(19) Bundesrepublik Deutschland
Deutsches Patent- und Markenamt

(10) **DE 102 51 112 A1** 2004.05.19

(12) # Offenlegungsschrift

(21) Aktenzeichen: **102 51 112.8**
(22) Anmeldetag: **02.11.2002**
(43) Offenlegungstag: **19.05.2004**

(51) Int Cl.[7]: **G10L 15/06**

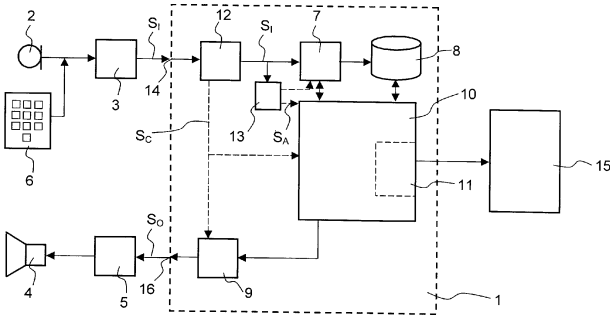
(71) Anmelder:
Philips Intellectual Property & Standards GmbH, 20099 Hamburg, DE

(72) Erfinder:
Kooiman, Albert R.R., Drs., 52070 Aachen, DE

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

(54) Bezeichnung: **Verfahren und System zur Spracherkennung**

(57) Zusammenfassung: Es wird ein Verfahren zur Spracherkennung beschrieben, bei welchem ein Sprachsignal eines Benutzers zur Erkennung einer im Sprachsignal enthaltenen Sprachinformation analysiert wird. Innerhalb einer Prüfprozedur wird das mit größter Wahrscheinlichkeit passende Erkennungsergebnis wieder in ein Sprachsignal umgewandelt und zur Verifikation und/oder Korrektur an den Nutzer ausgegeben. Bei der Analyse wird eine Anzahl von alternativen Erkennungsergebnissen generiert, welche mit den nächstgrößten Wahrscheinlichkeiten zu dem zu erkennenden Sprachsignal passen. Die Ausgabe innerhalb der Prüfprozedur erfolgt derart, dass der Benutzer bei einer Fehlerhaftigkeit des ausgegebenen Erkennungsergebnisses die Ausgabe unterbrechen kann. Es werden dann automatisch für einen vor einer Unterbrechung zuletzt ausgegebenen Abschnitt des betreffenden Erkennungsergebnisses jeweils entsprechende Abschnitte der alternativen Erkennungsergebnisse für eine Auswahl durch den Benutzer ausgegeben. Der betreffende Abschnitt im ausgegebenen Erkennungsergebnis wird anschließend auf Basis des entsprechenden Abschnitts eines ausgewählten alternativen Erkennungsergebnisses korrigiert. Die Prüfprozedur wird schließlich für die verbleibenden nachfolgenden Abschnitte des zu erkennenden Sprachsignals fortgesetzt. Darüber hinaus wird ein entsprechendes Spracherkennungssystem beschrieben.



Beschreibung

[0001] Die Erfindung betrifft ein Verfahren zur Spracherkennung, bei dem ein Sprachsignal eines Benutzers zur Erkennung einer im Sprachsignal enthaltenen Sprachinformation analysiert wird und ein mit einer größten Wahrscheinlichkeit passendes Erkennungsergebnis innerhalb einer Prüfprozedur wieder in eine Sprachinformation umgewandelt und zur Verifikation und/oder Korrektur an den Nutzer ausgegeben wird. Außerdem betrifft die Erfindung ein Spracherkennungssystem mit einer Vorrichtung zur Erfassung eines Sprachsignals eines Benutzers, mit einer Spracherkennungseinrichtung, um das erfasste Sprachsignal zur Erkennung einer im Sprachsignal enthaltenen Sprachinformation zu analysieren und ein mit einer größten Wahrscheinlichkeit passendes Erkennungsergebnis zu ermitteln, sowie mit einer Sprachausgabeeinrichtung, um das wahrscheinlichste Erkennungsergebnis innerhalb einer Prüfprozedur wieder in eine Sprachinformation umzuwandeln und zur Verifikation und/oder Korrektur an den Nutzer auszugeben.

[0002] Spracherkennungssysteme arbeiten in der Regel in der Weise, dass zunächst das Sprachsignal spektral oder zeitlich analysiert wird und das analysierte Sprachsignal dann abschnittsweise mit verschiedenen Modellen von möglichen Signalfolgen mit bekannten Sprachinformationen verglichen wird. Hierzu weist die Sprachausgabeeinrichtung in der Regel eine ganze Bibliothek von verschiedenen möglichen Signalfolgen – beispielsweise der sinnvollerweise in einer Sprache vorkommenden Wörter – auf. Anhand des Vergleichs des empfangenen Sprachsignals mit den zur Verfügung stehenden Modellen wird jeweils das am besten für einen bestimmten Abschnitt des Sprachsignals passende Modell ausgesucht, um zu einem Erkennungsergebnis zu gelangen. Dabei wird üblicherweise die Wahrscheinlichkeit berechnet, mit der jedes Modell zu dem jeweils zugehörigen Abschnitt des Sprachsignals passt. Sofern es sich bei dem Sprachsignal um längere Texte, beispielsweise einen oder mehrere Sätze handelt, werden bei der Analyse und bei der Berechnung der Wahrscheinlichkeit, wie gut die einzelnen Modelle zu den betreffenden Abschnitten eines Sprachsignals passen, auch grammatikalische und/oder linguistische Regeln hinzugezogen. Dadurch wird vermieden, dass die einzelnen Abschnitte des längeren Sprachsignals nicht nur isoliert gut zu den jeweiligen zur Verfügung stehenden Modellen passen, sondern es wird auch der Kontext berücksichtigt, innerhalb dessen sich die Sprachsignalabschnitte befinden, um insgesamt zu einem sinnvollen Erkennungsergebnis zu kommen und so die Fehlerquote zu verringern. Dennoch bleibt immer noch eine Rest-Fehlerwahrscheinlichkeit bestehen, dass von einem gesprochenen Text einige Sätze, Satzteile oder Wörter falsch verstanden wurden.

[0003] Bei den meisten Anwendungen ist es daher erforderlich, dass ein Benutzer des Spracherkennungssystems die Gelegenheit erhält, das Erkennungsergebnis zu überprüfen und gegebenenfalls zu korrigieren. Dies ist insbesondere in den Fällen notwendig, in denen der jeweilige Benutzer kein direktes Feedback auf eine Eingabe erhält, wie z. B. in Anwendungen, in denen der Benutzer einen längeren Text spricht, der dann in einer Schrifttextform oder in einer auf sonstige Weise maschinenlesbaren Textform (im Folgenden kurz „Textform" genannt) gespeichert wird. Typische Beispiele hierfür sind Diktiersysteme oder Anwendungen, bei denen Nachrichten zunächst in eine Textform umgewandelt werden und dann in dieser Textform verarbeitet bzw. über ein Kommunikationsnetz, beispielsweise als E-Mail, als Fax oder als SMS, weitergeleitet werden. Eine weitere solche Anwendungsform ist ein automatisches Übersetzungssystem, bei dem ein Sprachsignal zunächst in die Textform umgewandelt wird, dann auf Basis dieser Textform eine Übersetzung in eine andere Sprache erfolgt und schließlich der übersetzte Text wieder in ein Sprachsignal umgewandelt und mittels einer Sprachausgabeeinrichtung ausgegeben wird. Bei den klassischen Diktiersystemen an PCs ist es möglich, dass das Erkennungsergebnis unmittelbar in Textform auf einem Bildschirm des PCs dargestellt wird und dann der Benutzer den Text mit den üblichen Editierfunktionen korrigieren kann. Dieses Korrekturverfahren eignet sich jedoch nicht für solche Anwendungen, bei denen keine Möglichkeit für eine Sichtanzeige des erkannten Textes gegeben ist, beispielsweise bei einer Verwendung von Geräten ohne geeignete Anzeigeeinrichtung, wie „normale" Telefone oder bei Anwendungen für sehbehinderte Personen. In diesen Fällen ist es notwendig, das ermittelte Erkennungsergebnis über eine automatische Sprachausgabeeinrichtung, beispielsweise einen Text-to-Speech-Generator, wieder so in Sprachform an den Benutzer auszugeben, dass dieser die Möglichkeit hat, das Erkennungsergebnis zu bestätigen oder zu korrigieren.

[0004] Ein solches Verfahren wird beispielsweise in der US 6, 219,628 B1 beschrieben. In dieser Schrift werden verschiedene Möglichkeiten der Korrektur genannt. Eine Möglichkeit sieht vor, dass dem Benutzer die gesamte erkannte Nachricht vorgespielt wird und dieser – sofern das Erkennungsergebnis nicht der tatsächlich gesprochenen Nachricht entspricht – die Nachricht noch einmal spricht. Dieses Verfahren ist insbesondere unter solchen Bedingungen, bei denen die Erkennungs-Fehlerquote relativ hoch ist – beispielsweise bei Aufnahme eines Textes unter vielen Nebengeräuschen – nur wenig zufriedenstellend, da der Benutzer gegebenenfalls mehrfach die komplette Nachricht noch einmal sprechen muss, um letztlich das gewünschte Ergebnis zu erzielen. Bei einer zweiten Variante werden während der Analyse des Sprachsignals automatisch für bestimmte Abschnitte des Sprachsignals jeweils Gewissheitsfaktoren ermittelt. Anschließend werden dann nur die Ab-

schnitte des Textes wieder an den Benutzer ausgegeben, welche einen geringen Gewissheitsfaktor haben, d. h. bei denen die Wahrscheinlichkeit, dass ein Fehler aufgetreten ist, am höchsten ist. Auf diese Weise ist jedoch eine vollständige Überprüfung des Textes nicht möglich. Bei einer dritten Variante ist vorgesehen, den Text in Abschnitten, beispielsweise wort- oder satzweise vorzuspielen und jeweils am Ende eines Abschnitts abzuwarten, wobei der Benutzer dann die Gelegenheit hat, jeden einzelnen Abschnitt individuell zu bestätigen oder abzulehnen, beispielsweise durch die Worte „ja" oder „nein". Wenn der Benutzer während der Pause eine längere Zeit schweigt, so wird dies als Zustimmung interpretiert. Sofern der Benutzer einen ausgegebenen Abschnitt ablehnt, hat er die Gelegenheit, diesen kompletten Abschnitt neu zu sprechen.

[0005] Diese dritte Variante ist zwar für den Benutzer schon erheblich zeitsparender und komfortabler als die erste Variante, bei der der Text komplett neu gesprochen werden muss. Dennoch besteht auch hier der Nachteil, dass der Benutzer insbesondere unter schwierigen Erkennungsbedingungen, bei denen eine höhere Fehlerquote auftritt, gegebenenfalls mehrfach den zu korrigierenden Abschnitt neu sprechen muss. Ein weiteres Problem dieser Methode tritt dann auf wenn beispielsweise bei einer besonders außergewöhnlichen Aussprache eines Textteils durch den Benutzer – z. B. wegen eines Dialekts des Benutzers – das Spracherkennungssystem nicht über die hierfür optimalen Modelle verfügt und daher auch bei mehrmaligem Sprechen immer wieder ein falsches Erkennungsergebnis als wahrscheinlichstes Erkennungsergebnis erhält.

[0006] Es ist Aufgabe der vorliegenden Erfindung, ein Verfahren zur Spracherkennung und ein Spracherkennungssystem der eingangs genannten Art dahingehend zu verbessern, dass die Korrektur eines falsch verstandenen Sprachsignals einfacher, schneller und für den Benutzer komfortabler durchgeführt werden kann.

[0007] Diese Aufgabe wird dadurch gelöst, dass bei der Analyse gleich eine Anzahl von alternativen Erkennungsergebnissen, d. h. mindestens eine Alternative generiert wird, welche mit den nächst größten Wahrscheinlichkeiten zu dem zu erkennenden Sprachsignal passen. Die Ausgabe innerhalb der Prüfprozedur erfolgt dabei derart, dass der Benutzer bei einer Fehlerhaftigkeit des ausgegebenen Erkennungsergebnisses die Ausgabe unterbrechen kann. Es werden dann automatisch für einen vor einer Unterbrechung zuletzt ausgegebenen Abschnitt des betreffenden Erkennungsergebnisses jeweils die entsprechenden Abschnitte der alternativen Erkennungsergebnisse – ebenfalls in Sprachform – für eine Auswahl durch den Benutzer ausgegeben. Anschließend wird der betreffende Abschnitt im ausgegebenen Erkennungsergebnis auf Basis des vom Benutzer ausgewählten Abschnitts eines der alternativen Erkennungsergebnisse korrigiert. Schließlich wird die Prüfprozedur für die verbleibenden nachfolgenden Abschnitte des zu erkennenden Sprachsignals fortgesetzt.

[0008] Bei diesem Verfahren wird ausgenutzt, dass ohnehin von der Spracherkennungseinrichtung mehrere alternative Erkennungsergebnisse bezüglich ihrer Wahrscheinlichkeiten überprüft werden müssen, um das wahrscheinlichste Erkennungsergebnis zu ermitteln. Anstatt während der Analyse die unwahrscheinlicheren Ergebnisse wieder zu verwerfen, generiert hier die Spracherkennungseinrichtung die n-besten Sätze bzw. Worthypothesengraphen als Erkennungsergebnis-Alternativen und hinterlegt diese beispielsweise in einem Zwischenspeicher für die weitere Prüfprozedur. Der Mehraufwand für die Spracherkennungseinrichtung ist hierbei nur sehr gering. Während der Prüfprozedur können dann diese zusätzlichen Informationen dazu verwendet werden, um dem jeweiligen Benutzer alternative Angebote für den falsch erkannten Abschnitt des Erkennungsergebnisses zu machen. Da sich in vielen Fällen die Wahrscheinlichkeiten der verschiedenen Alternativen nur geringfügig unterscheiden, besteht oft eine relativ hohe Wahrscheinlichkeit, dass der Benutzer unter den Alternativen das richtige Erkennungsergebnis findet. Er kann diese richtige Alternative dann auf einfache Weise auswählen, ohne dass er den Textabschnitt neu sprechen muss. Die Gefahr, dass der zur Korrektur neu gesprochene Textabschnitt noch einmal falsch erkannt wird, besteht somit nicht mehr.

[0009] Die Ausgabe des Erkennungsergebnisses während der Prüfprozedur kann derart erfolgen, dass immer nach bestimmten Abschnitten eine kurze Pause gemacht und in diesen Pausen geprüft wird, ob der Benutzer beispielsweise durch die Worte „stopp" oder „nein" den letzten Abschnitt des Erkennungsergebnisses ablehnt. Vorzugsweise wird jedoch während der Ausgabe des Erkennungsergebnisses permanent die Sprachaktivität des Benutzers überwacht. Sobald der Benutzer in die Ausgabe hinein einen Kommentar abgibt, wird die Ausgabe unterbrochen. D. h. es wird ein sogenanntes „Barge-In-Verfahren" genutzt. Auf diese Weise kann auf unnötige Pausen während der Ausgabe verzichtet werden, so dass die Prüfprozedur sehr schnell abgeschlossen werden kann.

[0010] Um zu vermeiden, dass auch in den Fällen, in denen der Benutzer während der Sprachausgabe eine Äußerung gemacht hat, die zu einer Unterbrechung der Ausgabe geführt hat, obwohl sie an sich nicht dazu dienen sollte, die Sprachausgabe des Erkennungsergebnisses zu unterbrechen, sondern die beispielsweise anderen Personen im Raum galt, ist vorgesehen, dass der Benutzer durch Sprechen eines bestimmten Befehls wie z. B. „weiter" sofort die Ausgabe fortsetzen kann, ohne sich zunächst die verschiedenen alternativen Erkennungsergebnisse anzuhören.

[0011] Bei einem besonders bevorzugten Verfahren

wird, sofern der Benutzer keinen Abschnitt der alternativen Erkennungsergebnisse auswählt, weil beispielsweise alle Erkennungsergebnisse falsch waren, ein Anforderungssignal an den Benutzer ausgegeben, damit dieser den betreffenden Abschnitt für eine Korrektur neu spricht.

[0012] Für die Auswahl der ausgegebenen alternativen Erkennungsergebnisse bestehen verschiedene Möglichkeiten.

[0013] Bei einer ersten Variante werden die Erkennungsergebnisse der Reihe nach ausgegeben und anschließend wird jeweils abgewartet, ob der Benutzer das Erkennungsergebnis bestätigt. Im Falle einer Bestätigung wird das alternative Erkennungsergebnis als richtig akzeptiert. Anderenfalls wird das nächste alternative Erkennungsergebnis ausgegeben.

[0014] Bei einer zweiten Variante werden alle alternativen Erkennungsergebnisse bzw. die betreffenden Abschnitte der alternativen Erkennungsergebnisse kontinuierlich aufeinanderfolgend ausgegeben und der Benutzer wählt anschließend das passende Erkennungsergebnis aus. Vorzugsweise wird dabei jedes alternative Erkennungsergebnis gemeinsam mit einem Indikator, beispielsweise einer Ziffer oder einem Buchstaben, ausgegeben, welcher dem jeweiligen Erkennungsergebnis zugeordnet ist. Die Auswahl des betreffenden Abschnitts der verschiedenen alternativen Erkennungsergebnisse kann dann durch eine Eingabe des Indikators erfolgen, indem der Benutzer einfach beispielsweise die betreffende Ziffer oder den Buchstaben spricht.

[0015] Bei einem weiteren bevorzugten Ausführungsbeispiel ist dem Indikator ein Tastensignal eines Kommunikationsendgeräts, beispielsweise ein DTMF-Signal eines Telefongeräts, zugeordnet. Die Auswahl eines der Abschnitte erfolgt dann durch Betätigung der entsprechenden Taste des Kommunikationsendgeräts. Dies hat den Vorteil, dass die Auswahl des Erkennungsergebnisses ohne Zwischenschaltung einer erneuten Spracherkennung erfolgt und dadurch bedingte weitere mögliche Fehler ausgeschlossen werden.

[0016] Alternativ kann auch bei der Ausgabe der alternativen Erkennungsergebnisse ein Barge-In-Verfahren verwendet werden. D. h. es werden dann die Abschnitte der alternativen Erkennungsergebnisse ohne Pause hintereinander ausgegeben und der Benutzer sagt einfach „stopp" oder „ja" oder dgl., wenn das richtige Erkennungsergebnis ausgegeben wird.

[0017] Bei einem besonders bevorzugten Ausführungsbeispiel werden nach einer Korrektur eines Abschnitts die verschiedenen Erkennungsergebnisse bezüglich ihrer Wahrscheinlichkeiten, mit denen sie jeweils zu dem erkennenden Sprachsignal passen, unter Berücksichtigung des korrigierten Abschnitts sowie aller zuvor bereits bestätigten oder korrigierten Abschnitte neu bewertet. Die Prüfprozedur wird dann mit der Ausgabe der nachfolgenden Abschnitte des Erkennungsergebnisses fortgesetzt, welches nach der Neubewertung die höchste Wahrscheinlichkeit aufweist. Durch die Neubewertung auf Basis aller bereits korrigierten bzw. bestätigten Teile des zu erkennenden Sprachsignals kann bei einer kontextabhängigen Wahrscheinlichkeitsanalyse das Erkennungsergebnis noch im Lauf der Prüfprozedur permanent verbessert werden und damit die Wahrscheinlichkeiten für notwendige Korrekturen in nachfolgenden Abschnitten vermindert werden.

[0018] Sofern längere Texte bzw. Nachrichten erkannt werden sollen, gibt es für die Durchführung der Prüfprozedur verschiedene Möglichkeiten.

[0019] Bei einer Variante erfolgt die Prüfprozedur erst nach Eingabe eines vollständigen Textes durch den Benutzer. Dass der gewünschte Text vollständig gesprochen wurde, kann beispielsweise durch den Benutzer mittels eines entsprechenden Befehls wie „Ende" oder dergl. signalisiert werden.

[0020] Bei einer anderen Variante erfolgt die Prüfprozedur jeweils bereits nach Eingabe eines Teils eines vollständigen Textes. Dies hat den Vorteil, dass bereits verifizierte bzw. korrigierte Teile des Textes gegebenenfalls in anderen Komponenten der Applikation weiter verarbeitet oder in einem Speicher hinterlegt werden können, ohne dass das Spracherkennungssystem hierdurch noch belastet wird. So kann beispielsweise immer dann eine Prüfprozedur für einen zuvor eingegebenen Textteil erfolgen, sobald eine bestimmte Länge des Textteils bzw. Sprachsignals erreicht ist und/oder wenn eine Sprechpause mit einer bestimmten Länge vorliegt und/oder wenn der Benutzer dies mit einem besonderen Befehl vorgibt.

[0021] Ein erfindungsgemäßes Spracherkennungssystem muss zur Durchführung des erfindungsgemäßen Verfahrens eine Spracherkennungseinrichtung aufweisen, die derart ausgebildet ist, dass sie bei der Analyse eine Anzahl von alternativen Erkennungsergebnissen generiert und ausgibt bzw. speichert, die jeweils – bezogen auf das mit der größten Wahrscheinlichkeit passende, ohnehin ausgegebene Erkennungsergebnis – mit den nächstgrößten Wahrscheinlichkeiten zu dem zu erkennenden Sprachsignal passen. Darüber hinaus benötigt das Spracherkennungssystem Mittel zur Unterbrechung der Ausgabe innerhalb der Prüfprozedur durch den Benutzer sowie eine Dialog-Steuereinrichtung, welche automatisch für einen vor einer Unterbrechung zuletzt ausgegebenen Abschnitt des betreffenden Erkennungsergebnisses jeweils die entsprechenden Abschnitte der alternativen Erkennungsergebnisse ausgibt. Weiterhin muss das Spracherkennungssystem Mittel zur Auswahl eines der ausgegebenen Abschnitte der alternativen Erkennungsergebnisse sowie eine Korrektureinrichtung zur Korrektur des betreffenden Abschnitts im zunächst ausgegebenen Erkennungsergebnis auf Basis des entsprechenden Abschnitts des ausgewählten alternativen Erkennungsergebnisses aufweisen.

[0022] Sofern die Auswahl des alternativen Erkennungsergebnisses mittels eines Tastensignals eines

Kommunikationsendgeräts erfolgen soll, ist es notwendig, dass das Spracherkennungssystem auch eine entsprechende Schnittstelle aufweist, um ein solches Tastensignal zu empfangen, zu erkennen und zur Auswahl eines der ausgegebenen Abschnitte zu verwenden.

[0023] Das erfindungsgemäße Spracherkennungssystem kann vorzugsweise im Wesentlichen mittels geeigneter Software auf einem Computer bzw. in einer Sprachsteuerung eines Geräts realisiert werden. So können z. B. die Spracherkennungseinrichtung und die Dialog-Steuereinrichtung vollständig in Form von Softwaremodulen realisiert werden. Auch eine Einrichtung zur Generierung von Sprache anhand von computerlesbaren Texten, beispielsweise ein sogenannter TTS-Konverter (Text-to-Speech-Konverter), ist ebenfalls mittels geeigneter Software realisierbar. Es ist lediglich erforderlich, dass das System eine Möglichkeit zur Spracheingabe, beispielsweise ein Mikrofon mit einem entsprechenden Verstärker, und zur Sprachausgabe, beispielsweise einen Lautsprecher mit einem entsprechenden Verstärker, umfasst.

[0024] Dabei kann sich das Spracherkennungssystem auch auf einem über ein übliches Kommunikationsnetz, beispielsweise ein Telefonnetz oder das Internet, erreichbaren Server befinden. In diesem Fall reicht es aus, wenn sich die Spracheingabeeinrichtung und Sprachausgabeeinrichtung, d. h. Mikrofon, Lautsprecher und entsprechende Verstärker, in einem Kommunikationsendgerät des Benutzers befinden, das über das betreffende Netz mit dem Server des Spracherkennungssystems verbunden ist. Weiterhin ist es auch möglich, dass das Spracherkennungssystem nicht innerhalb eines einzelnen Geräts, beispielsweise auf einem einzelnen Server, realisiert ist. Stattdessen können verschiedene Komponenten des Systems auch an verschiedenen Orten angeordnet sein, welche über ein entsprechendes Netzwerk untereinander verbunden sind. Das erfindungsgemäße Spracherkennungssystem kann einer ganz bestimmten Applikation zugeordnet sein, beispielsweise einer Anwendung, welche VoiceMail-Nachrichten innerhalb eines Kommunikationssystems in SMS-Nachrichten oder E-Mails umwandelt. Es ist aber auch möglich, dass das Spracherkennungssystem mehreren verschiedenen Anwendungen als dienstleistendes System zur Verfügung steht und so für mehrere Applikationen eine Schnittstelle zu den Benutzern der jeweiligen Applikation bildet.

[0025] Die Erfindung wird im Folgenden unter Hinweis auf die beigefügten Figuren anhand eines Ausführungsbeispiels näher erläutert. Es zeigen:

[0026] **Fig.** 1 ein schematisches Blockdiagramm für ein erfindungsgemäßes Spracherkennungssystem,

[0027] **Fig.** 2 ein Ablaufdiagramm zur Erläuterung des Korrekturverfahrens.

[0028] Das in **Fig.** 1 dargestellte Ausführungsbeispiel eines erfindungsgemäßen Spracherkennungssystems **1** weist einen Eingang **14** auf, an den ein Mikrofon **2** über einen Verstärker **3** angeschlossen ist. Außerdem weist das Spracherkennungssystem **1** einen Ausgang **16** auf, an den über einen Verstärker **5** ein Lautsprecher **4** zur Ausgabe von Sprachsignalen angeschlossen ist. Das Mikrofon **2** mit dem zugehörigen Verstärker **3** sowie der Lautsprecher **4** mit dem zugehörigen Verstärker **5** sind hierbei Teil eines vom Spracherkennungssystem **1** entfernten Geräts, welches über ein Kommunikationsnetz, beispielsweise ein Telefonnetz, mit dem Spracherkennungssystem **1** in Verbindung steht.

[0029] Das Kommunikationsendgerät weist außerdem eine Tastatur **6** auf, über die akustische Signale, beispielsweise DTMF-Signale (Dual Tone Multi Frequency), erzeugt werden können, die ebenfalls über den Sprachsignalkanal zum Eingang **14** des Spracherkennungssystems übertragen werden.

[0030] Vom Mikrofon **2** über den Verstärker **3** am Eingang **14** ankommende Sprachsignale S_l werden vom Spracherkennungssystem **1** in einen lesbaren bzw. maschinenlesbaren Text umgewandelt und an eine Applikation **15**, beispielsweise zur Versendung von SMS oder E-Mail weitergeleitet, welche dann die Textdaten entsprechend bearbeitet und/oder weiter versendet.

[0031] Eingangsseitig gelangt das akustisches Signal hierzu zunächst zu einem sog. „Voice-Activity-Detector" (VAD) **12**, der das ankommende Signal nur daraufhin überprüft, ob tatsächlich ein Sprachsignal S_l eines Benutzers ankommt oder ob es sich bei dem Signal nur um Hintergrundgeräusche etc. handelt. Das Sprachsignal S_l wird dann an eine Spracherkennungseinrichtung **7** weitergeleitet, die das Sprachsignal S_l zur Erkennung einer darin enthaltenen Sprachinformation in üblicher Weise analysiert und ein mit größter Wahrscheinlichkeit passendes Erkennungsergebnis ermittelt.

[0032] Erfindungsgemäß ist die Spracherkennungseinrichtung **7** hierbei so ausgerüstet, dass zusätzlich zu dem Erkennungsergebnis, welches mit größter Wahrscheinlichkeit zum zu erkennenden Sprachsignal S_l passt, auch eine Anzahl von alternativen Erkennungsergebnissen generiert wird, welche mit den nächstgrößten Wahrscheinlichkeiten zum zu erkennenden Sprachsignal S_l passen.

[0033] Das Erkennungsergebnis, das mit größter Wahrscheinlichkeit zu dem zu erkennenden Sprachsignal S_l passt, wird dann in Textform an eine Dialog-Steuereinrichtung **10** übermittelt, welche dieses wahrscheinlichste Erkennungsergebnis wieder an einen Text-To-Speech-Generator (TTS-Generator) **9** weiterleitet. Die alternativen Erkennungsergebnisse können ebenfalls sofort an die Dialog-Steuereinrichtung **10** weitergeleitet und dort zwischengespeichert werden oder von der Spracherkennungseinrichtung **7** in einem separaten Speicher **8** hinterlegt werden, auf welchen die Dialog-Steuereinrichtung **10** jederzeit Zugriff hat. Mit Hilfe des TTS-Generators **9** wird das wahrscheinlichste Erkennungsergebnis dann in ein Sprachsignal umgewandelt und innerhalb einer Prüf-

prozedur zur Verifikation und/oder Korrektur durch den Benutzer über den Verstärker **5** und den Lautsprecher **4** in Sprachform ausgegeben.

[0034] Der genaue Ablauf dieser Prüfprozedur wird im Folgenden anhand von **Fig.** 2 erläutert.

[0035] Das Verfahren beginnt zunächst in Verfahrensschritt I mit der bereits beschriebenen Spracheingabe. Anschließend werden in Verfahrensschritt II die verschiedenen alternativen Erkennungsergebnisse ermittelt und schließlich im Verfahrensschritt III bewertet, um festzustellen, welches Erkennungsergebnis am besten zu dem zu erkennenden Sprachsignal S_I passt. Anschließend erfolgt in Verfahrensschritt IV abschnittsweise die Ausgabe des wahrscheinlichsten Erkennungsergebnisses, wobei diese abschnittsweise Ausgabe kontinuierlich erfolgt, so dass für den Benutzer die einzelnen Abschnitte an sich nicht erkennbar sind. Bei den einzelnen Abschnitten kann es sich beispielsweise um die einzelnen Wörter eines Satzes oder eines Worthypothesengraphs oder auch um Satzteile bzw. Teile eines Worthypothesengraphs handeln.

[0036] Nach jedem Abschnitt wird in Verfahrensschritt V geprüft, ob ein Abbruch der Ausgabe durch den Benutzer erfolgt. Dies ist beispielsweise möglich, indem sich der Benutzer während der Ausgabe des Erkennungsergebnisses entsprechend äußert. Die Sprachaktivität des Benutzers wird von dem VAD **12** sofort erkannt, welcher über ein entsprechendes Steuersignal S_C den TTS-Generator **9** stoppt und gleichzeitig das Steuersignal S_C auch an die Dialog-Steuereinrichtung **10** übermittelt, so dass diese ebenfalls den Abbruch der Ausgabe durch den Nutzer registriert. Erfolgt kein Abbruch, so wird dann geprüft, ob das Ende des eingegebenen Textes erreicht ist (Verfahrensschritt VI). Ist dies der Fall, so gilt das Erkennungsergebnis als vom Benutzer verifiziert und das Erkennungsergebnis wird an die Applikation **15** übergeben (Verfahrensschritt VII). Ist das Ende des Textes noch nicht erreicht, so wird die Ausgabe des wahrscheinlichsten Erkennungsergebnisses fortgesetzt.

[0037] Wird dagegen in Verfahrensschritt V ein Abbruch registriert, so wird in Verfahrensschritt VIII zunächst ermittelt, um welchen falschen Abschnitt es sich handelt. Der Einfachheit halber wird hier angenommen, dass es sich um den Abschnitt handelt, der zuletzt ausgegeben wurde, unmittelbar bevor die Ausgabe vom Benutzer unterbrochen wurde.

[0038] Die Dialog-Steuereinrichtung **10** greift dann – sofern die alternativen Erkennungsergebnisse nicht innerhalb der Dialog-Steuereinrichtung **10** selbst zwischengespeichert wurden – auf den Zwischenspeicher **8** zu und ermittelt die entsprechenden Abschnitte der alternativen Erkennungsergebnisse, welche zu dem in Verfahrensschritt VIII ermittelten falschen Abschnitt korrespondieren. Den entsprechenden Abschnitten bzw. den alternativen Erkennungsergebnissen werden dann Indikatoren, beispielsweise die Ziffern **1** bis **0**, zugeordnet.

[0039] Über den TTS-Generator **9** werden dann die zur Verfügung stehenden alternativen Abschnitte jeweils gemeinsam mit den zugehörigen Indikatoren in Sprachform an den Benutzer ausgegeben (Verfahrensschritt IX).

[0040] In Verfahrensschritt X kann dann der Benutzer schließlich einen passenden Abschnitt der alternativen Erkennungsergebnisse auswählen, indem er eine dem Indikator entsprechende Zifferntaste eines Tastenfelds **6** drückt. Durch Druck auf diese Taste wird ein DTMF-Signal erzeugt, welches über den Sprachkanal an den Eingang **14** des Spracherkennungssystems **1** geleitet wird. Dieses DTMF-Signal wird dann von einem DTMF-Erkenner **13** erkannt, welcher parallel zur Spracherkennungseinrichtung **7** geschaltet ist. Der DTMF-Erkenner **13** gibt ein entsprechendes Auswahlsignal S_A an die Dialog-Steuereinrichtung **10** aus, welche dann eine Korrektureinheit **11** veranlasst, den falsch erkannten Abschnitt durch den betreffenden Abschnitt des ausgewählten alternativen Erkennungsergebnisses zu ersetzen (Verfahrensschritt XI). Die DTMF-Erkennungseinheit **13** kann außerdem bei Erkennung eines DTMF-Signals ein Signal an die Spracherkennungseinrichtung **7** übermitteln, damit diese beispielsweise außer Kraft gesetzt wird und nicht unnötigerweise versucht, das DTMF-Signal zu analysieren.

[0041] Nach erfolgter Korrektur wird in Verfahrensschritt XII eine Neubewertung aller Erkennungsergebnisse, d. h. des wahrscheinlichsten Erkennungsergebnisses und der alternativen Erkennungsergebnisse durchgeführt. Diese Neubewertung erfolgt vorzugsweise in der Spracherkennungseinrichtung **7**, welche ebenfalls in der Lage ist, auf den Zwischenspeicher **8** zuzugreifen bzw. welche die notwendigen Daten hierzu von der Dialog-Steuereinrichtung **10** erhält. Bei dieser kontextabhängigen Neubewertung der Erkennungsergebnisse werden alle bereits verifizierten bzw. korrigierten Abschnitte berücksichtigt, d. h. es wird die Tatsache berücksichtigt, dass für diese betreffenden Abschnitte jeweils die Wahrscheinlichkeit 100 % ist, und für alle alternativen Abschnitte dagegen die Wahrscheinlichkeit bei 0 % liegt. Auf diese Weise kann es beispielsweise erreicht werden, dass auf Basis der bereits bekannten Abschnitte solche Hypothesen, die ohne dieses Vorwissen eine hohe Wahrscheinlichkeit haben, verworfen werden und dafür andere Hypothesen, welche ursprünglich eine geringe Wahrscheinlichkeit hatten, nun sehr wahrscheinlich werden. Dadurch wird die Fehlerquote bei der Ausgabe der nachfolgenden Abschnitte deutlich reduziert und somit das gesamte Korrekturverfahren beschleunigt. Zusätzlich oder alternativ können die bereits sicher erkannten Teile der Äußerung des Nutzers auch für eine Adaption der Sprachmodelle und/oder der akustischen Modelle herangezogen werden.

[0042] Es wird noch einmal darauf hingewiesen, dass es sich bei dem vorbeschriebenen Spracherkennungssystem bzw. Verfahrensablauf nur um ein

spezielles Ausführungsbeispiel der Erfindung handelt und der Fachmann die Möglichkeit hat, das Spracherkennungssystem und das Verfahren auf verschiedene Weisen zu modifizieren. So ist es z. B. insbesondere möglich und auch sinnvoll, innerhalb des Verfahrens einen Schritt einzuführen, dass der Benutzer, sofern er keinen der Abschnitte der alternativen Erkennungsergebnisse für richtig hält, die Gelegenheit erhält, den Abschnitt neu zu sprechen. Ebenso ist es auch möglich, dass anstelle der Auswahl mittels einer DTMF-fähigen Tastatur **6** die Auswahl mittels Spracheingabe erfolgt oder dass die Tastatur andere Signale aussendet, welche über einen separaten Datenkanal an das Spracherkennungssystem **1** übermittelt werden, das diese Signale dann entsprechend weiter bearbeiten kann. Ebenso kann auch der Abbruch der Sprachausgabe innerhalb der Prüfprozedur mittels eines bestimmten DTMF-Signals oder dergl. erfolgen.

Patentansprüche

1. Verfahren zur Spracherkennung, bei welchem ein Sprachsignal eines Benutzers zur Erkennung einer im Sprachsignal enthaltenen Sprachinformation analysiert wird und ein mit einer größten Wahrscheinlichkeit passendes Erkennungsergebnis innerhalb einer Prüfprozedur wieder in ein Sprachsignal umgewandelt und zur Verifikation und/oder Korrektur an den Nutzer ausgegeben wird,
dadurch gekennzeichnet,
dass bei der Analyse eine Anzahl von alternativen Erkennungsergebnissen generiert wird, welche mit den nächstgrößten Wahrscheinlichkeiten zu dem zu erkennenden Sprachsignal passen,
und dass die Ausgabe innerhalb der Prüfprozedur derart erfolgt, dass der Benutzer bei einer Fehlerhaftigkeit des ausgegebenen Erkennungsergebnisses die Ausgabe unterbrechen kann, und dann automatisch für einen vor einer Unterbrechung zuletzt ausgegebenen Abschnitt des betreffenden Erkennungsergebnisses jeweils entsprechende Abschnitte der alternativen Erkennungsergebnisse für eine Auswahl durch den Benutzer ausgegeben werden, und schließlich der betreffende Abschnitt im ausgegebenen Erkennungsergebnis auf Basis des entsprechenden Abschnitts eines ausgewählten alternativen Erkennungsergebnisses korrigiert wird und dann die Prüfprozedur für verbleibende nachfolgende Abschnitte des zu erkennenden Sprachsignals fortgesetzt wird.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass bei der Ausgabe des Erkennungsergebnisses innerhalb der Prüfprozedur die Sprachaktivität des Benutzers permanent überwacht wird und bei Empfang eines Sprachsignals des Benutzers die Ausgabe unterbrochen wird.

3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, dass, falls kein Abschnitt der alternativen Erkennungsergebnisse ausgewählt wird, ein Anforderungssignal an den Benutzer ausgegeben wird, den betreffenden Abschnitt für eine Korrektur neu zu sprechen.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, dass jedem alternativen Erkennungsergebnis ein Indikator zugeordnet wird und bei der Prüfprozedur die betreffenden Abschnitte der alternativen Erkennungsergebnisse jeweils gemeinsam mit dem zugehörigen Indikator ausgegeben werden und die Auswahl eines Abschnitts eines alternativen Erkennungsergebnisses durch eine Eingabe des Indikators erfolgt.

5. Verfahren nach Anspruch 4, dadurch gekennzeichnet, dass der Indikator eine Ziffer oder ein Buchstabe ist.

6. Verfahren nach Anspruch 4 oder 5, dadurch gekennzeichnet, dass dem Indikator ein Tastensignal eines Kommunikationsendgeräts zugeordnet ist und die Auswahl eines Abschnitts eines alternativen Erkennungsergebnisses durch Betätigung der entsprechenden Taste des Kommunikationsendgeräts erfolgt.

7. Verfahren nach einem der Ansprüche 1 bis 6, dadurch gekennzeichnet, dass nach einer Korrektur eines innerhalb der Prüfprozedur ausgegebenen Abschnitts die verschiedenen Erkennungsergebnisse bezüglich ihrer Wahrscheinlichkeiten, mit denen sie jeweils zu dem zu erkennenden Sprachsignal passen, unter Berücksichtigung des zuletzt korrigierten Abschnitts und/oder der bereits zuvor bestätigen oder korrigierten Abschnitte neu bewertet werden und die Prüfprozedur mit der Ausgabe der nachfolgenden Abschnitte des Erkennungsergebnisses fortgesetzt wird, welches nach der Neubewertung die höchste Wahrscheinlichkeit aufweist.

8. Verfahren nach einem der Ansprüche 1 bis 7, dadurch gekennzeichnet, dass die Prüfprozedur erst nach Abschluss der Eingabe eines vollständigen Texts durch den Benutzer erfolgt.

9. Verfahren nach einem der Ansprüche 1 bis 7, dadurch gekennzeichnet, dass die Prüfprozedur bereits nach Eingabe eines Teils eines vollständigen Texts durch den Benutzer erfolgt.

10. Spracherkennungssystem (**1**) mit
– einer Vorrichtung (**2**) zur Erfassung eines Sprachsignals eines Benutzers
– einer Spracherkennungseinrichtung (**7**), um das erfasste Sprachsignal (S_l) zur Erkennung einer im Sprachsignal (S_l) enthaltenen Sprachinformation zu analysieren und ein mit einer größten Wahrscheinlichkeit passendes Erkennungsergebnis zu ermitteln,

– und einer Sprachausgabeeinrichtung (**9**), um das wahrscheinlichste
Erkennungsergebnis innerhalb einer Prüfprozedur wieder in eine Sprachinformation umzuwandeln und zur Verifikation und/oder Korrektur an den Nutzer auszugeben,
dadurch gekennzeichnet,
dass die Spracherkennungseinrichtung (**7**) derart ausgebildet ist, dass sie bei der Analyse eine Anzahl von alterativen Erkennungsergebnissen generiert, welche mit den nächstgrößten Wahrscheinlichkeiten zu dem zu erkennenden Sprachsignal (S_I) passen, und dass das Spracherkennungssystem (**1**)
– Mittel (**12**) zur Unterbrechung der Ausgabe innerhalb der Prüfprozedur durch den Benutzer,
– eine Dialog-Steuereinrichtung (**10**), welche automatisch für einen vor einer Unterbrechung zuletzt ausgegebenen Abschnitt des betreffenden Erkennungsergebnisses jeweils entsprechende Abschnitte der alternativen Erkennungsergebnisse ausgibt,
– Mittel (**6**, **13**) zur Auswahl eines der ausgegebenen Abschnitte der alternativen Erkennungsergebnisse
– und eine Korrektureinheit (**11**) zur Korrektur des betreffenden Abschnitts im zunächst ausgegebenen Erkennungsergebnis auf Basis des entsprechenden Abschnitts eines ausgewählten alternativen Erkennungsergebnisses
aufweist.

11. Computerprogrammprodukt mit Programmcode-Mitteln, um alle Schritte eines Verfahrens nach einem der Ansprüche 1 bis 9 auszuführen, wenn das Programm auf einem Computer ausgeführt wird.

Es folgen 2 Blatt Zeichnungen

Anhängende Zeichnungen

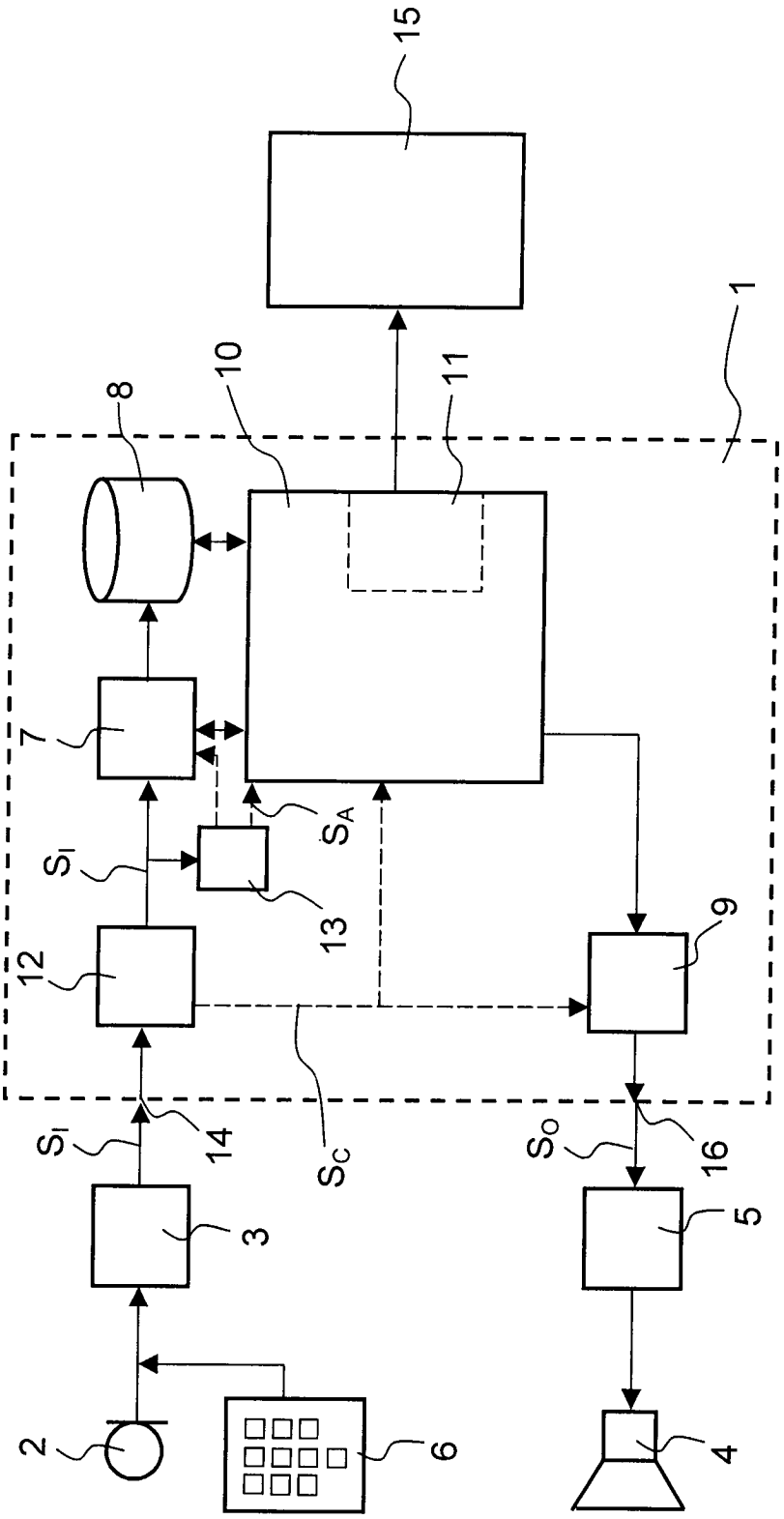


Fig. 1

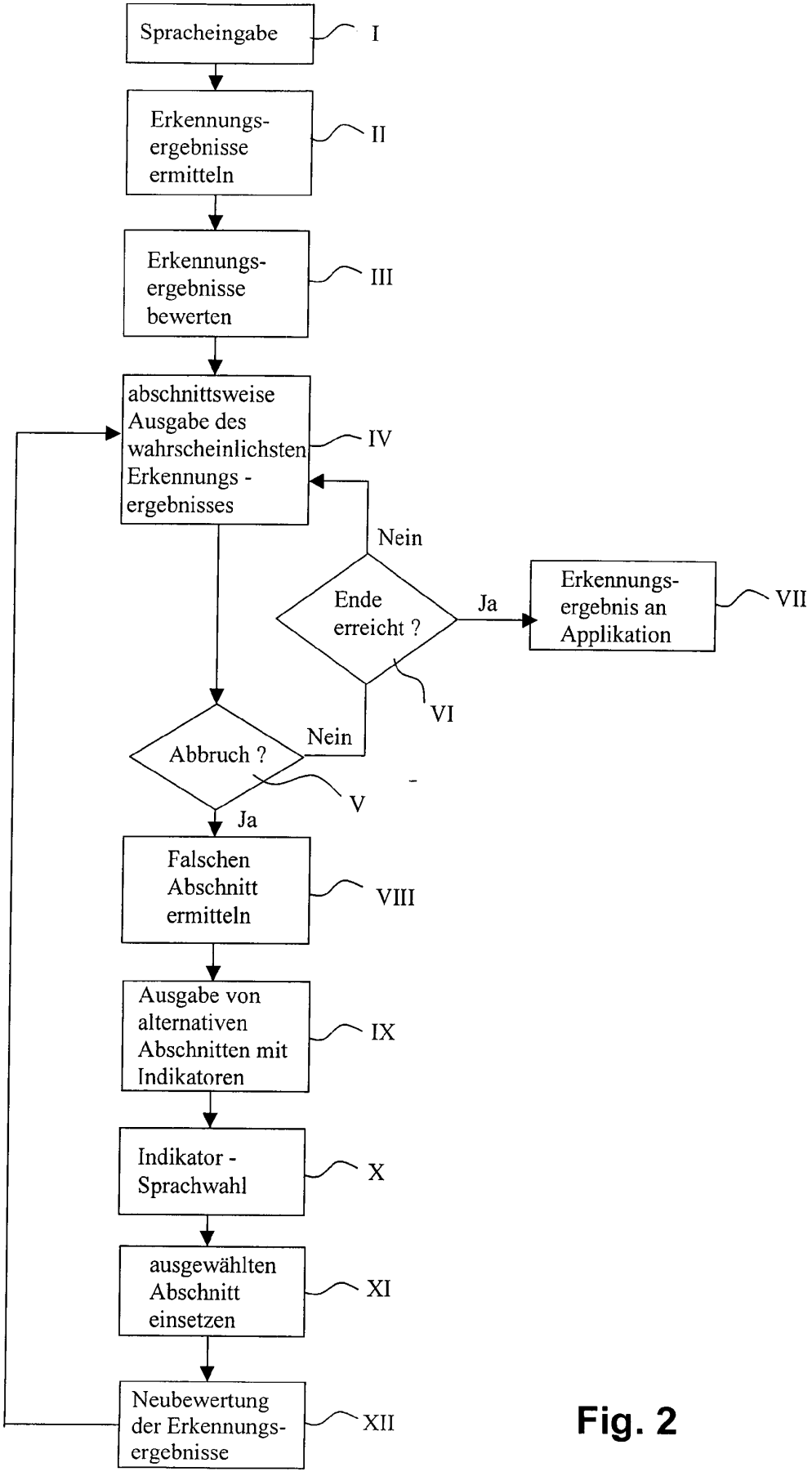

Spracheingabe
I
Erkennungs-ergebnisse ermitteln
II
Erkennungs-ergebnisse bewerten
III
abschnittsweise Ausgabe des wahrscheinlichsten Erkennungs - ergebnisses
IV
Nein
Ende erreicht ?
Ja
Erkennungs-ergebnis an Applikation
VII
VI
Abbruch ?
Nein
V
Ja
Falschen Abschnitt ermitteln
VIII
Ausgabe von alternativen Abschnitten mit Indikatoren
IX
Indikator - Sprachwahl
X
ausgewählten Abschnitt einsetzen
XI
Neubewertung der Erkennungs-ergebnisse
XII

Fig. 2