Universität Karlsruhe
Institut für Logik, Komplexität
und Deduktionssysteme
Am Fasanengarten 5
D–76128 Karlsruhe

Adaptive bimodale Sensorfusion für automatische Spracherkennung und Lippenlesen

Studienarbeit von

Wolfgang Hürst

Betreuer:

Dr. Paul Duchnowski

Mai 1995

Vorwort

In dieser Arbeit soll der Versuch unternommen werden, verschiedene Verfahren und Ebenen der bimodalen Sensorfusion am Beispiel der akustischen Spracherkennung, kombiniert mit einer visuellen Verarbeitung des Sprachsignals, zu testen und gegeneinander abzuwägen. Grundlage bildet ein System zur Erkennung von kontinuierlich gesprochenen Buchstaben-Sequenzen des deutschen Alphabets. Dieses System erhält als Eingabe nicht nur das rein akustische Sprachsignal, sondern auch visuelle Information, das heißt Information über die entsprechenden Lippenbewegungen des Sprechers. Hiervon verspricht man sich eine Verbesserung in der Erkennungsleistung, insbesondere unter schlechten Aufnahmebedingungen für das akustische Signal.
Der verwendete Erkenner basiert auf einem sogenannten Multi-State Time Delay Neural Network und wird in Kapitel 1 beschrieben. Ferner werden dort verschiedene Ebenen des Neuronalen Netzwerks, die sich zur Integration der beiden Eingabemodalitäten anbieten, vorgestellt. In den folgenden Kapiteln werden diese verschiedenen Ebenen genauer untersucht. Kapitel 2 betrachtet verschiedene Verfahren bei der Kombination auf Phonemebene (diese entspricht der Ausgabeschicht des Time Delay Neural Networks). Kapitel 3 beinhaltet einige Ansätze zur Kombination auf der Hidden-Ebene der Netzes, während in Kapitel 4 die Verschmelzung der Signale auf der Eingabe-Ebene angesprochen wird. In Kapitel 5 folgt schließlich eine Zusammenfassung und abschließende Bewertung der betrachteten Verfahren.

An dieser Stelle möchte ich mich noch bei allen bedanken, die mir bei der Erstellung dieser Arbeit mit Rat und Tat zur Seite standen. Besonders erwähnt seien hier Uwe Meier, sowie mein Betreuer Paul Duchnowski, der sich bereit erklärt hatte, diese Arbeit auch nach seiner Rückkehr an das Massachusetts Institute of Technology weiter zu betreuen.

Wolfgang Hürst 31. Mai 1995

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Grundlagen

In den letzten Jahren wurden im Bereich der maschinellen Spracherkennung große Fortschritte erzielt. Die guten Erkennungsraten aktueller Systeme sinken jedoch mitunter beträchtlich, wenn man sie nicht unter optimalen „Laborbedingungen", das heißt in einer relativ rauschfreien Umgebung, einsetzt. Rauschen, Zwischenrufe und ähnliche Nebengeräusche, wie sie z.B. an einem Arbeitsplatz häufig auftreten, verhindern eine gute Erkennung.
Motiviert dadurch und durch die Tatsache, daß auch der Mensch beim „Erkennen" von Sprache neben der gesprochenen zusätzlich visuelle Information benutzt [10], hat man vor einigen Jahren begonnen, Systeme zu entwickeln, die neben dem reinen Sprachsignal auch ein visuelles Signal (in der Regel Lippenbewegungen) als Eingabe bekommen. Die erhoffte Verbesserung in der Erkennungsrate, insbesondere unter schlechten akustischen Aufnahmebedingungen, wurde in einer Vielzahl von Experimenten bestätigt. Ein guter Überblick über bisherige Ansätze ist in [12] zu finden.
Bei der Nutzung von visuellen Daten als zusätzliche Informationsquelle und zur Unterstützung der akustischen Signale stellt sich die Frage einer geeigneten Verschmelzung der beiden Eingabemodalitäten. Dieser Frage soll im folgenden nachgegangen werden. Als Grundlage dient hierfür der Anfang der neunziger Jahre am Institut für Logik, Komplexität und Deduktionssysteme an der Universität Karlsruhe entwickelte Lippenleser. Hierbei handelt es sich um ein System, das kontinuierlich gesprochene deutsche Buchstaben-Sequenzen erkennen soll. Es basiert auf einem rein akustischen Erkenner, der in Kapitel 1.1 beschrieben wird. Die Erweiterung um einen zusätzlichen visuellen Teil wird in Kapitel 1.2 erläutert. Anschließend werden in Kapitel 1.3 alternative Architekturen bei der kombinierten Erkennung angesprochen.

1.1 Der akustische Buchstabenerkenner

Der an der Universität Karlsruhe entwickelte Buchstabenerkenner ist ein System, das beliebige, kontinuierlich gesprochene Sequenzen deutscher Buchstaben erkennt. Bei diesem Spracherkennungsproblem handelt es sich um einen kleinen, nur 26 Wörter umfassenden, Task[1], der allerdings hochgradig verwechselbar ist. Der Erkenner arbeitet sprecherunabhängig

[1] Mit Wort ist in diesem Fall ein einzelner Buchstabe gemeint, da es sich bei dem Erkennungsproblem um buchstabierte Sequenzen handelt. Ein gesprochener Buchstabe setzt sich somit wie ein Wort aus mehreren Phonemen (= kleinste akustisch unterscheidbare Einheit) zusammen, z.B. ist der Buchstabe „Ypsilon" ein relativ großes Wort. Buchstabierte Sequenzen sind somit Wortfolgen und werden dementsprechend als Sätze bezeichnet.

und liefert Erkennungsraten von über 90%. Grundlage des Erkenners ist ein sogenanntes *Multi-State Time Delay Neural Network* (MS-TDNN). Hier soll kurz auf die Funktionsweise eingegangen werden, soweit sie für die folgenden Kapitel von Bedeutung ist. Für eine ausführlichere Beschreibung des Systems sei auf [6, 7] verwiesen.
Abbildung 1.1 zeigt den schematischen Aufbau eines MS-TDNN. Es besteht aus einem *Time Delay Neural Network* (TDNN) mit zwei zusätzlichen Schichten. Bei einem TDNN handelt es sich um ein Neuronales Netzwerk, bei dem jedes Neuron zu einem bestimmten Zeitpunkt t nicht nur das aktuell anliegende Signal als Eingabe bekommt, sondern auch noch die Signale, die unmittelbar zuvor und unmittelbar danach eingegeben wurden, also die Werte zu den Zeitpunkten $t-1, t-2, ..., t-d$ bzw. $t+1, t+2, ..., t+d$. Den Wert d bezeichnet man als *Time-Delay*. Durch diesen Time-Delay wird es dem Netz ermöglicht, Zusammenhänge zwischen Eingabesignalen über die Zeit hinweg, wie sie insbesondere in der Sprachverarbeitung von Bedeutung sind, zu erlernen und zu erkennen. Eine ausführliche Beschreibung des TDNNs und seines Einsatzes bei der Spracherkennung ist in [20] zu finden.

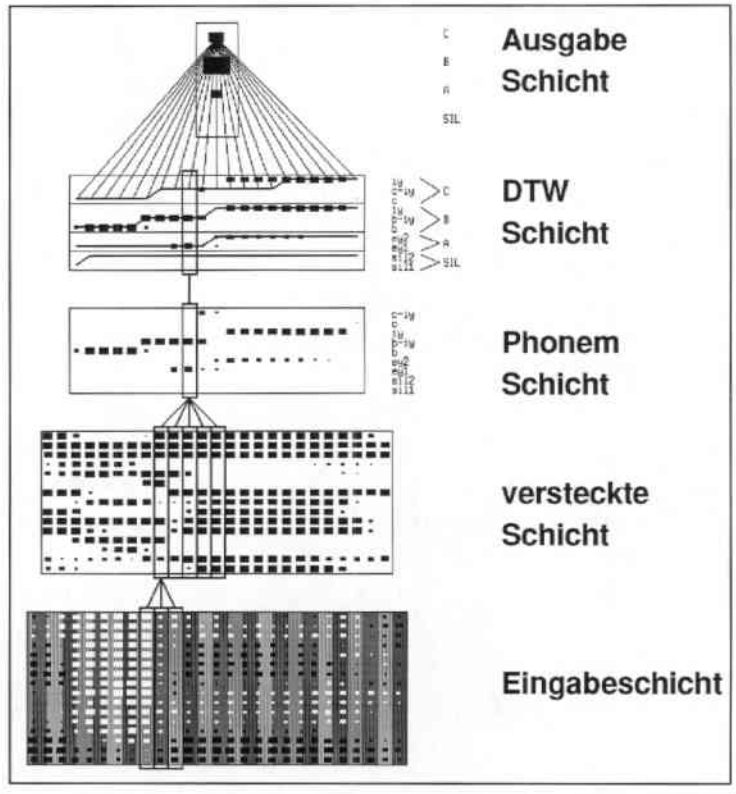


Abbildung 1.1: Exemplarische Darstellung eines MS-TDNN (aus [7]) für die Buchstaben A, B und C, sowie SIL („silence", d.h. es wurde nicht gesprochen). In diesem Beispiel ist die Aktivierung des Buchstaben B in der Ausgabeschicht (dargestellt durch ein schwarzes Quadrat, dessen Größe relativ zum Wert der Aktivierung ist) am größten. Demzufolge würde man hier schlußfolgern, daß ein B gesprochen wurde.

Mit dem TDNN wird eine Erkennung auf phonetischer Ebene betrieben. Die Ausgabeschicht des TDNNs liefert für jede Zeiteinheit und jedes Phonem einen Wert, der ungefähr die Wahrscheinlichkeit dafür angibt, daß dieses Phonem zu diesem Zeitpunkt gesprochen wurde. An sie schließt sich die sogenannte *DTW-Schicht* an. In dieser Schicht wird durch einen auf dem Prinzip des dynamischen Programmierens beruhenden sogenannten *One Stage Dynamic Time Warping Algorithmus* [13] der optimale Pfad durch die Phonem-Zustände gesucht, das

heißt, der Algorithmus versucht, die wahrscheinlichste Satzhypothese zu finden. Diese wird in die Ausgabeschicht geschrieben. Hier wird eine Entscheidung gefällt, welches Wort gesprochen wurde, indem das Wort mit der höchsten Aktivierung ausgewählt wird.
Als Eingabe für den Erkenner werden aus dem akustischen Signal mit der Technik der Melscale Filterbank Kodierung alle 10 ms sechzehn sogenannten Melscale Fourier Koeffizienten berechnet, die den Eingabevektor für den jeweiligen Zeitpunkt bilden.

1.2 Bimodale Erkennung

Mit dem Ziel, die bei äußeren Störeinflüssen sinkende Erkennungsrate zu verbessern, wurde der existierende Buchstabenerkenner um Komponenten erweitert, die visuelle Information über das Sprachsignal verarbeiten und damit die akustische Erkennung unterstützen sollen. Das System arbeitet wie der akustische Erkenner auf Buchstabensequenzen des deutschen Alphabets und ist momentan noch sprecherabhängig. Eine Umstellung auf sprecherunabhängige Erkennung wird jedoch zur Zeit durchgeführt.
Für eine ausführliche Übersicht über den Erkenner sei auf [2, 3, 4] oder [12] verwiesen. Hier soll nur ein kurzer Überblick gegeben werden, der insbesondere auf die Komponenten eingeht, die für das weitere Verständnis von Bedeutung sind.
Die akustischen Eingabedaten bestehen, wie bei der rein akustischen Erkennung, aus einem 16-dimensionalen Vektor der berechneten Melscale Koeffizienten. Auf der visuellen Seite werden zunächst Videobilder mit einer Auflösung von 144x80 Pixeln tiefpaßgefiltert auf eine Auflösung von 24x16 Pixel und anschließend vom Intervall $[0, 255]$ auf das Intervall $[-1, 1]$ normalisiert. Trotz dieser Merkmalreduktion von 11520 auf 384 Parameter kann es vorteilhaft oder erforderlich sein, eine weitere Verringerung der Parameteranzahl durchzuführen (vgl. Anfang von Kapitel 4). Aus diesem Grund wurden hierzu in [5, 11] verschiedene Verfahren untersucht. Beste Ergebnisse wurden mit den normalen, 384-dimensionalen Graustufenbildern und mit durch das Verfahren der *Linear Discriminant Analysis* (kurz LDA) auf 32 Dimensionen reduzierten Daten erreicht. Aus diesem Grund wurden die Ansätze in dieser Arbeit mit Graustufen- und LDA-Bildern als visuelle Eingabe durchgeführt. Da die Framerate der akustischen Daten ungefähr 3,5 mal höher als die der visuellen Daten ist, d.h. da pro visuellem drei bis vier akustische Zeitframes existieren, wird jedem akustischen Frame der zeitlich am nächsten liegende visuelle Frame zugeordnet[2].
Für die Integration der visuellen Erweiterungen boten sich mehrere Alternativen an (siehe [5]). Man entschloß sich dafür, für die visuelle Eingabe ein weiteres TDNN zu nehmen und die Ausgabe dieses Netzes mit der Ausgabe des akustischen Netzwerks in einer zusätzlichen Schicht geeignet zu verknüpfen. Das visuelle TDNN liefert als Ausgabe die Wahrscheinlichkeit, daß ein bestimmtes Visem gesprochen wurde. Viseme sind, analog zu den Phonemen auf der akustischen Seite, definiert als kleinste Spracheinheiten, die visuell unterschieden werden können. An die beiden Netze schließt sich eine Kombinationsschicht an, die für jedes Phonem zu jedem Zeitpunkt ein Ausgabeneuron enthält. Die Kombinationschicht entspricht also quasi der Ausgabeschicht des akustischen Netzes mit dem Unterschied, daß die Ausgabe dieser Schicht nicht nur von der akustischen, sondern auch von der visuellen Eingabe abhängt. Eine Übersicht über das Gesamtsystem befindet sich in Abbildung 1.2.

[2] mit Frame sind hier die Merkmale der jeweiligen Schicht des neuronalen Netzes zu einem bestimmten Zeitpunkt t gemeint

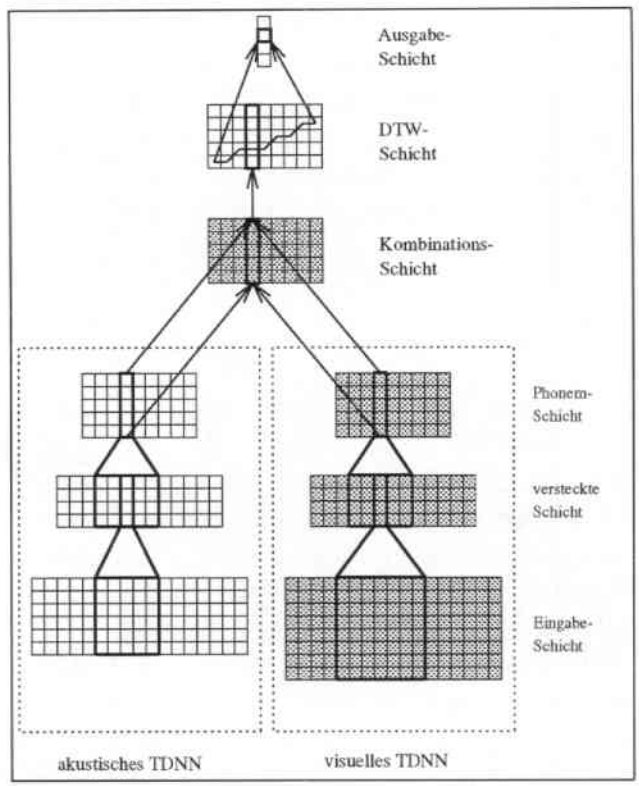


Abbildung 1.2: Netzarchitektur für die Kombination auf phonetischer Ebene. Die Erweiterungen gegenüber der Architektur des rein akustischen Erkenners (also das visuelle TDNN und die zusätzliche Kombinationsschicht) sind grau hervorgehoben.

Die Zuordnung von Visemen zu Phonemen ist nicht eindeutig, denn es existieren mehr Phoneme als Viseme[3]. Zum Beispiel können die Phoneme /p/, /b/ und /m/ visuell nicht unterschieden werden. Bei der Kombination der von den beiden TDNN gelieferten Phonem- und Visemwahrscheinlichkeiten wird deshalb wie folgt vorgegangen: für jedes Phonem wird anhand einer Phonem-Visem-Tabelle der zugehörige Visem-Wert herausgesucht und die geeignet gewichtete Summe aus Phonem- und Visem-Aktivierung in die zusätzliche Kombinationsschicht eingetragen.
Für die Gewichtung wurden zwei Möglichkeiten realisiert: die Gewichte werden gemäß der Verrauschung des akustischen Signals von Hand eingestellt (also bei starker Verrauschung eine höhere Gewichtung der visuellen Ausgabe), oder es wird eine sogenannte *Entropy* (siehe [14]), die ein Maß für die Sicherheit der Ausgabe eines Neuronalen Netzes darstellt, verwendet.
Die Entropy E zu einem bestimmten Zeitpunkt t wird wie folgt berechnet:

$$E = -\sum_i \frac{hyp_i}{\sum_j hyp_j} \log \frac{hyp_i}{\sum_j hyp_j}, \tag{1.1}$$

wobei hyp_i die Aktivierung des i. Ausgabeneurons des Neuronalen Netzes ist. Der Entropywert E ist am kleinsten, falls genau ein Ausgabeneuron die Aktivierung „Eins“ hat und alle anderen Werte Null sind. Er ist am größten, wenn alle Neuronen die gleiche Aktivierung

[3] Beim deutschen Alphabet und der hier zugrundeliegenden Visemdefinition hat man 63 Phoneme und 42 Viseme.

besitzen. Je größer also die berechnete Entropy ist, je unsicherer ist die Ausgabe des entsprechenden Netzes. Dies wird bei der Kombination nun wie folgt genutzt:
Für jeden Zeitpunkt t wird für die entsprechenden Ausgabeneuronen des akustischen und des visuellen Netzes die jeweilige Entropy E_A bzw. E_V berechnet. Ist die Entropy E_A des akustischen Netzes hoch, nimmt man an, daß die Ausgabe relativ unzuverlässig ist. Ist E_A jedoch klein, geht man von einer sicheren Ausgabe aus. Das entsprechende gilt für das visuelle Netz. Anhand dieser beiden Entropywerte bestimmt man nun den kombinierten Ausgabewert durch gewichtete Addition, wobei die Gewichte w_a für die akustische Aktivierung bzw. w_v für die visuelle Aktivierung nach folgendem Schema berechnet werden:

$$w_a = b + \frac{E_A - E_V}{2K}, \qquad (1.2)$$
$$w_v = 1 - w_a.$$

K steht hier für die maximale Entropy im aktuellen Satz und dient als Normierungsfaktor. b stellt einen Schwellwert dar, der von Hand einzustellen ist und bewirkt, daß eine bestimmte Modalität bevorzugt wird. Z.B. ist es bei starker Verrauschung der akustischen Daten naheliegend, die visuelle Ausgabe zu „bevorzugen". Der Wert hyp_i^B eines Neurons i der Kombinationsschicht zu einem bestimmten Zeitpunkt t berechnet sich schließlich durch

$$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V, \qquad (1.3)$$

mit hyp_i^A = Aktivierung des entsprechenden Neurons i in der Phonemschicht des akustischen TDNN und hyp_i^V = Aktivierung des entsprechenden, durch die Phonem-Visem-Tabelle bestimmten Neurons in der Visemschicht des visuellen TDNN.
Die weitere Erkennung verläuft wie bei dem in Kapitel 1.1 beschriebenen Buchstabenerkenner.

1.3 Alternative Architekturen

Die Verarbeitung der akustischen und visuellen Eingabe in getrennten Netzen mit anschließender Kombination bietet gewisse Vorteile, die in Kapitel 2 beschrieben werden. Für die Integration der beiden Eingabesignale bieten sich jedoch auch weitere Alternativen an, z.B. eine Kombination auf einer tieferen Ebene des Neuronalen Netzwerkes. In Abbildung 1.3 ist die Architektur für eine Verschmelzung der Signale auf der versteckten Schicht, in Abbildung 1.4 für eine Verschmelzung auf der Eingabeschicht dargestellt. Bei letzterer Architektur existieren eigentlich keine getrennten visuellen Komponenten mehr. Es handelt sich vielmehr um ein einfaches TDNN, das neben den akustischen auch die visuellen Signale als Eingabe erhält. Eine Differenzierung der beiden Modalitäten ist jedoch auch hier möglich, beispielsweise durch unterschiedliche Time-Delays (z.B. drei Frames für die Akustik und fünf für die visuellen Merkmale). In dieser Arbeit wurden jedoch stets die gleichen Time-Delays für beide Signalarten verwendet. Auf die Vor- und Nachteile der beiden Architekturen wird in den entsprechenden Kapiteln noch genauer eingegangen.
Eine weitere Kombinationsmöglichkeit wäre, zunächst akustische und visuelle Erkennung getrennt durchzuführen und anschließend die Ausgabe der beiden MS-TDNN zu verknüpfen, also eine Kombination zwischen der DTW-Schicht und der Ausgabeschicht durchzuführen. Auf diese Möglichkeit wird jedoch im weiteren nicht eingegangen.

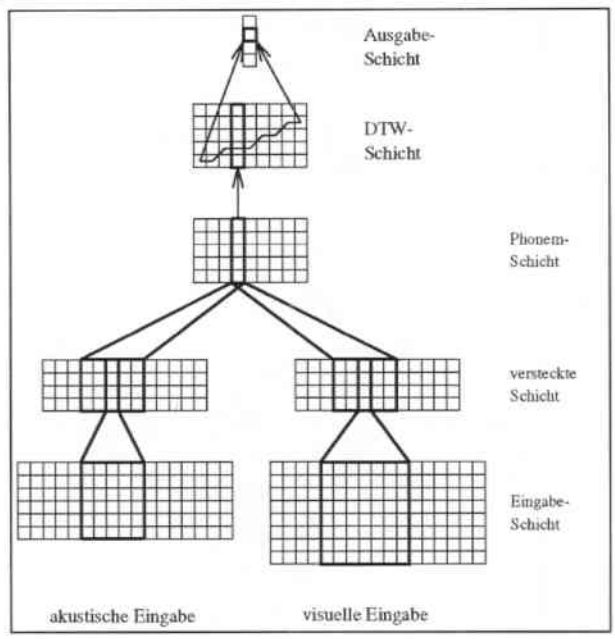


Abbildung 1.3: Netzarchitektur für die Kombination auf der Ebene der versteckten Einheiten.

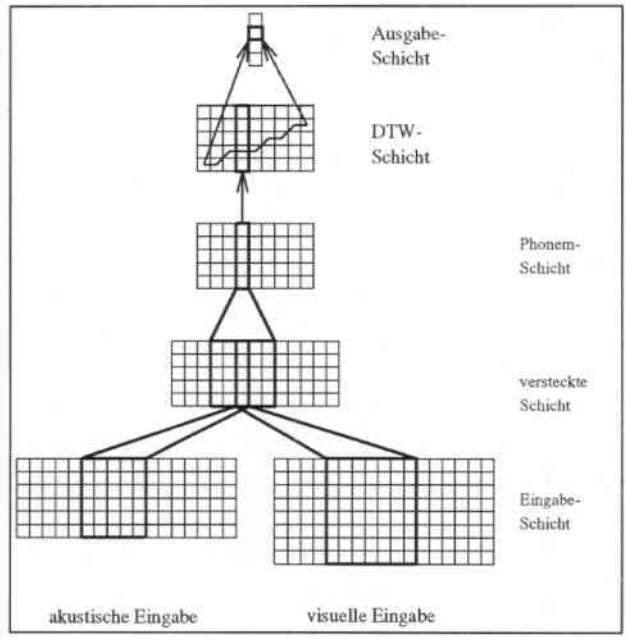


Abbildung 1.4: Netzarchitektur für die Kombination auf der Ebene der Eingabeeinheiten.

Kapitel 2

Kombination auf Phonemebene

Die Kombination der akustischen und visuellen Information durch gewichtete Addition auf Phonemebene bietet gewisse Vorteile gegenüber anderen Verfahren. Zunächst erlaubt diese Realisierung auf einfache Weise eine rein akustische bzw. rein visuelle Erkennung. Durch Setzen der Gewichte w_v bzw. w_a auf Null ist es möglich, den visuellen bzw. akustischen Teil der Gesamtarchitektur bei Bedarf „auszuschalten". Ferner ist man in der Lage, die beiden Eingabemodalitäten auf unterschiedliche, auf das jeweilige Signal angepaßte Art und Weise zu behandeln. In [12] wurde beispielsweise der Versuch unternommen, statt zweier nahezu identischer TDNN für die beiden unterschiedlichen Eingaben, auf der visuellen Seite ein dem veränderten Problem angepaßtes Netzwerk zu verwenden. Durch den Einsatz eines sogenannten MS-TDNN^{3d} wurde versucht, die Schwierigkeiten zu umgehen, die sich dadurch ergeben, daß es sich bei der visuellen Eingabe um zweidimensionale Bilder und nicht um ein eindimensionales Signal handelt. Desweiteren ist durch die Trennung des visuellen Netzes vom akustischen Netzwerk ein separates Training der beiden TDNNs möglich, was insbesondere dann einen Vorteil darstellt, wenn von einer Eingabemodalität mehr Daten vorliegen als von der anderen. Die Tatsache, durch Setzen der entsprechenden Gewichte „von außen" in die Kombination eingreifen zu können, bietet ferner die Möglichkeit, auf unvorhergesehene oder selten vorkommende Ereignisse (insbesondere solche, die in der Trainingsmenge selten oder nie vorkommen) gesondert zu reagieren. Auf die verschiedenen Nachteile, die sich bei dieser Architektur ergeben, wird in Kapitel 3 näher eingegangen.

Im folgenden werden zunächst verschiedene Kombinationsalternativen behandelt und getestet. Daran anschließend wird in Kapitel 2.2 versucht, die Verfahren, bei denen eine Einstellung bestimmter Parameter von Hand erforderlich ist, zu automatisieren. In Kapitel 2.3 wird schließlich ein weiterer Ansatz besprochen, der sich von den anderen dadurch unterscheidet, daß die Gewichte für die Kombination sowohl über die Zeit, als auch über die verschiedenen Merkmale, d.h. Phoneme, variieren.

2.1 Verschiedene Kombinationsalternativen

2.1.1 Bisher verwendete Verfahren

Bei den bisher verwendeten Verfahren wird die Kombination hyp_i^B der Ausgabe des akustischen mit der Ausgabe des visuellen TDNNs durch gewichtete Addition nach folgender Formel

durchgeführt:

$$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V, \tag{2.1}$$

wobei hyp_i^A die Aktivierung eines Neurons i in der Phonemschicht des akustischen TDNNs und hyp_i^V die Aktivierung des entsprechenden, durch die Phonem-Visem-Tabelle bestimmten Neurons in der Visemschicht des visuellen TDNNs ist. Die Gewichte w_a und w_v werden, wie in Kapitel 1.2 beschrieben, entweder von Hand gesetzt oder in Abhängigkeit des Entropywertes eines Zeitframes bestimmt (siehe Gleichungen 1.1 und 1.2 auf den Seiten 12 und 13).
Aus einer Menge von 200 Buchstabensequenzen eines männlichen Sprechers (mum1/mum2) wurden 170 Sequenzen für das Training des Neuronalen Netzes ausgewählt. Zum Testen der verschiedenen Verfahren wurden die restlichen 30 Wortfolgen verwendet. Diese Testsamples wurden auf verschiedene Art und Weise künstlich verrauscht, um äußere Störeinflüsse zu simulieren. Als Maßstab für den Grad der Verrauschung der akustischen Daten dient die sogenannte *Signal-to-Noise-Ratio*, kurz SNR, die sich wie folgt berechnen läßt:

$$SNR = 10 \log_{10} \frac{S^2}{N^2}, \tag{2.2}$$

wobei S die Signalwerte des reinen Sprachsignals und N die der Stör- und Hintergrundgeräusche angibt. Die Einheit der SNR ist Dezibel [dB].
Im einzelnen ist die Testdatenbasis folgendermaßen aufgebaut:

- CLEAN: Menge der 30 Testsamples, die unter optimalen Bedingungen aufgenommen wurden (optimale, also rauschfreie Aufnahmebedingungen entsprechen einer SNR von ungefähr 30 dB)

- NOISE 1 und NOISE 2: Die 30 Testsamples wurden mit einer konstanten SNR von ungefähr 16 bzw. 8 dB verrauscht

- INCREASE 1 und INCREASE 2: Die 30 Testsamples wurden durch eine von cirka 30 auf 16 dB bzw. von cirka 30 auf 8 dB abfallende SNR verrauscht

- RADIO 1 und RADIO 2: Die 30 Testsamples wurden durch die Geräusche eines laufenden Radios (Musik) mit einer um ungefähr 20 bzw. 17 dB schwankenden SNR verrauscht

- MOTOR 1 und MOTOR 2: Die 30 Testsamples wurden durch die Geräusche des Motors einer aktiven Kamerasteuerung mit einer um ungefähr 25 bzw. 11 dB schwankenden SNR verrauscht

Das Rauschsignal wurde jeweils auf die unverrauschten Daten aufaddiert. Bei NOISE 1, NOISE 2 und INCREASE 1, INCREASE 2 handelt es sich um ein künstlich erzeugtes sogenanntes weißes Rauschen, während das Rauschen bei RADIO und MOTOR mit einem Mikrofon aufgenommen wurde.
Die Ergebnisse der Verfahren auf der Testdatenbasis mit Graustufenbildern sind in Tabelle 2.1 dargestellt. Als Maß für die Leistung des Erkenners wird die sogenannte *Word Accuracy* angegeben, die wie folgt definiert ist:

$$WA = 100\%(1 - \frac{\#\text{SubstitutionError} + \#\text{InsertionError} + \#\text{DeletionError}}{\#\text{Buchstaben}}). \tag{2.3}$$

Der besseren Übersichtlichkeit wegen sind in der Tabelle nur die Ergebnisse mit den jeweils besten Werten für die von Hand einzustellenden Parameter aufgeführt. Da bei diesen Verfahren eine manuelle Einstellung der Werte für w_a bzw. b erforderlich war, waren mehrere Tests nötig, deren Ergebnisse in Anhang A in Tabelle A.1 zu finden sind. Eine grafische Darstellung der Ergebnisse befindet sich in Abbildung 2.1.

Graustufen-bilder	rein akustische Erkennung	von Hand gew. Add.		handoptim. Bias und Entropy	
		W.A.	wa	W.A.	Bias
CLEAN (30 dB)	94.1%	95.9%	0.8	96.5%	0.8
NOISE 1 (16 dB)	67.6%	75.9%	0.7	75.9%	0.8
NOISE 2 (8 dB)	49.4%	62.4%	0.7	62.9%	0.7
INCREASE 1 (30-16 dB)	74.7%	78.8%	0.6	85.9%	0.7
INCREASE 2 (30-8 dB)	44.1%	71.2%	0.5	69.4%	0.5
RADIO 1 ($\approx$ 20 dB)	92.4%	92.9%	0.8	93.5%	0.8
RADIO 2 ($\approx$ 17 dB)	72.9%	77.1%	0.8	78.2%	0.7
MOTOR 1 ($\approx$ 25 dB)	93.5%	94.7%	0.9	94.7%	0.8
MOTOR 2 ($\approx$ 11 dB)	50.6%	67.6%	0.8	68.2%	0.7

Tabelle 2.1: Testergebnisse mit den bisherigen Verfahren und Graustufenbildern als visuelle Eingabe. Angegeben ist jeweils die Word-Accuracy (W.A.) in Prozent, sowie das von Hand eingestellte, optimale Gewicht w_a bei der gewichteten Addition bzw. der von Hand eingestellte, beste Schwellwert (Bias b aus Gleichung 1.2, Seite 13) bei der Gewichtung über die Entropy. Die erste Spalte enthält zum Vergleich die Ergebnisse bei rein akustischer Erkennung.

LDA-Daten	rein akustische Erkennung	von Hand gew. Add.		handoptim. Bias und Entropy	
		W.A.	wa	W.A.	Bias
CLEAN (30 dB)	94.1%	97.6%	0.7	97.6%	0.7
NOISE 1 (16 dB)	67.6%	79.4%	0.6	80.0%	0.6
NOISE 2 (8 dB)	49.4%	75.3%	0.6	71.2%	0.7
INCREASE 1 (30-16 dB)	74.7%	84.7%	0.6	87.1%	0.8
INCREASE 2 (30-8 dB)	44.1%	78.2%	0.5	76.5%	0.5
RADIO 1 ($\approx$ 20 dB)	92.4%	95.3%	0.8	92.9%	0.8
RADIO 2 ($\approx$ 17 dB)	72.9%	87.1%	0.7	87.1%	0.7
MOTOR 1 ($\approx$ 25 dB)	93.5%	95.9%	0.8	95.3%	0.8
MOTOR 2 ($\approx$ 11 dB)	50.6%	75.9%	0.7	75.3%	0.7

Tabelle 2.2: Testergebnisse mit den bisherigen Verfahren und LDA-Daten als visuelle Eingabe.

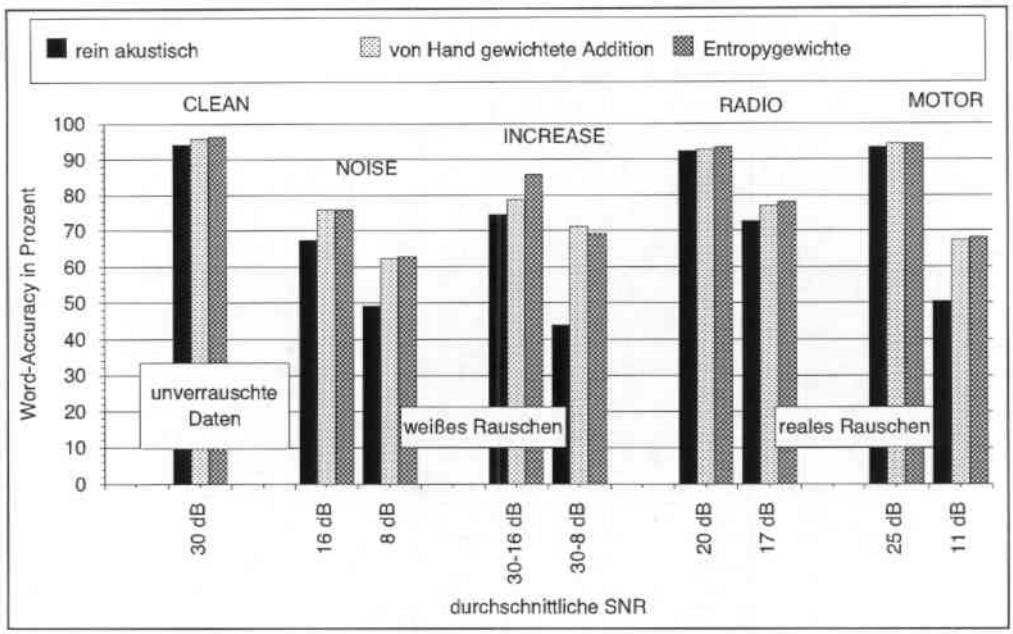


Abbildung 2.1: Grafische Darstellung der Testergebnisse der bisherigen Verfahren mit Graustufenbildern als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Prozent in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Es sei hier darauf hingewiesen, daß die angegebenen SNR-Werte nur Durchschnittswerte sind. Die tatsächliche SNR des jeweiligen akustischen Signals schwankt, insbesondere bei den RADIO- und MOTOR-Datenmengen, innerhalb gewisser Schranken um diese Mittelwerte.

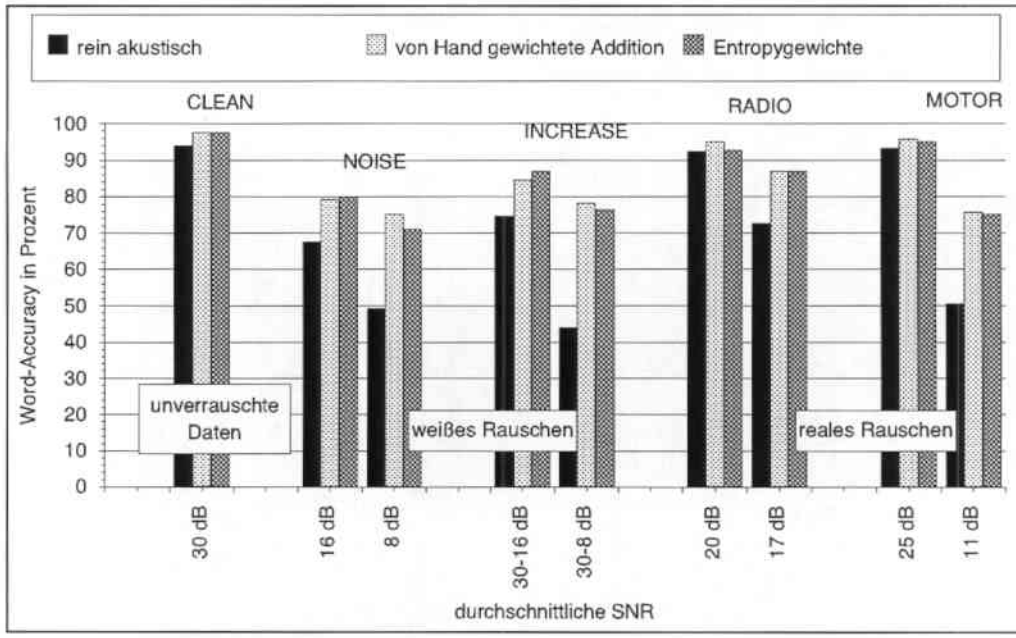


Abbildung 2.2: Grafische Darstellung der Testergebnisse der bisherigen Verfahren mit LDA-Daten als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Prozent in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16).

Man sieht, daß alle Kombinationsmöglichkeiten eine wesentliche Verbesserung gegenüber der

rein akustischen Erkennung (erste Spalte in der Tabelle) liefern, insbesondere bei den stark verrrauschten Eingabesignalen. Beide Verfahren erreichen annähernd die gleichen Ergebnisse. Ferner fällt der relativ kleine Schwellwert bei den Daten mit ansteigendem Rauschen (INCREASE 1 und 2) auf. Die besten Ergebnisse werden hier bei relativ hoher Gewichtung der visuellen Seite erzielt. Insbesondere im Vergleich zu den jeweiligen Schwellwerten der Datenmengen NOISE 1 und 2, die konstant mit einer SNR verrauscht sind, die in etwa dem maximalen SNR-Wert bei den Mengen INCREASE 1 und 2 mit ansteigendem Rauschen entspricht, erscheint der Wert für das Gewicht w_a bzw. den Bias b sehr gering. Es wäre eigentlich naheliegender, daß ein Signal, das zunächst unverrauscht ist und dessen Störgeräuschanteil im Laufe der Zeit immer mehr zunimmt, über die gesamte Zeit hinweg gesehen, akustisch besser zu interpretieren sein sollte als das gleiche Signal, dessen Störgeräuschanteil ständig den maximalen Wert besitzt. Auch die rein akustische Erkennung liefert für die Datenmenge INCREASE 2 eine schlechtere Word-Accuracy als bei den Wortfolgen aus der Testmenge NOISE 2. Eine Erklärung dieses Sachverhaltes ist möglicherweise in der Architektur des Erkenners zu finden. Die einzelnen Neuronen des TDNNs erhalten ja nicht nur die Aktivierungen der Neuronen der tieferen Schicht zum gleichen Zeitpunkt als Eingabe, sondern auch die der unmittelbar zuvor und danach anliegenden Werte. Durch diese Betrachtung der Aktivierungen über ein Zeitfenster könnte es möglich sein, daß das TDNN das ansteigende Rauschen als ein zusätzliches Signal ansieht, was zu Fehlinterpretationen führen kann.
Tabelle 2.2 enthält die entsprechenden Ergebnisse, wenn als visuelle Eingabe LDA-Daten verwendet werden. Grafisch ist die jeweilige Word-Accuracy in Abbildung 2.2 veranschaulicht. Eine ausführliche Ergebnistabelle mit allen durchgeführten Tests befindet sich in Anhang A (siehe Tabelle A.2).
Die erzielten Erkennungsraten entsprechen in etwa den Ergebnissen mit Graustufenbildern, sind sogar geringfügig besser. Überraschend ist jedoch die Tatsache, daß bei der Kombination mit entropyabhängiger Gewichtung bei den Daten NOISE 1 (SNR ungefähr 16 dB) mit einem von Hand eingestellten Schwellwert von 0.6 die besten Resultate erzielt wurden, während bei den NOISE 2-Daten (SNR ungefähr 8 dB) ein Schwellwert von 0.7 die beste Erkennungsrate lieferte. Bei schlechterer Qualität der akustischen Daten und gleichbleibender Qualität der visuellen Eingabe bringt also eine höhere Gewichtung der akustischen Seite ein besseres Ergebnis. Diese auf den ersten Blick völlig irreführende Tatsache macht auf ein Problem aufmerksam, das während dieser Arbeit bestand und das eine kritische Betrachtung der Ergebnisse nötig macht: die 30 zur Verfügung stehenden Testsamples sind zu wenig, um wirklich fundierte Aussagen zu machen. Wenn man bedenkt, daß bereits zwei falsch erkannte Buchstaben die Erkennungsrate um ungefähr ein Prozent sinken lassen können, werden auch scheinbar unerklärliche Ergebnisse, wie das eben angesprochene, verständlich. Da der Erkenner nur sprecherabhängig arbeitet, war ein Datensammeln und damit ein Erweitern der Trainings- und Testmenge im Laufe dieser Arbeit nicht möglich. Die aufgeführten Ergebnisse sind deshalb nur als Schätzungen und erste Richtlinien zu verstehen, sie erheben keinesfalls den Anspruch vollständiger Korrektheit bei der Ausdehnung auf größere Datenmengen.

2.1.2 Weitere Möglichkeiten der Kombination

In einem Ansatz von Stork, Wolff und Levine [18] wurde, ähnlich wie beim Karlsruher Ansatz, der Versuch unternommen, mit Hilfe von Neuronalen Netzen kombinierte akustische und visuelle Spracherkennung zu betreiben. Allerdings wurden anstelle eines MS-TDNNs zwei einfache TDNNs, je eines für eine der beiden Eingabemodalitäten, verwendet. Statt auf

Phonemebene wurde eine Erkennung auf Wortebene durchgeführt. Da sich die Aktivierungen der Ausgabeneuronen eines neuronalen Netzwerkes als Wahrscheinlichkeiten interpretieren lassen, wurden die jeweiligen Ausgabewerte normalisiert und frameweise multipliziert.
Eine analoge Vorgehensweise wurde nun auf den Karlsruher Erkenner übertragen. In einem ersten Ansatz wurden die Werte der Phonem- und der Visem-Schicht normalisiert und das Produkt der sich entsprechenden Neuronenaktivierungen in die Kombinationsschicht eingetragen. Die Ergebnisse der Tests sind in der zweiten Spalte von Tabelle 2.3 zu finden. Die erzielten Erkennungsraten liegen zwar über denen der rein akustischen Erkennung, sind aber schlechter als die bisherigen Verfahren, weshalb auch nur Graustufenbilder und nur ein Teil der Testdatenmenge ausgetestet wurden. Das Problem bei dieser Art der Kombination liegt darin, daß eine „Bevorzugung" einer Eingabemodalität durch Setzen des entsprechenden Gewichts nicht möglich ist. Deshalb wurde in einem weiteren Ansatz versucht, die Aktivierungen von Hand entsprechend der Güte der akustischen Eingabedaten zu beeinflussen. Vor der Normierung wurde auf die Neuronenaktivierungen des akustischen Netzes ein Faktor s_a, auf die visuellen ein Faktor s_v mit $s_v = 1 - s_a$ aufaddiert. Die Testergebnisse dieses Verfahrens sind in Tabelle 2.3 in der dritten Spalte aufgeführt. Sie entsprechen in etwa den Ergebnissen, die mit den bisherigen Verfahren erzielt wurden. Weitere Testergebnisse mit anderen Werten für s_a und s_v befinden sich in Anhang A, Tabelle A.3. Ein voneinander unabhängiges Setzen der Werte s_a und s_v wurde auch getestet, brachte aber keine zufriedenstellenden Ergebnisse.

Graustufen-bilder	rein akustische Erkennung	Multiplikation der Ausgaben	modifizierte Multiplikation	
			Wordacc.	s_a
CLEAN (30 dB)	94.1%	91.8%	95.3%	0.1
NOISE 1 (16 dB)	67.6%	71.8%	75.3%	0.3
NOISE 2 (8 dB)	49.4%	59.4%	64.7%	0.3

Tabelle 2.3: Testergebnisse bei Multiplikation der einzelnen Aktivierungen.

In einem weiteren Ansatz wurde eine Kombination nach folgendem Schema vorgenommen:

$$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V , \tag{2.4}$$

wobei die Aktivierungen hyp_i^A bzw. hyp_i^V der Ausgabeneuronen des akustischen bzw. visuellen TDNNs vorher frameweise normiert wurden. Die Idee zu dieser Vorgehensweise kam vom Expertensystem MYCIN, bei dem auf analoge Art gleiche Aussagen, die mit einem unterschiedlichen Sicherheitsmaß vorliegen, verknüpft werden. Desweiteren wurde ein ähnlicher Ansatz mit folgender Kombination untersucht:

$$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V . \tag{2.5}$$

Auch hier wurden die entsprechenden Werte vor der Verknüpfung normiert.
Motiviert durch die guten Ergebnisse wurde auch eine einfache gewichtete Addition (analog Gleichung 2.1 auf Seite 16) mit vorheriger Normierung der einzelnen Frames untersucht. Die Ergebnisse mit den besten, von Hand optimierten Gewichten für diese Verfahren sind in Tabelle 2.4 zusammengestellt. Ausführlichere Tabellen mit Testergebnissen für weitere Werte von w_a befinden sich in Anhang A (siehe die Tabellen A.4 und A.5). Die Abbildungen 2.3

Graustufen-bilder	Verfahren 1		Verfahren 2		Verfahren 3	
	Wordacc.	w_a	Wordacc.	w_a	Wordacc.	w_a
CLEAN (30 dB)	96.5%	0.55	95.9%	0.7	95.9%	0.6
NOISE 1 (16 dB)	76.5%	0.5	77.1%	0.4	77.1%	0.4
NOISE 2 (8 dB)	67.1%	0.4	65.9%	0.4	66.5%	0.4
INCREASE 1 (30-16 dB)	85.3%	0.4	86.5%	0.4	87.6%	0.4
INCREASE 2 (30-8 dB)	65.3%	0.3	67.1%	0.2	65.9%	0.3
RADIO 1 ($\approx$ 20 dB)	92.4%	0.5	92.4%	0.5	92.4%	0.5
RADIO 2 ($\approx$ 17 dB)	82.9%	0.5	80.6%	0.5	80.0%	0.5
MOTOR 1 ($\approx$ 25 dB)	95.9%	0.6	95.3%	0.6	95.3%	0.5
MOTOR 2 ($\approx$ 11 dB)	71.2%	0.5	71.2%	0.5	71.2%	0.5

mit

Verfahren 1: $$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V \quad (2.4)$$

Verfahren 2: $$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V \quad (2.5)$$

Verfahren 3: $$hyp_i^B = w_a hyp_i^A + w_v hyp_i^V \quad (2.1)$$

wobei die Hypothesen hyp_i jeweils frameweise normalisiert wurden

LDA-Daten	Verfahren 1		Verfahren 2		Verfahren 2	
	Wordacc.	w_a	Wordacc.	w_a	Wordacc.	w_a
CLEAN (30 dB)	96.5%	0.7	97.1%	0.5	97.1%	0.5
NOISE 1 (16 dB)	77.6%	0.6	77.6%	0.5	77.6%	0.4
NOISE 2 (8 dB)	72.9%	0.5	72.9%	0.5	71.8%	0.5
INCREASE 1 (30-16 dB)	86.5%	0.5	87.6%	0.6	87.6%	0.6
INCREASE 2 (30-8 dB)	71.2%	0.4	75.3%	0.4	74.7%	0.4
RADIO 1 ($\approx$ 20 dB)	93.5%	0.6	94.1%	0.6	94.1%	0.6
RADIO 2 ($\approx$ 17 dB)	87.6%	0.6	87.1%	0.6	87.1%	0.6
MOTOR 1 ($\approx$ 25 dB)	96.5%	0.6	97.1%	0.7	96.5%	0.5
MOTOR 2 ($\approx$ 11 dB)	76.5%	0.6	78.8%	0.6	78.2%	0.6

Tabelle 2.4: Testergebnisse mit weiteren Verfahren. Bei den Ergebnissen der oberen Tabelle wurden Graustufenbilder, bei denen der unteren LDA-Daten als visuelle Eingabe verwendet. Das Gewicht w_a wurde jeweils von Hand gesetzt.

und 2.4 stellen die Ergebnisse grafisch dar. Die erzielten Erkennungsraten entsprechen in etwa denen, die mit den bisherigen Verfahren erreicht wurden (vgl. Tab. 2.1 für Graustufenbilder und Tab. 2.2 für LDA-Daten).

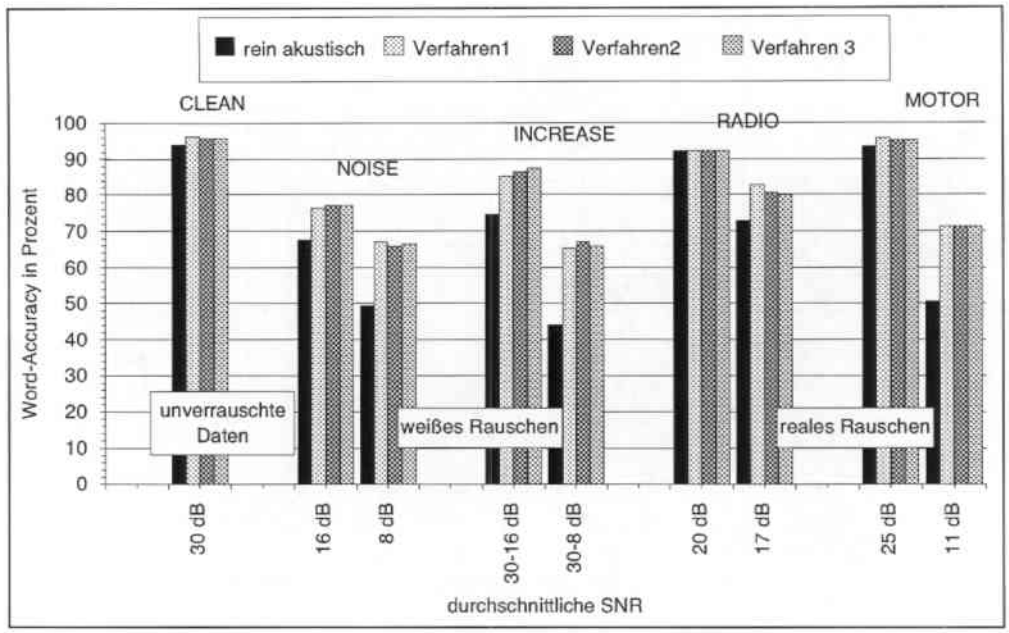


Abbildung 2.3: Grafische Darstellung der Testergebnisse der Verfahren aus Tabelle 2.4 mit Graustufenbildern als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Für die Kombinationsschemata der Verfahren 1 bis 3 siehe Tabelle 2.4.

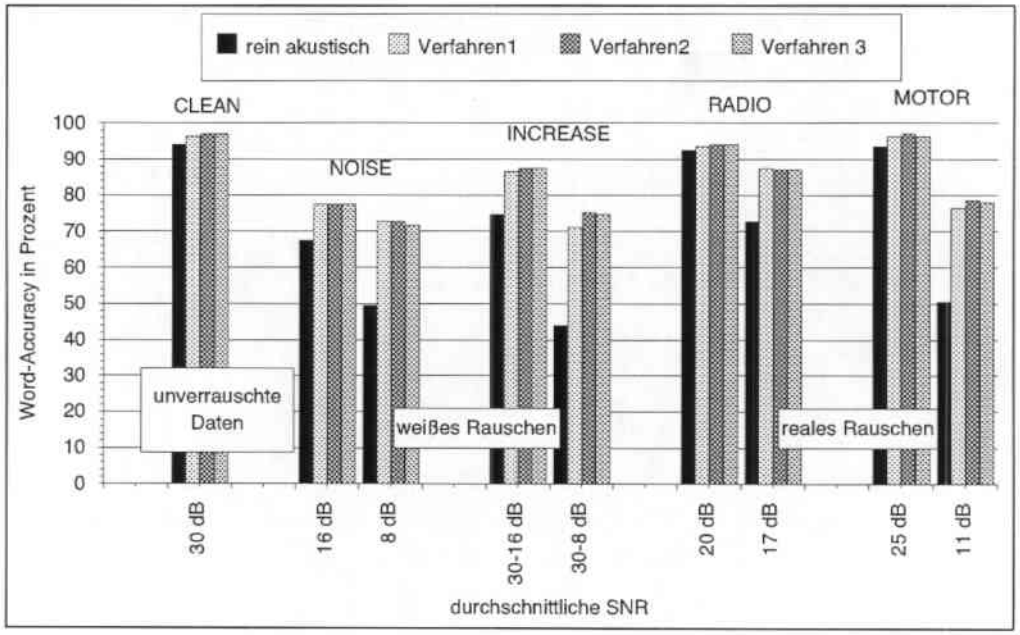


Abbildung 2.4: Grafische Darstellung der Testergebnisse der Verfahren aus Tabelle 2.4 mit LDA-Daten als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Für die Kombinationsschemata der Verfahren 1 bis 3 siehe Tabelle 2.4.

Ein weiteres System zur bimodalen Spracherkennung wurde von P.L. Silsbee [16] entwickelt. Bei diesem System handelt es sich allerdings nicht um einen neuronalen Erkenner, es basiert

	$hyp_i^A + hyp_i^{V^{w_v}}$		$hyp_i^{A^{w_a}} + hyp_i^{V^{w_v}}$	
	Wordacc.	w_v	Wordacc.	w_a
CLEAN (30 dB)	94.7%	0.1	94.7%	0.9
NOISE 1 (16 dB)	71.8%	0.8	75.9%	0.3
NOISE 2 (8 dB)	59.4%	0.9	65.9%	0.2

Tabelle 2.5: Testergebnisse mit weiteren Verfahren und Graustufenbildern als visuelle Eingabe. Die Werte w_a und w_v wurden jeweils von Hand gesetzt.

vielmehr auf einem sogenannten *Hidden-Markov-Modell.* Aufbauend auf der dort verwendeten Sensorfusion wurden die folgenden weiteren Kombinationsalternativen getestet:

$$hyp_i^B = hyp_i^A + hyp_i^{V^{w_v}}, \tag{2.6}$$

mit w_v von Hand gesetzt und

$$hyp_i^B = hyp_i^{A^{w_a}} + hyp_i^{V^{w_v}}, \tag{2.7}$$

wobei w_a von Hand gesetzt und w_v durch $w_v = 1 - w_a$ berechnet wird. Die Ergebnisse mit den besten Werten für w_a und w_v befinden sich in Tabelle 2.5, Ergebnisse mit weiteren Werten für die von Hand einzustellenden Gewichte sind in Anhang A in Tabelle A.6 dargestellt. Da die Ergebnisse keine wesentlichen Verbesserungen gegenüber den bisherigen Verfahren lieferten, wurden die Tests nur auf einem Teil der Testmenge und nur mit Graustufenbildern durchgeführt.

Zusammenfassend läßt sich an den verschiedenen verwendeten Methoden erkennen, daß unterschiedliche Vorgehensweisen existieren, die alle gute Ergebnisse liefern. Jedoch gibt es unter den getesteten Verfahren keines, dessen Erkennungsraten sich wesentlich von den anderen abheben. Dies war mit ein Grund dafür, weshalb in Kapitel 2.3 ein weiterer Ansatz getestet wurde, der sich prinzipiell von den hier verwendeten unterscheidet.

2.2 Automatisierung der Verfahren

Die bisher beschriebenen Verfahren besitzen alle einen großen Nachteil: für eine optimale Erkennung unter verschiedenen äußeren Bedingungen ist eine Anpassung einiger Parameter von Hand nötig. Dieser erforderliche manuelle Eingriff verhindert insbesondere eine optimale Online-Erkennung. Deshalb wird hier der Versuch unternommen, die Kombinationsalternativen, die im vorangegangenen Kapitel die besten Ergebnisse erzielt haben, zu automatisieren, so daß ein manuelles Einstellen der Parameter nicht mehr erforderlich ist.
Dies geschieht unter Zuhilfenahme der bereits in Kapitel 2.1.1 erwähnten *Signal-to-Noise Ratio* SNR (vgl. Gleichung 2.2 auf Seite 16). Für die Bestimmung des SNR-Wertes eines gegebenen akustischen Signals wird eine von M. Schoch im Rahmen einer Studienarbeit [15] durchgeführte Implementierung eines Algorithmus von H.G. Hirsch verwendet. Der Algorithmus basiert auf einer statistischen Analyse der Energieverteilung des akustischen Sprachsignals. Für eine genaue Beschreibung des Verfahrens sei auf [8] verwiesen.

Der geschätzte SNR-Wert wurde nun dazu benutzt, die Gewichtungsfaktoren der entsprechenden Verfahren automatisch einzustellen, anstatt sie von Hand zu setzen. Ausgehend von der Überlegung, daß ab einem bestimmten SNR-Wert SNR_{max} die Qualität der akustischen Signale so gut ist, daß eine stärkere Gewichtung der visuellen Seite wohl keine Verbesserung mehr bewirkt, wurden für zwei geeignete Werte SNR_{max} und SNR_{min} je ein Gewicht $w_{a_{max}}$ bzw. $w_{a_{min}}$ vorgegeben. Für alle SNR-Werte größer SNR_{max}, bzw. für alle SNR-Werte kleiner SNR_{min} wurde w_a konstant auf $w_{a_{max}}$ bzw. $w_{a_{min}}$ gehalten. Zwischen den beiden Eckdaten wurde eine lineare Interpolation durchgeführt. Eine grafische Darstellung der so entstandenen Funktion ist in Abbildung 2.5 zu finden. Der visuelle Gewichtungsfaktor w_v wird durch $w_v = 1 - w_a$ bestimmt. Die Werte für die Parameter SNR_{max} und SNR_{min}, sowie für $w_{a_{max}}$ und $w_{a_{min}}$ wurden in Abhängigkeit von den einzelnen Verfahren zu Beginn von Hand gesetzt. Die so entstandene Abbildungsfunktion der SNR-Werte auf die Gewichte w_a wurde dann unabhängig von den jeweiligen Eingabedaten für die entsprechenden Ansätze verwendet.

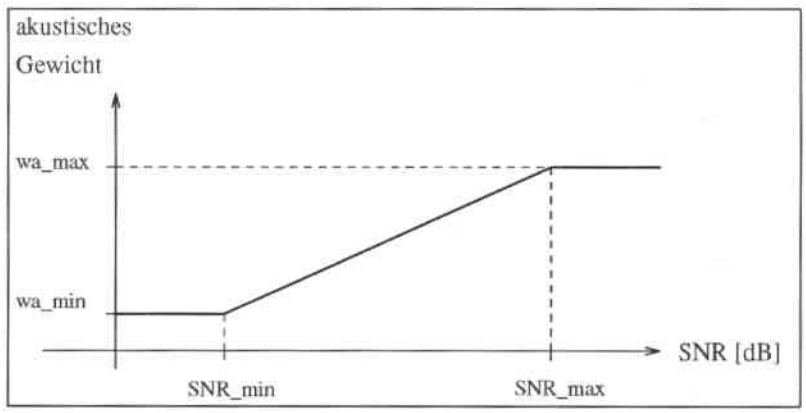


Abbildung 2.5: Grafische Darstellung der Funktion, die für einen gegebenen SNR-Wert eines akustischen Signals den Gewichtungsfaktor w_a für die akustischen Daten bei der kombinierten Erkennung berechnet.

Der SNR-Algorithmus liefert alle 500 ms eine über 1000 ms gemittelte Schätzung für die SNR des akustischen Signals im jeweiligen Zeitabschnitt. Da der Erkenner jedoch alle 10 ms ein Eingangssignal erhält, sind zusätzliche SNR-Werte erforderlich. Zwei Alternativen wurden für die Werte zwischen zwei SNR-Schätzungen, die vom Algorithmus geliefert werden, realisiert:

- es wurde der jeweils zeitlich am nächsten liegende SNR-Wert verwendet,
- es wurde zwischen den beiden zeitlich am nächsten liegenden SNR-Werten linear interpoliert.

Beide Vorgehensweisen sind an einem Beispiel in Abbildung 2.6 exemplarisch dargestellt. Da die Verwendung der linear interpolierten Werte immer, wenn auch nur geringfügig, bessere Ergebnisse lieferte, wurde im folgenden nur noch diese Vorgehensweise verwendet.
Für die Ermittlung der SNR-Werte durch den Algorithmus standen zwei Implementierungen zur Verfügung: einmal wurde für die Berechnung der SNR, also des Verhältnisses von Energie des Sprachsignals zur Energie des Störgeräuschs, die Signalenergie jeweils pro Frame bestimmt, zum anderen wurde eine Mittelung der Sprachenergie über den gesamten gesprochenen Satz verwendet. Wie erwartet führte die zweite Variante zu besseren Ergebnissen und wurde deshalb im folgenden ausschließlich angewandt.

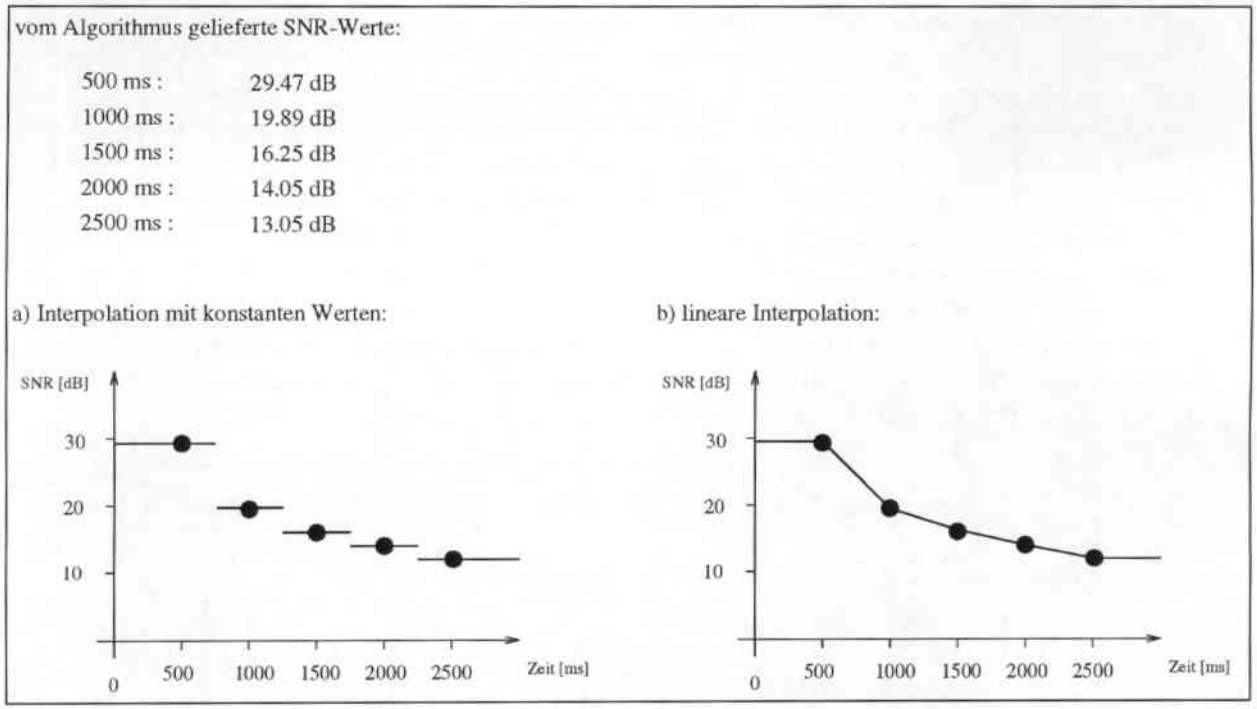


Abbildung 2.6: Beispiel für die beiden Verfahren, die für die Interpolation der benötigten SNR-Werte verwendet wurden. Die SNR-Werte wurden auf einer Buchstabenfolge aus der Testmenge INCREASE 2 (von ca. 30 dB auf ca. 8 dB abfallende SNR) berechnet.

Für die vier besten Kombinationsalternativen aus Kapitel 2.1 wurde in einem ersten Ansatz versucht, die dort von Hand einzustellenden Parameter durch die oben beschriebene Funktion automatisch aus dem jeweiligen SNR-Wert zu bestimmen. Im einzelnen waren dies folgende Verfahren:

- $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$, (vgl. Gleichung (2.1))
- $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$
  mit frameweise normierten Aktivierungen hyp_i,
- $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (vgl. Gleichung (2.4))
  mit frameweise normierten Aktivierungen hyp_i,
- $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (vgl. Gleichung (2.5))
  mit frameweise normierten Aktivierungen hyp_i.

Die Parameter für die jeweiligen Funktionen zur Bestimmung von w_a aus dem SNR-Wert wurden anhand der Ergebnisse der bisher verwendeten Verfahren von Hand festgelegt. Dazu wurden diverse Tests mit manuell gesetzten Gewichten durchgeführt. Anhand der besten Ergebnisse, die mit unverrauschten (CLEAN, 30 dB), leicht verrauschten (NOISE 1, 16 dB) und stark verrauschten (NOISE 2, 8 dB) Datenmengen erzielt wurden, wurden die Werte für die Parameter $w_{a_{min}}$, $w_{a_{max}}$, sowie für SNR_{min} und SNR_{max} derart festgelegt, daß man für die SNR-Werte 30, 16 und 8 dB jeweils in etwa den gleichen Wert für w_a erhält, mit dem man bei Handoptimierung jeweils die besten Ergebnisse erzielen konnte.

Da die Erkennungsraten extrem abhängig von der Wahl dieser Parameter sind, wurden Tests mit verschiedenen Werten durchgeführt. Die besten Ergebnisse sind zusammen mit den jeweiligen Parameterwerten in Tabelle 2.6 aufgeführt. Ergebnisse mit anderen Parametereinstellungen sind in Anhang B in den Tabellen B.1 und B.2 zu finden. Die besten Werte für die beiden Parameter SNR_{min} und SNR_{max} der Abbildungsfunktion von der SNR auf das Gewicht w_a (vgl. Abbildung 2.5) lagen bei den meisten Verfahren ungefähr bei 0 dB, bzw. 30 dB. Eine grafische Darstellung der erreichten Word-Accuracy in Abhängigkeit von der jeweiligen SNR ist in Abbildung 2.7 und 2.9 dargestellt. Mit Ausnahme der Erkennungsraten für die INCREASE 2-Daten, die auch schon bei den bisherigen Verfahren schlechte Ergebnisse lieferten, entspricht die Word-Accuracy in etwa den Ergebnissen, die man mit den von Hand eingestellten Parametern erhalten hat (vgl. Kapitel 2.1, Tabelle 2.1 (Graustufenbilder) bzw. Tabelle 2.2 (LDA-Daten), sowie Tabelle 2.4).

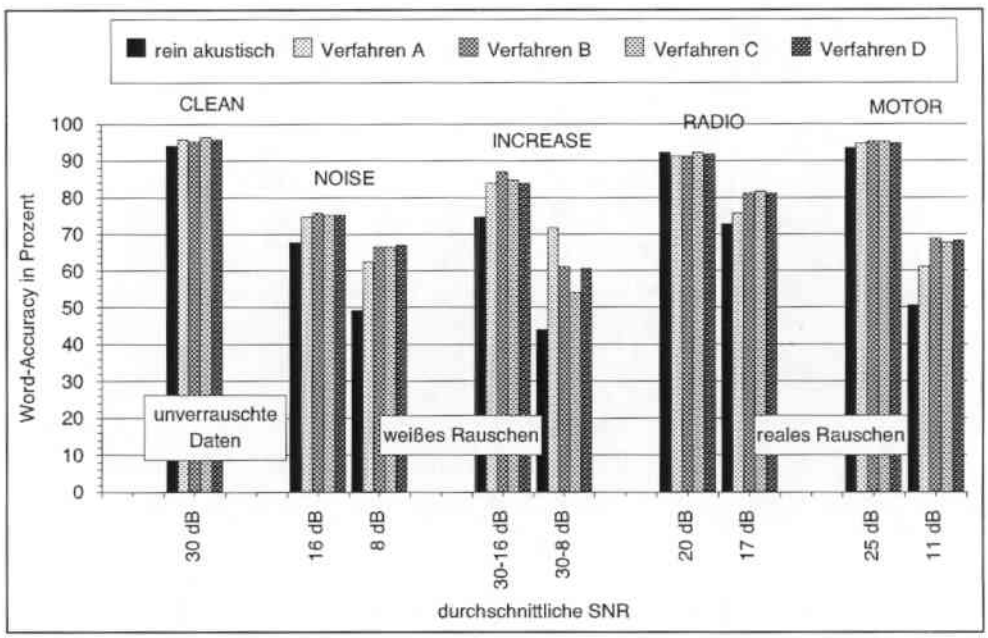


Abbildung 2.7: Grafische Darstellung der Testergebnisse der automatisierten Verfahren mit Graustufenbildern als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Die Kombinationsschemata der einzelnen Verfahren waren wie folgt:
- Verfahren A: $w_a hyp_i^A + w_v hyp_i^V$,
- Verfahren B: $w_a hyp_i^A + w_v hyp_i^V$ mit Normierung,
- Verfahren C: $w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung,
- Verfahren D: $w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung.

Die jeweiligen Gewichte w_a wurden automatisch über die SNR der akustischen Eingabesignale bestimmt.

Ein Beispiel für die automatische Gewichtsanpassung in Abhängigkeit von der SNR ist in Abbildung 2.8 dargestellt.

In einem weiteren Ansatz wurde der Versuch unternommen, die zusätzliche Information, die die Entropy über die Zuverlässigkeit einer Netzausgabe liefert, zu nutzen. Die Gewichte w_a und w_v werden nun in Abhängigkeit von der SNR und der Entropy des jeweiligen Frames automatisch gesetzt. Statt den Bias b aus Gleichung 1.2 auf Seite 13 von Hand zu setzen, wird er nun analog zu der Vorgehensweise im vorherigen Ansatz über die SNR bestimmt.

Graustufen-bilder	$w_a hyp_i^A + w_v hyp_i^V$	$w_a hyp_i^A + w_v hyp_i^V$ mit Normierung	$w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung	$w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung
	Word-Accuracy	Word-Accuracy	Word-Accuracy	Word-Accuracy
CLEAN (30 dB)	95.9%	95.3%	96.5%	95.9%
NOISE 1 (16 dB)	74.7%	75.9%	75.3%	75.3%
NOISE 2 (8 dB)	62.4%	66.5%	66.5%	67.1%
INCREASE 1 (30-16 dB)	84.1%	87.1%	84.7%	84.1%
INCREASE 2 (30-8 dB)	71.8%	61.2%	54.1%	60.6%
RADIO 1 ($\approx$ 20 dB)	91.2%	91.2%	92.4%	91.8%
RADIO 2 ($\approx$ 17 dB)	75.9%	81.2%	81.8%	81.2%
MOTOR 1 ($\approx$ 25 dB)	94.7%	95.3%	95.3%	94.7%
MOTOR 2 ($\approx$ 11 dB)	61.2%	68.8%	67.6%	68.2%
	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen
SNR_{min}, SNR_{max}	-5, 25	2, 30	0, 32	1, 33
$w_{a_{min}}$, $w_{a_{max}}$	0.2, 0.8	0.35, 0.5	0.35, 0.55	0.3, 0.6

LDA-Daten	$w_a hyp_i^A + w_v hyp_i^V$ (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V$ mit Normierung (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung (Word-Accuracy)
CLEAN (30 dB)	97.1%	97.1%	96.5%	97.1%
NOISE 1 (16 dB)	78.8%	77.1%	73.5%	78.2%
NOISE 2 (8 dB)	71.2%	70.6%	72.9%	71.2%
INCREASE 1 (30-16 dB)	82.9%	85.3%	82.9%	86.5%
INCREASE 2 (30-8 dB)	74.7%	68.8%	60.6%	66.5%
RADIO 1 ($\approx$ 20 dB)	94.1%	93.5%	93.5%	93.5%
RADIO 2 ($\approx$ 17 dB)	87.6%	85.9%	85.9%	85.3%
MOTOR 1 ($\approx$ 25 dB)	94.1%	96.5%	95.3%	96.5%
MOTOR 2 ($\approx$ 11 dB)	73.5%	74.1%	72.9%	72.9%
	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen
SNR_{min}, SNR_{max}	0, 33	-2, 30	-1, 33	-2, 28
$w_{a_{min}}$, $w_{a_{max}}$	0.5, 0.75	0.4, 0.55	0.35, 0.8	0.4, 0.55

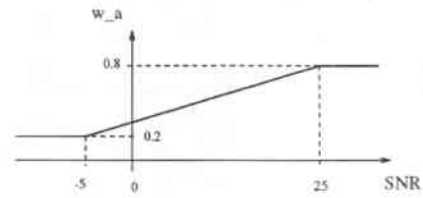
Tabelle 2.6: Testergebnisse der verschiedenen Verfahren bei automatischer Gewichtung der beiden Komponenten in Abhängigkeit vom jeweiligen SNR-Wert der akustischen Daten. Die obere Tabelle enthält die Ergebnisse bei Verwendung von Graustufenbildern als visuelle Eingabe, die untere enthält die Erkennungsraten bei LDA-Daten.

Automatisch bestimmte Gewichte aus der SNR der akustischen Daten am Beispiel der Buchstaben-Sequenz " T H O R A L F" aus der Testdatenmenge RADIO 1

Kombinationsschema:

$$hyp\_B = w\_a\ hyp\_A + w\_v\ hyp\_V$$

Funktionen zur automatischen Bestimmung der Gewichte aus den SNR-Werten der akustischen Daten:



$$w\_v = 1 - w\_a$$

Fuer die Buchstabensequenz "T H O R A L F" gelieferte SNR-Werte und daraus berechnete Gewichte w_a:

500 ms:	24.923 dB	w_a = 0.798
1000 ms:	22.683 dB	w_a = 0.753
1500 ms:	20.470 dB	w_a = 0.709
2000 ms:	21.836 dB	w_a = 0.737
2500 ms:	24.512 dB	w_a = 0.790
3000 ms:	24.813 dB	w_a = 0.796
3500 ms:	27.181 dB	w_a = 0.800

Grafische Darstellung der automatisch bestimmten Gewichte:

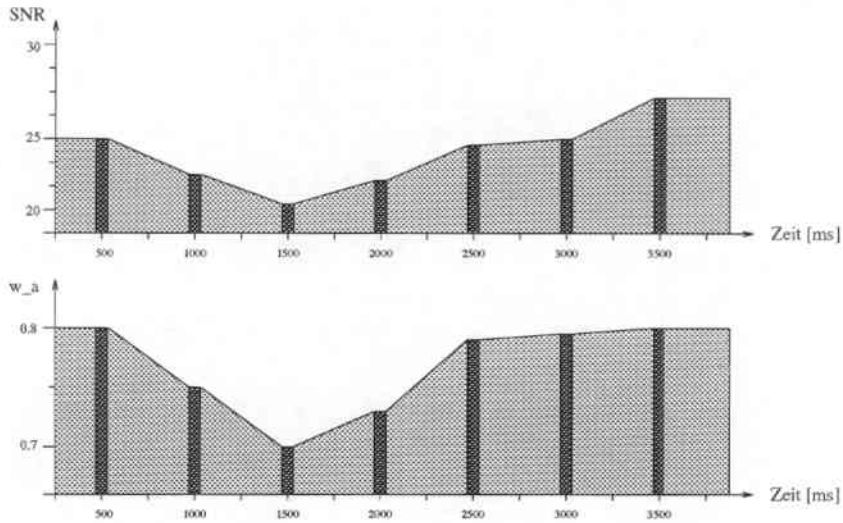


Abbildung 2.8: Beispiel für die automatische Bestimmung der Gewichte w_a und w_v in Abhängigkeit von den jeweiligen SNR-Werten der akustischen Daten. Der besseren Übersichtlichkeit wegen sind nur die SNR-Werte, die vom Algorithmus geliefert werden, dargestellt. Die interpolierten Werte sind im Diagramm nur angedeutet.

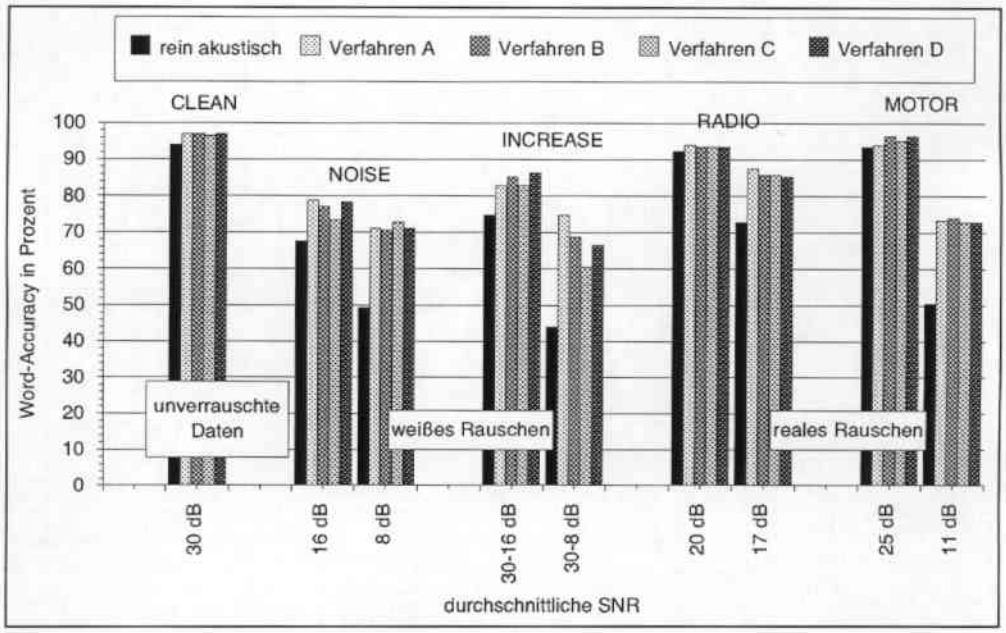


Abbildung 2.9: Grafische Darstellung der Testergebnisse der automatisierten Verfahren mit LDA-Daten als visuelle Eingabe. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Für die Kombinationsschemata der einzelnen Verfahren siehe Abbildung 2.7 auf Seite 26.

Es wurden wieder die vier besten Verfahren aus dem vorangegangenen Kapitel getestet. Die Ergebnisse befinden sich in Tabelle 2.7. Weitere getestete Parametereinstellungen sind in Anhang B in den Tabellen B.3 und B.4 zu finden. In den Abbildungen 2.10 und 2.11 sind die erzielten Ergebnisse graphisch veranschaulicht. Wie man sieht, wurden annähernd die gleichen Ergebnisse wie bisher erzielt. Die zusätzliche Verwendung der Entropy brachte folglich keine wesentliche Verbesserung der Erkennungsleistung.

Vergleicht man die hier untersuchten Ansätze, bei denen die Gewichtung jeweils automatisch bestimmt wurde, mit den Verfahren, bei denen die Gewichte von Hand eingestellt werden mußten, so läßt sich feststellen, daß, abgesehen von den beiden Datensets INCREASE 1 und 2, in etwa die gleichen Ergebnisse erzielt wurden. Es sei hier noch einmal darauf hingewiesen, daß die Parameter der Funktion, die die Gewichte w_a aus den SNR-Werten der akustischen Daten berechnet, anhand der handoptimierten Ergebnisse der Datenmengen NOISE 1 und 2 eingestellt wurden (vgl. Seite 25). Trotzdem konnten auf den Testdatensets RADIO 1 und 2, sowie MOTOR 1 und 2 in etwa die gleichen Ergebnisse erzielt werden, wie bei den vorherigen Verfahren, obwohl es sich hier um eine völlig andere Art des Rauschens handelt.

2.3 Phonemabhängige Gewichtung

Die bisher verwendeten Verfahren hatten eines gemeinsam: die Gewichte w_a und w_v wurden für jeden Frame neu gesetzt, waren jedoch für die einzelnen Merkmale eines Frames konstant. Auf diese Weise hatte ein Phonem, das beispielsweise akustisch sehr gut zu erkennen ist, visuell jedoch eine hohe Verwechselbarkeit mit anderen Visemen besitzt, die gleiche Gewichtung

Graustufen-bilder	$w_a hyp_i^A + w_v hyp_i^V$	$w_a hyp_i^A + w_v hyp_i^V$ mit Normierung	$w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung	$w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung
	Word-Accuracy	Word-Accuracy	Word-Accuracy	Word-Accuracy
CLEAN (30 dB)	96.5%	95.9%	96.5%	96.5%
NOISE 1 (16 dB)	75.9%	76.5%	76.5%	76.5%
NOISE 2 (8 dB)	64.1%	67.1%	65.9%	65.9%
INCREASE 1 (30-16 dB)	82.9%	86.5%	84.7%	85.9%
INCREASE 2 (30-8 dB)	58.2%	58.8%	58.2%	58.8%
RADIO 1 ($\approx$ 20 dB)	93.5%	90.6%	92.4%	91.8%
RADIO 2 ($\approx$ 17 dB)	79.4%	78.8%	80.0%	79.4%
MOTOR 1 ($\approx$ 25 dB)	94.7%	95.3%	95.3%	95.3%
MOTOR 2 ($\approx$ 11 dB)	67.1%	66.5%	66.5%	65.9%
	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen
SNR_{min}, SNR_{max}	-1, 18	2, 30	0, 32	1, 33
$w_{a_{min}}$, $w_{a_{max}}$	0.55, 0.8	0.35, 0.5	0.35, 0.55	0.3, 0.6

LDA-Daten	$w_a hyp_i^A + w_v hyp_i^V$ (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V$ mit Normierung (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung (Word-Accuracy)	$w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung (Word-Accuracy)
CLEAN (30 dB)	97.1%	95.9%	95.9%	95.9%
NOISE 1 (16 dB)	81.2%	77.6%	75.9%	77.6%
NOISE 2 (8 dB)	70.6%	71.2%	70.6%	71.8%
INCREASE 1 (30-16 dB)	84.1%	83.5%	82.9%	84.1%
INCREASE 2 (30-8 dB)	74.1%	59.4%	57.1%	60.0%
RADIO 1 ($\approx$ 20 dB)	92.4%	94.1%	93.5%	94.1%
RADIO 2 ($\approx$ 17 dB)	86.5%	87.6%	85.9%	87.6%
MOTOR 1 ($\approx$ 25 dB)	92.9%	95.3%	95.9%	95.3%
MOTOR 2 ($\approx$ 11 dB)	71.8%	78.2%	75.9%	78.8%
	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen	Parameter-einstellungen
SNR_{min}, SNR_{max}	0, 33	0, 33	0, 33	0, 33
$w_{a_{min}}$, $w_{a_{max}}$	0.5, 0.75	0.5, 0.75	0.5, 0.75	0.5, 0.75

Tabelle 2.7: Testergebnisse der verschiedenen Verfahren bei automatischer Gewichtung der beiden Komponenten in Abhängigkeit vom jeweiligen SNR-Wert der akustischen Daten und den Entropywerten der entsprechenden Frames der Ausgabeschichten der beiden TDNNs. In der oberen Tabelle sind die Erkennungsraten bei Verwendung von Graustufenbildern als visuelle Eingabe dargestellt, in der unteren die Ergebnisse bei LDA-Daten.

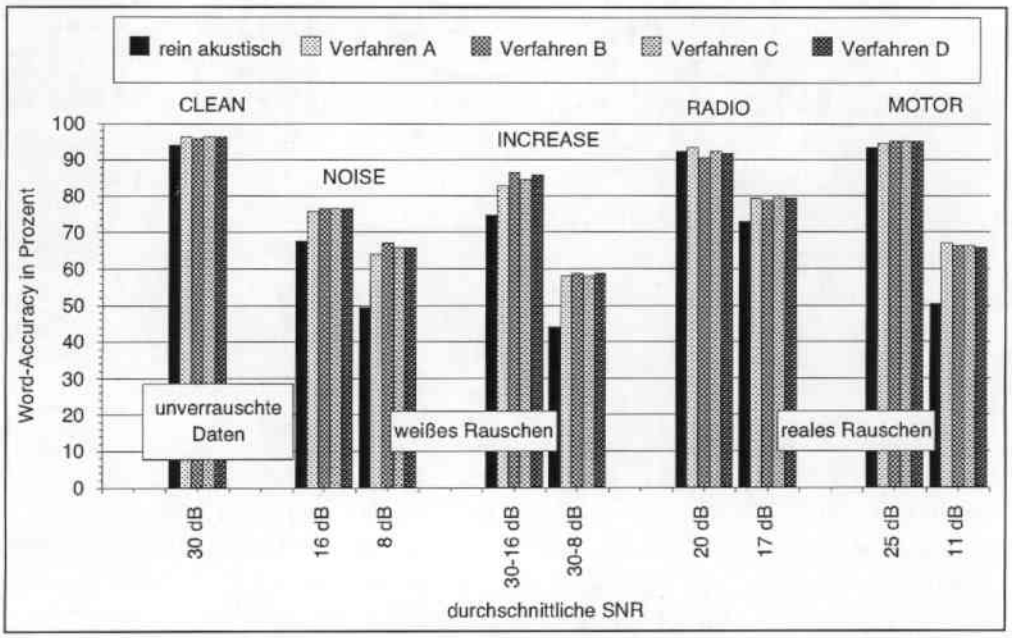


Abbildung 2.10: Grafische Darstellung der Testergebnisse der automatisierten Verfahren mit Graustufenbildern als visuelle Eingabe und zusätzlicher Verwendung der Entropy. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Die Kombinationsschemata der einzelnen Verfahren waren wie folgt:
- Verfahren A: $w_a hyp_i^A + w_v hyp_i^V$,
- Verfahren B: $w_a hyp_i^A + w_v hyp_i^V$ mit Normierung,
- Verfahren C: $w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ mit Normierung,
- Verfahren D: $w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ mit Normierung.

Die Gewichte w_a wurden automatisch über die SNR der akustischen Eingabesignale und die Entropy des jeweiligen Ausgabeframes bestimmt.

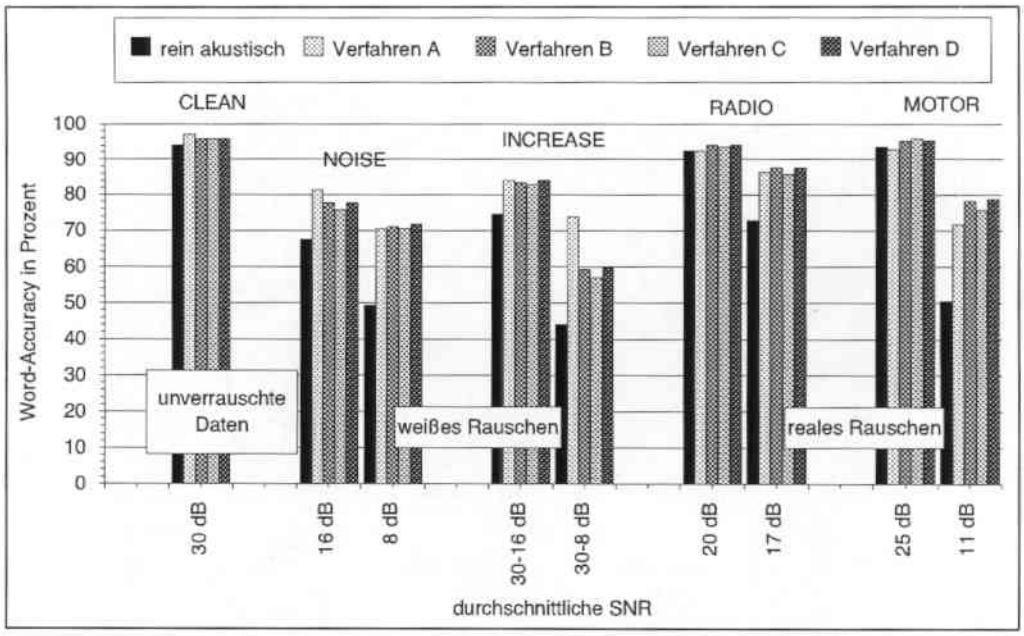


Abbildung 2.11: Grafische Darstellung der Testergebnisse der automatisierten Verfahren mit LDA-Daten als visuelle Eingabe und zusätzlicher Verwendung der Entropy. Dargestellt ist die erzielte Word-Accuracy in Abhängigkeit von der jeweiligen durchschnittlichen SNR der einzelnen Testsets der Testdatenbasis (vgl. Seite 16). Für das Kombinationsschema der einzelnen Verfahren siehe Abbildung 2.10 auf Seite 31.

wie ein anderes Phonem, bei dem sich die akustische und visuelle Verwechselbarkeit gerade umgekehrt verhalten. Beispielsweise wurde das Phonem /ehI/ auf einer hier verwendeten Testmenge vom akustischen TDNN in 63.5% der Fälle richtig erkannt, während das zugehörige Visem /_ehI/ nur in 12.9% der Fälle auch als solches erkannt wurde. Dagegen wurde das Phonem /bI/ nur in 35.2% der Fälle, das entsprechende Visem hingegen in 70.0% der Fälle richtig klassifiziert. Eine nicht nur über die Frames sondern auch über die einzelnen Merkmale variierende Gewichtung, die also nicht nur vom Grad der Verrauschung der akustischen Daten, sondern auch von der Verwechselbarkeit der einzelnen Phoneme bzw. Viseme abhängt, läßt somit auf eine weitere Verbesserung der Erkennungsrate hoffen.

Eine solche phonemabhängige Gewichtung wurde im folgenden auf der Basis sogenannter Konfusionsmatrizen realisiert. Bei den Konfusionsmatrizen handelt es sich um (nxn)-Matrizen (n = Anzahl der Phoneme), deren Spalten die gesprochenen Phoneme einer bestimmten Testmenge angeben und in deren Zeilen die tatsächlich erkannten Phoneme aufgetragen sind. Ein Eintrag in die Matrix an der Stelle (j, i) gibt an, wie oft ein bestimmtes Phonem p_i als ein Phonem p_j erkannt wurde. Die Diagonale dieser Matrix enthält, bei Normalisierung der Spalten, somit die relative Häufigkeit, mit der ein gesprochenes Phonem p_i richtig als p_i erkannt wurde. Anhand dieser Daten wurden nun für jedes Phonem p_i die Kombinationsgewichte w_{a_i} und w_{v_i} wie folgt festgelegt:

$$w_{a_i} = \min\{1,\ b + \frac{corr_{phone}}{corr_{phone} + corr_{visem}}\} \tag{2.8}$$

und

$$w_{v_i} = 1 - w_{a_i},$$

wobei $corr_{phone}$ die aus der Konfusionsmatrix ermittelte relative Häufigkeit von richtig erkannten Phonemen p_i und $corr_{visem}$ die relative Häufigkeit der korrekten Erkennung des entsprechenden Visems angibt. b stellt einen von Hand einzustellenden Schwellwert dar. In Abbildung 2.12 ist das Verfahren anhand der Phoneme /bI/, /dI/ und /ehI/ beispielhaft dargestellt.

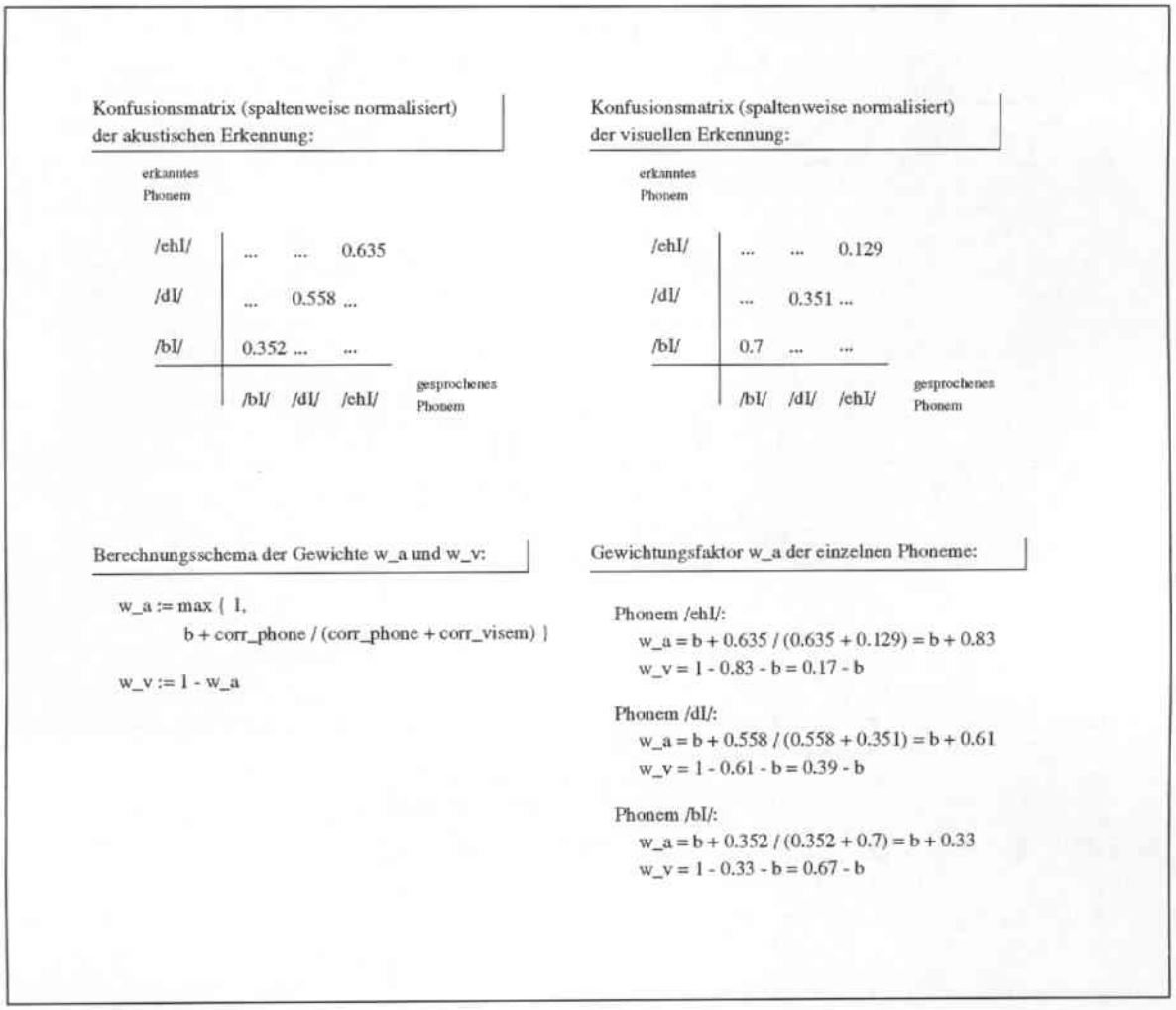


Abbildung 2.12: Beispiel für die Bestimmung der Gewichte w_{a_i} und w_{v_i} für ein Phonem p_i in Abhängigkeit von der Konfusionsmatrix. Es ist jeweils nur der zur Bestimmung wesentliche Teil der Matrix angegeben. Man sieht, daß beispielsweise das Phonem /ehI/, das visuell eine sehr hohe Verwechselbarkeit besitzt, akustisch jedoch wesentlich sicherer richtig erkannt wird, nun auch akustisch stärker gewichtet wird, während das Phonem /bI/, bei dem sich die Verwechselbarkeit gerade umgekehrt verhält, eine stärkere Gewichtung auf der visuellen Seite erfährt.

Als Daten zur Bestimmung der Konfusionsmatrizen wurden auf der akustischen Seite 1020 Sequenzen verwendet. Für die visuelle Konfusionsmatrix standen leider nur 170 Samples zur Verfügung. Die Ergebnisse für die Daten aus der Testdatenbasis aus Kapitel 2.1.1 mit Graustufenbildern als visuelle Eingabe und unterschiedlichen Werten für den Schwellwert b sind in Tabelle 2.8 dargestellt.
Da die Erkennungsrate, insbesondere für die verrauschten Daten, wesentlich unter den mit den bisherigen Kombinationsverfahren erzielten Ergebnissen liegt, wurde folgendes versucht: Die

Graustufen-bilder	rein akustisch	verschiedene Schwellwerte b							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
CLEAN (30 dB)	94.1%	82.4%	88.8%	95.3%	95.9%	**96.5%**	95.9%	95.9%	94.1%
NOISE 1 (16 dB)	67.6%	62.4%	65.9%	68.8%	**69.4%**	67.6%	67.6%		
NOISE 2 (8 dB)	49.4%	**51.8%**	**51.8%**	48.8%	50.0%	**51.8%**	48.2%		
INCREASE 1 (30-16 dB)	74.7%	69.4%	75.3%	78.8%	**79.4%**	77.6%	77.6%	75.9%	
INCREASE 2 (30-8 dB)	44.1%	59.4%	**61.8%**	59.4%	57.1%	51.8%			
RADIO 1 ($\approx$ 20 dB)	92.4%	74.1%	82.4%	90.0%	91.2%	90.6%	**92.3%**	91.8%	
RADIO 2 ($\approx$ 17 dB)	72.9%	65.3%	70.6%	75.9%	**77.1%**	76.5%	75.3%		
MOTOR 1 ($\approx$ 25 dB)	93.5%	77.6%	82.9%	91.2%	**93.5%**	92.9%	92.9%	92.9%	
MOTOR 2 ($\approx$ 11 dB)	50.6%	51.2%	52.4%	55.3%	**57.1%**	55.9%	52.4%		

Tabelle 2.8: Testergebnisse des ersten Ansatzes mit Graustufenbildern und phonemabhängigen Gewichten als visuelle Eingabe für unterschiedliche Schwellwerte b. Für die Werte der fehlenden Tabelleneinträge wurden keine Tests durchgeführt, da keine Ergebnisverbesserung zu erwarten war. Die maximalen Erkennungsraten der einzelnen Testdatenmengen sind jeweils hervorgehoben.

Graustufen-bilder	verschiedene Schwellwerte b									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CLEAN (30 dB)	82.4%	88.8%	95.3%	95.9%	**96.5%**	95.9%	95.9%	94.1%	94.1%	
NOISE 1 (16 dB)	47.1%	48.2%	55.3%	62.4%	70.6%	67.6%	**71.8%**	**71.8%**	69.4%	67.1%
NOISE 2 (8 dB)	42.9%	44.1%	47.1%	45.9%	45.3%	51.2%	**56.5%**	53.5%	53.5%	50.0%

Tabelle 2.9: Erkennungsraten bei phonemabhängiger Gewichtung und Graustufenbildern. Die Gewichte für die unterschiedlich verrauschten Daten wurden aus verschiedenen Konfusionsmatrizen berechnet. Diese Konfusionsmatrizen wurden mit Trainingsdaten erstellt, die jeweils genauso künstlich verrauscht wurden, wie die Daten, auf denen später getestet wurde. Die jeweils besten Ergebnisse sind hervorgehoben dargestellt.

1020 Sequenzen, die zur Erstellung der akustischen Konfusionsmatrix dienten, wurden mit ca. 16 dB und mit ca. 8 dB SNR künstlich verrauscht (analog zu den beiden Testsets NOISE 1 und NOISE 2). Auf den so erhaltenen verrauschten Datensets wurde jeweils eine Konfusionsmatrix erstellt. Anhand dieser wurden die Gewichte w_{a_i} und w_{v_i} nach obigem Verfahren für die beiden Verrauschungsstufen berechnet und auf den Testdatenmengen (NOISE 1 und NOISE 2, beide mit 30 Samples) getestet. Die Ergebnisse mit Graustufenbildern als visueller Eingabe befinden sich in Tabelle 2.9. Die Resultate, die mit dem optimalen, von Hand eingestellten Bias b (siehe Gleichung 2.8, Seite 32) erzielt wurden, sind in Abbildung 2.13 grafisch dargestellt.
Die Erkennungsraten bei Graustufenbildern liegen wesentlich unter den mit den bisherigen Verfahren erzielten Ergebnissen (vgl. Kapitel 2.1). Auffällig sind auch die hohen Schwellwerte, die bei verrauschten Daten erforderlich waren, um die beste Word-Accuracy zu erzielen. Hier macht sich wieder einmal die zu geringe Datenmenge, insbesondere für den visuellen Teil, bemerkbar. Während auf der akustischen Seite 1020 Datensequenzen zur Verfügung standen, um anhand der Konfusionsmatrizen die Gewichte w_{a_i} und w_{v_i} zu bestimmen, waren auf der visuellen Seite nur 170 Samples vorhanden[1]. Dies führte dazu, daß einige Phoneme in

[1] Die große Differenz der Anzahl der zur Verfügung stehenden Daten liegt darin begründet, daß zwar genauso viele visuelle wie akustische Daten vorlagen, die visuellen jedoch mit unterschiedlichen Aufnahmeverfahren (insbesondere mit bzw. ohne *Face Tracker*, vgl. [4, 9]) aufgenommen wurden. Die Verwendung dieser „verschiedenen" visuellen Daten war somit nicht möglich.

LDA-Daten	verschiedene Schwellwerte b									
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CLEAN (30 dB)	92.4%	94.7%	**96.5%**	**96.5%**	**96.5%**	95.9%	95.9%	94.1%	94.1%	
NOISE 1 (16 dB)	75.9%	75.3%	73.5%	**78.2%**	73.5%	70.6%	67.1%	67.1%	67.6%	67.6%
NOISE 2 (8 dB)	64.7%	**67.6%**	65.3%	61.8%	57.1%	52.9%	50.6%	48.8%	49.4%	49.4%

Tabelle 2.10: Erkennungsraten bei phonemabhängiger Gewichtung und LDA-Daten. Die Gewichte für die unterschiedlich verrauschten Daten wurden aus verschiedenen Konfusionsmatrizen berechnet. Die besten Ergebnisse sind wieder hervorgehoben dargestellt.

den Datenbasen, die zur Erstellung der Konfusionsmatrizen benutzt wurden, nur äußerst selten vorkamen. Dementsprechend sind die Werte, die als Basis für die phonemabhängige Gewichtung dienen sollen, mitunter sehr unzuverlässig. Bestätigt wird diese Vermutung durch eine genauere Betrachtung der falsch erkannten Worte:
Bei unverrauschten Daten wurden sechs Worte falsch klassifiziert. Dabei wurde viermal der Buchstabe „b" statt dem Buchstaben „w" erkannt. Die Anzahl der Vorkommen in der Trainingsmenge zur Berechnung der Konfusionsmatrizen war für die einzelnen Phoneme, aus denen sich diese Buchstaben zusammensetzen, wie folgt:

- Buchstabe „b":

  akustisch: /bI/: 1027, /b-eh/: 2255, /ehF/: 25963
  visuell: /_b/: 928, /_b-eh/:630, /_ehF/:4571

- Buchstabe „w":

  akustisch: /vI/: 315, /v-eh/: 667, /ehF/: 25963
  visuell: /_f/: 339, /_f-eh/: 98, /_ehF/:4571

Das auffallend niedrige Vorkommen der Phoneme für den Buchstaben „w" legt nahe, daß die entsprechenden Werte der Konfusionsmatrizen zur Ermittlung der Gewichte relativ unzuverlässig sind, was die häufige Fehlklassifikation dieses Buchstabens erklären könnte.
Ferner sollte bei der Interpretation der Ergebnisse beachtet werden, daß die 170 Buchstabensequenzen zur Bestimmung der visuellen Konfusionsmatrix den Daten entsprechen, mit denen das TDNN trainiert wurde. Geschickter wäre es hier, als Basis für die Konfusionsmatrix andere, zur Trainingsmenge disjunkte Samples zu verwenden.

Die Ergebnisse mit LDA-Daten als visuelle Eingabe sind in Tabelle 2.10 zu finden. Eine grafische Darstellung der besten erzielten Erkennungsraten ist in Abbildung 2.14 dargestellt. Der Mangel an Trainingsdaten macht sich bei Verwendung von LDA-Daten für die mit ca. 16 dB SNR verrauschten Daten nicht so stark bemerkbar. Die Ergebnisse entsprechen hier in etwa den bisher erzielten Erkennungsraten. Bei den stärker verrauschten Daten (NOISE 2, ca. 8 dB SNR) sinkt die Erkennungsleistung jedoch wieder deutlich.

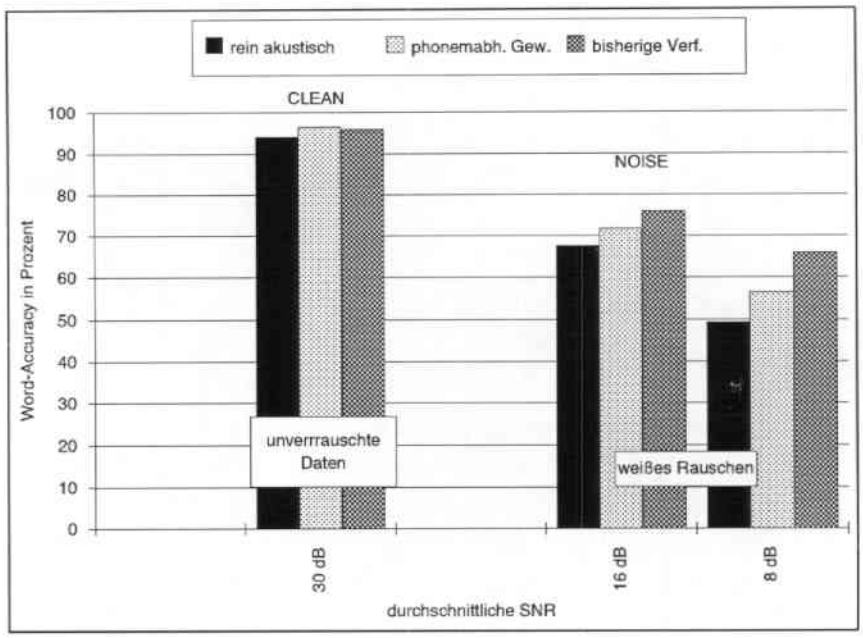


Abbildung 2.13: Testergebnisse bei Verwendung von Graustufenbildern als visuelle Eingabe und phonemabhängigen Gewichten, die über die jeweiligen Konfusionsmatrizen bestimmt wurden. Dargestellt sind nur die mit dem besten, von Hand eingestellten Bias (Wert b aus Gleichung 2.8, Seite 32) erzielten Ergebnisse. Zum Vergleich ist auch die Word-Accuracy aufgetragen, die in etwa mit den bisherigen Verfahren erzielt wurde. Da sich kein Verfahren besonders hervorgehoben hat, sind hier nur die Erkennungsraten angegeben, die ungefähr mit den besten Kombinationsalternativen erzielt wurden.

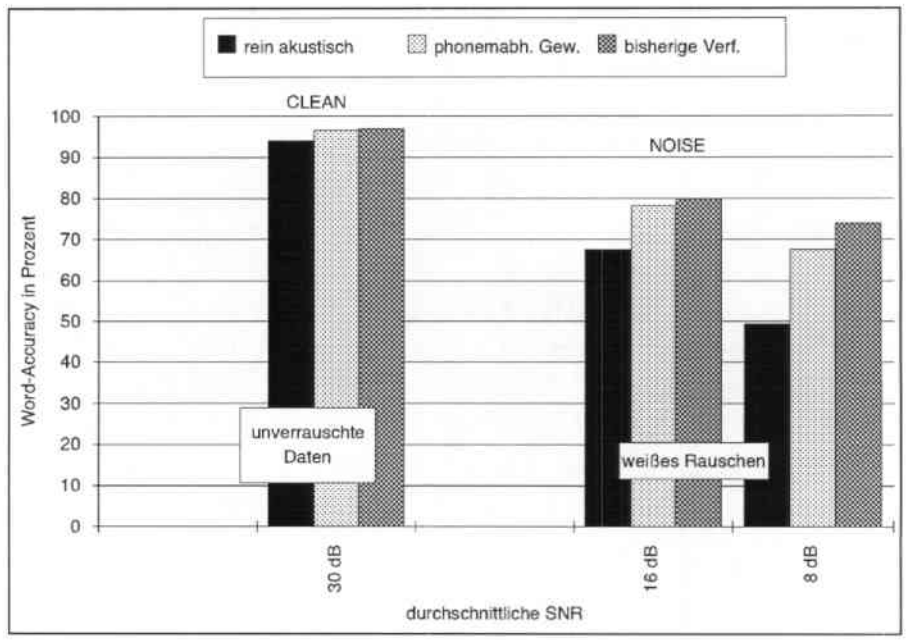


Abbildung 2.14: Testergebnisse bei Verwendung von LDA-Daten als visuelle Eingabe und phonemabhängigen Gewichten, die über die jeweiligen Konfusionsmatrizen bestimmt wurden. Dargestellt sind nur die mit dem besten, von Hand eingestellten Bias (Wert b aus Gleichung 2.8, Seite 32) erzielten Ergebnisse. Zum Vergleich wurde wieder die in etwa mit den bisherigen Verfahren erzielte Word-Accuracy ebenfalls aufgetragen.

Da im Laufe dieser Arbeit keine zusätzlichen Daten zur Verfügung gestellt werden konnten, wurde dieser Ansatz nicht weiter verfolgt. Auf einer größeren Datenmenge und bei guten Ergebnissen mit dem zuvor durchgeführten Verfahren wäre folgendes weitere Vorgehen naheliegend: die verschiedenen Gewichte für die einzelnen Phoneme werden wie oben beschrieben für unverrauschte, sowie für leicht und stark verrauschte Daten (z.B. Daten mit einer SNR von 30, 16 und 8 dB) anhand der entsprechenden Konfusionsmatrizen bestimmt. Bei der Erkennung wird dann, wie bei den Verfahren in Kapitel 2.2, auf den akustischen Daten die jeweilige SNR berechnet. Die Gewichte für die Kombination werden schließlich anhand der SNR durch lineare Interpolation zwischen den für die SNR-Werte 30, 16 und 8 dB berechneten Gewichten bestimmt.

Kapitel 3

Kombination auf Hiddenebene

Die im vorherigen Kapitel besprochene Kombination auf phonetischer Ebene besitzt neben den bereits erwähnten Vorteilen gegenüber anderen Netzarchitekturen auch diverse Nachteile. Durch die Tatsache, daß bestimmte Phoneme visuell nicht unterschieden werden können, war es erforderlich, sich Viseme zu definieren. Anhand einer Phonem-Visem-Tabelle wurde dann den einzelnen Phonemen jeweils genau ein Visem zuordnet. Das Problem der Definition solcher Viseme ist jedoch nicht trivial und bildet somit eine zusätzliche potentielle Fehlerquelle. Für weitere Anmerkungen zum Problem der Visemdefinition siehe [17].
Desweiteren gibt es Untersuchungen [1, 19], die vermuten lassen, daß der Mensch, der ja ebenfalls sowohl akustische als auch visuelle Information zum Erkennen von Sprache nutzt, die verschiedenen Signale bereits auf einer „tieferen Ebene" kombiniert. Das heißt, man nimmt an, daß der Mensch die beiden Eingabequellen verschmelzt und dann erst eine Klassifikation durchführt. Bei der in Kapitel 2 beschriebenen Vorgehensweise der Kombination auf Phonemebene wird dagegen erst eine Klassifikation und anschließend eine Verschmelzung der unterschiedlichen Eingabemodalitäten durchgeführt. Dieser Ansatz entspricht somit nicht dem „natürlichen" Vorbild.
Eine Trennung von akustischer und visueller Erkennung in zwei voneinander unabhängige Teilnetze verhindert ferner ein Erlernen von unter Umständen bestehenden Korrelationen zwischen dem akustischen Signal auf der einen und den Lippenbewegungen auf der anderen Seite.
In diesem Kapitel soll nun versucht werden, diese Probleme durch die Verwendung einer modifizierten Architektur des zugrundeliegenden TDNNs zu umgehen. In Kapitel 3.1 wird ein einfacher Ansatz mit der bereits in Kapitel 1.3 beschriebenen Netzarchitektur versucht. In Kapitel 3.2 wird ein weiterer Ansatz untersucht, bei dem zusätzlich zu den akustischen und visuellen Signalen auch noch die SNR als Netzeingabe verwendet wird.

3.1 Kombination ohne zusätzliche Information

Anstelle einer getrennten Erkennung von Phonemen und Visemen mit einer anschließenden Kombination der Ergebnisse wurde hier der Versuch unternommen, die Fusion der beiden Eingabemodalitäten auf einem tieferen Level, genauer auf der Hidden-Ebene des TDNNs, durchzuführen (vgl. Kapitel 1.3). Die Abbildung der entsprechend modifizierten Architektur des Erkenners ist in Abbildung 3.1 noch einmal dargestellt.
Mit dem Ziel, die Erkennungsleistung des Netzes zu verbessern, wurde zunächst versucht,

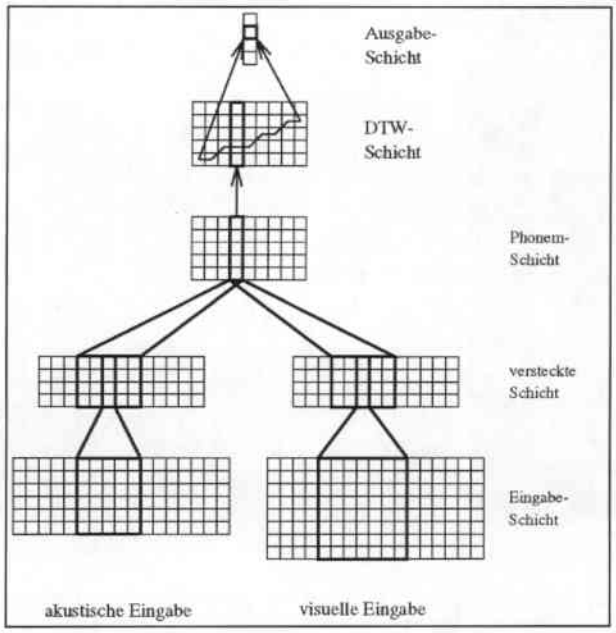


Abbildung 3.1: Netzarchitektur für die Kombination auf der Ebene der versteckten Einheiten.

Word-Accuracy in % auf CLEAN mit	Anzahl der visuellen Hidden-Units										
	5	9	10	11	12	13	14	15	16	20	25
Graustufenbildern	95.3	95.9	98.8	95.3	94.3	96.5	95.9	97.1	95.3	94.7	92.3
LDA-Daten	97.6		95.3					93.5		95.9	93.5

Tabelle 3.1: Testergebnisse bei Kombination auf Hidden-Ebene bei einer unterschiedlichen Anzahl von Hidden-Units. Angegeben ist jeweils die auf der Testmenge CLEAN (siehe Seite 16) erzielte Word-Accuracy in Prozent. Die Tests der fehlenden Tabelleneinträge bei Verwendung von LDA-Daten wurden nicht durchgeführt.

durch Variation der Anzahl der Hidden-Units des visuellen Teilnetzes deren optimale Anzahl zu bestimmen[1]. Das Netz wurde, wie auch die Netze bei der Kombination auf Phonemebene, mit einer unverrauschten Datenmenge (mum1&2, 170 Sequenzen) trainiert und anschließend auf den entsprechenden 30 Samples aus der Testdatenbank (CLEAN, vgl. Seite 16) getestet. Die verschiedenen Erkennungsraten in Abhängigkeit von der Anzahl der Hidden-Units sind in Tabelle 3.1 aufgeführt. Eine grafische Darstellung der Ergebnisse befindet sich in den Abbildungen 3.2 und 3.3.

[1] Die Anzahl der Hidden-Units im akustischen Teilnetz wurde bereits im Rahmen der Konstruktion des rein akustischen Erkenners durchgeführt. Aufgrund der dort erzielten Ergebnisse wurden bisher bei der kombinierten Erkennung sowohl auf der akustischen, als auch auf der visuellen Seite fünfzehn Hidden-Units verwendet.

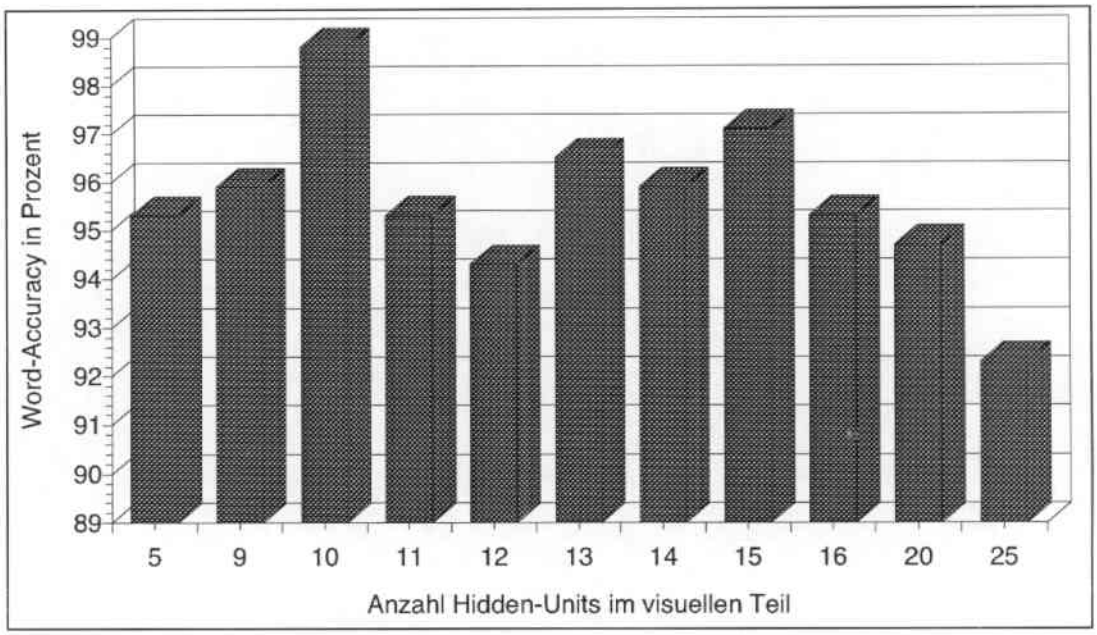


Abbildung 3.2: Testergebnisse bei unterschiedlicher Anzahl von Hidden-Units in der visuellen Teilschicht bei Graustufenbildern als visuelle Eingabe.

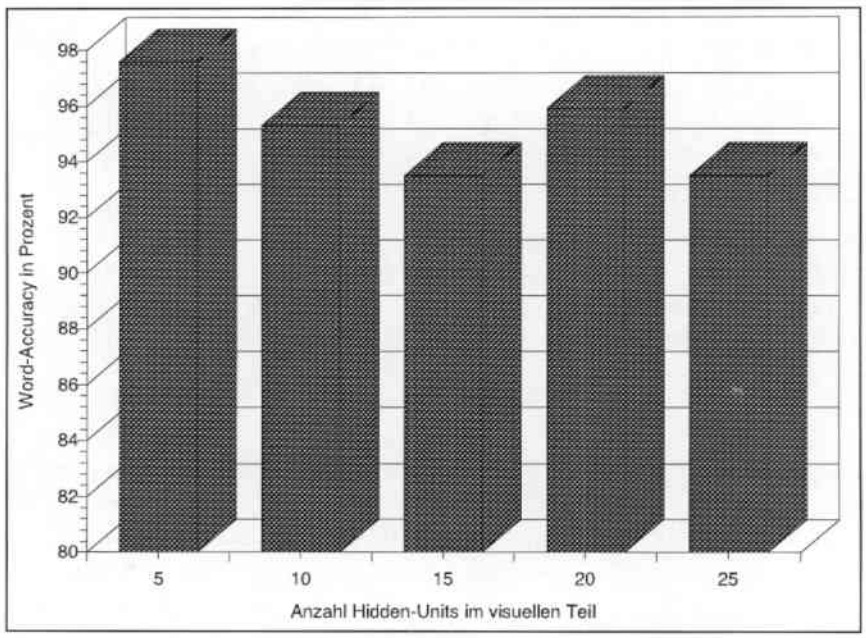


Abbildung 3.3: Testergebnisse bei unterschiedlicher Anzahl von Hidden-Units in der visuellen Teilschicht bei LDA-Daten als visuelle Eingabe.

Auffallend an den Ergebnisse, die mit Graustufenbildern als visuelle Eingabe erzielt wurden, ist, daß mit zehn und fünfzehn Hidden-Units die besten Erkennungsraten erreicht wurden, wohingegen alle Tests mit einer Hidden-Unit-Anzahl zwischen zehn und fünfzehn schlechtere Ergebnisse lieferten. Dies ist wieder einmal durch den geringen Umfang der Trainings- und Testdatenmengen zu erklären. Aus diesem Grund wurde für die LDA-Daten keine feinere

Graustufen-bilder	rein akustische Erkennung	Hidden-Kombination mit 10 Hidden-Units	15 Hidden-Units	ungefähre Ergebnisse bei Phonem-Kombin.
CLEAN (30 dB)	94.1%	98.8%	97.1%	≈96%
NOISE 1 (16 dB)	67.6%	77.1%	78.2%	≈76%
NOISE 2 (8 dB)	49.4%	52.9%	58.2%	≈66%
INCREASE 1 (30-16 dB)	74.7%	86.5%	84.1%	≈86%
INCREASE 2 (30-8 dB)	44.1%	50.6%	55.3%	≈70%
RADIO 1 (≈ 20 dB)	92.4%	88.8%	93.5%	≈93%
RADIO 2 (≈ 17 dB)	72.9%	58.8%	75.9%	≈80%
MOTOR 1 (≈ 25 dB)	93.5%	94.7%	94.7%	≈95%
MOTOR 2 (≈ 11 dB)	50.6%	44.7%	60.6%	≈70%

LDA-Daten	rein akustische Erkennung	Hidden-Kombination mit 5 Hidden-Units	10 Hidden-Units	ungefähre Ergebnisse bei Phonem-Kombin.
CLEAN (30 dB)	94.1%	97.6%	95.3%	≈97%
NOISE 1 (16 dB)	67.6%	76.5%	78.2%	≈80%
NOISE 2 (8 dB)	49.4%	44.7%	58.2%	≈74%
INCREASE 1 (30-16 dB)	74.7%	85.9%	85.9%	≈87%
INCREASE 2 (30-8 dB)	44.1%	60.0%	70.6%	≈76%
RADIO 1 (≈ 20 dB)	92.4%	93.5%	91.8%	≈95%
RADIO 2 (≈ 17 dB)	72.9%	71.2%	87.6%	≈87%
MOTOR 1 (≈ 25 dB)	93.5%	97.6%	94.7%	≈96%
MOTOR 2 (≈ 11 dB)	50.6%	57.6%	64.7%	≈78%

Tabelle 3.2: Testergebnisse bei der Kombination auf Hidden-Ebene. Die Netze wurden mit einer unverrauschten Datenmenge trainiert und mit der auf Seite 16 beschriebenen Datenbasis getestet. Angegeben ist die Word-Accuracy in Prozent. Um die Ergebnisse mit den jeweiligen Erkennungsraten, die bei der Kombination auf Phonem-Ebene erzielt wurden, vergleichen zu können, wurde eine Spalte angefügt, in der die Ergebnisse stehen, die in etwa mit den Verfahren aus Kapitel 2 erreicht wurden. Da sich kein Verfahren besonders hervorgehoben hat, sind hier nur die Erkennungsraten angegeben, die ungefähr mit den besten Kombinationsalternativen erzielt wurden.

Optimierung durchgeführt.

Trotz allem läßt sich jedoch ein gewisser Trend erkennen: die besten Ergebnisse bei der Verwendung von Graustufenbildern wurden mit ca. zehn bis fünfzehn Hidden-Units erzielt, bei der Eingabe von LDA-Daten erreichte man eine maximale Erkennungsrate bei ungefähr fünf bis zehn Hidden-Units. Dies ist verständlich, wenn man sich die unterschiedliche Größe der beiden Eingabevektoren vor Augen hält: den 384 Parameterwerten der Graustufenbilder stehen die 32 Parameter der LDA-Daten gegenüber.

Die optimierten Netze wurden nun mit der auf Seite 16 beschriebenen Datenbasis, die ja auch zum Testen der Verfahren im letzten Kapitel verwendet wurde, getestet. Die erzielten Erkennungsraten befinden sich in Tabelle 3.2. In Abbildung 3.4 sind die Testergebnisse bei Verwendung von Graustufenbildern grafisch dargestellt, Abbildung 3.5 zeigt die entsprechenden Ergebnisse beim Gebrauch von LDA-Daten.

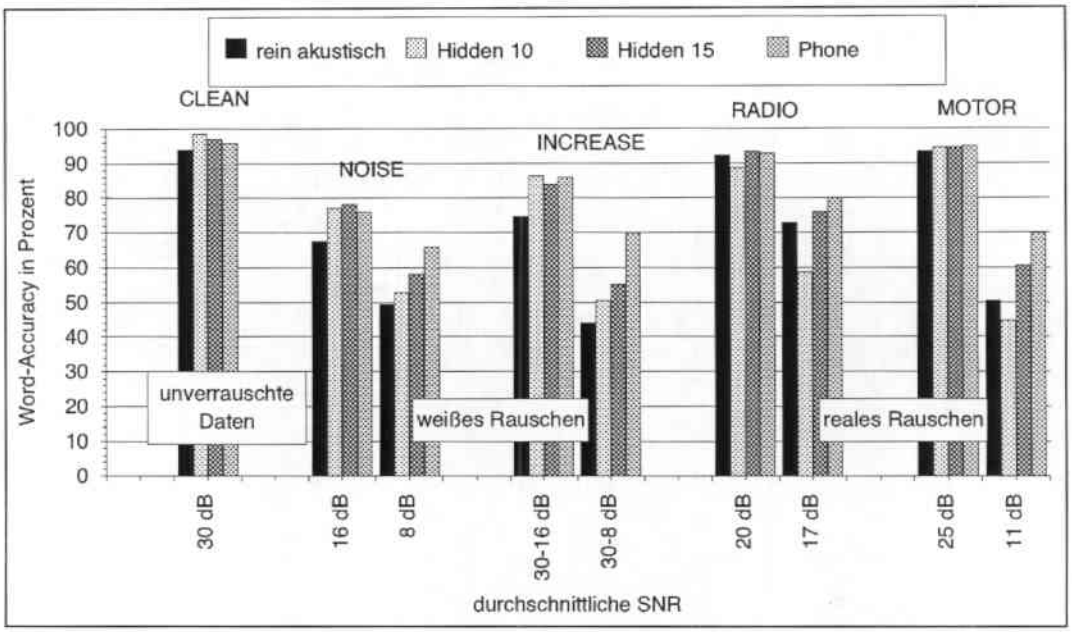


Abbildung 3.4: Testergebnisse bei der Kombination auf Hidden-Ebene nach dem Training mit einem unverrauschten Datenset. Als visuelle Eingabe wurden Graustufenbilder verwendet. „Hidden 10" stellt die Ergebnisse bei Verwendung von zehn Hidden-Units, „Hidden 15" die bei fünfzehn Hidden-Units dar. „Phone" entspricht den ungefähren Ergebnissen, die bei Kombination auf Phonemebene erzielt wurden.

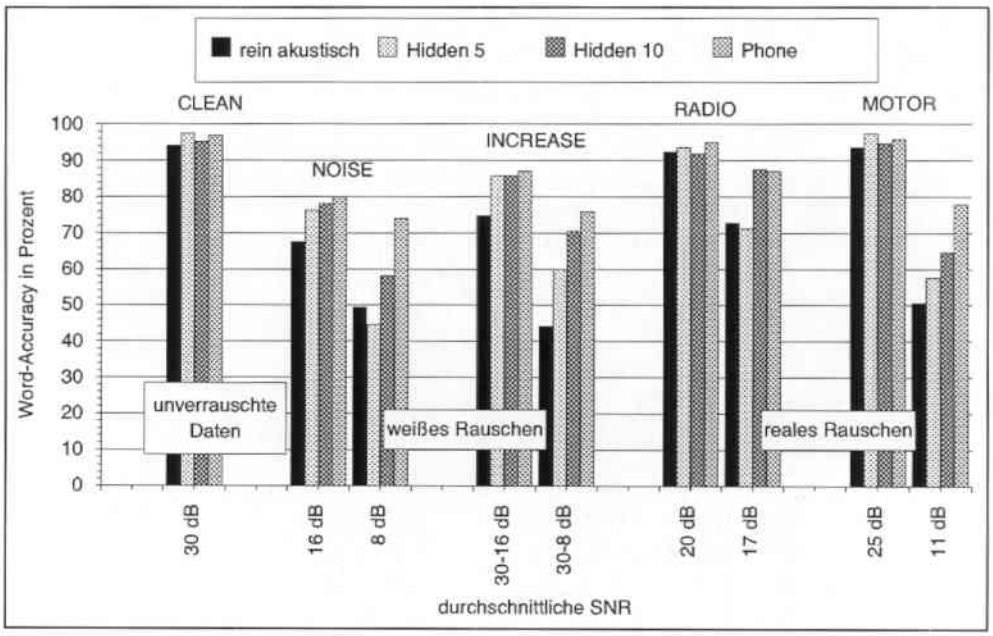


Abbildung 3.5: Testergebnisse bei der Kombination auf Hidden-Ebene nach dem Training mit einem unverrauschten Datenset. Als visuelle Eingabe wurden LDA-Daten verwendet. „Hidden 5" stellt die Ergebnisse bei Verwendung von fünf Hidden-Units, „Hidden 10" die bei zehn Hidden-Units dar. „Phone" entspricht den ungefähren Ergebnissen, die bei Kombination auf Phonemebene erzielt wurden.

Obwohl bei der Optimierung der Anzahl der Hidden-Units mit weniger Hidden-Units (10 bei Graustufenbildern, 5 bei LDA-Daten) auf sauberen Daten bessere Ergebnisse erzielt wurden,

als bei den Netzen mit mehr Hidden-Units (15 bei Graustufenbildern, 10 bei LDA-Daten), brachten die Netze mit einer höheren Anzahl von Hidden-Units beim Test auf verrauschten Daten mitunter eine beträchtliche Steigerung in der Erkennungsrate.
Im Vergleich zu den Erkennungsraten, die bei der Kombination auf Phonemebene erzielt wurden, fällt auf, daß die Ergebnisse bei weniger stark verrauschten Daten in etwa gleich gut sind, während die Erkennungsleistung bei den Daten mit starken Störgeräuschen (NOISE 2 (8 dB), MOTOR 2 ($\approx$ 11 dB) und INCREASE 2 (30-8 dB)) erheblich sinkt. Dieser Rückgang in der Erkennungsrate ist dadurch zu erklären, daß das Netz, das ja nur auf sauberen Daten trainiert wurde, Probleme bei der Erkennung stark verrauschter Daten bekommt, da sie wohl zu sehr von den Daten der Trainingsmenge abweichen, während es bei leicht verrauschten Daten hingegen noch gut generalisierungsfähig ist. Eine Möglichkeit, nachträglich in den Erkennungsprozeß einzugreifen und eine datenabhängige Gewichtung der beiden Eingabemodalitäten vorzunehmen, wie sie ja bei der Kombination auf phonetischer Ebene existiert, ist bei dieser Architektur nicht gegeben. Aus diesem Grund wurde im folgenden ein weiterer Ansatz untersucht, bei dem das TDNN zusätzliche Information über die Qualität der akustischen Eingabedaten erhält und damit eine angepaßte Gewichtung vornehmen kann.

3.2 Kombination mit der SNR als zusätzliche Eingabe

Die Netzarchitektur aus Kapitel 3.1 wurde im folgenden um ein zusätzliches Modul erweitert (siehe Abbildung 3.6), das die SNR-Werte der jeweiligen akustischen Daten als Eingabe erhält.

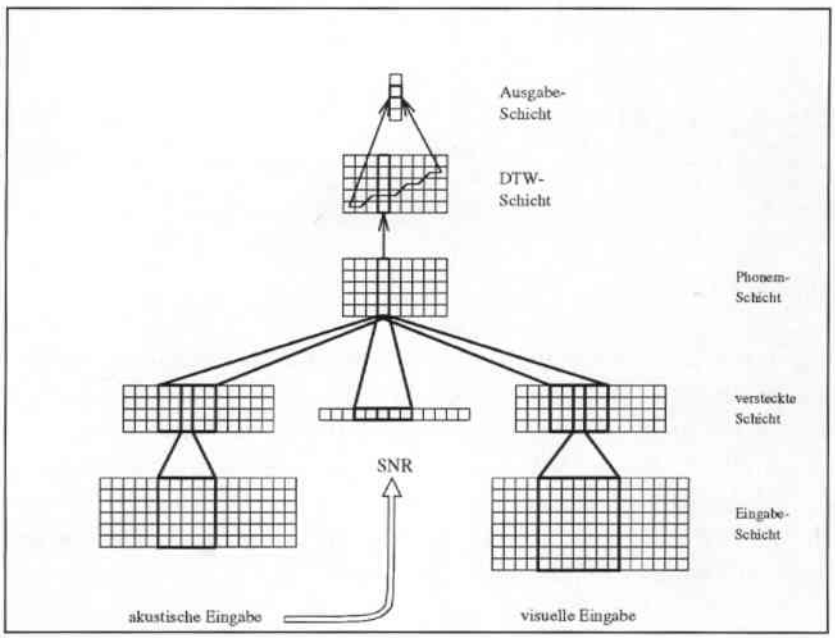


Abbildung 3.6: Netzarchitektur für die Kombination auf der versteckten Schicht mit einer Erweiterung der Hidden-Ebene um zusätzliche Einheiten zur Eingabe und Verarbeitung der SNR-Werte, die aus den akustischen Eingabedaten berechnet werden.

Zur Berechnung dieser Werte wird wieder der bereits in Kapitel 2.2 erwähnte Algorithmus von H.G. Hirsch verwendet. Die hier verwendete Implementierung des Algorithmus liefert alle 500 ms einen über 100 Zeitframes gemittelten SNR-Wert. Zwischen zwei vom Algorithmus

berechneten Werten wird eine lineare Interpolation (vgl. Kapitel 2.2, Abbildung 2.6) durchgeführt, so daß zu allen akustischen Werten, die ja im 10 ms Takt vorliegen, ein zugehöriger SNR-Wert existiert. Diese SNR-Werte werden durch folgende Funktion auf das Intervall [0, 1] abgebildet:

$$snr_{norm}(SNR) = \begin{cases} 1 & \text{falls } SNR > 35 \\ 0 & \text{falls } SNR < 0 \\ \frac{1}{35} snr & \text{sonst} \end{cases} . \tag{3.1}$$

Die Werte für die „Eckdaten" ($SNR < 0$ und $SNR > 35$) wurden heuristisch aufgrund der Ergebnisse im vorangegangenen Kapitel bestimmt. Eine grafische Darstellung dieser Funktion befindet sich in Abbildung 3.7. Anschließend werden diese Werte in die entsprechende Schicht (siehe Abbildung 3.6) eingetragen. Die Werte dieser Schicht gehen ebenfalls in das Training und in die Testphase mit ein, so daß das Netz bei der Kombination der beiden Eingabemodalitäten Ton und Bild auch noch zusätzliche Information erhält, die die Güte der akustischen Daten jeweils eines Frames angibt. In Anbetracht der Ergebnisse aus Kapitel 3.1 wurden im folgenden auf der visuellen Seite, sowohl bei Verwendung von Graustufenbildern, als auch bei LDA-Daten, fünfzehn Hidden-Units verwendet.

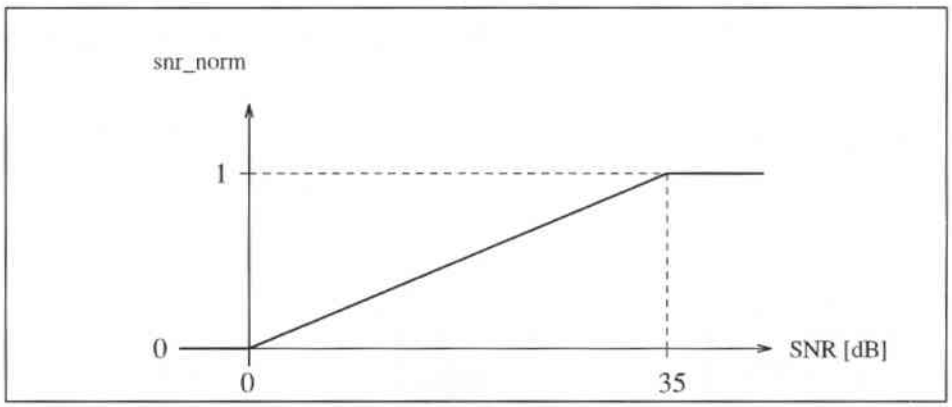


Abbildung 3.7: Grafische Darstellung der Funktion, die einen gegebenen SNR-Wert eines akustischen Signals auf das Intervall [0, 1] abbildet.

Ziel ist es nun, daß das TDNN seine Gewichte in der Trainingsphase durch die Zusatzinformation, die es durch die Eingabe der SNR-Werte erhält, derart einstellt, daß eine bessere Erkennung, insbesondere der stark verrauschten Daten, erfolgen kann. Um dem Netz das Erlernen einer „SNR-abhängigen" Gewichtung zu ermöglichen, ist es allerdings erforderlich, das Netz nicht nur wie bisher mit sauberen Daten, sondern auch mit verrauschten Daten zu trainieren. Die Datenbasis, mit der das Netz trainiert wurde, beinhaltet zum einen 60 der 170 unverrauschten Sequenzen, die bisher zum Training verwendet wurden. Ferner wurden diese 60 Sequenzen mit unterschiedlichen Rauscharten künstlich verrauscht. Diese künstliche Verrauschung geschah, wie auch beim Anlegen der Testdatenbasis (vgl. Kapitel 2.1.1, Seite 16), additiv. Im einzelnen wurden folgende Rauscharten und -stärken verwendet: NOISE 1 und 2 (16 bzw. 8 dB), RADIO 1 und 2 (≈20 bzw. 17 dB), sowie MOTOR 1 bzw. 2 (≈25 und 11 dB). Die gesammte Trainingsmenge umfaßte somit 420 Samples, davon 60 unverrauschte und 350 verrauschte.

Um die so erzielten Ergebnisse richtig einschätzen zu können, wurde der rein akustische Erkenner, der ja bisher nur mit rauschfreien Daten trainiert wurde, mit dem gleichen (ver-

Graustufen-bilder, verrauschte Trainingsmenge	rein akustische Erkennung	Hidden-kombination ohne SNR	Hidden-kombination mit SNR
CLEAN (30 dB)	95.3%	97.1%	98.2%
NOISE 1 (16 dB)	92.9%	94.1%	96.5%
NOISE 2 (8 dB)	92.9%	89.4%	92.4%
INCREASE 1 (0-16 dB)	80.0%	94.7%	92.4%
INCREASE 2 (0-8 dB)	54.7%	78.2%	62.9%
RADIO 1 ($\approx$ 20 dB)	95.9%	94.1%	98.2%
RADIO 2 ($\approx$ 17 dB)	94.1%	90.6%	97.1%
MOTOR 1 ($\approx$ 11 dB)	95.3%	94.7%	98.2%
MOTOR 2 ($\approx$ 25 dB)	91.8%	91.8%	92.4%

LDA-Daten, verrauschte Trainingsmenge	rein akustische Erkennung	Hidden-kombination ohne SNR	Hidden-kombination mit SNR
CLEAN (30 dB)	95.3%	93.5%	94.7%
NOISE 1 (16 dB)	92.9%	92.9%	93.5%
NOISE 2 (8 dB)	92.9%	91.2%	91.2%
INCREASE 1 (30-16 dB)	80.0%	92.4%	95.3%
INCREASE 2 (30-8 dB)	54.7%	81.2%	79.4%
RADIO 1 ($\approx$ 20 dB)	95.9%	95.3%	92.9%
RADIO 2 ($\approx$ 17 dB)	94.1%	92.4%	94.1%
MOTOR 1 ($\approx$ 25 dB)	95.3%	94.1%	96.5%
MOTOR 2 ($\approx$ 11 dB)	91.8%	91.8%	92.4%

Tabelle 3.3: Testergebnisse bei der Kombination auf Hidden-Ebene mit der SNR als zusätzlicher Eingabe. Die Netze wurden mit einer Datenmenge trainiert, die sich zusammensetzte aus sauberen Daten und Daten, die mit dem gleichem Rauschen wie NOISE 1 und 2, RADIO 1 und 2, sowie MOTOR 1 und 2 künstlich verrauscht wurden.

rauschten) Datenset trainiert wie die Architektur, die die SNR als zusätzliche Eingabe erhält. Die Ergebnisse stehen in der ersten Spalte von Tabelle 4.2. Ferner wurde die Architektur aus Kapitel 3.1 (siehe Abbildung 3.1) ebenfalls auf diesem Datenset trainiert, um eine bessere Einschätzung davon zu bekommen, wie groß der Einfluß der SNR auf die Erkennungsrate ist. Die Ergebnisse sind in der zweiten Spalte der Tabelle dargestellt. Die dritte Spalte dieser Tabelle enthält die jeweilige Word-Accuracy, die mit der Architektur mit Kombination auf Hidden-Ebene und zusätzlicher Eingabe der SNR erzielt wurde. Abbildung 3.8 bzw. 3.9 enthält eine grafische Darstellung der erreichten Erkennungsraten bei Verwendung von Graustufenbilder bzw. LDA-Daten.

Auffallend sind die überraschend guten Ergebnisse bei der rein akustischen Erkennung. Auf allen Testdatensets, die mit einem Störgeräusch verrauscht wurden, das auch in der Trainingsmenge enthalten war, wurde eine Word-Accuracy von über 90% erreicht. Offensichtlich ist das Netz gut in der Lage, bei entsprechender Erweiterung der Trainingsdatenmenge, auch auf unterschiedlich verrauschten Daten gut zu generalisieren. Die entsprechenden Ergebnisse bei der kombinierten Erkennung sind nur unwesentlich besser. Auch die Verwendung der SNR als zusätzliche Netzeingabe brachte gegenüber der Architektur, die diese Zusatzinformation nicht nutzt, keine großen Verbesserungen.

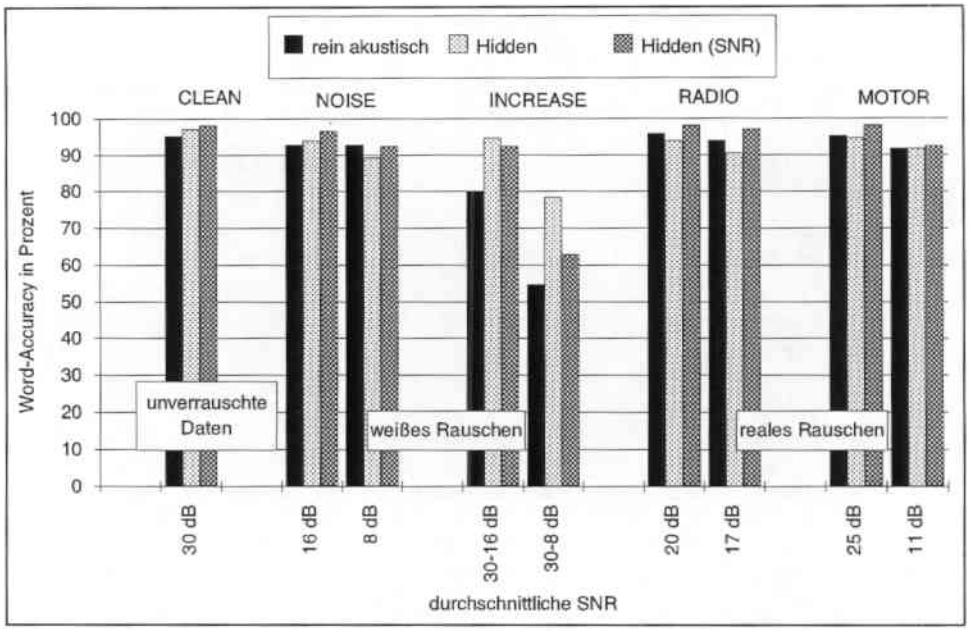


Abbildung 3.8: Grafische Darstellung der Testergebnisse bei der Kombination auf Hidden-Ebene und Graustufenbildern als visuelle Eingabedaten. Die Netze wurden jeweils auf verrauschten Daten trainiert. „Hidden" zeigt die Ergebnisse an, die mit der Architektur ohne zusätzliche Eingabe der SNR erzielt wurden, „Hidden (SNR)" die Ergebnisse bei Verwendung der SNR als zusätzliche Eingabe.

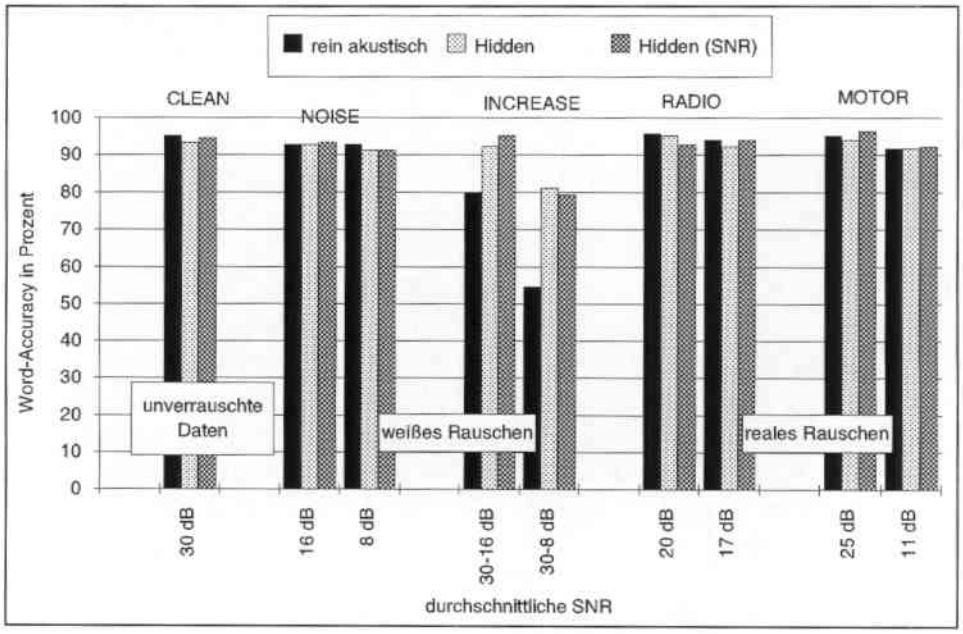


Abbildung 3.9: Grafische Darstellung der Testergebnisse bei der Kombination auf Hidden-Ebene und LDA-Daten als visuelle Eingabedaten. Die Netze wurden jeweils auf verrauschten Daten trainiert. „Hidden" zeigt die Ergebnisse an, die mit der Architektur ohne zusätzliche Eingabe der SNR erzielt wurden, „Hidden (SNR)" die Ergebnisse bei Verwendung der SNR als zusätzliche Eingabe.

Interessant sind nun die Ergebnisse der Datenmengen INCREASE 1 und 2, die ein Rauschen enthalten, das nicht in der Trainingsmenge vorkam. Wie bereits in Kapitel 2.1.1 auf Seite 19

erwähnt, kommt es durch das ansteigende Hintergrundrauschen hier zu Fehlklassifikationen des akustischen Erkenners. Diese wirken sich auch auf die Erkennungsraten bei einem verrauschten Trainingsset aus. Bei der bimodalen Erkennung kann hier jedoch der visuelle Teil des Netzes dazu beitragen, dieser Fehlinterpretation entgegenzuwirken und somit eine wesentlich bessere Word-Accuracy zu erzielen.

Ein weiterer Schritt wäre nun, zusätzliche Tests auf Daten mit Rauschanteilen, die nicht in der Trainingsmenge vorkamen, durchzuführen. Der rein akustische Erkenner liefert gute Ergebnisse, wenn das MS-TDNN auf den gleichen Rauscharten trainiert wurde wie die Testdaten. Es stellt sich jetzt die Frage, inwieweit das Netz in der Lage ist, auf anderen Arten von Störgeräuschen zu generalisieren. Die Tests haben gezeigt, das die Erkennungsleistung bei Daten mit ansteigendem Rauschen relativ stark sinkt. Sollte die Generalisierungsfähigkeit des Erkenners bei unterschiedlichem Hintergrundrauschen nachlassen, wäre es durchaus möglich, wie bei den Ergebnissen der Tests mit den INCREASE 1 und INCREASE 2 Datensets gesehen, daß man durch die zusätzliche visuelle Information ein zu starkes Absinken der Word-Accuracy verhindern kann. Aufgrund der zeitlichen Begrenzung konnten diese weiteren Untersuchungen im Rahmen dieser Arbeit leider nicht durchgeführt werden.

Kapitel 4

Kombination auf Inputebene

Eine weitere Alternative zur Kombination der beiden Modalitäten Sprach- und Bildsignal ist, wie bereits in Kapitel 1.3 erwähnt, die Kombination auf der Eingabe-Ebene. Bei diesem Fall der Sensorfusion geht man von einem einzigen TDNN aus, das als Eingabedaten sowohl das akustische als auch das visuelle Signal erhält. Die entsprechende Architektur ist in Abbildung 4.1 noch einmal dargestellt.

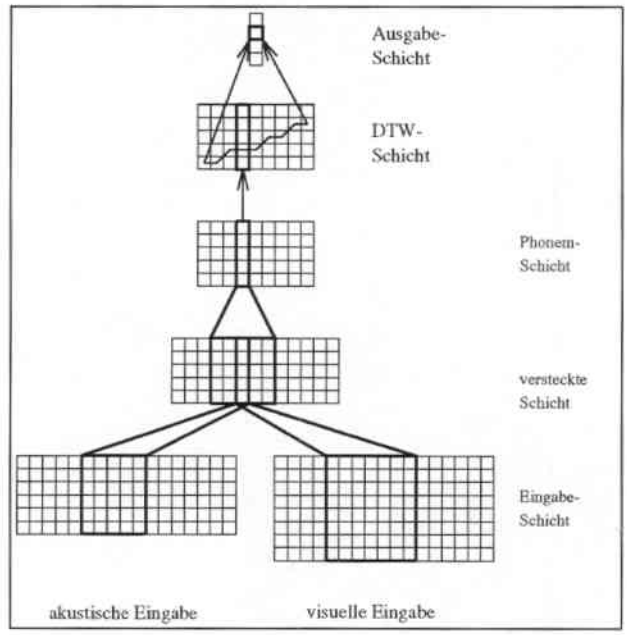


Abbildung 4.1: Netzarchitektur für die Kombination auf der Ebene der Eingabeeinheiten.

Diese Art der Kombination besitzt gegenüber einer Verschmelzung auf phonetischer Ebene die gleichen Vorteile wie die in Kapitel 3 verwendete Architektur (vgl. Seite 38). Es ergibt sich hier jedoch eine zusätzlich zu berücksichtigende Besonderheit: da hier die akustischen und visuellen Merkmale auf der gleichen Ebene direkt in die Netzarchitektur eingegeben werden, könnte der Fall eintreten, daß die 384 Parameter der Graustufenbilder die 16 Parameter der akustischen Daten derart dominieren, daß unter Umständen Probleme beim Training und bei der Erkennung auftreten (für weitere Ausführungen zu dieser Problematik siehe [11]). Aus diesem Grund wurden in Kapitel 4.2 nur Tests mit LDA-Daten als visuelle Eingabe

unternommen.

Desweiteren wurde nur eine geringe Anzahl der möglichen Ansätze untersucht, da ein ausführlicheres Vorgehen den zeitlichen Rahmen dieser Arbeit gesprengt hätte. Zunächst wurde in Kapitel 4.1 das Netz auf unverrauschten Daten trainiert und auf der Testdatenbasis getestet. Anschließend wurde in Kapitel 4.2 ein Ansatz untersucht, bei dem das Netz auch auf verrauschten Daten trainiert wurde und die SNR-Werte der jeweiligen akustischen Daten als zusätzliche Eingabe erhielt.

4.1 Kombination ohne zusätzliche Information

Das Netz aus Abbildung 4.1 wurde zunächst auf dem unverrauschten Datenset (mum1&2, 170 Sequenzen) trainiert und anschließend auf der Testdatenbasis (siehe Seite 16) getestet. Die dabei erzielten Erkennungsraten stehen in Tabelle 4.1. Eine grafische Darstellung der Ergebnisse befindet sich in Abbildung 4.2 (Graustufenbilder), bzw. in Abbildung 4.3 (LDA-Daten).

Graustufen-bilder	rein akustische Erkennung	Input-kombination (25 H.-Units)	Hidden-kombination (15 H.-Units)	Kombin. auf Phonemebene (ca.-Werte)
CLEAN (30 dB)	94.1%	92.9%	97.1%	≈96%
NOISE 1 (16 dB)	67.6%	72.4%	78.2%	≈76%
NOISE 2 (8 dB)	49.4%	50.0%	58.2%	≈66%
INCREASE 1 (30-16 dB)	74.7%	76.5%	84.1%	≈86%
INCREASE 2 (30-8 dB)	44.1%	46.5%	55.3%	≈70%
RADIO 1 (≈ 20 dB)	92.4%	80.0%	93.5%	≈93%
RADIO 2 (≈ 17 dB)	72.9%	58.8%	75.9%	≈80%
MOTOR 1 (≈ 25 dB)	93.5%	89.4%	94.7%	≈95%
MOTOR 2 (≈ 11 dB)	50.6%	57.1%	60.6%	≈70%

LDA-Daten	rein akustische Erkennung	Input-kombination (20 H.-Units)	Hidden-kombination (10 H.-Units)	Kombin. auf Phonemebene (ca.-Werte)
CLEAN (30 dB)	94.1%	92.9%	95.3%	≈97%
NOISE 1 (16 dB)	67.6%	76.5%	78.2%	≈80%
NOISE 2 (8 dB)	49.4%	53.5%	58.2%	≈74%
INCREASE 1 (30-16 dB)	74.7%	87.1%	85.9%	≈87%
INCREASE 2 (30-8 dB)	44.1%	77.1%	70.6%	≈76%
RADIO 1 (≈ 20 dB)	92.4%	91.2%	91.8%	≈95%
RADIO 2 (≈ 17 dB)	72.9%	87.6%	87.6%	≈87%
MOTOR 1 (≈ 25 dB)	93.5%	91.8%	94.7%	≈96%
MOTOR 2 (≈ 11 dB)	50.6%	67.1%	64.7%	≈78%

Tabelle 4.1: Testergebnisse bei der Verwendung einer unverrauschten Trainingsmenge und Kombination auf der Input-Ebene im Vergleich zu den besten Ergebnissen bei der Kombination auf Hidden- und Phonem-Ebene mit gleicher Trainingsdatenmenge. In der oberen Tabelle sind die mit Graustufenbildern erzielten Ergebnisse dargestellt, die untere enthält die Erkennungsraten bei Gebrauch von LDA-Daten.

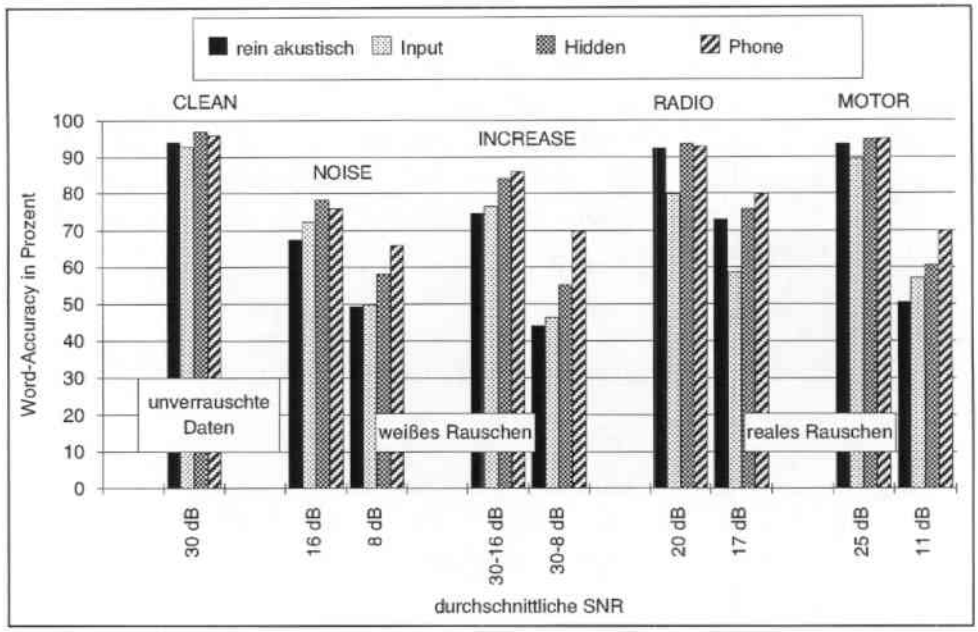


Abbildung 4.2: Testergebnisse bei Gebrauch von Graustufenbildern und einer unverrauschten Trainingsmenge bei der Kombination auf der Input-Ebene im Vergleich zu den besten Ergebnissen bei der Kombination auf Hidden- und Phonem-Ebene mit gleicher Trainingsdatenmenge.

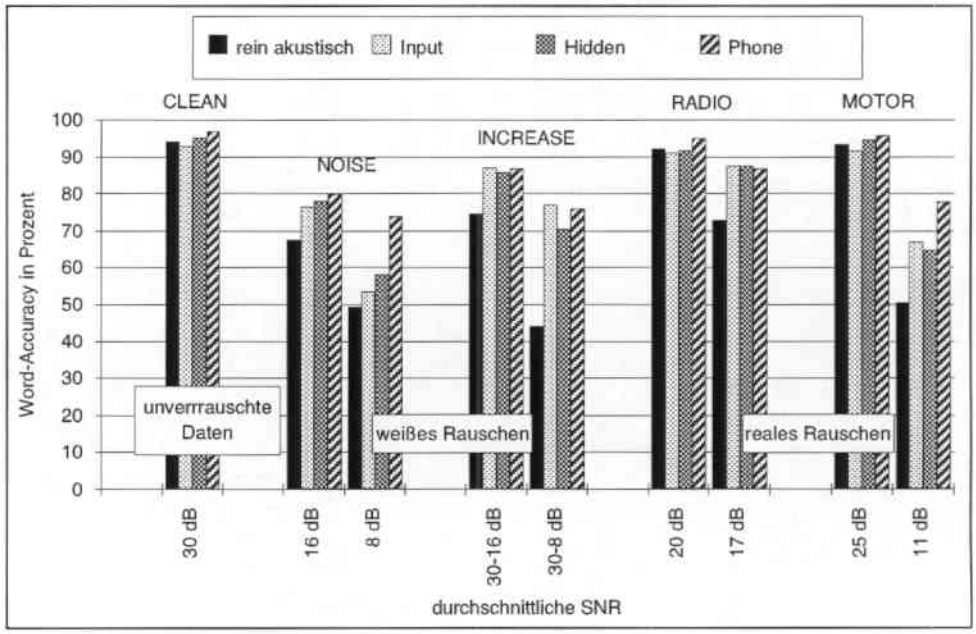


Abbildung 4.3: Testergebnisse bei Gebrauch von LDA-Daten und einer unverrauschten Trainingsmenge bei der Kombination auf der Input-Ebene im Vergleich zu den besten Ergebnissen bei der Kombination auf Hidden- und Phonem-Ebene mit gleicher Trainingsdatenmenge.

Bei der Verwendung von Graustufenbildern ist die erreichte Word-Accuracy wesentlich schlechter als bei Kombination auf Hidden- und Phonem-Ebene. Teilweise wird sogar ein schlechteres Ergebnis erzielt als bei einer rein akustischen Erkennung. Die Ergebnisse mit LDA-Daten als

visuelle Eingabe sind zwar um einiges besser, reichen aber im allgemeinen ebenfalls nicht an die bei der Kombination auf Phonemebene erzielten Erkennungsraten heran.

Bei der Interpretation der Ergebnisse ist jedoch zu beachten, daß hier nicht wie bei der Architektur aus Kapitel 3.1 eine Optimierung der Anzahl der Hidden-Units vorgenommen wurde. Es wurden vielmehr intuitiv 25 bzw. 20 Einheiten festgelegt. Ferner sei hier nocheinmal auf das ungleiche Verhältnis der akustischen zu den visuellen Eingabeparametern bei Verwendung von Graustufenbildern hingewiesen (16 akustische im Gegensatz zu 384 visuellen Merkmalen), was sich offensichtlich in einer schlechteren Erkennungsrate bemerkbar macht.

4.2 Kombination mit der SNR als zusätzliche Eingabe

Im folgenden wurde, analog zum Vorgehen in Kapitel 3.2, versucht, das Netz auf einer verrauschten Datenmenge zu trainieren und die SNR-Werte der jeweiligen akustischen Eingabesignale als zusätzliche Netzeingabe zu verwenden. Die entsprechende Architektur des Erkenners ist in Abbildung 4.4 dargestellt.

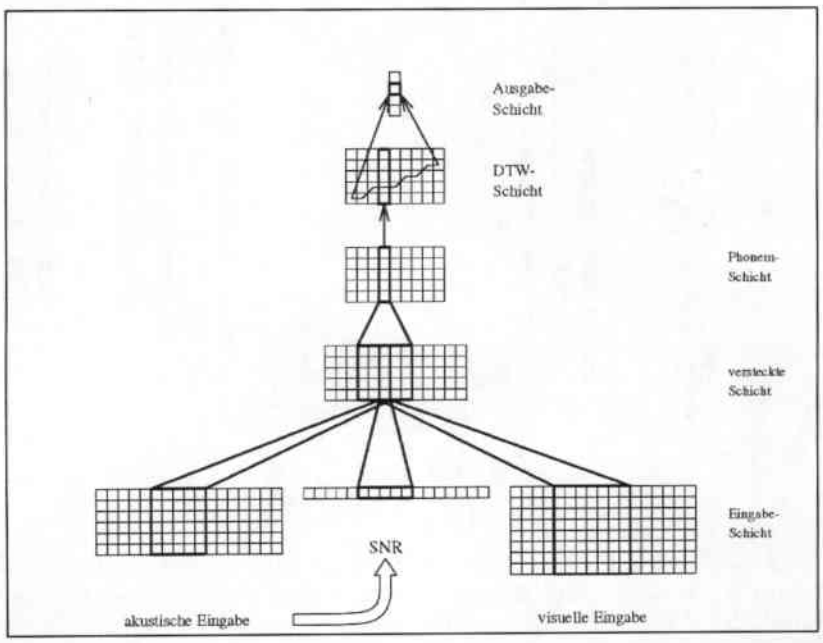


Abbildung 4.4: Architektur des Erkenners bei der Kombination auf der Eingabeschicht mit der aus den akustischen Daten berechneten SNR als zusätzliche Eingabe.

Das Netz wurde auf dem verrauschten Datenset trainiert, das auch schon in Kapitel 3.2 (siehe Seite 44) für das entsprechende Netz bei der Kombination auf Hidden-Ebene verwendet wurde. Für die visuelle Eingabe wurden ausschließlich LDA-Daten verwendet. Die Ergebnisse sind in Tabelle 4.2 aufgelistet. Ein Diagramm mit den jeweiligen Erkennungsraten befindet sich in Abbildung 4.5.

Mit Ausnahme der beiden Datenmenge INCREASE 1 und INCREASE 2 sind die erzielten Ergebnisse sogar schlechter als die mit der rein akustischen Erkennung erreichten.

LDA-Daten	rein akustische Erkennung	Input-kombination mit SNR	Hidden-kombination ohne SNR	Hidden-kombination mit SNR
CLEAN (30 dB)	95.3%	92.9%	93.5%	94.7%
NOISE 1 (16 dB)	92.9%	88.2%	92.9%	93.5%
NOISE 2 (8 dB)	92.9%	88.2%	91.2%	91.2%
INCREASE 1 (30-16 dB)	80.0%	90.6%	92.4%	95.3%
INCREASE 2 (30-8 dB)	54.7%	87.1%	81.2%	79.4%
RADIO 1 ($\approx$ 20 dB)	95.9%	91.8%	95.3%	92.9%
RADIO 2 ($\approx$ 17 dB)	94.1%	87.6%	92.4%	94.1%
MOTOR 1 ($\approx$ 25 dB)	95.3%	91.2%	94.1%	96.5%
MOTOR 2 ($\approx$ 11 dB)	91.8%	86.5%	91.8%	92.4%

Tabelle 4.2: Ergebnisse bei der Kombination auf der Input-Ebene und zusätzlicher Eingabe der SNR, sowie Verwendung von LDA-Daten für die visuelle Eingabe.

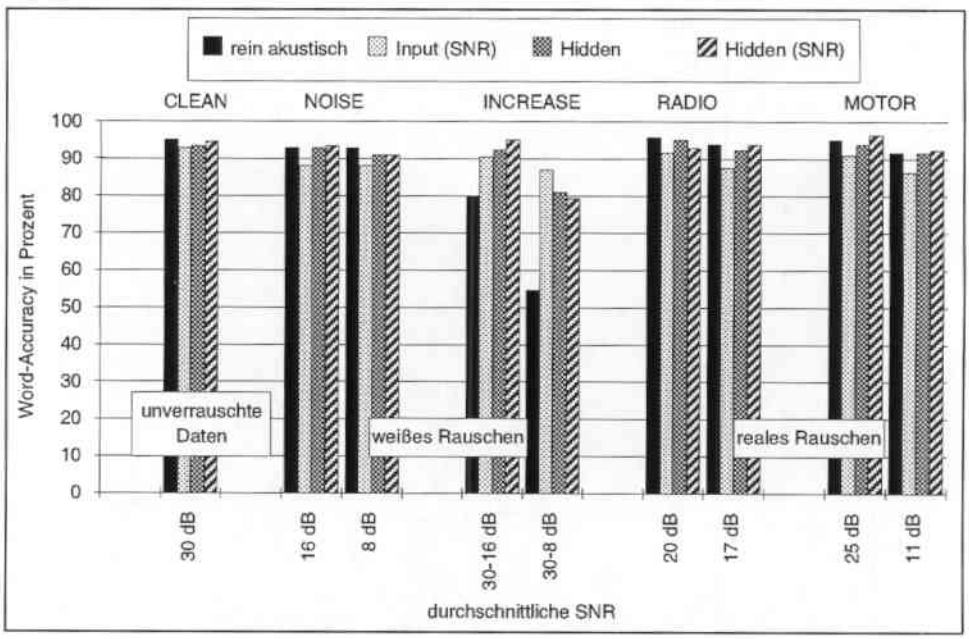


Abbildung 4.5: Testergebnisse bei der Kombination auf der Eingabeebene („Input"). Zum Vergleich sind die Ergebnisse bei gleicher Trainings- und Testdatenmenge bei der Kombination auf Hidden-Ebene ohne („Hidden") und mit („Hidden (SNR)") zusätzlicher Eingabe der SNR angegeben.

Wenn hier auch nur ein Bruchteil der sich anbietenden Verfahren für die Kombination durchgeführt und getestet wurde, so läßt sich anhand der erhaltenen Ergebnisse doch sagen, daß die Verschmelzung der beiden Modalitäten bereits auf der Eingabeebene wohl keine Alternative zu den beiden anderen untersuchten Integrationsmöglichkeiten darstellt. Eine fundiertere Bewertung ist jedoch erst nach einer Optimierung des Netzes und weiteren Tests möglich.

Kapitel 5

Zusammenfassung

In den vorangegangenen Kapiteln wurden verschiedene Methoden der bimodalen Sensorfusion am Beispiel der automatischen Spracherkennung mit Lippenlesen untersucht. Dabei wurden insbesondere verschiedene Integrationsebenen in der neuronalen Architektur des zugrundeliegenden Erkenners getestet. Das Training des Neuronalen Netzes erfolgte zum einen mit einem unverrauschten, zum anderen mit einem verrauschten Datenset. Ferner wurde auf der akustischen Eingabe die jeweilige SNR als eine Maßzahl für den Grad der Verrauschung dieser Daten berechnet und versucht, diese zusätzliche Information im Erkennungsprozeß auszunutzen. Eine Übersicht über die verschiedenen untersuchten Ansätze ist in Tabelle 5.1 zu finden.

Trainingsdatenmenge	unverrauschte Daten		unverrauschte und verrauschte Daten	
zusätzliche Infomation	keine	SNR-Werte	keine	SNR-Werte
rein akustische Erkennung	Kapitel 2.1		Kapitel 3.2	
Kombination auf Input-Ebene	Kapitel 4.1			Kapitel 4.2
Kombination auf Hidden-Ebene	Kapitel 3.1		Kapitel 3.2	Kapitel 3.2
Kombination auf Phonem-Ebene	Kapitel 2.1	Kapitel 2.2		

Tabelle 5.1: Übersicht über die untersuchten Ansätze

Zunächst wurden diverse Verfahren zur Kombination auf Phonemebene getestet, bei denen gewisse Parameter in Abhängigkeit von der Qualität der akustischen Eingabedaten von Hand eingestellt werden mußten. Diese Verfahren wurden im folgenden automatisiert, indem die vorher manuell festzulegenden Gewichte über die aus der akustischen Eingabe berechnete SNR gesetzt wurden. Obwohl nur eine kleine Untermenge (NOISE 1 und NOISE 2) der verschiedenen Rauscharten, die in der Testdatenbasis vorkamen, die Grundlage für die Umsetzung der

SNR-Werte in die entsprechenden Gewichte bildete, wurden auf den meisten Datensets der Testmengen annähernd die gleichen Ergebnisse wie bei einer von Hand optimierten Gewichtung erzielt.
Zum Training des MS-TDNN wurden bei dieser Kombinationsebene nur unverrauschte Daten verwendet. Das mit dieser Trainingsmenge trainierte Netz mit Kombination auf der Hidden-Ebene brachte bei starken Verrauschungen keine zufriedenstellenden Ergebnisse. Deshalb wurde versucht, das Netz mit einer Datenmenge zu trainieren, die auch verrauschte Samples enthält. Ferner wurde ein weiterer Ansatz untersucht, bei dem das Netz Zusatzinformation in Form der SNR der akustischen Daten erhielt. Beide Verfahren konnten bei den meisten Daten keine wesentlichen Verbesserungen gegenüber den bei rein akustischer Erkennung mit gleicher Trainingsdatenmenge erzielten Ergebnissen bringen. Desweiteren stellt sich bei dieser Art der Architektur und bei einer verrauschten Trainingsdatenmenge die Frage, inwieweit ein derartig trainierter Erkenner in der Lage ist, auf Rauscharten, die nicht in der Trainingsmenge vorkommen, zu generalisieren. Dies wurde in der vorliegenden Arbeit nicht mehr untersucht. Die hier erzielten Ergebnisse legen jedoch durchaus die Vermutung nahe, daß eine mangelnde Generalisierungsfähigkeit des rein akustischen Erkenners unter Umständen durch eine kombinierte Erkennung verhindert oder zumindest eingeschränkt werden kann.
Die Ergebnisse bei der Kombination auf Input-Ebene, die hier allerdings nur ansatzweise untersucht wurde, waren wesentlich schlechter als bei den beiden anderen Kombinationsalternativen. Die Hidden-Architektur scheint vorteilhafter, da das Netz hier die Möglichkeit besitzt, zunächst individuelle Merkmale der beiden Eingabemodalitäten getrennt zu lernen und anschließend unter Umständen bestehende Zusammenhänge der beiden Signalquellen zu erkennen.
Diese Möglichkeit besteht bei der Kombination auf Phonemebene nicht. Deshalb wäre eine weitere mögliche Alternative zu den hier untersuchten Ansätzen bei der Verschmelzung der unterschiedlichen Modalitäten auf dieser Ebene, ein zusätzliches Neuronales Netz zur Kombination einzusetzen. Dies hätte den Vorteil, daß für unterschiedliche Phoneme auch unterschiedliche Gewichte erlernt werden könnten, ähnlich wie es in Kapitel 2.3 bereits auf der Basis von Konfusionsmatrizen versucht wurde. Dieser Ansatz führte aufgrund der zu geringen Datenmenge zur Erstellung der Konfusionsmatrizen zu keinen befriedigenden Ergebnissen. Bei einer größeren Menge an Beispieldaten sind hier jedoch durchaus weitere Ergebnisverbesserungen zu erwarten.

Anhang A

Ergebnistabellen zu Kapitel 2.1

Da bei einigen Verfahren ein Einstellen und Optimieren gewisser Parameter von Hand vorgenommen werden mußte, waren mehrere Tests mit unterschiedlichen Parameterwerten notwendig, um die Einstellungen zu erhalten, bei denen die beste Erkennungsrate erzielt wird. In den vorangegangenen Kapiteln sind nur diese jeweils besten Einstellungen aufgeführt. Die folgenden Tabellen enthalten die Ergebnisse der restlichen Tests. Die jeweiligen Maximalwerte sind durch Fettdruck hervorgehoben. Zu den Werten bei fehlenden Tabelleneinträgen wurden keine Tests durchgeführt, da eine Verbesserung der Erkennungsleistung nicht zu erwarten war.

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
1.0	94.1%	67.6%	49.4%	74.7%	44.1%	92.4%	72.9%	93.5%	50.6%
0.9	**95.9%**				47.1%	91.2%	75.9%	**94.7%**	62.4%
0.8	**95.9%**	74.4%	60.0%	75.9%	49.4%	**92.9%**	**77.1%**	94.1%	**67.6%**
0.7	94.1%	**75.9%**	**62.4%**	75.9%	66.5%	92.4%	76.5%	92.9%	66.5%
0.6		71.2%	61.2%	**78.8%**	69.4%	85.3%	74.7%	87.6%	64.1%
0.5	83.5%		**62.4%**	74.1%	**71.2%**	80.0%	67.1%		
0.4			55.3%	67.6%	64.1%				
0.3					56.5%				

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a durch Bias b (manuell gesetzt) und die Entropy bestimmt)								
b	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.9	95.9%	72.9%	50.6%	78.2%	47.6%	91.2%	76.5%	**94.7%**	62.9%
0.8	**96.5%**	**75.9%**	58.8%	82.9%	50.6%	**93.5%**	77.1%	**94.7%**	67.1%
0.7	94.1%	74.7%	**62.9%**	**85.9%**	63.5%	91.2%	**78.2%**	92.9%	**68.2%**
0.6	92.9%	70.0%	60.6%	80.0%	66.5%	85.3%	75.9%	88.8%	62.4%
0.5	88.8%	66.5%	54.1%	74.1%	**69.4%**		72.4%	81.8%	60.6%
0.4		64.7%	58.2%		67.1%				
0.3		60.6%	57.1%		62.9%				
0.2			53.5%						
0.1			54.7%						

Tabelle A.1: Testergebnisse mit den bisherigen Verfahren und Graustufenbildern als visuelle Eingabe. Angegeben ist jeweils die Word-Accuracy in Prozent, sowie das von Hand eingestellte, optimale Gewicht w_a bei der gewichteten Addition bzw. der von Hand eingestellte, beste Schwellwert (Bias b aus Gleichung 1.2, Seite 13).

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
1.0	94.1%	67.6%	49.4%	74.7%	44.1%	92.4%	72.9%	93.5%	50.6%
0.9	95.9%	70.6%		60.6%	47.1%	92.9%	77.6%	94.7%	60.6%
0.8	97.1%	72.9%	64.1%	78.8%	50.6%	**95.3%**	85.3%	**95.9%**	72.4%
0.7	**97.6%**	77.6%	70.0%	83.5%	63.5%	**95.3%**	**87.1%**	94.1%	**75.9%**
0.6	95.9%	**79.4%**	**75.3%**	**84.7%**	75.3%	92.4%	**87.1%**	92.4%	72.9%
0.5		77.6%	71.8%	82.4%	**78.2%**	88.2%	82.9%		71.8%
0.4		72.9%	69.4%		73.5%		77.1%		
0.3					69.4%				
0.0	94.1%	67.6%	49.4%	74.7%	44.1%	92.4%	72.9%	93.5%	50.6%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a durch Bias b (manuell gesetzt) und die Entropy bestimmt)								
b	saubere Daten	NOISE 1	NOISE 2	INCR. 2	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.9	95.9%	71.8%		78.8%		**92.9%**	80.0%	**95.3%**	61.2%
0.8	97.1%	75.3%	65.9%	**87.1%**	52.4%	**92.9%**	85.9%	**95.3%**	74.1%
0.7	**97.6%**	78.8%	**71.2%**	82.9%	64.7%	**92.9%**	**87.1%**	94.1%	**75.3%**
0.6	94.7%	**80.0%**	70.6%	83.5%	74.1%	90.0%	85.3%	89.4%	72.4%
0.5		76.5%	68.8%	82.4%	**76.5%**	85.9%	81.8%		71.2%
0.4		71.8%	68.2%		75.3%				
0.3			65.9%		69.4%				
0.2					67.6%				

Tabelle A.2: Testergebnisse mit den bisherigen Verfahren und LDA-Daten als visuelle Eingabe. Angegeben ist jeweils die Word-Accuracy in Prozent, sowie das von Hand eingestellte, optimale Gewicht w_a bei der gewichteten Addition bzw. der von Hand eingestellte, beste Schwellwert (Bias b aus Gleichung 1.2, Seite 13).

Graustufen-bilder	Kombinationsschema: $hyp_i^B = (hyp_i^A + s_a) * (hyp_i^V + s_v)$ (nach der Addition frameweise normiert, s_a von Hand gesetzt)		
s_a	saubere Daten	NOISE 1	NOISE 2
0.7	80.6%		
0.6	87.1%		
0.5	89.4%	71.2%	62.4%
0.4	92.9%	73.5%	62.4%
0.3	93.5%	**75.3%**	**64.7%**
0.2	95.3%	73.5%	61.8%
0.1	**95.9%**		

Tabelle A.3: Testergebnisse bei modifizierter Multiplikation der einzelnen Aktivierungen.

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.7	94.7%	71.2%		77.6%	43.5%				
0.6	95.9%	72.9%		77.1%	47.1%	91.8%	81.8%	**95.9%**	70.0%
0.55	**96.5%**	75.3%							
0.5	95.9%	**76.5%**	63.5%	84.7%	52.4%	**92.4%**	**82.9%**	95.3%	**71.2%**
0.45		**76.5%**	64.1%						
0.4	94.7%	75.9%	**67.1%**	**85.3%**	59.4%	90.6%	80.6%	92.9%	65.9%
0.35			64.7%						
0.3	90.6%	70.6%	62.9%	78.2%	**65.3%**	80.0%	68.2%	84.7%	62.4%
0.2		58.8%	54.1%		57.1%				
0.1			42.4%						

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.9	94.7%					**92.4%**			
0.8	94.7%			77.1%		91.8%	78.8%	92.9%	58.8%
0.7	**95.9%**	70.6%		76.5%		91.8%	78.8%	94.1%	63.5%
0.65	**95.9%**								
0.6	**95.9%**	73.5%		78.2%	46.5%	91.2%	78.8%	**95.3%**	68.2%
0.5	95.3%	75.9%	64.1%	84.7%	52.4%	**92.4%**	**80.6%**	94.7%	**71.2%**
0.45		75.9%	64.7%						
0.4	95.3%	**77.1%**	**65.9%**	**86.5%**	61.8%	89.4%	**80.6%**	92.9%	68.8%
0.35		73.5%	64.7%						
0.3		71.8%	64.7%	81.8%	65.9%	81.8%	72.4%		64.1%
0.2		64.7%	58.2%		**67.1%**	72.9%	65.9%		61.8%
0.1		55.9%	52.4%		56.5%				

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
1.0		68.2%	48.2%		42.9%	91.8%			
0.8	94.7%			75.9%		91.8%			
0.7	**95.9%**	71.2%		76.5%		91.8%	78.8%	93.5%	63.5%
0.6	**95.9%**	73.5%		78.2%	46.5%	91.2%	78.2%	**95.3%**	68.8%
0.55	**95.9%**								
0.5	95.3%	75.9%	64.1%	84.7%	51.8%	**92.4%**	**80.0%**	**95.3%**	**71.2%**
0.45		75.9%	65.3%						
0.4	95.3%	**77.1%**	**66.5%**	**87.6%**	61.8%	89.4%	**80.0%**	92.9%	70.0%
0.35			64.1%						
0.3		72.4%	65.3%	83.5%	**65.9%**	82.9%	80.0%		64.7%
0.2		65.9%	57.6%		**65.9%**				
0.1			52.4%		60.0%				

Tabelle A.4: Testergebnisse mit weiteren Verfahren und Graustufenbildern

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.9	93.5%			71.8%		88.8%			
0.8	**96.5%**	70.6%	49.4%	76.5%	43.5%	90.6%	78.8%	94.7%	64.1%
0.7	**96.5%**	72.4%	57.1%	80.6%	47.1%	92.9%	81.8%	95.9%	70.0%
0.6	**96.5%**	**77.6%**	67.1%	84.1%	54.7%	**93.5%**	**87.6%**	**96.5%**	**76.5%**
0.5	**96.5%**	74.7%	**72.9%**	**86.5%**	62.4%	**93.5%**	84.1%	**96.5%**	72.4%
0.4	92.4%	74.1%	71.8%	81.2%	**71.2%**	84.1%	79.4%	87.6%	
0.3					66.5%				
0.2					60.0%				

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.8	96.5%			78.8%		92.9%		94.7%	
0.7	96.5%	72.9%		81.2%		**94.1%**	84.7%	**97.1%**	72.4%
0.6	96.5%	77.1%	68.8%	**87.6%**	57.6%	**94.1%**	**87.1%**	95.9%	**78.8%**
0.5	**97.1%**	**77.6%**	**72.9%**	85.3%	64.1%	92.9%	85.9%		74.1%
0.4	94.7%	76.5%	71.2%	83.5%	**75.3%**	87.6%	81.8%		72.4%
0.3		76.5%	67.6%		70.6%				
0.2					67.6%				

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a von Hand gesetzt)								
w_a	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
0.9	96.5%								
0.8	96.5%			78.8%		92.9%		95.3%	
0.7	96.5%	72.9%		82.4%	%	**94.1%**	84.7%	95.9%	72.4%
0.6	96.5%	77.1%	68.8%	**87.6%**	58.2%	**94.1%**	**87.1%**	95.9%	**78.2%**
0.5	**97.1%**	77.1%	**71.8%**	85.3%	65.3%	92.9%	85.9%	**96.5%**	74.7%
0.4	95.3%	**77.6%**	71.2%	84.1%	**74.7%**	88.2%	82.4%	90.6%	72.9%
0.3	88.2%	74.7%	68.2%	80.0%	71.2%			86.5%	
0.2		68.8%	64.1%		68.8%				

Tabelle A.5: Testergebnisse mit weiteren Verfahren und LDA-Daten

Graustufen-bilder	Kombinationsschema: $hyp_i^B = hyp_i^A + hyp_i^{V^{w_v}}$ (w_v von Hand gesetzt)		
w_a	saubere Daten	NOISE 1	NOISE 2
1.2			53.5%
1.1		70.0%	53.5%
1.0		**71.8%**	**59.4%**
0.9		**71.8%**	**59.4%**
0.8		**71.8%**	58.8%
0.7		70.6%	
0.6		68.2%	
0.5	94.1%	70.6%	57.6%
0.4	94.1%	70.0%	
0.3	94.1%	70.0%	
0.2	94.1%	70.6%	
0.1	**94.7%**	70.0%	
0.0	**94.7%**	68.2%	

Graustufen-bilder	Kombinationsschema: $hyp_i^B = hyp_i^{A^{w_a}} + hyp_i^{V^{w_v}}$ (w_a von Hand gesetzt)		
w_a	saubere Daten	NOISE 1	NOISE 2
1.0	**94.7%**		
0.9	**94.7%**		
0.8	94.1%		
0.7	94.1%		
0.6	93.5%	70.0%	
0.5	93.5%	72.9%	
0.4	90.0%	74.1%	
0.35		74.7%	
0.3	84.7%	**75.9%**	62.9%
0.25		74.1%	64.1%
0.2		73.5%	**65.9%**
0.15	79.4%		63.5%
0.1			63.5%

Tabelle A.6: Testergebnisse bei Gewichtung durch unterschiedliche Exponenten.

Anhang B

Ergebnistabellen zu Kapitel 2.2

Im folgenden werden die Ergebnisse, die mit den Verfahren aus Kapitel 2.2 bei unterschiedlichen Parametereinstellungen erzielt wurden, tabellarisch aufgelistet. Es handelt sich hierbei um die Verfahren, bei denen die Parameter, die für die entsprechende Kombination jeweils erforderlich sind, über die SNR-Werte der akustischen Daten automatisch gesetzt wurden.

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a über SNR gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-5/ 25/ 0.2 / 0.8)	95.9%	74.7%	62.4%	84.1%	71.8%	91.2%	75.9%	94.7%	61.2%
(0/ 32/ 0.6 / 0.85)	96.5%	74.7%	61.8%						
(0/ 32/ 0.45 / 0.85)	95.5%	76.5%	61.2%		71.8%	91.2%	75.9%		

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a über SNR gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 32/ 0.3 / 0.6)	95.9%	75.3%	65.9%						
(2/ 33/ 0.3/ 0.6)	95.9%	75.3%	65.9%						
(2/ 30/ 0.35/ 0.5)	95.3%	75.9%	66.5%	87.1%	61.2%	91.2%	81.2%	95.3%	68.8%
(0/ 32/ 0.15/ 0.6)	95.9%	76.5%	64.7%		67.6%	91.2%			

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 32/ 0.35/ 0.55)	96.5%	75.3%	66.5%	84.7%	54.1%	92.4%	81.8%	95.3%	67.6%
(0/ 28/ 0.35/ 0.55)	96.5%	74.7%	65.9%						
(-5/ 28/ 0.35/ 0.55)		75.9%	63.5%						
(0/ 32/ 0.35/ 0.6)	95.9%	75.9%	65.3%						
(0/ 32/ 0.3 / 0.55)	96.5%	75.9%	65.9%						
(0/ 32/ 0.3 / 0.6)	95.9%	74.7%	66.5%						
(2/ 36/ 0.35/ 0.6)	95.9%	75.3%	66.5%						
(1/ 30/ 0.35/ 0.55)	96.5%	75.3%	66.5%		57.1%				
(-2/ 23/ 0.3 / 0.55)	96.5%	76.5%	65.9%	84.1%	57.1%	92.9%	81.8%	95.3%	67.1%
(0/ 20/ 0.3 / 0.55)	96.5%	75.3%	65.3%						
(-1/ 23/ 0.25/ 0.55)	96.5%	75.3%	66.5%						
(-4/ 29/ 0.25/ 0.55)	95.9%	75.9%	65.3%		61.2%	92.4%			

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(1/ 33/ 0.3 / 0.6)	95.9%	75.3%	67.1%	84.1%	60.6%	91.8%	81.2%	94.7%	68.2%
(2/ 30/ 0.35/ 0.5)	95.3%	75.9%	66.5%						
(0/ 31/ 0.2 / 0.65)	95.9%	75.3%	62.9%		61.2%	91.8%			

Tabelle B.1: Testergebnisse mit automatisierten Verfahren (Gewichtung in Abhängigkeit von der SNR) und Graustufenbildern.

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a über SNR gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 33/ 0.5 / 0.75)	97.1%	78.8%	71.2%	82.9%	74.7%	94.1%	87.6%	94.1%	73.5%
(0/ 33/ 0.6 / 0.85)	96.5%	74.7%	72.9%	86.5%	65.9%	95.9%	86.5%	95.9%	74.1%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a über SNR gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-2/ 30/ 0.4/ 0.55)	97.1%	77.1%	70.6%	85.3%	68.8%	93.5%	85.9%	96.5%	74.1%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-1/ 33/ 0.35/ 0.8)	96.5%	73.5%	72.9%	82.9%	60.6%	93.5%	85.9%	95.3%	72.9%
(-1/ 33/ 0.45/ 0.9)	95.9%	73.5%	65.9%	79.4%	50.0%	91.2%	82.9%	95.3%	75.9%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a über gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-2/ 28/ 0.4/ 0.55)	97.1%	78.2%	71.2%	86.5%	66.5%	93.5%	85.3%	96.5%	72.9%

Tabelle B.2: Testergebnisse mit automatisierten Verfahren (Gewichtung in Abhängigkeit von der SNR) und LDA-Daten.

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a über SNR und Entropy gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-12/ 18/ 0.5/ 0.8)	96.5%	75.3%	63.5%						
(-1/ 18/ 0.55/ 0.8)	96.5%	75.3%	64.1%	82.9%	58.2%	93.5%	79.4%	94.7%	67.1%
(5/ 29/ 0.3 / 0.8)	95.9%	69.4%	58.2%		71.2%	88.2%			
(0/ 29/ 0.6 / 0.8)	96.5%	75.3%	62.4%		62.9%	93.5%			

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 31/ 0.65/ 0.8)	94.7%	71.2%	51.8%						
(0/ 31/ 0.5 / 0.7)	95.9%	74.7%	61.2%						
(0/ 31/ 0.35/ 0.55)	96.5%	76.5%	66.5%						
(2/ 30/ 0.35/ 0.5)	95.9%	76.5%	67.1%	86.5%	58.8%	90.6%	78.8%	95.3%	66.5%
(0/ 29/ 0.6 / 0.8)	94.7%	71.2%	53.5%		46.5%				
(-1/ 18/ 0.55/ 0.8)	94.7%	70.0%	51.8%						

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 32/ 0.35/ 0.55)	96.5%	76.5%	65.9%	84.7%	58.2%	92.4%	80.0%	95.3%	66.5%
(-1/ 18/ 0.55/ 0.8)	94.7%	69.4%	49.4%		43.5%	91.8%	80.0%	92.9%	61.2%

Graustufen-bilder	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
(SNR_{min}, SNR_{max}, $w_{a_{min}}, w_{a_{max}}$)	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(1/ 33/ 0.35/ 0.65)	95.9%	77.1%	65.3%						
(1/ 33/ 0.3 / 0.6)	96.5%	76.5%	65.9%	85.9%	58.8%	91.8%	79.4%	95.3%	65.9%
(1/ 33/ 0.25/ 0.55)	96.5%	74.7%	65.3%						
(2/ 30/ 0.35/ 0.5)	95.9%	75.9%	66.5%						
(0/ 29/ 0.6 / 0.8)	94.7%	71.8%	54.1%		45.9%				
(-1/ 18/ 0.55/ 0.8)	94.7%	70.0%	51.8%						

Tabelle B.3: Testergebnisse mit automatisierten Verfahren (Gewichtung in Abhängigkeit von der SNR und der Entropy) und Graustufenbildern.

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (w_a über SNR und Entropy gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(0/ 33/ 0.6/ 0.85)	96.5%	77.1%	71.2%	85.9%	65.3%	95.9%	87.1%	95.9%	74.7%
(0/ 33/ 0.5/ 0.75)	97.1%	81.2%	70.6%	84.1%	74.1%	92.4%	86.5%	92.9%	71.8%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-2/ 30/ 0.4/ 0.55)	97.1%	77.1%	71.8%		68.2%	91.8%	84.7%	94.7%	71.2%
(0/ 33/ 0.5/ 0.75)	95.9%	77.6%	71.2%	83.5%	59.4%	94.1%	87.6%	95.3%	78.2%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-1/ 33/ 0.35/ 0.8)	95.3%	73.5%	70.6%		60.0%	92.4%	83.5%	95.9%	71.2%
(0/ 33/ 0.5/ 0.75)	95.9%	75.9%	70.6%	82.9%	57.1%	93.5%	85.9%	95.9%	75.9%

LDA-Daten	Kombinationsschema: $hyp_i^B = w_a hyp_i^A + w_v hyp_i^V - w_a w_v hyp_i^A hyp_i^V$ (frameweise normiert, w_a über SNR und Entropy gesetzt)								
$(SNR_{min}, SNR_{max}, w_{a_{min}}, w_{a_{max}})$	saubere Daten	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(-2/ 28/ 0.4/ 0.55)	97.1%	76.5%	70.6%		68.2%	92.4%	82.9%	94.7%	71.2%
(0/ 33/ 0.5/ 0.75)	95.9%	77.6%	71.8%	84.1%	60.0%	94.1%	87.6%	95.3%	78.8%

Tabelle B.4: Testergebnisse mit automatisierten Verfahren (Gewichtung in Abhängigkeit von der SNR und der Entropy) und LDA-Daten.

Literaturverzeichnis

[1] L.D. Braida. Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology*, 43A(3):647–677, 1991.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. *Proc. ICASSP*, 1993. Minneapolis.

[3] C. Bregler, S. Manke, H. Hild, and A. Waibel. Bimodal sensor integration on the example of speech-reading. *ICNN*, 1993.

[4] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel. Toward movement-invariant automatic lip-reading and speech recognition. *Proc. ICASSP*, pages 109–112, 1995.

[5] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. *International Conference on Spoken Language Processing, ICSLP*, pages 547–550, 1994.

[6] H. Hild and A.Waibel. Multi-speaker / speaker-independent architectures for the multi-state time delay neural network. *Proc. Intern. Conference on Acoustics, Speech and Signal Processing ,IEEE*, 1993.

[7] Hermann Hild and Alex Waibel. Speaker-Independent Connected Letter Recognition With a Multi-Sate Time Delay Neural Network. In *3rd European Conference on Speech, Communication and Technology (EUROSPEECH) 93*, September 1993.

[8] 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.*

[9] M. Hunke and A. Waibel. Face localisation and tracking for human-computer interaction. *28th Annual Asimolar conference on Signal speech and Computers.*

[10] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 1976.

[11] U. Meier. Lippenlesen: verschiedene Methoden der visuellen Vorverarbeitung und Merkmalsextraktion. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1994.

[12] U. Meier. Robuste Systemarchitekturen für automatisches Lippenlesen. Diplom-Arbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[13] H. Ney. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, Signal Processing ASSP*, 32(2):263–271, April 1984.

[14] D.A. Pomerleau. *Neural Network Perception for Mobile Robot Guidance.* Phd thesis, Carnegie Mellon University, Pittsburgh, February 1992.

[15] M. Schoch. Schätzung des Signal-Rausch-Abstandes. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[16] P.L. Silsbee. Sensory integration in audiovisual automatic speech recognition. *28th Annual Asimolar conference on Signal speech and Computers.*

[17] R. Stiefelhagen. Automatische Bestimmung von Visemen für das maschinelle Lippenlesen. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[18] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. *IJCNN*, June 1992.

[19] Q. Summerfield. Audio-visual speech perception, lipreading and artificial stimulation. *Hearing Science and Hearing Disorders*, pages 131–182, 1983. London.

[20] A. Waibel, T. Hanazawa, G. Hinton, and K. Shikano. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328–339, 1989.