— Studienarbeit —
Schätzung des Signal-Rausch-Verhältnisses

Michael Schoch

Betreuer: Dr. Paul Duchnowski

Institut für Logik, Komplexität und Deduktionssysteme
Lehrstuhl Professor A. Waibel
Fakultät für Informatik
Universität Karlsruhe (TH)
D-76128 Karlsruhe

2. Februar 1995

Inhaltsverzeichnis

1. Einleitung

Bei der Entwicklung maschineller Spracherkennung wurden in letzter Zeit erhebliche Fortschritte erzielt. Beim Einsatz dieser Systeme in einer Umgebung in der Hintergrundgeräusche auftreten, was beim praktischen Einsatz der Normalfall sein wird, zeigt sich jedoch, daß die Erkennungsleistung stark von der Qualität des akustischen Signals abhängt. Deshalb wird seit einiger Zeit zur Realisierung robuster Spracherkennung Lippenlesen als Unterstützung eingesetzt. Die dabei gewonnenen zusätzlichen visuellen Informationen sind unabhängig von der jeweiligen akustischen Aufnahmesituation.

Auf welche Art die visuellen und akustischen Informationen kombiniert werden, ist zur Zeit Gegenstand weiterer Untersuchungen. Dabei ist die Gewichtung der visuellen Informationen abhängig von der jeweiligen Aufnahmesituation. Stehen saubere akustische Signale zur Verfügung, sollten hauptsächlich diese verwendet werden. Nur bei starken Hintergrundgeräuschen sollte auf die visuellen Informationen ein stärkeres Gewicht gelegt werden. Um dies zu ermöglichen muß die jeweilige Aufnahmesituation beurteilt werden, die Stärke der Hintergrundgeräusche muß eingeschätzt werden. Diese Einschätzung soll natürlich das System selbst vornehmen, und nicht vom Anwender gefordert werden.

Die Zielsetzung dieser Arbeit war es, ein Verfahren zu implementieren, das die Qualität der akustischen Signale einschätzt. Dieses Verfahren soll dann innerhalb des Lippenlesersystems [DUC94], [DUC95], [Mei], das zur Zeit am Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe, entwickelt wird, eingestetzt werden. Das eingesetzte Verfahren sollte ohne zusätzlichen Hardwareaufwand realisierbar sein. Diese Einschränkung bezieht sich vor allem auf den Einsatz eines Mikrophones. Verfahren die auf Autokoorelation des Sprachsignals aufbauen wurden deshalb nicht betrachtet. Das Verfahren sollte auch keine komplizierten Berechnungen erfordern, da das gesamte Lippenlesersystem auf Realzeiteinsatz ausgelegt ist. Es wurden die folgenden drei Ansätze untersucht:

- Bestimmung in Sprechpausen
  Die Stärke der Hintergrundgeräusche wird in den Sprechpausen ermittelt, und bis zur nächsten Pause als konstant angenommem. Dazu müssen aber die Wortgrenzen ermittelt werden.[RAB]

- Schätzung mit Hilfe eines neuronalen Netzes
  Basierend auf [TAM] in dem ein neuronales Netzes zur Rauschunterdrückung einge-

setzt wird, wurde ein neuronales Netz zur Schätzung der Hinergrundgeräusche trainiert.

- Schätzung mittles statistischer Analyse der Amplitudenverteilung
  H.G. Hirsch stellt in [HIR] ein Verfahren vor, das mittles statistischer Analyse der Energieverteilung in Unterbändern, die Stärke der Hintergrundgeräusche schätzt.

Diese Ausarbeitung ist wiefolgt gegliedert: In Kapitel 2 werden die Grundlagen über die verwendeten Signalrepräsentationen und Kenngrößen dargestellt. In Kapitel 3 werden die verschiedenen Ansätze vorgestellt und teilweise die Implementierungen erläutert. Kapitel 4 beinhaltet einen Vergleich der Verfahren. Kapitel 5 enthält eine kurze Zusammenfassung und einen Ausblick.

2. Verwendete Sprachsignale, Störgeräusche

Sämtliche verwendete Sprachdaten wurden mit einem A/D-Wandler der Firma Desklab (Gradientbox) digitalisiert. Dabei wurde eine Samplerate von 16000 Hz verwendet, wobei jeder Signalwert mit 8-Bit kodiert wurde. Die zur Zeit im Erkenner verwendeten Sprachdaten, bestehen aus kontinuierlich buchstabierten deutschen Wörtern. Abbildung 2.1 zeigt den Verlauf der digitalisierten Signalwerte X(t) für die Sequenz W-I-E-N. Diese Daten wurden in einer relativ ruhigen Umgebung aufgenommen, in der nur Störgeräusche auftreten sollten, die durch die Computeranlage verursacht werden. Diese Daten werden nun idealisiert als reine Sprachdaten S(t) interpretiert. Zu diesen Daten werden nun künstlich erzeugte oder natürliche Störgeräusche N(t) addiert:

$$X(t)=S(t)+N(t)$$

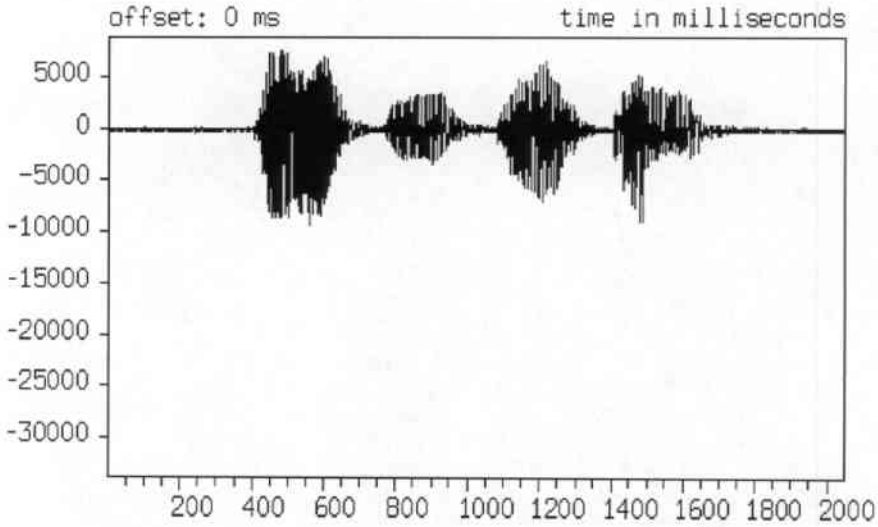

Abbildung 2.1: W-I-E-N 25db

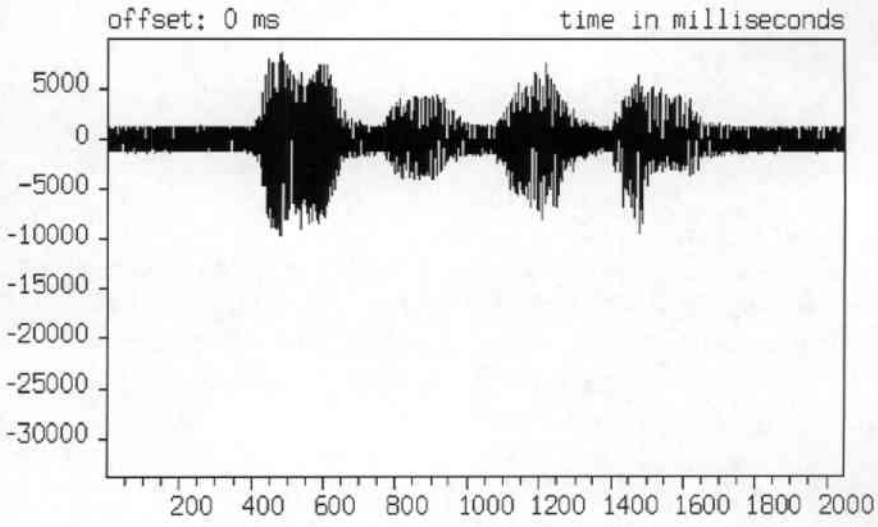

Abbildung 2.2: W-I-E-N 7db

Abbildung 2.2 zeigt die, durch Addition von künstlichen gleichverteilten weißen Rauschen entstandene, Buchstabensequenz W-I-E-N. Da die Energie von N(t) und S(t) bekannt berechnet werden können, kann nun mit Hilfe dieser Daten X(t) die Ergebnisse der einzelnen Verfahren, die X(t) als Eingabe erhalten, überprüft werden. Veränderungen des Signals R(t), die durch eine Faltung das Signal verändern (z.B. Einfluß eines Übertragungskanals), werden in dieser Arbeit nicht betrachtet. Als Kenngröße für die Qualität der akustischen Signale wird die Signal-to-Noise-Ratio (SNR) in einem Zeitabschnitt [t1,t2] bestimmt:

$$SNR = 10 * log_{10} \frac{\sum_{t=t1}^{t2} N(t)^2}{\sum_{t=t1}^{t2} S(t)^2} [\text{db}] \quad (1)$$

S(t)=Signalwerte des Sprachsignals
N(t)=Signalwerte der Hintergrundgeräusche

Analog kann die Energie im Zeitabschnitt [t1,t2] des Signals bestimmt werden:

$$E(S)_{[t1,t2]} = \sum_{t=t1}^{t2} S(t)^2$$

Damit kann die Gleichung (1) durch die Energien ausgedrückt werden:

$$SNR = \frac{E(N)_{[t1,t2]}}{E(S)_{[t1,t2]}} [db]$$

Die Sprachdaten wurden mit folgenden Störgeräuschen belegt:

- Konstantes weißes Rauschen mit 19-5db
- Radiohintergrundgeräusche (Musik) 10-25db
- Stellmotor der Kameraführung, die für das Lippenlesersystem verwendet wird, mit 10-25db
- In der Stärke nicht konstantes weißes Rauschen

Dabei wurden die Radio bzw Kamerastörgeräusche mit demselben System aufgenommen, mit dem auch die eigentlichen Sprachdaten für das Gesamtsystem aufgenommen werden. Diese Daten wurden dann in verschiedenen Stärken auf die Sprachdaten addiert. Das weiße Rauschen wurde mit Hilfe des Programms doNOISE von Christoph Bregler erzeugt, das ein auf dem gesamten Frequenzbereich gleichverteiltes weißes Hintergrundrauschen erzeugt.

Es konnte auf eine breite Basis von Sprachdaten zurück gegriffen werden, die zum Testen und Trainieren des Lippenlesen-Systems angelegt wurde.

3. Beschreibung der verschiedenen Ansätze

3.1 Bestimmung der SNR in Sprechpausen

3.1.1 Prinzipieller Ansatz

Bei diesem Ansatz wird versucht, Zeitabschnitte zu finden, in denen keine Sprache vorhanden ist. Diese treten zum Beispiel zwischen einzelnen Wörtern oder einzelnen Buchstaben auf. Die, in diesen Abschnitten vorhandenen, Signale werden den Störgeräuschen zugeordnet. Aufgrund der Stärke dieser Signale wird eine Schätzung der aktuellen SNR durchgeführt.

Die Stärke der Störgeräusche wird nun als konstant angenmonnem, bis in der nächsten Sprechpause wieder eine Bestimmung möglich ist.
Die Hauptschwierigkeit besteht nun darin, die Pausen zu bestimmen. Dies konnte bisher nicht befiedigend gelöst werden. Für die Bestimmung der SNR müssen die Wortgrenzen nicht exakt bestimmt werden. Es ist ausreichend, einen Abschnitt zu finden in der sicher keine Sprache vorhanden ist. Dies bereitet weniger Schwierigkeiten als das Auffinden exakter Wortgrenzen, da mit einem großen Toleranzbereich gearbeitet werden kann, falls die Sprechpausen ausreichend lange sind.

3.1.2 Ein Algorithmus zur Bestimmung von Sprechpausen

Nachstehend wird ein Algorithmus vorgestellt, der versucht Sprechpausen zu bestimmen, indem zwei Größen des Sprachsignals, Energie und Nulldurchgangsrate des Signals, betrachtet werden. Dieser Algorithmus wurde ursprünglich zur Bestimmung der Start-/Endpunkte von isolierten Sprachsequenzen entwickelt.

- Energie des Signals X im Intervall n:
  E(n)=$E(X)_{[n*10ms-5ms,n*10ms+5ms]}$
- Nulldurchgangsrate z(n):
  z(n) ist als Anzahl der Nulldurchgänge des Signals im Intervall [n*160-80,n*160+80] definiert.

Die Energie und Nulldurchgangsrate lassen sich durch einfache Berechnungen aus den Signalwerten bestimmen. Von diesen Größen werden Grenzwerte berechnet, die dann eine Unterscheidung Sprache/Stille ermöglichen.
Die Abbildung 3.2 zeigt die Verläufe der beiden Größen Energie und Nulldurchgangsrate für die Buchstabenfolge p-a. An diesem Beispiel zeigt sich, daß , neben der Energiewerte, die Nulldurchgangsrate als zweiter Indikator für das Vorhandensein von Sprache betrachtet werden muß. Vor dem Ansteigen der Energiewerte zum Zeitpunkt T2=300ms kann ein Anstieg der Nulldurchgangsrate zum Zeitpunkt T1=210ms beobachtet werden. Dort liegt auch der wirkliche Anfang des Buchstabens. Es gibt weitere Beispiele bei denen die alleinige Betrachtung der Energie nicht ausreicht:

- Schwache Reibelaute (z.B. f,h), am Anfang oder Ende von Sprachsegmenten.
- Schwache Explosivlaute (z.B. p,t,k), am Anfang oder Ende von Sprachsequenzen.
- Von stimmhaft zu stimmlos wechselnde Reibelaute am Ende von Sprachsequenzen

Durch die Betrachtung der Nulldurchgangsrate, als Indikator für stimmlose Sprache, kann eine Start-/Endpunktbestimmung auch in diesen Fällen durchgeführt werden.

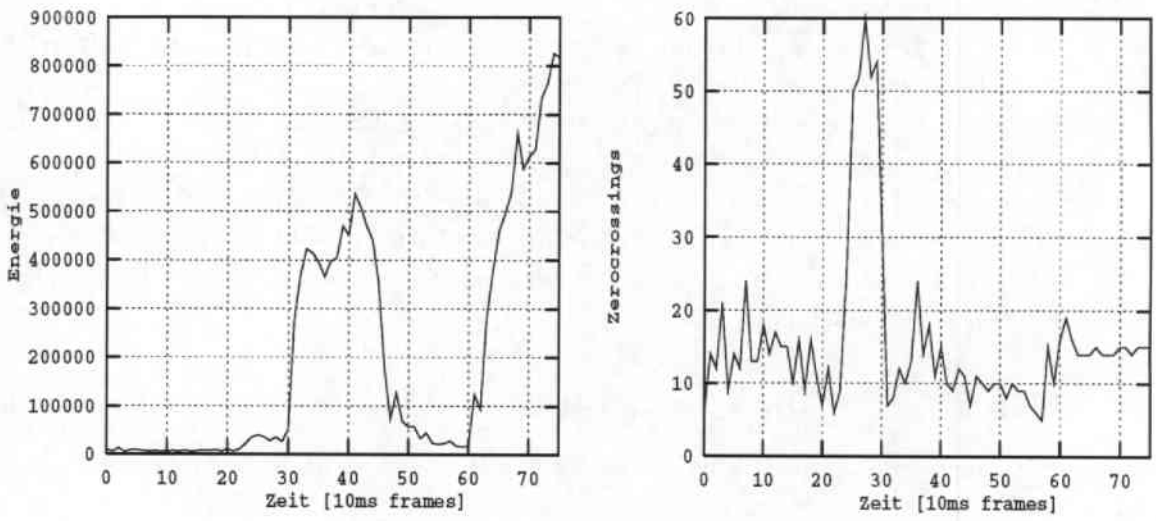

Abbildung 3.1: Zerocrossingrate und Energie

Der Algorithmus besteht im wesentlichen aus zwei Schritten. Sprachgrenzen werden aufgrund der Energiewerte bestimmt, diese werden dann nach Untersuchung der Nulldurchgangsraten in ihrer Umgebung korrigiert.

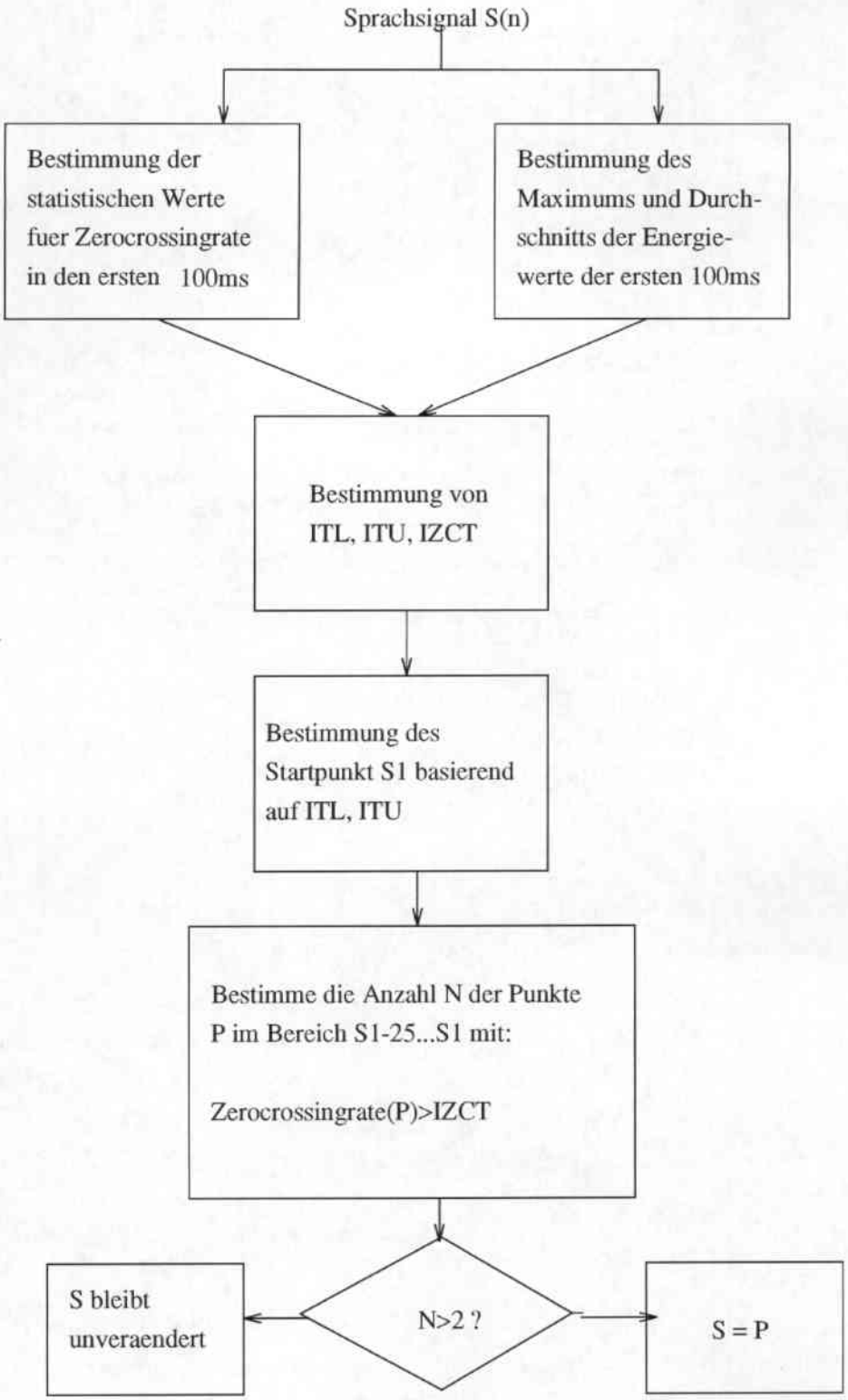

Abbildung 3.2: Startpunktbestimmung

Nachstehend werden die einzelnen Schritte zum Auffinden des Anfangs eines Sprachsegments beschrieben. Der Ablauf wird in Abbildung 3.1 noch einmal graphisch aufgezeigt. Die Suche nach dem Ende einer Sprachsegments verläuft analog.

Der Algorithmus besteht aus folgenden Schritten:

1. Berechnung der Energie E(n)

2. Initiale Bestimmung der Parameter
   Aus den ersten 100ms des Sprachsegments, die keine Sprache enthalten dürfen, werden folgende Parameter bestimmt, mit deren Hilfe dann Wortgrenzen gesucht werden:

   - Grenzwert für die Nulldurchgangsrate:

     $$IZCT = MIN(25, \overline{IZC} + 2\sigma_{IZC})$$

     IZC=Durchschnittliche Nulldurchgangsrate der ersten 10 10ms-Frames

   - Oberer und unterer Grenzwert für die Energiewerte:

     ITL=MIN(I1,I2)
     ITU=5*ITL
     I1=0.03*(IMX-IMN)+IMN
     I2=4*IMN
     IMX=Maximum der Energiewerte im gesamten Intervall
     IMN=Durchschnittlicher Energiewert der ersten 100ms

3. Suche nach dem ersten Punkt P, der als Wortanfang in Frage kommt, für den gilt: $(E(P) > ITL) \wedge (\exists P' : P' > P \wedge E(P') \geq ITU) \wedge (\forall P'' : P < P'' < P' : ITL < E(P'') < ITU)$

4. Ausgehend von P werden die letzten 25 Punkte untersucht. Es werden Punkte gesucht für die gilt:

   $$z(P) \geq IZCT$$

   Werden mehr als zwei Punkte gefunden, die diese Bedingung erfüllen, wird der letzte gefundene Punkt, als entgültiger Wortanfang bestimmt.

In den gefundenen Segmenten ohne Sprache lassen sich die Stärke und Charakteristika der vorhandenen Störgeräusche sehr genau bestimmen. Trotzdem wurde dieses Verfahren nicht bis zum praktischen Einsatz im Erkenner entwickelt, da sich ein paar grundsätzliche Probleme stellen:

- Stille am Anfang des Eingabesatzes vorrausgesetzt
  Um die initiale Parameterbestimmung ausreichend genau durchführen zu können, muß am Anfang des Sprachsegment ein Bereich von 100ms ohne Sprache vorhanden sein, um die benötigten Grenzwerte berechnen zu können.

- Nicht stationäre Hintergrundgeräusche
  Bei stark ansteigenden Hintergrundgeräuschen können Sprechpausen nicht mehr ermittelt werden, da die ursprünglich bestimmten Grenzwerte für die Energie in Sprechpausen, auch durch die Hintergrundgeräusche überschritten werden.

Verbesserungen des beschriebenen Algorithmus finden sich in [LAM], [REA], [MAK]. Diese Verbesserungen beziehen sich vor allem auf eine genauere Bestimmung der Grenzwerte für die Unterscheidung Sprache/Keine Sprache. Dies wurde aber in dieser Arbeit aus Zeitgründen nicht mehr untersucht.

3.2 Neuronales Netz

Aufbauend auf [TAM] wurde ein Neuronales Netz trainiert, um eine SNR Schätzung zu ermöglichen. Die folgenden Abschnitte enthalten eine kurze Beschreibung des verwendeten Neuronalen Netzes. Es wird nicht auf die Theorie der neuronalen Netze eingegangen, bzw. bei Benutzung von Standartverfahren diese nicht beschrieben.

3.2.1 Vorverarbeitung

Als Eingabewerte für das Netz wurde ein Vektor aus 17 Elementen bestimmt, der für 100ms Zeitabschnitte berechnet wurde. Dieser setzt sich wiefolgt zusammen:

- 16 Melscale Koeffizenten [WAI]
- Energie des Signals E(X) (normalisiert)

Die Melscale Koeffizenten werden auch als Eingabe für den Buchstabiererkenner verwendet. Es sind deshalb keine zusätzlichen Berechnungen notwendig. Die Energie E(X) kann ohne großen Zeitaufwand berechnet werden.

3.2.2 Netzbeschreibung

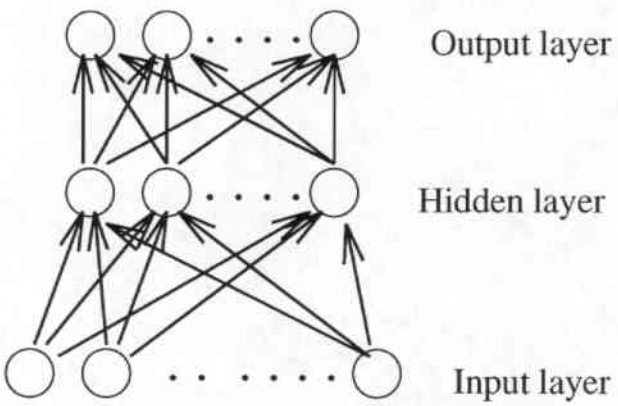

Abbildung 3.3: Netzarchitektur

Abbildung 3.3 zeigt die Architektur des Netzes. Es besteht aus 17 Eingabeunits, 8 Hiddenunits, und 6 Ausgabeunits. Die Ausgabeunits repräsentieren eine Klassifizierung des Signals. Jedem Ausgabeunit wird einer bestimmten Klasse zugeordnet. Die Klasseneinteilung ist in Tabelle 3.1 beschrieben.

Tabelle 3.1: Klasseneinteilung

Klasse	0	1	2	3	4	5
SNR S [db]	$S < 0$	$0 < S \leq 5$	$5 < S \leq 10$	$10 < S \leq 15$	$15 < S \leq 20$	$S > 20$

Als Ausgabefunktion der einzelnen Neuronen wurde die Sigmoidfunktion gewählt:

$$sigm(x) = \frac{1}{1+e^{-x}}$$

3.2.3 Training

Für das Training des Netzes wurden digitalisierte Sprachsegmente in 100ms Frames aufgeteilt, die jeweilige SNR bestimmt, und die Klassifizierung nach Tabelle 3.1 vorgenommen. Die Trainingsdaten wurden dann aus diesen Frames gebildet, wobei auf eine Gleichverteilung der Klassen innerhalb der Trainingsdaten geachtet wurde. Trainingsläufe, bei denen ein Sprachsegment komplett in die Trainingsdaten aufgenommen wurde, lieferten keine nutzbaren Ergebnisse, da die Klassen 0 und 5 dabei überproportional häufig vertreten sind. Während des Trainings wurden mit, in der Trainingsdaten nicht vertretenden, Testdaten in 4-Iterationen-Schritten der noch auftretende Fehler berechnet. Abbildung 3.4 zeigt den Verlauf des Fehlerwertes abhängig von den Iterationen.

Es wurden Trainigsläufe mit zwei verschiedenen Grundmengen für die Trainingsdaten durchgeführt:

1. Alle zur Verfügung stehenden Hintergrundgeräusche wurden mittrainiert.
2. Nur künstliches weißes Rauschen in verschiedenen Stärken war in den trainingsdaten vertreten.

Dabei zeigten sich Unterschiede in der Erkennungsfähigkeit bei Kamera bzw Radiohintergrundgeräuschen (siehe Abschnitt Resultate).

3.2.4 Resultate

Abbildung 3.5 zeigt den berechneten und geschätzten Verlauf der SNR für die Buchstabenfolge p-a-u-l. Dabei wurde künstliches weißes Hintergrundrauschen in zwei Stärken auf die Sprachdaten addiert.
Ähnliche Ergebnisse wurden bei allen getesteten Sprachsegmenten beobachtet. Die Klassifizierung des Netzes kann den SNR-Verlauf qualitativ sehr genau beschreiben, die erreichte Genauigkeit ist aber nicht sehr hoch.

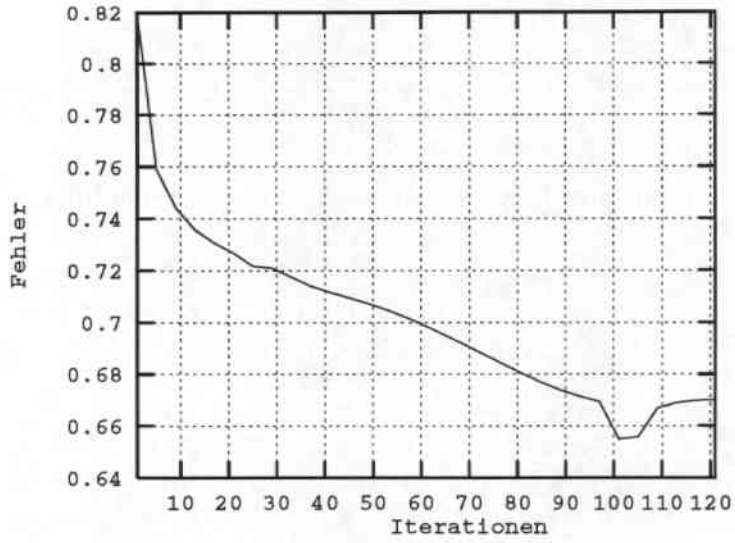

Abbildung 3.4: Trainingsverlauf

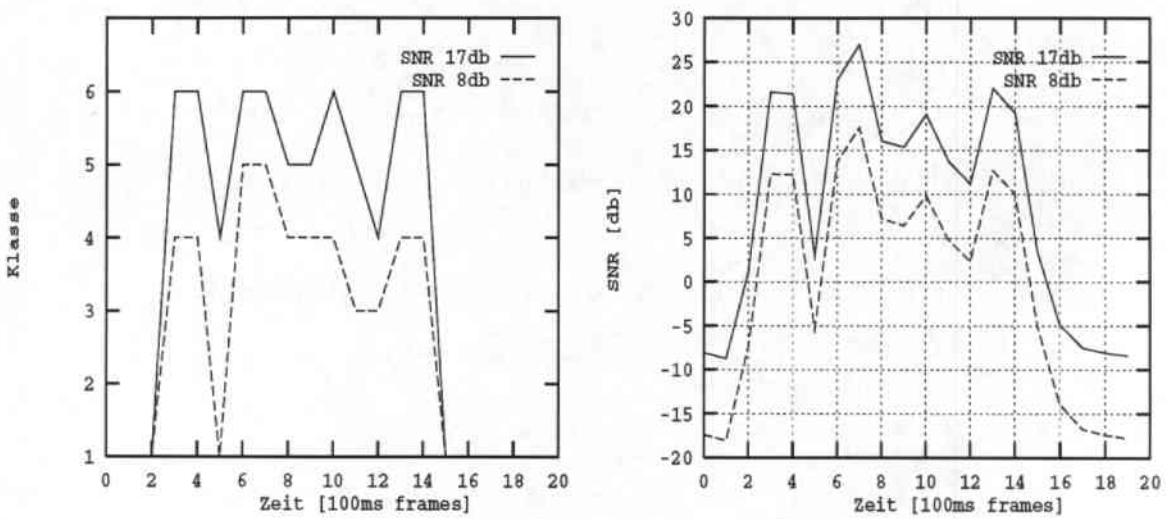

Abbildung 3.5: Geschätzter und berechneter SNR Verlauf bei 8db SNR und 17db SNR (p-a-u-l)

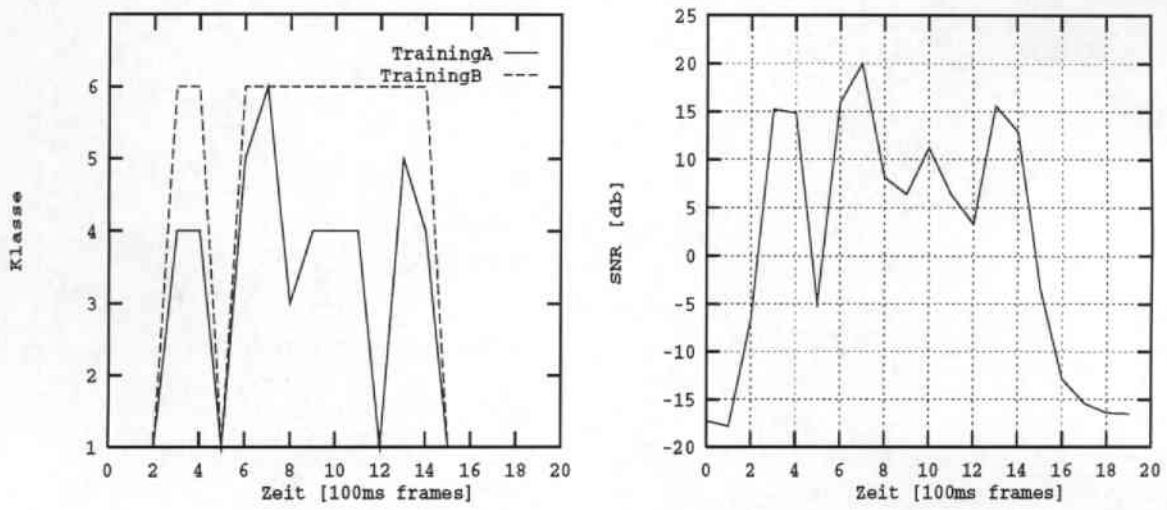

Abbildung 3.6: Camera p-a-u-l

Abbildung 3.6 zeigt die SNR-Schätzungen für die in Abschnitt 3.2.3 beschriebenen Trainingsmengen. Trainingsmenge B enthält keine Daten die mit Kamera-/Radiohintergrundgeräuschen versehen wurden. Es zeigen sich deutliche Unterschiede in der Genauigkeit der SNR-Bestimmung.

3.3 Bestimmung aufgrund statistischer Analyse der Energieverteilungen

[HIR]

3.3.1 Prinzipieller Ansatz

Die grundlegende Idee um die Stärke der Hintergrundgeräusche zu bestimmen basiert auf einer Analyse der Verteilung der Energiewerte von Unterbändern. Diese Unterbänder werden durch Einteilung des betrachteten Frequenzbereichs in gleich große Abschnitte gebildet.

In jedem Unterband wird die Verteilung der Energiewerte des Bandes, abhängig von der Zeit, bestimmt. Abbildung 3.7 zeigt den Energieverlauf und die korrespondierende Verteilung der Energiewerte eines Bands mit einer mittleren Frequenz von 1000 Hz und einer hohen SNR. Die Energiewerte wurden über den gesamten Satz (Dauer 3s) berechnet. Die meisten auftretenden Energiewerte sind Null oder annähernd Null. Das Maximum der Verteilungsfunktion liegt deshalb bei Null.

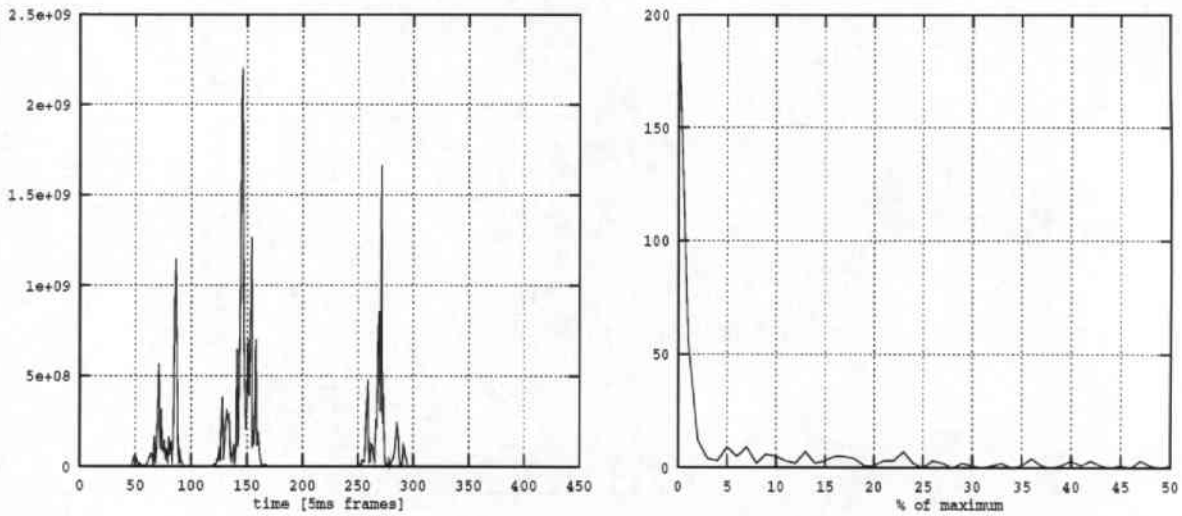

Abbildung 3.7: Spectral envelope 30db

Abbildung 3.8 zeigt dasselbe Band bei einer SNR von 15db. Der Satz wurde mit künstlichen weißen Rauschen versehen, das über das gesamte Frequenzspektrum gleichverteilt ist. Das Maximum der Verteilungsfunktion nimmt nun einen Wert an, der deutlich größer als Null

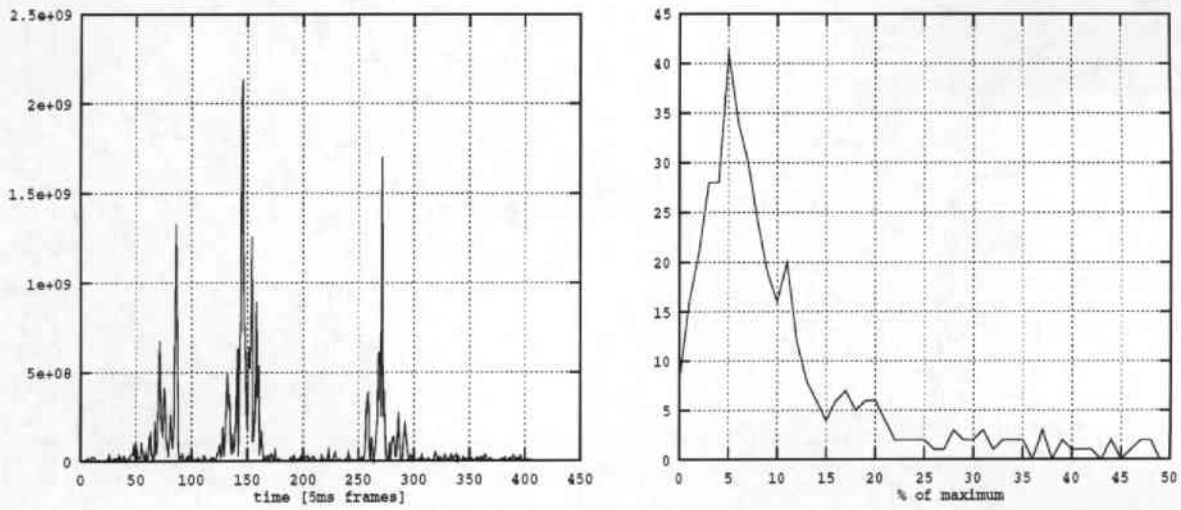

Abbildung 3.8: Spectral envelope 15db

ist. Der Maximumwert der Verteilungsfunktion steigt mit sinkender SNR an. In der Abbildung 3.9 läßt sich auch eine größere Streuung der Werte der Verteilungsfunktion um das Maximum beobachten. Beides, ansteigender Maximumwert und Streuung der Verteilungsfunktion, kann zur SNR-Bestimmung verwendet werden. In den meisten Fällen kann der Wert des Maximums direkt als Energiewert der Hintergrundgeräusche verwendet werden. Dabei muß aber die größere Streunung bei niedriger SNR berücksichtigt werden. Eine Glättung der Verteilungsfunktion, durch Reduktion der Genauigkeit, bei einer niedrigen SNR, verbessert die Maximumbestimmung. Entsprechend muß, bei Bändern mit einer hohen SNR, die Genauigkeit bei der Bildung der Verteilungsfunktion hoch sein um eine vernünftige Schätzung des Rauschen zu ermöglichen. Deshalb wird die Genauigkeit für die Berechnung der Verteilungsfunktion abhängig von der momentanen SNR innerhalb eines Bands abhängig gemacht. Wie dies im einzelnen realisiert wird, ist in Abschnitt 3.3.2 beschrieben.

Leider zeigen nicht alle Frequenzbereiche ein solch ideales Verhalten. Es lassen sich im wesentlichen zwei Problemfälle unterscheiden:

1. Bänder mit hoher Sprachenergie und niedriger Frequenz
   Abb 3.10 zeigt ein Band mit 100Hz. Die Verteilungskurve zeigt keinen stetigen Verlauf. Die hohen Werte werden durch die Energie des Sprachsignals und nicht durch Störgeräusche gebildet.

2. Bänder mit hohem Störgeräuschanteil und hoher Frequenz
   Abb 3.11 zeigt ein Band mit 5000 Hz und 7db. Die Verteilungskurve zeigt eben-

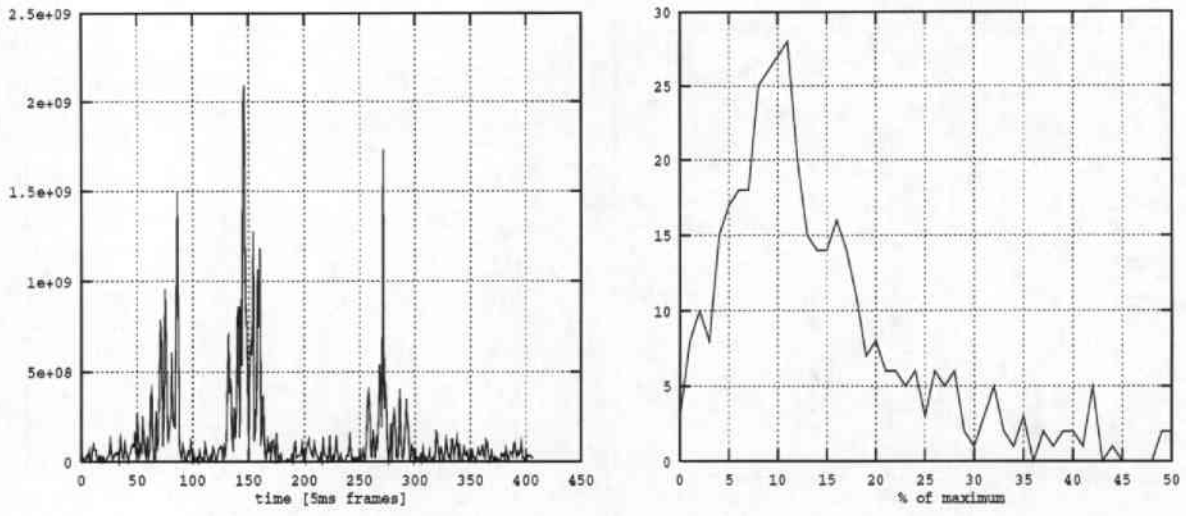

Abbildung 3.9: Spectral envelope 7db

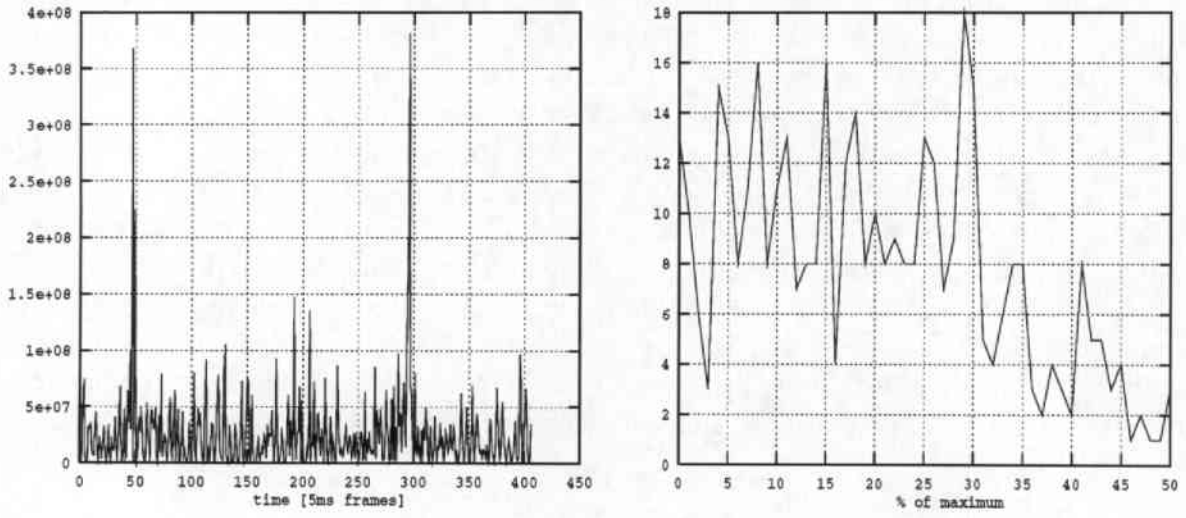

Abbildung 3.10: 100Hz 30db

falls keinen stetigen Verlauf. Vor der Bestimmung des Maximums muß die Funktion geglättet werden.

Wie diese beiden Fälle behandelt werden wird in Abschnitt 3.3.2 beschrieben.

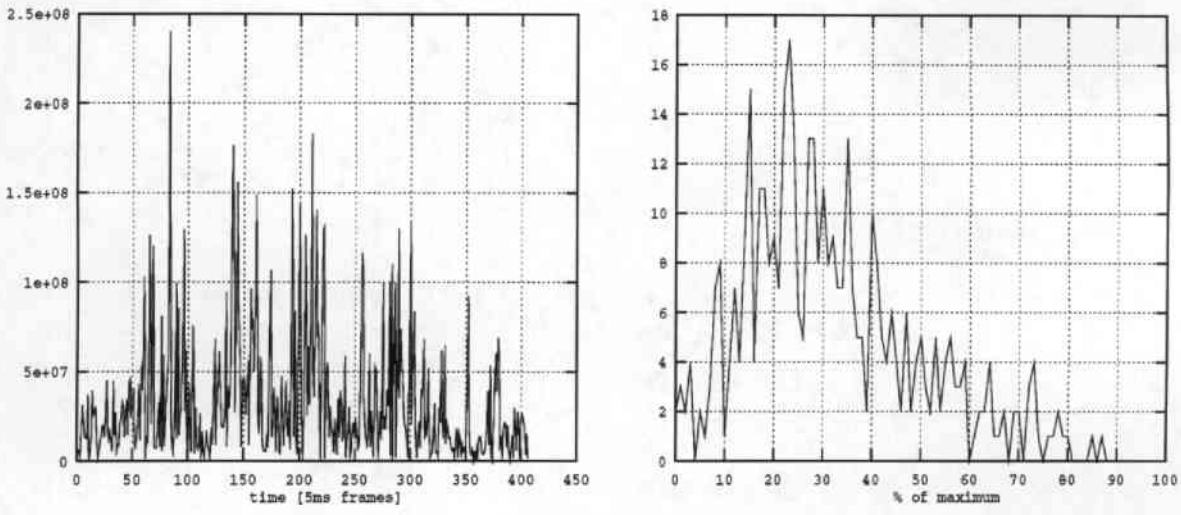

Abbildung 3.11: 7000Hz 9db

Um nun eine SNR-Schätzung für den gesamten Frequenzbereich zu erhalten, werden die Energiewerte in den einzelnen Bändern aufsummiert und durch die Anzahl der Subbänder geteilt.

$$N = \frac{1}{nSB} \sum_{i=0}^{nSB} N_{spec}(i\Delta f)$$

N=Gesamtenergie des Störsignals
$N_{spec}(f)$ = Energie des Störsignals in einem Band mit mittlerer Frequenz f
$\Delta f = \frac{f_{max}}{nSB}$
nB = Anzahl der untersuchten Bänder

Genauso kann die Energie des reinen Sprachsignals berechnet werden:

$$S = \frac{1}{nSB} \sum_{i=0}^{nSB} X_{spec} i\Delta F - N_{spec}(i\Delta f)$$

S=Gesamtenergie des Sprachsignals
$X_{spec}(f)$ = Energie des Gesamtsignals in einem Band mit mittlerer Frequenz f

3.3.2 Praktische Realisierung

Im praktischen Einsatz soll das Gesamtsystem auch mit nicht stationären Rauschen flexibel umgehen. Daher ist es nicht ausreichend, wenn das Verfahren nur einen SNR-Wert für den gesamten Eingabesatz liefert. Vielmehr sollte in bestimmten Zeitabständen ein aktueller SNR-Wert bestimmt werden. Die Länge dieses Zeitabstandes wird durch zwei Bedingungen bestimmt:

- Auftreten von nicht-stationären Hintergrundgeräuschen:
  Wird ein zu großer Zeitabstand gewählt, können kurzfristig auftretende Störgeräusche nicht erfaßt, bzw lokalisiert werden.

- Genauigkeit der Bestimmung
  Die Länge des Abstands bestimmt die Anzahl der Energiewerte, die für die Berechnung der Verteilungsfunktion zur Verfügung stehen. Bei zu wenigen Werten entsteht kein eindeutiges Maximum.

Es hat sich gezeigt, daß eine Fenstergröße von 1s für eine robuste Bestimmung ausreicht. Die Spektrumanalyse wurde alle 5ms durchgeführt, dadurch stehen für jedes Fenster 200 Werte für die Berechnung der Verteilungsfunktion zur Verfügung. Die Berechnung der aktuellen SNR wird alle 500ms mit überlappenden Fenstern durchgeführt.

Die Spektralanalyse wurde mit Hilfe schon bestehender Routinen durchgeführt. Für dir Fourieranalyse wurde eine Fenstergröße, die die Anzahl der Bänder bestimmt, von 256 gewählt. Durch die Samplefrequenz von 16000Hz ist die Bandbreite des Sprachsignals auf 8000Hz beschränkt. Dadurch ergibt sich die Bandbreite der einzelnen Unterbänder zu

$$\Delta f = \frac{8000Hz}{\frac{256}{2}}$$

Die Verteilungsfunktion wird für die Energierwerte für jedes Band berechnet. Dabei wird der Wertebereich von 0 bis zum Durchschnitt der Energiewerte in diesem Band gesetzt. Zuerst wird diese Verteilungsfunktion mit einer Genauigkeit von 0.25 Prozent des Durschnitts berechnet. Diese erste Verteilung wird dann grob geglättet, indem 8 Nachbarwerte aufsummiert werden. Dies ergibt dann eine zweite Verteilungsfunktion mit einer Genauigkeit von 2 Prozent. Mit dieser Verteilungsfunktion wird dann eine erste Bestimmung des Maximums M durchgeführt. Es werden nun 4 Fälle unterschieden:

1. M> 10 %:
   Der Wert des Maximums wird direkt übernommen.

2. $5\% < M \leq 10\%$
   Eine Genauigkeit von 1 Prozent wird verwendet. Die entsprechende Funktion wird durch glätten der ursprünglichen Verteilungsfunktion berechnet. Das Maximum dieser Funktion wird dann verwendet.

3. $2.5\% < M \leq 5\%$: Eine Genauigkeit von 0.5 Prozent wird verwendet. analog zu Fall 2.

4. M$\leq$ 2.5% : Eine Genauigkeit von 0.25 Prozent wird verwendet. Das Maximum wird nun in der ursprünglichen Verteilungsfunktion gesucht.

Mit diesem Vorgehen wird der in Abschnitt 3.3.1 beschreibenen Forderung, nach einer von der jeweiligen SNR abhängingen Genauigkeit Rechnung getragen.

Das oben beschriebene Verfahren kann nicht für die fünf Bänder mit den niedrigsten Frequenzen benutzt werden. Dies betrifft einen der in 3.3.1 beschreibenen Problemfälle. Die berechneten Maxima würden Energiewerte ergeben, bedingt durch die hohe Energie des reinen Sprachsignals, die die gesamte SNR Schätzung stark verfälschen würden. Deshalb wird für diese fünf Bänder der Durschschnittswert der Energie für die Hintergrundgeräusche der übrigen Bänder übernommen.

Mit dem Wert des Maximums M [in %] kann nun direkt die Energie des Unterbandes $E_{spec}(N)$ berechnet werden.

$$E(N) = M[\%] * \overline{E_{spec}(X)}$$

$\overline{E_{spec}(X)}$= Durchscnittsenergie des Signals im jeweiligen Unterband

Die Gesamtenergien können nun mit den in Abschnitt 3.3.1 Formalen berechnet werden. Damit läßt sich nun eine SNR Bestimmung für den betrachteten Zeitabschnitt berechnen. Es wurden zwei Funktionen mit verschiedenen Ausgaben implementiert:

- $\overline{S}NR$
  Es wird keine zeitabschnittsbezogene SNR berechnet sondern das Verhältnis der aktuellen Energie der Hintergrundgeräusche zur durchschnittlichen Energie der Sprachsignals im gesamten Sprachsegment bestimmt.
- Abschnittsbezogene SNR
  Für die Bestimmung der SNR werden für Sprache und Hintergrundgeräusche nur die aktuellemn Energiewerte betrachtet.

3.3.3 Testresultate

Die Abbildungen 3.12, 3.13, 3.15 und 3.14 zeigen Resultate mit verschiedenen Hintergrundgeräuschen, bzw künstlichen weißen Rauschen in verschiedenen Stärken. Tests mit etwa 50 Sätzen mit den verschiedenen Hintergrundgeräuschen zeigten ebenfalls keine größeren Abweichungen. Die Tests zeigten, daß das System auch mit nicht staionären Hintergrundgeräuschen zuverlässige Schätzungen der SNR liefert.

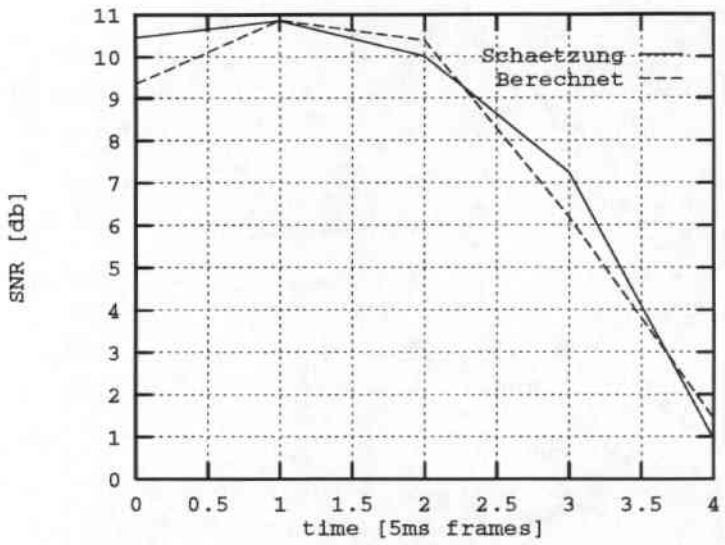

Abbildung 3.12: weises Rauschen

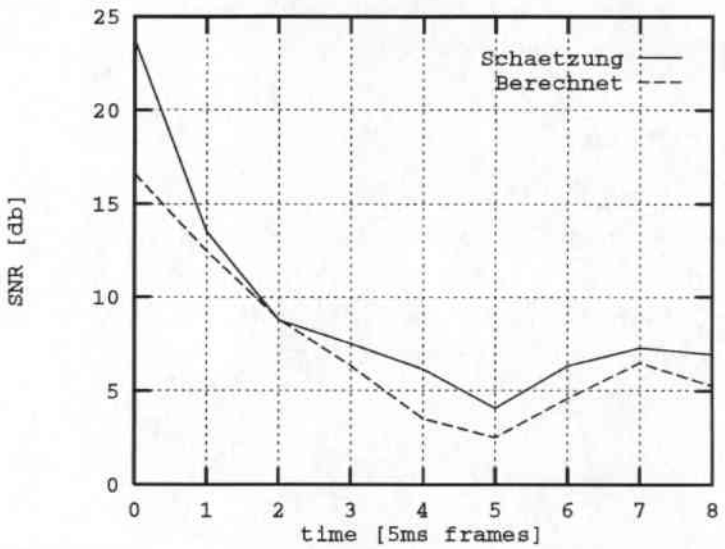

Abbildung 3.13: weises Rauschen;ansteigend/abfallend

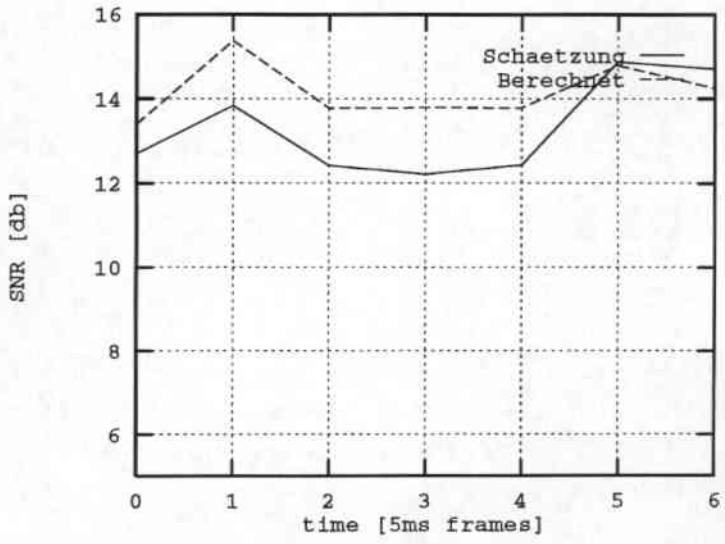

Abbildung 3.14: Kamera

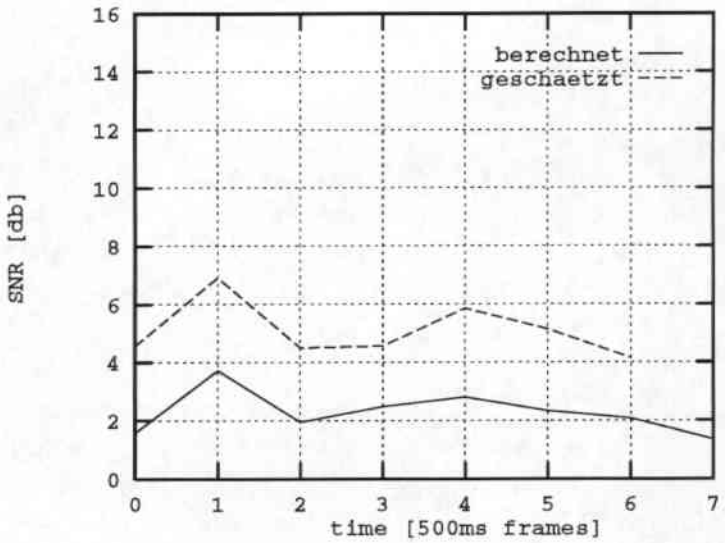

Abbildung 3.15: SNR-Verlauf geschätzt/berechnet bei Radiohintergrundgeräuschen

4. Vergleich der verschiedenen Verfahren

Ein direkter Vergleich der Verfahren gestaltet sich schwierig, da die Verfahren zum Teil unterschiedliche Ergebnisse liefern, oder nicht soweit implementiert wurden, daß vergleichbare Ergebnisse vorliegen. Nachfolgend werden die Art der Ergebnisse aufgeführt:

- Bestimmung der SNR in Sprechpausen:
  Bei der Bestimmung der SNR in Sprechpausen, kann eine sehr genaue Bestimmung erfolgen. Dabei muß aber die SNR bis zur nächsten Sprechpause als konstant angenommen werden. Mit dem betrachteten Algorithmus ist die Ermittlung der Sprechpausen bei sich stark ändernden Hintergrundgeräuschen nur beschränkt möglich.

- Neuronales Netz:
  Durch das neuronale Netz wird eine Klassifizierung des Signals erreicht. Ein aktueller Wert der SNR wird, mit einer Genauigkeit von 2.5 db, in 100ms Abständen bestimmt. Die Empfindlichkeit des Netzes gegenüber Hintergrundgeräuschen die nicht in den Trainingsdaten vorhanden sind, schränkt den Einsatz dieses Verfahrens ein. Auftretende Störgeräusche müßen bekannt sein, und in die Trainingsdaten aufgenommen werden zu können.

- Verfahren mittels Analyse der Energieverteilungen:
  Das Verfahren liefert alle 500ms eine relativ genaue Bestimmung der SNR. Dabei müssen keine Annahmen über die Art der Hintergrundgeräusche gemacht werden. Auch mit nicht stationären Hintergrundgeräuschen kann eine Bestimmung durchgeführt werden.

Die Entscheidung für den Einsatz im bestehenden Lippenlesersystem fiel zugunsten des Verfahrens mittels statistischer Analyse der Energieverteilungen. Die Beurteilung der verschiedenen Ansätze sollte unter dem Aspekt der Zielsetzung dieser Studienarbeit gesehen werden. Es sollte kein grundsätzlicher Vergleich der Ansätze durchzuführt werden. Aufgabe war es ein Verfahren zu finden, das hinreichend genau arbeitet und dieses zu implementieren. Nach der Entscheidung für einen Ansatz wurde natürlich versucht die Genauigkeit zu verbessern. Deshalb sind die Ergebnisse nicht unbedingt vergleichbar, da auch bei den anderen Verfahren durchaus noch Verbesserungen denkbar sind. Dies war aber, durch den beschränkten Zeitrahmen einer Studienarbeit, nicht möglich.

5. Zusammenfassung, Ausblick

In dieser Arbeit wurde versucht ein Verfahren zur verläßlichen Einschätzung der SNR eines Sprachsegments zu finden. Es wurden drei unterschiedliche Ansätze betrachtet, und soweit möglich miteinander verglichen. Mit der, mittels statistischer Analyse der Energiewerte in Unterbändern gewonnene, Bestimmung konnten die besten Ergebnisse erzielt werden. Bei diesem Ansatz werden gute Ergebnisse für alle getesteten Hintergrundgeräusche erreicht. Auch mit nicht-stationären Hintergrundgeräuschen konnte ein Bestimmung durchgeführt werden, die einen genauen Wert der aktuellen SNR zum Ergebniss hat.

Die Implementierung dieses Verfahrens wird zur Zeit von Wolfgang Huerst in seiner Studienarbeit,im Rahmen deselben Projekts, verwendet. In dieser Arbeit werden die verschiedenen Möglichkeiten der Kombination visueller und akustischen Informationen untersucht.

Abbildungsverzeichnis

Literaturverzeichnis

[DUC94] P. Duchnowski, U. Meier, A. Waibel: *See Me, Hear me: Integrating Automatic Speech Recognition and Lip-Reading.* International Conference on Spoken Language Processing, ICSLP, 1994.

[DUC95] P. Duchnowski, M .Hunke, D. Büsching, U. Meier, A. Waibel: *Toward Movement-Invariant Automatic lip-reading and speech recognition*, Proc. ICASSP , 1995.

[HER93] Hermann Hild, Alex Waibel: *Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network,* EUROSPEECH 93, Berlin, Germany, September 1993, Volume 2, pp. 1481-1484.

[HER93-2] Hermann Hild, Alex Waibel: *Connected Letter Recognition with a Multi-State Time Delay Neural Network* NIPS 5, San Marino, CA:Morgan Kaufmann Publishers, 1993.

[HIR] H. Günther Hirsch: *Estimation of Noise Spektrum and its Application to SNR-Estimation and Speech Enhancement.* Technical Report. International Computer Science Institute, Berkeley, California, USA.

[LAM] L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon: *An Improved Endpoint Dedector for Isolated Word Recognition.* IEEE ASSP, 29, 1981.

[MAK] Brian Mak, Jean-claude Junqua, Ben Reaves: *A Robust Speech/Non-Speech Dedection Algorithm Using Time and Frequency-Based Features.* Speech Technology Laboratory, Division of Panasonic Technologies, Inc, 3888 State Street, Santa Babara, California.

[Mei] Uwe Meier:*Robuste Systemarchitekturen füer maschinelles Lippenlesen,* Diplomarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[TAM] Shin'ichi Tamura, Alex Waibel: *Noise Reduction using Connectionist Models.* ATR Interpreting Telephony Research Laboratories, 2-1-61 Shiromi, Higashi-ku, Osaka, 540 Japan.

[RAB] L. R. Rabiner, M. R. Samur: *An Algorithm for Determining the Endpoints of isolated Utterances*, The Bell System Technical Journal, Vol. 54, No. 2, February 1975.

[REA] Ben Reaves: *Comments on "An improved Endpoint Dedector for isolated Word Recognition"*. IEEE Transactions on Signal Processing, Vol. 39,No. 2, February 1991.

[WAI] Alex Waibel:
Hier fehlt noch der Eintrag eines Papers daß die Melscalekoeffizienten beschreibt.

[HUE] Wolfgang Huerst:
Studienarbeit. Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.