(19) Republik Österreich Patentamt

(10) Nummer: AT 006 921 U1

(12) GEBRAUCHSMUSTERSCHRIFT

(21) Anmeldenummer: GM 805/02

(22) Anmeldetag: 28.11.2002

(42) Beginn der Schutzdauer: 15. 4.2004

(45) Ausgabetag: 25. 5.2004

(51) Int.Cl.[7] : G10L 15/22
G06F 3/16

(73) Gebrauchsmusterinhaber:

SAIL LABS TECHNOLOGY AG
A-1040 WIEN (AT).

(72) Erfinder:

PFANNERER NORBERT DIPL.ING.
WIEN (AT).
BACKFRIED GERHARD DIPL.ING.
PURKERSDORF, NIEDERÖSTERREICH (AT).

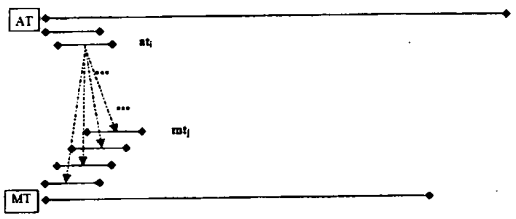
(54) VERFAHREN ZUR AUTOMATISCHEN ÜBEREINSTIMMUNG VON AUDIO-SEGMENTEN MIT TEXTELEMENTEN

(57) Verfahren zur automatischen Übereinstimmung von Audio-Segmenten mit Textelementen in einem manuell aus der Audioaufnahme erzeugten Transkript (MT), wobei aus der Audio-Aufnahme ein automatisches Transkript (AT) erstellt wurde, das die zu Textelementen geformten Audiosegmente zusammen mit einem Zeitbezug enthält, weiters umfassend:

das Unterteilen des automatischen Transkripts (AT) und des manuellen Transkripts (MT) in Passagen (at_i, mt_j) definierter Länge, die jeweils mehrere Textelemente umfassen, das Verschieben jeder Passage im automatischen Transkript (AT) und im manuellen Transkript (MT) über das gesamte automatische und manuelle Transkript, wobei jede Passage sich mit der vorhergehenden Passage überlappt, und Ermitteln einer bestimmten Passagen-Eigenschaft für jede Passage,

das Vergleichen der Passagen-Eigenschaft jeder Passage (at_i) im automatischen Transkript (AT) mit jeder Passage (mt_j) im manuellen Transkript (MT),

das Zuordnen einer jeweiligen Passage im automatischen Transkript (AT) zu jener Passage im manuellen Transkript (MT), so dass sich über die Summe der Passagenvergleiche gesehen der optimale Pfad an Zuordnungen ergibt.



AT 006 921 U1

Die Erfindung betrifft die automatische Erkennung natürlicher Sprache. Im Detail handelt es sich dabei um ein neuartiges Verfahren zur automatischen Übereinstimmung von in einer Audioaufnahme enthaltenen Audio-Segmenten mit Textelementen in einem manuell aus der Audioaufnahme erzeugten Transkript, wobei zunächst aus der Audio-Aufnahme, vorzugsweise durch einen automatischen Spracherkenner, ein automatisches Transkript erstellt wird, das die zu Textelementen geformten Audiosegmente zusammen mit einem Zeitbezug, an welcher Stelle in der Audio-Aufnahme sich das jeweilige automatisch erstellte Textelement befindet, enthält.

Ein automatischer Spracherkenner kann aus eingegebenen Audiodaten ein automatisches Transkript erzeugen, welches den in den Audiodaten vorkommenden Wörtern entspricht. Die Audiodaten können dabei von einer Vielfalt an Quellen kommen, so z.B. aus Videoaufnahmen oder Audio-Clips. Das manuelle Transkript wird typischerweise von einem Transkriptionisten erstellt, der eine Audio-Aufnahme oder ein Stenogramm als Referenz verwendet. Das automatische Transkript wird mit dem manuellen Transkript mittels des Programmierverfahrens des *Dynamic Alignment* verglichen und einander entsprechende Passagen gefunden.

Das erfindungsgemäße Verfahren eignet sich jedoch im Prinzip gleichermaßen auch für Texte, die nicht von einem automatischen Spracherkenner produziert wurden.

Das Gebiet der multimedialen Datenverarbeitung hat in den letzten Jahren stark an Bedeutung gewonnen. Die Mengen an Aufnahmen, die zur Verarbeitung bereitstehen, hat nicht zuletzt Dank der sprunghaften Entwicklung in der Verarbeitungs- und Speicherkapazität enorm zugenommen. Immer mehr stellt sich aber das Problem, aus diesen riesigen Mengen an Daten die gewünschten und relevanten Informationen effizient zu extrahieren. Speziell im Bereich der Aufnahmen von Gerichtsverhandlungen, von Vorträgen oder von Konferenzen stellt die Extraktion relevanter Daten eine besondere Herausforderung dar. Dieser Herausforderung wird einerseits durch Automatisierung des Transkriptionsprozesses mittels automatischer Sprachverarbeitung begegnet, andererseits werden die Aufnahmen nach wie vor manuell transkribiert, da die Qualität automatischer Verfahren bislang nur in den wenigsten Fällen als ausreichend betrachtet wird. Die manuelle Transkription erlaubt ein verlässliches Auffinden von Information in textueller Form. Da beim manuellen Transkribieren aber in den seltensten Fällen annotiert wird, wann ein Wort oder ein Satz genau gesagt wurde, fehlt die zeitliche Verbindung vom Text zum multimedialen Medium. Man muss also, etwa um den exakten Wortlaut einer Zeugenaussage zu überprüfen oder um eine Aussage im Video ansehen zu können, sequentiell auf dem

Medium suchen (unter Zuhilfenahme des Textes). Dies ist natürlich umständlich und bei längeren Passagen äußerst zeitaufwändig.

Zur Lösung dieses Problems wurden bereits Verfahren entwickelt, um eine exakte Verknüpfung zwischen den transkribierten Wörtern und dem multimedialen Medium herzustellen. Diese Verknüpfung erlaubt eine punktgenaue Verbindung zwischen Text und Audio (oder Video), was einen direkten Zugriff gestattet und langwieriges Suchen überflüssig macht.

Dabei wird die Zeitinformation des automatisch erkannten Textes (die jedem erkannten Wort genau einen Zeitpunkt im zu Grunde liegenden Audio/Video zuweist) auf die manuell transkribierten Wörter übertragen. Dies erlaubt ein effizientes Auffinden der entsprechenden Audio- oder Videosequenzen ausgehend vom manuell transkribierten Text. Fig. 1 stellt den Gesamtprozess schematisch dar.

Die Transkription multimedialer Daten (bzw. der darin enthaltenen Audiodaten) stellt aber noch einen Technologie-Bereich dar, der sich im Moment an der Schwelle von der Forschung in den kommerziellen Sektor befindet. Bestehende Verfahren, wie etwa in EP 0 649 144 "Automatic indexing of audio using speech recognition", US 5,649,060 "Automatic indexing and aligning of audio and text using speech recognition" und US 6,076,059 "Method for aligning text with audio signals" geoffenbart, zielen auf die Lösung des hier beschriebenen Problems ab. Allerdings sind diese Verfahren auf einzelnen Wörtern basiert, was sie anfälliger für schlechte Erkennungsraten der automatischen Spracherkennung macht. Den aus den zitierten Patentschriften bekannten Verfahren ist gemein, dass sie zudem auf der Erkennung und dem Finden identischer Wörter (im Kontext) basieren und diese gefundenen Paare als „Ankerpunkte" verwenden.

Zum besseren Verständnis ist anzumerken, dass man in der Sprachverarbeitung unter der Bezeichnung des „*forced-alignment*", wie in US 6,076,059 verwendet, ein Verfahren versteht, das einen bereits bekannten Text mit einer Aufnahme in Einklang bringt (d.h. zwischen Text und Audio ein *alignment* herstellen soll). Dieses Verfahren ist allerdings mit einer Unzahl an Problemen behaftet: der transkribierte Text weist quasi niemals die dazu notwendige Genauigkeit in der Transkription auf. Besonders im Falle von überlagerten Störungen kommt es leicht zu Problemen mit dem Spracherkenner. Längere Stücke an Audio enthalten u.U. gar nicht das gesamte Transkript und auch die Länge der zu verwendenden Fenster kann schwer im voraus bestimmt werden. Es ist zu erwähnen, dass der Begriff „Fenster" im Zusammenhang mit „forced alignment" sich auf Fenster der Audio-Datei

bezieht. Man nimmt z.B. die nächsten 20 Sekunden der Audio-Datei (d.h. ein Fenster der Länge 20s) die nächsten zehn noch nicht im Prozess verwendeten Wörter des manuellen Transkripts und lässt den Spracherkenner die Zuordnung dieser Wörter zum enthaltenen Audio feststellen.

Ziel der vorliegenden Erfindung ist es, mittels eines neuartigen Verfahrens eine automatische Übereinstimmung zwischen einem automatisch (vorzugsweise mittels automatischer Spracherkennung) produzierten Text bzw. Transkript und einem manuell erzeugten Text bzw. Transkript herzustellen, wobei das Verfahren wesentlich robuster gegenüber Fehlern und Unvollständigkeiten im automatisch erzeugten Text sein soll als die bekannten Verfahren. Weiters soll das erfindungsgemäße Verfahren den Aufwand in der Verarbeitung des automatisch erzeugten Textes wesentlich mindern.

Zur Lösung dieser Aufgabe sieht die Erfindung ein Verfahren zur automatischen Übereinstimmung zwischen einem automatisch erzeugten Text bzw. Transkript und einem manuell erzeugten Text bzw. Transkript vor, wie in Anspruch 1 definiert.

Vorteilhafte Ausgestaltungen und Weiterbildungen dieses Verfahrens sind in den von Anspruch 1 abhängigen Ansprüchen definiert.

Anders als bei den bekannten Verfahren werden beim erfindungsgemäßen Verfahren nicht die einzelnen Wörter selbst, sondern ganze Text-Passagen, welche fensterartig (*sliding window*) und überlappend über den gesamten Text verschoben werden, verwendet. Die Passagen werden dabei durch Eigenschaften der ihnen entsprechenden Wörter (ähnlich jenen, die auch auf dem Gebiet des „*Information Retrieval*" zum Einsatz kommen) repräsentiert, wodurch Fehler der Spracherkennung kompensiert werden können. Das Resultat ist eine Zuordnung von Passagen und der darin enthaltenen Wörter des automatischen Transkripts mit jenen des manuellen Transkripts. Da erfindungsgemäß Passagen (Textfenster) als Einheit des Übereinstimmungsprozesses und auf den in diesen Passagen enthaltenen Wörtern definierte Eigenschaften verwendet werden, ist eine exakte Übereinstimmung von Wörtern nicht mehr erforderlich, und das Verfahren wird somit gegen Fehler der automatischen Texterstellung wesentlich robuster.

Durch die erfindungsgemäße Verwendung eines auf Textpassagen basierenden Ansatzes anstelle des bekannten, auf einzelnen Worten basierendes Ansatzes wird auch der Aufwand der Verarbeitung erheblich gemindert.

Im Gegensatz zu solchen Verfahren, welche auf dem Prinzip des „*forced-alignment*" basieren, ist es beim erfindungsgemäßen Ansatz nicht notwendig, den Text schon vor der eigentlichen Erkennung zur Verfügung zu haben. Weiters erfolgt die Spracherkennung ohne Zuhilfenahme des *forced-alignment* (und der damit verbundenen Probleme). Das vorliegende Verfahren beschränkt sich ausschließlich auf die Verwendung des durch den Spracherkenner erzeugten Textes und des manuell erzeugten Gegenstückes.

Das vorliegende Verfahren erlaubt somit, eine beispielsweise durch einen automatischen Spracherkenner generierte Transkription (AT) mit einer manuellen Transkription (MT) derselben Audio- oder Videodatei automatisch und dynamisch in Einklang zu bringen (d.h. zwischen ihnen ein *Alignment*, eine Zuordnung herzustellen).

Im vorliegenden Verfahren produziert der automatische Spracherkenner ein automatisches Transkript, welches den in den eingegebenen Audiodaten vorkommenden Wörtern entspricht. Zusammen mit jedem Wort der Transkription wird auch ein Zeitstempel (*Time-tag*) des Wortes generiert. Dieser Zeitstempel gibt an, wann genau dieses Wort im Audiostrom erkannt wurde (relativ zum Begin der Datei). Die Audiodaten selbst können dabei von einer Vielfalt an Quellen kommen, so z.B. aus Videoaufnahmen oder Audio-Clips. Das manuelle Transkript wird typischerweise von einem Transkriptionisten erstellt, der eine Aufnahme oder ein Stenogramm als Referenz verwendet. Die Qualität der Transkription, und wie exakt diese das eigentliche Audio wiedergibt, variiert dabei sehr stark. Da bei der manuellen Transkription die Verständlichkeit im Vordergrund steht, und nicht eine möglichst exakte Transkription der Audiodaten geliefert werden soll, werden dabei außersprachliche Phänomene wie Räuspern, Husten, Atemgeräusche, Schmatzen der Lippen u.Ä., oder sprachliche Phänomene wie Stottern, Versprecher, Behebung von Fehlern und Mehrfachstarts einer Phrase (z.B. „ich ich ich möchte Sie ah Ihnen folgendes Angebot machen und ...") nicht berücksichtigt. Diese werden jedoch vom automatischen Spracherkenner erkannt und transkribiert (möglicherweise auch „falsch" erkannt und transkribiert). Sie stellen folglich ein Problem bei der Zuordnung der beiden Transkripte dar; ihre Berücksichtigung allerdings erlaubt eine genauere Zuordnung von Wörtern und deren Zeitstempel.

Die Erfindung wird im Folgenden unter Bezugnahme auf die Zeichnungen näher erläutert, in denen Fig. 1 ein allgemeines Schema der Zuordnung von Text aus einem automatisch aus einer Audio-Aufzeichnung erzeugten Transkript zu Text aus einem manuell aus der Audio-Aufzeichnung generierten Transkript darstellt, Fig. 2 schematisch einen Überblick über das erfindungsgemäße Verfahren zeigt, Fig. 3 eine im erfindungsgemäßen Verfahren verwendete

Auswertungs-Matrix zeigt, Fig. 4 einen bei der Durchführung des Verfahrens erstellten Worthäufigkeits-Vektor darstellt, Fig. 5 darstellt, wie der manuell erstellte Text Wort für Wort mit Zeitstempeln der automatisch erkannten Wörter versehen wird, und Fig. 6 die ersten Schritte des Ergebnisses des dynamischen Vergleichs in einem Ausführungsbeispiel des erfindungsgemäßen Verfahrens zeigt.

Das vorliegende Verfahren basiert auf der Unterteilung der beiden Texte in Passagen (Fenster), deren Länge durch einen Parameter, welcher angepasst werden kann, bestimmt ist. Jede Passage wird um einen anzugebenden Wert (an Worten) im Text nach hinten verschoben. Dies passiert in beiden Dateien gleich und überlappend, wobei jede Passage von AT mit jeder Passage von MT verglichen wird (siehe Fig. 2). Die Länge der Passagen muss dabei nicht gleich groß sein. Durch Variieren der Parameter und mehrfaches Erstellen einer Zuordnung kann diejenige Zuordnung, die die beste Gesamtbewertung erhielt, ausgewählt werden. Dies ist ein spezieller Fall des Verfahrens der dynamischen Programmierung auf Basis von Text-Passagen anstatt von Einzelworten. Die dynamische Programmierung stellt für sich ein allgemeines Programmierwerkzeug dar, das häufig zur Anwendung kommt, wenn der Suchraum eines Problems sich als Abfolge von Zuständen darstellen lässt. Die Zustände müssen dabei folgende Bedingungen erfüllen:

- der Initialzustand enthält triviale Lösungen von Sub-Problemen
- jede Teillösung eines späteren Zustandes kann aus einer eingeschränkten Anzahl an bereits errechneten Teillösungen eines früheren Zustandes ermittelt werden.
- der letzte Zustand enthält die Lösung des Gesamtproblems

Diese Voraussetzungen sind in unserem Fall erfüllt: die beiden Sequenzen von Textpassagen stellen die Achsen einer Matrix dar (Fig. 3). Die Spalten der Matrix stellen dabei die Zustände dar, ein Matrix-Eintrag stellt eine Teillösung dar, welche nur aus den Teillösungen der vorhergehenden und derselben Spalte ermittelt wird (z.B. stellt der Eintrag am_{ij} die bestmögliche Anordnung von Sequenzen bis inklusive einer Übereinstimmung von Sequenz a_i und Sequenz m_j dar). Das letzte Elemente am_{nm} enthält die Gesamtlösung des Problems. Durch Rückverfolgung des Pfades durch die Matrix ergibt sich dann die bestmögliche Zuordnung. In unserem Fall handelt es sich um einen Vergleich zweier Text-Passagen, der eine Bewertung über die Ähnlichkeit dieser Passagen liefert (welche wiederum Eingang in die dynamische Zuordnung findet). Dieser Vergleich wird unter Verwendung von Eigenschaften, die auf diesen Passagen definiert sind (wie sie z.B. im Bereich des „Information Retrieval" angewendet werden) durchgeführt. Die Passagen werden als Vektoren dieser Eigenschaften (definiert über den in der Passage enthaltenen Wörtern)

dargestellt. In der bevorzugten Ausführungsform wird dabei jeder Komponente des Vektors ein Wort und dessen Häufigkeit zugewiesen (Fig. 4).

Die Verwendung dieser Darstellung erlaubt eine Reihe von Möglichkeiten der Repräsentation. So kann etwa auch die auf dem Gebiet des „Information Retrieval" bekannte TF-IDF (term frequency / inverse document frequency) verwendet werden, deren wesentlichste Begriffe im Folgenden zusammengefasst sind:

term frequency / inverse document frequency (tf/idf):

term frequency:	$tf_{i,j}$	wie oft Wort w_i in Dokument d_j vorkommt
document frequency:	df_i	Anzahl der Dokumente im Korpus in denen w_i vorkommt
collection frequency:	cf_i	Gesamtanzahl der Vorkommen von w_i im gesamten Korpus

Üblicherweise werden diese Größen über Wörter und Dokumente in einem Korpus definiert. In unserem Fall kann man die Passagen (Fenster) als Dokumente und das Gesamtdokument als Korpus ansehen. Man kann aber auch das Gesamtdokument als Dokument ansehen und das Korpusmodell aus einem größeren Textkorpus erstellen. Man könnte diese beiden Ansätze auch miteinander kombinieren.

tf/idf sagt nur aus, dass die eigentliche Wortfrequenz (term frequency) und die Dokumentfrequenz (document frequency) miteinander kombiniert werden, um den Wert für ein bestimmtes Wort zu ermitteln. Es gibt zahlreiche Varianten, wie diese Werte miteinander kombiniert werden, z.B.

$weight(i,j) = (1+\log(tf_{i,j}))(\log (N/df_i))$ wobei N die Gesamtanzahl an Dokumenten ist.

Weitere Möglichkeiten der Repräsentation sind Wortstämme (Lemmas) anstatt der Vollformen, es können auch lautliche Ähnlichkeit von Wörtern oder Stopwortlisten zum Ausnehmen bestimmter Wörter eingesetzt werden. Durch Anwendung normierter Vektoren (d.h. deren Länge 1 ist) kann der Vergleich zweier Vektoren als Bestimmung des Cosinus des Winkels zwischen ihnen betrachtet werden. Dies dient als Maß der Ähnlichkeit der Vektoren und somit der durch sie repräsentierten Textpassagen. Alternative Maße sind etwa der Abstand der Endpunkte der Vektoren oder die Anzahl der unterschiedlichen Dimensionen. Das Verfahren der dynamischen Programmierung liefert die bestmögliche Kette von Zuordnungen zwischen Passagen aus AT und Passagen aus MT (*bestmöglich* im Sinne einer Minimierung der Kosten des Zuordnungsprozesses, wobei identische Passagen

einen Wert von 0 bekommen und unterschiedliche Passagen Werte zwischen 0 und 1, entsprechend ihrer Distanz, d.h. dem Winkel zwischen ihnen).

Die möglichen Zuordnungen zweier Passagen sind dabei:

1) Passage aus AT wird Passage aus MT zugeordnet
2) Passage aus AT kann keiner Passage aus MT zugeordnet werden
3) Passage aus MT kann keiner Passage aus AT zugeordnet werden

Fall 2) wird in der Literator als „*insertion error*" bezeichnet und entspricht einer Textpassage, die vom automatischen Spracherkenner transkribiert wurde. Es ist an dieser Stelle in der Audiodatei also Sprache vorhanden, die allerdings nicht manuell transkribiert wurde (möglicherweise Stimmen und Audio während einer Verhandlungspause oder Zwischenrufe etc.), weil sie vom Transkriptionisten überhört wurden oder als unwichtig betrachtet wurden (was für einen menschlichen Leser/Hörer auch stimmen mag, jedoch für eine automatische Verarbeitung ein Problem darstellt.)

Fall 3) wird als „*deletion error*" bezeichnet und entspricht einer manuell transkribierten Passage, die allerdings nicht vom automatischen Spracherkenner transkribiert wurde (z.B. Ergänzungen oder ungenaue manuelle Transkription.

Zuordnungen beiden Typs können vom gegenständlichen Verfahren berücksichtigt werden, „*insertion errors*" indem der „zusätzliche Text" anders dargestellt wird und „*deletion errors*", indem der nicht direkt zugeordnete Text an entsprechender Stelle eingefügt wird (siehe dazu auch die nachfolgende Beschreibung einer bevorzugten Ausführungsform).

Der dynamische Zuordnungsprozess liefert zudem einen Gesamtwert, der die Gesamtqualität der Zuordnung beschreibt. Unterschreitet dieser Werte eine Schranke, so kann man die Zuordnung als nicht sinnvoll verwerfen. Stehen mehrere Zuordnungen zur Verfügung, so kann diejenige mit der besten Bewertung gewählt werden.

Wird eine erfolgreiche (sinnvolle) Zuordnung von Passagen erstellt, dann kann dadurch eine direkte Beziehung der in diesen Passagen enthaltenen Wörter hergestellt werden. Dies erlaubt, den manuell erstellten Text Wort für Wort mit Zeitstempeln der automatisch erkannten Wörter zu versehen (Fig. 5). Dadurch ist es möglich, für jedes Wort die entsprechende Stelle in der zu Grunde liegenden Audio- bzw. Videodatei zu finden wodurch ein effizientes Auffinden von Audio ermöglicht wird. Die Genauigkeit dieses Vorgangs wird

dabei durch die Länge des Fensters (d.h. der Passage), der Zuweisung im Fenster aufgetretener Wörter und Interpolation zwischen den Fenstern bestimmt.

Im folgenden Beispiel einer derzeit bevorzugten Ausführungsform des erfindungsgemäßen Verfahrens wird ein automatisch erzeugtes Transkript (AT), welches durch einen automatischen Spracherkenner erzeugt wird, mit einem manuell erstellten Transkript (MT) derselben Audiodatei verglichen und die darin enthaltenen Textpassagen und Wörter miteinander in Einklang gebracht.

Die beiden Textdateien werden dazu in Textpassagen gleicher Länge unterteilt. Diese werden pro Vergleichsschritt um eine vorgegebene Zahl an Worten nach hinten verschoben, und zwar so, dass einander angrenzende Passagen sich um eine vorgegebene Zahl an Worten überlappen. Diese beiden Werte sind frei wählbar und können dem konkreten Text angepasst werden (z.B. kann Wissen um die Beschaffenheit der automatischen und/oder manuellen Transkription oder deren Qualität einfließen).

Mittels des Verfahrens der dynamischen Programmierung werden alle Passagen miteinander verglichen. Als Metrik in diesem Vergleich dient der Cosinus des Winkels (der *Abstand*) zwischen den die jeweiligen Textpassagen repräsentierenden Vektoren. Diese Vektoren werden aus den in der Passage enthaltenen Wörtern erzeugt. In der bevorzugten Ausführungsform werden hierfür die Wörter selbst verwendet, wobei jedes Wort und seine Häufigkeit eine Komponente des Vektors darstellt. Das gegenständliche Verfahren ist jedoch keineswegs auf diese Darstellung der Vektoren beschränkt, sondern eignet sich gleichwertig für andere Darstellungen. So können etwa auch TF/IDF oder auf lautlicher Ähnlichkeit oder auf anderen Eigenschaften beruhende Verfahren, wie etwa die Grundform eines Wortes oder seine phonetische Repräsentation verwendet werden. Gleichfalls können Wörter zu Komposita zusammengesetzt oder Komposita in ihre Bestandteile zerlegt werden. Das Resultat dieses Prozesses ist eine Zuordnung von Passagen aus den beiden Eingangstexten. Diese Zuordnung wird in einem nächsten Schritt nun zur Zuordnung auf Wortbasis verwendet. Vom letzten zugeordneten Paar von Textpassagen (die dem Ende der beiden Eingabedateien entsprechen) ausgehend wird folgendes Verfahren angewendet:

- entspricht die Passage aus MT jener aus AT, dann wird jedem Wort aus MT der Zeitstempel des entsprechenden Wortes aus AT zugewiesen;
- ist die Passage aus AT eine „insertion" Passage, d.h. entspricht sie vom automatischen Spracherkenner erkannten Text, welchem keine Passage in MT entspricht, so wird dieser Text verworfen;

- ist die Passage aus MT eine „deletion" Passage, d.h. entspricht ihr keine Passage in AT, dann werden die Wörter aus MT vor dem zuletzt belegten Zeitstempel entsprechend der Beschaffenheit ihrer Wörter (z.B. der Länge in Phonemen), eingefügt. Dazu werden Zeitstempel künstlich generiert, die in dem entsprechenden Zeitintervall liegen.
- Wörter, die im Bereich der Überlappung zweier benachbarter Passagen liegen, werden gesondert behandelt. Es werden dabei die Zeitstempel derjenigen Passage verwendet, die den besseren Vergleichswert besitzt (d.h. den kleineren Wert der Distanz). Korrekte miteinander in Einklang gebrachte Passagen (mit Abstand 0) werden so immer bevorzugt behandelt.

Die obigen Schritte werden so lange angewendet, bis die erste Passage in jedem der beiden Eingabedateien erreicht ist. Zu diesem Zeitpunkt wurde allen Wörtern in allen Passagen der MT jeweils ein Zeitstempel zugeordnet. Dieser wird nun gemeinsam mit dem Wort ausgegeben und kann in der Suche nach dem Wort, bzw. der Lokalisierung des Wortes in der Mediadatei direkt und effizient verwendet werden.

Wir wollen diese Schritte nun anhand eines konkreten Beispiels betrachten.

Als manuell transkribierter Text soll folgendes Beispiel dienen:

„der automatische Spracherkenner produziert dabei ein automatisches Transkript welches den in den Eingabedaten vorkommenden Wörtern entspricht"

Dieser Text entspricht der von einem menschlichen Transkriptionisten erstellten Version (MT).

Derselbe Text in der vom Spracherkenner ausgegebenen Fassung (AT) könnte beispielsweise folgendermaßen lauten:

<time start="00000011" end="00000045">lauter</time>
<time start="00000046" end="00000084">Bitte</time>
<time start="00000085" end="00000101">eine</time>
<time start="00000102" end="00000212">automatische</time>
<time start="00000213" end="00000253">Sprache</time>
<time start="00000254" end="00000281">erkennen</time>
<time start="00000282" end="00000325">produziert</time>
<time start="00000326" end="00000370">dabei</time>

<time start="00000371" end="00000387">eine</time>
<time start="00000388" end="00000410">[AHM]</time>
<time start="00000411" end="00000458">Transkript</time>
<time start="00000459" end="00000607">welches</time>
<time start="00000608" end="00000747">den</time>
<time start="00000748" end="00000772">in</time>
<time start="00000773" end="00000797">den</time>
<time start="00000798" end="00000825">Eingabe</time>
<time start="00000826" end="00000925">Dateien</time>
<time start="00000926" end="00000962">vor</time>
<time start="00000963" end="00001053">kommen</time>
<time start="00001054" end="00001074">den</time>
<time start="00001075" end="00001096">Wörtern</time>
<time start="00001097" end="00001160">Ende</time>
<time start="00001831" end="00001995">spricht</time>

Zusammen mit jedem Wort ist der jeweilige Zeitstempel angegeben, d.h. der Zeitpunkt relativ zum Begin der Audioeingabedaten (in 1/100 s), an dem der Spracherkenner das jeweilige Wort erkannte. Im obigen Beispiel wurde einmal ein Zögern im Audiostrom erkannt ([AHM]) Möglicherweise hat der Sprecher an dieser Position mit dem Fortsetzen des Satzes tatsächlich gezögert. Die beiden ersten Wörter der Passage könnten zum Beispiel von einem Zwischenruf stammen, den der Spracherkenner transkribierte, den jedoch der menschliche Transkriptionist nicht berücksichtigte.

Der Vergleich dieser beiden Texte (MT und AT) wird nun mittels einander überlappender Fenster durchgeführt. Im gegenständlichen Beispiel handelt es sich um Fenster der Länge 4 (Wörter), die jeweils um 2 Wörter nach hinten verschoben werden. Alle Fenster aus AT werden mit allen Fenstern aus MT verglichen. Dabei wird die durch die Text-Passagen definierte Matrix Schritt für Schritt, von links nach rechts (d.h. in der Zeit fortschreitend) und oben nach unten ausgefüllt. Jedes Matrix-Element entspricht dem bis zu ihm besten (kostengünstigsten) Pfad. Diese Schritte werden weiter durchgeführt, bis alle Passagen miteinander verglichen wurden und damit alle Elemente der Matrix einen Wert zugewiesen bekommen haben. Anschliessend wird, ausgehend vom kostengünstigsten Element der letzten Spalte der Weg, welcher zu diesem Element führte verfolgt (*back-tracking*), wodurch sich die eindeutige Sequenz von Aktionen und Zuweisungen zwischen Passagen ergibt.

Fig. 6 zeigt die ersten Schritte des Ergebnis des dynamischen Vergleichs. Das erste Fenster in AT wurde als „*insertion*“ ausgewiesen, d.h. als zwar erkannter, jedoch nicht dem manuell transkribierten Text entsprechender Text. Das zweite Fenster in AT wurde dem ersten Fenster aus MT zugewiesen, wodurch sich die Übertragung der dazugehörigen Zeitstempel im Resultat ergibt. Die weiteren Fenster wurden einander gemäß den in diesem Verfahren dargestellten Regeln zugewiesen.

Das Resultat des Verfahrens ist eine Zuweisung der Wörter aus MT zu den Zeitstempeln der Wörter aus AT entsprechend der einander zugewiesenen Fenster (siehe obige Abbildung)

<time start="00000085" end="00000101">der</time>
<time start="00000102" end="00000212">automatische</time>
<time start="00000213" end="00000253">Spracherkenner</time>
<time start="00000254" end="00000281">produziert</time>
<time start="00000282" end="00000325">dabei</time>
<time start="00000326" end="00000370">ein</time>
<time start="00000371" end="00000387">automatisches</time>
<time start="00000388" end="00000410">Transkript</time>
<time start="00000411" end="00000458">welches</time>
<time start="00000459" end="00000607">den</time>
<time start="00000608" end="00000747">in</time>
<time start="00000748" end="00000772">den</time>
<time start="00000963" end="00001053">Eingabedaten</time>
<time start="00001054" end="00001074">vorkommenden</time>
<time start="00001097" end="00001160">Wörtern</time>
<time start="00001831" end="00001995">entspricht</time>

Ansprüche:

1. Verfahren zur automatischen Übereinstimmung von in einer Audioaufnahme enthaltenen Audio-Segmenten mit Textelementen in einem manuell aus der Audioaufnahme erzeugten Transkript (MT), wobei zunächst aus der Audio-Aufnahme, vorzugsweise durch einen automatischen Spracherkenner, ein automatisches Transkript (AT) erstellt wird, das die zu Textelementen geformten Audiosegmente zusammen mit einem Zeitbezug, an welcher Stelle in der Audio-Aufnahme sich das jeweilige automatisch erstellte Textelement befindet, enthält, dadurch gekennzeichnet, dass das Verfahren folgende weitere Schritte umfasst:

das Unterteilen des automatischen Transkripts (AT) und des manuellen Transkripts (MT) in Passagen (at_i, mt_j) definierter, aber nicht notwendigerweise gleicher Länge, die jeweils mehrere Textelemente umfassen,

das Verschieben jeder Passage um einen anzugebenden Wert an Textelementen im automatischen Transkript (AT) und im manuellen Transkript (MT) über das gesamte automatische und manuelle Transkript, wobei jede Passage sich mit der vorhergehenden Passage überlappt, und Ermitteln einer bestimmten Passagen-Eigenschaft für jede Passage,

das Vergleichen der Passagen-Eigenschaft jeder Passage (at_i) im automatischen Transkript (AT) mit jeder Passage (mt_j) im manuellen Transkript (MT),

das Zuordnen einer jeweiligen Passage im automatischen Transkript (AT) zu jener Passage im manuellen Transkript (MT), so dass sich über die Summe der Passagenvergleiche gesehen der optimale Pfad an Zuordnungen ergibt.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, dass ein Textelement ein oder mehrere Wörter oder Bestandteile von Wörtern oder Wortstämmen umfasst.

3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, dass die Passagen-Eigenschaft die Häufigkeit des Auftretens der in der Passage enthaltenen Textelemente, oder lautlich ähnlicher Einheiten ist.

4. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, dass die Passagen-Eigenschaft die term frequency / inverse document frequency (TF-IDF) ist.

5. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass zur Ermittlung der Passagen-Eigenschaft Stopwortlisten zum Ausnehmen bestimmter Wörter verwendet werden.

6. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass die Passagen-Eigenschaft durch einen Vektor, vorzugsweise einen normierten Vektor mit einer Einheitslänge, dargestellt wird, und vorzugsweise der Vergleich von Passagen-Eigenschaften zweier Passagen anhand des von den Vektoren gebildeten Winkels oder des Abstandes der Spitzen der Vektoren voneinander oder der Anzahl der unterschiedlichen Dimensionen der Vektoren oder einer Funktion der obigen Maßzahlen erfolgt.

7. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass die Länge der Passagen und/oder die Weite ihrer Verschiebung in mehreren Durchläufen des Verfahrens variiert werden und in jedem Durchlauf die Passagen-Eigenschaft ermittelt, verglichen und einer jeweiligen Passage im automatischen Transkript (AT) jene Passage im manuellen Transkript (MT) zugeordnet wird, so dass sich über die Summe der Passagenvergleiche gesehen der optimale Pfad an Zuordnungen ergibt, wobei als endgültige Zuordnung jene ausgewählt wird, die für alle Passagen die beste Gesamtbewertung erzielt.

8. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass die Zuordnung durch die Mittel der dynamischen Programmierung getroffen wird.

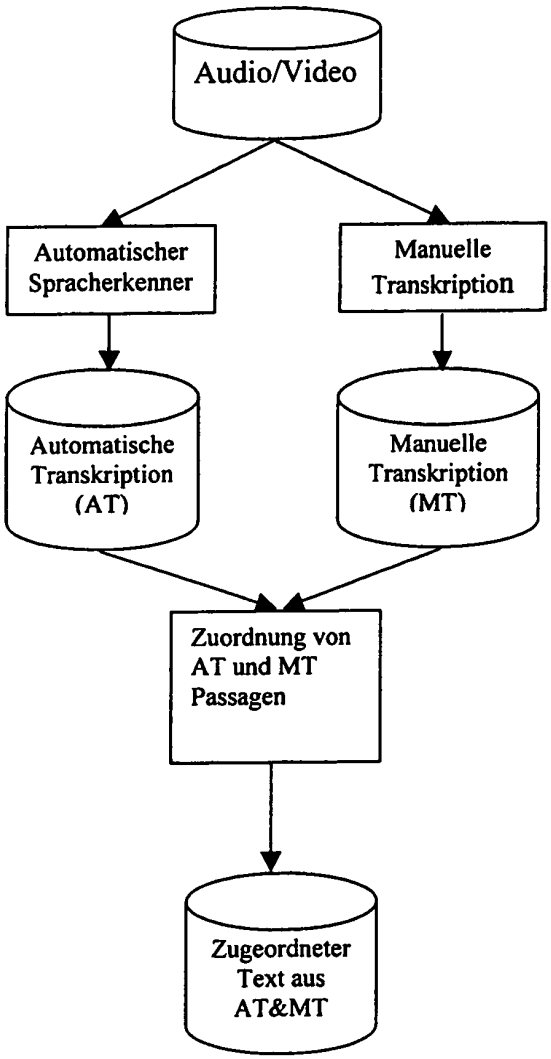

Audio/Video
Automatischer Spracherkenner
Manuelle Transkription
Automatische Transkription (AT)
Manuelle Transkription (MT)
Zuordnung von AT und MT Passagen
Zugeordneter Text aus AT&MT

Fig. 1

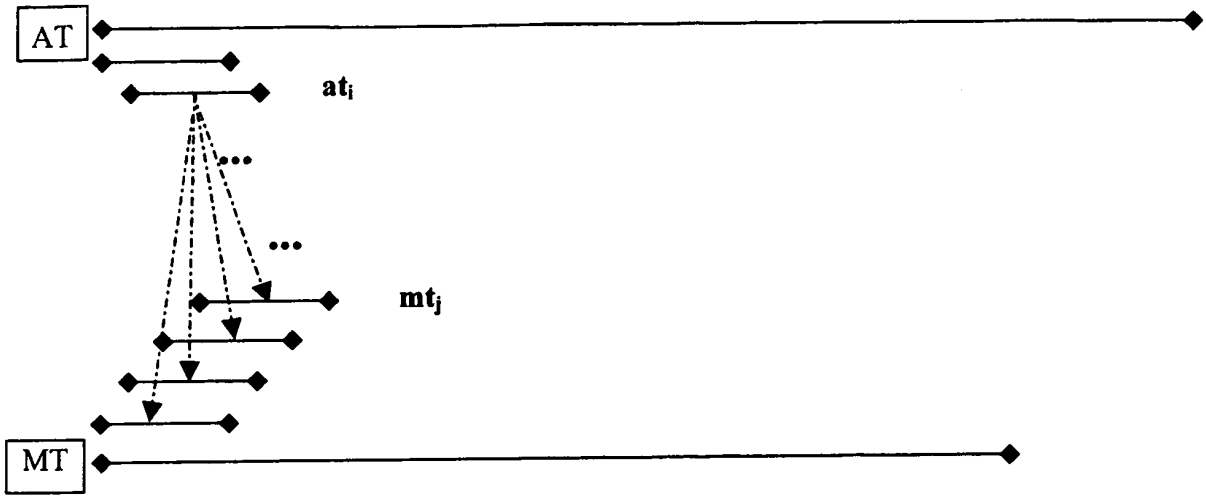


Fig. 2

	m_1	m_2	m_3	m_4			m_m
a_1							
a_2							
a_3		am_{32}					
a_4							
a_n							am_{nm}

Fig. 3

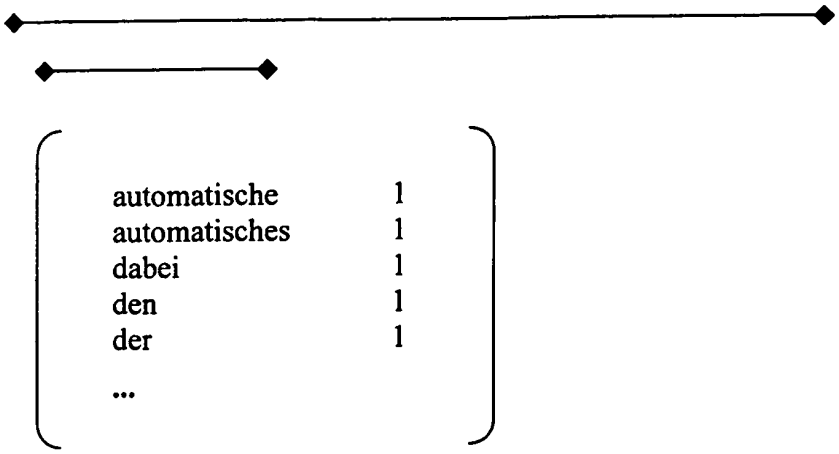


Fig. 4

AT: eine*(t1) automatische(t2) Sprache(t3) Erkennen(t4)* ...

MT: *der automatische Spracherkenner produziert* ...

Ausgabe: *der(t1) automatische(t2) Spracherkenner(t3) produziert(t4)* ...

Fig. 5

*MT: *(null)* *############################ (Distanz: 1.00)* *AT: lauter Bitte eine automatische* **********
MT: der automatische Spracherkenner produziert *############################ (Distanz: 0.38)* *AT: eine automatische Sprache erkennen* **********
MT: Spracherkenner produziert dabei ein *############################ (Distanz: 0.25)* *AT: Sprache erkennen produziert dabei*
MT: dabei ein automatisches Transkript *############################ (Distanz: 0.38)* *AT: produziert dabei eine [AHM]*
...

Fig. 6

ÖSTERREICHISCHES PATENTAMT

Recherchenbericht zu GM 805/2002

Klassifikation des Anmeldungsgegenstands gemäß IPC[7]: G 10 L 15/22; G 06 F 3/16
Recherchierter Prüfstoff (Klassifikation): G 10 L; G 06 F; G 09 B
Konsultierte Online-Datenbank: Wpi; Epodoc; Paj
Dieser Recherchenbericht wurde zu den **am 28.11.2002 eingereichten** Ansprüchen erstellt. Die in der Gebrauchsmusterschrift veröffentlichten Ansprüche könnten im Verfahren geändert worden sein (§ 19 Abs. 4 GMG), sodass die Angaben im Recherchenbericht, wie Bezugnahme auf bestimmte Ansprüche, Angabe von Kategorien (X, Y, A), nicht mehr zutreffend sein müssen. In die dem Recherchenbericht zugrundeliegende Fassung der Ansprüche kann beim Österreichischen Patentamt während der Amtsstunden Einsicht genommen werden.

Kategorie*)	Bezeichnung der Veröffentlichung: Ländercode*), Veröffentlichungsnummer, Dokumentart (Anmelder), Veröffentlichungsdatum, Textstelle oder Figur soweit erforderlich	Betreffend Anspruch
A	WO 01/95292 A2 (Werth) 13. Dezember 2001 (13.12.2001) Ansprüche 1-5,21-22;	1
A	DE 44 09 205 A (Zangs) 21. September 1995 (21.09.95) Gesamtes Dokument	1
P	US 2003028375 A (Philips) 6. Feber 2003 (06.02.2003) Zusammenfassung	1

Datum der Beendigung der Recherche:	Prüfer(in):
12. Juni 2003	Dipl.-Ing. MIHATSEK

*) Bitte beachten Sie die Hinweise auf dem Erläuterungsblatt!

☐ Fortsetzung siehe Folgeblatt

ÖSTERREICHISCHES PATENTAMT

Erläuterungen zum Recherchenbericht

Die **Kategorien** der angeführten Dokumente dienen in Anlehnung an die Kategorien der Entgegenhaltungen bei EP- bzw. PCT-Recherchenberichten nur zur raschen Einordnung des ermittelten Stands der Technik. Sie stellen keine Beurteilung der Erfindungseigenschaft dar:

"A" Veröffentlichung, die den **allgemeinen Stand der Technik** definiert.

"Y" Veröffentlichung **von Bedeutung**: der Antragsgegenstand kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren weiteren Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese **Verbindung für einen Fachmann naheliegend** ist.

"X" Veröffentlichung **von besonderer Bedeutung**: der Antragsgegenstand kann allein aufgrund dieser Druckschrift nicht als neu bzw. auf erfinderischer Tätigkeit beruhend betrachtet werden.

"P" Dokument, das **von besonderer Bedeutung** ist (Kategorie „X"), jedoch **nach dem Stichtag**, auf den das Gutachten abzustellen war, **veröffentlicht** wurde.

"&" Veröffentlichung, die Mitglied derselben **Patentfamilie** ist.

Ländercodes:

AT = Österreich; **AU** = Australien; **CA** = Kanada; **CH** = Schweiz; **DD** = ehem. DDR; **DE** = Deutschland; **EP** = Europäisches Patentamt; **FR** = Frankreich; **GB** = Vereinigtes Königreich (UK); **JP** = Japan; **RU** = Russische Föderation; **SU** = Ehem. Sowjetunion; **US** = Vereinigte Staaten von Amerika (USA); **WO** = Veröffentlichung gem. PCT (WIPO/OMPI); weitere Codes siehe **WIPO ST. 3**.

Die genannten Druckschriften können in der Bibliothek des Österreichischen Patentamtes während der Öffnungszeiten (Montag bis Freitag von 8 bis 12 Uhr 30, Dienstag von 8 bis 15 Uhr) unentgeltlich eingesehen werden. Bei der von der Teilrechtsfähigkeit des Österreichischen Patentamts betriebenen Kopierstelle können **Kopien** der ermittelten Veröffentlichungen bestellt werden.

Auf Bestellung gibt die von der Teilrechtsfähigkeit des Österreichischen Patentamts betriebene Serviceabteilung gegen Entgelt zu den im Recherchenbericht genannten Patentdokumenten allfällige veröffentlichte **"Patentfamilien"** (den selben Gegenstand betreffende Patentveröffentlichungen in anderen Ländern, die über eine gemeinsame Prioritätsanmeldung zusammenhängen) bekannt.

Auskünfte und Bestellmöglichkeit zu diesen Serviceleistungen erhalten Sie unter der Telefonnummer

01 / 534 24 - 738 bzw. 739;

Schriftliche Bestellungen:

per FAX Nr. 01 / 534 24 – 737 oder per E-Mail an Kopierstelle@patent.bmvit.gv.at