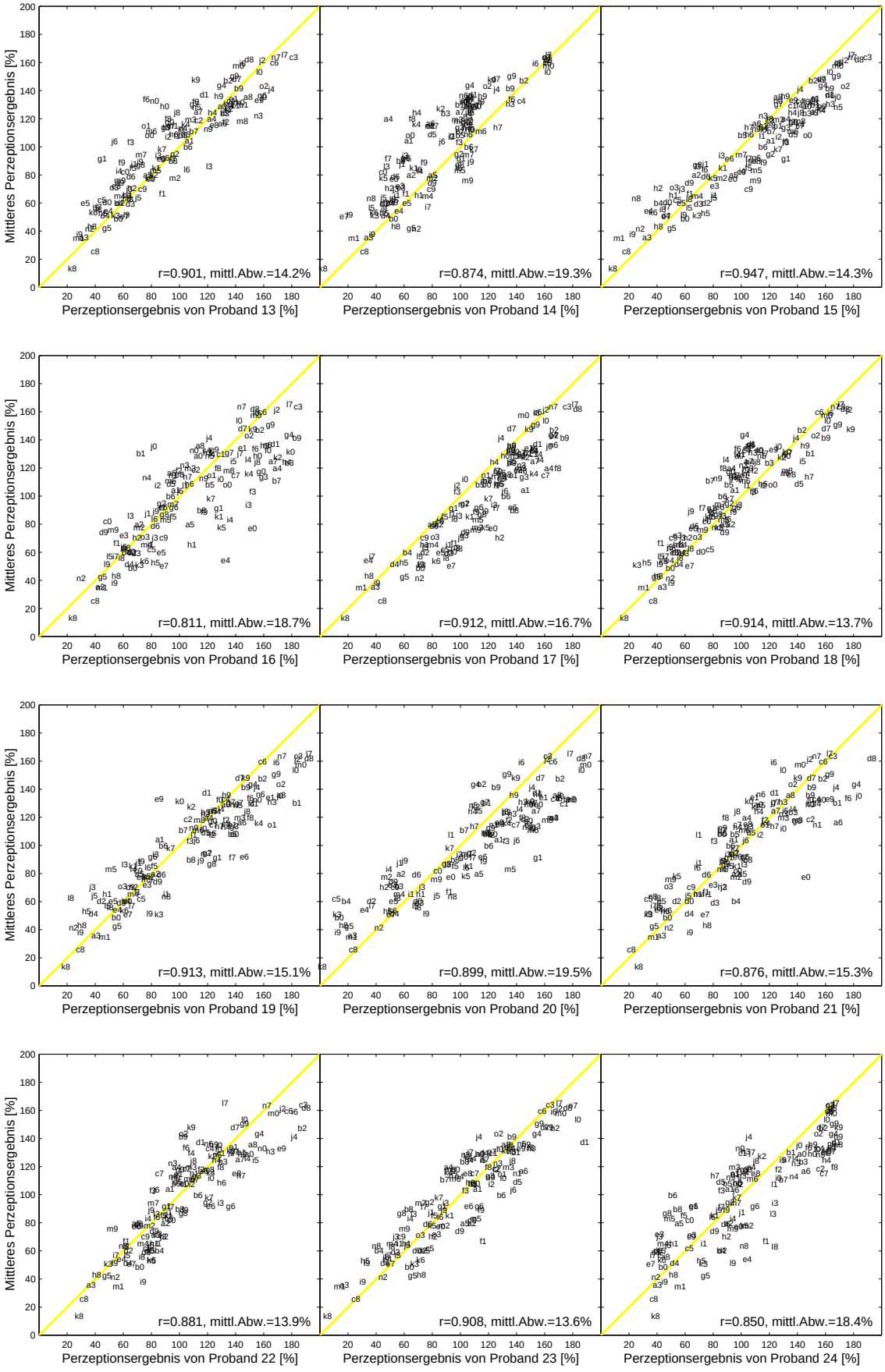

Abb. A.2: Einzelergebnisse der Probanden 13 bis 24 korreliert mit dem Gruppenergebnis.

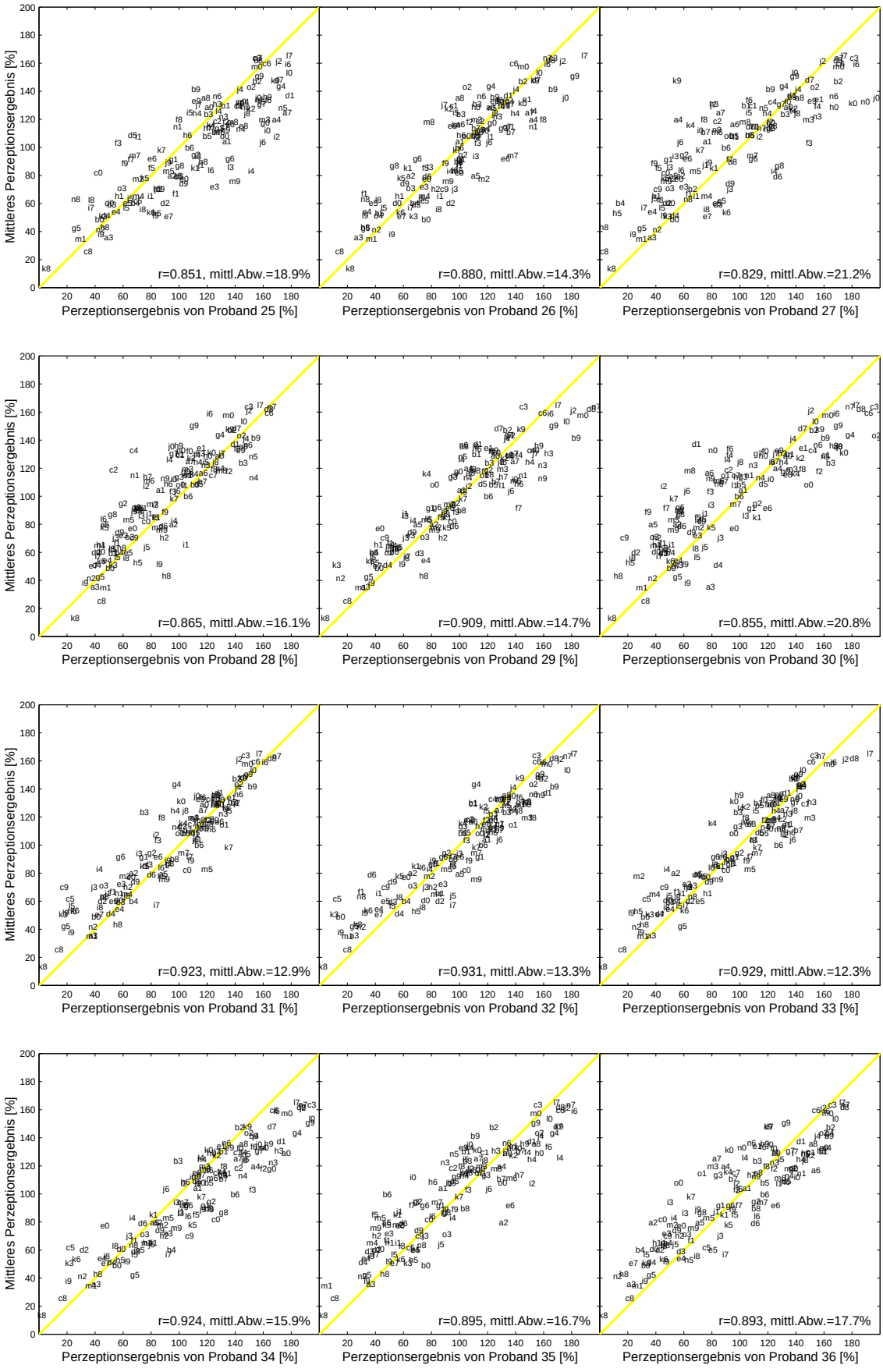

Abb. A.3: Einzelergebnisse der Probanden 25 bis 36 korreliert mit dem Gruppenergebnis.

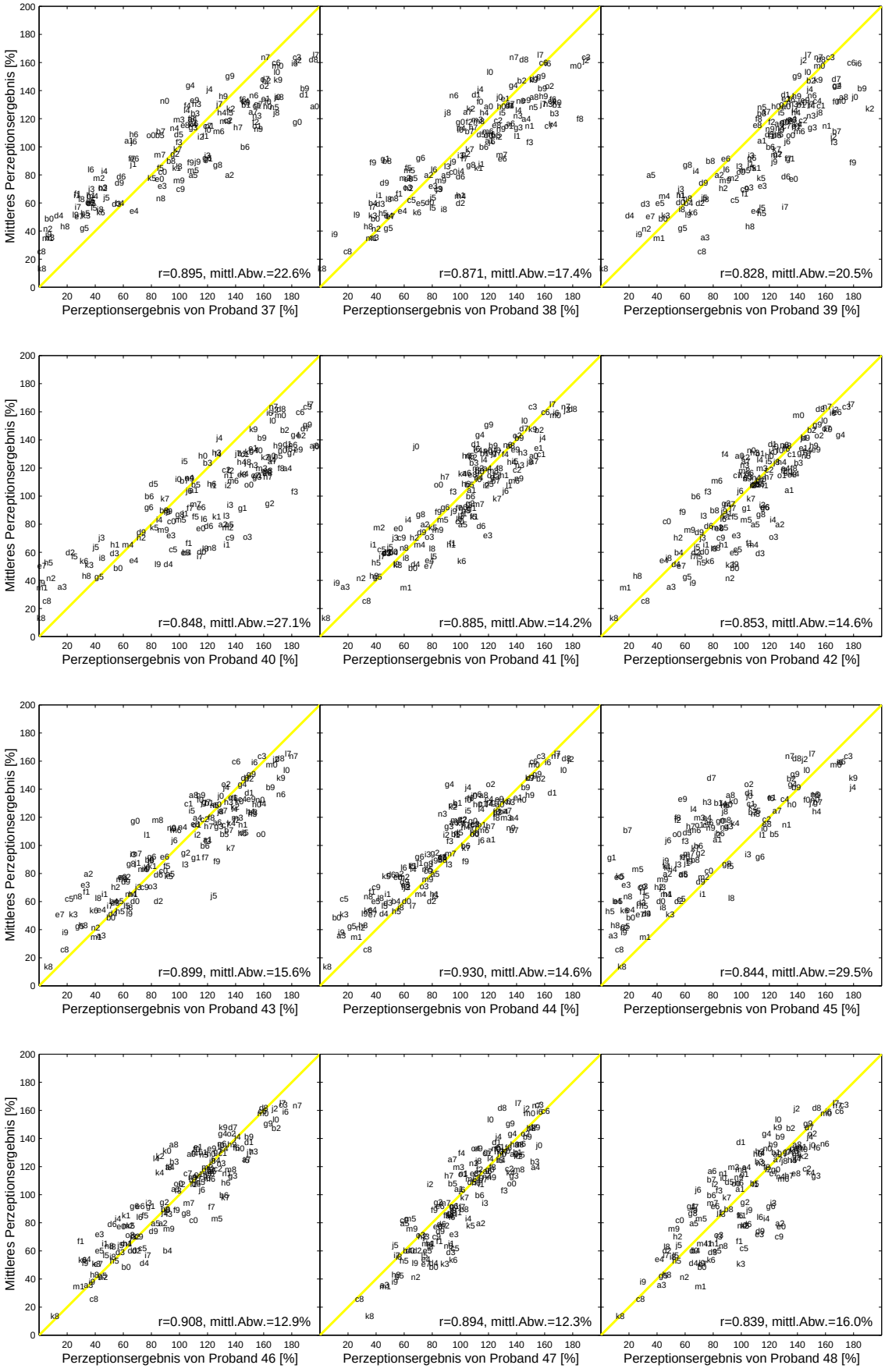

Abb. A.4: Einzelergebnisse der Probanden 37 bis 48 korreliert mit dem Gruppenergebnis.

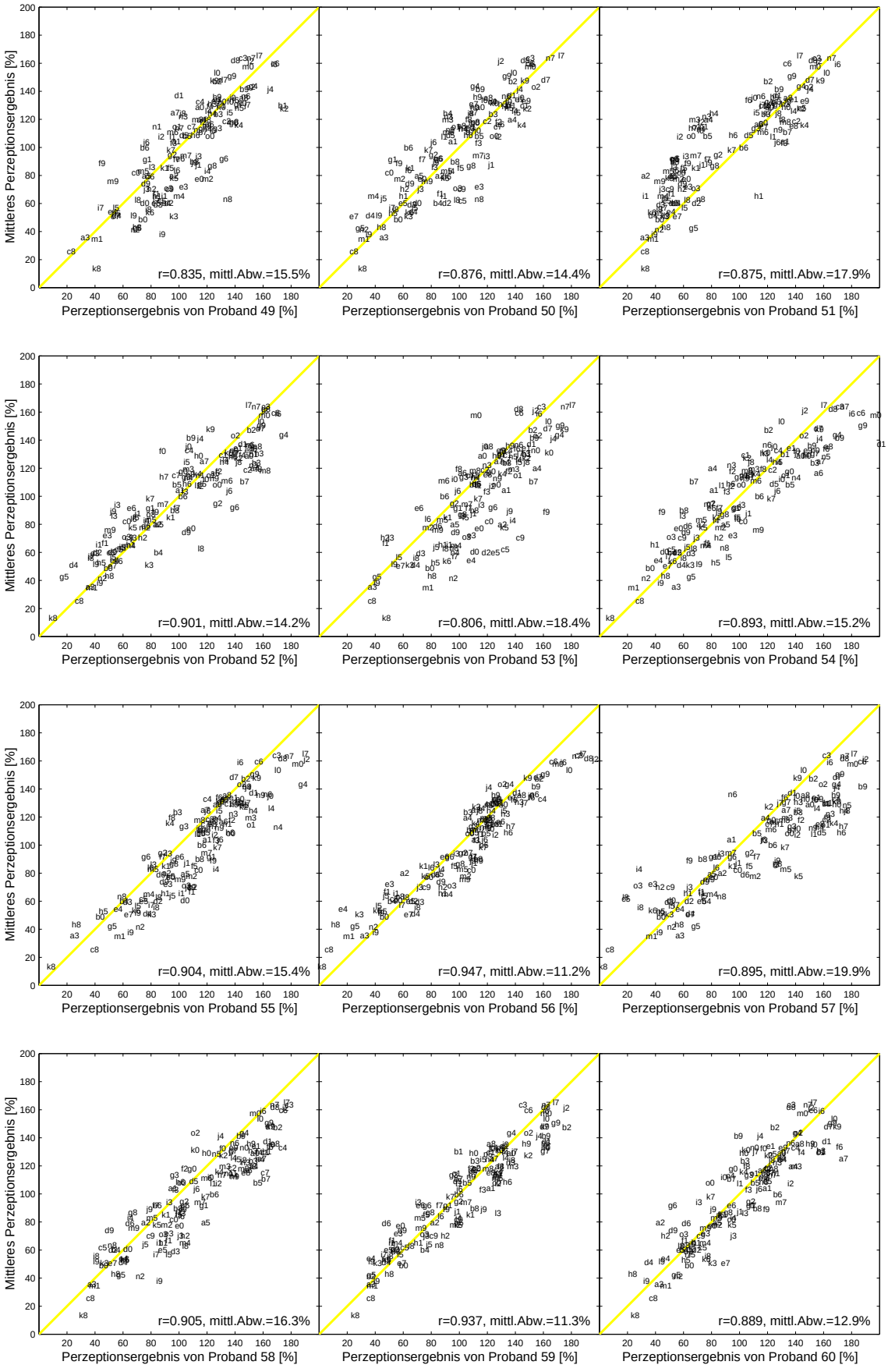

Abb. A.5: Einzelergebnisse der Probanden 49 bis 60 korreliert mit dem Gruppenergebnis.

A.3 Einzelergebnisse sortiert nach Stimuli

In den nachfolgenden Abbildungen A.6 bis A.11 sind die Histogramme aller 141 Stimuli dargestellt, die jeweils über 60 Urteile berechnet wurden, da von jeder der 60 Versuchspersonen genau ein Urteil zu jedem Stimulus vorliegt.

Auf den X-Achsen der Histogramme sind die Häufigkeiten der durch das Programm aus Kap. 7.4 erhaltenen Rohdaten abgetragen. Hierbei entsprechen 100% dem zweiten, in Abb. 7.6 mit „1.“ gekennzeichneten Ankerschall. Zusätzlich ist in jedem Diagramm der Mittelwert als Sternchen sowie die Standardabweichung als Balken auf der X-Achse eingezeichnet.

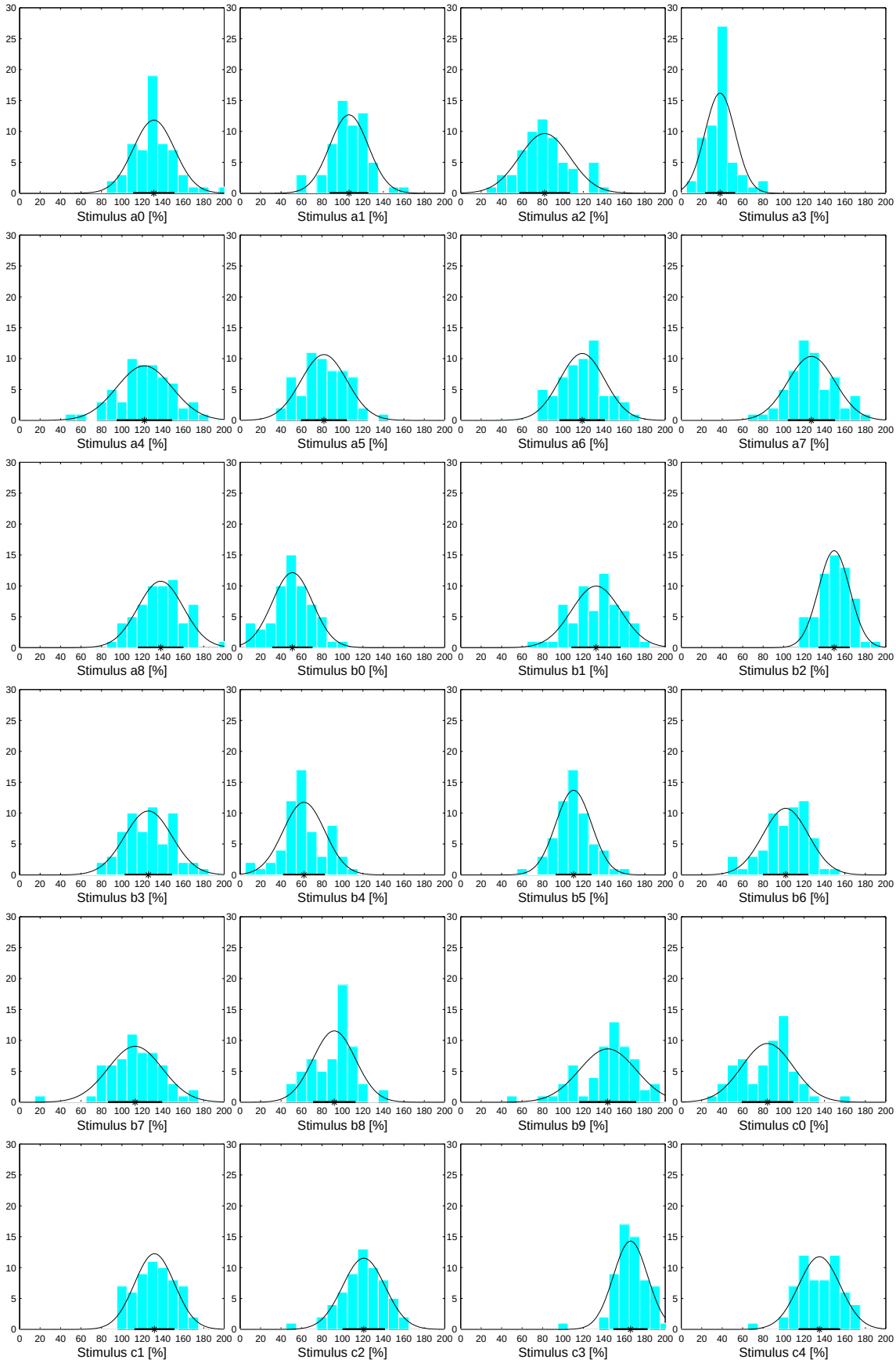

Abb. A.6: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *a0* bis *c4*.

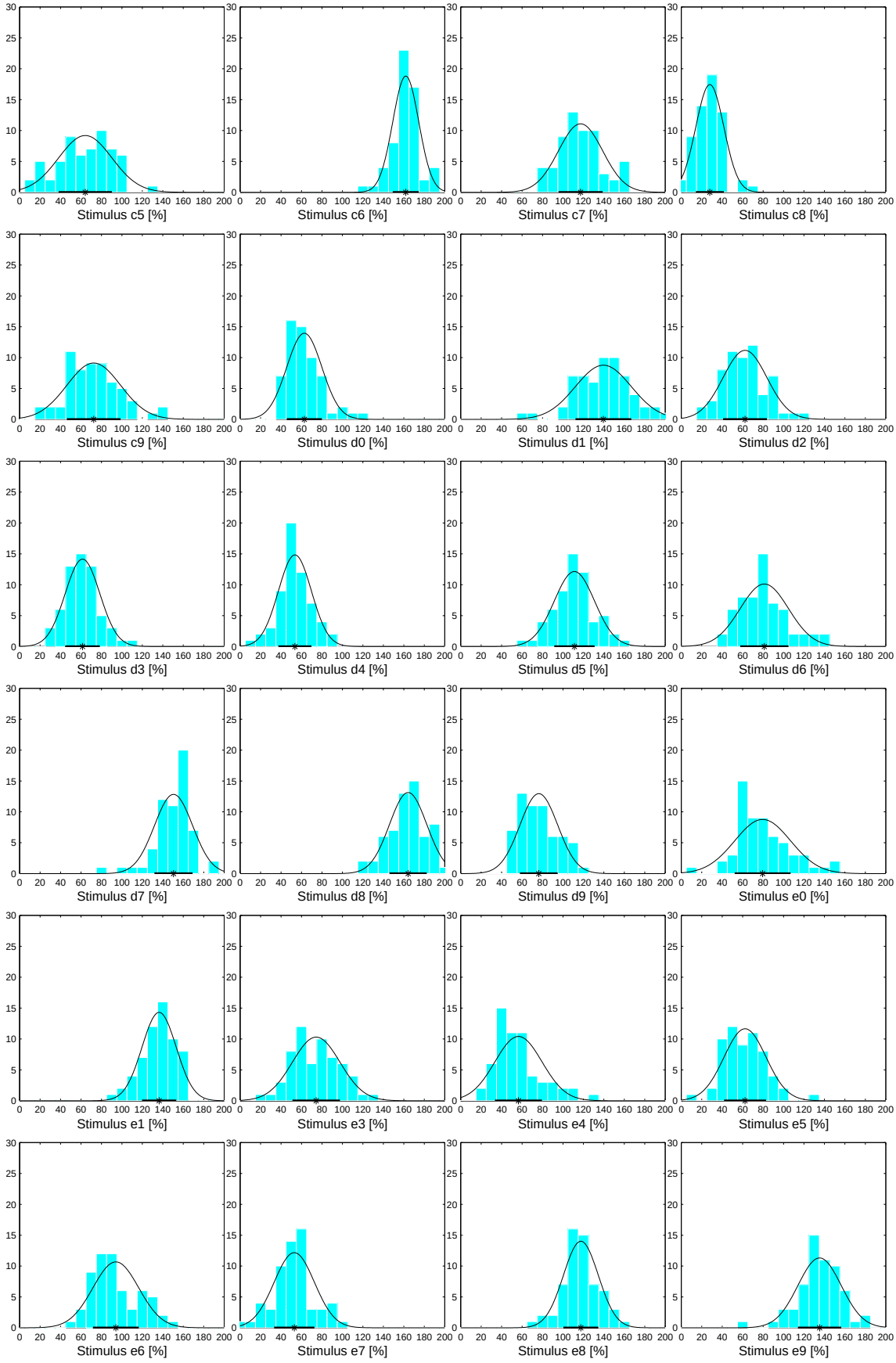

Abb. A.7: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *c5* bis *e9*.

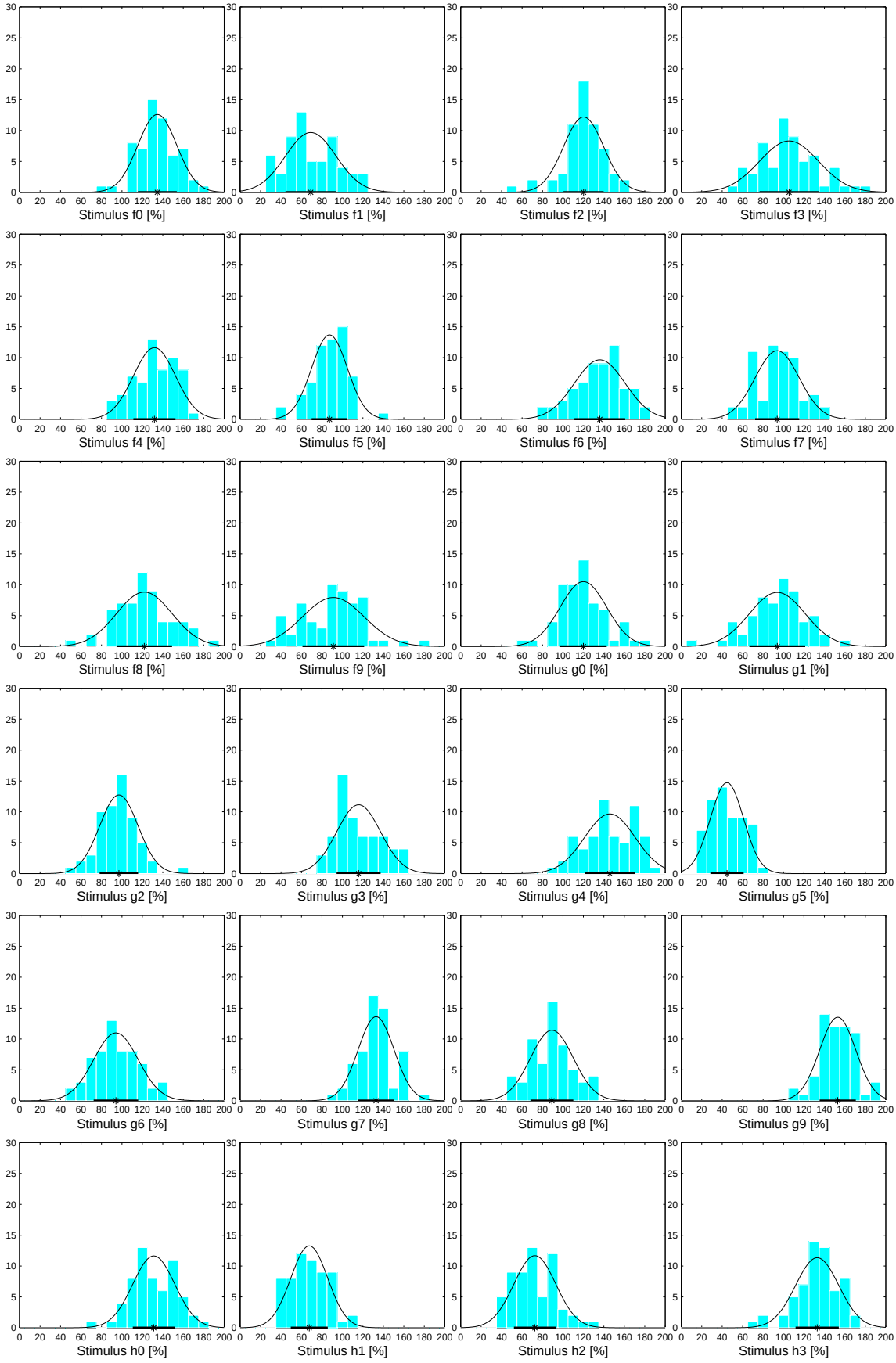

Abb. A.8: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *f0* bis *h3*.

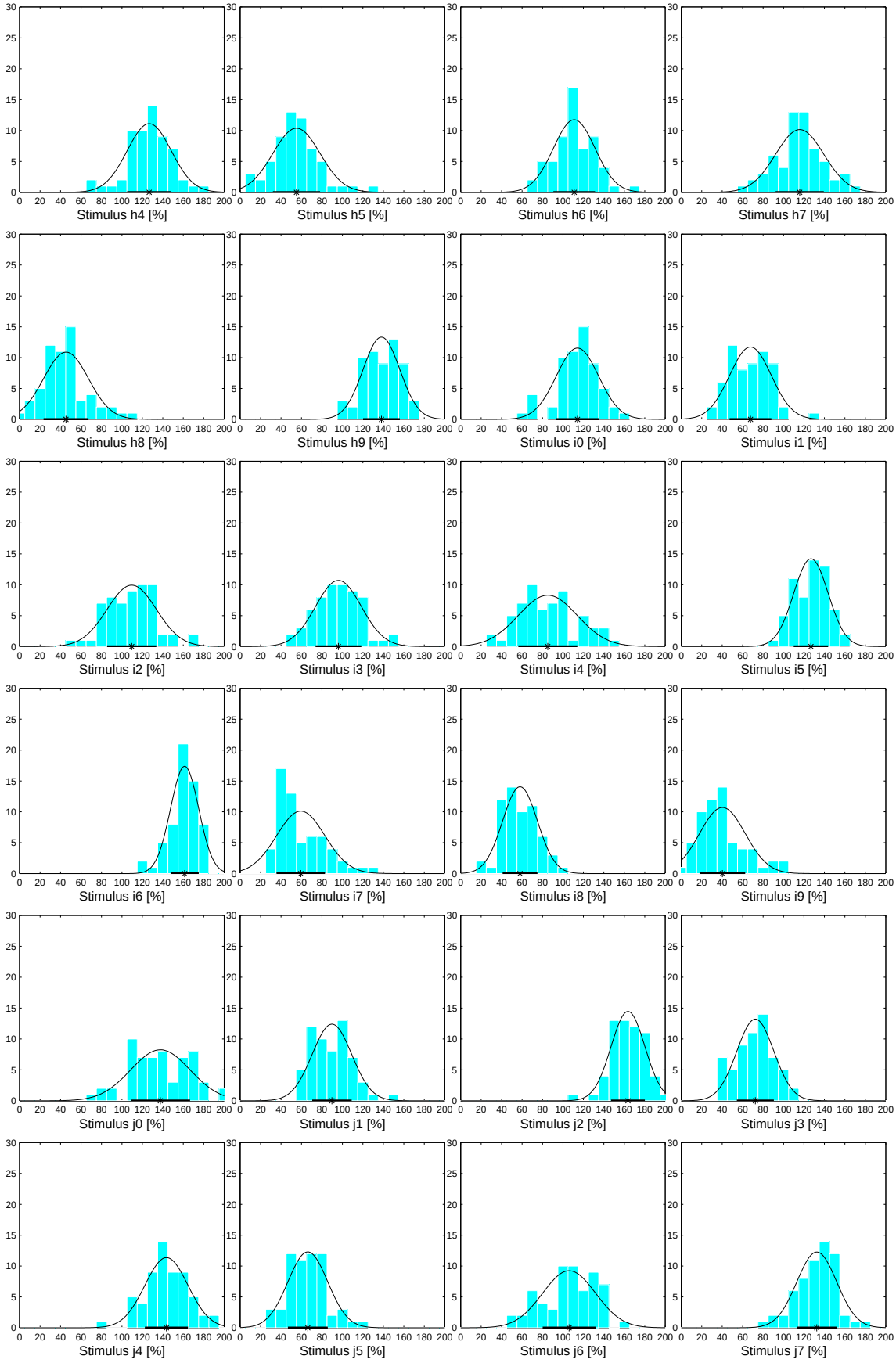

Abb. A.9: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *h4* bis *j7*.

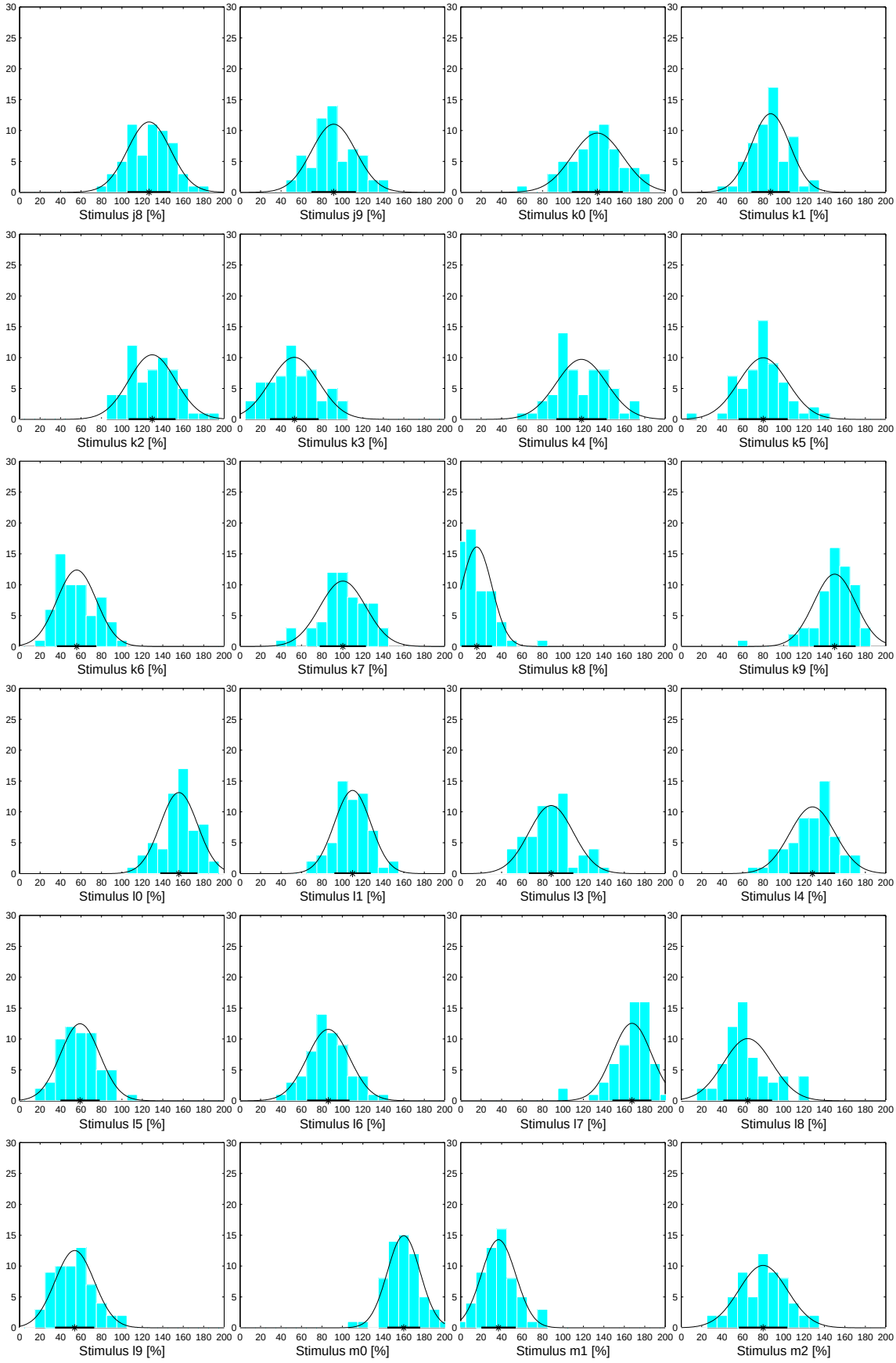

Abb. A.10: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *j8* bis *m2*.

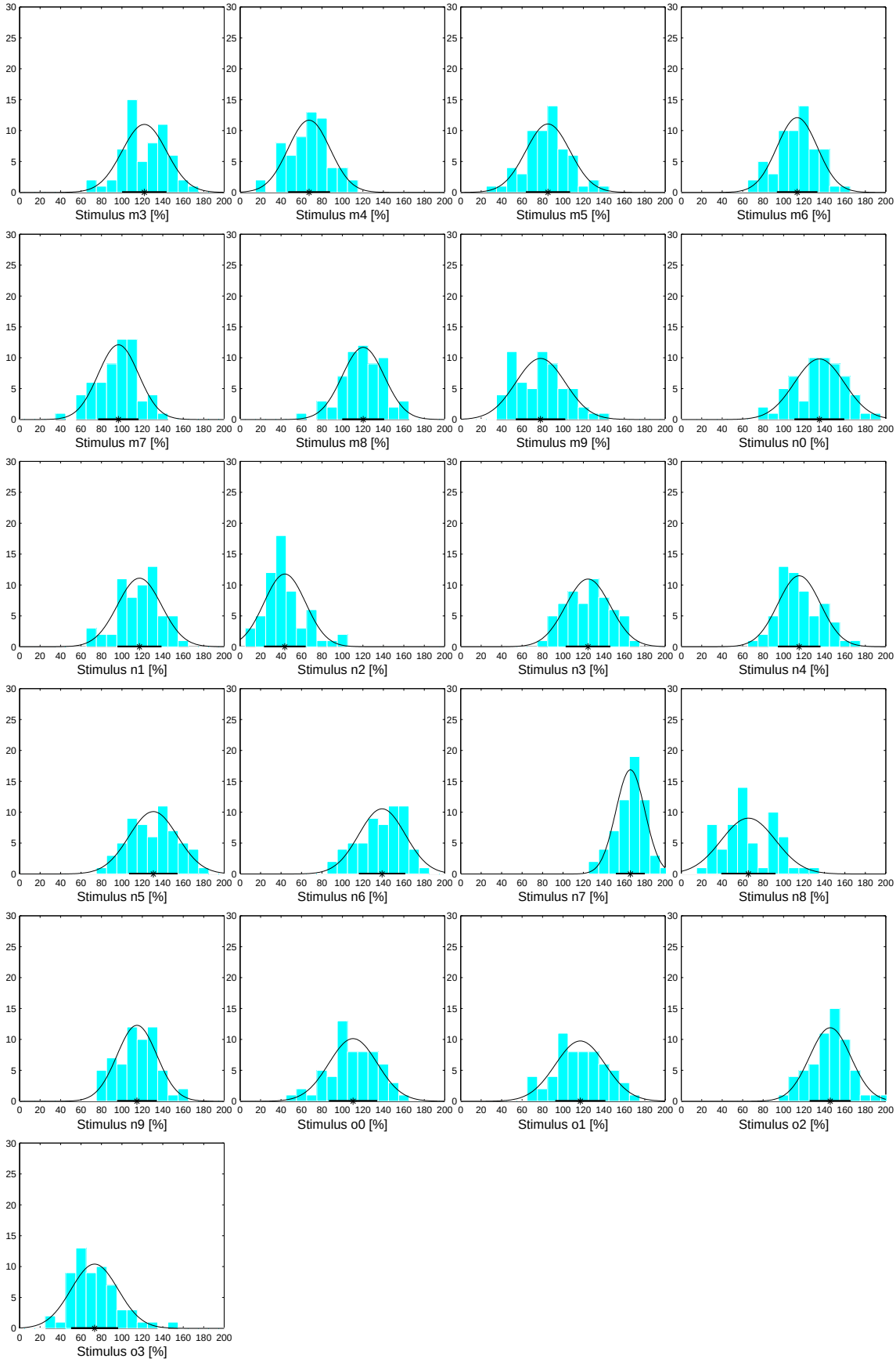

Abb. A.11: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *m3* bis *o3*.

B

Perzeptionsexperiment 5

Bei diesem zweiten Experiment zur Einschätzung der Sprechgeschwindigkeit erhielten die 30 teilnehmenden Versuchspersonen die Instruktionen des vorangegangenen Experiments (siehe Anhang A.1), allerdings mit zwei wesentlichen inhaltlichen Veränderungen: Die Anzahl der Stimuli (100 statt 141) und die damit verbundene Dauer des Experiments (30–50 Minuten statt 45–65 Minuten) waren geringer als beim Vorgängerexperiment.

B.1 Einzelergebnisse sortiert nach Probanden

In den nachfolgenden Abbildungen B.1 bis B.3 sind die Streuungsdiagramme zwischen den Perzeptionsergebnissen jeder einzelnen der 30 Versuchspersonen und den mittleren Perzeptionsergebnissen über alle 30 Versuchspersonen dargestellt. Auf den X-Achsen der Diagramme sind demnach die durch das Programm aus Kap. 7.4 erhaltenen Rohdaten abgetragen. Hierbei entsprechen 100% dem zweiten, in Abb. 7.6 mit „1.“ gekennzeichneten Ankerschall. Desweiteren ist in jedem Diagramm der Korrelationskoeffizient r sowie die prozentuale mittlere Abweichung der Stimuli jeder Versuchspersonen von den Gruppenmittelwerten (*mittl. Abw.*) angegeben.

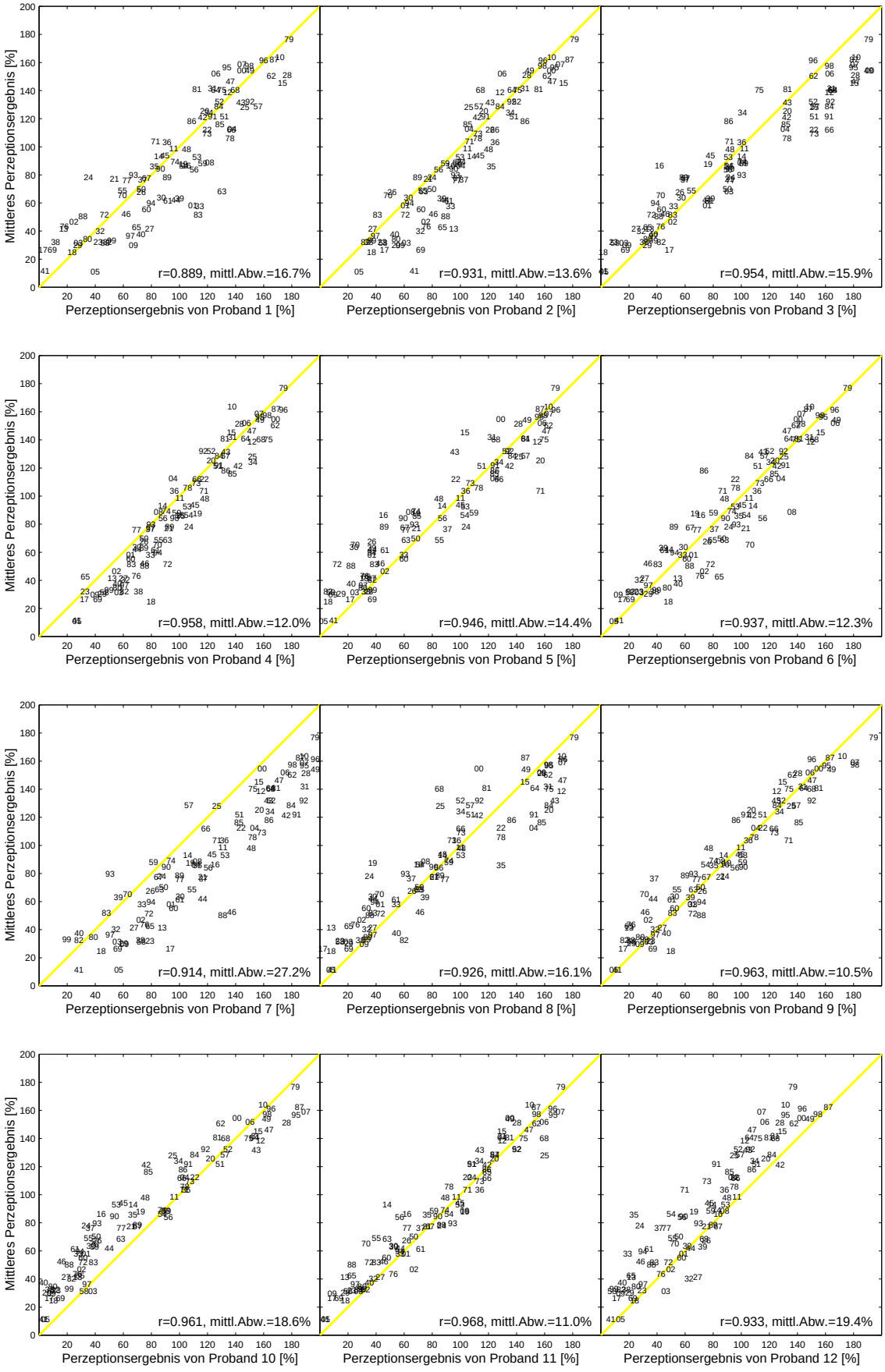

Abb. B.1: Einzelergebnisse der Probanden 1 bis 12 korreliert mit dem Gruppenergebnis.

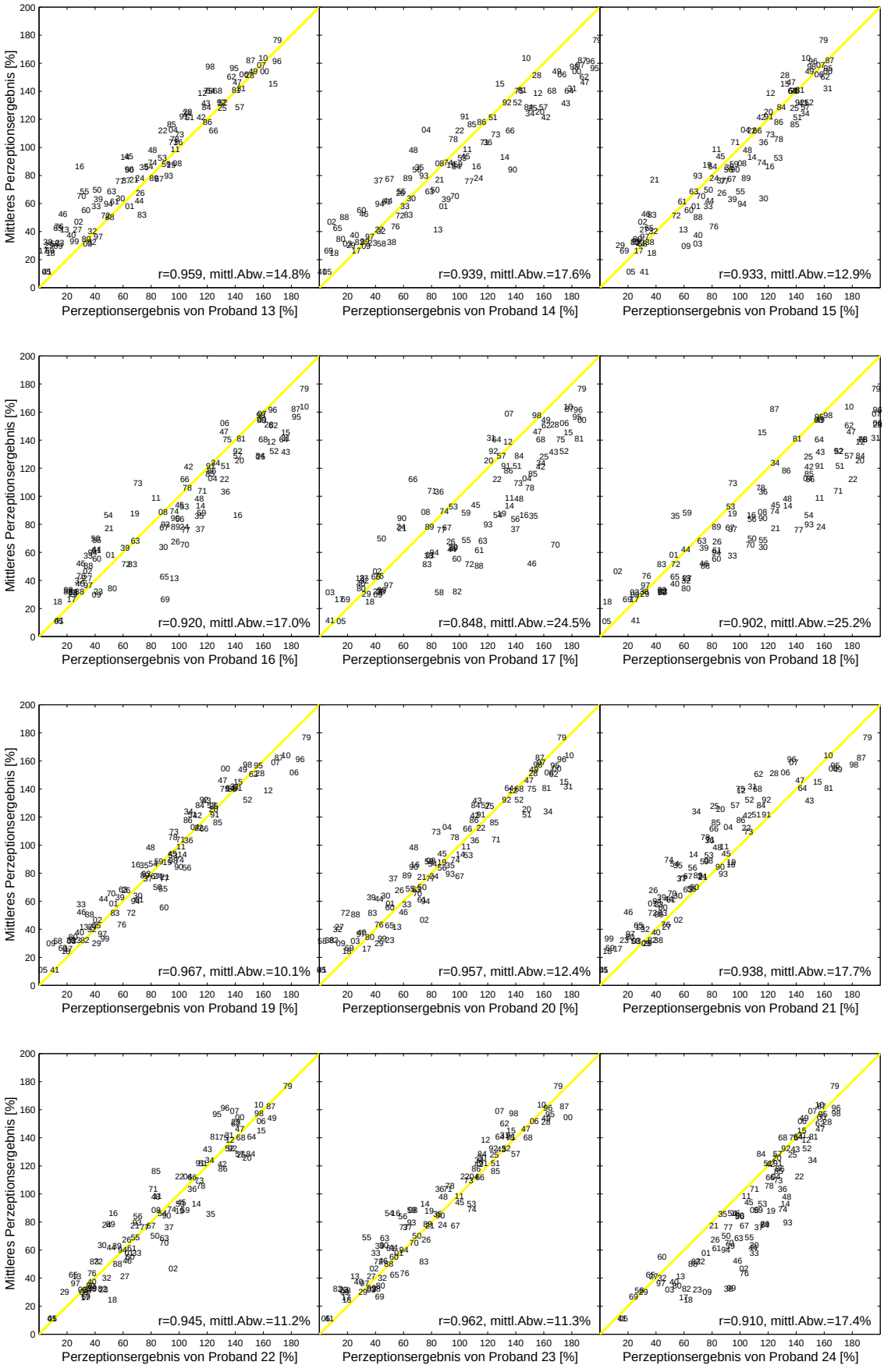

Abb. B.2: Einzelergebnisse der Probanden 13 bis 24 korreliert mit dem Gruppenergebnis.

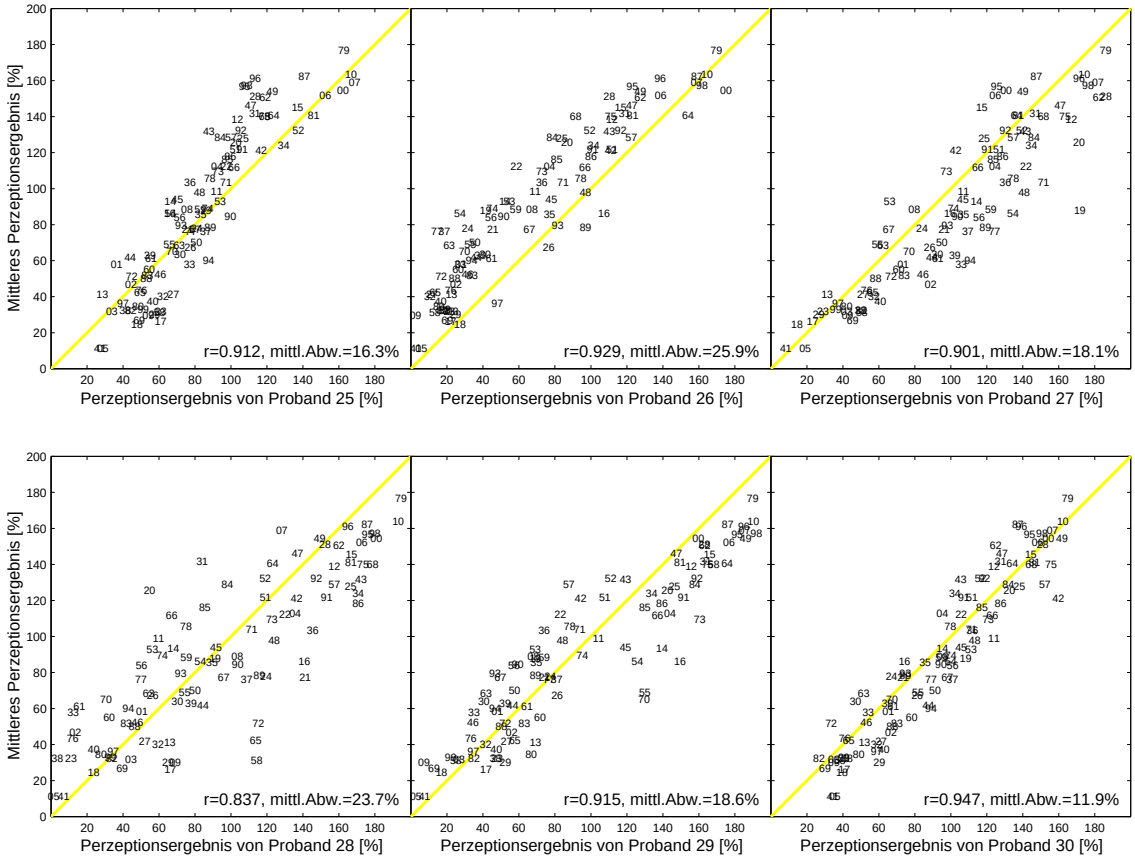

Abb. B.3: Einzelergebnisse der Probanden 25 bis 30 korreliert mit dem Gruppenergebnis.

B.2 Einzelergebnisse sortiert nach Stimuli

In den nachfolgenden Abbildungen B.4 bis B.8 sind die Histogramme der 100 Stimuli des fünften Experiments dargestellt, die jeweils über 30 Urteile berechnet wurden, da von jeder der 30 Versuchspersonen genau ein Urteil zu jedem Stimulus vorliegt.

Auf den X-Achsen der Histogramme sind die Häufigkeiten der durch das Programm aus Kap. 7.4 erhaltenen Rohdaten abgetragen. Hierbei entsprechen 100% dem zweiten, in Abb. 7.6 mit „1.“ gekennzeichneten Ankerschall. Zusätzlich ist in jedem Diagramm der Mittelwert als Sternchen sowie die Standardabweichung als Balken auf der X-Achse eingezeichnet.

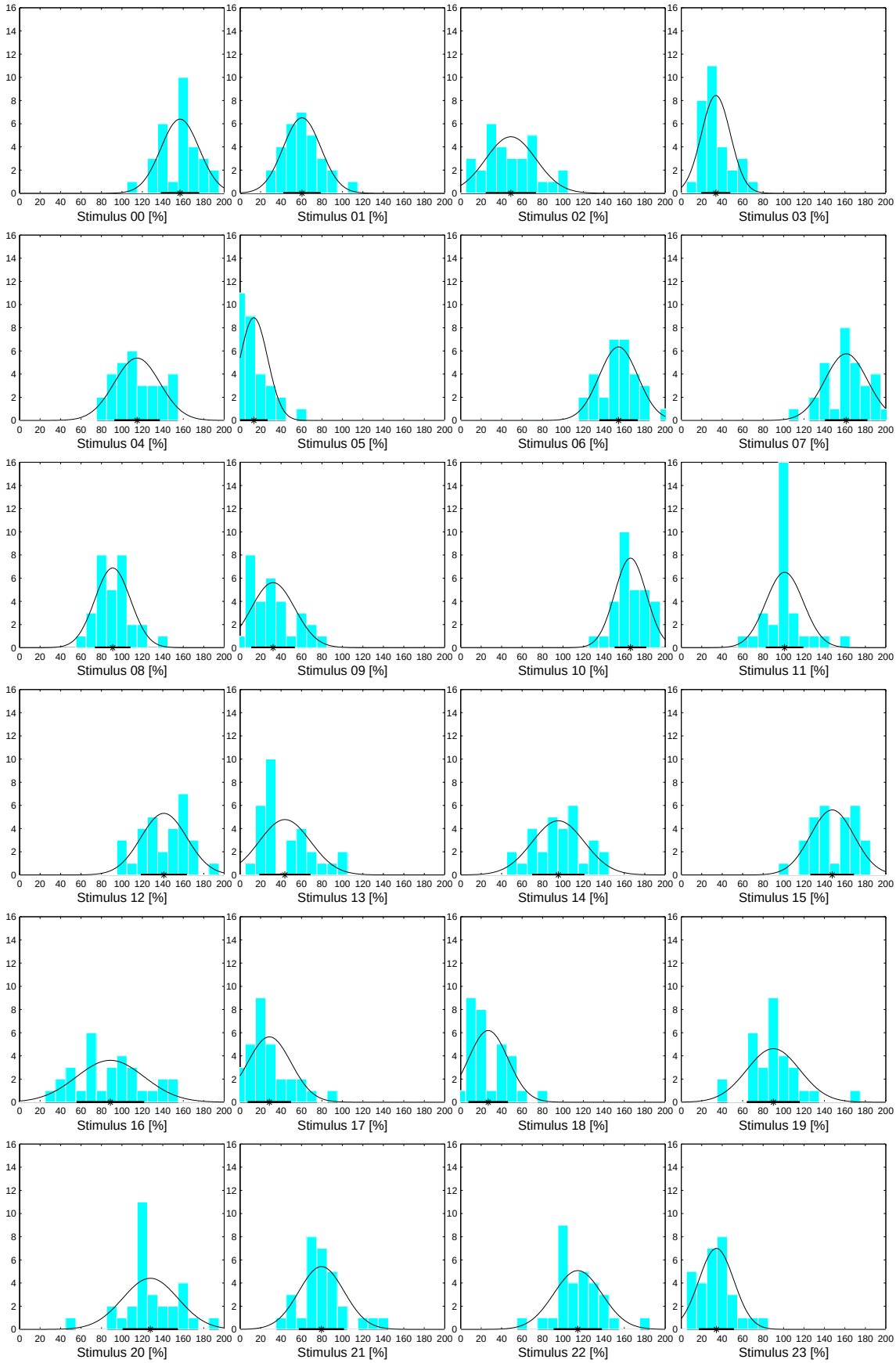

Abb. B.4: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *00* bis *23*.

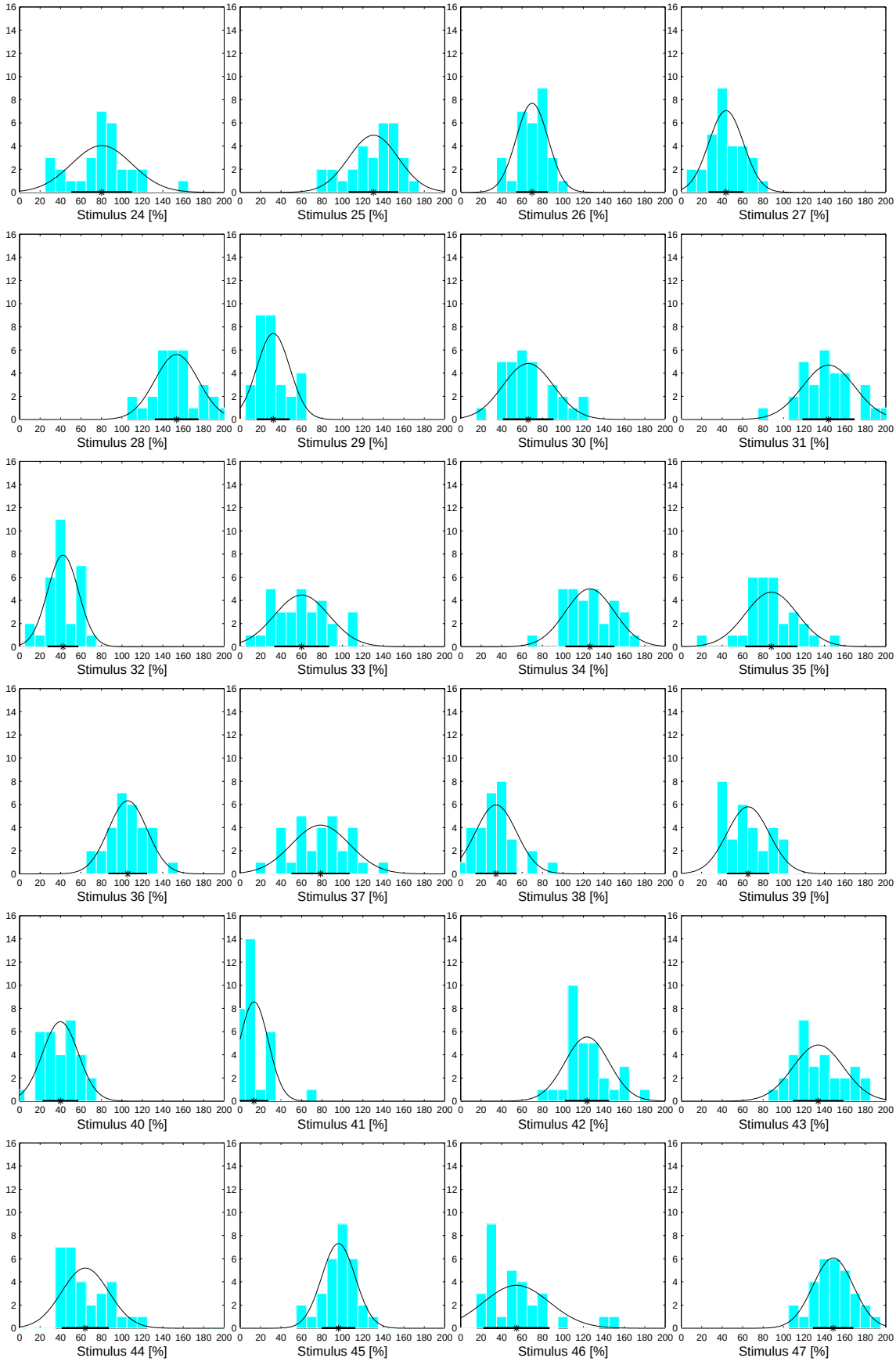

Abb. B.5: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *24* bis *47*.

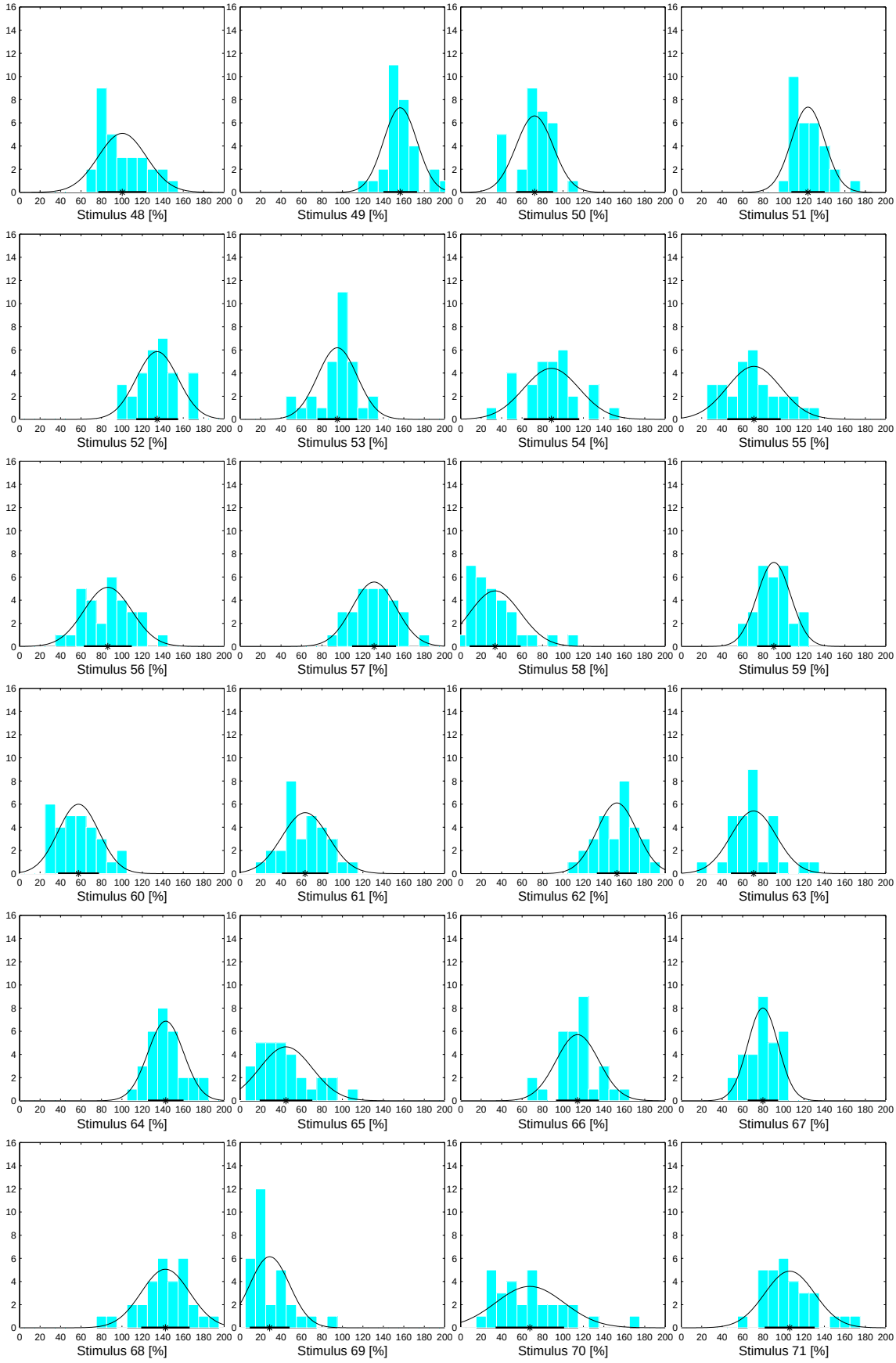

Abb. B.6: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *48* bis *71*.

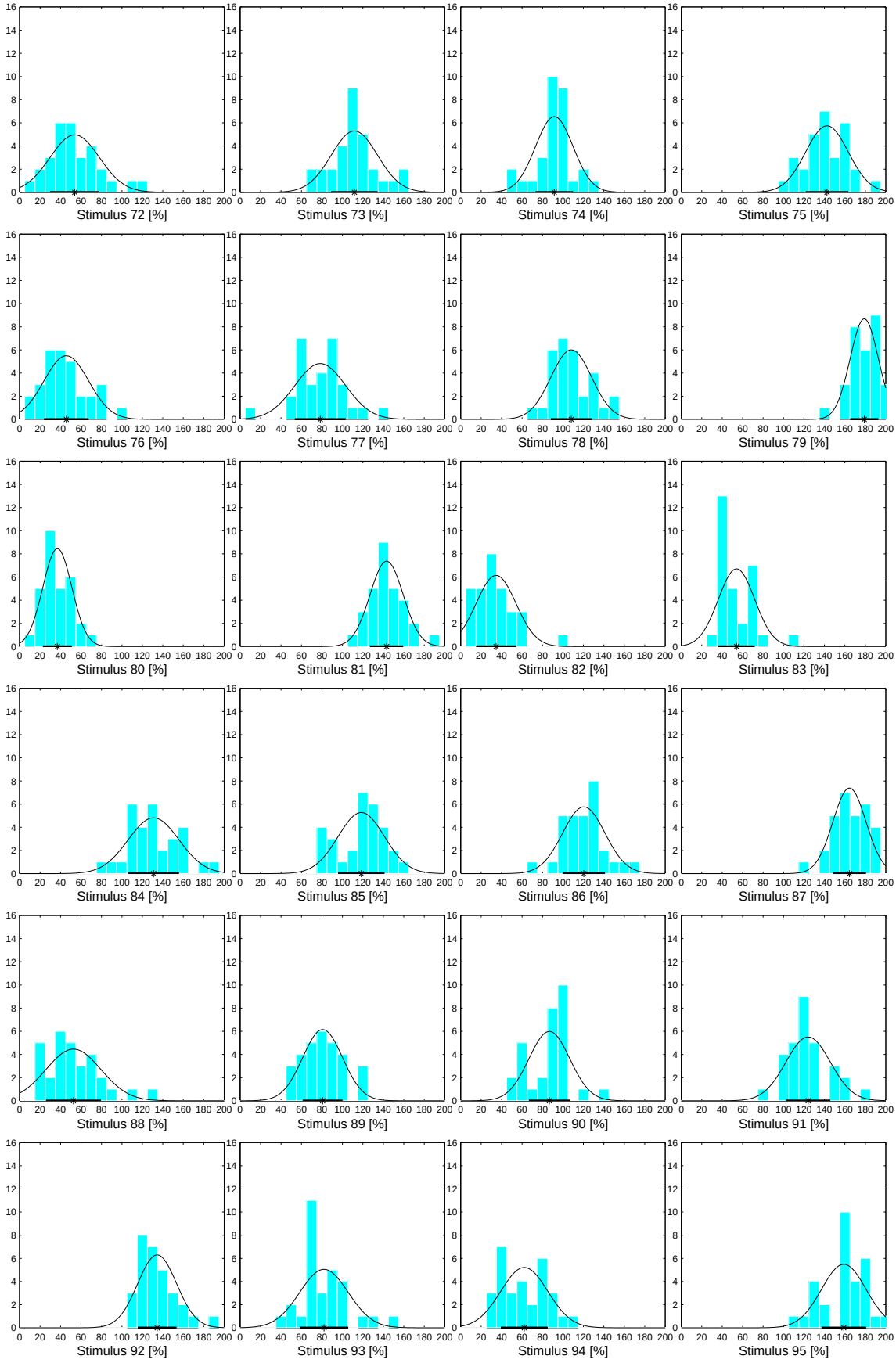

Abb. B.7: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *72* bis *95*.

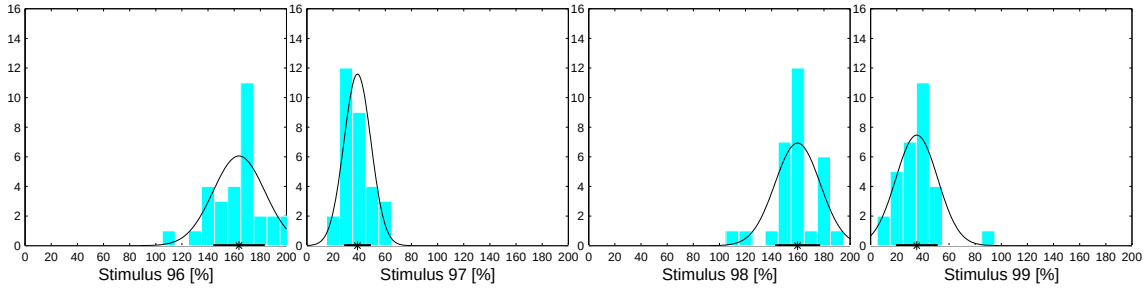

Abb. B.8: Histogramme, Mittelwerte und Standardabweichungen der Stimuli *96* bis *99*.

C

Perzeptionsexperiment 6

Bei diesem Wiederholungsexperiment zur Einschätzung der Sprechgeschwindigkeit erhielten die 10 teilnehmenden Versuchspersonen die Instruktionen des Experiments 4 (siehe Anhang A.1), allerdings mit den gleichen beiden inhaltlichen Veränderungen wie bei Experiment 5: Die Anzahl der Stimuli (100 statt 141) und die damit verbundene Dauer des Experiments (30–50 Minuten statt 45–65 Minuten) waren geringer als beim Experiment 4.

C.1 Einzelergebnisse sortiert nach Probanden

In der nachfolgenden Abbildung C.1 sind die Streuungsdiagramme zwischen den Perzeptionsergebnissen jeder einzelnen der 10 Versuchspersonen und den mittleren Perzeptionsergebnissen über 9 Versuchspersonen dargestellt, wobei jeweils diejenige Versuchsperson nicht in das Gruppenergebnis einbezogen wird, die gerade mit ihm korreliert wird. Bei nur 10 Probanden beträgt der Einfluß des Einzelnen auf das Gruppenergebnis immerhin 10%, so daß im Fall des Einbeziehens seiner Ergebnisse in die Gruppenmittelwerte und einer anschließenden Korrelation mit seinen Ergebnissen die Resultate des Korrelationskoeffizienten und der mittleren Abweichung um bis zu 10% besser erscheinen könnten. Die dargestellten Werte sind also konservativ ermittelt.

Auf den X-Achsen der nachfolgenden Diagramme sind wieder die durch das Programm aus Kap. 7.4 erhaltenen Rohdaten abgetragen. Hierbei entsprechen 100% dem zweiten, in Abb. 7.6 mit „1.“ gekennzeichneten Ankerschall. Desweiteren ist in jedem Diagramm der Korrelationskoeffizient r sowie die prozentuale mittlere Abweichung der Stimuli jeder Versuchspersonen von den konservativen Gruppenmittelwerten (*mittl. Abw.*) angegeben.

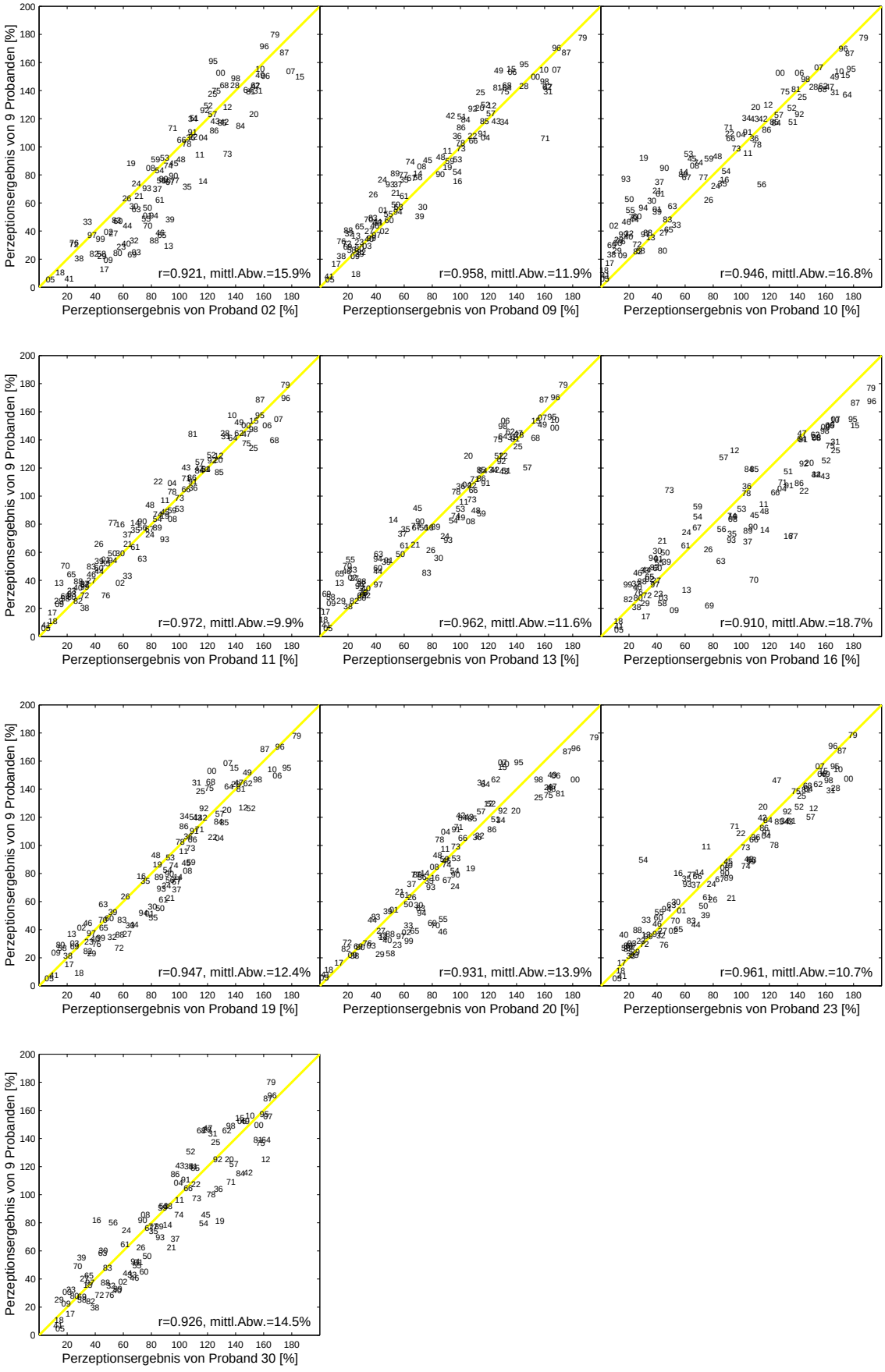

Abb. C.1: Einzelergebnisse der 10 Probanden korreliert mit dem konservativen Gruppenergebnis.

D

Programm zum Kopieren der Prosodie eines Sprechers auf einen anderen

Das hier abgedruckte Shell-Skript verwendet einerseits bisher unveröffentlichte Signalverarbeitungsprogramme auf Unix-Basis aus dem PHD-System und andererseits das undokumentierte .mip-Signaldatei-Format, das allerdings in etwa dem NIST-Format entspricht. Dennoch ist es wegen seiner Struktur sowie der verwendeten Parameterwerte sehr aufschlußreich:

```
# Batch zum Kopieren der PLSR und des Grundfrequenzverlaufs eines neuen Signals auf
# ein phon- und silbensegmentiertes Sprachsignal (in diesem Fall Papagenos Synthesesignal)
#
# 2.3.2001 Hartmut Pfitzinger
#
#
# Zuerst PLSR des Synthesesignals berechnen, invertieren und das normalisierte Signal berechnen:
#
promod/pro_seg2sr -cl8.140 -gl500 -xl150 -ws625 -ss1.000 -ck3.310 -gk450 -xk0 -c36.070 -Ipromod/.mip01 -Ipromod/.mip02 -Opromod/.mip04
promod/pro_sr_mod -iON -m100 -s100 -Ipromod/.mip04 -Opromod/.mip05
promod/pro_sr_chg -iOsz -mHQ2 -s12 -d6 -Ipromod/.mip00 -Ipromod/.mip05 -Opromod/.mip06
#AUTO Sig=5 Part=0 Moth=-1
#AUTO Sig=6 Part=0 Moth=-1
#
# Mit Hilfe von DTW Phon- und Silbensegmentation für das menschliche Sprachsignal berechnen:
#
promod/pro_dtw -b5 -s5 -w25 -r300 -o1.Wrp -mOFF -Ipromod/.mip00 -Ipromod/.mip03 -Opromod/.mip07
promod/pro_warp -iSeg -oData -Ipromod/.mip01 -Ipromod/.mip07 -Opromod/.mip08
promod/pro_warp -iSeg -oData -Ipromod/.mip02 -Ipromod/.mip07 -Opromod/.mip09
#AUTO Sig=7 Part=0 Moth=-1
#AUTO Sig=8 Part=0 Moth=3 Col=blue
#AUTO Sig=9 Part=0 Moth=3 Col=cyan
#
# PLSR des menschlichen Sprachsignals berechnen und auf normierte Länge umrechnen:
#
promod/pro_seg2sr -cl8.140 -gl700 -xl150 -ws625 -ss1.000 -ck3.310 -gk450 -xk0 -c36.070 -Ipromod/.mip08 -Ipromod/.mip09 -Opromod/.mip10
promod/pro_srwrp  -Ipromod/.mip10 -Opromod/.mip11
#AUTO Sig=11 Part=0 Moth=-1
#
# PLSR-normalisiertes Synthesesignal nach vorne verschieben und mit neuer PLSR versehen:
#
promod/pro_func -tx -iOFF -d-50 -g0.000 -oOFF -Ipromod/.mip06 -Opromod/.mip12
promod/pro_sr_chg -iOsz -mHQ2 -s12 -d6 -Ipromod/.mip11 -Ipromod/.mip12 -Opromod/.mip13
#AUTO Sig=12 Part=0 Moth=-1
#
# PLSR des neuen Synthesesignals durch DTW und Warping der Phon- und Silbenrate berechnen:
#
promod/pro_dtw -s10 -w25 -b10 -t1.000 -o1.Wrp -r300 -mOFF -xa0 -xl51 -ya0 -yl75 -Ipromod/.mip03 -Ipromod/.mip13 -Opromod/.mip14
promod/pro_warp -iSeg -oData -Ipromod/.mip08 -Ipromod/.mip14 -Opromod/.mip15
promod/pro_warp -iSeg -oData -Ipromod/.mip09 -Ipromod/.mip14 -Opromod/.mip16
promod/pro_seg2sr -cl8.140 -gl600 -xl150 -ws625 -ss1.000 -ck3.310 -gk450 -xk0 -c36.070 -Ipromod/.mip15 -Ipromod/.mip16 -Opromod/.mip17
#AUTO Sig=14 Part=0 Moth=-1
#AUTO Sig=15 Part=0 Moth=13 Col=blue
#AUTO Sig=16 Part=0 Moth=13 Col=cyan
#
# Grundfrequenzverlauf des PLSR-normalisierten Synthesesignals berechnen:
#
promod/pro_f0insel -ouSeg -mx500 -mn60 -pe0.100 -pl0.550 -pt0.200 -wnPar20 -Ipromod/.mip13 -Opromod/.mip18
promod/pro_seg2f -ws15 -ss1.000 -gs20 -mn20 -ex0 -Ipromod/.mip18 -Opromod/.mip19
#AUTO Sig=18 Part=0 Moth=-1 Col=green
#
```

```
# Verlauf in Stimmtonblöcke umrechnen:
#
promod/pro_sr_mod -iOFF -m0 -s100 -Ipromod/.mip19 -Opromod/.mip20
promod/pro_func -tx -iON -d0.000 -g-42 -oON -Ipromod/.mip20 -Opromod/.mip21
#AUTO Sig=20 Part=0 Moth=-1
#AUTO Sig=21 Part=0 Moth=-1
#
# Grundfrequenzverlauf des menschlichen Signals berechnen:
#
promod/pro_f0insel -ouSeg -mx500 -mn60 -pe0.100 -pl0.550 -pt0.200 -wnPar20 -Ipromod/.mip03 -Opromod/.mip22
promod/pro_seg2f -ws15 -ss1.000 -gs20 -mn40 -ex0 -inLin -Ipromod/.mip22 -Opromod/.mip23
#AUTO Sig=22 Part=0 Moth=-1 Col=green
#
# Neuen Verlauf mit Stimmtonblöcken kombinieren, neue Glottisimpulspositionen bestimmen und Signal ermitteln:
#
promod/pro_func -tx*y -iON -d0.000 -g0 -oON -Ipromod/.mip21 -Ipromod/.mip23 -Opromod/.mip24
promod/pro_f2seg -f16000.000 -Ipromod/.mip24 -Opromod/.mip25
promod/pro_f0_chg -g50 -tHQ -f5 -wHann -Ipromod/.mip13 -Ipromod/.mip18 -Ipromod/.mip25 -Opromod/.mip26
#AUTO Sig=25 Part=0 Moth=-1 Col=green
```

Literaturverzeichnis

[1] Abercrombie, D. (1967). *Elements of general phonetics*. University Press, Edinburgh.

[2] Adams, S. G.; Weismer, G.; Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *J. of Speech and Hearing Research*, 36: S. 41–54.

[3] Allen, G. (1970). The location of rhythmic stress-beats in English: An experimental study. *UCLA Working Papers in Phonetics*, 14: S. 80–132.

[4] Allen, G. (1972). The location of rhythmic stress beats in English: An experimental study I. *Language & Speech*, 15: S. 72–100.

[5] Allen, G. (1972). The location of rhythmic stress beats in English: An experimental study II. *Language & Speech*, 15: S. 179–195.

[6] Allen, G. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *J. of Phonetics*, 3: S. 75–86.

[7] Anderson, S.; Liberman, N.; Gillick, L.; Foster, S.; Hama, S. (1999). The effects of speaker training on ASR accuracy. In: *Proc. of EUROSPEECH '99*, Band 1, S. 403–406, Budapest.

[8] Arvaniti, A. (1999). Effects of speaking rate on the timing of single and geminate sonorants. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 1, S. 599–602, San Francisco.

[9] Auer, P.; Couper-Kuhlen, E.; Müller, F. (1999). *Language in time. The rhythm and tempo of spoken interaction*. Oxford University Press, New York, Oxford.

[10] Auer, P.; Uhmann, S. (1988). Silben- und akzentzählende Sprachen: Literaturüberblick und Diskussion. *Z. für Sprachwissenschaft*, 7(2): S. 214–259.

[11] Augustinus, A. (397/398). *Confessiones*. Buch XI.

[12] Batliner, A.; Kießling, A.; Kompe, R.; Niemann, H.; Nöth, E. (1997). Tempo and its change in spontaneous speech. In: *Proc. of EUROSPEECH '97*, Band 2, S. 763–766, Rhodes; Greece.

[13] Bruhn, H.; Oerter, R.; Rösing, H. (Hrsg.). (1993). *Musikpsychologie*. Rowohlt, Reinbeck.

[14] Butcher, A. (1981). Aspects of the speech pause: Phonetic correlates and communicative functions. Arbeitsberichte (AIPUK) 15, IPDS, Christian-Albrechts-University, Kiel.

[15] Butcher, A. (1981). Phonetic correlates of perceived tempo in reading and spontaneous speech. Work in Progress 3, Phonetics Laboratory, University of Reading, S. 105–117.

[16] Byrd, D.; Tan, C. C. (1996). Saying consonant clusters quickly. *J. of Phonetics*, 24: S. 263–282.

[17] Campbell, W. N. (1988). Extracting speech-rate values from a real-speech database. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP88)*, S. 683–686, New York.

[18] Campbell, W. N. (1988). Speech-rate variation and the prediction of duration. In: *12th Int. Conf. on Computational Linguistics (COLING)*, S. 93–95, Budapest.

[19] Campbell, W. N. (1990). Measuring speech-rate in the Spoken English Corpus. In: Aarts, J.; Meijs, W. (Hrsg.), *Theory and practice in corpus linguistics*, Band 4, S. 61–81. Rodopi, Amsterdam, Atlanta.

[20] Campbell, W. N. (1996). CHATR: A high-definition speech re-sequencing system. In: *ASA and ASJ Third Joint Meeting. Proc. of ASJ*, S. 1223–1228, Honolulu.

[21] Campbell, W. N.; Isard, S. D. (1991). Segment durations in a syllable frame. *J. of Phonetics*, 19: S. 37–47.

[22] Cedergren, H. J.; Perreault, H. (1994). Speech rate and syllable timing in spontaneous speech. In: *Proc. of ICSLP '94*, Band 3, S. 1087–1090, Yokohama.

[23] Chomsky, N.; Halle, M. (1968). *The sound pattern of English*. MIT Press, Cambridge; Massachusetts.

[24] Cook, P. R. (Hrsg.). (1999). *Music, cognition, and computerized sound. An introduction to psychoacoustics*. MIT Press, Cambridge; Massachusetts, London.

[25] Couper-Kuhlen, E. (1993). *English speech rhythm. Form and function in everyday verbal interaction*. John Benjamin, Amsterdam.

[26] Covell, M.; Withgott, M.; Slaney, M. (1998). Mach1: Nonuniform time-scale modification of speech. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP98)*, Band 1, S. 349–352, Seattle.

[27] Cruttenden, A. (1997). *Intonation*. Cambridge University Press, Cambridge, 2. Aufl.

[28] Crystal, T. H.; House, A. S. (1988). The duration of American-English stop consonants: An overview. *J. of Phonetics*, 16: S. 285–294.

[29] Crystal, T. H.; House, A. S. (1988). The duration of American-English vowels: An overview. *J. of Phonetics*, 16: S. 263–284.

[30] Crystal, T. H.; House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *J. of the Acoustical Society of America*, 88(1): S. 101–112.

[31] Cummins, F.; Port, R. (1998). Rhythmic constraints on stress timing in English. *J. of Phonetics*, 26: S. 145–171.

[32] Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *J. of Phonetics*, 25: S. 287–312.

[33] Dankovičová, J. (1999). Articulation rate variation within the intonation phrase in Czech and English. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 1, S. 269–272, San Francisco.

[34] Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *J. of Phonetics*, 11: S. 51–62.

[35] Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. In: *Proc. of the XIth Int. Congress of Phonetic Sciences*, Band 5, S. 447–450, Tallinn.

[36] Delattre, P. C. (1966). A comparison of syllable length conditioning among languages. *Int. Review of Applied Linguistics*, 4: S. 183–198.

[37] Demany, L.; McKenzie, B.; Vurpillot, F. (1977). Rhythm perception in early infancy. *Nature*, 266: S. 718–719.

[38] den Os, E. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica*, 42: S. 124–134.

[39] Deterding, D. (2001). The measurement of rhythm: A comparison of Singapore and British English. *J. of Phonetics*, 29: S. 217–230.

[40] Dimitrova, S. (1998). Bulgarian speech rhythm: Stress-timed or syllable-timed? *J. of the Int. Phonetic Association*, 27: S. 27–33.

[41] Donovan, A.; Darwin, C. J. (1979). The perceived rhythm of speech. In: *Proc. of the IXth Int. Congress of Phonetic Sciences (Copenhagen 1979)*, Band II, S. 268–274.

[42] Duez, D. (1999). Effects of articulation rate on duration in read French speech. In: *Proc. of EUROSPEECH '99*, Band 2, S. 715–718, Budapest.

[43] Edwards, J.; Beckman, M. E.; Fletcher, J. (1991). The articulatory kinematics of final lengthening. *J. of the Acoustical Society of America*, 89(1): S. 369–382.

[44] Ehrlich, S. (1958). Le mécanisme de la synchronisation sensori-motrice. *L'Année Psychologique*, 58: S. 7–23.

[45] Eisen, B.; Tillmann, H. G.; Draxler, C. (1992). Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases. In: *Proc. of ICSLP '92*, Band 2, S. 871–874, Banff; Kanada.

[46] Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *J. of the Acoustical Society of America*, 83(5): S. 1863–1875.

[47] Faltlhauser, R.; Pfau, T.; Ruske, G. (1999). Creating Hidden Markov Models for fast speech by optimized clustering. In: *Proc. of EUROSPEECH '99*, Band 1, S. 407–410, Budapest.

[48] Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, S. 127–152.

[49] Flege, J. E. (1988). Effects of speaking rate on tongue position and velocity of movement in vowel production. *J. of the Acoustical Society of America*, 84(3): S. 901–916.

[50] Fon, J. (1999). Speech rate as a reflection of variance and invariance in conceptual planning in storytelling. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 1, S. 663–666, San Francisco.

[51] Fosler-Lussier, E.; Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in conversational speech. *Speech Communication*, 29(2–4): S. 137–158.

[52] Fougeron, C.; Jun, S.-A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *J. of Phonetics*, 26: S. 45–69.

[53] Foulke, E. (1971). The perception of time-compressed speech. In: Kjeldergaard, P. M.; Horton, D. L.; Jenkins, J. J. (Hrsg.), *The perception of language*, Kapitel 4, S. 79–107. Charles E. Merrill Publishing Company, Columbus; Ohio.

[54] Fowler, C. A. (1977). *Timing control in speech production*. Dissertation, Indiana University, Bloomington.

[55] Fowler, C. A. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, 25: S. 375–388.

[56] Fraisse, P. (1982). Rhythm and tempo. In: Deutsch, D. (Hrsg.), *The psychology of music*, Series in Cognition and Perception, S. 149–180. Academic Press, New York.

[57] Francis, A. L.; Nusbaum, H. C. (1996). Paying attention to speaking rate. In: *Proc. of ICSLP '96*, Band 3, S. 1537–1540, Philadelphia.

[58] Friberg, A.; Sundberg, J. (1995). Time discrimination in a monotonic, isochronous sequence. *J. of the Acoustical Society of America*, 98: S. 2524–2531.

[59] Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *J. of the Acoustical Society of America*, 44(6): S. 1570–1573.

[60] Gay, T. (1978). Effects of speaking rate on vowel formant movements. *J. of the Acoustical Society of America*, 63: S. 223–230.

[61] Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica*, 38: S. 148–158.

[62] Gay, T.; Ushijima, T.; Hirose, H.; Cooper, F. S. (1974). Effect of speaking rate on labial consonant-vowel articulation. *J. of Phonetics*, 2: S. 47–63.

[63] Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, London, New York.

[64] Goldstein, B. E. (1997). *Wahrnehmungspsychologie: Eine Einführung*. Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford.

[65] Gopal, H. S. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations. *J. of Phonetics*, 18: S. 497–518.

[66] Gottfried, T. L.; Miller, J. L.; Payton, P. E. (1990). Effect of speaking rate on the perception of vowels. *Phonetica*, 47: S. 155–172.

[67] Grégoire, A. (1899). Variations de durée da la syllabe française. *La Parole*, 1: S. 161–418.

[68] Greisbach, R. (1992). Reading aloud at maximal speed. *Speech Communication*, 11(4/5): S. 469–473.

[69] Grosjean, F.; Collins, M. (1979). Breathing, pausing and reading. *Phonetica*, 36: S. 98–114.

[70] Grosjean, F.; Lane, H. (1976). How the listener integrates the components of speaking rate. *J. of Experimental Psychology: Human Perception and Performance*, 2: S. 538–543.

[71] Grosjean, F. H.; Lass, N. J. (1977). Some factors affecting the listener's perception of reading rate in English and French. *Language & Speech*, 20: S. 198–208.

[72] Halle, M.; Vergnaud, J.-R. (1978). *Metrical structures in phonology*. MIT Press, Cambridge; Massachusetts.

[73] Handel, S. (1989). Rhythm. In: *Listening. An introduction to the perception of auditory events*, S. 383–459. MIT Press, Cambridge; Massachusetts.

[74] Harris, K. S.; Tuller, B.; Kelso, J. A. S. (1986). Temporal invariance in the production of speech. In: Perkell, J. S.; Klatt, D. H. (Hrsg.), *Invariance and variability in speech processes*, Kapitel 12, S. 243–252. Lawrence Erlbaum Associates, Hillsdale.

[75] Hayes, B. (1981). *A metrical theory of stress rules*. Indiana University Linguistics Club, Indiana. MIT Dissertation.

[76] Hayes, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry*, 15: S. 33–74.

[77] Heid, S. J. G. G. (1998). Phonetische Variation: Untersuchungen anhand des PhonDat2-Korpus. Forschungsberichte (FIPKM) 36, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 193–368.

[78] Heinitz, W. (1921). Die Bewertung der Dauer in phonetischen Aufnahmen. *Vox*, 31(4): S. 153.

[79] Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices*. Springer-Verlag, Berlin, Heidelberg, New York.

[80] Hess, W. (1993). *Digitale Filter*. Teubner, Stuttgart, 2. Aufl.

[81] Hildebrandt, B. (1961). Die arithmetische Bestimmung der durativen Funktion. *Z. für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14(4): S. 328–336.

[82] Hildebrandt, B. (1963). Effektives Sprechtempo, reflexives Sprechtempo und Lautzahlminderung. *Z. für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 16(1–3): S. 63–76.

[83] Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. In: Horne, M. (Hrsg.), *Prosody: Theory and experiment. Studies presented to Gösta Bruce*, S. 335–350. Kluwer Academic Publishers, Dordrecht.

[84] Hoequist, C. E. (1983). Durational correlates of linguistic rhythm categories. *Phonetica*, 40(1): S. 19–31.

[85] Hoequist, C. E. (1983). Parameters of speech rate perception. Arbeitsberichte (AIPUK) 20, IPDS, Christian-Albrechts-University, Kiel, S. 99–138.

[86] Hoequist, C. E. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, 40(3): S. 203–237.

[87] Hoequist, C. E. (1984). Parameters of speech rate perception: Further results. Arbeitsberichte (AIPUK) 21, IPDS, Christian-Albrechts-University, Kiel, S. 149–191.

[88] Hoequist, C. E. (1984). Wahrgenommenes Sprechtempo und signalphonetische Parameter. Forschungsberichte (FIPKM) 19, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 278–286.

[89] Hoequist, C. E.; Kohler, K. J. (1986). Further experiments on speech rate perception with logatomes. Arbeitsberichte (AIPUK) 22, IPDS, Christian-Albrechts-University, Kiel, S. 29–136.

[90] Huggins, A. W. F. (1972). Just noticeable differences for segment duration in natural speech. *J. of the Acoustical Society of America*, 51(4, Pt. 2): S. 1270–1278.

[91] Huggins, A. W. F. (1972). On the perception of temporal phenomena in speech. *J. of the Acoustical Society of America*, 51(4, Pt. 2): S. 1279–1290.

[92] Huggins, A. W. F. (1975). On isochrony and syntax. In: Fant, G.; Tatham, M. A. A. (Hrsg.), *Auditory analysis and perception of speech*, S. 455–464. Academic Press, London, New York, San Francisco.

[93] IPA (Hrsg.). (1999). *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

[94] Janker, P. M. (1989). Der Einfluß von Segmentdauer- und Amplitudenmanipulation auf die P-center-Position einfacher CV-Silben. Forschungsberichte (FIPKM) 27, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 71–139.

[95] Janker, P. M. (1995). Sprechrhythmus, Silbe, Ereignis: Eine experimentalphonetische Untersuchung zu den psychoakustisch relevanten Parametern zur rhythmischen Gliederung sprechsprachlicher Äußerungen. Forschungsberichte (FIPKM) 33, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München.

[96] Jones, D. (1962). *An outline of English phonetics*. W. Heffer & Sons Ltd., Cambridge, 9. Aufl.

[97] Jones, M.; Woodland, P. C. (1993). Using relative duration in large vocabulary speech recognition. In: *Proc. of EUROSPEECH '93*, Band 1, S. 311–314, Technische Universität Berlin.

[98] Kato, H.; Tsuzaki, M.; Sagisaka, Y. (1997). Measuring temporal compensation effect in speech perception. In: Sagisaka, Y.; Campbell, W. N.; Higuchi, N. (Hrsg.), *Computing prosody. Computational models for processing spontaneous speech*, Kapitel 16, S. 251–270. Springer-Verlag, New York, Berlin, Heidelberg.

[99] Kegel, G. (1990). Sprach- und Zeitverarbeitung bei sprachauffälligen und sprachunauffälligen Kindern. In: Kegel, G.; Arnhold, T.; Dahlmeier, K.; Schmid, G.; Tischer, B. (Hrsg.), *Sprechwissenschaft und Psycholinguistik*, Band 4. Westdeutscher Verlag, Opladen.

[100] Kegel, G.; Dames, K.; Veit, S. (1988). Die zeitliche Organisation sprachlicher Strukturen als Sprachentwicklungsfaktor. In: Kegel, G.; Arnhold, T.; Dahlmeier, K.; Schmid, G.; Tischer, B. (Hrsg.), *Sprechwissenschaft und Psycholinguistik*, Band 2, S. 311–335. Westdeutscher Verlag, Opladen.

[101] Kessinger, R. H.; Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *J. of Phonetics*, 25: S. 143–168.

[102] Kessinger, R. H.; Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *J. of Phonetics*, 26: S. 117–128.

[103] Kim, Y.-j.; Oh, Y.-h. (1996). Prediction of prosodic phrase boundaries with considering variable speaking rate. In: *Proc. of ICSLP '96*, Band 3, S. 1505–1508, Philadelphia.

[104] Kohler, K. J. (1981). Perceptual centres? Arbeitsberichte (AIPUK) 16, IPDS, Christian-Albrechts-University, Kiel, S. 207–228.

[105] Kohler, K. J. (1983). Stress-timing and speech rate in German. a production model. Arbeitsberichte (AIPUK) 20, IPDS, Christian-Albrechts-University, Kiel, S. 5–53.

[106] Kohler, K. J. (1986). Invariance and variability in speech timing: From utterance to segment in German. In: Perkell, J. S.; Klatt, D. H. (Hrsg.), *Invariance and variability in speech processes*, Kapitel 13, S. 268–289. Lawrence Erlbaum Associates, Hillsdale.

[107] Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: Duration, F_0 movement, and F_0 level. *Language & Speech*, 29: S. 115–139.

[108] Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: Duration, pitch movement, and pitch level. Arbeitsberichte (AIPUK) 22, IPDS, Christian-Albrechts-University, Kiel, S. 137–177.

[109] Kohler, K. J.; Schäfer, K.; Thon, W.; Timmermann, G. (1981). Sprechgeschwindigkeit in Produktion und Perzeption. Arbeitsberichte (AIPUK) 16, IPDS, Christian-Albrechts-University, Kiel, S. 137–206.

[110] Köhlmann, M. (1982). Sprachsegmentierung mit Hilfe der Rhythmuswahrnehmung. In: *Fortschritte der Akustik (FASE/DAGA '82)*, S. 903–906. DPG, Bad Honnef.

[111] Köhlmann, M. (1984). Bestimmung der Silbenstruktur von fließender Sprache mit Hilfe der Rhythmuswahrnehmung. *Acustica*, 56: S. 120–125.

[112] Koopmans-van Beinum, F. J.; van Donzel, M. E. (1996). Relationship between discourse structure and dynamic speech rate. In: *Proc. of ICSLP '96*, Band 3, S. 1724–1727, Philadelphia.

[113] Kröger, B. J. (1996). Zur phonetischen Realisierung von Sprechtempoänderungen unter Einbeziehung von artikulatorischer Reorganisation: Artikulatorische und perzeptive Untersuchungen. In: Gibbon, D. (Hrsg.), *Natural language processing and speech technology. Results of the 3rd KONVENS conference, Bielefeld*, Kapitel 18, S. 171–185. Mouton de Gruyter, Berlin, New York.

[114] Kuehn, D. P.; Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *J. of Phonetics*, 4: S. 303–320.

[115] Kühnert, B. (1989). Das Timing der artikulatorischen Gesten bei der Produktion rhythmisch gleichmäßiger Sequenzen. Forschungsberichte (FIPKM) 27, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 141–209.

[116] Kühnert, B. (1996). Die alveolar-velare Assimilation bei Sprechern des Deutschen und des Englischen: Kinematische und perzeptive Grundlagen. Forschungsberichte (FIPKM) 34, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 175–392.

[117] Kuwabara, H. (1998). Perceptual and acoustic properties of phonemes in continuous speech for different speaking rate. In: *Proc. of ICSLP '98*, Band 3, S. 1039–1042, Sydney.

[118] Ladd, D. R.; Campbell, W. N. (1991). Theories of prosodic structure: Evidence from syllable duration. In: *Proc. of the XIIth Int. Congress of Phonetic Sciences*, Band 2, S. 290–293, Aix-en-Provence.

[119] Lane, H.; Grosjean, F. (1973). Perception of reading rate by speakers and listeners. *J. of Experimental Psychology*, 97: S. 141–147.

[120] Lass, N. J. (1970). The significance of intra- and intersentence pause times in perceptual judgements of oral reading rate. *J. of Speech and Hearing Research*, 13: S. 777–784.

[121] Laver, J. (1994). *Principles of phonetics*. Cambridge University Press, Cambridge.

[122] Lehiste, I. (1977). Isochrony reconsidered. *J. of Phonetics*, 5: S. 253–263.

[123] Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press, Cambridge; Massachusetts.

[124] Liberman, M. (1975). *The intonational system of English*. Garland, New York, London. Dissertation, publ. 1979.

[125] Liberman, M.; Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8: S. 249–336.

[126] Lindblom, B. E. F. (1990). Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, W. J.; Marchal, A. (Hrsg.), *Speech production and speech modelling*, Nr. 55 in Nato ASI series D: Behavioural and social sciences, S. 403–439. Kluwer Academic Publishers, Dordrecht, Boston, London.

[127] Lisker, L. (1974). On time and timing in speech. In: Sebeok, T. (Hrsg.), *Current trends in linguistics*, Band 12, S. 2387–2418. Mouton & Co., The Hague; Niederlande.

[128] Lloyd James, A. (1940). *Speech signals in telephony*. London.

[129] Lösener, H. (1999). *Der Rhythmus in der Rede. Linguistische und literaturwissenschaftliche Aspekte des Sprachrhythmus*. Max Niemeyer Verlag, Tübingen.

[130] Low, E. L. (1994). *Intonation patterns in Singapore English*. Dissertation, Cambridge University, Dept. of Linguistics.

[131] Low, E. L.; Grabe, E.; Nolan, F. (2000). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. *Language & Speech*, 43(4): S. 377–401.

[132] Marcus, S. M. (1976). *Perceptual centres*. Cambridge University, Cambridge. Dissertation.

[133] Marey, E. J. (1878). *La méthode graphique dans les sciences expérimentales*. Masson, Paris. 2me tirage avec supplément.

[134] Martínez, F.; Tapias, D.; Álvarez, J.; León, P. (1997). Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In: *Proc. of EUROSPEECH '97*, Band 1, S. 469–472, Rhodes; Greece.

[135] Menzerath, P.; de Lacerda, A. (1933). *Koartikulation, Steuerung und Lautabgrenzung: Eine experimentelle Untersuchung*. Ferdinand Dümmlers Verlag, Berlin, Bonn.

[136] Menzerath, P.; de Oleza S. J., J. M. (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung*. Walter de Gruyter, Berlin, Leipzig.

[137] Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *J. of the Acoustical Society of America*, 58(4): S. 880–883.

[138] Meyer, E. A. (1898). Die Silbe. *Die neueren Sprachen*, 6: S. 479–493.

[139] Meyer, E. A. (1903). *Englische Lautdauer. Eine experimentalphonetische Untersuchung*. Harrassowitz, Uppsala, Leipzig.

[140] Miller, J. L. (1981). Some effects of speaking rate on phonetic perception. *Phonetica*, 38: S. 159–180.

[141] Miller, J. L.; Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *J. of the Acoustical Society of America*, 73(5): S. 1751–1755.

[142] Miller, J. L.; Grosjean, F.; Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41: S. 215–225.

[143] Miller, J. L.; O'Rourke, T. B.; Volaitis, L. E. (1997). Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, 54: S. 121–137.

[144] Miller, J. L.; Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, 46(6): S. 505–512.

[145] Mirghafori, N.; Fosler, E.; Morgan, N. (1995). Fast speakers in large vocabulary continuous speech recognition: Analysis & antidotes. In: *Proc. of EUROSPEECH '95*, Band 1, S. 491–494, Madrid.

[146] Morgan, N.; Fosler, E.; Mirghafori, N. (1997). Speech recognition using on-line estimation of speaking rate. In: *Proc. of EUROSPEECH '97*, Band 4, S. 2079–2082, Rhodes; Greece.

[147] Morgan, N.; Fosler-Lussier, E. (1998). Combining multiple estimators of speaking rate. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP98)*, Band 2, S. 729–732, Seattle.

[148] Morton, J.; Marcus, S. M.; Frankish, C. R. (1976). Perceptual centers (P-centers). *Psychological Review*, 83: S. 405–408.

[149] Moulines, E.; Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6): S. 453–467.

[150] Nishinuma, Y.; Duez, D. (1989). Perceptual optimization of syllable duration in short French sentences. In: *Proc. of EUROSPEECH '89*, Band 2, S. 694–697, Paris.

[151] Nooteboom, S. G. (1979). Perceptual adjustment to speech rate: A case of backward perceptual normalisation. In: *Anniversaries in Phonetics: Studia gratulatoria dedicated to Hendrik Mol*, S. 255–269. Institute of Phonetic Sciences, Amsterdam.

[152] Nooteboom, S. G. (1981). Speech rate and segmental perception or the role of words in phoneme identification. In: Myers, T.; Laver, J.; Anderson, J. R. (Hrsg.), *The cognitive representation of speech*, S. 143–150. North-Holland, Amsterdam.

[153] Nooteboom, S. G. (1991). Some observations on the temporal organisation and rhythm of speech. In: *Proc. of the XIIth Int. Congress of Phonetic Sciences*, Band 1, S. 228–237, Aix-en-Provence.

[154] Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In: Hardcastle, W. J.; Laver, J. (Hrsg.), *The Handbook of Phonetic Sciences*, Nr. 5 in Blackwell Handbooks in Linguistics, Kapitel 21, S. 640–673. Blackwell, Oxford.

[155] Nord, L.; Kruckenberg, A.; Fant, G. (1989). Some timing studies of prose, poetry, and music. In: *Proc. of EUROSPEECH '89*, Band 2, S. 690–693, Paris.

[156] Ohno, S.; Fujisaki, H. (1995). A method for quantitative analysis of the local speech rate. In: *Proc. of EUROSPEECH '95*, Band 1, S. 421–424, Madrid.

[157] Ohno, S.; Fujisaki, H.; Hara, Y. (1998). On the effects of speech rate upon parameters of the command-response model for the fundamental frequency contours of speech. In: *Proc. of ICSLP '98*, Band 3, S. 659–662, Sydney.

[158] Ohno, S.; Fujisaki, H.; Taguchi, H. (1997). A method for analysis of the local speech rate using an inventory of reference units. In: *Proc. of EUROSPEECH '97*, Band 1, S. 461–464, Rhodes; Greece.

[159] Ohno, S.; Fujisaki, H.; Taguchi, H. (1998). Analysis of effects of lexical accent, syntax, and global speech rate upon the local speech rate. In: *Proc. of ICSLP '98*, Band 3, S. 655–658, Sydney.

[160] Ohno, S.; Fukumiya, M.; Fujisaki, H. (1996). Quantitative analysis of the local speech rate and its application to speech synthesis. In: *Proc. of ICSLP '96*, Band 4, S. 2254–2257, Philadelphia.

[161] Okadome, T.; Kaburagi, T.; Honda, M. (1999). Relations between utterance speed and articulatory movements. In: *Proc. of EUROSPEECH '99*, Band 1, S. 137–140, Budapest.

[162] Osaka, Y.; Makino, S.; Sone, T. (1994). Spoken word recognition using phoneme duration information estimated from speaking rate of input speech. In: *Proc. of ICSLP '94*, Band 1, S. 191–194, Yokohama.

[163] Osser, H.; Peng, F. (1964). A cross cultural study of speech rate. *Language & Speech*, 7: S. 120–125.

[164] Ostry, D. J.; Munhall, K. G. (1985). Control of rate and duration of speech movements. *J. of the Acoustical Society of America*, 77(2): S. 640–648.

[165] Peeters, W. J. M. (1991). *Diphthong dynamics*. Dissertation, Rijksuniversiteit te Utrecht, Utrecht; Niederlande, Oktober.

[166] Peterson, G. E.; Lehiste, I. (1960). Duration of syllable nuclei in English. *J. of the Acoustical Society of America*, 32(6): S. 693–703.

[167] Pfau, T. (2000). *Methoden zur Erhöhung der Robustheit automatischer Spracherkennungssysteme gegenüber Variationen der Sprechgeschwindigkeit*. Dissertation, Technische Universität München, September.

[168] Pfau, T.; Faltlhauser, R.; Ruske, G. (2000). A combination of speaker normalization and speech rate normalization for automatic speech recognition. In: *Proc. of ICSLP 2000*, Band 4, S. 362–365, Beijing.

[169] Pfitzinger, H. R. (1996). Two approaches to speech rate estimation. In: *Proc. of the sixth Australian Int. Conf. on Speech Science and Technology (SST '96)*, S. 421–426, Adelaide, Dezember.

[170] Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In: *Proc. of ICSLP '98*, Band 3, S. 1087–1090, Sydney.

[171] Pfitzinger, H. R. (1999). Local speech rate perception in German speech. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 2, S. 893–896, San Francisco.

[172] Pfitzinger, H. R.; Burger, S.; Heid, S. (1996). Syllable detection in read and spontaneous speech. In: *Proc. of ICSLP '96*, Band 2, S. 1261–1264, Philadelphia.

[173] Pickett, J. M.; Pollack, I. (1963). The intelligibility of excerpts from fluent speech: Effect of rate of utterance and duration of excerpts. *Language & Speech*, 6: S. 151–164.

[174] Pike, K. L. (1945). *The intonation of American English*. The University of Michigan Press, Ann Arbor; Michigan.

[175] Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *J. of Phonetics*, 17: S. 175–192.

[176] Pompino-Marschall, B. (1990). *Die Silbenprosdie: Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo*. Linguistische Arbeiten, 247. Max Niemeyer Verlag, Tübingen.

[177] Pompino-Marschall, B. (1992). PhonDat: Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch. Forschungsberichte (FIPKM) 30, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 99–128.

[178] Pompino-Marschall, B.; Piroth, H.-G.; Hoole, P.; Tillmann, H. G. (1984). 'Koartikulation' and 'Steuerung' as factors influencing the perception of 'momentary tempo'. In: Broecke, M. P. R. v. d.; Cohen, A. (Hrsg.), *Proc. of the Xth Int. Congress of Phonetic Sciences (Utrecht 1983)*, Band IIB, S. 537–540, Dordrecht.

[179] Pompino-Marschall, B.; Piroth, H.-G.; Hoole, P.; Tillmann, H. G. (1984). 'Koartikulation' und 'Steuerung' in der Wahrnehmung des 'momentanen Tempos'. Forschungsberichte (FIPKM) 19, Inst. für Phonetik und Sprachliche Kommunikation der Univ. München, S. 306–314.

[180] Pompino-Marschall, B.; Piroth, H.-G.; Tilk, K.; Hoole, P.; Tillmann, H. G. (1982). Does the closed syllable determine the perception of 'momentary tempo'? *Phonetica*, 39(6): S. 358–367.

[181] Pöppel, E. (1978). Time perception. In: Held, R.; Leibowitz, H. W.; Teuber, H.-L. (Hrsg.), *Handbook of sensory physiology*, Band VIII: *Perception*, Kapitel 23, S. 713–729. Springer-Verlag, Berlin, Heidelberg, New York.

[182] Pöppel, E. (1979). Temporal constraints in speech perception. Arbeitsberichte (AIPUK) 12, IPDS, Christian-Albrechts-University, Kiel, S. 221–235.

[183] Rallo Fabra, L. (1999). Speaking-rate effects in voiceless stops produced by Catalan speakers of English. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 2, S. 1417–1420, San Francisco.

[184] Ramus, F.; Nespor, M.; Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 72: S. 1–28.

[185] Rapp, K. (1971). A study of syllable timing. Speech Transmission Lab., Quarterly Progress and Status Report 1, KTH, Stockholm, S. 14–19.

[186] Reichl, W.; Ruske, G. (1993). Syllable segmentation of continuous speech with artificial neural networks. In: *Proc. of EUROSPEECH '93*, Band 3, S. 1771–1774, Technische Universität Berlin.

[187] Richardson, M.; Hwang, M.; Acero, A.; Huang, X. D. (1999). Improvements on speech recognition for fast talkers. In: *Proc. of EUROSPEECH '99*, Band 1, S. 411–414, Budapest.

[188] Roudet, L. (1910). *Eléments de phonétique générale*. Librairie Universitaire H. Walter, Paris.

[189] Rousselot, P. J. (1891). Les modifications phonétiques du langage. *Revue des patois gallo-romans*, 4: S. 65–208.

[190] Rousselot, P. J. (1897/1908). *Principes de phonétique expérimentale*. H. Welter, Paris. 2 Bände.

[191] Sachs, C. (1953). *Rhythm and tempo. A study in music history*. Norton & Company, New York.

[192] Sagisaka, Y.; Campbell, W. N.; Higuchi, N. (Hrsg.). (1997). *Computing prosody. Computational models for processing spontaneous speech*. Springer-Verlag, New York, Berlin, Heidelberg.

[193] Samudravijaya, K.; Singh, S. K.; Rao, P. V. S. (1998). Pre-recognition measures of speaking rate. *Speech Communication*, 24(1): S. 73–84.

[194] Saul, L.; Rahim, M. (1999). Modeling the rate of speech by Markov processes on curves. In: *Proc. of EUROSPEECH '99*, Band 1, S. 415–418, Budapest.

[195] Scripture, E. W. (1902). *The elements of experimental phonetics*. Charles Scribner's Sons, New York.

[196] Scripture, E. W. (1929). *Grundzüge der englischen Verswissenschaft*. Marburg.

[197] Shaiman, S.; Adams, S. G.; Kimelman, M. D. Z. (1995). Timing relationships of the upper lip and jaw across changes in speaking rate. *J. of Phonetics*, 23: S. 119–128.

[198] Shannon, C. E.; Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana, Chicago, London.

[199] Siegler, M. A.; Stern, R. M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In: *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP95)*, Band 1, S. 612–615, Detroit.

[200] Smith, B. L.; Sugarman, M. D.; Long, S. H. (1983). Experimental manipulation of speaking rate for studying temporal variability in children's speech. *J. of the Acoustical Society of America*, 74(3): S. 744–749.

[201] Sonoda, Y. (1987). Effect of speaking rate on articulatory dynamics and motor event. *J. of Phonetics*, 15: S. 145–156.

[202] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J. of Experimental Psychology: Human Perception and Performance*, 7(5): S. 1074–1095.

[203] Tajima, K.; Port, R.; Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *J. of Phonetics*, 25: S. 1–24.

[204] Tennant, J. (1999). A microprosodic approach to analyzing speech rate effects in sociophonetic variation. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 2, S. 1405–1408, San Francisco.

[205] Terhardt, E.; Schütte, H. (1976). Akustische Rhythmus-Wahrnehmung: Subjektive Gleichmäßigkeit. *Acustica*, 35: S. 122–126.

[206] Thon, W. (1992). Struktur eines Datenverarbeitungssystems für das Kieler PhonDat-Projekt: Von der Aufnahme ASL-PhonDat 92 zur Datenanalyse. Arbeitsberichte (AIPUK) 26, IPDS, Christian-Albrechts-University, Kiel, S. 111–173.

[207] Tillmann, H. G. (1964). *Das phonetische Silbenproblem: Eine theoretische Untersuchung*. Dissertation, Universität Bonn.

[208] Tillmann, H. G. (1972). Silbischer Ausprägungskode und Intonation. *Acta Universitatis Carolinae: Philologica 1, Phonetica Pragensia*, III: S. 261–265.

[209] Tillmann, H. G. (1994). Early modern phonetics, especially instrumental and experimental work. In: Asher, R. E.; Simpson, J. M. Y. (Hrsg.), *The Encyclopedia of Language and Linguistics*, Band 6, S. 3082–3095. Pergamon Press, Oxford, New York, Seoul, Tokyo.

[210] Tillmann, H. G.; Mansell, P. (1980). *Phonetik: Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Verlagsgemeinschaft Ernst Klett–J. G. Cotta, Stuttgart.

[211] Tillmann, H. G.; Pfitzinger, H. R. (2000). Parametric High Definition (PHD) speech synthesis-by-analysis: The development of a fundamentally new system creating connected speech by modifying lexically-represented language units. In: *Proc. of ICSLP 2000*, Band 3, S. 295–297, Beijing.

[212] Trouvain, J.; Grice, M. (1999). The effect of tempo on prosodic structure. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, Band 2, S. 1067–1070, San Francisco.

[213] Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague VII. Prag.

[214] Tuerk, A.; Young, S. (1999). Modelling speaking rate using a between frame distance metric. In: *Proc. of EUROSPEECH '99*, Band 1, S. 419–422, Budapest.

[215] Tuller, B. (1980). *Coordination among articulators in speech*. Dissertation, University of Connecticut. unpublished.

[216] Tuller, B.; Fowler, C. A. (1980). Some articulatory correlates of perceptual isochrony. *Perception & Psychophysics*, 27(4): S. 277–283.

[217] Tuller, B.; Fowler, C. A. (1981). The contribution of amplitude to the perception of isochrony. Status Report on Speech Research SR-65, Haskins Laboratories, S. 245–250.

[218] Tuller, B.; Harris, K. S.; Kelso, J. A. S. (1981). Articulatory motor events as a function of speaking rate and stress. Status Report on Speech Research SR-65, Haskins Laboratories, S. 33–62.

[219] Tuller, B.; Harris, K. S.; Kelso, J. A. S. (1982). Stress and rate: Differential transformations of articulation. *J. of the Acoustical Society of America*, 71(6): S. 1534–1543.

[220] Uldall, E. T. (1971). Isochronous stresses in R.P. In: Hammerich, L. L.; Jakobson, R.; Zwirner, E. (Hrsg.), *Form and Substance: Phonetic and Linguistic Papers presented to Eli Fischer-Jørgensen*, S. 205–210. Akademisk Forlag, Copenhagen.

[221] Uldall, E. T. (1978). Rhythm in very rapid R.P. *Language & Speech*, 21(4): S. 397–402.

[222] van Dommelen, W. A. (1992). Segmentieren und Etikettieren im Kieler PhonDat-Projekt. Arbeitsberichte (AIPUK) 26, IPDS, Christian-Albrechts-University, Kiel, S. 197–223.

[223] van Santen, J. P. H. (1997). Segmental duration and speech timing. In: Sagisaka, Y.; Campbell, W. N.; Higuchi, N. (Hrsg.), *Computing prosody. Computational models for processing spontaneous speech*, Kapitel 15, S. 225–249. Springer-Verlag, New York, Berlin, Heidelberg.

[224] van Son, R. J. J. H.; Pols, L. C. W. (1993). How does speaking rate influence vowel formant track parameters? In: van Heuven, V. J.; Pols, L. C. W. (Hrsg.), *Analysis and Synthsis of Speech. Strategic Research towards High-Quality Text-to-Speech Generation*, S. 171–191. Mouton de Gruyter, Berlin, New York.

[225] Vennemann, T. (1995). Der Zusammenbruch der Quantität im Spätmittelalter und sein Einfluß auf die Metrik. *Amsterdamer Beiträge zur älteren Germanistik*, 42: S. 185–223.

[226] Ventsov, A. V. (1981). Temporal information processing in speech perception. *Phonetica*, 38: S. 193–203.

[227] Ventsov, A. V. (1983). What is the reference that sound durations are compared with in speech perception? *Phonetica*, 40(2): S. 135–144.

[228] Verhasselt, J. P.; Martens, J.-P. (1996). A fast and reliable rate of speech detector. In: *Proc. of ICSLP '96*, Band 4, S. 2258–2261, Philadelphia.

[229] von Essen, O. (1949). Sprechtempo als Ausdruck des psychischen Geschehens. *Z. für Phonetik*, S. 317–341.

[230] Wagner, P. (1891). Über die Verwendung des Grützner-Marey'schen Apparates und des Phonographen zu phonetischen Untersuchungen. *Phonetische Studien*, 4: S. 69–83.

[231] Wagner, P. (1892). Französische Quantität. *Phonetische Studien*, 6: S. 1–19.

[232] Wahlster, W. (Hrsg.). (2000). *Verbmobil: Foundations of speech-to-speech translation*. Springer-Verlag, Berlin.

[233] Warner, N.; Arai, T. (2001). Japanese mora-timing: A review. *Phonetica*, 58: S. 1–25.

[234] Weitkus, K. (1931). *Experimentelle Untersuchung der Laut- und Silbendauer im deutschen Satz*. Dissertation, Universität Bonn.

[235] Wells, J. C.; Barry, W. J.; Fourcin, A. J. (1989). Transcription, labelling and reference. In: Fourcin, A. J.; Harland, G.; Barry, W. J.; V., H. (Hrsg.), *Speech technology assessment. Towards standards and methods for the European community*, S. 141–159. Ellis Harwood, Chichester.

[236] Wenk, B. J.; Wioland, F. (1982). Is French really syllable-timed? *J. of Phonetics*, 10: S. 193–216.

[237] Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. *Z. für Physiologie*, 61: S. 161–265.

[238] Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung*, 4: S. 301–350.

[239] Wieneke, G.; Janssen, P.; Belderbos, H. (1987). The influence of speaking rate on the duration of jaw movements. *J. of Phonetics*, 15: S. 111–126.

[240] Wightman, C. W. (1992). *Automatic detection of prosodic constituents*. Dissertation, Boston University Graduate School.

[241] Wood, S. (1973). What happens to vowels and consonants when we speak faster? Working Papers 9, Phonetics Laboratory Lund University, Lund, S. 8–39, Oktober.

[242] Woodrow, H. (1951). Time perception. In: Stevens, S. S. (Hrsg.), *Handbook of experimental psychology*, S. 1224–1236. John Wiley & Sons, New York.

[243] Wu, S.-L.; Kingsbury, B. E. D.; Morgan, N.; Greenberg, S. (1998). Performance improvements through combining phone- and syllable-scale information in automatic speech recognition. In: *Proc. of ICSLP '98*, Band 2, S. 459–462, Sydney.

[244] Zwicker, E.; Feldtkeller, R. (1967). *Das Ohr als Nachrichtenempfänger*. S. Hirzel Verlag, Stuttgart, 2. Aufl.

Autorenverzeichnis