Mixed-Type-Multi-Class Audio Klassifikation und Segmentation für die Inhaltserfassung von Broadcast Medien

Katja Madany

FB Informatik und Medien, Labor Computeranimation und Multimedia,
Beuth Hochschule für Technik Berlin
madany@beuth-hochschule.de

Abstract: Angestrebt ist eine Vorverarbeitung von Audiosignalen aus Mediadaten durch Klassifikation und Segmentation. Die in den Signalen vorkommende Sprache soll von den anderen Elementen wie Musik und Hintergrundgeräuschen getrennt werden, um eine ressourcen-optimierte Spracherkennung mit einer hohen Erkennungsrate zu gewährleisten. Mediadaten bestehen häufig aus Elementen, die nicht eindeutig einer Klasse zuzuordnen sind, sondern Mischformen aus den eben genannten Audioklassen darstellen. Die gängig für die Unterscheidung von Sprache und Musik eingesetzten akustischen Merkmale wurden daher in einer Mixed-Type-Multi-Class-Klassifikation angewendet, um darin ihre Aussagekraft zu überprüfen. Bei den verwendeten Merkmalen handelt es sich um statistische Kennwerte der zeitlichen und spektralen Struktur der Signale: Zero-Crossing-Rate, RMS-Energie, Spectral Centroid, Spectral Rolloff-Point, Spectral Flux und MFCCs. Während Sprache auch in den Mischformen mit Musik oder Hintergrundgeräuschen gut als diese klassifiziert werden konnte, stellt die Diskrimination der Hintergrundgeräusche ein Problem und eine Herausforderung an zukünftige Lösungsansätze dar.

1 Einleitung

Die automatische Klassifikation und Segmentation von Audiodaten stellt einen grundlegenden Schritt in der Vorverarbeitung der multimedialen Contenterfassung dar. Die Analyse und das Retrieval von Inhalten aus Audio- und Videodaten gewinnt zunehmend an Bedeutung, da nur so eine inhaltsbasierte Suche von Radio- und Fernsehbeiträgen, Videos und Podcasts (zusammen im Folgenden als Mediadaten bezeichnet) möglich ist. Die steigende Bedeutung der automatischen Inhaltserfassung wird deutlich, wenn man sich einige Zahlen vergegenwärtigt: Laut YouTube werden jeden Tag zwei Milliarden Videos abgerufen und hunderttausende von Videos hochgeladen und zwar in jeder Minute 24 Stunden Videomaterial.[1] Multimedia- Inhalte wie Podcasts oder Videos laufen mittlerweile on demand nahezu überall: auf der Homepage, dem Laptop, dem Video-iPod oder Handy-Farbdisplay. Heute gibt es allein im deutschsprachigen Raum unzählige kleinere wie größere Anbieter von solchen Mediainhalten, die ein gezieltes Auswählen bzw. Finden von Inhalten sehr schwierig machen. Eine solche Bereitstellung bzw. Speicherung von Wissen in multimedialer Form ist jedoch nur sinnvoll, sofern entsprechende Informationen auch wiedergefunden werden können. Dieser Aspekt wird von der ACM SIGMM[2] als eine große Herausfordung für die zukünftige Richtung der Multimediaforschung formuliert: „make capturing, storing, finding, and using digital media an everyday occurrence in our computing environment" [1].

Der etablierte Ansatz zur Erfassung und Verwaltung der Mediainhalte ist der Einsatz von Spracherkennungstechnologien, mit denen der Sprachinhalt erfasst wird. Die Inhalte werden dann mit Zusatzinformationen verknüpft, den sogenannten Meta-Daten. Für die automatische Spracherkennung (ASR: *Automatic Speech Recognition*) stellt das Auftreten von Sprache im

1 http://www.youtube.com/t/fact_sheet

2 Association for Computing Machinery (ACM) Special Interest Group on Multimedia (SIGMM), http://sigmm.utdallas.edu/

Zusammenhang mit anderen Audioklassen wie Hintergrundgeräuschen oder Musik ein großes Problem dar, das sich in einer starken Minderung der Erkennungsleistung der ASR ausdrückt. Mediadaten sind hinsichtlich ihrer akustischen Natur äußert vielfältig. Das Audiosignal der Video- und Audiodaten kann von Studioqualität sein oder wurde ursprünglich über ein Telefon oder einen anderen rauschhaften Kanal übertragen. Es zeichnet sich durch den erwähnten Wechsel von Sprachanteilen, Musik und Umgebungsgeräuschen, aber auch durch eine Mischung derer aus. Ein Übergang zwischen Segmenten kann graduell erfolgen wenn z.B. Hintergrundmusik mit steigender oder abschwingender Lautstärke auftritt, während der Wechsel zwischen verschiedenen Sprechern häufig abrupt erscheint oder Personen sogar gleichzeitig sprechen. Um in einem solchen Audiostream den Sprachinhalt transkribieren zu können, ist es notwendig den Datenstrom zunächst in homogene Abschnitte der auftretenden Audioklassen zu unterteilen, um vor allem eine Trennung der Sprache von den anderen Audioklassen zu vollziehen. Dies kann mittels Klassifikation und Segmentation des Audiostreams erreicht werden. Die Notwendigkeit für diese Entwicklungsarbeiten bedingt sich daraus, dass nur mit Hilfe dieser Trennung eine angestrebte Verarbeitungseffizienz erreicht werden kann, da nichtsprachliche Elemente nicht weiter analysiert werden müssen, d.h. die ASR nur auf Segmenten mit Sprache angewendet wird. Nach der Segmentation kann zudem das Signal von Störfaktoren bereinigt und somit eine Reduzierung des Signalrauschens vorgenommen werden. Inakkurate Segmentgrenzen sind eine entscheidende Ursache von Fehlern der ASR. Die Fehlerquote der Spracherkennungstechnologie kann also bereits durch die Segmentation signifikant reduziert werden, da diese die Grenzen nicht nach starren Regeln, sondern an tatsächlich auftretenden Segmentwechseln setzt. Viele praktische Applikationen können von einer solchen Audio-Vorverarbeitung profitieren. Diese sollte allerdings leicht zu integrieren sein und die Gesamverarbeitungszeit berücksichtigen. Einige Herausforderungen, die für die ASR zutreffen, gelten auch für die Klassifikation der Daten. So stellen die Mischformen und die Anzahl der verschiedenen Audioklassen, die in Radio- und Fernsehbeiträgen vorkommen, ein Mixed-Type-Multi-Class Klassifikationsproblem dar.

Im folgenden Abschnitt werden die bisher am häufigsten in der Literatur zur automatischen Klassifikation und Segmentation von Sprache und Musik verwendeten akustischen Merkmale beschrieben und verwandte Arbeiten vorgestellt. In Abschnitt 3 werden ausgewählte Verfahren getestet und evaluiert und in Abschnitt 4 diskutiert.

2 State of the Art

Die inhaltsbasierte Audioklassifikation ist vorrangig ein Mustererkennungsproblem, für das zwei Hauptaufgaben gelöst werden müssen. Zum einen müssen akustische Merkmale (Features) extrahiert werden, die effektiv die signifikantesten Eigenschaften der Audiosignale repräsentieren und dabei gegenüber unterschiedlichen Bedingungen robust sind, jedoch allgemein genug, um verschiedene Audioklassen zu beschreiben. Zum anderen ist ein geeigneter Kassifikator zu wählen, der anhand der extrahierten Merkmale die zu testenden Signalabschnitte den entsprechenden Klassen zuordnet. In den letzten Jahren wurden verschiedene Verfahren mittels unterschiedlicher akustischer Features und Klassifikationsmethoden vorgeschlagen. Ein Großteil davon befasst sich mit der Diskriminierung von Sprache und Musik [2-5] oder Sprache und Nicht-Sprache [6,7]. Nur wenige beschreiben die Segmentation mehrerer Klassen [8,9] oder sogar Mixed-Type [10]. Häufig wird eine Klassifikation auf Sprache und Musik beschränkt, da für diese Klassen deutliche Unterscheidungskriterien vorliegen.

2.1 Häufig verwendete akustische Features

Die ersten beiden der folgend aufgeführten Features werden auf Grundlage der zeitlichen Ebene berechnet, während die übrigen der spektralen Domäne entstammen und ihnen damit zumindest eine Fouriertransformation zugrunde liegt. Da das Spektrum von Musik in der Regel von Frame zu Frame deutlich weniger (d.h. langsamer) variiert, als das von Sprache,

sollten demzufolge die Delta-Werte der Features eine gute Diskriminationsperformanz bieten, weshalb sie häufig ebenfalls verwendet werden.

Zero-Crossing-Rate (ZCR): die ZCR repräsentiert die Anzahl von Nulldurchgängen und gibt damit die Rate der Vorzeichenwechsel eines Signals innerhalb eines vordefinierten Fensters an. Die ZCR ist eine einfache Messung des Frequenzgehaltes eines Frames. Kedem [11] bezeichnet sie als Maß der „*dominant frequency*". Saunders verwendet u.a. die Standardabweichung der ersten Ableitung und das dritte zentrale Moment des Mittelwerts (Schiefe) der ZCR [2]. Chen et al. 2006 verwenden die Varianz [10]. Lu et al. stellen fest, dass die Variation der ZCR ein größeres Disriminationspotential liefert, als deren exakter Wert. Sie definieren daher ein Merkmal namens High Zero Crossing Rate Ratio (HZCRR) als das Verhältnis der Anzahl der Frames, deren ZCR höher als das 1,5-fache der durchschnittlichen ZCR eines Fensters von einer Sekunde ist. [12] Sprache zeichnet sich dadurch aus, dass stimmhafte Laute (geringe ZCR) und stimmlose Laute (hohe ZCR) auf Silbenebene alternieren. Die Variation der ZCR ist daher für Sprache höher als für Musik. Entsprechendes gilt für die HZCRR.

Prozentsatz von Frames geringer Energie (PLEF: Percentage of Low Energy Frames [3] bzw. LSTER: Low Short Time Energy Ratio [12] bzw. SR: Silence Ratio [10]): Dieses Merkmal ist definiert als der Anteil von Frames, deren RMS Energie kleiner als 50% der durchschnittlichen RMS Energie ist, innerhalb eines vordefinieren Fensters von gewöhnlich einer Sekunde. Der Prozentsatz der Frames geringerer Energie ist gewöhnlich in Sprachsignalen höher, da darin häufiger als in Musik stille Momente vorkommen.

Spectral Centroid (SC): Der SC repräsentiert das Zentrum der Energieverteilung des Frequenzspektrums. Musiksignale erreichen gegenüber Sprache höhere Werte des SC aufgrund von hochfrequenten Rauschanteilen und Percussion Sounds.

Spectral Rolloff Point (SRP): Der Spectral Rolloff Point ist ähnlich dem SC ein Maß zur Beschreibung der Form des Spektrums. Ermittelt wird die Frequenz, bis zu der sich α% der Frequenzverteilung aufsummieren. Häufig wird der SRP bei α=95% gesetzt.

Spectral Flux (SF) oder Delta Spectrum Magnitude: Das Merkmal Spectral Flux dient dazu, das Ausmaß an Veränderungen (Fluktuation) des Spektrums zu messen. Es gibt die durchschnittliche Variation des Spektrums zwischen zwei benachbarten Frames an und wird errechnet als Euklidsche Norm der Delta Spektrum Magnitude.

Mel Frequency Cepstral Coefficients (MFCCs): Die auch in der Spracherkennung häufig verwendeten MFCCs werden auch für die Klassifikation eingesetzt. Kos et al. verwenden als Features die Variation der MFCCs [13]. Geht die Klassifikation und Segmentation einer ASR voraus, in der MFCCs verwendet werden, werden je nach Verfahren nur wenige weitere Verarbeitungsschritte benötigt und die Berechnungszeit des Gesamtsystems nur gering erhöht.

2.2 Verwandte Arbeiten

In diesem Abschnitt werden einige verwandte Arbeiten vorgestellt, in denen zumindest einige der oben genannten Merkmale in den vorgestellten Verfahren zum Einsatz kommen.

Saunders stellt einen Sprach-Musik-Klassifikator für die Segmentierung von Radiosendungen vor, der auf einfachen statistischen Merkmalen wie der ZCR und der Energiekontur beruht. Unter Einsatz eines multivariater Gaussian Classifier erreicht er eine korrekte Klassifikation von 98%. Saunders betont, dass durch den strikten Gebrauch von Merkmalen der Zeitdomäne die Notwendigkeit der Verwendung eines FFT-Prozessors vermieden wird, wodurch eine relativ geringe Berechnungszeit gewährleistet ist. Für dieses Verfahren ist es jedoch nötig Schwellenwerte zu setzen. [2]

Scheirer und Slaney vergleichen insgesamt 13 verschiedene Features in einem Musik-/Sprache-Unterscheidungssystem. Es handelt sich um acht Merkmale auf Frame Ebene: 4 Hz Modulationsenergie (ME), PLEF, Pulse Metric, ZCR, SC, SF, SRP, Cepstrum Resynthesis Residual Magnitude, sowie jeweils die Varianzen der fünf letztgenannten Merkmale innerhalb eines ein Sekunden langen Fensters. Am effektivsten hat sich eine Kombination der 4 Hz Modultionsenergie mit der Varianz des SF und dem Pulse Metric, das die Rhythmikkeit eines Fensters von 5 Sekunden länge ausdrückt, gezeigt. Unter Verwendung eines Klassifikators, der auf einem k-d-Baum beruht, wird eine Genauigkeit von 94,2% erreicht. Zur Vergleichbakeit mit der Arbeit von Saunders, berechnen die Autoren eine Klassifikation auf Basis von nicht-überlappenden 2,4 Sekunden Fenstern, in dem sie den Durschnitt der Ergebnisse der frame-by-frame Klassifikation bilden. Damit erreichen sie eine Klassifikationsgenauigkeit von 98,6%. [3]

Auch **Carey et al.** führen einen Vergleich verschiedener Merkmale durch, bei dem sich zeigt, dass auch die Tonhöhe (Pitch) und Amplitude (Filterbank Energie) gute Unterscheidungskriterien für Sprache und Musik sind. Neben Pitch und Amplitude zählen die ZCR und Cepstral Koeffizienten sowie jeweils die abgeleiteten Delta-Werte zu den Merkmalen, wobei jeweils ein Feature mit seinem Delta-Wert kombiniert und mit den anderen Merkmalspaaren verglichen wird. Die Klassifizierung wird mit Hilfe von Gaussian Mixture Models (GMM) durchgeführt. Das beste Ergebnis erzielen die Cepstral Koeffizienten und Delta Cepstral Koeffizienten mit einer equal error rate (EER) von 1,2%. Unter Verwendung der Amplitudenwerte kann eine EER von 1,7% erreicht werden, obwohl es das weit weniger komplexe Modell ist. Pitch mit einer ERR von 4% und ZCR mit einer ERR von 6% konnten diese Genauigkeit nicht erzielen. [14]

Ein System zur Segmentierung eines Audiosignals in die Klassen Musik, Sprache, Umgebungsgeräusch und Stille beschreiben **Lu et al.** unter Verwendung der folgenden sechs Features: HZCRR, LSTER, SF, LSP Distance, Band Periodicity (BP) und Noise Frame Ratio (NFR). Die LSP Distance oder Linear Spectral Pairs - Vector Quantisierung (LSP-VQ) stellt eine Analyse dar, bei der aus den Trainingsdaten ein *speech codebook* erstellt und dann der Abstand zwischen diesem und der LSP Kovarianz Matrix des zu testenden Audioclips verglichen wird. Die BP gibt die Periodizität in definierten Subbändern an, die aus einer Subband-Korreltionsanalyse hergeleitet wird. Das NFR ist das Verhältnis von Noise Frames zur Gesamtzahl der Frames eines Audioclips, wobei ein Frame als ein Noise Frame definiert ist, wenn das lokale Maximum der normalisierten Korrelationsfunktion unter einem festgelegten Schwellenwert liegt. Die Klassifikation erfolgt in zwei Schritten. Zunächst wird unter Verwendung der Merkmale HZCRR, LSTER und SF mittels eines k-Nearest-Neigbour (k-NN) Klassifikators und der Analyse per LSP-VQ eine Unterscheidung der Klassen Sprache und Nicht-Sprache durchgeführt. Im zweiten Schritt werden die Nicht-Sprach-Segmente nach einem regelbasierten Schema weiter in die Klassen Musik, Umgebungsgeräusche oder Stille unterteilt. Anhand der Festlegung von Schwellenwerten für die Short-Time Energy und die ZCR wird zunächst entschieden, ob es sich um Stille handelt. Für die restlichen Segmente erfolgt die Unterscheidung in Musik und Umgebungsgeräusche mit Hilfe von BP, SF und NFR. Hierbei wird entschieden, ob es sich um Umgebungsgeräusche handelt, alle übrigen Segmente müssen dann Musik sein. Auch in diesem Schritt kommen vordefinierte Schwellenwerte zum Einsatz. Bei der Auswahl der verwendeten Merkmale, gilt der SF einzig als in der Lage eine Unterscheidung von tonalen Umgebungsgeräuschen und Musik vorzunehmen, denn BP und NFR bei tonalen Umgebungsgeräuschen ähneln denen von Musik, nur der SF liegt über dem durch Experimente festgelegten Schwellenwert. Zur Segmentierung eines Audiostreams klassifizieren Lu et al. Segmente einer Länge von einer Sekunde. Um die Klassifikationsgenauigkeit zu erhöhen, werden mittels Smoothing unplausible Segmente entfernt. Mit diesem System erreichen die Autoren für die Zuordnung der Klassen Sprache,

Musik und Umgebungsgeräusch eine Klassifikationsgenauigkeit von 96,51%. Betrachten sie lediglich die Klassen Musik und Sprache erreichen sie eine Genauigkeit von 98,03%. Das System eigne sich auch für die Echzeitverarbeitung von Multimedia Applikationen. [12]

Ein System zur online Unterscheidung von Musik und Sprache im Broadcastbereich wird von **Kos et al.** erläutert. Die verwendeten Merkmale basieren auf der Varianz der MFCCs (MFCCV). Die Idee hinter dieser Merkmalskonstruktion liegt in der Energievariation in einem schmalen Frequenzsubband, wobei die Variation für Sprache größer ist als die für Musik. Verglichen werden vier Feature-Sets: 1.) ein Set bestehend aus fünf Mermalen (je den Varianzen von ZCR, PLEF, SF, SC, SRP), 2.) MFCCs [0-12], 3.) MFCCV und 4.) MFCCV in Kombination der fünf Merkmale des 1. Sets. Das beste Ergebnisse liefert das 4. Featur-Set (MFCCV + mix) mit 98,30%, aber auch die MFCCV allein erreicht 98,06%. Dabei wird jeweils die Sprache etwas schlechter als bei den herkömmlichen MFCCs erkannt, Musik jedoch besser. Damit ist die Gesamtklassifikation genauer als bei Verwendung der MFCC (96,77%). Ein Vorteil der MFCCV besteht in der leichten Implementierung und der geringen zusätzlichen Rechenzeit für ASR-Systeme, die MFCCs verwenden, da nur ein weiterer Berechnungsschritt nötig ist. [13]

Ein Vergleich der einzelnen Ansätze untereinander ist kaum aussagekräftig, da unterschiedliche Testdaten verwendet werden. Dennoch sind in Tabelle 1 die jeweils effektivsten Merkmale und Klassifikatoren der beschriebenen Verfahren nocheinmal zusammengefasst, um einen Überblick über die verschiedenen Umsetzungsmöglichkeiten und die dabei erzielte Klassifikationsgenauigkeit zu bieten.

Autor/en	**Features**	**Klassifikator**	**Klassifikations-genauigkeit**
Saunders	ZCR, Energiekontur	Multivariate Gaussian Classifier	98,00%
Scheirer & Slaney	4 Hz ME, Varianz SF, Pulse Metric	Spatial partitioning scheme based on k-d-trees	98,60%
Carey et al.	Cepstral Coefficient, Delta Cepstral Coeff.	GMM (EM-Algorithmus)	98,80%
Lu et al.	HZCRR, LSTER SF, LSP Distance BP, NFR	k-NN	98,03% (für Sprache, Musik & Umgebungsgeräusche 96,51%)
Kos et al.	MFCCV	GMM (Log-Likelihood-Estimation)	98,06%

Tabelle 1: Die effektivsten Features und Klassifikatoren aus den in den verwandten Arbeiten vorgenommenen Vergleiche sowie deren Klassifikationsgenauigkeit.

3 Test und Evaluation

Da die verschiedenen in der Literatur vorgestellten Verfahren sich entweder mit der Klassifikation von nur zwei Klassen (Musik und Sprache) befassen oder eine Datenbasis verwenden, die jeweils Klassen eines eindeutigen Typus enthalten und keine Mischformen wie sie in Radio- und Fernsehsendungen vorkommen, wurden in der vorliegenden Arbeit einige der beschreiebenen Features in einer Klassifikation mit mehreren Klassen (Muti-Class) und mit Klassen, die Mischformen (Mixed-Type) enthalten, getestet.

3.1 Datenbasis

Die Datenbasis wurde aus Fernsehbeiträgen gewonnen, die von verschiedenen öffentlichen und privaten Fernsehsendern ausgestrahlt wurden. Die Audiodaten wurden mit einer Samplerate von 16 kHz, 16 bit Auflösung, mono extrahiert. Jeweils 100 Audioclips der Länge 2-10 Sekunden wurden für die folgenden Klassen gesammelt: Sprache, Musik, Hintergrundgeräusche bzw. Umgebungsgeräusche (Noise), Stille, Sprache mit Musik (SpraMus), Sprache

mit Noise (SpraNoi). Die Daten enthalten sowohl männliche als auch weibliche Sprecher, mit Studio- und Telefonkanal-Aufnahmebedingungen. Möglichst viele verschiedene Musikgenres sind vertreten sowohl mit Gesang als auch ohne. Da die Audioclips aus Fernsehsendungen gesammelt wurden, gehören zu der Klasse Musik auch verschiedene Arten von Jingles. Die Hintergrundgeräusche waren unterschiedlicher Art, so wurden dieser Klasse z.B. sowohl Applaus und Jubel, Straßenlärm, das Schlagen von Türen oder Maschinenlärm als auch eher tonale Geräusche wie Vogelgesang zugeordnet.

3.2 Verwendete Feature-Sets

In der hier vorgestellten Arbeit wurden zunächst Klassifikationen auf Grundlage der im Abschnitt 2.1 beschriebenen gängigsten Features berechnet und verglichen. Die Feature-Extraktion wurde mit Hilfe des opensource Toolkits openSMILE [15] vorgenommen. Die berechneten Features werden auch als low-level Deskriptoren (LLDs) bezeichnet. Die LLDs wurden zunächst mittels eines Moving Average Filters der Festerlänge 3 geglättet. Um die Konturen der jeweiligen LLDs in einem Merkmalsvektor zu vereinen, wurden verschiedene statistische Funktionswerte berechnet. Für jedes Merkmal wurden zudem die Delta-Werte bestimmt. Diese sind jeweils zusätzlich Bestandteil der Feature-Sets. Die verwendeten Features sowie die darauf angewandten statistischen Funktionen sind in Tabelle 2 aufgelistet.

Name des Feature-Sets	Features	Funktionen
rms_zcr	RMS-Energie ZCR	Varianz Standardabweichung (stddev) Skewness Percencentile1.0 Percentile99.0
spectral	Spectral Centroid Spectral Rolloff Point (95%) Spectral Flux Spectral Entropy	Varianz Standardabweichung (stddev) Skewness arithmetisches Mittel (amean)
mfcc	MFCC [0-14]	

Tabelle 2: Name der Feature-Sets, darin verwendete Feature und statistischen Funktionen

Das Merkmal Spectral Entropy stellt ebenfalls ein Maß der Verteilung der Frequenzkomponenten dar und wird berechnet aus dem Power Spektrum. Die Funktionen Percentile1.0 und Percentile99.0 geben jeweils den minimalen bzw. maximalen Ausreißer robusten Wert der LLD-Kontur an, der von der 1%- bzw. 99%-Perzentilen repräsentiert wird. Die Feature-Sets wurden einzeln und in jeweiliger Kombination getestet. Für die Klassifikation wurde ein Naive-Bayes-Klassifikator gewählt, der bei schneller Berechnungszeit gute Klassifikationsraten liefert. Sie wurde unter Verwendung des Weka Toolkits [16] ausgeführt. Zur Überprüfung der Klassifikationsgenauigkeit wurde eine 10-fach Kreuzvalidierung vorgenommen.

3.3 Ergebnisse und Evaluation

Die Klassifikationsgenauigkeit der einzelnen Feature-Sets für die Klassifikation von Sprache und Musik einerseits und der Multi-Class-Klassifikation von Sprache, Musik, Noise und Stille andererseits ist in Tabelle 3 dargestellt. Angegeben ist jeweils die Gesamtklassifikationsrate aller Klassen. Aufgrund der geringen Datenbasis sind nur vorsichtige Schlüsse zu ziehen. Dennoch konnte für die Unterscheidung von Sprache und Musik eine vergleichbare Klassifikationsgenauigkeit erreicht werden wie sie in den oben vorgestellten Studien jeweils dargestellt wurde. Dies war zunächst als Basis für weitere Untersuchungen zu testen. Die verschiedenen Feature-Sets unterscheiden sich in der Klassifikationsrate nur gering. Vorzuziehen ist daher das Verfahren mit der geringsten Berechnungszeit, das in diesem Fall das *rms_zcr* Feature-Set ist.

	Genauigkeit der Klassifikation von	
Feature-Sets	**Sprache | Musik**	**Sprache | Musik | Noise | Stille**
rms_zcr	97,5%	85,5%
spectral	89,5%	74,0%
mfcc	96,5%	79,0%
Feature-Set-Kombination		
rms_zcr + spectral	97,0%	85,5%
rms_zcr + mfcc	97,5%	84,8%
rms_zcr + spectral + mfcc	97,0%	85,2%

Tabelle 3: Gesamtklassifikationsrate für die Unterscheidung von Sprache/Musik sowie Sprache/Musik/Noise/Stille (Multi-Class-Klassifikation) unter Verwendung der jeweiligen Feature-Sets.

Die Klassifikation von mehreren Klassen, in diesem Fall von Sprache, Musik, Noise und Stille, ergibt deutlich geringere Gesamterkennungsraten. Betrachtet man jedoch die Konfusionsmatrix der Klassifikation des *rms_zcr* Feature-Sets wird deutlich, dass es zu häufigen Konfusionen zwischen den Klassen Musik und Noise kommt. Das war zu erwarten, da aufgrund der Verschiedenartigkeit von Hintergrundgeräuschen die Audioclips der Klasse Noise viele musikähnliche Komponenten enthalten (z.B. Vogelgesang oder rhythmisches Klatschen). Sprache wird gut erkannt mit einer False Positive Rate (FP)=0.007, Precision=0.98 und Recall=0.97. Stille wurde bewußt als Klasse ausgewählt, um eine vorausgehende separate Trennung der stillen Segmente eines Audioclips und damit einen vorausgehenden Arbeitsschritt zu vermeiden. In weiteren Untersuchungen könnte überprüft werden, ob dieser vorangestellte Arbeitsschritt sinnvoll ist. Realisiert werden könnte dieser durch das Setzen von Schwellenwerten von Merkmalen wie der ZCR oder Energiewerten.

Auch für das Mixed-Type-Multi-Class-Klassifikationsproblem, bei dem die Klassen Sprache, Musik, Noise sowie die Mix-Klassen Sprache mit Musik (SpraMus) und Sprache mit Noise (SpraNoi) zuzuordnen waren, ergaben die Feature-Sets *rms_zcr*, *spectral* und *mfcc* ähnliche Klassifikationsraten, wobei das Set bestehend aus den 15 MFCCs und deren Delta-Werten mit 66,48% die höchste, wenn auch eine sehr schlechte Gesamtgenauigkeit erzielte (vergleichbar mit [10]). Betrachtet man jedoch die Konfusionsmatrix, ist zu erkennen, dass Sprache mit 85,6% als solche klassifiziert und ausschließlich mit den Klassen in denen Sprache mit Musik oder Noise vorkommt verwechselt wurde. Auch die Mix-Klassen werden mit Sprache, aber hauptsächlich untereinander verwechselt, was konsistent ist mit der Verwechslung der Klassen Musik und Noise, da die Sprache in den Mix-Klassen zusammen mit Musik oder Noise auftritt. Definiert man die Mix-Klassen als Sprache und führt erneut eine Klassifikation durch, wird die Klasse Sprache gegenüber Musik und Noise zu 92,88% korrekt klassifiziert (FP=0.055, Precision=0.966, Recall=0.929 für die Klasse Sprache).

4 Diskussion und Ausblick

Die in der vorliegenden Arbeit verwendeten Features sind nur bedingt geeignet, um Mediadaten zu klassifizieren. Es konnte gezeigt werden, dass Sprache auch in der Mixed-Type-Multi-Class-Klassifikation relativ gut erkannt wird. Verbesserungsbedarf besteht jedoch bei der Abgrenzung von Umgebungsgeräuschen zu den anderen Audioklassen, vor allem wenn sie musikähnlich sind. Das Hauptproblem liegt darin, dass Hintergrundgeräusche so unterschiedlicher akustischer Art sind, dass für sie anhand der bisher eingesetzten Features keine klaren Zuordnungskriterien festgesetzt werden können. Ebenfalls [10] konstatiert, dass die Klassifikation von Noise das Hauptproblem darstellt. Auch auf dynamischer Ebene, auf der sich die Variationen von Merkmalen zwischen Sprache und Musik unterscheiden, ist für Noise keine allumfassende Aussage zu treffen. Der Vergleich der Energiekonzentrationen

einzelner Frequenzbänder untereinander in Kombintion mit der Variation über die Zeit könnte auf Merkmalsebene zu besseren Ergebnissen führen. Ein weiterer Lösungsansatz könnte darin liegen die Klassifikation in mehreren Schritten auszuführen ähnlich wie in [12] und weitere Klassifikatoren zu testen. Diese Arbeit stellt erste Ergebnisse zu Beginn einer fortlaufenden Untersuchung dar. Für weitere Analysen muss die gesammelte Datenbasis stark erweitert werden.

Das vorrangige Ziel der Klassifikation und Segmentation für die geplante Applikation, die zur Medienbeobachtung eingesetzt wird, liegt in der Identifikation der Sprachsegmente eines Audiostreams, so dass nur diese zur Verringerung der Verarbeitungszeit an die Speech-to-Text-Engine weitergegeben werden müssen. So könnten die vorgestellten Feature-Sets in einer solchen Applikation zunächst für eine Unterscheidung von Sprache und Nicht-Sprache eingesetzt werden. Für zukünftig geplante Umsetzungen wäre das jedoch nicht ausreichend.

Acknowledgement
Die Ergebnisse entstanden im Rahmen des vom BMWi geförderten ZIM-Projektes „Entwicklung einer Plattformtechnologie zur automatisierten Verarbeitung und Verschlagwortung IP-basierter Multimediainhalte“, Förderkennzeichen KF2213301ED9.

Literatur

[1] Rowe, L.A., Jain, R.: ACM SIGMM Retreat Report on Future Directions in Multimedia Research. ACM Transactions on Multimedia Computing, Communications, and Application 1, 2005, pp. 3-13.
[2] Saunders, J.: Real-time Discrimination of Broadcast Speech/Music. In: Proc. of the IEEE ICASSP, 1996, pp. 993–996.
[3] Scheirer, E.; Slaney, M.: Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In: Proc. of ICASSP, Munich, 1997, pp. 1331–1334.
[4] Ajmera, J.; McCowan, I.; Bourlard, H.: Speech/Music Segmentation using Entropy and Dynamism Features in a HMM Classification Framework. In: Speech Communication, Vol. 40, Issue 3, 2003, pp. 351–363.
[5] Fu, Z.-H.; Wang, J.-F.; Xie, L.: Noise Robust Features for Speech/Music Discrimination in Real-Time Telecommunication. In: Proc. of the IEEE International Conference on Multimedia & Expo (ICME), New York, 2009, pp. 574–577.
[6] Maganti, H. K.; Motlicek, P.; Gatica-Perez, D.: Unsupervised Speech/Non-Speech Detection for Automatic Speech Recognition in Meeting Rooms. In: Proc. of ICASSP, Vol.4, 2007, pp. 1037–1040.
[7] Pigeon, S.; Verlinde, P.: Fusing Fast Algorithms to Achieve Efficient Speech Detection in FM Broadcast. In: Proc. of INTERSPEECH, Brighton, UK, 2009, pp. 2235–2238.
[8] Lu, L.; Zhang, H.-J.; Hao, J.: Content Analysis for Audio Classification and Segmentation. In: IEEE Transactions on Speech and Audio Processing, Vol.10, Issue 7, 2002, pp. 504–516.
[9] Lu, L.; Zhang, H.-J.; Li, S. Z.: Content-Based Audio Classification and Segmentation by Using Support Vector Machines. In: Multimedia Systems, Vol.8, Issue 6, 2003, pp. 482–492.
[10] Chen, L.; Gündüz, S.; Özsu, M. T.: Mixed Type Audio Classification with Support Vector Machine. In: Proc. of the IEEE International Conference on Multimedia & Expo (ICME), Toronto, 2006, pp. 781–784.
[11] Kedem, B.: Spectral analysis and discrimination by zero-crossings. Proc. IEEE, 74(11): 1477–1493, 1986.
[12] Lu, L.; Jiang, H.; Zhang, H.-J.: A Robust Audio Classification and Segmentation Method. In: Proc. of the ninth ACM international conference on Multimedia, New York, 2001, pp. 203–211.
[13] Kos, M.; Grasic, M.; Vlaj, D.; Kacic, Z.: On-Line Speech/Music Segmentation for Broadcast News Domain. In: Proc. of the International Conference on Systems, Signals and Image Processing (IWSSIP), 2009, pp. 1–4.
[14] Carey, M. J.; Parris, E. S.; Lloyd-Thomas, H.: A Comparison of Features for Speech, Music Discrimination. In: Proc. ICASSP, Vol.1, Issue 1, 1999, pp. 149–152.
[15] Eyben, F.; Wöllmer, M.; Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, Amsterdam, 2009.
[16] Hall, M.; Eibe, F.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Vol. 11, Issue 1, 2009.