(19) Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 10 2004 048 707 B3** 2005.12.29

(12) **Patentschrift**

(21) Aktenzeichen: **10 2004 048 707.3**
(22) Anmeldetag: **06.10.2004**
(43) Offenlegungstag: **–**
(45) Veröffentlichungstag
der Patenterteilung: **29.12.2005**

(51) Int Cl.[7]: **G10L 13/00**

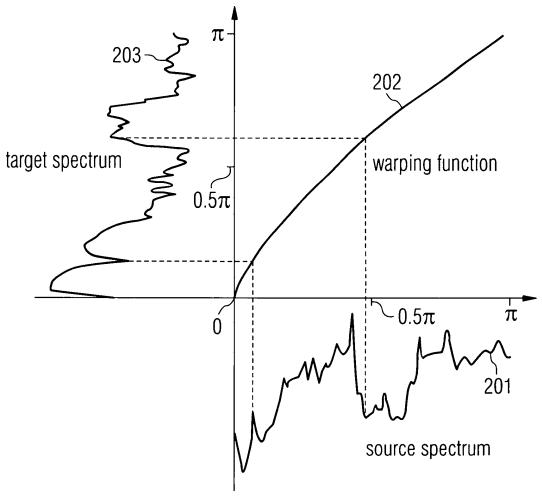
(73) Patentinhaber:
Siemens AG, 80333 München, DE

(72) Erfinder:
Sündermann, David, 06846 Dessau, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:
DE 698 11 656 T2
US 66 15 174 B1
US 64 63 412 B1
US 53 27 521
E. Eide und H. Gish: "A Parametric Approach to Vocal Tract Length Normalization", in Proc. of the ICASSP '96, Atlanta, USA, 1996;
D. Sündermann, H. Ney und H. Höge: "VTLN-Based Cross-Language Voice Conversion", in Proc. of the ASRU'03, St. Thomas, USA, 2003;

(54) Bezeichnung: **Verfahren zur Stimmenkonversion für ein Sprachsynthesesystem**

(57) Zusammenfassung: Verfahren zur Stimmenkonversion arbeiten üblicherweise im Frequenzbereich. Eine Anwendung dieser Verfahren zur Stimmenkonversion in konkatenativen Sprachsynthesesystemen ist daher mit zusätzlichem Rechenaufwand verbunden, da diese Sprachsynthesesysteme gängigerweise im Zeitbereich operieren.
Aufgabe der vorliegenden Erfindung ist es daher, ein Verfahren anzugeben, mit dem der erforderliche Speicherplatz und die benötigte Rechenleistung verringert werden. Errfindungsgemäß wird diese Aufgabe durch ein Verfahren gelöst, welches die Stimme im Zeitbereich manipuliert.



Beschreibung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zur Stimmenkonversion für ein Sprachsynthesesystem.

[0002] Als Stimmenkonversion bezeichnet man die Manipulation einer Stimme, so dass entweder eine bestimmte andere Stimme daraus entsteht oder die Eigenschaften der Stimme gezielt geändert werden. In der Sprachsynthese ist diese Technik von Interesse, um gezielt die Stimmcharakteristik eines gewünschten Sprechers zu imitieren. Die Stimmenkonversion kann auch eine interessante Funktion eines Mobilfunktelefons darstellen, mit dem man eine Stimme derart verändern kann, dass beispielsweise eine männliche Stimme wie eine weibliche Stimme klingt oder eine Stimme jünger oder erwachsener wirkt.

Stand der Technik

[0003] Aus der Spracherkennung ist das Verfahren der Vocal Tract Lenght Normalization (VTLN) bekannt, welches in E. Eide und H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", in Proc. of the ICASSP'96, Atlanta, USA, 1996 beschrieben ist. Hierbei wird durch eine Normalisierung der individuellen Stimmcharakteristik eine Angleichung von unterschiedlichen sprecherabhängigen Sprachsignalen erreicht, um die Spracherkennungsleistung zu verbessern. Das gleiche Verfahren wurde für die Stimmenkonversion umgesetzt und in D. Sündermann, H. Ney und H. Höge, "VTLN-Based Cross-Language Voice Conversion", in Proc. of the ASRU'03, St. Thomas, USA, 2003 beschrieben. Dabei kann eine Standardsynthesestimme in unterschiedliche Synthesestimmen transformiert werden. Hierzu wird das digitale Sprachzeitsignal einer Äußerung eines Quellsprechers mittels eines Pitch-Tracking Algorithmus in quasi-periodische Segmente unterteilt, die nun als Grundlage für die weitere Signalverarbeitung dienen. Mittels der diskreten Fourier-Transformation (DFT) wird das komplexwertige Spektrum jedes quasi-periodischen Segments berechnet. Durch Multiplikation mit einer Verzerrungsfunktion wird das Spektrum unter Verwendung der Spline-Interpolation in der gewünschten Weise verzerrt. Das resultierende Spektrum wird mit Hilfe der inversen diskreten Fourier-Transformation (IDFT) in den Zeitbereich transformiert. Das verzerrte Sprachzeitsignal wird dann durch Verkettung der transformierten quasi-periodischen Segmente erhalten.

[0004] Im Gegensatz zu Spracherkennungssystemen arbeiten konkatenative (verkettende) Sprachsynthesesysteme üblicherweise im Zeitbereich. Die Anwendung des oben beschriebenen Verfahrens zur Stimmenkonversion in einem Sprachsynthesesystem erfordert daher zusätzliche Hin- und Rücktransformationsoperationen in den Frequenzbereich. Dies erschwert einen breiten Einsatz dieses Verfahrens zur Stimmenkonversion in Sprachsynthesesystemen im Bereich integrierter Technologien, da hierbei jede vermeidbare zusätzliche Operation aufgrund der limitierten Prozessorressourcen vermieden werden muss.

Aufgabenstellung

[0005] Aufgabe der vorliegenden Erfindung ist es daher, ein Verfahren zur Stimmenkonversion anzugeben, mit dem der erforderliche Speicherplatz und die benötigte Rechenleistung verringert werden.

[0006] Erfindungsgemäß wird diese Aufgabe durch ein Verfahren mit den in Anspruch 1 angegebenen Merkmalen gelöst. Vorteilhafte Weiterbildungen der vorliegenden Erfindung sind in den abhängigen Ansprüchen angegeben.

[0007] Gemäß der vorliegenden Erfindung wird in einem Verfahren zur Stimmenkonversion für ein Sprachsynthesesystem ein erstes Sprachzeitsignal in zeitlich aufeinander folgende Segmente aufgeteilt. Die Segmente werden mit einer Verzerrungszeitfunktion gefaltet. Durch Zusammensetzen der gefalteten Segmente wird ein zweites Sprachzeitsignal erzeugt. Durch die Manipulation der Segmente zur Stimmenkonversion im Zeitbereich, entfallen die Hin- und Rücktransformationen in den Frequenzbereich. Dadurch kann die Rechenkomplexität und der benötigte Speicherplatz erheblich reduziert werden, wodurch eine schnellere Verarbeitung ermöglicht wird.

[0008] Erstes und/oder zweites Sprachzeitsignal können dabei als analoge und/oder digitale Sprachzeitsignale ausgeführt sein.

[0009] Gemäß einer vorteilhaften Ausführungsvariante der vorliegenden Erfindung wird das erste Sprachzeitsignal in quasiperiodische Segmente aufgeteilt. Die quasiperiodischen Segmente können beispielsweise mit Hilfe eines Pitch-Tracking Algorithmus ermittelt werden. Dieser ermittelt periodische Strukturen innerhalb eines Sprachzeitsignals mit einer speziellen Autokorrelationsfunktion.

[0010] Nach einer weiteren vorteilhaften Ausgestaltung der vorliegenden Erfindung ist mindestens ein Parameter der Verzerrungszeitfunktion vorgebbar. Dadurch wird auf schnelle und effiziente Weise eine Verfremdung der Stimme ermöglicht.

[0011] Entsprechend einer weiteren vorteilhaften Weiterbildung der vorliegenden Erfindung wird die Verzerrungszeitfunktion durch Transformation einer Verzerrungsfrequenzfunktion in den Zeitbereich ermittelt. Hierdurch ist es möglich, die Stimme gemäß einer gewünschten Frequenzcharakteristik gezielt zu verändern. Die Ermittlung der Verzerrungszeitfunkti-

on nach dieser Methode kann dabei vorab oder während der Ausführung des erfindungsgemäßen Verfahrens erfolgen.

[0012] Entsprechend einer weiteren vorteilhaften Ausgestaltung der vorliegenden Erfindung wird mindestens ein Parameter der Verzerrungsfrequenzfunktion durch Minimierung des euklidischen Abstandes zwischen einem Sprachfrequenzsignal eines Zielsprechers und einem mit dem mindestens einen Parameter der Verzerrungsfrequenzfunktion transformierten Sprachfrequenzsignal eines Quellsprechers ermittelt. Durch die so ermittelte Verzerrungsfrequenzfunktion kann die Stimme eines Quellsprechers an die Stimme eines Zielsprechers adaptiert werden.

[0013] Die Verzerrungsfrequenzfunktion kann beispielsweise als lineare Verzerrungsfrequenzfunktion mit einem Parameter ausgeführt sein. Dies hat den Vorteil, dass die Rechenkomplexität für die Stimmenkonversion mit der linearen Verzerrungsfrequenzfunktion erheblich reduziert wird (siehe Ausführungsbeispiel).

[0014] Nach weiteren vorteilhaften Ausführungsvarianten der vorliegenden Erfindung wird die Verzerrungsfrequenzfunktion als stückweise, lineare Verzerrungsfunktion mit einem Parameter oder als nichtlineare Verzerrungsfunktion mit einem Parameter ausgeführt. Außerdem ist die Verzerrungsfrequenzfunktion als stückweise, lineare Verzerrungsfunktion mit mehreren Parametern oder als nichtlineare Verzerrungsfunktion mit mehreren Parametern ausführbar. Die Flexibilität bei der Auswahl einer geeigneten Verzerrungsfunktion zur Stimmenkonversion hat die vorteilhafte Wirkung, dass je nach qualitativer Anforderung die Stimmenkonversion durch eine daran angepasste Verzerrungsfunktion erfolgen kann.

[0015] Gemäß einer weiteren Ausbildung der vorliegenden Erfindung werden durch die Faltung der Segmente mit der Verzerrungszeitfunktion entstandene Werte, welche außerhalb eines darstellbaren Wertebereichs liegen, interpoliert. Hierzu kann beispielsweise eine Spline-Interpolation oder eine lineare Interpolation eingesetzt werden.

[0016] Gemäß einer weiteren Ausbildungsform der vorliegenden Erfindung wird das Zusammensetzen der gefalteten Segmente mit einem TD-PSOLA Algorithmus ausgeführt. Dieser Algorithmus ermöglicht eine weitgehend artefakt- und verzerrungsfreie Verkettung von Segmenten, wobei gleichzeitig die Tonhöhe und Sprechgeschwindigkeit manipulierbar ist. Dieses Verfahren wurde 1988 bzw. 1989 unter verschiedenen Namen durch die France Telecom angemeldet und später patentiert (EP 0 363 233 B1 bzw. US 005327498 A).

[0017] In vorteilhafter Weise sind die Segmente des ersten Sprachzeitsignals durch Multiplikation mit einem vorgebbaren Verzerrungsfaktor im Zeit- und/oder Amplitudenbereich stauchbar bzw. streckbar. Somit wird eine erhebliche Reduzierung der benötigten Speicher- und Rechenkapazitäten erreicht.

Ausführungsbeispiel

[0018] Die vorliegende Erfindung wird nachfolgend an einem Ausführungsbeispiel anhand der Zeichnungen näher erläutert. Es zeigt

[0019] Fig. 1 ein Diagramm mit zwei Beispielen einer Verzerrungsfunktion,

[0020] Fig. 2 ein Beispiel eines Quellspektrums, das mit Hilfe einer Verzerrungsfunktion in ein Zielspektrum transformiert wird.

[0021] In Fig. 1 sind zwei Beispiele einer Verzerrungsfrequenzfunktion abgebildet. Auf der Abszissenachse sind die Frequenzen des Quellspektrums und auf der Ordinatenachse die Frequenzen des Zielspektrums aufgelistet. Die durchgezogene Linie zeigt eine stückweise, lineare Verzerrungsfrequenzfunktion **101** mit mehreren Parametern. Die gestrichelte Linie stellt eine lineare Verzerrungsfrequenzfunktion **102** mit einem Parameter dar.

[0022] Fig. 2 zeigt, wie im Frequenzbereich ein Quellspektrum **201** (dargestellt im rechten unteren Quadranten) mit Hilfe einer Verzerrungsfrequenzfunktion **202** in ein Zielspektrum **203** (dargestellt im linken oberen Quadranten) transformiert wird. Die Frequenzachse des Zielspektrums wird dabei in diesem Beispiel gestaucht bzw. gestreckt.

[0023] Gemäß der vorliegenden Erfindung wird in diesem Ausführungsbeispiel in einem ersten Schritt ein erstes digitales Sprachzeitsignal mit Hilfe eines Pitch-Tracking Algorithmus in zeitlich aufeinander folgende quasi-periodische Segmente aufgeteilt. Die quasi-periodischen Segmente entsprechen weitestgehend den Phonemen des ersten Sprachzeitsignals.

[0024] In einem zweiten Schritt werden die quasi-periodischen Segmente mit einer Verzerrungszeitfunktion gefaltet. Entsprechend der vorliegenden Erfindung soll die Manipulation des ersten Sprachzeitsignals durch die Verzerrungszeitfunktion im Zeitbereich stattfinden. Eine Multiplikation eines quasi-periodischen Segments eines Sprachfrequenzsignals mit einer Verzerrungsfrequenzfunktion im Frequenzbereich (Fig. 2) entspricht im Zeitbereich einer Faltung eines quasi-periodischen Segments eines Sprachzeitsignals mit einer Verzerrungszeitfunktion. In dem vorliegenden Ausführungsbeispiel soll als Verzerrungsfrequenzfunktion eine im Frequenzbereich line-

are Verzerrungsfrequenzfunktion mit einem Parameter angewandt werden. Eine solche Verzerrungsfrequenzfunktion wird durch ihren einen Parameter vollständig bestimmt, der durch einen Anwender frei vorgebbar sein soll. Transformiert man diese Verzerrungsfrequenzfunktion in den Zeitbereich und faltet sie mit den quasi-periodischen Segmenten des ersten Sprachzeitsignals, erhält man die quasi-periodischen Segmente des resultierenden zweiten Sprachzeitsignals. Für diese einfache Verzerrungsfunktion lassen sich die quasi-periodischen Segmente des resultierenden zweiten Sprachzeitsignals auch direkt aus einer Multiplikation eines Verzerrungsfaktors mit den Amplituden- und Zeitwerten der quasi-periodischen Segmente des ersten Sprachzeitsignals ermitteln. Der Verzerrungsfaktor ergibt sich aus einer Transformation der linearen Verzerrungsfrequenzfunktion in den Zeitbereich und entspricht hier dem einen Parameter der Verzerrungsfrequenzfunktion. Die Ergebnisse der Multiplikation, die außerhalb eines darstellbaren Werte-/Definitionsbereichs des digitalen Sprachzeitsignals liegen, werden mit einer Spline-Interpolation ermittelt. Durch die lineare Verzerrungsfunktion mit einem Parameter kann das erste Sprachzeitsignal also im Amplituden- und Zeitbereich skaliert werden. Die geringe Rechenkomplexität gewährleistet eine hohe Reduktion des Speicherbedarfs und der Rechenzeit.

[0025] In einem dritten Schritt werden die neu ermittelten quasi-periodischen Segmente mit dem TD-PSOLA Algorithmus zusammengesetzt. Dieser basiert im Wesentlichen auf einem Overlap-and-Add Verfahren im Zeitbereich, wobei man beispielsweise die quasi-periodischen Segmente zeitversetzt zusammensetzt und die überlappenden Abtastwerte addiert.

Patentansprüche

1. Verfahren zur Stimmenkonversion für ein Sprachsynthesesystem bei dem
– ein erstes Sprachzeitsignal in zeitlich aufeinander folgende Segmente aufgeteilt wird,
– die Segmente mit einer Verzerrungszeitfunktion gefaltet werden,
– ein zweites Sprachzeitsignal durch Zusammensetzen der gefalteten Segmente erzeugt wird.

2. Verfahren nach Anspruch 1, wobei das erste und/oder das zweite Sprachzeitsignal als analoge Sprachzeitsignale ausgeführt sind.

3. Verfahren nach Anspruch 1, wobei das erste und/oder das zweite Sprachzeitsignal als digitale Sprachzeitsignale ausgeführt sind.

4. Verfahren nach mindestens einem der Ansprüche 1 bis 3, wobei das erste Sprachzeitsignal in quasiperiodische Segmente aufgeteilt wird.

5. Verfahren nach Anspruch 4, wobei die quasiperiodischen Segmente mit Hilfe eines Pitch-Tracking Algorithmus ermittelt werden.

6. Verfahren nach mindestens einem der Ansprüche 1 bis 5, wobei mindestens ein Parameter der Verzerrungszeitfunktion vorgebbar ist.

7. Verfahren nach mindestens einem der Ansprüche 1 bis 6, wobei die Verzerrungszeitfunktion durch Transformation einer Verzerrungsfrequenzfunktion in den Zeitbereich ermittelt wird.

8. Verfahren nach Anspruch 7, wobei mindestens ein Parameter der Verzerrungsfrequenzfunktion im Frequenzbereich durch Minimierung des euklidischen Abstandes zwischen einem Sprachfrequenzsignal eines Zielsprechers und einem mit dem mindestens einen Parameter der Verzerrungsfrequenzfunktion transformierten Sprachfrequenzsignal eines Quellsprechers ermittelt wird.

9. Verfahren nach Anspruch 7, wobei mindestens ein Parameter der Verzerrungsfrequenzfunktion vorgebbar ist.

10. Verfahren nach Anspruch 7, wobei die Verzerrungsfrequenzfunktion als lineare Verzerrungsfrequenzfunktion mit einem Parameter ausgeführt ist.

11. Verfahren nach Anspruch 7, wobei die Verzerrungsfrequenzfunktion als stückweise, lineare Verzerrungsfrequenzfunktion mit einem Parameter ausgeführt ist.

12. Verfahren nach Anspruch 7, wobei die Verzerrungsfrequenzfunktion als nichtlineare Verzerrungsfrequenzfunktion mit einem Parameter ausgeführt ist.

13. Verfahren nach Anspruch 7, wobei die Verzerrungsfrequenzfunktion als stückweise, lineare Verzerrungsfrequenzfunktion mit mehreren Parametern ausgeführt ist.

14. Verfahren nach Anspruch 7, wobei die Verzerrungsfrequenzfunktion als nichtlineare Verzerrungsfrequenzfunktion mit mehreren Parametern ausgeführt ist.

15. Verfahren nach mindestens einem der Ansprüche 1 bis 14, wobei durch die Faltung der Segmente mit der Verzerrungszeitfunktion entstandene Werte, welche außerhalb eines darstellbaren Wertebereichs liegen, interpoliert werden.

16. Verfahren nach Anspruch 15, wobei für die Interpolation eine Spline-Interpolation eingesetzt wird.

17. Verfahren nach Anspruch 15, wobei für die In-

terpolation eine lineare Interpolation eingesetzt wird.

18. Verfahren nach mindestens einem der Ansprüche 1 bis 17, wobei das Zusammensetzen der gefalteten Segmente mit einem TD-PSOLA Algorithmus ausgeführt wird.

19. Verfahren nach mindestens einem der Ansprüche 1 bis 18, wobei die Segmente des ersten Sprachzeitsignals durch Multiplikation mit einem vorgebbaren Faktor im Zeit- und/oder Amplitudenbereich gestaucht bzw. gestreckt werden.

Es folgt ein Blatt Zeichnungen

Anhängende Zeichnungen

FIG 1

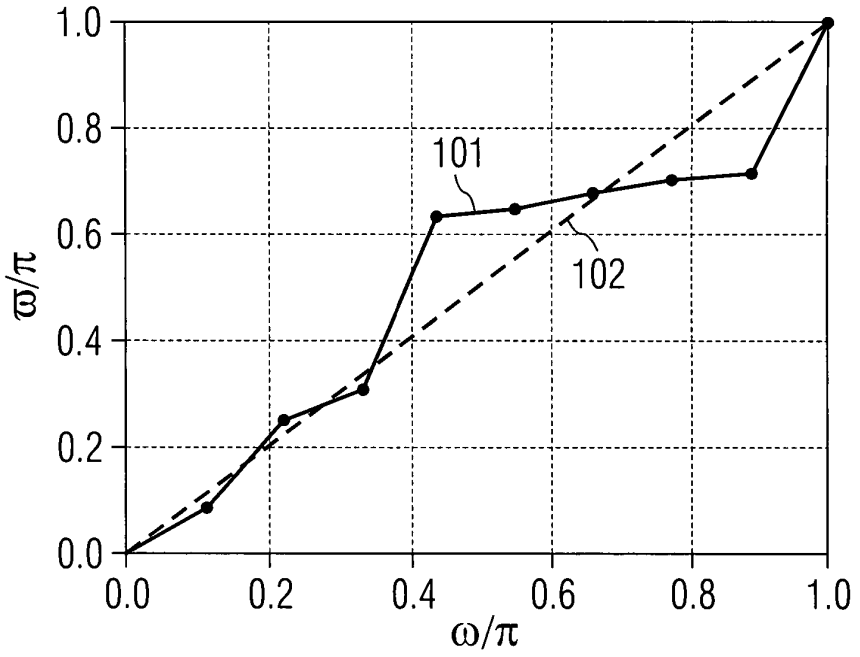


FIG 2