ulm university universität ulm

Universität Ulm | 89069 Ulm | Germany

Fakultät für Ingenieurwissenschaften und Informatik
Institut für Neuroinformatik

Deep feature-based speech emotion recognition for smart affective services

Seminararbeit an der Universität Ulm

Vorgelegt von:
Salih Bedelce
salih.bedelce@uni-ulm.de

Gutachter:
Prof. Dr. Friedhelm Schwenker

Betreuer:
Prof. Dr. Friedhelm Schwenker

2021

Inhaltsverzeichnis

Zusammenfassung

Diese kleine Einleitung soll dem Nutzer helfen selbst die eigene Arbeit mit LaTeX zu schreiben.
Sie enthält zu den wichtigsten Themen Beispiele.

[3] [7] [9] [2]

1 Einleitung

Die Sprache des Menschen ist die natürlichste Art und Weise um miteinander zu kommunizieren [2]. Durch die technische Entwicklung in den letzten Jahren kommen immer mehr Interaktionen zwischen Mensch und Maschine durch die Sprache zustande. Die Bezeichnung dafür sind die sogenannten intelligent personal assistants (IPAs) [3] wie zum Beispiel Amazon Alexa, Apple Siri und Google Assistant. Google Home, Amazon Echo und Apple HomePod sind Home-Assistant Systeme, die primär Sprachsignale als Interaktionsmöglichkeit besitzen. Diese IPAs sind sehr stark verbreitet und auf vielen Geräten verfügbar [5].

1.1 Was ist speech emotions recognition (SER)

SER ist ein Forschungsgebiet, welches sich mit der Analyse von Audiosignalen, durch das Extrahieren von typischen Merkmalen und anschließender Klassifikation der Emotionen, beschäftigt [2]. In den letzten Jahren wurden in diesem Gebiet von Spracherkennung signifikante Fortschritte gemacht. Die Audiosignale enthalten nicht nur die gesprochenen Wörter, sondern vielmehr auch den emotionalen Zustand des Sprechers [3, 5]. Das Ziel hierbei ist es durch verschiedene Klassifikationsverfahren das menschliche Befinden, die Gefühle und die Emotionen aus den Audiosignalen zu extrahieren und somit den emotionalen Zustand des Sprechers vorherzusagen [2]. SER kann zum Beispiel als diagnostisches Werkzeug für Therapeuten verwendet werden. Eine weitere Einsatzmöglichkeit bietet SER bei Notrufzentralen, um die Ernsthaftigkeit der Situation des Anrufenden maschinell durch die Stimme erkennen zu können [3]. Bei SER-Systemen gibt es verschiedene Herausforderungen wie zum Beispiel Robustheit bei Tonänderungen, unterschiedliche Sprechstile, Anzahl der Wörter und kulturell bedingte Ausdrucksweise der Emotionen beim Sprechen [2]. Allgemein werden Emotionen bei SER in 2 Hauptkategorien unterteilt - bewusst und

unbewusst ausgedrückte Emotionen [9]. Bewusst ausgedrückte Emotionen sind offensichtlicher als unbewusst ausgedrückte Emotionen. Wenn jemand beim Sprechen mit der Stimme lauter wird, so ist dies ein Indikator dafür, dass der Sprecher wütend ist. Dies ist ein Beispiel für bewusst ausgedrückte Emotionen. Der Sprecher kann aber auch wütend sein ohne seine Stimme zu erheben, da müssen dann andere Indikatoren hergezogen werden wie zum Beispiel die Knappheit der Wörter [9]. Somit ist das Hauptproblem von SER-Systemen die Erkennung von affektorientierte Unterscheidungsmerkmale der Sprachsignale und welche Kriterien zur Vorhersage für die Emotionen des Sprechers dienen [3].

1.2 Spektrogramme

Spektrogramme kommen in verschiedenen Sprachanalyse-Tools zum Einsatz wie zum Beispiel bei Sound Event Classification (SEC), Sprecher Erkennung, Spracherkennung und SER [2]. Bei SER-Systemen dienen Spektrogramme als Input in das Convolutional Neural Network (CNN) [7], welche aus den Sprachsignalen generiert werden. Ein Spektrogramm ist eine visuelle Darstellung der Audiosignale in Form von zweidimensionalen Graphen [2]. Dieser Graph hat folgende geometrische Dimensionen: die horizontale Achse repräsentiert die Zeit t und die vertikale Achse bildet die zeitabhängige Frequenz des Audiosignals [3]. Die Amplitude der Frequenz wird durch unterschiedliche Farben in der Darstellung gezeigt: niedrige Amplituden werden durch dunkelblaue Farben und hohe Amplituden durch hellere Farben bis hin zu Rot dargestellt (siehe Abbildung 1.1) [2]. Somit entsteht eine Wellenform über die Zeit, was die Audiosignale des Sprechers repräsentiert und wichtige Informationen enthält.

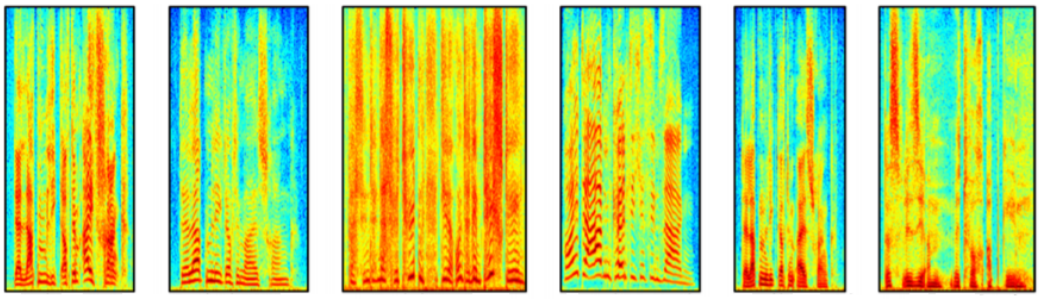

Abbildung 1.1: Spektogramme für verschiedene Emotionen [3]

1.2.1 STFT und FFT

STFT steht für short term Fourier transform und wird für elektronisch aufgezeichnete Tonaufnahmen verwendet, um aus den Signalen ein Spektrogramm zu generieren [3]. Bei fast Fourier transform (FFT) handelt es sich um einen digitalen Prozess, welcher die Signale über ein sich gleitendes Fenster berechnet und damit ein Spektrogramm erstellt [3]. Bei der Erstellung der Spektrogramme im Paper [3] wurden Sprachaufnahmen von Emo-DB (Berlin Database of Emotional Speech [4]) mit dem STFT-Algorithmus in Spektrogrammen umgewandelt, um später sie als Input für das Convolutional Neural Network zu verwenden.

2 Convolutional Neural Networks

In diesem Kapitel wird der Aufbau und Funktionsweise von CNN Architekturen erklärt und erläutert.

In den letzten Jahren haben Convolutional Neural Networks (CNN) in Zusammenhang mit Mustererkennung wie zum Beispiel Bildklassifizierung oder Spracherkennung bahnbrechende Ergebnisse erzielt [1]. Bei Bildklassifizierungen liefern CNN Architekturen die besten Ergebnisse [8]. Ein großer Vorteil bei CNN Architekturen ist, dass eine Reduzierung der Parameter Schicht für Schicht stattfindet und dadurch größere Modelle mit komplexen Aufgaben im Anschluss besser klassifiziert werden können [1]. CNN ist ein hierarchisches neuronales Netz, welches aus unterschiedlichen Schichten (layers) besteht [3]. Diese Schichten kann man in drei Hauptkomponenten aufteilen (siehe Abbildung 3.2).

- **convolutional layers** [3]
- **pooling layers** [3]
- **fully connected layers** [3]

Bei den **convolutional layers** kommt ein Faltungsfilter zum Einsatz, welcher auf den Input (Bild/Spektrogramm) angewendet wird. [3].
pooling layers: In den ersten Schichten werden einfache Merkmale wie zum Beispiel einzelne Bildpixel und Kanten in Betracht gezogen. Danach werden Schicht für Schicht Merkmalsextraktion und Abstraktionen auf höherer Ebene durchgeführt und Unterscheidungsmerkmale identifiziert [3]. Hier findet also eine Reduzierung der Dimensionalität statt, welche im Einführungstext in diesem Kapitel als ein Vorteil von CNN erwähnt wurde. Der am häufigsten verwendete Pooling-Algorithmus ist der sogennante max pooling, welcher die Maximalwerte behält. [3].

Fully connected layers sind für die globale Repräsentation der Faltungsmerkmale und Klassifikationen zuständig [3]. Diese Merkmale werden dann in einem sogenannten Softmax-Klassifikator übergen, welcher dann für jede Klasse der Faltungen die Wahrscheinlichkeiten generiert.

3 Modellarchitektur und Ablauf

In diesem Kapitel werden die zwei Phasen bei SER, die Modellarchitektur und die Vorgehensweise des Frameworks bei einem SER-System erläutert. Hierbei können unterschiedliche Kalssifikationsverfahren verwendet werden, welche in 3.1

3.1 Zwei Phasen bei SER

Die große Herausforderung für SER-Systemen ist die Unterscheidung der verschiedenen Emotionen durch die Sprache zu ermöglichen. Jeder Sprecher hat individuelle und kulturell bedingte Sprechstile, Sprechgeschwindigkeit, unterschiedliche Tonhöhe und Energiekontur im Spektrogramm, was das Extrahieren der Merkmale erschwert [3]. Diese Umstände werden in der Verarbeitungseinheit behandelt, um später beim Klassifizieren der Sprachsignale bessere Ergebnisse zu erhalten.

3.1.1 Verarbeitungseinheit (processing unit)

Um die in 3.1 genannten Probleme anzugehen wird in der Verarbeitungseinheit das Spektrogramm in mehrere Blöcke (chunks) aufgeteilt, die Frames genannt werden (siehe Abbildung 3.1) [3]. Diese einzelnen Frames werden dann im nächsten Schritt Klassifikator) analysiert.

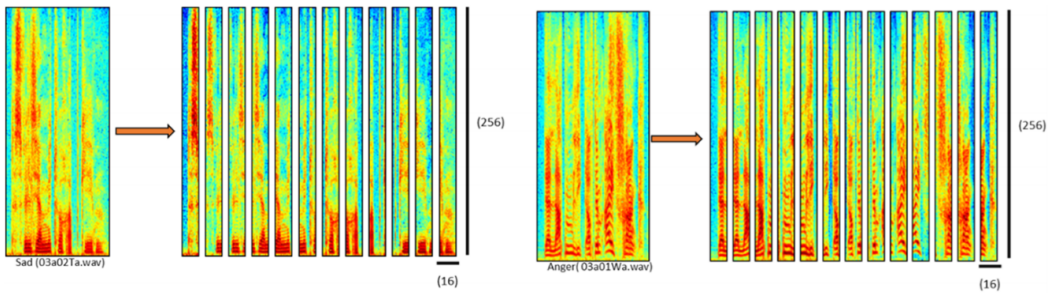

Abbildung 3.1: Spektrogramme, die in Frames aufgeteilt werden [3]

3.1.2 Klassifikator (classifier)

Nachdem das Spektrogramm in mehreren Frames aufgeteilt wurde findet die Klassifikation mit Hilfe von Machine Learning Algorithmen statt [3]. Hierbei werden die Frames einzeln untersucht und es können verschiedene Arten von Klassifikationsverfahren zum Einsatz kommen. Die gängigsten Algorithmen sind Hidden-Markov-Modelle (HMM), Gaußsches Mischungsmodell, Support Vector Machine (SVM), künstliche neuronale Netze und K-nearest neighbor, wobei SVM und HMM die am weitesten verbreitete Lernalgorithmen für sprachbezogene Anwendungen sind [3].
SVM ist weit verbreitet für Mustererkennungs- und Klassifikationsprobleme [10]. Im Vergleich zu den anderen Klassifikationsverfahren ist der Vorteil bei diesem Algorithmus, dass nicht so viele Traningsdaten benötigt werden um gute Klassifikationsleistung aufzuweisen [10].
HMM ist ein statistischer Modellierungsverfahren, welcher bei Klassifikationsproblemen in Spracherkennungsanwendungen häufig eingesetzt werden [6]. Die Grundidee bei HMM ist es, dass ein endliches Modell, welches einer Wahrscheinlichkeitsverteilung über eine unendliche Anzahl möglicher Folgen beschreibt.[6].

Wichtig ist es den Klassifikator zu trainieren. Dazu benötigt man gelabelte Datensätze, die als Tranings-Daten dienen. Hier bietet sich zum Beispiel die Emo-Db an [4]. In der Datenbank befinden sich gelabelte Tonaufnahmen von fünf weibliche und fünf männliche Schauspieler [4]. Die gelabelten Emotionen sind Ärger, neutral, Angst, Freude, Trauer, Ekel und Langeweile [4]. Nachdem der Klassifikator trainiert wurde können Testläufe durchlaufen werden, um die Richtigkeit und Robustheit der Vorhersagen des Modells zu testen [9].

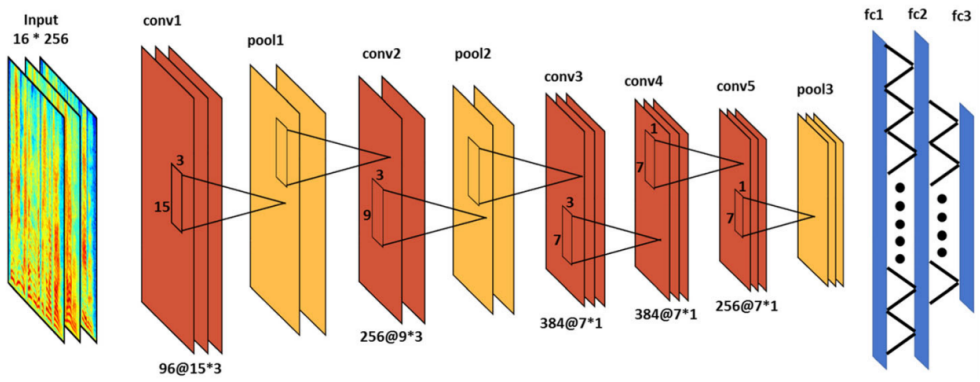

Abbildung 3.2: CNN Architektur mit den unterschiedlichen Schichten [3]

3.2 Aufbau der Modellarchitektur

In Abbildung 3.2 wird der genau Aufbau der Modellarchitektur dargestellt. Als Input (links) für das CNN dienen die Sprachsignale in Form von Spektrogrammen. Die Architektur hat insgesamt fünf convolutional layer, drei pooling layer und drei fully connected layer (rechts) [3]. Der Ausgang des letzten fully connected layer wird dem Softmax-Klassifikator weitergegeben, welcher für die Berechnung der Ausgangswahrscheinlichkeit und der Zuordnung der jeweiligen Emotions-Klasse zuständig ist [3].

3.3 Der Ablauf bei SER

In Abbildung 3.3 ist der komplette Ablauf des SER-Systems abgebildet. Der Zyklus beginnt mit einem Audiosignal, welches durch STFT ein Spektrogramm erstellt wird. Dieses Spektrogramm wird im nächsten Schritt in Frames aufgeteilt. Dies ist die Verarbeitungseinheit des Inputes (siehe 3.1.1). Im nächsten Schritt werden die einzelnen Frames weitergeleitet an den Klassifikator, welcher

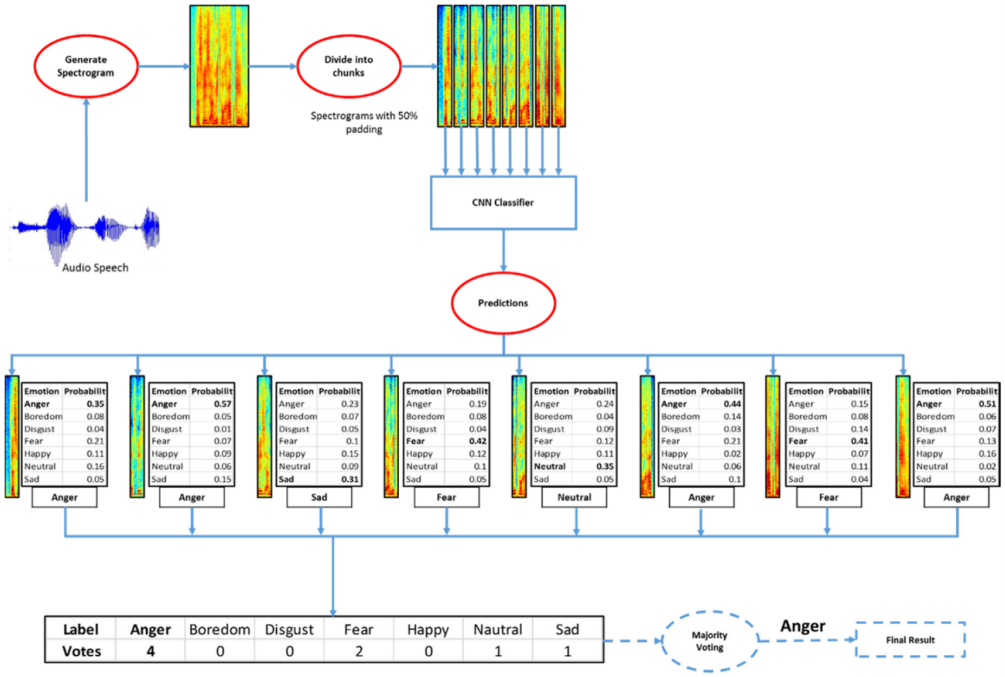

Abbildung 3.3: spezifischer Schema und Ablauf bei SER [3]

4 Schlussfolgerung

SER bietet ein breites Spektrum an Einsatzmöglichkeiten an und die Anwendungen von diesen Spracherkennungssystemen nehmen rasant zu. In der Automobilbranche kann so ein System bei der Erkennung des mentalen Zustands des Fahrers hilfreich sein [3]. Des weiteren kann SER eine wichtige Komponente bei der Entwicklung intelligenter Dienste in der Gesundheitsversorgung, Audio-Forensik und Mensch-Maschine Interaktion sein [3].

Literaturverzeichnis

[1] ALBAWI, Saad ; MOHAMMED, Tareq A. ; AL-ZAWI, Saad: Understanding of a convolutional neural network. In: *2017 International Conference on Engineering and Technology (ICET)*, 2017, S. 1–6

[2] BADSHAH, Abdul M. ; AHMAD, Jamil ; RAHIM, Nasir ; BAIK, Sung W.: Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: *2017 International Conference on Platform Technology and Service (PlatCon)*, 2017, S. 1–5

[3] BADSHAH, Abdul M. ; RAHIM, Nasir ; ULLAH, Noor ; AHMAD, Jamil ; MUHAMMAD, Khan ; LEE, Mi Y. ; KWON, Soonil ; BAIK, Sung W.: Deep features-based speech emotion recognition for smart affective services. In: *Multimedia Tools and Applications* 78 (2019), Nr. 5, S. 5571–5589

[4] BURKHARDT, Felix ; PAESCHKE, Astrid ; ROLFES, Miriam ; SENDLMEIER, Walter F. ; WEISS, Benjamin: A database of German emotional speech. In: *Ninth european conference on speech communication and technology*, 2005

[5] CLARK, Leigh ; DOYLE, Philip ; GARAIALDE, Diego ; GILMARTIN, Emer ; SCHLÖGL, Stephan ; EDLUND, Jens ; AYLETT, Matthew ; CABRAL, João ; MUNTEANU, Cosmin ; EDWARDS, Justin ; R COWAN, Benjamin: The State of Speech in HCI: Trends, Themes and Challenges. In: *Interacting with Computers* 31 (2019), 09, Nr. 4, 349-371. http://dx.doi.org/10.1093/iwc/iwz016. – DOI 10.1093/iwc/iwz016. – ISSN 0953–5438

[6] EDDY, Sean R.: Hidden markov models. In: *Current opinion in structural biology* 6 (1996), Nr. 3, S. 361–365

[7] HUANG, Zhengwei ; DONG, Ming ; MAO, Qirong ; ZHAN, Yongzhao: Speech Emotion Recognition Using CNN. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. New York, NY, USA : Association for Computing Machinery, 2014 (MM '14). – ISBN 9781450330633, 801–804

[8] KRIZHEVSKY, Alex ; SUTSKEVER, Ilya ; HINTON, Geoffrey E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems* 25 (2012), S. 1097–1105

[9] LI, Wu ; ZHANG, Yanhui ; FU, Yingzi: Speech Emotion Recognition in E-learning System Based on Affective Computing. In: *Third International Conference on Natural Computation (ICNC 2007)* Bd. 5, 2007, S. 809–813

[10] PAN, Yixiong ; SHEN, Peipei ; SHEN, Liping: Speech emotion recognition using support vector machine. In: *International Journal of Smart Home* 6 (2012), Nr. 2, S. 101–108