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

(11) **EP 3 962 115 B1**

(12) **EUROPÄISCHE PATENTSCHRIFT**

(45) Veröffentlichungstag und Bekanntmachung des Hinweises auf die Patenterteilung:
18.12.2024 Patentblatt 2024/51

(21) Anmeldenummer: **21190918.9**

(22) Anmeldetag: **12.08.2021**

(51) Internationale Patentklassifikation (IPC):
H04R 25/00 *(2006.01)* ***G10L 25/60*** *(2013.01)*
G10L 25/15 *(2013.01)*

(52) Gemeinsame Patentklassifikation (CPC):
H04R 25/505; G10L 25/15; G10L 25/60;
H04R 2225/43

(54) VERFAHREN ZUR BEWERTUNG DER SPRACHQUALITÄT EINES SPRACHSIGNALS MITTELS EINER HÖRVORRICHTUNG

METHOD FOR EVALUATING THE SPEECH QUALITY OF A SPEECH SIGNAL BY MEANS OF A HEARING DEVICE

PROCÉDÉ D'ÉVALUATION DE LA QUALITÉ DE PAROLE D'UN SIGNAL VOCAL AU MOYEN D'UN DISPOSITIF AUDITIF

(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

(30) Priorität: **28.08.2020 DE 102020210919**

(43) Veröffentlichungstag der Anmeldung:
02.03.2022 Patentblatt 2022/09

(73) Patentinhaber: **Sivantos Pte. Ltd.**
Singapore 539775 (SG)

(72) Erfinder:
- **THIEMT, Jana**
  91052 Erlangen (DE)
- **LUGGER, Marko**
  91365 Weilersbach (DE)

(74) Vertreter: **FDST Patentanwälte**
Nordostpark 16
90411 Nürnberg (DE)

(56) Entgegenhaltungen:
US-A1- 2004 167 774
US-A1- 2018 255 406
US-B2- 7 165 025

- **ASGER HEIDEMANN ANDERSEN ET AL: "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, USA, vol. 26, no. 10, 1 October 2018 (2018-10-01), pages 1925 - 1939, XP058416624, ISSN: 2329-9290, DOI: 10.1109/TASLP.2018.2847459**

EP 3 962 115 B1

Beschreibung

[0001] Die Erfindung betrifft ein Verfahren zur Bewertung der Sprachqualität eines Sprachsignals mittels einer Hörvorrichtung, wobei mittels eines akusto-elektrischen Eingangswandlers der Hörvorrichtung ein das Sprachsignal enthaltender Schall aus einer Umgebung der Hörvorrichtung aufgenommen und in ein Eingangs-Audiosignal umgewandelt wird, wobei durch Analyse des Eingangs-Audiosignals mittels einer Signalverarbeitung mindestens Eigenschaft des Sprachsignals quantitativ erfasst wird.

[0002] Eine wichtige Aufgabe in der Anwendung von Hörvorrichtungen, wie z.B. von Hörgeräten, aber auch von Headsets oder Kommunikationsgeräten, besteht oftmals darin, ein Sprachsignal möglichst präzise, also insbesondere akustisch möglichst verständlich an einen Benutzer der Hörvorrichtung auszugeben. Oftmals werden hierzu in einem Audiosignal, welches anhand eines Schalls mit einem Sprachsignal erzeugt wird, Störgeräusche aus dem Schall unterdrückt, um die Signalanteile, welche das Sprachsignal repräsentieren, hervorzuheben und somit dessen Verständlichkeit zu verbessern. Oftmals kann jedoch durch Algorithmen zur Rauschunterdrückung die Klangqualität eines resultierenden Ausgangssignals verringert werden, wobei durch eine Signalverarbeitung des Audiosignals insbesondere Artefakte entstehen können, und/oder ein Höreindruck generell als weniger natürlich empfunden wird.

[0003] Meist wird eine Rauschunterdrückung hierbei anhand von Kenngrößen durchgeführt, welche vorrangig das Rauschen oder das Gesamtsignal betreffen, also z.B. ein Signal-zu-Rausch-Verhältnis ("signal-to-noise-ratio", SNR), ein Grundrauschpegel ("noise foor"), oder auch einen Pegel des Audiosignals. Dieser Ansatz für eine Steuerung der Rauschunterdrückung kann jedoch letztlich dazu führen, dass die Rauschunterdrückung auch dann angewandt wird, wenn dies, obwohl merkliche Störgeräusche vorliegen, infolge von trotz der Störgeräusche weiter gut verständlichen Sprachanteilen gar nicht erforderlich wäre. In diesem Fall wird das Risiko einer nachlassenden Klangqualität, z.B. durch Artefakte der Rauschunterdrückung, ohne echte Notwendigkeit eingegangen. Umgekehrt kann ein Sprachsignal, welches nur von geringem Rauschen überlagert ist, und insofern das zugehörige Audiosignal ein gutes SNR aufweist, bei einer schwachen Artikulation des Sprechers auch eine geringe Sprachqualität aufweisen.

[0004] Dies könnte vermieden werden, wenn in einer Hörvorrichtung Algorithmen zur Rauschunterdrückung im Besonderen, aber auch die Signalverarbeitung im Allgemeinen, in Abhängigkeit einer Qualität eines Sprachsignalanteils im zu verarbeitenden Audiosignal gesteuert würden. Hierfür ist jedoch erforderlich, eine solche Qualität überhaupt mess- und erfassbar bar zu machen.

[0005] Die US 2004 / 0 167 774 A1 nennt ein Verfahren zum Analysieren und Bewerten von Stimmen. Hierbei wird ein Testsprachsignal unter Verwendung eines Hörmodells verarbeitet, wenigstens ein Merkmal einer Sprachqualität aus dem Testsprachsignal ermittelt, und das besagte Merkmal der Sprachqualität mit einem entsprechenden Basis-Merkmal der Sprachqualität. Darauf basierend kann ein Maß für eine Sprachqualität des Testsprachsignals ermittelt werden.

[0006] In A. H. Andersen et al., "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks", IEEE/ACM Transactions on Audio, Speech and Language Processing, IEEE, USA, Bd. 26, Nr. 10, 01.10.2018, Seiten 1925-1939, wird ein sog. Convolutional Neural Network (CNN) für eine nicht-intrusive Vorhersage der Sprachverständlichkeit ("Speech Intelligibility Prediction", SIP), welche kein sauberes (also rauschfreies) Sprachsignal für Vorhersagen benötigt. Die verwendete CNN-Architektur weist dabei Ähnlichkeit zu bestehenden SIP-Algorithmen auf, wodurch die trainierten Gewichte des CNN einfach und aussagekräftig zu interpretieren sind. Die vorgeschlagene Methode zeigt eine hohe Vorhersageleistung im Vergleich zu bestehenden intrusiven und nicht-intrusiven SIP-Algorithmen. Dies demonstriert das Potenzial von Deep Learning für die Vorhersage der Sprachverständlichkeit Vorhersage.

[0007] Die US 2018 / 0 255 406 A1 nennt ein Hörgerät, das eine Anzahl an Mikrofonen mit einem ersten Mikrofon zum Erzeugen eines ersten Mikrofoneingangssignals sowie und einen Prozessor zum Verarbeiten von Eingangssignalen zu einem elektrischen Ausgangssignal umfasst. Das Hörgerät umfasst weiter einen Receiver zum Umwandeln des elektrischen Ausgangssignals in ein Ausgangsschallsignal und eine Steuereinheit, die operativ mit der Anzahl an Mikrofonen verbunden ist, wobei die Steuereinheit basierend auf einem oder mehreren Mikrofoneingangssignale einen Indikator einer Sprachverständlichkeit schätzt. Die Steuereinheit ist weiter dazu eingerichtet, den Prozessor basierend auf dem Indikator der Sprachverständlichkeit zu steuern. Für das Schätzen des besagten Indikators wird ein Tonhöhenparameter einer ersten Audioquelle geschätzt. Der Indikator der Sprachverständlichkeit basiert auf dem Tonhöhenparameter und einer Richtung der ersten Audioquelle.

[0008] Die US 7,165,025 nennt eine Artikulationsanalyse zur Verwendung bei der Sprachqualitätsbewertung. Die Artikulationsanalyse basiert auf einem Vergleich zwischen Leistungen von Artikulations- und Nicht-Artikulations-Frequenzbereichen eines Sprachsignals, wobei die Sprachqualität basierend auf dem besagten Vergleich bewertet wird. Weder das Ausgangssprachsignal noch eine Schätzung desselben wird in der Artikulationsanalyse verwendet.

[0009] Der Erfindung liegt daher die Aufgabe zugrunde, ein Verfahren anzugeben, mittels dessen ein Sprachanteil in einem von einer Hörvorrichtung zu verarbeitenden Audiosignal objektiv in seiner Qualität bewertet werden kann. Der Erfindung liegt weiter die Aufgabe zugrunde, eine Hörvorrichtung anzugeben, welche dazu eingerichtet ist, für ein internes Audiosignal eine Qualität eines darin enthaltenen Sprachanteils objektiv zu bewerten.

[0010] Die erstgenannte Aufgabe wird erfindungsgemäß gelöst durch ein Verfahren zur Bewertung der Sprachqualität eines Sprachsignals mittels einer Hörvorrichtung, wobei mittels eines akusto-elektrischen Eingangswandlers der Hörvorrichtung ein das Sprachsignal enthaltender Schall aus einer Umgebung der Hörvorrichtung aufgenommen und in ein Eingangs-Audiosignal umgewandelt wird, wobei durch Analyse des Eingangs-Audiosignals mittels einer Signalverarbeitung, insbesondere einer Signalverarbeitung der Hörvorrichtung und/oder einer mit der Hörvorrichtung verbindbaren Hilfsvorrichtung, mindestens eine artikulatorische Eigenschaft des Sprachsignals quantitativ erfasst wird, und wobei in Abhängigkeit von der mindestens einen artikulatorischen Eigenschaft ein quantitatives Maß für die Sprachqualität abgeleitet wird. Vorteilhafte und teils für sich gesehen erfinderische Ausgestaltungen sind Gegenstand der Unteransprüche und der nachfolgenden Beschreibung.

[0011] Die zweitgenannte Aufgabe wird erfindungsgemäß gelöst durch eine Hörvorrichtung, welche einen akusto-elektrischen Eingangswandler und eine insbesondere einen Signalprozessor aufweisende Signalverarbeitungseinrichtung umfasst, wobei der akusto-elektrischen Eingangswandler dazu eingerichtet ist, einen Schall aus einer Umgebung der Hörvorrichtung aufzunehmen und in ein Eingangs-Audiosignal umzuwandeln, und wobei die Signalverarbeitungseinrichtung dazu eingerichtet ist, durch eine Analyse des Eingangs-Audiosignals mindestens eine artikulatorische Eigenschaft eines im Eingangs-Audiosignal enthaltenen Anteils eines Sprachsignals quantitativ zu erfassen und in Abhängigkeit von der mindestens einen artikulatorischen Eigenschaft gemäß dem vorbeschriebenen Verfahren ein quantitatives Maß für die Sprachqualität abzuleiten.

[0012] Die erfindungsgemäße Hörvorrichtung teilt die Vorzüge des erfindungsgemäßen Verfahrens, welches insbesondere mittels der erfindungsgemäßen Hörvorrichtung durchführbar ist. Die für das Verfahren und für seine Weiterbildungen nachfolgend genannten Vorteile können hierbei sinngemäß auf die Hörvorrichtung übertragen werden.

[0013] Unter einem akusto-elektrischen Eingangswandler ist hierbei insbesondere jedweder Wandler umfasst, welcher dazu eingerichtet ist, aus einem Schall der Umgebung ein elektrisches Audiosignal zu erzeugen, sodass durch den Schall hervorgerufene Luftbewegungen und Luftdruckschwankungen am Ort des Wandlers durch entsprechende Oszillationen einer elektrischen Größe, insbesondere einer Spannung im erzeugten Audiosignal wiedergegeben werden. Insbesondere kann der akusto-elektrische Eingangswandler durch ein Mikrofon gegeben sein.

[0014] Die Signalverarbeitung erfolgt insbesondere mittels einer entsprechenden Signalverarbeitungseinrichtung, welche mittels wenigstens eines Signalprozessors zur Durchführung der für die Signalverarbeitung vorgesehenen Berechnungen und/oder Algorithmen eingerichtet ist. Die Signalverarbeitungseinrichtung ist dabei insbesondere auf der Hörvorrichtung angeordnet. Die Signalverarbeitungseinrichtung kann jedoch auch auf einer Hilfsvorrichtung angeordnet sein, welche für eine Verbindung mit der Hörvorrichtung zum Datenaustausch eingerichtet ist, also z.B. ein Smartphone, eine Smartwatch o.ä. Die Hörvorrichtung kann dann z.B. das Eingangs-Audiosignal an die Hilfsvorrichtung übertragen, und die Analyse wird mittels der durch die Hilfsvorrichtung bereitgestellten Rechenressourcen durchgeführt. Abschließend kann als Ergebnis der Analyse das quantitative Maß an die Hörvorrichtung zurück übertragen werden.

[0015] Die Analyse kann dabei direkt am Eingangs-Audiosignal durchgeführt werden, oder anhand eines vom Eingangs-Audiosignal abgeleiteten Signals. Ein solches kann hierbei insbesondere durch den isolierten Sprachsignalanteil gegeben sein, aber auch durch ein Audiosignal, wie es z.B. in einer Hörvorrichtung durch eine Rückkopplungsschleife mittels eines Kompensationssignals zur Kompensation einer akustischen Rückkopplung erzeugt werden kann o.ä., oder durch ein Richtsignal, welches anhand eines weiteren Eingangs-Audiosignals eines weiteren Eingangswandlers erzeugt wird.

[0016] Unter einer artikulatorischen Eigenschaft des Sprachsignals sind hierbei eine Präzision von Formanten, besonders von Vokalen, sowie eine Dominanz von Konsonanten, besonders von Frikativen und/oder Plosiven, umfasst. Hierbei lässt sich die Aussage treffen, dass eine Sprachqualität als umso höher anzusetzen ist, je höher die Präzision der Formanten ist bzw. je höher die Dominanz und/oder Präzision von Konsonanten ist. Unter einer prosodischen Eigenschaft des Sprachsignals sind insbesondere eine Zeitstabilität einer Grundfrequenz des Sprachsignals und eine relative Schallintensität von Akzenten umfasst.

[0017] Klangerzeugung umfasst üblicherweise drei physikalische Bestandteile einer Schallquelle: Einen mechanischen Oszillator wie z.B. eine Saite oder Membran, welcher eine den Oszillator umgebende Luft in Schwingungen versetzt, eine Anregung des Oszillators (z.B. durch ein Zupfen oder Streichen), und einen Resonanzkörper. Der Oszillator wird durch die Anregung in Oszillationen versetzt, sodass die den Oszillator umgebende Luft durch die Schwingungen des Oszillators in Druckschwingungen versetzt wird, welche sich als Schallwellen ausbreiten. Hierbei werden im mechanischen Oszillator meist nicht nur Schwingungen einer einzigen Frequenz angeregt, sondern Schwingungen verschiedener Frequenzen, wobei die spektrale Zusammensetzung der propagierenden Schwingungen das Klangbild bestimmt. Die Frequenzen von bestimmten Schwingungen sind dabei oft als ganzzahlige Vielfache einer Grundfrequenz gegeben, und werden als "Harmonische" oder als Obertöne dieser Grundfrequenz bezeichnet. Es können sich jedoch auch komplexere spektrale Muster herausbilden, sodass nicht alle erzeugten Frequenzen als Harmonische derselben Grundfrequenz darstellbar sind. Für das Klangbild ist hierbei auch die Resonanz der erzeugten

Frequenzen im Resonanzraum relevant, da oftmals bestimmte, vom Oszillator erzeugte Frequenzen im Resonanzraum relativ zu den dominanten Frequenzen eines Klangs abgeschwächt werden.

[0018] Auf die menschliche Stimme angewandt bedeutet dies, dass der mechanische Oszillator gegeben ist durch die Stimmbänder, und deren Anregung in der aus den Lungen an den Stimmbändern vorbeiströmenden Luft, wobei der Resonanzraum v.a. durch den Rachen- und Mundraum gebildet wird. Die Grundfrequenz einer männlichen Stimme liegt dabei meist im Bereich von 60 Hz bis 150 Hz, für Frauen meist im Bereich von 150 Hz bis 300 Hz. Infolge der anatomischen Unterschiede zwischen einzelnen Menschen sowohl hinsichtlich ihrer Stimmbänder, als auch insbesondere hinsichtlich des Rachen- und Mundraums bilden sich zunächst unterschiedliche klingende Stimmen aus. Durch eine Veränderung des Volumens und der Geometrie des Mundraums durch entsprechende Kiefer- und Lippenbewegungen kann dabei der Resonanzraum derart verändert werden, dass sich für die Erzeugung von Vokalen charakteristische Frequenzen ausbilden, sog. Formanten. Diese liegen jeweils für einzelne Vokale in unveränderlichen Frequenzbereichen (den sog. "Formantenbereichen"), wobei ein Vokal meist durch die ersten zwei Formanten F1 und F2 einer Reihe von oftmals vier Formanten bereits klar hörbar gegen andere Laute abgegrenzt ist (vgl. "Vokaldreieck" und "Vokaltrapez"). Die Formanten bilden sich hierbei unabhängig von der Grundfrequenz, also der Frequenz der Grundschwingung aus.

[0019] Unter einer Präzision von Formanten ist in diesem Sinn insbesondere ein Grad einer Konzentration der akustischen Energie auf voneinander abgrenzbare Formantenbereiche, insbesondere jeweils auf einzelne Frequenzen in den Formantenbereichen, und eine hieraus resultierende Bestimmbarkeit der einzelnen Vokale anhand der Formanten zu verstehen.

[0020] Für ein Erzeugung von Konsonanten wird der an den Stimmbändern vorbeiströmende Luftstrom an wenigstens einer Stelle teilweise oder ganz blockiert, wodurch u.a. auch Turbulenzen des Luftstroms gebildet werden, weswegen nur manchen Konsonanten eine ähnlich klare Formantenstruktur zugeordnet werden kann wie Vokalen, und andere Konsonanten eine eher breitbandige Frequenzstruktur aufweisen. Jedoch lassen sich auch Konsonanten bestimmte Frequenzbänder zuordnen, in welchen die akustische Energie konzentriert ist. Diese liegen infolge der eher perkussiven "Geräuschartigkeit" von Konsonanten allgemein oberhalb der Formantenbereiche von Vokalen, nämlich vorrangig im Bereich von ca. 2 bis 8 kHz, während die Bereiche der wichtigsten Formanten F1 und F2 von Vokalen allgemein bei ca. 1,5 kHz (F1) bzw. 4 kHz (F2) enden. Die Präzision von Konsonanten bestimmt sich dabei insbesondere aus einem Grad der Konzentration der akustischen Energie auf die entsprechenden Frequenzbereiche und eine hieraus resultierende Bestimmbarkeit der einzelnen Konsonanten.

[0021] Die Unterscheidbarkeit der einzelnen Bestandteile eines Sprachsignals, und damit die Möglichkeit, diese Bestandteile auflösen zu können, hängt jedoch nicht nur ab von artikulatorischen Aspekten. Während diese vorrangig die akustische Präzision der kleinsten isolierten Klangereignisse von Sprache, der sog. Phoneme, betreffen, bestimmten auch prosodische Aspekte die Sprachqualität, da hier durch Intonation und Akzentsetzung insbesondere über mehrere Segmente, also mehrere Phoneme oder Phonemgruppen hinweg, einer Aussage ein besonderer Sinn aufgeprägt werden kann, wie z.B. durch das Anheben der Tonhöhe am Satzende zum Verdeutlichen einer Frage, oder durch das Betonen einer konkreten Silbe in einem Wort zur Unterscheidung verschiedener Bedeutungen (vgl. "umfahren" vs. "umfahren") oder das Betonen eines Wortes zu seiner Hervorhebung. Insofern lässt sich eine Sprachqualität für ein Sprachsignal auch anhand prosodischer Eigenschaften, insbesondere wie den eben genannten, quantitativ erfassen, indem z.B. Maße für eine zeitliche Variation der Tonhöhe der Stimme, also ihrer Grundfrequenz, und für die Deutlichkeit einer Abhebung der Amplituden- und/oder Pegelmaxima bestimmt werden.

[0022] Anhand einer oder mehrerer der genannten und/oder weiterer, quantitativ erfassten artikulatorischen und/oder prosodischen Eigenschaften des Sprachsignals lässt sich somit das quantitative Maß für die Sprachqualität ableiten.

[0023] Erfindungsgemäß wird dabei als artikulatorische Eigenschaft des Sprachsignals eine mit der Präzision von vorgegebenen Formanten von Vokalen in dem Sprachsignal korrelierte Kenngröße, eine mit der Dominanz von Konsonanten, insbesondere Frikativen, in dem Sprachsignal korrelierte Kenngröße und/oder eine mit der Präzision der Übergänge von stimmhaften und stimmlosen Lauten korrelierte Kenngröße erfasst. Das quantitative Maß für die Sprachqualität kann dann jeweils unmittelbar durch die besagte erfasste Kenngröße gegeben sein, oder anhand dieser gebildet werden, z.B. durch Gewichtung zweier Kenngrößen für unterschiedliche Formanten o.ä., oder auch durch die Gewichtung, also durch eine gewichtete Mittelwertbildung, von wenigstens zwei verschiedenen der genannten Kenngrößen zueinander. Das quantitative Maß für die Sprachqualität bezieht sich dabei also auf die Sprachproduktion eines Sprechers, welcher von einer als "sauber" empfundenen Aussprache Defizite (wie z.B. Lispeln oder Nuscheln) bis hin zu Sprachfehlern aufweisen kann, welche die Sprachqualität entsprechend reduzieren.

[0024] Im Unterschied zu Größen, welche auf eine Propagation der Sprache in einer Umgebung bezogen sind, wie z.B. der Sprachverständlichkeitsindex ("Speech Intellegibility Index", SII), welcher bandweise die einzelnen Sprach- und Rauschanteile gewichtet, oder der Sprachübertragungsindex ("Speech Transmission Index", STI), welcher mittels eines die Modulationen de menschlichen Sprache nachbildenden Testsignals die Auswirkung ei-

nes Übertragungskanals auf die Modulationstiefe erfasst, ist hier das vorliegende Maß für die dabei insbesondere unabhängig von den externen Eigenschaften eines Übertragungskanals wie z.B. einer Propagation in einem möglicherweise nachhallenden Raum oder einer lauten Umgebung, sondern bevorzugt nur abhängig von den intrinsischen Eigenschaften der Spracherzeugung durch den Sprecher.

[0025] Dies bedeutet insbesondere, dass in leisen Umgebungen und/oder Umgebungen mit nur geringem Rauschhintergrund eine reduzierte Sprachqualität (bezogen auf einen Referenzwert, welcher bevorzugt für als "sehr gut" empfundenen Sprachqualität festgelegt wird) erkannt wird.

[0026] Im Rahmen der Erfindung wird dabei für eine Erfassung der mit der Dominanz von Konsonanten in dem Sprachsignal korrelierte Kenngröße eine in einem niedrigen Frequenzbereich beinhaltete erste Energie berechnet, eine in einem über dem niedrigen Frequenzbereich liegenden höheren Frequenzbereich beinhaltete zweite Energie berechnet, und die korrelierte Kenngröße anhand eines Verhältnisses und/oder eines über die jeweiligen Bandbreiten der genannten Frequenzbereiche gewichteten Verhältnisses der ersten Energie und der zweiten Energie gebildet. Insbesondere kann hierbei vorab eine zeitliche Glättung des Sprachsignals erfolgen. Für die Berechnung der ersten und der zweiten Energie kann insbesondere das Eingangs-Audiosignal in den niedrigen und den höheren Frequenzbereich aufgeteilt werden, z.B. mittels einer Filterbank und ggf. mittels einer entsprechenden Auswahl einzelner resultierender Frequenzbänder. Hierbei wird der niedere Frequenzbereich derart gewählt, dass er innerhalb des Frequenzintervalls [0 Hz, 2,5 kHz], besonders bevorzugt innerhalb des Frequenzintervalls [0 Hz, 2 kHz] liegt. Zudem wird der höhere Frequenzbereich derart gewählt, dass er innerhalb des Frequenzintervalls [3 kHz, 10 kHz], besonders bevorzugt innerhalb des Frequenzintervalls [4 Hz, 8 kHz] liegt.

[0027] Im Rahmen der Erfindung wird für eine Erfassung der mit der Präzision der Übergänge von stimmhaften und stimmlosen Lauten korrelierten Kenngröße anhand eine Korrelationsmessung und/oder anhand einer Nulldurchgangsrate des Eingangs-Audiosignals oder eines vom Eingangs-Audiosignal abgeleiteten Signals eine Unterscheidung von stimmhaften und stimmlosen Zeitsequenzen durchgeführt wird, wobei ein Übergang von einer stimmhaften Zeitsequenz zu einer stimmlosen Zeitsequenz oder von einer stimmlosen Zeitsequenz zu einer stimmhaften Zeitsequenz ermittelt wird, für wenigstens einen Frequenzbereich die vor dem Übergang in der stimmhaften bzw. stimmlosen Zeitsequenz enthaltene Energie ermittelt wird, und für den wenigstens einen Frequenzbereich die nach dem Übergang in der stimmlosen bzw. stimmhaften Zeitsequenz enthaltene Energie ermittelt wird, und die Kenngröße anhand der Energie vor dem Übergang und anhand der Energie nach dem Übergang ermittelt wird.

[0028] Dies bedeutet insbesondere: Es werden zunächst die stimmhaften und stimmlosen Zeitsequenzen des Sprachsignals im Eingangs-Audiosignal ermittelt, und hieraus ein Übergang von stimmhaft nach stimmlos oder von stimmlos nach stimmhaft identifiziert. Für wenigstens einen, insbesondere anhand empirischer Erkenntnisse für die Präzision der Übergänge vorgegebenen Frequenzbereich wird nun die Energie vor dem Übergang im Frequenzbereich für das Eingangs-Audiosignal oder für ein hieraus abgeleitetes Signal ermittelt. Diese Energie kann z.B. genommen werden über die stimmhafte bzw. stimmlose Zeitsequenz unmittelbar vor dem Übergang. Ebenso wird die Energie im betreffenden Frequenzbereich nach dem Übergang ermittelt, also z.B. über die dem Übergang nachfolgende stimmlose bzw. stimmhafte Zeitsequenz.

[0029] Anhand dieser beiden Energien kann nun ein Kennwert ermittelt werden, welcher insbesondere eine Aussage über eine Änderung der Energieverteilung am Übergang ermöglicht. Dieser Kennwert kann beispielsweise bestimmt werden als ein Quotient oder eine relative Abweichung der beiden Energien vor und nach dem Übergang. Der Kennwert kann aber auch gebildet werden als ein Vergleich der Energie vor bzw. nach dem Übergang mit der gesamten (breitbandigen) Signalenergie. Insbesondere können jedoch auch für einen weiteren Frequenzbereich jeweils vor und nach dem Übergang die Energien ermittelt werden, sodass der Kennwert zusätzlich anhand der Energien vor und nach dem Übergang im weiteren Frequenzband ermittelt werden kann, z.B. als eine Änderungsrate der Energieverteilung auf die beteiligten Frequenzbereiche über den Übergang hinweg (also einen Vergleich der Verteilung der Energien in beiden Frequenzbereichen vor dem Übergang mit der Verteilung nach dem Übergang).

[0030] Anhand des besagten Kennwertes kann dann die mit der Präzision der Übergänge korrelierte Kenngröße für das Maß der Sprachqualität ermittelt werden. Hierzu kann der Kennwert direkt verwendet werden, oder der Kennwert kann mit einem vorab für ein gute Artikulation insbesondere anhand entsprechender empirischer Kenntnisse ermittelten Referenzwert verglichen werden (z.B. als Quotient oder relative Abweichung). Die konkrete Ausgestaltung, insbesondere hinsichtlich der zu verwendenden Frequenzbereiche und Grenz- bzw. Referenzwertekann generell anhand empirischer Ergebnisse über eine entsprechende Aussagekraft der jeweiligen Frequenzbänder bzw. der Gruppen von Frequenzbändern erfolgen. Als der wenigstens eine Frequenzbereich können hierbei insbesondere die Frequenzbänder 13 bis 24, bevorzugt 16 bis 23 der Bark-Skala verwendet werden. Als ein weiterer Frequenzbereich kann insbesondere ein Frequenzbereich von niedrigeren Frequenzen verwendet werden.

[0031] Im Rahmen der Erfindung werden für eine Erfassung der mit der Präzision von vorgegebenen Formanten von Vokalen in dem Sprachsignal korrelierten Kenngrö-ße die in wenigsten zwei verschiedenen Formantenbereichen konzentrierten akustischen Energien

des Sprachsignals (oder mit besagten Energien korrelierte Grö-ßen) miteinander verglichen. Besonders bevorzugt wird ein Signalanteil des Sprachsignals in wenigstens einem Formantenbereich im Frequenzraum ermittelt, für den Signalanteil des Sprachsignals im wenigstens einen Formantenbereich eine mit dem Pegel korrelierte Signalgröße ermittelt wird, und die Kenngröße anhand eines Maximalwertes und/oder anhand einer Zeitstabilität der mit dem Pegel korrelierten Signalgröße ermittelt. Insbesondere kann hierbei als der wenigstens Formantenbereich der Frequenzbereich der ersten Formanten F1 (bevorzugt 250 Hz bis 1 kHz, besonders bevorzugt 300 Hz bis 750 Hz) oder der zweiten Formanten F2 (bevorzugt 500 Hz bis 3,5 kHz, besonders bevorzugt 600 Hz bis 2,5 kHz) gewählt werden, oder es werden zwei Formantenbereiche der ersten und zweiten Formanten gewählt. Insbesondere können auch mehrere, unterschiedliche Vokalen zugeordnete erste und/oder zweite Formantenbereiche (also die Frequenzbereiche, welche dem ersten bzw. zweiten Formanten des jeweiligen Vokals zugeordnet sind) gewählt werden. Für den oder die gewählten Formantenbereiche wird nun der Signalanteil ermittelt, und eine mit dem Pegel korrelierte Signalgröße des jeweiligen Signalanteils bestimmt. Die Signalgröße kann dabei durch den Pegel selbst, oder auch durch die ggf. geeignet geglättete maximale Signalamplitude gegeben sein. Anhand einer Zeitstabilität der Signalgröße, welche sich wiederum durch eine Varianz der Signalgröße über ein geeignetes Zeitfenster ermitteln lässt, und/oder anhand einer Abweichung der Signalgröße von ihrem Maximalwert über ein geeignetes Zeitfenster lässt sich nun eine Aussage über die Präzision von Formanten dahingehend treffen, dass eine geringe Varianz und geringe Abweichung vom Maximalpegel für einen artikulierten Laut (die Länge des Zeitfensters kann insbesondere abhängig von der Länge eines artikulierten Lautes gewählt werden) für eine hohe Präzision sprechen.

[0032] Vorteilhafterweise wird durch Analyse des Eingangs-Audiosignals mittels der Signalverarbeitung weiter mindestens eine prosodische Eigenschaft des Sprachsignals quantitativ erfasst wird, und das quantitative Maß für die Sprachqualität zusätzlich in Abhängigkeit von der mindestens einen prosodischen Eigenschaft des Sprachsignals (18) ermittelt. Bevorzugt wird dabei die Grundfrequenz des Sprachsignals zeitaufgelöst erfasst, und als prosodische Eigenschaft des Sprachsignals eine für die Zeitstabilität der Grundfrequenz charakteristische Kenngröße ermittelt. Diese Kenngröße kann z.B. anhand vor einer über die Zeit kumulierten relative Abweichung der Grundfrequenz ermittelt werden, oder über das Erfassen einer Anzahl an Maxima und Minima der Grundfrequenz über einen vorgegebenen Zeitraum. Die Zeitstabilität der Grundfrequenz ist v.a. für eine Monotonie der Sprachmelodie und -akzentuierung von Bedeutung, weswegen eine quantitative Erfassung auch eine Aussage über die Sprachqualität des Sprachsignals erlaubt.

[0033] Bevorzugt wird für das Sprachsignal, insbesondere durch eine entsprechende Analyse des Eingangs-Audiosignals oder eines hiervon abgeleiteten Signals, eine mit der Lautstärke korrelierte Größe, insbesondere eine Amplitude und/oder ein Pegel, zeitaufgelöst erfasst, wobei über einen vorgegebenen Zeitraum ein Quotient eines Maximalwertes der mit der Lautstärke korrelierten Größe zu einem über den vorgegebenen Zeitraum ermittelten Mittelwert der besagten Größe gebildet wird, und wobei als prosodische Eigenschaft des Sprachsignals eine Kenngröße in Abhängigkeit von besagtem Quotienten ermittelt wird, welcher aus dem Maximalwert und dem Mittelwert der mit der Lautstärke korrelierten Größe über den vorgegebenen Zeitraum gebildet wird. Auf diese Weise lässt sich anhand der mittelbar erfassten Lautstärkendynamik des Sprachsignals eine Aussage über eine Definition der Akzentuierung treffen.

[0034] In einer vorteilhaften Ausgestaltung werden anhand der Analyse des Eingangs-Audiosignals wenigstens zwei jeweils für artikulatorische und/oder prosodische Eigenschaften charakteristische Kenngrößen ermittelt, wobei das quantitative Maß für die Sprachqualität anhand von einem Produkt dieser Kenngrößen und/oder anhand von einem gewichteten Mittelwert und/oder eines Maximal- oder Minimalwertes dieser Kenngrößen gebildet wird. Dies ist insbesondere dann vorteilhaft, wenn ein einziges Maß für die Sprachqualität erfordert oder gewünscht ist, oder wenn ein einziges Maß, welches alle artikulatorischen oder alle prosodischen Eigenschaften erfassen soll, gewünscht ist.

[0035] Bevorzugt wird vor einem Erfassen der mindestens einen artikulatorische und/oder prosodischen Eigenschaft des Sprachsignals eine Sprachaktivität detektiert und/oder ein SNR im Eingangs-Audiosignal ermittelt, wobei eine Analyse hinsichtlich der mindestens einen artikulatorischen und/oder prosodischen Eigenschaft des Sprachsignals in Abhängigkeit der detektierten Sprachaktivität bzw. des ermittelten SNR durchgeführt wird. Hierdurch kann die Analyse der Sprachqualität des Sprachsignals auf diejenigen Fälle beschränkt werden, in welchen tatsächlich ein Sprachsignal vorliegt bzw. in welchen das SNR insbesondere oberhalb eines vorgegebenen Grenzwertes liegt, sodass davonausgegangen werden darf, dass eine hinreichend gute Erkennung der Signalanteile des Sprachsignals im Eingangs-Audiosignal überhaupt erst möglich ist, um eine entsprechende Bewertung vorzunehmen. Umgekehrt wird bei einer herkömmlichen Signalverarbeitung für ein hinreichend hohes SNR meist keine Maßnahme zur Hervorhebung o.ä. eines Sprachsignals getroffen, obwohl eine mangelhafte Sprachqualität, also bei schwacher Artikulation und/oder geringer Ausprägung prosodischer Merkmale wie Betonungen, von einer Verbesserung mittels der Signalverarbeitung profitieren würde.

[0036] Bevorzugt ist die Hörvorrichtung als ein Hörgerät ausgestaltet. Das Hörgerät kann dabei durch ein monaurales Gerät, oder durch ein binaurales Gerät mit zwei lokalen Geräten gegeben sein, welche vom Benutzer des Hörgerätes jeweils an seinem rechten bzw. linken Ohr

zu tragen sind. Insbesondere kann das Hörgerät zusätzlich zum genannten Eingangswandler auch noch mindestens einen weiteren akusto-elektrischen Eingangswandler aufweisen, welcher den Schall der Umgebung in ein entsprechendes weiteres Eingangs-Audiosignal umwandelt, sodass die quantitative Erfassung der mindestens einen artikulatorischen und/oder prosodischen Eigenschaft eines Sprachsignals durch eine Analyse einer Mehrzahl von beteiligten Eingangs-Audiosignalen erfolgen kann. Im Fall eines binauralen Gerätes können zwei der verwendeten Eingangs-Audiosignale jeweils in unterschiedlichen lokalen Einheiten des Hörgeräts (also jeweils am linken bzw. am rechten Ohr) erzeugt werden. Die Signalverarbeitungseinrichtung kann hierbei insbesondere Signalprozessoren beider lokaler Einheiten umfassen, wobei bevorzugt jeweils lokal erzeugte Maße für die Sprachqualität je nach betrachteter artikulatorischer und/oder prosodischer Eigenschaft in geeigneter Weise durch Mittelwertbildung oder einen Maximal- oder Minimalwert für beide lokalen Einheiten vereinheitlicht werden.

[0037] Nachfolgend wird ein Ausführungsbeispiel der Erfindung anhand einer Zeichnung näher erläutert. Hierbei zeigen jeweils schematisch:

Fig. 1 in einem Schaltbild ein Hörgerät, welches einen Schall mit einem Sprachsignal erfasst, und

Fig. 2 in einem Blockdiagramm ein Verfahren zum Ermitteln eines quantitativen Maßes für die Sprachqualität des Sprachsignals nach Fig. 1.

[0038] Einander entsprechende Teile und Größen sind in allen Figuren jeweils mit denselben Bezugszeichen versehen.

[0039] In Figur 1 ist schematisch in einem Schaltbild eine Hörvorrichtung 1 dargestellt, welche vorliegend als ein Hörgerät 2 ausgestaltet ist. Das Hörgerät 2 weist einen akusto-elektrischen Eingangswandler 4 auf, welcher dazu eingerichtet ist, einen Schall 6 der Umgebung des Hörgerätes 2 in ein Eingangs-Audiosignal 8 umzuwandeln. Eine Ausgestaltung des Hörgerätes 2 mit einem weiteren Eingangswandler (nicht dargestellt), welcher ein entsprechendes weiteres Eingangs-Audiosignal aus dem Schall 6 der Umgebung erzeugt, ist hierbei ebenso denkbar. Das Hörgerät 2 ist vorliegend als ein alleinstehendes, monaurales Gerät ausgebildet. Ebenso denkbar ist eine Ausgestaltung des Hörgerätes 2 als ein binaurales Hörgerät mit zwei lokalen Geräten (nicht dargestellt), welche vom Benutzer des Hörgerätes 2 jeweils an seinem rechten bzw. linken Ohr zu tragen sind.

[0040] Das Eingangs-Audiosignal 8 wird einer Signalverarbeitungseinrichtung 10 des Hörgerätes 2 zugeführt, in welcher das Eingangs-Audiosignal 8 insbesondere gemäß den audiologischen Anforderungen des Benutzers des Hörgerätes 2 entsprechend verarbeitet und dabei zum Beispiel frequenzbandweise verstärkt und/oder komprimiert wird. Die Signalverarbeitungseinrichtung 10 ist hierfür insbesondere mittels eines entsprechenden Signalprozessors (in Figur 1 nicht näher dargestellt) und eines über den Signalprozessor adressierbaren Arbeitsspeichers eingerichtet. Eine etwaige Vorverarbeitung des Eingangs-Audiosignals 8, wie z.B. eine A/D-Wandlung und/oder Vorverstärkung des erzeugten Eingangs-Audiosignals 8, soll hierbei als Teil des Eingangswandlers 4 betrachtet werden.

[0041] Die Signalverarbeitungseinrichtung 10 erzeugt hierbei durch die Verarbeitung des Eingangs-Audiosignals 8 ein Ausgangs-Audiosignal 12, welches mittels eines Elektro-akustischen Ausgangswandlers 14 in eine Ausgangsschallsignal 16 des Hörgerätes 2 umgewandelt wird. Der Eingangswandler 4 ist hierbei vorzugsweise gegeben durch ein Mikrofon, der Ausgangswandler 14 beispielsweise durch einen Lautsprecher (wie etwa einen Balanced Metal Case Receiver), kann aber auch durch einen Knochenleithörer o.ä. gegeben sein.

[0042] Der Schall 6 der Umgebung des Hörgerätes 2, welcher vom Eingangswandler 4 erfasst wird, beinhaltet unter anderem ein Sprachsignal 18 eines nicht näher dargestellten Sprechers, sowie weitere Schallanteile 20, welche insbesondere durch gerichtete und/oder diffuse Störgeräusche (Störschall bzw. Hintergrundrauschen) umfassen können, aber auch solche Geräusche beinhalten können, welche je nach Situation als ein Nutzsignal angesehen werden könnten, also beispielsweise Musik oder die Umgebung betreffende, akustische Warn- oder Hinweis-Signale.

[0043] Die in der Signalverarbeitungseinrichtung 10 zur Erzeugung des Ausgangs-Audiosignals 12 erfolgende Signalverarbeitung des Eingangs-Audiosignals 8 kann insbesondere eine Unterdrückung der Signalanteile umfassen, welche die im Schall 6 enthaltenen Störgeräusche unterdrücken, bzw. eine relative Anhebung der das Sprachsignal 18 repräsentierenden Signalanteile gegenüber den die weiteren Schallanteile 20 repräsentierenden Signalanteil. Insbesondere können hierbei auch eine frequenzabhängige oder breitbandige Dynamik-Kompression und/oder Verstärkung sowie Algorithmen zur Rauschunterdrückung angewandt werden.

[0044] Um die Signalanteile im Eingangs-Audiosignal 8, welche das Sprachsignal 18 repräsentieren, im Ausgangs-Audiosignal 12 möglichst gut hörbar zu machen, und dem Benutzer des Hörgerätes 2 im Ausgangsschall 16 dennoch einen möglichst natürlichen Höreindruck vermitteln zu können, soll in der Signalverarbeitungseinrichtung 10 zur Steuerung der auf das Eingangs-Audiosignal 8 anzuwendenden Algorithmen ein quantitatives Maß für die Sprachqualität des Sprachsignals 18 ermittelt werden. Dies ist anhand von Figur 2 beschrieben.

[0045] Figur 2 zeigt in einem Blockdiagramm eine Verarbeitung des Eingangs-Audiosignals 8 des Hörgerätes 2 nach Figur 2. Zunächst wird für das Eingangs-Audiosignal 8 eine Erkennung einer Sprachaktivität VAD durchgeführt. Liegt keine nennenswerte Sprachaktivität vor (Pfad "n"), so erfolgt die Signalverarbeitung des Eingangs-Audiosignals 8 zur Erzeugung des Ausgangs-Au-

diosignals 12 anhand eines ersten Algorithmus 25. Der erste Algorithmus 25 bewertet dabei in einer vorab vorgegebenen Weise Signalparameter des Eingangs-Audiosignals 8 wie z.B. Pegel, Rauschhintergrund, Transienten o.ä., breitbandig und/oder insbesondere frequenzbandweise, und ermittelt hieraus einzelne Parameter, z.B. frequenzbandweise Verstärkungsfaktoren und/oder Kompressions-Kenndaten (also v.a. Kniepunkt, Verhältnis, Attack, Release), welche auf das Eingangs-Audiosignal 8 anzuwenden sind.

[0046] Insbesondere kann der erste Algorithmus 25 auch eine Klassifizierung einer Hörsituation vorsehen, welche im Schall 6 realisiert ist, und in Abhängigkeit der Klassifizierung einzelne Parameter einstellen, ggf. als entsprechend für eine konkrete Hörsituation vorgesehenes Hörprogramm. Überdies können für den ersten Algorithmus 25 auch die individuellen audiologischen Anforderungen des Benutzers des Hörgerätes 2 berücksichtigt werden, um durch die Anwendung des ersten Algorithmus 25 auf das Eingangs-Audiosignal 8 eine Hörschwäche des Benutzers möglichst gut kompensieren zu können.

[0047] Wird jedoch bei der Erkennung einer Sprachaktivität VAD eine nennenswerte Sprachaktivität festgestellt (Pfad "y" der), so wird als nächstes ein SNR ermittelt, und mit einem vorgegebenen Grenzwert TH_{SNR} verglichen. Liegt das SNR nicht oberhalb des Grenzwertes, also $SNR \leq TH_{SNR}$, so wird auf das Eingangs-Audiosignal 8 zur Erzeugung des Ausgangs-Audiosignals 12 erneut der erste Algorithmus 25 angewandt. Liegt jedoch das SNR oberhalb des vorgegebenen Grenzwertes TH_{SNR}, also $SNR > TH_{SNR}$, so wird für die weitere Verarbeitung des Eingangs-Audiosignals 8 in nachfolgend beschriebener Weise ein quantitatives Maß 30 für die Sprachqualität des im Eingangs-Audiosignal 8 enthaltenen Sprachanteils 18 ermittelt. Hierfür werden artikulatorische und/oder prosodische Eigenschaften des Sprachsignals 18 quantitativ erfasst. Unter dem Begriff des im Eingangs-Audiosignal 8 enthaltenen Sprachsignalanteils 26 sind hierbei diejenigen Signalanteile des Eingangs-Audiosignals 8 zu verstehen, welche den Sprachanteil 18 des Schalls 6 repräsentieren, aus dem das Eingangs-Audiosignal 8 mittels des Eingangswandlers 4 erzeugt wird.

[0048] Zum Ermitteln des besagten quantitativen Maßes 30 wird das Eingangs-Audiosignal 8 in einzelne Signalpfade aufgeteilt.

[0049] Für einen ersten Signalpfad 32 des Eingangs-Audiosignals 8 wird zunächst eine Schwerpunktwellenlänge λc ermittelt, und mit einem vorgegebenen Grenzwert für die Schwerpunktwellenlänge Th_{λ} verglichen. Wird anhand des besagten Grenzwertes für die Schwerpunktwellenlänge Th_{λ} festgestellt, dass die Signalanteile im Eingangs-Audiosignal 8 hinreichend hochfrequent sind, so werden im ersten Signalpfad 32, ggf. nach einer geeignet zu wählenden zeitlichen Glättung (nicht dargestellt), für einen niedrigen Frequenzbereich NF und einen über dem niedrigen Frequenzbereich NF liegenden, höheren Frequenzbereich HF die Signalanteile ausgewählt. Eine mögliche Aufteilung kann beispielsweise derart sein, dass der niedrige Frequenzbereich NF alle Frequenzen $f_N \leq 2500Hz$, insbesondere $f_N \leq 2000$ Hz umfasst, und der höhere Frequenzbereich HF Frequenzen f_H mit $2500 \text{ Hz} < f_H \leq 10000$ Hz, insbesondere $4000 \text{ Hz} \leq f_H \leq 8000$ Hz oder $2500 \text{ Hz} < f_H \leq 5000$ Hz umfasst.

[0050] Die Auswahl kann unmittelbar im Eingangs-Audiosignal 8 durchgeführt werden, oder auch derart erfolgen, dass das Eingangs-Audiosignal 8 mittels einer Filterbank (nicht dargestellt) in einzelne Frequenzbänder aufgeteilt wird, wobei einzelne Frequenzbänder in Abhängigkeit der jeweiligen Bandgrenzen dem niedrigen oder höheren Frequenzbereich NF bzw. HF zugeordnet werden.

[0051] Anschließend werden für das im niedrigen Frequenzbereich NF enthaltene Signal eine erste Energie E1 und für das im höheren Frequenzbereich HF enthaltene Signal eine zweite Energie E2 ermittelt. Es wird nun ein Quotient QE aus der zweiten Energie als Zähler und der ersten Energie E1 als Nenner gebildet. Der Quotient QE kann nun bei geeignet gewähltem niederen und höheren Frequenzbereich NF, HF als eine Kenngröße 33 herangezogen werden, welche mit Dominanz von Konsonanten im Sprachsignal 18 korreliert ist. Die Kenngröße 33 ermöglicht somit eine Aussage über eine artikulatorische Eigenschaft der Sprachsignalanteile 26 im Eingangs-Audiosignal 8. So kann z.B. für einen Wert des Quotienten $QE >> 1$ (also $QE > Th_{QE}$ mit einem vorgegebenen, nicht näherdargestellten Grenzwert $Th_{QE} >> 1$) eine hohe Dominanz für Konsonanten gefolgert werden, während für einen Wert $QE < 1$ eine geringe Dominanz gefolgert werden kann.

[0052] In einem zweiten Signalpfad 34 wird im Eingangs-Audiosignal 8 anhand von Korrelationsmessungen und/oder anhand einer Nulldurchgangsrate des Eingangs-Audiosignals 8 eine Unterscheidung 36 in stimmhafte Zeitsequenzen V und stimmlose Zeitsequenzen UV durchgeführt. Anhand der stimmhaften und stimmlosen Zeitsequenzen V bzw. UV wird ein Übergang TS von einer stimmhaften Zeitsequenz V zu einer stimmlosen Zeitsequenz UV ermittelt. Die Länge einer stimmhaften oder stimmlosen Zeitsequenz kann z.B. zwischen 10 und 80 ms, insbesondere zwischen 20 und 50 ms betragen.

[0053] Es wird nun für wenigstens einen Frequenzbereich (z.B. eine als geeignet ermittelte Auswahl an besonders aussagekräftigen Frequenzbändern, z.B. die Frequenzbänder 16 bis 23 der Bark-Skala, oder die Frequenzbänder 1 bis 15 der Bark-Skala) jeweils ein Energie Ev für die stimmhafte Zeitsequenz V vor dem Übergang TS und eine Energie En für die stimmlose Zeitsequenz UV nach dem Übergang TS ermittelt. Insbesondere können hierbei auch für mehr als einen Frequenzbereich jeweils getrennt entsprechende Energien vor und nach dem Übergang TS ermittelt werden. Es wird nun bestimmt, wie sich die Energie am Übergang TS verändert, z.B. durch eine relative Änderung ΔE_{TS} oder durch einen Quotienten (nicht dargestellt) der Energien Ev, En vor

und nach dem Übergang TS.

[0054] Das Maß für die Änderung der Energie, also vorliegend die relative Änderung wird nun mit einem vorab für eine gute Artikulation ermittelten Grenzwert Th_E für Energieverteilung an Übergängen verglichen. Insbesondere kann eine Kenngröße 35 anhand eines Verhältnisses aus der relative Änderung ΔE_{TS} und dem besagten Grenzwert Th_E oder anhand einer relativen Abweichung der relative Änderung ΔE_{TS} vom diesem Grenzwert Th_E gebildet werden. Besagte Kenngröße 35 ist mit der Artikulation der Übergänge von stimmhaften und stimmlosen Lauten im Sprachsignal 18 korreliert ist, und ermöglicht somit einen Aufschluss über eine weitere artikulatorische Eigenschaft der Sprachsignalanteile 26 im Eingangs-Audiosignal 8. Generell gilt hierbei die Aussage, dass eine Übergang zwischen stimmhaften und stimmlosen Zeitsequenzen umso präziser artikuliert ist, je schneller, also zeitlich abgrenzbarer ein Wechsel der Energieverteilung über die für stimmhafte und stimmlose Laute relevanten Frequenzbereiche erfolgt.

[0055] Für die Kenngröße 35 kann jedoch auch eine Energieverteilung in zwei Frequenzbereichen (z.B. die oben genannten Frequenzbereichen gemäß der Bark-Skala, oder auch im niederen und höheren Frequenzbereich NF, HF) betrachtet werden, z.B. über einen Quotienten der jeweiligen Energien oder einen vergleichbaren Kennwert, und eine Veränderung des Quotienten bzw. des Kennwertes über den Übergang hinweg für die Kenngröße herangezogen werden. So kann z.B. eine Änderungsrate des Quotienten bzw. der Kenngröße bestimmt und mit einem vorab als geeignet ermittelten Referenzwert für die Änderungsrate verglichen werden.

[0056] Zur Bildung der Kenngröße 35 können auch Übergänge von stimmlosen Zeitsequenzen in analoger Weise betrachtet werden. Die konkrete Ausgestaltung, insbesondere hinsichtlich der zu verwendenden Frequenzbereiche und Grenz- bzw. Referenzwertekann generell anhand empirischer Ergebnisse über eine entsprechende Aussagekraft der jeweiligen Frequenzbänder bzw. der Gruppen von Frequenzbändern erfolgen.

[0057] In einem dritten Signalpfad 38 wird im Eingangs-Audiosignal 8 zeitauflöst eine Grundfrequenz f_G des Sprachsignalanteils 26 erfasst, und für besagte Grundfrequenz f_G eine Zeitstabilität 40 anhand einer Varianz der Grundfrequenz f_G ermittelt. Die Zeitstabilität 40 kann als eine Kenngröße 41 verwendet werden, welche eine Aussage über eine prosodische Eigenschaft der Sprachsignalanteile 26 im Eingangs-Audiosignal 8 ermöglicht. Eine stärkere Varianz in der Grundfrequenz f_G kann dabei als ein Indikator für eine bessere Sprachverständlichkeit herangezogen werden, während eine monotone Grundfrequenz f_G eine geringere Sprachverständlichkeit aufweist.

[0058] In einem vierten Signalpfad 42 wird für das Eingangs-Audiosignal 8 und/oder für den darin enthaltenen Sprachsignalanteil 26 zeitaufgelöst ein Pegel LVL erfasst, und über einen insbesondere anhand entsprechender empirischer Erkenntnisse vorgegebenen Zeitraum 44 ein zeitlicher Mittelwert MN_{LVL} gebildet. Des Weiteren wird über den Zeitraum 44 das Maximum MX_{LVL} des Pegels LVL ermittelt. Das Maximum MX_{LVL} des Pegels LVL wird nun durch den zeitlichen Mittelwert MN_{LVL} des Pegels LVL dividiert, und so eine mit einer Lautstärke des Sprachsignals18 korrelierte Kenngröße 45 ermittelt, welche eine weitere Aussage über eine prosodische Eigenschaft der Sprachsignalanteile 26 im Eingangs-Audiosignal 8 ermöglicht. Anstatt des Pegels LVL kann hierbei auch eine andere mit der Lautstärke und/oder dem Energieinhalt des Sprachsignalanteils 26 korrelierte Größe verwendet werden.

[0059] Die jeweils im ersten bis vierten Signalpfad 32, 34, 38, 42 wie beschrieben ermittelten Kenngrößen 33, 35, 41 bzw. 45 können nun jeweils einzeln als das quantitative Maß 30 für die Qualität des im Eingangs-Audiosignal 8 enthaltenen Sprachanteils 18 herangezogen werden, in dessen Abhängigkeit das Eingangs-Audiosignal nun ein zweiter Algorithmus 46 auf das Eingangs-Audiosignal 8 zur Signalverarbeitung angewandt wird. Der zweite Algorithmus 46 kann hierbei aus dem ersten Algorithmus 25 durch eine in Abhängigkeit des betreffenden quantitativen Maßes 30 erfolgende, entsprechende Veränderung eines oder mehrerer Parameter der Signalverarbeitung hervorgehen, oder ein gänzlich eigenständiges Hörprogramm vorsehen.

[0060] Insbesondere kann als quantitatives Maß 30 für die Sprachqualität auch ein einzelner Wert anhand der wie beschrieben ermittelten Kenngrößen 33, 35, 41 bzw. 45 bestimmt werden, z.B. durch einen gewichteten Mittelwert oder ein Produkt der Kenngrößen 33, 35, 41, 45 (in Fig. 2 schematisch durch das Zusammenführen der Kenngrößen 33, 35, 41, 45 dargestellt). Die Gewichtung der einzelnen Kenngrö-ßen kann hierbei insbesondere anhand von vorab empirisch ermittelten Gewichtungsfaktoren erfolgen, welche anhand einer Aussagekraft der durch die jeweilige Kenngröße erfasste artikulatorische bzw. prosodische Eigenschaft für die Sprachqualität bestimmt werden können.

[0061] Obwohl die Erfindung im Detail durch das bevorzugte Ausführungsbeispiel näher illustriert und beschrieben wurde, so ist die Erfindung nicht durch die offenbarten Beispiele eingeschränkt und andere Variationen können vom Fachmann hieraus abgeleitet werden, ohne den Schutzumfang der Erfindung zu verlassen.

Bezugszeichenliste

[0062]

1	Hörvorrichtung
2	Hörgerät
4	Eingangswandler
6	Schall der Umgebung
8	Eingangs-Audiosignal
10	Signalverarbeitungseinrichtung
12	Ausgangs-Audiosignal

14 Ausgangswandler
16 Ausgangsschall
18 Sprachsignal
20 Schallanteile
25 erster Algorithmus
26 Sprachsignalanteil
30 quantitatives Maß für Sprachqualität
32 erster Signalpfad
33 Kenngröße
34 zweiter Signalpfad
35 Kenngröße
36 Unterscheidung
38 dritter Signalpfad
40 Zeitstabilität
41 Kenngröße
42 vierter Signalpfad
44 Zeitraum
45 Kenngröße
46 zweiter Algorithmus

ΔE_{TS} relative Änderung (der Energie am Übergang)
λ_C Schwerpunktwellenlänge

E1 erste Energie
E2 zweite Energie
Ev Energie (vor dem Übergang)
En Energie (nach dem Übergang)

f_G Grundfrequenz
LVL Pegel
HF höherer Frequenzbereich
MN_{LVL} zeitlicher Mittelwert (des Pegels)
MX_{LVL} Maximum des Pegels
NF niedriger Frequenzbereich
QE Quotient
SNR Signal-zu-Rausch-Verhältnis (SNR)
Th_λ Grenzwert (für die Schwerpunktwellenlänge)
Th_E Grenzwert (für relative Änderung der Energie)
TH_{SNR} Grenzwert (für das SNR)
TS Übergang
V stimmhafte Zeitsequenz
VAD Erkennung einer Sprachaktivität
UV stimmlose Zeitsequenz

Patentansprüche

1. Verfahren zur Bewertung der Sprachqualität eines Sprachsignals (18) mittels einer Hörvorrichtung (1),

- wobei mittels eines akusto-elektrischen Eingangswandlers (4) der Hörvorrichtung (1) ein das Sprachsignal (18) enthaltender Schall (6) aus einer Umgebung der Hörvorrichtung (1) aufgenommen und in ein Eingangs-Audiosignal (8) umgewandelt wird,
- wobei durch Analyse des Eingangs-Audiosignals (8) mittels einer Signalverarbeitung mindestens eine artikulatorische Eigenschaft des Sprachsignals (18) quantitativ erfasst wird, und
- wobei in Abhängigkeit von der mindestens einen artikulatorischen Eigenschaft ein quantitatives Maß (30) für die Sprachqualität abgeleitet wird

wobei als artikulatorische Eigenschaft des Sprachsignals (18)

- eine mit der Präzision von vorgegebenen Formanten von Vokalen in dem Sprachsignal (18) korrelierte Kenngröße, und/oder
- eine mit der Dominanz von Konsonanten, insbesondere Frikativen, in dem Sprachsignal (18) korrelierte Kenngröße (31) und/oder
- eine mit der Präzision von Übergängen von stimmhaften und stimmlosen Lauten korrelierte Kenngröße (35)

erfasst wird,
dadurch gekennzeichnet,
dass für eine Erfassung der mit der Dominanz von Konsonanten in dem Sprachsignal (18) korrelierte Kenngröße (33)

- eine in einem niedrigen Frequenzbereich (NF) beinhaltete erste Energie (E1) berechnet wird, wobei der niedrige Frequenzbereich (NF) innerhalb des Frequenzintervalls [0 Hz, 2,5 kHz] gewählt wird,
- eine in einem über dem niedrigen Frequenzbereich (E2) liegenden höheren Frequenzbereich (HF) beinhaltete zweite Energie (E2) berechnet wird, wobei der höheren Frequenzbereich (HF) innerhalb des Frequenzintervalls [3 kHz, 10 kHz] gewählt wird,
- und die Kenngröße anhand eines Verhältnisses (QE) und/oder eines über die jeweiligen Bandbreiten der genannten Frequenzbereiche (NF, HF) gewichteten Verhältnisses der ersten Energie (E1) und der zweiten Energie (E2) gebildet wird, bzw.

dass für eine Erfassung der mit Präzision der Übergänge von stimmhaften und stimmlosen Lauten korrelierten Kenngröße (35)

- anhand eine Korrelationsmessung und/oder anhand einer Nulldurchgangsrate eine Unterscheidung (36) von stimmhaften Zeitsequenzen (V) und stimmlosen Zeitsequenzen (UV) durchgeführt wird,
- ein Übergang (TS) von einer stimmhaften Zeitsequenz (V) zu einer stimmlosen Zeitsequenz (UV) oder von einer stimmlosen Zeitsequenz (UV) zu einer stimmhaften

Zeitsequenz (V) ermittelt wird,
- für wenigstens einen Frequenzbereich die vor dem Übergang (TS) in der stimmhaften bzw. stimmlosen Zeitsequenz (V, UV) enthaltene Energie (Ev) ermittelt wird, und für den wenigstens einen Frequenzbereich die nach dem Übergang (TS) in der stimmlosen bzw. stimmhaften Zeitsequenz (UV, V) enthaltene Energie (En) ermittelt wird, und
- die Kenngröße (35) anhand der Energie (Ev) vor dem Übergang (TS) und anhand der Energie (En) nach dem Übergang (TS) ermittelt wird, bzw.

dass für eine Erfassung der mit der Präzision von vorgegebenen Formanten von Vokalen in dem Sprachsignal (18) korrelierten Kenngröße

- ein Signalanteil des Sprachsignals (18) in wenigstens einem Formantenbereich im Frequenzraum ermittelt wird,
- für den Signalanteil des Sprachsignals (18) im wenigstens einen Formantenbereich eine mit dem Pegel korrelierte Signalgröße ermittelt wird, und
- die Kenngröße anhand eines Maximalwertes und/oder anhand einer Zeitstabilität der mit dem Pegel korrelierten Signalgröße ermittelt wird.

2. Verfahren nach Anspruch 1,

- wobei durch Analyse des Eingangs-Audiosignals (8) mittels der Signalverarbeitung weiter mindestens eine prosodische Eigenschaft des Sprachsignals (18) quantitativ erfasst wird, und
- wobei das quantitative Maß (30) für die Sprachqualität zusätzlich in Abhängigkeit von der mindestens einen prosodischen Eigenschaft des Sprachsignals (18) ermittelt wird.

3. Verfahren nach Anspruch 2,

wobei die Grundfrequenz (f_G) des Sprachsignals (18) zeitaufgelöst erfasst wird, und
wobei als prosodische Eigenschaft des Sprachsignals (18) eine für die Zeitstabilität (40) der Grundfrequenz (f_G) charakteristische Kenngröße (41) ermittelt wird.

4. Verfahren nach Anspruch 2 oder Anspruch 3,

wobei für das Sprachsignal (18) eine mit der Lautstärke korrelierte Größe (LVL) zeitaufgelöst erfasst wird,
wobei über einen vorgegebenen Zeitraum (44) ein Quotient eines Maximalwertes (MX_{LVL}) der mit der Lautstärke korrelierten Größe (LVL) zu einem über den vorgegebenen Zeitraum (44) ermittelten Mittelwert (MN_{LVL}) der besagten Größe (LVL) gebildet wird, und
wobei als prosodische Eigenschaft des Sprachsignals (18) eine Kenngröße (45) in Abhängigkeit von besagtem Quotienten ermittelt wird, welcher aus dem Maximalwert (MX_{LVL}) und dem Mittelwert (MN_{LVL}) der mit der Lautstärke korrelierten Größe (VL) über den vorgegebenen Zeitraum (44) gebildet wird.

5. Verfahren nach einem der vorhergehenden Ansprüche,

wobei anhand der Analyse des Eingangs-Audiosignals (18) wenigstens zwei jeweils für artikulatorische und/oder prosodische Eigenschaften charakteristische Kenngrößen (33, 35, 41, 45) ermittelt werden, und
wobei das quantitative Maß (30) für die Sprachqualität anhand von einem Produkt dieser Kenngrößen (33, 35, 41, 45) und/oder anhand von einem gewichteten Mittelwert dieser Kenngrößen (33, 35, 41, 45) gebildet wird.

6. Verfahren nach einem der vorhergehenden Ansprüche,

wobei vor einem Erfassen der mindestens einen artikulatorischen und/oder prosodischen Eigenschaft des Sprachsignals eine Sprachaktivität (VAD) detektiert und/oder ein Signal-zu-Rausch-Verhältnis (SNR) im Eingangs-Audiosignal (18) ermittelt wird, und
wobei eine Analyse hinsichtlich der mindestens einen artikulatorischen und/oder prosodischen Eigenschaft des Sprachsignals (18) in Abhängigkeit der detektierten Sprachaktivität (VAD) bzw. des ermittelten Signal-zu-Rausch-Verhältnisses (SNR) durchgeführt wird.

7. Hörvorrichtung (1), umfassend:

- einen akusto-elektrischen Eingangswandler (4), welcher dazu eingerichtet ist, einen Schall (6) aus einer Umgebung der Hörvorrichtung (1) aufzunehmen und in ein Eingangs-Audiosignal (8) umzuwandeln, und
- eine Signalverarbeitungseinrichtung (10), welche dazu eingerichtet ist, anhand einer Analyse des Eingangs-Audiosignals (8) mindestens eine artikulatorische Eigenschaft eines im Eingangs-Audiosignal (8) enthaltenen Anteils eines Sprachsignals (18) quantitativ zu erfassen und in Abhängigkeit von der mindestens einen artikulatorischen Eigenschaft ein quantitatives Maß (30) für die Sprachqualität gemäß dem Verfahren nach einem der vorhergehenden Ansprüche

abzuleiten.

8. Hörvorrichtung (1) nach Anspruch 7, ausgestaltet als ein Hörgerät (2).

Claims

1. Method for rating the speech quality of a speech signal (18) by way of a hearing device (1),

- wherein an acousto-electric input transducer (4) of the hearing device (1) records a sound (6) containing the speech signal (18) from surroundings of the hearing device (1) and converts it into an input audio signal (8),
- wherein at least one articulatory property of the speech signal (18) is quantitatively acquired through analysis of the input audio signal (8) by way of a signal processing operation, and
- wherein a quantitative measure (30) of the speech quality is derived on the basis of the at least one articulatory property

wherein

- a characteristic variable correlated with the precision of predefined formants of vowels in the speech signal (18), and/or
- a characteristic variable (31) correlated with the dominance of consonants, in particular fricatives, in the speech signal (18), and/or
- a characteristic variable (35) correlated with the precision of transitions from voiced and unvoiced sounds is acquired as articulatory property of the speech signal (18),

characterized

in that, in order to acquire the characteristic variable (33) correlated with the dominance of consonants in the speech signal (18),

- a first energy (E1) contained in a low frequency range (NF) is calculated, wherein the low frequency range (NF) is selected within the frequency interval [0 Hz, 2.5 kHz],
- a second energy (E2) contained in a frequency range (HF) higher than the low frequency range (E2) is calculated, wherein the higher frequency range (HF) is selected within the frequency interval [3 Hz, 10 kHz],
- and the characteristic variable is formed based on a ratio (QE), and/or a ratio weighted over the respective bandwidths of said frequency ranges (NF, HF), of the first energy (E1) and the second energy (E2), or

in that, in order to acquire the characteristic variable (35) correlated with precision of the transitions from voiced and unvoiced sounds,

- a distinction (36) is made between voiced temporal sequences (V) and unvoiced temporal sequences (UV) based on a correlation measurement and/or based on a zero crossing rate,
- a transition (TS) from a voiced temporal sequence (V) to an unvoiced temporal sequence (UV) or from an unvoiced temporal sequence (UV) to a voiced temporal sequence (V) is ascertained,
- the energy (Ev) contained in the voiced or unvoiced temporal sequence (V, UV) prior to the transition (TS) is ascertained for at least one frequency range, and the energy (En) contained in the unvoiced or voiced temporal sequence (UV, V) following the transition (TS) is ascertained for the at least one frequency range, and
- the characteristic variable (35) is ascertained based on the energy (Ev) prior to the transition (TS) and based on the energy (En) following the transition (TS), or

in that, in order to acquire the characteristic variable correlated with the precision of predefined formants of vowels in the speech signal (18),

- a signal component of the speech signal (18) in at least one formant range in the frequency space is ascertained,
- a signal variable correlated with the level is ascertained for the signal component of the speech signal (18) in the at least one formant range, and
- the characteristic variable is ascertained based on a maximum value and/or based on a temporal stability of the signal variable correlated with the level.

2. Method according to Claim 1,

- wherein furthermore at least one prosodic property of the speech signal (18) is quantitatively acquired through analysis of the input audio signal (8) by way of the signal processing operation, and
- wherein the quantitative measure (30) of the speech quality is additionally ascertained on the basis of the at least one prosodic property of the speech signal (18).

3. Method according to Claim 2,

wherein the fundamental frequency (f_G) of the speech signal (18) is acquired in a temporally resolved manner, and

wherein a characteristic variable (41) characteristic of the temporal stability (40) of the fundamental frequency (f_G) is ascertained as prosodic property of the speech signal (18).

4. Method according to Claim 2 or Claim 3,

wherein a variable (LVL) correlated with the volume is acquired in a temporally resolved manner for the speech signal (18),
wherein a quotient of a maximum value (MX_{LVL}) of the variable (LVL) correlated with the volume to a mean (MN_{LVL}) of said variable (LVL), ascertained over a predefined time interval (44), is formed over the predefined time interval (44), and
wherein a characteristic variable (45) is ascertained as prosodic property of the speech signal (18) on the basis of said quotient that is formed from the maximum value (MX_{LVL}) and the mean (MN_{LVL}) of the variable (VL) correlated with the volume over the predefined time interval (44).

5. Method according to one of the preceding claims,

wherein at least two characteristic variables (33, 35, 41, 45) each characteristic of articulatory and/or prosodic properties are ascertained based on the analysis of the input audio signal (18), and
wherein the quantitative measure (30) of the speech quality is formed based on a product of these characteristic variables (33, 35, 41, 45) and/or based on a weighted mean of these characteristic variables (33, 35, 41, 45).

6. Method according to one of the preceding claims,

wherein speech activity (VAD) is detected and/or a signal-to-noise ratio (SNR) in the input audio signal (18) is ascertained before the at least one articulatory and/or prosodic property of the speech signal is acquired, and
wherein analysis is performed with regard to the at least one articulatory and/or prosodic property of the speech signal (18) on the basis of the detected voice activity (VAD) or the ascertained signal-to-noise ratio (SNR).

7. Hearing device (1) comprising:

- an acousto-electric input transducer (4) that is designed to record a sound (6) from surroundings of the hearing device (1) and to convert it into an input audio signal (8), and
- a signal processing apparatus (10) that is designed to quantitatively acquire at least one articulatory property of a component, contained in the input audio signal (8), of a speech signal (18) based on analysis of the input audio signal (8) and to derive a quantitative measure (30) of the speech quality on the basis of the at least one articulatory property according to the method according to any of the preceding claims.

8. Hearing device (1) according to Claim 7, designed as a hearing aid (2).

Revendications

1. Procédé d'évaluation de la qualité vocale d'un signal vocal (18) à l'aide d'un dispositif auditif (1),

- un son (6) contenant le signal vocal (18) provenant d'un environnement du dispositif auditif (1) étant enregistré et converti en un signal audio d'entrée (8) au moyen d'un convertisseur d'entrée acousto-électrique (4) du dispositif auditif (1),
- une analyse du signal audio d'entrée (8) à l'aide d'un traitement de signal permettant de détecter quantitativement au moins une propriété articulatoire du signal vocal (18), et
- une mesure quantitative (30) reflétant la qualité vocale étant dérivée en fonction de l'au moins une propriété articulatoire,
- une grandeur corrélée à la précision de formantes spécifiées de voyelles dans le signal vocal (18), et/ou
- une grandeur (31) corrélée à la dominance de consonnes, notamment fricatives, dans le signal vocal (18) et/ou
- une grandeur corrélée à la précision de transitions entre des sons voisés et non voisés (35) étant détectée comme propriété articulatoire du signal vocal (18),

caractérisé en ce que

pour détecter une grandeur (33) corrélée à la dominance de consonnes dans le signal vocal (18)

- une première énergie (E1) contenue dans une gamme de fréquences basses (NF) est calculée, la gamme de fréquences basses (NF) étant sélectionnée dans l'intervalle de fréquence [0 Hz, 2,5 kHz],
- une deuxième énergie (E2) contenue dans une gamme de fréquences plus élevées (HF) au-dessus de la gamme de fréquences basses (E2) est calculée, la gamme de fréquences plus élevées (HF) étant sélectionnée dans l'intervalle de fréquence [3 kHz, 10 kHz],

- et la grandeur est formée sur la base d'un rapport (QE) et/ou d'un rapport de la première énergie (E1) et de la deuxième énergie (E2) pondéré sur les bandes passantes respectives des gammes de fréquences mentionnées (NF, HF), ou

pour détecter la grandeur corrélée à la précision de transitions entre des sons voisés et non voisés (35)

- une distinction (36) entre les séquences temporelles voisées (V) et les séquences temporelles non voisées (UV) est effectuée sur la base d'une mesure de corrélation et/ou d'un taux de passage par zéro,
- une transition (TS) d'une séquence temporelle voisée (V) à une séquence temporelle non voisée (UV) ou d'une séquence temporelle non voisée (UV) à une séquence temporelle voisée (V) est déterminée,
- l'énergie (Ev) contenue avant la transition (TS) dans la séquence temporelle voisée ou non voisée (V, UV) est déterminée pour au moins une gamme de fréquences et l'énergie (En) contenue après la transition (TS) dans la séquence temporelle voisée ou non voisée (UV, V) est déterminée pour l'au moins une gamme de fréquences, et
- la grandeur (35) est déterminée sur la base de l'énergie (Ev) avant la transition (TS) et sur la base de l'énergie (En) après la transition (TS), ou

pour détecter la grandeur corrélée à la précision de formantes spécifiées de voyelles dans le signal vocal (18),

- une composante du signal vocal (18) est déterminée dans au moins une gamme de formantes dans l'espace fréquentiel,
- une dimension de signal corrélée au niveau est déterminée pour la composante du signal vocal (18) dans au moins une gamme de formantes, et
- la grandeur est déterminée sur la base d'une valeur maximale et/ou sur la base d'une stabilité temporelle de la dimension du signal corrélée au niveau.

2. Procédé selon la revendication 1,

- une analyse du signal audio d'entrée (8) à l'aide d'un traitement de signal permettant en outre de détecter quantitativement au moins une propriété prosodique du signal vocal (18), et
- la mesure quantitative (30) reflétant la qualité vocale étant en outre déterminée en fonction de l'au moins une propriété prosodique du signal vocal (18).

3. Procédé selon la revendication 2,

la fréquence fondamentale (f_G) du signal vocal (18) étant détectée de manière résolue dans le temps, et
une grandeur (41) caractéristique de la stabilité temporelle (40) de la fréquence fondamentale (f_G) étant déterminée comme propriété prosodique du signal vocal (18) .

4. Procédé selon la revendication 2 ou la revendication 3,

une amplitude (LVL) corrélée au volume sonore étant détectée de manière résolue dans le temps pour le signal vocal (18),
sur une période spécifiée (44), un quotient d'une valeur maximale (MX_{LVL}) de l'amplitude (LVL) corrélée au volume sonore à une valeur moyenne (MN_{LVL}) de ladite amplitude (LVL) déterminée sur la période spécifiée (44) étant formé, et
une grandeur (45) étant déterminée comme propriété prosodique du signal vocal (18) en fonction dudit quotient qui est formé de la valeur maximale (MX_{LVL}) et de la valeur moyenne (MN_{LVL}) de l'amplitude (VL) corrélée au volume sonore sur la période spécifiée (44) .

5. Procédé selon l'une des revendications précédentes,

au moins deux grandeurs (33, 35, 41, 45) caractéristiques de propriétés articulatoires et/ou prosodiques, étant déterminées sur la base de l'analyse du signal audio d'entrée (18), et
la mesure quantitative (30) reflétant la qualité vocale étant formée sur la base d'un produit de ces grandeurs (33, 35, 41, 45) et/ou sur la base d'une valeur moyenne pondérée de ces grandeurs (33, 35, 41, 45).

6. Procédé selon l'une des revendications précédentes,

avant de détecter l'au moins une propriété articulatoire et/ou prosodique du signal vocal, une activité vocale (VAD) étant détectée et/ou un rapport signal sur bruit (SNR) étant déterminé dans le signal audio d'entrée (18), et
une analyse concernant au moins une propriété articulatoire et/ou prosodique du signal vocal (18) étant effectuée en fonction de l'activité vocale détectée (VAD) ou du rapport signal sur bruit (SNR) déterminé.

7. Dispositif auditif (1), comprenant :

- un transducteur d'entrée acousto-électrique (4), qui est conçu pour acquérir un son (6) provenant d'un environnement du dispositif auditif (1) et le convertir en un signal audio d'entrée (8), et
- un module de traitement de signal (10) qui est conçu pour détecter quantitativement au moins une propriété articulatoire d'une composante d'un signal vocal (18) contenue dans le signal audio d'entrée (8) sur la base d'une analyse du signal audio d'entrée (8) et pour déduire, en fonction d'au moins une propriété articulatoire, une mesure quantitative (30) reflétant la qualité vocale selon le procédé selon l'une des revendications précédentes.

8. Dispositif auditif (1) selon la revendication 7, conçu comme une aide auditive (2).

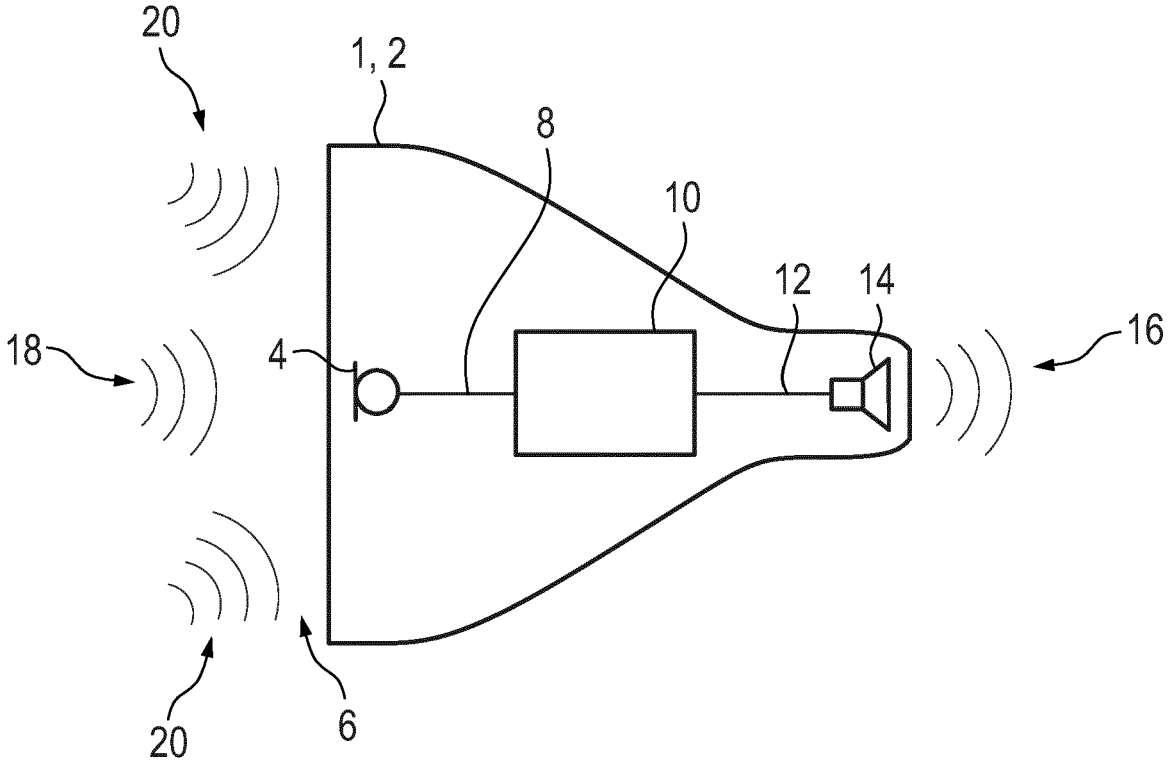

Fig. 1

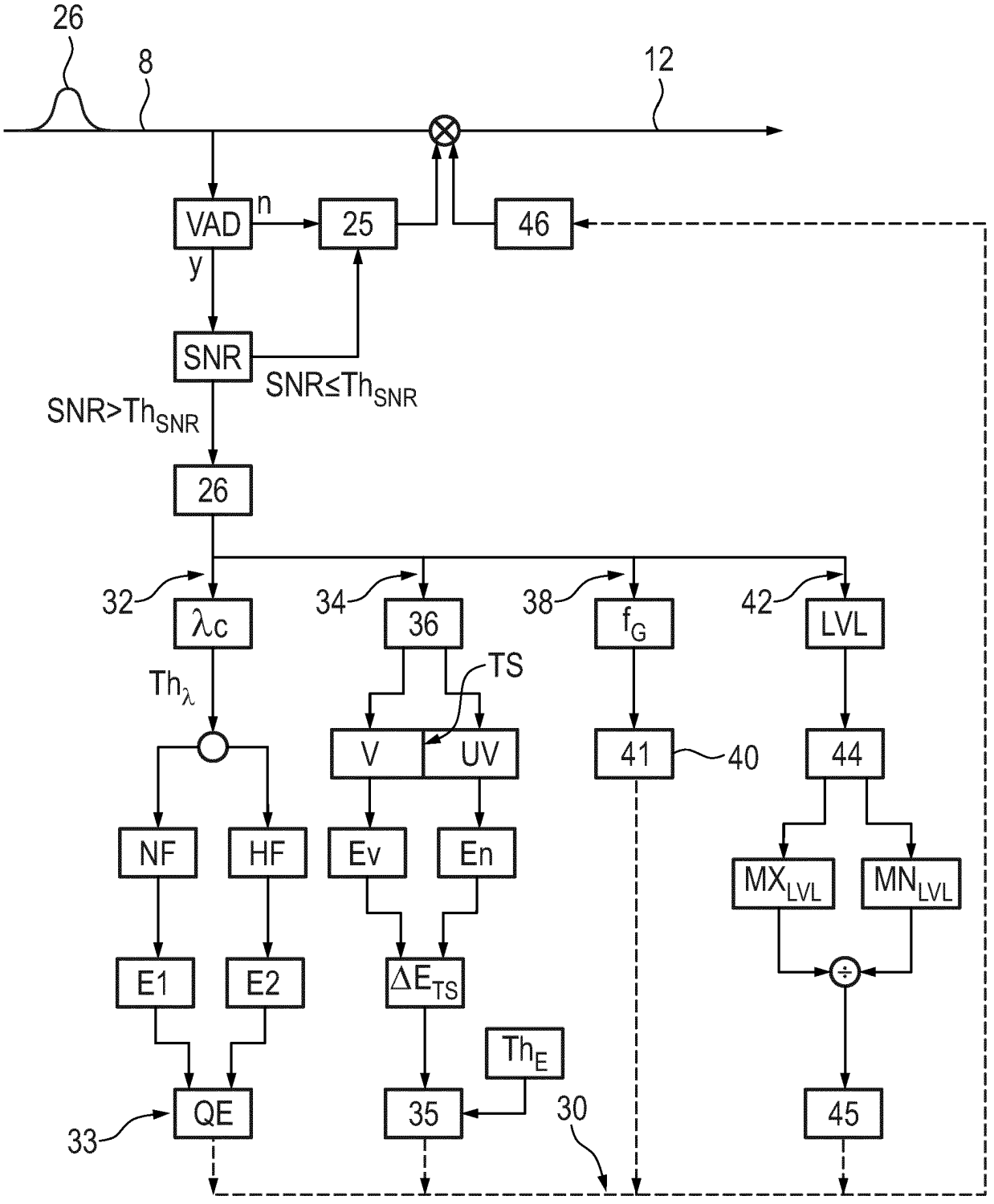
26
8
12
VAD
n
25
46
y
SNR
SNR≤Th_SNR
SNR>Th_SNR
26
32
λc
Th_λ
34
36
TS
V
UV
38
f_G
41
40
42
LVL
44
NF
HF
Ev
En
MX_LVL
MN_LVL
E1
E2
ΔE_TS
Th_E
÷
33
QE
35
30
45

Fig. 2

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 Patentdokumente

- US 20040167774 A1 **[0005]**
- US 20180255406 A1 **[0007]**
- US 7165025 B **[0008]**

In der Beschreibung aufgeführte Nicht-Patentliteratur

- Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks. **A. H. ANDERSEN et al.** IEEE/ACM Transactions on Audio, Speech and Language Processing. IEEE, 01. Oktober 2018, vol. 26, 1925-1939 **[0006]**