(19) österreichisches patentamt

(10) **AT 414 060 B 2006-08-15**

(12) **Patentschrift**

(21) Anmeldenummer: A 260/2004
(22) Anmeldetag: 2004-02-19
(42) Beginn der Patentdauer: 2005-11-15
(45) Ausgabetag: 2006-08-15

(51) Int. Cl.[7]: **G10L 15/10**

(56) Entgegenhaltungen:

(L1) JYH-MING KUO; PRINCIPE, J.C.: SPEECH CLASSIFICATION USING A MODIFIED FOCUSED GAMMA NETWORK. NEURAL NETWORKS, 1996. IEEE INTERNATIONAL CONFERENCE ON, VOLUME: 4, 3-6 JUNE 1996, P.: P 1877-1882 VOL. 4.
(L2) SU-LIN WU; KINGSBURY, E.D.; MORGAN, N.; GREENBERG, S.: INCORPORATING INFORMATION FROM SYLLABLE-LENGTH TIME SCALES INTO AUTOMATIC SPEECH RECOGNITION.

(73) Patentinhaber:

HICKERSBERGER HELMUT DIPL.ING.
A-1200 WIEN (AT).

(72) Erfinder:

HICKERSBERGER HELMUT DIPL.ING.
WIEN (AT).

(54) KLANGFOLGEN-ERKENNER

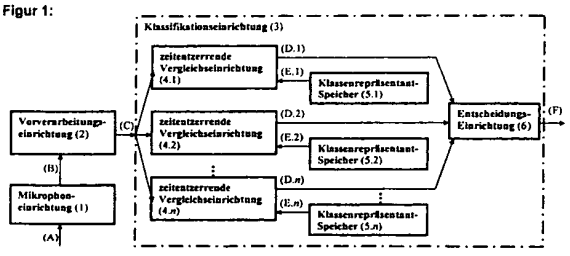
(57) Es handelt sich um ein Verfahren zur robusten Klassifikation von Schallschwingungen (A). Das Klassifikationsergebnis (F) dient zur Steuerung von technischen Geräten.

Anwendungen: Sprachsteuerungen für Mensch-Maschine-Schnittstellen sowie Klassifikation von nichtsprachlichen Schallschwingungen (beispielsweise Fahrzeuggeräusche in der Militärtechnik).

Vorteile: Geringer Speicherbedarf der Klassenrepräsentanten (E.1)-(E.*n*). Bei Erhöhung der Schallschwingungsklassenanzahl, können neue Klassenrepräsentant-Speicher (5.1)-(5.*n*) samt Vergleichseinrichtungen (4.1)-(4.*n*), ohne Modifikation der bereits gespeicherten Klassenrepräsentanten (E.1)-(E.*n*), hinzugefügt werden. Die Klassifikationszeit bleibt dabei nahezu gleich.

Deltamerkmale bestehen einerseits in der Art der für die Klassenrepräsentanten (E.1)-(E.*n*) abgespeicherten Information (für jeden Klassenrepräsentanten wird eine Folge von „Klängen" und für jeden Klang die für Plausibilität zulässige Minimalzeitdauer und Maximalzeitdauer abgespeichert) und andererseits in der Anwendung des „Run-Length-Limited-Dynamic-Programming-Zeitentzerrers" (4.1)-(4.*n*) zum Vergleich je eines Klassenrepräsentanten (E.1)-(E.*n*) mit der zu klassifizierenden Merkmalsvektorfolge (C). Dieser Zeitentzerrer nutzt die abgespeicherte Intervall-Information optimal: Klassenrepräsentanten, für die der zeitliche Rhythmus der zu klassifizierenden Schallschwingung plausibel ist, liefern bessere Vergleichsergebnisse (D.1)-(D.*n*).

Figur 1:



AT 414 060 B 2006-08-15

1. Technisches Gebiet

Pattern Recognition. Automatic Speech Recognition. Es handelt sich um ein Verfahren zur robusten Klassifikation von Schallschwingungen. Das erfindungsgemäße Verfahren eignet sich zur Anwendung für Sprachsteuerungen für Mensch-Maschine-Schnittstellen sowie zur Klassifikation von nichtsprachlichen Schallschwingungen (beispielsweise Fahrzeuggeräusche in der Militärtechnik).

2. Nächstliegender Stand der Technik:

Hidden Markoff Model Speech Recognizer *[Rabiner1989]*, Predictive Neural Network Speech Recognizer *[Iso1990]*, Hidden Control Neural Network Speech Recognizer *[Levin1993]*. Diese üblichen Verfahren repräsentieren - wie auch das erfindungsgemäße - Schallschwingungsklassen mittels Zustandsfolgen, wobei jedem Zustand ein Subword-Unit-Modell zugeordnet wird. Die Subword-Unit-Modelle sind verschieden und im allgemeinen relativ kompliziert aufgebaut (Gaussian Mixtures *[Rabiner1989]*, Predictive Neural Networks *[Iso1990]*, Hidden Control Neural Network *[Levin1993]*). Das erfindungsgemäße Verfahren verwendet hingegen sehr einfache Subword-Unit-Modelle, wobei nur ein einziger Merkmalsvektor samt plausiblem Zeitdauerintervall pro Subword-Unit-Modell gespeichert wird.

Die Zeitentzerrung beim Mustervergleich wird bei den üblichen Verfahren mittels des relativ einfachen „Dynamic-Programming-Zeitentzerrers" *[Levin1993]* beziehungsweise des „Viterbi-Zeitentzerrers" *[Forney1973]* bewerkstelligt. Das erfindungsgemäße Verfahren verwendet hingegen Zeitentzerrer, die zusätzliche Eigenschaften aufweisen müssen, nämlich daß sie die plausiblen Zeitdauerintervalle berücksichtigen können. Vorzugsweise wird der sogenannte „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" verwendet.

[Levin1993] Levin, E.: Hidden control neural architecture modeling of nonlinear time varying systems and its applications. Transactions on Neural Networks vol.4 (1993), p.109 - 116.

[Iso1990] Iso, K.; Watanabe, T.: Speaker-independent word recognition using a neural prediction model. Proceedings of the ICASSP (1990), p.441 - 444.

[Rabiner1989] Rabiner, L. R.: A Tutorial on hidden markov models and selected applications in speech recognition. IEEE Proceedings vol.77, no.2 (1989), p.257 - 286.

[Forney1973] Forney, G. D.: The viterbi algorithm. Proceedings of the IEEE vol. 61 (1973), p.268 - 278.

[L1] Jyh-Ming Kuo; Principe, J.C.: Speech classification using a modified focused gamma network. Neural Networks, 1996. IEEE International Conference on, Volume: 4, 3-6 June 1996, Seiten: p 1877 - 1882 vol. 4.

[L2] Su-Lin Wu; Kingsbury, E.D.; Morgan, N.; Greenberg, S.: Incorporating information from syllable-length time scales into automatic speech recognition. Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on, Volume: 2, 12-15 May 1998, Seiten:721 - 724 vol. 2.

[L3] Yfantis, E.A.; Lazarakis, T.; Angelopoulos, A.; Elison, J.D.; Zhang, Y.: On time alignment and metric algorithms for speech recognition. Information Intelligence and Systems, 1999. Proceedings. 1999 International Conference on, 31 October - 3 November 1999, Seiten: 423 - 428.

[L4] Komori, T.; Katagiri, S.: Application of a generalized probablistic descent method to dynamic time warping-based speech recognition. Acoustics, Speech, and Signal Processing,

1992. ICASSP-92., 1992 IEEE International Conference on, Volume: 1, 23-26 March 1992, Seiten: 497 - 500 vol. 1.

[L5] Katagiri, S.; Lee, C.-H.; Juang, B.-H.: New discriminative training algorithms based on the generalized probablistic descent method. Neural Networks for Signal Processing [1991]., Proceedings of the 1991 IEEE Workshop, 30 September - 1 October 1991, Seiten: 299 - 308.

3. Kurzbeschreibung

§1: Es handelt sich um ein Verfahren zur automatischen Klassifikation von Schallschwingungen, insbesondere von Geräuschen und gesprochenen Worten. Eine zu klassifizierende Schallschwingung wird zunächst mittels einer Mikrophoneinrichtung in eine elektrische Größe und mittels einer üblichen Vorverarbeitungseinrichtung zu einem zeitlich-diskreten vektorwertigen Signal, der so genannten Merkmalsvektorfolge verarbeitet. Diese wird vorzugsweise in der Klassifikationseinrichtung abgespeichert und wird mit mehreren Klassenrepräsentanten verglichen, wobei zeitentzerrende Vergleichseinrichtungen verwendet werden. Jede Vergleichseinrichtung liefert eine skalare Größe als Vergleichsergebnis. Die Vergleichsergebnisse werden mittels einer Entscheidungseinheit zu einem Klassifikationsergebnis verarbeitet, das heißt, das die Entscheidungseinheit jene Klasse angibt, die das beste Vergleichsergebnis aufzuweisen hat. Das Klassifikationsergebnis dient zur Steuerung eines technischen Gerätes vorzugsweise im Rahmen einer Mensch-Maschine-Schnittstelle.

§2: Jeder Klassenrepräsentant ist in einem zugehörigen Klassenrepräsentant-Speicher, vorzugsweise datenkomprimiert abgespeichert. Folgende Information wird gespeichert: Die Anzahl der Klänge, vorzugsweise 3-15, die einzelnen Klänge in richtiger Reihenfolge repräsentiert durch je einen einzigen Merkmalsvektor und für jeden Klang die jeweils zulässige Minimalzeitdauer und die jeweils zulässige Maximalzeitdauer.

§3: Die zeitentzerrenden Vergleichseinrichtungen werden als „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" realisiert. Dieser Zeitentzerrer wählt beim Vergleich der Merkmalsvektorfolge mit einem Klassenrepräsentanten die Zeitdauern der „Klänge" des Klassenrepräsentanten so, dass sich ein Vergleichsergebnis ergibt, das so gut wie möglich ist. Das Vergleichsergebnis ergibt sich als Summe der Abstandsmaße - vorzugsweise der quadrierten euklidischen Distanz - jedes Merkmalsvektors der zu klassifizierenden Merkmalsvektorfolge zum zugehörigen Klang des Klassenrepräsentanten.

§4: Das Deltamerkmal zum üblichen Dynamic-Programming-Zeitentzerrer mit Left-to-Right-Zustandsmodell besteht darin, dass der „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" die Zeitdauern der „Klänge" zum Zwecke des Vergleichs so anpasst, dass sie die Bedingungen erfüllen, innerhalb der ihnen jeweils zugeordnet abgespeicherten Intervalle aus Minimalzeitdauer und Maximalzeitdauer zu liegen. Es gelten weiters folgende Randbedingungen: Der erste Merkmalsvektor gehört zum ersten Klang; der letzte Merkmalsvektor gehört zum letzten Klang; die Klänge folgen aufeinander; kein Klang darf ausgelassen werden; die Dauern der Klänge müssen innerhalb der plausiblen Zeitdauerintervalle liegen. Unter diesen Randbedingungen passt der „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" die Zeitdauern der „Klänge" für ein optimales Vergleichsergebnis an.

Vorteile des erfindungsgemäßen Verfahrens: Geringer Speicherbedarf der Klassenrepräsentanten sowie der skalierbare, modulare Aufbau: Bei Erhöhung der Schallschwingungsklassenanzahl, können neue Klassenrepräsentant-Speicher samt Vergleichseinrichtungen, ohne Modifikation der bereits gespeicherten Klassenrepräsentanten hinzugefügt werden. Die Klassifikationszeit bleibt dabei nahezu gleich.

4. Auflistung der Figuren

Figur 1 zeigt eine Übersicht über den Klassifikationsvorgang: Die Aufnahme der Schallschwingung (A) mittels der Mikrophoneinrichtung (1); die Vorverarbeitung der elektrischen Größe (B) mittels einer Vorverarbeitungseinrichtung (2); die Klassifikation der Merkmalsvektorfolge (C) mittels der Klassifikationseinrichtung (3); das Klassifikationsergebnis (F).

Figur 2 zeigt beispielsweise die Zuordnung der Klänge (Z1)-(Z5) eines Klassenrepräsentanten (E.*n*) zu den Vektoren (M1)-(M14) der zu klassifizierenden Merkmalsvektorfolge (C). Die Zuordnung wird von der zeitentzerrenden Vergleichseinrichtung (4.*n*) getroffen. Im dargestellten Beispiel handelt es sich um einen Klassenrepräsentanten mit 5 Klängen, der mit einer zu klassifizierenden Merkmalsvektorfolge von 14 Merkmalsvektoren verglichen wird.

Figur 3 zeigt das sogenannte Left-to-Right-Zustandsmodell, dessen Zustandsfolge (Klangfolge) folgende Bedingungen erfüllt: Der erste Klang ist dem ersten Merkmalsvektor zugeordnet; der letzte Klang ist dem letzten Merkmalsvektor zugeordnet; kein Klang darf ausgelassen werden; die Reihenfolge der Klänge darf nicht vertauscht werden. Zusätzlich sind beispielhafte Lauflängenbedingungen für die einzelnen Zustände (Klänge) notiert.

Figur 4 zeigt das transformierte Zustandsdiagramm, welches der „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" aus den abgespeicherten, zulässigen Zeitdauerintervallen der Klänge berechnet und zur Zeitanpassung verwendet. In der Figur ist ein Zustandsdiagramm für folgende Lauflängenbedingungen dargestellt: Für den Klang Z1 gilt 1 bis unendlich. Für den Klang Z2 gilt 1 bis 2. Für den Klang Z3 gilt 2 bis 3. Für den Klang Z4 gilt 2 bis 4. Für den Klang Z5 gilt 1 bis unendlich. Einen der möglichen Pfade zeigt Figur 2.

5. Detailbeschreibung

Der Erkennungsvorgang im Überblick

Siehe Kurzbeschreibung §1. Siehe Figur 1 samt Beschreibung. Mittels einer üblichen Vorverarbeitungseinrichtung (2), beispielsweise einer Filterbank, wird die elektrische Größe (B) der zu klassifizierenden Schallschwingung (A) in eine Folge von Merkmalsvektoren (C) verarbeitet, die den zeitlichen Verlauf üblicher Signaleigenschaften angibt. Der Vorrat an Klassen wird durch mehrere Klassenrepräsentant-Speicher (5.1)-(5.*n*) zusammen mit den zugehörigen zeitentzerrenden Vergleichseinrichtungen (4.1)-(4.*n*) gebildet, welche voneinander vollständig unabhängig arbeiten. Der Vorrat an Klassen ist daher modular erweiterbar. Die Vergleichsergebnisse (D.1)-(D.*n*) repräsentieren Maße für die Übereinstimmung der Klassenrepräsentanten mit der beobachteten Merkmalsvektorfolge (C). In der Entscheidungseinrichtung (6) wird die Klasse mit dem besten Vergleichsergebnis als Klassifikationsergebnis (F) gewählt. Jede Klasse wird vorzugsweise durch genau einen Klassenrepräsentanten repräsentiert.

Der Klassenrepräsentant-Speicher

Siehe Kurzbeschreibung §2. Je nach Implementierung kann der zur Speicherung des Klassenrepräsentanten notwendige Speicherplatz bemerkenswert niedrig sein.

Die zeitentzerrenden Vergleichseinrichtungen

Siehe Kurzbeschreibung §3. Siehe Figur 2 samt Beschreibung. Jedes Quadrat in Figur 2 bedeutet die Berechnung eines Abstandsmaßes, vorzugsweise der quadrierten euklidischen Distanz, je eines Klanges und eines Merkmalvektors. Die schwarz gefärbten Quadrate repräsentieren die beispielhafte optimale Zuordnung unter dem Kriterium, dass die Gesamtsumme der Abstandsmaße minimal ist.

Siehe Figur 3 samt Beschreibung. Das übliche Left-to-Right-Zustandsmodell nach Figur 3 ohne Lauflängenbedingungen kann jedoch auch Zustandsfolgen produzieren, die nach dem erfin-

dungsgemäßen Verfahren nicht zulässig sind. Zulässig sind beim erfindungsgemäßen Verfahren nur Zustandsfolgen, welche die in Figur 3 unterhalb der Zustände beispielsweise notierten Lauflängenbedingungen erfüllen. Diese zusätzlichen Bedingungen können aber mit dem üblichen Dynamic-Programming-Zeitentzerrer nicht ohne weiteres berücksichtigt werden. Daher ist der „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" notwendig.

Realisierung des Run-Length-Limited-Dynamic-Programming-Zeitentzerrers

Siehe Kurzbeschreibung §4. Siehe Figur 4 samt Beschreibung. Der „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" berücksichtigt die Randbedingungen für die Lauflängen der Zustände, in dem das Zustandsdiagramm nach einem im Folgenden beschriebenen Verfahren in eines ohne Lauflängenbedingungen transformiert wird. Sodann wird der übliche Dynamic-Programming-Zeitentzerrer angewendet und das Resultat auf das ursprüngliche Zustandsdiagramm zurück übertragen. Das Verfahren zur Transformation des Zustandsdiagramms ist wie folgt festgelegt:

Die Zustände des transformierten Zustandsmodells werden im Folgenden als Sub-Zustände bezeichnet. Jedem Subzustand ist eindeutig aber im Allgemeinen nicht umkehrbar ein Zustand zugeordnet, wie Figur 4 zeigt.

Jeder Zustand, für den nicht beliebig lange Lauflängen zugelassen sind, im Beispiel (Z2)-(Z4), wird in genau so viele aufeinander folgende Subzustände aufgeteilt, wie seiner Maximallauflänge entspricht. Die Minimallauflängenbedingung wird dadurch berücksichtigt, dass ab der Minimallauflänge Zustandsübergänge zum ersten Subzustand des nächsten Zustandes möglich sind. Zunächst erfolgt also eine bestimmte Anzahl von Subzuständen, welche nacheinander durchlaufen werden müssen, ohne dass Sprünge auf einen anderen Hauptzustand möglich sind. Diese Anzahl ist gleich der minimalen Lauflänge minus Eins. Darauf folgt eine bestimmte Anzahl von Subzuständen, von welchen aus jeweils Sprünge auf den ersten Subzustand des folgenden Zustands möglich sind. Diese Anzahl ist gleich der maximalen Lauflänge minus der minimalen Lauflänge plus Eins. Daher ergibt sich insgesamt als Anzahl der Subzustände die maximale Lauflänge.

Für den ersten und letzten Zustand werden vorzugsweise beliebig lange Lauflängen zugelassen, da diese beiden Zustände hauptsächlich die Nebengeräusche vor und nach dem interessierenden Schallereignis repräsentieren. In diesem Fall wird zur Realisierung der Maximallauflängenbedingung ein einziger Subzustand verwendet, von dem aus Sprünge auf sich selbst möglich sind (Siehe Beispiel Figur 4, Zustände Z1 und Z5). In diesem Fall ist die Anzahl der Subzustände für den Zustand gleich der minimalen Lauflänge. Vorzugsweise werden weiters Minimallauflängen größer Null gefordert, so dass kein Zustand ausgelassen werden kann.

Der optimale Weg durch das transformierte Zustandsdiagramm wird mittels des üblichen Dynamic-Programming-Zeitentzerrers bestimmt, wobei die Subzustände das berechnete Abstandsmaß des Zustandes, dem sie zugeordnet sind übernehmen. Die Lösung wird schließlich wieder auf das ursprüngliche, einfache Left-to-Right-Zustandsdiagramm zurückübertragen.

Der Vorteil der Verwendung des „Run-Length-Limited-Dynamic-Programming-Zeitentzerrers" gegenüber der üblichen „Dynamic-Programming-Zeitentzerrers" *[Levin1993]* besteht darin, dass den Vergleichseinrichtungen bestimmte Flexibilität bei der Zeitentzerrung genommen wird, die sie zur Repräsentation der ihnen zugehörigen Schallschwingungsklasse nicht benötigen. Dadurch liefern sie bei Schallschwingungen, welche nicht zu ihrer Schallschwingungsklasse gehören schlechtere Vergleichsergebnisse, was die Chance erhöht, dass die richtige Vergleichseinrichtung das beste Vergleichsergebnis liefert.

Patentanspruch:

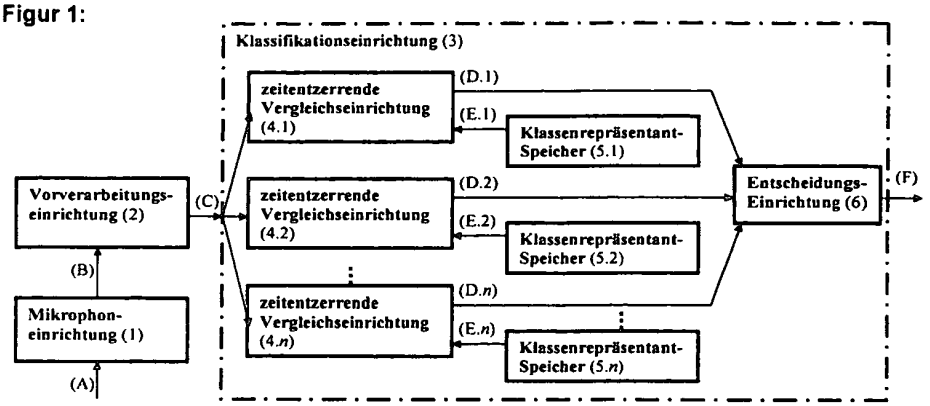
Verfahren zur automatischen Klassifikation von Schallschwingungen, insbesondere Geräuschen und gesprochenen Worten, bei dem eine zu klassifizierende Schallschwingung (A) zunächst mittels einer Mikrophoneinrichtung (1) in eine elektrische Größe (B) und mittels einer Vorverarbeitungseinrichtung (2) zu einer zeitlich-diskreten Merkmalsvektorfolge (C) verarbeitet wird, die anschließend in einer Klassifikationseinrichtung (3) mit den Klassenrepräsentanten (E.1)-(E.*n*) mittels zeitentzerrender Vergleichseinrichtungen (4.1)-(4.*n*) verglichen wird, und die skalaren Vergleichsergebnisse (D.1)-(D.*n*) mittels einer Entscheidungseinrichtung (6) zu einem Klassifikationsergebnis (F) verarbeitet wird, welches zur Steuerung eines technischen Gerätes dient, *dadurch gekennzeichnet*, dass

jeder Klassenrepräsentant (E.1)-(E.*n*) in genau einem zugehörigen Klassenrepräsentant-Speicher (5.1)-(5.*n*), vorzugsweise datenkomprimiert abgespeichert ist und zwar als endliche Folge von Merkmalsvektoren, vorzugsweise 3-15, welche im folgenden „Klänge" genannt werden, sowie für jeden „Klang" die jeweils zulässige Minimalzeitdauer und die jeweils zulässige Maximalzeitdauer,

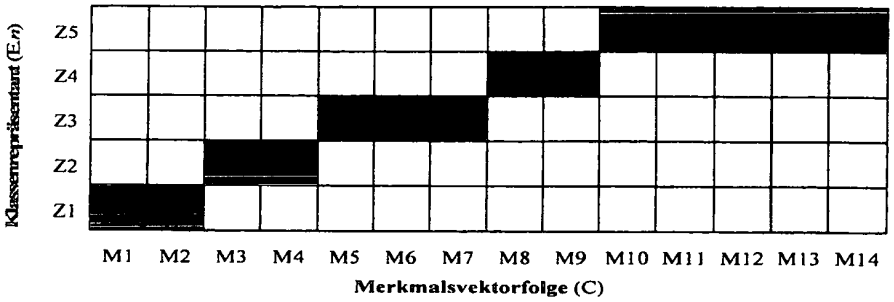
und jede zeitentzerrende Vergleichseinrichtung (4.1)-(4.*n*) durch einen so genannten „Run-Length-Limited-Dynamic-Programming-Zeitentzerrer" realisiert wird, der beim Vergleich der Merkmalsvektorfolge (C) mit einem der Klassenrepräsentanten (E.1)-(E.*n*) die Zeitdauern der „Klänge" des jeweiligen Klassenrepräsentanten derart anpasst, dass das jeweilige Vergleichsergebnis der (D.1)-(D.*n*) so gut wie möglich ist, wobei das Deltamerkmal zum üblichen Dynamic-Programming-Zeitentzerrer darin besteht, dass die Zeitdauern der „Klänge" die Bedingungen erfüllen müssen, innerhalb der ihnen jeweils zugeordnet abgespeicherten Intervalle aus Minimalzeitdauer und Maximalzeitdauer zu liegen.

Hiezu 1 Blatt Zeichnungen

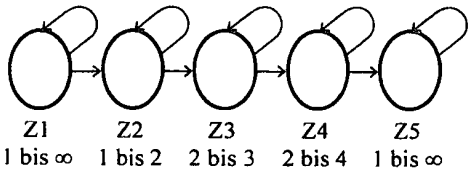
Figur 1:



Figur 2:



Figur 3:



Figur 4: