Ein Verfahren zur Erkennung und Trennung akustischer Objekte

UWE BAUMANN

Fachgebiet Akustische Kommunikation, Technische Universität München
(jetzt: Fachbereich Audiologie, HNO-Poliklinik der Ludwig-Maximilians-Universität München)

Das im folgendem dargestellte Verfahren hat die Aufgabe, durch Anwendung von psychoakustischen und gestaltpsychologischen Mechanismen die Fähigkeit des menschlichen Gehörs zur Stimmentrennung nachzubilden. Den Kernpunkt bildet dabei das Auffinden von *akustischen Objekten*, also die Gruppierung und Zuordnung von Spektralkomponenten zu einer der menschlichen Wahrnehmung entsprechenden Einheit. Dabei werden die Ergebnisse aus den in [1] dargestellten Hörversuchen verwendet. Die Verkettung der Abfolge zusammengehörig empfundener akustischer Objekte wird im Englischen „auditory stream" genannt [2]. Aus diesem Begriff abgeleitet soll diese Abfolge als *akustischer Strom* bezeichnet werden.

Verfahren

Ausgehend von einem Modell der Informationsaufnahme biologischer Systeme nach Terhardt [5] wurde ein hierarchisches Verfahren entwickelt. Fig. 1 gibt einen Überblick über das Verfahren. Eine Reihe sequentiell angeordneter Verarbeitungsstufen bildet von Stufe zu Stufe immer höher abstrahierte Informationen aus dem jeweiligen Eingangssignal heraus. Die Anordnung der Verarbeitungsschritte ist dabei streng hierarchisch und hält sich an das „bottom up" Prinzip: Aus dem Eingangssignal wird durch nacheinander stattfindende Anwendung von relativ einfachen Bearbeitungsvorschriften die dem Signal inhärente Information immer feiner und besser erfaßt und an die nächstfolgende Stufe weitergegeben. Die einzelnen Verarbeitungsstufen sind miteinander rückwirkungsfrei verknüpft, d.h., daß ein in der Verarbeitungshierarchie übergeordneter Prozeß keine untergeordneten Prozesse in ihrem Ablauf steuert.

Nach der analog/digital Wandlung des Zeitsignals findet eine Spektralanalyse statt, welche die Funktion der Frequenz-Ortstransformation des Innenohres nachbilden soll. An das Verfahren der Spektralanalyse werden höchste Anforderungen gestellt, denn die Nachbildung der Stimmentrennungsfähigkeit erfordert sowohl eine hohe Frequenz- als auch eine hohe Zeitauflösung. Eingeschlossen in den Block der Spektralanalyse sind Funktionen zum Gleichrichten und Quadrieren der Kurzzeitspektren sowie eine separat wählbare zeitliche Glättung durch Tiefpaßfilterung. Die Ausgabe der durch den Funktionsblock Spektralanalyse generierten Kurzzeitspektren muß nicht zwangsweise im Zeitraster der durch das Abtastintervall T_s angegebenen Zeitspanne erfolgen. Heinbach [4] wies nach, daß für eine gehörgerechte Signaldarstellung eine Ausgabe der Kurzzeitspektren mit einem Auswerteintervall T_A zwischen 1,25 ms und 7,5 ms ausreicht.

An die Spektralanalyse schließt sich eine in Fig. 1 mit der Bezeichnung Konturierung versehene Verarbeitungsstufe an. Für jedes Betragsquadrat-Kurzzeitspektrum erfolgt eine Suche nach lokalen Maxima („peak picking"). Jedes detektierte Maximum wird durch ein Frequenz/Pegelpaar repräsentiert und als *Teilton* (TT) bezeichnet. Die Gültigkeitsdauer eines jeden Teiltons ist durch ein Auswerteintervall mit der Dauer T_A definiert. Der Satz aller zu einem bestimmten Zeitpunkt gehörenden Teiltöne wird *Teiltonmuster* (TTM) genannt. Die fortlaufende Abfolge von TTM wird als *Teiltonzeitmuster* (TTZM) bezeichnet [4].

Im Block Konturierung sind zusätzlich Verfahren zur Berücksichtigung psychoakustischer Bewertungsfunktionen integriert; neben einem Verfahren zur Pegelüberschuß-Berechnung zur Berück-

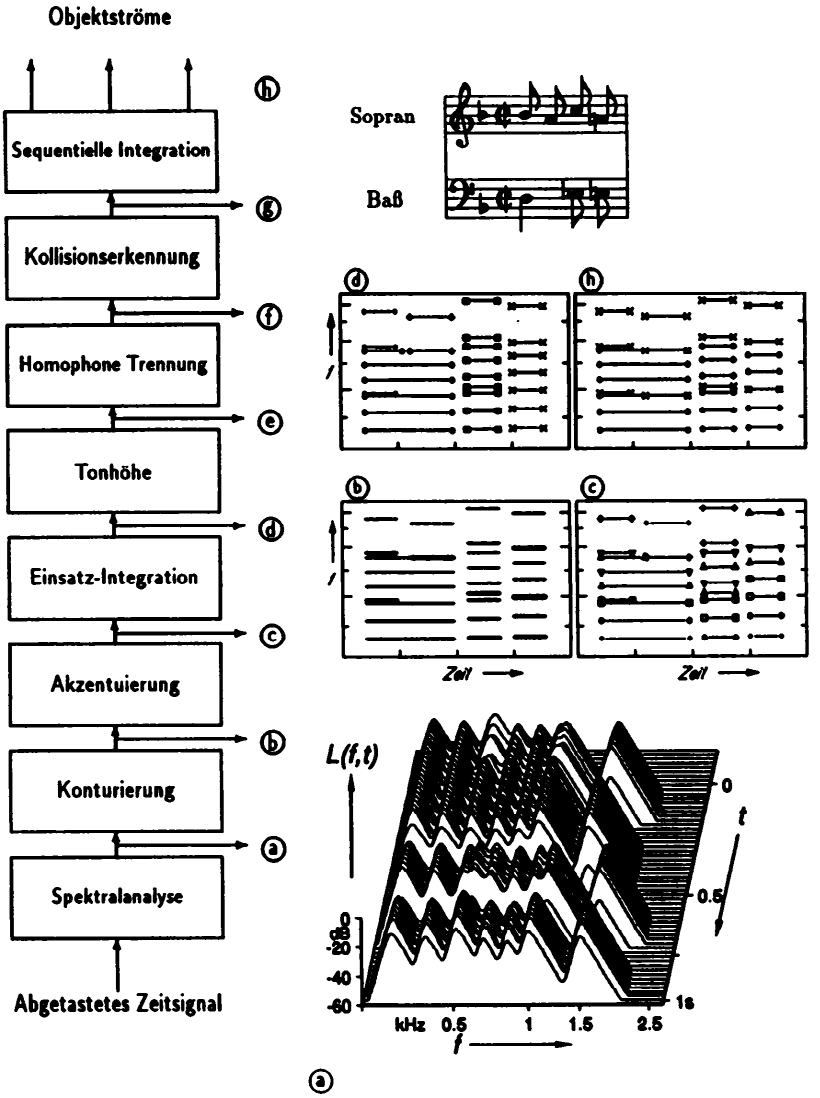

Fig. 1: *Überblick über das Verfahren zur Trennung und Erkennung akustischer Objekte.* ***Links:*** *Blockbild.* ***Rechts:*** *Zwischenergebnisse der Verarbeitungsstufen für oben notiertes Beispiel, dargeboten mit harmonisch-komplexen Tönen (Bass 6 Harmonische, Sopran 3 Harmonische).*

sichtigung spektraler Maskierungseffekte wird eine Bestimmung des Spektraltonhöhengewichtes nach [6] vorgenommen, um erste Hinweise über die Heraushörbarkeit eines TT zu gewinnen. TT mit ausgesprochen hohem Spektraltonhöhengewicht sind eher aus einem Klanggemisch heraushörbar als TT mit einem niedrigem Gewicht. Ebenfalls im Block Konturierung befindet sich ein Verarbeitungsschritt, dessen Aufgabe die Verkettung einzelner zeitlich aufeinanderfolgender Teiltöne unter Berücksichtigung der Gestaltgesetze der Nähe und Ähnlichkeit ist. Das Ergebnis dieser Verkettung sind *Teiltonlinien* (TTL), welche die überwiegend tonalen Anteile des Signales repräsentieren. Die Zusammenfassung aller TTL führt zum *Teiltonlinien-Zeitmuster* (TTLZM).

Psychoakustische Experimente [3] zeigen, daß die Ausgeprägtheit der Tonhöhe von Sinustönen abhängig von deren Dauer ist: kurze Töne zeigen eine geringe, längere Töne eine größere Ausgeprägtheit der Tonhöhe. Der mit der Bezeichnung Akzentuierung versehene Block führt im wesentlichen eine zeitabhängige Bewertung des Spektraltonhöhengewichtmusters durch. Dieser Prozeß stellt eine Neuerung gegenüber dem bisher nur für stationäre Signale geeigneten Verfahren dar.

Die Weiterverarbeitung des akzentuierten TTLZM erfolgt in dem mit Einsatz-Integration bezeichneten Prozeß. Seine Aufgabe ist die Nachbildung der „simultanen Integration", die bei der Wahrnehmung von fast gleichzeitig auftretenden Spektralkomponenten auftritt. Dabei werden Teiltonlinien mit einem nahezu gleichen Start- und Endzeitpunkt zusammengefaßt. Die Grenzen für diese Operation ergeben sich aus den Ergebnissen der in [1] diskutierten Hörversuche zur Erkennung asynchroner Teiltöne. Das Ergebnis dieser Verarbeitungsstufe soll als *auditives Objektmuster* (AOM) bezeichnet werden.

In dem sich daran anschließenden Verarbeitungsschritt Tonhöhe werden für jedes Objekt des AOM eine oder — je nach Ausgeprägtheit — mehrere Tonhöhenverläufe mittels eines Tonhöhenberechnungsverfahrens [6] bestimmt. Das Resultat dieser Bearbeitung ist das *tonhöhenbewertete auditive Objekt Muster* (TAOM).

Der Funktionsblock Homophone Trennung hat die Aufgabe, für eine Auftrennung von denjenigen akustischen Objekten zu sorgen, die bei der vorangegangenen Tonhöhenbestimmung mit mehrdeutigen Tonhöhenverläufen versehen worden sind. Dieser Verarbeitungsschritt wird in der Regel dann nötig, wenn mehrere Stimmen gleichzeitig auftreten und damit in der musikalischen Terminologie *homophon*[1] miteinander klingen.

Immer wenn mehrere komplexe Klänge gleichzeitig ertönen, kann es, wie beispielsweise im Beispiel aus Fig. 1 ersichtlich, zu einer Überlagerung von Teiltönen kommen. Im Block Kollisionserkennung wird für die Dauer aller nun mit eindeutigen Tonhöhenkonturen versehenen akustischen Objekten nach harmonischen TT gesucht, die bereits mit einem anderen Objekt verbunden sind. Ergeben sich derartige TT, so erfolgt eine gesonderte Behandlung dieser Kollisionen.

Der abschließende Funktionsblock Sequentielle Integration hat die Aufgabe, die durch die bisherigen Funktionsschritte gewonnenen einzelnen akustischen Objekte des TAOM in ihrer zeitlichen Abfolge in einer Weise zu verketten, die möglichst der Wahrnehmung der *akustischen Ströme* entspricht, die ein Zuhörer bei der Darbietung des mehrstimmigen Eingangs-Signales heraushören könnte.

Erprobung und Anwendung des Verfahrens

Als Testsignal wurde das in Fig. 2 im Notentext abgebildete zweistimmige Musikstück verwendet, welches auf einem Synthesizer, den ein Computer über eine MIDI Schnittstelle ansteuer-

[1] homophon: gleichstimmig, melodiebetont.

te, abgespielt wurde. Das Stück wurde von zwei Stimmen dargeboten, welche jeweils auf sechs Harmonische beschränkt wurden. Das Signal wurde mit einer Abtastrate von $f_s = 8\,\text{kHz}$ aufgezeichnet und mit einer Auflösung von 16 Bit abgespeichert. Am Ende des in Fig. 1 dargestell-

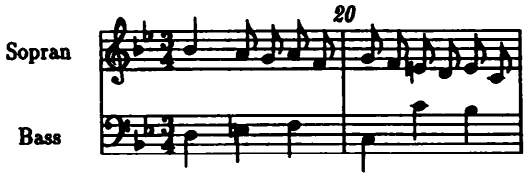

Fig. 2: *Takt 19–20 aus J. S. Bachs Minuet in B-moll (Kleines Notenbüchlein f. Anna Magdalena Bach).*

ten Verfahrens steht die Auftrennung der auditiven Objekte in zu separaten Stimmen gehörende akustische Ströme. Fig. 3 zeigt die getrennten Einzelstimmen als TTZM. Bis auf die sech-

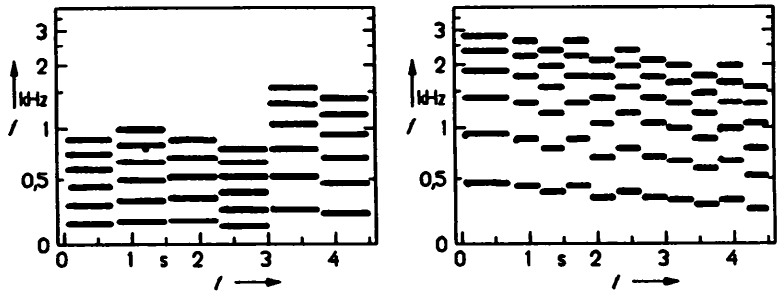

Fig. 3: *Ergebnis des Stimmentrennungs-Verfahrens für das in Fig. 2 dargestellte Musikstück. Links: Baß-Stimme. Rechts: Sopran.*

ste Harmonische des dritten Tones der Baß-Stimme konnte eine korrekte Zuordnung erfolgen. Die Resynthese der Einzelstimmen entspricht der durch die mehrstimmige Darbietung entstandenen Vorstellung vom Klang der Einzelstimmen. Es wäre denkbar, die Tonhöhenverläufe der getrennten Stimmen durch einen weiteren Verarbeitungsprozeß automatisch zu transkribieren oder in das MIDI-Format zur Ansteuerung eines Synthesizers zu übertragen.

[1] **Baumann, U. (1994).** *Über die Wahrnehmung von Teiltönen in Melodien aus harmonisch komplexen Klängen.* In: Fortschritte der Akustik – DAGA '94, Seiten 1017–1021, Bad Honnef. DPG-GmbH.

[2] **Bregman, A. S. (1990).** *Auditory Scene Analysis.* MIT, Cambridge, Massachusetts.

[3] **Fastl, H. (1989).** *Pitch strength of pure tones.* In: 13th Intern. Conf. on Acoustics, Belgrade, Yugoslavia 1989, Seiten 11–14, 1989.

[4] **Heinbach, W. (1988).** *Aurally adequate signal representation: The part-tone-time-pattern.* Acustica, 67: 113–121.

[5] **Terhardt, E. (1992).** *From speech to language: On auditory information processing.* In: Schouten, M. E. H., Editor, The Auditory Processing of Speech: from Sounds to Words, Seiten 363–380. Mouton de Gruyter, Berlin.

[6] **Terhardt, E., Stoll, G., Seewann, M. (1982).** *Algorithm for extraction of pitch and pitch salience from complex tonal signals.* J. Acoust. Soc. Am., 71(3): 679–688.