(19) **Bundesrepublik Deutschland**
**Deutsches Patent- und Markenamt**

(10) **DE 697 20 861 T2** 2004.02.05

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 0 813 184 B1**
(21) Deutsches Aktenzeichen: **697 20 861.3**
(96) Europäisches Aktenzeichen: **97 870 079.7**
(96) Europäischer Anmeldetag: **29.05.1997**
(97) Erstveröffentlichung durch das EPA: **17.12.1997**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **16.04.2003**
(47) Veröffentlichungstag im Patentblatt: **05.02.2004**

(51) Int Cl.[7]: **G10L 13/06**
**G10L 21/04**

(30) Unionspriorität:
**9600524** **10.06.1996** **BE**

(84) Benannte Vertragsstaaten:
**BE, CH, DE, DK, ES, FR, GB, IT, LI, NL, SE**

(73) Patentinhaber:
**Faculté Polytechnique de Mons, Mons, BE**

(74) Vertreter:
**Zeitler & Dickel Patentanwälte, 80539 München**

(72) Erfinder:
**Dutoit, Thierry, 7332 Sirault, BE; Pagel, Vincent, 54000 Nancy, FR; Pierret, Nicolas, 7030 Saint-Symphorien, BE**

(54) Bezeichnung: **Verfahren zur Tonsynthese**

## Beschreibung

[0001] Die hier beschriebene Erfindung betrifft ein Verfahren zur Synthese von Tonsignalen. Um die Beschreibung zu vereinfachen wird die Hauptaufmerksamkeit auf Sprachtöne gelegt, wobei man allerdings in Erinnerung behält, dass die Erfindung genauso gut auf das Gebiet der Musiksynthese angewandt werden kann.

### Hintergrund der Erfindung

[0002] Im Rahmen der sogenannten „verketteten" Synthesetechniken, die in steigendem Maße angewandt werden, erzeugt man eine synthetische Sprache aus einer Datenbank von Sprachsegmenten. Die Segmente können zum Beispiel Diphone sein, die von der Mitte des stationären Teiles eines Phons an beginnen (wobei das Phon die akustische Verwirklichung eines Phonems ist) und die in der Mitte des stationären Teiles des nächsten Phons enden. Französisch zum Beispiel ist aus 36 Phonemen zusammengesetzt, die annähernd 1240 Diphonen entsprechen (tatsächlich sind einige Kombinationen von Phonemen unmöglich). Andere Typen von Segmenten können verwendet werden, etwa Triphone, Polyphone, Halbsilben usw. Verkettete Synthesetechniken erzeugen irgendeine Folge von Phonemen durch eine Verkettung der geeigneten Segmente. Die Segmente selbst werden aus der Segmentierung eines Sprachkorpus gewonnen, das von einem menschlichen Sprecher gelesen wird.

[0003] Zwei Probleme müssen während des Verkettungsprozesses gelöst werden, um ein Sprachsignal zu erhalten, welches mit der menschlichen Sprache vergleichbar ist.

[0004] Das erste Problem entsteht aus den Disparitäten der phonemischen Zusammenhänge, aus welchen die Segmente extrahiert wurden, was im allgemeinen zu einer gewissen fehlenden Übereinstimmung der spektralen Hüllkurve an den beiden Enden der Segmente führt, die verkettet werden sollen. Als ein Ergebnis führt eine bloße Verkettung von Segmenten zu scharfen Übergängen zwischen den Einheiten und zu einer weniger flüssigen Sprache.

[0005] Das zweite Problem besteht darin, die Prosodie der synthetischen Sprache zu steuern, d. h. ihren Rhythmus (Phonem und Pausenlängen) und ihre Grundfrequenz (die Schwingungsfrequenz der Stimmbänder). Der Punkt ist der, dass die in dem Korpus aufgezeichneten Segmente ihre eigene Prosodie haben, die nicht notwendigerweise der Prosodie entspricht, die zum Zeitpunkt der Synthese auferlegt worden ist.

[0006] Folglich besteht ein Bedarf, ein Mittel zur Kontrolle der prosodischen Parameter zu finden und zur Erzeugung von weichen Übergängen zwischen den Segmenten, ohne dass die Natürlichkeit der Sprachsegmente beeinträchtigt wird.

[0007] Man unterscheidet zwei Familien von Verfahren, um solche Probleme zu lösen: diejenigen, welche ein Spektralmodell des Vokaltraktes implementieren, und diejenigen, welche die Segmentwellenformen direkt in dem Zeitbereich verändern.

[0008] In der ersten Kategorie von Verfahren werden Übergänge zwischen verketteten Segmenten ausgeglichen, indem man den Unterschied zwischen den spektralen Hüllkurven auf beiden Seiten des Verkettungspunktes berechnet und diesen Unterschied in dem spektralen Bereich auf beiden Segmenten verbreitet. Die Art und Weise, wie die erste Kategorie von Verfahren die Tonhöhe und die Dauer der Segmente steuert, hängt von dem besonderen Modell ab, welches für die Abschätzung der spektralen Hüllkurve verwendet wird. Alle diese Verfahren erfordern eine hohe Rechenleistung zum Zeitpunkt der Synthese, was diese Verfahren daran hindert, in Echtzeit auf Prozessoren mit einem geringen Preis implementiert zu werden.

[0009] Im Gegensatz dazu zielt die zweite Familie der Syntheseverfahren darauf ab, eine Veränderung der Verkettung und Prosodie direkt in dem Zeitbereich mit Hilfe einer sehr begrenzten Rechenleistung zu erzeugen. Alle diese Verfahren ziehen einen Vorteil aus dem sogenannten „Poissonschen Summentheorem", welches unter den Spezialisten der Signalverarbeitung gut bekannt ist und welches zeigt, dass es möglich ist, aus irgendeiner endlichen Wellenform mit einer gegebenen spektralen Hüllkurve eine unendliche Wellenform mit derselben spektralen Hüllkurve für eine willkürlich gewählte (und konstante) Tonhöhe zu bauen. Dieses Theorem kann auf die Veränderung der Grundfrequenz der Sprachsignale angewandt werden. Vorausgesetzt das Spektrum der elementaren Wellenformen liegt nahe genug an der spektralen Hüllkurve des Signals, das man zu verändern wünscht, dann kann die Tonhöhe auferlegt werden, indem man die Verschiebung zwischen den elementaren Wellenformen zu der gezielt angesteuerten Tonhöheperiode einstellt, und indem man die resultierenden überlappenden Wellenformen addiert. In dieser zweiten Familie unterscheiden sich die Syntheseverfahren hauptsächlich nach der Art und Weise, wie sie die elementaren Wellenformen aus den vorher aufgezeichneten Segmenten ableiten. Um eine synthetische Sprache von hoher Qualität zu erzeugen, müssen jedoch die überlappenden elementaren Wellenformen, welche die Verfahren verwenden, eine Dauer von mindestens dem Zweifachen der Grundfrequenz der ursprünglichen Segmente aufweisen. Zwei Klassen von Techniken in dieser zweiten Familie der Syntheseverfahren werden in dem was nun folgt beschrieben.

[0010] Die erste Klasse bezieht sich auf Verfahren, welche im dem was folgt als 'PSOLA' Verfahren bezeichnet werden (Pitch Synchronous Overlap Add = Synchrone Überlappungs-Addition der Tonhöhe) und gekennzeichnet sind durch die direkte Extraktion der Wellenformen aus den kontinuierlichen Tonsignalen. Die verwendeten Tonsignale sind entweder identisch

mit den originalen Signalen (die Segmente) oder sie werden nach einiger Transformation aus diesen originalen Signalen erhalten. Elementare Wellenformen werden aus den Tonsignalen extrahiert, indem man die Signale mit Gewichtungsfenstern von endlicher Dauer multipliziert, welche synchron mit der Grundfrequenz des originalen Signals angeordnet sind. Da die Größe der elementaren Wellenformen mindestens das Zweifache der ursprünglichen Periode ausmachen muss, und wenn man davon ausgeht, dass es eine Wellenform für jede Periode des originalen Signals gibt, dann werden dieselben Sprachmuster in mehreren aufeinanderfolgenden Wellenformen verwendet: die Gewichtungsfenster überlappen sich in den Tonsignalen.

[0011] Beispiele solcher PSOLA Verfahren sind jene, welche in den Dokumenten EP-0363233, US-5479564, EP-0706170 definiert worden sind. Ein spezifisches Beispiel ist auch das MBR-PSOLA Verfahren, wie es von T. Dutoit und H. Leich veröffentlicht worden ist, und zwar in Speech Communication, Elsevier Publisher, November 1993, Vol. 13, No. 3–4, 1993. Das in dem Dokument US-5479564 beschriebene Verfahren schlägt ein Hilfsmittel zur Veränderung der Frequenz eines Tonsignals mit einer konstanten Grundfrequenz durch eine Überlappungsaddition kurzfristiger Signale vor, die aus diesem Signal extrahiert worden sind. Die Länge der Gewichtungsfenster, die verwendet werden, um die kurzfristigen Signale zu gewinnen, ist annähernd gleich dem Zweifachen der Periode des Tonsignals und ihre Position innerhalb der Periode kann auf irgendeinen Wert eingestellt werden (vorausgesetzt, dass die Zeitverschiebung zwischen aufeinanderfolgenden Fenstern die gleiche ist wie die Periode des Tonsignals). Das Dokument US-5479564 beschreibt auch eine Vorrichtung zur Interpolation von Wellenformen zwischen Segmenten, die zu verketten sind, um so Diskontinuitäten auszugleichen. Dies wird durch eine Veränderung der Perioden entsprechend dem Ende des ersten Segmentes und dem Beginn des zweiten Segmentes in solch einer Weise erreicht, dass der Unterschied zwischen der letzten Periode des ersten Segmentes und der ersten Periode des zweiten Segmentes verbreitet wird.

[0012] Die zweite Klasse von Techniken, die in dem was nun folgt als die „analytische Techniken" bezeichnet wird, basiert auf einer Veränderung des Zeitbereiches von Wellenformen, die sich ihre Muster nicht teilen, sogar nicht einmal teilweise. Der Syntheseschritt verwendet noch eine Verschiebung und eine Überlappungs-Addition der gewichteten Wellenformen, welche die Information der spektralen Hüllkurve tragen. Diese Wellenformen werden nicht länger mit Hilfe von überlappenden Gewichtungsfenstern aus einem kontinuierlichen Sprachsignal extrahiert. Beispiele dieser Techniken sind jene, die sowohl in den Dokumenten US-5369730 und GB-2261350 definiert sind, als auch jene, die von T. Yazu, K. Yamada in „The speech synthesis system for an unlimited Japanese vocabulary", in den Proceedings IEEE ICASSP 1986, Tokyo, S. 2019–2022 beschrieben worden sind. Die Europäische Patentanmeldung EP-A-0527527 und die Internationale Patentanmeldung WO 90/03027 offenbaren zwei weitere Beispiele von PSOLA Techniken.

[0013] In all diesen „analytischen" Techniken sind elementare Wellenformen Impulsantworten des Vokaltraktes, welche aus gleichmäßig entfernt angeordneten Rahmen von Sprachsignalen ermittelt worden sind und welche über ein Spektralmodell erneut synthetisiert, d. h. resynthetisiert worden sind. Die vorliegende Erfindung fällt in diese Klasse von Verfahren.

[0014] Ein Vorteil der analytischen Verfahren gegenüber den PSOLA Verfahren besteht darin, dass die Wellenformen, die von denselben verwendet werden, sich aus einem wahren Spektralmodell des Vokaltraktes ergeben. Daher können sie intrinsisch die Information der augenblicklichen spektralen Hüllkurve mit einer größeren Genauigkeit und Präzision abbilden als die PSOLA Techniken, welche einfach ein Zeitbereichssignal mit einem Gewichtungsfenster gewichten. Darüber hinaus ist es mit analytischen Verfahren möglich, die periodischen (stimmhaft) und die aperiodischen (stimmlos) Komponenten einer jeden Wellenform zu trennen und deren Ausgleich während des Schrittes der Resynthese zu verändern, um die Sprachqualität (weich, rau, flüsternd usw.) zu verändern.

[0015] In der Praxis wird dieser Vorteil durch einen Anstieg der Größe der resynthetisierten Segmentdatenbank (typischerweise ein Faktor 2, da die aufeinanderfolgenden Wellenformen sich keine Muster teilen, während deren Dauer die gleiche sein muss wie noch mindestens zweimal diejenige der Tonhöheperiode des Tonsignals) ausgeglichen. Das von Yazu und Yamada beschriebene Verfahren zielt genau auf eine Verminderung der Anzahl der Muster ab, die gespeichert werden müssen, durch ein Resynthetisieren von Impulsantworten, in denen die Phasen der spektralen Hüllkurve gleich Null gesetzt werden. Nur die Hälfte der Wellenform braucht in diesem Fall gespeichert zu werden, da eine Nullsetzung der Phase zu vollständig symmetrischen Wellenformen führt. Der Hauptnachteil dieses Verfahrens besteht darin, dass es in einem großen Maße die Natürlichkeit der synthetischen Sprache beeinträchtigt. Es ist in der Tat gut bekannt, dass die Vornahme von bedeutenden Phasenverzerrungen eine starke Auswirkung auf die Qualität der Sprache hat.

## Ziel der Erfindung

[0016] Die vorliegende Erfindung zielt darauf ab, ein Verfahren für die Tonsynthese vorzuschlagen, welches die im Zusammenhang mit dem Stand der Technik dargestellten Nachteile vermeidet und welches einen begrenzten Speicher für die Wellenformen erfordert, während es dabei bedeutende Verzerrungen der natürlichen Phase der akustischen Signale ver-

meidet.

## Wesentliche kennzeichnende Elemente der Erfindung

[0017] Die vorliegende Erfindung betrifft ein Verfahren zur Tonsynthese von Wellenformen, die in einem Wörterbuch gespeichert sind, wobei die Wellenformen erzielt werden durch eine Spektralanalyse eines Wörterbuches von Tonsegmenten, und wobei das Verfahren die folgenden Schritte aufweist:

– die Wellenformen sind unendlich und vollkommen periodisch, und sie sind gespeichert als eine Periode derselben, welche selbst dargestellt ist als eine Sequenz von Tonmustern von a priori irgendeiner Länge;

– eine Synthese wird durchgeführt durch Überlappen und Addieren der Wellenformen, multipliziert durch ein Gewichtungsfenster, dessen Länge ungefähr zweimal die Periode der originalen Wellenform ausmacht, und dessen Position in Bezug auf die Wellenform auf irgendeinen festen Wert eingestellt werden kann;

– die aufeinanderfolgenden Wellenformen teilen keine Muster;

– wodurch die Zeitverschiebung zwischen zwei aufeinanderfolgenden gewichteten Signalen, die durch Gewichtung der originalen Wellenformen erzielt worden sind, die gleiche ist wie die fundamentale Periode, die für das synthetische Signal erfordert ist, dessen Wert auferlegt ist. Dieser Wert kann kleiner oder größer sein als derjenige der originalen Wellenformen.

[0018] Das Verfahren gemäß der vorliegenden Erfindung unterscheidet sich grundlegend von irgendeinem anderen „analytischen" Verfahren durch die Tatsache, dass die elementaren Wellenformen, die verwendet werden, keine Impulsantworten des Vokaltraktes sind, sondern unendliche, periodische Signale, multipliziert durch ein Gewichtungsfenster, um ihre Länge endlich zu halten, und dass sie dieselbe spektrale Hüllkurve tragen wie die originalen Tonsignale. Ein Spektralmodell (hybrid harmonisches/stochastisches Modell zum Beispiel, obwohl die Erfindung nicht ausschließlich irgendein besonderes Spektralmodell betrifft) wird für die Resynthese verwendet, um periodische Wellenformen zu erhalten (anstelle der symmetrischen Impulsantworten von Yazu und Yamada), welche die Information der augenblicklichen spektralen Hüllkurve tragen. Wegen der Periodizität der erzeugten elementaren Wellenformen braucht nur die erste Periode gespeichert zu werden. Die durch dieses Verfahren gewonnene Tonqualität ist dem Verfahren von Yazu und Yamada unvergleichlich überlegen, da die Berechnung der periodischen Wellenformen den spektralen Hüllkurven keine Phasenbeschränkungen auferlegt, wodurch man die damit zusammenhängende qualitative Entwertung vermeidet.

[0019] Die Perioden, die gespeichert werden müssen, werden durch eine Spektralanalyse eines Wörterbuches von Tonsegmenten gewonnen (z. B. Diphone in dem Fall einer Sprachsynthese). Die Spektralanalyse erzeugt Abschätzungen der spektralen Hüllkurve über jedes Segment hinweg. Harmonische Phasen und Amplituden werden dann aus der spektralen Hüllkurve berechnet und aus der Zielperiode (d. h. die spektrale Hüllkurve wird mit der angesteuerten Grundfrequenz abgetastet).

[0020] Die Länge einer jeden resynthetisierten Periode kann in einer vorteilhaften Weise für alle Perioden von allen Segmenten gleich gewählt werden. In diesem besonderen Fall erlauben klassische Techniken der Wellenformkompression (z. B. ADPCM) sehr hohe Kompressionsverhältnisse (etwa 8) mit sehr begrenzten Berechnungskosten für die Decodierung. Die bemerkenswerte Wirksamkeit solcher Techniken auf die gewonnenen Wellenformen rührt hauptsächlich her von der Tatsache, dass:

– alle Perioden, die in der Segmentdatenbank gespeichert sind, dieselbe Länge haben, was zu einem sehr wirksamen unterscheidenden Codierungsschema von Periode zu Periode führt;

– die Verwendung eines Spektralmodels für die Abschätzung der spektralen Hüllkurve die Trennung der harmonischen und der stochastischen Komponenten der Wellenformen erlaubt. Wenn die Energie der stochastischen Komponente klein genug ist, verglichen mit derjenigen der harmonischen Komponente, dann kann sie vollständig weggelassen werden, in welchem Fall nur die harmonische Komponente resynthetisiert wird. Dies führt zu Wellenformen, die ausgeprägter rein und rauscharm sind und die eine höhere Regelmäßigkeit zeigen als das originale Signal, was zusätzlich die Wirksamkeit der ADPCM Techniken der Codierung heraufsetzt.

[0021] Um die Wirksamkeit der Codierungstechniken weiter zu vergrößern, können die Phasen der Harmonischen unterer Ordnung (d. h. niedrigerer Frequenz) einer jeden gespeicherten Periode fest sein (ein Phasenwert fest eingestellt für jede Harmonische der Datenbank) für den Schritt der Resynthese. Das Frequenzband, bei dem diese Einstellung annehmbar ist, reicht von 0 bis annähernd 3 kHz. In diesem Fall führt der Arbeitsschritt der Resynthese zu einer Folge von Perioden mit einer konstanten Länge, in welcher der Zeitbereichunterschied zwischen zwei aufeinanderfolgenden Perioden hauptsächlich auf Unterschiede der spektralen Hüllkurve zurückzuführen ist. Da die spektrale Hüllkurve von Tonsignalen sich im allgemeinen langsam mit der Zeit verändert, wenn man von der gegebenen Trägheit des physikalischen Mechanismus ausgeht, der sie erzeugt, dann wird sich die Gestalt der auf diesem Wege gewonnenen Perioden auch langsam ändern. Dies wiederum ist besonders wirksam, wenn es zu Codierungssignalen auf der Grundlage von Unter-

schieden von Periode zu Periode kommt.

[0022] Unabhängig von der Verwendung für die Segmentcodierung führt die Idee, einen Satz von festen Werten für die Phasen der Harmonischen der niedrigeren Frequenzen aufzuerlegen, zu der Implementation einer zeitlichen Glättungstechnik zwischen aufeinanderfolgenden Segmenten, um die fehlende spektrale Übereinstimmung zwischen den Perioden abzuschwächen. Der zeitliche Unterschied zwischen der letzten Periode des ersten Segmentes und der ersten Periode des zweiten Segmentes wird berechnet und wird ausgleichend auf beiden Seiten des Verkettungspunktes verbreitet mit einem Gewichtungskoeffizienten, der ständig zwischen –0,5 und +0,5 variiert (abhängig davon, auf welcher Seite des Verkettungspunktes verarbeitet wird).

[0023] Es sollte angemerkt werden, dass, obwohl die oben erwähnten wirksamen Eigenschaften zur Codierung und Fähigkeiten zum Ausgleich bereits in der MBR-PSOLA Technik verfügbar waren, wie in dem Stand der Technik beschrieben, ihre Wirkung in der vorliegenden Erfindung drastisch verstärkt wird, weil im Gegensatz zu den Wellenformen, die von der MBR-PSOLA Technik verwendet werden, die hier verwendeten Perioden keine ihrer Muster teilen, wodurch eine vollkommene Trennung zwischen harmonisch gereinigten Wellenformen und Wellenformen, die im wesentlichen stochastisch sind, erlaubt ist.

[0024] Schließlich macht es die vorliegende Erfindung noch möglich, die Qualität der synthetisierten Tonsignale zu erhöhen, indem man mit einem jeden resynthetisierten Segment („Basissegment") einen Satz von Ersatzsegmenten, ähnlich aber nicht identisch zu dem Basissegment, verbindet. Jedes Basissegment wird in derselben Weise verarbeitet wie das entsprechende Basissegment und eine Folge von Perioden wird resynthetisiert. Für jedes Ersatzsegment zum Beispiel kann man zwei Perioden halten entsprechend jeweils zu dem Beginn und dem Ende des Ersatzsegmentes zum Zeitpunkt der Synthese. Wenn zwei Segmente dabei sind, verkettet zu werden, dann ist es möglich, die Perioden des ersten Basissegmentes so zu verändern, um an den letzten Perioden dieses Segmentes den Unterschied zwischen der letzten Periode des Basissegmentes und der letzten Periode von einer ihrer Ersatzsegmente zu verbreiten. Ähnlich ist es möglich, die Perioden des zweiten Basissegmentes so zu verändern, um an den ersten Perioden dieses Segmentes den Unterschied zwischen der ersten Periode des Basissegmentes und der ersten Periode von einer ihrer Ersatzsegmente zu verbreiten. Die Verbreitung dieser Unterschiede wird einfach durch Multiplizieren der Unterschiede durch einen ständig von 1 bis 0 (von Periode zu Periode) variierenden Gewichtungskoeffizienten und Addieren der gewichteten Unterschiede zu den Perioden der Basissegmente durchgeführt.

[0025] Solch eine Veränderung der Perioden des Zeitbereiches eines Basissegmentes, um es ertönen zu lassen, wie eines seiner Ersatzsegmente, kann vorteilhaft eingesetzt werden, um freie Varianten zu einem Grundton zu erzeugen, wodurch die Eintönigkeit vermieden wird, welche aus dem wiederholten Gebrauch eines Grundtones entsteht. Es kann auch für die Erzeugung von linguistisch motivierten Tonvarianten (z. B. betonte /unbetonte Vokale, angespannte/ weiche Stimme, usw.) verwendet werden.

[0026] Der grundlegende Unterschied zwischen dem in dem Stand der Technik beschriebenen Verfahren, welches gemäß unserer Klassifizierung ein „PSOLA" Verfahren ist, und dem Verfahren gemäß der vorliegende Erfindung hat seinen Ursprung in der besonderen Art und Weise, wie die verwendeten Perioden abgleitet werden. Im Gegensatz zu den Wellenformen, die von einem kontinuierlichen Signal extrahiert werden, wie es in dem Stand der Technik vorgeschlagen wird, teilen die in der vorliegenden Erfindung verwendeten Wellenformen keine ihrer Muster (daher überlappen sie nicht). Das Verfahren profitiert daher von den typischen Vorteilen anderer analytischer Verfahren:

– sehr effiziente Codierungstechniken, welche die Tatsache mit berücksichtigen, dass:
– Perioden harmonisch rein sein können durch eine vollständige Beseitigung ihrer stochastischen Komponente;
– wenn die Perioden resynthetisiert werden, die Phase der Harmonischen der niedrigeren Frequenz konstant gesetzt werden kann (d. h. ein fester Wert für jede Harmonische durch die ganze Segmentdatenbank hindurch)
– Fähigkeit, Tonvarianten durch Interpolation zwischen Basis- und Ersatzsegmenten zu erzeugen. Für jedes Basissegment zum Beispiel werden zwei zusätzliche Perioden gespeichert, entsprechend dem Beginn und dem Ende des Segmentes und genommen aus einem Ersatzsegment. Dies ermöglicht die Synthese von Stimmen, die natürlicher klingen.

## Kurze Beschreibung der Zeichnungen

[0027] Das Verfahren gemäß der vorliegenden Erfindung soll präziser beschrieben werden, indem es mit den folgenden Verfahren nach dem Stand der Technik verglichen wird:

[0028] **Fig.** 1 stellt die verschiedenen Schritte der Sprachsynthese gemäß einem PSOLA Verfahren dar,

[0029] **Fig.** 2 beschreibt die verschiedenen Schritte der Sprachsynthese gemäß dem Verfahren, das von Yazu und Yamada vorgeschlagen worden ist,

[0030] **Fig.** 3 beschreibt die verschiedenen Schritte der Sprachsynthese gemäß der vorliegenden Erfindung.

## Beschreibung einer bevorzugten Ausführungsform der Erfindung

[0031] **Fig.** 1 zeigt eine klassische Darstellung ei-

nes PSOLA Verfahrens, das durch die folgenden Schritte gekennzeichnet ist.

1. Mindestens bei den stimmhaften Teilen von Sprachsegmenten wird eine Analyse durch Gewichtung der Sprache mit einem Fenster durchgeführt, das annähernd zentriert auf den Beginn einer jeden Impulsantwort des Vokaltraktes ist, der durch die Stimmbänder angeregt worden ist. Das Gewichtungsfenster hat eine Form, welche an ihren Rändern bis auf Null herunter abnimmt, und es hat eine Länge, die mindestens annähernd zweimal so groß ist wie die Grundperiode der originalen Sprache oder zweimal so groß wie die Grundperiode der Sprache, die synthetisiert werden soll.
2. Die Signale, welche aus dem Arbeitsschritt der Gewichtung resultieren, werden gegenseitig in Bezug aufeinander verschoben, dabei wird die Verschiebung an die Grundperiode der Sprache angepasst, die synthetisiert werden soll, kleiner oder größer als die originale Grundperiode, entsprechend der prosodischen Information, die sich auf die Grundperiode zu dem Zeitpunkt der Synthese bezieht.
3. Die synthetische Sprache wird durch ein Summieren dieser verschobenen Signale erzielt.

[0032] **Fig.** 2 zeigt das von Yazu und Yamada beschriebene Verfahren gemäß dem Stand der Technik, welches 3 Schritte umfasst:

1. Die originale Sprache wird zu jeder festen Rahmenperiode (daher nicht synchron zur Tonhöhe) herausgeschnitten und das Spektrum eines jeden Rahmens wird durch eine Cepstralanalyse berechnet. Phasenkomponenten werden auf Null gesetzt, so dass nur spektrale Amplituden zurückgehalten werden. Eine symmetrische Wellenform wird dann für jeden anfänglichen Rahmen durch eine inverse FFT erzielt. Diese symmetrische Wellenform wird mit einem Fenster fester Länge gewichtet, welches an seinen Begrenzungen fast auf den Wert Null abnimmt.
2. Die Signale, die aus dem Arbeitsschritt der Gewichtung resultieren, werden gegenseitig in Bezug aufeinander verschoben, dabei wird die Verschiebung an die Grundperiode der Sprache angepasst, die synthetisiert werden soll, kleiner oder größer als die originale Grundperiode, entsprechend der prosodischen Information, die sich auf die Grundperiode zu dem Zeitpunkt der Synthese bezieht.
3. Die synthetische Sprache wird durch ein Summieren dieser verschobenen Signale erzielt.

[0033] Bei dieser letzten Technik werden die Schritte 1 und 2 oft ein für alle Mal verwirklicht, was den Unterschied ausmacht zwischen den analytischen Verfahren und solchen, die auf einem spektralen Modell des Vokaltraktes beruhen. Die verarbeiteten Wellenformen werden in einer Datenbank gespeichert, welche in einem rein zeitlichen Format alle die Informationen zentralisiert, welche mit der Veränderung der spektralen Hüllkurve der Sprachsegmente zusammenhängen.

[0034] Was die bevorzugte Implementation der hier beschriebenen Erfindung anbetrifft, so beschreibt **Fig.** 3 die folgenden Schritte:

1. Den Analyserahmen wird eine feste Länge und Verschiebung (durch S bezeichnet) zugewiesen. Anstelle einer Abschätzung der spektralen Hüllkurve eines jeden Analyserahmens durch eine Cepstralanalyse und eine Berechnung ihre inversen FFT (wie von Yazu und Yamada durchgeführt) wird der Analysealgorithmus des mächtigen MBE (Multi-Banderregung) Modells angewandt, welches die Frequenz, die Amplitude und die Phase einer jeden Harmonischen des Analyserahmens berechnet. Die spektrale Hüllkurve wird dann für einen jeden Rahmen abgeleitet und eine Abänderung der Frequenzen und Amplituden der Harmonischen findet statt, ohne dass sich diese Hüllkurve ändert, um so eine feste Grundfrequenz zu erzielen, die gleich groß ist wie die Analyseverschiebung S (d. h. das Spektrum wird in dem Frequenzbereich „re-harmonisiert"). Phasen der niedrigeren Harmonischen werden auf einen Satz von festen Werten eingestellt (d. h. ein Wert, der ein für alle Mal für eine gegebene Ordnungszahl der Harmonischen gewählt wird). Die Wellenformen des Zeitbereichs werden dann aus den Harmonischen durch Berechnung einer Summe von Sinuskurven gewonnen, die Frequenzen, Amplituden und Phasen werden denen der Harmonischen gleichgesetzt. Im Gegensatz zu der Erfindung von Yazu und Yamada sind die Wellenformen nicht symmetrisch, da die Phasen nicht auf Null gesetzt worden sind (es gab keine andere Wahl bei dem vorhergehenden Verfahren). Weiterhin werden die erzielten präzisen Wellenformen nicht durch den Algorithmus auferlegt, weil sie streng von den festen Phasenwerten abhängen, die vor der Resynthese auferlegt werden. Anstatt die vollständige Wellenform in einer Segmentdatenbank zu speichern, wird nur eine Periode der Wellenform festgehalten, da sie durch den Aufbau (Summe der Harmonischen) vollkommen periodisch ist. Diese Periode kann auseinander gefaltet werden, um die entsprechende unendliche Wellenform zu erzielen, wie sie für den nächsten Schritt erfordert ist.
2. Bei den stimmhaften Teilen von Sprachsegmenten wird eine Analyse durch Gewichtung der zuvor erwähnten resynthetisierten Wellenform (gewonnen aus dem Durchlaufen einer ihrer Perioden, die als eine Summe von Harmonischen berechnet worden ist) mit einem Fenster mit einer festen Länge durchgeführt. Das Gewichtungsfenster hat eine Form, welche an ihren Rändern bis auf Null herunter abnimmt, und seine Länge ist genau zweimal so groß wie der Wert von S, und

daher auch zweimal so groß wie die Grundperiode der resynthetisierten Sprache, die in Schritt 1 erzielt worden ist. Eines von solchen Fenstern wird aus einer jeden der in dem Schritt 1 abgeleiteten unendlichen Wellenform genommen.
3. Die Signale, die aus dem Arbeitsschritt der Gewichtung resultieren, werden überlappt und gegenseitig in Bezug aufeinander verschoben, dabei wird die Verschiebung an die Grundperiode der Sprache angepasst, die synthetisiert werden soll, kleiner oder größer als S, entsprechend der prosodischen Information, die sich auf die Grundperiode zu dem Zeitpunkt der Synthese bezieht. Eine synthetische Sprache wird durch das Summieren dieser verschobenen Signale erzielt.

[0035] Die Erfindung macht es möglich, in dem Zeitbereich spektrale Diskontinuitäten auf Grund des festen Satzes an Phasen auszugleichen, die auf die Perioden während des Schrittes der Resynthese für die Harmonischen unterer Ordnung angewandt werden, da eine Interpolation zwischen zwei solchen Perioden in dem Zeitbereich dann gleichwertig zu einer Interpolation in dem Frequenzbereich ist.

**Patentansprüche**

1. Verfahren zur Tonsynthese von Wellenformen, die in einem Wörterbuch gespeichert sind, wobei die Wellenformen durch eine Spektralanalyse eines Wörterbuches von Tonsegmenten erzielt werden, wobei das Verfahren die folgenden Schritte aufweist:
– die Wellenformen sind unendlich und vollkommen periodisch, und sie sind gespeichert als eine Periode derselben, welche selbst dargestellt ist als eine Sequenz von Tonmustern von a priori irgendeiner Länge;
– eine Synthese wird durchgeführt durch Überlappen und Addieren der Wellenformen, multipliziert durch ein Gewichtungsfenster, dessen Länge ungefähr zweimal die Periode der originalen Wellenform ausmacht, und dessen Position in Bezug auf die Wellenform auf irgendeinen festen Wert eingestellt werden kann;
– die aufeinanderfolgenden Wellenformen teilen keine Muster;
– wodurch die Zeitverschiebung zwischen zwei aufeinanderfolgenden gewichteten Signalen, die durch Gewichtung der originalen Wellenformen erzielt worden sind, die gleiche ist wie die fundamentale Periode, die für das synthetische Signal erfordert ist, dessen Wert auferlegt ist.

2. Verfahren zur Tonsynthese gemäss Anspruch 1, **dadurch gekennzeichnet**, dass die fundamentale Periode des synthetischen Signals größer oder kleiner ist als die originale Periode in dem Wörterbuch.

3. Verfahren zur Tonsynthese gemäss Anspruch 1 oder 2, dadurch gekennzeichnet, dass die Längen der Perioden, die in dem Wörterbuch gespeichert sind, alle identisch sind.

4. Verfahren zur Tonsynthese gemäss Anspruch 3, dadurch gekennzeichnet, dass die Phasen der Harmonischen der niedrigeren Frequenz (typischerweise von 0 bis 3 kHz) der gespeicherten periodischen Wellenformen einen festen Wert pro Harmonische durch das ganze Wörterbuch hindurch aufweisen.

5. Verfahren zur Tonsynthese gemäss irgendeinem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass die gespeicherten Wellenformen durch die Spektralanalyse eines Wörterbuches von Segmenten von Tonsignalen erzielt werden, wie etwa von Diphonen im Fall der Sprachsynthese, wodurch eine Spektralanalyse in regelmäßigen Zeitintervallen eine Abschätzung der augenblicklichen spektralen Hüllkurve in jedem Segment, von der die Wellenformen berechnet worden sind, liefert.

6. Verfahren zur Tonsynthese gemäss Anspruch 5, dadurch gekennzeichnet, dass wenn zwei Segmente verkettet werden, die letzten Perioden des ersten Segmentes und die erste Periode des zweiten Segmentes modifiziert werden, um den Zeitbereichunterschied auszugleichen, der zwischen der letzten Periode des ersten Segmentes und der ersten Periode des zweiten Segmentes gemessen wird, wobei dieser Zeitbereichunterschied jeder modifizierten Periode hinzugefügt wird, und das mit einem Gewichtungskoeffizienten, der zwischen –0,5 und 0,5 variiert, je nach der Position der modifizierten Periode in Bezug auf den Verkettungspunkt.

7. Verfahren zur Tonsynthese gemäss Anspruch 6, dadurch gekennzeichnet, dass für jedes Basissegment Ersatzsegmente gespeichert werden, wodurch zum Zeitpunkt der Synthese, wenn zwei Segmente dabei sind, verkettet zu werden, die Perioden des ersten Basissegmentes so modifiziert werden, um an den letzten Perioden dieses Segmentes den Unterschied zwischen der letzten Periode des Basissegmentes und der letzten Periode von einer seiner Ersatzsegmente zu verbreiten, und wodurch die Perioden des zweiten Basissegmentes so modifiziert werden, um an den ersten Perioden dieses Segmentes den Unterschied zwischen der ersten Periode des Basissegmentes und der ersten Periode von einer seiner Ersatzsegmente zu verbreiten, wobei die Verbreitung dieser Unterschiede durch Multiplizieren der gemessenen Unterschiede mit einem ständig von 1 bis 0 (von Periode zu Periode) variierenden Gewichtungskoeffizienten und Addieren der gewichteten Unterschiede zu den Perioden der Basissegmente durchgeführt wird.

Es folgen 3 Blatt Zeichnungen

**Anhängende Zeichnungen**

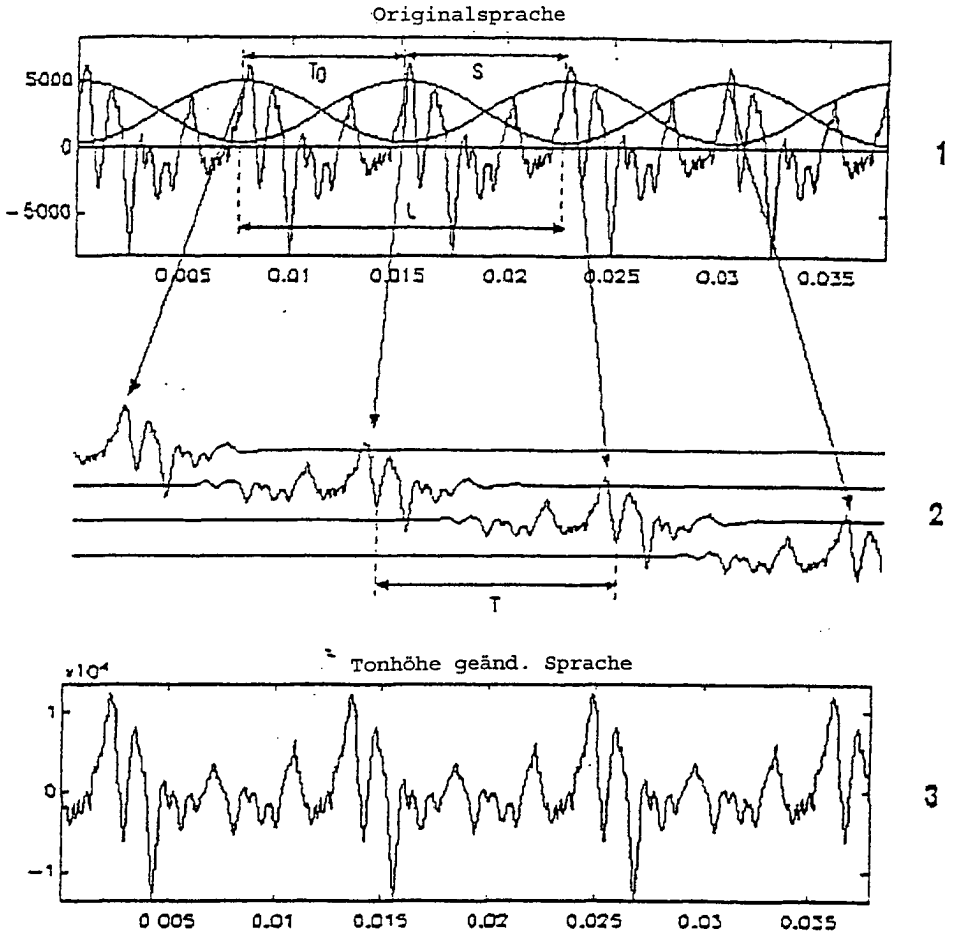


Fig. 1

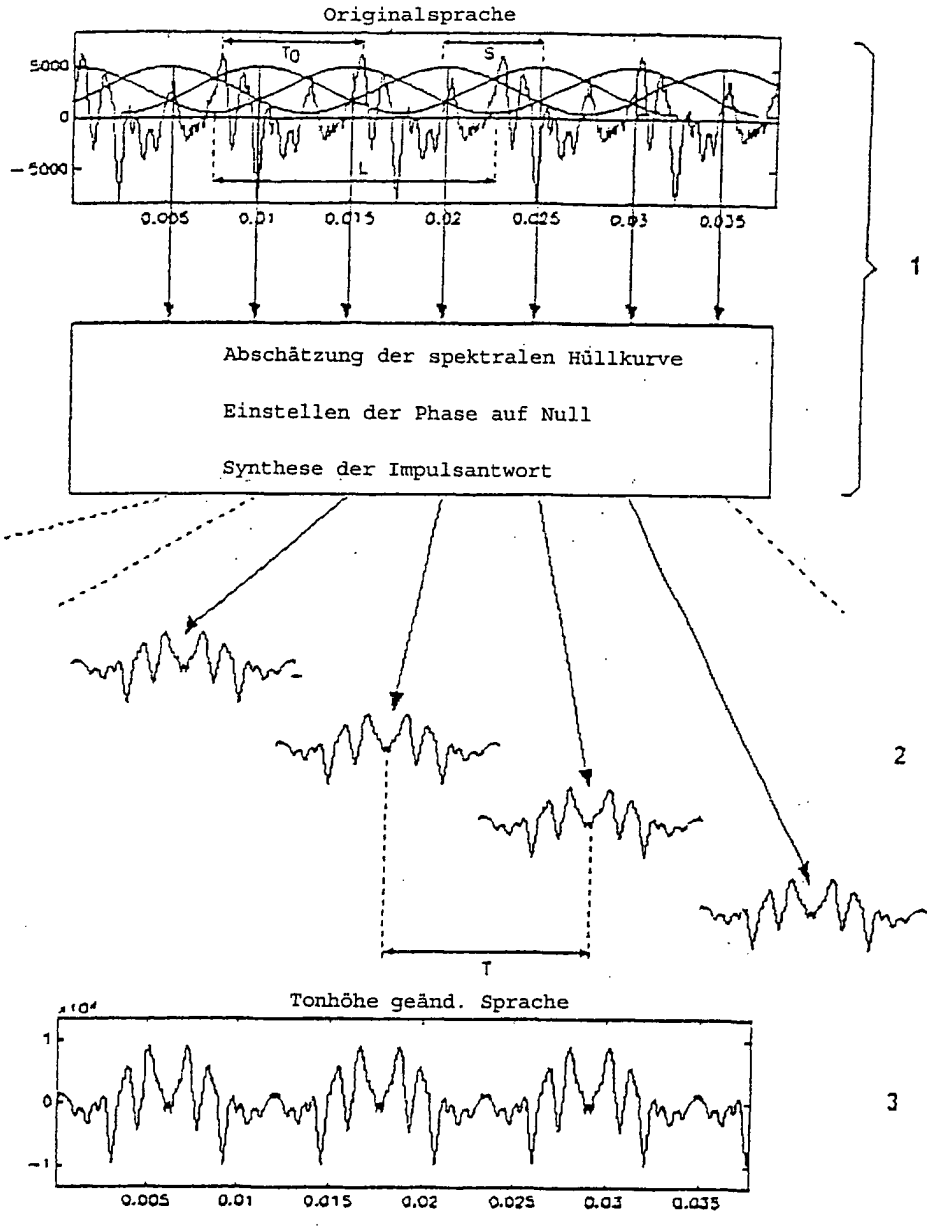
Originalsprache
5000
0
−5000
T0
S
L
0.005
0.01
0.015
0.02
0.025
0.03
0.035
1
Abschätzung der spektralen Hüllkurve
Einstellen der Phase auf Null
Synthese der Impulsantwort
2
T
Tonhöhe geänd. Sprache
x 10^4
1
0
−1
0.005
0.01
0.015
0.02
0.025
0.03
0.035
3

Fig. 2

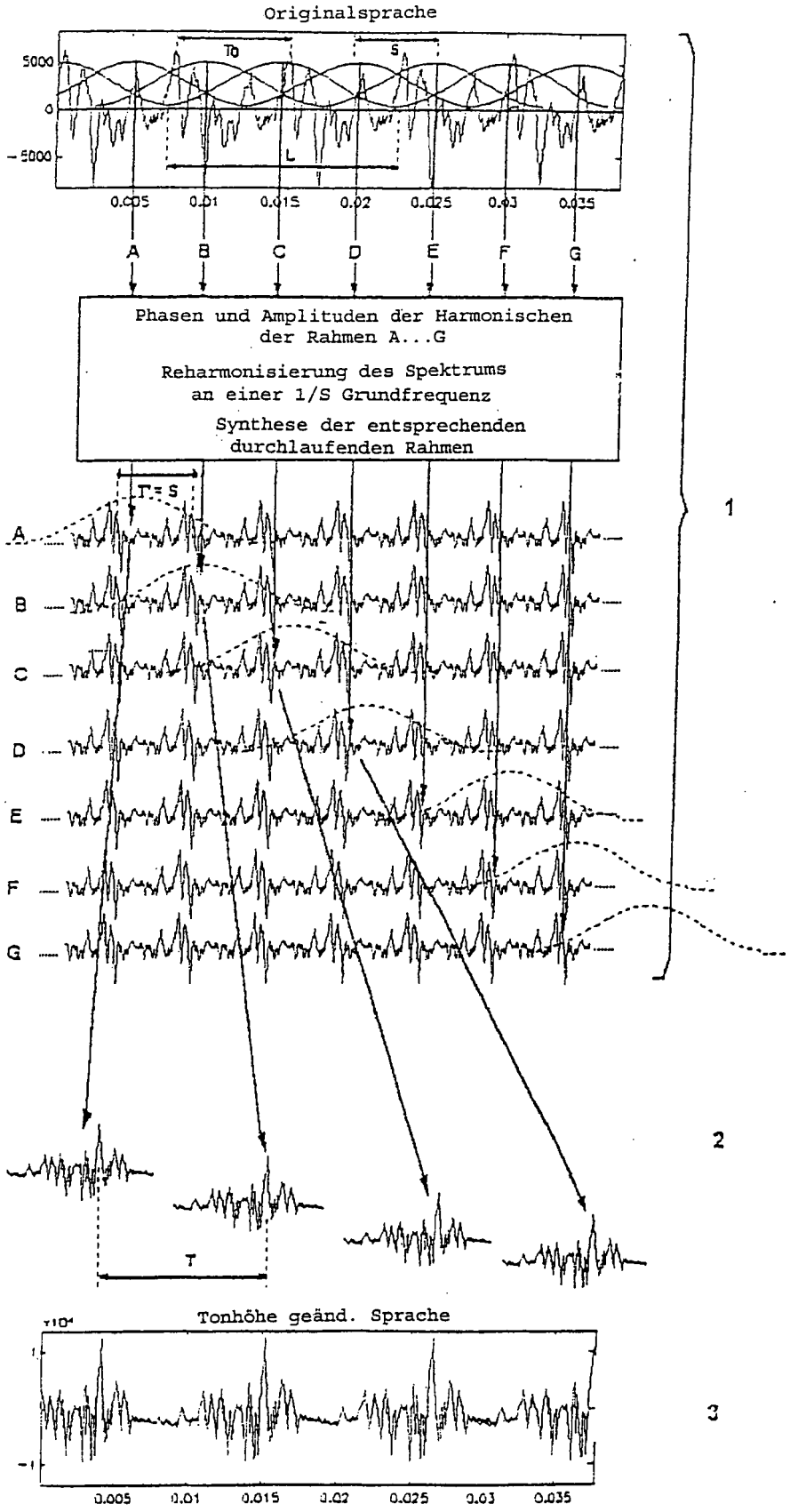
Originalsprache
5000
0
-5000
To
S
L
0.005
0.01
0.015
0.02
0.025
0.03
0.035
A
B
C
D
E
F
G
Phasen und Amplituden der Harmonischen der Rahmen A...G
Reharmonisierung des Spektrums an einer 1/S Grundfrequenz
Synthese der entsprechenden durchlaufenden Rahmen
T = S
1
2
T
Tonhöhe geänd. Sprache
3

Fig. 3