Gehörmodelle für die Sprach- und Audiocodierung: Eine Einführung

Birger Kollmeier
Medizinische Physik, Universität Oldenburg, D-26111 Oldenburg

Einleitung

Trotz großer Fortschritte in der Sprach- und Audiosignalverarbeitung innerhalb der letzten Jahrzehnte verbleiben Herausforderungen, die mit den bisherigen, auf technischen Ansätzen beruhenden Signalverarbeitungs-Strategien nicht befriedigend gelöst werden konnten (z. B.die „optimale“ Signalkodierung mit geringer Datenrate bei höchstmöglicher Wiedergabequalität, robuste Spracherkennung für unterschiedliche Störgeräusche und Sprachsignale, und „intelligente“ Hörgeräte). In letzter Zeit hat es nun erfolgversprechende Ansätze gegeben, derartige Probleme durch Anlehnung an die Signalverarbeitung lösen zu können, die im menschlichen Hörsystem realisiert wird. Dafür benötigt man Hörmodelle, die die Umsetzung des akustischen Signals in die „interne Repräsentation“ im auditorischen System beschreiben. Diese interne Repräsentation kann entweder durch psychoakustische Modellgrößen (z. B. Erregungspegel, Lautheits-Zeitmuster, Teiltonmuster oder temporal masking patterns nach Zwicker und Fastl (1990)) oder durch Modellvorstellungen über die „effektive“ Signalverarbeitung (z.B. Auditory Image Modell nach Pattersen et al. , 1995 oder Perzeptions-Modell nach Dau et al., 1997) beschrieben werden. Charakteristisch für derartige Abbildungen ist, daß durch die nichtlinearen Verarbeitungsstufen und die Berücksichtigung der statistischen Fehlern („internes Rauschen“) immer ein „Informationsverlust“ auftritt. Er führt zu einer nicht-eindeutigen Abbildung des akustischen Eingangssignals auf die zugehörige interne Repräsentation, d.h. unterschiedliche akustische Signale führen zu demselben subjektiven Eindruck. Die quantitativen Modellierung dieses Informationsverlusts kann einerseits psychoakustisch ausmeßbare Phänomene (z.B. Maskierung eines Tons in Anwesenheit eines Rauschens) korrekt beschreiben und andererseits für eine Reihe von Anwendungen nutzbringend eingesetzt werden:

- Signalkodierung: Ein kodiertes Signal wird nach der Dekodierung vom Hörer als identisch mit dem Original-Signal wahrgenommen, 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.
- Signalqualitäts-Bewertung: 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 (z. B. PSQM-Verfahren nach Beerends und Stemerdink, 1994, Verfahren nach Hansen und Kollmeier (2000)) sowie Ansätze zur Beurteilung von Audio-Übertragungsqualität (z.B. Beerends, 1995).
- Sprach- und Mustererkennung: Der auf der Ebene der internen Repräsentation dieselben Ähnlichkeitsbeziehungen auftreten wie beim menschlichen Hören, liegt es nahe, die hohe Leistungsfähigkeit der menschlichen auditiven Mustererkennung (z. B. Sprachwahrnehmung unter Störgeräusch) nachzubilden, indem eine Mustererkennung auf dieser Ebene der internen Repräsentation durchgeführt wird. Derartige Ansätze der gehörbasierten Spracherkennung haben sich bereits als vorteilhaft erwiesen (Hermansky, 1990, Hermansky und Morgan, 1994, Tchorz und Kollmeier, 1999).
- Hörgeräte: 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 individuellen Schwerhörenden möglichst gleich dem Ausgang des Normalhörenden-Modells für das unmodifizierte Signal ist. Ein erster Ansatz dazu wurde von Hohmann (1993) beschrieben und findet sich in modifizierter Form in digitalen Hörgeräte-Systemen z.B. der Firma Phonak wieder.

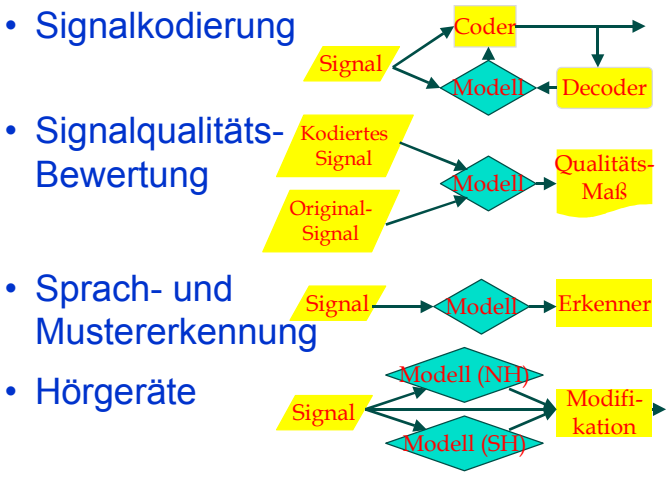

Abb. 1: Prinzip der Anwendung von Gehörmodellen

Nachbildung spektraler Eigenschaften

Die bisher in der gehörbasierten Signalverarbeitung implementierten Verfahren basieren vorwiegend auf einer Nachbildung von spektralen Auflösungs- und Verdeckungseigenschaften des Gehörs (z. B. DCC-Verfahren, Mini-Disk, digitaler Rundfunk, MP3-Kodierung). Sie sind durch die Frequenz-Orts-Transformation im Innenohr motiviert und die damit zusammenhängende spektralen Maskierungs-Eigenschaften und Lautheitsmodelle, die für stationäre Signale ausgiebig untersucht wurden (Zwicker und Fastl, 1990). Auf demselben Prinzip beruht auch die spektrale Sprachperzeptions-Theorie, die insbesondere für die Wahrnehmung von Vokalen die spektrale Detektion von Formanten als wesentlich voraussetzt. Trotz der unstreitbaren Erfolge dieser spektralen Ansätze hat sich jedoch in letzter Zeit die Kenntnis durchgesetzt, daß die zeitliche Verarbeitung im auditorischen System eine mindestens ebenso wichtige Rolle spielt, so daß entsprechende Zeit- Eigenschaften bei der technischen Umsetzung von Gehör- Eigenschaften ebenfalls berücksichtigt werden müssen.

Berücksichtigung von Zeiteffekten

Zu den wichtigen zeitlichen Eigenschaften des Gehörs, die durch psychoakustische Experimente charakterisiert und durch Gehörmodelle adäquat beschrieben werden sollten, gehören:

- Vor- und Nachverdeckung: Die Hörbarkeit eines Testsignals ist ab ca. 10 ms vor und bis zu 200 msec nach der Präsentation eines Maskierungssignals herabgesetzt.
- Zeitliche Integration: Mit zunehmender Länge eines Testsignals (bis zu ca. 200 msec) nimmt seine Detektierbarkeit zu.
- Modulationswahrnehmung: Abhängig von der Bandbreite eines Trägersignals zeigt die Detektierbarkeit einer aufgeprägten Modulation für verschiedene Modulationsfrequenzen eine Tiefpaß, Hochpaß – oder Bandpaß-Charakteristik (Fleischer, 1982, Dau et al., 1997). Dabei spricht eine große Zahl experimenteller Befunde für eine Modulationsfrequenz-Selektivität im auditorischen System, d. h. die Aufspaltung von Einhüllenden-Fluktuationen in verschiedene Modulations-Frequenzen.

Modell der „effektiven“ Signalverarbeitung (Oldenburger PEMO)

Zur adäquaten Nachbildung primär der zeitlichen Eigenschaften des auditorischen Systems in psychakustischen Experimenten hat sich das zunächst in Göttingen und ab 1993 in Oldenburg weiterentwickelte Modell der „effektiven“ Signalverarbeitung im auditorischen System nach Dau et al. (1996, 1997) bewährt. Es basiert auf einer geringen Zahl von Annahmen und Parametern, die durch physiologische Gegebenheiten und psychoakustische Erkenntnisse vorgegeben sind und ermöglicht die quantitative Vorhersage einer vergleichsweise großen Zahl unterschiedlicher psychoakustischer Experimente. Insbesondere lassen sich die o. a. Zeiteffekte, Modulationswahrnehmungsaspekte und –mit den Einschränkungen, die eine lineare Gammatone-Filterbank als erste Verarbeitungsstufe bedingt – auch spektrale Effekte befriedigend quantitativ vorhersagen. Einen Überblick über die Anwendung des Modells in der Psychoakustik findet sich bei Dau (1996). Bei der Anwendung auf die Audiosignal- und Sprachverarbeitung konnten mit diesem Modell die folgenden Ergebnisse erzielt werden:

- Vorhersage von Sprachqualität: In der Dissertation von M Hansen (1999) konnte das Modell erfolgreich zur Vorhersage der Gesamt-Übertragungsqualität von Sprachkodierungsverfahren eingesetzt werden (Hansen, 1998, Hansen und Kollmeier, 2000b). Ebenso konnte eine Vorhersage von zeitlich variierenden Qualitätsansprüchen erreicht werden (Hansen und Kollmeier, 1999) und von dementspechenden Urteilen bei frequenzabhängiger Qualitätsbeeinflussung (Hansen und Kollmeier, 2000a).
- Automatische Spracherkennung: Tchorz und Kollmeier (1999) konnten die Überlegenheit eines Spracherkenners mit einer Vorverarbeitung durch das Perzeptionsmodell gegenüber einer Standard-Vorverarbeitung mit Mel-Frequenz-basierten Cepstralkoeffizienten (MFCC) zeigen. Dabei ging die Spracherkennungsrate erst bei ungünstigeren Signal-Rauschabständen zurück. Diese zusätzliche Robustheit motivierten die Umsetzung dieses Modells in eine Hardware-Realisation (Beitrag Nebel et al. , dieser Band). Besonders wichtig für die erreichte Robustheit erwies sich die „effektive“ Übertragungscharakteristik für Amplituden-Modulationen, die ungefähr dem Modulationsspektrum von natürlicher Sprache angepaßt ist.
- Anwendung des Modulationsspektrogramms: Die Aufspaltung von Einhüllenden-Fluktuationen nach verschiedenen Modulationsfrequenzen, die sowohl psychoakustisch als auch physiologisch (Arbeiten von Langner, 1992) motiviert ist, bietet den Vorteil, daß akustische Quellen selbst dann vom auditorischen System getrennt werden können, wenn sie Energie im selben Spektralbereich aufweisen (aber ein unterschiedliches Modulationsspektrum besitzen). Die daraus folgenden psychoakustischen Effekte „Comodulation Masking Release“ (Detektionsvorteil bei kohärent modulierten, weit spektral auseinanderliegenden Frequenzbändern) und „Modulation Detection Interference“ (Verschlechterung der Modulations-Detektion, wenn

ein spektral weit entfernter Modulator ein ähnliches Modulationsspektrum aufweist) wurden experimentell und theoretisch von der Dissertation von Jesko Verhey (1999) untersucht.

Eine erste technische Umsetzung des Prinzips des Modulationsspektrogramms (d. h. Aufspaltung jedes Frequenzbandes in verschiedene Modulations-Frequenzbänder und Analyse bzw. Modifikation dieser zweidimensionalen Darstellung von Mittenfrequenz vs. Modulationsfrequenz) wurde von Kollmeier und Koch (1994) für die Störgeräuschunterdrückung beschrieben. Eine weitere Anwendung dieses zweidimensionalen Modulationsspektrogramms zur Schätzung des Signal-Rauschabstandes in einem rauschbehafteten Sprachsignal wurden von Tchorz und Kollmeier (dieser Band) beschrieben. Mit Hilfe eines neuronalen Netzes als Auswerteeinheit wird aus dem Modulationsspektrogramm der Signal-Rauschabstand geschätzt. Dabei erweist es sich als notwendig, die vollständige zweidimensionale Verbundverteilung von Mittenfrequenz und Modulaitonsfrequenz als Eingang anzubieten. Wird dagegen nur das Modulationsspektrum (gemittelt über alle Frequenzbänder) oder das Leistungsspektrum (gemittelt über alle Modulationsfrequenzbänder) oder die Kombination dieser beiden (als eindimensionale Vektoren darstellbare) Verteilungen als Eingangsgröße für die Schätzung verwendet, nimmt der Schätzfehler des Signal-Rauschabstandes deutlich zu. Dies weist darauf hin, daß charakteristische zweidimensionale Muster im Modulationsspektrogramm vorteilhaft für die Auswertung und Verarbeitung von Sprachsignalen sind, so daß die im auditorischen System realisierte Kombination von spektraler Analyse und Modulations-Frequenzanalyse motiviert erscheint.

Ausblick

Obwohl der derzeitige Stand der Anwendung von gehörgerechter Signalverarbeitung für die Sprach- und Audiosignalverarbeitung dominiert ist durch spektrale Modelle, erscheint die adäquate Modellierung von Zeiteffekten (einschließlich der Modulationsverarbeitung) notwendig und erfolgversprechend. Ein Beispiel dafür sind die hier angeführten Arbeiten zum Oldenburger Perzeptionmodell, das eine Nachbildung von Zeiteffekten und der zeitlichen Integration ebenso beinhaltet wie die Auswertung des Modulationsspektrogramms, d. h. der zweidimensionalen Aufspaltung in spektralen Gehalt und Modulationsspektrum. Teile dieser Prinzipien sind ebenfalls in anderen Hörmodellen (z.B. Auditory Image Modell nach Pattersen et al., 1995) oder in Sprach(vor-) verarbeitungs-Algorithmen (z.B. RASTA-Algorithmus nach Hermansky und Morgan, 1994) realisiert. Ein weiterer Ausbau und eine Validierung der genannten Hörmodelle erscheint jedoch ebenso notwendig wie ihre weitere Erprobung und konsequenter Einsatz für verschiedene Bereiche der Sprachtechnologie.

Literatur

Beerends, J. G. and J. A. Stemerdink (1994). "A Perceptual Speech Quality Measure based on a Psychoacoustic Sound Perception." J. Audio Eng. Soc. **42**(3): 115--123.

Beerends, J. G. (1995). Measuring the Quality of Speech and Music Codecs, an Integrated Psychoacoustic Approach. 98th AES Convention Paris (Ed.). New York.

Dau, T. (1996). Modeling auditory processing of amplitude modulations. Oldenburg, BIS-Verlag, Universität Oldenburg.

Dau, T., D. Püschel, et al. (1996). "A quantitative model of the `effective' signal processing in the auditory system: I. Model structure." J. Acoust. Soc. Am. **99**: 3615--3622.

Dau, T., B. Kollmeier and A. Kohlrausch (1997). "Modeling auditory processing of amplitude modulation: I. Detection and masking with narrow-band carriers." J. Acoustical Soc. Am. **102**(5): 2892-2905.

Fleischer, H. (1982). "Modulationsschwellen von Schmalbandrauschen." Acustica **51**: 154-161.

Hansen, M. (1998). Assessment and prediction of speech transmission quality with an auditory processing model. Fachbereich Physik, Universität Oldenburg.

Hansen, M. and B. Kollmeier (1999). "Continuous Assessment of time-varying speech quality." J.Acoust. Soc. Am. **106**: 2888-2899.

Hansen, M. and B. Kollmeier (2000a). "Perception of band-specific speech quality distortions: detection and pairwise comparison." Acustica united with acta acustica: (in press).

Hansen, M. and B. Kollmeier (2000b). "Objective modeling of speech quality with a psychoacoustically validated auditory model." J. Audio Eng. Soc.: (in press).

Hermansky, H. (1990). "Perceptual Linear Predictive (PLP) analysis of speech." J. Acoust. Soc. Am. **87(4)**: 1738--1752.

Hermansky, H. and N. Morgan (1994). RASTA Processing of Speech. IEEE Trans. on Speech and Audio Processing **:** 578--589.

Hohmann, V. (1993). Dynamikkompression für Hörgeräte -- Psychoakustische Grundlagen und Algorithmen. Düsseldorf, VDI-Verlag.

Kollmeier, B. and R. Koch (1994). "Speech Enhancement Based on Physiological and Psychoacoustical Models of Modulation Perception and Binaural Interaction." J. Acoust. Soc. Am. **95**: 1593-1602.

Langner, G. (1992). "Periodicity coding in the auditory system." Hear.Res. **60**: 115--142.

Patterson, R. D., A. M. H. and C. Giguère (1995). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform." J. Acoust. Soc. Am. **98(4)**: 1890--1894.

Tchorz, J. and B. Kollmeier (1999). "A model of auditory perception as front end for automatic speech recognition." Journal of the Acoustical Society of America **106**(4): 2040-2050.

Verhey, J. (1999). Psychoacoustics of spectro-temporal effects in masking and loudness perception. Fachbereich Physik, BIS-Verlag, Universität Oldenburg.

Zwicker, E. and H. Fastl (1990). Psychoacoustics - Facts and Models. , Springer.