Experimentelle Ansätze zur Schlagzeuganalyse in Musiksignalen

Stephan Hübler und Rüdiger Hoffmann

TU Dresden - Institut für Akustik und Sprachkommunikation
stephan.huebler@tu-dresden.de

Kurzfassung: Das an der TU Dresden entwickelte UASR System (*Unified Approach for Signal Analysis and Recognition*), ursprünglich entwickelt für die Analyse und Synthese von Sprache, wird zunehmend auch für andere Signalarten verwendet. Der vorliegende Beitrag adaptiert das hierarchisch aufgebaute Systemkonzept für die Verwendung von Musiksignalen, insbesondere für westliche populäre Musik mit Schlagzeug. Nach den grundlegenden theoretischen Überlegungen werden zwei konkrete Entwürfe von Experimenten vorgestellt. Diese beschäftigen sich jeweils mit einem Teilgebiet des Gesamtsystems und dienen dessen schrittweiser Entwicklung. Mögliche Anwendungen sind die Transkription, Manipulation und Empfehlung von Musik.

1 Einleitung

Die Transkription von Musik ist Gegenstand aktueller Forschungen. Der vorliegende Beitrag geht insbesondere auf die Transkription von Schlagzeug in populärer westlicher Musik ein. Hauptaugenmerk liegt auf der Adaption des UASR Systems der TU Dresden, welches ursprünglich für Sprache entwickelt wurde. Sprach- und Musiksignale besitzen Gemeinsamkeiten, welche die Verwendung ähnlicher Technologien ermöglicht [6].

In der Literatur finden sich unterschiedliche Ansätze der Schlagzeugtranskription, welche übersichtlich in der Dissertation von Jouni Paulus aufgeführt sind [9]. Bei der *Segmentierung und Klassifikation* wird das Musiksignal mittels Anschlagserkennung oder eines musikalischen Rasters in bedeutungsvolle Einheiten geteilt. Diese einzelnen Teile werden dann klassifiziert. Der Ansatz der *Trennung und Klassifikation* trennt zunächst die einzelnen zu detektierenden Schlagzeuginstrumente im Signal in einzelne Audioströme. Dafür wird zumeist das Verfahren der Independent Component Analysis (ICA) oder der Prior Subspace Analysis (PSA) verwendet. Anschließend werden in den einzelnen Datenströmen zumeist mittels Anschlagerkennung und Klassifikation die einzelnen Noten erkannt. Eine *Simultane Segmentierung und Klassifikation* ist möglich, wenn die verschiedenen Hidden Markov Modelle der einzelnen Instrumente zu einem Erkennungsnetzwerk verknüpft werden, da dann bei der freien Erkennung die Anschlagszeitpunkte der einzelnen Instrumente und deren Klassifikation simultan durchgeführt werden. Zuletzt ist noch der Ansatz *Vergleichen und Anpassen* zu nennen, bei dem zunächst spektrale Templates der einzelnen Instrumente erstellt werden und diese dann mit dem Audiosignal verglichen werden. Gleichzeitig findet eine Anpassung der Templates statt. Die Einbeziehung des zeitlichen Kontext durch musikalische Modelle wird überwiegend durch die Anwendung von N-Grammen realisiert. Es kommt zu den bekannten Schwierigkeiten der großen nötigen Datenmenge für diese Modelle.

2 Adaption des UASR-Systemkonzeptes

Mithilfe des UASR-Systems der TU Dresden soll eine Schlagzeugtranskription realisiert werden. Der vorgestellte Systementwurf in Abbildung 1 ermöglicht dabei mit verschiedenen Ansätzen der Schlagzeugtranskription zu experimentieren. Es werden Überlegungen bis hin zur Einbindung von musikalischen Modellen vorgestellt. Kerngedanke des UASR-Systems ist die Analyse und Synthese von Signalen mit gleichen Datenbanken [7]. Schwerpunkt der vorliegenden Betrachtungen ist die Analyseseite des Systems. Die Synthese mit den gleichen Datenbanken eröffnet ein weiteres interessantes Forschungsgebiet. Durch die Möglichkeit der Midisynthese ergibt sich allerdings eine einfache Möglichkeit, Erkennungsergebnisse als Audio hörbar zu machen.

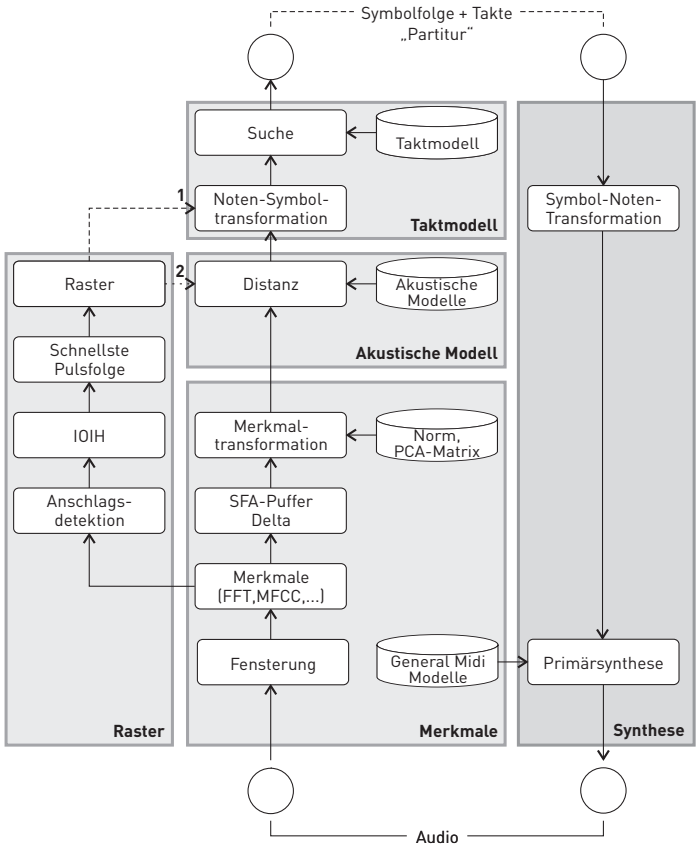

Abbildung 1 - Blockschaltbild von UASR für die Transkription von Schlagzeug unter Einbezug eines musikalischen Taktmodells.

Die einzelnen Blöcke Merkmale, akustisches Modell, Taktmodell und Raster werden in den nächsten Unterabschnitten näher erläutert. Bei alleiniger Verwendung der Blöcke Merkmal und akustisches Modell ist eine freie Erkennung, mithilfe eines Netzwerkes der akustischen Modelle, der einzelnen Schlaginstrumente möglich. Dies entspricht dem Ansatz *Simultane Segmentierung und Klassifikation*. Eine Verwendung von Bigrammen für die Übergangswahrscheinlichkeiten zwischen den einzelnen Modellen ist denkbar, wenn auch noch nicht in der Systemübersicht integriert. Bei Hinzunahme des Blockes Raster bei der Zuordnung der akustischen Modelle (Weg 2) kann von dem Ansatz *Segmentierung und Klassifikation* gesprochen werden. Der Block Raster ermittelt ein zeitliches Raster mit weitgehend gleichen Abständen zwischen den Rasterzeitpunkten. Die einzelnen Rasterpunkte stimmen idealerweise mit den Anschlagszeitpunkten der meisten musikalischen Ereignisse im Musikstück überein. Als musikalisches Modell wird

ein Taktmodell vorgeschlagen. Der Ansatz ist begründet in der These, dass die Abfolge von Schlagzeugsymbolen einen Takt kennzeichnet. Für die Verwendung dieser Grammatik ist ein musikalisches Raster zwingend erforderlich.

2.1 Merkmale

Jede Erkennung von Ereignissen in Audiosignalen erfordert die Extraktion von Merkmalen, welche die Ereignisse möglichst gut beschreiben und voneinander trennen. In der Literatur sind die Mel Frequency Cepstral Coefficients (MFCC) als Merkmale weit verbreitet [9, 2]. Außerdem werden Fast Fourier Transformation [1], Spectral Centroid, Spectral Width und Spectral Flatness [2] verwendet. Die Merkmalextraktion erfordert zunächste eine Fensterung des Ausgangssignals für die Berechnung der spektralen Eigenschaften. In einem weiteren Schritt kann die Berechnung von sekundären Merkmalen (SFA: secondary feature analysis) erfolgen. Es werden die Δ und $\Delta\Delta$ Merkmale berechnet und damit bereits der unmittelbare zeitliche Verlauf berücksichtigt. Die Merkmaltransformation dient der Entfernung von redundanten Informationen. Dies kann zum Beispiel mithilfe der Hauptkomponentenanalyse (PCA: Principal component analysis) geschehen. Diese benötigt eine aus Trainingsmaterial berechnete Transformationsmatrix.

2.2 Akustisches Modell

In der Literatur werden für die Klassifikation zumeist Support Vector Machine (SVM) oder Hidden Markov Modelle (HMM) verwendet (eine Übersicht findet sich in [9]). Beide haben die Möglichkeit eine Wahrscheinlichkeit für die verschiedenen möglichen Klassen von Schlagzeugnoten auszugeben und so jeweils eine Distanz zu jedem Modell abzubilden. Die Modelle für jede einzelne Note müssen trainiert werden und erfordern annotiertes Audiomaterial. Zunächst werden Detektorklassifikatoren erstellt. Ein Klassifikator für die Base Drum zeigt an, ob eine Base Drum im aktuellen Zeitabschnitt vorliegt oder nicht. Wenn im Signal mehrere Noten zur gleichen Zeit gespielt werden, dann können mehrere Detektorklassifikatoren gleichzeitig anspringen. Alternativ wären einzelne Klassifikatoren für die Kombination von Schlaginstrumenten denkbar. Dies stellt allerdings besondere Anforderungen an das Trainingsmaterial, da jede zu erkennende Kombination ausreichend oft im Trainingsmaterial auftreten muss.

2.3 Taktmodell

Die Abfolge von Schlagzeugnoten ist nicht willkürlich sondern durch die Metrik vorgegeben. Insbesondere populäre Musik mit konstantem Tempo hat eine ausgeprägte regelmäßige rhythmische Struktur. In der Literatur werden für die Abbildung solcher musikalischen Eigenschaften N-Gramme verwendet [2, 9]. Das vorliegende System geht von der Annahme aus, dass sich insbesondere über einen Takt regelmäßige rhythmische Muster ausbilden. Daher werden einzelne Takte für das Training eines Taktmodells verwendet. Dazu dienen die aus der Sprachverarbeitung bekannten endlichen Automaten. Für die Verwendung eines solchen Taktmodells müssen zwei Voraussetzungen erfüllt werden. Zum einen hat jeder Zustandsübergang auch eine musikalische Entsprechung in Form eines Notenwertes (z.Bsp. Achtel-, Sechzehntel,- etc.) und zum andern kann immer nur ein Symbol pro Zeitpunkt auftreten. Ersteres macht die Verwendung eines Rasters nötig (siehe Abschnitt 2.4). Zweiteres erfordert das zusammenlegen von gleichzeitig gespielten Schlagzeugnoten zu einem Symbol (Noten-Symbol-Transformationen). Für das Training eines Taktmodells sind wiederum annotierte Daten nötig, welche die gespielten Noten, deren Zeitpunkte, Taktgrenzen, Taktart und Tempoinformationen besitzen müssen.

2.4 Raster

Der zeitliche Abstand von Schlagzeugnoten unterliegt metrischen Beschränkungen, welche in populärer Musik verstärkt auftreten [3]. Diese musikalische Eigenschaft kann wertvolle Informationen für die Schlagzeugtranskription als auch für mögliche nachgeschaltete Algorithmen bieten. Als schnellste Pulsfolge wird dabei das Raster bezeichnet, dessen Rasterpunkte mit möglichst vielen Anschlagszeitpunkten der musikalischen Ereignisse im Musiksignal zusammen treffen. Bei gleichbleibendem Tempo ist der Abstand der Rasterzeitpunkte konstant. Ausgangspunkt des Systems ist die Anschlagsdetektion, welche alle Zeitpunkte im Audiosignal markiert an denen ein Instrument angeschlagen oder angezupft wird [5]. Auf der Grundlage der Anschläge kann ein Inter-Onset-Intervall Histogramm (IOIH) erstellt werden. Dieses Histgramm bildet die Häufigkeit der zeitlichen Abständer verschiedener Anschläge ab. Dies kann wiederrum als Grundlage für die Ermittlung der schnellsten Pulsfolge verwendet werden [4, 11]. Für die Verwendung eines Taktmodells werden besondere Anforderungen an das Raster gestellt, welche in Abschnitt 4.2 erläutert werden.

3 Anwendung

Unter der Voraussetzung eines funktionsfähigen Systems ergeben sich vielfältige Anwendungsmöglichkeiten. Es können globale rhythmische Eigenschaften wie Taktart, Taktgrenzen, Tempo und schnellste Pulsfolge ermittelt werden. Die Transkription kann als Grundlage für Ähnlichkeitsmaße von Schlagzeugrhythmen über einen Takt, bis hin zu rhythmischen Ähnlichkeiten ganzer Titel dienen. Weitere rhythmische Aspekte sind statistische Merkmale über die Verwendung der einzelnen Noten, rhythmische Genreklassifikation sowie Popularität und Komplexität des Rhythmus. Nicht zu vergessen ist die Möglichkeit der sinnvollen Segmentierung von Musikstücken zum Beispiel an den Taktgrenzen.

4 Experimentelle Ansätze

Im Folgenden werden zwei Experimente für verschiedene Teilgebiete des Systems vorgestellt. Sie beruhen zunächst auf der Verwendung von Midi Dateien. Dieses Dateiformat kann als Transkription von Musik aufgefasst werden und ermöglicht die Synthese von kompletten Musikstücken. Es ist damit möglich den kompletten Prozess der Analyse abzubilden und jegliche Teilergebnisse künstlich zu erzeugen. Die Firma midiart[1] stellt freundlicherweise für die Forschung 14.203 professionell erzeugte Mididateien zur Verfügung.

4.1 Schlagzeugnotenerkennung

Experiment 1 ist für die Untersuchung der akustischen Ebene konzipiert. Es handelt sich um die Erkennung von Schlagzeugnoten aus polyphoner Musik. Abbildung 2 zeigt zwei Varianten. Die erste Variante (durchgezogenen Linien) bestimmt ohne Einschränkung der möglichen zeitlichen Positionen, welches Schlaginstrument wann gespielt wird. Das entspricht dem Ansatz (*Simultane Segmentierung und Klassifikation*). Variante zwei nutzt die metrischen Eigenschaften von Musiksignalen, nach der die Noten nicht an beliebigen Zeitpunkten im Signal auftreten können (gestrichelte Linie). Dies entspricht dem Ansatz *Segmentierung und Klassifikation*, da das Audiosignal in Rasterstücke aufgeteilt wird und diese dann einzeln klassifiziert werden.

[1] www.midi.de

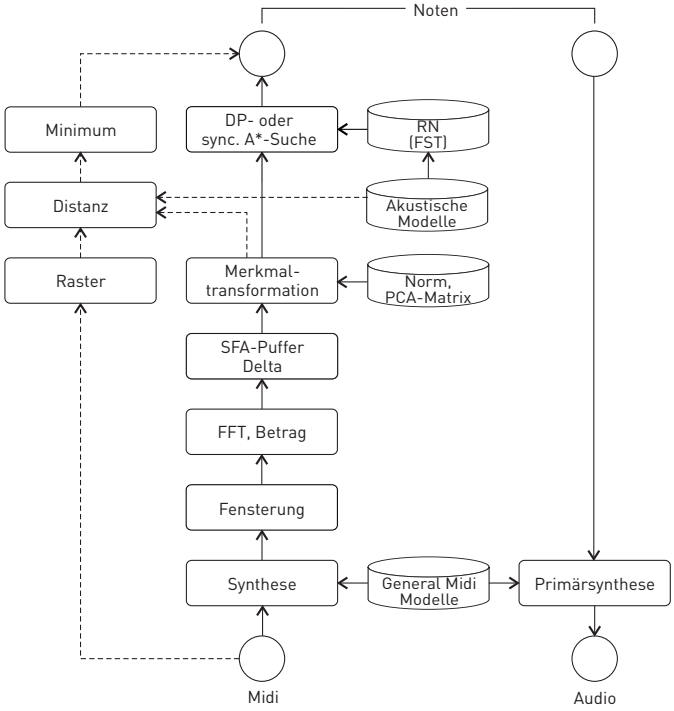

Abbildung 2 - Blockschaltbild für die Erkennung von Schlagzeugnoten

Um ausreichend Trainings- und Testdaten zu erhalten soll das Experiment zunächst mit synthetisierten Mididaten durchgeführt werden. Für die Synthese sind verschiedene Soundbibliotheken zu verwenden. Die Annotation der einzelnen Schlagzeugnoten im synthetisierten Audiomaterial kann direkt aus der Mididateien gewonnen werden. Hauptaugenmerk liegt auf dem Training der akustischen Modelle. Als Klassifikatoren werden Hidden Markov Modelle verwendet.

Variante eins lässt Wechsel innerhalb der Detektormodelle zwischen den beiden Ausgaben beliebig oft zu. Es entsteht ein Erkennungsnetzwerk für jedes mögliche Schlaginstrument. Statt der Verwendung von Detektormodellen ist auch ein Netzwerk aus Modellen für verschiedene Kombinationen von gleichzeitigen Schlagzeugnoten denkbar, dies führt bei Paulus allerdings zu schlechteren Ergebnissen [10]. Dies ist vor allem in der Größe der Trainingsdatenbank begründet, da auch alle Kombinationen von gleichzeitigen Schlaginstrumenten häufig genug auftreten müssen. Für die Verknüpfung der verschiedenen Modelle wäre ein Bigramm denkbar.

In Variante zwei werden die Merkmalvektorfolgen der einzelnen Rasterelemente mit jedem HMM verglichen und das am besten passende Modell ausgewählt. Kritisch sind Schlagzeugereignisse, welche nicht auf ein Rasterelement fallen und damit kurz vorher oder nachher beginnen.

4.2 Taktgrenzenerkennung mittels Symbolverarbeitung

Experiment 2 liegt auf der Ebene der Symbolverarbeitung. Es stellt die These auf, dass die Taktart und die Taktgrenzen eines Musikstückes anhand der Abfolge von Schlagzeugnoten ermittelt werden können. Dafür wird eine Grammatik entwickelt, welche die verschiedenen Schlagzeugmuster eines Taktes von Musiksignalen abbildet. Die technische Umsetzung erfolgt mit endlichen Automaten. Ein Beispielautomat für den Refrain des Titels *Last Resort von Pa-*

pa Roach ist in Abbildung 3 dargestellt. Dabei ist jeder Zustandsübergang mit dem Tripel aus Eingabesymbol (B: Base Drum, S: Snare Drum, H: Hi Hat, C: Cymbel, T: Tom Tom), Ausgabesymbol (4/4: Taktart) und Übergangswahrscheinlichkeit beschriftet.

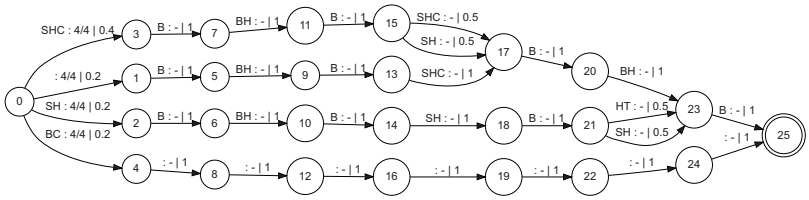

Abbildung 3 - Taktgrammatik des Refrains Last Resort von Papa Roach

Da dieses Experiment auf der Symbolebene arbeitet und die Schlagzeugnoten momentan noch nicht automatisiert aus Audiomaterial extrahiert werden können, dienen Midi Dateien als Ausgangspunkt der Untersuchungen. Eine Übersicht über das Experiment ist in Abbildung 4 zu sehen. Zunächst werden die Mididaten quantisiert und dann ternäre Dateien entfernt. Für unterschiedliche Phrasierungen sind unterschiedliche Grammatikmodelle nötig, da sich das Raster unterscheidet. Für die Verwendung eines Taktmodells werden besondere Anforderungen an das Raster gestellt. Es liegen zunächst Noten vor, welche in zeitlicher Abfolge gespielt werden. Die zeitliche Rasterschrittweite im Modell und Anfragesong müssen übereinstimmen.

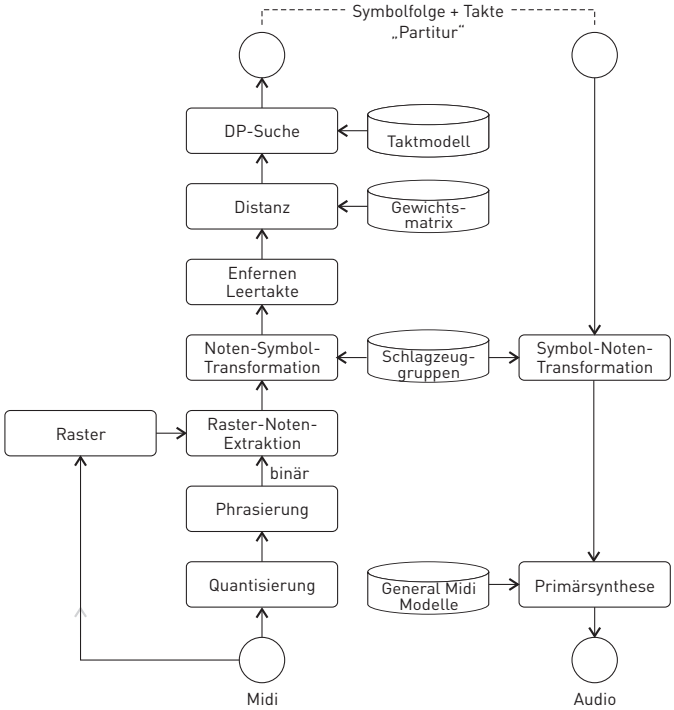

Abbildung 4 - Blockschaltbild für die Taktgrenzenerkennung auf der Grundlage von Schlagzeugnoten

Um der Anforderung gleicher zeitlicher Schrittweiten gerecht zu werden, ergeben sich zwei Möglichkeiten:

1. Modelle mit verschiedenen musikalischen Rasterschrittweiten eines Titels trainieren (4tel, 8tel,16tel ...). Wobei je ein Modell für gleiche zeitliche Rasterschrittweiten erforderlich ist. Bei der Anfrage wird dann das Modell mit der ähnlichsten zeitlichen Rasterschrittweite ausgewählt.

2. Sowohl Modell als auch Anfrage werden in eine genügend kleine zeitliche Rasterschrittweite transferiert. Diese ist dem menschlichen Gehör anzupassen.

Es soll die Variante zwei weiter verfolgt werden. Die zeitliche Auflösung des Rasters kann anhand von psychoakustischen Kriterien vorgenommen werden. Die Untergrenze der menschlichen Wahrnehmung für die Erkennung zweier getrennter Töne mit Reihenfolge liegt bei 20 ms[2]. Mit 100 ms wird die schnellste Abfolge musikalischer Ereignisse angegeben, welche noch perzeptuell durch den Menschen separierbar sind [8]. Das Raster wird deswegen von 320 - 640 BPM festgelegt (187,5 -93,8_ms). Jedes ermittelte schnellste Pulsraster wird mit einem Teiler dividiert bis das Raster dem vorgegebenen Bereich entspricht. Dieser Teiler darf sich zumindest nicht innerhalb eines Taktes ändern, da sonst die Taktgrenzen durch die Grammatik nicht mehr richtig bestimmt werden können. Eine Tempoänderung führt zu einer Veränderung der zeitlichen Auflösung des Rasters, aber nicht zwangsläufig zu einer Veränderung des Teilers. Die Anzahl der Zustände pro Takt hängen damit vom Musikstück ab.

Nachdem die Noten den Rasterelemente zugeordnet sind, erfolgt eine Zusammenfassung von gleichzeitig auftretenden Noten zu einem Symbol. Herzstück ist das Taktmodell, das mit möglichst vielen Takten trainiert wird und so die möglichen Schlagzeugmuster verschiedener Taktarten enthält. Da nicht alle Taktmuster der Erkennung zwangsweise auch im Training vorkommen, ist es nötig andere Symbole zuzulassen. Die Wahrscheinlichkeit aller Symbole zu einem Zeitpunkt wird durch die Gewichtsmatrix gekennzeichnet. Nach diesen Schritten der Vorverarbeitung kann eine unbekannte Symbolfolge eines Titels mithilfe des Taktmodells in Takte unterteilt werden. Die Ausgabe des Automaten entspricht den Taktarten der einzelnen Takte und beschriftet damit gleichzeitig die Taktgrenzen.

5 Ausblick

Als nächster Arbeitsschritt erfolgt die Umsetzung der vorgestellten Experimente. Die Verwendung von Mididaten in Experiment eins ermöglicht eine recht große annotierte Datenbasis. Der Übergang zum richtig Audiomaterial ist allerdings unerlässlich, da die klangliche Vielfalt und der Gesang bei der Verwendung der Midisynthese fehlt. Insbesondere die Verwendung eines Taktmodells als musikalisches Modell im Experiment zwei scheint ein neuer interessanter Ansatz zu sein. Erste Erfahrungen bei der Umsetzung dieses musikalischen Modells machen die Herausforderungen deutlich. Diese liegen unter anderem in der Möglichkeit auch ungesehene Taktmuster richtig einzuordnen. Eine Verwendung des UASR Systems für Musiksignale erscheint prinzipiell möglich.

Literatur

[1] DITTMAR, C. und C. UHLE: *Further Steps towards Drum Transcription of Polyphonic Music*. In: *Proc. of the Audio Engineering Society 116th Convention*, Berlin, Germany, 2004.

[2] GILLET, O. und G. RICHARD: *Automatic transcription of drum loops*. In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, Bd. 4, S. iv–269 – iv–272 vol.4, may 2004.

[2] http://musicog.ohio-state.edu/Music839D/Notes/timeline.html

[3] GOUYON, F. und S. DIXON: *A review of automatic rhythm description systems*. Computer Music Journal, 29(1):34–35, Mar 2005.

[4] GOUYON, F., P. HERRERA und P. CANO: *Pulse-Dependent Analyses of Percussive Music*. In: *AES, 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[5] HÜBLER, S. und R. HOFFMAN: *Evaluation of Onset Detection Algorithms in Popular Polyphonic Music on a Large Scale Database*. In: *Proc. of the Audio Engineering Society 130th Convention*, London, UK, 2011.

[6] HÜBLER, S. und R. HOFFMANN: *Comparing the Rhythmical Characteristics of Speech and Music – Theoretical and Practical Issues*. In: ESPOSITO, A., A. HUSSAIN, M. F. ZANUY und A. NIJHOLT (Hrsg.): *Proceedings of the 3rd COST 2102- EUCOGII International School on Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, Bd. 6456 d. Reihe *Lecture Notes in Computer Science (LNCS)*, S. 376–386. Springer-Verlag, Caserta, Italy, 2011.

[7] HOFFMANN, R., M. EICHNER und M. WOLFF: *Analysis of Verbal and Nonverbal Acoustic Signals with the Dresden UASR System*. In: ESPOSITO, A., M. FAUNDEZ-ZANUY, E. KELLER und M. MARINARO (Hrsg.): *Verbal and Nonverbal Communication Behaviours. COST Action 2102 International Workshop, Vietri sul Mare*, Bd. 4775 d. Reihe *Lecture Notes in Artificial Intelligence*, S. 200–218. Springer-Verlag, 2007.

[8] LONDON, J.: *Hearing in Time – Psychological Aspects of Musical Meter*. Oxford University Press, 2004.

[9] PAULUS, J.: *Signal Processing Methods for Drum Transcription and Music Structure Analysis*. Doktorarbeit, Tampere University of Technology, 2009.

[10] PAULUS, J. und A. KLAPURI: *Drum Sound Detection in Polyphonic Music with Hidden Markov Models*. EURASIP Journal on Advances in Signal Processing, S. doi:10.1155/2009/497292, 2009.

[11] UHLE, C. und J. HERRE: *Estimation of Tempo, Micro Time and Time Signature from Percussive Music*. In: *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03)*, London, UK, Sep. 2003.