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.

[2] [4] [5] [1]

1 Einleitung

Die Sprache des Menschen ist die natürlichste Art und Weise um miteinander zu kommunizieren. Durch die technische Entwicklung in den letzten Jahren kommen immer mehr Interaktionen zwischen Mensch und Maschine durch die Sprache zustande [2]. Die Bezeichnung dafür sind die sogenannten intelligent personal assistants (IPAs) 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 [3].

1.1 Was ist speech emotions recognition (SER)

SER ist ein Forschungsgebiet, welches sich mit der Analyse von Audiosignalen in Form von Spektrogrammen beschäftigt. In den letzten Jahren wurden in dem Gebiet von Spracherkennung signifikante Fortschritte gemacht. Die Audiosignale enthalten nicht nur die gesprochenen Wörter, sondern vielmehr auch den emotionalen Zustand des Sprechers [2, 3]. Das Ziel hierbei ist es durch verschiedene machine-learning Algorithmen das menschliche Befinden, die Gefühle und Emotionen aus den Audiosignalen zu extrahieren und somit den emotionalen Zustand des Sprechers zu erkennen. 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 [2]. Allgemein werden Emotionen bei SER in 2 Hauptkategorien unterteilt - bewusst und unbewusst ausgedrückte Emotionen [5]. 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.

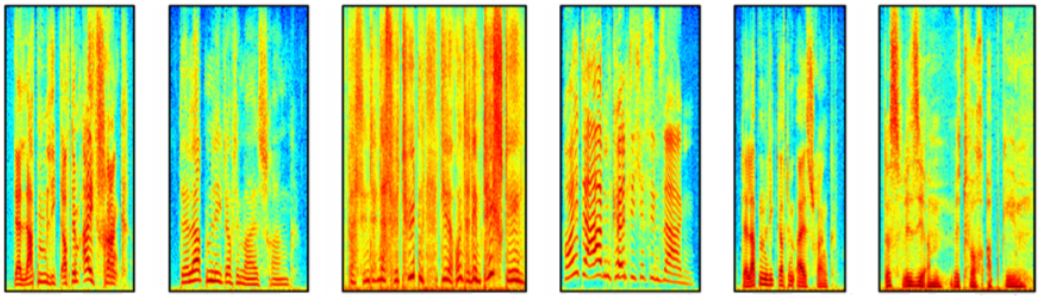

Abbildung 1.1: Spektogramme für verschiedene Emotionen [2]

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 [5]. 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 [2].

1.2 Spektrogramme

Spektrogramme spielen bei SER eine wichtige Rolle, denn sie dienen als Input.

1.2.1 STFT und FFT

2 Convolutional Neural Networks

Convolutional Neural Network (CNN) ist ein hierarchisches neuronales Netz, welches aus unterschiedlichen Schichten (layers) besteht. Diese Schichten kann man in drei Hauptkomponenten aufteilen.[2]

- convolutional layers
  diese Schicht ist für das Filtern des Inputs zuständig
- pooling layers
- fully connected layers

3 Modellarchitektur und Ablauf

In diesem Abschnitt wird der Aufbau und die Vorgehensweise des Algorithmuses bei SER-Systemen erläutert.

3.1 2 Phasen der 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 [2]. Diese Umstände werden in der Verarbeitungseinheit behandelt, um später beim Klassifizieren der Sprachsignale gute 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) [2].

3.1.2 Klassifikator (classifier)

Nachdem das Spektrogramm in mehreren Frames aufgeteilt wurde findet die Klassifikation mit Hilfe von Machine Learning Algorithmen statt [2]. Hierbei werden die Frames einzeln

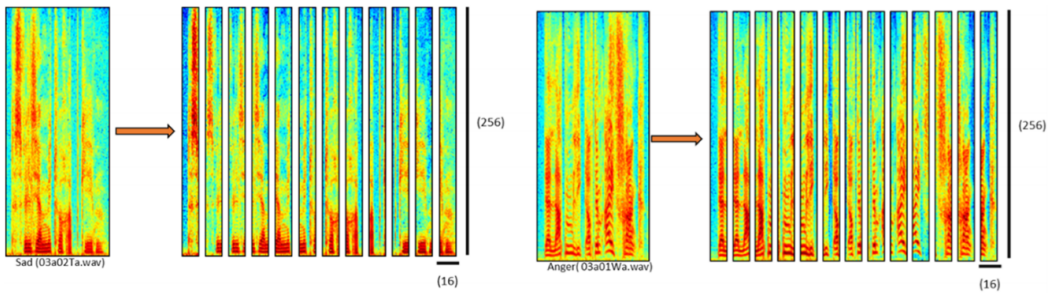

Abbildung 3.1: in Frames ausgeteilte Spektrogramme [2]

untersucht und es kommen verschiedene Arten von Klassifikatoren zum Einsatz. 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 verbreiteten Lernalgorithmen für sprachbezogene Anwendungen sind [2].

3.2 Aufbau der Modellarchitektur

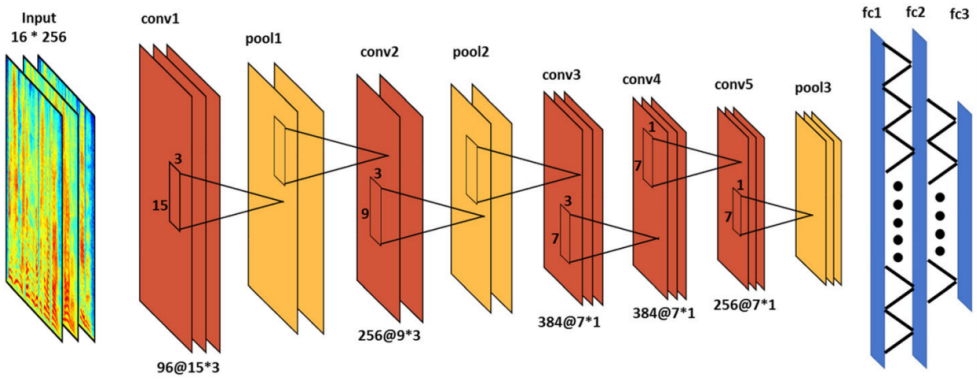

Abbildung 3.2: CNN Architektur mit den unterschiedlichen Schichten [2]

Mit Hilfe eines Labels kann man sich dann im Text auf diese Grafik (3.2) beziehen.

3.3 Der Ablauf bei SER

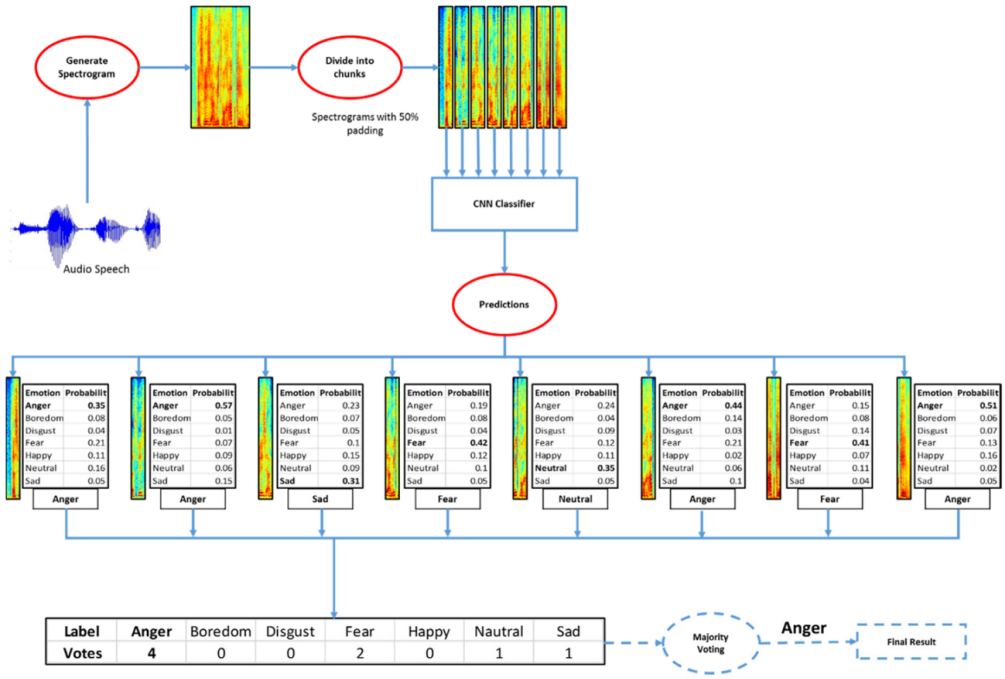

Abbildung 3.3: spezifischer Schema und Ablauf bei SER [2]

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 [2]. Des weiteren kann SER eine wichtige Komponente bei der Entwicklung intelligenter Dienste in der Gesundheitsversorgung, Audio-Forensik und Mensch-Maschine Interaktion sein [2].

Literaturverzeichnis

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

Name: Salih Bedelce Matrikelnummer: 1038226

Erklärung

Ich erkläre, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Ulm, den .....

Salih Bedelce