Steuerung sprechernormalisierender Abbildungen durch künstliche neuronale Netzwerke

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Knut Müller
aus Frankfurt am Main

Göttingen 2000

D 7
Referent: Prof. Dr. M. R. Schroeder
Korreferent: Prof. Dr. D. Ronneberger
Tag der mündlichen Prüfung: 1.11.2000

Für Karl-Heinz, Tino, Jürgen und Karin.

Inhaltsverzeichnis

1. Einführung

Dieses Kapitel gibt eine kurze Übersicht über den Bereich der automatischen Spracherkennung und die Problematik der Erkennungssicherheit, so daß hiernach die vorliegende Arbeit thematisch eingeordnet werden kann. Daran anschließend findet man eine Beschreibung des Aufbaus und der Gliederung des Textes.

1.1. Die Struktur automatischer Spracherkennungssysteme

Ein automatisches Spracherkennungssystem dient der Identifikation von Lauten oder Äußerungen in gesprochener Sprache. Das Wissen um die Identität eines Lautes kann z. B. zur Steuerung irgendeiner Apparatur verwendet werden. Andere typische Anwendungen sind Diktier- oder Dialogsysteme. Bei Informationssystemen, die häufig in einem eingeschränkten Kontext verwendet werden, kann man sich gegebenenfalls mit der Identifikation von Schlüsselwörtern begnügen, um unabhängig von der genauen Formulierung die gewünschte Information bereitzustellen *(Word Spotting)*. Klassisches Beispiel hierfür ist ein automatisches System zur Fahrplanauskunft. Eine verwandte denkbare (wenngleich auch wenig wünschenswerte) Anwendung sind Überwachungssysteme, die beispielsweise den Telefonverkehr nach Schlüsselwörtern durchsuchen und dadurch Gespräche identifizieren, die für den Überwacher von potentiellem Interesse sind.

Spracherkennungssysteme bestehen aus einer Reihe von Konstituenten. In der Anwendung benötigen sie naturgemäß ein Aufnahmesystem (Telefon, Headset, freies Mikrophon) inklusive der unvermeidlichen Vorverstärker und Filter. Daran kann sich eine für die Spracherkennung geeignete analoge Aufbereitung der Signale anschließen. Da derzeit alle Spracherkennungssysteme auf Digitalrechnern basieren, folgt ein Analog-Digital-Wandler, und daran schließt sich eine Stufe zur weiteren Vorverarbeitung des digitalisierten Sprachsignals an. Unter Umständen wird auf die analoge Aufbereitung auch verzichtet. Deren Aufgaben werden dann vollständig von der digitalen Vorverarbeitung übernommen. Diese besteht nicht nur aus einer einfachen Filterung des Signals, sondern aus einer komplizierteren Transformation in einen vektorwertigen Datenstrom, der dem eigentlichen Erkenner zugeführt wird. Die Art der Vorverarbeitung orientiert sich gewöhnlich an der Signalverarbeitung im menschlichen Gehör, da die Annahme, daß diese Art der Vorverarbeitung grundsätzlich auch

für die automatische Spracherkennung förderlich ist, durch zahlreiche Experimente gestützt wird.[1]

Der anschließende Erkennungsschritt besteht im einfachsten Falle aus einer Klassifizierung der lautlichen Äußerung als eine von N bekannten Äußerungen. Das Ergebnis steht hiernach zu einer weiteren Verarbeitung zur Verfügung. Diese »höhere« Verarbeitung kann z. B. aus einer Wahrscheinlichkeits- und Plausibilitätsbewertung des klassifizierten Lautstromes anhand von Gegebenheiten wie etwa der Silbenbauregeln der erwarteten Sprache, der Grammatik und des Kontextes bestehen.

1.2. Der Erkennungsfehler

Leider hat jedes Spracherkennungssystem (wie auch der Mensch) einen gewissen Erkennungsfehler, was bedeutet, daß manche Laute nicht richtig klassifiziert werden. Der Erkennungsfehler eines Spracherkennungssystems hängt von vielen Faktoren ab, wie z. B. der Architektur des Systems, der Art der Vorverarbeitung des Eingangssignals und dem »Training« des Systems, also letztlich der Wahl der Klassifizierungsparameter.

Die Wahl der Klassifizierungsparameter ist deswegen einer der wesentlichen Gründe für den Erkennungsfehler, da »gleiche Laute«, womit phonetisch äquivalente Sprachsignale gemeint sein sollen, in ihren Realisierungen eine große akustische Variabilität aufweisen können. Wegen der nach wie vor unvollkommenen Kenntnis über die akustischen Invarianten dieser gleichen Laute führen die auftretenden Variabilitäten zu einer deutlichen Streuung in den in der Literatur benannten Klassifizierungsräumen, so daß sich die Klassen im allgemeinen überlappen. Dies führt zwangsläufig zu fehlerhaften Klassifizierungen und damit einem gewissen Erkennungsfehler.

Bei Klassifizierungsaufgaben stellen sich grundsätzlich zwei Probleme: Die Klassifizierung wird umso schwieriger,

1. je mehr Klassen existieren (z. B. für Worterkennung: Je größer der zu erkennende Wortschatz ist) und

2. je größer die Streuung innerhalb der Klassen ist, d. h. je größer die Variabilität der zu erkennenden Sprachsignale ist.

Die Anzahl der Klassen ist gemeinhin durch das spezielle Erkennungsproblem, also durch die projektierte Anwendung des Systems gegeben und kann nicht ohne weiteres reduziert werden. Daher kann man sich zur Verbesserung der Erkennungsleistungen nur auf den zweiten Aspekt konzentrieren, also

1. auf die Verminderung der Streuung innerhalb des Klassifizierungsraumes, was äquivalent mit der Verbesserung der Vorverarbeitung ist, und

2. auf die Verbesserung der Klassifizierung selbst, was auf eine Anpassung der Parameter des Klassifikators hinausläuft.

[1] Eine Diskussion hierüber findet man beispielsweise in (Hermansky, 1998).

1.3. Thema dieser Arbeit

Die vorliegende Arbeit beschäftigt sich mit der Verbesserung der Vorverarbeitung mit dem Ziel, die Variabilität in der Darstellung gleicher Laute zu reduzieren. Dies wird durch eine *Sprechernormalisierung* erreicht, also eine Transformation der aufbereiteten Sprachsignale eines Sprechers in die eines gegebenen Zielsprechers.

Der Versuch einer Sprechernormalisierung ist ein etwas heikler Ansatz, da letztlich – wie bei der gesamten Aufbereitung des Sprachsignals – Information vernichtet werden soll. Jene nämlich, die letztlich zu der sprecherspezifischen Streuung im Klassifizierungsraum führt. Dabei muß aber die lautspezifische Variation des Sprachsignals, die letztlich die Klassifikation der Laute erlaubt, weitestmöglich erhalten bleiben. Möglicherweise ist dies ein Grund dafür, daß über Sprechernormalisierung viel weniger Literatur existiert als über *Sprecheradaptation* (siehe Abschnitt 3.2.1).

Ziel dieser Arbeit ist nicht die Optimierung bestehender Methoden. Vielmehr stellen im Bereich der Sprechernormalisierung die Konstruktion einer Abbildung ohne ein bekanntes Labeling der Äußerungen und/oder ohne größere verfügbare Mengen von Datenmaterial des aktuellen Testsprechers besondere Herausforderungen dar. Beides wird in dieser Arbeit versucht. Dabei werden hauptsächlich Sprechernormalisierungsmethoden betrachtet, die auf der Abbildung durch simulierte neuronale Netze (*Perzeptrone*) basieren. Besondere Berücksichtigung erfährt hierbei die Frage nach der Interpolierbarkeit zwischen diesen normalisierenden Abbildungen, um den Datenaufwand zur Ermittlung einer neuen Abbildung für einen bis dahin unbekannten Quellsprecher zu reduzieren. Dazu müssen Ähnlichkeitsbeziehungen zwischen den Sprechern gefunden werden. Auch hierfür werden Perzeptrone eingesetzt. Die dabei erhaltenen Einblicke in die Struktur des durch die Vorverarbeitung gebildeten Merkmalsraumes werden diskutiert.

1.4. Aufbau dieser Arbeit

Im *nächsten Kapitel* wird ein großer Bogen über die automatische Spracherkennung gezogen. Dabei wird kurz auf die Sprachproduktion und den Aufbau des Gehörs eingegangen, um dann einen Blick auf die Sprachperzeption und die Signaleigenschaften bestimmter Lautklassen zu werfen. Dadurch wird die Art der Vorverarbeitung des Sprachsignals motiviert, die anschließend besprochen wird. Hierauf werden zwei wichtige Methoden (DTW und HMM) zur Klassifikation der solcherart aufbereiteten Sprachsignale erklärt. Dieses relativ umfangreiche Kapitel soll auch dem Leser, der mit der Sprachsignalverarbeitung nicht so vertraut ist, einen Einblick in dieses Thema gewähren und, wie die Anhänge, späteres Nachschlagen der Konzepte erlauben.

Kapitel 3 geht auf den engeren Bereich der Thematik dieser Arbeit ein, nämlich die Sprechernormalisierung. Zunächst werden die Ursachen der Inter- und Intrasprechervarianzen der Merkmalsvektoren gleicher Äußerungen kurz diskutiert, dann die wesentlichen Ansätze und zugehörigen Methoden zur Verringerung der Intersprechervarianz besprochen und so der Kontext für die folgenden Experimente hergestellt.

In *Kapitel 4* wird das verwendete Datenmaterial und seine Verarbeitung vorgestellt. Anschließend wird eine Methode zur unüberwachten Findung sprechernormalisierender Abbildungen entwickelt und die Ergebnisse dieser Methode dargestellt. Darauf folgt die Beschreibung einer Sprechernormalisierungsmethode durch einfache Perzeptrone, deren Eignung belegt wird. Hierauf folgt ein Versuch zur Interpolation der auf diese Weise konstruierten Abbildungen für einen unbekannten Sprecher, dessen Position im »Sprecherraum« zuvor durch die Berechnung von mittleren DTW-Abständen und anschließende multidimensionale Skalierung ermittelt wurde. Nachdem sich dieses Vorgehen als erfolgreich erwiesen hat, wird nun eine Methode zur automatischen Ermittlung der Sprecherposition durch ein mehrschichtiges Perzeptron beschrieben. Als Steuerparameter werden hierfür Barkspektrogramme und artikulatorische Parameter verwendet.

Eine Zusammenfassung der Ergebnisse und einen Ausblick findet man in *Kapitel 5*. Der *Anhang* schließlich informiert über die wichtigsten Algorithmen, die verwendete Lautschrift und das Lautsystem.

2. Aufbau eines Spracherkennungssystems

In Abbildung 2.1 ist der allgemeine Aufbau eines Spracherkennungssystems skizziert. Die Spracheingabe des Benutzers wird von einer Worterkennungseinheit klassifiziert und in der Form eines Stromes von Hypothesen an die höhere Verarbeitung weitergereicht.

Syntaktisches[1], semantisches[2] und pragmatisches[3] Wissen wird zur Bewertung der Hypothesen herangezogen und wirkt auf die Worterkennung und die höhere Verarbeitung zurück, wodurch ggf. auch die Komplexität der Erkennungsaufgabe reduziert wird. Teil dieses Wissens ist die Aufgabenbeschreibung des Spracherkennungssystems und der Kontext, der aus bereits analysierten Bestandteilen der Äußerung gebildet wird.

Vorverarbeitung, Mustererkennung und höhere Verarbeitung werden im folgen-

[1]Syntax (Satzlehre): Beschreibung der formal zulässigen Verbindungen von Wörtern zu Wortgruppen und Sätzen (nach Digel und Kwiatkowski (1987)).
[2]Semantik (Bedeutungslehre): Bescheibung der mit den Wörtern verbundenen Vorstellungen und Konzepte (nach Digel und Kwiatkowski (1987)).
[3]Pragmatik (Lehre vom sprachlichen Handeln): Beschreibung der Beziehungen zwischen den Wörtern und den von den Benutzern der Wörter intendierten und erzielten Wirkungen (nach Digel und Kwiatkowski (1987)).

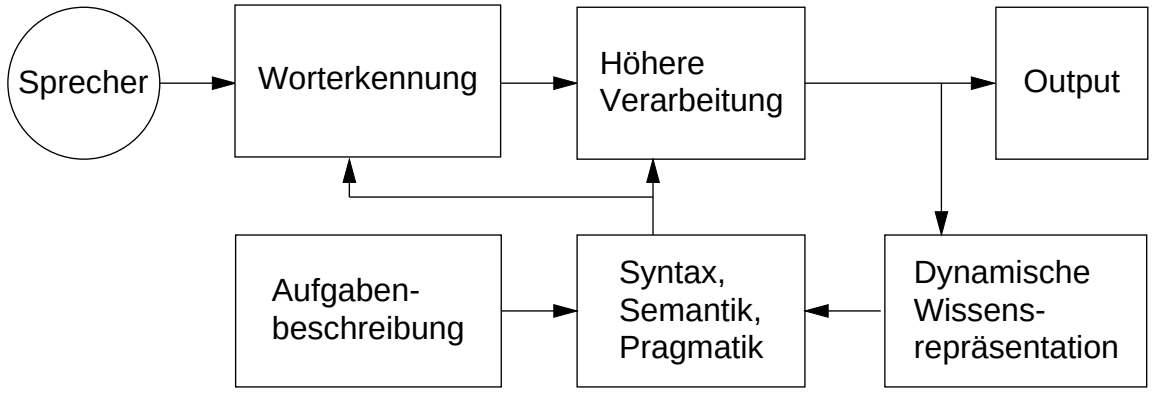

Abbildung 2.1.: Genereller Aufbau eines aufgabenorientierten Spracherkennungssystems (Nach Rabiner und Juang (1993))

den diskutiert. Dabei wird keine Allgemeingültigkeit beansprucht, jedoch sind die beschriebenen Verfahren verbreitet.

Die Vorverarbeitung des Sprachsignals hat zum Ziel, das akustische Sprachsignal in eine Form zu transformieren, auf der der eigentliche Klassifikationsprozeß möglichst effizient durchgeführt werden kann.

Um die Entscheidung darüber treffen zu können, welche Art der Vorverarbeitung denn wohl für die automatische Spracherkennung geeignet ist, ist es zweckmäßig, zunächst die Art der menschlichen Sprachproduktion und die Signalverarbeitung im menschlichen Gehör zu betrachten. Die Motivation hierfür ist zunächst die Tatsache, daß Abschätzungen für das akustische Sprachsignal (im Sinne der Informationstheorie) einen Informationsfluß von etwa 30 kb/s ergeben, wohingegen das symbolische Äquivalent hierfür nur eine Rate von etwa 50 b/s erfordert (Flanagan, 1965).[4] Natürlich ist in dem Symbolstrom nicht alle durch das Sprachsignal übermittelte Information enthalten. Aspekte wie Prosodie werden vernachlässigt.[5] Andererseits ist die im Symbolstrom enthaltene Information für die Belange der Spracherkennung normalerweise ausreichend. In diesem Sinne ist das Sprachsignal offenbar hochredundant. Daher stellt sich die Frage, welche Aspekte des Sprachsignals für die automatische Spracherkennung ignoriert werden können. Aus diesem Grund muß man einen Blick auf die Sprachproduktion und die Sprachperzeption werfen, denn nur diejenigen Geräusche, die produziert und auch wahrgenommen werden, berühren die menschliche Spracherkennung, welche letztlich den Maßstab für die maschinelle Spracherkennung bildet.

2.1. Sprachproduktion

Bei der menschlichen Sprachproduktion wird Luft aus der Lunge durch die *Glottis*[6] in den Vokaltrakt gedrückt[7]. Dieser besteht im wesentlichen aus dem Rachenraum, der Mund- und der Nasenhöhle und endet mit den Lippen bzw. mit den Nasenlöchern. Innerhalb des Vokaltraktes befinden sich die Artikulationsorgane, also das Gaumensegel *(Velum)*, die Zunge und die Lippen. Die Lautgebung erfolgt durch die Stimmlippen (stimmhafte Laute), durch turbulentes Rauschen an Engstellen (Frikative), durch das plötzliche Aufgeben von Verschlüssen im Vokaltrakt (Plosive) oder Kombinationen davon (stimmhafte Frikative, aspirierte Plosive usw.).

[4] Flanagan betrachtet hierfür einen Übertragungskanal mit 30 dB Signal-Rausch-Abstand *(SNR)* und setzt für die Informationsrate des Symbolstroms die Phone des Englischen mit ihren relativen Häufigkeiten an.

[5] Unter *Prosodie* versteht man silben-, wort- oder satzübergreifende lautliche Merkmale wie Satzmelodie, Segmentdauer und -betonung. Unter Einbeziehung der Prosodie steigt der Informationsfluß auf etwa 200 b/s, auf der Ebene der neuromuskulären Umsetzung in Artikulatorbewegungen beträgt der Informationsfluß etwa 2 kb/s (Rabiner und Juang, 1993).

[6] Glottis: Der Zwischenraum zwischen den Stimmlippen.

[7] Die Produktion von Sprachlauten während des Einatmens ist sehr selten.

2.1.1. Eigenschaften des Sprachsignals

Relativ zu den Stimmlippen haben alle Artikulationsorgane eine große Masse, so daß ihre Positionierung im Vergleich zu der Frequenz der Lautgebung recht langsam vonstatten geht. Die Frequenz der gebildeten Laute ist also relativ gering. Bei vokalischer Anregung des Stimmkanals werden Fluktuationen in der Anregungsfrequenz als nichtbedeutungstragend angenommen, weshalb das Sprachsignal innerhalb kurzer Zeitfenster (in der Größenordnung von 5–100 ms) als annähernd stationär gelten kann (Flanagan, 1965). In der Praxis wird daher die Analyse des Sprachsignals auf Signalstücken von kurzer Dauer durchgeführt. Der Frequenzbereich, innerhalb dessen das Sprachsignal signifikante Energie aufweist, liegt etwa zwischen 100 und 8 000 Hz. Das Signal hat eine stark schwankende Amplitude, welche sich unter normalen Umständen zwischen 30 und 90 dB SPL[8] (in 1 m Abstand von den Lippen) bewegt.

2.2. Sprachperzeption

2.2.1. Aufbau des Gehörs

Das menschliche Gehör ist ein bemerkenswertes Organ. Dennoch unterliegt es naturgemäß gewissen Beschränkungen in seiner Fähigkeit, Sprachlaute (oder Geräusche überhaupt) zu analysieren. Und da das, was vom Gehör nicht wahrgenommen wird, vermutlich auch für die automatische Spracherkennung nicht förderlich ist, sollen hier zunächst der Aufbau und die Funktionsweise des Gehörs dargestellt werden. Im nächsten Abschnitt wird dann auf die sich daraus ergebenden Fähigkeiten und Beschränkungen eingegangen.

Die Sprachperzeption basiert auf der Analyse des akustischen Sprachsignals durch das Ohr. Das Ohr besteht aus dem Außenohr (Ohrmuschel und Gehörgang), dem Mittelohr (Paukenhöhle mit Gehörknöchelchen) und dem Innenohr (Schnecke oder *Cochlea* und Bogengänge des Gleichgewichtsorgans). Die Ohrmuschel dient der Schallortung und erhöht die Empfindlichkeit des Gehörs gegenüber frontalen Schallereignissen (O'Shaughnessy, 1987). Der Gehörgang bildet einen Resonator und erhöht so die Empfindlichkeit des Ohres im Bereich von 3–5 kHz um bis zu 12–15 dB. Der Gehörgang wird durch das Trommelfell abgeschlossen, an das sich das luftgefüllte Mittelohr anschließt.

Im Mittelohr werden die Schwingungen des Trommelfells über die drei Gehörknöchelchen (Hammer, Amboß und Steigbügel) auf das ovale Fenster übertragen, welches die Begrenzung zum mit Lymphe gefüllten Innenohr bildet. Die Funktion des Mittelohres liegt in der Impedanzanpassung an das flüssigkeitsgefüllte Innenohr. Die Konzentration der Schallenergie von der relativ großen Fläche des Trommelfells (etwa 55 mm^2) auf das kleinere ovale Fenster (etwa 3,2 mm^2) trägt wesentlich zur Erhöhung des Schalldruckes (pro Einheitsfläche) im Mittelohr um etwa 30 dB bei. Dabei verhält

[8] SPL *(Sound Pressure Level)*: Schalldruckpegel. 0 dB SPL entsprechen einem Schalldruck von 20 μPa, was in etwa der Hörschwelle bei 1 kHz entspricht.

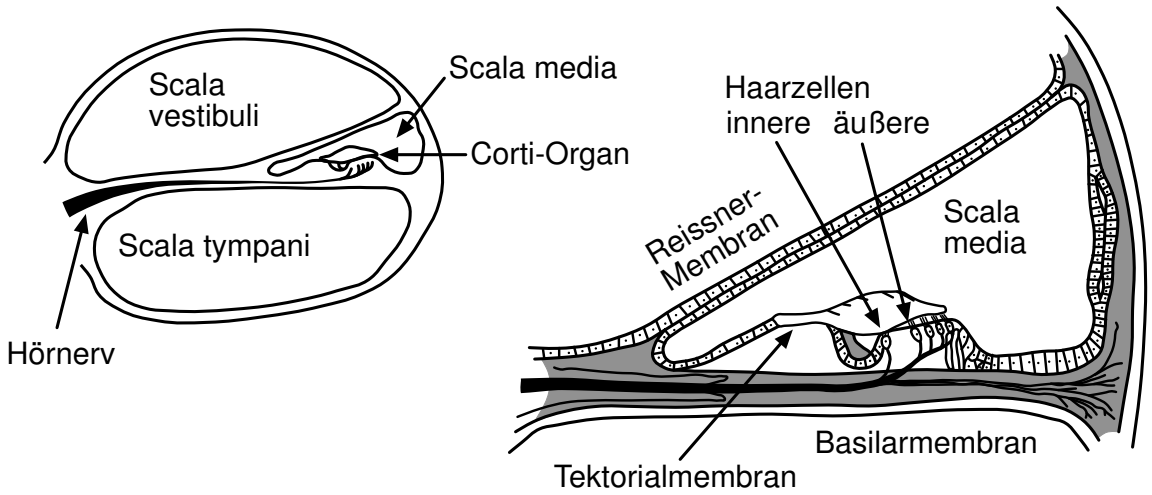

Abbildung 2.2.: Querschnitt durch eine Windung der Cochlea und Ausschnittsvergrößerung um das Corti-Organ (Nach (O'Shaughnessy, 1987))

sich das Mittelohr wie ein Tiefpaß-Filter mit einem Abfall von etwa 15 dB/Oktave oberhalb von 1 kHz.

Im Innenohr schließlich werden die mechanischen Schwingungen am ovalen Fenster in Reize des Hörnerves umgesetzt. Dies geschieht in der Cochlea, einer gewundenen, sich verjüngenden flüssigkeitsgefüllten Röhre, welche durch zwei Membranen längs in drei Kanäle aufgeteilt ist. Die mechanischen Schwingungen des ovalen Fensters werden in einen der beiden äußeren Kanäle (die *Scala vestibuli*) eingekoppelt. Da die Bewandung der Cochlea knöchern ist und die Flüssigkeit inkompressibel, bewirkt die Schwingung eine Bewegung der die Scala vestibuli begrenzenden Reissner[9]-*Membran*. Zwischen der Scala vestibuli und dem anderen äußeren Kanal, der *Scala tympani*, ist im *Apex* (der Spitze) der Cochlea (dem *Helicotrema*) ein Flüssigkeitsaustausch möglich. Am äußeren Ende der Scala tympani ist ein Druckausgleich über das runde Fenster möglich. Der mittlere Kanal, die *Scala media*, ist gegen die Scala tympani durch die Basilarmembran, gegen die Scala vestibuli durch die Reissner-*Membran* getrennt (Abb. 2.2). Die Basilarmembran ändert in ihrem Verlauf ihre Breite und Steifigkeit, und zwar ist sie am basalen Ende der Cochlea schmaler und steifer, im Apex der Cochlea dagegen breiter und weicher. Dadurch reagiert sie an verschiedenen Orten mehr oder weniger stark auf die verschiedenen Anregungsfrequenzen und führt so eine Frequenz-Ortstransformation durch.

Innerhalb der Scala media liegt das Corti[10]-*Organ* auf der Basilarmembran, welches die Sinneszellen trägt, auf denen der Hörnerv endet. Diese Zellen werden von der *Tektorialmembran* bedeckt. Jede dieser Sinneszellen trägt feine Härchen (*Stereozilien*), die gegen die Tektorialmembran gerichtet sind. Es existieren zwei Arten von Haarzellen, die *äußeren* und die *inneren* Haarzellen. Auf den etwa 5 000 inneren Haarzellen

[9] Ernst Reissner, Anatom, Dorpat, 1824–1878.
[10] Alfonso de Corti, Pathologe, Wien, 1822–1876.

enden 95 % der afferenten[11] Nerven des Hörnervs, wohingegen die restlichen 5 % auf den viel zahlreicheren äußeren Haarzellen enden. Bei den efferenten[12] Nerven verhält es sich gerade umgekehrt.

Die Härchen der äußeren Haarzellen sind in die Tektorialmembran eingebettet, die der inneren Haarzellen berühren sie nur. Man vermutet daher, daß die inneren Haarzellen durch die Scherung der Härchen aus den Schwingungen zwischen der Basilarmembran und der Tektorialmembran gereizt werden und im wesentlichen für den Sinneseindruck verantwortlich sind, die äußeren Haarzellen dagegen in Umkehrung des Prinzips die Tektorialmembran positionieren und auf diese Weise zu einer positiven Rückkopplung und dadurch zu einer weiteren Steigerung der Empfindlichkeit des Gehörs beitragen (Spector, 2000; Spector et al., 1999).

2.2.2. Eigenschaften des Gehörs

Nachdem im letzten Abschnitt der Aufbau und die Funktion des Gehörs erläutert wurden, sollen hier die wichtigsten Auswirkungen, Fähigkeiten und Einschränkungen dargestellt werden. Dabei muß berücksichtigt werden, daß die im folgenden genannten Zusammenhänge teilweise stark von der Art des Lautreizes abhängen. Eine ausführliche Darstellung findet sich beispielsweise bei Zwicker (1982).

Wahrnehmbarer Frequenz- und Amplitudenbereich

Die Bandbreite der wahrnehmbaren Frequenzen ist im wesentlichen durch die Beschaffenheit der Basilarmembran auf etwa 16–18 000 Hz begrenzt. Die Empfindlichkeit des menschlichen Gehörs ist durch die Filtereigenschaften des Außen- und des Mittelohres sowie durch die geringere Anzahl der Sinneszellen für die Wahrnehmung sehr tiefer und sehr hoher Frequenzen stark frequenzabhängig mit einem Maximum bei etwa 4 kHz. Innerhalb einer Bandbreite von 700–7 000 Hz ist die Empfindlichkeit des Gehörs allerdings fast konstant. Die Hörschwelle ändert sich bei einem Schalldruckpegel von 0 dB SPL in diesem Frequenzbereich nur um etwa 3 dB. Innerhalb dieses Bandes können Geräusche über einen Dynamikbereich von mehr als 100 dB wahrgenommen werden.

Der Zusammenhang zwischen der Empfindung *Lautheit* N ($[N]$=Sone) und dem physikalischen Schalldruck P ist sehr komplex. Er kann durch ein Potenzgesetz

$$N \propto P^k \tag{2.1}$$

angenähert werden, wobei der Exponent je nach Art des Geräusches und der Versuchsperson zwischen 0,2 und 0,5 liegt (Schroeder et al., 1979; Warren, 1970; Zwicker, 1982).

[11] afferent: zum Zentralnervensystem (ZNS) hinführend.
[12] efferent: vom ZNS herausführend.

Phase

Die Haarzellen feuern gewöhnlich kurz vor der maximalen Auslenkung der Stereozilien nach außen (Warren, 1982). Wegen ihrer Refraktärzeit von 1–3 ms können sie dies bis zu Frequenzen von maximal 1 kHz leisten. Auf den Stimulus synchronisierte Histogramme der Feuerzeitpunkte *(Post-Stimulus Time Histograms)* geben annähernd das Bild der gleichgerichteten Basilarmembranschwingung wieder (O'Shaughnessy, 1987). Bei höheren Frequenzen werden von einzelnen Neuronen einzelne Halbwellen ausgelassen, die Summe der Feuerzeitpunkte eines Ensembles von sich an einer bestimmten Stelle der Basilarmembran befindlichen Haarzellen erhält aber insgesamt dieses phasensynchrone Muster. Ab etwa 4 kHz gerät allerdings der *Jitter*[13] der Haarzellen in die Größenordnung der Anregungsfrequenz, so daß das Muster verschmiert wird und keine Information über die Phasenlage mehr aufweist (Warren, 1982).

Das Gehör ist relativ unsensibel gegenüber Phasenverzerrungen, solange die Gruppenlaufzeitverzögerungen einige Millisekunden nicht überschreiten (Schroeder, 1975). Bei zeitinvarianten Verzerrungen wäre eine bessere Empfindlichkeit gewöhnlich auch nicht sinnvoll, da Phasenverschiebungen beispielsweise durch Wandreflexionen (Hall) allenthalben auftreten. Nichtsdestoweniger können Phasenverzerrungen den *Klang* von Sprachsignalen durchaus beeinflussen. Bei zeitvarianten Verzerrungen liegen die Dinge etwas anders. So kann die Änderung der zeitlichen Struktur synthetischer vokalartige Reize zur Wahrnehmung von Formantverschiebungen[14] führen, ohne daß das Amplitudenspektrum tatsächlich geändert wurde (Schroeder und Strube, 1986).

Verdeckungseffekte

Weiterhin treten im Gehör Verdeckungseffekte auf: Ein kurz nach einem lauten Signal folgendes sehr leises Signal, welches für sich genommen wohl wahrgenommen würde, wird von dem lauten Signal maskiert (zeitliche Nachverdeckung). Dies gilt interessanterweise auch umgekehrt: Ein früheres leises Signal kann durch ein späteres lautes Signal ebenfalls maskiert werden (Zwicker, 1982). Dies wirkt wie ein akausaler Effekt, hat aber seine Ursache einfach in der nichtinstantanen Verarbeitung des Reizes. Auch im Frequenzbereich werden leise Komponenten eines komplexen Geräusches von benachbarten lauteren Komponenten maskiert (Frequenzverdeckung) (Zwicker, 1982).

Alle Arten von Verdeckungseffekten kann man sich beispielsweise für verlustbehaftete Audiokodierungstechniken zunutze machen, um zu erreichen, daß das rekonstruierte Signal bei geringerem Informationsgehalt den gleichen perzeptiven Eindruck hinterläßt wie das Originalsignal (DCC, MiniDisc, MP3. Für eine Übersicht siehe (Brandenburg und Bosi, 1997), für Vergleiche (Soulodre et al., 1998)).

Im Zusammenhang mit den beschriebenen Frequenzverdeckungseffekten findet man ein anderes interessantes Phänomen, nämlich das der *Frequenzgruppen* (*critical*

[13] Jitter: Statistische Abweichungen von der Synchronizität.
[14] Formanten heißen die Resonanzen des Stimmkanals.

Bands. Man stellt fest, daß die beschriebene gegenseitige Beeinflussung der Wahrnehmung benachbarter Frequenzanteile nur innerhalb gewisser Bandbreiten stattfindet. Sind die Signalkomponenten so weit voneinander entfernt, daß sie eine kritische Bandbreite überschreiten, endet die Beeinflussung relativ abrupt. Die Breite dieser Frequenzgruppen ist frequenzabhängig und definieret eine perzeptive Frequenzskala, die Barkskala[15], welche auf der fortlaufenden Numerierung der kritischen Bänder basiert. Es wurden verschiedene analytische Approximationen für diese Skala gegeben, z. B. von Schroeder (1977)

$$z = 7 \operatorname{arcsinh}\left(\frac{f}{650\,\mathrm{Hz}}\right), \tag{2.2}$$

von Zwicker und Terhardt (1980)

$$z = 13 \arctan\left(\frac{7\,600 f}{\mathrm{Hz}}\right) + 3{,}5 \arctan\left(\frac{f}{7\,500\,\mathrm{Hz}}\right) \tag{2.3}$$

und von Traunmüller und Lacerda (1987)

$$z = \frac{26{,}81 f}{1\,960\mathrm{Hz} + f} - 0{,}53. \tag{2.4}$$

Dabei wird die Frequenz f in Hertz angegeben, z ist die resultierende Tonheit[16] in Bark. Vgl. zu den verschiedenen analytischen Ausdrücken auch (Traunmüller, 1988).

Wahrnehmung von Sprachlauten

Nicht der ganze dem Ohr zugängliche Frequenzumfang wird für die Wahrnehmung von Sprachlauten benötigt. Experimente haben gezeigt, daß die wichtigsten zum Sprachverständnis nötigen Beiträge im Bereich von 200–5 600 Hz liegen (Pavlovic und Studebaker, 1984). Man kann dies täglich bei der Benutzung eines Telefons neu erfahren. Dieses Frequenzband entspricht ebenfalls etwa dem Bereich der höchsten Empfindlichkeit des Gehörs (s. Seite 9) und dem Bereich der höchsten Intensitäten des Sprachsignals.

Innerhalb dieses Frequenzbereiches führen Störungen durch additives Rauschen oder Filterungen zu typischen Erkennungsfehlern: Innerhalb des Systems der Laute (Anhang B) werden benachbarte Laute verwechselt. Aus den spektralen Eigenschaften dieser Laute läßt sich schließen, daß die Stimmhaftigkeit eines Lautes anhand seiner harmonischen Struktur erkannt wird, welche in tiefen Frequenzbereichen besonders prominent ist (O'Shaughnessy, 1987). Auch die Hinweise auf die Artikulationsart[17] liegen offenbar bei tiefen Frequenzen. Der Artikulationsort hingegen scheint im wesentlichen anhand von Merkmalen erkannt zu werden, die oberhalb von 1 kHz, insbesondere in der Gegend des zweiten Formanten liegen (O'Shaughnessy, 1987).

[15] nach H. G. Barkhausen, dt. Physiker (1881–1956).

[16] Tonheit: Wahrnehmungsgröße für die Frequenz, wird in Bark oder Mel ($m = 2\,595 \lg\left(1 + \frac{f}{700\ \mathrm{Hz}}\right)$ (O'Shaughnessy, 1987)) angegeben.

[17] Auf die Begriffe *Artikulationsart* und *-ort* wird auf Seite 12 noch näher eingegeangen.

2.3. Lautspezifische Merkmale

Im letzten Abschnitt wurde bereits angesprochen, daß die Wahrnehmung bestimmter phonemspezifischer Eigenschaften (Artikulationsart und -ort, etc.) an spezifische Signaleigenschaften gebunden ist. Es liegt daher nahe, in der auditorischen Verarbeitung Detektoren für diese Eigenschaften zu vermuten bzw. solche für die automatische Spracherkennung zu implementieren. Welche Eigenschaften das sind, soll im folgenden kurz skizziert werden.

2.3.1. Vokale

Vokale zeichnen sich durch stimmhafte Anregung (annähernd periodisches Signal), starke und (bei nahezu stabiler Vokaltraktkonfiguration) relativ schmalbandige Formanten sowie eine gewisse Dauer aus. Zur Diskriminierung stationärer Vokale ist die Kenntnis der Lage der ersten zwei oder drei Formanten ausreichend (Klein et al., 1970). So zeichnen sich die vorderen Vokale durch relativ hohe Energie im zweiten und dritten Formaten aus, während die Formantmuster der zentralen[18] Vokale gut getrennte Formanten ausgeglichener Energie aufweisen und die Energie der Formanten oberhalb des zweiten bei den hinteren Vokalen stark abfällt. Darüberhinaus werden anhand der Anregungsfrequenz Annahmen über die Lage der Formanten getroffen. Höhere Anregungsfrequenzen lassen auf einen weiblichen Sprecher oder ein Kind schließen, so daß wegen der gegenüber Männern kürzeren Vokaltrakte eine Verschiebung der Formantfrequenzen nach oben erwartet wird. Um also beispielsweise Verwechselungen von /i/ und /ɪ/ zu vermeiden, muß auch die Anregungsfrequenz bekannt sein (O'Shaughnessy, 1987).

Innerhalb eines natürlichen Kontextes werden Vokale mit ihren Nachbarlauten *koartikuliert*[19] und selbst im Silbenkern nicht so prägnant artikuliert wie im Falle isolierter Vokale, d. h. die Formanten erreichen normalerweise nicht ihre Endpositionen (O'Shaughnessy, 1987).

Den Vokalen zuordnen lassen sich auch die *Diphthonge* (Doppellaute), die sich dadurch charakterisieren lassen, daß sie in der Nähe der Artikulatorkonfiguration eines Vokals starten und stetig in die Konfiguration eines anderen Vokals übergehen.

2.3.2. Konsonanten

Konsonanten werden normalerweise nach Artikulationsart und -ort und als stimmhaft/stimmlos klassifiziert (vgl. Anhang B).

Die Artikulationsart, welche es erlaubt, die Lautklassen Vokal (inklusive Liquid), Nasal, Plosiv und Frikativ zu bestimmen, kann an spezifischen spektralen und Amplitudenmerkmalen unterschieden werden.

[18] Damit ist hier die Zungenposition bei der Artikulation des Vokals gemeint. Vgl. auch Abbildung B.1.
[19] Koartikulation: Gegenseitige Beeinflussung der Artikulation zeitlich benachbarter Laute.

Nasale werden stimmhaft artikuliert. Die Mundhöhle wird an einem Punkt vollständig geschlossen und das Velum abgesenkt, um den Luftstrom durch den Nasaltrakt und die Nasenlöcher entweichen zu lassen. Dadurch, daß die Mundhöhle trotz des Verschlusses akustisch an den Rachenraum gekoppelt bleibt, wirken ihre Resonanzfrequenzen als Antiresonanzen in der Vokaltraktübertragungsfunktion. Nasale zeichnen sich gegenüber Vokalen durch eine plötzliche geringere Amplitude, breitere (stärker gedämpfte) Formanten und eine Energiekonzentration bei Frequenzen um 250 Hz aus (O'Shaughnessy, 1987).

Signale mit relativ langer Dauer und einer Energiekonzentration bei hohen Frequenzen werden als Frikative erkannt. Stimmlose Frikative werden gebildet, indem der Vokaltrakt an einer Stelle verengt wird, so daß dort der Luftstrom turbulent wird. Auch hier wirken die Resonanzen der Höhle vor der Einschnürung als Antiresonanzen. Stimmhafte Frikative gleichen den entsprechenden stimmlosen Frikativen bis auf die zusätzliche stimmhafte Anregung.

Plosive unterbrechen den Signalfluß kurzzeitig oder beginnen ein Signal mit einem plötzlichen Rauschimpuls. Sie werden durch einen vollständigen Verschluß des Vokaltraktes und anschließende plötzliche Druckfreisetzung artikuliert. Plosive können naturgemäß nur im Kontext mit anderen Lauten erkannt werden. Wichtig ist hierbei die Verschlußdauer und (im vokalischen Kontext) der Verlauf der anschließenden Formanten und die spektrale Lage des assoziierten Rauschimpulses. Stimmlose Plosive unterscheiden sich von ihren stimmhaften Gegenstücken im wesentlichen durch eine kurze Pause (die behaucht sein kann) nach der Verschlußöffnung, nach der erst die Stimmlippenschwingung wieder einsetzt.

2.3.3. Relevanz von Vokalen und Konsonanten für die Darstellung des übermittelten Textes

Angesichts der großen Energie und Dauer, welche die vokalischen Anteile des akustischen Signals enthalten, sind die Vokale für den übermittelten Text von überraschend geringer Bedeutung. Betrachtet man beispielsweise den Titel dieses Abschnittes, so kann dieser (in seiner Textform) anhand der enthaltenen Konsonanten leicht rekonstruiert werden (1), wohingegen dies nur aus den enthaltenen Vokalen schwerlich gelingt (2):

1. `R_l_v_nz v_n V_k_l_n _nd K_ns_n_nt_n f_r d__ D_rst_ll_ng d_s _b_rm_tt_lt_n T_xt_s`

2. `_e_e_a__ _o_ _o_a_e_ u__ _o__o_a__e_ _ü_ _ie _a___e__u__ _e_ ü_e__i__e__e_ _e__e_`

Dies führt so weit, daß beispielsweise in der hebräischen Schrift Vokale nicht bezeichnet werden – der Leser muß die Vokale anhand ihm bekannter Vokalismusregeln ergänzen (Miller, 1996).

In der Spracherkennung sind Vokale allerdings von großer Bedeutung (Cole et al., 1996). Beispielsweise sind die Vokale in vielen Sprachen die silbenbildenden Laute und

Träger der prosodischen Merkmale. Bei der Anwesenheit mehrerer Sprecher tragen die Vokale wesentlich zur Trennbarkeit der Äußerungen der verschiedenen Sprecher bei. Ähnliches gilt für andere Störgeräusche.

2.4. Merkmalsbildung

Die Ausführungen der letzten Abschnitte legen bereits nahe, daß die direkte Klassifizierung von Lauten auf dem Zeitsignal keine adäquate Methode für die automatische Spracherkennung ist. Das Sprachsignal enthält zuviel Redundanz und Information, die auch dem menschlichen Hörer nicht zugänglich ist bzw. nicht von ihm genutzt wird. Außerdem variiert das Signal auch für gleichlautende Äußerungen stark von Fall zu Fall.

Der erste Schritt zur automatischen Spracherkennung besteht daher in einer Analyse und Aufbereitung des Sprachsignals zu Merkmals- oder *Feature*-Vektoren, die die Redundanz und Variabilität des Signals reduzieren, ohne dabei für die Lautklassifikation wichtige Merkmale zu eliminieren.

Wie ebenfalls oben schon dargelegt, kann das Sprachsignal innerhalb kurzer Zeiten als annähernd stationär betrachtet werden. Nach einer geeigneten Aussteuerung des Signals, möglicherweise unter Verwendung einer *AGC (Automatic Gain Control)* und einer Höhenanhebung, um den mittleren spektralen Abfall des Sprachsignals von etwa 6 dB/Oktave[20] auszugleichen, ist der nächste Schritt zur Analyse daher eine Zerlegung des Sprachsignals in kurze Abschnitte.

2.4.1. Fensterung

An die Zerlegung eines Sprachsignals in einen Strom von gefensterten Signalstücken werden verschiedene Anforderungen gestellt. Zunächst sollten die Fenster eine zeitlich begrenzte Länge besitzen, d. h. für Fensterfunktionen $w(t) : \mathcal{R} \rightarrow \mathcal{R}$ soll gelten: $\exists T_1, T_2 \in \mathcal{R}$, $-\infty < T_1 < T_2 < \infty$ mit

$$w(t) = 0 \text{ für } t < T_1 \text{ und } t > T_2 \tag{2.5}$$

Weiterhin sollen alle Anteile des Signals gleich gewichtet bleiben:

$$\sum_{n=-\infty}^{\infty} w_n(t)x(t) = x(t), \tag{2.6}$$

d. h., das Signal wird durch Addition der gefensterten Signalstücke wieder identisch rekonstruiert. Daraus folgt insbesondere, daß die Fensterfunktionen beschränkt sein

[20] Dieser spektrale Abfall tritt nur bei stimmhaften Signalteilen auf – stimmlose Signalteile haben eher ein flaches Spektrum, so daß eine einheitliche Präemphase nur ein Kompromiß ist.

müssen und die Summe der Fenster nirgends verschwinden darf. Für alle t gilt also:

$$\exists c \in \mathcal{R} : w_n(t) < c \tag{2.7}$$

$$\sum_{n=-\infty}^{\infty} w_n(t) \neq 0 \tag{2.8}$$

$$\text{und mit 2.6:} \quad \sum_{n=-\infty}^{\infty} w_n(t) = 1 \tag{2.9}$$

In der Praxis wird zudem gewöhnlich gefordert, daß Fensterfunktionen nicht-negativ und symmetrisch sein sollen. Außerdem wird normalerweise die gleiche Fensterfunktion für alle Fenster verwendet, und das Signal wird äquidistant gefenstert, d. h. für alle Zeiten t gilt:

$$w(t) \geq 0 \tag{2.10}$$

$$w(t) = w(-t) \tag{2.11}$$

$$w_n(t) = w(t - nT), \tag{2.12}$$

wobei T der Fenstervorschub ist.

Fensterlänge

Um die annähernde Stationarität des Sprachsignals auszunutzen, darf ein Zeitfenster nicht zu lang sein. Andererseits sollte ein Fenster eine gewisse Mindestlänge aufweisen, damit mindestens eine komplette (Quasi-)Periode des zu analysierenden Signals im Fenster liegt. Ein typischer Kompromiß besteht darin, die Fenster eher etwas länger (etwa 20 ms) zu wählen, sich dafür aber überlappen zu lassen, indem man einen geringeren Vorschub (etwa 5 ms) wählt. Um jetzt die Forderungen der Gleichungen 2.9 und 2.12 zu erfüllen, muß die Fensterlänge ein ganzzahliges Vielfaches des Vorschubs betragen. Die Wahl einer bestimmten Fensterform kann die möglichen Verhältnisse von Fenstervorschub und Fensterlänge weiter eingeschränken, wenn die Bedingung 2.9 eingehalten werden soll.

Spektrale Eigenschaften

In den vorangegangenen Abschnitten war häufiger von spektralen Eigenschaften die Rede, die bestimmte Laute von anderen unterscheidbar machen. Um solche Eigenschaften sichtbar zu machen, werden die gefensterten Signalstücke in der weiteren Verarbeitung oft fouriertransformiert. In diesem Zusammenhang sind die Eigenschaften des Spektrums der Fensterfunktionen interessant.[21]

[21] Nach den gemachten Annahmen ist jede Fensterfunktion ein Element des Raumes der quadratintegrablen Funktionen (L_2). Da der L_2 sein eigener Dualraum ist, besitzt jede Fensterfunktion eine Fouriertransformierte.

Die Forderung nach zeitbeschränkten Fenstern (Gleichung 2.5) erzwingt ein unendliches Spektrum des Fensters. Die Forderung eines nicht-negativen Fensters (Gleichung 2.10) bewirkt, daß das Fenster die Charakteristik eines Tiefpaßes besitzt. Die Multiplikation des Originalsignals führt nach dem *Faltungssatz* zu einer Faltung des Signalspektrums mit dem Spektrum des Fensters, so daß jede Harmonische des Signals mit dem Fensterspektrum »verschmiert« wird. Man wünscht sich daher ein möglichst gut lokalisiertes Fensterspektrum mit steilem Abfall zu höheren Frequenzen. Auskunft über das asymptotische Verhalten des Fensterspektrums gibt das RIEMANNsche Lemma:

Sei $f(t)$ eine zeitbeschränkte integrierbare Funktion. Sei $F(\omega) := \int_a^b f(t)e^{-i\omega t}\,\mathrm{d}t$ ihre Fouriertransformierte. Dann gilt

$$F(\omega) = \mathcal{O}\left(\frac{1}{\omega}\right) \text{ für } |\omega| \to \infty. \tag{2.13}$$

Sei nun $f(t)$ eine n-mal differenzierbare quadratintegrable Funktion. Dann ist das Spektrum der ersten Ableitung $i\omega F(\omega)$ und wegen des RIEMANNschen Lemmas von der Ordnung $\mathcal{O}\left(\frac{1}{\omega}\right)$. Also muß $F(\omega)$ wie $\frac{1}{\omega^2}$ abfallen. Durch wiederholte Anwendung ergibt sich

$$F(\omega) = \mathcal{O}\left(\frac{1}{\omega^{n+1}}\right). \tag{2.14}$$

Viele der gemachten Vorgaben stehen im Widerspruch zueinander, so daß jede Fensterfunktion nur einen Kompromiß darstellen kann. Optimal im Sinne einer maximalen Energiekonzentration innerhalb einer gegebenen Bandbreite sind die *Prolaten Sphäroidfunktionen* (Stearns und Hush, 1999). Das Zeit-Bandbreiten-Produkt wird durch *Gaußfunktionen* minimiert (die allerdings nicht zeitbegrenzt sind). Einen guten Kompromiß stellt das *Hann*-Fenster dar, welches durch eine Cosinusperiode realisiert wird. Um die Leistung des ersten Seitenbandes zu reduzieren, hat R. W. HAMMING dieses Fenster mit einem kleinen Rechtecksockel modifiziert, wodurch sich allerdings das asymptotische Verhalten des Fensters wieder verschlechtert. Dieses *Hamming*-Fenster ist in der Sprachsignalanalyse besonders verbreitet:

$$h(t) = \begin{cases} 0{,}54 - 0{,}46\cos\left(\pi\frac{t}{T}\right) & \text{für } -T < t < T \\ 0 \text{ sonst.} \end{cases} \tag{2.15}$$

Alle diese Cosinus-Fenster summieren sich wieder zu 1 auf, wenn (bei geeigneter Normierung) der Fenstervorschub ganzzahlige Bruchteile der halben Fensterlänge (Fensterlänge hier: $2T$) beträgt.

Details zu den genannten Fensterfunktionen findet man beispielsweise bei Stearns und Hush (1999) oder bei Schroeder (1990).

2.4.2. Parametrisierung

Nachdem das Sprachsignal gefenstert worden ist, ist es für die meisten Anwendungen sinnvoll und üblich, das Signal innerhalb der Fenster zu parametrisieren. Die dazu

eingesetzten Transformationen haben häufig die Form

$$y(n) = \int_{-\infty}^{\infty} \mathbf{T}x(t)w(t - nT)\,\mathrm{d}t \tag{2.16}$$

Auf diese Weise gewonnene Parameter können Kurzzeit-Mittelwert der Energie oder Amplitude sein (mit $\mathbf{T}x = \frac{1}{T}x^2$ oder $\mathbf{T}x = \frac{1}{T}|x|$, T: Fensterlänge), mittlere Nulldurchgangsraten (etwas lax: $\mathbf{T}x = \frac{1}{2T}|\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{sgn} x|$) oder Autokorrelationen (mit $\mathbf{T}x(t) = x(t)x(t-\tau)$).

Neben diesen zeitlichen Charakteristika sind insbesondere spektrale Parameter von besonderem Interesse. Solche Parameter können durch Operatoren der Form $\mathbf{T} = e^{-i\omega t}$ gewonnen werden. Durch diese Transformation erhält man für jedes gefensterte Signalsegment ein Spektrum. Ein *Spektrogramm* als Funktion der Zeit ergibt sich dann durch die zeitliche Aneinanderreihung der Spektren der einzelnen Signalsegmente. Dabei wird die Phaseninformation gewöhnlich verworfen, da sie wenig zum Sprachverständnis beiträgt.

Es hat sich in der Praxis außerdem durchgesetzt, die entstehenden Spektrogramme in ihrer Dynamik zu komprimieren. Dies stellt sich allgemein als vorteilhaft heraus und steht im Einklang mit der Funktionsweise des Gehörs. Üblich ist eine Potenzierung analog Gleichung 2.1 oder eine Logarithmierung. Besonders die Logarithmierung wird auch zur Datenreduktion eingesetzt: Wenn man einerseits davon ausgeht, daß die Bewegung der Artikulatoren langsamer vonstatten geht als die Anregung des Vokaltraktes durch die Glottisschwingungen (Abschnitt 2.1.1), und man andererseits annimmt, daß das Signal der Glottisschwingungen durch den Vokaltrakt gefiltert wird, dann kann das Gesamtsignal *entfaltet* werden, indem man das Produkt von Glottisspektrum und Vokaltraktspektrum durch Logarithmierung in eine Summe transformiert. Wendet man auf diese Summe eine weitere Fouriertransformation an,[22] so gibt deren *niederquefrenter* Anteil die Vokaltraktfilterung wieder, wohingegen der *hochquefrente* Anteil die Stimmlippenschwingungen repräsentiert. Durch eine *Tiefpaßlifterung* lassen sich dann die *cepstralen* Anteile der anregenden Schwingungen eliminieren.

Ein weiterer verbreiteter Vorverarbeitungsschritt besteht in der Anpassung der Spektren an eine gehörorientierte Frequenzskala, wie sie etwa durch die Gleichungen 2.4 oder 2.3 gegeben ist. Dies geschieht durch Mittelung über die Kanäle der FFT[23], etwa über die Bandbreite jeweils eines Barkkanals. Das resultierende Spektrogramm besitzt dann – je nach Abtastfrequenz – beispielsweise 19–21 Barkkanäle. Welche Frequenzskala nun genau verwendet wird, oder ob man auf Fourier-, LPC-[24], Filterbank- oder Cepstrumskoeffizienten arbeitet, ist von Fall zu Fall verschieden.

[22] Diese Transformation wird nach J. W. Tukey als *Cepstrum* bezeichnet. Die Entsprechung der Frequenz nannte er konsequenterweise *Quefrenz* und Filteroperationen auf dem Cepstrum *Lifterungen*.

[23] FFT: *Fast Fourier Transform*. FFT ist ein schneller Algorithmus zur Berechnung einer Diskreten Fouriertransformation (DFT).

[24] LPC: *Linear Predictive Coding*. Die LPC-Methode versucht, durch eine Linearkombination von n vergangenen Signalwerten auf den nächsten Signalwert zu schließen (Lineare Prädiktion). Die Anzahl n der Koeffizienten ist die Ordnung der Analyse. Die Methode fußt auf der Annahme eines annähernd stationären Signals. Vgl. hierzu (Markel und Grey, Jr., 1976).

Zusätzlich zu den beschriebenen zeitlichen und spektralen Parametern können auch Merkmale höherer Ordnung generiert und ergänzend oder ausschließlich zur Bildung von Merkmalsvektoren verwendet werden. Dazu gehören Parameter wie die Ausgabe eines Stimmhaft/Stimmlos-Klassifikators, eines Signal/Stille-Detektors oder eines Formant-Trackers. Diese Parameter verlagern etwas von der Kompetenz des nachfolgenden Laut-Klassifikators in die Vorverarbeitung. Solche Parameter sind allerdings nicht immer sicher zu bestimmen, und fehlerhafte Parameter dieser Art führen später naturgemäß fast zwangsläufig zu einer Fehlklassifikation der zugehörigen Laute.

2.5. Mustererkennung

Nachdem das Sprachsignal in der beschrieben Weise vorbereitet worden ist, stellt es sich als ein Strom von n-dimensionalen Merkmalsvektoren dar, wobei n gewöhnlich im Bereich von 10–30 anzusiedeln ist. Die Strom ist zeitlich zumeist äquidistant und hat eine Frequenz von typischerweise 200 Vektoren pro Sekunde.

Man könnte nun erwarten, daß die Vorverarbeitung eine brauchbare Trennung der Vektoren nach Lautklassen innerhalb des durch die Vektoren aufgespannten Raumes ermöglichen würde. Leider ist dies im allgemeinen nicht der Fall, nicht zuletzt, weil die Vektoren unter der Annahme der annähernden Stationarität des Sprachsignals gebildet wurden, andererseits aber ein großer Teil der im Sprachsignal enthaltenen Information in der Form dynamischer Eigenschaften codiert ist. Dies führt beispielsweise für lange Vokale zu einer recht guten Clusterung, Plosive aber können nur im Kontext mehrerer aufeinanderfolgender Vektoren erkannt werden. Um dies zu unterstützen, werden häufig sogenannte Delta-Merkmale (Differenzen zweier aufeinanderfolgender Merkmalsvektoren) gebildet. Auch Delta-Delta-Merkmale werden benutzt (Differenzen der Differenzenvektoren) (Rabiner und Juang, 1993). Diese Parameter werden zu einem neuen, höherdimensionalen Vektor kombiniert, der dann als neuer Merkmalsvektor dem Klassifikator zugeführt wird.

Eine andere Methode zur Berücksichtigung der dynamischen Eigenschaften des Sprachsignals ist die Arbeit auf mehreren aufeinanderfolgenden Merkmalsvektoren (Lang et al., 1990). Dies schließt die gleichzeitige Verwendung der eben beschriebenen Deltamerkmale nicht aus.

2.5.1. Segmentierung

Um den Kontext der einzelnen Merkmalsvektoren berücksichtigen zu können, kann der Strom der Vektoren in geeignete Einheiten segmentiert werden. Die Wahl der Segmentierung hängt von der Aufgabenstellung des Erkennungssystems ab. So ist die natürliche Segmentierung von Einzelworterkennungssystemen durch die Wortgrenzen gegeben, die sich durch Pausen zwischen Wörtern relativ gut bestimmen lassen. Einzelworterkenner werden häufig für Kontroll- und Steuersysteme eingesetzt, die einen

relativ beschränkten Wortschatz benötigen. Dadurch ist die Anzahl der zu klassifizierenden Äußerungen relativ gering.

Systeme zur Erkennung fließender Sprache sind häufig für einen größeren Wortschatz ausgelegt, was die Klassifikation erschwert. Außerdem treten in fließender Sprache keine deutlich erkennbaren Wortgrenzen auf. Deswegen arbeiten diese Systeme gewöhnlich auf Subworteinheiten (SWE) wie Silben, Halbsilben, Triphonen, Diphonen oder Phonen. Silben lassen sich recht gut im Silbenkern durch ein Energiekriterium segmentieren. Hier sind auch die Koartikulationseffekte im allgemeinen geringer als am Silbenrand. Allerdings existieren beispielsweise im Deutschen oder im Englischen tausende verschiedener Silben. Auf der anderen Seite weist eine phonweise Segmentierung die geringste Anzahl von Klassen auf (im Englischen etwa 40). Diese sind aber schwieriger zu separieren und zeigen deutliche Koartikulationseffekte. Die übrigen genannten Segmentierungen stellen Kompromisse zwischen diesen beiden Extremen dar. Darüberhinaus ist auch eine selbstorganisierte Segmentierung aufgrund von Ähnlichkeiten von – und der Dynamik zwischen – aufeinanderfolgenden Merkmalsvektoren in SWE möglich, die nicht notwendigerweise einer der genannten linguistischen Kategorien entsprechen müssen (Behme und Brandt, 1993).

Auch für das Training eines Spracherkennungssystems ist eine – wie auch immer geartete – Segmentierung verbreitet. Diese basiert letztlich immer auf einem *Labeling*[25] der einzelnen Segmente der Trainings- oder Vergleichsmustermenge durch einen Menschen. Weitere Trainingsdaten können automatisch oder halbautomatisch durch ein existierendes Spracherkennungssystem gelabelt werden. In der Anwendung lösen viele Spracherkennungssysteme das Problem der Segmentierung, indem sie diese nicht explizit vornehmen, sondern auf einem größeren Zeitfenster fester Länge (Lang et al., 1990) oder gleich auf ganzen Sätzen arbeiten.

Ein Segmentlabeling kann auch vollständig vermieden werden bzw. vollständig automatisch generiert werden: Angenommen, die Erkennung soll auf der Basis geeigneter SWE stattfinden. Eine ausreichende Zahl im Wortlaut bekannter Trainingsäußerungen stehe zur Verfügung, sowie ein Lexikon, mit dessen Hilfe der Wortlaut der Äußerungen in die gegebenen SWE zerlegt werden kann. Dann kann eine Aneinanderreihung der »richtigen«, also den den SWE des Lexikons entsprechenden HMMs,[26] zur Modellierung der bekannten Äußerungen verwendet werden. Die Äußerungen werden in eine Anzahl Abschnitte gleicher Länge zerlegt, die der Anzahl der SWE in der Äußerung entspricht. Jedem HMM werden die zugehörigen Segmente zugewiesen. Im Prinzip (auf Details soll hier nicht eingegangen werden) wird nun jedes HMM mit den entsprechenden Segmenten trainiert. Daraufhin kann die HMM-Kette zur Neusegmentierung der Äußerungen verwendet und das Verfahren bis zur Konvergenz iteriert werden. Da die jeweiligen HMMs durch das Lexikon mit den zugehörigen SWE-Lautungen assoziiert sind, kann so im Bedarfsfall gleichzeitig ein Labeling der

[25] Labeling: Bezeichung eines Segmentes mit seinem Lautwert.

[26] HMM: *Hidden-Markov-Modell*. Sollte der geneigte Leser mit der Funktionsweise von HMMs noch nicht vertraut sein, möge er diesen Absatz überspringen oder sich zunächst in Abschnitt 2.5.3 informieren.

Äußerungen gewonnen werden. Details zu dem beschriebenen Verfahren findet man in (Rabiner und Juang, 1993).

2.5.2. Einzelworterkennung durch DTW

DTW steht für *Dynamic Time Warping* (Sakoe und Chiba, 1978) und ist ein *Pattern-Matching*-Algorithmus aus der Gruppe der *Dynamischen Programmierung* (Bronstein und Semendjajew, 1982). Die Methode macht den Ansatz, daß gleichlautende Äußerungen sich voneinander im wesentlichen durch die Sprechgeschwindigkeit unterscheiden und ihre Merkmalsvektorenströme daher durch eine nichtlineare Verzerrung der Zeitachsen einander angeglichen werden können. Der Algorithmus findet diejenige Verzerrung, die bezüglich eines Abstandsmaßes[27] optimal ist. Der Erkennungsprozeß besteht darin, daß eine zu erkennende Äußerung an die Prototypen der Äußerungen des Vergleichswortschatzes angeglichen wird. Im einfachsten Falle gilt diejenige Äußerung als erkannt, zu der der verbleibende (akkumulierte) Abstand *(DTW-Abstand)* am geringsten ist. Das genaue Verfahren ist in Anhang A.1 erklärt.

Diese Methode kommt ohne Training aus, wird aber – obwohl der Algorithmus auch auf die Erkennung fließender Sprache ausgedehnt werden kann – trotz guter Erkennungsraten heute kaum noch zur Spracherkennung verwendet, da der Rechenaufwand bei der Erkennung weit größer ist, als etwa bei dem im folgenden Abschnitt beschriebenen Verfahren. In dieser Arbeit wird der Algorithmus zur Einzelworterkennung und zur Berechnung eines Ähnlichkeitsmaßes zwischen Sprechern (Abschnitt 4.4.1) eingesetzt.

2.5.3. Erkennung fließender Sprache durch Hidden-Markov-Modelle

Bei der im folgenden beschriebenen Methode macht man die Annahme, daß bei der Gesamtheit möglicher Folgen von Merkmalsvektoren die Wahrscheinlichkeit für das Auftreten eines bestimmten Merkmalsvektors in einer dieser Folgen allein durch den unmittelbar zuvor beobachteten Merkmalsvektor gegeben ist. Aufgrund dieser Annahme werden für die in Betracht kommenden Äußerungen statistische Modelle gebildet. Einem Strom von Merkmalsvektoren werden auf der Basis dieser Modelle jeweils Wahrscheinlichkeiten zugeordnet. Die wahrscheinlichste Äußerung wird als erkannt angenommen. Derzeit sind Spracherkennungssysteme auf der Basis von Hidden-Markov-Modellen (HMM) die erfolgreichsten und verbreitetsten Systeme. Daher soll im folgenden etwas näher auf das Prinzip eingegangen werden.

[27] Dieser Abstand ist gewöhnlich kein Abstand im mathematischen Sinne, sondern ein Fehlermaß wie die mittlere quadratische Abweichung oder die Korrelation (das Skalarprodukt) der verglichenen Merkmalsvektoren.

Markov-Ketten

Unter einem (zeitdiskreten) Markov-Prozeß oder einer Markov-Kette (erster Ordnung) versteht man einen stochastischen Prozeß, dessen zukünftige Erscheinung allein von seinem gegenwärtigen Zustand bestimmt wird. Eine solche (endliche) Markov-Kette besitzt eine Anzahl N von Zuständen, Anfangswahrscheinlichkeiten $\pi_i = P(q_1 = i)$ zur Zeit 1 den Zustand i eingenommen zu haben und Übergangswahrscheinlichkeiten $a_{ij} = P(q_{t+1} = j | q_t = i)$ vom Zustand i in den Zustand j zu wechseln. Ein beliebtes Beispiel ist das folgende Wettersystem. Das Wetter möge die drei Zustände 1=*Regen*, 2=*Bewölkt* und 3=*Sonnig* annehmen können. Es gelten folgende Übergangswahrscheinlichkeiten:

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} 0{,}4 & 0{,}3 & 0{,}3 \\ 0{,}2 & 0{,}6 & 0{,}2 \\ 0{,}1 & 0{,}1 & 0{,}8 \end{pmatrix} \tag{2.17}$$

Damit ist die Wahrscheinlichkeit, daß es morgen regnet (Zustand 1), wenn heute die Sonne scheint (Zustand 3) $a_{31} = 0{,}1$.

Hidden-Markov-Prozesse

In obigem Beispiel ist der Zustand, in dem sich der Markov-Prozeß befindet, direkt beobachtbar. Ist hingegen an den Zustand selbst ein stochastischer Prozeß geknüpft, dessen Ergebnis die einzige Observable ist, spricht man von einem Hidden-Markov-Prozeß. Ein Beispiel hierfür sei der folgende Prozeß: Ein Mensch würfelt hinter einem Vorhang mit einer Anzahl Würfel, die sich aus einem Markov-Prozeß bestimmt. Die einzige Observable sei die Gesamtzahl der geworfenen Augen. Lautet das Ergebnis beispielsweise 4, so ist unbekannt, ob er mit einem Würfel eine 4 oder beispielsweise mit zwei Würfeln eine 1 und eine 3 geworfen hat. Um diesen Prozeß zu beschreiben, benötigt man zusätzlich zu den Parametern des einfachen Markov-Prozesses die Emissionswahrscheinlichkeiten $b_j(k) = P(o_t = v_k | q_t = j)$, mit denen im Zustand j das Symbol v_k emittiert wird. (o_t ist hier das beobachtete Symbol zur Zeit t).

Werden HMMs in der automatischen Spracherkennung eingesetzt, so besteht die Beobachtungssequenz zunächst aus einer Folge von Merkmalsvektoren. Jeder Zustand j eines HMM kann – entsprechend den zugehörigen Emissionswahrscheinlichkeiten $b_j(k)$ – Merkmalsvektoren v_k bzw. deren Index emittieren. Da die $b_j(k)$ zunächst eine diskrete Verteilung darstellen, kann man den Raum der Merkmalsvektoren durch eine Vektorquantisierung[28] in eine *Voronoi*-Zerlegung mit beispielsweise 64–256 Elementen diskretisieren.[29] Dadurch entsteht ein Codebuch, also eine indizierte Menge der typischen, jeweils einen diskreten Bereich des Merkmalsraumes repräsentierenden Vektoren. Eine klassische Methode hierfür ist der *LBG*-Algorithmus (benannt nach seinen Erfindern Linde, Buzo und Grey (Buzo et al., 1980; Linde et al., 1980)). Die

[28] Zur Vektorquantisierung findet man bei Makhoul et al. (1985) eine schöne Darstellung.

[29] Auf die heute zumeist verwendeten kontinuierlichen oder semikontinuierlichen Emissionswahrscheinlichkeiten wird auf Seite 23 noch näher eingegangen.

$b_j(k)$ stellen dann die Wahrscheinlichkeiten dar, im Zustand j einen durch den Codebuchvektor k repräsentierten Merkmalsvektor zu beobachten. Die einzelnen Zustände eines HMM bezeichnen dann lautliche Untereinheiten, die im Verlauf einer Äußerung auftreten. Da die Laute einer Äußerung eine zeitliche Ordnung besitzen, ist die Struktur eines solchen HMM die eines Links-Rechts- oder Bakis-Modells (Jelinek, 1976), welche gleichfalls eine Ordnung besitzen: Man kann, von einem ausgezeichneten Anfangszustand startend, nur zu Zuständen höherer Ordnung gelangen, oder anders gesagt: Die Übergangswahrscheinlichkeiten a_{ij} sind gleich Null für $j < i$. In der automatischen Spracherkennung haben sich Modelle mit $a_{ij} = 0$ für $j > i + 2$ als geeignet und ausreichend flexibel erwiesen, d. h., es sind Übergänge von einem Zustand auf sich selbst, den nächsten und den übernächsten Zustand erlaubt, alle anderen Wechsel finden nicht statt.

Die Anwendung dieses Konzeptes auf die automatische Spracherkennung geschieht nun für den einfacheren Fall einer Einzelworterkennung mit begrenztem Wortschatz in solcher Weise, daß für jede in Frage kommende Äußerung w ein eigenes HMM gebildet wird, gegebenenfalls mit einer je nach Wortlänge verschiedenen Anzahl von Zuständen und mit Parametern $\pi_i^w = 1$ für $i = 1$ (den Startzustand) und $\pi_i^w = 0$ für alle anderen Zustände sowie für das jeweilige Wort geeigneten $\mathbf{A}^w$ und $\mathbf{B}^w$ (der Matrix der $b_j^w(k)$), wobei der Index w das jeweilige Wort bezeichnet. Für jede hereinkommende Äußerung wird für jedes Wortmodell die höchste Wahrscheinlichkeit berechnet, mit der es die jeweilige Beobachtungssequenz emittiert haben kann. Im einfachsten Fall wird dasjenige Wort erkannt, welches durch das HMM mit dem höchsten Ergebnis modelliert wird.

Für das Training von HMM-Erkennern existieren zahlreiche Beschreibungen (z. B. (Rabiner, 1989)). Die Problematik soll hier daher nur kurz angerissen werden. Bei der Beschreibung eines stochastischen Prozesses durch HMMs steht man vor drei Fragestellungen:

1. Gegeben sei die Beobachtungssequenz $\mathbf{o} = (o_0, \ldots, o_T)$ und ein HMM $\lambda(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ mit $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. Wie berechnet man die Wahrscheinlichkeit $P(\mathbf{o}|\lambda)$ für das Auftreten der Beobachtungssequenz bei gegebenem Modell?

2. Gegeben sei die Beobachtungssequenz $\mathbf{o} = (o_0, \ldots, o_T)$ und ein HMM $\lambda(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$. Welches ist dann diejenige Zustandssequenz $\mathbf{q} = (q_1, \ldots, q_T)$, welche die Beobachtung – für ein geeignetes Optimierungskriterium – in optimaler Weise beschreibt?

3. Wie müssen die Parameter des Modells $\lambda(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ gewählt werden, damit $P(\mathbf{o}|\lambda)$ maximiert wird?

Für alle diese Probleme existieren effektive Algorithmen. Klassischerweise wird das Problem 1 durch den sogenannten *Forward-Backward*-Algorithmus gelößt, Problem 2 durch den *Viterbi*-Algorithmus und Problem 3 durch den Baum-Welch-Algorithmus. Auch auf diese Algorithmen soll hier nicht weiter eingegangen werden. Für Details vgl. z. B. wiederum (Rabiner, 1989).

Die diskreten Emissionswahrscheinlichkeiten $\mathbf{B}$ erfordern eine Quantisierung des Merkmalsraumes. Dabei geht notwendigerweise ein Teil der in den Merkmalsvektoren enthaltenen Information verloren. Ein naheliegendes alternatives Verfahren wäre die Zuordnung einer kontinuierlichen Verteilungsfunktion $b_j(\mathbf{x})$ zu jedem Zustand j, welche direkt die Wahrscheinlichkeit beschreibt, den Merkmalsvektor $\mathbf{x}$ im Zustand j zu beobachten. Da die Schätzung dieser Wahrscheinlichkeiten wiederum unrealistische Datenmengen benötigen würde, wird statt dessen gemeinhin eine parametrisierte Verteilung wie die Gaußverteilung angenommen, für die nur der Mittelwert und die Varianz zu schätzen ist, oder Mischungen (Linearkombinationen) von Verteilungen (Rabiner und Juang, 1993). Ein rechenzeitfreundlicher Kompromiß zwischen dem eben beschriebenen kontinuierlichen und dem diskreten Fall ist die Abdeckung des Merkmalsraumes mit für alle Zustände identischen (Gaussschen) Wahrscheinlichkeitsverteilungen, die ähnlich wie bei der Zerlegung des Raumes durch die Vektorquantisierung gefunden werden (Rabiner und Juang, 1993). Die Zustände emittieren dann wiederum die Parameter für Linearkombinationen der verschiedenen Gaußverteilungen. Dieser Ansatz wird häufig als *semikontinuierlich* bezeichnet.

Für den Fall eines Spracherkennungssystems für fließende Sprache und großen Wortschatz ist die Verwendung von Wortmodellen aus den in Abschnitt 2.5.1 beschriebenen Gründen ungeeignet. Statt dessen wählt man hier HMMs, die geeignetere SWE modellieren, wie beispielsweise Di- und Triphone.

Um nun einen ganzen Satz erkennen zu können, kann man naheliegenderweise nicht jede denkbare Aneinanderreihung der HMMs in Betracht ziehen. Statt dessen wird z. B. folgendermaßen vorgegangen (Rabiner und Juang, 1993): Die in Frage stehende Äußerung besitze die Länge T. Für jedes SWE-Modell und jeden Zeitpunkt $0 < t \leq T$ wird die höchste Wahrscheinlichkeit berechnet, mit der das gegebene Modell die zugehörige Merkmalsvektorfolge emittiert haben kann. Im nächsten Schritt wird, ausgehend vom frühest möglichen Startpunkt (welcher durch die Mindestlänge der von allen Modellen generierbaren Beobachtungssequenzen gegeben ist) bis zum Ende der Äußerung wiederum die nämliche Berechnung durchgeführt. Die sich ergebenden höchsten Wahrscheinlichkeiten werden mit denen des vorherigen Schrittes multipliziert (bzw. addiert, wenn man, wie in der Praxis üblich, mit logarithmierten Wahrscheinlichkeiten arbeitet). Das Verfahren wird fortgesetzt, bis eine vorher festgelegte Maximalzahl von Schritten (entsprechend einer maximalen SWE-Folgenlänge) erreicht ist. In jedem Schritt merkt man sich für jeden Zeitpunkt dasjenige Modell, welches mit der höchsten Wahrscheinlichkeit dorthin geführt hat. Zum Schluß wird die SWE-Folge zurückverfolgt, welche den Endpunkt der Äußerung mit der höchsten akkumulierten Wahrscheinlichkeit erreicht. Auch dieses Verfahren ist eine Methode der Dynamischen Programmierung und ähnelt deutlich dem DTW-Algorithmus.

2.6. Höhere Verarbeitung

Das in Abschnitt 2.5.3 beschriebene Verfahren kann durch die Berücksichtigung mehrerer alternativer Kandidaten (SWE-Folgen) erweitert werden. Häufig werden auch

noch mehr oder weniger einfache Grammatiken hinzugefügt, die beschreiben, wie wahrscheinlich es in der jeweiligen Sprache ist, daß die erkannte SWE-Folge auftritt. Diese Sprachmodelle können ebenfalls als HMMs konzipiert werden. Besonders populär sind Digramm- oder Trigramm-Grammatiken, die sich darauf beschränken, die Wahrscheinlichkeit des Auftretens einer SWE in ihrem unmittelbaren Kontext zu beschreiben, da sich hierfür Statistiken noch mit vertretbarem Datenaufwand berechnen lassen. Anschließend können dann beispielsweise Lexika herangezogen werden, um aus dem erkannten Symbolstrom zulässige Wörter abzuleiten. Auf dem resultierenden Wortstrom kann wiederum analog verfahren werden: Auch hier können durch Bestimmung der jeweiligen Wortklasse und unter Einbeziehung grammatischen und semantischen Wissens Modelle zur Bewertung der Wortfolge gebildet werden. Die jeweiligen Bewertungen können ggf. als Gewichtungen in den Schritten des auf Seite 23 beschriebenen Suchprozeßes eingebracht werden.

Die hier beschriebenen Verfahren sind exemplarisch, aber realistisch und typisch. Es gibt allerdings zahllose Variationen und alternative Konzepte, wie beispielsweise die Merkmalsbildung auf der Basis akustisch-artikulatorischer Parameter anstelle von spektralen Parametern oder die Phonemerkennung auf der Basis künstlicher neuronaler Netzwerke, sowie hybride Systeme, bei denen etwa die Parameter der HMMs durch Neuronale Netze modifiziert werden. Letzten Endes illustriert die vorherrschende Verwendung von spektralen Merkmalen in Verbindung mit wie auch immer gearteten stochastischen Modellen eine trotz aller Anstrengungen nach wie vor unvollständige Kenntnis von der Struktur der akustischen Realisation sprachlicher Äußerungen.

3. Sprechernormalisierung

In der Einleitung wurde dargelegt, daß der Erkennungsfehler eines Spracherkennungssystems bei gegebenem Merkmalsraum sowohl mit der Anzahl der zu unterscheidenden Lautklassen (Wörter, SWE) steigt als auch mit der Streuung, die diese Einheiten innerhalb ihrer Klasse besitzen. Im letzten Kapitel wurde auf die Vorverarbeitung und Aufbereitung des Sprachsignals eingegangen, welche zum Ziel hat, redundante und für die Klassifikation nicht verwertbare Information aus dem Sprachsignal zu entfernen. Im folgenden sollen Zusammenhänge dargestellt werden, die zu Streuungen innerhalb der jeweiligen Klassen führen und von den beschriebenen Vorverarbeitungsmethoden nicht bzw. nicht vollständig erfaßt werden.

3.1. Variabilität des Sprachsignales

Die Ursachen für die unterschiedlichen Signale, die ein Erkennungssystem auch bei gleichen Lauten registriert, sind vielfältig und teilweise miteinander verquickt. Sie lassen sich grob unterteilen in solche, die vom Übertragungskanal herrühren, und in solche, die in den unterschiedlichen Realisationen der Laute durch die Sprecher selbst begründet sind.

3.1.1. Übertragungskanalinduzierte Ursachen

Der Übertragungsweg zwischen dem von den Lippen (und der Nase) des Sprechers abgestrahlten Schall und dem digitalisierten akustischen Signal, welches letztlich dem Erkenner zugeführt wird, besteht aus dem Raum, in dem der Sprecher spricht, und der Aufnahmeapparatur. Wichtige Ursachen für die durch den Übertragungsweg eingeführten Variabilitäten liegen in der nicht vollständig reproduzierbaren Position des Sprechers, welche zu verschiedenen Raumübertragungsfunktionen führt; genausowenig ist im allgemeinen die Entfernung zum Mikrofon und die Sprechrichtung konstant. Modifikationen im Aufnahmesystem, wie z. B. der Wechsel des Mikrophons, führen ebenfalls zu Variationen des Signals. In der praktischen Anwendung sind Störgeräusche (Lüfterrauschen, Kaffeemaschine) so wenig zu vermeiden wie ein gewisses Grundrauschen in der Aufnahmeapparatur.

3.1.2. Sprecherinduzierte Ursachen

Die unterschiedliche Realisation gleicher Laute durch einen oder mehrere Sprecher ist in der verschiedenartigen Anregung, Beschaffenheit und Konfiguration der Vokaltrakte der Sprecher begründet. Auch diese Ursachen können wieder in zwei Kategorien zerlegt werden, und zwar in Ursachen für die Intra- und solche für die Intersprechervariabilitäten. Diese Einteilung und die folgende Aufzählung der Ursachen ist allerdings recht unscharf und soll nur illustrieren, wie vielfältig die Gründe für die Variabilitäten des Sprachsignals sind.

3.1.3. Variabilitäten in den Äußerungen eines Sprechers

Ursachen, die bei einem gegebenen Sprecher zu verschiedenen Sprachsignalen für gleiche Laute führen, sind z. B.

- die Verfassung des Sprechers (Gesundheit, emotionaler Zustand, Atemaufwand etc.),
- die Sprechweise (flüstern, schreien, verwendete Lautung etc.) und
- die Umgebung. (Der Sprecher paßt seine Sprechweise der Umgebung an.[1])

Dies sind im wesentlichen Unterschiede in der Artikulation und Phonation.

Dagegen wirken sich Erkältungen beispielsweise zwar ebenfalls auf die Realisation eines Lautes aus, haben aber wegen der mit ihnen verbundene Änderung der Vokaltraktübertragungsfunktion durch das Zuschwellen des Nasen- und Rachenraumes aus der Sicht der Spracherkennung eher den Rang eines Sprecherwechsels.

3.1.4. Variabilitäten in den Äußerungen verschiedener Sprecher

Sprecherspezifische Variabilitäten, also solche, die individuelle Eigenschaften des Sprechers widerspiegeln, führen bei der Betrachtung der Äußerungen mehrerer Sprecher zu zusätzlichen Streuungen in den realisierten Äußerungen. Die Ursachen hierfür umfassen beispielsweise

- das Geschlecht der Sprecher (Länge des Vokaltraktes, Anregungsfrequenz),
- die individuelle Anatomie des Vokaltraktes und
- das Alter (Mikrotremor).

Die beiden ersten Aspekte sind Auswirkungen anatomischer Unterschiede zwischen den Sprechern. Genauso wie die unterschiedlichen Anregungsfrequenzen bei Kindern, Frauen und Männern wirkt sich auch ein eventueller Mikrotremor bei älteren Menschen durch die Verbreiterung der Harmonischen grundlegend auf das Spektrum

[1] In einer geräuschvollen Umgebung tut er dies, um seine Stimme aus dem Geräusch hervorzuheben (Lombard-Effekt, (van Summers et al., 1988)).

stimmhafter Laute aus. Verschiedenartige Anatomie führt auch zu verschiedenartiger Sprachqualität (Klang, Timbre), welche im wesentlichen durch das Spektrum der Anregung sowie die Übertragungsfunktion des Vokaltraktes bestimmt wird. Parameter, die durch die Anatomie des Sprechers beeinflußt werden, sind beispielsweise

- der zeitliche Verlauf des anregenden glottalen Flusses innerhalb einer Schwingungsperiode.
- der zeitliche Verlauf der Anregungsfrequenz,
- die mittlere Anregungsfrequenz,
- die Absolutwerte der Formantfrequenzen,
- die Formantbreiten,
- die Formant-Trajektorien,
- die Gestalt der spektralen Einhüllenden und mit ihr der spektrale Abfall sowie
- das mittlere Spektrum.

Ebenfalls sprecherspezifisch sind Akzentfärbungen, Dialektlautungen oder sonstige Sprechgewohnheiten. Die sich daraus ergebenden Aussprachevarietäten sind sehr vielfältig.

Bei verschiedenem Akzent oder Dialekt der Sprecher sind die Auswirkungen auf die Sprechäußerung ebenfalls lautspezifisch. Manche dieser Effekte resultieren in einer sprecherabhängigen Verschiebung der Lautklassen; beispielsweise können Vokale mehr oder weniger offen gesprochen werden, ohne daß dadurch ein anderes Phon realisiert würde (Disner, 1980).[2] Anders liegt die Problematik etwa bei dem deutschen Phonem /ʀ/, welches je nach Kontext und Lautung als [ʀ], [ʁ], [r] oder [ɾ] artikuliert wird, oder gar bei Allophonen, also verschiedenen Lauten des gleichen Phonems, wie etwa [ç] und [x] Allophone eines Phonems /χ/ sind (Drosdowski, 1974). Solcherlei Variationen müssen auf einer höheren Stufe des Erkennungssystems verarbeitet werden, und daher soll darauf hier nicht weiter eingegangen werden.

Naturgemäß werden alle diese Ursachen für die Variabilität des Sprachsignals durch eine *sprecherabhängige* Spracherkennung weitgehend vermieden. Daher ist der Erkennungsfehler eines sprecherunabhängigen Spracherkennungssystems notwendigerweise größer als bei sprecherabhängiger Spracherkennung. Hazen (1998) beispielsweise berichtet über mehr als eine Halbierung des Fehlers beim sprecherabhängigen Betrieb eines Erkennungssystems im Unterschied zum sprecherunabhängigen Betrieb.

[2]Damit ist gemeint, daß den Lauten beispielsweise nach der in diesem Text verwendeten IPA-Lautschrift (Anhang B.1, (IPA)) dennoch das gleiche Symbol zugeordnet werden muß.

3.2. Sprecheradaptation und Sprechernormalisierung

Für manche Anwendungen, wie z. B. für automatische Auskunftssysteme, stellt eine sprecherabhängige Spracherkennung keine Option dar. Insofern ist die Wahl eines sprecherunabhängigen Erkennungssystems eine Frage der Notwendigkeit, genauso, wie die gewünschte Anwendung die Zahl der zu erkennenden Kategorien festlegt. Da, wo eine sprecherabhängige Spracherkennung nicht möglich ist, kann man den Versuch unternehmen, die sich hieraus ergebende zusätzliche Variabilität zu vermindern. Dazu stehen prinzipiell zwei Wege offen: die sprecherabhängige Variation des Klassifikators und die sprecherabhängige Variation der Vorverarbeitung. Dies sind keine Alternativen. Vielmehr kann man in der Anwendung beides unternehmen.

Die Begriffe *Sprecheradaptation* und *Sprechernormalisierung* werden in der Literatur uneinheitlich verwendet. Hier soll jedoch die Modifikation der Parameter des Klassifikators *Sprecheradaptation* genannt werden, die Modifikation der Vorverarbeitung hingegen *Sprechernormalisierung*.

3.2.1. Sprecheradaptation

Da bei der Sprecheradaptation der Erkenner selbst modifiziert wird, sind die gewählten Methoden vom speziellen Klassifikatortyp abhängig. Man kann daher diese Adaptationsmethoden nicht immer ohne weiteres für eine andere Erkennerstruktur verwenden. In der Praxis ist das nicht unbedingt eine Einschränkung, da der größte Teil der derzeit eingesetzten Erkenner augenscheinlich durch irgendeine Form von Hidden-Markov-Modellen realisiert ist. Insofern sind die für diesen Erkennertypus gefundenen Adaptationsmethoden durchaus übertragbar.

Verbreitet sind hier Methoden zur sprecherabhängigen Neuschätzung der Mittelwerte der Gaussschen Emissionswahrscheinlichkeiten in sprecherunabhängigen (semi-) kontinuierlichen HMMs (Seite 23). Dies geschieht typischerweise durch Varianten einer *Maximum-Likelihood*-Schätzung, von denen wiederum MAP (*Maximum-a-posteriori*) (Gauvain und Lee, 1994; Lee und Gauvain, 1993) und MLLR (*Maximum-Liklihood-Linear-Regression*) (Legetter und Woodland, 1995) bzw. Hybride davon derzeit am häufigsten verwendet werden (beispielsweise Chesta et al. (1999); Chou (1999); Goronzy und Kompe (1999)).

Kuhn et al. (1999) verfolgen den Ansatz, die Mittelwerte aus bereits trainierten prototypischen Einzelwort-HMMs als Vektoren zur Konstruktion eines Sprecherraumes zu verwenden. Sie unterziehen diese Vektoren einer Hauptkomponentenanalyse, um eine geringe Zahl von »Eigenstimmen« zu finden, welche den Sprecherraum definieren. Anschließend schätzen sie die Position eines unbekannten Sprechers in diesem Raum, um daraus neue Mittelwerte für entsprechende sprecherunabhängig trainierte HMMs zu ermitteln. Sie berichten über deutlich bessere Ergebnisse für kleine Trainingsmengen als bei der Verwendung von MAP oder MLLR. Nguyen et al. (1999) verbinden in diese Methode wiederum u. a. mit MLLR.

Einen ähnlichen Ansatz verfolgen Wang und Liu (1999), die für gegebene Referenzsprecher sowohl jeweils sprecherabhängige als auch je ein gemeinsames sprecherun-

abhängiges HMM bilden. Hier wird ebenfalls ein neuer Mittelwert für die Emissionswahrscheinlichkeitsverteilung jedes Zustandes ermittelt, der sich als eine gewichtete Summe der Mittelwerte der sprecherabhängigen Modelle ergibt, wobei die Gewichtung durch eine *Maximum-Likelihood*-Schätzung aus den Adaptionsdaten ermittelt wird.

Davon unabhängig und auch für die Sprechernormalisierung brauchbar sind Methoden wie *Speaker Clustering*. Hazen (2000) beschreibt eine solche Methode, bei der sprecherabhängige Markov-Modelle entsprechend gewisser Gemeinsamkeiten der zugehörigen Sprecher (Geschlecht, Sprechgeschwindigkeit) gruppiert werden. Ein unbekannter Sprecher wird gegebenfalls in mehreren Durchgängen einer der Gruppen zugeordnet. Adaptions- und Normalisierungstechniken können dann begrenzt auf die Modelle dieser Sprechergruppe angewandt werden.

3.2.2. Sprechernormalisierung

In gewisser Weise stellt sich die Situation bei der Sprechernormalisierung komplementär zur Sprecheradaptation dar: Die Normalisierungsmethoden sind teilweise spezifisch für eine gewisse Darstellung des Sprachsignals und können nicht immer auf andere Darstellungen übertragen werden. Da die Wahl des eigentlichen Erkenners die Wahl der Darstellung festlegen oder zumindest einschränken kann, ist nicht jede Normalisierungsmethode für jeden Erkennertypus geeignet. Andererseits ähneln sich die etablierten Darstellungen in ihrer Struktur und gestatten daher weitgehend eine Übertragung der Methoden.

In der Sprechernormalisierung werden Abbildungen von Merkmalsvektoren durch lineare Regressionsmethoden oder auch durch neuronale Netze (Fukuzawa et al., 1991) versucht, es herrschen aber eindeutig Methoden vor, die derzeit im weitesten Sinne als *Vokaltraktnormalisierung* (VTN) bezeichnet werden. Damit ist nicht etwa die Normalisierung auf der Basis von Vokaltraktparametern gemeint (Freienstein, 2000; Naito et al., 1999), sondern im wesentlichen eine zumeist nichtlineare Transformation der Frequenzachsen der Spektrogramme, die den Merkmalsvektoren zugrundeliegen, auf die eines Zielsprechers. Dazu wurden verschiedene Ansätze untersucht, wie beispielsweise die Verzerrung der Frequenzachse mittels Dynamischer Programmierung (Blomberg und Elenius, 1986) oder Polynom-Anpassungen der Sprachspektren (Zahorian und Jagharghi, 1991). In jüngerer Zeit werden solche Verzerrungen auch mit Hilfe vorhandener sprecherunabhängig trainierter HMMs optimiert. So beschreiben Welling et al. (1999) einen Ansatz, der dies für eine stückweise lineare Frequenzskalenverzerrung erreicht. Dabei wird von einem durchschnittlichen Sprechermodell ausgegangen, woraufhin für eine Anzahl Trainingssprecher eine geeignete Verzerrung ermittelt wird. Ein unbekannter Sprecher wird einem Erkennungsprozeß durch die durchschnittlichen Modelle unterzogen. Anschließend wird bestimmt, welche der trainierten Verzerrungen die Wahrscheinlichkeit für die erkannte Äußerung maximiert. Daraufhin wird mit der gefundenen Verzerrung eine endgültige Erkennung durchgeführt.

VTN-Methoden können ebenfalls zur Berechnung einer Verschiebung der Mittelwerte der Emissionswahrscheinlichkeiten von HMMs verwendet werden, wie dies

beispielsweise von McDonough und Byrne (1999) untersucht wird.

3.3. Bereitstellung sprecherspezifischer Information

Der Grundgedanke bei allen Versuchen der Sprecheradaptation oder -normalisierung ist der, daß man ein eigentlich sprecherabhängiges Erkennungssystem an den jeweils aktuellen Sprecher anpaßt und so zu einem sprecherunabhängigen System gelangt, ohne die beschriebenen Nachteile (Seite 27) eines wirklich sprecherunabhängigen Systems in Kauf nehmen zu müssen. Für diesen Ansatz ist es offensichtlich erforderlich, daß die dafür zuständigen Verarbeitungsstufen Informationen über den aktuellen Sprecher erhalten. Dieses Wissen kann auf verschiedene Weise bereitgestellt werden, nämlich

- durch die Identifikation des Sprechers, wonach die gewünschte Information aus einem Lexikon ausgewählt werden kann, oder
- durch die Schätzung grundlegender sprecherspezifischer Eigenschaften aus der (längerfristigen) Analyse des Sprachsignals des aktuellen Sprechers.

Die Identifikation des Sprechers findet entweder direkt durch den Sprecher (z. B. über die Tastatur des Computers) statt oder wird von einem *Sprechererkenner* vorgenommen. Ist der Sprecher identifiziert, kann der Erkenner spezifisch und für den Sprecher optimal arbeiten, und man hat eigentlich wieder einen sprecherabhängigen Spracherkenner. Die Nachteile der Sprecheridentifikation liegen für die Selbstidentifikation des Sprechers in einer potentiellen Unzumutbarkeit. Außerdem läßt die konkrete Anwendung eine solche direkte Steuerung des Systems unter Umständen nicht zu. Im Falle der Identifikation des Sprechers durch einen Sprechererkenner verlagert man den möglichen Klassifizierungsfehler auf den Sprechererkenner, denn dieser kann den aktuellen Sprecher verwechseln.[3] Die Erhöhung der Sprechererkennungssicherheit durch die Einführung einer obligatorischen Schlüsseläußerung führt wieder zu einer Unbequemlichkeit für den Benutzer, die diesem möglicherweise nicht zugemutet werden kann oder soll. In jedem Fall ist man auf eine gewisse Zahl bekannter Sprecher beschränkt, was diese Methoden für eine Reihe von Anwendungen disqualifiziert.

Die Schätzung spezifischer Eigenschaften des aktuellen Sprechers kann verwendet werden, um den Erkenner beispielsweise zumindest an die durch das Geschlecht des Sprechers gegebenen Charakteristika des Sprachsignals anzupassen. In dieser Arbeit wird untersucht, ob bzw. inwieweit eine solche Analyse zu einer Auswahl oder Mischung von prototypischen Parametersätzen für die Vorverarbeitung des Sprachsignals herangezogen werden kann. Sie beschäftigt sich also im Sinne der obigen Definition mit Sprechernormalisierung.

[3] Dieses Risiko ist beispielsweise bei Diktiersystemen relativ klein, da dem Sprechererkennungssystem hier eine Menge Datenmaterial für seine Entscheidung zur Verfügung steht. Andererseits sind Diktiersysteme gerade Anwendungen, die nach einem sprecherabhängigen Spracherkennungssystem verlangen.

3.4. Ansätze zur Sprechernormalisierung

Wenn die Struktur und die Parameter des Klassifikators nicht verändert werden sollen, muß eine Sprechernormalisierungsstufe das Sprachsignal zwangsläufig in einer Weise manipulieren, daß das Ergebnis der Normalisierung die gleiche Form besitzt wie das Eingangssignal der Erkennungsstufe. Genauer kann man formulieren: Ist $\mathcal{X}$ der $\mathcal{R}$-Vektorraum der Merkmalsvektoren, dann kann das vorverarbeitete Sprachsignal als eine vektorwertige Funktion $\mathbf{s} : \mathcal{N} \rightarrow \mathcal{X}$ der (diskreten) Zeit t aufgefaßt werden. Ohne eine Normalisierung sind ihre Werte $\mathbf{s}(t)$ (die Merkmalsvektoren) das Eingangssignal des Klassifikators, und eine Normalisierung ist dann eine Abbildung $\mathbf{N} : \times_{i=1}^{n} \mathcal{X} \rightarrow \mathcal{X}$, wobei $\times_{i=1}^{n} \mathcal{X}$ das kartesische Produkt von n Exemplaren von $\mathcal{X}$ ist. Die Normalisierung kann also auch auf der Basis einiger aufeinanderfolgender Merkmalsvektoren erfolgen.

Die Sprechernormalisierung wird häufig vorgenommen, indem man die Sprachsignale $\mathbf{s}(t)$ des aktuellen Sprechers auf die gleichen Laute eines Referenzsprechers abbildet, für den der Erkenner trainiert worden ist (Knohl und Rinscheid, 1993; Müller und Strube, 1993).[4]

Nun stellt sich also die Frage, wie so eine normalisierende Abbildung $\mathbf{N}$ gefunden werden kann. Die Antwort darauf steckt in den Unterschieden, die sich in den Merkmalsvektoren finden, wenn verschiedene Sprecher die gleiche Äußerung realisieren. Welcher Art die Unterschiede im Sprachsignal sind, wurde schon in Abschnitt 3.1.4 umrissen.

Von den dort erwähnten Parametern sind idealerweise solche, die (bei vokalischer Anregung) mit der Anregungsfrequenz des Vokaltraktes verbunden sind, in den Merkmalsvektoren nicht mehr prominent, da es zu den Zielen der Vorverarbeitung gehört, diese Eigenschaften aus dem Signal zu entfernen. Langfristige Parameter wie mittlere spektrale Einhüllende können dagegen auch nach der Vorverarbeitung noch leicht erfaßt werden. Sie sind zwar mitunter nur schwer von den übertragungskanalinduzierten Variabilitäten zu trennen, aber dies ist für die Anwendung (anders möglicherweise als für die Forschung) ohne Belang, da diese Variabilitäten in der gleichen Weise behandelt werden können. Mittelfristige Parameter wie Formanttrajektorien und prosodische Parameter sind vergleichsweise schwer zu bestimmen. Wegen der nötigen Datenmenge können manche typische prosodische Eigenschaften des aktuellen Sprechers selbst bei Systemen zur Erkennung fließender Sprache nur dann gewonnen werden, wenn der Sprecher hinreichend lange mit diesen Systemen interagiert.

Da die spätere Klassifikation der Äußerung nur durch solche sprecherspezifischen Eigenschaften beeinträchtigt werden kann, die sich auch in den Merkmalsvektoren wiederfinden lassen, ist es naheliegend, die Merkmalsvektoren selbst zur Bestimmung der Abbildung $\mathbf{N}$ zu verwenden. Relativ einfach liegt der Fall, wenn man $\mathbf{N}$ zwischen bekannten Sprechern auf der Basis bekannter Äußerungen finden will. In diesem Fall

[4] Eine Alternative ist die Transformation der Merkmalsvektoren aller Sprecher in eine sprecherunabhängige Darstellung. Dies erfordert dann aber einen explizit für diese Darstellung trainierten Klassifikator (Class et al., 1990).

kann jede Methode, die eine Abbildung von gegebenen Eingangsvektoren auf bekannte Ausgangsvektoren generiert, verwendet werden. In der Literatur werden verschiedene lineare und nichtlineare Methoden verwendet, die auch je nach Anforderung (Art des Spracherkennungssystems, Anwendung/Forschung) variieren. Verbreitet sind Algorithmen, die eine stetige, zumeist nichtlineare Abbildung zwischen den Signalen eines Testsprechers und denen des Referenzsprechers konstruieren. Häufig werden hierfür *Perzeptrone*[5] in diversen Modifikationen eingesetzt (Fukuzawa et al., 1991; Huang et al., 1991). Ebenfalls angewendet wird die Vektorquantisierung (VQ) der Sprachsignale. Die Abbildung besteht dann darin, den Repräsentanten des Eingangssignals die (phonetisch) zugehörigen Codebuchvektoren des Referenzsprechers zuzuordnen (Knohl und Rinscheid, 1993). Ein Nachteil dieser Methode liegt hier wieder darin, daß durch die Quantisierung Information verlorengeht. Prinzipiell kann die Quantisierung als eine weitere Klassifikation betrachtet werden, die die schon erwähnten Fehlermöglichkeiten in sich birgt. Arbeiten auf diesem Gebiet beschäftigen sich damit, Methoden zu finden, diesen Fehler so klein wie möglich zu halten (z. B. Fuzzy-VQ, (Nakamura und Shikano, 1990)).

[5] Perzeptrone sind ein spezieller Typ neuronaler Netzwerke, auf die ab Seite 38 und vor allem in Anhang A.3 noch ausführlich eingegangen wird.

4. Experimente zur Sprechernormalisierung

Im folgenden Abschnitt wird das in dieser Arbeit verwendete Datenmaterial und seine Aufbereitung vorgestellt. Die anschließenden Abschnitte behandeln dann Experimente zur Sprechernormalisierung auf den beschriebenen Datenbasen.

4.1. Datenmaterial

Die im folgenden angeführten Experimente fanden auf zwei verschiedenen Datenbasen statt, deren Inhalte aber in ähnlicher Weise vorverarbeitet wurden. Eine der Datenbasen (SPINA) wurde am Dritten Physikalischen Institut der Universität Göttingen und an der Universität Bochum (Lehrstuhl für allgemeine Elektrotechnik und Akustik) im Zusammenhang mit dem BMFT[1]-Projekt SPINA[2] erstellt, die andere Datenbank ist die kommerzielle PHONDAT-1-Datenbasis, welche an den Universitäten München, Kiel, Bonn und Bochum aufgenommen wurde. Beide Datenbasen sind deutschsprachig.

4.1.1. Die SPINA-Datenbasis

Die SPINA-Datenbasis umfaßt 62 Einzelwörter[3] und 10 Sätze, die von über 20 Sprechern (jeweils zur Hälfte Männer und Frauen) gesprochen wurden. Jede der Äußerung liegt dabei in 5 Wiederholungen vor, die während mindestends zweier verschiedener Sitzungen in einem reflexionsarmen Raum mithilfe eines DAT[4]-Rekorders aufgezeichnet wurden. Die Signale wurden vom DAT analog ausgegeben, bei 8 kHz tiefpaßgefiltert und mit 16 Bit Wortbreite bei einer Abtastrate von 16 kHz redigitalisiert. Das digitalisierte Zeitsignal wurde durch Hammingfenster mit einer Fensterlänge von 40 ms bei einem Vorschub von 10 ms gefenstert. Die Signalstücke wurden durch eine FFT in 256kanalige Amplitudenspektren transformiert. Durch die Zusammenfassung der entsprechenden Kanäle wurden die Spektren in 21kanalige Barkspektren umgewandelt.

[1] BMFT: Bundesministerium für Forschung und Technologie, inzwischen BMBF: Bundesministerium für Bildung und Forschung.
[2] SPINA: *SPrachverstehen In Neuronaler Architektur*, FKZ: 413-4001-01 IN 108 A/2.
[3] In dieser Arbeit fanden nur die in der Basis enthaltenen Einzelwortäußerungen Verwendung.
[4] DAT: *Digital Audio Tape*.

Die Zusammenfassung wurde jeweils innerhalb von Fenstern vorgenommen, die auf der Barkskala nach Gleichung 2.4 die Form von gleichschenkligen Trapezen hatten. Die Basisbreite der Trapeze betrug 1,5 Bark, die der oberen Grundlinie 0,5 Bark. Der Überlapp der Fenster war 0,5 Bark. Um dem durch Gleichung 2.1 gegebenen Zusammenhang zwischen Signalamplitude und Lautheit Rechnung zu tragen, wurden alle Signale einer Dynamikkompression mit einem Exponenten 0,5 unterzogen, d. h. aus jedem einzelnen Element der Merkmalsvektoren wurde die Quadratwurzel gezogen. Abschließend wurden die Spektrogramme jeder Äußerung auf einen Maximalwert von 1,0 normiert.

4.1.2. Die PhonDat-Datenbasis

Die PhonDat-1-Datenbasis umfaßt 450 Äußerungen, und zwar im wesentlichen Sätze, aber auch Einzelwörter und zwei kurze Textpassagen. 201 verschiedene Sprecher haben jeweils Teilmengen des Korpus gelesen. Die Aufnahme erfolgte unter Studiobedingungen an den genannten vier Universitäten mit verschiedenem Gerät. Die Abtastrate bei der Digitalisierung betrug 48 kHz, die Auflösung 16 Bit. Die Signale wurden bei 8 kHz digital tiefpaßgefiltert und auf 16 kHz heruntergetastet. Für die im Korpus enthaltenen Sätze existieren Phonemlabels. Das weitere Verfahren entspricht dem oben für die Spina-Daten beschriebenen, allerdings wurden die Merkmalsvektoren zunächst nicht normiert.

4.2. Abbildung ohne gelabelte Quellsprecherdaten

Für die Bestimmung einer sprechernormalisierende Abbildung ohne die Kenntnis über den Wortlaut der getätigten Äußerungen stehen zwei Wege zur Verfügung:

1. Man verwendet den gegebenen Spracherkenner zur Erzeugung eines Labelings, oder

2. man verwendet langfristige Parameter wie das mittlere Spektrum.

Da man sich im ersten Fall bildlich an seinen eigenen Haaren aus dem Sumpf zieht *(Bootstrap)*, stellt sich das Problem der Zuverlässigkeit des Klassifikators auf der Grundlage nichtnormalisierter Daten. Dies mag das folgende (extreme) Szenario illustrieren: Ein Klassifikator kategorisiere jede getane Äußerung fälschlich in eine bestimmte, immer gleiche Lautklasse. Eine normalisierende Abbildung, die möglicherweise als identische Abbildung initialisiert worden ist, wird dann durch einen iterativen Prozeß zu einer Abbildung werden, die jede Äußerung auf genau den prototypischen Merkmalsvektor dieser Lautklasse transformiert. So wäre sie am Ende des Prozesses schlechter, als sie in ihrer Startkonfiguration war. Um ein solches Bootstrapping zu ermöglichen, muß daher der Klassifikator entweder auch ohne Normalisierung schon hinreichend gut sein, oder man muß aus den Hypothesen des Klassifikators solche zum Training der Normalisierung verwenden, die hinreichend sicher (im Sinne

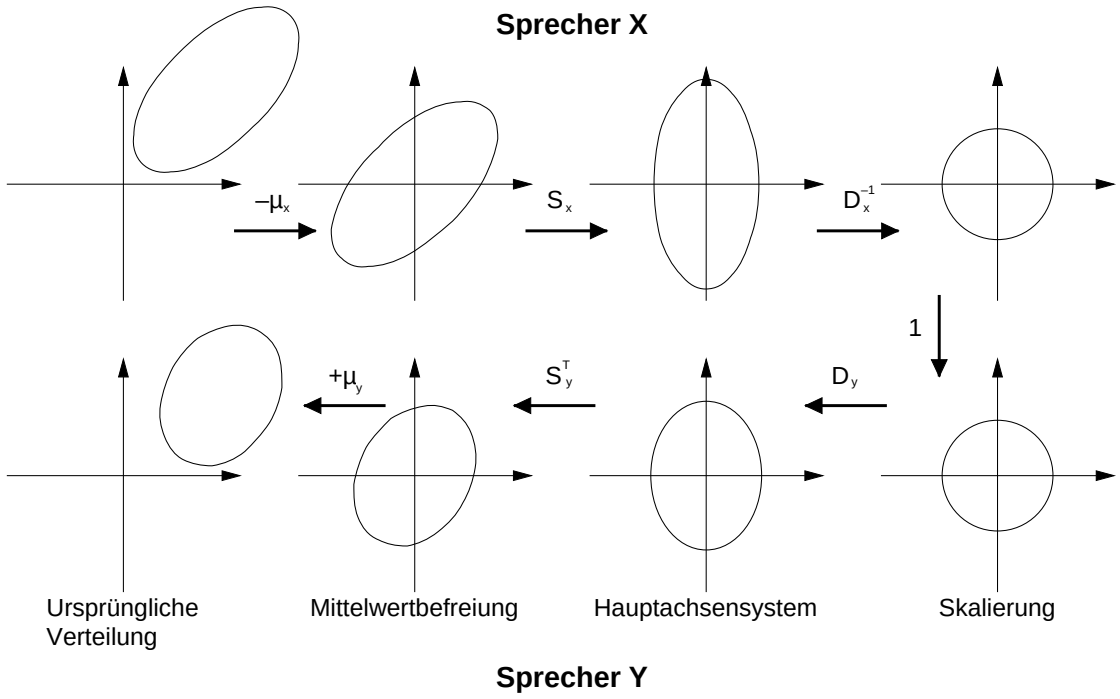

Abbildung 4.1.: Skizze des Verfahrens zur Sprechernormalisierung durch Hauptkomponentenanalyse der Merkmalsvektoren. Die eingezeichnete Abbildung folgt den Gleichungen 4.1 und 4.6. Wird (ganz rechts) statt der Einheitsmatrix die Matrix $\tilde{\mathbf{A}}$ *(Seite 37) eingesetzt, erhält man eine identische Abbildung.*

einer Wahrscheinlichkeitsbewertung durch den Klassifikator) erkannt wurden und alle Zweifelsfälle verwerfen.

Die zweite Methode wird im allgemeinen weniger effektiv sein, da die zur Verfügung stehende Information geringer ist, ist dafür aber völlig unabhängig von dem anschließenden Klassifikator. Wie im vorherigen Kapitel erklärt wurde, liegt das Hauptaugenmerk dieser Arbeit nicht auf dieser Problematik. Es wird im folgenden aber ein Experiment dargestellt, das diesen Ansatz verfolgt.

Das Verfahren basiert auf einer Hauptkomponentenanalyse. Diese wurde schon mehrfach erfolgreich angewendet, um verschiedene Lautklassen zu normalisieren, z. B. von Klein et al. (1970) und Zahorian und Jagharghi (1992). Hier geschieht das aber gewissermaßen für alle Klassen gleichzeitig. Dem Verfahren liegt folgender Gedanke zu Grunde: Faßt man die einzelnen Merkmalsvektoren einer ausreichenden (repräsentativen) Auswahl von Äußerungen eines Sprechers als Vektoren im $\mathcal{R}^n$ auf, so bilden sie eine Punktwolke, die das für diesen Sprecher im Merkmalsraum zugängliche Gebiet definieren. Geht man davon aus, daß die sich so für zwei Sprecher ergebenden Verteilungen näherungsweise durch eine affine Abbildung[5] ineinander überführt werden können, kann die gewünschte Abbildung unter gewissen Voraussetzungen (vgl.

[5]Eine affine Abbildung ist eine Kombination der geometrischen Operationen Translation, Spiegelung, Rotation und Dilatation.

die Diskussion weiter unten) durch zwei Hauptachsentransformationen approximiert werden. Das Verfahren ist in Bild 4.1 skizziert.

Die Abbildung geschah in folgender Weise: Zunächst wurden die Merkmalsvektoren $\mathbf{x}_i$ mittelwertbefreit und die Mittelwerte $\boldsymbol{\mu}$ gespeichert:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j, \quad \mathbf{x}_k \rightarrow \mathbf{x}_k - \boldsymbol{\mu}. \tag{4.1}$$

Um die Daten nun in ihr Hauptkoordinatensystem zu überführen, wurde jetzt die Kovarianzmatrix $\mathbf{C}$ der Daten berechnet

$$\mathbf{C} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T \tag{4.2}$$

und die Eigenwerte λ_k sowie die Eigenvektoren $\mathbf{s}_k$ von $\mathbf{C}$ bestimmt. Die Eigenwerte und die zugehörigen Eigenvektoren wurden ihrer Größe nach sortiert und die Eigenvektoren normiert:

$$\mathbf{s}_k \rightarrow \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|}. \tag{4.3}$$

Aus den neuen $\mathbf{s}_k$ wurde die Basiswechsel-Matrix gebildet

$$\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n) \tag{4.4}$$

und eine eine Streckungs-Matrix bestimmt, um die Varianzen der Merkmalsvektoren einander anzugleichen:

$$\mathbf{D} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}. \tag{4.5}$$

In der gleichen Weise wurden auch die Merkmalsvektoren des zweiten Sprechers verarbeitet. Die Abbildung der mittelwertbefreiten Daten des Quellsprechers geschah nun durch Multiplikation mit der Matrix

$$\mathbf{M} = \mathbf{S}_y^T\mathbf{D}_y\mathbf{A}\mathbf{D}_x^{-1}\mathbf{S}_x \tag{4.6}$$

und anschließende Additon der Mittelwerte $\boldsymbol{\mu}_y$:

$$\tilde{\mathbf{x}}_i = \boldsymbol{\mu}_y + \mathbf{M}(\mathbf{x}_i - \boldsymbol{\mu}_x). \tag{4.7}$$

Dabei stehen die Indizes x und y für den Quell- bzw. Zielsprecher. Wählt man hier für die Matrix $\mathbf{A}$ die Einheitsmatrix, transformiert man die Daten von dem einen Hauptkoordinatensystem auf das andere. Dies ist nicht so ohne weiteres möglich, weswegen $\mathbf{A}$ speziell konstruiert werden muß. Darauf wird gleich näher eingegangen werden.

Tabelle 4.1.: Erkennungsraten mit und ohne Sprechernormalisierung auf der Basis einer Hauptachsentransformation der Merkmalsvektoren und Adaptionsgrad in Prozent.

Sprecherpaare	Erkennungsraten (%)		Adaptionsgrad α
	vor Abbildung	nach Abbildung	%
Gesamt	78,6	82,9	20,1
Männlich → Männlich	86,9	86,7	−1,5
Männlich → Weiblich	78,3	82,3	18,4
Weiblich → Männlich	70,9	80,6	33,3
Weiblich → Weiblich	79,9	82,4	12,4

Da hier als Merkmalsvektoren Barkspektren verwendet wurden, waren die Dimensionen der Vektoren stark korreliert und die höheren Eigenwerte sehr klein. Um numerische Probleme zu vermeiden, wurden daher nur die »unteren« Dimensionen berücksichtigt, d. h. die Dimensionalitäten von $\mathbf{S}$, $\mathbf{D}$ und $\mathbf{A}$ wurden beschränkt: $\mathbf{S} \in \mathrm{M}(n \times p)$, $\mathbf{D} \in \mathrm{M}(p \times p)$, $\mathbf{A} \in \mathrm{M}(p \times p)$, mit $p < n$. Weil auch die Zuordnung der Hauptkomponenten für die höheren Dimensionen zunehmend unsicher wurde, wurden nur die größten q (mit $q < p$) Komponenten abgebildet, d. h. $\mathbf{A}$ hat folgende Gestalt:

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{\mathrm{red}} \end{pmatrix}. \tag{4.8}$$

Dabei ist $\mathbf{A}^{\mathrm{red}}$ die untere rechte $(p-q) \times (p-q)$-Submatrix von $\tilde{\mathbf{A}} = \mathbf{D}_y^{-1}\mathbf{S}_y\mathbf{S}_x^T\mathbf{D}_x$ und $\mathbf{1}_q$ die $q \times q$-Einheitsmatrix. Durch diese Transformation wurden die höheren Komponenten quasi »im Kreis«, also identisch abgebildet, wie man sieht, wenn man $\tilde{\mathbf{A}}$ in 4.6 einsetzt.

Mit dem dargestellten Verfahren wurde folgendes Experiment durchgeführt: Als Datenbasis wurden die SPINA-Daten verwendet. Anders als oben beschrieben wurden die Barkspektrogramme zusätzlich logarithmiert, um die Verteilung der Barkspektren zu entzerren. Jeweils ein Sprecher von sechs männlichen und sechs weiblichen Sprechern wurde als Quell- bzw. Zielsprecher gewählt. Alle Sprecherpaarungen wurden untersucht. Die Mittelwerte $\boldsymbol{\mu}$ und Abbildungsmatrizen $\mathbf{M}$ wurden aus sämtlichen Einzelwörtern aller fünf vorhandenen Versionen berechnet. In die Abbildung gingen die ersten zehn Hauptkomponenten ein, wovon zwei effektiv transformiert wurden. Für die Erkennungsexperimente wurde nur die Version 0 sowohl der Quell- als auch der Zielsprecher verwendet. Der Quellwortschatz bestand aus jedem zweiten Wort (in alphabetischer Folge), also aus 31 Wörtern, die durch eine DTW-Erkennung gegen den kompletten Wortschatz des Zielsprechers getestet wurden. Dies führte zu den in

Tabelle 4.1 angegebenen Erkennungsraten.[6,7]

Das beschriebene Verfahren verbessert die Erkennungsraten nur wenig. Die Ursache dafür liegt vermutlich hauptsächlich in den implizierten Annahmen über die Verteilung der Merkmalsvektoren: Die Anwendung einer Hauptkomponentenanalyse setzt eine Normalverteilung der Daten voraus, da hierdurch nur die ersten zwei Momente der Verteilungen berücksichtigt werden. Diese ist aber für die Merkmalsvektoren insbesondere wegen der Betragsbildung bei ihrer Berechnung nicht gegeben. Dies ist auch der Grund, weswegen in dem beschriebenen Experiment die Merkmalsvektoren durch Logarithmierung entzerrt wurden. Es steht zu erwarten, daß eine weitere Untersuchung mit einer Betragsbildung *nach* der Normalisierung deutlich bessere Ergebnisse ergeben würde. Dann sollte auch auf die Logarithmierung verzichtet werden. Desweiteren könnte sich eine Transformation von dreien oder aber nur einer Hauptkomponente als wirkungsvoller erweisen.

4.3. Sprechernormalisierende Abbildung durch Perzeptrone

Im folgenden werden Versuche zur sprechernormalisierenden Abbildung durch Perzeptrone beschrieben. Über deren Aufbau und Eigenschaften informiert Anhang A.3. Das Training der Netze erfolgte ebenfalls wie dort angegeben, nämlich nach der *Deltaregel* für die einfachen und nach dem *Backpropagation*-Algorithmus (BP) für mehrschichtige Netze. Anders als im Anhang angegeben, wurde allerdings noch ein sogenannter »Trägheitsterm« verwendet. Dabei wird Gleichung A.12

$$\Delta w_{ij} = -\epsilon \frac{\partial E}{\partial w_{ij}} \tag{4.10}$$

um einen Term erweitert, der den Gradienten des letzten Zyklus verkörpert:

$$\Delta w_{ij}^{n} = -\epsilon \frac{\partial E}{\partial w_{ij}^{n}} + \alpha \Delta w_{ij}^{n-1}. \tag{4.11}$$

[6] Der dort angegebene *Adaptionsgrad* α ist ein Maß für die Qualität der Normalisierung und berechnet sich wie folgt:

$$\alpha = \frac{R_{qz\,N} - R_{qz_0}}{R_{zz} - R_{qz_0}}. \tag{4.9}$$

Dabei sind R_{qz} die Erkennungsraten für die Äußerungen des Quellsprechers q bei *Targets* des Zielsprechers z. Die Indizes 0 und N bedeuten ohne bzw. mit Normalisierung. Für dieses Maß gilt:

- α=0 Die Erkennungsraten vor und nach Normalisierung sind gleich.
- α>0 Die Normalisierung steigert die Erkennungsraten.
- α=1 Die Äußerungen des Quellsprechers werden nach der Normalisierung genauso gut erkannt wie die Äußerungen des Zielsprechers selbst.
- α<0 Die Normalisierung verschlechtert die Erkennungsraten.

[7] Wegen des gewählten Verfahrens beträgt die Erkennungsrate für den Zielsprecher hier 100%.

Tabelle 4.2.: Für das Training der Perzeptrone verwendete Parameter. Die Parameter erwiesen sich in weiten Grenzen als unproblematisch.

Netzwerktyp	Initialisierung	Initialisierungs-intervall	Lernparameter ϵ	Trägheitsparameter α
einfach	zufällig	$\pm 0{,}001$	0,01	0,9
mehrschichtig				
linear	zufällig	$\pm 0{,}001$	0,01	0,8
nichtlinear	zufällig	$\pm 0{,}001$	0,01	0,8

Hier bedeuten die Indizes n den jeweiligen Zeitschritt. Analoge Veränderungen können an den Gleichungen A.20 und A.21 vorgenommen werden. Durch diesen Term erhält der Gradientenabstieg der Delta- oder BP-Regeln eine gewisse Trägheit, so daß kleine Störungen in der Fehlerlandschaft der Netze gewissermaßen gemittelt werden. Diese Maßnahme erlaubt die Verwendung größerer Schrittweiten des Lernparameters ϵ. Wie Phansalkar und Sastry (1994) zeigen konnten, beinflußt dieser Term das Ergebnis des Abstieges qualitativ nicht. Bei den im weiteren beschriebenen Versuchen wurden Parameter entsprechend Tabelle 4.2 verwendet. Diese erwiesen sich in weiten Bereichen als unkritisch und wurden, da es hier nicht auf eine optimale Konvergenzgeschwindigkeit ankam, in keiner Weise optimiert. Das Gleiche gilt für die Lernregeln. Insbesondere der BP-Algorithmus konvergiert zwar langsam, dafür muß er als einfaches Gradientenverfahren[8] (für vernünftige ϵ) aber konvergieren. Wiederum aus dem gleichen Grund wurde das Berechnen der neuen Gewichte nach jedem einzelnen Mustervektor vorgenommen.[9] Die Muster wurden in ihren natürlichen Einheiten (je ein Wort oder SWE) abgearbeitet und ihre Reihenfolge dann zufällig neu festgelegt.

4.3.1. Einfache Netze – mehrschichtige Netze

In früheren eigenen Arbeiten konnte gezeigt werden, daß eine sprechernormalisierende Abbildung bereits durch eine einfache fensterweise lineare Abbildung von Barkspektren erfolgreich etabliert werden kann (Müller und Strube, 1993). Daher ist die Annahme berechtigt, daß eine solche Abbildung durch Perzeptrone ebenfalls möglich, und wegen deren Fähigkeit, auch kompliziertere Abbildungen darzustellen, mindestens genauso erfolgreich sein wird. Diese Annahme kann jedoch zunächst nur für einfache Perzeptrone gemacht werden. Kompliziertere Netzwerke sind natürlich in der Lage, die gleichen Abbildungen zu leisten, können aber prinzipbedingt zu suboptimalen Lösungen führen.

[8] Details über Gradientenverfahren findet man beispielsweise in (Press et al., 1988).

[9] Man kann auch die Änderung der Gewichte Δw_{ij} über einige Mustervektoren kumulieren, bevor man endlich die Gewichte w_{ij} anpaßt. Dadurch erreicht man eine schnellere Simulation, läuft aber Gefahr, die Annahmen des Gradientenverfahrens zu verletzen.

Interessant ist aber insbesondere, welche Anzahl versteckter Zellen bei mehrschichtigen Perzeptronen benötigt wird, um eine adäquate Abbildung zu ermöglichen. Hierzu wurde folgendes Experiment unternommen: Ein Perzeptron mit einer versteckten Schicht wurde darauf trainiert, eine identische Abbildung zwischen den Barkspektren der Einzelwortäußerungen des Zielsprechers *weg* zu lernen. Daß dies im Falle der 21kanaligen Spektren mit 21 versteckten Zellen möglich sein muß, liegt auf der Hand, denn dann können zu Einheitsmatrizen proportionale Gewichtsmatrizen die gewünschte Abbildung approximieren. Die Spektrogramme wurden wortweise auf das Intervall [0,1 . . . 0,9] transformiert, um der Ausgangsdynamik des Netzes zugänglich zu sein. Alle Anzahlen von versteckten Zellen zwischen 1 und 22 wurden getestet. Um dem schlechten Schnitt der PhonDat-Einzelwortäußerungen gerecht zu werden, wurden die Spektrogramme anhand eines Energiekriteriums neu geschnitten. Dadurch wurde das Grundrauschen zu Beginn und am Ende der Äußerungen eliminiert. Insgesamt standen 13 Einzelwortäußerungen zur Verfügung, also wurden für jede Zellenzahl 13 Netzwerke jeweils unter Auslassung eines Wortes trainiert, welches dann zum Testen der Abbildung verwendet wurde. Der Trainingsprozeß wurde in jeweils 5 000 Schritten iteriert. Danach wurde das Spektrogramm des verbleibenden Wortes abgebildet und der mittlere euklidische Fehler zwischen der abgebildeten und der ursprünglichen Äußerung über alle Elemente aller Merkmalsvektoren berechnet. Bei der gegebenen geringen Datenmenge ist ein »Auswendiglernen« der Muster nicht auszuschließen, denn ein Netz mit je 21 Eingangs-, Zwischenschicht- und Ausgangszellen besitzt bereits 924 Parameter (22×21 pro Schicht, 22 wegen des Schwellwertvektors), und die Merkmalsvektoren sind hochkorreliert.

Erwartungsgemäß nimmt der Abbildungsfehler kontinuierlich mit der Zellenzahl ab. Daß dies auch bei höheren Zellenzahlen noch deutlich der Fall ist, ist etwas überraschend, da wenigstens zur Darstellung bzw. Diskriminierung von Vokalen eine zwei- bis dreidimensionale Repräsentation ausreichen sollte, wie in Abschnitt 2.3.1 beschrieben. Da die Vokale auch von ihrer Dauer und Energie das Sprachsignal dominieren, hätte man eine geringere Anzahl von Zwischenschichtneuronen für ausreichend halten können. Diese Annahme wird auch durch den erfolgreichen Einsatz von Kohonenschen Merkmalskarten (Kohonen, 1988) in der automatischen Spracherkennung (z. B. Behme und Brandt (1993); Brandt (1991)) gestützt. So aber liegt die Vermutung nahe, daß sich die Konsonanten einer niedrigdimensionalen Darstellung entziehen und auf diese Weise das Erfordernis einer großen Zahl versteckter Zellen begründen.

Um nun zu prüfen, ob ein mehrschichtiges Netz bei all seinen Nachteilen Vorteile gegenüber einfachen Netzen bietet, wurde das gleiche Experiment mit einem einfachen Perzeptron unternommen. Seiner schnelleren Konvergenz wegen wurden hier nur 1 000 Iterationen durchgeführt. Das Ergebnis ist als (untere) waagerechte Linie in der gleichen Abbildung aufgetragen. Man erkennt, daß der Fehler hier geringer ist, als bei dem komplizierteren Netz selbst bei 21 versteckten Zellen. Um noch einen anderen Vergleichswert zu ermitteln, wurde das Experiment noch ein weiteres Mal durchgeführt, wobei diesmal nur die mittleren Spektren der Äußerungen angepaßt wurden. Dies geschah so, daß jeweils 12 Äußerungen über die Zeit gemittelt wurden. Dann wurde die Vergleichsäußerung mittelwertbefreit und das mittlere Spektrum der 12

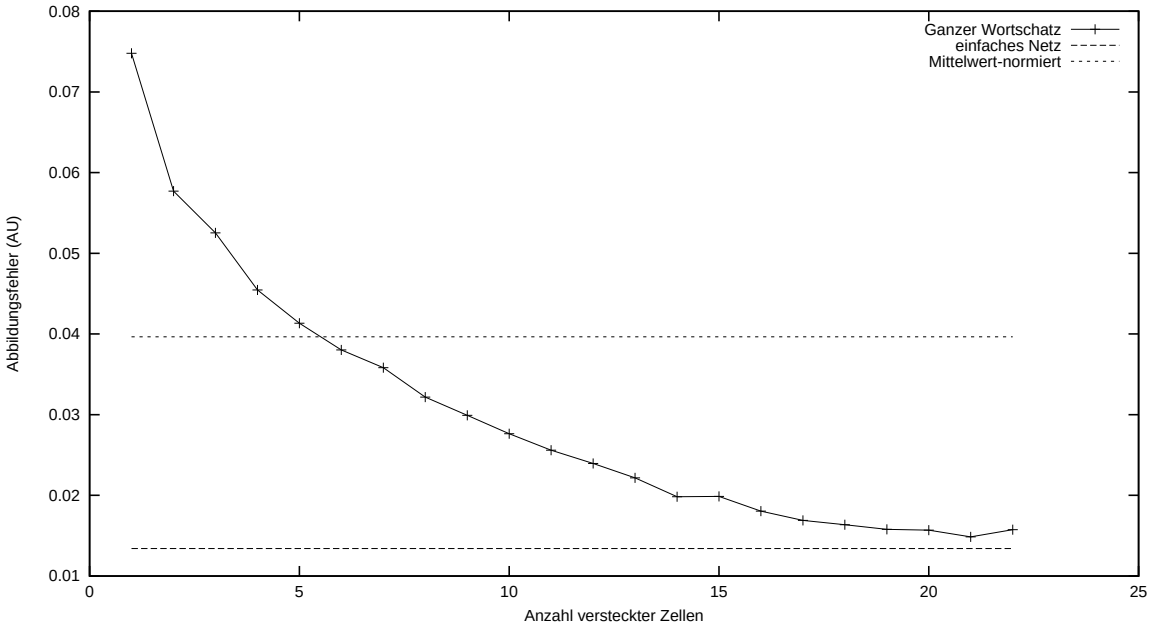

Abbildung 4.2.: Mittlerer Abbildungsfehler bei Abbildung des Einzelwortschatzes von Sprecher weg *auf sich selbst. Die Abbildung erfolgte durch ein Perzeptron mit einer versteckten Schicht. Der Abbildungsfehler ist als Funktion der Zellenzahl der Zwischenschicht aufgetragen. Die untere waagerechte Kurve stellt das Ergebnis für die Abbildung durch ein einfaches Perzeptron dar, die obere das Ergebnis für eine einfache Anpassung der Mittelwerte.*

Äußerungen zur Vergleichsäußerung addiert. Der mittlere Fehler bei dieser Art der Abbildung ist durch die obere waagerechte Linie in Abbildung 4.2 gegeben. Wie man sieht, wird selbst diese Linie von dem mehrschichtigen Netz erst bei 6 versteckten Zellen unterschritten.

Wie sich die verschiedenen Netzwerktypen bei der Abbildung von Spektrogrammen von ungleichen Quell- und Zielsprechern verhalten, sollte das nächste Experiment zeigen. Dazu wurden aus der Menge derjenigen Sprecher, die die 13 Einzelwörter aufgenommen haben, zufällig fünf (*w20*, *kpr*, *rot*, *m75*, *hud*) ausgewählt. Diesmal wurden sowohl einfache Perzeptrone als auch Perzeptrone mit einer versteckten Schicht von 11 Zellen darauf trainiert, die Spektrogramme der Einzelwortäußerungen von jeweils einem dieser Sprecher auf die des Zielsprechers *weg* abzubilden. Da die Äußerungen der verschiedenen Sprecher im allgemeinen verschiedene Längen besitzen, wurde zwischen den Äußerungen jeweils ein DTW-Alignment durchgeführt, wie es in Anhang A.1 beschrieben ist. Für den DTW-Algorithmus wurde als Fehlermaß der mittlere quadratische Fehler verwendet. Obwohl sich hierfür eine Kompression von $k = 0{,}2 \ldots 0{,}3$ (vgl. Gleichung 2.1) als optimal erwiesen hat (Müller, 1993), wurde darauf verzichtet, die ursprüngliche Kompression der Spektren ($k = 0{,}5$) zu ändern, da nicht zu erwarten war, daß die Ergebnisse davon qualitativ beeinflußt werden sollten.

Durch das Alignment werden die verschiedenen Merkmalsvektoren des einen Spek-

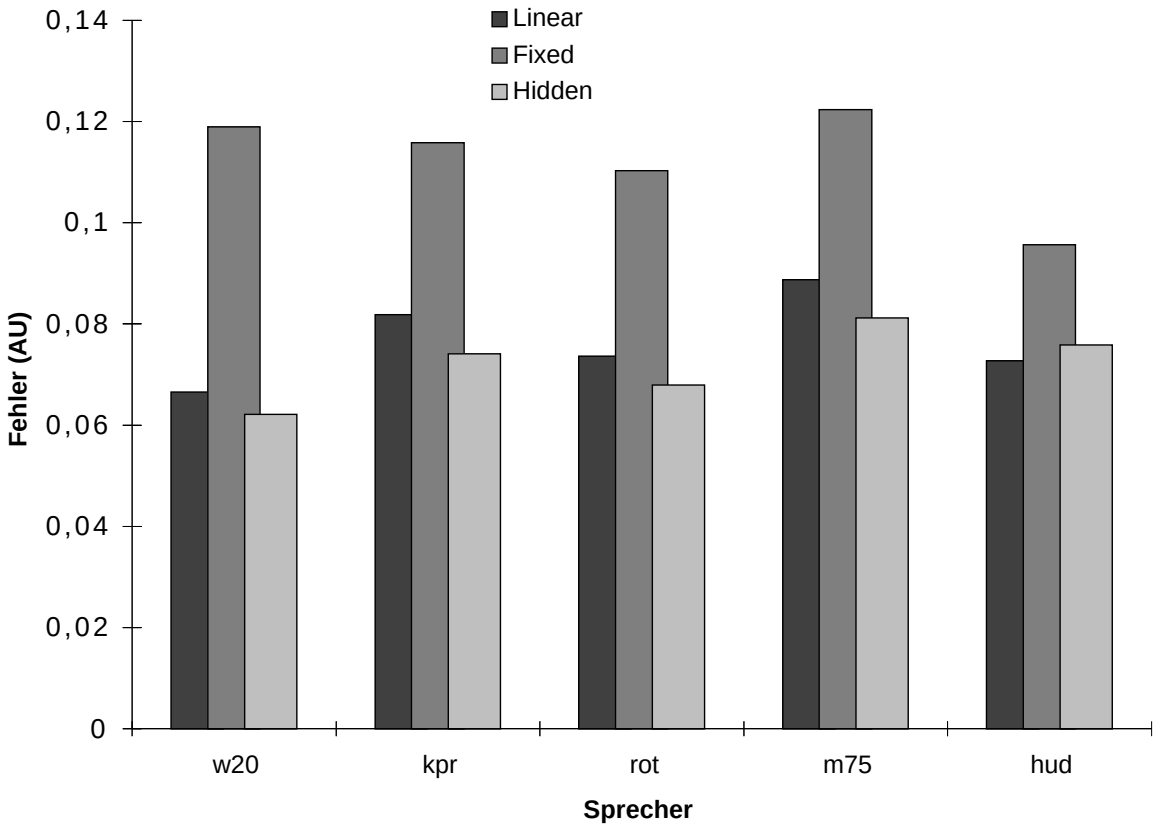

Abbildung 4.3.: Mittlerer Abbildungsfehler über 13 Einzelwortäußerungen für 5 Sprecher, jeweils für ein einfaches Perzeptron (Linear)*, für ein Perzeptron mit 11 versteckten Zellen* (Hidden) *und für ein Perzeptron mit 11 versteckten Zellen, welches für eine Abbildung* weg-weg *trainiert und dann unter Festhalten der Gewichte zwischen Zwischen- und Ausgabeschicht für die jeweiligen Zielsprecher nachtrainiert worden ist* (Fixed)*.*

trogramms nicht immer auch verschiedenen Merkmalsvektoren des anderen Spektrogramms zugeordnet (vgl. Abbildung A.1). Daher muß man sich entscheiden, ob man Mehrfachzuordnungen verwirft oder mittelt (das eine Spektrogramm also kürzt), oder ob man sie durch Wiederholung des einzelnen Zielvektors auflöst (und damit das andere Spektrogramm verlängert). Hier wurde das zweite Vorgehen gewählt, welches ebenfalls insgesamt zu besseren Ergebnissen führte. Wiederum wurden reihum 12 Äußerungen zum Training und eine zum Testen der Abbildung verwendet. Die resultierenden Abbildungsfehler sind in Abbildung 4.3 aufgetragen. Hier sind die Ergebnisse – außer für den Sprecher *hud* – für das mehrschichtige Perzeptron besser (*Hidden:* $\bar{x} = 0{,}072 \pm 0{,}007$) als für das einfache (*Linear:* $\bar{x} = 0{,}077 \pm 0{,}009$).

In der gleichen Abbildung sind auch die Ergebnisse eines weiteren Experimentes *(Fixed)* aufgetragen. Hier wurde untersucht, ob sich bei einem Perzeptron mit einer versteckten Schicht in der Zwischenschicht eine Darstellung etabliert, die anhand der Aktivitäten in der Eingabeschicht prototypische Spektrogramme an der

Ausgabeschicht emittiert. Dazu wurde – wie oben beschriebene – ein Netz mit 11 versteckten Zellen auf die identische Abbildung des Zielsprechers *weg* trainiert. Anschließend wurden die Gewichte zwischen der Zwischenschicht und der Ausgabeschicht festgehalten und die Eingabeschicht für die jeweiligen Zielsprecher nachtrainiert. Die Ergebnisse sind allerdings durchwegs so deutlich schlechter als bei den anderen Netzen ($\bar{x} = 0{,}113 \pm 0{,}010$), daß man schließen muß, daß die Codierung der Spektren nicht allein in der Gewichtsmatrix zwischen der Zwischen- und der Ausgabeschicht geleistet wird.

Die besseren Ergebnisse bei der Verwendung mehrschichtiger Netze zeigen, daß auch bei der relativ geringen verwendeten Datenmenge keine Probleme wegen mangelhafter Generalisierungsfähigkeit auftraten. Die stetige Abnahme des Abbildungsfehlers bis hinauf zu 21 versteckten Neuronen deutet darauf hin, daß durch das Training keine lokalen Minima in der Fehlerlandschaft der mehrschichtigen Netze gefunden wurden. Insgesamt scheint sich jedoch der Mehraufwand für das Training mehrschichtiger Netze kaum zu lohnen. Für die mehrschichtigen Netze muß die Hypothese einer Codierung von prototypischen Spektren in den Gewichten zwischen Zwischen- und Ausgabeschicht abgelehnt werden.

4.4. Abbildung durch Interpolation zwischen Merkmalsvektoren verschiedener Sprecher

In Abschnitt 3.3 wurde bereits darauf eingegangen, daß es wünschenswert wäre, eine Sprechernormalisierung *ohne* explizite Kenntnis des in Frage stehenden Sprechers durchführen zu können. In diesem Fall kann die normalisierende Abbildung nicht auf bereits gesammeltem Sprachmaterial des Quellsprechers gebildet werden, da dieser ja nicht identifiziert werden kann. Statt dessen muß die sprecherspezifische Information zunächst allein aus dem hereinkommenden Sprachsignal gewonnen werden. Man kann sich aber behelfen, indem man versucht, den Sprecher als einem bekannten Sprecher »ähnlich« zu bewerten und eine bereits existierende Abbildung dieses bekannten Sprechers für den unbekannten Sprecher zu verwenden. Dies impliziert eine Nachbarschaftsbeziehung und ein Abstandsmaß zwischen den Sprechern. Ein Abstandsmaß, welches sich sofort anbietet, wäre beispielsweise der euklidische Abstand der mittleren Spektren. Da es hier aber um automatische Spracherkennung geht, scheint die Wahl eines aus der Spracherkennung entlehnten Maßes angemessener. Daher wird hier der in Abschnitt A.1 erklärte DTW-Abstand verwendet.

4.4.1. Analyse der Nachbarschaftsbeziehungen zwischen Sprechern durch multidimensionale Skalierung

Um einen Eindruck über die Struktur der Nachbarschaftsbeziehung zu erhalten, wurde folgendermaßen verfahren: Von 137 Sprechern standen auf der PhonDat-Datenbasis je 12 Einzelwort-Aufnahmen zur Verfügung. Zwischen den jeweils 137 gleichlautenden

Wörtern wurde sprecherpaarweise der DTW-Abstand berechnet. Die so erhaltenen Abstände (hier Fehlerquadrate) wurden über die 12 Wörter gemittelt, um auf diese Weise eine Distanzmatrix[10] der Sprecher zu gewinnen. Aus einer Distanzmatrix kann man durch eine multidimensionale Skalierung (MDS) eine Darstellung der Lage der Sprecher im durch die Abstände definierten Raum erhalten.

Durch MDS behandelt man Probleme der folgenden Art: Gegeben seien n Objekte, die zueinander in irgendeiner Ähnlichkeitsbeziehung stehen. Gesucht ist nun eine möglichst getreue q-dimensionale Darstellung ($q \leq n$) der n Objekte. Die Methoden der MDS können auf eine Vielzahl von Fragestellungen angewendet werden. Diese können so einfache Ähnlichkeitsbeziehungen beinhalten wie die Entfernungen zwischen Städten, oder so schwer faßbare wie etwa die unterschiedlichen »Blumigkeiten« von Weinen, die Tester auf einer subjektiven Skala vergeben könnten.

Die durch die MDS gefundenen Räume lassen sich häufig schwer interpretieren. Im Beispiel mit der Entfernungstabelle ist die Interpretation klar: Das Ergebnis einer zweidimensionalen MDS wäre eine Art Landkarte, die allerdings nicht orientiert sein würde, da sich aus den Abständen nicht auf die Himmelsrichtungen schließen läßt. Bei den Weinen hingegen würde eine zweidimensionale MDS eine Projektion auf eine Hyperebene durch einen Raum subjektiver Bewertungen (»Blume«, »Frucht«, »Ausdruck« o. ä.) erzeugen, deren Achsen irgendwelche Linearkombinationen dieser vermutlich stark korrelierten Parameter wäre.

Die Algorithmen zur MDS können in metrische und nichtmetrische Methoden unterteilt werden. Metrische MDS erfordert eine Abstandsmatrix der Objekte, welche einer Metrik[11] genügt. Gewöhnlich wird von euklidischen oder wenigstens von Minkowskiabständen[12] ausgegangen. Die nichtmetrischen Algorithmen benötigen keine Abstände, sondern arbeiten auch auf Ähnlichkeitsmaßen, welche nicht die Eigenschaften einer Metrik erfüllen, wie z. B. Rangfolgen und/oder unvollständigen Matrizen. Da durch das oben beschriebene Verfahren eine Abstandsmatrix für die Beziehung zwischen den Sprechern angegeben werden kann, ist hier eine metrische MDS-Methode ausreichend. Verwendet wurde die denkbar einfachste Methode, nämlich der Algorithmus von Torgerson, der im Anhang A.2 dargestellt wird.

[10] Die Elemente d_{ij} einer Distanzmatrix $\mathbf{D}$ geben jeweils den Abstand $d(x_i, x_j)$ zwischen zwei Punkten (hier: Sprechern) x_i und x_j an.

[11] Eine *Metrik* ist eine Abbildung $d : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ mit

1. $d(\mathbf{x}, \mathbf{y}) = 0 \iff x = y$
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$

[12] Minkowski- oder L_p-Metriken sind solche der Form

$$d_p(\mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_i |x_i - y_i|^p}, \qquad p \in \mathcal{R}, \quad p > 0 \tag{4.12}$$

Als Spezialfall ergibt sich für $p = 1$ die sog. *City-Block-Metrik*, für $p = 2$ die *euklidische Metrik*, und für $\lim_{p \to \infty} d_p$ definiert man die *Maximums-* oder *tschebyscheffsche Metrik*.

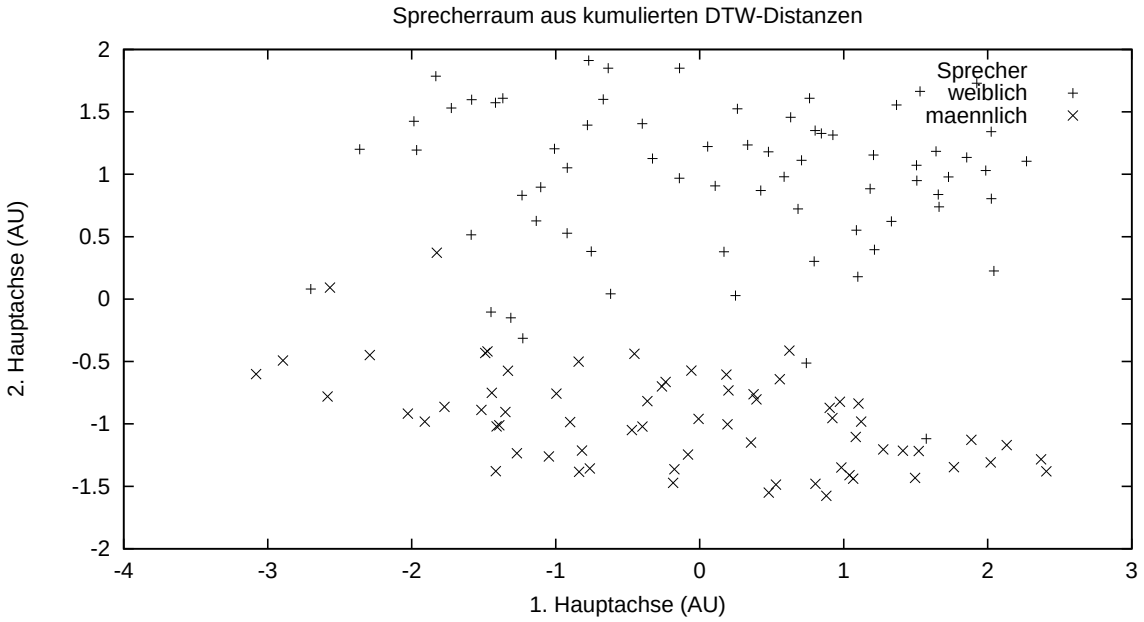

Abbildung 4.4.: Ergebnis einer zweidimensionalen MDS der DTW-Abstände. Die Symbole unterscheiden männliche und weibliche Sprecher.

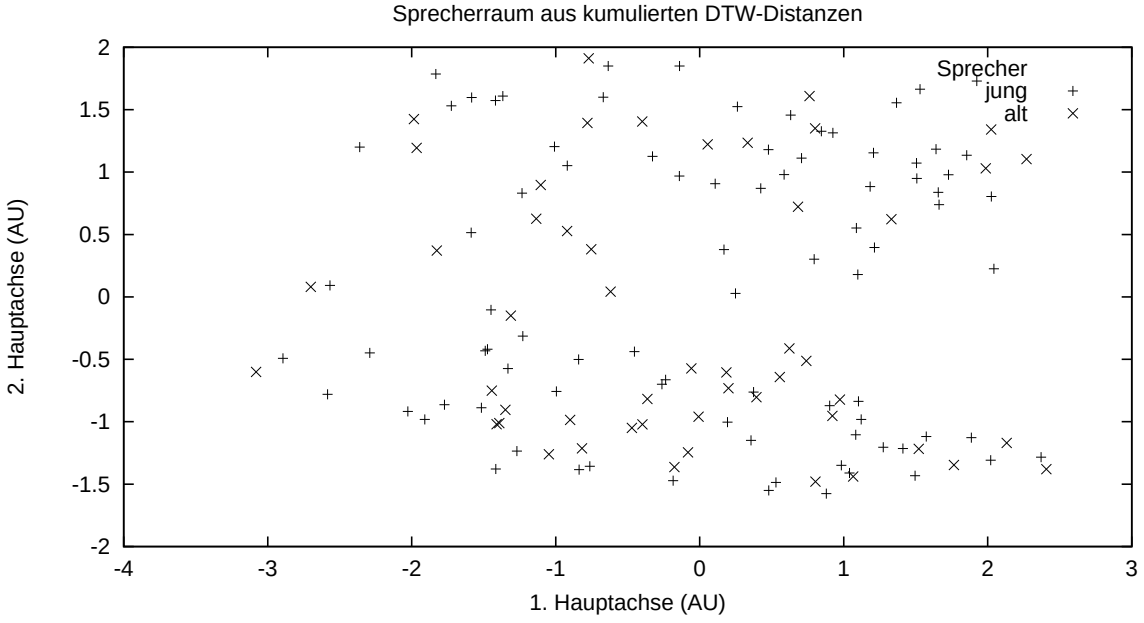

Abbildung 4.5.: Ergebnis einer zweidimensionalen MDS der DTW-Abstände. Die Symbole unterscheiden jüngere und ältere Sprecher.

Wie erwähnt, sind multidimensionale Skalierungen generell schwierig zu interpretieren. Die PHONDAT-1-Datenbasis informiert über zwei Eigenschaften der aufgenommenen Sprecher: Das *Geschlecht* und das *Alter*. Das Alter wird allerdings nur als *jung* oder *alt* angegeben, wobei unbekannt ist, ab welchem Alter ein Sprecher als *alt* gilt. Beide Kategorien wurden in den Abbildungen 4.4 und 4.5 eingetragen. Man erkennt, daß die Projektion die Sprecher gut nach ihrem Geschlecht trennt, so daß dieses offenbar auch für einen großen Teil der Streuung in diesem Raum verantwortlich ist. Das Alter der Sprecher scheint andererseits keinen wesentlichen Einfluß auf die Ähnlichkeit der Sprecher zueinander zu haben.

Ein anderer Gewinn, der aus der Analyse der Abstände gezogen werden kann, ist ein Kriterium für die Wahl eines (oder mehrerer) geeigneter Zielsprecher, denn dieser sollte derjenige Sprecher sein, der im Mittel zu allen anderen Sprechern den geringsten Abstand aufweist. Im beschriebenen Experiment war dies ein männlicher Sprecher mit dem Kürzel *weg*.

4.4.2. Interpolation der Abbildungen

Um nun die Hypothese der Interpolierbarkeit von sprechernormalisierenden Abbildungen zu erfüllen, muß zunächst eine Nachbarschaftsbeziehung zwischen den verschiedenen Sprechern bestehen. Diese wurde in der in Abschnitt 4.4.1 beschriebenen Weise bestimmt. Anhand der durch die MDS ermittelten Koordinaten im »Raum der Sprecher« wurden nun Sprecher gesucht, die sich innerhalb dieses Raumes – vom Zielsprecher *weg* aus – annähernd auf einer Linie befinden. Hierfür wurden drei Sprecherpaare gefunden: Die Sprecher *cha* und *erl*, *was* und *alz*, und die Sprecher *cha* und *swt*, wobei der zweite Sprecher immer der von Zielsprecher aus gesehen entferntere Sprecher ist.

Für die entfernteren (»äußeren«) Sprecher wurden auf der Basis der Einzelwortäußerungen einfache Perzeptrone als Abbildung auf den Zielsprecher *weg* trainiert. Da sich inzwischen herausgestellt hatte, daß der DTW-Algorithmus auf logarithmierten Daten besser arbeitet, wurden die Merkmalsvektoren der Sprecher diesmal logarithmiert. Um für die anschließende Mischung der Abbildungen Artefakte durch die nichtlineare Ausgangskennlinie der Netzwerk-Neuronen zu vermeiden, wurde außerdem eine identische Abbildung des Zielsprechers *weg* auf sich selbst trainiert.

Anschließend wurden die Einzelwortäußerungen der dem Zielsprecher *weg* nähergelegenen (»inneren«) Sprecher durch die Netze sowohl des Zielsprechers als auch des zugehörigen äußeren Sprechers abgebildet. Die so gewonnenen Datensätze wurden linear interpoliert und das Ergebnis in einem DTW-Erkennungsexperiment mit den Datensätzen des Zielsprechers verglichen. Dabei hat sich gezeigt, daß die DTW-Erkennung besser auf dem Skalarprodukt der Merkmalsvektoren als Fehlermaß als auf dem quadratischen Fehler funktioniert (mutmaßlich wegen der Unempfindlichkeit dieses Maßes gegen Skalierungen), so daß dieses Fehlermaß verwendet wurde.

Bei diesem Experiment stellte sich das Problem, daß die Erkennungsraten auch ohne Abbildung auf dem gegebenen kleinen Wortschatz bereits 100 % betrugen. Daher wurde ein anderes Qualitätskriterium für die Güte der Abbildung verwendet, und zwar

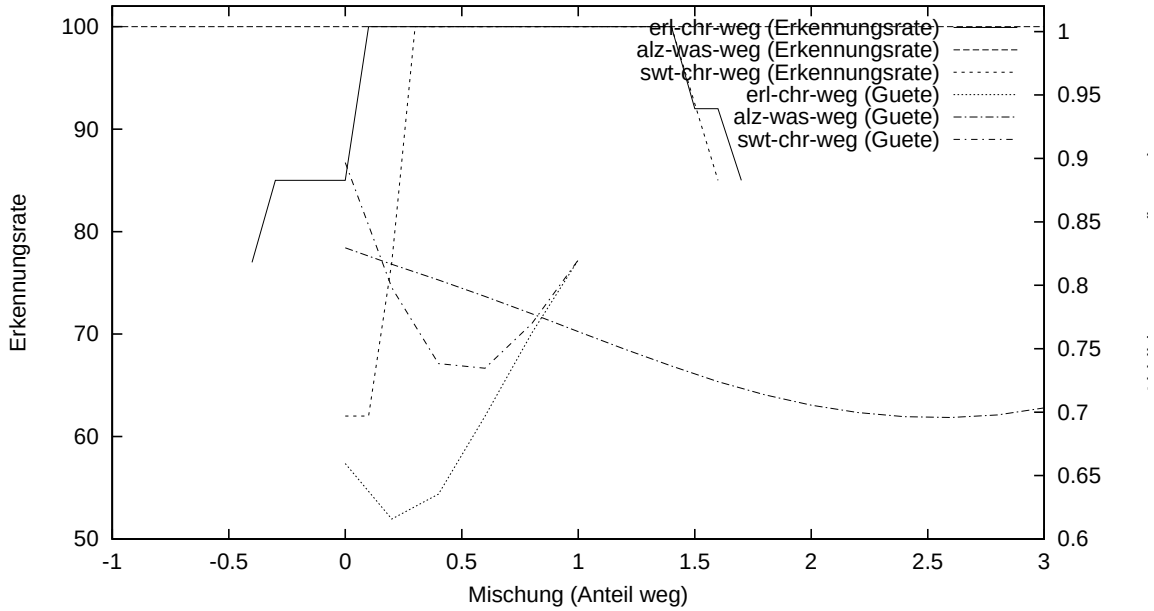

Abbildung 4.6.: Erkennungsraten und inverse Güte der Abbildung für drei Quellsprecher. Die Abbildung bestand jeweils aus einer Mischung zweier sprechernormalisierender Abbildungen auf Sprecher, »zwischen« denen sich die Quellsprecher im Raum der Sprecher jeweils befanden. Das Mischungsverhältnis der Abbildungen wurde variiert und ist an der Abszisse aufgetragen. Lineare Interpolation hätte alle Extrema näher bei Null erwarten lassen.

das Verhältnis des DTW-Abstandes zur richtigen Äußerung zu dem des nächstbesten Kandidaten, was ein Maß für die Sicherheit der Erkennung darstellt.

Die Ergebnisse dieses Experimentes sind in Abbildung 4.6 aufgetragen. Insgesamt ergab sich folgendes Bild:

1. Die Güte der interpolierten Abbildung war immer größer als die der beiden »Randabbildungen«, d. h. der des Zielsprechers oder der des äußeren Sprechers für sich.

2. Die Maxima der Güte der gemischten Abbildungen lagen nicht an der Stelle, wo sie eine lineare Interpolation erwarten ließe.

Daraus ergibt sich, daß

1. eine Interpolation von Abbildungen zwischen Sprechern möglich ist,

2. diese aber nur in schlechter Näherung linear angesetzt werden kann.

4.4.3. Bestimmung der Mischungen gegebener Abbildungen für unbekannte Sprecher

Da nun gezeigt werden konnte, daß sich gegebene Abbildungen für unbekannte Sprecher mit Erfolg interpolieren lassen, stellt sich die Frage, ob bzw. wie sich diese Mischungen automatisch ermitteln lassen. Der Gedanke liegt nahe, auch hierfür Perzeptrone zu verwenden. Die Eingangsdaten für das Perzeptron sind wieder Barkspektrogramme, die Ausgabemuster sollten Mischungskoeffizienten für die gegebenen Zielsprecher sein. Das Problem hierbei liegt nun in der Beschaffung der zu trainierenden Zielmuster, da die Mischungen unbekannt sind.

Der Lösungsansatz besteht nun darin, das Perzeptron mit den Mustern gegebener Zielsprecher zu trainieren, da für diese das Mischungsverhältnis durch den Ansatz als euklidischer Einheitsvektor gegeben ist, d. h. in diesen Fällen muß das Ergebnis lauten, daß die Abbildung des jeweiligen aktuellen Zielsprechers mit Gewichtung »1«, die anderen Abbildungen mit dem Gewicht Null verwendet werden sollen; das Netz wird gewissermaßen als Sprechererkenner trainiert. Nach den Ergebnissen des vorherigen Experimentes sollte dieser Ansatz dann auch für unbekannte Sprecher zu geeigneten Mischungen führen, sofern die Zielsprecher den Raum der Sprecher genügend gut abdecken und das Netz genügend gut generalisiert, zumal die Maxima der Güte der Abbildung hinreichend breit sind (Abbildung 4.6).

Es wurden verschiedene Vorversuche durchgeführt, bei denen mit logarithmierten und nichtlogarithmierten Daten, einfachen und mehrschichtigen, *Time-Delay-* (Lang et al., 1990) und rekurrenten (Olurotimi, 1994) Perzeptronen sowie unterschiedlichen Auswahlen von Lautklassen (stimmhafte Laute, vokalische Laute) experimentiert wurde. Die Ergebnisse waren insgesamt unbefriedigend. Brauchbare Resultate ergaben sich erst bei der Beschränkung auf einzelne Vokale. Um hierfür geeignete Mengen an Trainingsmaterial zu erhalten, mußte eine neue Sprecherauswahl getroffen

Tabelle 4.3.: Klassifikationsfehler der Sprechererkennung bei verschiedenen Vokalen in Prozent, jeweils auf der Trainingsmenge und einer unabhängigen Testmenge.

Datenmenge	Vokal					
	/aː/	/eː/	/iː/	/oː/	/uː/	/əː/
Trainingsmenge	1,2	1,8	1,5	0,7	1,9	12,8
Testmenge	3,8	5,7	7,9	6,9	10,5	19,8

werden, da die PhonDat-1-Datenbank nur sechs Sprecher enthält, die einen großen (und gleichen) Teil des Korpus gesprochen haben. Diese Sprecher sind *gvn* (weiblich), *hei* (männlich), *kko* (männlich), *m02* (männlich), *ror* (männlich) und *rtd* (weiblich). Für alle diese Sprecher wurden anhand der Phonemlabels diejenigen Signalanteile ausgesucht, die als einer der Randvokale /aː/, /eː/, /iː/, /oː/ oder /uː/, oder als der neutrale Laut *Schwa* (/ə/) bezeichnet sind.[13] Diese Daten wurden in zwei willkürliche Teilmengen aufgeteilt, und zwar in diejenigen, die aus den Äußerungen bis Nr. 200 stammen, und in diejenigen, die den Äußerungen ab Nr. 201 entstammen.

Die in Tabelle 4.3 zusammengestellten Ergebnisse zeigen den Klassifikationsfehler einfacher Perzeptrone, welche in der beschriebenen Weise auf die Zuordnung der Vokalspektrogramme zu den sechs Zielsprechern trainiert wurden. Dabei wurde für jeden Vokal ein eigenes Netz verwendet. Die Spektrogramme wurden logarithmiert und zeitgemittelt. Trainiert wurde jeweils mit einer der Teilmengen. Getestet wurde sowohl auf der Trainingsmenge als auch auf der zweiten Teilmenge. Die Ergebnisse wurden über beide Trainings- bzw. Testmengen gemittelt. Wie man erkennt, sind die Ergebnisse auf den unabhängigen Testmengen deutlich schlechter. Die besten Resultate auf der Trainingsmenge wurden für den hinteren halbgeschlossenen Vokal /oː/ erreicht. Allerdings sollten für die Verwendung zur Sprechererkennung die unabhängigen Testmengen maßgeblich sein. Die besten Ergebnisse wurden hier für den vorderen offenen Eckvokal /aː/ erzielt. Am schlechtesten sind die Werte für den neutralen Laut /ə/. Dies ist einerseits überraschend, da dieser Laut definitionsgemäß ohne Spannung gesprochen wird und insofern für den Vokaltrakt des jeweiligen Sprechers charakteristisch sein sollte. Andererseits hat dieser Laut immer nur eine kurze Dauer, so daß sich hier Koartikulationseffekte verstärkt niederschlagen, und da der Laut entspannt gesprochen wird, sollte er keine sprecherspezifische Dialekt- oder Akzentfärbung aufweisen. Desweiteren kann nicht ausgeschlossen werden, daß dieser Kategorie beim Labeling auch weniger ausgeprägte Vokale zugeschlagen wurden, die nicht eigentlich ein Schwa sind. Insgesamt kann eine Sprechererkennung unter den verwendeten Lauten also am sichersten auf den /aː/-Segmenten durchgeführt werden.

[13] Zu den Vokalen vgl. Abbildung B.1.

Tabelle 4.4.: Klassifikationsfehler der Sprechererkennung bei Verwendung verschiedener Merkmalsvektoren für die Vokalen /a:/ und /e:/ in Prozent, jeweils auf der Trainingsmenge und einer unabhängigen Testmenge.

Datenmenge	nur Spektrum Vokal		Spektrum & AP Vokal		nur AP Vokal	
	/a:/	/e:/	/a:/	/e:/	/a:/	/e:/
Trainingsmenge	1,2	1,8	0,6	1,0	38,5	25,4
Testmenge	3,8	5,7	3,5	4,4	41,0	35,1

Einbeziehung artikulatorischer Parameter

Im nächsten Experiment sollten zusätzlich zu den bisher allein verwendeten spektralen Parametern artikulatorische Parameter (AP) in die Merkmalsvektoren aufgenommen werden, wovon man sich eine Verbesserung der Klassifikationsleistung versprechen konnte. Bei diesen Parametern handelt es sich um geometrische Merkmale, die die Skalierung des Vokaltraktes im Vergleich zu einem Referenzsprecher beschreiben. Sie wurden unter Verwendung eines Vokaltraktmodells anhand der Formatlagen innerhalb der Spektren des Sprachsignals geschätzt. Zwei der Parameter beschreiben eine stückweise lineare Längenskalierung, vier weitere Parameter Querschnittsskalierungen. Das Verfahren wird ausführlich in (Freienstein, 2000) beschrieben.

Diese sechs Parameter wurden für die Vokale /a:/ und /e:/ ermittelt und als zusätzliche Elemente an den bisherigen Merkmalsvektor angefügt. Das sonstige Vorgehen blieb unverändert. Die Ergebnisse sind in Tabelle 4.4 dargestellt. Zum Vergleich sind die Ergebnisse des Experimentes ohne AP noch einmal aufgeführt, sowie die Resultate des gleichen Experimentes unter ausschließlicher Verwendung der AP. Die durch die Hinzunahme der AP erzielte Verbesserung muß mit einer gewissen Vorsicht betrachtet werden, da hier auf der Seite des Vokaltraktschätzers eine gewisse Vorauswahl der Signalabschnitte stattgefunden hat, auf denen das Experiment durchgeführt wurde. Außerdem ist die zugrundeliegende Datenmenge zu gering, um den Gewinn als statistisch signifikant bewerten zu können. Andererseits zeigen die Ergebnisse bei der ausschließlichen Verwendung der AP, daß diese durchaus sprecherspezifische Informationen tragen. (Die Ratewahrscheinlichkeit liegt hier bei 16,7 %.) Das die Ergebnisse nicht besser ausfallen, kann darauf zurückgeführt werden, daß die verwendeten neuronalen Netze selbst in der Lage sind, die in den AP kodierten Sprecherspezifika aus den Barkspektrogrammen zu extrahieren.

4.4.4. Abbildung der Vokale

Im folgenden soll die automatische Bestimmung einer Abbildung für einen unbekannten Sprecher untersucht werden. Reihum wurde einer der 6 Referenzsprecher *(gvn, hei, kko, m02, ror* und *rtd)* als Zielsprecher der Abbildung verwendet. Wie im letzten Abschnitt beschrieben wurden jeweils Sprechererkennungs-Netzwerke für die übrigen 5

Referenzsprecher trainiert. Abbildung 4.6 kann man entnehmen, daß auch Mischungskoeffizienten größer 1 unter Umständen sinnvoll sind, so daß hier neben Fermifunktionen auch lineare Aktivierungsfunktionen für die Neuronen der Ausgangsschicht gewählt wurden. Für das Sprechererkennungs-Perzeptron wurde eine versteckte Schicht mit 10 Zellen verwendet, die nach wie vor sigmoide Transferfunktionen besaßen.[14]

Als eigentliche normalisierende Abbildungen wurden ebenfalls Perzeptrone mit 10 versteckten Zellen gewählt, die jeweils darauf trainiert worden waren, die Spektren aller benannten Vokale eines der Referenzsprecher auf die Spektren gleichlautender Vokale eines anderen Referenzsprechers zu transformieren. Die Vokale aller 159 Sprecher der PhonDat-1-Datenbasis, für die die gewünschten Vokale vorhanden waren und die nicht zu den 6 Referenzsprechern gehörten, wurden wie oben beschrieben extrahiert. Da hier nicht – wie oben – eine DTW-Erkennung durchgeführt werden konnte, bestand die Erkennungsaufgabe nun darin, die Vokale der ersten Teilmenge (bis Äußerung 200) aller 159 Sprecher durch eine lineare (Fishersche) Diskriminanzanalyse (vgl. Anhang A.4) zu trennen.

Hierfür mußte zuallererst die Trennbarkeit der Vokale *ohne* Abbildung untersucht werden, denn dieses Maß sollte durch die Normalisierung nicht unterschritten werden. Anschließend wurden die 5 Referenzsprecher selbst mit den für sie berechneten Abbildungen reihum auf den sechsten Sprecher abgebildet. Dies sollte das Maß sein, welches von der Abbildung der unbekannten Sprecher nicht überschritten werden kann. Als letzte Vergleichsgröße wurden die Vokale der 6 Referenzsprecher jeweils einzeln direkt einer Diskriminanzanalyse unterzogen. Durch den Vergleich der Diskriminierbarkeit zwischen den abgebildeten Vokalen der Referenzsprecher und der direkten Analyse kann die Qualität der Abbildung beurteilt werden.

Nun wurden die Vokale der 159 Sprecher durch das Sprechererkennungsnetzwerk bewertet. Dessen Ausgangssignal wurde für alle Vokale getrennt gemittelt. Dadurch wurden Gewichtsvektoren gebildet, die jeweils die Ähnlichkeit des aktuellen Testsprechers zu den 5 Referenzsprechern wiederspiegelt. Die Spektren der Vokale der 159 Sprecher wurden durch die 5 Abbildungen der Referenzsprecher auf den Zielsprecher transformiert. Die resultierenden 5 transformierten Spektren wurden entsprechend der Bewertung durch die ihrer Vokalklasse zugehörigen Gewichtsvektoren aufaddiert. Die so gewonnenen normalisierten Spektren wurden ebenfalls einer Diskriminanzanalyse unterzogen.

Die Ergebnisse sind in Tabelle 4.5 zusammengestellt. Wie man erkennt, wird durch die vom Sprechererkennungs-Netzwerk ermittelte Mischung der Abbildungen der 5 Referenzsprecher die Diskriminierung der Vokale der 159 Testsprecher deutlich verbessert, und zwar soweit, daß sogar die Werte für die Diskriminierung der Referenzsprecher selbst überschritten wird. Der Adaptionsgrad (Gleichung 4.9) für das Kontrollnetz ohne lineare Ausgangskennlinie liegt noch deutlich vor dem des Netzes

[14] Man kann sich leicht überlegen, daß dies immer so sein muß, denn wenn die Aktivierungsfunktionen der versteckten Schicht(en) linear sind, erkennt man an den Gleichungen A.14 und A.15, daß man anderenfalls effektiv ein einfaches Perzeptron konstruiert. Sind – wie hier – die Ausgabeaktivierungsfunktionen ebenfalls linear, reduziert sich das Ganze auf eine affine Abbildung.

Tabelle 4.5.: Ergebnisse der FISHER*schen Diskriminanzanalyse der Vokalbarkspektrogramme. Die Sprechergruppe* 159 *bezeichnet alle untersuchten Sprecher ohne die 6 Referenzsprecher, Die Gruppe* 6 *die der Referenzsprecher unter sich. Die Behandlung »—« bezeichnet eine direkte Diskriminanzanalyse der Merkmalsvektoren der ganzen Sprechergruppe ohne Normalisierung, »einzeln« das Gleiche mit den Merkmalsvektoren nur jeweils eines einzigen Sprechers, »(nicht)linear« eine Normalisierung mit Steuerung der Abbildung durch ein Netz mit (nicht)linearen Ausgangskennlinien.*

Experiment		Güte der Diskrimination	Adaptionsgrad α
Sprecherset	Behandlung	%	%
159	—	70,72	
159	linear	91,41	150,4
159	nichtlinear	97,66	195,8
6	einzeln	84,48	
6	—	77,85	
6	linear	92,52	221,3
6	nichtlinear	92,20	220,7

mit linearer Ausgangskennlinie. Dies deutet darauf hin, daß durch das Verfahren der Mischung eine Interpolation der Abbildungen eher möglich ist als eine Extrapolation.

Für die Referenzsprecher selbst ist der Adaptionsgrad noch größer, allerdings zeigt hier das nichtlineare Kontrollnetz keine besseren Ergebnisse. Betrachtet man sich die Ergebnisse im einzelnen, dann sieht man, daß dies im wesentlichen auf die schlechteren Ergebnisse für die Sprecherin *rtd* zurückzuführen ist. Es kann vermutet werden, daß diese Sprecherin nur durch eine Extrapolation erreicht werden kann, sie also relativ weit vom Schwerpunkt der Verteilung entfernt ist. Dies erscheint auch daher plausibel, als sich diese Sprecherin auch bei anderen Experimenten als Ausnahmefall erwiesen hat (Freienstein, 2000).

4.5. Zusammenfassung und Diskussion der Ergebnisse

In Abschnitt 4.2 wurde eine Methode zur unüberwachten Sprechernormalisierung auf der Basis einer Hauptkomponentenanalyse der Merkmalsvektoren von Quell- und Zielsprecher vorgestellt. Schwierigkeiten ergeben sich bei dieser Methode daraus, daß die gefundenen Hauptkoordinaten der beiden Sprecher einander nicht entsprechen müssen. Daher wurden nur die beiden größten Hauptkoordinaten berücksichtigt. Desweiteren ist bei der gewählten Vorverarbeitung die Annahme einer Gaußverteilung der Merkmalsvektoren nicht erfüllt. Dennoch konnte insgesamt ein Adaptionsgrad von 20 % erreicht werden.

Abschnitt 4.3 beschrieb die sprechernormalisierende Abbildung durch Perzeptro-

ne. Es wurden Netze mit und ohne versteckte Schicht verwendet. Der Abbildungsfehler konnte durch mehrschichtige Netze nicht nennenswert verringert werden. Die aufgestellte Hypothese, daß sich in der Zwischenschicht eines mehrschichtigen Perzeptrons eine niedrigdimensionale Repräsentation des Sprecherraumes etablieren würde, wurde untersucht und mußte verworfen werden.

In Abschnitt 4.4 ging es nun um die Ermittlung eines »Sprecherraumes« auf der Basis einer multidimensionalen Skalierung von gemittelten DTW-Abständen. So konnten Nachbarschaftsbeziehungen zwischen Sprechern und – in der Mitte der Verteilung – »Normalsprecher« gefunden werden. Dabei zeigte sich, daß der Sprecherraum eine deutliche Separierung zwischen männlichen und weiblichen Sprechern wiedergibt, wohingegen das Alter der Sprecher offenbar keinen großen Einfluß auf die Anordnung der Sprecher hat.

Im Unterabschnitt 4.4.2 wurde dann untersucht, ob sich die so gefundenen Nachbarschaftsbeziehungen mit Gewinn für die Abbildung unbekannter Sprecher nutzen lassen. Hierfür wurden im Sprecherraum zwei Sprecher gesucht, »zwischen« denen sich der unbekannte Sprecher befand, und die Merkmalsvektoren des unbekannten Sprechers durch die gegebene Abbildungen dieser zwei Referenzsprecher abgebildet. Die Ergebnisse wurden mit verschiedenen Gewichten addiert. Da die Erkennungsraten in diesem Experiment bei 100 % lagen wurde ein Trennmaß definiert, daß die Qualität der Abbildung beschreibt. Dieses konnte durch das gewählte Vorgehen gesteigert werden. Es zeigte sich aber auch, daß die Maxima des Trennmaßes gegenüber der durch lineare Interpolation gefundenen Gewichte verschoben war. Dies weist darauf hin, daß der Sprecherraum nur lokal linear interpoliert werden kann.

Die automatische Positionsbestimmung eines unbekannten Sprecheres durch mehrschichtige Perzeptrone wurde in Unterabschnitt 4.4.3 vorgestellt. Hierzu wurden Perzeptrone als Sprechererkenner für eine Auswahl von sechs Referenzsprechern trainiert. Für die sechs Sprecher ergab sich dabei eine Erkennungsrate von 96,2 % auf der Testmenge bei Verwendung des Vokals /aː/. Da hierfür zunächst eine Erkennung des /aː/ erforderlich ist, handelt es sich um ein *Bootstrap*-Verfahren mit den in Abschnitt 4.2 geschilderten Gefahren. Auf der anderen Seite ist die sprecherunabhängige Erkennung von Vokalen relativ unproblematisch. Abgesehen davon können für die Zwecke der Sprechererkennung alle Zweifelsfälle verworfen werden.

Im gleichen Abschnitt wurde auch die Einbeziehung von artikulatorischen Parametern zur weiteren Steigerung der Erkennungsleistung untersucht. Die dabei erreichte Verbesserung ist allerdings nicht signifikant. Hier liegt die Vermutung nahe, daß die verwendeten Netzwerke die gleiche Information bereits aus den Merkmalsvektoren extrahieren konnten.

Unterabschnitt 4.4.4 beschreibt nun die Verwendung der Sprechererkennungsnetze zur Normalisierung unbekannter Sprecher. Dabei wird die Ausgabe der im Prinzip 1-aus-N-kodierten Sprechererkenner für unbekannte Sprecher als Gewichtung der Abbildungen von fünf der sechs Referenzsprecher auf den sechsten – als Zielsprecher – betrachtet. Für das Sprechererkennungsnetzwerk wurden sowohl lineare wie sigmoide Ausgangskennlinien betrachtet. Die Merkmalsvektoren der abgebildeten Vokale wurden schließlich einer linearen Diskriminanzanalyse unterworfen. Der Adaptionsgrad

betrug für alle 159 verfügbaren Sprecher 150 % bei linearen und 196 % bei sigmoiden Ausgangskennlinien. Das bessere Abschneiden der Netze mit nichtlinearen Aktivierungsfunktionen kann man darauf zurückführen, daß durch den gewählten Ansatz negative Ausgangswerte oder solche größer Eins nicht trainiert wurden und die Netze in diesem Bereich ungeeignete Mischungen erzeugen. Da andererseits für die Referenzsprecher unter sich die Ergebnisse kaum voneinander verschieden waren (der Adaptionsgrad betrug hier jeweils 221 %), liegt die Vermutung nahe, daß einer der sechs Sprecher (*rtd*) nur durch eine Extrapolation erreichbar gewesen wäre.

Insgesamt belegen diese Ergebnisse in beeindruckender Weise die Hypothese von der Interpolierbarkeit der sprechernormalisierenden Abbildungen, selbst bei der Abdeckung des Sprecherraumes mit nur 6 Referenzsprechern.

5. Zusammenfassung und Ausblick

5.1. Zielsetzung

Diese Arbeit beschäftigte sich mit der Verbesserung der Vorverarbeitung von Sprachsignalen zur Steigerung der Leistung sprecherunabhängiger automatischer Spracherkennungssysteme. Dies geschah durch die Einführung einer Sprechernormalisierungsstufe. Deren Aufgabe ist es, die zusätzliche Streuung, die durch die Berücksichtigung unterschiedlicher Sprecher in die Merkmalsvektoren eingeführt wird, durch Transformation der Merkmalsvektoren auf die eines speziellen Referenzsprechers zu reduzieren. Für die Konstruktion einer solchen Transformation müssen gewisse Charakteristika aus den Merkmalsvektoren des jeweiligen Quellsprechers bestimmt werden.

Da hierfür eine gewisse Datenmenge erforderlich ist, lag das Hauptaugenmerk dieser Arbeit auf der Frage, ob die gewünschte Transformation nicht unter Verwendung gegebener Transformationen bekannter Sprecher gewonnen werden kann. Auf diese Weise wäre eine Datenmenge ausreichend, welche die Auswahl dieser bereits im Vorfeld bestimmten Transformationen erlaubt.

5.2. Zusammenfassung

Nach einer Einführung in die Problematik der automatischen Spracherkennung und der Sprechernormalisierung wurde ein Verfahren zur unüberwachten Sprechernormalisierung auf der Basis einer Hauptkomponentenanalyse vorgestellt, welches im Mittel einen Adaptionsgrad von 20 % erreichte..

Daran anschließend wurden verschiedene Versuche zur Konstruktion einer sprechernormalisierenden Abbildung unternommen. Dabei konnte bestätigt werden, daß die gewünschte Abbildung mithilfe von gerichteten künstlichen neuronalen Netzen (Perzeptronen) erzeugt werden kann. In diesem Zuge wurde auch die Anzahl der versteckten Zellen eines mehrschichtigen Perzeptrons untersucht, die benötigt werden, um eine identische Abbildung zwischen den Spektren eines Sprechers zu konstruieren. Der Abbildungsfehler fiel auch bei größeren Zellenzahlen kontinuierlich mit der Zunahme der Zellenzahl. Da es gleichzeitig nicht möglich war, eine befriedigende Abbildung allein durch Nachtrainieren der Eingabe-/Zwischenschichtgewichte zu erzeugen, kann die ursprüngliche Annahme, daß sich in der Zwischenschicht des Perzeptrons eine niederdimensionale Repräsentation der Merkmale etablieren würde, nicht bestätigt werden. Vielmehr deuten die Ergebnisse darauf hin, daß die Codierung der

Zielspektren im ganzen Netz verteilt erfolgt.

Hierauf wurden Untersuchungen über die Existenz und Struktur von Nachbarschaftsbeziehungen zwischen unterschiedlichen Sprechern angestellt. Die Anordnung der Sprecher im »Raum der Sprecher« wurde aus den Merkmalsvektoren ihrer Äußerungen bestimmt. Die Ergebnisse dieser Untersuchungen belegen die Möglichkeit einer linearen Interpolierbarkeit von sprechernormalisierenden Abbildungen anhand der gefundenen Nachbarschaftsbeziehungen. Es zeigten sich hierbei allerdings Diskrepanzen zwischen der tatsächlichen Lage des Optimums der Interpolation gegenüber der Position, die sich aus der Annahme einer linearen Beziehung zwischen dieser Optimallage und der räumlichen Distanz der Sprecher ergibt. Es muß daher gefolgert werden, daß diese Beziehung nur lokal als linear approximiert werden kann.

Hierauf folgte die Untersuchung der Eignung von Perzeptronen zur Ermittlung der Nachbarschaftsbeziehungen eines Sprechers. Dafür wurde eine Menge von Referenzsprechern ausgewählt, für die ein mehrschichtiges Netz als Sprechererkenner ausgebildet wurde. Dies geschah unter der Annahme, daß sich die Nachbarschaftsbeziehungen eines unbekannten Sprechers zu den trainierten Referenzsprechern auf diese Weise nach Anbieten der zugehörigen Merkmalsvektoren in der Ausgabe des Sprechererkennungsnetzes wiederfinden läßt. Wegen der Gegebenheiten der verfügbaren Datenbasis konnten diese Experimente nur auf einzelnen Lauten und nur für sechs Referenzsprecher durchgeführt werden. Daher wurde eine Auswahl der Vokale der verfügbaren Sprecher extrahiert und sowohl zum Training als auch zum Testen sowohl der normalisierenden Abbildungen als auch des Sprechererkenners verwendet. So konnte nebenbei gezeigt werden, daß ein solcherart konstruiertes Sprechererkennungs-Netzwerk auf den verwendeten Randvokalen eine zuverlässige Sprechererkennung leisten kann.

Zuletzt wurde nun die Eignung der Mischung sprechernormalisierender Abbildungen untersucht, welche durch die Sprechererkenner-Netzwerke geschätzt wurde. Hier konnte durch die gewählte Methode die Klassifizierung der Vokale durch die gefundene Abbildung deutlich verbessert werden.

Insgesamt konnte in dieser Arbeit die Möglichkeit einer Sprechernormalisierung unbekannter Sprecher auf der Basis der Interpolation bekannter normalisierender Abbildungen gezeigt und die Verbesserung der Erkennungsleistungen für unbekannte Sprecher auf der Basis der so gewonnenen Abbildungen belegt werden.

5.3. Ausblick

Für die Rückführung einer sprecherunabhängigen Spracherkennung auf eine sprecherabhängige Spracherkennung zur Verminderung des Klassifikationsfehlers existieren grundsätzlich zwei Methoden: Die Adaption des Klassifikators an den unbekannten Sprecher oder die Transformation der Signale eines unbekannten Sprechers auf die des Referenzsprechers, für den der Erkenner trainiert wurde.

Gegenüber dem sprecherspezifischen Training eines Spracherkennungssystems lohnen sich beide Methoden nur dann, wenn insbesondere der Datenaufwand zur Bestim-

mung der Parameter des Erkennungssystems deutlich höher ist als für die Adaption oder die Normalisierung.

Die vorgestellte Methode zur Interpolation gegebener Abbildungen reduziert den Datenaufwand zur Bestimmung einer normalisierenden Abbildung dadurch, daß bereits berechnete Abbildungen verwendet werden können, beträchtlich. Wenige Worte sind ausreichend, um die Position des unbekannten Sprechers in bezug auf diejenigen Sprecher zu ermitteln, deren Abbildungen letztlich für die Normalisierung verwendet werden.

Das Verfahren wird durch eine geeignete Auswahl der Referenzsprecher verbessert werden können. Diese sollten so gewählt sein, daß sie den »Raum der Sprecher« möglichst vollständig und gleichmäßig abdecken. Zwischen diesen Sprechern wird eine Interpolation der Abbildungen um so leichter möglich sein, je dichter sie liegen. Es ist denkbar, eine Hierarchie von Steuernetzen aufzubauen, die die für die Mischung der Abbildungen in Frage kommenden Sprecher stufenweise eingrenzen *(Speaker-Clustering)*.

Stehen mehr Sprecher zur Verfügung als Referenzsprecher benötigt werden, können im Vorfeld Mischungsverhältnisse von Referenzsprecherabbildungen für diese Sprecher ermittelt werden. Es sollte dann eine weitere Verbesserung des Verfahrens zu erreichen sein, indem man diese Verhältnisse zum Training des bisherigen Sprechererkennungs-Netzwerks hinzunimmt. Damit würde dem offenbar nichtlinearen Zusammenhang zwischen den Abständen der Sprecher voneinander und den optimalen Mischungsverhältnissen Rechnung getragen.

Eine weitere Option ist die Steuerung der eigentlichen Abbildung durch geeignete – von der speziellen Normalisierungsmethode abhängige – Parameter anstelle der einfachen Überlagerung der Abbildungen. Desweiteren wird es sich vermutlich lohnen, die Eignung verschiedener Merkmale als Eingabemuster des Steuernetzes zu untersuchen. In dieser Arbeit wurden im wesentlichen Merkmale verwendet, die für die Spracherkennung optimiert waren. Dies ist nicht notwendigerweise auch die beste Wahl für die Steuerung einer Adaption oder normalisierenden Abbildung.

Aufgrund der schon in dieser Arbeit gezeigten guten Ergebnisse kann erwartet werden, daß die vorgestellte Verfahrensweise mit den beschriebenen Ausarbeitungen und Erweiterungen zu einer praktisch nutzbaren und leistungsfähigen Erweiterung existierender Spracherkennungssysteme führen wird.

A. Algorithmen

A.1. Dynamic Time Warping

Die Sprachsignale zweier gleichlautender Äußerungen unterscheiden sich normalerweise unter anderem dadurch, daß sie verschiedene Dauer besitzen. Die zeitliche Zuordnung von Signalabschnitte der beiden Äußerungen erfordert im allgemeinen nichtlineare Zeitachsenskalierungen, also monotone Verzerrungen der beiden Zeitachsen. *Dynamic Time Warping* (DTW) (Sakoe und Chiba, 1978) ist ein Algorithmus zur Bestimmung dieser Verzerrungen. Diese werden so gewählt, daß der kumulierte Abstand (»DTW-Abstand«) der einander durch die Verzerrung zugeordneten Signalelemente (*Alignment*) minimiert wird.

Der Algorithmus lautet folgendermaßen: Gegeben seien zwei Folgen von Merkmalsvektoren $(\mathbf{x}_i)$ mit $0 \leq i \leq I$ und $(\mathbf{y}_j)$ mit $0 \leq j \leq J$ sowie ein Abstandsmaß $d(\mathbf{x}_i, \mathbf{y}_j)$. Gesucht wird dann eine Folge von Indexpaaren $((i,j)_n)$ mit $0 \leq n \leq N$, so daß $S_{IJ} := \min \sum_{n=0}^{N} d(\mathbf{x}_{i_n}, \mathbf{y}_{j_n})$. Für die Spracherkennung ist dabei nur S_{IJ} als Resultat des Algorithmus von Bedeutung. Dabei gelten folgende Randbedingungen:

- $i_0 = j_0 = 0$, $i_N = I$, $j_N = J$, d. h. die Anfangs- und Endpunkte der Signale müssen einander zugeordnet werden.
- $i_n \leq i_{n+1}, j_n \leq j_{n+1}$, d. h., die zeitliche Reihenfolge der Signalwerte muß erhalten bleiben.

S_{IJ} wird wie folgt bestimmt: (Der Einfachheit halber werden keine Fallunterscheidungen für negative Indizes vorgenommen. S_{ij}-Terme, die negative Indizes enthalten, werden im Algorithmus nicht berücksichtigt.)

1. Setze $S_{0,0} = 0$. Beginne mit $i = 0$, $j = 1$.
2. Berechne $S_{ij} = \min\left(S_{i-1,j},\ S_{i,j-1},\ S_{i-1,j-1}\right) + d(\mathbf{x}_i, \mathbf{y}_j)$.
3. Inkrementiere j. Solange $j \leq J$, fahre fort mit 2.
4. Setze $j = 0$ und inkrementiere i. Solange $i \leq I$, fahre fort mit 2.

Im Laufe des Verfahrens werden zum aktuellen $\mathbf{x}$-Index k nur die Werte S_{ij} mit $i = \{k-1,\ k\}$ benötigt. »Frühere« S_{ij} $(i < k-1)$ können wieder freigegeben werden.

Der Algorithmus wird heute kaum noch für Spracherkennung eingesetzt. Häufig wird aber für andere Zwecke ein Alignment zwischen verschiedenen Exemplaren

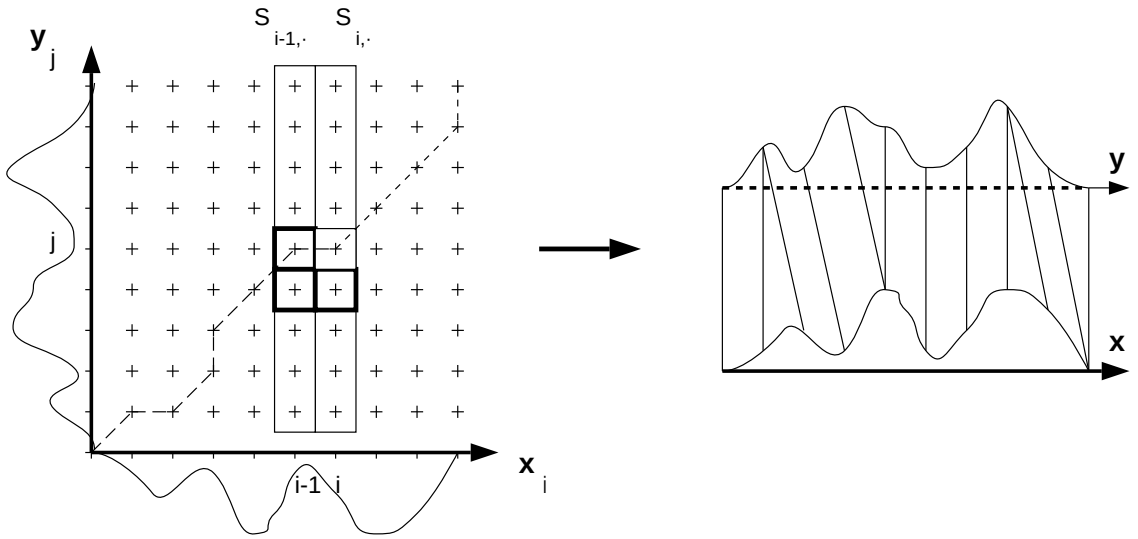

Abbildung A.1.: Links ist der DTW-Algorithmus skizziert, rechts die Zuordnung (das Alignment*), das durch das Verfahren hergestellt wird.*

gleichlautender Äußerungen gewünscht. Hierfür kann in Schritt 2 der Pfad, der zum aktuellen S_{ij} führte, gespeichert werden. Der Weg, der letztlich zu S_{IJ} geführt hat, kann dann nach dem Ablauf des Algorithmus zurückverfolgt werden.

A.2. Torgersons Algorithmus (Hauptkoordinatenmethode)

Die Hauptkoordinatenmethode ist ein Algorithmus zur metrischen multidimensionalen Skalierung und wurde von Torgerson (1952, 1958) angegeben. Im Prinzip wird eine Hauptkomponentenanalyse der Daten durchgeführt, die einer gegebenen Distanzmatrix zugrundeliegen. Dafür wird jedoch die Kovarianzmatrix $\mathbf{C}$ der Daten benötigt, die zunächst nicht zur Verfügung steht.

Bekannt seien also die euklidischen Abstände d_{ij} zwischen den Daten $\mathbf{x}_i$ und $\mathbf{x}_j$. Die Kovarianzen c_{ij} sind definitionsgemäß durch

$$c_{ij} := (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_j - \boldsymbol{\mu}) \tag{A.1}$$

gegeben (Papoulis, 1991), wobei $\boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$ der Mittelwert der Datenvektoren ist. Durch Einsetzen der Mittelwerte und Ausmultiplizieren erhält man

$$c_{ij} = \mathbf{x}_i^T \mathbf{x}_j - \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_i^T \mathbf{x}_n - \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_j^T \mathbf{x}_n + \frac{1}{N^2} \sum_{n=1}^{N} \sum_{k=1}^{N} \mathbf{x}_n^T \mathbf{x}_k. \tag{A.2}$$

Wegen

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^2 = \mathbf{x}_i^2 - 2\mathbf{x}_i^T\mathbf{x}_j + \mathbf{x}_j^2 \quad \text{bzw.} \tag{A.3}$$

$$\mathbf{x}_i^T\mathbf{x}_j = \frac{1}{2}\left(\mathbf{x}_i^2 + \mathbf{x}_j^2 - d_{ij}^2\right) \tag{A.4}$$

findet man nach Einsetzen und einigem Kürzen:

$$c_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{N}\sum_{n=1}^{N} d_{ik}^2 - \frac{1}{N}\sum_{n=1}^{N} d_{kj}^2 + \frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{N} d_{kl}^2\right). \tag{A.5}$$

Man kann also die gewünschte Kovarianzmatrix aus den Quadraten der Abstände bestimmen. Auf Grund dieser Kovarianzmatrix wird nun eine Hauptkomponentenanalyse durchgeführt. Durch eine Eigenwertzerlegung (Kiełbasiński und Schwetlick, 1988) mit $\mathbf{C} = \mathbf{UVU}^T$ erhält man die Eigenvektoren der Kovarianzmatrix in den Spalten von $\mathbf{U}$ und die zugehörigen Eigenwerte in der Diagonalen von $\mathbf{V}$. Aus der $r \leq n$ Eigenvektoren zu von Null verschiedenen Eigenwerten ergeben sich die rekonstruierten r-dimensionalen Koordinaten (*Hauptkoordinaten*) als Zeilenvektoren $\mathbf{Y} = \mathbf{UV}^{\frac{1}{2}}$.

A.3. Perzeptrone

Perzeptrone sind ein spezieller Typus künstlicher neuronaler Netze. Algorithmen aus der Klasse der neuronalen Netze simulieren Mechanismen der Informationsverarbeitung, welche bei natürlichen (biologischen) neuronalen Netzwerken gefunden wurden. Auf diese Weise werden Probleme angegangen, die sich bei der Behandlung durch »klassische« Algorithmen als hartnäckig erweisen, aber von natürlichen neuronalen Netzen sehr erfolgreich gelöst werden. Hierzu gehören insbesondere zahlreiche Fragestellungen der Mustererkennung.

A.3.1. Das Neuronenmodell

Natürliche Nervenzellen sind komplexe biologische Gebilde mit einer großen Zahl von Eigenschaften. Für die Simulation der Funktion eines natürlichen Neuronenverbandes ist aber eine reduzierte *funktionale* Beschreibung von Nervenzellen erwünscht – es geht also bei der Beschreibung einer Nervenzelle um das »Was« und nicht das »Wie«. In diesem Sinne ist eine Nervenzelle ein »Schaltelement«. Sie empfängt Reize aus dem Außenraum oder von anderen Neuronen, und wenn die Summe dieser Reize ein gewisses Maß übersteigt, entlädt sich das zwischen dem Außenraum und dem Zellinneren bestehende Potential. Ein Neuron kann in diesem Sinne als ein Integrator mit einer gewissen Zeitkonstante beschrieben werden. Die Entladung des Membranpotentials setzt sich über das Axon der Nervenzelle fort und führt an dessen Ende zu einer Ausschüttung von Neurotransmittern, welche den Reiz zu einem anderen Neuron oder einer Muskelzelle weitertragen. Nachdem die Nervenzelle »gefeuert« hat, muß das Potential wieder aufgebaut werden, was zu einer Refraktärzeit

von einigen Millisekunden führt, innerhalb derer ein Neuron nicht wieder feuern kann. Das resultierende Signal wird also durch einen *Spike Train* beschrieben. Darin ist die »Stärke« der Eingangsreize durch die Feuerrate kodiert. Außerdem enthält das Signal durch die jeweiligen Feuerzeitpunkte Information über die zeitliche Feinstruktur Reize, welche das Feuern ausgelöst haben. Obwohl diese Feinstruktur augenscheinlich wichtig ist, wird sie in der Simulation häufig vernachlässigt, genauso wie sich auch die Existenz verschiedener Klassen von Nervenzellen gewöhnlich in der Simulation nicht niederschlägt. Insgesamt wird eine Nervenzelle daher gewöhnlich als ein Summierer beschrieben, dessen Ausgang eine monotone Funktion der Summe der Eingangssignale ist und als eine mittlere Feuerrate betrachtet werden kann. Bezeichnet man die an den »Dendriten« des simulierten Neurons anliegenden Reizstärken mit n_i, so postuliert man die Stärken der »synaptischen Kopplungen« als plastische Gewichte w_i. Überschreitet die Summe der entsprechend gewichteten Eingangsreize einen gewissen Schwellwert ϑ, so führt das zu einer Aktivierung $f(\sum_i w_i - \vartheta)$ am Ausgang des Neurons. f ist dabei die beschriebene monotone Transferfunktion des Neurons und wird oft als Fermifunktion

$$f_F(x) = \frac{1}{1 + e^{-2\beta x}} \tag{A.6}$$

angesetzt. In vielen Fällen ist es einfacher, die Transferfunktion nicht auf Werte zwischen 0 und 1 zu beschränken, sondern statt dessen symmetrische Funktionen mit Werten zwischen -1 und 1 zu verwenden. Dann entspricht der Fermifunktion der Tangens hyperbolicus. Verschiedentlich wird auch ganz auf eine nichtlineare Transferfunktion verzichtet, so daß dann $f_L(x) = ax$ mit $a \in \mathcal{R}^+$ gilt.

A.3.2. Neuronale Netzwerke

Künstliche neuronale Netzwerke simulieren einen ganzen Neuronenverband. Meistens werden die Transferfunktionen der einzelnen simulierten Nervenzellen einheitlich festgelegt. Die Aktivitäten (s_k) der Neuronen k werden initialisiert und anschließend gegebenfalls in einem iterativen Prozeß neu festgelegt, bis der Verband in einem stabilen Zustand angelangt ist. Dies geschieht beispielsweise durch folgende Iterationsvorschrift:

$$\mathbf{s}_{n+1} = f(\mathbf{W}\mathbf{s}_n - \boldsymbol{\vartheta}). \tag{A.7}$$

(Hier und im folgenden wird f und nicht die Vektorschreibweise $\mathbf{f}$ verwendet, um zu illustrieren, daß die Transferfunktion für alle Zellen die gleiche sein soll – Nichtsdestoweniger *ist* f natürlich vektorwertig.) Die Endkonfiguration des Netzes, also die gewünschte Ausgabe, wird im wesentlichen durch die Gewichtsmatrix $\mathbf{W}$ festgelegt, deren Elemente w_{ij} das synaptische Gewicht von Neuron i zu Neuron j bezeichnet. Diese wird durch eine *Lernregel* bestimmt. Zusammengefaßt definiert sich also ein künstliches neuronales Netz gewöhnlich durch ein verwendetes Neuronenmodell, seine Architektur (Anzahl der Zellen und Definition der möglichen Verbindungen) und durch seine Lernregel. Auf die verschiedenen Typen künstlicher neuronaler Netzwerke soll hier nicht weiter eingegangen werden. Allgemeine Darstellungen bieten beispielsweise Lippmann (1987), Müller und Reinhardt (1990) oder Rumelhart et al. (1986b).

A.3.3. Perzeptrone

Perzeptrone sind gerichtete (unidirektionale oder *feed-forward*), mehrschichtige Netzwerke. Das heißt, wenn es einem Neuron i erlaubt ist, ein Neuron j zu beinflussen, dann ist dies Neuron j umgekehrt nicht erlaubt. Im einfachsten Falle bestehen Perzeptrone aus einer Eingabeschicht, welche ein Eingangssignal aufnimmt, und einer Ausgabeschicht, welche das gewünschte, von der Eingabe abhängige Ergebnis präsentieren soll. Der Name *Perzeptron* wurde solchen Systemen von Rosenblatt (1959) gegeben, der diese Systeme (für binäre Neuronen) zuerst detailliert untersuchte. Weil diese Struktur gerichtet ist, ist es nicht nötig, die Ausgabezellen zu initialisieren oder die Abbildung zu iterieren (Gleichung A.7). Das Ergebnis liegt statt dessen nach einem Schritt vor.

Perzeptrone werden zur Approximation einer gesuchten Abbildung oder zur Klassifikation von Daten eingesetzt. Beide Anwendungen werden in dieser Arbeit genutzt.

Einfache Perzeptrone

Die durch ein einfaches[1] Perzeptron gegebene Abbildung lautet

$$\mathbf{s} = f(\mathbf{W}\boldsymbol{\sigma} - \boldsymbol{\vartheta}). \tag{A.8}$$

Hier bezeichnet $\boldsymbol{\sigma}$ den Vektor der Aktivitäten in der Eingabeschicht, $\mathbf{s}$ den der Ausgabeschicht. Man sieht leicht, daß man bei der Berechnung der Abbildung durch die formale Einführung einer zusätzlichen Eingabezelle mit der statischen Aktivität -1 den Schwellwertvektor $\boldsymbol{\vartheta}$ in die Gewichtsmatrix verlagern kann. Dies soll im folgenden so gehandhabt werden, um die Rechnungen übersichtlicher zu gestalten:

$$\begin{aligned} \boldsymbol{\sigma} &= (\sigma_i) \rightarrow (\sigma_i | -1)^T \\ \text{und} \quad \mathbf{W} &= (\mathbf{w}_i) \rightarrow (\mathbf{w}_i | \boldsymbol{\vartheta}). \end{aligned} \tag{A.9}$$

Die Aufgabe des Perzeptrons ist die Approximation einer gewünschten Abbildung zwischen einer Anzahl von M Eingabemustern $\mathbf{u}^m$ auf gegebene zugehörige Ausgabemuster $\mathbf{v}^m$.[2] Hierzu benötigt man eine Lernregel, welche die Gewichtsmatrix $\mathbf{W}$ in geeigneter Weise festlegt. Erreicht werden kann dies durch einen Gradientenabstieg im Raum der Gewichte. Um diese Regel zu finden, definiert man zunächst ein geeignetes Fehlermaß

$$E = \frac{1}{2} \sum_{m=1}^{M} (\mathbf{s}^m - \mathbf{v}^m)^2. \tag{A.10}$$

[1] Die hier als »einfache« Perzeptrone bezeichneten Netze werden auch häufig als *einschichtig* bezeichnet, da die Eingabeschicht keine Operation durchführt. Da andererseits bei mehrschichtigen Perzeptronen die Eingabeschicht gewöhnlich mitgezählt wird, entsteht eine Inkonsistenz, die durch die hier gewählte Nomenklatur vermieden werden soll.

[2] Für den Fall einer Klassifikation wählt man normalerweise eine 1-aus-N-Kodierung für die Ausgabeschicht, so daß jede Ausgabezelle mit einer Klasse korrespondiert. Die Zielmuster $\mathbf{v}$ wären dann euklidische Einheitsvektoren. Da die Werte 0 und 1 von der Fermifunktion nie erreicht werden, wählt man für die Zielmuster allerdings besser Werte wie 0,1 und 0,9.

Dieses Maß ist eine monotone Funktion der Abweichung der Antwort des Netzes auf ein Eingabemuster vom gewünschten Zielmuster. Der Faktor $\frac{1}{2}$ dient hier nur der Bequemlichkeit – er kürzt sich gleich wieder heraus. Das Fehlermaß ist nach Gleichung A.8 eine Funktion der Gewichte. Wie der Fehler minimiert werden kann, erkennt man durch Substitution von Gleichung A.8 und Ableitung nach den Gewichten mit Hilfe der Kettenregel:

$$\begin{aligned}\frac{\partial E}{\partial w_{ij}} &= \frac{1}{2}\frac{\partial}{\partial w_{ij}}\sum_{m=1}^{M}\left(f(\mathbf{W}\boldsymbol{\sigma}^m)-\mathbf{v}^m\right)^2 \\ &= \sum_{m=1}^{M}(s_i^m - v_i^m)f'\left(\sum_k w_{ik}\sigma_k^m\right)\sigma_j^m .\end{aligned} \tag{A.11}$$

Da der Gradient in Richtung des steilsten Anstieges der Funktion zeigt, kann der Abbildungsfehler iterativ minimiert werden, indem man die Gewichte in genügend kleinen Schritten in die entgegengesetzte Richtung adaptiert:

$$\Delta w_{ij} = -\epsilon\frac{\partial E}{\partial w_{ij}} = -\epsilon\sum_{m=1}^{M}(s_i^m - v_i^m)f'\left(\sum_k w_{ik}\sigma_k^m\right)\sigma_j^m \tag{A.12}$$

Die Ableitung der Fermifunktion A.6 ist übrigens

$$f_F'(x) = -2\beta\left(1 - f_F^2(x)\right). \tag{A.13}$$

Mehrschichtige Perzeptrone

Leider sind einfache Perzeptrone beschränkt, was ihre Fähigkeit zur Approximation einer Abbildung angeht (Minsky und Papert, 1969). So können prinzipiell nur solche Klassifikationsaufgaben gelöst werden, deren Elemente im Raum der Muster linear separabel sind, d. h. durch Hyperebenen voneinander getrennt werden können. Fügt man hingegen noch eine (oder mehrere) weitere gleichartige Schichten zwischen die Eingabe- und die Ausgabeschicht eines einfachen Perzeptrons ein, so kann gezeigt werden, daß bereits eine einzelne »versteckte« Schicht es (prinzipiell) ermöglicht, jede stetige Funktion zu repräsentieren (de Figueiredo, 1980; Hecht-Nielsen, 1987). Die klassische Lernregel für mehrschichtige Perzeptrone ist der *Backpropagation*-Algorithmus (Rumelhart et al., 1986a), welcher eine Verallgemeinerung der eben beschriebenen sogenannten *Deltaregel* ist. Das Verfahren soll hier an einem Netz mit einer versteckten Schicht entwickelt werden. Die Erweiterung auf mehrere versteckte Schichten ist dann leicht zu sehen. Wie oben (Gleichung A.8 mit A.9) sei für die versteckte Schicht (deren Größen hier mit einer Tilde bezeichnet werden):

$$\tilde{\mathbf{s}} = f(\tilde{\mathbf{W}}\boldsymbol{\sigma}) \tag{A.14}$$

und für die Ausgabeschicht

$$\mathbf{s} = f(\mathbf{W}\tilde{\mathbf{s}}) \tag{A.15}$$

Wiederum sei die Forderung

$$E = \frac{1}{2}\sum_{m=1}^{M}(\mathbf{s}^m - \mathbf{v}^m)^2. \tag{A.16}$$

Betrachtet man zunächst nur den Gradientenabstieg im Raum der Gewichte der Ausgabeschicht, ergibt sich ähnlich wie Gleichung A.12

$$\Delta w_{ij} = -\epsilon\frac{\partial E}{\partial w_{ij}} = -\epsilon\sum_{m=1}^{M}(s_i^m - v_i^m)f'\left(\sum_k w_{ik}\tilde{s}_k^m\right)\tilde{s}_j^m. \tag{A.17}$$

Die Berechnung der Änderung der Gewichte von der Eingabe- zur Zwischenschicht ist vollkommen analog, allerdings muß die Kettenregel noch einmal mehr angewandt werden:

$$\begin{aligned}
\Delta\tilde{w}_{ij} &= -\epsilon\frac{\partial E}{\partial\tilde{w}_{ij}} = -\frac{\epsilon}{2}\frac{\partial}{\partial\tilde{w}_{ij}}\sum_{m=1}^{M}\left(f(\mathbf{W}\tilde{\mathbf{s}}^m) - \mathbf{v}^m\right)^2 \\
&= -\epsilon\sum_{m=1}^{M}\sum_l(s_l^m - v_l^m)f'\left(\sum_k w_{lk}\tilde{s}_k^m\right)\frac{\partial\sum_k w_{lk}\tilde{s}_k^m}{\partial\tilde{s}_i^m}\frac{\partial\tilde{s}_i^m}{\partial\tilde{w}_{ij}} \\
&= -\epsilon\sum_{m=1}^{M}\sum_l(s_l^m - v_l^m)f'\left(\sum_k w_{lk}\tilde{s}_k^m\right)w_{li}\ f'\left(\sum_k\tilde{w}_{ik}\sigma_k^m\right)\frac{\partial}{\partial\tilde{w}_{ij}}\sum_k\tilde{w}_{ik}\sigma_k^m \\
&= -\epsilon\sum_{m=1}^{M}\sum_l(s_l^m - v_l^m)f'\left(\sum_k w_{lk}\tilde{s}_k^m\right)w_{li}\ f'\left(\sum_k\tilde{w}_{ik}\sigma_k^m\right)\sigma_j^m.
\end{aligned} \tag{A.18}$$

Durch Vergleich mit Gleichung A.17 und mit der Abkürzung

$$D_i^m := (s_i^m - v_i^m)f'\left(\sum_k w_{ik}\tilde{s}_k^m\right) \tag{A.19}$$

wird aus Gleichung A.17

$$\Delta w_{ij} = -\epsilon\sum_{m=1}^{M}D_m^i\tilde{s}_j^m \tag{A.20}$$

und aus Gleichung A.18

$$\Delta\tilde{w}_{ij} = -\epsilon\sum_{m=1}^{M}\left(\sum_l D_l^m w_{li}\right)f'\left(\sum_k\tilde{w}_{ik}\sigma_k^m\right)\sigma_j^m. \tag{A.21}$$

Im Gegensatz zur Fehlerlandschaft des einfachen Perzeptrons ist die eines mehrschichtigen Netzes multimodal, so daß man hier damit rechnen muß, daß der Gradientenabstieg in einem lokalen Minimum endet – es empfiehlt sich daher, die Proze-

dur gegebenfalls mit verschiedenen Startkonfigurationen zu wiederholen.[3] Es ist keine Methode bekannt, mit der sich generell bestimmen ließe, wieviele Zwischenschichtneuronen zur Bewältigung einer gegebenen Aufgabe nötig sind. Die Interpretation der Abbildung in den versteckten Schichten häufig schwierig.

Ein anderes Problem bei mehrschichtigen Netzen ist die Anzahl der freien Parameter des Netzes, welche durch die Zwischenschichtneuronen deutlich wächst. Zum einen ist dadurch der Rechenaufwand für das Training mehrschichtiger Netze bedeutend höher als bei einfachen Perzeptronen. Zum anderen benötigt man eine entsprechend große Menge von Trainingsmusterpaaren, um diese Parameter zu bestimmen. Stehen diese nicht zur Verfügung (oder sind sie zu stark korreliert), läuft man Gefahr, daß das Netz die Trainingsmenge »auswendig« lernt: Wenn die Zahl der freien Parameter des Netzes gegenüber der unabhängigen Variablen der Trainingsmenge groß genug ist, neigen Perzeptrone dazu, die Trainingsmenge zu »speichern«. Für neue Muster werden dann »undefinierte« Ergebnisse produziert. Erst wenn sich dieses Verhältnis umkehrt, werden Netzwerkonfigurationen gefunden, die den dann unvermeitlichen Abbildungsfehler minimieren und dann auch »gutartige« Ergebnisse für unbekannte Daten ausgeben. Dieses verhalten nennt man die *Generalisierungsfähigkeit* des Netzes. Überlegungen und Experimente für den Fall binärer Neuronen findet man bei Anshelevich et al. (1989).

A.4. Diskriminanzanalyse

Besitzt man einen Datensatz, dessen Elemente verschiedenen Klassen angehören, so ist häufig so, daß die Elemente einer Klasse gewisse zusammenhängende Gebiete innerhalb des Parameterraumes definieren. Durch eine Diskriminanzanalyse möchte man nun Funktionen finden, die diese Gebiete voneinander trennen. Anhand dieser Diskriminanzfunktionen kann man beispielsweise entscheiden, welcher Klasse ein bisher unbekanntes Element zugeordnet werden muß. Wenn es einen gewissen Überlapp in den Klassen gibt, kann man von einer Diskriminanzfunktion nicht in allen Fällen eine richtige Entscheidung erwarten, sondern allenfalls eine, welche in gewisser Weise statistisch optimal ist.

Prinzipiell kann man die Methoden zur Diskriminanzanalyse in parametrische und in nichtparametrische Methoden aufteilen. Nichtparametrische Methoden (Sortieren nach Koordinaten, Methode der konvexen Hülle (Wilf, 1977)) arbeiten direkt auf den Daten und versuchen, im Parameterraum Oberflächen zu finden, welche die Klassen so gut wie möglich voneinander trennen. Parametrische Methoden versuchen dagegen, die Parameter der den Daten zugrundeliegenden Verteilungen zu schätzen, so daß zukünftige Werte anhand dieser Verteilungen zugeordnet werden können.

Für parametrische Diskriminanzanalysemethoden verwendet man sinnvollerweise die Verteilungsfunktionen der zu analysierenden Daten, sofern diese bekannt sind.

[3] Bei hochdimensionalen Problemen tritt dieses Problem allerdings eher selten auf, da dann – bildlich gesprochen – die Chance, daß in irgendeiner Dimension noch ein Türchen offen steht, relativ groß ist.

Anderenfalls ist es wegen des *Zentralen Grenzwertsatzes* (Papoulis, 1991) ein vernünftiger Ansatz, eine Normalverteilung anzusetzen, solange man annehmen kann, daß die Streuung innerhalb der Klassen aus einer großen Zahl additiver Zufallseffekte resultiert. Die n-dimensionale Normalverteilung einer Variablen $\mathbf{x}$ ist gegeben durch

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{A.22}$$

Dabei ist $\mathbf{C}$ die Kovarianzmatrix der Zufallsvariablen $\mathbf{x}$ und $\boldsymbol{\mu}$ deren Mittelwert (Papoulis, 1991).

A.4.1. Lineare Diskriminanzanalyse

Die lineare Diskriminanzanalyse gehört zu den parametrischen Verfahren (Wilf, 1977). Sie macht die Annahme einer Normalverteilung der zu analysierenden Daten. Desweiteren wird davon ausgegangen, daß die Kovarianzen der verschiedenen Verteilungen identisch sind. Es ist dann leicht einzusehen, daß ein unbekanntes Element $\mathbf{x}$ nun der Klasse i zugeordnet werden kann, dessen Verteilungsfunktion f_i für $\mathbf{x}$ maximal ist, denn dadurch werden offenbar die zu erwartenden Kosten einer Fehlklassifikation minimiert, wenn die Häufigkeiten des Auftretens der Elemente der verschiedenen Klassen sowie die Kosten für eine Fehlklassifikation für alle Klassen gleich sind.

Unter den getroffenen Annahmen wird aus der Frage

$$f_i(\mathbf{x}) > f_j(\mathbf{x}) \tag{A.23}$$

zur Klassifizierung eines unbekannten Elementes $\mathbf{x}$ die Frage

$$-(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) > -(\mathbf{x}-\boldsymbol{\mu}_j)^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j), \tag{A.24}$$

wobei die $\boldsymbol{\mu}_k$ die Mittelwerte der jeweiligen Verteilungen sind. Durch Ausmultiplizieren und Kürzen ergibt sich daraus:

$$2\mathbf{x}_i^T \mathbf{C}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^T \mathbf{C}^{-1}\boldsymbol{\mu}_i > 2\mathbf{x}_j^T \mathbf{C}^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T \mathbf{C}^{-1}\boldsymbol{\mu}_j. \tag{A.25}$$

Mit den Abkürzungen

$$\mathbf{a}_k := 2\mathbf{C}^{-1}\boldsymbol{\mu}_k \tag{A.26}$$

und

$$b_k := \boldsymbol{\mu}_k^T \mathbf{C}^{-1}\boldsymbol{\mu}_k \tag{A.27}$$

wird daraus

$$\mathbf{a}_i^T \mathbf{x}_i - b_i > \mathbf{a}_j^T \mathbf{x}_j - b_j. \tag{A.28}$$

So kann also die Entscheidung, welcher Klasse ein Datenpunkt $\mathbf{x}$ angehören soll, anhand linearer Diskriminanzfunktionen

$$h_k(\mathbf{x}) := \mathbf{a}_k^T \mathbf{x} - b_k \tag{A.29}$$

getroffen werden, im Unterschied zu Gleichung A.24, in der $\mathbf{x}$ in quadratischen Termen auftritt.

Nun kann man sich die Frage stellen, ob der Gewinn, den man durch eine lineare Diskriminanzfunktion erhält, nicht durch die getroffenen Annahmen zunichte gemacht wird. Man gewinnt aber noch einen weiteren Vorteil: Da die Parameter der Verteilung nur geschätzt werden können, läßt sich die Frage, ob die Kovarianzen zweier Verteilungen *tatsächlich* verschieden sind, gerade bei kleinen Datenmengen oft nicht mit statistischer Signifikanz beantworten. Wenn aber die Kovarianzen statistisch nicht unterscheidbar sind, erhält man durch die Vereinigung der verfügbaren Daten offenkundig eine bessere Schätzung der gemeinsamen Kovarianzen.

B. Lautschrift und System der Laute

Die in dieser Arbeit verwendete Lautschrift ist das *Internationale Phonetische Alphabet* (IPA), welches im folgenden dargestellt wird, soweit es für diese Arbeit von Belang ist. Dort werden auch einige weitere verwendete Konventionen erklärt. Der anschließende Abschnitt B.2 erläutert die verschiedenen Klassifikationen von Vokalen und Konsonanten. Die Darstellung folgt im wesentlichen (Drosdowski, 1995).

B.1. Verwendete Lautschrift

Tabelle B.1.: Konsonanten und Vokale. (Wird auf der nächsten Seite fortgesetzt.)

IPA-Symbol	Beispiel	Anmerkung	IPA-Symbol	Beispiel
Konsonanten			*Vokale*	
[b]	**B**all		[a]	k**a**lt
[ç]	Be**ch**er		[ɑ]	K**ah**n
[d]	**D**ampf		[ɐ]	Schieb**er**
[f]	**F**rosch		[ã]	Gourm**and**
[g]	**G**ans		[æ]	n**ä**hme
[h]	**H**aus		[e]	R**e**h
[j]	**J**acke		[ɛ]	B**e**tt
[k]	**K**amm		[ɛ̃]	T**eint**
[l]	**L**ist		[ə]	Rab**e**
[m]	**M**ilch		[i]	Br**ie**f
[n]	**N**apf		[ɪ]	S**i**nn
[ŋ]	Ri**ng**		[o]	H**o**f
[p]	**P**ult		[ɔ]	T**o**pf
[r]	**R**and	alveolar	[õ]	Balk**on**
[ʀ]	**R**and	uvular	[ø]	F**ö**hn
[s]	Mu**s**e		[œ]	K**ö**rner
[ʃ]	**Sch**al		[œ̃]	Parf**um**
[t]	**T**eer		[u]	M**u**t
[v]	**W**ald		[ʊ]	H**u**nd

IPA-Symbol	Beispiel	Anmerkung	IPA-Symbol	Beispiel
Konsonanten			*Vokale*	
[x]	Ka**ch**el		[y]	s**ü**ß
[z]	**S**inn		[ʏ]	S**ü**nde
[ʒ]	**G**enie			
[ʔ]	Uhr	Glottaler Verschluß vor [u]		

Tabelle B.3.: Andere Konventionen

Symbol	Bedeutung
[·]	phonetische Schreibweise
/·/	phonologische/phonemische Schreibweise

B.2. Lautsystem

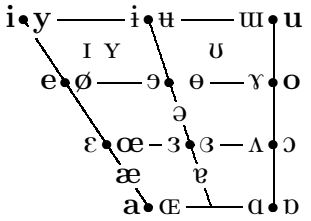

Abbildung B.1: IPA-Vokalviereck. Die im Deutschen vorkommenden Vokale sind fett gesetzt. Die Zungenstellung ist von links vorn, zentral, hinten*, die Lippen sind von oben* geschlossen, halbgeschlossen, halboffen, offen*. Bei Vokalpaaren ist der rechte Vokal gerundet.*

Tabelle B.4.: Konsonanten

Artikulationsart		Artikulierendes Organ						
		labial		koronal		dorsal		glottal
		stl	sth	stl	sth	stl	sth	
obstruent	plosiv	p	b	t	d	k	g	ʔ
	frikativ	f	v	s ʃ	z ʒ	ç	j	h
sonorant	nasal		m		n		ŋ	
	oral				l		ʀ	

In dieser Tabelle bedeutet »sth« stimmhaft, »stl« stimmlos. Orale Sonoranten heißen zumeist *Liquide*.

Tabelle B.5.: Vokale

Mundöffnung	Zungenposition				
	vorn				hinten
	Lippenrundung				
	ungerundet		gerundet		
geschlossen	i	ɪ	y	ʏ	u ʊ
halbgeschlossen	e	ɛ	ø	œ	o ɔ
offen		æ a			ɑ

Die innerhalb der Spalten jeweils links stehenden Vokale sind *gespannte* Vokale, rechts stehen die *ungespannten* Vokale. Die gespannten Vokale liegen im Vokalviereck weiter außen, d. h. je gespannter ein Vokal ist, desto weiter ist er von dem neutralen (entspannten) Schwa-Laut /ə/ entfernt.

Literaturverzeichnis

V. V. Anshelevich, B. R. Amirikian, A. V. Lukashin und M. D. Frank-Kamenetskii. On the ability of neural networks to perform generalization by induction. *Biol. Cybern.*, 61:125–128, 1989.

Holger Behme und Wolf-Dieter Brandt. Speech recognition by hierarchical segment classification. In *Proc. Internat. Conf. Artificial Neural Networks*, 1993.

M. Blomberg und K. Elenius. Nonlinear frequency warp for speech recognition. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.*, Band 4, Seiten 2631–2634. IEEE, 1986.

Karlheinz Brandenburg und Marina Bosi. Overview of MPEG audio: Current and future standards for low-bit-rate audio coding. *J. Audio Eng. Soc.*, 45(1/2):4–21, 1997.

Wolf-Dieter Brandt. Bildung von Merkmalen zur Spracherkennung mittels phonotopischer Karten. Diplomarbeit, Drittes Physikal. Inst., Universität Göttingen, 1991.

I. N. Bronstein und K. A. Semendjajew. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun, 23 Auflage, 1982.

Andrés Buzo, Augustine H. Gray, jr., Robert M. Gray und John D. Markel. Speech coding based upon vector quantization. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(4):562–574, 1980.

Christina Chesta, Olivier Siohan und Chin-Hui Lee. Maximum a posteriori linear regression for hidden markov model adaptation. In *Proceedings of Eurospeech '99*, Band 1, Seiten 211–214, 1999.

Wu Chou. Maximum a posterior linear regression with elliptically symmetric matrix variate priors. In *Proceedings of Eurospeech '99*, Band 1, Seiten 1–4, 1999.

F. Class, A. Kaltenmeier, P. Regel und K. Trottler. Fast speaker adaptation for speech recognition systems. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.*, Band 1, Seiten 133–136. IEEE, 1990.

Ronald A. Cole, Yonghong Yan, Brian Mak, Mark Fanty und Troy Bailey. The contribution of consonants versus vowels to word recognition in fluent speech. In

Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process., Band 2, Seiten 853–856. IEEE, 1996.

R. J. P. de Figueiredo. Implications and applications of Kolmogorov's superposition theorem. *J. Math. Anal. Appl.*, 38:1227, 1980.

Werner Digel und Gerhard Kwiatkowski, Herausgeber. *Mayers Großes Taschenlexikon in 24 Bänden*. Bibliographisches Institut & F. A. Brockhaus AG, Mannheim, zweite Auflage, 1987.

S. Disner. Evaluation of vowel normalization procedures. *J. Acoust. Soc. Am.*, 67(1): 253–261, Januar 1980.

Günther Drosdowski, Herausgeber. *Aussprachewörterbuch*, Band 6 von *Der Große Duden in 10 Bänden*. Bibliographisches Institut AG, Mannheim, zweite Auflage, 1974.

Günther Drosdowski, Herausgeber. *Grammatik der deutschen Gegenwartssprache*, Band 4 von *Der Duden in 12 Bänden*. Dudenverlag, Mannheim, fünfte Auflage, 1995.

James L. Flanagan. *Speech Analysis Synthesis and Perception*. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag, Berlin, erste Auflage, 1965.

Heiko Freienstein. *Vokaltraktmodellbasierte Schätzung von Steuerparametern eines Moduls zur Sprechernormalisierung*. Dissertation, Universität Göttingen, Drittes Physikalisches Institut, 2000.

K. Fukuzawa, H. Sawai und M. Sugiyama. Segment-based speaker adaptation by neural network. In *Proc. IEEE Workshop Neural Networks for Signal Process.*, Seiten 440–451. IEEE, 1991.

J. L. Gauvain und C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian observations of Markov chains. *IEEE Trans. Speech Audio Proc.*, 2(2): 291–298, April 1994.

S. Goronzy und R. Kompe. A MAP-like weighting scheme for MLLR speaker adaptation. In *Proceedings of Eurospeech '99*, Band 1, Seiten 5–8, 1999.

Timothy J. Hazen. *The use of speaker correlation information for automatic speech recognition*. Dissertation, Massachusetts Institute of Technology, 1998. Zitiert nach Hazen (2000).

Timothy J. Hazen. A comparison of novel techniques for rapid speaker adaptation. *Speech Communication*, 31(1):15–33, Mai 2000.

R. Hecht-Nielsen. Kolmogorov's mapping neural network existence theorem. In *IEEE First Annual Internat. Conf. Neural Networks*. IEEE, 1987. Paper III-11.

Hynek Hermansky. Should recognizers have ears? *Speech Communication*, 25(1–3): 3–27, August 1998.

X. D. Huang, K. F. Lee und A. Waibel. Connectionist speaker normalization and its application to speech recognition. In *Neural Networks for Signal Process., Proc. of the 1991 IEEE Workshop*, Seiten 357–366. IEEE, 1991.

IPA. The international phonetic alphabet. Leeds, 1989.

F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64: 532–556, 1976.

A. Kiełbasiński und H. Schwetlick. *Numerische lineare Algebra*. Verlag Harri Deutsch, Thun, 1988.

W. Klein, R. Plomp und L. Pols. Vowel spectra, vowel spaces and vowel identification. *J. Acoust. Soc. Am.*, 48(4):999–1009, 1970.

L. Knohl und A. Rinscheid. Verfahren zur kontinuierlichen Merkmalsadaption mittels selbstorganisierender, topologie-erhaltender Merkmalskarten. In *Fortschritte der Akustik – DAGA'93*, Seiten 1004–1007, Bad Honnef, 1993. DAGA, DPG-GmbH.

Teuvo Kohonen. *Self-Organization and Associative Memory*, Band 8 von *Springer Series in Information Sciences*. Springer-Verlag, Berlin, zweite Auflage, 1988.

Roland Kuhn, Patrick Nguyen, Jean-Claude Junqua, Robert C. Boman, Nancy A. Niedzielski, Steven Fincke, Kenneth L. Field und Matteo Contolini. Fast speaker adaptation using *a priori* knowledge. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.* IEEE, 1999. Beitrag Nummer 1587.

K. J. Lang, A. H. Waibel und G. E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23–43, 1990.

C.-H. Lee und J.-L. Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.*, Band 2, Seiten 558–561. IEEE, 1993.

C. Legetter und P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Comp. Speech Lang*, 9(2):171–185, 1995.

Yoseph Linde, Andrés Buzo und Robert M. Gray. An algorithm for vector quantizer design. *IEEE Trans. Communications*, 28:84–95, 1980.

R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, Seiten 4–22, April 1987.

J. Makhoul, S. Roucos und H. Gish. Vector quantization in speech coding. *Proc. IEEE*, 73(11):1551–1588, 1985.

J. D. Markel und A. H. Grey, Jr. *Linear Prediction of Speech*. Springer-Verlag, Berlin, 1976.

John W. McDonough und William J. Byrne. Speaker adaptation with all-pass transforms. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.* IEEE, 1999. Beitrag Nummer 2093.

George A. Miller. *Wörter – Streifzüge durch die Psycholinguistik*. Zweitausendeins, Frankfurt am Main, zweite Auflage, 1996.

Marvin Minsky und Seymour Papert. *Perceptrons – An introduction to computational geometry*. The MIT Press, Cambridge, Mass., 1969.

B. Müller und J. Reinhardt. *Neural Networks – An Introduction*. Springer-Verlag, Berlin, 1990.

Knut Müller. Sprechernormalisierung auf Barkspektrogrammen durch lineare Abbildung im Sinne des kleinsten quadratischen Fehlers. Diplomarbeit, Drittes Physikal. Inst., Universität Göttingen, 1993.

Knut Müller und Hans Werner Strube. Sprechernormalisierende Abbildung von Barkspektrogrammen. In *Fortschritte der Akustik – DAGA'93*, Seiten 986–989, Bad Honnef, 1993. DAGA, DPG-GmbH.

Masati Naito, Li Deng und Yoshinori Sagisaka. Model-based speaker normalization methods for speech recognition. In *Proceedings of Eurospeech '99*, Band 6, Seiten 2515–2518, 1999.

S. Nakamura und K. Shikano. Speaker adaptation applied to HMM and neural networks. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.*, Band 1, Seite 612. IEEE, 1990.

Patrick Nguyen, Christian Wellekens und Jean-Claude Junqua. Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. In *Proceedings of Eurospeech '99*, Band 6, Seiten 2519–2522, 1999.

Oluṣeyi Olurotimi. Recurrent neural network training with feedforward complexity. *IEEE Trans. Neural Networks*, 5(2):185–197, 1994.

Douglas O'Shaughnessy. *Speech Communication – Human and Machine*. Addison-Wesley Series in Electrical Engineering: Digital Signal Processing. Addison-Wesley Publishing Company, Reading, 1987.

Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Series in Electrical Engineering. McGraw-Hill, Inc., New York, dritte Auflage, 1991.

Chaslav V. Pavlovic und Gerald A. Studebaker. An evaluation of some assumptions underlying the articulation index. *J. Acoust. Soc. Am.*, 75:1606–1612, 1984.

V. V. Phansalkar und P. S. Sastry. Analysis of the back-propagation algorithm with momentum. *IEEE Trans. Neural Networks*, 5(3):505–506, Mai 1994.

William H. Press, Brian P. Flannery, Saul A. Teukolsky und William T. Vetterling. *Numerical Recipies in C*. Cambridge University Press, Cambridge, Mass., erste Auflage, 1988.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989.

Lawrence R. Rabiner und Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series. PTR Prentice Hall, Inc., Englewood Cliffs, New Jersey, erste Auflage, 1993.

F. Rosenblatt. *Principles of Neurodynamics – Perceptrons and the Theorie of Brain Mechanisms*. Spartan Books, New York, 1959.

David E. Rumelhart, G. E. Hilton und R. J. Williams. *Learning Internal Representations by Error Propagation*, Kapitel 8. Band 1 von Rumelhart, Rumelhart et al. (1986b), 1986a.

David E. Rumelhart, James L. McClelland und the PDP Research Group. *Parallel Distributed Processing – Explorations in the Microstructure of Cognition*. The MIT Press, Cambridge, Mass., 1986b.

H. Sakoe und S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process.*, 26:43–49, 1978.

Manfred R. Schroeder. Models of hearing. *Proc. IEEE*, 63:1332–1350, 1975.

Manfred R. Schroeder. Recognition of complex acoustic signals. In T. H. Bullock, Herausgeber, *Life Sciences Research Report*, Band 5, Seiten 323–328. Abakon Verlag, Berlin, 1977. Zitiert nach Schroeder et al. (1979).

Manfred R. Schroeder. *Computer Speech – Recognition, Compression, Synthesis*, Band 35 von *Springer Series in Information Sciences*. Springer-Verlag, Berlin, 1990.

Manfred R. Schroeder, B. S. Atal und J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am.*, 66(6): 1647–1652, 1979.

Manfred R. Schroeder und Hans Werner Strube. Flat-spectrum speech. *J. Acoust. Soc. Am.*, 79:1580–1583, 1986.

Gilbert A. Soulodre, Theodore Grusec, Michael Lavoie und Louis Thibault. Subjective evaluation of state-of-the-art two-channel audio codecs. *J. Audio Eng. Soc.*, 46(3): 164–177, 1998.

Alexander A. Spector. On the mechanoelectrical coupling in the cochlear outer hair cell. *J. Acoust. Soc. Am.*, 107(3):1435–1441, März 2000.

Alexander A. Spector, William E. Brownell und Aleksander S. Popel. Nonlinear active force generation by cochlear outer hair cell. *J. Acoust. Soc. Am.*, 105(4):2414–2420, April 1999.

Samual D. Stearns und Don R. Hush. *Digitale Verarbeitung analoger Signale.* R. Oldenbourg Verlag, München, 1999.

W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17:401–419, 1952.

W. S. Torgerson. *Theory and Methods of Scaling.* Wiley & Sons, New York, 1958.

H. Traunmüller und F. Lacerda. Perceptual relativity in identification of two-formant vowels. *Speech Communication*, 6:143–157, 1987.

Hartmut Traunmüller. Analytical expressions for the tonotopic sensory scale. *PERILUS*, 8:93–102, Dezember 1988.

W. van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow und M. A. Stockes. Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.*, 84(3):917–928, September 1988.

Zuoying Wang und Feng Liu. Speaker adaptation using maximum likelihood model interpolation. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.* IEEE, 1999. Beitrag Nummer 1368.

R. M. Warren. Elimination of biases in loudness judgement for tones. *J. Acoust. Soc. Am.*, 48:1397–1403, 1970.

Richard M. Warren. *Auditory Perception.* Pergamon General Psychology Series. Pergamon Press Inc., New York, 1982.

Lutz Welling, Stephan Kanthak und Hermann Ney. Improved methods for vocal tract normalization. In *Proc. IEEE Internat. Conf. Acoust., Speech, Signal Process.* IEEE, 1999. Beitrag Nummer 1436.

Herbert S. Wilf. A method of coalitions in statistical discriminant analysis. In Kurt Enslein, Anthony Ralston und Herbert S. Wilf, Herausgeber, *Statistical Methods for Digital Computers*, Band 3 von *Mathematical Methods for Digital Computers*, Kapitel 6. John Wiley & Sons, Inc., New York, zweite Auflage, 1977.

S. A. Zahorian und A. J. Jagharghi. Speaker normalization of static and dynamic vowel spectra features. *J. Acoust. Soc. Am.*, 1(90):67–75, 1991.

S. A. Zahorian und A. J. Jagharghi. Minimum mean-square error transformation of categorical date to target positions. *IEEE Trans. Signal Proc.*, 1(40):13–23, 1992.

E. Zwicker. *Psychoakustik*. Springer-Verlag, Berlin, 1982.

E. Zwicker und E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, 68:1523–1525, 1980.

Danksagung

Bedanken möchte ich mich vor allem bei den vielen Menschen, die die besondere freie und produktive Atmosphäre des Dritten Physikalischen Institutes geschaffen und weitergetragen haben. Dazu gehört vor allem Prof. Dr. Manfred R. Schroeder, der mir die Möglichkeit zu dieser Arbeit gegeben hat und mir durch seine ungewöhnliche Synthese vielfältiger Interessen, fachlichen Wissens und persönlicher Bekanntschaften immer wieder erstaunliche Einblicke in Hintergründe, unbekannte Details und seltsame Zusammenhänge eröffnet hat. Ähnliches gilt für Dr. Hans Werner Strube, der diese Arbeit betreut hat, und für den es im weiten Umfeld unseres Arbeitsbereiches kaum eine physikalische oder mathematische Frage zu geben scheint, zu der er nicht Wesentliches beizutragen wüßte.

Dazu gehören auch die Diplomanden und Doktoranden der Arbeitsgruppen Professor Schroeders, die mir in vielen Dingen hilfreich zur Seite standen und von deren Kooperation, Wissen und Arbeit ich häufig profitieren konnte. Namentlich erwähnen möchte ich Dr. Kyrill Fischer, ohne den ich vermutlich nicht nach Göttingen gekommen wäre, Dr. Andreas Eilemann und Joachim Kaufmann, ohne die wir womöglich immer noch mit LaTeX2.09 arbeiten würden, Dr. René Koch und Dr. Dirk Püschel, ohne die das Signalverarbeitungssystem SI++, mit dem die Versuche dieser Arbeit durchgeführt wurden, nicht existieren würde, und Hansjörg Klock, der das System erweitert und gepflegt hat.

Außerdem möchte ich mich bei den derzeitigen und ehemaligen Mitgliedern der Arbeitsgruppe »Sprache und neuronale Netzwerke«, insbesondere Dr. Heiko Freienstein, und neben den bereits genannten, bei Dr. Dirk Michaelis, Rainer Blome, Jan Lessing und Olaf Schreiner für Freundschaft und gute Zusammenarbeit bedanken, speziell bei Joachim Kaufmann und Dr. Matthias Fröhlich für das Korrekturlesen dieser Arbeit und bei Dr. Holger Behme-Jahns für hilfreiche Anregungen und Auskünfte.

Ebenfalls entscheidenden Anteil an der Fertigstellung dieser Arbeit hat das »Sekretariat«, vor allem Edith Rohrmoser, Karin in der Beek und Martina Schulz, die jedes administrative Problem souverän lösen oder aus der Welt schaffen konnten.

Mein ganz spezieller Dank gilt den Mensabetrieben des Studentenwerks, die mich nicht nur mit Nahrung, sondern auch mit absolut faszinierenden Einblicken in die kulinarische Welt versorgten – Kreationen wie die dänische Frühlingsrolle, das vegetarische Chili-con-Carne oder die geräucherte Biotofuschnitte werde ich nie vergessen, genauswenig wie die zahlreichen innovativen Fischgerichte, die, ob verbrannt oder noch gefroren, »Dichtungsringe« (sog. Tintenfisch) oder »Trangeschmack in kugelsicherem Zerealienmantel« (Scholle), radikal mit dem Vorurteil aufräumten, daß Fisch

und Zitrone zusammengehören. Bedauerlich finde ich es, daß speziell die Küche der Zerntralmensa ihre Versuche der Findung der ultimativen Panade (sog. »Tagessuppe«, einem mehr oder weniger aromatisiertes Salz-Mehl-Konglomerat), noch immer nicht erfolgreich abschließen konnte, weswegen Pommes generell ungesalzen bleiben mußten. Umsomehr bewunderte ich immer die Kunstfertigkeit des Darreichungspersonals, welches beständig Wege fand, die Speisenkomponenten derart auf den Tabletts zu arrangieren, daß jeder Griff in die pastösen Nahrungsbestandteile gehen mußte, und deren Geschick Dinge wie »Eierteigwaffel unter Senf-Zwiebel-Sauce« oder der »Waterloo-Salat«, niedergestreckte Gurken mit Tomatendesaster an einer Kelle eigenen Saftes, zu verdanken sind. Als besonders rücksichtsvoll empfand ich hingegen immer das Quarkangebot, welches mit dem fiesesten Gericht versöhnen und dessen geschmackliche Signatur durch erfreulichen Genuß zu beseitigen verstand.

Index

Lebenslauf

Ich wurde am 11. April 1967 als zweiter Sohn von Peter und Helene Müller, geb. Schroth, in Frankfurt am Main geboren.

Nach dem Besuch verschiedener Grundschulen (Grundschule Frankfurt-Sindlingen 1973/74, Ludwig-Uhland-Schule Neu-Isenburg 1974/75, Grundschule Frankfurt-Nieder-Erlenbach 1975–77) wechselte ich auf die Otto-Hahn-Schule (Frankfurt), eine additive Gesamtschule ohne Oberstufe, die ich bis 1983 besuchte. Hiernach wechselte ich zum Besuch der Oberstufe an die Ernst-Reuter-Schule I (Frankfurt), wo ich 1986 mein Abitur erlangte.

Nachdem ich von 1986–88 meinen Zivildienst geleistet hatte, schrieb ich mich im April 1988 an der Frankfurter Johann-Wolfgang-Goethe-Universität zum Studium der Physik ein. Nach meiner Diplom-Vorprüfung im November 1989 setzte ich mein Studium an der Georg-August-Universität Göttingen fort. Hier fertigte ich am Dritten Physikalischen Institut (Schwingungsphysik) meine Diplomarbeit des Titels »Sprechernormalisierung auf Barkspektrogrammen durch lineare Abbildung im Sinne des kleinsten quadratischen Fehlers« an. Im Juli 1993 legte ich die Diplom-Hauptprüfung ab.

Seitdem bearbeitete ich das Thema weiter im Rahmen verschiedener Anstellungen als wissenschaftliche Hilfskraft und wissenschaftlicher Mitarbeiter, insbesondere im Rahmen des BMFT-Projektes SPINA bis 1994 und innerhalb eines von der DFG geförderten Projektes 1996 bis Mai 2000. Seit 1995 pflegte ich die am Institut entwickelte Signalverarbeitungssoftware SI++, für das ich auch im Akustikbüro Göttingen ein Selbstdokumentationsprogramm entwickelte. Außerdem war ich für das am Institut abgehaltene Blockpraktikum zur Digitalen Signalverarbeitung als Betreuer und Redakteur des zugehörigen Protokollbandes tätig.