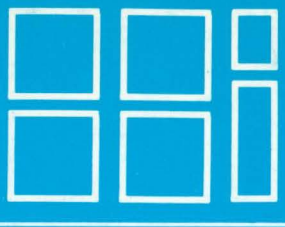


Heinrich-Hertz-Institut
für Nachrichtentechnik
Berlin GmbH

Technischer Bericht Nr. 206

Sprachparameterkonturen und deren Zeitnormalisierung bei der Sprecherverifizierung

Bearbeiter: Bernhard Kriener

Heinrich-Hertz-Institut für Nachrichtentechnik Berlin GmbH
Einsteinufer 37, 1000 Berlin 10

Technischer Bericht Nr. 206

Sprachparameterkonturen und deren Zeitnormalisierung bei der Sprecherverifizierung

Bearbeiter: Bernhard Kriener

AUTOR
.....B. Kriener.....
(Bernhard Kriener)

ABTEILUNGSLEITER
.....A. Lacroix.....
(i.V. Dr.-Ing. A. Lacroix)

WISS.TECHN.GESCHÄFTSFÜHRER
.....H. Ohnsorge.....
(Dr. H. Ohnsorge)

Berlin, 30.11.1979

ISBN 0172-8873

INHALT

1. Einleitung

In diesem Bericht wird untersucht, inwieweit Sprachparameterkonturen als Merkmalsvektoren bei der Sprecherverifizierung verwendet werden können. Außerdem wird über verschiedene Methoden zur Zeitnormalisierung berichtet. Zunächst werden die wichtigsten Begriffe dieses Themas und das im HHI aufgebaute Sprecher-Verifizierungs-System kurz erläutert. Eine ausführlichere Darstellung derselben ist in /1/ und /3/ zu finden.

1.1 Was ist Sprecherverifizierung?

Bei der automatischen Sprechererkennung unterscheidet man zwei Anwendungsarten, die Sprecheridentifizierung und die Sprecherverifizierung.

Bei der Identifizierung muß die Zuordnung Sprachprobe-Sprecher, d.h. die Identität des Sprechers dadurch ermittelt werden, daß die aktuelle Sprachprobe mit den gespeicherten Referenzproben aller Sprecher verglichen wird. Bei der Verifizierung dagegen gibt der Sprecher seine Identität vorher bekannt, sie muß nur bestätigt werden. Deshalb genügt es in diesem Fall, die Sprachprobe nur mit den gespeicherten Referenzdaten dieses einen Sprechers zu vergleichen. Es muß dann die Entscheidung getroffen werden, ob eine genügende Ähnlichkeit vorliegt, d.h. ob die Merkmale der aktuellen Sprachprobe mit den Referenzmerkmalen hinreichend übereinstimmen. Dabei wird vorausgesetzt, daß die verwendeten Merkmale sprecherspezifisch sind. Sie sollten bei dem gleichen Sprecher möglichst reproduzierbare Werte annehmen und bei verschiedenen Sprechern stark streuen.

1.2 Was sind sprecherspezifische Merkmale?

In einer sprachlichen Äußerung lassen sich sprecherspezifische Eigenschaften im wesentlichen auf die unterschiedliche Klangfärbung der Stimme und die unterschiedlichen Sprechweisen (abgehacktes oder fließendes Sprechen, monoton oder mit Betonung usw.) der Sprecher zurückführen. Während die Klangfärbung der Stimme sowohl durch die Anatomie der Stimmbänder und des Vokaltraktes

(Mund, Nasen- und Rachenraum) als auch durch die Gewohnheiten beeinflußt wird, hängt die Sprechweise, d.h. die zeitliche Struktur einer Äußerung, nur von den erlernten Gewohnheiten ab. Trotzdem sind der genauen Nachahmung wahrscheinlich Grenzen gesetzt. Andererseits werden bei einer Telefonübertragung die Merkmale der Sprechweise unempfindlicher gegenüber den dort auftretenden Störungen sein als die der Klangfärbung.

1.3 Wie werden sprecherspezifische Merkmale gewonnen?

Bei der automatischen Merkmalsgewinnung müssen die beschriebenen sprechertypischen Größen aus einer sprachlichen Äußerung extrahiert werden. Wird das Sprachsignal z.B. durch eine Filterbank in seine spektralen Anteile zerlegt (zur Spektralanalyse vgl. /2/), so erhält man eine Folge von Spektren, die eine Zeit-Frequenz-Matrix darstellen. In dieser Matrix beinhaltet jede Spalte ein Spektrum zu einer bestimmten Zeit, d.h. die momentane Klangfärbung der Stimme. Andererseits stellt jede Zeile der Matrix den zeitlichen Verlauf eines Spektralkanals dar und repräsentiert damit die Sprechweise. Bild 1 veranschaulicht diese Matrix.

Eine Art der Merkmalsgewinnung, die auf den sprachlichen Inhalt der Äußerung keinen Bezug nimmt, ist die statistische Analyse. Sie mittelt alle Spektralkanäle über der Zeit und liefert so das Langzeitspektrum, d.h. die mittlere Klangfärbung einer Stimme. Eine andere Art, sprecherspezifische Merkmale zu gewinnen, ist die Verwendung von Kurzzeit-Spektren von bestimmten Lauten. In dieser Segmentanalyse wird also die momentane oder lautbezogene Klangfärbung der Stimme ausgewertet (/3/).

In einer weiteren Art der Merkmalsgewinnung, der Konturenanalyse, werden die zeitlichen Verläufe von Spektralkanälen und daraus abgeleiteten Parameterkonturen verwendet, d.h. es werden im wesentlichen die Sprechgewohnheiten als Merkmale benutzt.

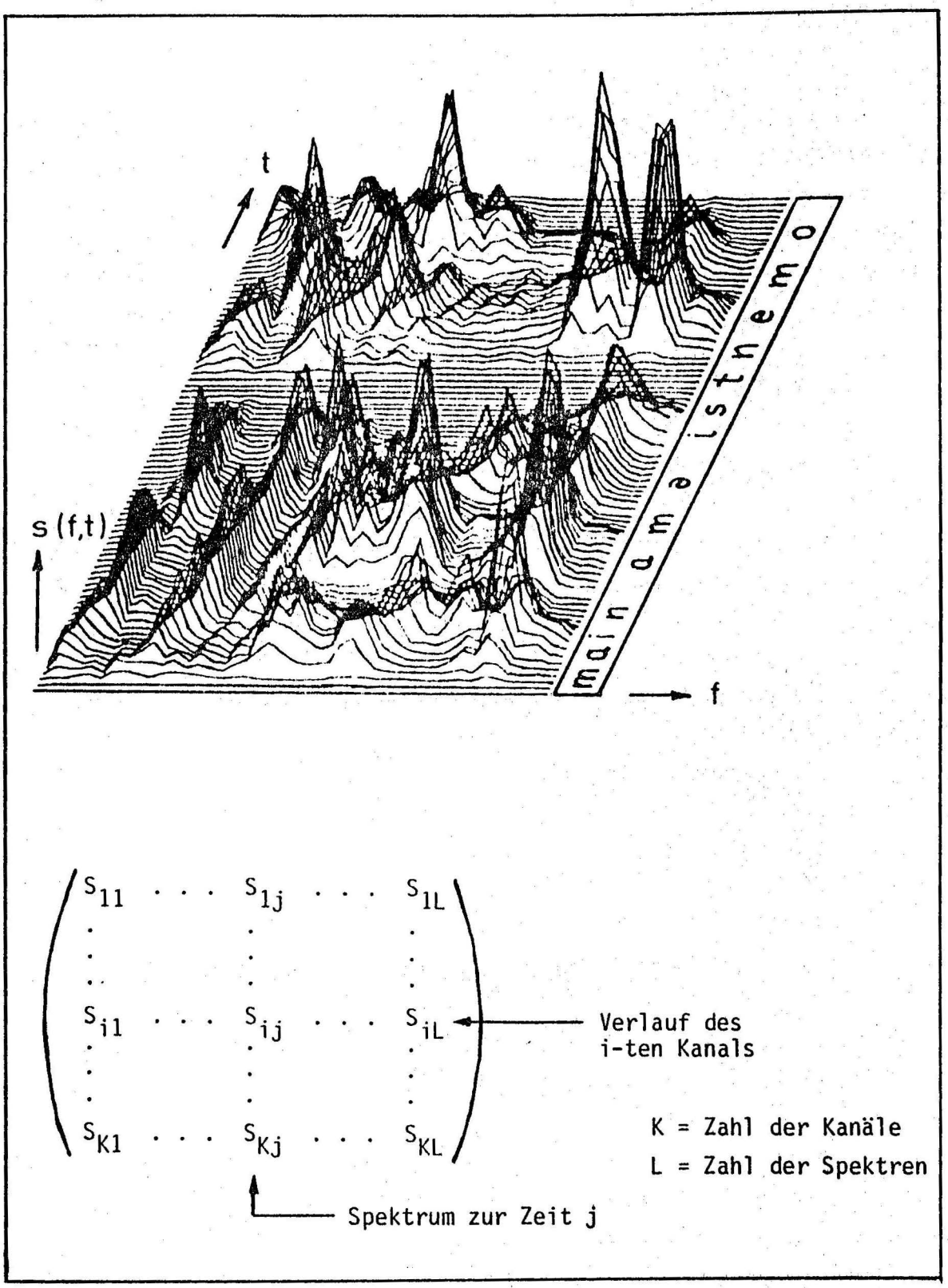


Bild 1: Zeit-Frequenz-Matrix

1.4 Was ist Zeitnormalisierung?

Sowohl die Segmentanalyse als auch die Konturanalyse sind von dem sprachlichen Inhalt der Äußerung abhängig. In der Segmentanalyse müssen die gewünschten sprachlichen Ereignisse (bestimmte Laute oder Lautübergänge) gefunden werden. Dies erfordert bei unbekanntem Text ein Verfahren zur Lokalisation von Sprachereignissen, d.h. eine Art Spracherkennung. Auf Grund der Kooperationsbereitschaft des Sprechers kann bei der Verifizierung jedoch ein fester Codesatz vereinbart werden. Durch die Kenntnis der Reihenfolge der Sprachlaute wird dann das Auffinden bestimmter Segmente erheblich erleichtert. In diesem Fall wird die Testäußerung einer Referenzäußerung, in der die Lage der Laute bekannt ist, zugeordnet. Diese Zuordnung erfordert in der Regel eine nichtlineare Zeitnormalisierung, wie in Abschnitt 2.1 gezeigt wird.

Bei der Konturanalyse ist zwar zunächst nur eine lineare Zeitnormalisierung (Dehnung oder Stauchung) notwendig, um für die Klassifizierung eine konstante Länge der Konturen zu erzeugen. Trotzdem trägt auch hier eine nichtlineare Zeitnormalisierung dazu bei, die Merkmale sprechertypischer zu machen, wie in Abschnitt 3 gezeigt wird.

1.5 Das Sprecher-Verifizierungs-System (SVS)

An dieser Stelle soll das System nur in Umrissen dargestellt werden, soweit es für das Verständnis dieses Berichtes notwendie ist.

Das im HHI aufgebaute Sprecher-Verifizierungs-System besteht aus einer Filterbank für die Vorverarbeitung und einem Minicomputer für die Merkmalsextraktion und Klassifizierung. Der Teilnehmer wird nach Eingabe einer Codenummer aufgefordert, einen bestimmten Codesatz nachzusprechen. Innerhalb des anschließenden Aufnahmefensters erzeugt dann die Filterbank aus dem ankommenden Signal eine Folge von Kurzzeitspektren, die an den Rechner übergeben

werden. Dort werden die Merkmalsextraktion und die Klassifizierung durchgeführt. Der Teilnehmer hat bei einem zunächst erfolglosen Verifizierungsversuch noch maximal zwei weitere Möglichkeiten, entweder doch noch akzeptiert oder endgültig zurückgewiesen zu werden.

2. Verfahren zur Zeitnormalisierung

Wenn im Rahmen einer Merkmalsextraktion eine Lautzuordnung zwischen zwei Codesätzen hergestellt werden soll, so muß zunächst geklärt werden, ob für die Zuordnung die gesamte Zeit-Frequenz-Matrix berücksichtigt werden muß, oder ob es eine bestimmte Kontur gibt, die in zeitlicher Richtung repräsentativ für die Matrix ist. Dies ist nicht zuletzt eine Frage des Rechen- bzw. Zeitaufwandes, der bei der Merkmalsextraktion akzeptabel erscheint.

Bei der Konturanalyse bietet sich zunächst diejenige Frequenzkontur an, die dann auch klassifiziert werden soll. Sollen allerdings verschiedene Frequenzkonturen ausgewertet werden, so ist es sinnvoll, eine gemeinsame Kontur zu suchen, auf die eine Zeitnormalisierung angewendet wird.

Bei der Segmentanalyse erscheint es analog dazu sinnvoll, die Spektren, die anschließend klassifiziert werden, direkt aus der Zeit-Frequenz-Matrix herauszusuchen. Bei diesem sog. Pattern-Matching werden ein oder mehrere typische Spektren (Referenzspektren) des gesuchten Lautes über die Kurzzeit-Spektren der aktuellen Äußerung bzw. des Aufnahmefensters geschoben und jeweils der Abstand gebildet. Dieser Abstand hat dann dort ein Minimum, wo der Laut in dem Aufnahmefenster vorkommt. Bild 2 veranschaulicht dies an dem Wort 'jedes'. Je nach Zahl der zu suchenden Spektren erhöht sich dabei der Rechenaufwand.

Bei dem implementierten Sprecher-Verifizierungs-System sollten nun beide Analysearten möglich sein. Aus diesem Grund ist es sinnvoll, die Zeitnormalisierung auf eine Kontur anzuwenden, die in beiden Fällen als charakteristisch erscheint. Die einfachste Möglichkeit bietet hier die Intensitätskontur, im folgenden kurz Pegel genannt.

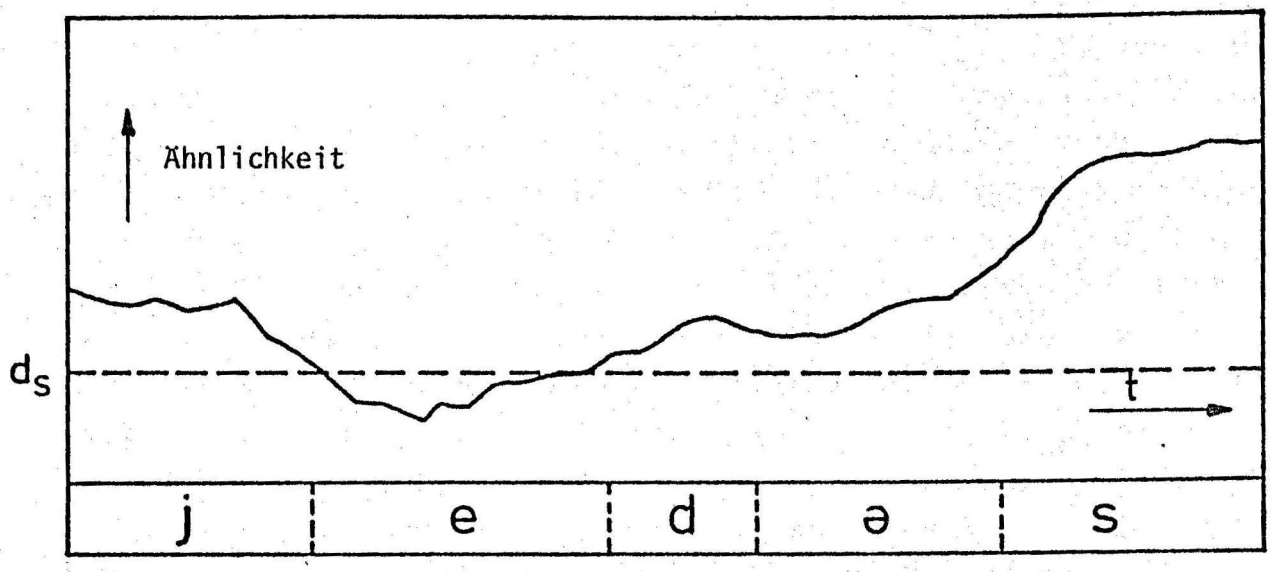


Bild 2: Ähnlichkeit des Referenzspektrums 'e' mit den Spektren des Wortes 'jedes'

Nun haben stimmlose Laute in der Regel einen niedrigeren Pegel als stimmhafte Laute. In der normalen Sprache läßt er sich oft kaum vom Rauschpegel unterscheiden. Berücksichtigt man jedoch den Energieanteil der höheren Frequenzen (> 1000 Hz) durch Höhenanhebung stärker als er in der natürlichen Sprache vorhanden ist, so kann man auch stimmlose Phoneme lokalisieren. Dadurch bietet der Pegelverlauf der gesamten Energie eine ziemlich stabile Kontur, die sowohl für die Segment- als auch die Konturanalyse als repräsentativer Parameterverlauf zur Zeitnormalisierung verwendet werden kann.

2.1 Prinzipielle Methoden der nichtlinearen Zeitnormalisierung

Die Pegelverläufe von verschiedenen Realisierungen ein und desselben Codesatzes haben in der Regel nicht nur unterschiedliche Länge je nach Sprechgeschwindigkeit, sondern weisen auch nach einer linearen Zeitnormalisierung, die die Konturen auf gleiche Länge interpoliert, noch zeitliche Verzerrungen auf, wie Bild 3a zeigt.Um eine Lautzuordnung zwischen diesen Codesätzen herzustellen, ist daher eine nichtlineare Zeitnormalisierung notwendig. Mathematisch ausgedrückt, bedeutet dies, daß eine Testkontur $S(\tau)$, die an den M Stützstellen $\underline{\tau} = (\tau_1, \ldots \tau_M)$ die Werte $S(\tau_1)$ bis $S(\tau_M)$ hat, sowohl in der Amplitude als auch in der Zeit Abweichungen zu der an den L Stützstellen $\underline{t} = (t_1, \ldots t_L)$ definierten Referenz-Kontur R (t) aufweist. Daraus ergibt sich die Beziehung

$$S(\tau(t)) = R(t) + E(t) \qquad (1)$$

mit der Zuordnungskontur

$$\tau(t) = t + q(t), \qquad (2)$$

wobei E(t) die Differenz in der Amplitude und q(t) die zeitl. Verzerrung zwischen dem Wert der Referenz-Kontur an der Stelle t und dem zugeordneten Wert der Test-Kontur an der Stelle $\tau(t)$ darstellen. Die Verzerrungskontur q(t) hat dann ebenfalls diskrete Werte.

Die Zuordnung $t \rightarrow \tau$ muß im allgemeine Fall keine eindeutige Abbildung sein (vgl. Bild 3b), da sowohl in der Referenz- wie auch in der Testkontur Punkte herausfallen können.
Das bedeutet, daß $\tau(t)$ und q(t) nur für die zugeordneten Referenzpunkte t_z definiert sind, deren Anzahl L_z kleiner als L sein kann. Entsprechend gibt es dann L_z zugeordnete Testpunkte τ_z.

In der praktischen Anwendung ist es aber sinnvoll, durch zusätzliche Randbedingungen (vgl. Abschnitt 2.5) z.B. eine eindeutige Abbildung mit $L = L_z$ zu ermöglichen, d.h. daß jedem Referenzpunkt ein Testpunkt zugeordnet werden kann.

Eine optimale Anpassung beider Konturen bedeutet nun, daß möglichst zu jedem Punkt der Referenz-Kontur derart ein Punkt in der Testkontur gefunden wird, daß ein Abstandsmaß zwischen den zugeordneten Amplitudenwerten zu einem Minimum wird.

Wird als Abstandsmaß z.B. der mittlere quadratische Fehler genommen, dann bedeutet dies

$$A_{MQF} = \sum_{t \in t_z} E(t) = \sum_{t \in t_z} [S(\tau(t)) - R(t)]^2 \stackrel{!}{=} \text{Min.}$$

Entsprechendes gilt für den mittleren absoluten Fehler A_{MAF} bzw. die Korrelation A_{KOR}.

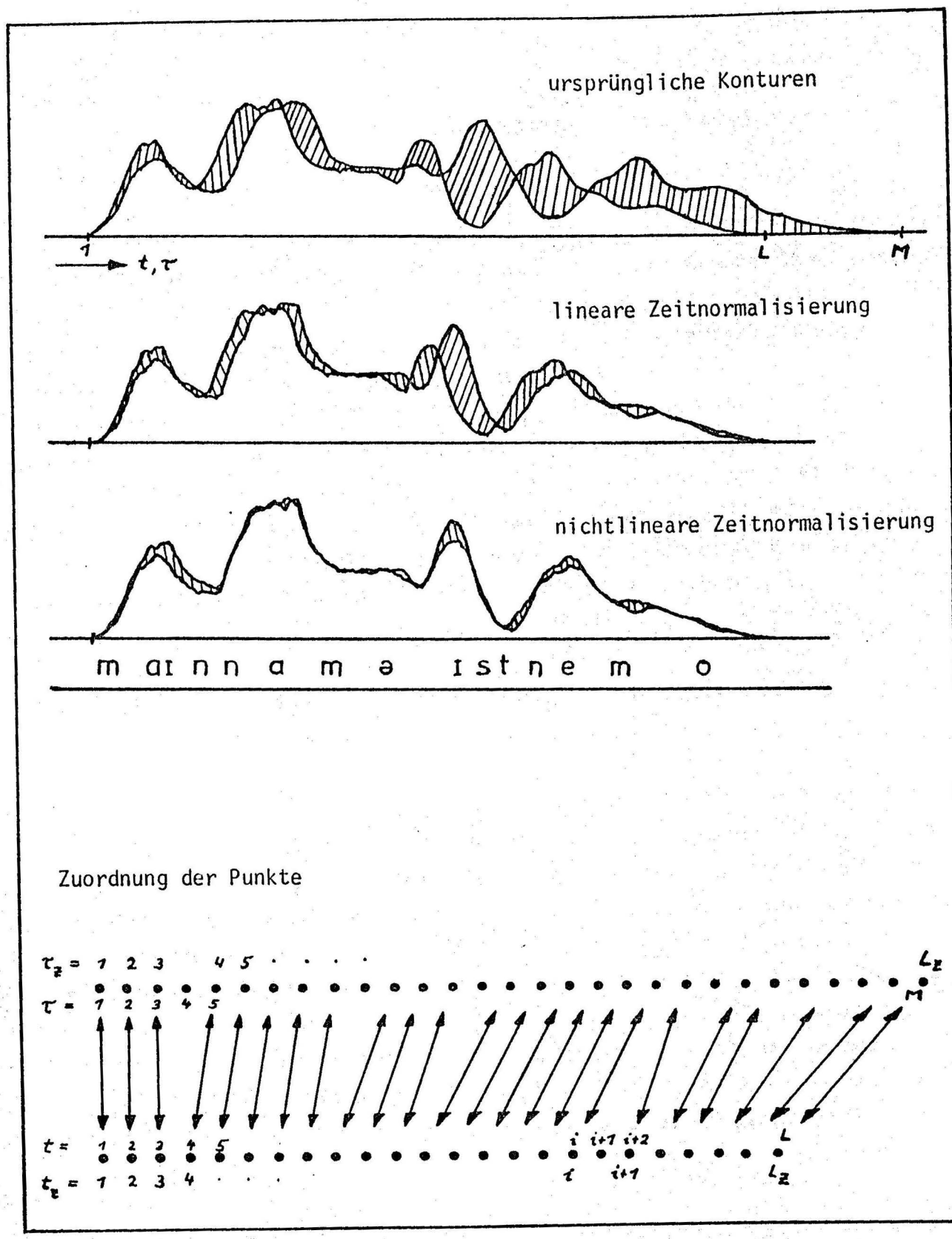


Bild 3: Zeitnormalisierung

2.2 Normierung der Konturen

Bei der Anwendung der Abstandsmaße A_{MQF} oder A_{MAF} müssen die Konturen in ihrer Amplitude auf gleichen Mittelwert normiert werden.

Diese Normierung erfordert zunächst eine Anfang-Ende-Erkennung, damit der mittlere Pegel der Testkontur bestimmt werden kann. Dazu kann der Energiepegel mit einer Rauschschwelle verglichen werden. Die Bestimmung dieser Rauschschwelle muß i.a. adaptiv erfolgen, d.h. sie muß von der Lautstärke, mit der das Signal aufgenommen wurde, abhängig sein. Eine Möglichkeit besteht darin, zu Beginn des Aufnahmefensters in einem kurzen Zeitabschnitt von ca. 100 ms den Pegel zu messen und den Mittelwert davon als Rauschschwelle zu nehmen. Gegebenenfalls kann das gleiche auch am Ende des Fensters durchgeführt werden. Das setzt allerdings voraus, daß in diesem Zeitabschnitt kein Nutzsignal vorhanden ist.

Eine andere Methode der Normierung besteht darin, den Energiepegel über das ganze Aufnahmefenster zu mitteln.

Sofern sichergestellt werden kann, daß das Verhältnis der Testkonturlänge zur Länge des gesamten Aufnahmefensters etwa gleich bleibt, kann dieser mittlere Pegel zur Normierung verwendet werden. Bei der Anwendung der Dynamischen Programmierung im Sprecherverifizierungssystem hat sich diese Normierung als ausreichend erwiesen, da durch das Vorsprechen des Codesatzes die Länge ziemlich konstant bleibt. Zwar kann sie durch Hinzufügen weiterer Äußerungen bzw. entsprechender Störungen verfälscht werden. Eine derartige Verfälschung der Normierung kann jedoch nur eine Zurückweisung der Sprachprobe bewirken, was in diesem Fall angebracht erscheint.

Andererseits kann ein bestimmter prozentualer Anteil dieses mittleren Signalpegels als Rauschschwelle für die Anfang-Ende-Erkennung verwendet werden.

Pausen bei Plosivlauten bzw. zwischen einzelnen Worten innerhalb des Codesatzes können als zum Signal gehörig erkannt werden, wenn eine bestimmte maximale Pausenlänge vorgegeben wird.

Auf analoge Weise lassen sich vor und hinter dem Signal kurze Störungen (Knacken etc.) durch Festsetzung einer minimalen Signallänge eliminieren. Beide Variablen können dem sprachlichen Inhalt der Codesätze angepaßt werden.

2.3 Gradientenverfahren

Doddington /7/ hat 1974 ein Verfahren veröffentlicht, in dem das Abstandsmaß zwischen zwei Konturen mit Hilfe des Gradienten minimisiert wird. In einem iterativen Prozeß wird der Gradient von A ($\bar{\nabla}A$) bezüglich der Variablen $\underline{t}$ und $\underline{q}$ gebildet. Diese Variablen werden dann proportional dazu schrittweise erhöht. Beim i-ten Schritt sind demnach die Inkremente

$$\underline{\Delta t_i} = C_t \, (\bar{\nabla}_t A)_i$$

$$\underline{\Delta q_i} = C_q \, (\bar{\nabla}_q A)_i$$

$$\text{mit} \quad (\nabla_p A) = \left(\frac{\delta A}{\delta p_1}, \frac{\delta A}{\delta p_2}, \ldots \frac{\delta A}{\delta p_L} \right)$$

Dadurch wird immer in Richtung der größten Änderung von A eine Änderung der Variablen $\underline{t}$ bzw. $\underline{q}$ vorgenommen. Die Proportionalitätsfaktoren C_t und C_q können davon abhängig gemacht werden, ob die Richtung der Änderung beibehalten wird oder nicht.

Wie bei allen derartigen Minimisierungsverfahren hat die Schrittweite der Inkrementierung erheblichen Einfluß darauf, ob das Verfahren das absolute Minimum findet oder ob es in einem relativen Minimum hängenbleibt. Die Iteration wird dann abgebrochen, wenn die Änderung von A eine bestimmte Schwelle unterschreitet, d.h. wenn A nicht mehr signifikant kleiner wird.

Das Verfahren ist sehr rechenintensiv. Deshalb wurde bei einer Implementierung die Zahl der möglichen Verzerrungspunkte, die ursprünglich gleich der Zahl der gegebenen Stützstellen ist, auf 10 äquidistant über die gesamte Länge verteilte Punkte reduziert. Die Arbeitsweise des Verfahrens zeigt Bild 4 an dem Codesatz "Mein Name ist Nemo".

Nach der 15. Iteration wurde der mittlere quadratische Fehler nicht mehr kleiner. In Bild 4 unten ist die Verzerrung aufgezeichnet, die sich danach ergeben hatte.

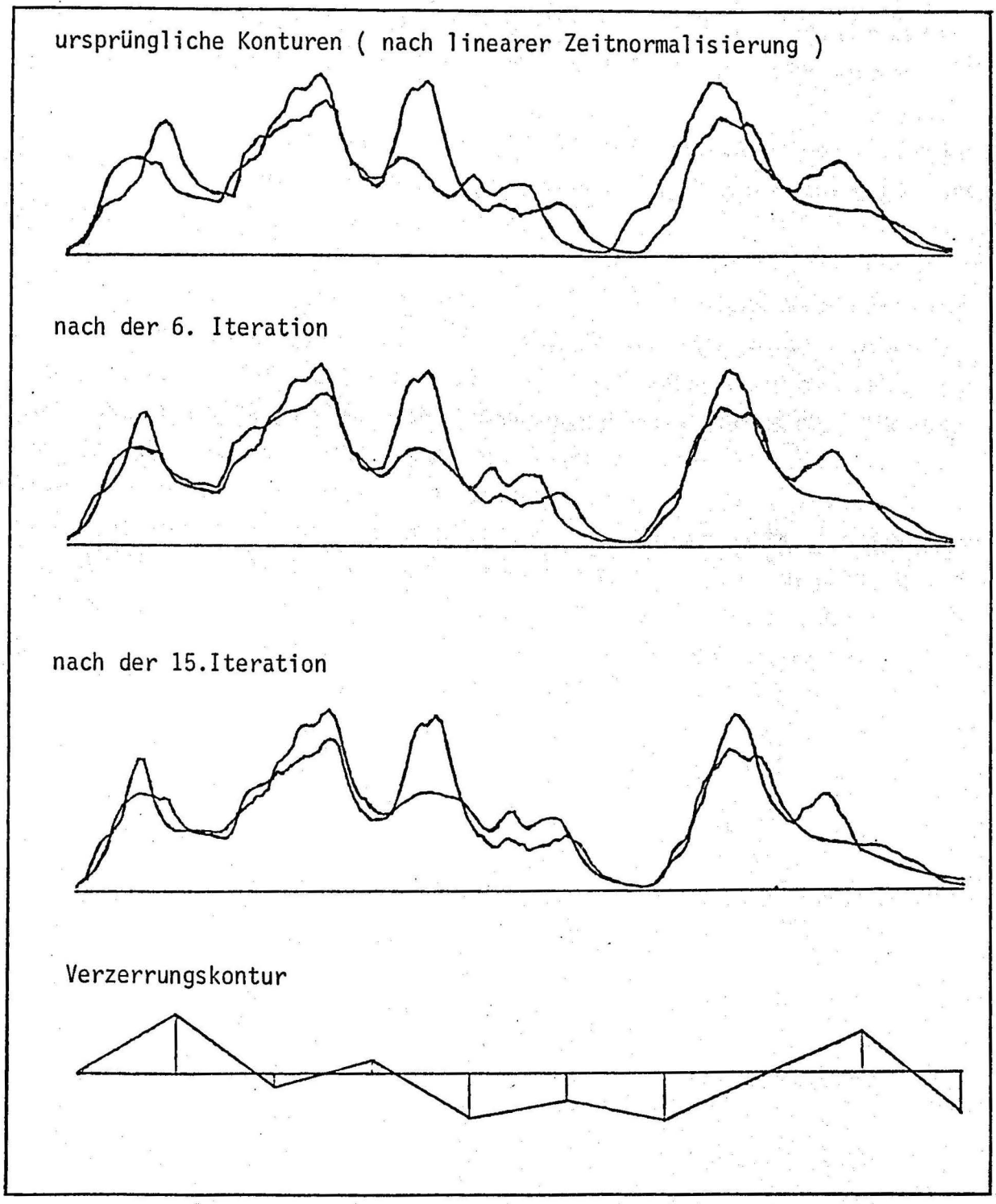


Bild 4: Zeitnormalisierung mit Hilfe des Gradienten-verfahrens

Bild 5 zeigt ebenfalls ein Beispiel für die Normalisierung von zwei Konturen, die mit diesem Verfahren erzielt wurde.

An beiden Ergebnissen wird deutlich, daß zwischen mathematisch optimaler Anpassung und einer Anpassung im Sinne einer Lautzuordnung ein Unterschied sein kann.

Innerhalb von stimmhaften Bereichen haben Vokale und vokalähnliche Laute einen höheren Pegel als stimmhafte Konsonanten*). Das bedeutet, daß im Pegelverlauf einer sprachlichen Äußerung die Maxima in der Regel Vokale anzeigen und die Minima Konsonanten oder Pausen, die auch bei Plosivlauten auftreten. Deshalb müßten bei einer Lautzuordnung diese Extrema einander zugeordnet werden. Eine Verschiebung der Maxima, die das Verfahren jeweils am Ende des Codesatzes (bei "o" von "Nemo") durchgeführt hat, ist daher nicht sinnvoll. Diese Überlegungen führten zur Entwicklung des nachfolgend beschriebenen MML-Verfahrens.

---

*) Da bei den ersteren die Mundöffnung größer ist als bei Konsonanten, kann mehr Energie der Anregungsfunktion eingestrahlt werden.

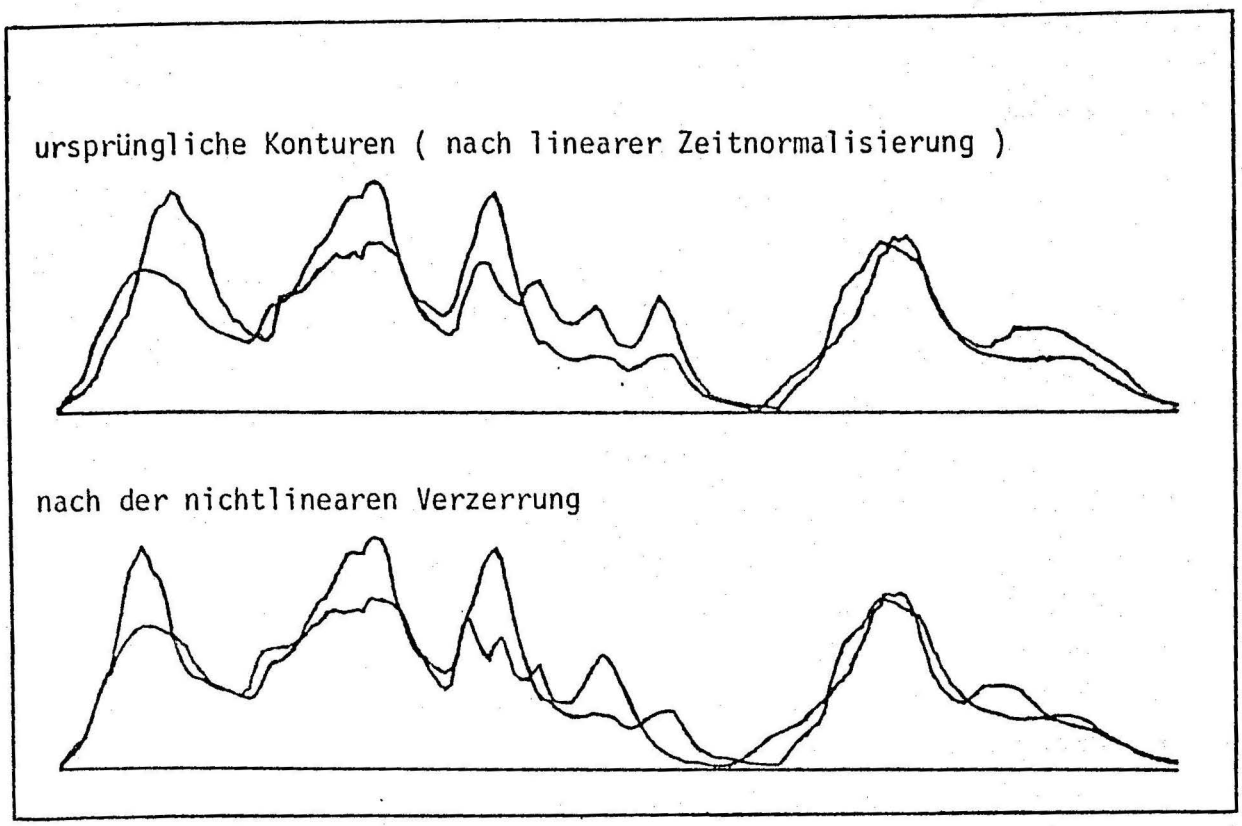


Bild 5: Zeitnormalisierung mit Hilfe des Gradienten-verfahrens

2.4 MML-(Maximum-Minimum-Lokalisations-)Verfahren

Dieses Verfahren führt genau die im vorhergehenden Abschnitt geforderte Extremwertzuordnung durch. Es ist weniger ein (Amplituden-) Abstandsminimisierungsverfahren wie in Abschnitt 2.1 beschrieben, sondern läßt sich eher als Lautverifizierungsverfahren charakterisieren. Es beruht auf der Entscheidung, ob ein bestimmtes sprachliches Ereignis, das an einem Pegelextremum erkenntlich ist, vorliegt oder nicht.

Dadurch wird nur an den Orten der Extrema eine nichtlineare Verzerrung durchgeführt. Die Zahl der Verzerrungspunkte wird somit auf die Zahl der vorhandenen Extrema reduziert, bzw. es werden nicht mehr alle Stützpunkte der Konturen optimal angepaßt, sondern nur noch wenige markante Punkte. Bei der Zuordnung derselben werden dann die Verzerrungen q(t) minimisiert. Das bedeutet, daß die Punkte einander zugeordnet werden, die unter Berücksichtigung der unterschiedlichen Länge der Konturen den geringsten Abstand haben. Das Verfahren setzt also voraus, daß sich die Konturen nicht wesentlich voneinander unterscheiden. Diese Voraussetzung ist bei dem implementiertenSprecherverifizierungssystem gegeben, da die Codesätze vorgesprochen werden.

Zwar kann bei dem MML-Verfahren auf die Amplitudennormierung verzichtet werden, wenn die Auswahl der markanten Punkte (Extrema) nicht durch Amplitudenvergleich getroffen wird,jedoch erfordert das Verfahren eine Anfang-Ende-Bestimmung des Nutzsignals, da die ZuOrdnung dieser Punkte wie erwähnt von der Länge des Codesatzes abhängt.Die dazu notwendige Rauschschwelle kann aus dem im Abschnitt 2.2 erwähnten gesamten mittleren Signalpegel ermittelt werden.

Die Zuordnung von Extrema erfordert nun, zwischen eigentlichen (signifikanten) und zufälligen (Neben-) Extrema zu unterscheiden, d.h. die markanten Punkte so auszuwählen, daß sie möglichst stabil sind. Es wurden dazu verschiedene Methoden der Signifikanzbestimmung erprobt. Am besten erwies sich ein Verfahren, das die

absoluten Extrema einer Differenz-Kontur bestimmt, die dadurch gewonnen wird, daß von der ursprünglichen Kontur eine stark geglättete Version derselben subtrahiert wird (s.Bild 6).

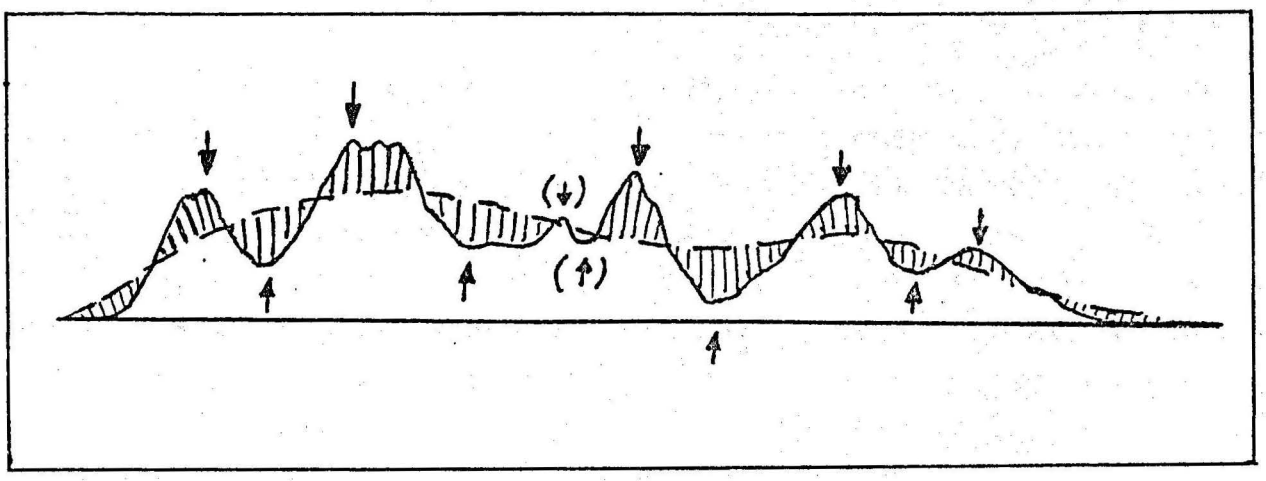


Bild 6: Signifikanz-Bestimmung beim MML-Verfahren

Nur wenn die Fläche der schraffierten Bereiche eine bestimmte Schwelle überschreitet, wird in diesem Bereich das absolute Extremum gesucht. Kleine Extrema, die zwar noch als solche in der Differenz-Kontur auftauchen, werden vernachlässigt. Das gleiche gilt für die relativen Extrema innerhalb der schraffierten Bereiche, die in der Regel ebenfalls keine signifikanten Extrema der ursprünglichen Kontur darstellen.

Die so gewonnenen signifikanten Extrema der Testkontur werden nun mit denen der Referenzkontur verglichen. Da die Zahl dieser Extrema unterschiedlich sein kann, wird die Zuordnung derart vorgenommen, daß zu jedem Referenzextremum das auf die gesamte Länge bezogen nächstliegende Testextremum gesucht wird. Dies ist analog zum Referenzextremum ein Minimum oder ein Maximum.

Da der minimale Abstand abwechselnd von beiden Seiten gesucht wird, d.h. da sowohl zu jedem Referenzextremum das nächstliegende Testextremum, als auch zu jedem aktuellen Testextremum das nächstliegende Referenzextremum gesucht wird, ist eine Falschzuordnung bzw. eine Fehlerfortpflanzung als Folge derselben nicht möglich.

Es gibt aber auch die Möglichkeit, daß ein Codesatz mehrere relativ schwach ausgeprägte Extrema hintereinander hat, von denen je nach Aussprache mal das eine, mal das andere Extremum als signifikant definiert wird. Die Folge davon, daß in diesem Fall zwar signifikante Extrema, aber nicht die gleichen Laute einander zugeordnet werden, kann dadurch verhindert werden, daß der gefundene Abstand zwischen den Extrema (immer bezogen auf die gesamte Länge des Codesatzes) noch daraufhin überprüft wird, ob er kleiner als eine maximal mögliche Verzerrung ist.

Dieses MML-Verfahren, das in der 2. Projektphase eingesetzt wurde, brachte bei Aufnahmen in ruhiger Umgebung gute Normalisierungsergebnisse. Eine wesentliche schwache Stelle des Verfahrens bleibt allerdings durch die Anfang-Ende-Bestimmung des Codesatzes erhalten.Treten an diesen Stellen Störgeräusche auf, so werden sie als zum Nutzsignal gehörig betrachtet. Die dadurch falsch gesetzten Grenzen bzw. die sich daraus ergebende falsche Gesamtlänge des Codesatzes können dann zu erheblichen Falschzuordnungen von Lauten führen.

Aus diesem Grund wurde in der letzten Projektphase das Verfahren der Dynamischen Programmierung zur Zeitnormalisierung verwendet.

2.5 Dynamische Programmierung

Die optimale Anpassung einer Kontur an eine andere kann auch dadurch erreicht werden, daß das Abstandmaß A für alle möglichen Zuordnungen berechnet wird und dann nach dem Minimum gesucht wird. Die bereits genannten Abstandsmaße sind prinzipiell alle auch hier anwendbar. Am häufigsten findet man in der Literatur die Dynamische Programmierung mit dem mittleren absoluten Fehler A_{MAF}. In dieser Art wurde sie auch im SVS implementiert. Deshalb soll diese Version hier näher erläutert werden.

Das Verfahren berechnet eine Matrix aller möglichen Zuordnungen der jeweiligen Stützpunkte, in der jedes Element die bis dahin aufsummierten Amplitudendifferenzen, d.h. das Abstandsmaß bis zu diesem Element angibt (s. Bild 7).

Das bedeutet, daß jedes Element sich zusammensetzt aus der aktuellen Differenz der beiden Amplitudenwerte und dem kleinsten Element in der Umgebung. Was die Umgebung ist, hängt von den Einschränkungen ab, die zusätzlich gemacht werden.

Im vorliegenden Fall der Anwendung im Sprecherverifizierungssystem wird, wie schon mehrfach erwähnt, der Codesatz vorgesprochen. Dies hat zur Folge, daß die aktuelle Länge des nachgesprochenen Codesatzes nur zwischen der einfachen und zweifachen Minimallänge variiert. Wird nun die Referenzkontur so gestaltet, daß sie gerade diese Minimallänge hat, dann kann der Pfad der Zuordnungen nur die Steigung 1 oder 1/2 haben.

In Bild 3 bedeutet dies: Zu jedem Referenzpunkt existiert eine eindeutige Abbildung auf die Testpunkte. Andererseits können nicht zwei Testpunkte hintereinander herausfallen.

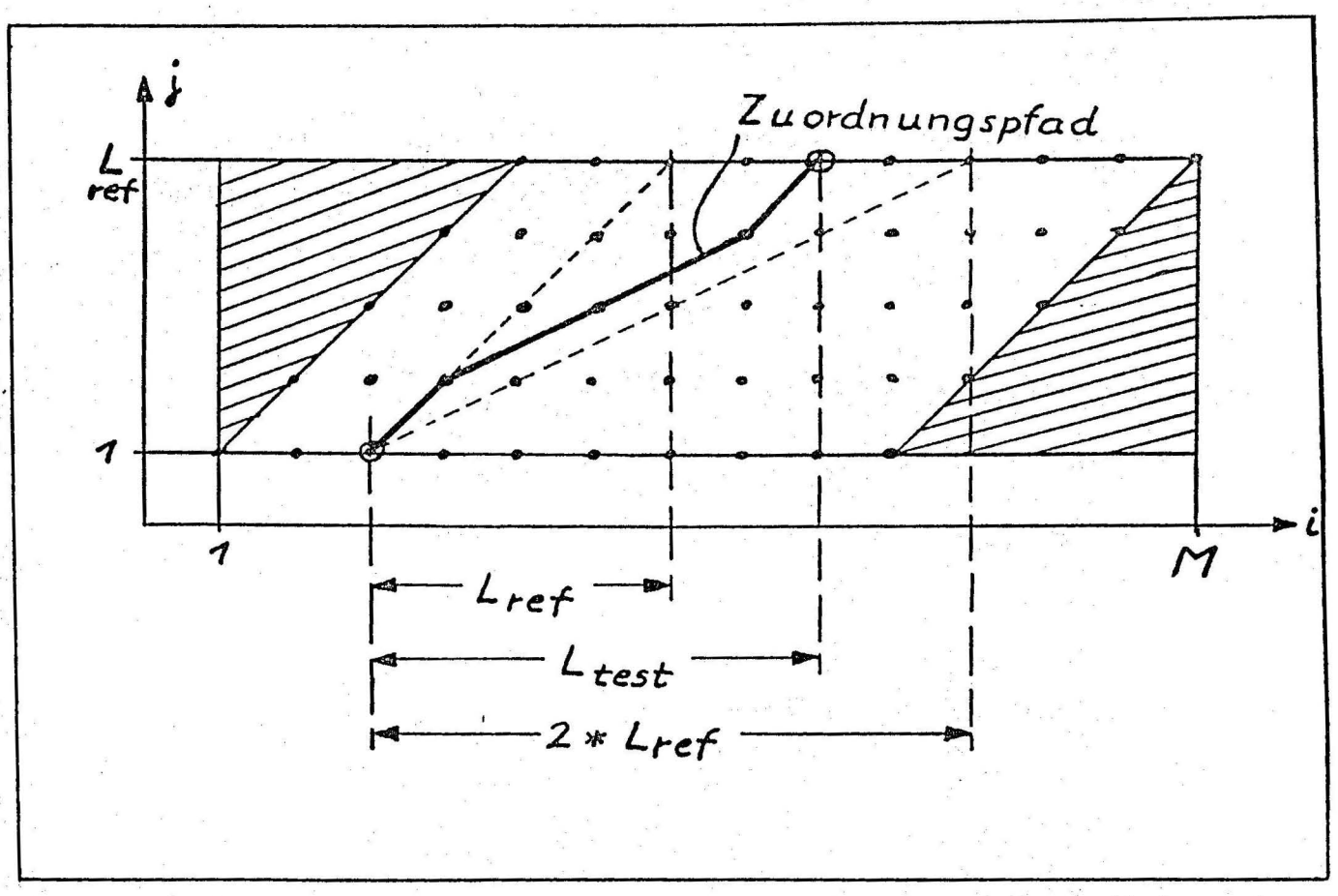


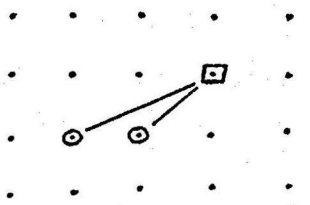
Bild 7: Dynamische Programmierung

Daraus ergibt sich folgende Berechnungsvorschrift für die Matrixelemente:

$$A(i,j) = DIS(i,j) + \text{Min}\,|A(i-1,j-1),\ A(i-2,j-1)|; \quad \begin{matrix} i = 2,N \\ j = 2,L \end{matrix}$$

mit $DIS(i,j) = |S(i) - R(j)|$; N = Länge des Aufnahmefensters
L = Länge der Referenzkontur

Die Umgebung besteht dann nur aus den beiden vorhergehenden Elementen der letzten Reihe:



Die erste Zeile der Matrix beinhaltet nur die entsprechenden Differenzen

$$A(i,1) = DIS(i,1), \qquad i = 1,N$$

Wenn alle Elemente der Matrix berechnet sind (auf Grund der erwähnten Einschränkungen braucht der schraffierte Teil der Matrix von Bild 7 nicht berechnet zu werden), gibt das Minimum in der obersten Zeile das kleinste gesamte Abstandsmaß an, d.h. den Endpunkt des Zuordnungspfades mit den geringsten Abständen. Wird von diesem Punkt ausgehend jeweils der vorhergehende Punkt mit dem geringsten (Teil-)Abstandsmaß gesucht, so führt dieser Pfad zum Anfangspunkt des Codesatzes und beinhaltet die optimale Anpassung der Konturen.

3. Auswertung von Sprachparameter-Konturen

Bei dem implementierten Sprecherverifizierungssystem ist es möglich, alle die in der Einleitung genannten Arten der Merkmalsgewinnung zu benutzen. In diesem Bericht sollen nun die Möglichkeiten und Grenzen der Konturanalyse für die Sprecherverifizierung untersucht werden. Als Merkmalsvektoren werden dabei verwendet

a) die (Intensitäts-)Pegelkontur
b) eine Stationaritätskontur, die die Änderung aufeinanderfolgender Spektren beinhaltet
c) die Verzerrungskontur, die sich aus der nichtlinearen Zeitnormalisierung ergibt
d) Pegel-Konturen von verschiedenen einzelnen Frequenzkanälen
e) (zum Vergleich) das Langzeitspektrum

Bei allen Untersuchungen wurde, soweit nicht anders erwähnt, eine nichtlineare Zeitnormalisierung der Intensitätskonturen mit Hilfe der Dynamischen Programmierung durchgeführt. Dabei wurde eine sprecherspezifische Referenzkontur verwendet. Die anderen Parameterkonturen wurden ggf. an die normalisierte Pegelkontur angepaßt. Zur Klassifizierung der Merkmalsvektoren wurde eine gewichtete Korrelation verwendet.
Die angegebenen Fehlerraten sind so ermittelt, daß die Raten für Falscherkennung und Falsch-Rückweisung gleich groß sind.

Die Merkmalsvektoren wurden mit Hilfe einer Stichprobe von ca. 360 Realisierungen des Codesatzes "Sesam öffne dich" getestet. Die Stichprobe wurde im Rahmen eines Feldversuchs, der sich über mehrere Tage hin erstreckte, mit dem implementierten Sprecherverifizierungssystem gewonnen. Es nahmen 6 Sprecher daran teil, die zunächst je 10 Aufnahmen für die Lernphase machten. Im weiteren Verlauf der Stichprobenerstellung wurden dann abwechselnd Erkennungsversuche des wahren Sprechers und Täuschungsversuche durchgeführt. Bei den letzteren ahmten die Sprecher sich gegenseitig nach. Jeder Täuschungsversuch wurde dabei einerseits durch

Vorsprechen des wahren Sprechers, andererseits durch genaue Auskunft über die bereits erreichte Ähnlichkeit unterstützt. Dadurch sind die Täuschungsversuche nicht mehr zufällig, sondern gezielt vorgenommen. Auf diese Weise wurden ca. 50 Erkennungs- und 50 Täuschungsversuche, bestehend aus jeweils 3 Sprachproben gesammelt. Diese sind notwendig, da das Sprecher-Verifizierungs-System mit einer sequentiellen Entscheidungsstrategie arbeitet (Näheres dazu in /3/).

3.1 Die Pegelkontur der gesamten Energie

Diese Pegelkontur stellt die gesamte Energie des Sprachsignals über der Zeit dar. Wie oben erwähnt, ist es jedoch sinnvoll, die Energie im höheren Frequenzbereich anzuheben.

Bild 8 zeigt Realisierungen dieses Merkmalsvektors. Um zu ermitteln, wie die nichtlineare Zeitnormalisierung der Pegelkontur die Sprecherunterscheidung beeinflußt, wurde sowohl die zeitlich nichtlinear verzerrte als auch die linear auf Referenzkonturlänge gestauchte* Pegelkontur als Merkmalsvektor genommen.

In beiden Fällen wurde die Anfang-Ende-Bestimmung des Satzes mit der Zeitnormalisierung durchgeführt, damit nicht dadurch zusätzliche Unterschiede in den verwendeten Konturen erzeugt wurden.

Dabei zeigte es sich (vgl. Bild 12), daß eine nichtlinear auf eine Referenzkontur angepaßte Pegelkontur mit 8% Fehlerrate eine bessere Sprecherunterscheidbarkeit ergibt als eine linear interpolierte Kontur (12%). Wird die Anpassung jedoch nicht, wie oben erwähnt, auf eine sprecherspezifische, sondern auf eine sprecherunabhängige Referenzkontur durchgeführt, so verdoppelt sich in beiden Fällen die Fehlerrate.

---

*) Wie im Abschnitt 2.5 erwähnt, ist bei der im Sprecher-Verifizierungs-System implementierten Dynamischen Programmierung die Referenzkonturlänge die minimal zulässige Testkonturlänge. Aus diesem Grund ist die aktuelle Kontur immer länger als die Referenzkontur.

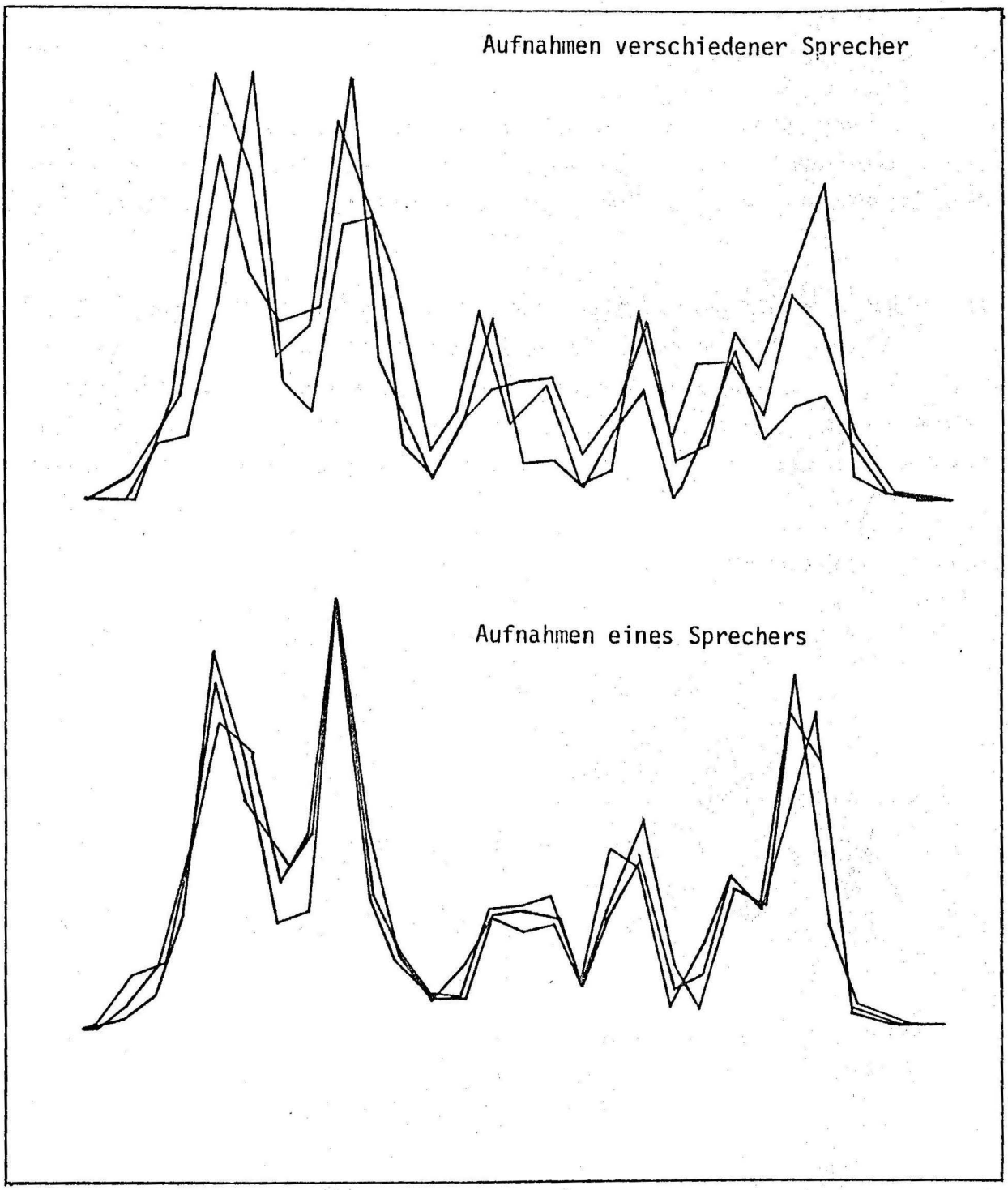


Bild 8: Inter- und Intrasprechervarianz der Pegelkontur

3.2 Stationaritätskontur

Um die Änderungen aufeinanderfolgender Spektren als Merkmal auswerten zu können, wurde eine Stationaritätskontur untersucht, die die zeitlichen Änderungen der einzelnen Spektralkanäle berechnet und dann die Kanaldifferenzen summiert. Vor der Differenzenbildung wurden die Kanalwerte auf den entsprechenden Gesamt-Pegelwert normiert.

Bild 9 zeigt Realisierungen dieses Merkmalsvektors. Zum Vergleich wurden auch hier die Abstände sowohl an den Stützpunkten der nichtlinear, als auch an denen der linear normalisierten Pegelkontur gebildet. Wie aus Bild 12 zu entnehmen ist, wurden dabei jedoch keine signifikanten Unterschiede festgestellt. Die Fehlerrate lag bei ca. 9%.

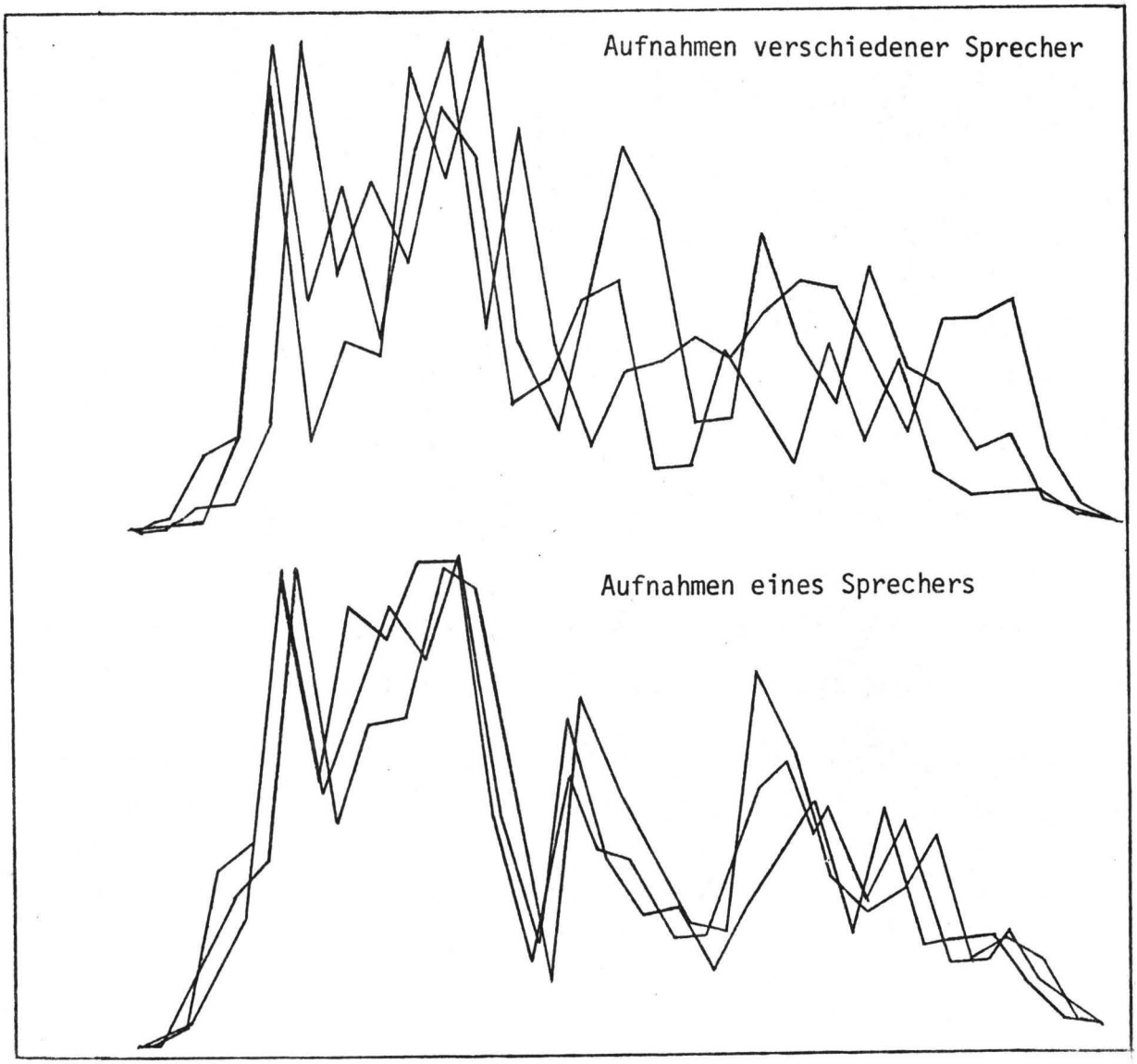


Bild 9: Inter- und Intrasprechervarianz der Stationaritätskontur

3.3 Die Verzerrungskontur

Bild 10 zeigt Verzerrungskonturen, wie sie bei der Zeitnormalisierung berechnet werden. Ihre Sägezahnstruktur entsteht durch die Randbedingungen (Steigung 1 oder 2) bei der Dynamischen Programmierung.

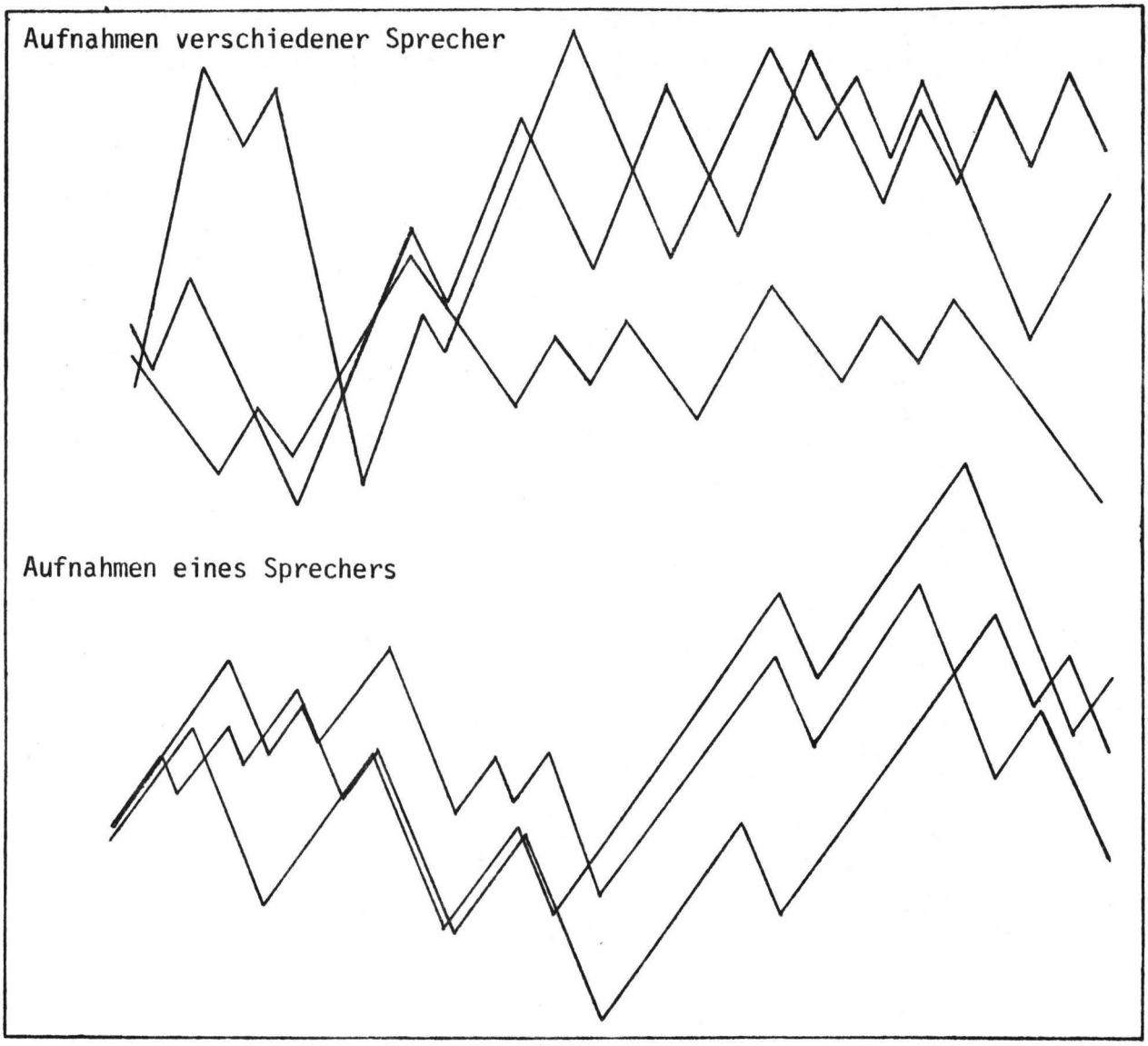


Bild 10: Inter- und Intrasprechervarianz der Verzerrungskontur

Es hat sich gezeigt, daß diese Verzerrungskontur relativ wenig zur Sprecherunterscheidung beitragen kann, da ihre Streuungen sehr groß sind. Entsprechend hoch sind die damit erzielten Fehlerraten (vgl. Bild 12). Auch eine Glättung der Sägezahnstruktur brachte keine wesentlich besseren Ergebnisse.

3.4 Kanalkonturen

Es wurde nun untersucht, was einzelne Kanalkonturen, d.h. Energieanteile bestimmter Frequenzbereiche über der Zeit (in Bild 1 einzelne Zeilen der Matrix), zur Sprecherunterscheidung beitragen. Gemäß der bei der Analyse verwendeten Filterbank ergeben sich für die einzelnen Kanäle die in Tabelle 1 angegebenen Frequenzbereiche.

Tabelle 1:

Kanal:	Frequenzbereich:	Fehler:
7	316 - 383	14%
8	383 - 464	9%
9	464 - 562	6%
10	562 - 681	5%
11	681 - 835	6%
12	835 - 1000	4%
13	1000 - 1211	19%
14	1211 - 1463	6%
15	1463 - 1778	4%
16	1778 - 2154	6%
17	2154 - 2610	4%
18	2610 - 3162	4%
19	3162 - 3831	5%
20	3831 - 4642	4%
21	4642 - 5623	12%

Die Fehlerraten, die diese Kanalkonturen bei der Klassifizierung der erwähnten Stichprobe ergaben, sind ebenfalls in Tabelle 1 zusammengestellt. Fast alle verwendeten Konturen zwischen 500 Hz und 4 kHz schwanken zwischen 4 und 6%. Einige Testergebnisse deuten allerdings darauf hin, daß der Frequenzbereich zwischen 1,5 kHz und 4 kHz sprechertypischer ist als der darunterliegende Frequenzbereich.

Bild 11 zeigt Realisierungen der Kanalkontur 15. Auch hier wurden einerseits die linear, andererseits die nichtlinear normalisierten Konturen getestet. Die in Tabelle 1 angegebenen Fehlerraten beziehen sich auf die letzteren. Ohne Anwendung der nichtlinearen Zeitnormalisierung auf den Kanalverlauf waren die Ergebnisse signifikant schlechter (ca. 6-10%).
In Bild 12 ist zum Vergleich neben der mittleren Fehlerrate einer Kanalkontur die Fehlerrate des Langzeitspektrums eingetragen, die ebenfalls ca. 5% beträgt.

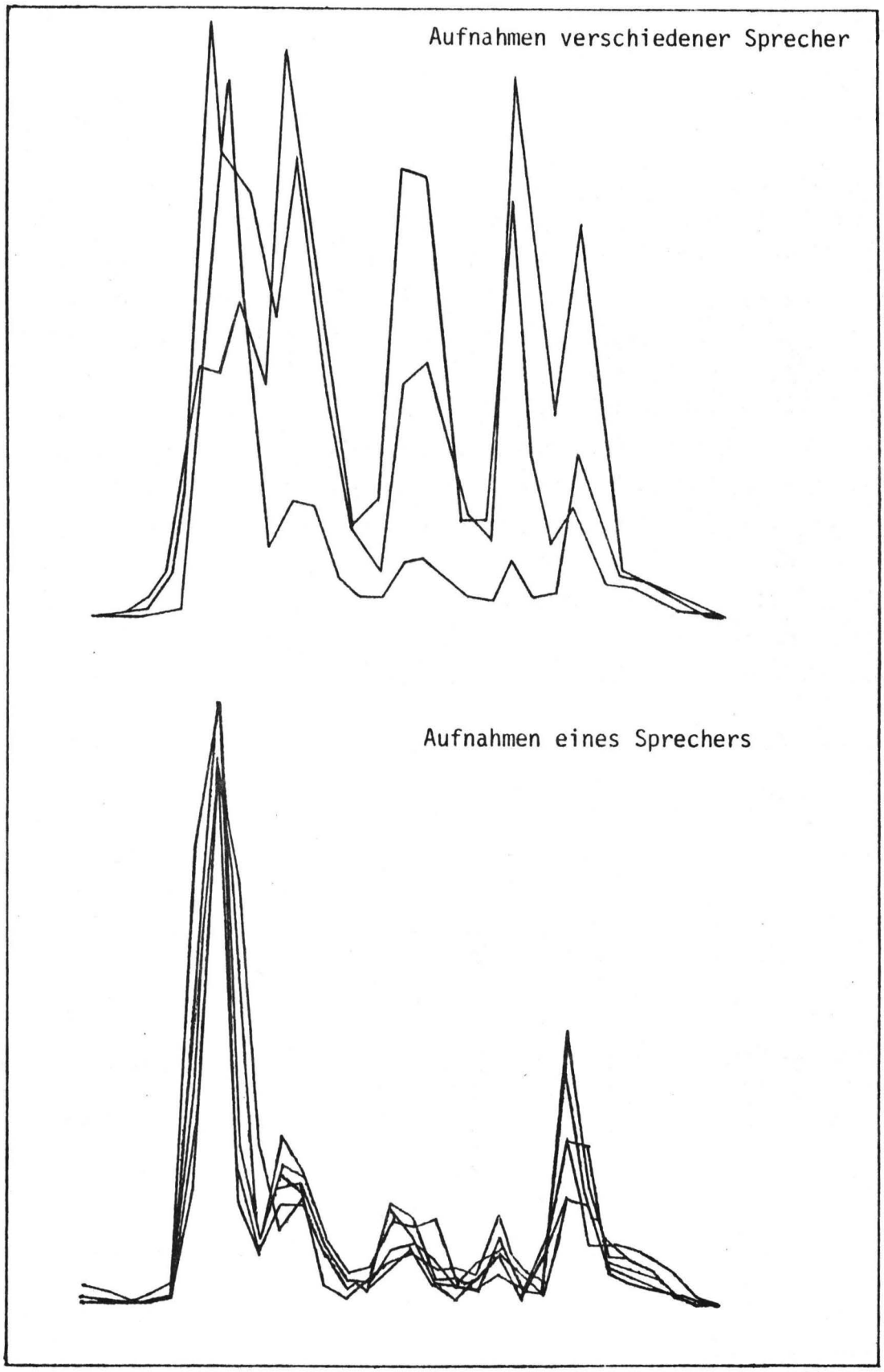


Bild 11: Inter- und Intrasprechervarianz einer Kanalkontur

3.5 Kombination der Konturen

Wie schon erwähnt, ermöglicht das implementierte Sprecher-Verifizierungs-System verschiedene Merkmalsvektoren zu kombinieren und einzeln zu gewichten.So wurden verschiedene Kombinationen getestet. Entsprechend den Fehlerraten der einzelnen Konturen wurde u.a. eine Kombination aus Pegelkontur
Stationaritätskontur
Verzerrungskontur
Konalkontur 14 - 19
zusammengestellt, wobei die ersten drei Merkmalsvektoren geringer gewichtet wurden. Die nichtlinear normalisierten Kanalkonturen wurden so ausgwählt, daß im wesentlichen nur Teile des beim Telefon verwendeten Frequenzbereichs (0,3-3,4 kHz) benutzt wurden. Derartige Kombinationen von Merkmalsvektoren können die erwähnte Stichprobe ohne Fehler klassifizieren.

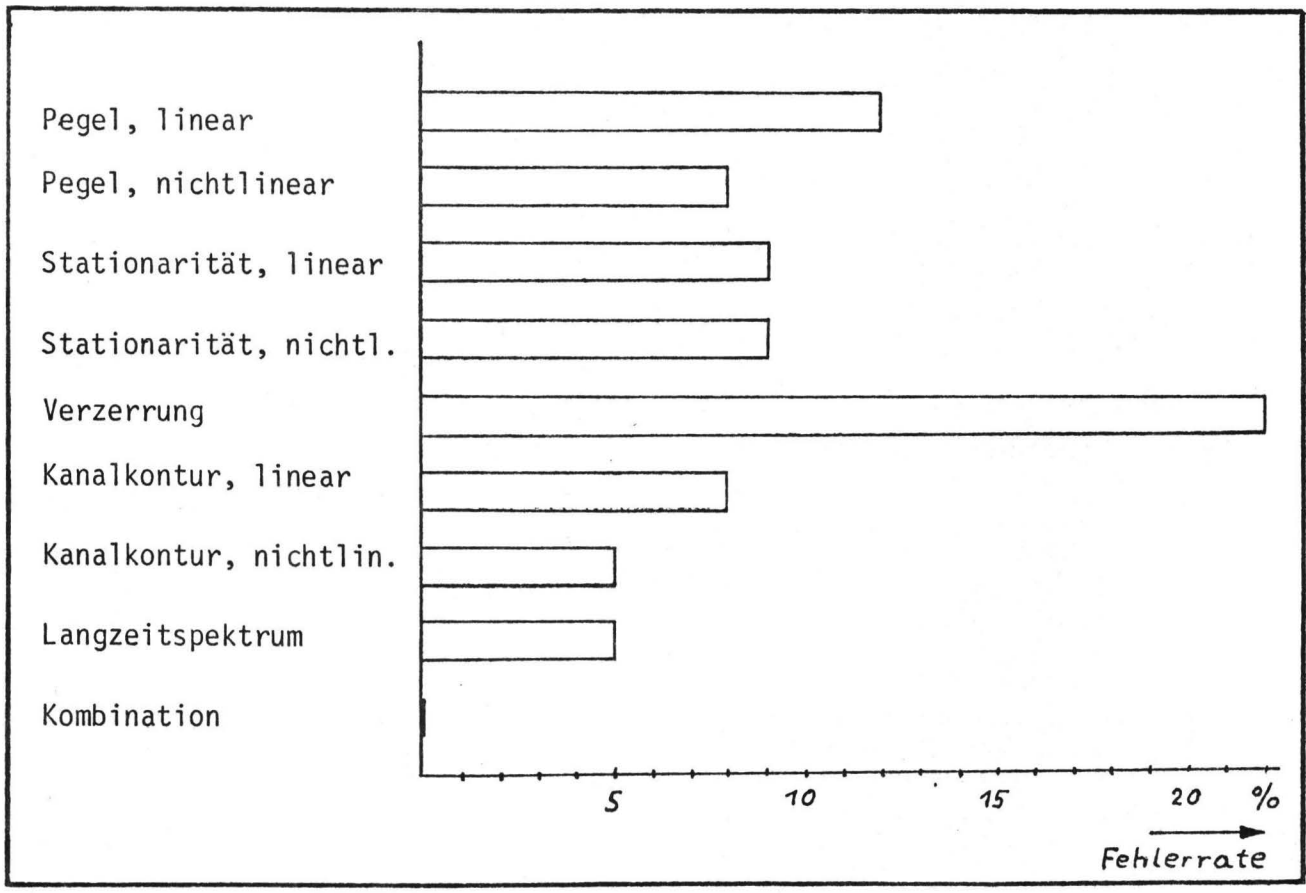


Bild 12: Güte der Teilvektoren

4. Zusammenfassung

In diesem Bericht wurden verschiedene nichtlineare Zeitnormalisierungsverfahren verglichen.
Im vorliegenden Fall hat sich dafür die Dynamische Programmierung als geeignet erwiesen. Eine derartige nichtlineare Anpassung bietet nicht nur die Möglichkeit, für die Segmentanalyse bestimmte Kurzzeitspektren zu lokalisieren, sondern verbessert auch bei der Konturanalyse die Merkmalsvektoren.

Einzelne Konturen bestimmter Frequenzbereiche sind wie die einzelnen Spektren zu bestimmten Zeiten (vergl. /3/) gute Merkmalsvektoren zur Sprecherunterscheidung. Insbesondere bei der Telefonübertragung können sie als wesentliche Merkmale verwendet werden, da die zeitliche Struktur durch den Übertragungsweg kaum beeinflußt wird.

5. Literatur

1. JESORSKY, P.: "Möglichkeiten und Grenzen der automatischen Sprechererkennung", Technischer Bericht Nr. 204 des Heinrich-Hertz-Instituts, Berlin, 1979

2. TALMI, M.: "Spektrale Vorverarbeitung von Sprachsignalen", Technischer Bericht Nr. 205 des Heinrich-Hertz-Instituts für Nachrichtentechnik, in Vorbereitung

3. HÖFKER, U.: "Die Merkmalsgewinnung bei der Sprechererkennung", Technischer Bericht Nr. 207 des Heinrich-Hertz-Instituts für Nachrichtentechnik, in Vorbereitung

4. DODDINGTON, G.R.: "Speaker Verification" Technical Report, RADC-TR, April 1974

Heinrich-Hertz-Institut
für Nachrichtentechnik
Berlin GmbH