Fußbodenidentifizierung mittels Schrittgeräuschen

Ansatz zur Sortierung einer Sound Library

Philipp Matalla[1], Silke Bögelein[1], Adam Kujawski[1], Jonas Oertel[1], Athanasios Lykartsis[1]
[1] *TU Berlin, FG Audiokommunikation, 10587 Berlin, Deutschland*
Email: {philipp.matalla, silke.boegelein, adam.kujawski, jonas.oertel}@stud.tu-berlin.de, alykartsis@win.tu-berlin.de

1. Einleitung

Der erste vertonte Film "Don Juan" im Jahre 1926 war die Geburtsstunde mehrerer heute etablierter Berufe wie z. B. dem des Sound Designers und des Foley Artists. Denkt man an Kassenschlager wie "Transformers" oder "Star Wars", ist deren Erfolg mitunter auch auf die außergewöhnlich guten Vertonungen zurückzuführen. Dazu gehört nicht nur, die Sprache der Protagonisten möglichst authentisch und synchron zur visuellen Ebene abzubilden. Mindestens ebenso wichtig ist die geschmackvolle Vertonung der Umgebungsgeräusche einer Szene und das designen spannender Special Effects. Dabei sind die Werkzeuge des Filmvertoners Sounddatenbanken aller Art, wobei die Kunst darin besteht, individuell passende Geräusche aus der Datenbank zu filtern. Die Suche nach den richtigen Geräuschen kann im kreativen Entstehungsprozess jedoch sehr viel Zeit in Anspruch nehmen, besonders dann, wenn bestimmte Eigenschaften der Audiodatei nicht aus dem Namen hervorgehen. Um diese Aufgabe zu erleichtern, ist eine automatische Sortierung unbekannter Samples in einzelne Kategorien von Vorteil. Die folgende Arbeit bezieht sich ausschließlich auf die Klassifizierung verschiedener Untergründe anhand von Schrittgeräuschen, als Beispiel der Sortierung einer Sounddatenbank mittels Methoden des Maschinellen Lernens.

Im Bereich der Schrittgeräuscherkennung gibt es bisher noch wenige Studien, wobei sich die Mehrzahl auf eine Schrittgeräuscherkennung zur Identifikation von Personen im Bereich der Wachsysteme beziehen [1] [2] [3]. Rodríguez et al. [1] verwenden hierfür Merkmale (Features) basierend auf der "Ground Reaction Force" in Verbindung mit Support Vector Machines (SVMs). Geiger et al. [3] extrahierten ihre Daten aus Videomaterialien. In einer auf akustischen Ereignissen basierenden Studie der Waseda Universität in Japan [2], wurde ein Mikrophon-Line-up zur Personenerkennung benutzt. Die akustischen Daten wurden mit Mel-Frequency Cepstral Coefficients modelliert und der jeweilige Personentyp mit Hilfe von Hidden Markov Models (HMMs) klassifiziert. Cai et al. [4] verwendeten SVMs, um drei unterschiedliche Personen anhand von Schrittgeräuschen erkennen zu können.

Die bisher genannten Studien identifizieren bereits erfolgreich Personen an Hand von Schrittgeräuschen. Es liegt deshalb nahe, die bereits gewonnenen Erkenntnisse als Ansatz zur Klassifizierung ähnlicher Anforderungen zu verwenden.

Eine Studie der Technischen Universität Berlin zusammen mit der Universität Pompeu Fabra in Barcelona [5], klassifizierten aus Audiodaten zum einen die Art des Untergrundes, wie auch des Schuhwerks. Hierfür verwendeten sie Gammatonfilterbanken, um diese zum einen im Frequenzbereich mittels Hilbert-Transformation und im Zeitbereich mittels Inner Haircell Model zu analysieren. Als Klassifikator wurden zum einen SVMs und zum anderen das HMM verwendet. Somit ergaben sich vier Ansätze (Zeitbereich - SVMs, Frequenzbereich - SVMs, Zeitbereich - HMMs, Frequenzbereich - HMMs) die sie gegenüberstellten und verglichen. Hierbei ergab sich für die Klassifizierung der Schuhsolen (getestet auf verschiedenen Untergründen) das beste Ergebnis von 93,6 % bzw. 95,3 % (je Untergrund) für die Kombination Frequenzbereich (Gammatonfilterbank + Hilbert-Transformation) mit SVMs. Bei der Klassifizierung des Bodens ergaben sich 96,2 % richtige Erkennung für die Kombination Zeitbereich (Gammatonfilterbank + Inner Haircell Model) und Support Vector Machine.

Eine Studie der Aichi Prefectural Universität in Japan [6] konzentrierte sich bei der Schritt-Klassifizierung auf gewöhnliche Audio-Features. Das Ziel der Klassifizierung bestand darin, eine von zehn Personen jeweils richtig zuzuordnen. Dabei verwendeten sie drei cepstrale Features (Reales Kurzzeit-Cepstrum, Linear Predictive Coding Cepstrum und Mel-Frequency-Scaled Cepstrum). Zusätzlich glichen sie die Spektren mittels Dynamic Time Warping an, um unterschiedliche Zeitverläufe auszugleichen. Hierbei erzielten sie je Feature zwischen 97 % und 98 % richtige Erkennung der Person.

Im Nachfolgenden wird der Ansatz zur Klassifizierung verschiedener Bodenuntergründe mittels Schrittgeräusche dargestellt.

2. Methoden

2.1. Features

Ein Schrittgeräusch kann als ein in seiner klanglichen Ausprägung vielfältiges und teils tonales, teils perkussives Signal charakterisiert werden. Gleiches gilt sowohl für Sprachsignale als auch für eine Reihe von Musiksignalen. Daher werden in dieser Arbeit zum einen aus der Spachanalyse bekannte Features zur Klassifizierung verschiedener Untergründe verwendet. Wie die Aichi Perfectural Univerität in Japan [6] bereits zeigte, lassen sich diese auch auf nicht sprachliche Audiodateien erfolgreich anwenden. Außerdem finden diese Features auch häufig Anwendung bei der Analyse von Musik. Zusätzlich werden diese Features in dieser Arbeit um das Linear Predictive Coding Cepstrum (LPCC) erweitert. Das LPCC wird ebenfalls in der Sprachsignalverarbeitung verwendet, um das Anregesignal der Glottis vom Vokaltraktfil-

ter zu trennen. Im Übertragenen Sinne würde dies eine Anregung des Bodens durch den Fuß bedeuten, der von der darauffolgenden Klangfilterung der unterschiedlichen Bodenmaterialien getrennt wird. Als Klassifizierungsverfahren dienen uns Support Vector Machines, da diese im Vergleich [5] zu anderen Klassifizierern die besten Ergebnisse erzielten.
Zunächst werden alle Samples vorverarbeitet (Mittelwert aller Kanäle, normalisiert) und hinsichtlich diverser zeitlicher, spektraler und cepstraler Features analysiert, die im Nachfolgenden aufgelistet sind:

- Time-Based Features
  - Rise[3]
  - Decay[3]
- Spectral Features
  - Fundamental Frequency[1]
  - High Frequency Content[1]
  - Spectral Centroid[1]
  - Spectral Flux[1]
  - Spectral Flatness[1]
  - Spectral Spread[1]
- Cepstral Features
  - Linear Predictive Coding Cepstrum[2]

Verwendete Features basierend auf: [1] nach [8], [2] basierend auf Toolbox matlab speech features [9], [3] nach [7].
Auf Grund der bereits vorliegenden Segmentierung der Audiodateien (meist unter einer Sekunde) wurden sämtliche Features (außgenommen der zeitlichen Features und dem Spectral Flux) ohne Unterteilung in Sampleblöcke ermittelt. Die Rise- und Decay-Time wurden an Hand von blockweisen quadratischen Mitteln (root mean square, RMS) berechnet, wobei 10-ms-Blöcke mit 5 ms Überlappung genutzt werden. Der Rise, sowie der Decay wurden an Hand der Maximas für eine Anstiegs- sowie Abstiegsspanne von -3 dB bis -10 dB ermittelt. Der Spectral Flux ist ein Maß für die zeitliche Veränderung der Form im Spektrum und bedarf somit einer Einteilung in Blöcke (Blocklänge = 256 Samples). Die Fundamental Frequency wurde mittels Autokorrelation berechnet. Der Spectral Centroid ist ein Kennwert für das Verhältnis von hohen und tiefen Anteilen im Signal. Die Spectral Flatness errechnet sich aus dem Verhältnis des geometrischen Mittelwerts des logarithmisch skalierten Betragsspektrums zum arithmetischen Mittelwert des linear skalierten Betragsspektrums. Das Feature High Friquency Content ist ein Maß für den Anteil hoher Frequenzen im Amplitudenspektrum. Die Hervorhebung hoher Frequenzen erfolgt mittels gewichtetem Leistungsspektrum. Der Spectral Spread eines Signals beschreibt die spektrale Verteilung um den spektralen Schwerpunkt. Dies erfolgt durch Berechnung des Mittelwertes, woraufhin sich die spektrale Energie auf einen spezifischen Bereich im Leistungsdichtespektrum konzentriert. Als cepstrales Feature werden Koeffizienten aus dem Linear Predictive Coding Cepstrum (LPCC) berechnet. Diese wurden jedoch in einer vereinfachten Form verwendet. So wurde auf die für Spachsignale übliche Vorverstärkung mittels Hochpassfilter, also auch auf eine Uterteilung in Blöcke durch ein Hanningfenster verzichtet. Zwölf Linear Predictive Coding Cepstrum Koeffizienten wurden berechnet.

2.2. Datenbank

Um auf eine ausreichend große Anzahl an Samples verschiedener Schrittgeräusche zurückgreifen zu können, wurde eine individuelle Sammlung aus verschiedenen Foleydatenbanken erstellt. Diese ergab sich unter anderem aus nicht öffentlichen, temporären Bereitstellungen (Tonfabrik Köln, Sony Pictures SFX Library, Hollywood Edge Complete, BBC Sound Effects Original Series, Digiffects Sound Effects Library). Ingesammt fasst die Datenbank eine große Zahl von Aufnahmen unterschiedlicher Personen und deren Schritten auf verschiedensten Untergründen mit variierendem Schuhwerk (z. B. Sneakers, Stöckelschuhe, Lederschuhe), sowie variierender Laufgeschwindigkeit (gehen, rennen, usw.). Wie in Foleydatenbanken üblich, waren mehrere Ausprägungen von Schrittgeräuschen auf einem Untergrund in einer Audiodatei zusammengefasst. In diesem Fall wurden die Schritte mittels DAW (Digital Audio Workstation) in einzelne Audiodateien getrennt und, insofern nicht im WAV Format und einer Aptastrate von 44,1 kHz vorhanden, entsprechend konvertiert. Die Länge der Dateien entspricht der Länge des Schrittsignals.
Die einzelnen Samples sind bewusst in unterschiedlichen Qualitäten hinsichtlich des Signalrauschabstandes und Resamplings komprimierter Datenformate gewählt worden. Die Verteilung der Samples war unbalanciert, wodurch einige Böden stärker bzw. schwächer repräsentiert wurden. Damit bestand potenziell die Gefahr beim Training der Daten einzelne Böden zu häufig zu klassifizieren. Insgesamt ergab sich ein Datensatz von 2381 Audiodateien, die in folgende Kategorien unterteilt werden konnten:

- Gras (89 Samples)
- Holz (293 Samples)
- Kiesel (340 Samples)
- Metall (289 Samples)
- Schlamm (325 Samples)
- Schnee (285 Samples)
- Stein (335 Samples)
- Teppich (124 Samples)
- Wasser (301 Samples)

2.3. Klassifizierung

Zur Klassifizierung wurde in dieser Arbeit die Machine-Learning-Methode der Support Vector Machines (SVMs) verwendet, um an die damit verbundenen erfolgreichen Resultate in [5] anzuknüpfen.
Die Entscheidung über die Zuordnung eines zu klassifizierenden Fußbodens findet über einen Featurevektor statt,

der die dafür nötigen extrahierten Informationen aus den ursprünglichen Audiodatein beinhaltet. Ziel der SVM ist die Trennung zweier Klassen durch eine Abgrenzung (Gerade, Ebene, Hyperebene), die so gewählt wird, dass der Abstand der Klassen zueinander maximal wird.
Eine Klasse im Merkmalraum anhand ihres Featurevektors von anderen zu unterscheiden, bedarf häufig einer Abgrenzung abweichend einer einfachen Geraden. In einer höherdimensionalen Abbildung des Merkmalraums mittels diverser Kernelfunktionen, lässt sich jedoch eine alternative Trennfläche, eine sogenannte Hyperebene, ermöglichen, die eine Klasse gegenüber den restlichen trennt (One-vs-All). Verwendet wurde in diesem Fall der Gaussian Radial Basis Function Kernel. Dieser kann zudem mit dem Parameter der Kernel Scale in seiner Empfindlichkeit variiert werden. Hier wurde ein kleiner Wert genutzt, der eine große Varianz zulässt, dadurch jedoch nur geringe systematische Fehler auslöst. Der genaue Wert wurde von Matlabs Classification Learner automatisch heuristisch optimiert. Zudem wurden alle Werte der Featurevektoren vor dem Training mittels z-Transformation standardisiert.
Das Trainingsset wurde mittels fünffacher Kreuzvalidierung gegeneinander getestet. Dabei wurde in jedem Durchlauf die betrachtete Klasse entgegen dem gesamten heterogenen Rest gestellt (One-vs-All).
In der Trainingsphase wurden verschiedene Kombinationen von Features verwendet, um herauszufinden, welchen Beitrag diese im Prozess der Klassifizierung hinsichtlich der Erkennungsleistung erbringen.
Nach erfolgreichem Training, wurde ein Testset mit jeweils 25 Samples pro Klasse gebildet und anschließend getestet.

3. Ergebnisse

Die unterschiedlichen Kombinationen der verwendeten Features ergaben, dass die besten Ergebnisse im Training und Test bei Verzicht auf zeitlichliche Features und das spektrale Feature Grundfrequenz erreicht werden konnten.
Wie in Abbildung 1 zu erkennen ist, lassen sich die unterschiedlichen Fußbodenklassen im Training mit einer hohen Zurordnungswahrscheinlichkeit trennen. Besonders positive Ergebnisse konnten für die Untergründe Holz, Schlamm und Stein erreicht werden, mit Werten von 95 % und mehr. Ausschließlich bei Schrittgeräuschen im Wasser findet sich eine Klassifizierung unter 80 %. Bei Fehlklassifizierung werden hauptsächlich die Untergründe Teppich und Schnee ermittelt. Dennoch ergibt sich eine hohe Gesamtgenauigkeit bei der Klassifizierung im Training mit 88,7 %.
Die Ergebnisse des Testsdurchlaufs sind in Abbildung 2 dargestellt. Dieser Ergab bei sechs von neun Klassen eine korrekte Zuordnung von über 92 % der Samples. Die Kategorie Holzboden wurde sogar zu 100 % richtig erkannt. Werte unter 92 % ergaben sich für die Klassen Teppich, Stein und Kiesel mit jeweils 88 %. Wobei die Kategorien Stein und Kiesel jeweils untereinander vertauscht wurden. Insgesamt konnte somit eine Gesamtgenauigkeit von 92,6 % in der Klassifizierung der verschiedenen Fußböden

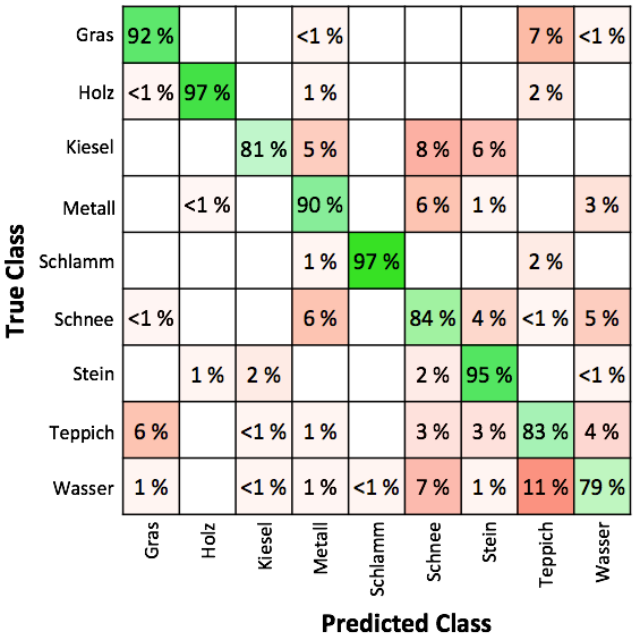

Abbildung 1: Ergebnisse des Trainings.

erzielt werden.

4. Diskussion

Die Klassifizierung der Fußböden liefert sehr gute Ergebnisse. Dennoch gibt es auffällige Unterschiede bei der Erkennung einzelner Klassen. So ist es zum Beispiel nachvollziehbar, dass eine Verwechselung von Kiesel- und Steinboden aufgrund der ähnlichen Materialeigenschaften durchaus erwartbar ist. Allerdings gilt die Verwechslung Steinboden oder Teppichboden mit den Schrittgeräuschen auf Wasser als nicht plausibel. Es fehlt daher eventuell ein weiteres Feature, welches eine eindeutigere Trennung der Klassen im Training und damit eine Verbesserung der Zuordnung im Test möglich macht. Unter den hier verwendeten Features sind bereits große Unterschiede in der Eignung zur Fußbodenklassifizierung erkennbar. So sind beispielsweise zeitliche Features aus dem finalen Training ausgeschlossen wurden, da sie die eindeutige Trennung einiger Klassen verhinderten. Das Linear Predictive Coding Cepstrum und einige spektrale Features wie Spectral Flatness und Centroid begünstigten hingegen eine korrekte Unterscheidung der Böden enorm.
Eine weitere Erklärung für mögliche Fehlklassifizierungen einiger Untergründe könnte die ungenügende Anzahl an Stichproben sein. Zwar ließ sich der Untergrund Teppich mit einer hohen Genauigkeit eindeutig klassifizieren, dennoch ist das Ergebnis gegenüber den anderen Klassen als schlecht einzuordnen. Mit 124 Samples war es die am zweitgeringsten repräsentierte Klasse in diesem Versuch. Allerdings kann diese Vermutung unter Beachtung der Stichprobengröße des Untergrundes Gras (89 Samples) nicht weiter belegt werden.
Zusätzlich kann ergänzt werden, dass das Trainingsset aus jeweils 25 Samples pro Klasse relativ gering ist. Die Fehlkategorisierung eines einzelnen Samples entspricht demnach bereits einer Fehlerrate von 4 %. Um detailier-

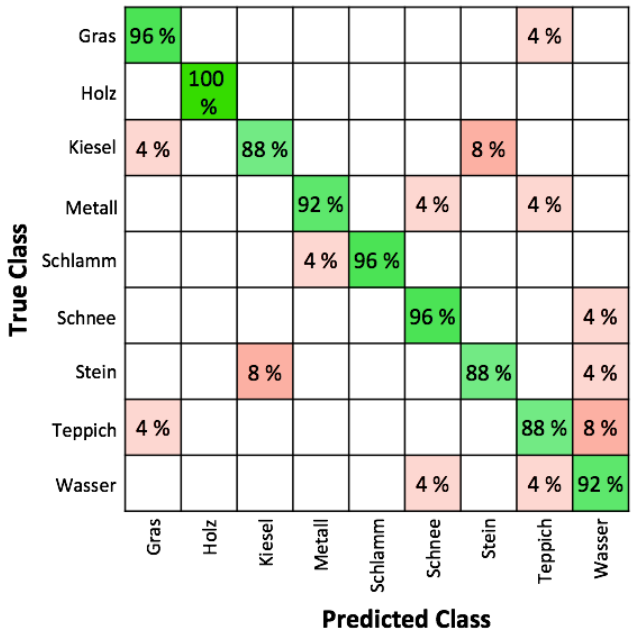

Abbildung 2: Ergebnisse des Tests.

tere Aussagen über Fehlzuordnungen geben zu können müssen weitere, wie auch größere Trainingsdatensätze verwendet werden.
Zusammenfassend lässt sich bereits eine sehr gute Klassifizierung verschiedener Fußböden erzielen, jedoch müssen die Schwächen, sowie Erweiterungen des Systems im Weiteren detailiert analysiert werden.
Der hier verwendete Ansatz gilt ohne weiteres als echtzeitfähiges System und verfügt, gemessen an seinem Klassifizierungsumfang, über vergleichbare Klassifizierungsergebnisse wie [5].

5. Ausblick

Bevor eine automatische Sortierung innerhalb einer Sound Library realisiert werden kann, ist noch eine Reihe von weiteren Untersuchungen und Experimenten nötig.
So könnte z. B. der hier dargelegte Ansatz der Klassifizierung durch weitere Untergründe oder andere Geräusche (Türgeräusche, Motorengeräusche) erweitert werden.
Des Weiteren wäre eine weitere Unterteilung der Schrittgeräusche nach unterschiedlichem Schuhwerk und der Schrittgeschwindigkeit von Vorteil.
Eine vollständige Feature Selection könnte zu weiteren Verbesserungen der Klassifizierung führen, ebenso wie Versuche mit weiteren Klassifizierungsalgorithmen des maschinellen Lernens.

Literatur

[1] Vera Rubén Rodríguez, Lewis, Richard P., Mason, John S.D., and Nicholas W.D. Evans. Footstep recognition for a smart home environment. In International Journal of Smart Home, 2008.

[2] Kazuhiro Nakadai, Fujii, Yuta, and Shigeki Sugano. Footstep detection and classification using distributed microphones. In Image Analysis for Multimedia Interactive Services (WIAMIS) 14th International Workshop on Image Analysis for Multimedia Interactive Services, 2013.

[3] Jürgen T. Geiger, Kneißl, Maximilian, Schuller, Björn, and Gerhard Rigoll. Acoustic gait-based person identification using hidden markov models. In Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge andWorkshop, pages 35–30, 2014.

[4] Francisco Cai, Philipson, David, and Salik Syed. A step-by-step approach to footstep detection. In Social and Information Network Analysis Stanford University, 2010.

[5] Robert Anniés, Martínez Hernández, Elena, Adiloglu, Kamil, Purwins, Hendrik, and Klaus Obermayer. Classification schemes for step sounds based on gammatone-filters. In Proceedings of the IEEE Conference on Web Intelligence, 2007.

[6] Akitoshi Itai and Hiroshi Yasukara. Footstep classfication using simple speech recognition technique. In IEEE International Symposium on Circuits and Systems, page 3237, 2008.

[7] Wilfried Weißgerber. Elektrotechnik für Ingenieure - Formelsammlung. Springer Vieweg, 4. Auflage, 2013.

[8] Alexander Lerch. An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. JohnWiley & Sons, 2012.

[9] James Lyons. Matlab speech features (matlabtoolbox). MIT, 2013. URL: `https://github.com/jameslyons/matlab_speech_features`