Universität Hamburg
Fakultät für Mathematik,
Informatik und Naturwissenschaften

Hausarbeit im Proseminar:
Verarbeitung gesprochener Sprache

Vergleich von openEAR und EmoVoice

Jan Mägdefrau

9maegdef@informatik.uni-hamburg.de
Studiengang Informatik
Matr.-Nr. 7260915
Fachsemester 2

Inhaltsverzeichnis

1 Einleitung

Abstract - Sowohl EmoVoice, als auch OpenEAR stellen Programme zur Klassifikation von Emotionen in der digitalen Sprachverarbeitung dar. Beide verfügen dabei über Gemeinsamkeiten und Unterschiede. Diese Seminararbeit wird die beiden Systeme hinsichtlich Genauigkeit und Integrierbarkeit in eigene Anwendungen untersuchen.

1.1 Motivation der Arbeit

Seit dem Jahr 2000 wird jährlich der BigBrotherAward an Firmen und Organisationen verliehen, welche in besonderem Maße gegen die Privatsphäre von Menschen verstoßen [Big]. Im Jahr 2019 gewann in der Kategorie Kommunikation, die aus Aachen stammende Firma Precire Technologies GmbH. Sie entwickelte eine Software, welche unter anderem anhand eines fünfzehnminütigen Telefon-Interviews, Bewerber für einen Job beurteilte. Das Interessante daran: der Inhalt des Gesprochenen, also das „was“, war dabei nicht von Bedeutung, lediglich die Stimme, das „wie“, wurde analysiert. Die Stimmprobe wurde angeblich in über 500.000 Bestandteile zerlegt, wobei unter anderem Eigenschaften wie Stimmhöhe, Lautstärke, Modulationsfähigkeit und Sprechtempo extrahiert wurden. Zur Analyse verglich Precire Technologies die gewonnen Daten mit denen von Testpersonen aus ihrem Datenpool und erstellte so eine psychologische Beurteilung des Bewerbers. [Big19]

Dass die Analyse von paralinguistischen Phänomenen und Emotionen aus Sprache auch seriös in der Wissenschaft Anwendung findet, zeigt die jährlich im Rahmen der INTERSPEECH, der größten internationalen Konferenz mit Fokus auf Sprachverarbeitung, stattfindende Interspeech Computational Paralinguistics ChallengE (ComParE) [SB19]. Dort treten jedes Jahr Universitäten gegeneinander an, um in verschiedenste Challenges, bestmöglich paralinguistische Phänomene aus einem Sprachbeispiel zu analysieren. So galt es beispielsweise aus einem Sprachbeispiel zu erkennen, ob der Sprecher müde ist, einen bestimmten Dialekt spricht [SBB+19] oder welche Emotion [DGJ+13] er während des Sprechens verkörpert.

Auch zwei deutsche Universitäten entwickelten Frameworks mit welchen die Analyse von Emotionen aus Sprache ermöglicht werden kann. Die Universität Augsburg veröffentlichte 2008 das Framework zur online Erkennung von Emotionen aus Sprache – EmoVoice [VAB08].

2009 entwickelte die Technische Universität München, dann das emotionserkennende Open-Source Toolkit openEAR [EWS09].

1.2 Zielsetzung

Die Frameworks openEAR und EmoVoice wurden beide entwickelt, um die Analyse von Emotionen aus gesprochener Sprache zu erleichtern. Das Ziel dieser Arbeit ist es dabei, dem Leser beide Systeme bestmöglich nahe zu bringen und hinsichtlich der Aspekte Genauigkeit und Integrierbarkeit in Anwendungen zu vergleichen.

1.3 Aufbau der Arbeit

Die Arbeit besteht aus drei Teilen. Zunächst werden im Kapitel 2, die beiden Frameworks vorgestellt damit der Leser einen Überblick über beide Systeme erhält. Dabei wird der grundsätzliche Aufbau der Software beschrieben sowie die verwendeten Algorithmen und Methoden zur Feature Extraktion und anschließenden Klassifizierung vorgestellt. Im anschließenden Kapitel 3 folgt der Vergleich in den beschriebenen Aspekten Genauigkeit sowie Integrierbarkeit in Anwendungen. Abschließend nehmen wir in Kapitel 4 eine Einordnung vor und ziehen ein Fazit, welches System in welchem Bereich besser abschneidet.

2 Vorstellung Frameworks

Damit eine Vergleich beider Systeme möglich wird, werden nun im folgenden Abschnitt die beiden System vorgestellt. Dabei wird das Hauptaugenmerk auf den technischen Aufbau der jeweiligen Software gelegt, damit im folgenden Kapitel ein Vergleich von Genaugikeit und Integrierbarkeit in Anwendungen vorgenommen werden kann.

2.1 EmoVoice

EmoVoice wurde im Jahr 2008 von der Universität Augsburg als Framework zur offline und online Echtzeit Erkennung von Emotionen aus Sprache entwickelt. Das System kann dabei in zwei Module unterteilt werden: ein Modul zur offline Erkennung von Emotionen aus einem Sprachkorpus und ein Modul zur online Echtzeit-Analyse während des Sprechens.

Das erste Modul besteht neben einer grafischen Nutzerschnittstelle zum Aufnehmen von Sprachbeispielen und zur Erzeugung von Klassifikatoren, auch aus einer Vielzahl von Werkzeugen. Diese Werkzeuge werden dabei genutzt, um die Schritte Audiosegmentierung, Merkmalsextration sowie anschließende Klassifizierung in eine der gewünschten Emotionen zu durchlaufen. Die Audiosegmentierung dient dabei dazu, das Audiosignal in kleine Teile zerlegen zu können. Da EmoVoice eine Echtzeitanalyse vornimmt, kann kein zeitaufwändiger Automatic Speech Recogniser verwendet werden, sondern es wurde auf eine Voice Activity Detection zurückgegriffen. Bei diesem Schritt ist vor allem die Länge der Segmente von großer Bedeutung. Sind die Segmente zu kurz gibt es wohlmöglich zu wenig Daten, um eine Klassifizierung vornehmen zu können. Ist das Segment zu lang könnte sich die Emotion während des Segmentes verändern. Auf Grund dieser Einschränkungen und weil im fließenden Sprechen längere Pausen selten sein können, legt EmoVoice eine maximale Länge der Segmente von 2 bis 3 Sekunden fest.

In der Merkmalsextraktion gilt es nun jene Eigenschaften im akustischen Signal zu extrahieren, welche eine bestimmte Emotion bestmöglich charakterisieren. Das Ziel dabei ist es Feature-Vektoren zu finden, welche während der Klassifikation mit einem Label versehen werden können. Für jedes Segment wird dabei ein Feature-Vektor erzeugt. Da vor allem die Veränderung des Signals mit der Zeit wichtig ist, werden mit Hilfe von statistischen Methoden wie dem Durchschnitt, diese Informationen im Vektor enkodiert. Da EmoVoice eine Echtzeit-

Analyse vornimmt, können auch nur Eigenschaften genutzt werden, welche vollautomatisch extrahiert werden können. Die dabei grundlegend verwendeten Eigenschaften sind: logarithmierte Tonhöhe, Signalenergie, Mel Frequency Cepstral Coefficients, Kurzzeit-Frequenz-Spektrum und Harmonics-to-noise ratio. Neben weiteren extrahierten Eigenschaften werden auf alle grundlegenden Features statistische Methoden, wie z.B. Maximum, Minimum, Varianz und Median angewandt, so dass insgesamt der Vektor aus 1302 Features besteht. Da dies für eine schnelle Echtzeitanalyse zu viele sind, wird mit Hilfe einer Correlation-based Feature Selection eine Reduktion auf 50-200 Features vorgenommen.

In der abschließenden Klassifizierung wird nun dem Feature-Vektor ein Label zugewiesen. EmoVoice hat dafür zwei verschiedenen Algorithmen integriert: den Naive Bayes (NB) classifier und den Support Vector Machine (SVM) classifier. Der Naive Bayes classifier liefert auch für große Feature-Vektoren schnell Ergebnisse und ist damit gut für eine Echtzeitanalyse geeignet. Im Gegensatz zum SVM classifier sind die Ergebnisse allerdings ungenauer. Mit einem kleinen Feature-Vektor liefert allerdings auch der SVM classifier schnelle Ergebnisse.

2.2 openEAR

Im Jahr 2009 veröffentliche der Fachbereich Mensch-Maschine-Kommunikation der Technischen Universität München das Toolkit zur Emotions- und Affekterkennung „openEAR“. Ihr Ziel war es dabei eine stabile und effiziente Software, welche vollständig Open-Source ist, zu erschaffen. Das Toolkit besteht dabei aus drei Hauptkomponenten. Einem Tool zur Signalverarbeitung und Merkmalsextraktion, SMILE (Speech and Music Interpretation by Large-Space Extraction). Dieses Werkzeug kann Features in Echtzeit sowohl aus Live-Audio als auch aus Offline-Medien extrahieren. Dabei werden Low-Level Audiomerkmale extrahiert, auf welche verschiedenste statistische Funktionen und Transformationen angewandt werden können. Des Weiteren gibt es eine Komponente, welche die Möglichkeit bietet, verschiedene Klassifikatoren auf die extrahierten Merkmale anzuwenden. Die dritte Komponente ist ein Werkzeug, welches sowohl die Möglichkeit bietet, eigenen Modelle zu trainieren, als auch eine von vier bereits trainierten Modellen anzuwenden. Die mitgelieferten Modelle umfassen dabei: die Erkennung von 6 Emotionen, die Einordnung in einen drei-dimensionalen Raum aus „activitation“, „valence“ und „dominance“, die Erkennung des Levels an Interesse und die Erkennung von affektiven Zuständen wie Trunkenheit.

openEAR ist dabei sehr modular aufgebaut und kann dadurch leicht angepasst werden. Die zentrale Einheit bildet das „Data Memory“, in welches durch belie-

bige „Data Sources“ – wie zum Beispiel eine Audiodatei – Daten geladen werden können. Diese können dann mit verschiedenen „Date Processors“ verarbeitet werden und anschließend durch „Data Sinks“ an einen Klassifikator oder einen Datenexport angeschlossen werden.

3 Vergleich

3.1 Genauigkeit

Zu den wichtigsten Eigenschaften einer Emotions-Erkennungssoftware gehört die Genauigkeit, sprich die Wahrscheinlichkeit, mit welcher einer Sprachsequenz die korrekte Emotion zugeordnet werden kann. Um diese Eigenschaft vergleichen zu können, wurde sowohl bei openEAR als auch bei EmoVoice die Genauigkeit anhand mehrerer Sprachkorpusse gemessen. Ein Sprachkorpus ist eine Datenbank von Sprach-Audiodateien, welche sowohl aus geschauspielerten bzw. eingesprochenen Sprachbeispielen als auch aus spontanen Sprachbeispielen bestehen kann. [Wik] Für einen Vergleich der beiden Programme werden hier die Sprachkorpusse Berlin Speech Emotion Database und SmartKom verwendet. Berlin Speech Emotion Database, kurz Emo-DB, umfasst jeweils 10 Sätze in 7 Emotionen, welche von 5 weiblichen und 5 männlichen Sprechern eingesprochen wurden. Die geschauspielerten Emotionen umfassen dabei „neutral“, „Ärger“, „Angst“, „Freude“, „Trauer“, „Ekel“ und „Langeweile“. [BPR+05] SmartKom ist eine Datenbank, welche spontane Emotionen beinhaltet, welche während einer sogenannten „Wizard-Of-Oz“-Studie aufgenommen wurden. Die Probanden mussten Aufgaben lösen, während sie dabei aufgenommen wurden. Die Emotionen können dort unterteilt werden in: „neutral“, „Freude“, „Ärger“, „Hilflosigkeit“, „nachdenklich“ und „überrascht“. [SSDR02]

Nutzt man Emo-DB als Sprachkorpus erreicht openEAR nach [SVE+09] beim Nutzen einer Support-Vector-Machine als Classifier eine Genauigkeit von 84,6% (unweighted average) bzw. 85,6% (weighted average). Im originalen Paper [EWS09] werden mit 88,8% (unweighted average) und 89,5% (weighted average) die Genauigkeiten, um 4,2 Prozentpunkte bzw. 3,6 Prozentpunkte höher angegeben. Mit dem von EmoVoice verwendeten Algorithmus wird eine Genauigkeit von 81,14% bei einer „overall recognition rate“ und 79,18% bei einer „class-wise recognition rate“ erreicht. [VA06]

Für den SmartKom-Sprachkorpus erreicht openEAR ebenfalls beim Nutzen ei-

ner Support-Vector-Machine eine Genauigkeit von 23,5% (unweighted average) sowie 39,9% (weighted average). [SVE+09] EmoVoice erreicht bei diesem Sprachkorpus mit einer leicht veränderten Festlegung der zu klassifizierenden Emotionen („nachdenklich“ wird hier nicht berücksichtigt) eine Genauigkeit von 37,5%. [VA05]

Vergleicht man nun diese Ergebnisse so zeigt sich, dass wenn zudem auch noch die leicht unterschiedlichen Testszenarien berücksichtigt werden, die Genauigkeit von EmoVoice und openEAR sehr ähnlich sind. Auffällig ist dabei, dass bei geschauspielerten Emotionen (Emo-DB), die Genauigkeit wesentlich höher ausfällt als bei spontanen Emotionen (SmartKom). Diese Abweichung kann unteranderem damit erklärt werden, dass SmartKom ein hohes Noise-Level beinhaltet und die Emotionen der Probanden multimodal, also beispielsweise über Mimik, Gestik und Sprache ausgedrückt werden und damit die Emotion nicht immer eindeutig über die Sprache allein erfasst werden kann. [SVE+09]

Zudem zeigt sich, dass je weiter die Emotionen zusammengefasst werden, desto genauer sind die Ergebnisse: Fasst man bei SmartKom-Sprachkorpus die Emotionen „Neutral“, „Freude“ und „Überrascht“ zu „kein Problem“ und „Hilflosigkeit“ sowie „Ärger“ zu „Problem“ zusammen, so kann mit EmoVoice eine Genauigkeit von 68,3% erreicht werden. [VA05]

3.2 Integrierbarkeit in Anwendungen

Allein die Information, welche Emotion ein Sprecher während des Redens verkörpert ist per se nicht von großer Bedeutung. Diese Bedeutung wird erst durch die Integration in eine reale Anwendung erlangt. EmoVoice und openEAR sind beide in zahlreiche Projekte integriert.

Die mit Hilfe von EmoVoice analysierten Emotionen können direkt durch die Verwendung einer Socket-Verbindung an eine Anwendung weitergeleitet werden. Dies passiert bei mehreren Projekten, welche Versuchen einen Roboter, durch das Reagieren auf und zeigen von Emotionen, glaubhafter erscheinen lassen. So sollten Versuchspersonen dem Roboter BARTHOC, sehr emotional das Märchen Rotkäppchen vorlesen. EmoVoice analysierte die Emotion des Lesers und der Roboter versuchte auf diese Emotion zu reagieren bzw. sie zu spiegeln. Anschließend mussten die Versuchspersonen Fragen beantworten, ob die vom Roboter gezeigten Emotionen zu den gelesenen Buchstellen passten. [HSW+06] Eine ähnliche Anwendung ist die virtuelle Agentin Greta, welche ebenfalls mit Hilfe von EmoVoice den emotionalen Zustand des Nutzer analysiert und diesen in ihrem digitalen Gesicht wiederspiegelt und so dem Nutzer emotionsbehaftetes Feed-

back gibt und versucht Empathie zu zeigen [DRPP+03]

Auch openEAR findet in mehrere Projekten Anwendung. So zum Beispiel bei der bereits in der Einleitung erwähnten INTERSPEECH Paralinguistic Challenge 2010. Dort wurde der Feature-Extraktor openSMILE aus dem openEAR Toolkit verwendet, um die Features für die einzelnen Aufgaben zu finden. [SSB+10] Auch bei der Audio-Visual Emotion Challenge (AVEC) 2014, eine zur INTERSPEECH Paralinguistic Challenge ähnliche Veranstaltung, wurde der Feature-Extraktor angewandt. [VSS+14] Zudem wurde mit Hilfe von statistischen Funktionen aus dem openEAR-Toolkit in einem Projekt, welches die Ablenkung eines Fahrzeugführers versucht zu analysieren, die bereits extrahierten Low-Level-Features weiterverarbeitet. [WBS+11]

Diese Beispiele zeigen gut, die unterschiedlichen Zielsetzungen der Systeme. EmoVoice wurde entwickelt, um direkt in der Mensch-Computer-Interaktion, in vollem Umfang (Audioaufnahme bis Klassifizierung), Anwendung zu finden. [VAB08] openEAR hingegen sollte ein Toolkit zur weiteren Softwareentwicklung werden [EWS09], wobei die Projekte gebrauch vom modularen Aufbau machen und so nur einzelne Komponenten, wie zum Beispiel die Feature-Extraktion, in eigenen Anwendungen integrieren.

4 Fazit

Der Vergleich hat gezeigt, dass beide Systeme eine ähnliche Genauigkeit aufweisen, welche bei realen Bedingungen meist noch wesentlich geringer ist als bei geschauspielerten Sprachbeispielen. Dies zeigt einmal mehr, wie gefährlich heutzutage noch der Einsatz von Systemen, wie das in der Einleitung erwähnte Analysetool der Firma Precire sein kann. Menschen die sich zum Beispiel für einen Beruf bewerben, laufen so immer Gefahr, da das System keine Genauigkeit von 100% besitzt, auf Grund eines Klassifizierungsfehlers falsch beurteilt und so eventuell unberechtigt abgelehnt zu werden.

Eine Entscheidung zwischen beiden Systemen kann also unabhängig von der Genauigkeit getroffen werden, da beide dort ähnliche Werte aufweisen. Vielmehr sollte bei der Wahl das gewünschte Einsatzszenario berücksichtigt werden. Möchte man eine Software, welche man ohne tiefergreifendes Verständnis über die eigentliche Klassifizierung der Emotionen in eine eigene Anwendung integrieren kann, so bietet sich EmoVoice als Lösung an.

Möchte man selber eine Analyse von paralinguistischen Phänomenen vorneh-

men und benötigt dafür nur ein oder mehrere spezielle Werkzeuge, so bietet das openEAR-Toolkit aufgrund seiner Modularität eine gute Lösung dafür an. Hierbei sollte allerdings beachtet werden, dass zur Integration ein grundlegendes Verständnis über die Analyse von Emotionen aus Sprache benötigt wird und sich openEAR somit mehr an ein Fachpublikum als an ein breites Anwenderspektrum richtet.

Literaturverzeichnis

[Big] BIGBROTHERAWARDS: *Was sind die BigBrotherAwards?* Internet. https://bigbrotherawards.de/ueber-uns. – Abgerufen: 17.08.2020

[Big19] BIGBROTHERAWARDS: *Kommunikation: Precire Technologies GmbH | BigBrotherAwards*. Internet. https://bigbrotherawards.de/2019/kommunikation-precire-technologies-gmbh. Version: 2019. – Abgerufen: 2020-08-17

[BPR+05] 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

[DGJ+13] DHALL, Abhinav ; GOECKE, Roland ; JOSHI, Jyoti ; WAGNER, Michael ; GEDEON, Tom: Emotion recognition in the wild challenge 2013. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, S. 509–516

[DRPP+03] DE ROSIS, Fiorella ; PELACHAUD, Catherine ; POGGI, Isabella ; CAROFIGLIO, Valeria ; DE CAROLIS, Berardina: From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. In: *International journal of human-computer studies* 59 (2003), Nr. 1-2, S. 81–118

[EWS09] EYBEN, Florian ; WÖLLMER, Martin ; SCHULLER, Björn: OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: *2009 3rd international conference on affective computing and intelligent interaction and workshops* IEEE, 2009, S. 1–6

[HSW+06] HEGEL, Frank ; SPEXARD, Torsten ; WREDE, Britta ; HORSTMANN, Gernot ; VOGT, Thurid: Playing a different imitation game: Interaction with an Empathic Android Robot. In: *2006 6th IEEE-RAS International Conference on Humanoid Robots* IEEE, 2006, S. 56–61

[SB19] SCHULLER, Björn ; BATLINER, Anton: *Intro Interspeech Computational Paralinguistics ChallengE*. Internet. http://www.compare.openaudio.eu/. Version: 2019. – Abgerufen: 17.08.2020

[SBB+19] SCHULLER, Björn W. ; BATLINER, Anton ; BERGLER, Christian ; POKORNY, Florian B. ; KRAJEWSKI, Jarek ; CYCHOSZ, Margaret ; VOLLMANN, Ralf ; ROELEN, Sonja-Dana ; SCHNIEDER, Sebastian ; BERGELSON, Elika u. a.: The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In: *Interspeech*, 2019, S. 2378–2382

[SSB+10] SCHULLER, Björn ; STEIDL, Stefan ; BATLINER, Anton ; BURKHARDT, Felix ; DEVILLERS, Laurence ; MÜLLER, Christian ; NARAYANAN, Shrikanth S.: The INTERSPEECH 2010 paralinguistic challenge. In: *Eleventh Annual Conference of the International Speech Communication Association*, 2010

[SSDR02] STEININGER, Silke ; SCHIEL, Florian ; DIOUBINA, Olga ; RAUBOLD, S: Development of user-state conventions for the multimodal corpus in smartkom. In: *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, 2002, S. 33–37

[SVE+09] SCHULLER, Björn ; VLASENKO, Bogdan ; EYBEN, Florian ; RIGOLL, Gerhard ; WENDEMUTH, Andreas: Acoustic emotion recognition: A benchmark comparison of performances. In: *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* IEEE, 2009, S. 552–557

[VA05] VOGT, Thurid ; ANDRÉ, Elisabeth: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: *2005 IEEE International Conference on Multimedia and Expo* IEEE, 2005, S. 474–477

[VA06] VOGT, Thurid ; ANDRÉ, Elisabeth: Improving Automatic Emotion Recognition from Speech via Gender Differentiaion. In: *LREC*, 2006, S. 1123–1126

[VAB08] VOGT, Thurid ; ANDRÉ, Elisabeth ; BEE, Nikolaus: EmoVoice—A framework for online recognition of emotions from voice. In: *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems* Springer, 2008, S. 188–199

[VSS+14] VALSTAR, Michel ; SCHULLER, Björn ; SMITH, Kirsty ; ALMAEV, Timur ; EYBEN, Florian ; KRAJEWSKI, Jarek ; COWIE, Roddy ; PANTIC, Maja: AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In: *Proceedings of the 4th International Workshop on Au-*

dio/Visual Emotion Challenge. New York, NY, USA : Association for Computing Machinery, 2014 (AVEC '14). – ISBN 9781450331197, 3–10

[WBS+11] WOLLMER, M. ; BLASCHKE, C. ; SCHINDL, T. ; SCHULLER, B. ; FARBER, B. ; MAYER, S. ; TREFFLICH, B.: Online Driver Distraction Detection Using Long Short-Term Memory. In: *IEEE Transactions on Intelligent Transportation Systems* 12 (2011), Nr. 2, S. 574–582

[Wik] WIKIPEDIA: *Speech corpus*. `https://en.wikipedia.org/wiki/Speech_corpus`. – Abgerufen: 24.08.2020