BMBF

Verbmobil
Verbundvorhaben

# Adaptiver stochastischer Sprache/Pause-Detektor

Manfred Beham
Günther Ruske

**Technische Universität München**

Vm **Report 125**
**Mai 1996**

Mai 1996

Manfred Beham
Günther Ruske

Forschungsgruppe Sprachverarbeitung
Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
Arcisstr.21
80290 München

Tel.: (089) 2892 - 8563
e-mail: ruske@e-technik.tu-muenchen.de

**Gehört zum Antragsabschnitt:** TP3 Spracherkennung und Sprecheradaption

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 C/6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

# Adaptiver stochastischer Sprache/Pause-Detektor

Manfred Beham und Günther Ruske

Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
Arcisstr. 21, 80290 München

## 1 Einführung

Bei der automatischen Spracherkennung muß eine Reihe von einzelnen Teilaufgaben gelöst werden, zu denen auch die Anzeige von Beginn und Ende der gesprochenen Äußerung gehört. Diese Aufgabe ist durchaus nichttrivial, besonders wenn eine fehlerhafte Anzeige zu einer Beschneidung der Äußerung und damit letztlich zu Erkennungsfehlern führt. Daher wird in vielen Spracherkennungssystemen angestrebt, vor Beginn der Äußerung und nach dessen Ende möglichst einen Signalbereich mit aufzunehmen, der eine Sprachpause enthält. Es wird dann versucht, während der Erkennung mit Hilfe spezieller Pausenmodelle die Pausenbereiche zu kennzeichnen und zu eliminieren. Dieses Vorgehen kann man als ”indirekte” Sprache/Pause-Detektion bezeichnen. Sinnvoll einsetzen läßt sich diese Methode vor allem im sogenannten Offline-Betrieb, bei dem die Sprachdaten in Dateien gespeichert sind. Problematisch ist die indirekte Sprache/Pause-Detektion aber im Online-Betrieb, bei dem das eingehende Sprachsignal unmittelbar und schritthaltend verarbeitet werden soll. Hier wirkt sich nachteilig aus, daß die Feststellung des Sprachbeginns erst nach der Erkennung geliefert wird, so daß ein zu spät angezeigter Beginn nicht mehr korrigiert werden kann.

Demgegenüber arbeiten Verfahren zur ”direkten” Sprache/Pause-Detektion ohne Einbeziehung der Spracherkennungsstufe. Bei dieser Lösung wird versucht, die Grenzen der Äußerung unmittelbar aus dem Signalverlauf zu bestimmen. In der Praxis werden hierfür aber meist nur Pegelwerte und Pegeländerungen des Signals verwendet, was zu keiner besonders sicheren Anzeige führt. Insbesondere werden Störgeräusche mit speziellem spektralen Verlauf nicht ausgeschlossen. Im vorliegenden Beitrag wird daher ein Sprache/Pause-Detektor vorgestellt, der unmittelbar nach der spektralen Vorverarbeitungsstufe eingesetzt wird, und der somit die spektrale Verteilung des Signals einbezieht. Durch eine geeignete Zeit- bzw. Längenmodellierung im Detektor lassen sich kurze Sprechpausen innerhalb von Wörtern oder Sätzen überbrücken. Ebenso lassen sich kurze Störgeräusche in Pausenabschnitten ausblenden. Da der Detektor so ausgelegt ist, daß er nur verhältnismäßig wenig Rechenzeit beansprucht, kann er im Prinzip permanent aktiv sein. In diesem Fall kann das Spracherkennungssystem ohne einen externen Start/Stop-Schalter verwendet werden, denn der Sprache/Pause-Detektor ist in der Lage, den Erkennungsvorgang selbständig zu starten und zu beenden; diese Möglichkeit der automatischen Steuerung kann sicher auch für den Einsatz im Telefonbereich bedeutsam sein.

Ein wichtiger Anwendungsbereich ist aber vor allem bei der Steuerung der Kanalkompensation gegeben, die sich inzwischen in vielen Spracherkennungssystemen bewährt hat. Bei der Kanalkompensation werden die logarithmierten Pegelwerte der einzelnen spektralen Kanäle von ihrem Mittelwert befreit, der seinerseits jeweils über ein gewisses Zeitfenster adaptiv nachgeschätzt wird. Hierbei ist es sinnvoll, die Nachschätzung nur im Bereich der Sprache vorzu-



erschienen in: 17. DAGM-Symposium „Mustererkennung“, Bielefeld, 13.-15.5.1995, Springer-Verlag, S.60-67

nehmen und nicht im Bereich von Pausen. Wird diese Unterscheidung nicht vorgenommen, so wird die Kanalkompensation sich in Pausen langsam an den Pausenpegel adaptieren, was beim Einsetzen des Sprachsignals zwangsläufig zu einem Übersteuerungseffekt führt. Bei der Einzelworterkennung kann dies eventuell noch hingenommen werden, nämlich dann, wenn diese Anfangsübersteuerung in jedem Wortmodell mit repräsentiert ist. Bei der Erkennung fließender Sprache auf der Basis von Phonem-Modellen ist dies aber wenig sinnvoll, da nun die Phoneme am Anfang der Äußerung anders repräsentiert wären als dieselben Phoneme im Innern der Äußerung. Diese Effekte lassen sich vermeiden, wenn grundsätzlich nur in Sprachabschnitten adaptiert wird. Hierfür ist es nun ebenfalls sinnvoll, die direkte Sprache/Pause-Detektion einzusetzen.

Der entwickelte Sprache/Pause-Detektor geht von einem stochastischen Prozeß für "Sprache" und "Pause" aus, der durch ein ergodisches "Hidden-Markov"-Modell (HMM) nachgebildet wird. Das HMM verwendet die Kurzzeitspektren und den Signalpegel, die von der Vorverarbeitung geliefert werden. Das ergodische HMM verwendet 2 multivariate Normalverteilungen, die die Emissionswahrscheinlichkeiten für Sprache und Pause widerspiegeln; im Prinzip könnten hier natürlich auch Mixturen aus mehreren überlagerten Normalverteilungen zum Einsatz kommen. Die Zeitmodellierung sowohl für Sprache als auch für Pause wird durch die spezielle Struktur des HMMs erreicht. Ein Bayes'scher Entscheidungsprozeß, der dem HMM übergeordnet ist, berechnet zu jedem Zeitpunkt die Rückschlußwahrscheinlichkeiten für "Sprache" und "Pause" und gibt diese Entscheidung aus. Abhängig von dieser Entscheidung werden die entsprechenden Emissionswahrscheinlichkeiten adaptiv nachgeführt (entsprechend einem "Nachtraining" der Dichtefunktionen). Auf diese Weise bleibt z.B. in langen Pausen die Verteilung für "Sprache" unverändert erhalten, während ebenso innerhalb von Sprachabschnitten die Verteilung für "Pause" nicht verändert wird. Die experimentellen Ergebnisse zeigen, daß dieses Verfahren tatsächlich sehr robust arbeitet.

Mit dieser Sprache/Pause-Anzeige läßt sich die Kanalkompensation so steuern, daß die Adaption nur im Bereich der Sprache wirksam wird, wodurch ein Absinken der Kanalmittelwerte während der Pausenabschnitte vollständig vermieden wird.

# 2 Grundstruktur des Spracherkennungssystems

Im folgenden soll das Spracherkennungssystem nur soweit in groben Zügen beschrieben werden, wie das für den Einsatz des Sprache/Pause-Detektors notwendig ist. Die Vorverarbeitungsstufe berechnet im Abstand von 10 ms aus dem Sprachsignal 30 cepstral geglättete Komponenten des mel-Spektrums, sowie die Nulldurchgangsrate und die Energie (Bild 1). In der anschließenden Kanalkompensation werden die logarithmierten Kanal-Pegelwerte $\log(P_k)$ der Kanäle $k$ vom Mittelwert befreit:

$$\log(P_k)_{komp} = \log(P_k) - E\{\log(P_k)\}, \qquad k = 1 \ldots 30$$

Die laufende Mittelwertbildung pro Kanal erfolgt über ein kürzeres Zeitfenster (running average), allerdings nur im Bereich von Sprache. Nachdem die kompensierten Kanal-Pegelwerte $\log(P_k)_{komp}$ einer linearen Diskriminanzanalyse (LDA) unterworfen wurden, werden diese Merkmalsvektoren dem Spracherkenner angeboten, der mit Hilfe von silbenteil-basierten HMM-Modellen die Wort- und Satzerkennung durchführt [1]. Hierbei kann zwar die Information über Sprache/Pause abgeleitet werden (in Bild 1 gestrichelt gezeichnet), was aber zu unerwünschten Verzögerungen und Rückkopplungseffekten zwischen Erkenner und Kanalkompensation führt. Der hier vorgestellte Sprache/Pause-Detektor verwendet deshalb direkt die mel-Spektren und steuert die Kanalkompensation im Sinne einer Vorwärtsregelung.

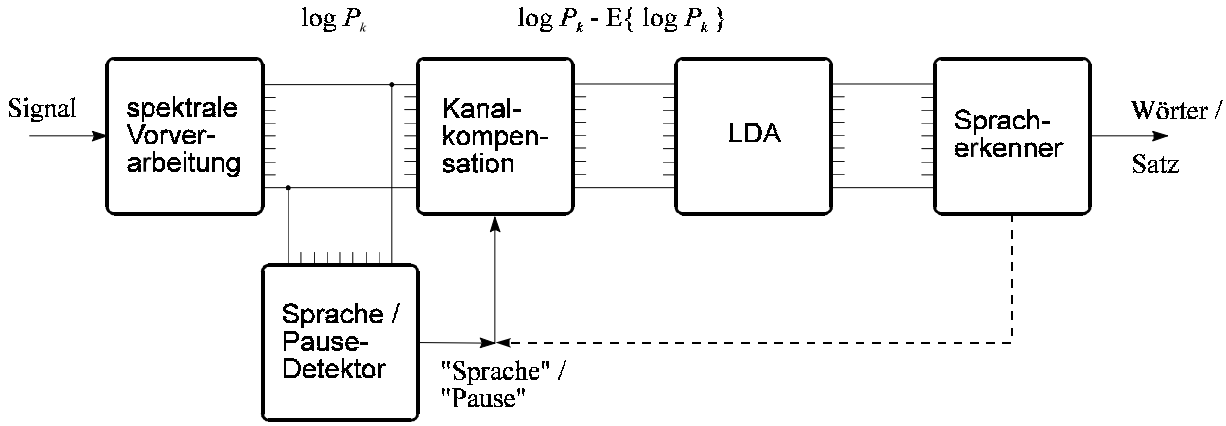


**Bild 1.** Grundstruktur des Spracherkennungssystems.

# 3 Realisierung des Sprache/Pause-Detektors

Die zeitliche Abfolge der Merkmalsvektoren $\vec{x}$ für Sprache und Pause kann mit Hilfe eines stochastischen Erzeugungsprozesses beschrieben werden. Dieser Prozeß wird durch ein HMM für Sprache und ein HMM für Hintergrundgeräusche nachgebildet [2]. Durch Verkettung dieser beiden Modelle zu einem ergodischen HMM können beliebige Abfolgen von {Pause, Sprache, Pause, ... } modelliert werden.

Das Sprach-HMM (S) und Pause-HMM (P) werden jeweils als Modell mit kontinuierlicher Verteilungsfunktion realisiert. Um den Rechenaufwand für den Detektor zu begrenzen, wird jeweils nur eine Normalverteilung mit diagonal besetzter Kovarianzmatrix für Sprache $p(\vec{x}|S)$ und eine für Pause $p(\vec{x}|P)$ verwendet und adaptiv nachgeschätzt. Es wird aber gefordert, daß in die aktuelle Entscheidung zum Zeitpunkt $t$ mehrere Merkmalsvektoren $\vec{x}_{t-N},\ldots,\vec{x}_{t-1},\vec{x}_t$ einbezogen werden; daher müssen die Rückschlußwahrscheinlichkeiten

$$p(\text{"Sprache"}|\vec{x}_{t-N_S},\ldots,\vec{x}_{t-1},\vec{x}_t) \quad \text{und} \quad p(\text{"Pause"}|\vec{x}_{t-N_P},\ldots,\vec{x}_{t-1},\vec{x}_t)$$

für Sprache und Pause berechnet werden. $N_S$ und $N_P$ sind jetzt die gewünschten Beobachtungszeitdauern (in Frames) für Sprache und Pause. Da der Detektor schritthaltend Signale beliebiger Länge verarbeiten soll, darf in die aktuelle Entscheidung nur eine zeitlich begrenzte Vergangenheit bis $t$-$N$ einfließen; die weiter als $t$-$N$ zurückliegenden Merkmalsvektoren $\vec{x}_t$ dürfen dagegen keinen Einfluß auf die aktuelle Entscheidung haben. Das wird durch zwei Übergangsmodelle (P') von S nach P und (S') von P nach S erreicht, die eine dem Fergusson-Modell vergleichbare Struktur haben. Damit können folgende - für die Sprache/Pause-Detektion wichtige - Eigenschaften in der zeitlichen Abarbeitung explizit modelliert werden:

- Kurze Störungen im Hintergrundgeräusch werden nicht als Sprache identifiziert.
- Sprachabschnitte im Signal müssen eine einstellbare Mindeslänge (Frameanzahl) $N_S$ aufweisen.
- Pausen innerhalb von Sprachabschnitten, die eine maximale Länge von $N_P$ nicht überschreiten, werden als Sprache gekennzeichnet (Überbrückung von Pausen).

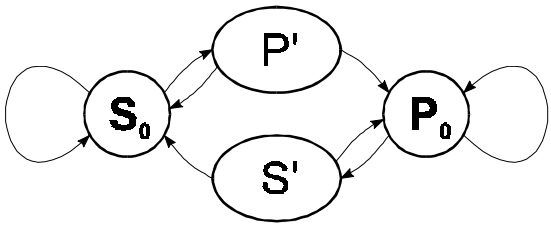


**Bild 2.** Grundstruktur des ergodischen Sprache/Pause-HMMs.

Bild 2 zeigt die grundlegende Modellstruktur bestehend aus der Verkettung der vier Modelle P, S, P' und S' zu einem ergodischen HMM. Die Modelle P und S bestehen jeweils nur aus einem Zustand, während die Übergangsmodelle aus mehreren Zuständen aufgebaut sind, die aber ebenfalls nur die beiden Emissionen $p(\vec{x}|S)$ und $p(\vec{x}|P)$ beinhalten.

### 3.1 Merkmalsvektor des Sprache/Pause-Detektors

Die 30 Komponenten des mel-Spektrums werden durch Zusammenfassen von jeweils 5 benachbarten Kanälen auf 6 reduziert, was für eine Sprache/Pause-Unterscheidung völlig ausreichend ist. Von diesen 6 spektralen Pegeln, der Energie und der Nulldurchgangsrate werden noch die zeitlichen Ableitungen (Delta-Merkmale) gebildet, was schließlich einen 16-dimensionalen Merkmalsvektor $\vec{x}_t$ ergibt.

### 3.2 Modellstruktur

Die Forderung nach einer Mindestverweildauer $N \cdot \Delta t$ ($\Delta t = 10$ ms) in einem Zustand des HMMs kann durch eine Aufteilung dieses Zustands in eine Abfolge von *N* Subzuständen, die keinen Selbstübergang haben, realisiert werden. In jedem dieser Subzustände erfolgt die gleiche Emission, d.h. alle Subzustände haben die gleiche Verteilungsfunktion. Diese Abfolge von Zuständen mit gleicher Emission ist in den Übergängen zwischen den Hauptzuständen $S_0$ und $P_0$ realisiert, siehe Bild 3. Der einmalige, erste Einsprung erfolgt in $P_0$ bei *e*. Für den Übergang von $P_0$ zu $S_0$ (das ist *S'*) bedeutet das z.B., daß $N_S$-mal ein Vektor für Sprache mit $p(\vec{x}|S)$ emittiert werden muß, bevor der Hauptzustand für Sprache erreicht wird, der dann durch einen Selbstübergang mit der Übergangswahrscheinlichkeit $p_{So,So} = 1{,}0$ beliebig lang andauern kann. Analog dazu ist der Übergang von $S_0$ nach $P_0$ durch $N_P$ Zustände mit der Emission $p(\vec{x}|P)$ realisiert. Durch die fehlenden Selbstübergänge in den Zwischenzuständen und der damit erzwungenen Zustandsabfolge sind diese HMMs nicht mehr von 1. Ordnung.

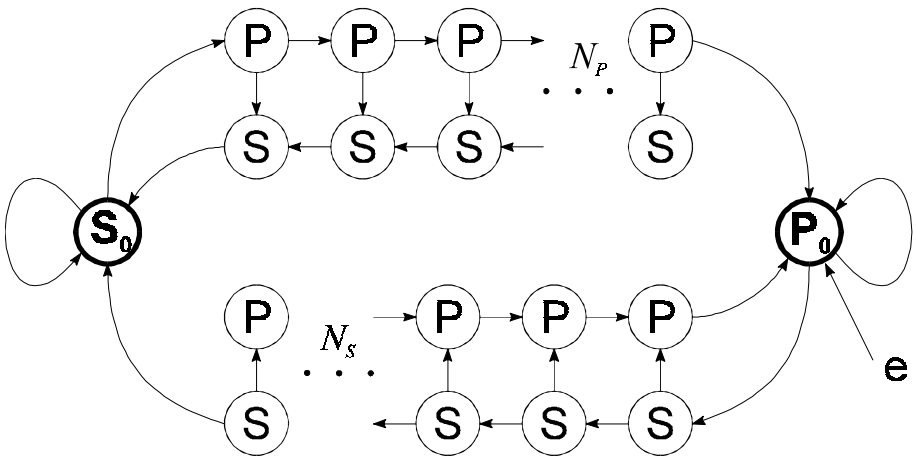


**Bild 3.** Vollständige Modellstruktur des realisierten Sprache/Pause-HMMs.

Zusätzlich existiert innerhalb des Übergangs jeweils noch ein Rückpfad in den entsprechenden Hauptzustand, der dafür sorgt, daß eine kurze Störung des momentan andauernden Hauptzustands nicht zu einem Wechsel zwischen den Hauptzuständen führt. Anschaulich kann man sich die Wirkung dieses Rückpfads so vorstellen, daß eine kurzzeitige Störung des jeweiligen Hauptzustands - also die Emission von $n$ Vektoren des jeweiligen anderen Zustands - wieder kompensiert werden kann, indem die gleiche Anzahl von Vektoren passend zum jeweiligen Hauptzustand emittiert wird. Solange die Zahl der Vektoren $n$ in der Störung kleiner ist als die geforderte Mindestanzahl $N_S$ bzw. $N_P$, erfolgt kein Wechsel in der Anzeige des Detektors.

### 3.3 Abarbeitung und Backtracking

Zur Vereinfachung soll im folgenden nur noch der Übergang von $P_0$ zu $S_0$ (das ist $S'$) beispielhaft betrachtet werden, da der andere Übergang dazu symmetrisch aufgebaut ist. Bild 4 zeigt einen entsprechenden Ausschnitt aus dem Trellis-Diagramm zu diesem Modell. Da alle möglichen Pfade in diesem Modell in den beiden Hauptzuständen $P_0$ und $S_0$ rekombinieren, müssen für die Berechnung der Wahrscheinlichkeiten die Pfade nur bis zum letzten Hauptzustand verfolgt werden. Ohne Berücksichtigung des Rückpfades ergibt sich zu einem Zeitpunkt $t$ die Wahrscheinlichkeit einer Folge von $N_S$ Vektoren im Zustand $S_0$ nach Viterbi zu

$$p(\vec{x}_{t-N_S},\ldots,\vec{x}_t|S_0) = \max\left\{\begin{array}{c} p(\vec{x}_{t-N_S-1},\ldots,\vec{x}_{t-1}|S_0)p(\vec{x}_t|S)\ , \\ p(\ldots,\vec{x}_{t-N_S-1}|P_0)p(\vec{x}_{t-N_S}|S)\ldots p(\vec{x}_{t-1}|S)p(\vec{x}_t|S) \end{array}\right\}$$

In den Zwischenzuständen $S_i$ ergibt sich analog dazu diejenige Wahrscheinlichkeit, die jeweils nur eine Folge von $n < N_S$ Vektoren berücksichtigt. Für eine Entscheidung des Detektors auf Sprache wird nun gefordert, daß zu einem Zeitpunkt $t$

$$p(\vec{x}_{t-N_S},\ldots,\vec{x}_t|S_0) \ \rightarrow \ \text{Maximum}$$

wird. Das bedeutet, daß von allen in Konkurrenz laufenden Pfaden derjenige mit der gewünschten Mindestlänge $N_S$ gegenüber allen kürzeren die größte Wahrscheinlichkeit haben muß. Falls diese Forderung nicht erfüllt ist, wird die endgültige Entscheidung auf einen späteren Zeitpunkt verschoben.

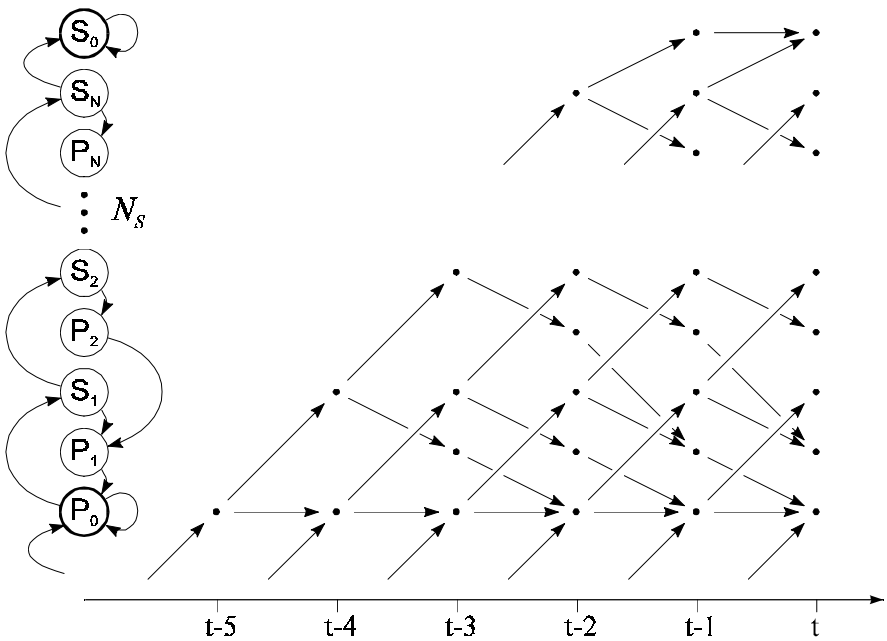


**Bild 4.** Ausschnitt aus dem Trellis-Diagramm für den Pause-Sprache-Übergang.

In der Viterbi-Rekursion müssen natürlich alle möglichen Pfade durch alle Zustände $z_i$ der gezeigten Modellstruktur abgearbeitet werden. Für jeden Zeitpunkt $t$ erfolgt dann anhand der log-Wahrscheinlichkeiten ($\alpha$-Score) eine Entscheidung nach folgender Regel:

$$\left["\mathbf{P}\text{ause}"\right]_t \quad \text{falls} \quad \max_i\{\alpha_t(z_i)\} = \alpha_t(P_0)$$

$$\left["\mathbf{S}\text{prache}"\right]_t \quad \text{falls} \quad \max_i\{\alpha_t(z_i)\} = \alpha_t(S_0)$$

$$\left["\mathbf{A}\text{ufgeschoben}"\right]_t \quad \text{sonst}$$

Das heißt, für eine Anzeige muß das Maximum in den Hauptzuständen $P_0$ oder $S_0$ liegen. Die spezielle Modellstruktur in den Übergängen sorgt dafür, daß die Entscheidung "Aufgeschoben" nur für eine begrenzte zeitliche Dauer erfolgt, da alle möglichen Pfade nach endlicher Dauer ($N = \max(2N_S, 2N_P)$) wieder in einem Hauptzustand $P_0$ oder $S_0$ rekombinieren. Deshalb kann die Viterbi-Suche schritthaltend arbeiten und zum aktuellen Zeitpunkt $t$ eine Entscheidung für die vergangenen $N$ Frames fällen, ohne von einem absoluten Endzeitpunkt alle Frames im Backtracking rückzuverfolgen. Es ist sichergestellt, daß ausgehend vom letzten Hauptzustand in die Berechnung der bedingten Wahrscheinlichkeit und damit letztlich auch in die noch zu berechnende Rückschlußwahrscheinlichkeit eine bestimmte Mindestanzahl von Vektoren einbezogen worden ist.

Die endgültige Entscheidung für [”Sprache”/”Pause”]$_{t-N}$ eilt also dem aktuellen Zeitpunkt $t$ um einen maximalen Versatz von $N$ Frames nach. Deshalb wird ein Puffer (Bild 5) für die Vektoren $x_{t-N}, \ldots x_{t-1}, x_t$ (FIFO) benötigt. Die aufgeschobenen Entscheidungen können dann durch ein vereinfachtes Backtracking nachträglich gefällt und die entsprechenden Vektoren als Sprache oder Pause klassifiziert werden. Dazu ist es nicht notwendig, den Pfad in der Viterbi-Trellis zu speichern und anschließend rückzuverfolgen, da ausgehend von der letzten gültigen Entscheidung (”Sprache”/”Pause”) immer eindeutig bekannt ist, welcher Übergang zur aktuellen gültigen Entscheidung geführt hat. Folgende Tabelle zeigt die möglichen Zuordnungen der aufgeschobenen Entscheidung (”Aufg.”):

| *t-N-1* | *t-N* | ... | *t-1* | *t* | Zuordnung |
|---|---|---|---|---|---|
| ”Pause” | ”Aufg.” | ... | ”Aufg.” | ”Pause” | ”Aufg.” → ”Pause” |
| ”Sprache” | ”Aufg.” | ... | ”Aufg.” | ”Pause” | ”Aufg.” → ”Pause” |
| ”Pause” | ”Aufg.” | ... | ”Aufg.” | ”Sprache” | ”Aufg.” → ”Sprache” |
| ”Sprache” | ”Aufg.” | ... | ”Aufg.” | ”Sprache” | ”Aufg.” → ”Sprache” |

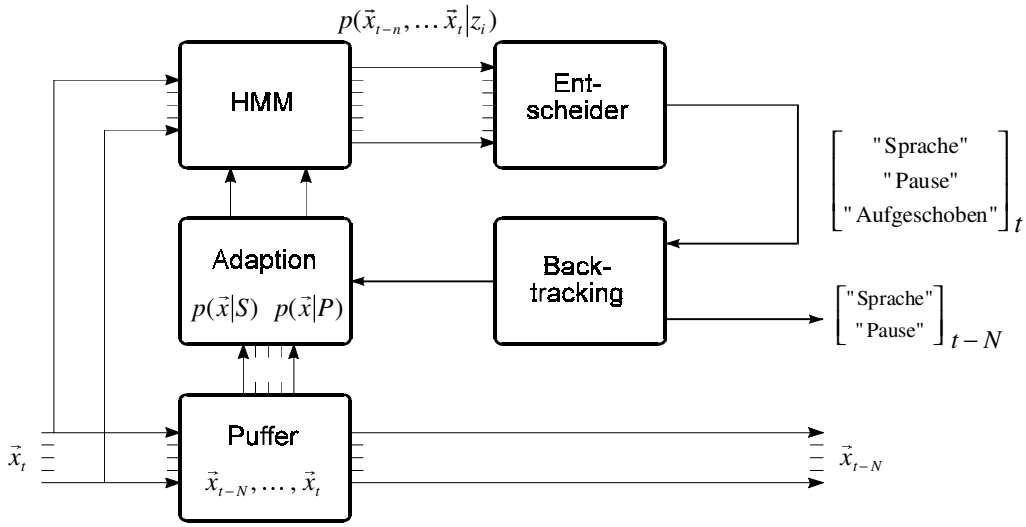


**Bild 5.** Aufbau des gesamten Sprache/Pause-Detektors.

Die gesamte Struktur des Sprache/Pause-Detektors ist in Bild 5 dargestellt. Für die Viterbi-Rekursion wird jeweils nur das aktuelle Muster $\vec{x}_t$ benötigt. Nach dem Backtracking steht die Klassifikation (”Sprache”/”Pause”) für den Zeitpunkt *t-N* zur Verfügung. Das gespeicherte Muster $\vec{x}_{t-N}$ kann im Spracherkennungssystem weiter verarbeitet sowie als Ausgangspunkt für die Adaption der Verteilungen des Sprache/Pause-Detektors verwendet werden.

### 3.4 Adaptive Bestimmung der Verteilungen

Für den gesamten Sprache/Pause-Detektor müssen nur die beiden Verteilungen $p(\vec{x}|S)$ und $p(\vec{x}|P)$ bestimmt werden. Wie bei allen nicht-überwachten Lernverfahren kann man unter der Annahme, daß die Erkennungsrate des Detektors besser als 50 % ist, - also im Mittel häufiger richtig auf Sprache bzw. Pause entschieden wird - diese Verteilungen mit der eigenen Entscheidung des Detektors adaptiv nachschätzen. Um Rückkopplungseffekte zu vermeiden, kann man zum Adaptieren der Verteilungen nur solche Muster verwenden, die mit ausreichender Sicherheit klassifiziert wurden. Dazu ist es notwendig, die Rückschlußwahrscheinlichkeit für ”Sprache”

$$p(\text{"Sprache"}\,|\,\vec{x}_{t-N_S},\ldots,\vec{x}_{t-1},\vec{x}_t) = \frac{\exp(\alpha_t(S_0))}{\exp(\alpha_t(S_0))+\exp(\alpha_t(P_0))}$$

zu berechnen (analog für ”Pause”) und mit einem vorgegebenen Schwellwert zu vergleichen. Nur falls die Rückschlüsse diese Schwelle überschreiten, werden die entsprechenden Muster zur gleitenden Mittelwerts- und Varianzanzberechnung der Emissionsverteilungen verwendet; ansonsten bleiben die Verteilungen unverändert. Für die gleitende Adaption hat sich eine Zeitkonstante von ca. 2s bewährt.

## 4 Ergebnisse

Die Zuverlässigkeit des Sprache/Pause-Detektors kann am besten ermittelt werden, indem Beginn und Ende der angezeigten Sprachabschnitte mit gekennzeichneten (tatsächlichen) Grenzen verglichen werden. Der mittlere Versatz dieser Anfangs- und Endpunkte ist ein Maß für die Genauigkeit des Detektors.

Bild 6 zeigt die relativen Häufigkeiten der Abweichung von gekennzeichnetem zu detektiertem Beginn und Ende einer Sprachäußerung. Der Test wurde mit fließender Sprache (ganze Sätze) durchgeführt. Das Testmaterial beinhaltet ca. 100 Äußerungen von 100 verschiedenen Sprechern. Der Beginn und das Ende der Äußerung wurden in diesem Material automatisch mit einem Spracherkennungssystem gekennzeichnet, das ein Pausenmodell beinhaltet. Die beiden Verteilungen zeigen, daß der Detektor sehr zuverlässig arbeitet. Durch die vorgegebene Modellstruktur wird erreicht, daß Störgeräusche wirkungsvoll unterdrückt werden und im Mittel die detektierten Grenzen sehr gut mit den gekennzeichneten übereinstimmen. Der realisierte Detektor hat für das gesamte Spracherkennungssystem die gutmütige Eigenschaft, den Anfangspunkt einer Äußerung eher zu früh und den Endpunkt eher zu spät anzuzeigen. Der umgekehrte Fall tritt praktisch nicht (ca. 1%) auf. Dadurch ist sichergestellt, daß die Sprachäußerung nicht beschnitten wird.

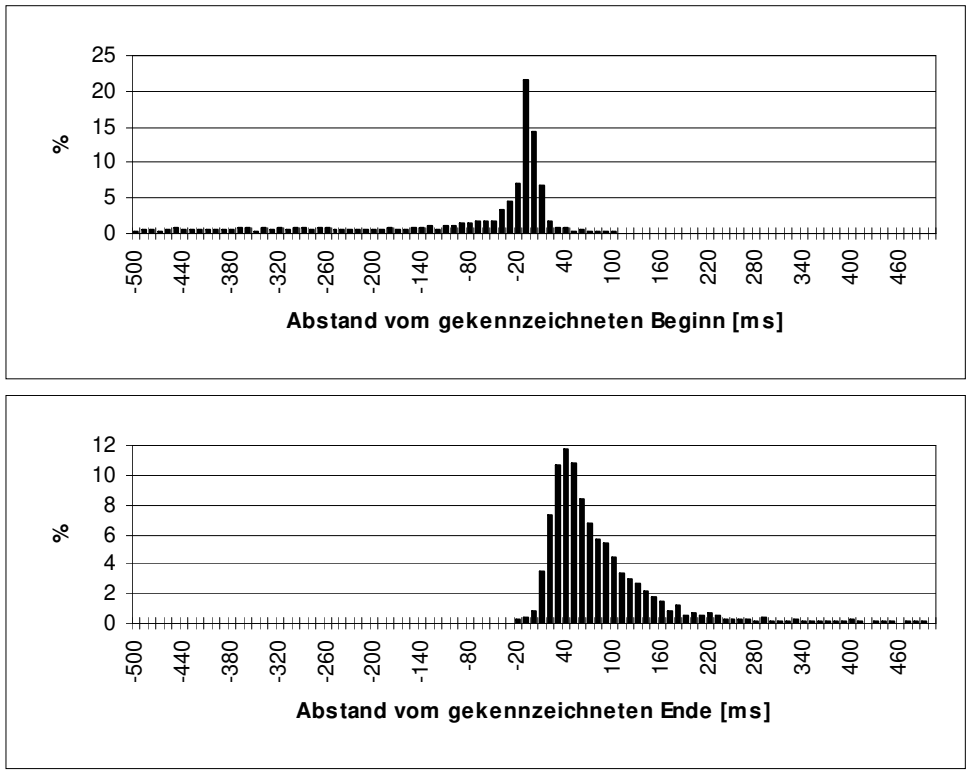


**Bild 6.** Relative Häufigkeit des Abstandes der Detektoranzeige vom gekennzeichneten Start- und Endpunkt.

# 5 Ausblick

Es konnte gezeigt werden, daß der Sprache/Pause-Detektor in der bestehenden Realisierung bereits sehr zuverlässig und robust arbeitet. Eine weitere Verbesserung wäre bei der Verwendung von multimodalen Normalverteilungen denkbar. Z.B. könnten dann verschiedene Arten von stationären Hintergrundgeräuschen in verschiedenen Moden des HMMs repräsentiert werden. Ein Problem bereitet dann allerdings die schritthaltende Adaption dieser Verteilungen, da zusätzlich noch entschieden werden müßte, welcher Mode nachgeschätzt werden soll. Außerdem wächst damit der Rechenaufwand und es ist fraglich, ob der Detektor mit aufwendigeren Verteilungen noch echtzeitfähig ist.

# 6 Literatur

[1] Plannerer, B., Einsele, T., Beham, M. und Ruske, G., A continuous speech recognition system integrating additional knowledge sources in a data-driven beam search algorithm. Int. Conference on Spoken Language Processing ICSLP-94, Yokohama/Japan, 18.-22. Sept. 1994, S01-5.1 - S01-5.4.

[2] Acero, A., Crespo, C., de la Torre, C. und Torrecilla, J.C, Robust HMM-Based Endpoint Detector. Eurospeech 1993, Berlin, 1551 - 1554.

[3] Rabiner, L.R., A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, Vol. 77, No. 2, 1989, 257 - 286.