



(11)

**EP 2 543 035 B1**

(12)

## EUROPÄISCHE PATENTSCHRIFT

(45) Veröffentlichungstag und Bekanntmachung des Hinweises auf die Patenterteilung:  
**11.12.2013 Patentblatt 2013/50**

(51) Int Cl.:  
**G10L 25/90** <sup>(2013.01)</sup>

(21) Anmeldenummer: **11708975.5**

(86) Internationale Anmeldenummer:  
**PCT/AT2011/000088**

(22) Anmeldetag: **22.02.2011**

(87) Internationale Veröffentlichungsnummer:  
**WO 2011/106809 (09.09.2011 Gazette 2011/36)**

(54) **VERFAHREN ZUR ERMITTLUNG VON GRUNDFREQUENZ-VERLÄUFEN MEHRERER SIGNALQUELLEN**

METHOD FOR MULTIPITCH TRACKING

PROCÉDÉ POUR LA DÉTERMINATION DE LA FRÉQUENCE FONDAMENTALE DE PLUSIEURS SOURCES DE SIGNAL

(84) Benannte Vertragsstaaten:  
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(74) Vertreter: **Patentanwaltskanzlei Matschnig & Forsthuber OG**  
**Siebensterngasse 54**  
**1071 Wien (AT)**

(30) Priorität: **01.03.2010 AT 3152010**

(43) Veröffentlichungstag der Anmeldung:  
**09.01.2013 Patentblatt 2013/02**

(73) Patentinhaber: **Technische Universität Graz**  
**8010 Graz (AT)**

(72) Erfinder:  
• **WOHLMAYR, Michael**  
**A-8010 Graz (AT)**  
• **STARK, Michael**  
**A-8010 Graz (AT)**  
• **PERNKOPF, Franz**  
**A-8043 Graz (AT)**

(56) Entgegenhaltungen:  
• **WOHLMAYR M ET AL: "A mixture maximization approach to multipitch tracking with factorial hidden Markov models", ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), 2010 IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 14. März 2010 (2010-03-14), Seiten 5070-5073, XP031697023, ISBN: 978-1-4244-4295-9**  
• **Michael Wohlmayr ET AL: "Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model", ISCA Interspeech 2009, 6. September 2009 (2009-09-06), 10. September 2009 (2009-09-10), Seiten 1079-1082, XP55002778, Brighton [gefunden am 2011-07-14]**

Anmerkung: Innerhalb von neun Monaten nach Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents im Europäischen Patentblatt kann jedermann nach Maßgabe der Ausführungsordnung beim Europäischen Patentamt gegen dieses Patent Einspruch einlegen. Der Einspruch gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist. (Art. 99(1) Europäisches Patentübereinkommen).

**EP 2 543 035 B1**

**Beschreibung**

**[0001]** Die Erfindung betrifft ein Verfahren zur Ermittlung von Grundfrequenz-Verläufen mehrerer Signalquellen aus einer einkanaligen Audioaufnahme eines Mischsignals.

**[0002]** Verfahren zur Verfolgung bzw. Trennung von einkanaligen Sprachsignalen über die wahrgenommene Grundfrequenz (der englische Fachbegriff "Pitch" wird im Rahmen der folgenden Ausführungen gleichbedeutend mit der wahrgenommenen Grundfrequenz verwendet) werden in einer Reihe von Algorithmen und Applikationen in der Sprach- und Audio-signalverarbeitung verwendet, wie z.B. bei der einkanaligen Quellentrennung (Single-Channel Blind Source Separation; SCSS) (D. Morgan et al., "Cochannel speaker separation by harmonic enhancement and suppression", IEEE Transactions on Speech and Audio Processing, Vol. 5, pp. 407-424, 1997), der Computational Auditory Scene Analysis (CASA) (DeLiang Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis", P. Divenyi [Ed], Speech Separation by Humans and Machines, Kluwer Academic, 2004) und der Sprachkompression (R. Salami et al., "A toll quality 8 kb/s speech codec for the personal communications system (PCS)", IEEE Transactions on Vehicular Technology, Vol. 43, pp. 808-816, 1994). Typische Anwendungen solcher Verfahren sind beispielsweise Konferenzsituationen, wo während eines Vortrags manchmal mehrere Stimmen hörbar sind und dadurch die Erkennungsrate einer automatischen Spracherkennung stark sinkt. Auch eine Anwendung in Hörgeräten ist möglich. Der Artikel "Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model", Michael Wohlmayr et al, ISCA Interspeech 2009, 6.-10. September 2009, Seiten 1079-1082, Brighton, offenbart ein Verfahren zur Ermittlung von Grundfrequenz-Verläufen mehrerer Signalquellen aus einer einkanaligen Audioaufnahme eines Mischsignals, bei dem in der Trainingsphase Mischspektren mit den jeweiligen Einzelsprecheranteilen trainiert werden. Zum Training werden gemischte Signale verwendet.

**[0003]** Die Grundfrequenz ist eine fundamentale Größe in der Analyse, Erkennung, Codierung, Kompression und Darstellung von Sprache. Sprachsignale lassen sich durch die Überlagerung von sinusförmigen Schwingungen beschreiben. Für stimmhafte Laute wie z.B. Vokale ist dabei die Frequenz dieser Schwingungen entweder die Grundfrequenz oder ein Vielfaches der Grundfrequenz, die sog. Oberschwingungen bzw. Obertöne. Damit lassen sich Sprachsignale durch Identifizierung der Grundfrequenz des Signals bestimmten Signalquellen zuordnen.

**[0004]** Während für den Fall eines einzelnen Sprechers bei rauscharmer Aufnahme bereits eine Reihe von erprobten Methoden für die Abschätzung bzw. die Verfolgung (Tracking) der Grundfrequenz in Verwendung ist, gibt es nach wie vor Probleme bei der Bearbeitung von minderwertigen (also mit Störgeräuschen wie Rauschen versehenen) Aufnahmen mehrerer gleichzeitig sprechender Personen.

**[0005]** Mingyang Wu et al. schlagen in "A Multipitch Tracking Algorithm for Noisy Speech" (IEEE Transactions on Speech and Audio Processing, Volume 11, Issue 3, pp. 229-241, May 2003) eine Lösung für robustes mehrfaches Grundfrequenztracking bei Aufnahmen mit mehreren Sprechern vor. Die Lösung basiert auf dem unitären Modell für Grundfrequenzwahrnehmung, für das verschiedene Verbesserungen vorgeschlagen werden, um eine Wahrscheinlichkeitstheoretische Darstellung der Periodizitäten des Signals zu erhalten. Die Nachverfolgung der Wahrscheinlichkeiten der Periodizitäten unter Verwendung des Hidden Markov Model (HMM) ermöglicht die Darstellung semikontinuierlicher Grundfrequenzverläufe. Nachteilig an dieser Lösung ist zum einen der hohe Rechenaufwand und die dadurch benötigten Rechnerressourcen, andererseits die Tatsache, dass eine ordnungsgemäße Zuordnung der Grundfrequenzen zu den passenden Signalquellen, bzw. Sprechern nicht möglich ist. Der Grund dafür ist die Tatsache, dass in diesem System keine sprecherspezifischen Informationen eingebunden werden bzw. zur Verfügung stehen, die eine derartige Verknüpfung von gemessenen Pitchwerten und Sprechern ermöglichen würde.

**[0006]** Es ist daher eine Aufgabe der Erfindung, ein Verfahren für mehrfaches Grundfrequenztracking bereitzustellen, das eine sichere Zuordnung der ermittelten Grundfrequenzen zu Signalquellen bzw. Sprechern erlaubt und gleichzeitig eine geringe Speicher- und Rechenintensivität aufweist.

**[0007]** Diese Aufgabe wird mit einem Verfahren der eingangs genannten Art erfindungsgemäß durch die folgenden Schritte gelöst:

- a) Ermitteln der Spektrogramm-Eigenschaften der Pitchzustände einzelner Signalquellen unter Benutzung von Trainingsdaten;
- b) Ermitteln der Wahrscheinlichkeiten der möglichen Grundfrequenzkombinationen der in dem Mischsignal enthaltenen Signalquellen durch Kombination der in a) ermittelten Eigenschaften mittels eines Interaktionsmodells;
- c) Tracken der Grundfrequenzverläufe der einzelnen Signalquellen im Mischsignal unter Verwendung der in b) ermittelten Wahrscheinlichkeiten.

**[0008]** Dank der Erfindung lässt sich eine hohe Genauigkeit des Trackens der mehrfachen Grundfrequenzen erreichen, bzw. können Grundfrequenzverläufe besser den jeweiligen Signalquellen bzw. Sprechern zugeordnet werden. Durch eine Trainingsphase a) unter Verwendung sprecherspezifischer Informationen und die Wahl eines geeigneten Interaktionsmodells in b) wird der rechnerische Aufwand deutlich minimiert, sodass die Methode rasch und ressourcenarm

durchgeführt werden kann. Dabei werden nicht Mischspektren mit den jeweiligen Einzelsprecheranteilen (im einfachsten Fall zwei Sprecher und ein entsprechendes Grundfrequenz-Paar) trainiert, sondern die jeweiligen Einzelsprecheranteile, was den Rechenaufwand und die Zahl der durchzuführenden Trainingsphasen weiter minimiert. Da je Signalquelle Pitchzustände aus einem abgegrenzten Frequenzbereich (z.B. 80 bis 500 Hz) betrachtet werden, ergibt sich bei Kombination der Zustände in Schritt b) eine begrenzte Anzahl von Grundfrequenzkombinationen, die als "mögliche" Grundfrequenzkombinationen bezeichnet werden. Der Begriff Spektrum steht in weiterer Folge für das Magnituden-Spektrum; je nach Wahl des Interaktionsmodells in b) werden das Kurzzeit-Magnitudenspektrum oder das logarithmische Kurzzeit-Magnitudenspektrum (log-Spektrum) verwendet.

**[0009]** Die Anzahl der zu trainierenden Pitchzustände ergibt sich aus dem beobachteten Frequenzbereich und dessen Unterteilung (siehe weiter unten). Bei Sprachaufnahmen beträgt ein solcher Frequenzbereich beispielsweise 80 bis 500 Hz.

**[0010]** Aus Sprachmodellen einzelner Sprecher kann mit Hilfe des in b) angewandten Interaktionsmodells ein Wahrscheinlichkeitsmodell aller im oben genannten Frequenzbereich möglichen Pitchkombinationen, bzw. für ein gewünschtes Sprecherpaar (also z.B. für eine Aufnahme, auf der zwei Sprecher zu hören sind), erhalten werden. Bei der Annahme von zwei Sprechern mit jeweils A Zuständen bedeutet das also, dass eine  $A \times A$ -Matrix mit den Wahrscheinlichkeiten für alle möglichen Kombinationen ermittelt wird. Für die einzelnen Sprecher können auch Sprachmodelle verwendet werden, die eine Vielzahl von Sprechern beschreiben, beispielsweise, indem das Modell auf geschlechtsspezifische Merkmale abstellt (speaker-independent, bzw. gender-dependent).

**[0011]** Für das Tracking in c) kann eine Reihe von Algorithmen verwendet werden. Beispielsweise kann die zeitliche Abfolge der geschätzten Pitchwerte durch ein Hidden Markov Model (HMM) oder auch durch ein Factorial Hidden Markov Model (FHMM) modelliert werden, und auf diese graphischen Modelle können der Max-Sum Algorithmus, der Junction-Tree Algorithmus oder der Sum-Product-Algorithmus zum Einsatz kommen. In einer Variante der Erfindung ist es auch möglich, die auf isolierten Zeitfenstern geschätzten Pitchwerte unabhängig voneinander zu betrachten und auszuwerten, ohne einen der oben genannten Tracking-Algorithmen anzusetzen.

**[0012]** Für die Beschreibung der Spektrogramm-Eigenschaften kann ein allgemeines, parametrisches oder auch nicht-parametrisches statistisches Modell verwendet werden. Günstigerweise werden in a) die Spektrogramm-Eigenschaften mittels eines Gaussian Mixture Modells (GMM) ermittelt.

**[0013]** Vorteilhafterweise wird die Anzahl der Komponenten eines GMM durch Anwendung des Minimum-Description-Length (MDL) Criterion ermittelt. Das MDL-Criterion dient zur Wahl eines Modells aus einer Vielzahl möglicher Modelle. Beispielsweise unterscheiden sich die Modelle, wie im vorliegenden Fall, nur durch die Anzahl der verwendeten Gaussskomponenten. Neben dem MDL-Criterion ist beispielsweise auch die Verwendung des Akaike Information Criterion (AIC) möglich.

**[0014]** In b) werden als Interaktionsmodell ein lineares Modell oder das Mixture-Maximization (MixMax)-Interaktionsmodell oder das ALGONQUIN-Interaktionsmodell verwendet.

**[0015]** Günstigerweise erfolgt das Tracken in c) mittels des Factorial Hidden Markov Modells (FHMM).

**[0016]** Zur Durchführung des Trackings auf einem FHMM können eine Reihe von Algorithmen verwendet werden, beispielsweise werden in Varianten der Erfindung der Sum-Product Algorithmus oder der Max-Sum-Algorithmus verwendet.

**[0017]** Im Folgenden wird die Erfindung anhand eines nicht einschränkenden Ausführungsbeispiels, das in der Zeichnung dargestellt ist, näher erläutert. In dieser zeigt schematisch:

Fig. 1 einen Faktorgraphen der grundfrequenzabhängigen Erzeugung eines aus zwei Einzelsprecher(log)spektren resultierenden (log-)Spektrums  $y$  eines Mischsignals,

Fig. 2 eine Darstellung des FHMM, und

Fig. 3 ein Blockdiagramm des erfindungsgemäßen Verfahrens.

**[0018]** Die Erfindung betrifft ein einfaches und effizientes Modellierungsverfahren für das Grundfrequenztracking von mehreren gleichzeitig emittierenden Signalquellen, beispielsweise Sprechern in einer Konferenz- oder Besprechungssituation. Nachfolgend wird die erfindungsgemäße Methode aus Gründen der Nachvollziehbarkeit anhand von zwei Sprechern dargestellt, allerdings lässt sich das Verfahren auf eine beliebige Anzahl von Subjekten anwenden. Dabei sind die Sprachsignale einkanalig, werden also mit nur einem Aufnahmemittel - z.B. Mikrophon - aufgenommen.

**[0019]** Das Kurzzeitspektrum eines Sprachsignals bei gegebener Sprachgrundfrequenz lässt sich mit Hilfe von Wahrscheinlichkeitsverteilungen wie der Gauß'schen Normalverteilung beschreiben. Eine einzelne Normalverteilung, gegeben durch die Parameter Mittelwert  $\mu$  und Varianz  $\sigma^2$ , reicht dabei zumeist nicht aus. Zur Modellierung allgemeiner, komplexer Wahrscheinlichkeitsverteilungen verwendet man üblicherweise Mischverteilungen wie beispielsweise das Gaussian Mixture Model (bzw. Gauß'sches Mischverteilungsmodell - GMM). Das GMM setzt sich additiv aus mehreren

einzelnen Gauß'schen Normalverteilungen zusammen. Dabei kann eine M-fache Gaußverteilung mit  $3M-1$  Parametern beschrieben werden - Mittelwert, Varianz und Gewichtungsfaktor für jede der M Gaußverteilungen (der Gewichtungsfaktor der M-ten Gaußkomponente ist redundant, daher das "-1"). Für die Modellierung beobachteter Datenpunkte durch ein GMM wird häufig ein Spezialfall des "Expectation Maximization"-Algorithmus verwendet, wie weiter unten

beschreiben ist.

**[0020]** Der Verlauf der Pitchzustände eines Sprechers kann näherungsweise durch eine Markovkette beschrieben werden. Die Markov-Eigenschaft dieser Zustandsketten besagt, dass der Folgezustand nur vom gegenwärtigen Zustand abhängt und nicht von vorangegangenen Zuständen.

**[0021]** Bei der Analyse eines Sprachsignals zweier gleichzeitig sprechender Subjekte ist nur das resultierende Spektrum  $y^{(t)}$  der Mischung der beiden einzelnen Sprachsignale verfügbar, nicht aber die Pitchzustände  $x_1^{(t)}$  und  $x_2^{(t)}$  der Einzelsprecher. Der tiefgestellte Index bei den Pitchzuständen bezeichnet dabei Sprecher 1 und 2, während der hochgestellte Zeitindex von  $t=1, \dots, T$  verläuft. Diese einzelnen Pitchzustände sind verborgene Variablen. Zur Auswertung wird beispielsweise ein Hidden Markov Modell (HMM) verwendet, bei dem aus den beobachtbaren Zuständen (hier also aus dem resultierenden Spektrum  $y^{(t)}$  der Mischung) auf die verborgenen Variablen bzw. Zustände geschlossen wird.

**[0022]** Jede verborgene Variable hat im beschriebenen Ausführungsbeispiel  $|X|=170$  Zustände mit Grundfrequenzen aus dem Intervall von 80 bis 500 Hz. Natürlich können auch mehr oder weniger Zustände aus anderen Grundfrequenzintervallen verwendet werden.

**[0023]** Der Zustand "1" bedeutet "kein Pitch" (stimmlos bzw. keine Sprachaktivität), während Zustandswerte "2" bis "170" verschiedene Grundfrequenzen zwischen den oben genannten Werten bezeichnen. Im Speziellen wird der Pitch-

Wert  $f_0$  für die Zustände  $x>1$  nach der Formel  $f_0 = \frac{f_s}{30+x}$  ermittelt. Die Samplingrate beträgt  $f_s=16$  kHz. Das Pitch-

Intervall wird also uneinheitlich aufgelöst; niedrige Pitch-Werte haben eine feinere Auflösung als hohe Pitch-Werte: Die Zustände 168, 169 und 170 haben Grundfrequenzen von 80,80 Hz ( $x=168$ ), 80,40 Hz ( $x=169$ ) und 80,00 Hz ( $x=170$ ), während die Zustände 2, 3 und 4 die Grundfrequenzen 500,00 Hz ( $x=2$ ), 484,84 Hz ( $x=3$ ) und 470,58 Hz ( $x=4$ ) haben.

**[0024]** Das erfindungsgemäße Verfahren umfasst im beschriebenen Ausführungsbeispiel die folgenden Schritte:

- **Trainingsphase:** Trainieren eines sprecherabhängigen GMM zur Modellierung des Kurzzeitspektrums für jeden der 170 Zustände (169 Grundfrequenzzustände sowie der Zustand "kein Pitch") jedes Einzelsprechers;
- **Interaktions-Modell:** Ermitteln einer wahrscheinlichkeitstheoretischen Darstellung für die Mischung der beiden Einzelsprecher unter Anwendung eines Interaktionsmodells, z.B. des MixMax-Interaktionsmodells; abhängig von der Wahl des Interaktionsmodells wird in der Trainingsphase entweder das Kurzzeit-Magnitudenspektrum oder das logarithmische Kurzzeit-Magnituden-Spektrum modelliert.
- **Tracking:** Ermitteln der Grundfrequenztrajektorien der beiden Einzelsprecher unter Verwendung eines geeigneten Tracking Algorithmus, z.B. Junction-Tree oder Sum-Product (im vorliegenden Ausführungsbeispiel wird die Anwendung des Factorial Hidden Markov Modells (FHMM) beschrieben).

## Trainingsphase

**[0025]** In der erfindungsgemäßen Methode wird ein überwachtes Szenario angenommen, in dem die Sprachsignale der Einzelsprecher unter Ausnutzung von Trainingsdaten modelliert werden. Dabei können im Prinzip alle überwachten Trainingsrhythmen eingesetzt werden, also generative und diskriminative. Die Spektrogramm-Eigenschaften lassen sich durch ein allgemeines, parametrisches oder auch nicht-parametrisches statistisches Modell  $p(s_i|x_i)$  beschreiben. Die Verwendung von GMMs stellt also einen Spezialfall dar.

**[0026]** Im vorliegenden Ausführungsbeispiel werden unter Verwendung des EM (Expectation-Maximization)-Algorithmus für jeden Sprecher 170 GMMs trainiert (ein GMM pro Pitch-Zustand). Bei den Trainingsdaten handelt es sich beispielsweise um Tonaufnahmen von Einzelsprechern, also einen Satz von  $N_i$  log-Spektren von  $i$  Einzelsprechern,

$S_i = \{s_i^{(1)}, \dots, s_i^{(N_i)}\}$ , zusammen mit den zugehörigen Pitch-Werten  $\{x_i^{(1)}, \dots, x_i^{(N_i)}\}$ . Diese Daten können automatisch mit einem Pitchtracker aus Einzelsprecheräußerungen erzeugt werden.

**[0027]** Der EM-Algorithmus ist eine iterative Optimierungsmethode zum Abschätzen unbekannter Parameter beim Vorliegen von bekannten Daten wie Trainingsdaten. Dabei wird iterativ durch abwechselnde Klassifikation (Expectation-Schritt) und eine anschließende Anpassung der Modellparameter (Maximization-Schritt) die Wahrscheinlichkeit für das Auftreten eines stochastischen Prozesses bei einem vorgegebenen Modell maximiert.

**[0028]** Da der stochastische Prozess - im vorliegenden Fall das Spektrum des Sprachsignals - durch die Trainingsdaten gegeben ist, müssen zur Maximierung die Modellparameter angepasst werden. Die Voraussetzung für das Auffinden

dieses Maximums ist, dass nach jedem Induktionsschritt und der Berechnung eines neuen Modells der Likelihood des Modells ansteigt. Zur Initialisierung des Lernalgorithmus wird eine Anzahl überlagerter Gaussverteilungen und ein GMM mit beliebigen Parametern (z.B. Mittelwert, Varianz und Gewichtungsfaktoren) gewählt.

**[0029]** Durch die iterative Maximum-Likelihood (ML)-Schätzung des EM erhält man also ein repräsentatives Modell

für das Einzelsprecher-Sprachsignal, im vorliegenden Fall ein sprecherabhängiges GMM  $p(s_i | \Theta_{i,x_i}^{M_{i,x_i}})$ . Damit müssen für jeden Sprecher 170 GMMs trainiert werden, also ein GMM für jeden Pitch-Zustand  $x_i$  entsprechend der oben definierten Anzahl von Zuständen.

**[0030]** Die Modellierung der zustandsabhängigen log-Einzelspektren der Sprecher mittels GMM im vorliegenden Ausführungsbeispiel erfolgt also gemäß

$$p(s_i | x_i) = p(s_i | \Theta_{i,x_i}^{M_{i,x_i}}) = \sum_{m=1}^{M_{i,x_i}} \alpha_{i,x_i}^m NV(s_i | \theta_{i,x_i}^m), \text{ mit } i \in \{1,2\}.$$

**[0031]**  $M_{i,x_i} \geq 1$  bezeichnet dabei die Anzahl der Mischungskomponenten (also der Normalverteilungen, die zur Darstellung des Spektrums notwendig sind), ist der Gewichtungsfaktor jeder Komponente  $m=1, \dots, M_{i,x_i}$  "NV" bezeichnet die Normalverteilung.

**[0032]** Der Gewichtungsfaktor  $\alpha_{i,x_i}^m$  muss positiv sein –  $\alpha_{i,x_i}^m \geq 0$  – und die Normierungsbedingung  $\sum_{m=1}^{M_{i,x_i}} \alpha_{i,x_i}^m = 1$

erfüllen. Das zugehörige GMM ist vollständig bestimmt durch die Parameter  $\Theta_{i,x_i}^{M_{i,x_i}} = \{\alpha_{i,x_i}^m, \theta_{i,x_i}^m\}_{m=1}^{M_{i,x_i}}$  mit

$\theta_{i,x_i}^m = \{\mu_{i,x_i}^m, \Sigma_{i,x_i}^m\}$ ;  $\mu$  steht dabei für den Mittelwert,  $\Sigma$  bezeichnet die Kovarianz.

**[0033]** Nach der Trainingsphase liegen also GMMs für alle Grundfrequenzwerte aller Sprecher vor. Im vorliegenden Ausführungsbeispiel bedeutet das: Zwei Sprecher mit je 170 Zuständen aus dem Frequenzintervall 80 bis 500 Hz. Es sei noch einmal darauf hingewiesen, dass es sich hier um ein Ausführungsbeispiel handelt und das Verfahren auch auf mehrere Signalquellen und andere Frequenzintervalle anwendbar ist.

### Interaktions-Modell

**[0034]** Zur Analyse werden die aufgenommenen und mit einer Samplingfrequenz von beispielsweise  $f_s=16\text{kHz}$  gesampelten einkanaligen Sprachsignale zeitabschnittsweise betrachtet. In jedem Zeitabschnitt  $t$  wird das beobachtete (log-)Spektrum  $y^{(t)}$  des Mischsignals, also der Mischung der beiden Einzelsprechersignale, mit der Beobachtungswahrscheinlichkeit  $p(y^{(t)} | x_1^{(t)}, x_2^{(t)})$  modelliert. Anhand dieser Beobachtungswahrscheinlichkeit können beispielsweise die zu jedem Zeitpunkt wahrscheinlichsten Pitch-Zustände beider Sprecher ermittelt werden, oder die Beobachtungswahrscheinlichkeit dient direkt als Input für den in Schritt c) verwendeten Tracking-Algorithmus.

**[0035]** Im Prinzip lassen sich die (log-)Spektren der einzelnen Sprecher, bzw.  $p(s_1 | x_1)$  und  $p(s_2 | x_2)$ , zum Mischsignal  $y$  addieren; die Magnituden-Spektren addieren sich näherungsweise, daher gilt für die log-Magnitudenspektren:  $y = \log(\exp(s_1) + \exp(s_2))$ . Die Wahrscheinlichkeitsverteilung des Mischsignals ist also eine Funktion der beiden Einzelsignale,  $p(y) = f(p(s_1), p(s_2))$ . Die Funktion hängt nun davon ab, welches Interaktionsmodell gewählt wird.

**[0036]** Dafür sind mehrere Herangehensweisen möglich. Beim linearen Modell werden im Magnitudenspektrogramm die Einzelspektren gemäß der oben angegebenen Form addiert, das Mischsignal ist also näherungsweise die Summe der Magnitudenspektren der Einzelsprecher. Vereinfacht ausgedrückt bildet also die Summe der Wahrscheinlichkeitsverteilungen der beiden Einzelsprecher,  $NV(s_1 | \mu_1, \Sigma_1)$  und  $NV(s_2 | \mu_2, \Sigma_2)$ , die Wahrscheinlichkeitsverteilung des Mischsignals  $NV(y | \mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$ , wobei hier nur aus Gründen der besseren Verständlichkeit Normalverteilungen angeführt sind – gemäß der erfindungsgemäßen Methode handelt es sich bei den Wahrscheinlichkeitsverteilungen um GMMs.

**[0037]** Im dargestellten Ausführungsbeispiel des erfindungsgemäßen Verfahrens wird ein weiteres Interaktionsmodell verwendet: Nach dem MixMax-Interaktionsmodell kann das log-Spektrogramm zweier Sprecher durch das elementweise Maximum der log-Spektren der Einzelsprecher approximiert werden. Damit ist es möglich, rasch ein gutes Wahrscheinlichkeitsmodell des beobachteten Mischsignals zu erhalten. Dadurch werden auch Dauer und Rechenaufwand der Lernphase drastisch reduziert.

**[0038]** Für jeden Zeitabschnitt  $t$  gilt  $y^{(t)} \equiv \max(s_1^{(t)}, s_2^{(t)})$ , wobei  $s_i^{(t)}$  das log-Magnitudenspektrum des Sprechers  $i$  ist. Das log-Magnitudenspektrum  $y^{(t)}$  wird also erzeugt mittels eines stochastischen Modells, wie es in Fig. dargestellt ist.

**[0039]** Darin produzieren die beiden Sprecher ( $i=1, 2$ ) je ein log-Magnitudenspektrum  $s_i^{(t)}$  in Abhängigkeit des Grundfrequenzzustandes  $x_i^{(t)}$ . Das beobachtete log-Magnitudenspektrum  $y^{(t)}$  des Mischsignals wird approximiert durch die elementweisen Maxima beider Einzelsprecher-log-Magnitudenspektren. Mit anderen Worten: Für jeden Frame des Zeitsignals (Samples des Zeitsignals werden in Frames zusammengefasst, und aus Samples eines Frames wird dann mittels FFT (Fast Fourier Transformation) und unter Ausschluss der Phaseninformation das Kurzzeitmagnitudenspektrum berechnet) wird das logarithmische Magnitudenspektrogramm des Mischsignals angenähert durch das elementweise Maximum beider logarithmischer Einzelsprecher-Spektren. Anstatt die nicht zugänglichen Sprachsignale der Einzelsprecher zu betrachten, werden die Wahrscheinlichkeiten der Spektren betrachtet, die zuvor einzeln gelernt werden konnten.

**[0040]** Für einen fixen Grundfrequenz-Wert in Bezug auf einen Zustand  $x_i^{(1)}$  erzeugt Sprecher  $i$  ein log-Spektrum,  $s_i^{(t)}$ , das eine Realisierung der durch das Einzelsprechermodell  $p(s_i^{(t)}|x_i^{(t)})$  beschriebenen Verteilung darstellt.

**[0041]** Die beiden log-Spektren werden dann durch den elementweisen Maximum-Operator kombiniert, um das beobachtbare log-Spektrum  $y^{(t)}$  zu formen. Damit ist also  $p(y^{(t)} | s_1^{(t)}, s_2^{(t)}) = \delta(y^{(t)} - \max(s_1^{(t)}, s_2^{(t)}))$ , wobei  $\delta(\cdot)$  die Dirac'sche Deltafunktion bezeichnet.

**[0042]** Bei Verwendung des MixMax-Interaktionsmodells müssen also die GMMs für jeden Zustand jedes Sprechers ermittelt werden, also zwei Mal die Kardinalität der Zustandsvariablen. Bei herkömmlichen Modellen resultieren bei den angenommenen 170 verschiedenen Grundfrequenzzuständen für jeden Sprecher insgesamt 28900 verschiedene Grundfrequenzpaarungen, was einen deutlich erhöhten Rechenaufwand bewirkt.

**[0043]** Neben dem linearen Modell und dem MixMax-Interaktionsmodell können auch noch andere Modelle verwendet werden. Ein Beispiel dafür ist das Algonquin-Modell, wie es beispielsweise Brendan J. Frey et al. in "ALGONQUIN - Learning dynamic noise models from noisy speech for robust speech recognition" (Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, pp. 1165-1172, January 2002) beschreiben.

**[0044]** Wie auch beim MixMax-Interaktionsmodell wird mit dem Algonquin-Modell das log-Magnitudenspektrum der Mischung zweier Sprecher modelliert. Während beim MixMax-Interaktionsmodell  $y = \max(s_1, s_2)$  gilt, hat das Algonquin-Modell folgende Form:  $y = s_1 + \log(1 + \exp(s_2 - s_1))$ . Daraus kann wiederum die Wahrscheinlichkeitsverteilung des Mischsignals aus der Wahrscheinlichkeitsverteilung der Einzelsprechersignale abgeleitet werden.

**[0045]** Wie schon erwähnt wird im dargestellten Ausführungsbeispiel des erfindungsgemäßen Verfahrens nur das MixMax-Interaktionsmodell behandelt.

## Tracking

**[0046]** Die Aufgabe des Tracking beinhaltet im Prinzip das Suchen einer Sequenz von verborgenen Zuständen  $x^*$ , die die bedingte Wahrscheinlichkeitsverteilung  $x^* = \arg \max_x p(x|y)$  maximiert. Für das Tracking der Pitchverläufe über die Zeit wird im beschriebenen Ausführungsbeispiel des erfindungsgemäßen Verfahrens ein FHMM verwendet. Das FHMM erlaubt die Zustände mehrerer zeitlich parallel verlaufender Markovketten zu verfolgen, wobei die verfügbaren Beobachtungen als gemeinsamer Effekt aller einzelnen Markovketten betrachtet werden. Dabei werden die unter dem Punkt "Interaktionsmodell" beschriebenen Ergebnisse verwendet.

**[0047]** Bei einem FHMM werden also mehrere Markov-Ketten parallel betrachtet, wie es beispielsweise in dem beschriebenen Ausführungsbeispiel der Fall ist, wo zwei Sprecher gleichzeitig sprechen. Die Situation, die sich damit ergibt, ist in Fig. 2 dargestellt.

**[0048]** Wie oben erwähnt, werden die verborgenen Zustandsvariablen der Einzelsprecher mit  $x_k^{(t)}$  bezeichnet, wobei  $k$  die Markovketten (und damit die Sprecher) bezeichnet und der Zeitindex  $t$  von 1 bis  $T$  verläuft. Die Markovketten 1, 2 sind in Fig. 2 die horizontal verlaufend dargestellt. Die Annahme lautet, dass alle verborgenen Zustandsvariablen die Kardinalität  $|X|$  haben, im beschriebenen Ausführungsbeispiel also 170 Zustände. Die beobachtete Zufallsvariable wird mit  $y^{(t)}$  bezeichnet.

**[0049]** Die Abhängigkeit der verborgenen Variablen zwischen zwei aufeinander folgenden Zeitabschnitten ist definiert mit der Übergangswahrscheinlichkeit  $p(x_k^{(t)}|x_k^{(t-1)})$ . Die Abhängigkeit der beobachteten Zufallsvariablen  $y^{(t)}$  von den verborgenen Variablen desselben Zeitabschnitts ist definiert mit der Beobachtungswahrscheinlichkeit  $p(y^{(t)}|x_1^{(t)}, x_2^{(t)})$ , die, wie weiter oben bereits erwähnt, mittels eines Interaktionsmodelles erstellt werden kann. Die Ausgangswahrscheinlichkeit der verborgenen Variablen in jeder Kette ist gegeben als  $p(x_k^{(1)})$ .

**[0050]** Die gesamte Sequenz der Variablen lautet  $x = \prod_{t=1}^T \{x_1^{(t)}, x_2^{(t)}\}$  und  $y = \prod_{t=1}^T \{y^{(t)}\}$ , ergibt sich für die gemeinsame Verteilung aller Variablen der folgende Ausdruck:

$$p(x, y) = p(y | x) p(x) = \prod_{k=1}^2 p(x_k^{(1)}) \prod_{t=2}^T p(x_k^{(t)} | x_k^{(t-1)}) \prod_{t=1}^T p(y^{(t)} | x_1^{(t)}, x_2^{(t)}).$$

**[0051]** Beim FHMM ergeben sich je Markov-Kette eine  $|X| \times |X|$  Übergangsmatrix zwischen zwei verborgenen Zuständen - beim HMM wäre eine  $|X^2| \times |X^2|$  Übergangsmatrix erlaubt, also ungleich grösser.

**[0052]** Die Beobachtungswahrscheinlichkeit  $p(y^{(t)} | x_1^{(t)}, x_2^{(t)})$  ergibt sich im Allgemeinen mittels Marginalisierung über die unbekannten (log-)Spektren der Einzelsprecher:

$$p(y^{(t)} | x_1^{(t)}, x_2^{(t)}) = \iint p(y^{(t)} | s_1^{(t)}, s_2^{(t)}) p(s_1^{(t)} | x_1^{(t)}) p(s_2^{(t)} | x_2^{(t)}) d s_1^{(t)} d s_2^{(t)} \quad (1),$$

wobei  $p(y^{(t)} | s_1^{(t)}, s_2^{(t)})$  das Interaktionsmodell darstellt.

**[0053]** Damit ergibt sich für (1) bei Verwendung sprecherspezifischer GMMs, Marginalisierung über  $s_i$  und unter Verwendung des MixMax-Modells die folgende Darstellung:

$$p(y | x_1, x_2) = \sum_{m=1}^{M_{1,x_1}} \sum_{n=1}^{M_{2,x_2}} \alpha_{1,x_1}^m \alpha_{2,x_2}^n \prod_{d=1}^D \{ NV(y_d | \theta_{1,x_1}^{m,d}) \phi(y_d | \theta_{2,x_2}^{n,d}) + \phi(y_d | \theta_{1,x_1}^{m,d}) NV(y_d | \theta_{2,x_2}^{n,d}) \},$$

wobei  $y_d$  das d-te Element des log-Spektrums  $y$  ergibt,  $\theta_{i,x_i}^{m,d}$  gibt das d-te Element des zugehörigen Mittelwerts und

der Varianz, und  $\phi(y | \theta) = \int_{-\infty}^{\infty} NV(x | \theta) dx$  stellt die univariate kumulative Normalverteilung dar.

**[0054]** Gleichermassen ergibt sich für (1) bei Verwendung des linearen Interaktionsmodells die folgende Darstellung:

$$p(y | x_1, x_2) = \sum_{m=1}^{M_{1,x_1}} \sum_{n=1}^{M_{2,x_2}} \alpha_{1,x_1}^m \alpha_{2,x_2}^n NV(y | \mu_{1,x_1}^m + \mu_{2,x_2}^n, \Sigma_{1,x_1}^m + \Sigma_{2,x_2}^n),$$

wobei  $y$  das Spektrum des Mischsignals ist.

**[0055]** Fig. 3 zeigt in einer schematischen Darstellung den Ablauf des erfindungsgemäßen Verfahrens anhand eines Blockdiagramms.

**[0056]** Ein Sprachsignal, bzw. ein Signalgemisch aus mehreren Einzelsignalen, wird einkanalig aufgenommen, beispielsweise mit einem Mikrophon. Dieser Verfahrensschritt ist im Blockdiagramm mit 100 bezeichnet.

**[0057]** In einem unabhängigen Verfahrensschritt, der beispielsweise im Vorfeld der Anwendung des Verfahrens durchgeführt wird, werden in einer Trainingsphase 101 die Sprachsignale der Einzelsprecher unter Ausnutzung von Trainingsdaten modelliert. Unter Verwendung des EM (Expectation-Maximization)-Algorithmus wird für jeden der 170 Pitch-Zustände jeweils ein sprecherabhängiges GMM trainiert. Die Trainingsphase erfolgt für alle möglichen Zustände - im beschriebenen Ausführungsbeispiel sind das für zwei Sprecher je 170 Zustände zwischen 80 und 500 Hz. Mit anderen Worten wird also ein grundfrequenzabhängiges Spektrogramm von jedem Sprecher mittels GMM trainiert, wobei das MDL-Criterion angewendet wird, um die optimale Anzahl von Gauss-Komponenten aufzufinden. In einem weiteren Schritt 102 werden die GMMs, bzw. die zugehörigen Parameter, abgelegt, beispielsweise in einer Datenbank.

**[0058]** 103: Um eine wahrscheinlichkeitstheoretische Wiedergabe des Mischsignals von zwei oder mehr Sprechenden bzw. der Einzelsignalanteile des Mischsignals zu erhalten, wird ein Interaktionsmodell, vorzugsweise das MixMax-Interaktionsmodell, angewandt. Anschließend wird im Rahmen des Trackens 104 der Grundfrequenzverläufe das FHMM angewandt. Mittels FHMM ist es möglich, die Zustände mehrerer verborgener Markov-Prozesse zu tracken, die gleichzeitig ablaufen, wobei die verfügbaren Beobachtungen als Auswirkungen der einzelnen Markov-Prozesse betrachtet werden.

## Patentansprüche

1. Verfahren zur Ermittlung von Grundfrequenz-Verläufen mehrerer Signalquellen aus einer einkanaligen Audioaufnahme eines Mischsignals, umfassend die folgenden Schritte:

- a) Ermitteln der Spektrogramm-Eigenschaften der Pitchzustände einzelner Signalquellen unter Benutzung von Trainingsdaten;
- b) Ermitteln der Wahrscheinlichkeiten der möglichen Grundfrequenzkombinationen der in dem Mischsignal enthaltenen Signalquellen durch Kombination der in a) ermittelten Eigenschaften mittels eines Interaktionsmodells;
- c) Tracken der Grundfrequenzverläufe der einzelnen Signalquellen im Mischsignal unter Verwendung der in b) ermittelten Wahrscheinlichkeiten.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet, dass** in a) die Spektrogramm-Eigenschaften mittels einem Gaussian Mixture Modell (GMM) ermittelt werden.

3. Verfahren nach Anspruch 2, **dadurch gekennzeichnet, dass** weiters das Minimum-Description-Length Criterion angewandt wird, um die Anzahl der Komponenten des GMM zu ermittelt.

4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet, dass** in b) als Interaktionsmodell ein lineares Modell oder das MixMax-Interaktionsmodell oder das ALGONQUIN-Interaktionsmodell verwendet werden.

5. Verfahren nach einem der Ansprüche 1 bis 4, **dadurch gekennzeichnet, dass** das Tracken in c) mittels des Factorial Hidden Markov Modells (FHMM) erfolgt.

6. Verfahren nach Anspruch 5, **dadurch gekennzeichnet, dass** zur Lösung des FHMM der Sum-Product Algorithmus oder der Max-Sum Algorithmus verwendet werden.

## Claims

1. A method for establishing fundamental frequency progressions of a plurality of signal sources from a single-channel audio recording of a mix signal, said method comprising the following steps:

- a) establishing the spectrogram properties of the pitch states of individual signal sources with use of training data;
- b) establishing the probabilities of the possible fundamental frequency combinations of the signal sources contained in the mix signal by a combination of the properties established in a) by means of an interaction model;
- c) tracking the fundamental frequency progressions of the individual signal sources in the mix signal using the probabilities established in b).

2. The method according to Claim 1, **characterised in that** the spectrogram properties are established in a) by means of a Gaussian mixture model (GMM).

3. The method according to Claim 2, **characterised in that** the minimum-description-length criterion is also applied so as to establish the number of components of the GMM.

4. The method according to one of Claims 1 to 3, **characterised in that** a linear model or the mix-max interaction model or the ALGONQUIN interaction model is used in b) as the interaction model.

5. The method according to one of Claims 1 to 4, **characterised in that** the tracking in c) is carried out by means of the factorial hidden Markov model (FHMM).

6. The method according to Claim 5, **characterised in that** the sum-product algorithm or the max-sum algorithm is used to solve the FHMM.

## Revendications

1. Procédé servant à déterminer des évolutions de fréquence fondamentale de plusieurs sources de signal à partir d'un enregistrement audio monocanal d'un signal mélangé, comprenant les étapes suivantes consistant à :

- a) déterminer les caractéristiques de spectrogramme des états de pas de diverses sources de signal en utilisant des données d'apprentissage ;
- b) déterminer les probabilités des combinaisons possibles de fréquence de base des sources de signal contenues dans le signal mélangé en combinant les caractéristiques déterminées lors de l'étape a) au moyen d'un modèle d'interaction ;
- c) suivre les évolutions de la fréquence de base des diverses sources de signal dans le signal mélangé en utilisant les probabilités déterminées à l'étape b).

2. Procédé selon la revendication 1, **caractérisé en ce que** les caractéristiques de spectrogramme déterminées lors de l'étape a) sont déterminées au moyen d'un modèle de mélange gaussien (ou GMM, Gaussian Mixture Model).

3. Procédé selon la revendication 2, **caractérisé en ce qu'**en outre on applique le critère de la longueur minimale de description (minimum-description-length criterion) afin de déterminer le nombre de composants du GMM.

4. Procédé selon l'une quelconque des revendications 1 à 3, **caractérisé en ce qu'**on applique lors de l'étape b) comme modèle d'interaction un modèle linéaire ou un modèle d'interaction MixMax ou le modèle d'interaction ALGONQUIN.

5. Procédé selon l'une quelconque des revendications 1 à 4, **caractérisé en ce que** le suivi à l'étape c) est effectué au moyen du modèle de Markov caché pondéré (factorial hidden Markov model (FHMM)).

6. Procédé selon la revendication 5, **caractérisé en ce qu'**on utilise pour résoudre le FHMM, l'algorithme somme-produit ou l'algorithme max-somme.

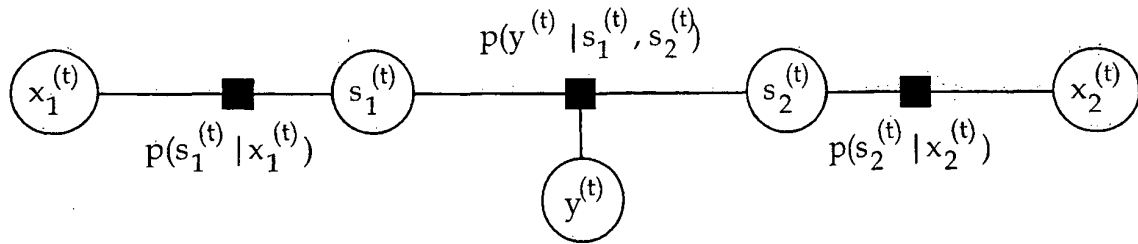


Fig. 1

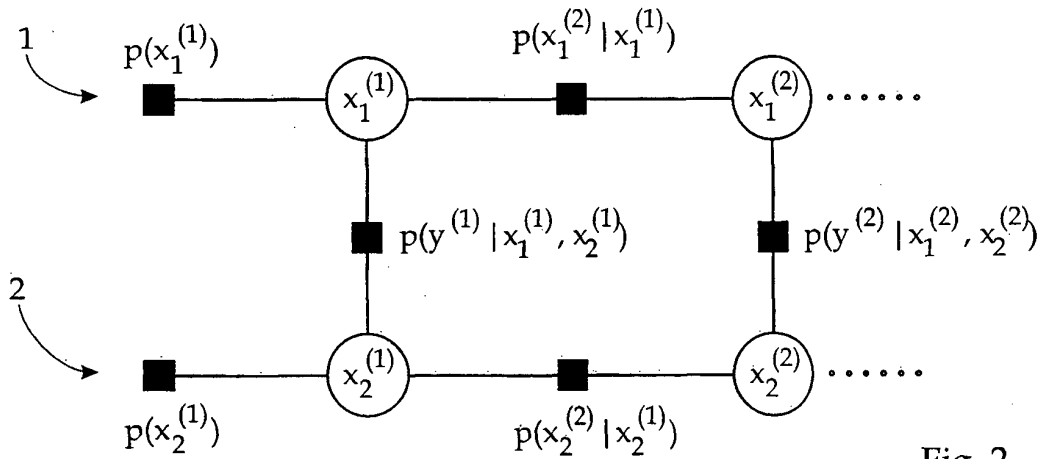


Fig. 2

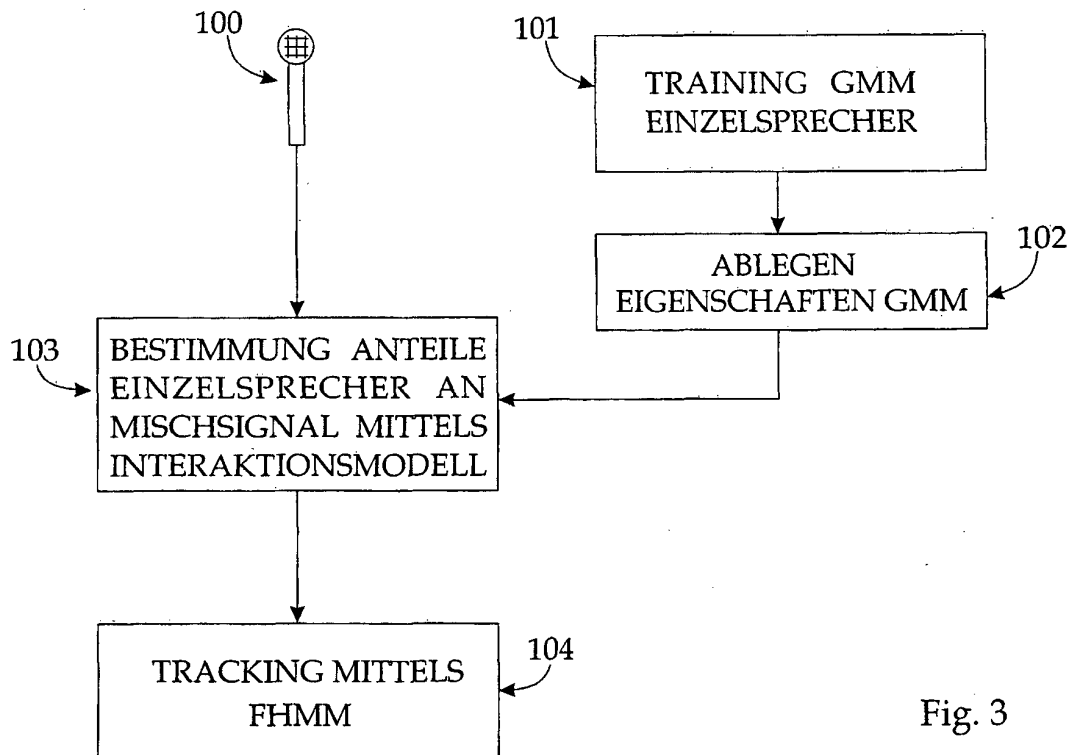


Fig. 3

## IN DER BESCHREIBUNG AUFGEFÜHRTE DOKUMENTE

*Diese Liste der vom Anmelder aufgeführten Dokumente wurde ausschließlich zur Information des Lesers aufgenommen und ist nicht Bestandteil des europäischen Patentdokumentes. Sie wurde mit größter Sorgfalt zusammengestellt; das EPA übernimmt jedoch keinerlei Haftung für etwaige Fehler oder Auslassungen.*

### In der Beschreibung aufgeführte Nicht-Patentliteratur

- **D. MORGAN et al.** Cochannel speaker separation by harmonic enhancement and suppression. *IEEE Transactions on Speech and Audio Processing*, 1997, vol. 5, 407-424 [0002]
- On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. **DELIANG WANG**. Speech Separation by Humans and Machines. Kluwer Academic, 2004 [0002]
- **R. SALAMI et al.** A toll quality 8 kb/s speech codec for the personal communications system (PCS. *IEEE Transactions on Vehicular Technology*, 1994, vol. 43, 808-816 [0002]
- **MICHAEL WOHLMAYR et al.** Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model. *ISCA Interspeech*, 06. September 2009, 1079-1082 [0002]
- **MINGYANG WU et al.** A Multipitch Tracking Algorithm for Noisy Speech. *IEEE Transactions on Speech and Audio Processing*, Mai 2003, vol. 11 (3), 229-241 [0005]
- ALGONQUIN - Learning dynamic noise models from noisy speech for robust speech recognition. **BRENDAN J. FREY et al.** Advances in Neural Information Processing Systems. MIT Press, Januar 2002, vol. 14, 1165-1172 [0043]