(19) Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 10 2004 011 426 B3** 2005.05.25

(12) **Patentschrift**

(21) Aktenzeichen: **10 2004 011 426.9**
(22) Anmeldetag: **09.03.2004**
(43) Offenlegungstag: **–**
(45) Veröffentlichungstag der Patenterteilung: **25.05.2005**

(51) Int Cl.[7]: **G10L 11/00**
G10L 15/24

Innerhalb von 3 Monaten nach Veröffentlichung der Erteilung kann Einspruch erhoben werden.

(71) Patentinhaber:
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., 80686 München, DE

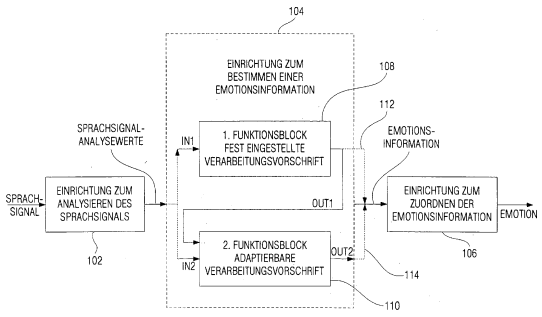
(74) Vertreter:
Schoppe, Zimmermann, Stöckeler & Zinkler, 82049 Pullach

(72) Erfinder:
Kim, Dong-Hak, 81549 München, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:
US 64 80 826 B2
US2003/00 28 384 A1
US 62 19 657 B1
EP 13 18 505 A1
EP 12 56 937 A2

(54) Bezeichnung: **Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion und Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion**

(57) Zusammenfassung: Eine Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion umfasst eine Einrichtung zum Bestimmen (104) einer Emotionsinformation aus Sprachsignalanalysewerten, die von dem Sprachsignal abgeleitet sind, wobei die Einrichtung zum Bestimmen (104) einen ersten Funktionsblock zum Liefern einer Ausgabe aus einem Ergebnis gemäß einer festeinstellbaren Verarbeitungsvorschrift umfasst, wobei die festeingestellte Verarbeitungsvorschrift eine Emotionserkennungsstandardeinstellung ist. Ferner umfasst die Einrichtung zum Bestimmen (104) einen zweiten Funktionsblock (110) zum Liefern einer Ausgabe aus einer Eingabe gemäß einer adaptierbaren Verarbeitungsvorschrift, wobei der zweite Funktionsblock (110) so ausgebildet ist, dass die adaptierbare Verarbeitungsvorschrift eine individuelle Adaption der Standardeinstellung der festeingestellten Verarbeitungsvorschrift an ein Individuum liefert, wenn eine Adaption mit einem Individuum ausgeführt wird, wobei der erste Funktionsblock (108) mit dem zweien Funktionsblock (110) so gekoppelt ist, dass eine Ausgabe des ersten Funktionsblocks (108) als Eingabe des zweiten Funktionsblocks (110) verwendbar ist. Eine derartig ausgestaltete Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion bietet somit den Vorteil, eine hochgradig genaue und einfache Emotionserkennung gegenüber einer herkömmlichen Emitionserkennung bereitstellen zu können.



Beschreibung

[0001] Die vorliegende Erfindung bezieht sich auf eine Vorrichtung zur Signalverarbeitung und insbesondere bezieht sich die vorliegende Erfindung auf eine Vorrichtung zur Parametrisierung der Emotion anhand der Stimmfarbe.

[0002] Das Mensch-Maschine-Interface (MMI) ist eine funktionale Schnittstelle zwischen Menschen und Maschine. Das MMI ermöglicht, dass Benutzer auf die Funktionen, die in der Maschine entweder hard-wired oder als Softwareprogramm realisiert sind, durch bestimmte Mechanismen zugreifen. Dieses wird im allgemeinen als Interaktivität bezeichnet, da die Funktion nicht autonom stattfindet, sondern durch aktive Teilnahme des Benutzers zustande kommt. Diese Bestätigung der Funktion kann quasi als interaktive Absprache zwischen Menschen und Maschine betrachtet werden.

[0003] Allerdings basiert diese Kommunikation zwischen Menschen und Maschine auf einer eher technisch orientierten Methode. Daher verliert der Mensch oft den Überblick darüber, welche Funktionen beispielsweise ein Gerät anbietet oder wie man eine bestimmte Funktion aktiviert. Aus diesem Grund wird immer öfter versucht, eine menschenfreundliche Schnittstelle zu definieren und zu entwerfen. Dabei werden die Sinnesorgane der Menschen oft als Vorbild genommen, da sie eine über lange Zeit optimierte „Schnittstelle" zur Umgebung darstellen und somit bereits erfolgreich evaluiert sind. Die Sprachkommunikation ist hier von besonderem Interesse.

[0004] Unter Sprachkommunikation im Bereich MMI versteht man eine verbale Interaktivität oder kurz ein Dialogsystem. Dieses Dialogsystem hat in der letzten Zeit viele Anwendungsbereiche neu entdeckt, wie z. B. im Bereich Telefonbanking, in Call-Centern, in Speech-to-Text-Anwendungen usw. Dabei hat man immer die non-verbale Komponente im Dialogsystem vernachlässigt. Dies liegt zum Teil daran, dass das bisher entdeckte Einsatzgebiet allein mit dem Dialogsystem gut beschäftigt bzw. ausgelastet ist, und zum anderen Teil daran, dass kein entsprechendes Einsatzgebiet für die nonverbale Komponente bisher entdeckt werden konnte.

[0005] Hierbei ist zunächst klarzustellen, dass diese non-verbale Komponente, die oft in der Forschung „Prodosy" genannt wird, aus Klarheitsgründen an dieser Stelle einschränkend erneut definiert wird. Insbesondere umfasst die non-verbale Komponente der Stimme (NVC = non-verbal component) zunächst, dass alle akustischen Eindrücke inklusive der Stimme Informationsquellen sind. Weiterhin werden die semantischen und lexikongrammatischen Elemente nicht berücksichtigt und zusätzlich wird auch ein Ausruf (Exclamation) als Informationsquelle betrachtet.

[0006] Die menschliche Stimme bzw. ein akustischer Eindruck enthält nicht nur die Information, die ihr lexikongrammatisch zugeordnet werden kann, sondern auch Elemente, die die emotionale Lage der Sprechenden verraten. Diese emotionale Lage kann beispielsweise eine Aufgeregtheit, eine Traurigkeit, eine Glücklichkeit, eine Deprimiertheit usw. umfassen. Diese Elemente werden für die MMI-Entwicklung bisher eher als Hindernis angesehen, weil beispielsweise eine emotionale Aufregung die Erkennung des gesprochenen Wortes erschwert. Die non-verbale Komponente (NVC) wird aus diesem Grund bisher als „Geräusch" oder „Interferenz" behandelt.

[0007] Für die Erkennung eines gesprochenen Wortes wird in herkömmlichen Verfahren im wesentlichen die menschliche Stimme durch technische Methoden analysiert, die in ihren Grundzügen nachfolgend näher erläutert werden.

[0008] Zunächst ist in diesem Zusammenhang eine Spektrumsanalyse der Stimmfarbe (Formant-Frequenz) zu nennen. Eine Formant-Frequenz ist ein charakteristischer Resonanzbereich, der z.B. für die Klangfarbe eines musikalischen Instruments oder der menschlichen Stimme (Vokal, stimmhafte Konsonanten) verantwortlich ist. Jedoch wird die ursprüngliche Definition heutzutage modifiziert und oft als Synonym für einen phonetischen Frequenzbereich verwendet.

[0009] Die menschliche Stimme hat ein breites Frequenzspektrum. Daher stellt z. B. ein Vokal mehrere Formant-Frequenzen dar. Bei der Spracherkennung (oft auch als Speech-Erkennung bezeichnet = semantische Erkennung unabhängig vom Sprecher) sind in der Regel die erste Frequenz (niedrigere Frequenz F1) und die zweite Frequenz (höhere Frequenz F2) relevant, wie in Fig. 4 dargestellt ist. Die weiteren Obertöne (höhere Formant-Frequenzen F3, F4,...) haben mit dem semantischen Kontext wenig zu tun, aber dafür eine hohe Bedeutung bei der Bestimmung von Stimmfarben wie z. B. männlich, weiblich und kindlich. Diese weiteren Formant-Frequenzen sind außerdem deshalb bisher wenig fokussiert, weil deren niedrige Amplituden die weitere Verarbeitung schwierig machen. Allerdings sind die höheren Formant-Frequenzen bei der Synthese einer realistischen Stimme sehr wichtig, da diese die Stimmfarbe stark beeinflussen.

[0010] In Fig. 4 ist dieser Zusammenhang nochmals dargestellt, wobei zu erkennen ist, dass, je höher eine Formant-Frequenz ist, desto niedriger ihre Amplitude ausfällt. Die Spektralanalyse hat jedoch einen Schwachpunkt, der darin besteht, dass nur diejenigen Amplituden, die buchstäblich „laut schreien" (d. h. eine höhere Amplitude haben), berücksichtigt werden, während die niedrigeren Amplituden der Formant-Frequenzen leicht vernachlässigt werden, weil

„laut" bzw. „groß" psychologisch und auch bei der Signalverarbeitung oftmals für wichtiger gehalten werden als Frequenzen mit niedrigen Amplituden. Da jedoch vor allem eine zu erkennende Gefühlslage aus einer Stimme aus den niedrigeren Amplituden bzw. höheren Formant-Frequenzen erkannt werden kann, weist die herkömmliche Spektralanalyse der Stimmfarbe einen diesbezüglichen Nachteil auf.

[0011] Ein weiterer Ansatz zur Erkennung von Emotionen ist die semi-spektrale bzw. non-spektrale Analyse. Hierbei wird ausgenutzt, dass akustisch ausgedrückte Emotionen nicht immer durch sprachliche Komponenten begleitet werden, sondern sich oft in Form von „Bursts" oder Pausen ausdrücken. Weiterhin zeigt sich die Emotion auch durch Intanation, die beim Sprechen verwendet wird. Diese Komponenten haben eigentlich nichts mit dem semantischen Kontext zu tun. Trotzdem werden sie in allen Sprachen zur Verdeutlichung des Kontextes eingesetzt. Der Ausdruck „wie bitte" kann beispielsweise je nach der genutzten Intonation entweder als Bitte um die Wiederholung des gesprochenen Satzes oder als aggressive Reaktion interpretiert werden.

[0012] Derartige Kamponenten und Elemente sind bei der semi-spektralen bzw. non-spektralen Analyse in Betracht zu ziehen, um eine präzisere Einschätzung der Gefühlslage eines Sprechers zu erzielen. Die genannten Komponenten werden aus diesem Grund oftmals kompensatorisch zur Analyse der Formant-Frequenzen verwendet, wie sie vorstehend näher beschrieben wurde.

[0013] Um nun die mit den vorstehend beschriebenen Verfahren (d.h. der Spektrums- oder Spektralalanalyse sowie der semi-spektralen bzw. non-spektralen Analyse) erhaltenen Analyseergebnisse zu verarbeiten, um hieraus eine Emotion des Sprachsignals zu erkennen, werden oftmals neuronale Netze eingesetzt. Neuronale Netze sind ein technischer Ansatz, der bei heuristischer Problemlösung oder vorzüglich bei Realisierung eines impliziten Mechanismus eingesetzt wird. Die Grundidee liegt darin, dass man auf eine Flowchart-Logik oder deterministische Algorithmen verzichtet und ein System selber die Lösung finden lässt. Das Vorbild dieses Ansatzes ist das Nervensystem.

[0014] Dabei wird die Informationsverarbeitung in die nachstehend näher aufgeführten zwei Untergruppen untergliedert. Die eine Untergruppe umfasst die Sensorik (= Input; afferentes Signal), wobei die andere Untergruppe die Motorik (= Output; efferentes Signal) ist. Unter dem Begriff „Sensorik" versteht man die Informationsverarbeitung, die mit Hilfe von Sensoren/Sinnesorganen stattfindet. Über diesen Weg sammelt ein System Informationen aus der Umwelt. Dies stellt also die Eingänge der Information ins System dar. Die Untergruppe der Motorik ist dagegen für eine Informationsverarbeitung verantwortlich, in der eine Lösung gefunden wird und nach außen weitergegeben wird.

[0015] Ein derartiger Zusammenhang ist in Fig. 5A dargestellt, in der ein neuronales Netz **500** eine Eingangsnetzwerkschicht **502** mit mehreren Eingangsneuroiden **504**, eine versteckte (= verborgene) Netzwerkschicht **506** mit mehreren versteckten Neuroiden **508** sowie eine Ausgangsnetzwerkschicht **510** mit mehreren Ausgangsneuroiden **512** umfasst. Die einzelnen Neuroiden der verschiedenen Netzwerkschichten (d. h. der Eingangsnetzwerkschicht **502**, der versteckten Netzwerkschicht **506** sowie der Ausgangsnetzwerkschicht **510**) sind durch Verknüpfungen **514** miteinander verknüpft, wobei die einzelnen Verknüpfungen durch unterschiedliche Gewichtungsfaktoren oder Verarbeitungsalgorithmen beschrieben werden können. Wird nun an die Neuroiden **504** der Eingangsnetzwerkschicht **502** eine Eingabe **516** angelegt, erfolgt durch das neuronale Netz **500** eine Verarbeitung derart, dass die Eingabe **516** in einer Ausgabe **518** resultiert, die von den Neuroiden **512** der Ausgangsnetzwerkschicht **510** bereitgestellt wird. Weiterhin wird aus Fig. 5A ersichtlich, dass zwischen dem Input des neuronalen Netzes **500** (d. h. der Eingangsnetzwerkschicht **502**) und dem Output des neuronalen Netzes **500** (d. h. der Ausgangsnetzwerkschicht **510**) eine (oder auch mehrere) weitere Schicht(en) (d.h. entsprechende weitere versteckte Schichten wie die versteckte Schicht **506**) einlegen. Dadurch wächst zwar die Komplexität der Struktur, aber die neuronalen Netze sind dann zur Lösung einer Aufgabe einer (mathematisch) höheren Dimension fähig.

[0016] Neben der in Fig. 5A dargestellten Struktur eines neuronalen Netzes als Feed-Forward-Netzstruktur (= Feed-forward-Netz) als allgemeine Struktur existiert weiterhin eine Netzstruktur, die als rekurrentes Netz bezeichnet wird und in Fig. 5B dargestellt ist. Im Unterschied zu dem in Fig. 5A dargestellten Feed-Forward-Netz sind nunmehr im rekurrenten Netz die Ausgänge direkt mit den Eingängen rückgekoppelt.

[0017] Seitdem das Forschungsgebiet der neuronalen Netze (d.h. der künstlichen Intelligenz) in den 60er Jahren des letzten Jahrhunderts begann, sind verschiedenste architektonische Strukturen dieser Netze vorgeschlagen worden. Insbesondere die Self-Organizing-Map SOM (= selbstorganisierende Karte), wie sie in Fig. 6 dargestellt ist, und das rekurrente Netz, wie es in Fig. 5B dargestellt ist, sind von besonderem Interesse. Unabhängig von der Struktur des neuronalen Netzes ist es allen neuronalen Netzen gemeinsam, dass dieselben aus Neuronen bzw. Neuroiden (Knoten) bestehen und jedes Neuron ein Gewicht enthält. Das Gewicht ist dabei ein Maßstab, wie wichtig ein Neuron für den nächsten Schritt ist.

Das Gewicht symbolisiert somit die Verbindungsstärke zwischen den einzelnen Neuronen im „Nervensystem".

[0018] Die SOM (Self-Organizing-Map) ist ein biologisch plausibles Modell. Der Input-Raum (= Input-Space) X ist in Fig. 6 als ein kontinuierlicher Raum dargestellt. Die Punkte **602** in diesem Input-Raum X werden als Input ins Netz **609**, das rechts in Fig. 6 dargestellt ist, eingefüttert. Mit der Zeit bildet das Netz bzw. die Neuronen oder Neuroiden die Ordnung/Mannigfaltigkeit des Inputraums implizit in Gewichten der Neuronen ab. Hierbei kennzeichnet die Variable a* ein Neuron, die Variable W_{a^*} das Gewicht des Neurons a* und die Variable X den Input-Space. Ein Input x aktiviert eine bestimmte Gruppe von Neuronen. Das zentrale Neuron, das für die nachfolgende Informationsverarbeitung den Hauptbeitrag liefert, wird nach dem folgenden Prinzip ausgewählt:

$$a^* = \arg\min \forall a \in A \|W_{a^*} - x\|$$

[0019] Das Gewicht der Neuronen wird anfangs mathematisch zufällig verteilt. Ein iteratives Lernen mit Daten bringt dann das Netz irgendwann zu einem „Equilibrium". Wenn ein Equilibrium erreicht ist, dann bedeutet dies, dass das neuronale Netz in der Lage ist, das vorgegebene Problem zu lösen.

[0020] Hat ein Netz mit bestimmter Topologie eine Lösung gefunden, dann heißt es, dass das Netz eine Lösung nicht algorithmisch darstellt, sondern implizit in Gewichten von Neuronen enthält. Diese Gewichte sind mathematisch als eine Matrix formulierbar, deren Elemente aus jeweiligen Gewichten besteht. Es ist hierbei jedoch anzumerken, dass die Verknüpfung der einzelnen Neuroiden, wie sie in Fig. 5A unter dem Bezugszeichen **519** dargestellt ist, durch einen mathematischen Zusammenhang, d. h. eine Verarbeitungsvorschrift, gekennzeichnet ist. Durch die Vielzahl von (zum Teil mathematisch unterschiedlichen) Verknüpfungsvorschriften **514** (wie beispielsweise linearen oder quadratischen Zusammenhängen) lässt sich eine Ausgabe des neuronalen Netzes jedoch aufgrund der Komplexität des neuronalen Netzes nicht als einfacher deterministischer Algorithmus beschreiben. Hierbei ist zu beachten, dass die Verknüpfungsvorschriften zumeist nicht-linear gewählt werden, da ansonsten die Gewichte der einzelnen Neuroiden unendlich wachsen könnten. Die Gewichte werden daher zumeist durch eine bestimmte Schwellenfunktion oder durch eine Normierung in einem bestimmten Bereich gehalten.

[0021] Ausgehend von diesen grundsätzlichen Methoden und Verfahren zum Analysieren eines Sprachsignals sind im Stand der Technik bereits einige Ansätze zur Erkennung einer Emotion in einem Sprachsignal vorgeschlagen worden.

Stand der Technik

[0022] So schlägt beispielsweise die Schrift US 6480826 B2 einen Ansatz vor, aus der Stimme verschiedene Parameter wie Vokalenergie, Frequenzspektrumsmerkmale, Formanten oder zeitliche Merkmale wie Sprachrate und Sprachpausen zu extrahieren, um ausgehend durch eine Trainingssequenz ein neuronales Netz derart zu trainieren, dass es in der Lage ist, emotionale Kategorien wie den Normalzustand, den Glückszustand, einen Ärger-Zustand, einen Traurigkeitszustand oder einen Angstzustand zu erkennen. Dieses System kann in Telefonzentren als Emotionsberater oder als Warneinrichtung für Geschäftstreffen eingesetzt werden. Nachteilhaft bei dem in Schrift US 6480826 B2 vorgeschlagenen Ansatz ist jedoch, dass das System und das Verfahren nicht für einzelne Sprecher individuell konfigurierbar ist und weiterhin eine lange Trainingszeit braucht.

[0023] Weiterhin wurde in der Schrift US 6219657 B1 vorgeschlagen, ein neuronales Netz mit Sprachinformationen aus einem Mikrophon zu füttern, wobei insbesondere die Lautstärke der Stimme und die Sprachintervalle für die Verarbeitung in einem neuronalen Netz zur Erkennung von Freude, Ärger, Traurigkeit und Überraschung herangezogen werden können. Dadurch, dass auf die Auswertung von Frequenzanteilen der Stimme verzichtet wird, ist jedoch lediglich eine suboptimale Erkennung von Emotionszuständen möglich, da gerade in verschiedenen Frequenzanteilen deutliche Hinweise auf den Emotionszustand enthalten sind. Hierbei ist anzumerken, dass zumeist in herkömmlichen neuronalen Netzen, die für eine Sprachanalyse verwendet werden, eine Netztopologie in der Form eines „feed-forward"-Netzes ausgebildet ist.

[0024] Weiterhin wird in der Schrift EP 1318505 A1 vorgeschlagen, ein Sprachsignal zuerst nach verschiedenen Charakteristika wie beispielsweise Frequenzkomponenten, Phonem-Segmenten oder Wortsegmenten zu analysieren und diese nachfolgend anhand eines Emotionsmusters aus einer Datenbasis zur Erkennung der derzeitigen Emotion des Sprechers zu erkennen. Ein derartiges Verfahren weist jedoch den Nachteil auf, dass immer eine (teils umfangreiche) Datenbank zur Verfügung stehen muss, wodurch ein derartiges System nicht speichereffizient und somit kostengünstig realisierbar ist.

[0025] Weiterhin wird in der Schrift EP 1256937 A2 vorgeschlagen, ein Sprachsignal zunächst durch einen Tiefpass zu filtern und nachfolgend aus dem Originalsignal und dem tiefpassgefilterten Signal einzelne Merkmale wie beispielsweise Spektralmerkmale, Varianzen, Mittelwerte, Maxima und Minima und ähnliches zu extrahieren und diese durch eine Emotionserkennungseinheit auf eine in dem Sprachsignal enthaltene Emotion hin zu überprüfen. Die Emotionser-

kennungseinheit kann hierbei unterschiedlich ausgestaltet sein, wobei auch ein neuronales Netz als Ausgestaltungsform der Emotionserkennungseinheit möglich ist. In der Schrift EP 1256937 A2 kann diese Emotionserkennungseinheit entweder off-line ausgestattet sein, wobei in dieser Offline-Ausgestaltung feste Parameter eingestellt sind, unter deren Verwendung aus einem Sprachsignal eine Emotion erkannt werden kann. Weiterhin ist es möglich, die Emotionserkennung durch eine Online-Emotionserkennungseinheit durchführen zu lassen, die mit einer Datenbasis und einem Lernalgorithmus an persönliche Sprachcharakteristikmerkmale eines Sprechers angepasst werden kann. Hierbei bietet sich zwar der Vorteil, dass bereits eine Anpassung an individuelle Sprachmerkmale eines Sprechers möglich ist, und somit eine exaktere Bestimmung der Emotion aus einem Sprachsignal erfolgen kann, dadurch, dass jedoch entweder die Offline-Emotionserkennung oder die Online-Emotionserkennung aktiviert ist, muss bei der sprecherindividuellen Emotionserkennung zunächst ein zeitaufwendiges Training des in der Emotionserkennungseinheit ausgeführten Emotionserkennungsalgorithmus an den Sprecher erfolgen. Dies kann mitunter sehr zeitaufwendig sein und somit die Einsetzbarkeit des in Schrift EP 1256937 A2 dargestellten Ansatzes behindern.

[0026] Die US 2003/0028384 A1 zeigt ein Verfahren zum Erkennen von Emotionen aus Sprache unter Verwendung einer Sprecheridentifikation. Hierzu wird zunächst aus einem empfangenen Spracheingangssignal ein Sprecher identifiziert. Hieran anschließend erfolgt ein Abgleich mit einer Datenbank, ob bezüglich des identifizierten Sprechers Emotionserkennungsdaten für diesen identifizierten Sprecher in der Datenbank vorliegen. Liegen solche Emotionserkennungsdaten in der Datenbank für den identifizierten Sprecher vor, wird zur Erkennung einer Emotion des identifizierten Sprechers eine Sprecher-spezifische Emotionserkennungseinrichtung zum Erkennen einer Emotion aus dem Sprachsignal verwendet. Ergibt die Identifizierung des Sprechers und der nachfolgende Abgleich mit der Datenbank das Ergebnis, dass für den identifizierten Sprecher noch keine Emotionserkennungsdaten in der Datenbank vorliegen (das bedeutet, dass für den identifizierten Sprecher noch keine Sprecher-spezifische Emotionserkennungseinrichtung trainiert wurde), wird zur Erkennung einer Emotion dieses Sprechers eine Sprecherunabhängige Emotionserkennungseinrichtung (beispielsweise mit einer Emotionserkennungsstandardeinstellung) verwendet. Aus der Sprecher-spezifischen oder alternativ der Sprecher-unabhängigen Emotionserkennungseinrichtung werden dann aus den erhaltenen Daten der bzw. die emotionalen Zustände des Sprechers abgeleitet.

Aufgabenstellung

[0027] Ausgehend von diesem Stand der Technik liegt der vorliegenden Erfindung die Aufgabe zugrunde, eine Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion und ein Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion zu schaffen, welche die Möglichkeit bieten, auf einfache und zeitsparende Weise eine Emotion, die in einem Sprachsignal enthalten ist, erkennen zu können.

[0028] Diese Aufgabe wird durch eine Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion gemäß Anspruch 1 und einem Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion gemäß Anspruch 18 gelöst.

[0029] Die vorliegende Erfindung schafft eine Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion mit folgendem Merkmal:
einer Einrichtung zum Bestimmen einer Emotionsinformation aus Sprachsignalanalysewerten, die von dem Sprachsignal abgeleitet sind, wobei die Einrichtung zum Bestimmen folgende Merkmale umfasst:
einen ersten Funktionsblock zum Liefern einer Ausgabe aus einem Ergebnis gemäß einer festeingestellten Verarbeitungsvorschrift, wobei die festeingestellte Verarbeitungsvorschrift eine Emotionserkennungsstandardeinstellung ist; und
einen zweiten Funktionsblock zum Liefern einer Ausgabe aus einer Eingabe gemäß einer adaptierbaren Verarbeitungsvorschrift, wobei der zweite Funktionsblock so ausgebildet ist, dass die adaptierbare Verarbeitungsvorschrift eine individuelle Adaption der Standardeinstellung der festeingestellten Verarbeitungsvorschrift an ein Individuum liefert, wenn eine Adaption mit einem Individuum ausgeführt wird, wobei der erste Funktionsblock mit dem zweiten Funktionsblock so gekoppelt ist, dass eine Ausgabe des ersten Funktionsblocks als Eingabe des zweiten Funktionsblocks verwendbar ist.

[0030] Ferner schafft die vorliegende Erfindung ein Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion mit folgendem Schritt:
Bestimmen einer Emotionsinformation aus Sprachsignalanalysewerten, die von dem Sprachsignal abgeleitet sind, wobei das Bestimmen folgende Schritte umfasst:
Liefern einer Ausgabe aus einem Ergebnis gemäß einer festeingestellten Verarbeitungsvorschrift in einem ersten Funktionsblock, wobei die festeingestellte Verarbeitungsvorschrift eine Emotionserkennungsstandardeinstellung ist; und
Liefern einer Ausgabe aus einer Eingabe gemäß einer adaptierbaren Verarbeitungsvorschrift in einem zweiten Funktionsblock, wobei der zweite Funktionsblock so ausgebildet ist, dass die adaptierbare Verarbeitungsvorschrift eine individuelle Adaption der

Standardeinstellung der festeingestellten Verarbeitungsvorschrift an ein Individuum liefert, wenn eine Adaption mit einem Individuum ausgeführt wird, wobei der erste Funktionsblock mit dem zweiten Funktionsblock so gekoppelt wird, dass eine Ausgabe des ersten Funktionsblocks als Eingabe des zweiten Funktionsblocks verwendet wird.

[0031] Der vorliegenden Erfindung liegt die Erkenntnis zugrunde, dass durch das Verwenden eines ersten Funktionsblocks und des Verwendens eines zweiten Funktionsblocks, der mit dem ersten Funktionsblock gekoppelt ist, eine einfache und hochgradig genaue Erkennung einer in einem Sprachsignal enthaltenen Emotion möglich ist. Dies resultiert insbesondere daraus, dass im ersten Funktionsblock eine festeingestellte Verarbeitungsvorschrift implementiert ist, die auf einer allgemein gültigen Emotionserkennung beruht. Diese allgemeine Emotionserkennung kann beispielsweise dadurch bereitgestellt werden, dass eine große Anzahl von Sprechern eine Datenreferenz bereitstellen, in denen Sprachsignale mit unterschiedlichen hinterlegten Emotionszuständen abgelegt sind. Aus dieser Datenreferenz können somit wesentliche Merkmale für eine Emotion in einem Sprachsignal extrahiert werden, die somit eine nahezu Allgemeingültigkeit besitzen. Ist beispielsweise bei einer Verärgerung eine deutliche Frequenzverschiebung des Grundtons der Sprache bei allen Referenzpersonen erkennbar, kann eine derartige Frequenzverschiebung als Merkmal für eine Verärgerung herangezogen werden, unabhängig, wie stark eine derartige Grundverschiebung bei individuellen Sprechern ausgebildet ist. Die festeingestellte Verarbeitungsvorschrift bietet somit die Möglichkeit, individuumunabhängig eine Emotion aus dem Sprachsignal des Individuums erkennen zu können.

[0032] Wird nun der zweite Funktionsblock verwendet, in dem die adaptierbare Verarbeitungsvorschrift ausführbar ist, kann eine Adaption der Emotionserkennung an einen individuellen Sprecher (oder eine individuelle Sprecherin) erfolgen, so dass eine weitere Verfeinerung der Emotionserkennung auf der Basis der festeingestellten Verarbeitungsvorschrift des ersten Funktionsblocks möglich ist. Dadurch, dass eine Ausgabe des ersten Funktionsblocks als Eingabe des zweiten Funktionsblocks verwendet wird, ist somit eine derartige Verfeinerung der Emotionserkennung möglich, so dass sich im Endeffekt eine Emotionserkennung einer in einem Sprachsignal enthaltenen Emotion realisieren lässt, die individuell auf eine sprechende Person einstellbar ist.

[0033] Der erfindungsgemäße Ansatz bietet den Vorteil, dass einerseits auf einen in der Wissenschaft bekannten Datensatz zur Erkennung von Emotionszuständen zurückgegriffen werden kann, der nicht von einzelnen Sprechern abhängig ist, und andererseits eine individuell anpaßbare Emotionserkennung möglich wird. Hierbei ist insbesondere anzumerken, dass durch die erfindungsgemäße Verkopplung der festeingestellten Verarbeitungsvorschrift und der adaptierbaren Verarbeitungsvorschrift eine deutlich beschleunigte Adaption, d. h.

[0034] Anpassung an die Sprachcharakteristik des individuellen Sprechers, möglich ist, da bereits auf die Grundmuster der Emotionserkennung zurückgegriffen werden kann. Durch den erfindungsgemäßen Ansatz ist es daher möglich, eine Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion sowie ein Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion bereitzustellen, die gegenüber dem Stand der Technik ein deutlich beschleunigtes Adaptionsverhalten an einen individuellen Sprecher (oder eine individuelle Sprecherin) ermöglichen, wobei der erfindungsgemäße Ansatz technisch einfach realisierbar ist, da keine Referenzdatenbasen notwendig sind, sondern lediglich auf funktionale Zusammenhänge zurückgegriffen werden kann, die in der festeingestellten Verarbeitungsvorschrift implementiert werden können.

[0035] Gemäß einem bevorzugten Ausführungsbeispiel der vorliegenden Erfindung umfasst die Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion ferner eine Einrichtung zum Analysieren des Sprachsignals, um Sprachsignalanalysewerte zu erhalten. Diese Einrichtung zum Analysieren kann beispielsweise eine Einrichtung zum Ausführen einer Fourier-Transformation umfassen, so dass beispielsweise aus einem Sprachsignal in Zeitbereichsdarstellung eine Frequenzbereichsdarstellung erzeugt wird, die als Eingabe für den ersten Funktionsblock und/oder den zweiten Funktionsblock verwendbar ist. Eine derartige Verwendung einer Einrichtung zum Analysieren des Sprachsignals bietet somit den Vorteil, dass die Vorrichtung zum Erkennen beispielsweise direkt an ein Mikrophon oder eine andere Sprachsignalquelle anschließbar ist, und nicht auf bereits vorbearbeitete Sprachsignalanalysewerte zurückgreifen braucht.

[0036] Ferner kann beispielsweise der zweite Funktionsblock als neuronales Netz implementiert sein, das ausgebildet ist, die adaptierbare Verarbeitungsvorschrift auszuführen. Dies bietet den Vorteil, dass bereits auf die umfangreichen Forschungsarbeiten des Teilgebiets der künstlichen Intelligenz zurückgegriffen werden kann, und somit auf einfache Art und Weise eine Adaption eines Emotionserkennungsalgorithmus an einen individuellen Sprecher möglich ist.

[0037] Vorzugsweise kann das neuronale Netz des zweiten Funktionsblocks eine Eingangsnetzschicht, eine mit der Eingangsnetzschicht gekoppelte verborgene Netzschicht und eine mit der verborgenen Netzschicht gekoppelte Ausgangsnetzschicht umfassen,

wobei die Ausgabe des ersten Funktionsblocks als Eingabe der verbogenen Netzschicht verwendbar ist. Dies bietet den Vorteil, dass bei einem Betrieb der Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion ein Ergebnis der festeingestellten Verarbeitungsvorschrift direkt in eine Verarbeitung mit der adaptierbaren Verarbeitungsvorschrift berücksichtigt werden kann, ohne dass beispielsweise eine an der Eingangsnetzschicht anliegende Eingabe verzerrt wird. Durch die Eingabe des Ergebnisses des ersten Funktionsblocks (d. h. der Ausgabe des ersten Funktionsblocks) in die verborgene Netzschicht wird somit erreicht, dass die Ausgabe des neuronalen Netzes des zweiten Funktionsblocks in wesentlichem Maße durch eine Ausgabe des ersten Funktionsblocks beeinflußt wird.

[0038] Hierdurch ergibt sich wiederum der vorstehend angeführte Vorteil, einer individuell an einen Sprecher adaptierbaren Sprachcharakteristik.

[0039] Gemäß einem weiteren Aspekt der vorliegenden Erfindung ist die Ausgangsnetzschicht des neuronalen Netzes des zweiten Funktionsblocks mit der Eingangsnetzschicht des neuronalen Netzes des zweiten Funktionsblocks gekoppelt, derart, dass eine Ausgabe der Ausgangsnetzschicht als Eingabe der Eingangsnetzschicht verwendbar ist. Durch ein derartiges Verkoppeln der Ausgangsnetzschicht mit der Eingangsnetzschicht lässt sich somit ein rekurrentes neuronales Netz herstellen, das in der Lage ist, eine adaptierbare Verarbeitungsvorschrift mit einer höheren Komplexität zu bewältigen, als dies ein neuronales Netz in Feed-Forward-Struktur ermöglicht. Hierdurch ergibt sich der Vorteil, dass durch eine einfache Verkopplung der einzelnen Netzschichten des neuronalen Netzes eine deutliche Erhöhung der verarbeitbaren Komplexität möglich ist.

[0040] Gemäß einem weiteren Aspekt der vorliegenden Erfindung können die Sprachsignalanalysewerte zumindest teilweise als Eingabe des zweiten Funktionsblocks verwendet werden. Dies bietet den Vorteil, dass beispielsweise auch im ersten Funktionsblock nicht benötigte Informationen aus den Sprachsignalanalysewerten (beispielsweise höheren Formant-Frequenzen) verwendet werden, indem beispielsweise der im ersten Funktionsblock nicht benötigte Anteil der Sprachsignalanalysewerte im zweiten Funktionsblock verwendet wird und somit möglichst die vollständige, in dem Sprachsignal enthaltene Emotion, verwendet werden kann.

[0041] Gemäß einem weiteren Aspekt der vorliegenden Erfindung ist der zweite Funktionsblock ausgebildet, um ansprechend auf ein Haltesignal eine Adaption der adaptierbaren Verarbeitungsvorschrift zu verhindern. Dies bietet den Vorteil, dass beispielsweise nach einem erfolgten Training der adaptierbaren Verarbeitungsvorschrift an eine Sprachcharakteristik eines Individuums das Adaptieren der adaptierbaren Verarbeitungsvorschrift abgeschaltet werden kann und somit durch das Vermeiden des kontinuierlichen Trainings eine deutliche Beschleunigung der Signalverarbeitung im zweiten Funktionsblock möglich ist.

[0042] Gemäß einem weiteren Aspekt der vorliegenden Erfindung können die Sprachsignalanalysewerte zumindest teilweise als Eingabe des ersten Funktionsblocks verwendet werden. Hierdurch bietet sich die Möglichkeit, beispielsweise nur einzelne, für eine sprecherunabhängige Emotionserkennung notwendige charakteristische Merkmale in dem Sprachsignal im ersten Funktionsblock zu verarbeiten. Durch eine derartige exemplarische Beschränkung auf wesentliche, für die personenunabhängige Emotionserkennung notwendige Merkmale lässt sich ferner eine weitere Reduzierung der Komplexität für eine Bearbeitung der festeinstellbaren Verarbeitungsvorschrift im ersten Funktionsblock realisieren.

[0043] Gemäß einem weiteren Aspekt der vorliegenden Erfindung kann der erste Funktionsblock als ein neuronales Netz ausgebildet sein, das ausgebildet ist, um die festeinstellbare Verarbeitungsvorschrift auszuführen. Vorzugsweise kann das neuronale Netz des ersten Funktionsblocks eine selbstorganisierende Karte sein. Hierdurch bietet sich die Möglichkeit, unter Ausnutzung der Erkenntnisse aus dem Teilgebiet der künstlichen Intelligenz, insbesondere dem Teilgebiet der neuronalen Netze, eine Implementierung einer Emotionserkennung zu ermöglichen, die eine ausreichende Möglichkeit bietet, eine Grobklassifikation der in einem Sprachsignal enthaltenen Emotion vornehmen zu können. Insbesondere durch die Wahl des neuronalen Netzes als selbstorganisierende Karte ist es möglich, eine strukturell einfache Emotionserkennungseinrichtung zu realisieren.

[0044] Gemäß einem weiteren Aspekt der vorliegenden Erfindung ist die Emotionsinformation teilweise aus der Ausgabe des ersten Funktionsblocks bestimmbar. Dies bietet die Möglichkeit, bereits ein erstes Grobergebnis über eine zu erwartende Emotion aus dem ersten Funktionsblock zu erhalten und somit bereits eine schnell verfügbare Vorabinformation über die zu erwartende Emotionsinformation zu erhalten.

[0045] Gemäß einem weiteren Aspekt der vorliegenden Erfindung umfasst das neuronale Netz des ersten Funktionsblocks eine Mehrzahl von Neuroiden, wobei eine Verknüpfung der Neuroiden von einer Sprachcharakteristik einer Mehrzahl von Individuen abhängig ist. Unter Ausnutzung der beispielsweise in Laborversuchen erstellten Zusammenhänge zwischen einer Emotion und einer Sprechcharakteristik, wobei die Zusammenhänge allgemeingültig

sind, lässt sich somit die Verknüpfung von den Neuroiden bereits beispielsweise werksseitig einstellen. Hierdurch lässt sich vorteilhaft die Verknüpfung der Neuroiden derart einstellen, dass bereits eine Grobinformation über die Emotionsinformation erhalten werden kann, bevor die Vorrichtung an eine Person adaptiert wurde.

[0046] Gemäß einem weiteren Aspekt der vorliegenden Erfindung umfasst die Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion eine Einrichtung zum Zuordnen der Emotionsinformation zu einem vordefinierten Emotionstyp. Der Emotionstyp kann beispielsweise ein Emotionszustand wie Glücklichkeit, Traurigkeit, Ärger oder ähnliches sein. Ist die Emotionsinformation beispielsweise ein Zahlenwert, kann durch die Einrichtung zum Zuordnen der Emotionsinformation dieser Zahlenwert zu dem vordefinierten Emotionstyp zugeordnet werden. Hierdurch bietet sich dem Nutzer einer derartigen Vorrichtung der Vorteil, beispielsweise auf einer Skala eines Anzeigegerätes direkt den Emotionszustand und die Intensität des Emotionszustands ablesen zu können.

[0047] Gemäß einem weiteren Aspekt der vorliegenden Erfindung kann die Einrichtung zum Zuordnen ausgebildet sein, um eine Ausgabe des ersten Funktionsblocks und eine Ausgabe des zweiten Funktionsblocks zum Zuordnen der Emotionsinformation zu dem vordefinierten Emotionstyp zu verwenden. Hierdurch bietet sich der Vorteil, die Ausgaben des ersten Funktionsblocks und des zweiten Funktionsblocks zum Bereitstellen der Emotionsinformation zu verwenden, und hierdurch eine hochpräzise Aussage über den Emotionszustand des Sprechers des Sprachsignals bereitzustellen.

[0048] Gemäß einem weiteren Aspekt der vorliegenden Erfindung kann das neuronale Netz des zweiten Funktionsblocks in das neuronale Netz des ersten Funktionsblocks eingebettet sein. Dies bietet den Vorteil, durch eine variable Verknüpfung von Neuroiden des ersten Netzes mit Neuroiden des zweiten Netzes den erfindungsgemäßen Ansatz in platzsparender Weise umzusetzen. Dies kann beispielsweise dadurch realisiert werden, dass einzelne Neuroiden des neuronalen Netzes durch die festeinstellbare Verarbeitungsvorschrift miteinander verknüpft sind und somit das neuronale Netz des ersten Funktionsblocks bilden, während andere Neuroiden des neuronalen Netzes in adaptierbarer Weise miteinander verknüpft sind und somit das neuronale Netz des zweiten Funktionsblocks bilden. Hierbei können die einzelnen Neuroiden der neuronalen Netze des ersten und zweiten Funktionsblocks jedoch auch in physikalischer Sicht nebeneinander benachbart sein, wodurch sich dann sagen lässt, dass das neuronale Netz des zweiten Funktionsblocks in das neuronale Netz des ersten Funktionsblocks eingebettet ist. Wesentlich ist hierbei lediglich die Ausgestaltung der Verknüpfungen der einzelnen Neuroiden, um die festeinstellbare sowie die adaptierbare Verarbeitungsvorschrift auszubilden.

Ausführungsbeispiel

[0049] Ein bevorzugtes Ausführungsbeispiel der vorliegenden Erfindung wird nachfolgend anhand der beiliegenden Zeichnungen näher erläutert. Es zeigen:

[0050] Fig. 1 ein Blockschaltbild eines Ausführungsbeispiels der erfindungsgemäßen Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion;

[0051] Fig. 2 eine schematische Darstellung einer Ausführungsreihenfolge des erfindungsgemäßen Verfahrens gemäß einem bevorzugten Ausführungsbeispiel;

[0052] Fig. 3A und Fig. 3B zwei unterschiedliche Darstellungen eines struk turell gleichen rekurrenten neuronalen Netzes;

[0053] Fig. 4 ein Diagramm, in dem exemplarisch das Ergebnis einer Spektralanalyse von in einem Sprachsignal auftretenden Frequenzen dargestellt sind;

[0054] Fig. 5A und Fig. 5B zwei Darstellungen von strukturell unterschiedli chen neuronalen Netzen; und

[0055] Fig. 6 eine Darstellung eines neuronalen Netzes als selbstorganisierende Karte.

[0056] In der nachfolgenden Beschreibung der bevorzugten Ausführungsbeispiele der vorliegenden Erfindung werden für die in den verschiedenen Zeichnungen dargestellten und ähnlich wirkenden Elemente gleiche oder ähnliche Bezugszeichen verwendet, wobei auf eine wiederholte Beschreibung dieser Elemente verzichtet wird.

[0057] Fig. 1 zeigt ein Blockschaltbild eines Ausführungsbeispiels der erfindungsgemäßen Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion. Die erfindungsgemäße Vorrichtung kann hierbei eine Einrichtung zum Analysieren des Sprachsignals **102**, eine Einrichtung zum Bestimmen einer Emotionsinformation **104** und eine Einrichtung zum Zuordnen der Emotionsinformation **106** umfassen. Die Einrichtung zum Bestimmen einer Emotionsinformation **104** kann ferner einen ersten Funktionsblock **108** und einen zweiten Funktionsblock **110** umfassen, wobei der erste Funktionsblock ausgebildet ist, um eine festeingestellte Verarbeitungsvorschrift abzuarbeiten und der zweite Funktionsblock

110 ausgebildet ist, eine adaptierbare Verarbeitungsvorschrift abzuarbeiten. Die Einrichtung zum Analysieren des Sprachsignals **102** umfasst einen Eingang zum Empfangen eines Sprachsignals und einen Ausgang zum Ausgeben von Sprachsignalanalysewerten. Die Einrichtung zum Bestimmen einer Emotionsinformation **104** umfasst einen Eingang zum Empfangen der Sprachsignalanalysewerte von der Einrichtung zum Analysieren des Sprachsignals **102** und einen Ausgang zum Ausgeben einer Emotionsinformation. Die Einrichtung zum Zuordnen der Emotionsinformation **106** umfasst einen Eingang zum Empfangen der Emotionsinformation von der Einrichtung **104** zu Bestimmen der Emotionsinformation und einen Ausgang zum Darstellen der Emotion. Ferner umfasst der erste Funktionsblock **108** der Einrichtung zum Bestimmen einer Emotionsinformation **104** einen Eingang IN1 zum Empfangen von Eingangswerten und einen Ausgang OUT1 zum Ausgeben einer Ausgabe des ersten Funktionsblocks. Der zweite Funktionsblock **110** umfasst einen ersten Eingang zum Empfangen der Ausgabe des ersten Funktionsblocks **108**, wobei der erste Eingang des zweiten Funktionsblocks **110** mit dem Ausgang OUT1 des ersten Funktionsblocks **108** verbunden ist.

[0058] Ferner umfasst der zweite Funktionsblock **110** einen zweiten Eingang IN2 zum Empfangen einer weiteren Eingabe in den zweiten Funktionsblock **110**. Weiterhin umfasst der zweite Funktionsblock **110** einen Ausgang OUT2 zum Ausgeben einer Ausgabe des zweiten Funktionsblocks **110**. Weiterhin sind der Eingang IN1 des ersten Funktionsblocks **108** und der zweite Eingang IN2 des zweiten Funktionsblocks **110** mit dem Eingang der Einrichtung zum Bestimmen einer Emotionsinformation **104** verbunden. Außerdem sind der Ausgang OUT1 des ersten Funktionsblocks **108** und der Ausgang OUT2 des zweiten Funktionsblocks **110** mit dem Ausgang der Einrichtung zum Bestimmen einer Emotionsinformation **104** verbunden.

[0059] Wird nun ein Sprachsignal, das beispielsweise durch ein Mikrophon bereitgestellt werden kann, an den Eingang der Einrichtung zum Analysieren des Sprachsignals **102** angelegt, kann die Einrichtung zum Analysieren des Sprachsignals aus dem Sprachsignal Sprachsignalanalysewerte extrahieren und diese am Ausgang der Einrichtung zum Analysieren des Sprachsignals **102** ausgeben. Die Sprachsignalanalysewerte können beispielsweise eine Information über den Anteil von einzelnen Frequenzen in dem Sprachsignal sein. Diese Sprachsignalanalysewerte lassen sich dann beispielsweise durch eine Digitalisierung des vom Mikrofon empfangenen Sprachsignals und eine nachfolgende Fourier-Transformation des Sprachsignals erhalten. Weiterhin kann ein Teil der Sprachsignalanalysewerte (beispielsweise die Amplituden der niedrigeren Frequenzhälfte der Sprachsignalanalysewerte) über den Eingang IN1 dem ersten Funktionsblock **108** mit der festeingestellten Verarbeitungsvorschrift zugeführt werden. Weiterhin kann die aus dem ersten Funktionsblock **108** resultierende Ausgabe über dessen Ausgang OUT1 in den zweiten Funktionsblock **110** mittels dessen ersten Eingang eingespeist werden. Neben der Ausgabe des ersten Funktionsblocks **108** kann der zweite Funktionsblock **110** beispielsweise mit einem Teil der Sprachsignalanalysewerte (beispielsweise den Amplitudenwerten der höheren Frequenzhälfte der Sprachsignalanalysewerte) beaufschlagt werden, die an dem zweiten Eingang IN2 des zweiten Funktionsblocks **110** angelegt werden. Der zweite Funktionsblock **110** kann hierbei aus der über den ersten Eingang zugeführten Ausgabe des ersten Funktionsblocks **108** und dem am zweiten Eingang IN2 anliegenden Teil der Signalanalysewerte nunmehr eine Adaption an eine individuelle Sprachcharakteristik einer Person ausführen und ein Ergebnis der ausgeführten adaptierbaren Verarbeitungsvorschrift an dem Ausgang OUT2 des zweiten Funktionsblocks **110** ausgeben.

[0060] Die in Fig. 1 dargestellte Emotionsinformation, die von der Einrichtung zum Bestimmen einer Emotionsinformation **104** ausgegeben wird, kann dabei einen ersten Anteil **112** umfassen, der von der Ausgabe des ersten Funktionsblocks **108** abhängig ist, und einen zweiten Anteil **114** umfassen, der von der Ausgabe des zweiten Funktionsblocks **110** abhängig ist. Die derart zusammengesetzte Emotionsinformation kann nachfolgend durch die Einrichtung zum Zuordnen der Emotionsinformation **106** einem vordefinierten Emotionszustand, beispielsweise einer Traurigkeit, einem Angstzustand oder einem Glückszustand des Sprechers des Sprachsignals zugeordnet werden und beispielsweise an einer Ausgabeeinheit dargestellt werden.

[0061] Fig. 2 zeigt eine schematische Darstellung eines Ausführungsbeispiels des erfindungsgemäßen Verfahrens. Fig. 2 zeigt dabei schematisch, wie die Parametrisierung der Emotion aus der Stimme technisch zustande kommt. Die in Fig. 2 dargestellten technischen Schritte werden im folgenden genauer beschrieben.

[0062] Zuerst wird dem Verfahren ein akustisches Signal (Stimme) zugeführt, das aus mehreren Frequenzen besteht. Dies erfolgt in dem ersten Schritt, der in Fig. 2 als Input bezeichnet wird und beispielsweise in einem Voice-Kollektor erfolgen kann. Das akustische Signal wird nachfolgend zunächst in die einzelnen Frequenzen unter Verwendung von geeigneten mathematischen Verfahren (z. B. FFT = Fast Fourier-Transformation oder Wavelet-Transformation) zerlegt. Anschließend werden die Formant-Frequenzen daraus extrahiert. Dies erfolgt vorzugsweise im in Fig. 2 dargestellten Schritt der Formant-Analyse/FFT. Da Formant-Frequenzen, z. B. Resonanzbe-

reiche, beim Menschen relativ gut bekannt sind und hierbei die semantische Verarbeitung wie Sprach-/Worterkennung nicht in Frage kommt, kann man das Frequenzbündel, das sich aus der exemplarisch gewählten FFT ergibt, ohne aufwendiges Verfahren direkt als Eingang für die nachfolgenden Schritte verwenden. Eine Filterung bzw. Extraktion von einzelnen Frequenzen unter Unterdrückung beispielsweise der niedrigeren Frequenzen kann somit entfallen. Durch ein derartiges Vorgehen wird das zentrale Interesse des vorliegenden Ansatzes im Bereich der höheren Frequenzen (d.h. der für die Stimmfarbe relevanten Frequenzen) untermauert.

[0063] Die mit Hilfe der FFT zerlegten Frequenzen können zunächst in verschiedene Gruppen unterteilt werden, die jeweils einen bestimmten Frequenzbereich (Bündel von mehreren Nachbarfrequenzen) beinhalten. Diese Gruppen werden dann als Eingänge für das neuronale Netz eingesetzt.

[0064] Als dritter Schritt werden die neuronalen Netze (d. h. die künstliche Intelligenz) miteinander vernetzt, derart, dass beispielsweise der erste Funktionsblock als selbstorganisierende Karte (SOM) und der zweite Funktionsblock als rekurrentes neuronales Netz ausgebildet und miteinander vernetzt werden. Hierbei wird der als SOM ausgebildete erste Funktionsblock als „Frontend"-Netz ausgebildet, in das einzelne Frequenzbündel direkt eingefüttert werden.

[0065] Dies soll dann eine interne Topologie anhand von zur Verfügung gestellten Informationen (hier die Formanten) organisieren und die implizite Ordnung der Frequenzen in die Metaebene (Gewichte der Neuronen) projizieren. Die selbstorganisierende Karte (SOM) hat dabei einen Vorteil, dass das Lernen nicht „reinforced" ist, d. h. dass die Ordnung selbst aus den eingegebenen Baten gebildet wird. Eine Vielzahl von Eingängen/Daten ist jedoch notwendig, um eine gute interne Topologie auszubilden, was dank des Frequenzprofils bei der Stimme nicht problematisch ist. Wenn die SOM ein ausgeglichenes „Plateau" (= Equilibrium) erreicht hat, dann heißt dies, dass sich der Input-Space (= Eingangsraum) im neuronalen Netz eingebettet hat. Dies kann durch eine dem Fachmann bekannte sogenannte Energiefunktion überprüft werden.

[0066] Das rekurrente Netz ist mathematisch gesehen ein iteratives Verfahren. Die Fig. 3A und Fig. 3B verdeutlichen diesen Zusammenhang. Diese Figuren sind zwei unterschiedliche Darstellungen eines strukturell gleichen rekurrenten Netzes. Die Fig. 3A entspricht hierbei der Fig. 5B. Insbesondere umfasst das rekurrente Netz aus Fig. 3A wiederum eine Eingangsnetzwerkschicht **502** mit Eingangsneuroiden **504**, eine versteckte Schicht **506** mit versteckten Neuroiden **508** sowie eine Ausgangsschicht **510** mit Ausgangsneuroiden **512**. Da, wie in Fig. 3A ersichtlich ist, die Ausgänge, d. h. die Ausgangsneuroiden **512**, mit den Eingängen, d. h. den Eingangsneuroiden **504**, direkt verbunden sind, kann dieses Netz umformuliert werden und gemäß der Darstellung in Fig. 3B wiedergegeben werden. Hierbei fallen die Eingangsneuroiden **504** mit den Ausgangsneuroiden **512** zusammen, so dass sich die in Fig. 3B dargestellte Struktur ergibt. Die Eingangsneuroiden bzw. Ausgangsneuroiden **512** sind somit lediglich mit den versteckten Neuroiden **508** der versteckten Schicht **506** verbunden. Gemäß dem in Fig. 2 dargestellten Ausführungsbeispiel des erfindungsgemäßen Verfahrens wird die versteckte Schicht (d. h. die Neuroiden der versteckten Schicht **508**) mit der SOM verbunden. Hierdurch ist sichergestellt, dass die in Fig. 2 dargestellten beiden neuronalen Netze in Form der SOM (obere Darstellung in der Spalte der neuronalen Netze) und in der Form des rekurrenten Netzes (untere Darstellung der Spalte der neuronalen Netze aus Fig. 2) implementiert werden können.

[0067] Eine weitere Eigenschaft des rekurrenten neuronalen Netzes ist, dass durch die nichtlineare Eigenschaft des Netzes ein unendliches Wachstum verhindert wird. Wie in Fig. 3A zu sehen ist, werden die Ausgänge aus dem Netz wieder als Eingänge genutzt. Dies zeigt trotzdem kein deterministisches Verhalten, weil das Netz aus mehreren untereinander vernetzten Knoten/Schichten besteht.

[0068] In einem weiteren Schritt können die in der Fig. 2 aus der künstlichen Intelligenz erhaltenen Parameter in zwei Kategorien aufgeteilt werden. Die Parameter stellen hierbei eine Gewichtsmatrix von neuronalen Netzen dar. Zunächst ist ein „globaler" Parameter zu nennen, der die Gewichtsmatrix der SOM darstellt. Dieser „globale" Parameter ist in Fig. 2 in der Parametrisierungsspalte (letzte Spalte) als obere globale Matrix zu erkennen. Diese Matrix ist nicht personenspezifisch und soll sich wie eine globale Konstante verhalten. Hieraus wird ersichtlich, dass in der globalen Matrix bereits Zusammenhänge implementiert sind, die einer Emotionsstandardeinstellung entsprechen. Der andere Parameter ist ein „lokaler" Parameter. Der lokale Parameter ist die Gewichtsmatrix des rekurrenten Netzes, wie sie in Fig. 2 in der Spalte Parametrisierung als lokale Matrix gekennzeichnet ist. Diese Matrix soll die weitere Änderung/Anpassung (z.B. additiv oder subtraktiv) vornehmen, je nachdem, wer das Interface bedient (d.h. speaker specific = sprecherspezifisch sein) und wie lang eine Person dieses Interface nutzt (d.h. adaptiv sein). Dies bedeutet, dass sich das Gerät mit diesem Interface an seinen Besitzer anpasst. Je länger man das Interface nutzt, desto besser versteht das Gerät den Besitzer. Alternativ kann auch die Adaption durch ein Haltesignal ausgeschaltet werden, wenn das Gerät bereits ausreichend an eine Sprachcharakteristik des Benutzers adaptiert ist. Dies weist den Vorteil

auf, dass nunmehr nicht ein numerisch aufwendiger Adaptionsalgorithmus, sondern lediglich ein Analysealgorithmus bzw. eine Analysevorschrift durch das rekurrente Netz auszuführen ist.

[0069] Die Emotionsinformation kann nachfolgend beispielsweise aus einer additiven oder subtraktiven Verknüpfung von Signalen erfolgen, die nach Beaufschlagen der SOM und des rekurrenten Netzes mit Sprachsignalanalysewerten aus einem Ergebnis der SOM und einem Ergebnis des rekurrenten Netzes abgeleitet sind. Hierdurch wird eine sprecher-spezifische Emotionserkennung ermöglicht, die auf einer sprecher-unspezifischen Emotionserkennung als Standardeinstellung und einer adaptierbaren sprecher-spezifischen Emotionserkennung basiert.

[0070] Das vorstehend beschriebene Verfahren ist als „Add-On"-Interface bezeichnet, da dieses Verfahren in allen möglichen Einsatzgeräten als Software realisiert werden kann, solange ein Audioeingang zur Verfügung steht. Als Zielplattform kommen unter anderem die folgenden Möglichkeiten in Frage:

1. Ein mobiles Gerät wie ein Handy oder PDA oder eine Anwendung im Automobil
   Ein portables Gerät oder ein Automobil lässt sich leicht „personalisieren", da man den Eindruck hat, das Gerät wirklich zu „besitzen". Der Einsatz des Emotion-Sensitiv-Interfaces (EI) soll den Besitzer zur weiteren psychologischen Empfindung bringen, dass das Gerät „lebt".
2. Business-Applikationsplattform
   Das vorstehend beschriebene Verfahren kann überall dort eingesetzt werden, wo das Dialogsystem auch eingesetzt ist. Dies kann beispielsweise in einem Call-Center, beim Diktat oder als eigenes Anwendungsgerät zur Emotionswarnung bei Geschäftsbesprechungen eingesetzt werden.

[0071] Weiterhin kann durch die Parametrisierung der Emotion eine Emotionserkennung „klonfähig" gemacht werden. Das Interface EI (EI = emotion-sensitive-interface) bietet unter anderem die Möglichkeit an, die Parameter (beispielsweise die lokale Matrix) zu einem weiteren Gerät zu übertragen. Da der gelernte bzw. angepasste „Charakter" des EI als Gewicht in Form einer Matrix abgespeichert ist, kann diese Matrix ohne großen Aufwand zu verschiedenen Geräten hin kopiert werden. Das neue Gerät, auf das die Parameter übertragen wurden, braucht dann keine zusätzliche Lernphase und ist sofort in der Lage, sich auf den Besitzer einzustellen. Dieses Gerät ist außerdem weiterhin lernfähig.

[0072] Die Übertragung der Parameter ist ein reiner Kopiervorgang, aber es wirkt für Menschen, als wäre der „Charakter" des Geräts geklont, da der „Charakter" (d.h. die Parameter des EI) gleich bleibt und noch zur weiteren Anpassung fähig sind und benutzt werden können.

[0073] In den nachfolgenden Abschnitten werden ein paar Anwendungsgebiete präsentiert, um die Idee leichter verständlich zu gestalten.

[0074] Zunächst soll ein Emotion-Sensitive-Interface (EI) im Automobil vorgestellt werden. Das EI ist lernfähig, d. h. das EI versucht ständig, sich an den Menschen, der das Gerät bedient, anzupassen. Auf den ersten Blick zeigt dies eine Ähnlichkeit mit gängigen biometrischen Verfahren (Sprechererkennung) in dem Aspekt, dass das Gerät mit dem Emotion-Sensitive-Interface den Besitzer erkennt. Dieses Merkmal ist jedoch kein Hauptziel, sondern eine Erscheinung, die durch das Charakteristikum von Formant-Frequenzen zustande kommt.

[0075] Der Hauptfokus liegt darauf, die emotionale Lage des Besitzers mit Hilfe des Emotion-Sensitive-Interfaces zu erfassen. Allerdings wäre die Ermittlung der Emotion anfangs nicht immer akkurat, das Gerät passt sich mit der Zeit jedoch an den Besitzer an. Dieser adaptive Charakter differenziert sich deutlich vom allgemeinen Sprachdialogsystem, das hauptsächlich deterministisch aufgebaut ist, d. h. dass die Interaktivität vom Prozess/Gerät von Anfang an fest definiert ist (hard-wired) und hinterher nicht zu ändern ist. Beim Emotion-Sensitive-Interface ist jedoch die Interaktivität nicht festverdrahtet, sondern dynamisch.

[0076] Als Einsatzbereich ist hier beispielsweise ein Dialogsystem oder eine Fahrererkennung im Automobil denkbar. Das Emotion-Sensitive-Interface kann mit der derzeitigen Personalisierungstechnik oder einem Profilsystem kombiniert werden. Folgendes Szenario ist hierbei denkbar: Der Fahrer steigt ein und gibt einen verbalen Befehl „Musik". Der Unterschied zwischen einem deterministischen Dialogsystem (Variante A) und einem System mit Emotion-Sensitive-Interface (Variante B) zeigt sich beispielsweise durch einen nachfolgenden exemplarischen Dialog:
A fragt nach: „Welches Genre wollen Sie gerne hören?" Ein Dialog dieser Art muss weiter durchgeführt werden, bis der Benutzer eine eindeutige Musik auswählt.

[0077] Ein Gerät gemäß Variante B fragt nach: „Wollen Sie eine ruhige Musik hören? Sie hören sich etwas traurig an".

[0078] Der größte Unterschied besteht darin, dass die Auswahl nicht an einer logischen Eingrenzung liegt (d. h. dass man einen Entscheidungsbaum durchgeht und am Ende zu einem bestimmten Zielwert gelangt), sondern dass das Gerät mit dem „Geschmack" des Besitzers mit der Zeit vertraut wird.

[0079] Falls ein derzeit bestehendes Profilsystem mit einem EI ausgestattet wird, kann eine assoziative

Funktion realisiert werden. Dies bedeutet: Variante B' (Variante B + Profilsystem) fragt nach: „Sie hören sich etwas traurig an. Wollen Sie die Musik hören, die Sie zum letzten Mal hörten?"

[0080] Als weiteres Anwendungsbeispiel kommt auch eine Emotionsberaterfunktion in Frage. Je nach der emotionalen Lage kann man eine unterschiedliche Entscheidung treffen, selbst wenn die umgebende Bedingung/Situation gleich bleibt. Man kann in einer aufgeregten bzw. aggressiven Kondition eine Fehlentscheidung treffen. Wenn dies z. B. eine unternehmerische Entscheidung betrifft oder vergleichbar wichtige Angelegenheiten angeht, dann folgt eine ungemütliche Konsequenz hinterher. Ein emotionssensitiver „PDA" kann dem Besitzer im solchen Fall signalisieren, dass die Gefühlslage am Rand oder außerhalb des normalen Status liegt. Ein solcher psychologischer Überwachungsmechanismus entspricht in Analogie z. B. einem medizinischen Gerät wie beispielsweise für einen Diabetiker.

[0081] Weiterhin ist sehr relevant, die allgemeine Atmosphäre bei einem Gruppenverhalten zu verstehen, wie dies beispielsweise bei einem Meeting notwendig ist. Eine Protokollierung von emotionalen Abläufen in einer Sitzung kann ein sachliches Protokoll (manuell oder via ASR erstellt) ergänzen (vergleichbar einem Tonfilm und eine Stummfilm ohne Musik).

[0082] Eine weitere Anwendung der vorstehend beschriebenen Erfindung ist beispielsweise in einem Call-Center denkbar. Manche Call-Center sind mit einem Dialogsystem ausgestattet, um die Mitarbeiter zu entlasten und die Arbeit innovativ zu gestalten. Das Dialogsystem bedeutet eine Reihe von Dialoglisten, die per Spracherkennung bzw. Wiedergabe/Synthese mit dem Anrufer interagiert. Dabei geht es darum, den breiten Umfang von Service in einer logischen Reihenfolge zu formulieren, damit der Anrufer durch sukzessive Dialoge ans Ziel/Menü gelangt. Falls der Serviceumfang relativ groß ist, passieren auch Fälle, dass Anrufer innerhalb eines Labyrinths vom Dialogsystem „gefangen" werden. Wenn man dabei das Dialogsystem mit einem emotionssensitiven Interface ergänzt, ist das kombinierte System in der Lage, auf die Emotion des Anrufers dynamisch zu reagieren. Angenommen, dass die Stimme eines Anrufers erkennbar aggressiv klingt, dann kann das emotionssensitive Interface ein Signal auslösen, dass der Dialog nicht mehr weitergeführt wird, sondern von einem Mitarbeiter übernommen wird.

[0083] Ein Dialogsystem fragt z. B. „Sie wirken sehr aufgeregt. Wollen Sie lieber mit unserem Mitarbeiter verbunden werden?" Dann kommt die Warteschleife mit der entsprechenden Musik. Es kann aber auch sein, dass der Mitarbeiter bzw. Angerufene die emotionale Lage des Anrufers erfährt, bevor er ans Telefon kommt.

[0084] Zusammenfassend lässt sich sagen, dass sich das emotionssensitive Interface vorzugsweise als ein Add-On-Prinzip ausgestalten lässt. Die meisten Geräte mit Audioeingängen können daher mit dem emotionssensitiven Interface ausgestattet werden.

[0085] Als Vorteil des emotionssensitiven Interfaces lässt sich nennen, dass es erstens lernfähig ist, zweitens auch ohne zusätzliches Lernen die Emotion aus der Stimme erkennen kann, jedoch dann nicht sprecher- (d. h. personen-)spezifisch ist, und drittens in einer Kombination mit einem Profil-System zur kategorischen Assoziation fähig ist (Geschmack gegenüber Musik-Genre).

[0086] Weiterhin ist zu nennen, dass das emotionssensitive Interface sowohl als Software als auch in Form von Hardware realisierbar ist. Daher kann es in einem sehr breiten Umfeld eingesetzt werden, solange die Audioeingänge am Zielgerät existieren. Als weitere Vorteile sind somit zu nennen, dass erstens die Erkennung der Emotion eine erhöhte Personalisierung bietet, zweitens ein neuartiges Interface mit menschenfreundlicher Komponente bereitgestellt wird und drittens eine Übertragbarkeit durch Parametrisierung erfolgen kann.

[0087] Abhängig von den Gegebenheiten kann das erfindungsgemäße Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion in Hardware oder in Software implementiert werden. Die Implementierung kann auf einem digitalen Speichermedium, insbesondere einer Diskette oder CD mit elektronisch auslesbaren Steuersignalen erfolgen, die so mit einem programmierbaren Computersystem zusammenwirken können, dass das entsprechende Verfahren ausgeführt wird. Allgemein besteht die Erfindung somit auch in einem Computerprogrammprodukt mit einem auf einem maschinenlesbaren Träger gespeicherten Programmcode zur Durchführung des erfindungsgemäßen Verfahrens, wenn das Computerprogrammprodukt auf einem Rechner abläuft. Mit anderen Worten ausgedrückt, kann die Erfindung somit als ein Computerprogramm mit einem Programmcode zur Durchführung des Verfahrens realisiert werden, wenn das Computerprogramm auf einem Computer abläuft.

Patentansprüche

1. Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion mit folgendem Merkmal:
einer Einrichtung zum Bestimmen (**104**) einer Emotionsinformation aus Sprachsignalanalysewerten, die von dem Sprachsignal abgeleitet sind, wobei die Einrichtung zum Bestimmen (**104**) folgende Merkmale umfasst:
einen ersten Funktionsblock (**108**) zum Liefern einer

Ausgabe aus einem Ergebnis gemäß einer festeingestellten Verarbeitungsvorschrift, wobei die festeingestellte Verarbeitungsvorschrift eine Emotionserkennungsstandardeinstellung ist; und
einen zweiten Funktionsblock (**110**) zum Liefern einer Ausgabe aus einer Eingabe gemäß einer adaptierbaren Verarbeitungsvorschrift, wobei der zweite Funktionsblock so ausgebildet ist, dass die adaptierbare Verarbeitungsvorschrift eine individuelle Adaption der Standardeinstellung der festeingestellten Verarbeitungsvorschrift an ein Individuum liefert, wenn eine Adaption mit einem Individuum ausgeführt wird, wobei der erste Funktionsblock (**108**) mit dem zweiten Funktionsblock (**110**) so gekoppelt ist, dass eine Ausgabe des ersten Funktionsblocks (**108**) als Eingabe des zweiten Funktionsblocks (**110**) verwendbar ist.

2. Vorrichtung zum Erkennen einer in einem Sprachsignal enthaltenen Emotion gemäß Anspruch 1, die ferner folgendes Merkmal umfasst:
eine Einrichtung zum Analysieren des Sprachsignals (**102**), um Sprachsignalanalysewerte zu erhalten.

3. Vorrichtung gemäß Anspruch 2, bei der die Einrichtung zum Analysieren des Sprachsignals (**102**) eine Einrichtung zum Ausführen einer Fourier-Transformation umfasst, wobei die Einrichtung zum Ausführen der Fourier-Transformation ausgebildet ist, um die Sprachsignalanalysewerte bereitzustellen.

4. Vorrichtung gemäß einem der Ansprüche 1 bis 3, bei der der zweite Funktionsblock (**110**) ein neuronales Netz ist, das ausgebildet ist, um die adaptierbare Verarbeitungsvorschrift auszuführen.

5. Vorrichtung gemäß Anspruch 4, bei der das neuronale Netz eine Eingangsnetzwerkschicht (**502**), eine mit der Eingangsnetzwerkschicht (**502**) gekoppelte verborgene Netzwerkschicht (**506**) und eine mit der verborgenen Netzwerkschicht (**506**) gekoppelte Ausgangsnetzwerkschicht (**510**) umfasst, wobei die Ausgabe des ersten Funktionsblocks (**108**) als Eingabe der verbogenen Netzwerkschicht (**506**) verwendbar ist.

6. Vorrichtung gemäß Anspruch 5, bei der die Ausgangsnetzwerkschicht (**510**) mit der Eingangsnetzwerkschicht (**502**) gekoppelt ist, derart, dass eine Ausgabe der Ausgangsnetzwerkschicht (**510**) als Eingabe der Eingangsnetzwerkschicht (**502**) verwendbar ist.

7. Vorrichtung gemäß einem der Ansprüche 1 bis 6, bei der die Sprachsignalanalysewerte zumindest teilweise als Eingabe des zweiten Funktionsblocks (**110**) verwendbar sind.

8. Vorrichtung gemäß einem der Ansprüche 5 bis 7, bei der der zweite Funktionsblock ausgebildet ist, um ansprechend auf ein Haltesignal ein Adaptieren der adaptierbaren Verarbeitungsvorschrift anzuhalten.

9. Vorrichtung gemäß einem der Ansprüche 5 bis 8, bei der die Emotionsinformation zumindest teilweise aus einer Ausgabe der Ausgangsnetzwerkschicht (**510**) bestimmbar ist.

10. Vorrichtung gemäß einem der Ansprüche 1 bis 9, bei der die Sprachsignalanalysewerte zumindest teilweise als Eingabe des ersten Funktionsblocks (**108**) verwendbar sind.

11. Vorrichtung gemäß einem der Ansprüche 1 bis 10, bei der der erste Funktionsblock (**108**) ein neuronales Netz ist, das ausgebildet ist, um die festeinstellbare Verarbeitungsvorschrift auszuführen.

12. Vorrichtung gemäß Anspruch 11, bei der das neuronale Netz des ersten Funktionsblocks (**108**) als eine selbstorganisierende Karte ausgebildet ist.

13. Vorrichtung gemäß Anspruch 11 oder 12, bei der das neuronale Netz des ersten Funktionsblocks (**108**) eine Mehrzahl von Neuroiden umfasst, wobei eine Verknüpfung der Neuroiden von einer Sprachcharakteristik einer Mehrzahl von Individuen abhängig ist.

14. Vorrichtung gemäß einem der Ansprüche 1 bis 13, bei der die Emotionsinformation teilweise aus der Ausgabe des ersten Funktionsblocks (**108**) bestimmbar ist.

15. Vorrichtung gemäß einem der Ansprüche 1 bis 14, die ferner folgendes Merkmal umfasst:
eine Einrichtung zum Zuordnen der Emotionsinformation (**106**) zu einem vordefinierten Emotionstyp.

16. Vorrichtung gemäß Anspruch 15, bei der die Einrichtung zum Zuordnen (**106**) ausgebildet ist, um eine Ausgabe des ersten Funktionsblocks (**108**) und eine Ausgabe des zweiten Funktionsblocks (**110**) zum Zuordnen der Emotionsinformation zu dem vordefinierten Emotionstyp additiv oder subtraktiv zu verknüpfen.

17. Vorrichtung gemäß Anspruch 4 und 11, bei der das neuronale Netz des zweiten Funktionsblocks (**110**) in das neuronale Netz des ersten Funktionsblocks (**108**) eingebettet ist.

18. Verfahren zum Erkennen einer in einem Sprachsignal enthaltenen Emotion mit folgendem Schritt:
Bestimmen einer Emotionsinformation aus Sprachsignalanalysewerten, die von dem Sprachsignal abgeleitet sind, wobei das Bestimmen folgende Schritte

umfasst:
Liefern einer Ausgabe aus einem Ergebnis gemäß einer festeingestellten Verarbeitungsvorschrift in einem ersten Funktionsblock (**108**), wobei die festeingestellte Verarbeitungsvorschrift eine Emotionserkennungsstandardeinstellung ist; und
Liefern einer Ausgabe aus einer Eingabe gemäß einer adaptierbaren Verarbeitungsvorschrift in einem zweiten Funktionsblock (**110**), wobei der zweite Funktionsblock (**110**) so ausgebildet ist, dass die adaptierbare Verarbeitungsvorschrift eine individuelle Adaption der Standardeinstellung der festeingestellten Verarbeitungsvorschrift an ein Individuum liefert, wenn eine Adaption mit einem Individuum ausgeführt wird, wobei der erste Funktionsblock (**108**) mit dem zweiten Funktionsblock (**110**) so gekoppelt wird, dass eine Ausgabe des ersten Funktionsblocks (**108**) als Eingabe des zweiten Funktionsblocks verwendet wird.

19. Computerprogramm mit Programmcode zur Durchführung des Verfahrens gemäß Anspruch 18, wenn das Programm auf einem Computer durchgeführt wird.

Es folgen 6 Blatt Zeichnungen

Anhängende Zeichnungen

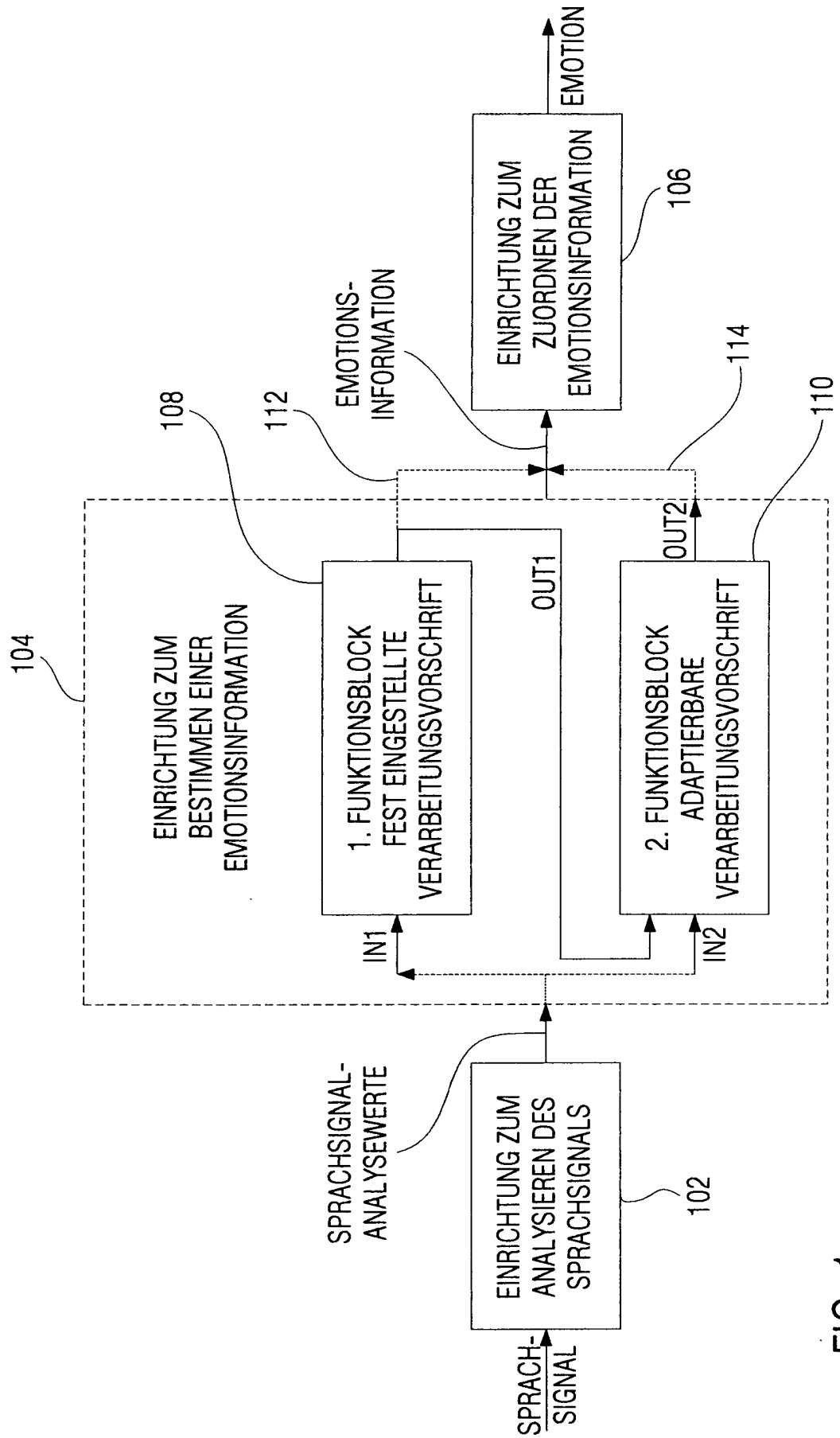


FIG. 1

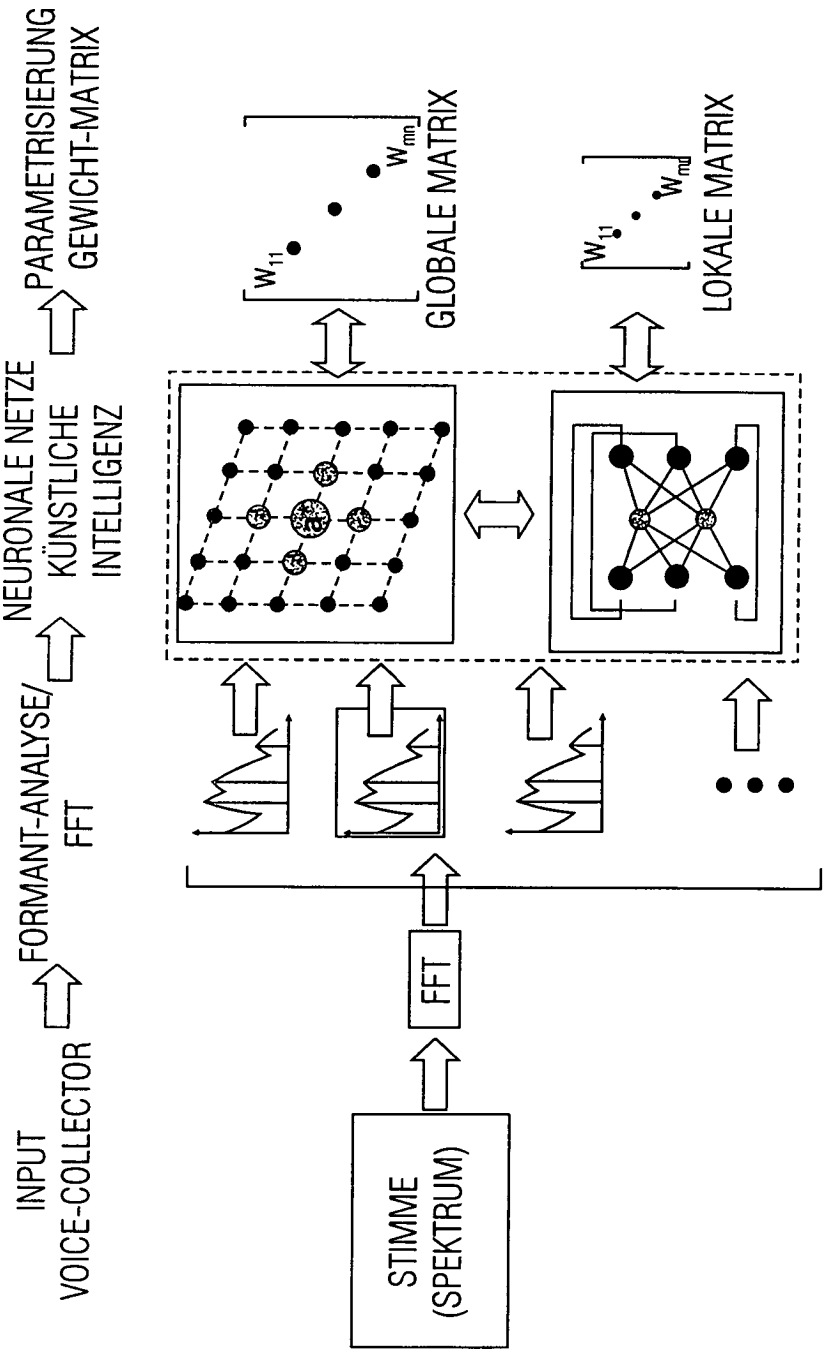

INPUT
VOICE-COLLECTOR
FORMANT-ANALYSE/
FFT
NEURONALE NETZE
KÜNSTLICHE
INTELLIGENZ
PARAMETRISIERUNG
GEWICHT-MATRIX
STIMME
(SPEKTRUM)
FFT
GLOBALE MATRIX
LOKALE MATRIX

FIG. 2

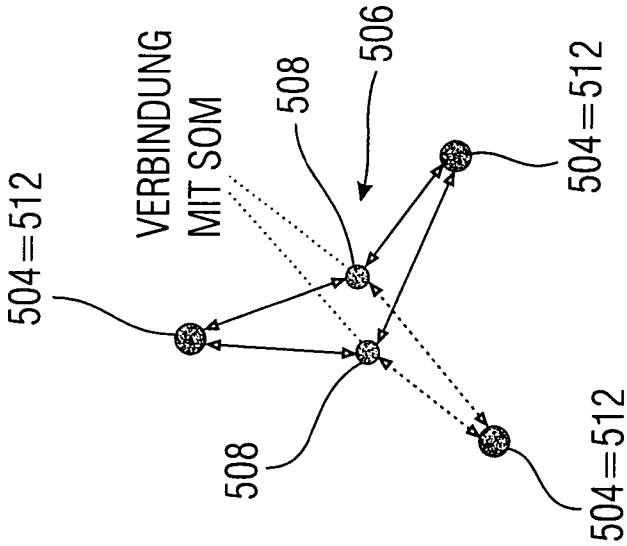

504=512
VERBINDUNG MIT SOM
508
506
504=512
504=512
508

FIG. 3B

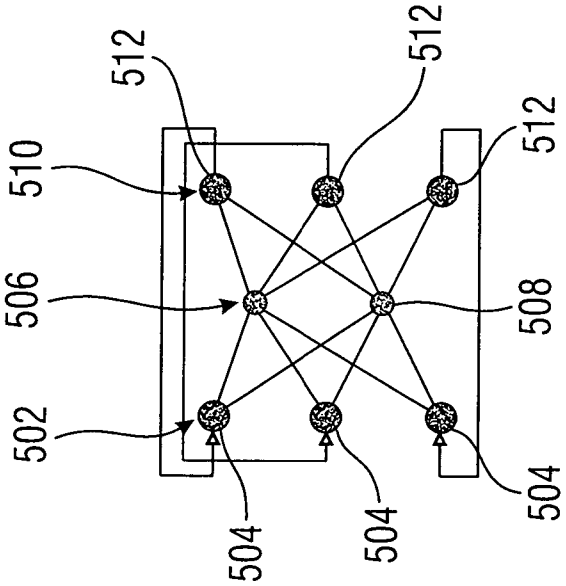

512
512
512
510
506
508
502
504
504
504

FIG. 3A

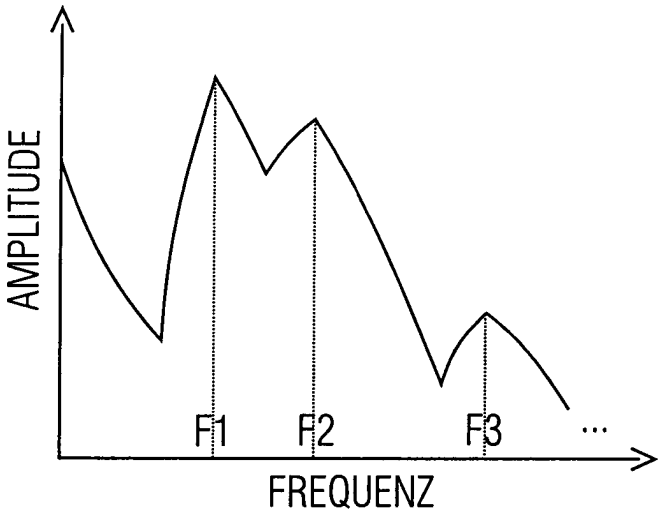


FIG. 4

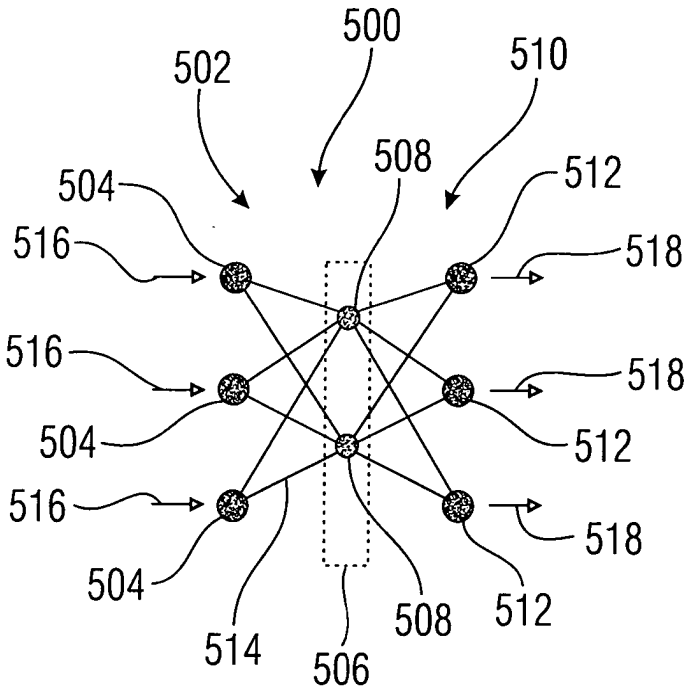

500
502
510
508
504
512
516
518
516
518
504
512
516
518
504
514
506
508
512

FIG. 5A

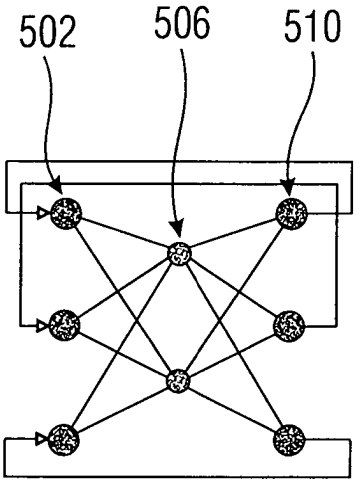

502
506
510

FIG. 5B

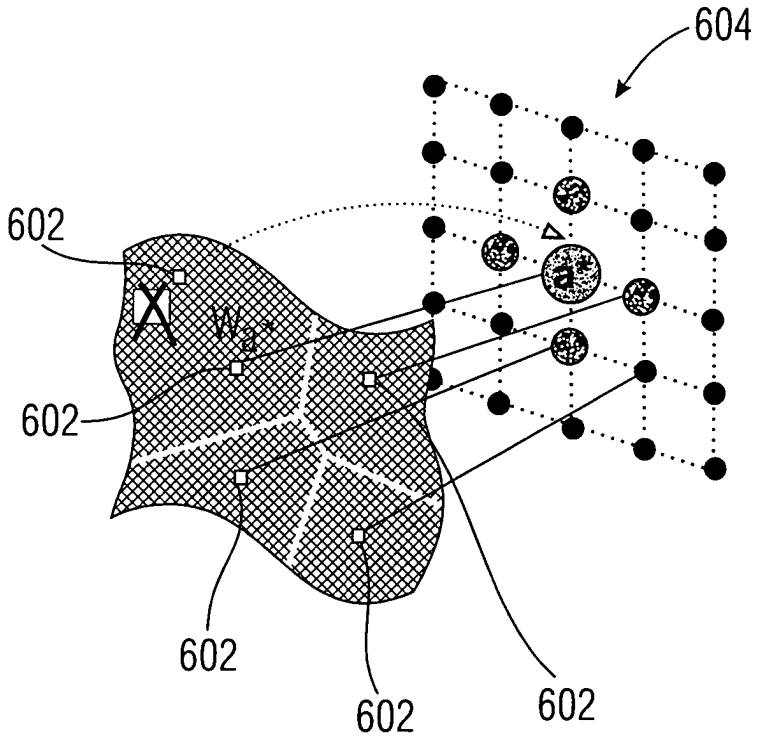


FIG. 6