# MARKOV-ENTROPIE BASIERTE AUSWAHL GEEIGNETER SPRACHSEGMENTE FÜR KORPUSBASIERTE SPRACHSYNTHESE-SYSTEME

*Christian Weiss*

*Institut für Kommunikationsforschung und Phonetik, Universität Bonn*
*cwe@ikp.uni-bonn.des*

**Abstract:** Konkatenative Text-to-Speech Systeme, welche auf der Grundlage des „variable size non-uniform unit-selection“ Algorithmus aufgebaut sind, haben die Aufgabe adäquate Sprachbausteine auszuwählen, die die qualitativen und quantitativen Zielvorgaben maximal approximieren. Hieraus ergibt sich die Problemstellung ein geeignetes Sprachdatenkorpus zu erstellen, welches maximale Abdeckung aufweist und minimal redundant ist. In bestehenden Systemen wird die Segmentauswahl zur Laufzeit mittels einer Einheitenkosten-Funktion und Übergangskosten-Funktion berechnet.

Als Alternative zur Einheiten-Kostenfunktion wird in diesem Beitrag ein statistisches Verfahren vorgestellt, welches auf abhängigen Wahrscheinlichkeiten der jeweiligen parametrisierten Sprachsegmente beruht. In der Abfolge der zu synthetisierenden Äußerung wird aus diesen ein Markov-Graph der auszuwählenden Sprachsegmente aufgebaut. Die Auswahl eines geeigneten Sprachsegments wird mit Hilfe des Entropie-Maßes ermittelt. Ist dadurch die Auswahl eines bestimmten Sprachsegments nicht eindeutig identifizierbar, kommt die Mahalonobis-Distanz zwischen den spektralen Eigenschaften der Sprachbausteine abhängig vom Kontext zum Tragen. Das hier beschriebene Verfahren eignet sich sowohl zur Überprüfung der in Sprachdatenkorpora verwendeten Sprachsegmente als auch im Speziellen für eine Reduktion nicht verwendeter Sprachbausteine. Aus diesem Grund kann das Verfahren auch zur Korpuserstellung für die konkatenative Sprachsynthese eingesetzt werden, da nicht relevante Sprachbausteine eine geringe Auswahlwahrscheinlichkeit haben und somit aus der Datenbasis entfernt werden können.

## 1 Einleitung

Sprachsynthese-Systeme, im Besonderen Unit-Selection basierte TTS-Systeme, haben meist das Problem, für eine linguistisch motivierte Zieläußerung adäquate Sprachbausteine aus dem zugrunde liegenden Sprachdatenkorpus auszuwählen. Dabei sollen die linguistischen, quantitativen, prosodischen und spektralen Zielvorgaben durch die ausgewählten Sprachbausteine maximal angenähert werden. Bei Systemen die keinerlei Ressourcenbeschränkung unterliegen, kann eine Erhöhung des verwendeten Sprachdaten-Inventars eine Lösung bieten, um eine hohe Abdeckungswahrscheinlichkeit zu erreichen, bei der durch die Sprachsegmentauswahl das entsprechende lautsprachliche Segment nach Zielvorgaben ausgewählt werden kann und die prädizierten Vorgaben maximal approximiert werden. Eine hohe Abdeckung der einzelnen sprachlichen Phänomene kann durch diese Ausweitung des Sprachdatenkorpus allerdings nur beschränkt erreicht werden. Dies ist auf das bekannte LNRE - (large number of rare events) – Theorem [7] zurückzuführen. Will man zudem ein System, welches Ressourcen schonend ist, ist dieser Ansatz der Inventarvergrößerung nicht als sinnvoll zu betrachten. In bisherigen Arbeiten wurde die Problemstellung der geeigneten Sprachsegmentauswahl über eine mehrdimensionale Kostenfunktion gelöst. Dieser Ansatz geht zurück auf [1,4]. Hierbei wurden die kontextabhängigen Einheiten-Kosten sowie die Übergangskosten der jeweiligen potentiellen

Sprachbausteine berücksichtigt. Die Kostenfunktionen errechnen für die Sprachsegmente, welche in Betracht kommen, die Abstände zu den Zielvorgaben voraus. Es werden die Sprachsegmente ausgewählt, bei denen die Einheiten-Kosten und die Übergangskosten in der Summe minimal sind, also den geringsten Abstand zu den Zielvorgaben haben.

In dieser Arbeit wird eine Alternative zu dem bestehenden Ansatz vorgestellt. Die Einheiten-Kostenfunktion wird ersetzt durch ein Wahrscheinlichkeitsmodell, welches auf abhängigen Wahrscheinlichkeiten aufbaut. In der zeitlichen Abfolge, in der lautsprachliche Einheiten in einem Sprachdatenkorpus auftreten, kommt ein probabilistisches Sprachmodell zum Tragen, das die kontextabhängigen Wahrscheinlichkeiten der Sprachsegmente bestimmt. Die zeitliche Abfolge der potentiellen Sprachsegmente wird durch einen Markov-Graphen repräsentiert. Die Auswahl der Sprachsegmente für die linguistisch, prosodisch motivierte Zieläußerung basiert auf dem Entropie-Maß. Das Segment, welches die meiste Information liefert wird als lautsprachliches Äquivalent für die zu synthetisierende Äußerung ausgewählt. In unserer Arbeit hat sich gezeigt, dass es dennoch vorkommen kann, dass mehrere Sprachsegmente mit gleich hoher Wahrscheinlichkeit zur Auswahl zur Verfügung stehen. Hier wird der Abstand hinsichtlich der spektralen Eigenschaften, die mittels *Mel-frequency cepstral coefficients* (MFCC) repräsentiert sind [6, 12] berechnet. Als Abstandsmaßberechung haben wir die Mahalanobis-Distanz zwischen den aufeinander folgenden Sprachbausteinen verwendet. Eine zweite Sichtweise, welche sich aus den Arbeiten ergab, war die Reduktion der Korpus-Einheiten. Sprachsegmentbausteine, die mehrfach vorhanden sind und den gleich hohen Informationsgehalt haben, konnten bei minimalen Abweichungen der spektralen Eigenschaften aus dem Korpus entfernt werden, da sich perzeptiv keine Verbesserung ergab. Dadurch ist die Erstellung einer ressourcenschonenden Datenbasis gegeben. Das resultierende Unit-Selection basierte TTS-Modul wurde in unsere bestehende Software AVISS zur Generierung Audio-Visueller-Synthese integriert.

Dieser Beitrag gliedert sich wie folgt: In Abschnitt 2 wird die Parameter-Extraktion dargestellt, welche die Sprachsegmente durch einen Merkmalsvektor beschreibt und so einer parametrischen Darstellung der jeweiligen Sprachsegmente entspricht. In Abschnitt 3 wird der Aufbau des Wahrscheinlichkeitsmodells für den parametrisierten Korpus erläutert und der Informationsgehalt der potentiellen Sprachbausteine, welche für die Synthese in Betracht kommen, durch das Entropie-Maß herausgearbeitet. In Abschnitt 4 kommt das Abstandsmaß der spektralen Eigenschaften zum Einsatz, welches durch die Mahalanobis-Distanz der MFCC Koeffizienten errechnet wird. Abschnitt 5 enthält einen kurzen Ausblick auf weitere Arbeiten.

## 2 Parameter Extraktion

Unit-Selection basierte Sprach-Synthese Systeme basieren auf Sprachdatenkorpora, aus denen zur Laufzeit Sprachbausteine ausgewählt werden. Die Sprachsegmente, welche in unserem System zur Auswahl stehen gliedern sich in Bigramme, Wörter, Silben, Triphone, Diphone und Phone. Als Auswahlkriterien zur Laufzeit dienen linguistisch motivierte wie auch prosodische und spektrale Vorgaben, welche die auszuwählenden Segmente determinieren. Die zur Auswahl stehenden Sprachsegmente werden in unserer Arbeit parametrisch dargestellt, was sich in Form eines Merkmalvektors für das entsprechende Sprachsegment widerspiegelt. Zur Beschreibung der Sprachsegmente durch die jeweiligen Parameter muss eine Unterteilung der einzelnen Parameterkategorien vorgenommen werden. Zum einen sind die aus dem Sprachdatenkorpus direkt zu gewinnende Parameter in der Kategorie quantitative und linguistische Parameter zusammengefasst. Zum anderen werden die Parameter Dauer und F0 über ein Entscheidungsbaumverfahren vorhergesagt und bestimmt. Die spektralen Parameter werden durch *MFCC* repräsentiert und wurden zuvor aus dem bestehenden Sprachdatenkorpus für die jeweiligen Segmente extrahiert.

Der verwendete Sprachdatenkorpus besteht aus 3300 Sätzen Spontansprache, das wurde von einer nicht profesionellen Sprecherin gesprochen.

## 2.1 Quantitative und linguistische Parameter

Die quantitative Parametrisierung der Sprachsegmente ergibt sich aus der Position der jeweiligen Sprachbausteine innerhalb des Sprachdatenkorpus. Wortbasierte Segmente werden innerhalb eines Satzes und Segmente unterhalb der Wortebene werden zusätzlich innerhalb eines Wortes lokalisiert. Die quantitativen Parameter beschränken sich auf Wortsegmente, Silbensegmente sowie Triphon-Sequenzen, Diphone und Phone. Es wird jeweils der rechte und linke Abstand zum Satzende ermittelt. Sind die Segmente unterhalb der Wortebene, wird zusätzlich der rechte und linke Abstand im Wort ermittelt und als Segmentmerkmal verwendet. Hieraus ergibt sich ein quantitativer Merkmalsvektor der das Sprachsegment hinsichtlich der Position im Satz und Wort beschreibt. Als linguistische Parameter werden zu den quantitativen Merkmalen Part-of-Speech, Satz- und Wortakzent sowie Satztyp hinzugefügt. Bei Phon-Segmenten wird zusätzlich das binäre Merkmal stimmhaft, stimmlos berücksichtigt.

## 2.2 F0-Prädiktion und Dauer-Prädiktion

Die Natürlichkeit und Verständlichkeit der synthetisierten Äußerung hängt stark von den prosodischen Eigenschaften der Sprachsegmente im Verlauf einer Äußerung ab. Eine geeignete Sprachbaustein-Auswahl wird somit sehr stark durch die prosodischen Parameter F0 und Dauer bedingt. Je besser die akustischen Parameter F0 und Dauer als Zielwerte approximiert werden, desto genauer sind die Zielvorgaben hinsichtlich der Sprachsegmente, die ausgewählt werden. Dadurch wird die nachträgliche Manipulation der Sprachbausteine verhindert bzw. minimiert. Die Zielwerte für die Sprachsegmente werden über ein Entscheidungsbaumverfahren approximiert. Die Prädiktion des prosodischen Parameters F0 wird auf Silbenebene, Triphon-, Diphon- und Phonebene generiert. Bei Silben wird der mittlere F0-Wert des Silbenvokals vorhergesagt, bei Triphonen und Diphonen wird der mittlere F0-Wert der Phonemsequenz prädiziert. Auf Phonebene wird ebenso der mittlere F0-Wert für das entsprechende Phonem ermittelt. Die Vorhersage der Zielwerte für die Dauer der Segmente wird ebenfalls auf Wort-, Silben-, Triphon-, Diphon- und Phonemebene durchgeführt. Die Merkmalsvektoren für die Dauer-Prädiktion setzten sich aus den quantitativen und linguistischen Parametern zusammen. Als zusätzliche Parameter kommen noch stimmhaft, stimmlos hinzu. Der Kontext erstreckt sich über je zwei rechts und zwei links stehende Segmente. Für die Vorhersage der F0-Werte wurde der Merkmalsvektor um die Segmentdauer erweitert. Nachfolgend sind die Parameter zusammenfassend aufgeführt, die sowohl für das Entscheidungsbaumlernen als auch als Beschreibung für die Sprachbausteine herangezogen werden.

| **Quantitative Parameter** | **Linguistische Parameter** | **Prosodische Parameter** |
|---|---|---|
| Position im Satz rechts | Part-of-Speech | Dauer |
| Position im Satz links | Satztyp | F0-Wert |
| Position im Wort rechts | Satzakzent | |
| Position im Wort links | Wortakzent | **Spektrale Parameter** |
| | stimmhaft/stimmlos | MFCC |

**Tabelle 1 – Parameter zum Aufbau der Merkmalsvektoren der Satzsegmente**

Die Dauervorhersage und ebenso die F0-Vorhersage wird mittels eines CART [2] durchgeführt, wobei die zu vorhersagenden Klassen, jeweils als logarithmierter F0-Wert und logarithmierte Dauer repräsentiert sind.

### 2.3 Spektrale Parameter

Bei der Konkatenation der Sprachsegmente bestehen oftmals Störungen an den Lautgrenzen, die auf unterschiedliche spektrale Eigenschaften der Sprachbausteine zurückzuführen sind. Geeignete spektrale Parameter, die aus der Spracherkennung kommen, sind die *Mel-frequency cepstral coefficients*. Diese spektralen Parameter werden auch hier als spektrale Eigenschaften der Sprachsegmente verwendet. Um geeignete Bausteine auszuwählen, welche an den Konkatenationsstellen keine Sprünge aufweisen, hat es sich als sinnvoll erwiesen, ein Abstandsmaß zwischen den spektralen Eigenschaften der Spachsegmente einzuführen [10]. Das in dieser Arbeit verwendete Abstandsmaß wird mittels der Mahalanobis-Distanz berechent. Eine detaillierte Ausführung zur Abstandsmaßberechnung findet sich in Abschnitt 4 wieder.

## 3 Markov-Graph und Entropie-Maß

In diesem Beitrag wird eine Alternative zu der bekannten Einheiten-Kostenfunktion dargestellt. Das Verfahren ist zweistufig aufgebaut. Wie zuvor beschrieben wird in Unit-Selection basierten Sprachsynthese-Systemen zur Laufzeit mittels der Einheiten-Kostenfunktion ein Abstandsmaß errechnet, wie gut ein Sprachsegment aus dem Sprachdatenkorpus zu den Zielvorgaben der entsprechenden lautsprachlichen Einheit passt. Diese Einheiten-Kostenfunktion wird in unserer Arbeit durch ein probabilistisches Modell ersetzt, welches die Wahrscheinlichkeiten der Segmente und der Segmentabfolge wiedergibt. Auf Basis der parametrisierten Sprachsegemente des Korpus werden abhängige Wahrscheinlichkeiten berechnet. Die abhängigen Wahrscheinlichkeiten, der potentiell in Betracht kommenden parametrisierten Sprachsegmente, werden in einer Übergangsmatrix abgelegt. Im Verlauf der zu synthetisierenden Äußerung wird der Pfad für die wahrscheinlichste Segmentsequenz ermittelt. Bei der Auswahl der Sprachsegmente wird von der größten Einheit aus, hier Bigramme, die Suche gestartet. Die Segmente werden mit einer top-down Strategie ausgewählt. Haben die Segmente gleiche Übergangswahrscheinlichkeiten kommt die zweite Stufe zum Einsatz, wobei der spektrale Abstand zwischen den Elementen als Auswahlkriterium hinzugenommen wird. Mittels des Entropie-Maßes wird die Sprachsegmentabfolge bestimmt, welche die maximale Entropie aufweist und kann daher als beste Segmentsequenz angesehen werden.

### 3.1 Markov-Graph

Abbildung 1 zeigt einen potentiellen Markov-Graphen für die entsprechenden parametrisierten Sprachbausteine in ihrer zeitlichen Abfolge, die für die zu synthetisierende Äußerung in Betracht kommen. Die Sprachbausteine wurden mit ihren Wahrscheinlichkeiten in einer Übergangsmatrix abgelegt. Zur Laufzeit werden die Segmente, welche die Zielvorgaben maximal approximieren, in der zeitlichen Abfolge mittels eines Markov-Graphen dargestellt. $S_{(ij)}$ repräsentiert in Abbildung 1 jeweils ein in Frage kommendes Sprachsegment. Die Kanten des Graphen sind mit den Übergangswahrscheinlichkeiten gewichtet. Es wurden jeweils für die verschiedenen Segmentebenen die Übergangswahrscheinlichkeiten (2) zwischen den Segmenten voraus errechnet. Mittels der Übergangswahrscheinlichkeiten wird eine Vorauswahl der Sprachbausteine vollzogen. Die Sprachbausteine die unterhalb eines Schwellenwertes liegen kommen nicht in Betracht.

$$a_{ij} = P(X_t = s_i \mid X_{t-1} = s_j) \tag{2}$$

$$\prod_i = P(x_l = s_i) \tag{3}$$

Die Übergangswahrscheinlichkeiten der Sprachbausteine ergeben sich aus (2). Die Startwahrscheinlichkeiten der einzelnen parametrisierten Sprachsegmente werden hinsichtlich ihrer Verteilung im Korpus errechnet.

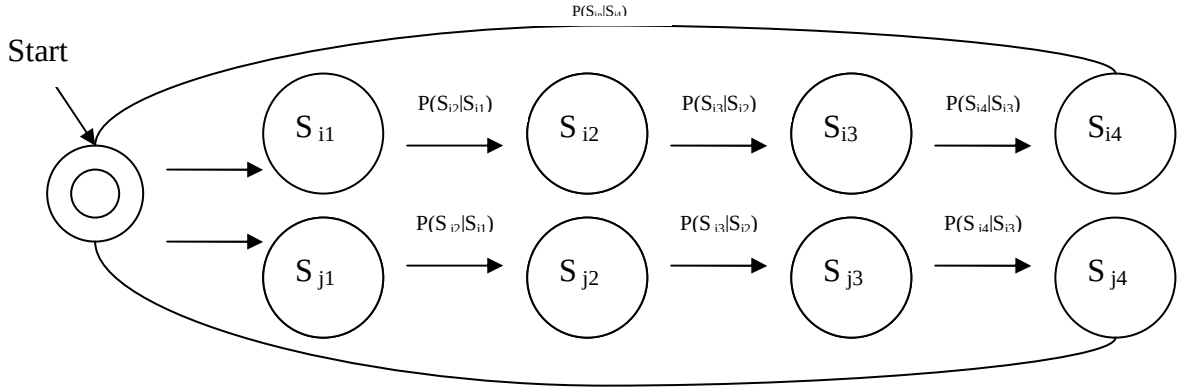


**Abbildung 1 –** Markov-Graph potentieller Sprachbausteine

Die für die zu synthetisierende Äußerung in Betracht kommenden Sprachbausteine bilden für die Auswahl einen Markov-Graphen mit den jeweiligen Segmenten als Knoten. Es wurden nur diejenigen Sprachsegmente als Knoten repräsentiert, die ihrer Wahrscheinlichkeit zu Folge als potentielle Segmente in Frage kommen. Segmente die eine geringe Wahrscheinlichkeit aufweisen, werden nicht als Knoten repräsentiert. So wird eine Reduktion der Segmentauswahlmöglichkeiten erreicht.

### 3.2 Entropie potentieller Sprachsegmente

Das Entropie-Maß (4) wird verwendet, um den Informationsgehalt wiederzugeben. In dieser Arbeit wird die Markov-Entropie (5) verwendet, um eine geeignete Bausteinfolge auszuwählen, die das beste Ergebnis hinsichtlich der zu synthetisierenden Äußerung liefert.

$$H = -\sum p_v \ \mathrm{ld}\ p_y \tag{4}$$

$$H_{\mathrm{Markov}_n} = 1 / (n + 1) * H(x_1, x_2, x_3, \ldots, x_n) \tag{5}$$

Mit (5) wird der Informationsgehalt der einzelnen Knoten über den Gesamt-Graphen errechnet. Daraus ergibt sich die Entropie der Segmentfolge. Es wird diejenige Segmentfolge zu einer Äußerung konkateniert die den maximalen Wert aufweist. Haben die Sprachsegmente die gleiche Wahrscheinlichkeit und demnach den gleichen Informationsgehalt an der entsprechenden Position kommt die Mahalanobis-Distanz der spektralen Eigenschaften der aufeinander folgenden Sprachbausteine zu Tragen.

## 4 Spektrales Abstandsmaß

Im Bereich der Lautübergänge wird der MFCC Abstand als Glattheitskriterium [1] eingesetzt. Die Berechnung dieses Abstandsmaßes ist nachfolgend angeführt.

### 4.1 Malahanobis-Abstand spektraler Eigenschaften

Die Mahalanobis-Distanz wird in dieser Arbeit als Distanzmaß zwischen dem mehrdimensionalen MFCC-Vektorraum benutz. Die Mahalanobis-Distanz wird somit als Abstandsmaß der spektralen Eigenschaften der ausgewählten Sprachsegmente verwendet, falls mehrere Auswahlmöglichkeiten vorhanden sind.

Die *m* Koeffizienten werden als m-dimensionaler Spaltenvektor dargestellt. Diese werden als Realisation eines Zufallsvektors $\underline{X}$ mit der Kovarianzmatrix $\underline{\Sigma}$ verwendet. Die Kovarianzmatrix wird durch die Matrix $\underline{S}$ berechnet. Der Abstand zweier so verteilter $\underline{x}$ und $\underline{y}$ werden dann bestimmt durch die Mahalanobis-Distanz

$$d(x, y) = (\underline{x} - \underline{y})^T \underline{S}^{-1} (\underline{x} - \underline{y}). \tag{6}$$

Je geringer der Abstand zwischen den Sprachbausteinen, um so geeigneter sind sie für die Konkatenation.

## 5 Schlussbetrachtung und Ausblick

In diesem Beitrag wurde ein Verfahren für die Unit-Selection basierte Sprachsynthese vorgestellt, welche sich als Alternative zu der bestehenden Einheitenkosten-Funktion anbietet. Mittels eines auf abhängigen Wahrscheinlichkeiten erstellten Markov-Graphen werden die geeigneten Sprachbausteine nach ihrer maximalen Entropie ausgewählt. Der Suchraum für die auszuwählenden Segmente hat sich durch dieses auf Wahrscheinlichkeiten beruhende Vorauswahlverfahren verkleinert. Das Verfahren wurde als TTS-Modul in unserem audio-visuellen Synthese-System eingesetzt. Erste Tests haben gute Ergebnisse erzielt. Eine Evaluation des Sprachsynthese-Moduls, welches auf diesem Verfahren beruht wird in zukünftiger Arbeit durchgeführt. Ebenso eine vergleichende Evaluation hinsichtlich Laufzeit und Perzeption mit einem Unit-Selection, welches mit der herkömmlichen Einheitenkosten-Funktion arbeitet. Testsätze können unter http://www.ikp.uni-bonn.de/~cwe/aviss bezogen werden.

## Literatur


[1] Black, A. and Campbell, N.: Optimizing selection of units from speech databases for concatenative synthesis. In: Eurospeech, Vol 1, Madrid, Spain 1995, pp 581-584.

[2] Breiman, L., Friedman, J., Olshen, R., Stone C.: Classification and regression trees, Chapman Hall, 1984.

[3] Donovan, R.E.: A new distance measure for costing spectral discontinuities in concatenative speech synthesizers, The 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[4] Hunt, A. and Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP 1996, Vol 1, Atlanta, Georgia, pp 373-376.

[6] Klabbers, E., Veldhuis, R.: Reducing audible spectral discontinuities, IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, January 2001, pp 39-51.

[7] Möbius, B.: German and Multilingual Speech Synthesis. Habilitationsschrift, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 7 (4) 2001.

[8] Stöber, K., Hess, W.: Additional Use of Phoneme Duration Hypothesis in Automatic Speech Segmentation. In: Proceedings of the ICSLP 1998, Sydney, Vol. 4, pp. 1595-1598

[9] Stöber, K., Wagner, P., Helbig, J., Köster, S., Stall, D., Thomae, M., Blauert, J., Hess, W.,

Hoffmann, R., Mangold, H.: Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora. In: Wolfgang Wahlster (ed.), Verbmobil: Foundations of Speech-to-Speech Translation, Symbolic Computation, Springer, Berlin, pp. 519-537

[10] Stöber, K.: Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese. Dissertation. IKP, Universität Bonn 2002

[11] Vary, P.; Heute, U.; Hess, W.: Digitale Sprachsignalverarbeitung. Teubner, Stuttgart, 1998.

[12] Wouters, J., Macon,M.: A perceptual evaluation of distance measures for concatenative speech synthesis, in Proc. 5th Int. Conf. Spoken Language Processing (ICSLP'98), vol. 6, Sydney, Australia, 1998, pp. 2747–2750.

[13] Tinsley, H., E.: Handbook of applied multivariate statistics and mathematical modeling. Academic Press, San Diego, 2000