Stimmentrennung in polyphoner Musik

U. Baumann, Lehrstuhl für Elektroakustik, TU München

Einleitung

Die Fähigkeit des menschlichen Gehörs, unterschiedliche musikalische Stimmen als separate Melodielinien wahrnehmen zu können, führte zu einer Reihe von Theorien. Es wurden dabei häufig Analogien aus dem Bereich der Psychologie verwendet, die sich mit der Erklärung der Figur/Hintergrund Zuordnung des visuellen Systems beschäftigen. Insbesondere werden Prinzipien der Gestaltpsychologie wie *Proximity, Similarity, Good Continuation* oder *Common Fate* herangezogen, um Wahrnehmungsstrategien zur Gruppierung akustischer Ergebnisse zu untersuchen [2].

Im folgenden soll gezeigt werden, daß durch die Verarbeitung von Tonhöhe und Einsatz von Teiltönen eine Ausgangsbasis zur Trennung von Einzelstimmen aus polyphoner Musik geschaffen werden kann. Es wurde ein hierarchisches Modell entwickelt und nach seiner Implementation auf einem Rechner mit verschiedenen Beispielen polyphoner Musik getestet. Eine akustische Überprüfung des Verfahrens erfolgte durch Resynthese der herausgetrennten Einzelstimmen.

Modell

Das in Fig. 1 dargestellte Verfahren setzt sich aus sieben Stufen zusammen. Die Einzeldiagramme aus Fig. 3 sollen die Ergebnisse der einzelnen Verarbeitungsstufen für ein kurzes Melodiebeispiel (Fig. 2) verdeutlichen.

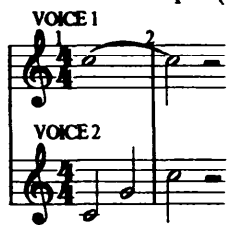


Fig. 2: Notation der kurzen Beispielsequenz.

Die Beispielmelodie ist aus zwei einfachen Stimmen zusammengesetzt: ein aus drei Harmonischen (500, 1000, und 1500 Hz, 70 dB) zusammengesetzter, lang ausgehaltener komplexer Ton und eine aus drei komplexen Tönen (Grundfrequenzen bei 250, 375 und 500 Hz) zusammengesetzte aufsteigende Melodielinie. Jeder Ton der aufsteigenden Tonsequenz besteht aus sechs Harmonischen mit einem Pegel von je 60 dB und einer Dauer von je etwa 300 ms.

Stufe 1: FTT — Spektraltransformation

Die Vorverarbeitung des diskreten Zeitsignals wird durch eine spezielle zeitvariante Kurzzeit–Fourier–Transformation nach Terhardt [6] vorgenommen. Die FTT (Fourier–Time–Transformation) zählt zur Gruppe der gehörangepaßten Spektraltransformationen, da die Verteilung von Analysestützfrequenzen und Bandbreiten an psychoakustische Ergebnisse angepaßt wurde. Fig. 3.1 zeigt das Ergebnis der FTT für die in Fig. 2 angegebene Beispielmelodie.

Stufe 2: Konturisierung

Heinbach [3] zeigte, daß die von ihm entwickelte spektrale Konturisierung durch das *Teiltonzeitmuster* eine gehörgerechte Repräsentation von Audiosignalen ermöglicht. Vom Autor [1] wurde das Verfahren während des Restaurationsvorgang einer historischen Kirchenorgel zu objektiven Beurteilungen der Klangqualität benutzt.

Der Konturisierungsprozeß wird in wählbaren Zeitabständen an den von der FTT erzeugten Kurzzeitspektren durchgeführt. Der in [3] beschriebene Maximumsdetektions–Algorithmus wird dabei zur Erkennung von spektralen Peaks benutzt und liefert Frequenz/Pegelpaare – die zugehörige Phaseninformationen wird nicht berücksichtigt.

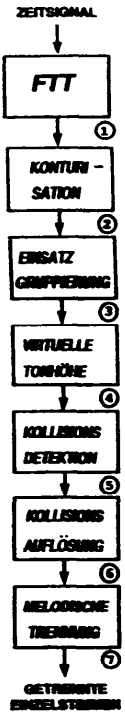
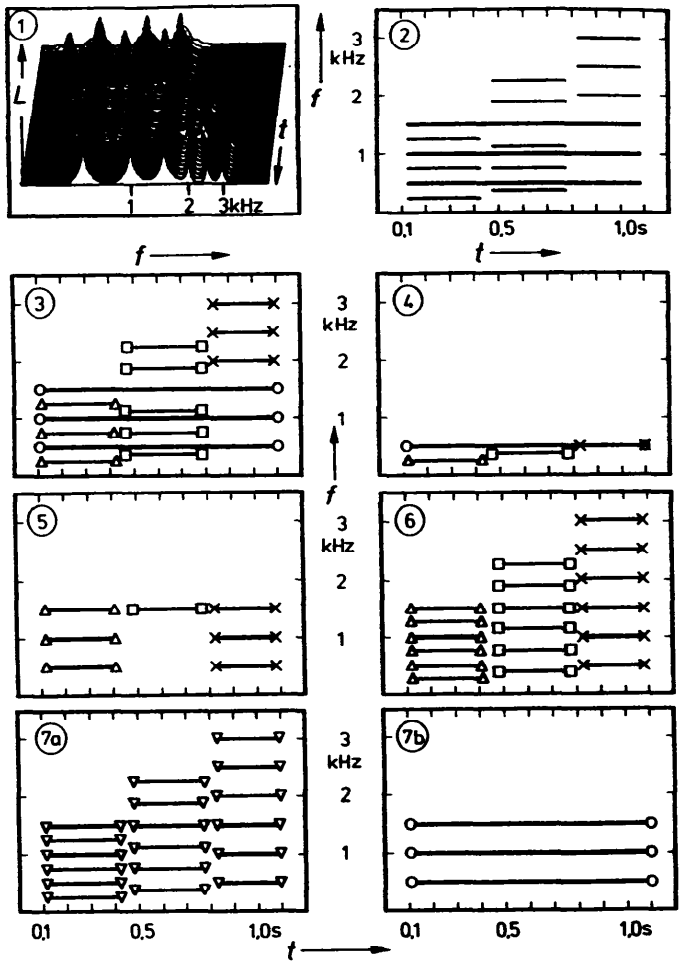


Fig. 1 (oben): Schema zur Trennung musikal. Stimmen.

Fig.3.1-3.7 (rechts): Ergebnisse der Zwischenstufen für die Beispielsequenz.



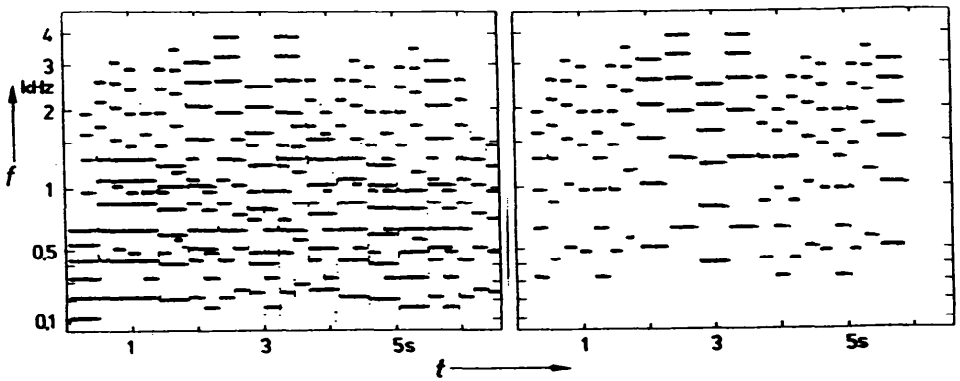


Fig. 5a (links): Ausgabe der 2.Stufe des Verfahrens für das Musikstück aus Fig. 4.
Fig. 5b (rechts): Ausgabe der automatisch extrahierten Oberstimme nach der siebten Stufe.

Fig. 3.2 zeigt stilisiert das Ergebnis der zweiten Stufe – ein sogenanntes Teiltonzeitmuster (TTZM). Der Teiltonpegel ist bei dieser Darstellungsart als Linienstärke kodiert.

Stufe 3: Einsatz-Gruppierung

Eine wesentliche Hilfe zur Gruppierung von akustischen Ereignissen bildet die Erkennung *gleichzeitiger* Veränderungen von Frequenz oder Pegel simultaner Teiltöne. In der dritten Stufe des Modells wird daher ein Gruppierungsmechanismus eingesetzt, der auf den Gestaltprinzipien *Common Fate* und *Similarity* beruht. Die Erkennung gleichzeitiger Veränderungen (im folgenden als *Einsatz* bezeichnet) wird dabei in verschiedene Teilaufgaben untergliedert. Zunächst wird ein von Mummert in [4] beschriebener Tonlinien–Verfolgungsalgorithmus verwendet, um Teiltonlinien zu erkennen, die den tonalen Anteil des Signals repräsentieren. Danach werden Pegel– und Frequenzmittelwerte dieser Tonlinien berechnet und zusammen mit der Startzeit und der Dauer der jeweiligen Linie abgespeichert. Nun erfolgt eine Zusammenfassung gleichzeitig einsetzender Tonlinien zu einer Einsatzgruppe. Im weiteren erfolgt die Aufteilung dieser Einsatzgruppe in Untergruppen mit ähnlicher durchschnittlicher Liniendauer. Fig. 3.3 zeigt das Ergebnis dieser Gruppierungsprozedur für die Beispielsequenz. Es werden vier akustische Objekte erkannt, die durch unterschiedliche Symbole ($\bigcirc, \square, \triangle, \times$) gekennzeichnet sind.

Stufe 4: Virtuelle Tonhöhenbestimmung

Ein wesentlicher Schlüssel zur Erkennung einer musikalischen Einzelstimme aus einem komplexen, zeitveränderlichen Frequenzgemisch ist die Detektion einer Fundamentalfrequenz. Die Abschätzung dieser speziellen Frequenz ist für die Rekonstruktion und die Erkennung der melodischen Linie einer musikalischen Stimme wichtig.

Eine Schwierigkeit bei der Bewältigung dieser Aufgabe bildet die Erkennung der zeitlichen Überlagerung der Spektralkomponenten unterschiedlicher Stimmen. Die in der vorangegangenen Stufe erfolgte Vorgruppierung nach gemeinsamer Linienlängendauer ermöglicht jedoch, ausschließlich Mitglieder der gleichen Einsatzgruppe zur Grundfrequenzabschätzung heranzuziehen. Zur Tonhöhenbestimmung aller so gewonnenen Einsatzgruppen wird das Verfahren zur Bestimmung der virtuellen Tonhöhe nach Terhardt [5] benutzt. Fig. 3.3 zeigt ein Beispiel zur Erkennung einer virtuellen Tonhöhe: die durch Kreuze markierte Einsatzgruppe mit Harmonischen bei 2, 2.5 und 3 kHz generiert durch das Tonhöhenberechnungsverfahren einen virtuellen Tonhöhenkandidaten bei 500 Hz, welcher in Fig. 3.4 auf der mit Kreisen beginnenden Linie bei 500 Hz durch Kreuze dargestellt ist.

Stufe 5 + 6 : Kollisionserkennung und -auflösung

Immer wenn mehrere komplexe Klänge gleichzeitig ertönen, kann es zu einer Überlagerung von Teiltönen kommen. Um eine möglichst genaue Rekonstruktion des Obertonaufbaues einer musikalischen Stimmen zu erhalten, ist die Detektion dieser Kollisionen unumgänglich, da schwache Teiltöne gegebenenfalls maskiert wurden und deren Pegel somit nicht mehr ermittelbar ist. Beispielsweise sind die tiefsten drei Teiltöne des durch Kreuze markierten Objektes (L_i = 60 dB) in Fig. 3.3 durch stärkere Teiltöne des durch Kreise gekennzeichneten Objektes (L_i = 70dB) verdeckt.

Die vorangegangene Abschätzung der virtuellen Tonhöhe eines jeden Objektes ermöglicht die Erkennung überlagerter Teiltöne (Kollisionen) zwischen verschiedenen Einsatzgruppen. Fig. 3.5 veranschaulicht die in der Beispielsequenz als Kollision erkannten Teiltonlinien. Nach dem Ermitteln der Kollisionen erfolgt eine Zuordnung dieser Teiltonlinien zu denjenigen Objekten, die in der Umgebung der jeweiligen durchschnittlichen Teiltonlinienfrequenz noch keine harmonischen Komponenten besitzen. Fig. 3.6 zeigt die Ergebnisse der Kollisionszuordnung für die Beispielsequenz.

Stufe 7: Melodische Trennung

Die letzte Aufgabe des Modells bildet die Erkennung der Melodielinien, die in einem polyphonen Stimmengemisch enthalten sind. In den vorangegangenen Verarbeitungsstufen wurden Einsätze detektiert und in akustische Objekte zusammengefaßt, deren Obertonaufbau und Tonhöhe bestimmt wurde. Die letzte Verarbeitungsstufe dieses Verfahrens versucht nun diese akustischen Objekte so miteinander zu verbinden, daß der sich hieraus ergebende Melodieeindruck der Wahrnehmung eines Zuhörers entspricht. Eine sehr einfache Strategie, die Kriterien bezüglich der Klangfarbe einer Stimme völlig außer acht läßt, ist die Suche nach dem in der Zeit und virtueller Tonhöhe am nächsten liegenden Nachbarobjekt. Diese Methode läßt sich natürlich nur auf einfache Tonfolgen anwenden; Stimmenkreuzungen beispielsweise führen aufgrund ihrer Mehrdeutigkeit zu Problemen. Fig. 3.7 zeigt das Ergebnis der melodischen Trennung der Beispielsequenz: die aufsteigende Stimme (▽) und einen aus drei Harmonischen bestehenden Ton, der ausgehalten wird (○).

Polyphones Beispiel

Das in dieser Arbeit vorgestellte Verfahren wurde unter anderem mit dem in Fig. 4 notierten zweistimmigen, polyphonen Musikstück getestet. Das Stück wurde zur besseren Reproduzierbarkeit von einem Computer auf einem Synthesizer abgespielt. Der temperiert gestimmte Synthesizer generierte für jeden Ton sechs Harmonische gleichen Pegels mit einfacher Rechteck-Amplitudenhüllkurve. Das so gewonnene Zeitsignal wurde mit 16 Bit quantisiert, mit einer Rate von 12.8 kHz abgetastet und danach mit dem vorgestellten Modell untersucht. Fig. 5.a zeigt das Ergebnis der zweiten Stufe für das Beipiel aus Fig. 4, Fig. 5.b die durch das Verfahren automatisch herausgetrennte Oberstimme.

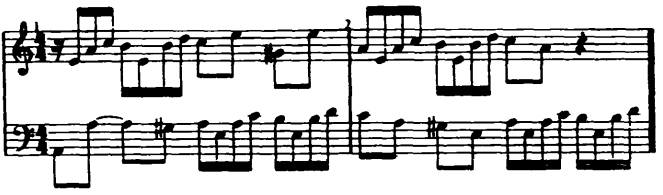


Fig. 4: Notentext des Beginns der Invention Nr. 13 (J.S. Bach). Fig. 5.b zeigt das Ergebnis der automatischen Stimmentrennung für die Oberstimme.

Zusammenfassung

Die Implementierung des vorliegenden Modells zur musikalischen Stimmentrennung hat gezeigt, daß sich die Fähigkeit des menschlichen Zuhörers zur Separation bereits auf einer relativ niedrigen Ebene der auditiven Informationsverarbeitung simulieren läßt. Weitere Untersuchungen sollen die Leistungsfähigkeit des Verfahrens bezüglich der Trennung von Stimmen realer Musikinstrumente erweitern.

Literatur

[1] Baumann, U., Vergleich akustischer Daten einer Barockorgel vor und nach Restaurierung. In *Fortschritte der Akustik* (DAGA 91). Bad Honnef – Bochum, 1991, S. 873–877.

[2] Deutsch, D., Grouping Mechanisms in Music. In *The Psychologie of Music.* New York: Academic Press, 1982, 99–134.

[3] Heinbach, W., Aurally adequate signal representation: The Part-Tone-Time-Pattern. *Acustica* 1988, 67, 113–121.

[4] Mummert, M., Trennung von tonalen und geräuschhaften Signalen im Sprachsignal. In *Fortschritte der Akustik* (DAGA 90). Bad Honnef – Wien, 1990, S. 1047–1050.

[5] Terhardt, E., Calculating virtual pitch. *Hearing Research* 1979, 1,155–182.

[6] Terhardt, E., Fourier Transformation of Time Signals: Conceptual Revision. *Acustica* 1985, 57 , 242–256.