Von heiseren Handys bis zum Ohr als Chip

Von Jürgen Tchorz, Michael Kleinschmidt, Volker Hohmann und Birger Kollmeier

Bei der Mobilkommunikation soll die übertragene Sprache möglichst natürlich und unverfälscht klingen, obwohl nur ein kleiner Teil der Sprachinformation zum Empfänger gesendet werden kann. Damit das Handy nicht "heiser" klingt, weil die falsche Information gesendet wird, benötigt man in jedem Handy ein (vereinfachtes) Modell des menschlichen Gehörs.

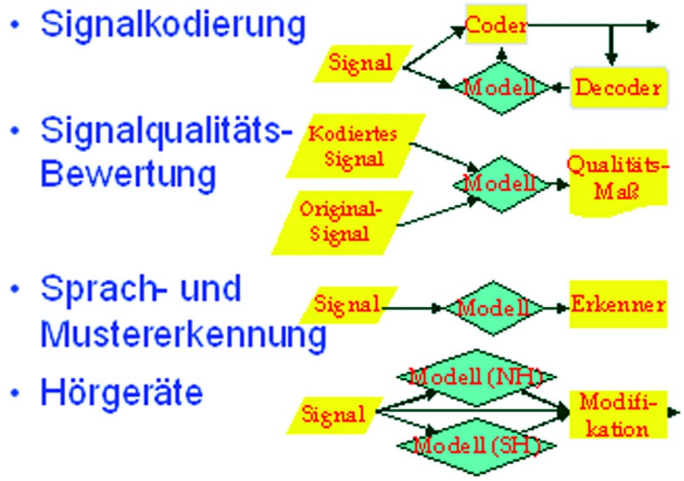
Die Anwendung von Kenntnissen unseres Hörvorgangs bringt für die sprachliche Mensch-Maschine-Kommunikation und die Telekommunikation einige Vorteile. Objektive Beurteilung der Sprachübertragung bei Handys, Datenkompression für die Musik- und Sprachübertragung im Internet und robuste automatische Spracherkennung sind einige Beispiele, bei denen der Computer zuerst das “richtige Hören” lernen muss.

A detailed knowledge of the signal processing in our ear is advantageous for human-machine communication with natural speech and telecommunication: Objective assessment of speech transmission quality in cellular phones, audio and speech data compression for the internet as well as robust automatic speech recognition are examples where computers first have to learn to ”hear correctly”.

Die digitale Verarbeitung von Sprache und Musik begegnet uns in vielen Formen und ist aus unserem täglichen Leben kaum noch wegzudenken. Mobiltelefone beispielsweise wandeln die Stimme des Sprechers digital um und versenden sie in einer komprimierten Form an den Empfänger, wo sie wieder in hörbare Sprache umgewandelt wird. Auf CDs wird Musik in hoher Qualität digital abgespeichert und kann von Abspielgeräten wiedergegeben werden. In diese technischen Anwendungen fließt dabei stets Wissen über unser Gehör und seine Eigenschaften ein. Beispielsweise darüber, welche Töne und Frequenzen wir hören können, und welche nicht mehr. So werden auf CDs keine Frequenzen abgespeichert, die oberhalb von etwa 20.000 Hertz liegen. Technisch wäre das möglich, aber wir können solch hohe Frequenzen nicht mehr wahrnehmen. Daher wäre es unsinnig, dafür Speicherplatz zu verbrauchen. Soll bei gleichbleibender Klangqualität der Speicherbedarf weiter verringert werden, so muss noch tiefer in die Trickkiste der gehörorientierten Musikverarbeitung gegriffen werden. Das derzeit sehr populäre MP3-Verfahren zur Komprimierung von Musikdateien nutzt dabei sogenannte Maskierungseffekte aus. Von Maskierung spricht man, wenn ein lautes Geräusch ein leiseres Geräusch überdeckt und quasi "unhörbar" macht. Dadurch, dass bei MP3 nur die gehörrelevanten Musikanteile extrahiert und gespeichert werden, wird eine Verringerung des Speicherbedarfs um etwa 90 Prozent erreicht, bei (fast) unhörbaren Qualitätsverlusten. Und so ist MP3 ein gutes Beispiel dafür, wie Erkenntnisse aus der Gehörforschung erfolgreich in alltägliche Anwendungen umgesetzt werden können.

Das Ohr als Computermodell

Auch am Fachbereich Physik der Universität Oldenburg wird seit mehreren Jahren Gehörforschung betrieben, einerseits Grundlagenforschung, andererseits aber auch in Hinblick auf die Anwendung in verschiedenen Bereichen der digitalen Sprach-



Schematischer Einsatz von Gehör-Modellen für verschiedene Anwendungen der Computer-unterstützten Kommunikation mit Audio- und Sprach-signalen

verarbeitung. Ein Beispiel dafür ist ein Gehörmodell (Perzeptionsmodell), welches im Oldenburger Graduiertenkolleg "Psychoakustik" entwickelt wurde (vgl. auch Torsten Dau: "Modell der effektiven Signalverarbeitung im Gehör", EINBLICKE 29, 1999). Dabei handelt es sich quasi um eine Computersimulation der ersten Verarbeitungsschritte der Gehörbahn, vom Innenohr über den Hörnerv bis hin zur "internen Repräsentation" im Gehirn, d.h. das Muster von Nervenimpulsen, mit denen die Information über den akustischen Schall im Gehirn dargestellt und weiterverarbeitet werden. Die Computersimulation extrahiert dafür diejenigen Informationen und Merkmale aus dem Eingangsschall, die auch für unser Gehirn wichtig sind zur Bewältigung der verschiedenen "akustischen Aufgaben" des Alltags, wie z. B. Sprachverstehen oder das Erkennen bestimmter Geräusche. Ein derartiges Modell lässt sich für eine Reihe von technischen Anwendungen einsetzen (vgl. Abb. oben).

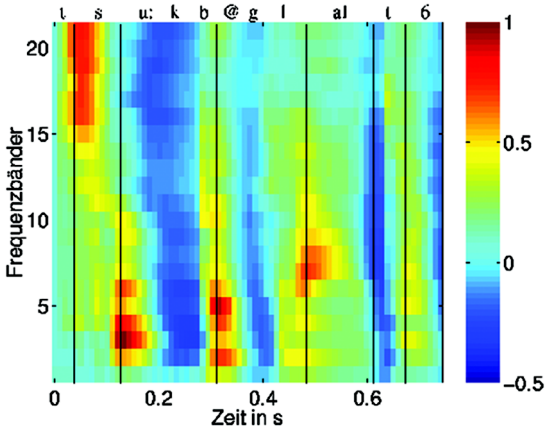
Bei der am Beispiel von MP3 schon erwähnten *Signalkodierung* soll ein kodiertes (z.B: datenkomprimiertes und über das Internet übertragenes) Signal nach der Dekodierung vom Hörer als identisch mit dem Original-Signal wahrgenommen werden. Ein anderes Beispiel ist die Entwicklung von neuen Verfahren zur Kodierung und Übermittlung von Sprache in Mobiltelefonen (Handys). Hier soll einerseits die Datenrate möglichst niedrig sein, um die zur Verfügung stehenden Kanäle optimal auszunutzen, andererseits aber muss die Sprachqualität, die beim Empfänger ankommt, möglichst hoch sein, d.h. es darf keine störenden Klangverzerrungen ("heiseres Handy") ausweisen.

Dies ist der Fall, wenn seine (durch das Modell berechnete) interne Repräsentation mit der des Originalsignals übereinstimmt. Als Fehlermaß für einen optimalen Kodierer sollte daher der Abstand auf der Ebene der internen Repräsentation (d. h. am Ausgang des Modells) verwendet werden. Dasselbe grundlegende Schema kann auch für die (objektive) Beurteilung eines (durch ein Übertragungssystem verfälschtes) Signal angewandt werden, bei dem die Abweichung zum Original-Signal auf der Ebene der internen Repräsentation ein Maß für die subjektiv empfundenen Qualitätseinbußen darstellt. Auf diesem Prinzip beruhen objektive Verfahren zur Beurteilung der Sprachübertragungsqualität sowie Ansätze zur Beurteilung von Audio-Übertragungsqualität, bei denen das Oldenburger Perzeptionsmodell sich inzwischen hervorragend bewährt hat. Hier ist eine automatische und verlässliche Beurteilung der erreichten Sprachqualität sehr wichtig, da im Entwicklungsstadium solcher Verfahren ständige Beurteilungen durch eine große Anzahl von Versuchspersonen viel zu aufwendig sind. Individuelle Beurteilungen einzelner Entwickler haben jedoch nur eine geringe Aussagekraft. Hier hat sich gezeigt, dass ein gehörbasiertes, automatisches Verfahren die Beurteilung vieler Versuchspersonen im Durchschnitt gut vorhersagen kann und damit die Entwicklung von Verfahren zur Sprachkodierung deutlich erleichtert.

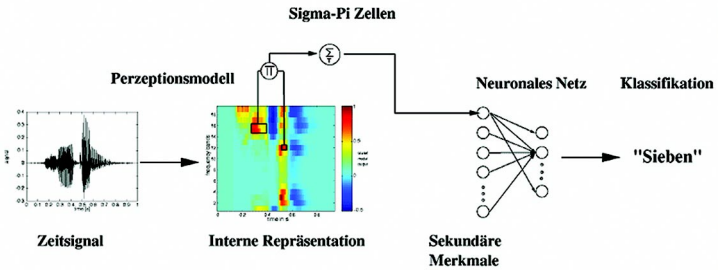
Neben einer automatischen Bestimmung der Übertragungsqualität einer Handy-Unterhaltung erlaubt das Modell auch die Vorhersage der Klangqualität von Hörgeräten. Gerade bei Hörgeräten erschließt sich noch eine weitere, Gehörmodell-basierte Anwendung: Um dem individuellen Schwerhörenden approximativ dieselbe interne Repräsentation des akustischen Signals wie dem mittleren Normalhörenden zu vermitteln, kann in einem Hörgeräte-Algorithmus versucht werden, das Eingangssignal so zu verändern, dass der Ausgang des nachgeschalteten Modells für den Schwerhörenden möglichst gleich dem Ausgang des Normalhörenden-Modells für das unmodifizierte Signal ist (vgl. den Beitrag von Volker Hohmann). Ein erster von unserer Arbeitsgruppe entwickelter Ansatz (Dissertationen Hohmann und Launer) findet sich bereits in modifizierter Form in kommerziellen digitalen Hörgeräte-Systemen (z. B. Firma Phonak) wieder.

Brauchen Computer Hörgeräte?

Eines der größten Probleme von automatischen Spracherkennern ist die mangelhafte Erkennungsleistung, wenn zusätzlich zur Sprache Störgeräusche zu hören sind. Herkömmliche Spracherkenner (z. B. in PC-Diktiersystemen) versagen in solchen Situationen oftmals kläglich, selbst bei niedrigen Störgeräuschpegeln, bei denen Menschen überhaupt keine Probleme beim Sprachverstehen haben. Daher liegt es nahe, das menschliche Gehör als quasi "perfekten" Spracherkenner in Teilen zu simulieren. Tatsächlich hat sich in vielen Experimenten gezeigt, dass der Einsatz des Gehörmodells in automatischen Spracherkennern eine deutliche Verbesserung der Erkennungsleistung in Störgeräuschen ermöglicht, im Vergleich zu den (kaum am Gehör orientierten) Standardverfahren. Dazu wird eine Mustererkennung auf der Ebene der internen Repräsentation (am Ausgang des Gehörmodells) durchgeführt, da im Idealfall auf dieser Ebene dieselben Ähnlichkeitsbeziehungen auftreten wie beim menschlichen Hören. Interessanterweise funktioniert diese gehörgerechte Aufarbei-



Interne Repräsentation des gesprochenen Wortes "Zugbegleiter". Das Perzeptionsmodell berechnet aus dem akustischen Signal ein solches Abbild der neuronalen Repräsentation auf der Hörbahn.



Beispiel für den Ablauf einer automatischen Spracherkennung. Besonders gut funktionieren Systeme, bei denen die Komponenten Teile der menschlichen Hörbahn modellieren.

tung des Sprachsignals für die Spracherkennung am besten, wenn als eigentlicher Sprachmuster-Erkenner am Ausgang des Perzeptionsmodells ein neuronales Netz verwendet wird. D. h. auch hier ist ein "Biologie-naher" Ansatz in der Kombination erfolgversprechender als die Standard-Verfahren, die auf rein statistischen Ansätzen beruhen (so genannte "Hidden Markov Modelle"). Neueste Erkenntnisse aus der Neurobiologie und Psychoakustik deuten darauf hin, dass einzelne Neuronen im Kortex (der Großhirnrinde) Informationen aus bestimmten Bereichen der Internen Repräsentationen verknüpfen und der Mensch so z. B. Sprachlaute unterscheiden kann. Dies lässt sich technisch mit so genannten Sigma-Pi Zellen nachbilden, deren Ausgang wiederum sekundäre Merkmale darstellt. Aktuelle Arbeiten zeigen, dass die Leistung von Spracherkennungssystemen durch die Verwendung dieser sekundären Merkmale weiter verbessert werden kann (vgl. Abb oben). Zudem läßt die Analyse der Relevanz einzelner Sigma-Pi Zellen Rückschlüsse über die "Verschaltung" der Neuronen im primären auditorischen Kortex erwarten.

Neben der Möglichkeit, eine robuste automatische Spracherkennung durch eine geeignete gehörbasierte Repräsentation der akustischen Muster zu erreichen, bietet sich auch die Option, zunächst die Störgeräusche im akustischen Eingangssignal zu unterdrücken und anschließend die Spracherkennung auf die derartig "verbesserten" Signale anzuwenden. Dabei kann prinzipiell dieselbe Störgeräuschunterdrückung vorgenommen werden wie bei "intelligenten" Hörgeräten - gewissermaßen hat der Computer dasselbe Problem in akustisch ungünstigen Umgebungen wie ein hörgestörter Mensch! Aber auch hier greift der gehörorientierte Ansatz: In einer dem menschlichen Gehör nachempfundenen Repräsentation des Sprachsignals als sog. Amplituden-Modulations-Spektrogramm (AMS) lässt sich ein Störgeräusch viel leichter von Sprache unterscheiden als mit herkömmlichen Analysemethoden. Dadurch können die Signalanteile, die von einem Störgeräusch stammen, im Eingangssignal unterdrückt werden und die sprachlichen Signalanteile ungehindert durchgelassen werden. Eine akustische Demonstration eines derartigen Störgeräusch-Unterdrückungssystems ist im Internet abrufbar unter http://medi.uni-oldenburg.de/members/juergen/ams.html. Tatsächlich wird mit dieser Vorverarbeitung die Erkennungsleistung eines Spracherkenners in Störgeräusch deutlich verbessert. Erste Versuche mit Schwerhörigen sind ebenfalls vielversprechend - ein Beispiel dafür, dass dieselbe Idee in zwei sehr unterschiedlichen Anwendungen erfolgreich sein kann!

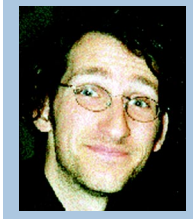
Kleiner und leistungsfähiger: Ohr aus Silizium

Die bisher beschriebenen Anwendungen gehörgerechter Signalverarbeitung mit einem Computermodell des Ohres haben leider einen Haken: Die Berechnung kostet so viel Rechnerleistung, dass eine Echtzeit-Anwendung mit den derzeit modernsten Mikroprozessoren gerade noch möglich ist, die viel Strom "fressen" und große Netzteile und Kühleinrichtungen in entsprechend großen, fest installierten Geräten erfordern. Dabei sind die Anwendungen für die computerunterstützte akustische Mensch-Mensch- und Mensch-Maschine-Kommunikation besonders für kleine und mobile Geräte interessant, z. B. Handys, Hörgeräte, Laptops mit Diktiergerät-Funktion, MP3-Player und zukünftige "personal digital assistents (PDA)", bei denen die Funktionalität sämtlicher vorgenannter Geräte in einem vereinigt wird. Als Lösung bietet sich die Umsetzung der gehörbasierten Audio-Signalverarbeitung in einen spezialisierten, mit minimalem Stromverbrauch und kleiner Spannung arbeitender Silizium-Chip an - d. h. eine ähnliche Lösung, wie sie bei der derzeitigen Handy- und Hörgeräte-Technologie eingesetzt wird, wo nicht etwa ein freiprogrammierbarer Universalprozessor (wie im PC), sondern ein speziell konfigurierter Signalprozessor-Chip mit großem Entwicklungsaufwand entworfen und gefertigt wird.

Glücklicherweise haben wir in der Informatik an der Universität Oldenburg und der Universität Hamburg Partner gefunden, die genau auf dieses Geschäft spezialisiert sind: Die Arbeitsgruppe "Entwicklung integrierter Schaltungen" (Prof. Dr.-Ing. Wolfgang Nebel) besitzt besondere Kompetenz auf dem Gebiet stromsparender "Öko-Chips" (vgl. W. Nebel: "Recyclebares Mikrochip-Design", EINBLICKE Nr. 26, 1997) und die Arbeitsgruppe "Informatikmethoden und Anwendungen" (Prof. Dr.-Ing. Bärbel Mertsching) verfügt über besonderes Know-how bei der Implementation Biologie-naher Rechenverfahren in integrierte Schaltungen. Zusammen wurden wir von der Deutschen Forschungsgemeinschaft (DFG) in zwei Schwerpunktprogrammen gefördert, um eine stromsparende Hardware-implementation des Perzeptionsmodells und der darauf aufbauenden gehörgerechten Signalverarbeitung zu entwickeln. Der derzeitige Prototyp beschränkt sich noch auf einen frei programmierbaren Chip (FPGA = field programmable gate array), der freilich noch nicht sehr stromsparend arbeitet. Wichtig ist jedoch, dass die prinzipiellen Hürden auf dem Weg zum Ohr als "Öko-Chip" genommen wurden, so dass einer Umsetzung der hier vorgestellten Konzepte für eine gehörgerechte, mobile, computergestützte Sprachkommunikation nichts mehr im Weg steht. Es bleibt zu hoffen, dass damit das "Oldenburger Perzeptionsmodell" irgendwann in jedem Handy zu finden ist!

Die Autoren

Dr. Jürgen Tchorz, Physikstudium in Oldenburg 1990-96 (1992 unterbrochen von einem zehnmonatigem Studium der mathematischen Physik am University College Galway/Irland), anschließend wissenschaftlicher Mitarbeiter am Fachbereich Physik, AG Medizinische Physik, Promotion 2000. Seit Ende 2000 bei Phonak Hearing Systems in Stäfa (Schweiz) tätig. Forschungsschwerpunkte: Automatische Spracherkennung und Störgeräuschunterdrückung.



Michael Kleinschmidt, wissenschaftlicher Mitarbeiter und Doktorand am Fachbereich Physik, AG Medizinische Physik. Physikstudium 1992-95 in Göttingen und 1996-99 in Oldenburg. 1995/1996 einjähriger Studienaufenthalt in Wellington/Neuseeland. Assoziiertes Mitglied des europäischen Graduiertenkollegs Neurosensorik. Forschungsschwerpunkt: Automatische Sprach- und Signalklassifikation. (Dr. Volker Hohmann s. S. 26, Prof. Dr. Dr. Birger Kollmeier s. S. 4)