Vortrag im Graduiertenkolleg “Sprachliche Repräsentationen und ihre Interpretation” an der Universität Stuttgart
Stuttgart, den 30. November 2000

Artikulatorische, akustische und perzeptive Untersuchungen zum Vokalsystem des Deutschen

Christine Mooshammer
in Zusammenarbeit mit Christian Geng, Peter Janker und Daniel Pape

Zentrum für Allgemeine Sprachwissenschaften, Typologie und Universalienforschung
Berlin

Vorbemerkungen:
Mitarbeiter
z.T. sehr neue und vorläufige Ergebnisse, Bitte um Anregungen
Bitte um Zwischenfragen

Überblick

I Einleitung: Darstellung von Vokalsystemen und Probleme

II Datenmaterial

III Sprechernormalisierung: Procustes Methode

IV Ergebnisse zur Akustik

V Ergebnisse zur Artikulation

VI Perzeptionstest

VII Zusammenfassung

I Einleitung: Darstellung von Vokalsystemen und Probleme

Entsprechend dem Internationalen Phonetischen Alphabet werden Vokale in Vokaltrapezen dargestellt, d.h. die Symbole sind entsprechend ihrer Lage innerhalb des Vokaltrapez definiert.
Das Vokaltrapez selbst ist nach Daniel Jones einerseits artikulatorisch und andererseits perzeptiv bestimmt. Dabei ist zu beachten, daß die Begriffe ARTIKULATORISCH und PERZEPTIV zum Großteil nicht auf Messungen sondern auf den Gehörseindruck basieren. Es handelt sich also um einen abstrakten Vokalraum.

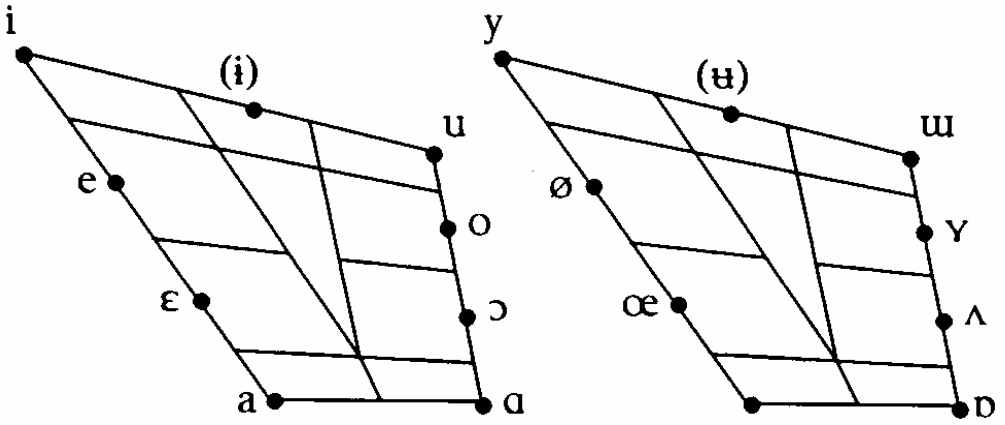
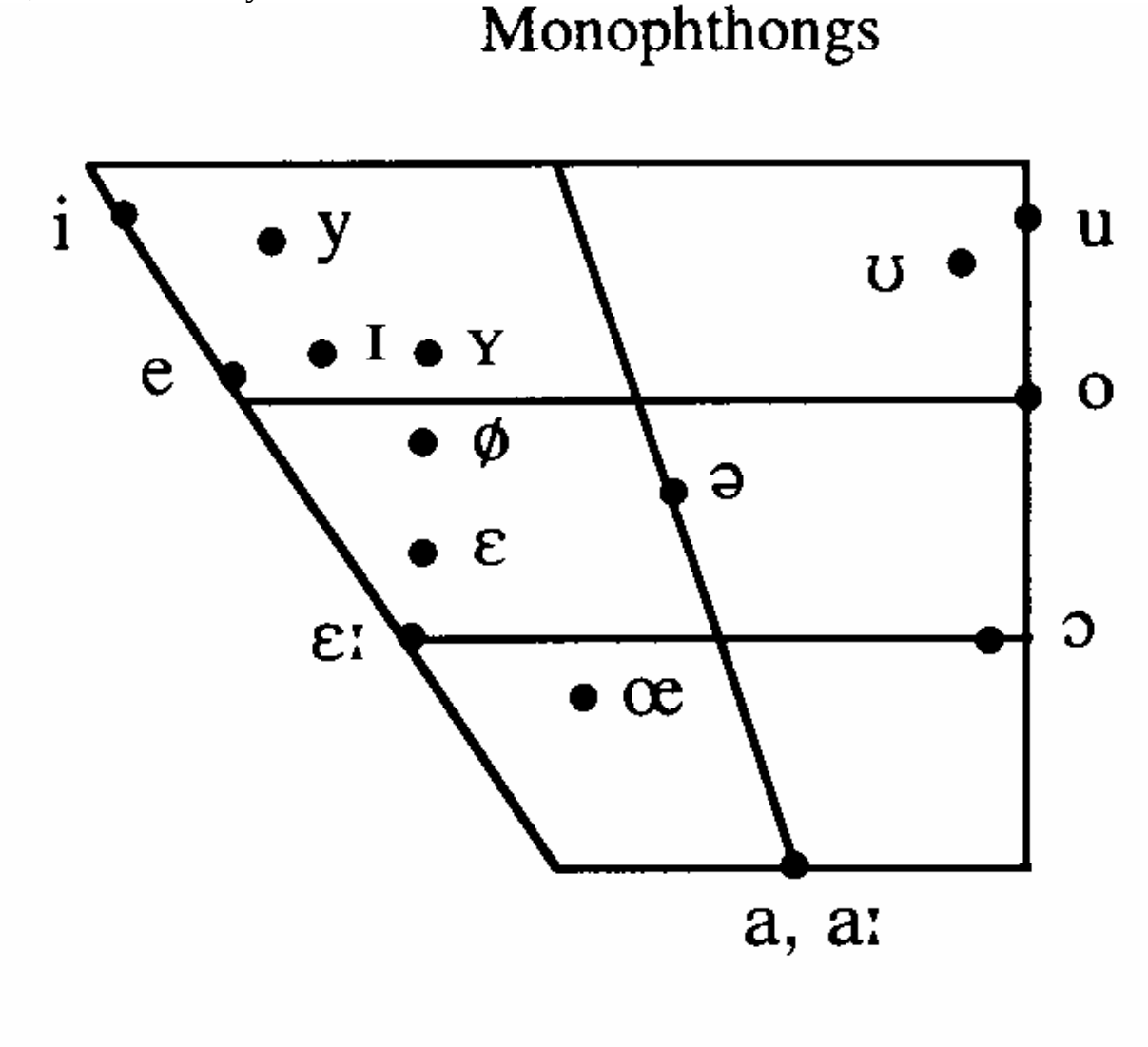


Abb. 97: Die primären (links) und sekundären (rechts) Kardinalvokale.

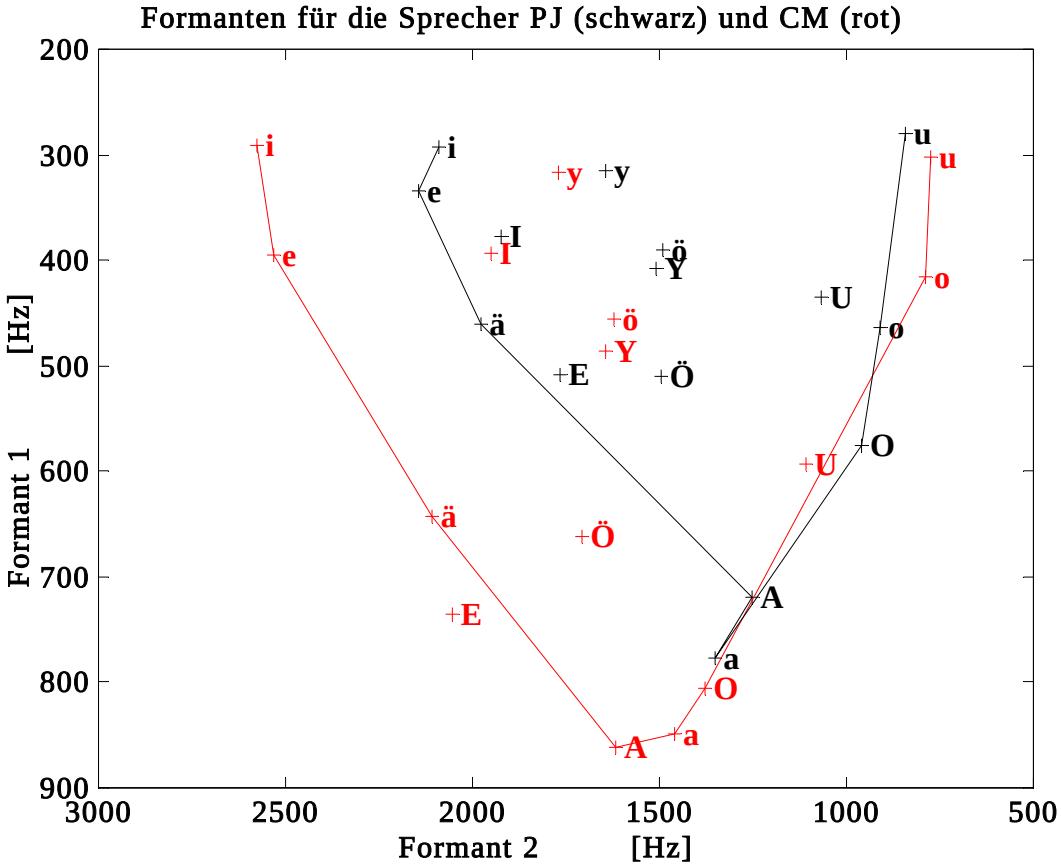
Nach Daniel Jones sind artikulatorisch definiert:
Kardinalvokal 1 /i/: Der mit der höchsten und vordersten Zungenstellung produzierbare Vokal
Kardinalvokal 5 /ɑ/: Der mit der tiefsten und am weitesten nach hinten verlagerten Zungenstellung produzierbare Vokal
Die restlichen primären Kardinalvokale sind durch ihre gleichförmige auditive Distanz voneinander gekennzeichnet, d.h. der Kardinalvokal /e/ unterscheidet sich auditiv vom Kardinal /i/ im gleichen Maße wie der Kardinalvokal /ɛ/ vom /e/. Daniel Jones selbst spricht jedoch von „approximately equal degrees of acoustic separation between each vowel and the next“, wobei er hierzu meines Wissens keine akustischen Messungen durchgeführt hat.
Die Kardinalvokale und der durch sie definierte Raum dient als abstraktes Referenzsystem zur Bestimmung der Vokale in den Einzelsprachen. Sie kommen also in keiner Sprache vor und können streng genommen nur von Daniel Jones selbst bzw. Aufnahmen seiner Realisationen gelernt werden. Siehe hierzu die Homepage des Phonetikinstituts in Utrecht:
http://www.let.uu.nl/~audiufon/data/e_cardinal_vowels.html

Die Lage der Vokale einzelsprachlicher Systeme wird in Bezug auf die Kardinalvokale definiert, d.h. z.B. das deutsche /i/ liegt tiefer als das Kardinal /i/, das deutsche /y/ zentraler und tiefer.



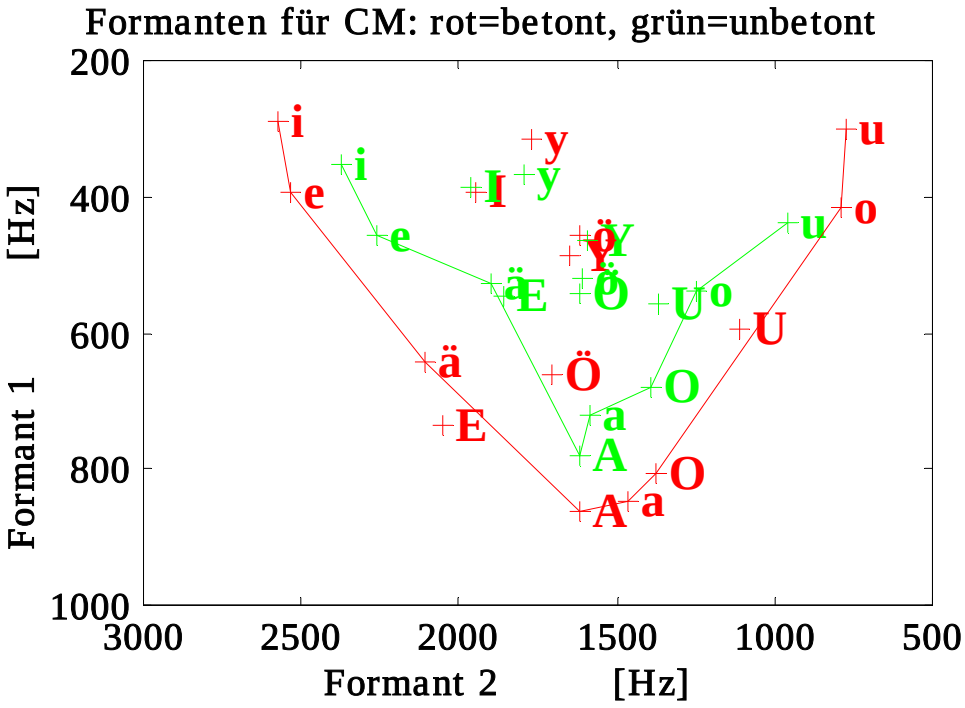
Aussagen dieser Art basieren auf den Gehörseindrücken des transkribierenden Phonetikers. Dabei wird der wahrgenommene Klangunterschied mit artikulatorischen Begriffen wiedergegeben, wie vorne – hinten, hoch – tief, geschlossen – offen.
Die Frage, die sich nun stellt, ist, auf welchen akustischen und artikulatorischen Korrelaten basiert diese Art der phonetischen Repräsentation? D.h. welche Beziehungen bestehen zwischen den Messungen der Realisierungen dieser Symbole und ihrer Lage im Vokaltrapez?

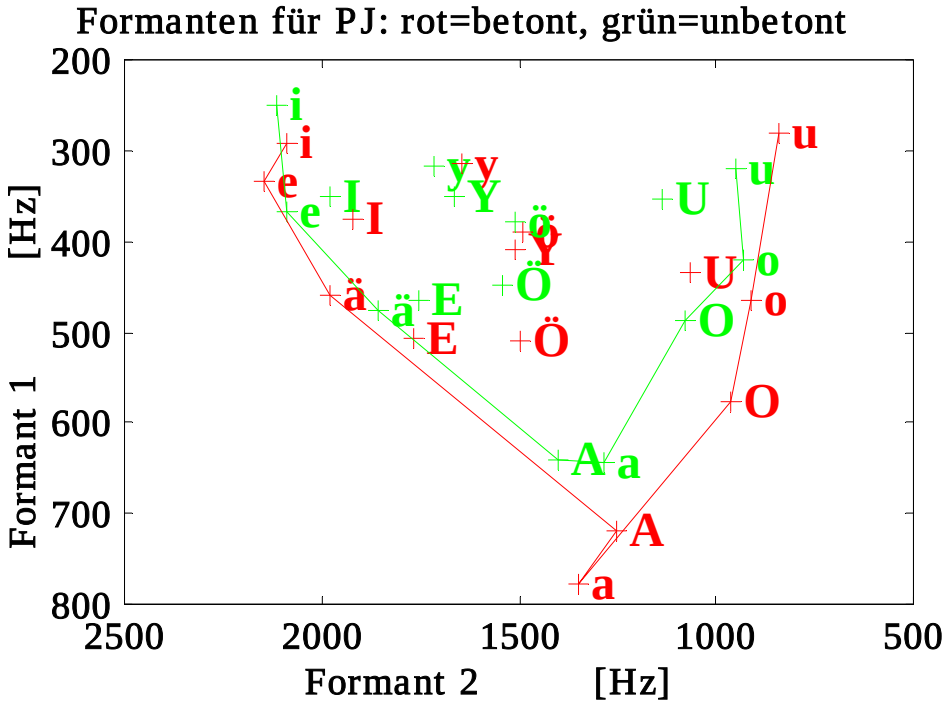
Dabei ist die Lage der Symbole für die einzelsprachliche Repräsentation wiederum eine Abstraktion von den Einzelsprechern.



Beispiel: Formantkarte von zwei Sprechern PJ (männlich) und CM (weiblich), gemittelt über jeweils sechs Wiederholungen
Die Werte der ersten beiden Formanten sind entscheidend für die Klangeigenschaften der Vokale.
Die Kleinbuchstaben entsprechen den gespannten, die Großbuchstaben den ungespannten Vokalen.
Wie die Abbildung zeigt, unterscheiden sich die beiden Sprecher nicht nur in der Größe des Raumes, den ihre Vokale aufspannen, sondern auch in der relativen Verteilung zueinander. Vergleiche z.B. die Lage von gespanntem und ungespanntem /o/ zueinander.

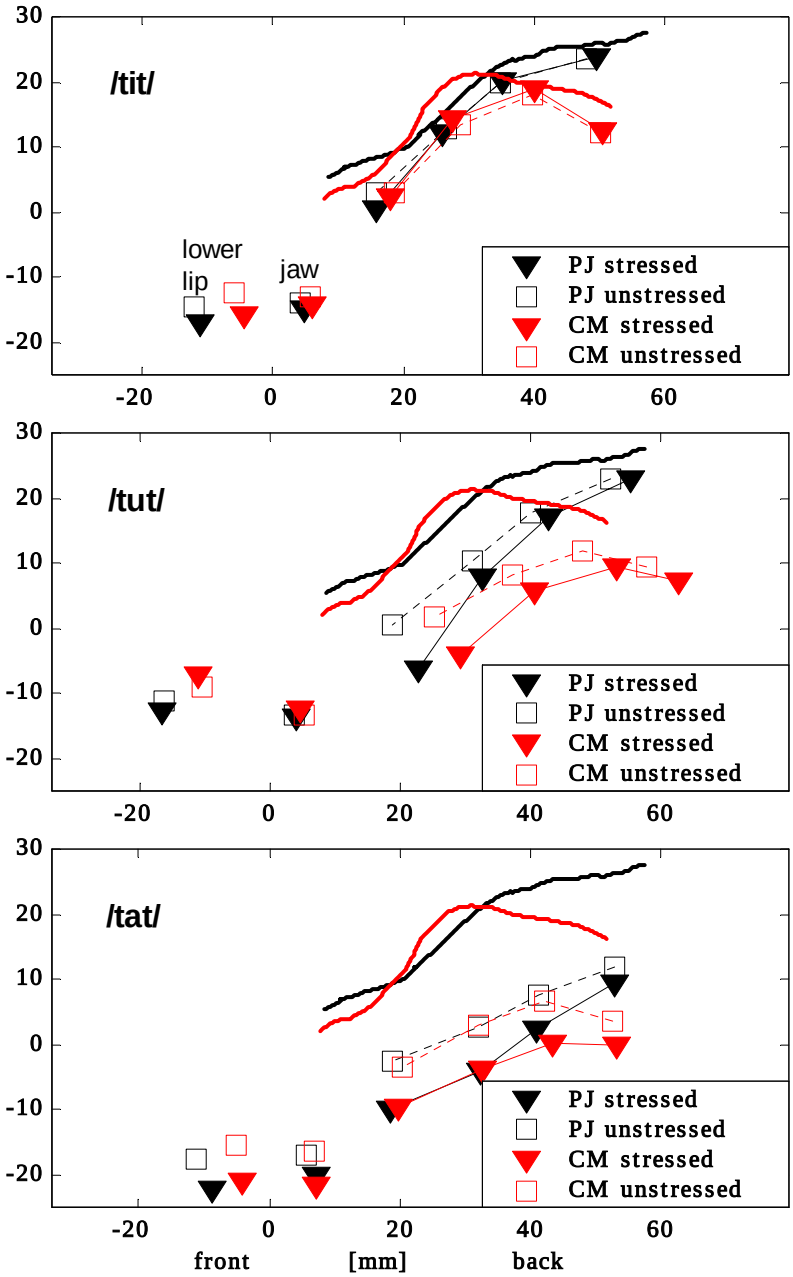
Neben den sprecherspezifischen Unterschieden werden die akustischen und artikulatorischen Eigenschaften der Vokale auch durch linguistische und paralinguistische Faktoren beeinflußt. So werden z.B. Vokale in schnell gesprochenen Äußerungen zentralisierter, d.h. näher am Schwa, realisiert als in normalem Tempo. Ebenso ergibt sich ein Unterschied zwischen betonten und unbetonten Vokalen: auch hier spricht man bei den unbetonten Vokalen von einer Zentralisierung bzw. von Target undershoot.





Oberes Bild: Zentralisierung der unbetonten Vokale bei Sprecherin CM, Formantwerte der betonten Vokale in rot, Formantwerte der unbetonten Vokale in grün. Es zeigt sich deutlich, daß die unbetonten Vokale näher zusammen-rücken.
Vergleich mit Sprecher PJ: weniger starke Zentralisierung, z.T. Höher-verlagerung.
➔ Sprecherspezifische Unterschiede in der Realisierung des Akzents.

Artikulation:



Gaumenkontur:
/tit/
Beide Sprecher produzieren das /i/ mit einer engen Konstriktion entlang des harten Gaumens. Deshalb weist Sprecher PJ wiederum eine sehr flache Zungenkonfiguration auf und Sprecherin CM eine sehr gewölbte. Nach IPA ist für die Klassifikation der Vokale der höchste Zungenpunkt ausschlaggebend. Danach hätte also Sprecher PJ ein höheres /i/ als Sprecherin CM, was nicht sinnvoll ist, da diese Kategorisierung rein von anatomischen Strukturen abhängen würde.

Wie sich beim /i/ deutlich zeigt, hat die Gaumenkontur einen starken Einfluß auf die Zungenform. Sprecher PJ sehr flache Gaumenkontur (schwarz), Sprecherin CM extrem gewölbter Gaumen.
Abgesehen von der Unterlippe für beide Sprecher keine wesentlichen Unterschiede zwischen betont und unbetont.
/tut/
Auch hier ist die Zungenkonfiguration von Sprecherin CM gewölbt und die von Sprecher PJ flach.
Akzentunterschied: Zungenblatt ist für beide Sprecher angehoben und leicht nach vorne verlagert bei unbetonten Silben, d.h. näher an der apikalen Konsonantartikulation.
Ebenso /a/ aber hier auch Unterschied in der Unterkieferstellung.

Wie ich bisher gezeigt habe, steckt der Teufel wie immer in der Experimental-phonetik im Detail, d.h. die Sprecher unterscheiden sich nicht nur in den akustischen und artikulatorischen Zielkonfigurationen sondern auch in ihren Strategien, um linguistische Unterschiede wie Akzent zu markieren.
D.h. um ein allgemeineres Modell der Vokalproduktion im Deutschen zu erhalten, muß erst eine Sprechernormalisierung durchgeführt werden.

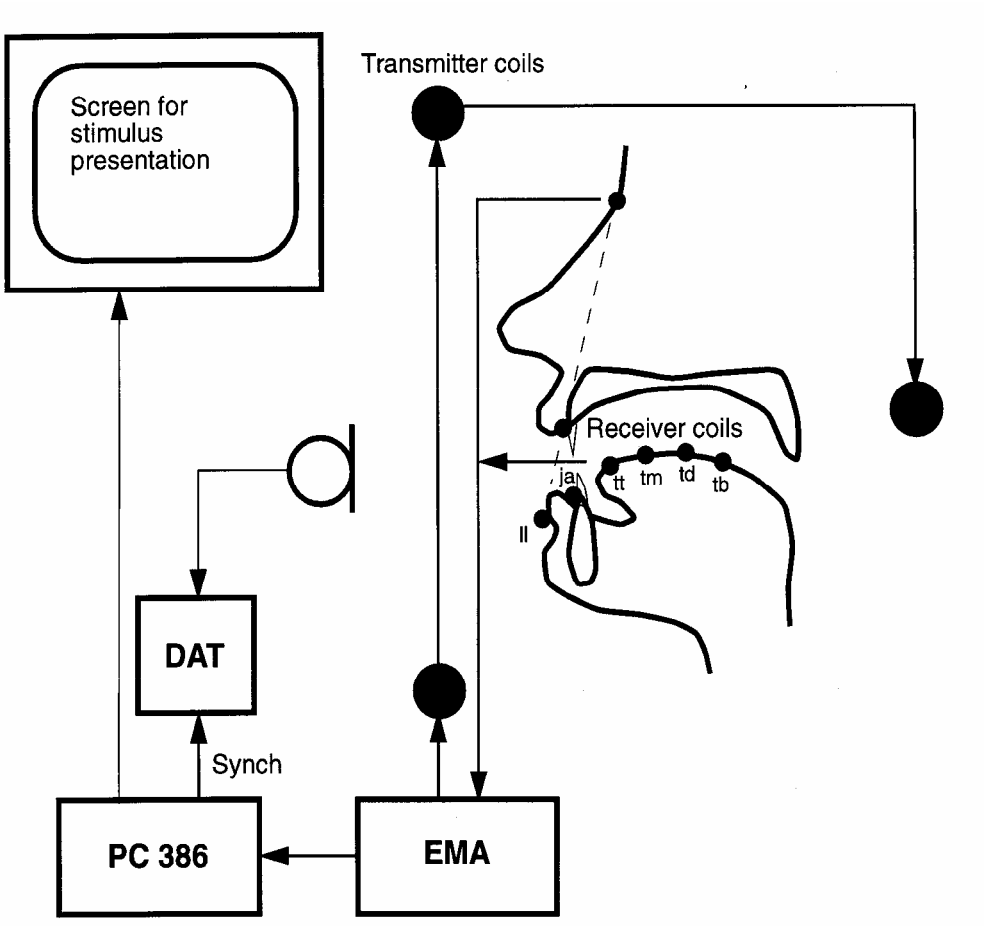
Zusammenfassung:
Ziele dieser Untersuchung

- Ermittlung sprecherunabhängiger Strategien zur Realisierung des Wortakzents: akustisch und artikulatorisch (z.B. was bedeutet Target undershoot?)
- Zusammenhang zwischen Akustik und Artikulation
- Werden die ermittelten Unterschiede in der Vokalqualität vom Hörer wahrgenommen?

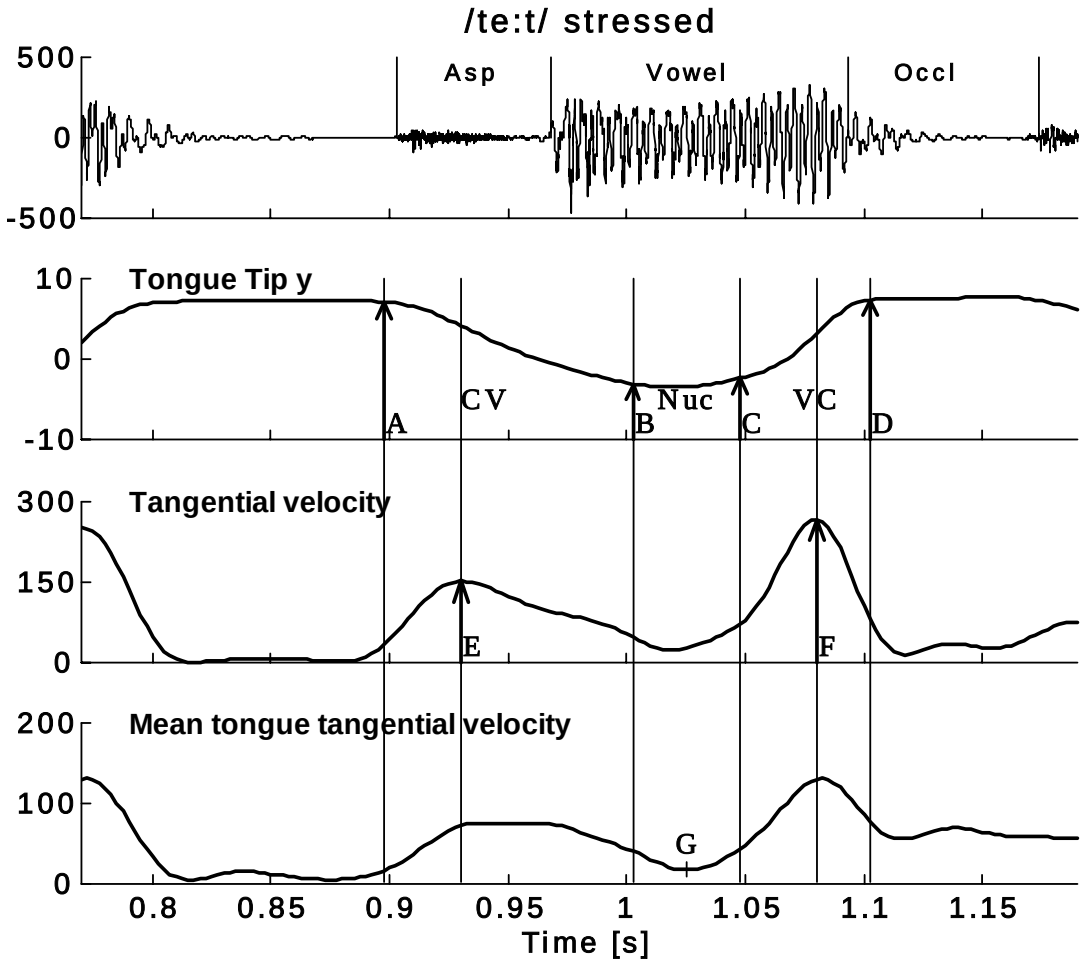
Bevor ich auf die von uns gewählte Methode zur Sprechernormalisierung eingehe, beschreibe ich erst mal das aufgenommene Datenmaterial.

II Datenerfassung

- ✓ Sieben Sprecher des Deutschen
- ✓ CVC Sequenzen mit symmetrischem Konsonantkontext /t/
  Vokale: /iː, ɪ, yː, ʏ, eː, ɛ, ɛː, øː, œ, aː, a, oː, ɔ, uː, ʊ/
- ✓ Zielwörter: /ˈtVtɐ/ und /tVˈtaːl/, z.B.. *tater, tatal, teeter, teetal, tutter, tuttal*
- ✓ Rahmensatz: “Ich habe *tieter,* nicht *tietal* gesagt.”
- ✓ Sechs bzw. zehn Wiederholungen
- ✓ Elektromagnetische Artikulographie EMMA (AG100, Carstens Medizin-elektronik) und DAT Recorder
- ✓ Sensorplazierung: 4 Sensoren auf der Zunge, je ein Sensor auf der Unterlippe und den unteren Schneidezähnen (Referenzsensoren: obere Schneidezähne und Nasenwurzel)



Messungen



Wichtig: Zeitpunkt G entspricht Umkehrpunkt der Zungentrajektorie bei möglichst vielen Zungensensoren.
Messungen der ersten drei Formanten mittels Signalize zum Zeitpunkt G

III Sprechernormalisierung: Procrustes Methode

Verschiedene Arten von Sprechernormalisierungen:

- Z-Transformation: Zentrierung auf den Mittelwert und Division durch die Standardabweichung
  Lobanov-Normalisierung
  Für akustische Sprechernormalisierung sehr erfolgreich
  Nicht auf artikulatorische Daten anwendbar
- Perzeptive Reskalierung: z.B. Nearey
  Die Formanten werden in auditiv relevante Skalen transformiert
  Nachteil: nicht auf artikulatorische Daten anwendbar
- Methoden aus der artikulatorischen Synthese: Perrier
  Formanten werden mit einem Ansatzrohrmodell synthetisiert und anschließend die Differenzen zu gemessenen Formanten berechnet, d.h. sprecherspezifische Formanten werden auf einen Normsprecher angepaßt.
  Für unseren Zweck nicht geeignet
- Faktorenanalyse: PARAFAC
- Geometrische Normalisierung: Procrustes Methoden

Kriterien für Auswahl der Methode:

- gleichzeitig auf akustische (2D) und artikulatorische (12D) Daten anwendbar
- Reduktion der sprecherspezifischen Variabilität
- Trennung der Kategorien

Procrustes Methode

Bösewicht aus der griechischen Sage, der als Räuber in Attica sein Unwesen trieb. Er hatte zwei Eisenbetten, in die er seine Opfer legte. Waren sie kürzer als das Bett, wurden sie gestreckt. Längeren Opfern wurden die Beine abgehackt.

Encyclopaedia Britannica:

„In either event the victim died .“

„The bed of Procrustes has become proverbial for arbitrarily – and perhaps ruthlessly – forcing someone or something to fit into an unnatural scheme or pattern.“

SEHR VIELVERSPRECHEND!!!!

Die einfachste Form der Procrustes Analyse besteht aus drei Schritten:
1. Zentrierung
2. Skalierung mit einem konstanten Faktor
3. Rotation

➔ Konsensus-Objekt

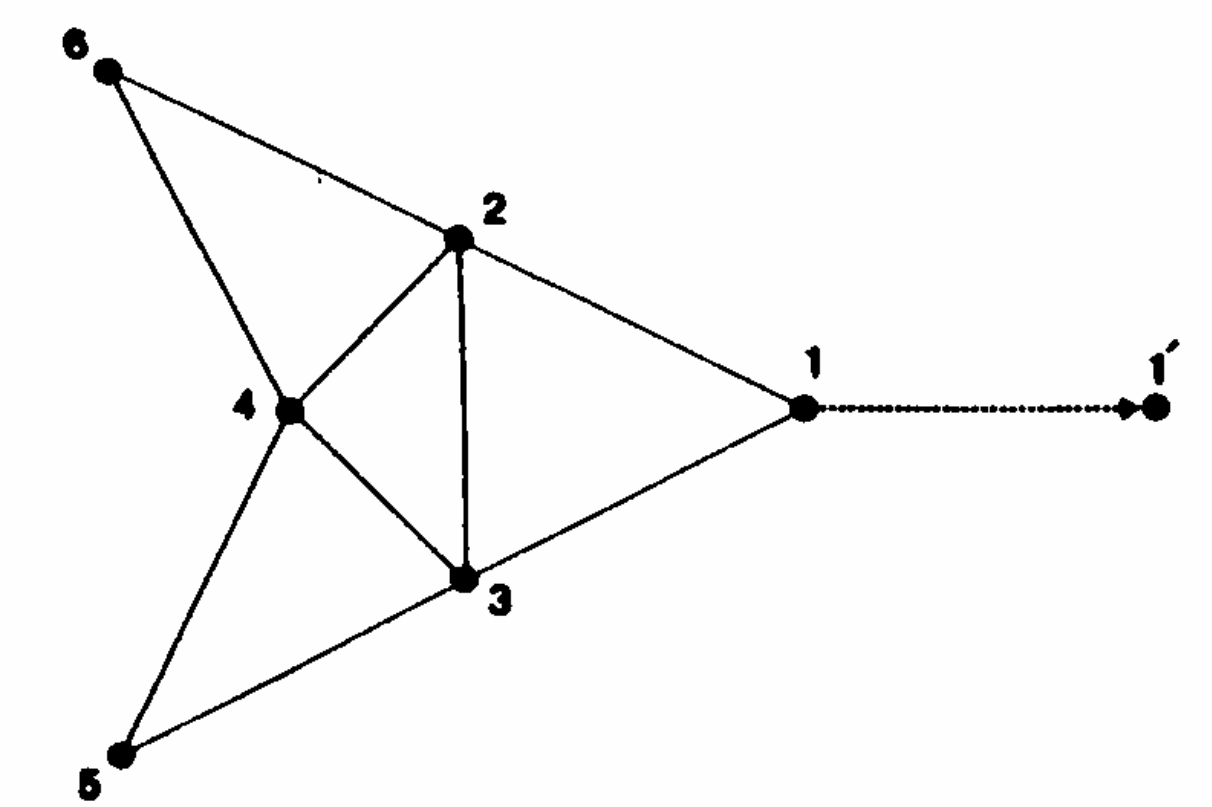


FIG. 1. Diagram showing configuration of landmarks for two artificial "organisms" differing only in the location of landmark number 1.

Aus James Rohlf und Dennis Slice (1990). Journal of Systematic Zoology.
6 x/y Koordinaten von 20 artifiziellen Organismen
Einfachste Form des Algorithmus basiert auf Least-Squares Fit, d.h. die Summe der Abweichungen der Landmarks pro Objekt wird iterativ minimiert.
Diese Methode erweist sich zwar als unstabil gegenüber einzelnen Outliers, reduziert aber die Variabilität am stärksten.

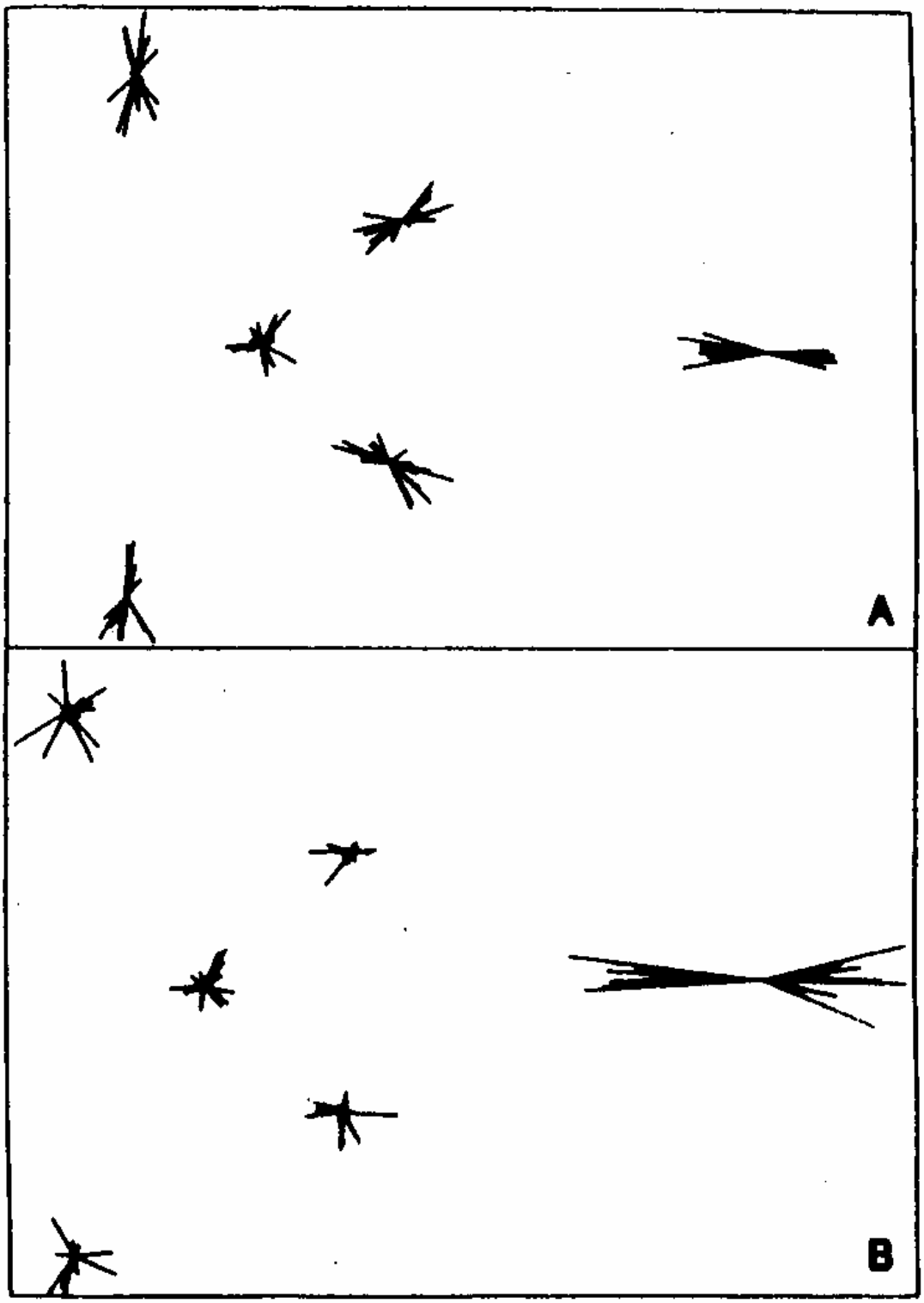


FIG. 2. Results of (A) generalized Procrustes and (B) resistant-fit analyses for a sample of 20 from the population defined in Figure 1. Vectors emanating from the centroid of each landmark show the deviation of each landmark from the consensus configuration.

Die transformierten Daten bestehen aus den Koordinaten des Konsensusobjekts und aus den organismus- bzw. sprecherspezifischen neuen Koordinaten (siehe Abbildung 2A).
Für die einfache, d.h. orthogonale Variante dieses Algorithmus gilt, daß sich die Winkel der neuen Objekte nicht von denen der Rohdaten unterscheiden.
D.h. die einzelnen Objekte werden in alle Richtungen proportional gestreckt oder gestaucht.
Es gibt jedoch auch sogenannte affine Transformationen, bei denen die Objekte in nur eine Richtung gestreckt werden können (siehe Abbildung unten). Hier verändern sich die Winkel der Kanten.

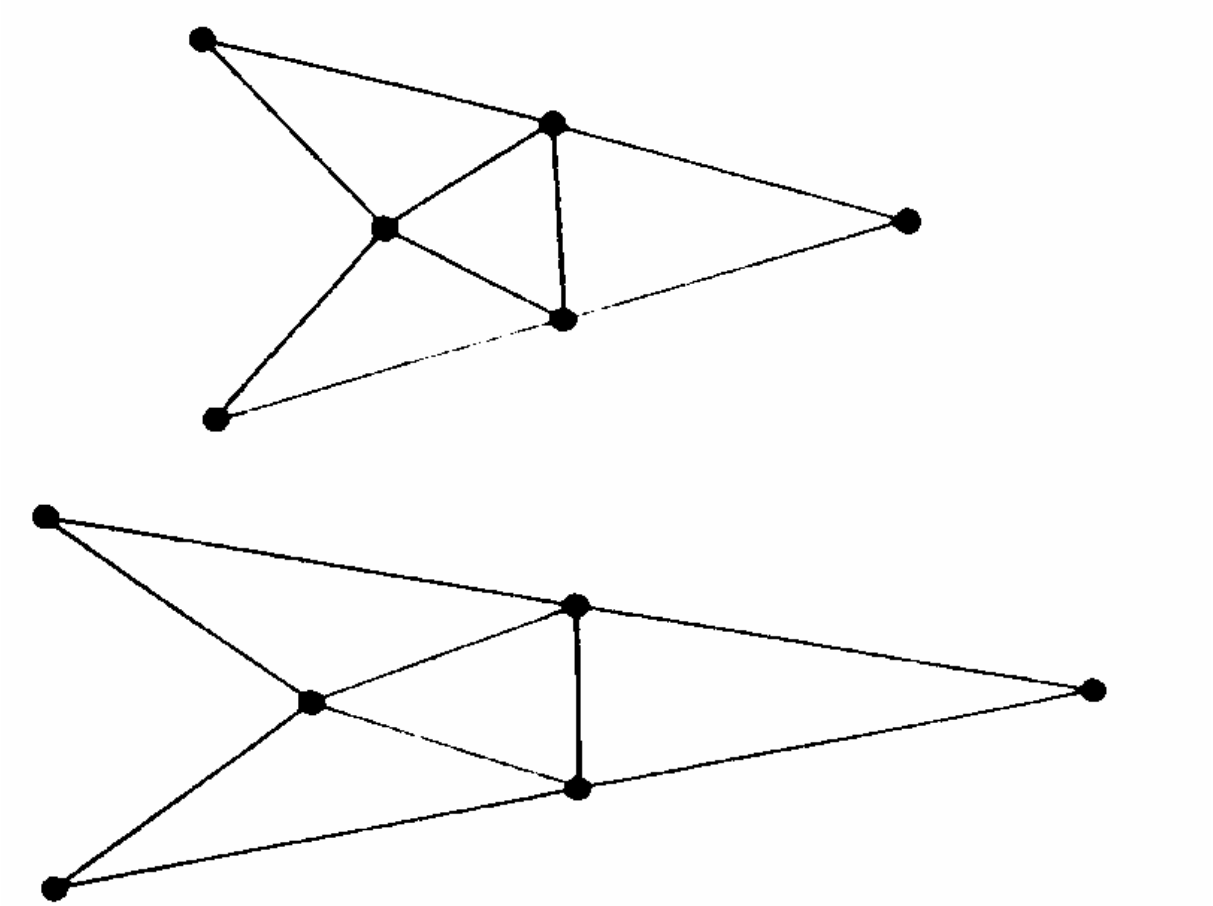


FIG. 3. Affine deformation. The configurations are identical except that the lower one was uniformly stretched in the horizontal direction.

Nun ein Beispiel aus der systematischen Zoologie

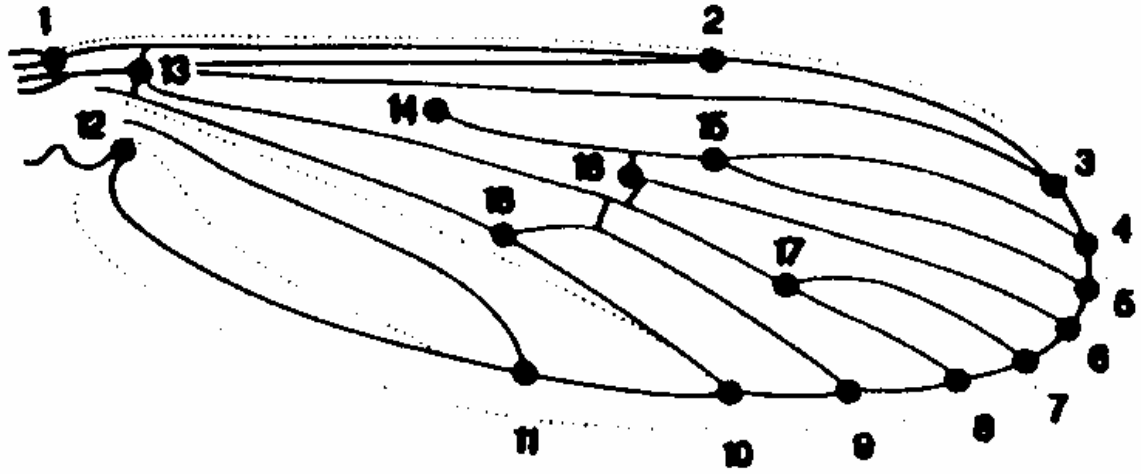


FIG. 7. Diagram of the wing venation of a mosquito (Culicidae). Dotted line around outside shows usual extent of fringe.

127 Spezies der nordamerikanischen Moskitos

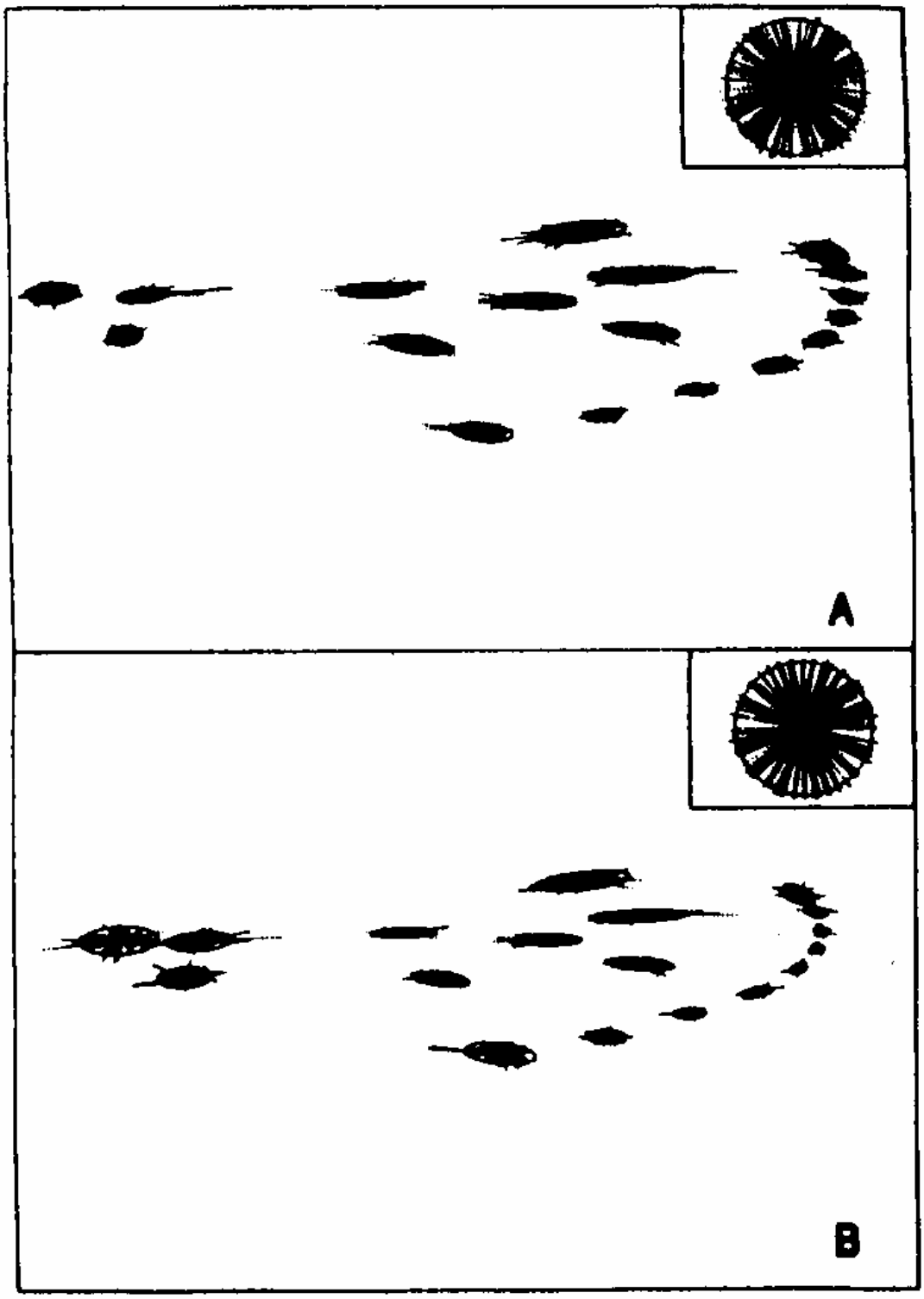
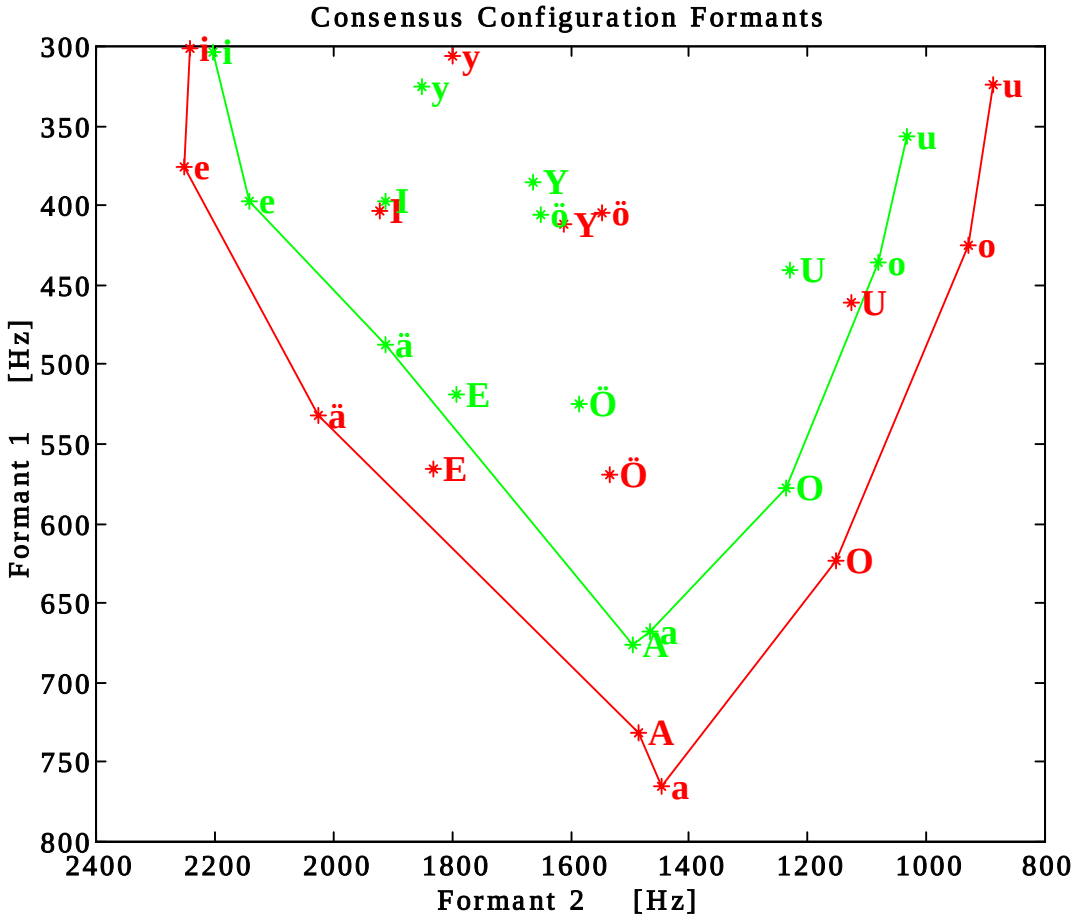


FIG. 9. Results of generalized affine (A) least-squares and (B) resistant-fit analysis of the wings of 127 species of North American mosquitoes. Upper right corner in each shows the inverse of the scaled minor strain cross for each species.

IV Ergebnisse Akustik

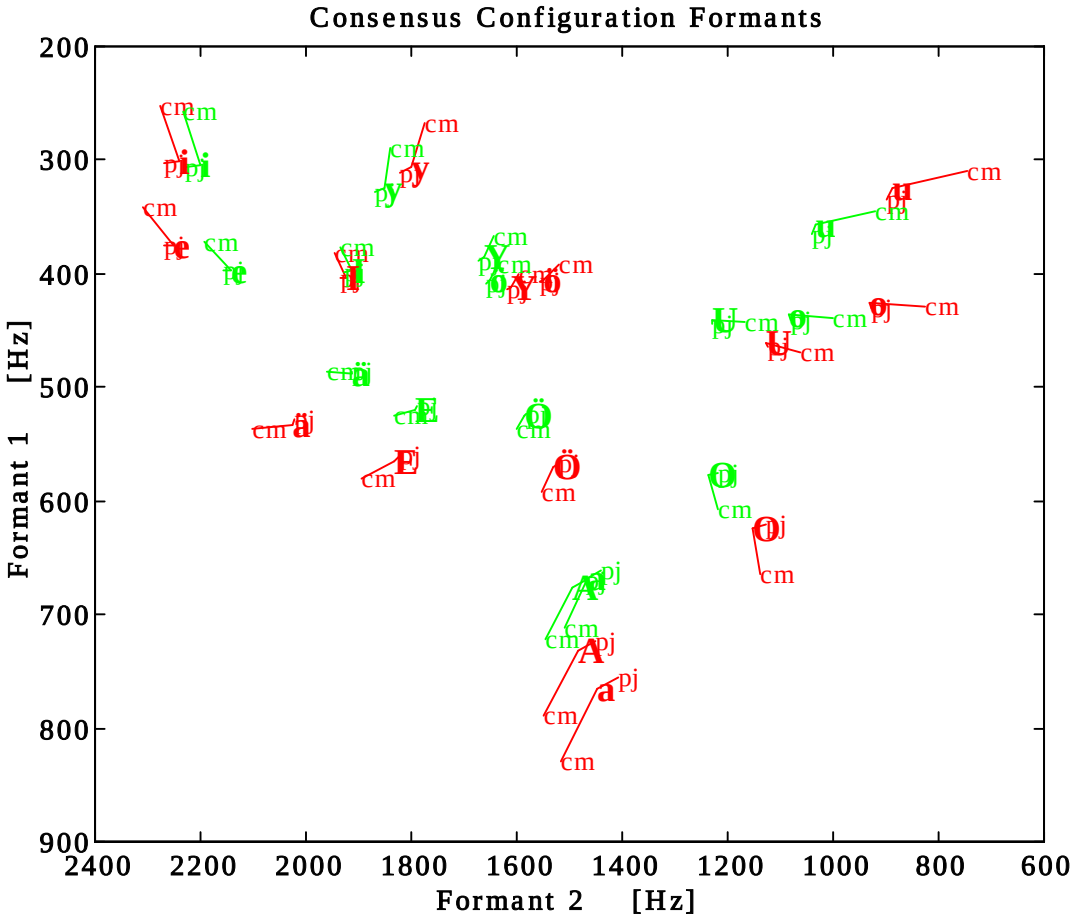
Eingabedaten: 30 Formantpaare von sieben Sprechern

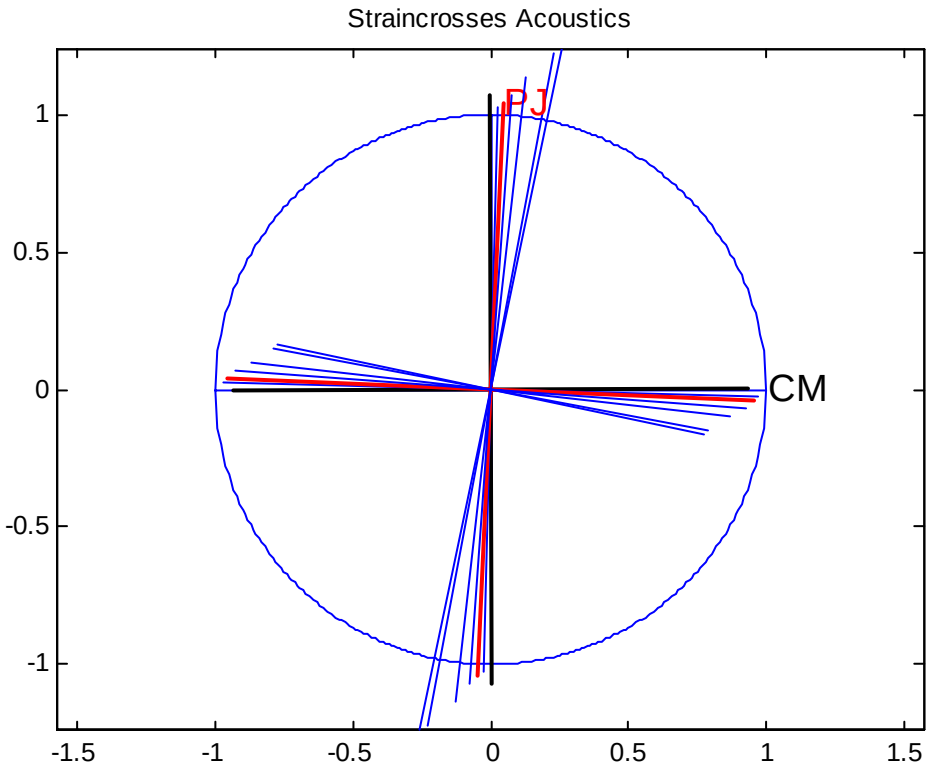


Konsensusobjekt: Rot betont, grün unbetont.

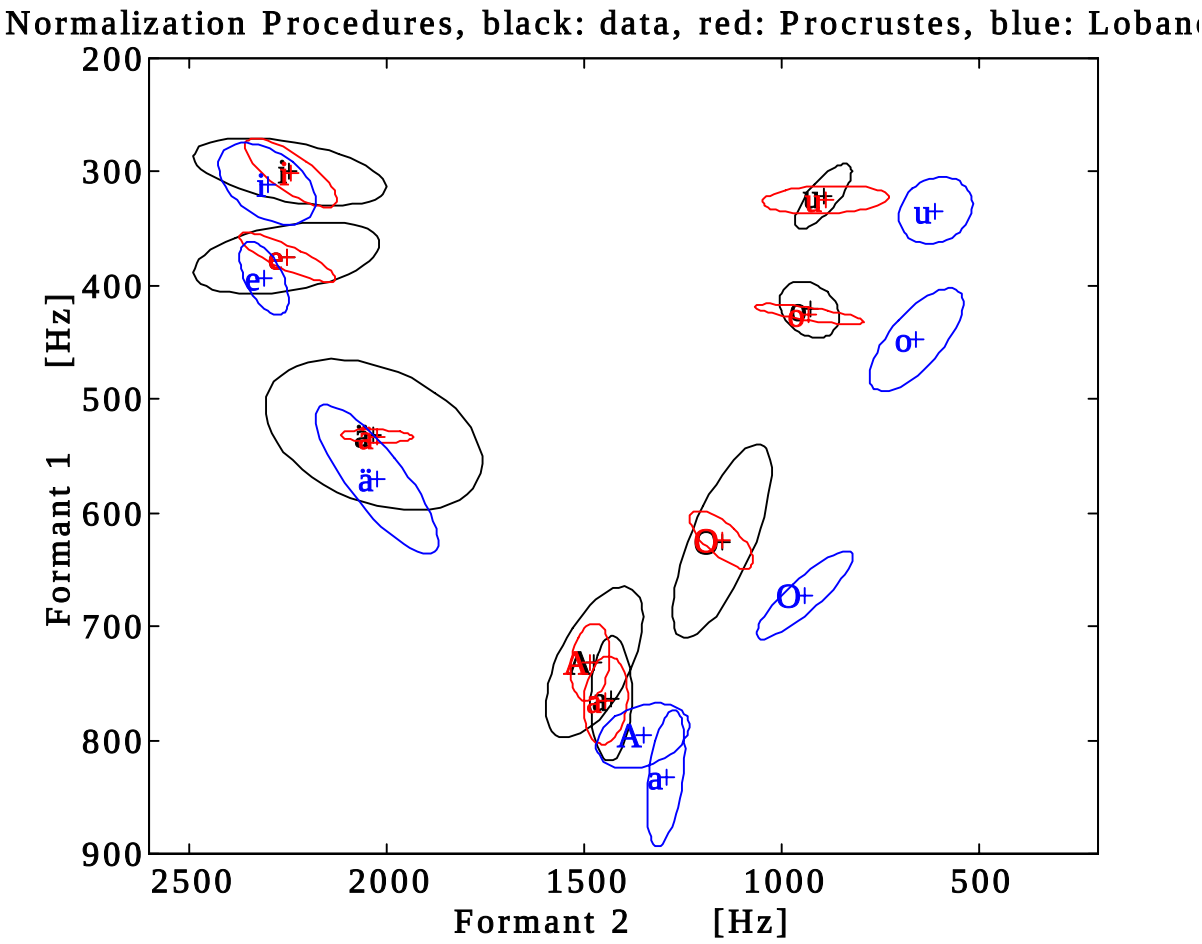
➔ Target undershoot bei Deakzentuierung entspricht stärkerer Koartikulation.

Modellierte Sprecherunterschiede





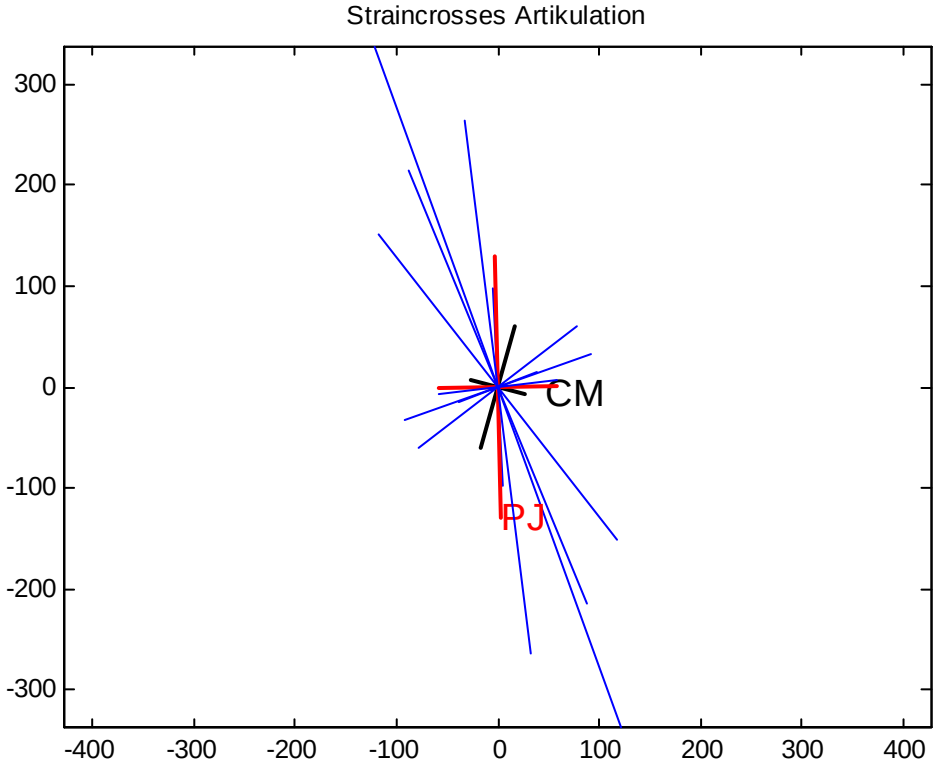
Vergleich Rohdaten – Lobanov - Procustes



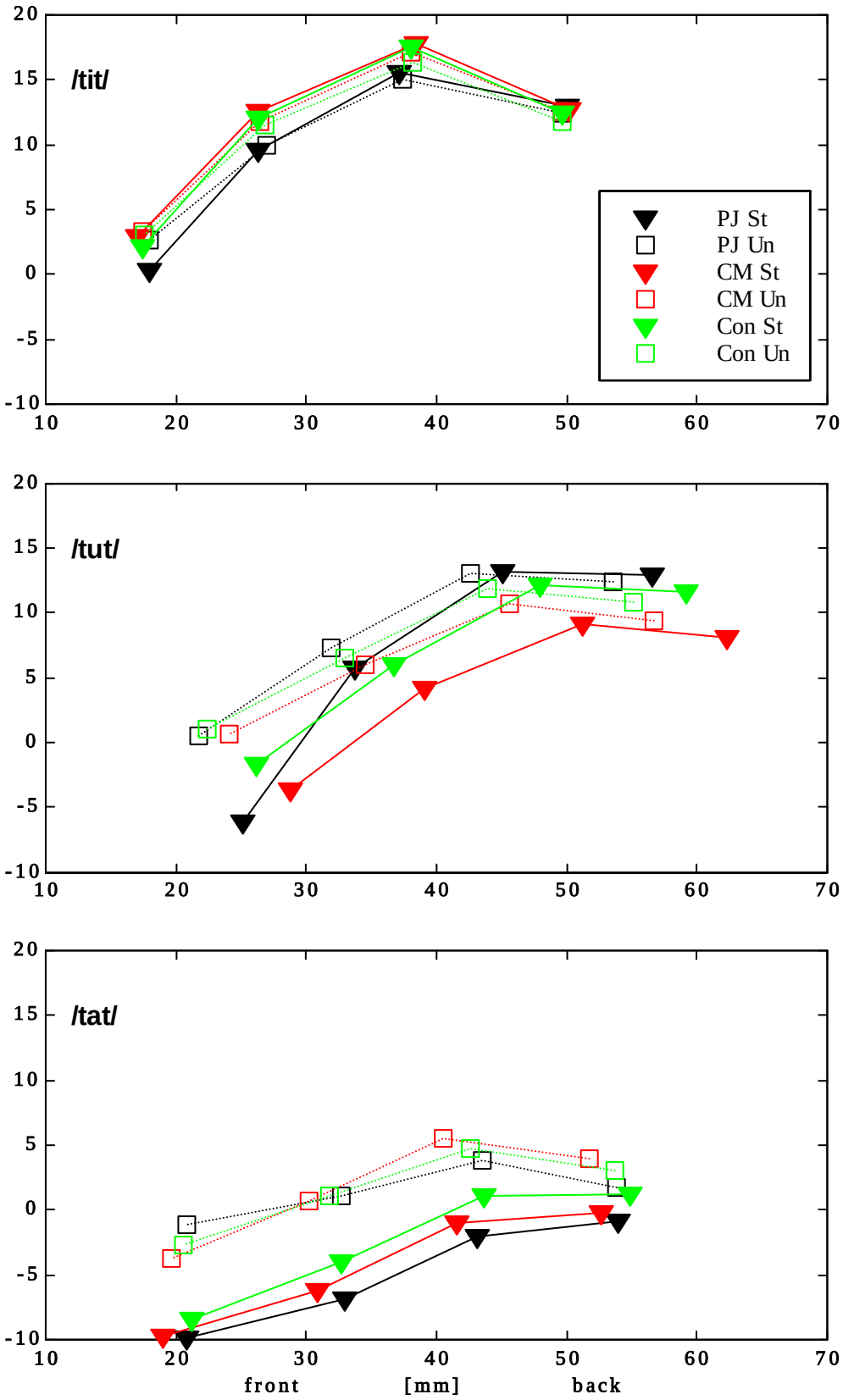
Procustes: Reduktion auf 15%
Lobanov: Reduktion auf 47%

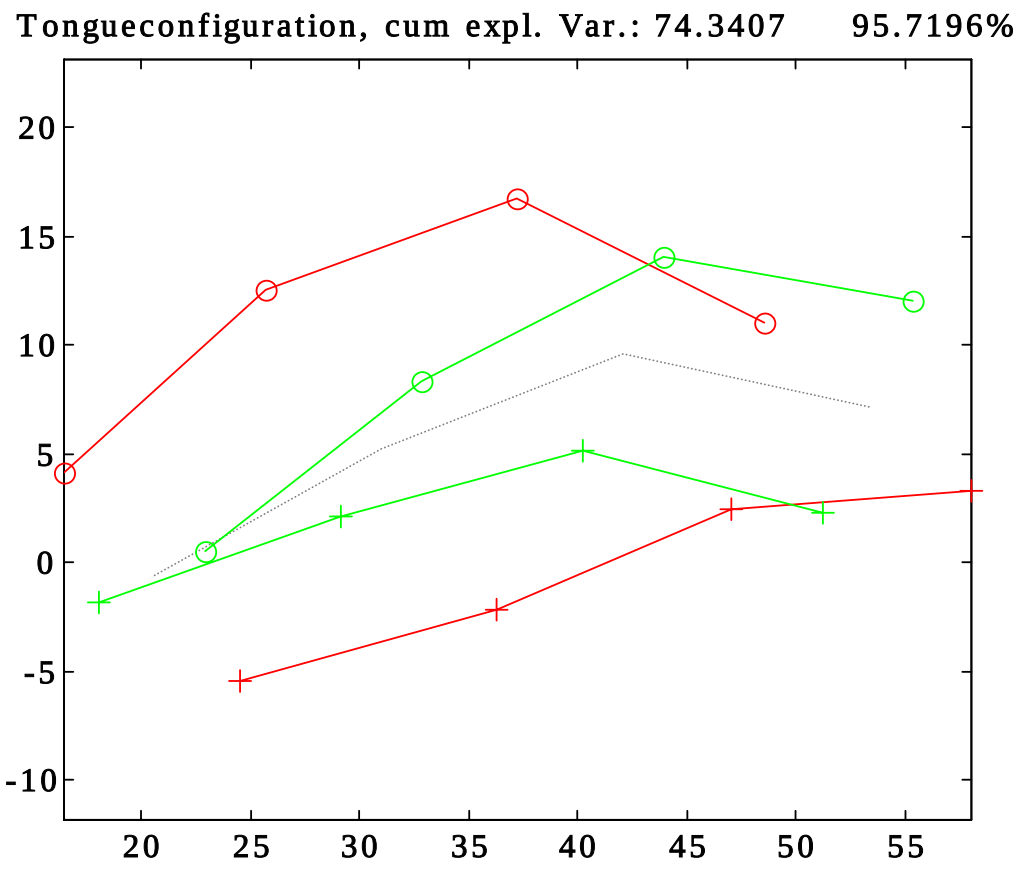
V Ergebnisse zur Artikulation

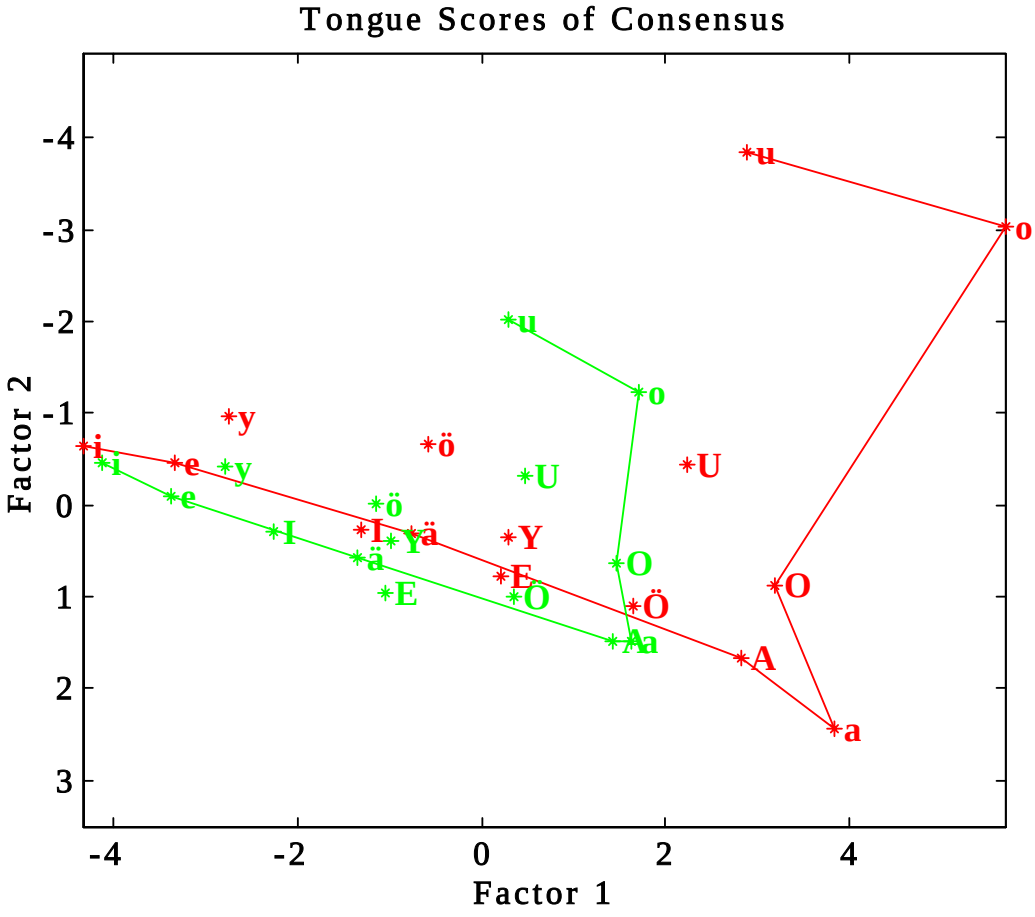
Staincrosses



Zungenkonfigurationen



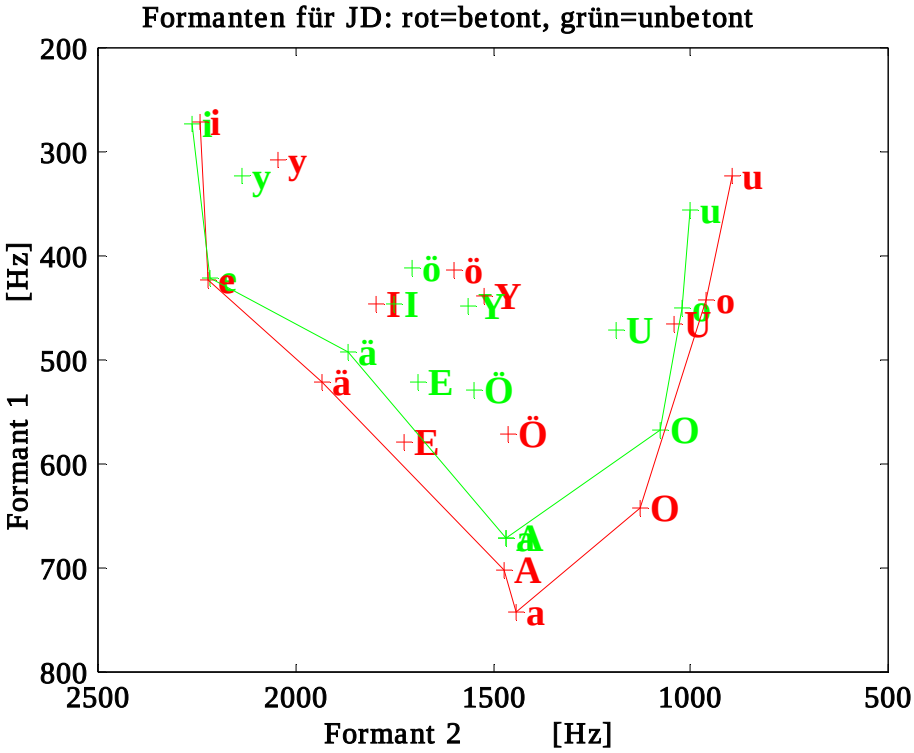

Tongueconfiguration, cum expl. Var.: 74.3407 95.7196%
20
15
10
5
0
-5
-10
20
25
30
35
40
45
50
55


Tongue Scores of Consensus
Factor 2
Factor 1
-4
-3
-2
-1
0
1
2
3
-4
-2
0
2
4
u
o
y
ö
U
i
e
I
ä
Y
E
Ö
O
A
a

VI Perzeptionstest

Es sollte keinen Einfluß der Dauer auf die Kategorisierung geben, d.h. uns interessierte hauptsächlich der Einfluß des Klangunterschiedes. Deshalb wurden die Dauern der einzelnen Vokale auf den Mittelwert der betonten gespannten und betonten ungespannten Vokale normiert. D.h. die Dauer des Stimulus für das gespannte /i/ wurde gekürzt, während die Dauern für die unbetonten und ungespannten <i>‘s gelängt wurden. Die Mittelwerte wurden pro Kategorie berechnet, d.h. einzeln für <i, e, ü> usw. Durchschnittlich waren die Stimuli ca. 50 ms lang. Entnommen von Sprecher JD, der nur sehr geringe akzentbedingte Unterschiede zeigte.
Für die Ä bzw. E Laute ergeben sich deshalb im ungespannten Fall zwei Stimulusgruppen: Längung bzw. Kürzung auf die Dauer vom Mittelwert von /e/ betont und /E/ betont (Gruppe 1);
Längung bzw. Kürzung auf die Dauer vom Mittelwert von /ä/ betont und /E/ betont (Gruppe 1); /ä/ länger

17 Hörer, 5 Blocks, 32 Stimuli.



Den Hörern wurden diese Stimuli einzeln in randomisierter Reihenfolge vorgespielt. Sie konnten sich für eine der folgenden orthographischen Repräsentationen entscheiden:
s. Tabelle
Antworthäufigkeiten

STI	Ant.	tiet	titt	tüt	tütt	tet	tett	tät	tätt	töt	tött	tat	tatt	tot	tott	tut	tutt	NE	Ant..	STI
i	S	**72.9**	*23.5*	1.2								1.2						1.2	i	S
	U	**63.5**	*31.8*	1.2	1.2	1.2												1.2		U
y	S			**68.2**	*27.1*					1.2							1.2	2.4	y	S
	U	2.4	1.2	**75.3**	*17.6*					1.2			2.4							U
eː	S		*32.9*			**42.4**	16.5	5.9						1.2				1.2	e	S
	U	1.2	*34.1*		1.2	**44.7**	9.4	5.9	2.4	1.2										U
ɛː	S					8.2	**44.7**	*28.2*	15.3						1.2			2.4	ɛː ɛ	S
	U					18.8	*23.5*	**37.6**	15.3									4.7		U
ø	S			8.2	*18.8*					**63.5**	8.2							1.2	ø	S
	U				*23.5*					**63.5**	11.8							1.2		U
ɑ	S											**81.2**	*16.5*		1.2			1.2	ɑ	S
	U					1.2					1.2	**77.6**	*14.1*	1.2	4.7					U
o	S													**68.2**		*16.5*	14.1	1.2	o	S
	U													**43.5**	14.1	3.5	*37.6*	1.2		U
u	S	1.2										1.2				**74.1**	*22.4*	1.2	u	S
	U	5.9	3.5	14.1	15.3	1.2					1.2					**35.3**	*23.5*			U
Voc.		tiet	titt	tüt	tütt	tet	tett	tät	tätt	töt	tött	tat	tatt	tot	tott	tut	tutt	NE	Voc.	

Gespannte Vokale: kaum Unterschiede für die Vorderzungenvokale,
Hinterzungenvokale: unbetontes /o/ wird häufig als /ʊ/ klassifiziert; unbetontes /u/ auch als /ʊ/ und interessanter weise als gespanntes oder ungespanntes /y/. D.h. die Zentralisierung der unbetonten Hinterzungenvokale wird durchaus wahrgenommen, wobei die Rundung beibehalten wird.

STI	Ant.	tiet	titt	tüt	tütt	tet	tett	tät	tätt	töt	tött	tat	tatt	tot	tott	tut	tutt	NE	Ant..	STI
ɪ	S		**55.3**	1.2	*24.7*	1.2	7.1		2.4	3.5	2.4						1.2	1.2	ɪ	S
	U		**48.2**	1.2	7.1	8.2	*20.0*		3.5		8.2						1.2	2.4		U
ʏ	S		1.2	2.4	**85.9**		1.2			3.5	*4.7*		1.2						ʏ	S
	U		11.8	3.5	**48.2**	1.2	3.5	1.2		10.6	*18.8*							1.2		U
eː ɛ	S		1.5		1.5	2.9	**38.2**	5.9	*26.5*	1.5	17.6				1.5			2.9	eː ɛ	S
	U	1.2				1.2	*22.4*	2.4	17.6	2.4	**45.9**		1.2		3.5		2.4			U
ɛː ɛ	S					2.4	**38.8**	4.7	*25.9*	1.2	16.5		3.5		2.4		1.2	3.5	ɛː ɛ	S
	U		2.4		1.2	5.9	*28.2*	4.7	10.6	3.5	**36.5**		1.2		4.7		1.2			U
œ	S						*11.8*	3.5	4.7	4.7	**61.2**		1.2		10.6		1.2	1.2	œ	S
	U					5.9	8.2		5.9	5.9	**34.1**	1.2		1.2	*30.6*		5.9	1.2		U
a	S			1.2		1.2			1.2		1.2	**78.8**	*14.1*		1.2			1.2	a	S
	U		2.4				1.2		2.4		3.5	**54.1**	*30.6*		3.5			2.4		U
ɔ	S		1.2								1.2	2.4		3.5	**89.4**	1.2		1.2	ɔ	S
	U					2.4	3.5		2.4	2.4	4.7	1.2	4.7	*9.4*	**63.5**		2.4	3.5		U
ʊ	S		1.2												*9.4*	2.4	**85.9**	1.2	ʊ	S
	U		3.5		2.4		2.4	1.2			9.4	1.2		5.9	**42.4**	1.2	*28.2*	2.4		U
		tiet	titt	tüt	tütt	tet	tett	tät	tätt	töt	tött	tat	tatt	tot	tott	tut	tutt	NE		

Ungespannte Vokale: Verwechslungsmatrix sieht gleich auf den ersten Blick wesentlich chaotischer aus.
Gelängte ungespannte Vokale werden insgesamt sehr selten mit gespannten Vokalen verwechselt, und die Lippenrundung wird häufiger verwechselt als bei den gespannten Vokalen: s. unbetontes /eɛ/
Wesentlich stärkerer Effekt der Deakzentuierung: unbetonte ungespannte Vokale werden sehr häufig mit anderen ungespannten Vokalen verwechselt s. z.B. /ʊ/
Hier spielt sicherlich eine Rolle, daß als Antworten keine Schwas möglich waren und durch die zusätzliche Zentralisierung durch die Deakzentuierung große Unsicherheit bei den
Interessant ist der Vokal /a/, der sich in seiner Qualität nur geringfügig unterscheidet und auch gekürzt fast immer als langes a klassifiziert wird.
Ausblick: Synthetisierung der Vokale nach akustischem Procustes Konsensusobjekt

VII Zusammenfassung

Akzentunterschied:
Target undershoot bedeutet hier nicht in Richtung des Zentralvokals sondern in Richtung des konsonantischen Kontexts, d.h. unbetonte Silben werden stärker koartikuliert. Bei /t/-Kontext werden die Hinterzungenvokale nach vorne und die tiefen Vokale nach oben verlagert.
Zeigt sich in allen drei Bereichen Akustik, Artikulation und tendenziell auch in der Perzeption

Sprechernormalisierung:
Procustes Methoden liefern sinnvolle und interpretierbare Ergebnisse, sowohl für Akustik als auch für Artikulation. (Kiefer-Lippe fehlt noch)

Zusammenhang Artikulation – Akustik:
Formantkarten verschiedener Sprecher können durch relativ einfache lineare Transformationen aufeinander bezogen werden, während für die Artikulation massive nicht-lineare Streckungen und Dehnungen nötig sind. D.h. der Sprecher hat wahrscheinlich im Laufe des Spracherwerbs gelernt, die artikulatorischen Targets den akustischen anzupassen.

Zusammenhang Meßdaten – Vokaltrapez:
Relativ nah an der Akustik
Unterschiede zu Perzeption basieren zum Teil darauf, daß zwar Vokalhöhe und Zungenlage kontinuierlich veränderbar sind, die Lippenrundung jedoch nicht.
➔ Darstellung als Würfel