

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

EP 1 273 003 B1

(12)

EUROPÄISCHE PATENTSCHRIFT

(45) Veröffentlichungstag und Bekanntmachung des Hinweises auf die Patenterteilung:

07.12.2005 Patentblatt 2005/49

(51) Int Cl.7: **G10L 13/08**

(86) Internationale Anmeldenummer:

PCT/DE2001/001394

(21) Anmeldenummer: **01940136.3**

(87) Internationale Veröffentlichungsnummer:

WO 2001/078063 (18.10.2001 Gazette 2001/42)

(22) Anmeldetag: **09.04.2001**

(54) **VERFAHREN UND VORRICHTUNG ZUM BESTIMMEN PROSODISCHER MARKIERUNGEN**

METHOD AND DEVICE FOR THE DETERMINATION OF PROSODIC MARKERS

PROCEDE ET DISPOSITIF DE DETERMINATION DE MARQUAGES PROSODIQUES

(84) Benannte Vertragsstaaten:

DE FR GB IT

(56) Entgegenhaltungen:

GB-A- 2 325 599

(30) Priorität: **12.04.2000 DE 10018134**

(43) Veröffentlichungstag der Anmeldung:

08.01.2003 Patentblatt 2003/02

(73) Patentinhaber: **SIEMENS**

AKTIENGESELLSCHAFT

80333 München (DE)

(72) Erfinder:

- **HOLZAPFEL, Martin**
80933 München (DE)
- **MÜLLER, Achim**
81739 München (DE)

- **YUKIKO YAMAGUCHI ET AL: "A NEURAL NETWORK APPROACH TO MULTI-LANGUAGE TEXT-TO-SPEECH SYSTEM" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP),JP,TOKYO, ASJ, 18. November 1990 (1990-11-18), Seiten 325-328, XP000503375**
- **CHEN S -H ET AL: "AN RNN-BASED PROSODIC INFORMATION SYNTHESIZER FOR MANDARIN TEXT-TO-SPEECH" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING,US,IEEE INC. NEW YORK, Bd. 6, Nr. 3, 1. Mai 1998 (1998-05-01), Seiten 226-239, XP000785353 ISSN: 1063-6676**
- **MULLER A F ET AL: "Robust generation of symbolic prosody by a neural classifier based on autoassociators" IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. PROCEEDINGS (ICASSP 2000), ISTANBUL, TURKEY, 5. - 9. Juni 2000, Seiten 1285-1288 vol.3, XP002173691 IEEE, Piscataway, NJ, USA ISBN: 0-7803-6293-4**
- **PALMER D D ET AL: "ADAPTIVE MULTILINGUAL SENTENCE BOUNDARY DISAMBIGUATION" COMPUTATIONAL LINGUISTICS,CAMBRIDGE, MA,US, Bd. 23, Nr. 2, Juni 1997 (1997-06), Seiten 241-267, XP000997368**

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist. (Art. 99(1) Europäisches Patentübereinkommen).

EP 1 273 003 B1

Beschreibung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zum Bestimmen prosodischer Markierungen und eine Vorrichtung zur Umsetzung des Verfahrens.

[0002] Bei der Aufbereitung von unbekanntem Text für die Sprachsynthese in einem TTS-System, ('text to speech'-Systemen) bzw. Text/Sprache-Umsetzungssystemen, ist ein wesentlicher Schritt die Aufbereitung und Strukturierung des Textes für die nachfolgende Generierung der Prosodie. Um prosodische Parameter für Sprachsynthesesysteme zu erzeugen, wird ein zweistufiger Ansatz verfolgt. Dabei werden in der ersten Stufe zunächst prosodische Marker bzw. prosodische Markierungen erzeugt, die dann in der zweiten Stufe in physikalische Parameter umgesetzt werden.

[0003] Als prosodische Markierungen können insbesondere Phrasengrenzen und Wortakzente (pitch-accent) dienen. Unter Phrasen werden Gruppierungen von Wörtern verstanden, die innerhalb eines Textes in der Regel zusammen gesprochen werden, also ohne dazwischen eingeschoben liegende Sprechpausen. Sprechpausen liegen erst an den jeweiligen Enden der Phrasen, den Phrasengrenzen, an. Durch das Einlegen derartiger Pausen an den Phrasengrenzen der synthetisierten Sprache wird deren Verständlichkeit und Natürlichkeit wesentlich gesteigert.

[0004] In der Stufe 1 eines derartigen zweistufigen Ansatzes bereiten sowohl die stabile Vorhersage bzw. Bestimmung von Phrasengrenzen als auch die von Akzenten Probleme.

[0005] In einer Veröffentlichung ist unter dem Titel "A hierarchical stochastic model for automatic prediction of prosodic boundary location" von M. Ostendorf und N. Veilleux in Computational Linguistics, 1994, ein Verfahren veröffentlicht worden, in dem zur Bestimmung von Phrasengrenzen "Classification and Regression Trees" (CART) verwendet werden. Die Initialisierung eines solchen Verfahrens erfordert ein hohes Maß an Expertenwissen. Der Aufwand steigt bei diesem Verfahren mit der angestrebten Genauigkeit überproportional.

[0006] Auf der Konferenz Eurospeech 1997 ist unter dem Titel "Assigning phase breaks from part-of-speech sequences" von Alan W. Black und Paul Taylor ein Verfahren veröffentlicht worden, in dem die Phrasengrenzen mit einem "Hidden-Markov-Modell" (HMM) bestimmt werden. Zur Erzielung einer guten Vorhersage-Genauigkeit für eine Phrasengrenze ist ein Trainingstext mit beträchtlichem Umfang notwendig. Die Erstellung dieser Trainingstexte ist teuer, da hierzu Expertenwissen notwendig ist.

[0007] Der Artikel "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", von Sin-Hong Chen et al. in "IEEE Transactions on Speech and Audio Processing", US, IEEE Inc. New York, Bd. 6, Nr. 3, 1. Mai 1998, Seiten 226-239, offenbart ein Verfahren zum Bestimmen prosodischer Markierungen, bei dem die Markierungen auf der Basis linguistischer Kategorien durch ein neuronales Feed-Forward-Netzwerk bestimmt werden.

[0008] Aus Yuyiko Yamaguchi et al.: "A Neural Network Approach to Multi-Language Text-to-Speech System", in: "Proceedings of the International Conference on Spoken Language Processing (ICSLP)", JP, Tokyo, ASJ, 18. November 1990, Seiten 325-328, ist ein Verfahren bekannt, bei dem mit Hilfe eines vernetzten neuronalen Feed-Forward-Netzwerks syntaktische Grenzen bestimmt werden.

[0009] Demnach ist es die Aufgabe der vorliegenden Erfindung, ein Verfahren zur Aufbereitung und Strukturierung eines unbekannten gesprochenen Textes zu schaffen, das mit einem kleineren Trainingstext trainiert werden kann und etwa ähnliche Erkennungsraten wie bekannte Verfahren erzielt, die mit größeren Texten trainiert werden.

[0010] Diese Aufgabe wird durch ein Verfahren gemäß Patentanspruch 1 und ein neuronales Netzwerk gemäß Patentanspruch 6 gelöst.

[0011] Demnach werden in einem erfindungsgemäßen Verfahren prosodische Markierungen durch ein neuronales Netzwerk auf der Basis linguistischer Kategorien bestimmt. In Abhängigkeit von der jeweiligen Sprache eines Textes sind Unterteilungen der Wörter in verschiedene linguistische Kategorien bekannt. Bei der deutschen Sprache werden im Rahmen dieser Erfindung beispielsweise 14 Kategorien, für die englische Sprache z.B. 23 Kategorien vorgesehen. Unter Kenntnis dieser Kategorien wird ein neuronales Netzwerk so trainiert, daß es Strukturen erkennen kann und so auf der Basis von Gruppierungen von z.B. 3 bis 15 aufeinander folgenden Wörtern eine prosodische Markierung vorhersagt bzw. bestimmt.

[0012] Weiterhin wird für ein erfindungsgemäßes Verfahren ein zweistufiger Ansatz gewählt, der das Erfassen der Eigenschaften jeder prosodischen Markierung durch neuronale Autoassoziatoren und das Auswerten der von jedem der Autoassoziatoren ausgegebenen detaillierten Ausgangsinformationen, die als sogenannter Fehler-Vektor vorliegt, in einem neuronalen Klassifikator beinhaltet.

[0013] Durch die erfindungsgemäße Anwendung von neuronalen Netzen wird ermöglicht, bei der Erzeugung prosodischer Parameter für Sprachsynthesysteme Phrasengrenzen genau vorherzusagen.

[0014] Das erfindungsgemäße neuronale Netz ist robust gegenüber "wenigem" bzw. einem geringen Umfang von Trainingsmaterial (engl.: sparse training material).

[0015] Die Verwendung neuronaler Netzwerke gestattet zeit- und kostensparende Trainingsverfahren und eine flexible Anwendung eines erfindungsgemäßen Verfahrens und eine entsprechende Vorrichtung auf beliebige Sprachen. Es ist wenig zusätzlich aufbereitete Information und wenig Expertenwissen zum Initialisieren eines solchen Systems einer bestimmten Sprache erforderlich. Das erfindungsgemäße neuronale Netzwerk ist deshalb gut geeignet, um mit

einem multilingualen TTS-System Texte aus mehreren Sprachen zu synthetisieren. Da die erfindungsgemäßen neuronalen Netzwerke ohne Expertenwissen trainiert werden können, können sie kostengünstiger als bekannte Verfahren zum Bestimmen von Phrasengrenzen initialisiert werden.

[0016] In einer Weiterbildung umfaßt die zweistufige Struktur mehrere Autoassoziatoren, die jeweils auf eine Phrasierungsstärke für alle auszuwertenden linguistischen Klassen trainiert werden. So sind Teile des neuronalen Netzwerkes klassenspezifisch ausgebildet. Das Trainingsmaterial ist in der Regel statistisch asymmetrisch ausgebildet, d. h., daß viele Wörter ohne Phrasengrenzen, aber nur wenige mit Phrasengrenzen vorhanden sind. Im Gegensatz zu Verfahren nach dem Stand der Technik wird eine Dominanz innerhalb eines neuronalen Netzes dadurch vermieden, daß ein klassenspezifisches Training der jeweiligen Autoassoziatoren durchgeführt wird.

[0017] Vorteilhafte Weiterbildungen eines erfindungsgemäßen Verfahrens sind der Gegenstand von Unteransprüchen.

[0018] Das vorliegende Verfahren wird im folgenden unter Bezugnahme auf die zugehörigen Zeichnungen näher erläutert.

[0019] In den Zeichnungen zeigt:

Fig. 1 schematisch ein neuronales Netzwerkwerk gemäß der Erfindung;

Fig. 2 eine Ausgabe bei einfacher Phrasierung anhand eines Beispieltexsts;

Fig. 3 ein Beispiel für eine Ausgabe mit ternärer Bewertung der Phrasierung anhand eines Textbeispiels;

Fig. 4 schematisch eine bevorzugte Ausführungsform eines neuronalen Netzwerkes;

Fig. 5 schematisch einen Autoassoziator während des Trainings (links) und während des Betriebs (rechts);

Fig. 6 schematisch ein Blockschaltbild des neuronalen Netzwerkes nach Fig. 4 mit den mathematischen Zusammenhängen; und

Fig. 7 schematisch einen erweiterten Autoassoziator, und

Fig. 8 ein Computersystem zum Ausführen des erfindungsgemäßen Verfahrens in einem Blockschaltbild.

[0020] In der Figur 1 ist schematisch ein erfindungsgemäßes neuronales Netzwerkwerk 1 mit einem Eingang 2, einer Zwischenschicht 3 und einem Ausgang 4 zum Bestimmen prosodischer Markierungen dargestellt. Der Eingang 2 ist aus neun Eingangsgruppen 5 zur Durchführung einer 'part-of-speech'- (POS-) Sequenz Untersuchung aufgebaut. Jede der Eingangsgruppe 5 umfaßt in Anpassung an die deutsche Sprache 14 Neuronen 6, die aus Gründen der Übersichtlichkeit nicht alle in Fig. 1 dargestellt sind. Es ist also je ein Neuron 6 für eine der linguistischen Kategorie vorhanden. Die linguistischen Kategorien sind beispielsweise folgendermaßen unterteilt:

Tabelle 1:

linguistische Kategorien	
Kategorie	Beschreibung
NUM	Numerale
VERB	Verben
VPART	Verbpartikel
PRON	Pronomen
PREP	Präpositionen
NOMEN	Nomen, Eigennamen
PART	Partikel
DET	Artikel
CONJ	Konjunktionen
ADV	Adverben
ADJ	Adjektive
PDET	PREP+DET
INTJ	Interjektionen

Tabelle 1: (fortgesetzt)

linguistische Kategorien	
Kategorie	Beschreibung
PUNCT	Satzzeichen

[0021] Der Ausgang 4 ist durch ein Neuron mit einem kontinuierlichen Verlauf ausgebildet, das bedeutet, daß die Ausgangswerte alle Werte eines bestimmten Zahlenbereiches, der z.B. alle reellen Zahlen zwischen 0 und 1 umfaßt, annehmen können.

[0022] Bei dem in Fig. 1 gezeigten Ausführungsbeispiel sind neun Eingangsgruppen 5 zum Eingeben der Kategorien der einzelnen Wörter vorgesehen. An die mittlere Eingangsgruppe 5a wird die Kategorie des Wortes angelegt, von dem bestimmt werden soll, ob am Ende des Wortes eine Phasengrenze vorliegt oder keine Phasengrenze vorliegt. An die vier Eingangsgruppen 5b auf der linken Seite der Eingangsgruppe 5a werden die Kategorien von den Vorläufern des zu untersuchenden Wortes und an die auf der rechten Seite angeordneten Eingangsgruppen 5c die Nachfolger des zu untersuchenden Wortes angelegt. Vorläufer sind alle Wörter, die im Kontext unmittelbar vor dem zu untersuchenden Wort angeordnet sind. Nachfolger sind alle Wörter, die im Kontext unmittelbar nachfolgend auf das zu untersuchende Wort angeordnet sind. Hierdurch wird mit dem erfindungsgemäßen neuronalen Netzwerk 1 nach Fig. 1 ein Kontext von max. neun Wörtern ausgewertet.

[0023] Bei der Auswertung wird die Kategorie des zu untersuchenden Wortes an die Eingangsgruppe 5a angelegt, das heißt, daß an das Neuron 6, das der Kategorie des Wortes entspricht, der Wert +1 und an die übrigen Neuronen 6 der Eingangsgruppe 5a der Wert -1 angelegt wird. In entsprechender Weise werden die Kategorien der vier zu dem zu untersuchenden Wort vorhergehenden bzw. nachfolgenden Wörter an die Eingangsgruppen 5b, bzw. 5c angelegt. Sollten keine entsprechenden Vorläufer bzw. Nachfolger vorhanden sein, wie es z.B. am Anfang und am Ende eines Textes der Fall ist, werden an die Neuronen 6 der entsprechenden Eingangsgruppen 5b, 5c der Wert 0 angelegt.

[0024] Eine weitere Eingangsgruppe 5d ist zum Eingeben der vorhergegangenen Phrasengrenzen vorgesehen. An dieser Eingangsgruppe 5d können die letzten neun Phrasengrenzen eingegeben werden.

[0025] Für die deutsche Sprache - mit 14 linguistischen Kategorien - weist der Eingangsraum eine beachtliche Dimension m von 135 ($m = 9 \cdot 14 + 9$) auf. Eine zweckmäßige Unterteilung der linguistischen Kategorien der englischen Sprache umfaßt 23 Kategorien, so daß die Dimension des Eingangsraumes 216 beträgt. Die Eingangsdaten bilden einen Eingangsvektor x mit der Dimension m .

[0026] Das erfindungsgemäße neuronale Netzwerk wird mit einer Trainingsdatei trainiert, die einen Text und die Informationen zu den Phrasengrenzen des Textes umfaßt. Diese Phrasengrenzen können rein binäre Werte enthalten, das heißt, lediglich Informationen, ob eine Phrasengrenze vorliegt oder ob keine Phrasengrenze vorliegt. Wird das neuronale Netzwerk mit einer derartigen Trainingsdatei trainiert, so ist die Ausgabe am Ausgang 4 binär. Der Ausgang 4 erzeugt an sich kontinuierliche Ausgangswerte, die jedoch mittels einer Schwellwertentscheidung diskreten Werten zugeordnet werden.

[0027] In Fig. 2 ist ein Beispielsatz dargestellt, der hinter den Begriffen "Wort" und "Phrasengrenze" jeweils eine Phrasengrenze aufweist. Hinter den anderen Wörtern dieses Beispielsatzes gibt es keine Phrasengrenze.

[0028] Für bestimmte Anwendungen ist es vorteilhaft, wenn die Ausgabe nicht nur binäre Werte, sondern mehrstufige Werte enthält, das heißt, daß Informationen über die Stärke der Phrasengrenze berücksichtigt werden. Hierzu ist das neuronale Netzwerk mit einer Trainingsdatei zu trainieren, die mehrstufige Informationen zu den Phrasengrenzen umfaßt. Die Abstufung kann von zwei Stufen bis an sich beliebig viele Stufen umfassen, so daß eine quasi kontinuierliche Ausgabe erzielt werden kann.

[0029] In Fig. 3 ist ein Beispielsatz mit einer dreistufigen Auswertung mit den Ausgangswerten 0 für keine Phrasengrenze, 1 für eine primäre Phrasengrenze und 2 für eine sekundäre Phrasengrenze dargestellt. Nach dem Begriff "sekundären" befindet sich eine sekundäre Phrasengrenze und nach den Begriffen "Phrasengrenze" und "erforderlich" eine primäre Phrasengrenze.

[0030] In Fig. 4 ist eine bevorzugte Ausführungsform des erfindungsgemäßen neuronalen Netzes dargestellt. Dieses neuronale Netz umfaßt wiederum einen Eingang 2, der in Fig. 4 lediglich schematisch als ein Element dargestellt ist, aber genauso wie der Eingang 2 aus Fig. 1 aufgebaut ist. Die Zwischenschicht 3 besteht bei diesem Ausführungsbeispiel aus mehreren Autoassoziatoren 7 (AA1, AA2, AA3) die jeweils ein Modell für eine vorbestimmte Phrasierungsstärke darstellen. Die Autoassoziatoren 7 sind Teilnetzwerke, die zum Detektieren einer bestimmten Phrasierungsstärke trainiert werden. Der Ausgang der Autoassoziatoren 7 ist mit einem Klassifikator 8 verbunden. Der Klassifikator 8 ist ein weiteres neuronales Teilnetzwerk, das auch den anhand von Fig. 1 bereits beschriebenen Ausgang umfaßt.

[0031] Das in Fig. 4 gezeigte Ausführungsbeispiel umfaßt drei Autoassoziatoren, wobei mit jedem Autoassoziator eine bestimmte Phrasierungsstärke detektiert werden kann, so daß dieses Ausführungsbeispiel zur Detektion zweier unterschiedlicher Phrasierungsstärken und dem Vorliegen keiner Phrasierungsgrenze geeignet ist.

[0032] Jeder Autoassoziator wird mit den Daten der Klasse, die er darstellt, trainiert. D.h., daß jeder Autoassoziator mit den zu der von ihm repräsentierten Phrasierungsstärke gehörenden Daten trainiert wird.

[0033] Die Autoassoziatoren bilden den m-dimensionalen Eingangsvektor $\mathbf{x}$ auf einen n-dimensionalen Vektor $\mathbf{z}$ ab, wobei $n \ll m$ gilt. Der Vektor $\mathbf{z}$ wird auf einen Ausgangsvektor $\mathbf{x}'$ abgebildet. Die Abbildungen erfolgen mittels Matrizen $\mathbf{w}_1 \in \mathbb{R}^{n \times m}$ und $\mathbf{w}_2 \in \mathbb{R}^{n \times m}$. Die gesamte in den Autoassoziatoren ausgeführte Abbildung kann durch folgende Formel dargestellt werden:

$$\mathbf{x}' = \mathbf{w}_2 \tanh(\mathbf{w}_1 \cdot \mathbf{x}),$$

wobei $\tanh$ elementweise angewendet wird.

[0034] Die Autoassoziatoren werden so trainiert, daß ihre Ausgangsvektoren $\mathbf{x}'$ so genau wie möglich mit den Eingangsvektoren $\mathbf{x}$ übereinstimmen (Fig. 5 linke Seite). Hierdurch wird die Information des m-dimensionalen Eingangsvektors $\mathbf{x}$ auf den n-dimensionalen Vektor $\mathbf{z}$ komprimiert. Hierbei wird davon ausgegangen, daß keine Informationen verlorengehen und das Modell die Eigenschaften der Klasse erfaßt. Das Kompressionsverhältnis m:n der einzelnen Autoassoziatoren kann unterschiedlich sein.

[0035] An die Eingangs- und Ausgangsseite der einzelnen Autoassoziatoren werden beim Training nur die Eingangsvektoren $\mathbf{x}$ angelegt, die den Zuständen entsprechen, bei denen die den jeweiligen Autoassoziatoren zugeordneten Phrasengrenzen auftreten.

[0036] Beim Betrieb wird für jeden Autoassoziator ein Fehler-Vektor $\mathbf{e}_{\text{rec}} = (\mathbf{x} - \mathbf{x}')^2$ berechnet (Fig. 5, rechte Seite). Die Quadrierung erfolgt hierbei elementweise. Dieser Fehler-Vektor $\mathbf{e}_{\text{rec}}$ ist ein "Abstandsmaß", das dem Abstand des Vektors $\mathbf{x}'$ zum Eingangsvektor $\mathbf{x}$ entspricht und somit indirekt proportional zur Wahrscheinlichkeit ist, daß die dem jeweiligen Autoassoziator zugeordnete Phrasengrenze vorliegt.

[0037] Das vollständige die Autoassoziatoren und den Klassifikator umfassende neuronale Netzwerk ist schematisch in Fig. 6 dargestellt. Es zeigt Autoassoziatoren 7 für k Klassen.

[0038] Die Elemente p_i des Ausgangsvektors $\mathbf{p}$ werden gemäß folgender Formel berechnet:

$$p_i = \frac{e^{(\mathbf{x} - A_i(\mathbf{x}))^T \text{diag}(\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_m^{(1)}) (\mathbf{x} - A_i(\mathbf{x}))}}{\sum_{j=1}^k e^{(\mathbf{x} - A_j(\mathbf{x}))^T \text{diag}(\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_m^{(1)}) (\mathbf{x} - A_j(\mathbf{x}))}},$$

wobei $A_i(\mathbf{x}) = \mathbf{w}_2^{(i)} \tanh(\mathbf{w}_1^{(i)} \mathbf{x})$ gilt und $\tanh$ als elementweise Operation ausgeführt wird und $\text{diag}(\mathbf{w}_1^{(i)}, \dots, \mathbf{w}_m^{(i)}) \in \mathbb{R}^{m \times m}$ eine Diagonalmatrix mit den Elementen $(\mathbf{w}_1^{(i)}, \dots, \mathbf{w}_m^{(i)})$ darstellt.

[0039] Die einzelnen Elemente p_i des Ausgangsvektors $\mathbf{p}$ geben die Wahrscheinlichkeit an, mit welcher eine Phrasengrenze am Autoassoziator i detektiert worden ist.

[0040] Wenn die Wahrscheinlichkeit p_i größer als 0,5 ist, wird dies als Vorliegen einer entsprechenden Phrasengrenze i bewertet. Ist die Wahrscheinlichkeit p_i kleiner als 0,5, so bedeutet dies, daß die Phrasengrenze i hier nicht vorliegt.

[0041] Hat der Ausgangsvektor $\mathbf{p}$ mehr als zwei Elemente p_i , so ist es zweckmäßig, den Ausgangsvektor $\mathbf{p}$ derart zu bewerten, daß diejenige Phrasengrenze vorliegt, deren Wahrscheinlichkeit p_i am größten im Vergleich zu den übrigen Wahrscheinlichkeiten p_i des Ausgangsvektors $\mathbf{p}$ ist.

[0042] In einer Weiterbildung der Erfindung kann es zweckmäßig sein, falls eine Phrasengrenze ermittelt wird, deren Wahrscheinlichkeit p_i im Bereich um 0,5, z.B. im Bereich von 0,4 bis 0,6, liegt, eine weitere Routine durchzuführen, mit der das Vorliegen der Phrasengrenze überprüft wird. Diese weitere Routine kann auf einem regelgetriebenen als auch auf einem datengetriebenen Ansatz beruhen.

[0043] Beim Training mit einer Trainingsdatei, die entsprechende Phrasierungsinformationen umfaßt, werden in einer ersten Trainingsphase die einzelnen Autoassoziatoren 7 jeweils auf ihre vorbestimmte Phrasierungsstärke trainiert. Wie es oben angegeben ist, werden hierbei an die Eingangs- und die Ausgangsseite der einzelnen Autoassoziatoren 7 die Eingangsvektoren $\mathbf{x}$ angelegt, die der Phrasengrenze entsprechen, die dem jeweiligen Autoassoziator zugeordnet ist.

[0044] In einer zweiten Trainingsphase werden die Gewichtungselemente der Autoassoziatoren 7 festgehalten und der Klassifikator 8 trainiert. An die Eingangsseite des Klassifikators 8 werden die Fehler-Vektoren $\mathbf{e}_{\text{rec}}$ der Autoassoziatoren und an der Ausgangsseite die Vektoren, die die Werte für die unterschiedlichen Phrasengrenzen enthalten,

angelegt. In dieser Trainingsphase lernt der Klassifikator aus den Fehler-Vektoren die Ausgangsvektoren **p** zu bestimmen.

[0045] In einer dritten Trainingsphase wird eine Feineinstellung aller Gewichtungselemente des gesamten neuronalen Netzwerkes (der *k* Autoassoziatoren und des Klassifikators) durchgeführt.

[0046] Durch die oben beschriebene Architektur eines neuronalen Netzwerkes mit mehreren jeweils auf eine bestimmte Klasse trainierten Modellen (hier: den Autoassoziatoren) und einem übergeordneten Klassifikator ist es möglich, einen Eingangsvektor mit sehr großer Dimension auf einen Ausgangsvektor mit kleiner Dimension bzw. ein Skalar zuverlässig korrekt abzubilden. Diese Netzwerkarchitektur kann auch vorteilhaft bei anderen Anwendungen eingesetzt werden, bei welchen Elemente unterschiedlicher Klassen behandelt werden müssen. So kann es z.B. zweckmäßig sein, diese Netzwerkarchitektur auch bei der Spracherkennung zum Detektieren von Wort und/oder Satzgrenzen einzusetzen. Hierfür sind die Eingangsdaten entsprechend anzupassen.

[0047] Der in Figur 6 gezeigte Klassifikator 8 weist Gewichtungsmatrizen GW auf, die jeweils einem Autoassoziator 7 zugeordnet sind. Die dem *i*-ten Autoassoziator 7 zugeordnete Gewichtungsmatrix GW weist in der *i*-ten Zeile Gewichtungsfaktoren w_n auf. Die übrigen Elemente der Matrix sind gleich Null. Die Anzahl der Gewichtungsfaktoren w_n entspricht der Dimension des Eingangsvektors, wobei jeweils ein Gewichtungselement w_n in Beziehung zu einer Komponente des Eingangsvektors steht. Besitzt ein Gewichtungselement w_n einen größeren Wert als die übrigen Gewichtungselemente w_n der Matrize, so bedeutet dies, daß die korrespondierende Komponente des Eingangsvektors von großer Bedeutung für die Bestimmung der Phrasengrenze, die durch den Autoassoziator ermittelt wird, dem die entsprechende Gewichtungsmatrix GW zugeordnet ist.

[0048] Bei einer bevorzugten Ausführungsform werden erweiterte Autoassoziatoren verwendet (Fig. 7), die eine bessere Erfassung von Nichtlinearitäten erlauben. Diese erweiterten Autoassoziatoren führen folgende Abbildung aus:

$$x' = w_2 \tanh(\cdot) + w_3 (\tanh(\cdot))^2,$$

wobei $(\cdot) := (\mathbf{w}_1 \cdot \mathbf{x})$ gilt, und die Quadrierung $(\cdot)^2$ und der tanh elementweise ausgeführt werden.

[0049] Bei Versuchen ist ein erfindungsgemäßes neuronales Netzwerk mit einem vorbestimmten englischen Text trainiert worden. Der gleiche Text wurde zum Trainieren eines HMM-Erkenner verwendet. Als Leistungskriterien wurden im Betrieb die Prozentzahl der korrekt erkannten Phrasengrenzen (B-corr), der insgesamt korrekt beurteilten Wörter, unabhängig davon, ob eine oder keine Phrasengrenze folgt (Gesamt), und der nicht-korrekt erkannten Wörter ohne Phrasengrenze (NB-ncorr) ermittelt. Bei diesen Versuchen wurde ein neuronales Netzwerk mit den Autoassoziatoren nach Fig. 6 und ein neuronales Netzwerk mit den erweiterten Autoassoziatoren verwendet. Es wurden folgende Ergebnisse erzielt:

Tabelle 2

	B-corr	Gesamt	NB-ncorr
erw. Autoass.	80,33%	91,68%	4,72%
Autoass.	78,10%	90,95%	3,93
HMM	79,48%	91,60%	5,57%

[0050] Die in der Tabelle aufgeführten Ergebnisse zeigen, daß die erfindungsgemäßen neuronalen Netzwerke bzgl. der korrekt erkannten Phrasengrenzen und der insgesamt korrekt erkannten Wörter etwa gleiche Ergebnisse wie ein HMM-Erkenner liefern. Jedoch sind die erfindungsgemäßen neuronalen Netzwerke bzgl. der fehlerhaft detektierten Phrasengrenzen, an Stellen, wo es an sich keine Phrasengrenze gibt, wesentlich besser als der HMM-Erkenner. Diese Art von Fehler ist bei der Sprache-zu-Text-Umsetzung besonders gravierend, da diese Fehler eine dem Zuhörer sofort auffallende Fehlbetonung erzeugen.

[0051] In weiteren Versuchen wurde eines der erfindungsgemäßen neuronalen Netzwerke mit einem Bruchteil des bei den obigen Versuchen verwendeten Trainingstextes (5%, 10%, 30%, 50%) trainiert. Hierbei wurden folgende Ergebnisse erzielt:

Tabelle 3

Bruchteil des Trainingstextes	B-corr	Gesamt	NB-ncorr
5%	70,50%	89,96%	4,65%

Tabelle 3 (fortgesetzt)

Bruchteil des Trainingstextes	B-corr	Gesamt	NB-ncorr
10%	75,00%	90,76%	4,57%
30%	76,30%	91,48%	4,16%
50%	78,01%	91,53%	4,44%

[0052] Mit Bruchteilen von 30% und 50% des Trainingstextes wurden ausgezeichnete Erkennungsraten erzielt. Mit einem Bruchteil von 10% und 5% des ursprünglichen Trainingstextes sind zufriedenstellende Erkennungsraten erzielt worden. Dies zeigt, daß die erfindungsgemäßen neuronalen Netzwerke auch bei geringem Trainingsumfang gute Erkennungsraten liefern. Dies stellt einen wesentlichen Fortschritt gegenüber bekannten Phrasengrenzen-Erkennungsmethoden dar, da das Aufbereiten von Trainingsmaterial kostenintensiv ist, da hier Expertenwissen eingesetzt werden muß.

[0053] Das oben beschriebene Ausführungsbeispiel weist k Autoassoziatoren auf. Für eine präzise Bewertung der Phrasengrenzen kann es zweckmäßig sein, eine große Anzahl Autoassoziatoren zu verwenden, wobei bis zu 20 Autoassoziatoren zweckmäßig sein können. Hierdurch wird eine quasi kontinuierlicher Verlauf der Ausgabewerte erzielt.

[0054] Die oben beschriebenen neuronalen Netzwerke sind als Computerprogramme realisiert, die selbständig auf einem Computer zum Umsetzen der linguistischen Kategorie eines Textes in dessen prosodischer Marker ablaufen. Sie stellen somit ein automatisch ausführbares Verfahren dar.

[0055] Das Computerprogramm kann auch auf einen elektronisch lesbaren Datenträger gespeichert werden und so auf ein anderes Computersystem übertragen werden.

[0056] Ein zur Anwendung des erfindungsgemäßen Verfahrens geeignetes Computersystem ist in Fig. 8 gezeigt. Das Computersystem 9 weist einen internen Bus 10 auf, der mit einem Speicherbereich 11, einer zentralen Prozessoreinheit 12 und einem Interface 13 verbunden ist. Das Interface 13 stellt über eine Datenleitung 14 eine Datenverbindung zu weiteren Computersystemen her. An dem internen Bus sind ferner eine akustische Ausgabeeinheit 15, eine grafische Ausgabeeinheit 16 und eine Eingabeeinheit 17 angeschlossen. Die akustische Ausgabeeinheit 15 ist mit einem Lautsprecher 18, die grafische Ausgabeeinheit 16 mit einem Bildschirm 19 und die Eingabeeinheit 17 mit einer Tastatur 20 verbunden. An das Computersystem 9 können über die Datenleitung 14 und das Interface 13 Text übertragen werden, die im Speicherbereich 11 abgespeichert werden. Der Speicherbereich 11 ist in mehrere Bereiche unterteilt, in denen Texte, Audiodateien, Anwendungsprogramme zum Durchführen des erfindungsgemäßen Verfahrens und weitere Anwendungs- und Hilfsprogramme gespeichert sind. Die als Textdatei abgespeicherten Texte werden mit vorbestimmten Programmpaketen analysiert und die jeweiligen linguistischen Kategorien der Wörter ermittelt. Danach werden mit dem erfindungsgemäßen Verfahren aus den linguistischen Kategorien die prosodischen Marker ermittelt. Diese prosodischen Marker werden wiederum in ein weiteres Programmpaket eingegeben, das die prosodischen Marker zur Erzeugung von Audiodateien verwendet, die über den internen Bus 10 zur akustischen Ausgabeeinheit 15 übertragen und von dieser am Lautsprecher 18 als Sprache ausgegeben werden.

[0057] In den hier dargestellten Beispielen ist nur eine Anwendung des Verfahrens auf die Vorhersage von Phrasengrenzen beschrieben worden. Das Verfahren kann bei ähnlichem Aufbau einer Vorrichtung und einem angepaßten Training aber auch zur Auswertung eines unbekannten Textes hinsichtlich einer Vorhersage von Betonungen, z.B. gemäß den international standardisierten ToBI-Labeln (tones and breaks indices), und/oder der Satzmelodie genutzt werden. Diese Anpassungen haben in Abhängigkeit von der jeweiligen Sprache des zu verarbeitenden Textes zu erfolgen, da die Prosodie stets sprachspezifisch ist.

Patentansprüche

1. Verfahren zum Bestimmen prosodischer Markierungen, wobei als prosodische Markierungen Phrasengrenzen und Wortakzente dienen, wobei

- auf der Basis linguistischer Kategorien prosodische Markierungen durch ein neuronales Netzwerk (1) bestimmt

gekennzeichnet durch die Schritten

- Erfassen der Eigenschaften jeder prosodischen Markierung **durch** neuronale Autoassoziatoren (7), die auf jeweils eine bestimmte prosodische Markierung trainiert sind, und
- Auswerten der von jedem der Autoassoziatoren (7) ausgegebenen Ausgangsinformationen in einem neuro-

nen Klassifikator (8).

2. Verfahren nach Anspruch 1,
dadurch gekennzeichnet,
5 **daß** als prosodische Markierungen Phrasengrenzen bestimmt und vorzugsweise auch ausgewertet und/oder bewertet werden.

3. Verfahren nach Anspruch 1 und/oder Anspruch 2,
dadurch gekennzeichnet,
10 **daß** am Eingang (2) des Netzwerkes (1) die linguistischen Kategorien von zumindest drei Wörtern eines zu synthetisierenden Textes angelegt werden.

4. Verfahren nach einem der vorhergehenden Ansprüche,
dadurch gekennzeichnet,
15 **daß** die Autoassoziatoren (1) für eine jeweilige vorbestimmte Phrasengrenze trainiert sind.

5. Verfahren nach Anspruch 4,
dadurch gekennzeichnet,
20 **daß** das Training des neuronalen Klassifikators (8) nach dem Training aller Autoassoziatoren (7) erfolgt.

6. Neuronales Netzwerk zum Bestimmen prosodischer Markierungen, wobei als prosodische Markierungen Phrasengrenzen und Wortakzente dienen, mit
 einem Eingang (2), einer Zwischenschicht (3) und einem Ausgang (4), wobei der Eingang zum Erfassen von linguistischen Kategorien von Wörtern eines zu analysierenden Textes ausgebildet ist,
25 **dadurch gekennzeichnet,**
 daß Eigenschaften jeder prosodischen Markierung durch neuronale Autoassoziatoren (7) erfassbar sind, die auf jeweils eine bestimmte prosodische Markierung trainiert sind, und
 daß die von jedem der Autoassoziatoren (7) ausgegebenen Ausgangsinformationen in einem neuronalen Klassifikator (8) auswertbar sind.
30

7. Neuronales Netzwerk nach Anspruch 6,
dadurch gekennzeichnet,
 daß die Zwischenschicht (3) zumindest zwei Autoassoziatoren (7) aufweist.

8. Neuronales Netzwerk nach Anspruch 6 oder 7,
dadurch gekennzeichnet,
35 **daß** der Eingang (2) Eingangsgruppen (5) aufweist, welche mehrere Neuronen (6) besitzen, die jeweils einer linguistischen Kategorie zugeordnet sind, und jede Eingangsgruppe zum Erfassen der linguistischen Kategorie eines Wortes des zu analysierenden Textes dient.
40

9. Neuronales Netzwerk nach einem der Ansprüche 6 bis 8,
dadurch gekennzeichnet,
 daß das Netzwerk zum Ausgeben einer binären, tertiären oder quaternären Phrasierungsstufe ausgebildet ist.

10. Neuronales Netzwerk nach einem der Ansprüche 7 bis 9,
dadurch gekennzeichnet,
45 **daß** das Netzwerk zum Ausgeben eines quasikontinuierlichen Phrasierungsbereichs ausgebildet ist.

11. Verfahren nach einem der Ansprüche 1 bis 5,
gekennzeichnet durch,
50 die Verwendung eines Neuronalen Netzwerkes nach einem der Ansprüche 6 bis 10.

12. Vorrichtung zum Bestimmen prosodischer Markierungen mit einem Computersystem (9), das einen Speicherbereich (11) aufweist, in dem ein Programm zum Ausführen eines Neuronalen Netzwerkes nach einem der Ansprüche 6 bis 10 gespeichert ist
55

Claims

1. Method for determining prosodic markers, phrase boundaries and word accents serving as prosodic markers,
 5 - prosodic markers being determined by a neural network (1) on the basis of linguistic categories,
 characterized by the following steps;
 - acquisition of the properties of each prosodic marker by neural autoassociators (7) which are trained to in
 10 each case one specific prosodic marker, and
 - evaluation of the output information output by each of the autoassociators (7) in a neural classifier (8).
2. Method according to Claim 1,
 15 **characterized**
 in that, as prosodic markers, phrase boundaries are determined and preferably also evaluated and/or assessed.
3. Method according to Claim 1 and/or Claim 2,
 characterized
 20 **in that** the linguistic categories of at least three words of a text to be synthesized are applied to the input (2) of
 the network (1).
4. Method according to one of the preceding claims,
 characterized
 25 **in that** the autoassociators (1) are trained for a respective predetermined phrase boundary.
5. Method according to Claim 4,
 characterized
 in that the neural classifier (8) is trained after the training of all the autoassociators (7).
- 30 6. Neural network for determining prosodic markers, phrase boundaries and word accents serving as prosodic mark-
 ers, having an input (2), an intermediate layer (3) and an output (4), the input being designed for acquiring linguistic
 categories of words of a text to be analysed,
 characterized
 in that properties of each prosodic marker can be acquired by neural autoassociators (7) which are trained to in
 35 each case one specific prosodic marker, and
 in that the output information output by each of the autoassociators (7) can be evaluated in a neural classifier (8) .
7. Neural network according to Claim 6,
 characterized
 40 **in that** the intermediate layer (3) has at least two autoassociators (7).
8. Neural network according to Claim 6 or 7,
 characterized
 45 **in that** the input (2) has input groups (5) having a plurality of neurons (6) each assigned to a linguistic category,
 and each input group serves for acquiring the linguistic category of a word of the text to be analysed.
9. Neural network according to one of Claims 6 to 8,
 characterized
 50 **in that** the network is designed for outputting a binary, tertiary or quaternary phrasing stage.
10. Neural network according to one of Claims 7 to 9,
 characterized
 in that the network is designed for outputting a quasi-continuous phrasing region.
- 55 11. Method according to one of Claims 1 to 5,
 characterized by
 the use of a neural network according to one of Claims 6 to 10.

12. Device for determining prosodic markers having a computer system (9), which has a memory area (11) in which a program for executing a neural network according to one of Claims 6 to 10 is stored.

5 Revendications

1. Procédé pour déterminer des marquages prosodiques, des limites de phrase et des accents de mot servant de marquages prosodiques,

- des marquages prosodiques étant déterminés par un réseau (1) neuronal sur la base de catégories linguistiques, **caractérisé par** les étapes suivantes
- enregistrement des propriétés de chaque marquage prosodique par des auto-associateurs (7) neuronaux, qui sont entraînés, sur respectivement un marquage prosodique défini, et
- analyse des informations de sortie éditées par chacun des auto-associateurs (7) dans un classificateur (8) neuronal.

2. Procédé selon la revendication 1, **caractérisé en ce que** des limites de phrases sont définies comme des marquages prosodiques et analysées et/ou évaluées également de préférence.

3. Procédé selon la revendication 1 et/ou la revendication 2, **caractérisé en ce que** à l'entrée (2) du réseau (1), on établit les catégories d'au moins trois mots d'un texte à synthétiser.

4. Procédé selon l'une quelconque des revendications, **caractérisé en ce que** les auto-associateurs (1) sont entraînés pour une limite de phrase prédéfinie respective.

5. Procédé selon la revendication 4, **caractérisé en ce que** l'entraînement du classificateur (8) neuronal s'effectue après l'entraînement de tous les auto-associateurs (7).

6. Réseau neuronal pour la détermination de marquages prosodiques, des limites de phrases et des accents de mot servant de marquages prosodiques, avec une entrée (2), une couche intermédiaire (3) et une sortie (4), l'entrée pour l'enregistrement de catégories linguistiques de mots d'un texte à analyser étant réalisée, **caractérisé en ce que** des propriétés de chaque marquage prosodique peuvent être enregistrées par des auto-associateurs (7) neuronaux, qui sont entraînés sur respectivement un marquage prosodique défini, et **en ce que** les informations de sortie éditées par chacun des auto-associateurs (7) peuvent être analysées dans un classificateur (8) neuronal.

7. Réseau neuronal selon la revendication 6, **caractérisé en ce que**, la couche intermédiaire (3) présente au moins deux auto-associateurs (7).

8. Réseau neuronal selon la revendication 6 ou 7, **caractérisé en ce que** l'entrée (2) présente des groupes d'entrée (5) qui ont plusieurs neurones (6), lesquels sont attribués respectivement à une catégorie linguistique, et chaque groupe d'entrée sert à l'enregistrement de la catégorie linguistique d'un mot du texte à analyser.

9. Réseau neuronal selon l'une quelconque des revendications 6 à 8, **caractérisé en ce que** le réseau est conçu pour l'édition d'un niveau de phrasage binaire, tertiaire ou quaternaire.

10. Réseau neuronal selon l'une quelconque des revendications 7 à 9,
caractérisé en ce que
le réseau est réalisé pour l'édition d'une zone de phrasage quasi continue.

5 11. Procédé selon l'une quelconque des revendications 1 à 5,
caractérisé par
l'utilisation d'un réseau neuronal selon l'une quelconque des revendications 6 à 10.

10 12. Dispositif pour la détermination de marquages prosodiques avec un système informatique (9), qui présente une
zone de mémoire (11) dans laquelle est mémorisé un programme pour l'exécution d'un réseau neuronal selon
l'une quelconque des revendications 6 à 10.

15

20

25

30

35

40

45

50

55

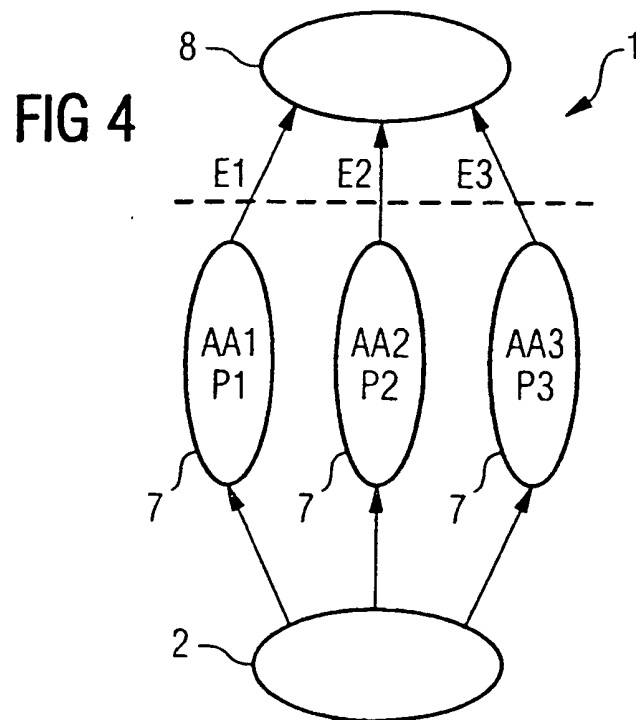
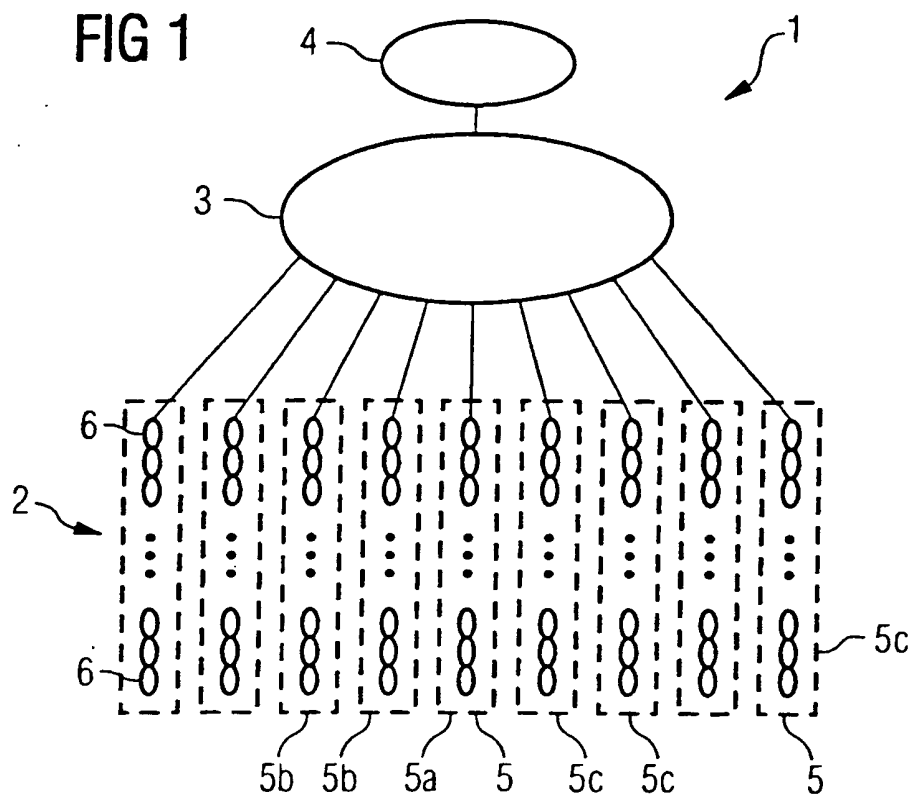


FIG 2

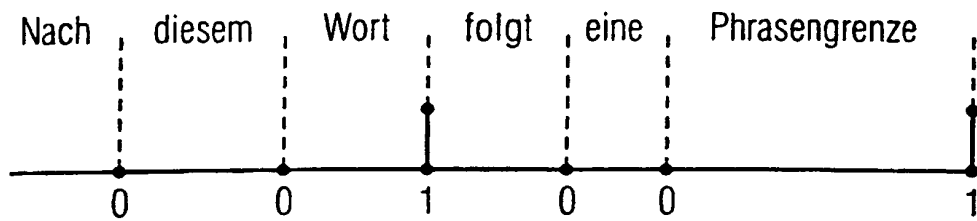


FIG 3

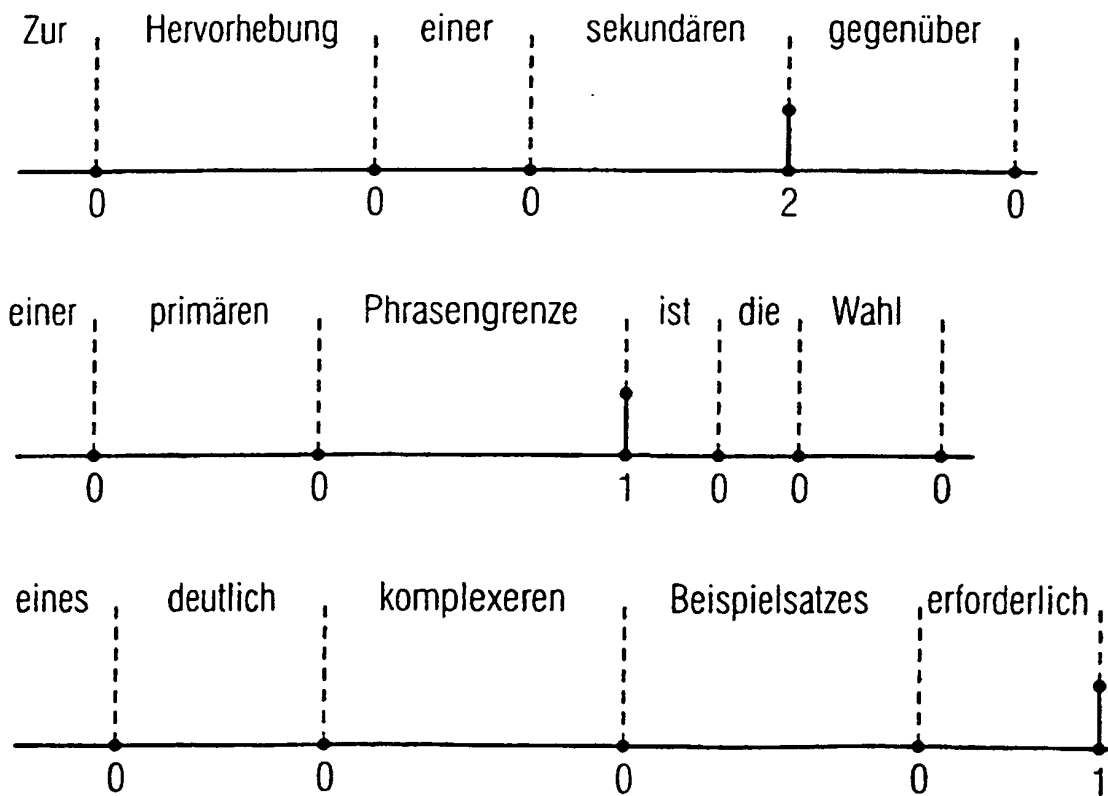


FIG 5

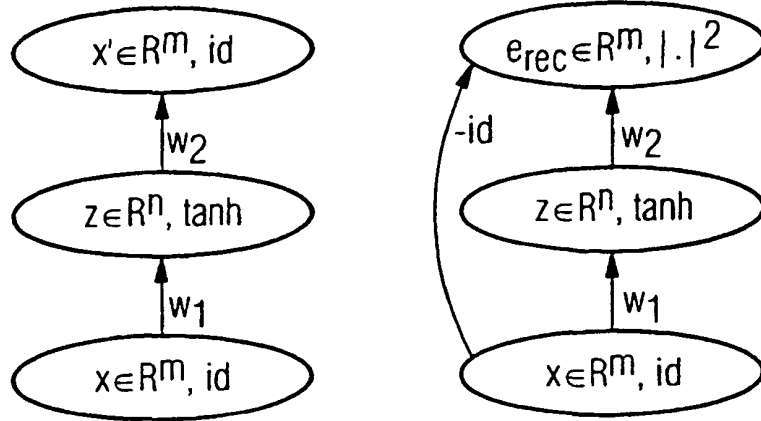


FIG 6

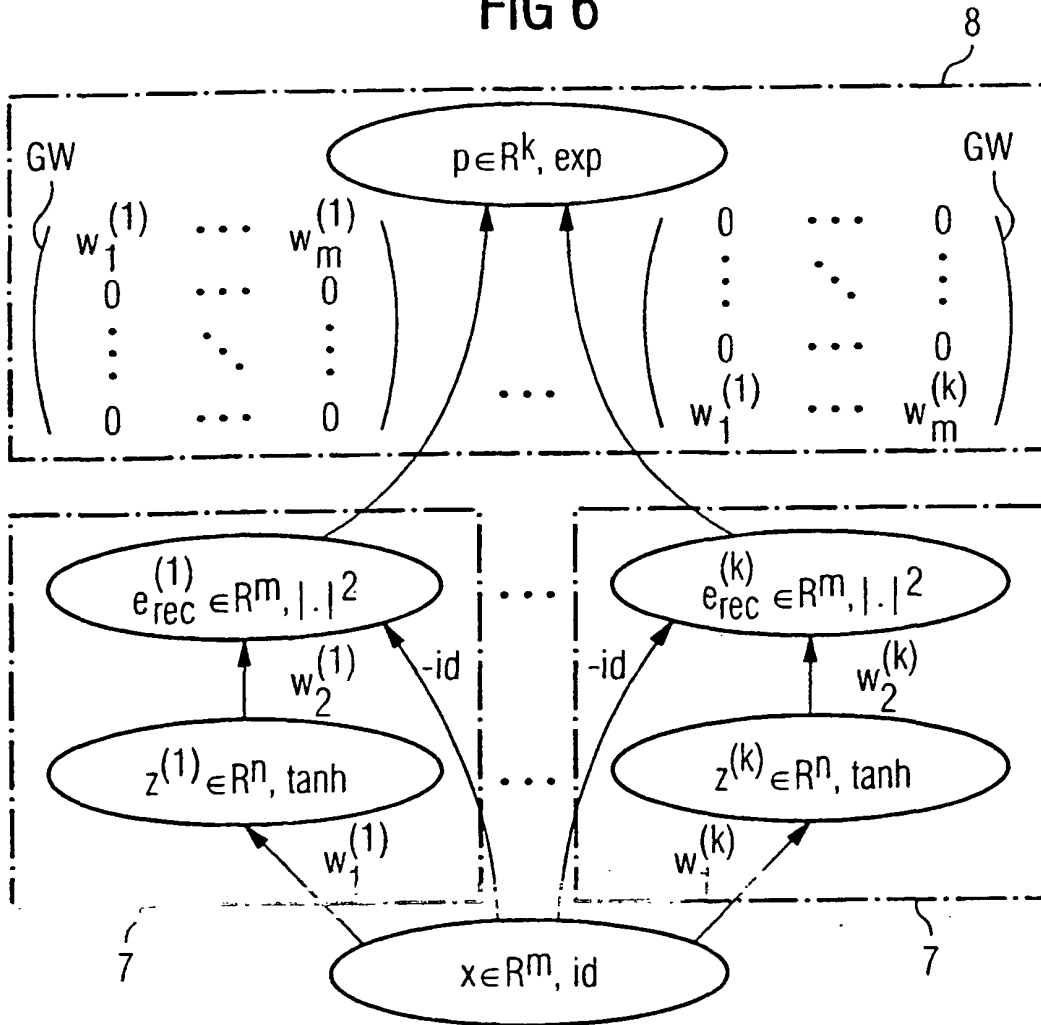


FIG 7

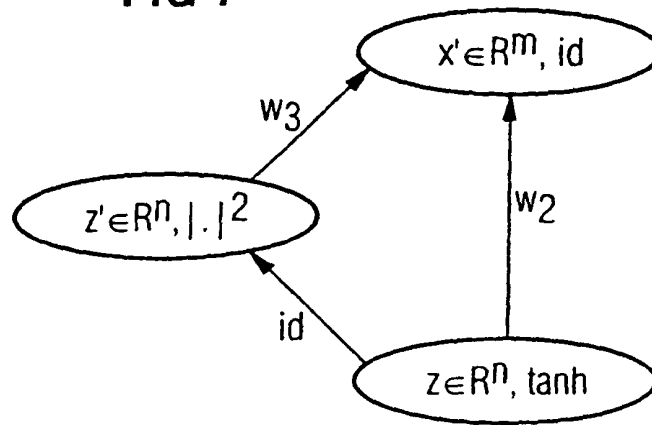


FIG 8

