

# Automatische Musiktranskription (AMT)

Roland Stigge  
Institut für Informatik  
Humboldt-Universität zu Berlin  
stigge@informatik.hu-berlin.de

16. Juni 2003

## Zusammenfassung

Im vorliegenden Artikel wird ein Überblick über den Stand der Forschung im Bereich der maschinellen Erkennung musikalischer Noten und Einblick in ein Verfahren zur Notenbestimmung gegeben. Es werden Vergleiche zwischen bekannten Algorithmen angestellt und diese auf ihre Tauglichkeit hin überprüft.

## 1 Einleitung

Die Disziplin der Automatischen Musiktranskription (AMT) beschäftigt sich mit dem weiten Feld der maschinellen Analyse von Audiosignalen, welche als Samples, aber auch als MIDI-Daten o.ä. vorliegen können und hinsichtlich musikalischer Strukturen und Eigenschaften hin untersucht werden sollen. Dabei ist insbesondere ein formatiertes Notenbild (siehe Abb. 1) als Ergebnis interessant, wofür allerdings eine Reihe verschiedener Verfahren, welche beispielsweise versuchen, Instrumente, Tonhöhen und Metren zu erkennen, angewendet werden müssen.

Die Ursprünge der AMT lassen sich bis in die 1970er Jahre zurückverfolgen. Beginnend mit der einfachen monophonen Analyse [Moorer75] [Piszcalski77] bis hin zu komplexen Systemen [Klapuri98] wurden in den letzten 30 Jahren einige interessante Ergebnisse publiziert, wobei jedoch wenige den Weg in die praktische (und damit kommerzielle) Verwirklichung gefunden haben. Für polyphone Analyse ist im Moment nur ein kommerzielles System bekannt [IntelliScore], welches selbst noch mit verschiedenen Problemen zu kämpfen hat.

Die AMT kann als musikalisches Pendant zur maschinellen Spracherkennung angesehen werden und steht eher im Schatten letzterer. Auch diese blickt auf eine jahrzehntelange Geschichte zurück, wobei sich größere Erfolge in der Praxis erst in den letzten Jahren einstellen.

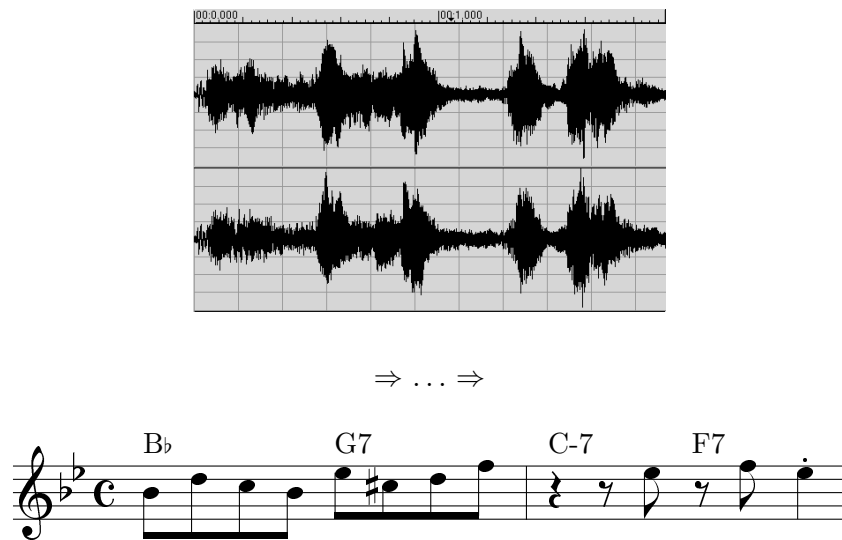


Abbildung 1: (Muster- bzw. Wunsch-) Beispiel einer kompletten Analyse: Oben: Audiosignal im Sampleformat, unten: (nach Rechnung) Ergebnis im Notenbild mit Taktart, Tonart, Noten (Höhen, Längen), Pausen, Akzenten, Akkorden

Konkrete Anwendung finden Teile der Ergebnisse der AMT heute schon im Musikerequipment wie Stimm- oder Effektgeräten. Komplexere und leistungsfähigere Systeme könnten neben der Erstellung kompletter Transkriptionen (d.h. Notation “abgehörter” Stücke ohne ursprüngliches Vorliegen von Noten) in der Titelerkennung und -kategorisierung (vgl. GEMA) und der Ähnlichkeitserkennung (vgl. Urheberrecht) Anwendung finden. Dies klingt besonders vor dem fragwürdigen aber unumgänglichen Hintergrund der Bestrebungen der Musikindustrie, möglichst frühzeitig Charakteristika von “Hits” zu bestimmen und wiederzuverwenden, interessant.

## 2 Problemseparation

Da das Forschungsgebiet eine sehr allgemeine Sicht auf die konkrete Analyse darstellt, muss die Automatische Musiktranskription in viele Teilgebiete aufgeteilt werden, wovon hier die wichtigsten genannt werden sollen:

- Notenanfangs-Erkennung (“Onset Detector”)
- Rhythmische Analyse (Taktart, “Beat”, Notenarten)
- Tonhöhenenerkennung (“Pitch Estimation”)
  - einstimmig (monophon)

– mehrstimmig (polyphon)

- Harmonische Analyse (“Akkorde”)
- Instrumentenerkennung

Die Notenanfangserkennung umfasst die Erkennung der Zeitpunkte, zu denen Instrumente bzw. Stimmen einen Ton beginnen lassen zu erklingen. Analog wird die sog. “Offset Detection” behandelt, welche das Ende eines gespielten Tones finden soll. Für die Erkennung von höheren musikalischen Strukturen ist jedoch vor allem die “Onset Detection” von Bedeutung, da damit das Metrum bestimmt wird.

Darauf aufbauend folgt die Rhythmische Analyse, welche höhere temporale Strukturen wie Tatum, Beat, Metrum, Form erkennen soll (siehe weiter unten).

Bei der Tonhöhenenerkennung werden diskrete Tonhöhen aus einem möglicherweise schwierigen (z.B. Vibrato, Verstimmung) Eingangssignal erkannt. Richtlinie ist hierbei die in der westlichen Musik übliche temperierte Skala, welche pro Oktave jew. die Frequenz verdoppelt und dabei 12 Halbtöne umfasst, wobei benachbarte Halbtöne immer in einem festen Verhältnis ( $2^{1/12}$ ) zueinander stehen. Analog könnten aber auch natürliche Stimmungen oder Skalen mit Vierteltönen usw. behandelt werden. Besondere Schwierigkeiten bereitet im Gegensatz zur monophonen (“einstimmigen”) die polyphone (“mehrstimmige”) Analyse, da sich hierbei sowohl Obertöne als auch weitere Klänge wie z.B. perkussive Instrumente (Schlagzeug) nicht-eindeutig überlagern.

Anhand der Tonhöhen kann nun eine Harmonische Analyse vorgenommen werden, wobei gleichzeitig erklingende Töne zu Akkorden zusammengefasst werden können. Hierbei kann bei bekannter Stilistik (Jazz / Pop / Klassik) bei fehlenden Akkordtönen auf fehlende Töne geschlossen werden.

Relativ separat kann die Technik der Instrumentenerkennung angesehen werden, wobei sowohl die Klasse des Instrumentes (Saiteninstrument (Streicher / “Zupfer”), Blasinstrument (Blech / Blatt), Perkussion) als auch konkrete Instrumente erkannt werden sollen [Eronen01]. Letzteres stellt sich als besonders schwierig dar, und die besten beschriebenen Systeme leisten in praktisch relevanten Fällen kaum mehr als eine Erkennungsrate von rund 35%. Hinzu kommt die Problematik gleichzeitig erklingender Instrumente, welche ihre Charakteristika schwer trennbar vermischen. Was der Mensch mit seinem komplexen System aus Ohr und Gehirn leicht erkennen kann, stellt sich für Maschinen als besonders schwierig heraus.

Auf die Instrumentenerkennung soll hier nicht näher eingegangen werden.

### 3 Entwicklung

Im Rahmen einer Studienarbeit am Lehrstuhl für Signalverarbeitung und Mustererkennung am Institut für Informatik (HU) wird hier ein System skizziert, welches die relevanten Felder des Problembereiches abdeckt.

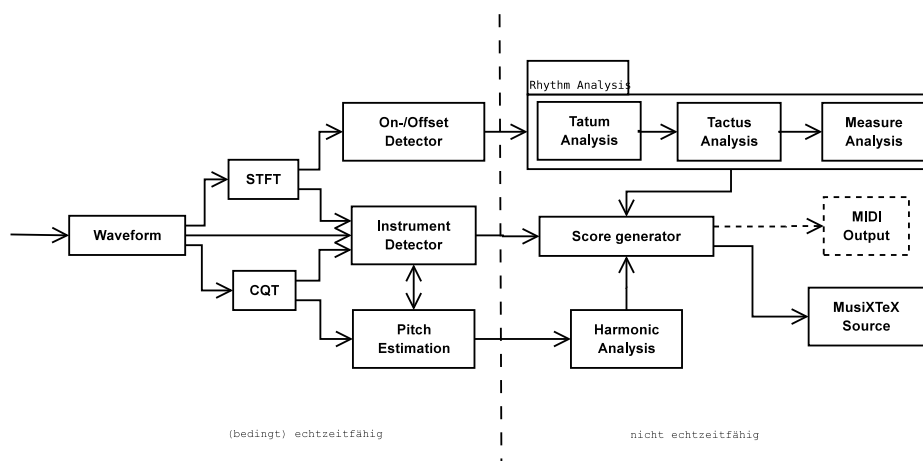


Abbildung 2: Signalflussdiagramm eines AMT-Systems. Links: echtzeitfähige Komponenten, rechts: Verarbeitung bei schon bekannten längeren zeitlichen Bereichen

In Abb. 2 wird der Signalfluss im System gezeigt. Auf der linken Seite werden grundlegende Eigenschaften des Signals analysiert, welches gespeichert vorliegen, aber auch "live" verarbeitet werden kann. Letzteres ist besonders vor dem Hintergrund der deutlichen Datenreduktion interessant, welche durch Instrumenten-, Tonhöhen-, und -anfangserkennung (*On-/Offset Detector*, *Instrument Detector*, *Pitch Estimation*) vorgenommen wird. Somit ist lediglich die Zwischenspeicherung der relevanten Merkmale anstelle des Sample-Eingangssignales nötig. Als Zwischentransformationen werden hierbei die wohlbekannte Diskrete Fouriertransformation in ihrer Ausprägung als Fast Fourier Transform (FFT; per Kurzzeit-Fouriertransformation, STFT) und die Constant-Q-Transformation (CQT, siehe nächster Abschnitt) als wichtige Vertreter in typischen AMT-Systemen angegeben, wodurch das Eingangssignal in eine für den vorliegenden Problembereich geeignete Darstellung gebracht wird.

Rechts in Abb. 2 wird die weitere Verarbeitung der analysierten Basismerkmale dargestellt. Mit dem Ergebnis der Notenanfangserkennung kann die Erkennung höherer zeitlicher Strukturen vorgenommen werden. Dieser Vorgang wird wiederum in mehrere Stufen untergliedert.

Grundlegend ist die Tatum-Erkennung. Dieses Kunstwort wurde von "Time Quantum" abgeleitet und bezieht sich auf das kleinste im Eingangs-

signal gefundene Zeitintervall. Hierauf baut der Beat (*Tactus*) auf, welcher mit dem wichtigsten wahrnehmbaren Puls gleichgesetzt wird (Dies sind z.B. im 3/4 oder 4/4-Takt die Viertel). Noch höhere Strukturen bestehen aus der dem zu analysierenden Stück zugrunde liegenden Taktart (*Measure*) und der Form (bestimmte Anzahl von Takten).

Zusammen mit dem Ergebnis der oben beschriebenen harmonischen Analyse können die Noten generiert werden (*Score Generation*). Als Ergebnis liegt dann eine Datei vor, welche die Noten in digitaler Form angemessen repräsentieren kann. Historisch ist hierfür besonders das MIDI-Format populär, welches jedoch manche Eigenschaften der Partitur nicht ausdrücken kann (deshalb die gestrichelte Darstellung). Daher wird hier ein ausdrucksstärkeres Format, wie z.B. MusiX $\text{\TeX}$  vorgeschlagen (siehe Abb. 3). Dies ist vor allem für das Druckbild interessant. Um Interoperabilität mit kommerziellen Notensatzprogrammen zu gewährleisten und Weiterverarbeitung zu ermöglichen, sind Formate wie das hier nicht dargestellte MusicXML sinnvoll.

```

1  \input musixtex
2  \Largemusicsize
3  \bigtype
4  \generalmeter\meterC
5  \generalsignature{-2}
6  \nostartrule
7  \nobarnumbers
8  \startpiece
9  \Notes\upertext{B$\scriptstyle\flat$}\Qqbl ikji\upertext{G7}\Qqbl l{^j}km\en
10 \bar
11 \Notes\upertext{C-7}\qp\ds\ca l\upertext{F7}\ds\ca m \upz l\qa l\en
12 \endpiece
13 \bye

```



Abbildung 3: Oben: Beispiel einer MusiX $\text{\TeX}$ -Datei, unten: deren Druckbild

## 4 Anwendung der CQT

An dieser Stelle soll exemplarisch der Bereich der Tonhöhenenerkennung näher erläutert werden.

Zusätzlich zur Diskreten Fourier-Transformation (DFT), welche in der Signalverarbeitung sehr verbreitet ist und effizient implementiert werden kann, ist die Q-Transformation mit konstantem Q (auch: *Constant Q Transform*, CQT [Brown91]) besonders auf die Bedürfnisse der Automatischen Musiktranskription zugeschnitten. Hierbei handelt es sich um eine Vari-

ante der Kurzzeit-Fouriertransformation (*Short Time Fourier Transform*, STFT), wobei eine logarithmische Frequenzskala eingeführt wird, auf der die Oktaven, welche jeweils die Frequenz verdoppeln, äquidistant verteilt sind. Die CQT betrachtet anstelle einer festen Fensterbreite (die für die DFT essenziell ist), eine variable Fensterbreite mit einer konstanten Anzahl an Schwingungen (die Konstante  $Q$ ) der jew. zu modellierenden bzw. analysierenden Teilfrequenz. (Die Fenstermitten werden hierbei um den betreffenden Zeitpunkt zentriert).

Aus der Formel für die Fourier-Transformation (hier ohne Normalisierung):

$$X(k) = \sum_{n=0}^{N-1} W_n x(n) \exp\left(\frac{-j2\pi kn}{N}\right)$$

wobei  $N$  die feste Fensterbreite,  $W_n$  die Fensterfunktion und  $k$  die jew. betrachtete Frequenz ist, kann unter Einführung der Normalisierung in Abhängigkeit einer Fensterbreite  $N(k)$  die Vorschrift für die Berechnung der CQT hervorgehen:

$$X(k) = \frac{1}{N(k)} \sum_{n=-\lfloor N(k)/2 \rfloor}^{\lceil N(k)/2 \rceil - 1} W_n(k) x(n) \exp\left(\frac{-j2\pi Qn}{N(k)}\right)$$

wobei  $W_n(k) = \alpha + (1 - \alpha) \cos\left(\frac{2\pi n}{N(k)}\right)$  für  $n = -\lfloor N(k)/2 \rfloor, \dots, \lceil N(k)/2 \rceil - 1$  die ursprungszentrierte Fensterfunktion (hier: Hamming-Fenster mit  $\alpha = 0.54$ ) darstellt, welche aus dem Eingangssignal einen auswertbaren Bereich herauschneidet und das Verhalten an den potentiellen Sprungstellen verbessert (siehe Abb. 4).

$N(k) = \left(\frac{N_{max}}{2^{k/12m}}\right)$  ist die variable Fensterbreite in Abhängigkeit von der niedrigsten zu analysierenden Frequenz bzw. der dazugehörigen Fensterbreite  $N_{max}$ , der Anzahl  $m$  der zu betrachtenden diskreten Frequenzen pro Halbton und natürlich  $k$ , der Nummer der jew. betrachteten Frequenz.  $m$  wird hierbei meist auf 2 oder 4 gesetzt, um die Frequenzauflösung und damit die Möglichkeiten der Analyse auszuweiten. Die Konstante  $Q$  ersetzt das  $k$  im Exponenten der DFT und kann auch als Frequenz je Bandbreite aufgefasst werden:  $Q = \lfloor f/\delta f \rfloor$ .

Die angepassten Grenzen der Summe tragen der Zentrierung der unterschiedlich großen Fenster Rechnung.

In Abb. 5 werden DFT (hier aus Effizienzgründen: FFT) und CQT gegenübergestellt. Während die lineare Frequenzskala auf der Ordinate bei der FFT dazu führt, dass sich die Tonhöhenunterschiede mehr in den Abständen der (bei natürlichen Instrumenten immer vorkommenden) Oberschwingungen als in der eindeutigen Lokalisierung der Grundfrequenz des Tones widerspiegeln, treten durch den logarithmischen Charakter der Frequenzachse bei der CQT die Tonhöhen als Positionen der Grundschwingungen hervor.

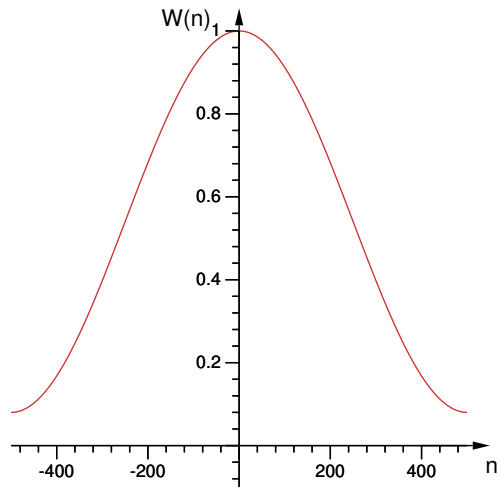


Abbildung 4: Hamming-Fenster (Beispiel:  $N(k) = 1000$ )

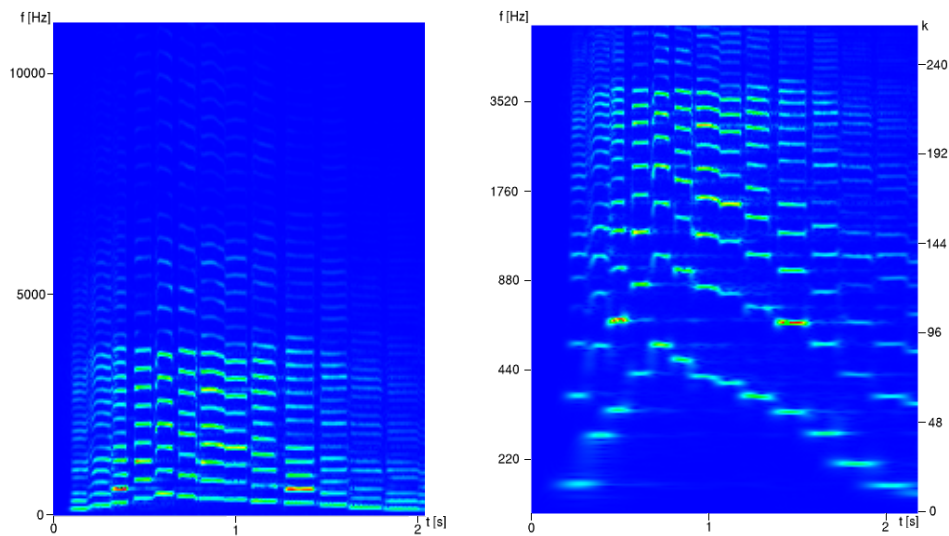


Abbildung 5: Vergleich der DFT mit der CQT. Links: FFT mit Fensterbreite  $N=1024$  bei 44.1kHz Samplingfrequenz, rechts: gleiches Signal nach CQT mit  $Q = 68$ ,  $m = 4$ , d.h. Achteltonabstand. Zusätzlich zur logarithmierten Darstellung auf der Ordinate ermöglicht die CQT für jede Frequenz eine optimierte Zeitunschärfe.

Ein wichtiges Merkmal der CQT ist, dass sie ermöglicht, die übliche Zeit-/Frequenzunschärfe zu optimieren. Wie im Beispiel sichtbar, sind niedrige

Frequenzen durch das große Fenster nur recht unscharf bzw. verschwommen zu erkennen, wohingegen hohe Frequenzen zeitlich gut lokalisiert sind. Die Frequenzauflösung bleibt pro Halbton konstant.

Man beachte den interessanten Effekt, dass bei der CQT zwar die Oberschwingungen nicht mehr äquidistant auf der Ordinate anzutreffen sind, sich jedoch bei verschiedenen Tonhöhen (d.h. Grundfrequenzen) die gleichen vertikalen Muster ergeben, welche jeweils lediglich um die senkrechte Position der Grundschwingung verschoben sind. Dadurch wird eine einfache Tonhöhenerkennung bereits durch Korrelation mit dem "idealen Muster" einer Oberschwingungsreihe ermöglicht [Brown92a]. Bei einem deutlichen Maximum wird hierbei wahrscheinlich ein entsprechend auftretender Ton gefunden.

Das Ergebnis einer solchen Berechnung ist in Abb. 6 zu sehen, wobei die gefundenen Grundschwingungen der Töne und die jeweils dazugehörigen Halbtöne der (temperierten) westlichen Skala bereits eingezeichnet wurden.

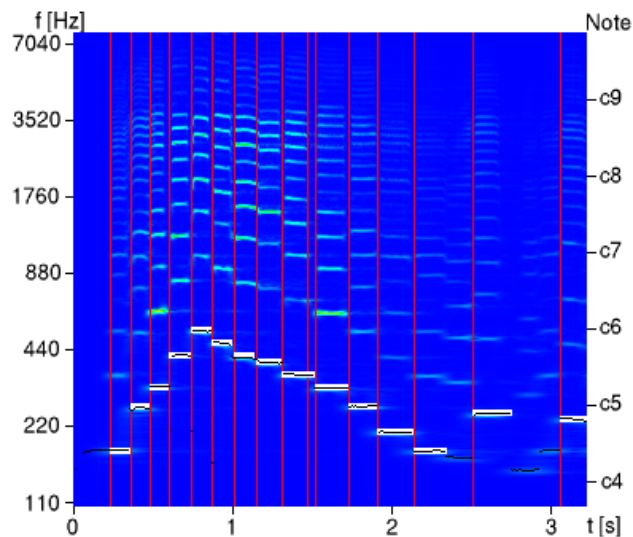


Abbildung 6: Das Ergebnis einer Tonhöhenerkennung. Zugrunde liegt das obige Ergebnis der CQT, welches mit den schwarzen jew. gefundenen Grundfrequenzen der Töne überlagert wurde. Die weißen Balken zeigen die Rundung zum wahrscheinlichsten korrespondierenden Ton. Zusätzlich wurden rot die erkannten Onsets markiert.

Leider ist bis jetzt für die CQT kein effizienter Algorithmus in der Ordnung der FFT bekannt, bei dem keine deutlichen Nachteile durch Optimierungen hervortreten. Beispielsweise hat die ursprüngliche Autorin einen schnellen Algorithmus vorgeschlagen [Brown92b], der jedoch durch die vorgeschaltete FFT die Zeitunschärfe für alle betrachteten diskreten Frequenzen mit der Unschärfe bei der niedrigsten zu betrachtenden Frequenz gleichsetzt.

Dadurch ist dieses Verfahren für die Erkennung schneller musikalischer Passagen ungeeignet, da hierbei oft im Bereich höherer Töne im betrachteten Fenster durchaus acht oder mehr Töne “gleichzeitig” erklingen würden.

Es besteht allerdings die Hoffnung, dass im Laufe der Arbeit am vorliegenden Projekt bessere Verfahren gefunden und integriert werden können, die die Vorteile der CQT bewahren.

In Abb. 6 ist desweiteren das Ergebnis der Onset Detection zu sehen, welche hier ein heuristisches Verfahren anwendet, um die Anfänge einzelner Oberschwingungen eines Tones zu finden und dann mehrere Oberschwingungen möglichst gut zusammenfasst. Dabei tritt immer eine gewisse Fehler auf, wie z.B. übersehene Tonanfänge bei leisen Tönen oder falsch erkannte Anfänge.

## 5 Ausblick

Durch die Diskrepanz zwischen den möglichen Anwendungen der entwickelten Technologie und dem internationalen Stand der Forschung auf dem Gebiet der AMT ergibt sich noch viel Arbeit in diesem Bereich. Insbesondere die Bereiche der polyphonen Tonhöhenenerkennung und der Instrumentenerkennung befinden sich noch stark in der Entwicklung. Bei der Tonerkennung wird mit moderneren Verfahren wie Neuronalen Netzen zur Erkennung überlagerter Klänge und Akkorde, Wavelettransformationen, Relativer Tonhöhenenerkennung und Stereosignalauswertung gearbeitet.

Bei der Instrumentenerkennung gibt es viele verschiedene Ansätze [Eronen01], um Eigenschaften aus Klängen zu extrahieren, wie z.B. Hidden-Markov-Modelle (HMM) zur Klassifikation, Support Vector Machines, Eigenschaften der spektralen und zeitlichen Hüllen von Klängen und deren Zusammenhänge, Untersuchung nicht-harmonischer Eigenschaften (d.h. Frequenzanteile ausserhalb der Oberschwingungsreihe), Eigenschaften der Phasen des Klanges (Attack, Decay, Sustain, Release). In der Praxis werden viele dieser Verfahren miteinander kombiniert, um die Erkennungsraten, welche zum gegenwärtigen Zeitpunkt noch verbesserungswürdig sind, zu erhöhen.

Ein dringendes Problem ist die Effizienzsteigerung bei der CQT, wobei alternative Verfahren mit ähnlichen Eigenschaften oder Verbesserungen des Algorithmus in Betracht gezogen werden. Eine vorgeschlagene Variante reduziert die Komplexität auf kleinere Fensterbreiten, indem zuerst nur die oberste zu analysierende Oktave untersucht wird und anschließend das gleiche Verfahren auf darunterliegende Oktaven angewendet wird. Allerdings führt dies zu großen Fehlern in den unteren Registern, da für die gewünschte Effizienzsteigerung “unsaubere” Filtertechniken beim “Downsampling” eingesetzt werden müssen.

Durch den Austausch mit dem Bereich Jazz der Universität der Künste Berlin wird versucht, die Anwendungsfälle der Musiker im Bereich der

künstlerischen Praxis auszuloten.

## Literatur

- [Brown91] Brown (1991), *Calculation of a constant Q spectral transform*, J. Acoust. Soc. Am., Januar 1991
- [Brown92a] Brown (1992), *Musical fundamental frequency tracking using a pattern recognition method*, J. Acoust. Soc. Am., September 1992
- [Brown92b] Brown (1992), *An efficient algorithm for the calculation of a constant Q transform*, J. Acoust. Soc. Am., November 1992
- [Eronen01] Eronen (2001), *Automatic musical instrument recognition*, MSc thesis, Tampere University of Technology
- [IntelliScore] Innovative Music Systems, Inc., Intelliscore:  
<http://www.intelliscore.net/>
- [Klapuri98] Klapuri (1998), *Automatic transcription of music*, MSc thesis, Tampere University of Technology
- [Moorer75] Moorer (1975), *On the Transcription of Musical Sound by Computer*, Computer Music Journal, Nov. 1977
- [Piszczałski77] Piszczałski, Galler (1977), *Automatic Music Transcription*, Computer Music Journal, 1(4), 1977