(19) Deutsches Patent- und Markenamt

(10) **DE 10 2013 022 596 B3** 2020.02.27

(12) # Patentschrift

(21) Aktenzeichen: **10 2013 022 596.5**
(22) Anmeldetag: **25.01.2013**
(45) Veröffentlichungstag der Patenterteilung: **27.02.2020**

(51) Int Cl.: ***G10L 15/22*** (2006.01)
G10L 15/26 (2006.01)

Innerhalb von neun Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten (§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 1 Patentkostengesetz).

(62) Teilung aus:
10 2013 001 219.8

(73) Patentinhaber:
inodyn NewMedia GmbH, 69151 Neckargemünd, DE

(72) Erfinder:
Pantel, Lothar, 69151 Neckargemünd, DE

(56) Ermittelter Stand der Technik:
WO **2014/ 093 238** **A1**

(54) Bezeichnung: **Verfahren und System zur Sprachaktivierung mit Aktivierungswort am Satzanfang, innerhalb des Satzes oder am Satzende**

(57) Zusammenfassung: Die Erfindung betrifft ein Verfahren und ein System zur Sprachaktivierung eines Software-Agenten aus einem Standby-Modus. Audio-Daten (11) werden in einem Audio-Puffer (6) zwischengespeichert, so dass der Audio-Puffer (6) stets die Audio-Daten (11) der jüngsten Vergangenheit enthält. Gleichzeitig werden die Audio-Daten (11) einer sekundären Spracherkennung (7) zugeführt, die z.B. energetisch sparsam ist. Beim Erkennen eines Aktivierungsworts (18) durch die sekundäre Spracherkennung (7), wandelt ein primärer Spracherkennungs-Prozess (8) den Inhalt des Audio-Puffers (6) in Text (13) um, und zwar ab dem Satzanfang, der im Audio-Puffer (6) anhand einer Sprechpause (16) identifiziert wird. Der Text (13) wird einem Dialogsystem (9) zuführt. Das beschriebene Verfahren und System ist in der Lage, ein Aktivierungswort (18) nicht nur am Satzanfang zu erkennen, sondern auch innerhalb des Satzes und insbesondere am Satzende.

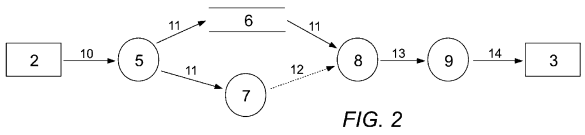
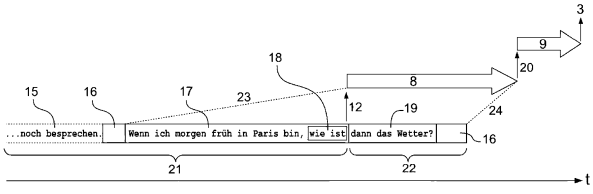


FIG. 2



Beschreibung

Technisches Gebiet

[0001] Die Erfindung betrifft das Gebiet der Spracherkennung, insbesondere die Aktivierung von Vorgängen per Sprache.

Stand der Technik

[0002] Die Spracherkennung, also das Umwandeln von akustischen Sprachsignalen in Text, konkret, das Umwandeln in eine digitale Text-Darstellung mittels einer Zeichenkodierung, ist bekannt. Es ist möglich, Systeme ohne haptische Bedienung zu steuern. Die Verfahren und Systeme der Patente US 8,260,618 B2 und US 7,953,599 B2 und der Offenlegungsschriften US2013/0289994 A1 und US2014/0163978 A1 beschreiben, wie sich Geräte per Sprache steuern und auch aktiveren lassen.

[0003] Smartphones (Mobiltelefone mit Computer-Funktionalität) haben aufgrund ihrer geringen Größe eine stark eingeschränkte Ergonomie, wenn sie per Touchscreen bedient werden. Eine Alternative sind digitale Sprachassistenten, bei denen das Smartphone mit Sprachkommandos gesteuert werden kann, zum Teil auch mit natürlicher Sprache ohne spezielle Steuerbefehle. Ein bekanntes Beispiel ist das System „Siri“ auf dem Smartphone „iPhone“ vom Hersteller-Apple (Fundstelle: http://www.apple.com).

[0004] Ein Sprachassistent kann eine eigenständige App auf dem Smartphone sein oder in das Betriebssystem integriert sein. Die Spracherkennung, Auswertung und Reaktion kann lokal auf der Hardware des Smartphones erfolgen. In der Regel wird aber wegen der größeren Rechenleistung ein Server-Verbund im Internet („in the Cloud“) verwendet, mit dem der digitale Sprachassistent kommuniziert. D.h. es werden komprimierte Sprach- bzw. Tonaufnahmen an den Server bzw. Server-Verbund geschickt, und die per Sprachsynthese generierte verbale Antwort wird zurück auf das Smartphone gestreamt.

[0005] Digitale Sprachassistent-Systeme sind eine Teilmenge der Software-Agenten. Man kann unterscheiden zwischen verschiedenen Interaktionsmöglichkeiten: z.B. das Abfragen von Fakten oder Wissen, das Abfragen von Status-Updates in Sozialen Netzwerken oder das Diktieren von E-Mails. In den meisten Fällen kommt auf der Seite des digitalen Sprachassistenten ein Dialogsystem (bzw. ein sogenannter Chatbot) zum Einsatz, welches zum Teil mit semantischer Analyse oder mit Ansätzen von Künstlicher Intelligenz ein realitätsnahes Gespräch zu einem Thema simuliert.

[0006] Ein weiteres Beispiel für einen digitalen Sprachassistenten ist das als „S Voice“ bezeichnete System auf dem Smartphone „Galaxy S III“ des Herstellers Samsung (Fundstelle: http://www.samsung.com). Dieses Produkt verfügt über die Möglichkeit, das Smartphone aus einem Standby- bzw. Schlafzustand zu wecken, und zwar per Sprachbefehl, ohne einen Touchscreen zu berühren oder eine Taste zu drücken. Dazu kann der Benutzer in den Systemeinstellungen eine gesprochene Phrase hinterlegen, die zum Aufwecken dient. Werkseitig voreingestellt ist „Hi Galaxy“. Der Benutzer muss die akustische Überwachung explizit freischalten und später wieder deaktivieren, da der Stromverbrauch für einen tagelangen Betrieb zu hoch wäre. Laut Hersteller ist das System für Situationen vorgesehen, in denen eine Aktivierung per Hand keine Option ist, z.B. beim Autofahren. In diesem Fall gibt der Autofahrer das verbale Kommando „Hi Galaxy“, worauf hin sich je nach Einstellung „S Voice“ z.B. mit der Begrüßung meldet: „What would you like to do?“. Erst jetzt, in einem zweiten Schritt und nachdem der Benutzer bereits unproduktiv Zeit durch sein erstes Kommando und durch das Abwarten der benötigten Zeit zum Aufwecken sowie durch den Begrüßungsspruch verloren hat, kann er seine eigentliche Frage stellen, z.B. „Wie ist das Wetter in Paris?“

[0007] In der Systemsteuerung des Smartphones „Galaxy S III“ ist es möglich, eine stark begrenzten Anzahl von weiteren Phrasen zu hinterlegen, mit denen dann im Anschluss sehr einfache Aktionen ausgelöst werden können. Durch das Kommando „Foto machen“ könnte z.B. die Kamera-App gestartet werden. Es ist jedoch nicht möglich, dem Smartphone bzw. „S Voice“ komplexe Fragen zu stellen oder das Smartphone zu komplexen Aktionen aufzufordern, solange sich das System im Standby- bzw. Schlafzustand befindet. Eine Frage, wie z.B. „Brauche ich übermorgen eine Regenjacke in Paris?“, kann von dem System - trotz akustischer Überwachung - aus dem Standby- bzw. Schlafzustand heraus nicht beantwortet werden. Dazu muss das Smartphone zuvor explizit aufgeweckt werden.

[0008] Die vom Smartphone „Galaxy S III“ genutzte Sprachaktivierungs-Technologie stammt vom Hersteller Sensory Inc. (Fundstelle: http://www.sensoryinc.com). Der Hersteller betont die extrem niedrige Falsch-Positiv-Rate bei der akustischen Überwachung mittels seiner „TrulyHandsfree“ Technologie. „Falsch-Positiv“ bezieht sich auf das fälschliche Interpretieren von anderen Geräuschen als Erkennungs-Phrase und ein daraus resultierendes unerwünschtes Auslösen des Triggers. In seinen Beschreibungen beschränkt sich der Hersteller auf einen sequentiellen Ablauf, bei dem das Gerät zunächst per Schlüsselwort aufgeweckt wird und erst dann über weitere Kommandos gesteuert werden kann. Zitat: „TrulyHandsfree can be alwayson and listening for dozens of keywords that will bring the device to life to be controlled via further voice comman-

ds.“ Eine andere, davon abweichende Vorgehensweise wird nicht offenbart.

[0009] Die nachveröffentlichte Patentanmeldung WO 2014/093238 A1 beschreibt ein System, in dem Audio-Daten in einem „Memory Buffer“-Modul zwischengespeichert werden und bei Erkennung eines Aktivierungsworts mittels eines „Speech Detection“-Moduls und eines „Speech Processing“-Moduls einem „Speech Recognition“-Server zugeführt werden.

Darstellung der Erfindung

[0010] Der vorliegenden Erfindung liegt die Aufgabe zu Grunde, ein Verfahren zu schaffen, mit dem es möglich ist, einem Software-Agenten oder einem digitalen Sprachassistenten, der sich in einem Standby- bzw. Schlafzustand befindet, per „natürlicher“ Sprache Fragen zu stellen oder auch Mitteilungen und Aufforderungen.

[0011] Erfindungsgemäß wird die voranstehende Aufgabe mit den Merkmalen aus den unabhängigen Ansprüchen 1 und 8 gelöst. Vorteilhafte Ausgestaltungen, mögliche Alternativen und optionale Funktionalitäten sind in den Unteransprüchen angegeben.

[0012] Ein Software-Agent bzw. ein digitaler Sprachassistent befindet sich in einem stromsparenden Standby-Modus bzw. Schlafzustand, wobei die von einem oder mehreren Mikrofonen aufgenommenen Umgebungsgeräusche - beispielsweise Sprache - digitalisiert und kontinuierlich in einem Audio-Puffer zwischengespeichert werden, so dass der Audio-Puffer stets die Umgebungsgeräusche (bzw. die Sprache) der jüngsten Vergangenheit enthält, beispielsweise jene der letzten 30 Sekunden.

[0013] Außerdem werden die von dem Mikrofon (oder den Mikrofonen) aufgenommenen digitalisierten Umgebungsgeräusche (bzw. die Sprache) ohne nennenswerte Verzögerung einem energiesparenden, sekundären Spracherkennungs-Prozess zugeführt, welcher beim Erkennen eines Schlüsselworts oder einer Phrase einen primären Spracherkennungs-Prozess startet oder aus einem Ruhezustand aktiviert. Dieses Schlüsselwort (bzw. die Phrase) wird häufig auch als „Aktivierungswort“ bezeichnet.

[0014] Der Energie-intensivere, primäre Spracherkennungs-Prozess wandelt nun den gesamten Audio-Puffer oder den jüngsten Teil ab einer erkannten Sprechpause, die typischerweise den Satzanfang einer Frage kennzeichnet, in Text um, wobei der primäre Spracherkennungs-Prozess anschließend nahtlos mit einer Umwandlung der „Liveübertragung“ vom Mikrofon fortfährt. Der per Spracherkennung erzeugte Text, sowohl aus dem Audio-Puffer, als auch aus der anschließenden „Liveübertragung“, wird einem Dialogsystem (bzw. Chatbot) zugeführt, welches ebenfalls gestartet wird oder aus dem Ruhezustand aktiviert wird.

[0015] Das Dialogsystem analysiert den Inhalt des Textes darauf hin, ob er eine Frage, eine Mitteilung und/oder eine Aufforderung enthält, die vom Benutzer an den Software-Agenten (bzw. an den digitalen Sprachassistenten) gerichtet wird, beispielsweise mittels semantischer Analyse.

[0016] Falls im Text eine Aufforderung oder ein Thema erkannt wird, für welche oder für welches der Software-Agent (bzw. digitale Sprachassistent) zuständig ist, wird vom Dialogsystem eine passende Aktion ausgelöst oder eine passende Antwort generiert und dem Benutzer per Ausgabevorrichtung (z.B. Lautsprecher und/oder Display) mitgeteilt.

[0017] Am Ende des Vorgangs kehrt die Kontrolle dann wieder zum sekundären Spracherkennungs-Prozess zurück, welcher die Umgebungsgeräusche (bzw. die Sprache) nach weiteren Schlüsselwörtern oder Phrasen überwacht.

Figurenliste

[0018] Weitere Ziele, Merkmale, Vorteile und Anwendungsmöglichkeiten der vorliegenden Erfindung ergeben sich aus den Zeichnungen und der nachfolgenden Beschreibung. Dabei bilden, unabhängig von der Zusammenfassung in einzelnen Ansprüchen oder deren Rückbeziehung, alle beschriebenen und/oder bildlich dargestellten Merkmale für sich oder in beliebiger Kombination den Gegenstand der Erfindung.

Fig. 1 zeigt ein Smartphone mit Mikrofon und Lautsprecher, auf dem ein digitaler Sprachassistent als Software läuft. (Stand der Technik)

Fig. 2 zeigt ein Datenflussdiagramm des grundlegenden Verfahrens.

Fig. 3 zeigt eine schematische Darstellung des zeitlichen Ablaufs des Verfahrens auf einer Zeitachse t mit Text-Beispiel und der Schlüsselwort-Phrase „wie ist“.

Fig. 4 veranschaulicht eine Ausführungsform, bei der sich sowohl der primäre Spracherkennungs-Prozess (ausgeführt auf einem Prozessor) als auch der sekundäre Spracherkennungs-Prozess (als Hardware-Schaltung) im lokalen Endgerät befinden.

Fig. 5 veranschaulicht eine einfache Ausführungsform, bei der sowohl der primäre Spracherkennungs-Prozess als auch der sekundäre Spracherkennungs-Prozess auf dem selben Single- oder Mehrkern-Prozessor ausgeführt werden.

Fig. 6 veranschaulicht eine bevorzugte Ausführungsform, bei der sich der sekundäre Spracherkennungs-Prozess (als Hardware-Schaltung) im lokalen Endgerät befindet und bei der der primäre Spracherkennungs-Prozess auf dem Prozessor eines Servers ausgeführt wird, der mit dem Endgerät über ein Netzwerk verbunden ist.

Fig. 7 zeigt einen Programmablaufplan (Flussdiagramm) des Verfahrens einschließlich der Erkennung von Satz-Anfang, Satz-Ende und irrelevanten Audio-Aufnahmen.

Grundlegende Ausführung der Erfindung

[0019] Ein Endgerät kann als mobiles Computersystem oder als stationäres, kabelgebundenes Computersystem realisiert werden. Das Endgerät ist über ein Netzwerk mit einem Server verbunden und kommuniziert nach dem Client-Server-Modell. Mobile Endgeräte sind per Funk mit dem Netzwerk verbunden. Bei dem Netzwerk handelt es sich typischerweise um das Internet. In **Fig. 1** ist das Endgerät **1** ein Smartphone.

[0020] Auf dem Endgerät **1** läuft die Software eines digitalen Sprachassistenten. Unter Bezugnahme auf **Fig. 2** verfügt das Endgerät **1** über eine Vorrichtung zur digitalen Tonaufnahme und Wiedergabe: typischerweise ein oder mehrere Mikrofone **2** und ein oder mehrere Lautsprecher **3** samt zugehörigen A/D-Wandler **5** und D/A-Wandler Schaltungen. Im regulären Vollbetrieb wird die digitale Tonaufnahme **11** (mit den Umgebungsgeräuschen bzw. der Sprache) einem primären Spracherkennungs-Prozess **8** zugeführt. Der primäre Spracherkennungs-Prozess **8** kann je nach Ausführungsform als Software oder als Hardware-Schaltkreis realisiert werden. Außerdem kann sich die Spracherkennung je nach Ausführungsform im lokalen Endgerät **1** befinden oder auf einem Server **28**, wobei die digitale Tonaufnahme **11** dann kontinuierlich über ein Netzwerk **29** zum Server **28** übertragen wird. Eine typische Ausführungsform verwendet zur Spracherkennung den Server **28**, wobei die Spracherkennung als Software implementiert ist.

[0021] Bei dem primären Spracherkennungs-Prozess **8** handelt es sich um eine hochwertige Spracherkennung, welche während des Dialogs mit dem Benutzer die akustischen Informationen möglichst vollständig in Text **13** umsetzt und dabei typischerweise den gesamten unterstützten Wortschatz des Spracherkennungs-Systems verwendet. Dieser Betriebszustand wird im Folgenden als Vollbetrieb bezeichnet. Vor und nach dem Dialog mit dem Benutzer kann sich das Endgerät **1** in einen Ruhezustand bzw. Standby-Modus versetzen, um Energie zu sparen.

[0022] Neben der Spracherkennung für den Vollbetrieb verfügt das System gemäß **Fig. 2** über einen zweiten Spracherkennungs-Prozess für den Ruhezustand bzw. Standby-Modus. Dieser sekundäre Spracherkennungs-Prozess **7** ist auf geringen Ressourcen-Verbrauch optimiert und kann ebenfalls je nach Ausführungsform als Software oder als Hardware-Schaltkreis realisiert werden. Bei einer Realisierung in Hardware ist auf geringe Leistungsaufnahme zu achten und bei einer Software-Implementierung auf eine geringe Beanspruchung von Ressourcen, wie Prozessor oder Arbeitsspeicher. Der sekundäre Spracherkennungs-Prozess **7** kann je nach Ausführung auf dem lokalen Endgerät **1** ausgeführt werden oder auf einem Server, wobei die digitale Tonaufnahme **11** dann zum Server übertragen wird.

[0023] Eine bevorzugte Ausführungsform verwendet zur Spracherkennung im Standby-Modus das lokale Endgerät **1**, wobei der sekundäre Spracherkennungs-Prozess **7** als FPGA (Field Programmable Gate Array) oder als ASIC (Application-Specific Integrated Circuit) realisiert ist und auf geringe Leistungsaufnahme optimiert ist.

[0024] Um den geringen Ressourcen-Verbrauch des sekundären Spracherkennungs-Prozesses **7** realisieren zu können, verfügt dieser über einen stark begrenzten Wortschatz. Der sekundäre Spracherkennungs-Prozess **7** kann somit nur wenige Wörter oder kurze Ausschnitte aus Redewendungen (Phrasen) verstehen. Diese Schlüsselwörter und Phrasen sind so zu wählen, dass sie die typischen Merkmale bei einer Kontaktaufnahme oder einer Frage an den digitalen Sprachassistenten enthalten. Die gewählten Schlüsselwörter und Phrasen müssen sich dabei nicht notwendigerweise am Anfang eines Satzes befinden. Geeignet sind z.B. alle Schlüsselwörter und Phrasen, die eine Frage vermuten lassen, beispielsweise „hast du“, „gibt es“, „brauche ich“, „habe ich“.

[0025] Unter Bezugnahme auf **Fig. 2** werden im Standby-Modus alle ankommenden Audio-Daten **11** für eine gewisse Zeit in einem Audio-Puffer **6** zwischengespeichert. Im einfachsten Fall wird für diesen Zweck der Arbeitsspeicher verwendet. Wenn sich der sekundäre Spracherkennungs-Prozess **7** im Endgerät **1** befindet, dann sollte sich auch der Audio-Puffer **6** im Endgerät **1** befinden. Wenn die Standby-Spracherkennung über den Server abgewickelt wird, sollte der Audio-Puffer **6** vom Server vorgehalten werden. Die Länge des Audio-Puffers **6** ist so zu wählen, dass mehrere gesprochene Sätze hineinpassen. Praxistaugliche Werte liegen zwischen 15 Sekunden und 2 Minuten.

[0026] Sobald der sekundäre Spracherkennungs-Prozess **7** ein potentiell relevantes Schlüsselwort **18** oder eine Phrase erkannt hat, z.B. „weißt du ob“, veranlasst dieser ein vorübergehendes Aufwachen des primären Spracherkennungs-Prozesses **8**; siehe Trigger-Signal **12** in **Fig. 2**. Dem primären Spracherkennungs-Prozess **8** wird der Inhalt des Audio-Puf-

fers **6** übergeben: In einer einfachen Ausführungsform befindet sich der Audio-Puffer **6** im Arbeitsspeicher des Endgeräts **1**. Wenn auch der primäre Spracherkennungs-Prozess **8** auf dem Endgerät **1** ausgeführt wird, ist lediglich ein Zugriff auf den Audio-Puffer **6** im Arbeitsspeicher erforderlich. Wenn der primäre Spracherkennungs-Prozess **8** auf dem Server **28** ausgeführt wird, wird der Inhalt des Audio-Puffers **6** über das Netzwerk **29** zum Server **28** übertragen.

[0027] Durch den Audio-Puffer **6** liegt dem primären Spracherkennungs-Prozess **8** nun die Vergangenheit des potentiellen Gesprächs vor, beispielsweise die letzten 30 Sekunden. Der primäre Spracherkennungs-Prozess **8** muss in der Lage sein, die eintreffenden Audio-Daten **11** mit hoher Priorität zu verarbeiten: Ziel ist es, den Audio-Puffer **6** zeitnahe zu leeren, um bald möglichst „Live-Audio"-Daten 22 zu verarbeiten. Weitere Details können der Zeichnung **Fig. 3** und der Bezugszeichenliste entnommen werden. Das Resultat des primären Spracherkennungs-Prozesses **8** ist der gesprochene Text **13** der jüngsten Vergangenheit bis zur Gegenwart.

[0028] Dieser Text **13** wird dem Dialogsystem **9** zugeführt, welches mit semantischer Analyse oder ggf. Künstlicher Intelligenz analysiert, inwiefern tatsächlich eine Anfrage an den digitalen Sprachassistenten vorliegt. Es ist auch möglich, dass das von dem sekundären Spracherkennungs-Prozess **7** erkannte Schlüsselwort **18** im nun vorliegenden Text **13** nicht mehr vorkommt, da die Spracherkennung im Vollbetrieb (d.h. der primäre Spracherkennungs-Prozess **8**) höherwertiger ist und sich der sekundäre Spracherkennungs-Prozess **7** somit geirrt hat.

[0029] In allen Fällen, in denen sich die im Audio-Puffer **6** befindliche Tonaufnahme **11** als irrelevant erweist, veranlasst das Dialogsystem **9** eine unmittelbare Rückkehr in den Standby-Modus, insbesondere wenn nur Störgeräusche vorliegen oder wenn der Sinn des Textes vom Dialogsystem **9** nicht erkannt wurde. Falls das Dialogsystem **9** jedoch zu dem Ergebnis kommt, dass die im Audio-Puffer **6** enthaltene Frage, Mitteilung oder Aufforderung relevant ist, so verbleibt das Endgerät **1** im Vollbetrieb, und das Dialogsystem **9** wird mit dem Benutzer interagieren. Sobald keine weiteren Anfragen oder Mitteilungen vom Benutzer erfolgen, wechselt das Endgerät **1** wieder in den Standby-Modus und übergibt somit die Kontrolle an den sekundären Spracherkennungs-Prozess **7**. Weitere Details können dem in **Fig. 7** dargestellten Flussdiagramm entnommen werden.

Bevorzugte Ausführungen der Erfindung

[0030] Im folgenden werden bevorzugte Ausführungsformen beschrieben. In einigen Fällen werden auch Alternativen oder optionale Funktionen erwähnt.

[0031] Gemäß der Erfindung wird nach dem Erkennen eines Schlüsselworts **18** oder einer Phrase durch den sekundären Spracherkennungs-Prozess **7** zunächst im Audio-Puffer **6** der Anfang eines Satzes mit einer Frage, Mitteilung oder Aufforderung gesucht. Wie in **Fig. 3** dargestellt, kann zumeist davon ausgegangen werden, dass sich vor dem Anfang des Satzes ein kurzer Zeitabschnitt **16** ohne Sprache (d.h. mit relativer Stille, bezogen auf die Umgebungsgeräusche) befindet, da die meisten Menschen kurz inne halten, wenn sie eine konkrete, wohl formulierte Frage, Mitteilung oder Aufforderung an den digitalen Sprachassistenten richten wollen.

[0032] Um den Anfang des relevanten Satzes zu finden, wird der Audio-Puffer **6**, ausgehend von der zeitlichen Position des erkannten Schlüsselworts **18** bzw. der Phrase, zeitlich rückwärts durchsucht, bis ein Zeitabschnitt gefunden wird, welcher sich als Stille bzw. Sprechpause **16** interpretieren lässt. Typischerweise sollte die Länge dieses Zeitabschnitts mit der Sprechpause **16** mindestens eine Sekunde betragen.

[0033] Sobald eine Position mit (relativer) Stille bzw. der Sprechpause **16** gefunden wird und somit der wahrscheinliche Anfang eines Satzes feststeht, wird dem nachfolgend gestarteten bzw. aktivierten primären Spracherkennungs-Prozess **8** dieser Inhalt **17** des Audio-Puffers **6** übergeben.

[0034] Falls bei der Auswertung durch das Dialogsystem **9** kein Sinn im Text **13** erkannt wird, möglicherweise weil der Satzanfang falsch gedeutet wurde, kann optional in einem zweiten Schritt der gesamte Inhalt **21** des Audio-Puffers **6** zusammen mit der nachfolgenden „Liveübertragung" 22 in Text **13** umgewandelt werden und vom Dialogsystem **9** analysiert werden.

[0035] Falls es nicht gelingt, eine Position mit (relativer) Stille bzw. einer Sprechpause **16** im gesamten Audio-Puffer **6**, **21** zu lokalisieren, liegt wahrscheinlich keine Frage, Mitteilung oder Aufforderung an den digitalen Sprachassistenten vor, sondern ein Störgeräusch oder ein Gespräch zwischen Menschen. In diesem Fall ist es nicht notwendig, den primären Spracherkennungs-Prozess **8** zu starten oder zu aktivieren.

[0036] Damit ein Anwender nicht übermäßig lange auf eine Antwort **14** (oder Aktion) warten muss, ist es vorteilhaft, dass nach dem Auslösen **12** durch ein Schlüsselwort **18** oder durch eine Phrase, der primäre Spracherkennungs-Prozess **8** mit hoher Priorität ausgeführt wird und in kurzer Zeit **23**, **24** abgeschlos-

sen ist. Dies wird in **Fig. 3** durch die gestrichelten Linien **23** und **24** dargestellt.

[0037] Da erfindungsgemäß eine vollwertige Spracherkennung durch den primären Spracherkennungs-Prozess **8** erfolgt, darf der sekundäre Spracherkennungs-Prozess **7** beim Erkennen von Schlüsselwörtern **18** bzw. Phrasen eine erhöhte Falsch-Positiv-Rate aufweisen, d.h. der Auslöser oder Trigger **12** des sekundären Spracherkennungs-Prozesses **7** reagiert empfindlich: Bei der Überwachung der Umgebungsgeräusche wird nur extrem selten ein Schlüsselwort **18** bzw. eine Phrase übersehen. Werden hingegen andere Geräusche oder andere Wörter fälschlicherweise als Schlüsselwort **18** bzw. Phrase interpretiert, so werden diese Fehler dann vom primären Spracherkennungs-Prozess **8** korrigiert: Sobald erkannt wird, dass der Trigger **12** fälschlicherweise ausgelöst worden ist, beendet bzw. deaktiviert sich der primäre Spracherkennungs-Prozess **8** umgehend.

[0038] Die stark eingeschränkte Erkennungsleistung des sekundären Spracherkennungs-Prozesses **7** ermöglicht es, diesen besonders energiesparend zu gestalten; beispielsweise als Software auf einem langsam getakteten Prozessor mit geringer Leistungsaufnahme oder auf einem digitalen Signalprozessor **25**, ebenfalls optimiert auf geringe Leistungsaufnahme. Ebenso möglich ist ein FPGA oder ein ASIC oder generell eine stromsparende Hardware-Schaltung **25**; siehe hierzu auch das Blockdiagramm gemäß **Fig. 4**.

[0039] Falls sowohl der primäre als auch der sekundäre Spracherkennungs-Prozess **7**, **8** auf der lokalen Hardware, d.h. auf dem Endgerät **1**, ausgeführt werden, können, wie in **Fig. 5** dargestellt, beide Spracherkennungs-Prozesse **7**, **8** auch den selben Single- oder Mehrkem-Prozessor **27** verwenden, wobei der sekundäre Spracherkennungs-Prozess **7** in einem besonders Ressourcen-schonenden Betriebsmodus läuft, welcher mit geringem Speicherbedarf und geringer Stromaufnahme auskommt.

[0040] Besonders vorteilhaft ist es jedoch, wenn der primäre Spracherkennungs-Prozess **8** und das Dialogsystem **9** auf einem externen Server **28** oder auf einem Serververbund ausgeführt werden, wie in **Fig. 6** dargestellt. Dabei wird der gesamte oder der jüngste Inhalt **17**, **21** des Audio-Puffers **6** sowie im Anschluss auch eine „Liveübertragung“ der Audio-Daten **11**, **19**, **22** über ein Netzwerk **29** bzw. Funknetzwerk zum Server **28** oder Serververbund übertragen. Typischerweise handelt es sich bei dem Netzwerk **29** um das Internet.

[0041] Es entsteht eine Latenz bzw. Übertragungsverzögerung, sobald nach einer Sprachaktivierung **12** (ausgelöst durch den sekundären Spracherkennungs-Prozess **7**) der Inhalt des Audio-Puffers **6** über das Netzwerk **29** zum Server **28** bzw. zum Server-Verbund übertragen werden muss, damit der primäre Spracherkennungs-Prozess **8** und das Dialogsystem **9** den Inhalt auswerten können. Um diese Latenz zu vermeiden, kann ein „vorauseilender Standby-Modus“ verwendet werden: Sobald die Anwesenheit eines Benutzers detektiert wird, überträgt das System im „vorauseilenden Standby-Modus“ den Inhalt **21** des Audio-Puffers **6** und die sich daran anschließende „Liveübertragung“ 22 der Umgebungsgeräusche bzw. Sprache an den externen Server **28** oder Serververbund. Die Audio-Daten **11** werden dort zwischengespeichert, so dass im Fall einer Sprachaktivierung **12** der primäre Spracherkennungs-Prozess **8** nahezu latenzfrei auf die Audio-Daten **11** zugreifen kann.

[0042] Von der Anwesenheit eines Benutzers kann ausgegangen werden, wenn Benutzeraktivitäten vorliegen; beispielsweise Eingaben per Touchscreen oder Bewegungen und Lageänderungen des Endgeräts **1**, welche mittels eines Beschleunigungs- und Lagesensors erfasst werden. Ebenfalls möglich ist das Erkennen von Änderungen in der Helligkeit mittels eines Lichtsensors, das Erkennen von Positionsänderungen per Satellitennavigation, beispielsweise GPS, sowie eine Gesichtserkennung per Kamera.

[0043] Optional kann der sekundäre Spracherkennungs-Prozess **7** die Überwachung der Umgebungsgeräusche auf Schlüsselwörter **18** bzw. Phrasen intensivieren, solange sich das System im „vorauseilenden Standby-Modus“ befindet.

[0044] Grundsätzlich lassen sich die Einträge im Schlüsselwort- und Phrasen-Katalog einteilen in:

- Fragewörter und fragende Phrasen: z.B. „wer hat“, „was ist“, „wie kann“, „wie ist“, „wo gibt es“, „gibt es“, „weißt du ob“, „kann man“.

- Aufforderungen und Befehle: Beispielsweise die Aufforderung: „Bitte schreibe eine E-Mail an Hans“. Erkannt wird in diesem Beispiel die Phrase „schreibe eine E-Mail“. Ein weiteres Beispiel: „Ich möchte ein Foto machen.“ Erkannt wird die Phrase „Foto machen“.

- Substantive zu Themen, zu denen es Informationen in der Datenbank des Dialogsystems **9** gibt: z.B. „Wetter“, „Termin“ und „Fußball“.

- Produktnamen, Spitznamen und Gattungsbegriffe zur direkten Ansprache des digitalen Sprachassistenten. Beispiele für Gattungsbegriffe: „Handy“, „Smartphone“, „Computer“, „Navi“.

[0045] Die Verwendung eines Produktnamens als Schlüsselwort **18** hat den Vorteil, dass sich im Vergleich zu einem Katalog mit Fragewörtern die Häu-

figkeit reduzieren lässt, mit der das System unnötigerweise in den Vollbetrieb wechselt. Bei Verwendung eines Produktnamens kann davon ausgegangen werden, dass der digitale Sprachassistent zuständig ist. Zum Beispiel: „Hallo **<**Produktname**>**, bitte berechne die Quadratwurzel aus **49**“ oder „Wie spät ist es, **<**Produktname**>**?“.

[0046] In einer vorteilhaften Ausführungsform lässt sich der Schlüsselwort- und Phrasen-Katalog vom Anwender ändern. Wenn die Sprachaktivierung per Produktname oder Gattungsbegriff erfolgt, so könnte der Benutzer beispielsweise einen Spitznamen für das Endgerät **1** als weiteres, alternatives Schlüsselwort **18** festlegen. Der Benutzer könnte auch einige Schlüsselwörter **18** oder Phrasen aus dem Katalog streichen, z.B. wenn sich der digitale Sprachassistent seltener melden soll oder nur noch zu bestimmten Themen.

[0047] Sobald der sekundäre Spracherkennungs-Prozess **7** ein Schlüsselwort **18** oder eine Phrase erkannt hat, muss der Benutzer einige Augenblicke warten, bis der primäre Spracherkennungs-Prozess **8** und das Dialogsystem **9** eine Antwort **14** (oder Aktion) generiert haben. In einer besonders vorteilhaften Ausführungsform wird beim Erkennen eines Schlüsselworts **18** oder einer Phrase durch den sekundäre Spracherkennungs-Prozess **7** umgehend ein optisches, akustisches und/oder haptisches Signal an den Benutzer ausgegeben, beispielsweise ein kurzes Piepsen oder Vibrieren des Endgeräts **1**, eine Anzeige auf dem Display **4** oder das Einschalten der Hintergrundbeleuchtung des Displays **4**. Der Benutzer ist dann informiert, dass seine Anfrage bei dem Endgerät **1** angekommen ist. Gleichzeitig stört diese Form von Signalisierung nur minimal, falls das Schlüsselwort **18** oder die Phrase irrtümlich erkannt wurde. In diesem Fall, wenn im Audio-Puffer **6** bzw. aus dem daraus resultierenden Text **13** kein relevanter oder kein auswertbarer Inhalt erkannt werden kann, ist es vorteilhaft, ein weiteres optisches, akustisches oder haptisches Signal auszugeben, welches sich zweckmäßigerweise von dem ersten Signal unterscheidet, beispielsweise ein Doppelton (erst hoch, dann tief) oder das Ausschalten der Hintergrundbeleuchtung, welche zuvor eingeschaltet wurde.

[0048] In einer weiteren Ausführungsform kann der digitale Sprachassistent verschiedene Sprecher an der Stimme auseinander halten, so dass nur Fragen, Mitteilungen und Aufforderungen vom Dialogsystem **9** beantwortet werden, die von einer berechtigten Person ausgehen, beispielsweise nur Fragen vom Besitzer. Da der primäre Spracherkennungs-Prozess **8** eine deutlich größere Erkennungsleistung hat, kann erfindungsgemäß nur dieser Prozess verschiedene Sprecher an der Stimme unterscheiden. Der sekundäre Spracherkennungs-Prozess **7** kann in dieser Ausführungsform verschiedene Sprecher hingegen nicht unterscheiden: Beim Vorliegen eines Schlüsselworts **18** bzw. einer Phrase eines noch nicht identifizierten Sprechers wird von dem sekundären Spracherkennungs-Prozess **7** die Ausführung des primären Spracherkennungs-Prozesses **8** veranlasst. Der primäre Spracherkennungs-Prozess **8** erkennt an der Stimme, ob der Sprecher berechtigt ist, den digitalen Sprachassistenten zu nutzen. Falls keine entsprechende Berechtigung vorliegt, beendet sich der primäre Spracherkennungs-Prozess **8** (bzw. er kehrt wieder in den Ruhezustand zurück), und die Kontrolle wird wieder dem sekundären Spracherkennungs-Prozess **7** übergeben. Bei diesem Vorgang kann das Dialogsystem **9** im Ruhezustand verbleiben.

[0049] In einer vorteilhaften Ausführungsform berücksichtigt das Dialogsystem **9** den Kontext einer Unterhaltung: Bei der Überwachung einer Unterhaltung zwischen Personen taucht im Gespräch ein Schlüsselwort **18** bzw. eine Phrase aus dem Schlüsselwort- und Phrasen-Katalog auf (beispielsweise „Fußball“), so dass der primäre Spracherkennungs-Prozess **8** und das Dialogsystem **9** gestartet bzw. aktiviert werden. Das Dialogsystem **9** prüft, ob es für den Inhalt **21**, **22** des aktuellen Gesprächs zuständig ist, insbesondere, ob eine Frage, Mitteilung oder Aufforderung an den digitalen Sprachassistenten gerichtet wurde. Falls das Dialogsystem **9** nicht zuständig ist, speichert das Dialogsystem **9** den Kontext und/oder das Thema und/oder die Schlüsselwörter bzw. Phrasen für einen späteren Rückbezug und kehrt zusammen mit dem primären Spracherkennungs-Prozess **8** in den Ruhezustand zurück. Wird jetzt zu einem etwas späteren Zeitpunkt das Dialogsystem **9** erneut durch ein weiteres Schlüsselwort **18** bzw. Phrase (z.B. „wer hat“) gestartet bzw. reaktiviert, so können die zuvor gesicherten Informationen als Kontext berücksichtigt werden. Beispielsweise kann auf die Frage „Wer hat heute das Spiel gewonnen?“ mit den Fußballergebnissen des aktuellen Spieltages geantwortet werden.

[0050] Da die vollständigen Sätze der auszuwertenden Fragen, Mitteilungen oder Aufforderungen des Benutzers im Audio-Puffer **6** vorliegen, ist es auch möglich, die Spracherkennung im Rahmen des primären Spracherkennungs-Prozesses **8** mehrfach durchzuführen. Zunächst könnte die Spracherkennung mit einem besonders schnellen Algorithmus durchgeführt werden, der die Wartezeit des Benutzers verkürzt. Falls der resultierende Text **13** für das Dialogsystem **9** nicht stichhaltig ist bzw. nicht auswertbar ist, kann der Audio-Puffer **6** erneut in Text **13** umgewandelt werden, und zwar mit einem oder mehreren anderen Verfahren der Spracherkennung, die beispielsweise besonders resistent gegenüber Störgeräuschen sind.

[0051] In den Ansprüchen wird für das Schlüsselwort **18** (bzw. die Phrase) der Begriff „Aktivierungswort“ verwendet.

Bezugszeichenliste

1 Smartphone (Endgerät)

2 Mikrofon

3 Lautsprecher

4 Display (Anzeige)

5 Analog-Digital Wandler (A/D)

6 Audio-Puffer

7 Sekundärer Spracherkennungs-Prozess

8 Primärer Spracherkennungs-Prozess

9 Dialogsystem

10 Analoge Mikrofon-Signale

11 Digitale Audio-Daten

12 Trigger-Signal nach erkanntem Schlüsselwort (bzw. Phrase)

13 Text (digitale Darstellung mittels Zeichenkodierung)

14 Antwort (oder Aktion) des Dialogsystems

15 Tonaufnahme des zuvor gesprochenen Satzes im Audio-Puffer

16 Tonaufnahme der Sprechpause (Stille)

17 Tonaufnahme des aktuellen Satzes (erster Teil) im Audio-Puffer

18 Schlüsselwort (bzw. Phrase)

19 „Liveübertragung“ des aktuellen Satzes (zweiter Teil)

20 Start des Dialogsystems

21 Audio-Daten der jüngsten Vergangenheit im Audio-Puffer

22 „Liveübertragung“ der Audio-Daten

23 Verzögerung der Bearbeitung bezogen auf den Satz-Anfang

24 Reduzierte Verzögerung am Satz-Ende

25 Hardware-Schaltung (Digitaler Signalprozessor, FPGA oder ASIC)

26 Hauptprozessor

27 Single- oder Mehrkern-Prozessor mit Stromsparfunktion

28 Server (oder Server-Verbund)

29 Netzwerk (Funk, Internet)

30 Mikrofon-Signale per A/D-Wandler digitalisieren;

31 Live-Audio-Daten im Audio-Puffer zwischenspeichern;

32 Sekundären Spracherkennungs-Prozess mit Live-Audio-Daten ausführen;

33 Schlüsselwort oder Phrase gefunden?

34 Audio-Puffer rückwärts nach Sprechpause durchsuchen;

35 Sprechpause gefunden?

36 Primären Spracherkennungs-Prozess und Dialogsystem starten/aktivieren;

37 Primären Spracherkennungs-Prozess anwenden auf Audio-Puffer ab Sprechpause;

38 Primären Spracherkennungs-Prozess anwenden auf neue Live-Audio-Daten;

39 Sprechpause des Satzendes gefunden?

40 Den Text des Satzes im Dialogsystem analysieren;

41 Text enthält relevante Frage, Mitteilung oder Befehl?

42 Antwort generieren oder Aktion auslösen;

43 Gibt es weitere Fragen/Befehle vom Benutzer?

44 Primären Spracherkennungs-Prozess und Dialogsystem beenden/deaktivieren;

Patentansprüche

1. Verfahren zur Sprachaktivierung eines Software-Agenten mittels eines Aktivierungsworts **dadurch gekennzeichnet**,
a) dass ein Aktivierungswort (18) am Satzanfang, innerhalb des Satzes und/oder am Satzende erkannt wird,
b) dass Audio-Daten (11) mit mindestens einem Mikrofon (2) aufgenommen werden,
c) dass die Audio-Daten (11) kontinuierlich in mindestens einem Audio-Puffer (6) zwischengespeichert werden, so dass der Audio-Puffer (6) stets die Audio-Daten (11) der jüngsten Vergangenheit enthält,
d) dass die Audio-Daten (11) zeitnahe mindestens einem sekundären Spracherkennungs-Prozess (7) zugeführt werden,
e) dass beim Erkennen eines Aktivierungsworts (18) durch den sekundären Spracherkennungs-Prozess (7) mindestens die nachfolgenden Vorgänge ausgelöst werden,
f) dass im Audio-Puffer (6), ausgehend von der zeitlichen Position des erkannten Aktivierungsworts (18), rückwärts gesucht wird, bis ein geeigneter Zeitabschnitt gefunden wird, welcher sich als Sprechpause (16) interpretieren lässt,
g) dass mindestens einem primären Spracherkennungs-Prozess (8) der Inhalt (17) des Audio-Puffers (6) ab der erkannten Sprechpause (16) übergeben

wird, sowie eine sich daran anschließende Liveübertragung (22) der Audio-Daten (11),
h) dass der primäre Spracherkennungs-Prozess (8) die Audio-Daten (11) in Text (13) umwandelt, und zwar bis eine Sprechpause (16) am Satzende gefunden wird,
i) dass der Text (13) mindestens einem Dialogsystem-Prozess (9) zuführt wird, welcher den Inhalt des Textes (13) darauf hin analysiert, ob dieser eine Frage, eine Mitteilung und/oder einen Befehl enthält, die bzw. der vom Benutzer an den Software-Agenten gerichtet wurde, und mindestens falls dies bejaht wird, der Dialogsystem-Prozess (9) eine passende Aktion auslöst oder eine passende Antwort (14) generiert und mit dem Benutzer per Ausgabevorrichtung (3, 4) in Kontakt tritt und
j) dass nach Abschluss der Interaktion mit dem Benutzer die Ausführung des Dialogsystem-Prozesses (9) und spätestens dann auch die Ausführung des primären Spracherkennungs-Prozesses (8) beendet oder deaktiviert werden und die Kontrolle wieder dem sekundären Spracherkennungs-Prozess (7) zurückgeben wird.

2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, dass falls der Dialogsystem-Prozess (9) bei der Auswertung keinen Sinn im Text (13) erkennt, in einem zweiten Schritt der gesamte Inhalt (21) des Audio-Puffers (6) zusammen mit der nachfolgenden Liveübertragung (22) der Audio-Daten (11) vom primären Spracherkennungs-Prozess (8) in Text (13) umgewandelt wird und vom Dialogsystem-Prozess (9) analysiert wird, so dass eine möglicherweise falsch gedeutete Sprechpause (16) am Satzanfang kompensiert wird.

3. Verfahren nach einem der Ansprüche 1 oder 2, **dadurch gekennzeichnet**, dass von einer Ausgabevorrichtung ein optisches Signal an den Benutzer ausgegeben wird, sobald vom sekundären Spracherkennungs-Prozess (7) ein Aktivierungswort (18) erkannt wird.

4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet**, dass der sekundäre Spracherkennungs-Prozess (7) im Vergleich zum primären Spracherkennungs-Prozess (8) eine geringere Leistungsaufnahme hat.

5. Verfahren nach einem der Ansprüche 1 bis 4, **dadurch gekennzeichnet**, dass nach dem Auslösen (12) durch ein Aktivierungswort (18) der primäre Spracherkennungs-Prozess (8) mit hoher Priorität ausgeführt wird und nach kurzer Zeit (24) abgeschlossen wird.

6. Verfahren nach einem der Ansprüche 1 bis 5, **dadurch gekennzeichnet**, dass der primäre Spracherkennungs-Prozess (8) und der Dialogsystem-Prozess (9) auf einem externen Server (28) oder auf einem Serververbund ausgeführt werden, wobei die Audio-Daten (11) über ein Netzwerk (29) und/oder Funknetzwerk zum Server (28) oder Serververbund übertragen werden.

7. Verfahren nach einem der Ansprüche 1 bis 5, **dadurch gekennzeichnet**, dass der primäre Spracherkennungs-Prozess (8) und der sekundäre Spracherkennungs-Prozess (7) auf dem selben Single- oder Mehrkern-Prozessor (27) ausgeführt werden, wobei der sekundäre Spracherkennungs-Prozess (7) in einem besonders Ressourcen-schonenden Betriebsmodus ausgeführt wird, insbesondere mit geringer Leistungsaufnahme.

8. System zur Sprachaktivierung eines Software-Agenten per Aktivierungswort, mit mindestens einem Mikrofon (2), mindestens einem Audio-Puffer (6), mindestens einer Ausgabevorrichtung (3, 4) und einer Hardware-Infrastruktur (25, 26, 27, 28, 29), welche Prozesse (7, 8, 9) ausführen kann, **dadurch gekennzeichnet**, dass die Hardware-Infrastruktur (25, 26, 27, 28, 29) so konfiguriert oder programmiert ist,
a) dass ein Aktivierungswort (18) am Satzanfang, innerhalb des Satzes und/oder am Satzende erkannt wird,
b) dass Audio-Daten (11) mit dem mindestens einen Mikrofon (2) aufgenommen werden,
c) dass die Audio-Daten (11) kontinuierlich in dem mindestens einen Audio-Puffer (6) zwischengespeichert werden, so dass der Audio-Puffer (6) stets die Audio-Daten (11) der jüngsten Vergangenheit enthält,
d) dass die Audio-Daten (11) zeitnahe mindestens einem sekundären Spracherkennungs-Prozess (7) zugeführt werden,
e) dass beim Erkennen eines Aktivierungsworts (18) durch den sekundären Spracherkennungs-Prozess (7) mindestens die nachfolgenden Vorgänge ausgelöst werden,
f) dass im Audio-Puffer (6), ausgehend von der zeitlichen Position des erkannten Aktivierungsworts (18), rückwärts gesucht wird, bis ein geeigneter Zeitabschnitt gefunden wird, welcher sich als Sprechpause (16) interpretieren lässt,
g) dass mindestens einem primären Spracherkennungs-Prozess (8) der Inhalt (17) des Audio-Puffers (6) ab der erkannten Sprechpause (16) übergeben wird, sowie eine sich daran anschließende Liveübertragung (22) der Audio-Daten (11),
h) dass der primäre Spracherkennungs-Prozess (8) die Audio-Daten (11) in Text (13) umwandelt, und zwar bis eine Sprechpause (16) am Satzende gefunden wird,
i) dass der Text (13) mindestens einem Dialogsystem-Prozess (9) zuführt wird, welcher den Inhalt des Textes (13) darauf hin analysiert, ob dieser eine Frage, eine Mitteilung und/oder einen Befehl enthält, die bzw. der vom Benutzer an den Software-Agenten gerichtet wurde, und mindestens falls dies bejaht wird,

der Dialogsystem-Prozess (9) eine passende Aktion auslöst oder eine passende Antwort (14) generiert und mit dem Benutzer per Ausgabevorrichtung (3, 4) in Kontakt tritt und
j) dass nach Abschluss der Interaktion mit dem Benutzer die Ausführung des Dialogsystem-Prozesses (9) und spätestens dann auch die Ausführung des primären Spracherkennungs-Prozesses (8) beendet oder deaktiviert werden und die Kontrolle wieder dem sekundären Spracherkennungs-Prozess (7) zurückgeben wird.

9. System nach Anspruch 8, **dadurch gekennzeichnet**, dass die Umwandlung des Inhalts (17, 21) des Audio-Puffers (6) in Text (13) in einer Zeitspanne erfolgt, die kürzer ist als es für den Benutzer gedauert hat, den entsprechenden Inhalt (17, 21) zu sprechen.

10. System nach Anspruch 8 oder 9, **dadurch gekennzeichnet**, dass der sekundäre Spracherkennungs-Prozess (7) im Vergleich zum primären Spracherkennungs-Prozess (8) eine geringere Leistungsaufnahme hat.

11. System nach einem der Ansprüche 8 bis 10, **dadurch gekennzeichnet**, dass von einer Ausgabevorrichtung ein optisches Signal an den Benutzer ausgegeben wird, sobald vom sekundären Spracherkennungs-Prozess (7) ein Aktivierungswort (18) erkannt wird.

12. System nach einem der Ansprüche 8 bis 11, **dadurch gekennzeichnet**,
a) dass das mindestens eine Mikrofon (2), der mindestens eine Audio-Puffer (6) und die mindestens eine Ausgabevorrichtung (3, 4) Bestandteile eines lokalen Endgeräts (1) sind und dass der sekundären Spracherkennungs-Prozess (7) auf dem lokalen Endgerät (1) ausgeführt wird und
b) dass der primäre Spracherkennungs-Prozess (8) und der Dialogsystem-Prozess (9) auf einem externen Server (28) oder auf einem Serververbund ausgeführt werden, wobei die Audio-Daten (11) über ein Netzwerk (29) und/oder Funknetzwerk vom lokalen Endgerät (1) zum Server (28) oder Serververbund übertragen werden.

13. System nach einem der Ansprüche 8 bis 11, **dadurch gekennzeichnet**, dass das mindestens eine Mikrofon (2), der mindestens eine Audio-Puffer (6) und die mindestens eine Ausgabevorrichtung (3, 4) Bestandteile eines lokalen Endgeräts (1) sind und dass sowohl der sekundären Spracherkennungs-Prozess (7) als auch der primäre Spracherkennungs-Prozess (8) auf dem lokalen Endgerät (1) ausgeführt werden.

14. System nach Anspruch 12 oder 13, **dadurch gekennzeichnet**, dass sich der Audio-Puffer (6) in einem Arbeitsspeicher des lokalen Endgeräts (1) befindet.

15. System nach einem der Ansprüche 8 bis 14, **dadurch gekennzeichnet**,
a) dass das Aktivierungswort (18) ein Produktname, ein Spitzname und/oder ein Gattungsbegriff ist,
b) dass der Software-Agent ein digitaler Sprachassistent ist und
c) dass die Ausgabevorrichtung (3, 4) mindestens ein Lautsprecher (3) ist.

Es folgen 4 Seiten Zeichnungen

Anhängende Zeichnungen

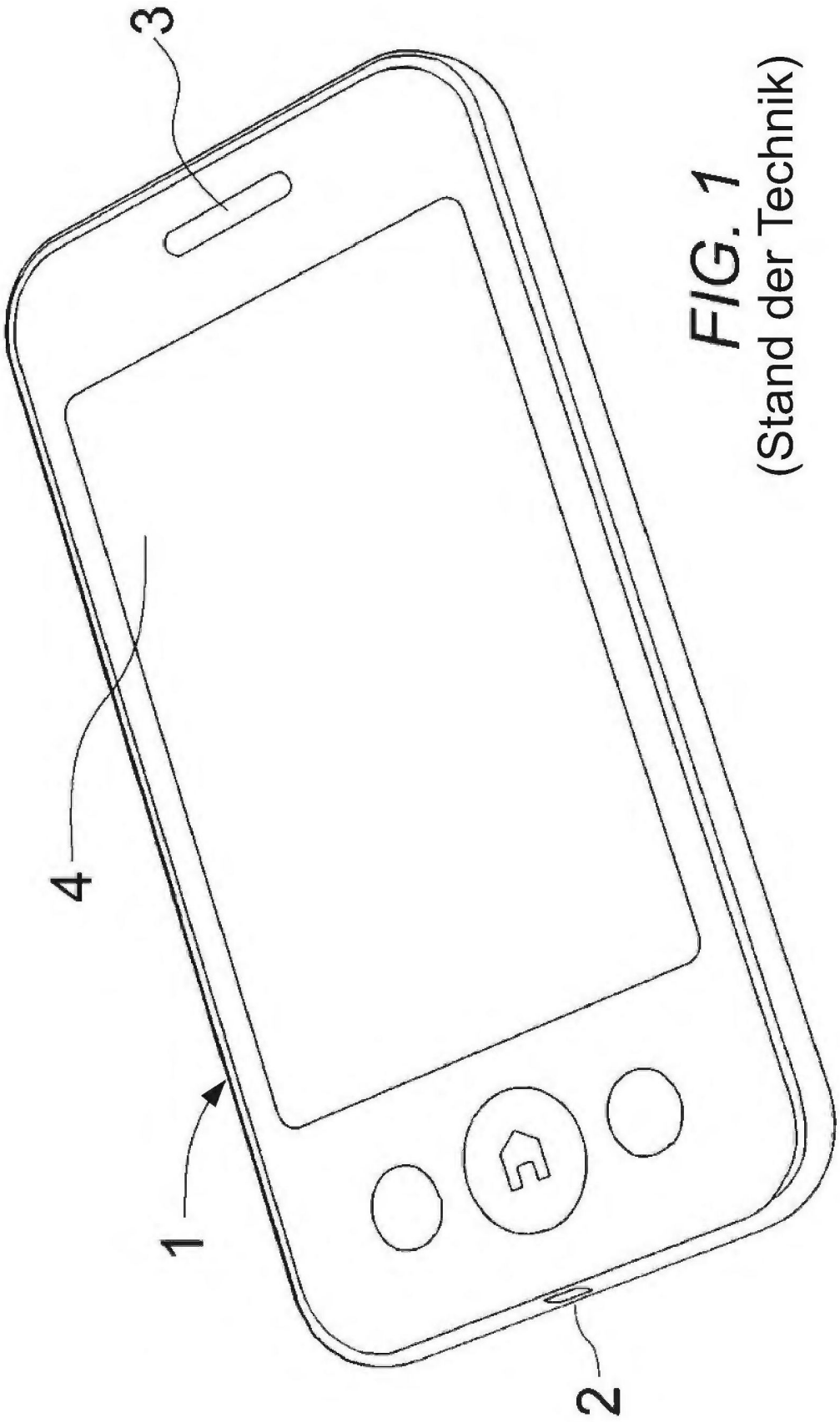


FIG. 1
(Stand der Technik)

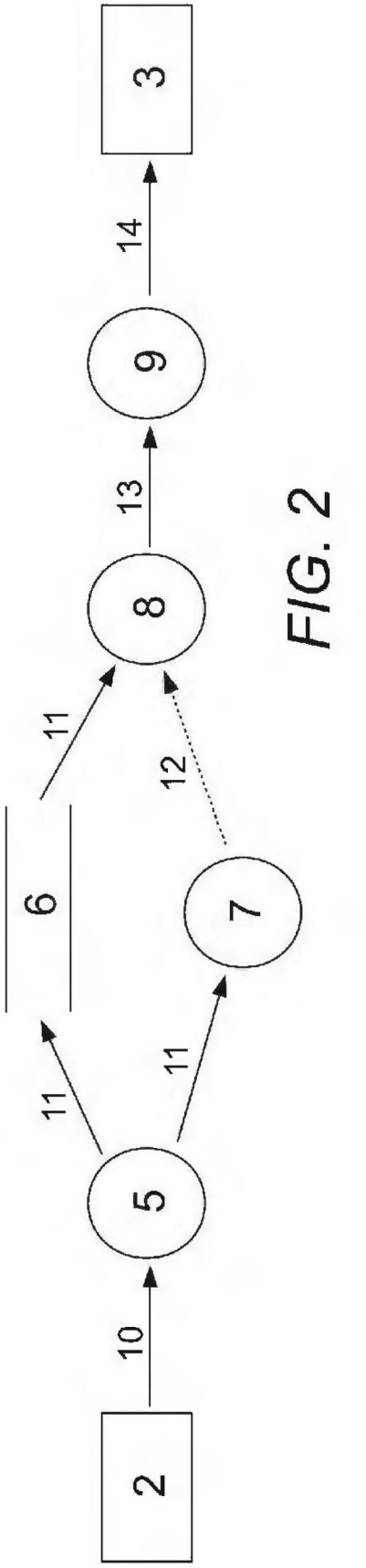

2
3
5
6
7
8
9
10
11
11
11
12
13
14

FIG. 2

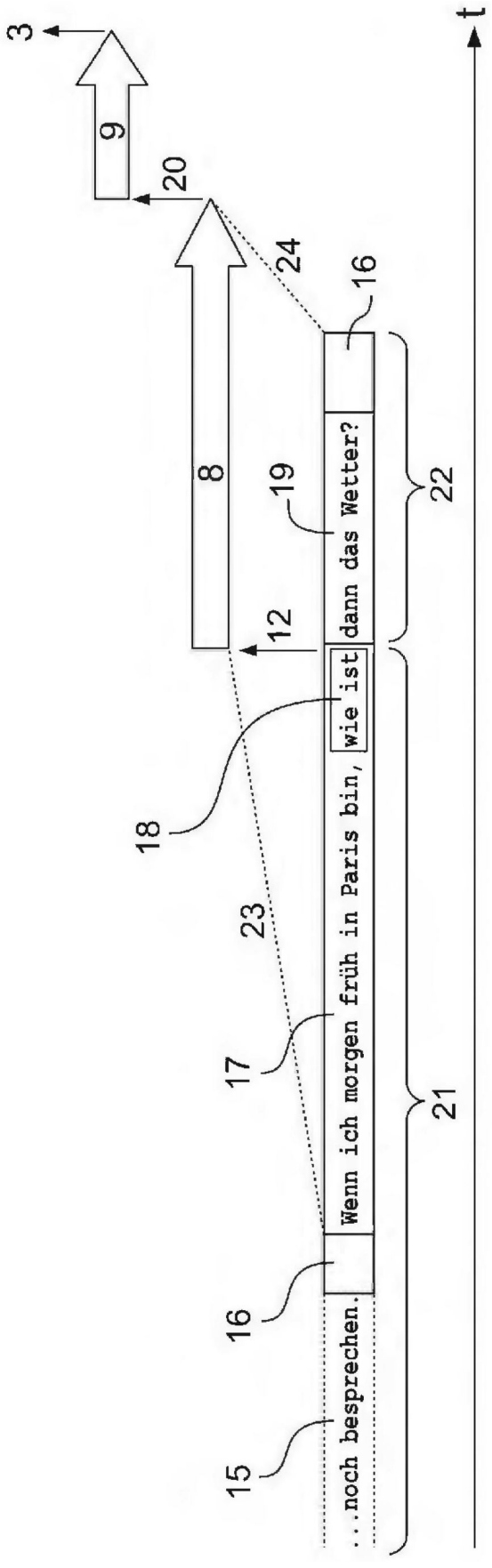

...noch besprechen.
Wenn ich morgen früh in Paris bin, wie ist dann das Wetter?
t
3
8
9
12
15
16
16
17
18
19
20
21
22
23
24

FIG. 3

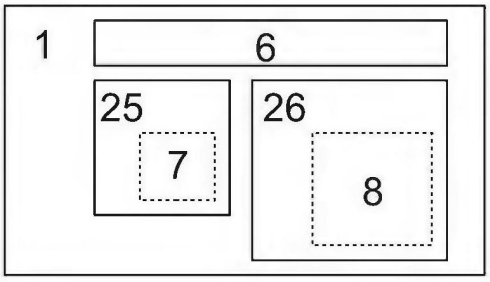


FIG. 4

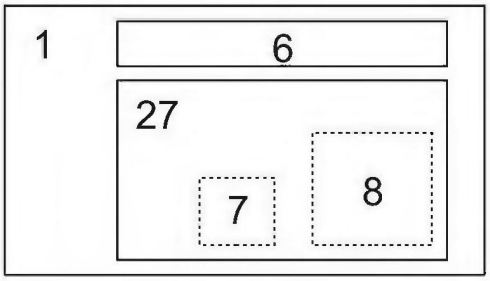


FIG. 5

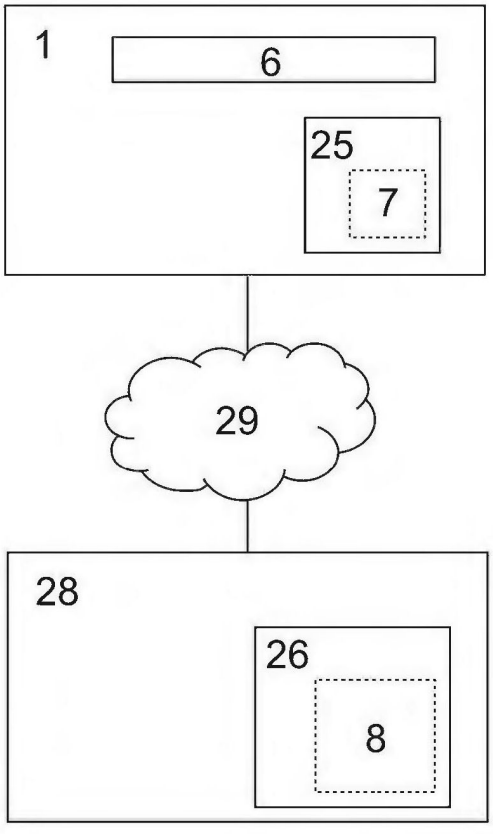


FIG. 6

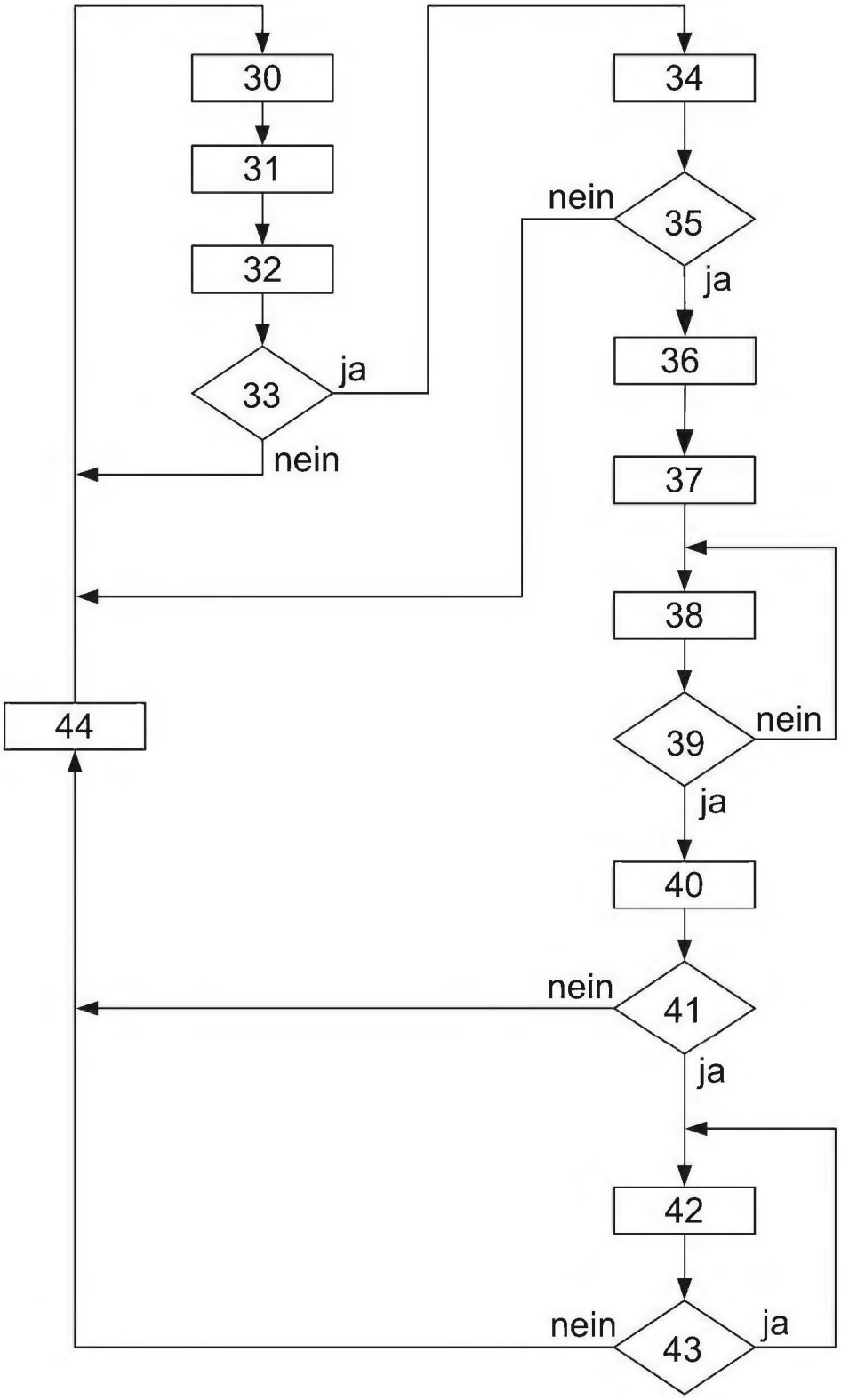

30
31
32
33
ja
nein
34
nein
35
ja
36
37
38
39
nein
ja
40
41
nein
ja
42
43
nein
ja
44

FIG. 7