(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 3 144 929 A1**

(12) EUROPÄISCHE PATENTANMELDUNG

(43) Veröffentlichungstag:
22.03.2017 Patentblatt 2017/12

(51) Int Cl.:
G10L 13/033 *(2013.01)* ***G10L 13/047*** *(2013.01)*

(21) Anmeldenummer: **15185879.2**

(22) Anmeldetag: **18.09.2015**

(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
Benannte Erstreckungsstaaten:
BA ME
Benannte Validierungsstaaten:
MA

(71) Anmelder: **Deutsche Telekom AG**
53113 Bonn (DE)

(72) Erfinder: **BURKHARDT, Felix**
10551 Berlin (DE)

(74) Vertreter: **Brandt & Nern Patentanwälte**
Kekuléstrasse 2-4
12489 Berlin (DE)

(54) SYNTHETISCHE ERZEUGUNG EINES NATÜRLICH KLINGENDEN SPRACHSIGNALS

(57) Die Erfindung bezieht sich auf eine Lösung Sprachsynthese, nämlich auf die Erzeugung eines synthetischen Sprachsignals in einem automatisierten Ablauf. Zur Erzeugung eines möglichst natürlich klingenden synthetischen Sprachsignals wird vorgeschlagen, dass ein während der Sprachsynthese erzeugtes, noch nicht emotionsbehaftetes Sprachrohsignal mit einem Parametergemisch moduliert wird, welches Parameter mehrerer, mit Melodiemerkmalen, mit Dauermerkmalen, mit Stimmmerkmalen oder mit der Artikulationsgenauigkeit der Sprache korrespondierender Merkmalsgruppen umfasst, die entsprechend mindestens zwei vorgegebenen, mit voneinander verschiedenen der vorgenannten Merkmalsgruppen assoziierten Zielemotionen eingestellt werden. Das dazu vorgeschlagene System (1) besteht insbesondere aus einer Eingangsstufe (2) mit einer Phonemisierungskomponente (3), aus einem Emotionssimulator (4) und aus einer Ausgangsstufe (5) mit einer Syntheseeinheit (6). Die Eingangsstufe (2) ist zur Entgegennahme von Informationen über mindestens zwei Zielemotionen ausgebildet. Der Emotionssimulator (4) ist ausgebildet zur Auswertung dieser Informationen, zur Einstellung der Parametereigenschaften mindestens zweier verschiedener Sprachmerkmalsgruppen entsprechend den Zielemotionen, zum Mischen der eingestellten Parameter und zur Modulation des Sprachrohsignals mit dem Parametergemisch.

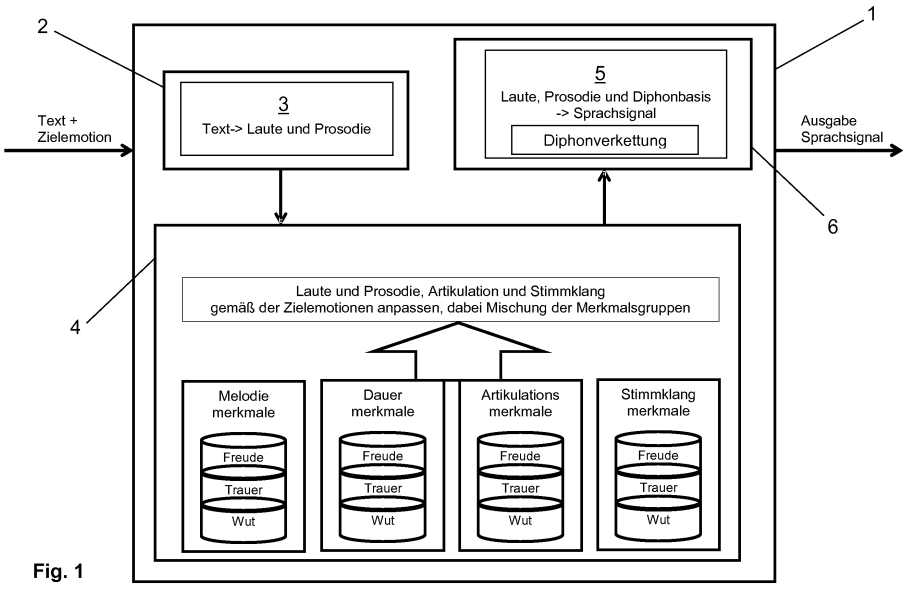

Fig. 1

EP 3 144 929 A1

Beschreibung

[0001] Die Erfindung betrifft eine Lösung zur automatisierten Sprachsynthese, welche es ermöglicht, Sprachsequenzen zu erzeugen, die aufgrund der Simulation einer emotionalen Sprechweise für einen durch diese Sprachsequenzen angesprochen Zuhörer besonders natürlich klingen. Sie bezieht sich auf eine Lösung, welche durch die Einbeziehung extralinguistischer und/oder paralinguistischer Merkmale deutlich über die rein linguistische Erzeugung synthetischer Sprache hinausgeht. Gegenstände der Erfindung sind ein entsprechendes Verfahren und ein zur Durchführung dieses Verfahrens geeignetes System.

[0002] Insbesondere zur Verbesserung der Mensch-Maschine-Kommunikation, aber auch der von Mensch-zu-Mensch unter Zwischenschaltung einer Maschine erfolgenden Kommunikation ist es bereits seit längerem bekannt, mittels technischer Einrichtungen, welche auch als Sprachsynthetisierer bezeichnet werden, Sprache künstlich zu erzeugen, also zu synthetisieren. Aus technischer Sicht geht es bei der Sprachsynthese um die künstliche Generierung von Sprachsignalen aus beliebigem Text mittels entsprechender computerbasierter Systeme. Die bekannten Syntheseverfahren lassen sich zumeist einem der vier, nachfolgend in historisch aufsteigender Reihenfolge genannten Hauptverfahren zuordnen:

- Formant Synthese: Bei diesem Verfahren wird ein Sprachsignal anhand physikalischer Modelle berechnet. Die Resonanz-Frequenzen im Sprechtrakt bezeichnet man hierbei als Formanten. Das Verfahren ist sehr flexibel und hat geringste Ressourcen-Anforderungen. Allerdings klingen auf diese Weise erzeugte Sprachsignale bislang noch sehr unnatürlich.

- Diphonsynthese: Hierbei wird das Sprachsignal durch Verkettung von Diphonen (Nachbar-Laut Kombinationen) erzeugt. Die Prosodie-Anpassung (Rhythmus und Melodie) geschieht durch Signal-Manipulation, deren Art abhängig von der Kodierung der Diphone ist. Dieses Verfahren benötigt verhältnismäßig wenige Ressourcen und eignet sich damit auch für Embedded Anwendungen. Aber auch nach diesem Verfahren erzeugte Sprachsignale klingen bislang nicht sehr natürlich.

- Non-uniform unit-selection: Ausgehend von einer sehr großen Datenbasis werden die am besten passenden Sprachteile (units) miteinander zu Ketten variabler Länge (non-uniform) verkettet. Dabei wird eine doppelte Kostenfunktion minimiert: Die Stücke sollen gut aneinander passen (Verkettungs-Kosten) und die Vorgaben der Ziel-Prosodie erfüllen (Target-Kosten). Dies klingt sehr natürlich, nämlich ähnlich dem Originalsprecher, ist aber unflexibel bezüglich Situationen, die nicht in der Datenbasis abgebildet sind. Das Verfahren erfordert große Ressourcen und wird meist bei serverbasierten Anwendungen eingesetzt.

- HMM Synthese: Hierbei handelt es sich um die Synthese basierend auf dem so genannten Hidden Markov Model. Es handelt sich um ein stochastisches Verfahren zum Modellieren der Übergangswahrscheinlichkeiten akustischer Parameter in Bezug auf die zu erzeugende Sprache. Auf diesem Modell basierende Verfahren lernen auf einem relativ großem Sprachkorpus, brauchen aber für die Synthese nur relativ geringe Ressourcen, da sie nicht direkt auf dem Wellensignal (wave signal) basieren sondern parametrisierte Repräsentationen verwenden (zum Beispiel LPC = Linear predictive coding). Allerdings tendieren entsprechende Verfahren deshalb auch zu Artefakten.

[0003] Neben Verfahren, die zu den vorgenannten vier Hauptverfahren gehören, sind in neuerer Zeit auch Verfahren bekannt geworden, welche sich künstlicher neuronaler Netze zur Generierung künstlicher Sprachsignale bedienen. Es handelt sich hierbei, ebenso wie bei der HMM Synthese um so genannte statistische Verfahren, bei denen, wollte man sie mit der HMM Synthese vergleichen, die HMM's gewissermaßen durch künstliche neuronale Netze ersetzt werden, indem basierend auf einer sehr großen Trainingsmenge die akustischen Parameter für eine Eingangsmenge an phonetisch-linguistischen Merkmalen gelernt werden.

[0004] Wenn Menschen miteinander sprechen, geht jedoch die zwischen ihnen auf der sprachlichen Ebene geführte Kommunikation über den rein linguistischen Bedeutungsgehalt der zwischen ihnen ausgetauschten Wörter und Sätze klar hinaus. Beim Sprechen drücken Menschen immer auch Emotionen aus. Bei den entsprechenden, derartige tatsächliche oder gegebenenfalls auch nur scheinbare Emotionen zum Ausdruck bringenden Merkmalen spricht man von extralinguistischen oder paralinguistischen Merkmalen. Die extralinguistischen Merkmale werden dabei durch die Natur beziehungsweise Stimmeigenschaft eines sprechenden Menschen bestimmt. Beispiele hierfür sind unterschiedliche Stimmeigenschaften von Männern und Frauen, aber auch andere, im Grunde unbeeinflussbare Stimmeigenschaften von Menschen, welche sich beispielsweise in einer eher getragenen Sprechweise oder in einer stark akzentuierten Aussprache manifestieren. Entsprechende auffällige Sprechweisen von Menschen können von anderen, sie nicht kennenden Menschen auch als ein Ausdruck vermeintlicher Emotionen angesehen beziehungsweise interpretiert werden. Darüber hinaus werden durch einen Menschen gesprochene Wörter oder Sätze hinsichtlich ihrer Wahrnehmung durch Zuhörer noch durch paralinguistische Merkmale beeinflusst, welche tatsächliche augenblickliche Emotionen, wie Freude, Trauer oder Wut,

des Sprechenden vermitteln.

[0005] Ein Ziel bei der Entwicklung von Lösungen für die Sprachsynthese ist es, auch die zuvor erläuterten extralinguistischen und paralinguistischen Merkmale in die Synthese einzubeziehen und im Ergebnis natürlich wirkende, synthetisch erzeugte Sprachsequenzen zu erhalten. Hierbei lässt man sich davon leiten, dass es auf diese Weise möglich ist, das Verhalten von Menschen während eines im Zuge einer Mensch-Maschine-Kommunikation geführten Dialogs besser und letztlich auch stärker zielorientiert zu beeinflussen.

[0006] Zur künstlichen Generierung emotionsbehafteter Sprachsignale mittels entsprechender Sprachsynthetisierer werden unterschiedliche Merkmalsgruppen der Sprache moduliert. Hierbei handelt es sich insbesondere um den Tonhöhen- oder Melodieverlauf der Sprache (Melodiemerkmale), um die Dauer der einzelnen Laute (Dauermerkmale), um die Auswahl und Aussprache der einzelnen Laute (Artikulationsgenauigkeit) und um den Stimmklang (Stimmmerkmale). Diese Merkmalsgruppen selbst haben jeweils nicht unbedingt einen Einfluss auf die Bedeutung des Gesprochenen, verleihen aber rein linguistisch betrachteten identischen textlichen Aussagen unter Umständen eine sehr unterschiedliche Wirkung. Daher wird durch eine gezielte Veränderung beziehungsweise Beeinflussung der vorgenannten Merkmalsgruppen im Zusammenhang mit der Sprachsynthese eine emotionsbehaftete Sprechweise simuliert. Nach dem Stand der Technik werden dabei jeweils für eine bestimmte Zielemotion durchschnittliche Ausprägungen der vorgenannten Merkmalsgruppen gemessen und diese Merkmalsgruppen bei der Synthese durch Veränderung entsprechender Parameter gezielt angepasst. Beispielsweise werden für die Emotion "jubelnde Freude" die Lautdauern der stimmhaften Frikative verlängert oder es wird zum Erreichen einer Zielemotion "Trauer" der Stimmklang von "modal" auf "behaucht" verändert.

[0007] Aber auch insoweit ist die natürliche menschliche Sprechweise erheblich komplexer und vielschichtiger. Die aus dem Stand der Technik bekannt gewordenen Lösungen zur Sprachsynthese zur Erzeugung emotionsbehafteter Sprachsequenzen, welche sich der vier eingangs genannten grundsätzlichen Ansätze der Sprachsynthese bedienen und diese unter Berücksichtigung einer bestimmten Zielemotion beeinflussen, führen zwar bereits zu einer verhältnismäßig natürlich wirkenden Aussprache synthetisch erzeugter Sprachsequenzen, sind aber in dieser Hinsicht immer noch verbesserungsfähig.

[0008] Bezüglich des Standes der Technik sei in diesem Zusammenhang auf die nachfolgend genannten Dokumente verwiesen:

1. Felix Burkhardt, "Simulation emotionaler Sprechweise mit Sprachsynthesesystemen", Shaker Verlag, 2001,

2. Felix Burkhardt und W. F. Sendlmeier, "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis", Proceedings ISCA Workshop (ITRW) on Speech and Emotion, Belfast 2000,

3. T. Dutoit, V. Pagel, N. Pierret, F. Bataille und O. Van der Vreken, "The Mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396, 1996,

4. M. Schröder und J. Trouvain, "The German text-to-speech synthesis system mary: A tool for research, development and teaching", International Journal of Speech Technology, pp. 365-377, 2003

5. Felix Burkhardt, "Emofilt: the Simulation of Emotional Speech by Prosody-Transformation", Interspeech 2005.

[0009] Aufgabe der Erfindung ist es, eine Lösung für die automatisierte Sprachsynthese bereitzustellen, welche zu Sprachsequenzen führt, die bei ihrer phonetischen Wiedergabe im Hinblick auf eine emotionsbehaftete Sprechweise noch natürlicher wirken, als dies nach dem Stand der Technik der Fall ist. Hierzu sind ein Verfahren und ein zur Durchführung des Verfahrens geeignetes System anzugeben.

[0010] Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Patentanspruchs 1 gelöst. Ein die Aufgabe lösendes, zur Durchführung des Verfahrens geeignetes System wird durch den ersten Sachanspruch charakterisiert. Vorteilhafte Ausbeziehungsweise Weiterbildungen der Erfindung sind durch die Unteransprüche gegeben.

[0011] Nach dem zur Lösung der Aufgabe vorgeschlagenen Verfahren wird ein im Rahmen einer automatisierten Sprachsynthese erzeugtes Sprachsignal zur Simulation einer emotionalen Sprechweise beeinflusst. Erfindungsgemäß geschieht dies dadurch, dass ein während der Sprachsynthese erzeugtes Sprachrohsignal, welches noch nicht emotionsbehaftet ist, vor der Ausgabe des generierten Sprachsignals gezielt mit einem Parametergemisch moduliert wird. Dieses Parametergemisch umfasst erfindungsgemäß Parameter mehrerer, mit Melodiemerkmalen, mit Dauermerkmalen, mit Stimmmerkmalen oder mit Merkmalen zur Artikulationsgenauigkeit der Sprache korrespondierender Merkmalsgruppen. Hierbei werden die Parameter zumindest zweier, voneinander verschiedener der vorgenannten Merkmalsgruppen eingestellt und miteinander gemischt. Bei den insoweit eingestellten Parametern handelt es sich um Parameter aus Merkmalsgruppen, mit denen mindestens zwei vorgegebene Zielemotionen assoziiert werden.

[0012] Gemäß dem der Beschreibung und den Patentansprüchen zugrundeliegenden Verständnis handelt es sich bei einem Sprachrohsignal um ein aus Phonemen bestehendes Signal mit einer bezüglich der vorstehend genannten, Merkmalsgruppen, das heißt insbesondere im Hinblick auf die Prosodie und den Stimmklang neutralen Parameterbelegung, welches noch nicht die Form eines akustisch wiedergebbaren Audiosignals aufweist.

[0013] Vorzugsweise vollzieht sich bei der Umsetzung des zuvor grundsätzlich charakterisierten Verfahrens folgender Verfahrensablauf:

- An ein dazu ausgebildetes, hard- und softwarebasiertes System wird ein in Form von synthetisch erzeugter Sprache auszugebender Text übergeben. Mit dem Text werden an das System außerdem Informationen über mindestens zwei für das auszugebende Sprachsignal gewünschte Zielemotionen, wie beispielsweise freudige Erregung oder wütende Trauer, beigefügt.
  Der an ein entsprechendes System übergebene Text wird dann phonemisiert. Die dabei erzeugten, mit einer neutralen Prosodiebeschreibung versehenen Phoneme bilden ein Sprachrohsignal als Grundlage zur Erzeugung eines hinsichtlich seiner linguistischen Bedeutung mit dem vorgegebenen Text korrespondierenden synthetischen Sprachsignals unter Anwendung eines der bekannten Syntheseverfahren, nämlich insbesondere der Diphonesynthese, der Formantsynthese, der HMM Synthese oder eines mittels künstlicher neuronaler Netze umgesetzten statistischen Verfahrens.

- Mittels eines Emotionssimulators werden die dem übergebenen Text zu den gewünschten Zielemotionen beigefügten Informationen ausgewertet. Unter Rückgriff auf eine entsprechende Datenbasis, vorzugsweise eine Datenbank, werden Parametereigenschaften zweier verschiedener Merkmalsgruppen (Melodiemerkmale, Dauermerkmale, Artikulationsmerkmale oder Stimmklang), mit denen die vorgegebenen Zielemotionen assoziiert werden, entsprechend den vorgegebenen Zielemotionen eingestellt und zu einem Parametergemisch zusammengeführt.

- Im Zuge seiner Übergabe an die eigentliche, nach einem der vorgenannten Syntheseverfahren arbeitende Syntheseeinheit wird das in der Phonemisierungskomponente erzeugte Sprachrohsignal mit dem die mindestens zwei Zielemotionen abbildenden Parametergemisch aus den Merkmalsgruppen moduliert.

- In der Syntheseeinheit wird schließlich aus dem mit dem Parametergemisch modulierten Sprachrohsignal das auszugebende synthetische Sprachsignal, beispielsweise im Wege einer Diphonverkettung, erzeugt.

[0014] Bei dem zur Lösung der Aufgabe vorgeschlagenen und zur Umsetzung des zuvor erläuterten Verfahrens geeigneten System handelt es sich um einen speziell ausgebildeten, hard- und softwarebasierten Sprachsynthetisierer. Dieser besteht aus einer von einer Eingangsstufe des Synthetisierers umfassten Phonemisierungskomponente, aus einem Emotionssimulator sowie aus der eigentlichen, von einer mit Mitteln zur akustischen Ausgabe des erzeugten Sprachsignals ausgestatteten Ausgabestufe umfassten Syntheseeinheit. Entscheidend für die Lösung der Aufgabe ist die Art und Weise, wie die vorgenannten Komponenten, welche als solches aus dem Stand der Technik durchaus bekannt sein können, in dem erfindungsgemäßen Sprachsynthetisierer miteinander in eine Wirkverbindung gebracht sind. Die Ausbildung der vorgenannten Komponenten und die zwischen ihnen zur Umsetzung des vorgeschlagenen Verfahrens bestehende Wirkverbindung wird auf der Grundlage der hard- und softwarebasierten Struktur des Sprachsynthetisierers erreicht, welche insoweit über eine Verarbeitungseinheit verfügt und mit einer Software ausgestattet ist, bei deren Verarbeitung durch die Verarbeitungseinheit die entsprechenden, ineinandergreifenden Funktionen der Komponenten des Synthetisierers bereitgestellt werden.

[0015] Bei dem vorgeschlagenen System werden die von der Phonemisierungskomponente erzeugten, zunächst mit einer neutralen Prosodiebeschreibung versehenen Phoneme nicht unmittelbar oder, hinsichtlich von Sprachparametern, welche insbesondere die Prosodie und/oder den Stimmklang bestimmen, lediglich in Richtung einer Zielemotion beeinflusst an die Syntheseeinheit übergeben. Vielmehr werden die mit der neutralen Prosodiebeschreibung versehenen Phoneme zur Übergabe an die Syntheseeinheit mittels eines Parametergemisches moduliert. Das entsprechende Parametergemisch wird durch den zu dem Sprachsynthetisierer gehörenden Emotionssimulator erzeugt. Dieser wertet die dem übergebenen Text, welcher in synthetische Sprache umzusetzen ist, beigefügten Informationen zu wenigstens zwei Zielemotionen aus und beeinflusst die Parameter mindestens zweier unterschiedlicher Sprachmerkmalsgruppen, welche mit den Vorgaben assoziiert werden, gemäß diesen Vorgaben. Die insoweit hinsichtlich ihrer jeweiligen Ausprägung beeinflussten Parameter, mit welchen Emotionen wie Freude, Trauer und Wut assoziiert werden, werden durch den Emotionssimulator zudem gemischt, das heißt, zu einem Parametergemisch zusammengefügt und auf das Sprachrohsignal - dieses modulierendend - angewendet. Die Synthesekomponente erzeugt nach einem der bekannten Verfahren aus dem mit dem Parametergemisch modulierten Sprachrohsignal das gewünschte synthetische Sprachsignal und gibt dieses aus.

[0016] Die Erfindung soll nochmals anhand von zwei Ausführungsbeispielen zu ihren Einsatzmöglichkeiten erläutert werden. Die in den nachfolgenden Erläuterungen enthaltenen Verweise auf den Stand der Technik beziehen sich dabei auf die fünf eingangs genannten Dokumente.

[0017] Das erste Beispiel bezieht sich auf einen Roboter, der parallel mehrere Emotionen ausdrücken kann. Es sei angenommen, für den Besitzer dieses Roboters ist eine neue Nachricht eingetroffen. Diese hat, basie-

rend auf einer Inhaltsanalyse oder einer Markierung des Absenders, einen freudigen Inhalt. Der Roboter nutzt dann die Möglichkeit mehrere Emotionen gleichzeitig auszudrücken und mischt "Überraschung" (über das Eintreffen der Nachricht) mit "Freude", um den Besitzer auf deren erfreulichen Inhalt vorzubereiten. Der Roboter nutzt für seine Sprachausgabe beziehungsweise für die eigentliche Sprachsynthese eine Software des aus dem Stand der Technik bekannten, auf der Diphonsynthese basierenden Mbrola-Projekts (siehe dazu [3]).

[0018] Die Diphonsynthese erzeugt Sprache durch Verkettung einzelner Doppellaute (Diphone) aus einer Datenbasis und nachträglicher Prosodie-Anpassung (Beeinflussung von Tonmelodie und Lautdauern) nach dem PSOLA-Verfahren (Pitch Synchronous Overlap and Add). Der an dieser Stelle beispielhaft erläuterte Roboter verfügt über drei Diphon-Datenbasen für unterschiedliche Stimmqualitäten, nämlich über jeweils ein Inventar für entspannte, normale und angespannte Sprechweise. Dabei nutzt der Roboter konkret seine Parameterdaten für die beiden Zielemotionen "Überraschung" und "Freude".

[0019] Da es im Falle der Diphonsynthese vier manipulierbare Parametergruppen gibt werden jeweils zwei für die Modellierung der ersten und zwei für die der zweiten Zielemotion verwendet. Konkret bedeutet dies:

- Gemäß der ersten Zielemotion "Überraschung" wird die Merkmalsgruppe "Stimmqualität" so angepasst, dass die Diphonbasis mit angespanntem Stimmklang ausgewählt wird. Die zweite Merkmalsgruppe "Artikulationsgenauigkeit" wird auf "Vowel Target Undershoot" gestellt, das heißt die Aussprache ist eher verschliffen.
- Gemäß der zweiten Zielemotion "Freude" wird das Melodiemodell mit einem auf die hauptbetonten Silben hin sanft an- und absteigenden Verlauf modelliert. Beim Dauermodell werden die stimmhaften Frikative um 20 % verlängert, da Messungen und Resynthese-Experimente ergeben haben, das die entstehende Sprache dann eher als "freudig" wahrgenommen wird (siehe dazu [1]).

[0020] Dem Grundgedanken der Erfindung folgend, wird durch eine Phonemisierungskomponente, zum Beispiel "Mary" (siehe dazu [4]), zunächst ein Sprachrohsignal als neutrale Version der Zieläußerung (synthetische Sprache, mit den gleichzeitig zum Ausdruck gebrachten Emotionen "Überraschung" und "Freude") berechnet, bestehend aus den Phonemen die verwendet werden sollen und einer neutralen Prosodiebeschreibung. Dieses Sprachrohsignal wird dann gemäß den zuvor beschriebenen Änderungen modifiziert beziehungsweise moduliert, zum Beispiel mittels der "Emofilt" Software (siehe dazu [5]). Dem, wie bereits angegeben, auf "Mbrola" (siehe dazu [3]) basierenden Synthesizer wird dann die modifizierte Sprachsignalbeschreibung sowie der Hinweis auf die zu verwendende Diphonbasis übergeben und daraus das Sprachsignal erzeugt welches sowohl Überraschung als auch Freude ausdrückt.

[0021] Ein weiteres Ausführungsbeispiel betrifft einen virtuellen Agenten in einem Spiel. Der Nutzer interagiert mit dem Agenten unter anderem per Sprache. Um das Spiel interessant zu machen und den Agenten lebensähnlicher, wechselt dessen Stimmung basierend auf Ereignissen, die im Spiel geschehen. Der Agent hat zum Beispiel morgens erfahren, dass er im Lotto gewonnen hat und ist grundsätzlich guter Stimmung. Nun wirft er aktuell in einer Spielsituation eine wertvolle Vase um, die dabei zerbricht und macht einen wütenden Ausruf, der allerdings von dem Einfluss der positiven Stimmung gemildert wird.

[0022] Die Erzeugung des gemischten Emotionsausdrucks in der Sprache soll hierbei zum Beispiel durch Formantsynthese geschehen, wie dies beispielsweise in [2] beschrieben wird. Die wiederum auf "Mary" (siehe dazu [4]) basierende Phonemisierungskomponente erzeugt aus dem zu sprechenden Text als Sprachrohsignal eine kanonische Aussprachevariante sowie eine neutrale Prosodiebeschreibung. Diese wird von der Emotionalisierungskomponente "Emofilt" (siehe dazu [5]) dahingehend modifiziert, dass ein für die Emotion "Wut" typischer Dauer und Melodieverlauf entsteht, das heißt, der gesamte Grundfrequenzverlauf wird um 150 % angehoben, der Range (der frequenzbezogene Stimmumfang) um 40 % verbreitert und die Kontur bekommt einen finalen Anstieg der Frequenz. Alle Silben werden um 30 % beschleunigt und die starkbetonten nochmal um 20 %.

[0023] Die Rolle des Synthesizers übernimmt in diesem Falle nicht "Mbrola" (siehe dazu [3]), sondern die "EmoSyn" Software (siehe dazu [2]), welche einen Klatt-Formantsynthesizer durch eine Kombination aus Parametervorlagen für stimmhafte Laute und Regeln für stimmlose Laute ansteuert. Dabei lassen sich sämtliche Aspekte des akustischen Sprachsignals als Ergebnis eines Quelle-Filter Systems modellieren, das heißt die akustische Beschaffenheit von Kehlkopf-Anregungssignal und Mundraum-Ansatzrohr lassen sich parametrisch steuern.

[0024] Die Merkmale der Gruppen "Artikulationsgenauigkeit" und "Stimmklang" werden gemäß der zweiten Zielemotion, also "Zufriedenheit", modelliert. Das heißt, für den Stimmklang wird eine eher behauchte Sprechweise verwendet, wobei (gemäß [1], S. 218) das Anregungssignal durch den Liljencranz-Fant Parameter Öffnungsquotient, die spektrale Dämpfung, die Bandbreite des ersten Formanten, die Amplitude des stimmhaften Anteils und die Amplitude des rauschhaften Anteils angepasst wird. Die Formanten werden leicht angehoben da dies einer lächelnden Sprechweise entspricht. Die Artikulationsgenauigkeit ist erhöht, es wird also ein "Vowel-Target overshoot" modelliert durch Dezentralisierung der ersten beiden Formanten.

[0025] Anhand zugehöriger Zeichnungen soll an dieser Stelle noch die grundsätzliche Struktur zweier möglicher Ausbildungsformen des erfindungsgemäßen Sys-

tems dargestellt werden. Die Zeichnungen zeigen im Einzelnen:

Fig. 1: einen erfindungsgemäßen Sprachsynthetisierer mit einer nach dem Prinzip der Diphonverkettung arbeitenden Syntheseeinheit,

Fig. 2: einen Sprachsynthetisierer mit einer das auszugebende Sprachsignal im Wege der Formantsynthese erzeugenden Syntheseeinheit.

[0026] Die Fig. 1 zeigt eine erste mögliche Ausbildungsform des erfindungsgemäßen Systems 1, nämlich eines Sprachsynthetisierers, in einer sehr stark vereinfachten schematischen Darstellung. Bei allen dargestellten Komponenten des Sprachsynthetisierers handelt es sich jeweils um soft- und hardwarebasierte Komponenten.

[0027] Demgemäß besteht der Sprachsynthetisierer 1 im Wesentlichen, wie aus der Fig. 1 ersichtlich, aus der Eingangsstufe 2 mit der Phonemisierungskomponente 3, dem Emotionssimulator 4 und der Ausgangsstufe 5 mit der eigentlichen Syntheseeinheit 6 und mit - hier allerdings nicht dargestellten - Mitteln (zum Beispiel Lautsprecher) für die akustische Ausgabe. Bei dem gezeigten Beispiel ist die Phonemisierungskomponente 3 beispielsweise durch das hierfür bekannte Phonemisierungssystem "Mary" ausgebildet. Der Emotionssimulator 4 basiert auf der zur Erzeugung emotionaler Sprache geschaffenen Software "Emofilt". Die Komponente stützt sich auf eine Datenbank ab, in welcher Parameter für die "Emotionen "Freude", "Trauer" und "Wut" in vier voneinander unabhängigen Merkmalsgruppen, nämlich den Melodiemerkmalen, den Dauermerkmalen, den Artikulationsmerkmalen und Merkmalen des Stimmklangs, gehalten werden. Die betreffenden Parameter, nämlich Parameter aus mindestens zwei der vorgenannten Merkmalsgruppen, werden mittels der "Emofilt" Software entsprechend mindestens zweier, an das System 1 im Zusammenhang mit der Übergabe eines in Sprache umzusetzenden Textes übergebener Zielemotionen beeinflusst. Zur Umsetzung des erfindungsgemäßen Verfahrens werden dann die entsprechend eingestellten beziehungsweise eingeregelten Parameter durch dafür als Bestandteil des Emotionssimulators 4 zusätzlich vorgesehene Programmsequenzen zu einem Gemisch zusammengefügt, mittels welchem die in der Phonemisierungskomponente 3 erzeugten, ein Sprachrohsignal bildenden Phoneme moduliert werden. Als Syntheseeinheit 5 dient bei dem in der Fig. 1 gezeigten Ausführungsbeispiel eine Komponente, in welcher die Software "Mbrola" implementiert ist. Diese Komponente, basierend auf "Mbrola", erzeugt aus dem wie zuvor angegeben modulierten Sprachrohsignal mit neutraler Sprachsignalparameterbelegung - also aus den hinsichtlich des Stimmklangs ohnehin neutralen (Mary versieht die Phoneme nicht mit Informationen oder Parametern zum Stimmklang) Phonemen mit gleichfalls neutraler Prosodie, welche entsprechend den zwei Zielemotionen moduliert wurden - im Wege der Diphoneverkettung schließlich das von den (nicht gezeigten) Ausgabemitteln des Sprachsynthetisierers 1 auszugebende synthetische Sprachsignal.

[0028] Die Fig. 2 zeigt eine weitere Ausführungsform eines gemäß der Erfindung gestalteten Sprachsynthetisierers 1. Dieser umfasst die grundsätzlich gleichen Komponenten wie der Sprachsynthetisierer 1 gemäß dem zuvor erläuterten Ausführungsbeispiel. Die betreffenden Komponenten arbeiten auch in derselben Weise zusammen. Bei dem in der Fig. 2 gezeigten Sprachsynthetisierer 1 ist gegenüber dem Ausführungsbeispiel nach der Fig. 1 lediglich die Sprachsyntheseeinheit 5 durch eine andere Ausbildungsform für eine derartige Einheit ersetzt worden. Die eigentliche Sprachsynthese erfolgt mittels dieser Komponente, in welcher beispielsweise die Software "EmoSyn" implementiert wurde. Diese erzeugt ein synthetisches Sprachsignal nach dem Prinzip der Formantsynthese.

Patentansprüche

1. Verfahren zur Sprachsynthese, nach welchem durch ein dafür ausgebildetes System (1) in einem automatisierten Ablauf ein synthetisches Sprachsignal erzeugt wird, dessen Parametereigenschaften zur Simulation einer emotionalen Sprechweise beeinflusst werden, **dadurch gekennzeichnet, dass** ein während der Sprachsynthese erzeugtes, noch nicht emotionsbehaftetes Sprachrohsignal zur Generierung des auszugebenden, emotionsbehafteten Sprachsignals mit einem Parametergemisch moduliert wird, welches Parameter mehrerer, mit Melodiemerkmalen, mit Dauermerkmalen, mit Stimmmerkmalen oder mit der Artikulationsgenauigkeit der Sprache korrespondierender Merkmalsgruppen umfasst, die entsprechend mindestens zwei vorgegebenen, mit voneinander verschiedenen der vorgenannten Merkmalsgruppen assoziierten Zielemotionen eingestellt werden.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet, dass** der automatisierte Ablauf zur Sprachsynthese folgende Verfahrensschritte umfasst:

   a.) Übergabe eines als Audiosprachsignal auszugebenden Textes und von Informationen über mindestens zwei für das als Audiosprachsignal auszugebende synthetische Sprachsignal gewünschte Zielemotionen an das zur Sprachsynthese ausgebildete System (1),
   b.) Erzeugung eines Sprachrohsignals durch Phonemisierung des übergebenen Textes,
   c.) Auswertung der mit dem Text übergebenen Zielemotionen und Einstellung der Parametereigenschaften mindestens zweier verschiede-

ner, mit Melodiemerkmalen, mit Dauermerkmalen, mit Stimmmerkmalen oder mit der Artikulationsgenauigkeit der Sprache korrespondierender Merkmalsgruppen entsprechend den vorgegebenen Zielemotionen,

d.) Zusammenführen der entsprechend den mindestens zwei Zielemotionen eingestellten Parametereigenschaften zu einem Parametergemisch,

e.) Modulierung des erzeugten Sprachrohsignals mit dem aus den eingestellten Parametereigenschaften gebildeten Parametergemisch,

f.) Erzeugung eines synthetischen Sprachsignals aus dem mit dem Parametergemisch modulierten Sprachrohsignal,

g.) Ausgabe des emotionsbehafteten synthetischen Sprachsignals als Audiosprachsignal durch das System (1).

3. Verfahren nach Anspruch 3, **dadurch gekennzeichnet, dass** die Erzeugung des synthetischen Sprachsignals aus dem mit dem Parametergemisch modulierten Sprachrohsignal durch Formatsynthese erfolgt.

4. Verfahren nach Anspruch 3, **dadurch gekennzeichnet, dass** die Erzeugung des synthetischen Sprachsignals aus dem mit dem Parametergemisch modulierten Sprachrohsignal durch Diphonsynthese erfolgt.

5. Verfahren nach Anspruch 3, **dadurch gekennzeichnet, dass** die Erzeugung des synthetischen Sprachsignals aus dem mit dem Parametergemisch modulierten Sprachrohsignal durch HHM-Synthese, nämlich durch Synthese basierend auf dem Hidden Markov Model, erfolgt.

6. Verfahren nach Anspruch 3, **dadurch gekennzeichnet, dass** die Erzeugung des synthetischen Sprachsignals aus dem mit dem Parametergemisch modulierten Sprachrohsignal mittels neuronaler Netze erfolgt.

7. System (1) zur Sprachsynthese, nämlich hard- und softwarebasierter Sprachsynthetisierer zur automatisierten Erzeugung eines eine emotionale Sprechweise simulierenden synthetischen Sprachsignals, bestehend aus einer Eingangsstufe (2) zur Entgegennahme eines als synthetische Sprache auszugebenden Textes mit einer Phonemisierungskomponente (3) zur Erzeugung eines Sprachrohsignals durch Phonemisierung entgegengenommenen Textes, aus einem Emotionssimulator (4) zur Beeinflussung des Sprachrohsignals für die Simulation einer emotionalen Sprechweise und aus einer Ausgangsstufe (5) mit einer Syntheseeinheit (6) zur Erzeugung eines emotionsbehafteten synthetischen Sprachsignals aus dem mittels des Emotionssimulators (4) beeinflussten Sprachrohsignal und mit Mitteln zur akustischen Ausgabe des erzeugten synthetischen Sprachsignals, **dadurch gekennzeichnet, dass** die Eingangsstufe (2) zur Entgegennahme von Informationen über mindestens zwei Zielemotionen für das zu erzeugende synthetische Sprachsignal ausgebildet ist und dass der Emotionssimulator (4) ausgebildet ist zur Auswertung von der Eingangsstufe (2) entgegengenommener Informationen zu Zielemotionen und zur Einstellung der Parametereigenschaften mindestens zweier verschiedener, mit Melodiemerkmalen, mit Dauermerkmalen, mit Stimmmerkmalen oder mit der Artikulationsgenauigkeit der Sprache korrespondierender Merkmalsgruppen entsprechend den der Auswertung der Informationen ermittelten Zielemotionen sowie zum Mischen der für die Parametereigenschaften des zu erzeugenden Sprachsignals eingestellten Parameter und zur Modulation des Sprachrohsignals mit diesem Parametergemisch.

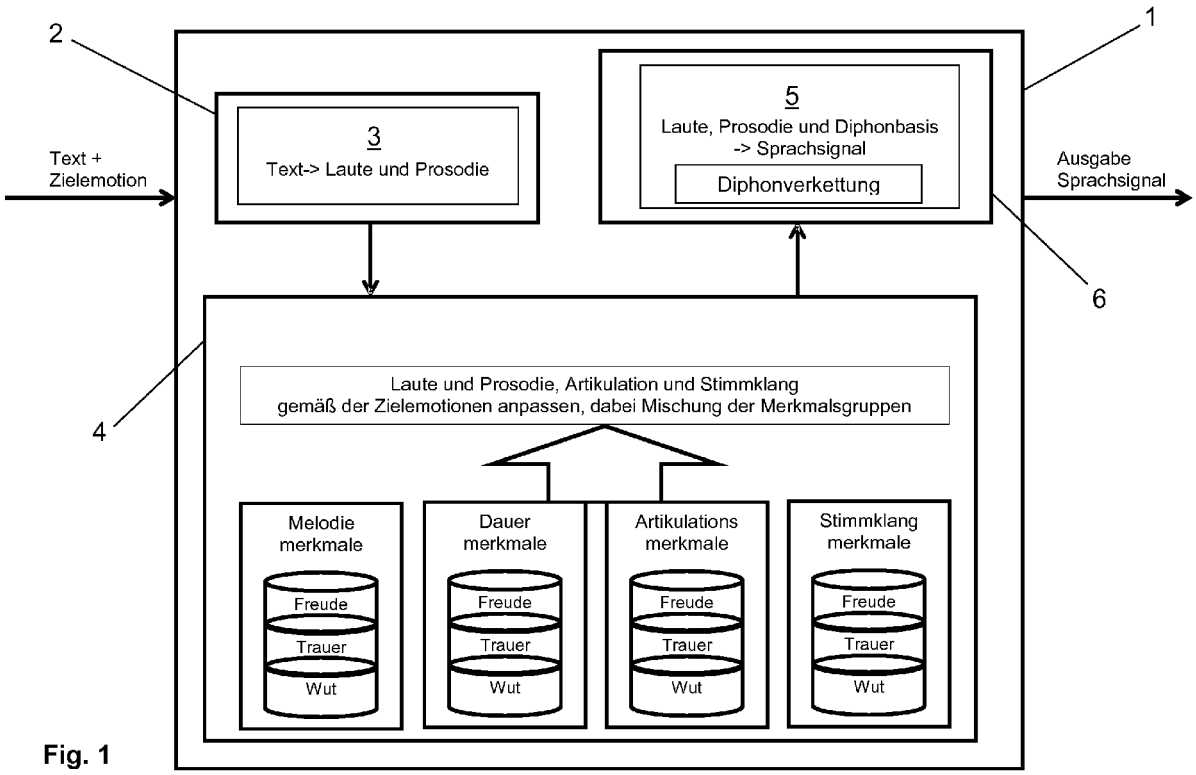
2
1
3
Text-> Laute und Prosodie
5
Laute, Prosodie und Diphonbasis
-> Sprachsignal
Diphonverkettung
Text +
Zielemotion
Ausgabe
Sprachsignal
6
4
Laute und Prosodie, Artikulation und Stimmklang
gemäß der Zielemotionen anpassen, dabei Mischung der Merkmalsgruppen
Melodie
merkmale
Dauer
merkmale
Artikulations
merkmale
Stimmklang
merkmale
Freude
Trauer
Wut

Fig. 1

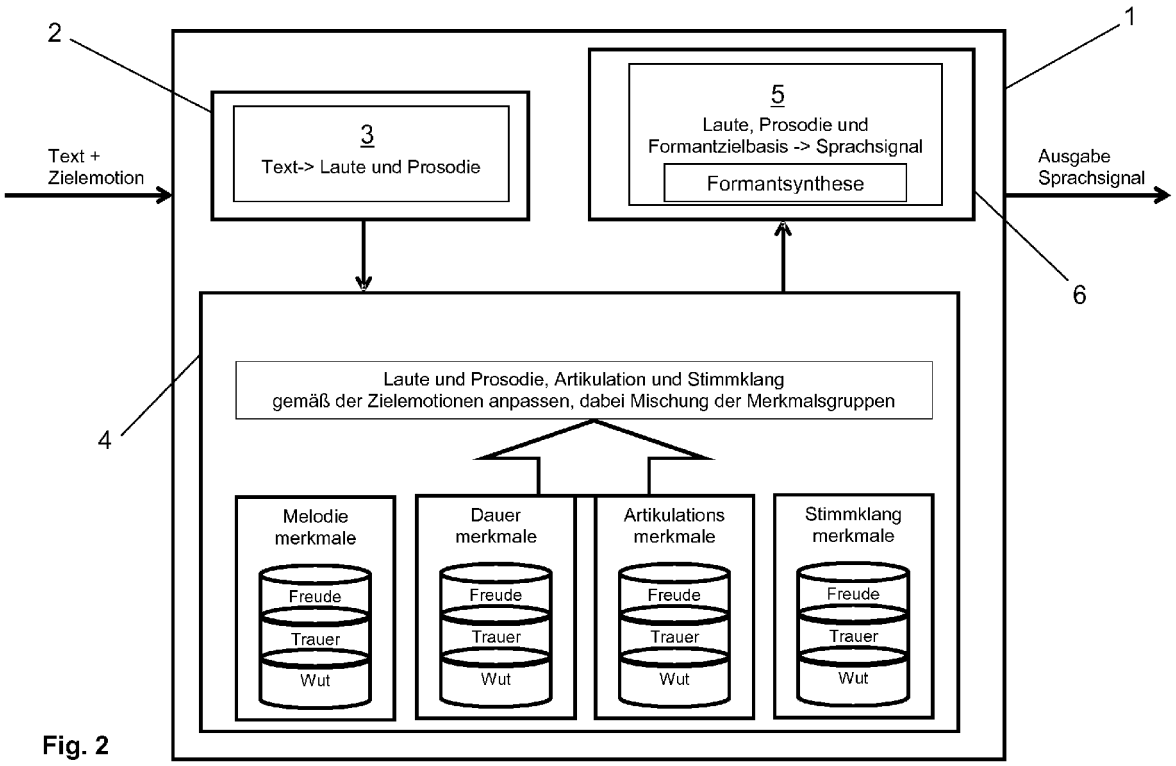
2
1
3
Text-> Laute und Prosodie
5
Laute, Prosodie und
Formantzielbasis -> Sprachsignal
Formantsynthese
Text +
Zielemotion
Ausgabe
Sprachsignal
6
4
Laute und Prosodie, Artikulation und Stimmklang
gemäß der Zielemotionen anpassen, dabei Mischung der Merkmalsgruppen
Melodie
merkmale
Dauer
merkmale
Artikulations
merkmale
Stimmklang
merkmale
Freude
Trauer
Wut

Fig. 2

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPÄISCHER RECHERCHENBERICHT

Nummer der Anmeldung
EP 15 18 5879

EINSCHLÄGIGE DOKUMENTE

Kategorie	Kennzeichnung des Dokuments mit Angabe, soweit erforderlich, der maßgeblichen Teile	Betrifft Anspruch	KLASSIFIKATION DER ANMELDUNG (IPC)
X	US 6 226 614 B1 (MIZUNO OSAMU [JP] ET AL) 1. Mai 2001 (2001-05-01) * Abbildungen 1,3 * * Spalte 6, Zeilen 38-40, 54, 55 * * Spalte 8, Zeilen 55, 56 * * Spalte 9, Zeile 27 - Spalte 10, Zeile 11 * -----	1-7	INV. G10L13/033 G10L13/047
X,D	Felix Burkhardt: "Emofilt: the Simulation of Emotional Speech by Prosody-Transformation", Proc. Interspeech, 4. September 2005 (2005-09-04), Seiten 1-4, XP055225958, Gefunden im Internet: URL:http://felix.syntheticspeech.de/publications/emofiltInterspeech05.pdf [gefunden am 2015-11-04] * Zusammenfassung * * Abschnitt 1, zweiter Absatz * * Abschnitt 2 * * Abschnitt 3, erster Satz * * Abschnitt 5.1 * * Abschnitt 6, letzter Satz * -----	1-7	

RECHERCHIERTE SACHGEBIETE (IPC)

G10L

Der vorliegende Recherchenbericht wurde für alle Patentansprüche erstellt

Recherchenort	Abschlußdatum der Recherche	Prüfer
München	26. November 2015	Tilp, Jan

KATEGORIE DER GENANNTEN DOKUMENTE

X : von besonderer Bedeutung allein betrachtet
Y : von besonderer Bedeutung in Verbindung mit einer anderen Veröffentlichung derselben Kategorie
A : technologischer Hintergrund
O : nichtschriftliche Offenbarung
P : Zwischenliteratur

T : der Erfindung zugrunde liegende Theorien oder Grundsätze
E : älteres Patentdokument, das jedoch erst am oder nach dem Anmeldedatum veröffentlicht worden ist
D : in der Anmeldung angeführtes Dokument
L : aus anderen Gründen angeführtes Dokument

& : Mitglied der gleichen Patentfamilie, übereinstimmendes Dokument

EPO FORM 1503 03.82 (P04C03)

ANHANG ZUM EUROPÄISCHEN RECHERCHENBERICHT ÜBER DIE EUROPÄISCHE PATENTANMELDUNG NR. EP 15 18 5879

In diesem Anhang sind die Mitglieder der Patentfamilien der im obengenannten europäischen Recherchenbericht angeführten Patentdokumente angegeben.
Die Angaben über die Familienmitglieder entsprechen dem Stand der Datei des Europäischen Patentamts am
Diese Angaben dienen nur zur Unterrichtung und erfolgen ohne Gewähr.

26-11-2015

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 6226614 B1	01-05-2001	CA 2238067 A1	21-11-1998
		DE 69821673 D1	25-03-2004
		DE 69821673 T2	05-01-2005
		EP 0880127 A2	25-11-1998
		US 6226614 B1	01-05-2001
		US 6334106 B1	25-12-2001

EPO FORM P0461

Für nähere Einzelheiten zu diesem Anhang : siehe Amtsblatt des Europäischen Patentamts, Nr.12/82

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

- **FELIX BURKHARDT.** Simulation emotionaler Sprechweise mit Sprachsynthesesystemen. Shaker Verlag, 2001 **[0008]**
- **FELIX BURKHARDT ; W. F. SENDLMEIER.** Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis. *Proceedings ISCA Workshop (ITRW) on Speech and Emotion,* 2000 **[0008]**
- **T. DUTOIT ; V. PAGEL ; N. PIERRET ; F. BATAILLE ; O. VAN DER VREKEN.** The Mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96,* 1996, vol. 3, 1393-1396 **[0008]**
- **M. SCHRÖDER ; J. TROUVAIN.** The German text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology,* 2003, 365-377 **[0008]**
- **FELIX BURKHARDT.** Emofilt: the Simulation of Emotional Speech by Prosody-Transformation. *Interspeech,* 2005 **[0008]**