FORSCHUNG

AUTOMATISCHE SPRACH-ERKENNUNG FÜR GESANG

DR. ANNA KRUSPE

Die Suche nach Musikstücken auf Basis ihrer gesungenen Textinhalte kann im Privat-, Broadcast- und Forschungsbereich eine große Unterstützung sein. Mit ihrer Hilfe können z. B. Lieder in bestimmten Sprachen oder mit bestimmten gesungenen Themen gefunden werden. Trotz der zunehmenden Verbreitung von Spracherkennungstechnologien in den letzten Jahren gibt es bisher aber fast noch keine Möglichkeiten, diese auch für Gesang einzusetzen. In diesem Artikel wird aufgezeigt, was dabei die größten Hindernisse sind und wie diese sich lösen lassen. Vorgestellt werden Ansätze für die Erkennung der Landessprache von Musikstücken, für die Schlagwortsuche, für die automatische Zuordnung von bekanntem Text zu Gesangsaufnahmen sowie für die Suche von Musikstücken anhand von gesungenen Ausschnitten. Den Ausgangspunkt dafür bildet die Erkennung von Lauten (Phonemen), welche auf Technologien des maschinellen Lernens beruht.

Searching for musical pieces on the basis of their sung lyrics can be helpful for private users, broadcasters, and researchers. Such search methods allow finding songs in certain languages or about certain topics. Despite the widespread use of speech recognition technologies in recent years, there are few advances in their application to singing. This article summarizes the largest obstacles to this and how to overcome them. Presented applications include language recognition in musical pieces, keyword spotting, automatic lyrics-to-audio alignment, and retrieval of songs by sung excerpts. These technologies are based on the recognition of phonemes, which is implemented via machine learning methods.

Motivation

Seit einigen Jahren ist die automatische Spracherkennung im Alltag vieler Menschen angekommen. Siri, Alexa und „Okay Google" kennt mittlerweile fast jeder. Gleichzeitig hat der Bereich der Musikdistribution eine massive Wandlung erfahren. Viele Musikhörer nutzen tagtäglich digitale Streamingdienste wie Spotify oder Apple Music. Solche Portale bieten oft auch neue Möglichkeiten, Musik zu entdecken, die oft auf Methoden aus der Künstlichen Intelligenz oder allgemeiner der Data Science basieren. Da stellt sich die Frage: Warum wird Spracherkennung eigentlich nicht mehr für Gesang eingesetzt? Darum geht es in diesem Artikel, der die Dissertation, „Application of automatic speech recognition technologies to singing" zusammenfasst [1]. Die Nutzungsmöglichkeiten sind vielfältig: Im Privatbereich könnten Musikhörer leichter Stücke mit bestimmten gesungenen Schlagworten, Themen oder Sprachen finden. So können beispielsweise bekannte Stücke anhand des Textes wiedergefunden werden oder automatisch Playlists für verschiedene Anlässe erzeugt werden. Die Suche nach Liedern in bestimmten Sprachen ist z. B. auch für den Sprachlernbereich sehr interessant. Eine andere Anwendungsmöglichkeit liegt in der automatischen Zuordnung von Gesangstexten zu Audioaufnahmen (das sogenannte Alignment), um damit z. B. automatisch Karaoketracks zu erzeugen.

Aber auch im Broadcast- und Distributionsbereich sind solche Ansätze von großem Nutzen. Analog zur Suche bei Privatnutzern können Redakteure oder Werbetreibende hier beispielsweise schnell passende Musik für Beiträge finden, indem sie in einem vorhandenen Katalog nach Sprachen und Schlagworten suchen. Dabei können sich wertvolle Kombinationen mit anderen Entwicklungen auf dem Gebiet des Music Information Retrieval ergeben, wie z. B. die Suche von Musikstücken mit bestimmten Stimmungen, Instrumenten, Stimmen oder Stilen.

Möglicherweise stellt sich nun die Frage: Wenn Spracherkennungssysteme mittlerweile so weit entwickelt sind wie von Alexa und Co. aufgezeigt, warum können sie nicht einfach für Gesang eingesetzt werden? Die Gründe dafür liegen in den starken Signalunterschieden zwischen Sprache und Gesang. Im Gegensatz zur Sprache weisen Gesangsaufnahmen viel größere Schwankungen in der Tonhöhe auf, mit denen herkömmliche Spracherkennungssysteme meist nicht umgehen können. Ähnlich verhält es sich mit den verschiedenen Aussprachedauern von Lauten. In der Sprache sind diese relativ konstant, während Sänger durch die musikalischen Rahmenbedingungen Laute oftmals sehr viel länger ziehen oder abkürzen. Auch werden an vielen Stellen Laute anders ausgesprochen oder an die sängerischen Gegebenheiten angepasst. Ein häufiges Beispiel hierfür ist die Verschiebung von längeren „i"-Lauten hin zu „e", da dies angenehmer klingt. Weitere Aspekte sind das Vibrato bei Sängerstimmen und Unterschiede in Themen und Vokabular gegenüber Sprache. Ein großes Problem liegt weiterhin im instrumentalen Hintergrund der meisten Musikstücke, der für die Spracherkennung einen signifikanten Störfaktor bildet.

In der vorgestellten Arbeit werden Möglichkeiten aufgezeigt, um Spracherkennungsalgorithmen dennoch für Gesang einzusetzen. Prinzipiell werden dafür zwei Ansatzpunkte beleuchtet: Die Algorithmen können entweder robuster gegenüber den erwähnten Gesangscharakteristika gemacht werden, oder diese Charakteristika können gezielt ausgenutzt werden, um die Erkennung zu verbessern. Aufgezeigt wird dies an fünf Themen: Der Erkennung von Lauten

(Phonemen) in Gesang, der Erkennung von Sprachen, der Schlagwortsuche sowie dem Text-zu-Audio-Alignment und, damit verwandt, dem Auffinden (Retrieval) von Liedern.

Technischer Hintergrund

In der Spracherkennung werden seit langem Techniken des maschinellen Lernens, insbesondere künstliche neuronale Netze, eingesetzt. Diese Technologien erleben in den letzten Jahren unter dem Namen Deep Learning einen neuen Boom und werden mittlerweile auch in vielen anderen Feldern mit großen Erfolg genutzt. Grundlegend werden dabei Modelle mit Daten trainiert, für die das gewünschte Ergebnis bereits bekannt ist, z. B. indem es manuell annotiert wurde. Nach diesem Training kann das Modell dann für neue Daten eingesetzt werden, um Vorhersagen zu treffen. Diese Prozesse sind in Abbildung 1 dargestellt. Für die Erkennung der (Landes-)Sprache von Gesang heißt das beispielsweise, dass zunächst einmal ein Datensatz aus mehreren Tausend Musikstücken von Nöten ist, für die die Sprache bekannt ist. Auf dieser Basis wird dann ein Modell erzeugt, das für neue Musikstücke ebenfalls eine Zuordnung treffen kann, ohne dass ein Nutzer noch etwas dazu beitragen muss. Für ein einzelnes Musikstück wäre das übertrieben; die Vorteile werden aber klar, wenn Sammlungen mehrerer Tausend Stücke vorliegen, für die nun auf einen Schlag alle Sprachen bestimmt werden können.

Wie funktioniert ein solches Modell nun? Ein neuronales Netz ist im Grunde ein mathematisches Modell, das aus einer Vielzahl künstlicher Neuronen besteht, die in Schichten angeordnet sind. Jedes Neuron beinhaltet dabei eine Gewichtsmatrix, mit der seine Inputs multipliziert werden. Dann werden sie aufaddiert und mit einer nichtlinearen Aktivierungsfunktion belegt, z. B. einer Sigmoidfunktion oder einer Gleichrichtefunktion. Eine Visualisierung ist in Abbildung 2 gezeigt. Ein einzelnes solches Neuron ist noch nicht sehr leistungsfähig. Die eigentlichen Fähigkeiten eines neuronalen Netzes ergeben sich erst aus der Verknüpfung vieler solcher Neuronen über mehrere Schichten hinweg, die (im einfachsten Fall) vorwärts aufeinander aufbauen wie in Abbildung 3 dargestellt. Die erste und die letzte Schicht haben dabei die Funktion eines Inputs bzw. Outputs, diejenigen dazwischen werden als verdeckte Schichten bezeichnet. Neuronale Netze gelten als „tief" (siehe Deep Learning), wenn sie drei oder mehr verdeckte Schichten besitzen. Ein Netz zur Sprachenerkennung würde beispielsweise an seiner vordersten Schicht eine Signalrepräsentation des Musikstückes anliegen haben und an seiner letzten die Wahrscheinlichkeiten der verschiedenen Sprachen für dieses Musikstück ausgeben, wie ebenfalls in Abbildung 3 gezeigt.

Wie fließen nun die Informationen aus bekannten Daten ein? Die oben erwähnten Gewichtsmatrizen der Neuronen im Modell sind nicht von vornherein bekannt. Sie werden zunächst mit zufälligen Werten initialisiert. Dann erfolgt ein Training, in dem immer wieder verglichen wird, inwieweit das Netz von den Inputs auf die bekannten Outputs schließen kann. Anhand des dabei berechneten Fehlers werden die Gewichtsmatrizen schrittweise rückwärts angepasst. Sobald das Netz zufriedenstellend Vorhersagen für die bekannten Daten treffen kann, werden die Gewichte festgesetzt. Neue Musikstücke können nun in das Modell hineingegeben werden, um anhand der trainierten Gewichte ihre Wahrscheinlichkeiten für verschiedene Sprachen vorherzusagen.

Phonemerkennung in Gesang

Die Grundlage für die Spracherkennung bildet meist die Erkennung einzelner Laute, der so genannten Phoneme. Beispiele für solche Phoneme wären verschiedene kurze und lange Vokale, Plosive (z. B. „p", „t", „k"), Nasale (z. B. „m", „n", „ng"), oder Frikative (z. B. hartes und weiches „s", „sch", „f"). Diese Phoneme sind abhängig von der jeweiligen Sprache und werden von Linguisten definiert, wobei es keine einheitliche Festlegung gibt. Für die englische Sprache werden in den meisten Ansätzen der Spracherkennung um die 40 Phoneme angenommen, wobei teilweise noch einmal in Subphoneme unterteilt wird.

Für die Spracherkennung werden nach dem oben beschriebenen Prinzip Modelle trainiert *(acoustic modeling)*. Der Ausgangspunkt dafür ist meist Audiomaterial, für das händisch Phoneme annotiert wurden. Teilweise reichen auch Wortannotationen aus. Das Audiomaterial wird oft vor dem Training vorverarbeitet, indem eine Merkmalsextraktion darauf durchgeführt wird, z. B. von *mel-frequency cepstral coefficients (MFCC)*. Die erzeugten Modelle können dann wiederum Phonemwahrscheinlichkeiten in neuem Audiomaterial bestimmen. Diese können in einem Phonemposteriorgramm

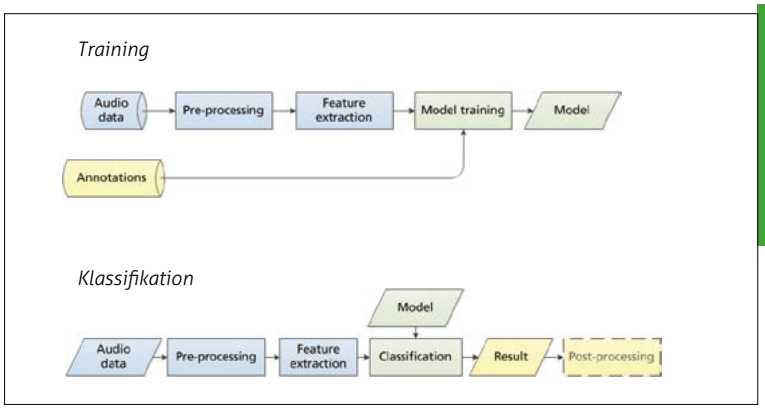

Abbildung 1. Trainings- und Klassifikationsprozesse

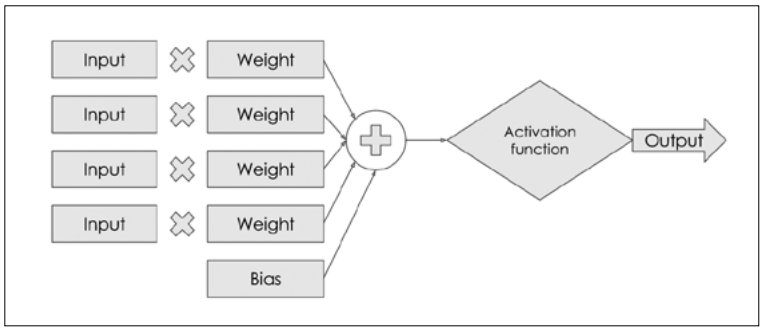

Abbildung 2. Ein einzelnes Neuron eines neuronalen Netzes [26]

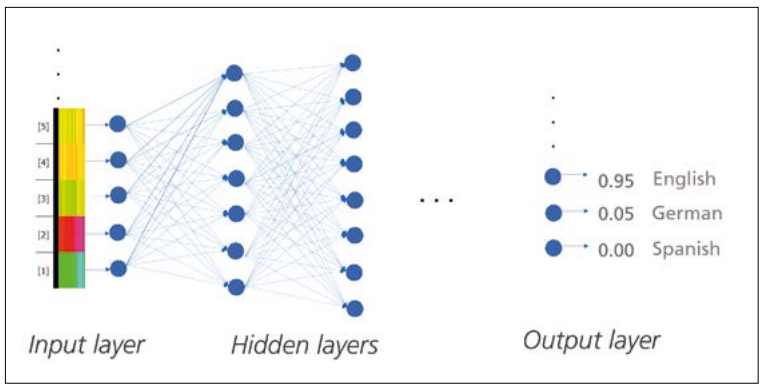

Abbildung 3. Beispiel für ein neuronales Netz zur Bestimmung der Sprache

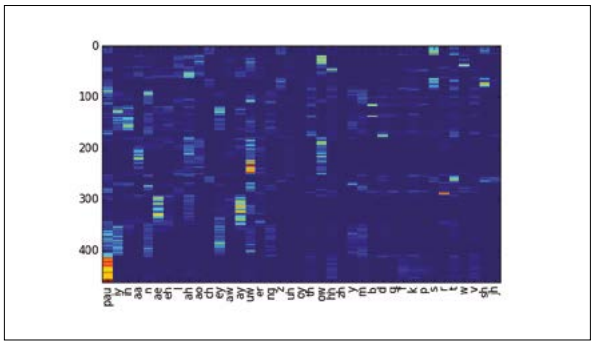

Abbildung 4. Beispiel für Phonemposteriorgramm, das die Wahrscheinlichkeiten von Phonemen (x-Achse) über die Zeit (in Frames, y-Achse) darstellt

dargestellt werden, siehe z. B. Abbildung 4. Daraus ergibt sich noch nicht eindeutig eine Folge von Worten und Sätzen, sodass meist ein zweiter Schritt nachgeschaltet wird. Dieser modelliert die Wahrscheinlichkeiten, mit denen bestimmte Phoneme und Worte aufeinander folgen, um aus dem Ergebnis des *acoustic modeling* plausible Sequenzen zu erzeugen *(language modeling)*. In neueren Ansätzen werden oft beide Schritte in einem Modell vereint, so dass nicht mehr explizit Phoneme vorgegeben oder erkannt werden müssen. Zudem wird auch die Merkmalsextraktion vom Modell mitgelernt, so dass das Audiomaterial nur in Form von Waveforms oder Spektrogrammen bereitgestellt werden muss. Das Modell lernt dann direkt eine Abbildung von Audiomaterial auf Textsequenzen *(end-to-end modeling)*.

Die ersten Ansätze zur Phonemerkennung für Gesang beruhen auf Modellen, die auf Sprachmaterial nach diesem Prinzip trainiert wurden [2,3,4,5]. Wie leicht vorstellbar ist, ist die phonem- oder wortweise Annotation von Audiomaterial sehr aufwändig und setzt einiges Expertenwissen voraus. Wie oben erwähnt, sind für das Training von neuronalen Netzen sehr viele solcher Daten vonnöten. Für Sprache existieren schon länger solche Datensätze (z. B. vom Linguistic Data Consortium[1]). Für Gesang hingegen gibt es diese nicht. Zusätzlich zum Ressourcenbedarf für die Erzeugung eines solchen Datensatzes kommt hier auch noch die Rechteproblematik im Musikbereich hinzu. Daher werden für das Training oft Sprachdatensätze genutzt. Ein sehr bekannter davon heißt *TIMIT*. Er besteht aus 4620 Sprachbeispielen *(utterances)* von wenigen Sekunden Länge in englischer Sprache, die von verschiedenen Sprechern realisiert werden. Erste Tests mit Modellen, die nur auf diesen Sprachdaten trainiert wurden, lieferten jedoch keine guten Ergebnisse beim Test auf Gesangsdaten.

Ein Trick, um der Charakteristik von Gesangsaufnahmen näher zu kommen, besteht darin, vorhandene Sprachdaten künstlich gesangsähnlicher zu machen. Wie oben erwähnt, sind im Gesang die Tonhöhen und die Lautdauern sehr viel variabler als in Sprache. Mit Signalverarbeitungsalgorithmen kann Pitch Shifting, also eine Tonhöhenveränderung, auf Sprachaufnahmen angewandt werden, und Vokale können zeitlich gestreckt oder gestaucht werden. Auch Vibrato kann hinzugefügt werden. Solche Ansätze, die vorhandenes Trainingsmaterial künstlich erweitern, werden auch als *data augmentation* bezeichnet und finden vor allem in der Computervision viel Anwendung. Modelle, die mit so verändertem Material trainiert werden, liefern tatsächlich bessere Ergebnisse beim Test auf Gesangsdaten [6].

Wie erwähnt wäre der ideale Trainingsdatensatz einer, der aus realen Gesangsaufnahmen mit Phonem- oder Wortannotationen besteht. Datensätze aus Gesangsaufnahmen ohne Annotationen existieren immerhin, so zum Beispiel *DAMP* von der Stanford University[2]. *DAMP* enthält 34.000 Aufnahmen, die durch die Karaoke-App *Smule Sing!* erzeugt wurden. Zu hören sind hier Amateur-Sänger. Die Audioqualität ist in den meisten Fällen recht gut und die Aufnahmen beinhalten nur den puren Gesang ohne Background-Musik, was für das Training der Modelle sehr hilfreich ist. Annotiert ist dieser Datensatz nicht, aber die App beinhaltet nur eine bestimmte Anzahl von Liedern, für die die Texte über die *Smule Sing!*-Website verfügbar sind.

Hier kann nun ein weiterer Trick zum Einsatz kommen. Wenn sowohl Audioaufnahmen als auch Texte vorhanden sind, kann mit so genannten *alignment*-Methoden zugeordnet werden, an welcher Stelle im Audiomaterial welches Wort oder welcher Laut zu hören ist. Solche Methoden basieren ebenfalls auf der Phonemerkennung (mehr dazu später), allerdings ist dieses Problem viel leichter zu lösen als wenn vorher gar nichts über den Text bekannt ist. Nutzt man nun Modelle, die auf den bekannten Sprachdaten trainiert sind (z. B. auf *TIMIT)*, funktioniert das *alignment* schon recht gut. Hierdurch können also automatisch Phonemannotationen für den *DAMP*-Datensatz erzeugt werden.

Mit diesem neuen Datensatz aus Gesangsaufnahmen und den zugehörigen Annotationen können nun neue Modelle wie oben beschrieben trainiert werden. Diese sind sehr viel besser für die Phonemerkennung in Gesang geeignet, da die Trainingsdaten ebenso Gesang beinhalten und die Modelle dadurch robust gegenüber den beschriebenen Charakteristika werden. Das notwendige *alignment* funktioniert zwar nicht immer perfekt, da es sich jedoch um einen so großen Datensatz handelt, gleichen sich Fehler auf Dauer im Modelltraining wieder aus. Experimente belegen, dass auf diese Art trainierte Modelle sehr viel besser für Gesang funktionieren als solche, die auf reiner Sprache trainiert wurden, und auch besser als solche, für die wie oben beschrieben Sprachaufnahmen gesangsähnlicher gemacht wurden [7].

Automatische Sprachenerkennung

Eine Forschungsrichtung mit vielen praktischen Anwendungen ist die automatische Erkennung der Sprache von gesungenen Aufnahmen, also z. B. Englisch, Deutsch oder Spanisch. Hier gibt es wiederum schon viele Ansätze für Sprachmaterial, aber fast keine für Gesang [8,9,10,11]. Grundsätzlich lässt sich das Problem auf zwei Arten lösen: Entweder durch die Erstellung von Modellen, die direkte Zusammenhänge zwischen Audiomaterial und Sprache herstellen, oder über den Umweg der Phonemerkennung.

In ersterem Fall werden Modelle wie oben beschrieben trainiert, um direkt von Gesangsmaterial auf die Sprache abzubilden. Die Schwierigkeit liegt hier in der Merkmalsextraktion. Die Sprache lässt sich im Allgemeinen nicht aus einzelnen, kurzen Frames wie in einem Spektrogram vorhanden ableiten, sondern ergibt sich aus den Zusammenhängen über längere Zeiträume. Zudem sind viele Komponenten des Signals über Sprachen hinweg gleich oder sehr ähnlich und tragen daher nichts zur Erkennung bei. Ein beliebter Ansatz aus der Spracherkennung für solche Fälle ist die so genannte *i-vector*-Extraktion. Sie wird z. B. auch für die Erkennung von Sprechern eingesetzt. Es handelt sich dabei um eine Dimensionsreduktion, für die die universellen Hintergrundkomponenten entfernt werden und das Restsignal in

1) https://catalog.ldc.upenn.edu/
2) https://ccrma.stanford.edu/damp/

eine Variabilitätsmatrix und ihre Aktivierungen zerlegt wird. Die Aktivierungen, welche unabhängig von der Signaldauer immer die gleiche Dimensionalität haben, heißen *i-vectors* und bilden die Inputs für das Modelltraining. Analog zu den Ergebnissen auf Sprachmaterial zeigt sich, dass dieser Ansatz auch für Gesang sehr gut funktioniert [12].

Der zweite Ansatz beruht darauf, dass der Vorrat von Phonemen und ihre Auftrittswahrscheinlichkeiten sich von Sprache zu Sprache unterscheiden. Auf Basis der zuvor beschriebenen Phonemerkennung können Statistiken gebildet werden und daraus Rückschlüsse auf die Sprache gezogen werden. Einzelne gesprochene Sätze sind für die statistische Berechnung nicht ausreichend, auf einem ganzen Musikstück funktioniert sie jedoch [13].

Schlagwortsuche in Gesang

Mit Hilfe der Phonemerkennung lassen sich auch bestimmte Schlagworte in gesungenem Text auffinden. Wie oben beschrieben ist das Ergebnis der Phonemerkennung ein Posteriorgramm wie in Abbildung 4 dargestellt. Dabei ist erkennbar, dass dieses Ergebnis aufgrund des schwierigen Audiomaterials relativ verrauscht ist, so dass es nicht direkt auf eine Sequenz von Phonemen heruntergebrochen werden kann. Die meisten bestehenden Ansätze nutzen daher Zusatzinformationen wie vorhandene Texte [14,15] oder Aufnahmen der gesuchten Schlagworte oder Phoneme [16,17].

Für die direkte Erkennung von Schlagworten auf Basis der Posteriorgramme ohne Zusatzinformationen kann ein weiterer Modelltyp genutzt werden: Das Hidden Markov-Modell (HMM). Diese Methode erlaubt die Modellierung von (meist zeitlichen) Sequenzen von Zuständen mit bestimmten Übergangswahrscheinlichkeiten. Im Falle von Schlagworten besteht ein solches HMM aus zwei Teilen: Einerseits einem Unter-HMM, das die Phonemsequenz des gesuchten Wortes *(keyword)* modelliert, andererseits einem Unter-HMM, das alle Phoneme beinhaltet *(filler)*. Dargestellt ist dies in Abbildung 5. Der Parameter β bestimmt dabei die Übergangswahrscheinlichkeit vom *filler* zum *keyword*, also die Empfindlichkeit des Modells. Als Input für ein solches Modell dienen wieder die Ergebnisse der Phonemerkennung. Dieser Ansatz ist prinzipiell in der Lage, Schlagworte zu finden, funktioniert jedoch noch nicht zuverlässig für Gesangsmaterial [18]. Dafür gibt es zwei Gründe. Zum einen ist, wie erwähnt, die Phonemerkennung als Ausgangspunkt noch relativ instabil. Zum anderen sind im Gesang meist recht kurze Schlagworte mit wenigen Phonemen von Interesse, wie z. B. „love", „time" oder „heart". Wie zu erwarten zeigt sich eine Abhängigkeit der Erkennungsgenauigkeit von der Anzahl der Phoneme (längere Worte bieten mehr Anhaltspunkte zur Erkennung) und es ist zu vermuten, dass die Suche für längere Phrasen statt Worte besser funktionieren würde. Auch in Ansätzen, die für Sprache eingesetzt werden, werden üblicherweise erst Worte mit fünf oder mehr Phonemen zuverlässig gefunden.

Eine Erweiterungsmöglichkeit besteht darin, dem Algorithmus zusätzlich Informationen über die möglichen Dauern von Phonemen mitzugeben. Wie beschrieben schwanken diese zwar im Gesang mehr als in Sprache, allerdings trifft das vor allem auf Vokale zu. Konsonanten sind in ihrer Dauer eingeschränkt. Dieses Wissen kann, modelliert mit Gammaverteilungen, in die Erkennung integriert werden und führt zu etwas besseren Ergebnissen [19].

Lyrics-to-audio-Alignment und Song Retrieval

Eine weitere Nutzung der Ergebnisse der Phonemerkennung besteht in ihrer Anwendung für die zeitliche Zuordnung

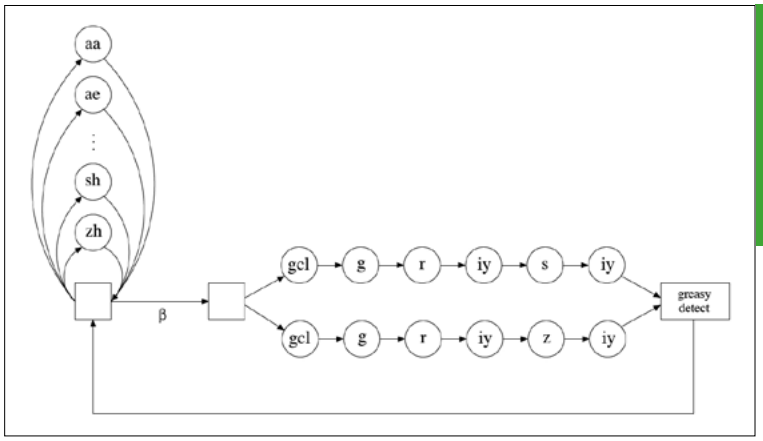

Abbildung 5: Hidden Markov-Modell mit keyword-Zweig für das Wort „greasy" (rechts) und filler-Zweig (links). Die Übergangswahrscheinlichkeit ist β. [27]

von bekanntem Text zu Audioaufnahmen, das erwähnte *alignment*. Da oft der Text schon bekannt ist, nur nicht die Zeitpunkte der einzelnen Worte, kann hiermit beispielsweise die Schlagwortsuche vereinfacht werden oder eine dynamische Suche ermöglicht werden. Auch die synchrone Wiedergabe von Musik und Text wird so ermöglicht, z. B. für Karaokeanwendungen.

Dafür stellt sich die Frage, wie die zeitbasierten Ergebnisse der Phonemerkennung zu den bekannten Wort- und damit Phonemsequenzen des Textes zugeordnet werden können. Es gibt dafür ebenfalls HMM-basierte Algorithmen sowie solche, die die Methode des Dynamic Time Warping nutzen [20]. Durch das Rauschen, das in den Posteriorgrammen vorhanden ist (siehe z. B. Abbildung 4) schlagen diese jedoch des öfteren fehl. Eine neue Methode beruht darauf, solche Schwierigkeiten in diesen Posteriorgrammen zu umgehen und bekannte Charakteristika auszunutzen. Dazu werden zunächst die Posteriorgramme entlang der Zeitachse geglättet und auf Basis der Wahrscheinlichkeiten und Dauern Phoneme für jeden Zeitframe ausgewählt. Dann werden diese aufgrund von bekannten Konfusionen des Modells gefiltert, so dass am Ende eine Phonemsequenz plausibler Länge übrig bleibt. Diese kann dann mit dem bekannten Text in Übereinstimmung gebracht werden. Die beschriebene Methode schnitt 2017 am besten bei der *MIREX*-Challenge ab [21].

Diese Nachverarbeitung der Posteriorgramme kann jedoch auch für einen anderen Zweck genutzt werden: Es können damit auch längere Textstellen wiedergefunden werden *(Retrieval)*. Die erzeugte Phonemsequenzen wird dabei mit einer Datenbank aus vielen Liedtexten abgeglichen und die wahrscheinlichste Stelle zurückgeliefert. Dieser Ansatz ermöglicht es, Lieder aufgrund des gesungenen Textes wiederzufinden, egal ob vom Originalsänger oder von einem Nutzer. Damit realisiert er einen ähnlichen Anwendungsfall wie beispielsweise die Software *Shazam*[3], jedoch auf Basis des Textes anstelle der Melodien und Harmonien [22, 23].

Die Ergebnisse des *alignment* können auch die Grundlage für viele weitere Anwendungen bilden. Eine davon ist die automatische Suche nach Schimpfworten in Musik. Für die Ausstrahlung werden diese vor allem im englischsprachigen Raum oft ausgeblendet oder mit einem „Bleep" vertuscht. Dies wird meist händisch umgesetzt. Die vorgestellte zeit-

3) https://www.shazam.com/

Quelle: Bild: rbb/ARD/ZDF/Claudius Pflug

Für die Dissertation „Application of automatic speech recognition technologies to singing" wurde Dr. Anna Kruspe mit dem ARD/ZDF-Förderpreis Frauen+Medientechnologie ausgezeichnet. Zudem erhielt sie für ihre Arbeit den ICT Dissertation Award des Fraunhofer-Verbunds IUK-Technologie.

liche Zuordnung von Texten zu Musik ermöglicht auch hier eine automatische Umsetzung, so dass die entsprechenden Gesangsstellen direkt z. B. per Kanalsubtraktion entfernt werden können [24].

Zusammenfassung & Zukünftige Arbeiten

In diesem Artikel wurden Möglichkeiten aufgezeigt, um Methoden der automatischen Spracherkennung für Gesang einzusetzen. Dies bietet vielfältige Anwendungsmöglichkeiten im Privat-, Broadcast- und Distributionsbereich. Die Schwierigkeit liegt darin, dass Gesang und Sprache sehr unterschiedliche Signalcharakteristika aufweisen und daher auf Sprache entwickelte Methoden nicht direkt einsetzbar sind.

Die Grundlage bildet die Erkennung von Phonemen. Diese wird mit Methoden des maschinellen Lernens, genauer mit neuronalen Netzen, umgesetzt. Dafür sind große Trainingsdatensätze nötig. Eine vielversprechende Möglichkeit zur Erzeugung solcher Datensätze besteht darin, bekannte Songtexte mit vortrainierten Modellen zeitlich den entsprechenden Audioaufnahmen zuzuordnen *(alignment)*. Die Phonemerkennung kann dann beispielsweise eingesetzt werden für die Erkennung der Sprache (wobei dort auch ein direktes Modelltraining ohne Phonemerkennung möglich ist), für die Schlagwortsuche, für ein besseres *alignment* sowie für das Auffinden von Liedern anhand ihres gesungenen Textes.

Die Forschung in diesem Bereich ist noch nicht abgeschlossen [25]. An vielen Stellen bieten sich noch Verbesserungsmöglichkeiten. Ein besondere Schwierigkeit entsteht u.a. durch Hintergrundmusik und im speziellen durch instrumentale Soli. Hier könnte Abhilfe geschaffen werden, indem die beschriebenen Methoden mit einer vorgeschalteten Gesangsdetektion oder Systemen zur Quellentrennung kombiniert werden. Aber auch verschiedene Aussprachen und Akzente erschweren die Erkennung. Hinzu kommt, dass die einzelnen Ansätze auf Gesang mehr oder weniger schwerwiegende Fehler machen. Hier wären robustere Modelle sowie verlässlichere Trainingsdaten notwendig. Mit einem nachgeschalteten *language modeling* könnten die Methoden an spezifische Anwendungsfälle angepasst werden und auch dadurch stabilisiert werden.

Wie oben erwähnt gibt es mittlerweile auch Modelle, die nicht mehr explizit Phoneme erkennen, sondern direkt auf Text abbilden können. Es wäre interessant, dieses Prinzip auch auf die Erkennung im Gesang anzuwenden. Umgekehrt können aus diesem Anwendungsfall hervorgegangene Algorithmen auch genutzt werden, um die Erkennung auf Sprachmaterial robuster zu machen.

Ein besonders spannender weiterer Schritt wäre die Kombination mit Algorithmen zur automatischen Erkennung von Melodien und Harmonien, um eine gegenseitige Verbesserung zu bewirken. Abschließend lässt sich sagen, dass es Anwendungsseite viele interessante Einsatzmöglichkeiten gibt. Manche davon wurden hier schon aufgezeigt, andere ergeben sich sicher zukünftig.

Bild: Fraunhofer IDMT

DR. ANNA KRUSPE

ist kommissarische Gruppenleiterin am Institut für Datenwissenschaften Datenmanagement und -analyse des Deutschen Zentrums für Luft und Raumfahrt (DLR)

www.dlr.de

Referenzen

[1] A. M. Kruspe, "Application of automatic speech recognition technologies to singing". Dissertation, Technische Universität Ilmenau, 2018.

[2] M. Gruhne, K. Schmidt, and C. Dittmar, "Phoneme recognition in popular music," in International Society for Music Information Retrieval Conference (ISMIR), 2007.

[3] A. Mesaros and T. Virtanen, "Adaptation of a speech recognizer for singing voice," in European Signal Processing Conference (EUSIPCO), 2009.

[4] H. Fujihara, M. Goto, and H. G. Okuno, "A novel framework for recognizing phonemes of singing voice in polyphonic music," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2009.

[5] J. K. Hansen, "Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients," in Sound and Music Computing Conference (SMC), 2012.

[6] A. M. Kruspe, "Training phoneme models for singing with "songified" speech data", in 16th International Society for Music

Information Retrieval Conference (ISMIR), Malaga, Spain, 2015.
[7] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing", in 17th International Society for Music Information Retrieval Conference (ISMIR), New York, NY, USA, 2016.
[8] J. Schwenninger, R. Brueckner, D. Willett, and M. E. Hennecke, "Language Identification in Vocal Music," in International Society for Music Information Retrieval Conference (ISMIR), 2006.
[9] W.-H. Tsai and H.-M. Wang, "Towards Automatic Identification Of Singing Language In Popular Music Recordings," in International Society for Music Information Retrieval Conference (ISMIR), 2004.
[10] M. Mehrabani and J. H. L. Hansen, "Language identification for singing," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.
[11] A. M. Kruspe, Jakob Abesser, Christian Dittmar, "A GMM approach to singing language identification", in Proc. of the AES Conference on Semantic Audio, London, UK, 2014.
[12] A. M. Kruspe, "Improving singing language identification through i-vector extraction", in Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany, 2014.
[13] A. M. Kruspe, "Phonotactic Language Identification for Singing", in Interspeech, San Francisco, CA, USA, 2016.
[14] H. Fujihara, M. Goto, and J. Ogata, "Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics," in International Society for Music Information Retrieval Conference (ISMIR), 2008.
[15] T. Nakano and M. Goto, "LyricListPlayer: A consecutive-query-by-playback interface for retrieving similar word sequences from different song lyrics," in Sound and Music Computing Conference (SMC), 2016.
[16] C. Dittmar, P. Mercado, H. Grossmann, and E. Cano, "Towards lyrics spotting in the SyncGlobal project," in 3rd International Workshop on Cognitive Information Processing (CIP), 2012.
[17] G. Dzhambazov, S. Sentürk, and X. Serra, "Searching lyrical phrases in a-capella Turkish Makam recordings," in International Society for Music Information Retrieval Conference (ISMIR), 2015.
[18] A. M. Kruspe, "Keyword spotting in a-capella singing", in 15th International Society for Music Information Retrieval Conference (ISMIR), Taipei, Taiwan, 2014.
[19] A. M. Kruspe, "Keyword spotting in singing with duration-modeled HMMs", in European Signal Processing Conference (EUSIPCO), Nice, France, 2015.
[20] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its applications," in Multimodal Music Processing, M. Müller, M. Goto, and M. Schedl, Eds., vol. 3. Dagstuhl Follow-Ups, 2012.
[21] A. M. Kruspe, "Lyrics alignment using HMMs, posteriorgram-based DTW, and phoneme-based Levenshtein alignment", in 18th International Society for Music Information Retrieval Conference (ISMIR) (MIREX submission), Suzhou, China, 2017.
[22] A. M. Kruspe, "Retrieval of textual song lyrics from sung inputs", in Interspeech, San Francisco, CA, USA, 2016.
[23] A. M. Kruspe, M. Goto, "Retrieval of song lyrics from sung queries", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, 2018.
[24] A. M. Kruspe, "Automatic B**** Detection", in 17th International Society for Music Information Retrieval Conference (ISMIR) (Late-breaking demo), New York, NY, USA, 2016.
[25] E. J. Humphrey, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Kruspe, A. Kumar, S. Reddy, P. Seetharaman, L. Yang, "Signal Processing for Singing Voice Analysis: Significance, Applications, and Methods". IEEE Signal Processing Magazine 36(1):82-94, 2019.
[26] R. Campbell, "Demystifying Deep Neural Nets". 2017. https://medium.com/@RosieCampbell/demystifying-deep-neural-nets-efb726eae941.
[27] A. Jansen and P. Niyogi, "An experimental evaluation of keyword-filler Hidden Markov Models," Tech. Rep., 2009.

DORTMUNDER FORSCHER ENTWICKELN 5G-CAMPUSNETZPLANER

Die 5G-Funktechnologie ermöglicht neben den öffentlichen, flächendeckenden Mobilfunknetzen auch den Betrieb lokaler Funknetze. Forscher der TU Dortmund haben jetzt einen Campusnetzplaner entwickelt, mit dem Unternehmen in wenigen intuitiven Schritten die voraussichtlich anfallende Frequenzgebühr ermitteln können, die für das unternehmensspezifische Campusnetz fällig wird.

Nach der Versteigerung der 5G-Funkfrequenzen an die Mobilfunknetzbetreiber im Frühjahr 2019 hat die Bundesnetzagentur im November 2019 das Antragsverfahren für zusätzliche lokale 5G-Campusnetze im Frequenzbereich 3.700 bis 3.800 MHz gestartet. Diese Funkfrequenzen bilden die Grundlage für 5G-Anwendungen, beispielsweise in den Bereichen Industrie 4.0, Smart Farming oder Smart City. Die Gebühren für die Frequenzen berechnen sich nach einer Formel, die vor allem die Größe des abgedeckten Gebiets, die Bandbreite und die Laufzeit der Frequenz berücksichtigt.

Den Campusnetzplaner hat die TU Dortmund als Konsortialpartner des Competence Center 5G.NRW entwickelt. Der Campusnetzplaner ist kostenfrei unter https://5G.nrw/campusnetzplaner/ verfügbar. Dort lassen sich die Daten für den 5G-Antrag an die Bundesnetzagentur passend aufbereiten und können im Anschluss direkt in das Antragsformular eingetragen werden. Das Competence Center 5G.NRW wird vom Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen gefördert.

Den Campusnetzplaner kann die TU Dortmund auch für ihren eigenen Campus einsetzen. Foto: TU Dortmund

https://5g.nrw www.tu-dortmund.de