Universität Ulm, Abteilung für Neuroinformatik

Neuronale Fusionsarchitekturen in der audio-visuellen Sprachverarbeitung

Dissertation zur Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für Informatik der Universität Ulm

vorgelegt von

Barbara Helga Talle
aus Thuine

Ulm 2002

Tag der Promotion:	15. Januar 2003
amtierender Dekan:	Prof. Dr. Friedrich von Henke
Gutachter:	Prof. Dr. Günther Palm Prof. Dr. Heiko Neumann Prof. Dr. Heinrich Niemann (Universität Erlangen-Nürnberg)

Inhaltsverzeichnis

1 Einleitung **1**

2 Sprachproduktion und Sprachverarbeitung **3**
2.1 Phone und Phoneme 3
2.2 Anregung und Lautformung 4
2.2.1 Anregung 5
2.2.2 Lautformung 6
Vokale 7
Konsonanten 9
2.2.3 Quelle-Filter-Theorie 12
2.3 Das Ohr 13

3 Mustererkennung und Neuronale Netze **19**
3.1 Mustererkennung 19
3.1.1 Lineare Diskriminanzanalyse 23
3.2 Grundlagen der Neuronalen Netze 25
3.3 Modelle der Neuronalen Netze 28
3.3.1 Multilayer-Perzeptrone und Backpropagation 29
3.3.2 Time-Delay Netzwerke 32
3.3.3 Boltzmann-Maschinen 33
3.3.4 Assoziativspeicher 34
Lernen 35
Retrieval 36
Kodierung 37
Rückwärtsprojektion zur Entscheidungsfindung 38
3.4 Sprachverarbeitung in technischen Systemen 39
3.4.1 Verarbeitung und Darstellung von Sprachsignalen 39
Analyse im Zeitbereich 41
Analyse im Frequenzbereich 41
3.4.2 Kodierung von Sprachsignalen-Vektorquantisierung 44
3.4.3 Klassifikation von Sprachsignalen 44
Hidden Markov Modelle 44
Dynamische Zeitanpassung 45

4 Audio-visuelle Spracherkennung beim Menschen 47
4.1 Nutzen der Kombination 48
4.2 Komplementarität 49
4.3 Viseme 52
4.4 McGurk-Effekt 53
4.5 Visuelle Merkmale 56
4.6 Synchronität 58
4.7 Psychologische Untersuchungen der Integration 58
4.7.1 Welche Metrik könnte zur Integration genutzt werden ? . 59
4.7.2 Wo findet die Integration statt ? 60
Experimentelle Ergebnisse: frühe Integration 60
Experimentelle Ergebnisse: späte Fusion 63
4.8 Diskussion 67

5 Automatische audio-visuelle Spracherkennung 69
5.1 Späte Fusionsarchitekturen 70
5.2 Fusion auf einer Zwischenstufe 73
5.3 Frühe Fusionsarchitekturen 75
5.4 Vergleich von Fusionsarchitekturen 78
5.5 Diskussion 89

6 Das Worterkennungssystem 91

7 Durchführung der Fusionsexperimente 97
7.1 Datenmaterial 97
7.2 Auswahl der Trainings- und Testdaten 99
7.3 Spezifikation und Training des Worterkennungssystems 102
7.4 Fusionsarchitekturen 103

8 Fusionsresultate für Originalmerkmale 109
8.1 Optimierung der Netztopologien 109
8.2 Fusionsarchitekturen 111
8.3 Modulare Architekturen 114
8.4 Fusion mit modularen Architekturen 116
8.5 Experimente mit gestörtem akustischen Kanal 117
8.5.1 Experimente mit akustischen Daten 118
8.5.2 Fusionsarchitekturen mit gestörtem akustischen Kanal . . 119
8.5.3 Ergebnisse für die Architektur L 121
8.6 Kapitelzusammenfassung 123

9 Fusion mit dimensionsreduzierten Daten 125
9.1 Bestimmung der Anzahl der LDA-Koeffizienten 125
9.2 Optimierung der Netztopologien 127
9.3 Fusionsarchitekturen 132
9.4 Experimente mit gestörtem akustischen Kanal 132
9.4.1 Experimente mit akustischen Daten 132
9.4.2 Fusionsarchitekturen mit gestörtem akustischen Kanal . . 134

Vergleich der Ergebnisse für Originaldaten und dimensionsreduzierte Daten 135
Vergleich der Fusionsarchitekturen 138
Vergleich unterschiedlicher visueller Merkmale 141
9.4.3 Ergebnisse für die Architektur L 144
9.5 Kapitelzusammenfassung 148

10 Fusion der gestörten akustischen Daten **151**
10.1 Vergleich unterschiedlicher visueller Merkmale 154
10.2 Ergebnisse für die Architektur L 157
10.3 Vergleich der Fusionstrategien 159
10.4 Kapitelzusammenfassung 165

11 Zusammenfassung und Diskussion **167**

12 Ausblick **173**

A Ergebnisse für Originaldaten **175**

B Ergebnisse für dimensionsreduzierte Daten **185**

C Ergebnisse für gemeinsame Dimensionsreduktion **209**

Abbildungsverzeichnis

2.1 Sprechapparat des Menschen 5
2.2 Vokalbildung in der Mundhöhle 7
2.3 Vokalviereck 8
2.4 Konsonatenbildung in der Mundhöhle 9
2.5 Schematische Darstellung des Quelle-Filter-Modells 12
2.6 Aufbau des menschlichen Ohrs 13
2.7 Schematische Darstellung der aufgerollten Schnecke 15
2.8 Querschnitt durch eine Schneckenwindung 15

3.1 Prinzipieller Aufbau eines Klassifikationssystems 19
3.2 Aufteilung eines zweidimensionalen Musterraums durch mehrere Diskriminanten 21
3.3 Projektion zweidimensionaler Muster auf eindimensionale Unterräume 24
3.4 Schematische Abbildung eines biologischen Neurons 27
3.5 Schematische Abbildung eines künstlichen Neurons 27
3.6 Mehrschichtiges neuronales Netzwerk (MLP) 29
3.7 Konvergenzverhalten des BP-Algorithmus bei verschiedenen Lernraten 31
3.8 Aufbau eines Assoziativspeichers 35
3.9 Lernen im Assoziativspeicher 36
3.10 Schematische Darstellung der Vorverarbeitung in der Sprachverarbeitung 40

4.1 Audio-visuelle Erkennungsleistung von normal hörenden Versuchspersonen 49
4.2 Übertragene Information bei verschiedenen SNR-Verhältnissen unter Nutzung des akustischen und des akustisch-visuellen Signals 50
4.3 Auditorischer Verwechslungsbaum für Konsonanten 51
4.4 Visueller Verwechslungsbaum für Konsonanten 52

6.1 Worterkennungssystem für Buchstabenwörter 92

7.1 Diskriminationsvermögen für Einzelworte in Beziehung zum Signal-Rauschverhältnis 99
7.2 Schematischer Aufbau eines Crossvalidierungsexperiments 100
7.3 Übersicht über den prinzipiellen Aufbau der untersuchten Fusionsarchitekturen 105

8.1 Übersicht über die Fusionsarchitekturen für hochdimensionale Daten . . . 112
8.2 Fusionsergebnisse für ungestörte hochdimensionale Daten 113
8.3 Modularisierungsarchitektur für einzelne akustische Zeitscheiben 115
8.4 Modularisierungsarchitektur für einzelne Frequenzkanäle 115
8.5 Darstellung der modularen Architektur H2 117
8.6 Fusionsergebnisse mit Originaldaten, vollständige und modularisierte Architekturen 118
8.7 Ergebnisse für akustische Originaldaten, Training und Testen mit unterschiedlich gestörten Daten 119
8.8 Fusionsergebnisse mit unterschiedlich gestörten Originaldaten, modulare Architekturen 120
8.9 Fusionsergebnisse für Architektur L mit unterschiedlich gestörten Originaldaten 122

9.1 Ergebnisse für dimensionsreduzierte akustische Daten bei Variation der Anzahl der versteckten Neuronen 127
9.2 Ergebnisse für dimensionsreduzierte visuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung einer Zeitscheibe) 128
9.3 Ergebnisse für dimensionsreduzierte visuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung des ganzen Fensters) 128
9.4 Fusionsergebnisse für dimensionsreduzierte audiovisuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung einer Zeitscheibe, Architektur E) 129
9.5 Fusionsergebnisse für dimensionsreduzierte audiovisuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung einer Zeitscheibe, Architektur I1) 130
9.6 Fusionsergebnisse für dimensionsreduzierte audiovisuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung des ganzen Fensters, Architektur E) 131
9.7 Fusionsergebnisse für dimensionsreduzierte audiovisuelle Daten bei Variation der Anzahl der versteckten Neuronen (Verwendung des ganzen Fensters, Architektur I1) 131
9.8 Übersicht über die Fusionsarchitekturen für dimensionsreduzierte Daten 133
9.9 Ergebnisse für dimensionsreduzierte akustische Daten, Training und Testen mit unterschiedlich gestörten Daten 134
9.10 Vergleich der Fusionsergebnisse für hochdimensionale und dimensionsreduzierte Daten für den ungestörten akustischen Kanal (Verwendung einer Zeitscheibe als visuelles Merkmal) 135
9.11 Vergleich der Fusionsergebnisse für hochdimensionale und dimensionsreduzierte Daten für den mit 3dB gestörten akustischen Kanal (Verwendung einer Zeitscheibe als visuelles Merkmal) 136

9.12 Vergleich der Fusionsergebnisse für hochdimensionale und dimensionsreduzierte Daten für den mit -3dB gestörten akustischen Kanal (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 136
9.13 Vergleich der Fusionsergebnisse für hochdimensionale und dimensionsreduzierte Daten für den mit -11dB gestörten akustischen Kanal (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 137
9.14 Fusionsergebnisse bei Verwendung dimensionsreduzierter ungestörter und gestörter Daten (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 139
9.15 Fusionsergebnisse bei Verwendung dimensionsreduzierter ungestörter und gestörter Daten (Verwendung des ganzen Fensters als visuelles Merkmal) . . . 140
9.16 Fusionsergebnisse für dimensionsreduzierte Daten bei Verwendung ungestörter akustischer Daten, Vergleich der beiden visuellen Merkmale . . . 142
9.17 Fusionsergebnisse für dimensionsreduzierte Daten bei Verwendung der mit 3dB gestörten akustischen Daten, Vergleich der beiden visuellen Merkmale . . . 142
9.18 Fusionsergebnisse für dimensionsreduzierte Daten bei Verwendung der mit -3dB gestörten akustischen Daten, Vergleich der beiden visuellen Merkmale . . . 143
9.19 Fusionsergebnisse für dimensionsreduzierte Daten bei Verwendung der mit -11dB gestörten akustischen Daten, Vergleich der beiden visuellen Merkmale . . . 143
9.20 Fusionsergebnisse für Architektur L mit unterschiedlich gestörten dimensionsreduzierten Daten (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 144
9.21 Fusionsergebnisse für Architektur L mit unterschiedlich gestörten dimensionsreduzierten Daten (Verwendung des ganzen Fensters als visuelles Merkmal) . . . 145
9.22 Fusionsergebnisse für ausgewählte Variationen der Architektur L mit unterschiedlich gestörten dimensionsreduzierten Daten (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 146
9.23 Fusionsergebnisse für ausgewählte Variationen der Architektur L mit unterschiedlich gestörten dimensionsreduzierten Daten (Verwendung des ganzen Fensters als visuelles Merkmal) . . . 147

10.1 Architektur für eine gemeinsame Dimensionsreduktion der gestörten akustischen Daten . . . 152
10.2 Architektur für eine gemeinsame Dimensionsreduktion der gestörten audio-visuellen Daten . . . 152
10.3 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, ungestörte und mit 3 dB gestörte Daten (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 153
10.4 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, mit -3 dB und -11 dB gestörte Daten (Verwendung einer Zeitscheibe als visuelles Merkmal) . . . 154

10.5 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, ungestörte und mit 3 dB gestörte Daten (Verwendung des ganzen Fensters als visuelles Merkmal) 155
10.6 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, mit -3 dB und -11 dB gestörte Daten (Verwendung des ganzen Fensters als visuelles Merkmal) 155
10.7 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, ungestörte und mit -3 dB gestörte Daten, Vergleich der beiden visuellen Merkmale 156
10.8 Fusionsergebnisse bei gemeinsamer Dimensionsreduktion, mit 3 dB und -11 dB gestörte Daten, Vergleich der beiden visuellen Merkmale 156
10.9 Fusionsergebnisse für Architektur L bei gemeinsamer Dimensionsreduktion (Verwendung einer Zeitscheibe als visuelles Merkmal) 157
10.10 Fusionsergebnisse für Architektur L bei gemeinsamer Dimensionsreduktion (Verwendung des ganzen Fensters als visuelles Merkmal) 158
10.11 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung einer Zeitscheibe als visuelles Merkmal, ungestörte akustische Daten) 160
10.12 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung einer Zeitscheibe als visuelles Merkmal, mit 3 dB gestörte akustische Daten) . . . 160
10.13 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung einer Zeitscheibe als visuelles Merkmal, mit -3 dB gestörte akustische Daten) . . . 161
10.14 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung einer Zeitscheibe als visuelles Merkmal, mit -11 dB gestörte akustische Daten) . . 161
10.15 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung des ganzen Fensters als visuelles Merkmal, ungestörte akustische Daten) 162
10.16 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung des ganzen Fensters als visuelles Merkmal, mit 3 dB gestörte akustische Daten) . . . 162
10.17 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung des ganzen Fensters als visuelles Merkmal, mit -3 dB gestörte akustische Daten) . . . 163
10.18 Vergleich der Ergebnisse für die gemeinsame Dimensionsreduktion und für spezialisierte Netze (Verwendung des ganzen Fensters als visuelles Merkmal, mit -11 dB gestörte akustische Daten) . . 163

Kapitel 1

Einleitung

Spracherkennungssysteme finden in den letzen Jahren immer weitere Verbreitung, so zum Beispiel in Diktiersystemen, in Telefoniesystemen oder bei der Bedienung von Handys. Obwohl sich die Erkennungsleistungen dieser Systeme enorm verbessert haben, zeigen sie eine mangelnde Robustheit in unterschiedlichen Geräuschumgebungen. Auch ist ihre Leistungsfähigkeit verglichen mit der des Menschen immer noch geringer [Lip96], [Lip97]. Verfahren, die eine robuste Spracherkennung auch in geräuschvollen Umgebungen gewährleisten können, kommt daher eine große Bedeutung zu [Gon95].
Ein vergleichsweise neuer Ansatz, die Leistung von Spracherkennungssystemen zu steigern bzw. robust zu machen, besteht darin, auf andere Informationen als auf das akustische Sprachsignal alleine zurückzugreifen und eine multimodale Erkennung durchzuführen [Ovi00] [CMB97]. Diese Informationen werden durch akustische Störungen nicht beeinflußt. Aus Untersuchungen der menschlichen Sprachwahrnehmung ist bekannt, daß auch visuelle Informationen zur Spracherkennung herangezogen werden [MM76],[Sum92]. Daher liegt es nahe, Informationen zum Beispiel über die Lippenbewegungen des Sprechers in den Spracherkennungsprozeß in technischen Systemen mit einzubeziehen. Erst in den letzten Jahren wurde eine Anzahl solcher audio-visueller Spracherkennungssysteme entwickelt und das Forschungsgebiet der audio-visuellen Spracherkennung etabliert.
Eine bei der Entwicklung eines audio-visuellen Spracherkennungssystems noch ungelöste Fragestellung ist dabei die Problematik der bestmöglichen Sensorfusion. Die Fragestellung dabei lautet, auf welcher Ebene der Informationsverarbeitung die Fusion beider Kanäle am besten stattfinden sollte und wie diese optimal kombiniert werden. An eine gute Sensorfusionsstrategie wird die Anforderung gestellt, daß das resultierende System zumindest eine gleich gute Performanz wie die Einzelsysteme erbringt. Dabei sind die Ergebnisse des Gesamtsystems aufgrund der Performanzen der Einzelkanäle nicht vorhersagbar, da die unimodalen Informationen unter Umständen nicht optimal integriert werden. Andererseits können aber bei der Integration im Gesamtsystem auch intermodale Informationen verwendet werden, die bei der unimodalen Verarbeitung nicht vorhanden sind [Bra91].
In der vorliegenden Arbeit wird diese zentralen Fragestellung der Sensorfusion

behandelt. Zunächst wird der Prozeß der audio-visuellen Sprachereknnung beim Menschen näher betrachtet, indem die wichtigsten Ergebnisse psycho-physischer Experimente aus der Literatur vorgestellt werden. Dies erlaubt erste Hinweise auf eine gute Fusionsstrategie. Im weiteren Verlauf der Arbeit wird die automatische audio-visuelle Sprachereknnung behandelt. Nachdem mit Hilfe eines Literaturüberblicks noch offene Fragestellungen identifiziert werden, werden diese mit Hilfe eines neuartigen Worterkennungssystems untersucht, welches in dieser Arbeit entwickelt wird. Die verwendete Architektur ist neuronal aufgebaut und erlaubt daher die Realisierung unterschiedlicher Fusionsarchitekturen, in denen die Integration auf verschiedenen Verarbeitungsebenen stattfinden kann. Dieses Worterkennungssystem wird anhand der Klassifikation von Buchstabenwörtern getestet. Buchstabenwörter sind hochgradig verwechselbar und stellen daher ein schwieriges Klassifikationsproblem dar. Um statistische Aussagen über die Performanzen einzelner Fusionsstrategien gewinnen zu können, werden Cross-Validierungsexperimente durchgeführt.

Die Arbeit gliedert sich in drei Teile:
Im ersten Teil werden zunächst in Kapitel 2 die Grundzüge der Sprachproduktion und -verarbeitung erläutert. In Kapitel 3 werden die Grundlagen der Mustererkennung und der neuronalen Netze behandelt. Damit dient der erste Teil dem besseren Verständnis der folgenden Kapitel.
Der zweite Teil der Arbeit beschäftigt sich mit der audio-visuellen Sprachverarbeitung. Zunächst wird in Kapitel 4 der audio-visuelle Sprachverarbeitungsprozeß beim Menschen dargestellt, indem die wichtigsten Aspekte bei der audio-visuellen Sprachwahrnehmung des Menschen erläutert werden. Weiterhin werden Experimente vorgestellt, die die Integrationsproblematik bei der menschlichen Sprachwahrnehmung zum Thema haben. In Kapitel 5 werden technische Realisierungen audio-visueller Sprachverarbeitungssysteme beschrieben. Das Hauptaugenmerk liegt dabei auf denjenigen Systemen, die einen Vergleich unterschiedlicher Fusionsstrategien erlauben.
Der dritte Teil umfaßt schließlich die Darstellung der Untersuchungen an dem in dieser Arbeit entwickelten audio-visuellen Worterkennungssystem. In Kapitel 6 wird der Aufbau des Worterkennungssystems erläutert. Kapitel 7 umfaßt eine Darstellung des verwendeten Datenmaterials, der Trainingsprozedur und der untersuchten Fusionsarchitekturen. Im Kapitel 8 werden die Ergebnisse der Experimente mit hochdimensionalen Daten dargelegt, gefolgt von Kapitel 9, welches die Experimente mit niedrigdimensionalen Daten und deren Ergebnisse beschreibt. In Kapitel 10 werden Experimente vorgestellt, die auf einer Fusion der gestörten akustischen Daten beruhen. Die Zusammenfassung der Ergebnisse und ein Ausblick auf weitere Arbeiten erfolgen in Kapitel 11 und 12.

Kapitel 2

Sprachproduktion und Sprachverarbeitung

Dieses Kapitel beschäftigt sich mit der Sprachproduktion und der Sprachverarbeitung. Dabei soll zunächst auf den Begriff des Lautes näher eingegangen werden. Im folgenden Abschnitt wird beschrieben, wie die Laute vom Menschen während des Sprechens hervorgebracht werden, gefolgt von einem Abschnitt, der ein gängiges Modell der Sprachproduktion beschreibt. Der daran anschließende Abschnitt beschäftigt sich mit dem Aufbau des menschlichen Ohres und den Vorgängen, die sich beim Hören abspielen. Der letzte Abschnitt dieses Kapitels behandelt grundlegende Methoden der Sprachverarbeitung in technischen Systemen.

2.1 Phone und Phoneme

Der Begriff 'Laut' ist sehr allgemein und umfaßt zunächst alle Geräusche, die durch koordinierte Bewegungen des Sprechapparates hervorgerufen werden können. Obwohl der Mensch in der Lage ist, eine im Grunde genommen unendliche Anzahl verschiedener Laute hervorzubringen, wird nur ein kleiner Bruchteil davon zur Kommunikation genutzt.
In der Sprachwissenschaft bezeichnet man als Laut oder Phon die kleinste Einheit (minimales Lautsegment) der gesprochenen Sprache, welche in der geschriebenen Sprache ihre Analogie im Buchstaben hat. Eine grobe Unterteilung der Laute kann zunächst hinsichtlich der Bewegung der Artikulationsorgane getroffen werden. Bei der Artikulation einiger Laute verändert sich die Stellung der Artikulatoren nicht, so daß diese Laute über einen langen Zeitraum angehalten werden können. Auf der anderen Seite existieren Laute, bei deren Äußerung die Artikulationsorgane einen bestimmten Bewegungsablauf ausführen, so daß diese prinzipiell nicht lange angehalten werden können. Zur weiteren Unterscheidung verschiedener Laute dienen die Klangfarbe, die zeitliche Länge, die Stärke der Betonung und die Tonhöhe. Phone werden mit den Abgrenzungszeichen [..] zitiert. Da es streng genommen nicht möglich ist, einen Laut zweimal auf genau dieselbe Art und Weise zu äußern, ist mit [a] daher nicht ein ganz bestimmter Laut gemeint, sondern eine unendliche Anzahl an Lauten, die je-

doch aufgrund gemeinsamer charakteristischer Eigenschaften alle als [a] erkannt werden können.
Während Laute oder Phone an sich keine bedeutungsunterscheidende Funktion haben, bezeichnet man als Phonem die kleinste bedeutungsunterscheidende Einheit einer Sprache. Als Beispiel des Begriffs 'Phonem' soll das Wort 'Hase' dienen. Das Phon [a] in Hase kann laut oder leise, hoch oder tief ausgesprochen werden, die Bedeutung des Wortes ändert sich dadurch nicht. Wird hingegen das Phon [a] durch das Phon [o] ersetzt, ergibt sich plötzlich ein anderes Wort: 'Hose', die Bedeutung hat sich geändert. Die Phone [a] bzw. [o] sind bedeutungsunterscheidende Einheiten und bilden daher verschiedene Phoneme [1]. Phoneme werden durch die Begrenzungszeichen /../ dargestellt. Abhängig von der lautlichen Umgebung kann ein Phonem durch verschiedene Phone repräsentiert werden. Verschiedene Phone, die dasselbe Phonem repräsentieren, werden als Allophone bezeichnet. Als Beispiel sollen die Phone [R] (= akzentfreies Zäpfchen-r) und [r] (= Zungen-r) dienen. Sie bilden zusammen das Phonem /r/ und sind damit Allophone des Phonems /r/ [2]. Die Beschreibung von Phonen ist von der Sprache unabhängig, während Phoneme immer hinsichtlich einer bestimmten Sprache definiert werden, ebenso wie der Satz an distinktiven Merkmalen, die eine Sprache verwendet [3]. Als Beispiel soll das Phonem /p/ dienen, welches behaucht (aspiriert), d.h. mit einem geringen, aber hörbaren Luftausstoß (geschrieben als $[p^h]$) oder unbehaucht (geschrieben als $[p^=]$) gesprochen werden kann. Beide Versionen sind in der deutschen Sprache Allophone des Phonems /p/, d.h. die Aspiration ist im Deutschen kein distinktives Merkmal. In der chinesischen Sprache dagegen werden behauchte und unbehauchte Formen unterschieden, so daß hier die Phone $[p^h]$ und $[p^=]$ verschiedene Phoneme bilden. Die Lehre von den Phonen bezeichnet man als Phonetik, während die Phonologie (auch Phonematik oder Phonemik genannt) als lautliche Einheiten der Sprache die Phoneme annimmt. Die Sprachen der Welt verwenden zwischen 25 und 45 Phoneme [Str96].

2.2 Anregung und Lautformung

Die an der Sprachproduktion beteiligten Organe waren ursprünglich für andere Körperfunktionen angelegt, wie das Atmen, Essen und Riechen oder den Schutz der Luftröhre vor Speiseresten. Entwicklungsgeschichtlich betrachtet wurde die Sprache erst zu einem späten Zeitpunkt 'erfunden'. Sprache kann daher als eine überlagerte Funktion verstanden werden, da sie auf dem Atmen beruht und die Hauptaufgabe des Atmens die Aufnahme von Sauerstoff ist. Sprache wird durch ein Zusammenspiel von Lunge, Kehlkopf, Mund-, Rachen- und Nasenraum erzeugt (vgl. Abb. 2.1).

[1] Bemerkenswert an diesem Beispiel ist noch die Tatsache, daß beide Wörter sich durch genau ein Phonem unterscheiden. Beide Wörter bilden damit ein minimales Paar.

[2] Solche Allophone mit freien Varianten, die beliebig austauschbar sind, bezeichnet man als freie Allophone. Im Gegensatz dazu stehen sogenannte stellungsbedingte Allophone, welche nicht in derselben lautlichen Umgebung auftreten.

[3] Das Wörterbuch der Kognitionswissenschaft [Str96] gibt u.a. als weitere Definition für 'Phonem' auch 'Bündel distinktiver Merkmale' an.

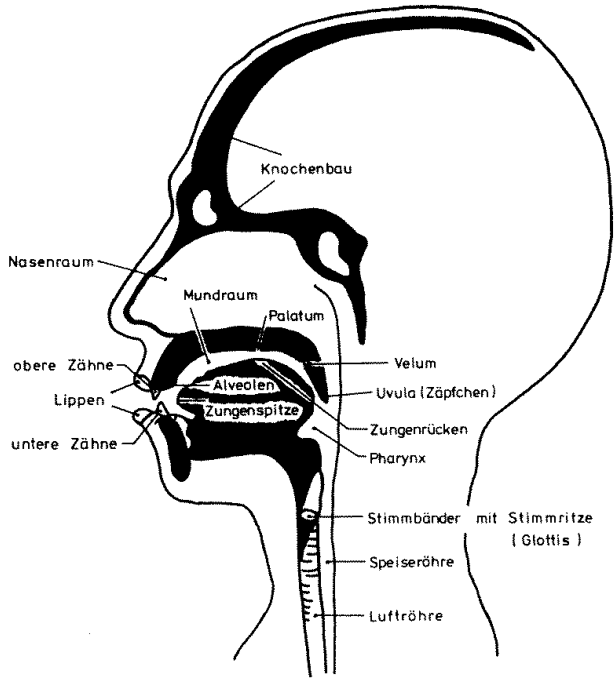

Abbildung 2.1: Sprechapparat des Menschen, nach [Fel84]

2.2.1 Anregung

Bei der Hervorbringung von Lauten (Vokalisieren) wird die Luft kontrolliert durch den Kehlkopf (*Larynx*), den Rachen (*Pharynx*) und den Mund-Nasenraum ausgestoßen. Die aus den Lungen ausströmende Luft wird im Kehlkopf durch die Stimmbänder in ein Anregungssignal umgeformt. Die eigentliche Lautformung (Artikulation) geschieht anschließend im Artikulationstrakt.
Bei der Erzeugung des Anregungssignals im Kehlkopf lassen sich verschiedene Fälle unterscheiden:
Ist der Spalt zwischen den Stimmbändern, die Stimmritze (*Glottis*), weitgehend geschlossen, so wird der aus der Lunge kommende Luftstrom vor den durch Muskelkraft angespannten Stimmbändern angestaut, bis der Druck so hoch ist, daß die Stimmbänder auseinandergedrückt werden und die Luft passieren kann. Dadurch sinkt der Druck bis sich die Stimmbänder wieder schließen und der Vorgang sich zu wiederholen beginnt. Es entsteht eine periodische Schwingung der Stimmbänder und eine periodische Druckwelle als Anregungssignal für die sogenannten stimmhaften Laute [4]. Die Frequenz des Anregungssignals wird als Sprachgrundfrequenz oder Grundfrequenz bezeichnet (oft abgekürzt mit F0) und liegt zwischen 80 Hz (tiefe Männerstimme) und 350 Hz (hohe Kinderstimme). Da die Grundfrequenz von der Anspannung der Stimmbänder abhängig ist (bei großer Anspannung ist sie höher, bei kleiner Anspannung niedriger), un-

[4]Nach [EH93] gibt es für die Entstehung des periodischen Anregungssignals bei stimmhaften Lauten allerdings auch eine Theorie, die auf eine aktive Anregung der Stimmbänder durch die Stimmbandmuskeln zurückgreift.

terliegt sie bei jedem Sprecher Änderungen, die sich ungefähr über eine Oktave erstrecken [O'S87]. Das Anregungssignal bei stimmhaften Lauten entspricht in etwa einer periodischen Dreiecksimpulsfolge mit sehr obertonreichem Spektrum, welches proportional zum Quadrat der Frequenz abnimmt. Auch die Lautstärke eines Tons ist nicht unabhängig von der Tonhöhe, da für die größeren Schwingungsamplituden eines lauteren Tons ein größerer Druck der angestauten Luft notwendig ist, der durch eine höhere Anspannung der Stimmbandmuskeln erzeugt wird, welches zu einer höheren Frequenz des Tones führt.
Ist die Stimmritze geöffnet und die Stimmbänder kaum angespannt, wird der Luftstrom der Lungen an den Kanten und Ritzen des Artikulationstraktes gebrochen, wodurch ein rauschförmiges Anregungssignal mit kontinuierlichem Spektrum entsteht, welches noch relativ hohe Energieanteile bei hohen Frequenzen besitzt. Die durch dieses Anregungssignal gebildeten Laute werden als stimmlos bezeichnet.
Ist die Stimmritze etwas geöffnet, entsteht als Anregungssignal eine Mischung aus stimmhaften und stimmlosen Signalen, da die Stimmbandränder noch in Schwingungen versetzt werden. Sie kann als 'gemischte' Anregung bezeichnet werden [MS91].

2.2.2 Lautformung

Die eigentliche Lautformung geschieht im Artikulations- oder Vokaltrakt, der im wesentlichen aus Rachenraum, Nasen- und Mundhöhle, Zunge, Zähnen und Lippen besteht. Der sich den Stimmbändern anschließende Luftraum bildet dabei einen Hohlraumresonator mit ausgeprägten Resonanzeigenschaften, der in hohem Maße veränderbar ist, z.B. durch die Stellung der Zunge oder der Lippen. Die Hauptresonatoren sind dabei der Rachen-, Mund- und Nasenraum [5]. Die Resonanzeigenschaften dieses Hohlraums beeinflussen das Spektrum des Anregungssignals, so daß bestimmte Frequenzen verstärkt, andere dagegen unterdrückt werden. Die Resonanzfrequenzen nennt man nach L. Hermann (1890) auch Formanten oder Formantfrequenzen. Die meisten Laute werden im Mundraum gebildet, wobei die Zunge eine herausragende Stellung einnimmt. Bei der Lautbildung müssen sich alle Artikulatoren schnell und gleichzeitig bewegen, was dem Sprecher eine komplexe motorische Fertigkeit abverlangt. Dabei ist vor allem die zeitliche Koordination der einzelnen Artikulatoren ein zu lösendes Problem.
Da die Sprachlaute nicht durch eine Abfolge getrennter Schritte gebildet werden, sondern immer im Zusammenhang mit den ihnen nachfolgenden oder vorangehenden Lauten, beeinflussen diese die Merkmale der einzelnen Laute. Z.B. ist die Lippenstellung bei der Äußerung des [k] in den Wörtern 'Kultur' und 'Kiste' aufgrund des nachfolgenden Vokals unterschiedlich. Diese Beeinflussung der Laute wird als Koartikulation bezeichnet.
Die Phonetik unterteilt die Laute in Konsonanten und Vokale. Als weitere Unterteilungskriterien werden zusätzlich noch Artikulationsart (wie wird die Luft

[5] Die unveränderlichen Teile des Vokaltraktes wie Nasennebenhöhlen oder die Luftröhre führen zu sprechertypischen Merkmalen in der Sprache, die bei der Sprechererkennung verwendet werden können.

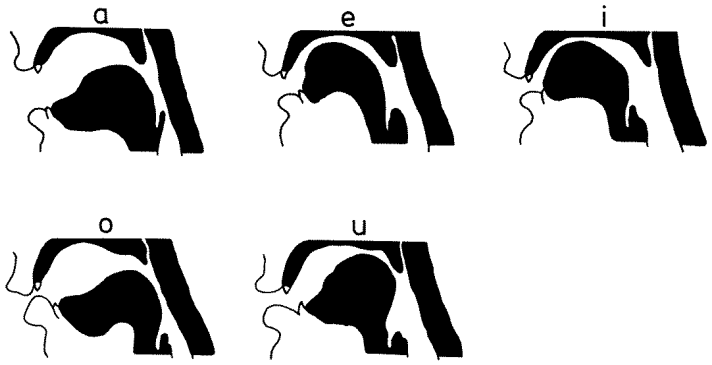

Abbildung 2.2: Vokalbildung in der Mundhöhle, nach [Fel84]

in ihrer Bewegung beeinflußt) und Artikulationsort (wo geschieht die Lautbildung, welche Organe sind daran beteiligt) verwendet.

Vokale

Alle Vokale werden durch eine stimmhafte Anregung des Vokaltraktes gebildet, wobei die Zunge den Gaumen nicht berührt (Öffnungslaute). Ist alleine die Mundhöhle bei der Bildung des Vokals beteiligt, bezeichnet man diese als Oralvokale [EH93]. Dabei kann man zwischen Vokalen unterscheiden, die mit offenem, halbgeschlossenem und fast geschlossenem Mund gesprochen werden (man vergleiche [a], [o] und [u]). Ebenfalls üblich ist eine Unterteilung der Vokale nach der Form, die die Lippen bei der Artikulation bilden. So wird ein [u] als gerundeter Vokal, ein [e] dagegen als ungerundet bezeichnet [Fel84]. Entweicht die Luft dagegen hauptsächlich durch den Nasenraum, bezeichnet man diese Vokale als Nasalvokale. Diese kommen z.B. in der französischen und polnischen Sprache vor. Auch die aus den Lautübergängen zwischen zwei Vokalen bestehenden Diphtonge werden zu den Vokalen gezählt. Diese entstehen dadurch, daß die Zunge (evtl. zusammen mit den Lippen) eine Gleitbewegung von einem Vokal zu einem anderen ausführt (z.B. [au] in Laub oder [ai] in Mais).

Vokale werden durch die Resonanzeigenschaften des Artikulationstraktes gebildet, was zu einer ausgeprägten Formantenbildung führt. Dabei spielt die Stellung der Zunge und der Lippen eine zentrale Rolle. Abbildung 2.2 zeigt die Mundhöhle während der Bildung einiger wichtiger Vokale. Man erkennt, daß die horizontale und vertikale Position des Zungenrückens wesentlich zur Unterscheidung der Vokale beiträgt. Ist die Position des Zungenrückens in der Nähe der Zähne (z.B. beim [i]), klingt der Vokal hell und wird auch als Vorderzungen- oder palataler Vokal bezeichnet. Beim [u] dagegen befindet sich der Zungenrücken in der Nähe des Zäpfchens und klingt dunkel. Man bezeichnet ihn dann als

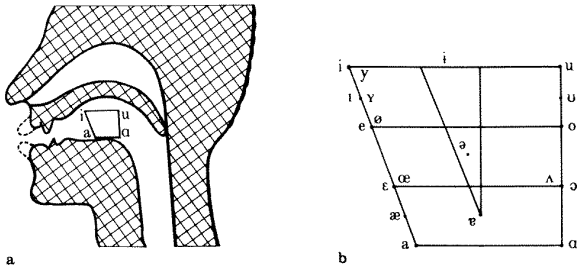

Abbildung 2.3: Vokalviereck, nach [Fel84]

Hinterzungen- oder velaren Vokal. Auch die Vertikalposition des Zungenrückens trägt zur Unterscheidung bei. Zusätzlich ist auch wie erwähnt die Form der Lippen, die Lippenrundung, für eine Beschreibung des Vokals wichtig [O'S87]. Die verschiedenen horizontalen und vertikalen Stellungen der Zunge werden häufig zum sogenannten Vokalviereck (vgl. Abb. 2.3) zusammengefaßt, dem man die 'Lage' der wichtigsten Vokale entnehmen kann.

Vokale haben meist eine hohe Signalenergie bei tiefen Frequenzen und sind maßgeblich an der Bildung der Satzmelodie beteiligt. Im Zeitverlauf zeigen sie eine periodische Signalform aufgrund des periodischen Anregungssignals und des offenen Vokaltraktes. Zur Unterscheidung der Vokale ist das Verhältnis der ersten zwei bis drei Formanten ausreichend, bei einer künstlichen Spracherzeugung wächst aber die Natürlichkeit der Sprache, wenn mehr Formanten verwendet werden. Die Lage der Formanten ist dabei von der Höhe der Grundfrequenz abhängig und kann stark variieren [6]. Menschen setzen beim Hören die Formanten in Beziehung zur Grundfrequenz und gleichen so Unterschiede aus. Das Sprachverstehen von Vokalen hängt aber nicht allein von der Frequenz der Formanten ab.

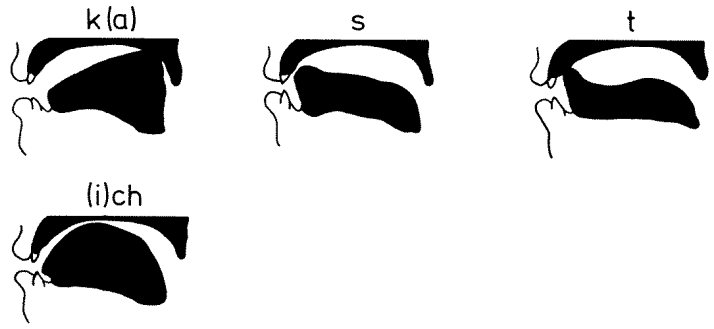

Abbildung 2.4: Konsonatenbildung in der Mundhöhle, nach [Fel84]

Konsonanten

Während die Erzeugung von Vokalen relativ einfach durch stimmhafte Anregung und die jeweilige Gestaltung und die daraus resultierenden Resonanzeigenschaften des Artikulationstraktes (insbesondere Zunge und Lippen) erklärt werden kann, spielen bei der Bildung von Konsonanten auch andere Artikulatoren wie Zähne und Gaumensegel eine Rolle. Zudem sind auch stimmlose Anregungssignale möglich.
Konsonanten werden in erster Linie durch Engstellen im Luftstrom gebildet. Unterschieden werden Konsonanten dabei durch die Artikulationsart und den Artikulationsort. Bei der Frage nach dem Artikulationsort stehen die Artikulatoren, die an der Lautbildung beteiligt sind, im Vordergrund, z.B. bei der Zunge der Ort, an dem sie sich bei der Artikulation befinden. Die Artikulationsart dagegen beschreibt die Art und Weise, wie die Artikulatoren dann die Lautbildung ausführen.
Als Artikulationsart versteht man demnach die Art und den Umfang, wie der Luftstrom auf seinem Weg behindert wird, z.B. durch Verschließen des Luftweges mit anschließender plötzlichen Öffnung oder Umlenkungen des Luftstroms. Man kann folgende Artikulationsarten unterscheiden:

- Verschlußlaute (Plosive)
  Dabei wird durch Gaumen, Zunge oder Lippen der Luftstrom so gestoppt, daß er weder durch Mund noch durch den Nasenraum entweichen kann. Nach einem Druckaufbau wird der Verschluß plötzlich geöffnet. Bei stimm-

[6]Dabei fällt die Frequenz des ersten Formanten (F1), umgekehrt zur Höhe des Zungenrückens (vgl. [a] zu [i]), während die Frequenz des zweiten Formanten (F2) von dem Verschieben der Zunge abhängig ist. So steigt F2 bei einer Vorwärtsbewegung der Zunge an ([o] zu [i]) [LE99][O'S87]. Das Verhältnis von F1 zu F2 wird oft als sogenanntes Vokaldreick dargestellt

haften Verschlußlauten (wie [b] und [d]) schwingen die Stimmbänder kurz an, stimmlose Verschlußlaute sind z.B. [p] und [t].

- Reibelaute (Frikative)
  Die Einengung des Luftstroms erfolgt hier im Mund- oder Rachenraum, so daß ein rauschförmiger Laut entsteht . Dieser kann stimmhaft ([j]) oder stimmlos ([f]) sein.

- Nasale
  Hierbei werden wie bei den Verschlußlauten Verschlüsse im Mundraum erzeugt, allerdings bleibt der Nasenraum geöffnet, so daß hier die Luft entweichen kann. Die deutschen Nasale (z.B. [m] und [n]) sind stimmhaft.

- Seitenlaute (Laterale)
  Die Zunge blockiert hierbei die Mittelpassage der Mundhöhle, so daß die Luft seitlich der Zunge vorbeiströmt. Ein Beispiel ist [l].

- Schwinglaute (Vibranten)
  Durch den Luftstrom werden die Zungenspitze oder das Zäpfchen zum Schwingen angeregt. Beispiele sind das Zungenspitzen-[r] und das Zäpfchen-[R].

Bei der Einteilung der Konsonanten nach dem Artikulationsort interessiert wie erwähnt nicht nur der Ort, an dem die Lautbildung stattfindet, sondern auch die Artikulatoren wie z.B. Lippen, Zunge und Zähne, die daran beteiligt sind. Man unterscheidet folgende Artikulationsorte bei der Lautbildung:

- bilabial:
  durch Ober- und Unterlippe. Beispiele: [m],[b] und [p]

- labiodental:
  durch Unterlippe und obere Schneidezähne. Beispiele: [f] und [v]

- dental:
  durch Zungenspitze und obere Schneidezähne. Beispiele: [s] und [t]

- interdental:
  die Zungenspitze wird zwischen oberen und unteren Zähnen hervorgestreckt. Beispiel: das englische [θ]

- alveolar:
  zwischen Zungenspitze und oberem Zahnfächer (*Alveolen*). Beispiel: [d]

- palatal:
  zwischen Zunge und hartem Gaumen (*Palatum*). Beispiel: [j]

- velar:
  zwischen Zunge und weichem Gaumen (*Velum*). Beispiel: [k]

- uvular:
  durch die Zunge und Zäpfchen (*Uvula*). Beispiel: Zäpfchen-[R]

- pharyngeal:
  durch Verengen der Rachenhöhle

- glottal:
  in der Stimmritze (*Glottis*) dadurch, daß diese nur zum Teil geschlossen ist und so eine rauschförmige Anregung an den Stimmbändern erzeugt wird. Beispiel: [h]

Eine Einteilung der deutschen Konsonanten nach Artikulationsart und -ort sowie Stimmhaftigkeit findet sich z.B. bei [EH93].
Konsonanten stellen im Gegensatz zu den Vokalen eine sehr heterogene Gruppe dar. So besitzen einige stimmhafte Konsonanten wie die Vokale ausgeprägte Formanten, während andere nicht statisch sondern nur durch den zeitlichen Verlauf der Frequenzanteile beschrieben werden können (z.B. Plosive) und wiederum andere zwar einen statischen Frequenzverlauf aufweisen, dieser jedoch nicht aus unterscheidbaren Formanten sondern aus quasi 'verwischten' Frequenzanteilen z.B. im oberen Frequenzband (Reibelaute) besteht.
Nasale besitzen einen ausgeprägten Formanten im niedrigen Frequenzbereich, die anderen Formanten sind nur schwach. Da Nasale durch die Kopplung der Resonatoren des Mund- und Nasenraums gebildet werden, können dabei sogar Antiformanten (Frequenzbereiche mit keinen oder nur geringen Energieanteilen) auftreten.
Plosive können stimmhaft (z.B. [b,d,g]) oder stimmlos ([p,t,k]) gebildet werden. Sie können anhand der Zeitdauer unterschieden werden, die der Mund verschlossen bleibt sowie der Zeitdauer, die zwischen dem Öffnen des Mundes und dem Einsetzen der Schwingung der Stimmbänder liegt. Diese nennt man voice onset time (VOT). Als Beispiel dafür sollen die Phoneme /b/ und /p/ genannt werden. Fällt der Zeitpunkt, an dem die Stimmbänder zu schwingen beginnen, mit dem Auslösen des Verschlußlautes zusammen, wird ein /b/ gehört, hinkt der Voice Onset um ca. eine Zentelsekunde hinterher, wird ein /p/ verstanden.
In Untersuchungen, die den VOT bei synthetischer Spracherzeugung in kleinen Zeitabständen (0.01s) variierten, konnte man feststellen, daß bei einer zufälligen Präsentation die Versuchspersonen fast alle Stimuli fast immer genau einer Kategorie (/p/ oder /b/) zuordneten, Unsicherheiten herrschten nur in einem extrem schmalen Bereich von ca. 0.05 s Breite, in dem die Stimuli etwa zu gleichen Teilen zweifelsfrei einer der beiden Kategorien zugeordnet wurden (Phonemgrenze). Dieses wird als kategorische Wahrnehmung bezeichnet und ist ein wichtiges Merkmal der menschlichen Sprachwahrnehmung.
Frikative haben ein rauschförmiges Spektrum, welches im oberen Frequenzbereich über 4000 Hz liegt. Während einige Frikative aufgrund dieses Spektrums unterschieden werden können, sind für die Erkennung anderer die Formantenübergänge von bzw. zu benachbarten Lauten maßgeblich.
Auch bei den sogenannten Halbvokalen [w,r,l,j] spielen die Formantenübergänge, genauer gesagt deren Dauer, eine wesentliche Rolle. Ist z.B. die Übergangsdauer mit denselben Formanten kürzer als 150 ms, so wird anstelle eines [u] ein [w] , für noch kürzere Zeiten (<50 ms) ein [b] gehört. Gegenüber Störungen sind die Konsonanten unterschiedlich empfindlich, stimmhafte Laute und Nasale sind dabei eher unempfindlich.

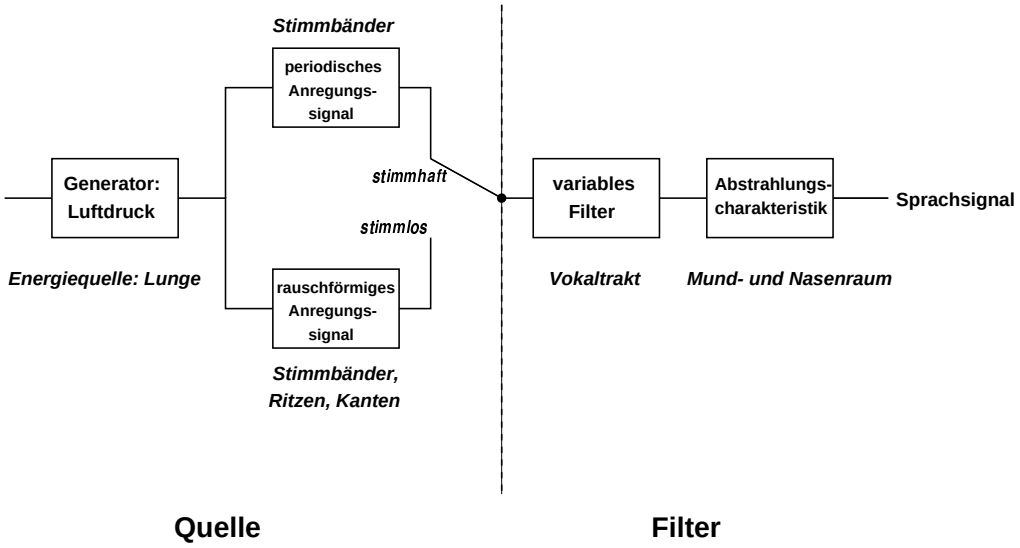

Abbildung 2.5: Schematische Darstellung des Quelle-Filter-Modells

2.2.3 Quelle-Filter-Theorie

Die Quellen-Filter-Theorie der Lautproduktion, vom deutschen Physiologen Johannes Müller (1801-1858) vorgeschlagen, versucht, die Sprachproduktion beim Menschen durch ein relativ einfaches Model nachzubilden, welches aus der Aneinanderreihung verschiedener Komponenten besteht (vgl. Abb. 2.5).

Die durch den Luftstrom der Lunge (= Energiequelle) in Schwingungen versetzten Stimmbänder werden dabei als Quelle der Lautproduktion betrachtet. Die Resonatoreigenschaften des Vokaltraktes und die Abstrahlungscharakteristik des Mund- und Nasenraums werden durch einen variablen Filter beschrieben, welcher aus dem Kehlkopfton der Quelle das endgültige Ausgabespektrum erzeugt. Dieses Modell kann den tatsächlichen Vorgang der Sprachproduktion nur unvollständig beschreiben, da es die Kopplungen zwischen der Quelle und dem Filter nicht berücksichtigt. So verändert sich z.B. durch die Auf- und Abbewegung des Kehlkopfes beim Sprechen die Größe des Vokaltraktes und des noch unterhalb der Stimmbänder liegenden Resonanzraumes. Auch die Rückkopplung des Sprachsignals über das Gehör zum Gehirn findet in diesem Model keine Beachtung. Trotzdem findet das Quelle-Filter-Model auch heute noch bei der Sprachanalyse und -synthese weite Verwendung.

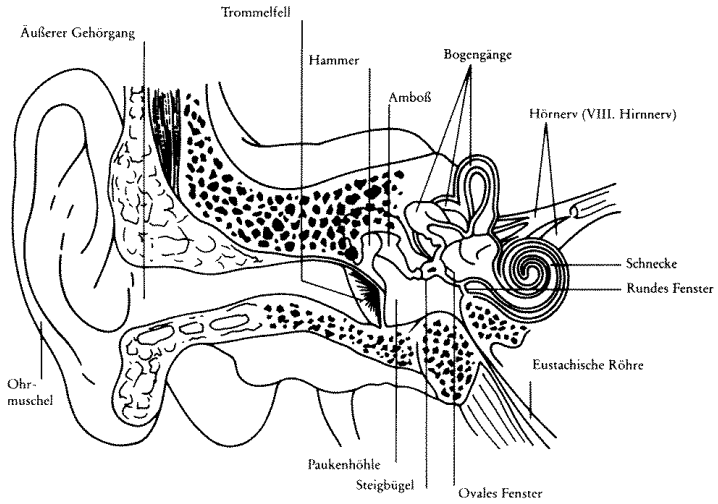

Abbildung 2.6: Aufbau des menschlichen Ohrs, nach [Cry95]

2.3 Das Ohr

Der Hörvorgang vollzieht sich in mehreren Stufen. Zunächst wird der Schall aufgenommen und weitergeleitet. Anschließend erfolgt eine Verstärkung und Anpassung, bis sich schließlich im Innenohr die Transformation der akustischen Reize in elektrische Aktionspotentiale und deren neuronale Weiterleitung vollzieht. Der endgültige Höreindruck entsteht erst im Gehirn.
Neben dem Hören dient das Hörsystem auch der Schallokalisation, wobei die Laufzeitunterschiede der Schallwellen, die zu verschiedenen Zeitpunkten an beiden Ohren eintreffen, verwendet werden.
Der Hörbereich des Menschen liegt zwischen 20 Hz und 20 kHz, wobei die größte Empfindlichkeit den Bereich von 200-5600 Hz umfaßt, in dem auch die für die Spracherkennung wichtigsten Frequenzen liegen. Der wahrnehmbare Lautstärkebereich erstreckt sich zwischen 0 dB (Hörschwelle) [7] und 120 dB (Schmerzgrenze). Die Wahrnehmung der Lautstärke eines Tones ist dabei nicht nur vom Schalldruck sondern auch von der Frequenz des Tones abhängig. Das Gehör erreicht eine zeitliche Auflösung, die es z. B. erlaubt, Zeitverzögerungen von 11 μs zwischen zwei klickenden Lauten wahrzunehmen. Sehr leise Töne können schon 10 ms nach Abklingen eines sehr lauten Tones wahrgenommen werden, d.h. die Adaptationsgeschwindigkeit des Gehörs ist sehr hoch.
Der Aufbau des Ohres läßt sich anatomisch in drei verschiedene Bereiche auf-

[7] Dabei liegt der Schalldruckpegel nur wenig über dem durch die Bewegung der Luftmoleküle erzeugten thermischen Rauschen.

gliedern, welche verschiedene Aufgaben beim Hörvorgang übernehmen: Das Außen-, Mittel- und das Innenohr. Abb. 2.6 zeigt einen Schnitt durch das menschliche Ohr.
Das Außenohr umfaßt die Ohrmuschel und den Gehörgang und endet am Trommelfell. Die Ohrmuschel (*Aurikula*) dient zur Weiterleitung der Schallwellen in den Gehörgang und hilft bei der Schallokalisation, indem sie von vorne kommende Geräusche verstärkt, während rückwärtige Geräusche abgeschwächt werden. Der Gehörgang besteht im wesentlichen aus einem ca. 2,5 cm langen Rohr, welches an einem Ende durch das Trommelfell verschlossen wird. Er wirkt als Hohlraumresonator mit einem Resonanzbereich von ca. 2 bis 4 kHz, in dem die Verstärkungs- (und damit die Hörleistung) am größten ist.
Das Mittelohr besteht aus einem kleinen luftgefüllten Raum, der Paukenhöhle. Der Druckausgleich, der für eine freie Beweglichkeit des Trommelfells zwischen der Paukenhöhle und dem Gehörgang notwendig ist, wird durch die sogenannte Eustachische Röhre gebildet, die eine Verbindung zwischen Paukenhöhle und Nasen-Rachen-Raum herstellt. Das Mittelohr dient der Übertragung der Trommelfellschwingungen auf das Innenohr. Dabei findet eine Druckverstärkung und Impedanzanpassung an den höheren Wellenwiderstand der Flüssigkeit im Innenohr statt, ohne die es zu einer massiven Reflexion der Schallenergie kommen würde. Für den Frequenzbereich von 1 bis 3 kHz ist diese Anpassung am besten. Zudem hat das Mittelohr die Aufgabe, die empfindlichen Sinnesrezeptoren des Innenohrs vor sehr lautem Schall zu schützen. Bewerkstelligt werden alle drei Aufgaben durch die Gehörknöchelchen Hammer (*Malleus*), Amboß (*Incus*) und Steigbügel (*Stapes*) (vgl. Abb. 2.6). Der Hammer ist am Trommelfell befestigt und überträgt dessen Schwingungen auf den Amboß, der diese wiederum an den Steigbügel weitergibt. Der Steigbügel ist mit dem sogenannten ovalen Fenster verbunden, durch das die Schallschwingungen an das Innenohr weitergeleitet werden. , so daß über 90 % davon für den Hörvorgang verloren gingen.
Das Innenohr liegt im Knochen des Felsenbeins und enthält die Schnecke (*Cochlea*) und die Bogengänge, die für den Gleichgewichtssinn des Menschen zuständig sind. Im Innenohr findet die Transformation der Schalleinwirkung in elektrische Aktionspotentiale statt, die vom Gehirn weiterverarbeitet werden können. Im Gegensatz zu Außen- und Mittelohr sind die Kenntnisse über das Innenohr noch immer lückenhaft.

Abb. 2.7 zeigt die für den Hörvorgang wichtigen Teile des Innenohrs. Die Schnecke (*Cochlea*) hat eine Länge von ca. 32mm und ist in 2 $\frac{1}{2}$ Windungen aufgewickelt. Im Schnittbild erkennt man, daß die Schnecke in drei Abteilungen untergliedert werden kann, die drei parallele Kanäle bilden: Die Vorhoftreppe (*scala vestibuli*), die Paukentreppe (*scala tympani*) und der dazwischen liegende Schneckengang (*ductus cochlearis*, auch *scala media* genannt). Zwischen Vorhof- und Paukentreppe existiert an der Schneckenspitze eine Verbindung in Form einer kleinen Öffnung (*Helicotrema*). Beide sind mit derselben Flüssigkeit (*Perilymphe* gefüllt, während der Schneckengang die chemisch unterschiedliche *Endolymphe* enthält. Die Paukentreppe wird zum Mittelohr hin durch eine feine Membran verschlossen, das runde Fenster, das dem Druckausgleich dient. Der Schneckengang wird durch die Reissnersche Membran von der Vorhoftreppe getrennt, zwischen der Paukentreppe und dem Schneckengang liegt die Basi-

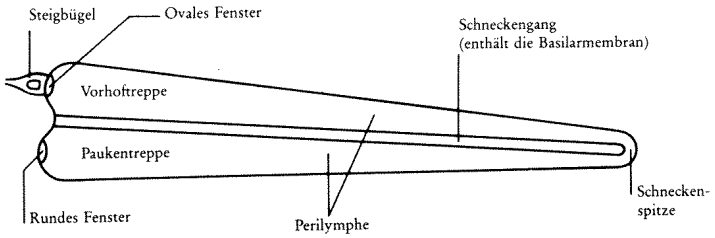

Abbildung 2.7: Schematische Darstellung der aufgerollten Schnecke, nach [Cry95]

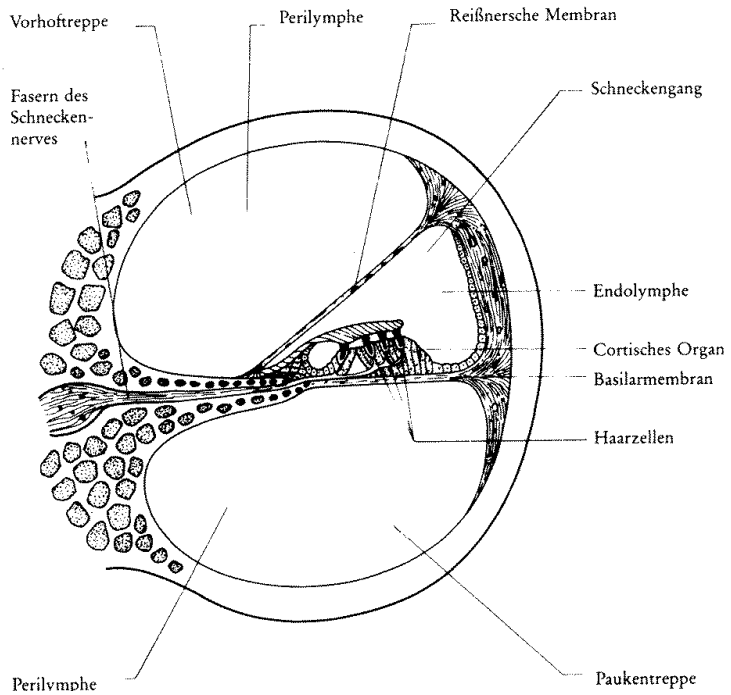

Abbildung 2.8: Querschnitt durch eine Schneckenwindung, nach [Cry95]

larmembran, oberhalb derer sich das Cortische Organ befindet (vgl. Abb. 2.8). Die Basilarmembran ist in der Nähe des ovalen Fensters nur schmal und straff gespannt, während sie sich zur Schneckenspitze hin verbreitert und schlaffer wird.
Das eigentliche Hörorgan wird vom Cortischen Organ gebildet, welches die Haarzellen enthält (vgl. Abb. 2.8). Diese besitzen an ihrer Oberfläche feine Härchen (*Stereocoliten*) und stehen in synaptischem Kontakt mit auditorischen Nervenfasern, die wiederum mit dem Gehörnerv (*nervus acusticus*) verbunden sind. [8] Man unterscheidet die äußeren und die inneren Haarzellen. Erstere sind in drei Reihen angeordnet und weitaus zahlreicher (ca. 25.000 Zellen), während letztere (ca. 3500) nur eine Reihe bilden. Beide Gruppen werden getrennt mit Nervenfasern versorgt, wobei etwa 95 % der Nervenfasern mit den inneren Haarzellen in Verbindung stehen, während die äußeren Haarzellen mit dem restlichen Anteil an Nervenfasern verbunden sind. Es wird angenommen [MS91], daß die inneren Haarzellen der Frequenzanalyse und einer Art Kontrastverstärkung dienen, während die äußeren Haarzellen eine Lautstärkeadaptation vornehmen [Mae90]. Das Cortische Organ wird durch die Tectorialmembran bedeckt, welche auf den Spitzen der Haarzellen aufliegt (vgl. Abb. 2.8). Durch Bewegungen der Basilarmembran relativ zur Tectorialmembran und den Haarzellen wird durch die Abbiegung der Haarzellen ein Reiz der entsprechenden Nervenfasern erzeugt.
Die Schallwellen werden vom Mittelohr aus vom Steigbügel durch das ovale Fenster auf das Innenohr übertragen und regen dessen Flüssigkeitssystem zum Schwingen an. Der Druck wird von der *scala vestibuli* über die *scala media* zur *scala tympani* weitergeleitet. Über die Art und Weise, wie nun die Bewegungen der Basilarmembran die verschiedenen Frequenzen der Schallwellen zu den entsprechenden Hörreizen überführen, herrscht allerdings noch keine vollkommene Klarheit. Gemeinhin anerkannt ist die Theorie von G. von Békésy, nach der sich durch die Bewegung des Steigbügels eine fortschreitende Welle (Wanderwelle) auf der Basilarmembran in Richtung *Helicotrema* ausbreitet. Die Einhüllende dieser Wanderwelle hat ein relativ scharfes Maximum, dessen Lage von der Frequenz abhängig ist. Durch dieses Maximum wird eine besonders starke Erregung des Cortischen Organs erzeugt, so daß die Frequenz der Schwingung durch den Ort der Erregung kodiert wird. Hohe Frequenzen werden in der Nähe des ovalen Fensters, tiefe in der Nähe des Helicotremas abgebildet. Durch die relativ scharfen Maxima der Einhüllenden ist ein hohe Frequenzauflösung möglich.
Mit Hilfe psychoakustischer Experimente, z.B. mit Maskierungsexperimenten (vgl. [ZF67][O'S87]) konnte festgestellt werden, daß das Gehör bei der Wahrnehmung von Klängen diese in Frequenzgruppen wahrnimmt, die allerdings keinen festen Platz auf der Frequenzskala einnehmen. Vielmehr ist die Stelle, an der sie das Gehör bildet, vom wahrgenommenen Schall selbst abhängig. Die Wahrnehmung eines Klanges ändert sich abrupt, sobald dieser Frequenzen enthält, die außerhalb der Bandbreite einer Frequenzgruppe liegen. Eine solche Bandbreite wird als kritische Bandbreite bezeichnet und ist abhängig von

[8] Eine genaue Beschreibung der Reizweiterleitung und Umwandlung in einen Nervenimpuls kann z.B. [Ade87] entnommen werden.

der Mittenfrequenz der entsprechenden Frequenzgruppe. Für Mittenfrequenzen kleiner als 1 kHz beträgt die Bandbreite etwa 100 Hz, für höhere Frequenzen steigt die kritische Bandbreite in etwa logarithmisch mit der Frequenz auf Werte bis zu 700 Hz (bei 4kHz), d.h. die Auflösung bei niedrigeren Frequenzen ist höher. Insgesamt läßt sich daher das wahrgenommene Frequenzspektrum des Gehörs in 24 Frequenzgruppen zerlegen [Zwi61]. Die so gewonnene Skala wird als Bark-Skala [9] bezeichnet.
Als weitere Skala soll noch die Mel-Skala erwähnt werden. Die Mel-Skala benutzt ein ähnliches Maß wie die Bark-Skala, nur ist die Abbildung näherungsweise linear für Frequenzen bis zu 1 kHz und logarithmisch für höhere Frequenzen (vgl. [O'S87]).

[9] nach Heinrich Barkhausen, der sich als erster mit Messungen der Lautstärkeempfindung beschäftigte

Kapitel 3

Mustererkennung und Neuronale Netze

In diesem Kapitel sollen die Grundlagen der Mustererkennung besprochen werden, bevor Neuronale Netze als eine Realisierungsmöglichkeit von Klassifikatoren eingeführt werden. Zunächst wird kurz auf die historische Entwicklung und die biologische Motivation bei der Entwicklung der Neuronalen Netze eingegangen, bevor deren Grundlagen und zwei verschiedene Vertreter von Neuronalen Netzen näher besprochen werden.

3.1 Mustererkennung

Bei einem Mustererkennungssystem handelt es sich um ein System, das in der Lage ist, ihm zugeleitete Daten festgelegten Kategorien zuzuordnen. Solche Systeme werden in einem breitem Aufgabenspektrum verwendet, welches von der Handschriftenerkennung bis zur Sprach- und Bildverarbeitung reicht (vgl. zum Beispiel [EH93], [DH73]).
Abbildung 3.1 zeigt den prinzipiellen Aufbau eines Mustererkennungssystems (vgl. auch [SK92]). Es besteht aus einem Sensor, einer Vorverarbeitungseinheit, einer Merkmalsextraktionseinheit und einem Klassifikator. Dabei können verschiedene Einheiten u. U. auch zusammengefaßt werden, z.B. die Vorverarbeitungsmit der Merkmalsextraktionseinheit.
Die Arbeitsweise eines solchen Systems soll anhand eines Beispiels aus der Handschriftenerkennung erläutert werden. Die Aufgabe bestehe darin, Abbildungen der beiden Buchstaben X und D voneinander zu unterscheiden. Zunächst werden die Abbildungen z. B. mit Hilfe eines Scanners eingelesen. Die Vorverar-

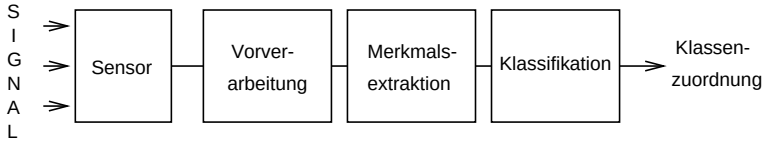

Abbildung 3.1: Prinzipieller Aufbau eines Klassifikationssystems

beitungseinheit kann nun die Aufgabe haben, eventuell existierende Schmutzeffekte in den eingelesenen Abbildungen zu eliminieren, Kanten zu verstärken oder auch den Umfang der Abbildung durch Zusammenfassung von Pixelgebieten bestimmter Größe zu verringern oder ein Graustufenbild in ein Binärbild umzuwandeln. Die Merkmalsextraktionseinheit reduziert anschließend die Bilder auf für das Erkennungsproblem signifikante Merkmale. Im vorliegenden Fall könnten die Merkmale z. B. aus der Lage und Form der Ecken, Kanten oder Rundungen im Bild bestehen, aber es ist auch möglich, die Graustufen- bzw. Binärwerte der einzelnen Pixel des Bildes selbst als Merkmale zu betrachten. Die anschließende Klassifikationseinheit hat die Aufgabe, anhand des Merkmalsatzes die Kategorie (auch Klasse genannt) zu bestimmen, zu der das Bild gehört. Sie soll im folgenden mit C bezeichnet werden. Im vorliegenden Beispiel soll eine Zuordnung zu den Klassen 'X' oder 'D' vorgenommen werden.
Um die Funktionsweise eines Klassifikators besser beschreiben zu können, soll das Konzept des Musterraums einführt werden. Alle von der Merkmalsextraktionseinheit gelieferten Mermale können zu einem sogenannten Merkmalsvektor zusammengefaßt werden, der oft auch als Mustervektor bezeichnet wird. Dabei bezeichnet jede Komponente des Mustervektors ein bestimmtes Merkmal, dessen Ausprägung als Zahlenwert kodiert wird. Werden bespielsweise die einzelnen Pixel einer Abbildung des Buchstabens als Merkmale gewählt, dann enthält jede Mustervektorkomponente den Graustufenwert eines bestimmten Pixels als Merkmal, die Länge des Mustervektors entspricht der Anzahl der Pixel des Bildes. Jeder Mustervektor kann durch einen Punkt in einem n-dimensionalen Musterraum dargestellt werden, wobei n die Länge des Mustervektors bezeichnen soll. Den einzelnen Achsen des Musterraums können die einzelnen Merkmale zugeordnet werden.
Um nun jedem Mustervektor eine Klasse zuordnen zu können, wird der Musterraum vom Klassifikator in Regionen eingeteilt, in denen jeweils nur Muster einer bestimmten Klasse liegen. Ein Muster kann dann aufgrund seiner Zugehörigkeit zu einer bestimmten Region klassifiziert werden. Die Entscheidungsregionen im Musterraum können durch reellwertige Funktionen $g_i(\mathbf{x})$ definiert werden, die auch als Diskriminierungsfunktionen bezeichnet werden und so gewählt werden, daß die Bedingung

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall \, j \neq i \tag{3.1}$$

erfüllt ist, wenn das Muster $\mathbf{x}$ der Klasse C_i zugeordnet werden soll. Für den Fall daß $g_i(\mathbf{x}) = g_j(\mathbf{x})$ gilt, kann keine Aussage zu einer Klassenzugehörigkeit getroffen werden. Dieser Fall definiert im Musterraum eine Entscheidungsgrenze. Für den Fall, daß $g_i(\mathbf{x})$ eine lineare Funktion ist, ist die Entscheidungsgrenze eine Hyperebene.
Eine Trennung zweier Musterklassen durch nur eine Hyperebene im Musterraum ist nicht immer möglich, wie die Abb. 3.2 zeigt. In diesem Fall muß entweder auf nichtlineare Funktionen als Diskriminierungsfunktionen zurückgegriffen werden, oder es muß mehr als eine Hyperebene verwendet werden (vgl. Abb 3.2). Eine andere Möglichkeit besteht darin, ein weiteres geeignetes Merkmal für die Klassifikation zu finden, so daß die Muster im nun n+1-dimensionalen Musterraum linear separabel sind, d.h. die Muster können durch eine Hyper-

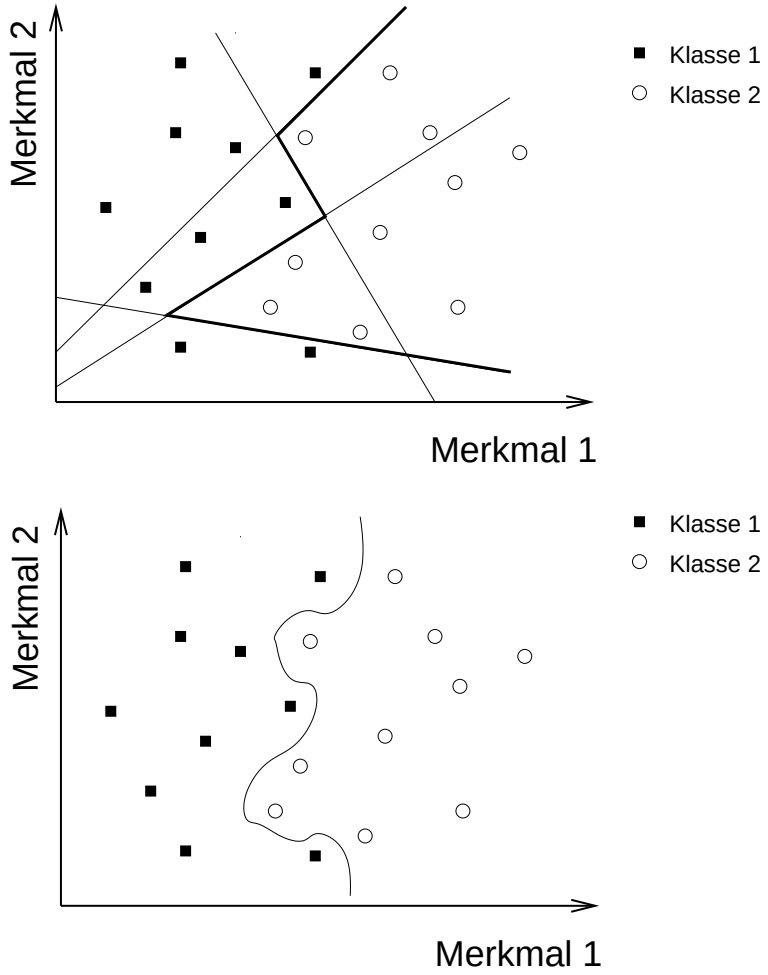

Abbildung 3.2: Aufteilung eines zweidimensionalen Musterraums durch mehrere Hyperebenen bzw. eine nichtlineare Diskriminante

ebene voreinander getrennt werden.
Man erkennt, daß zwischen der Wahl der Merkmale und des Klassifikators ein Zusammenhang besteht: Falls die Merkmalsextraktionseinheit gute Merkmale liefert, reicht zur Klassifikation eine einfache Diskriminierungsfunktion aus. Sind die besten Merkmale aber nicht bekannt, was i. a. die Regel ist, müssen an die Klassifikationseinheit höhere Anforderungen gestellt werden.
Ein Klassifikator wird durch einen Satz von Parametern beschrieben. Besteht der Klassifikator z.B. aus einer Hyperebene im Musterraum, die zwei Musterklassen voneinander trennt, so kann dieser Klassifikator durch die die Hyperebene beschreibenden Koordinaten dargestellt werden. Um die Parameter eines Klassifikators zu bestimmen, werden i.a. Beispielmuster mit bekannter Klassenzugehörigkeit verwendet, die auch als Trainingsmuster bezeichnet werden. Klassifikatoren, die solche Trainingssätze mit bekannten Klassenzugehörigkeiten für die Aufteilung des Musterraums in Entscheidungsregionen verwenden, werden auch als überwachte Klassifikatoren bezeichnet. Dabei kann man verschiedene Ansätze unterscheiden.
Bei den adaptiven Klassifikatoren werden zunächst die Entscheidungsgrenzen willkürlich in den Musterraum hineingelegt. Anschließend wird überprüft, welche Trainingsmuster durch diese Einteilung falsch klassifiziert werden und die Entscheidungsgrenzen dementsprechend verschoben. Ein Beispiel für einen solchen Klassifikator bietet das Feedforwardnetzwerk, welches z.B. mit dem Backpropagation-Algorithmus den Trainingsdaten angepaßt werden kann (vgl. Kapitel 3.3.1, [Bis95]). Die Anweisung, nach der die Entscheidungsgrenzen in Abhängigkeit der falsch klassifizierten Muster verschoben werden, wird auch als Lernregel bezeichnet. Solche Klassifikatoren können auch als nichtdeterministische Klassifikatoren bezeichnet werden, da bei ihnen die endgültige Lage der Entscheidungsgrenzen nicht vorherbestimmt ist, sondern z.B. von der Anfangslage der Entscheidungsgrenzen, von der gewählten Lernregel und evtl. auch von der Reihenfolge, in der die Muster bei der Verschiebung der Entscheidungsgrenzen berücksichtigt wurden, abhängig ist.
Statistische Klassifikatoren benutzen dagegen statistische Werte der Trainingsmuster, wie z.B. Mittelwerte der Trainingsmuster einer Klasse, so daß mit der Auswahl der Trainingsmuster der Klassifikator festgelegt ist. Solche Klassifikatore können daher auch als deterministisch bezeichnet werden. Ein Beispiel für einen deterministischen Klassifikator bietet der k-Nearest-Neighbor-Klassifikator '(vgl. [The89], [Han81], [Nie74]). Hierbei werden alle Trainingsmuster direkt dazu benutzt, ein Testmuster zu klassifizieren, indem die k zum Testmuster im Musterraum nächsten Trainingsmuster bestimmt werden. Das Testmuster wird anschließend der Klasse zugerechnet, die unter diesen k nächsten Nachbarn am häufigsten repräsentiert wird.
Zu erwähnen bleiben die sogenannten unüberwachten Verfahren. Bei diesen werden Muster unbekannter Klassenzugehörigkeit im Musterraum in verschiedenen Clustern zusammengefaßt, wobei man sich z.B. die Dichte der Muster im Musterraum zunutze macht. Das Ergebnis ist eine Strukturierung des Musterraums in Gebiete, wobei allerdings nicht sichergestellt werden kann, daß nur Muster einer Klasse in einem Gebiet liegen, da zur Einteilung der Gebiete die tatsächliche Klassenzugehörigkeit der Muster nicht herangezogen werden kann.

Solche Verfahren können aber z.B. dazu dienen, die Anzahl der möglichen Klassen abzuschätzen.
Abschließend soll kurz auf zwei Probleme hingewiesen werden, die im Zusammenhang mit realen Problemstellungen oftmals auftreten.
Die Hinzunahme neuer Merkmale um eine lineare Separierbarkeit der Daten zu erreichen, wie auf S. 22 beschrieben, und die damit verbundene Vergrößerung der Dimensionalität des Musterraums hat nicht nur Vorteile. So läßt sich oftmals feststellen, daß ab einem bestimmten Punkt die Hinzunahme neuer Merkmale zu einem schlechteren Klassifikationsergebnis führt, was in der Literatur als 'curse of dimensionality' bezeichnet wird (vgl. [Bis95]). Dies liegt daran, daß die Anzahl an Trainingsmustern i. a. begrenzt ist, so daß bei Vergrößerung der Dimensionalität die Verteilung der Trainingsmuster im Musterraum nur noch spärlich ist und damit keine gute Repräsentation der zu lernenden Musterverteilung mehr darstellt. Eine Möglichkeit, diese Schwierigkeit zu umgehen, liegt darin, die Muster vor dem Training eines Klassifikators einer Dimensionsreduktion zu unterziehen (vgl. Kapitel 3.1.1).

3.1.1 Lineare Diskriminanzanalyse

Ein Verfahren zur Dimensionsreduktion der Daten besteht in der linearen Projektion der Datenpunkte auf einen Unterraum. Ziel muß es dabei sein, eine möglichst gute Trennbarkeit der projezierten Datenpunkte zu gewährleisten. Die Lineare Diskriminanzanalyse (oft auch als Fishers Lineare Diskriminante bezeichnet) erhebt den Anspruch, genau dieses zu bewerkstelligen. Dazu werden die Klassenzugehörigkeiten der Datenpunkte verwendet, im Gegensatz zur Hauptkomponentenanalyse.
Das Verfahren soll am Beispiel von Datenpunkten $\mathbf{x}$ erläutert werden [1], die zu zwei verschiedenen Klassen C_1 und C_2 gehören (Die Anzahl der Muster für die Klassen betrage N_1 bzw. N_2). Dazu wird eine Projektion der Datenpunkte auf einen eindimensionalen Unterraum betrachtet:

$$y = \mathbf{w}^t \mathbf{x} \tag{3.2}$$

Ziel ist es nun, dasjenige $\mathbf{w}$ zu finden, durch das eine maximale Trennbarkeit der beiden Klassen im eindimensionalen Unterraum erreicht wird. Eine Möglichkeit, die Trennbarkeit der beiden Klassen im Unterraum zu bestimmen, liegt darin, die Differenz der projezierten Mittelwerte der beiden Klassen zu betrachten:

$$\tilde{m}_2 - \tilde{m}_1 = \mathbf{w}^t(\mathbf{m}_2 - \mathbf{m}_1) \tag{3.3}$$

mit

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{x \in C_i} \mathbf{x}^n \tag{3.4}$$

als den Klassenmitteln der ursprünglichen Datenpunkte aus Klasse C_i und

$$\tilde{m}_k = \mathbf{w}^t \mathbf{m}_k \tag{3.5}$$

[1] Die Darstellung richtet sich dabei im wesentlich nach [Bis95].

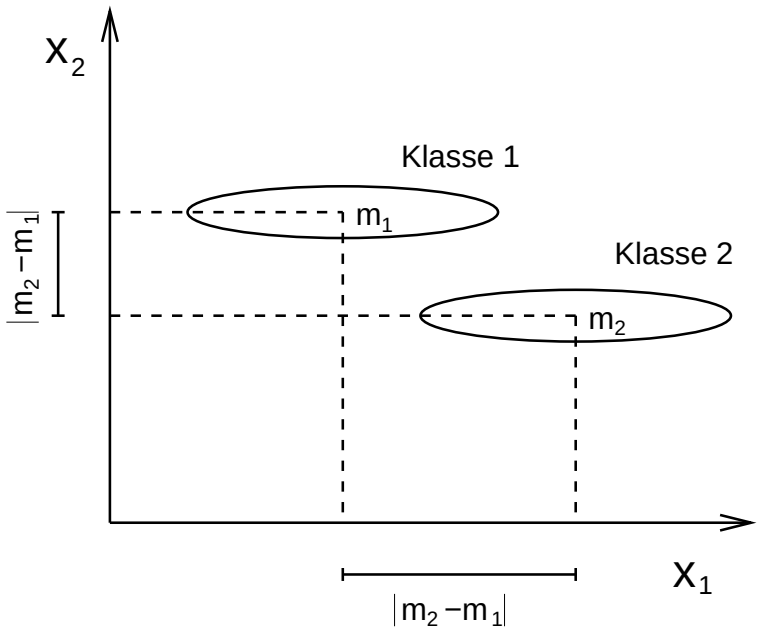

Abbildung 3.3: Projektion zweidimensionaler Muster auf eindimensionale Unterräume

als dem Klassenmittel der projezierten Daten von Klasse C_k. Gleichung 3.3 soll nun maximiert werden, wobei die Differenz aber schon allein durch Vergrößerung von $\mathbf{w}$ beliebig groß gemacht werden kann, so daß $\mathbf{w}$ auf die Einheitslänge beschränkt werden sollte.

Ein weiteres Problem dieses Ansatzes wird in Abbildung 3.3 deutlich. Die Projektion der Daten auf die x_1-Achse liefert die größte Differenz zwischen den beiden projezierten Klassenmitteln, allerdings sind beide Klassen dann nicht mehr trennbar. Die optimale Projektion wäre in diesem Falle eine Projektion auf die x_2-Achse. Es ist daher notwendig, auch die Streuungen innerhalb der beiden Klassen zu berücksichtigen. Fisher schlug vor, eine Funktion zu maximieren, bei der die Differenz der projezierten Klassenmittel durch ein Maß normalisiert wird, welches die Streuung innerhalb der projezierten Datenpunkte einer Klasse angibt. Dazu wird die Klassenkovarianz der projezierten Daten innerhalb einer Klasse verwendet (welche auch als within-class covariance bezeichnet wird):

$$s_k^2 = \sum_{n \in C_k} (y^n - \tilde{m}_k)^2. \tag{3.6}$$

Für das gesamte Datenset kann dann die totale within-class covariance als $s_1^2 + s_2^2$ bestimmt werden. Es ergibt sich daraus als zu maximierende Funktion

$$J(\mathbf{w}) = \frac{(\tilde{m}_2 - \tilde{m}_1)^2}{s_1^2 + s_2^2}, \tag{3.7}$$

die auch als Fishers Kriterium bekannt ist. Die Linearfunktion, die Gleichung 3.7 maximiert, wird als Fishers lineare Diskriminante bezeichnet. Gleichung 3.7

kann mit Hilfe der Gleichungen 3.2, 3.5 und 3.6 umgeformt werden, um die Abhängigkeit von $\mathbf{w}$ explizit zu machen:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}. \tag{3.8}$$

Dabei bezeichnet $\mathbf{S}_B$ die sogenannte 'Zwischen-Klassen-Kovarianzmatrix' (between-class covariance matrix) und ist gegeben durch:

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^t, \tag{3.9}$$

wobei festgestellt werden kann, daß $\mathbf{S}_B\mathbf{w}$ immer in der Richtung von $(\mathbf{m}_2 - \mathbf{m}_1)$ liegt. $\mathbf{S}_W$ ist die totale Kovarianzmatrix innerhalb der Klassen (total within-class covariance matrix) und wird berechnet durch

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{3.10}$$

mit

$$\mathbf{S}_i = \sum_{n \in C_i} (\mathbf{x}^n - \mathbf{m}_i)(\mathbf{x}^n - \mathbf{m}_i)^t. \tag{3.11}$$

Gleichung 3.8 wird unter Ausnutzung der Tatsache, daß $\mathbf{S}_B\mathbf{w}$ immer in der Richtung von $(\mathbf{m}_2 - \mathbf{m}_1)$ liegt und Vernachlässigung skalarer Faktoren durch

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{3.12}$$

maximiert, wodurch sich die lineare Diskriminante von Fisher ergibt. Die Erweiterung des Verfahrens auf mehr als zwei Klassen kann [Bis95] oder [DH73] entnommen werden.
Die lineare Diskriminanzanalyse wird auf dem Gebiet der audio-visuellen Sprachverarbeitung neben der Hauptkomponentenanalyse verwendet, um eine Reduktion der Dimensionalität der Datenvektoren zu erreichen. Dabei wird das Verfahren in ersten Linie dazu verwendet, die Dimensionalität des visuellen Datenmaterials zu verringern, das im allgemeinen in Form von Grauwertbildern der Lippenregion vorliegt, aber auch zur Bearbeitung des akustischen Datenmaterials wird die LDA verwendet (vgl. [Sio95]).
Die mit Hilfe der LDA erzielten Ergebnisse sind in der Regel besser als diejenigen, die unter Zuhilfenahme einer Hauptkomponentenanalyse erzielt werden können (vgl. [PG98],[MSYW99], [DHB+95]).

3.2 Grundlagen der Neuronalen Netze

Das Forschungsgebiet der künstlichen neuronalen Netze kann auf eine wechselvolle Geschichte zurückblicken. Einen Überblick über die ersten Anfänge gibt ein Artikel von J. D. Cowan [Cow90] beginnend mit den Arbeiten von McCulloch und Pitts 1943, die neurologische Netzwerke aus McCulloch-Pitts-Neuronen einführten [MP43] und zeigen konnten, daß neuronale Netze prinzipiell jede arithmetische oder logische Funktion berechnen können, über Hebb, der 1949 in [Heb49] die Hebb'sche Lernregel für Neuronen beschrieb und Taylor [Tay56] und Steinbuch [Ste61], die die Grundzüge der assoziativen Speicher

entwickelten, bis hin zu Rosenblatts Werk über das Perzeptron [Ros58] [Ros62]. Als Minsky und Papert in ihrem Buch *Perceptrons* 1969 [MP69] das Perzeptron einer genauen mathematischen Analyse unterzogen und zeigen konnten, daß dieses einige wichtige Problem, z.B. das XOR-Problem, nicht lösen kann, führte u.a. dies zunächst für fast 15 Jahre zu einer gewissen Stagnation des Gebiets. Erst Arbeiten von John Hopfield Anfang der achtziger Jahre [HT85] und vor allem das Buch *Parallel Distributed Processing* [RM86b], durch das das Backpropagation-Verfahren eine weite Verbreitung fand, führte zu einer Wiederauflebung des Gebietes. Künstliche neuronale Netze haben seitdem eine weite Verbreitung in verschiedensten Bereichen gefunden, so z.B. in der Sprachverarbeitung [EM88] [SYMHGJ90], der Mustererkennung [Bis95] [Rip96] und bei der Zeitreihenanalyse [WG94] [2].

Inspiriert durch das biologische Vorbild des Gehirns versucht man bei der Entwicklung Neuronaler Netzwerkmodelle, dessen Vorteile gegenüber herkömmlichen VonNeumann-Rechnern, wie Parallelität der Berechnungen, Fehlertoleranz bzw. Generalisierungsfähigkeit, Lernfähigkeit und Flexibilität in Algorithmen umzusetzen. Im Unterschied zu herkömmlichen Algorithmen werden daher neuronale Netze nicht 'fest' programmiert, sondern sie 'lernen' anhand von Beispielen, ein Problem zu lösen.

Anstatt nur einen Prozessor für eine serielle Verarbeitung der Daten zu benutzenm verwenden neuronale Netze viele einfache Verarbeitungseinheiten, die auf verschiedene Art und Weise miteinander vernetzt werden und die Daten parallel verarbeiten können.

Vorbild für eine solche Verarbeitungseinheit ist das biologische Neuron. Dieses besteht aus einem Zellkörper, dem Axon und den Dendriten (vgl. Abbildung 3.4). Die Kopplungsstellen zwischen dem Axon des einen Neurons und den Dendriten anderer Neurone sind die Synapsen, wobei es neben erregenden auch hemmende Synapsen gibt, die das Erregunspotential des nachfolgenden Neurons verringern anstatt es zu vergrößern. Über die Synapsen der Dendriten erhält das Neuron Erregungsmuster von anderen Neuronen, die (vereinfacht gesagt) summiert werden, bis ein Schwellwert im Zellkörper erreicht ist. Ist dies der Fall, beginnt das Neuron zu 'feuern', d.h. selbst eine Erregung über sein Axon und die Synapsen an die Dendriten anderer Neuronen weiterzugeben.

Eine Verrechnungseinheit eines neuronalen Netzes (auch Knoten oder *unit* genannt) bildet die Verhaltensweise eines biologischen Neurons nach. Das einfachste Modell ist in Abb. 3.5 abgebildet. Es besteht aus Dendriten, die über gewichtete Synapsen w_{ij} (diese werden auch als Verbindungsgewichte bezeichnet) die Eingabe e_i der anderen Neuronen in das Neuron weiterleiten. Im Zellkörper findet anschließend eine Summation dieser Eingaben statt, gefolgt von einer Übertragungsfunktion $g(x)$ mit einem Schwellwert T, durch die die Ausgabe des Neurons bestimmt wird:

$$o_j = g(\sum_i w_{ij} e_i - T). \tag{3.13}$$

[2] Weitere Anwendungsgebiete können z. B. dem Übersichtsartikel von Kemsley und Martinez [KM92] entnommen werden.

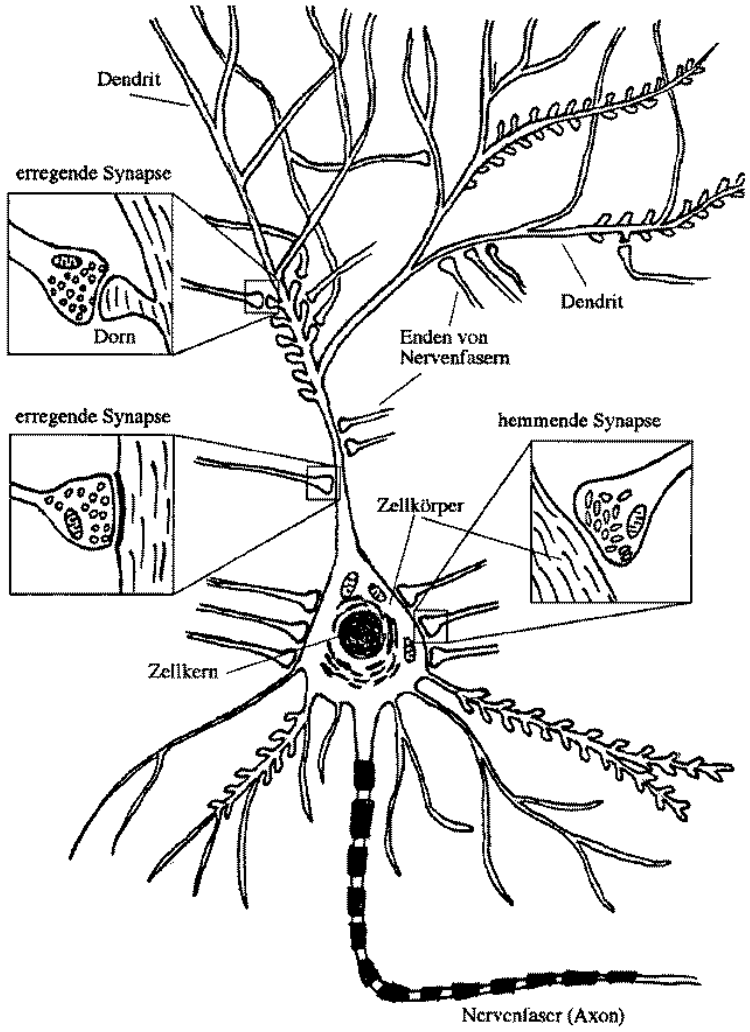

Abbildung 3.4: Schematische Abbildung eines biologischen Neurons, nach [Zel94]

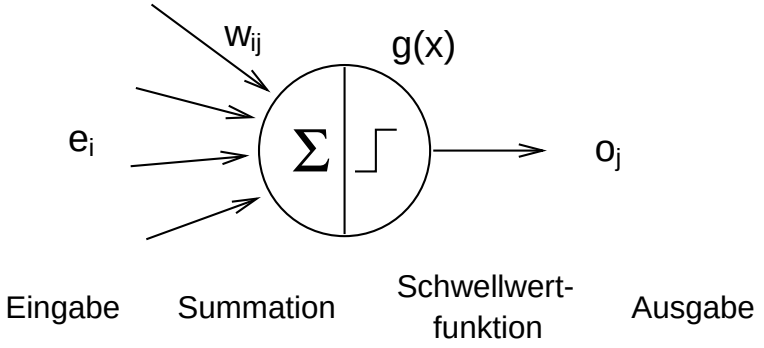

Abbildung 3.5: Schematische Abbildung eines künstlichen Neurons

Populäre Funktionen für Übertragungsfunktionen sind z.B. binäre Schwellenfunktionen (McCulloch-Pitts-Neuronen)

$$g(x) = \begin{cases} 1 & , \quad wenn \quad x \geq 0 \\ 0 & , \quad sonst \end{cases} \tag{3.14}$$

oder sigmoide Funktionen, wie die logistische Aktivierungsfunktion

$$g(x) = \frac{1}{1 + e^{-x}} \, . \tag{3.15}$$

Ein Neuron dieser Art beschreibt im Musterraum eine Hyperebene und ist daher in der Lage, bei linear separierbaren Mustern als Klassifikator zu dienen. Die verschiedenen neuronalen Netzwerkmodelle unterscheiden sich hinsichtlich der Verknüpfungsstruktur ihrer Neuronen (mit und ohne Rückkopplung), der Art der Neuronen und des Algorithmus, mit dem die veränderlichen Parameter des Netzes (das sind i.a. die Verbindungsgewichte und der Schwellwert der Übertragungsfunktion) angepaßt werden, um eine bestimmte Problemstellung zu lösen. Die Anpassung der veränderlichen Parameter an die Problemstellung wird auch als 'Lernen' bezeichnet und erfolgt mit Hilfe von Beispielen, die die Problemstellung charakterisieren. Diese Beispiele werden derart kodiert, daß sie als Eingabe eines neuronalen Netzes dienen können, wobei unter Umständen auch die für jede Eingabe gewünschte Ausgabe mit angegeben werden kann.
Allgemein unterscheidet man dabei drei Arten des Lernens:
1. überwachtes Lernen (*supervised learning*)
Dabei werden dem Netz neben den Eingaben auch die erwünschten Ausgaben (Lehrersignal) präsentiert und die veränderlichen Parameter so angepaßt, daß das Netz auf möglichst jede Eingabe die korrekte Ausgabe erzeugt.
2. bestärkendes Lernen (*reinforcemnet learning*)
Hier wird dem Netzwerk zu jedem Eingabemuster nur mitgeteilt, ob es dieses richtig oder falsch klassifiziert hat. Die korrekte Ausgabe muß das Netz selber finden.
3. unüberwachtes Lernen (*unsupervised learning*)
Dabei gibt es überhaupt keinen externen Lehrer. Das Lernen geschieht hierbei durch Selbstorganisation des Netzes, wobei das Netz versucht, ähnliche Eingaben in ähnliche Kategorien zu ordnen.
Um festzustellen wie gut das gelernte Netz an die gestellte Problematik angepaßt ist, ist es üblich, diesem nach dem Training der Beispiele (auch Trainingsmuster genannt) weitere unbekannte Beispiele zu präsentieren, die sogenannten Testmuster. Dadurch kann festgestellt werden, in wieweit das Netz in der Lage ist zu generalisieren, d.h. unbekannte Muster korrekt zu klassifizieren.

3.3 Modelle der Neuronalen Netze

Im folgenden Abschnitt soll auf verschiedene neuronale Netzwerke näher eingegangen werden, wobei insbesondere diejenigen Modelle beschrieben werden sollen, die in dieser Arbeit näher betrachtet werden beziehungsweise Erwähnung finden.

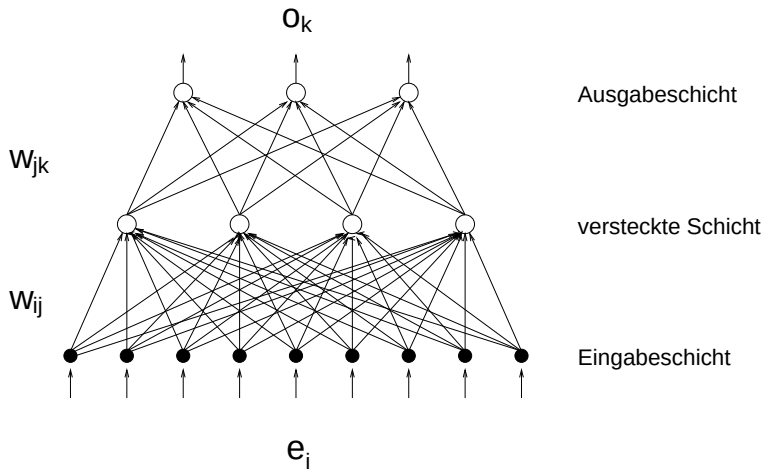

Abbildung 3.6: Mehrschichtiges neuronales Netzwerk (MLP)

3.3.1 Multilayer-Perzeptrone und Backpropagation

Feedforward-Netzwerke sind Netzwerke von Neuronen mit mehr als einer Schicht adaptiver Gewichte. Die Neuronen haben keine Rückkopplungen zu Neuronen früherer Schichten, und es sind meist nur die Neuronen zweier benachbarter Schichten miteinander verbunden.j Sind die verwendeten Neuronen im Netzwerk von der Art, wie sie Gleichung 3.13 beschreibt, wird das Netzwerk auch als Multilayerperzeptron bezeichnet (dies soll zukünftig auch als MLP abgekürzt werden). Für solche Netzwerke mit differenzierbaren Aktivierungsfunktionen existieren Lernverfahren, die die Ableitungen dieser Aktivierungsfunktionen benutzen. Das wichtigste Verfahren dabei wird error backpropagation (kurz Backpropagation) genannt und soll hier kurz skizziert werden.

Abbildung 3.6 zeigt ein zweischichtiges Feedforward-Netzwerk. Es besteht aus einer Eingabe-, einer sogenannten versteckten (hidden bzw. intermediate [HKP91]) und einer Ausgabeschicht. In der Abbildung sind die Bezeichnungen und Indizes angegeben, die im folgenden bei der Herleitung der Lernregeln benutzt werden sollen. Das Backpropagationverfahren ist ein überwachtes Lernverfahren, daß Lehrersignale zum Training des Netzes verwendet. Dabei soll der Fehler, den das Netz bei der Klassifikation der Trainingsmuster macht, für alle Trainingsmuster minimiert werden, indem die Gewichte des Netzes modifiziert werden. Die zu minimierende Fehlerfunktion wird daher als Energiefunktion in Abhängigkeit der Gewichte formuliert. Die Optimierung erfolgt mit Hilfe eines Gradientenabstiegs. Als Energiefunktion wird oft der quadratische euklidische Abstand zwischen erwarteter Ausgabe (Lehrersignal, t) und tatsächlicher Ausgabe (O) des Netzes gewählt:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{pk} \left(t_k^p - O_k^p\right)^2 \tag{3.16}$$

Dabei bezeichnet p den Index der Trainingsmuster, k ist der Index der Ausgabeschicht und $\mathbf{w}$ ist der Gewichtsvektor des Netzes. Die Änderung der Gewichte

des Netzes wird nun proportional zum negativen Gradienten dieser Fehlerfunktion bestimmt, Der Proportionalitätsfaktor ist die sogenannte Lernrate oder Schrittweite η, :

$$\Delta w_{ij} = -\eta \frac{\partial}{\partial w_{ij}} E \tag{3.17}$$

Um die Gewichtsänderungen für jede Schicht des Netzes zu bestimmen, soll zunächst die Energiefunktion als Funktion der Gewichte explizit ausgedrückt werden. Die Ausgabe der verstecken Schicht, H_j ergibt sich zu:

$$H_j = g(\underbrace{\sum_i w_{ij} e_i}_{h_j}) \tag{3.18}$$

und die Ausgabe des Netzes, O_k, zu:

$$O_k = \underbrace{g\left(\sum_j w_{jk} h_j\right)}_{o_k}. \tag{3.19}$$

Damit erhält man als Energiefunktion $E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{pk} \left(t_k^p - g\left(\sum_j w_{jk} g\left(\sum_i w_{ij} e_i \right) \right) \right)^2 \tag{3.20}$$

Als update-Regel für die Gewichte zwischen der versteckten und der Ausgabeschicht ergibt sich damit:

$$\begin{aligned} \Delta w_{jk} &= -\eta \frac{\partial E}{\partial w_{jk}} && (3.21)\\ &= \eta \sum_p \left(t_k^p - o_k^p\right) g'\left(h_k^p\right) H_j^p && (3.22)\\ &= \eta \sum_p \delta_k^p H_j^p && (3.23) \end{aligned}$$

mit

$$\delta_k^p = g'\left(h_k^p\right)\left(t_k^p - O_k^p\right)$$

und für die Berechnung der update-Regel für die Gewichte zwischen der Eingabe- und der versteckten Schicht:

$$\begin{aligned} \Delta w_{ij} &= -\eta \frac{\partial E}{\partial w_{ij}} && (3.24)\\ &= -\eta \sum_p \frac{\partial E}{\partial H_j^p} \frac{\partial H_j^p}{\partial w_{jk}} && (3.25)\\ &= \eta \sum_{pk} \left(t_k^p - O_k^p\right) g'\left(h_k^p\right) w_{jk} g'\left(h_j^p\right) e_i^p && (3.26)\\ &= \eta \sum_{pk} \delta_k^p w_{jk} g'\left(h_j^p\right) e_i^p && (3.27)\\ &= \eta \sum_p \delta_j^p e_i^p && (3.28) \end{aligned}$$

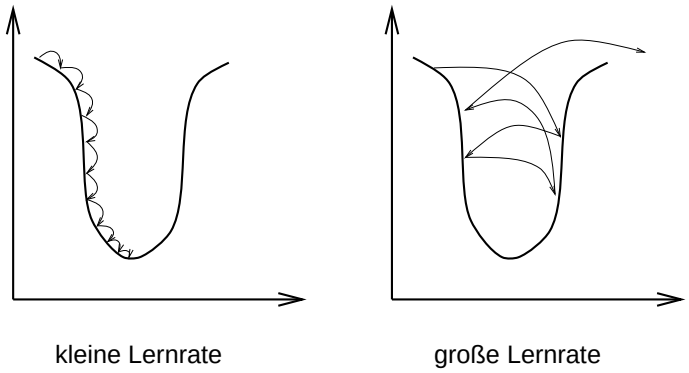

Abbildung 3.7: Konvergenzverhalten des Backpropagation-Algorithmus bei Lernraten η unterschiedlicher Größe

mit

$$\delta_j^p = g'\left(h_j^p\right) \sum_k w_{jk} \delta_k^p \,.$$

Die Neuberechnung der Gewichte w_{ij} erfolgt, indem alle Trainingsmuster dem Netz präsentiert und die entsprechenden w_{ij} berechnet und gespeichert werden. Nach Präsentation aller Muster werden die Gewichte angepaßt, was man auch als Off-Line-Lernen oder Batch-Lernen bezeichnet und der Berechnung des exakten Gradienten entspricht. Alternativ dazu wird oftmals ein On-Line-Lernen verwendet, bei dem die Gewichtsanpassung nach jeder Präsentation eines Musters erfolgt, so daß hierbei auch die Reihenfolge der Musterpräsentationen eine Rolle spielt. Diese werden daher in zufälliger Reihenfolge präsentiert.

Die richtige Wahl der Lernrate η ist ebenfalls für das Erreichen eines Minimums und eine schnelle Konvergenz wichtig. Ist sie zu klein, sind die Gewichtsänderungen entsprechend klein und das Verfahren benötigt eine größere Anzahl an Schritten, um ein Minimum zu finden. Ist sie zu groß, kann unter Umständen das Minimum nicht erreicht werden, da der Algorithmus immer wieder aus dem Minimum herausspringt (vgl. Abbildung 3.7). Aus diesem Grund wurden verschiedene Modifikationen des Backpropagation-Algorithmus entwickelt, die die Lernrate automatisch anpassen und i. a. zu einer schnelleren Konvergenz führen. Dazu gehören z. B. die Verwendung eines Momentum-Terms oder das Verfahren Super-SAB (vgl. [Zel94]). Allerdings verwenden diese Verfahren weitere Parameter zur Steuerung der Anpassung der Lernrate, deren optimale Werte ebenfalls wieder vom Problem abhängig sind.

Ein weiterer Faktor, der Einfluß auf das Trainingsergebnis eines Netzes hat, ist die Initialisierung der Gewichte, die i.a. zufällig erfolgt. Ist diese extrem ungünstig, liegt der Startpunkt der Optimierung sehr weit vom (globalen) Minimum entfernt, und das Netz findet unter Umständen keine geeignete Lösung des Problems.

Schließlich ist auch die für die Problemstellung günstigste Wahl der Netztopologie schwierig zu bestimmen. Da durch die Festlegung der Eingabe- und Ausgabekodierung die Anzahl der Eingabe- und Ausgabeneuronen feststeht, verbleibt

noch die Wahl der Anzahl an versteckten Schichten und der jeweiligen Neuronenanzahl pro Schicht. Hornik et al. [HSW89] konnten zeigen, daß Netze schon mit einer versteckten Schicht in der Lage sind, jede kontinuierliche Funktion zu approximieren, solange genug versteckte Neuronen zur Verfügung stehen. Es bleibt, die notwendige Anzahl an versteckten Neuronen zu bestimmen. Sind es zu wenige, ist das Netz nicht in der Lage, die Abbildung ausreichend genau zu lernen. Bei Verwendung zu vieler Neuronen wird die Abbildung zu exakt gelernt, was dazu führt, daß dem Netz unbekannte Muster nur noch schlecht klassifiziert werden, d.h. die Generalisierungsfähigkeit des Netzes sinkt. Dieses Problem wird als 'Overfitting' bezeichnet und kann auch auftreten, wenn das Netz zu lange trainiert wird, so daß das globale Minimum gefunden wird und dadurch die Trainingsdaten zu exakt abgebildet werden.
Neben der Wahl einer günstigen Kodierung der Eingabemuster stellt sich ebenso die Frage nach einer geeigneten Kodierung der Ausgabe. Besteht die Problemstellung darin, zu den Eingabemustern die jeweilige Klasse zu bestimmen, wird häufig die sogenannte 1-aus-m-Kodierung gewählt, bei der jede Klasse durch ein Ausgabeneuron repräsentiert wird. Der zu einer Klasse C_i gehörende Ausgabevektor ist demnach ein Binärvektor mit einer '1' an der i-ten Stelle, die übrigen Stellen sind Null. Richard und Lippmann [RL91] konnten zeigen, daß Netzwerke, die eine 1-aus-m-Kodierung und als Energiefunktion die sogenannte 'sum-of-squared-error'-Funktion verwenden, in der Lage sind, die a posteriori-Wahrscheinlichkeit der Klassenzugehörigkeiten mit ihren Ausgabewerten abzuschätzen.

3.3.2 Time-Delay Netzwerke

Bei einem TDNN (Time-delay Neural Network) handelt es sich um ein vorwärtsgerichtetes neuronales Netz, welches der Erkennung von Merkmalen in zeitlich veränderlichen Mustern, z.B. Zeitsignalen dient (vgl. [Wai89],[LWH90]). Dabei können die Muster unterschiedliche Längen aufweisen und die Merkmale sich an unterschiedlichen Positionen innerhalb des Musters befinden. Als Eingabe in das Netz dient ein Fenster des Zeitsignals, im Falle von Sprachsignalen werden demnach mehrere Zeitscheiben gleichzeitig dem Netz als Eingabe präsentiert. Nur ein jeweils begrenzter Ausschnitt aus diesem Eingabefenster wird als Input für ein festgelegtes Neuron der versteckten Schicht interpretiert, das damit ein Art rezeptives Feld besitzt. Die Eingabe- und die verstecke Schicht sind nicht vollständig miteinander vernetzt. Die Neuronen der versteckten Schicht kann man sich in mehreren Zeilen angeordnet vorstellen, wobei die Neuronen einer Zeile als Eingabe jeweils ein anderes verschobenes rezeptives Feld besitzen, so daß die Neuronen einer Zeile insgesamt den ganzen Zeitverlauf der Eingabe abdecken. Diese Neuronen teilen sich dieselben Gewichtswerte. Ein TDNN besitzt zwei versteckte Schichten. Das Training erfolgt mit einem modifizierten Backpropagation-Algorithmus. Ein MS-TDNN (Multistate-TDNN) ist eine Erweiterung des TDNNs, welches nicht nur Buchstaben oder Phoneme sondern eine Folge davon, das heißt zum Beispiel Worte, erkennen kann. Dies wird durch Hinzufügen einer weiteren versteckten Schicht erreicht, die hinter die zweite versteckte Schicht des TDNNs (die auch als 'Phonemschicht', vgl.

[Zel94] bezeichnet wird) eingeschoben wird und die zeitliche Variationen der verschiedenen Phoneme berücksichtigen kann.

3.3.3 Boltzmann-Maschinen

Eine Boltzmannmaschine besteht aus einem rekurrenten neuronalen Netz, in dem alle Neuronen miteinander aber nicht mit sich selbst verbunden sind. Die Gewichte sind symmetrisch, die Aktivierungen der Neuronen sind 1 und 0. Eine Gruppe der Neuronen ist sichtbar, die übrigen Neuronen sind versteckt. Über die sichtbaren Neuronen können Eingaben an das Netz angelegt werden. Die Aktivierungsfunktion der Neuronen ist stochastischer Natur und enthält einen sogenannten 'Temperaturterm'. Das Lernverfahren für die Boltzmannmaschine kann als simulated annealing bezeichnet werden, da die 'Temperatur' verringert wird, um ein Energieminimum im Zustandsraum zu finden.

3.3.4 Assoziativspeicher

Die Grundzüge des Assoziativspeichers wurden in den 60er Jahren von K. Steinbuch [Ste61]und [Ste71] beschrieben, auch wenn die Realisierung der darin beschriebenen Lernmatrix nicht explizit neuronal war. Diese Grundzüge wurden von Willshaw et al. ([WBLH69]) und Palm ([Pal80], [Pal82]) weiterentwickelt und auf eine neuronale Basis gestellt.
Bei einem Assoziativspeicher handelt es sich um ein neuronales Netz mit folgenden Eigenschaften:

- Ein Assoziativspeicher ist fehlertolerant, d.h. auch fehlerhafte Eingaben von gelernten Mustern erzeugen korrekte Ausgaben.
- Als ein Spezialfall der Fehlertoleranz ist die Musterergänzung zu sehen: Ein Assoziativspeicher ist in der Lage, unvollständige Eingabemuster zu vervollständigen.
- Ein Assoziativspeicher kann 'interpolieren', d.h. falls ein nicht gespeichertes Muster eingegeben wird, wird das ähnlichste gespeicherte Muster bestimmt und die entsprechende Ausgabe generiert.

Die Mustervektoren eines Assoziativspeichers bestehen aus binären Vektoren $x^\mu \in X = \{0,1\}^n$. Dabei läßt sich eine '1' als das Vorhandensein eines bestimmten fest gelegten Merkmals interpretieren. Auch der Ausgabevektor $y^\mu \in Y = \{0,1\}^m$ ist binär kodiert. Die Dimension des Eingabevektors sei dabei n, die des Ausgabevektors m. Die Aufgabe des Assoziativspeichers besteht darin, M binäre Vektorpaare $\{(x^\mu, y^\mu)\}_{\mu=1}^M$ so zu speichern, daß der Assoziativspeicher bei Eingabe eines x^μ als Ausgabevektor das gewünschte Ausgabemuster y^μ erzeugt. Diese Anfrage an den Assoziativspeicher wird auch als Retrieval bezeichnet.
Abbildung 3.8 zeigt einen Assoziativspeicher, der aus m Neuronen aufgebaut ist, welche ein einschichtiges Netz bilden. Die Gewichte der Neuronen sind dabei als Punkte gezeichnet. Der Eingabevektor $\mathbf{x}$ der Dimension n wird an die Neuronen angelegt, wobei die einzelnen Komponenten x_i mit den Gewichten w_{ij} multipliziert werden. Anschließend erfolgt eine Summation in jedem Neuron j, mit nachfolgender Schwellwertentscheidung, die im Bild mit einem T angedeutet wird. Dabei gilt für die Ausgabe $\mathbf{y}$

$$y_j = \begin{cases} 1 & : \quad \sum_{i=1}^n w_{ij} x_i \geq T \\ 0 & : \quad \sum_{i=1}^n w_{ij} x_i < T \end{cases} . \tag{3.29}$$

Jedes Neuron erzeugt somit eine Ausgabe y_j, alle Ausgaben können als Ausgabevektor $\mathbf{y}$ zusammengefaßt werden.
Bei der Art der Speicherung lassen sich zwei unterschiedliche Aufgabenstellungen unterscheiden:
Bei der Autoassoziation bestehen Eingabe- und Ausgabevektoren aus denselben Binärvektoren, d.h. vom Assoziativspeicher soll die Abbildung

$$x^\mu \to x^\mu \quad \text{mit} \quad \mu = 1, .., M \tag{3.30}$$

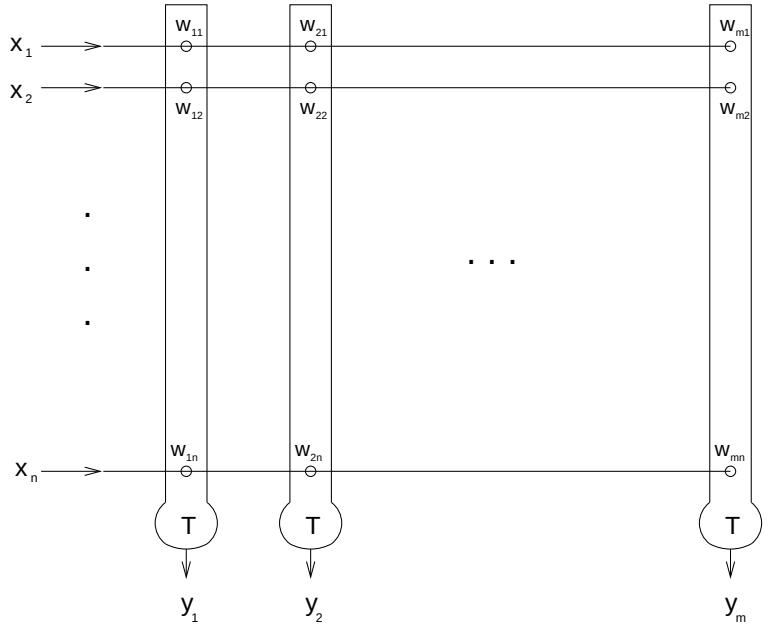

Abbildung 3.8: Aufbau eines Assoziativspeichers

gelernt werden.
Die Hauptaufgabe der Autoassoziation besteht in der Vervollständigung bzw. Fehlerkorrektur eines eingegebenen Musters $\tilde{x^\mu}$:

$$\tilde{x^\mu} \rightarrow x^\mu \quad \text{mit} \quad \mu = 1, .., M \tag{3.31}$$

Bei der Heteroassoziation unterscheiden sich Eingabe- und Ausgabevektor, so daß dieser Fall z. B. der Klassifikation des eingegebenen Musters dienen kann:

$$x^\mu \rightarrow y^\mu \quad \text{mit} \quad \mu = 1, .., M \quad \text{und} \quad x^\mu \neq y^\mu \tag{3.32}$$

In beiden Fällen soll das Retrieval fehlertolerant in dem Sinne sein, daß alle $\tilde{\mathbf{x}}$ auf das $\mathbf{y}$ abgebildet werden für das gilt, daß die $\tilde{\mathbf{x}}$ bezüglich einer definierten Metrik dem zu $\mathbf{y}$ ursprünglich gehörenden Eingabemuster $\mathbf{x}$ ähnlich sind.

Lernen

In der Lernphase werden alle Gewichte des Netzes zunächst auf Null gesetzt und anschließend die zueinandergehörigen Eingabe- und Ausgabemuster gleichzeitig an den Assoziativspeicher angelegt. Jedes Musterpaar wird nur einmal präsentiert. Ein- und Ausgabemuster erzeugen im Assoziativspeicher gewissermaßen prä- und postsynaptische Signale. Die Gewichte des Assoziativspeichers werden unter Zuhilfenahme dieser Signale eingestellt, indem lokale Lernregeln verwendet werden.
Die bekannteste ist die sogenannte Hebbsche Lernregel, bei der ein Gewicht verstärkt wird, wenn sowohl ein prä- als auch ein postsynaptisches Signal am

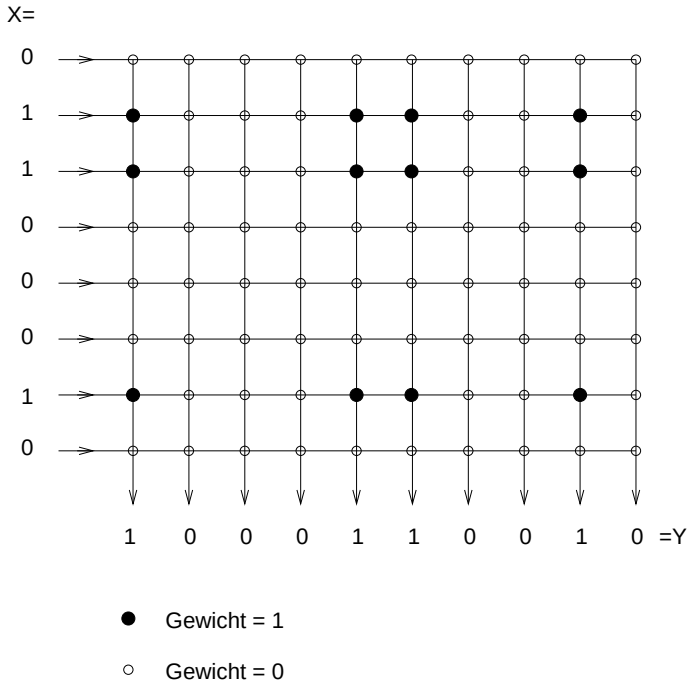

Abbildung 3.9: Lernen im Assoziativspeicher

Gewicht anliegt (vgl. graphische Illustration in Abbildung 3.9):

$$w_{ij}^{neu} = w_{ij}^{alt} + x_i y_j. \tag{3.33}$$

Diese Lernregel wird auch als additive Lernregel bezeichnet, da bei jeder Speicherung eines neuen Musterpaares das Gewicht erhöht wird.
Eine weitere lokale Lernregel ist die sogenannte Hopfield-Lernregel, bei der bei übereinstimmenden prä- und postsynaptischen Signalen die Gewichtswerte erhöht und bei nichtübereinstimmenden Signalen erniedrigt werden.
Als geclippte oder binäre Hebb-Lernregel bezeichnet man eine Lernregel, nach der der Wert der Gewichte durch eine nichtlineare Operation begrenzt wird:

$$w_{ij}^{neu} = \begin{cases} 1 & : \text{ falls } \quad x_i y_j = 1 \\ w_{ij}^{alt} & : \quad sonst \end{cases} . \tag{3.34}$$

Retrieval

Beim Retrieval berechnen alle Schwellenneurone gleichzeitig ihre Antwort auf die Eingabe $\mathbf{x}$, indem die gewichtete Summe aus den Komponenten des Eingabevektors berechnet wird, welche anschließend einer Schwellwertentscheidung unterzogen wird (vgl. Gleichung 3.29). Der Schwellwert wird dabei global vorgegeben:

$$T = T_j \quad \forall j. \tag{3.35}$$

Der Schwellwert kann als die Anzahl der Einsen im Anfragemuster festgelegt werden. Dieser Wert wird auch als Aktivität des Musters bezeichnet:

$$T = \sum_{i=1}^{n} x_i. \tag{3.36}$$

Bei dieser sogenannten harten Schwellwertentscheidung werden unter Umständen nur die gelernten Muster wiedererkannt, da unbekannte Muster von Null verschiedene Komponenten besitzen können, die keine Korrelationen zu den gelernten Ausgabemustern haben. In diesem Fall kann kein Ausgabevektor mit dieser Strategie bestimmt werden, und die Schwelle muß solange gesenkt werden, bis eine Ausgabe bestimmt werden kann. Andererseits können bei gelernten Mustern zusätzliche Einsen im Ausgabevektor auftreten, die durch die gelernten Korrelationen der anderen Eingabe-Ausgabevektor-Paare zustande kommen.
Bei einer weichen Schwellwertstrategie wird dagegen bei Verwendung der geclippten Hebb-Lernregel die Schwelle auf das Maximum der gewichteten Summen, die für alle Neuronen berechnet wurden, gesetzt:

$$T = max_j(\sum_i w_{ij} x_i). \tag{3.37}$$

In diesem Fall erhält man nur dann keinen Antwortvektor, wenn alle Komponenten des Eingabevektors zu keinem der gelernten Ausgabevektoren korreliert sind.

Kodierung

Um möglichst viele Muster fehlerfrei abspeichern zu können,, d.h. um eine hohe Speicherkapazität zu erhalten, ist es sinnvoll, für die Vektoren eine Kodierung zu wählen, durch die die Eingabe- und Ausgabevektoren nur spärlich mit Einsen besetzt sind [PSS94]. Es wird also eine bijektive Abbildung benötigt, die die ursprünglichen Merkmalsvektoren, die nicht binär kodiert sein müssen, derart auf binäre Vektoren abbildet, daß diese möglichst wenig Einsen enthalten:

$$f(\mathbf{x}) : \mathbf{x} \rightarrow \hat{\mathbf{x}} \tag{3.38}$$

Eine solche Abbildungsvorschrift bezeichnet man als spärliche Kodierung. Um eine spärliche Kodierung zu finden, kann man z. B. ein Muster durch einen kleinen Satz von Merkmalen repräsentieren, welche aus einem sehr großen Satz an möglichen Merkmalen bestimmt werden. Jedes Merkmal wird dabei durch eine Eins-Komponente des kodierten Vektors repräsentiert. Zusätzlich zur hohen Speicherkapazität wird eine gute Fehlertoleranz vom Assoziativspeicher erwartet, so daß die Abbildung außerdem die ursprünglichen Merkmalsvektoren ähnlichkeitserhaltend in die spärlichen Binärvektoren überführen soll. Als ähnlichkeitserhaltend bezeichnet man eine Kodierung, wenn folgendes gilt: Seien $\mathbf{x_1}$, $\mathbf{x_2}$ und $\mathbf{x_3}$ drei ursprüngliche beliebige Mustervektoren, wobei $\mathbf{x_1}$ dem Muster $\mathbf{x_2}$ bezüglich eines Ähnlichkeitsmaßes ähnlicher ist als $\mathbf{x_3}$. Dann muß für die kodierten Mustervektorengelten, daß $\hat{\mathbf{x_1}}$ dem Muster $\hat{\mathbf{x_2}}$ ähnlicher ist als $\hat{\mathbf{x_3}}$. Als Ähnlichkeitsmaß kann z.B. der Hammingabstand oder der Überlapp

bei binären Vektoren dienen (vgl. Seite 39).
In dieser Arbeit wurde eine n-Gramm-Kodierung verwendet, die ähnlichkeitserhaltend und spärlich ist (vgl. [PSSS97], [Wic69], [RM86a]). Die n-Gramm-Kodierung führt Zeichenketten beliebiger Länge, deren Zeichen aus einem Alphabet der Länge L stammen, in spärlich kodierte binäre Vektoren über. Dazu benutzt man die in den Zeichenketten enthaltenen Zeichen-n-Gramme für die Kodierung. Aus den L Zeichen des Alphabets lassen sich L^n verschiedene n-Gramme bilden. In der n-Gramm-Kodierung wird jedes mögliche n-Gramm durch eine Komponente des transformierten Vektors kodiert. Ist ein bestimmtes n-Gramm in der Zeichenkette enthalten, wird dies durch eine Eins in der entsprechenden Komponente kodiert. Da eine Zeichenkette in der Regel nur einen sehr geringen Bruchteil der tatsächlich möglichen n-Gramme enthält, ist diese Kodierung spärlich. Durch Vergrößerung des n kann die Spärlichkeit der Kodierung erhöht werden, gleichzeitig erhöht sich aber auch die Dimensionalität der Binärvektoren. Gleichzeitig werden gleiche Merkmale, d.h. gleiche n-Gramme gleich kodiert, die Kodierung ist ähnlichkeitserhaltend bzgl. vieler Distanzmaße [PSS94]. Über die n-Gramme können Informationen über die Reihenfolge der Zeichen in der Zeichenkette kodiert werden. Gleichzeitig ist sie unabhängig von der Länge der Zeichenkette, da mehrmals auftretende n-Gramme nur durch eine, nämlich dieselbe, Komponente des transformierten Vektors kodiert werden. Sinnvolle Werte für n sind 2,3 oder 4, wobei Trigramme (n=3) auch als Wickelmerkmale bezeichnet werden [RM86a].

Rückwärtsprojektion zur Entscheidungsfindung

Bei einer 1-aus-n-Kodierung der Ausgabevektoren ist es möglich, daß der Assoziativspeicher mehr als einen Ausgabevektor als Antwort auf einen Eingabevektor produziert. Daher ist es unter Umständen notwendig, eine Entscheidung zwischen diesen Antwortvektoren zu erzwingen. Wie in [Wic00] beschrieben, ist dies unter Ausnutzung eine Rückwärtprojektion der Ausgabevektoren möglich. Dazu werden die Ausgabevektoren $\mathbf{y}$ als Eingabevektoren in einen Assoziativspeicher betrachtet, dessen Gewichtsmatrix aus der transponierten Gewichtsmatrix des Assoziativspeichers besteht. Man erhält einen Antwortvektor, der mit $\mathbf{x_b}$ bezeichnet werden soll. Für seine Berechnung gilt

$$\mathbf{x_{ib}} = \begin{cases} 1 & : \quad \sum_{j=1}^{m} w_{ij} y_j \geq T \\ 0 & : \quad \sum_{j=1}^{m} w_{ij} y_j < T \end{cases} \tag{3.39}$$

Dabei wird für die Bestimmung des Schwellwertes T dieselbe Schwellwertstrategie wie für die Vorwärtprojektion verwendet. Um ein Vertrauensmaß für den Ausgabevektor $\mathbf{y}$ zu erhalten, wird in einem weiteren Schritt die Ähnlichkeit des Vektor $\mathbf{x_b}$ zum ursprünglichen Eingabevektor $\mathbf{x}$ bestimmt. Als Ähnlichkeitsmaße können dazu z.B. der Hammingabstand (Anzahl unterschiedlicher Bits der beiden Vektoren)

$$d_H(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \quad \text{mit} \quad \mathbf{x}, \mathbf{y} \in \{0, 1\}^n \tag{3.40}$$

oder der Überlapp (Anzahl übereinstimmender Einsen)

$$d_U(x,y) = \sum_{i=1}^{n} x_i y_i, \quad \text{mit} \quad \mathbf{x}, \mathbf{y} \in \{0,1\}^n \tag{3.41}$$

verwendet werden. Der Ausgabevektor, dessen zugehöriger Vektor $\mathbf{x_b}$ die größte Ähnlichkeit mit dem ursprünglichen Eingabevektor $\mathbf{x}$ hat, wird dann als eindeutiger Ausgabevektor gewählt.
Bei der konkreten Realisierung des Assoziativspeichers wurde eine n-Gramm-Kodierung mit $n = 3$ verwendet. Diese wird auch als Tripelkodierung bezeichnet. Zusätzlich zu den 30 Segmentklassen, die in jeder zu speichernden Sequenz vorkommen können, wurde ein Anfangs- und Endzeichen #, welches an den Anfang und an das Ende jeder Sequenz angefügt wurde, zu den Segmentklassen hinzugenommen, so daß die Größe des Alphabets $L = 31$ beträgt. Dies geschieht, um evtl. Minderberücksichtigungen der ersten und letzten Segmente jeder Sequenz zu vermeiden. Für die Ausgabevektoren wurde eine 1-aus-n-Kodierung gewählt, bei der jede Komponente des Ausgabevektors für eine mögliche Ausgabeklasse steht. Die Klassenzugehörigkeit eines Musters wird also durch eine Eins an entsprechender Stelle des Ausgabevektors kodiert. Die Dimension des Ausgabevektors betrug in diesem Falle 25. Für das Lernen der Muster wurde eine geclippte Hebb-Lernregel verwendet (vgl. Gleichung 3.34), beim Retrieval wurde eine weiche Schwellwertstrategie mit Berechnung des Maximums verwendet (vgl. Gleichung 3.37).

3.4 Sprachverarbeitung in technischen Systemen

Dieser Abschnitt behandelt in kurzer Form grundlegende Methoden der Sprachverarbeitung in technischen Systemen. Eine ausführliche Darstellung der Methoden kann z.B. [EH93], [O'S87] entnommen werden. Zunächst wird auf die Darstellung und Verarbeitung der Sprachsignale eingegangen, abgeschlossen wird dieser Abschnitt durch die Beschreibung von zwei Methoden zur Klassifikation eines Sprachsignals.

3.4.1 Verarbeitung und Darstellung von Sprachsignalen

Bevor die Klassifizierung eines Sprachsignals in einem Spracherkennungssystem erfolgen kann, muß dieses zunächst einer Vorverarbeitung unterzogen werden, die zu einer für die jeweilige Aufgabenstellung geeigneten Darstellung des Sprachsignals führt.
Abbildung 3.10 zeigt eine schematische Darstellung der üblichen Vorverarbeitungsschritte.
Zunächst wird das Sprachsignal mit einem Mikrofon aufgenommen, welches das akustische Signal zu einem elektrischen Signal transformiert. Dieses liegt anschließend als zeitkontinuierliches und wertekontinuierliches Analogsignal vor. Als nächster Schritt erfolgt eine Abtastung und Quantisierung des Analogsignals, wodurch ein zeitdiskretes und wertediskretes Signal entsteht. Dazu wird das Analogsignal zunächst durch einen Tiefpaßfilter der oberen Grenzfrequenz

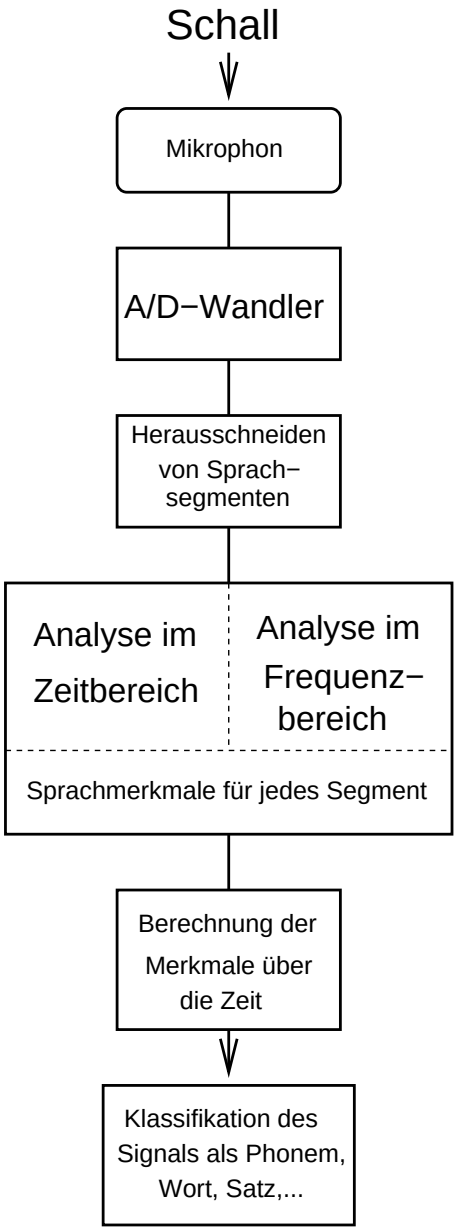

Abbildung 3.10: Schematische Darstellung der Vorverarbeitung in der Sprachverarbeitung

f_g bandbegrenzt, anschließend erfolgt eine Abtastung des Signals in gleichbleibendem zeitlichen Abstand mit der Abtastfrequenz f_0. Dabei muß gemäß dem Abtasttheorem beachtet werden, daß $f_0 >= 2f_g$ gelten muß [Lü90], in der Praxis werden Abtastfrequenzen gewählt, die mehr als doppelt so hoch wie die Grenzfrequenz f_g sind, da die ansonsten notwendigen unendlich steilen Flanken des Tiefpaßfilters technisch nicht realisiert werden können [Fel84]. Eine anschließende Quantisierung des Signals führt zu einer wertediskreten Darstellung, so daß insgesamt das Signal in digitalisierter Form vorliegt.
Je nach Aufgabenstellung erfolgt nun eine Weiterverarbeitung des Sprachsignals. Dabei kann die Zeitdarstellung des Signals zugrundegelegt werden, oder es wird eine Transformation des Signals in die Frequenzebene durchgeführt. Weiterhin kann zwischen einer Signalformdarstellung, wie sie für alle Analogsignale möglich ist, und einer parametrischen Darstellung unterschieden werden. Die parametrischen Verfahren basieren dabei auf dem Quelle-Filter-Modell (vgl. Abschnitt 2.2.3) und versuchen, die Parameter des Anregungssignals sowie die Vokaltraktparameter zu bestimmen. Sie sind daher nur auf Sprachsignale anwendbar.
In allen Fällen wird die Analyse des Signals nur innerhalb eines kurzen Intervalls durchgeführt, welches mit Hilfe einer Fensterfunktion aus dem Signal herausgeschnitten wird. Dem liegt die Annahme zugrunde, daß das Sprachsignal innerhalb des Fensters quasistationär ist. Üblich sind Rechteckfenster, Hammingfenster u.ä. (s. z.B. [O'S87])mit einer Länge von 5 - 25ms [MS91].

Analyse im Zeitbereich

Eine einfache Möglichkeit der Analyse im Zeitbereich liegt in der Ermittlung der Nulldurchgangsrate (zero crossing rate). Ein Nulldurchgang ist dabei als Vorzeichenwechsel zweier aufeinanderfolgender Abtastwerte definiert. Aus der Nulldurchgangsrate lassen sich beschränkte Aussagen über den Frequenzbereich, in dem ein Signal liegt und damit über die Stimmhaftigkeit gewinnen. In manchen Fällen kann sie daher für die Einteilung des Signals in Sprachsegmente verwendet werden.
Wichtiger für eine Analyse des Zeitsignals ist die Bestimmung der Autokorrelationsfunktion (AKF), welche ein Maß für die Ähnlichkeit zwischen zwei Signalen ist. Im vorliegenden Fall wird die Ähnlichkeit zwischen zwei Segmenten desselben Zeitsignals ermittelt. Die Segmente werden z.B. mit einer Rechteckfunktion aus dem Sprachsignal ausgeschnitten. Mit Hilfe der AKF können stimmlose Laute von stimmhaften unterschieden werden, da die Aperiodizität der stimmlosen Laute zu kleinen Werten der AKF führt, während die periodischen Signale der stimmhaften Laute zu einer periodischen AKF mit derselben Periode wie das Sprachsignal führt. Die AKF kann auch zur Bestimmung der Grundfrequenz genutzt werden.

Analyse im Frequenzbereich

Eine Signalanalyse im Frequenzbereich hat das Ziel, die spektralen Eigenschaften des Signals zu ermitteln und weiterzuverarbeiten.

Eine Möglichkeit der Frequenzanalyse liegt in der Anwendung einer Filterbank. Dabei wird das Signal durch eine parallele Anordnung von Bandfiltern geleitet, wobei jedes Filter als Ausgangssignal die Energie der entsprechenden Frequenzen liefert. Die Anwendung einer Filterbank führt damit zu einer Aufsplittung des Signals in seine Frequenzanteile, die als Merkmale für eine spätere Klassifikation oder als Grundlage für weitere Datenreduktionsverfahren dienen können. In Spracherkennungssystemen werden üblicherweise keine Filterbänke verwendet, deren Filter alle diesselbe Bandbreite haben. Da das menschliche Ohr die tiefen Frequenzbereiche besser auflösen kann als die der hohen Frequenzen [ZF67], liegt es nahe, die Filter einer Filterbank gemäß den Frequenzgruppen des menschlichen Ohres zu wählen. Üblich ist z.B. eine Anordnung von Filtern gemäß der Barkskala (vgl. S. 2.3). Es werden i.a. Filter mit sich überlappenden Filterkurven eingesetzt, da die Energie benachbarter Bänder stark korreliert ist [EH93]. Die Anzahl der Filter in einer Filterbank liegt üblicherweise zwischen 10 und 24 Filtern.
Eine weitere Möglichkeit, ein Sprachsignal in seine spektralen Anteile aufzuspalten, ist die Anwendung der Diskreten Fouriertransformation (DFT), die in den meisten Spracherkennungssystemen mit Hilfe der Fast Fourier Transformation (FFT) (vgl. z.B. [PTVF92]) realisiert wird. Die Fouriertransformation einer zeitdiskreten Folge von Signalwerten führt zu einem diskreten Spektrum (Linien-Spektrum) in der Frequenzebene, wobei die Werte symmetrisch konjugiert komplex zur Mitte des Frequenzintervalls, d.h. zur halben Abtastfrequenz, liegen. Dabei entspricht die Anzahl der Werte im Frequenzbereich der Anzahl der Abtastwerte im Zeitbereich. Da eine DFT nur auf periodische Signale angewendet werden kann, wird auch hier wieder auf die Hilfe von Signalsegmenten zurückgegriffen, die durch Anwendung einer Fensterfunktion aus dem Signal herausgeschnitten werden. Innerhalb des Segments wird das Signal als periodisch angenommen. Als Fensterfunktion ist dabei die Rechteckfunktion ungeeignet, da es durch den harten Übergang an beiden Seiten zu Unstetigkeiten an den Grenzen kommt, die zu zusätzlichen hohen Frequenzanteilen im Spektrum führen. Günstiger ist z.B. die Verwendung des Hammingfensters. Oft werden überlappende Segmente für die Berechnung der DFT verwendet.
Als weiteres Verfahren der Signalverarbeitung im Frequenzbereich soll die homomorphe oder Cepstralanalyse erwähnt werden. Es handelt sich dabei um ein parametrisches Verfahren, welches das Quelle-Filter-Modell der Sprachproduktion als Grundlage heranzieht. Danach besteht ein Sprachsignal aus einem (periodischen) Anregungsignal, welches durch den Vokaltrakt moduliert wird, so daß das Signal als Faltung der Anregungsfunktion und der Impulsantwort des Vokaltraktes erscheint. Eine DFT des Signals liefert daher als Fouriertransformierte des Signals das Produkt der Fouriertransformierten der Anregungsfunktion und der Impulsantwort. Um beide Anteile voneinander trennen zu können, wird eine Logarithmierung der Fouriertransformierten durchgefuhrt, die das obige Produkt in eine Summe überführt. Bei einer Fourier-Rücktransformation bleibt die Summe erhalten. Das resultierende Ausgangssignal wird als Cepstrum [3] des Sprachsignals bezeichnet und enthält die Summe der Einzel-

[3] Eine Umordnung der Buchstaben von 'spectrum'

Cepstra des Anregungssignals und der Impulsantwort. Für periodische Signale erhält man durch Fouriertransformation aufgrund der Grundfrequenz und ihrer Harmonischen ebenfalls ein Signal mit periodischen Anteilen, welche den Anteil der Anregungsfunktion wiedergeben. Durch die Logarithmierung wird dieser als Summand abgespalten und erscheint durch anschließende Fourier-Rücktransformation als einzelner Peak im Cepstrum, während der Einfluß des Vokaltraktes als regelloser Beitrag in der Umgebung des Nullpunktes dargestellt wird. Für stimmlose Laute treten dagegen keine Spitzen im Cepstrum auf. Eine Cepstrum-Analyse dient daher z.B. der Grundfrequenzanalyse, der Unterscheidung zwischen stimmhaften und stimmlosen Lauten, sowie der Ermittlung der reinen Übertragungsfunktion des Vokaltraktes.
Schließlich soll noch auf die Analyse durch Lineare Prädiktion eingegangen werden, welche inzwischen eines der am häufigsten verwendeten Verfahren ist. Zugrunde gelegt wird bei dieser Methode wiederum die Annahme, daß sich die Signalcharakteristik innerhalb des zu analysierenden Zeitraumes nicht ändert, so daß auch hier mit Segmenten des Zeitsignals gearbeitet wird. Bei dem Prädiktionsverfahren versucht man, den nächsten Signalwert aus den vorhergegangenen Abtastwerten zu berechnen, indem diese mit entsprechenden Koeffizienten multipliziert und dann addiert werden. Die Bestimmung dieser Koeffizienten, die die Charakteristik eines periodischen Signals eindeutig festlegen, ist die Aufgabe des Prädiktionsverfahrens. Mit Hilfe der Koeffizienten, in der Regel 12-18 Stück, läßt sich dann das Sprachsegment beschreiben. Die Koeffizienten müssen für jedes Sprachsegment neu berechnet werden, indem z.B. die Summe der Fehlerquadrate zwischen vorhergesagtem und tatsächlichem Signal für das gesamte Segment durch ein Gradientenverfahren minimiert wird. Mit den Prädiktionskoeffizienten lassen sich spektrale Eigenschaften des Sprachsignals beschreiben, allerdings werden die bei Nasalen auftretenden Antiformanten sowie rauschförmige Signale nicht adäquat erfaßt ([MS91]).
Um den dynamischen Verlauf der Sprache wiedergeben zu können, ist eine dreidimensionale Darstellung des Sprachsignals üblich. Dabei wird auf der horizontalen Achse die Zeit wiedergegeben, während auf der vertikalen Achse die Frequenz aufgetragen ist. Für jeden diskreten Zeitschritt auf der Zeitskala wird das Ergebnis einer Spektralanalyse, z.B. durch eine FFT, eines entsprechenden Sprachsegments aufgetragen, indem die Intensität in den jeweiligen Frequenzbändern durch eine Graustufendarstellung (die dritte Dimension) wiedergegeben wird. Solche Diagramme werden auch als Visible Speech Diagramme oder Spektrogramme bezeichnet und erlauben eine sehr anschauliche und gut interpretierbare Darstellung des Sprachsignals. Man unterscheidet Schmalband- (Bandbreite der Filter ca. 50 Hz) und Breitbandspektrogramme (Bandbreite der Filter ca. 300 Hz), wobei die Breitbandanalyse in den meisten Fällen bevorzugt wird, da sie die Formanten und spektrale Änderungen, die auf Bewegungen des Vokaltraktes zurückzuführen sind (z.B bei Plosiven), erkennbar werden läßt [O'S87].

3.4.2 Kodierung von Sprachsignalen-Vektorquantisierung

Eine weit verbreitete Methode zur Kodierung von Sprachsignalen besteht in der sogenannten Vektorquantisierung. Bei der Vektorquantisierung handelt es sich um eine Möglichkeit, die Datenrate zur Übertragung von Sprachsignalen zu verkleinern (vgl. [EH93]). Dazu wird eine Folge von Vektoren mit kontinuierlichen der diskreten Komponenten in eine digitale Folge übersetzt. Die Komponenten der Vektoren können zum Beispiel Abtastwerte des Sprachsignals sein, so daß ein Zeitabschnitt des Signals als Vektor interpretiert werden kann. Einem Vektor wird anschließend ein Zahlenwert zugeordnet, indem man einen Mustervergleich durchführt. Der Quantisierer besitzt ein sogenanntes Codebuch mit Referenzvektoren, die alle mit dem zu quantisierenden Vektor anhand eines Ähnlichkeitsmaßes verglichen werden. Die Referenzvektoren teilen also den Musterraum, in dem sie liegen, in Segmente auf. Alle Mustervektoren, die in einem Segment liegen, erhalten bei der Quantisierung den zugehörigen Wert (das heißt die zugehörige Codebuch-Nummer des Referenzvektors) zugewiesen. Der Referenzvektor ist meistens der Mittelwert des Segments des Eingangsraums Bei der Datenübertragung wird nur diese Nummer übertragen. Sender und Empfänger benutzen dieselben Codebücher.

3.4.3 Klassifikation von Sprachsignalen

In diesem Abschnitt sollen aus der Vielzahl der Verfahren zur Klassifikation von Sprachsignalen zwei Verfahren näher erläutert werden, da es sich bei diesen um sehr geläufige Verfahren der Mustererkennung handelt.

Hidden Markov Modelle

Bei einem Hidden Markov Modell (HMM) handelt es sich um ein Markov-Modell erster Ordnung. Dieses generiert eine Sequenz von Zuständen, die dem Beobachter verborgen sind, und in einem zweiten Zufallsprozeß Beobachtungen, die vom jeweiligen Zustand des Modells abhängig sind. Diese Beobachtungen werden gemäß einer Wahrscheinlichkeitsverteilung generiert, die dem Zustand des Modells zugeordnet ist. Die Trainingsdaten werden dazu benutzt, die optimalen Parameter des Modells zu bestimmen.
Ein Markov-Modell ist ein System, das sich zu jeder Zeit in einem von N Zuständen, N endlich, befindet. Die Zustände sollen im folgenden mit $s_1, ..., s_N$ bezeichnet werden.In regelmäßigen Zeitintervallen kann das System seinen Zustand ändern oder auch im momentanen Zustand verbleiben. Die Wahrscheinlichkeit für das Modell, den Zustand zu ändern, hängt nur vom Zustand des Modells zum vorausgegangenen Zeitpunkt ab. Sei q_t der Zustand des Modells zum Zeitpunkt t und a_{ij} die Übergangswahrscheinlichkeit von Zustand i zu Zustand j mit

$$a_{ij} = P(q_t = s_j | q_{t-1} = s_i). \tag{3.42}$$

Die Übergangswahrscheinlichkeiten können in einer Matrix

$$A = \begin{bmatrix} a_{11} & .. & a_{1N} \\ .. & .. & .. \\ a_{N1} & .. & a_{NN} \end{bmatrix} \tag{3.43}$$

zusamengefaßt werden.
π_i bezeichne nun die Wahrscheinlichkeitsverteilung für den initialen Zustand

$$\pi_i = P(q_a = s_i) \quad \text{mit} \quad 1 \leq i \leq N \tag{3.44}$$

Sei Q_T die Zustandssequenz des Modells nach T Zeitschritten:

$$Q_T = \{q_1, q_2, ..., q_T\}. \tag{3.45}$$

Bei jedem Zeitschritt wird gemäß einer Wahrscheinlichkeitsverteilung, die vom Zustand des Modells abhängig ist, eine Beobachtung x zufällig ausgewählt, d.h. jedem Zustand s_i wird eine Wahrscheinlichkeitsverteilung $b_i(x), B = \{b_i(x)\}$, zugeordnet. Diese kann diskret oder kontinuierlich sein.
Ein HMM wird üblicherweise mit folgender Notation beschrieben: $\lambda = \{A, B, \pi\}$. Während des Trainings eines Hidden-Markov-Modells mit einem Satz an Beobachtungssequenzen wird ein optimales Set von Modellparametern λ bestimmt. Dazu dient z.B. der Baum-Welch-Algorithmus. Während der Erkennungsphase soll eine Beobachtungssequenz erkannt werden, indem z.B. mit dem Forward-Backward-Algorithmus das wahrscheinlichste Modell aus einer Anzahl von konkurrierenden Modellen ausgewählt wird oder indem die wahrscheinlichste Zustandssequenz in einem großen Modell bestimmt wird (dazu wird oft der Viterbi-Algorithmus verwendet).
Bei der Einzelworterkennung wird z.B. ein HMM für jedes Wort trainiert, zur Erkennung wird das Modell bestimmt, das am wahrscheinlichsten die Beobachtungssequenz generiert hat.
Eine genauere Beschreibung von Hidden-Markov-Modellen ist z.B. in [RJ93], [RJ86] oder [Jel97] zu finden.

Dynamische Zeitanpassung

Bei dem Verfahren des Dynamic Time Warping (DTW) handelt es sich um eine nichtlineare Zeitanpassung unter Zuhilfenahme der Dynamischen Programmierung (vgl. [EH93]). Das Prinzip soll am Beispiel der Wortklassifikation erläutert werden. Wenn ein Wort mehrmals geäußert wird, ist sein Sprachsignal für jede Äußerung unterschiedlich lang und hat unterschiedliche Charakteristika. Um einen Vergleich eines zu klassifizierenden Testmusters (in diesem Fall ein geäußertes Wort) mit einem Referenzmuster durchführen zu können, werden beide Muster synchronisiert. Jedes Muster ist mit derselben Frequenz abgetastet und in Zeitscheiben unterteilt. Zur Synchronisation der Muster wird eine Abbildungsfunktion zwischen beiden Mustern bestimmt, so daß gleiche akustische Ereignisse einander zugeordent werden. Verschiedene Randbedingungen bei der Generierung der Abbildungsfunktion dienen u. a. dazu, den Grad der

Nichtlinearität möglichst gering zu halten und eine sinnvolle Abbildungsfunktion zu generieren (Randbedingungen: Anfangs- und Endpunkte beider Muster müssen identisch sein, die chronologische Reihenfolge der Zeitscheiben muß erhalten bleiben, keine Zeitscheibe darf ausgelasen werden). Die Bewertung einer Abbildungsfunktion, die das Testmuster dem jeweiligen Referenzmuster zuordnet, wird anhand eines lokalen Abstandsmaßes bestimmt. Mit Hilfe der dynamischen Programmierung wird die optimale Funktion ermittelt, die den Gesamtabstand minimiert. Zur Klassifikation eines Wortes/Testmusters wird für jedes Testmuster-Referenzmuster-Paar dieser Abstand bestimmt. Das Referenzmuster mit dem kleinsten Abstand stellt das erkannte Wort dar.

Kapitel 4

Audio-visuelle Spracherkennung beim Menschen

Daß visuelle Stimuli die Wahrnehmung von Sprachsignalen in fast allen ihren Dimensionen beeinflussen können, konnte in zahlreichen Experimenten gezeigt werden. Dies gilt für:

- Ort der Artikulation für Plosive [MM76],[Sum79],[FD91],[GK91],[MCG+93]
- Ort der Artikulation für Vokale [SM84],[MC93]
- Abschätzung der audiovisuellen VOT [RFM81],[GAKS85],[BP86]
- Wahrnehmung der Sprechgeschwindigkeit [Gre87]
- Interaktionen zwischen Stimmhaftigkeit und Ort der Artikulation [GK89], [GK91]

Auch Säuglinge sind schon empfänglich für die Korrespondenz zwischen auditorischer und visueller Information [KM82].
Es bleibt die Frage, warum die akustische und die visuelle Information kombiniert wird, da die Performanz bei der Spracherkennung in rauschfreien Umgebungen für den akustischen Kanal alleine nahezu perfekt ist. Eine Vorstellung wäre, die visuelle Sprache als eine Art Back-up-Device zu betrachten, auf welches in lauter Umgebung zurückgegriffen wird. Der visuelle Kanal liefert dann komplementäre Informationen und ist unabhängig von der gehörten Sprache. Der McGurk-Effekt (vgl. Abschnitt 4.4) zeigt aber, daß auch in rauschfreien Umgebungen der akustische und der visuelle Kanal fusioniert wird.
Das visuelle Signal kann auch dazu dienen, die Aufmerksamkeit der akustischen Analyse auf das Sprachsignal anstelle der umgebenden Störungen zu lenken (vgl. [Rei78]). Dabei kann die Information über die Person des Sprechers, d.h. sein Alter, Geschlecht, usw., dazu verwendet werden, die Anzahl der möglichen akustischen Muster zu begrenzen. Die Information, wo sich der Sprecher befindet, kann die Lokalisation der Stimme unterstützen, während sich die Information,

wann der Sprecher spricht, dazu verwenden läßt, das Sprachsignal von Hintergrundgeräuschen zu trennen. Weiterhin liefert das visuelle Signal redundante Informationen und kann dadurch Ausfälle beim akustischen Kanal kompensieren. Dies gilt vor allem für Vokale. Die redundante Information kann außerdem dazu genutzt werden, beide Kanäle in gegenseitiger Synchronität zu halten (vgl. auch [Sum87]).
In diesem Kapitel werden einige Aspekte der audio-visuellen Verarbeitung und Integration näher beleuchtet, wobei zunächst der Frage des quantitativen Nutzens der visuellen Informationen und der Komplementarität der Signale nachgegangen wird. Anschließend werden grundlegende Themen der audio-visuellen Spracherkennung wie die Visembildung und der McGurk-Effekt besprochen. Im Hinblick auf die Realisierung eines automatischen audio-visuellen Spracherkenners wird weiterhin darauf eingegangen, welche visuellen Merkmale für die Erkennung von Bedeutung sind und in welchem Maße beide Kanäle synchronisiert sein müssen. Abgeschlossen wird das Kapitel mit einer Darstellung der Ergebnisse psychologischer Untersuchungen, die die Frage der Integration beider Kanäle untersuchen.

4.1 Nutzen der Kombination akustischer und visueller Informationen

Schon sehr früh wurde festgestellt, daß sich die Verständlichkeit der Sprache verbessert, wenn zusätzlich zum akustischen Signal auch auf das visuelle zurückgegriffen werden kann [Cot35], insbesondere dann, wenn das akustische Signal gestört ist. Verschiedene psychophysische Experimente zeigen, daß in Umgebungen, in denen das akustische Signal gestört ist, die Sprachverständlichkeit durch die zusätzlichen Informationen über die Lippenbewegungen des Sprechers auch bei normalhörenden Menschen gesteigert wird. Der visuelle Anteil der audio-visuellen Spracherkennung ist dabei umso größer, je kleiner das Signal-Rausch-Verhältnis ist. Für sehr kleine Signal-Rausch-Verhältnisse [1] (d.h. ab ca. -24 dB [Erb69]) erfolgt im Prinzip eine rein visuelle Spracherkennung, so daß ab diesem Punkt der Anteil der visuellen Spracherkennung gleich bleibt [SP54] [BMK94] [Erb69] [Erb75] [BMJ74] [GB91] [Nee56] [Sum79][MS87],[RRSLE98],[O'N54].
Dabei kann bei audio-visueller Spracherkennung ein in etwa um 10 dB schlechteres Signal-Rausch-Verhältnis bei gleicher Spracherkennungsleistung toleriert werden gegenüber einer rein akustischen Spracherkennung. In hochverrauschten Umgebungen ist der Nutzen des visuellen Signals noch größer und kann einen Unterschied von etwa 15 dB im Signal-Rausch-Verhältnis ausgleichen [Erb69] [SP54], wobei die Größe des Nutzeffektes positiv mit der Lippenlesefähigkeit des Hörers korreliert ist [MS87],[EN71]. Die audio-visuelle Erkennungsleistung ist allerdings unter den Hörern wesentlich variabler als die rein akustische Erkennungsleistung, da die Lippenlesefähigkeiten bei untrainierten Versuchspersonen sehr unterschiedlich sein können [Erb69] (vgl. Abb. 4.1). So spielt auch das Alter der Versuchspersonen eine Rolle [Far59]. Auch sind die Ergebnisse für einzelne

[1] Das Signal-Rausch-Verhältnis soll zukünftig durch SNR (=signal-to-noise-ratio) abgekürzt werden.

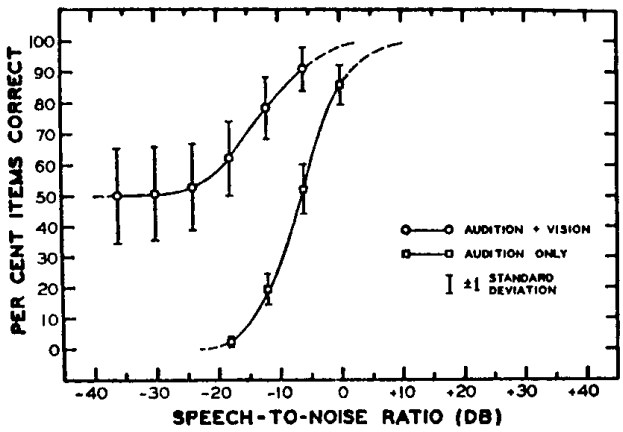

Abbildung 4.1: Audio-visuelle Erkennungsleistung von 5 normal hörenden Versuchspersonen, der akustische Kanal ist durch weißes, bei 6000 Hz tiefpaßgefiltertes Rauschen gestört, nach [Erb69]

Sprecher unterschiedlich [GB91],[Les88].
Das visuelle Signal wird allerdings auch dann genutzt, wenn das akustische völlig ungestört und intakt ist. Reisberg et al. [RMG87] führten dazu vier verschiedene Experimente durch, wobei die Erkennung der Sprache nicht dadurch erschwert wurde, daß das akustische Signal gestört wurde, sondern dadurch daß die zu erkennenden Texte schwer verständlich waren ('easy to hear, but hard to understand'). In zwei Experimenten sollten Texte in fremden Sprachen (Deutsch und Französisch) von englischsprachigen Versuchspersonen nachgesprochen werden, in weiteren Experimenten wurden englische Texte mit fremdsprachigem Akzent sowie ein semantisch komplexer Text verwendet. In allen Fällen konnte gezeigt werden, daß durch die zusätzliche visuelle Information die Performanz verbessert wurde. Die Autoren schließen aus diesen Ergebnissen, daß Sprachwahrnehmung ein einheitlich amodaler Prozeß ist, welcher immer die visuelle Wahrnehmung des Sprachsignals miteinbezieht, nur wird der visuelle Beitrag unter normalen Umständen nicht bemerkt, da im Allgemeinen das akustische Signal für eine fast perfekte Spracherkennungsleistung ausreicht.

4.2 Komplementarität des akustischen und des visuellen Sprachsignals

Ein Grund für den Nutzen der visuellen Information bei der Spracherkennung liegt in der Komplementarität des akustischen und visuellen Signals.
1955 untersuchten Miller und Nicely [MN55] die Verständlichkeit 16 verschiede-

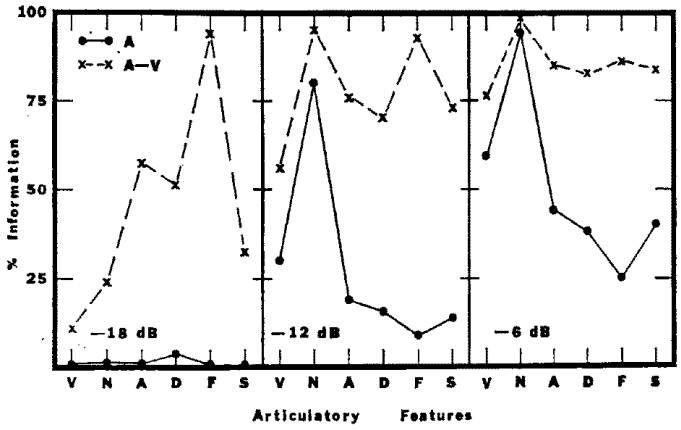

Abbildung 4.2: Prozentsatz an übertragener Information bei drei verschiedenen Signal-zu-Rausch-Verhältnissen unter Nutzung des akustischen Signals (A) und des akustisch-visuellen Signals (AV). Die Merkmale sind Stimmhaftigkeit (V), Nasalität (N), Affrikation (A), Dauer (D), Front Place of Artikulation (F) und Stop (S), nach [BMJ74]

ner englischer Konsonanten, deren Sprachsignal gestört war. Die Konsonanten ließen sich gemäß folgender artikulatorischer Merkmale in fünf Gruppen aufgliedern: Stimmhaftigkeit, Nasalität, Affrikation, Dauer und Ort der Artikulation. Dabei konnten sie feststellen, daß die Merkmale Stimmhaftigkeit sowie Nasalität durch die Störung am wenigsten in der Verständlichkeit litten, während das Merkmal 'Ort der Artikulation' am meisten durch die Störung in Mitleidenschaft gezogen wurde. Miller und Nicely mutmaßten, daß genau dieses Merkmal durch Beobachtung der Lippen eines Sprechers am leichtesten zu erkennen sei und schlossen mit der Feststellung: "Lip reading, therefore, is a valuable skill for listeners who are partially deafened because it provides just the information that the noise or deafness removes." ([MN55], S. 352).

Binnie et al. [BMJ74] führten daraufhin 1974 Experimente durch, um zu zeigen, in welchem Maß der visuelle Kanal tatsächlich zur Verbesserung der Verständlichkeit beiträgt. Abb. 4.2 zeigt ihre Resultate. Es zeigt sich, daß die Merkmale 'Stimmhaftigkeit' und 'Nasalität' am wenigsten durch zusätzliches Rauschen gestört wird, während das Merkmal 'Ort der Artikulation' am meisten durch Störungen beeinträchtigt wird. Die visuelle Information kann Verwechslungen

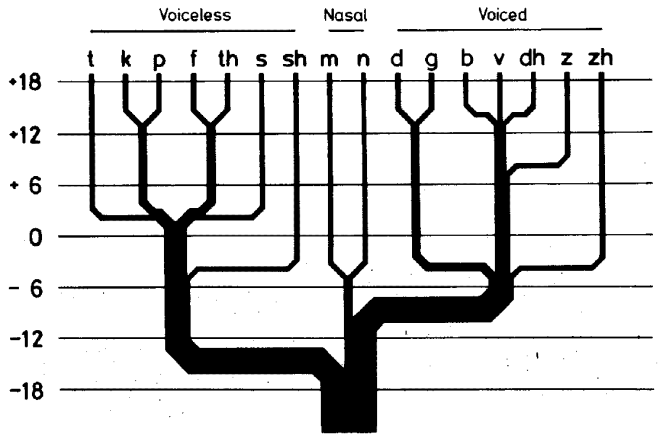

Abbildung 4.3: Auditorischer Verwechslungsbaum für Konsonanten. Die Konsonanten wurden dafür in einem CV-Kontext präsentiert mit unterschiedlichen Signal-zu-Rausch-Verhältnissen (-18 dB bis 18 dB), aufgetragen als Ordinatenachse. Gezeigt ist, ab welchem SNR Unterscheidungen zwischen den jeweiligen Konsonantengruppen möglich sind (in [Sum87], nach [MN55]).

signifikant reduzieren, wenn sich die verwechselten Phoneme im Ort der Artikulation unterscheiden. Bei niedrigem Signal-Rausch-Verhältnis ist der visuelle Beitrag zur Verständlichkeit der Phoneme am größten.
Walden et al. [WPW75] konnten im wesentlichen die Ergebnisse von Miller und Nicely sowie Binnie et al. durch Untersuchungen mit in verschiedenem Grad hörgeschädigten Versuchspersonen bestätigen.
In den Abbildungen 4.3 und 4.4 werden die Verwechslungsergebnisse für Konsonanten bei auditorischer [MN55] und bei visuellen Wahrnehmung [WPM+77] zusammengefaßt.
Beim auditorischen Verwechslungsbaum (vgl. Abb. 4.3) läßt sich erkennen, daß mit wachsendem Signal-Rausch-Verhältnis als erstes das Merkmal 'Stimmhaftigkeit' zur Unterscheidung der Konsonanten beiträgt gefolgt vom Merkmal 'Nasalität'. Erst bei einem SNR von etwa 15 dB können alle Konsonanten aufgrund des Merkmals 'Ort der Artikulation' sicher unterschieden werden. Dieses hat im wesentlichen drei Gründe: Zunächst sind 'Ortsanteile' im Sprachsignal weniger intensiv, als die Stimmhaftigkeits- und Nasalitätsanteile und werden daher zuerst durch Rauschen überdeckt. Weiterhin liegen die 'Ortsanteile' hauptsächlich in den mittleren und hohen Frequenzen des Sprachsignals und sind daher z. B. bei Hörverlust als erste betroffen. Zudem haben die 'Ortsanteile' im allgemeinen eine feinere Spektralstruktur und sind zum Teil auch sehr kurz, so daß zur Erkennung eine gute Frequenz- und Zeitauflösung notwendig ist.

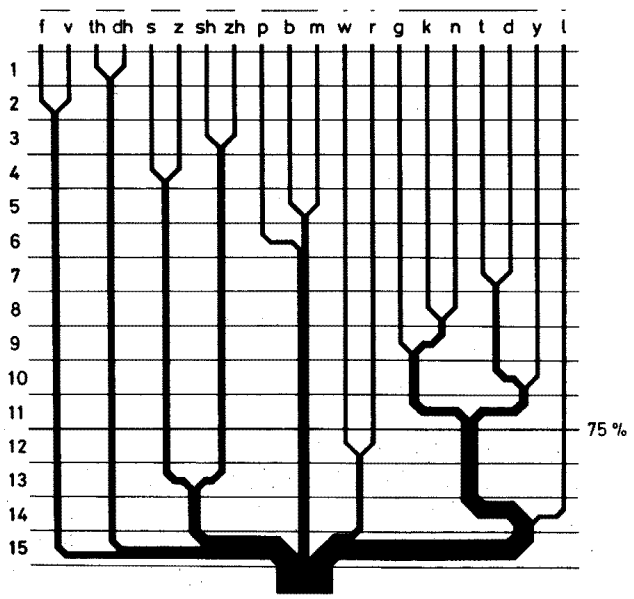

Abbildung 4.4: Visueller Verwechslungsbaum für Konsonanten. Die Konsonanten wurden dafür in einem CV-Kontext präsentiert. Die Darstellung zeigt die Ergebnisse einer hierarchischen Clusteranalyse der visuellen Verwechslungen bei trainierten hörgeschädigten Versuchspersonen. Die 75 % Markierung kann zur Bestimmung von Visemklassen genutzt werden, d.h. bei 75 % der Präsentationen werden Konsonanten, die oberhalb dieser Markierung verbunden sind, als Mitglieder ihrer Gruppe identifiziert (in [Sum87], nach [WPM+77].

Im Gegensatz dazu können die 'Ortsinformationen' bei visueller Sprachwahrnehmung besonders leicht unterschieden werden, wie Abb 4.4 zeigt (vgl. /k/ und /p/, diese werden akustisch leicht verwechselt), während Konsonanten, die sich nur durch die Merkmale 'Stimmhaftigkeit' oder 'Art der Artikulation' (z. B. Nasalität) unterscheiden, verwechselt werden (vgl. /p/, /b/ und /m/).

4.3 Viseme

Schon E. Nitchie [Nit16] bemerkte, daß sich einige Laute visuell nicht von anderen unterscheiden lassen. In den darauffolgenden Jahren wurden einige Experimente durchgeführt, um minimale Einheiten des Lippenlesens zu finden, so z.B. von [WB60]. Fisher prägte schließlich 1968 in [Fis68] für solche Gruppen von visuell nicht zu unterscheidenden Phonemen den Begriff "visual phonemes" oder kürzer Visem.
Um Viseme zu bestimmen, können z.B. Verwechslungsbäume oder Verwechslungsmatrizen verwendet werden wie sie z. B. in Abbildung 2.4. gezeigt sind. Viseme werden als Phonemgruppen angegeben, die untereinander verwechselt werden, so besteht z. B. ein Visem, das in fast allen Untersuchungen gefunden

wurde, aus den Phonemen /p/,/b/ und /m/.
Die meisten Untersuchungen beschäftigen sich mit der Aufteilung von Konsonanten in Viseme. Bisher konnte allerdings noch kein Standardsatz gefunden werden, da die Anzahl der Viseme von unterschiedlichen Faktoren abhängt.
So stellten Walden et al. [WPM+77] fest, daß ein Training der Versuchspersonen die Genauigkeit der Erkennung als auch die Anzahl der wahrgenommenen Viseme von 5 auf 9 erhöhen konnte. Sie wiesen allerdings darauf hin, daß die Verbesserungen der Fähigkeiten auch durch eine größere Vertrautheit der Versuchspersonen mit Sprecher und Versuchsbedingungen zustande kommen könnte [2].
Auch der Sprecher hat einen Einfluß auf das Visemset, wobei die Anzahl der Viseme von 4 bis 8 je nach Sprecher schwanken kann (vgl. [KL82],[KL85],[Les88]).
Weiterhin wurden Konsonantencluster auf ihre Visemgruppierungen untersucht [FK72], es ergaben sich 7 verschiedene Cluster.
Untersuchungen mit verschiedenen Vokalen zeigen, daß Vokale zwar unterschiedlich gut visuell erkannt werden können, daß sich aber nur schlecht distinktive Gruppen von Vokalen bilden lassen, da die sichtbaren Unterschiede zwischen den Vokalen eher graduell und kontinuierlich sind [MSB84], Diphthonge lassen sich besser als Vokale voneinander unterscheiden, da dort auch die zum Teil extreme Gleitbewegung von einem Vokal zum anderen visuell erkannt werden kann [WJ79]. Die individuellen Unterschiede bei den Lippenlesefähigkeiten bei Vokalen sind größer als bei Konsonanten, dementsprechend sind Viseme schwieriger zu bestimmen [MSB84]. Zudem sind Vokale im Kontext mit gut sichtbaren Konsonanten schlecht unterscheidbar [MWP87], aber auch die Lippenlese-Performanz bei Konsonanten erweist sich als kontextabhängig [BPF82].
Weitere Einflußmöglichkeiten sind der Blickwinkel, unter dem das Gesicht beim Sprechen gesehen wird [WJ79] und der Vokalkontext, in dem der Konsonant präsentiert wird [OB85]. So unterscheiden sich die Viseme im Vokalkontext mit /u/ beträchtlich von denen anderer Vokalkontexte.
Einen Überblick über die Visemgruppierungen der unterschiedlichen Untersuchungen ist in [Jac88] zu finden.
Abschließend sei noch bemerkt, daß Berger in [Ber72] feststellen konnte, daß es beim Lippenlesen nicht möglich ist, ein initiales /h/ zu erkennen, das somit keiner Visemklasse zugeordnet werden kann.

4.4 McGurk-Effekt

Im Jahr 1976 untersuchten Harry McGurk und John MacDonald in einem mittlerweile berühmt gewordenen Experiment [MM76], wie sich widerstreitende akustische und visuelle Stimuli auf die Sprachwahrnehmung normalhörender Versuchspersonen auswirken. Zu diesem Zweck wurden den Versuchspersonen Videobänder einer Sprecherin vorgeführt, die einsilbige Wörter der Art [ba,ga,ka,pa] äußerte. Sowohl das Bild- als auch das Sprachmaterial waren störungsfrei und die Silben konnten aufgrund des akustischen Materials nahezu

[2] Die Frage, ob ein Training wirklich die Lippenlesefähigkeiten eines Hörers steigern kann, ist immer noch umstritten (vgl. Diskussion bei [Gre98], S. 17 und [Sum92]).

fehlerfrei erkannt werden. Um zu untersuchen wie sich die zusätzliche visuelle Information auf die Sprachwahrnehmung auswirkt, wurden die akustischen und visuellen Aufnahmen von verschiedenen Wörtern miteinander kombiniert und vorgeführt. So wurde z.B. ein akustisches [ba] mit einem visuellen [ga] verbunden. Als Versuchspersonen fungierten Kinder im Alter von 3 bis 8 Jahren und Erwachsene. Die widerstreitende Information der Kanäle hatte einen grundlegenden Effekt auf die Erkennungsleistung der einzelnen Silben, die nur noch zum geringeren Teil korrekt [3] klassifiziert wurden. Vielmehr wurden Kombinationen beider Silben oder Silben erkannt, die weder im akustischen noch im visuellen Signal enthalten waren. So wurde z. B. die Kombination aus auditorischem [ba] und visuellem [ga] überwiegend als [da] empfunden, der umgekehrte Fall wurde hingegen meistens als eine Kombination aus beiden Klassen, z.B. als [gabga] erkannt [4]. Weiterhin konnten McGurk und MacDonald feststellen, daß die erwachsenen Versuchspersonen bei ihrer Entscheidung eher durch den visuellen Input und die Kinder mehr durch den akustischen Input beeinflußt wurden, sobald sie keine kombinatorische Antwort gaben. Die Wahrnehmung beim McGurk-Effekt wird von den Versuchspersonen als Höreindruck beschrieben und tritt selbst dann auf, wenn die Versuchspersonen sich über die unterschiedlichen Informationen der beiden Kanäle bewußt waren und aufgefordert wurden, den visuellen Kanal zu ignorieren [SM84],[MC83].
McGurk und MacDonald stellten die Hypothese auf, daß bei audio-visueller Spracherkennung die Art der Artikulation und die Stimmhaftigkeit bei Konsonanten akustisch bestimmt wird, während der Ort der Artikulation visuell erkannt wird. Am Beispiel des 'klassischen' McGurk-Effektes (visuelles /ga/, akustisches /ba/) bedeutet dies, daß der Artikulationsort des visuellen Signals (/ga/ ist nonlabial) die auditorische Information für den Ort (/ba/ ist labial) ersetzt; fusioniert mit der verbleibenden akustischen Information (/ba/ ist plosiv und stimmhaft) ergibt sich als Wahrnehmung ein nonlabialer, plosiver und stimmhafter Laut: /da/. In ihren Untersuchungen konnten sie diese Hypothese nur zum Teil untermauern (für labial-akustisch/nichtlabial-visuelle Präsentationen), sie stellten aber auch dieser Theorie widersprechende Resultate fest.
Der McGurk-Effekt wird mittlerweile in zahlreichen Untersuchungen der audiovisuellen Wahrnehmung beim Menschen eingesetzt. Er konnte auch mit synthetischen Stimuli nachvollzogen werden, in denen das Videoband durch einen sogenannten 'Talking Head' ersetzt wurde [MC83].[MC95].
Der McGurk-Effekt wurde auch bei den Vokalen /a/,/i/ und /u/ gefunden [SM84], und zwar wurde jeweils ein Vokal wahrgenommen, der eher dem visuell präsentierten entsprach.
Die Abhängigkeit des Effekts von der Vokalumgebung der Konsonanten untersuchte Green in [GKM88]. Dabei ergab sich, daß der Effekt für eine /u/-Umgebung schwächer als für /a/ und /i/ ist. Für die Vokalumgebungen /a/ und /i/ ergaben sich zudem unterschiedliche Antwortmuster: für /a/ erhielt man mehr /th/-Fusionsantworten, während in der /i/-Umgebung die /d/-Antworten überwiegen.

[3] Korrekt bedeutet hier, daß die Klassifikation gemäß der akustischen Information erfolgte.
[4] Die Kombination aus auditorischem /pa/ und visuellem /ka/ wurde als /ta/ erkannt.

Eine Inkongruenz zwischen akustischem und visuellem Kanal hat hingegen keine Einfluß auf die Größe des McGurk-Effekts, wie Green et al. [GKMS91] zeigen konnten. Sie benutzten dazu eine weibliche Stimme zusammen mit Videoaufnahmen eines männlichen Sprechers und umgekehrt [5].
Auch hinsichtlich einer zeitlichen Asychronität zwischen akustischen und visuellem Signal ist der Effekt robust. Green und Gerdeman [GG95] konnten zeigen, daß eine Asynchronität von bis zu 225 ms zwischen beiden Kanälen den Effekt nicht beeinflußt, obwohl die Asynchronität von den Versuchspersonen durchaus beobachtet wird. Munhall et al. [MGSW96] kamen zu ähnlichen Ergebnissen. Präsentiert man hingegen die Konsonanten mit nicht übereinstimmenden Koartikulationscharakteristika, indem man z.B. ein akustisches /bi/ mit einem visuellen /ga/ kombiniert, beeinflußt dies den McGurk-Effekt dahingehend, daß er bei den inkongruenten Stimuli kleiner wird [GG95].
Die unterschiedliche Ausprägung des McGurk-Effekt bei Hörern unterschiedlichen Alters, wie sie in der Original-Untersuchung festgestellt wurde, war Gegenstand mehrerer Untersuchungen. Es zeigte sich, daß allein die Versuchsdurchführung zu beträchtlichen Unterschieden führten, obgleich es durchgehend der Fall ist, daß jüngere Kinder einen schwächeren Effekt als ältere zeigen [Gre98].
Sekiyama und Tohkura [ST91] stellten fest, daß der McGurk-Effekt im Japanischen bei vollständiger Verständlichkeit des akustischen Signals nur sehr schwach ausgeprägt ist und erst, wenn der akustische Kanal gestört wird, stark zum Tragen kommt. Dabei kommen Kombinationsantworten nur sehr selten vor. Vergleichende Untersuchungen zwischen japanischen und amerikanischen Sprechern und Hörern zeigte, daß bei den amerikanischen Hörern der McGurk-Effekt stärker auftrat als bei den japanischen. Zudem war der Effekt mit 'fremdsprachlichen' Stimuli bei jeder der beiden Gruppen größer, als bei Stimuli der eigenen Sprache [ST93]. Verschiedene Erklärungen wurden für dieses Phänomen angegeben: Zunächst ist es möglich, daß kulturelle Einflüsse eine Rolle spielen, da Japaner dazu tendieren, das Gesicht des Sprechers nicht anzusehen, weil dies als unhöflich angesehen wird. Diese Erklärung wird durch zusätzliche Untersuchungen mit chinesischen Versuchspersonen gestärkt, die einen ähnlich schwachen McGurk-Effekt zeigen [STU96] [Sek97]. Weiterhin wäre es möglich, daß im Japanischen die Integration visueller Information nicht notwendig ist, da es keine Konsonantcluster, nur 5 Vokale und keine visuell unterschiedlichen Konsonanten gibt [ST93]. Auch sind japanische Versuchspersonen empfänglicher für die audio-visuellen Diskrepanzen und integrieren daher beide Informationen nicht [Sek94].
Green und Gerdeman [GG95] schlugen einen Zusammenhang mit dem Vokalkontext, in dem die Stimuli präsentiert werden, vor, da der verwendete /a/-Kontext im Englischen eher zu /th/-Antworten führt, welche im Japanischen nicht vorkommen.
Ähnliche Untersuchungen wurden auch für andere Sprachen durchgeführt: Finnisch (starker McGurk-Effekt) [SSHK97], [SMS+98], Niederländisch und Kan-

[5] Walker et al. [WBO95] konnten allerdings zeigen, daß im Fall von inkongruenten Stimuli der McGurk-Effekt reduziert wird, wenn dem Hörer der Sprecher bekannt ist.

tonesisch [dGBVC95], Spanisch und Deutsch [Dur95],[FD96], Niederländisch und Englisch [MCS95], wobei hier einige der Ergebnisse durch Unterschiede in den phonetischen Systemen der untersuchten Sprachen erklärt werden könnten [FD96],[MCS95].
Untersuchungen des McGurk-Effektes mit verschiedenen Sprachen bzw. Sprachhintergründen werden mittlerweile zur Untersuchung verschiedener Fragestellungen in der audio-visuellen Spracherkennung benutzt und lassen sich nach [Bur98] in drei Gruppen einteilen:

- cross-cultural studies: Dabei wird das Auftreten des McGurk-Effektes in unterschiedlichen Sprachen untersucht, um festzustellen, ob es z.B. Unterschiede bei der Bewertung der akustischen und visuellen Komponenten in verschiedenen Sprachen gibt.

- inter-lingual studies: In diesen Studien wird untersucht, welchen Effekt die Präsentation von audio-visueller Sprache von Sprechern einer fremden Sprache hat im Vergleich zu audio-visueller Sprache von Sprechern der eigenen Muttersprache.

- cross-language studies: Bei diesen Untersuchungen werden die phonetischen Unterschiede der einzelnen Sprachen dazu genutzt, um Fragestellungen wie die Art und den Zeitpunkt der Integration der akustischen und visuellen Information in der menschlichen Sprachwahrnehmung zu untersuchen.

Viele Untersuchungen nutzen den McGurk-Effekt, auch deswegen, weil er bei intakten Stimuli auftritt. An der Größe des Effekts (d.h. an der Anzahl der Antworten, die nicht dem akustischen Stimulus entsprechen) läßt sich erkennen, ob die visuelle Information die phonetische Analyse des auditorischen Signals beeinflußt, und das Muster der Antworten gibt Hinweise darauf, welche Art der Informationen in den zwei Modalitäten genutzt werden.

4.5 Visuelle Merkmale

Ein weiterer wichtiger Gesichtspunkt des Lippenlesens, der auch für die Entwicklung von audio-visuellen Spracherkennungssystemen wichtig ist, ist die Frage, welche Teile des Gesichts für die audio-visuelle Sprachwahrnehmung die größte Bedeutung haben.
Summerfield untersuchte in [Sum79] verschiedene Arten der visuellen Information: das ganze Gesicht, nur die Lippen, sowie nur vier Punkte, die die Mundwinkel sowie die Mitten der Ober- und Unterlippe kennzeichneten. Dabei wurde die Performanz der Lippenleser nur für die ersten beiden Merkmale signifikant besser, wobei die Ergebnisse für das ganze Gesicht die besten waren. Bei der Verwendung der reinen Lippenbilder führte insbesondere das Fehlen der Zähne und Zungenbewegungen zu dem schlechteren Ergebnis. Im Falle der vier Punkte war die visuelle Information zu unspezifisch, da diese Repräsentation keine Information über die Öffnung zwischen den Lippen lieferte, woraus geschlossen werden kann, daß dies ein wichtiges Merkmal ist.

Der Winkel, unter dem der Hörer das Gesicht des Sprechers sieht, kann ebenfalls die Verständlichkeit der lippengelesenen Sprache beeinflussen, doch besteht in diesem Punkt noch keine Einigung. Neely [Nee56] stellte fest, das die Erkennungsleistung lippengelesener Sprache bei einer 'Face-to-face'-Situation (d.h. ein Sehwinkel von 0 Grad) am besten ist, bei 45 und 90 Grad dagegen wird sie immer schlechter. Die Entfernung des Hörers vom Sprecher dagegen spielt in einem Bereich von 1 bis 3 m keine Rolle. Wozniak und Jackson [WJ79] konnten dagegen keine Unterschiede zwischen den Erkennungsleistungen bei 0 und bei 90 Grad feststellen.
Jackson et al. [JMB76] stellten bei Untersuchungen mit Vokalen und Diphthongen fest, daß die Lippenrundung bzw. das Hervorstrecken der Lippen sowie die vertikale Lippenöffnung die wichtigsten Merkmale sind, gefolgt von dem Gebiet der maximalen Lippenöffnung. Bei den Diphthongen kamen noch die vertikalen Bewegungen und die Lippenöffnung gegen Ende der Äußerung hinzu.
Summerfield et al. untersuchten in [SMMB89] den Beitrag der Lippen und Zähne gegenüber dem des ganzen Gesichts (dieser Beitrag entspricht 100 %). Sie stellten fest, daß der Beitrag der Lippen zur visuellen Erkennungsleistung bei ca. 50 % liegt, die zusätzliche Information über die Zähne liefert einen weiteren Beitrag von ca. 6 % und hilft hauptsächlich bei der Erkennung der Frontvokale.
In [LM94], [VBEY94] und [VBMH+95] wurden die Augenbewegungen des Hörers beobachtet, um daraus auf wichtige Merkmale der visuellen Sprachwahrnehmung zu schließen. Es konnte festgestellt werden, daß neben dem Mund den Augen große Aufmerksamkeit geschenkt wurde, wobei bei verrauschtem akustischen Kanal der Mund mehr beachtet wurde, allerdings betrug der zeitliche Anteil nie mehr als 55 % [MVB98]. Der Hörer nutzt wahrscheinlich globalere Informationen als nur diejenigen über die Mundbewegungen alleine. Aufgrund der engen Beziehungen zwischen den Bewegungen von Gesichtsregionen, die den Lippen benachbart sind (z.B. das Kinn) und den eigentlichen Bewegungen der Artikulatoren, ist es nicht notwendig für den Hörer, die Lippen direkt zu fixieren. Möglicherweise nutzt der Hörer auch die Informationen über die Position der Augen, um bei Kopfbewegungen des Sprechers auf die Kopfposition und daraus auf die relativen Lippenbewegungen rückschließen zu können.
Benoit et al. [BGMGA95] und Le Goff et al. [GGMB95] untersuchten mit Hilfe eines synthetischen Sprechers, welche Anteile des Gesichts (Lippen, Lippen und Kiefer, synthetisches bzw. natürliches Gesicht) unter verschiedenen Störbedingungen des akustischen Kanals am meisten zur visuellen Spracherkennung beitragen. Es stellte sich heraus, daß die Spracherkennung umso besser ist, umso mehr Informationen über das Gesicht gegeben werden. Dabei war außerdem eine natürliche Präsentation wichtig: ein Schädel mit Lippenbewegungen, der exakt dieselben Informationen wie ein synthetisches Gesicht enthielt, wurde trotzdem schlechter klassifiziert. Für eine maschinelle Erkennung war eine Frontalansicht des Gesichts besser geeignet als eine Seitenansicht.
Zusammenfassend kann gesagt werden, daß das ganze Gesicht des Sprechers während des Lippenlesens als Informationsquelle herangezogen wird, wobei allerdings die Lippen den größten Teil der Information bereitstellen. Die Lippenbewegungen selbst können durch sorgfältige Auswahl von Merkmalen wie Lippenrundung oder vertikale Lippenöffnung beschrieben werden, doch ist die

richtige Auswahl dieser Merkmale schwierig.

4.6 Synchronität von akustischem und visuellem Kanal

Daß eine Asynchronität zwischen akustischem und visuellem Kanal vom Hörer bemerkt wird, ist z.B. aus schlecht synchronisierten Filmen bekannt. Quantitative Untersuchungen dieses Phänomens waren Gegenstand einer Untersuchung von Campbell und Dodd [CD80]. Versuchspersonen sollten audio-visuell präsentierte Worte in weißem Rauschen identifizieren, wobei die lippengelesene Information der auditorischen um 0, 400, 800 oder 1600 ms voranging. Die Ergebnisse zeigten, daß das Sehen des Sprechers die Erkennungsleistung im Vergleich zur auditorischen allein verbesserte, wobei die Länge des Delays die Performanz nicht beeinflußte.
Dixon und Spitz [DS80] untersuchten auch den Effekt, der hervorgerufen wird, wenn das akustische Signal dem visuellen vorangeht. Es stellte sich heraus, daß beim Vorausgehen des akustischen Signals um bis zu 130 ms die Desynchronität nicht detektiert wurde, beim umgekehrten Fall wurde eine Desynchronität um bis zu 250 ms toleriert. Diese Asymmetrie der Delays konnte auch an anderer Stelle bestätigt werden mit Werten von 79 ms für ein vorangehendes akustisches Signal und 138 ms für ein folgendes akustisches Signal (vgl. [Sum92],[MS85]).

4.7 Psychologische Untersuchungen der Integrationsfrage

Die Integration akustischer und visueller Signale bei der Sprachwahrnehmung ist Gegenstand zahlreicher psychologischer und psychophysischer Untersuchungen.
Dabei stehen folgende Fragestellungen im Mittelpunkt:

- Welche Metrik, wenn überhaupt, wird bei der Integration beider Kanäle genutzt ?
- An welcher Stelle werden der akustische und der visuelle Kanal kombiniert, d.h. an welcher Stelle der Sprachverarbeitung geht die modalitätsspezifische Verarbeitung in eine modalitätsunabhängige Verarbeitung über ?

Hierbei werden im wesentlichen zwei mögliche Arten der Integration unterschieden: Die frühe und die späte Integration. Man versteht dabei unter früher Integration ein Modell, in dem beide Informationen zusammengeführt werden, bevor eine phonetische oder phonemische Verarbeitung des Sprachsignals erfolgt (z.B. indem phonetische Merkmale, wie der Ort der Artikulation extrahiert werden). In einem Modell der späten Integration werden die Informationen zunächst in eine linguistische Repräsentation, sei diese phonetisch, phonemisch oder auch syllabisch, überführt, bevor die Integration stattfindet [Gre98]. Über die Art

dieser linguistischen Repräsentation besteht zur Zeit noch keine Übereinstimmung.
Beide Fragestellungen sind dabei eng miteinander verknüpft. In den folgenden Abschnitten sollen verschiedene Überlegungen und Untersuchungen unterschiedlicher Arbeitsgruppen vorgestellt werden.

4.7.1 Welche Metrik könnte zur Integration genutzt werden ?

In [Sum87] untersuchte Summerfield fünf verschiedene Repräsentationsmöglichkeiten, welche für eine Fusion der akustischen und der visuellen Informationen in Frage kommen. Dabei handelte es sich um

- phonetische Merkmale: Nach der VPAM-Hypothese [6] werden zunächst für jede Modalität getrennt phonetische Merkmale bestimmt. Anschließend werden diese selektiv kombiniert in der Art, daß die 'Art-der-Artikulation'-Merkmale sowie das Stimmhaftigkeitsmerkmal und das Nasalitätsmerkmal des akustischen Kanals und die 'Ort-der-Artikulation'-Merkmale des visuellen Kanals übernommen werden. Summerfield fand in seiner Arbeit Evidenzen, die gegen ein solches Modell sprechen. So kann z.B. dieses Modell nicht die Kombinationsantworten des McGurk-Effektes erklären, wenn ein akustisches [g] kombiniert mit visuellem [b] z.B. als [bg] wahrgenommen wird [Sum79].

- die Filterfunktion des Vokaltraktes: Die Quellencharakteristik kann dabei aus dem akustischen Signal und die Filterfunktion selbst aus dem akustischen und dem visuellen abgeschätzt werden. Dabei weist Summerfield daraufhin, daß aus diesen Informationen allein die Filterfunktion nicht exakt genug abgeschätzt werden kann, sondern daß der Hörer vermutlich weiteres Wissen über die audio-visuelle Struktur der Phoneme in der benutzten Sprache nutzen muß. Bei der Integration beider Kanäle wird wahrscheinlich keine einfache Durchschnittsbildung der beiden Transferfunktionen vorgenommen, da Untersuchungen gezeigt haben, daß die Wahrnehmung eines Konsonanten davon abhängt, in welcher Modalität die entsprechende Filterfunktion berechnet wurde. Diese Metrik entspricht einer auditorischen Repräsentation.

- akustische Spektren und visuelle Muster: Bei dieser Metrik werden der akustische und der visuelle Kanal dadurch integriert, daß das akustische Spektrum und die visuellen Muster einfach aneinandergehängt werden, ohne daß eine weitere Vorverarbeitung stattfindet. [7]

- (aufeinanderfolgende) statische artikulatorische Konfigurationen: Bei dieser Metrik werden die Informationen des akustischen und des visuellen

[6] VPAM = vision: place, audition: manner

[7] Robert-Ribes et al. [RRSE95b] bemerken dazu, daß ohne gemeinsame intermediäre Repräsentation kein Vergleich zwischen beiden Modalitäten möglich wäre und daß Menschen diesen Vergleich ziehen könnten, da sie Asynchronitäten detektieren und auch Inkompabilitäten zwischen beiden Kanälen registrieren könnten. Trotzdem würden beide Kanäle noch fusioniert.

Kanals dazu genutzt, ein 3D-Modell des Vokaltraktes zu erstellen, wobei die visuelle Information zur Berechnung des vorderen und die akustische zur Berechnung des hinteren Vokaltraktes hinzugezogen wird.

- zeitveränderliche kinematische Muster, welche Evidenzen über die artikulatorische Dynamik liefern: Bei diesem Verfahren werden aus den akustischen und den visuellen Informationen Parameter bestimmt, die die artikulatorische Dynamik des Sprechvorganges beschreiben. Diese Metrik ist im Gegensatz zu den anderen vorgestellten Metriken modalitätenunabhängig.

Die Überlegungen Summerfields deuten eher auf ein Modell der frühen Fusion hin, da die Fusion eher in einer Metrik stattfindet, die noch vor einer phonetischen Kategorisierung liegt.

4.7.2 Wo findet die Integration statt ?

In den folgenden Abschnitten werden die wichtigsten Untersuchungen zu diesem Thema kurz vorgestellt. Zunächst sollen die Experimente erläutert werden, die eher auf eine frühe Fusion beider Kanäle hindeuten, gefolgt von denen, die durch eine späte Integration gedeutet werden. Eine kurze Diskussion schließt diesen Abschnitt ab.

Experimentelle Ergebnisse: frühe Integration

In den Untersuchungen von Risberg und Lubker [RL78], Rosen et al [RFM81], Grant et al. [GAKS85] und Breeuwer und Plomp [BP86] wurden Informationen über die Grundfrequenz des Sprachsignals (präsentiert z.B. als Pulstöne) mit visuellen Informationen gekoppelt. Obwohl die Grundfrequenzinformationen für sich genommen völlig unverständlich waren und auch aus den visuellen Informationen alleine keine Stimmhaftigkeitsinformationen gezogen werden konnten, konnten die Versuchspersonen die Stimmhaftigkeit der Stimuli identifizieren, wenn beide Modalitäten gekoppelt wurden. Dies kann dadurch erklärt werden, daß die Versuchspersonen die zeitliche Koordination zwischen beiden Signalen extrahieren können und damit einen audio-visuellen VOT bestimmen können. Diese Untersuchungen liefern einen deutlichen Hinweis darauf, daß die Integration der auditorischen und visuellen Signale vor einer phonetischen Kategorisierung stattfindet, zumindest was die Stimmhaftigkeit betrifft.
In ihren Untersuchungen von 1985 [GM85] gingen Green und Miller von der Beobachtung aus, daß sich die VOT-Grenze z.b. in einem /bi/-/pi/-Kontinuum mit der Sprechgeschwindigkeit ändert. Sie untersuchten daraufhin die Frage, ob auch eine visuell präsentierte Sprechgeschwindigkeit Einfluß auf die Lage der VOT-Grenze hat. Dazu wurden auditorische /bi-/pi/-Silben, die mit mittlerer Sprechgeschwindigkeit geäußert wurden, mit visuellen /bi/ oder /pi/-Silben gekoppelt, die mit einer höheren oder niedrigeren Sprechgeschwindigkeit geäußert wurden, aber alleine nicht korrekt als /bi/ oder /pi/ klassifiziert werden konnten. Sie konnten zeigen, daß sowohl die Sprechgeschwindigkeit verändert wahrgenommen wurde als auch, daß sich die VOT-Grenze verschob. Dies zeigt, daß

auch die visuelle Information der Sprechgeschwindigkeit extrahiert und mit der auditorischen Information kombiniert wird, bevor eine Entscheidung über die Stimmhaftigkeit der Äußerung getroffen wird.
Braida [Bra91] bewertete mit Hilfe der Ergebnisse anderer Studien, die die audio-visuelle Spracherkennungsleistung von Versuchspersonen mit z. T. verrauschtem akustischen Signal untersuchten (vgl. [BMJ74] etc.), drei Integrationsmodelle, die jeweils aus der unimodalen Performanz der Versuchspersonen die bimodale Performanz vorhersagten. Diese Vorhersage wurde mit der tatsächlichen Performanz verglichen und danach die Modelle bewertet. Bei den Modellen handelte es sich um das FLMP (vgl. Seite 65), ein pre-labelling-Modell, das (mit Hilfe abstrakter Merkmale) kontinuierliche Sensordaten kombiniert, bevor die Daten kategorisiert werden, und ein post-labelling-Modell, bei dem die Antworten, die in den unimodalen Konditionen gemacht wurden, kombiniert werden zu einer gemeinsamen Antwort. Es ergab sich, daß das pre-labelling-Modell die Vorhersage am besten treffen konnte, gefolgt vom FLMP. Das post-labelling-Modell schnitt am schlechtesten ab.
Der Ausgangspunkt der Untersuchungen von Green und Kuhl [GK89] war die Beobachtung, daß sich in einem VOT-Kontinuum die wahrgenommene Grenze zwischen Stimmhaftigkeit und Stimmlosigkeit verschiebt, wenn sich der Ort der Artikulation des phonetischen Segmentes verändert. Zum Beispiel liegt die Stimmhaftigkeitsgrenze in einem bilabialen /b-p/-Kontinuum bei einem kürzeren VOT als die Stimmhaftigkeitsgrenze in einem alveolaren /d-t/-Kontinuum. Im Experiment wurde ein akustisches /ibi/-/ipi/ bzw. /idi/-/iti/-Kontinuum mit visuellen Bildern von /igi/ gekoppelt, beides wurde vom Hörer als /idi-iti/-Kontinumm wahrgenommen. Sie verglichen, ob sich die VOT-Grenze bei der rein akustischen Präsentation von der der akustisch und visuellen Präsentation unterschied. Sie stellten fest, daß sich bei einer Änderung der Wahrnehmung auch die VOT-Grenze verschob, d.h. im Fall des /ibi-ipi/-Kontinuums gekoppelt mit dem visuellen /igi/ verschob sich die VOT-Grenze wie es bei einem alveolaren akustischen Kontinuum der Fall wäre. Daraus konnten sie schließsen, daß der Ort der Artikulation aus der Kombination der akustischen und der visuellen Information bestimmt wird und zwar, bevor eine Stimmhaftigkeitsentscheidung gefällt wird.
Green und Gerdeman (s. [GG95]) variierten dieses Experiment, indem die auditorischen Stimuli asynchron zu den visuellen dargeboten wurden, und zwar wurden sie um bis zu 250 ms **vor** den visuellen präsentiert, so daß die Information über die Stimmhaftigkeit dem Wahrnehmungssystem schon zur Verfügung steht, bevor kombiniert werden kann. Die VOT-Grenze wurde wie im Experiment von Green und Kuhl [GK89] verschoben, d.h die auditorischen Informationen über die Stimmhaftigkeit und die visuelle Informationen über den Ort der Artikulation werden integral und nicht separat verarbeitet.
In einer weiteren Untersuchung [GK91] wurde ein selektives Aufmerksamkeitsschema dazu verwendet, zu bestimmen, ob die akustisch präsentierte phonetische Information separat oder integral mit der visuellen verarbeitet wird. Unter Ausnutzung des McGurk-Effektes wurden Stimuli variiert, wobei das Ortsmerk-

mal visuell und das Stimmhaftigkeitsmerkmal akustisch präsentiert wurde [8]. Es zeigte sich, daß die Entscheidung über die Stimmhaftigkeit durch Variation der Ortskomponente beeinflußt wurde und umgekehrt, d.h. beide Dimensionen wurden integral verarbeitet, obwohl sie in zwei verschiedenen Modalitäten präsentiert wurden. Dieses Ergebnis stützt die Ergebnisse der beiden vorher beschriebenen Experimente[9].
Strand und Johnson [SJ96] untersuchten, wie die Kategoriegrenze zwischen /s/ und /ʃ/ durch die Geschlechtscharakteristik des Sprechers beeinflußt wird, der den nachfolgenden Vokal äußert. Dazu wurde ein von einem männlichen bzw. weiblichen Sprecher geäußertes auditorisches /s-ʃ/-Kontinuum mit nachfolgendem Vokal den Versuchspersonen präsentiert, wobei sich eine unterschiedliche Grenze zwischen /s/ und /ʃ/ ergab. Wurde ein hinsichtlich des Geschlechts des Sprechers zweideutiges /s-ʃ/-Kontinuum mit einem männlichen bzw. weiblichen Gesicht gekoppelt, konnte diese Verschiebung reproduziert werden. Dies deutet auf eine frühe Fusion beider Informationskanäle hin. Green und Gerdemann nutzten in [GG95] die Tatsache, daß ein Mismatch in der Koartikulationsinformation zwischen einem Konsonanten und dem nachfolgenden Vokal die Zeitspanne vergrößert, die zur Klassifikation des Vokals benötigt wird. In ihrem Versuch koppelten sie auditorische /ba/- und /bi/-Silben mit visuellen /gi/- und /ga/-Silben. Bei konsistenten Vokalen betrug die Stärke des McGurk-Effektes etwa 75 %, dieser Wert fiel für nichtkonsistente Paarungen auf 44 %, d.h. die Versuchspersonen waren sensitiv für die Koartikulationsinformationen, wenn akustische und visuelle Signale kombiniert wurden.
Norrix und Green nutzten 1996 Koartikulationseffekte zwischen Konsonantclustern in einem aufwendigen Versuch, deren Ergebnisse die Annahme einer frühen Fusion beider Kanäle stützen ([NG96],[Gre96a]).
Robert-Ribes et al. [RRSE95b], [RRSE94], [RRSE95a], [RRPSE96] stellten in ihren Arbeiten eine Taxonomie von vier verschiedenen Integrationsmodellen auf Sie untersuchten dabei, ob eine gemeinsame Repräsentation existieren muß, in der beide Kanäle fusioniert werden, und ob die gemeinsame Repräsentation eher linguistischer Natur ist (was einer späten Fusionsarchitektur entspricht) oder ob es sich um prä-linguistische Repräsentationen handelt und wie diese aussehen könnten.
Beim ersten Integrationsmodell handelte es sich um das Direct Identification Model (DI), wobei die akustischen und visuellen Eingabesignale direkt einem Klassifikator zugeleitet wurde, der das bimodale Signal erkannte (frühe Fusion). Dieses Modell hatte keine intermediäre Repräsentation und entsprach z.B. der dritten von Summerfield vorgeschlagenen Metrik. Ein weiteres Integrationsmodell war das Separate Identification Model (SI), welches für jede Modalität einen einzelnen Erkennungsprozeß vorsah, der als Ausgabe Phoneme oder phonemische Merkmale lieferte, welche in einem nächsten Schritt fusioniert wurden (späte Fusion). Dieses Modell entsprach Summerfields erster Metrik. Bei den letzten beiden Modellen wurde ebenfalls angenommen, daß die Fusion in einer

[8] Eine genauere Beschreibung des Verfahrens und der Experimente kann z. B. auch [Gre96b] entnommen werden.

[9] Es bleibt anzumerken, daß Vroomen in [Vro92], S. 116 ff darauf hinweist, daß dieses Experiment auch zugunsten einer späten Integration gedeutet werden kann.

gemeinsamen Metrik erfolge, allerdings wird die Integration bewerkstelligt, bevor eine Entscheidung über phonetische Merkmale erfolgt (frühe Fusion). Beim Dominant Recoding Model (DR) wurde die auditorische Modalität als die dominante Modalität angenommen, so daß die visuelle Information in eine Repräsentation dieser dominanten Modalität transferiert wurde. Dabei konnte es sich z. B. um die Transferfunktion des Vokaltraktes handeln. Dieses Modell entsprach Summerfields zweiter Metrik. Beim Motor Recoding Model (MR) wurden beide Informationen in eine amodale Metrik projeziert, bevor sie fusioniert wurden. Dabei konnte es sich z.B. um die Konfigurationen des Vokaltraktes oder auch dessen 'Motorprogramme' handeln. Mögliche Metriken für dieses Modell sind Summerfields vierte und fünfte Metrik.
Aufgrund einiger der Ergebnisse der zuvor vorgestellten Untersuchungen schlossen die Autoren das SI- und das DI-Modell als unplausibel aus. Im Vergleich des DR- und MR-Modells tendierten die Autoren zum MR-Modell, da vermutlich die Audition nicht komplett dominant für die Sprachwahrnehmung ist. Als Indiz dafür wurden Ergebnisse von Lisker und Rossi [LR92] angesehen, die zeigen konnten, daß ein visueller Stimulus die Wahrnehmung eines auditorischen Stimulus negativ beeinflussen kann, der weder verrauscht noch widerstreitend zum visuellen ist. Die Versuchspersonen sollten dabei die Rundung eines Vokals bestimmen. Bei einem bestimmten Vokal ergab sich für den auditorischen Stimulus, daß er zu 60 % von den Versuchspersonen als gerundet empfunden wurde, für den visuellen Stimulus waren es nur 1 %. Eine gleichzeitige Präsentation beider Stimuli ergab einen Prozentsatz von 25 %.
Sams et al. [SSHK97],[SMS+98] untersuchten den McGurk-Effekt bei muttersprachlich finnisch sprechenden Versuchspersonen, wobei es sich bei den Stimuli um Silben, isolierte Wörter und Wörter in kurzem Satzzusammenhang handelte. Es wurde erwartet, daß die audiovisuellen Kombinationen als bedeutungsvolle Wörter oder Nichtwörter bzw. zum Kontext passend oder nichtpassend wahrgenommen würden.
Der McGurk-Effekt bei den Versuchspersonen war stark ausgeprägt. Es ergab sich, daß die Stärke des McGurk-Effektes nicht durch Wortbedeutung oder Satzkontext beeinflußt wurde.
Sams et al. schlossen daraus, daß die audiovisuelle Integration zu einem relativ frühen Stadium stattfindet, bevor die Wortbedeutung extrahiert wird.

Experimentelle Ergebnisse: späte Fusion

In einer der ersten Experimente untersuchten Roberts und Summerfield in [RS81] die Fragestellung, ob die selektive Adaptation in der Sprachwahrnehmung auditorischer oder phonetischer Natur ist. Bei der selektiven Adaptation handelt es sich um folgenden Effekt: Einem Hörer werden in zufälliger Reihenfolge Silben präsentiert, die zu einem Kontinuum zwischen zwei phonetischen Kategorien gehören, z.B. liegen sie zwischen [ba] und [pa]. Anhand der Antworten kann nun die Phonemgrenze bestimmt werden, bei der gleichwahrscheinlich die Silben einer der beiden Phonemen zugeordnet werden. Im Adaptationsfall werden dem Hörer vor der Präsentation der eigentlichen Silben mehrere Wiederholungen von Adaptationssilben vorgespielt. Falls die Adaptationssilbe

unzweideutig zu einer der beiden Phonemklassen gehört, z.B. zu [ba], verschiebt sich die Phonemgrenze in Richtung dieser Klasse, d.h. insbesondere in der Region der Phonemgrenze werden mehr Silben als [pa] identifiziert. Um festzustellen ob durch die selektive Adaptation eher der auditorische oder der phonetische Level der Verarbeitung beeinflußt wird, verwendeten Roberts und Summerfield den McGurk-Effekt. Dadurch wird die akustische Struktur des Sprachsignals von der phonetischen Wahrnehmung dissoziiert, da durch den McGurk-Effekt die phonetische Wahrnehmung geändert wird, ohne daß die akustische Struktur der Silbe geändert wird.
Sie präsentierten zu Silben aus einem [bε]-[dε]-Kontinuum als Adaptoren Stimuli, die sich aus akustischem [bε] und visuellem [gε] zusammensetzten, und somit überwiegend als [dε] wahrgenommen wurden. Falls die Adaptation hauptsächlich akustischer Natur ist, sollte sich die Phonemgrenze in Richtung des [bε] verschieben, anderenfalls in Richtung des [dε]. Roberts und Summerfield konnten feststellen, daß sich die Adaptation nur auf die akustischen und nicht auf die phonetischen Beziehungen zwischen Adaptoren und Testsilben gründeten und schlossen daraus, daß selektive Adaptation von der akustischen Struktur des Stimulus abhängt und nicht von der phonetischen Wahrnehmung des Hörers.
Da die Versuchspersonen dieser Untersuchung nicht alle den McGurk-Effekt bei der audio-visuellen Präsentation des Adaptors zeigten, wiederholten Saldana und Rosenblum [SR94] den Versuch mit einer 'zwingenderen' audio-visuellen Kombination als Adaptor (akustisch /ba/, visuell /va/, wahrgenommen als /va/) und konnten damit die Ergebnisse bestätigen und erhärten.
Die Interpretation dieses Versuchs ist umstritten. Vroomen ([Vro92], S. 42) weist darauf hin, daß dieser Versuch nicht im Widerspruch zu einer späten Integration sei. Er zeige zunächst nur, daß die selektive Adaptation auf einem Level stattfinde, der der auditorischen und visuellen Sprachintegration vorausgehe. Es sei allerdings nicht klar, wo dieser Level liege, da auch der Standpunkt vertreten werden könne, daß die Adaptation im phonetischen Modul selbst stattfinde, wo akustische und visuelle Informationen noch nicht integriert seien, was bei einer späten Integration möglich wäre. Robert-Ribes et al. weisen in [RRSE95b] darauf hin, daß die Ergebnisse zwar die Existenz eines modalitätsspezifischen Effekts in der phonetischen Verarbeitung eines auditorischen Inputs zeigten, der eher in Übereinstimmung mit einer späten Integration sei, aber der auch auf einer rein auditorischen Basis verstanden werden könne und verweisen auf die neurale Adaptation des auditorischen Nervs.
Green et al. untersuchten in [GSKM90] und [GKMS91], ob inkompatible visuelle und akustische Informationen über den Sprecher einen Einfluß auf die Größe des McGurk-Effektes haben. Dazu wurden visuelle Aufnahmen männlicher Sprecher mit den akustischen Aufnahmen weiblicher Sprecher gekoppelt und umgekehrt. Obwohl die Inkompabilität von den Versuchspersonen bemerkt wurde, hatte dies doch keine Einfluß auf die Größe des McGurk-Effektes. Die Autoren schlossen daraus, daß Informationen über die Stimmcharakteristik des Sprechers extrahiert und zur Normalisierung des Sprachsignals auf einem frühen Level der phonetischen Verarbeitung benutzt werden, bevor eine Integration beider Signale stattfindet. Da (vgl. [Vro92]) die Sprechernormalisierung aber als ein früher Schritt in der phonetischen Verarbeitung betrachtet wird, könnte

man aus diesen Ergebnisse daraus schließen, daß beim Beginn der phonetischen Verarbeitung der Signale die auditorischen und visuellen Informationen noch nicht kombiniert sind.
Massaro entwickelte mit dem Fuzzy Logical Model of Perception (FLMP) ein Modell der bimodalen Sprachwahrnehmung [Mas87],[Mas96], [MCS95], dessen zentrale Annahmen wie folgt sind: Für jede Informationsquelle wird separat der Grad bestimmt, mit dem sie mögliche Alternativen der Klassifikation unterstützt. Die Informationsquellen werden unabhängig voneinander ausgewertet. Die Quellen werden integriert, um einen Gesamtheitsgrad der Unterstützung für jede Alternative zu liefern. Die Identifikation erfolgt gemäß dem relativen Grad der Unterstützung zwischen den Alternativen.
Das Modell arbeitet in drei Stufen: Merkmalsauswertung von kontinuierlichen Merkmalen, Integration, Vergleich mit Prototypbeschreibungen im Gedächtnis des Modells und Entscheidung über die Identifikation. Um den Ergebnissen von Breeuwer und Plomp [BP86] Rechnung zu tragen, wurde das Modell in [Mas96] dahingehend erweitert, daß das Stimmhaftigkeitsmerkmal eines Stimulus schon im ersten Schritt der Merkmalsauswertung bimodal bestimmt wird (frühe Fusion für Stimmhaftigkeit). Dieses wird aber unabhängig von den akustischen und visuellen Merkmalen als separate Informationsquelle betrachtet. Das Merkmal 'Art der Artikulation' wird dagegen immer noch unabhängig in beiden Modalitäten berechnet. Die Stimmhaftigkeitsinformation wird mit den modalitätsspezifischen Informationsquellen dann wie gewohnt integriert.
Die Anwendbarkeit des FLMP wurde durch Experimente überprüft, in denen Versuchspersonen z.B. ein audio-visuelles /ba/-/da/-Kontinuum klassifizieren sollten. Die Vorhersagbarkeit der Ergebnisse mit dem FLMP wurde untersucht, wobei sich gute Übereinstimmungen ergaben [Mas96], [CM95], auch z.B. für japanische Sprecher [MCG+93].
Das FLMP wird von einigen Autoren als späte Integration bezeichnet ([Vro92], S. 27, [RRSE95b], S. 328), da der erste Schritt der Merkmalsauswertung auf einer Anpassung an phonetische Prototypen basiert. Dabei ist wichtig, daß es **zwei** unabhängige Merkmalsextraktionsmechanismen für den Ort der Artikulation gibt, die auf einem modalitätsspezifischem Signal arbeiten.
Ein Argument gegen das FLMP weist darauf hin, daß für die Erklärung der Kombinationsantworten beim McGurk-Effekt ein /bda/-Prototyp im Modell notwendig sei, die Existenz eines solchen aber plausibel gemacht werden müsse, da z.B. /bda/ keine mögliche Silbe im Englischen ist ([FD91].[Vro92], S. 19). Auch daß das Modell keine modalitätsspezifischen Unterscheidungen treffen kann, wird bemängelt. ([Vro92], S. 17). Zudem sei das FLMP nicht in der Lage, den Zeitverlauf der Sprache hinreichend zu modellieren [FD91].
Massaro untersuchte auch das Categorical Model of Perception (CMP). Bei diesem wird von jedem Kanal nur die kategorische Information, d.h. eine Entscheidung darüber, ob das Signal z.B. ein /ba/ ist oder nicht, geliefert, bevor integriert wird. Dieses Modell wurde als nicht geeignet befunden, experimentelle Resultate zu modellieren [MCG+93].
Vroomen untersuchte die Integrationsfrage anhand mehrerer Experimente [Vro92]. In ersten Versuchen wurde der sogenannte recency effect in serial recall verwendet. Dabei werden einer Versuchsperson eine Liste von Items präsentiert, welche

diese in der präsentierten Reihenfolge wiederholen soll. Normalerweise werden dabei das erste und das letzte Item besser als die mittleren erinnert, was als primacy bzw. recency effect bezeichnet wird. Die Größe des recency effects ist dabei größer, wenn die Items gehört werden, als wenn sie still gelesen werden oder nichtsprachliche Stimuli sind. Man spricht beim recency effect auch von einem sprachspezifischen Effekt. Auch der suffix effect wurde bei den Untersuchungen benutzt: ein auditorischer verbaler Suffix, der ans Ende der präsentierten Liste angehängt wird, beeinflußt den recency effect in der Art, daß dieser schwächer auftritt. Derselbe Suffix am Ende einer still gelesenen Liste hat dagegen nahezu keinen Einfluß auf den suffix effect.
Vroomen untersuchte den recency und den suffix effect mit akustischen, visuellen und akustisch-visuellen Items, gefolgt von Suffixen in einer dieser Modalitäten. Ohne Suffix ergab sich in allen Modalitäten dieselbe Größe des recency-Effekts. Die Präsentation eines Suffixes hatte hingegen nur dann den suffix effect zur Folge, wenn Suffix und präsentierte Liste diegleiche Modalität besaßen, was darauf hindeutet, daß es eine modalitätspezifische Komponente in der Gedächtnisrepräsentation gibt. Dies spricht dagegen, daß recency und suffix effect auf einem gemeinsamen Sprachprozeß basieren und weist auf eine späte Fusion hin. Allerdings ist auch eine andere Erklärung dieser Ergebnisse möglich.
Aufgrund dieser Interpretationsschwierigkeiten untersuchte Vroomen den recency effect noch mit Hilfe unterschiedlicher Konsonanten bzw. Vokale in der präsentierten Liste, da bekannt ist, daß der recency effect in Listen mit variierenden Konsonanten (z.B. /ba,da,ga/) kleiner ist als bei Listen mit variierenden Vokalen (z.B. /bi,ba,bo/). Die Listen wurden bei diesen Untersuchungen nur lippengelesen oder gehört-und-lippengelesen. Es ergab sich, daß der recency effect auch hier bei vokalvariierten Listen größer war als bei konsonantvariierten. Ein Extrabeitrag zum recency effect ergab sich für vokalvariierte und gehört-und-lippengelesene Listen verglichen mit nur lippengelesenen Listen. Diese Ergebnisse konnten ebenfalls nicht eindeutig interpretiert werden.
In einer zweiten Untersuchungsreihe wurde daher versucht festzustellen, ob lippengelesene Sprache Zugang zu einem frühen Stadium der auditorischen Sprachverarbeitung hat, indem der Prozeß der Koartikulation verwendet wurde. Bei der Koartikulation handelt es sich um den Effekt, daß Phoneme beträchtlich in der Zeit überlappen, so daß der Hörer versucht, den phonetischen Kode aus dem koartikulierten Sprachsignal wiederzugewinnen, indem er die koartikulatorischen Einflüsse eines Phonems auf ein anderes berücksichtigt. Das Phänomen kann z.B. beobachtet werden, wenn auf ein /s/- oder /ʃ/-Klang ein /t/-/k/ Kontinuum folgt. Aufgrund der Kinematik des Vokaltraktes 'zieht' das /s/ den Ort der Artikulation des folgenden /k/ weiter nach vorne als ein /ʃ/, so daß die Phonemgrenze im Kontinuum zum /t/-Ende hingezogen wird, wenn zuvor ein /s/ gehört wurde. Vroomen untersuchte diesen Effekt, indem er den frikativen Kontext lippengelesen präsentierte. Dabei koppelte er einen zweideutigen, zwischen /s/ und /ʃ/ liegenden akustischen Reiz mit einem visuellen Reiz, der entweder /s/ oder /ʃ/ entsprach, so daß der Hörer entweder /s/ oder /ʃ/ wahrnahm. Diesem Reiz folgte ein Laut aus einem /ta/-/ka/-Kontinuum. Es ergab sich keine Verschiebung der Phonemgrenze. Damit fand Vroomen einen modalitätsspezifischen auditorischen Effekt, der im phonetischen Modul lokali-

siert ist und schloß daraus, daß lippengelesene Information keinen Zugang zum Bestimmungsprozeß der auditorischen Ortsmerkmale hat, was für ein spätes Fusionsmodell spricht.

4.8 Diskussion

Die Frage nach der Integrationsstufe, auf der akustischer und visueller Kanal kombiniert werden, kann noch nicht abschließend beantwortet werden. Nur in sehr begrenzten Teilgebieten konnten eindeutige Ergebnisse erzielt werden. So herrscht hinsichtlich der Frage, ob zur Bestimmung der Stimmhaftigkeit akustischer und visueller Kanal auf früher oder später Ebene integriert werden, kaum noch Uneinigkeit; die Ansicht, daß dies auf früher Stufe passiert, ist allgemein akzeptiert, auch von Vertretern eines ansonsten 'späten' Modells (vgl. dazu [Mas96], [Vro92], S. 26-27) [10]. Bei dem phonetischen Merkmal des Artikulationsortes dagegen ist dies nicht der Fall. Andere phonetische Merkmale wie Nasalität wurden in die Untersuchungen noch nicht einbezogen. Nichtsdestotrotz scheint sich die Ansicht, daß die Integration auf einer frühen Ebene geschieht, eher durchzusetzen (vgl. die Arbeiten von Green et al., Robert-Ribes et al, Burnham et al.). Die Frage der Metrik kann ebenfalls noch nicht beantwortet werden, doch geht die Tendenz zu einer artikulatorischen Metrik (vgl. [Sum87], [Gre98], [SRRE98]).

[10] Vroomen weist allerdings darauf hin, daß lippengelesene Sprache an sich keine Informationen über die Stimmhaftigkeit beinhaltet, sondern hauptsächlich über den Ort der Artikulation, so daß dieses Ergebnis an sich noch keinen allgemeingültigen Beweis für die Natur des Integrationsprozesses liefert.

Kapitel 5

Audio-visuelle Sprachverarbeitung in technischen Systemen

Visuelle Informationen werden in verschiedenster Art und Weise auf dem Gebiet der Sprachverarbeitung verwendet. Neben der Kombination des visuellen Kanals mit den akustischen Sprachsignalen, die in diesem Kapitel behandelt werden soll, werden auch Systeme entwickelt, die ausschließlich mit den visuellen Informationen arbeiten ([BMS99], [MM99], [GGP94], ([RM94]), [GMS96], [MP91], [MP89], [ITF01], [FM88], [KFMT89], [Nis86]). Weiterhin wird die visuelle Information auch auf dem Gebiet der Sprecheridentifikation und -verifikation hinzugezogen ([MNS99], [SNM99], [MCC+98], [PLH00], [JLGW97a], [JLGW97b], [Mas97], [LTB96a]).
Bei der Realisierung eines audio-visuellen Spracherkennungssystems sind mehrere kritische Punkte zu beachten: Zunächst ist die Problematik der Synchronisation beider Kanäle zu nennen, auch das Problem, daß beide Kanäle mit zumeist unterschiedlichen Aufnahmeraten aufgezeichnet werden, gehört in diesen Themenkreis. Daneben ist auch zu erwähnen, daß die Anzahl der verwendeten Merkmale für jeden Kanal unterschiedlich groß sein kann. So ist oft der visuelle Kanal als Graustufenbild der Lippenregion oder des Gesichts gegeben, welches ohne ein Merkmalsreduktionsverfahren eine wesentlich größere Dimensionalität als der akustische Kanal aufweist. Die Extraktion von viusellen Merkmalen ist daher ein viel diskutiertes Thema in der audio-visuellen Sprachverarbeitung. Dabei bezieht sich die Merkmalsextraktion meistens ausschließlich auf die Lippenregion, wobei zum Beispiel die Lippenkonturen modelliert werden (vgl. [BO94],[LTB96b],[LTB96c],[LD98],[Vog95],[BOSK95],[DKB95]) oder lippenspezifische Parameter bestimmt werden (vgl. [NPZ99],[AB95b]). Eine weitere Möglichkeit zur Datenreduktion besteht in der Anwendung einer LDA oder PCA auf die Graustufenbilder. Als weitere wichtige und bislang noch nicht endgültig geklärte Problematik kann die Integration beider Kanäle genannt werden, d.h. wie und auf welcher Stufe der Informationsverarbeitung erfolgt die Fusion beider Informationskanäle. In diesen Themenkreis gehört auch die Frage nach einer Gewichtung beider Kanäle.

In diesem Kapitel sollen die wichtigsten technischen Systeme beschrieben werden, die visuelle Informationen zur Verbesserung der Spracherkennungsleistung einsetzen. Ein Vergleich verschiedener Systeme gestaltet sich dabei schwierig, da die jeweiligen Ergebnisse abhängig von der Komplexität der Aufgabe (z.B. kontinuierliche Worterkennung oder Zahlenerkennung) sind und auch davon, welches akustische System und welche visuellen Merkmale benutzt wurden. In zwei Tabellen werden die wichtigsten Merkmale der vorgestellten Systeme aufgelistet, wobei Tabelle 5.1 nur Systeme vorstellt, die eine einzige Fusionsarchitektur implementiert haben, wohingegen sich Tabelle 5.2 auf die Systeme stützt, die einen Vergleich mehrerer Fusionsstrategien durchführen.
Nur ein Teil der Systeme erlaubt zudem die Untersuchung verschiedener Fusionsstrategien. Auf diese Arbeiten soll im kommenden der Schwerpunkt gelegt werden.
Als 'frühe Fusion' soll dabei eine Architektur bezeichnet werden, bei der die Merkmalsvektoren beider Kanäle integriert werden, bevor sie gemeinsam als Eingabe in einen Klassifikator dienen. Im Gegensatz dazu wird bei einer späten Fusionsarchitektur zunächst jeder Kanal separat klassifiziert bzw. kategorisiert, anschließend findet eine Integration beider Kanäle statt. Im dritten Fall, der ein Zwischenglied zwischen diesen beiden Extremen bildet und der als 'Fusion auf einer Zwischenstufe' bezeichnet werden soll, findet die Integration der beiden Informationskanäle innerhalb des Klassifikators statt.
Bei der frühen Fusionsmöglichkeit ist eine schlechte Synchronisation der Kanäle kritisch, da dadurch unter Umständen nicht zusammen gehörende akustische und visuelle Merkmale zu einer Eingabe kombiniert werden. Gleichzeitig kann aber das Erkennungssystem von Abhängigkeiten zwischen beiden Kanälen Nutzen ziehen. Allerdings ist der Eingabevektor in das System oft sehr hochdimensional und der Ausfall eines Kanals bei der Erkennung ist kritisch.
Bei der späten Fusion können Abhängigkeiten der Kanäle nicht berücksichtigt werden, da beide Kanäle separat klassifiziert werden. Damit ist diese Fusionsmöglichkeit unempfindlich gegenüber nicht synchronen Signalen. Ein spätes Fusionsmodell ermöglicht die Wahl unterschiedlicher Klassifikatoren für beide Kanäle. Zudem kann der Ausfall eines Kanals im Regelfall leicht kompensiert werden.
Das Fusionsmodell auf der Zwischenstufe kann im besten Fall die Vorteile der frühen und späten Fusion in sich vereinen: Es können Abhängigkeiten zwischen beiden Kanälen berücksichtigt werden, ein Ausfall eines Kanals kann kompensiert werden.

5.1 Späte Fusionsarchitekturen

Eines der ersten Systeme [1], in denen die Verwendbarkeit zusätzlicher visueller Informationen in Spracherkennungssystemen getestet wurde, war ein sprecherabhängiges System, welches isolierte Äußerungen erkennen konnte. Dieses System wurde von Petajan 1984 entwickelt ([Pet84]). Als visuelle Informationen

[1] Das vermutlich allererste System für automatisches Lippenlesen wurde in den sechziger Jahren von Nassimbene bei IBM entwickelt, vgl. [Pet84].

wurden dabei binäre Bilder der Lippenregion verwendet aus denen 4 visuelle Merkmale extrahiert wurden. Für die Erkennung der akustischen Daten diente ein kommerzielles Spracherkennungssystem. Die Erkennung des visuellen Kanals erfolgte über eine Distanzberechnung. Eine späte Fusion beider Kanäle führte zu einer deutlichen Steigerung der Erkennungsleistung um bis zu 13%, wobei der visuelle Kanal im wesentlichen zur Korrektur des akustischen verwendet wurde, der eine Vorauswahl möglicher Klassifikationen lieferte. Das Vokabular bestand aus Zahlen und Buchstabenwörtern.
Bregler et al. [BMHW93], [BHMW93] untersuchten unter Zuhilfenahme eines modularen MS-TDNN, welches durch einen Backpropagation-Algorithmus trainiert wurde, eine späte Fusionsarchitektur. Die Aufgabe bestand in der Erkennung von Sequenzen von Buchstabenwörtern. Die akustischen Daten wurden zunächst mit 16 kHz aufgenommen und alle 10 ms 16 Melscale-Fourierkoeffizienten berechnet, indem ein Hammingfenster benutzt wurde. Die visuellen Daten wurden alle 33 ms aufgenommen und bestanden aus Graustufenbildern der Lippenregion, die manuell zentriert wurden. Das Graustufenbild wurde tiefpaßgefiltert und einer histogrammnormalisierten Grauwertkodierung unterzogen. Jeder Kanal wurde separat mit Hilfe eines TDNNs gelernt. Auf der Phonem- (65 Phoneme) bzw. Visemebene (42 Viseme) wurde die Fusion durchgeführt, indem die gewichtete Summe beider Kanäle gebildet wurde, wobei die relative Entropie zwischen beiden Kanälen zur Gewichtung genutzt wurde. Anschließend erfolgte ein DTW (Dynamic Time Warping), die Ausgabeschicht des Gesamtnetzes lieferte dann eine Wortklassifikation. Das System wurde mit den Daten von zwei Sprechern trainiert und getestet. Dabei zeigte sich in jedem Fall eine deutliche Steigerung der Erkennungsleistung für das fusionierte System, insbesondere, wenn die akustischen Daten verrauscht wurden.
Silsbee und Bovic [SB93], [SB96] entwickelten ein sprecherabhängiges System zur Erkennung von 14 Vokalen in unterschiedlichen Wortkontexten. Für das akustische System wurde zunächst eine perzeptuell motivierte Merkmalsextraktion angewendet, gefolgt von einer Vektorquantisierung. Die entstehende Sequenz von diskreten Symbolen wurde als Eingabe in ein HMM verwendet, dessen Ausgabe als Bewertung interpretiert wurde, daß ein bestimmtes Wort beziehungsweise ein bestimmter Vokal gesprochen worden war. Ähnlich wurde für den visuellen Kanal zunächst ein modifizierter Vektorquantisierungsalgorithmus auf Bilder der Lippenregion angewendet. Die daraus resultierende Sequenz an diskreten Symbolen diente wiederum als Eingabe in ein weiteres HMM. Auch beim visuellen Kanal wurde so eine Wortliste mit entsprechenden Bewertungen erzeugt. Die Integration erfolgte durch eine gewichtete Summierung der Wertungen jedes Kanals für jeweils ein Wort. Das Wort mit der besten Gesamtwertung wurde als Klassifikation gewählt. Es wurden Versuche mit verrauschtem akustischen Kanal durchgeführt, wobei das System auch mit verrauschten Daten trainiert wurde. Es zeigte sich, daß der Gewichtsparameter für die Integration beider Kanäle nahezu unabhängig von dem Signal-Rausch-Verhältnis gewählt werden konnte. Wiederum zeigte sich insbesondere für die verrauschten Daten ein Verbesserung der Erkennungsleistung, wenn visuelle Informationen für die Klassifikation hinzugezogen wurden.
Cox et al [CMB97], [MCC+98] untersuchten eine späte Fusionsarchitektur, bei

der die Ausgaben der Einzelklassifikatoren, die aus Hidden-Markov-Modellen bestanden, gewichtet addiert wurden. Das Datenmaterial bestand aus Buchstabenwörtern, die von zehn Sprechern jeweils dreimal geäußert wurden. Die visuellen Merkmale bestanden aus einem 80 x 60 Pixelbild der Lippenregion, aus dem ein 20-dimensionaler Merkmalsvektor extrahiert wurde. Die akustischen Daten wurden mit einer Filterbank mit 24 Kanälen weiterverarbeitet. Durch Addition mit Gaußschem Rauschen wurden gestörte akustische Daten mit verschiedenem SNR erzeugt (SNR=20 dB bis 0 dB). Der Wichtungsfaktor für die Integration wurde sowohl empirisch bestimmt und hinsichtlich des SNR optimiert als auch mit Hilfe eines entropiebasierten Konfidenzmaßes automatisch bestimmt. Die audio-visuelle Erkennungsrate des Systems lag deutlich über derjenigen des akustischen Erkenners.
Verma et al. [VFNB99] untersuchten eine späte Fusionsarchitektur anhand von sprecherunabhängigen Daten. Die Daten bestanden aus 700 Sätzen, gesprochen von 6 Sprechern. Aus dem akustischen Signal wurden 24 Mel-Cepstral-Koeffizienten extrahiert, die anschließend einer LDA unterzogen wurden. Zur Gewinnung der visuellen Merkmale wurde ein 45 x 30 Pixelbild der Mundregion zunächste einer PCA unterzogen. Die resultierenden 100 Koeffizienten wurden mit Hilfe einer LDA noch einmal dimensionsreduziert zu 35 Merkmalen. Der Erkenner bestand aus zwei getrennten Hidden-Markov-Modellen. Es wurden drei Methoden der Klassifikation untersucht. Bei der ersten Methode wurden beide Kanäle auf Phonembasis klassifiziert und die Ergebnisse gewichtet addiert, um eine endgültige Klassifikation zu erreichen. Bei der zweiten Methode wurde die Phonembasis für den visuellen Kanal durch eine Visembasis ersetzt. Bei der dritten Methode wurden beiden Kanäle auf Visembasis und anschließend auf den noch in Frage kommenden Phonemen klassifiziert. Alle drei Methoden zeigten eine verbesserte Klassifikationsleistung für alle untersuchten Rauschstufen gegenüber beiden Einzelkanälen, die dritte Methode erwies sich dabei als die beste.
Die späte Fusionsarchitektur von Gurbuz et al. [GTPG01] bestand aus zwei völlig getrennten Erkennern für den akustischen und den visuellen Kanal. Die Daten waren sprecherabhängig und bestanden aus isolierten Wörtern. Die vom Audio-System benutzten Merkmale beruhten auf diskreten Mel-Frequency-Koeffizienten. Die Basis für die visuellen Merkmale waren Daten über die äußeren Lippenkonturen des Sprechers und weitere Merkmale wie die normalisierte äußere Lippenweite oder Höhe. Weiter wurden auch die zeitlichen Ableitungen dieser Merkmale verwendet. Zur Klassifikation der Einzelkanäle dienten Hidden-Markov-Modelle. Deren log-likelihood-Werte wurden gewichtet addiert, wobei die Gewichte je nach SNR der akustischen Daten gewählt wurde. Die Klassifikation erfolgte gemäß dem maximalen kombinierten log-likelihood-Wert. Der audio-visuelle Erkenner zeigte in jedem Fall eine bessere Leistung als jeder der beiden Einzelkanäle.
Heckmann et al. [HBK01] untersuchten anhand von kontinuierlich gesprochenen Ziffern ein sprecherabhängiges Erkennungssystem. Die Audiomerkmale bestanden aus 12 Cepstral-Koeffizienten und der log-Energie. Die visuellen Merkmale bestanden aus 6 Parametern der Lippen, wie äußere Lippenweite und -höhe und innere Lippenweite und -höhe. Die Fusionsarchitektur benutzte neurona-

le Netzwerke, um für jeden Informationskanal separat die a posteriori Wahrscheinlichkeiten auf Phonemebene zu bestimmen. Mehrere Zeitscheiben wurden den neuronalen Netzen gleichzeitig präsentiert, um zeitliche Informationen zu berücksichtigen. Die Ausgaben der Netze wurden direkt als Eingaben in ein HMM verwendet, welches jeweils ein einzelnes Phonem modellierte. Dabei wurden beide Ausgaben je nach SNR des akustischen Kanals mit verschiedenen Methoden gewichtet. Wortmodelle wurden durch Aneinanderreihung der entsprechenden Phonemmodelle generiert. Für alle betrachteten Wichtungsschemata konnte eine Verbesserung der Erkennungsleistung der Fusionsarchitektur gegenüber beiden Einzelkanälen für alle untersuchten Rauschstufen festgestellt werden.

5.2 Fusion auf einer Zwischenstufe

In den Arbeiten von Yuhas et al. [YMHGSJ90], [SYMHGJ90], [YMHGS89] wird die Fusion beider Kanäle vor einer Kategorisierung durchgeführt, um einen möglichen Informationsverlust durch eine frühe Kategorisierung zu vermeiden. Das sprecherabhängige System erkennt stationäre Vokale. Als akustisches Merkmal wird ein geglättetes Kurzzeit-Energiespektrum benutzt, aus welchem 32 akustische Merkmale extrahiert werden. Als visuelle Merkmale werden statische Graustufenbilder der Lippen verwendet. Für die Integration wurde das visuelle Signal auf die akustische Repräsentation, die eng mit der Vokaltrakt-Transfer-Funktion verbunden ist, abgebildet. Das visuelle Signal wurde also dafür benutzt, eine Abschätzung des zugehörigen akustischen Spektrums zu erhalten. Diese Abbildung erfolgte mit Hilfe eines Feedforward-Netzes, welches mittels des Backpropagationalgorithmus daraufhin trainiert wurde, aus den Lippenbildern die zugehörigen akustischen Merkmale abzuschätzen. Die Integration beider Kanäle erfolgte durch eine Gewichtung und Addition der aus dem akustischen und visuellen Kanal ermittelten Merkmale, wobei der Wichtungsparameter vom Signal-zu-Rausch-Verhältnis abhängig ist. Die lineare Abhängigkeitsfunktion wurde dabei empirisch bestimmt. Die abschließende Vokalklassifikation erfolgte wiederum durch ein Feedforward-Netz. Um das System zu testen, wurde der akustische Kanal künstlich verrauscht, wobei die verrauschten Daten vom System gelernt wurden. Es zeigte sich in allen Fällen eine Verbesserung der Erkennungsleistung, wenn das visuelle Signal zur Erkennung hinzugezogen wurde. Dabei war die Größe der Verbesserung vom Signal-zu-Rauschverhältnis abhängig und lag bei bis zu 40 %.
Girin et al. [GFS95], [GFS97] benutzten das visuelle Signal, um ein Modell des Audiosignals daraus abzuleiten. Dieses Modell diente anschließend zur Filterung des gestörten akustischen Signals. Zunächst wurden mit Hilfe einer LPC-Analyse aus den akustischen Daten spektrale Parameter ermittelt. Die Abbildung der visuellen Merkmale (Weite innerhalb der Lippen, Höhe und Lippenregion) auf die spektralen Parameter erfolgte mit Hilfe einer linearen Regressionsmethode. Die spektralen Parameter des gestörten akustischen Signals wurden schließlich mit den aus den visuellen Merkmalen ermittelten Parametern kombiniert und aus den neuen Parametern ein 'gereinigtes' akustisches Signal

ermittelt. Das System wurde mit einer Datenbasis aus Vokalen getestet und zeigte eine deutliche Verbesserung der Klassifikationsergebnisse.
Niyogi et al. [NPZ99] schlugen die distinktiven Merkmale (vgl. Seite 4) als die Zwischenstufe vor, auf der eine Integration beider Kanäle stattfinden könnte. Distinktive Merkmale besitzen eine 'natürliche' artikulatorische Interpretation, z.B. sind bei den labialen Lauten die Lippen geschlossen, jedes Merkmal besitzt eine akustische und visuelle Entsprechung. Die Autoren beschrieben mehrere Detektoren, mit deren Hilfe die distinktiven Merkmale sowohl aus den akustischen als auch den visuellen Informationen extrahiert werden könnten. Eine Realisierung eines audio-visuellen Erkenners wurde allerdings nicht vorgestellt.
Cosi und Caldognetto [CC96] testeten ein audio-visuelles Spracherkennungssystem auf der Basis eines rekurrenten neuronalen Netzwerkes, welches mit Hilfe eines backpropagationähnlichen Verfahrens trainiert wurde. Die Fusion beider Kanäle fand dabei auf der Ebene der versteckten Neuronen des Netzes und damit auf einer Zwischenstufe der Informationsverarbeitung statt. Sie verwendeten bei ihren Untersuchungen ein Vokabular aus Nonsense-Wörtern der Form VCV wobei C=/p,t,k,b,d,g/ und V=/a,i,u/ waren. Die akustischen Merkmale bestanden aus 40 spektralähnliche Parametern. Die 14 visuellen Merkmale wurden mit Hilfe von auf dem Gesicht und den Lippen platzierten Markern aufgenommen, die der Extraktion einer Reihe von Parametern wie z.B. die vertikale Bewegung der Lippen oder des Kinns dienten. Sprecherabhängige und -unabhängige Versuche zeigten eine deutliche Verbesserung der Erkennungsleistung durch Hinzunahme der artikulatorischen Merkmale sowohl in rauschfreier als auch in gestörter Umgebung.
Luettin und Dupont [LD98] untersuchten Modelle mit Integration auf einer Zwischenstufe anhand von französischen Zahlenwörtern, die von mehreren Sprechern gesprochen wurden. Die akustischen Daten wurden mit 8 kHz abgetastet, anschließend wurden 25 Merkmale extrahiert. Gestörte akustische Daten wurden durch Addition von weißem Rauschen erzeugt. Die visuellen Daten bestanden aus Graustufenbildern. Aus diesen wurden 24 Parameter, die die Lippenkonturen sowie Intensitätswerte der Lippenregion beschrieben, sowie deren Ableitungen als Merkmale extrahiert. Die Klassifikation der Einzelkanäle erfolgte durch Hidden-Markov-Modelle. Die Fusionsarchitekturen basierten auf den Hidden-Markov-Modellen der Einzelkanäle und bestanden aus Multi-Stream-Modellen. In diesen wurden die Übergangswahrscheinlichkeiten der einzelnen Hidden-Markov-Modelle kombiniert. Das erste Modell gestattete eine asynchrone Verarbeitung beider Kanäle, wobei die Rekombination beider Informationskanäle auf Wortebene erfolgte. In einem weiteren Modell wurde die Rekombination auf Zustandsebene durchgeführt, wobei keinerlei Asynchronität berücksichtigt werden konnte. Die Gewichtungsparameter beider Kanäle wurden empirisch bestimmt, indem Trainingsdaten mit demselben Störungsgrad wie diejenige der Testdaten verwendet wurden. Das audio-visuelle System zeigte dabei bei allen Rauschstufen eine bessere Performanz als die beiden Einzelerkenner. Die Rekombination auf Wortebene war dabei am besten.
Zhang et al. [ZLH00] benutzten in ihrem Modell intermodale Korrelationen, um eine bessere Entscheidung treffen zu können. Dazu wurde ein Fusionsmodell auf der Zwischenstufe realisiert, indem mit akustischen und visuellen

Daten jeweils ein HMM getrennt trainiert wurde, und anschließend aus den schon trainierten Modellen ein gemeinsames akustisch-visuelles HMM gebildet wurde, welches dem 'state-synchronous' Modell von Luettin et al. (vgl. S. 87) entsprach. Die Aufgabe des Systems bestand in der sprecherunabhängigen Erkennung von Zahlenreihen, wobei Daten von 100 Sprechern mit jeweils 110 Zahlen zur Verfügung standen. Die visuellen Daten wurden durch einen Tracking-Algorithmus des Mundumrisses ermittelt, welcher 8 visuelle Merkmale sowie deren erste und zweite zeitliche Ableitung lieferte. Als akustische Daten wurden 13 Mel-Frequency-Cepstral-Koeffizienten und deren erste und zweite zeitliche Ableitung verwendet. Das System wurde mit unverrauschten Daten trainiert, getestet wurde es auch mit verrauschten akustischen Daten, wobei unterschiedliche Noisearten verwendet wurden. Eine Gewichtung der Einzelkanäle wurde festgelegt, wobei der akustische Kanal mit einem Faktor von 0.7 wesentlich stärker als der visuelle (Faktor 0.3) gewichtet wurde. Beim Testen fand ein leaving-one-out-Crossvalidierungsverfahren Anwendung. Das Fusionssystem zeigte bei Experimenten mit verrauschten akustischen Daten bei 0 dB SNR eine bessere Leistung als das akustische System alleine.

5.3 Frühe Fusionsarchitekturen

Wu et al. [WTM+91] untersuchten die Erkennungsleistung bei der Vokalerkennung von fünf japanischen Vokalen, indem verschiedene visuelle Merkmale gemeinsam mit den akustischen Informationen verwendet wurden. Die Erkennung war sprecherunabhängig. Als visuelle Merkmale wurden Graustufenbilder der Lippen, daraus abgeleitete binäre Lippenbilder einmal inklusive der Zähne, einmal ohne Zähne und geometrische Merkmale (horizontale Länge, vertikale Breite, Größe der Lippenregion) verwendet. Zur Klassifikation der Daten diente ein dreischichtiges Feedforwardnetzwerk, das mit Hilfe des Backpropagation-Algorithmus trainiert wurde. Beide Informationskanäle wurden vor der Eingabe in das Netzwerk fusioniert. Die Spracherkennungsleistung ließ sich mit Hilfe der visuellen Informationen um bis zu 12% steigern, wobei die binären Lippenbilder das beste Ergebnis erzielten, gefolgt von den Graustufenbildern. Bei der Verwendung der geometrischen Merkmale ergaben sich noch schlechtere Klassifikationsleistungen als bei Verwendung der akustischen Daten alleine.
Bregler et al. [BOK94],[BK94], [BO95] untersuchten ein audio-visuelles Spracherkennungssystem, das die Aufgabe hatte, kontinuierlich gesprochene Buchstabenwörter zu klassifizieren. Das hochdimensionale Graustufen-Lippenbild wurde dabei mit Hilfe des sogenannten 'Surface-learning' auf nur wenige visuelle Merkmale abgebildet. Als visuelle Merkmale wurden entweder 10 dieser sogenannten Eigenlip-Koordinaten verwendet oder die 10 Eigenlip-Koordinaten zuzüglich ihrer 10 Deltamerkmale, da die zeitlichen Änderungen der Lippenkonturen relevante Informationen beinhalten. Die Fusion beider Kanäle fand statt, bevor der gemeinsame Merkmalsvektor als Eingabevektor in ein MLP verwendet wurde. Dabei wurden 9 akustische und 10 visuelle Merkmale konkateniert, die fehlenden visuellen Daten wurden interpoliert. Das verwendete Zeitfenster wurde mit 190 ms Breite sehr groß gewählt, da festgestellt werden

konnte, daß schon vor Einsetzen des akustischen Signals visuelle Beiträge vorhanden waren. Die größte Korrelation zwischen Daten beider Kanäle konnte bei einer Zeitverschiebung von 120 ms festgestellt werden, wobei die visuelle Information der akustischen vorausging. Als Ausgabe des MLP erhielt man die a posteriori-Wahrscheinlichkeit auf Phon-Ebene, die Worterkennung erfolgte durch ein HMM. Es wurden Versuche mit und ohne verrauschten akustischen Daten durchgeführt. Dabei wurde festgestellt, daß bei ungestörten akustischen Daten keine Verbesserung der Erkennungsleistung durch die zusätzlichen visuellen Informationen erzielt werden konnte, wohl aber bei verrauschten akustischen Daten.

Dalton et al. [DKB95] benutzten einen sprecherabhängigen Datensatz, der 20 Wiederholungen von jeweils 40 isolierten Wörtern umfaßte. Die visuellen Merkmale bestanden aus 6 Parametern, die aus einer Modellierung der Lippenkonturen abgeleitet wurden. Als akustische Merkmale wurden 8 Melscale-Filterbankkoeffizienten gewählt, die alle 20 ms berechnet wurden. Die konkatenierten Merkmale wurden mit Hilfe eines auf DTW basierenden Erkenners klassifiziert und lieferten bessere Erkennungsleistungen als die jeweiligen Einzelkanäle über mehrere SNR-Stufen hinweg.

Potamianos et al. [PCGR97] untersuchten eine frühe Fusionsarchitektur auf der Basis von Hidden-Markov-Modellen. Die Erkennung war sprecherabhängig, und das Datenmaterial bestand aus 1500 kontinuierlich gesprochenen Zahlwörtern. Die akustischen Daten bestanden aus 39 Mel-Frequency-Cepstral-Koeffizienten, die visuellen aus 12 dimensionalen Merkmalsvektoren mit geometrischen Merkmalen. Die Merkmalsvektoren beider Kanäle wurden konkateniert und von einem einzelnen HMM klassifiziert. Das audio-visuelle System zeigte dabei eine bessere Performanz als jeder der beiden Einzelkanäle.

Autoren	Architek-turen (F,Z,S)	Methode	Trainings-methode	Anz. akust. Merkmale	Anz. vis. Merkmale	Sprecher -abhängig	Corpus	Gewichtung der Kanäle	Gewichtungs-methode
Petajan '84	S	kommerz. System (ak) Bewertung durch Distanzmessung (vis)	TT, keine	k. A.	4	ja	Ziffern Buchstaben-worte	nein	-
Bregler '93	S	NN	TT k.A.	16	384	ja	Buchstaben-worte	ja	autom.
Silsbee '93, '96	S	HMM	TT st	3	1	ja	Vokale im Wortkontext	ja	SNR(M)
Cox '97, '98	S	HMM	TT oh	24	20	nein	Buchstaben-worte	ja	SNR,SNR(M), autom.
Verma '99	S	HMM	k.A. oh	24	35	nein	Sätze	ja	SNR(M)
Gurbuz '01	S	HMM	TT oh	k. A.	k. A.	ja	isolierte Worte	ja	SNR(M)
Heckmann '01	S	NN und HMM	TT st	13	6	ja	kontinuierl. Ziffern	ja	SNR(M)
Yuhas '89, '90	Z	NN	TT oh	32	500	ja	Vokale	ja	SNR(M)
Girin '95, '97	Z	Gaußklass.	TT st	50	3	ja	Vokale	nein	-
Cosi '96	Z	NN	TT oh	40	14	ja,nein	VCV	nein	-
Luettin '98	Z	HMM	CV oh	25	48	nein	Ziffern	ja	SNR(M)
Zhang '00	Z	HMM	CV oh	39	24	nein	Ziffer-reihen	ja	SNR(M)
Wu '91	F	NN	TT keine	64	50, 3	nein	Vokale	nein	-
Bregler '94, '95	F	NN und HMM	TTV oh	9	10	nein	Buchstaben-worte	nein	-
Dalton '95	F	DTW	TT oh	8	6	ja	Worte	ja	SNR(M)
Potamianos '99	F	HMM	TT oh	39	12, 45	ja	kontinuierl. Ziffern	nein	-

Tabelle 5.1: Fusionsarchitekturen in der Literatur, keine Vergleiche, F=frühe Fusion, Z= Fusion auf Zwischenstufe, S= späte Fusion, k.A.= keine Angaben, Trainingsmethode: TT=Datensatz wird in Trainings- und Testdaten aufgespalten, TTV= Datensatz wird in Trainings-, Test- und Validierungssatz aufgespalten, CV=Datensatz wird in Cross-Validierungssätze aufgespalten, keine=keine gestörten Daten verwendet, st=Training mit gestörten Daten, oh=Training nur mit ungestörten Daten; Gewichtungsmethode: autom.=automatische Gewichtung ohne Bestimmung des SNR, SNR(M)= empirische Ermittlung des für den jeweiligen SNR geeigneten Parameterwert des Gewichts, SNR=automatische Gewichtung unter Berücksichtigung des SNR, VCV= Vokal-Konsonant-Vokal-Datensatz

5.4 Vergleiche von unterschiedlichen Fusionsarchitekturen

Stork et al. [SWL92] untersuchten mit Hilfe einer neuronalen Architektur unterschiedliche Fusionsmöglichkeiten. Dazu verwendeten sie ein bzw. zwei Time-Delay Netze (TDNN), in das sowohl die akustischen als auch die visuellen Daten eingespeist wurden. Das System war sprecherunabhängig und wurde mit verrauschten und ungestörten akustischen Daten verwendet. Zur Erzeugung der visuellen Daten wurden 10 Reflexionspunkte auf dem Gesicht des Sprechers befestigt. Aus den 10 Punkten wurden 5 Merkmale berechnet, z.B. die Höhe der Mundöffnung. Die Merkmale wurden anschließend normalisiert. Die akustischen Daten bestanden aus den Ausgaben von 14 Melscale-Filtern. Alle Äußerungen waren 1 Sekunde lang und bestanden aus 10 gesprochenen Buchstaben, die aufgrund ihrer großen Verwechselbarkeit gewählt wurden. Basierend auf Time-Delay-Netzen wurden verschiedene Architekturen realisiert: Ein TDNN, welches nur die visuellen Daten verarbeitete, benutzte ein Eingabefenster von 200 ms Länge und enthielt 800 zu trainierende Gewichte. Das akustische TDNN hatte 1280 Gewichte.
Ein Modell für eine späte Fusion wurde aufgrund der schon trainierten Einzelnetze für die akustischen und visuellen Daten realisiert, indem die Ausgabewahrscheinlichkeiten der Einzelnetze miteinander und mit einer Normalisierungskonstante multipliziert wurden, wobei die einzelnen Kanäle nicht gewichtet wurden. Das Gesamtnetz besaß demnach 1280 Gewichte, die allerdings getrennt trainiert wurden. Ein weiteres Fusionsnetzwerk realisierte eine Fusion auf der Zwischenstufe, indem in einem TDNN die Fusion auf der Ebene der versteckten Neuronen durchgeführt wurde, so daß diese Neuronen sowohl akustische als auch visuelle Eingaben erhielten. Das initiale Netz hatte 3280 Gewichte und wurde dabei aus den schon trainierten Einzelnetzen zusammengefügt und anschließend komplett nachtrainiert.
Das visuelle Einzelnetz erreichte eine Erkennungsrate von 51 %, das akustische von 64 %. Das späte Fusionsmodell klassifizierte 91 % der Daten korrekt, während das Modell mit Fusion auf einer Zwischenstufe nur 87 % erreichte. Eine Erklärung dafür kann aufgrund der Anzahl der Gewichte gegeben werden, da diese bei dem Zwischenstufen- Modell wesentlich größer als beim späten Fusionsmodell war.
Die Verwendung verrauschter akustischer Daten führte zu einem Klassifikationsergebnis von 43 % für die akustischen Daten alleine und von 75 % für die fusionierten Daten bei Verwendung der späten Fusion.
Robert-Ribes et al. untersuchten in [RRLES93], [RRSE95a], [RRPSE96] und [RRSE95b] insgesamt vier verschiedene Fusionsarchitekturen mit demselben Datensatz. Dabei handelte es sich um jeweils ein Modell für eine frühe Fusion (genannt DI-Modell, direct identification) und ein spätes Fusionsmodell (SI-Modell, separate identification) sowie um zwei verschiedene Realisierungen von Fusionsarchitekturen auf der Zwischenstufe der Informationsverarbeitung. Dabei fand die Integration jeweils in einem gemeinsamen Merkmalsraum statt. Im ersten Fall war dieser auditiver Natur, da man annimmt, daß der auditive

Kanal der bei der Erkennung dominante Kanal sei (DR-Modell, 'dominant recoding'). Im zweiten Fall war der gemeinsame Merkmalsraum artikulatorischer Natur, d.h. man führte die Integration mit Repräsentationen der Vokaltraktkonfigurationen durch (MR-Modell, motor recoding).
Der Datensatz bestand aus 10 französischen Vokalen, die von einem Einzelsprecher jeweils 34 mal ausgesprochen wurden, so daß jeder Stimulus in etwa gleiche Länge hatte. Das Signal wurde anschließend an 3 festgelegten zeitlichen Positionen analysiert. Die akustischen Daten wurden mit 16 kHz abgetastet. Als Merkmale wurden 20 jeweils 1-Bark-breite Kanäle zwischen 0 und 5 kHz benutzt. Für die visuellen Merkmale wurden aus Aufnahmen des Gesichts folgende Lippenparameter extrahiert: horizontale Weite, vertikale Höhe, Lippenregion. Gestörte akustischen Daten mit unterschiedlichem SNR wurden durch Hinzufügung von Gaußschem Rauschen vor der weiteren Spektralanalyse erzeugt.
Beim DI-Modell wurden alle Merkmalsvektoren zu einem Merkmalsvektor zusammengefaßt und dieser 23-dimensionale Vektor als Eingabe in einen Gaußklassifikator verwendet. Beim SI-Modell wurde für jeden Kanal ein Gaußklassifikator verwendet, die Ausgaben der beiden Klassifikatoren wurden anschließend gewichtet und multiplikativ integriert zu einer Gesamtklassifikation. Beim DR-Modell wurde als gemeinsamer Merkmalsraum das Spektrum des Signals verwendet, welches für den akustischen Kanal unmittelbar bereit steht. Für den visuellen Kanal wurde ein Polynomialklassifikator trainiert, der aus den visuellen Merkmalen ein zugehöriges Spektrum schätzte. Beide Spektren wurden anschließend gewichtet addiert, wobei der Wichtungsfaktor abhängig vom SNR war. Der Wichtungsfaktor war eine sigmoidale Funktion in Abhängigkeit des Signal-Rausch-Verhältnisses und wurde mit Hilfe der gestörten Daten aller SNR-Stufen gelernt. Das resultierende Gesamtspektrum wurde wiederum mit Hilfe eines Gaußklassifikators klassifiziert, welcher nur mit den akustischen Daten trainiert wurde. Beim Modell MR wurden zu jedem Kanal 3 artikulatorische Merkmale bestimmt: X und Y (höchster Punkt der Zunge, horizontal und vertikal gesehen) sowie S (Maß für die Lippenregion). Diese wurden aus den akustischen Daten mit Hilfe eines Polynomialklassifikators gewonnen. Die Abbildung wurde dabei mit Hilfe prototypischen Lehrervorgaben bestimmt, die von phonetischen Experten vorgeschlagen wurden. Für die visuellen Daten lag S direkt vor, Y wurde ebenfalls durch einen Polynomialklassifikator bestimmt, X konnte aus den visuellen Daten nicht bestimmt werden. Die Fusion erfolgte durch eine gewichtete Summe der Merkmale, wobei für jedes Merkmal ein anderes Gewicht verwendet wurde. Die endgültige Klassifikation erfolgte wieder durch einen Gaußklassifikator.
Die audio-visuellen Modelle zeigten in fast allen Fällen eine bessere Performanz als der akustischen Kanal alleine, für sehr kleine SNR-Werte waren sie allerdings schlechter als der visuelle Kanal alleine. Es zeigte sich, daß Modell MR bessere Ergebnisse als das Modell DR lieferten und die Performanz für die Modelle SI, MR und DI nahezu gleich war [RRPSE96].
Teissier et al. [TRRSGD99], [TGDS98] untersuchten die von Robert-Ribes et al. vorgeschlagenen Architekturen genauer, indem zum einen Crossvalidierungsexperimente durchgeführt wurden und zum anderen die verwendeten Modelle

so modifiziert wurden, daß die Anzahl der Merkmale in allen Modellen dieselbe war (dieses wurde auch schon ansatzweise in [RRSE95b] untersucht). Für die akustischen Daten wurde eine PCA verwendet, um die Anzahl der Merkmale auf 3 zu beschränken. Das verwendete Datenmaterial war dasselbe wie das von Robert-Ribes et al. benutzte. Trainiert wurden die Modelle mit gering verrauschten Daten, getestet wurde auf allen Rauschstufen. Zusätzlich wurde auch für Modell DI eine Gewichtung der Kanäle je nach SNR eingeführt, die direkt auf den Eingabevektoren erfolgte. Bei der Gewichtung beider Kanäle wurde entweder die Information über den SNR dem System direkt mitgeteilt ('mit Kontext') oder die Wichtungsparameter wurden mit Hilfe aller verrauschten Daten gelernt ('ohne Kontext'). Bei den Versuchen ohne Kontextinformation konnten die audio-visuellen Systeme nicht immer die Leistung des akustischen bzw. visuellen Systems übertreffen, so war das akustische System bei hohen SNR-Werten, das visuelle bei sehr niedrigen SNR-Werten besser. Die beste Performanz insgesamt zeigte das SI-Modell, gefolgt von den Modellen DI, MR und DR, welches in etwa eine Performanz wie das visuelle System zeigte. Bei den Versuchen mit Kontextinformation zeigten die fusionierten Systeme insgesamt eine bessere Leistung als die Einzelsysteme bei allen SNR-Werten. Die Modelle SI und DI zeigten die beste Performanz, gefolgt von Modell MR, DR lieferte wiederum die schlechtesten Werte.
Die schlechte Leistung von Modell DR im Vergleich zum Modell MR kann darauf zurückgeführt werden, daß Modell MR die Komplementarität der beiden Informationsquellen ausnutzen kann, wohingegen es beim Modell DR erforderlich ist, eng zusammenliegende visuelle Merkmale (z.B. für Daten, die zur selben Visemklasse gehören) auf sehr unterschiedliche Spektren abzubilden (d.h. die Daten gehören zu unterschiedlichen Phonemen).
In der Arbeitsgruppe um A. Waibel wurden verschiedene Versuche mit einem audio-visuellen Spracherkennungssystem unternommen, welches aus einem modularen MS-TDNN bestand. Eine erste Beschreibung dieses Systems ist in [BHMW93] enthalten. Die Daten bestanden aus Aufnahmen eines Sprechers. Als visuelle Daten wurden Graustufenbilder der Lippenregion verwendet, für die akustischen Merkmale wurden alle 10 msec 16 Melscale-Fourierkoeffizienten berechnet. In ersten Untersuchungen [DMW94] wurden verschiedene Möglichkeiten einer weiteren visuellen Vorverarbeitung untersucht, die einer Dimensionsreduktion der visuellen Eingabemerkmale dienen sollte. Dies geschah, um für beide Kanäle in etwa gleichviele Merkmale zu verwenden. Eine erste Repräsentation der Daten bestand in einer Tiefpaßfilterung und einem Downsampling, so daß das Lippenbild noch 384 Pixel enthielt. Anschließend erfolgte eine Histogrammnormalisierung. Eine weitere Vorverarbeitung dieses Graustufenbildes bestand in einer PCA mit 32 Koeffizienten. Die optimale Anzahl der Koeffizienten wurde dabei experimentell bestimmt. Weiterhin wurde eine LDA mit ebenfalls 32 Koeffizienten verwendet. Außerdem wurde die Anwendung einer Fouriertransformation untersucht. Für die Verwendung der Graustufenbilder sowie der LDA wurden die besten Ergebnisse erzielt. Es wurden unterschiedliche Fusionmöglichkeiten untersucht: In einer späten Fusionsarchitektur wurde die Gewichtung der Kanäle durch sogenannte Entropiegewichte durchgeführt. Dabei entspricht ein großer Entropiewert einer gleichmäßigen Verteilung der

Ausgabeaktivitäten, d.h. es gibt eine große Mehrdeutigkeit bei der Entscheidung. Weiterhin wurde auch eine frühe Fusion auf der Ebene der Eingabeschicht des Netzes sowie eine Fusion auf einer Zwischenstufe, nämlich auf der Ebene der versteckten Schichten des Netzes untersucht. Für die frühe und die späte Fusion ergaben sich ähnliche Performanzen, wobei für die späte Fusion immer ein besseres Ergebnis als für die Einzelkanäle alleine erzielt werden konnte, was nicht auf die beiden anderen Architekturen zutraf. Die besten Ergebnisse wurden für die Fusion auf einer Zwischenebene erzielt, wobei dies nicht für Versuche mit stark gestörten akustischen Daten der Fall war. Allerdings war der Unterschied zu den Performanzen der anderen Fusionsarchitekturen relativ klein. Es ist dazu anzumerken, daß die frühe Fusion und die Fusion auf der Zwischenstufe nicht separat optimiert wurden hinsichtlich der optimalen Anzahl der versteckten Neuronen. Zudem hatten diese beiden Architekturen keine gestörten Daten als Trainingsmaterial, während die späte Fusionsarchitektur aufgrund der Gewichtung der Kanäle in der Lage war, auf unterschiedliche Signal-zu-Rausch-Verhältnisse einzugehen.
In späteren Versuchen konnte ein automatisches Finden und Verfolgen der Lippenregion im Bild entwickelt werden (vgl. dazu auch [DHB+95]). In diesen Versuchen [MHD96], [MSYW99] konnte festgestellt werden, daß eine Vorverarbeitung der visuellen Daten mit Hilfe einer LDA bessere Resultate als mit Graustufenbildern erbrachte, wobei die LDA sich als empfindlicher gegenüber wechselnden Beleuchtungsbedingungen herausstellte.
In weiteren Versuchen [MSYW99] wurden unterschiedliche Möglichkeiten der Gewichtung der Einzelkanäle in der späten Fusionsarchitektur näher untersucht, da bei der späten Fusion eine unabhängige Kontrolle der beiden Modalitäten möglich ist. Die Aktivierungen der Neuronen des Phonem- beziehungsweise Visemlayers wurden dazu mit Hilfe einer gewichteten Summation fusioniert. Die Gewichte dieser Summation wurden mit Hilfe unterschiedlicher Methoden bestimmt: Zunächst wurde wie schon zuvor die Entropie beider Kanäle dazu benutzt, ergänzt mit einer Parameter, der einen der beiden Kanäle previligierte und per Hand je nach Qualität der Daten eingestellt wurde. Weiterhin war eine Wahl der Gewichte als Funktion des SNRs möglich, der automatisch alle 500 ms aus den akustischen Daten bestimmt wurde und für die dazwischenliegenden Zeitscheiben linear interpoliert wurde. Schließlich wurden die Gewichte auch mit Hilfe eines einschichtigen neuronalen Netzwerkes aus den Trainingsdaten bestimmt, d.h. sie wurden nicht während der Erkennungsphase berechnet und waren fest. Allerdings lieferte diese Methode eine Möglichkeit, für jedes Merkmal (d.h. Phonem oder Visem) ein unterschiedliches Gewicht zu wählen. Für die frühe und die Fusionsmöglichkeit auf der Zwischenstufe wurde als weiterer Versuch auch die Verwendung des SNR als zusätzliche Eingabe in die Fusionsebene des Netzes untersucht. Es bleibt zu bemerken, daß bei der frühen und der Fusion auf der Zwischenstufe sowie bei der Bestimmung der Gewichte mit Hilfe eines neuronalen Netzes keine explizite Information über die Qualität der akustischen Daten verwendet wurde.
Die verschiedenen Fusionsarchitekturen wurden mit ungestörten akustischen Daten trainiert und mit unterschiedlich gestörten akustischen Daten getestet. In jedem Fall zeigte sich eine Verbesserung der Klassifikationsleistung des akustisch-

visuellen Netzes gegenüber dem rein akustischen Netz, insbesondere bei gestörten Daten. Die besten Ergebnisse konnten mit der späten Fusionsarchitektur bei Verwendung des neuronalen Netzes zur Gewichtung der Kanäle erzielt werden. Bei stark gestörtem akustischen Kanal erzielte die Gewichtung mit Hilfe der Entropiemethode leicht bessere Ergebnisse. Nahezu in allen Fällen war die frühe Fusionsarchitektur die schlechteste, wobei bemerkt wurde, daß die präkategoriellen Fusionsarchitekturen in diesem Versuch keine Informationen über den Signal-Rausch-Abstand und somit einer geeigneten Gewichtung der Kanäle hatten. In einem letzten Versuch wurde in das Training auch gestörtes Datenmaterial mit einbezogen, so daß hier auch der SNR bei den präkategoriellen Architekturen mit einbezogen werden konnte, was mitunter zu leichten Verbesserungen führte. Insgesamt konnte durch das Training mit gestörten Daten die Performanz der System für gestörte Daten gesteigert werden, auch wenn die zum Testen verwendete Art der Störung nicht mittrainiert worden war. Auch hier zeigte die späte Fusionsarchitektur die besten Ergebnisse.
Adjoundani und Benoit [AB95a], [AB95b] untersuchten an einem sprecherabhängigen System, dessen Aufgabe die Erkennung isolierter Nonsense-Wörter war, zwei unterschiedliche Fusionsarchitekturen, eine frühe und eine späte Fusion. Als Klassifikator wurde dabei ein HMM eingesetzt. Für die Erstellung der visuellen Daten wurden einer Versuchsperson die Lippen geschminkt und aus den aufgenommenen Bildern der Lippenregion 6 unterschiedliche Merkmale wie z.B. die Höhe der Mundöffnung, sowie deren erste Ableitungen extrahiert. Als akustisches Merkmal wurden 12 Cepstral-Koeffizienten, deren erste Ableitungen sowie die Energie verwendet. Die Merkmalsvektoren wurden für beide Kanäle alle 20 ms bestimmt. Das System wurde mit unverrauschten akustischen Daten trainiert und mit verrauschten Daten getestet. Beim Training wurden Crossvalidierungsexperimente durchgeführt, verglichen wurden die gemittelten Testergebnisse. Bei der frühen Fusionsarchitektur wurden die akustischen und die visuellen Merkmalsvektoren konkateniert, bevor sie vom HMM klassifiziert wurden, bei der späten Fusionsarchitektur wurden zwei getrennte Hidden-Markov-Modelle mit jeweils den akustischen oder visuellen Daten trainiert. Bei der Integration beider Kanäle wurden drei verschiedene Strategien untersucht:

- Die Klassifikation des besten Kanals wurde übernommen. Dazu wurden beide Kanäle bewertet, indem die Dispersion der Ausgabewahrscheinlichkeiten aller HMMe für einen Kanal betrachtet wurde. Falls alle HMMe eine in etwa gleiche Ausgabewahrscheinlichkeit für eine Äußerung erzeugen, war diese Dispersion eher niedrig, so daß der Kanal mit der höchsten Dispersion ausgewählt wurde. Dessen bester Kandidat wurde als Klassifikation übernommen.

- Die Ausgabewahrscheinlichkeiten beider Kanäle wurden multipliziert, und es wurde die Klassifikation übernommen, die das Produkt maximiert.

- Es wurde die zweite Regel angewendet, allerdings dahingehend modifiziert, daß die beiden Faktoren gemäß der Dispersion der Ausgabewahrscheinlichkeiten (vgl. Methode a)) gewichtet wurden.

Es zeigte sich, daß der akustische Kanal alleine bei sinkendem SNR eine immer schlechtere Performanz zeigte, bei kleineren Werten als 18 dB war sie sogar geringer als die des visuellen Kanals. Die frühe Fusionsarchitektur zeigte immer eine bessere Leistung als die Klassifikation des akustischen Kanals alleine, sie war allerdings nur bis zu einem SNR von 12 dB auch besser als der visuelle Kanal. Bei geringeren Werten wurde sie schlechter und erfüllte damit nicht die an ein audio-visuelles Sprachsystem gesetzte Bedingung, immer besser als beide Einzelkanäle zu sein. Auch bei der späten Fusionarchitektur erfüllte die zweite Fusionsmethode diese Anforderung nicht, bei einem SNR kleiner als 6 dB wurde auch sie schlechter als der visuelle Kanal. Allerdings waren alle späten Architekturen immer besser als der akustische Kanal alleine. Die erste und dritte Fusionsmethode, die beide eine Gewichtung der Einzelkanäle unter impliziter Berücksichtigung des SNR realisierten, blieben beide immer über den Leistungen der Einzelkanäle, wobei die dritte Methode insgesamt die beste Performanz zeigte. Dementsprechend zeigte insgesamt die späte Fusionarchitektur eine bessere Leistung als die frühe Architektur.
Hennecke et al. [HSP95] [Hen97] untersuchten die Integrationsproblematik mit Hilfe eines Systems, welches Hidden-Markov-Modelle zur Klassifikation von isolierten Wörtern verwendete. Die akustischen Merkmale bestanden aus 14 Melscale-Koeffizienten, die alle 16 ms berechnet wurden. Zur Bestimmung der visuellen Merkmale wurden Farbbilder verwendet, in denen zunächst Gesicht und Mund gesucht wurden. Anschließend wurden die Lippen mit Hilfe von eines simplen Modell des Mundes nachgebildet. Die Änderungen der Parameter dieses Modells über die Zeit dienten als visuelle Merkmale. Da die Zeitintervalle, in denen akustische Merkmale ermittelt wurden, kleiner als diejenigen für die visuellen Daten waren, wurden die visuellen Merkmale interpoliert, indem die vorhergegangenen Daten dupliziert wurden. Als Datensatz dienten japanische Bahnhofsnamen, die unverauscht aufgenommen wurden. Es wurde eine frühe und eine späte Integrationsstruktur aufgebaut, indem zum einen die Merkmalsvektoren konkateniert wurden und zusammen die Eingabe in ein HMM bildeten, zum anderen wurden zwei verschiedene HMMs mit jeweils den akustischen und den visuellen Daten trainiert und anschließend deren Ausgaben gewichtet und additiv kombiniert. Dabei zeigte die späte Fusionsarchitektur bessere Leistungen als die frühe, die sogar schlechtere Leistungen als das Einzelsystem für die akustischen Daten zeigte.
Mit einem weiteren Datensatz von Nonsense-Wörtern, die einen hohen Grad von Koartikulation zeigten und verrauscht aufgenommen wurden, wurde zusätzlich zur frühen und späten Architektur auch eine Integration auf einer Zwischenstufe untersucht. Dazu wurden sogenannte Boltzmann-Zippers verwendet. Ein Boltzmann-Zipper besteht aus zwei Boltzmann-Ketten (spezielle Boltzmann-Maschinen), die durch zusätzliche Gewichte miteinander verbunden werden, woraus ein neues Netzwerk resultiert. Im Vergleich der Architekturen zeigte diese Integration auf einer Zwischenebene die besten Resultate.
Movellan und Chadderdon [MC96] verglichen eine frühe mit einer späten Fusionsarchitektur. Der Datensatz bestand aus 24 Beispielen von vier verschiedenen englischen Zahlwörten, gesprochen von 12 Sprechern. Die visuellen Daten basierten auf einem Graustufenbild der Lippenregion, welches symmetrisiert wur-

de. Die Differenzbilder zweier aufeinanderfolgender Zeitscheiben wurden berechnet, dann tiefpaßgefiltert und durch eine Schwellwertfunktion weiterverarbeitet. Für die akustischen Daten wurde alle 30 ms eine Cepstral-Analyse durchgeführt, wobei 12 Cepstral-Koeffizienten und deren Ableitungen die Merkmale bildeten, ergänzt durch den Energiewert und dessen Ableitung. Die späte Fusionarchitektur wurde durch zwei separate Hidden-Markov-Modelle realisiert, deren Ausgabewerte für jede Wortkategorie addiert wurden. Die Wortkategorie mit der höchstens Summe bestimmte die Klassifikation. Für das frühe Fusionsmodell wurden die akustischen und visuellen Merkmale konkateniert und mit Hilfe eines HMMs klassifiziert. Beide Modelle wurden mittels verrauschter Daten getestet, wobei der akustische Kanal mit Gaußschem Rauschen gestört wurde. Die Evaluation der beiden Architekturen erfolgte mittels der leaving-one-out-Methode. Die späte Fusionsarchitektur zeigte in den meisten Fällen eine bessere Performanz als die frühe Fusionsarchitektur, insbesondere für kaum und stark gestörte akustische Daten. Beide Fusionsarchitekturen lieferten bessere Ergebnisse als der akustische Kanal alleine.
Tomlinson et al. [TRB96] untersuchten ebenfalls ein System mit früher Integration und Integration auf einer Zwischenstufe. Der Datensatz bestand aus Zahlenreihen. Die akustischen Daten wurden mit Hilfe einer Filterbank mit 26 Filtern, die die Frequenzen von 60 Hz bis 10 kHz umfaßte, weiterverarbeitet. Die Merkmale wurden alle 10 ms berechnet. Zudem wurden künstlich verrauschte akustische Daten verwendet. Die visuellen Daten lagen alle 40 ms vor und bestanden aus einem Bild der Lippenregion des Sprechers. Dieser Vektor wurde mit Hilfe einer PCA auf 10 Dimensionen reduziert. Um eine Synchronisation der akustischen und visuellen Daten zu erreichen, wurden die visuellen Daten zeitlich in einem Verhältnis von 1:4 repliziert. Die akustischen und visuellen Daten wurden zu einem Merkmalsvektor konkateniert, die Klassifikation erfolgte durch ein HMM (frühe Fusion). Aufgrund der schlechten Erkennungsleistungen, die auf eine gewisse Asynchronität der akustischen und visuellen Daten zurückgeführt wurde, wurde eine weitere Architektur, basierend auf Hidden-Markov-Modellen, die auf den Einzelkanälen trainiert worden waren, untersucht. Diese Architektur wurde als Cross-Produkt-Topologie bezeichnet und realisierte eine Fusion auf einer Zwischenstufe. Dieses zweite Modell zeigte sowohl gegenüber dem reinen visuellen als auch dem akustischen System eine verbesserte Performanz bei allen untersuchten Rauschstufen.
Rogozan et al. [ADR96], [RDA97], [RD98] untersuchten verschiedene Systeme zur audio-visuellen Spracherkennung, in denen die Fusion beider Kanäle auf unterschiedlichen Ebenen durchgeführt wurde. Zudem wurde eine Wichtung beider Kanäle vollzogen. Die untersuchten Modelle verwendeten kontinuierliche Hidden-Markov-Modelle zur Klassifikation. Als Daten wurden sprecherabhängige Aufnahmen von kontinuierlich gesprochenen französischen Buchstaben verwendet. Für den akustischen Kanal wurden alle 10ms 13 Parameter (12 MFCC und die Gesamtenergie des Fensters, dieses war 30ms breit) sowie deren erste und zweite Ableitung berechnet. Die akustischen Daten wurden künstlich verrauscht, und zum Training hinzugezogen. Für die Verarbeitung des visuellen Kanals wurde ein geometrisches Lippenkonturen-Modell verwendet, welches die verwendeten Merkmale unabhängig von Umgebungseffekten extrahierte und

durch einen kleinen Satz an Parametern beschrieben werden konnte [RD97]. Die Berechnung der Merkmale erfolgte alle 20ms, so daß für eine Synchronisation des visuellen mit dem akustischen Kanal eine Interpolation der visuellen Merkmale notwendig wurde. Im ersten untersuchten Modell erfolgte die Fusion auf der Ebene der Eingabevektoren für das System. Eine Gewichtung beider Kanäle erfolgte mittels eines Gewichtsvektors, so daß der Beitrag des visuellen Anteils des Eingabevektors geändert werden konnte. Diese Gewichtung wurde empirisch bestimmt. In einem weiteren Modell erfolgte die Integration erst nachdem eine Klassifikation durchgeführt wurde. Dabei wurden die akustischen Klassifikationen als Ausgangspunkt verwendet, die vom visuellen Kanal weiter bewertet wurden. Beide Bewertungen wurden anschließend multipliziert und die Klassifikation mit der höchsten Bewertung wurde übernommen. Dabei erfolgte eine Gewichtung der beiden Kanäle, indem die Dispersion der Kanäle (vgl. [AB95b]) als Maß für die Störung des akustischen Kanals verwendet wurde. Als Entscheidungseinheit, auf die klassifiziert wurde, diente hierbei zunächst das Phonem, welches allerdings für den visuellen Kanal inadequat ist. In einem weiteren Versuch wurde daher die vom akustischen Kanal gemachte Entscheidung auf Phonemebene auf Visemebene abgebildet und dort die visuelle Bewertung berechnet. Die Viseme wurden dabei automatisch mit Hilfe eines selbstorganisierenden Kohonennetzes bestimmt [RD97]. Schließlich wurde noch eine hybride Architektur untersucht, indem im zweiten Modell mit später Fusion der akustische Anteil durch das Modell mit früher Fusion ersetzt wurde. Es fand also eine Integration beider Kanäle auf zwei Ebenen statt, auf einer frühen präkategoriellen, und einer späten, bei der die Ausgabe der frühen Fusion nochmals durch den visuellen Kanal bewertet wurde. Die Gewichtung beider Kanäle erfolgte wie in den jeweiligen Einzelsystemen. Als Ergebnis konnte festgehalten werden, daß die zusätzliche visuelle Information das Klassifikationsergebnis in jedem Fall verbesserte. Weiterhin konnten durch die automatische Gewichtung der Kanäle gute Ergebnisse erzielt werden, die vergleichbar mit denjenigen der empirisch ermittelten Gewichtungsfaktoren waren. Das hybride System zeigte mit die besten Ergebnisse, jedoch waren diese auch vom SNR abhängig, so lieferte für 0 dB das späte Fusionsmodell die besten Ergebnisse, für 10 db das frühe Fusionsmodell, während für stark gestörte Daten das späte Fusionsmodell nur sehr schlechte Ergebnisse erzielte. Dies ließe sich dadurch erklären, daß bei diesem System der visuelle Kanal nur zur Bewertung von akustischen Klassifikationen hinzugezogen wurde, ohne darüberhinausgehende Klassifikationmöglichkeiten beisteuern zu können Die akustischen Klassifikationen waren aber bei einem kleinen SNR schon von vorneherein schlechter, während beim frühen und beim hybriden System (welches das frühe System ja beinhaltet) der visuelle Kanal direkt die Klassifikation beeinflussen konnte. Der Gebrauch von Visemen zeigte sich als vorteilhaft.

Nakamura et al [NNS97] verwendeten zum Vergleich einer frühen und einer späten Fusionsarchitektur 1997 ebenfalls HMMe. Die Aufgabe bestand in der Klassifikation von Wörtern bei verschiedenem SNR. Dabei wurde ein Datensatz von insgesamt 5230 Wörtern aufgenommen. Die Aufnahmerate der akustischen Daten betrug 8ms pro frame, es wurden 33 akustische Merkmale generiert. Die visuelle Aufnahmerate betrug 33.3 ms pro frame. Das Graustufenbild wurde

einer nichtparametrischen Merkmalsextraktion unterzogen, resultierend in einem 48-dimensionalen Merkmalsvektor. Für den visuellen Kanal alleine konnte eine Erkennungsrate von etwa 85 % erzielt werden, die Erkennungsrate für den akustischen Kanal war abhängig vom SNR und lag bei 100 % bei 25 dB, ca. 70 % bei 20 dB, 15 % bei 15 dB und erreichte bei 0 db und darunter nur noch etwa 3 %. Bei der untersuchten frühen Fusionsarchitektur bestand die Eingabe in ein gemeinsames HMM aus dem konkatenierten akustischen und visuellen Vektor. Beim späten Fusionsmodell wurden getrennte HMMe für beide Kanäle trainiert, deren Ausgaben gewichtet multiplikativ kombiniert wurden. In beiden Fällen war das audio-visuelle System besser als der akustische Kanal alleine, für das frühe Fusionsmodell fiel die Performanz allerdings bei kleinem SNR (kleiner als 5dB) unter diejenige des visuellen Kanals alleine. Die Performanz des späten Fusionsmodells erreichte bei Signal-zu-Rausch-Verhältnissen kleiner als 5dB auch nur diejenige des visuellen Kanals. Zusammenfassend ließ sich feststellen, daß die späte Fusionsarchitektur eine bessere Performanz als die frühe Architektur zeigte.
Kumatani et al. [KNS01] untersuchten ebenfalls Fusionsarchitekturen auf der Grundlage von Hidden-Markov-Modellen. Sie verwendeten eine Datenbasis aus 4940 Worten, von denen 200 als Testmaterial verwendet wurden und die von einem Sprecher geäußert worden waren. Als akustische Merkmale wurden 16 Mel-Frenquency-Cepstralkoeffizienten und deren Deltawerte verwendet, die alle 8 ms berechnet wurden, d.h. es wurden 32 akutische Merkmale verwendet. Es wurden zudem gestörte akustische Daten verwendet. Die visuellen Informationen bestanden in einem Graustufenbild der Lippenregion, welches histogrammnormalisiert wurde und anschließend einer zweidimensionalen FFT unterzogen wurde. Als Merkmale wurden 35 logarithmierte Werte des Energiespektrums sowie deren Deltawerte verwendet. Zur Synchronisation beider Kanäle wurden die visuellen Merkmale entsprechend oft kopiert. Die untersuchten Fusionsarchitekturen waren eine frühe Fusion (Konkatenation beider Merkmalsvektoren, Verwendung eines einzelnen Hidden-Markov-Modells), ein spätes Fusionsmodell (getrennte HMMe für beide Kanäle, Zusammenführung auf der Ebene der Klassifikationen) und ein Modell auf einer Zwischenstufe. Dieses bestand aus einem Produkt-HMM (vgl. S. 87). Die dabei zugrundeliegenden einzelnen HMMs wurden entweder einzeln trainiert, oder das sich daraus ergebende gesamte System wurde nachtrainiert. Eine Gewichtung beider Kanäle erfolgte nicht. Versuche mit gestörten akustischen Daten bei verschiedenen Signal-Rausch-Verhältnissen zeigten tendenziell, daß das nachtrainierte Produkt-HMM die besten Ergebnisse erbrachte, gefolgt vom späten und vom frühen Fusionsmodell. Das einfache Produkt-HMM zeigte die schlechteste Performanz. Bei sinkendem SNR zeigte das frühe Fusionsmodell eine bessere Leistung als das späte, bei großen SNR-Werten kehrte sich dieser Trend um.
Luettin et al. untersuchten in [LPN01] unterschiedliche Fusionsmodelle, die auf Hidden-Markov-Modellen basierten. Das sprecherunabhängige System verarbeitete kontinuierliche Sprache. Im visuellen Kanal wurde zunächst das ganze Gesicht aufgenommen. Anschließend wurde eine Kosinustransformation der Lippenregion durchgeführt, deren Ergebnis einer LDA unterzogen wurde. Nach einer anschließenden Lineartransformation, ergaben sich 41 visuelle Merkma-

le. Diese wurden noch extrapoliert, damit zu jeder akustischen Zeitscheibe auch visuelle Merkmale vorhanden waren. Die akustischen Daten wurden alle 10 msec neu berechnet. Die 60-dimensionalen akustischen Merkmale bestanden aus Melfrequency-Cepstral-Koeffizienten (MFCC), die einer diskreten Kosinustransformation unterzogen wurden. Jeweils 9 aufeinanderfolgende Zeitscheiben wurden anschließend einer LDA unterzogen. Die Merkmalsextraktion wurde wie bei den visuellen Merkmalen mit einer Lineartransformation beschlossen. Die Verwendung der LDA auf mehreren konkatenierten Zeitscheiben diente einer Reduktion der Dimensionalität sowie der Modellierung von dynamischer Information in den Kanälen. Eine genauere Beschreibung der Vorverarbeitung kann [PLN01] entnommen werden. Die akustischen Daten wurden künstlich gestört. Das Training des Systems erfolgte immer mit Daten derselben Rauschstufe wie beim Testen. Der Erkenner bestand aus Hidden-Markov-Modellen, die Phoneme modellierten. In einem frühen Fusionsmodell (genannt 'frame synchronous integration') wurden die akustischen und visuellen Merkmalsvektoren konkateniert und anschließend von einem HMM klassifiziert. Bei der Verwendung von sogenannten 'multi-stream' Hidden-Markov-Modellen wurden beide Informationskanäle zunächst durch jeweils ein Einzel-HMM modelliert. Anschließend wurde ein neues HMM kreiert, indem die Übergangswahrscheinlichkeiten der einzelnen Hidden-Markov-Modelle für jeden Zustand kombiniert wurden. Bei dieser Kombination auf Zustandslevel konnten beide Informationskanäle gewichtet werden. Es handelte es sich um eine Fusionsarchitektur auf einer Zwischenstufe. Neben einem getrennten Training der Hidden-Markov-Modelle war auch ein gemeinsames synchrones Training des gesamten Systems mit dem konkatenierten audio-visuellen Vektor möglich. Dabei wurden beide Hidden-Markov-Modelle als ein System betrachtet. Beide Systeme werden als 'state synchronous' bezeichnet, da die Synchronisation auf Zustandslevel erfolgt. In beiden Fällen konnten die Exponenten, die der gewichteten Fusion beider Kanäle dienen, bestimmt werden, indem die trainierten Systeme auf einem unabhängigen Testsatz getestet wurden und dadurch die besten Werte durch Ausprobieren empirisch bestimmt wurden. In [GVN+01] wurden unterschiedliche Gewichtungsmethoden untersucht, wobei festgestellt wurde, daß der Gebrauch unterschiedlicher Gewichte sowohl für die akustischen Phonemklassen als auch für die visuellen Visemklassen am vorteilhaftesten war.
Um eine asynchrone Verarbeitung beider Kanäle zu ermöglichen, wurde ein sogenanntes Produkt-HMM verwendet. Dieses besteht aus zusammengesetzten Zuständen, die gemeinsame audio-visuelle Ausgabewahrscheinlichkeiten beinhalten und somit eine Fusion auf einer Zwischenstufe implementieren. Eine Synchronisation des ganzen Systems erfolgte auf Phonemebene. Ein Vergleich aller vier vorgestellten Systeme ergab, daß das frühe Fusionsmodell die schlechteste Performanz zeigte, gefolgt von den beiden zustandssynchronen Fusionsmodellen. Dabei zeigte ein gemeinsames Training das bessere Ergebnis. Am besten im Gesamtvergleich schnitt das Produkt-HMM ab.

Autoren	Architekturen (F,Z,S)	Gewinner	Trainings-methode	Methode	Anz. akust. Merkmale	Anz. vis. Merkmale	Sprecher -abhängig	Corpus	Gewichtung der Kanäle	Gewichtungs-methode
Stork '92	Z,S	S	NN	TT keine	14	5	ja	Buchstaben-worte	nein	-
Robert-Ribes '93, '95, '96	F,Z,S	alle in etwa gleich	Gaußklassi-fikator	TT st	20	3	ja	Vokale	F: nein Z,S: ja	SNR(M)
Teissier '98, '99	F,Z,S	S≥F>Z	Gaußklassi-fikator	CV st	3	3	ja	Vokale	ja	SNR(M), SNR, autom.
Duchnowski '94	F,Z,S	Z (*)	NN	TTV oh	16	384,32	ja	Buchstaben-worte	F,Z: nein S: ja	SNR(M)
Meier '96, '99	F,Z,S	S>Z>F	NN	TT st,oh	16	384,32	ja	Buchstaben-worte	F,Z: nein S: ja	SNR(M), autom.
Adjoudani '95	F,S	S	HMM	CV oh	25	6	ja	Nonsense-Worte	F: nein S: ja	autom.
Hennecke '95, '97	F,Z,S	Z>S>F	HMM	TT keine	14	12	nein	isolierte Worte	F,Z: nein S: ja	SNR(M)
Movellan '96	F,S	S	HMM	CV oh	26	300	nein	Ziffern	nein	-
Tomlinson '96	F,Z	Z	HMM	TT oh	26	10	ja	Ziffern	nein	-
Rogozan '96,'97 '98	F,S, Kombina-tion aus F u. S	Komb. (*)	HMM	TT oh	39	9	ja	Buchstaben-worte	ja	SNR(M), autom.
Nakamura '97	F,S	S	HMM	TTV oh	33	48	ja	Worte	ja	SNR(M)
Kumatani '01	F,Z,S	Z>S(*)>F(*)	HMM	TTV k.A.	32	70	ja	Worte	ja	SNR(M)
Luettin '01	F,Z	Z	HMM	TTV st	60	41	nein	kontinuierl. Sprache	F: nein Z: ja	SNR(M)

Tabelle 5.2: Tabelle: Fusionsarchitekturen in der Literatur, Vergleiche, F=frühe Fusion, Z= Fusion auf Zwischenstufe, S= späte Fusion, (*)= für die meisten Signal-Rausch-Verhältnisse, > = ist besser als, Trainingsmethode: TT=Datensatz wird in Trainings- und Testdaten aufgespalten, TTV= Datensatz wird in Trainings-, Test- und Validierungssatz aufgespalten, CV=Datensatz wird in Cross-Validierungssätze aufgespalten, keine=keine gestörten Daten verwendet, st=Training mit gestörten Daten, oh=Training nur mit ungestörten Daten; Gewichtungsmethode: autom.=automatische Gewichtung ohne Bestimmung des SNR, SNR(M)= empirische Ermittlung des für den jeweiligen SNR geeigneten Parameterwert des Gewichts, SNR=automatische Gewichtung unter Berücksichtigung des SNR

5.5 Diskussion

Der überwiegende Anteil der vorgestellten Systeme implementiert eine sprecherabhängige Erkennung, wobei der verwendete Datensatz größtenteils aus einem eng umrissenen Wortschatz besteht (z. B. Buchstabenworte, Vokale, Ziffern oder eine begrenzte Anzahl an Worten). Die Anzahl der visuellen Merkmale variiert von nur wenigen (z.B. 3 beschreibende Parameter) bis zu hochdimensionalen (z.B. 500 Pixel eines Graustufenbildes). Die Dimensionalität der akutischen Merkmale variiert weit weniger und liegt zwischen 3 und 64, nicht immer ist die Anzahl der visuellen Merkmale größer als diejenige der akustischen. Nur in einigen Fällen sind beide Anzahlen in etwa gleich groß.
Die meisten vorgestellten Systeme basieren auf Hidden-Markov-Modellen, wobei eine Fusion auf einer Zwischenstufe in fast allen Fällen durch Multi-Stream-Modelle realisiert wurde.
Die Systeme werden größtenteils ohne gestörte Daten trainiert, das Testen findet mit gestörten Daten statt. Zum Training wird der Datensatz zumeist lediglich in einen Trainings- und einen Testdatensatz geteilt (TT). Ein weiterer Anteil des Datensatzes als Validierungssatz wird nur selten verwendet (TTV) wie auch wenige Cross-Validierungsexperimente (CV) durchgeführt werden.
Eine Gewichtung der einzelnen Kanäle wird für späte Fusionsarchitekturen fast immer durchgeführt, für die anderen Fusionsmöglichkeiten in etwa der Hälfte der Fälle. Die Gewichtung erfolgt dabei fast immer unter Verwendung empirisch ermittelter Parameterwerte, die je nach Signal-Rausch-Abstand der Daten gewählt werden (SNR(M)). Falls eine automatische Gewichtung beider Kanäle durchgeführt wird, findet fast immer die Dispersion, wie sie in [AB95a] eingeführt wurde, Verwendung.
In etwa der Hälfte der besprochenen Arbeiten werden mehrere Fusionsmöglichkeiten miteinander verglichen. Dabei ist die späte Fusionsmöglichkeit fast immer dabei, wie diese auch in den meisten der Arbeiten Verwendung findet, die nur eine Möglichkeit der Fusion betrachten. In den Vergleichsuntersuchungen werden nur in insgesamt sechs Arbeiten alle Fusionsmöglichkeiten betrachtet, wobei die Arbeiten aus der Arbeitsgruppe um A. Waibel neuronale Netze als Grundlage verwenden, die Arbeiten aus der Arbeitsgruppe um J. Robert-Ribes Gaußklassifikatoren und zwei unabhängige Arbeitsgruppen (Kumatani, Hennecke) Hidden-Markov-Modelle.

Kapitel 6

Das Worterkennungssystem

Für die Untersuchung der Sensorfusionsproblematik wurde ein Worterkennungssystem für die Erkennung sogenannter Buchstabenwörter entwickelt (siehe auch [TKP97],[KTWP97]). Dabei handelt es sich um das gesprochene Alphabet der deutschen Sprache. Der Aufbau des Worterkennungssystems sowie seine Komponenten werden dabei möglichst einfach gestaltet, um dessen Einfluß auf einen Vergleich der Fusionsarchitekturen überschaubar zu halten und gleichzeitig die Realisierung möglichst vieler Fusionsarchitekturen zu ermöglichen.

In dem entwickelten Worterkennungssystem wird die Problematik der unterschiedlichen Länge der Signale durch die Aufspaltung des Klassifikationsprozesses in zwei Module gelöst (vgl. Abb. 6.1). Das erste Modul dieses Systems dient der Klassifikation eines einzelnen Zeitfensters und liefert eine sogenannte Segmentklassifikation. Für die Klassifikation des ganzen Signals wird das Zeitfenster über das Signal geschoben, und die aufeinanderfolgenden Zeitfenster des Signals erzeugen an der Ausgabe des ersten Moduls eine Sequenz von Segmentklassifikationen (vgl. Abb. 6.1). Die abschließende Zuordnung dieser Sequenz zu einem Buchstabenwort wird vom zweiten Modul übernommen, so daß die Zeitintegration über das Signal im zweiten Modul auf symbolischer Ebene erfolgt. Die Realisierung beider Module ist beliebig, im Rahmen der vorliegenden Arbeit wurden allerdings neuronale Netze für die Realisierung verwendet, da diese eine große Bandbreite an Fusionsmöglichkeiten bieten.

In der vorliegenden Arbeit wird für das erste Modul ein MLP eingesetzt (vgl.

Anfangssegmentklassen:	Endsegmentklassen:
/a:/, /e:/, /i:/, /o:/, /u:/	/a:/, /e:/, /i:/, /o:/, /u:/
/b/, /c/, /d/, /g/, /p/, /t/, /w/	/e:/
/e/	/f/, /l/, /m/, /n/, /r/, /s/
/h/, /k/	/a:/
/q/	/u:/
/v/, /z/, /j/	/au/, /et/, /ot/
/i/	/ks/

Tabelle 6.1: Übersicht über die Anfangs- und Endsegment-Klassen

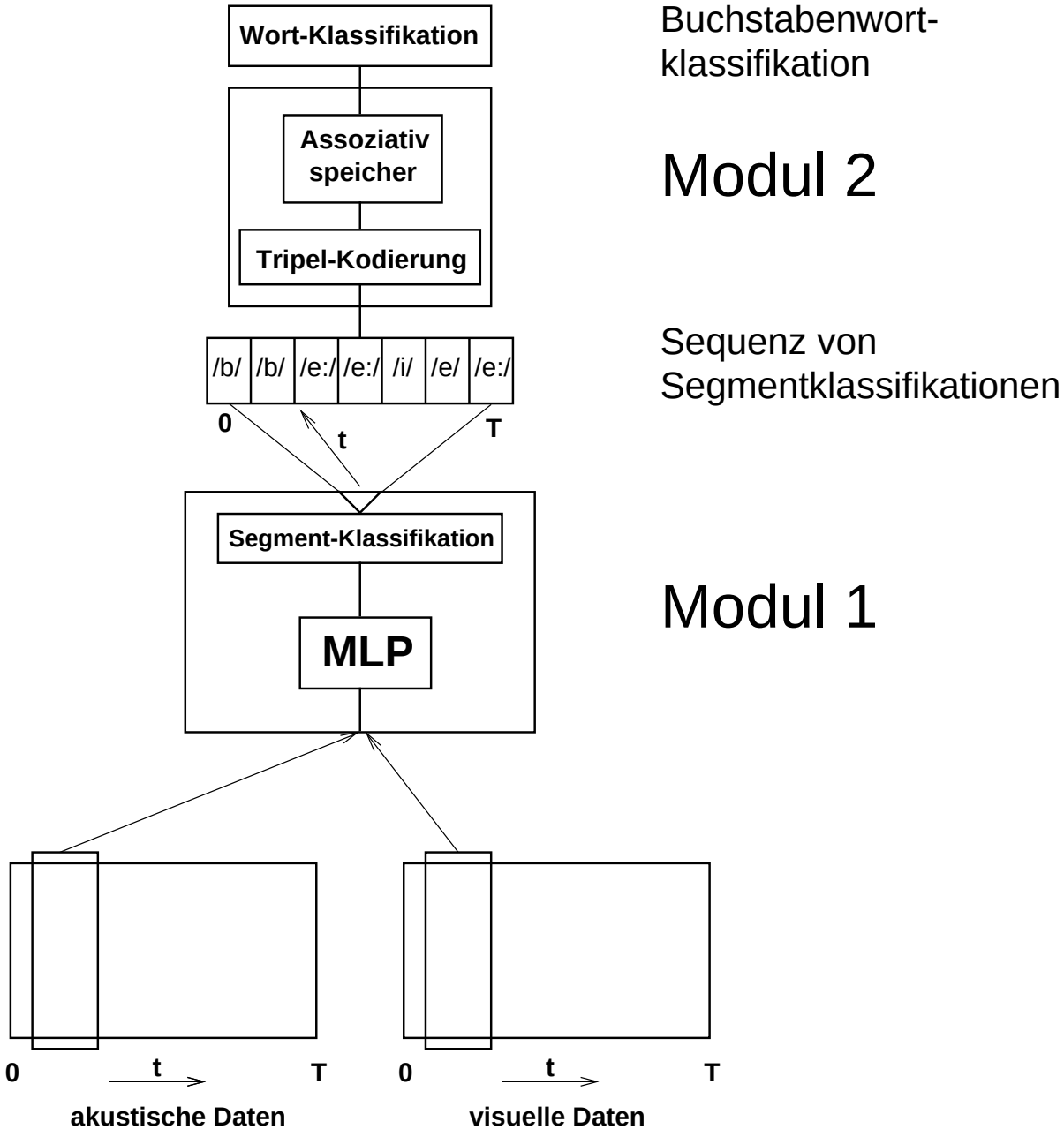

Abbildung 6.1: Worterkennungssystem für Buchstabenwörter

Buchstabenwort	Anfangssegmentklasse	Endsegmentklasse
A	/a:/	/a:/
B	/b/	/e:/
C	/c/	/e:/
D	/d/	/e:/
E	/e:/	/e:/
F	/e/	/f/
G	/g/	/e:/
H	/h/	/a:/
I	/i:/	/i:/
J	/j/	/ot/
K	/k/	/a:/
L	/e/	/l/
M	/e/	/m/
N	/e/	/n/
O	/o:/	/o:/
P	/p/	/e:/
Q	/q/	/u:/
R	/e/	/r/
S	/e/	/s/
T	/t/	/e:/
U	/u:/	/u:/
V	/v/	/au/
W	/w/	/e:/
X	/i/	/ks/
Z	/z/	/et/

Tabelle 6.2: Segmentklassenzuordnungen für die einzelnen Buchstabenwörter

3.3.1). Für das Training eines MLPs werden gelabelte Daten benötigt, d.h. für jedes Trainingsmuster, das in diesem Fall aus einem Zeitfenster des Signals besteht, wird ein Lehrersignal gebraucht. Jedes Zeitfenster wird daher einer von 30 sogenannten Segmentklassen zugeordnet. Eine Übersicht dieser Segmentklassen ist der Tabelle 6.1 zu entnehmen, wobei das polysyllabische 'Y' ausgenommen wurde. Für das Training des MLPs werden das jeweils erste und letzte Zeitfenster eines Buchstabenworts verwendet, da eine exakte Zuordnung eines Zeitfensters zu einer Segmentklasse nur für diese gemacht werden kann. So wird dem ersten Zeitfenster des Buchstabenwortes 'B' die Klasse /b/ und dem letzten Zeitfenster die Klasse /e:/ zugeordnet. Die Klassenzuordnungen der Anfangs- und Endzeitfenster der einzelnen Buchstabenwortklassen sind der Tabelle 6.2 zu entnehmen. Da nur die jeweils ersten und letzten Zeitfenster zum Training des Netzes verwendet werden, bleibt anzumerken, daß dieses bei der Klassifikation aller Zeitfenster eines Buchstabenwortes über die zwischen dem ersten und letzten Zeitfenster liegenden Zeitfenster generalisieren muß.
Das zweite Modul wird mit Hilfe eines Hetero-Assoziativspeichers realisiert, wie er in Kapitel 3.3.4 beschrieben wird. Um die zum Trainingsdatensatz gehören-

den Buchstabenwörter im Assoziativspeicher abzuspeichern, wird zunächst die Sequenz der Zeitfenster des Buchstabenwortes vom (mit den Anfangs- und Endzeitfenstern des Buchstabenwortes) trainierten MLP klassifiziert und so eine Sequenz von Segmentklassifikationen erzeugt. Die Sequenzen einzelner Buchstabenworte sind dabei unterschiedlich lang.
Diese Problematik kann durch Anwendung einer Tripelkodierung und eines Assoziativspeichers (vgl. Kapitel 3.3.4) gelöst werden. Durch die Verwendung einer Tripelkodierung können die Sequenzen auf Binärvektoren gleicher Länge abgebildet und so eine Zeitintegration hergestellt werden. Die Länge des Binärvektors ist durch die Anzahl der möglichen auftretenden Tripel von vorneherein festgelegt, da jede Vektorkomponente genau einem Tripel, welches auftreten kann, zugeordnet wird. Die in der Sequenz auftretende Tripel werden unabhängig von ihrer Position innerhalb der Sequenz (Translationsinvarianz) an jeweils gleicher Stelle innerhalb des Binärvektors durch eine '1 dargestellt. Wiederholt auftretende Tripel werden daher nur einmal kodiert. Die Binärvektoren können im Assoziativspeicher gespeichert werden, indem als gewünschte Ausgabe die Klassenzugehörigkeit des Buchstabenwortes in einer 1-aus-m-Kodierung angegeben wird.
Die vorgestellte Methode der Zeitintegration des Signals durch Verwendung einer Tripelkodierung und eines Assoziativspeichers bietet eine Alternative zu herkömmlichen Verfahren wie z.B. dem DTW [1] oder Hidden-Markov-Modellen. Gleichzeitig gelingt eine homogene Realisierung eines Worterkennungssystems, da dieses komplett neuronal aufgebaut ist. Dies erlaubt zudem eine Vereinfachung des Vergleichs der unterschiedlichen Fusionsarchitekturen.
Um die Performanz des trainierten Systems zu bestimmen, werden anschließend die Testmuster klassifiziert. Dazu wird wiederum ein Zeitfenster über das Signal des Testmusters geschoben, wobei jedes Zeitfenster vom MLP einer Segmentklasse zugeordnet wird. Nach einer anschließenden Tripel-Kodierung der erzeugten Segmentklassensequenz und Eingabe in den Assoziativspeicher liefert dieser die Klassenzugehörigkeit zu einem Buchstabenwort.
Diese Vorgehensweise soll nochmal anhand eines Beispiels erläutert werden:
Das vorverarbeitete Signal einer Äußerung des Buchstabenworts 'B' wird in das erste Modul hineingegeben. Es bestehe aus 9 aufeinanderfolgenden Zeitfenstern. Das MLP klassifiziert jedes Zeitfenster separat (Segmentklassifikation) und liefert als Ausgabe zum Beispiel folgende Sequenz von Segmentklassifikationen:

/b//b//e://e://e://e://e://i://e:/

Diese Sequenz besteht aus folgenden unterschiedlichen Tripeln:

/b//b//e:/, /b//e://e:/, /e://e://e:/, /e://e;//i:/ und /e://i://e:/.

Die Kodierung jedes dieser Tripel erfolgt, indem an jeweils entsprechender Stel-

[1] So wird in den Arbeiten der Arbeitsgruppe Waibel (vgl. Kapitel 5) die Zeitintegration mit Hilfe des DTW gelöst. Das in diesen Versuchen benutzte Datenmaterial ist zum Teil dasselbe wie das in dieser Arbeit verwendete (vgl. Kapitel 7). Die Performanzen beider Systeme sind ähnlich gut, wenn auch ein genauer Vergleich nicht möglich ist, da in diesen Arbeiten die Performanz für Sequenzen von Buchstabenwörtern ermittel wird.

le im Binärvektor eine 1 eingefügt wird. Das wiederholt auftretende Tripel /e://e://e:/ wird dabei nur einmal gezählt. Dieser Binärvektor wird wie in Kapitel 3.3.4 dargestellt vom Assoziativspeicher schließlich einem Buchstabenwort zugeordnet.

Kapitel 7

Durchführung der Fusionsexperimente

In diesem Kapitel sollen die Einzelheiten der Experimentdurchführungen beschrieben werden:
Zunächst wird das verwendete Datenmaterial vorgestellt. Anschließend wird die Auswahl der Trainings- und Testmuster sowie die genaue Spezifikation des Worterkennungssystems und der verwendeten Netze beschrieben. Abschließend werden die zu untersuchenden Fusionsarchitekturen vorgestellt.

7.1 Datenmaterial

Der in dieser Arbeit verwendete Datensatz wurde im Rahmen des Landesforschungsschwerpunktes Neuroinformatik aufgenommen und von einem der Projektpartner, der Forschungsgruppe von Prof. Alex Waibel, Karlsruhe, zur Verfügung gestellt. Eine genaue Beschreibung des Datensatzes findet sich in [Mei95]. Es handelt sich dabei um buchstabierte Namen und Zufallssequenzen, wobei ein einzelner dieser Buchstaben als Buchstabenwort bezeichnet wird. Der Wortschatz enthält Vornamen, Städtenamen und Äußerungen einer zufälligen Aneinanderreihung von Buchstaben. Die buchstabierten Sequenzen haben unterschiedliche Längen, wobei die Buchstaben des deutschen Alphabets nicht in gleicher Häufigkeit auftreten. Die akustischen und visuellen Daten wurden gleichzeitig mit einer NTSC-Kamera und einem Standardmikrophon aufgenommen, wobei die Synchronisation mit Hilfe von Zeitstempeln erfolgte [1]. Die Labelung der Daten erfolgte durch forced alignment auf akustischer Ebene durch einen Spracherkenner (MS-TDNN). Die Label wurden dann für die visuellen Daten übernommen. Die Labelung und der Spracherkenner werden genauer in [HW93] beschrieben. Zur Verfügung stehen die Sprach- und Bilddaten eines Sprechers.
Die akustischen Daten wurden mit einer 16 kHz Samplingrate aufgenommen und mit einer Präzision von 2 Byte pro Abtastpunkt abgespeichert. Um diese akustischen Daten in der vorliegenden Arbeit zu verwenden, wurde zur weiteren

[1] Jedesmal, wenn ein neues Bild vom Framegrabber übertragen wurde, wurde die Systemzeit mit abgespeichert, was als Zeitstempel bezeichnet werden soll.

Vorverarbeitung alle 10 ms eine Kurzzeitfouriertransformation mit einem Hammingwindow von 20 ms Breite durchgeführt. Eine anschließende Filterung der Frequenzwerte mit einer Filterbank führt zu einer Frequenzrepräsentation mit 16 Werten. Die in der Filterbank verwendeten Filter haben Mittenfrequenzen, welche gleiche Abstände auf der Bark-Skala haben, die Bandbreiten der Filter, gemessen in Bark, sind gleich (vgl. dazu Kapitel 2.3)[2].
Als visuelle Daten wurden nur die Lippenregion des Sprechers aufgenommen. Dazu wurde das Kamerabild des Sprechers diesem auf einem Bildschirm zusammen mit einem Rechteck gezeigt, und der Sprecher mußte bei der Aufnahme darauf achten, daß die Lippenregion innerhalb des Rechteckes blieb. Die Daten wurden anschließend handgelabelt, d.h. für jede Sequenz wurde zu Anfang der Mittelpunkt der Lippenregion von Hand markiert und diese Position dann für den Rest der Sequenz als konstant angenommen. Anschließend wurde ein Ausschnitt von 144 x 80 Pixeln aus dem Bild herausgeschnitten. Die Aufnahmen wurden ohne Tageslicht (d.h. bei abgedunkeltem Fenster) und mit Hintergrundbeleuchtung gemacht. Jedes Pixel des Bildes wurde als 8bit-Graustufeninformation gespeichert. Um eventuelle Unterschiede der Beleuchtungsbedingungen zu eliminieren, wurde als weitere Vorverarbeitung eine histogrammnormalisierte Graustufenkodierung durchgeführt. Dabei wird jedes Pixel als Durchschnittswert eines kleinen Quadrates des Originalbildes berechnet (was einer Tiefpassfilterung des Originalbildes entspricht und zu einer Bildgröße von 24 x 16 Pixeln führt). Die Grauwerte werden anschließend so reskaliert, daß das dunkelste bzw. hellste 5% der Pixel im Histogramm mit 0 bzw. 256 kodiert wird. Die dazwischenliegenden Pixel werden anschließend linear angepaßt. Die visuellen Daten wurden mit einer Frequenz von 30 Hz aufgenommen. Damit auch für den visuellen Kanal alle 10 ms Daten zur Verfügung stehen, wurde das aktuelle Bild über 3-4 akustische Zeitscheiben lang wiederholt.
Um eine weitere Dimensionsreduktion der visuellen Daten zu erhalten, wurde in dieser Arbeit die Symmetrie des Lippenbildes ausgenutzt (vgl. auch [Sto97], [Mei95], [Mov95]). Dazu wurde das Bild in der Mitte vertikal geteilt, eine Hälfte gespiegelt und der Mittelwert aus beiden Lippenhälften berechnet. Zusätzlich wurde noch der Rand des Bildes abgeschnitten, der keine Bildinformationen über die Lippen enthielt, so daß das schließlich verwendete Lippen-Halbbild eine Größe von 10 x 14 Pixeln besaß. Zunächst (Originaldaten) keine weitere Vorverarbeitung der Daten durchgeführt, um keine Informationen zu verlieren.
Zusätzlich wurden in dieser Arbeit auch Versuche mit einem gestörten akustischen Kanal durchgeführt. Dazu wurde zu dem akustischen Zeitsignal ein Gaußsches Rauschen addiert. Das Signal-Rausch-Verhältnis (SNR) beträgt bei den drei erzeugten gestörten akustischen Datensätzen 3 dB, -3 dB und -11 dB.
Abbildung 7.1 zeigt zum Vergleich die Abhängigkeit der Verständlichkeit für Logatome (Einsilbler) vom Signal-Rausch-Verhältnis für normalhörende Versuchspersonen. Bei einem SNR von 3 dB liegt die Verständlichkeit noch bei ca. 80%, während sie bei einem SNR von -3dB schon auf nur 50% abgesunken ist. Einem SNR von -11 dB entspricht eine Verständlichkeit von weniger als 10%.

[2]Die Vorverarbeitungssoftware für die akustischen Daten wurde von A. Wichert in PV-WAVE implementiert.

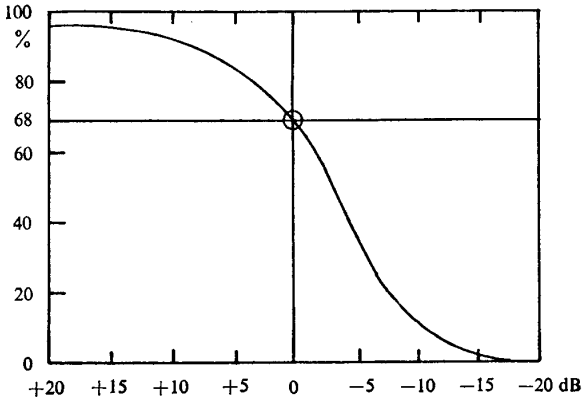

Abbildung 7.1: Diskriminationsvermögen für Einzelworte in Beziehung zum Signal-Rauschverhältnis für normalhörende Versuchspersonen, nach [Gro68]

7.2 Auswahl der Trainings- und Testdaten

Um die Performanz einer Architektur zu bestimmen, ist es notwendig, diese mit Hilfe eines vom Trainingsdatensatz unabhängigen Testdatensatzes zu testen. Die Performanz auf dem Testdatensatz wird auch als Generalisierungsleistung bezeichnet. Der zur Verfügung stehende Datensatz wird daher in Trainings- und Testdatensatz aufgespalten. Da bei den vorliegenden Untersuchungen nur ein vergleichsweise kleiner Datensatz zur Verfügung stand, wurden Crossvalidierungsexperimente durchgeführt, bei denen der gesamte Datensatz zum Training verwendet werden kann (vgl. [Rip96], [Bis95]). Dazu wird der Datensatz D, welcher aus N Datenpunkten besteht, zufällig in k gleich große disjunkte Segmente ν_j aufgeteilt, so daß gilt:

$$D = \bigcup_{j=1}^{k} \nu_j \tag{7.1}$$

und

$$\forall i \neq j : \nu_i \cap \nu_j = \emptyset \tag{7.2}$$

Zum Training wird ein Trainingsdatensatz T_j verwendet, der aus $k-1$ Segmenten besteht:

$$T_j = D \setminus \nu_j, \tag{7.3}$$

getestet wird mit dem Segment ν_j. Dieser Prozeß wird k-fach wiederholt, wobei immer ein anderes Segment als Testdatensatz dient. Der Prozeß der Aufspaltung des Datensatzes und der Durchführung der Versuche ist in Abbildung 7.2 illustriert.

Anschließend werden die Testfehler über alle k Ergebnisse gemittelt, um die Generaliserungsleistung des Netzes zu ermitteln.

Ein Vorteil des Verfahrens liegt darin begründet, daß ein großer Teil des Datenmaterials zum Training verwendet werden kann, gleichzeitig aber alle Daten auch für das Testen benutzt werden. Nachteilig wirkt sich dabei aber die große

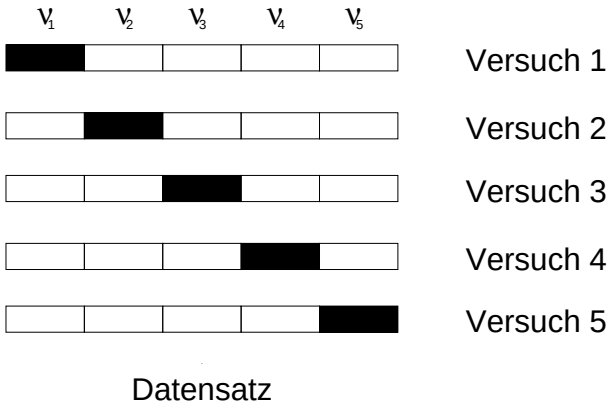

Abbildung 7.2: Schematischer Aufbau eines Crossvalidierungsexperiments mit k=5. Gezeigt ist die Aufteilung des Datensatzes in 5 Teilsegmente sowie ihre Verwendung bei den einzelnen Versuchen. Das jeweils zum Testen verwendete Datensegment ν_j ist schwarz eingezeichnet, die zum Trainingsdatensatz gehörenden Segmente sind weiß.

Anzahl an Versuchen aus, die unter Umständen bei Wahl eines großen k durchgeführt werden müssen.
Für k sind daher Werte von 5 oder 10 üblich (vgl. [Moo98]). Wird $k = N$ gewählt, bezeichnet man das Verfahren auch als leaving-one-out.
Bei der Bestimmung der Performanz einer Architektur ist außerdem zu beachten, welche Parameter auf die Klassifikationsleistung einen Einfluß haben. Beim Backpropagation-Algorithmus sind dies außer der Auswahl der Trainings- und Testdatensätze auch die Initialisierung der Netze (vgl. Kapitel 3.3.1, [KP91]). Daher muß auch diese berücksichtigt werden, indem die Lernprozedur einige Male mit unterschiedlichen Initialisierungen durchgeführt wird, und die Variation der Genauigkeit der resultierenden Klassifikatoren bestimmt wird (vgl. dazu [Die98]).
Bei den hier durchgeführten Experimenten wurde $k = 5$ gewählt. Zudem wurden für jede untersuchte Architektur 5 verschiedene Initialierungen verwendet. Durch die Kombination aller Crossvalidierungssätze mit allen Initialisierungen ergaben sich somit 25 Versuche pro Experiment.
Als Ergebnis werden der Mittelwert der Klassifikationsleistungen x_i aller 25 Realisierungen R einer Architektur und das 95%ige Konfidenzintervall angegeben. Der Mittelwert $\bar{x}$ berechnet sich gemäß

$$\bar{x} = \frac{1}{R} \sum_{i=1}^{R} x_i. \tag{7.4}$$

Um die Ergebnisse der Experimente miteinander vergleichen zu können, wurde für jedes Experiment außerdem das 95%ige Konfidenzinterval bestimmt. Mit einer 95%igen Wahrscheinlichkeit liegt der wahre Mittelwert der Verteilung μ innerhalb dieses Intervalls. Das Konfidenzintervall kann mit Hilfe des Standard-

fehlers des Mittelwertes

$$\hat{\sigma_{\bar{x}}} = \sqrt{\frac{\hat{\sigma^2}}{R}} \tag{7.5}$$

mit

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{R}(x_i - \bar{x})^2}{R-1} \tag{7.6}$$

als der erwartungstreuen Schätzung der Varianz berechnet werden. Dazu wird der Standardfehler mit denjenigen z-Werten multipliziert, die in der Standardnormalverteilung die mittleren 95% der Fläche begrenzen. Da der Stichprobenumfang bei den Untersuchungen kleiner als 30 war, wurde als Multiplikator anstelle des z-Wertes derjenige t-Wert verwendet, der von der t-Verteilung mit R-1 Freiheitsgraden 95% der Fläche ausschneidet. Dieser beträgt im vorliegenden Falle 2.064 für $R-1 = 24$, so daß sich für die Bestimmung des Konfidenzintervalls

$$\sigma_{\bar{x}KI} = 2.064\frac{\hat{\sigma}}{\sqrt{R}} \tag{7.7}$$

ergibt (vgl. [Bor93]).
Nach [Coh95] kann gezeigt werden, daß ein zweiseitiger t-Test zwei Mittelwerte als unterschiedlich einstufen würde, wenn sich die Konfidenzintervalle der beiden Mittelwerte nicht überlappen, d.h. wenn die Differenz der beiden Mittelwerte größer als die Summe ihrer Konfidenzintervalle ist.
Da die Verteilung der Klassen innerhalb des Datensatzes nicht gleichverteilt war, (und es hinsichtlich der Segmentklassen auch nicht möglich war, diese Verteilung auszugleichen, da die Klasse /e:/ z.B. in sehr viel mehr Buchstabenwörtern vorkommt als etwa die Klasse /i/), wurde zur Bewertung der Performanz nicht nur die 'allgemeine' Performanz bestimmt, bei der alle richtig klassifizierten Buchstabenwörter aufsummiert und durch die Gesamtanzahl der Klassifikationen geteilt wurde, sondern zusätzlich ein anderes Fehlermaß berechnet, welches im folgenden als Klassenspezifisches Mittel (Abkürzung: KSPM) bezeichnet werden soll. Dabei wird der Prozentsatz der richtig klassifizierten Muster f_i für jede Klasse i getrennt bestimmt und anschließend über alle Klassen gemittelt:

$$\bar{f_{KSPM}} = \frac{1}{m}\sum_{i=1}^{m} f_i\,, \tag{7.8}$$

wobei m die Anzahl der Klassen und f_i die klassenspezifische Performanz bezeichnen soll, die für jede Klasse i getrennt ermittelt wird.
Für alle untersuchten Architekturen werden die allgemeine und die klassenspezifische Performanz auf dem Trainings- und dem Testdatensatz sowohl für die Segmentklassifikation (die am Ausgang des ersten Moduls des Worterkennungssystems berechnet wird) als auch für die Klassifikationen der Buchstabenwörter bestimmt und im Anhang angegeben.

7.3 Spezifikation des Worterkennungssystems und Training der Netze

Als Eingabe in das Worterkennungssystem werden Zeitfenster der Länge 100 ms verwendet, um auch zeitliche Änderungen im Signal noch erfassen zu können [3]. Daraus resultiert für die akustischen Daten ein Eingabevektor der Dimension 160, da 100 ms 10 Zeitscheiben entsprechen und für jede Zeitscheibe Werte für 16 Frequenzfilter existieren. Der entsprechende Eingabevektor des visuellen Kanals ist dagegen mit mindestens 560 Dimensionen relativ hochdimensional: da nur alle 3-4 Zeitscheiben ein neues Lippenbild zur Verfügung steht, genügt es, jeweils das erste, vierte, siebte und zehnte Bild eines 100 ms-Zeitfensters zu benutzen, um alle möglichen Lippenbilder, die im Zeitfenster vorkommen können, zu erfassen, so daß sich $4 \times 140 = 560$ Eingabewerte ergeben. Daher wurde für den visuellen Kanal bei Verwendung der Originaldaten nur das Lippenbild der ersten Zeitscheibe als Eingabevektor verwendet. Das erste Lippenbild wurde verwendet, weil bei der Ausprache die Artikulatoren ihre zur Artikulation erforderliche Position erst einnehmen müssen, bevor sie 'artikulieren' können, d.h. das visuelle Signal liegt zeitlich immer gleichzeitig oder etwas vor dem dazugehörenden akustischen Signal (vgl. auch [BOK94],[BK94]).
Dieser Datensatz soll im folgenden als 'hochdimensional' oder als 'Originaldaten' bezeichnet werden. Bei der Verwendung einer Dimensionsreduktion sowohl für die akustischen als auch für die visuellen Daten wurden beide Möglichkeiten betrachtet, sowohl nur die Verwendung des ersten Lippenbildes als auch aller vier zum Zeitfenster gehörenden Lippenbilder. Diese Daten werden im folgenden als 'niedrigdimensional' oder 'LDA-reduziert' bezeichnet, da zur Dimensionsreduktion eine LDA verwendet wurde.
Bei der Klassifikation des Buchstabenwortes wurde das Zeitfenster immer um eine Zeitscheibe, d.h. um 10 ms weitergeschoben.
Für das Training des MLPs wird, wie schon erwähnt, das jeweils erste und letzte Zeitfenster eines Buchstabenwortes verwendet. Die Gewichte des Netzes werden zufällig im Intervall [-1;1] initialisiert und die Eingabedaten ebenfalls dementsprechend skaliert. Als Aktivierungsfunktion der Neuronen wird die logistische Funktion gemäß Gleichung 3.15 verwendet. Die Ausgabekodierung ist eine 1-aus-m-Kodierung. Um die Segmentklassifikation eines Zeitfensters zu bestimmen, wird die maximale Ausgabe des Netzes bestimmt, und das Ausgabeneuron mit der maximalen Ausgabe bestimmt die Segmentklassenzugehörigkeit. Die Anzahl der Ausgabeneuronen beträgt daher in jedem Fall 30.
Das Training des MLPs erfolgt mit dem Standard-Backpropagation-Algorithmus (vgl. 3.3.1). Erste Versuche mit Variationen des Backpropagation-Algorithmus (z.B. Verwendung des Momentum-Terms) zeigten eine starke Abhängigkeit der Klassifikationsleistung einzelner Architekturen von der Wahl der Parameter, was für das Standard-Backpropagation-Verfahren nicht im selben Umfang beobachtet wurde. Hier waren die Ergebnisse für unterschiedliche Architekturen bei gleicher Wahl der Lernrate immer zufriedenstellend, so daß folgen-

[3] Nach [Fur86] liegt die Länge des für eine Erkennung von CV-Silben notwendigen Signalstücks bei bis zu 100ms.

Lernrate η	Anzahl der Epochen
0.8	200
0.6	200
0.4	100
0.3	100
0.25	100
0.2	100
0.15	100
0.1	100

Tabelle 7.1: Lernraten für das Training

des Trainingsparadigma für alle Architekturen verwendet wurde: Das Training wird nach 1000 Epochen oder bei Erreichen eines Minimalfehlers abgebrochen, wobei die Lernrate schrittweise abgesenkt wird (vgl. Tabelle 7.1). Der minimale Fehler betrug 0.005 SSEPU (Sum-Squared-Error-per-Unit). Die verwendeten Feedforwardnetze wurden mit Hilfe des Stuttgarter-Neuronale-Netze-Simulators (SNNS) simuliert. Eine Beschreibung dieses Simulators, der ebenfalls im Rahmen des Landesforschungsschwerpunkts Baden-Württemberg zur Verfügung gestellt wurde, findet sich in [Zel94].
Als zu minimierende Energiefunktion wurde Gleichung 3.16 gewählt.
Bei der konkreten Realisierung des Assoziativspeichers wurde eine n-Gramm-Kodierung mit $n = 3$ verwendet. Diese wird auch als Tripelkodierung bezeichnet. Zusätzlich zu den 30 Segmentklassen, die in jeder zu speichernden Sequenz vorkommen können, wurde ein Anfangs- und Endzeichen #, welches an den Anfang und an das Ende jeder Sequenz angefügt wurde, zu den Segmentklassen hinzugenommen, so daß die Größe des Alphabets $L = 31$ beträgt. Dies geschieht, um evtl. Minderberücksichtigungen der ersten und letzten Segmente jeder Sequenz zu vermeiden. Für die Ausgabevektoren wurde eine 1-aus-m-Kodierung gewählt, bei der jede Komponente des Ausgabevektors für eine mögliche Ausgabeklasse steht. Die Klassenzugehörigkeit eines Musters wird also durch eine Eins an entsprechender Stelle des Ausgabevektors kodiert. Die Dimension des Ausgabevektors betrug in diesem Falle 25. Für das Lernen der Muster wurde eine geclippte Hebb-Lernregel verwendet (vgl. Gleichung 3.34), beim Retrieval wurde eine weiche Schwellwertstrategie mit Berechnung des Maximums verwendet (vgl. Gleichung 3.37). Zur endgültigen Entscheidungsfindung wurden als Ähnlichkeitsmaß sowohl der Hammingabstand als auch der Überlapp verwendet. Für die Simulationen wird eine Implementierung des Assoziativspeichers von A. Wichert verwendet.

7.4 Fusionsarchitekturen

Im Rahmen des vorgestellten Worterkennungssystems können verschiedene Fusionsmöglichkeiten untersucht werden. Bevor dies geschieht, ist es allerdings notwendig, eine für die Klassifikation der Daten der Einzelkanäle günstige Netz-

topologie des jeweiligen Feedforwardnetzes zu bestimmen, genauergesagt, die Anzahl der Neuronen der versteckten Schicht zu ermitteln. Dazu werden zunächst Versuche mit beiden Kanälen separat durchgeführt, aus denen die optimale Anzahl an versteckten Neuronen für jeden Kanal einzeln bestimmt wird. Die Anzahl der versteckten Neuronen der Kombinationsarchitekturen ist dann die Summe der Neuronenanzahlen der einzelnen Kanäle. Durch parallele Versuche mit vollverbundenen Netztopologien, die die fusionierten akustischen und visuellen Daten verarbeiten, läßt sich feststellen, ob diese ermittelte Anzahl an versteckten Neuronen für die Fusionsarchitekturen sinnvoll sind. Gegebenfalls können die Ergebnisse dieser Experiment auch dazu genutzt werden, bei gleichwertigen Netztopologien der Einzelkanäle zwischen diesen eine Entscheidung zu fällen.
Sobald die Anzahl der versteckten Neuronen ermittelt wurde, lassen sich alle anderen Fusionsarchitekturen darauf aufbauen. Eine Übersicht mit den verwendeten Bezeichnungen der einzelnen Architekturen ist der Abbildung 7.3 zu entnehmen.
Als frühe Fusionsarchitekturen sollen Architekturen bezeichnet werden, bei denen die akustischen und die visuellen Daten zusammengeführt werden, bevor sie in den Klassifikator gegeben werden. Die Fusion geschieht also schon auf der Ebene der Eingabeschicht des Netzes, und im folgenden sollen diese Architekturen mit I1 (Integration auf '**I**nputlayer', **eine** versteckte Schicht) bezeichnet werden.
Wird die Fusion innerhalb des Klassifikators vorgenommen, bevor eine Klassifikation in einer wie auch immer gearteten Art und Weise geschieht, wird dies als Fusion auf einer Zwischenstufe bezeichnet. Die Fusion geschieht dabei auf der Ebene der versteckten Schicht und soll im folgenden mit H1 (Integration auf dem '**H**iddenlayer', **eine** versteckte Schicht) bezeichnet werden.
Weiterhin ist es möglich, die Verarbeitung der akustischen und visuellen Daten getrennt zu halten, bis beide unabhängig voneinander als Segmentklassifikationen klassifiziert wurden. Zur Fusion wurden allerdings nicht die Segmentklassenzugehörigkeiten verwendet, sondern die Ausgaben der Einzelnetze. Realisiert wurde die Fusion mithilfe eines einschichtigen Netzes, welches wiederum trainiert wurde. Diese Fusionsmöglichkeit ist eine späte Fusion und soll im folgenden mit C (Integration nach '**C**lassification') bezeichnet werden.
Betrachtet man diese Fusionsmöglichkeit als ein einziges neuronales Netz, welches aus drei verschiedenen, getrennt trainierten Netzen zusammengesetzt ist, so liegt es nahe, zum Vergleich auch ein weiteres Netz hinzuzuziehen, welches dieselbe Netztopologie aufweist, allerdings als ein Ganzes trainiert werden soll. Die beiden Ausgabeschichten der Einzelnetze bei Architektur C bilden somit eine weitere versteckte Schicht, deren Neuronenanzahl 60 ist. Diese Fusionsarchitektur weist ebenfalls einen Zusammenschluß der beiden Information auf der Ebene der versteckten Schichten auf und wird demzufolge als H2 bezeichnet (Integration auf dem '**H**iddenlayer', **zwei** versteckte Schichten).
Im Vergleich zu dieser Architektur wird auch eine weitere frühe Fusionsmöglichkeit mit zwei versteckten Schichten untersucht, die mit I2 (Integration auf dem '**I**nputlayer', **zwei** versteckte Schichten) bezeichnet werden soll.
Als letzte verbleibende Möglichkeit bleibt eine Fusion, nachdem sowohl die aku-

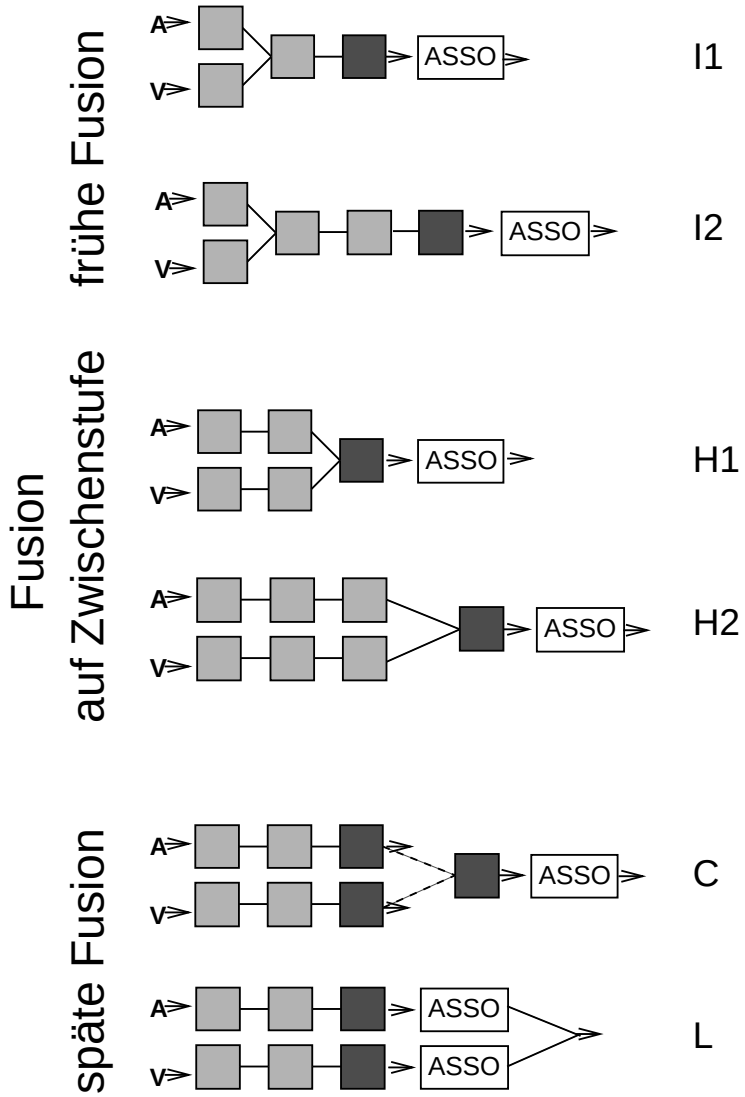

Abbildung 7.3: Übersicht über den prinzipiellen Aufbau der untersuchten Fusionsarchitekturen. Die Schichten der Feedforwardnetze werden als Quadrate dargestellt, die Ausgabeschichten sind dunkler gezeichnet.

stischen als auch die visuellen Daten einzeln vom Worterkennungssystem klassifiziert worden sind. Bei dieser Fusionsmöglichkeit werden also 'binäre' Klassifikationen zusammengeführt, im Gegensatz zu den kontinuierlichen Ausgabewerten bei Architektur C. Diese Fusionsarchitektur soll mit L gekennzeichnet werden (Integration auf der letzten Stufe). Für diese Fusion ergeben sich folgende Möglichkeiten:

- Beide Kanäle entscheiden sich für dieselbe Klasse, dies ergibt insgesamt eine eindeutige Entscheidung.
- Für beide Kanäle ergibt sich keine eindeutige Klassifikation. Besteht in diesem Fall die Schnittmenge beider Klassifikationen aus einem Element, kann eine dementsprechende Entscheidung getroffen werden. Ist die Schnittmenge leer oder hat mehr als ein Element, kann keine Entscheidung getroffen werden.
- Für einen der beiden Kanäle ergab sich eine eindeutige Entscheidung, die Entscheidung des anderen Kanals war mehrdeutig. In diesem Fall kann die eindeutige Entscheidung des einen Kanals übernommen werden (die evtl. auch in der Schnittmenge beider Entscheidungsmengen liegt).
- Beide Kanäle liefern eine eindeutige Entscheidung, die allerdings unterschiedlich ist. In diesem Fall stellt sich die Frage, wie beide Kanäle relativ zueinander gewichtet werden sollen. Eine Möglichkeit besteht darin, daß in diesem Fall die Entscheidung des akustischen Kanals übernommen wird, da im rauschfreien Fall der akustische Kanal im allgemeinen die genauere Klassifikation liefert. Dies ist bei verrauschtem akustischem Kanal nicht unbedingt gegeben.
  In der Literatur (vgl. Kapitel 5) werden unterschiedliche Möglichkeiten genutzt, um dieses Problem zu lösen. Meist wird dazu das SNR zu Hilfe genommen, um die Gewichtung beider Kanäle zu bestimmen. Im vorliegenden Fall gibt es eine andere Möglichkeit, die das vom Assoziativspeicher berechnete Ähnlichkeitsmaß benutzt. Dieses Ähnlichkeitsmaß ist für jeden Kanal gegeben und soll im folgenden mit h_{ak} für den akustischen Kanal und mit h_{vis} für den visuellen Kanal bezeichnet werden. Falls als Ähnlichkeitsmaß z.B. der Hammingabstand gewählt wird, kann die Entscheidung danach gefällt werden, wessen Hammingabstand kleiner ist:

$$\begin{array}{lcl} h_{ak} \leq h_{vis} & \Rightarrow & C_{BSTW} = c_{ak} \\ h_{ak} > h_{vis} & \Rightarrow & C_{BSTW} = c_{vis}. \end{array} \tag{7.9}$$

  Ebenfalls ist eine Wichtung der einzelnen Kanäle möglich, indem eine Schwelle b zum Hammingabstand des visuellen Kanals addiert wird, so daß die Klassifikation des akustischen Kanals bevorzugt als Gesamtklassifikation übernommen wird, auch wenn der Hammingabstand des akustischen Kanals größer als der des visuellen Kanals ist:

$$\begin{array}{lcl} h_{ak} \leq h_{vis} + b & \Rightarrow & C_{BSTW} = c_{ak} \\ h_{ak} > h_{vis} + b & \Rightarrow & C_{BSTW} = c_{vis}. \end{array} \tag{7.10}$$

Je größer der Wert für b eingestellt wird, umso mehr wird der akustische Kanal gewichtet. Der Wert für b wurde experimentell variiert.

Kapitel 8

Fusionsresultate für Originalmerkmale

Im folgenden Kapitel sollen die Versuchsergebnisse dargestellt werden, die mit den Originaldaten erzielt wurden (siehe auch [TKP97]). Dabei wird die Performanz der Architekturen auf der Ebene der Buchstabenwortklassifikation betrachtet. Soweit nicht anders gesagt, werden immer die Ergebnisse auf dem Testdatensatz dargestellt, wobei als Fehlerkriterium das klassenspezifische Mittel gewählt wird. Dargestellt ist jeweils der Mittelwert über die 25 Crossvalidierungsversuche, die für jede Architektur durchgeführt wurden, sowie das 95%ige Konfidenzintervall. Alle Ergebnisse, d.h. die Performanz der Netze auf dem Trainings- und Testsatz in Form der nichtgewichteten Trainings- und Testfehler sowie der klassenspezifischen Mittel, als auch die Klassifikationsleistung auf der Segmentklassifikationsebene können dem Anhang A entnommen werden.
Ebenfalls im Anhang A stehen die Ergebnisse, die bei der Verwendung unterschiedlicher Abstandsmaße als Entscheidungskriterium bei der Rückprojektion des Assoziativspeichers erzielt wurden. Untersucht wurde in diesem Zusammenhang die Verwendung des Hammingabstandes sowie des Überlapps.
Als Ergebnis für die Architektur L wird immer das beste erzielte Ergebnis angegeben. Die unterschiedlichen Fusionsmöglichkeiten bei Architektur L sind dem Abschnitt 8.5.3 zu entnehmen.

8.1 Optimierung der Netztopologien

Da es keine theoretischen Aussagen darüber gibt, welche Topologie ein Feedforward-Netz bei gegebenen Daten haben sollte, um das bestmögliche Klassifikationsergebnis zu erzielen [Rip96], muß diese experimentell ermittelt werden. Im vorliegenden Fall steht die Dimension der Eingabe- und Ausgabeschicht fest: Die Eingabeschicht hat im Fall des akustischen Netzes die Dimension 160, da ein 100 msec langes Zeitfenster benutzt wird. Dieses entspricht 10 Zeitscheiben des Spektrogramms und für jede Zeitscheibe werden die Intensitäten für 16 Frequenzbänder bestimmt, so daß sich 160 Merkmale für jedes akustische Eingabemuster ergeben. Für den visuellen Kanal ergeben sich 140 Merkmale für jedes Eingabemuster (vgl. Abschnitt 7.3). Da jedes Zeitfenster auf eine von

Netz (Anzahl der Gewichte)	Architektur	KSPM (test)
A (1900)	160-10-30	61.35 (5.50)
A (3800)	160-20-30	80.70 (1.63)
A (5700)	160-30-30	82.19 (1.49)
A (7600)	160-40-30	81.64 (1.87)

Tabelle 8.1: Ergebnisse für Netztopologien mit unterschiedlicher Anzahl an versteckten Neuronen für die akustischen Daten. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und in Klammern das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

Netz (Anzahl der Gewichte)	Architektur	KSPM (test)
V (1700)	140-10-30	32.04 (1.69)
V (3400)	140-20-30	35.69 (2.02)
V (5100)	140-30-30	33.15 (2.36)
V (6800)	140-40-30	32.15 (2.44)

Tabelle 8.2: Ergebnisse für Netztopologien mit unterschiedlicher Anzahl an versteckten Neuronen für die visuellen Daten. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und in Klammern das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

30 Segmentklassen abgebildet wird und eine 1-aus-n-Kodierung für die Ausgabeschicht gewählt wird, enthält jede Ausgabeschicht 30 Neuronen. Die optimale Anzahl der versteckten Neuronen soll experimentell bestimmt werden. Dazu werden sowohl für die akustischen als auch für die visuellen Daten zweischichtige Feedforwardnetze mit einer unterschiedlichen Anzahl an versteckten Neuronen getestet. Die Ergebnisse können den Tabellen 8.1 und 8.2 entnommen werden. Da stets die Verwendung des Hammingabstandes als Entscheidungskriterium des Assoziativspeichers die besten Ergebnisse liefert (vgl. Tabellen A.1 und A.3 in Anhang A), sind nur diese Ergebnisse aufgelistet.

Obwohl die Performanzunterschiede der Architekturen (bis auf Architektur 160-10-30 im akustischen Fall) nicht statistisch signifikant sind, kann anhand des Mittelwertes eine Entscheidung gefällt werden. Im visuellen Fall erhält man mit der Architektur 140-20-30 das beste Ergebnis, welches um etwa 2.5 % besser als die Ergebnisse der anderen Architekturen ist. Auch kann festgestellt werden, daß die Performanz mit einer weiter steigender Anzahl an versteckten Neuronen fällt. Geringere Unterschiede sind im akustischen Fall zu erkennen, hier unterscheidet sich die Performanz der Architekturen 160-30-30 und 160-40-30 nur

Netz (Anzahl der Gewichte)	Architektur	KSPM (test)
AV (6600)	300-20-30	67.50 (4.21)
AV (9900)	300-30-30	61.93 (6.24)
AV (13200)	300-40-30	63.44 (6.11)
AV (16500)	300-50-30	64.17 (4.95)
AV (19800)	300-60-30	65.03 (5.02)
AV (23100)	300-70-30	61.74 (4.34)
AV (26400)	300-80-30	62.64 (4.48)

Tabelle 8.3: Ergebnisse für Netztopologien mit unterschiedlicher Anzahl an versteckten Neuronen für die akustischen und die visuellen Daten. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und in Klammern das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

um etwa ein halbes Prozent. Um zwischen diesen Architekturen auswählen zu können, sollen die Ergebnisse für Netztopologien zu Hilfe genommen werden, die mit den akustischen und den visuellen Daten zusammen trainiert wurden und somit die Fusionsarchitektur I1 realisieren. Die Ergebnisse können der Tabelle 8.3 entnommen werden.

Es zeigt sich, daß auch hier keine statistisch signifikanten Unterschiede zwischen den Architekturen festgestellt werden können. Man kann aber feststellen, daß nach der Topologie mit nur 20 versteckten Neuronen die Topologie mit 60 versteckten Neuronen die beste Performanz bietet, mit einer Differenz von etwa 0.9 % gefolgt von der Architektur mit 50 versteckten Neuronen. Bei der Wahl der Architektur mit 20 versteckten Neuronen als Fusionsarchitektur müßten entsprechend für die Einzelkanäle jeweils 10 versteckte Neuronen gewählt werden, die aber auf den Einzelkanälen jeweils die schlechteste Performanz zeigen, und deshalb nicht verwendet werden sollen. Setzt man für den visuellen Kanal die Anzahl an versteckten Neuronen mit 20 an, so kann man aufgrund der Ergebnisse mit den fusionierten Daten für den akustischen Kanal eine versteckten Neuronenanzahl von 40 festlegen.

8.2 Fusionsarchitekturen

Auf der Grundlage der optimierten Einzelnetze können nun daraus die in Abbildung 8.1 dargestellten Fusionsarchitekturen abgeleitet und untersucht werden (vgl. Kapitel 7.4).

In Abbildung 8.2 sind die Ergebnisse für ausgewählte Architekturen als Mittelwerte des klassenspezifischen Mittels dargestellt. Die Fehlerbalken bezeichnen die 95%igen Konfidenzintervalle. Alle Ergebnisse können der Tabelle A.2 des Anhangs A entnommen werden.

Zunächst kann festgestellt werden, daß die Hinzunahme des visuellen Kanals in

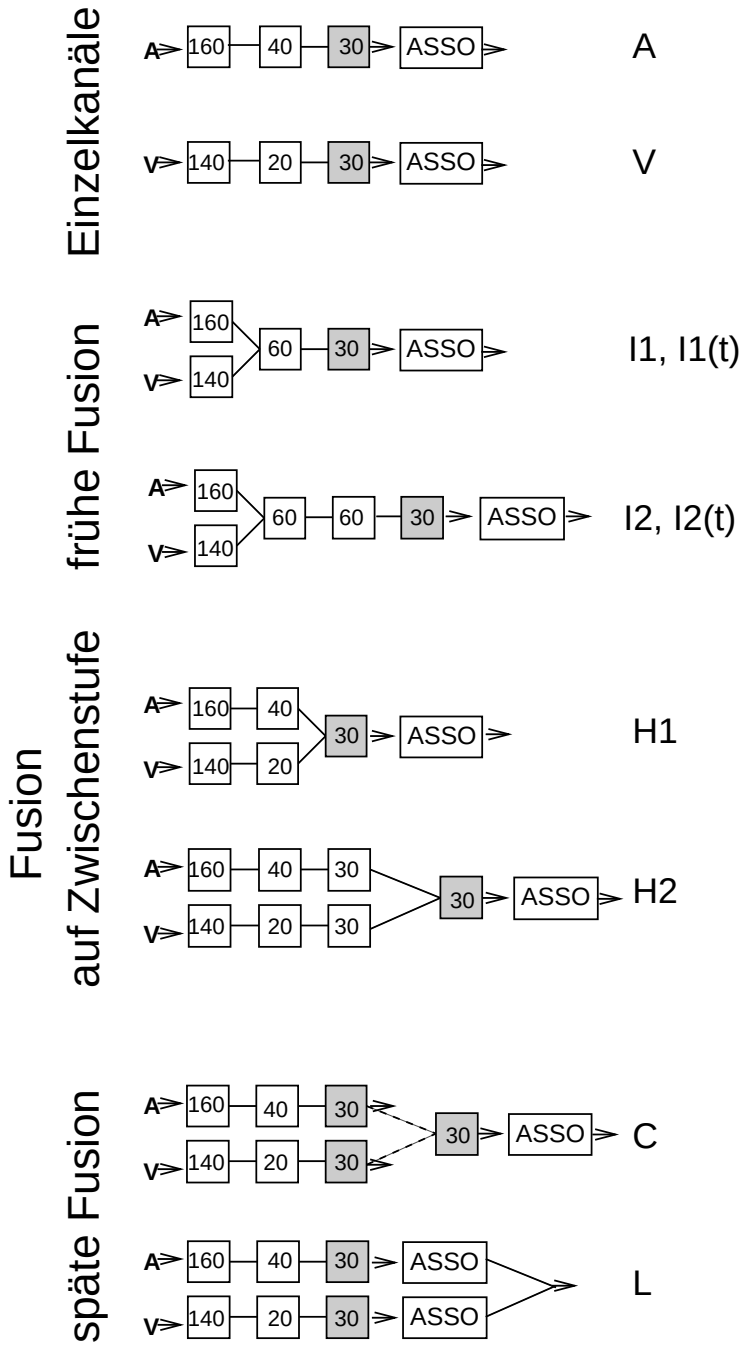

Abbildung 8.1: Übersicht über die für die hochdimensionalen Daten verwendeten Fusionsarchitekturen. Die Ziffern geben die Anzahl der Neuronen in der jeweiligen Schicht an.

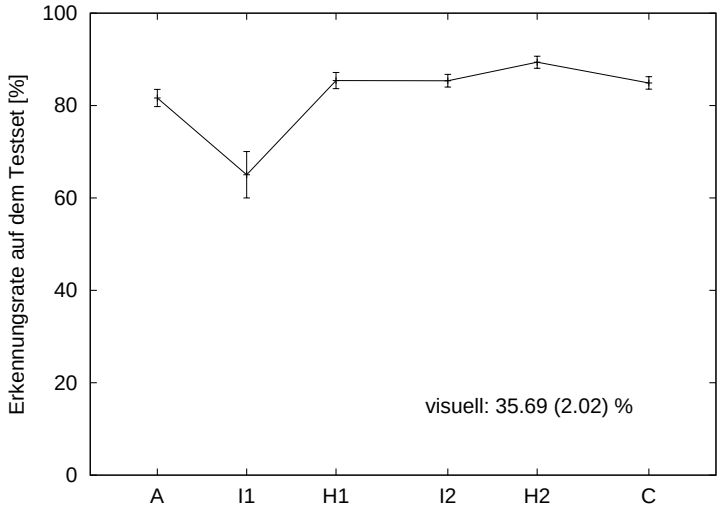

Abbildung 8.2: Fusionsergebnisse für ungestörte hochdimensionale Daten dargestellt als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

Architektur	Anzahl der Gewichte	Anzahl der Gewichte
	normal	akustisch modular
A	7600	1840
V	3400	3400
I1	19800	19800
I1(t)	11000	5240
H1	11000	5240
I2	23400	23400
I2(t)	12800	7040
H2	12800	7040
C	7600 + 3400 + 1800	1840 + 3400 + 1800
L	7600 + 3400	1840 + 3400

Tabelle 8.4: Übersicht über die verwendeten Fusionsarchitekturen und die Anzahl ihrer Gewichte

den meisten Fällen eine Verbesserung der Klassifikationsleistung im Vergleich zu der Leistung des akustischen Kanals alleine bringt. Als beste Architektur ergibt sich die Architektur H2, welche eine Fusion auf der Zwischenstufe realisiert und zwei versteckte Schichten besitzt. Der Performanzunterschied zu den anderen Architekturen ist statisch signifikant.
Auffällig ist das schlechte Abschneiden der Architektur I1, deren Performanz noch unter der Performanz des akustischen Kanals liegt. Gleichzeitig fällt auf, daß das Konfidenzintervall sehr groß ist, was seine Ursache unter Umständen darin haben könnte, daß die Netze noch nicht austrainiert waren. Ein weiteres Training der Netze zeigt aber, daß sich die Performanz nicht weiter bessert. Als weitere Ursache läßt sich vermuten, daß die Anzahl der Gewichte der Architektur zu groß ist (vgl. Tabelle 8.4). Um diese Vermutung zu überprüfen, sollen weitere Experimente durchgeführt werden, in denen die Anzahl der Gewichte einer Architektur auf verschiedene Art und Weise variiert werden.

8.3 Modulare Architekturen

Im Falle der akustischen Daten lassen sich unterschiedliche Modularisierungsarchitekturen untersuchen. So ist es möglich, daß jede einzelne Zeitscheibe des Zeitfensters nur von separaten Neuronen der versteckten Schicht verarbeitet wird, erst nach der versteckten Schicht findet eine Zusammenführung aller Zeitscheiben statt (vgl. Abbildung 8.3). Dies entspricht einer Fusionsarchitektur auf der Zwischenstufe, nur daß hier nicht unterschiedliche Informationskanäle verwendet werden, sondern verschiedene Zeitscheiben als Eingaben. Ebenso kann die Modularisierung hinsichtlich der Frequenzkanäle im Zeitfenster stattfinden, wie es in Abbildung 8.4 gezeigt wird.
Als Topologie für diese Versuche wurde das Netz 160-30-30 ausgewählt, da sich hierfür beide Modularisierungsarten mit ähnlicher Gewichtsanzahl und Anzahl der versteckten Neuronen realisieren lassen: Für die Modularisierung anhand der Zeitscheiben werden die jeweils zu einer Zeitscheibe gehörenden 16 Komponenten des Eingabevektors mit jeweils 3 Neuronen der versteckten Schicht verbunden, dies resultiert in einer 160-30-30 Architektur mit 1380 Gewichten, die mit 160[mod10]-30-30 bezeichnet werden soll. Bei der Modularisierung hinsichtlich der Frequenzkanäle werden die jeweils zu einem Frequenzkanal gehörenden 10 Komponenten des Eingabevektors mit jeweils 2 Neuronen der versteckten Schicht verbunden, was in einer 160-32-30 Architektur mit 1280 Gewichten resultiert. Diese Architektur soll mit 160[mod16]-30-30 bezeichnet werden. Zusätzlich wird eine 160-30-30 Architektur untersucht, deren Gewichte zufällig auf einen Wert von 1380 ausgedünnt wurden (bezeichnet mit 160[ranmod]-30-30). Die Ergebnisse dieser Versuche können der Tabelle 8.5 entnommen werden. Die Performanzunterschiede zwischen allen Architekturen sind nicht statistisch signifikant, beide Architekturen mit 30 versteckten Neuronen liefern nahezu gleiche Ergebnisse, die Modularisierungsarchitektur bezüglich der Frequenzkanäle ist um etwa 2 % schlechter, besitzt allerdings auch weniger Gewichte und mehr versteckte Neuronen. Im Vergleich zu dem nichtausgedünnten 160-30-30 Netz kann festgestellt werden, daß alle Modularisierungsarchitekturen besser sind,

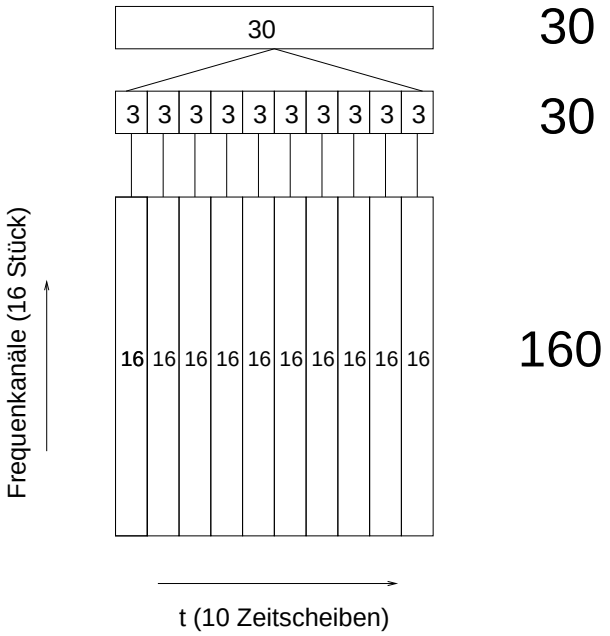

Abbildung 8.3: Modularisierungsarchitektur für den akustischen Kanal. Die Modularisierung wird für die einzelnen Zeitscheiben durchgeführt.

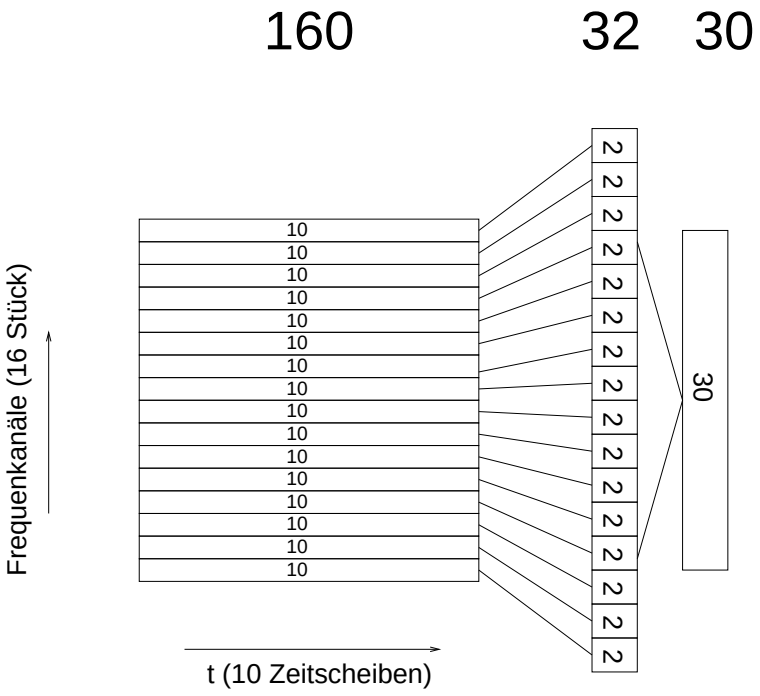

Abbildung 8.4: Modularisierungsarchitektur für den akustischen Kanal. Die Modularisierung wird für die einzelnen Frequenzkanäle durchgeführt.

Netz (Anzahl der Gewichte)	Architektur	KSPM (test)
A (5700)	160-30-30	82.19 (1.49)
A (1380)	160[mod10]-30-30	84.99 (1.53)
A (1380)	160[ranmod10]-30-30	85.05 (1.40)
A (1280)	160[mod16]-32-30	83.02 (1.09)
A (7600)	160-40-30	81.64 (1.87)
A (1840)	160[mod10]-40-30	86.32 (1.04)

Tabelle 8.5: Ergebnisse für Netztopologien mit unterschiedlicher Modularisierungsart für die akustischen Daten. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und in Klammern das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

wenngleich auch diese Unterschiede nicht statistisch signifikant sind.
Versuche mit einem modularisierten 160-40-30 Netz, welches gemäß der Modularisierung gemäß der Zeitscheiben ausgedünnt wurde, zeigen hingegen eine statistisch signifikante Verbesserung der Performanz und insgesamt die beste Klassifizierungsleistung, wie Tabelle 8.5 zu entnehmen ist.

8.4 Fusion mit modularen Architekturen

In einer neuen Versuchsreihe werden zusätzlich zu den bisher untersuchten Fusionsarchitekturen (vgl. Abschnitt 8.3) auch ausgedünnte Netze betrachtet, bei denen bei den frühen Fusionsarchitekturen I1 und I2 zufällig soviele Gewichte entfernt werden, bis deren Anzahl gleich jener der entsprechenden Fusionsarchitektur auf der Zwischenstufe H1 bzw. H2 ist. Dies geschieht, um auszuschließen, daß die beste Performanz der Architektur H2 aufgrund der geringeren Anzahl der Gewichte zustande kommt. Die entsprechenden Architekturen sollen H1(t) und H2(t) genannt werden.
In einer weiteren Versuchsreihe werden auch Architekturen H1 und H2 betrachtet, bei denen der 'akustische' Teil des Netzes modular ausgedünnt wird, so daß die resultierende Architektur von vorneherein weniger Gewichte hat. (vgl. Abb. 8.5). Dazu wird die Modularisierung anhand der Zeitscheiben gewählt, da diese im SNNS sehr leicht implementierbar ist und für den Fall mit 30 versteckten Neuronen zusammen mit der zufälligen Ausdünnung die besten Ergebnisse erzielt. Zusätzlich zu dieser Architektur werden auch die entsprechend zufällig ausgedünnten H1(t) und H2(t)-Architekturen untersucht, die dieselbe Anzahl an Gewichten, wie die akustisch modularisierten H1 und H2-Architekturen haben. Es bleibt anzumerken, daß die vollverbundenen frühen Architekturen I1 und I2 für beide Versuchsreihen diesselben sind. Alle anderen Architekturen unterscheiden sich. Ein Überblick über die Anzahl der Gewichte der einzelnen Architekturen ist der Tabelle 8.4 zu entnehmen.

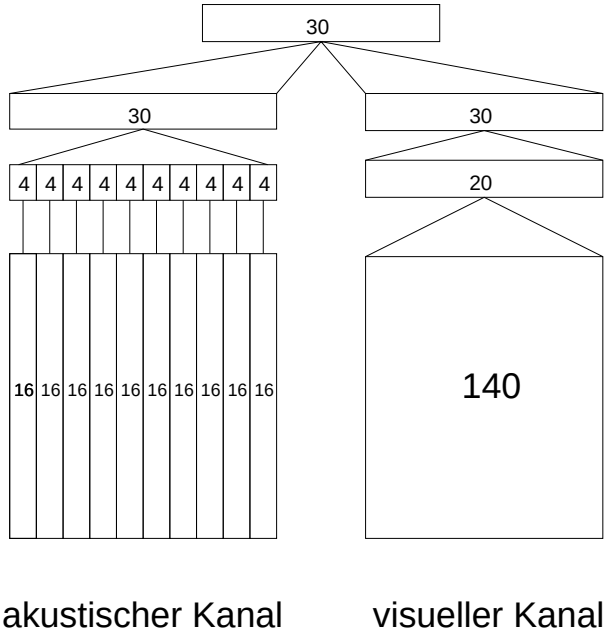

Abbildung 8.5: Darstellung der Architektur H2, bei der der akustische Teil des Netzes modular ausgedünnt wird.

Die entsprechenden Versuchsergebnisse können der Abbildung 8.6 entnommen werden. Wiederum wird als Entscheidungskriterium des Assoziativspeichers der Hammingabstand gewählt, da damit die bessere Performanz erzielt werden kann (vgl. Tabelle A.5, A). Zunächst ist erkennbar, daß die auf dem akustischen Kanal weiter ausgedünnten Architekturen die besseren Ergebnisse liefern, außer für die Architekturen mit zwei versteckten Schichten sind diese Unterschiede statistisch signifikant. Man erkennt, daß jeweils die ausgedünnten Architekturen I1(t) und I2(t) in etwa eine gleich gute Performanz wie die Architekturen H1 und H2 liefern, der Unterschied kommt also durch die größere Anzahl an Gewichten zustande. Desweiteren läßt sich feststellen, daß für die auf akustischer Seite noch weiter modularisierten Netze die Unterschiede der Performanz zwischen den Netzen mit einer und zwei versteckten Schichten nicht mehr wahrnehmbar sind. Die Architekturen C und L mit einer späten Fusion liefern die schlechtesten Ergebnisse.
Da durch die Modularisierung eine Verbesserung der Performanz erzielt werden konnte, sollen im nächsten Kapitel, in dem auch verrauschte akustische Daten betrachtet werden, nur die Architekturen mit modularem akustischen Anteil betrachtet werden.

8.5 Experimente mit gestörtem akustischen Kanal

Zusätzlich zu den ungestörten Daten wurden auch Daten mit verrauschtem akustischen Kanal verwendet (vgl. Abschnitt 7.1), um die Performanz der verschiedenen Fusionsarchitekturen zu testen. Es wurden Versuche mit drei unter-

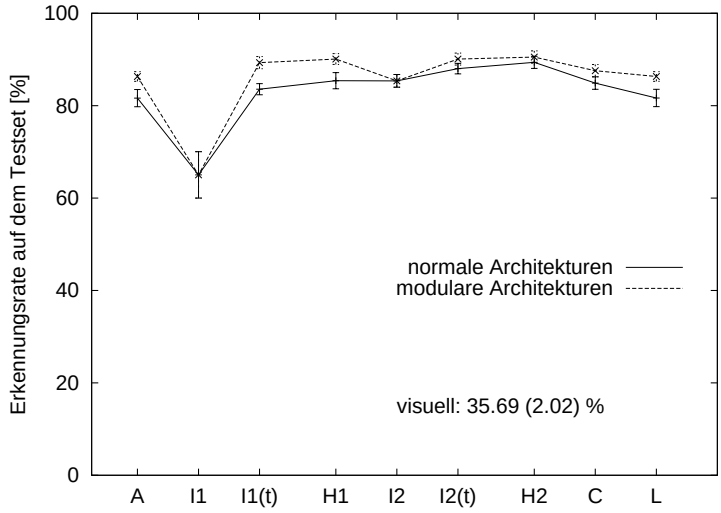

Abbildung 8.6: Ergebnisse für zwei Versuchsreihen mit hochdimensionalen Daten: vollständige Architekturen, sowie mit auf akustischer Seite modularisierten Architekturen (vgl. Text). Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und als Fehlerbalken das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

schiedlichen Signal-zu-Rausch-Verhältnissen durchgeführt.

8.5.1 Experimente mit akustischen Daten

In ersten Experimenten werden zunächst nur die akustischen Daten verwendet. Ein modular ausgedünntes Netz wird dazu mit jeweils einem der vier Datensätze (nicht gestörte akustische Daten, SNR = 3 dB, -3 dB, -11 dB) trainiert und mit den jeweils anderen drei Datensätzen getestet.

Die Ergebnisse sind der Abbildung 8.7 zu entnehmen. Wiederum wird als Entscheidungskriterium des Assoziativspeichers der Hammingabstand gewählt, da damit die bessere Performanz erzielt werden kann (vgl. Tabelle A.10, Anhang A). Zunächst läßt sich feststellen, daß die Netze jeweils dann die beste Erkennungsleistung zeigen, wenn sie mit der Rauschstufe getestet wurden, mit der sie auch trainiert wurden. Für die einzelnen Rauschstufen ergibt sich aber ein unterschiedliches Bild. Wird nur mit den unverrauschten Daten trainiert, aber mit gestörten Daten getestet, sinkt die Erkennungsleistung dramatisch ab. Wird mit verrauschten Daten trainiert, sind die Unterschiede nicht so gravierend. Wird z.B. mit Daten mit einem SNR von -3 dB trainiert, ist die Erkennungsleistung für die Daten mit einem SNR von 3 dB nicht signifikant schlechter, dies gilt auch für das Training mit einem SNR von -11 dB und dem Testen bei -3 dB. Die Erkennungsleistungen für die mit -11 dB verrauschten Daten sind sogar nahezu gleich, gleichgültig, ob mit demselben SNR von -11 dB oder nur mit einem SNR von -3 dB trainiert wurde. Wird mit -3 dB oder mit -11 dB trainiert, erge-

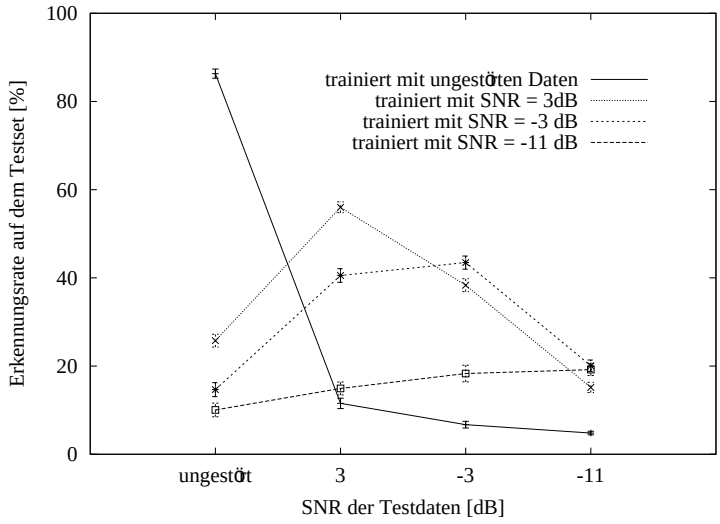

Abbildung 8.7: Versuche mit verrauschten akustischen Daten, wobei zum Training und Testen unterschiedlich gestörte Daten verwendet werden. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und als Fehlerbalken das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

ben sich die schlechtesten Performanzen für das Testen mit den unverrauschten Daten, bei einem Training mit Daten mit einem SNR von 3 dB sind die Testergebnisse für die unverrauschten Daten wenig besser als diejenigen für die mit -11 dB verrauschten Daten.
Insgesamt läßt sich feststellen, daß mit gestörten Daten trainiert werden muß, wenn auch mit gestörten Daten getestet werden soll. Dabei sollte das SNR von Trainings- und Testdaten möglichst gleich groß sein.

8.5.2 Fusionsarchitekturen mit gestörtem akustischen Kanal

Aufgrund der Ergebnisse des vorangegangenen Kapitels werden für das Training der Fusionsarchitekturen diesselben Rauschstufen verwendet, die auch für das Testen genommen werden. Wie schon zuvor beschrieben, werden in dieser Versuchsreihe wiederum diejenigen Fusionsarchitekturen untersucht, bei denen der akustische Teil der Netze modular ausgedünnt worden ist, außerdem liefert die Verwendung des Hammingabstandes als Entscheidungskriterium des Assoziativspeichers wiederum die beste Performanz (vgl. Tabellen A.6, A.7 und A.8 im Anhang A). Die Ergebnisse können der Abbildung 8.8 entnommen werden.

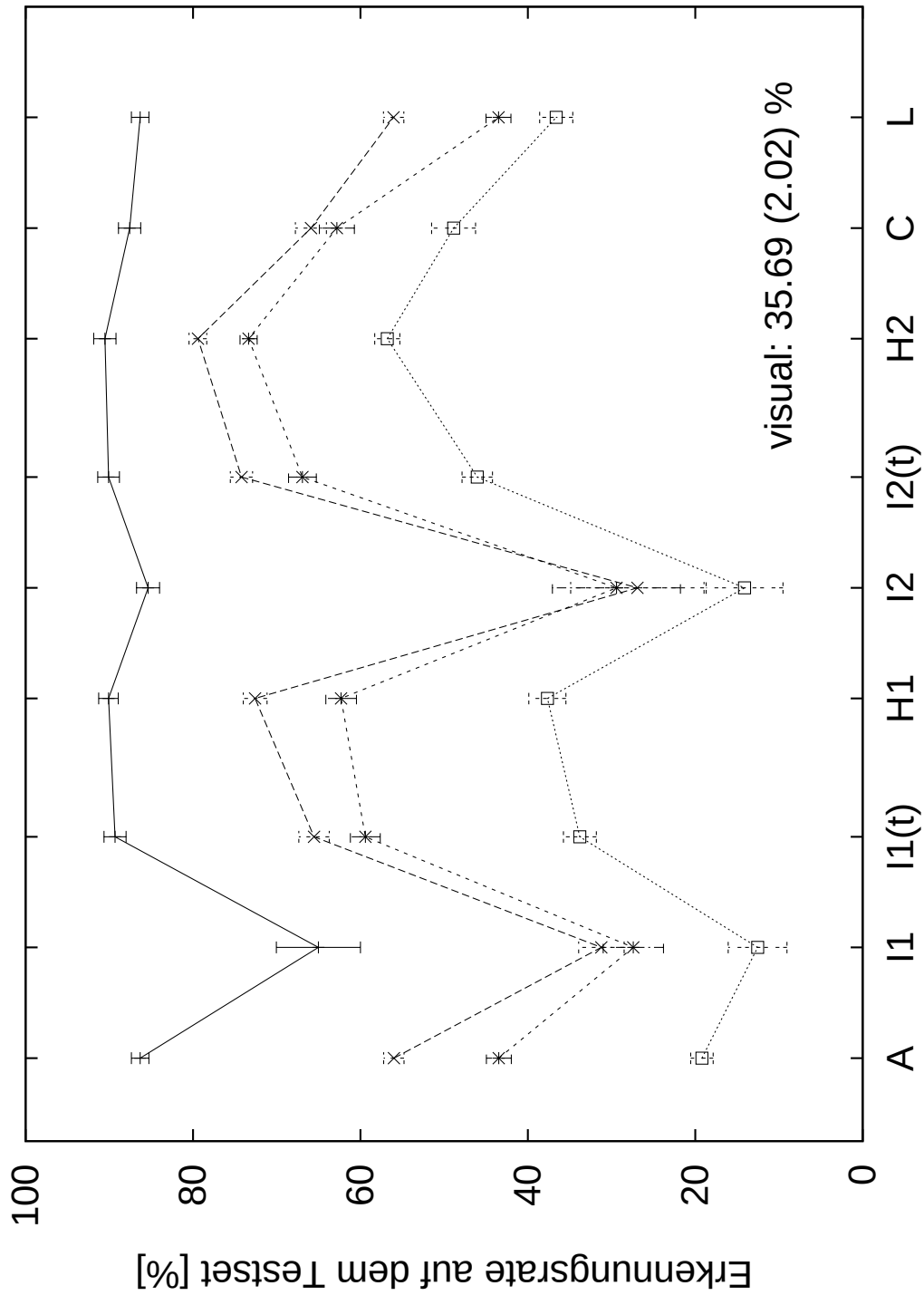

Abbildung 8.8: Ergebnisse mit den hochdimensionalen Daten bei Verwendung unterschiedlich gestörter akustischer Daten. Es werden modulare Fusionsarchitekturen verwendet. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und als Fehlerbalken das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

Als erstes ist zu erkennen, daß auch für kleinere Signal-Rausch-Verhältnisse die Architekturen I1 und I2 zu viele Gewichte besitzen, da ihre Performanz sehr schlecht ist und auch die Konfidenzintervalle entsprechend groß sind. Die zufällig ausgedünnten Architekturen I1(t) und I2(t) liefern deutlich bessere Ergebnisse, die allerdings zumindest im Fall eines SNR=3 dB und ansonsten für zwei versteckte Schichten immer noch signifikant schlechter als die Ergebnisse für die Architekturen H1 bzw. H2 sind.
Weiterhin läßt sich feststellen, daß die Performanz der Fusionsnetze beim Einsatz eines verrauschten akustischen Kanals immer größer als die für den akustischen Kanal alleine ist, d.h. die Hinzunahme der visuellen Information verbessert das Klassifikationsergebnis. Ausnahmen hierfür sind wie zuvor die Architekturen I1 und I2 sowie die Architektur L mit einer späten Fusion.
Falls kein verrauschter akustischer Kanal verwendet wird, ist die Performanz der Fusionsarchitekturen ebenfalls besser als die des akustischen Kanals alleine, ausgenommen die späten Fusionsarchitekturen C und L, die in etwa eine gleich gute Performanz bieten. Statistisch signifikante Unterschiede ergeben sich zwischen der Architektur H2 und den späten Architekturen C und L.
Für den Fall eines verrauschten akustischen Kanals ist für die Architekturen mit zwei versteckten Schichten die Architektur H2 statistisch signifikant besser als alle anderen Architekturen, für die Architekturen mit nur einer versteckten Schicht ist dies nur für ein SNR=3 dB gültig.
Insgesamt läßt sich allerdings sagen, daß die Architekturen mit einer Fusion auf der Zwischenstufe in beiden Fällen die besten Ergebnisse liefern, wobei Architektur H2 statistisch signifikant bessere Performanzen als Architektur H1 liefert. Ausnahme hiervon ist die Performanz für die ungestörten akustischen Daten.
Bei allen Verrauschungstufen erweisen sich die späten Fusionsarchitekturen als die schlechtesten, wobei die Architektur C im Falle eines verrauschten akustischen Kanals immer noch eine statistisch signifikant bessere Leistung als Architektur L erbringt.
Weiterhin läßt sich feststellen, daß bei einem kleinen SNR von -11 dB die Fusionsarchitekturen mit nur einer versteckten Schicht sowie die Architekturen L und I2 nicht besser sind, als ein Netz, daß nur die visuellen Daten allein verarbeitet.
Zusammenfassend läßt sich zwischen den Architekturen folgende Rangordnung aufstellen: Die beste Performanz erbringt Architektur H gefolgt von der gewichtsreduzierten Architektur I(t). Diese ist besser als Architektur C, für kleines SNR sind beide in etwa gleich gut. Architektur C ist besser als Architektur L, nur für unverrauschte Daten ist der Unterschied nicht statistisch signifikant.

8.5.3 Ergebnisse für die Architektur L

Wie schon in Abschnitt 7.4 erläutert, können verschiedene Arten der Fusion auf Ebene der Buchstabenworterkennung untersucht werden, bei der die jeweiligen Buchstabenwortklassifikationen der einzelnen Kanäle miteinander fusioniert werden. Der in Abschnitt 7.4 eingeführte Parameter b dient dazu, den Einfluß des akustischen Kanals unterschiedlich gewichten zu können; je größer

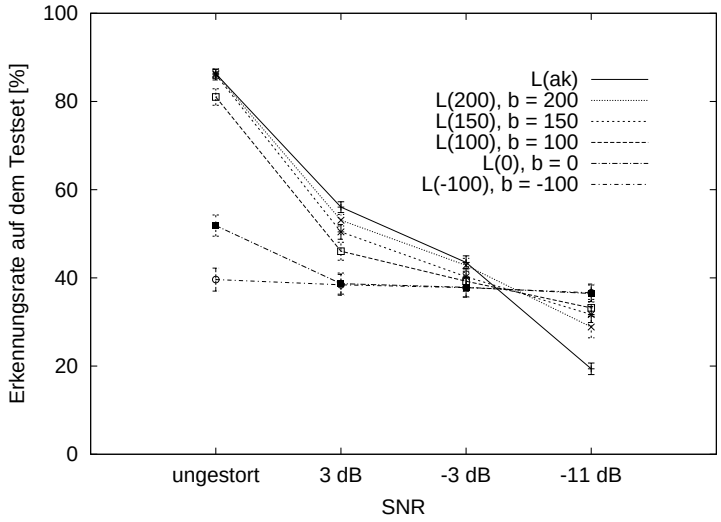

Abbildung 8.9: Ergebnisse mit hochdimensionalen Daten, verschiedene Realisierungen der Architektur L. Dargestellt ist der Mittelwert des klassenspezifischen Mittels auf den Testdaten und als Fehlerbalken das 95%ige Konfidenzintervall. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand gewählt.

b ist, desto stärker wird die Klassifikation des akustischen Kanals berücksichtigt.

Dies läßt sich auch an den Klassifikationsergebnissen mit unterschiedlichen Werten für b erkennen, die in Abbildung 8.9 zu sehen sind. Wiederum wird als Entscheidungskriterium des Assoziativspeichers der Hammingabstand gewählt, da damit schon zuvor die bessere Performanz erzielt werden konnte. Alle Ergebnisse dieses Versuchs sind der Tabelle A.9 im Anhang A zu entnehmen. Zusätzlich sind auch die Ergebnisse für den Fall aufgetragen, daß dem akustischen Kanal im Zweifelsfall immer der Vorzug gegeben wird. Diese Möglichkeit soll im folgenden als L(ak) bezeichnet werden. Die Klassifikationsleistungen der Architektur L(200) ist der Leistung der Architektur L(ak) am nächsten, wobei L(ak) für ungestörte Daten sowie für einen SNR von 3 dB und -3 dB die besten Ergebnisse liefert. Für die Architektur L(ak) ist der Abfall der Performanz bei kleinem SNR am stärksten, wohingegen für die Architektur mit b=-100 die Performanz auch bei kleinem SNR nahezu konstant ist, wobei die Ergebnisse hier aber insgesamt auf niedrigem Niveau liegen. Allein für ein SNR von -11 dB sind die Ergebnisse für die Architekturen L(0) bzw. L(-100) am besten. Daran läßt sich erkennen, daß eine starke Gewichtung des visuellen Kanals erst für ein SNR von -11 dB notwendig wird, da sich hier die Performanzergebnisse zwischen den untersuchten Alternativen umkehren.

Insgesamt läßt sich feststellen, daß auch bei den Fusionsmöglichkeiten, die den Hammingabstand als Entscheidungskriterium verwenden, der Parameter b gemäß des SNR gewählt werden muß, falls in jedem Fall das beste Ergebnis für diese Architektur erzielt werden soll. Geht es allein um eine Verbesserung

gegenüber der naiven Strategie bei L(ak), ohne daß das SNR bekannt ist, kann L(200) als bessere Strategie gewählt werden, da hier für größere SNR nahezu gleiche Performanz wie bei der Architektur L(ak) erzielt wird (eine leichte Abweichung bei SNR = 3 dB), für SNR = -11 dB diese aber bedeutend besser ist.

8.6 Kapitelzusammenfassung

Erste Versuche mit hochdimensionalen Daten zeigen, daß die Hinzunahme der visuellen Informationen die Klassifikationsleistung in den meisten Fällen steigern kann. Das Versagen der Architektur mit früher Fusion I1 in dieser Hinsicht kann auf die zu große Anzahl an Gewichten dieses Netzes zurückgeführt werden. Die beste Performanz bieten Architekturen, bei denen die Fusion auf einer Zwischenstufe erfolgt (Architekturen H1 und H2), die schlechtesten Leistungen erbringt die Architektur L, bei der die Fusion auf der Ebene der Buchstabenworterkennung erfolgt. Diese Ergebnisse werden sowohl für ungestörte als auch für verrauschte akustische Daten erzielt. Dabei läßt sich feststellen, daß mit gestörten akustischen Daten trainiert werden sollte, wenn auch mit ebensolchen Daten getestet wird.
Bei allen Untersuchungen kann festgestellt werden, daß als Entscheidungskriterium des Assoziativspeichers der Hammingabstand dem Überlapp vorzuziehen ist.
Ebenfalls läßt sich den Tabellen des Anhangs A entnehmen, daß die Klassifikationsleistung hinsichtlich des klassenspezifischen Mittels für die Buchstabenworterkennung besser als jene auf der Segmentklassifikationsebene ist, was dadurch erklärt werden kann, daß der Assoziativspeicher in der Lage ist, systematische Fehlklassifikationen auf der Segmentebene auszugleichen.
Schließlich kann für Architektur L gezeigt werden, daß die Verwendung eines Schwellwertes für den Hammingabstand gemäß Kapitel 7.4 die Klassifikationsleistungen gegenüber der einfachen Strategie, der akustischen Klassifikation den Vorzug zu geben, zu einer Verbesserung führt, insbesondere bei stark gestörtem akustischem Kanal. Die Wahl dieser Schwelle sollte idealerweise vom SNR des akustischen Kanals abhängig gemacht werden.

Kapitel 9

Fusion mit dimensionsreduzierten Daten

Im vorangegangenen Kapitel konnte gezeigt werden, daß eine zu große Anzahl an Gewichten die Klassifikationsleistung eines Netzes neagtiv beeinflußt. Um einen besseren Vergleich der Fusionsarchitekturen mit Hilfe kleinerer Netze durchführen zu können, ist es sinnvoll, eine Dimensionsreduktion der Eingangsdaten der Netze durchzuführen. Zu diesem Zweck wird eine lineare Diskriminanzanalyse (LDA) gewählt. In diesem Kapitel werden die entsprechenden Versuchsergebnisse dargestellt und diskutiert (siehe auch [TW99]). Die Darstellung der Ergebnisse folgt wiederum den schon in der Einleitung zu Kapitel 8 dargelegten Beschreibungen.
Die Verwendung eines dimensionsreduzierenden Verfahrens ermöglicht auch die Verwendung des vollständigen 100ms-Fensters als Merkmal für den visuellen Kanal, was bisher nur für den akustischen Kanal möglich war. Da die Videobilder etwa alle 33 msec neu aufgenommen werden, reicht es, nur die zu der ersten, vierten, siebten und zehnten Zeitscheibe des 100ms-Fensters gehörenden Bilder zu betrachten, um alle sich unterscheidenden Bilder des Zeitfensters abzudecken. Diese werden zu einem 560-dimensionalen Vektor zusammengefügt und einer linearen Diskriminanzanalyse unterzogen.
Schließlich ermöglicht die weitere Vorverarbeitung der Daten auch die Realisierung einer weiteren Fusionsarchitektur, bei der beide Kanäle zusammengeführt werden, bevor die lineare Diskriminanzanalyse durchgeführt wird.

9.1 Bestimmung der Anzahl der LDA-Koeffizienten

Um die Dimension der hochdimensionalen Daten zu reduzieren, wird als weitere Vorverarbeitung das Verfahren der Linearen Diskriminanzanalyse gewählt, welches in Kapitel 3.1.1 erläutert wird.
Die lineare Diskriminanzanalyse wird auf dem Gebiet der audio-visuellen Sprachverarbeitung neben der Hauptkomponentenanalyse verwendet, um eine Reduktion der Dimensionalität der Datenvektoren zu erreichen. Dabei wird das Verfahren in ersten Linie dazu verwendet, die Dimensionalität des visuellen Datenmaterials zu verringern, das im allgemeinen in Form von Grauwertbildern der

Anzahl an LDA-Koeef.	AKUSTISCH KSPM (Standardabweichung)	VISUELL KSPM (Standardabweichung)
5	56.86 (2.34)	30.24 (2.27)
7	67.49 (1.88)	32.55 (3.64)
9	72.02 (2.54)	33.43 (2.20)
10	71.38 (3.01)	33.26 (2.95)
15	75.78 (2.03)	33.87 (1.40)
20	76.72 (2.84)	35.32 (1.93)
25	77.57 (2.39)	35.58 (2.06)

Tabelle 9.1: Klassifikationsergebnisse für einen k-Nearest-Neighbour-Klassifikator mit unterschiedlicher Anzahl an LDA-Koeffizienten, k=7 (vgl. Kapitel 3.1.1). Es sind die gemittelten Ergebnisse auf den Testdatensätzen als klassenspezifische Mittel und in Klammern die Standardabweichungen angegeben.

Lippenregion vorliegt, aber auch zur Bearbeitung des akustischen Datenmaterials wird die LDA verwendet (vgl. [Sio95]).[1] Die mit Hilfe der LDA erzielten Ergebnisse sind in der Regel besser als diejenigen, die unter Zuhilfenahme einer Hauptkomponentenanalyse erzielt werden können (vgl. [PG98],[MSYW99], [DHB+95]).
Zur Bestimmung der Anzahl der LDA-Koeffizienten für den akustischen und den visuellen Kanal werden Vorversuche unternommen. Dazu wird für jeden Crossvalidierungssatz einzeln aus den Trainingssätzen die Projektionsmatrix der LDA berechnet und diese dann auf den Testsatz des jeweiligen Crossvalidierungssatzes angewendet. Untersucht werden unterschiedliche Anzahlen an Koeffizienten (5,7,9,10,15,20 und 25 Koeffizienten) für jeden Kanal separat. Dabei ist zu beachten, daß die Anzahl der LDA-Koeffizienten maximal 29 betragen kann (vgl. Kapitel 3.1.1). Die mit der LDA vorverarbeiteten Trainingsdatensätze werden für die Konstruktion eines k-Nearest-Neighbour-Klassifikators mit k=7 verwendet, der der Klassifikation der entsprechenden Testsätze dient. Die Ergebnisse werden über die fünf Crossvalidierungssätze gemittelt und sind der Tabelle 9.1 zu entnehmen.
Für die Bestimmung einer geeigneten Anzahl an Koeffizienten muß berücksichtigt werden, daß jeder Kanal zumindest dieselbe Klassifikationsleistung wie bei Verwendung der hochdimensionalen Daten liefern sollte, die für den akustischen Kanal bei 74.48 % und für den visuellen Kanal bei 27.17 % liegt. Für eine Koeffizientenanzahl ab 15 ist dies für den akustischen Kanal der Fall, der visuelle Kanal übertrifft auch schon bei kleiner Koeffizientenanzahl den vorgegebenen

[1] Im Falle der visuellen Daten werden auch spezifischere Verfahren verwendet, wie z.B. die Modellierung der Lippenkonturen (vgl. [BO94],[LTB96b],[LTB96c],[LD98],[Vog95],[BOSK95],[DKB95]) oder die Verwendung lippenspezifischer Parameter (vgl. [NPZ99],[AB95b]), um eine Dimensionsreduktion zu erreichen.

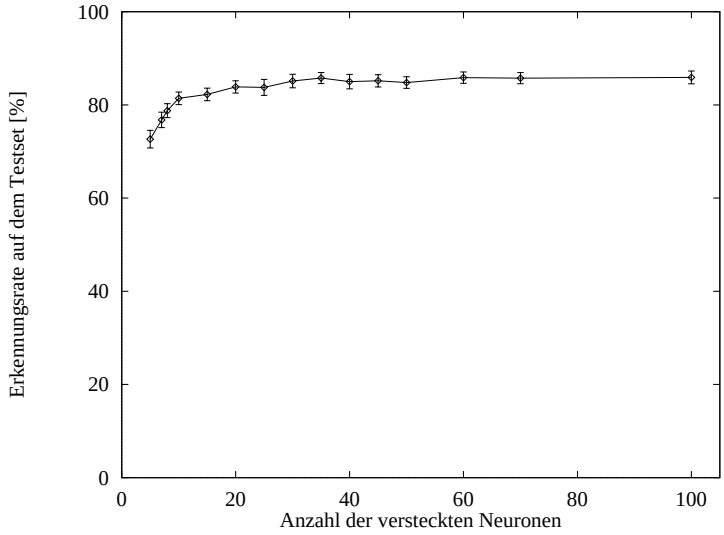

Abbildung 9.1: Klassifikationsergebnisse für den akustischen Kanal bei Variation der Anzahl der versteckten Neuronen. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

Wert.
Da in den folgenden Versuchen auch untersucht werden soll, wie sich eine LDA-Vorverarbeitung auf den gekoppelten akustischen und visuellen Kanal auswirkt, ist die Gesamtanzahl der akustischen und visuellen Koeffizienten auf 29 beschränkt. Die insgesamt beste Klassifikationsleistung ergibt sich für den akustischen Kanal, der daher durch eine möglichst große Anzahl an Koeffizienten repräsentiert werden soll. Die beste Leistung wird mit 25 Koeffizienten erreicht. Die Leistung des visuellen Kanals ist allerdings bei 5 Koeffizienten am niedrigsten, während sie für 9 oder 10 Koeffizienten um ca. 3 % größer ist. Insgesamt werden daher 20 Koeffizienten für den akustischen Kanal und 9 Koeffizienten für den visuellen Kanal gewählt.

9.2 Optimierung der Netztopologien

Um die zu verwendende Anzahl an versteckten Neuronen zu ermitteln, werden wiederum sowohl Versuche für die Einzelkanäle als auch für die Kombination der Kanäle durchgeführt.
Für alle Versuche zeigt sich wiederum, daß die Verwendung des Hammingabstandes der Verwendung des Überlapps als Entscheidungskriterium des Assoziativspeichers überlegen ist (vgl. Anhang B), so daß im folgenden nur die Ergebnisse bei Verwendung des Hammingabstandes angegeben sind.
Zunächst sollen die Ergebnisse der Einzelkanäle betrachtet werden. Die Abbildungen 9.1, 9.2 und 9.3 zeigen die entsprechenden Ergebnisse.
Es läßt sich erkennen, daß auch bei einer hohen Anzahl an versteckten Neuronen

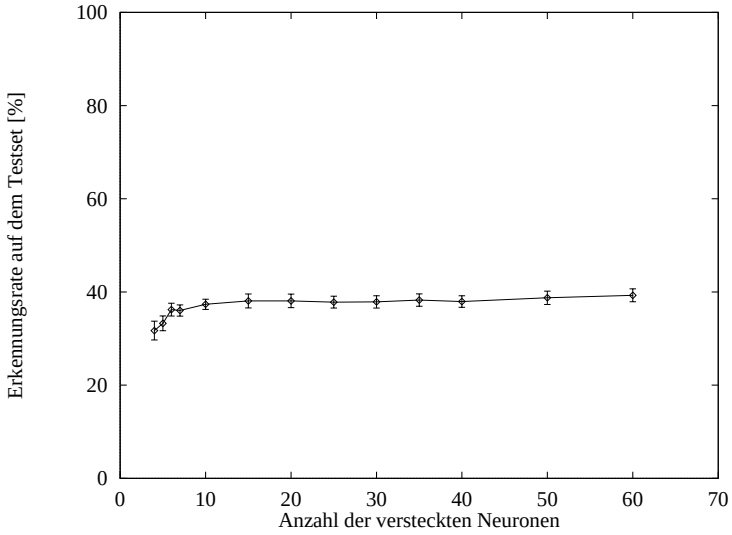

Abbildung 9.2: Klassifikationsergebnisse für den visuellen Kanal bei Variation der Anzahl der versteckten Neuronen, Verwendung nur eines Bildes für jedes Zeitfenster als visuelles Merkmal. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

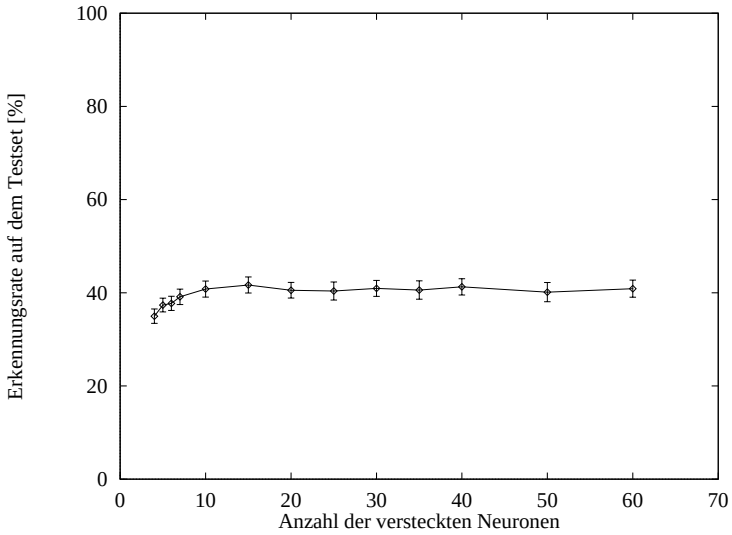

Abbildung 9.3: Klassifikationsergebnisse für den visuellen Kanal bei Variation der Anzahl der versteckten Neuronen, Verwendung des ganzen Zeitfensters als visuelles Merkmal. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

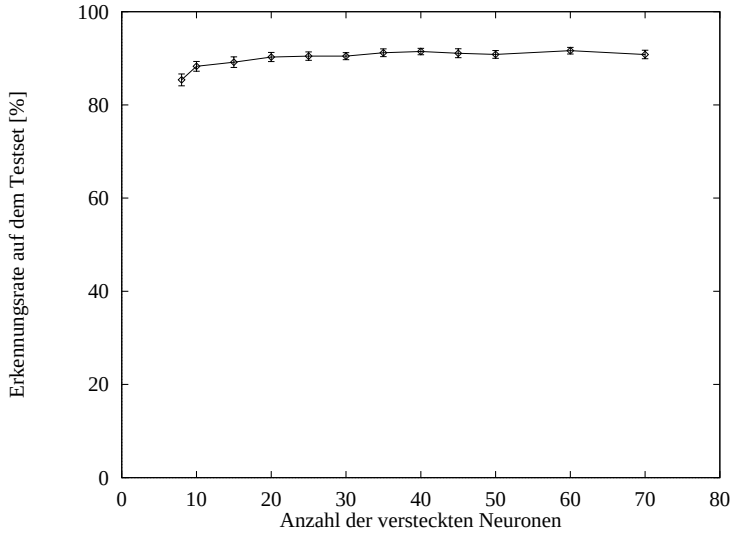

Abbildung 9.4: Klassifikationsergebnisse für die Kombination aus akustischem und visuellem Kanal und Verwendung der Architektur E bei Variation der Anzahl der versteckten Neuronen, für den visuellen Kanal wird nur ein Bild pro Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

kein Overfitting-Effekt auftritt, sondern die Klassifikationsleistung annähernd konstant bleibt. Bei Verwendung der akustischen Daten ergibt sich ab einer Neuronenanzahl von 20 kein statistisch signifikanter Unterschied zum besten Wert mehr, der bei einer Neuronenanzahl von 35 liegt und später wieder für höhere Neuronenanzahlen von 60 und mehr erreicht wird. Bei den visuellen Daten liegt der beste Wert bei einer Neuronenanzahl von etwa 15 Neuronen. Auch hier sind die Unterschiede zu den Klassifikationsleistungen für 6 (bzw. 7 bei der Verwendung des ganzen Zeitfensters als visuelle Eingabe) und mehr Neuronen nicht mehr statistisch signifikant.
Dieser Plateau-Effekt zeigt sich auch, wenn die Kombination beider Kanäle betrachtet wird. Neben der einfachen Aneinanderhängung der dimensionsreduzierten akustischen und visuellen Daten, wie sie für die Architektur I notwendig ist, wird auch eine Kombination betrachtet, bei der die akustischen und visuellen Daten vor der Dimensionsreduktion aneinandergehängt werden. Diese Fusionsmöglichkeit soll im folgenden als 'E' bezeichnet werden.
Die Abbildungen 9.4 und 9.5 zeigen die Ergebnisse für die Architekturen E und I1 bei Verwendung nur eines Bildes des Eingabefensters als visuelle Daten.
Die beste Leistung der Architektur E wird bei einer Neuronenanzahl von 40 und später nochmal bei 60 versteckten Neuronen erreicht, obwohl auch hier die Ergebnisse für 20 und mehr versteckte Neuronen nicht mehr statistisch signifikant unterschiedlich zum besten Ergebnis sind. Für Architektur I1 wird die beste Leistung für 70 gefolgt von 50 Neuronen erreicht, wobei auch hier die Unterschiede schon bei einer Neuronenanzahl von 25 und mehr nicht mehr sta-

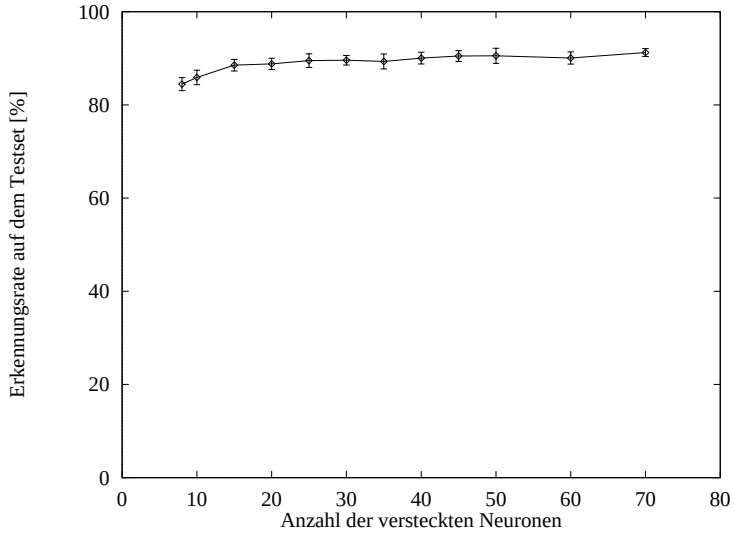

Abbildung 9.5: Klassifikationsergebnisse für die Kombination aus akustischem und visuellem Kanal und Verwendung der Architektur I1 bei Variation der Anzahl der versteckten Neuronen, für den visuellen Kanal wird nur ein Bild pro Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

tistisch signifikant sind.
Dieselben Effekte lassen sich feststellen, wenn für den visuellen Kanal das ganze 100ms-Fenster verwendet wird. Die entsprechenden Ergebnisse lassen sich den Abbildungen 9.6 und 9.7 entnehmen.
Das beste Ergebnis für Architektur E liegt hier bei 40 bzw. 45 Neuronen und schon ab einer Neuronenanzahl von 20 unterscheiden sich die Klassifikationsleistungen nicht mehr statistisch signifikant vom besten Ergebnis. Bei der Architektur I1 läßt sich das beste Ergebnis für eine Neuronenanzahl von 50 Neuronen finden, wobei der Unterschied zu anderen Klassifikationsleistungen schon für eine Neuronenanzahl von 15 nicht mehr statistisch signifikant ist.
Insgesamt läßt sich feststellen, daß sich für die Wahl der Anzahl der versteckten Neuronen für beide Kanäle ein breiter möglicher Wertebereich ergibt. Als erster Ansatzpunkt soll daher eine Gesamtneuronenanzahl von 40 bzw. 45 Neuronen näher untersucht werden, da in diesem Bereich die zum Teil besten Klassifikationsergebnisse erzielt werden und gleichzeitig die Anzahl der Neuronen noch relativ klein ist. Dazu werden die entsprechenden Architekturen I2 untersucht, bei denen es wie in Kapitel 8 gesehen, ebenfalls zu Problemen aufgrund der zu hohen Anzahl an Gewichten kommen kann. Die Ergebnisse sind der Tabelle 9.2 zu entnehmen.
Vergleicht man die Klassifikationsergebnisse bei Verwendung nur einer Zeitscheibe als visuelles Merkmal, läßt sich erkennen, daß schon bei einer Gesamtneuronenanzahl von 45 Neuronen in der ersten Schicht Schwierigkeiten aufgrund der höheren Anzahl der Gewichte zu erwarten sind, weshalb eine Gesamtneuro-

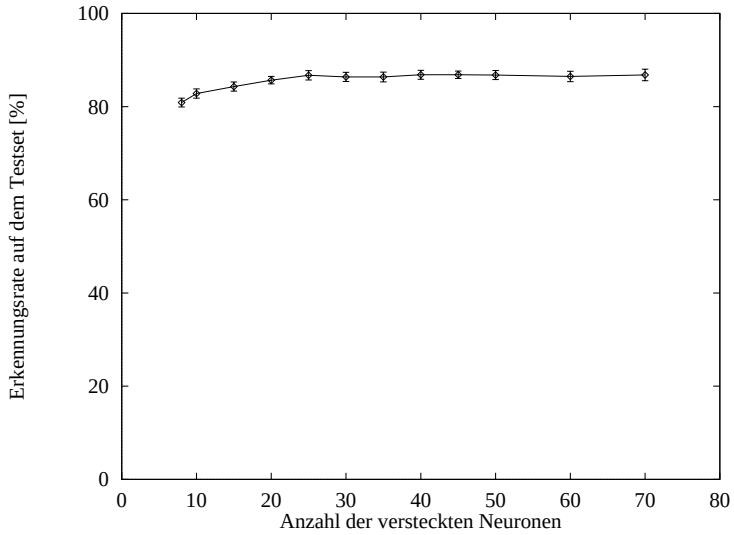

Abbildung 9.6: Klassifikationsergebnisse für die Kombination aus akustischem und visuellem Kanal und Verwendung der Architektur E bei Variation der Anzahl der versteckten Neuronen, für den visuellen Kanal wird das ganze Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

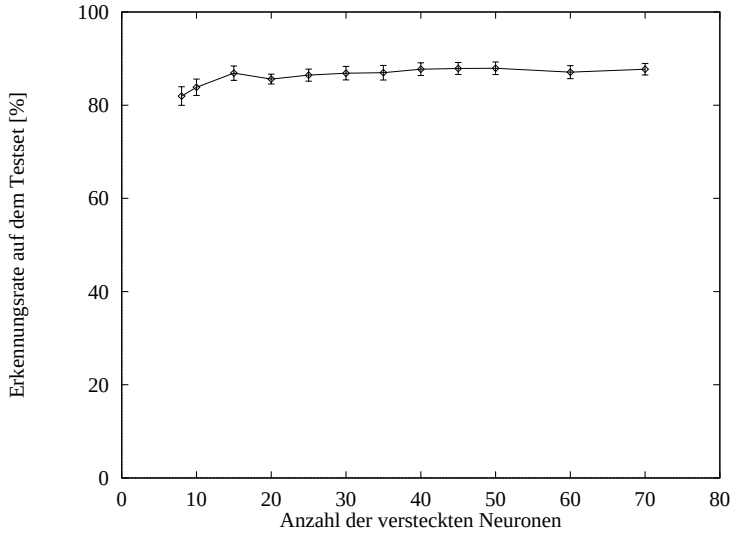

Abbildung 9.7: Klassifikationsergebnisse für die Kombination aus akustischem und visuellem Kanal und Verwendung der Architektur I1 bei Variation der Anzahl der versteckten Neuronen, für den visuellen Kanal wird das ganze Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

Architektur (Anzahl der Gewichte)	Architektur (Anzahl der Gewichte)
29-40-60-30 (5360)	29-45-60-30 (5805)
eine Zeitscheibe als visuelles Merkmal	
89.41 (1.44)	83.87 (6.33)
gesamtes Zeitfenster als visuelles Merkmal	
86.81 (1.39)	86.75 (1.08)

Tabelle 9.2: Klassifikationsergebnisse für die Architektur I2 bei Verwendung von 40 bzw. 45 versteckten Neuronen der ersten Schicht. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und in Klammern die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

nenanzahl von 40 gewählt wird. Für den visuellen Kanal wird eine Neuronenanzahl von 15 Neuronen übernommen, die sowohl bei der Verwendung des ganzen Fensters als auch bei Verwendung nur einer Zeitscheibe das jeweils bestes Ergebnis erzielt. Daraus ergibt sich für den akustischen Kanal eine Neuronenanzahl von 25 Neuronen für die versteckte Schicht.

9.3 Fusionsarchitekturen

Abbildung 9.8 zeigt die sich aus den Netztopologien für die Einzelkanäle ergebenden Fusionsarchitekturen. Durch die Verwendung einer Dimensionsreduktion als Vorverarbeitung ist wie schon erwähnt die Realisierung einer weiteren Fusionsarchitektur möglich. Dabei werden die Daten beider Kanäle zusammengeführt, bevor sie der linearen Diskriminanzanalyse unterzogen werden. Es handelt sich also um eine Fusionsart auf dem frühestmöglichen Level, daher soll diese Architektur im folgenden auch mit 'E' (für **E**arly) gekennzeichnet werden. Die Einordnung dieser Architektur innerhalb der anderen Architekturen kann man ebenfalls der Abbildung 9.8 entnehmen.

9.4 Experimente mit gestörtem akustischen Kanal

In diesem Kapitel sollen Experimente beschrieben werden, in denen die in Abbildung 9.8 dargestellten Fusionsarchitekturen mit Hilfe von ungestörten und gestörten akustischen Daten untersucht werden. Dazu werden für die gestörten akustischen Daten jeweils getrennt die Projektionen auf den Unterraum mit Hilfe der linearen Diskriminanzanalyse bestimmt.

9.4.1 Experimente mit akustischen Daten

Wie schon in Kapitel 8.5.1 beschrieben, wird ein Netz mittels eines der vier unterschiedlich gestörten niedrigdimensionalen Datensätze trainiert, um anschließend mit den übrigen drei Datensätzen getestet zu werden. Beim Testen mit

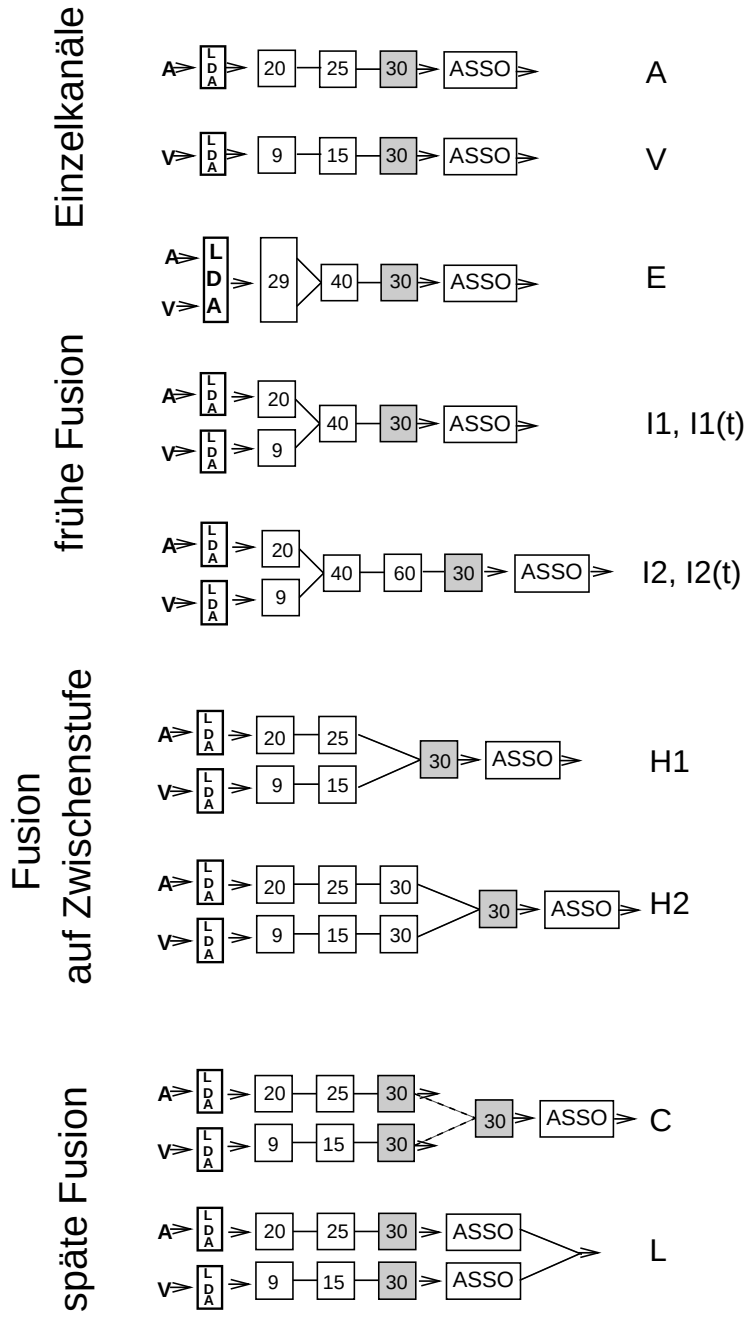

Abbildung 9.8: Übersicht über die Fusionsarchitekturen für dimensionsreduzierte Daten

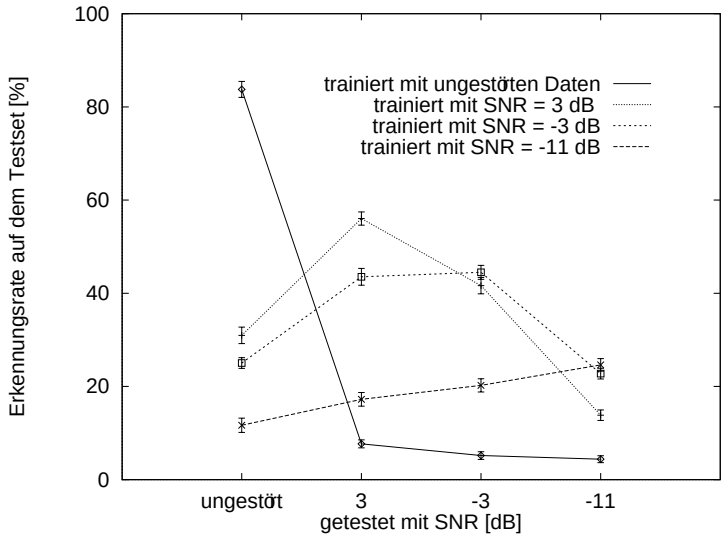

Abbildung 9.9: Versuche mit unterschiedlichen Signal-zu-Rauschverhältnissen für den akustischen Kanal. Es werden nur die akustischen Daten verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

einem anderen Datensatz wird für die Erzeugung des niedrigdimensionalen Testdatensatzes die Projektionsmatrix verwendet, die auch für die Erstellung des Trainingssatzes verwendet wird. Abbildung 9.9 zeigt die Ergebnisse. Man erkennt diesselben Tendenzen, die auch für die hochdimensionalen Daten zu erkennen sind.
Weiterhin laßt sich feststellen, daß die Ergebnisse im Vergleich zum gleichen Versuch mit hochdimensionalen Daten fast immer besser sind, wenn gestörte akustische Daten zum Training verwendet werden, wenngleich die Unterschiede nicht immer statistisch signifikant sind. Bei der Verwendung von ungestörten Daten als Trainingsmaterial ergeben sich für die hochdimensionalen Daten bessere Ergebnisse.
Wiederum läßt sich feststellen, daß es notwendig ist, mit gestörten Daten zu trainieren, wenn auch gestörte Daten zum Testen verwendet werden sollen.

9.4.2 Fusionsarchitekturen mit gestörtem akustischen Kanal

Dieser Abschnitt beschreibt die Experimente, die mit den Fusionsarchitekturen durchgeführt werden und führt einen Vergleich mit den entsprechenden Ergebnissen für das hochdimensionale Datenmaterial durch, bevor genauer auf den Vergleich der Fusionsarchitekturen eingegangen wird. Es schließt sich ein Vergleich der Ergebnisse für die verschiedenen visuellen Merkmale, die verwendet werden, an.
Wiederum zeigt sich, daß der Hammingabstand als Entscheidungskriterium des Assoziativspeichers eine bessere Klassifikationsleistung erbringt, als der Über-

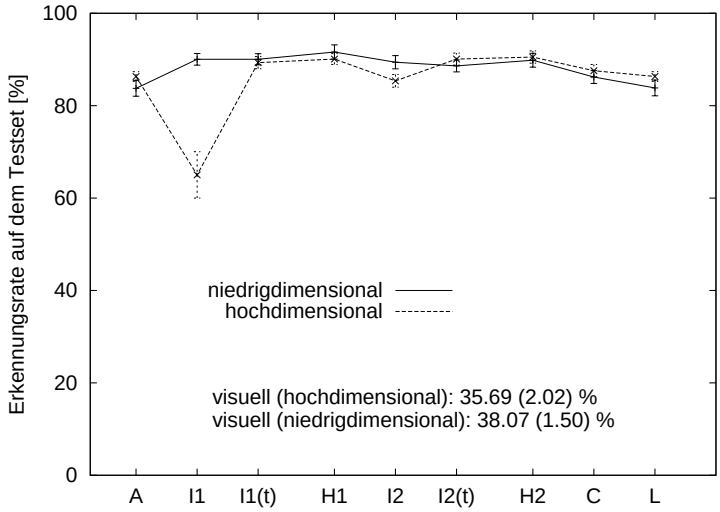

Abbildung 9.10: Vergleich der Ergebnisse für hoch- und niedrigdimensionale Daten für einen ungestörten akustischen Kanal (Verwendung einer Zeitscheibe als visuelles Merkmal). Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen. Als visuelles Merkmal wird jeweils nur eine Zeitscheibe des Zeitfensters verwendet.

lapp, so daß nur die Ergebnisse bei Verwendung des Hammingabstandes angegeben werden.

Vergleich der Ergebnisse für Originaldaten und dimensionsreduzierte Daten

Die Abbildungen 9.10 bis 9.13 zeigen die Ergebnisse, die mit den hoch- und niedrigdimensionalen Daten erzielt werden, jeweils für einen Datensatz mit unterschiedlicher Störung des akustischen Kanals. Als visuelles Merkmal wird jeweils nur eine Zeitscheibe des Zeitfensters verwendet.
Zunächst läßt sich feststellen, daß die Performanz für den visuellen Kanal für die niedrigdimensionalen Daten meistens besser ist, wenn auch der Unterschied nicht statistisch signifikant ist.
Aus Abbildung 9.10 läßt sich erkennen, daß die Performanz der Fusionsarchitekturen für beide Datensätze nicht statistisch signifikant unterschiedlich ist. Nur bei den Architekturen I1 und I2 ergeben sich für die niedrigdimensionalen Daten bessere Leistungen. Für die Klassifikation der rein akustischen Daten ist die Verwendung der hochdimensionalen Daten leicht besser.
Bei einem SNR von 3 dB für den akustischen Kanal (vgl. Abbildung 9.11) ergibt sich ein anderes Bild. Hier zeigen für den akustischen Kanal alleine beide Datensätze eine nahezu identische Leistung, während für die zweischichtigen Fusionsarchitekturen für die niedrigdimensionalen Daten die eindeutig besseren Ergebnisse erzielt werden. Bei den dreischichtigen Netzen ist die Leistung

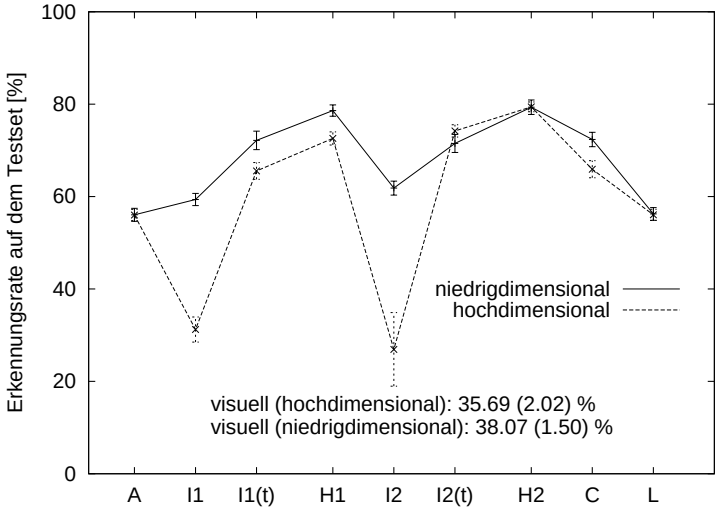

Abbildung 9.11: Vergleich der Ergebnisse für hoch- und niedrigdimensionale Daten (Verwendung einer Zeitscheibe als visuelles Merkmal). Das SNR für den akustischen Kanal beträgt 3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen. Als visuelles Merkmal wird jeweils nur eine Zeitscheibe des Zeitfensters verwendet.

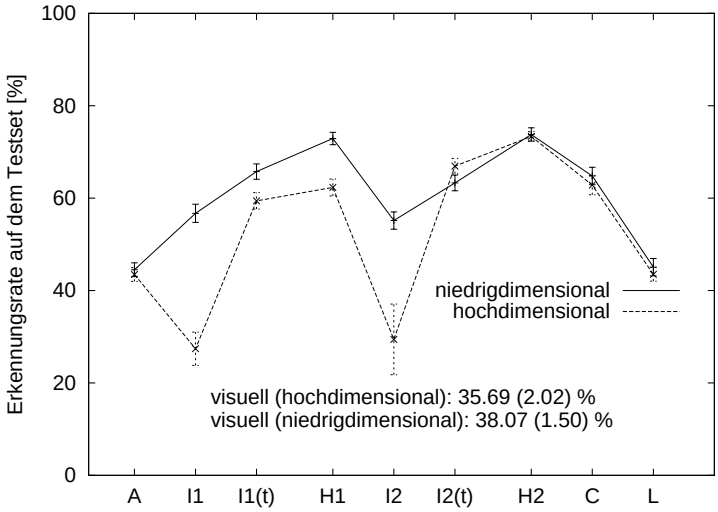

Abbildung 9.12: Vergleich der Ergebnisse für hoch- und niedrigdimensionale Daten (Verwendung einer Zeitscheibe als visuelles Merkmal). Das SNR für den akustischen Kanal beträgt -3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen. Als visuelles Merkmal wird jeweils nur eine Zeitscheibe des Zeitfensters verwendet.

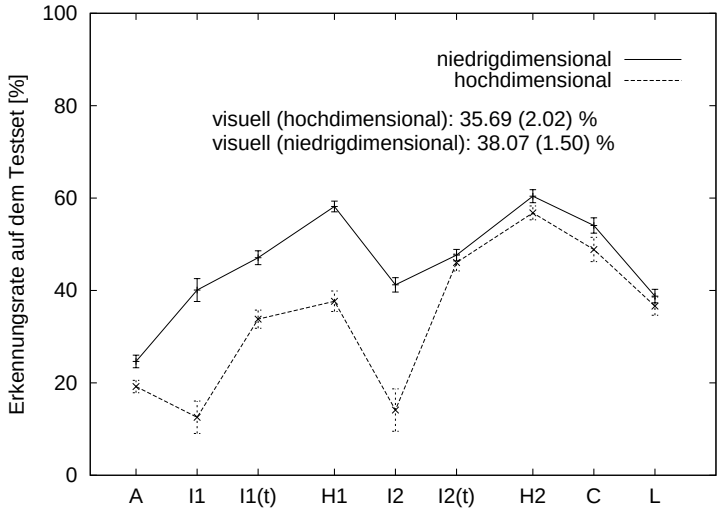

Abbildung 9.13: Vergleich der Ergebnisse für hoch- und niedrigdimensionale Daten (Verwendung einer Zeitscheibe als visuelles Merkmal). Das SNR für den akustischen Kanal beträgt -11 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen. Als visuelles Merkmal wird jeweils nur eine Zeitscheibe des Zeitfensters verwendet.

hingegen wieder nahezu gleich, nur für die Architektur C liefern die niedrigdimensionalen Daten wiederum bessere Ergebnisse.
Für einen SNR von -3 dB des akustischen Kanals ergibt sich fast ein identisches Bild (vgl. Abbildung 9.12), nur zeigt hier auch die Architektur C eine nahezu gleiche Performanz für beide Datensätze, wohingegen sich bei der Architektur I2(t) ein leichter statistisch signifikanter Unterschied zugunsten der hochdimensionalen Daten ergibt.
Bei einem SNR von -11 dB schließlich liefert der niedrigdimensionale Datensatz durchgängig die besten Ergebnisse, wie der Abbildung 9.13 zu entnehmen ist, wobei der Unterschied nur bei Architektur I2(t) und L nicht statistisch signifikant ist. Zum ersten Mal ist auch die Performanz für den akustischen Kanal alleine bei Verwendung des niedrigdimensionalen Datensatzes besser als Verwendung des hochdimensionalen Datensatzes.
Zusammenfassend kann festgestellt werden, daß die Dimensionsreduktion vor allem bei stark gestörtem akustischen Kanal eine Verbesserung der Klassifikationsleistung erbringt, was für durchgängig alle Architekturen gültig ist. Die Unterschiede bei den dreischichtigen Netzen sowie den späten Fusionsarchitekturen sind allerdings weit weniger stark ausgeprägt und zum Teil nicht statistisch signifikant bei gering oder nicht gestörten akustischen Daten.

Vergleich der Fusionsarchitekturen

Die Abbildungen 9.14 und 9.15 zeigen alle Ergebnisse, die mit Hilfe der niedrigdimensionalen Daten für die einzelnen Fusionsarchitekturen erzielt werden. Die Auftragung ist wiederum analog der Abbildung 8.8 bei Verwendung der hochdimensionalen Daten.
Zunächst sollen die Ergebnisse betrachtet werden, die bei Verwendung nur einer Zeitscheibe für die visuellen Daten erzielt werden. Diese sind Abbildung 9.14 zu entnehmen. Als Realisierungen der Architektur E werden die Architekturen I1, I1(t), H1, I2, I2(t) und H2 mit den zusammengenommen projezierten akustischen und visuellen Daten trainiert. Dabei bilden die Architekturen I1(t) und H1 sowie I2(t) und H2 in diesem Fall zwei Realisierungen des ausgedünnten Netzes mit unterschiedlichen Initialisierungen, da die Trennung der H-Netze in einen akustischen und einen visuellen Anteil bis zu ihrer Zusammenführung auf der Ebene der versteckten Schichten im Falle der gemeinsamen Vorverarbeitung beider Kanäle keine sinnvolle Struktur ist. Wie den Tabellen in Anhang B zu entnehmen ist, ergeben sich bei gleichem SNR zwischen den einzelnen Realisierungen der Architektur E für jeweils eine oder zwei versteckte Schichten keine statistisch signifikanten Unterschiede, so daß die jeweils beste Leistung der Architektur E als horizontale Linie in Abbildung 9.14 eingezeichnet ist. Die Architekturen I1(t) und H1 sowie I2(t) und H2 zeigen wie erwartet diesselbe Performanz.
Man erkennt, daß für einen ungestörten akustischen Kanal die Performanz aller Architekturen nahezu gleich ist, nur die späten Fusionsarchitekturen C und L liefern signifikant schlechtere Ergebnisse. Die besten Ergebnisse werden von den Architekturen H1 und E erzielt. Bis auf Architektur L liefern alle Fusionsarchitekturen auch bessere Leistungen als der Einzelkanal alleine.
Für einen gestörten akustischen Kanal verstärken sich die zuvor beschriebenen Tendenzen noch weiter: Die besten und dabei nahezu gleich guten Ergebnisse werden von den Architekturen H und E erbracht, gefolgt und statisch signifikant deutlich unterschiedlich von den Architekturen I1(t) bzw. I2(t) und C, wobei für ein SNR von -11 dB Architektur C bessere Ergebnisse erbringt. Noch schlechtere Ergebnisse liefert Architektur I1 bzw. I2 wiederum mit deutlich statistisch signifikantem Unterschied. Das Schlußlicht bildet Architektur L, die nur für ein SNR von -11 dB eine gleich gute Leistung wie Architektur I1 bzw. I2 erbringen kann.
Zu bemerken bleibt weiterhin, daß bei -11 dB die Leistung des akustischen Kanals schlechter als die des visuellen ist, die Fusionsarchitekturen aber alle wenigstens die Performanz des visuellen Kanals zeigen.
Für diesen Datensatz ergibt sich in etwa folgende Rangordnung der Architekturen: E und H erbringen die beste Leistung, gefolgt von I(t) und C. Das Schlußlicht bilden die Architekturen I und L.

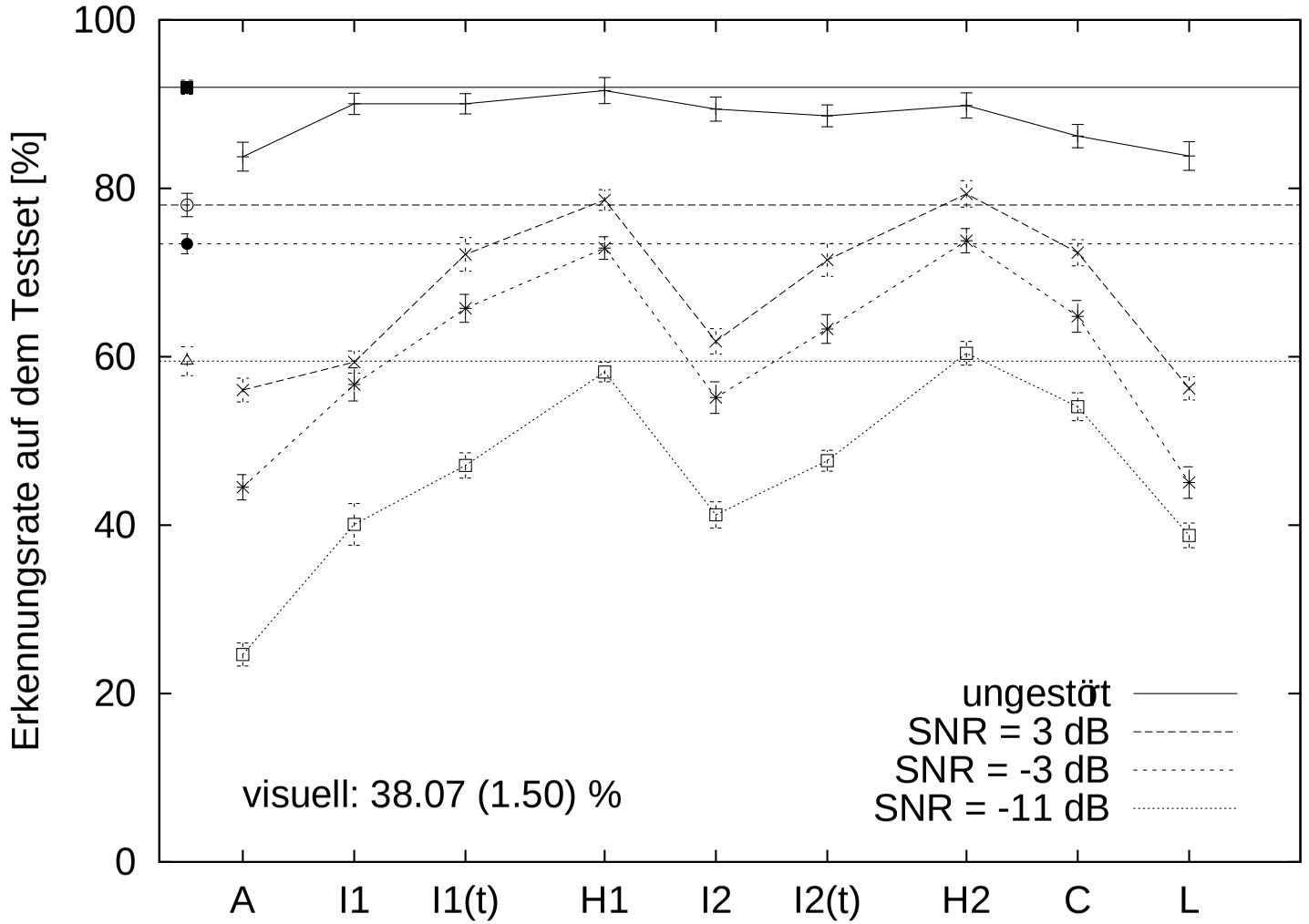

Abbildung 9.14: Ergebnisse für die Fusionsarchitekturen bei Verwendung niedrigdimensionalen Daten. Als visuelle Daten wird nur jeweils eine Zeitscheibe pro Zeitfenster verwendet. Als horizontale Linie ist die jeweils beste Leistung für die Architektur E dargestellt. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

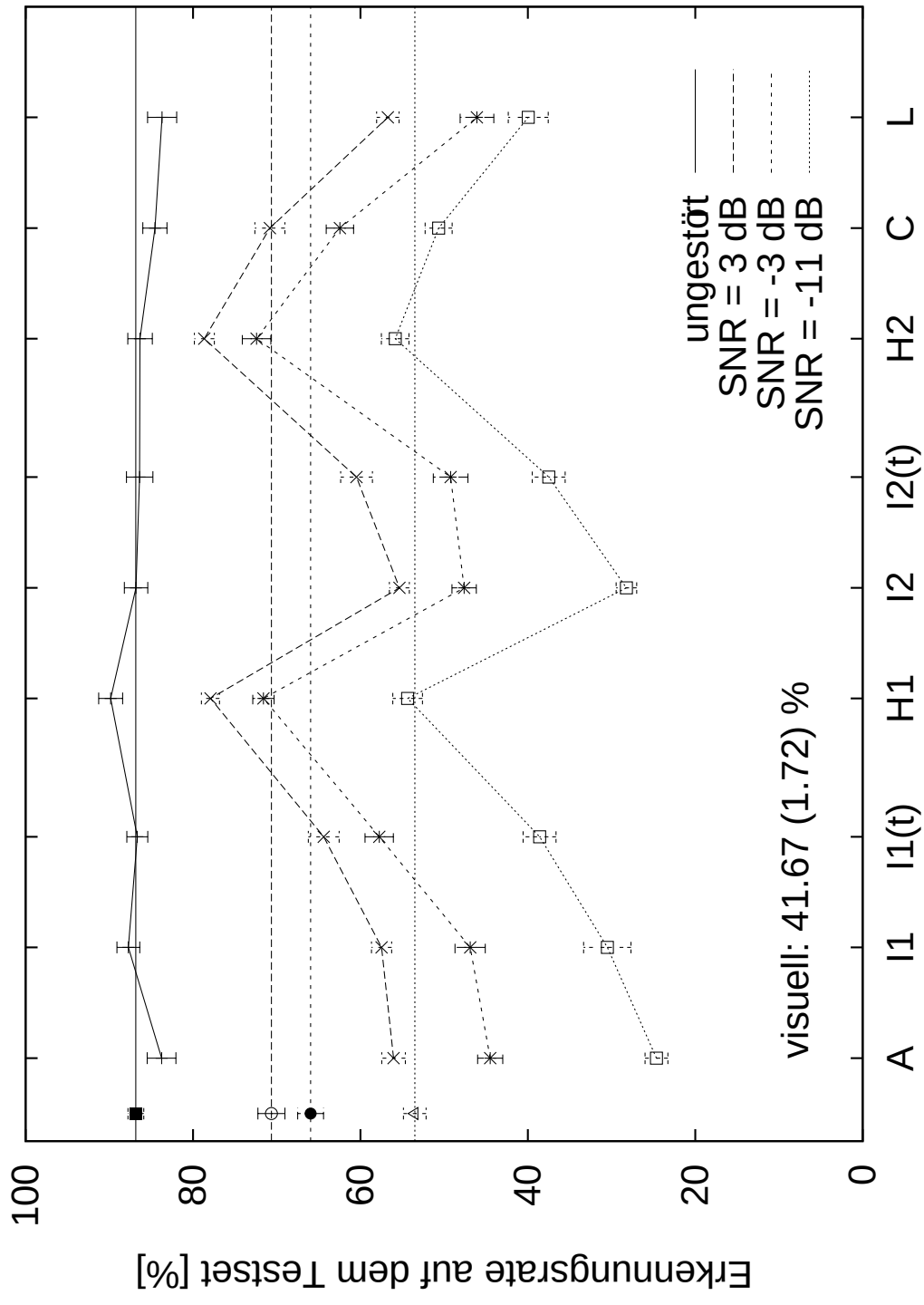

Abbildung 9.15: Ergebnisse für die Fusionsarchitekturen bei Verwendung niedrigdimensionalen Daten. Als visuelle Daten werden jeweils vier Zeitscheiben pro Zeitfenster verwendet. Als horizontale Linie ist die jeweils beste Leistung für die Architektur E dargestellt. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

Abbildung 9.15 zeigt die Ergebnisse für die einzelnen Fusionsarchitekturen, wenn als visuelle Daten das ganze Zeitfenster verwendet wird.
Zunächst läßt sich feststellen, daß die Performanz für den visuellen Kanal alleine signifikant besser ist als für den Fall, daß nur eine Zeitscheibe pro Zeitfenster als visueller Datensatz verwendet wird.
Bei Betrachtung der Ergebnisse der einzelnen Realisierungen für die Architektur E fällt auf, daß auch hier alle Realisierungen nahezu gleiche Performanz erbringen, nur für den Fall eines ungestörten akustischen Kanals ergibt sich für die Realisierungen mit Hilfe der Architekturen mit zwei versteckten Schichten ein signifikant schlechteres Ergebnis als für die beste Realisierung I1.
Die Fusionsarchitekturen zeigen insgesamt ein sehr ähnliches Leistungsbild wie im Falle des Datensatzes mit nur einer visuellen Zeitscheibe pro Zeitfenster. Auffällig ist hier, daß die Architektur E im Vergleich zur Architektur H1 bzw. H2 ein zum Teil signifikant schlechteres Ergebnis erzielt, insbesondere für einen SNR von 3 db und -3 dB. Die späten Fusionsarchitekturen liefern wiederum schlechtere Ergebnisse als die besten Architekturen, auch wenn diese Unterschiede im Falle eines ungestörten akustischen Kanals nicht signifikant sind. Weiterhin auffällig ist, daß die Architekturen I1, I1(t), I2 und I2(t) diesmal die schlechtesten Ergebnisse zusammen mit Architektur L liefern, Architektur C ist immer signifikant besser und erreicht sogar in etwa die Leistung der Architektur E.
Weiterhin läßt sich erkennen, daß bei Verwendung dieses Datensatzes und bei einem SNR von -11 dB die Fusionsarchitekturen I1, I1(t), I2 und I2(t) eine schlechtere Leistung erbringen als bei Verwendung des visuellen Kanals alleine. Nur die Architekturen E, H1, H2 und C erbringen signifikant bessere Leistungen.
Zusammenfassend läßt sich für diesen Datensatz folgende Rangordnung der Fusionsarchitekturen aufstellen: Architektur H erbringt die besten Leistungen, dicht gefolgt von Architektur E. Die nächstbeste Architektur ist C, gefolgt von den Architekturen I(t), I und L.

Vergleich unterschiedlicher visueller Merkmale

Wie schon im vorherigen Kapitel gesehen, ergibt sich bei der Verwendung des gesamten Zeitfensters als visuelles Merkmal für den visuellen Kanal alleine ein signifikant besseres Erkennungsergebnis. Fusioniert man allerdings den visuellen mit dem akustischen Kanal, kehrt sich dieses um: Die Daten, die nur eine visuelle Zeitscheibe verwenden, ergeben zum Teil die bessere Erkennungsleistung. Die Abbildungen 9.16 bis 9.19 zeigen diese Ergebnisse nochmals im Detail für die verschieden gestörten akustischen Daten.
Bei Verwendung des ungestörten akustischen Kanals sieht man, daß die Klassifikationsleistungen nahe beieinander liegen, ein statistisch signifikant besseres Ergebnis bei der Verwendung nur einer visuellen Zeitscheibe ergibt sich bei den Architekturen E, I1(t) und H2.
Bei Verwendung eines gestörten akustischen Kanals zeichnen sich gewisse Tendenzen ab: Hier ist die Leistung der späten Fusionsarchitekturen C und L, sowie die Leistung für die Fusion auf der Zwischenebene (Architekturen H1 und H2)

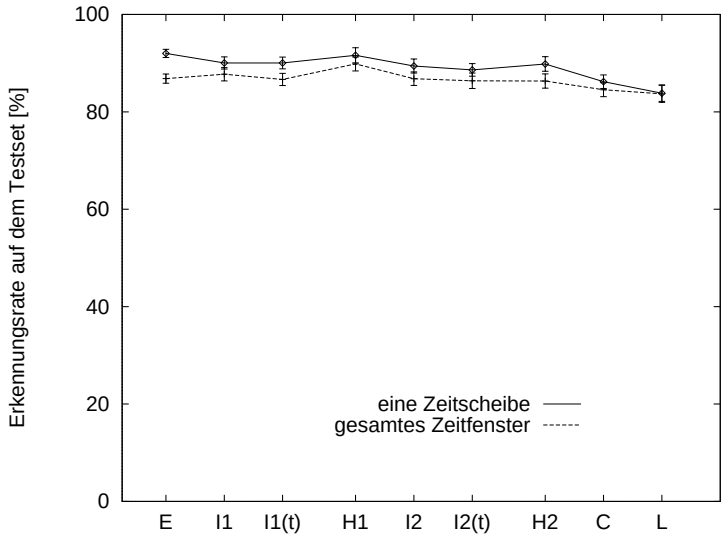

Abbildung 9.16: Vergleich der Ergebnisse der niedrigdimensionalen Daten bei der Verwendung nur einer visuellen Zeitscheibe pro Zeitfenster und der Verwendung des ganzen Zeitfensters als visuelles Merkmal. Der akustische Kanal ist ungestört. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

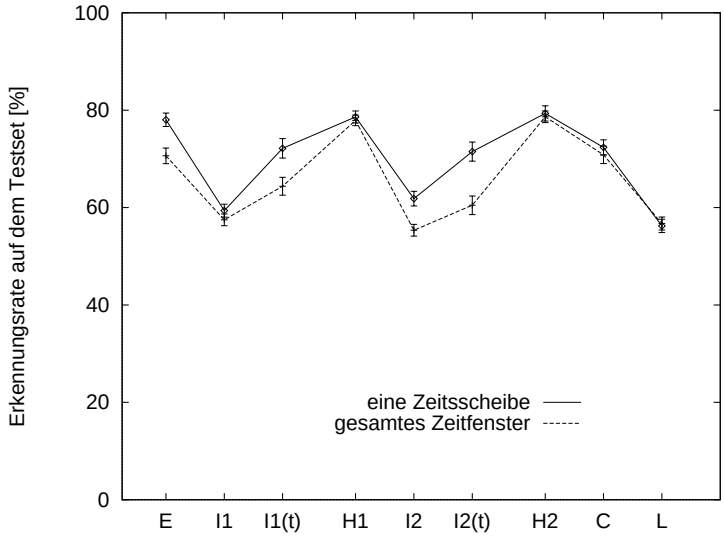

Abbildung 9.17: Vergleich der Ergebnisse der niedrigdimensionalen Daten bei der Verwendung nur einer visuellen Zeitscheibe pro Zeitfenster und der Verwendung des ganzen Zeitfensters als visuelles Merkmal. Der akustische Kanal ist gestört, das SNR beträgt 3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

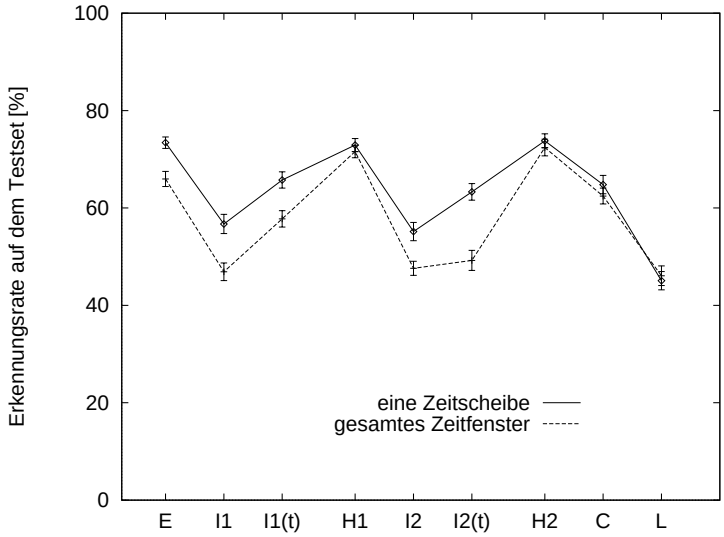

Abbildung 9.18: Vergleich der Ergebnisse der niedrigdimensionalen Daten bei der Verwendung nur einer visuellen Zeitscheibe pro Zeitfenster und der Verwendung des ganzen Zeitfensters als visuelles Merkmal. Der akustische Kanal ist gestört, das SNR beträgt -3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

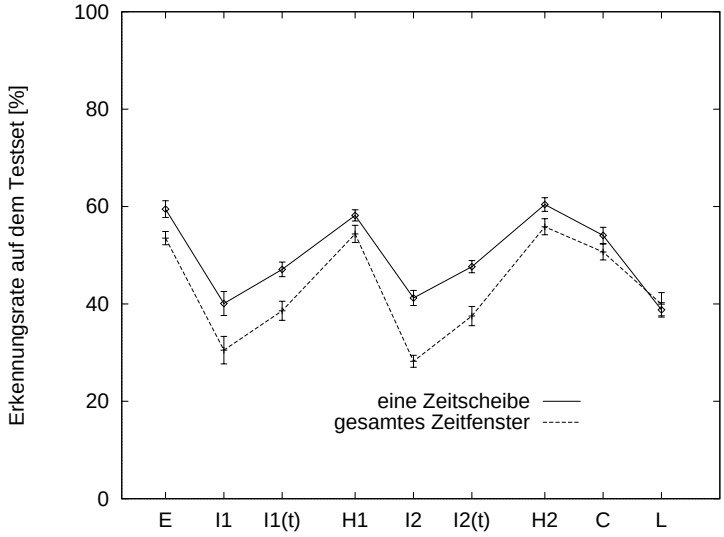

Abbildung 9.19: Vergleich der Ergebnisse der niedrigdimensionalen Daten bei der Verwendung nur einer visuellen Zeitscheibe pro Zeitfenster und der Verwendung des ganzen Zeitfensters als visuelles Merkmal. Der akustische Kanal ist gestört, das SNR beträgt -11 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

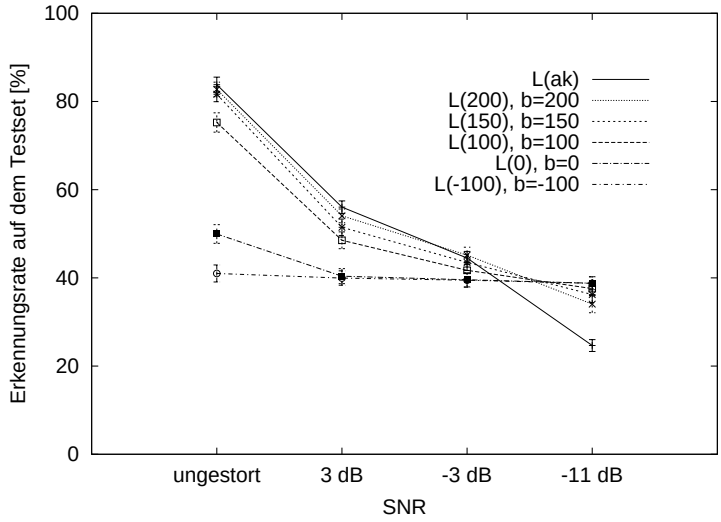

Abbildung 9.20: Ergebnisse mit niedrigdimensionalen Daten, nur die erste Zeitscheibe wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

nahezu gleich für beide Datensätze, wenn man gestörte akustische Daten mit einem SNR von 3 dB bzw. -3 dB verwendet. Für die frühen Fusionsarchitekturen E, I und I(t) ergeben sich größere Unterschiede in der Performanz. Bei der Verwendung eines stark gestörten akustischen Kanals (SNR=-11 dB) sackt auch die Leistung der Architekturen H1 und H2 ab, während sie für die Architekturen C und L nicht statistisch signifikant unterschiedlich sind. Bei Architektur L läßt sich im Gegenteil bemerken, daß die Leistung bei Verwendung des ganzen visuellen Zeitfensters gegenüber der Verwendung nur einer visuellen Zeitscheibe besser wird, je kleiner der SNR.
Zusammengefaßt läßt sich sagen, daß sich für die späten Architekturen C und L kein signifikant schlechteres Ergebnis feststellen läßt, bei den frühen Fusionsarchitekturen tritt dagegen ein deutlicher Unterschied hervor. Die Fusionsarchitekturen auf der Zwischenebene zeigen nur bei stark gestörtem akustischen Kanal eine schwächere Leistung.

9.4.3 Ergebnisse für die Architektur L

In diesem Abschnitt sollen die Ergebnisse besprochen werden, die mit Hilfe unterschiedlicher Schwellenstrategien für die Architektur L erzielt werden können. Die Abbildung 9.20 zeigt die Ergebnisse, die mit verschiedenen Schwellwerten b mit der Architektur L erzielt werden. Als visuelles Merkmal wird dabei nur die erste Zeitscheibe des Zeitfensters genutzt.
Man erkennt dieselben Tendenzen, die auch im Fall der hochdimensionalen Daten auftreten: Eine stärkere Gewichtung des visuellen Kanals ist erst bei einem

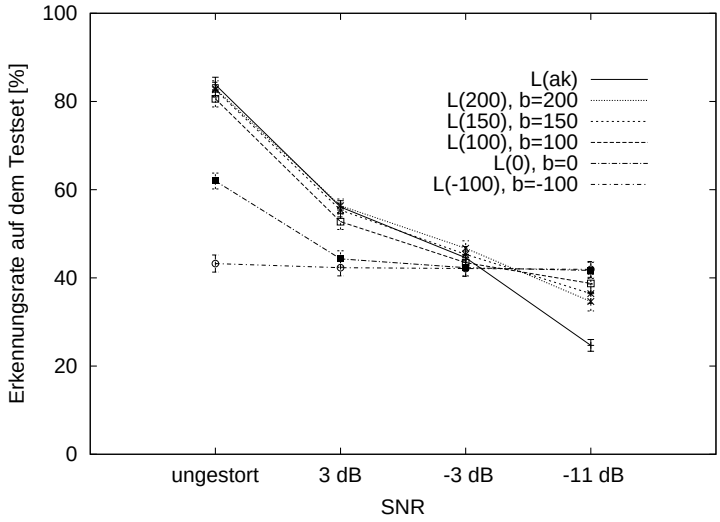

Abbildung 9.21: Ergebnisse mit niedrigdimensionalen Daten, das gesamte Zeitfenster wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

SNR von -11 dB notwendig, zuvor ergeben sich für die Architekturen L(ak) und L(200) die besten Ergebnisse, wobei im Falle eines SNR von -3 dB sogar Architektur L(200) leicht besser ist.
Auch in Abbildung 9.21, die die entsprechenden Ergebnisse bei Benutzung des ganzen Zeitfensters als visuelles Merkmal zeigt, erkennt man, daß die starke Gewichtung des visuellen Kanals erst bei einem SNR -11 dB wünschenswert ist. Jetzt zeigt allerdings die Architektur L(200) bei einem SNR von -3 dB schon eine deutlich bessere Performanz als Architektur L(ak), bei den anderen Signal-zu-Rausch-Verhältnissen sind die Ergebnisse nahezu identisch.
Um für jedes SNR das beste Klassifikationsergebnis zu erhalten, sollte der Schwellwert b also möglichst dem SNR angepaßt gewählt werden. Um dies zu erreichen, kann man sich z.B. das Verhältnis der berechneten Hammingabstände für den visuellen und den akustischen Kanal zu nutze machen, indem man den Wert für die Schwelle b abhängig von diesem Verhältnis wählt. Um eine geeignete Zuordnung zwischen den Verhältnissen und den Schwellwerten zu finden, sollen die Mittelwerte der Verhältnisse des 'akustischen' Hammingabstandes zum 'visuellen' Hammingabstand betrachtet werden (vgl. Tabelle 9.3). Für den Fall, daß die Architektur L(ak) die besten Ergebnisse erzielt, wird dies in der Tabelle mit einem Schwellwert $b > 200$ angezeigt.
Eine geeignete Wahl für die gesuchte Abbildung ist z.B. eine logistische Funktion, deren Parameter entsprechend den Mittelwerten für die Trainingsdaten aus Tabelle 9.3 gewählt und empirisch feiner abgestimmt werden können:

$$g(x) = \frac{k}{1 + e^{-\beta(x-v)}}. \tag{9.1}$$

SNR [dB]	Trainingsdaten	Testdaten	bester Schwellwert b
hochdimensionale Daten			
ungestört	1.59 (0.30)	1.58 (0.29)	> 200
3	5.05 (0.98)	5.20 (1.02)	> 200
-3	6.66 (1.36)	7.02 (1.48)	> 200
-11	5.87 (1.96)	6.55 (2.37)	0, -100
niedrigdimensionale Daten, eine Zeitscheibe als visuelles Merkmal			
ungestört	2.34 (0.30)	2.32 (0.36)	> 200
3	4.20 (0.51)	4.25 (0.56)	> 200
-3	5.80 (0.51)	5.83 (0.49)	200
-11	8.44 (1.57)	9.35 (1.82)	0, -100
niedrigdimensionale Daten, das ganze Zeitfenster als visuelles Merkmal			
ungestört	1.23 (0.16)	1.20 (0.17)	> 200
3	2.21 (0.25)	2.20 (0.25)	200
-3	3.05 (0.21)	3.03 (0.16)	200
-11	4.46 (0.83)	4.88 (0.96)	0, -100

Tabelle 9.3: Mittelwerte und Standardabweichungen der Verhältnisse von 'akustischem' Hammingabstand zum 'visuellen' Hammingabstand. Zusätzlich ist die Schwelle angegeben, für die das beste Klassifikationsergebnis erzielt wurde.

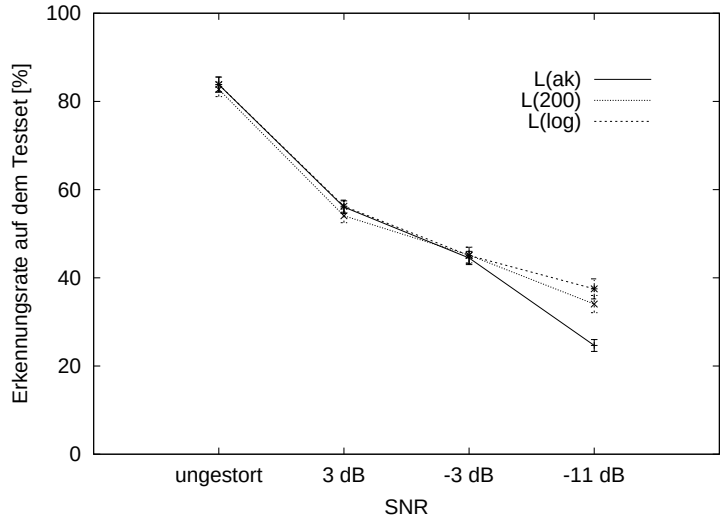

Abbildung 9.22: Ergebnisse mit niedrigdimensionalen Daten, nur die erste Zeitscheibe wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

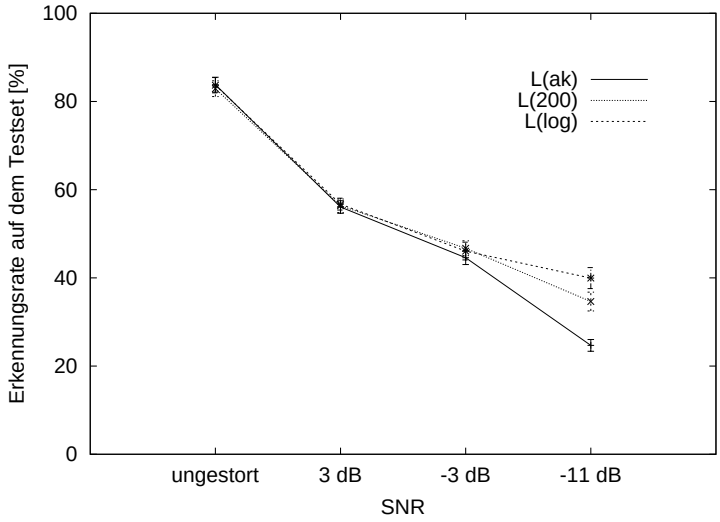

Abbildung 9.23: Ergebnisse mit niedrigdimensionalen Daten, das ganze Fenster wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

Dabei entspricht x dem für einen Datensatz ermittelten mittleren Verhältnis zwischen akustischem und visuellem Hammingabstand. Anhand der Funktion 9.1 kann dann der Schwellwert berechnet werden, der für diesen Datensatz verwendet wird. Der Schwellwert wird also ohne Kenntnis des Signal-Rausch-Verhältnisses der Daten berechnet. Die Architektur, die die logistische Funktion 3.15 als Entscheidungskriterium verwendet, soll mit L(log) bezeichnet werden. Die Abbildungen 9.22 und 9.23 zeigen die entsprechenden Ergebnisse im Vergleich zu den Klassifikationsleistungen der Architekturen L(ak) und L(200) bei Verwendung der niedrigdimensionalen Daten. Wie aus Tabelle 9.3 zu ersehen, liegt der Verhältnis-Mittelwert für die hochdimensionalen Daten bei einem SNR von -11 dB noch unter dem von -3 dB, gleichzeitig muß genau zwischen diesen beiden Signal-zu-Rauschverhältnnissen der Schwellwert gewechselt werden. Eine sinnvolle Anpassung der logistischen Funktion ist daher nicht möglich und soll hier nicht weiter betrachtet werden.
Für die niedrigdimensionalen Daten werden als Parameter der logistischen Funktion die Werte $k = 400$, $\beta = 1.1$ und $v = 6.0$ verwendet, falls nur die erste Zeitscheibe als visuelles Merkmal benutzt wird, bei Verwendung des gesamten Zeitfensters als visuelles Merkmal wird $v = 2.8$ verwendet, die Wahl der anderen Parameter bleibt gleich.
Man erkennt, daß durch den Gebrauch der Verhältnisse der Hammingabstände das Klassifikationsergebnis immer mindestens ebenso gut wie bei den Architekturen L(200) und L(ak) ist, obwohl der für jeden Datensatz neu ermittelte Schwellwert nicht unbedingt immer im jeweils günstigsten Schwellwertbereich liegen muß. Bei einem SNR von -11 dB ist die Architektur L(log) immer am be-

sten; für den Fall daß das gesamte Zeitfenster als visuelles Merkmal verwendet wird, ist der Unterschied zu den Leistungen der anderen Architekturen sogar statistisch signifikant. Allerdings ist zu erwähnen, daß die Architektur L(log) für einen SNR von -11 dB nicht ganz die Perfomanz erreicht, die bei Wahl eines Schwellwertes von $b = -100$ erzielt wird. Dies läßt sich wiederum damit erklären, daß nicht für jeden Datensatz dieser Schwellwert bei Verwendung der logistischen Funktion ermittelt wird, so daß insgesamt ein allerdings nicht statisch signifikant schlechteres Ergebnis erzielt wird.

9.5 Kapitelzusammenfassung

Dieses Kapitel beschreibt die Ergebnisse, die durch Verwendung einer Dimensionsreduzierung der Eingangsdaten mit Hilfe einer LDA erzielt werden können. Durch die Verwendung der LDA ist es zudem möglich, das ganze Zeitfenster als visuelles Merkmal zu verwenden, zudem kann als weitere Fusionsmöglichkeit die Architektur E realisiert werden, bei der beide Datensätze zusammengeführt werden, bevor die LDA angewendet wird.
Bei allen Versuchen kann wiederum festgestellt werden, daß der Hammingabstand gegenüber dem Überlapp die besseren Ergebnisse erzielt.
Bei der Optimierung der Architekturen sowohl für die Einzelkanäle als auch für die kombinierten Daten zeigt sich auch für eine große Anzahl an versteckten Neuronen kein Overfitting-Effekt.
Experimente mit gestörtem akustischem Kanal zeigen wiederum, daß mit gestörten Daten trainiert werden sollte, wenn mit ebensolchen getestet wird.
Die Dimensionsreduktion führt insbesondere bei stark gestörtem akustischen Kanal zu einer Verbesserung der Klassifikationsleistung.
Ein Vergleich der Fusionsarchitekturen zeigt erneut, daß die Architektur H die besten Leistungen erbringt, welche für den Fall, daß nur eine Zeitscheibe als visuelles Merkmal verwendet wird, auch von der Architektur E erreicht werden. Die schlechtesten Leistungen erbringen die Architekturen I und L, Architektur C erbringt mittlere Leistungen, wobei diese im Fall, daß nur eine Zeitscheibe als visuelles Merkmal verwendet wird, auch von Architektur I(t) erreicht wird, welche ansonsten ähnlich schlechte Leistungen wie I und L erbringt. Insgesamt wird die Klassifikationsleistung durch Verwendung der visuellen Merkmale insbesondere im Falle gestörter akustischer Daten deutlich erhöht.
Ein Vergleich der unterschiedlichen visuellen Merkmale zeigt bessere Klassifikationsleistungen insbesondere für die frühen Fusionsarchitekturen, wenn nur eine Zeitscheibe des Zeitfensters verwendet wird, obwohl in diesem Fall die Leistung für den visuellen Kanal alleine hinter derjenigen zurückbleibt, die mit Verwendung des gesamten Zeitfensters erreicht wird. Für die Architekturen C und L, sowie für Architektur H bei höherem Signal-Rausch-Verhältnissen sind die Leistungen aber in etwa gleich gut.
Ein Vergleich der Werte des Klassenspezifischen Mittels auf den Testdaten für die Klassifikation auf Segmentebene und auf Buchstabenwortebene zeigt erneut eine Verbesserung der Leistungen auf Buchstabenwortebene.
Für Architektur L kann gezeigt werden, daß eine Anpassung der Schwellwerte

je nach Datensatz möglich ist, indem etwa auf das Verhältnis der 'akustischen' zu den 'visuellen' Hammingabständen zurückgegriffen wird, ohne daß explizit das Signal-zu-Rausch-Verhältnis bekannt sein muß.

Kapitel 10

Fusion der gestörten akustischen Daten

Die in Kapitel 9 mit der Fusionsarchitektur E eingeführte Möglichkeit, eine Fusion mit Hilfe des vorangeschalteten Dimensionsreduktionsverfahrens durchzuführen, kann auch auf die unterschiedlich gestörten akustischen Daten ausgedehnt werden. Die verschieden gestörten akustischen Daten werden zunächst zusammengeführt und dann gemeinsam einer LDA unterzogen. Das Training der Netze erfolgt mit allen akustischen Daten gemeinsam, getestet werden die Architekturen mit Hilfe der Testdatensätze auch für die einzelnen Verrauschungsstufen getrennt.

Die Abbildungen 10.1 und 10.2 zeigen diese Fusionsmöglichkeit. Im Falle der Fusionsmöglichkeit GEM-I erfolgt diese Zusammenführung nur für die akustischen Daten alleine (vgl. Abbildung 10.1), bevor die LDA berechnet wird. Die mit Hilfe dieser berechneten Projektion dimensionsreduzierten akustischen Daten werden anschließend mit den visuellen Daten gekoppelt, wie es in Kapitel 8 und 9 für die Architektur I der Fall war. Dieselbe Art der Vorverarbeitung wird ebenfalls für die anderen Architekturen abgesehen von der Architektur E verwendet.

Bei der Fusionsmöglichkeit GEM-E werden akustische und visuelle Daten miteinander kombiniert, bevor die Dimensionsreduktion für beide zusammen erfolgt (vgl. Abbildung 10.2). Dies entspricht der Architektur E der vorangegangenen Kapitel.

Die folgenden Abschnitte beschreiben die mit dieser Fusion erzielten Klassifikationsergebnisse.

Zunächst kann festgestellt werden, daß als Entscheidungskriterium für den Assoziativspeicher der Hammingabstand gewählt werden sollte, da dieser bessere Ergebnisse als der Überlapp gewährleistet (vgl. Anhang C). Die im folgenden dargestellten Ergebniswerte beziehen sich daher auf diesen.

Die Abbildungen 10.3 und 10.4 zeigen die Ergebnisse, die mit der gemeinsamen Verarbeitung der unterschiedlichen Verrauschungsstufen erzielt werden für den Fall, daß nur eine Zeitscheibe als visuelles Merkmal verwendet wird. Die Fusionsmöglichkeiten GEM-E und GEM-I werden wieder als Architekturen E bzw. I dargestellt.

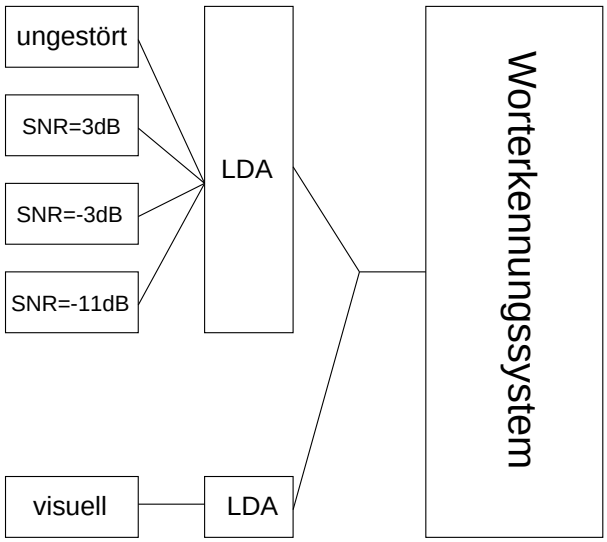

Abbildung 10.1: Fusionsarchitektur für eine gemeinsame Dimensionsreduktion der gestörten akustischen Daten. Fusionsmöglichkeit GEM-I.

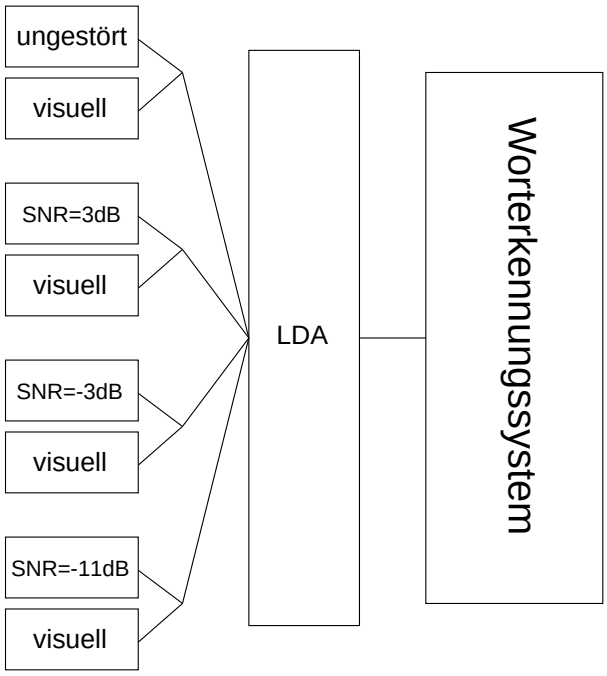

Abbildung 10.2: Architektur für eine Fusion der gestörten akustischen mit den visuellen Daten, bevor sie einer LDA unterzogen werden. Fusionsmöglichkeit GEM-E.

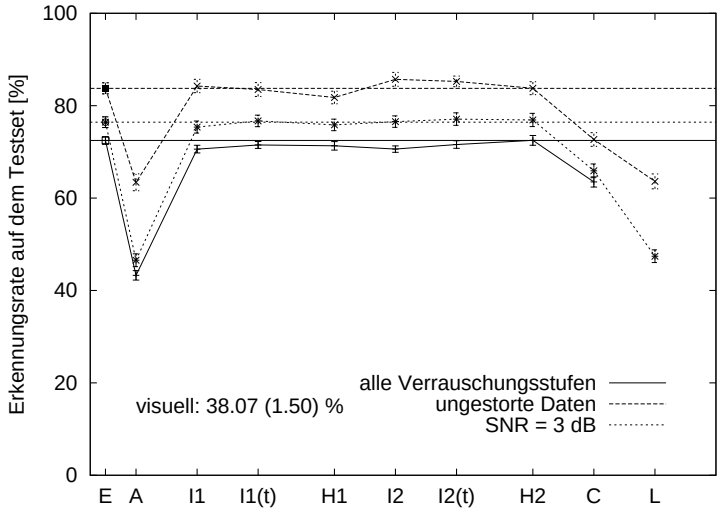

Abbildung 10.3: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für die ungestörten akustischen Daten und für gestörte Daten mit SNR = 3 dB, sowie die Performanz auf allen Rauschstufen. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

Man erkennt, daß erneut alle Fusionsarchitekturen außer der späten Architektur L eine bessere Leistung erzielen, als der akustische Kanal alleine. Nur für ein SNR von -3 dB und -11 dB ist auch die Architektur L besser als der akustische Kanal.
Für die ungestörten Daten, sowie für die mit einem SNR von 3 dB und -3 dB gestörten Daten sind alle anderen Architekturen in etwa gleich gut, nur die Architektur C liefert signifikant schlechtere Ergebnisse. Falls ungestörte akustische Daten zum Testen verwendet werden, ergibt sich bei den Architekturen mit zwei versteckten Schichten eine tendenziell bessere Performanz. Werden stark gestörte akustische Daten (SNR = -11 dB) zum Testen verwendet, ergeben sich die besten Leistungen für die Architekturen E, H1 und H2. Insgesamt läßt sich feststellen, daß die Architektur E immer mit die besten Leistungen erbringt.
Die Abbildungen 10.5 und 10.6 zeigen die Ergebnisse, die erzielt werden, wenn als visuelles Merkmal das vollständige Zeitfenster verwendet wird.
Es zeigen sich erneut dieselben Tendenzen: Alle Fusionsarchitekuren bis auf Architektur L zeigen bessere Leistungen als die Architektur, die ausschließlich die akustischen Daten verwendet, nur bei einem Signal-zu-Rausch-Verhältnis von -3 dB und -11 dB ist Architektur L besser als diese. Als zweitschlechteste Architektur erweist sich wiederum die späte Architektur C, für niedrige Signal-zu-Rausch-Verhältnisse ist Architektur C in etwa gleich gut wie die frühen Architekturen I, für SNR = -3 dB ist die Performanz ähnlich der der Architektur I2, für SNR = -11 dB ist sie in etwa gleich der der Architektur I1 und besser

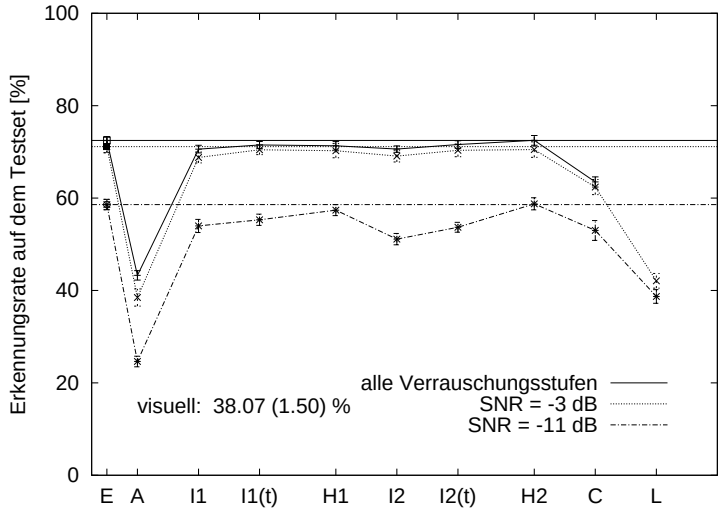

Abbildung 10.4: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für gestörte Daten mit SNR = -3 dB und -11 dB, sowie die Performanz auf allen Rauschstufen. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

als die der Architekturen I2 und I2(t).
Die besten Ergebnisse werden von den Architekturen H1 und H2 zusammen mit der Architektur I1(t) erzielt, während Architektur E einen deutlichen Abfall insbesondere für weniger oder gar nicht gestörte akustische Daten aufweist. Die Überlegenheit der Architektur H2 gegenüber anderen Architekturen mit zwei versteckten Schichten wird insbesondere im Falle eines niedrigen Signal-zu-Rausch-Verhältnisses deutlich, während die Performanz der Fusionsarchitekturen mit nur einer versteckten Schicht bei allen Signal-zu-Rauschverhältnissen ähnlich gut ist und nur für ein sehr niedriges SNR von -3 dB und -11 dB eine schlechtere Leistung der Architektur I1 sichtbar wird. Architektur E zeigt für alle Signal-zu-Rauschverhältnisse eine signifikant schlechtere Leistung als die jeweils beste Fusionsarchitektur, nur für ein SNR von -11 dB ist die Leistung die gleiche. Es kann weiterhin festgestellt werden, daß der Leistungsunterschied zwischen Architektur E und den jeweils besten Architekturen tendenziell umso größer ist, umso größer das Signal-zu-Rauschverhältnis ist. Bei fallendem SNR werden die Unterschiede immer geringer.

10.1 Vergleich unterschiedlicher visueller Merkmale

Die Abbildungen 10.7 und 10.8 zeigen einen Vergleich der Ergebnisse, die bei Verwendung unterschiedlicher visueller Merkmale erzielt werden. Dabei ist zu beachten, daß aus Darstellungsgründen in diesem Fall zum einen die Ergebnis-

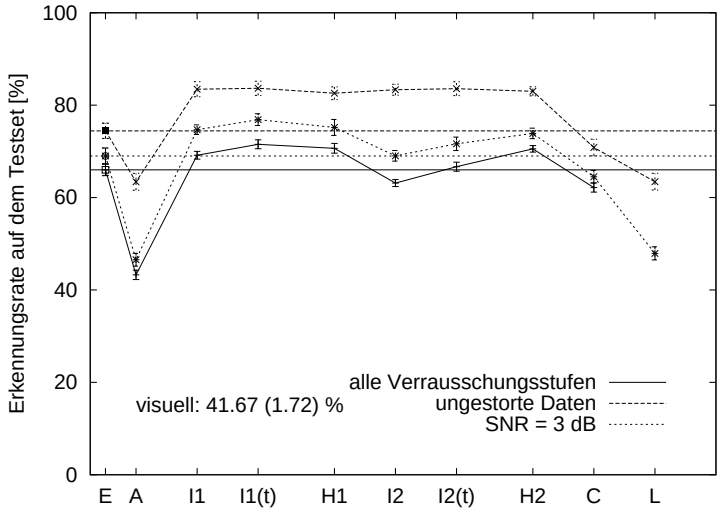

Abbildung 10.5: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für die ungestörten akustischen Daten und für gestörte Daten mit SNR = 3 dB, sowie die Performanz auf allen Rauschstufen. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

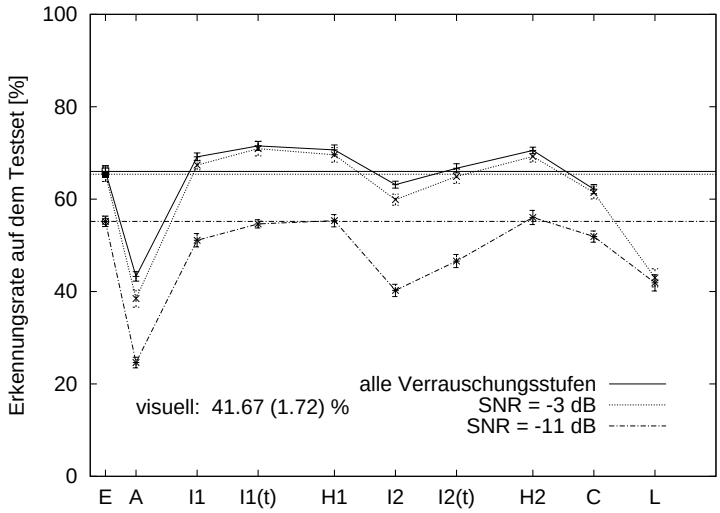

Abbildung 10.6: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für gestörte Daten mit SNR = -3 dB und -11 dB, sowie die Performanz auf allen Rauschstufen. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

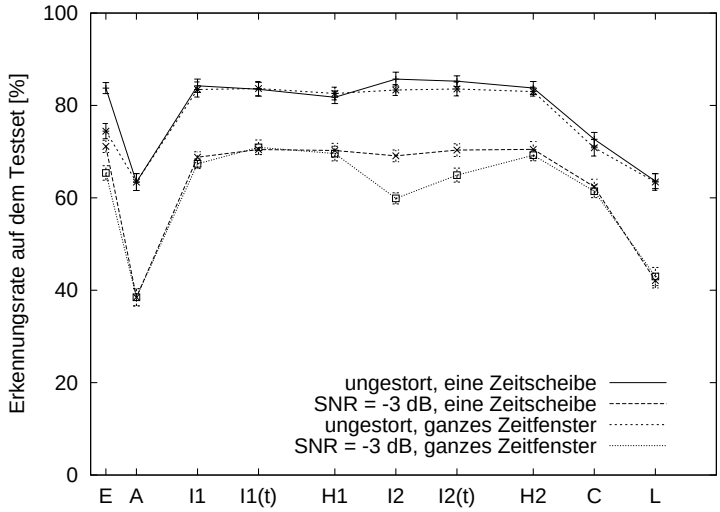

Abbildung 10.7: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für ungestörte Daten sowie mit SNR = -3 dB. Vergleich der Ergebnisse bei Verwendung unterschiedlicher visueller Merkmale. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

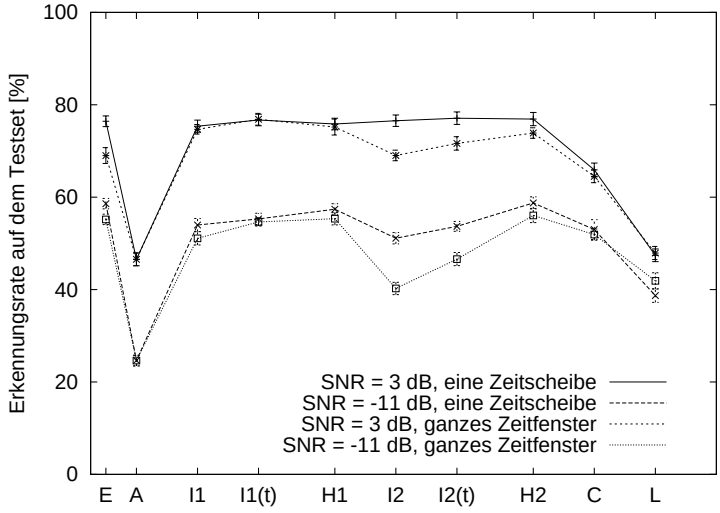

Abbildung 10.8: Klassifikationsergebnisse für die Fusionsmöglichkeit mit gemeinsamer Dimensionsreduktion, dargestellt für gestörte Daten mit SNR = 3 dB und SNR = -11 dB. Vergleich der Ergebnisse bei Verwendung unterschiedlicher visueller Merkmale. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz und als Fehlerbalken die 95%igen Konfidenzintervalle für die einzelnen Architekturen.

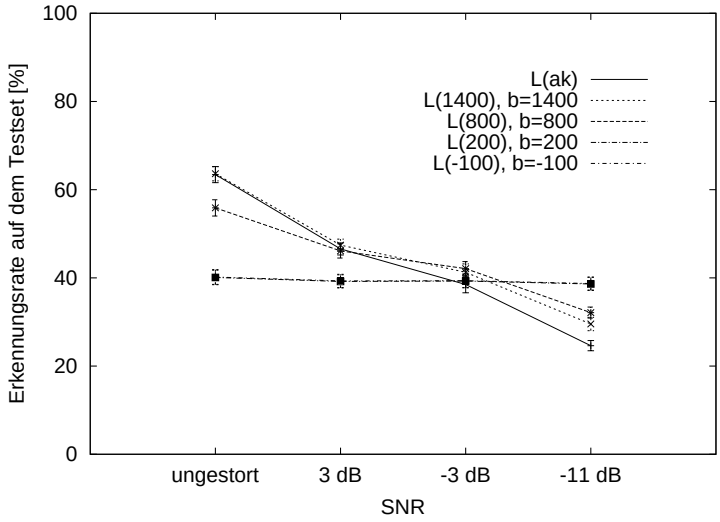

Abbildung 10.9: Ergebnisse mit gemeinsam fusionierten Daten, nur die erste Zeitscheibe wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

se für ungestörte Daten und für gestörte Daten mit einem Signal-zu-Rausch-Verhältnis von -3 dB, zum anderen die Ergebnisse für gestörte Daten mit einem Signal-zu-Rauschverhältnis von 3 dB und -11 dB gemeinsam abgebildet sind. Betrachtet man die Ergebnisse, die für ungestörte akustische Daten erzielt werden, zeigen sich zwischen beiden visuellen Merkmalen keine statistisch signifikanten Unterschiede bis auf Architektur E. Bei einem Signal-zu-Rausch-Verhältnis von 3 dB weisen die Architekturen I2, I2(t), E und H2 schlechtere Leistungen auf, wenn als visuelles Merkmal nur eine Zeitscheibe des Zeitfensters verwendet wird, ansonsten ergeben sich keine Unterschiede. Dieses läßt sich ebenfalls für einen SNR von -3 dB feststellen, nur daß hier auch Architektur H2 für beide Merkmale diesselbe Leistung erbringt, hier sind nur die Architekturen I2, I2(t) und E deutlich schlechter. Auch für ein SNR von -11 dB können diese Tendenzen festgestellt werden, wobei sich die Unterschiede in den Performanzen bei fallendem SNR noch verstärken, nur bei Architektur E ist ein genau umgekehrter Trend zu beobachten. Weiterhin läßt sich feststellen, daß für sinkendes Signal-zu-Rausch-Verhältnis die Performanz der Architektur L bei der Verwendung des gesamten Zeitfensters als visuelles Merkmal immer besser gegenüber der Verwendung nur einer Zeitscheibe des Zeitfensters wird.

10.2 Ergebnisse für die Architektur L

Die Abbildungen 10.9 und 10.10 zeigen die Ergebnisse, die mit Architektur L bei Verwendung verschiedener Schwellwerte erzielt werden, wenn die akustischen Daten unterschiedlicher Störungsgrade gemeinsam fusioniert werden. Es

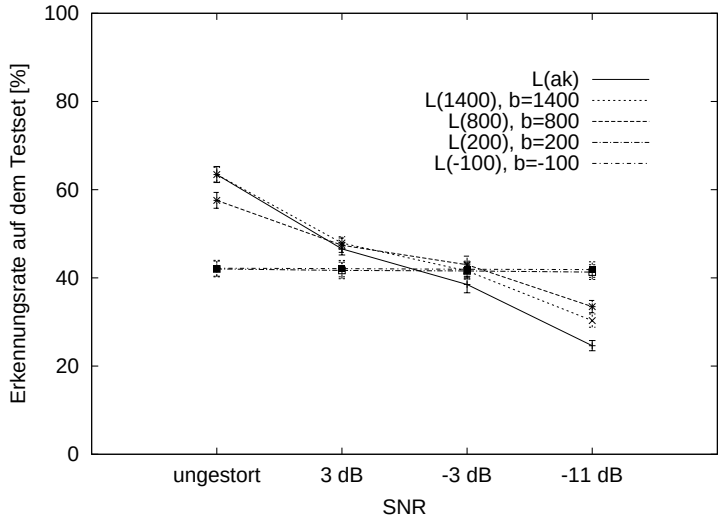

Abbildung 10.10: Ergebnisse mit gemeinsam fusionierten Daten, das ganze Zeitfenster wird als visuelles Merkmal genutzt. Verschiedene Realisierungen der Architektur L, Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

bleibt zu bemerken, daß die Ergebniskurven für die Schwellwerte $b = -100$ und $b = 200$ gleich und daher übereinander gezeichnet sind. Zunächst läßt sich feststellen, daß die Schwellwerte wesentlich höher gewählt werden müssen, um eine unterschiedliche Gewichtung der einzelnen Kanäle zu erreichen. Insgesamt lassen sich dieselben Tendenzen ausmachen, wie sie auch in Kapitel 8 und 9 ermittelt werden konnten, nur daß hier für die Wahl der Schwelle mit b = 1400 immer bessere Ergebnisse erzielt werden, als wenn dem akustischen Kanal der Vorzug gegeben wird (Architektur L(ak)). Für ein SNR von -3 dB ergibt sich für einen Schwellwert von b = 800 die beste Leistung, für ein niedriges Signal-zu-Rausch-Verhältnis von -11 dB ist die Wahl einer niedrigen Schwelle von b = -100 am günstigsten. Diese Ergebnisse sind für beide verwendeten visuellen Merkmale gültig.
Wiederum soll versucht werden, eine Anpassung der Schwellwerte mit Hilfe des Verhältnisses von akustischem und visuellem Hammingabstand zu erreichen (vgl. Kapitel 9.4.3). Tabelle 10.1 zeigt die entsprechenden Mittelwerte.
Es zeigt sich zwar ein Trend zu größeren Verhältniswerten bei fallendem SNR, allerdings sind die Unterschiede zwischen den einzelnen Verhältnissen sehr viel kleiner, als dies bei den Experimenten mit niedrigdimensionalen Daten der Fall war, zudem sind die Standardabweichungen größer. Weiterhin spiegeln die Verhältnisse für den Testdatensatz ungenau diejenigen des Trainingsdatensatzes wieder. Für Versuche mit der logistischen Anpassungsfunktion

$$g(x) = \frac{1500}{1 + e^{-\beta(x-v)}} - 100. \tag{10.1}$$

SNR [dB]	Trainingsdaten	Testdaten	bester Schwellwert b
LDA, gesamt, 1ZS			
ungestört	17.71 (1.58)	17.23 (1.40)	>= 1400
3	18.42 (1.64)	18.55 (1.55)	1400
-3	18.74 (1.67)	19.22 (1.61)	800
-11	19.08 (1.72)	20.76 (1.82)	-100
LDA, gesamt, 4 ZS			
ungestört	9.36 (0.91)	8.97 (0.67)	>= 1400
3	9.73 (0.94)	9.67 (0.83)	1400
-3	9.90 (0.93)	10.02 (0.93)	800
-11	10.08 (0.94)	10.82 (1.02)	-100

Tabelle 10.1: Mittelwerte und Standardabweichungen der Verhältnisse von 'akustischem' Hammingabstand zum 'visuellen': Mittel(ak/vis) zusätzlich ist die Schwelle angegeben, für die das beste Klassifikationsergebnis erzielt wurde.

werden für eine Zeitscheibe pro Zeitfenster als visuelles Merkmal die Parameter $\beta = 10$ und $v = 18.77$ und für das ganze Zeitfenster als visuelles Merkmal die Parameter $\beta = 25$ und $v = 9.915$ empirisch angepaßt. Es zeigt sich allerdings keine weitere Verbesserung der Performanz gegenüber der Verwendung eines Schwellwertes von b = 1400 (vgl. Tabellen C.9 und C.10 in Anhang C).

10.3 Vergleich der Fusionstrategien für dimensionsreduzierte Daten

Abschließend sollen die verschiedenen Fusionsmöglichkeiten für die niedrigdimensionalen Daten miteinander verglichen werden. Neben der gemeinsamen Dimensionsreduktion für alle verrauschten akustischen Daten, für die gemeinsam ein Netz trainiert wird, ist noch die Dimensionsreduktion für jede Rauschstufe einzeln zu nennen, die jeweils mit einem spezialisierten Netz gelernt wird, wie es in Kapitel 9 der Fall war. Auf diesen Fall soll im folgenden mit 'spezialisiertes Netz' Bezug genommen werden. Die einzelnen Ergebnisse sind zum Vergleich nochmals in den Abbildungen 10.11 bis 10.18 dargestellt. Für den Fall, daß die ungestörten akustischen Daten zum Testen der Netze verwendet werden, erkennt man für beide visuellen Merkmale, daß mit dem spezialisierten Netz die statistisch signifikant besten Ergebnisse erzielt werden (vgl. Abbildungen 10.11 und 10.15).
Bei Verwendung von akustischen Daten mit einem SNR von 3 dB erhält man wiederum für beide visuellen Merkmale einen sehr ähnlichen Verlauf. Betrachtet man alle Fusionsarchitekturen, so erweist sich die gemeinsame Fusion als für alle Architekturen in etwa gleich gutes Verfahren, wobei die Benutzung von spezialisierten Netzen nur für die Architektur H einen Vorteil bringt, welche dann auch bei Verwendung einer Zeitscheibe pro Zeitfenster als visuelles Merkmal das mit Architektur E zusammen insgesamt beste Ergebnis erzielen. Die Architektur E für eine gemeinsame Fusion erzielt ein ähnlich gutes Er-

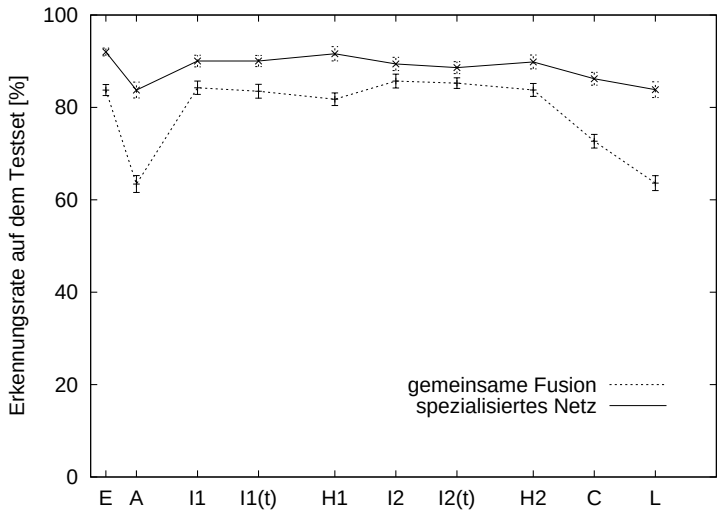

Abbildung 10.11: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Dimensionsreduktion und spezialisierte Netze. Visuelles Merkmal ist eine Zeitscheibe des Zeitfensters, der akustische Kanal ist ungestört. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

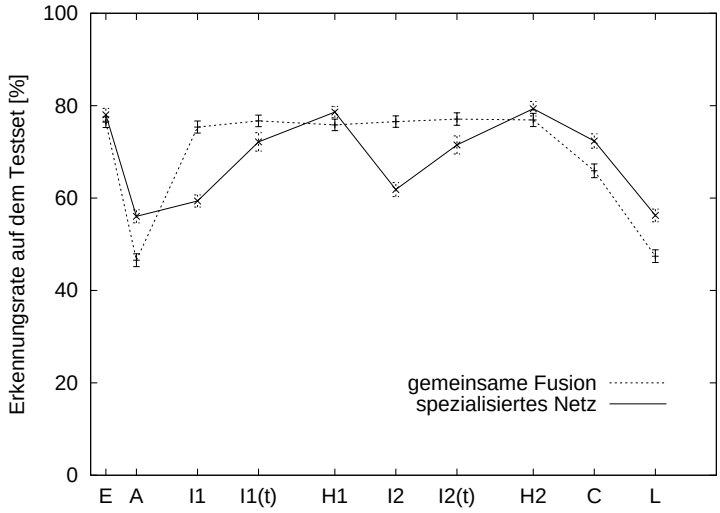

Abbildung 10.12: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist eine Zeitscheibe des Zeitfensters, der akustische Kanal ist gestört mit SNR = 3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

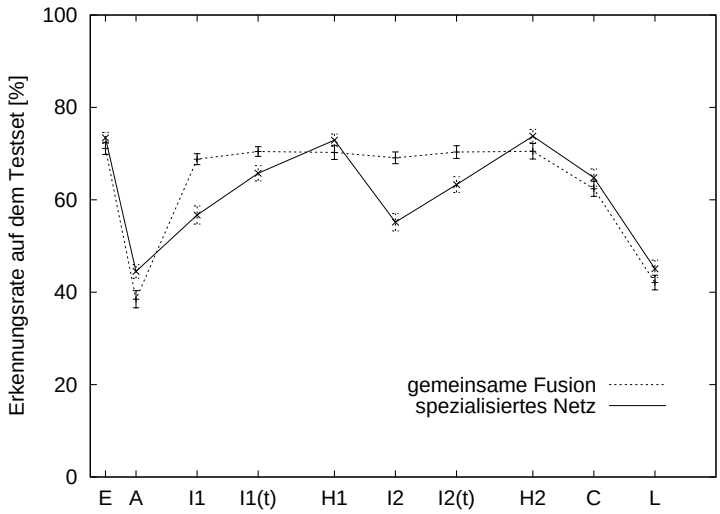

Abbildung 10.13: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist eine Zeitscheibe des Zeitfensters, der akustische Kanal ist gestört mit SNR = -3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

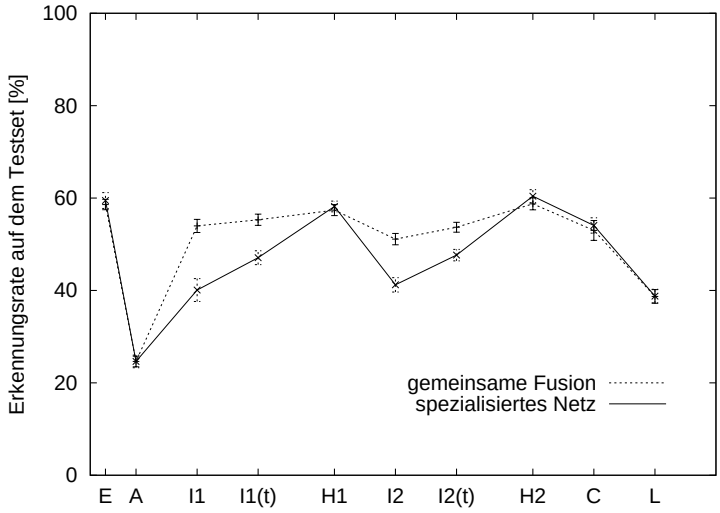

Abbildung 10.14: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist eine Zeitscheibe des Zeitfensters, der akustische Kanal ist gestört mit SNR = -11 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

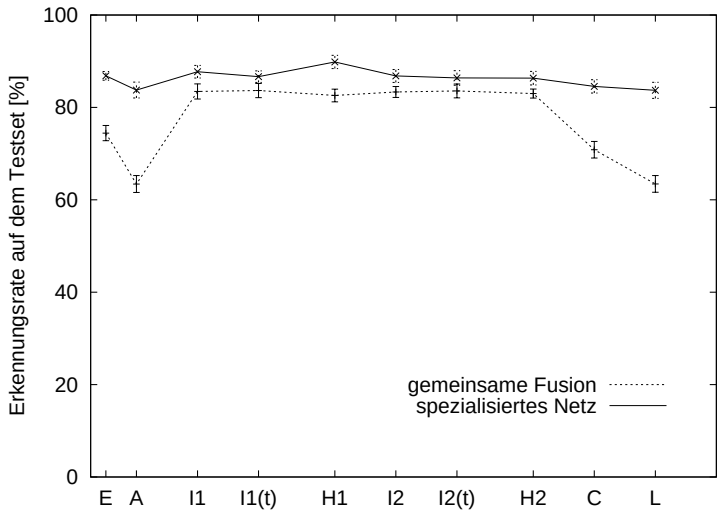

Abbildung 10.15: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist das ganze Zeitfenster, der akustische Kanal ist ungestört. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

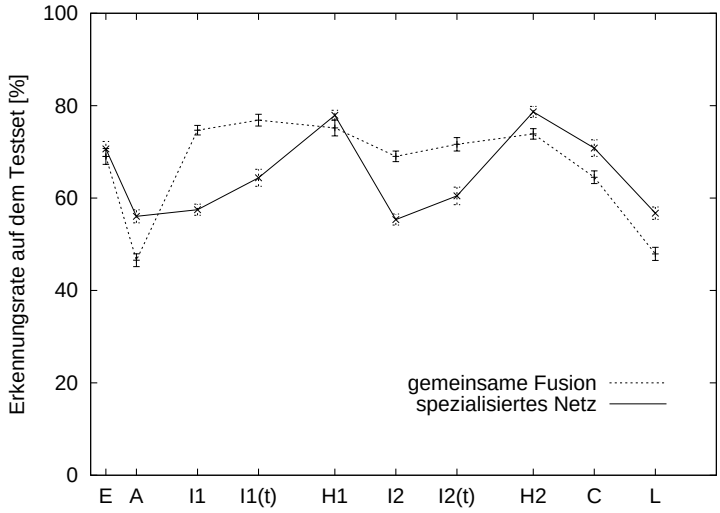

Abbildung 10.16: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist das ganze Zeitfenster, der akustische Kanal ist gestört mit SNR = 3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

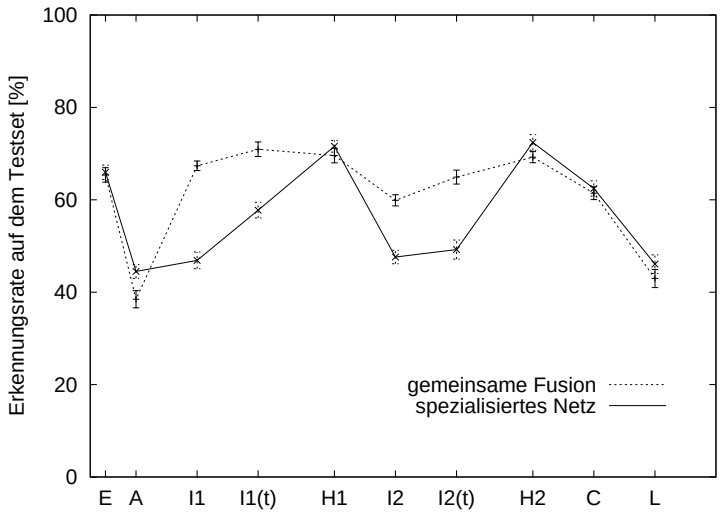

Abbildung 10.17: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist das ganze Zeitfenster, der akustische Kanal ist gestört mit SNR = -3 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

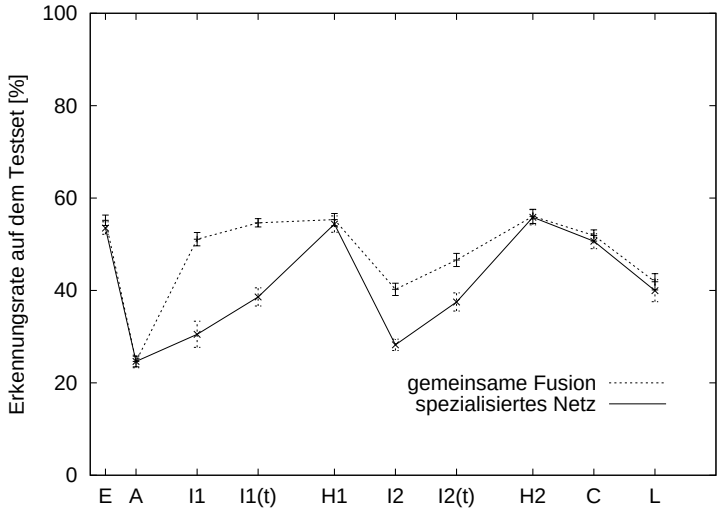

Abbildung 10.18: Vergleich der Ergebnisse für verschiedene Fusionsarten: Gemeinsame Fusion und spezialisierte Netze. Visuelles Merkmal ist das ganze Zeitfenster, der akustische Kanal ist gestört mit SNR = -11 dB. Ergebnisse als klassenspezifische Mittel auf dem Testdatensatz. Dargestellt ist die mittlere Performanz, als Fehlerbalken die 95%igen Konfidenzintervalle.

gebnis. Wiederum zeigt die gemeinsame Fusion für alle Fusionsarchitekturen ein insgesamt ausgeglicheneres Bild, bei den spezialisierten Netzen ist die gute Performanz der Architektur H dagegen ein Sonderfall. Bei den späten Fusionsarchitekturen C und L erweisen sich die spezialisierten Netze als die beste Lösung. Für den Fall, daß das ganze Zeitfenster als visuelles Merkmal verwendet wird, ist die herausgehobene Position der Architektur H der spezialisierten Netze noch deutlicher, während die Performanz der Architektur E für beide Fusionsmöglichkeiten abfällt.
Werden akustische Daten mit einem SNR von -3 dB zum Testen verwendet, erkennt man wiederum diesselben Tendenzen wie im vorangegangenen Fall, für die Verwendung nur der akustischen Daten (Architektur A) ist das spezialisierte Netz weiterhin besser. Die besten Ergebnisse werden von den Architekturen H der spezialisierten Netze erbracht. Für den Fall, daß nur eine Zeitscheibe als visuelles Merkmal verwendet wird, sind diese auch in etwa gleich demjenigen für die Architektur E. Für die späten Fusionsarchitekturen C und L erweist sich auch das spezialisierte Netz als am geeignetsten. Alle Unterschiede für die späten Fusionsarchitekturen sind allerdings nicht statistisch signifikant. Wird das ganze Zeitfenster als visuelles Merkmal verwendet, werden die besten Ergebnisse von den spezialisierten H-Architekturen gemeinsam mit der Architektur I1(t) und H für die gemeinsame Fusion erzielt. Bei der frühen Architektur E sind die Ergebnisse jeweils für beide Fusionsarten für beide visuellen Merkmale fast identisch.
Werden schließlich stark gestörte akustische Daten für das Testen der Netze verwendet, zeigt sich im Fall, daß nur eine Zeitscheibe als visuelles Merkmal verwendet wird, folgendes Bild: Die Performanz für den akustischen Kanal alleine ist für beide Möglichkeiten gleich gut. Die Klassifikationsleistungen für die Architektur E sind für beide Fusionsmöglichkeiten nahezu gleich, auch bei den späten Fusionsarchitekturen C und L sowie bei Architektur H2 lassen sich kaum noch Unterschiede wahrnehmen. Die Architekturen H erbringen eine fast gleich gute Performanz wie Architektur E. Falls frühe Fusionsarchitekturen I verwendet werden sollen, liefert die gemeinsame Fusion die besten Ergebnisse. Dies gilt auch für alle Signal-zu-Rausch-Verhältnisse, bis auf die ungestörten Daten.
Verwendet man als visuelles Merkmal das ganze Zeitfenster ergibt sich ein sehr ähnliches Bild: Die Performanzen der Architekturen E und H sind nahezu gleich gut für beide Fusionsarten, bei den späten Fusionsmöglichkeiten C und L liefert hingegen zum ersten Mal die gemeinsame Fusionsart die leicht besseren Ergebnisse. Dieser Unterschied ist aber nicht statistisch signifikant. Bei Verwendung der frühen Architektur I ist wiederum die gemeinsame Fusion vorzuziehen. Für den akustischen Kanal alleine erbringen die gemeinsame Fusion und die spezialisierten Netze eine gleich gute Leistung. Insgesamt läßt sich als Tendenz feststellen, daß sowohl bei der Verwendung nur der akustischen Daten als auch für die Architekturen E, C und L die Unterschiede zwischen den Fusionsmöglichkeiten immer geringer werden, je kleiner das Signal-zu-Rausch-Verhältnis ist.

10.4 Kapitelzusammenfassung

In diesem Abschnitt wird die Fusion durch gemeinsame Dimensionsreduktion auf die verschieden gestörten akustischen Daten ausgedehnt, indem die LDA für alle Verrauschungsstufen gemeinsam bestimmt wird.
Erneut liefert die Verwendung des Hammingabstandes als Entscheidungskriterium des Assoziativspeichers die besten Ergebnisse. Es zeigt sich wiederum, daß alle Fusionsarchitekturen außer Architektur L eine bessere Performanz leisten, als bei Verwendung ausschließlich akustischer Daten zu erreichen ist. Architektur L übertrifft die Leistung für den akustischen Kanal nur für ein kleines Signal-zu-Rausch-Verhältnis.
Für die übrigen Fusionsarchitekturen zeigt sich, daß diese alle in etwa gleich gut sind, bis auf die späten Fusionsarchitekturen C und L, bei Verwendung des ganzen Zeitfensters als visuelles Merkmal fällt auch Architektur E in der Performanz ab. Bei sinkendem Signal-zu-Rausch-Verhältnis zeigt sich dagegen, daß die Architekturen H und E eine im Vergleich zu den übrigen Architekturen immer bessere Performanz bieten. Architektur E ist im Fall, daß das ganze Zeitfenster als visuelles Merkmal verwendet wird, gegenüber der besten Architektur umso schlechter, je größer das Signal-zu-Rausch-Verhältnis ist. Eine Anpassung der für Architektur L verwendeten Schwellwerte mittels einer logistischen Funktion ist in diesem Fall nicht erfolgreicher als die Verwendung eines festen Schwellwerts.
Ein Vergleich der Leistungen für unterschiedliche visuelle Merkmale zeigt, daß bis auf Architektur E alle Fusionsarchitekturen in etwa gleiche Leistungen zeigen, bei gestörten akustischen Daten steigt der Performanzunterschied bei Verwendung des ganzen Zeitfensters als visuelles Merkmal gegenüber dem visuellen Merkmal, nur eine Zeitscheibe zu verwenden, für die Architektur I2 für sinkendes Signal-zu-Rausch-Verhältnis an, für Architektur E sinkt er hingegen ab.
Ein abschließender Vergleich der gemeinsamen Dimensionsreduktion für alle akustischen Daten mit dem Verfahren, jedes Signal-zu-Rausch-Verhältnis von einem spezialisierten Netz lernen zu lassen, zeigt, daß für ungestörte akustische Daten die spezialisierten Netze immer die beste Performanz zeigen, die Ergebnisse für eine gemeinsame Reduktion sind dagegen ausgeglichener. Bei gestörten Daten sind bei Verwendung nur einer Zeitscheibe pro Zeitfenster die spezialisierten Architekturen E und H am besten, die gemeinsame Reduktion liefert fast ebenso gute Ergebnisse wie Architektur H, für stark gestörte akustische Daten sind beide Fusionsmöglichkeiten nahezu gleich, außer für die Architektur I. Wird das ganze Zeitfenster als visuelles Merkmal verwendet, ergeben sich ähnliche Tendenzen, nur Architektur E ist auszunehmen, diese liefert nun erst bei einem Signal-zu-Rausch-Verhältnis von -11 dB mit das beste Ergebnis. Tendenziell verschwinden die Unterschiede zwischen den Fusionsmöglichkeiten bei sinkendem Signal-zu-Rausch-Verhältnis für die Architekturen A, E, C und L.

Kapitel 11

Zusammenfassung und Diskussion

Ziel der vorliegenden Arbeit ist es, die Sensorfusionsproblematik in der audio-visuellen Spracherkennung näher zu untersuchen, die bei der Kombination des akustischen mit dem visuellen Kanal auftritt.
Dazu wird zunächst der Prozeß der audio-visuellen Sprachverarbeitung beim Menschen näher betrachtet. Es werden die wichtigsten Ergebnisse psycho--physischer Experimente aus der Literatur zu dem Thema der Integration vorgestellt (vgl. Kapitel 4.7). Dabei werden eine frühe und eine späte Fusionsmöglichkeit gegeneinander abgewogen, wobei bei der frühen Fusion beide Informationen zusammengeführt werden, bevor eine phonetische Verarbeitung des Signals erfolgt. Bei einer späten Fusion erfolgt diese Zusammenführung nachdem eine linguistische Repräsentation für beide Kanäle bestimmt wurde. Es zeigt sich, daß die Mehrzahl der Ergebnisse für eine frühe Fusion sprechen.
Die Sensorfusion in automatischen audio-visuellen Spracherkennern nimmt den zweiten Teil der Arbeit ein. Dabei wird neben der frühen (Integration vor Eingabe in den Spracherkenner) und späten Fusion (Integration nach Einzelklassifikation der Kanäle) auch eine Fusion auf einer Zwischenebene betrachtet, bei der die Kanäle innerhalb des Erkenners integriert werden, bevor eine Klassifikation erfolgt. Ein Literaturüberblick über vorhandene Systeme (vgl. Kapitel 5) zeigt, daß nur in wenigen Fällen alle Fusionsmöglichkeiten miteinander verglichen werden. Dies liegt zum Teil daran, daß die meisten audio-visuellen Spracherkenner auf Hidden-Markov-Modellen basieren, bei denen eine Fusion auf der Zwischenstufe nicht einfach zu realisieren ist. Weiterhin fehlen bei den meisten Arbeiten Angaben darüber, ob sich die untersuchten Architekturen statistisch signifikant voneinander unterscheiden oder ob ein Unterschied zwischen zwei Architekturen z.B. durch eine spezielle Auswahl der Trainings- und Testdaten zustande kommen könnte.
In der vorliegenden Arbeit wird daher ein neues audio-visuelles Erkennungssystem entwickelt, welches aus neuronalen Modulen zusammengesetzt wird. Verwendet werden Daten sogenannter Buchstabenwörter. Der neuronale Aufbau ermöglicht die Realisierung einer großen Anzahl an Fusionsmöglichkeiten. Das erste Modul dient der Klassifikation eines Zeitfensters, das über das Signal ge-

schoben wird, zu einer von 30 definierten Segmentklassen. In diesem Modul wird ein Multilayer-Perzeptron als Klassifikator eingesetzt. Als Ausgabe liefert dieses Modul eine Sequenz von Segmentklassifikationen, die je nach Signallänge unterschiedlich lang sein kann. Die anschließende Zeitintegration dieser Sequenz wird in einem zweiten Modul mit Hilfe einer Tripelkodierung und des Einsatzes eines Assoziativspeichers erreicht. Zum Training des Systems ist nur die Labelung des jeweils ersten und letzten Zeitfensters eines Buchstabenwortes zu einer der 30 Segmentklassen, sowie die Labelung des gesamten Buchstabenwortes notwendig, was mit minimalem Aufwand durchgeführt werden kann. Bei der Erkennung eines Buchstabenwortes ist das System trotzdem in der Lage, über alle dazwischenliegenden Zeitfenster zu generalisieren.
Die verwendeten akustischen Daten werden mit Hilfe einer Fouriertransformation und einer Filterbank, die auf der Bark-Skala beruht, weiterverarbeitet. Es werden weiterhin gestörte akustische Daten mit unterschiedlichen Signal-zu-Rauschverhältnissen betrachtet. Die visuellen Daten werden durch Grauwert-Halbbilder der Lippenregion gebildet.
Als visuelle Merkmale werden sowohl die Grauwertbilder des ganzen Zeitfensters als auch nur das zur ersten Zeitscheibe des Zeitfensters gehörende Grauwertbild betrachtet. Weiterhin werden nicht nur die ursprünglich vorliegenden hochdimensionalen Daten für die Experimente verwendet, sondern auch mit Hilfe einer LDA dimensionsreduzierte Daten.
Um statistische Aussagen über die Performanz der untersuchten Fusionsarchitekturen zu erhalten, werden Crossvalidierungsversuche durchgeführt, wobei sowohl unterschiedliche Aufteilungen des Datensatzes in Trainings- und Testdaten als auch verschiedene Initialisierungen des Multilayer-Perzeptrons betrachtet werden.

Mit Hilfe des Worterkennungssystems werden Fusionsarchitekturen auf jeder möglichen Stufe der Informationsverarbeitung realisiert. Dazu werden zunächst Netze für die Einzelkanäle optimiert, auf denen aufbauend anschließend für die Verarbeitung der kombinierten akustischen und visuellen Daten unterschiedliche Fusionsarchitekturen entwickelt werden. Der Aufbau der einzelnen Fusionsarchitekturen und deren Bezeichnungen können den Abbildungen 8.1 und 9.8 entnommen werden.
Es lassen sich dabei folgende Fusionsstufen unterscheiden: Bei einer frühen Fusion findet die Zusammenführung beider Informationskanäle statt, bevor diese in das Worterkennungssystem eingegeben werden. Dies kann vor einer Dimensionsreduktion erfolgen (Architektur E), oder durch einfache Konkatenation der beiden Datenströme (Architektur I). Bei einer Fusion auf einer Zwischenstufe erfolgt die Zusammenführung innerhalb des Worterkennungssystems, aber noch vor einer Klassifikation (Architektur H). Schließend können beide Informationskanäle auch auf einem späten Level erfolgen, d.h. nachdem eine Klassifikation der Kanäle separat stattgefunden hat. Geschieht dies auf der Ebene der Segmentklassifikationen, wird die entsprechende Architektur mit C bezeichnet. Für die Zusammenführung nach der Buchstabenworterkennung durch den Assoziativspeicher (Architektur L) kann eine Gewichtung der Kanäle mit einer Fusionsstrategie erreicht werden, die das beim Assoziativspeicher verwendete

Entscheidungskriterium benutzt. Diese Gewichtung kann automatisch erfolgen, ohne daß das Signal-Rausch-Verhältnis des akustischen Kanals bekannt sein muß (vgl. Kapitel 9.4.3).
Die Fusion wird in einem abschließenden Experiment auch auf die Fusion unterschiedlich stark gestörter akustischer Daten ausgedehnt (gemeinsame Fusion).

Die bezüglich des Trainings und des Aufbaus des Worterkennungssystems erzielten Ergebnisse sollen nun stichpunktartig zusammengefaßt werden:

- Durch Versuche mit akustischen Daten wurde festgestellt, daß für das Testen des Systems mit gestörten Daten dieses auch mit Daten einer möglichst ähnlichen Verrauschungsstufe trainiert werden sollte.

- Die Dimensionsreduktion der Daten mit Hilfe einer LDA führt zu mindestens gleich guten Klassifikationsleistungen, im Fall von stark gestörten akustischen Daten führt sie zu einer Verbesserung, wodurch der Nutzen der LDA bei der Vorverarbeitung bestätigt werden konnte (vgl. [PG98])

- Bei der Optimierung der Netze für die Klassifikation der Einzelkanäle zeigt sich, daß im Falle der dimensionsreduzierten Daten, für die die Anzahl der Gewichte der Netze wesentlich geringer als im Falle der ursprünglichen Daten ist, auch bei einer stark wachsenden Anzahl an versteckten Neuronen kein Overfitting-Effekt auftritt. Dies bestätigt Untersuchungen von Lawrence et al [LGT97]

- Bei der sehr späten Fusion (Architektur L) ist eine Gewichtung der Kanäle mittels der bei der Klassifikation mit den Assoziativspeichern berechneten Hammingabstände möglich, indem ein Schwellwert für den Vergleich der Hammingabstände verwendet wird. Die Anpassung dieses Schwellwertes kann unter Umständen ohne Informationen über den SNR der akustischen Daten erfolgen, indem das Verhältnis des akustischen zum visuellen Hammingabstand verwendet wird. Dieses neue Verfahren implementiert damit eine automatische Gewichtung ohne Kenntnis des SNR, was in der Literatur sonst nur unter Verwendung der Dispersion gelingt (vgl. 89). Für die anderen Fusionsmöglichkeiten (bis auf Architektur E) kann eine Gewichtung der Kanäle von den neuronalen Netzen gelernt werden.

- Als Entscheidungskriterium des Assoziativspeichers ist der Hammingabstand dem Überlapp vorzuziehen, da er bessere Ergebnisse erzielt.

- Die Leistung des Systems ist auf der Ebene der Buchstabenwortklassifikation immer größer als auf der Ebene der Segmentklassifikation, da eventuell auftretende systematische Fehler des ersten Moduls (z.B. Verwechslungen innerhalb einer Visemklasse für den visuellen Kanal) durch das zweite Modul gelernt und ausgeglichen werden können.

Aus dem Vergleich verschiedener Fusionsarchitekturen lassen sich folgende Ergebnisse gewinnen:

- Die Verwendung des visuellen Kanals verbessert die Klassifikationsleistung des Worterkennungssystems insbesondere bei stark gestörten akustischen Daten. Dies gilt für fast alle Fusionsarchitekturen, bis auf Architektur L, die nur bei stark gestörtem akustischen Kanal eine Verbesserung der Leistung erbringt und in manchen Fällen bis auf Architektur I.

- Werden verschiedene Fusionsarchitekturen miteinander verglichen, so ist es notwendig, die Anzahl der Gewichte der Architekturen in etwa gleich groß zu halten, da eine zu große Gewichtsanzahl zu einer erheblich schlechteren Klassifikationsleistung führen kann, ohne daß die Architektur selbst ungeeignet wäre. Dies läßt sich bei Versuchen mit den Originaldaten feststellen.

- Bei Verwendung des ganzes Zeitfensters als visuelles Merkmal ergibt sich für den visuellen Kanal alleine eine bessere Klassifikationsleistung als bei Verwendung nur einer Zeitscheibe des Zeitfensters. Bei den Fusionsarchitekturen ergeben sich dagegen keine Leistungsverbesserungen für die Architekturen H, C und L. Für die Architekturen I und E sind die Leistungen schlechter.
  Bei der Konstruktion eines Sensorfusionssystems sollte daher beachtet werden, daß eine Zusammenführung der besten Einzelmerkmale nicht zwingendermaßen zum besten Fusionssystem führt.

- Bei einer gemeinsamen Dimensionsreduktion aller verrauschten akustischen Daten (gemeinsame Fusion) ist die Klassifikationsleistung weniger architekturabhängig und zeigt für die meisten Fusionsarchitekturen eine gleichmäßig gute Leistung. Diese reicht im Fall von gestörten akustischen Daten an diejenige der spezialisierten Netze (Architektur H) heran, die nur auf ein SNR hin trainiert werden. Im Fall stark gestörter akustischer Daten sind beide Möglichkeiten annähernd gleich gut. Für ungestörte Daten sind die spezialisierten Netze besser.

- Architektur E, die nur mit Hilfe einer Dimensionsreduktion der Daten realisiert werden kann, zeigt je nach visuellem Merkmal eine unterschiedliche Performanz: Bei Verwendung nur einer Zeitscheibe pro Zeitfenster gehört sie zu den besten Architekturen, bei Verwendung des ganzen Zeitfensters fällt die Leistung dagegen ab, nur bei stark gestörten akustischen Daten ist sie wieder in etwa gleich gut wie die beste Architektur.

- Architektur H gehört in allen Fällen zu den besten Fusionsarchitekturen, sie wird nur in sehr wenigen Einzelfällen von Architektur E übertroffen, welche aber unter Umständen auf besonders gute Merkmale aufgrund der gemeinsamen Vorverarbeitung beider Kanäle zurückgreifen kann. Bei Verwendung derselben Merkmale ist Architektur H am besten. Die schlechtesten Leistungen erbringt Architektur L, gemeinsam mit Architektur C

im ungestörten Fall sowie Architektur I im Fall gestörter akustischer Daten. Bei gestörten Daten ist C (zum Teil zusammen mit I(t)) dagegen die zweitbeste Architektur. Architektur I(t) ist der Architektur I aufgrund der geringen Gewichtsanzahl überlegen. Architektur C hat bei der Fusion noch mehr Informationen zur Verfügung als Architektur L, da die Eingaben in das Kombinationsnetz die a posteriori Wahrscheinlichkeiten der Segmentklassenzuordnungen sind und keine kategorischen Klassenzuordnungen.

Zusammenfassend läßt sich als allgemeine Rangfolge vereinfacht
H > E,C,I(t) > I,L angeben. Damit erbringt eine möglichst späte Fusion vor einer Klassifikation die besten Ergebnisse (Fusion auf der Zwischenstufe). Nicht ganz so gute Ergebnisse werden mit Hilfe von frühen Fusionsarchitekturen nur erzielt, wenn diese entweder eine reduzierte Gewichtsanzahl haben oder auf eine gemeinsame Vorverarbeitung beider Kanäle zurückgreifen können (I(t),E > I). Für späte Fusionsarchitekturen werden ebenfalls schlechtere Ergebnisse erzielt. Dabei ist Architektur C der Architektur L überlegen, da sie bei der Fusion noch auf mehr Informationen zurückgreifen kann.
Die beste Performanz der Architektur H deckt sich mit den Ergebnissen aus psychophysischen Experimenten, die eine präkategorielle Fusion bevorzugen.

Abschließend soll die hier festgestellte Reihenfolge der Fusionsarchitekturen mit denen anderer Untersuchungen verglichen werden. Ein Vergleich einzelner Arbeiten gestaltet sich dabei schwierig, da die einzelnen Aufgabenstellungen (verwendete Datensätze, Sprecherabhängigkeit, Trainingsparadigma, Auswahl der akustischen und visuellen Merkmale,...) voneinander abweichen. Auch ist eine statistische Aussage nicht in allen Fällen gegeben.
Bei allen bisherigen Untersuchungen mit automatischen Spracherkennern zeigte die frühe Fusion nie die besten Ergebnisse. Die Fusion auf einer Zwischenstufe sowie die späte Fusion zeigen in etwa gleich häufig die besten Ergebnisse, sowohl im Vergleich mit der frühen Fusion als auch in den Vergleichen mit allen Fusionsmöglichkeiten.
Bei einem direkten Vergleich von späten Fusionsarchitekturen und Fusionen auf einer Zwischenstufe ist zu beachten, daß in vielen Fällen bei der späten Fusion eine Gewichtung der Einzelkanäle durchgeführt wird, was nicht immer auch für die Fusion auf einer Zwischenstufe der Fall ist. Nur in Fällen, wo für beide Architekturen dieselben Bedingungen vorgegeben waren, ist ein echter Vergleich möglich, d.h. Architektur L soll in dieser Diskussion ausgeschlossen werden.
Im Hinblick darauf bleiben somit noch vier Arbeiten zu diskutieren: Stork et al. [SWL92] verwendeten neuronale Netze als Klassifikatoren. Sie erzielten mit einer späten Fusionsarchitektur die besten Ergebnisse. Es bleibt allerdings anzumerken, daß die Anzahl der Gewichte in den einzelnen Realisierungen der Fusionsarchitekturen unterschiedlich groß war. Für die späte Fusion waren weniger Gewichte notwendig, als für die Fusion auf einer Zwischenebene, so daß das Ergebnis auf diesen Unterschied zurückgeführt werden könnte.
Robert-Ribes et al. [RRLES93], [RRSE95a], [RRPSE96], [RRSE95b] konnten keiner der Architekturen einen eindeutigen Vorzug geben.
Bei den Ergebnissen von Teissier et al. [TRRSGD99], [TGDS98] fällt auf, daß

die frühe Fusionsarchitektur zusammen mit der späten die besten Ergebnisse erbringt. Bei der Fusion auf einer Zwischenstufe werden hingegen beide Kanäle auf einen gemeinsamen Merkmalsraum abgebildet, welcher unter Umständen ungünstig gewählt wurde und daher zu einem schlechten Resultat führte.
Kumatani erzielte mit der Fusion auf einer Zwischenstufe die besten Ergebnisse, wenngleich dies auch nur für die meisten der untersuchten Signal-Rausch-Abstände der Fall war.
Damit können diese Ergebnisse in Einklang mit der in der vorliegenden Arbeit ermittelten Reihenfolge gebracht werden [1].

In dieser Arbeit wurde die theoretisch und praktisch sehr wichtige Problematik der Sensorfusion anhand eines neuartigen audio-visuellen Worterkennungssystems untersucht. Dabei wurde die Realisierung und systematische Untersuchung aller möglichen Fusionsarchitekturen zum Ziel genommen, was bei der Entwicklung des Worterkennungssystems berücksichtigt werden mußte. Das Worterkennungssystem ist daher vollständig neuronal aufgebaut und erlaubt im Gegensatz zu herkömmlichen Verfahren (z.B. Hidden-Markov-Modellen) die Untersuchung von Fusionsarchitekturen auf allen Stufen der Informationsverarbeitung. Auch die Realisierung und Bewertung einer Fusion auf einer Zwischenebene innerhalb des Systems ist leicht möglich. Obwohl die Parameter des Worterkennungssystems nicht explizit auf die bestmögliche Erkennungsleistung optimiert werden, ist die Performanz ähnlich gut wie die anderer Systeme. Zudem kann bei diesem Worterkennungssystem auf eine Segmentlabelung der Daten, die sehr aufwendig und nicht unproblematisch ist, aufgrund des geschickt gewählten Aufbaus verzichtet werden. Dennoch stand auch bei der vorliegenden Untersuchung nur eine vergleichsweise geringe Datenmenge zur Verfügung, was eine gründliche Behandlung der Problematik erschwert. Da es bisher keine frei verfügbare große audio-visuelle Datenbasis gibt, trifft dieser Umstand auch alle anderen Autoren, so daß ein direkter Vergleich unterschiedlicher Untersuchungen nicht möglich ist. Allerdings kann gesagt werden, daß in den bisherigen Arbeiten ein Vergleich der Fusionsmöglichkeiten nicht methodisch zufriedenstellend untersucht worden konnte, da entweder eine statistische Absicherung der Ergebnisse fehlte, oder bei weitem nicht alle Fusionsmöglichkeiten in Betracht gezogen wurden. Diese Lücke wird durch die vorliegende Arbeit geschlossen.

[1] Auch Hennecke et al. [HSP95] [Hen97] konnten mit der Fusion auf einer Zwischenstufe die besten Ergebnisse erreichen, obwohl dort keine Gewichtung der Kanäle durchgeführt wurde, wie dies für die späte Fusionsarchitektur der Fall war.
Dasselbe läßt sich über die Arbeit von Duchnowski et al. [DMW94] sagen. Hier kann erst bei der Wahl einer ausgefeilteren Gewichtung bei der späten Fusion (Arbeit von Meier et al. [MSYW99]) diese ein besseres Ergebnis erzielen, wobei anzumerken ist, daß in dieser Arbeit keinerlei Validierungsdaten zum Einsatz kamen.

Kapitel 12

Ausblick

Das entwickelte audio-visuelle Worterkennungssystem bietet einige Erweiterungsmöglichkeiten. Zunächst kann es auf Wortebene erweitert werden, so daß buchstabierte Wörter erkannt werden. Dies kann implementiert werden, indem ein zweiter Assoziativspeicher auf das zweite Modul aufgesetzt wird. Für akustische Daten wurde dieses System im Rahmen des Landesforschungsschwerpunktes auch schon realisiert und ergab eine gute Performanz.
Interessanter wäre eine Erweiterung des Systems auf Wortebene, bei der Wörter die Eingabe des Systems bilden. Dazu ist allerdings eine aufwendigere und feinere Labelung der Daten auf Phonebene notwendig, wozu z.B. ein Clusterungsverfahren genutzt werden könnte.
Der Aufbau des Worterkennungssystem erlaubt den Einsatz auch anderer Klassifikationsverfahren (z.B. Gaußklassifikatoren, etc.) im ersten Modul, so daß die Anzahl der Parameter des Systems verkleinert und der damit verbundene Aufwand beim Training des Systems verringert werden könnte.
Bezüglich der audio-visuellen Erkennung erhebt sich die Frage, ob der Gebrauch von Visemen für die Performanz des Systems vorteilhaft ist, da dadurch eventuell die Leistung des visuellen Kanals und damit des ganzen Systems gesteigert werden könnte.
Weiterhin wäre ein Vergleich verschiedener Gewichtungsverfahren von großem Interesse. So könnte durch den Gebrauch der Dispersion als Wichtungsfaktor z.B. die Performanz der späten Architektur C gesteigert werden.

Anhang A

Ergebnisse für Originaldaten

In diesem Anhang sind alle Ergebnisse aufgelistet, die mit den hochdimensionalen Originaldaten erzielt wurden. Aufgetragen sind die Trainings- und Testfehler sowie die klassenspezifischen Mittel auf Trainings- und Testdatensatz für die einzelnen Architekturen sowohl auf Segmentklassifikationsebene als auch auf der Ebene der Buchstabenwortklassifikationen. In Klammern sind jeweils die 95%igen Konfidenzintervalle angegeben. Für jede Architektur ist die Bezeichnung, die Anzahl der Gewichte sowie die Topologie dargestellt. Die Tabellen für die Klassifikationsleistungen auf Buchstabenwortebene geben für jede Architektur in der ersten Zeile die Ergebnisse wieder für den Fall, daß der Hammingabstand als Entscheidungskriterium des Assoziativspeichers verwendet wird, die zweite Zeile enthält die Ergebnisse, die mit dem Überlapp als Entscheidungskriterium erreicht werden.

Netz (Anzahl der Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A (1900)	160-10-30	77.09 (4.13)	43.95 (7.80)	73.16 (3.43)	38.31 (6.01)
A (3800)	160-20-30	94.27 (0.62)	85.70 (1.82)	85.16 (0.54)	67.28 (1.38)
A (5700)	160-30-30	94.62 (0.65)	87.04 (1.64)	85.67 (0.65)	68.57 (1.35)
A (7600)	160-40-30	94.51 (0.46)	86.85 (1.42)	85.86 (0.69)	68.43 (1.37)
Erkennungsleistung auf Buchstabenwortebene					
A (1900)	160-10-30	75.78 (5.92)	75.66 (6.00)	64.91 (5.50)	61.35 (5.50)
		58.44 (8.85)	60.92 (8.23)	47.84 (7.85)	46.73 (7.12)
A (3800)	160-20-30	95.85 (0.99)	95.56 (1.01)	83.68 (1.40)	80.70 (1.63)
		84.23 (3.79)	87.82 (2.40)	69.51 (3.76)	70.00 (2.42)
A (5700)	160-30-30	96.04 (0.97)	96.34 (1.00)	84.46 (1.50)	82.19 (1.49)
		84.45 (3.16)	88.75 (2.24)	71.44 (3.38)	72.61 (2.31)
A (7600)	160-40-30	95.62 (1.10)	95.80 (1.00)	84.48 (1.79)	81.64 (1.87)
		82.57 (3.23)	87.19 (2.00)	69.50 (3.11)	70.56 (2.14)

Tabelle A.1: Klassifikationsleistungen für die akustischen Daten.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A (1380)	160[mod10]-30-30	96.84 (0.18)	92.59 (0.47)	86.92 (0.61)	71.14 (1.14)
A (1380)	160[ranmod10]-30-30	96.49 (0.16)	92.00 (0.37)	87.27 (0.42)	72.14 (0.81)
A (1280)	160[mod16]-32-30	96.25 (0.17)	92.72 (0.38)	86.46 (0.51)	72.79 (1.31)
A (1840)	160[mod10]-40-30	97.67 (0.15)	94.57 (0.42)	88.49 (0.59)	74.48 (0.80)
Erkennungsleistung auf Buchstabenwortebene					
A (1380)	160[mod10]-30-30	98.80 (0.39)	98.69 (0.36)	88.10 (1.10)	84.99 (1.53)
		93.83 (1.87)	95.36 (1.08)	79.60 (2.01)	77.39 (1.57)
A (1380)	160[ranmod10]-30-30	98.86 (0.25)	98.64 (0.35)	88.41 (1.00)	85.05 (1.40)
		95.01 (1.66)	95.62 (1.06)	80.76 (1.95)	77.72 (1.72)
A (1280)	160[mod16]-32-30	98.15 (0.84)	98.42 (0.48)	85.74 (1.57)	83.02 (1.09)
		89.16 (2.46)	93.54 (0.99)	73.21 (2.98)	73.87 (2.01)
A (1840)	160[mod10]-40-30	99.01 (0.33)	98.88 (0.32)	89.45 (1.12)	86.32 (1.04)
		94.08 (1.83)	95.44 (0.93)	80.80 (2.04)	78.83 (1.35)

Tabelle A.2: Klassifikationsleistungen für die akustischen Daten. Die Anzahl Gewichte der Netze wird auf unterschiedliche Art und Weise verringert.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
V (1700)	140-10-30	59.69 (0.79)	29.55 (1.30)	53.39 (0.92)	25.01 (0.87)
V (3400)	140-20-30	61.76 (1.40)	33.69 (2.30)	54.46 (1.16)	27.17 (1.40)
V (5100)	140-30-30	59.11 (2.9)	29.94 (2.04)	53.05 (1.27)	25.27 (1.40)
V (6800)	140-40-30	57.33 (1.70)	28.36 (2.29)	51.85 (1.41)	24.81 (1.78)
Erkennungsleistung auf Buchstabenwortebene					
V (1700)	140-10-30	39.55 (2.37)	50.77 (2.02)	25.67 (1.72)	32.04 (1.69)
		22.29 (1.68)	30.53 (1.97)	10.84 (1.34)	15.47 (1.45)
V (3400)	140-20-30	43.55 (2.79)	55.17 (2.53)	29.23 (2.68)	35.69 (2.02)
		25.05 (2.21)	34.96 (2.87)	13.23 (1.49)	18.85 (1.98)
V (5100)	140-30-30	40.88 (1.82)	51.60 (1.79)	27.30 (2.06)	33.15 (2.36)
		22.44 (1.65)	31.24 (2.25)	11.60 (0.98)	16.64 (1.53)
V (6800)	140-40-30	37.50 (2.42)	50.42 (2.60)	24.81 (1.72)	32.15 (2.44)
		21.49 (2.02)	29.57 (2.82)	11.51 (1.63)	15.69 (2.33)

Tabelle A.3: Klassifikationsleistungen für die visuellen Daten.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
$I1$ (6600)	300-20-30	79.53 (3.20)	47.47 (7.43)	75.17 (2.49)	41.83 (5.36)
$I1$ (9900)	300-30-30	76.58 (4.13)	41.87 (7.48)	72.43 (3.24)	37.63 (5.97)
$I1$ (13200)	300-40-30	78.51 (4.38)	47.09 (8.97)	74.38 (3.44)	41.40 (6.86)
$I1$ (16500)	300-50-30	77.69 (3.71)	44.41 (7.43)	73.92 (2.98)	40.00 (5.94)
$I1$ (19800)	300-60-30	76.77 (3.53)	42.94 (6.58)	73.49 (2.84)	38.23 (5.19)
$I1$ (23100)	300-70-30	74.77 (3.24)	39.26 (6.44)	71.69 (2.44)	35.74 (5.24)
$I1$ (26400)	300-80-30	77.06 (3.27)	42.25 (7.02)	73.72 (2.69)	38.30 (5.54)
Erkennungsleistung auf Buchstabenwortebene					
$I1$ (6600)	300-20-30	77.94 (5.01)	78.08 (4.77)	69.63 (4.75)	67.50 (4.21)
		59.62 (6.34)	62.41 (6.49)	50.80 (5.84)	51.06 (5.38)
$I1$ (9900)	300-30-30	72.95 (7.06)	71.58 (7.34)	64.97 (5.80)	61.93 (6.24)
		57.47 (7.95)	55.63 (8.50)	48.86 (6.57)	45.93 (6.95)
$I1$ (13200)	300-40-30	75.47 (6.92)	74.40 (6.71)	65.74 (6.34)	63.44 (6.11)
		61.35 (7.36)	60.65 (8.32)	51.71 (6.04)	49.50 (7.15)
$I1$ (16500)	300-50-30	76.29 (5.55)	74.82 (5.62)	67.33 (5.00)	64.17 (4.95)
		61.08 (6.46)	59.65 (6.89)	52.06 (5.52)	49.48 (5.97)
$I1$ (19800)	300-60-30	77.02 (5.30)	75.58 (5.53)	69.04 (4.87)	65.03 (5.02)
		59.71 (6.60)	59.28 (7.27)	51.30 (5.74)	48.04 (6.37)
$I1$ (23100)	300-70-30	73.22 (4.81)	70.84 (4.90)	65.71 (4.16)	61.74 (4.34)
		57.25 (7.05)	55.74 (6.61)	48.48 (5.77)	46.11 (5.37)
$I1$ (26400)	300-80-30	75.30 (4.74)	73.94 (4.76)	66.61 (4.21)	62.64 (4.48)
		60.69 (6.16)	59.22 (6.45)	51.17 (5.36)	47.89 (5.78)

Tabelle A.4: Klassifikationsleistungen für die gemeinsame Verarbeitung der akustischen und der visuellen Daten, die Fusionsarchitektur ist I1.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
V (3400)	140-20-30	61.76 (1.4)	33.69 (2.30)	54.46 (1.16)	27.17 (1.40)
A (7600)	160-40-30	94.51 (0.46)	86.85 (1.42)	85.86 (0.69)	68.43 (1.37)
A_{mod} (1840)	160[mod10]-40-30	97.67 (0.15)	94.57 (0.42)	88.49 (0.59)	74.48 (0.80)
$I1$ (19800)	300-60-30	76.77 (3.53)	42.94 (6.58)	73.49 (2.84)	38.23 (5.19)
$I1(t)$ (11000)	300-60-30	94.6 (0.70)	88.30 (1.34)	85.41 (0.84)	69.89 (1.57)
H1 (11000)	160-40 \ 30 140-20 /	94.76 (0.50)	89.11 (1.03)	86.96 (0.97)	72.84 (1.75)
$I1(T)_{mod}$ (5240)	300-60-30	98.80 (0.12)	97.62 (0.25)	92.14 (0.41)	82.78 (1.40)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140————20 /	98.39 (0.15)	96.98 (0.36)	91.22 (0.43)	81.08 (0.97)
$I2$ (23400)	300-60-60-30	98.45 (0.33)	96.60 (0.88)	88.3 (0.63)	74.21 (1.65)
$I2(t)$ (12800)	300-60-60-30	99.06 (0.08)	98.10 (0.21)	90.14 (0.41)	78.55 (1.25)
$H2$ (12800)	160-40-30 \ 30 140-20-30 /	98.84 (0.10)	97.95 (0.21)	90.27 (0.57)	80.31 (1.33)
$I2(t)_{mod}$ (7040)	300-60-60-30	99.22 (0.07)	98.42 (0.20)	91.88 (0.50)	81.94 (1.09)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140————20-30 /	99.31 (0.06)	98.48 (0.19)	91.63 (0.47)	81.35 (1.14)
C (1800)	160-40-30 \ 30 140-20-30 /	97.60 (0.24)	94.45 (0.80)	87.3 (0.48)	72.55 (1.20)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140————20-30 /	99.03 (0.12)	97.79 (0.29)	88.4 (0.63)	75.08 (1.08)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Buchstabenwortebene					
V (3400)	140-20-30	43.55 (2.79)	55.17 (2.53)	29.23 (2.68)	35.69 (2.02)
		25.05 (2.21)	34.96 (2.87)	13.23 (1.49)	18.85 (1.98)
A (7600)	160-40-30	95.62 (1.10)	95.80 (1.00)	84.48 (1.79)	81.64 (1.87)
		82.57 (3.23)	87.19 (2.00)	69.50 (3.11)	70.56 (2.14)
A_{mod} (1840)	160[mod10]-40-30	99.01 (0.33)	98.88 (0.32)	89.45 (1.12)	86.32 (1.04)
		94.08 (1.83)	95.44 (0.93)	80.80 (2.04)	78.83 (1.35)
$I1$ (19800)	300-60-30	77.02 (5.30)	75.58 (5.53)	69.04 (4.87)	65.03 (5.02)
		59.71 (6.60)	59.30 (5.74)	51.30 (5.74)	48.04 (6.37)
$I1(t)$ (11000)	300-60-30	96.71 (0.88)	96.75 (0.70)	85.99 (1.38)	83.57 (1.20)
		83.31 (3,31)	88.65 (1.82)	70.25 (3.29)	72.75 (2.39)
$H1$ (11000)	160-40 \ 30 140-20 /	96.66 (1.04)	96.49 (0.88)	87.57 (1.42)	85.41 (1.75)
		81.13 (2.63)	86.86 (1.57)	70.77 (2.44)	73.72 (1.91)
$I1(t)_{mod}$ (5240)	300-60-30	98.98 (0.37)	99.18 (0.24)	91.31 (1.11)	89.32 (1.31)
		96.11 (1.46)	97.06 (0.91)	85.69 (1.60)	84.06 (1.36)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140————20 /	99.13 (0.39)	98.95 (0.41)	92.10 (0.84)	90.10 (1.17)
		94.98 (2.35)	96.04 (1.55)	85.77 (2.23)	84.48 (2.03)
$I2$ (23400)	300-60-60-30	98.32 (0.66)	98.23 (0.55)	88.32 (1.30)	85.37 (1.38)
		94.18 (2.25)	94.89 (1.59)	81.77 (2.03)	79.12 (1.81)
$I2(t)$ (12800)	300-60-60-30	98.99 (0.52)	99.16 (0.43)	90.02 (0.81)	88.02 (1.14)
		96.62 (1.40)	97.56 (0.74)	84.24 (1.58)	82.47 (1.47)
$H2$ (12800)	160-40-30 \ 30 140-20-30 /	99.36 (0.34)	99.35 (0.29)	90.80 (1.08)	89.38 (1.31)
		93.59 (2.27)	96.17 (1.10)	83.55 (2.42)	83.89 (2.00)
$I2(t)_{mod}$ (7040)	300-60-60-30	99.10 (0.44)	99.12 (0.35)	92.05 (0.88)	90.09 (1.29)
		97.84 (0.59)	97.97 (0.51)	88.19 (1.00)	86.24 (1.38)
$H2_{mod}$(7040)	160[mod10]-40-30 \ 30 140————20-30 /	99.31 (0.31)	99.30 (0.36)	92.53 (0.90)	90.53 (1.33)
		97.59 (1.11)	98.06 (0.61)	88.22 (1.20)	85.96 (1.60)
C (1800)	160-40-30 \ 30 140-20-30 /	98.03 (0.64)	97.95 (0.71)	87.57 (1.14)	84.89 (1.35)
		88.80 (2.65)	92.75 (1.45)	76.18 (2.72)	76.91 (1.94)
C_{mod}(1800)	160[mod10]-40-30 \ 30 140————20-30 /	99.30 (0.25)	99.26 (0.29)	90.36 (1.02)	87.57 (1.33)
		97.41 (1.31)	97.52 (0.92)	84.99 (1.67)	82.33 (1.69)

Tabelle A.5: Klassifikationsleistungen für die gemeinsame Verarbeitung der akustischen und visuellen Daten, die akustischen Daten sind ungestört, Verwendung verschiedener Fusionsmöglichkeiten.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A_{mod} (1840)	160[mod10]-40-30	85.32 (0.39)	70.12 (0.73)	65.66 (1.02)	36.43 (0.86)
$I1$ (19800)	300-60-30	56.48 (1.04)	14.51 (0.97)	54.80 (1.11)	14.04 (0.84)
$I1(t)_{mod}$ (5240)	300-60-30	79.51 (1.64)	57.23 (3.29)	67.90 (1.54)	40.87 (2.37)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———/	90.39 (0.77)	81.17 (1.61)	74.49 (0.93)	52.64 (1.33)
$I2$ (23400)	300-60-60-30	46.86 (7.30)	13.04 (3.05)	45.71 (6.83)	12.53 (2.79)
$I2(t)_{mod}$ (7040)	300-60-60-30	97.26 (0.21)	95.05 (0.44)	76.65 (0.79)	54.89 (1.38)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140———20-30 /	97.41 (0.19)	95.01 (0.45)	78.18 (0.88)	59.39 (1.75)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———20-30 /	92.30 (0.39)	82.75 (0.91)	70.66 (1.03)	44.52 (1.53)
Erkennungsleistung auf Buchstabenwortebene					
A_{mod} (1840) H	160[mod10]-40-30	97.85 (0.69)	97.76 (0.43)	66.34 (1.67)	56.01 (1.22)
		83.84 (1.99)	91.01 (1.01)	46.68 (1.78)	42.85 (1.22)
$I1$ (19800)	300-60-30	38.82 (4.38)	37.99 (3.91)	34.06 (4.21)	31.22 (2.72)
		19.11 (2.73)	15.63 (2.54)	16.96 (3.39)	12.13 (1.81)
$I1(t)_{mod}$ (5240)	300-60-30	96.33 (0.96)	95.72 (0.90)	72.30 (1.33)	65.54 (1.82)
		78.38 (2.18)	85.94 (1.62)	51.63 (2.47)	52.07 (2.15)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———/	97.79 (0.38)	97.39 (0.41)	78.49 (1.36)	72.58 (1.43)
		81.30 (1.73)	89.12 (0.83)	59.19 (1.90)	60.71 (1.54)
$I2$ (23400)	300-60-60-30	33.94 (10.95)	33.95 (10.52)	28.18 (8.93)	26.92 (7.97)
		15.14 (5.95)	15.60 (6.32)	10.79 (4.11)	10.08 (3.79)
$I2(t)_{mod}$ (7040)	300-60-60-30	99.49 (0.19)	99.33 (0.23)	80.38 (1.11)	74.21 (1.33)
		93.82 (2.43)	96.52 (1.10)	69.98 (2.29)	66.16 (1.54)
$H2_{mod}$(7040)	160[mod10]-40-30 \ 30 140———20-30 /	99.14 (0.28)	99.03 (0.24)	83.35 (0.78)	79.42 (1.05)
		93.36 (2.82)	95.85 (1.18)	72.81 (2.83)	70.66 (1.64)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———20-30 /	98.57 (0.21)	98.38 (0.23)	74.28 (1.09)	65.92 (1.85)
		89.98 (1.97)	92.96 (1.08)	60.64 (1.91)	55.19 (1.73)

Tabelle A.6: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, das SNR der akustischen Daten beträgt 3 dB. Verwendung verschiedener Fusionsarchitekturen. Für den akustischen Teil der Fusionsarchitekturen werden modularisierte Netze verwendet.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A_{mod} (1840)	160[mod10]-40-30	68.98 (2.06)	44.38 (3.83)	53.71 (1.14)	23.69 (1.47)
$I1$ (19800)	300-60-30	51.17 (2.42)	12.63 (0.76)	49.45 (2.46)	12.24 (0.73)
$I1(t)_{mod}$ (5240)	300-60-30	70.90 (1.86)	42.91 (3.21)	60.53 (1.14)	30.61 (1.44)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———–/	72.51 (1.66)	50.04 (2.76)	61.51 (1.23)	35.43 (1.58)
$I2$ (23400)	300-60-60-30	48.70 (6.00)	14.05 (3.04)	46.69 (5.57)	13.25 (2.59)
$I2(t)_{mod}$ (7040)	300-60-60-30	96.32 (0.22)	92.75 (0.54)	68.89 (0.78)	45.84 (1.75)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140———–20-30 /	96.34 (0.19)	93.33 (0.38)	72.28 (0.79)	52.75 (1.54)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———–20-30 /	84.97 (1.26)	69.19 (2.65)	65.55 (0.87)	39.48 (2.00)
Erkennungsleistung auf Buchstabenwortebene					
A_{mod} (1840)	160[mod10]-40-30	93.71 (1.40)	95.97 (0.78)	52.13 (2.22)	43.47 (1.49)
		85.44 (2.09)	90.77 (1.41)	39.03 (1.97)	33.49 (1.38)
$I1$ (19800)	300-60-30	34.63 (4.68)	33.05 (4.39)	29.88 (4.06)	27.40 (3.61)
		15.94 (2.26)	12.02 (1.86)	13.27 (1.99)	9.57 (1.49)
$I1(t)_{mod}$ (5240)	300-60-30	94.42 (1.32)	94.69 (1.03)	66.33 (1.60)	59.42 (1.78)
		78.04 (2.04)	84.50 (1.86)	46.66 (1.76)	45.00 (2.00)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———–/	95.15 (1.42)	95.65 (0.80)	67.92 (2.40)	62.31 (1.83)
		79.57 (1.74)	86.33 (1.10)	48.92 (2.16)	48.80 (2.02)
$I2$ (23400)	300-60-60-30	41.57 (11.79)	42.07 (1.24)	33.00 (9.39)	29.42 (7.64)
		18.25 (6.87)	20.52 (8.03)	11.44 (4.14)	10.98 (4.01)
$I2(t)_{mod}$ (7040)	300-60-60-30	99.51 (0.12)	99.34 (0.16)	73.47 (1.18)	66.93 (1.65)
		94.38 (2.23)	96.60 (1.04)	61.58 (2.25)	57.67 (1.73)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140———–20-30 /	99.60 (0.13)	99.52 (0.16)	77.82 (0.81)	73.36 (1.03)
		93.84 (2.51)	96.63 (1.06)	66.18 (2.48)	64.07 (1.49)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———–20-30 /	96.37 (1.30)	96.38 (0.83)	69.15 (1.75)	62.82 (2.09)
		78.66 (1.69)	86.37 (1.17)	48.02 (1.99)	48.19 (2.02)

Tabelle A.7: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, das SNR der akustischen Daten beträgt -3 dB. Verwendung verschiedener Fusionsarchitekturen. Für den akustischen Teil der Fusionsarchitekturen werden modularisierte Netze verwendet.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A_{mod} (1840)	160[mod10]-40-30	49.56 (1.28)	19.46 (2.02)	38.68 (0.62)	11.02 (0.56)
$I1$ (19800)	300-60-30	38.06 (3.60)	8.69 (0.97)	35.45 (3.46)	8.26 (0.90)
$I1(t)_{mod}$ (5240)	300-60-30	57.24 (1.54)	24.55 (2.11)	48.58 (1.09)	18.66 (1.21)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———/	63.22 (1.45)	36.09 (2.51)	51.27 (1.14)	21.82 (1.44)
$I2$ (23400)	300-60-60-30	38.12 (5.85)	8.86 (1.74)	36.08 (4.97)	8.37 (1.60)
$I2(t)_{mod}$ (7040)	300-60-60-30	89.64 (0.71)	80.04 (1.54)	53.08 (0.88)	29.96 (1.44)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140———20-30 /	92.59 (0.38)	87.48 (0.70)	58.02 (0.90)	37.01 (1.30)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———20-30 /	74.03 (1.61)	50.38 (3.16)	57.78 (0.91)	30.16 (1.92)
Erkennungsleistung auf Buchstabenwortebene					
A_{mod} (1840)	160[mod10]-40-30	80.08 (3.55)	86.63 (2.56)	21.42 (1.72)	19.20 (1.34)
		66.83 (4.76)	72.83 (4.74)	12.40 (1.50)	11.40 (1.22)
$I1$ (19800)	300-60-30	15.18 (4.08)	18.62 (4.69)	10.39 (3.11)	12.55 (3.51)
		3.08 (1.29)	2.48 (1.04)	1.59 (0.83)	1.03 (0.48)
$I1(t)_{mod}$ (5240)	300-60-30	70.12 (4.17)	77.97 (3.21)	33.35 (2.53)	33.79 (1.97)
		51.62 (4.37)	60.50 (4.71)	17.63 (1.67)	20.14 (1.76)
$H1_{mod}$ (5240)	160[mod10]-40 \ 30 140-20———/	74.59 (2.78)	82.75 (2.02)	37.78 (2.42)	37.67 (2.21)
		59.55 (2.68)	69.98 (2.68)	22.23 (1.79)	25.23 (2.39)
$I2$ (23400)	300-60-60-30	20.26 (6.45)	22.64 (7.24)	14.20 (4.50)	14.11 (4.59)
		6.12 (2.48)	6.41 (2.88)	2.61 (1.26)	2.56 (1.40)
$I2(t)_{mod}$ (7040)	300-60-60-30	97.52 (1.14)	98.39 (0.52)	51.22 (1.89)	46.05 (1.82)
		89.54 (2.14)	94.35 (1.07)	37.77 (1.98)	36.17 (1.46)
$H2_{mod}$ (7040)	160[mod10]-40-30 \ 30 140———20-30 /	97.92 (1.29)	98.61 (0.54)	60.18 (1.52)	56.80 (1.50)
		90.82 (1.64)	94.51 (0.80)	48.21 (1.58)	47.52 (1.76)
C_{mod} (1800)	160[mod10]-40-30 \ 30 140———20-30 /	83.10 (2.46)	87.47 (1.48)	49.71 (2.41)	48.87 (2.62)
		65.45 (2.23)	75.15 (1.89)	30.76 (2.48)	34.13 (2.85)

Tabelle A.8: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, das SNR der akustischen Daten beträgt -11 dB. Verwendung verschiedener Fusionsarchitekturen. Für den akustischen Teil der Fusionsarchitekturen werden modularisierte Netze verwendet.

Trainingsdaten	KSPM (trai)	Testdaten	KSPM (test)	SNR
Schwelle -100				
48.63 (3.60)	59.14 (2.96)	34.18 (3.22)	39.38 (2.41)	ungestört, nichtmodular
48.84 (3.75)	59.36 (3.10)	34.42 (3.36)	39.61 (2.61)	ungestört
48.71 (3.69)	59.19 (3.04)	33.26 (3.07)	38.42 (2.36)	SNR = 3 dB
48.14 (3.44)	58.98 (2.96)	32.24 (2.88)	37.79 (2.18)	SNR = -3 dB
47.18 (3.17)	58.40 (2.88)	30.34 (2.65)	36.61 (1.98)	SNR = -11 dB
Schwelle 0				
63.12 (3.77)	66.12 (2.92)	52.97 (3.86)	51.04 (2.59)	ungestört, nichtmodular
64.72 (3.94)	67.22 (3.17)	54.28 (3.37)	51.85 (2.40)	ungestört
48.72 (3.66)	59.11 (2.99)	34.06 (3.20)	38.70 (2.40)	SNR = 3 dB
48.08 (3.45)	58.92 (2.97)	32.29 (2.94)	37.86 (2.18)	SNR = -3 dB
47.00 (3.16)	58.19 (2.87)	30.20 (2.63)	36.45 (1.93)	SNR = -11 dB
Schwelle 100				
92.27 (0.91)	93.85 (0.65)	81.50 (1.65)	79.60 (1.60)	ungestört, nichtmodular
95.14 (1.44)	95.52 (1.11)	84.97 (1.74)	81.02 (1.87)	ungestört
69.30 (2.74)	72.62 (2.38)	50.33 (2.64)	46.04 (2.03)	SNR = 3 dB
57.02 (3.79)	64.57 (2.96)	38.02 (3.15)	39.23 (2.28)	SNR = -3 dB
45.56 (2.69)	59.65 (2.15)	26.28 (2.29)	33.18 (1.94)	SNR = -11 dB
Schwelle 150				
94.82 (0.99)	95.41 (0.90)	83.90 (1.71)	81.41 (1.87)	ungestört, nichtmodular
98.92 (0.38)	98.83 (0.35)	89.35 (1.17)	86.07 (1.18)	ungestört
78.62 (1.74)	80.16 (1.59)	56.73 (2.13)	50.43 (1.66)	SNR = 3 dB
64.27 (2.86)	71.33 (0.41)	40.62 (2.70)	40.26 (1.85)	SNR = -3 dB
52.19 (2.49)	67.58 (1.91)	25.24 (1.91)	31.68 (1.81)	SNR = -11 dB
Schwelle 200				
95.63 (1.09)	95.81 (0.90)	84.53 (1.80)	81.68 (1.87)	ungestört, nichtmodular
99.02 (0.33)	98.90 (0.31)	89.45 (1.11)	86.32 (1.04)	ungestört
83.31 (1.18)	86.13 (1.05)	59.30 (1.75)	53.07 (1.31)	SNR = 3 dB
70.85 (2.14)	78.68 (1.40)	44.16 (2.35)	42.84 (1.59)	SNR = -3 dB
59.65 (2.70)	73.82 (1.69)	24.37 (2.01)	28.86 (2.46)	SNR = -11 dB
akustische Entscheidung hat Vorrang				
95.63 (1.09)	95.81 (0.99)	84.51 (1.80)	81.67 (1.87)	ungestört, nichtmodular
99.02 (0.33)	98.90 (0.31)	89.45 (1.11)	86.32 (1.04)	ungestört
97.86 (0.69)	97.80 (0.41)	66.37 (1.67)	56.04 (1.22)	SNR = 3 dB
93.76 (1.41)	95.99 (0.78)	52.19(2.22)	43.51 (1.48)	SNR = -3 dB
80.29 (3.53)	86.81 (2.54)	21.56 (1.71)	19.38 (1.31)	SNR = -11 dB

Tabelle A.9: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, Verwendung der Architektur L, verschiedene Fusionsstrategien. Alle Ergebnisse, wenn nicht anders vermerkt, beziehen sich auf die Verwendung der modularen Architektur für die akustischen Daten. Als Entscheidungskriterium für den Assoziativspeicher wird der Hammingabstand verwendet. Die Klassifikationsergebnisse beziehen sich auf die Buchstabenwortebene.

ungestört	SNR = 3 dB	SNR = -3 dB	SNR = -11 dB	Getestet mit ...
mit ungestörten Daten trainiert				
99.01 (0.33)	12.36 (1.27)	6.77 (1.01)	4.53 (0.81)	Training
98.88 (0.32)	11.51 (0.88)	7.08 (0.60)	4.76 (0.30)	KSPM(trai)
89.45 (1.12)	12.84 (1.65)	6.61 (1.19)	4.25 (0.71)	Test
86.32 (1.04)	11.54 (1.18)	6.71 (0.76)	4.81 (0.40)	KSPM(test)
mit gestörten Daten trainiert: SNR = 3 dB				
35.23 (1.90)	97.85 (0.69)	60.85 (0.82)	20.21 (0.94)	Training
26.40 (1.53)	97.76 (0.43)	54.89 (0.95)	20.25 (0.72)	KSPM(trai)
34.37 (2.33)	66.34 (1.67)	48.38 (1.54)	16.59 (0.95)	Test
25.76 (1.46)	56.01 (1.22)	38.34 (1.47)	15.15 (1.15)	KSPM(test)
mit gestörten Daten trainiert: SNR = -3 dB				
21.07 (2.07)	58.00 (1.75)	93.71 (1.40)	34.06 (1.25)	Training
14.80 (1.31)	52.07 (1.50)	95.97 (0.78)	32.46 (1.58)	KSPM(trai)
21.27 (2.64)	49.45 (2.27)	52.13 (2.22)	25.39 (1.56)	Test
14.66 (1.58)	40.55 (1.55)	43.47 (1.49)	20.08 (1.30)	KSPM(test)
mit gestörten Daten trainiert: SNR = -11 dB				
11.35 (2.19)	19.57 (2.47)	28.16 (3.02)	80.08 (3.55)	Training
9.79 (1.19)	18.24 (1.50)	27.59 (2.21)	86.63 (2.56)	KSPM(trai)
11.70 (2.24)	17.75 (2.24)	21.50 (2.81)	21.42 (1.72)	Test
10.05 (1.54)	14.92 (1.47)	18.31 (1.88)	19.20 (1.34)	KSPM(test)

Tabelle A.10: Klassifikationsleistungen für den akustischen Kanal bei Verwendung der Architektur 160[mod10]-40-30 (A_{mod}). Dabei wird jeweils ein akustischer Datensatz zum Training verwendet (vgl. die Blöcke in der Tabelle), während zum Testen auch die Datensätze mit anderem SNR hinzugezogen werden (vgl. Spalten in der Tabelle). Die Klassifikationsleistungen beziehen sich auf die Buchstabenwortebene. Als Entscheidungskriterium für den Assoziativspeicher wird der Hammingabstand verwendet.

Anhang B

Ergebnisse für dimensionsreduzierte Daten

In diesem Anhang sind alle Ergebnisse aufgelistet, die mit den niedrigdimensionalen Daten (LDA-reduzierte Merkmale) erzielt wurden. Aufgetragen sind die Trainings- und Testfehler sowie die klassenspezifischen Mittel auf Trainings- und Testdatensatz für die einzelnen Architekturen sowohl auf Segmentklassifikationsebene als auch auf der Ebene der Buchstabenwortklassifikationen. In Klammern sind jeweils die 95%igen Konfidenzintervalle angegeben. Für jede Architektur ist die Bezeichnung, die Anzahl Gewichte sowie die Topologie dargestellt.
Die Tabellen für die Klassifikationsleistungen auf Buchstabenwortebene geben für jede Architektur in der ersten Zeile die Ergebnisse wieder für den Fall, daß der Hammingabstand als Entscheidungskriterium des Assoziativspeichers verwendet wird, die zweite Zeile enthält die Ergebnisse, die mit dem Überlapp als Entscheidungskriterium erreicht werden.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A (250)	20-5-30	85,54 (0.80)	60,51 (2.14)	78,15 (1.06)	50,41 (1.68)
A (350)	20-7-30	91,96 (0.54)	77,94 (1.54)	81,72 (0.63)	60,03 (0.80)
A (400)	20-8-30	94,28 (0.27)	85,18 (0.86)	83,45 (0.65)	65,59 (1.30)
A (500)	20-10-30	95,92 (0.21)	89,63 (0.21)	83,69 (0.74)	66,39 (1.34)
A (750)	20-15-30	97,78 (0.15)	94,85 (0.43)	85,56 (0.56)	70,63 (0.94)
A (1000)	20-20-30	98,43 (0.09)	96,76 (0.27)	86,86 (0.14)	73,01 (1.53)
A (1250)	20-25-30	98,66 (0.09)	97,50 (0.31)	87,64 (0.54)	74,21 (1.31)
A (1500)	20-30-30	98,76 (0.07)	97,74 (0.23)	88,21 (0.36)	75,67 (1.07)
A (1750)	20-35-30	98,83 (0.06)	97,96 (0.21)	88,18 (0.54)	75,78 (1.21)
A (2000)	20-40-30	98,90 (0.08)	98,12 (0.18)	88,29 (0.43)	76,04 (1.32)
A (2250)	20-45-30	98,93 (0.06)	98,18 (0.22)	87,93 (0.50)	75,04 (1.37)
A (2500)	20-50-30	98,91 (0.06)	98,11 (0.23)	88,27 (0.46)	76,15 (1.18)
A (3000)	20-60-30	98,97 (0.06)	98,34 (0.15)	88,26 (0.45)	76,04 (1.12)
A (3500)	20-70-30	98,96 (0.06)	98,28 (0.17)	88,48 (0.36)	76,43 (1.22)
A (5000)	20-100-30	98,96 (0.07)	98,37 (0.19)	88,41 (1.12)	76,47 (1.12)
Erkennungsleistung auf Buchstabenebene					
A (250)	20-5-30	95.45 (0.81)	95.34 (0.83)	77.24 (1.44)	72.64 (1.89)
		77.01 (2.95)	84.57 (1.97)	55.38 (3.17)	57.11 (2.35)
A (350)	20-7-30	97.17 (0.59)	97.31 (0.54)	80.73 (1.33)	76.78 (1.65)
		84.32 (2.58)	89.90 (1.28)	63.87 (3.23)	64.90 (2.07)
A (400)	20-8-30	98.07 (0.40)	98.06 (0.44)	83.16 (1.16)	78.76 (1.50)
		91.53 (2.25)	93.58 (1.17)	72.43 (2.47)	70.33 (1.79)
A (500)	20-10-30	98.25 (0.46)	98.45 (0.43)	84.48 (1.11)	81.41 (1.35)
		92.61 (1.88)	94.74 (0.94)	74.36 (1.88)	72.92 (1.39)
A (750)	20-15-30	98.73 (0.37)	99.05 (0.27)	85.21 (0.90)	82.24 (1.35)
		96.70 (1.25)	97.26 (0.73)	78.63 (1.70)	75.71 (1.62)
A (1000)	20-20-30	98.80 (0.30)	99.04 (0.24)	86.45 (0.98)	83.86 (1.32)
		98.06 (0.51)	98.13 (0.45)	81.56 (1.13)	78.70 (1.26)
A (1250)	20-25-30	99.16 (0.29)	99.25 (0.24)	87.08 (1.37)	83.76 (1.71)
		97.63 (0.87)	97.83 (0.54)	81.07 (1.66)	78.11 (1.84)
A (1500)	20-30-30	99.01 (0.31)	99.25 (0.27)	87.81 (1.01)	85.12 (1.46)
		98.24 (0.53)	98.30 (0.46)	82.45 (1.32)	79.12 (1.76)
A (1750)	20-35-30	98.94 (0.31)	99.23 (0.22)	88.02 (1.05)	85.76 (1.19)
		97.61 (0.76)	97.92 (0.44)	82.07 (1.41)	79.03 (1.53)
A (2000)	20-40-30	98.98 (0.44)	99.21 (0.35)	87.63 (1.16)	84.99 (1.55)
		98.14 (0.56)	98.20 (0.52)	82.69 (1.21)	79.72 (1.52)
A (2250)	20-45-30	99.11 (0.28)	99.27 (0.25)	87.89 (0.88)	85.17 (1.32)
		97.71 (0.91)	97.91 (0.52)	82.34 (1.10)	79.22 (1.26)
A (2500)	20-50-30	99.10 (0.23)	99.28 (0.22)	87.60 (0.98)	84.79 (1.23)
		98.53 (0.30)	98.52 (0.29)	82.82 (1.10)	79.81 (1.28)
A (3000)	20-60-30	99.37 (0.22)	99.48 (0.23)	88.59 (0.94)	85.85 (1.21)
		98.58 (0.33)	98.48 (0.33)	83.67 (1.14)	80.07 (1.42)
A (3500)	20-70-30	98.89 (0.38)	99.10 (0.31)	88.03 (0.87)	85.74 (1.20
		98.02 (0.51)	98.11 (0.44)	82.79 (1.18)	80.06 (1.46)
A (5000)	20-100-30	99.18 (0.36)	99.38 (0.29)	88.27 (0.85)	85.90 (1.37)
		98.35 (0.49)	98.30 (0.47)	83.30 (1.02)	80.47 (1.46)

Tabelle B.1: Klassifikationsleistungen für die akustischen Daten.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
V (156)	9-4-30	58,56 (0.29)	31,27 (0.94)	52,63 (0.80)	25,85 (0.94)
V (195)	9-5-30	60,03 (0.17)	34,65 (0.77)	52,88 (0.78)	27,19 (0.85)
V (234)	9-6-30	61,23 (0.21)	38,44 (0.75)	53,40 (0.79)	28,37 (0.87)
V (273)	9-7-30	62,09 (0.18)	40,63 (0.66)	54,14 (0.71)	30,46 (0.95)
V (390)	9-10-30	63,17 (0.15)	43,76 (0.45)	54,46 (0.70)	32,08 (0.92)
V (585)	9-15-30	64,20 (0.20)	46,29 (0.49)	54,91 (0.85)	33,38 (1.26)
V (780)	9-20-30	64,48 (0.21)	46,80 (0.50)	55,00 (0.68)	33,05 (1.45)
V (975)	9-25-30	64,68 (0.17)	47,21 (0.42)	55,03 (0.82)	33,03 (1.24)
V (1170)	9-30-30	64,70 (0.22)	47,39 (0.45)	55,09 (0.77)	33,13 (1.36)
V (1365)	9-35-30	64,67 (0.19)	47,07 (0.44)	55,27 (0.80)	33,64 (1.56)
V (1560)	9-40-30	64,71 (0.19)	47,26 (0.46)	54,98 (0.62)	33,30 (1.34)
V (1950)	9-50-30	64,78 (0.16)	47,28 (0.41)	55,24 (0.64)	33,62 (1.24)
V (2340)	9-60-30	64,78 (0.19)	47,22 (0.41)	55,24 (0.71)	33,28 (1.28)
Erkennungsleistung auf Buchstabenebene					
V (156)	9-4-30	37.45 (2.27)	50.90 (1.64)	25.26 (2.46)	31.71 (2.02)
		22.74 (1.26)	32.26 (1.26)	13.31 (1.26)	17.98 (1.36)
V (195)	9-5-30	37.71 (2.01)	52.03 (1.48)	25.74 (1.80)	33.26 (1.59)
		23.42 (1.01)	33.38 (1.05)	13.56 (1.13)	18.95 (1.10)
V (234)	9-6-30	41.23 (1.95)	54.60 (1.11)	28.45 (1.66)	36.21 (1.37)
		25.28 (0.60)	36.06 (0.63)	14.41 (1.10)	20.52 (0.86)
V (273)	9-7-30	39.86 (1.65)	54.18 (1.25)	27.44 (1.69)	36.02 (1.20)
		25.59 (1.00)	36.76 (1.02)	15.60 (1.33)	22.92 (1.22)
V (390)	9-10-30	41.55 (1.62)	55.91 (1.31)	28.67 (1.58)	37.35 (1.09)
		26.15 (0.69)	37.10 (0.86)	15.31 (0.96)	22.29 (1.32)
V (585)	9-15-30	44.20 (1.42)	57.48 (1.09)	30.22 (1.52)	38.07 (1.50)
		27.92 (0.82)	39.02 (0.92)	16.14 (1.15)	23.32 (1.59)
V (780)	9-20-30	44.84 (1.68)	57.64 (1.53)	30.66 (2.26)	38.09 (1.46)
		28.26 (0.88)	38.95 (1.05)	16.27 (1.43)	23.80 (1.43)
V (975)	9-25-30	43.90 (1.27)	57.57 (1.03)	29.59 (1.56)	37.82 (1.28)
		28.44 (0.66)	39.30 (0.94)	16.02 (1.35)	23.34 (1.67)
V (1170)	9-30-30	44.50 (1.43)	58.19 (1.18)	29.56 (1.94)	37.87 (1.32)
		28.57 (0.94)	39.36 (1.08)	16.05 (1.32)	23.34 (1.60)
V (1365)	9-35-30	44.22 (1.59)	57.61 (1.20)	29.98 (1.81)	38.26 (1.33)
		27.90 (1.00)	38.83 (1.31)	15.70 (1.43)	23.20 (1.44)
V (1560)	9-40-30	44.65 (1.66)	58.36 (0.94)	29.78 (1.80)	37.94 (1.24)
		28.64 (0.91)	40.03 (1.16)	16.10 (1.34)	23.61 (1.46)
V (1950)	9-50-30	44.41 (1.51)	58.09 (1.17)	30.47 (1.94)	38.75 (1.43)
		28.29 (0.96)	39.30 (1.24)	16.67 (1.40)	23.97 (1.34)
V (2340)	9-60-30	44.75 (1.84)	57.96 (1.38)	30.76 (2.30)	39.28 (1.39)
		28.39 (1.01)	39.12 (1.30)	16.46 (1.38)	23.62 (1.63)

Tabelle B.2: Klassifikationsleistungen für die visuellen Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
V (156)	9-4-30	68,03 (0.18)	40,40 (0.78)	50,27 (0.95)	24,29 (0.89)
V (195)	9-5-30	69,79 (0.23)	45,65 (0.23)	51,35 (0.83)	27,25 (0.65)
V (234)	9-6-30	70,98 (0.29)	49,02 (0.81)	51,57 (0.80)	28,70 (0.89)
V (273)	9-7-30	71,58 (0.28)	50,79 (0.58)	51,98 (0.78)	29,45 (0.94)
V (390)	9-10-30	73,19 (0.24)	55,11 (0.59)	52,57 (0.72)	31,53 (0.84)
V (585)	9-15-30	73,67 (0.25)	56,22 (0.57)	53,12 (0.72)	32,58 (0.95)
V (780)	9-20-30	73,90 (0.24)	56,99 (0.54)	53,28 (0.76)	32,80 (0.83)
V (975)	9-25-30	74,01 (0.23)	57,15 (0.51)	53,12 (0.67)	32,72 (0.97)
V (1170)	9-30-30	74,00 (0.21)	57,06 (0.50)	53,24 (0.75)	32,81 (0.85)
V (1365)	9-35-30	73,87 (0.27)	57,01 (0.71)	53,29 (0.63)	32,85 (0.79)
V (1560)	9-40-30	73,75 (0.27)	56,74 (0.68)	53,17 (0.69)	32,57 (0.85)
V (1950)	9-50-30	73,88 (0.21)	56,79 (0.57)	53,13 (0.69)	32,33 (0.88)
V (2340)	9-60-30	73,92 (0.21)	57,01 (0.62)	53,24 (0.71)	33,00 (0.91)
Erkennungsleistung auf Buchstabenebene					
V (156)	9-4-30	50.93 (1.38)	65.66 (0.96)	28.45 (1.56)	34.98 (1.55)
		36.79 (0.89)	48.28 (0.93)	16.19 (0.97)	22.00 (0.83)
V (195)	9-5-30	54.21 (1.34)	68.10 (0.90)	31.27 (1.33)	37.37 (1.48)
		39.27 (0.58)	51.41 (0.76)	17.59 (0.96)	23.62 (1.59)
V (234)	9-6-30	54.28 (1.23)	68.96 (1.19)	31.00 (1.30)	37.73 (1.53)
		40.31 (1.04)	52.72 (1.02)	18.34 (1.18)	24.45 (1.81)
V (273)	9-7-30	55.16 (1.34)	69.54 (1.13)	32.62 (1.15)	39.13 (1.65)
		41.23 (1.03)	53.82 (0.95)	19.60 (0.98)	25.60 (1.82)
V (390)	9-10-30	58.35 (1.10)	72.39 (1.02)	33.18 (1.52)	40.80 (1.72)
		44.20 (0.90)	56.90 (0.97)	19.78 (0.92)	25.77 (1.67)
V (585)	9-15-30	59.84 (1.67)	73.28 (1.16)	34.76 (1.62)	41.67 (1.72)
		45.20 (1.10)	58.35 (0.99)	20.88 (0.76)	27.87 (1.91)
V (780)	9-20-30	59.56 (1.22)	73.22 (1.02)	34.01 (0.92)	40.54 (1.68)
		46.06 (1.02)	59.31 (0.97)	21.24 (0.81)	27.49 (1.63)
V (975)	9-25-30	59.02 (1.01)	72.88 (1.05)	33.61 (1.44)	40.38 (1.94)
		46.00 (0.93)	59.48 (0.97)	20.61 (1.05)	27.42 (2.09)
V (1170)	9-30-30	58.86 (1.03)	72.96 (1.15)	33.47 (1.29)	40.94 (1.72)
		45.88 (1.13)	59.01 (1.03)	20.88 (0.85)	27.35 (1.69)
V (1365)	9-35-30	59.84 (1.04)	73.32 (1.03)	34.15 (1.46)	40.59 (1.97)
		45.78 (1.18)	59.01 (1.15)	20.62 (1.05)	27.11 (2.10)
V (1560)	9-40-30	59.51 (1.20)	72.88 (1.11)	34.57 (1.40)	41.27 (1.74)
		45.33 (0.88)	58.80 (0.92)	20.72 (0.97)	26.57 (1.96)
V (1950)	9-50-30	58.20 (1.31)	72.20 (1.26)	33.50 (1.20)	40.14 (2.06)
		45.33 (0.91)	58.80 (1.10)	20.88 (0.81)	27.00 (1.96)
V (2340)	9-60-30	59.69 (0.85)	72.93 (1.02)	33.86 (1.34)	40.88 (1.84)
		45.39 (1.01)	58.74 (0.99)	20.22 (0.86)	26.20 (2.19)

Tabelle B.3: Klassifikationsleistungen für den visuellen Kanal. Als visuelles Merkmal wird das ganze Zeitfenster verwendet.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
$I1$ (472)	29-8-30	96,08 (0.33)	89,42 (1.11)	85,50 (0.62)	69,58 (1.06)
$I1$ (590)	29-10-30	97,79 (0.12)	94,97 (0.40)	86,83 (0.54)	73,33 (1.29)
$I1$ (885)	29-15-30	98,80 (0.07)	97,73 (0.24)	89,01 (0.43)	77,22 (1.17)
$I1$ (1180)	29-20-30	99,07 (0.07)	98,49 (0.15)	89,58 (0.59)	79,61 (1.22)
$I1$ (1475)	29-25-30	99,16 (0.06)	98,72 (0.17)	89,75 (0.62)	79,70 (1.31)
$I1$ (1770)	29-30-30	99,22 (0.07)	98,92 (0.15)	90,08 (0.64)	80,39 (1.42)
$I1$ (2065)	29-35-30	99,23 (0.06)	98,93 (0.12)	90,14 (0.58)	80,24 (1.13)
$I1$ (2360)	29-40-30	99,25 (0.06)	99,06 (0.09)	90,05 (0.57)	80,09 (1.40)
$I1$ (2655)	29-45-30	99,26 (0.07)	99,04 (0.13)	90,37 (0.52)	80,97 (1.10)
$I1$ (2950)	29-50-30	99,28 (0.06)	99,05 (0.11)	90,51 (0.57)	81,21 (1.41)
$I1$ (3540)	29-60-30	99,27 (0.06)	99,05 (0.11)	90,55 (0.54)	81,36 (1.27)
$I1$ (4130)	29-70-30	99,26 (0.05)	99,04 (0.08)	90,57 (0.59)	81,20 (1.16)
Erkennungsleistung auf Buchstabenebene					
$I1$ (472)	29-8-30	98.65 (0.34)	98.74 (0.30)	87.25 (1.11)	84.45 (1.40)
		90.87 (2.39)	94.30 (1.31)	76.10 (2.44)	75.48 (1.92)
$I1$ (590)	29-10-30	98.82 (0.29)	98.93 (0.29)	88.11 (0.99)	85.90 (1.56)
		93.47 (1.99)	95.63 (0.98)	78.85 (1.87)	78.04 (1.63)
$I1$ (885)	29-15-30	99.14 (0.62)	99.41 (0.37)	90.04 (1.07)	88.52 (1.26)
		97.95 (0.91)	98.21 (0.54)	84.62 (1.04)	82.81 (1.27)
$I1$ (1180)	29-20-30	99.29 (0.32)	99.43 (0.26)	90.25 (0.88)	88.81 (1.20)
		98.08 (0.89)	98.27 (0.57)	85.64 (1.29)	84.24 (1.39)
$I1$ (1475)	29-25-30	99.51 (0.24)	99.58 (0.25)	90.96 (1.06)	89.50 (1.47)
		98.51 (0.60)	98.47 (0.54)	86.42 (1.21)	85.14 (1.44)
$I1$ (1770)	29-30-30	99.46 (0.20)	99.57 (0.20)	90.95 (0.68)	89.58 (1.03)
		98.75 (0.42)	98.67 (0.41)	86.74 (1.03)	85.28 (1.34)
$I1$ (2065)	29-35-30	99.31 (0.34)	99.42 (0.31)	90.92 (0.99)	89.32 (1.61)
		98.33 (0.58)	98.31 (0.51)	86.50 (0.87)	84.74 (1.40)
$I1$ (2360)	29-40-30	99.41 (0.35)	99.58 (0.28)	90.88 (0.95)	90.04 (1.26)
		98.72 (0.50)	98.73 (0.40)	86.96 (1.23)	85.59 (1.60)
$I1$ (2655)	29-45-30	99.28 (0.38)	99.44 (0.30)	91.52 (0.91)	90.48 (1.15)
		98.61 (0.43)	98.57 (0.37)	87.38 (0.84)	85.76 (1.05)
$I1$ (2950)	29-50-30	99.30 (0.35)	99.46 (0.29)	91.38 (1.26)	90.54 (1.62)
		98.80 (0.39)	98.76 (0.36)	87.09 (1.36)	86.22 (1.77)
$I1$ (3540)	29-60-30	99.26 (0.35)	99.43 (0.28)	91.24 (0.99)	90.06 (1.31)
		98.56 (0.44)	98.51 (0.41)	87.24 (1.06)	85.92 (1.41)
$I1$ (4130)	29-70-30	99.23 (0.44)	99.41 (0.33)	91.24 (0.86)	91.24 (0.86)
		98.25 (0.70)	98.24(0.63)	87.00 (0.93)	85.50 (1.27)

Tabelle B.4: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitektur ist I1.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
$I1$ (472)	29-8-30	95,23 (0.41)	87,17 (1.29)	84,26 (0.79)	66,49 (1.66)
$I1$ (590)	29-10-30	97,17 (0.17)	93,10 (0.61)	85,80 (0.66)	70,60 (1.30)
$I1$ (885)	29-15-30	98,41 (0.12)	96,65 (0.37)	87,89 (0.49)	75,18 (1.03)
$I1$ (1180)	29-20-30	98,82 (0.09)	97,92 (0.22)	88,18 (0.66)	75,88 (1.18)
$I1$ (1475)	29-25-30	98,97 (0.08)	98,27 (0.21)	88,59 (0.49)	76,57 (0.99)
$I1$ (1770)	29-30-30	99,05 (0.05)	98,52 (0.12)	88,90 (0.53)	77,17 (1.21)
$I1$ (2065)	29-35-30	99,12 (0.07)	98,64 (0.15)	89,16 (0.54)	77,91 (1.13)
$I1$ (2360)	29-40-30	99,15 (0.07)	98,79 (0.16)	88,83 (0.49)	77,33 (1.26)
$I1$ (2655)	29-45-30	99,16 (0.07)	98,82 (0.14)	89,21 (0.51)	78,58 (0.93)
$I1$ (2950)	29-50-30	99,19 (0.07)	98,84 (0.16)	89,31 (0.46)	78,24 (1.14)
$I1$ (3540)	29-60-30	99,15 (0.06)	98,79 (0.15)	89,58 (0.56)	79,26 (1.06)
$I1$ (4130)	29-70-30	99,19 (0.05)	98,90 (0.12)	89,65 (0.52)	79,18 (1.16)
Erkennungsleistung auf Buchstabenebene					
$I1$ (472)	29-8-30	98.32 (0.55)	98.40 (0.48)	85.24 (1.74)	81.96 (1.99)
		88.30 (2.54)	92.90 (1.40)	71.28 (3.08)	72.34 (2.38)
$I1$ (590)	29-10-30	98.87 (0.43)	99.05 (0.33)	87.00 (1.15)	83.84 (1.76)
		93.12 (2.33)	95.56 (1.24)	76.89 (2.43)	75.40 (1.93)
$I1$ (885)	29-15-30	99.06 (0.26)	99.15 (0.25)	88.88 (1.09)	86.88 (1.54)
		97.37 (1.08)	97.75 (0.61)	83.03 (1.45)	80.74 (1.42)
$I1$ (1180)	29-20-30	99.25 (0.30)	99.36 (0.25)	88.53 (0.86)	85.60 (1.06)
		97.91 (0.89)	98.07 (0.56)	82.45 (1.45)	79.43 (1.38)
$I1$ (1475)	29-25-30	99.35 (0.36)	99.47 (0.29)	89.07 (0.88)	86.45 (1.29)
		98.58 (0.55)	98.55 (0.48)	84.10 (1.08)	81.39 (1.51)
$I1$ (1770)	29-30-30	99.44 (0.22)	99.52 (0.24)	89.15 (0.94)	86.87 (1.45)
		98.21 (0.88)	98.35 (0.56)	83.24 (1.49)	81.20 (1.57)
$I1$ (2065)	29-35-30	99.22 (0.33)	99.32 (0.32)	89.27 (0.98)	86.97 (1.56)
		98.06 (0.57)	98.06 (0.46)	83.78 (0.94)	80.94 (1.46)
$I1$ (2360)	29-40-30	99.42 (0.25)	99.55 (0.24)	89.35 (0.94)	87.73 (1.37)
		98.66 (0.36)	98.59 (0.34)	84.76 (0.75)	82.51 (1.13)
$I1$ (2655)	29-45-30	99.55 (0.15)	99.64 (0.18)	89.64 (0.77)	87.88 (1.29)
		98.75 (0.48)	98.74 (0.42)	84.84 (0.91)	82.62 (1.23)
$I1$ (2950)	29-50-30	99.27 (0.30)	99.37 (0.28)	89.39 (0.87)	87.92 (1.35)
		98.47 (0.36)	98.38 (0.35)	84.62 (0.94)	82.91 (1.35)
$I1$ (3540)	29-60-30	99.07 (0.38)	99.23 (0.31)	89.10 (0.98)	87.09 (1.40)
		98.33 (0.46)	98.31 (0.41)	84.67 (0.93)	82.32 (1.49)
$I1$ (4130)	29-70-30	99.52 (0.22)	99.63 (0.22)	89.72 (0.91)	87.72 (1.24)
		98.70 (0.33)	98.57 (0.34)	84.99 (1.08)	82.53 (1.47)

Tabelle B.5: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitektur ist I1.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
E (472)	29-8-30	97,86 (0.19)	94,70 (0.62)	86,48 (0.35)	72,10 (1.27)
E (590)	29-10-30	98,98 (0.09)	98,19 (0.26)	88,32 (0.35)	75,94 (1.22)
E (885)	29-15-30	99,42 (0.06)	99,19 (0.12)	89,71 (0.49)	79,10 (1.28)
E (1180)	29-20-30	99,56 (0.05)	99,48 (0.12)	90,14 (0.47)	80,34 (0.91)
E (1475)	29-25-30	99,59 (0.05)	99,56 (0.09)	90,49 (0.51)	80,96 (1.06)
E (1770)	29-30-30	99,61 (0.04)	99,57 (0.10)	90,50 (0.43)	80,81 (1.03)
E (2065)	29-35-30	99,59 (0.05)	99,57 (0.08)	90,38 (0.48)	80,84 (1.21)
E (2360)	29-40-30	99,60 (0.05)	99,59 (0.09)	90,29 (0.45)	80,52 (1.07)
E (2655)	29-45-30	99,62 (0.06)	99,63 (0.09)	90,33 (0.45)	80,39 (1.18)
E (2950)	29-50-30	99,61 (0.05)	99,56 (0.09)	90,23 (0.53)	80,60 (1.17)
E (3540)	29-60-30	99,58 (0.08)	99,54 (0.12)	90,28 (0.41)	80,27 (0.86)
E (4130)	29-70-30	99,55 (0.08)	99,50 (0.12)	90,21 (0.46)	80,25 (1.16)
Erkennungsleistung auf Buchstabenebene					
E (472)	29-8-30	98.90 (0.33)	98.95 (0.31)	88.00 (0.75)	85.37 (1.27)
		89.89 (2.69)	94.06 (1.36)	76.94 (2.30)	77.37 (2.06)
E (590)	29-10-30	99.15 (0.38)	99.27 (0.27)	90.14 (0.76)	88.28 (1.05)
		96.27 (1.48)	97.31 (0.82)	83.43 (1.64)	82.23 (1.16)
E (885)	29-15-30	99.26 (0.44)	99.50 (0.31)	90.96 (0.78)	89.17 (1.14)
		98.65 (0.61)	98.80 (0.45)	86.74 (1.07)	85.01 (1.12)
E (1180)	29-20-30	98.84 (0.67)	99.24 (0.46)	90.74 (0.91)	90.27 (0.98)
		97.43 (1.77)	98.23 (0.94)	86.23 (1.66)	86.40 (1.30)
E (1475)	29-25-30	99.20 (0.49)	99.49 (0.33)	91.39 (0.72)	90.47 (0.91)
		98.86 (0.50)	98.93 (0.35)	88.24 (0.81)	87.35 (0.90)
E (1770)	29-30-30	99.26 (0.50)	99.51 (0.34)	91.28 (0.69)	90.46 (0.75)
		98.93 (0.50)	98.98 (0.35)	88.03 (0.91)	87.17 (0.91)
E (2065)	29-35-30	99.26 (0.43)	99.54 (0.30)	91.22 (1.06)	91.19 (0.83)
		98.90 (0.44)	98.95 (0.31)	87.67 (1.08)	87.77 (0.79)
E (2360)	29-40-30	99.43 (0.37)	99.63 (0.26)	91.70 (0.86)	91.45 (0.69)
		98.71 (0.62)	98.80 (0.43)	88.27 (0.97)	87.75 (0.80)
E (2655)	29-45-30	99.29 (0.48)	99.56 (0.32)	91.82 (0.92)	91.08 (0.96)
		98.91 (0.48)	98.96 (0.32)	88.30 (1.00)	87.57 (0.94)
E (2950)	29-50-30	99.64 (0.07)	99.78 (0.07)	91.33 (0.76)	90.83 (0.83)
		99.12 (0.41)	99.11 (0.29)	87.52 (0.95)	86.92 (0.89)
E (3540)	29-60-30	99.48 (0.34)	99.67 (0.23)	91.86 (0.70)	91.63 (0.72)
		99.09 (0.33)	99.09 (0.22)	88.51 (0.82)	88.48 (0.71)
E (4130)	29-70-30	99.10 (0.60)	99.43 (0.41)	91.35 (0.98)	90.81 (0.91)
		98.73 (0.60)	98.82 (0.42)	87.90 (1.01)	87.48 (0.80)

Tabelle B.6: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitektur ist E.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
E (472)	29-8-30	98,68 (0.19)	95,43 (0.79)	82,28 (0.51)	64,01 (1.00)
E (590)	29-10-30	99,46 (0.08)	98,47 (0.33)	84,39 (0.41)	69,78 (1.25)
E (885)	29-15-30	99,86 (0.03)	99,88 (0.07)	85,90 (0.36)	71,91 (1.54)
E (1180)	29-20-30	99,89 (0.03)	99,94 (0.02)	86,41 (0.41)	72,27 (1.41)
E (1475)	29-25-30	99,90 (0.02)	99,96 (0.02)	86,84 (0.38)	73,73 (1.52)
E (1770)	29-30-30	99,91 (0.02)	99,97 (0.02)	86,82 (0.34)	73,71 (1.46)
E (2065)	29-35-30	99,91 (0.02)	99,97 (0.01)	86,87 (0.45)	73,51 (1.59)
E (2360)	29-40-30	99,91 (0.02)	99,97 (0.01)	86,85 (0.32)	73,30 (1.17)
E (2655)	29-45-30	99,91 (0.03)	99,96 (0.02)	86,89 (0.35)	73,79 (1.28)
E (2950)	29-50-30	99,91 (0.02)	99,97 (0.01)	86,94 (0.42)	73,74 (1.50)
E (3540)	29-60-30	99,92 (0.02)	99,98 (0.01)	86,87 (0.38)	73,68 (1.43)
E (4130)	29-70-30	99,91 (0.02)	99,97 (0.01)	86,88 (0.35)	73,58 (1.28)
Erkennungsleistung auf Buchstabenebene					
E (472)	29-8-30	99.49 (0.21)	99.42 (0.29)	83.46 (0.77)	80.87 (0.94)
		92.76 (2.43)	95.85 (1.34)	72.13 (2.06)	72.47 (1.24)
E (590)	29-10-30	99.67 (0.06)	99.71 (0.10)	84.86 (0.85)	82.80 (0.99)
		97.06 (1.92)	98.17 (0.84)	77.23 (2.10)	76.52 (1.23)
E (885)	29-15-30	99.73 (0.08)	99.81 (0.08)	85.90 (0.76)	84.30 (0.98)
		99.04 (0.72)	99.09 (0.32)	80.32 (1.36)	79.15 (1.10)
E (1180)	29-20-30	99.67 (0.07)	99.81 (0.04)	86.74 (0.88)	85.67 (0.80)
		99.29 (0.09)	99.17 (0.08)	82.12 (1.01)	80.95 (0.72)
E (1475)	29-25-30	99.72 (0.08)	99.80 (0.08)	87.55 (1.01)	86.72 (1.00)
		99.32 (0.12)	99.17 (0.09)	82.92 (1.08)	82.07 (1.05)
E (1770)	29-30-30	99.73 (0.08)	99.83 (0.06)	87.09 (1.00)	86.37 (0.98)
		99.33 (0.11)	99.18 (0.08)	82.76 (1.03)	81.93 (0.99)
E (2065)	29-35-30	99.75 (0.08)	99.84 (0.06)	87.35 (0.92)	86.36 (1.05)
		99.38 (0.10)	99.23 (0.07)	82.90 (1.11)	81.50 (1.30)
E (2360)	29-40-30	99.66 (0.10)	99.79 (0.08)	87.60 (1.02)	86.83 (0.95)
		99.29 (0.11)	99.17 (0.08)	83.08 (1.03)	81.65 (1.08)
E (2655)	29-45-30	99.73 (0.08)	99.84 (0.04)	87.62 (0.76)	86.83 (0.79)
		99.35 (0.10)	99.21 (0.08)	83.19 (0.99)	81.99 (1.16)
E (2950)	29-50-30	99.72 (0.09)	99.82 (0.07)	87.43 (0.85)	86.78 (0.97)
		99.32 (0.11)	99.16 (0.09)	82.85 (1.06)	82.01 (1.29)
E (3540)	29-60-30	99.75 (0.09)	99.86 (0.05)	87.43 (1.08)	86.47 (1.11)
		99.38 (0.10)	99.23 (0.07)	83.11 (1.05)	81.77 (1.21)
E (4130)	29-70-30	99.71 (0.10)	99.84 (0.05)	87.62 (1.20)	86.79 (1.24)
		99.32 (0.11)	99.17 (0.08)	82.73 (1.17)	81.12 (1.24)

Tabelle B.7: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitektur ist E.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
	ungestört				
A (1250)	20-25-30	98,66 (0.09)	97,50 (0.31)	87,64 (0.54)	74,21 (1.31)
	SNR = 3dB				
A (1250)	20-25-30	78,03 (1.02)	49,20 (2.35)	64,57 (0.95)	31,95 (1.13)
	SNR = -3 dB				
A (1250)	20-25-30	84,22 (1.12)	70,83 (2.63)	57,02 (0.71)	28,11 (1.06)
	SNR = -11 dB				
A (1250)	20-25-30	66,48 (2.61)	46,62 (5.66)	39,89 (0.74)	15,51 (1.01)
	eine Zeitscheibe des Zeitfensters				
V (585)	9-15-30	64,2 (0.20)	46,29 (0.49)	54,91 (0.85)	33,38 (1.26)
	ganzes Zeitfenster				
V (585)	9-15-30	73,67 (0.25)	56,22 (0.57)	53,12 (0.72)	32,58 (0.95)
Erkennungsleistung auf Buchstabenebene					
	ungestört				
A (1250)	20-25-30	99.16 (0.29)	99.25 (0.24)	87.08 (1.37)	83.76 (1.71)
		97.63 (0.87)	97.83 (0.54)	81.07 (1.66)	78.11 (1.84)
	SNR = 3 dB				
A (1250)	20-25-30	95.90 (0.79)	95.50 (0.73)	65.64 (1.40)	56.04 (1.41)
		76.83 (1.92)	85.52 (1.30)	43.28 (1.70)	41.96 (1.21)
	SNR = -3 dB				
A (1250)	20-25-30	95.69 (1.22)	97.28 (0.59)	54.52 (1.83)	44.50 (1.50)
	SNR = -11 dB				
A (1250)	20-25-30	91.22 (3.19)	95.01 (2.34)	30.07 (1.70)	24.63 (1.36)
		85.46 (4.41)	90.04 (4.46)	21.59 (1.62)	17.59 (1.29)
	eine Zeitscheibe des Zeitfensters				
V (585)	9-15-30	44.20 (1.42)	57.48 (1.09)	30.22 (1.52)	38.07 (1.50)
		27.92 (0.82)	39.02 (0.92)	16.14 (1.15)	23.32 (1.59)
	ganzes Zeitfenster				
V (585)	9-15-30	59.84 (1.67)	73.28 (1.16)	34.76 (1.62)	41.67 (1.72)
		45.20 (1.10)	58.35 (0.99)	20.88 (0.76)	27.87 (1.91)

Tabelle B.8: Übersicht über die Klassifikationsleistungen für die Einzelkanäle bei unterschiedlichen Signal-zu-Rausch-Verhältnissen des akustischen Kanals sowie für unterschiedliche visuelle Merkmale.

ungestört	SNR = 3 dB	SNR = -3 dB	SNR = -11 dB	Getestet mit ...
trainiert mit ungestörten Daten				
99.16 (0.29)	7.18 (1.07)	4.02 (0.55)	3.18 (0.50)	Training
99.25 (0.24)	8.72 (1.05)	5.62 (0.64)	4.44 (0.46)	KSPM(trai)
87.08 (1.37)	6.22 (0.90)	3.81 (0.66)	3.04 (0.50)	Test
83.76 (1.71)	7.69 (0.86)	5.18 (0.82)	4.40 (0.74)	KSPM(test)
trainiert mit gestörten Daten: SNR = 3 dB				
42.80 (2.07)	95.90 (0.79)	59.89 (1.09)	16.88 (0.85)	Training
31.65 (1.61)	95.50 (0.73)	53.86 (1.20)	17.37 (0.86)	KSPM(trai)
42.58 (2.71)	65.64 (1.40)	50.79 (1.60)	14.49 (1.03)	Test
30.97 (1.76)	56.04 (1.41)	41.64 (1.76)	13.83 (1.13)	KSPM(test)
trainiert mit gestörten Daten: -3 dB				
35.11 (1.42)	68.05 (1.64)	95.69 (1.22)	42.20 (0.93)	Training
25.51 (0.97)	62.30 (1.23)	97.28 (0.59)	41.05 (0.99)	KSPM(trai)
35.54 (2.16)	54.53 (2.21)	54.52 (1.83)	28.91 (1.33)	Test
25.03 (1.15)	43.53 (1.80)	44.50 (1.50)	22.75 (1.15)	KSPM(test)
trainiert mit gestörten Daten: -11 dB				
11.76 (1.23)	19.79 (1.39)	26.26 (1.26)	91.22 (3.19)	Training
11.10 (1.14)	18.28 (1.25)	23.31 (1.19)	95.01 (2.34)	KSPM(trai)
12.27 (1.43)	19.06 (1.58)	23.36 (1.45)	30.07 (1.70)	Test
11.68 (1.53)	17.24 (1.46)	20.24 (1.42)	24.63 (1.36)	KSPM(test)

Tabelle B.9: Klassifikationsleistungen für den akustischen Kanal bei Verwendung der Architektur 20-25-30. Dabei wird jeweils ein akustischer Datensatz zum Training verwendet (vgl. die Blöcke in der Tabelle), während zum Testen auch die Datensätze mit anderem SNR hinzugezogen werden (vgl. Spalten in der Tabelle). Die Klassifikationsleistungen beziehen sich auf die Buchstabenwortebene. Als Entscheidungskriterium für den Assoziativspeicher wird der Hammingabstand verwendet.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
	trainiert mit ungestörten Daten				
E (2360)	29-40-30	99,6 (0.05)	99,59 (0.09)	90,29 (0.45)	80,52 (1.07)
E (1835)	29-40-30	99,55 (0.06)	99,47 (0.12)	90,65 (0.45)	81,20 (1.18)
E (1835)	20-25 \ 30 9-15 /	99,56 (0.05)	99,42 (0.12)	90,67 (0.46)	81,97 (0.98)
E (5360)	29-40-60-30	99,75 (0.03)	99,72 (0.05)	90,33 (0.56)	80,25 (1.34)
E (3635)	29-40-60-30	99,7 (0.04)	99,63 (0.07)	90,32 (0.49)	80,74 (1.03)
E (3635)	20-25-30 \ 30 9-15-30 /	99,62 (0.04)	99,38 (0.14)	89,47 (0.50)	78,16 (1.11)
	trainiert mit gestörten Daten: SNR = 3 dB				
E (2360)	29-40-30	98,65 (0.07)	98,38 (0.12)	77,4 (0.44)	58,73 (1.13)
E (1835)	29-40-30	98,28 (0.07)	97,95 (0.14)	77,11 (0.54)	58,89 (1.28)
E (1835)	20-25 \ 30 9-15 /	98,38 (0.09)	97,97 (0.21)	77,76 (0.46)	59,16 (1.20)
E (5360)	29-40-60-30	99,46 (0.06)	99,13 (0.14)	77,56 (0.65)	59,84 (1.50)
E (3635)	29-40-60-30	99,3 (0.07)	98,86 (0.20)	77,24 (0.55)	59,69 (1.40)
E (3635)	20-25-30 \ 30 9-15-30 /	98,9 (0.10)	98,07 (0.24)	76,57 (0.61)	58,09 (1.40)
	trainiert mit gestörten Daten: SNR -3 dB				
E (2360)	29-40-30	97,57 (0.09)	97,51 (0.11)	72,29 (0.62)	52,77 (1.86)
E (1835)	29-40-30	97,11 (0.08)	96,95 (0.12)	72,3 (0.66)	52,46 (1.60)
E (1835)	20-25 \ 30 9-15 /	97,20 (0.10)	96,90 (0.13)	72,5 (0.61)	52,61 (1.54)
E (5360)	29-40-60-30	99,12 (0.10)	98,68 (0.16)	72,41 (0.81)	53,45 (1.99)
E (3635)	29-40-60-30	98,82 (0.08)	98,42 (0.11)	71,63 (0.69)	52,69 (1.81)
E (3635)	20-25-30 \ 30 9-15-30 /	98,35 (0.13)	97,20 (0.13)	71,44 (0.76)	52,01 (1.58)
	trainiert mit gestörten Daten: SNR -11 dB				
E (2360)	29-40-30	94,85 (0.12)	95,67 (0.20)	61,63 (0.49)	42,42 (1.44)
E (1835)	29-40-30	94,05 (0.11)	94,66 (0.17)	61,08 (0.49)	42,17 (1.26)
E (1835)	20-25 \ 30 9-15 /	94,17 (0.12)	94,62 (0.16)	62,29 (0.52)	43,58 (1.25)
E (5360)	29-40-60-30	98,23 (0.12)	97,79 (0.20)	61,04 (0.57)	42,07 (1.51)
E (3635)	29-40-60-30	97,39 (0.16)	97,12 (0.21)	60,65 (0.83)	41,87 (1.32)
E (3635)	20-25-30 \ 30 9-15-30 /	95,98 (0.17)	94,56 (0.37)	60,70 (0.49)	41,18 (1.49)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Buchstabenebene					
	mit ungestörten Daten trainiert				
E (2360)	29-40-30	99.43 (0.37)	99.63 (0.26)	91.70 (0.86)	91.45 (0.69)
		98.71 (0.62)	98.80 (0.43)	88.27 (0.97)	87.75 (0.80)
E (1835)	29-40-30	99.01 (0.83)	99.44 (0.38)	91.51 (0.99)	91.02 (0.67)
		98.30 (1.11)	98.74 (0.50)	88.33 (1.65)	87.96 (1.00)
E (1835)	20-25 \ 30 9-15 /	99.53 (0.07)	99.71 (0.07)	92.51 (0.72)	92.00 (0.85)
		99.19 (0.07)	99.16 (0.06)	89.16 (0.76)	88.52 (0.83)
E (5360)	29-40-60-30	99.62 (0.08)	99.76 (0.07)	91.92 (0.66)	90.51 (0.97)
		99.23 (0.07)	99.14 (0.08)	88.70 (0.80)	87.16 (1.12)
E (3635)	29-40-60-30	99.57 (0.09)	99.73 (0.09)	91.76 (0.63)	90.23 (0.97)
		98.75 (0.81)	98.89 (0.44)	88.00 (1.04)	86.54 (1.12)
E (3635)	20-25-30 \ 30 9-15-30 /	99.47 (0.16)	99.63 (0.18)	91.08 (0.76)	89.58 (1.08)
		98.50 (0.85)	98.69 (0.45)	86.80 (1.12)	84.82 (1.25)
	mit gestörten Daten trainiert: SNR = 3 dB				
E (2360)	29-40-30	99.44 (0.11)	99.35 (0.17)	81.27 (0.99)	78.00 (0.96)
		95.71 (2.27)	97.21 (1.06)	73.43 (1.62)	70.65 (1.19)
E (1835)	29-40-30	98.93 (0.31)	98.83 (0.31)	80.25 (0.93)	76.57 (1.14)
		93.40 (2.67)	96.02 (1.32)	70.84 (1.94)	68.92 (1.45)
E (1835)	20-25 \ 30 9-15 /	99.08 (0.31)	98.99 (0.30)	80.99 (0.94)	76.97 (0.83)
		91.15 (3.01)	95.08 (1.45)	69.64 (2.72)	69.06 (1.65)
E (5360)	29-40-60-30	99.53 (0.13)	99.41 (0.16)	81.90 (0.88)	78.00 (1.04)
		97.61 (1.57)	98.32 (0.71)	74.95 (1.49)	71.52 (1.25)
E (3635)	29-40-60-30	99.59 (0.07)	99.48 (0.10)	81.78 (1.05)	78.02 (1.38)
		96.78 (1.81)	98.03 (0.78)	74.20 (1.77)	71.45 (1.38)
E (3635)	20-25-30 \ 30 9-15-30 /	99.45 (0.18)	99.31 (0.17)	80.81 (0.77)	76.91 (1.00)
		96.88 (1.73)	97.93 (0.77)	73.02 (1.70)	69.95 (0.98)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
	mit gestörten Daten trainiert: SNR = -3 dB				
E (2360)	29-40-30	98.74 (0.47)	98.89 (0.33)	76.02 (1.27)	71.91 (1.19)
		87.57 (1.07)	93.96 (0.60)	59.70 (1.06)	61.45 (1.23)
E (1835)	29-40-30	98.57 (0.57)	98.79 (0.36)	76.30 (1.31)	72.22 (1.34)
		87.41 (1.47)	93.85 (0.74)	61.26 (1.72)	62.50 (1.29)
E (1835)	20-25 \ 30 9-15 /	98.94 (0.23)	98.99 (0.20)	76.86 (0.71)	72.27 (0.88)
		88.04 (0.91)	94.24 (0.51)	60.94 (1.27)	62.28 (0.99)
E (5360)	29-40-60-30	99.59 (0.11)	99.45 (0.15)	77.48 (1.04)	73.41 (1.18)
		98.25 (0.89)	98.57 (0.39)	69.48 (1.17)	66.52 (1.29)
E (3635)	29-40-60-30	99.57 (0.11)	99.46 (0.14)	76.21 (0.98)	71.62 (1.09)
		96.51 (1.71)	97.85 (0.71)	66.69 (1.77)	63.93 (1.07)
E (3635)	20-25-30 \ 30 9-15-30 /	99.39 (0.17)	99.25 (0.22)	76.48 (1.12)	72.64 (0.96)
		92.45 (2.42)	95.93 (1.03)	63.70 (2.45)	63.37 (1.24)
	mit gestörten Daten trainiert: SNR = -11 dB				
E (2360)	29-40-30	97.04 (1.09)	98.26 (0.44)	62.18 (1.44)	59.47 (1.71)
		81.19 (0.80)	91.06 (0.45)	43.67 (1.15)	48.26 (1.39)
E (1835)	29-40-30	96.56 (1.31)	98.14 (0.55)	61.39 (1.28)	58.49 (1.33)
		81.43 (0.90)	91.08 (0.54)	43.01 (1.03)	47.47 (1.16)
E (1835)	20-25 \ 30 9-15 /	97.57 (0.91)	98.48 (0.43)	62.39 (1.10)	59.20 (1.19)
		81.19 (1.40)	90.85 (0.70)	43.63 (1.58)	47.72 (1.27)
E (5360)	29-40-60-30	99.33 (0.71)	99.62 (0.30)	62.87 (1.12)	58.94 (1.25)
		92.12 (2.09)	96.23 (0.87)	49.51 (2.18)	49.79 (1.45)
E (3635)	29-40-60-30	99.29 (0.67)	99.56 (0.28)	62.15 (1.24)	58.18 (1.31)
		88.78 (2.20)	94.75 (0.97)	46.61 (1.97)	48.26 (1.43)
E (3635)	20-25-30 \ 30 9-15-30 /	99.14 (0.34)	99.31 (0.18)	61.88 (1.29)	57.76 (1.55)
		86.18 (1.83)	93.29 (0.80)	44.43 (1.96)	46.95 (1.45)

Tabelle B.10: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die akustischen Daten sind unterschiedlich stark gestört. Die Fusionsarchitektur ist E.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
	mit ungestörten Daten trainiert				
E (2360)	29-40-30	99,91 (0.02)	99,97 (0.01)	86,85 (0.32)	73,30 (1.17)
E (1835)	29-40-30	99,9 (0.02)	99,93 (0.02)	86,42 (0.33)	72,56 (1.41)
E (1835)	20-25 \ 30 9-15 /	99,88 (0.02)	99,90 (0.02)	86,49 (0.55)	73,11 (1.89)
E (5360)	29-40-60-30	99,94 (0.02)	99,96 (0.01)	84,73 (0.36)	71,90 (1.32)
E (3635)	29-40-60-30	99,93 (0.02)	99,90 (0.05)	84,69 (0.49)	72,02 (1.92)
E (3635)	20-25-30 \ 30 9-15-30 /	99,87 (0.04)	99,82 (0.07)	82,81 (0.59)	68,66 (1.69)
	mit gestörten Daten trainiert: SNR = 3 dB				
E (2360)	29-40-30	99,65 (0.06)	99,77 (0.04)	72,95 (0.55)	52,80 (2.25)
E (1835)	29-40-30	99,62 (0.07)	99,71 (0.06)	72,96 (0.58)	53,75 (1.59)
E (1835)	20-25 \ 30 9-15 /	99,5 (0.07)	99,48 (0.09)	73,11 (0.70)	52,91 (2.06)
E (5360)	29-40-60-30	99,95 (0.02)	99,94 (0.04)	71,90 (0.66)	52,89 (2.30)
E (3635)	29-40-60-30	99,89 (0.03)	99,88 (0.06)	72,56 (0.51)	53,47 (2.00)
E (3635)	20-25-30 \ 30 9-15-30 /	99,48 (0.08)	98,79 (0.20)	70,71 (0.51)	51,41 (1.99)
	mit gestörten Daten trainiert: SNR = -3 dB				
E (2360)	29-40-30	99,45 (0.07)	99,61 (0.07)	67,55 (0.68)	47,58 (1.99)
E (1835)	29-40-30	99,38 (0.08)	99,55 (0.09)	67,56 (0.68)	47,81 (1.45)
E (1835)	20-25 \ 30 9-15 /	99,12 (0.11)	98,60 (0.29)	67,96 (0.59)	47,90 (2.26)
E (5360)	29-40-60-30	99,89 (0.04)	99,84 (0.07)	67,03 (0.72)	46,97 (2.31)
E (3635)	29-40-60-30	99,85 (0.04)	99,86 (0.05)	66,78 (0.66)	46,49 (1.98)
E (3635)	20-25-30 \ 30 9-15-30 /	98,92 (0.10)	97,60 (0.27)	66,33 (0.73)	47,11 (2.53)
	mit gestörten Daten trainiert: SNR = -11 dB				
E (2360)	29-40-30	98,92 (0.08)	99,21 (0.12)	57,98 (0.81)	39,83 (2.05)
E (1835)	29-40-30	98,78 (0.08)	99,02 (0.14)	57,90 (0.95)	39,44 (2.07)
E (1835)	20-25 \ 30 9-15 /	98,49 (0.12)	97,91 (0.27)	59,39 (0.74)	41,11 (2.02)
E (5360)	29-40-60-30	99,72 (0.07)	99,56 (0.14)	58,14 (0.68)	39,32 (2.34)
E (3635)	29-40-60-30	99,58 (0.07)	99,55 (0.13)	58,41 (0.70)	39,66 (1.80)
E (3635)	20-25-30 \ 30 9-15-30 /	97,84 (0.16)	96,63 (0.27)	57,20 (0.73)	39,15 (2.53)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Buchstabenebene					
	mit ungestörten Daten trainiert				
E (2360)	29-40-30	99.66 (0.10)	99.79 (0.08)	87.60 (1.02)	86.83 (0.95)
		99.29 (0.11)	99.17 (0.08)	83.08 (1.03)	81.65 (1.08)
E (1835)	29-40-30	99.63 (0.08)	99.79 (0.05)	87.70 (0.78)	86.54 (0.78)
		99.23 (0.10)	99.11 (0.08)	83.24 (1.00)	81.82 (0.93)
E (1835)	20-25 \ 30 9-15 /	99.68 (0.09)	99.80 (0.06)	87.76 (0.71)	86.63 (1.15)
		98.29 (1.17)	98.76 (0.51)	81.59 (1.63)	81.56 (1.03)
E (5360)	29-40-60-30	99.57 (0.07)	99.51 (0.13)	86.55 (0.79)	83.70 (1.00)
		99.27 (0.08)	99.19 (0.08)	81.07 (0.93)	78.58 (1.18)
E (3635)	29-40-60-30	99.62 (0.07)	99.59 (0.13)	86.71 (0.73)	84.90 (0.91)
		99.22 (0.11)	99.14 (0.10)	81.77 (0.88)	80.22 (1.19)
E (3635)	20-25-30 \ 30 9-15-30 /	99.47 (0.10)	99.44 (0.14)	85.63 (0.79)	83.62 (1.02)
		99.09 (0.16)	98.97 (0.19)	79.63 (0.96)	77.29 (1.12)
	mit gestörten Daten trainiert: SNR = 3 dB				
E (2360)	29-40-30	99.65 (0.50)	99.41 (0.10)	73.61 (0.84)	69.32 (1.17)
		97.56 (1.59)	98.49 (0.62)	64.54 (1.62)	61.87 (1.43)
E (1835)	29-40-30	99.64 (0.05)	99.43 (0.09)	74.41 (1.13)	69.89 (1.17)
		96.72 (1.85)	98.20 (0.71)	64.27 (1.75)	61.64 (1.77)
E (1835)	20-25 \ 30 9-15 /	99.57 (0.04)	99.34 (0.07)	74.20 (1.09)	69.72 (1.31)
		93.97 (2.03)	97.06 (0.79)	61.88 (2.18)	60.27 (1.81)
E (5360)	29-40-60-30	99.63 (0.04)	99.38 (0.07)	75.68 (1.45)	70.63 (1.61)
		99.39 (0.05)	99.21 (0.05)	68.06 (1.50)	63.47 (1.51)
E (3635)	29-40-60-30	99.65 (0.05)	99.42 (0.09)	75.08 (1.20)	70.46 (1.76)
		98.98 (0.86)	99.09 (0.33)	67.81 (1.34)	63.84 (1.69)
E (3635)	20-25-30 \ 30 9-15-30 /	99.59 (0.06)	99.35 (0.09)	74.36 (1.30)	69.08 (1.39)
		97.03 (1.71)	97.93 (0.85)	65.20 (2.17)	60.94 (1.63)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
	mit gestörten Daten trainiert: SNR = -3 dB				
E (2360)	29-40-30	99.27 (0.29)	99.02 (0.31)	68.81 (1.00)	63.93 (0.99)
		97.21 (1.52)	98.13 (0.61)	59.51 (1.55)	56.73 (1.16)
E (1835)	29-40-30	99.25 (0.27)	99.04 (0.27)	69.24 (1.01)	64.95 (1.12)
		96.32 (1.83)	97.80 (0.75)	59.46 (1.67)	56.65 (1.33)
E (1835)	20-25 \ 30 9-15 /	99.41 (0.10)	99.21 (0.11)	69.58 (1.00)	65.41 (1.62)
		94.46 (1.91)	97.12 (0.76)	57.74 (1.82)	56.86 (1.53)
E (5360)	29-40-60-30	99.62 (0.10)	99.37 (0.10)	70.57 (1.50)	65.04 (2.01)
		99.43 (0.13)	99.26 (0.12)	62.14 (1.64)	57.33 (1.99)
E (3635)	29-40-60-30	99.53 (0.11)	99.26 (0.11)	70.10 (1.34)	64.67 (1.51)
		98.99 (0.76)	99.02 (0.32)	61.58 (1.57)	57.12 (1.46)
E (3635)	20-25-30 \ 30 9-15-30 /	99.45 (0.10)	99.19 (0.12)	70.61 (1.37)	65.96 (1.55)
		95.40 (1.83)	97.18 (0.75)	59.63 (2.21)	57.60 (1.72)
	mit gestörten Daten trainiert: SNR = -11 dB				
E (2360)	29-40-30	99.40 (0.27)	99.32 (0.29)	56.80 (1.26)	52.79 (1.61)
		90.67 (2.14)	95.57 (0.84)	44.25 (1.69)	44.15 (1.73)
E (1835)	29-40-30	99.41 (0.26)	99.38 (0.28)	56.94 (1.18)	52.81 (1.35)
		89.99 (2.21)	95.27 (0.93)	43.35 (1.71)	43.00 (1.81)
E (1835)	20-25 \ 30 9-15 /	98.90 (0.65)	99.08 (0.28)	57.80 (1.14)	53.42 (1.36)
		85.40 (1.34)	93.14 (0.58)	40.81 (1.17)	43.19 (1.59)
E (5360)	29-40-60-30	99.68 (0.12)	99.57 (0.14)	58.36 (1.67)	53.02 (1.90)
		96.66 (1.79)	98.26 (0.73)	46.96 (1.95)	44.52 (1.82)
E (3635)	29-40-60-30	99.14 (0.80)	99.30 (0.32)	58.10 (1.63)	53.51 (1.36)
		94.34 (2.21)	97.18 (0.93)	45.83 (2.10)	44.34 (1.37)
E (3635)	20-25-30 \ 30 9-15-30 /	98.80 (0.26)	98.66 (0.26)	58.02 (1.25)	52.86 (1.64)
		88.21 (2.02)	93.82 (0.83)	43.24 (1.66)	42.53 (1.79)

Tabelle B.11: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die akustischen Daten sind unterschiedlich stark gestört. Die Fusionsarchitektur ist E.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
	mit ungestörten Daten trainiert				
$I1$ (2360)	29-40-30	99,25 (0.06)	99,06 (0.09)	90,05 (0.57)	80,09 (1.40)
$I1(t)$ (1835)	29-40-30	99,1 (0.09)	98,73 (0.17)	90,27 (0.62)	80,62 (1.17)
$H1$ (1835)	20-25 \ 30 9-15 /	99,17 (0.06)	98,86 (0.14)	90,75 (0.65)	82,50 (1.33)
$I2$ (5360)	29-40-60-30	99,53 (0.05)	99,38 (0.09)	90,01 (0.59)	79,61 (1.49)
$I2(t)$ (3635)	29-40-60-30	99,42 (0.06)	99,14 (0.12)	89,78 (0.50)	79,51 (1.14)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99,35 (0.06)	98,95 (0.14)	90,18 (0.64)	80,50 (0.92)
C (1800)	20-25-30 \ 30 9-15-30 /	99,58 (0.04)	99,39 (0.09)	87,43 (0.60)	74,59 (1.45)
	mit gestörten Daten trainiert: SNR = 3 dB				
$I1$ (2360)	29-40-30	79,85 (1.17)	53,76 (2.71)	66,97 (0.94)	35,56 (1.44)
$I1(t)$ (1835)	29-40-30	94,64 (0.24)	90,96 (0.50)	72,8 (0.63)	49,40 (1.09)
$H1$ (1835)	20-25 \ 30 9-15 /	92,2 (0.63)	84,75 (1.37)	75,99 (0.76)	56,36 (1.47)
$I2$ (5360)	29-40-60-30	96,78 (0.41)	92,96 (0.99)	70,1 (0.95)	42,83 (1.31)
$I2(t)$ (3635)	29-40-60-30	98,04 (0.19)	96,70 (0.43)	73,06 (0.87)	50,35 (1.18)
$H1$ (3635)	20-25-30 \ 30 9-15-30 /	97,81 (0.10)	95,49 (0.29)	77,55 (0.65)	59,29 (1.42)
C (1800)	20-25-30 \ 30 9-15-30 /	92,66 (0.47)	84,19 (1.16)	74,43 (0.78)	52,93 (1.63)
	mit gestörten Daten trainiert: SNR = -3 dB				
$I1$ (2360)	29-40-30	85,61 (2.32)	72,20 (5.23)	63,2 (0.86)	35,17 (1.60)
$I1(t)$ (1835)	29-40-30	94,09 (0.28)	91,76 (0.56)	66,72 (0.99)	42,21 (1.10)
$H1$ (1835)	20-25 \ 30 9-15 /	94,54 (0.44)	91,36 (0.99)	70,37 (0.69)	49,27 (1.59)
$I2$ (5360)	29-40-60-30	97,55 (0.16)	95,86 (0.32)	64,17 (0.90)	36,84 (1.15)
$I2(t)$ (3635)	29-40-60-30	97,26 (0.24)	95,83 (0.42)	66,02 (0.88)	41,14 (1.25)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97,72 (0.10)	96,00 (0.33)	71,01 (0.79)	50,93 (1.68)
C (1800)	20-25-30 \ 30 9-15-30 /	94,31 (0.31)	89,58 (0.65)	66,68 (0.88)	43,50 (1.66)
	mit gestörten Daten trainiert: SNR = -11 dB				
$I1$ (2360)	29-40-30	70,2 (3.81)	47,09 (7.66)	49,86 (0.74)	23,39 (1.55)
$I1(t)$ (1835)	29-40-30	87,82 (0.95)	84,55 (1.35)	52,92 (0.75)	30,51 (1.46)
$H1$ (1835)	20-25 \ 30 9-15 /	88,38 (1.05)	83,52 (1.76)	58,96 (0.93)	39,37 (1.66)
$I2$ (5360)	29-40-60-30	94,98 (0.65)	91,87 (1.18)	49,38 (0.70)	26,59 (1.66)
$I2(t)$ (3635)	29-40-60-30	94,97 (0.24)	93,04 (0.40)	51,7 (0.71)	31,22 (1.12)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	95,15 (0.26)	92,52 (0.47)	59,52 (0.85)	39,54 (1.60)
C (1800)	20-25-30 \ 30 9-15-30 /	87,11 (1.25)	78,89 (2.69)	57,74 (1.02)	36,42 (1.54)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Buchstabenebene					
	mit ungestörten Daten trainiert				
$I1$ (2360)	29-40-30	99.41 (0.35)	99.58 (0.28)	90.88 (0.95)	90.04 (1.26)
		98.72 (0.50)	98.73 (0.40)	86.96 (1.23)	85.59 (1.60)
$I1(t)$ (1835)	29-40-30	99.36 (0.25)	99.49 (0.25)	91.46 (0.84)	90.04 (1.21)
		98.36 (1.00)	98.54 (0.53)	86.84 (1.19)	85.29 (1.36)
$H1$ (1835)	20-25 \ 30 9-15 /	99.42 (0.27)	99.52 (0.26)	92.46 (1.03)	91.61 (1.55)
		98.38 (0.82)	98.46 (0.46)	88.10 (1.61)	87.42 (1.69)
$I2$ (5360)	29-40-60-30	99.59 (0.15)	99.69 (0.16)	90.69 (1.04)	89.41 (1.44)
		99.14 (0.24)	99.08 (0.25)	86.58 (1.01)	84.62 (1.72)
$I2(t)$ (3635)	29-40-60-30	99.42 (0.27)	99.49 (0.29)	90.33 (0.88)	88.61 (1.29)
		98.52 (0.75)	98.63 (0.42)	85.96 (1.14)	84.17 (1.36)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.60 (0.21)	99.67 (0.22)	91.33 (1.01)	89.83 (1.50)
		99.05 (0.34)	98.92 (0.37)	87.30 (1.23)	85.77 (1.71)
C (1800)	20-25-30 \ 30 9-15-30 /	99.67 (0.16)	99.75 (0.15)	89.00 (0.79)	86.20 (1.39)
		98.45 (1.09)	98.74 (0.56)	83.24 (1.48)	80.85 (1.58)
	mit gestörten Daten trainiert: SNR = 3 dB				
$I1$ (2360)	29-40-30	95.80 (0.72)	95.11 (0.76)	68.76 (1.31)	59.37 (1.32)
		76.44 (1.72)	84.68 (1.29)	45.32 (1.51)	44.92 (1.29)
$I1(t)$ (1835)	29-40-30	98.63 (0.23)	98.37 (0.27)	78.22 (1.49)	72.16 (2.01)
		83.23 (2.02)	91.20 (1.04)	58.14 (2.27)	59.22 (1.89)
$H1$ (1835)	20-25 \ 30 9-15 /	97.08 (0.89)	97.05 (0.57)	81.66 (1.17)	78.62 (1.23)
		78.01 (1.83)	86.87 (0.97)	60.45 (2.03)	65.89 (1.35)
$I2$ (5360)	29-40-60-30	99.47 (0.13)	99.28 (0.17)	71.16 (1.72)	61.83 (1.51)
		92.17 (2.18)	96.06 (0.95)	57.00 (2.64)	51.58 (1.81)
$I2(t)$ (3635)	29-40-60-30	99.50 (0.17)	99.36 (0.18)	77.53 (1.88)	71.50 (1.96)
		92.44 (2.30)	96.22 (1.00)	64.48 (2.73)	61.04 (2.02)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.08 (0.36)	99.04 (0.39)	82.85 (1.18)	79.34 (1.56)
		90.73 (2.65)	94.68 (1.30)	71.01 (2.09)	70.56 (1.58)
C (1800)	20-25-30 \ 30 9-15-30 /	96.15 (1.23)	96.19 (0.73)	77.63 (1.70)	72.36 (1.56)
		78.21 (1.50)	86.75 (0.96)	56.78 (1.89)	58.97 (1.48)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
	mit gestörten Daten trainiert: SNR = -3 dB				
$I1$ (2360)	29-40-30	97.56 (0.82)	97.34 (0.75)	64.56 (1.73)	56.71 (1.97)
		84.54 (2.71)	90.82 (1.78)	45.67 (2.78)	43.77 (2.51)
$I1(t)$ (1835)	29-40-30	98.94 (0.28)	98.90 (0.29)	72.57 (1.30)	65.75 (1.67)
		87.61 (2.07)	93.82 (0.86)	55.75 (2.34)	54.71 (2.20)
$H1$ (1835)	20-25 \ 30 9-15 /	98.91 (0.28)	98.87 (0.23)	77.51 (0.95)	72.91 (1.35)
		88.91 (2.67)	93.63 (1.26)	63.14 (2.63)	62.84 (1.94)
$I2$ (5360)	29-40-60-30	99.67 (0.11)	99.53 (0.13)	64.37 (1.87)	55.15 (1.88)
		96.32 (1.52)	97.97 (0.62)	53.42 (2.06)	46.33 (1.74)
$I2(t)$ (3635)	29-40-60-30	99.59 (0.13)	99.44 (0.14)	70.73 (1.81)	63.30 (1.71)
		94.37 (2.19)	97.07 (0.91)	58.29 (2.76)	53.37 (2.20)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.59 (0.14)	99.65 (0.14)	78.23 (1.00)	73.78 (1.45)
		96.22 (1.93)	97.66 (0.98)	69.37 (2.18)	66.43 (1.97)
C (1800)	20-25-30 \ 30 9-15-30 /	98.63 (0.26)	98.52 (0.25)	71.54 (1.79)	64.80 (1.89)
		86.75 (2.34)	92.26 (1.17)	55.52 (2.64)	53.20 (2.18)
	mit gestörten Daten trainiert: SNR = -11 dB				
$I1$ (2360)	29-40-30	86.49 (4.32)	90.77 (3.02)	42.49 (3.47)	40.09 (2.47)
		73.03 (5.69)	80.13 (5.60)	27.60 (2.79)	28.60 (2.38)
$I1(t)$ (1835)	29-40-30	95.72 (1.20)	97.70 (0.54)	51.98 (2.13)	47.10 (1.50)
		86.40 (1.49)	92.85 (0.77)	36.96 (2.01)	36.24 (1.17)
$H1$ (1835)	20-25 \ 30 9-15 /	95.07 (1.60)	96.88 (0.79)	61.02 (1.30)	58.18 (1.16)
		80.77 (0.91)	88.99 (1.01)	43.09 (1.07)	46.07 (1.19)
$I2$ (5360)	29-40-60-30	98.20 (0.79)	99.13 (0.36)	47.30 (1.35)	41.22 (1.55)
		95.90 (1.34)	97.78 (0.59)	36.56 (1.64)	32.18 (1.69)
$I2(t)$ (3635)	29-40-60-30	98.29 (0.93)	99.15 (0.39)	52.32 (1.80)	47.65 (1.24)
		93.19 (1.49)	96.54 (0.64)	40.30 (1.95)	38.19 (1.40)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97.89 (1.28)	98.93 (0.53)	62.79 (2.25)	60.41 (1.40)
		91.68 (2.12)	95.61 (0.84)	51.51 (2.27)	51.01 (1.57)
C (1800)	20-25-30 \ 30 9-15-30 /	92.82 (1.92)	95.01 (1.19)	56.33 (1.61)	54.07 (1.66)
		78.31 (1.85)	86.47 (1.63)	38.61 (1.11)	41.15 (1.23)

Tabelle B.12: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die akustischen Daten sind unterschiedlich stark gestört. Die Fusionsarchitektur sind unterschiedlich.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
	mit ungestörten Daten trainiert				
$I1$ (2360)	29-40-30	99,15 (0.07)	98,79 (0.16)	88,83 (0.49)	77,33 (1.26)
$I1(t)$ (1835)	29-40-30	98,95 (0.10)	98,41 (0.21)	89,12 (0.50)	78,01 (1.08)
$H1$ (1835)	20-25 \ 30 9-15 /	99,06 (0.06)	98,67 (0.12)	89,69 (0.63)	79,75 (1.14)
$I2$ (5360)	29-40-60-30	99,39 (0.05)	99,13 (0.13)	89,2 (0.54)	77,67 (1.36)
$I2(t)$ (3635)	29-40-30-30	99,26 (0.06)	98,78 (0.17)	88,73 (0.50)	76,88 (1.26)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99,15 (0.06)	98,56 (0.15)	88,24 (0.61)	76,57 (1.36)
C (1800)	20-25-30 \ 30 9-15-30 /	99,68 (0.04)	99,49 (0.12)	86,94 (0.65)	73,52 (1.67)
	mit gestörten Daten trainiert: SNR = 3 dB				
$I1$ (2360)	29-40-30	77,67 (1.30)	48,88 (2.59)	64,83 (0.95)	33,10 (1.23)
$I1(t)$ (1835)	29-40-30	92,58 (0.37)	86,00 (0.99)	70,56 (0.89)	43,69 (1.11)
$H1$ (1835)	20-25 \ 30 9-15 /	93,67 (0.44)	88,15 (1.09)	75,08 (0.73)	54,19 (1.89)
$I2$ (5360)	29-40-60-30	95,66 (0.41)	89,68 (1.06)	67,51 (0.72)	38,94 (1.13)
$I2(t)$ (3635)	29-40-30-30	96,71 (0.21)	93,56 (0.59)	69,28 (0.72)	42,42 (1.14)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	98,17 (0.19)	96,57 (0.48)	74,95 (0.64)	55,59 (1.80)
C (1800)	20-25-30 \ 30 9-15-30 /	94,42 (0.40)	88,74 (1.04)	72,04 (0.64)	50,31 (1.93)
	mit gestörten Daten trainiert: SNR = -3 dB				
$I1$ (2360)	29-40-30	80,54 (2.29)	61,64 (5.16)	58,49 (1.05)	29,20 (1.58)
$I1(t)$ (1835)	29-40-30	92,32 (0.28)	88,48 (0.57)	64,17 (0.87)	37,29 (1.13)
$H1$ (1835)	20-25 \ 30 9-15 /	94,99 (0.26)	92,17 (0.60)	68,92 (0.65)	47,51 (1.41)
$I2$ (5360)	29-40-60-30	96,14 (0.18)	92,64 (0.40)	59,29 (0.88)	31,76 (1.07)
$I2(t)$ (3635)	29-40-30-30	95,1 (0.29)	91,73 (0.62)	60,54 (1.06)	33,05 (1.35)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	98,23 (0.13)	97,19 (0.34)	69,37 (0.76)	48,09 (1.48)
C (1800)	20-25-30 \ 30 9-15-30 /	95,46 (0.26)	92,03 (0.53)	65,43 (0.72)	42,08 (1.50)
	mit gestörten Daten trainiert: SNR = -11 dB				
$I1$ (2360)	29-40-30	65,26 (4.38)	40,84 (8.56)	44,1 (1.02)	17,60 (1.86)
$I1(t)$ (1835)	29-40-30	84,86 (1.12)	80,63 (1.51)	48,44 (0.70)	25,42 (1.46)
$H1$ (1835)	20-25 \ 30 9-15 /	90,7 (0.90)	86,17 (1.61)	57,15 (0.72)	36,87 (2.13)
$I2$ (5360)	29-40-60-30	91,11 (0.64)	85,38 (1.23)	41,26 (0.85)	20,43 (1.51)
$I2(t)$ (3635)	29-40-30-30	91,59 (0.54)	87,85 (0.90)	45,75 (0.98)	24,22 (1.37)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97,11 (0.14)	95,39 (0.33)	57,67 (0.62)	38,35 (1.89)
C (1800)	20-25-30 \ 30 9-15-30 /	90,21 (0.92)	83,87 (1.80)	56,57 (0.78)	36,93 (1.75)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Buchstabenebene					
	mit ungestörten Daten trainiert				
$I1$ (2360)	29-40-30	99.42 (0.25)	99.55 (0.24)	89.35 (0.94)	87.73 (1.37)
		98.66 (0.36)	98.59 (0.34)	84.76 (0.75)	82.51 (1.13)
$I1(t)$ (1835)	29-40-30	99.32 (0.23)	99.48 (0.20)	89.07 (0.81)	86.65 (1.25)
		98.08 (0.85)	98.27 (0.43)	84.38 (1.14)	81.60 (1.47)
$H1$ (1835)	20-25 \ 30 9-15 /	99.49 (0.25)	99.56 (0.27)	90.80 (1.14)	89.84 (1.43)
		98.29 (0.78)	98.41 (0.45)	85.91 (1.55)	84.94 (1.52)
$I2$ (5360)	29-40-60-30	99.55 (0.18)	99.63 (0.19)	89.18 (0.99)	86.81 (1.39)
		99.22 (0.20)	99.12 (0.22)	85.04 (0.97)	82.29 (1.38)
$I2(t)$ (3635)	29-40-30-30	99.23 (0.39)	99.42 (0.31)	88.88 (1.24)	86.38 (1.58)
		98.18 (0.75)	98.39 (0.40)	83.70 (1.48)	81.09 (1.73)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.49 (0.22)	99.60 (0.21)	88.41 (0.91)	86.33 (1.46)
		98.95 (0.29)	98.88 (0.29)	83.76 (0.94)	81.14 (1.38)
C (1800)	20-25-30 \ 30 9-15-30 /	99.67 (0.14)	99.68 (0.15)	87.68 (1.09)	84.55 (1.44)
		98.14 (1.14)	98.39 (0.63)	81.34 (1.78)	78.27 (1.87)
	mit gestörten Daten trainiert: SNR = 3 dB				
$I1$ (2360)	29-40-30	95.89 (0.64)	95.52 (0.61)	67.12 (1.36)	57.48 (1.20)
		78.20 (2.39)	85.67 (1.55)	45.18 (2.25)	42.53 (1.15)
$I1(t)$ (1835)	29-40-30	98.48 (0.32)	98.22 (0.38)	72.59 (1.67)	64.38 (1.84)
		83.85 (1.72)	91.43 (0.83)	52.03 (2.25)	50.45 (1.81)
$H1$ (1835)	20-25 \ 30 9-15 /	97.87 (0.39)	97.52 (0.44)	81.47 (0.72)	77.91 (1.08)
		78.78 (1.45)	88.12 (0.82)	59.59 (2.03)	64.58 (1.60)
$I2$ (5360)	29-40-60-30	99.45 (0.18)	99.26 (0.20)	66.77 (1.24)	55.35 (1.20)
		92.76 (2.34)	96.40 (0.98)	53.26 (2.29)	45.43 (1.53)
$I2(t)$ (3635)	29-40-30-30	99.40 (0.14)	99.22 (0.18)	70.47 (1.46)	60.47 (1.90)
		89.95 (1.87)	95.03 (0.75)	53.72 (1.74)	49.06 (1.58)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.35 (0.19)	99.17 (0.24)	81.63 (1.23)	78.65 (1.20)
		92.50 (2.59)	95.98 (1.22)	69.96 (2.86)	69.63 (2.18)
C (1800)	20-25-30 \ 30 9-15-30 /	97.35 (0.91)	97.28 (0.62)	76.02 (1.58)	70.82 (1.79)
		80.59 (1.35)	89.15 (0.65)	55.65 (1.51)	57.83 (1.90)

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
	mit gestörten Daten trainiert: SNR = -3 dB				
$I1$ (2360)	29-40-30	97.35 (0.90)	97.84 (0.48)	57.24 (1.79)	46.89 (1.81)
		85.79 (2.48)	91.89 (1.33)	40.14 (1.94)	35.58 (1.77)
$I1(t)$ (1835)	29-40-30	98.94 (0.25)	98.76 (0.29)	66.36 (1.63)	57.76 (1.69)
		88.27 (1.96)	93.70 (0.79)	48.24 (2.30)	45.35 (1.77)
$H1$ (1835)	20-25 \ 30 9-15 /	99.10 (0.27)	98.93 (0.30)	75.94 (1.10)	71.59 (1.28)
		88.79 (2.34)	94.01 (1.08)	60.21 (2.02)	60.57 (1.50)
$I2$ (5360)	29-40-60-30	99.67 (0.09)	99.51 (0.12)	59.19 (1.39)	47.61 (1.45)
		98.20 (1.13)	98.70 (0.56)	49.50 (1.42)	39.18 (1.43)
$I2(t)$ (3635)	29-40-30-30	99.23 (0.47)	99.25 (0.25)	60.02 (1.80)	49.23 (2.06)
		93.78 (2.35)	96.73 (0.98)	46.45 (2.48)	39.70 (2.16)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.67 (0.08)	99.51 (0.12)	76.99 (1.22)	72.41 (1.71)
		97.45 (1.28)	98.31 (0.57)	67.98 (1.61)	65.08 (1.26)
C (1800)	20-25-30 \ 30 9-15-30 /	98.84 (0.25)	98.74 (0.27)	69.83 (1.39)	62.45 (1.63)
		87.16 (2.20)	93.14 (0.97)	53.21 (2.58)	50.36 (1.48)
	mit gestörten Daten trainiert: SNR = -11 dB				
$I1$ (2360)	29-40-30	85.06 (4.17)	90.13 (3.26)	33.71 (2.91)	30.51 (2.83)
		73.92 (6.23)	79.90 (6.08)	22.26 (2.62)	20.84 (2.65)
$I1(t)$ (1835)	29-40-30	94.92 (1.13)	97.44 (0.52)	42.37 (1.92)	38.60 (1.96)
		88.07 (1.40)	93.43 (0.74)	29.42 (1.77)	28.58 (1.80)
$H1$ (1835)	20-25 \ 30 9-15 /	95.14 (1.96)	97.17 (0.86)	57.94 (2.27)	54.37 (1.77)
		82.16 (1.08)	90.32 (0.83)	41.34 (1.30)	42.62 (1.32)
$I2$ (5360)	29-40-60-30	97.57 (0.89)	98.95 (0.33)	34.17 (1.40)	28.23 (1.23)
		96.09 (1.18)	98.07 (0.47)	25.74 (1.25)	21.16 (1.18)
$I2(t)$ (3635)	29-40-30-30	97.74 (1.18)	98.93 (0.50)	42.04 (2.05)	37.51 (1.97)
		95.42 (1.36)	97.43 (0.60)	31.76 (1.80)	28.84 (1.76)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	99.33 (0.64)	99.60 (0.26)	60.02 (1.31)	55.85 (1.65)
		92.59 (2.01)	96.01 (0.79)	47.65 (1.43)	45.90 (1.47)
C (1800)	20-25-30 \ 30 9-15-30 /	94.11 (1.55)	96.32 (0.72)	54.47 (1.88)	50.66 (1.62)
		81.03 (1.22)	89.26 (1.06)	36.94 (1.31)	37.08 (1.82)

Tabelle B.13: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die akustischen Daten sind unterschiedlich stark gestört. Die Fusionsarchitektur sind unterschiedlich.

Trainingsdaten	KSPM (trai)	Testdaten	KSPM (test)	SNR
Schwelle -100				
47.76 (2.34)	60.55 (1.82)	33.43 (2.27)	41.01 (1.93)	ungestört
47.69 (2.29)	60.47 (1.77)	32.48 (1.92)	39.96 (1.63)	3 dB
47.57 (2.30)	60.42 (1.80)	32.11 (1.89)	39.47 (1.59)	-3 dB
47.35 (1.99)	60.36 (1.69)	31.22 (1.57)	38.76 (1.47)	-11 dB
Schwelle 0				
60.68 (2.32)	67.12 (1.87)	49.91 (2.60)	49.95 (2.11)	ungestört
48.31 (2.57)	60.52 (1.83)	33.85 (2.12)	40.39 (1.71)	3 dB
47.56 (2.30)	60.42 (1.80)	32.27 (1.89)	39.56 (1.60)	-3 dB
47.31 (1.99)	60.32 (1.70)	31.24 (1.57)	38.78 (1.47)	-11 dB
Schwelle 100				
88.40 (1.13)	91.74 (0.94)	77.20 (1.72)	75.24 (2.19)	ungestört
73.99 (1.85)	77.05 (1.59)	54.18 (1.70)	48.52 (1.91)	3 dB
62.69 (2.38)	70.16 (2.12)	41.77 (1.88)	41.72 (1.87)	-3 dB
47.24 (1.94)	60.69 (1.74)	30.23 (1.73)	37.57 (1.65)	-11 dB
Schwelle 150				
92.21 (0.76)	95.57 (0.56)	81.64 (1.61)	81.64 (1.69)	ungestört
80.73 (1.47)	83.20 (1.27)	58.33 (1.50)	51.48 (1.85)	3 dB
71.80 (2.59)	78.29 (1.92)	46.06 (2.37)	43.50 (1.83)	-3 dB
49.93 (2.43)	63.93 (2.23)	29.39 (1.72)	36.22 (1.85)	-11 dB
Schwelle 200				
93.80 (0.74)	96.60 (0.41)	83.03 (1.66)	82.74 (1.64)	ungestört
84.32 (0.95)	87.32 (0.90)	59.98 (1.45)	54.06 (1.55)	3 dB
77.90 (2.15)	83.37 (1.39)	49.05 (1.85)	45.17 (1.78)	-3 dB
57.12 (2.85)	71.58 (2.02)	28.19 (1.59)	34.05 (1.96)	-11 dB
Schwelle als logistische Funktion				
99.16 (0.33)	99.25 (0.24)	87.14 (1.36)	83.84 (1.70)	ungestört
91.23 (1.13)	93.13 (0.76)	63.31 (1.38)	56.24 (1.37)	3 dB
78.00 (2.91)	83.56 (2.22)	49.50 (2.26)	45.06 (1.88)	-3 dB
49.15 (3.21)	61.78 (2.74)	30.77 (1.82)	37.49 (2.27)	-11 dB
akustische Entscheidung hat Vorrang				
99.16 (0.28)	99.25 (0.24)	87.12 (1.36)	83.82 (1.70)	ungestört
95.93 (0.79)	95.52 (0.73)	65.64 (1.40)	56.04 (1.41)	3 dB
95.69 (1.22)	97.28 (0.59)	54.52 (1.83)	44.50 (1.50)	-3 dB
91.25 (3.19)	95.03 (2.34)	30.14 (1.68)	24.66 (1.35)	-11 dB

Tabelle B.14: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, Verwendung der Architektur L, verschiedene Fusionsstrategien. Als akustisches Netz wird die Architektur 20-25-30, als visuelles Netz die Architektur 9-15-30 verwendet. Als Entscheidungskriterium für den Assoziativspeicher wird der Hammingabstand verwendet. Die Klassifikationsergebnisse beziehen sich auf die Buchstabenwortebene. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet.

Trainingsdaten	KSPM (trai)	Testdaten	KSPM (test)	SNR
Schwelle -100				
60.87 (1.87)	74.06 (1.22)	36.86 (1.73)	43.24 (1.94)	ungestört
60.86 (1.87)	74.06 (1.21)	35.76 (1,61)	42.30 (1.84)	3 dB
60.73 (1.81)	74.03 (1.20)	35.35 (1.60)	42.12 (1.79)	-3 dB
60.82 (1.85)	74.07 (1.22)	35.08 (1.58)	41.92 (1.79)	-11 dB
Schwelle 0				
81.04 (1.78)	84.50 (1.30)	63.70 (2.00)	61.97 (1.80)	ungestört
65.03 (2.67)	74.83 (1.39)	42.79 (2.33)	44.30 (1.85)	3 dB
61.19 (2.17)	74.15 (1.33)	36.89 (1.98)	42.32 (1.86)	-3 dB
60.49 (1.98)	73.77 (1.30)	34.85 (1.67)	41.70 (1.82)	-11 dB
Schwelle 100				
93.10 (0.73)	95.80 (0.58)	81.08 (1.46)	80.59 (1.84)	ungestört
84.63 (1.25)	86.78 (1.10)	58.31 (1.56)	52.73 (1.75)	3 dB
75.44 (1.65)	82.71 (1.28)	44.45 (1.65)	43.46 (1.82)	-3 dB
59.79 (1.98)	73.59 (1.42)	32.18 (1.89)	38.73 (2.07)	-11 dB
Schwelle 150				
94.79 (0.46)	97.06 (0.29)	83.00 (1.54)	82.67 (1.68)	ungestört
86.97 (1.08)	89.20 (0.87)	60.40 (1.47)	55.39 (1.56)	3 dB
80.54 (1.98)	86.64 (1.35)	47.95 (1.54)	45.34 (1.79)	-3 dB
62.06 (2.18)	76.08 (1.71)	30.47 (1.64)	36.43 (2.20)	-11 dB
Schwelle 200				
96.35 (0.78)	97.87 (0.43)	84.40 (1.60)	82.91 (1.79)	ungestört
89.18 (0.87)	91.45 (0.71)	61.66 (1.45)	56.35 (1.54)	3 dB
85.10 (1.44)	89.76 (0.99)	50.73 (1.54)	46.72 (1.68)	-3 dB
68.40 (2.34)	80.59 (1.63)	30.15 (1.75)	34.64 (2.13)	-11 dB
Schwelle als logistische Funktion				
98.98 (0.47)	99.16 (0.31)	86.90 (1.44)	83.70 (1.74)	ungestört
91.28 (0.90)	93.07 (0.70)	62.87 (1.40)	56.72 (1.34)	3 dB
82.77 (2.09)	88.07 (1.49)	49.31 (1.70)	46.08 (2.02)	-3 dB
61.09 (2.24)	74.18 (1.53)	33.78 (1.97)	39.95 (2.38)	-11 dB
akustische Entscheidung hat Vorrang				
99.16 (0.28)	99.25 (0.24)	87.11 (1.36)	83.81 (1.70)	ungestört
95.94 (0.78)	95.55 (0.73)	65.67 (1.41)	56.07 (1.42)	3 dB
95.69 (1.22)	97.28 (0.59)	54.55 (1.84)	44.56 (1.54)	-3 dB
91.27 (3.19)	95.03 (2.34)	30.14 (1.67)	24.69 (1.34)	-11 dB

Tabelle B.15: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, Verwendung der Architektur L, verschiedene Fusionsstrategien. Als akustisches Netz wird die Architektur 20-25-30, als visuelles Netz die Architektur 9-15-30 verwendet. Als Entscheidungskriterium für den Assoziativspeicher wird der Hammingabstand verwendet. Die Klassifikationsergebnisse beziehen sich auf die Buchstabenwortebene. Als visuelles Merkmal wird das ganze Zeitfenster verwendet.

Anhang C

Fusionsergebnisse für die gemeinsame Dimensionsreduktion

In diesem Anhang sind alle Ergebnisse aufgelistet, die mit der gemeinsamen Fusion aller Rauschstufen erzielt wurden. Aufgetragen sind die Trainings- und Testfehler sowie die klassenspezifischen Mittel auf Trainings- und Testdatensatz für die einzelnen Architekturen sowohl auf Segmentklassifikationsebene als auch auf der Ebene der Buchstabenwortklassifikationen. In Klammern sind jeweils die 95%igen Konfidenzintervalle angegeben. Für jede Architektur ist die Bezeichnung, die Anzahl der Gewichte sowie die Topologie dargestellt.
Die Tabellen für die Klassifikationsleistungen auf Buchstabenwortebene geben für jede Architektur in der ersten Zeile die Ergebnisse wieder für den Fall, daß der Hammingabstand als Entscheidungskriterium des Assoziativspeichers verwendet wird, die zweite Zeile enthält die Ergebnisse, die mit dem Überlapp als Entscheidungskriterium erreicht werden.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
E (2360)	29-40-30	97.31 (0.09)	97.34 (0.12)	73.31 (0.46)	55.02 (1.57)
E (1835)	29-40-30	96.36 (0.06)	96.44 (0.11)	73.08 (0.50)	54.85 (1.38)
E (5360)	29-40-60-30	99.37 (0.07)	99.24 (0.14)	72.86 (0.50)	55.55 (1.56)
E (3635)	29-40-60-30	99.00 (0.06)	98.88 (0.10)	72.20 (0.38)	54.32 (1.47)
Erkennungsleistung auf Buchstabenebene					
E (2360)	29-40-30	94.26 (1.58)	96.09 (0.89)	72.56 (1.18)	69.89 (1.12)
		74.51 (1.49)	86.54 (0.98)	51.35 (1.45)	57.01 (1.25)
E (1835)	29-40-30	94.14 (1.50)	95.46 (1.07)	72.59 (1.07)	69.21 (1.33)
		70.85 (2.42)	83.65 (1.67)	48.79 (1.80)	54.97 (1.68)
E (5360)	29-40-60-30	99.22 (0.24)	99.30 (0.25)	75.75 (0.81)	72.48 (0.82)
		94.52 (2.17)	96.74 (0.89)	65.71 (2.03)	64.37 (1.30)
E (3635)	29-40-60-30	98.98 (0.60)	99.25 (0.35)	74.75 (0.70)	71.24 (0.71)
		84.71 (2.17)	92.20 (1.00)	57.33 (2.25)	60.15 (1.06)

Tabelle C.1: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitektur ist E. Trainiert und getestet wird mit allen Rauschstufen gemeinsam.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
E (2360)	29-40-30	99.52 (0.04)	99.69 (0.03)	68.86 (0.42)	49.97 (1.51)
E (1835)	29-40-30	99.41 (0.03)	99.60 (0.03)	68.65 (0.42)	49.87 (1.52)
E (5360)	29-40-60-30	99.98 (0.01)	99.99 (0.01)	69.15 (0.50)	49.91 (1.90)
E (3635)	29-40-60-30	99.96 (0.01)	99.97 (0.02)	68.79 (0.58)	49.85 (2.08)
Erkennungsleistung auf Buchstabenebene					
E (2360)	29-40-30	97.76 (1.62)	98.85 (0.65)	68.51 (1.58)	65.15 (0.61)
		90.85 (2.10)	95.65 (0.88)	55.25 (1.90)	55.36 (1.23)
E (1835)	29-40-30	97.83 (1.48)	98.73 (0.63)	68.62 (1.47)	65.17 (0.69)
		88.51 (2.47)	94.51 (1.04)	53.46 (2.11)	54.54 (1.39)
E (5360)	29-40-60-30	99.67 (0.03)	99.44 (0.06)	70.39 (0.90)	65.16 (1.08)
		99.50 (0.06)	99.33 (0.07)	62.33 (0.78)	57.86 (1.18)
E (3635)	29-40-60-30	99.71 (0.06)	99.56 (0.10)	70.31 (0.86)	66.01 (1.25)
		99.46 (0.09)	99.31 (0.07)	62.20 (0.85)	58.64 (1.28)

Tabelle C.2: Klassifikationsleistungen für die kombinierten akustischen und visuellen Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitektur ist E. Trainiert und getestet wird mit allen Rauschstufen gemeinsam.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
A (1250)	20-25-30	82.13 (0.20)	70.85 (0.40)	60.01 (0.60)	34.78 (0.66)
$I1$ (2360)	29-40-30	95.17 (0.12)	93.16 (0.26)	70.89 (0.61)	51.59 (0.62)
$I1(t)$ (1835)	29-40-30	93.99 (0.11)	91.71 (0.25)	71.87 (0.60)	52.84 (0.82)
$H1$ (1835)	20-25 \ 30 9-15 /	93.96 (0.10)	91.67 (0.25)	71.80 (0.49)	53.69 (1.13)
$I2$ (5360)	29-40-60-30	98.23 (0.08)	97.35 (0.15)	69.97 (0.49)	50.91 (0.89)
$I2(t)$ (3635)	29-40-60-30	97.32 (0.13)	96.00 (0.25)	70.50 (0.64)	51.22 (0.92)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97.13 (0.10)	95.66 (0.27)	71.35 (0.47)	53.40 (0.87)
C (1800)	20-25-30 \ 30 9-15-30 /	92.38 (0.09)	87.81 (0.23)	69.32 (0.59)	48.55 (1.10)
Erkennungsleistung auf Buchstabenebene					
A (1250)	20-25-30	84.70 (1.38)	90.74 (0.61)	47.30 (1.73)	43.27 (1.03)
		71.25 (0.96)	81.74 (0.63)	30.39 (0.93)	30.73 (0.87)
$I1$ (2360)	29-40-30	96.10 (1.09)	96.90 (0.61)	74.06 (1.15)	70.60 (0.83)
		76.83 (1.17)	86.76 (0.97)	51.52 (0.98)	56.09 (1.08)
$I1(t)$ (1835)	29-40-30	94.21 (1.14)	95.03 (0.80)	74.07 (1.42)	71.49 (0.75)
		72.93 (1.62)	83.38 (1.30)	50.37 (1.37)	56.02 (1.01)
$H1$ (1835)	20-25 \ 30 9-15 /	93.99 (1.63)	95.26 (0.82)	73.65 (1.68)	71.31 (0.92)
		72.91 (1.62)	83.15 (1.26)	50.86 (1.31)	56.55 (1.08)
$I2$ (5360)	29-40-60-30	98.70 (0.60)	98.92 (0.36)	74.26 (0.96)	70.61 (0.69)
		84.33 (1.98)	92.29 (0.89)	54.89 (1.80)	58.28 (0.93)
$I2(t)$ (3635)	29-40-60-30	98.74 (0.26)	98.76 (0.25)	75.22 (1.02)	71.59 (0.82)
		81.65 (1.56)	90.77 (0.78)	53.74 (1.22)	57.99 (0.99)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	96.04 (1.59)	97.36 (0.77)	74.40 (1.64)	72.49 (1.06)
		80.50 (1.94)	89.05 (1.19)	55.47 (1.52)	60.20 (1.28)
C (1800)	20-25-30 \ 30 9-15-30 /	89.77 (1.57)	92.10 (0.68)	66.18 (1.30)	63.49 (1.11)
		68.70 (1.05)	79.52 (0.76)	43.76 (0.92)	47.79 (1.41)

Tabelle C.3: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitekturen sind unterschiedlich. Trainiert und getestet wird mit allen Rauschstufen gemeinsam.

Netz (Anzahl Gewichte)	Architektur	Trainings-daten	KSPM (trai)	Testdaten	KSPM (test)
Erkennungsleistung auf Segmentebene					
$I1$ (2360)	29-40-30	95.03 (0.15)	92.68 (0.24)	70.23 (0.47)	50.42 (1.10)
$I1(t)$ (1835)	29-40-30	94.67 (0.17)	92.24 (0.27)	71.41 (0.45)	51.92 (0.97)
$H1$ (1835)	20-25 \ 30 9-15 /	94.84 (0.12)	92.43 (0.26)	70.62 (0.45)	51.87 (1.19)
$I2$ (5360)	29-40-60-30	97.22 (0.09)	95.73 (0.21)	67.02 (0.57)	46.29 (0.92)
$I2(t)$ (3635)	29-40-60-30	96.40 (0.15)	94.68 (0.23)	68.55 (0.65)	48.85 (0.95)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97.52 (0.09)	96.01 (0.23)	70.10 (0.45)	51.23 (1.26)
C (1800)	20-25-30 \ 30 9-15-30 /	94.00 (0.13)	90.43 (0.32)	67.02 (0.44)	47.06 (1.29)
Erkennungsleistung auf Buchstabenebene					
$I1$ (2360)	29-40-30	96.99 (0.80)	97.13 (0.54)	73.14 (0.87)	69.15 (0.84)
		76.75 (1.00)	86.20 (0.80)	50.08 (1.03)	54.32 (1.07)
$I1(t)$ (1835)	29-40-30	96.22 (0.77)	96.43 (0.53)	74.55 (0.93)	71.53 (0.96)
		74.49 (1.10)	84.72 (0.87)	50.35 (1.43)	55.75 (1.06)
$H1$ (1835)	20-25 \ 30 9-15 /	96.46 (0.70)	96.82 (0.54)	73.98 (0.88)	70.67 (1.06)
		74.98 (1.20)	85.12 (0.98)	50.30 (1.01)	55.01 (1.00)
$I2$ (5360)	29-40-60-30	98.40 (0.82)	98.54 (0.41)	68.01 (1.01)	63.13 (0.75)
		82.96 (1.30)	91.08 (0.66)	46.88 (1.39)	49.24 (1.19)
$I2(t)$ (3635)	29-40-60-30	97.63 (1.00)	97.88 (0.57)	70.82 (1.45)	66.68 (0.98)
		80.12 (1.08)	89.40 (0.74)	48.57 (1.29)	52.91 (0.91)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97.68 (1.07)	98.25 (0.53)	73.58 (1.33)	70.54 (0.71)
		80.04 (1.50)	89.41 (0.81)	52.39 (1.82)	56.99 (1.05)
C (1800)	20-25-30 \ 30 9-15-30 /	93.41 (1.56)	94.57 (0.81)	66.37 (1.21)	62.17 (0.98)
		72.53 (1.14)	83.69 (0.90)	43.82 (0.99)	47.31 (1.24)

Tabelle C.4: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitekturen sind unterschiedlich. Trainiert und getestet wird mit allen Rauschstufen gemeinsam.

Netz (Anzahl Gewichte)	Architektur	ungestort	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Trainingsdaten					
E (2360)	29-40-30	91.58 (2.13)	93.68 (1.83)	94.79 (1.53)	96.98 (1.00)
		69.50 (1.86)	72.63 (1.69)	74.84 (1.54)	81.06 (1.12)
E (1835)	29-40-30	91.71 (2.23)	93.48 (1.82)	94.65 (1.41)	96.71 (0.75)
		65.39 (3.20)	68.29 (2.86)	71.23 (2.34)	78.50 (1.49)
E (5360)	29-40-60-30	99.06 (0.32)	99.14 (0.27)	99.17 (0.26)	99.50 (0.14)
		93.73 (2.46)	94.04 (2.32)	94.46 (2.19)	95.84 (1.74)
E (3635)	29-40-60-30	98.61 (0.85)	98.94 (0.70)	99.02 (0.60)	99.37 (0.30)
		81.82 (2.53)	83.27 (2.33)	84.66 (2.23)	89.08 (1.77)
Erkennungsleistung auf Trainingsdaten, KSPM					
E (2360)	29-40-30	94.46 (1.23)	95.70 (1.03)	96.40 (0.85)	97.79 (0.58)
		83.55 (1.26)	85.44 (1.11)	86.81 (0.97)	90.35 (0.73)
E (1835)	29-40-30	93.82 (1.54)	95.00 (1.27)	95.76 (1.02)	97.25 (0.58)
		80.17 (2.17)	82.06 (1.97)	83.99 (1.60)	88.38 (1.07)
E (5360)	29-40-60-30	99.16 (0.33)	99.24 (0.29)	99.26 (0.27)	99.56 (0.15)
		96.23 (1.09)	96.45 (0.97)	96.69 (0.88)	97.59 (0.68)
E (3635)	29-40-60-30	99.01 (0.49)	99.19 (0.43)	99.26 (0.38)	99.52 (0.17)
		90.50 (1.28)	91.37 (1.12)	92.21 (1.02)	94.71 (0.74)
Erkennungsleistung auf Testdaten					
E (2360)	29-40-30	80.92 (1.95)	76.00 (1.46)	72.45 (1.19)	60.88 (1.15)
		57.34 (2.19)	54.02 (1.53)	51.04 (1.52)	43.01 (1.33)
E (1835)	29-40-30	81.80 (1.88)	76.54 (1.22)	71.78 (0.95)	60.24 (1.14)
		54.61 (2.83)	50.87 (1.81)	48.21 (1.66)	41.48 (1.37)
E (5360)	29-40-60-30	86.73 (0.78)	79.57 (1.14)	74.80 (1.20)	61.91 (1.15)
		77.91 (2.43)	69.50 (2.48)	64.06 (2.26)	51.35 (1.49)
E (3635)	29-40-60-30	85.75 (0.89)	78.71 (1.11)	72.89 (0.99)	61.66 (1.11)
		67.20 (2.64)	60.19 (2.69)	55.47 (2.49)	46.44 (1.89)
Erkennungsleistung auf Testdaten, KSPM					
E (2360)	29-40-30	79.91 (1.90)	72.81 (1.59)	69.28 (1.48)	57.57 (1.18)
		65.93 (1.83)	59.73 (1.55)	56.27 (1.46)	46.11 (1.68)
E (1835)	29-40-30	80.42 (1.79)	72.70 (1.07)	67.37 (1.61)	56.34 (1.91)
		63.92 (2.44)	57.36 (1.55)	53.64 (1.75)	44.97 (1.80)
E (5360)	29-40-60-30	83.76 (1.19)	76.43 (1.16)	71.13 (1.32)	58.60 (1.13)
		76.51 (1.55)	68.10 (1.63)	62.25 (1.81)	50.62 (1.14)
E (3635)	29-40-60-30	83.29 (1.41)	74.88 (1.32)	69.08 (1.14)	57.70 (1.20)
		71.65 (1.90)	63.37 (1.45)	57.91 (0.99)	47.65 (1.45)

Tabelle C.5: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitektur ist E. Die mit allen Rauschstufen trainierten Netze (vgl. Tabelle C.1 werden für die einzelnen Signal-zu-Rausch-Verhältnisse getrennt getestet. Dies geschieht mit den Trainings- und Testdaten.

Netz (Anzahl Gewichte)	Architektur	ungestört	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Trainingsdaten					
E (2360)	29-40-30	97.42 (1.76)	97.72 (1.70)	97.82 (1.61)	98.07 (1.42)
		90.06 (2.29)	90.60 (2.19)	90.83 (2.10)	91.92 (1.90)
E (1835)	29-40-30	97.32 (1.68)	97.80 (1.50)	98.00 (1.43)	98.22 (1.32)
		87.25 (2.70)	88.09 (2.57)	88.68 (2.48)	90.02 (2.17)
E (5360)	29-40-60-30	99.65 (0.03)	99.67 (0.04)	99.65 (0.03)	99.70 (0.04)
		99.47 (0.08)	99.47 (0.08)	99.46 (0.07)	99.59 (0.07)
E (3635)	29-40-60-30	99.69 (0.06)	99.70 (0.06)	99.72 (0.06)	99.74 (0.06)
		99.43 (0.11)	99.45 (0.10)	99.45 (0.09)	99.51 (0.08)
Erkennungsleistung auf Trainingsdaten, KSPM					
E (2360)	29-40-30	98.66 (0.71)	98.84 (0.69)	98.89 (0.64)	99.02 (0.56)
		95.23 (0.96)	95.53 (0.91)	95.64 (0.87)	96.20 (0.79)
E (1835)	29-40-30	98.42 (0.76)	98.69 (0.63)	98.83 (0.60)	98.98 (0.55)
		93.86 (1.16)	94.29 (1.08)	94.63 (1.04)	95.28 (0.93)
E (5360)	29-40-60-30	99.43 (0.06)	99.43 (0.07)	99.41 (0.06)	99.49 (0.08)
		99.32 (0.07)	99.31 (0.08)	99.28 (0.07)	99.41 (0.08)
E (3635)	29-40-60-30	99.54 (0.10)	99.55 (0.10)	99.56 (0.10)	99.59 (0.11)
		99.29 (0.08)	99.30 (0.08)	99.30 (0.07)	99.35 (0.07)
Erkennungsleistung auf Testdaten					
E (2360)	29-40-30	77.40 (1.93)	71.62 (1.49)	67.49 (1.77)	57.55 (1.56)
		63.95 (2.29)	58.01 (2.15)	54.07 (2.09)	44.96 (1.52)
E (1835)	29-40-30	77.02 (2.03)	71.57 (1.68)	67.79 (1.47)	58.09 (1.26)
		61.35 (2.66)	55.89 (2.54)	52.65 (2.17)	43.95 (1.66)
e (5360)	29-40-60-30	79.08 (0.83)	73.85 (1.01)	69.71 (1.31)	58.93 (1.15)
		72.05 (0.87)	65.78 (1.00)	61.48 (1.03)	50.01 (0.97)
E (3635)	29-40-60-30	78.77 (0.92)	73.35 (1.17)	69.77 (1.08)	59.33 (1.15)
		71.94 (1.02)	65.39 (1.09)	61.24 (1.05)	50.25 (0.97)
Erkennungsleistung auf Testdaten, KSPM					
E (2360)	29-40-30	73.69 (0.95)	68.03 (1.20)	63.91 (1.02)	54.99 (0.84)
		63.22 (1.49)	58.17 (1.64)	54.07 (1.40)	45.99 (1.35)
E (1835)	29-40-30	73.26 (1.16)	67.65 (1.05)	64.01 (1.13)	55.76 (1.24)
		62.31 (1.61)	56.92 (1.83)	53.58 (1.53)	45.33 (1.61)
E (5360)	29-40-60-30	73.61 (1.72)	68.70 (1.30)	63.97 (1.61)	54.34 (0.92)
		66.25 (1.61)	61.35 (1.41)	56.86 (1.56)	47.00 (1.17)
E (3635)	29-40-60-30	74.44 (1.64)	69.01 (1.71)	65.39 (1.57)	55.19 (1.12)
		67.40 (1.60)	61.65 (1.58)	57.76 (1.67)	47.77 (1.20)

Tabelle C.6: Klassifkationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitektur ist E. Die mit allen Rauschstufen trainierten Netze (vgl. Tabelle C.2) werden für die einzelnen Signal-zu-Rausch-Verhältnisse getrennt getestet. Dies geschieht mit den Trainings- und Testdaten.

Netz (Anzahl Gewichte)	Architektur	ungestört	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Trainingsdaten					
A (1250)	20-25-30	74.51 (2.02)	82.43 (1.77)	86.81 (1.42)	95.04 (0.60)
		54.77 (1.72)	67.40 (1.22)	74.26 (0.97)	88.56 (0.77)
$I1$ (2360)	29-40-30	93.59 (1.75)	95.81 (1.22)	96.74 (1.05)	98.28 (0.58)
		70.48 (1.79)	74.35 (1.39)	77.35 (1.24)	85.13 (0.70)
$I1(t)$ (1835)	29-40-30	90.70 (1.64)	94.05 (1.29)	95.12 (0.20)	96.98 (0.72)
		65.82(2.44)	70.66 (1.92)	73.50 (1.56)	81.74 (0.79)
$H1$ (1835)	20-25 \ 30 9-15 /	91.26 (2.17)	93.51 (1.83)	94.52 (1.59)	96.65 (1.07)
		66.10 (2.39)	70.29 (1.97)	73.57 (1.51)	81.69 (0.83)
$I2$ (5360)	29-40-60-30	97.51 (1.17)	98.53 (0.72)	99.10 (0.47)	99.66 (0.15)
		80.26 (2.51)	82.42 (2.18)	84.31 (2.07)	90.32 (1.33)
$I2(t)$ (3635)	29-40-60-30	97.73 (0.62)	98.68 (0.29)	99.05 (0.18)	99.50 (0.13)
		77.68 (2.01)	79.81 (1.74)	81.48 (1.52)	87.62 (1.09)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	94.67 (2.00)	95.71 (1.76)	96.29 (1.57)	97.49 (1.13)
		76.27 (2.09)	78.57 (2.15)	80.58 (2.01)	86.57 (1.64)
C (1800)	20-25-30 \ 30 9-15-30 /	84.20 (2.36)	89.21 (1.72)	91.15 (1.49)	94.51 (0.92)
		59.75 (1.60)	66.57 (1.12)	69.84 (1.06)	78.63 (0.86)
Erkennungsleistung auf Trainingsdaten, KSPM					
A (1250)	20-25-30	84.18 (0.97)	89.63 (0.80)	92.18 (0.65)	96.98 (0.34)
		69.45 (1.27)	79.72 (0.81)	84.47 (0.59)	93.32 (0.55)
$I1$ (2360)	29-40-30	94.85 (1.04)	96.72 (0.71)	97.43 (0.59)	98.61 (0.32)
		82.12 (1.44)	85.22 (1.22)	87.37 (1.02)	92.32 (0.43)
$I1(t)$ (1835)	29-40-30	92.15 (1.21)	94.82 (0.92)	95.75 (0.85)	97.39 (0.46)
		78.21 (1.85)	81.81 (1.53)	83.93 (1.31)	89.56 (0.70)
$H1$ (1835)	20-25 \ 30 9-15 /	93.14 (1.11)	94.88 (0.97)	95.64 (0.82)	97.37 (0.54)
		78.28 (1.75)	81.43 (1.49)	83.72 (1.23)	89.17 (0.75)
$I2$ (5360)	29-40-60-30	98.01 (0.76)	98.77 (0.42)	99.22 (0.26)	99.68 (0.13)
		89.87 (1.27)	91.26 (0.97)	92.48 (0.89)	95.53 (0.58)
$I2(t)$ (3635)	29-40-60-30	97.84 (0.53)	98.68 (0.29)	99.00 (0.20)	99.51 (0.13)
		88.26 (1.13	89.87 (0.87)	90.87 (0.74)	94.10 (0.48)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	96.31 (1.10)	97.12 (0.85)	97.59 (0.68)	98.43 (0.54)
		86.31 (1.36)	87.87 (1.34)	89.23 (1.19)	92.78 (0.97)
C (1800)	20-25-30 \ 30 9-15-30 /	88.12 (1.13)	91.81 (0.74)	93.05 (0.66)	95.43 (0.47)
		73.34 (1.10)	78.36 (0.85)	80.41 (0.84)	85.98 (0.65)

Netz (Anzahl Gewichte)	Architektur	ungestört	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Testdaten					
A (1250)	20-25-30	60.54 (2.40)	51.41 (2.33)	45.83 (2.26)	31.41 (1.46)
		39.57 (1.59)	32.89 (1.51)	28.67 (1.30)	20.41 (1.13)
$I1$ (2360)	29-40-30	85.64 (1.72)	79.14 (1.33)	73.16 (1.07)	58.28 (1.70)
		60.43 (1.52)	54.92 (1.24)	50.76 (1.07)	39.98 (1.51)
$I1(t)$ (1835)	29-40-30	84.13 (2.01)	79.28 (1.57)	74.25 (1.45)	58.63 (1.74)
		57.94 (2.43)	53.94 (1.91)	49.85 (1.35)	39.76 (1.38)
$H1$ (1835)	20-25 \ 30 9-15 /	82.39 (2.42)	78.31 (1.87)	73.74 (1.70)	60.16 (1.56)
		56.83 (1.73)	53.80 (1.83)	50.61 (1.41)	42.18 (1.30)
$I2$ (5360)	29-40-60-30	87.84 (1.61)	79.86 (1.20)	74.06 (1.34)	55.30 (1.27)
		68.29 (2.47)	59.31 (2.52)	53.85 (1.93)	38.10 (1.24)
$I2(t)$ (3635)	29-40-60-30	87.62 (1.13)	81.02 (1.30)	74.93 (1.47)	57.29 (1.18)
		65.08 (1.43)	57.64 (1.41)	52.14 (1.64)	40.10 (1.32)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	84.97 (1.82)	78.69 (1.96)	73.51 (1.57)	60.43 (1.82)
		64.48 (2.11)	58.09 (1.63)	54.23 (1.86)	45.08 (1.54)
C (1800)	20-25-30 \ 30 9-15-30 /	73.35 (2.20)	69.71 (1.59)	65.85 (1.41)	55.81 (1.73)
		48.00 (1.52)	46.02 (1.08)	43.41 (1.22)	37.61 (1.42)
Erkennungsleistung auf Testdaten, KSPM					
A (1250)	20-25-30	63.41 (1.83)	46.55 (1.38)	38.48 (1.86)	24.63 (1.15)
		47.35 (1.15)	32.32 (1.34)	26.33 (1.66)	16.95 (0.99)
$I1$ (2360)	29-40-30	84.26 (1.44)	75.37 (1.31)	68.79 (1.18)	53.97 (1.42)
		68.50 (1.86)	59.48 (1.45)	54.54 (1.53)	41.85 (1.48)
$I1(t)$ (1835)	29-40-30	83.50 (1.51)	76.71 (1.25)	70.45 (1.05)	55.30 (1.22)
		67.07 (2.08)	60.22 (1.40)	54.28 (1.27)	42.52 (1.15)
$H1$ (1835)	20-25 \ 30 9-15 /	81.77 (1.37)	75.84 (1.26)	70.26 (1.54)	57.39 (1.19)
		65.48 (1.20)	59.99 (1.66)	55.13 (1.49)	45.59 (1.29)
$I2$ (5360)	29-40-60-30	85.70 (1.50)	76.55 (1.24)	69.09 (1.28)	51.11 (1.22)
		74.25 (1.72)	63.63 (1.38)	56.00 (1.55)	39.23 (1.30)
$I2(t)$ (3635)	29-40-60-30	85.26 (1.15)	77.10 (1.35)	70.33 (1.38)	53.69 (1.07)
		72.29 (1.32)	62.05 (1.44)	55.86 (1.40)	41.76 (1.15)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	83.78 (1.40)	76.91 (1.43)	70.50 (1.68)	58.76 (1.31)
		71.20 (1.86)	63.72 (1.69)	58.06 (1.81)	47.83 (1.52)
C (1800)	20-25-30 \ 30 9-15-30 /	72.69 (1.48)	65.90 (1.48)	62.39 (1.63)	52.99 (2.15)
		55.39 (1.93)	49.54 (1.52)	46.54 (2.01)	39.70 (1.91)

Tabelle C.7: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet. Die Fusionsarchitekturen sind unterschiedlich. Die mit allen Rauschstufen trainierten Netze (vgl. Tabelle C.3) werden für die einzelnen Signal-zu-Rausch-Verhältnisse getrennt getestet. Dies geschieht mit den Trainings- und Testdaten.

Netz (Anzahl Gewichte)	Architektur	ungestört	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Trainingsdaten					
$I1$ (2360)	29-40-30	95.35 (1.32)	96.81 (0.90)	97.29 (0.77)	98.51 (0.37)
		71.29 (1.69)	74.16 (1.21)	76.62 (0.95)	84.92 (0.72)
$I1(t)$ (1835)	29-40-30	94.32 (1.23)	96.14 (0.96)	96.56 (0.77)	97.88 (0.30)
		68.80 (1.76)	72.25 (1.41)	74.78 (1.02)	82.13 (0.51)
$H1$ (1835)	20-25 \ 30 9-15 /	94.28 (1.31)	96.36 (0.85)	97.04 (0.58)	98.16 (0.30)
		68.95 (1.79)	72.78 (1.47)	75.57 (1.09)	82.60 (0.77)
$I2$ (5360)	29-40-60-30	97.63 (1.16)	98.13 (1.02)	98.49 (0.88)	99.36 (0.33)
		78.20 (1.62)	80.29 (1.64)	82.41 (1.44)	90.94 (0.65)
$I2(t)$ (3635)	29-40-60-30	96.78 (1.43)	97.18 (1.20)	97.62 (1.00)	98.94 (0.44)
		75.45 (1.59)	77.28 (1.28)	79.64 (0.98)	88.12 (0.69)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	96.70 (1.28)	97.64 (1.10)	97.90 (1.09)	98.48 (0.95)
		75.76 (1.92)	78.39 (1.62)	79.99 (1.48)	86.01 (1.24)
C (1800)	20-25-30 \ 30 9-15-30 /	89.80 (2.35)	93.39 (1.80)	94.43 (1.42)	96.01 (0.82)
		65.29 (1.72)	71.10 (1.37)	73.80 (1.13)	79.95 (0.73)
Erkennungsleistung auf Trainingsdaten, KSPM					
$I1$ (2360)	29-40-30	95.66 (0.93)	96.87 (0.62)	97.40 (0.53)	98.58 (0.27)
		82.21 (1.32)	84.39 (0.91)	86.40 (0.77)	91.78 (0.52)
$I1(t)$ (1835)	29-40-30	94.89 (0.82)	96.28 (0.66)	96.66 (0.59)	97.88 (0.27)
		80.71 (1.27)	83.20 (1.14)	85.05 (0.89)	89.92 (0.47)
$H1$ (1835)	20-25 \ 30 9-15 /	95.17 (0.95)	96.65 (0.61)	97.19 (0.48)	98.28 (0.28)
		81.04 (1.38)	83.72 (1.20)	85.62 (0.91)	90.11 (0.65)
$I2$ (5360)	29-40-60-30	97.84 (0.65)	98.29 (0.51)	98.62 (0.42)	99.41 (0.18)
		88.20 (0.82)	89.63 (0.88)	90.97 (0.77)	95.53 (0.32)
$I2(t)$ (3635)	29-40-60-30	97.19 (0.82)	97.44 (0.67)	97.84 (0.58)	99.07 (0.27)
		86.69 (1.10)	87.75 (0.92)	89.22 (0.69)	93.93 (0.41)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	97.53 (0.69)	98.21 (0.55)	98.41 (0.53)	98.84 (0.46)
		86.82 (1.05)	88.56 (0.90)	89.47 (0.82)	92.77 (0.68)
C (1800)	20-25-30 \ 30 9-15-30 /	91.93 (1.20)	94.57 (0.95)	95.25 (0.79)	96.53 (0.50)
		78.93 (1.19)	83.03 (1.42)	84.59 (0.95)	88.23 (0.69)

Netz (Anzahl Gewichte)	Architektur	ungestört	SNR 3 dB	SNR -3 dB	SNR -11 dB
Erkennungsleistung auf Testdaten					
$I1$ (2360)	29-40-30	85.47 (1.56)	79.08 (0.91)	72.61 (1.13)	55.43 (1.17)
		60.14 (1.80)	54.17 (1.32)	49.07 (1.42)	36.94 (1.16)
$I1(t)$ (1835)	29-40-30	85.02 (1.60)	79.70 (1.26)	74.47 (1.05)	59.00 (0.95)
		58.55 (2.18)	53.53 (1.92)	50.01 (1.45)	39.33 (1.19)
$H1$ (1835)	20-25 \ 30 9-15 /	84.21 (1.37)	78.65 (1.46)	74.14 (1.11)	58.93 (1.02)
		57.51 (1.71)	53.59 (1.23)	49.88 (1.15)	40.22 (1.21)
$I2$ (5360)	29-40-60-30	86.14 (1.04)	74.52 (1.37)	66.17 (1.48)	45.21 (1.16)
		63.54 (1.65)	51.43 (2.15)	43.55 (1.73)	29.02 (1.09)
$I2(t)$ (3635)	29-40-60-30	85.74 (1.59)	76.49 (1.78)	70.01 (1.89)	51.04 (1.34)
		61.64 (1.73)	51.95 (1.79)	46.74 (1.63)	33.96 (1.34)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	85.29 (1.45)	77.47 (1.34)	72.97 (1.71)	58.58 (1.67)
		61.02 (2.24)	54.61 (1.89)	51.65 (2.02)	42.26 (1.98)
C (1800)	20-25-30 \ 30 9-15-30 /	74.23 (2.14)	69.72 (1.44)	66.61 (1.33)	54.93 (1.20)
		48.05 (1.55)	46.17 (1.42)	43.87 (1.15)	37.20 (1.22)
Erkennungsleistung auf Testdaten, KSPM					
$I1$ (2360)	29-40-30	83.45 (1.64)	74.69 (1.03)	67.37 (1.03)	51.10 (1.45)
		68.24 (1.87)	58.07 (1.24)	52.28 (1.31)	38.71 (1.61)
$I1(t)$ (1835)	29-40-30	83.65 (1.53)	76.86 (1.28)	70.94 (1.58)	54.64 (0.90)
		67.06 (1.58)	59.57 (1.23)	54.58 (1.75)	41.80 (1.19)
$H1$ (1835)	20-25 \ 30 9-15 /	82.58 (1.37)	75.18 (1.73)	69.58 (1.59)	55.34 (1.33)
		65.13 (1.68)	58.84 (1.06)	53.73 (1.53)	42.33 (1.26)
$I2$ (5360)	29-40-60-30	83.35 (1.18)	69.03 (1.15)	59.89 (1.21)	40.24 (1.32)
		68.88 (1.62)	54.34 (1.78)	44.90 (1.40)	28.84 (1.39)
$I2(t)$ (3635)	29-40-60-30	83.56 (1.50)	71.63 (1.45)	64.92 (1.52)	46.60 (1.42)
		69.34 (1.47)	56.42 (1.52)	50.45 (1.40)	35.43 (1.37)
$H2$ (3635)	20-25-30 \ 30 9-15-30 /	83.00 (0.98)	73.88 (1.14)	69.23 (1.20)	56.04 (1.53)
		68.00 (1.22)	59.36 (1.22)	55.46 (1.53)	45.13 (1.69)
C (1800)	20-25-30 \ 30 9-15-30 /	70.84 (1.79)	64.51 (1.37)	61.43 (1.38)	51.90 (1.22)
		53.95 (1.98)	49.06 (1.35)	46.10 (1.41)	40.14 (1.48)

Tabelle C.8: Klassifikationsleistungen für kombinierte akustische und visuelle Daten. Als visuelles Merkmal wird das ganze Zeitfenster verwendet. Die Fusionsarchitekturen sind unterschiedlich. Die mit allen Rauschstufen trainierten Netze (vgl. Tabelle C.4) werden für die einzelnen Signal-zu-Rausch-Verhältnisse getrennt getestet. Dies geschieht mit den Trainings- und Testdaten.

Trainingsdaten	KSPM (trai)	Testdaten	KSPM (test)	SNR
Schwelle -100				
46.33 (1.84)	59.72 (1.34)	32.18 (1.86)	40.20 (1.67)	ungestört
46.65 (1.88)	59.93 (1.43)	31.68 (1.77)	39.31 (1.51)	3 dB
46.99 (1.94)	60.11 (1.52)	31.81 (1.75)	39.35 (1.53)	-3 dB
47.61 (2.20)	60.46 (1.74)	31.16 (1.60)	38.71 (1.48)	-11 dB
Schwelle 200				
46.24 (1.84)	59.61 (1.34)	32.10 (1.81)	40.10 (1.64)	ungestört
46.59 (1.89)	59.86 (1.42)	31.54 (1.71)	39.20 (1.48)	3 dB
46.92 (1.94)	60.04 (1.51)	31.71 (1.71)	39.28 (1.54)	-3 dB
47.57 (2.20)	60.40 (1.74)	31.11 (1.60)	38.63 (1.49)	-11 dB
Schwelle 800				
61.52 (1.99)	73.85 (1.23)	50.31 (2.00)	55.87 (1.84)	ungestört
67.21 (1.66)	78.16 (1.02)	45.34 (1.91)	46.12 (1.61)	3 dB
70.69 (1.62)	80.26 (0.97)	43.17 (1.96)	42.09 (1.60)	-3 dB
78.44 (1.56)	84.84 (2.40)	33.07 (1.37)	32.13 (1.25)	-11 dB
Schwelle 1400				
70.46 (2.11)	82.15 (1.05)	58.28 (2.41)	63.61 (1.61)	ungestört
76.98 (1.77)	86.91 (0.83)	48.94 (2.30)	47.42 (1.37)	3 dB
80.97 (1.48)	89.27 (0.72)	44.86 (2.37)	41.26 (1.90)	-3 dB
89.02 (0.76)	93.98 (0.40)	32.29 (1.46)	29.54 (1.51)	-11 dB
akustische Entscheidung hat Vorrang				
74.54 (2.01)	84.19 (0.97)	60.56 (2.40)	63.42 (1.82)	ungestört
82.45 (1.77)	89.64 (0.80)	51.43 (2.32)	46.56 (1.37)	3 dB
86.82 (1.42)	92.19 (0.65)	45.85 (2.25)	38.49 (1.86)	-3 dB
95.04 (0.60)	96.98 (0.34)	31.41 (1.46)	24.63 (1.15)	-11 dB
Schwelle als logistische Funktion				
66.81 (4.29)	78.44 (3.79)	58.28 (2.41)	63.61 (1.61)	ungestört
65.29 (6.40)	76.87 (5.46)	48.94 (2.30)	47.42 (1.37)	3 dB
64.06 (7.58)	75.07 (6.31)	44.86 (2.37)	41.26 (1.90)	-3 dB
67.88 (9.09)	76.89 (7.25)	32.29 (1.46)	29.54 (1.51)	-11 dB

Tabelle C.9: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, Verwendung der Architektur L, verschiedene Fusionsstrategien. Als akustisches Netz wird die Architektur 20-25-30, als visuelles Netz die Architektur 9-15-30 verwendet. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand verwendet. Die Klassifikationsergebnisse beziehen sich auf die Buchstabenwortebene. Als visuelles Merkmal wird eine Zeitscheibe des Zeitfensters verwendet.

Trainingsdaten	KSPM (trai)	Testdaten	KSPM (test)	SNR
Schwelle -100				
60.36 (1.69)	73.79 (1.14)	35.27 (1.61)	42.18 (1.80)	ungestört
60.50 (1.73)	73.89 (1.16)	35.25 (1.62)	42.09 (1.86)	3 dB
60.62 (1.78)	73.96 (1.18)	35.17 (1.61)	41.96 (1.82)	-3 dB
60.78 (1.84)	74.04 (1.21)	35.01 (1.62)	41.88 (1.76)	-11 dB
Schwelle 200				
59.94 (1.75)	73.38 (1.15)	35.06 (1.64)	42.02 (1.79)	ungestört
60.17 (1.75)	73.61 (1.12)	34.87 (1.65)	41.70 (1.88)	3 dB
60.31 (1.80)	73.68 (1.15)	34.80 (1.65)	41.58 (1.82)	-3 dB
60.62 (1.87)	73.89 (1.22)	34.61 (1.64)	41.31 (1.66)	-11 dB
Schwelle 800				
65.86 (1.51)	77.94 (0.78)	50.95 (1.91)	57.58 (1.80)	ungestört
72.27 (1.40)	82.63 (0.77)	46.07 (1.95)	47.41 (1.63)	3 dB
75.93 (1.39)	84.81 (0.83)	43.52 (1.87)	42.97 (1.97)	-3 dB
83.90 (1.39)	89.48 (0.86)	34.33 (1.49)	33.48 (1.39)	-11 dB
Schwelle 1400				
71.40 (2.07)	82.63 (1.02)	58.28 (2.54)	63.42 (1.66)	ungestört
78.29 (1.73)	87.57 (0.82)	49.40 (2.39)	47.91 (1.42)	3 dB
82.26 (1.45)	89.93 (0.72)	45.08 (2.33)	41.57 (1.82)	-3 dB
90.38 (0.78)	94.67 (0.43)	33.07 (1.54)	30.31 (1.50)	-11 dB
akustische Entscheidung hat Vorrang				
74.57 (2.01)	84.20 (0.96)	60.59 (2.39)	63.44 (1.81)	ungestört
82.47 (1.77)	89.65 (0.79)	51.44 (2.32)	46.57 (1.38)	3 dB
86.81 (1.42)	92.18 (0.65)	45.85 (2.25)	38.49 (1.86)	-3 dB
95.04 (0.60)	96.98 (0.34)	31.41 (1.46)	24.63 (1.15)	-11 dB
Schwelle als logistische Funktion				
67.83 (2.77)	80.00 (1.90)	58.28 (2.54)	63.42 (1.66)	ungestort
70.13 (3.81)	81.50 (2.95)	49.40 (2.39)	47.91 (1.42)	3 dB
71.27 (4.83)	81.77 (3.59)	45.08 (2.33)	41.57 (1.82)	-3 dB
74.14 (6.37)	83.30 (4.45)	33.07 (1.54)	30.31 (1.50)	-11 dB

Tabelle C.10: Klassifikationsleistungen für die gemeinsame Verarbeitung von akustischen und visuellen Daten, Verwendung der Architektur L, verschiedene Fusionsstrategien. Als akustisches Netz wird die Architektur 20-25-30, als visuelles Netz die Architektur 9-15-30 verwendet. Als Entscheidungskriterium des Assoziativspeichers wird der Hammingabstand verwendet. Die Klassifikationsergebnisse beziehen sich auf die Buchstabenwortebene. Als visuelles Merkmal wird das ganze Zeitfenster verwendet.

Literaturverzeichnis

[AB95a] A. Adjoudani and C. Benoit. Audio-visual speech recognition compared across two architectures. In *Proc. of the European Conf. on Speech Communication and Technology*, pages 1563–1566, Madrid, Spain, September 1995. ESCA.

[AB95b] A. Adjoudani and C. Benoit. On the integration of auditory and visual parameters in an HMM-based ASR. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 461–471. Springer Verlag, 1995.

[Ade87] G. Adelman, editor. *Encyclopedia of Neuroscience.* Birkhäuser Verlag, Boston, Basel, Stuttgart, 1987.

[ADR96] M. Alissali, P. Deleglise, and A. Rogozan. Asynchronous integration of visual information in an automatic speech recognition system. In *Proc.of the Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.

[Ber72] K. W. Berger. Three experiments in speechreading. *Journal of Communication Disorders*, 5:280–285, 1972.

[BGMGA95] C. Benoit, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani. Which components of the face do humans and machines best speechread ? In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 315–325. Springer Verlag, September 1995.

[BHMW93] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, pages 557–560, April 1993.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[BK94] C. Bregler and Y. Konig. 'Eigenlips' for robust speech recognition. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, 1994.

[BMHW93] C. Bregler, S. Manke, H. Hild, and A. Waibel. Bimodal sensor integration on the example of 'speech-reading'. In *Proc. of the Int. Conf. on Neural Networks*, 1993.

[BMJ74] C. A. Binnie, A. A. Montgomery, and P. L. Jackson. Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research*, 17:619–630, 1974.

[BMK94] C. Benoit, T. Mohamadi, and S. Kandel. Effects of phonetic context on audio-visual intelligibility of french. *Journal of Speech and Hearing Reseach*, 37:1195–1203, October 1994.

[BMS99] J. F. Baldwin, T. P. Martin, and M. Saeed. Automatic computer lip-reading using fuzzy set theory. In D. Massaro, editor, *Proc. of the AVSP'99*, pages 86–91, August 1999.

[BO94] C. Bregler and S. M. Omohundro. Surface learning with applications to lipreading. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*. Morgan Kaufmann, 1994.

[BO95] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In W. Eric and L. Grimson, editors, *Proc. of the Fifth Int. Conf. on Computer Vision*, pages 494–499, Los Alamitos, CA, 1995. IEEE Computer Society Press.

[BOK94] C. Bregler, S. M. Omohundro, and Y. Konig. A hybrid approach to bimodal speech recognition. In *28th Annual Asilomar Conf. on Signals, Systems and Computers*, pages 556–560. IEEE Computer Society Press, November 1994.

[Bor93] J. Bortz. *Statistik*. Springer Verlag, 1993.

[BOSK95] C. Bregler, S. M. Omohundro, J. Shi, and Y. Konig. Towards a robust speechreading dialog system. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 409–423. Springer Verlag, 1995.

[BP86] M. Breeuwer and R. Plomp. Speechreading supplemented with auditorily presented speech parameters. *J. Acoust. Soc. Amer.*, 79(2):481–499, February 1986.

[BPF82] A.-P. Benguerel and M. K. Pichora-Fuller. Coarticulation effects in lipreading. *Journal of Speech and Hearing Research*, 25:600–607, December 1982.

[Bra91] L. D. Braida. Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology*, 43A(3):647–677, 1991.

[Bur98] D. Burnham. Language specifity in the development of auditory-visual speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by Eye II*, chapter 2, pages 27–60. Psychology Press, 1998.

[CC96] P. Cosi and E. Magno Caldognetto. Lips and jaw movements for vowels and consonants: Spatio-temporal characteristics and bimodal recognition applications. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 291–313. Springer-Verlag, 1996.

[CD80] R. Campbell and B. Dodd. Hearing by eye. *Quarterly Journal of Experimental Psychology*, 32:85–99, 1980.

[CM95] M. M. Cohen and D. W. Massaro. Perceiving visual and auditory information in consonant-vowel and vowel syllables. In C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen, editors, *Levels in Speech Communication: Relations and Interactions*. Elsevier Science B. V., Amsterdam, 1995.

[CMB97] S. Cox, I. Matthews, and A. Bangham. Combining noise compensation with visual information in speech recognition. In C. Benoit and R. Campbell, editors, *Proceedings of the Audio-Visual Speech Processing 97*, pages 53–56. ESCA, September 1997.

[Coh95] P. R. Cohen. *Empirical Methods for Artificial Intelligence*. MIT Press, 1995.

[Cot35] J. C. Cotton. Normal 'visual hearing'. *Science*, 82(2138):592–593, December 1935.

[Cow90] J. D. Cowan. Neural networks: the early days. *Neural Information Processing Systems 2*, pages 828 – 842, 1990.

[Cry95] D. Crystal. *Die Cambridge Enzyklopädie der Sprache*. Camous Verlag, 1995.

[dGBVC95] B. de Gelder, P. Bertelson, J. Vroomen, and H. C. Chen. Interlanguage differences in the McGurk effect for dutch and cantonese listeners. In *Proc. of the 4th European Conf. on Speech Communication and Technology*, pages 1699–1702. ESCA, September 1995.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience Publication, 1973.

[DHB+95] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel. Toward movement-invariant automatic lip-reading and speech recognition. In *Proc. of the Int. Conf. on Acoustic, Speech and Signal Processing*, 1995.

[Die98] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[DKB95] B. Dalton, R. Kauric, and A. Blake. Automatic speechreading using dynamic contours. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 373–382. Springer Verlag, 1995.

[DMW94] P. Duchnowski, U. Meier, and A. Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. In *Int. Conf. on Spoken Language Processing*, pages 547–550, Yokohama, Japan, September 1994.

[DS80] N. F. Dixon and L. Spitz. The detection of auditory visual desynchrony. *Perception*, 9:719–721, 1980.

[Dur95] A. Fuster Duran. McGurk effect in spanish and german listeners: influences of visual cues in the perception of spanish and german conflicts audio-visual stimuli. In *Proc. of the 4th European Conf. on Speech Communication and Technology*, pages 295–298. ESCA, September 1995.

[EH93] B. Eppinger and E. Herter. *Sprachverarbeitung.* Carl Hanser Verlag, München, Wien, 1993.

[EM88] J. L. Elman and J. L. McClelland. Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, 27:143–165, 1988.

[EN71] H. W. Ewertsen and H. Birk Nielsen. A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta Otolaryng*, 72:201–205, 1971.

[Erb69] N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12:423–425, 1969.

[Erb75] N. P. Erber. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40:481–492, 1975.

[Far59] T. Farrimond. Age differences in the ability to use visual cues in auditory communications. *Language and Speech*, 2:179–192, 1959.

[FD91] C. A. Fowler and D. J. Dekle. Listening with eye and hand: Crossmodal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3):816–828, 1991.

[FD96] A. Fuster-Duran. Perception of conflicting audio-visual speech: An examination across spanish and german. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 135–143. Springer Verlag, 1996.

[Fel84] K. Fellbaum. *Sprachverarbeitung und Sprachübertragung.* Springer Verlag, 1984.

[Fis68] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11:796–804, 1968.

[FK72] J. R. Franks and J. Kimble. The confusion of english consonant clusters in lipreading. *Journal of Speech and Hearing Research*, 15:474–482, 1972.

[FM88] K. E. Finn and A. A. Montgomery. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8:159–164, October 1988.

[Fur86] S. Furui. On the role of spectral transition for speech perception. *J. Acoust. Soc. Am.*, 80(4):1016–1025, October 1986.

[GAKS85] K. W. Grant, L. H. Ardell, P. Kuhl, and D. W. Sparks. The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *J. Acoust. Soc. Am.*, 77(2):671–677, February 1985.

[GB91] K. W. Grant and L. D. Braida. Evaluating the articulation index for auditory-visual input. *J. Acoust. Soc. Am.*, 89(6):2952–2960, June 1991.

[GFS95] L. Girin, G. Feng, and J.-L. Schwartz. Noisy speech enhancement with filters estimated from the speaker's lips. In *Proc. of the European Conf. on Speech Communication and Technology*, pages 1559–1562. ESCA, September 1995.

[GFS97] L. Girin, G. Feng, and J. L. Schwartz. Noisy speech enhancement by fusion of auditory and visual information: a study of vowel transitions. In *Eurospeech 1997*. ESCA, September 1997.

[GG95] K. P. Green and A. Gerdeman. Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 21(6):1409–1426, 1995.

[GGMB95] B. Le Goff, T. Guiard-Marigny, and C. Benoit. Read my lips... and my jaw ! How intelligible are the components of a speakers face. In *Proc. of the 4th European Conf. on Speech Communication and Technology*, pages 291–294. ESCA, September 1995.

[GGP94] A. J. Goldschen, O. N. Garcia, and E. Petajan. Continuous optical automatic speech recognition by lipreading. In *28th Annual Asilomar Conf. on Signals, Systems and Computers*, pages 572–577, 1994.

[GK89] K. P. Green and P. K. Kuhl. The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, 45(1):34–42, 1989.

[GK91] K. P. Green and P. K. Kuhl. Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17(1):278–288, 1991.

[GKM88] K. P. Green, P. K. Kuhl, and A. N. Meltzoff. Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *J. Acoust. Soc. Am. Suppl. 1*, 84:S155, Fall 1988.

[GKMS91] K. P. Green, P. K. Kuhl, A. N. Meltzoff, and E. B. Stevens. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 50(6):524–536, 1991.

[GM85] K. P. Green and J. L. Miller. On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, 38(3):269–276, 1985.

[GMS96] M. S. Gray, J. R. Movellan, and T. J. Seijnowski. Dynamic features for visual speech-reading: A systematic comparison. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*. MIT Press of Cambridge, MA, 1996.

[Gon95] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[Gre87] K. P. Green. The perception of speaking rate using visual information from a talker's face. *Perception & Psychophysics*, 42(6):587–593, 1987.

[Gre96a] K. P. Green. Studies of the McGurk effect: Implications for theories of speech perception. In *Proc. of the Fourth Int. Conf. on Spoken Language Processing*, 1996.

[Gre96b] K. P. Green. The use of auditory and visual information in phonetic perception. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 55–77. Springer Verlag, 1996.

[Gre98] K. P. Green. The use of auditory and visual information during phonetic processing: implications for theories of speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by Eye II*, chapter 1, pages 3–25. Psychology Press, 1998.

[Gro68] J. J. Groen. *Slechthorendheid en Hoortostellen.* Stafleu's Wetenschappelijke Uitgeversmaatschappij N. V. Leiden, 1968.

[GSKM90] K. P. Green, E. B. Stevens, P. K. Kuhl, and A. M. Meltzoff. Exploring the basis of the 'McGurk effect': Can percievers combine information from a female face and a male voice ? *J. Acoust. Soc. Am. Suppl.*, 87(1):S125, Spring 1990.

[GTPG01] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy. Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition. In *Proc. of the Int. Conf. of Acoustics, Speech and Signal Processing.* IEEE, May 2001.

[GVN+01] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, USA, May 2001.

[Han81] D. J. Hand. *Discrimination and Classification.* John Wiley & Sons, 1981.

[HBK01] Martin Heckmann, Frederic Berthommier, and Kristian Kroschel. optimal weighting of posteriors for audio-visual speech recognition. In *Proceedings of the ICASSP 2001*, Salt Lake City, Utah, May 2001. IEEE.

[Heb49] D. O. Hebb. *The Organization of Behavior.* Wiley, New York, 1949.

[Hen97] M. Hennecke. *Audio-visual speech recognition preprocessing, learning and sensory integration.* PhD thesis, Stanford University, September 1997.

[HKP91] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation.* Addison Wesley Publishing Company, 1991.

[HSP95] M. E. Hennecke, D. G. Stork, and K. Venkatesh Prasad. Visionary speech: Looking ahead to practical speechreading systems. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 331–352. Springer Verlag, Berlin, 1995.

[HSW89] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[HT85] J. Hopfield and D. W. Tank. Neural computation of decisions in optimization problems. *Biological Cybernetics*, 52:141 – 152, 1985.

[HW93] H. Hild and A. Waibel. Speaker-independent connected letter recognition with a multi-state time delay neural network. In *Proc. of the 3rd European Conf. on Speech, Communication and Technology*, volume 2, pages 1481–1484, Berlin, Germany, September 1993.

[ITF01] K. Iwano, S. Tamura, and S. Furui. Bimodal speech recognition using lip movements measured by optical flow-analysis. In *Proc. of the Int. Workshop on Handsfree Speech Communication*, pages 187–190, April 2001.

[Jac88] P. L. Jackson. The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, 90(5):99–115, September 1988.

[Jel97] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, Massachusetts, 1997.

[JLGW97a] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. In J. Bigün, G. Chollet, and G. Borgefors, editors, *Proc. of the first Int. Conf. on Audio- and Video-based Biometric Person Authentication.* Springer Verlag, March 1997.

[JLGW97b] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Integrating acoustic and labial information for speaker identifikation and verification. In *Proc. of the European Conf. on Speech Communcation and Technology.* ESCA, September 1997.

[JMB76] P. L. Jackson, A. A. Montgomery, and C. A. Binnie. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, 19:796–812, 1976.

[KFMT89] K. Kurosu, T. Furuya, Y. Murayama, and S. Tamura. Lipreading by a neural network method. In *8th European Annual Conf. on Human Decision Making and Manual Control*, pages 329–339, 1989.

[KL82] P. B. Kricos and S. A. Lesner. Differences in visual intelligibility across talkers. *The Volta Review*, 84(4):219–225, May 1982.

[KL85] P. B. Kricos and S. A. Lesner. Effect of talker differences on the speechreading of hearing-impaired teenagers. *The Volta Review*, 87(1):5–14, January 1985.

[KM82] P. K. Kuhl and A. N. Meltzoff. The bimodal perception of speech in infancy. *Science*, 218:1138–1141, December 1982.

[KM92] D. H. Kemsley and T. R. Martinez. A survey of neural network research and fielded applications. *Int. Journal of Neural Networks: Research and Applications*, 2(2/3/4):123–133, 1992.

[KNS01] K. Kumatani, S. Nakamura, and K. Shikano. An adaptive integration method based on product HMM for bi-modal speech recognition. In *Proc. of the Int. Workshop on Hands-free Speech Communication*, pages 195–198, Kyoto, Japan, April 2001.

[KP91] J. F. Kolen and J. B. Pollack. Backpropagation is sensitive to initial conditions. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems*, pages 860–867. Morgan Kaufmann, San Mateo, CA, 1991.

[KTWP97] G. Krone, B. Talle, A. Wichert, and G. Palm. Neural architecture for sensor fusion in speech recognition. In C. Benoit and R. Campbell, editors, *Proc. of the AVSP*, pages 57–60, Rhodes, Greece, September 1997. ESCA.

[LD98] J. Luettin and S. Dupont. Continuous audio-visual speech recognition. In H. Burkhardt and B. Neumann, editors, *Proc. of the 5th European Conf. on Computer Vision*, pages 657–673. Springer Verlag, June 1998.

[LE99] T. C. Levin and M. E. Edgerton. Die Obertonsänger der Tuwa. *Spektrum der Wissenschaft*, (11):50 – 55, November 1999.

[Les88] S. A. Lesner. The talker. *The Volta Review*, 90(5):89–98, September 1988.

[LGT97] Steve Lawrence, C. Lee Giles, and Ah Chung Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97*, pages 540–545, Menlo Park, California, 1997. AAAI, AAAI-Press.

[Lip96] R. P. Lippmann. Recognition by humans and machines: miles to go before we sleep. *Speech Communication*, 18:247–248, 1996.

[Lip97] R. P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22:1–15, 1997.

[LM94] C. R. Lansing and G. W. McConkie. A new method for speechreading research: Tracking observers' eye movements. *Journal of the Academy of Rehabilitative Adiology (JARA)*, 27:25–43, 1994.

[LPN01] J. Luettin, G. Potamianos, and C. Neti. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, USA, May 2001.

[LR92] L. Lisker and M. Rossi. Auditory and visual cueing of the [+/- rounded] feature of vowels. *Language and Speech*, 35(4):391–417, 1992.

[LTB96a] J. Luettin, N. A. Thacker, and S. W. Beer. Speaker identification by lipreading. In *Proc. of the Int. Conf. on Spoken Language Processing*, 1996.

[LTB96b] J. Luettin, N. A. Thacker, and S. W. Beet. Locating and tracking facial speech features. In *Proc. of the Int. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

[LTB96c] J. Luettin, N. A. Thacker, and S. W. Beet. Statistical lip modeling for visual speech recognition. In *Proc. of the VIII European Signal Processing Conference*, Trieste, Italy, 1996.

[Lü90] H. D. Lüke. *Signalübertragung*. Springer Verlag, Berlin Heidelberg New York, 1990.

[LWH90] K. J. Lang, A. Waibel, and G. E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:23–43, 1990.

[Mae90] A. Maelicke. *Vom Reiz der Sinne*. VCH Verlagsgesellschaft mbH, Weinheim, New York, Basel, Cambridge, 1990.

[Mas87] D. W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[Mas96] D. W. Massaro. Bimodal speech perception: A progress report. In D. G. Stork and M. E. Hennecke, editors, *Speereading by Humans and Machines*, pages 79–102. NATO ASI series, Series F, Computer and Systems sciences, vol. 150, 1996.

[Mas97] J. S. Mason. 'Watch these lips' - adding to acoustic signals to improve speaker recognition. In J. Bigün, G. Chollet, and G. Borgefors, editors, *Proc. of the First Int. Conf. on Audio- and Video-based Biometric Person Authentication*, Springer Verlag, March 1997.

[MC83] D. W. Massaro and M. M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5):753–771, 1983.

[MC93] D. W. Massaro and M. M. Cohen. Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, 13:127–134, 1993.

[MC95] D. W. Massaro and M. M. Cohen. Perceiving talking faces. *Current Directions in Psychology*, 4:104–109, 1995.

[MC96] J. R. Movellan and G. Chadderdon. Channel separability in the audio visual integration of speech: A bayesian approach. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Man and Machine*, pages 473–488. NATO ASI series, Series F, Computer and systems sciences, vol. 150, Springer Verlag, 1996.

[MCC+98] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham. Lipreading using shape, shading and scale. In *Proc. of the AVSP '98*, 1998.

[MCG+93] D. W. Massaro, M. M. Cohen, A. Gesi, R. Heredia, and M. Tsuzaki. Bimodal speech perception: an examination across languages. *Journal of Phonetics*, 21:445–478, 1993.

[MCS95] Dominic W. Massaro, Michael M. Cohen, and Paula T. Smeele. Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, 23(1):113–131, 1995.

[Mei95] U. Meier. Robuste Systemarchitekturen für maschinelles Lippenlesen. Master's thesis, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, Am Fasanengarten 5, März 1995.

[MGSW96] K. G. Munhall, P. Gribble, L. Sacco, and M. Ward. Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58(3):351–362, 1996.

[MHD96] U. Meier, W. Hürst, and P. Duchnowski. Adaptive bimodal sensor fusion for automatic speechreading. In *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, 1996.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, December 1976.

[MM99] J.R. Movellan and P. Mineiro. A diffusion network approach to visual speech recognition. In D. W. Massaro, editor, *Proc. of AVSP 99*, pages 92–96, Santa Cruz, California, USA, August 1999.

[MN55] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352, March 1955.

[MNS99] B. Maison, C. Neti, and A. Senior. Audio-visual speaker recognition for video broadcast news: some fusion techniques. In *Multimedia Signal Processing*, Denmark, September 1999.

[Moo98] J. Moody. Forecasting the economy with neural nets: A survey of challenges and solutions. In G. B. Orr and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, chapter 16, pages 347–371. Springer Verlag, 1998.

[Mov95] J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro and D. Touretzky, editors, *Advances in Neural Information Processing Systems, Vol. 7*, pages 851–858. MIT Press, 1995.

[MP43] W. S. McCulloch and W. Pitts. A logical calculus of the idea immanent in nervous activity. *Bulletin of Mathematical Biophysics*, (5):115 – 133, 1943.

[MP69] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry.* MIT Press, Cambridge, MA, 1969.

[MP89] K. Mase and A. Pentland. Lip reading: Automatic visual recognition of spoken words. In *Image Understanding and Machine Vision*, pages 124–127. Optical Society of America, June 1989.

[MP91] K. Mase and A. Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991.

[MS85] M. McGrath and Q. Summerfield. Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acoust. Soc. Am.*, 77(2):678–685, February 1985.

[MS87] A. MacLeod and Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21:131–141, 1987.

[MS91] D. P. Morgan and C. L. Scofield. *Neural Networks and Speech Processing.* Kluwer Academic Publishers, Boston, Dordrecht, London, 1991.

[MSB84] M. McGrath, Q. Summerfield, and M. Brooke. Roles of lips and teeth in lipreading vowels. *Proc. of the Institute of Acoustics*, 6(4):401–408, 1984.

[MSYW99] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lip reading. In *Second Int. Conf. on Multimodal Interfaces*, 1999.

[MVB98] K. G. Munhall and E. Vatikiotis-Bateson. The moving face during speech communication. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by Eye II*, chapter 6, pages 123–139. Psychology Press, 1998.

[MWP87] A. A. Montgomery, B. E. Walden, and R. A. Prosek. Effects of consonantal context on vowel lipreading. *Journal of Speech and Hearing Research*, 30:50–59, March 1987.

[Nee56] K. K. Neely. Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America*, 28(6):1275–1277, November 1956.

[NG96] L. W. Norrix and K. P. Green. Auditory-visual context effects on the perception of /r/ and /l/ in a stop cluster. *J. Acoust. Soc. Am.*, 99(4, Pt. 2):2591, April 1996.

[Nie74] H. Niemann. *Methoden der Mustererkennung.* Akademische Verlagsanstalt, Frankfurt am Main, 1974.

[Nis86] S. Nishida. Speech recognition enhancement by lip-information. *ACM SIGCHI bulletin*, 17(4):198–204, April 1986.

[Nit16] E. B. Nitchie. The use of homophenous words. *The Volta Review*, 18(3):85–93, March 1916.

[NNS97] S. Nakamura, R. Nagai, and K. Shikano. Improved bimodal speech recognition using tied-mixture HMMs and 5000 word audio-visual synchronous database. *Proc. of the European Conf. on Speech Communication and Technology*, pages 1623–1626, 1997.

[NPZ99] P. Niyogi, E. Petajan, and J. Zhong. Feature based representation for audio-visual speech recognition. In D. W. Massaro, editor, *Proc. of the AVSP'99*, pages 97–102, Santa Cruz, CA, August 1999.

[OB85] E. Owens and B. Blazek. Visems observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28:381–393, 1985.

[O'N54] J. O'Neill. Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders*, 19:429–439, 1954.

[O'S87] D. O'Shaughnessy. *Speech Communication.* Addison-Wesley Publishing Company, New York, Amsterdam, Tokyo, 1987.

[Ovi00] S. Oviatt. Multimodal interface research: a science without borders. In *Proc. of the Int. Conf. on Spoken Language Processing*, 2000.

[Pal80] G. Palm. On associative memory. *Biol. Cybernetics*, 36:19–31, 1980.

[Pal82] G. Palm. *Neural Assemblies.* Springer Verlag, 1982.

[PCGR97] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe. Speaker independent audio-visual database for bimodal asr. In C. Benoit and R. Campbell, editors, *Proc. of the AVSP 97*, pages 65–68. ESCA, September 1997.

[Pet84] E. D. Petajan. Automatic lipreading to enhance speech recognition. In *Proc. of the Global Communications Conference*, pages 265–272. IEEE Communication Society, 1984.

[PG98] G. Potamianos and H. P. Graf. Linear discriminant analysis for speechreading. In *Proc. of the IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, pages 221–226, Los Angeles, CA, 1998.

[PLH00] H. Pan, Z.-P. Liang, and T. S. Huang. A new approach to integrate audio and visual features of speech. In *Proc. of the Int. Conf. on Multimedia and Expo*, New York, NY, USA, 30 July-2 August 2000. IEEE.

[PLN01] G. Potamianos, J. Luettin, and C. Neti. Hierarchical discriminant features for audio-visual LVCSR. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, USA, May 2001. IEEE.

[PSS94] G. Palm, F. Schwenker, and F. T. Sommer. Associative memory networks and sparse similarity preserving codes. In V. Cherkassky, editor, *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, pages 282–302. NATO ASI Series F, Springer Verlag, 1994.

[PSSS97] G. Palm, F. Schwenker, F. T. Sommer, and A. Strey. Neural associative memories. In A. Krikelis and C. C. Weems, editors, *Associative processing and processors*, pages 307–326. IEEE Computer Society Press, 1997.

[PTVF92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flamery. *Numerical Recipes in C - The Art of Scientific Computing.* Cambridge University Press, second edition, 1992.

[RD97] A. Rogozan and P. Deleglise. Continuous visual speech recognition using geometric lip-shape models and neural networks. In *Proc. of the European Conf. on Speech Communication and Technology*, Rhodes, Greece, September 1997. ESCA.

[RD98] A. Rogozan and P. Deleglise. Adaptive fusion of acoustic and visual sources for automatic speech recogntion. *Speech Communication*, 26:149–161, 1998.

[RDA97] A. Rogozan, P. Deleglise, and M. Alissali. Adaptive determination of audio and visual weights for automatic speech recognition. In C. Benoit and R. Campbell, editors, *Proc. of the AVSP'97*, pages 61–64, Rhodes, Greece, September 1997.

[Rei78] D. Reisberg. Looking where you listen: visual cues auditory attention. *Acta Psychologica*, 42:331–341, 1978.

[RFM81] S. M. Rosen, A. J. Fourcin, and B. C. J. Moore. Voice pitch as an aid to lipreading. *Nature*, 291:150–152, May 1981.

[Rip96] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.

[RJ86] L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, January 1986.

[RJ93] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[RL78] A. Risberg and J. L. Lubker. Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress & Status Report, Stockholm (STL-QPSR)*, 4:1–16, 1978.

[RL91] M. D. Richard and R. P. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3(4):461–483, 1991.

[RM86a] D. E. Rumelhart and J. L. McClelland. On learning the past tenses of english verbs. In J. L. McClelland, D. E. Rumelhart, and the PDP Group, editors, *Parallel distributed processing, Vol. 2: Psychological and Biological Models*, pages 216–271. MIT Press, 1986.

[RM86b] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundation. MIT Press, 1986.

[RM94] R. R. Rao and R. M. Mersereau. Lip modeling for visual speech recognition. In *28th Annual Asilomar Conference on Signals, Systems and Computers*, 1994.

[RMG87] D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd and R. Campbell, editors, *Hearing by Eye: The Psychology of Lip-reading*, chapter 4, pages 97–113. Lawrence Erlbaum Associates, Publishers, London, Hillsdale, New Jersey, 1987.

[Ros58] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386 – 408, 1958.

[Ros62] F. Rosenblatt. *Principles of Neurodynamics.* Spartan Books, New York, 1962.

[RRLES93] J. Robert-Ribes, T. Lallouache, P. Escudier, and J.-L. Schwarts. Integrating auditory and visual representations for audiovisual vowel recognition. In *Proc. of the European Conf. on Speech Communication and Technology*, pages 1753–1756. ESCA, 1993.

[RRPSE96] J. Robert-Ribes, M. Piquemal, J.-L. Schwartz, and P. Escudier. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 193–210. Springer Verlag, 1996.

[RRSE94] J. Robert-Ribes, J.-L. Schwartz, and P. Escudier. Audio-visual recognition of speech units: A tentative functional model compatible with psychological data. In R. Togneri, editor, *Proc. of the Fifth Australian Int. Conf. on Speech Science and Technology*, pages 448–453, December 1994.

[RRSE95a] J. Robert-Ribes, J.-L. Schwartz, and P. Escudier. Auditory, visual and audiovisual vowel representations: Experiments and modelling. In *XIIIth Int. Congress of Phonetic Sciences*, pages 114–121, 1995.

[RRSE95b] J. Robert-Ribes, J.-L. Schwartz, and P. Escudier. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, 9:323–346, 1995.

[RRSLE98] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier. Complementary and synergy in bimodal speech: Auditory, visual, and audio-visual identification of french oral vowels in noise. *J. Acoust. Soc. Am.*, 103(6):3677–3689, June 1998.

[RS81] M. Roberts and Q. Summerfield. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, 30(4):309–314, 1981.

[SB93] P. L. Silsbee and A. C. Bovic. Audio-visual speech recognition for a vowel discrimination task. *Proc. SPIE Symp. Visual Communications and Image Processing*, 1:84–95, November 1993.

[SB96] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, September 1996.

[Sek94] K. Sekiyama. Differences in auditory-visual speech perception between japanese and americans: McGurk effect as a function of incompability. *Journal of the Acoustical Society of Japan (E)*, 15(3):143–158, 1994.

[Sek97] K. Sekiyama. Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in chinese subjects. *Perception & Psychophysics*, 59(1):73–80, 1997.

[Sio95] O. Siohan. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. In *Proc. of the Int. Conf. Acoustics, Speech and Signal Processing*, pages 125–128, 1995.

[SJ96] E. A. Strand and K. Johnson. Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon, editor, *Natural Language Processing and Speech Technology - Results of the 3rd KONVENS Conference*, pages 14–26. Nouton de Gruyter, October 1996.

[SK92] J. Schürmann and U. Kreßel. Mustererkennung mit statistischen Methoden. Vorlesungsskript für das Sommersemester 1992 und 1993, Universität Ulm, 1992.

[SM84] Q. Summerfield and M. McGrath. Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, 36A:51–74, 1984.

[SMMB89] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke. Lips, teeth, and the benefits of lipreading. In A. W. Young and H. D. Ellis, editors, *Handbook of Research on Face Processing*, pages 223–233, North-Holland, 1989. Elsevier Science Publishers B. V.

[SMS+98] M. Sams, P. Manninen, V. Surakka, P. Helin, and R. Kättö. McGurk effect in finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, 26:75–87, 1998.

[SNM99] A. Senior, C.V. Neti, and B. Maison. On the use of visual information for improving audio-based speaker recognition. In D. W. Massaro, editor, *Proc. of the AVSP 99*, pages 108–111, Santa Cruz, California, USA, August 1999.

[SP54] W. H. Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustic Society of America*, 26(2):212–215, 1954.

[SR94] H. M. Saldana and L. D. Rosenblum. Selective adaptation in speech perception using a compelling audiovisual adaptor. *J. Acoust. Soc. Am.*, 95(6):3658–3661, June 1994.

[SRRE98] J.-L. Schwartz, J. Robert-Ribes, and P. Escudier. Ten years after summerfield: a taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, and D. Burnham, editors, *Hearing by Eye II - Advances in the Psychology of Speechreading and Auditory-visual Speech*, chapter 4, pages 85–108. Psychology Press, 1998.

[SSHK97] M. Sams, V. Surakka, P. Helin, and R. Kättö. Audiovisual fusion in finnish syllables and words. In C. Benoit and R. Campbell, editors, *Proc. of the Workshop on Audio-Visual Speech Processing*, pages 101–103. ESCA, September 1997.

[ST91] K. Sekiyama and Y. Tohkura. McGurk effect in non-english listeners: Few visual effects for japanese subjects hearing japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.*, 90(4):1797–1805, October 1991.

[ST93] K. Sekiyama and Y. Tohkura. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21:427–444, 1993.

[Ste61] K. Steinbuch. Die Lernmatrix. *Kybernetik 1*, 1:36 – 45, 1961.

[Ste71] K. Steinbuch. *Automat und Mensch.* Springer Verlag, 1971.

[Sto97] D. G. Stork, editor. *HAL's Legacy.* MIT Press, Cambridge, Massachusetts, 1997.

[Str96] G. Strube, editor. *Wörterbuch der Kognitionswissenschaft.* Klett-Cotta Verlag, Stuttgart, 1996.

[STU96] K. Sekiyama, Y. Tohkura, and M. Umeda. A few factors which affect the degree of incorporating lip-read information into speech perception. In *Proc. of the Fourth Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.

[Sum79] Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36:314–331, 1979.

[Sum87] Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In B Dodd and R. Campbell, editors, *Hearing by Eye: The Psychology of Lip-Reading*, chapter 1, pages 3–51. Lawrence Erlbaum Associates Publishers, London, Hillsdale, New Jersey, 1987.

[Sum92] Q. Summerfield. Lipreading and audio-visual speech perception. *Pil. Trans. R. Soc. Lond. B*, 335:71–78, 1992.

[SWL92] D. G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proc. of the Int. Joint Conf. on Neural Networks*, pages II–285–II–295, Baltimore, NY, June 1992.

[SYMHGJ90] T. J. Sejnowski, B. P. Yuhas, Jr. M. H. Goldstein, and R. E. Jenkins. Combining visual and acoustic speech signals with a neural network improves intelligibility. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 232–239, San Mateo, CA, 1990. Morgan Kaufmann.

[Tay56] W. K. Taylor. Electrical simulation of some nervous system functional activities. *Information Theory*, 3:314 – 328, 1956.

[TGDS98] P. Teissier, A. Guerin-Dugue, and J.-L. Schwartz. Models for audiovisual fusion in a noisy-vowel recognition task. *Journal of VLSI Signal Processing*, 20:25–44, 1998.

[The89] C. W. Therrien. *Decision Estimation and Classifikation.* John Wiley & Sons, 1989.

[TKP97] B. Talle, G. Krone, and G. Palm. Comparison of neural architectures for sensorfusion. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, volume 4, pages 3273–3276, Munich, Germany, April 1997.

[TRB96] M. J. Tomlinson, M. J. Russell, and N. M. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, pages 821–824, 1996.

[TRRSGD99] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guerin-Dugue. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7(6):629–642, November 1999.

[TW99] B. Talle and A. Wichert. Audio-visual sensorfusion with neural architectures. In D. Massaro, editor, *Proc. of the AVSP*, pages 103–107, Santa Cruz, California, USA, August 1999.

[VBEY94] E. Vatikiotis-Bateson, I.-M. Eigsti, and S. Yano. Listener eye movement behavior during audiovisual speech perception. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 527–530, 1994.

[VBMH+95] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. V. Lee, and D. Terzopoulos. The dynamics of audiovisual behavior in speech. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 221–232. Springer Verlag, September 1995.

[VFNB99] A. Verma, T. Faruquie, C. Neti, and S. Basu. Late integration in audio-visual continuous speech recognition. In *Proc. of the ASRU*, 1999.

[Vog95] M. Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination cinditions. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, pages 399–460. Springer Verlag, 1995.

[Vro92] J. H. M. Vroomen. *Hearing Voices and Seeing Lips - Investigations in the Psychology of Lipreading.* PhD thesis, Katholieke Universiteit Brabant, September 1992.

[Wai89] A. Waibel. Consonant recognition by modular construction of large phonemic time-delay neural networks. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 215–223. Morgan Kaufmann, 1989.

[WB60] M. F. Woodward and C. G. Barber. Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, 3:212–222, 1960.

[WBLH69] D. Willshaw, O. Buneman, and H. Longuet-Higgins. Nonholographic associative memory. *Nature*, 222:960–962, 1969.

[WBO95] S. Walker, V. Bruce, and C. O'Malley. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57(8):1124–1133, 1995.

[WG94] A. S. Weigend and N. A. Gershenfeld, editors. *Time Series Predicition: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1994.

[Wic69] W. A. Wickelgren. Context-sensitive coding, associative memory and serial order in (speech) behavior. *Psychological Review*, 76:1–15, 1969.

[Wic00] A. Wichert. *Associative Computation.* PhD thesis, Universität Ulm, 2000.

[WJ79] V. Wozniak and P. L. Jackson. Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research*, 22:354–365, 1979.

[WPM+77] B. E. Walden, R. A. Prosek, A. A. Montgomery, C. K. Scherr, and C. J. Jones. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20:130–145, 1977.

[WPW75] B. E. Walden, R. A. Prosek, and D. W. Worthington. Auditory and audiovisual feature transmission in hearing-impaired adults. *Journal of Speech and Hearing Research*, 18:272–280, 1975.

[WTM+91] J.-T. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K. Okazaki. Neural network vowel-recognition jointly using voice features and mouth shape image. *Pattern Recognition*, 24(10):921–927, 1991.

[YMHGS89] B. P. Yuhas, Jr. M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27:65–71, November 1989.

[YMHGSJ90] B. P. Yuhas, Jr. M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins. Neural network models of sensory integration for improved

vowel recognition. *Proc. of the IEEE*, 78(10):1658–1668, October 1990.

[Zel94] A. Zell. *Simulation neuronaler Netze.* Addison-Wesley Publishing Company, 1994.

[ZF67] E. Zwicker and R. Feldtkeller. *Das Ohr als Nachrichtenempfänger.* S. Hirzel Verlag, Stuttgart, 1967.

[ZLH00] Y. Zhang, S. Levinson, and T. Huang. Speaker independent audio-visual speech recognition. In *Proc. of the Int. Conf. on Multimedia and Expo.* IEEE, 30 July - 2 August 2000.

[Zwi61] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248, February 1961.

Danksagung

Mein erster Dank gilt Professor Dr. Günther Palm für seine Betreuung dieser Arbeit, die stets hilfreichen Diskussionen und die Übernahme des Hauptgutachtens. Auch den Professoren Dr. Heiko Neumann und Dr. Heinrich Niemann gilt mein Dank für die Übernahme der Gutachtertätigkeit dieser Arbeit.
Allen Kolleginnen und Kollegen der Abteilung Neuroinformatik der Universität Ulm danke ich für die freundschaftliche und angenehme Arbeitsatmosphäre. Stellvertretend möchte ich dabei Axel Baune, Dr. Andrea Bibbig, Dr. Guido Kolano, Gabi Krone, Dr. Andreas Küchler, Birgit Lonsinger, Dr. Friedhelm Schwenker, Dr. Fritz Sommer, Dr. Alfred Strey und Dr. Andreas Wichert nennen, die auf die eine oder andere Art nicht unwesentlich zum Gelingen dieser Arbeit beitrugen.
Mein besonderer Dank gilt meinen Freunden und meiner Familie, die mich stets unterstützten und mir mit Rat und Tat zur Seite standen.

Barbara Talle