Musikpsychologie Bd. 25, 55–71, © Hogrefe Verlag, Göttingen 2015

Verfahren zur Schätzung der Grundfrequenzverläufe von Melodiestimmen in mehrstimmigen Musikaufnahmen

Jonathan Driedger & Meinard Müller[1]

Zusammenfassung

Betrachtet man den einem Musikstück zugrunde liegenden Notentext, so lassen sich aus diesem Informationen wie der Verlauf einer Melodiestimme unmittelbar ablesen. Die Tonhöhen, Dauern und Einsatzzeiten der Melodietöne werden hierbei durch geeignete Notensymbole explizit kodiert. In einer Musikaufnahme sind die Verhältnisse wesentlich komplexer. Betrachtet man die Aufnahme eines auf einem Instrument gespielten oder gesungenen Tons, so kann man neben der Grundfrequenz auch eine Reihe von Obertönen und anderen Frequenzkomponenten messen. Weiterhin nehmen sich Musiker bei der Interpretation und Umsetzung eines Notentextes erhebliche Freiheiten. Neben der Anpassung von Tempo und Dynamik kann eine Melodiestimme zum Beispiel durch Verwendung von Vibrato und Glissando ausgestaltet werden. Bei der Betrachtung von mehrstimmiger Musik überlagern sich diese Phänomene und führen zu komplexen Klanggemischen. In diesem Beitrag diskutieren wir automatisierte Verfahren mit dem Ziel, den Grundfrequenzverlauf der dominanten Melodiestimme aus einer Musikaufnahme zu extrahieren. Hierbei wollen wir insbesondere auf Probleme eingehen, die bei der Betrachtung mehrstimmiger Musik entstehen. Weiterhin zeigen wir auf, wie sich die Resultate automatisierter Verfahren durch Integration von Zusatzwissen und durch Möglichkeiten der Benutzerinteraktion verfeinern lassen. Für die Anwendung der vorgeschlagenen Methoden stellen wir einen Prototyp für eine Benutzerschnittstelle zur Bestimmung von Grundfrequenzverläufen vor.

1 *Danksagung:* Die Autoren Jonathan Driedger und Meinard Müller arbeiten für die International Audio Laboratories Erlangen – eine Gemeinschaftseinrichtung der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) und des Fraunhofer-Institut für Integrierte Schaltungen IIS. Die vorgestellte Arbeit wurde durch die Deutsche Forschungsgemeinschaft gefördert (DFG MU 2686/6-1). Wir danken an dieser Stelle auch den Gutachtern dieses Artikels für die vielen konstruktiven Verbesserungsvorschläge.

Abstract

In the musical score, most information about the melody of a piece of music is given explicitly. The melody, which is often notated in a separate staff, is specified by a sequence of note symbols that encode the notes' pitches, onset times, and durations. A recording of the same piece of music constitutes a much more complex representation. The waveform of even a single note played on an instrument already consists of different frequency components, among them a fundamental frequency as well as a number of overtones. Furthermore, musicians make their performance more expressive by changing tempo and dynamics, by adding vibrato to played notes, or by smoothly connecting subsequent notes with glissandi. In a polyphonic music recording, all these effects and nuances superimpose to a complex sound mixture, in which the information about the melody is given only implicitly. As a result, the estimation of the sequence of fundamental frequency values, which represent the melody, is a challenging task. In this contribution, we discuss computational approaches for automatically estimating the fundamental frequency track of the dominant melodic voice in a polyphonic music recording. Furthermore, we show how the estimation result can be improved by integrating additional knowledge about the piece of music and by means of user interaction. Finally, we present a prototype of a user interface that integrates the presented techniques and functionalities for interactive fundamental frequency estimation.

1 Einleitung

Die Melodie ist in der Musik ein grundlegendes Konzept. Viele Musikstücke enthalten charakteristische Melodielinien, die leicht im Gedächtnis haften bleiben. In Abbildung 1a ist ein Auszug aus dem Notentext der Oper *Der Freischütz* von Carl Maria von Weber dargestellt. Die in diesem Abschnitt vorkommende Melodiestimme (die in grau unterlegte Gesangspassage für Sopran) lässt sich im Notentext unmittelbar ablesen, da die Tonhöhen und Längen der zu singenden Melodietöne explizit durch Notensymbole kodiert sind. Wird ein Musikstück aufgeführt, werden Noten zu Klängen. Beim Spielen oder Singen eines Tons entsteht ein Klangereignis, das sich als Überlagerung von unterschiedlichen Frequenzen darstellen lässt. Neben möglichen rauschartigen Frequenzkomponenten kann ein harmonischer Klang insbesondere durch eine Grundfrequenz und ihre Obertöne beschrieben werden.

Die automatisierte Schätzung von Grundfrequenzverläufen in Musikaufnahmen ist seit vielen Jahren Gegenstand regen Forschungsinteresses. Ziel dieser Aufgabenstellung ist es, den Verlauf der Grundfrequenz einer Melodiestimme, oft auch als *F0-Trajektorie* bezeichnet, in einer gegebenen Musikaufnahme zu bestimmen (Salamon et al., 2014).

Im Gegensatz zu notentextbasierten Darstellungsformen ist der Melodiestimmenverlauf in einer Musikaufnahme nicht direkt ersichtlich. Dies wird durch Abbildung 1b illustriert, die die *Wellenform* (Zeit-Amplituden Darstellung der

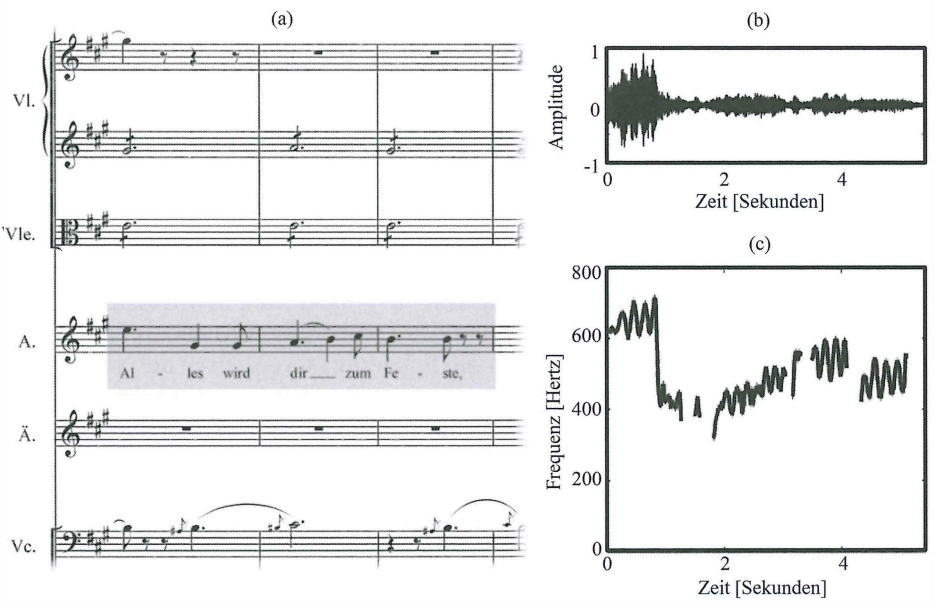

Abb. 1:
Auszug aus der Oper „Der Freischütz"
(a): Notentext, (b): Wellenform, (c): F0-Trajektorie

Schallwelle) einer Musikaufnahme unseres Beispiels zeigt. Die Wellenform entsteht durch Überlagerung der von den beteiligten Instrumenten erzeugten Schallwellen, von denen jede einzelne wiederum aus einer Vielzahl unterschiedlicher Frequenzkomponenten besteht. Weiterhin wird die Wellenform von Faktoren wie Resonanzen, Interferenzen und Raumakustik beeinflusst. Die Bestimmung der sich über die Zeit verändernden Grundfrequenz der Melodiestimme aus einer Wellenform stellt daher im Falle mehrstimmiger Musik eine schwierige Aufgabenstellung dar.

In diesem Beitrag widmen wir uns der Problemstellung der Schätzung des Grundfrequenzverlaufs für komplexe Musikaufnahmen. Das Ergebnis einer solchen Schätzung ist eine F0-Trajektorie wie sie in Abbildung 1c angedeutet wird. Zu jedem Zeitpunkt der zu analysierenden Musikaufnahme ist entweder die Grundfrequenz der Melodiestimme zu berechnen oder zu spezifizieren, dass die Melodiestimme zu diesem Zeitpunkt nicht aktiv ist. In der F0-Trajektorie unseres Beispiels lassen sich die Töne der Melodiestimme erahnen. Weiterhin sind auch periodische Veränderungen (Vibrato) sowie kontinuierliche Veränderungen (Glissando) der Grundfrequenz sichtbar. Diese Stilmittel werden von Sängern verwendet, um ihrem Gesang mehr Ausdruck zu verleihen. Im Gegensatz zum Notentext kann eine F0-Trajektorie somit auch interpretationsspezifische Feinheiten der Melodiestimme aufzeigen. Auf der anderen Seite ist die Übertragung einer F0-Trajektorie in eine Abfolge von Noten nicht trivial, da in

der F0-Trajektorie musikalische Notenparameter wie Tonhöhen, Dauern und Einsatzzeiten nicht explizit kodiert sind.

Ziel dieses Beitrages ist es, Herausforderungen und mögliche Lösungsansätze für das F0-Schätzproblem zu diskutieren. Hierbei betrachten wir das Problem eher aus Sicht der Signal- und Musikverarbeitung (und weniger aus Sicht der Musikpsychologie). Bei den vorgestellten Techniken werden häufig vereinfachende Modellannahmen getroffen, die sich in der Praxis jedoch als funktional erweisen. Zunächst gehen wir näher auf den Begriff der Grundfrequenz und des Grundfrequenzverlaufs ein (Abschnitt 2). Anschließend beschreiben wir ein grundlegendes Verfahren zur Schätzung von F0-Trajektorien in Musikaufnahmen und diskutieren Stärken und Schwächen des vorgestellten Ansatzes (Abschnitt 3). Abschließend zeigen wir, wie sich die Schätzungsergebnisse des vorgestellten Verfahrens durch die Integration von Zusatzwissen über die Musikaufnahme und durch Benutzerinteraktion verbessern lassen (Abschnitt 4). Hierzu stellen wir eine von uns entwickelte Benutzerschnittstelle vor.

Zum Abschluss dieses einleitenden Abschnitts wollen wir noch auf mögliche Anwendungen für die Grundfrequenzverlaufsschätzung eingehen. Eine prominente Anwendung aus dem Bereich der Musikverarbeitung stellt die *inhaltsbasierte Suche* in Musikdaten dar. Zum Beispiel wird in dem als *Query-by-Humming* bezeichneten Szenario eine gesummte oder gepfiffene Melodie als Suchanfrage verwendet. Das Ziel besteht dann darin, alle in einer Datenbank enthaltenen Aufnahmen, in denen die angefragte Melodie vorkommt, zu identifizieren. Um die einstimmige Anfrage und die typischerweise mehrstimmigen Musikaufnahmen vergleichbar zu machen, besteht ein Lösungsansatz darin, zunächst F0-Trajektorien für die Anfrage und die Melodiestimmen der Musikaufnahmen zu schätzen und diese dann abzugleichen (Salamon et al., 2013). Ein weiteres Anwendungsgebiet ist die *Quellentrennung* mit dem Ziel, eine bestimmte musikalische Stimme aus einer Musikaufnahme herauszutrennen, um auf diese Weise zum Beispiel eine Karaoke-Versionen zu generieren. In diesem Prozess ist die F0-Schätzung der abzutrennenden Stimme ein entscheidender Schritt (Ewert et al., 2014; Virtanen et al., 2008). Techniken zur Quellentrennung werden zum Beispiel im Bereich der Musikdidaktik bereits gewinnbringend eingesetzt (Dittmar et al., 2012). Bei der *automatisierten Transkription* von Musikaufnahmen geht es um die Überführung einer Musikaufnahme in eine Notenschriftdarstellung. Auch hier ist die Berechnung von F0-Trajektorien ein entscheidender Verarbeitungsschritt (Klapuri & Davy, 2006; Poliner et al., 2007). Im Anschluss an die F0-Schätzung werden die berechneten Trajektorien durch geeignete Segmentierungs- und Quantisierungsverfahren in diskrete Notenparameter transformiert. Hierbei müssen interpretationsabhängige Parameter (z. B. verursacht durch Rubato, Vibrato oder Glissando) herausgerechnet werden. Im Gegensatz dazu geht es bei der *Aufführungsanalyse* um die Erfassung solcher interpretationsabhängiger Feinheiten, die für einen bestimmten Musiker oder ein bestimmtes Genre charakteristisch sind. Auch hierbei ist eine exakte Schätzung von Grundfrequenzverläufen essenziell (Abeßer et al., 2014; Devaney & Ellis, 2008; Jers & Ternström, 2005).

2 Die Grundfrequenz

Ein auf einem Instrument gespielter Ton erzeugt eine im Allgemeinen schon recht komplexe Luftdruckschwankung. Versetzt man beispielsweise die Saite einer Geige durch das Streichen mit dem Bogen in Schwingung, so entsteht nicht eine rein sinusförmige Elementarschwingung, sondern eine Überlagerung ganz unterschiedlicher Phänomene. Neben den Obertonschwingungen werden zum Beispiel auch diverse andere Saiten und der Resonanzkörper zum Mitschwingen angeregt. Weiterhin wird der entstehende Klang auch durch die Raumakustik maßgeblich beeinflusst. Dennoch kann der Mensch dem Klang eines Tons eine bestimmte Tonhöhe zuordnen. Dieser Tonhöhe entspricht wiederum eine bestimmte Schwingungsfrequenz, die auch als *Grundfrequenz (F0)* des Klangs bezeichnet wird. Wie in Abbildung 2a illustriert, kann ein Klang als *Wellenform,* also als ein Zeit-Amplituden-Diagramm der gemessenen Luftdruckschwankungen, dargestellt werden.

Die menschliche Wahrnehmung eines Klangs und seiner Tonhöhe ist ein komplexes Phänomen (Cook, 2001; Gelfand, 2004; Hellbrück & Ellermeier,

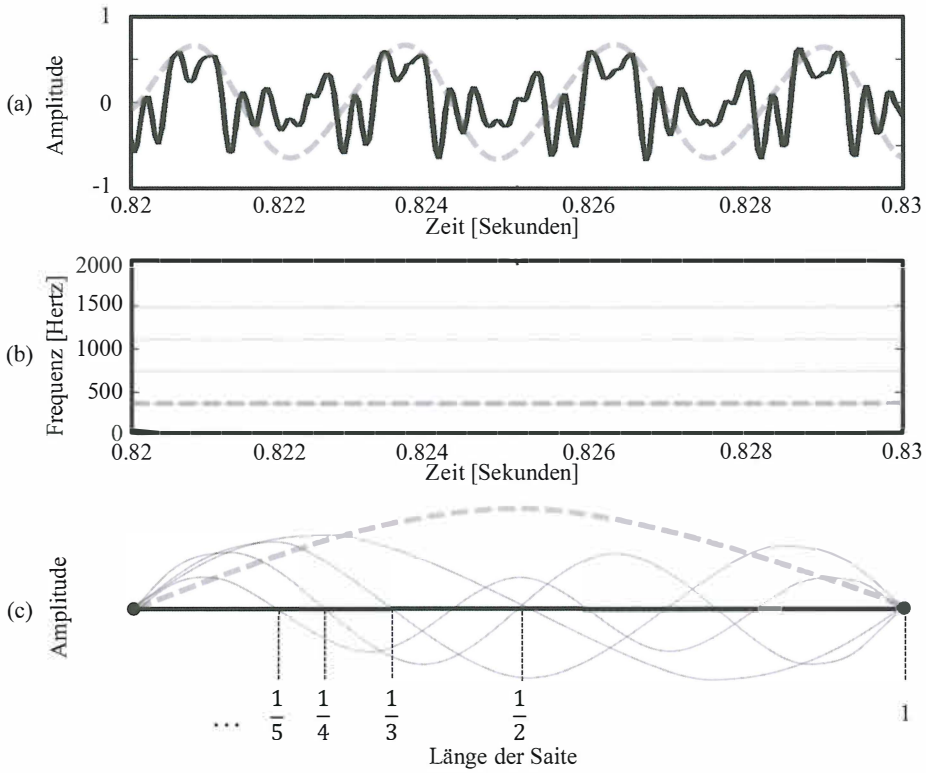

Abb. 2:
Geigenklang
(a): Wellenform, (b): Spektrogramm, (c): Schwingende Saite

2004; Yost, 2007). Zum Beispiel kann die Grundfrequenz völlig fehlen und dennoch wird eine Grundtonhöhe wahrgenommen (Residualton) (siehe Schouten, 1940). Bei vielen Instrumenten, wie etwa der Geige, sind die Grundfrequenzen weniger stark ausgeprägt als die der Obertöne, da diese Instrumente für die Abstrahlung ihrer tiefsten Frequenzen (der größten Wellenlängen) viel zu klein sind. Untersuchungen zeigen, dass weniger die Grundfrequenz, sondern eher die zeitabhängigen Periodizitäten in der Wellenform für die Tonhöhenempfindung verantwortlich sind. Im Allgemeinen entspricht die wahrgenommene Tonhöhe eines Klanges derjenigen Frequenz, die den größten gemeinsamen Teiler der Frequenzen aller an dem Klang beteiligten Teiltöne bildet (Meyer-Eppler et al., 1959). In der Literatur kommen daher häufig Methoden der *Autokorrelation* zur Anwendung, um Periodizitäten in der Wellenform zu messen (Hess, 1983). Aus den gemessenen Periodizitäten können dann Informationen zur Grundfrequenz und Tonhöhe abgeleitet werden.

Auf Autokorrelation basierende Verfahren liefern insbesondere für einstimmige Klänge robuste Schätzungen der Grundfrequenz (Cheveigné & Kawahara, 2002; Hess, 1983). Bei mehrstimmigen Klängen, wie zum Beispiel bei Aufnahmen polyphoner Musik, sind die Periodizitäten aufgrund der komplexen akustischen Überlagerungen der verschiedenen Stimmen stark gestört und kaum noch durch Autokorrelation zu erfassen. Um Klanggemische dieser Art zu analysieren, werden häufig Methoden der Fourieranalyse eingesetzt, bei denen die Wellenform in sinusartige *Elementarschwingung* zerlegt wird. Durch lokale Anwendung der Fourieranalyse (gefensterte Fouriertransformation) kann man die Wellenform in eine Zeit-Frequenz-Darstellung, ein sogenanntes *Spektrogramm*, überführen (siehe Abb. 2b). Bei der Berechnung eines Spektrogramms wird die Wellenform zunächst in kurze, sich überlappende Segmente konstanter Länge eingeteilt, die auch als *Frames* bezeichnet werden. Diese Frames sind für gewöhnlich nur wenige Millisekunden lang und repräsentieren somit lokale Eigenschaften der Wellenform. Die Wellenform eines jeden Frames wird anschließend durch Anwendung der Fouriertransformation mit einer Reihe von sinusartigen Elementarschwingungen verschiedener Frequenzen korreliert. Die Phasen dieser Schwingungen werden dabei so angepasst, dass die Korrelationen maximiert werden. Ein errechneter Korrelationswert gibt an, wie stark die entsprechende Frequenz in dem betrachteten Zeitfenster der Wellenform enthalten ist. Dieser Wert wird anschließend in einer Zeit-Frequenz-Ebene (als Grauwert kodiert) an der entsprechenden Stelle dargestellt. Das sich ergebende Spektrogramm gibt Aufschluss über die zeitabhängige Frequenzverteilung in der zugrunde liegenden Musikaufnahme.

Zur Illustration betrachten wir das in Abbildung 2 gezeigte Beispiel, das einen Ausschnitt eines Geigentons zeigt. Die Wellenform (Abb. 2a) lässt sich als Überlagerung in mehrere Elementarschwingungen auflösen. Die jeweiligen Beiträge (Magnituden) der beteiligten Elementarschwingungen werden durch das Spektrogramm kodiert (Abb. 2b). In unserem Beispiel beträgt die tiefste, in der Wellenform vorkommende Frequenz (grau gestrichelt unterlegt) 375 Hertz. Weiterhin zeigt das Spektrogramm, dass die Frequenzen der anderen dominant enthaltenen Schwingungen ganzzahlige Vielfache dieser Grundfrequenz sind.

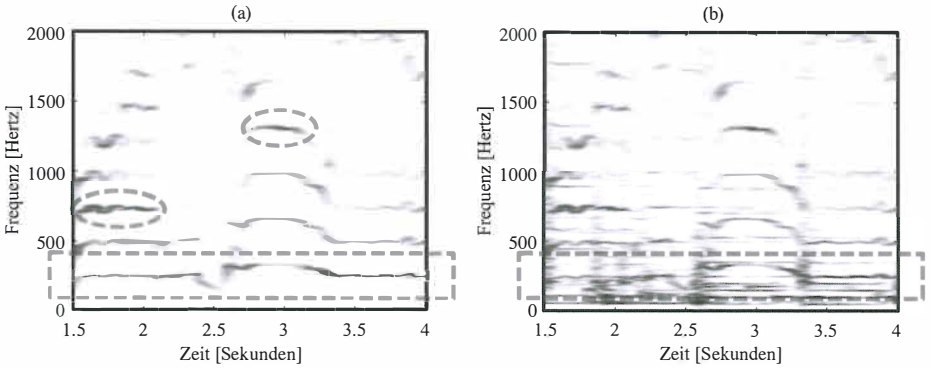

Abb. 3:
Spektrogramme einer Singstimme
(a): Einzelne Singstimme, (b): Singstimme mit Begleitung

Physikalisch kommen diese als *Obertöne* bezeichneten Schwingungen dadurch zustande, dass die Saite der Geige nicht nur auf ihrer gesamten Länge schwingt, sondern ebenfalls auf ihrer halben Länge, drittel Länge, viertel Länge, und so weiter (siehe Abb. 2c).

Wie bereits erwähnt, ist die wahrgenommene Tonhöhe mit der Grundfrequenz bzw. der entsprechenden Periodizität verknüpft; die Grundfrequenz muss jedoch nicht zwingend mehr Energie aufweisen als die Frequenzen der Obertöne. Zur Illustration dieses Phänomens zeigt Abbildung 3a ein Spektrogramm einer isoliert aufgenommenen Gesangsstimme, in dem sich deutlich die F0-Trajektorie (grau gestrichelt umrahmt) und die Trajektorien der Obertöne erkennen lassen. Man kann auch erkennen, dass einige der Obertöne in den Zeitbereichen um die zweite und dritte Sekunde (jeweils mit Ovalen markiert) deutlich mehr Energie aufweisen als die Grundfrequenz selbst. Dieses Phänomen tritt bei Singstimmen häufig auf, da der menschliche Sprachtrakt durch Resonanzen bestimmte Frequenzbereiche in der Stimme verstärkt. Fällt ein Oberton der Gesangsstimme in einen solchen Frequenzbereich, kann seine Energie die der Grundfrequenz übertreffen. Dieser Umstand ist ein Hauptgrund, warum es insbesondere bei fourierbasierten Methoden der Grundfrequenzschätzung zu Fehlern kommt. Allerdings können fourierbasierte Techniken auch im Falle von mehrstimmigen und komplexen Klanggemischen noch hilfreiche Information liefern.

Bevor wir in Abschnitt 3 auf ein solches fourierbasiertes Schätzverfahren eingehen, wollen wir die grundsätzlichen Schwierigkeiten der Grundfrequenzschätzung im Fall von mehrstimmiger Musik anhand eines Beispiels diskutieren. Hierzu betrachten wir die gleiche Gesangsstimme wie zuvor, diesmal allerdings begleitet von Gitarre, Klavier und Schlagzeug. In Abbildung 3b ist das Spektrogramm dieser mehrstimmigen Aufnahme dargestellt. Im Gegensatz zur isolierten Gesangsstimme (Abb. 3a), lässt sich die F0-Trajektorie im Falle der mehrstimmigen Aufnahme wegen der Überlagerungen mit den anderen Instrumenten kaum noch identifizieren. Dies liegt unter anderem daran, dass das Kla-

vier in einer tieferen und die Gitarre in einer ähnlichen Tonlage wie die Singstimme agieren. Dadurch wird die F0-Trajektorie der Gesangsstimme von zahlreichen Obertönen und Grundfrequenzen der begleitenden Instrumente überlagert. Allerdings fällt auf, dass sich die Obertöne der melodietragenden Gesangsstimme deutlich von den Frequenzkomponenten der anderen Instrumente abheben. Diese Beobachtung kann in einem grundlegenden Verfahren zur Grundfrequenzverlaufsschätzung, welches wir im folgenden Abschnitt beschreiben, ausgenutzt werden.

3 Grundlegendes Verfahren

In den vergangenen Jahren wurden unterschiedliche Verfahren zur Schätzung von Frequenztrajektorien entwickelt. Viele dieser Arbeiten – wie bereits in Abschnitt 2 angedeutet – messen die in der Wellenform vorkommenden Periodizitäten mittels Autokorrelation (siehe z. B. Cheveigné & Kawahara, 2002, oder Hess, 1983).

Die Autokorrelation liefert insbesondere bei einstimmigen Audioaufnahmen weitestgehend fehlerfreie Schätzungen von Grundfrequenzverläufen. Bei mehrstimmigen Musikaufnahmen sind diese Verfahren allerdings nicht ohne Weiteres anwendbar, da durch die Überlagerung unterschiedlicher Schallwellen keine klaren Periodizitäten in der Wellenform mehr vorliegen. Um die Verläufe von Melodiestimmen in mehrstimmigen Musikaufnahmen zu schätzen, haben sich Verfahren basierend auf Zeit-Frequenz-Darstellungen der Musikaufnahmen durchgesetzt (siehe z. B. Goto, 2004, Klapuri, 2008, und Salamon & Gómez, 2012). Das Ziel bei Goto (2004) ist die Berechnung der F0-Trajektorie für die Melodie und Bassstimme, bei Klapuri (2008) die Erkennung der Grundfrequenzverläufe aller an der Musikaufnahme beteiligten Instrumente und bei Salamon und Gómez (2012) die Berechnung der F0-Trajektorie der dominanten Melodiestimme. Den Verfahren ist gemein, dass sie, ausgehend von einer Zeit-Frequenz-Darstellung, eine sogenannte „Salienz-Darstellung" herleiten, aus der sich Grundfrequenzkandidaten ableiten lassen. Basierend auf der bei Salamon und Gómez (2012) vorgestellten Salienz-Darstellung stellen wir in diesem Abschnitt ein einfaches, aber effektives Verfahren zur Schätzung des Grundfrequenzverlaufes einer dominanten Melodiestimme vor.

Zur Illustration verwenden wir im Folgenden den schon in Abschnitt 1 vorgestellten Ausschnitt aus der Oper *Der Freischütz*. Ausgangspunkt unseres Verfahrens ist eine als Wellenform gegebene Musikaufnahme (siehe Abb. 4a). Wie bereits in Abschnitt 2 diskutiert, ist in der Wellenformdarstellung der Verlauf der Grundfrequenz nicht ersichtlich. Daher berechnen wir in einem ersten Verarbeitungsschritt ein Spektrogramm der Musikaufnahme (siehe Abb. 4b). In dieser Zeit-Frequenz-Darstellung kann man insbesondere im hochfrequenten Teil modulierende Strukturen erkennen, welche den Obertönen der mit Vibrato gesungenen Melodiestimme entsprechen. Allerdings ist im tieffrequenten Teil der Grundfrequenzverlauf der Melodiestimme aufgrund von Überlagerungen mit Frequenzkomponenten anderer Instrumente nur schlecht zu erkennen. Um eine

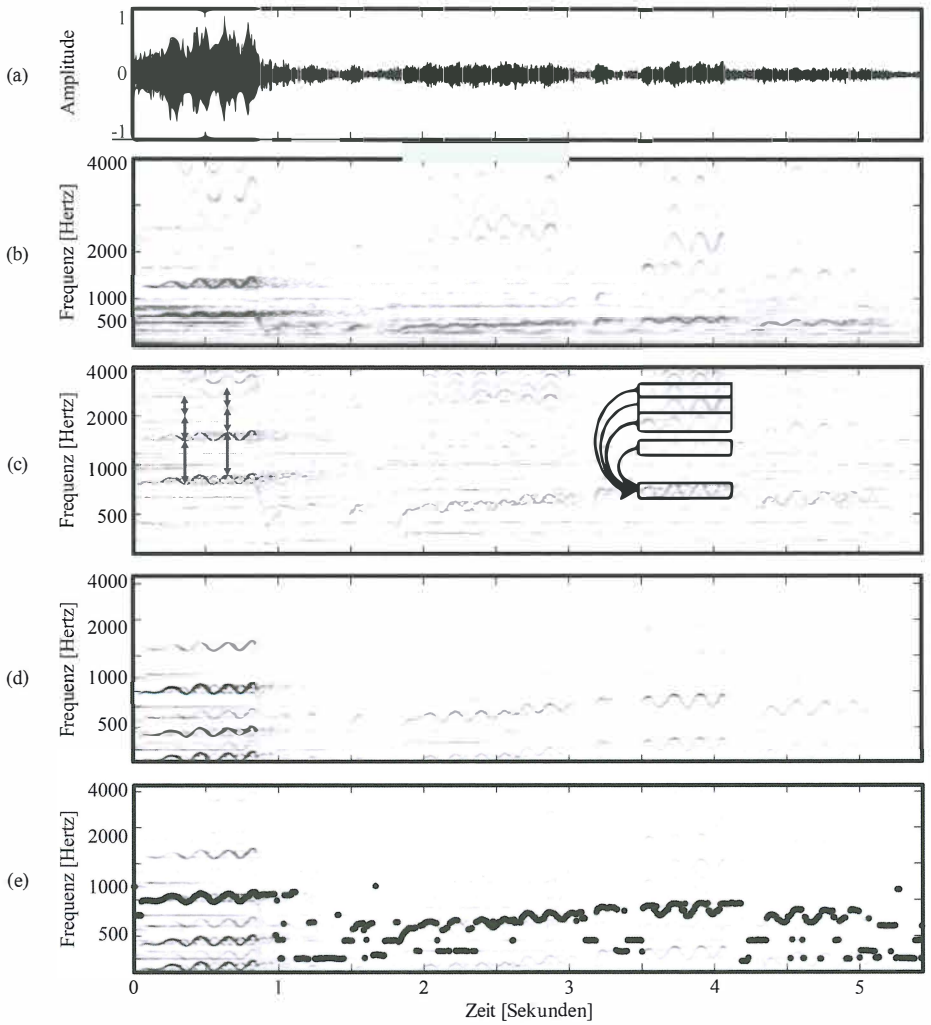

Abb. 4:
Grundlegendes Schätzungsverfahren. (a): Wellenform, (b): Spektrogramm, (c): Log-Spektrogramm, (d): Salienz-Darstellung, (e): F0-Trajektorie

weitere Verarbeitung zu erleichtern, wird das Spektrogramm in einem nächsten Schritt in ein sogenanntes *Log-Spektrogramm* überführt (siehe Abb. 4c). Hierzu wird die lineare Frequenzachse in eine logarithmische Frequenzachse transformiert. Für die Berechnung von F0-Trajektorien ist eine logarithmische Aufteilung der Frequenzachse vorteilhaft, da hierdurch der tieffrequente Bereich des Spektrogramms, in dem die Grundfrequenzverläufe zu erwarten sind, entzerrt wird. Ein weiterer Vorteil der logarithmischen Frequenzachse besteht darin, dass die multiplikativ von der Grundfrequenz abhängigen Obertonstruktur in eine

additiv abhängige Struktur übergeht (angedeutet mit Pfeilen im linken Teil von Abbildung 4c). Dies hat unter anderem zur Folge, dass die Frequenztrajektorien der Obertöne durch Translation (und nicht durch Skalierung) aus der Trajektorie der Grundfrequenz hervorgehen. Diese Eigenschaft wird insbesondere bei den aufgrund des Vibratos modulierenden Frequenztrajektorien sichtbar (vgl. Abb. 4b und Abb. 4c).

Um die Frequenzauflösung insbesondere im tieffrequenten Bereich des Log-Spektrogramms zu verfeinern, kann die bei der Berechnung der gefensterten Fouriertransformation mitgelieferte Informationen über die Phasen der Elementarschwingungen ausgenutzt werden. Indem man die Ableitung des Phasenverlaufs als instantane Frequenz interpretiert, können kleine, über aufeinanderfolgende Frames auftretende Phasenverschiebungen zur Verbesserung der Frequenzschätzung verwendet werden (Flanagan & Golden, 1966). Diese verbesserte Schätzung ermöglicht eine Verfeinerung der Frequenzauflösung, was letztendlich zu einer exakteren Grundfrequenzberechnung führt (Salamon & Gómez, 2012).

Im Log-Spektrogramm lässt sich im Allgemeinen der Verlauf der Grundfrequenz bereits besser erkennen als im ursprünglichen Spektrogramm (siehe auch Abb. 4c). Allerdings, wie bereits in Abschnitt 2 diskutiert, können die Grundfrequenzen weniger stark ausgeprägt sein als die Frequenzen der Obertöne. Dieses Phänomen führt bei fourierbasierten Verfahren häufig zu Verwechslungen der Grundfrequenz- mit den Oberton-Trajektorien. Zur Reduzierung solcher Verwechslungen besteht die Idee des folgenden Verarbeitungsschritts darin, die in den Obertönen enthaltenen Energien auf den Grundton zu übertragen. Hierzu verwenden wir eine Technik, die auch als *harmonische Summation* bezeichnet wird (Klapuri, 2006; Salamon & Gómez, 2012). Jedem Eintrag im Log-Spektrogramm entspricht eine bestimmte Frequenz. Bei der harmonischen Summation werden nun für jeden Eintrag des Log-Spektrogrammes die zu ganzzahligen Vielfachen dieser Frequenz korrespondierenden Einträge aufsummiert. Diese Summe bildet den Eintrag für eine neue Darstellung, die wir als *Salienz-Darstellung* bezeichnen. Der Summationsprozess ist in Abbildung 4c durch mit Pfeilen verbundene Rechtecke angedeutet, während die resultierende Salienz-Darstellung in Abbildung 4d dargestellt ist. Durch den Summationsprozess wird dem einer Grundfrequenz entsprechenden Eintrag die Energie all seiner Obertöne zugeordnet. Hierdurch erhält der Eintrag eine große Gewichtung, selbst wenn in der Grundfrequenz kaum Energie enthalten sein sollte. Durch den Übergang zu einer logarithmischen Frequenzachse kann die harmonische Summation sehr effizient durch geeignete Translationen des Log-Spektrogramms und der Summation über diese Translate realisiert werden. Für eine Diskussion der für die Berechnung der Salienz-Darstellung verwendeten Parameter verweisen wir auf Salamon und Gómez (2012).

Die Salienz-Darstellung stellt die Grundlage für die weitere Schätzung der F0-Trajektorie der Melodiestimme dar. Die Grundannahme für die Schätzung – die auch häufig in bestehender Literatur der automatisierten Musikverarbeitung anzutreffen ist (Poliner et al., 2007; Salamon & Gómez, 2012) – ist, dass die Melodiestimme in gewisser Weise die in der Aufnahme dominante Stimme ist.

Hierbei bezieht sich die Dominanz auf die im Grundton und all seinen Obertönen enthaltene Gesamtenergie, die gerade durch die Salienz-Darstellung erfasst wird. In einem ersten Ansatz zur Schätzung der gewünschten Grundfrequenz-Trajektorie betrachten wir zu jedem Zeitpunkt die Frequenz, die dem Eintrag mit dem höchsten Wert in der Salienz-Darstellung entspricht. In unserem Beispiel (siehe Abb. 4e) kann man erkennen, dass die sich ergebende Trajektorie den tatsächlichen Verlauf der Grundfrequenz (dargestellt in Abb. 1c) für die meisten Zeitpunkte gut abbildet. Allerdings gibt es auch einige Stellen, an denen die errechnete Trajektorie nicht der Gesangsstimme, sondern einer Begleitstimme folgt. Weiterhin weist das vorgestellte Verfahren jedem Zeitpunkt einen Grundfrequenzwert zu – unabhängig davon, ob die Melodiestimme zu diesem Zeitpunkt tatsächlich aktiv ist oder nicht. Im folgenden Abschnitt werden wir uns mit Möglichkeiten der Korrektur solcher Fehler auseinandersetzen.

4 Integration von Zusatzwissen und Benutzerinteraktion

Das im vorangegangenen Abschnitt diskutierte automatisierte Verfahren liefert Schätzungen des Grundfrequenzverlaufs der dominanten Melodiestimme – zumindest für die meisten Zeitpunkte. Allerdings ist für viele Anwendungen die Qualität der so berechneten Trajektorien noch nicht ausreichend. In diesem Abschnitt beschäftigen wir uns daher mit Strategien zur Verbesserung der Ergebnisse, die Zusatzwissen über den Grundfrequenzverlauf in den Schätzungsprozess einfließen lassen.

Betrachtet man das Beispiel in Abbildung 4e, so ist zu erkennen, dass die dort dargestellte F0-Trajektorie zahlreiche Sprungstellen und Ausreißer aufweist. Diese Art von Unstetigkeiten kommt dadurch zustande, dass das grundlegende Verfahren den Grundfrequenzwert zu einem Zeitpunkt unabhängig von seinem zeitlichen Kontext wählt. In der Praxis kommen solche abrupten Änderungen im Grundfrequenzverlauf nur selten vor. Insbesondere sind die Ausreißer schon aus rein physikalischen Gründen bei der Klangerzeugung auf realen Instrumenten nicht möglich. Im Allgemeinen ist daher anzunehmen, dass es sich bei den meisten Unstetigkeitsstellen um Fehler in der Trajektorien-Schätzung handelt. Um diese zu vermeiden, können zusätzliche Annahmen hinsichtlich der Stetigkeit von Trajektorien getroffen werden. Die Kernidee besteht darin, allen möglichen Sprungweiten in einer potenziellen Trajektorie „Kosten" zuzuordnen. Um die Stetigkeit von Trajektorien zu begünstigen, werden kleine Sprünge mit geringen Kosten belegt, während große Sprünge durch hohe Kosten bestraft werden. Mit Hilfe von Optimierungstechniken, die auf dem Paradigma der dynamischen Programmierung basieren (Müller, 2007), lässt sich im Anschluss eine kostenminimierende Trajektorie effizient berechnen. Diese Trajektorie weist im Regelfall nur noch wenige Sprungstellen auf, die typischerweise den Notenwechseln in der Melodiestimme entsprechen. Die meisten Ausreißer hingegen werden durch das Verfahren beseitigt.

In Abbildung 5a sehen wir eine solche kostenminimierende Trajektorie, die deutlich weniger Unstetigkeitsstellen aufweist als die vorherige Trajektorie in

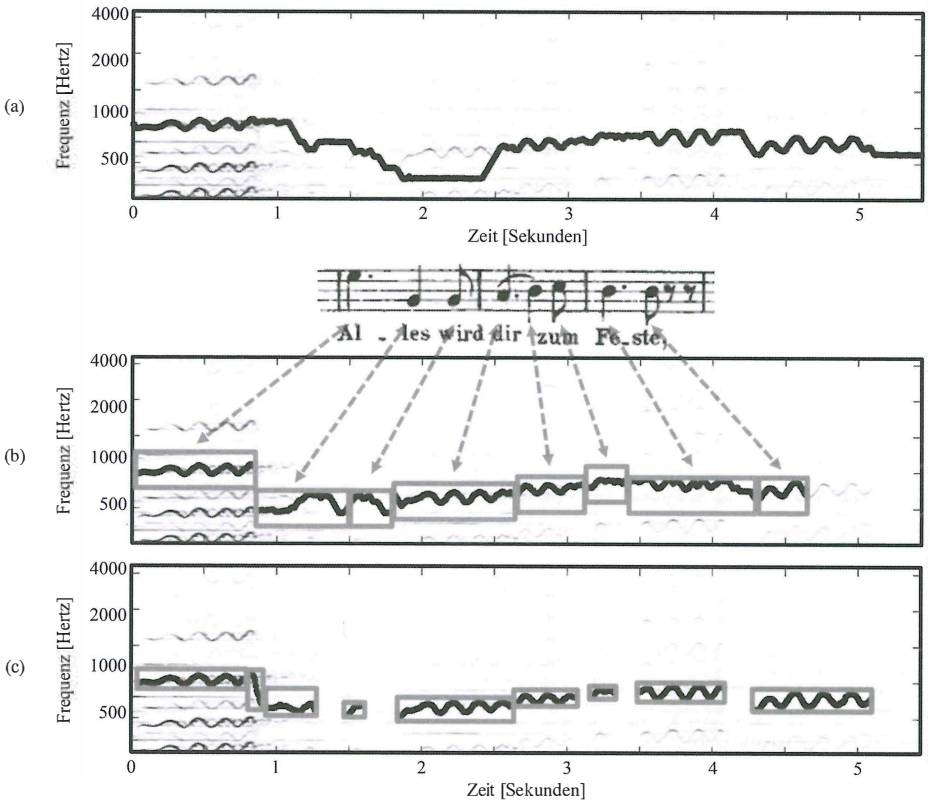

Abb. 5:
Verbesserung von Schätzungsergebnissen
(a): Trajektorie mit minimaler Anzahl von Sprungstellen, (b): Trajektorie in aus Notentext abgeleiteten Einschränkungsbereichen, (c): Trajektorie in manuell angepassten Einschränkungsbereichen

Abbildung 4e und den wahren Verlauf der Grundfrequenz wesentlich besser widerspiegelt. Allerdings ist die optimierte Trajektorie in einigen Passagen noch immer fehlerhaft. Zum Beispiel wird im Zeitbereich um die zweite Sekunde herum fälschlicherweise der Frequenzverlauf einer Begleitstimme erfasst. Diese Art von Verwechslungen ist in der Praxis häufig anzutreffen. Ein Grund liegt darin, dass häufig auch die Begleitstimmen starke harmonische Komponenten aufweisen können, die dann zu hohen Werten bei der harmonischen Summation in der Salienz-Berechnung führen.

Einen Ansatz zur Lösung dieses Problems stellt die Integration von zusätzlichem Wissen über die Melodiestimme dar. Hat man beispielsweise Zugriff auf den Notentext der Melodiestimme, so lassen sich anhand der Notensymbole geeignete Einschränkungsbereiche in der Salienz-Darstellung definieren, mittels derer die Trajektorien-Berechnung gesteuert werden kann (siehe Abb. 5b). Bei Vorliegen des Notentextes in einer computerlesbaren Form können diese Ein-

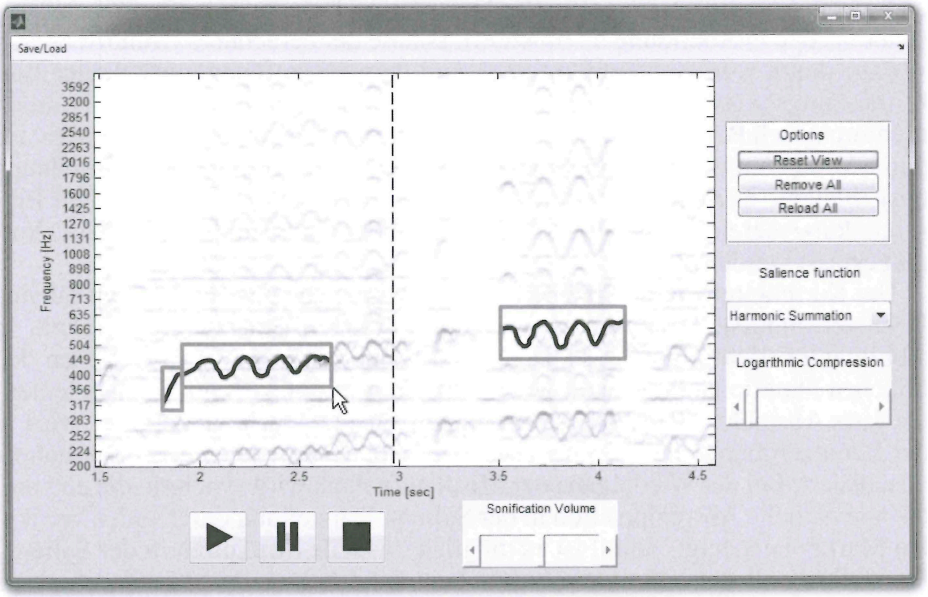

Abb. 6:
Grafische Oberfläche einer Benutzerschnittstelle zur interaktiven Schätzung von Grundfrequenzverläufen

schränkungsbereiche mit Hilfe von Techniken der Musiksynchronisation vollautomatisch bestimmt werden (Müller, 2007). Wie in unserem Beispiel angedeutet, wird die F0-Trajektorie lediglich innerhalb der (in grau angedeuteten) Einschränkungsbereiche berechnet, wodurch Verwechslungen mit anderen musikalischen Stimmen zu einem großen Teil vermieden werden. Zusätzlich können auf diese Weise diejenigen Bereiche, in denen die Melodiestimme nicht aktiv ist, von der Trajektorien-Berechnung explizit ausgeschlossen werden.

Der Einsatz von Einschränkungsbereichen kann jedoch auch zu Problemen führen. So hält sich beispielsweise die Sängerin in der betrachteten Aufnahme nicht genau an den Notentext und singt bei der vorletzten Note nicht das notierte *H4,* sondern ein *C#5*. Durch diese Abweichung liegt die gewünschte F0-Trajektorie außerhalb des vom Notentext abgeleiteten Einschränkungsbereichs und kann dadurch nicht richtig berechnet werden. Ein weiteres Beispiel für eine Abweichung vom Notentext zeigt sich bei der letzten Note. Hier ist lediglich eine Achtelnote notiert, die Sängerin hält diese jedoch wesentlich länger aus. Durch den aus der Achtelnote abgeleiteten Einschränkungsbereich wird die F0-Trajektorie in der Berechnung zu früh abgeschnitten.

Die Erkennung und Korrektur solcher Fehler mit automatisierten Verfahren ist im Allgemeinen schwierig, kann aber häufig vom Menschen mit wenig manuellem Aufwand durchgeführt werden. Zum Beispiel kann ein Benutzer die fehlerhaften Einschränkungsbereiche geeignet verschieben, verlängern, oder

anderweitig modifizieren, um so die Trajektorien-Berechnung interaktiv zu korrigieren. Wie in Abbildung 5c illustriert, konnte der berechnete Grundfrequenzverlauf durch wenige Modifikationen und dem Hinzufügen zusätzlicher Einschränkungsbereiche erheblich verbessert werden. Insbesondere konnten hierdurch auch Passagen, in denen die Melodiestimme kurzzeitig nicht aktiv ist, durch Verkürzen von Einschränkungsbereichen für die Trajektorien-Berechnung ausgeblendet werden. Weiterhin konnte durch Hinzufügen eines neuen Einschränkungsbereichs das Glissando zwischen der ersten und zweiten Note korrekt erfasst werden.

Um Korrekturen dieser Art zu erleichtern, wurde von uns ein Prototyp für eine Benutzerschnittstelle zur interaktiven Grundfrequenzschätzung entwickelt. In Abbildung 6 sieht man die grafische Oberfläche unserer Software. Neben den üblichen Funktionalitäten eines gewöhnlichen Audioplayers mit Bedienelementen zum Abspielen, Pausieren und Stoppen für die Musikwiedergabe, wird in der Benutzeroberfläche die Salienz-Darstellung der geladenen Musikaufnahme visualisiert. Bei der Wiedergabe der Musikaufnahme wird synchron die entsprechende aktuelle Abspielposition in der Salienz-Darstellung durch einen vertikalen Marker angezeigt. Somit ist es möglich, visuelle Strukturen in der Salienz-Darstellung direkt mit den akustischen Eindrücken der Musikaufnahme abzugleichen. Ein Anwender kann nun die Einschränkungsbereiche direkt in die gezeigte Salienz-Darstellung einzeichnen oder diese aus synchronisierten Notentextdaten automatisiert ableiten und in das Programm laden. Diese Bereiche können anschließend editiert und durch weitere Bereiche ergänzt werden. Nach jeder Modifikation wird die F0-Trajektorie neu berechnet und entsprechend den Vorgaben angepasst. Als eine weitere Funktionalität unserer Software kann die aktuell geschätzte Trajektorie beim Abspielen der Musikaufnahme mittels eines Sinustongenerators sonifiziert und synchron mit der Musikaufnahme wiedergegeben werden. Der Anwender kann dadurch auch akustisch überprüfen, wie genau die aktuelle Schätzung der F0-Trajektorie ist. Schließlich kann der Benutzer auch auf die Berechnung und Visualisierung der Salienz-Darstellung Einfluss nehmen und zum Beispiel durch einen Regler (siehe „Logarithmic Compression“ in Abb. 6) schwache spektrale Komponenten anheben oder starke Komponenten abschwächen. Die berechnete Trajektorie kann zu jedem Zeitpunkt abgespeichert werden, um sie entweder in anderen Anwendungen zu verwenden oder mit der Bearbeitung zu einem späteren Zeitpunkt fortzufahren.

Neben der von uns vorgestellten Benutzerschnittstelle existieren eine Anzahl weiterer Softwareprogramme zur interaktiven Berechnung von Frequenztrajektorien – sowohl aus dem kommerziellen als auch aus dem akademischen Umfeld. Bei der Software „Melodyne“ der Firma Celemony wird eine Musikaufnahme in notenartige Klangobjekte (sogenannte „Blobs“) zerlegt, die dann interaktiv herausgegriffen und verändert werden können. Unter anderem wird für jedes Klangobjekt der Verlauf einer Frequenztrajektorie berechnet und angezeigt, wodurch der Tonhöhenverlauf dieses Bausteines veranschaulicht wird. Die Zerlegung in „Blobs“ und damit auch die Grundfrequenzschätzungen können durch Integration von Zusatzwissen (zum Beispiel Tonart oder Analyseparameter) beeinflusst werden. Die Hauptfunktionalität von Melodyne besteht allerdings in

einer notenbasierten Modifikation der Musikaufnahme und weniger in der Bestimmung einer der Melodie entsprechenden Grundfrequenztrajektorie.

Im akademischen Umfeld wurden ebenfalls Benutzerschnittstellen zur interaktiven Berechnung von F0-Trajektorien entwickelt, wie zum Beispiel von Mauch et al. (2014) und Pant et al. (2010). Diese Programme bieten eine Vorauswahl von automatisch errechneten Trajektorien an. Aus dieser Vorauswahl kann dann ein Benutzer interaktiv gewisse Trajektorien aussuchen oder verwerfen. Diese Vorgehensweise ist sehr zeiteffizient, da die Auswahlmöglichkeiten beschränkt sind, und es nicht nötig ist, die gesamte Trajektorie feingranular auf ihre Korrektheit zu überprüfen. Allerdings kann es insbesondere bei mehrstimmigen Musikaufnahmen passieren, dass sich unter den vorausgewählten Trajektorien keine befindet, die den tatsächlichen Grundfrequenzverlauf der Melodiestimme angemessen widerspiegelt. Die Verwendung der von uns vorgestellten Benutzerschnittstelle zur Schätzung des Grundfrequenzverlaufs der Melodiestimme ist unter Umständen zeitaufwendiger. Gegebenenfalls müssen eine Vielzahl von Einschränkungsbereichen manuell definiert werden, um den Suchraum für die Trajektorien-Schätzung weit genug einzuschränken. Allerdings ist es hierdurch möglich, selbst für komplexe Musikaufnahmen gute Trajektorien-Schätzungen zu generieren.

Ein weiteres im akademischen Bereich verbreitetes Werkzeug ist das Programm „Praat“ (Boersma, 2001). Diese Software bietet insbesondere für die phonetische Analyse von Sprachaufnahmen eine Funktionalität für die Grundfrequenzschätzung. Hierbei kann ein Benutzer den Frequenzbereich, in dem die Trajektorie geschätzt werden soll, geeignet einschränken. So kann zum Beispiel berücksichtigt werden, ob es sich um eine männliche oder weibliche Stimme handelt. Für die Analyse von mehrstimmigen Musikaufnahmen scheint Praat allerdings nur bedingt geeignet zu sein, da die verwendeten Techniken auf der Autokorrelation basieren (Boersma, 1993).

5 Fazit

In diesem Beitrag wurde ein automatisiertes Verfahren vorgestellt, das den Grundfrequenzverlauf der dominanten Melodiestimme aus einer mehrstimmigen Musikaufnahme schätzt. Hierbei wurden zum einen typische Probleme, die sich bei der Schätzung ergeben, diskutiert und zum anderen Lösungsansätze angedeutet. Als ein Hauptbeitrag wurde eine interaktive Benutzerschnittstelle mit unterschiedlichen Visualisierungs-, Sonifikations- und Korrekturfunktionalitäten vorgestellt. Die in diesem Beitrag vorgestellten Techniken basieren aus psychoakustischer Sicht auf sehr stark vereinfachenden Annahmen. Wir haben gezeigt, wie auf Basis dieser Annahmen ein Verfahren zur Trajektorien-Berechnung realisiert werden kann, das zum einen praktikabel ist (die Trajektorien können zum Beispiel in Echtzeit modifiziert werden) und zum anderen eine einfache Integration von (Notentext-, Benutzer-) Vorwissen erlaubt. Für die zukünftige Arbeit ist zu testen, inwieweit das vorgestellte Verfahren die Arbeit in der Musikpsychologie zum Beispiel zum Zwecke der Interpretationsanalyse unterstützen kann.

Unsere bisherigen Ergebnisse zeigen (wie auch das Beispiel der Weber-Aufnahme illustriert), wie sich mit dem vorgestellten Werkzeug mit vertretbarem Aufwand hochqualitative F0-Trajektorien einer Melodiestimme erzeugen lassen.

Literatur

Abeßer, J., Pfleiderer, M., Frieler, K. & Zaddach, W. (2014). Score-informed tracking and contextual analysis of fundamental frequency contours in trumpet and saxophone jazz solos. In S. Disch, J. Herre, R. Rabenstein, B. Edler, M. Müller & S. Turowski (Eds.), *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx- 14)*, (pp. 181–186). Erlangen: International Audio Laboratories Erlangen.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, 17*, 97–11.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*(4), 1917–193. http://doi.org/1.1121/1.1458024

Cook, P. R. (2001). *Music, cognition, and computerized sound. An introduction to psychoacoustics.* Cambridge, UK: MIT Press.

Devaney, J. & Ellis, D. P. W. (2008). An empirical approach to studying intonation tendencies in polyphonic vocal performances. *Journal of Interdisciplinary Music Studies, 2*(1–2), 141–156.

Dittmar, C., Cano, E., Abeßer, J. & Grollmisch, S. (2012). Music information retrieval meets music education. In M. Müller, M. Goto & M. Schedl, (Eds.), *Multimodal music processing* (Dagstuhl Follow-Ups, 3, pp. 95–120). Wadern: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Ewert, S., Pardo, B., Müller, M. & Plumbley, M. (2014). Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine, 31*(3), 116–124. http://doi.org/1.1109/MSP.2013.2296076

Flanagan, J. L. & Golden, R. M. (1966). Phase vocoder. *Bell System Technical Journal, 45*, 1493–1509. http://doi.org/1.1002/j.1538-7305.1966.tb01706.x

Gelfand, S. a. (2004). *Hearing – An introduction to psychological and physiological acoustics.* New York: Informa Healthcare.

Goto, M. (2004). A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in realworld audio signals. *Speech Communication (ISCA Journal), 43*(4), 311–329. http://doi.org/1.1016/j.specom.2004.07.001

Hellbrück, J. & Ellermeier, W. (2004). *Hören – Physiologie, Psychologie und Pathologie*. Göttingen: Hogrefe.

Hess, W. (1983). *Pitch determination of speech signals: Algorithms and devices (Springer Series in Information Sciences)*. Berlin: Springer-Verlag. http://doi.org/1.1007/978-3-642-81926-1

Jers, H. & Ternström, S. (2005). Intonation analysis of a multi-channel choir recording. *Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR), 47*(1), 1–6.

Klapuri, A. P. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes. In K. Lemström, A. Tindale & R. Dannenberg (Eds.), *International Society for Music Information Retrieval Conference (ISMIR)* (pp. 216–221). Victoria: University of Victoria.

Klapuri, A. P. (2008). Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(2), 255–266. http://doi.org/1.1109/TASL.2007.908129

Klapuri, A. P. & Davy, M. (Hrsg.). (2006). *Signal processing methods for music transcription.* New York: Springer. http://doi.org/1.1007/0-387-32845-9

Mauch, M., Cannam, C. & Fazekas, G. (2014, April). *Efficient computer-aided pitch track and note estimation for scientific applications, April 2014*. Extended abstract accepted at SEMPRE 2014 conference, London, UK.

Meyer-Eppler, W., Sendhoff, H. & Rupprath, R. (1959). Residualton und Formantton. *Gravesaner Blätter, 14*, 70–83.

Müller, M. (2007). *Information retrieval for music and motion.* Berlin: Springer Verlag. http://doi.org/1.1007/978-3-540-74048-3

Pant, S., Rao, V. & Rao, P. (2010). A melody detection user interface for polyphonic music. *National Conference on Communications (NCC),* 1–5.

Poliner, G. E., Ellis, D. W. P., Ehmann, A., Gómez, E., Streich, S. & Ong, B. (2007). Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1247–1256. http://doi.org/1.1109/TASL.2006.889797

Salamon, J. & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(6), 1759–177. http://doi.org/1.1109/TASL.2012.2188515

Salamon, J., Gómez, E., Ellis, D. W. P. & Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine, 31*(2), 118–134. http://doi.org/1.1109/MSP.2013.2271648

Salamon, J., Serrà, J. & Gómez, E. (2013). Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval, 2*(1), 45–58. http://doi.org/1.1007/s13735-012-0026-0

Schouten, J. F. (1940). The residue, a new component in subjective sound analysis. In *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 43*, 356–365.

Virtanen, T., Mesaros, A. & Ryynänen, M. (2008). Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In B. Raj, P. Smaragdis, D. Ellis, P. Wolfe & S. Makino (Eds.), *Proceedings of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, (pp. 17–22). Brisbane: ISCA.

Yost, W. (2007). *Fundamentals of hearing* (5th ed.). San Diego, CA: Emerald.