Post Cochlea Processing

Mustererkennung zur Informationsextraktion
aus simulierten Aktionspotenzialen des Hörnervs

Dissertation
zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Technischen Universität Dortmund

Der Fakultät Statistik vorgelegt von
Gero Szepannek
Dortmund, im Oktober 2008

technische universität
dortmund

Gutachter:
Prof. Dr. C. Weihs
JProf Dr. U. Ligges

Tag der mündlichen Prüfung:
14. Oktober 2008

Vorwort

Die vorliegende Arbeit ist geprägt durch das Projekt „Post Cochlea Processing“, eine Kooperation des Fraunhofer Instituts für Digitale Medientechnologie (IDMT, Ilmenau) und des Lehrstuhls für Computergestützte Statistik der Technischen Universität Dortmund auf der einen Seite und andererseits das Projekt A4 „Statistik und maschinelles Lernen“ des DFG Sonderforschungsbereichs 475 zur „Komplexitätsreduktion in multivariaten Datenstrukturen“ an der Fakultät Statistik der Technischen Universität Dortmund.

Ein besonderer Dank gilt Prof. Dr. Claus Weihs für alle Möglichkeiten, die er mir eröffnet hat und die hilfreichen Ratschläge und Diskussionen auf dem Weg sowie Dr. Frank Klefenz, durch den dieses Projekt – und damit meine Berührung mit dem faszinierenden Thema der menschlichen Schallwahrnehmung – überhaupt erst ermöglicht wurde, für dessen unterstützende Begleitung. Dank gilt auch den weiteren Mitgliedern der Prüfungskommission: Prof. Dr. Katja Ickstadt, Dr. Matthias Arnold und dem Zweitgutachter JProf. Dr. Uwe Ligges, ohne den mein Rechner bestimmt nur halb so häufig das getan hätte, was ich von ihm wollte.

Besondere Erwähnung gilt auch meinen „DFÜ-Kollegen“ aus Ilmenau: Tamás Harczos, András Kátai und Stephan Werner sowie außerdem Yvonne Bäro für meine schönen Aufenthalte an der Ilm und zahlreiche fruchtbare Diskussionen, vor allem Tamás Harczos für unzählige Tips um Matlab- oder C-Programmierung – und für das Teilhaben-lassen an seinem schier unerschöpflichen „Readily Releasable Pool“ von Ideen. Bastian Schmitz, Olaf Mersmann und Sebastian Krey bin ich zu einem riesen Dank verpflichtet für wiederholte Tips bei den Versuchen, meinen Rechner dazu zu bekommen, das zu tun, was er soll. Karsten Lübke verdanke ich zahlreiche lehrreiche Diskussionen und eine nicht zu verachtende Excel-Tabelle. Matthias Budinger, Christian Röver, Kai Vogtländer und Günter Szepannek ist zu großen Anteilen die Lesbarkeit

dieses Werkes zu verdanken. Unbedingt zu erwähnen sind außerdem Kai Kammers, Gerd Kopp, Oliver Melsheimer, Tina Müller und besonders Nils Raabe wegen K^n (Kaffeetrinken, Kochen, Krökeln, ...) sowie diversen hilfreichen Korrekturvorschlägen. Und auch meine Kolleginnen Julia Schiffner, Katrin Sommer und Heike Trautmann dürfen an dieser Stelle nicht unerwähnt bleiben. Weiterhin bietet sich hier eine Gelegenheit, Ralf Schmidtke, Nadine Körwer, Michael Surmann und Matthias Budinger dafür zu erwähnen, mich in den vergangenen zehn Jahren mit Aspekten ihres Wesens nicht nur beeindruckt, sondern hoffentlich auch auf mich abgefärbt zu haben.
Ein Dank ganz anderer Natur geht an meine Eltern Margret und Günter Szepannek für die vielen Besuche in Dortmund, die mir den nötigen zeitlichen Freiraum zur Anfertigung des vorliegenden Werkes gaben und auch noch sehr viel anderes in den vergangenen 32 Jahren, sowie auch an Hans und Anneliese von Seggern. Ein abschließendes, dickes Dankeschön ist an meine Familie gerichtet: Leona, Cara und Katie, für das Ertragen zahlreicher geistesabwesender Minuten auch noch nach Feierabend.

Inhaltsverzeichnis

Zusammenfassung der wesentlichen Thesen

- Unterschiedliche Vorschläge auditorisch motivierter Merkmalsextraktion wurden auf den Output eines detaillierten, neurophysiologisch parametrisierten auditorischen Simulationsmodells übertragen und in einer Vergleichsstudie als Merkmale zur automatischen Spracherkennung (ASR) verwendet. Diese lassen sich untergliedern in
  - *Orts-Durchschnittsfeuerraten* (OD, Abschnitt 3.4),
  - *Phase Locking-* bzw. *Inter-Spike Intervall* basierte (ISI, Abschnitte 3.6 und 3.5) und
  - *Delay-Computing* basierte Merkmale (Abschnitt 3.9).

- Durch Betrachtung der Aktionspotenzial-Emissionszeitpunkte als Punktprozesse lassen sich die verschiedenen Merkmalsextraktionsansätze theoretisch motivieren (Abschnitt 3.8).

- Mit *regularisierter heteroskedastischer Diskriminanzanalyse* (RHDA, Abschnitt 5.3.3) wird eine Erweiterung zur *heteroskedastischen Diskriminanzanalyse* (vgl. Abschnitt 5.3.2) eingeführt, die eine Dimensionsreduktion und Dekorrelation der Merkmale ermöglicht und insbesondere zur Kombination verschiedener Merkmalssätze anwendbar ist.

- Der Einsatz von RHDA liefert eine deutliche Verbesserung der Spracherkennungsergebnisse (Abschnitt 6.3.6), insbesondere bei der Verwendung heteroskedastischer Kovarianzmatrizen und deren gleichzeitiger leichter Schrumpfung gegen Diagonalität.

- Die Untersuchungen ergeben, dass
  1. mit Phase Locking-basierten Merkmalen bessere Resultate erzielt werden als durch die Verwendung von Orts-Durchschnittsfeuerraten-Merkmalen (Abschnitt 6.3.7),
  2. jedoch weder Orts-Durchschnittsfeuerraten noch Phase Locking basierte Merkmale allein die besten Ergebnisse hervorrufen; sondern eine Kombination beider Merkmalstypen eine weitere Verbesserung der Erkennungsleistung liefert,
  3. anhand von Merkmalsextraktion auf Basis paralleler lokaler *Delay-Computing Netzwerke* (PLDCN) eine Verbesserung der Ergebnisse im Vergleich zur Verwendung einzelner *Delay-Computing Netzwerke* (DCN) erzielt werden kann,
  4. eine abermalige Verbesserung der Erkennungsergebnisse erzielt werden kann, kombiniert man PLDCN basierte Merkmalssätze mit Orts-Durchschnittsfeuerraten und Inter-Spike Intervall basierten Merkmalen.
  5. Hierbei erweist sich als Abwandlung des von Eisele u. a. (1996) vorgeschlagenen Vorgehens insbesondere eine dimensionsreduzierende Transformation *nach* $\Delta-$Koeffizientenbildung als effektiv.
  6. Abschließend kann gezeigt werden, dass sich durch *Kombination* von nicht-auditorisch-basierten MFCC-Standardmerkmalen der automatischen Spracherkennung *(Mel-Frequenz Cepstralkoeffizienten)* mit den aus dem auditorischen Simulationsmodell gewonnen Merkmalen eine deutliche Verbesserung der Erkennungsergebnisse erzielen lässt.
- Durchgeführte Signifikanztests (Abschnitt 6.3.11) stützen die oben beobachteten Hypothesen.

1 Einleitung

„Im Gegensatz zu bloßen Schallschwingungen entsteht beim Hören ein Geräusch im Kopf“, schreibt Jourdain (2001). Die menschliche Wahrnehmung von Schall ist bis heute nicht vollständig erforscht. Am Beispiel der Audiokodierung durch Formate wie das bekannte MPEG-1 Audio Layer 3 („`*.mp3`“, vgl. z.B. Brandenburg und Stoll, 1994) wird jedoch deutlich, dass für die menschliche Wahrnehmung nicht alle Aspekte der Originalzeitreihe erforderlich sind, um den Klang eines Musikstückes zu erfassen. Im Forschungsgebiet der Psychoakustik werden Hörexperimente durchgeführt, um darüber Aufschluss zu liefern, wie ein Schallsignal vom Menschen wahrgenommen wird (vgl. z.B. Howard und Angus, 2006).

Die Motivation dieser Arbeit hat ihre Wurzeln in der Vorstellung, dass die Art und Weise, in der Schall wahrgenommen wird, dadurch bedingt ist, wie das Schallsignal zunächst von der *auditorischen Peripherie* – d.h. Außen-, Mittel- und Innenohr bis hin zu den Hörnerven – verarbeitet wird.

Schwerpunkt der vorliegenden Arbeit bildet die Extraktion von Information (in Gestalt von Merkmalen) aus der simulierten Repräsentation von Schall am menschlichen Hörnerven. Von dieser Datengrundlage darf erhofft werden, dass sie diejenige Information nachempfindet, die auch für die menschliche Wahrnehmung von Schall vorliegt. Die Gestalt dieser Repräsentation manifestiert sich in Zeitpunkten, zu denen die verschiedenen Neuronen *Aktionspotenziale*, so genannte *Spikes*, freisetzen.

Als Untersuchungsgegenstand dient die Aufgabe der automatischen Erkennung (Klassifikation, Zuordnung) von gesprochener Sprache. Die Kernfragestellung dieser Arbeit befasst sich damit, in welchen Merkmalen oder Kenngrößen der neuronalen Antwort auf ein Sprachsignal die wesentliche Information verborgen liegt, die zur Erkennung von Sprache benötigt wird. Im Gegensatz zur Aufgabenstellung der reinen Extraktion des Grundtons aus einem Schallsignal, wie sie beispielsweise von Ligges (2006)

oder Heinz (2002) zur automatischen Melodietranskription durchgeführt wird, geht es bei der Erkennung von Sprache um Klassifikation auf Basis von Klangeigenschaften, da beispielsweise ein Vokal 'a' korrekt identifiziert werden soll, unabhängig davon, ob der Sprecher ein Mann, eine Frau oder ein Kind ist. Es ist jedoch zu erwarten, dass alle drei Sprecher Sprachsignale mit sehr unterschiedlichen Grundtönen produzieren.
Abbildung 1.1 veranschaulicht die verschiedenen Schritte, die, bis hin zur Bildung eines Spracherkennungsmodells (ASR Back Ends) durchlaufen werden, auf Basis von dessen Performance abschließend die unterschiedlichen Merkmale verglichen werden können. Diese Grafik wird in den folgenden Kapiteln wieder aufgegriffen, um eine leichtere Eingliederung der einzelnen Abschnitte in den Gesamtkontext der Arbeit zu ermöglichen.

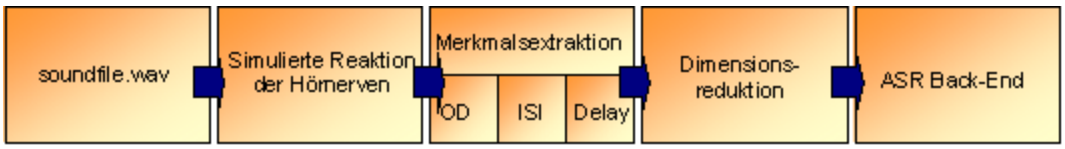


Abbildung 1.1: Entstehung eines Modells zur automatischen Spracherkennung (ASR) auf Basis des auditorischen Simulationsmodells.

Die Prozessschritte des in der Arbeit verwendeten auditorischen Simulationsmodells, sowie die Gestalt des entstehenden Output als Reaktion auf ein eingehendes Schallsignal werden kurz in Kapitel 2 beschrieben. Die einzelnen Verarbeitungsschritte sind ausführlich in Anhang A diskutiert.
Zum Zwecke der Spracherkennung wurden bislang zumeist nur sehr einfache auditorische Modelle eingesetzt, die spezielle Phänomene nachempfinden, deren Parameter jedoch keine unmittelbare neurophysiologische Interpretation besitzen (siehe z.B. Ghitza, 1988; Ali u. a., 2002). In den Arbeiten von Hemmert u. a. (2004) und Mamsch (2006) wird dagegen ein detailliertes, neurophysiologisches Simulationsmodell nach Sumner u. a. (2002) zur automatischen Spracherkennung verwendet, dessen wesentliche Prozessschritte auch für die Signalverarbeitung in den inneren Haarzellen im Simulationsmodell dieser Arbeit Verwendung finden. Untersucht werden dort allerdings lediglich durchschnittliche Feuerraten der Hörnerven als Merkmale zur weiteren Verarbeitung.
In Kapitel 3 werden unterschiedliche Merkmalssätze aus der auditorischen Repräsenta-

tion entwickelt und diskutiert. Hierfür werden in Abschnitt 3.2 zunächst gängige Hypothesen zur Informationskodierung vorgestellt, darunter insbesondere ein relativ neuer, durch Beobachtungen von Greenberg (1997) motivierter Ansatz basierend auf Information in der Gestalt der Wanderwellen-Delaytrajektorie entlang der Cochlea. Zunächst werden einige bekannte Ansätze zur Merkmalsextraktion (Ghitza, 1988; Seneff, 1988; Yang u. a., 1992; Ali, 1999; Hemmert u. a., 2004) auf die Gestalt des Outputs des detaillierten neurophysiologischen Simulationsmodells dieser Arbeit übertragen.

Eine Motivation der vorgestellten Merkmale ist möglich, fasst man die an den Hörnerven emittierten Aktionspotenziale als Punktprozesse auf und legt zudem eine weitere Klasse von Merkmalen nahe, wie sie in Abschnitt 3.9.1 vorgestellt wird: Mit Hilfe von *Delay-Computing Netzwerken* (DCNs, Brückmann u. a., 2004) wird Information über die Wanderwellengestalt extrahiert und damit für eine Weiterverarbeitung nutzbar gemacht. Eine Erweiterung zu *parallelen lokalen* Delay-Computing Netzwerken (PLDCNs, Szepannek und Weihs, 2006a) wird in Abschnitt 3.9.2 motiviert und eingeführt, und auf Basis dieser Repräsentationen werden weitere Merkmale vorgeschlagen.

Ein Aspekt dieser Arbeit von besonderer Relevanz besteht in der Kombination unterschiedlicher Merkmalssätze: Es stellt sich die Frage, ob die erzielten Resultate besser werden, wenn man zwei (oder mehrere) zu einem gemeinsamen Merkmalssatz zusammensetzt. Es kann untersucht werden, ob dieser ergänzende Information im Verhältnis zu den Originalmerkmalssätzen allein beinhaltet. Die Dimension des resultierenden gemeinsamen Merkmalsvektors ist jedoch i.d.R. deutlich höher und legt einen anschließenden Schritt einer Merkmalsdimensionsreduktion nahe. Verschiedene Methoden hierzu werden in Kapitel 5 beschrieben und diskutiert.

Sämtliche vorgestellten Merkmale werden in einer Vergleichsstudie zur automatischen Spracherkennung auf der TIMIT Sprachdatenbank eingesetzt. Die Erkennung einer Sprachsequenz erfolgt dabei mit Hilfe von *Hidden Markov Modellierung*, wie sie in Kapitel 4 vorgestellt wird.

Eine vergleichende Auswertung der unterschiedlichen Informationsextraktionsansätze aus Kapitel 3 auf Basis der mit ihnen erzielten Erkennungsraten erfolgt in Kapitel 6.

Zur Lektüre dieser Arbeit seien insbesondere die Kapitel 3, 5 und 6 nahegelegt, da dort die Merkmalsextraktion aus der simulierten Hörnervenantwort (Kapitel 3) so-

wie die anschließenden Merkmalstransformationen (Kapitel 5) beschrieben und verglichen (Kapitel 6) werden. Sie beinhalten damit die zentralen Ergebnisse dieser Arbeit. Für ein besseres Verständnis der Motivation der verschiedenen Merkmale ist die in Kapitel 2 und Anhang A beschriebene Gestalt der neuronalen Repräsentation eines Schallsignals am Hörnerven hilfreich. In Kapitel 4 werden Hidden Markov Modelle als nötiges Handwerkszeug zur Erstellung eines automatischen Spracherkennungssystems beschrieben.

2 Auditorisches Simulationsmodell

2.1 Einordnung

Von einem Simulationsmodell der *auditorischen Peripherie* – d.h. Außen- und Mittelohr, bis hin zu den Hörnerven – darf erhofft werden, dass es eingehenden Schall ähnlich erfasst, wie er dem Menschen zur Wahrnehmung vorliegt. Es existieren eine Reihe beobachteter Phänomene der menschlichen Wahrnehmung von Schall, die Abweichungen von Charakteristika des entsprechenden, physikalisch gemessenen Spektrums darstellen. Beispiele hierfür sind: Maskierungs- bzw. Verdeckungseffekte, wie sie z.B. zur Audiokompression ausgenutzt werden (vgl. z.B. Brandenburg und Stoll, 1994; Baumgarte, 2000), Adaption, Lautstärkewahrnehmung (vgl. z.B. Gold und Morgan, 2000, Kapitel 14 und 15) oder die Wahrnehmung eines virtuellen Grundtons (vgl. Anhang B.8). Viele dieser Phänomene lassen sich dabei durch Schritte in der neurophysiologischen Verarbeitungskette des Schallsignals entlang des Ohrs bis hin zum Hörnerven erklären.

Es existieren eine Reihe sogenannter *auditorischer Modelle*, die bekannte Phänomene auf Basis einfacher Transformationen der Originalzeitreihe oder des zugehörigen Spektrums phänomenologisch nachempfinden (Ghitza, 1988; Seneff, 1988; Slaney, 1988; Dau u. a., 1996; Ali, 1999). Beispiele für solche Modelle sind in den Abschnitten 3.5 und 3.6 beschrieben. Im Unterschied dazu wird in dieser Arbeit ein detailgetreues auditorisches Simulationsmodell verwendet, dessen einzelne Prozessschritte und Parameter eine neurophysiologische Entsprechung besitzen, bis hin zur Beschreibung einzelner Aktionspotenziale, die an den Hörnerven emittiert werden. Ein solches, neurophysiologisch

parametrisiertes Simulationsmodell wurde von Meddis (1986) vorgestellt und durch Sumner u. a. (2002) erweitert. Die dort vorgeschlagene Modellierung bildet die Grundlage des in dieser Arbeit verwendeten Simulationsmodells, dessen Gestalt im weiteren Verlauf dieses Kapitels kurz vorgestellt wird. In Anhang A erfolgt eine ausführliche Beschreibung und Diskussion der einzelnen Modellierungsschritte inklusive eigener Untersuchungen.
In Folgenden wird eine kurze, zusammenfassende Übersicht der modellierten Prozesskette gegeben. Weiterhin sind dort einige Beispiele für die Gestalt des Simulationsmodell-Outputs beschrieben, der im Weiteren die Basis zur Merkmalsextraktion bildet.

2.2 Anatomie des Ohrs

In Anatomiebüchern wird das Ohr in der Regel in drei große Abschnitte unterteilt: das Außenohr (Auris externa), das Mittelohr (Auris media), sowie das Innenohr (Auris interna). Abbildung 2.1 [1] zeigt den schematischen Aufbau. Das *Außenohr* umfasst die

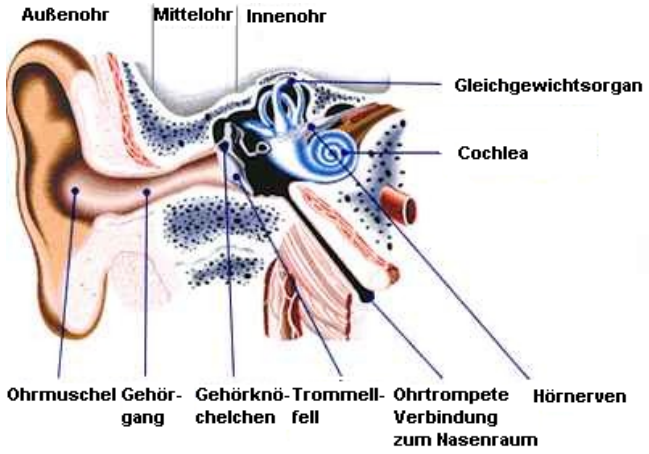


Abbildung 2.1: Schematische Darstellung des Ohrs.

Ohrmuschel, das Ohrläppchen und den äußeren Gehörgang und bildet einen Filter, der Schallwellen bis hin zum Trommelfell leitet. Der röhrenförmige Gehörgang bewirkt eine Verstärkung des Signals für einen Frenquenzbereich um etwa 2000-3000 Hz herum,

[1] Abbildung verwendet mit Erlaubnis von http://www.hoerforum.de/401.html.

die zu den charakteristischen frequenzspezifischen *Hörschwellen* führt (Abbildung 2.2[2], vgl. Szepannek u. a., 2005). Beim Menschen stellt die Ohrmuschel zudem einen richtungsselektiven Filter dar, der auch zur Lokalisation von Schallquellen genutzt wird (vergleiche Heinz, 2002, S. 41).

Das *Mittelohr* dient dazu, den Luftdruck der eingehenden Schallwellen am Trommelfell

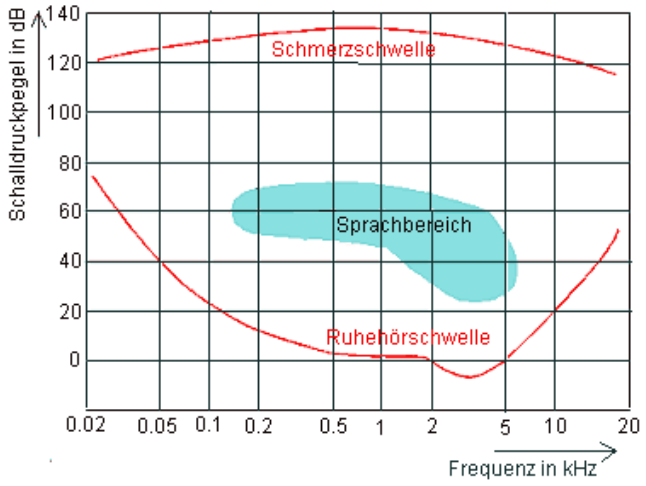


Abbildung 2.2: Ruhehörschwellen des Menschen.

auf die im Innenohr in der Gehörschnecke (Cochlea) befindliche Flüssigkeit (Lymphe) zu übertragen. Hierfür wird die Hebelwirkung einer Kombination von *Gehörknöchelchen* (*Hammer*, *Amboß* und *Steigbügel*) ausgenutzt, die an das *ovale Fenster* münden (vergleiche Heinz, 2002, S. 42 f).

Av ovalen Fenster wird die Schallwelle in die Cochlea übertragen. Das *Innenohr* lässt sich grob in zwei Bereiche unterschiedlicher Funktion unterteilen: das Gleichgewichtsorgan und die *Cochlea*. In der letzteren erfolgt die Umformung der eingehenden Schallwellen in Nervenimpulse an den *Hörnerven* (auditiven Nervenfasern, ANFs). Betrachtet man den Querschnitt der Cochlea (siehe Abbildung 2.3[3], links), so lässt sich diese in die drei Bereiche Scala Vestibuli, Scala Tympani und Scala Media unterteilen, die durch Membranen voneinander getrennt sind. In der inneren Scala Media ist das Cor-

[2]Abbildung verwendet mit Erlaubnis von www.dasp.uni-wuppertal.de
[3]Abbildung verwendet mit Erlaubnis von www.dasp.uni-wuppertal.de

tische Organ mit der *Basilarmembran* (BM) angesiedelt. Auf der BM liegen dabei drei Reihen *äußerer* (outer hair cells, OHCs) sowie eine Reihe *innerer Haarzellen* (inner hair cells, IHCs, Abbildung 2.3, rechts[4]). Während die Hauptaufgabe der OHCs sehr wahrscheinlich darin besteht, schwache Signale zu verstärken, ist die wesentliche Funktion der IHCs die Signalübermittlung in Richtung Gehirn (Gebeshuber, 2000). Hierzu muss die mechanische Schwingung in elektrische Impulse bzw. *Aktionspotenziale* (APs) oder auch *Spikes* umgewandelt werden.

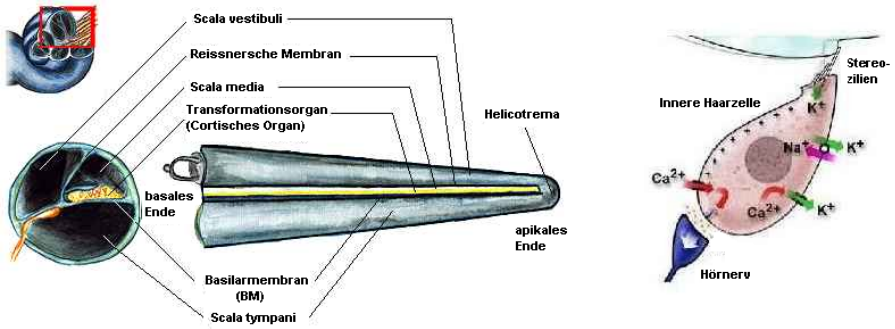


Abbildung 2.3: Cochlea (links oben) und ihre Querschnittsdarstellung (links unten) sowie der schematische Aufbau einer inneren Haarzelle (rechts).

Abbildung 2.4 stellt die Prozessschritte dar, mit denen eingehender Schall innerhalb des menschlichen Innenohrs weiterverarbeitet wird. Eine zusammenfassende Beschreibung der Verarbeitungsschritte findet sich in Szepannek u. a. (2005).

Abbildung 2.4: Schematische Darstellung der Signalverarbeitungs-Prozessschritte im Innenohr.

[4]Abbildung verwendet mit Erlaubnis von www.cochlee.org

2.3 Zusammenfassung der modellierten Prozessschritte

Die simulierte auditorische Verarbeitung eines eingehenden Schallsignals im Ohr erfolgt im Wesentlichen durch Modellierung der folgenden Prozessschritte:

1. Die Schallwelle wird von Außen- und Mittelohr einer Filterung unterzogen und durch die *Gehörknöchelchen* auf die im Innenohr befindliche Flüssigkeit übertragen.

2. Dort versetzen sie die *Basilarmembran, kurz BM* in Schwingung. Die Schallwelle wird entlang der Basliarmembran in viele unterschiedlich bandpass-gefilterte Signale zerlegt. Somit erfolgt eine *Frequenz-Orts-Transformation* des Schallsignals (vgl. Abb. 2.5).

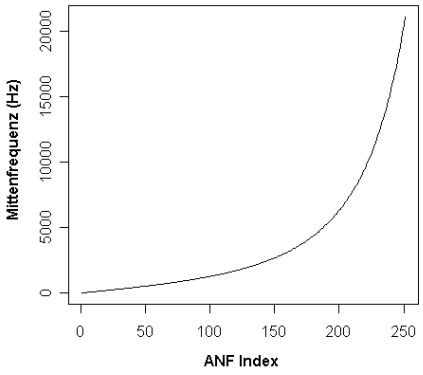


Abbildung 2.5: Frequenzauflösung entlang der Basilarmembran.

3. *Äußere Haarzellen* tragen zur Verstärkung des Signals bei und erhöhen dadurch den Dynamikbereich der menschlichen Wahrnehmung, während in den *inneren Haarzellen* eine Umwandlung der mechanischen Schwingungen in elektrische Impulse erfolgt, wie sie zur Generierung von elektrischen *Aktionspotenzialen* am Hörnerven erforderlich ist.

4. Hierzu schwingen zunächst auf der BM befindliche *Stereozilienbündel* (vgl. Abb. 2.3, rechts), die Kaliumkanäle öffnen und somit eine Veränderung des Membranpo-

tenzials der inneren Haarzelle (IHC) hervorrufen. Es kommt zu einem nichtlinearen Effekt saturierender Wellenamplituden bei starken Intensitätsleveln (vgl Abb. 2.6).

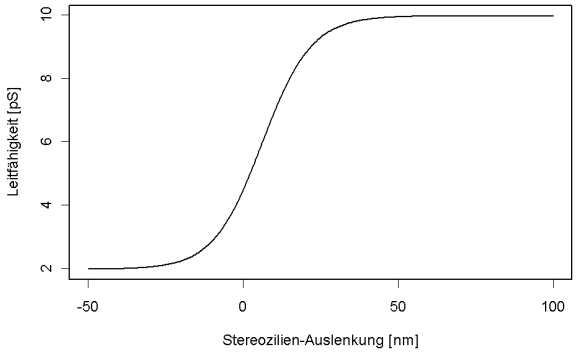


Abbildung 2.6: Funktionaler Zusammenhang zwischen Auslenkungsposition der Sterozilien und Ionenzufluss in die IHCs.

5. Das *Membranpotenzial* der inneren Haarzelle lässt sich durch einen Stromkreis beschreiben.

6. Am *präsynaptischen Ende* der inneren Haarzellen (vgl. Abb. 2.3, rechts) werden an sogenannten *aktiven Zonen* in Abhängigkeit der dortigen Konzentration von Calciumionen Neurotransmitter freigesetzt. Der Einfluss von Calciumionen ist dabei wiederum vom aktuellen Membranpotenzial abhängig und von höherer als linearer Ordnung.

7. Die Ausschüttung von Neurotransmittermolekülen in den *synaptischen Spalt* erfolgt aus sich verhältnismäßig langsam nachfüllenden *Pools*. Als Konsequenz stellt sich *Adaption* ein (siehe Abschnitt 2.5), d.h. auf ein neu einsetzendes Signal folgt zunächst unmittelbar eine hohe Ausschüttung von Neurotransmittervesikeln. Diese nimmt jedoch mit anhaltender Stimulusdauer ab.

8. Im synaptischen Spalt diffundieren die freigesetzten Neurotransmittermoleküle und binden an Rezeptoren an der postsynaptischen Membran des Hörnervs. Dort

führen sie zu einer schrittweisen Spannungserhöhung in sogenannten *miniature endplate potentials (mEPPs).*

9. Überschreitet das postsynaptsiche Potenzial einen Schwellenwert, so führt dies zur Freisetzung eines *Aktionspotenzials* (AP) oder *Spikes.*

10. Unmittelbar nach der Emission eines Spikes verfällt der Hörnerv in einen gehemmten Zustand neuronaler *Refraktärzeit*, der während einer Dauer von etwa 1-2 ms weitere Spikeemisionen verhindert und eine Tiefpassfilterung des Signals bewirkt.

Im Anhang B.8 ist anhand eines Beispiels des Phänomens des *virtuellen Grundtons* einer Sopranistin, wie ihn Ligges (2006) beobachtet, dargestellt, dass das auditorisch verarbeitete Schallsignal Rückschlüsse auf die Wahrnehmung von Frequenzen zulässt, die anhand des Spektrums der Originalschallwelle nicht unmittelbar erkennbar sind.

Im Folgenden sind einige Beispiele zur Veranschaulichung der auditorischen Signalverarbeitung dargestellt. Abbildung 2.7 zeigt die Veränderung einer eingehenden Schallwelle im Laufe der auditorischen Verarbeitung für das Beispiel eines Sinustons von 441 Hz Frequenz an der Basilarmembranposition (Nr. 55) mit maximaler beobachteter Auslenkung. Von der Auslenkung der BM (oben links) zur Stereozilienbewegung (oben rechts) ist eine Phasenverschiebung zu beobachten. Leicht zu erkennen ist die Halbwellengleichrichtung der Welle an der Leitfähigkeit $G(u(t))$ und der Membranspannung $V(t)$ der inneren Haarzelle (zweite Zeile) durch den saturierenden Ionenzufluss für hohe Auslenkungen der BM. Der nichtlineare Effekt des Potenzials der IHC auf Calciumzufluss ($ICa(t)$) und Neurotransmitteremissionswahrscheinlichkeit $k(t)$ ist in Zeile drei zu erkennen und äußert sich in zeitlich pointierten Peaks von $k(t)$. Diese haben eine recht präzise mit der Periode der Schallwelle einher gehende Neurotransmitterausschüttung zur Folge (unten links). Die resultierenden Spikes ergeben sich als Teilmenge der beobachteten Vesikel, so dass die Aktionspotenzialgenerierung eine probabilistische Tiefpassfilterung der Neurotransmittervesikelemission darstellt. Die aus den Inversen der mittleren Inter-Spike Intervalle dieses Fensters geschätzte Frequenz beträgt hier 433.7 Hz.

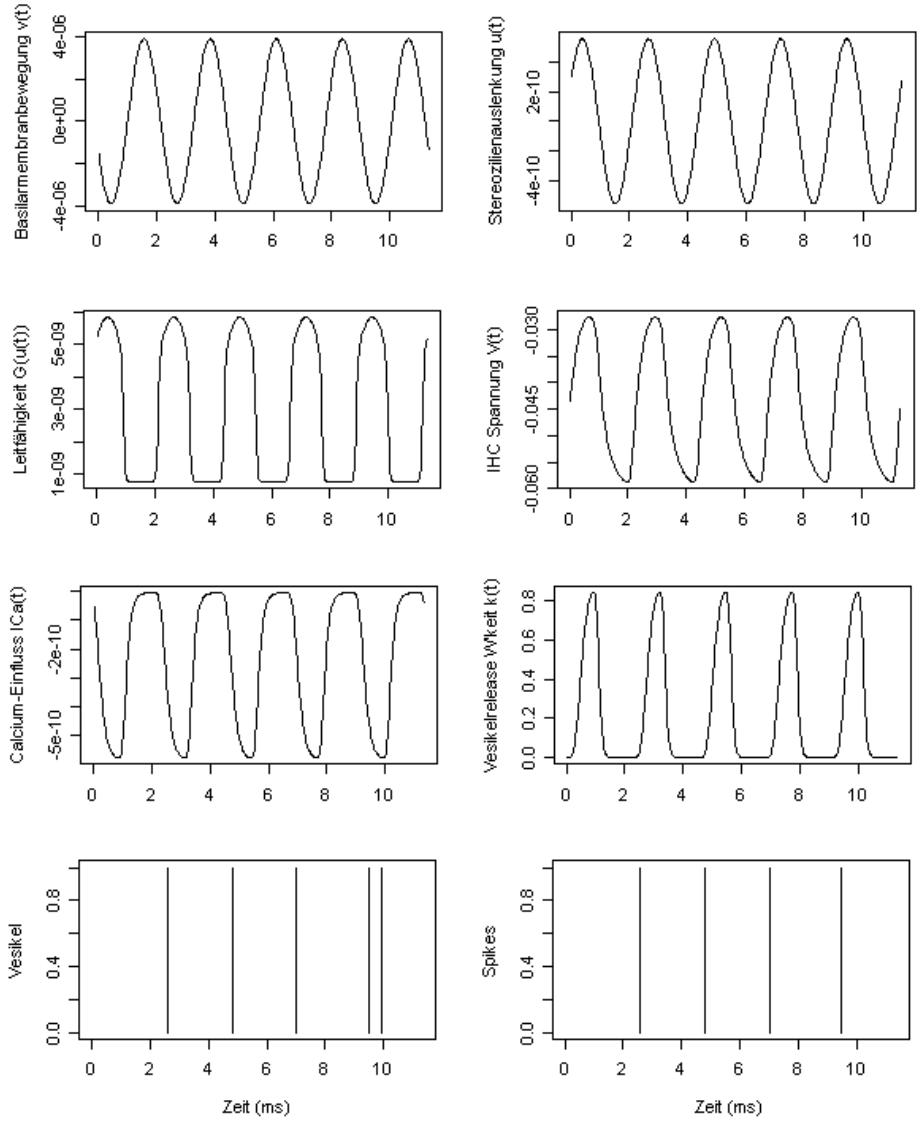


Abbildung 2.7: Output für unterschiedliche Prozessschritte des auditorischen Simulationsmodells auf einen Sinuston der Frequenz 441 Hz. (Für eine detaillierte Beschreibung der dargestellten Parameter siehe Anhang A.)

2.4 Gestalt des Simulationsmodell-Outputs

Die emittierten Aktionspotenziale bilden den Output $X_i(t)$ des Simulationsmodells. Dieser ist binär, es gilt $X_i(t) = 1$ wenn ANF i zum Zeitpunkt t ein AP emittiert und

0 sonst. Abbildung 2.8 zeigt das entstehende *auditorische Muster (oder auditorische Bild, engl. auditory image, AI)* für die Antwort der Hörnerven auf einen Sinuston der Frequenz 880 Hz von 100 ms Dauer. Nach 50 ms setzt ein zweiter Sinuston der Frequenz 440 Hz ein. Die Abszisse repräsentiert die Zeit während auf der Ordinate die Position der Cochlea (hohe Frequenzen oben, niedrige Frequenzen unten) abgetragen ist. Auftretende Aktionspotenziale sind durch helle Punkte markiert.

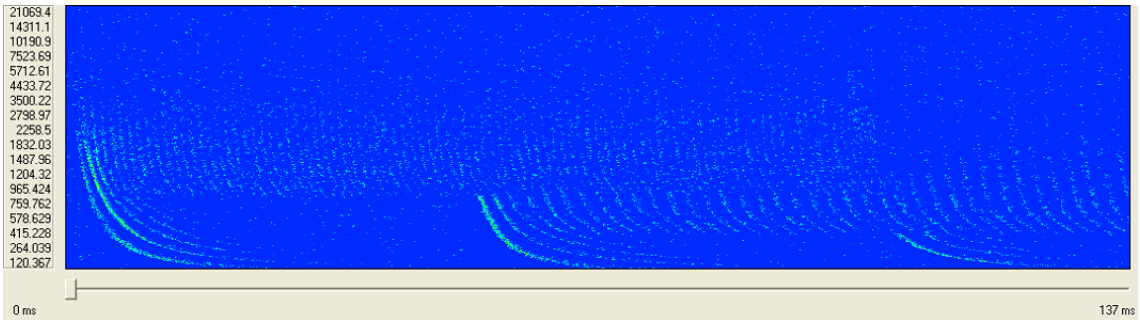


Abbildung 2.8: Simulierte Aktionspotenziale an den Hörnerven auf eine 880 Hz Sinuston von 100 ms Dauer. Nach 50 ms setzt ein zweiter Sinuston der Frequenz 440 Hz ein.

Zunächst lassen sich deutlich die *Delaytrajektorien* der *Wanderwellen* entlang der Cochlea als gekrümmte Linien identifizieren. Hierbei handelt es sich um die auf die in der Cochlea befindliche Flüssigkeit übertragenen Schallwellen, die die Cochlea entlang wandern (auf der Ordinate von oben nach unten). Auf der Abszisse ist ihr zeitlicher Verlauf zu erkennen, der von reziproker Gestalt ist. Weiterhin ist die *Frequenz-Orts-Transformation* erkennbar: Während des vorliegenden Tons von 880 Hz ist eine verstärkte AP Emission für andere ANFs zu beobachten als dies nach Einsetzen des 440 Hz Tones nach 50 ms der Fall ist. Das Antwortverhalten der Hörnerven verlagert sich hin zu ANFs mit niedrigerer Mittenfrequenz. Nach 100 ms – wenn der 880 Hz Sinuston verstummt – ändert sich dies erneut.
An den entsprechenden Hörnerven ist eine synchrone Antwort zu beobachten. Die Spikeemission erfolgt in nahezu konstanten Zeitabständen, die mit der Signalperiode einher gehen. Dieses Phänomen wird als *Phase Locking* bezeichnet (vgl. Abschnitt 3.2). Für die nicht angeregten Hörnerven ist die spontane Antwort in Form von zufällig

emittierten Aktionspotenzialen zu beobachten. Außerdem sind lange Trajektorien der Wanderwelle bei Signalveränderungen (Onset und Offset der einzelnen Sinustöne im Signal zu Beginn, und nach 50 und 100 ms) deutlich erkennbar.

2.5 Beispiel: Adaption

Aktionspotenziale werden jedoch nicht in jeder Signalperiode freigesetzt (vgl. Abb. 2.7, unten). Die dem Schallsignal zugrunde liegende Periode wird deutlich ersichtlich, bildet man ein *Post Stimulus Time Histogram* (PSTH, vgl. Anhang B.1) über viele Stimuluswiederholungen (vgl. Abb. 2.9, rechts, am Beispiel eines 100 Mal wiederholten 441 Hz Sinustons an der entsprechenden ANF). Dieses enthält den durchschnittlichen Anteil emittierter Spikes während kurzer Zeitabschnitte (Bins, hier je der Länge eines Zeitintegrationsschritts von 44100^{-1} s) über viele Wiederholungen des Stimulus (vgl. z.B. Johnson, 1980). In der Realität ist eine derartige Stimuluswiederholung natürlich nicht gegeben. Dafür besitzt der Mensch jedoch entlang der Cochlea etwa 3500 innere Haarzellen (anstelle von 251 inneren Haarzellen des verwendeten Simulationsmodells) mit denen jeweils acht bis zehn afferente Hörnerven verbunden sind (vgl. z.B. Allen, 1994; Yang u. a., 1992). Zur neuronalen Weiterverarbeitung steht dem Menschen somit auch ohne Stimuluswiederholung eine mehrfache Hörnervenantwort zur Verfügung.

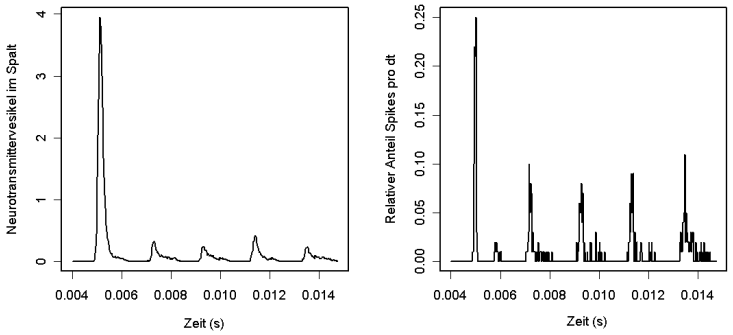


Abbildung 2.9: Adaption veranschaulicht am Beispiel des mittleren synaptischen Spaltinhalts $C(t)$ (links) und des PSTHs (rechts) über 100 Simulationen eines Sinustons von 441 Hz.

Abbildung 2.9 zeigt das Phänomen der Adaption zu Beginn eines 441 Hz Sinustons: Während der Verlauf der Vesikelemissionswahrscheinlichkeit $k(t)$ (vgl. Abb. 2.7) über die ersten Perioden konstant ist, leert sich in dieser Zeit der RRP. Dies ist am Inhalt des synaptischen Spalts zu beobachten (Abb. 2.9, links): Zu Stimulusbeginn werden mehrere Vesikel emittiert, die dem in der aktiven Zone befindlichen *Readily Releasable Pool (RRP)* bei weiterer Stimulusdauer fehlen (vgl. Anhang A.5). Folglich sinkt die durchschnittliche Neurotransmitterkonzentration in den folgenden Perioden. Insbesondere beträgt der durchschnittliche Neurotransmitterinhalt des synaptischen Spalts in den Folgeperioden weniger als ein Vesikel, was bedeutet, dass im Mittel nicht in jeder Stimulusperiode ein Spike emittiert wird. Diese beobachtete Adaption überträgt sich weiter auf das durchschnittliche Spikeverhalten der ANF (Abb. 2.9, rechts). Deutlich zu erkennen ist auch die neuronale Refraktärzeit unmittelbar nach der ersten Spitze des PSTH.

2.6 Beispiel: Maskierung

Neben der nichtlinearen Signalverstärkung bzw. Saturierung für hohe Intensitätslevel und der damit verbundenen Erweiterung des Dynamikbereichs scheint vor allem das auftretende Phänomen der Signal*maskierung* oder *Verdeckung* von Bedeutung. Die Kombination von Bandpassfilterung des Signals an der BM und sich leerenden Neurotransmitter-Pools (vgl. Abb. 2.9) führt dazu, dass ein schwächeres oder später einsetzendes Signal ähnlicher Frequenz von einem sogenannten *Maskierersignal* unterdrückt wird. Untersuchungen über Wahrnehmungs-Schwellenwerte existieren aus dem Gebiet der *Psychoakustik* und sind beispielsweise in Gold und Morgan (2000) oder Baumgarte (2000) beschrieben. Abbildung 2.10 (a) zeigt diesen Effekt in der neuronalen Antwort auf einen Grundton 'G' der Frequenz 392 Hz, sowie einen 50 ms später einsetzenden, 15 dB leiseren Ton 'A' der Frequenz 440 Hz. Der später einsetzende Ton ist in der Originalschallwelle (oben) deutlich zu erkennen. Im entstehenden auditorischen Muster (unten) ist nicht ersichtlich, dass hier zwei unterschiedliche Frequenzen präsent sind. Einzig erkennbar ist eine lange Wanderwelle, die auf den einsetzenden zweiten Ton folgt. Aus der Psychoakustik ist bekannt, dass bei maskierten Tönen häufig ein „Klicken“ die einzig wahrgenommene Veränderung im Schallsignal ist (vgl.

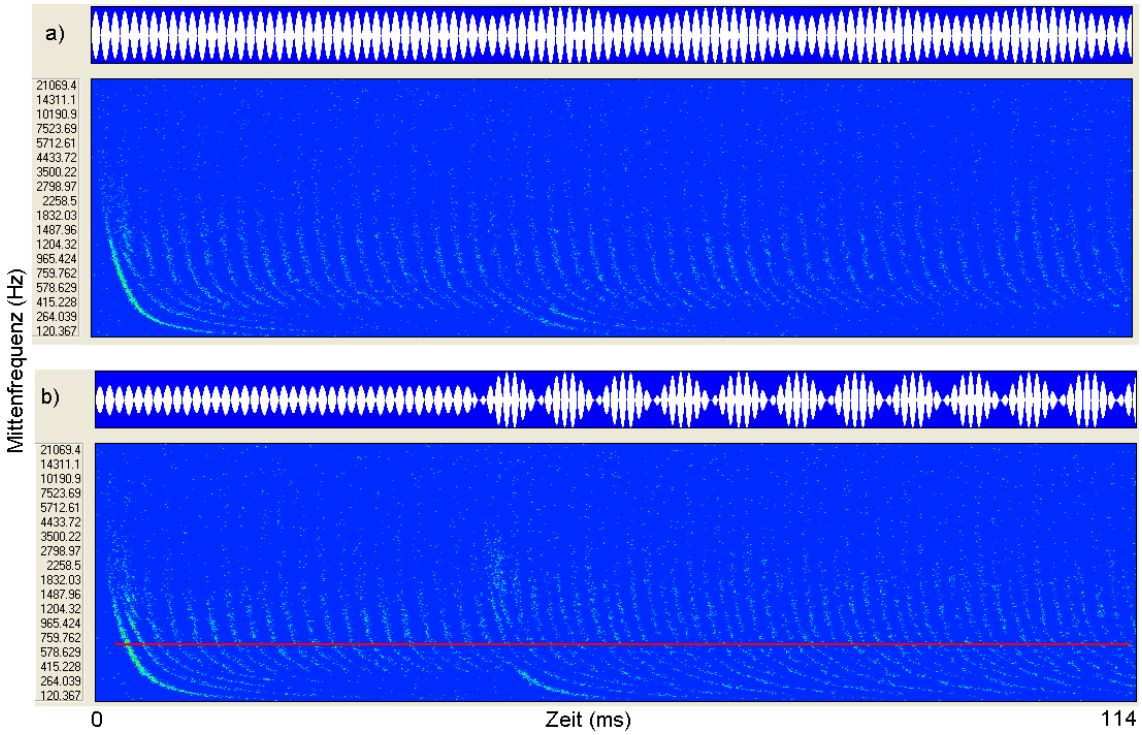


Abbildung 2.10: Beispiel des Maskierungseffekts. a) Auf einen Grundton G der Frequenz 392 Hz folgt nach 50 ms ein 15 dB leiserer Ton A der Frequenz 440 Hz. Originalschallwelle: oben, auditorisches Bild: unten. b) Auf einen Grundton G der Frequenz 392 Hz folgt nach 50 ms ein gleichlauter Ton C der Frequenz 523 Hz.

Gold und Morgan, 2000, S. 210). Der untere Teil (b) der Abbildung zeigt ein anderes Verhalten der Hörnerven: erneut schwingt zunächst der Grundton 'G'. Nach 50 ms setzt hier ein höherer Ton 'C' mit einer Frequenz von 523 Hz und gleicher Lautstärke ein. Es ist zu sehen, dass einige weiter vorne gelagerte Hörnerven in dessen Folge mit geringeren Zeitintervallen Aktionspotenziale freisetzen (oberhalb der roten Linie), während andere Neuronen mit niedrigerer Mittenfrequenz ihre Intervallänge beibehalten. Anhang B.7 beinhaltet eine kurze Beschreibung der Frequenzmaskierung.

2.7 Beispiel: Vokale

Als letztes Beispiel soll die neuronale Antwort innerhalb von 20 ms langen Zeitfenstern von vier verschiedenen Vokalen (/aa/, /ae/, /iy/ und /uw/) männlicher Sprecher der TIMIT Datenbank (vgl. Abschnitt 6.2.1) fungieren.

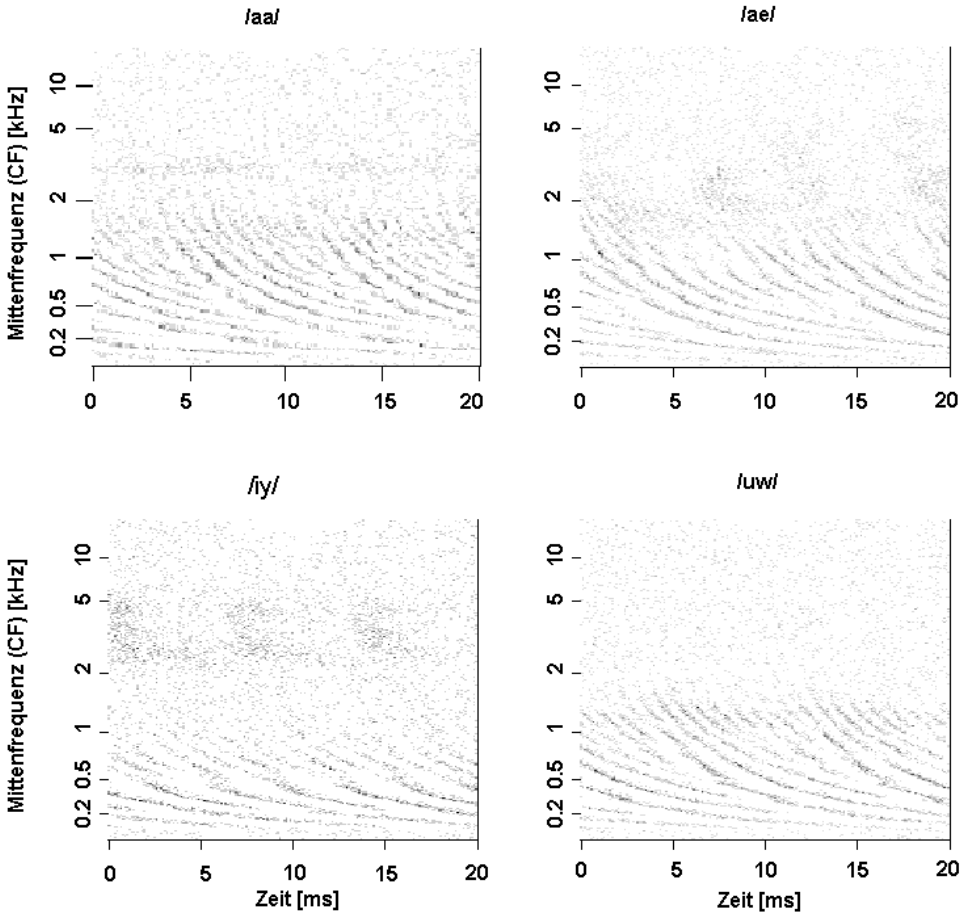


Abbildung 2.11: Output des auditorischen Simulationsmodells für vier unterschiedliche Vokale (/aa/, /ae/, /iy/ und /uw/) der TIMIT Datenbank (amerikanisches Englisch, siehe Kapitel 6.2.1) im 20 ms Fenster im Zentrum.

Es ist anzumerken, dass die zu beobachtenden Periodizitäten Information über die Grundfrequenz beinhalten, diese jedoch beispielsweise beeinflusst wird vom Geschlecht des Sprechers und stark variieren kann für verschiedene Repräsentationen des gleichen Vokals. Zur automatischen Spracherkennung sind eher Klangcharakteristika von Be-

deutung als Information über die Tonhöhe (vgl. z.B. Schukat-Talamazzini, 1995, S. 59). Betrachtet man die exemplarischen Repräsentationen der vier verschiedenen Vokale in Abbildung 2.11, so lassen sich auf den ersten Blick Unterschiede ausmachen. Beispielsweise zeigt sich im Simulationsmodell-Output auf den Vokal /iy/ eine verstärkte Aktivität der Hörnerven mit Mittenfrequenzen zwischen zwei und etwa fünf kHz. Die resultierenden neuronalen Muster (AIs) des auditorischen Simulationsmodells der Vokale /aa/ und /ae/ wirken einander ähnlicher als die der anderen beiden Vokale. Es entsteht außerdem der Eindruck, dass die beobachteten *Delaytrajektorien* der Wanderwellen entlang der Basilarmembran für die vier Vokale nicht den selben Verlauf aufweisen.
Das erklärte Ziel dieser Arbeit besteht darin, Merkmale im auditorischen Muster zu identifizieren, die relevante Information enthalten, die zur automatischen Erkennung von Sprachsignalen nutzbar sind. Hierzu werden im folgenden Kapitel Prinzipien und Verarbeitungsschritte beschrieben, die letztendlich zur Definition unterschiedlicher, aus dem auditorischen Muster (AI) extrahierter, Merkmalssätze führen.

3 Merkmalsextraktion

3.1 Übersicht

Es existieren verschiedene Hypothesen über die Kodierung von Information in der Aktivität der Hörnerven. Um diese zu untersuchen, werden in diesem Kapitel Merkmalsextraktions-Prinzipien vorgestellt, die sich diese zu Nutzen machen. Ihre Anwendung zum Zwecke *automatischer Spracherkennung* (engl. *automatic speech recognition*, ASR) soll in dieser Arbeit verwendet werden, um auf Basis der erzielten Erkennungsergebnisse einen Vergleich der Effizienz der auf den verschiedenen Hypothesen beruhenden Merkmale anstellen zu können.
Im Folgenden seien zunächst die Grundprinzipien zur Modellierung in der automatischen Spracherkennung beschrieben.
Unter den Prozessschritten der automatischen Spracherkennung wird zwischen *Front Ends* und *Back Ends* unterschieden. Als Back Ends werden diejenigen Methoden bezeichnet, innerhalb derer der eigentliche Erkennungsprozess vollzogen wird durch Klassifikation der Merkmalsrepräsentation in ein transkribierbares Alphabet, z.B. in Phonemsequenzen. In den vergangenen Jahren haben sich für diesen Zweck insbesondere *Hidden Markov Modelle (HMMs)* etabliert (siehe Kapitel 4). Front Ends dagegen erzeugen zunächst die hierfür verwendete Merkmalsrepräsentation aus der Input-Zeitreihe eines akustischen Signals. Dies erfolgt in gleichen Zeitabständen von etwa 10 ms für überlappende Zeitfenster von ca. 25 ms Dauer, während derer das Signal als näherungsweise stationär angesehen wird. Beide Größen entsprechen dabei typischen Werten im

Bereich der automatischen Spracherkennung (vgl. z.B. Gold und Morgan, 2000, S. 324) und sind für diese Arbeit entsprechend Young u. a. (2005), S. 56 f gewählt.
In diesem Kapitel werden Methoden vorgestellt, wie aus dem Output des auditorischen Simulationsmodells Merkmalsrepräsentationen als Front Ends zur Weiterverarbeitung mit Hilfe von Hidden Markov Modellen gewonnen werden können.
Der folgende Abschnitt 3.2 beschreibt zunächst die gängigen Hypothesen über die Kodierung von Information in der Aktivität der Hörnerven.
In den Abschnitten 3.4 bis 3.6 werden dann mehrere bekannte auditorisch basierte Ansätze zur Merkmalsextraktion vorgestellt und auf das in dieser Arbeit verwendete Simulationsmodell übertragen. Eine Motivation der vorgestellten Merkmalssätze ergibt sich durch den probabilistischen Ansatz der Auffassung der neuronalen Aktivität der Hörnerven als Punktprozesse (Abschnitt 3.8). In Abschnitt 3.9 werden weitere neue Alternativen zur Merkmalsextraktion vorgestellt: Dazu werden zunächst in Abschnitt 3.9.1 Delay-Computing Netzwerke eingeführt, die zur Identifikation von Verlaufs der vorliegenden Cochlea Wanderwellen dienen. Aus deren Output werden verschiedene Merkmale motiviert.
Vorab werden jedoch in Abschnitt 3.3 Cepstralkoeffizienten des Mel-skalierten Spektrums (MFCCs) eingeführt, deren Grundlage zwar nicht der Output des auditorischen Simulationsmodells darstellt, die aber im Bereich der automatischen Spracherkennung sehr populär sind aufgrund der guten, bei ihrere Verwendung erzielten Ergebnisse.

3.2 Information in der Hörnervaktivität

Es bestehen zwei wesentliche Theorien zur Kodierung von Information in den Aktionspotenzialen, die von den Hörnerven generiert werden: *Orts-Information* und Information, die in den *Inter-Spike Intervallen* enthalten ist.

Inter-Spike Intervalle (ISIs) zwischen aufeinanderfolgenden Aktionspotenzialen eines Hörnervens folgen, wie im letzten Kapitel beschrieben wurde, im Wesentlichen der Stimulusperiode. Dieses Phänomen des *Phase Lockings* wurde erstmals von Rose u. a. (1967) beschrieben und gilt zumindest für niedrige Frequenzen. Für höhere Frequenzen (oberhalb von 1 kHz) nimmt die Synchronisation der Hörnerven mit dem Ein-

gangssignal jedoch stark ab (Johnson, 1980). Aus diesem Grund kann Phase Locking zumindest nicht als alleiniger Informationsträger in Frage kommen. Zwei Ansätze, die auf dem Phänomen des Phase Lockings zur Merkmalsextraktion auf einfachen phänomenologischen auditorischen Modellen beruhen, gehen auf Ghitza (1988) und Seneff (1988) zurück. Beide werden in den Abschnitten 3.5 und 3.6 beschrieben und auf die Gestalt des Simulationsmodell-Outputs dieser Arbeit übertragen. Heinz (2002) am Fraunhofer IDMT nutzt Phase Locking eines auditorischen Simulationsmodells zur Konstruktion eines Algorithmus zum Information Retrieval in Musikdatenbanken.

Orts-Durchschnittsfeuerraten Kodierung: Die ursprüngliche Hypothese neuronaler Informationsübertragung, wie sie auf Adrian (1928) zurückgeht, beruht auf der Beobachtung, dass stimulierte Nervenzellen mehr Aktionspotenziale freisetzen als solche, die nicht stimuliert werden. Sie feuern Spikes mit einer höheren *durchschnittlichen Feuerrate*. Als Rate wird in diesem Kontext die Anzahl freigesetzter Spikes während eines kurzen Zeitintervalls angesehen. Während dieses Intervalls werden die exakten Zeitpunkte der Spikes nicht weiter berücksichtigt, im Englischen findet man aus diesem Grund auch häufig die Bezeichnung *mean rate* (siehe z.B. Seneff, 1988). Durch die Bandpass-Frequenzselektivität entlang der Basilarmembran kann jeder Hörnerv mit einem spezifischen Frequenzbereich, respektive mit seiner zugehörigen Mittenfrequenz (CF) assoziiert werden. Die Durchschnittsfeurraten der Hörnerven lassen sich somit als Amplituden einer spektralen Repräsentation interpretieren, die zuvor eine Filterung durch das Gehör erfahren hat. Sachs und Young (1979) untersuchten die Repräsentation einer solchen Durchschnittsfeuerraten-Repräsentation auf Stimuli unterschiedlicher Vokale an Hörnerven von Katzen. Sie beschreiben, dass die Formanten[1] in dieser Darstellung gut erkennbar sind, dass jedoch durch das Saturierungsverhalten für laute Signale das Spektrum stark geglättet wird (siehe auch Anhang A) und in Folge dessen spektrale Gipfel nicht mehr klar identifiziert werden können. Eine alleinige Kodierung von Frequenzen durch durchschnittliche Feuerraten an unterschiedlichen ANFs ist aus diesem Grund auch unwahrscheinlich.

[1] Als *Formanten* werden Bereiche des Frequenzspektrums mit erhöhter Resonanz bezeichnet. Diese werden erzeugt durch die Stellung des Vokaltrakts, den die an den Stimmbändern in Schwingung versetzte Schallwelle passiert. Im Gegensatz zur Grundfrequenz bleiben diese beobachtbaren Frequenzbereiche mit erhöhter Amplitude auch für unterschiedliche Sprecher beim selben Laut ähnlich, unterscheiden sich aber von Vokal zu Vokal (vgl. Schukat-Talamazzini, 1995, S. 51).

Eine weitere Art der Informationskodierung soll in dieser Arbeit untersucht werden und basiert auf der Gestalt der **Wanderwellen** entlang der **Cochlea**, wie sie in Greenberg u. a. (1997) und Greenberg (1997) beschrieben wird. Diese lassen sich exemplarisch in Abbildung 2.8 nachvollziehen. Die Zeitdauer bis zum Auftreten an einer bestimmten Position der Basilarmembran weist einen reziproken Verlauf auf. Am Anfang der Cochlea, in Bereichen, die stark auf sehr hohe Frequenzen ansprechen, besitzt die Welle noch eine hohe Geschwindigkeit. Sie verlangsamt dann stark nach Erreichen des Orts ihrer maximalen Amplitude. Die Amplitude der Welle nimmt ab diesem Punkt sehr schnell ab. Unterschiedliche Signale produzieren demnach ein verschiedenes Aussehen des Wanderwellenverlaufs. Pfeiffer und Kim (1975) und Kim und Molnar (1980) beobachten dieses Phänomen für Töne unterschiedlicher Frequenz und Lautstärke. Greenberg (1997) beschreibt die Gestalt der Welle als stabil, d.h. sehr ähnlich über unterschiedliche Schalldrucklevel und formuliert die Hypothese, dass die Verzögerungsstruktur der Welle über einen weiten Dynamikbereich und unterschiedliche akustische Bedingungen Information beinhaltet.

Im Abschnitt 3.9.1 wird ein künstliches neuronales *Delay-Computing Netzwerk* beschrieben, das dazu dienen soll, verschiedene Gestalten der Cochlea-Wanderwelle zu identifizieren und diese damit als Informationsträger zur Merkmalsextraktion nutzbar zu machen. Merkmale werden hieraus in Abschnitt 3.9.3 abgeleitet.

3.3 Mel-Frequenz Cepstral Koeffizienten (MFCC)

Cepstralkoeffizienten des Mel-skalierten Spektrums (Mel frequency cepstral coefficients, MFCCs, Davis und Mermelstein, 1980) sind aufgrund der guten Ergebnisse, die mit ihnen erzielt werden, die momentan meistverwendeten Merkmale zum Zweck der automatischen Spracherkennung. Deswegen soll hier vorab das Prinzip zu ihrer Bildung vorgestellt werden, obgleich ihre Entstehung weniger an der auditorischen Verarbeitung des Schallsignals orientiert ist als an der *Theorie der Sprachproduktion* (*Source-Filter Modell, Fant, 1960)*. Hiernach ergibt sich ein Sprachsignal als Faltung

einer Schallanregung an den *Glottis*[2] mit demjenigen Filter, den der – je nach Laut unterschiedlich geformte – Vokaltrakt bildet. Das zugehörige Spektrum eines Sprachsignals $A(\omega)$ ergibt sich damit als Produkt der Anregung $E(\omega)$ mit der Antwort des Resonanzsystems $R(\omega)$, den der Vokaltrakt darstellt:

$$A(\omega) = E(\omega) * R(\omega). \tag{3.1}$$

Von der Anregung wird hierbei angenommen, dass stimmlose Laute ein flaches Spektrum besitzen, sowie dass sonorante Laute zusätzlich die Grundfrequenz (je nach Sprecher zwischen 50 und 400 Hz), die durch die Glottisschwingung erzeugt wird (vgl. Schukat-Talamazzini, 1995, S. 32 ff) enthalten. Der für die Analyse interessante Teil des Signals spiegelt sich in den Resonanzfrequenzen des Vokaltrakts wider, die je nach gesprochenem Laut und damit verbundener Form des Vokaltrakts variieren. Eine Beschreibung der Verarbeitungsschritte zur Berechnung von MFCCs ist beispielsweise in Willett (2000), S. 117 ff, zu finden.
Durch Logarithmierung ändert sich zunächst das Produkt aus Anregung und Filter in eine Summe:

$$\log A(\omega) = \log E(\omega) + \log R(\omega). \tag{3.2}$$

Eine anschließende Invertierung der Fourier Transformation (üblicherweise durch die *Inverse diskrete Cosinus Transformation, IDCT,* berechnet die „spektralen Komponenten des logarithmierten Spektrums“. Durch diese Transformation werden die spektralen Erhöhungen der Anregung (in den oberen Koeffizienten) und die Vokaltraktresonanzen (niedrigere Koeffizienten) von einander separiert (siehe Schukat-Talamazzini, 1995, S. 58 f). Die sich ergebenden Koeffizienten besitzen zudem im Gegensatz zum ursprünglichen Sprachspektrum nur noch eine niedrige Korrelation (Merhav und Lee, 1993).
Die cepstralen Koeffizienten $C(q)$ ergeben sich als

$$C(q) = \sqrt{4/N} \sum_{\omega=0}^{N/2-1} \log(|A(\omega)|) \cos\left(\frac{\pi q(2\omega+1)}{N}\right), \tag{3.3}$$

$q = 1, \ldots, N/2$, wobei N die Sampleanzahl pro Zeitfenster darstellt (siehe Schukat-Talamazzini, 1995, S. 58) und somit $N/2$ die Anzahl an Fourier Frequenzen.

[2] Als Glottis wird die schmale Ritze bezeichnet, die zwischen den Stimmbändern im Kehlkopf entsteht, wenn diese sich in der Luftröhre verengen.

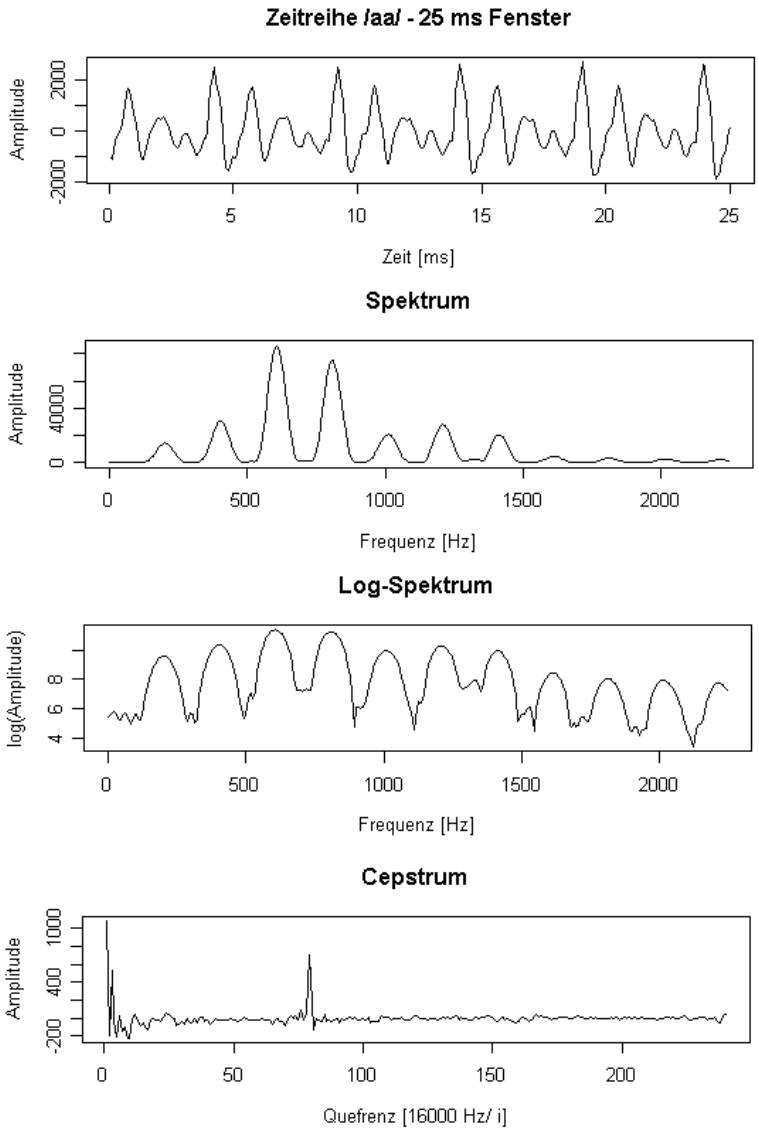


Abbildung 3.1: Cepstrale Transformation, von oben nach unten: Zeitreihe, Kurzzeitspektrum, logarithmiertes Kurzzeitspektrum sowie cepstrale Koeffizienten eines Vokals /aa/ einer weiblichen Sprecherin).

Der Begriff „Mel“ im Namen der MFCCs rührt dabei von der Tatsache her, dass das Spektrum des Sprachsignals zunächst – noch vor der Logarithmierung – einer nichtlinearen Transformation entlang der Frequenzachse unterzogen wird: die Frequenzachse

wird dabei äquidistant entlang der Mel-Skala (siehe Anhang B.5) in etwa 20 gleichgroße Abschnitte unterteilt. Die gewichteten Amplituden der Fourierfrequenzen werden zu „Mel-Filterbank Koeffizienten“ aufsummiert, wobei als Gewichte Dreiecksfilter verwandt werden, die zur Hälfte überlappen (siehe z.B. Young u. a., 2005, S. 59 f).

Häufig werden die ersten zwölf MFCC Koeffizienten als Front end zur automatischen Spracherkennung verwendet (Schukat-Talamazzini, 1995, S. 58 f), ergänzt um ihre ersten und zweiten Differenzen (über den Zeitverlauf von mehreren Zeitfenstern gebildet). Bei Hinzunahme der Signalenergie als zusätzlichem Merkmal ergibt sich in diesem Fall ein 39-dimensionaler Merkmalsraum.

Es ist anzumerken, dass MFCCs im Wesentlichen nicht auf Prinzipien der Sprachwahrnehmung, sondern genau entgegengesetzt auf solche der Spracherzeugung zurückgehen, wobei die Frequenzgruppenbildung entlang der Mel-Skala zusätzlich der humanen Frequenzunterschiedswahrnehmung Rechnung trägt.

Der oben beschriebene „cepstrale Trick“ aus Logarithmierung und anschließender IDCT kann prinzipiell genauso auf auditorisch basierte Merkmale angewandt werden, sofern diese im Sinne eines Spektrums interpretierbar sind.

Abbildung 3.1 veranschaulicht die Schritte der cepstralen Transformation. In der oberen Abbildung ist die Originalzeitreihe abgebildet. Die zweite Grafik von oben zeigt die zugehörige Spektraldarstellung: Gut zu erkennen sind die Harmonischen als ganzzahlige Vielfache der Grundfrequenz von etwa 200 Hz – ebenso wie die *spektrale Einhüllende*, die sich durch eine Glättung um die Amplituden der Harmonischen ergibt. Die dritte Grafik von oben zeigt das logarithmierte Kurzzeitspektrum. Auch höhere Vielfache der Grundfrequenz sind noch deutlich erkennbar. Die untere Grafik beschreibt die per IDCT aus dem logarithmierten Spektrum berechneten cepstralen Koeffizienten: Die durch ihre Vielfachen im Spektrum periodisch vorkommende Grundfrequenz ist deutlich als Gipfel für den Koeffizienten (= Quefrenz) 79 (entsprechend einer Frequenz von $16000/79 \sim 202$ Hz) auszumachen. Durch Betrachtung lediglich der niedrigen cepstralen Koeffizienten trennt man die Information der spektralen Einhüllenden von der Grundfrequenz und ihren Harmonischen.

3.4 Orts-Durchschnittsfeuerraten Merkmale

Die einfachste Art, spektrale Information über das Sprachsignal aus der neuronalen Antwort der ANFs zu extrahieren, besteht in einer *Orts-Durchschnittsfeuerraten Kodierung* (vgl. Abschnitt 3.2). Die Merkmale ergeben sich als durchschnittliche Feuerraten der in einem betrachteten Zeitfenster beobachteten Spikes, hochgerechnet auf 1 Sekunde:

$$X_i^{OD} = \sum_{t \in \text{Zeitfenster}} X_i(t)/\text{Fensterlänge} \tag{3.4}$$

mit $X_i(t)$ als Zustand der ANFs wie in Abschnitt 2.4 (1 für ein Spike von ANF i zum Zeitpunkt t und 0 sonst).

Durch die den ANFs zugehörigen Mittenfrequenzen $CF(i)$ (vergleiche Abschnitt A.2) lassen sich die X_i^{OD} den $CF(i)$ im Sinne von Amplituden zuordnen, und es entsteht eine spektrale Kurzzeitrepräsentation der vom auditorischen System verarbeiteten Schallwelle. Die Physiologie der Cochlea bewirkt dabei eine implizite Bark-Skalierung. Perdigao und Sa (1998) beschreiben, dass das Saturierungsverhalten der durchschnittlichen Feuerraten (vgl. Kapitel 2 bzw. Anhang A) neurophysiologisch einen Effekt erzielt, der mit der Logarithmierung des Spektrums innerhalb einer cepstralen Transformation vergleichbar ist.

Unterschiede möglicher Implementierungen bestehen im Umfang des betrachteten Frequenzbereichs oder der Art der Gruppierung der Hörnerven (ANFs) zu Merkmalen. Seneff (1988) und darauf aufbauend Ali u. a. (2002) schlagen die Betrachtung von durchschnittlichen Feuerraten für einen Frequenzbereich von $[200, 6400]$ Hz vor mit Bandpass-Filterbänken im Abstand von 0.5 Bark. Die Frequenzauflösung des verwendeten Simulationsmodells dieser Arbeit beträgt 0.1 Bark (vgl. Abschnitt A.2), so dass entsprechend der Mittelwert der Feuerraten von je fünf benachbarten ANFs als Merkmalsrepräsentation verwendet werden kann. Es ergibt sich als

$$X_k^{OD_1} = 1/5 \sum_{l=1}^{5} X_{15+k*5+l}^{OD} \quad k = 1, \ldots, 36 \tag{3.5}$$

ein 36-dimensionaler Merkmalsvektor, um in etwa den gewünschten Frequenzbereich abzubilden.
Perdigao und Sa (1998) schlagen 20 „Kanäle“ im Frequenzbereich zwischen 200 und 3400 Hz vor. Analog zu Gleichung 3.5 ergeben sich Merkmale:

$$X_k^{OD_2} = 1/8 \sum_{l=1}^{8} X_{10+k*8+l}^{OD} \quad k = 1, \ldots, 20. \tag{3.6}$$

Tchorz und Kollmeier (1999) verwenden 19 äquidistante Kanäle zwischen 300 und 4000 Hz, das entspricht sehr ähnlichen Merkmalen im Vergleich zu $X_k^{OD_2}$:

$$X_k^{OD_3} = 1/8 \sum_{l=1}^{8} X_{21+k*8+l}^{OD} \quad k = 1, \ldots, 29. \tag{3.7}$$

Allen (1994) beschreibt die Existenz von 20 unterschiedlichen Artikulationskanälen der menschlichen Sprachverarbeitung, die sich aus je etwa 114 benachbarten inneren Haarzellen zusammensetzen. Im menschlichen Ohr existieren ca. 3500 solcher IHCs, im Simulationsmodell ist diese Anzahl jedoch auf 251 reduziert. Dies motiviert Gruppen von acht benachbarten ANFs und bestätigt damit die durch $X_k^{OD_2}$ und $X_k^{OD_3}$ vorgeschlagene Zusammenfassung benachbarter ANFs.
Hemmert u. a. (2004) benutzen lediglich zwölf breitere, überlappende Kanäle. Diese liegen in einem Frequenzbereich zwischen etwa 50 und 15000 Hz (vgl. Holmberg und Hemmert, 2004). Als Merkmale entstehen:

$$X_k^{OD_4} = 1/36 \sum_{l=1}^{36} X_{-3+k*18+l}^{OD} \quad k = 1, \ldots, 12. \tag{3.8}$$

In einer Vorstudie von Szepannek und Weihs (2006a) zeigte eine auf Basis von parallelen lokalen Delay-Computing Netzwerken (vgl. Abschnitt 3.9.1) erzielte Ortskodierung gute Ergebnisse (vgl. Abschnitt 3.9.4). Diese ergibt sich durch

$$X_k^{OD_5} = 1/15 \sum_{l=1}^{15} X_{5+k*15+l}^{OD} \quad k = 1, \ldots, 12. \tag{3.9}$$

Die vorgestellten Merkmale $X_k^{OD_\cdot}$ repräsentieren allesamt Teilbereiche des Schallspektrums, es bietet sich somit die Möglichkeit der Anwendung einer anschließenden cepstralen Transformation (siehe Abschnitt 3.3) zur Erzeugung auditorisch basierter Cepstralkoeffizienten (AFCCs).

3.5 Generalisierte Synchronizitäts-Detektion (GSD)

Generalisierte Synchronizitäts-Detektion (GSD, Seneff, 1988, und in gerinfügig modifizierter Form Ali (1999); Ali u. a. (2002)), stellt eine von den Orts-Durchschnittsfeuerraten abweichende, spektrale Merkmalsrepräsentation für Sprachsignale dar, die ebenso auf auditorischer Signalverarbeitung beruht.

Hierbei sind insbesondere die zeitlichen Aspekte der neuronalen Aktivität zur Merkmalsbildung von Bedeutung. Das zugrunde liegende auditorische Modell unterscheidet sich von dem in der vorliegenden Arbeit verwendeten Modell jedoch durch eine lediglich phänomenologische Modellierung der auditorischen Verarbeitung des Schallsignals an Stelle der detaillierten neurophysiologischen Nachbildung des menschlichen Hörapparats, wie sie im Simulationsmodell aus Kapitel 2 beschrieben ist.

Die Prozessschritte des verwendeten phänomenologischen Modells umfassen eine Bandpassfilterung in verschiedene Kanäle (entsprechend unterschiedlichen Positionen entlang der Basilarmembran), einen nichtlinearen Kompressionsschritt (das Saturierungsverhalten an den Stereozilien imitierend), einen weiteren Prozessschritt zur Nachbildung der Kurzzeitadaption, sowie einen Tiefpassfilter (der eine Abnahme der Synchronizität der Feuerraten für hohe Frequenzen bewirkt). Eine abschließende *automatische Verstärkungsregelung (Automatic Gain Control, AGC)* fungiert als simples Modell zur Nachempfindung der neuronalen Refraktärzeit (für eine detaillierte Beschreibung, siehe Seneff, 1988; Ali, 1999).

Das Ergebnis ist eine Transformation der Originalzeitreihe der Schallwelle in unterschiedliche Kanäle und ist nach Seneff (1988) als zeitabhängige Feuerrate eines Post Stimulus Time Histogramms (PSTH [3] , siehe auch Anhang B.1) interpretierbar.

[3] Das *PSTH (Post Stimulus Time Histogramm)* $r(t)$ beschreibt die momentane Feuerrate als Funktion der Zeit. Miller und Sachs (1983) unterteilen hierfür zur Berechnung ein (rechteckiges) Zeitfenster von 20 ms in 256 gleichlange Zeitabschnitte (Bins). Als PSTH $\hat{r}(t)$ fungiert die gemittelte Anzahl an Spikes pro Bin über alle Stimuluswiederholungen. Als Stimulus wurden Vokale von 100 ms Dauer verwendet, die alle 250 ms wiederholt wurden, so lange bis im betrachteten Zeitfenster

Der GSD basiert auf einer Spektralzerlegung des zeitlichen Verlaufs der (durch Mittelwertbildung über mehrere Stimuluswiederholungen beobachteten) zeitabhängigen Feuerraten, wie sie von Miller und Sachs (1983) an Katzen für Vokal-Stimuli als *Average Localized Synchronized Rates, ALSR* untersucht worden ist. Dort wird mit einer diskreten Cosinus Transformation (DCT) eine Dekomposition des PSTHs $\widehat{r}(t)$, in Frequenzkomponenten durchgeführt, so dass

$$\widehat{r}(t) = R_0^l + \sum_{t=1}^{N/2-1} R_k^l \cos\left(\frac{2\pi kt}{N} + \phi_k^l\right), \quad t = 0, \ldots, N-1 \tag{3.10}$$

mit N als Bin-Anzahl (von 256) des PSTH während eines (20 ms) Zeitfensters und R_k^l als den Fourierfrequenzen zugehörigen Amplituden, wobei k den Index den Fourierfrequenz und l denjenigen der Nervenfaser bezeichnet. Man bemerke, dass R_0^l der vom Zeitverlauf unabhängigen *mittleren Feuerrate* entspricht, wie sie zur *Orts-Durchschnittsfeuerraten-Kodierung* verwendet wird.
Das ALSR Spektrum ist nun definiert durch (Miller und Sachs, 1983):

$$ALSR(k) = \frac{1}{M_k} \sum_{l \in c(k)} R_k^l \tag{3.11}$$

mit R_k^l wie in Gleichung 3.10 als Amplitude für die l^{te} ANF, $c(k)$ als derjenigen Menge von ANFs mit Mittenfrequenz von $\pm\frac{1}{4}$ Oktave um die k^{te} Fourierfrequenz und $M_k = |c(k)|$ und l dem ANF-Index.
Die GSD-Repräsentation ist nun, ähnlich dem von Miller und Sachs (1983) beobachteten ALSR, motiviert zur Detektion von Periodizitäten im Antwortverhalten der ANFs. Sie ist gegeben durch (Seneff, 1988):

$$X_k^{GSD} = A_s \tan^{-1}\left[\frac{1}{A_s}\left(\frac{\overline{(|X_k^{PSTH}(t) + X_k^{PSTH}(t-n_k)|)}_t - \delta}{\overline{(|X_k^{PSTH}(t) - \beta^{n_k} X_k^{PSTH}(t-n_k)|)}_t}\right)\right], \tag{3.12}$$

wobei $X_k^{PSTH}(t)$ den Output des auditorischen Simulationsmodells am Kanal k zur Zeit t bezeichnet. $n_k = f_s/CF(k)$ gibt die Anzahl von Zeitintegrationsschritten der Länge der Abtastperiode f_s pro Mittenfrequenz ($CF(k)$) der ANF an. A_s, β und δ sind

der ersten 20 ms mindestens 600 Spikes beobachtet wurden. Die Spikes wurden dabei mit einer Genauigkeit von $10\,\mu\,s$ gemessen. Eine solche Stimuluswiederholung entspricht natürlich nicht der Realität.

Konstanten, wobei $\beta := 0.99$ geringfügig kleiner als 1 gewählt wird um eine Null im Nenner zu verhindern. Der Parameter $\delta := 60 \cdot dt = 0.004$ ist klein zu wählen (knapp oberhalb der spontanen Rate der ANF) und dient zur Unterdrückung schwacher Signale. Für kleine Werte des hinteren Bruchs aus Gleichung 3.12 verläuft die GSD nahezu linear; die Steigung wird durch den Parameter $A_s(= 4)$ kontrolliert. Für sehr große Werte – entsprechend nahezu periodischer Antwort eines Hörnerven mit seiner Mittenfrequenz – saturiert die GSD. $\overline{(.)}_t$ symbolisiert Mittelwertbildung bezüglich der Zeit im aktuell betrachteten Fenster.

Die GSD ist leicht auf den Output des neurophysiologischen Simulationsmodells dieser Arbeit übertragbar, da sich eine PSTH-Schätzung der momentanen Feuerraten aus der Durchschnittsbildung mehrerer simulierter ANFs gleicher Mittenfrequenz gewinnen lässt. Die Feuerraten $X_k^{PSTH}(t)$ sollten jedoch einen möglichst glatten, nahezu stetigen Verlauf aufweisen. Diese Anforderung wird vom phänomenologischen Modell nach Seneff (1988) erfüllt. Um dies auch für geschätzte Feuerraten aus dem in dieser Arbeit verwendeten Modell zu gewährleisten, sollten $M >> 1$ wiederholte Simulationen durchgeführt werden.

Berücksichtigend, dass zur Berechnung des ALSR 600 Spikes in 20 ms verwendet werden, bedeutet das umgerechnet 30 000 Spikes je Sekunde. Ausgehend von einer maximalen durchschnittlichen Spikerate (R_0) von etwa 350 Spikes/s (Schoonhoven u. a., 1997) würde dies eine Anzahl von ca. 100 Repetitionen der Simulation bedeuten. Aufgrund der hohen Rechenzeit des Simulationsmodells werden lediglich 50 Wiederholungen pro ANF durchgeführt (MANF, vgl. Abschnitt 2.3). Seneff (1988) und Ali u. a. (2002) verwenden jedoch zur Merkmalsextraktion Kanäle im Abstand von 0.5 Bark. Die Abstände zwischen benachbarten Mittenfrequenzen des auditorischen Modells dieser Arbeit betragen je 0.1 Bark, so dass jeweils fünf benachbarte Kanäle per Mittelwertbildung zusammengefasst werden können, um die gleiche Bandbreitenresolution wie im Seneff-Modell zur erhalten. Es ergibt sich für die PSTHs als Input zur GSD-Berechnung:

$$X_k^{PSTH}(t) = 1/5 \sum_{l=1}^{5} X_{18+k*5+l}^{MANF}(t) \quad k = 1, \ldots, 36, \tag{3.13}$$

wobei $X_i^{MANF}(t)$ hier den Output von ANF i des auditorischen Simulationsmodells dieser Arbeit, gemittelt über 50 Repetitionen, bezeichnet.

Aus Gründen der Stabilität schlagen Ali u. a. (2002) *durchschnittliche lokalisierte Synchronizitäts Detektions-Merkmale* (englisch: *Average Localized Synchrony Detection*, ALSD) X_k^{ALSD} vor, mit

$$X_k^{ALSD}(t) = 1/3 \sum_{l=k-1}^{k+1} X_l^{GSD_k}(t), \tag{3.14}$$

wobei $X_l^{GSD_k}(t)$ hier das Merkmal $X_l^{GSD}(t)$ unter Verwendung der Periode n_k an Stelle von n_l bezeichnet.

3.6 Ensemble Intervall Histogramme (EIH)

Ein weiteres Modell einer von den Orts-Durchschnittsraten abweichenden Repräsentation frequenzbasierter Merkmale auf Basis auditorischer Modellierung stellen *Ensemble Intervall Histogramme* (EIH, Ghitza, 1988) dar. Die Art der Merkmalsextraktion weist dabei Parallelen auf zu Beobachtungen zur *dominanten Komponente* (vgl. Palmer u. a., 1986; Palmer, 1990; Holmes u. a., 2004). Die dominante Komponente ist definiert als diejenige Frequenz mit größter Amplitude im Spektrum des PSTH (siehe auch Anhang B.1 bzw. S. 30) innerhalb eines Bereichs von $\pm 1/4$ Oktave um die Mittenfrequenz einer bestimmten ANF (Delgutte, 1983):

$$DC_i = \arg\max_{\omega} A_i(\omega) I_{[2^{-1/4} * CF(i), 2^{1/4} * CF(i)]}(\omega) \tag{3.15}$$

wobei $A_i(\omega)$ das – beispielsweise durch eine Fast Fourier Transformation (FFT) gewonnene – Kurzzeitspektrum des PSTH von ANF i darstellt, $I(\cdot)$ die Indikatorfunktion bezeichnet und $CF(i)$ die zu ANF i gehörende Mittenfrequenz.
Delgutte und Kiang (1984) beobachteten an Katzen, dass sich die dominanten Komponenten der neuronalen Antwort der Hörnerven in fünf Bereiche von ANFs unterteilen lassen – abhängig von ihrer Mittenfrequenz. Diese Bereiche sind für Vokalstimuli durch

die Grundfrequenzen und deren Harmonischen sowie dessen ersten beiden *Formanten* bestimmt (vgl. S. 23).

Die Nachempfindung des menschlichen Hörapparats erfolgt von Ghitza (1988) zur Berechnung des EIH auf eine sehr simple und vereinfachende Weise in Form einer Bandpassfilterung in verschiedene Frequenzgruppen bzw. Kanäle, von denen jeweils diejenigen Zeitpunkte als Aktionspotenziale angesehen werden, an denen die gefilterte Originalwelle Schwellenwerte überschreitet.

Ghitzas Wahl der Filter beruht dabei auf Mittenfrequenzen der ANFs nach Greenwood (1990) und zugehörigen Filtern gemäß Goldstein (1990) (siehe Ghitza, 1994). Es entstehen unterschiedlich gefilterte Wellen für 85 verschiedene (äquidistante) Positionen entlang der Basilarmembran (Sandhu und Ghitza, 1995).

Die neuronale Aktivität (in Form von APs) wird nun modelliert durch diejenigen Zeitpunkte, an denen ein Kanal einen Schwellenwert (von unten nach oben) überschreitet (vgl. Abb 3.2). Für jeden Kanal existieren dabei fünf verschiedene solcher Schwellenwerte[4]. Solche verschiedenen Schwellenwerte pro Kanal ermöglichen eine diskretisierte Erfassung der Amplitudeninformation der Originalwelle und lassen sich dadurch motivieren, dass sich an jede reale innere Haarzelle mehrere ANFs anschließen sowie außerdem durch die Tatsache, dass Nervenfasern unterschiedlichen Typs existieren (vgl. Kapitel 2). Es ergeben sich $5 \cdot 85 = 425$ Folgen von Feuerzeitpunkten.

Die auditorische Modellierung erfolgt insgesamt jedoch auf eine sehr vereinfachte Art, so werden zum Beispiel das Phänomen der Adaption oder neuronale Refraktärzeiten (vgl. Kapitel 2) durch eine derartige Modellierung nicht berücksichtigt.

Aus der oben beschriebenen Menge von Feuerzeitpunkten wird das EIH nun wie folgt gebildet: Für jede Folge von AP-Zeitpunkten werden die jeweils letzten 20 Intervalle zwischen sukzessiven Feuerzeitpunkten berechnet (Ghitza, 1988). Deren Reziprokwert bezeichnet eine Frequenzschätzung einer wahrgenommenen Signalperiode.

Das EIH berechnet sich nun aus diesen beobachteten Frequenzschätzern als Histogramm mit einer Achseneinteilung in 128 gleichlange Abschnitte entlang der Frequenz-

[4]Diese werden randomisiert aus Normalverteilungen ermittelt mit Erwartungswerten, gleichverteilt entlang einer logarithmierten Amplitudenskala sowie Varianzen proportional zum Erwartungswert (siehe Ghitza, 1994). Diese Zufallskomponente in den Schwellenwerten soll die in der Natur leicht unterschiedlichen Größen von Zelldurchmesser und Synapsenverbindungen berücksichtigen.

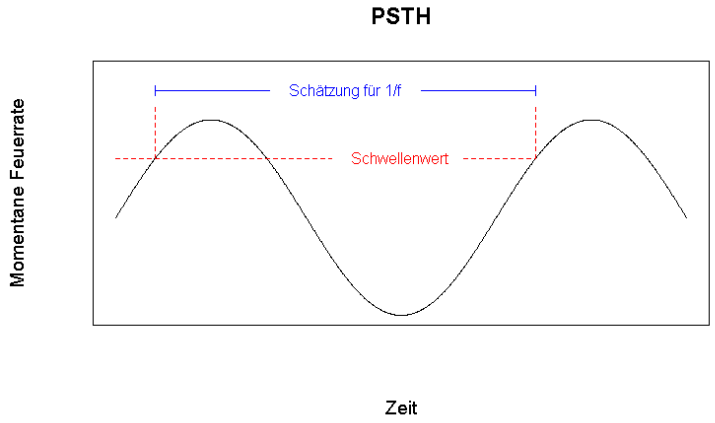


Abbildung 3.2: Beispiel für Frequenzschätzung aus Schwellenwertüberschreitungen des PSTH, wie sie zur EIH-Bildung genutzt wird.

achse (siehe Sandhu und Ghitza, 1995). Ghitza (1988) betont, dass die Betrachtung der jeweils letzten 20 Intervalle eine Fensterung mit nicht-äquidistanter Fensterbreite für die verschiedenen ANF-Kanäle bedeutet – vergleichbar mit einer Wavelet Analyse. Er beschreibt diese Art der Informationskodierung im Gegensatz zu der weiter oben beschriebenen Orts-Durchschnittsfeuerraten Kodierung als einen „rein temporalen Code“ (Ghitza, 1994).

Eine dem EIH ähnliche Merkmalsrepräsentation sind *Zero Crossings with Peak Amplitudes (ZCPA)* (Kim u. a., 1999), bei denen die Amplituden-Peaks der gefilterten Schallwellen in die EIH-Bildung mit einbezogen werden.

Ein Übertragen der den EIHs zugrunde liegenden Idee auf den Output des in Kapitel 2 beschriebenen Simulationsmodells ist naheliegend, da dieser gerade aus den Feuerzeitpunkten der auditiven Nervenfasern besteht. Das in dieser Arbeit verwendete Simulationsmodell empfindet zusätzlich die Phänomenologie der auditorischen Verarbeitungskette detailgetreuer nach, einschließlich der im ursprünglichen EIH nicht erfassten Adaption (vgl. die Abschnitte 2.3, bzw. A.5 und A.6). Andererseits besteht ein mögliches Manko der auf Basis des Modells dieser Arbeit erhaltenen spektralen EIH-Repräsentation in der dort fehlenden – und nicht direkt neurophysiologisch motivierbaren – Amplitudenerfassung.

Die so entstehenden Frequenzschätzungen $\hat{f} = \frac{1}{ISI}$ sind durch die neuronalen Refraktärzeiten tiefpassgefiltert mit einer cut off-Frequenz von 1.25 kHz. Die Inter-Spike Intervalle einzelner Nerven folgen häufig nur einem Vielfachen der Stimulusperiode. Diese Überlegung legt anstelle der Verwendung der direkten Aktionspotenzialemissionszeitpunkte die Bildung von Ensemble Intervall Histogrammen durch Schwellenwertüberschreitungen des aus wiederholten Simulationen gebildeten PSTHs, wie oben beschrieben nach Ghitza (1988) nahe. Diese sollten zur EIH Berechnung jedoch einen ausreichend glatten, nahezu stetigen Verlauf aufweisen (siehe auch Anhang B.1). In Experimenten mit PSTHs gebildet aus 50 wiederholten ANF-Simulationen bei Zusammenfassung von je fünf benachbarten ANFs zu einem PSTH und einer Bin-Länge 3/44100 s erwiesen sich die Frequenzschätzungen aus den beobachteten Perioden über zeitlichen Verlauf einer Sprachäußerung als unzureichend genau.
Secker-Walker und Searle (1990) beschreiben in ihren Untersuchungen zur periodischen Gestalt des PSTH in echten Messdaten von Katzen trotz bis zu 500 Stimuluswiederholungen die fehlende nötige Glattheit des PSTH. Auch Holmberg u. a. (2007) erwähnen unbefriedigende Ergebnisse bei auf Basis von weniger als 100 Wiederholungen gebildeten PSTHs.
Im Gegensatz zum EIH sind aufgrund der zeitlichen Mittelwertbildung in der GSD/ ALSD wesentlich stabilere Spektogramm-Schätzungen zu beobachten. Diese Beobachtung legt nahe, dass EIHs eher auf einer stetigen Zeitreihe, wie sie von Ghitza (1988) verwendet wurde, angemessen sind. Einzelne Aktionspotenziale stellen zwar einen effektiven neurophysiologischen Mechanismus dar, um Information zu übertragen, verlieren allerdings für einzelne ANFs an zeitlich präziser Information, die vermutlich durch die hohe Anzahl an in der Natur vorkommenden Hörnerven ausgeglichen wird. Die Auswertungen in Kapitel 6 beschränken sich bei der Untersuchung Phase Locking motivierter Merkmalssätze aus den beschriebenen Gründen auf Verwendung von Merkmalen des im vorigen Abschnitt beschriebenen Typs $X^{GSD/ALSD}$.

3.7 Kombination der bekannten Merkmalsextraktionsprinzipien

Alle bislang vorgestellten Merkmale lassen sich interpretieren im Sinne von Spektren und versuchen nachzuempfinden, was nach dem Prozess der auditorischen Verarbeitung vom ursprünglichen Schallsignal erhalten bleibt. Die Art der Merkmalsextraktion erfolgt jedoch nach unterschiedlichen Prinzipien: basierend auf dem Ort der Anregung entlang der Cochlea einerseits und den in der neuronalen Antwort enthaltenen Periodizitäten (durch Phase Locking entstehenden Inter-Spike Intervallen) andererseits.
Es kann versucht werden, die in den verschiedenen Merkmalen enthaltene Information zu kombinieren. Eine einfache Kombination kann in einer Konvexkombination

$$X_{\omega}^{OD,ISI}(t,\rho) = \rho X_{\omega}^{OD}(t) + (1-\rho)X_{\omega}^{ISI}(t) \tag{3.16}$$

mit $\rho \in [0,1]$ erfolgen. Der Subindex ω anstelle von k soll hierbei symbolisieren, dass die entsprechenden Merkmale beider Merkmalssätze die gleichen Frequenzbereiche (durch gleiche zusammengefasste Mittenfrequenzen) beschreiben. Dabei ist zu berücksichtigen, dass beide (Orts- bzw. Phase Locking-basierte) Merkmale vergleichbar skaliert sein sollten und von daher hierfür in dieser Arbeit je auf eine Standardabweichung von 1 normiert werden.
Durch die neuronale Refraktärzeit ist davon auszugehen, dass die ISI-basierten Merkmale schlecht geeignet sind zur Erfassung hoher Frequenzen des Signals (siehe z.B. Johnson, 1980) während Orts-Durchschnittsfeuerraten Merkmale $X_k^{OD\cdot}$ insbesondere niedrige Frequenzen nicht sehr präzise identifizieren können (vgl. z.B. Abb. 2.8 am Beispiel zweier Sinusschwingungen von 440 und 880 Hz). Verwendet man eine „Grenzfrequenz“ wie z.B. $\omega_0 = 1500Hz$ (anlehnend an die Beobachtungen abfallenden Phase Lockings für höhere Frequenzen nach Johnson, 1980), so lassen sich aus Merkmalsvektoren X_{ω}^{ISI} und X_{ω}^{OD} sehr einfache kombinierte Merkmale der Gestalt:

$$X_{\omega}^{OD,ISI}(t,\omega_0) = I_{[\omega_0,\infty]}(\omega)X_{\omega}^{OD}(t) + I_{[0,\omega_0]}(\omega)X_{\omega}^{ISI}(t) \tag{3.17}$$

gewinnen, wobei ω im Merkmalsindex diejenige mit dem Merkmal assoziierte Frequenz, im Fall von Orts-Durchschnittsfeuerraten die Mittenfrequenz, bezeichnet und $I_{[.]}(.)$ die Indikatorfunktion ist.

Zwischen $X_\omega^{OD,ISI}(t,\rho)$ und $X_\omega^{OD,ISI}(t,\omega_0)$ lässt sich eine Beziehung aufstellen, wenn man $\rho = \rho(\omega)$ merkmalsabhängig wählt.

3.8 Motivation der Merkmalsextraktionsansätze anhand eines Punktprozessmodells

Punktprozessmodell

Eine andere Perspektive stellt die Auffassung der Spike-Zeitpunkte einer ANF als Punktprozess dar. Betrachtet sei im Folgenden zunächst nur je eine ANF mit Index i. Seien $\{t_j,\ j = 1,\ldots,N_i\}$ die Zeitpunkte, an denen ANF i ein Aktionspotenzial emittiert, d.h. es gilt:

$$\{t_j : X_i(t_j) = 1\} \tag{3.18}$$

mit $X_i(t_j)$ als Zustand von ANF i zum Zeitpunkt t_j wie oben (siehe z.B. Abschnitt 2.4). Die bedingte Wahrscheinlichkeit der ANF i, Aktionspotenziale zu den Zeitpunkten $\{t_j\}$ zu generieren, gegeben einen zeitabhängigen Stimulus $s(t)$, wird mit $P(\{t_j\}|s(t))$ bezeichnet und der Einfachheit halber im Folgenden durch $P(\{t_j\})$ abgekürzt.
Während eines kurzen Zeitfensters $[t, t+\Delta]$ betrage die Wahrscheinlichkeit einer Spike-Emission $p([t, t+\Delta])$. Die *zeitabhängige Feuerrate* $r(t)$ sei nun definiert durch ihre asymptotische Proportionalität zur Spike-Emissionswahrscheinlichkeit für sehr kurze Zeitfenster:

$$p([t, t+\Delta]) \underset{\Delta\to 0}{=} r(t)\Delta. \tag{3.19}$$

Das einfachste Modell für $r(t)$ ist dessen Auffassung als *Poissonprozess*. Die erforderliche, vereinfachend hiermit implizit getroffene Annahme **(A1)** lautet, dass die Spike Rate $r(t)$ vollständig bedingt ist durch den zugrunde liegenden Stimulus $s(t)$ und nicht abhängt von Zeitpunkten vorangegangener Aktionspotenziale. Diese Annahme ist stark vereinfachend und entspricht nicht der Realität, da z.B. die Refraktärzeit nach vorangegangenen APs das Verhalten der ANF und damit die Rate $r(t)$ beeinflussen, sie stellt jedoch ein gutes Modell dar, wenn kurze Zeitabschnitte ($\leq 200\,ms$) betrachtet werden (Rieke u. a., 1997, S. 52 f). Dies ist für die als stationär angesehenen Zeitfenster zur Spracherkennung gegeben. Bei einer konstanten Feuerrate $r(t)$ würde sich für die Länge der Inter-Spike Intervalle eine Exponentialverteilung ergeben.

Es lässt sich nun die Likelihood-Funktion für $r(t)$ bei beobachteten Feuer-Zeitpunkten $\{t_j\}$ herleiten: Betrachtet man ein durch Fensterung gegebenes Zeitintervall $[0, T]$, unterteilt in kurze Abschnitte (Bins) der Dauer Δ, so ist die Wahrscheinlichkeit einer Sequenz von beobachteten Aktionspotenzialen zu den Zeitpunkten $t_1, \ldots, t_N$ gegeben durch

$$\begin{aligned} P(\{t_j\}) &= P(\text{aufgetretene Spikes in entsprechenden Bins}) \\ &\times \; P(\text{keine Spikes in übrigen Bins}). \end{aligned} \tag{3.20}$$

Die Wahrscheinlichkeit des Auftretens eines Spikes in einer Bin um den Zeitpunkt t ist gegeben durch $p(t) = r(t)\Delta$ (s.o.). Gemäß (A1) können die $\{t_j\}$ als unabhängig angenommen werden und das Produkt aller Auftretenswahrscheinlichkeiten aller einzelnen Spikes in den Bins um die t_j gebildet werden. Das resultierende Produkt ist proportional zu $(\Delta)^N$:

$$P(\text{aufgetretene Spikes in entsprechenden Bins}) = 1/N!(\Delta)^N \prod_{j=1}^{N} r(t_j). \tag{3.21}$$

Der Term $1/N!$ korrigiert um die Anzahl möglicher Permutationen.
Auf die gleiche Weise erhält man

$$P(\text{keine Spikes in übrigen Bins}) = \prod_{n \neq j} (1 - r(t_n)\Delta)\,, \tag{3.22}$$

wobei mit dem Ausdruck „$n \neq j$“ hier alle anderen Bins in $[0, T]$ gemeint sind, in denen kein Spike emittiert wird.
Aus beiden Gleichungen zusammen ergibt sich entsprechend Gleichung 3.20

$$P(\{t_j\})\Delta^N = 1/N! \prod_{n} (1 - r(t_n)\Delta) \prod_{j=1}^{N} \left(\frac{r(t_j)\Delta}{1 - r(t_j)\Delta} \right), \tag{3.23}$$

wobei $\prod_n$ hier das Produkt über alle Bins bezeichnet. Das erste Produkt kann umgeformt werden zu

$$\begin{aligned} \prod_{n} (1 - r(t_n)\Delta) &= \prod_{n} \exp\big(\log\big[1 - (r(t_n))\Delta\big]\big) \\ &= \exp\big(\sum_{n} \log\big[1 - (r(t_n))\Delta\big]\big). \end{aligned} \tag{3.24}$$

Für sehr kleine Bin-Größen Δ ergibt sich der Logarithmus von Werten nahe eins. Es gilt $\log(1) = 0$, und man kann die Taylor-Reihe des Logarithmus im Punkt 1 entwickeln; es ergibt sich

$$\log(1 - r(t_n)\Delta) = -r(t_n)\Delta - 1/2(r(t_n)\Delta)^2 + \dots \tag{3.25}$$

Die Likelihood lässt sich damit ausdrücken als

$$\begin{aligned} P(\{t_j\})(\Delta)^N &= \frac{1}{N!}\prod_{j=1}^{N}\frac{r(t_j)\Delta}{1-r(t_j)\Delta} \\ &\quad \exp\left(\sum_n \big(-r(t_n)\Delta\big) - \frac{1}{2}\sum_n \big(-r(t_n)\Delta\big)^2 + \dots\right) \\ &= \frac{1}{N!}\Delta^N \prod_{j=1}^{N}\frac{r(t_j)}{1-r(t_j)\Delta} \\ &\quad \exp\left(\sum_n \big(-r(t_n)\Delta\big) - \frac{1}{2}\sum_n \big(-r(t_n)\Delta\big)^2 + \dots\right). \end{aligned} \tag{3.26}$$

Weiterhin folgt aus der Definition des Riemann-Integrals für $\Delta \to 0$:

$$\lim_{\Delta\to 0}\sum_n r(t_n)\Delta = \int r(t)dt \tag{3.27}$$

Alle Summanden mit Termen höherer Ordnung von Δ verschwinden für $\Delta \to 0$.
Da zudem $(1 - r(t_i)\Delta) \to 1$ für $\Delta \to 0$, ergibt sich schließlich die Likelihood-Funktion

$$P(\{t_j\}) = 1/N! \prod_{j=1}^{N} r(t_j)\, \exp\Big(-\int_0^T r(t)dt\Big) \tag{3.28}$$

Parametrisierung

Für die beobachtete Antwort der Hörnerven im Speziellen wird folgende zusätzliche Annahme getroffen:

Annahme (A2):
Für zugrunde liegenden Schallsignale ist der Stimulus einer ANF als bandpass-gefilterte Schallwelle von sinusförmiger Natur. Die zeitabhängige Feuerrate $r(t)$ lässt sich nach Rieke u. a. (1997), S. 30, durch die Gleichung

$$r(t) = r_0 + A\sin(\omega t + \phi) \tag{3.29}$$

beschreiben. Für die kurzen, als stationär angenommen Zeitfenster ist der den beobachteten Spikes zugrunde liegende Prozess an einer ANF k damit vollständig charakterisiert durch das Parametertupel $\{r_{0,k}, \omega_k, A_k, \phi_k\}$, und die weiter oben berechnete Likelihood-Funktion liefert eine Grundlage zu deren identifikation auf Basis der beobachteten Spikes mit Hilfe des Maximum-Likelihood Ansatzes. Drückt man $r(t)$ entsprechend (A2) durch Gleichung 3.29 aus, so ergibt sich für die logarithmierte Likelihood-Funktion:

$$\begin{aligned} L(r_0, A, \omega, \phi, \{t_j\}) &= \sum_{j=1}^{N} \log\big(r_0 + A\sin(\omega t_j + \phi)\big) \\ &\quad - \int_0^T \Big(r_0 + A\sin(\omega t + \phi)\Big)dt - \log N! \\ &= \sum_{j=1}^{N} \log\big(r_0 + A\sin(\omega t_j + \phi)\big) \\ &\quad -r_0 T + \frac{A}{\omega}\Big(\cos(\phi) - \cos(\omega T + \phi)\Big) - \log N!. \end{aligned} \tag{3.30}$$

Nach Wolpert (2007) könnte dies über die gemeinsame Likelihood mehrerer benachbart positionierter Hörnerven erfolgen, denen eine ähnlich gefilterte Originalwelle zugrunde liegt. Die gemeinsame Log-Likelihood-Funktion ergibt sich dabei aufgrund von Unabhängigkeit als Produkt der individuellen $L_i(r_0, A, \omega, \phi, \{t_j\}_i)$ (mit i als ANF-Index). Eine geschlossene Lösung dieses Optimierungsproblems ist leider nicht möglich. Ebenso gestaltet sich eine iterative Optimierung bezüglich der vier Parameter als schwierig. Ligges (2006), S. 65 f, illustriert die Problematik einer gleichzeitigen Parameteroptimierung von Frequenz und Phase im Zeitbereich, bei der sich das Problem einer Vielzahl lokaler Optima stellt.

Zur Interpretation der Parameter lässt sich jedoch sagen, dass ω_i mit der *dominanten Komponente* (siehe Abschnitt 3.6) der ANF i vergleichbar ist. A_i liefert eine zusätzliche Amplitudeninformation. Diese beiden Parameter finden somit Entsprechung in den Inter-Spike Intervall basierten Merkmalen *EIH* und *GSD*.

Die $r_{0,i}$ repräsentieren zeitinvariante Feuerraten und sind von daher mit den beschriebenen *Orts-Durchschnittsfeuerraten* zu vergleichen.

Die Parameter ϕ_i schließlich beschreiben die Phasenverschiebung von $r_i(t)$ an der Po-

sition i der Basilarmembran. Diese wird durch die bisher vorgestellten Merkmale nicht erfasst. Die Entwicklung von Merkmalen, die auch derartige Information beinhalten, stellt den Inhalt der folgenden Abschnitte dar.

3.9 Delay-Computing basierte Ansätze der Merkmalsextraktion

3.9.1 Delay-Computing Netzwerke (DCN)

Prinzip

Künstliche *Delay-Computing Netzwerke* (DCN) wurden entwickelt von Brückmann u. a. (2004) zur Detektion von Verzögerungsstrukturen in Aktionspotenzialzeitpunkten mehrerer Neuronen. Diese repräsentieren charakteristische Muster für einen Klang (vgl. Moore, 2003).

DCNs bestehen aus drei unterschiedlichen Einheiten: *Input-Neuronen*, *Output-Neuronen* und *Delay-Neuronen*. Bezeichne $X_i(t),\ i = 1, \ldots, N$, den Wert von Input-Neuron i zum Zeitpunkt t, wobei t diskrete Werte annimmt und somit für einen Zeitintegrationsschritt steht. Es gilt: $X_i(t) = 1$, wenn das Neuron zum „Zeitpunkt“ t ein AP feuert und null sonst. Im vorliegenden Fall dieser Arbeit stellen die ANFs die Input-Neuronen für ein anschließendes Delay-Computing Netzwerk dar. Abbildung 3.3 (links, S. 44) stellt den Aufbau eines DCNs für das Beispiel von $N = 9$ Input-Neuronen dar. Die Input-Neuronen sind in der Vertikalen am linken Rand dargestellt.

In jedem Zeitintegrationsschritt $t \rightarrow t + \Delta t$ werden die Werte aus Layer (bzw. Schicht) j (Vertikale in Abbildung 3.3) entweder in den benachbarten nächsten Layer $j + 1$ überführt oder in den *Delay-Neuronen* D_{ij} genau einen Zeitintegrationsschritt Δt lang verzögert. Letzteres erfolgt mit einer Wahrscheinlichkeit $P_{ij}(t)$.

Entsprechend erreicht ein Spike (d.h. eine Eins) aus Input-Neuron i, das sich zum

Zeitpunkt t im Layer j befindet, mit einer Wahrscheinlichkeit von $1 - P_{ij}(t)$ direkt den benachbarten Layer $j + 1$.
Die *Output-Neuronen* befinden sich horizontal angeordnet am unteren Rand des Netzwerkes in Abbildung 3.3. Die Werte der Output-Neuronen $Y_j(t),\ j = i, \ldots, N$, berechnen sich durch Summation aller Einsen, die sich zum Zeitpunkt t im vertikalen *Layer* j oberhalb des Output-Neurons j befinden.
Die $P_{ij}(t)$ werden im Laufe der Zeit selbstlernend trainiert: Wenn ein Output-Neuron j zum Zeitpunkt t feuert, werden die Gewichte $w_{ij,direct}$ und $w_{ij,delay}$ (siehe Abbildung 3.3) der zugehörigen ausgewählten (bzw. nicht-ausgewählten) „Pfade“ in Layer j folgendermaßen trainiert:

$$\begin{aligned} w_{ij,selected}(t + \Delta t) &= w_{ij,selected}(t) + (1 - w_{ij,selected}(t) \cdot \alpha) \\ w_{ij,deselected}(t + \Delta t) &= w_{ij,deselected}(t) - w_{ij,deselected}(t) \cdot \alpha. \end{aligned} \tag{3.31}$$

α stellt dabei eine *Lernrate* dar und die Bezeichnungen *selected* bzw. *deselected* bezeichnen situational den direkten bzw. verzögerten Pfad. Ein Output-Neuron $Y_j(t)$ feuert genau dann, wenn sein Wert einen Schwellenwert überschreitet. Nach Brückmann u. a. (2004) geschieht dies, wenn alle ausgewählten Pfade aktiviert sind oder aber $Y_j(t)$ einen Schwellenwert überschreitet. Die neuen Verzögerungs-Wahrscheinlichkeiten $P_{ij}(t)$ ergeben sich durch

$$P_{ij}(t + \Delta t) = \frac{e^{w_{ij,delay}/T_{Boltzmann}}}{e^{w_{ij,delay}/T_{Boltzmann}} + e^{w_{ij,direct}/T_{Boltzmann}}}, \tag{3.32}$$

wobei $T_{Boltzmann}$ eine *Boltzmann-Temperatur* darstellt. Diese startet mit einem Wert von T_{max} und sinkt sukzessive Layer-spezifisch mit jedem Feuern des zugehörigen Output-Neurons j um einen konstanten Wert δT bis ein Minimalwert von T_{min} erreicht worden ist.
Der Lernprozesses schreitet voran, bis die Gewichte w_{ij} gegen Werte null oder eins konvergieren und letztendlich jedes Output-Neuron genau eine zugehörige Delaystruktur repräsentiert. Brückmann u. a. (2004) zeigen Konvergenz des Netzwerkes für Verzögerungsstrukturen in Gestalt von Geraden oder Sinuskurven.
Die Werte der Input-Neuronen $X_i(t)$ können dabei prinzipiell auch gebrochen rationale Zahlen annehmen, wie es im Fall von ANFs als Input-Neuronen und einer Durchschnittsbildung über mehrfache Simulation der Fall ist.

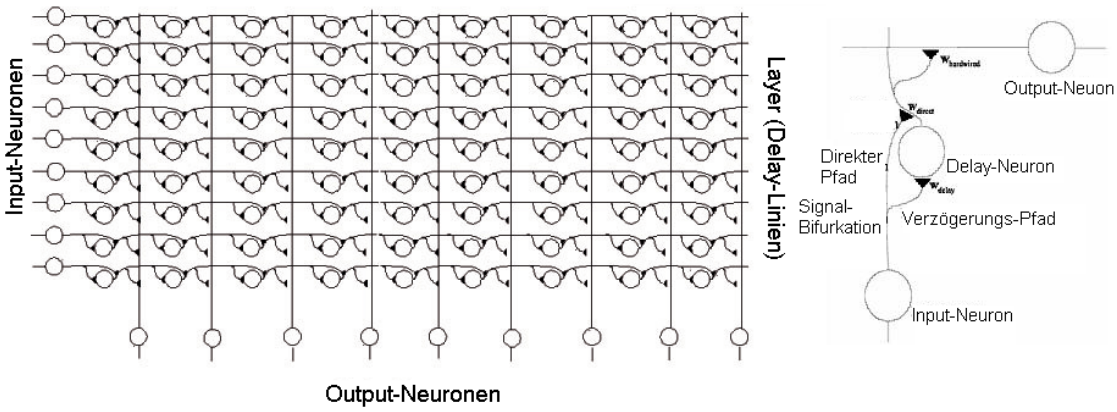


Abbildung 3.3: Links: Exemplarischer Aufbau eines 9x9 DCNs. Rechts: Schematische Darstellung einer Synapse (um 90 Grad gedreht). Abbildungen entnommen aus Brückmann u. a. (2004).

Anpassung an das auditorische Simulationsmodell

In einer Arbeit von Harczos u. a. (2006a) wurde ein solches Delay-Computing Netzwerk auf die Antwort der simulierten Hörnerven $X_i(t)$ als Netzwerk-Input-Neuronen angepasst, wobei i hier den Index der Position der Input-ANF auf der Basilarmembran bezeichnet. Aufgrund des eingeschränkten spektralen Bereichs, in dem sich Sprache bewegt, sind nicht alle simulierten Hörnerven zur Weiterverarbeitung erforderlich. Zudem verläuft die Antwort für benachbarte ANFs sehr hoher Frequenzen nicht mehr synchron und kann damit vermutlich nicht mehr zur Trajektorienidentifikation des Netzwerkes beitragen. Beginnend bei ANF Nr.26 (entsprechend einer CF von 219 Hz) wird eine Netzwerkgröße von 141 $\times$ 141 Neuronen vorgeschlagen, wobei genau jede ANF als ein Input-Neuron fungiert. Als Zeitintegrationsschritt wird eine Schrittgröße von $\Delta t = \frac{6 \text{ bis } 8}{44100}$ s – entsprechend einer Abtastfrequenz zwischen 5512.5 und 7350 Hz – vorgeschlagen, da so durch das DCN der Verlauf einer vollständigen Trakjektorie bis hin zum apikalen Ende der Cochlea erfasst werden kann (Harczos u. a., 2006a).

Das selbstlernende Trainieren der Kurven ist ein sehr zeitaufwändiger Prozess. Aus diesem Grund schlagen Harczos u. a. (2006a) einen schnellen und effektiven Algorithmus zum Lernen der mit den Output-Neuronen assoziierten Kurven vor, der die Gewichte entsprechend trainiert. Fasst man die Ort-Zeit-Ebene der Hörnervenaktivität

als zweidimensionales Gitter auf, wobei ein Pixel in x-Richtung bestimmt wird durch einen Zeitintegrationsschritt und in y-Richtung durch die Positionsveränderung von einer ANF entlang der Basilarmembran, so lässt sich die durch das DCN trainierte Kurvenschar darstellen durch:

$$\{(f_j(y), y) : y = 1, \ldots, N\},\ j = 1, \ldots, N \tag{3.33}$$

wobei

$$f_j(y) = \frac{j f_{min}(N - 1 - y)}{(y + f_{min})(N - 1)}. \tag{3.34}$$

j bezeichnet dabei den Index des Output-Neurons und f_{min} ist ein Parameter, der die „durchschnittliche Neigung“ der Kurven beschreibt: kleinere Werte von f_{min} bedeuten dabei eine höhere Konvexität der durch das Netzwerk berschriebenen Kurven. Harczos u. a. (2006a) schlagen für das beschriebene DCN einen Wert von $f_{min} = 30$ bis 40 als guten Kompromiss für sowohl männliche als auch weibliche Sprecher vor. Abbildung 3.4 illustriert den Effekt von f_{min} auf die durch das Netzwerk repräsentierten Trajektorien. Die Antwort des DC Netzwerkes ist in Abbildung 3.5 an einem künstli-

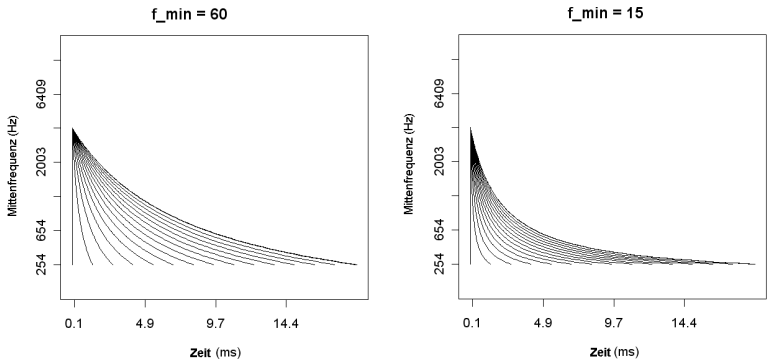


Abbildung 3.4: Durch das Netzwerk repräsentierte Trajektorien für unterschiedliche Werte $f_{min} = 60$ (links) und $f_{min} = 15$ (rechts). (Kurven nur dargestellt für j mit $mod(j, 10) = 1$).

chen Beispiel illustriert (vgl. Harczos u. a., 2006a): Kurven unterschiedlicher Neigung sind in weißes Rauschen gelegt. Die Gestalt des DCN Outputs manifestiert sich in der Aktivität der Output-Neuronen $Y_j(t)$ (entlang der Ordinate) zum Zeitpunkt t

(entlang der Abszisse), wobei jedes j eine spezifische Krümmung der Wanderwellen-Trajektorie repräsentiert. Die verschiedenen Trajektorien lassen sich leicht als Reaktion des Netzwerkes an unterschiedlichen Positionen j (Output-Neuronen) ausmachen. Im hier vorhandenen Idealfall liegen die Antworten aller Input-Neuronen auf derselben Trajektorie. Abbildung 3.6 zeigt die Reaktion des Netzwerkes auf Sinustöne

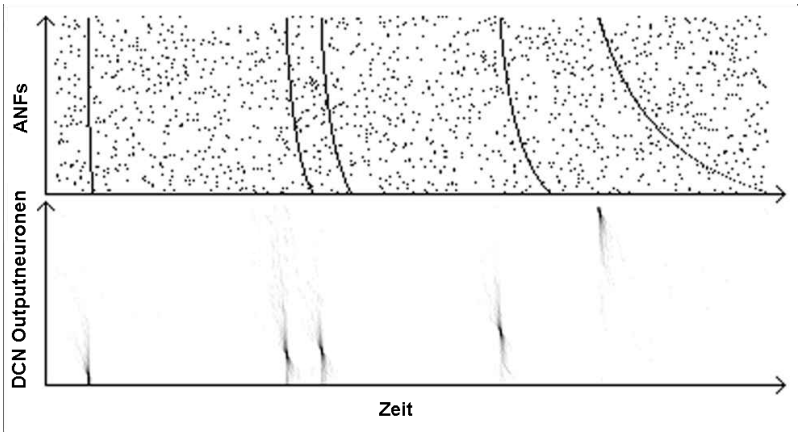


Abbildung 3.5: Beispiel des DCN Outputs: Identifikation verschiedener Kurven in weißem Rauschen. Oben: Auditory Image, exemplarische Reaktion des Hörnerven, unten: DCN Output (aus Harczos u. a., 2006a).

unterschiedlicher Frequenzen. Es ist erkennbar, dass je nach zugrunde liegender Signalfrequenz unterschiedliche Trajektorienverläufe der Wanderwellen vorliegen. Diese Beobachtung deckt sich mit den Aussagen von Greenberg u. a. (1997), wie sie in Abschnitt 3.2 beschrieben sind. Es ist aber auch zu sehen, dass die angenommenen Maximalwerte der Output-Neuronen (über die Zeit, hier berechnet auf dem durchschnittlichen Output von 50 HSR Hörnerven) Werte annehmen, die um ein Vielfaches unterhalb der Maximal-Reaktion von $Y_j(t) = 141$ für gleichzeitige Antwort aller Input-Neuronen liegt. Abbildung 3.7 zeigt beispielhaft die unterschiedliche Antwort des Delay-Computing Netzwerks für vier verschiedene Vokale. Insbesondere lässt sich eine feine, periodische Zeitstruktur des DCN-Outputs ausmachen. Die Output-Neuronen des DCNs fassen Aktivität derjenigen ANFs zusammen, die auf derselben Delaytrajektorie ihre Aktionspotenziale freisetzen. Diese Eigenschaft wird in Szepannek u. a. (2006) und Harczos u. a. (2006b) zur Merkmalsextraktion ausgenutzt. Auf

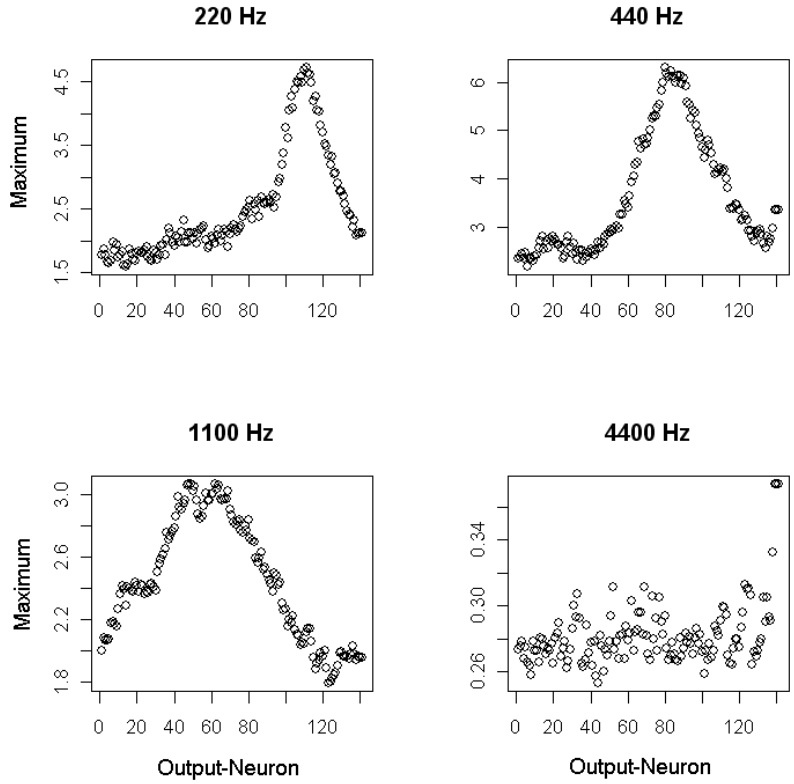


Abbildung 3.6: Maximale Antwort der DCN Output-Neuronen über die Zeit auf Sinustöne unterschiedlicher Frequenz.

den Aspekt der Merkmalsextraktion wird in Abschnitt 3.9.3 weiter eingegangen.

3.9.2 Parallele lokale Delay-Computing Netzwerke (PLDCN)

Ein Engpass in der unmittelbaren Anwendung der oben beschriebenen Delay-Computing Netzwerke auf den Output des auditorischen Simulationsmodells begründet sich in der Beobachtung von Greenberg u. a. (1997), dass die Amplitude der Wanderwelle stark abnimmt, nachdem sie an der Basilarmembran den Ort maximaler Anregung erreicht hat. Die Trajektorie der Welle ist nicht über den vollen Bereich von ANFs bis zum apikalen Ende der Cochlea zu beobachten. Mehr noch, durch die Bandpassfilterung des Signals entlang der Basilarmembran, wie sie in Kapitel 2 beschrieben wird, erzeugen Wellen einer spezifischen Frequenz eine Reaktion lediglich an einer (zusammenhängen-

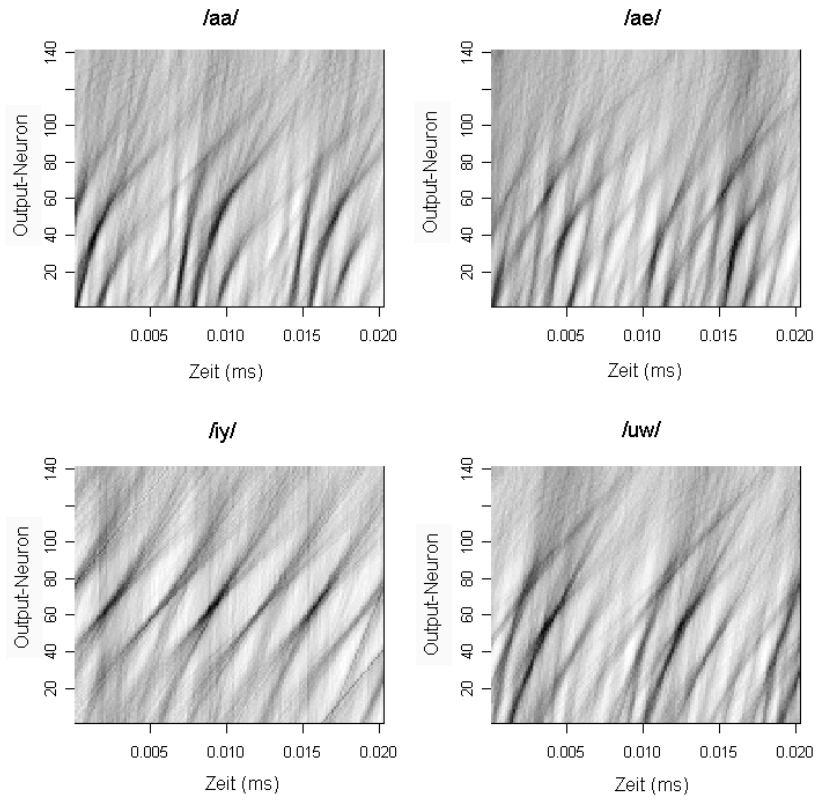


Abbildung 3.7: Exemplarischer Output des DCNs auf 20 ms Zeitsegmenten in der Mitte von vier unterschiedlichen Vokalen.

den) Teilmenge der Hörnerven. Abbildung 3.8 veranschaulicht dies für Sinustöne dreier verschiedener Frequenzen. Es ist erkennbar, dass auf jede der drei Stimulusfrequenzen nur ein eingeschränkter, zusammenhängender Bereich von ANFs eine verstärkte, bzw. synchrone Reaktion aufweist. Dieser umfasst bis zu etwa 50 benachbarte ANFs (Szepannek und Weihs, 2006a). Diese Beobachtungen motivieren die Implementierung von mehreren *parallelen lokalen Delay-Computing Netzwerken (PLDCNs)* auf jeweils eingeschränkten, entlang der Basilarmembran zusammenhängenden Bereichen benachbarter ANFs (Szepannek und Weihs, 2006a) für die gilt:
Wenn $i_1 < i_2$ *Indizes von Input-Neuronen des lokalen Netzwerks* m *bezeichnen und* $i_1 < i < i_2$*, so bezeichnet auch* $i \in \mathbb{N}$ *den Index eines Input-Neurons dieses Netzwerks.*
Sei M die Anzahl solcher lokaler Netzwerke und $N_m, m = 1, \ldots, M$, deren Anzahl an

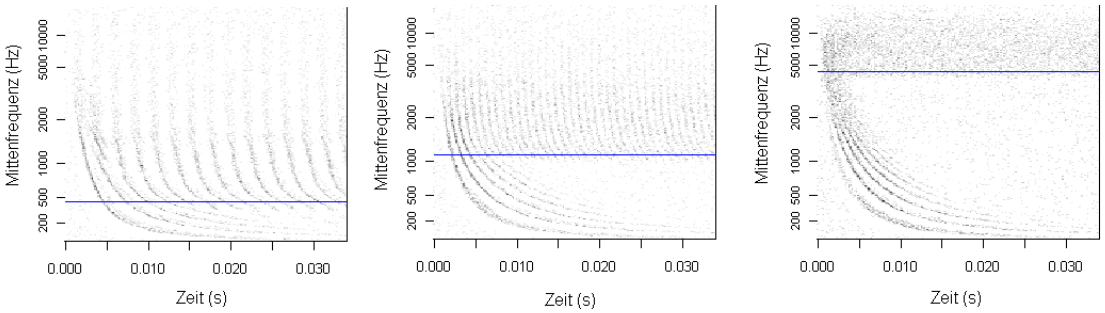


Abbildung 3.8: Simulierte Reaktion der Hörnerven auf Sinustöne unterschiedlicher Frequenz: 440 Hz (links), 1100 Hz (Mitte) und 4400 Hz (rechts). Die horizontale blaue Linie markiert die jeweilige Position mit entsprechender Mittenfrequenz auf der Basilarmembran.

Input- und Output-Neuronen, hier identisch für alle m gewählt. Dann ergeben sich die Input-Neuronen des lokalen DCNs m als $X_{a+(m-1)N_m+1}$ bis $X_{a+(m-1)N_m+N_m}$, wobei $a+1$ der Index der ANF mit niedrigster Mittenfrequenz des ersten lokalen Netzwerkes $m = 1$ bezeichnet. Die Input-Neuronen der Netzwerke sind also disjunkt und insbesondere ausschöpfend über den Bereich von $\sum_m N_m$ Hörnerven. Die Eigenschaft der Disjunktheit kann dabei prinzipiell auch zugunsten von überlappenden lokalen DCNs bezüglich der Input-Neuronen aufgeweicht werden.
Über die Wahl von N_m lässt sich Folgendes aussagen: Je kleiner die Größe der lokalen DCNs gewählt ist, desto frequenzspezifischer ist die Information, die das Netzwerk verarbeitet, da durch die CFs der Input-Neuronen diejenigen Frequenzbereiche im Schallsignal definiert werden, die zur Aktivierung der Netzwerke führen. Ein größeres Netzwerk mit bis zu 50 Input-ANFs dagegen, führt zu einer stabileren Identifikation der Trajektorienverläufe.

Zu f_{min} lässt sich die Überlegung anstellen, dass die lokalen Netzwerke nicht mehr den gesamten Trajektorienverlauf beschreiben, wie dies für die ursprünglichen Delay-Computing Netzwerke der Fall ist. Der vollständige Trajektorienverlauf ergibt sich vielmehr zusammengesetzt aus den einzelnen Kurven der lokalen Netzwerke. Im Bereich der „oberen“ Input-Neuronen ergibt sich damit nicht mehr notwendig für al-

le möglichen Kurven der charakteristische, fast senkrechte Kurvenverlauf (siehe z.B. Abb. 3.8). Dies eröffnet die Möglichkeit einer höheren Wahl des Parameters f_{min} (vergleiche Abb. 3.4). Im Rahmen dieser Arbeit werden neben dem in Abschnitt 3.9.1 beschriebenen DCN mit 144 Input-Neuronen (Harczos u. a., 2006a), sowohl „große“ PLDCNs mit $N_m = 35$ ($f_{min} = 15$) als auch „kleinere“ mit $N_m = 18$ ($f_{min} = 5$ bzw. 15) sowie zudem frequenzspezifischere PLDCNs der Größe $N_m = 8$ ($f_{min} = 4$ bzw. 7) implementiert (mit jeweils $a = 20$, siehe auch Szepannek und Weihs, 2006a).
Abbildung 3.9 zeigt exemplarisch die Antwort von PLDCNs der Größe $N_m = 18$ für die vier Vokale aus Abbildung 3.7. Es lassen sich bereits per Auge Unterschiede zwischen den Outputs für die verschiedenen Vokale ausmachen. Die gewonnenen Repräsentationen weisen eine starke periodische Struktur auf. Insbesondere scheinen unterschiedliche Netzwerke von den verschiedenen Vokalen unterschiedlich stark aktiviert zu werden. Es zeichnen sich aber auch Unterschiede zwischen den verschiedenen, von den lokalen Netzwerken repräsentierten Trajektorien ab (vgl. Abbildung 3.10). Werner (2008) beschreibt Möglichkeiten zur Nutzung von PLDCNs zur Lokalisation von Schallquellen. Die Generierung von Merkmalen zur automatischen Spracherkennung aus dem Output von (PL)DCNs wird im folgenden Abschnitt 3.9.3 behandelt und anschließend, in Abschnitt 3.9.4, werden erste vergleichende Ergebnisse zur Eignung unterschiedlicher aus DCNs und PLDCNs gewonnener Merkmale zur automatischen Vokalklassifikation präsentiert.

3.9.3 Orts-Trajektorienneigungs-Merkmale

Die in diesem Abschnitt vorgestellten Merkmale basieren auf dem Output der (parallelen lokalen) Delay-Computing Netzwerke und nutzen die Gestalt der Wanderwellen-Delaytrajektorie als Informationsträger (s.o.).
Zunächst sei festgestellt, dass jedes Output-Neuron eines (PL)DCNs von sämtlichen Input-Spikes durchlaufen wird. Eine zeitliche Integration liefert also denselben Wert für alle Output-Neuronen j und somit keinerlei verwertbare Information. Eine interessante Fragestellung ist vielmehr, ob die Input-Spikes zeitlich konzentriert auf einzelnen Delaytrajektorien liegen, die durch eines der Output-Neuronen repräsentiert werden, oder nicht. Als bekanntes Maß für Konzentration wird häufig die *Entropie* verwendet (siehe Shannon und Weaver, 1949). Szepannek und Weihs (2006a) motivieren von

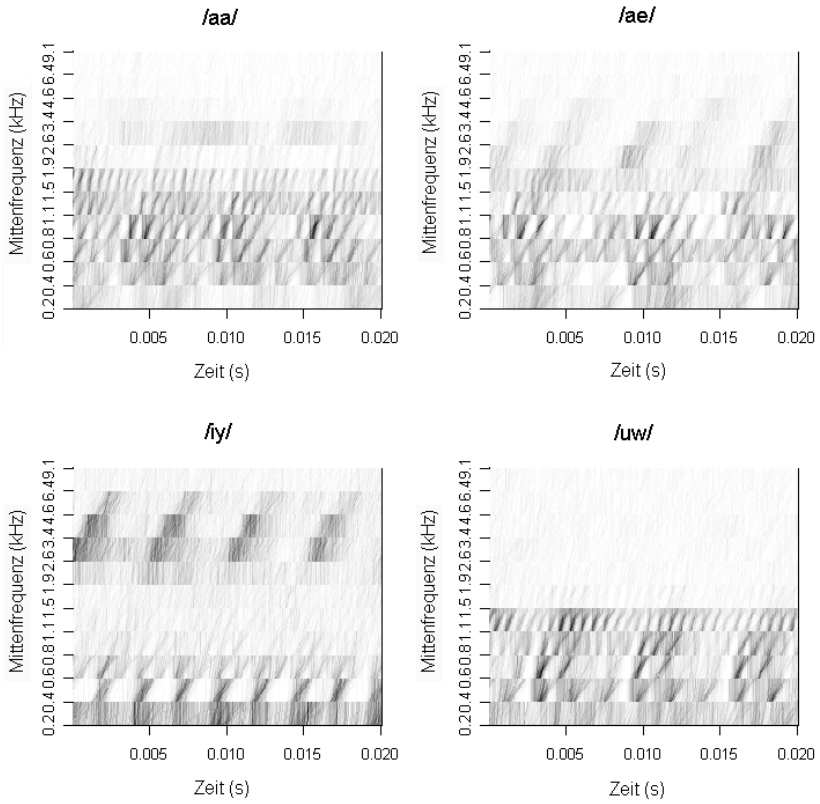


Abbildung 3.9: Exemplarischer Output der PLDCNs auf 20 ms Zeitsegmenten in der Mitte von vier unterschiedlichen Vokalen (vgl. Abbildung 2.11 und 3.7.)

daher als Merkmalsrepräsentation zur Charakterisierung eines Schallsignals auf Basis des DCN Outputs die Entropie der unterschiedlichen Output-Neuronen über den zeitlichen Verlauf eines Zeitfensters. Diese lässt sich beschreiben durch die Gleichung:

$$X_j^{Entropie} = -\sum_t p_j(t) \log_2(p_j(t)) \tag{3.35}$$

mit $p_j(t) := \frac{Y_j(t)}{\sum_u Y_j(u)}$.
Beispielhaft ist in Abbildung 3.11 die Antwort dreier unterschiedlicher Output-Neuronen auf ein Schallsignal während eines Zeitfensters von 20 ms dargestellt. Es ist deutlich die höhere zeitlich konzentrierte Aktivität des Neurons im rechten Fenster zu erkennen. Dies wird widergespiegelt durch eine niedrigere Entropie $X_{rechts}^{Entropie} = 7.76$ im Verhält-

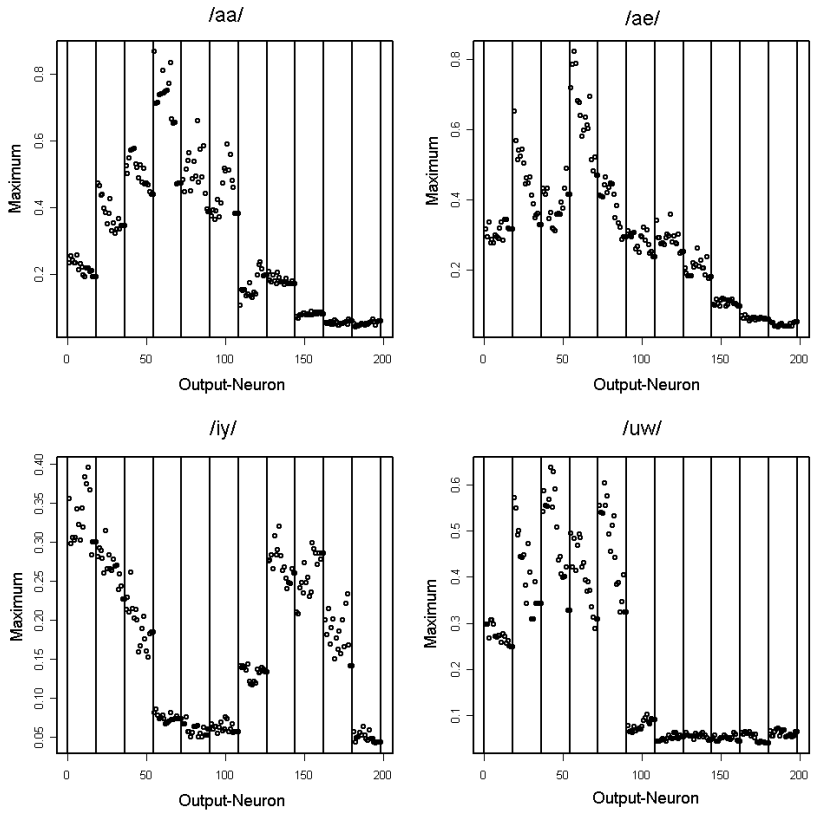


Abbildung 3.10: Beispiel: Maximale Werte $\max_t Y_j(t)$ der PLDCN Output-Neuronen auf 20 ms Zeitsegmenten in der Mitte von vier verschiedenen Vokalen (vgl. Abbildungen 2.11, 3.7 und 3.9.)

nis zur Entropie im linken Fenster von $X_{links}^{Entropie} = 8.00$ bei der keine derartige zeitlich synchrone Antwort der Input-Neuronen entlang der zugehörigen Delaytrajektorie beobachtbar ist. Die Aktivität in der mittleren Grafik liegt vom Synchronisationsgrad zwischen denen der beiden anderen. Dies ist auch anhand der Entropie zu erkennen.

Eine andere Merkmalsrepräsentation lässt sich aus einem neurophysiologischen Blickwinkel motivieren: In neuronalen Systemen setzen Neuronen immer dann ein Aktionspotenzial frei, wenn ihr Potenzial einen Schwellenwert θ überschreitet. Ihr Potenzial erhöht sich dabei durch die eingehenden APs der Input-Neuronen. Weiterhin ist bekannt, dass die Feuerzeitpunkte der ANFs (und damit auch der Wert der DCN Output-Neuronen) für niedrige Frequenzen der Stimulusperiode folgen und auch für

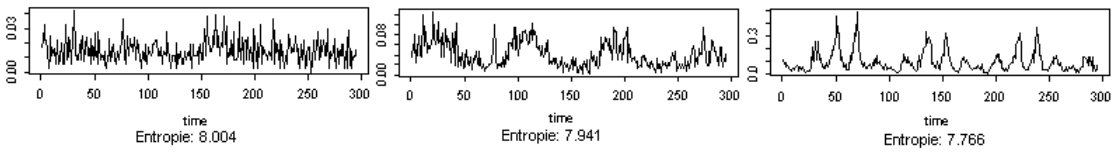


Abbildung 3.11: Beispiel: Vergleich der Entropie $X_j^{Entropie}$ dreier unterschiedlicher DCN Output-Neuronen auf einem 20 ms Fenster.

hohe Frequenzen die Refraktärzeit kein durchgängiges Feuern der ANFs ermöglicht. Unter der Annahme, eine bestimmte Trajektorienform sei repräsentativ für eine bestimmte Lautklasse, sollte der Schwellenwert θ bei Vorliegen dieses Lautes an einigen (aber nicht zu allen) Zeitpunkten t überschritten werden, nicht (oder nur selten) aber bei Vorliegen von anderen Lautklassen. Es sollte optimalerweise gelten

$$\max_t Y_j(t) < \theta \tag{3.36}$$

für das zur entsprechenden Trajektorie gehörende Output-Neuron j und alle Zeitpunkte t, in denen der entsprechende Laut nicht vorliegt. Szepannek und Weihs (2006a) motivieren auf diese Weise

$$X_j^{max} = \max_t Y_j(t) \tag{3.37}$$

als Merkmalsrepräsentation des (PL)DCN-Outputs eines Sprachsignals.
Benachbart gelegene Output-Neuronen (paralleler lokaler) Delay-Computing Netzwerke beschreiben ähnliche Kurvenverläufe der Wanderwelle, so dass zwischen den Ausprägungen positive Korrelation zu erwarten ist. Die Anzahl der Output-Neuronen in den implementierten Netzwerken liegt über 100, so dass eine Zusammenfassung mehrerer X_j^{max} bzw. $X_j^{Entropie}$ zu einem Merkmal X_k^{max} bzw. $X_k^{Entropie}$ nahe liegt. Die einfachste Möglichkeit der Zusammenfassung besteht in der Aggregation über jeweils gleich vielen benachbarte X_j^{max}, für DCNs z.B. zu insgesamt zwölf Merkmalen durch

$$X_k^{max} = \sum_{j=12(k-1)+1}^{12k} X_j^{max},\ k = 1, \ldots, 12. \tag{3.38}$$

Basierend auf der Beobachtung, dass die Hauptaktivität auf den inneren Output-Neuronen eines DCNs liegt, schlagen Harczos u. a. (2007b) eine breitere Gruppierung

an den Rändern vor, gemäß Gleichung

$$X_k^{max} = \sum_{j=\sum_{l<k} h_l+1}^{\sum_{l\leq k} h_l} X_j^{max},\ k = 1, \ldots, 12 \tag{3.39}$$

mit

$$h_l = \left[a \left(\sin(\pi + \pi \frac{l-1}{12-1})\right)^b + c\right] \tag{3.40}$$

und [.] der Gaußklammer, $a = 7, b = 3$ und $c = 15.13$ (vgl. Abb. 3.12).

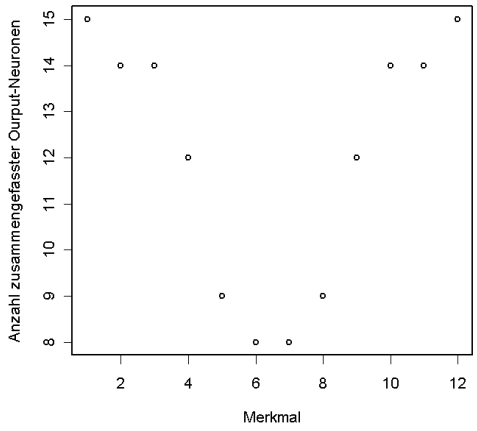


Abbildung 3.12: Anzahl h_l zu Merkmalen gruppierter Output-Neuronen.

Im Fall von PLDCNs bietet es sich an, die Output-Neuronen jedes lokalen Netzwerkes in 2/3/... Merkmale zusammenzufassen, die entsprechend einen mehr oder weniger steilen Trajektorienverlauf entlang der Netzwerk-Input-Neuronen repräsentieren.
Auf Basis der (PL)DCN basierten Merkmale lässt sich eine Erweiterung der Merkmale $X_j^{OD,ISI}$ vornehmen, indem der kombinierte Merkmalsvektor zusätzlich um die Delay-Computing basierten Merkmale $X_j^{Entropie}$ oder X_j^{max} ergänzt wird. Ein Standardvorgehen zur Kombination verschiedener Merkmalssätze als Front Ends zur automatischen Spracherkennung besteht in deren einfacher Konkatenation (vgl. Ellis, 2000). Ergänzt wird dieses Vorgehen in dieser Arbeit durch eine anschließende Dimensionsreduktion. Diese soll helfen, redundante Information, die in beiden einzelnen Merkmalssätzen enthalten ist, auf einen niedrigerdimensionalen Merkmalsvektor zu verdichten, dessen

Merkmale zudem eine geringe Korrelation untereinander aufweisen. Es entsteht der Merkmalsvektor $X^{OD,ISI,Delay}$ (siehe Kapitel 5) .

3.9.4 Voruntersuchung: Vokalerkennung mit Delay-Computing basierten Merkmalen

An dieser Stelle sollen die Ergebnisse einer ersten Untersuchung von Szepannek und Weihs (2006a) zum Vergleich der verschiedenen vorgestellten Merkmale auf Basis von (PL)DCNs wiedergegeben werden. Ziel war die automatische Klassifikation von vier verschiedenen Vokalen: /aa/, /ae/, /iy/ und /uw/ der TIMIT-Datenbank (siehe Kapitel 6.2.1 und Anhang C.1). Diese Auswahl bietet sich an, da die vier Vokale gerade die extremen Artikulationspositionen (des Mundes) repräsentieren (siehe z.B. Ali u. a., 2002; Schukat-Talamazzini, 1995, S. 25). Zum Vergleichen der verschiedenen Merkmale wurden ein Trainings- und ein Testdatensatz mit je 16 Vokalen aus jeder der vier Vokalklassen (randomisierte Auswahl je einer Frau und eines Mannes aus jeder von acht Dialektregionen der kompletten TIMIT Datenbank) erzeugt. Grundlage zur Merkmalsbildung stellte jeweils der Merkmalsvektor eines einzigen Zeitfensters von 20 ms genau in der Mitte des (PL)DCN-Outputs der entsprechenden Vokalrealisierung dar. Zur Klassifikation wurde *Penalisierte Diskriminanzanalyse (PDA, Hastie u. a., 1995)* verwendet, da diese sich in Untersuchungen von Szepannek u. a. (2006) als beste Methode zu diesem Zweck herausgestellt hat. Ein Hauptunterschied zwischen penalisierter- und linearer Diskriminanzanalyse besteht in der Regularisierung der Kovarianzmatrix der Daten durch *Shrinkage* (dt. *Schrumpfung*, vgl. z.B. Schmidt und Trenkler, 1998, S. 214 ff). Dieses Prinzip wird in Abschnitt 5.3.3 zur Dimensionsreduktion der Daten aufgegriffen.

Als Merkmale wurden X^{max} und $X^{Entropie}$ zum Vergleich jeweils auf dem Output eines einzigen großen DCNs (141 Input-Neuronen, nach Harczos u. a., 2006a, vgl. Abschnitt 3.9.1) sowie auf dem Output von elf PLDCNs der Größe 18 (vgl. Abschnitt 3.9.2) gebildet. Abbildung 3.7 auf Seite 48 zeigt an vier beispielhaften Vokalen deutliche Unterschiede in der Aktivität der Neuronen verschiedener Netzwerke (unterschiedliche Grautöne lassen die Zuordnungen der Neuronen zu den Netzwerken erkennen, unterscheiden sich aber von Vokal zu Vokal). Aufbauend auf dieser Beobachtung

wurde zusätzlich die *Aktivierung ganzer lokaler Netzwerke* als Merkmalsrepräsentation untersucht, die ausgedrückt wird durch die Merkmale

$$X_k^{Act} = \sum_j \sum_t Y_{kj}(t). \tag{3.41}$$

Die Bezeichnungen sind dabei analog zur Terminologie in Abschnitt 3.9.2 gewählt mit t als Zeit, k als Index des lokalen Netzwerkes und j als Index der Neuronen eines lokalen Netzwerks. Wie Abbildung 3.13 zeigt, ergeben sich deutlich erkennbare Unterschiede in den durchschnittlichen Merkmalsausprägungen X_k^{Act} auf den Trainingsdaten für die vier verschiedenen Vokale. Durch die unterschiedlichen Input-Neuronen der lokalen Netzwerke und deren Eigenschaft, zusammenhängend bezüglich der Input-Neuronen zu sein, lassen sich die X_k^{Act} spezifischen Frequenzbereichen zuordnen, die Merkmalsrepräsentation besitzt also erneut einen spektralen Charakter. Dies motiviert eine zusätzliche cepstrale Transformation (siehe Abschnitt 3.3). Abbildung 3.14 zeigt die Test-Erkennungsraten. Es ist zu sehen, dass sich die Klassifikationsergebnisse stark verbessern, wenn der Vokal /uw/ nicht in die Analyse einbezogen wird. Im Verhältnis zu einfachen Delay-Computing Netzwerken werden durch Verwendung von PLDCNs deutlich bessere Erkennungsraten erzielt. Die Informationskodierung über die Gestalt der Verzögerungsstruktur der Cochlea-Wanderwelle allein erreicht geringere Erkennungsraten als die frequenzinformation beinhaltenden Merkmale der Netzwerkaktivität. Eine anschließende cepstrale Transformation der Merkmale kann die Klassifikationsergebnisse nochmals verbessern.

Bemerkung 3.1 *Die folgende Überlegung skizziert, dass es sich bei der durch die lokale Netzwerkaktivität erfasste Information im Wesentlichen um Orts-Durchschnittsfeuerraten-Information handelt, wie sie in Abschnitt 3.4 beschrieben wird:*
Für die Merkmale X_k^{Act} gilt: Alle Input-Spikes erreichen jedes Output-Neuron eines lokalen DCNs k, allerdings zu unterschiedlichen Zeitpunkten (vergleiche Abschnitt 3.9.2). Somit gilt:

$$X_k^{Act} = \sum_j \sum_t Y_{kj}(t) = N_k \sum_t Y_{kl}(t); \tag{3.42}$$

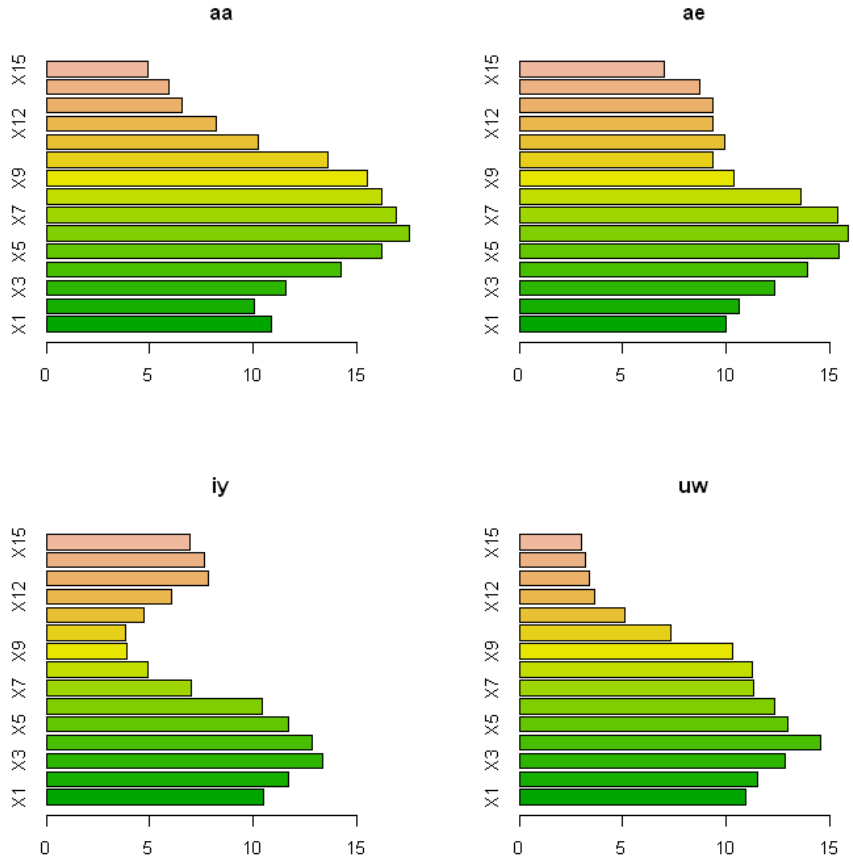


Abbildung 3.13: Durchschnittliche Merkmale X_k^{Act} für die vier verschiedenen Vokale auf den Trainingsdaten.

für ein beliebiges l[5]. Weiterhin gilt

$$\sum_t Y_{kl}(t) = \sum_i \sum_t X_i(t). \tag{3.43}$$

Der Laufindex i bezeichnet dabei sämtliche Input-Neuronen des lokalen Netzwerks k. Die Merkmale X_k^{Act} sind also äquivalent zu geglätteten durchschnittlichen Feuerraten über alle Netzwerk-Input-Neuronen und entsprechen somit den Merkmalen $X_k^{OD_5}$ in

[5]Kleine Unterschiede in $\sum_t Y_{kl}(t)$ für verschiedene l sind höchstens auf die – durch die Fensterbreite vorgegebenen – Schranken der Summe zurückzuführen, falls ein AP zwei Output-Neuronen eines (PL)DCNs in verschiedenen Zeitfenstern passiert.

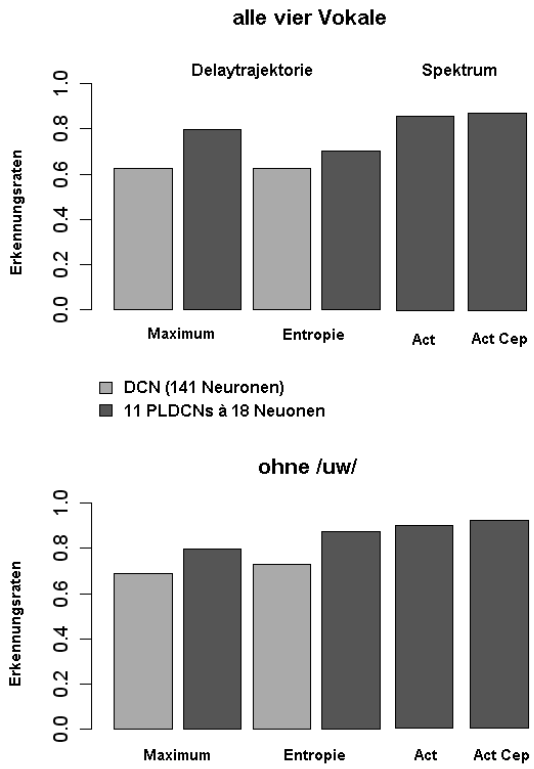
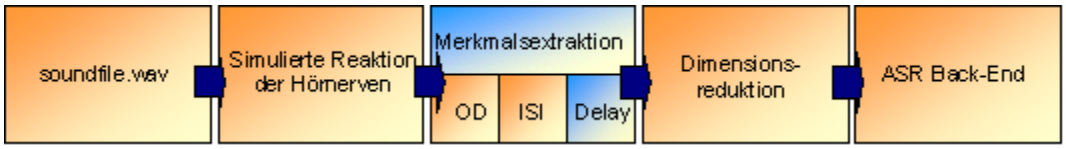


Abbildung 3.14: Ergebnis: Vokalerkennung auf Basis der unterschiedlichen (PL)DCN basierten Merkmale.

Abschnitt 3.4.

3.10 Lateral inhibitorische neuronale Netzwerke (LIN)



Das Phänomen der Verlangsamung der Cochlea-Wanderwelle am Ort maximaler Auslenkung der Basilarmembran wird auch von Shamma (1985a) beschrieben: Vor Erreichen dieser Position bewegen sich die Auslenkungen an benachbarten Positionen der BM näherungsweise in Phase, am Ort maximaler Auslenkung kommt es aufgrund der

beschriebenen Verlangsamung der Welle zu einer Phasenverschiebung.
Shamma (1985b) nutzt diese Beobachtung zur Motivation eines neurophysiologisch inspirierten *lateral inhibitorischen Neuronalen Netzwerkes (LINs)* der neuronalen Weiterverarbeitung.
Für Output-Neuronen $Y_k(t)$ eines einfachen LINs zum Zeitpuntk t gilt:

$$Y_k(t) = \alpha_{kk} X_k(t) - \sum_{j \neq k} \alpha_{jk} X_j(t), \tag{3.44}$$

mit $\alpha_{jk} \geq 0$ und $X_k(t)$ den Input-Neuronen, die in dieser Arbeit durch die Aktivität der Hörnerven (beschrieben durch das PSTH, vgl. Anhang B.1 bzw. S. 30) gegeben sind. Abbildung 3.15 zeigt den schematischen Aufbau eines lateral inhibitorischen neuronalen Netzwerks. Die Netzwerkstruktur bewirkt, dass jedes Output-Neuron mit ei-

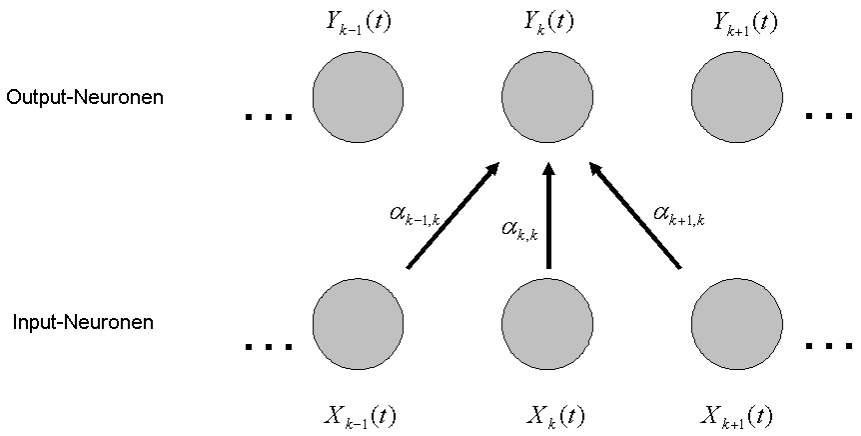


Abbildung 3.15: Schematische Darstellung eines LINs.

nem spezifischen Hörnerven assoziiert wird und dessen Aktivität beschreibt, gehemmt durch die gleichzeitige Aktivität der übrigen Hörnerven. Intuitiv sollte der inhibitorische Einfluss α_{jk} mit zunehmender Größe von $|j-k|$ sinken. Shamma (1985b) modelliert einen Einfluß $\alpha_{jk} > 0$ für einen Mittenfrequenzabstand innerhalb von etwa 1/3 Oktave. Giron (2006) verwendet ein sehr einfaches Netzwerk mit Parametern

$$\alpha_{jk} = \begin{cases} 1, & j = k \\ 0.5, & |j-k| = 1 \\ 0, & sonst. \end{cases} \tag{3.45}$$

Diese ist im Einklang zu den von Shamma (1985b) im Anhang beschriebenen durchgeführten Untersuchungen und soll auch in dieser Arbeit verwendet werden, wobei die PSTHs auf Basis von je fünf bzw. acht ANFs – beruhend auf den Arbeiten von Seneff (1988) bzw. Allen (1994) – über einen (CF-) Frequenzbereich zwischen etwa 200 und 6400 Hz gebildet werden (vgl. auch Abschnitt 3.4). Als Bingröße des PSTH ist entsprechend Shamma (1985b) ein Wert von etwa drei bis fünf Zeitintegrationsschritten des auditorischen Modells zu wählen; zur Verwendung kommen hier Bingrößen der Länge 1/14700 s. Eine mögliche Alternative besteht in einer frequenzabhängigen Bingrößenwahl der verschiedenen PSTHs an unterschiedlichen Positionen der Cochlea, die mehr dem Verlauf der cochleären Wanderwelle gerecht würde. Hierzu könnten Messungen der Impulsantwort-Verzögerung an den Hörnerven, wie sie von Harczos (2007) am Fraunhofer IDMT durchgeführt wurden, herangezogen – und mit dem allgemeinen, von Greenberg u. a. (1997) beschriebenen Zusammenhang $\text{Delay}(CF) = \frac{a}{CF} + b$ abgeglichen werden. Für die Untersuchungen dieser Arbeit wird jedoch die oben beschriebene kosntante Wahl der Bingrößen verwendet.
Zur Weiterverarbeitung der $Y_k(t)$ zu Merkmalen schlagen Yang u. a. (1992) zunächst eine Halbwellengleichrichtung der Form

$$Y_k^*(t) = \max(Y_k(t), 0), \tag{3.46}$$

d.h. weiter betrachtet wird nur die positive Aktivität der Output-Neuronen, motiviert durch die nichtlineare neuronale Signalverarbeitung, die erst bei Schwellenwertüberschreitung Aktivität zeigen. Die resultierenden Merkmale ergeben sich durch zeitliche Integration $X_k^{LIN} = \sum_t Y_k^*(t)$ über die Dauer eines Zeitfensters.
Über die mit den Output-Neuronen des LINs assoziierbaren CFs wird durch diese Merkmalsrepräsentation auch die Gestalt des Trajektorienverlaufs in Frequenzinformation überführt. Die entstehenden Merkmale sind damit entsprechend Abschnitt 3.7 kombinierbar mit den Merkmalen X_ω^{OD}, X_ω^{ALSD} oder $X_\omega^{OD,ISI}$.

3.11 Zusammenfassung

Es wurden drei Theorien über den Informationsgehalt in der neuronalen Antwort an den Hörnerven vorgestellt: basierend auf dem Ort der neuronalen Aktivität, auf Zeitin-

tervallen zwischen sukzessiven Aktionspotenzialen einer Nervenfaser (Phase Locking) sowie auf der Art der Verzögerung der Wanderwelle entlang der Basilarmembran.
Verschiedene Alternativen einer Merkmalsextraktion aus dem simulierten neuronalen Output wurden vorgestellt: Bekannte Verfahren der Merkmalsextraktion zur automatischen Spracherkennung aus auditorischen Modellen (GSD/ALSD, EIH, LIN sowie Orts-Durchschnittsfeuerraten) wurden auf den Output des in dieser Arbeit verwendeten, detaillierten auditorischen Simulationsmodells übertragen. All diese Merkmalssätze repräsentieren spektrale Signalkomponenten, die nach der auditorischen Verarbeitung in der Antwort der Hörnerven enthalten sind.
Ein probabilistischer Ansatz der Auffassung der neuronalen Aktivität der Hörnerven als Punktprozesse erlaubt die Motivation der vorgestellten Merkmale und motiviert weiterhin, auch in der neuronalen Antwort enthaltene Phaseninformation zu nutzen.
(Parallele lokale) neuronale Delay-Computing Netzwerke wurden beschrieben zur Verarbeitung des Outputs der Hörnerven mit dem Zweck einer gezielten Detektion von Wanderwellenverzögerungsstrukturen. Aus dem Output dieser Netzwerke wurden Merkmale motiviert, die einzig auf der Gestalt der vom Signal erzeugten Wanderwellen-Delaytrajektorien ohne explizite Ausnutzung von im Signal enthaltener Frequenzinformation basieren.

Merkmal	Ort	Phase Locking	Delay	M-ANF	(PL)DCN
$X^{OD.}$	+				
$X^{GSD/ALSD}$	(+)	+		+	
$X^{OD,ISI}$	+	+		+	
$X^{Entropie}$			+		+
X^{max}			+		+
X^{LIN}	(+)		+	+	
$X^{OD,ISI,Delay}$	+	+	+	+	+

Tabelle 3.1: Auflistung und Kategorisierung der vorgestellten Merkmalsrepräsentationen.

Tabelle 3.1 fasst die vorgestellten Merkmale zusammen und *charakterisiert* diese anhand der Art der Weiterverarbeitung des Outputs aus dem neuronalen Simulationsmodell: Die ersten drei Spalten beschreiben die Art der Informationskodierung; ein (+), z.B. im Falle von X^{GSD} in Spalte „Ort “ symbolisiert hierbei, dass Information über den Ort der Anregung zwar nicht explizit als Haupteigenschaft – jedoch impli-

zit über den Vergleich mit der CF in den Merkmalsausprägungen enthalten ist. Die folgenden beiden Spalten beschreiben die Datengrundlage der beschriebenen Merkmale: Werden (PL)DCNs verwendet? Ist eine wiederholte (M-ANF) Simulation der Hörnervenreaktion für die Merkmalsbildung erforderlich (vgl. Abschnitt 2.3)?
Eine anschließende Transformation der Merkmale auf Basis von Methoden der linearen Dimensionsreduktion ist prinzipiell für alle Merkmale möglich und wird in Kapitel 5 beschrieben.
Eine erste Voruntersuchung zur Vokalklassifikation zum Vergleich der (PL)DCN basierten Merkmale belegt den Nutzen von Frequenzinformation zur Vokalerkennung sowie insbesondere, dass gute Ergebnisse auf Basis einer Orts-Durchschnittsfeuerratenkodierung zu erwarten sind.
Eine vergleichende Evaluierung sämtlicher vorgestellter Merkmale erfolgt im Kapitel 6. Zur Erkennung werden dabei als Back Ends Hidden Markov Modelle (HMMs) verwendet; diese sind im folgenden Kapitel 4 beschrieben.

4 Erkennung

4.1 Klassifikation

Um die in Kapitel 3 beschriebenen Merkmale (Front Ends) zur automatischen Erkennung kontinuierlicher Sprache einsetzen zu können, sind zur Weiterverarbeitung sogenannte *Back Ends* nötig, die die erzeugten Merkmalsvektoren einem Alphabet zuordnen und so eine Transkription ermöglichen. Im Gegensatz zu der Klassifikationsaufgabe in Abschnitt 3.9.4 führen Back Ends die Mustererkennung nicht auf Basis eines einzelnen Merkmalsvektors X durch, sondern auf Grundlage einer Sequenz von Merkmalsvektoren $\{X_1, \ldots, X_T\}$, im Folgenden kurz mit X bezeichnet, wobei X_t die Vektoren zu den einzelnen (diskreten) Zeitpunkten (im Abstand von zehn ms) bezeichnet.

Bezeichne w eine zu erkennende Wortfolge bzw. ein zu erkennendes Wort. Als Erkennungsaufgabe stellt sich die Suche nach der wahrscheinlichsten $\hat{w}$, gegeben die beobachtete Merkmalssequenz X

$$\hat{w} = \arg\max_{w} P(w|X). \tag{4.1}$$

Aufgrund der Bayes-Regel (und da der Nenner der mittleren Gleichung konstant ist) ergibt sich:

$$\hat{w} = \arg\max_{w} \frac{P(X|w)P(w)}{P(X)} = \arg\max_{w} P(X|w)P(w). \tag{4.2}$$

Die rechte Seite der Gleichung setzt sich aus zwei Bestandteilen zusammen: dem *akustischen Modell* $P(X|w)$ sowie dem *Sprachmodell* $P(w)$. Motivierbar sind diese Modelle

durch das *Quelle-Kanal Modell* der Spracherzeugung (vgl. Jelinek, 1997, S. 9). Es wird davon ausgegangen, dass an der Sprachquelle (einem Sprecher) mit einer Wahrscheinlichkeit $P(w)$ die Wortfolge w generiert wird. Diese manifestiert sich in Form der entstehenden Schallwelle. Der akustische Kanal hat nun eine Vermittlerrolle zwischen Sprachquelle und Spracherkennungssystem: Die Schallwelle wird durch eine Merkmalssequenz X mit einer Wahrscheinlichkeit $P(X|w)$ repräsentiert.
Zur Lösung der Gleichung 4.2 sind nun ein *akustisches Modell* und ein *Sprachmodell* nötig, die in den folgenden Abschnitten beschrieben werden.

4.2 Akustische Modellierung: Hidden Markov Modelle

4.2.1 Einleitung

Der Einsatz von Hidden Markov Modellen (HMMs, Rabiner, 1989) zur akustischen Modellierung hat in den vergangenen Jahren die Verwendung von *Dynamischer Zeitverzerrung*[1] abgelöst.
Die Grundidee von HMMs stellen Markov-Prozesse von Zustandsfolgen $S_t, t = 1, \ldots, T$, dar. Die Menge möglicher Zustände bildet einen diskreten, endlichen Zustandsraum $\{q_1, \ldots, q_M\}$. Für die Wahrscheinlichkeit des Prozesses der Annahme eines Zustands zum Zeitpunkt t gilt:

$$P(S_t|S_1, \ldots, S_{t-1}) = P(S_t|S_{t-1}). \tag{4.3}$$

Diese vereinfachende Annahme wird als *Markov Annahme* bezeichnet: Die Wahrscheinlichkeit des Vorliegens eines Zustands q_j zum aktuellen Zeitpunkt t hängt lediglich vom Zustand des Prozesses zum vorangehenden Zeitpunkt $t-1$, nicht jedoch von der restlichen Vergangenheit des Prozesses ab. Die *Zustandsübergangswahrscheinlichkeiten* $P(S_t = q_j|S_{t-1} = q_i) =: a_{ij}$ bleiben über die Zeit konstant und lassen sich in einer Matrix

$$A = \{a_{ij} : i, j = 1, \ldots, M\} \tag{4.4}$$

[1] In der Dynamischen Zeitverzerrung (engl. Dynamic Time Warping, DTW) wird ein optimaler Zuordnungspfad zwischen zwei Merkmalsfolgen bestimmt. Mustererkennung erfolgt durch Vergleich einer neu beobachteten Merkmalsfolge mit einer Menge von Referenzmustern und deren Zuordnung zum nächstliegenden Repräsentanten bezüglich eines Abstandsmaßes (siehe Schukat-Talamazzini, 1995, S. 121 ff).

zusammenfassen. Für den Zustand des Systems zum Zeitpunkt $t = 1$ gelten die Zustandsstartwahrscheinlichkeiten

$$\pi_{q_i} := P(S_1 = q_i),\ i = 1, \ldots, M. \tag{4.5}$$

Der Begriff *Hidden* Markov Modell rührt von dem Sachverhalt her, dass nicht die Zustandsfolge $\{S_1, \ldots, S_T\}$ selbst beobachtet werden kann, sondern lediglich die Merkmalsfolge $X = \{X_1, \ldots, X_T\}$, die Zustandsfolge selbst bleibt also „verborgen“. Es wird jedoch angenommen, dass jeder Zustand q_i eine spezifische Merkmalsverteilung

$$P_{q_i}(X_t) = P(X_t|S_t = q_i), \forall\, t \tag{4.6}$$

besitzt, die Rückschlüsse über die vorliegende Zustandsfolge zu einer beobachteten Sequenz von Merkmalsvektoren $\{X_1, \ldots, X_T\}$ erlaubt. Ein *Hidden Markov Modell* λ ist durch das Tripel von Zustandsübergangswahrscheinlichkeitsmatrix, Zustandsstartwahrscheinlichkeitsvektor und die Emissionsverteilungen der Dichten Zustände charakterisiert.

4.2.2 Wahrscheinlichkeit für das Auftreten einer beobachteten Merkmalssequenz

Einzelworterkennung

Sei zunächst der Einfachheit halber die Problemstellung einer *Einzelworterkennung* gegeben, d.h. es existiert genau ein Modell λ_k je Wort eines zu erkennenden Wortschatzes und die Erkennungsaufgabe besteht darin, eine Sequenz von Merkmalsvektoren korrekt genau einem Wort zuzuordnen. Gesucht ist die Wahrscheinlichkeit für $P(X|\lambda_k)$ des *akustischen Modells*.
Bei gegebener Zustandsfolge $s = \{S_1 = s_1, \ldots, S_T = s_T\}$ gilt

$$P(X|s, \lambda_k) = \prod_{t=1}^{T} P_{s_t}^{(k)}(X_t). \tag{4.7}$$

Weiterhin lässt sich für das Durchlaufen einer beliebigen Zustandsfolge s aufgrund der Markov-Eigenschaft deren Auftretenswahrscheinlichkeit berechnen:

$$P(s|\lambda_k) = \pi_{s_1}^{(k)} \prod_{t=2}^{T} a_{s_{t-1}, s_t}^{(k)} \tag{4.8}$$

Daraus ergibt sich für die gemeinsame Wahrscheinlichkeit von einer Merkmalsfolge und einer zugehörigen Zustandssequenz:

$$P(X, s|\lambda_k) = \pi_{s_1}^{(k)} P_{s_1}^{(k)}(X_1) \prod_{t=2}^{T} a_{s_{t-1},s_t}^{(k)} P_{s_t}^{(k)}(X_t). \tag{4.9}$$

Die *Produktionswahrscheinlichkeit* resultiert hieraus als Summe über alle möglichen Zustandsfolgen s:

$$P(X|\lambda_k) = \sum_{s} P(X, s|\lambda_k). \tag{4.10}$$

Eine effektive Berechnung der Produktionswahrscheinlichkeiten in $T \cdot M$ Schritten ist mit dem *Forward-Algorithmus* möglich. Es werden Vorwärtsvariablen

$$\alpha_t(q_i^{(k)}) = P(X_1, \ldots, X_t, S_t = q_i^{(k)}|\lambda_k) \tag{4.11}$$

definiert als diejenigen Wahrscheinlichkeiten, dass λ_k zum Zeitpunkt t den Zustand $q_i^{(k)}$ erreicht und bis zu diesem Zeitpunkt die Merkmalsvektoren $X_1, \ldots, X_t$ beobachtet werden. Initialisiert werden die Vorwärtsvariablen durch:

$$\alpha_1(q_i^{(k)}) = \pi_{q_i}^{(k)} P_{q_i^{(k)}}(X_1)^{(k)}. \tag{4.12}$$

Für alle folgenden Zeitpunkte $t = 2, \ldots, T$ lassen sich die $\alpha_t(i)$ rekursiv aus den Vorwärtsvariablen des vorangegangenen Zeitpunkts, gegeben durch

$$\alpha_t(q_j^{(k)}) = \sum_{i=1}^{M} \alpha_{t-1}(q_i^{(k)}) a_{ij}^{(k)} P_{q_j^{(k)}}^{(k)}(X_t) \tag{4.13}$$

berechnen. Letztendlich ergibt sich die Produktionswahrscheinlichkeit als

$$P(X|\lambda_k) = \sum_{i=1}^{M} \alpha_T(q_i^{(k)}). \tag{4.14}$$

In der Praxis wird jedoch häufig der *Viterbi-Algorithmus* verwendet, der anstatt der Summation über alle möglichen Pfade lediglich den Pfad der wahrscheinlichsten Zustandsfolge bestimmt. Er ist sehr ähnlich zum Forward-Algorithmus. Ausgehend von initialen $\alpha_1^*(q_i^{(k)})$ genau wie in Gleichung 4.12 wird der jeweils wahrscheinlichste Pfad bis zum Zeitpunkt $t = 2, \ldots, T$ berechnet durch

$$\alpha_t^*(q_j^{(k)}) = \max_{i=1,\ldots,M} (\alpha_{t-1}^*(q_i^{(k)}) a_{ij}^{(k)} P_{q_j^{(k)}}^{(k)}(X_t)). \tag{4.15}$$

Schließlich gilt für die Wahrscheinlichkeit des optimalen Pfades:

$$P(X, s^*|\lambda_k)^* = \max_{i=1}^{M} \alpha_T^*(q_i^{(k)}). \tag{4.16}$$

Diese wird häufig an Stelle der mit dem Forward-Algorithmus berechneten Produktionswahrscheinlichkeit (Gleichung 4.14) verwendet. Als Argumentation für die Anwendung des Viterbi-Algorithmus wird oft angeführt, dass es sich bei fast allen Pfad-Wahrscheinlichkeiten um sehr kleine Werte handelt, die kaum Einfluss auf die Produktionswahrscheinlichkeit besitzen und diese i.d.R. durch die Pfadwahrscheinlichkeit der wahrscheinlichsten Zustandsfolge dominiert wird. Insbesondere lässt sich der Viterbi-Algorithmus aber auch leicht zur Erkennung kontinuierlicher Sprache verwenden (Young u. a., 2005, S. 9).

Erweiterung auf den Fall kontinuierlicher Spracherkennung

Besteht die Erkennungsaufgabe in der Klassifikation kontinuierlicher gesprochener Sprache wie z.B. ganzer Sätze, so lässt sich die Anwendung des Viterbi-Algorithmus leicht erweitern durch Verkettung mehrerer hintereinandergeschalteter HMMs. Um diese zu ermöglichen, werden künstliche Zwischenzustände $q_0^{(k)}$ und $q_{M+1}^{(k)}$ (ohne Merkmalsemission) definiert. Ein Übergang von HMM λ_k nach HMM λ_l, so dass der Anfangszustand $q_0^{(l)}(t)$ zum Zeitpunkt t angenommen wird, erfolgt, wenn

- entweder der Endzustand $q_{M+1}^{(k)}(t-1)$ des Vorgänger HMMs λ_k zum Zeitabschnitt $t-1$
- oder der erste Zustand $q_0^{(k)}(t)$ des Vorgänger HMMs λ_k zum Zeitabschnitt t gefolgt vom unmittelbaren Übergang $q_0^{(k)}(t) \rightarrow q_{M+1}^{(k)}(t)$ (mit Wahrscheinlichkeit $a_{0,M+1}^{(k)}$) während des Zeitabschnitts t erreicht wird.

Letzteres entspricht dem Überspringen bzw. Auslassen des HMMs λ_k, interpretierbar beispielsweise mit dem Verschlucken eines Lauts innerhalb einer Sprachsequenz.
Ein Anfangszustand $q_0^{(l)}$ kann also innerhalb des selben Zeitabschnitts aus einem anderen HMM λ_k nur durch das emissionslose Überspringen $q_1^{(k)}(t) \rightarrow q_{M+1}^{(k)}(t)$ erfolgen. Für einen beliebigen Zustand $q_j^{(k)}$ eines HMMs λ_k ergibt sich die Vorwärtswahrscheinlichkeit

$$\alpha_t(q_j^{(k)}) = \left(\alpha_t(q_0^{(k)}) a_{0,j}^{(k)} + \sum_{i=2}^{M} \alpha_{t-1}(q_i^{(k)}) a_{i,j}^{(k)} \right) P_{q_j^{(k)}}^{(k)}(X_t), \tag{4.17}$$

d.h. der Zustand $q_j^{(k)}$ kann zum selben Zeitabschnitt aus dem Zustand $q_0^{(k)}$ hervorgehen (mit Wahrscheinlichkeit $a_{0,j}^{(k)}$) oder aus allen anderen Zuständen $q_i^{(k)}, i = 1, \ldots, M$, sofern diese einen Zeitabschnitt zuvor vorlagen, nicht mehr jedoch aus Zustand $q_{M+1}^{(k)}$. Von dort aus erfolgt ein Übergang zum nachfolgenden HMM (für Details, siehe Young u. a., 2005, S. 120). Der obere Index $.^{(k/l)}$ bezeichnet hierbei das entsprechende HMM. Es wird hier die Anzahl M an Zuständen für alle HMMs als gleich vorausgesetzt, für unterschiedliche, HMM-spezifische Zustandsanzahlen ergibt sich lediglich eine Änderung der Notation.

Im Fall kontinuierlicher Spracherkennung sind die Hidden Markov Modelle in der Regel allerdings nicht durch ganze Wörter gegeben, da in diesem Fall zur Parameterschätzung aller HMMs zu viele Trainingsdaten nötig wären. Anstelle dessen repräsentieren die HMMs hier i.d.R. sogenannte *Unterworteinheiten* (beispielsweise Phoneme), aus deren Aneinanderreihung sich dann die Worte ergeben. Häufig werden Phoneme zudem noch je nach Kontext differenziert, d.h. es werden verschiedene HMMs für beispielsweise ein a gebildet, je nach vorangehendem und nachfolgendem Laut (vergleiche hierzu z.B. Schukat-Talamazzini, 1995, Kapitel 6).

4.2.3 Training

Da die nötigen HMM-Parameter zur oben beschriebenen Bestimmung der Produktionswahrscheinlichkeiten unbekannt sind, müssen sie trainiert werden. Dabei stellt sich das Problem, dass natürlich auch im Trainingsdatensatz die wahren Zustände s_t verborgen sind. Ebenso unbekannt sind die Verteilungsparameter der bedingten Verteilungen $P_{q_i}(X_t|S_t = q_i, \lambda)$ gegeben den Zustand q_i. Ein iteratives Verfahren – der sogenannte *Baum-Welch Algorithmus* (vgl. Young u. a., 2005, S. 6 ff) – wird verwendet, um abwechselnd Verteilungsparameter und Zustandszuordnung neu zu schätzen im Sinne der Maximierung der Likelihood-Funktion bei gegebenen Daten. Sei zunächst wieder der Fall von Einzelworterkennung und der Einfachheit halber die Parameterschätzung eines einzelnen HMMs λ aus einer einzelnen beobachteten Merkmalssequenz betrachtet: Ausgehend von einer *Initialisierung* der Zustandszuordnungen $\left(\hat{S}_t\right)_{t=1,\ldots,T}$, beispielsweise durch eine äquidistante Aufteilung der Beobachtungsvektoren in die natürliche Abfolge der Zustände, lässt sich eine erste Parameterschätzung vornehmen. Im einfa-

chen Fall einer zugrundegelegten Normalverteilungsannahme ergeben sich

$$\hat{\mu}_i^{(\lambda)} = \frac{1}{\sum_{t=1}^{T} w_i^{(\lambda)}(t)} \sum_{t=1}^{T} X_t * w_i^{(\lambda)}(t) \tag{4.18}$$

und

$$\hat{\Sigma}_i^{(\lambda)} = \frac{1}{\sum_{t=1}^{T} w_i^{(\lambda)}(t)} \sum_{t=1}^{T} w_i^{(\lambda)}(t) \cdot (X_t - \hat{\mu}_i^{(\lambda)})(X_t - \hat{\mu}_i^{(\lambda)})', \tag{4.19}$$

wobei $w_i^{(\lambda)}(t)$ Gewichte im Sinne von Indikatorfunktionen repräsentieren die 1 werden, wenn dem Zeitpunkt t der Zustand $q_i^{(\lambda)}$ zugeordnet wurde und 0 sonst.
Für die geschätzten Zustandübergangswahrscheinlichkeiten ergibt sich

$$\hat{a}_{ij}^{(\lambda)} = \frac{\sum_{t=1}^{T-1} w_{ij}^{(\lambda)}(t)}{\sum_{t=1}^{T-1} w_i^{(\lambda)}(t)} \tag{4.20}$$

mit $w_{ij}^{(\lambda)}(t) = 1$ wenn dem Zeitpunkt $S_t = q_i^{(\lambda)}$ und $S_{t+1} = q_j^{(\lambda)}$ zugeordnet wurden und 0 sonst.
Auf Basis der derart geschätzten Verteilungen lassen sich die Gewichte $w_i^{(\lambda)}(t)$ als Zustands-Zugehörigkeitswahrscheinlichkeiten neu schätzen. Es gilt:

$$w_i^{(\lambda)}(t) = \hat{P}(S_t = q_i^{(\lambda)}|X, \lambda) = \frac{\hat{P}(X, S_t = q_i^{(\lambda)}|\lambda)}{\hat{P}(X|\lambda)}. \tag{4.21}$$

Während der Nenner der auf Basis des Forward-Algorithmus und unter Ausnutzung der geschätzten Verteilungen $\hat{P}_i(X_t|\lambda)$ und $\hat{a}_{ij}^{(\lambda)}$ berechneten Produktwahrscheinlichkeit entspricht, ist zur Berechnung des Zählers ein weiterer Term erforderlich: Analog zu den Vorwärtsvariablen $\alpha_t(q_i^{(\lambda)})$ werden *Rückwärtsvariablen*

$$\beta_t(q_i^{(\lambda)}) := P(X_{t+1}, \ldots, X_T|S_t = q_i^{(\lambda)}, \lambda) \tag{4.22}$$

definiert, die die Wahrscheinlichkeit beschreiben, die restlichen Merkmalsvektoren zu beobachten (Young u. a., 2005, S. 8), bei gegebenem Zustand $S_t = q_j^{(\lambda)}$ zum Zeitpunkt t. Der Zähler aus Gleichung 4.21 ergibt sich durch

$$\hat{P}(X, S_t = q_i^{(\lambda)}|\lambda) = \hat{\alpha}_t(q_i^{(\lambda)})\hat{\beta}_t(q_i^{(\lambda)}). \tag{4.23}$$

Für die $w_{ij}^{(\lambda)}(t)$ zur Neuschätzung der Zustandsübergangswahrscheinlichkeiten nach Gleichung 4.20 ergibt sich mit Hilfe des Vorwärts-rückwärts Algorithmus

$$\begin{aligned} w_{ij}^{(\lambda)} &= \hat{P}(S_t = q_i^{(\lambda)}, S_{t+1} = q_j^{(\lambda)} | X, \lambda) \\ &= \frac{\hat{P}(X, S_t = q_i^{(\lambda)}, S_{t+1} = q_j^{(\lambda)} | \lambda)}{\hat{P}(X|\lambda)} \\ &= \frac{\hat{\alpha}_t(q_i^{(\lambda)}) \hat{P}_j(X_{t+1}|\lambda) \hat{\beta}_t(q_{j+1}^{(\lambda)})}{\hat{P}(X|\lambda)}. \end{aligned} \tag{4.24}$$

Wichtig ist hierbei: Die Topologie des HMMs (d.h. die Anzahl möglicher Zustandsübergänge) lässt sich durch die Initialisierung von A festgelegen. Beispielsweise determiniert eine äquidistante initiale Aufteilung der Merkmalssequenz in die Zustandsfolge $q_1, q_2, \ldots, q_M$, dass $a_{ij} = 0, \forall\, j \neq i+1$ und damit die häufig verwendete, sogenannte links-rechts Topologie (vgl. Schukat-Talamazzini, 1995, Kapitel 5), die nur Übergänge $q_i \rightarrow q_{i+1}$ gestattet.

Eine derartige abwechselnde iterative Schätzung von Zustandszugehörigkeitswahrscheinlichkeiten und Verteilungsparametern analog zum *EM-Algorithmus* (Dempster u. a., 1977) wird als *Baum-Welch Algorithmus* bezeichnet und führt zu einer monotonen Maximierung der Likelihood-Funktion die durch die Produktionswahrscheinlichkeit $P(X|\lambda)$ gegeben ist (Baum u. a., 1970). In der Regel werden jedoch nicht einzelne Normalverteilungen, sondern Mischverteilungen der Zustände angenommen. Für den Algorithmus besitzt diese Erweiterung keinerlei Auswirkungen, lediglich die zu schätzenden Verteilungsparameter werden zu $\mu_{hi}^{(\lambda)}, \Sigma_{hi}^{(\lambda)}$ und $w_{hi}^{(\lambda)}(t)$, wobei der zusätzliche Index h die Mischverteilungskomponente des jeweiligen Zustands $q_i^{(\lambda)}$ bezeichnet. Für den Fall kontinuierlicher Sprache ergeben sich durch die Einführung der zusätzlichen Zustände $q_0^{(\lambda)}$ und $q_{M+1}^{(\lambda)}$ (vgl. Abschnitt 4.2.2) zudem einige geringfügige Modifikationen des oben beschriebenen Vorwärts-rückwärts Algorithmus. Diese sind in Young u. a. (2005) auf Seite 128 beschrieben.

In der Regel erfolgt die HMM-Parameterschätzung natürlich nicht auf Basis einer einzelnen Observationsfolge, sondern auf Basis sämtlicher Beobachtungssequenzen, die im Trainingsdatensatz zur Verfügung stehen. Die Parameterschätzungen in Gleichung 4.18, 4.19 und 4.20 ergeben sich genau wie oben beschrieben, jedoch durch Summation über alle Beoachtungssequenzen (in Zähler und Nenner) (vgl. Schukat-Talamazzini, 1995, S. 146 f).

Zur Schätzung der HMM-Parameter ist man mit dem Problem einer Initialisierung konfrontiert. Eine einfache (und weitverbreitete) Methode hierzu ist eine lineare, äquidistante Unterteilung der Sprachsequenz in die verschiedenen Modellzustände (siehe Young u. a., 2005, S. 7).

4.3 Sprachmodellierung: n-Gramme

Während das Ziel der akustischen Modellierung darin besteht, Wahrscheinlichkeiten für das Auftreten einer beobachteten Merkmalssequenz bei gegebener Wortfolge w zu bestimmen, geht es in der Sprachmodellierung darum, die a priori Auftretenswahrscheinlichkeit einer bestimmten Wortfolge zu quantifizieren.
Sei $w = (w_1, \ldots, w_N)$, die einer Äußerung zugrunde liegende Folge von HMMs, wobei $w_i, i = 1, \ldots, N$ je ein HMM bezeichnet, nach deren zeitlicher Abfolge geordnet. Die Wahrscheinlichkeit für das Vorliegen der HMM-Folge w ist nach der Bayes-Regel gegeben durch

$$P(w) = P(w_N|w_1, \ldots, w_{N-1})P(w_{N-1}|w_1, \cdots, w_{N-2}) \ldots P(w_2|w_1). \tag{4.25}$$

Um die Anzahl möglicher Wortkombinationen einzuschränken, die exponentiell in der Folgenlänge N wächst, beschränkt man sich in der Praxis in der Regel bei den Faktoren auf der rechten Seite von Gleichung 4.25 auf HMM-Folgen der Länge n, sogenannte *n-Gramme*. Im Fall von *Bigrammen*, also $n = 2$, bedeutet dies:

$$P(w) = P(w_N|w_{N-1})P(w_{N-1}|w_{N-2}) \ldots P(w_2|w_1). \tag{4.26}$$

Für Bigramme müssen somit „lediglich“ die Wahrscheinlichkeiten aller HMM-Übergänge $P(w_i = \lambda_k|w_{i-1} = \lambda_j)$ geschätzt werden. Im einfachsten Fall von $n = 1$ ist sogar nur eine Schätzung von Komponenten $P(\lambda_k)$ nötig.
Die Schätzung gestaltet sich intuitiv durch Ermittlung der relativen Häufigkeiten des Vorkommens der verschiedenen HMM-Übergänge.
Bereits bei Bigrammen werden für gewöhnlich in den Trainingsdaten nicht alle möglichen HMM-Paar-Abfolgen realisiert. Dies zieht die unerwünschte Konsequenz einer Nullwahrscheinlichkeit nach sich und macht damit die Erkennung der entsprechenden Sequenz unmöglich. Ein Ausweg besteht in der Umverteilung der Wahrscheinlichkeitsmasse zwischen den verschiedenen Bigrammen durch sogenanntes *Discounting*, wie sie

auf Katz (1987) zurückgeht. Eine Möglichkeit dies zu tun (Young u. a., 2005, S. 267 f) besteht im *absoluten* Discounting, wo von den Auftretenshäufigkeiten der HMM-Übergänge $\lambda_j \rightarrow \lambda_k$ jeweils ein fester Wert d subtrahiert wird, so dass gilt

$$P(w_i = \lambda_k | w_{i-1} = \lambda_j) := \begin{cases} \frac{N_{\lambda_k|\lambda_j} - d}{N_{\lambda_j}}, & N_{\lambda_k|\lambda_j} \geq \theta \\ b(\lambda_j) P(\lambda_k), & sonst. \end{cases} \tag{4.27}$$

Dabei wird die durch Subtraktion von d freigegebene Wahrscheinlichkeitsmasse proportional zur Häufigkeit des Auftretens der HMMs allein aufgeteilt:

$$P(\lambda_k) = \frac{\max(N_{\lambda_k}, \xi)}{\sum_j \max(N_{\lambda_j}, \xi)}. \tag{4.28}$$

Die

$$b(\lambda_j) := \frac{\sum_{j,k} P(w_i = \lambda_k | w_{i-1} = \lambda_j) \cdot I_{[\theta,\infty]}(N_{\lambda_k|\lambda_j})}{\sum_j P(\lambda_j)} \tag{4.29}$$

ergeben sich zur Gewährleistung von $\sum_k P(w_i = \lambda_k | w_{i-1} = \lambda_j) = 1$. θ und ξ stellen minimale Schwellenwerte dar, unterhalb derer die aus der Häufigkeit geschätzten $P(w_i = \lambda_k | w_{i-1} = \lambda_j)$ nach oben korrigiert werden. $N_{\lambda_k|\lambda_j}$ bzw. N_{λ_k} bezeichnen die Häufigkeiten des Vorkommens der entsprechenden HMM-Kombination in den Trainingsdaten. $I_{[.]}(.)$ bezeichnet die Indikatorfunktion. Für die Parameter d, θ und ξ werden Standardwerte verwendet (siehe Young u. a., 2005, S. 267).

4.4 Implementierung

Zur Umsetzung des Back Ends wird das Hidden Markov Toolkit (HTK, Young u. a., 2005) verwendet. Phoneme bilden die verschiedenen HMMs, bestehend aus 3(+2) Zuständen je Phonem (vgl. z.B. Gold und Morgan, 2000, S. 365). Diese lassen sich durch zwei Übergangsphasen um das Phonemzentrum herum motivieren.
Analog zu Lee und Hon (1989) wurden die ursprünglich 61 Phoneme der TIMIT-Transkription (Garofolo u. a., 1993, vgl. Anhang C.1) zur Erkennung zu 39 Klassen nahezu gleich klingender Phoneme zusammengefasst. In einer Studie von Szepannek u. a. (2008) zur Vokalklassifikation mit lokalen Modellen konnte die vorgeschlagene Gruppierung für den Fall von Vokalen bestätigt werden. Als Architektur sämtlicher HMMs werden links-rechts Modelle gewählt (vgl. Gold und Morgan, 2000, S. 345), d.h. nur Übergänge zwischen unmittelbar „benachbarten“ Zuständen sind möglich.

Wie in den meisten Implementierungen zur Spracherkennung werden Normalverteilungen der Zustände mit diagonalen Kovarianzmatrizen modelliert (vgl. z.B. Gales, 1998; Young u. a., 2005, S. 95 f). Um eine Verletzung der damit verbundenen Annahme unkorrelierter Merkmale zu verhindern, werden im folgenden Kapitel 5 Methoden vorgestellt, die neben einer Dimensionsreduktion auch eine Dekorrelation der Merkmale bewirken.

Zum Parametertraining kommt der *Baum-Welch Algorithmus* zum Einsatz (siehe Abschnitt 4.2.3). Ausgehend von normalverteilten Merkmalen, gegeben den Zustand, werden wie in Young (1992) nach je drei Iterationen über den kompletten Trainigsdatensatz aus den Normalverteilungen Mischverteilungen gebildet. Gouws u. a. (2004) variieren für MFCCs die Anzahl verwendeter Mischkomponenten und erzielen die besten Resultate bei Verwendung von etwa sechs bis acht verwendete Komponenten je TIMIT-Phonem. Aus diesem Grund erfolgen sukzessive drei Verdopplungen der Anzahl an Mischkomponenten. Hierzu wird die von Young u. a. (2005) vorschlagene Heuristik verwendet: Die Mischverteilungskomponente mit dem jeweils größten Gewicht wird in zwei neue, normalverteilte Komponenten aufgeteilt, in dem deren Zugehörigkeitsgewichte als Gewichte der neuen Komponenten halbiert werden. Die Erwartungswerte der beiden neuen Komponenten bestimmen sich zunächst in jeder Variable aus dem geschätzten Parameter der ausgänglichen Normalverteilung gemäß der aktuell vorliegenden Iteration $\pm$ dem 0.2 fachen ihrer Standardabweichung. Abschließend werden sieben Iterationsdurchläufe des Baum-Welch-Algorithmus zur Stabilisierung der Parameterschätzungen durchgeführt.

Die Initialisierung der HMMs erfolgt über eine lineare, uniforme Unterteilung der Merkmalsvektorsequenzen des Trainingsmaterials entsprechend der zugehörigen Transkription in die Phonemklassen. Als Sprachmodell fungieren Bigramme (siehe Abschnitt 4.3) – zur Erkennung wird der *Viterbi Algorithmus* verwendet (siehe Abschnitt 4.2.2).

Das verwendete Back End entspricht somit einer Standardimplementierung. Experimente mit Verwendung generalisierter Triphone (vgl. Schukat-Talamazzini, 1995, Kapitel 6.3) oder eine Änderung der Anzahl verwendeter Mischkomponenten führten zu keiner Verbesserung der Ergebnisse. Während sich dies im ersten Fall vermutlich in der Menge vorhandenen Trainingsdatenmaterials begründet, ist dies auch für den zwei-

ten Fall naheliegend: Die durchschnittliche Anzahl an Mischverteilungskomponenten lässt sich als Anzahl latenter Sub-Ausprägungen eines Phonems intepretieren, die eine unterschiedliche Verteilung besitzen. Dies wird vermutlich für unterschiedliche Merkmalssätze gelten, damit also für die aus dem auditorischen Simulationsmodell extrahierten Repräsentationen ebenso wie für von Gouws u. a. (2004) untersuchten MFCCs. Eine weitere Optimierung der Back Ends – beispielsweise durch *Parametertying*[2] wie dies z.B. in Munich und Lin (2005) erfolgt – stellt nicht den Fokus dieser Arbeit dar, der auf dem Performancevergleich der verschiedenen Merkmalsrepräsentationen liegt, und wird von daher nicht vorgenommen.

[2] Als „Tying“ oder „Verklebung“ werden verschiedene Methoden bezeichnet, die durch geschicktes Zusammenfassen von Modellkomponenten mit ähnlichen Parametern eine stabilere Parameterschätzung ermöglichen (vgl. z.B. Schukat-Talamazzini, 1995, S. 147 ff).

5 Dimensionsreduktion

5.1 Motivation

Die in Kapitel 3 beschriebenen Merkmalsvektoren sind zum Teil sehr hoher Dimension, da das Simulationsmodell aus 251 Nervenfasern besteht und auch die modellierten Delay-Computing Netzwerke mehr als 100 Output-Neuronen besitzen. Für gängige, MFCC-basierte Spracherkennungsmodelle (vgl. Abschnitt 3.3) wird häufig eine Merkmalsvektordimension von 13 verwendet zuzüglich der ersten und zweiten (Zeit)differenzen dieser Vektoren, insgesamt also ein 39 dimensionaler Merkmalsvektor. Auch Schafföner u. a. (2003) erzielen nach Dimensionsreduktion optimale Ergebnisse für Merkmalsvektoren mit einer Dimension von etwa 35-40. Nimmt man diese Werte als ersten Anhaltspunkt auch für die auf dem auditorischen Simulationsmodell basierenden Merkmale dieser Arbeit, so erscheint eine Dimensionsreduktion erstrebenswert.

Eine einfache und leicht interpretierbare Methode, die Dimension eines Merkmalsvektors zu reduzieren, besteht in einer Variablenselektion. Szepannek und Weihs (2006b) zeigen im Bereich der Klassifikation, dass zum Teil deutliche Verbesserungen der Fehlerrate erzielt werden können, wenn die Variablenselektion klassenspezifisch für je Paare von zwei Klassen erfolgt. Dieser Ansatz ist jedoch sehr rechenzeitintensiv – insbesondere für das in dieser Anwendung vorliegende Erkennungsproblem einer hohen Anzahl von Phonemen (siehe Abschnitt 6.2.1) – und zudem nicht unmittelbar auf die Situation einer Klassifikation von Merkmals-Sequenzen (vgl. Abschnitt 4) übertragbar. Aus

diesem Grund stellt er für die Anwendung dieser Arbeit keine Alternative dar.

Nach einem Übersichtsartikel zur Variablenselektion von Guyon und Elisseeff (2003) lassen sich Variablenselektionsverfahren in zwei Gruppen unterteilen: in sogenannte *Wrapper-Variablenselektionsverfahren* einerseits und *korrelationsbasierte Variablenselektionsverfahren* andererseits.

Wrapper-Verfahren zeichnen sich dadurch aus, dass die Suche nach einer Untermenge an Variablen durch empirische Optimierung eines explizit definierten Zielkriteriums – im Fall dieser Arbeit beispielsweise der Erkennungsrate – erfolgt, durch wiederholtes Aufrufen des Verfahrens auf unterschiedlichen Variablenuntermengen. Ein Beispiel eines derartigen Variablenselektionsverfahrens im Bereich der Klassifikation stellt der *stepclass* Algorithmus (Weihs u. a., 2005) dar. Aufgrund der häufigen Wiederholung der vollständigen Prozedur im Rahmen der Variablenauswahl sind Wrapper Verfahren sehr Rechenzeit-aufwändig und stellen im Fall der vorliegenden Anwendung somit ebenfalls keine Option dar.

Im Gegensatz dazu sind korrelationsbasierte Verfahren unabhängig von der verwendeten Modellbildungsmethode: Es wird eine Untermenge von Variablen bestimmt, die in einem größtmöglichen Zusammenhang zur vorherzusagenden Variable stehen (im Fall dieser Arbeit sind dies die verschiedenen HMM-Zustände, die während des Durchlaufs der unterschiedlichen Phoneme vorliegen), jedoch untereinander möglichst wenig korreliert sind. Je nach Verfahren variieren die Kriterien zur Quantifizierung des Zusammenhangs zwischen den Merkmalen. Ein Beispiel einer derartigen Variablenselektionsmethode im Bereich der Klassifikation stellt die *Correlation Based Feature Subset Selection (CFS)* dar (vgl. z.B. Szepannek u. a., 2003).

Neben dem Ziel einer niedrigdimensionaleren Merkmalsrepräsentation ergibt sich ein weiterer Aspekt, der bei der Wahl der Dimensionsreduktionsmethode zu berücksichtigen ist: In der Verwendung von Hidden-Markov Modellen zur automatischen Spracherkennung, hat es sich etabliert, diagonale Kovarianzmatrizen der Klassen zu verwenden (Gales, 1998, siehe auch Abschnitt 4.4). Diese Annahme für die Modellierung hat eine enorme Reduktion der zu schätzenden Modellparameter zur Folge. Andererseits erfordert dies, dass die verwendeten Merkmale möglichst unkorreliert sein sollten. Für die verwendete Methode zur Reduktion der Merkmalsdimension ergibt sich somit als zusätzliche Anforderung eine Dekorrelation der extrahierten Merkmale.

Korrelationen der Merkmale sind jedoch implizit gegeben und durch die Neurophysiologie des Ohrs bedingt. Die auditorischen Nerven benachbarter ANFs weisen eine ähnliche Reaktion auf, sowohl hinsichtlich der Durchschnittsfeuerrate als auch hinsichlich der zeitlichen Struktur der Antwort. Dies ist z.B. zu erkennen im Beispiel in Abb. 2.8 auf S. 15. Auch in Merkmalsrepräsentationen, bei denen die Merkmale mit einer spezifischen Frequenz assoziierbar sind, ist eine höhere Korrelation für Merkmale ähnlicher Frequenzen zu erwarten. Diese Überlegungen spiegeln sich im Beispiel einer Darstellung der (absoluten) Korrelationen der Merkmalsvektoren X^{OD_1} wider (Abb. 5.1). Es gilt näherungsweise, dass für festes l $cor(X_l^{OD_1}, X_j^{OD_1})$ monoton fallend ist in $|l - j|$. Für Delay-Computing basierte Merkmale gilt das Gleiche: Benachbar-

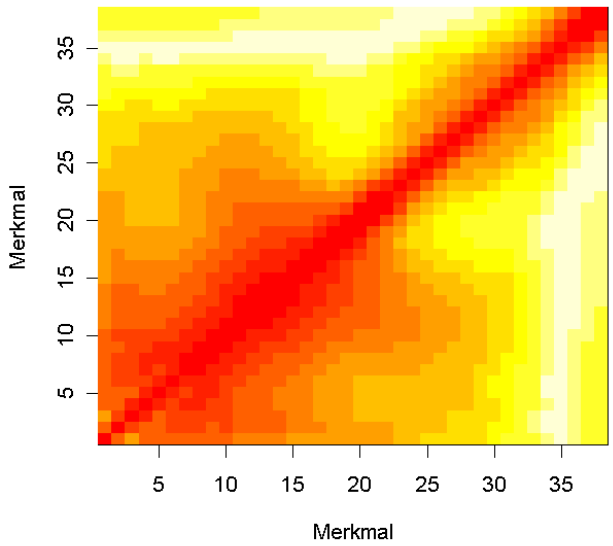


Abbildung 5.1: Absolute Korrelationen zwischen den Merkmalen $X_k^{OD_1}$ (rot:1, weiß:0).

te Output-Neuronen repräsentieren sehr ähnliche Verläufe der Cochlea-Wanderwelle, und große Teile ihrer Inputwerte stimmen überein. Die Ausprägungen benachbarter Output-Neuronen weisen damit auch per Konstruktion Korrelationen auf.

Variablenselektionsverfahren stellen einen Spezialfall einer *linearen Dimensionsreduk-*

tion der Gestalt

$$Y_l = \sum_{j=1}^{p} a_{jl} X_j \tag{5.1}$$

der Originalmerkmale $(X_j)_{(j=1,\ldots,p)}$ dar, mit $a_{jl} \in \{0,1\}$, $\sum_j a_{jl} = 1$, $\forall l$ und $\sum_l a_{jl} \leq 1$, $\forall j$.

Im Rest des Kapitels werden Verfahren zur Berechnung von Koeffizienten a_{jk} vorgestellt, die unkorrelierte Merkmale erzeugen und dabei ein größtmögliches Maß an Information für die weitere Verwendung zur Spracherkennung erhalten. Dabei bezeichnet $\mathbf{X}$ die Datenmatrix mit den Merkmalsvektoren $\mathbf{x_j}$, $j = 1, \ldots, p$, in den Spalten. Die Zeilen von $\mathbf{X}$ sind alle beobachteten Ausprägungen der Merkmale, die aus Anwendung eines der in Kapitel 3 beschriebenen Merkmalsextraktionsprinzipien auf eine Menge an Sprachäußerungen hervorgegangen sind. Die aus der Transformation resultierenden Merkmale ergeben sich als Spalten von $\mathbf{Y} = \mathbf{XA}$.

Abschnitt 5.2 beschreibt zunächst Standardverfahren, die eine „unüberwachte“ Dimensionsreduktion durchführen. Abschnitt 5.3 stellt alternativ verschiedene Methoden der Diskriminanzanalyse vor. Dies sind Fishers lineare Diskriminanzanalyse (Fisher, 1936), heteroskedastische Diskriminanzanalyse (Kumar und Andreou, 1998), sowie das Prinzip Friedmans regularisierter Diskriminanzanalyse (Friedman, 1989), wobei letzteres als Anwendung zum Zwecke linearer Dimensionsreduktion in dieser Arbeit erstmalig Verwendung auf dem Gebiet der automatischen Spracherkennung findet.

5.2 Unüberwachte Dimensionsreduktion

Als Standardverfahren der unüberwachten Dimensionsreduktion ist die *Hauptkomponentenanalyse* (engl.: *Principal Component Analysis*, PCA) bekannt. „Unüberwacht" bedeutet hierbei, dass die Dimensionsreduktion aufgrund der Beobachtungen selbst, ohne Hinzunahme zusätzlicher, externer Information erfolgt.
Die Idee der Hauptkomponentenanalyse besteht in der Identifikation einer Anzahl von $q < p$ unkorrelierten Komponenten mit maximaler Streuung, wobei nicht zu unterscheiden ist, ob die Streuung auf enthaltene Information oder auf Rauschen zurückzuführen ist. Es wird davon ausgegangen, dass mit einem Maximum an erhaltener Streuung der Originaldaten auch möglichst viel der vorhandenen Information in der niedrigdimensionalen Projektion enthalten bleibt.
Die standardisierten Beobachtungen des Merkmals X_j seien bezeichnet durch

$$z_{nj} = \frac{x_{nj} - \bar{x}_j}{\sqrt{N-1}\, s_j}, \tag{5.2}$$

mit n als Beobachtungsindex, N als Gesamtanzahl an Beobachtungen, $\bar{x}_j$ dem Mittelwert, $s_j^2 = \frac{\sum_{n=1}^{N}(x_{nj}-\bar{x}_j)^2}{N-1}$ der empirischen Varianz des Merkmals und x_{nj} der Ausprägung der Beobachtung n in Merkmal X_j. Standardisieren ist sinnvoll, da andernfalls die Streuung der Daten von denjenigen Merkmalen mit großer Varianz dominiert wird.
Sei weiter die *standardisierte Beobachtungsmatrix*

$$\mathbf{Z} = (z_{nj}). \tag{5.3}$$

Für die *empirische Korrelationsmatrix* R gilt

$$\mathbf{R} = \mathbf{Z}'\mathbf{Z}. \tag{5.4}$$

Gesucht sind die Spaltenvektoren $\mathbf{a_l} \in \mathbb{R}^p$ der *Ladungsmatrix* $\mathbf{A} \in \mathbb{R}^{p \times q}$, so dass zunächst für die erste Hauptkomponente $\mathbf{y_1}$ gilt:

$$Var(\mathbf{y_1}) = \mathbf{y_1'y_1} = (\mathbf{Za_1})'(\mathbf{Za_1}) = \mathbf{a_1'Z'Za_1} = \mathbf{a_1'Ra_1} \tag{5.5}$$

ist maximal. Für die weiteren Hauptkomponenten $\mathbf{y_2}, \ldots, \mathbf{y_q}$ soll jeweils gelten, dass diese innerhalb desjenigen orthogonalen Raums auf den zuvor identifizierten ersten Hauptkomponenten die maximale Streuung der Daten beinhalten. Diese Forderung

wird erfüllt von den (nach Größe der Eigenwerte geordneten) Eigenvektoren der Matrix **R** (vgl. z.B. Fahrmeir u. a., 1996, S. 597 f).

Eine andere Methode der unüberwachten Dimensionsreduktion, die für Merkmale des logarithmierten Spektrums beschrieben ist (Davis und Mermelstein, 1980), stellt die *cepstrale Transformation* dar (siehe Abschnitt 3.3), wie sie zur Berechnung der MFCCs verwendet wird. Diese erfolgt durch eine inverse diskrete Cosinus Transformation (IDCT) der spektralen Merkmale. Merhav und Lee (1993) zeigen asymptotische Unkorreliertheit der cepstralen Koeffizienten für eine Beobachtungssequenz unendlicher Länge, wenn der datengenerierende Prozess ein *stationärer Gaußprozess* oder ein *autoregressiver Prozess* ist.

Einige der in Kapitel 3 vorgeschlagenen, auditorisch basierten Merkmalssätze lassen sich durch implizit mit den Merkmalen assoziierbare Frequenzen, im Sinne von Spektralamplituden interpretieren, wobei Perdigao und Sa (1998) den nichtlinearsaturierenden Effekt an den Stereozilien (vgl. Abschnitt A.3) mit der vor der Cepstraltransformation durchgeführten Logarithmierung vergleichen. Für diese, aus dem auditorischen Simulationsmodell extrahierten Merkmalsvektoren lässt somit ebenfalls die Anwendung einer IDCT zur Dimensionsreduktion motivieren. Aufgrund der hohen Popularität von (auf Originalzeitreihen) berechneten MFCCs wird eine solche Transformation in *auditorische Frequenz-Cepstralkoeffizienten* (AFCCs) auf Basis der verschiedenen Repräsentationen durchgeführt. Dies gilt allerdings nicht für die Delay-Computing basierten Merkmale, die keine Interpretation im Spektralbereich besitzen, sondern Trajektorienverläufe der Cochleawanderwelle repräsentieren.

Alternativ zur cepstralen Transformation wurde von Nadeu und Gorricho (1995) die Methoden der *Frequenz-Filterung* (engl. *frequency filtering*) vorgestellt. Dort werden Merkmale der Form

$$X_k(t) = \sum_{k^*} h(k - k^*) X_{k^*}(t) \tag{5.6}$$

gebildet, wobei $X_k(t)$ die Originalmerkmale im Frequenzbereich (in der Originalanwendung sind dies logarithmierte Filterbank-Energien, wie sie auch zur Berechnung von MFCCs verwendet werden, vgl. Abschnitt 3.3) darstellen. Gute Ergebnisse unter unterschiedlichen Bedingungen (vgl. Jung, 2004; Nadeu und Gorricho, 1995) zeigt ein

Frequenz-Filter mit der Impulsantwort

$$h(k-k^*) = \begin{cases} 1, & (k-k^*) = 0, \\ -1, & (k-k^*) = -2, \\ 0, & sonst. \end{cases} \tag{5.7}$$

Die resultierenden Merkmale besitzen ähnlich hohe Varianzen. Im Gegensatz dazu sind diese für MFCCs sehr unterschiedlich, so weisen die niedrigen Koeffizienten eine deutlich höhere Streuung auf. Diese Eigenschaft ist bei der Verwendung von HMMs als Back Ends jedoch nicht von Bedeutung. Die entstehenden Merkmale sind im Vergleich zu den Originalmerkmalen (vgl. Giron, 2006) weniger stark korreliert. Zudem sind sie – im Gegensatz zu MFCCs – weiterhin im Spektralbereich interpretierbar. Die Merkmale besitzen jedoch den Nachteil die ursprüngliche Merkmalsdimension p auf exakt $q = p - 2$ zu verändern. Dies entspricht nicht dem Ziel einer Dimensionsreduktion. Von daher findet eine derartige Merkmalstransformation in dieser Arbeit keine Verwendung, sondern soll hier nur der Vollständigkeit halber erwähnt werden.

5.3 Diskriminanzanalyse

5.3.1 Lineare Diskriminanzanalyse

Einen Nachteil der unüberwachten Dimensionsreduktion stellt die Tatsache dar, dass die Bestimmung der Koeffizienten der Transformationsmatrix $\mathbf{A}$ nicht explizit an der zu lösenden Vorhersageaufgabe ausgerichtet ist. Bei der Hauptkomponentenanalyse beispielsweise besteht das Ziel darin, möglichst viel der Streuung im niedrigdimensionalen Raum zu erhalten, unabhängig davon, ob diese Streuung Information zur Vorhersage beinhaltet oder lediglich durch Rauschen verursacht ist. Aus diesem Grund hat sich in der automatischen Spracherkennung neben cepstraler Transformation insbesondere die auf Fisher (1936) zurückgehende lineare Diskriminanzanalyse (LDA) etabliert (vgl. z.B. Brown, 1987; Eisele u. a., 1996).

Zur Berechnung der *Diskriminanzkomponenten* ist neben der Matrix $\mathbf{X}$ der Merkmalsvektoren noch ein zusätzlicher Vektor $\mathbf{c} \in \{1, \ldots, K\}^N$ mit Klassenzugehörigkeiten (Supervisor) erforderlich. Im Folgenden bezeichne $k \in \{1, \ldots, K\}$ eine Klasse und c_n diejenige Klasse, der Beobachtung n angehört. Die Idee der linearen Diskriminanzanalyse besteht zunächst in der Identifikation derjenigen Diskriminanzkomponente $\mathbf{y_1} = \mathbf{X a_1}$, in der sich die Klassenmittelwerte $\bar{y}_{1,k}$ durchschnittlich am stärksten unterscheiden, unter Annahme gleicher Kovarianzmatrizen aller Klassen. Als intuitives, zu maximierendes Kriterium fungiert das Verhältnis der quadrierten Abstände der Klassenmittel vom Gesamtmittel der transformierten Merkmale, gewichtet mit der Klassenhäufigkeit und skaliert mit der Varianz der projizierten Daten:

$$\frac{\sum_{k=1}^{K} N_k (\bar{y}_{1,k} - \bar{y}_1)^2}{\sum_{k=1}^{K} s_{1,k}^2}, \tag{5.8}$$

mit K der Anzahl an Klassen, N_k der Klassenhäufigkeit, $s_{1,k}^2 = \sum_{n=1}^{N} I_{\{k\}}(c_n)(y_{1,n} - \bar{y}_{1,k})^2$ der Varianz von Klasse k, $\bar{y}_{1,k}$ den Klassenmittelwerten, c_n der Klasse, der Beobachtung n angehört und $I_{\{.\}}(.)$ der Indikatorfunktion.

Es ergibt sich äquivalent die Suche nach dem Vektor $\mathbf{a_1}$, für den gilt:

$$\frac{\mathbf{a_1' B a_1}}{\mathbf{a_1' W a_1}} \tag{5.9}$$

ist maximal (vgl. Fahrmeir u. a., 1996, S. 324), mit

$$\mathbf{B} = \frac{1}{N}\sum_{k=1}^{K} N_k(\bar{\mathbf{x}}_\mathbf{k} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\mathbf{k} - \bar{\mathbf{x}})' \tag{5.10}$$

der *Streuungsmatrix zwischen den Klassen* der gewichteten Abweichungsprodukte der Klassenmittel vom allgemeinen Mittel und

$$\mathbf{W} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \bar{\mathbf{x}}_{\mathbf{c_n}})(\mathbf{x_n} - \bar{\mathbf{x}}_{\mathbf{c_n}})', \tag{5.11}$$

der sogenannten *gepoolten Kovarianzmatrix* aller Klassen. Hierbei bezeichnen $\bar{\mathbf{x}}$ den (Spalten-)Vektor der arithmetischen Mittel aller Merkmale, $\bar{\mathbf{x}}_\mathbf{k}$ die klassenweisen Mittelwertsvektoren und $\mathbf{x_n}$ die Spaltenvektoren der Beobachtung n in allen p Merkmalen. Die Lösung des Maximierungsproblems ist nicht eindeutig, da für Vielfache $\mathbf{b} = \mathbf{ca}$, $\mathbf{c} \in \mathbb{R} \setminus \mathbf{0}$ gilt: $\mathbf{b'Bb/b'Wb} = \mathbf{(ca)'B(ca)/(ca)'W(ca)} = \mathbf{(c^2a'Ba)/(c^2a'Wa)} = \mathbf{a'Ba/a'Wa}$. Aus diesem Grund wird in der Regel die zusätzliche Restriktion $\mathbf{a_1'Wa_1} = \mathbf{1}$ getroffen. Diese Anforderung wird genau vom Eigenvektor der Matrix $\mathbf{W^{-1}B}$ mit maximalem Eigenwert erfüllt. Es existieren maximal $K-1$ orthogonaler Eigenvektoren a_l (d.h. mögliche Spalten der Projektionmatrix $\mathbf{A}$, siehe z.B. Weihs und Heilemann, 2000, S. 595) mit zugehörigen Eigenwerten $\lambda_k > 0$.

Eine äquivalente, häufig zu findende Formulierung ist, dass $\mathbf{A}$ die Matrix der Eigenvektoren $\mathbf{W^{-1}T}$ mit $\mathbf{T} = \mathbf{B} + \mathbf{W}$ darstellt: Für eine Matrix $\mathbf{A}$, deren Spalten Eigenvektoren von $\mathbf{W^{-1}T} = \mathbf{W^{-1}(B + W)}$ beinhalten, gilt, wenn $\mathbf{W}$ vollen Rang hat:

$$\begin{aligned} \mathbf{\Lambda A} &= \mathbf{W^{-1}TA} \\ &= \mathbf{W^{-1}BA + W^{-1}WA} \\ &= \mathbf{W^{-1}BA + IA} \end{aligned} \tag{5.12}$$

mit $\mathbf{\Lambda}$ der Diagonalmatrix der Eigenwerte von $\mathbf{W^{-1}T}$. Es gilt damit

$$\mathbf{(\Lambda - I)A = W^{-1}BA}, \tag{5.13}$$

d.h. die Eigenvektoren von $\mathbf{W^{-1}B}$ und $\mathbf{W^{-1}T}$ sind identisch.

Ein weiterer Vorteil der linearen Diskriminanzanalyse im Vergleich zur Hauptkomponentenanalyse wird durch die folgende Betrachtung deutlich: Die mit Hilfe einer

PCA gefundenen Hauptkomponenten sind unkorreliert bezüglich einer gemeinsamen Verteilung aller Daten. Würde es eine solche Verteilung geben, wäre allerdings die Möglichkeit zur Diskrimination nicht mehr gegeben. Die Diskriminanzkomponenten sind unkorelliert bezüglich aller klassenspezifischen Verteilungen, wobei für diese eine identische Kovarianzmatrix angenommen wird.
Zur Verwendung der LDA als Dimensionsreduktionsmethode ist der Vektor **c** der Klassenzugehörigkeiten erforderlich, die durch die gefundene Projektion der Originalmerkmale gut voneinander getrennt werden. Solch eine Klassenzuordnung ist für den Anwendungsfall dieser Arbeit nicht explizit gegeben, durch das gegebene Erkennungsproblem liegt jedoch nahe, die Klassen an der phonetischen Transkription der Merkmalssequenz zu orientieren. Hier bieten sich die verschiedenen zu erkennenden Phoneme an. Genau genommen besteht die Klassifikationsaufgabe der HMMs in der richtigen Zuordnung der einzelnen HMM-Zustände, und diese wiederum ist von einer korrekten Zuordnung in die Mischverteilungskomponenten abhängig. In einer Untersuchung von Duchateau u. a. (2001) auf Basis von Melspektralkoeffizienten kann eine Verbesserung der Erkennungsergebnisse erzielt werden, wenn anstatt der Phoneme die Zustände für die Klassenzuordnungen **c** gewählt werden. Schafföner u. a. (2003) erzielen keine wesentlichen Verbesserungen mehr durch Verwendung der einzelnen Mischkomponenten zur Klassendefinition. Insgesamt ergeben sich so $61 \cdot 3 = 183$ Klassen (vgl. S. 72 bzw. Anhang C.1), wobei anzumerken ist, dass die Zuordnung der Zustände zunächst nicht vorliegt (siehe Abschnitt 5.4). Im Anwendungsfall liegt zudem eine sehr große Menge an Beobachtungsvektoren vor, daher ist eine blockweise Berechnung der Kovarianzmatrizen erforderlich (vgl. Anhang B.9).

5.3.2 Heteroskedastische Diskriminanzanalyse

Eine Restriktion der linearen Diskriminanzanalyse stellt die Annahme gleicher Kovarianzmatrizen der Verteilung der verschiedenen Klassen dar, die natürlich nicht oder nur näherungsweise gegeben sein muss. Marks und Dunn (1974) untersuchen das Verhalten von LDA für unterschiedlich starke Verletzungen dieser Annahme im Bereich der Klassifikation. Aus diesem Grund ist zur Lösung von Klassifikationsproblemen als Erweiterung die *quadratische Diskriminanzanalyse* (QDA) entstanden, deren Annahme in normalverteilten Klassen mit klassenspezifischer Mittelwert- und Kovarianzstruktur

besteht (vgl. z.B. Weihs und Heilemann, 2000, S. 595). Diese liefert zwar eine Klassifikationsregel, nicht jedoch eine lineare Achsentransformation zur Dimensionsreduktion. Zu diesem Zweck wurde von Kumar und Andreou (1998) die *heteroskedastische lineare Diskriminanzanalyse* (HDA) entwickelt, die eine derartige Achsentransformation auf Basis einer Likelihood-Maximierung bestimmt.
Sei für die Verteilung der verschiedenen Klassen eine multivariate Normalverteilung (mit nicht notwendigerweise gleichen Kovarianzmatrizen der verschiedenen Klassen) angenommen. Für die Verteilung der Merkmale der verschiedenen Klassen ergibt sich nach Transformation $\mathbf{y} = \mathbf{A}'\mathbf{x}$, $\mathbf{A} \sim p \times p$ wiederum Normalverteilung. Diese sei für die Klasse k beschrieben durch die Parameter $\mathbf{y}|k \sim \mathrm{N}(\mu_\mathbf{k}, \mathbf{\Sigma_k})$, wobei für die Mittelwerte gelten soll:

$$\mu_\mathbf{k} = \begin{bmatrix} \mu_\mathbf{k}^\mathbf{q} \\ \mu_\mathbf{0}^\mathbf{p-q} \end{bmatrix}. \tag{5.14}$$

Das bedeutet: Die Klassenmittelwerte $\mu_\mathbf{k}$ unterscheiden sich nur in den ersten q Komponenten des transformierten Merkmalsraumes $\mathbf{y}$. Analog soll für die Kovarianzmatrizen $\mathbf{\Sigma_k}$ gelten

$$\mathbf{\Sigma_k} = \begin{pmatrix} \mathbf{\Sigma_k^q} & 0 \\ 0 & \mathbf{\Sigma_0^{p-q}} \end{pmatrix}. \tag{5.15}$$

Zusammen bedeuten (5.14) und (5.15), dass Verteilungen der Klassen sich im transformierten Merkmalsraum nur in den ersten $q < p$ Variablen unterscheiden, sich dort aber sowohl in ihren Mittelwerten, als auch in ihren Kovarianzmatrizen unterscheiden können. Eine zwar nicht erforderliche, aber bisweilen sinnvolle Restriktion stellt hierbei die Annahme einer diagonalen Gestalt von $\mathbf{\Sigma_k}$ dar. Auf diese Weise entsteht ein unkorrelierter Merkmalsvektor $\mathbf{y_1}, \ldots, \mathbf{y_p}$. Eine entsprechende Partitionierung der Ladungsmatrix lautet

$$\mathbf{A} = (\mathbf{a_1}, \ldots, \mathbf{a_q}, \mathbf{a_{q+1}}, \ldots, \mathbf{a_p}) = (\mathbf{A^q}\, \mathbf{A^{p-q}}), \tag{5.16}$$

wobei $\mathbf{A^q}$ diejenigen Spaltenvektoren beinhaltet, die die Merkmalsvektoren $\mathbf{x}$ in den q-dimensionalen Unterraum mit einer unterschiedlichen Verteilung der Klassen projizieren. Die Spaltenvektoren von $\mathbf{A^{p-q}}$ dagegen projizieren die Merkmale in einen Teilraum mit identischer Verteilung der Klassen, identifizieren also die Komponenten von $\mathbf{y}$, die unter den in (5.14) und (5.15) getroffenen Annahmen im Rahmen der Dimensionsreduktion vernachlässigt werden können.

Es lassen sich diejenige Transformation $\mathbf{A}$ und Verteilungsparameter $\mu_\mathbf{k}$, sowie $\mathbf{\Sigma_k}$ herleiten, die die Likelihood der ursprünglichen Beobachtungen $\mathbf{x_n}$ unter der für die Verteilung der $\mathbf{y_n}$ getroffenen Annahme, maximiert, mit $\mathbf{x_n}$ als n-tem Beobachtungsvektor des Trainingsdatensatzes.
Für die Likelihoodfunktion ergibt sich:

$$L(\mu_\mathbf{1}, \ldots, \mu_\mathbf{K}, \mathbf{\Sigma_1}, \ldots, \mathbf{\Sigma_K}, \mathbf{A}, \mathbf{x}) = \prod_{n=1}^{N} \frac{|\mathbf{A}|}{\sqrt{(2\pi)^p |\mathbf{\Sigma_{c_n}}|}} e^{\left(-\frac{1}{2}(\mathbf{A'x_n} - \mu_\mathbf{c_n})' \mathbf{\Sigma_{c_n}^{-1}} (\mathbf{A'x_n} - \mu_\mathbf{c_n})\right)}, \tag{5.17}$$

wobei der $|\cdot|$ Operator die Determinante einer Matrix bezeichnet. Hierzu muss angemerkt werden, dass – wie schon im vorangegangenen Abschnitt beschrieben – die einzelnen Beobachtungen der Sprachsequenzen natürlich nicht voneinander unabhängig sind, die Berechnung der Likelihood über Produktbildung hier somit also strenggenommen nicht angemessen ist. Dieser Sachverhalt findet in der gängigen Literatur der automatischen Spracherkennung zur HDA keine Erwähnung.
Im Fall angenommener Diagonalität der $\mathbf{\Sigma_k} = \mathrm{diag}((\sigma_k^1)^2, \ldots, (\sigma_k^q)^2, (\sigma_0^{q+1})^2, \ldots, (\sigma_0^p)^2)$ ergibt sich für die logarithmierte Likelihood

$$\begin{aligned} l(\mu_\mathbf{1}, \ldots, \mu_\mathbf{K}, \mathbf{\Sigma_1}, \ldots, \mathbf{\Sigma_K}, \mathbf{A}, \mathbf{x}) &= l(\mu_\mathbf{1}, \ldots, \mu_\mathbf{K}, \sigma_1, \ldots, \sigma_K, \mathbf{A}, \mathbf{x}) \\ &= \frac{-Np}{2} \log 2\pi + N \log |A| \\ &\quad - \frac{N}{2} \sum_{d=q+1}^{p} \log \sigma_0^d - \sum_{k=1}^{K} \frac{N_k}{2} \sum_{d=1}^{q} \log \sigma_k^d \\ &\quad - \frac{1}{2} \sum_{k=1}^{K} \sum_{c_n=k} \sum_{d=1}^{q} \frac{(\mathbf{a'_d x_n} - \mu_\mathbf{k}^\mathbf{d})^\mathbf{2}}{\sigma_k^d} \\ &\quad - \frac{1}{2} \sum_{k=1}^{K} \sum_{c_n=k} \sum_{d=q+1}^{p} \frac{(\mathbf{a'_d x_n} - \mu_\mathbf{0}^\mathbf{d})^\mathbf{2}}{\sigma_0^d}. \end{aligned} \tag{5.18}$$

Die Parameter $(\sigma_k^d)^2$ bzw. $(\sigma_0^d)^2$ bezeichnen dabei die Varianzen der Klassen k in Merkmal d. Die Korrelationen zwischen den transformierten Merkmalen betragen gemäß der getroffenen Annahme über die Gestalt von $\mathbf{y}$ null. Die Bezeichnung σ_0^d unterstreicht, dass die verschiedenen Klassen in den Merkmalen $d \geq q+1$ nach Voraussetzung identische Verteilungen besitzen. Analog bezeichnen μ_k^d und μ_0^d die Komponenten der (klassenspezifischen) Erwartungswerte der transformierten Merkmale. a_d bezeichnet den d-ten Spaltenvektor der Transformationsmatrix $\mathbf{A}$.

Durch Nullsetzen der partiellen Ableitungen ergeben sich für die Verteilungsparameter in Abhängigkeit von der Transformationsmatrix $\mathbf{A}$ die Maximum-Likelihood Schätzungen für Erwartungswert und Kovarianzmatrix (Kumar und Andreou, 1996):

$$\hat{\mu}_{\mathbf{k}}^{\mathbf{q}} = (\mathbf{A}^{\mathbf{p}})'\bar{\mathbf{x}}_{\mathbf{k}}, \tag{5.19}$$

$$\hat{\mu}_{\mathbf{0}}^{\mathbf{p-q}} = (\mathbf{A}^{\mathbf{p-q}})'(\bar{\mathbf{x}}), \tag{5.20}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{k}}^{\mathbf{q}} = \mathrm{diag}((\mathbf{A}^{\mathbf{q}})'\mathbf{W}_{\mathbf{k}}(\mathbf{A}^{\mathbf{q}})) \text{ und} \tag{5.21}$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{0}}^{\mathbf{p-q}} = \mathrm{diag}((\mathbf{A}^{\mathbf{p-q}})'(\mathbf{T})(\mathbf{A}^{\mathbf{p-q}})), \tag{5.22}$$

wobei

$$\mathbf{W}_{\mathbf{k}} = \frac{1}{N_k}\sum_{n=1}^{N} I_{\{k\}}(c_n)(\mathbf{x_n} - \bar{\mathbf{x}}_{\mathbf{c_n}})(\mathbf{x_n} - \bar{\mathbf{x}}_{\mathbf{c_n}})'. \tag{5.23}$$

Für $\hat{\mathbf{\Sigma}}_{\mathbf{0}}^{\mathbf{p-q}}$ sind dabei die klassenspezifischen Kovarianzmatrizen $\mathbf{W_k}$ der Originalmerkmale nicht von Bedeutung.
Setzt man die geschätzten Parameter in Gleichung 5.18 ein, ergibt sich nach Vereinfachung die Log-Likelihood als Funktion von $\mathbf{A}$ (Kumar, 1997, S. 67):

$$\begin{aligned} l(\mathbf{A}, \mathbf{x}) = & -\frac{N}{2}\log|\mathrm{diag}((\mathbf{A}^{\mathbf{p-q}})'(\mathbf{T})(\mathbf{A}^{\mathbf{p-q}}))| - \sum_{k=1}^{K}\frac{N_k}{2}\log|\mathrm{diag}((\mathbf{A}^{\mathbf{q}})'\mathbf{W_k}(\mathbf{A}^{\mathbf{q}}))| \\ & -\frac{Np}{2}(1+\log(2\pi)) + N\log|\mathbf{A}| \end{aligned} \tag{5.24}$$

und nach Vereinfachung in der über $\mathbf{A}$ zu optimierenden Funktion resultiert letztendlich

$$\begin{aligned} \mathbf{A_{HDA}} = \arg\max_{\mathbf{A}} \Bigg(& -\frac{N}{2}\log\left(|\mathrm{diag}\left((\mathbf{A}^{\mathbf{p-q}})'(\mathbf{W}+\mathbf{B})(\mathbf{A}^{\mathbf{p-q}})\right)|\right) \\ & -\sum_{k=1}^{K}\frac{N_k}{2}\log\left(|\mathrm{diag}\left((\mathbf{A}_{\mathbf{k}}^{\mathbf{q}})'(\mathbf{W_k})(\mathbf{A}_{\mathbf{k}}^{\mathbf{q}})\right)|\right) + N\log|\mathbf{A}|\Bigg). \end{aligned} \tag{5.25}$$

Anders als in der *linearen Diskriminanzanalyse* ist eine analytische Berechnung der Ladungsmatrix nicht möglich. Die Optimierung von Gleichung 5.24 erfolgt numerisch. Gales (1999) beschreibt hierfür einen effektiven Algorithmus (siehe auch Anhang B.11). Im Spezialfall identischer Klassenkovarianzmatrizen $\mathbf{W_k}$ stellt die Transformationsmatrix $\mathbf{A}$ einer linearen Diskriminanzanalyse auch eine Lösung der heteroskedastischen Diskriminanzanalyse dar (vergleiche hierzu Anhang B.10).

Erdogan (2005) schlägt eine regularisierende Restriktion der Transformationsmatrix dahingehend vor, dass $\mathbf{A}$ blockweise Nullen enthält und auf diese Weise durch die lineare Transformation explizit definierte Untergruppen von Merkmalen zusammengefasst werden (im Anwendungsfall sind dies die Gruppen der Originalmerkmale bzw. Δ-Koeffizienten, vgl. auch weiter unten, Abschnitt 5.4). Im Gegensatz dazu sollen im Fall dieser Arbeit die Transformationen explizit eingesetzt werden, um durch Linearkombinationen von Merkmalen **unterschiedlicher** Merkmalssätze gemeinsame Information in niedriger dimensionalen Merkmalssätzen zu verdichten. Aus diesem Grund wird im folgenden Abschnitt eine andere Art der Regularisierung vorgestellt.

5.3.3 Regularisierte heteroskedastische Diskriminanzanalyse

Die heteroskedastische Diskriminanzanalyse ermöglicht durch die Annahme klassenspezifischer Kovarianzmatrizen eine flexiblere Modellierung bei der Dimensionsreduktion im Gegensatz zur einfachen, homoskedastischen linearen Diskriminanzanalyse. Andererseits stehen für die Schätzung der jeweiligen $\mathbf{W_k}$ erheblich weniger Beobachtungen zur Verfügung als für die Schätzung von $\mathbf{W}$ im Fall einer einfachen LDA. In der Anwendung dieser Arbeit teilen sich die Beobachtungsvektoren auf 183 Klassen der HMM-Zustände auf (siehe Abschnitt 5.3.1). Burget (2004) schlägt von daher eine *geglättete heteroskedastische Diskriminanzanalyse* (engl. *Smoothed Heteroscedastic Discriminant Analysis*, SHDA) vor. Deren Idee besteht darin, eine Glättung von klassenindividueller und -gemeinsamer, gepoolter Kovarianzmatrix aller Klassen über eine Konvexkombination

$$\mathbf{W_k^{SHDA}}(\lambda) = (\lambda)\mathbf{W_k} + (1-\lambda)\mathbf{W} \tag{5.26}$$

vorzunehmen. Der Parameter $\lambda \in [0,1]$ legt dabei fest, wie stark die Kovarianzschätzung hinsichtlich des allgemeinen Mittels über alle Klassen geglättet werden soll: die Extremwerte $\lambda = 0$ bzw. $\lambda = 1$ repräsentieren die Extremfälle einer LDA bzw. HDA. Die verwendeten Kovarianzmatrizen $\mathbf{W_k^{SHDA}}$ werden dann analog zur HDA aus Abschnitt 5.3.2 weiterverwendet und ersetzen dort die $\mathbf{W_k}$. Der Vorteil eines solchen Vorgehens liegt in der Gewährleistung einer flexiblen Modellierung einerseits bei gleichzeitig höherer Stabilität im Verhältnis zu den beiden Originalverfahren.

Friedman (1989) schlägt für die Klassifikation eine leicht modifizierte Version der Kovarianzschätzung im Verhältnis zu Burget (2004) vor bei der auch eine entsprechende Skalierung der sich aus der Regularisierung ergebenden Beobachtungsanzahl vorgenommen wird:

$$\mathbf{W_k^{Fried}}(\lambda) = \frac{(\lambda)N_k\mathbf{W_k} + (1-\lambda)N\,\mathbf{W}}{(\lambda)N_k + (1-\lambda)N}. \tag{5.27}$$

Außerdem schlägt Friedman (1989) noch eine weitere Regularisierung hinsichtlich unkorrelierter Merkmale (d.h. Diagonalität) vor, so dass sich letztendlich

$$\mathbf{W_k^{RDA}}(\lambda,\gamma) = (\gamma)\mathbf{W_k^{Fried}}(\lambda) + (1-\gamma)\frac{tr(\mathbf{W_k^{Fried}}(\lambda))}{p}\mathbf{I} \tag{5.28}$$

ergibt. Die durch die Merkmalsdimension dividierte Spur von $\mathbf{W_k^{Fried}}(\lambda)$ im letzten Summanden entspricht der durchschnittlichen Varianz über alle Variablen. Eine derartige Regularisierung wird als *Shrinkage* (dt. Schrumpfung) bezeichnet.
Der Shrinkage-Ansatz für die Diskriminanzanalyse geht auf Di Pillo (1976) zurück. Im Falle von Klassifikationsproblemen bewirkt eine Regularisierung der Kovarianzschätzung eine verzerrte Klassifikationsregel einerseits, die aber auf der anderen Seite eine geringere Varianz besitzt (Di Pillo, 1976; Läuter, 1992). Dies kann insbesondere im Fall eines sehr geringen Verhältnisses von Beobachtungen je Klasse zu Variablen oder starker Kollinearität der Prädiktoren eine deutliche Verbesserung der Klassifikationsergebnisse bewirken (Di Pillo, 1979). Insgesamt stellt sich die Problematik des Abwägens zwischen erwartungstreuer Klassifikation und einer geringen Streuung der Schätzung der Diskriminanzfunktion, beeinflusst durch die Wahl der Regularisierungsparameter. In Di Pillo (1979), Läuter (1992) und Grüning und Kropf (2006) werden Vorschläge einer sinnvollen Wahl von γ zur Regularisierung von LDA gemacht. Friedman (1989) schlägt für die RDA die Evaluierung eines Gitters möglicher Parameterkombinationen hinsichtlich eines Testfehlers vor. Dieses Vorgehen kommt auch in dieser Arbeit zum Einsatz.
Di Pillo (1976) erzielt in Simulationsexperimenten konsistent bessere Ergebnisse auf Basis regularisierter LDA (d.h. für eine Wahl von $\gamma > 0$) im Vergleich zur einfachen, stichprobenbasierten Methode. Dieser Ansatz der Regularisierung aus der Klassifikation sei nun auf die in den vorangegangenen Abschnitten beschriebene Problemstellung der linearen Dimensionsreduktion des Originalmerkmalsraums übertragen:

Definition 5.1 *Es bezeichne* $\mathbf{A_{RHDA}}(\lambda, \gamma)$ *diejenige Transformationsmatrix, die sich auf Basis von heteroskedastischer linearer Diskriminanzanalyse ergibt, wenn als Schätzer der Klassenkovarianzmatrizen im Originalmerkmalsraum die Matrizen* $\mathbf{W_k^{RDA}}(\lambda, \gamma)$ *verwendet werden.*

Eine solche Regularisierung unter Einbezug der Shrinkage-Idee zur linearen Dimensionsreduktion findet bis dato auf dem Gebiet der automatischen Spracherkennung keine Anwendung.
Eine Regularisierung der Kovarianzmatrizen ist insbesondere im Fall hoher Kollinearität sowie einer geringen Anzahl an Beobachtungen im Verhältnis zur Merkmalsdimension sinnvoll (vgl. Grüning und Kropf, 2006). In den Überlegungen von Friedman (1989) wird stets von gleichen Klassenhäufigkeiten ausgegangen. Die durch die 183 HMM-Zustände bestimmten Klassen der Anwendung dieser Arbeit zeichnen sich durch eine hohe Unbalanciertheit aus, so existieren je nach Merkmal meist Zustände selten vorkommender Phoneme, denen nur wenige Beobachtungsvektoren zugewiesen werden; für die am häufigsten vorkommenden Pausen stehen dagegen zum Teil mehrere tausend Merkmalsvektoren zur Kovarianzschätzung zur Verfügung. Dies legt eine klassenweise unterschiedliche Wahl der Parameter λ und γ nahe, wobei diese mit wachsender Anzahl an Beobachtungen gegen eine unregularisierte Kovarianzschätzung $\mathbf{W_k}$ streben sollte. Dies führt zur Motivation der *Unbalancierten regularisierten heteroskedastischen Diskriminanzanalyse* (URHDA):

Definition 5.2 *Es bezeichne* $\mathbf{A_{URHDA}}(\beta_\lambda, \beta_\gamma, N_{0,\lambda}, N_{0,\gamma})$ *diejenige Transformationsmatrix, die auf Basis von heteroskedastischer linearer Diskriminanzanalyse erhalten wird, wenn als Schätzer der Klassenkovarianzmatrizen im Originalmerkmalsraum die Matrizen* $\mathbf{W_k^{RDA}}(\lambda_k, \gamma_k)$ *verwendet werden, mit*

$$\begin{aligned} \lambda_k(N_k, \beta_\lambda, N_{0,\lambda}) &= \frac{e^{\beta_\lambda(N_k - N_{0,\lambda})}}{1 + e^{\beta_\lambda(N_k - N_{0,\lambda})}} \quad \textit{und} \\ \gamma_k(N_k, \beta_\gamma, N_{0,\gamma}) &= \frac{e^{\beta_\gamma(N_k - N_{0,\gamma})}}{1 + e^{\beta_\gamma(N_k - N_{0,\gamma})}}. \end{aligned} \tag{5.29}$$

Eine solche Funktion nimmt für große Klassenhäufigkeiten N_k Werte von λ_k und γ_k nahe 1 an und strebt gegen die Anwendung einer HDA. Der Parameter $N_{0,\lambda/\gamma}$ beschreibt den Wendepunkt der Funktion mit $\lambda_k/\gamma_k(N_k) = 0.5$. Abbildung 5.2 beschreibt den

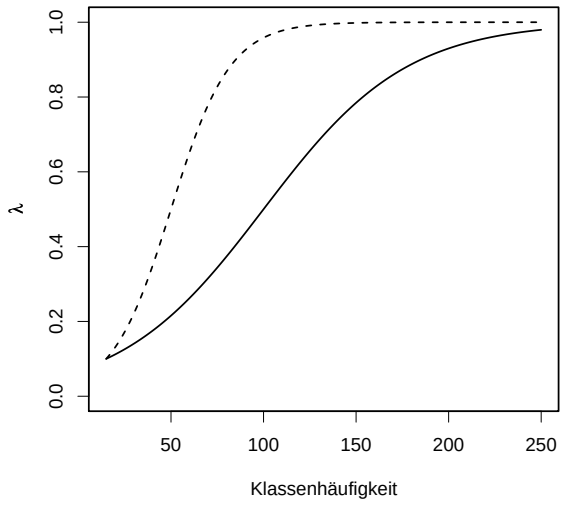


Abbildung 5.2: Exemplarischer Verlauf von λ_k (bzw. γ_k, Ordinate) in Abhängigkeit von N_k (Abszisse) für $N_0 = 50$ (gestrichelte Linie) und $N_0 = 100$ (durchgezogene Linie). Beide Kurven sind so gewählt, dass für $N_k = 15$ gilt: $\lambda_k = 0.1$.

Verlauf von $\lambda_k/\gamma_k(N_k)$ für zwei verschiedene Parameterkombinationen. Intuitiv sollte mit größerem N_0, d.h. einer größeren erforderlichen Klassenbeobachtungzahl bei unregularisierter Kovarianzschätzung, der Anstieg der Kurve einen weiteren Bereich von „Klassenhäufigkeiten“ umfassen, d.h. β kleiner werden. Ein solcher Zusammenhang könnte beispielsweise durch die Funktion $\beta_{\lambda/\gamma} = \frac{c_{\beta/\gamma}}{N_{0,\beta/\gamma}}$ beschrieben werden, mit einer Konstante c.

$$\begin{aligned} \lambda_k(N_k, c_\lambda, N_{0,\lambda}) &= \frac{e^{\frac{c_\lambda}{N_{0,\lambda}}(N_k - N_{0,\lambda})}}{1 + e^{\frac{c_\lambda}{N_{0,\lambda}}(N_k - N_{0,\lambda})}} = \frac{e^{\frac{c_\lambda N_k}{N_{0,\lambda}} - c_\lambda}}{1 + e^{\frac{c_\lambda N_k}{N_{0,\lambda}} - c_\lambda}} \quad \text{und} \\ \gamma_k(N_k, c_\gamma, N_{0,\gamma}) &= \frac{e^{\frac{c_\gamma}{N_{0,\gamma}}(N_k - N_{0,\gamma})}}{1 + e^{\frac{c_\gamma}{N_{0,\gamma}}(N_k - N_{0,\gamma})}} = \frac{e^{\frac{c_\gamma N_k}{N_{0,\gamma}} - c_\gamma}}{1 + e^{\frac{c_\gamma N_k}{N_{0,\gamma}} - c_\gamma}}. \end{aligned} \tag{5.30}$$

Die zur Schätzung „erforderliche“ Klassenhäufigkeit N_k wird mit der Merkmalsdimension wachsen. Die Anzahl zu schätzender Parameter wächst quadratisch mit der Merkmalsdimension (vgl. z.B. Hastie u. a., 2001, S. 89). Es kann von daher eine Umparame-

trisierung vorgenommen werden in das Verhältnis von Beobachtungen zur quadrierten Anzahl an Variablen $N_k^* = \frac{N_k}{p^2}$. Analog ist $N_0^* = \frac{N_0}{p^2}$. Für die klassenspezifischen Regularisierungsparameter gilt:

$$\begin{aligned} \lambda_k(N_k^*, c_\lambda, N_{0,\lambda}) &= \frac{e^{\frac{c_\lambda N_k^*}{N_{0,\lambda}^*} - c_\lambda}}{1 + e^{\frac{c_\lambda N_k^*}{N_{0,\lambda}^*} - c_\lambda}} \text{ und} \\ \gamma_k(N_k^*, c_\gamma, N_{0,\gamma}) &= \frac{e^{\frac{c_\gamma N_k^*}{N_{0,\gamma}^*} - c_\gamma}}{1 + e^{\frac{c_\gamma N_k^*}{N_{0,\gamma}^*} - c_\gamma}}. \end{aligned} \tag{5.31}$$

Auch die Konstante $c_{\lambda/\gamma}$ lässt sich weiter motivieren durch die Vorgabe eines Wertes $\lambda_k/\gamma_k(\frac{1}{p} = \frac{N_k}{p^2}) = r_{\lambda/\gamma}$, des Grads an Regularisierung, wenn die Beobachtungszahl genau der Variablenanzahl entspricht. Es gilt:

$$\log\left(\frac{r}{1-r}\right) = \frac{cN_k^*}{N_0^*} - c, \tag{5.32}$$

und damit weil in diesem Fall $N_k^* = \frac{1}{p}$

$$c = \frac{pN_0^*}{1 - pN_0^*} \log\left(\frac{r}{1-r}\right). \tag{5.33}$$

Da eine solche Form klassenspezifischer Regularisierung in der Literatur bisher nicht vorgeschlagen wurde, existieren auch keinerlei Vorgaben zur Wahl der freien Parameter $N_{0,\lambda}, N_{0,\gamma}, r_\lambda$ und r_γ. Für die Anwendungen in dieser Arbeit werden kreuzvalidiert Parameterkombinationen auf einem vorgegebenen Gitter plausibel erscheinender Werte anhand der mit ihnen erzielten Erkennungsrate ausgewertet und die optimale Parameterkombination bestimmt.

5.4 Implementierung

Zum Vergleich der Merkmale aus dem auditorischen Simulationsmodell-Output werden zur Dimensionsreduktion und Dekorrelation LDA und RHDA – wie in Abschnitt 5.3 beschrieben – sowie für die Merkmale spektraler Gestalt zusätzlich AFCCs (vgl. Abschnitte 5.2 bzw. 3.3) verwendet. Insbesondere können die Diskriminanzanalysen zur Kombination verschiedener Merkmalssätze verwendet werden, indem deren Vektoren

zusammengefügt werden und auf dem resultierenden, gemeinsamen Merkmalsvektor höherer Dimension anschließend eine Dimensionsreduktion durchgeführt wird. Es werden jeweils transformierte Merkmalssätze unterschiedlicher Dimension untersucht.
Eine Schwierigkeit bei der Anwendung von Diskriminanzanalysen (DA) besteht in der Zuordnung der Klassen (d.h. der HMM-Zustände, vgl. S. 84) zu den Merkmalsvektoren, die für das Trainingsmaterial natürlich nicht bekannt sind. Es muss daher zunächst eine Zuordnung der HMM-Zustände in den Trainingsdaten erfolgen, ehe die DA anwendbar ist. Der Vergleichbarkeit halber wird für sämtliche untersuchten Merkmale die gleiche Zustands-Labellung auf Basis eines MFCC basierten Standard-Erkenners verwendet.
Die transformierten Merkmale $\mathbf{y}(t)$ werden zusätzlich erweitert um ihre ersten und zweiten Differenzen $\mathbf{y}_\Delta(t) = \mathbf{y}(t) - \mathbf{y}(t - 20\,ms)$ und $\mathbf{y}_{\Delta\Delta}(t) = \mathbf{y}_\Delta(t) - \mathbf{y}_\Delta(t - 20\,ms)$. Da eine solche Differenzenbildung eine Linearkombination eines – über mehrere Zeitpunkte konkatenierten – Merkmalsvektors $\mathbf{y}_c(t)$ darstellt, ließe sich vermuten, dass eine lineare Diskriminanzanalyse in der Lage ist, diese effizient implizit aus $\mathbf{y}_c(t)$ zu extrahieren. In einer Studie zeigen Eisele u. a. (1996) jedoch, dass eine explizite Verwendung der Differenzenmerkmale (Δ-*Koeffizienten*) zu besseren Erkennungsergebnissen (auf Basis logarithmierter Spektralkoeffizienten) führt als deren implizite Bestimmung durch eine LDA.

6 Vergleichsstudie der verschiedenen Merkmalssätze

6.1 Übersicht

Im folgenden Kapitel werden die in Kapitel 3 motivierten verschiedenen Merkmale (vgl. S. 61) auf ihren Informationsgehalt (für die Erkennung von Sprache anhand der bei ihrem Einsatz erzielten Erkennungsrergebnisse) hin untersucht und miteinander kombiniert. Dazu werden sie als *Front End* eines Spracherkennungssystems verwendet. Als *Back Ends* werden *Hidden Markov Modelle* gebildet, wie in Abschnitt 4.4 beschrieben. Für einige der Merkmale bietet sich eine anschließende cepstrale Transformation zu AFCCs (vgl. Kapitel 5.2) an. Alternativ werden zudem Diskriminanzanalysen zur Dimensionsreduktion und Dekorrelation der Originalmerkmalssätze durchgeführt (vgl. Kapitel 5.3). Dies ist besonders sinnvoll bei kombinierten Merkmalssätzen, bei denen eine höhere Ausgangsdimension vorliegt. Die implementierten Merkmalstransformationen sind in Abschnitt 5.4 zusammengefasst. Es werden Erkennungsraten, sowohl auf den vorgeschlagenen originalen Merkmalen als auch auf transformierten Vektoren unterschiedlicher Dimension gebildet und miteinander verglichen.

Aus den wie in Kapitel 3 erzeugten Merkmalen werden Δ- und $\Delta\Delta$-Koeffizienten berechnet und mit diesen konkateniert. Diese definieren sich als $X_\Delta(t) = X(t) - X(t - 20\,ms)$, sowie $X_{\Delta\Delta}(t) = X_\Delta(t) - X_\Delta(t - 20\,ms)$. Hermansky (1997) betont die Bedeutung der Erfassung von zeitlichen Veränderungen des Sprachsignals für die automatische Spracherkennung und Eisele u. a. (1996) stellen speziell den Nutzen von Δ-Koeffizienten zum Erzielen höherer Erkennungsraten fest.

6.2 Beschreibung der durchgeführten Simulationsstudie

6.2.1 Datenbasis

Als Datenbasis zur Evaluierung der Spracherkennungssysteme mit verschiedenen Front Ends dient die TIMIT Sprachdatenbank (Garofolo u. a., 1993). Diese stellt eine häufig verwendete Standard-Datenbank dar. Sie enthält 6300 gesprochene Sätze amerikanischen Englischs (entsprechend etwa vier Stunden Sprachmaterial) von 630 unterschiedlichen Sprechern (beiden Geschlechts) aus acht verschiedenen Dialektregionen der USA.

Die Datenbank besteht aus drei Teilen, die als SA, SX und SI bezeichnet werden. Die SA Sätze wurden speziell ausgewählt um dialektspezifische Besonderheiten der unterschiedlichen Regionen herauszustellen. Jeder der 630 Sprecher spricht zwei solcher Sätze. Bei den SX Sätzen handelt es sich um 450 sogenannte „kompakte Sätze“, die ausgewählt wurden, die Menge an Phonempaarabfolgen gut abzubilden. Jeder der Sprecher spricht fünf dieser Sätze, so dass jeder Satz insgesamt sieben mal gesprochen wird. Die SI Sätze wurden aus bereits existierenden Quellen ausgewählt mit dem Ziel, eine hohe Diversität an allophonen Kontexten zu gewährleisten, d.h. die Phoneme in unterschiedlichen lautlichen (allophonen) Varianten zu repräsentieren. Jeder Sprecher spricht drei solcher Sätze. Dabei wird kein Satz zweimal gesprochen, so dass sich insgesamt 1890 SI Sätze ergeben.

Standardmäßig sind die 6300 Sätze in 4620 Trainings- und 1680 Testsätze aufgeteilt. Für die Evaluationsstudie in dieser Arbeit wurde aufgrund der hohen Rechendauer des auditorischen Simulationsmodells lediglich das TIMIT core Set verwendet, wie es von den Autoren vorgeschlagen wird: Demnach besteht das Testset aus 24 Sprechern (je zwei vorgegebene Männer und eine Frau aus jeder Dialektregion), von denen je die fünf kompakten SX- sowie die drei phonetisch diversen SI-Sätze verwendet werden – insgesamt 192 Sätze. Das Trainingset besitzt im Verhältnis zum Testset die dreifache Größe (576 Sätze), und es bestehen keine Überlappungen zwischen den Sprechern beider Sets.

Für die Auswertungen wurden die TIMIT-Files mit der Software `sox` (sourceforge.net,

2008) auf eine Abtastrate von 44.1 kHz konvertiert, da diese dem Inputformat des verwendeten auditorischen Simulationsmodells entspricht und mit Hilfe des `R` Paketes `tuneR` (vgl. Ligges, 2006, Kapitel 7) amplitudennormalisiert. Zunächst erfolgt eine erste, deskriptive Auswertung zur Bewertung der unterschiedlichen vorgestellten Merkmalssätze, die auch Aufschluss über die Wahl freier Parameter liefert. Als Gütemaß zur Beurteilung der Performance eines automatischen Spracherkennungssystems auf Basis der in dieser Arbeit beschriebenen Merkmale fungiert die *Correctness* (siehe Young u. a., 2005, S. 184). Diese beschreibt den Anteil richtig erkannter Phoneme im Testdatensatz und definiert sich wie folgt:

Definition 6.1 *Sei N die Summe aller im Testdatensatz auftretenden Phoneme. S bezeichne die Anzahl an Ersetzungsfehlern, d.h. in einer ansonsten korrekt erkannten Phonemabfolge wurde eines der Phoneme durch ein anderes, falsches ersetzt. D beschreibt die Anzahl gänzlich fehlender Phoneme in ansonsten korrekt erkannten Phonemsequenzen. Die* ***Correctness*** *definiert sich dann als*

$$corr = \frac{N - S - D}{N}, \tag{6.1}$$

d.h. sie beschreibt den Anteil Phonemen der korrekt identifiziert wurde.

6.2.2 Deskriptiver Vergleich der unterschiedlichen Merkmale

Zunächst werden alle in Kapitel 3 vorgestellten einzelnen Merkmalssätze anhand der bei ihrer Verwendung erzielten Correctness verglichen. Um jedoch die zentrale Fragestellung dieser Arbeit nach einer effektiven Informationsextraktion aus dem Output des auditorischen Simulationsmodells weiter zu erhellen, bietet es sich an, diese anhand der verschiedenen *Charakteristika* der Merkmale, wie sie in der Tabelle auf Seite 61 beschrieben sind, zu unterscheiden:
Dieses sind:

- Informationskodierung über den *Ort* der neuronalen Aktivierung,
- Informationskodierung über *Phase locking* sowie
- Informationskodierung durch Information über die *Delaystruktur* der Aktionspotenzialzeitpunkte verschiedener ANFs

(vgl. hierzu Tabelle 3.1). Insbesondere ist interessant, ob durch Kombination verschiedener Merkmale eine Verbesserung des Erkennungsergebnisses erzielt werden kann. Von Interesse ist auch ein Vergleich der Weiterverarbeitungsschritte des *auditorischen Musters (AI)*. Diese sind:

- Merkmale basierend auf einer mehrfach wiederholten Simulation im Vergleich zu den auf einmaliger Simulation extrahierten Merkmalssätze,
- parallele lokale vs. einfache Delay-Computing Netzwerke sowie
- eine abschließende Merkmalstransformation durch Diskriminanzanalysen (sowie deren regularisierter heteroskedastischer Erweiterung, vgl. Abschnitt 5.3.3) im Vergleich zu einer abschließenden cepstralen Transformation im Falle spektral interpretierbarer Merkmale oder keiner zusätzlichen Transformation.

Hinsichtlich der Recheneffizienz kann die Aussage getroffen werden, dass das mit großem Abstand langsamste Glied der Kette die Umwandlung der Originalschallwelle in die Reaktion der Hörnerven durch das auditorische Simulationsmodell darstellt. Diese erfolgt (für mono Signale, d.h. ein Ohr, auf AMD Athlon 64×2, 2.61 GHz) in etwa im 30-fachen der Echtzeit. Eine mehrfache Wiederholung beschränkt sich auf die Wiederholung der letzten Prozessschritte von Neurotransmitteremission und Aktionspotenzialgenerierung (vgl. Kapitel 2 bzw. Anhang A), die ca. 30% der Gesamtrechenzeit ausmachen, so dass sich für eine M-fache Simulation des auditorischen Modells etwa eine Rechenzeit von $30 + (M-1) \cdot 9$ multipliziert mit der Dauer der Originalschallwelle ergibt.

6.2.3 Auditorisch erweiterter Merkmalssatz

Abschließend wird untersucht, ob die aus dem auditorischen Modell extrahierten Merkmale Information enthalten, die in direkt aus dem Spektrum extrahierten MFCC-Merkmalen (vgl. Abschnitt 3.3) noch nicht enthalten ist. Die Ergebnisse aus den vorangehenden Untersuchungen der, auf Basis des auditorischen Simulationsmodells erhaltenen Merkmalsrepräsentationen werden genutzt, um einen *auditorisch erweiterten Merkmalssatz (aeM)* zu konstruieren. Dabei werden die Ergebnisse der deskriptiven

Auswertung verwendet, um zunächst eine Vorauswahl an Merkmalssätzen zu identifizieren. Die Merkmale mit den besten Resultaten werden **zusammen** mit Melfrequenz Cepstralkoeffizienten zu einem gemeinsamen Merkmalsvektor höherer Dimension zusammengefasst. Die Idee einer solchen Merkmalserzeugung ist ähnlich zu dem in Weihs u. a. (2007) in Abschnitt drei beschriebenen Vorgehen, anstelle der dort vorgeschlagenen Variablenselektion wird jedoch eine lineare Dimensionsreduktion mit Hilfe von HRDA bzw. URHDA (vgl. Abschnitt 5.3.3) durchgeführt. Es ist anzumerken, dass sich die in Weihs u. a. (2007) vorgeschlagene Variablenselektion als Spezialfall einer linearen Dimensionsreduktion $Y_i = \sum_j a_{ij} X_j$ interpretieren lässt, wobei sämtliche Koeffizienten a_{ij} entweder 1 oder 0 werden und zusätzlich gilt: $\sum_j a_{ij} = 1$ sowie $\sum_i a_{ij} \leq 1$.

Die Erkennungs-Performance auf Basis des auditorisch erweiterten Merkmalssatzes soll nun abschließend mit derjenigen bei einfacher Verwendung von MFCCs – die häufig in der Praxis zum Einsatz kommen (vgl. Anhang C.2) – verglichen werden.

6.2.4 Tests auf Signifikanz

Der Vergleich der Testfehlerraten erfolgt zunächst in rein deskriptiver Art. Er erlaubt zwar einen Eindruck der Performance, allerdings noch keine Signifikanzaussagen hinsichtlich der Ergebnisse. Für besonders interessante Performance-Vergleiche bietet sich weiterhin die Möglichkeit eines statistischen Tests an, wie er im Folgenden beschrieben wird. Für einen *Signifikanztest* der Correctness-Differenzen zweier Merkmalssätze i und j kann die Methode nach Dietterich (1998) auf Basis von 5×2-*facher Kreuzvalidierung* verwendet werden.
Dietterich schlägt zum Testen eine zweifache Kreuzvalidierung vor aus zwei Gründen:

- Im Gegensatz zur k-fachen Kreuzvalidierung mit $k > 2$ ist der Testdatensatz größer und die Annahme einer approximativer Normalverteilung der Differenzen $\Delta_i^{(m)}$ (s.u.) beider Methoden eher zu rechtfertigen.

- Im Gegensatz zur k-fachen Kreuzvalidierung mit $k > 2$ sind insbesondere die Trainingsdaten nicht überlappend und damit die Erkennungsraten auf beiden Testdaten der Kreuzvalidierung gegenseitig weniger abhängig.

Sei nun $\Delta_{ij,q}^{(m)} := \text{corr}_{i,q}^{(m)} - \text{corr}_{j,q}^{(m)}$ die Differenz der Correctness auf dem Testdatensatz $m \in \{1,2\}$ der $q \in \{1,\ldots,5\}$-ten Wiederholung der 2-fachen Kreuzvalidierung zwischen beiden Methoden. Sei weiterhin $s_{ij,q}^2 := \sum_m (\Delta_{ij,q}^{(m)} - \bar{\Delta}_{ij,q})^2$. Dann gilt unter angenommener Unabhängigkeit von $\Delta_{ij,q}^{(1)}$ und $\Delta_{ij,q}^{(2)}$: $\frac{s_{ij,q}^2}{\sigma^2} \sim \chi_1^2$ mit der echten Varianz σ^2 der Correctness-Differenzen.
Dietterich beschreibt jedoch eine hohe beobachtete Instabilität der $s_{ij,q}^2$ – bisweilen sogar den Wert 0 im Fall identischer Differenzen auf beiden Testdatensätzen einer Kreuzvalidierung. Aus diesem Grund schlägt er eine fünffache Wiederholung der zweifachen Kreuzvalidierung zur Stabilisierung der Varianzschätzung vor. Die vorgeschlagene Teststatistik lautet

$$T := \frac{\Delta_{ij,1}^{(1)}}{\sqrt{\frac{1}{5}\sum_{q=1}^{5} s_{ij,q}^2}} \sim t_5 \tag{6.2}$$

und ist unter der Nullhypothese gleicher Correctness für beide verwendeten Merkmalssätze approximativ t-verteilt (Dietterich, 1998).

Die Wahl freier Modellierungsparameter wie die der Regularisierungsparameter λ und γ (bzw. $\beta_{\lambda/\gamma}$ und $N_{0,\lambda/\gamma}$) im Falle der RHDA (bzw. URHDA) erfolgt dabei anhand der Ergebnisse der vorangehenden, deskriptiven Analyse auf einer separaten Unterteilung der Daten, um das Testergebnis nicht in Richtung einer positiven Performance der neu erzeugten Merkmale zu verzerren.

6.3 Ergebnisse

6.3.1 Ansteuerung

Eine erste Optimierung betrifft den Skalierungsfaktor für die Ansteuerung des Ohrmodells, damit das verarbeitete Sprachsignal weder der Lautstärke von Lärm entpricht noch unterhalb der Wahrnehmungsschwelle liegt. Diese wurde auf Basis von Orts-Durchschnittsfeuerraten-Merkmalen einer ANF-Gruppierung gemäß Seneff (1988) (vgl. Abschnitt 3.4) vorgenommen. Die Resultate sind in Anhang B.6 dargestellt. Optimale Ergebnisse wurden für eine Ansteuerung von -57.5 dB erzielt.

6.3.2 Orts-Durchschnittsfeuerraten basierte Merkmale

Abbildung 6.1 und Tabelle 6.1 zeigen die Erkennungsraten der verschiedenen Orts-Durchschnittsfeuerraten-Merkmale (siehe Abschnitt 3.4). Die besten Ergebnisse (57.68 %)

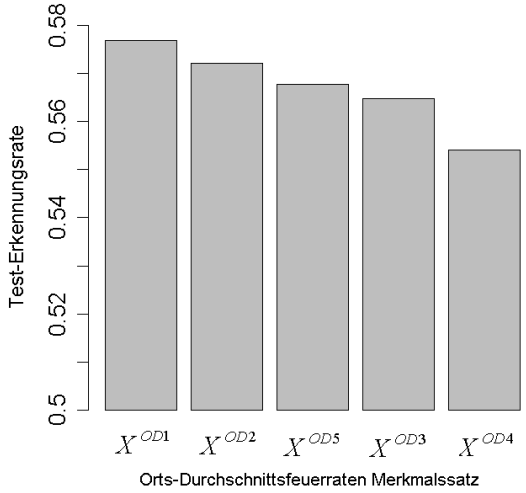


Abbildung 6.1: Phonem-Erkennungsraten auf den Testdaten.

Merkmalssatz	X^{OD_1}	X^{OD_2}	X^{OD_5}	X^{OD_3}	X^{OD_4}
Erkennungsrate	57.68	57.21	56.78	56.47	55.41

Tabelle 6.1: Phonem-Erkennungsraten für die unterschiedlichen vorgeschlagenen Orts-Durchschnittsfeuerraten Merkmalssätze (vgl. Abschnitt 3.4) auf den Testdaten.

sind für Zusammenfassung von ANFs gemäß Seneff (1988) (X^{OD_1}, je fünf ANFs, in einem Frequenzbereich von etwa 200 bis 6400 Hz) oder Perdigao und Sa (1998) (X^{OD_2}, je acht ANFs zwischen 200 und 3400 Hz) zu beobachten. Die schlechtesten Ergebnisse werden für die breitesten, zur Hälfte überlappenden Filter (X^{OD_4}) beobachtet.
Es ist zu bemerken, dass die Merkmalssätze X^{OD_2} und X^{OD_3} beide je acht benachbarte ANFs zusammenfassen, jedoch unterschiedliche Frequenzbereiche umfassen. Der besser abschneidende Merkmalssatz X^{OD_2} weist dabei ein Merkmal auf, das ausschließlich niedrigere Frequenzen erfasst, die im Merkmalsvektor X^{OD_3} keine Berücksichtigung

finden. Das gleiche gilt für die auf Seneff basierenden Merkmale X^{OD_1}, anhand derer die besten Ergebnisse erzielt werden. Beides zusammen spricht für eine niedrigere Wahl der unteren Grenze des in den Merkmalen berücksichtigten Frequenzbereichs von etwa 200 Hz.
Es liegt nun nahe, die beiden Gruppierungsweisen (je fünf ANFs nach Seneff oder je acht ANFs wie von Allen, 1994; Tchorz und Kollmeier, 1999; Perdigao und Sa, 1998, vorgeschlagen, im Folgenden als ANF Zusammenfassung „nach Allen“ bezeichnet) nochmals auszuwerten für unterschiedliche obere Grenzen des betrachteten Frequenzbereichs. Abbildung 6.2 und Tabelle 6.2 zeigen die Ergebnisse.

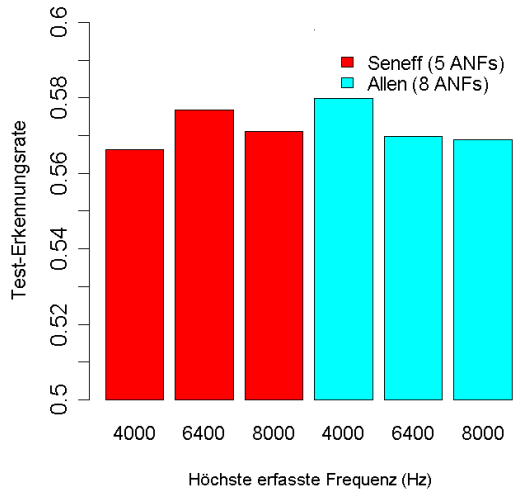


Abbildung 6.2: Phonem-Erkennungsraten auf den Testdaten.

ANF-Gruppierung	Maximale CF		
	4000	6400	8000
5 ANFs (Seneff)	56.62	57.68	57.10
8 ANFs (Allen)	57.98	56.98	56.88

Tabelle 6.2: Phonem-Erkennungsraten auf den Testdaten für unterschiedliche zur Merkmalsbildung berücksichtigte Frequenzbereiche.

Insgesamt lässt sich keine einheitliche Aussage über eine bessere Gruppierung von

ANFs treffen. Im Falle der Gruppierung nach Seneff stellt die vorgeschlagene obere Frequenzgrenze von etwa 6400 Hz die beste Wahl dar, während sich für eine Gruppierung von je acht ANFs nach Allen eine obere Frequenzgrenze von vier kHz als besser erweist. In beiden Fällen jedoch bewirkt die Wahl eines größeren Frequenzbereichs bis hin zu acht kHz einen leichten Abfall in den Test-Erkennungsraten.

6.3.3 Einfache vs. wiederholte Simulation

Vergleicht man für die am besten abschneidenden Merkmale X^{OD_1} Erkennungsraten, basierend auf einfacher Simulation (HSR), verglichen mit der ansonsten zugrundegelegten 50-fach wiederholten Simulation (H/M/LSR im Verhältnis 3:1:1, vgl. Anhang A), so fallen die Erkennungsraten auf den Testdaten von 57.68 % auf 55.30 % noch unter die schlechtesten beobachteten Ergebnisse (X^{OD_4}, 55.41 %) ab. Abbildung 6.3 veranschaulicht diese Ergebnisse am Beispiel der Merkmalsausprägungen für ca. eine Sekunde einer Sprachäußerung: In der rechten Grafik sind bei wiederholter Simulation wesentlich klarere Konturen der Merkmalsausprägungen über den zeitlichen Verlauf erkennbar. Diese Beobachtung lässt Zweifel an der modellierten Nachempfindung ein-

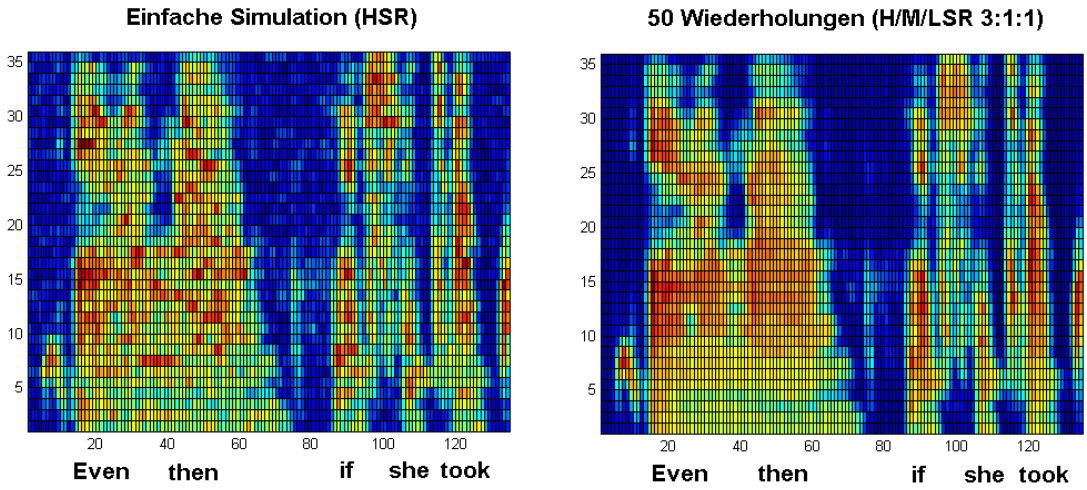


Abbildung 6.3: Exemplarische Merkmalsausprägungen von X^{OD_1} für eine Äußerung nach einfacher (links) oder mehrfacher (rechts) Simulation.

zelner Aktionspotenziale entstehen. Sie stellen zwar den neuronalen Mechanismus dar,

der dem Nervensystem zur Informationstransmission zur Verfügung steht. Es ist jedoch davon auszugehen, dass die hohe Anzahl an ANFs im menschlichen Ohr (vgl. Abschnitt A.7) eine gute Rekonstruktion momentaner Vesikelrelease-Raten zur neuronalen Weiterverarbeitung ermöglicht, wie sie durch entleerte Pools $q(t)$ einerseits und die Emissionswahrscheinlichkeit $k(t)$ gemeinsam bestimmt sind (vgl. Abschnitt A.5). Für das Simulationsmodell mit lediglich 251 ANFs ist dies nicht der Fall. Es ist denkbar an Stelle der binomialverteilten Neurotransmitterausschüttung zukünftig deterministische Ausschüttungs*raten* zu modellieren und diese anschließend einer Tiefpassfilterung (zur Nachbildung des Effekts neuronaler Refraktärzeiten) zu unterziehen.
Anhand der Orts-Durchschnittsfeuerraten-Merkmale werden nun auch exemplarisch die verschiedenen Möglichkeiten einer linearen Dimensionsreduktion untersucht.

6.3.4 Cepstrale Transformation

Die erste Untersuchung betrifft die cepstrale Transformation der extrahierten Merkmale zu AFCCs. Diese wurde für die Merkmale mit den bis dato besten Ergebnissen durchgeführt sowie zusätzlich für solche, die ein größeres Spektrum umfassen, für je unterschiedlich dimensionale cepstrale Merkmalsvektoren.
Tabelle 6.3 zeigt die Ergebnisse: Es ist ein leichtes Ansteigen der Performance für

	Seneff			Allen			X^{OD_4}	X^{OD_5}
$F_{max}(kHz)$	4	6.4	8	Perdigao	4	8		
Anzahl Koeffizienten	71	87	89	48	51	55	51	27
Test-Erkennungsrate	57.68	57.63	**58.38**	57.62	58.16	**58.71**	58.07	57.93
Vergleich: kein AFCC	56.62	57.68	57.10	57.21	57.98	56.88	55.41	56.78

Tabelle 6.3: Optimale Erkennungsraten und zugehörige AFCC-Vektordimensionen (vor der Δ-Koeffizienten-Bildung) für unterschiedliche Orts-Durchschnittsfeuerraten-Merkmale.

alle Merkmale (im Vergleich zu den Erkennungsraten auf den untransformierten Orts-Durchschnittsfeuerraten Merkmalen) zu verzeichnen. Die Erkennungsrate einiger Merkmalssätze überschreitet durch Bildung von AFCCs einen Wert von 58 %. Optimale Ergebnisse (bis hin zu 58.7 %) sind für recht große Anzahlen an Koeffizienten zu beobachten. Es ist zu erkennen, dass für die nach Seneff gebildeten Merkmale bei höher dimensionalen Merkmalsvektoren die besten Ergebnisse erzielt werden im Vergleich

zur ANF-Zusammenfassung nach Allen. Die optimale AFCC-Merkmalsdimension liegt im Bereich zwischen 50 und 55 Koeffizienten. Eine ANF-Zusammenfassung entsprechend X^{OD_4} bzw. X^{OD_5} liefert auch hier keine weitere Verbesserung der Resultate. Erstaunlicherweise werden die besten Ergebnisse – sowohl für Seneff- als auch für Allen-Merkmale – erzielt, wenn für die Originalmerkmale ein höherer Frequenzbereich gewählt wird. Dieses Resultat wird weiter bestätigt durch die (im Verhältnis zum vorigen Abschnitt) guten Ergebnisse der Gruppierung entsprechend X^{OD_4} bzw. X^{OD_5}, da auch diese je ein vergleichsweise breiteres Spektrum in die Merkmalsberechnung einbeziehen (vgl. Kapitel 3.4).

6.3.5 Lineare Diskriminanzanalyse

Die Anwendung von linearer Diskriminanzanalyse zur Merkmalstransformation liefert keine weitere Verbesserung gegenüber den auf Basis von AFCCs erzielten Erkennungsraten, jedoch ähnlich gute Ergebnisse bereits für niedriger dimensionale Merkmalsvektoren. Abbildung 6.4 zeigt die Ergebnisse für transformierte Merkmalsvektoren unterschiedlicher Größe der verschiedenen Originalmerkmale. Die maximale Vektorgröße ist durch die Dimension des Ausgangsmerkmalsraums beschränkt. Allgemein ist ein Anstieg der Erkennungsperformance für höhere Merkmalsdimensionen zu beobachten. Besonders auffällig ist, dass für identische ANF-Gruppierung auch bei unterschiedlichen berücksichtigten Frequenzbereichen nahezu gleiche Verläufe der Test-Erkennungsraten in Abhängigkeit von der Merkmalsdimension zu beobachten sind. Erneut sind für eine Gruppierung der ANFs nach Seneff und Allen die besten Ergebnisse zu beobachten. Eine Modifikation der Gruppierung nach Szepannek und Weihs (2006a) auf nur 10 ($X^{OD_5^*}$) anstelle von 15 ANFs (X^{OD_5}, SW) kann die Erkennungsraten verbessern. Am Schlechtesten schneiden erneut die auf Basis von X^{OD_4} gebildeten Merkmale ab.

Eine Berechnung von AFCCs oder Diskriminanzkoeffizienten auf logarithmierten Merkmalen erzielt – entsprechend der Behauptung von Perdigao und Sa (1998), dass der auditorische Verarbeitungsprozess bereits eine implizite, mit der Logarithmierung von Spektralamplituden vergleichbare Transformation beinhaltet (vgl. die Abschnitte 3.3 und 5.2) – keine Verbesserung der Ergebnisse.

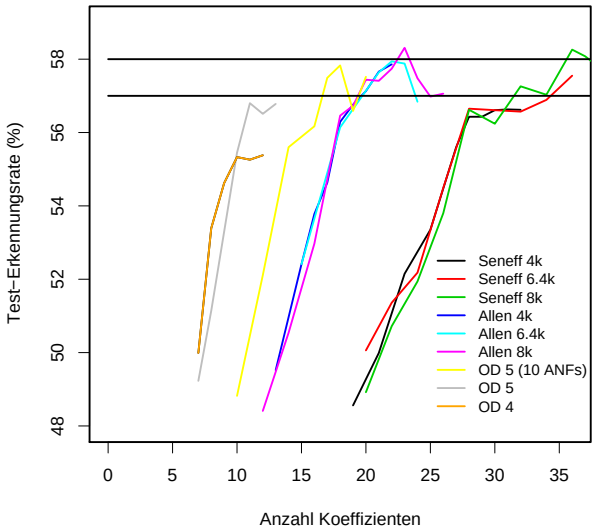


Abbildung 6.4: Phonem-Erkennungsraten auf den Testdaten nach LDA Transformation für unterschiedlich dimensionale Merkmalsvektoren.

6.3.6 Regularisierte heteroskedastische Diskriminanzanalyse

Führt man die beschriebene Erweiterung der LDA Transformation hinsichtlich der in Kapitel 5.3.3 eingeführten RHDA[1] durch, so können in allen untersuchten Merkmalen deutliche Verbesserungen erzielt werden, verglichen mit den bisher beobachteten Erkenungsraten.

	Seneff 4k	Seneff	Seneff 8k	Allen 4k	Allen 6.4k	Allen 8k	$X^{OD_5^*}$	X^{OD_5}	X^{OD_4}
LDA	56.63	57.55	58.26	57.86	57.94	58.31	57.83	56.80	55.38
Dimension	31	36	36	22	22	23	18	11	12
RHDA	59.25	59.38	59.66		**60.18**	59.08	59.69	58.57	57.98
Dimension	20	21	20		21	14	21	12	12
α	1.00	1.00	1.00		1.00	1.00	1.00	1.00	0.75
γ	0.50	0.50	0.25		0.50	0.25	0.50	0.33	0.75

Tabelle 6.4: Optimale Test-Erkennungsraten nach LDA und RHDA.

Tabelle 6.4 zeigt die Test-Erkennungsraten für eine optimal gewählte Merkmalsdimen-

[1] Ausgewertet auf einem Parametergitter $(\lambda, \gamma) \in \{0, 0.25, 0.33, 0.5, 0.66, 0.75, 1\}^2$.

sion nach Transformation. Nahezu alle Merkmale erreichen nach Transformation oberhalb von 59 % Correctness und erstmals wird auch der Wert von 60 % überschritten. Betrachtet man die Parameter λ und γ, die optimale Ergebnisse liefern, so werden die besten Resultate stets für die heteroskedastische Variante ($\lambda = 1$), bei gleichzeitiger Schrumpfung (Shrinkage) in Richtung von Unkorreliertheit der Kovarianzmatrizen, erzielt. Es ist zu bemerken, dass Burget (2004) für MFCCs zur Regularisierung zwar einen Kompromiss zwischen homoskedastischen und heteroskedastischen Klassenkovarianzmatrizen vorschlägt (entsprechend einem Parameter $\lambda < 1$), den Aspekt der Schrumpfung jedoch nicht berücksichtigt.

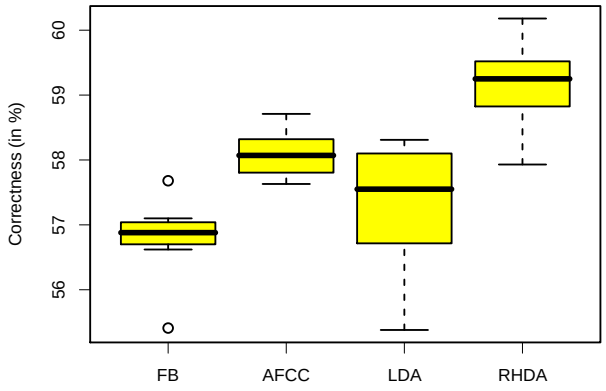


Abbildung 6.5: Vergleich der erzielten Erkennungsergebnisse (auf verschiedenen Merkmalssätzen) nach unterschiedlichen Transformationen.

Abbildung 6.5 zeigt zusammenfassend die optimalen erzielten Ergebnisse für die verschiedenen möglichen linearen Merkmalstransformationen auf den unterschiedlichen Originalmerkmalssätzen. Durch eine anschließende Transformationen kann die Correctness, im Vergleich zu den auf Basis der Originalmerkmale erzielten Werten, verbessert werden. Eine AFCC Transformation führt dabei zu besseren Ergebnissen als lineare Diskriminanzanalyse (allerdings bei im Durchschnitt mehr als doppelt so hoher Merkmalsdimension). Durch Regularisierung heteroskedastischer Diskriminanzanalyse wird die größte Verbesserung der Erkennungsergebnisse erzielt.

Die folgenden Analysen seien bezüglich der untersuchten ANF-Gruppierungen auf die beiden Ansätze nach Seneff bzw. Allen beschränkt mit einer maximalen CF von jeweils 6400 Hz, da diese in den bisherigen Untersuchungen konstant gute Ergebnisse aufwiesen.

6.3.7 Inter-Spike Intervall basierte Merkmale

Es werden nun Merkmale näher untersucht, die sich die Eigenschaft des *Phase Lockings* der Hörnerven zu Nutze machen um die an den ANFs vorliegenden Frequenzen zu beschreiben. Abbildung 6.6 zeigt exemplarisch den Unterschied zwischen *generalisierter Synchronizitätsdetektion* (GSD, Seneff, 1988) und deren Erweiterung zur *durchschnittlichen lokalisierten Synchronizitäts-Detektion* (ALSD, Ali u. a., 2002) anhand der gleichen Sprachäußerung wie in Abb. 6.3. Die ALSD-Erweiterung wirkt stabi-

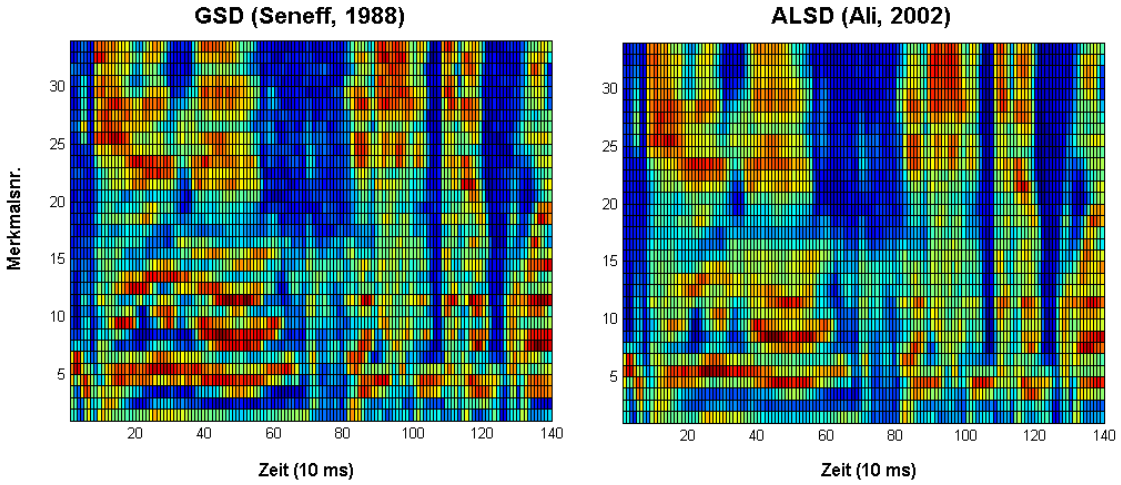


Abbildung 6.6: Exemplarische Merkmalsausprägungen von X^{GSD} (links) und X^{ALSD} (rechts) für die Sprachäußerung aus Abb. 6.3.

ler in der Hinsicht, dass aufgrund der örtlichen Glättung (entlang der Ordinate) auch glattere Konturen im zeitlichen Verlauf der Merkmale (Abszisse) auszumachen sind. Für die Auswertungen wird aus diesem Grund die ALSD-Repräsentation ausgewählt. Eine erste Untersuchung betrifft die Originalmerkmalsvektoren bei unterschiedlicher Zusammenfassung von Nachbar-ANFs: Verglichen werden Zusammenfas-

sungen von je fünf bzw. acht ANFs bis zu einer CF von etwa 6.4 kHz entsprechend Abschnitt 6.3.2. Abbildung 6.7 und Tabelle 6.5 zeigen die Ergebnisse im Vergleich zu den bei gleicher ANF-Gruppierung erzielten Test-Erkennungsraten auf Basis von Orts-Durchschnittsfeuerraten Merkmalen.

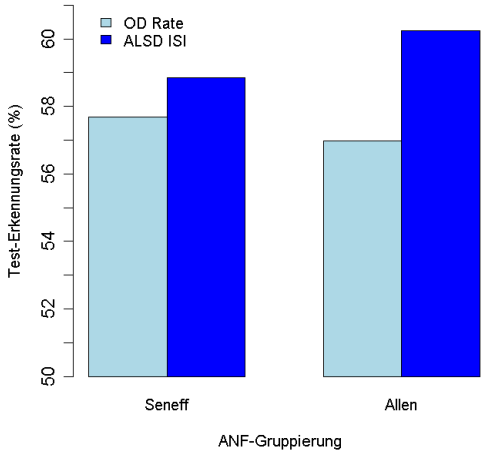


Abbildung 6.7: Test-Erkennungsraten für X^{ALSD} Merkmale (dunkelblau) unterschiedlicher ANF-Zusammenfassung (nach Seneff/Allen, s.o.) bei maximaler CF von 6.4 kHz im Vergleich zu den Ergebnissen auf Basis von X^{OD} (hellblau) bei gleicher ANF-Gruppenbildung.

	5 ANFs	8 ANFs
ISI	58.84	60.25
OD	57.68	56.98

Tabelle 6.5: Vergleich der Test-Erkennungsraten auf Basis von Inter-Spike Intervall basierten (ISI) und Orts-Durchschnittsfeuerraten basierten (OD) Merkmalen.

In beiden Fällen werden auf Grundlage der Inter-Spike Intervall basierten Merkmale bessere Ergebnisse erzielt, im Falle der Zusammenfassung nach Allen (acht ANFs) liegt die Erkennungsrate sogar erstmals bereits ohne zusätzliche Merkmalstransformation bei 60.25%.

Eine Kombination beider Merkmalsätze (ANF-Gruppierung gemäß Allen) kann die Erkennungsrate nochmals verbessern: Als optimal erweist sich dabei aber nicht wie erwartet die Verwendung von X_k^{ALSD} Merkmalen für kleine k und X_k^{OD} für große k, sondern eine simple Konvexkombination der (varianznormierten) Merkmale für gleiches k (d.h. übereinstimmende Mittenfrequenzen). Abbildung 6.8 zeigt die Erkennungsraten für unterschiedlich stark anteilige Gewichtung beider Merkmalssätze. Dabei wird sogar eine Test-Erkennungsrate von 61.11% erreicht. Keine weiteren Verbesserungen

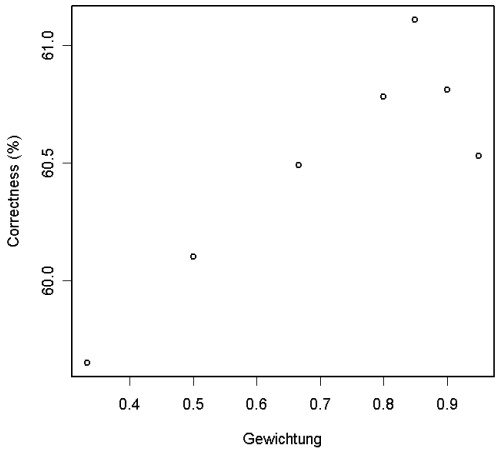


Abbildung 6.8: Erkennungsraten für verschiedene Parameter w und Konvexkombinationen $X_k^{OD/ISI,w} = w \cdot X_k^{ALSD} + (1-w) \cdot X_k^{OD}$ bei ANF-Zusammenfassung von je acht ANFs und maximaler CF von 6.4 kHz .

werden für merkmalsabhängige Gewichte $w(k)$ erzielt, wobei unter der Annahme sinkender Relevanz von X_k^{ALSD} bei höheren Frequenzen (d.h. größerem k) Funktionen der Gestalt $w(k) = \frac{exp(a(k-b))}{1+exp(a(k-b))}$ für unterschiedliche Parameterkombinationen (b in Größenordung der Merkmalsdimension, $a < 0$) verwendet werden.

6.3.8 Delay-Computing basierte Merkmale

Implementiert wurden die in Abschnitt 3.9.3 beschriebenen Maximum- und Entropie-Merkmale für Delay-Computing Netzwerke sowie parallele lokale DCNs unterschiedlicher Parametrisierung. Variiert wurden:

- die lokale Netzwerkgröße: zehn $\times$ 18, fünf $\times$ 35 sowie 24 lokale Netze mit je acht Input-Neuronen,
- der die mittlere Neigung der Delaytrajektorien beschreibende Parameter f_{min}: fünf bzw. 15 für Netzwerke der Größe 18 – wobei insbesondere der Wert 15 Neigungen zulässt, die bereits zwischen den „oberen“ Input-Neuronen eine ähnliche Zeitverzögerung zulassen wie zwischen den „unteren“, die durch die Lokalität der Netzwerke ermöglicht ist (vgl. Abschnitt 3.9.3) – 15 für die Netzwerke der Größe 35 und vier bzw. sieben für diejenigen der Größe acht sowie
- die Zeitintegrationschrittgröße: drei bzw. acht $\times$ 44100^{-1} s, wobei der erste Fall einer Abtastrate von 14.7 kHz entspricht und der zweite einer Standardparametrisierung, die im Falle von DCNs den vollständigen Trajektorienverlauf abbildet (vgl. Harczos u. a., 2007b).

Abbildung 6.9 zeigt beispielhaft die auf Basis von Maximalwerten pro Zeitfenster gebildeten Maximalwerte der einzelnen Output-Neuronen (noch nicht zu Merkmalen zusammengefasst). Unter den Output-Neuronen des DCNs (links, Ordinate) lassen sich insbesondere drei „Gebiete“ verstärkter Aktivität identifizieren. In der rechten Abbildung sind auf der Ordinate deutlich die Grenzen zwischen den lokalen Netzwerken auszumachen. Es sind auch Unterschiede der Maximalwerte von Output-Neuronen des gleichen Netzwerkes zu erkennen (z.B. im Bereich von Zeitfenster 85 und Output-Neuron 60). Tabelle 6.6 und Grafik 6.10 (links) zeigen die optimalen erzielten Ergebnisse bei variierter Anzahl an aus dem DCN- bzw. PLDCN-Output (Netzwerkgröße 18) zusammengefassten Merkmalen. Es ergeben sich schlechtere Erkennungsraten,

	DCN	PLDCNs	
f_{min}	40	5	15
Maximum	44.93	52.67	53.35
Entropie	36.19	52.00	50.68

Tabelle 6.6: Test-Erkennungsraten auf Basis von DCNs und PLDCNs der lokalen Netzwerkgröße 18 mit Zeitintegrationsschrittgröße acht.

verglichen mit den Orts-Durchschnittsfeuerraten oder den Phase Locking basierten

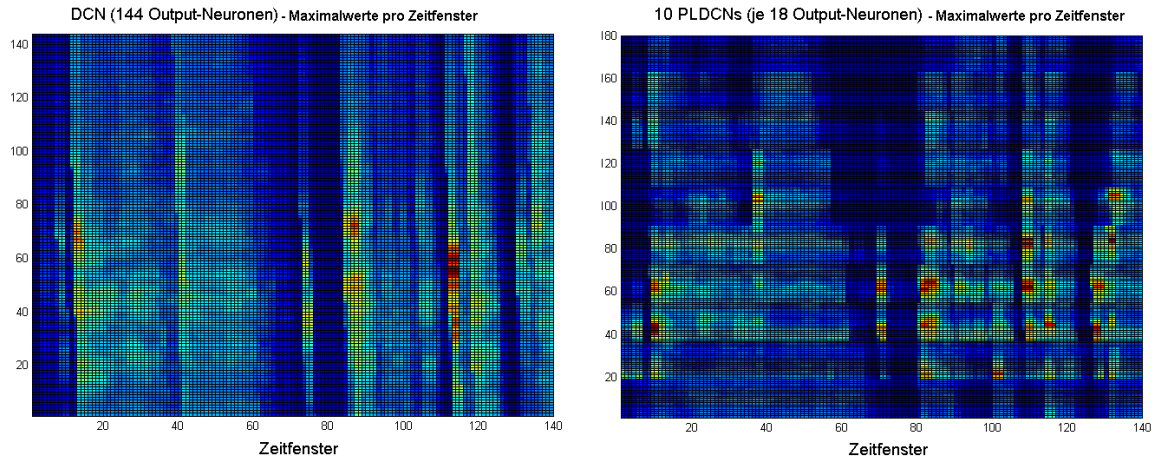


Abbildung 6.9: Exemplarische Merkmalsausprägungen von X^{Max} auf Basis eines DCNs (links) bzw. mehrerer PLDCNs der Größe 18 (rechts) für die Sprachäußerung aus Abb. 6.3.

Merkmalen. Dieses Ergebnis unterstreicht die Bedeutung von Frequenzinformation zur Spracherkennung. Wie schon in Szepannek und Weihs (2006a) wird auf Basis der PLDCN Merkmale eine Verbesserung der Ergebnisse erzielt, im Vergleich zur Verwendung einfacher DCNs. Die besten Erkennungsraten sind für eine hohe Wahl des Parameters f_{min} zu beobachten. Die Verwendung des Maximums zeigt konsistent bessere Resultate im Vergleich zur Entropie. Über die Zusammenfassung der Output-Neuronen zu Merkmalen lässt sich folgende Aussage treffen (siehe auch Abbildung 6.10, Mitte): Für DCNs werden die optimalen Ergebnisse sowohl auf Basis von Maximalwerten als auch auf Basis von Entropie bei einer Zusammenfassung (durch Summation) zu 16 Merkmalen von je neun benachbarten Output-Neuronen erzielt. Zwölf Merkmale ungleich vieler zusammengefasster Output-Neuronen (vgl. S. 54) liefern ähnlich gute Ergebnisse, jedoch keine Verbesserung (43.86 %). Eine derartige Zusammenfassung zu der als optimal identifizierten Anzahl von 16 Merkmalen (als Parameter ergeben sich: $a = 5, b = 3, c = 11.32$, vgl. ebenfalls S. 54) führt zu einem ähnlichen Ergebnis (43.97 %). Die auf der Grundlage von PLDCNs der Größe 18 erzielten Erkennungsraten werden optimal bei einer Zusammenfassung der lokalen Netzwerke zu je zwei Merkmalen gleich vieler Netzwerk-Output-Neuronen pro lokalem Netz. Während das

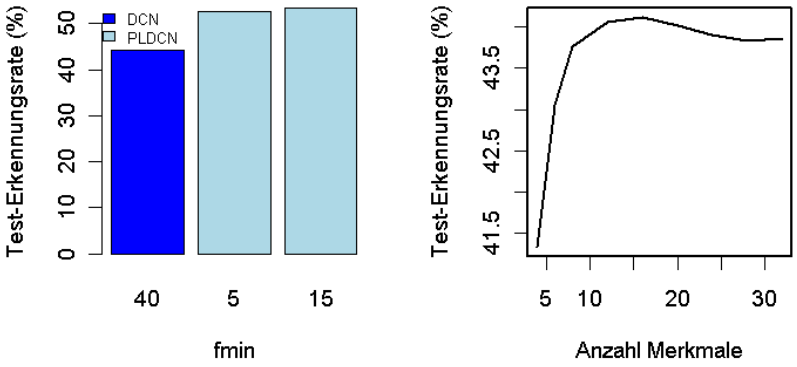


Abbildung 6.10: Links: Test-Erkennunsgrate maximumbasierter Merkmale auf Basis von DCNs und PLDCNs der Größe 18. Mitte: Test-Erkennungsrate für unterschiedliche Anzahlen an Merkmalen für DCN.

eine Merkmal die „steileren“ Trajektorienverläufe zusammenfasst und einen Index für die vorhandene Aktivität im Bereich der Input-ANFs allgemein darstellt, fasst das jeweils andere Merkmal diejenigen Trajektorien zusammen, die eine stärkere Verzögerung aufweisen, beschreibt also eine innerhalb des lokalen Netzwerkes einsetzende Verlangsamung der Wanderwelle, wie sie von Greenberg u. a. (1997) beschrieben wird. Bei zehn lokalen Netzwerken ergibt sich insgesamt ein 20 dimensionaler Merkmalsvektor.

Es ist zu bedenken, dass die Bildung lokaler Delay-Computing Netzwerke implizit Information über zugrundeliegende Signalfrequenzen innerhalb der (hier zehn) verschiedenen, durch die Netzwerk-Input-Neuronen charakterisierten Frequenzbänder beinhaltet. Eine Implementierung von 24 noch frequenzspezifischeren lokalen Netzwerken der Größe acht (entsprechend dem ANF Gruppierungsvorschlag basierend auf Allen, 1994, s.o.) liefert allerdings keine neuerliche Verbesserung der Erkennungsraten (52.8%)

Vergrößert man den Frequenzbereich, den die Input-Neuronen umfassen, durch fünf PLDCNs der Größe 35 ($f_{min} = 15$), ergibt sich eine Test-Erkennungsrate von 51.41 % (auf Basis der Maximalwerte), also weiterhin eine Verbesserung der auf der Grundlage einfacher Delay-Computing Netzwerke erzielten Resultate. Für die Merkmalsanzahl erweist sich eine Zusammenfassung zu vier Merkmalen pro Netz (drei mal zehn sowie die 15 „oberen“ benachbarten Output-Neuronen) als optimal. Im Falle der Verwendung

von Entropie zur Merkmalsbildung ergibt sich hier eine leicht schlechtere Erkennungsrate von 50.62 %. Abbildung 6.11 zeigt exemplarisch die Maximalwerte pro Zeitfenster der Output-Neuronen von lokalen Netzwerken der Größe 35 für die schon mehrfach verwendete Sprachäußerung. Deutlich lassen sich hier die Unterschiede der angenommenen Werte der Output-Neuronen innerhalb der einzelnen Netzwerke ausmachen.

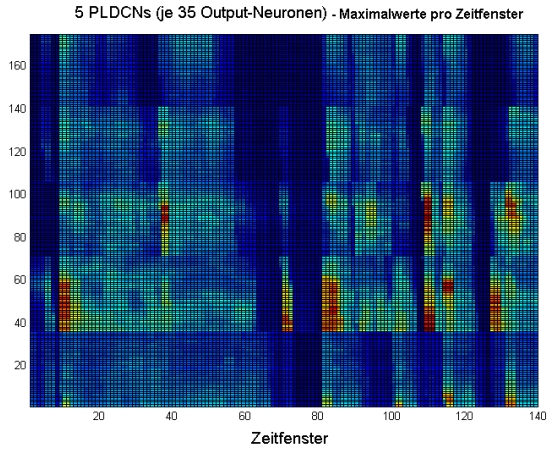


Abbildung 6.11: Exemplarische Merkmalsausprägungen von X^{Max} auf Basis von fünf PLDCNs der Größe 35 für die Sprachäußerung aus Abb. 6.3.

Es seien nun die verschiedenen Ansätze zur Merkmalsextraktion kombiniert: Als bisher beste Merkmalskombination im Sinne der erzielten Erkennungsraten stellte sich der Einbezug von sowohl Orts-Durchschnittsfeuerraten-, als auch Inter-Spike Intervall basierten Merkmalen im 22 dimensionalen Merkmalssatz $X_k^{OD/ISI,0.85}$ heraus. Dieser wird mit dem als am effektivsten identifizierten Delay-Computing Merkmalssatz (zehn PLDCNs der Netzwerkgröße 18, $f_{min} = 15$, Berechnung fensterweiser Maxima mit zwei Merkmalen pro lokalem Netz, insgesamt 20 Merkmalen) kombiniert und mit Hilfe von RHDA zu einen niedrigdimensionaleren Merkmalsvektor reduziert. Dies führt bei Parametern $\lambda = 1$ und $\gamma = 0.6$ und einem resultierenden 24 dimensionalen Merkmalsraum zu einer leichten Verbesserung der Erkennungsergebnisse auf 61.31 %. Eine Alternative zur Durchführung der RHDA Transformation und anschließenden Δ-Koeffizienten-Bildung, wie sie von Eisele u. a. (1996) vorgeschlagen wird, besteht darin, zunächst die Δ-Koeffizienten auf Basis des zusammengefassten Vektors der

Originalmerkmale zu berechnen und erst abschließend eine RHDA-Transformation durchzuführen. Dieses Vorgehen – implementiert für eine konstante Parameterwahl von $(\lambda, \gamma) = (1, 0.5)$, wie sie in den vorigen Untersuchungen gute Ergebnisse zeigte, und unterschiedliche resultierende Merkmalsgrößen, führt zu einer Test-Erkennungsrate von 61.91 %. Abbildung 6.12 zeigt die Test-Erkennungsrate für unterschiedlich große Merkmalsvektoren.

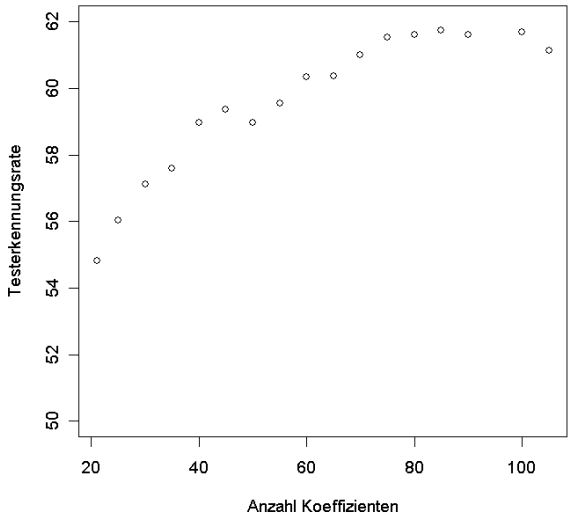


Abbildung 6.12: Test-Erkennungsrate für Merkmalssätze unterschiedlicher Dimension auf Basis von kombinierten Orts-Durchschnittsfeuerraten, Inter-Spike Intervall basierten und PLDCN basierten Merkmalen.

Die Verwendung von URHDA anstelle von RHDA führte zu ähnlich guten Ergebnissen (61.90%), bewirkte jedoch keine zusätzliche Verbesserung der Erkennungsraten. Aufgrund der hohen Rechenzeit und der – wegen der vier freien Parameter wachsenden Anzahl an zu evaluierenden Punkten eines Parametergitters – konnte hier jedoch keine umfassende Suche nach einer optimalen Parameterkombination vorgenommen werden. Es bleibt festzustellen, dass zumindest nicht auf Anhieb eine deutliche Verbesserung erreicht wird. Dieses Ergebnis ist in der Hinsicht erklärbar, dass das Erkennungsergebnis durch die am häufigsten vorkommenden Phoneme dominiert wird und damit die Wahl der Parameter λ und γ für diese Phonemklassen von entscheidender Bedeu-

tung ist. Eine umfassendere Untersuchung des Nutzens der Idee eines unbalancierten Ansatzes an einem weniger rechenintensiven Beispiel ist als Inhalt einer zukünftigen Studie denkbar.

6.3.9 LIN basierte Merkmalsextraktion

Abbildung 6.13 zeigt auf Basis von lateralen inhibitorischen Netzwerken erhaltene Merkmale: Deutlich sind auch hier die vorliegenden Frequenzen über den zeitlichen Verlauf zu identifizieren, wie sie bereits für die Merkmale X^{OD} oder X^{ALSD} zu erkennen waren.

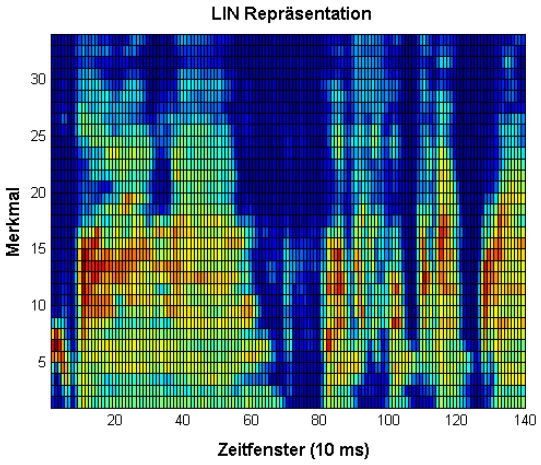


Abbildung 6.13: Exemplarische Merkmalsausprägungen von X_k^{LIN} für die Sprachäußerung aus Abb. 6.3.

Tabelle 6.7 zeigt die erzielten Erkennungsraten auf Basis von nach Seneff- bzw. Allen-Gruppierung von ANFs gebildeten PSTHs: Eine Kombination der Varianz-normierten

ANF-Gruppierung	Seneff	Allen
Erkennungsrate	56.93	57.85

Tabelle 6.7: Test-Erkennungsraten auf Basis von LINs.

Merkmale X_k^{LIN} und $X_k^{OD/ISI,0.85}$ mit der Gewichtung 0.1 zu 0.9 kann die erzielten

Erkennungsergebnisse weiter – allerdings nur leicht – verbessern auf 61.24 %. Wegen des geringen Performanceunterschieds wird im folgenden lediglich der kombinierte Merkmalsvektor $X^{OD/ISI,0.85}$ ohne weitere Berücksichtigung LIN basierter Merkmale verwendet.

6.3.10 Auditorisch erweiterte Merkmalsextraktion

Es werden nun zum Vergleich Melfrequenz Cepstralkoeffizienten (vgl. Abschnitt 3.3) als häufig verwendete Standardmerkmale (Young u. a., 2005, S. 61) zur automatischen Spracherkennung betrachtet. Tabelle 6.8 gibt die Test-Erkennungsraten bei unterschiedlichen Anzahlen mit vor der abschließenden Cepstraltransformation gebildeten Filterbankkoeffizienten (siehe Abschnitt 3.3) an. In der verwendeten Standardparametrisierung der MFCC-Merkmale werden 20 Koeffizienten verwendet. Zum Vergleich wurden auch MFCCs auf Basis von 24 bzw. 36 Koeffizienten im Frequenzbereich $[200, 6400]$ Hz berechnet, da diese Frequenzgruppen sich in den vorangegangenen Studien auf Basis des auditorischen Modells als effizient erwiesen haben. Beide Alternativen fallen hinsichtlich ihrer Ergebnisse allerdings leicht gegenüber den Standardwerten ab; insgesamt sind die Erkennungsergebnisse etwas niedriger als diejenigen des kombinierten Merkmalssatzes auditorisch-modellbasierter Merkmalsextraktion aus dem letzten Abschnitt. Um der Fragestellung nach unterschiedlicher Information

Anzahl Koeffizienten	20	24 (Allen)	36 (Seneff)
Test-Erkennungsrate	60.92	60.57	59.83

Tabelle 6.8: Test-Erkennungsraten auf Basis von MFCCs für unterschiedliche Anzahlen gebildeter Filterbankkoeffzienten vor Cepstraltransformation.

nachzugehen, werden MFCCs (mit 20 Filterbankkoeffizienten) und $X^{OD/ISI,0.85}$ (vgl. Abschnitt 6.3.7) einerseits sowie MFCCs mit $X^{OD/ISI,0.85}$ und zusätzlich X^{Max} (für PLDCNs der Größe 18 und $f_{min} = 15$) andererseits, zu *auditorisch erweiterten Merkmalssätzen (aeM)* kombiniert.
Wie sich im vorangegangen Abschnitt herausgestellt hat, werden bessere Ergebnisse erzielt, wenn die Δ-Koeffizientenbildung der RHDA-Transformation vorausgeht. Die kombinierten Merkmalssätze besitzen vor Dimensionsreduktion Dimensionen von

$(12+22)\cdot 3 = 102$ bzw. $(12+22+20)\cdot 3 = 162$. RHDA wird erneut für die Parameterkombination $(\lambda, \gamma) = (1, 0.5)$ und verschiedene Merkmalsdimensionen durchgeführt. Bereits ohne Verwendung der Delay-Computing Merkmale können die Ergebnisse deutlich verbessert werden auf 67.26 % (bei einer Merkmalsdimension von 95). Die zusätzliche Verwendung von PLDCN-Merkmalen erzielt eine weitere Verbesserung der Testerkenungsraten auf 67.82 % (Mekmalsdimension 90). Diese deutliche Verbesserung der Ergebnisse stützt die Vermutung, dass aus dem auditorischen Modell extrahierte Merkmale zusätzliche, zur Spracherkennung nützliche Information gegenüber der einfachen Verwendung von MFCCs enthalten.
Zur Berechung von MFCCs haben sich eine Reihe hilfreicher, vorangeschalteter Transformationsschritte etabliert (vgl. Young u. a., 2005, S. 56 ff). Einen solchen stellt die Verwendung von Hamming-Fensterung zur Berechnung der Kurzzeitspektren dar. Zur auditorisch-modellbasierten Merkmalsextraktion wurde diese nicht verwendet und zur Gewährleistung der Vergleichbarkeit der Merkmalssätze auch zur MFCC Berechung hier nicht eingesetzt. Verwendet man Hamming-Fensterung, verbessern sich die mit MFCCs allein erzielten Erkennungsraten auf 64.90 %. Die auf Basis des auditorisch erweiterten Merkmalssatzes erzielten Ergebnisse steigen nochmals an auf 68.39 % ohne- bzw. 68.72 % inklusive PLDCN-Merkmalen (siehe Tabelle 6.9).

	MFCCs	aeM	
Hamming		ohne PLDCN	mit PLDCN
nein	60.90	67.26	67.82
ja	64.90	68.39	68.72

Tabelle 6.9: Test-Erkennungsraten für auditorisch erweiterte Merkmalssätze.

6.3.11 Signifikanztests

In abschließenden Tests soll nun der Frage nach Signifikanz der in den vorangehenden Abschnitten beobachteten Performanceunterschiede nachgegangen werden. Es werden die folgenden vier Fragestellungen untersucht:

- Ergibt sich eine Verbesserung der Erkennungsergebnisse (am Beispiel der Orts-Durschnittsfeuerraten Merkmale nach Allen) durch die Verwendung von RHDA

(zu einem 21 dimensionalen Merkmalsvektor mit Regularisierungsparametern $\lambda = 1$ und $\gamma = 0.5$)?

- Ergibt sich eine Verbesserung der Erkennungsergebnisse für die Kombination von Orts-Durchschnittsfeuerraten (OD, nach Allen) und der Inter-Spike Intervall basierten durchschnittlichen lokalisierten Synchronizitätsdetektion (ALSD) (Kombination der standardisierten Merkmale mit Gewichtung 0.15 : 0.85) im Vergleich zur einfachen Verwendung von Orts-Durchschnittsfeuerraten Merkmalen?

- Ergibt sich durch die Erweiterung dieses kombinierten Merkmalssatzes um die Maximalwert basierten extrahierten Merkmale aus (zehn) parallelen lokalen Delay-Computing Netzwerken (PLDCN) der Größe 18 ($f_{min} = 15$, zwei aggregierten Merkmalen pro lokalem Netzwerk, Kombination mit RHDA aus dem konkatenierten Merkmalsvektor inklusive Δ-Koeffizienten zu einem 114 dimensionalen Merkmalssatz mit Regularisierungsparametern $\lambda = 1$ und $\gamma = 0.5$) ein besseres Erkennungsergebnis, im Vergleich zu der einfachen Merkmalskombination aus dem vorangehenden Test?

- Erzielt ein auditorisch erweiterter Merkmalssatz (aeM) von standard MFCCs (ohne Verwendung einer Hamming-Fensterung) um sowohl Delay-Computing Merkmale wie im letzten Test beschrieben als auch den gewichteten kombinierten Merkmalsvektor aus Orts-Durchschnittsfeuerraten und durchschnittlicher lokalisierter Synchronizitätsdetektion eine bessere Performance als MFCCs allein?

Tabelle 6.10 enthält die Testergebnisse auf Basis der in Abschnitt 6.2.3 beschriebenen Tests. Alle durchgeführten Tests zeigen signifikante Ergebnisse zu einem Niveau von 5%. Da mehr als ein Test durchgeführt wird ergibt sich ein multiples Testproblem. Eine konservative Korrektur nach Holm (1979) (vgl. z.B. Horn und Vollandt, 1995, S. 8) ergibt jedoch weiterhin, dass alle vier untersuchten Nullhypothesen gleicher Performance verworfen werden können.

Merkmalssatz 1	Merkmalssatz 2	p-Wert
OD	OD (RHDA)	0.0067
OD	OD + ISI	0.0046
OD + ISI	OD + ISI + PLDCN (RHDA)	0.0429
MFCC	MFCC + OD + ISI + PLDCN (RHDA)	$2.11 \cdot 10^{-5}$

Tabelle 6.10: Testerergebnisse für den Performanceunterschied verschiedener Merkmalssätze.

6.4 Zusammenfassung der Ergebnisse

In diesem Kapitel wurden verschiedene Arten der Merkmalsextraktion (vgl. Kapitel 3) aus dem Output des auditorischen Simulationsmodells (vgl. Kapitel 2) sowie anschließende Transformationen (vgl. Kapitel 5) anhand ihrer Ergebnisse bei der Verwendung als Front Ends zur automatischen Spracherkennung (vgl. Kapitel 4) miteinander verglichen. Tabelle 6.11 und Abbildung 6.14 fassen die Test-Erkennungsraten für die verschiedenen unterschiedlichen Merkmalssätze zusammen:

Merkmalssatz	%
OD	57.68
OD + AFCC	58.71
OD + RHDA	60.18
ISI	60.25
OD + ISI	61.11
DCN	44.91
PLDCN 35	51.41
PLDCN 18	53.35
LIN	57.85
OD + ISI + PLDCN 18 + RHDA	61.91
MFCC ohne Hamming	60.92
MFCC ohne Hamming + OD + ISI + PLDCN 18 + RHDA	67.82
MFCCs mit Hamming	64.90
MFCC mit Hamming + OD + ISI + PLDCN 18 + RHDA	68.72

Tabelle 6.11: Zusammenfassung der auf Basis der unterschiedlichen Merkmalssätze erhaltenen Ergebnisse.

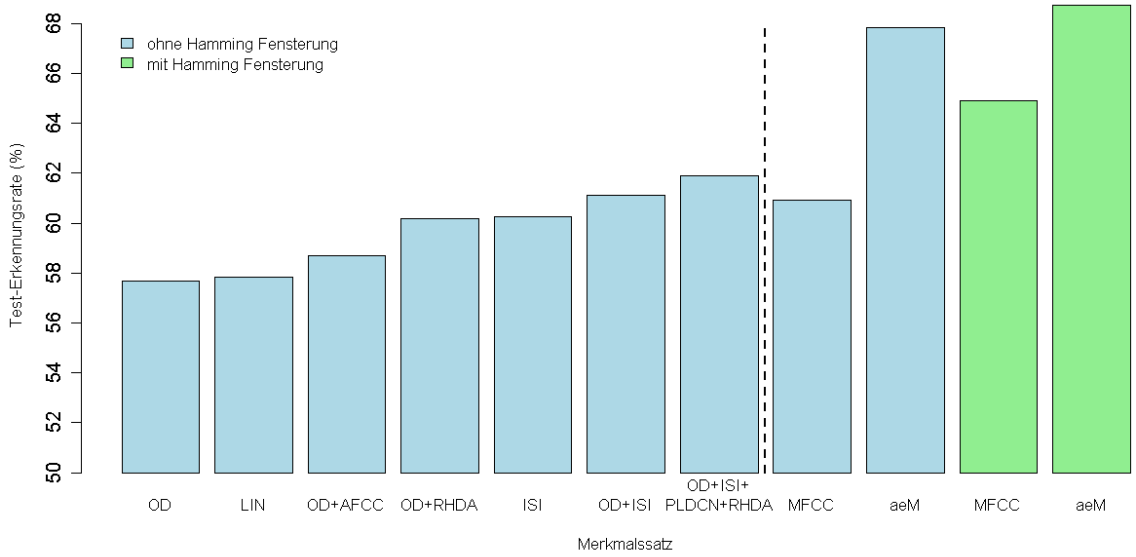


Abbildung 6.14: Zusammenfassung der Test-Erkennungsraten für unterschiedliche Merkmalssätze.

Vier der beobachteten Performanceunterschiede zwischen verschiedenen Merkmalssätzen wurden auf statistische Signifikanz untersucht; dabei konnten in allen Fällen signifikante Unterschiede der Erkennungsergebnisse für die verschiedenen Front Ends nachgewiesen werden.

Zur abschließenden Bewertung der Ergebnisse sei erwähnt, dass der Testdatensatz aus 7333 Phonemen in 39 verschiedenen Phonemklassen besteht. Die für standard MFCCs erzielten Ergebnisse entsprechen den von Young (1992) beobachteten Werten für eine ähnliche Back End Architektur. In den vergangenen Jahren sind zahlreiche Vorschläge zur Back End Optimierung entstanden; als Beispiel sei hierfür das Parametertying erwähnt (vgl. Kapitel 4). Diese fanden im Rahmen dieser Arbeit keine Berücksichtigung, da der Fokus auf der Untersuchung und dem Vergleich der alternativen Merkmalssätze liegt. Eine weitere Back End Optimierung ist aber auch für die vorgestellten Merkmale auf Basis des auditorischen Simulationsmodells vorstellbar und naheliegend. Die T9 Technologie aus der Mobilfunktechnik ist ein anschauliches Beispiel dafür, dass durch eine Tastenbelegung von drei Buchstaben je Taste

(vergleichbar einer Erkennungsrate 1/3) bereits eine gute Worterkennung möglich ist, wenn man die Menge möglicher Alternativen durch Verknüpfung mit einem Wörterbuch einschränkt und wäre für eine praktische Anwendung erforderlich. Eine mögliche Erweiterung um die Detektion semantischer *ConTexte* wird beispielsweise von Bordag und Bordag (2003) beschrieben.

7 Zusammenfassung

Am Fallbeispiel automatischer Spracherkennung wurde die auditorische Informationsübermittlung anhand unterschiedlicher Merkmalsextraktionsprinzipien untersucht. Insbesondere die Kombination verschiedener Merkmalssätze erwies sich hierbei als effizient. Vor allem die neu vorgeschlagene RHDA stellte sich in diesem Zusammenhang als effektives Mittel zur Dimensionsreduktion von kombinierten Merkmalsvektoren heraus: Auf Basis der Kombination von sowohl *Orts-Durchschnittsfeuerraten* als auch *Phase Locking* basierten Merkmalen konnte eine deutliche Verbesserung der Erkennungsergebnisse erzielt werden, im Vergleich zur jeweiligen Verwendung der einzelnen Merkmalssätze allein.

Diese Ergebnisse legen nahe, dass Untersuchungen zur Kodierung von Phase Locking Information auch für zukünftige Entwicklungen zur Ansteuerung von Cochlea Implantaten (CI) eine wichtige Rolle spielen könnten. Eine Stimulation der Hörnerven durch Phase Locking Information könnte einen Weg in die Richtung der bis heute schwierigen Ansteuerung niedrigerer Frequenzen (in der letzten Windung der Cochlea) ebnen und damit auch helfen, dem Ziel, Cochlea Implantat Trägern das Musik hören zu ermöglichen, näher zu kommen (vgl. auch Grayden u. a., 2004; Mc Dermott, 2004).

In den Experimenten dieser Arbeit stellte sich als signifikant heraus, dass auf Basis von *parallelen lokalen Delay-Computing Netzwerken* exrahierte Merkmale ergänzende Information im Vergleich zu frequenzbasierten Standard-Merkmalsextraktions-Prinzipien beinhalten.

Die guten, unter Verwendung von nicht auf dem auditorischen Simualtionsmodell basierenden *MFCCs* erzielten Ergebnisse stützen die Behauptung, dass die dort vorgenommene Logarithmierung des Spektrums bereits implizit einen Aspekt der menschlichen Wahrnehmung von Schall nachahmt (Perdigao und Sa, 1998).

Andere Prozessschritte der auditorischen Signalverarbeitung sind jedoch nicht durch

MFCCs nachempfunden. Als Beispiel sei hier die Adaption (vgl. Abschnitt 2.5) erwähnt, die als eine der Hauptursachen für Maskierungseffekte angesehen wird. Ivanov und Petrovsky (2004) betonen die Bedeutung von Maskierung in der menschlichen Spracherkennung. Harczos u. a. (2007a) konnten in einer Untersuchung auf Basis von Orts-Durchschnittsfeuerraten, extrahiert an unterschiedlichen Stellen der auditorischen Verarbeitungskette, jedoch keinen positiven Effekt der hinteren Verarbeitungsschritte der auditorischen Signalverarbeitung auf das Erkennungsergebnis feststellen, die für das Phänomen der Adaption verantwortlich zeichnen (vgl. Abschnitt A.5). Diese Resultate legen nahe, dass die Modellierung der Neurotransmitteremission nach Sumner u. a. (2002) wohlmöglich noch eine der Schwachstellen des neurophysiologischen Simulationsmodells darstellt.

Einzelne Aktionspotenziale stellen zwar den Mechanismus dar, der dem Nervensystem zur Informationsübertragung zur Verfügung steht. Durch die hohe Anzahl an Hörnerven kommt dem exakten Zeitpunkt eines einzelnen APs jedoch nur eine geringe Bedeutung zu (vgl. hierzu z.B. die Arbeiten zur Modellierung von Bushy Cells von Werner und Fodroczi, 2006, am Fraunhofer IDMT, Ilmenau). In den durchgeführten Studien zur automatischen Spracherkennung konnte kein positiver Effekt einer präzisen Nachmodellierung einzelner Aktionspotenzial-Zeitpunkte beobachtet werden. Für weiterführende Arbeiten scheint eine direkte Modelierung von *zeitabhängigen Feuerraten* durch die Neurotransmitteremissionsraten, gekoppelt mit einer anschließenden Tiefpassfilterung zur Berücksichtigung von durch Refraktärzeiten entstehenden Effekten als genügend und dabei rechenzeiteffizienter.

Die abschließende Kombination von MFCCs mit auditorisch basierten Merkmalen zu einem *auditorisch erweiterten Merkmalssatz (aeM)* bewirkte eine deutliche, signifikante Verbesserung der Erkennungsergebnisse, die als Nachweis angesehen werden kann, dass die aus dem auditorischen Simulationsmodell gewonnenen Merkmale zusätzliche, zur Spracherkennung nützliche Information enthalten.

A Verarbeitungsschritte im auditorischen Simulationsmodell

A.1 Übersicht

In Abschnitt 2.2 ist die Anatomie des Ohrs beschrieben. In diesem Abschnitt erfolgt eine detaillierte Beschreibung der einzelnen Prozessschritte des Simulationsmodells. Eine übersichtliche Zusammenfassung der auditorischen Signalverarbeitungskette findet sich in Szepannek u. a. (2005). Nach dem die (mechanische) Schallwelle zunächst am Trommelfell auf die Cochlea übertragen wird, erfolgt in den entlang der Basilarmembran angeordneten inneren Haarzellen eine Umwandlung der mechanischen Welle in elektrische Impulse (Aktionspotenziale). Hierbei wird die folgende Prozesskette durchlaufen:

Abbildung A.1: Schematische Darstellung der Signalverarbeitungs-Prozessschritte im Innenohr.

A.2 Außen-, Mittelohr, Basilarmembranauslenkung und äußere Haarzellen

Der erste Modellierungschritt umfasst den Effekt von Außen- sowie Mittelohr, die Hydromechanik der Cochlea, sowie denjenigen der äußeren Haarzellen (OHCs). Die in der Lymph-Flüssigkeit mitschwingende Basilarmembran (BM) wird durch ihre über

die Länge unterschiedlichen Durchmesser und Steifigkeiten an verschiedenen Positionen, von verschiedenen Frequenzen unterschiedlich stark angeregt. Dabei verhält sich die BM wie ein Bandpassfilter. In diesem Zusammenhang wird auch von der Frequenzselektivität der BM gesprochen. Bis zu einer Frequenz von etwa 1000 Hz verhält sich die Frequenzauflösung in etwa linear, darüber steigt sie logarithmisch an. Aus der Psychoakustik sind aus diesem Grund die Mel- bzw. Bark-Skala bekannt (siehe Anhang B.5 oder auch Schukat-Talamazzini, 1995, S. 41 f)).
Entlang der BM befinden sich ca. 3500 innere Haarzellen (IHCs), die bandpass-gefilterte Signale des eingehenden Schalls um eine Mittenfrequenz – je nach ihrer Position (hohe Frequenzen zu Beginn am basalen -, niedrige Frequenzen an ihrem apikalen Ende) - in Richtung Gehirn weiterleiten.
Im verwendeten Modell wird die Basilarmembran nach Baumgarte (2000) in 251 gleich große Sektionen unterteilt, wobei gleich groß bedeutet, dass der Frequenzabstand der Mittenfrequenzen (Center Frequencies, CFs) zweier aufeinanderfolgender Sektionen immer 0.1 Bark beträgt. Eine *Mittenfrequenz* bezeichnet hierbei diejenige Frequenz eines Schallsignals, die die Basilarmembran an der betrachteten Position maximal in Schwingungen versetzt. Für die Mittenfrequenzen der modellierten 251 Sektionen entlang der BM ergibt sich folgende Gleichung (siehe Baumgarte, 2000, S. 148):

$$\begin{aligned} CF_1 &= 5\,Hz \\ CF_2 &= 10\,Hz \\ CF_i &= CF_{i-1} + 0.1 * (25 + 75 * (1 + 1.4 * (CF_{i-1}/1000)^2)^{0.69}), i \geq 3 \end{aligned} \quad \text{(A.1)}$$

entsprechend einem Mittenfrequenzbereich von 0.05 bis 25 Bark bzw. 5 Hz bis etwa 20 kHz (siehe Abbildung A.2). Für das verwendete Simulationsmodell wählt Baumgarte (2000) zunächst aus mehreren bekannten Cochlea-Modellen dasjenige von Zwicker und Peisl (1990), wegen seiner Eigenschaft das aktive, nichtlinear signalverstärkende Verhalten der äußeren Haarzellen im Zeitbereich realistisch zu modellieren (Baumgarte, 2000, S. 42 f). Dieses Modell wird um eine weitere, zweite OHC-Verstärkerstufe über einen Parallelschwingkreis ergänzt, die zwar physiologisch nicht unmittelbar rechtfertigbar ist, jedoch nötig um in der Realität gemessene Verstärkungen von 60 dB zu ermöglichen (Baumgarte, 2000, S. 47).
Es ist möglich, das hydromechanische Modell der lokalen Nachgiebigkeit, Masse und

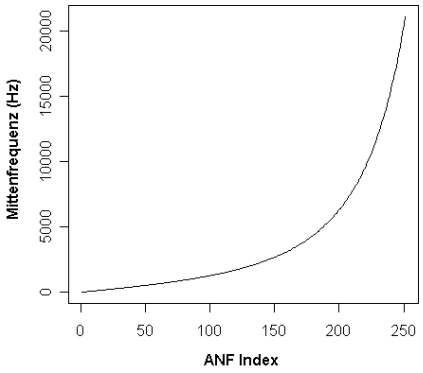


Abbildung A.2: Frequenzauflösung entlang der Basilarmembran.

Reibungsverluste der BM durch elektromechanische Ersatz-Schaltbilder in Serie geschalteter Sektionen longitudinaler Abschnitte der als eindimensional betrachteten Cochlea zu ersetzen. Hierzu wird eine Kraft-Strom Analogie verwendet, in der die Geschwindigkeit der Bewegung des ovalen Fensters durch die Eingangspannnung (vgl. Abbildung A.3, u_{of}) modelliert wird (siehe Baumgarte, 2000, S. 37 ff). Ströme entsprechen Kräften, Induktivitäten spiegeln die örtliche Nachgiebigkeit der BM wieder, Kapazitäten stehen für Massen und Widerstände für Reibungsverluste des schwingenden Systems. Zur Simulation wird die Methode der *Wellendigitalfilter* (Fettweis, 1986) verwendet, die im Gegensatz zu numerischen Integrationsmethoden Stabilität auch für einen hohen Dynamikbereich gewährleistet und nur geringe Empfindlichkeit gegenüber Toleranzen der Filterkoeffizienten besitzt (vgl. Baumgarte, 2000, S. 43). Auf diese Weise kann die Frequenzselektivität der BM aus einem Schallsignal technisch umgesetzt werden. Das Wirken der resultierenden Bandpassfilter entlang der Basilarmembran wird in Abbildung A.4 veranschaulicht. Die Filter besitzen eine sehr ähnliche Gestalt entlang der Bark-Skala (Abbildung rechts, oben und unten für je verschiedene Schalldrucklevel eines eingehenden Signals unterschiedlicher Frequenz), d.h. sie sind in etwa logarithmisch gestaucht für Frequenzen oberhalb von 1 kHz (vgl. Abbildung A.2). In Richtung der Frequenzen unterhalb der CF fallen die Filteramplituden stärker ab. Abbildung A.5 zeigt die Auslenkung der Basilarmembran des implemen-

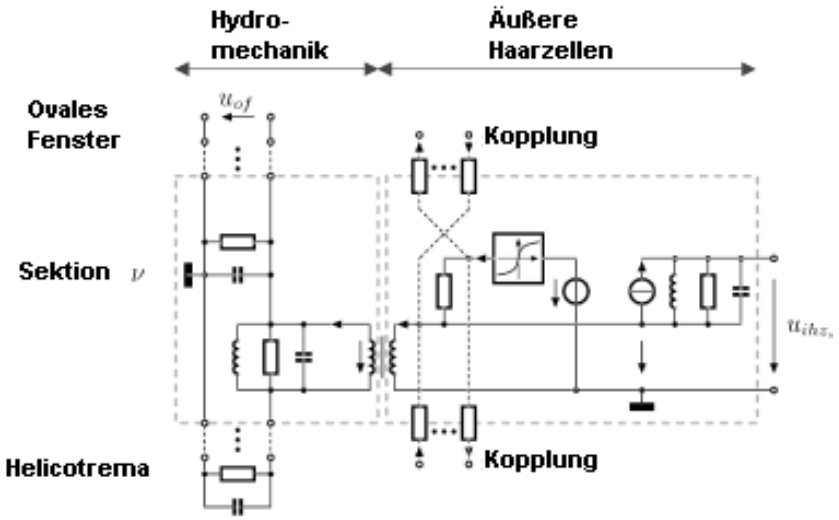


Abbildung A.3: Modellschaltbild eines Abschnitts der Cochlea (entnommen aus Heinz, 2002, S. 53).

tierten Simulationsmodells. Sie liegt zwischen plus- (türkis) und minus 80 nm (rot). An der Ordinate sind die 251 modellierten Sektionen der Basilarmembran abgetragen. Die Abszisse repräsentiert die Zeit. Auf den ersten Blick sind die Trajektorien der *Cochlea Wanderwelle* zu erkennen. Weiterhin ist augenscheinlich, dass die Periodizität des eingehenden Signals auf die Basilarmembran übertragen wird. Es ist auch zu erkennen, das die Amplitude der übertragenen Schwingung an unterschiedlichen Positionen der BM verschieden stark ist. Dies ist nochmals in Abbildung A.6 illustruiert. Hier ist die Reaktion der simulierten Basilarmembran auf einen 45 ms langen Sweep Ton, d.h. ein Schallsignal bestehend aus einem einzelnen Ton, mit einer, im Laufe der Signaldauer ansteigenden Frequenz, abgebildet.

Zur Generierung von neuronalen Aktionspotenzialen aus der simulierten BM-Bewegung wird anschließend die mechano-elektrische Signalumwandlung in den inneren Haarzellen modelliert. Die Modellierung der IHCs erfolgt in weiten Teilen entsprechend dem Modell von Sumner u. a. (2002), einer der aktuellsten Arbeiten auf diesem Gebiet. Die einzelnen verwendeteten Verarbeitungschritte, sowie Modifikationen, Erweiterungen und Untersuchungen werden in den folgenden Abschnitten beschrieben.

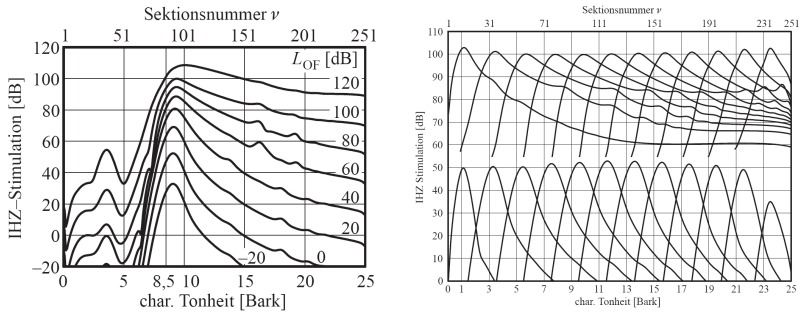


Abbildung A.4: Bandpass-Signalfilterung entlang der Basilarmembran, links: Geschwindigkeit $v(t)$ der BM (siehe Gleichung A.2) für eine eingehende 1000 Hz Sinusschwingung unterschiedlicher Intensität an verschiedenen Positionen der BM, wobei L_{OF} die Wellenamplitude am ovalen Fenster darstellt. Rechts: Gestalt der Filter für 0 (unten) und 100 dB Eingangssignalamplituden (oben) entlang der BM in 2 Bark-Abständen. Beide Abbildungen entnommen aus Baumgarte (2000).

A.3 Stereozilienauslenkung

Um die einheitslose Ausgabe des Wellendigitalfiltermodells aus dem Basilarmembran-Simulationsmodell von Baumgarte in einem neurophysiologisch parametrisierten Modell der inneren Haarzellen weiterverarbeiten zu können, ist es erforderlich, diese auf die Größe der Auslenkung der Membran umzuparametrisieren. Dies erfolgt durch eine Skalierung der Filteroutputwerte mit Hilfe einer multiplikativen Kopplungskonstante (siehe hierzu auch Anhang B.2). Am oberen Ende der inneren Haarzellen, auf der Basilarmembran, befinden sich drei Reihen feiner Härchen, die als *Stereozilien* bezeichnet werden und mit Flüssigkeit und BM mitschwingen. Durch ihre Auslenkung öffnen und schließen sich Ionenkanäle, die den Einfluss positiv geladener K^+ Ionen in die IHC regulieren.
Die Auslenkung der *Stereozilien* $u(t)$ erfolgt nach Shamma u. a. (1986) in Abhängigkeit von der Basilarmembrangeschwindigkeit $v(t)$ durch die Gleichung

$$\tau_c \frac{\mathrm{d}u(t)}{\mathrm{d}t} + u(t) = \tau_c C_{Stereozilien} v(t). \tag{A.2}$$

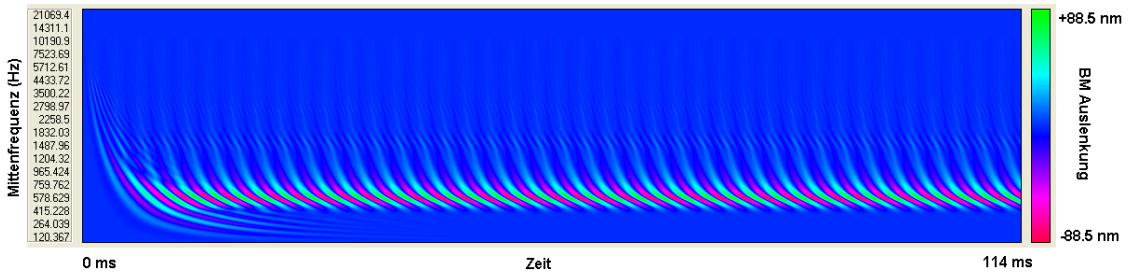


Abbildung A.5: Output des Simulationsmodells an der Basilarmembran für einen Sinuston von 440 Hz.

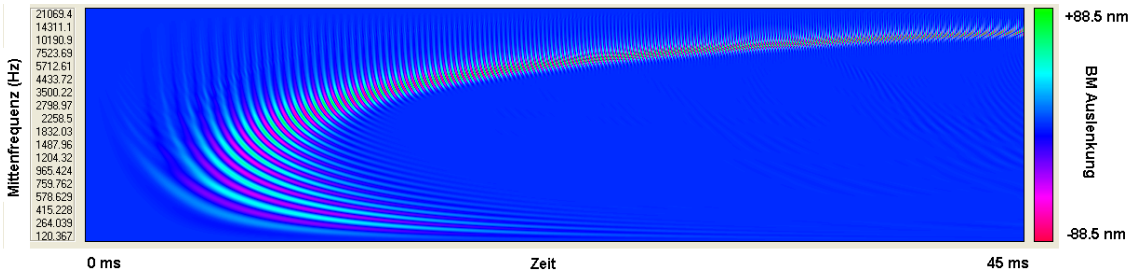


Abbildung A.6: Output des Simulationsmodells für einen Quicksweep Ton von 45 ms mit ansteigender Frequenz.

Dabei stellen τ_c eine Verzögerungskonstante und t die Zeit dar. Die verwendeten Werte sämtlicher Modellparameter befinden sind in Anhang B.4 aufgelistet. Die Stereozilien bewegen sich damit für niedrige Frequenzen in Phase mit der Geschwindigkeit der BM und für hohe Frequenzen phasenverschoben (Sumner u. a., 2002).

Neben der systematischen, signalinduzierten Wanderwellenbewegung der cochleären Flüssigkeit unterliegt diese in der Realität zudem geringen thermischen Fluktuationen, die physikalisch gemein durch eine Brown'sche Bewegung beschreibbar sind. Für deren Auswirkungen auf die Auslenkung der Stereozilien dagegen muss zusätzlich die durch die Steifigkeit der Sterozilien hervorgerufene Dämpfung berücksichtigt werden, deren Fußende an der IHC fixiert ist. Nach Gebeshuber (2000) ergibt sich für die Modellierung ein *Ornstein-Uhlenbeck Prozess* der Gestalt

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} = -\lambda u(t) + \lambda \epsilon(t). \tag{A.3}$$

$u(t)$ stellt den Prozess dar, $\epsilon(t)$ repräsentiert weißes Rauschen und aus λ ergibt sich eine Zeitkonstante $\tau = \lambda^{-1}$, die die Dämpfung berücksichtigt. Für τ gibt Gebeshuber (2000) einen Wert von $\tau = 10^{-15}$ s an. Dieser Wert ist sehr klein, bezogen auf die Abtastrate des Simulationsmodells von 44.1 kHz. Die Korrelation des Ohrnstein-Uhlenbeck Prozesses zwischen zwei Zeitpunkten t_1 und t_2 ergibt sich nach Longtin (1993) als

$$C(u(t_1), u(t_2)) = e^{(-\frac{t_2 - t_1}{\tau})}. \tag{A.4}$$

Durch die extrem kurze Zeitkonstante τ ergibt sich eine Korrelation von 0.01 für etwa $4.6 * 10^{-15}$ s. Dieser Wert liegt etliche Größenordnungen unter der gewählten Abtastrate von 44100 s^{-1}, so dass für diese Unkorreliertheit zwischen aufeinanderfolgenden Beobachtungen angenommen werden kann und damit die Modellierung der Stereozilienfluktuation durch ein einfaches additives *weißes Rauschen* möglich ist.
Die Menge der Hörnerven lässt sich unterteilen in drei unterschiedliche Typen je nach der Art ihrer spontanen Aktivität. Mit *spontaner Aktivität* wird das Auslösen von Aktionspotenzialen auch ohne zugrunde liegenden Stimulus bezeichnet. Als *LSR* (low spontaneous rate) Nerven werden diejenigen ANFs bezeichnet, die durchschnittlich weniger als 0.5 Spikes/s emittieren, ohne angeregt zu werden. Für *MSR* (medium spontaneous rate) ANFs liegt dieser Wert zwischen 0.5 und 18 Spikes/s, für *HSR* (high spontaneous rate) ANFs $>$ 18 Spikes/s (vgl. Abbildung A.7). Diese Einteilung ist beispielsweise zu finden in Geisler (1998). HSR, MSR und LSR teilen sich etwa auf im Verhältnis 60% : 25% : 15%. Die verschiedenen Nerventypen zeigen eine unterschiedliche Reaktion auf unterschiedliche Dynamikbereiche des eingehenden Schalls: Während die HSR Hörnerven bereits durch schwache (leise) Signale aktiviert werden, aber früh saturieren bezüglich ihrer durchschnittlichen Spikerate, zeigen die LSR Nerven auf sehr schwache Signale gar keine Antwort, dafür aber Unterschiede auch noch für sehr laute Signale. Abbildung A.7 zeigt die Histogrammverteilung der durchschnittlichen spontanen Feuerraten der ANFs. Der Effekt des thermischen Rauschens liegt bei sehr leisen Signalen im Bereich der Wahrnehmungsschwellen. Exemplarisch veranschaulicht ist er in Abbildung A.8. Sei vereinfacht das postsynaptische Potenzial am Hörnerven eine eins-zu-eins Übertragung einer einfachen Sinusschwingung (linkes Bild), bei der die Schwelle zum Auslösen von Aktionspotenzialen (horizontale Linie) jedoch nicht erreicht wird. Durch das additive weiße Rauschen werden zufällige Spikes

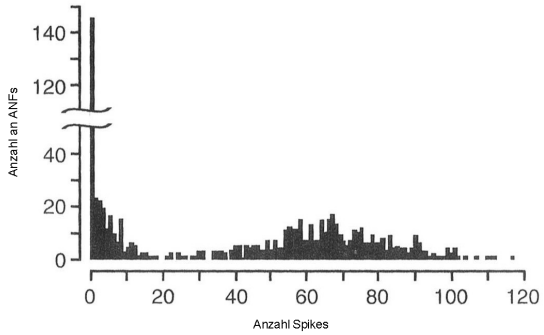


Abbildung A.7: Verteilung spontaner Spikeraten an den Hörnerven (aus Geisler, 1998).

ausgelöst (rechte Abbildung), die jedoch der Periode folgen. Dieser Effekt führt zu einer Herabsetzung der Wahrnehmungsschwellen und damit einer Erweiterung des Dynamikbereichs (vgl. Gebeshuber, 2000).

Für das vorliegende Simulationsmodell konnte jedoch ein solcher Effekt nicht beobachtet werden: Es wurden Simulationen für niedrigfrequente Stimuli von 441 Hz bei unterschiedlicher Intensität, im Lautstärkebereich abnehmender Synchronisation (quantifiziert durch den *Synchronization Index*, SI, vgl. Abschnitt B.2) durchgeführt. Für varriierende Varianz eines additiven Rauschens (zwischen 1/100 und der maximalen Stereozilienauslenkung) konnte dabei keine Zunahme des SI beobachtet werden. Im Gegenteil: eine weitere Erhöhung der Varianz des künstlich addierten Rauschens führte zu einem Synchronisationsverlust auch für höhere Intensitätslevel. Aus diesem Grund wird im Simulationsmodell dieser Arbeit auf die zusätzliche Simulation der thermal bedingten Stereozilienfluktuation verzichtet.

A.4 Modellierung des IHC-Potenzials

Die Auslenkung der Stereozilien bewirkt verstärkten Zufluss von positiv geladenen Kalium-(K^+)Ionen in die IHCs. Am oberen Ende der IHCs befinden sich drei Reihen von Stereozilien, die durch sogenannte *Tip Links* an ihren oberen Enden miteinander

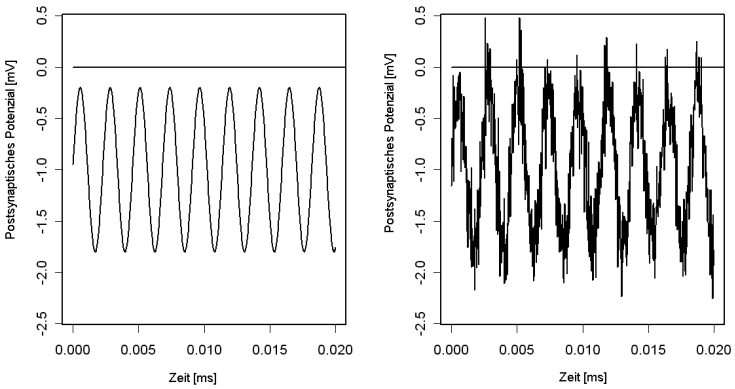


Abbildung A.8: Veranschaulichung der Wirkung von additivem weißen Rauschen.

verbunden sind (Hudspeth, 1989). Sie öffnen und verschließen Ionenkanäle, je nach Richtung der Auslenkung der Sterozilien-Haarbündel.
Nach Johnston und Wu (1995), Kapitel 10, lässt sich das Verhalten solcher Ionenkanäle durch *Chapman-Kolmogorov Differentialgleichungen* beschreiben. Dabei wird davon ausgegangen, dass sich die Ionenkanäle in einem von drei unterschiedlichen Zuständen befinden. Die Zustandswechsel der Kanäle werden als Markovprozesse aufgefasst. Diese Annahme deckt sich mit Messergebnissen für Haarzellen von Schildkröten von Crawford u. a. (1991), die die Öffungsdauern der Ionenkanäle als exponentialverteilt beschreiben. Im Falle von inneren Haarzellen schlagen Mountain und Cody (1999) ein drei-Zustands-Modell vor, das die Messergebnisse besser nachempfindet als ein Modell mit nur zwei Zuständen (siehe auch Markin und Hudspeth, 1995):

$$Close_1 \underset{B}{\overset{A}{\rightleftharpoons}} Close_2 \underset{D}{\overset{C}{\rightleftharpoons}} Open \tag{A.5}$$

Vom $Close_1$ Zustand aus kann sich ein Ionenkanal nicht direkt öffnen. A bis D entsprechen Ratenkonstanten für die Zustandsübergänge aus deterministischen Modellen. Für das Chapman-Kolmogorov Modell gilt:

$$P_{Close_1,Close_2}(\Delta) := P(Close_1(t) \rightarrow Close_2(t+\Delta)) = A\Delta + o(\Delta). \tag{A.6}$$

Entsprechend bezeichnen B, C und D die Übergänge $Close_2 \rightarrow Close_1$, $Close_2 \rightarrow$

Open und *Open* $\rightarrow$ $Close_2$. Für $\Delta \rightarrow 0$ beträgt also die Wahrscheinlichkeit für mehr als einen Zustandswechsel 0. Bezeichne $Q := \lim_{\Delta \rightarrow 0} \frac{P_{ij}(\Delta) - I_{\{i\}}(j)}{\Delta}$, mit $I_{\{.\}}(\cdot)$ der Indikatorfunktion, die infinitesimale Matrix zur Beschreibung der Zustandsübergangswahrscheinlichkeiten, dann gilt (analog zu Johnston und Wu, 1995, S. 255):

$$Q = \begin{pmatrix} -A & A & 0 \\ B & -(B+C) & C \\ 0 & D & -D \end{pmatrix} . \tag{A.7}$$

Für die Matrix $P(\Delta)$ der Zustandsübergangswahrscheinlichkeiten $P_{ij}(\Delta)$ gilt die *Chapman-Kolmogorov Gleichung.*

$$\frac{dP(\Delta)}{d\Delta} = P(\Delta)Q \tag{A.8}$$

Für eine Anwendung im Simulationsmodell ist jedoch Kenntnis der Parameter A bis D, als Funktion vom aktuellen Auslenkungsgrad der Stereozilien nötig. Hierfür sind jedoch keine Größen bekannt. Nach MacQueen (1996) erfolgt die Antwort der Ionenkanäle auf Haarzellenauslenkung zudem extrem schnell; aus diesem Grund wird für das verwendete Simulationsmodell eine unmittelbare Änderung der Ionenkanäle in Abhängigkeit von der Position der Stereozilien verwendet, die auf deren Arbeit zurückgeht. Modelliert wird dabei die aus geöffneten Ionenkanälen unmittelbar resultierende Leitfähigkeit $G(u)$ als Anteil einer größtmöglichen Leitfähigkeit $G^{max}_{Stereozilien} + G_a$ durch eine drei-Zustands Boltzmann-Funktion (vgl. Sumner u. a., 2002):

$$G(u(t)) = \frac{G^{max}_{Stereozilien}}{1 + e^{\frac{u(t)-u_0}{s_0}}(1 + e^{\frac{u(t)-u_1}{s_1}})} + G_a. \tag{A.9}$$

$u(t)$ stellt dabei die Auslenkung der Stereozilien aus dem vorangehenden Abschnitt dar. G_a ist die konstante passive Leitfähigkeit der Membran. u_0, u_1, s_0 und s_1 sind Konstanten, die den sigmoiden Verlauf der Funktion charakterisieren und im Anhang B.4 aufgelistet sind. Der Zusammenhang ist veranschaulicht in Abbildung A.9 (links). Als Effekt ergibt sich eine nichtlineare Verzerrung des an der BM vorliegenden Signals: Für geringe Amplituden ist ein starker Anstieg der Leitfähigkeit zu beobachten, für große Amplituden stellt sich Saturierung ein. Es wird auch von einer *Halbwellengleichrichtung (engl: half-way rectification)* eingehender Sinussignale gesprochen, da für negative Auslenkungen bereits früh Saturierung einsetzt. Diese ist in Abbildung A.10 am Beispiel einer sinusförmigen Schallwelle von 441 Hz veranschaulicht. Perdigao und Sa

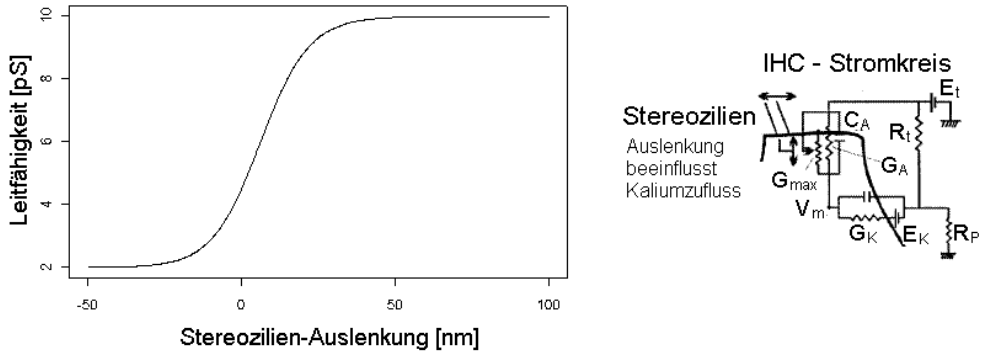


Abbildung A.9: Funktionaler Zusammenhang zwischen Auslenkungsposition der Sterozilien und Ionenzufluss in die IHCs (links) und schematische Darstellung des Stromkreises, der das Potenzial der IHC beschreibt (rechts, angelehnt an Sumner u. a., 2002, es gilt $E_k^* = E_k + E_t \frac{R_p}{R_p + R_t}$).

(1998) sprechen wegen dieser Saturierung von einer impliziten *Pseudo-Logarithmierung* des Signals durch den menschlichen Hörapparat, wie sie beispielsweise in den populären MFCC Merkmalen (siehe Abschnitt 3.3) zur automatischen Spracherkennung eingesetzt wird (siehe Kapitel 3).
Das Potenzial $V(t)$ innerhalb der inneren Haarzellen lässt sich nun entsprechend Shamma u. a. (1986) modellieren durch einen einfachen Stromkreis (vergleiche Abbildung A.9, rechts). Es gilt die Differentialgleichung

$$C_m \frac{dV(t)}{\delta t} + G(u)\left((V(t) - E_t) + G_k(V(t) - E_k^*)\right) = 0. \tag{A.10}$$

Die Stromstärke ist dabei gegeben durch die Spannungsdifferenz zwischen dem Potenzial der IHC $V(t)$ und den konstanten (da wesentlich größervolumigen) Flüssigkeiten außerhalb der IHC, E_t stellt hierbei das deutlich höhere *endocochleäre Potenzial* dar. Durch geöffnete Ionenkanäle der Stereozilien fließt positive Ladung in die (negativ geladene) IHC ein und führt zu einer *Depolarisierung* der IHC. Die Stärke der Depolarisierung hängt von $G(u(t))$, also der Anzahl geöffneter Ionenkanäle, ab. Durch die *basolateralen Ionenkanäle* herrscht ein stetiger (negativer) Ionenfluss mit konstanter Leitfähigkeit G_k in die IHC. E_k^* bezeichnet das basolaterale *Nernst Potenzial*. C_m ist die Kapazität der inneren Haarzelle.

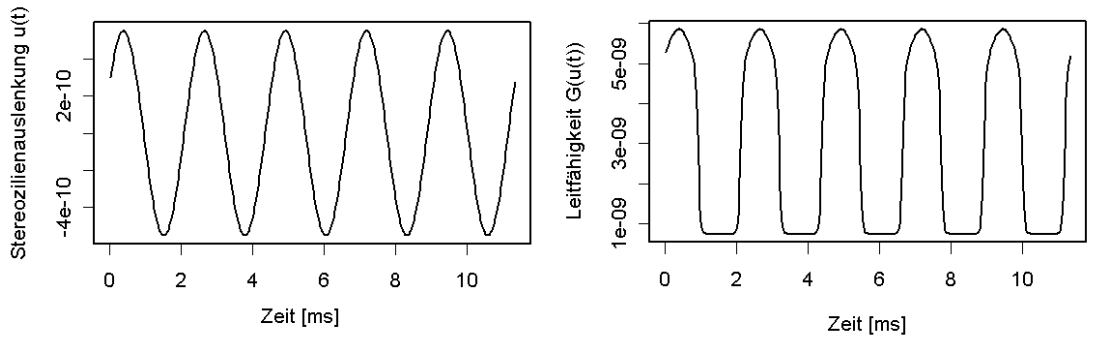


Abbildung A.10: Auslenkung der Stereozilien ($u(t)$, links) und Leitfähigkeit ($G(u(t))$, rechts) der Ionenkanäle für einen 441 Hz Sinuston.

Einen alternativen Ansatz zur Modellierung der IHC-Spannung stellen Rattay und Gitter (1998) vor. Dort wird der gleiche Stromkreis modelliert wie im oben beschriebenen Modell von Shamma u. a. (1986). Ein Unterschied liegt jedoch in der Modellierung des Ionenzuflusses am apikalen Ende der Sterozilien. Es wird angenommen, dass jede IHC 60 Ionenkanäle besitzt. Bei jeder Auslenkung der Stereozilien um weitere vier nm öffnet sich je ein zusätzlicher Ionenkanal. Von den Kanälen wird angenommen, dass diese für eine feste Zeit im geöffneten Zustand verharren und sich danach wieder schließen. Pro Kanal wird eine feste Leitfähigkeit $G_{channel}$ angegeben. Auf diese Weise lässt sich aus der momentan geöffneten Anzahl an Ionenkanälen die Gesamtleitfähigkeit $G(u(t))$ (siehe Gleichung A.9) zur Modellierung des Stromkreises bestimmen. Im Ruhezustand der IHC werden 15% offener Kanäle angenommen. Rattay und Gitter (1998) präsentieren Ergebnisse für Kanalöffnungsdauern von 0.1 ms, 1 ms und 18 ms. Nach Messungen von Crawford u. a. (1991) liegt die mittlere Öffnungszeit der Kanäle in Haarzellen von Schildkröten bei etwa 1.1 ms, so dass Zeiten von 18 ms als unrealistisch einzuschätzen sind. Eigene Simulationen für unterschiedliche Frequenzen zeigten die besten Resultate für mit der Position der BM – und damit der mit der CF – variierende Kanalöffnungsdauern von etwa $10/CF$ wobei CF die Mittenfrequenz der betrachteten IHC darstellt (vgl. Abbildung A.11).

Die auf diese Weise modellierte Abhängigkeit der Leitfähigkeit $G(u(t))$ von der Auslenkung der Basilarmembran ist linear. Ein Modell für das Saturierungsverhalten der

Anzahl offener Kanäle bei starker Auslenkung der Stereozilien wird von Rattay und Gitter (1998) nicht angegeben. Auch deckt sich der Wert von 4 nm zur Öffnung eines Ionenkanals nicht mit denjenigen Werten, die sich aus anderen Modellen ergeben (Crawford u. a., 1991; Sumner u. a., 2002) sondern liegt deutlich höher. Abbildung A.11 zeigt den Verlauf des IHC-Potenzials für Sinusschwingungen von 500 Hz (links) und 2000 Hz (rechts) bei unterschiedlichen angenommenen Kanalöffnungsdauern. Man erkennt gut die *Halbwellengleichrichtung* der zugrunde liegenden Schwingung. Die horizontale Linie stellt den Schwellwert zur Aktionspotenzialgenerierung dar wie ihn Rattay und Gitter (1998), unter der Vernachlässigung der weiteren Prozessschritte, angeben. Bereits bei einer Frequenz von zwei kHz kann bei dieser Art der Modellierung das Membranpotenzial nicht mehr „effektiv“ der ursprünglichen Schwingung folgen: Für kleine Öffnungszeiten wird der Schwellenwert nie erreicht; bei längeren Ionenkanal-Öffnungszeiten wird er dauerhaft überschritten und die angrenzende ANF feuert mit ihrer maximal möglichen Rate die nur noch durch *Refraktärzeiten* reguliert wird (siehe Abschnitt A.6).
Die zuvor beschriebene Modellierung nach MacQueen (1996) und Shamma u. a. (1986) besitzt diesen Nachteil nicht, da dort der Anteil geöffneter Ionenkanäle unmittelbar von der Auslenkung der Stereozilien abhängt und die Leitfähigkeit unterschiedlichen Perioden folgen kann. Die Membrankapazität C_m aus Gleichung A.10 wird so angepasst, dass das beobachtbare Verhältnis zwischen Gleich- und Wechselstomanteil des IHC-Potenzials eingehalten wird (siehe Sumner u. a., 2002).

A.5 Neurotransmitterausschüttung

Um letztendlich ein Aktionspotenzial (AP) am Hörnerven auszulösen, müssen am präsynaptischen Ende der IHC *Neurotransmitter*moleküle (Glutamat) in den *synaptischen Spalt* freigesetzt werden. Diese befinden sich gruppiert zu sogenannten *Vesikeln* à etwa 2000 Molekülen in als *aktive Zonen* bezeichneten Bereichen an dem den Hörnerven zugewandten Ende der IHCs (siehe z.B. Johnston und Wu, 1995, Kapitel 12). Zur Ausschüttung der Neurotransmittervesikel ist die Bindung von positiv geladenen Calcium-Ionen erforderlich. Die Anzahl emittierter Vesikel ergibt sich also in Abhängigkeit von der Calcium-Konzentration in der Nähe der aktiven Zonen. Diese

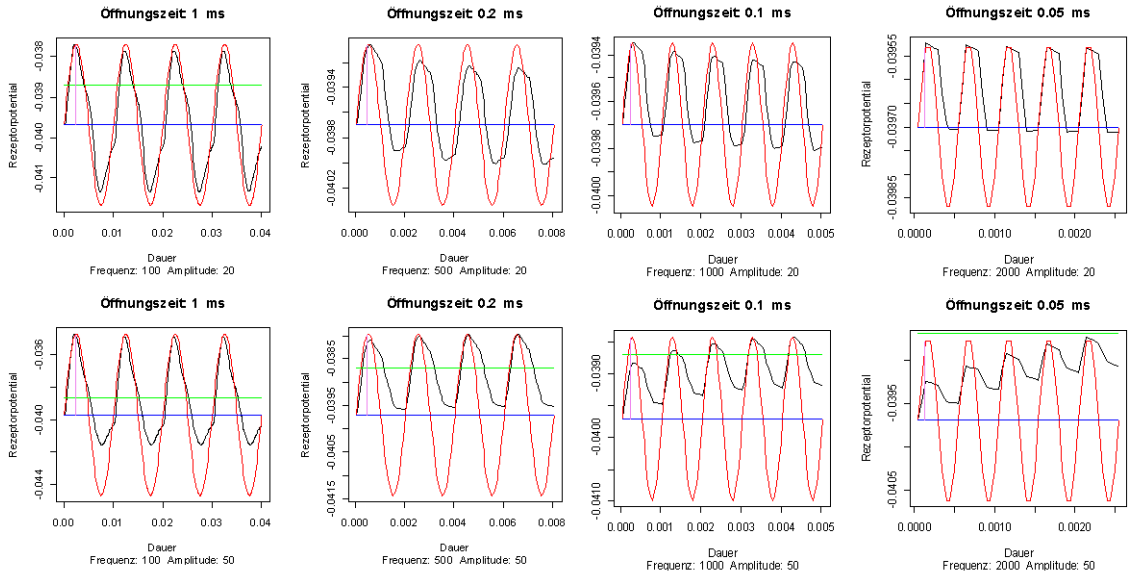


Abbildung A.11: Verlauf des IHC-Potenzials für Sinusschwingungen von 100 (links), 500, 1000 und 2000 Hz (rechts) bei kleiner (oben) und großer (unten) Signalamplitude, bei variierenden Kanalöffnungsdauern von 1/10 der Schwingungsperiode der BM. Die grüne horizontale Linie entspricht dem Schwellenwert für eine Aktionspotenzialemission.

kann nach Kidd und Weiss (1990) modelliert werden in Abhängigkeit von dem in der IHC vorliegenden Potenzial. Der Calciumstrom $I_{Ca}(t)$ wird dabei zunächst beschrieben durch

$$I_{Ca}(t) = G_{Ca}^{max} m_{I_{Ca}}^3(t) \left(V(t) - E_{Ca}\right) \tag{A.11}$$

wobei E_{Ca} das Calcium-Nernstpotenzial darstellt, G_{Ca}^{max} die maximale Durchlässigkeit der IHC-Membran für Calcium-Ionen, und $V(t)$ ist das in Gleichung (A.10) beschriebene Potenzial der inneren Haarzelle. $m_{I_{Ca}}(t)$ beschreibt den Anteil geöffneter Calcium-Ionenkanäle zum Zeitpunkt t. Dieser wirkt sich in dritter Ordnung auf die Leitfähigkeit aus und ergibt sich als tiefpassgefilterte Funktion

$$\tau_{I_{Ca}} \frac{dm_{I_{Ca}}(t)}{dt} + m_{I_{Ca}}(t) = m_{I_{Ca},\infty} \tag{A.12}$$

der spannungsabhängigen Grenzdurchlässigkeit

$$m_{I_{Ca},\infty} = \frac{1}{1 + \beta_{Ca}^{-1} e^{\gamma_{Ca} V(t)}}. \tag{A.13}$$

β_{Ca} und γ_{Ca} sind an neurophysiologische Messdaten angepasste Parameter (Hudspeth und Lewis, 1988) und in Anhang B.4 angegeben; $\tau_{I_{Ca}}$ ist die Zeitkonstante für den Verlauf der Calciumdurchlässigkeit.
Aus dem Calciumstrom (Gleichung A.11) lässt sich nun die an den aktiven Zonen vorliegende Calciumkonzentration bestimmen (Hudspeth und Lewis, 1988):

$$\tau_{[Ca]} \frac{\delta[Ca^{++}](t)}{\delta t} + [Ca^{++}](t) = I_{Ca}(t). \tag{A.14}$$

Die Ausschüttung von Neurotransmittern erfolgt nicht deterministisch sondern stochastisch und lässt sich durch eine Binomial- oder Poissonverteilung beschreiben (siehe z.B. Gabbiani, 2005). Die Wahrscheinlichkeit der Emission von Transmittervesikeln erfolgt in Anhängigkeit von der an der *aktiven Zone* vorliegenden Calciumkonzentration $[Ca^{++}]$. Beutner u. a. (2001) geben an, dass fünf Bindungsschritte von Calciumionen zur Emission eines Neurotransmittervesikels nötig sind. Für eine gegebene Bindungswahrscheinlichkeit eines einzelnen Calcium-Ions ergäbe sich damit ein Zusammenhang fünfter Ordnung. Es sind jedoch manche der Vesikel bereits ein- oder mehrfach gebunden (Hemmert, 2005), so dass sich effektiv ein Zusammenhang der Neurotransmittervesikelausschüttungswahrscheinlichkeit von der Calciumkonzentration mit kleinerer Ordnung ergibt. Augustine und Charlton (1985) geben einen effektiven Zusammenhang dritter Ordnung an, in dem sie gemessene postsynaptische Ströme an präsynaptsiche Ströme in Riesentintenfischen anpassen. Mit Bezug auf diese Arbeit modellieren Sumner u. a. (2002) einen Zusammenhang dritter Ordnung in der folgenden Art:

$$k(t) = \max\{([Ca^{++}]^3(t) - [Ca^{++}]^3_{thres})z, 0\}. \tag{A.15}$$

Dabei stellt $k(t)$ die momentane Emissionswahrscheinlichkeit für Neurotransmittervesikel dar. $[Ca^{++}]^3_{thres}$ und z sind an beobachtete Daten (in Meerschweinchen) angepasste Konstanten, die im Anhang B.4 tabelliert sind. Die explizite Form von Gleichung (A.15) lässt sich allerdings nicht unmittelbar neurophysiologisch motivieren sondern soll phänomenologisch den von Augustine und Charlton (1985) beobachteten Zusammenhang dritter Ordnung nachbilden.
Eine gängige Annahme ist, dass die Neurotransmitter bereits zu emittierbaren Vesikeln „verpackt“ in sogenannten *Readily Releasable Pools (RRPs)* in nächster Nähe zum synaptischen Spalt gruppiert sind und dort auf ihre Ausschüttung „warten“. Dies

ermöglicht eine derart schnelle Antwort der Nerven mit Hilfe von chemischen Prozessen, wie sie zu beobachten ist (für eine Beschreibung des Phänomens siehe z.B. Neher, 2003). Sumner u. a. (2002) schlagen ein drei-Pool-Modell vor (siehe Abbildung A.12), in dem die Vesikel in einem RRP zur Emission in den synaptischen Spalt bereitgestellt liegen. Die Vesikelemission wird binomialverteilt modelliert: $B(q(t), k(t))$ Neurotrans-

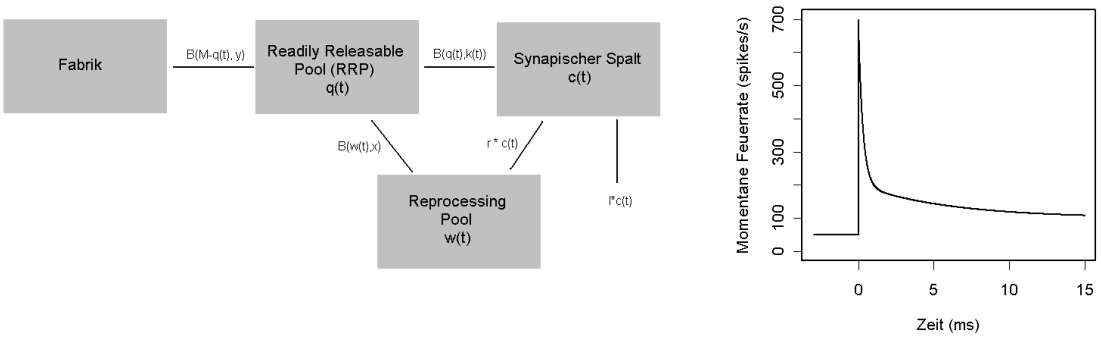


Abbildung A.12: Links: Schematische Darstellung des drei-Pool-Modells von Sumner u. a. (2002) der Neurotransmitterausschüttung. Rechts: Kurve zum Adaptionsverhalten der Releaserate in Anhängigkeit von der Stimulusdauer (ab $t = 0$) nach Westerman und Smith (1984).

mittervesikel werden während eines Zeitintegrationsschritts in den synaptischen Spalt emittiert, wobei q die Anzahl der aktuell im RRP befindlichen Vesikel darstellt. Der RRP wird (relativ langsam) aus einem „Fabrik"-Pool wieder aufgefüllt mit einer Rate y. Ein Anteil der im Spalt befindlichen Vesikel geht mit einer Rate l verloren, die restlichen Neurotransmitter gelangen mit einer Rate r wieder zurück in die IHC und werden dort in einem *Reprocessing Pool* wieder zu Vesikeln gebündelt und binomialverteilt $B(w(t), x)$ dem RRP zur Verfügung gestellt, wobei $w(t)$ die aktuell im Reprocessing Pool befindliche Anzahl an Vesikeln und x eine konstante Rate darstellen. Es ergibt sich ein System von drei Differentialgleichungen:

$$\begin{array}{rcl} \frac{dq(t)}{dt} & = & B([w(t)], x) + B(M - q(t), y) - B(q(t), k(t)) \\ \frac{dc(t)}{dt} & = & B(q(t), k(t)) - lc(t) - rc(t) \\ \frac{dw(t)}{dt} & = & rc(t) - B([w(t)], x) \end{array} \tag{A.16}$$

Hierbei stellen $q(t), c(t)$ und $w(t)$ die aktuell befindlichen Mengen von Neurotransmit-

tern (in Vesikeleinheiten ausgedrückt) in RRP, syaptischem Spalt und Reprocessing Pool dar. x, y, r, l und $k(t)$ geben die Raten für den Pool-Wechsel eines Neurotransmittervesikels an. $k(t)$ ist dabei die einzige Zeitveränderliche die durch die Vorgänge in der IHC beeinflusst wird und wird durch Gleichung (A.15) bestimmt. $[\cdot]$ ist der Gaußklammeroperator und nötig, da die Neurotransmitter nach der Emission in den Spalt nicht mehr zu Vesikeln gebündelt vorliegen. Es werden Anteile von Vesikeln in den Reprocessing Pool aufgenommen und erst dort wieder zu Vesikeleinheiten gebündelt. Auf einen Stimulus erfolgt zunächst eine hohe Vesikelausschüttung (bzw. als weitere Konsequenz auch eine starke Reaktion des Hörnerven). Mit zunehmender Stimulusdauer pendelt sich die Reaktion dann auf einem konstanten Level ein (siehe Abbildung A.12, rechts). Dieses Phänomen ergibt sich, da sich der RRP bei vorliegendem Stimulus entleert, aber nur relativ langsam dazu aus Fabrik und Reprocessing Pool wieder aufgefüllt wird. Nach Stimulusende werden aus dem entleerten RRP mit niedrigem $k(t)$ sogar weniger als die spontane Rate an Transmittern emittiert und der RRP füllt sich langsam wieder auf, bis die spontane Ausschüttungsrate wieder erreicht wird. Die Beschreibung mit einem solchen drei-Pool-Modell erweist sich als sinnvoll um das Phänomen der Adaption realistisch abzubilden (Meddis, 1986). Nach Westerman und Smith (1984) setzt sich Höhe der Feuerraten bei andauerndem Stimulus im Zeitverlauf aus drei unterschiedlichen Komponenten zusammen: der *rapid adaption- Komponente*, der *short-term adaption Komponente* sowie einem konstanten Anteil. Sie lässt sich phänomenologisch beschreiben durch die Gleichung

$$A(t) = A_R e^{\frac{-t}{\tau_R}} + A_{ST} e^{\frac{-t}{\tau_{ST}}} + A_{SS}. \tag{A.17}$$

Dabei steht $A(t)$ für den nicht-periodischen Anteil der Transmitteremission, der im Mittel über viele Stimuluswiederholungen zu beobachten ist und setzt sich aus den rapid-, short-term- und konstanten („steady state“) Anteilen A_R, A_{ST} und A_{SS} zusammen (siehe Abbildung A.12, rechts). Die Größe dieser Anteile im zeitlichen Verlauf werden durch die Zeitkonstanten τ_R (im Bereich weniger Millisekunden) und τ_{ST} (von einigen zig Millisekunden) charakterisiert.
Die Hinzunahme des dritten (Reprocessing) Pools ermöglicht eine schnelle Wiederaufnahme von Vesikeln aus dem synaptischen Spalt und damit eine schnellere Wiederauffüllung des RRPs als es durch die Fabrik allein der Fall wäre. Auf diese Wei-

se können beide Anteile der Neurotransmitteremission realistisch modelliert werden (Sumner u. a., 2003a).

Gemeinsam mit der Bandpassfilterung der Signale entlang der Basilarmembran wird angenommen, dass die Adaption der Feuerrate an den Hörnerven hauptursächlich zur Entstehung der beobachtbaren *Maskierung* beiträgt. Von Maskierung oder *Verdeckung* spricht man, wenn bei zwei Signalen ähnlicher Frequenz nur eines der beiden wahrgenommen und das andere unterdrückt wird (siehe hierzu Baumgarte, 2000, S. 26–29). Ist nun der RRP einer inneren Haarzelle durch ein Signal schon teilentleert wird dadurch die Antwort auf ein zweites Signal ähnlicher Frequenz gehemmt (siehe auch Anhang B.7).

Im vorangegangenen Abschnitt über Stereozilienauslenkung wurde beschrieben, dass nach drei Typen von Hörnervenfasern unterschieden wird (HSR, MSR und LSR), in Abhängigkeit von der Menge ihrer spontanen Vesikelemission im unangeregten Zustand, und dass diese es ermöglichen, einen großen Dynamikbereich eingehender Schallwellen wahrzunehmen. An jede innere Haarzelle münden jeweils etwa acht bis zehn afferente ANFs aller drei Typen, (gleichbedeutend mit etwa $30000 - 35000$ afferenten Hörnerven (vgl. Allen, 1994). Dies bedeutet, dass die beobachtbaren Unterschiede im ANF-Typ ihre Ursache nicht in unterschiedlichen Parametrisierungen der IHCs haben. Sumner u. a. (2003a) erwähnen, dass der Hörnerv-Typ wesentlich beeinflusst wird von der Anzahl an Ionenkanälen in Nähe der Synapse. Unterschiede zwischen HSR, MSR und LSR werden von daher im Wesentlichen durch verschiedene Parameter G_{Ca}^{max} (siehe Gleichung A.11) ausgedrückt. Zur Wahl der Parameter wurde das Saturierungsverhalten der Feuerrate der unterschiedlichen drei ANF-Typen an in Meerschweinchen beobachtete Kurven (vgl. Winter u. a., 1990) bei verschiedenen Intensitätsleveln für hohe Frequenzen von sieben und 18 kHz angepasst, bei denen die Transmitteremission den einzelnen Stimulusperioden nicht mehr folgen kann. Der Parameter M (siehe Gleichung A.16), im Ruhezustand im RRP befindlicher Neurotransmitter, zeigt lediglich Auswirkungen, die sich in einer Skalierung der Emissionsrate ausdrücken.

Abbildung A.13 zeigt exemplarisch die Neurotransmittervesikelausschüttung in den synaptischen Spalt für einen Sinuston von 440 Hz. Im Vergleich zur bandpass gefilterten und nichtlinear verzerrten Originalwelle wie in Abbildung 2.7 findet an dieser Stelle eine Diskretisierung des Signals statt, die vorliegende Daten sind binär: zu jedem

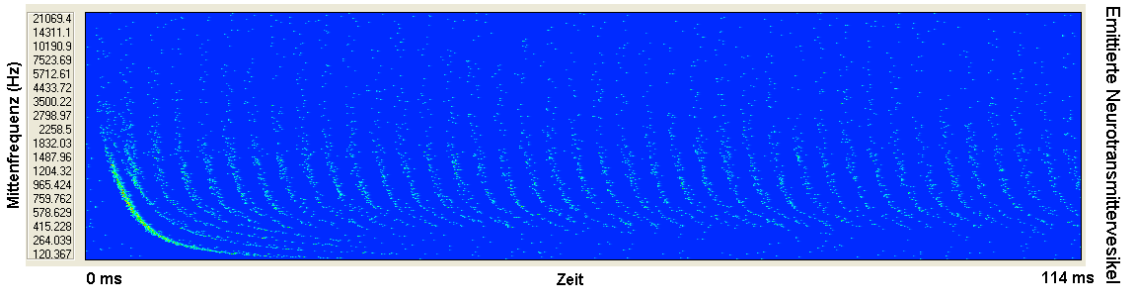


Abbildung A.13: Neurotransmittervesikelemissionen des Simulationsmodells für einen einfach Sinuston von 440 Hz.

Zeitpunkt wird an jeder Position der BM entweder ein Vesikel von Neurotransmittern emittiert (1) oder nicht (0). Die Trajektorien des Wanderwellenverlaufs sind weiterhin erkenntlich und folgen der Stimulusperiode an der entsprechenden Stelle der Basilarmembran.

A.6 Postsynaptische Aktionspotenzialgenerierung

Diffundiert ein Vesikel aus der präsynaptischen IHC in den Spalt, so bindet es an einem Rezeptorprotein der postsynaptischen Membran und setzt dort Ladung frei. Durch die freigesetzten Moleküle eines Vesikels erhöht sich das postsynaptische Potenzial um ein sogenanntes mEPP (miniature end-plate potential) von etwa 0.5–1 mV (vgl. Fatt und Huxley, 1952). Überschreitet die Depolarisation der postsynaptschen Nervenzelle einen bestimmten Schwellenwert ν, kommt es zur Generierung eines Aktionspotenzials: Die Membranspannung depolarisiert zunächst für eine sehr kurze Dauer von weniger als 1 ms extrem stark, danach hyperpolarisiert sie und ist für einen Zeitraum „blockiert“ – die sogenannte *Refraktärzeit* – währenddessen keine weiteren APs auftreten können.
In Abhängigkeit von der Konzentration $c(t)$ von Neurotransmittern im synaptischen Spalt fließt ein Reizstrom

$$I(t) = I_{Vesikel} \cdot c(t) \tag{A.18}$$

in die postsynaptische Zellmembran, wobei $I_{Vesikel}$ den durch ein Vesikel verursachten Stromfluss repräsentiert.
Die ursprüngliche neurophysiologisch präzise Modellierung des Spannungsverlaufs

in Nervenzellen geht zurück auf Hodgkin und Huxley (1952). Dabei sind Haupteinflussfaktoren der Austausch von Kalium- $[K^+]$ und Natrium-Ionen $[Na^+]$ sowie ein spezifisches „Leck“ [L] (im Wesentlichen von Cl^--Ionen) von Bedeutung, die sich durch unterschiedliche maximale Leitfähigkeiten (g_{Na}, g_K, g_L) und zeitliche Verläufe der Membrandurchlässigkeit (in Abhängigkeit vom vorliegenden Potenzial, modelliert durch die von $V_{post}(t)$ abhängigen Variablen m, h, n), auszeichnen. Die Membranspannung $V_{post}(t)$ an den Hörnerven verhält sich gemäß

$$\frac{dV_{post}(t)}{dt}C = -\sum_k I_k(t) + I(t) \tag{A.19}$$

$$\sum_k I_k(t) = (V_{post}(t) - V_{Na})m^3 h g_{Na} + (V_{post}(t) - V_K)n^4 g_K + (V_{post}(t) - V_L)g_L. \tag{A.20}$$

V_{Na}, V_K und V_L sind Nernst-Gleichgewichtspotenziale, die die Richtung des Ionenaustauschs in Abhängigkeit von der Membranspannung $V_{post}(t)$ festlegen. $I_k(t)$ stellt den Ionenaustausch durch die Zellmembran dar und C die Kapazität des Kondensators, den die Zellmembran bildet. Wenn ein externer Strom $I(t)$ – hier z.B. durch die von den IHCs emittierten Neurotransmitter – in die Zelle strömt, lädt sich der Kondensator und es entsteht Leck durch die Membran-Ionenkanäle. Das zeitliche Verhalten der Leitfähigkeiten in Abhängigkeit des Membranpotenzials wird von Hodgkin und Huxley (1952) durch ein System dreier Differentialgleichungen beschrieben:

$$\frac{d\,m(V_{post}(t))}{dt} = \alpha_m(V_{post}(t))(1 - m(V_{post}(t))) - \beta_m(V_{post}(t))m(V_{post}(t)) \tag{A.21}$$

$$\frac{d\,n(V_{post}(t))}{dt} = \alpha_n(V_{post}(t))(1 - n(V_{post}(t))) - \beta_n(V_{post}(t))n(V_{post}(t)) \tag{A.22}$$

$$\frac{d\,h(V_{post}(t))}{dt} = \alpha_h(V_{post}(t))(1 - h(V_{post}(t))) - \beta_h(V_{post}(t))h(V_{post}(t)). \tag{A.23}$$

Die Funktionen $\alpha_i(V_{post}(t))$ und $\beta_i(V_{post}(t))$ in Abhängigkeit von der Spannung $V_{post}(t)$ sind von Hodgkin und Huxley (1952) für $i = \{m, n, h\}$ angepasst worden. Es ergeben sich unterschiedlich schnelle Reaktionen der verschiedenen Ionenflüsse auf eine externe Spannungsänderung, die den Spannungsverlauf in der Membran charakterisieren, siehe Abbildung A.14. Das Verhalten einer auditiven Nervenfaser i lässt sich durch die Zeitpunkte gefeuerter Spikes

$$T_i = \{t_i^{(k)} | 1 \leq k \leq n_i\} = \{t | V_{post,i}(t) = \nu\} \tag{A.24}$$

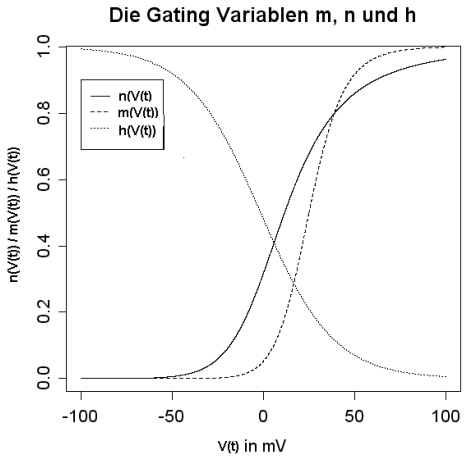


Abbildung A.14: Gating-Variablen h, m und n in Abhängigkeit der Membranspannung $V_{post}(t)$ nach Hodgkin und Huxley (1952).

beschreiben, wobei $V_{post,i}(t)$ die Spannung der postsynaptischen Membran der i-ten Hörnervenfaser und ν der zur AP-Auslösung nötige Schwellenwert sind und n_i die Anzahl von dieser emittierter Spikes bezeichnet. Im Allgemeinen wird $\nu = \nu(t)$ als dynamisch angenommen, da Neuronen unmittelbar nach Emission eines Spikes eine kurze Weile blockiert sind. Diese Spikeunterdrückung wird in der Regel modelliert mit Hilfe einer kurzen absoluten *Refraktärzeit* τ_{abs} von etwa 0.8 ms (vgl. z.B. Sumner u. a., 2002) gefolgt von einem exponentiellen Abklingen mit einer Zeitkonstante τ von etwa 0.25 ms, so dass die Schwellenwertspannung modelliert werden kann durch

$$\nu(t) = \nu_0 + \sum_{t_i^{(k)} \in T_i} \eta_0 e^{\left(\frac{-(t-t_i^{(k)}-\tau_{abs})}{\tau}\right)} \cdot I_{[\tau_{abs},\infty]}(t - t_i^{(k)}) \tag{A.25}$$

wenn das Neuron sich nicht im absolut refraktären Zustand befindet und $\nu(t) = \infty$, falls $t - t_i^{(k)} \in [0, \tau_{abs}]$ für ein $k \in \{1, \ldots, n_i\}$, mit $I(.)$ als Indikatorfunktion. Aus Recheneffizenzgründen wird die Summe meist auf den Einfluss der letzten ein oder zwei Spikes vereinfacht.

Eine einfachere Modellierung des Spikeauslösungsverhaltens kann mit Hilfe von *Integrate-And-Fire (IAF)* Modellen erfolgen (Gerstner, 1998). Die Spannung(sänderung) in der postsynaptischen Membran durch einen externen durch Neuroransmitterausschüttung

verursachten Strom $I(t)$ ergibt sich als

$$\frac{dV_{post}(t)}{dt} = \frac{I(t) - V_{post}(t)/R}{C}. \tag{A.26}$$

Der Leckstrom des IAF-Modells ergibt sich im Gegensatz zum komplizierteren Hodgkin-Huxley Modell durch eine einzelne Komponente

$$I(t) = I_R(t) + I_{cap}(t) \tag{A.27}$$

(siehe Schikowski, 2006), wobei $I_R(t) = V_{post}(t)/R$ den Leckstrom darstellt und $I_{cap}(t)$ den kapazitiven Strom, der die Membran auflädt. Aktionspotenzialgenerierung und Modellierung der Refraktärzeiten erfolgt genau wie im Hodgkin-Huxley Modell, durch einen Schwellenwert $\nu(t)$.
Eine noch einfachere Modellierung der Aktionspotenzialgenerierung im Hörnerven geht auf Beobachtungen von Siegel (1992) zurück, wonach bereits ein einzelnes emittiertes Vesikel in den synaptischen Spalt zur Auslösung eines APs ausreicht und letzteres lediglich verhindert wird durch dessen refraktären Zustand. Dieser kann nun modelliert werden durch eine Spikewahrscheinlichkeit bei Vesikelausschüttung zum Zeitpunkt t, gegeben den letzten Spike Feuerzeitpunkt t_{last}

$$P(Spike|Vesikel, t) = \begin{cases} 0, & t < t_{last} + \tau_{abs} \\ 1 - c_R e^{-\frac{t - t_{last} - \tau_{abs}}{\tau_R}}, & t \geq t_{last} + \tau_{abs} \end{cases} \tag{A.28}$$

dabei stellen τ_{abs} und τ_R Konstanten zur Charakterisierung der Refraktärzeit dar und c_R bestimmt den Einfluss der Refraktärzeit auf die modellierte Spikewahrscheinlichkeit (siehe auch MacQueen, 1996, S. 142). Eine Darstellung des Verlaufs von $P(\text{Spike}|\text{Vesikel}, t)$ ist in Abbildung A.15 (schwarze Linie) zu sehen. Diese Modellierung zeigt jedoch ein Defizit: Werden zwei Vesikel kurz hintereinander emittiert, das zweite noch während der absoluten Refraktärzeit der ANF, so werden vermutlich keine zwei Aktionspotenziale generiert sondern nur eines durch das erste Vesikel und das zweite bleibt völlig ohne Effekt, obwohl gemäß der detaillierteren und neurophysiologisch präziseren Modellierung nach Hodgkin und Huxley (1952) bzw. ebenso gemäß Integrate-And-Fire Modellierung sehr wohl eine Auswirkung auf das postsynaptische Membranpotenzial besteht. Die Wahrscheinlichkeit zur Auslösung eines Spikes durch ein weiteres, im Folgenden ausgeschüttetes Vesikel würde erhöht. Dieses Defizit des

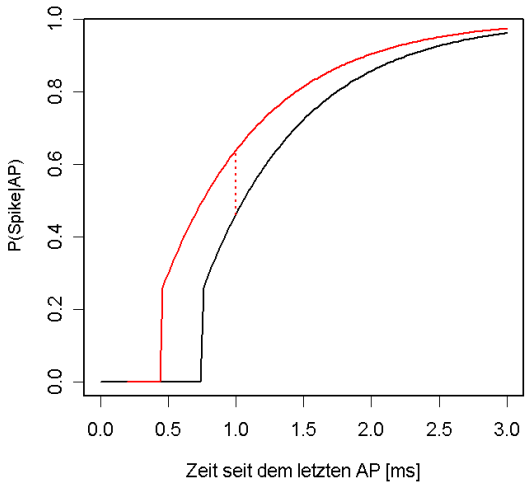


Abbildung A.15: Refraktärzeit: Wahrscheinlichkeit der Auslösung eins APs durch ein ausgeschüttetes Vesikel in Anhängigkeit von der vergangenen Zeit seit dem letzten Spike. Rote Linie: Veränderung von $P(Spike|Vesikel, t)$ durch Erweiterung mit einer Wahl von $\Delta = 0.3$ (vgl. Gleichung A.30) und einem zweiten Vesikel nach $t = 0.2$ ms.

Modells lässt sich durch eine leichte Modifikation beheben:

$$P(Spike|Vesikel, t)^{erweitert} = \begin{cases} 0, & t < t^*_{last} + \tau_{abs} \\ 1 - c_R e^{-\frac{t - t^*_{last} - \tau_{abs}}{\tau_R}}, & t \geq t^*_{last} + \tau_{abs} \end{cases} \tag{A.29}$$

wobei gelten soll:

$$t^*_{last} = \begin{cases} t_{last} & \text{kein weiteres Vesikel im Intervall } [t_{last}, t] \\ t_{last} - n \cdot \Delta & n\text{Vesikel im Intervall } [t_{last}, t] \end{cases} \tag{A.30}$$

Δ erhöht die bereits vergangene Refraktärzeit und damit die Wahrscheinlichkeit einer AP Generierung bei Auftreten eines weiteren Vesikels (vgl. Abbildung A.15, rote Linie). Der Effekt dieser Erweiterung beschränkt sich jedoch auf hochfrequente Stimuli. In einer Arbeit am Fraunhofer Institut für Digitale Medientechnologie (FIDMT, Schikowski, 2006) wurden alle drei Modelle der Aktionspotenzialgenerierung am Hörnerven implementiert. Als Basis zur Parametrisierung des durch Vesikelemission verursachten Stromflusses $I(t)$ diente dazu die Vorgabe, dass die Erhöhung des Potenzials durch

ein Vesikel zum Überschreiten des Schwellenwertpotenzials genügt (Siegel, 1992, s.o.). Untersuchungen der derart simulierten spontanen Raten sowie der Feuerraten bei maximaler Lautstärke zeigen realistische Werte von 47/177 (Hodgkin-Huxley-Modell) bzw. 35/170 (IAF). Die Ausgabe aller drei Modelle für einen Sinuston von 440 Hz ist in Abbildung A.16 dargestellt.

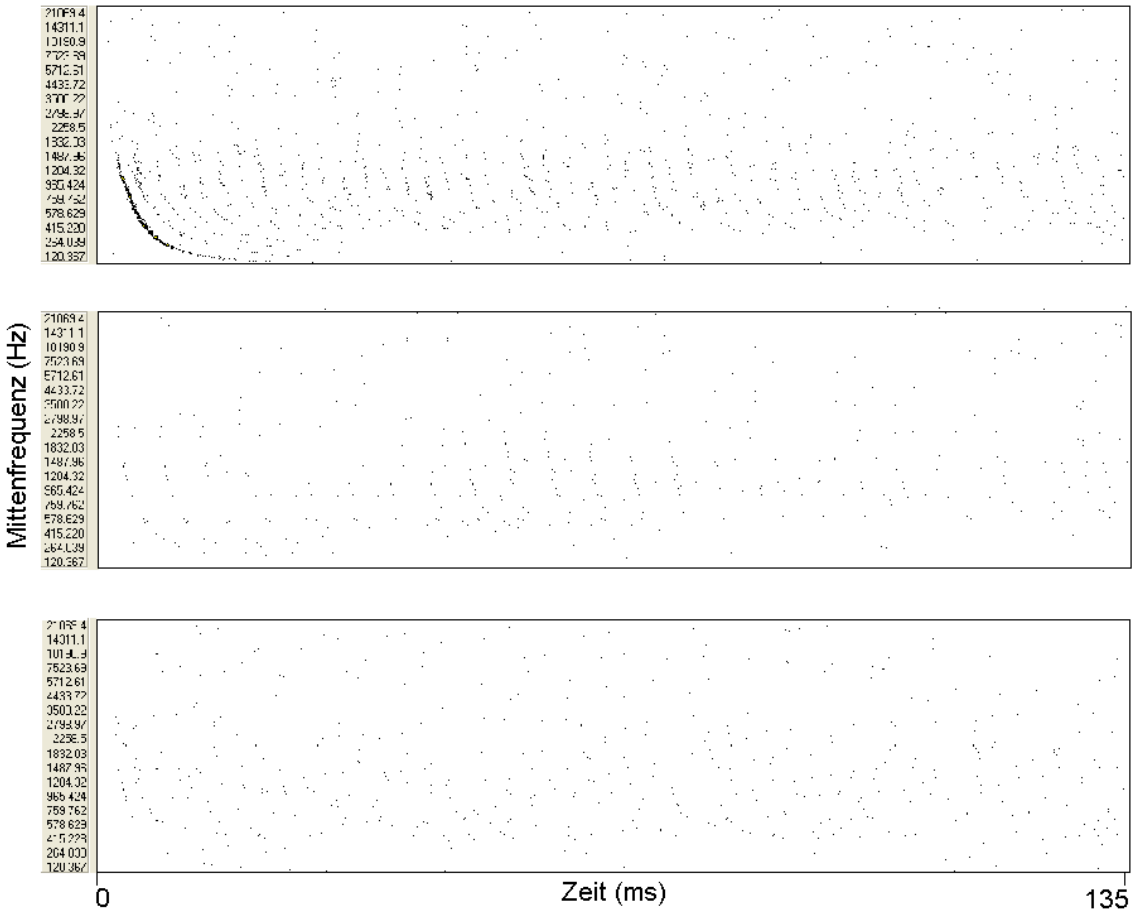


Abbildung A.16: Simulierte Aktionspotenziale an den Hörnerven als Antwort auf einen 440 Hz Sinuston. Oben: Modellierung nach Siegel (1992), Mitte und unten: Modellierung nach Schikowski (2006) für Integrate-And-Fire Neuronen (Mitte) und Hodgkin-Huxley Parametrisierung (unten). Ein Punkt repräsentiert ein Aktionspotenzial zum Zeitpunkt t (Abszisse) an der auf der Ordinate beschriebenen Position der Cochlea.

Die Struktur der Wanderwellen entlang der Basilarmembran, die der Stimulusperiode folgen, ist am besten in dem einfachen, auf Siegel (1992) basierenden Modell wiederzuerkennen. Insbesondere in der auf dem Hodgkin/Huxley-Modell basierenden Parametrisiergn der ANF ist zum Teil nicht einmal mehr die Periode erkenntlich. Für die Simulationen dieser Arbeit wird die einfache Variante der AP Generierung nach Siegel (1992) verwendet, zusammen mit einer Modellierung der Refraktärzeit durch Gleichung A.29. Der Prozessschritt der Aktionspotenzialgenerierung kann damit als probabilistische Tiefpassfilterung der Neurotransmitterausschüttung aufgefasst werden. Zusätzlich besitzt das verwendete AP Generierungs-Modell Vorzüge gegenüber beiden anderen hinsichtlich der Recheneffizienz.

A.7 Kritische Überlegung

Aktionspotenzialemission erfolgt nicht deterministisch. Diese Beobachtung ist zahlreicher in der Literatur beschrieben (siehe z.B. Sumner u. a., 2002; Geisler, 1998; Siegel, 1992). Basierend auf der Arbeit von Sumner u. a. (2002) erfolgt die Modellierung im verwendeten Simulationsmodell durch eine binomialverteilte Ausschüttung von Neurotransmittern in den synaptischen Spalt. Eine solche Modellierung entspricht der Realität in der neuronalen Informationsübertragung (vgl. z.B. Gabbiani, 2005; Johnston und Wu, 1995). In einem echten menschlichen Ohr existieren jedoch erheblich mehr als die modellierten 251 inneren Haarzellen (etwa 3500, jede verbunden mit etwa acht bis zehn afferenten Hörnerven, vgl. Yang u. a., 1992; Allen, 1994).
Es ist davon auszugehen, dass eine derart hohe Anzahl simulierter ANFs eine gute Rekonstruktion momentaner Vesikelrelease-Raten ermöglichen würde, wie sie die durch entleerte Pools $q(t)$ einerseits und die Emissionswahrscheinlichkeit $k(t)$ gemeinsam bestimmt sind (vgl. Abschnitt A.5). Für das Simulationsmodell mit lediglich 251 ANFs ist dies nicht der Fall.
Die einzige bekannte Hypothese über einen positiven Effekt von Rauschen besteht in einer Vergrößerung des Dynamikbereichs (z.B. durch die thermale Fluktuation der Stereozilien, vgl. Abschnitt A.3), die aber für das beschriebene Simulationsmodell nicht festgestellt werden konnte. Durch die binomialverteilte Modellierung der Neurotransmitteremission wird dieser Effekt unter Umständen sogar unterdrückt: In

einem entleerten RRP ist die Ausschüttung eines Neurotransmitters erst möglich, sobald wieder ein Vesikel an Transmittern (über Reprocessing aus dem synaptischen Spalt oder aus der Fabrik) in den Pool gelangt (siehe Abschnitt A.5). Die Auffüllung des Pools sowohl durch Reprocessing als auch aus der Fabrik erfolgen jedoch nur langsam mit konstanter Rate, unabhängig von der Stimulusperiode. Die – durch ein zusätzliches Sterozilienrauschen erzielte – Erhöhung der Neurotransmittervesikelemissionswahrscheinlichkeit $k(t)$, die der Signalperiode folgt, wird somit von der nichtperiodischen Wiederauffüllung des RRPs dominiert bzw. unterdrückt und besitzt keinen positiven Effekt mehr auf den Erhalt der Stimulusperiode in den Aktionspotenzial-Emissionszeitpunkten.
Es ist denkbar, an Stelle der binomialverteilten Neurotransmitterausschüttung zukünftig deterministische Ausschüttungs**raten** zu modellieren und diese anschließend einer Tiefpassfilterung (zur Nachbildung des Effekts neuronaler Refraktärzeiten) zu unterziehen.

B Ergänzungen zur auditorischen Modellierung

B.1 Post Stimulus Time Histogram

Das *Post Stimulus Time Histogram* (PSTH) dient zur Charakterisierung neuronaler Aktivität im zeitlichen Verlauf (im Gegensatz zu Durchschnittsfeuerraten, wie sie in Kapitel 3.4 beschrieben sind). Trotz seiner häufigen Verwendung ist nur selten eine explizite Definition in der Literatur angegeben. Das PSTH beschreibt die durchschnittliche Aktivität eines Neurons über viele Stimuluswiederholungen während kurzer Zeitbins. Diese Aggregierung von einzelnen Beobachtungen (hier den Aktionspotenzialzeitpunkten) entspricht einer Histogrammbildung.

Miller und Sachs (1983) unterteilen zur Berechnung ein (rechteckiges) Zeitfenster von 20 ms in 256 gleichlange Zeitabschnitte (den Bins). Als PSTH-Schätzung fungiert die gemittelte Anzahl an Spikes pro Bin über alle Stimuluspräsentationen (Vokale von 100 ms Dauer). Diese wurden alle 250 ms wiederholt, so lange, bis im ersten Zeitfenster mindestens 600 Spikes gezählt wurden. Die Spikes wurden dabei (an Katzen) mit einer Genauigkeit von $10\,\mu s$ gemessen.

Zur Beurteilung des zeitlichen Verlaufs der neuronalen Aktivität ist eine PSTH Schätzung auf Basis eines einzelnen Stimulus nicht sinnvoll. Im auditorischen Simulationsmodell besteht der Ausweg in einer mehrfachen Anwendung des Simulationsmodells, die jedoch eine erhebliche Erhöhung der Rechenzeit zur Folge hat.

Ein Beispiel eines PSTH für einen 441 Sinuston an der Cochlea-Position mit entsprechender Mittenfrequenz befindet sich in Abbildung B.1 (rechts).

B.2 Parametrisierung des auditorischen Simulationsmodells

Bei der Wahl der dem Simulationsmodell zugrunde liegenden Parameter kann ausgenutzt werden, dass Sumner u. a. (2002, 2003b) und Meddis (2006) bereits mehrere sehr ähnliche Parametersätze für ein Modell der inneren Haarzelle angeben und Baumgarte (2000) eine Parametrisierung für den Prozess bis hin zur schwingenden Basilarmembran durchführt. Die Schwierigkeit liegt in der Verknüpfung der beiden Modelle, was mit Hilfe einer Kopplungskonstante erfolgt (vgl. Anhang A.2). Für deren Wahl liegen am Fraunhofer IDMT Erfahrungswerte vor. Für die vorliegende Arbeit wurde sie dahingehend gewählt, realistische Werte hinsichtlich der maximalen Auslenkung der Basilarmembran sowie der Emissionswahrscheinlichkeit für Neurotransmitter zu generieren (vgl. Anhang B.4). Die vorgenommene Parametrisierung des Simulationsmodells wird im Folgenden kurz dahingehend evaluiert, realistische Ergebnisse des Simulationsmodells zu gewährleisten.

Die simulierten Antworten der Hörnerven sollen sowohl realistische durchschnittliche Feuerraten für hochfrequente Töne besitzen als auch *Phase Locking* bei niederfrequenten Tönen aufweisen. Beide Eigenschaften sind für die in Abschnitt 3.2 beschriebenen Arten der Informationskodierung von Bedeutung.

Als Basis zur Beurteilung der Modellgüte dienen Beobachtungswerte, wie sie in Schoonhoven u. a. (1997), Geisler (1998) und Johnson (1980) zu finden sind. Zur Bewertung des Phase Lockings wird ein Testton von 441 Hz und 0.35 s Dauer verwendet. Es es werden je zehn Wiederholungen für verschiedene Intensitätslevel (wobei die dB-Umrechnung im Simulationsmodell intern in double Genauigkeit erfolgt.) Abzüglich des Onsets in den ersten 40 ms (vgl. Holmes u. a., 2004) wird ein *Post Stimulus Time Histogram* (PSTH, vgl. Abschnitt B.1) für HSR- und LSR-Fasern an der Position mit entsprechender Mittenfrequenz gebildet durch Zählen aller Spikes, die in Bins der Größe 3/44100 s (anlehnend an Johnson, 1980) anfallen. Dies wird auf die Dauer einer Periodenlänge aggregiert, bei gleichzeitiger Optimierung über die Phasenverschiebung. Bezeichne h_m den relativen Anteil an Spikes in Bin m, dann berechnet sich der *Syn-*

chronization Index ((SI), Johnson, 1980, 1974)

$$SI = \sqrt{\Big(\sum_{m=0}^{M-1} h_m sin(2\pi m/M)\Big)^2 + \Big(\sum_{m=0}^{M-1} h_m cos(2\pi m/M)\Big)^2}, \tag{B.1}$$

als Maß für den Grad an Phase Locking des PSTH, mit M als Anzahl an Histogramm Bins während der betrachteten Stimulusperiode. Gefordert sei nun, dass der maximale SI über die verschiedenen Intensitätslevel im Idealfall etwa 0.8 beträgt (vgl. Johnson, 1980).

Zur Beurteilung der beobachteten durchschnittlichen Feuerraten für unterschiedliche Intensitätslevel werden Simulationen eines 6300 Hz-Sinustons für verschiedene Intensitätslevel durchgeführt. Sumner u. a. (2003b) betrachten hierzu die vollständige Dauer; als durchschnittliche Spikerate werden alle auftretenden Spikes für HSR und LSR-Nervenfasern an der ANF mit entsprechender Mittenfrequenz (ANF Nr. 209) summiert – hochgerechnet auf die Dauer einer Sekunde.

Im Sinne eines realistischen Simulationsmodells von besonderer Bedeutung ist zudem der Dynamikbereich, den das Simulationsmodell umspannt, denn die durchschnittlichen Feuerraten von HSR und LSR zeigen bei unterschiedlichen Intensitätsleveln Sättigung. Die Kurve für HSR-Nervenfaser saturiert bereits für geringe Intensitätslevel während diejenige für LSR Nervenfasern erst für etwa 25 dB höhere Dynamiklevel (vgl. z.B. Schoonhoven u. a., 1997) einen deutlichen Anstieg erkennen lässt.

Saturierung der Durchschnittsfeuerraten für hohe Intensitätslevel sollte sich für beide Nervenfasertypen bei etwa 300 Spikes/s einstellen (Schoonhoven u. a., 1997).

Für die spontanen Raten der ANFs gilt nach Geisler (1998), dass diese für HSR ANFs im Mittel etwa 60, jedoch deutlich über 18 sowie weniger als 100 und für LSR 0 Spikes/s betragen. Diese lassen sich leicht durch Mittelwertbildung über alle 251 ANFs der simulierten Antwort auf eine Stille von einer Sekunde Dauer schätzen.

Das *Raten-Intensitäts Diagramm* (Abb. B.1, links) zeigt Sättigung der durchschnittlichen Feuerraten für HSR- bereits bei etwa 20 dB niedrigeren Raten als dies für LSR-Nervenfasern der Fall ist. Dies entspricht in etwa dem weiter oben formulierten Unterschied von 25 dB, wobei das Simulationsmodell einen etwas geringeren Dynamikbereich abdeckt. Auffällig sind jedoch die niedrigeren Sättigungsraten für LSR-Nervenfasern sowie der Rückgang der Durchschnittsfeuerraten der LSR-ANFs für sehr

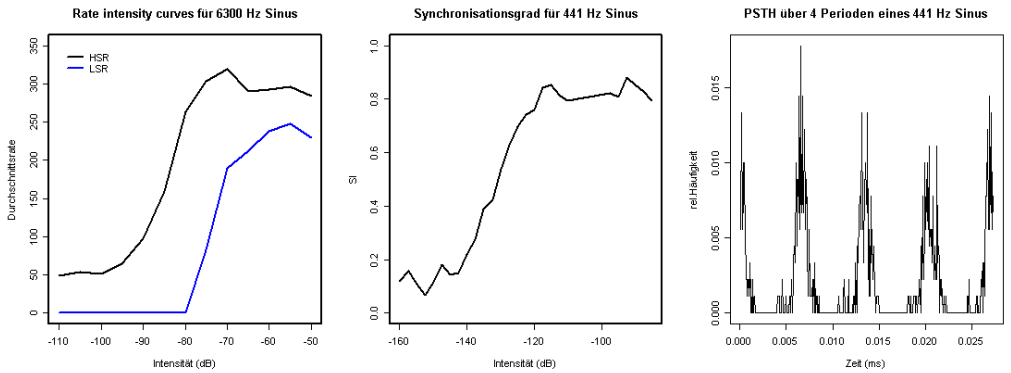


Abbildung B.1: Raten-Intensitäts Diagramme simulierter HSR- und LSR-Nervenfasern für einen 6300 Hz Sinuston (links), Synchronitätsindex einer HSR Nervenfaser für einen 441 Hz Sinuston (Mitte) und PSTH über 4 Stimulusperioden eines 441 Hz Sinustons (rechts).

hohe Intensitätslevel.

Die mittlere Grafik in Abbildung B.1 zeigt den Synchronisationsindex für HSR Nervenfasern bei einem Sinuston von 441 Hz unterschiedlicher Intensität. Dieser saturiert bei einem realistischen Wert von etwa 0.8. Abbildung B.1 (rechts) zeigt die durchschnittliche Aktivität einer HSR mit entprechender Mittenfrequenz über die Dauer von 4 Perioden im PSTH entsprechend der Beschreibung weiter oben. Es ist deutlich zu erkennen, dass das Spikeverhalten der Stimulusperiode folgt.

Die mittlere *spontane Rate* über alle 251 HSR Nervenfasern während eines Stimulus von einer Sekunde Stille liegt bei 48.7 Spikes/s. Für LSR Hörnerven liegt die spontane Aktivität bei 0 Spikes/s. Auch diese beiden Werte sind realistisch.

Mit leichten Abstrichen beim Sättigungsverhalten der LSR Nervenfasern für hohe Intensitätslevel und dem umspannten Dynamikbereich ist damit für die Parametrisierung des auditorischen Simulationsmodells ein realistisches Verhalten im Sinne der oben formulierten Anforderungen zu beobachten. Geringfügige Änderungen der Kopplungskonstante bewirken keine nennenswerte Änderung des beobachtbaren Spikeverhaltens. Die Parameter des Simulationsmodells sind in Abschnitt B.4 angegeben.

Abbildung B.2 zeigt die durchschnittlichen Feuerraten für Töne unterschiedlicher Frequenz (Abszisse, vgl. auch Anhang B.5) und Lautstärke (Ordinate). Es entsprechen

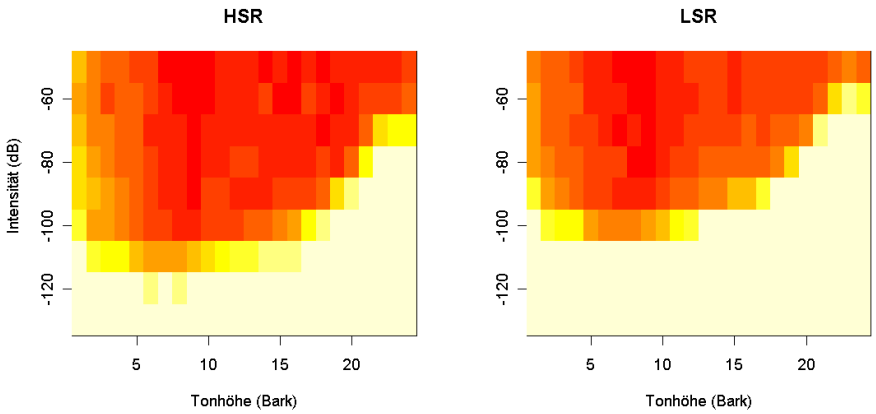


Abbildung B.2: Raten-Intensitäts Diagramme für Sinustöne unterschiedlicher Frequenz und Intensität.

fünf Bark in etwa 520 Hz, zehn Bark in etwa 1270 Hz, 15 Bark etwa 2700 Hz und 20 Bark etwa 6400 Hz. Der charakteristische Verlauf der Hörschwellen in Abhängigkeit von der Frequenz wird realistisch abgebildet (vgl. z.B. Schukat-Talamazzini, 1995, S. 40). Es ist zu erkennen, dass die LSR-Nervenfasern erst durch ca. 20 dB lautere Töne angeregt werden.

Für die Auswertungen dieser Arbeit werden im Falle einfacher simulation (S-ANF) HSR Nervenfasern zur Simulation verwendet. Zur wiederholten Simulation (M-ANF) werden alle drei Nervenfasertypen H/M/LSR im Verhältnis 3:1:1 simuliert, da dieses Verhältnis der Realität entspricht (vgl. Abschnitt A.3). Die Auswertung der Sprache erfolgt mit einer Intensität von -57.5 dB bezogen auf den Vollauschlag des original `*.wav`-Files (Umskalierung innerhalb des Simulationsmodell mit double-Präzision, vgl. hierzu Anhang B.6). Diese Lautstärkenormierung entspricht in etwa dem typischen Level gesprochener Sprache (vgl. Abschnitt B.6).

B.3 Zusammenfassung der Gleichungen des Simulationsmodells

Ausgehend von der Geschwindigkeit der Basilarmembran $v(t)$ erfolgt in diesem Abschnitt eine zusammenfassende Auflistung der verwendeten Modellgleichungen. Für Details sei auf den Abschnitt A verwiesen.

- Stereozilienauslenkung (vgl. Abschnitt A.3):

$$\tau_c \frac{\mathrm{d}u(t)}{\mathrm{d}t} + u(t) = \tau_c C_{Stereozilien} v(t)$$

  (τ_c und $C_{Stereozilien}$ entsprechen dabei `TAU_C` und `C_CILIA` in Abschnitt B.4.)

- Kaliumionenfluss in die IHC (vgl. S. 134):

$$G(u(t)) = \frac{G^{max}_{Stereozilien}}{1 + e^{\frac{u(t)-u_0}{s_0}}(1 + e^{\frac{u(t)-u_1}{s_1}})} + G_a$$

  (u_0, u_1, s_0, s_1, $G^{max}_{Stereozilien}$ und G_a entsprechen `U_0`, `U_1`, `S_0`, `S_1`, `G_CILIA_MAX` und `G0_CILIA` in Abschnitt B.4.)

- Membranpotenzial der IHC (vgl. S. 135):

$$C_m \frac{dV(t)}{\delta t} + G(u)\left((V(t) - E_t) + G_k(V(t) - E^*_k)\right) = 0$$

$$E^*_k = E_k + E_t E_{k-corr}$$

  (C_m, E_k, E_t und E_{k-corr} entsprechen `TOTCAP_CM`, `REV_POT_EK`, `ENDO_POT_ET` und `EK_CORR` in Abschnitt B.4.)

- Presynaptische Calciumkonzentration (vgl. S. 138f):

$$m_{I_{Ca},\infty} = \frac{1}{1 + \beta^{-1}_{Ca} e^{\gamma_{Ca} V(t)}}$$

$$\tau_{I_{Ca}} \frac{dm_{I_{Ca}}(t)}{dt} + m_{I_{Ca}}(t) = m_{I_{Ca},\infty}$$

$$I_{Ca}(t) = G^{max}_{Ca} m^3_{I_{Ca}}(t)\,(V(t) - E_{Ca})$$

$$\tau_{[Ca]} \frac{\delta[Ca^{++}](t)}{\delta t} + [Ca^{++}](t) = I_{Ca}(t)$$

  (G^{max}_{Ca}, E_{Ca}, $\tau_{I_{Ca}}$, β_{Ca}, γ_{Ca} und $\tau_{[Ca]}$ entsprechen `GCA_MAX_XSR`, `REV_POT_ECA`, `TAU_ICA`, `BETA_CA`, `GAMMA_CA` und `TAU_CA` in Abschnitt B.4.)

- Neurotransmitteremissionswahrscheinlichkeit (vgl. S. 139):

$$k(t) = \max\{([Ca^{++}]^3(t) - [Ca^{++}]^3_{thres})z, 0\}$$

  ($[Ca^{++}]^3_{thres}$ und z entsprechen `CA_THRES_XSR` und `SCALAR_Z` in Abschnitt B.4.)

- Neurotransmitterpool-Inhalte (vgl. S. 140):

$$\begin{array}{rcl} \frac{dq(t)}{dt} & = & B([w(t)], x) + B(M - q(t), y) - B(q(t), k(t)) \\ \frac{dc(t)}{dt} & = & B(q(t), k(t)) - lc(t) - rc(t) \\ \frac{dw(t)}{dt} & = & rc(t) - B([w(t)], x) \end{array}$$

  (M, l, r, x und y entsprechen `TPOOL_M_XSR`, `TPOOL_L`, `TPOOL_R`, `TPOOL_X` und `TPOOL_Y` in Abschnitt B.4.)

- Aktionspotenzialemission am Hörnerv (vgl. S. 146):

$$P(Spike|Vesikel, t) = \begin{cases} 0, & t < t_{last} + \tau_{abs} \\ 1 - c_R e^{-\frac{t - t_{last} - \tau_{abs}}{\tau_R}}, & t \geq t_{last} + \tau_{abs} \end{cases}$$

  (τ_{abs}, τ_R und c_R entsprechen dabei `MEDDIS_AN_RA`, `MEDDIS_AN_SR` und `MEDDIS_AN_CR` in Abschnitt B.4.)

B.4 Parameter des Simulationsmodells

Die beschriebene Parametrisierung des auditorischen Simulationsmodells, wie es in dieser Arbeit verwendet worden ist, entspricht in den Bezeichnungen denjenigen, wie sie in der Konfigurationsdatei des in Anhang C.3 beschriebenen, gemeinsam mit dem Fraunhofer Institut für Digitale Medientechnologie (IDMT) entwickelten Programms `EarAnalyzer` verwendet werden.

BM_NO_SECTIONS	251	MEDDIS_AN_RA	0.75e-3
BM_MAX_SECT	252	TPOOL_M_HSR	10.0
BM_MAX_LAT_COUPL	8	TPOOL_M_MSR	10.0
BM_MAX_BARK	25.0	TPOOL_M_LSR	8.0
BM_DELTA_Z	0.1	TPOOL_L	2580.0
BM_FACTOR	0.0000095	TPOOL_R	6580.0
BM_TYPE	BM_EZB	TPOOL_X	66.3
IHC_TYPE	TYPE_HSR	TPOOL_Y	10.0
C_CILIA	16.0	TPOOL_CTHRES	0.0
TAU_C	2.13e-3	KT_FACTOR	1.0
G_CILIA_MAX	8.0e-9	GANGLION_MODEL	MEDDIS_AN
G0_CILIA	1.974e-9	HH_U_SPIKETHRES	0.0
DISP_S0	85.0e-9	HH_DETECT_REFRAC_TIME	0.0015
DISP_S1	5.0e-7	HH_QUANTUM_CURR	-0.000115
DISP_U0	7.0e-9	HH_CAPACITY	0.000001
DISP_U1	7.0e-9	HH_MAXCOND_NA	120.0
ENDO_POT_ET	100.0e-3	HH_MAXCOND_K	36.0
REV_POT_EK	-70.45e-3	HH_MAXCOND_L	0.3
EK_CORR	0.04	COND_GK	1.8e-8
HH_TEMPERATURE	36.3	TOTCAP_CM	6.0e-12
HH_CONC_NA_IN	14.0	STEREOBROWN_RMS_HSR	0.0
HH_CONC_NA_OUT	140.0	REV_POT_ECA	0.066
HH_CONC_K_IN	124.0	TAU_ICA	1.0e-4
HH_CONC_K_OUT	5.0	BETA_CA	400.0
HH_TAU_FACTOR	0.1	GAMMA_CA	130.0
IF_RESTING_POTENTIAL	-0.070	TAU_CA	1.0e-4
IF_THRESHOLD_POTENTIAL	-0.054	SCALAR_Z	20.0e31
IF_SPIKE_POTENTIAL	0.040	Z_POWER	3.0
IF_REFRACTORY_TIME	0.0045	GCA_MAX_HSR	8.0e-9
IF_QUANTUM_CURR	0.12	GCA_MAX_MSR	4.5e-9
IF_CAPACITY	0.000001	GCA_MAX_LSR	2.75e-9
IF_COND_LEAK	0.000008	CA_THRES_HSR	4.48e-11
MEDDIS_AN_CR	0.55e-3	CA_THRES_MSR	3.2e-11
MEDDIS_AN_SR	0.80e-3	CA_THRES_LSR	4.0e-11

B.5 Tonhöhenwahrnehmung

Die menschliche Tonhöhenwahrnehmung erfolgt nicht proportional zur Frequenz-Skala in Hertz. Der Grund hierfür liegt in der Beschaffenheit der Basilarmembran (vgl. Abschnitt A.2).

Bereits vor relativ langer Zeit wurden anhand psychoakustischer Messungen Skalen erstellt, die die Tonhöhenwahrnehmung des Menschen (Tonheit) beschreiben. Ein doppelt so hoch wahrgenommener Ton soll die doppelte Tonheit bedeuten. Zur Beschreibung der Tonhöhenwahrnehmung wurde die *Bark-Skala* eingeführt, die nach Heinrich von Barkhausen benannt ist.

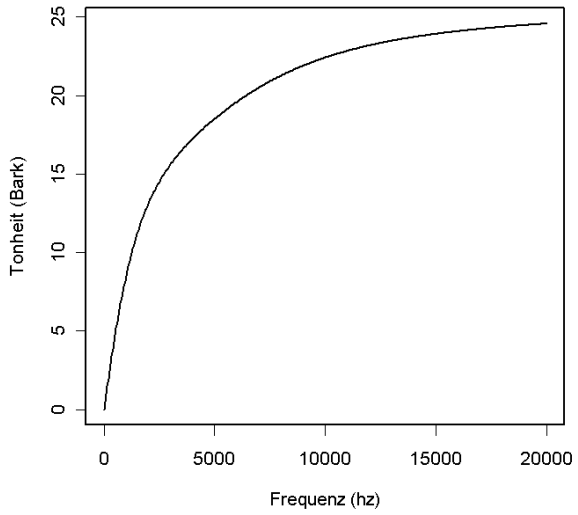


Abbildung B.3: Wahrgenommene Tonhöhe in Abhängigkeit von der zu Grunde liegenden Frequenz.

Eine andere Bezeichung für die Tonheit, die auf Stanley Smith Stevens, John Volkman und Edwin Newmann im Jahr 1937 zurückgeht, ist die Einheit „Mel“ , die sich aus dem Wort „melody“ ableitet. Es gilt 100 Mel = 1 Bark.

Zur Normierung der Barkskala dient der Ton C von 131 Hz. Er entspricht genau 131 Mel = 1.31 Bark.

Für die Umrechnung von Hz in Bark gilt der folgende Zusammenhang:

$$z = 13 \arctan(0.00076 f) + 3.5 \arctan((f/7500)^2), \tag{B.2}$$

wobei f die Frequenz in Hz und z die umgerechnete Frequenz in Bark bezeichnen. Abbildung B.3 veranschaulicht diesen Zusammenhang: tiefere Frequenzen können feiner wahrgenommen werden als hohe Frequenzen.
Die Mittenfrequenzen der Bandpassfilter des auditorischen Simulationsmodells sind äquidistant gewählt entlang der Basilarmembran in Abständen von 0.5 Bark, entsprechend Abständen etwa 0.65 mm (vgl. Abschnitt A.2).

B.6 Schalldruck

Die wahrgenommene Lautstärke eines Schallsignals hängt vom Druck der zugehörigen Schallwelle ab: je stärker der Schalldruck ist, desto stärker wird auch die Basilarmembran in Schwingung versetzt.
In der Regel beschreibt man Lautstärke durch Verhältniswerte. Diese werden in *Dezibel* (dB) angegeben. Es gilt für das Verhältnis zweier Schalldrucklevel p_1 und p_0:

$$1dB = 20 \log_{10} \left(\frac{p_1}{p_0} \right). \tag{B.3}$$

Als feste Bezugsgröße für Schalldruck wurde derjenige Schalldruck der menschlichen Hörschwelle für einen Ton von 1 kHz festgelegt (vgl. Gold und Morgan, 2000, S. 178). Dieser wird mit 0 dB *SPL* bezeichnet, wobei SPL für die englische Bezeichnung *sound pressure level* steht.

Die Wellenamplitude eines `*.wav`-Files ist einheitslos und kann prinzipiell beliebige Schalldrucklevel repräsentieren – von der kaum wahrnehmbaren Laustärke einer auf den Boden fallenden Stecknadel bis hin zur Lautstärke eines Düsenjets. Es ist erforderlich, die Amplitude eines, vom Ohrsimulationsmodell zu verarbeitenden `*.wav`-Files intern (mit double Präzision) auf einen Schalldrucklevel zu normieren. Abbildung B.4 (links) zeigt die durchschnittlichen Aktionspotenzial-Generierungsraten des Simulationsmodells für unterschiedlich normierte 1000 Hz-Sinustöne. Es ist zu erkennen, dass die Hörschwelle bei einer Skalierung von zwischen -130 und -120 dB liegt. Gesprochene Sprache besitzt einen Schalldruck von etwa 70 dB SPL (Gold und Morgan, 2000, S.

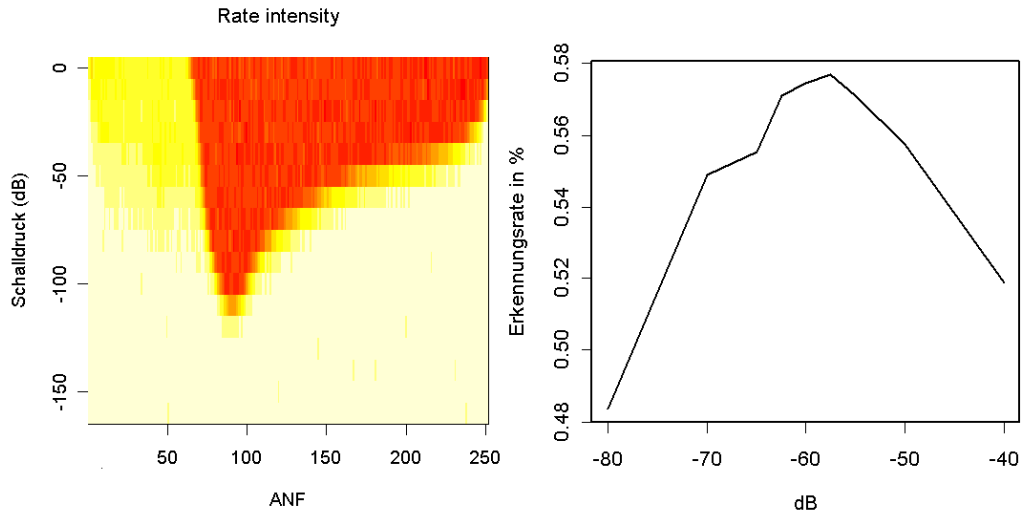


Abbildung B.4: Links: Rate intensity Kurven für unterschiedlich skalierte Sinustöne von 1000 Hz an den verschiedenen Hörnerven. Rechts: Erkennungsraten auf Basis von X^{OD_1}-Merkmalen bei unterschiedlich laut angesteuertem Ohrmodell.

178). Abbildung B.4 (rechts) zeigt Erkennungsraten unter Verwendung der Merkmale X^{OD_1} (vgl. S. 28) bei einem unterschiedlich laut angesteuerten Ohr-Simulationsmodell. Die besten Ergebnisse liegen im zu erwartenden Bereich, für deutlich lautere bzw. leisere Signale ist ein klares Abfallen der Erkennungsergebnisse zu beobachten. Optimale Ergebnisse werden für -57.5 dB erzielt. Aus diesem Grund findet eine solche Ansteuerung auch in dieser Arbeit Verwendung.

B.7 Frequenzmaskierung

In der menschlichen Wahrnehmung von Schall kommt es unter gewissen Umständen zur Unterdrückung von im Signal vorliegenden Frequenzen. Dieses Phänomen wird als *Maskierung* oder *Verdeckung* bezeichnet. Die Ursache für diesen Effekt liegt in der Mechanik des Innenohrs begründet: Die in der Cochlea befindliche Basilarmembran besitzt an ihrem Anfang Resonanzen hoher Frequenzen, während niedrige Frequenzen am apikalen Ende der Cochlea die stärksten Auslenkungen bewirken (vgl. Abschnitt A.2).

Erst nach Erreichen dieser Position wird die von einer Schwingung erzeugte Welle abgedämpft (vgl. Greenberg u. a., 1997). Bis ein Ton niedriger Frequenz die Position maximaler Resonanz am Ende der Cochlea erreicht, verläuft die von diesem erzeugte Welle entlang der Basilarmembran, passiert also auch die Resonanzpositionen für höherfrequente Töne. Ein mitschwingender höherer Ton muss somit so stark sein, dass er die vom tieferen Ton erzeugte Welle „übertönt“, um gehört zu werden.
Zur Bestimmung der *Mithörschwellen* wurde eine Vielzahl von psychoakustischen Experimenten durchgeführt (siehe z.B. Gold und Morgan, 2000, S. 208 ff). Es wird zwischen verschiedenen Maskierungseffekten unterschieden: Man spricht beispielsweise von *Vorwärtsverdeckung* (oder englisch: *forward masking*), wenn der Maskiererton dem maskierten Ton vorausgeht. Starke Effekte treten für einen Zeitabstand von 10-50 ms auf, aber auch für Verzögerungen von 100 ms können noch Maskierungseffekte beobachtet werden (Gold und Morgan, 2000, S. 212). Von *Rückwärtsverdeckung* (englisch: *backward masking*) spricht man, wenn der Maskiererton erst nach dem maskierten Signal einsetzt. Baumgarte (2000) gibt eine Beschreibung verschiedener Maskierungsarten.
Allgemein lässt sich aussagen, dass

- ... ein größerer Maskierungseffekt auftritt, je ähnlicher sich die Frequenzen zweier Signale sind.
- ... der Maskierungseffekt eines Tons größer wird, selbst für weiter entfernte Frequenzen, je lauter er ist.
- ... der Maskierungseffekt stärker ist in Richtung höherer Frequenzen.

Neurophysiologisch ist davon auszugehen, dass Maskierung ihre Hauptursachen im Wanderwellenverlauf entlang der Cochlea besitzt und in der auditorischen Prozesskette insbesondere die Bandpassfilterung des Signals entlang der Basilarmembran, sowie die sich entleerenden Neurotransmitterpools der inneren Haarzellen zu diesem beobachtbaren Phänomen beitragen.
Maskierungseffkete wurden z.B. von Brandenburg und Stoll (1994) bei der Entwicklung des populären MPEG-1 Audio Layer 3 Codecs ausgenutzt, um Musiksignale ohne hörbare Veränderung unter Verwendung von geringstmöglichen Ressourcen speichern zu können.

B.8 Virtuelle Tonhöhenwahrnehmung

Zur Wahrnehmung von Tönen schreibt Jourdain (2001): „Im Gegensatz zu bloßen Schallschwingungen entsteht beim Hören ein Geräusch im Kopf.“ Ein bekanntes Beispiel für den beschriebenen Unterschied stellt das Phänomen des *virtuellen Grundtons* (engl. *virtual pitch* oder *missing fundamental*) dar: In manchen Situationen wird ein Ton vom Hörer wahrgenommen, obwohl die Grundfrequenz dieses Tons im Spektrum der Schallwelle nicht auszumachen ist, sonderen nur deren Harmonische.
Ein solches Beispiel gibt Ligges (2006) auf S. 44 an: Die Grundfrequenz eines Tons **C** einer Sopranistin ($\sim$ 523 Hz) ist im zugehörigen Spektrum fast nicht auszumachen, sondern lediglich diejenige Frequenz des ersten Obertons eine Oktave darüber, etwas über 1000 Hz. Dieses Phänomen ist in Abbildung B.5 (links) dargestellt (erstellt mit dem `R` Paket `tuneR`, vgl. Ligges, 2006, Kapitel 7). Die Reaktion der Hörnerven im au-

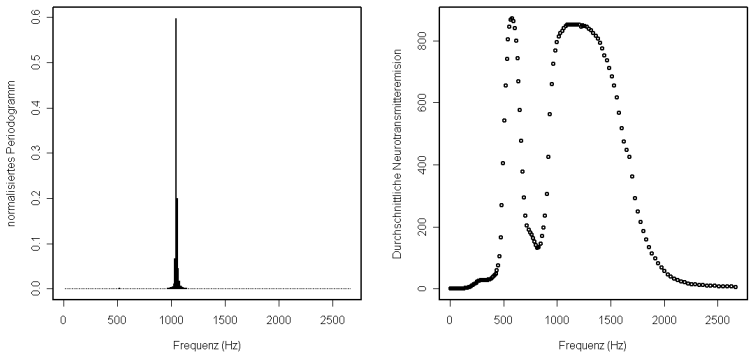


Abbildung B.5: Links: Spektrum des Tons C (Grundfrequenz $\sim$ 523 Hz, zehn ms Sample. Rechts: Reaktion der Hörnerven des audit. Modells: durchschnittliche Neurotransmitteremissionswahrscheinlichkeit an verschiedenen Positionen des Innenohrs (nach Mittenfrequenz abgetragen).

ditorischen Simulationsmodell (mit -50 dB angesteuert, Abbildung B.5, rechts) lässt den Grundton bei etwa 500 Hz klar erkennen. (Aktivität von Hörnerven mit Mittenfrequenzen im Bereich um 1000 Hz ist für einen breiten Bereich von Hörnerven zu beobachten: Dies ist vermutlich durch die Breite der entsprechenden Bandpassfilter bedingt und gestaltet eine exakte Frequenzschätzung für den ersten Oberton allein

auf Basis von Orts-Durchschnittsfeuerraten schwierig. Abbildung B.6 zeigt den Verlauf der Neurotransmitteremissionswahrscheinlichkeit an unterschiedlichen Positionen der Basilarmembran: In beiden Fällen ist ein periodischer Verlauf auszumachen. Die mittlere Periodenlänge in der linken Grafik (an ANF 55) entspricht einer Frequenz von 518 Hz, diejenige in der rechten Grafik (ANF 100) entspricht 1035 Hz.)

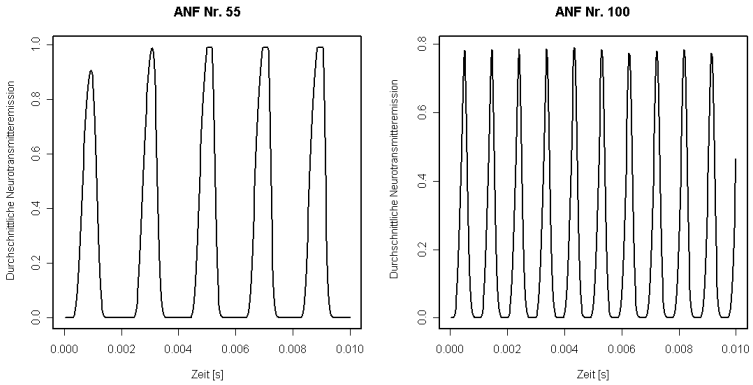


Abbildung B.6: Zeitlicher Verlauf der Neurotransmitteremissionswahrscheinlichkeit für ANF 55 mit einer Mittenfrequenz von etwa 500 Hz (links) sowie ANF 100 mit einer Mittenfrequenz von etwa 1000 Hz (rechts).

Nach Howard und Angus (2006), S. 121 f, lässt sich Tonhöhenwahrnehmung am besten durch das Prinzip einer Suche nach dem größten gemeinsamen Teiler der Peaks im Spektrum erklären. Im Fall des betrachteten Tons von Ligges (2006) handelt es sich gar nicht um einen fehlenden Grundton im eigentlichen Sinne – die Frequenz ist durchaus im Signal vorhanden jedoch nur mit geringer Amplitude. Dies wird ersichtlich, wenn man anstelle des Spektrums das logarithmierte Spektrum betrachtet (vgl. Abb. B.7). Es ist eine deutliche lokale Spitze im Periodogramm bei etwa 500 Hz auszumachen. Das logarithmierte Spektrum trägt dabei der menschlichen Lautheitswahrnehmung Rechnung, da Lautheitsunterschiede auch auf einer logarithmischen Skala beziffert werden (vgl. Abschnitt B.6). Im auditorischen Modell erfolgt diese Transformation implizit (vergleiche auch Perdigao und Sa, 1998): Bereits für sehr geringe Schalldrucklevel werden die Stereozilien ausgelenkt, für sehr hohe Intensitäten saturiert deren Auslenkung jedoch (vgl. Abschnitt A.3).

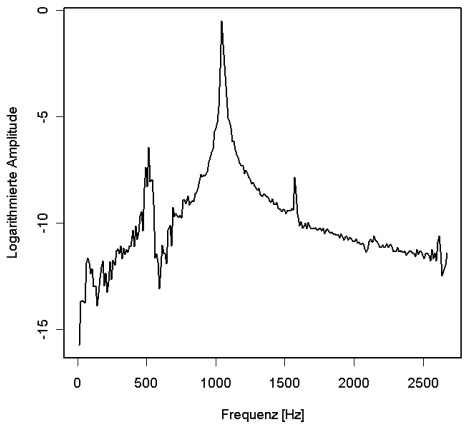


Abbildung B.7: Logarithmiertes Spektrum des betrachteten Tons.

Ausführlichere Untersuchungen zum Phänomen des virtuellen Grundtons sollen nicht den Fokus dieser Arbeit darstellen; das Beispiel legt jedoch nahe, dass Abweichungen der Schallwahrnehmung von der Präsenz von Frequenzen im Spektrum der Schallwelle – zumindest zu Teilen – auf die auditorische Signalverarbeitung zurückführbar sein können.

B.9 Blockweise Berechnung linearer Diskriminanzanalyse

Zur Durchführung einer linearen Diskriminanzanalyse ist die Bestimmung der Streuungsmatrizen innerhalb der Klassen $\mathbf{W}$ sowie zwischen den Klassenmittelwerten $\mathbf{B}$ nötig (vgl. S. 83). Stelle $\mathbf{X}$ die Datenmatrix dar (Zeilen sind Beobachtungen und Spalten sind Merkmale), so ergibt sich $\mathbf{W}$ durch

$$\mathbf{W} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{N_k} \mathbf{W_k} \tag{B.4}$$

mit N der Gesamtzahl an Beobachtungen, k dem Klassenindex, N_k den Klassenhäufigkeiten

$$\mathbf{W_k} = \frac{\mathbf{1}}{\mathbf{N_k}} (\mathbf{X_{|k}})'(\mathbf{X_{|k}}) - \bar{\mathbf{x}}_\mathbf{k} \bar{\mathbf{x}}'_\mathbf{k} \tag{B.5}$$

der empirischen Kovarianzmatrix von Klasse k, wobei hier $(\mathbf{X_{|k}})$ diejenige Teilmatrix von $\mathbf{X}$ darstellt, deren Zeilen Klasse k zugeordnet werden. $\bar{\mathbf{x}}_\mathbf{k}$ ist der Klassenmittelwerts(zeilen)vektor von Klasse k.
Analog gilt für $\mathbf{B}$:

$$\mathbf{B} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{N_k} \left(\bar{\mathbf{x}}_\mathbf{k} \bar{\mathbf{x}}_\mathbf{k}' - \bar{\mathbf{x}} \bar{\mathbf{x}}' \right) \tag{B.6}$$

mit $\bar{\mathbf{x}}$ entsprechend dem Gesamtmittelwertsvektor der Daten.
Wenn die Zahl der Beobachtungen sehr groß ist, kann es bei den Matrixmultiplikationen zur Berechnung von $\mathbf{B}$ und $\mathbf{W}$ zu Speicherkapazitätsproblemen kommen. Im Anwendungsfall dieser Arbeit liegen ~ 172000 Beobachtungsvektoren mit bis zu 251 Komponenten (im Falle der Verwendung der Reaktion der einzelnen Hörnerven als Basis-Merkmalsrepräsentation) vor.
Da die Beobachtungen auf verschiedene Files aufgeteilt sind (eines pro Sprachäußerung) bietet sich ein sukzessives Einlesen der Daten an. Dabei können die Matrizen $\mathbf{B}$ und $\mathbf{W}$ durch Summation einzelner, auf Basis von Teilmengen der Daten gebildeter Kovarianzmatrizen bestimmt werden, da für partitionierte Matrizen der Gestalt

$$\mathbf{A} = \begin{pmatrix} \mathbf{A_1} \\ \mathbf{A_2} \\ \cdots \\ \mathbf{A_p} \end{pmatrix}$$

gilt:

$$\mathbf{A'A} = \sum_{\mathbf{i=1}}^{\mathbf{p}} (\mathbf{A_i})'(\mathbf{A_i}). \tag{B.7}$$

Auf diese Weise können blockweise die Matrizen $(\mathbf{X_{|k}})'(\mathbf{X_{|k}})$ berechnet werden. Auch für die Vektoren $\bar{\mathbf{x}}_\mathbf{k}$ und $\bar{\mathbf{x}}$ sowie die Häufigkeiten N_k und N ist eine schrittweise Berechnung möglich.

B.10 Zusammenhang zwischen HDA und LDA

Die *heteroskedastische Diskriminanzanalyse* stellt eine Erweiterung der linearen Diskriminanzanalyse dar. Es lässt sich zeigen (Kumar, 1997, Anhang C), dass im Fall identisch angenommener Kovarianzmatrizen $\mathbf{W_k} = \mathbf{W}$, $\forall k$ (d.h. ebenso $\mathbf{\Sigma_k} = \mathbf{\Sigma}$, $\forall k$),

die LDA-Transformation die Gleichung 5.24 maximiert, wenn q gleich der Anzahl an LDA-Diskriminanzkomponenten gewählt wird:
Es sei nochmals die logarithmierte Likelihood der beobachteten Daten unter HDA Transformation $\mathbf{A}$ bei Annahme von Diagonalität der $\mathbf{\Sigma_k}$ aus Gleichung 5.24 rekapituliert:

$$\begin{aligned} l(\mathbf{A},\mathbf{x}) \quad = - \quad & \frac{N}{2}\log|\mathrm{diag}((\mathbf{A^{p-q}})'(\mathbf{T})(\mathbf{A^{p-q}}))| - \sum_{k=1}^{K}\frac{N_k}{2}\log|\mathrm{diag}((\mathbf{A^q})'\mathbf{W_k}(\mathbf{A^q}))| \\ & -\frac{Np}{2}(1+\log(2\pi)) + N\log|\mathbf{A}| \end{aligned} \tag{B.8}$$

Durch $\sum_{k=1}^{K}\frac{N_k}{2}\log|\mathrm{diag}((\mathbf{A^q})'\mathbf{W_k}(\mathbf{A^q}))| = \frac{N}{2}\log|\mathrm{diag}((\mathbf{A^q})'\mathbf{W}(\mathbf{A^q}))|$ vereinfacht sich diese zu

$$\begin{aligned} l(\mathbf{A},\mathbf{x}) \quad = - \quad & \frac{N}{2}\log|\mathrm{diag}((\mathbf{A^{p-q}})'\mathbf{T}(\mathbf{A^{p-q}}))| - \frac{N}{2}\log|\mathrm{diag}((\mathbf{A^q})'\mathbf{W}(\mathbf{A^q}))| \\ & -\frac{Np}{2}(1+\log(2\pi)) + N\log|\mathbf{A}|. \end{aligned} \tag{B.9}$$

Ableiten und Nullsetzen von (B.9) ergibt, dass für $\mathbf{A_{ML}}$ mit maximaler Likelihood gelten muss (Kumar, 1997, S. 93):

$$(\mathbf{A^q_{ML}})'\mathbf{W}(\mathbf{A^q_{ML}})\left(\mathrm{diag}((\mathbf{A^q_{ML}})'\mathbf{W}(\mathbf{A^q_{ML}}))\right)^{-1} = \mathbf{I_q}, \tag{B.10}$$

$$(\mathbf{A^{p-q}_{ML}})'\mathbf{T}(\mathbf{A^{p-q}_{ML}})\left(\mathrm{diag}((\mathbf{A^{p-q}_{ML}})'\mathbf{T}(\mathbf{A^{p-q}_{ML}}))\right)^{-1} = \mathbf{I_{p-q}}, \tag{B.11}$$

$$(\mathbf{A^{p-q}_{ML}})'\mathbf{W}(\mathbf{A^q_{ML}})\left(\mathrm{diag}((\mathbf{A^q_{ML}})'\mathbf{W}(\mathbf{A^q_{ML}}))\right)^{-1} = \mathbf{0} \ \textit{und} \tag{B.12}$$

$$(\mathbf{A^q_{ML}})'\mathbf{T}(\mathbf{A^{p-q}_{ML}})\left(\mathrm{diag}((\mathbf{A^{p-q}_{ML}})'\mathbf{T}(\mathbf{A^{p-q}_{ML}}))\right)^{-1} = \mathbf{0}. \tag{B.13}$$

Damit muss die Lösung $\mathbf{A_{ML}}$ der HDA

$$(\mathbf{A_{ML}})'\mathbf{W}(\mathbf{A_{ML}}) = \begin{bmatrix} \begin{bmatrix} \lambda_1^W & & 0 \\ & \cdots & \\ 0 & & \lambda_q^W \end{bmatrix} & 0 \\ 0 & \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda^W_q} & 0 \\ 0 & \mathbf{U} \end{bmatrix} \tag{B.14}$$

und

$$(\mathbf{A_{ML}})'\mathbf{T}(\mathbf{A_{ML}}) = \begin{bmatrix} \mathbf{V} & 0 \\ 0 & \begin{bmatrix} \lambda_{q+1}^T & & 0 \\ & \cdots & \\ 0 & & \lambda_p^T \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{V} & 0 \\ 0 & \mathbf{\Lambda^T_{p-q}} \end{bmatrix} \tag{B.15}$$

erfüllen, wobei $\mathbf{U}$, $\mathbf{V}$ positiv definite Matrizen bezeichnen.
Da $\mathbf{W}$ und $\mathbf{T}$ symmetrisch sind, gilt dies auch für $\mathbf{U}$ und $\mathbf{V}$. Es lassen sich somit Transformationen $\mathbf{D_1}$ und $\mathbf{D_2}$ finden, die $\mathbf{U}$ und $\mathbf{V}$ diagonalisieren (vgl. z.B. Fahrmeir u. a., 1996, Anhang A.11). Seien nun $\mathbf{D_1}$ und $\mathbf{D_2}$ der Art, dass diese sowohl $\mathbf{U}$ als auch $\mathbf{\Lambda^T_{p-q}}$ (bzw. $\mathbf{V}$ und $\mathbf{\Lambda^W_q}$) diagonalisieren und sei $\mathbf{A^*_{ML}} = \mathbf{A_{ML}D}$ eine Transformation, die (B.14) und (B.15) erfüllt und zusätzlich diagonale Matrizen $\mathbf{U}^*$ und $\mathbf{V}^*$ generiert, mit

$$\mathbf{D} = \begin{bmatrix} \mathbf{D_2} & 0 \\ 0 & \mathbf{D_1} \end{bmatrix}. \tag{B.16}$$

Kumar (1997) zeigt Skaleninvarianz von $\mathbf{A_{ML}}$, so dass o.B.d.A $|\mathbf{D_1}| = |\mathbf{D_2}| = \mathbf{1}$. Es ergibt sich für die Determinanten

$$|\mathbf{D_2'\Lambda^W_q D_2}| = |\mathbf{D_2'}||\mathbf{\Lambda^W_q}||\mathbf{D_2}| = |\mathbf{\Lambda^W_q}| \;\; und \tag{B.17}$$

$$|\mathbf{D_1'\Lambda^T_{p-q} D_1}| = |\mathbf{D_1'}||\mathbf{\Lambda^T_{p-q}}||\mathbf{D_1}| = |\mathbf{\Lambda^T_{p-q}}|, \tag{B.18}$$

und die sich ergebenden Matrizen auf der jeweils linken Gleichungsseite sind weiterhin Diagonalmatrizen. Der Wert der (Log-)Likelihood bleibt somit gleich.
Für den Kandidaten $\mathbf{A^*_{ML}}$ mit $(\mathbf{A^*_{ML}})'\mathbf{W}(\mathbf{A^*_{ML}})$ und $(\mathbf{A^*_{ML}})'\mathbf{T}(\mathbf{A^*_{ML}})$ diagonal, gilt dann

$$l(\mathbf{A^*_{ML}}, \mathbf{x}) = l(\mathbf{A_{ML}}, \mathbf{x}). \tag{B.19}$$

und $\mathbf{A^*_{ML}}$ diagonalisiert sowohl $\mathbf{W}$ als auch $\mathbf{T}$. Im Folgenden sei gezeigt, dass dies auch genau für die Lösung einer LDA der Fall ist. Es ergibt sich aus $\mathbf{W^{-1}W} = \mathbf{I}$ und $(\mathbf{A^*_{ML}})'\mathbf{W}\mathbf{A^*_{ML}} = \mathbf{\Lambda_W}$ durch sukzessive Postmultiplikationen der Gleichungen sowie unter Ausnutzung der Orthogonalität von $\mathbf{A^*_{ML}}$ (vgl. z.B. Schmidt und Trenkler, 1998, S. 46 und 78):

$$\begin{array}{lll} \mathbf{I} & = & \mathbf{W^{-1}A^*_{ML}\Lambda_W(A^*_{ML})'} \\ \Leftrightarrow \; ((\mathbf{A^*_{ML}})')^{-1} & = & \mathbf{W^{-1}A^*_{ML}\Lambda_W} \\ \Leftrightarrow \; ((\mathbf{A^*_{ML}})')^{-1}(\mathbf{\Lambda_W})^{-1} & = & \mathbf{W^{-1}A^*_{ML}} \\ \Leftrightarrow \; ((\mathbf{A^*_{ML}})')^{-1}(\mathbf{\Lambda_W})^{-1}(\mathbf{A^*_{ML}})^{-1} & = & \mathbf{W^{-1}}. \end{array}$$

Es sind (wegen $(\mathbf{A^*_{ML}})' = (\mathbf{A^*_{ML}})^{-1}$) also die Spalten von $\mathbf{A^*_{ML}}$ auch Eigenvektoren von $\mathbf{W^{-1}}$, und für die Diagonalmatrix der zugehörigen Eigenwerte gilt: $\mathbf{\Lambda_{(W^{-1})}} = (\mathbf{\Lambda_W})^{-1}$. Weiterhin gilt für die Eigenwerte von $\mathbf{W^{-1}T}$, da auch die Umkehrungen ($\mathbf{W} = (\mathbf{A^*_{ML}})'\mathbf{\Lambda_W}(\mathbf{A^*_{ML}})$) gelten (Fahrmeir u. a., 1996, Anhang 11) und erneut unter Aus-

nutzung der Orthogonalität von $\mathbf{A}^*_{\mathbf{ML}}$:

$$\mathbf{W^{-1}T} = \mathbf{A}'_{\mathbf{ML}}(\mathbf{\Lambda_W})^{-\mathbf{1}}\mathbf{A_{ML}}(\mathbf{A_{ML}})'\mathbf{\Lambda_T}\mathbf{A_{ML}} = (\mathbf{A_{ML}})'(\mathbf{\Lambda_W})^{-\mathbf{1}}\mathbf{\Lambda_T}\mathbf{A_{ML}} \tag{B.20}$$

und damit

$$\mathbf{\Lambda}_{(\mathbf{W^{-1}T})} = (\mathbf{\Lambda_W})^{-\mathbf{1}}\mathbf{\Lambda_T}. \tag{B.21}$$

Die Eigenwerte ergeben sich also aus Quotienten der Eigenwerte λ_i^T und λ_i^W von $\mathbf{T}$ und $\mathbf{W}$ und insbesondere beinhaltet $\mathbf{A}^*_{\mathbf{ML}}$ genau die Eigenvektoren von $\mathbf{W^{-1}T}$, was der Dimensionsreduktion einer linearen Diskriminanzanalyse entspricht.
Für die Log-Likelihood ergibt sich aus Gleichung B.9:

$$l(\mathbf{A}, \mathbf{x}) = -\frac{N}{2}\sum_{d=1}^{q}\log(\lambda_d^W) - \frac{N}{2}\sum_{d=q+1}^{p}\log(\lambda_d^T) - \frac{Np}{2}(1+\log(2\pi)). \tag{B.22}$$

Die Lösung $\mathbf{A_{LDA}}$ einer linearen Diskriminanzanalyse hat als Spaltenvektoren die Eigenvektoren von $\mathbf{W^{-1}T}$, geordnet nach Größe der zugehörigen Eigenwerte als Diagonalelemente von $(\mathbf{\Lambda_W})^{-\mathbf{1}}\mathbf{\Lambda_T}$ (vgl. Gleichung B.21). Ein Vertauschen zweier Spalten von $\mathbf{A_{LDA}}$, $\mathbf{a_i}$ mit $i \leq q$ und $\mathbf{a_j}$ mit $j \geq q+1$, ändert genau zwei Summanden der Likelihood: $-\log\lambda_j^T - \log\lambda_i^W$ zu $-\log\lambda_i^T - \log\lambda_j^W$. Aus der Anordnung gemäß der Größe der Eigenwerte $\frac{\lambda_i^T}{\lambda_i^W} \geq \frac{\lambda_j^T}{\lambda_j^W}$ ergibt sich $\lambda_i^T \geq \frac{\lambda_j^T\lambda_i^W}{\lambda_j^W}$. Hieraus folgt die Abschätzung

$$\begin{aligned} -\log\lambda_i^T - \log\lambda_j^W &\leq -\log\left(\tfrac{\lambda_j^T\lambda_i^W}{\lambda_j^W}\right) - \log\lambda_j^W \\ &= -\left(\log\lambda_j^T + \log\lambda_i^W - \log\lambda_j^W\right) - \log\lambda_j^W \\ &= -\log\lambda_j^T - \log\lambda_i^W + \log\lambda_j^W - \log\lambda_j^W \\ &= -\log\lambda_j^T - \log\lambda_i^W \end{aligned} \tag{B.23}$$

der beiden Summanden. Durch Vertauschen zweier Spalten von $\mathbf{A_{LDA}}$ kann die Likelihood somit nicht vergrößert werden, d.h. die Transformationsmatrix einer linearen Diskriminanzanalyse maximiert gleichzeitig die Likelihood der Beobachtungen, gegeben Normalverteilungen der Klassen mit identischen, diagonalen Kovarianzmatrizen im transformierten Zielraum. Die LDA Transformation ist jedoch nicht eindeutig in ihrer Eigenschaft, die Likelihood $l(\mathbf{A}, \mathbf{X})$ zu maximieren. Man stelle sich beispielsweise Vertauschungen von Zeilen $\mathbf{a_i}$ und $\mathbf{a_j}$ aus $\mathbf{A_{LDA}}$ mit $i, j \leq q$ vor.

B.11 Optimierung der Likelihood in der heteroskedastischen Diskriminanzanalyse

Ein effektiver Algorithmus zur Bestimmung der HDA-Transformationsmatrix, die die Likelihood der Daten optimiert, ist in Gales (1998) angegeben.
Gegeben seien die Inputparameter

- `A`: Startwert der Transformationsmatrix (die LDA-Transformation),
- `q`: die Dimension des Zielraums,
- `d`: die Dimension des Originalmerkmalsraums,
- `W.cl`: Array der klassenspezifischen Kovarianzschätzungen,
- `W`: gepoolte, gemeinsame Kovarianzmatrix für alle Klassen,
- `K`: die Anzahl an Klassen,
- `gamma`: Vektor mit Klassenbesetzungszahlen,
- `N`$= \sum_{k=1}^{K}$ `gamma(k)`: die Gesamtzahl an Beobachtungsvektoren,
- `iters`: Anzahl Optimierungsiterationen und
- `initers`: Anzahl Optimierungsiterationen in der inneren Schleife (10).

Der Algorithmus lautet wie folgt (entommen aus Burget, 2004, S. 155, die Notation entspricht MATLAB Syntax):

```
for iter = 1:iters,
   Q = tau * log(det(A')^2) - tau * d * (log(2 * pi) + 1);

   for i = 1:d,

      if i <= q,
         G = zeros(d,d);

         for k = 1:K,
            sigma.i = A(:,i)' * W.cl{k} * A(:,i);
            G = G + gamma(k) / sigma.i * W.cl{k};
            Q = Q - gamma(k) * log(sigma.i);
         end

       else
          sigma.i = A(:,i)' * W * A(:,i);
          G = tau / sigma.i * W;
          Q = Q - tau * log(sigma.i);
       end

       invG{i} = inv(G);
   end


   Q = Q/2;


   for initer = 1:initers,

      for i = 1:d,
         C = (inv(A') * det(A'))';
         ci_invG = C(i,:) * invG{i};
         A(:,i) = (ci_invG * sqrt(tau / (ci_invG * C(i,:)')))';
      end

   end

end


A = A/(det(A)^(1/d)); % Normalisierung auf det(A) = 1
```

C Ergänzungen zur Implementierung

C.1 TIMIT Phoneme

TIMIT-Label	Beispiel	TIMIT-Label	Beispiel
pcl	p (closure)	bcd	b (closure)
tcl	t (closure)	dcl	d (closure)
kcl	k (closure)	gcl	g (closure)
p	**p**ea	b	**b**ee
t	**t**ea	d	**d**ay
k	**k**ey	g	**g**ay
q	ba**t**	dx	dir**t**y
ch	**ch**oke	jh	**j**oke
f	**f**ish	v	**v**ote
th	**th**in	dh	**th**en
s	**s**ound	z	**z**oo
sh	**sh**out	zh	a**z**ure
m	**m**oon	n	**n**oon
em	botto**m**	en	butto**n**
ng	si**ng**	eng	Washi**ng**ton
nx	wi**nn**er	el	bott**le**
l	**l**ike	r	**r**ight
w	**w**ire	y	a**h**ead
hh	**h**ay	hv	bott**le**
er	b**ir**d	axr	butt**er**
iy	b**ee**t	ih	b**i**t
ey	b**ai**t	eh	b**e**t
ae	b**a**t	aa	f**a**ther
ao	b**ou**ght	ah	b**u**t
ow	b**oa**t	uh	b**oo**k
uw	b**oo**t	ux	t**oo**t
aw	ab**ou**t	ay	b**i**te
oy	b**oy**	ax-h	s**u**spect
ox	**a**bout	ix	deb**i**t
epi	(epenthetic sil.)	pau	(pause)
h#	(silence)		

Tabelle C.1: Phonemlabel der TIMIT Datenbank.

C.2 HTK Konfigurationsparameter für MFCC Merkmale

Zur Berechnung der Melfrequenz Cepstralkoeffizienten mit Hilfe von HTK wurden die folgenden Konfigurationsparameter verwendet:

```
# Coding parameters
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAVE
TARGETKIND = MFCC
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
ZMEANSOURCE = T
USEHAMMING = F
PREEMCOEF = 0.97
NUMCHANS = 20
CEPLIFTER = 22
NUMCEPS = 12
```

Δ-Koeffizienten erster- und zweiter Ordnung wurden gebildet durch Erweiterung des Merkmalssatzes $y(t)$ um dessen erste und zweite Differenzen wie in den Untersuchungen von Eisele u. a. (1996) beschrieben: $y_\Delta(t) = y(t) - y(t - 20\,ms)$ und $y_{\Delta\Delta}(t) = y_\Delta(t) - y_\Delta(t - 20\,ms)$.

C.3 Verwendete Software

Für die Berechnungen dieser Arbeit wurden verschiedene Softwarepakete eingesetzt: ein Großteil der Berechnungen und Auswertungen sowie Skripte zur automatischen Berechnung des auditorischen Simulationsmodells auf den TIMIT Sprachdaten wurden mit dem statistischen Softwarepaket `R` (Ihaka und Gentleman, 1996), Version 2.6.2 (R Development Core Team, 2006) durchgeführt.

Zur Bearbeitung von *.wav Files wurde das `R` Paket `tuneR` (Ligges, 2006, Kapitel 7) sowie die freie Software SoX (Sound eXchange sourceforge.net, 2008) eingesetzt.
Zur Hidden Markov Modellierung wurde das Hidden Markov Toolkit `HTK` (v3.3, vgl. Young u. a., 2005) verwendet.
Die Merkmalsextraktion, lineare Dimensionsreduktion, sowie Skripte zur Erstellung automatischer Spracherkennungssysteme mit Hilfe von `HTK` erfolgte in `Matlab` 11.
Die Berechnung des auditorischen Simulationsmodells erfolgte mit dem in Zusammenarbeit mit der Fraunhofer IDMT (auf Basis von `C++`) entwickelten `EarAnalyzer` (Version 1.22b) unter Verwendung des in B.4 angegebenen Konfigurations-Dateien.
Zur Implementierung von Delay-Computing Netzwerken wurde das von Tamás Harczos entwickelte Programm `hough` (v0.1) und zur Implementierung der PLDCNs dessen eigene Erweiterung `plhwn` verwendet.

Im Laufe des Promotionsstudiums enstandene Publikationen

1. Szepannek, G. und Luebke, K. (2005): Different subspace classification. In: Weihs, C. und Gaul, W. (eds.): *Classification – the Ubiquitous Challenge*, Springer, Heidelberg, 224–231.

2. Roever, C. und Szepannek, G. (2005): Application of a genetic algorithm to variable selection in fuzzy clustering. In: Weihs, C. und Gaul, W. (eds.): *Classification – the Ubiquitous Challenge*, Springer, Heidelberg, 674–681.

3. Szepannek, G., Luebke, K. und Weihs, C. (2005): Understanding patterns with different subspace classification. In: Perner, P. und Imiya, A. (eds.): *Machine Learning and Data Mining in Pattern Recognition*, Springer LNAI 3587, Heidelberg, 110–119.

4. Szepannek, G., Klefenz, F. und Weihs, C. (2005): Schallanalyse - Neuronale Repräsentation des Hörvorgangs als Basis. *Informatik Spektrum 28(5), 289–295.*

5. Szepannek, G. und Weihs, K. (2006): Variable selection for discrimination of more than two classes if the data are sparse. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A. und Gaul, W. (eds.): *From Data and Information Analysis to Knowledge Engineering*, Springer, Heidelberg, 700–707.

6. Weihs, C., Szepannek, G., Ligges, U., Luebke, K. und Raabe, N. (2006): Local models in register classification by timbre. In: Batagelij, V., Bock, H., Ferligoj, A. und Ziberna, A. (eds): *Data Science and Classification*, Springer, Heidelberg, 313–322.

7. Szepannek, G. und Weihs, C. (2006): Local modelling in classification on different feature subspaces. In: Perner, P. (ed.): *Advances in Data Mining*, Springer LNAI 4065, Heidelberg, 226–238.

8. Szepannek, G., Harczos, T., Klefenz, F., Katai, A., Schikowski, P. und Weihs, C. (2006): Vowel classification by a perceptually motivated neurophysiologically parameterized auditory model. In: Decker, R., und Lenz, H. (eds.): *Advances in Data Analysis*, Springer, Heidelberg, 653–660.

9. Harczos, T., Szepannek, G., und Klefenz, F. (2006): Auditory model based vowel classification. *Proceedings of IEEE Biomedical Circuits and Systems (BioCAS)*, London/UK.

10. Szepannek, G., Bischl, B. und Weihs, C. (2007): On the combination of locally optimal pairwise classifiers. In: Perner, P.(ed.): *Machine Learning and Data Mining in Pattern Recognition*, Springer LNAI 4571, Heidelberg, 104–116.

11. Harczos, T., Szepannek, G. und Klefenz, F. (2007): Towards automatic speech recognition based on cochlear travelling wave delay trajectories. *Proceedings of International Symposium on Auditory and Audiological Research (ISAAR)*, Helsingoer/DK.

12. Harczos, T., Nogueira, W., Szepannek, G. und Klefenz, F. (2007): Comparative evaluation of successive cochlear modelling stages as possible front-ends for automatic speech recognition. *Proceedings of International Congress on Acoustics (ICA)*, Madrid/E.

13. Szepannek, G. (2008): *Different Subspace Classification – Verfahren zur Datenanalyse, -interpretation, -visualisierung und Vorhersage in hochdimensionalen Räumen.* vdm-Verlag, Saarbrücken, ISBN 978-3-8364-6302-7.

14. Szepannek, G., Schiffner, J., Wilson, J. und Weihs, C. (2008): Local modelling in classification. In: Perner, P. (ed.): *Advances in Data Mining*, Springer LNAI 5077, Heidelberg, 153–164.

15. Szepannek, G., Bischl, B. und Weihs, C. (2008): On the combination of locally optimal pairwise classifiers. *Journal of Engineering Applications of Artificial Intelligence*, *in Druck*.

Abkürzungsverzeichnis

Abkürzung	Bedeutung
aeM	auditorisch erweiterter Merkmalssatz
AFCC	auditorische Frequenz Cepstralkoeffizienten
AI	Auditory Image (auditorisches Muster)
ALSD	Average Localized Synchrony Detection
ALSR	Average Localized Synchronized Rate
AP	Aktionspotenzial
ASR	Automatic Speech rRcognition
BF	Best Frequency
BM	Basilarmembran
CF	Center Frequency
CI	Cochlea Implantat
dB	Dezibel
DCT	Diskrete Cosinus Transformation
EIH	Ensemble Interval Histogram
FFT	Fast Fourier Transformation
GSD	Generalized Synchony Detection
HDA	Heteroscedastic Discriminant Analysis
HSR	High Spontaneous Rate (nerve fibres)
HTK	Hidden Markov Toolkit
IAF	Integrate-And-Fire
ISI	Inter-Spike Intervall
LDA	Linear Discriminant Analysis
LIN	Lateral inhibitorisches neuronales Netzwerk
LSR	Low Spontaneous Rate (nerve fibres)
M-ANF	Wiederholte Simulation der Hörnervenantwort
MFCC	Mel-Frequenz Cepstralkoeffizienten
MSR	Medium Spontaneous Rate (nerve fibres)
PCA	Principal Component Analysis
PSTH	Post Stimulus Time Histogram
PVS	(Klassen-)paarweise Variablenselektion
RDA	Regularized Discriminant Analysis
RHDA	Regularized Heteroscedastic Discriminant Analysis
RI	Rate vs. Intensity
RRP	Readily Releasable Pool
S-ANF	Einmalige Simulation der Hörnervenantwort
SHDA	Smoothed Heteroscedastic Discriminant Analysis
SI	Synchronization Index
SPL	Sound Pressure Level (Schalldruck)
URHDA	Unbalanced Regularized Heteroscedastic Discriminant Analysis
ZCPA	Zero crossings with peak amplitudes

Tabelle C.2: Übersicht der in der Arbeit verwendeten Abkürzungen.

Index

Literaturverzeichnis

[Adrian 1928] Adrian, E.: *The Basis of Sensation: The Action of Sense Organs.* New York : W.W.Norton, 1928

[Ali 1999] Ali, A.: *Auditory-Based Acoustic-Phonetic Signal Processing for Robust Continous Speech Recognition*, University of Pennsylvenia, Dissertation, 1999

[Ali u. a. 2002] Ali, A. ; Spiegel, J. van der ; Mueller, P.: Robust auditory-based speech recognition using the average localized synchrony-detection. In: *Vol. 10, no.5* Proceedings of IEEE Transactions on Speech and Audio Processing (Veranst.), 2002, S. 279–292

[Allen 1994] Allen, J.: How do humans process and recognize speech? In: *IEEE Transactions on Speech And Signal Processing* 2 (1994), Nr. 4, S. 567–577

[Augustine und Charlton 1985] Augustine, G. ; Charlton, S.: Calcium entry and transmitter release at voltage-clamped nerve terminals of squid. In: *Journal of Physiology* (1985), Nr. 369, S. 163–181

[Baum u. a. 1970] Baum, L. ; Petrie, T. ; Soules, G. ; Weiss, N.: A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. In: *Annals of Mathematical Statistics* 41 (1970), Nr. 1, S. 164–171

[Baumgarte 2000] Baumgarte, F.: *Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung*, Universität Hannover, Dissertation, 2000

[Beutner u. a. 2001] Beutner, T. ; Voets, T. ; Neher, E. ; Moser, T.: Calcium dependence of exocytosis and endocytosis at the cochlear inner hair cell afferent synapses. In: *Neuron* (2001), Nr. 29, S. 681–690

[Bordag und Bordag 2003] Bordag, S. ; Bordag, D.: Advances in automatic speech recognition by imitating spreading activation. In: *Text, Speech and Dialogue*. Heidelberg : Springer, 2003 (LNCS 2807), S. 158–164

[Brandenburg und Stoll 1994] Brandenburg, K. ; Stoll, G.: ISO-MPEG-1 audio: a generic standard for high-quality audio coding. In: *Journal of the Audio Engineering Society* (1994), S. 780–792

[Brown 1987] Brown, P.: *The Acoustic-Modelling Problem in Automatic Speech Recognition*, Carnegie Mellon University, Dissertation, 1987

[Brückmann u. a. 2004] Brückmann, A. ; Klefenz, F. ; Wünsche, A.: A neural net for 2d-slope and sinusoidal shape detection. In: *International Scientific Journal of Computing* 3 (2004), Nr. 1, S. 21–26

[Burget 2004] Burget, L.: *Complementarity of Speech Recognition Systems and System Combination*. Czech Republic, Faculty of Information Technology, Brno University of Technology, Dissertation, 2004

[Crawford u. a. 1991] Crawford, A. ; Evans, M. ; Fettiplace, R.: The actions of calcium on the mechano-electrical transducer current of turtle hair cells. In: *Journal of Physiology* (1991), Nr. 434, S. 369–398

[Dau u. a. 1996] Dau, T. ; Püschel, D. ; Kohlrausch, A.: A quantitative model of the effective signal processing in the auditory system. I. model structure. In: *Journal of the Acoustical Society of America* 99 (1996), S. 3615–3622

[Davis und Mermelstein 1980] Davis, K. ; Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continously cpoken sentences. In: *IEEE Transactions on Acoustics Speech and Signal Processing* 28 (1980), Nr. 4, S. 357–366

[Delgutte 1983] Delgutte, B.: Speech coding in the auditory nerve II: processing schemes for vowel-like sounds. In: *Journal of the Acoustical Society of America* 75 (1983), Nr. 3, S. 879–886

[Delgutte und Kiang 1984] Delgutte, B. ; Kiang, N.: Speech coding in the auditory nerve I: vowel-like sounds. In: *Journal of the Acoustical Society of America* 75 (1984), Nr. 3, S. 866–878

[Dempster u. a. 1977] Dempster, A. ; Laird, N. ; Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society B* 39 (1977), Nr. 1, S. 1–22

[Di Pillo 1976] Di Pillo, P.: The application of bias to discriminant analysis. In: *Communications in Statistics - Theory and Methods* 5 (1976), Nr. 9, S. 843–854

[Di Pillo 1979] Di Pillo, P.: Biased discriminant analysis: evaluation of the optimum probability of misclassification. In: *Communications in Statistics - Theory and Methods* 8 (1979), Nr. 14, S. 1447–1457

[Dietterich 1998] Dietterich, T.: Approximate statistical tests for comparing supervised classification learning algorithms. In: *Neural Computation* 10 (1998), Nr. 7, S. 1895–1923

[Duchateau u. a. 2001] Duchateau, J. ; Demunynck, K. ; Compernolle, D. ; Wambacq, P.: Class definition in discriminant feature analysis. In: *Proceedings of Eurospeech 2001, Genf (CH) vol. 2* ICASSP (Veranst.), 2001, S. 1621–1624

[Eisele u. a. 1996] Eisele, T. ; Haeb-Umbach, R. ; Langmann, D.: A comparative study of linear feature transformation techniques for automatic speech recognition. In: *Proceedings of the International Conference on Speech and Language Processing (ICSLP), Philadelphia* ICASSP (Veranst.), 1996

[Ellis 2000] Ellis, D.: Stream combination before and/or after the acoustic model / International Computer Science Institute, Berkeley. 2000. – Forschungsbericht

[Erdogan 2005] Erdogan, H.: Regularizing heteroscedastic discriminant analysis for speech recognition. In: *Proceedings of the* 13^{th} *IEEE Signal Processing and Communications Applications Conference*, 2005, S. 107–110

[Fahrmeir u. a. 1996] Fahrmeir, L. ; Hamerle, A. ; Tutz, G.: *Multivariate Statistische Verfahren*. 2nd. New York : de Gruyter, 1996

[Fant 1960] Fant, G.: *The Acoustic Theory of Speech Production.* The Hague : Mouton and Co., 1960

[Fatt und Huxley 1952] Fatt, P. ; Huxley, B.: Spontaneous subthreshold activity at motor nerve endings. In: *Journal of Physiology* 1 (1952), Nr. 117, S. 109–128

[Fettweis 1986] Fettweis, A.: Wave digital filters: theory and practice. In: *Proc. IEEE 74* IEEE (Veranst.), 1986, S. 270–324

[Fisher 1936] Fisher, R.: The use of multiple measurements in taxonomic problems. In: *Annals of Eugenics* 7 (1936), Nr. 2, S. 179–188

[Friedman 1989] Friedman, J.: Regularized discriminant analysis. In: *Journal of the American Statistical Asscociation* 84 (1989), S. 165–175

[Gabbiani 2005] Gabbiani, F.: *Quantal Hypothesis and Stochastic Models of Synaptic Release.* Lecture in theroretical Neuroscience, Department of Computational and Applied Mathematics, Rice Univ., `http://www.caam.rice.edu/~caam415/lec2/g1`. Spring 2005

[Gales 1998] Gales, M.: Semi-tied covariance matrices for hidden Markov models. In: *IEEE Transactions on Speech and Audio Processing* 7 (1998), S. 272–281

[Gales 1999] Gales, M.: Maximum likelihood multiple projection schemes for hidden Markov models / Cambridge University. 1999 (CUED/F-INFENG/TR.365). – Forschungsbericht

[Garofolo u. a. 1993] Garofolo, J. ; Lamel, L. ; Fiesher, W. ; Fiscus, J. ; Palett, D. ; Dahlgren, N.: DARPA TIMIT acoustic-phonetic continuous speech corpus / National Institute of Standards and Technology (NIST). NIST, Gaithersburgh, MD, 1993 (NISTIR 4930). – Forschungsbericht

[Gebeshuber 2000] Gebeshuber, I.: The influence of stochastic behavior on the human threshold of hearing. In: *Chaos Solitions and Fractals* 11 (2000), S. 1855–1868

[Geisler 1998] Geisler, J.: *From Sound to Synapse.* Nework : Oxford University Press, 1998

[Gerstner 1998] Gerstner, W.: Spiking neurons. In: Maass, W. (Hrsg.) ; Bishop, C. (Hrsg.): *Pulsed Neural Networks*. Cambridge, Massechusetts : MIT Press, 1998, Kap. 1, S. 3–53

[Ghitza 1988] Ghitza, O.: Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. In: *Journal of Phonetics* 16 (1988), S. 109–123

[Ghitza 1994] Ghitza, O.: Auditory models and human performance in tasks related to speech coding and speech recognition. In: *IEEE Transactions on Speech and Audio Processing* 2 (1994), Nr. 1, S. 115–132

[Giron 2006] Giron, F.: Correlation analysis for the derivation of speech recognition features based on an auditory model. In: *TC-STAR Workshop on Speech-to-Speech Translation*. Barcelona, Spain, 2006, S. 159–164. – URL `http://www.tc-star.org/ws06/pdfs/asr/tcstar06_giron.pdf`

[Gold und Morgan 2000] Gold, B. ; Morgan, N.: *Speech and Audio Signal Processing*. New York : Wiley, 2000

[Goldstein 1990] Goldstein, J.: Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass nonlinearity filtering. In: *Hearing research* 49 (1990), S. 39–60

[Gouws u. a. 2004] Gouws, E. ; Woolvaardt, K. ; Kleynhans, N. ; Barnard, E.: Appropriate baseline values for HMM-based speech recognition / Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa. 2004. – Forschungsbericht

[Grayden u. a. 2004] Grayden, D. ; Burkitt, A. ; Kenny, O. ; Clarey, J. ; Paolini, A. ; Clark, G.: A cochlear implant speech processing strategy based on an auditory model. In: *Proceedings of Intelligent Sensors, Sensor Networks and Information Processing Conference*, 2004, S. 491–496

[Greenberg 1997] Greenberg, S.: The significance of the cochlear travelling wave for theories of frequency analysis and pitch. In: Lewis, E. (Hrsg.) ; Steele,

C. (Hrsg.) ; Lyon, R. (Hrsg.): *Diversity in Auditory Mechanics*. World Scientific Publishing, 1997

[Greenberg u. a. 1997] Greenberg, S. ; Poeppel, D. ; Roberts, T.: *A space-time theory of pitch and timbre based on cortical expansion of the cochlear travelling wave*. Proceedings of the XIth International Symposium on Hearing, Grantham. 1997

[Greenwood 1990] Greenwood, D.: A cochlear frequency-position function for several species – 29 years later. In: *Journal of the Acoustical Society of America* 87 (1990), Nr. 6, S. 2592–2602

[Grüning und Kropf 2006] Grüning, M. ; Kropf, S.: A Ridge classification method for high dimensional observations. In: Spiliopolou, M. (Hrsg.) ; Kruse, R. (Hrsg.) ; Borgelt, C. (Hrsg.) ; Nürnberger, A. (Hrsg.) ; Gaul, W. (Hrsg.): *From Data and Information Analysis to Knowledge Engineering*. Heidelberg : Springer, 2006, S. 684–691

[Guyon und Elisseeff 2003] Guyon, I. ; Elisseeff, A.: An introduction to variable and feature selection. In: *The Journal of Machine Learning Research* 3 (2003), S. 1157 – 1182

[Harczos 2007] Harczos, T.: *Persönliches Gespräch*. 2007

[Harczos u. a. 2006a] Harczos, T. ; Klefenz, F. ; Katai, A.: A neurobiologically inspired vowel recognizer using Hough-transform – a novel approach to auditory image processing. In: *Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP), Vol.1*, 2006, S. 251–256

[Harczos u. a. 2007a] Harczos, T. ; Nogueira, W. ; Szepannek, G. ; Klefenz, F.: Comparative evaluation of successive cochlear modelling stages as possible front-ends for automatic speech recognition. In: *Proceedings of the International Congress on Acoustics (ICA)*. Madrid, 2007

[Harczos u. a. 2006b] Harczos, T. ; Szepannek, G. ; Katai, A. ; Klefenz, F.: Auditory model based vowel recognition. In: *Proceedings of IEEE Biomedical Circuits and Systems Conference, London (BioCAS)*, 2006

[Harczos u. a. 2007b] Harczos, T. ; Szepannek, G. ; Klefenz, F.: Towards automatic speech recognition based on cochlear travelling wave delay trajectories. In: *Proceedings of International Symposium on Auditory and Audiological Research (ISAAR)*. Helsingoer, 2007

[Hastie u. a. 1995] Hastie, T. ; Buja, A. ; Tibshirani, R.: Penalized discriminant analysis. In: *Annals of Statistics* 23 (1995), S. 73–102

[Hastie u. a. 2001] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The Elements of Statistical Learning*. Bd. 1. Springer, 2001

[Heinz 2002] Heinz, T.: *Ein physiologisch gehörgerechtes Verfahren zur automatisierten Melodietranskription*, TU Ilmenau, Dissertation, 2002

[Hemmert 2005] Hemmert, W.: *Persönliches Gespräch.* 2005

[Hemmert u. a. 2004] Hemmert, W. ; Holmberg, M. ; Gelbart, D.: Auditory-based automatic speech recognition workshop on statistical and perceptual audio processing, Jeju, Korea. In: *Proceedings of the International Conference of Acoustics, Speech and Signal Processing (ICASSP)*, 2004

[Hermansky 1997] Hermansky, H.: Should recognizers have ears ? In: *Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997, S. 1–10

[Hodgkin und Huxley 1952] Hodgkin, A. ; Huxley, A.: A quantitative description of ion currents and its application to conduction and excitation in nerve membranes. In: *Journal of Physiology* (1952), Nr. 117, S. 500–544

[Holm 1979] Holm, S.: A simple sequentially rejective multiple test proceedure. In: *Scandinavian Journal of Statistics* 6 (1979), S. 65–70

[Holmberg u. a. 2007] Holmberg, M. ; Gelbart, D. ; Hemmert, W.: Speech encoding in a model of peripheral auditory processing: quantitative assessment by means of automatic speech recognition. In: *Speech Communication* 49 (2007), S. 917–932

[Holmberg und Hemmert 2004] Holmberg, M. ; Hemmert, W.: An auditory model for coding speech into nerve action potentials. In: *Proceedings of Joint Congress of the Congrès Francais d'Acoustique and German Annual Conference on Acoustics (CFA/DAGA)*, 2004, S. 703–704

[Holmes u. a. 2004] Holmes, S. ; Sumner, C. ; O'Mard, L. ; Meddis, R.: The temporal representation of speech in a nonlinear model of the guniea pig cochlea. In: *Journal of the Acoustical Society of America* 116 (2004), Nr. 6, S. 3534–3545

[Horn und Vollandt 1995] Horn, M. ; Vollandt, R.: *Multiple Tests und Auswahlverfahren.* Stuttgart : Fischer, 1995

[Howard und Angus 2006] Howard, D. ; Angus, J.: *Acoustics and psychoacoustics.* 3. Amsterdam : Focal Press in Elsevier, 2006

[Hudspeth 1989] Hudspeth, A.: How the ear's works work. In: *Nature* (1989), October, Nr. 341, S. 397–404

[Hudspeth und Lewis 1988] Hudspeth, A.. ; Lewis, R.: Kinetic anaylsis of voltage- and ion-sependent conductances in saccular hair cells of the bull frog Rana Cateebebneiana. In: *Journal of Physiology* (1988), Nr. 400, S. 237–274

[Ihaka und Gentleman 1996] Ihaka, R. ; Gentleman, R.: R: a language for data analysis and graphics. In: *Journal of Computational and Graphical Statistics* 5 (1996), Nr. 3, S. 299–314

[Ivanov und Petrovsky 2004] Ivanov, A. ; Petrovsky, A.: Anthropomorphic feature extraction algorithm for speech recognition in adverse environments. In: *Proceedings of Speeecom'04 in St. Petersburg*, 2004, S. 166–173

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

[Johnson 1974] Johnson, D.: *The Response of Single Auditory-Nerve Fibres in Cats to Single Tones: Synchrony and Average Discharge Rate.* Cambridge, MA, Massachusetts Institute of Technology, Dissertation, 1974

[Johnson 1980] Johnson, D.: The relationship between spike rate and synchrony in responses of auditory-nerve fibres to single tones. In: *Journal of the Acoustical Society of America* 68 (1980), Nr. 4, S. 1115–1122

[Johnston und Wu 1995] Johnston, D. ; Wu, S.: *Foundations of Cellular Neurophysiology.* Cambridge, MA : MIT University Press, 1995

[Jourdain 2001] Jourdain, R.: *Das wohltemperierte Gehirn – Wie Musik im Kopf entsteht und wirkt (dt. Übersetzung).* Spektrum Akademischer Verlag, 2001

[Jung 2004] Jung, H.: Filtering of filter-bank energies for robust speech recognition. In: *ETRI Journal* 26 (2004), Nr. 3, S. 273–57

[Katz 1987] Katz, S.: Estimation of probabilities from sparse date for the language model component of a speech recognizer. In: *IEEE Transactions on Acoustics, Speech and Signal processing* 35 (1987), Nr. 3, S. 400–401

[Kidd und Weiss 1990] Kidd, R. ; Weiss, T.: Mechanisms that degrade timing information in the cochlea. In: *Hearing Research* (1990), Nr. 49, S. 181–208

[Kim u. a. 1999] Kim, D. ; Lee, S. ; Kil, R.: Auditory processing of speech signals for robust speech recognition in real-world noisy environments. In: *IEEE Transactions on Speech and Audio Signal Processing* 7 (1999), Nr. 1, S. 55–69

[Kim und Molnar 1980] Kim, D. ; Molnar, C.: Cochlear meachnics: nonlinear behaviour in two-tone responses and in ear canal sound pressure. In: *Journal of the Acoustical Society of America* 67 (1980), S. 1704–1721

[Kumar 1997] Kumar, N.: *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.* Baltimore, Maryland, USA, Johns Hopkins University, Dissertation, 1997

[Kumar und Andreou 1996] Kumar, N. ; Andreou, A.: On generalizations of linear discriminant analysis / Johns Hopkins University. 1996 (Tech. Rep. JHU/ECE-9607). – Forschungsbericht

[Kumar und Andreou 1998] Kumar, N. ; Andreou, A.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. In: *Speech Communication* 25 (1998), Nr. 4, S. 283–297

[Läuter 1992] Läuter, J.: *Stabile multivariate Verfahren. Diskriminanzanalyse, Regressionsanalyse, Faktoranalyse.* Berlin : Akademie Verlag, 1992

[Lee und Hon 1989] Lee, K. ; Hon, F.: Speaker-independent phone recognition using hidden markov models. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 37 (1989), Nr. 11, S. 1641–1648

[Ligges 2006] Ligges, U.: *Transkription monophoner Gesangszeitreihen*, Fachbereich Statistik, Universität Dortmund, Dissertation, 2006

[Longtin 1993] Longtin, A.: Stochastic resonance in neuron models. In: *Journal of Statistical Physics* 1/2 (1993), Nr. 70, S. 309–327

[MacQueen 1996] MacQueen, J.: Computational analysis of hair cell and auditory nerve processes. In: Hawkins, H. (Hrsg.) ; Mc Mullen, T. (Hrsg.) ; Popper, A. (Hrsg.) ; Fay, R. (Hrsg.): *Auditory Computation.* New York : Springer, 1996 (Handbook of Auditory Research), S. 121–156

[Mamsch 2006] Mamsch, M.: *Automatic Speech Recognition on Acoustical Features Generated by a Biologically Motivated Auditory Model.* Universität Magdeburg, Lehrstuhl für Kognitive Systeme, Fakultät für Elektro- und Informationstechnik, Diplomarbeit, 2006

[Markin und Hudspeth 1995] Markin, V. ; Hudspeth, A.: Gating-spring models of mechanoelectrical transduction by hair cells of the internal ear. In: *Annual Review of Biophysics and Biomolecular Structure* (1995), Nr. 24, S. 59–83

[Marks und Dunn 1974] Marks, S. ; Dunn, O.: Discriminant functions when the covariance matrices are unequal. In: *Journal of the American Statistical Association* 69 (1974), Nr. 346, S. 555–559

[Mc Dermott 2004] Mc Dermott, H.: Music perception with cochlear implants: a review. In: *Trends in Amplification* 8 (2004), S. 49–82

[Meddis 1986] Meddis, R.: Simulation of mechanical to neural transduction in the auditory receptor. In: *Journal of the Acoustical Society of America* 3 (1986), Nr. 79, S. 702–711

[Meddis 2006] Meddis, R.: Auditory-nerve first-spike latency and auditory absolute threshold: a computer model. In: *Journal of the Acoustical Society of America* 119 (2006), Nr. 1, S. 406–417

[Merhav und Lee 1993] Merhav, N. ; Lee, C.: On the asymptotic statistical behaviour of empirical cepstral coefficients. In: *IEEE Transactions on Signal Processing* 41 (1993), May, Nr. 5, S. 1990–1993

[Miller und Sachs 1983] Miller, M. ; Sachs, M.: Representation of stop consonants in the discharge patterns of auditory-nerve fibres. In: *Journal of the Acoustical Socitety of America* 74 (1983), Nr. 2, S. 502–517

[Moore 2003] Moore, B.: Coding of sounds in the auditory system and its relevance to signal processing and coding cochlear implants. In: *Otology and Neurotology* 24 (2003), S. 243–254

[Mountain und Cody 1999] Mountain, D. ; Cody, A.: Multiple modes of inner hair cell stimulation. In: *Hearing Research* (1999), Nr. 132, S. 1–14

[Munich und Lin 2005] Munich, M. ; Lin, Q.: Auditory image model features for automatic speech recognition. In: *Proceedings of Interspeech Conference 2005, Lisbon*, IEEE ICASSP, 2005

[Nadeu und Gorricho 1995] Nadeu, Hernando J. ; Gorricho, M.: On the decorrelation of filterbank-energies in speech recognition. In: *Proceedings of Eurospeech 1995 Conference* ISCA (Veranst.), 1995, S. 1381–1384

[Neher 2003] Neher, E.: Pegasus im Nevenland. In: *Georgia Augusta* (2003), Nr. 2, S. 53–57

[Palmer 1990] Palmer, A.: The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge

patterns of guinea pig cochlear-nerve fibres. In: *Journal of the Acoustical Society of America* 88 (1990), Nr. 3, S. 1412–1426

[Palmer u. a. 1986] Palmer, A. ; Winter, I. ; Darwin, C.: The representation of steady-state vowel sounds in the temproal discharge patterns of the guinea pig cochlear nerve and primarylike cochlear nucleus neurons. In: *Journal of the Acoustical Society of America* 79 (1986), S. 100–113

[Perdigao und Sa 1998] Perdigao, F. ; Sa, L.: Auditory models as front-ends for speech recognition. In: *Procedings of NATO ASI on Computational Hearing*, 1998

[Pfeiffer und Kim 1975] Pfeiffer, R. ; Kim, D.: Cochlear nerve fibre responses: distribution along the cochlear partition. In: *Journal of the Acoustical Society of America* 58 (1975), S. 867–869

[R Development Core Team 2006] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria: , 2006. – URL `http://www.R-project.org`. – ISBN 3-900051-07-0

[Rabiner 1989] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE* 77 (1989), Feb., Nr. 2, S. 257–285

[Rattay und Gitter 1998] Rattay, I. ; Gitter, A.: The mammalian auditory hair cell: a simple electric circuit model. In: *Journal of the Acoustical Society of America* (1998), Nr. 103, S. 1558–1565

[Rieke u. a. 1997] Rieke, F. ; Warland, D. ; de Ruyter van Stevenick, R. ; Bialek, W.: *Spikes - Exploring the Neural Code*. MIT Press, 1997

[Rose u. a. 1967] Rose, J. ; Brugge, J. ; Anderson, D. ; Hind, J.: Phase-locked response to low-frequency tones in single auditory nerve fibres of the squirrel monkey. In: *Journal of Neurophysiology* 30 (1967), S. 767–793

[Sachs und Young 1979] Sachs, M. ; Young, E.: Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. In: *Journal of the Acoustical Society of America* 66 (1979), Nr. 2, S. 470–479

[Sandhu und Ghitza 1995] Sandhu, S. ; Ghitza, O.: A comparative study of Mel cepstra and EIH for phone classification under adverse conditions ICASSP (Veranst.), 1995, S. 409–412

[Schafföner u. a. 2003] Schafföner, M. ; Katz, M. ; Krüger, S. ; Wendemuth, A.: Improved robustness of automatic speech recognition using a new class definition in linear discriminant analysis. In: *Proceedings of Eurospeech 2003 Conference, Genf (CH)* ICASSP (Veranst.), 2003, S. 2841–2844

[Schikowski 2006] Schikowski, P.: *Implementationen und Untersuchungen zur Aktionspotentialgenerierung des auditorischen Nervs durch gequantelte Transmitterausschüttungen der inneren Haarzelle.* Januar 2006. – Studienarbeit, Fraunhofer IDMT, Ilmenau

[Schmidt und Trenkler 1998] Schmidt, K. ; Trenkler, G.: *Moderne Matrix-Algebra.* Heidelberg : Springer, 1998

[Schoonhoven u. a. 1997] Schoonhoven, R. ; Prijs, V. ; Frijns, H.: Transmitter release in inner hair cell synapses: a model analysis of spontaneous and driven rate properties of cochlear nerve fibres. In: *Hearing Research* (1997), Nr. 113, S. 247–260

[Schukat-Talamazzini 1995] Schukat-Talamazzini, E.-G.: *Automatische Spracherkennung - Statistische Verfahren der Musteranalyse.* 1995

[Secker-Walker und Searle 1990] Secker-Walker, H. ; Searle, C.: Time-domain analysis of auditory-nerve fibre firing rates. In: *Journal of the Acoustical Society of America* 88 (1990), Nr. 3, S. 1427–1436

[Seneff 1988] Seneff, S.: A joint synchrony/mean-rate model of auditory speech processing. In: *Journal of Phonetics* 16 (1988), S. 55–76

[Shamma 1985a] Shamma, S.: Speech processing in the auditory system I: the representation of speech sounds in the response of the auditory nerve. In: *Journal of the Acoustical Society of America* 78 (1985), Nr. 5, S. 1612–1621

[Shamma 1985b] Shamma, S.: Speech processing in the auditory system II: lateral inhibition and the central processing of speech evoked activity in the auditory nerve. In: *Journal of the Acoustical Society of America* 78 (1985), Nr. 5, S. 1622–1632

[Shamma u. a. 1986] Shamma, S. ; Chadwick, R. ; Wilbur, J. ; Morrish, K. ; Rinzel, J.: A biophysical model of cochlear processing: intensity dependence of pure tone responses. In: *Journal of the Acoustical Society of America* 80 (1986), Nr. 1, S. 133–145

[Shannon und Weaver 1949] Shannon, C. ; Weaver, W.: *The Mathematical Theory of Communication*. Urbana : University of Illinois Press, 1949

[Siegel 1992] Siegel, J.: Spontaneous synaptic potentials from afferent terminals in the guinea pig cochlea. In: *Hearing Research* (1992), Nr. 59, S. 85–92

[Slaney 1988] Slaney, M.: Lyon's Cochlear Model / Advanced Technology Group, Apple. `citeseer.ist.psu.edu/442882.html`, 1988 (13). – Forschungsbericht

[sourceforge.net 2008] sourceforge.net: *Sound eXchange Version 14.0.1*. `http://sox.sourceforge.net/`. 2008

[Sumner u. a. 2003a] Sumner, C. ; Lopez-Poveda, E. ; O'Mard, L. ; Meddis, R.: Adaption in a revised inner-hair cell model. In: *Journal of the Acoustical Society of America* 2 (2003), Nr. 113, S. 893–901

[Sumner u. a. 2003b] Sumner, C. ; O'mard, L. ; Lopez-Poveda, E. ; Meddis, R.: A nonlinear filter-bank model of the guinea-pig cochlear nerve: rate responses. In: *Journal of the Acoustical Society of America* 113 (2003), Nr. 6, S. 3264–3274

[Sumner u. a. 2002] Sumner, C. ; Poveda, E. ; O'Mard, L. ; Meddis, R.: A revised model of the inner hair cell and auditory nerve complex. In: *Journal of the Acoustical Society of America* 111 (2002), Nr. 5, S. 2178–2188

[Szepannek u. a. 2006] Szepannek, G. ; Harczos, T. ; Klefenz, F. ; Katai, A. ; Schikowski, P. ; Weihs, C.: Vowel classification by a perceptually motivated neurophysiologically parameterized auditory model. In: Decker, R. (Hrsg.) ; Lenz, H. (Hrsg.): *Advances in Data Analysis*. Heidelberg : Springer, 2006, S. 653–660

[Szepannek u. a. 2005] Szepannek, G. ; Klefenz, F. ; Weihs, C.: Schallanalyse - Neuronale Repräsentation des Hörvorgangs als Basis. In: *Informatik Spektrum* 28 (2005), Nr. 5, S. 389–395

[Szepannek u. a. 2003] Szepannek, G. ; Luebke, K. ; Weihs, C.: Gruppierung der Spielwiesen von Vereinen in der Fussballbundesligasaison 2002/2003 mit Hilfe von Clusteranalysen / Fachbereich Statistik, Universität Dortmund. 2003 (4). – Forschungsbericht

[Szepannek u. a. 2008] Szepannek, G. ; Schiffner, J. ; Wilson, J. ; Weihs, C.: Local modelling in classification. In: Perner, P. (Hrsg.): *Advances in Data Mining*. Heidelberg : Springer LNAI 5077, 2008, S. 153–164

[Szepannek und Weihs 2006a] Szepannek, G. ; Weihs, C.: Explorative development of information extraction schemes for speech recognition from simulated auditory neural response data via parallel local Hubel-Wiesel networks / Fachbereich Statistik, Universität Dortmund. 2006 (2). – Forschungsbericht

[Szepannek und Weihs 2006b] Szepannek, G. ; Weihs, C.: Variable selection for discrimination of more than two classes if the data are sparse. In: Spiliopolou, M. (Hrsg.) ; Kruse, R. (Hrsg.) ; Borgelt, C. (Hrsg.) ; Nürnberger, A. (Hrsg.) ; Gaul, W. (Hrsg.): *From Data and Information Analysis to Knowledge Engineering*. Heidelberg : Springer, 2006, S. 700–707

[Tchorz und Kollmeier 1999] Tchorz, J. ; Kollmeier, B.: A model of auditory perception as front end for automatic speech recognition. In: *Journal of the Acoustical Society of America* 106 (1999), Nr. 4, S. 2040–2050

[Weihs und Heilemann 2000] Weihs, C. ; Heilemann, U.: Diskriminanzanalyse. In: Voss, W. (Hrsg.): *Taschenbuch der Statistik*. München : Fachbuchverlag Leipzig, Carl Hanser Verlag, 2000, S. 683–608

[Weihs u. a. 2005] Weihs, C. ; Ligges, U. ; Luebke, K. ; Raabe, N.: klaR – analyzing German business cycles. In: Baier, D. (Hrsg.) ; Becker, R. (Hrsg.) ; Schmidt-Thieme, L. (Hrsg.): *Data Analysis and Decision Support*. Berlin : Springer, 2005, S. 335–343

[Weihs u. a. 2007] Weihs, C. ; Ligges, U. ; Mörchen, F. ; Müllensiefen, D.: Classification in Music Research. In: *Adavances in Data Analysis and Classification* 1 (2007), Nr. 3, S. 255–291

[Werner 2008] Werner, S.: *Trennung von Nutzsignalen und Rauschsignalen auf Basis von Vesikelfilterung in einem neuronalen auditorischen Modell*. Fraunhofer Institut für Digitale Medientechnologie, Ilmenau, Fakultät für Elektotechnik und Informationstechnik, Institut für Medientechnik, TU Ilmenau, Diplomarbeit, 2008

[Werner und Fodroczi 2006] Werner, S. ; Fodroczi, Z.: 2006. – Arbeiten zur Modellierung von Bushy Cells am IDMT

[Westerman und Smith 1984] Westerman, L. ; Smith, R.: Rapid and short-term adaption in auditory nerve responses. In: *Hearing Research* (1984), Nr. 15, S. 249–260

[Willett 2000] Willett, D.: *Beiträge zur Statistischen Modellierung und Effizienten Dekodierung in der Automatischen Spracherkennung*, Universität Duisburg, Dissertation, 2000

[Winter u. a. 1990] Winter, I. ; Robertson, D. ; Yates, G.: Diversity of characteristic frequency rate-intensity functions in guinea pig auditory nerve fibres. In: *Hearing Research* (1990), Nr. 45, S. 191–202

[Wolpert 2007] Wolpert, R.: *Persönliches Gespräch*. 2007

[Yang u. a. 1992] Yang, X. ; Wang, K. ; Shamma, S.: Auditory representaions of acoustic signals. In: *IEEE Transactions on Information Theory* 38 (1992), Nr. 2, S. 824–839

[Young 1992] Young, S.: The general use of tying in phoneme-based HMM speech recognisers. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1992, S. 569–572

[Young u. a. 2005] Young, S. ; Evermann, G. ; Gales, M. ; Hain, T. ; Kershaw, D. ; Moore, G. ; Odell, J. ; Ollason, D. ; Povey, D. ; Valtchev, V. ; Woodland, P.: *The HTK Book (for HTK Version 3.3)*. http://htk.eng.cam.ac.uk/docs/docs.shtml : Cambridge University Engineering Department, 2005

[Zwicker und Peisl 1990] Zwicker, E. ; Peisl, W.: Cochlear processing in analog models, in digital models and in human inner ear. In: *Hearing Research* (1990), Nr. 44, S. 209–216