JÖNKÖPING UNIVERSITY

Exploring AI-Based Emotion Recognition in Swedish: Speech, Text, and Vocal Markers

Author(s): Sara LJUNG & Janna HAKKARAINEN
Main subject area: Computer Engineering
School: School of Engineering
Date: March 2025

This final thesis has been carried out at the School of Engineering at Jönköping University within Computer Engineering. The authors are responsible for the presented opinions, conclusions and results.

Examiner: Neziha AKALIN
Supervisor: Garrit SCHAAP
Scope: 15 Credits
Date: 2025-03-09

Contents

Abstract

Keywords:

Introduction

This thesis aims to explore emotion recognition and its effectiveness in the Swedish language. With the rapid advancement of the technology industry and artificial intelligence, emotion recognition has started to play an increasingly important role in the enhancement of human-computer interactions. These areas hold potential to transform and develop several important fields, but there are still challenges in the field. Much of the research has been focused on specific languages, notably English. This research focuses on emotion recognition across two distinct modalities in Swedish, speech-based emotion recognition and text-based emotion recognition and aim to contribute to broadening the field of emotion recognition in a non-English language.

1.1 Background

Emotion recognition has attracted increasing attention with the rapid advancement of technology and artificial intelligence. Emotions are experienced by all humans but are difficult to define precisely. They are an internal experience that are foundational to our sense of identity, our relationships, and moral judgement. Scientists have faced challenges in the effort to characterize how emotions are communicated. Emotions are internal but also expressed externally through voice and movements of the body. They are not only communicated through the words we say, but also how we say them. Tones of the voice is a source of varied emotional expressions where its states may alter patterns in vocalizations. It is considered that various emotion-related physiological changes influence acoustic features such as pitch, tempo, pitch variability, and loudness in the speech (Oatley et al., 2019). Scientists have developed different techniques to determine states of emotions as well as opinions, a field of Natural Language Processing (NLP) that intersects artificial intelligence, computer science, and linguistics (Kansara et al., 2020). With the development of Artificial Intelligence several techniques have accelerated in the recent years, including for NLP, even if its origins back to the 1950s when questions about whether a machine could learn and think to interact with humans (Núñez et al., 2024). NLP has remained as a significant contributor of AI. Some of the active research areas in the NLP domain is Machine Translation, Chatbots, recognizing speech, text summarization, and sentiment analysis (S. Kusal et al., 2023). Figure 1.1 demonstrates the different subdomains of AI.

Sentiment analysis is a computational branch in NLP that utilizes the detection and evaluation of people's emotions, opinions, and moods based on text, speech, facial expression, etc., without analyzing these feelings (Ermakova et al., 2023). The rise of sentiment analysis is associated with the growth of social media, which has generated vast amounts of digital option data recorded in digital forms. Since the early 2000's, the field has become one of the most researched parts in NLP (L. Zhang et al., 2018), expanding beyond computer science to fields like finance, marketing, political- and health science. Accordingly, sentiment analysis is valuable across different areas of society. Sentiment analysis is utilized in the popular index called the happy planet index (HappyPlanetIndex, n.d.), measuring sustainable well-being of different countries, even if it only can observe three feelings, positive, negative, or neutral. The happy planet index checks the happiness level calculated from a particular country, where emotion

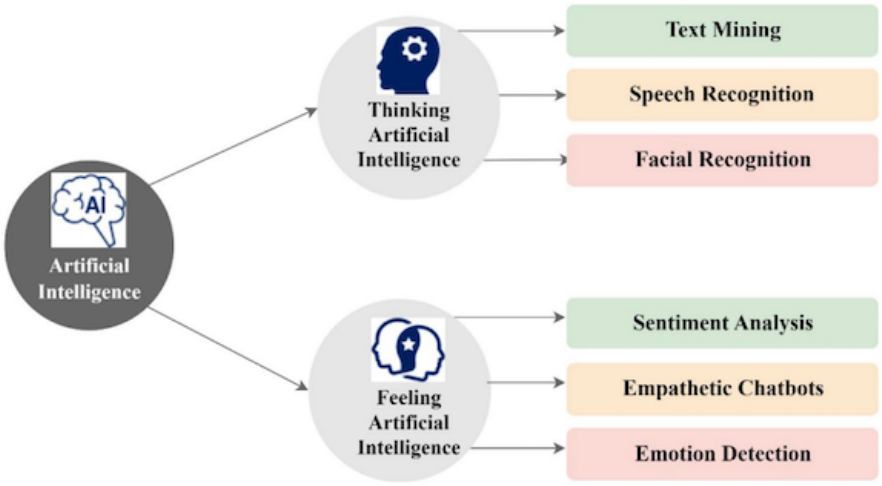


Figure 1.1: *Subdomains of AI (S. Kusal et al., 2023).*

detection is used with sentiment analysis (Madhuri & Lakshmi, 2021). With the evolution of deep learning networks, emotion detection has advanced. Sentiment analysis identifying positive, negative, or neutral states have progressed into recognizing the six basic emotions; joy, sadness, anger, disgust, fear, and surprise in text (Safari & Chalechale, 2023). The emotions categorization fluctuate depending on the research. These basic six where determined by Paul Ekman (S. Kusal et al., 2023; Oatley et al., 2019) who determined that these six fundamental emotions is shared in people of different cultures, characterized by facial features. However, Ekman's classification was made over 20 years ago when no agreement about what emotions should be considered as existed. Today, the agreement about evidence for universal emotional signals and evidence for five emotions is robust: anger, disgust, sadness, happiness, and fear (Ekman, 2016).

Emotion recognition is accomplished either through text or speech, with numerous studies in both areas. Text-based emotion detection is a complex field that requires a very clear understanding of the context. Today this technique is frequent in several fields, such as human-computer interaction, big data, data mining, e-commerce, online tutorials and psychology (Madhuri & Lakshmi, 2021). The first emotion dataset, The International Survey on Emotion Antecedents and Reactions (ISEAR) was developed in 1997, after it was stated that computers need the ability to interpret, express, and identify emotions if we want true intelligent computers and be able to communicate with them naturally (S. Kusal et al., 2023).

Emotion recognition from textual data is important in various domains such as customer reviews, social media analysis, public monitoring, and conversational agents. A systematic review (S. Kusal et al., 2023) shows that Deep Learning models outperform traditional Machine Learning models due to their ability to capture contextual dependencies. The review further demonstrates the highest accuracy is shown by transformer-based models such as bidirectional encoder representations from transformers (BERT), highlights challenges such as small or imbalanced datasets that can affect the model reliability, and notes that multimodal approaches with text, speech, and images improve emotion recognition. BERT-technology is leveraged with a model that outperformed other text-based emotion recognition models, with 76% validation accuracy (Madhuri & Lakshmi, 2021). However, text-based emotion detection (TBED) has challenges with identifying hidden emotions, and adapting to diverse languages. Datasets based on different languages than English, as Arabic and Hindi, are tested in a study (Maruf et al., 2024) that identifies challenges as limited resources for non-English languages. The authors underscore the potential of TBED but notes limitations as it is no universal solution for

challenges like sarcasm, dynamic emotions, and cultural variances.

Emotion detection research progressed with Speech Emotion Recognition (SER) (S. Kusal et al., 2023). It has shown that hearers can evaluate five emotions in speech-prosody, anger, happiness, sadness, fear, and tenderness, with 70 percent accuracy (Oatley et al., 2019). Speech emotion recognition focuses on how something is said rather than the words themselves. Acoustic features like amplitude, formants, and pitch help classify emotions. A common approach uses Mel-frequency cepstral coefficients (MFCCs), which capture vocal patterns and have proven effective in detecting emotions (Thaler et al., 2024). Those features offer invaluable insights into the subtle emotional expressions conveyed through speech, assisting the complicated process of emotion recognition. They are typically categorized into three primary groups: prosodic features, voice quality features, and spectral features (Lian et al., 2023).

Several studies distinguish different emotions through vocal features. Already in 2005, Lee & Narayanan investigated automatic recognition of emotions, especially positive vs. negative from spoken dialogs. In that study, acoustic, lexical, and discourse information were combined to enhance emotion detection and move beyond traditional acoustic-only ways. The authors analyzed acoustic features, lexical features, and discourse features. Linear Discriminant Classifiers were used and resulted in good performance for acoustic and lexical information (C. M. Lee & Narayanan, 2005). In 2014, one study (Bänziger et al., 2014) showed that regular people can rate how voices sound reliably, when actors pretend to feel emotions, sometimes better than machine sound measurements for pitch and volume. The human voice feature rating worked better than technical measures, especially for happy emotions. These acoustic features in our voice enable emotion recognition through speech using Deep Learning, which have many advantages for Speech Emotion Recognition over traditional sentiment methods. Deep Learning has the capability to detect complex structures and different features without requiring manual feature extraction (Khalil et al., 2019). SER's goal is identification of emotions in speech, unrelated to the semantic content (S. D. Kusal et al., 2024). Figure 1.2 represents a SER system.

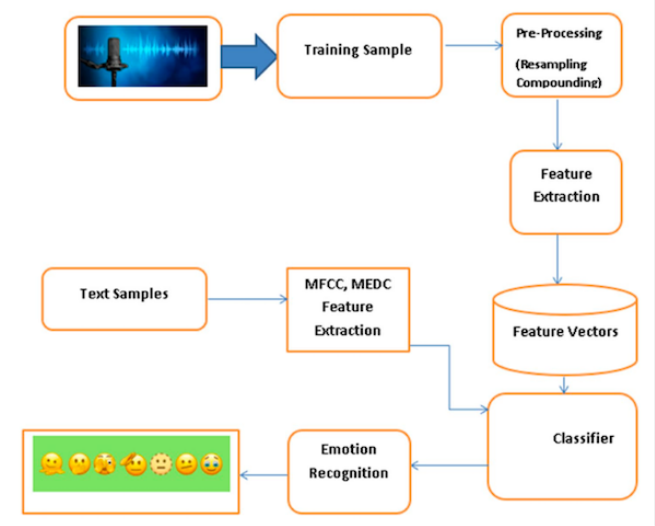


Figure 1.2: *Block diagram of SER (Tyagi & Szénási, 2024).*

In recent years, speech emotion recognition has emerged as a significant and complex research area within pattern recognition, speech signal processing, and artificial intelligence. Its growing importance is driven by its applications in human-computer interaction. Specifically, SER systems enable emotion-aware interactions through speech, eliminating the need for traditional input devices. This advancement has led to the development of intelligent affective services in fields such as call centers, healthcare, surveillance, and affective computing (S. Zhang

et al., 2021). The accuracy of models tested in recent years have improved significantly. Several studies conducted in the last year's show emotion detection accuracy results over 90% (Adebiyi et al., 2024; Praseetha & Joby, 2022; Rahman et al., 2024). A research paper (Abbaschian et al., 2021) compares several studies, detailing the results, databases used, and the time periods during which the studies were conducted. The reported accuracies depends greatly on which model was used, as well as the tested dataset. According to Abbaschian (2021), in 2014, one model showed 54.3% accuracy for the IEMOCAP dataset. The same dataset was used for the LSTM model in 2017 and 2018 with 63.5% respectively 64.93% accuracy. When the same dataset was tested with a CNN-model, the accuracy increased to 82.8% in 2019.

A study (Juslin et al., 2018) concluded in 2018 analyzed 1,877 voice clips from 23 datasets to compare spontaneous and posed emotions. Their findings highlighted key differences:

- Spontaneous expressions were rated as more genuine than posed ones, even when intensity was controlled.
- Posed expressions were more intense, but intensity alone did not fully explain perceived authenticity.
- Acoustic differences were small but present, mainly in pitch range, speech rate, and voice intensity.
- Highly intense spontaneous emotions conveyed emotions as clearly as posed ones, suggesting that emotion intensity plays a role in perception.

These findings suggest that posed and spontaneous emotions are not interchangeable and that datasets used in SER research must account for these differences to ensure reliable models.

One study (Rathi & Tripathy, 2024) presents a comprehensive review of SER which explores the impact of speech data corpora selection and speech feature extraction on the accuracy of emotion classification. Publicly available speech datasets from 2014-2023 are systematically analyzed and categorize speech datasets and speech features to evaluate their impact on the accuracy of emotion recognition. The study by Rathi & Tripathy (2024) includes the six emotions: happiness, anger, sadness, surprise, fear, and neutrality. According to Scherer, Frühholz & Belin (2018), most studies focus on only four broad emotion categories: anger, fear, happiness, and sadness, which is seen as a limitation by the authors (Scherer et al., 2018). However, emotion classification fluctuates in different studies. A database has been conducted that unravel this limitation. The database is called GoEmotions (Demszky et al., 2020) which is a big, detailed, and reliable dataset for recognizing 27 emotions in text. The researchers (Demszky et al., 2020) used a BERT model and got an average F1-score of 0.46 for these emotions, best for gratitude (0.86) and worst for grief (0.00). The model performed a score of 0.64 for the simpler 6 emotion grouping. The authors conclude that the model still needs advancements for tricky feelings but suggest the model as a good starting point. This research is included in the research that the AI-model Hume.ai builds upon (Hume, n.d.-a), which is important to acknowledge since it might affect biases. Most researchers focus on the fundamental and widely recognized emotions. The six fundamental emotion model by Paul Ekman is the most widely used for both text-based and speech-based emotion recognition (Maruf et al., 2024).

Datasets are essential for data-driven learning, enhancing model performance and robustness. Emotion recognition datasets are classified based on signal type, including speech (textual/audio), visual, physiological signals, and multimodal data. Speech emotion datasets are further classified by how they are collected:

- Performer-based datasets feature emotions acted by trained performers.

- Induced datasets capture emotions in controlled environments, making them less expressive but closer to real-life emotions.
- Natural datasets come from real conversations (e.g., public dialogues, call centers) and contain authentic emotional variations but are more challenging due to background noise and limited availability (Cai et al., 2023).

Most research in the field of SER investigates the accuracy of different Deep Learning algorithms on diverse public datasets. A review (Rathi & Tripathy, 2024) analyses 93 research papers where IEMOCAP and RAVDESS are among the most widely used datasets, chosen by 35.83% and 21.50% of researchers, respectively. Their popularity stems from well-annotated data and diverse range of emotional expressions, which enhance model performance. Figure 1.3 presents the usage of the datasets in this review. The paper (Rathi & Tripathy, 2024)

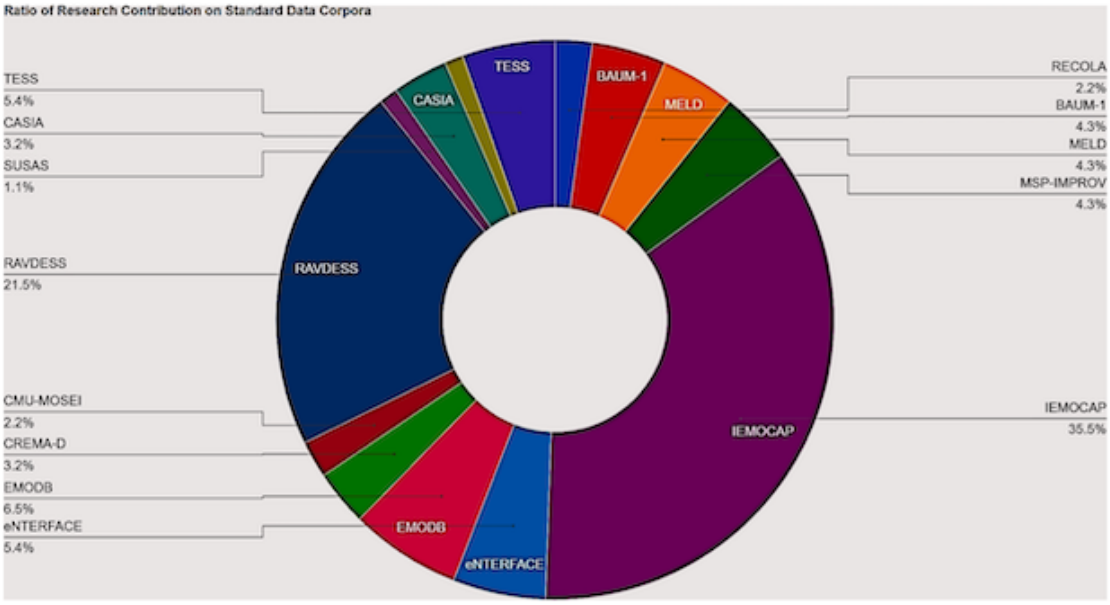


Figure 1.3: *Research contributions of standard data corpora in percentage*(Rathi & Tripathy, 2024).

includes varying datasets in terms of linguistic diversity, recording conditions (acted, induced, or natural speech), and the number of emotional categories. The results show that the choice of dataset significantly affects the model's performance. Key features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, intensity, prosodic cues, and spectral properties impact SER accuracy significantly. The authors discuss the balance between realistic datasets and classification accuracy, noting that natural speech is more difficult to classify due to its high variability and noise in natural speech.

The number of natural datasets is relatively limited (Cai et al., 2023), and many research papers test on acted datasets. An empirical analysis (Ahammed et al., 2024) of Machine Learning algorithms across diverse datasets, concluded in 2024, demonstrates high-accuracy SER system using an SVM classifier and advanced feature extraction techniques. The model performed exceptionally well across multiple datasets (RAVDESS, TESS, SAVEEE, and combined dataset), with the combined dataset achieving perfect scores (100%) in accuracy, precision, and F1-score. Two of these datasets, RAVDESS (Livingstone & Russo., 2019) and TESS (Pichora-Fuller & Dupuis, 2020), are acted and the SAVEE (of Surrey, n.d.) dataset was recorded from four male, postgraduate students and researchers. All three datasets are in recorded in English. The TESS dataset was used in a study from 2022 (Praseetha & Joby, 2022) that leveraged other techniques, with one original dataset and one augmented dataset. Achieved accuracy was

93% and 97% for the augmented dataset, resulting in a significant improvement for the processed data. A different study (Alroobaea, 2024) from 2024 investigates transformers for SER in cross-corpus scenarios. The recordings are preprocessed to remove noise, and augmented techniques are applied. Three datasets were used, two the same as the previous study, SAVEE and RAVDESS, with the Berlin Database of Emotional Speech, Berlin EMO-DB, including ten professional speakers, was additional. A combined cross-corpus dataset of the three sets were used to test generalizability across different datasets. This proposed transformer-based model outperformed traditional deep learning methods. The model showed high accuracy results with 95% for SAVEE, 94% for RAVDESS, 97% for EMO-DB, and 97% for the cross-corpus dataset. However, this model was not evaluated on spontaneous speech datasets. Comparing these three studies from 2022 to 2024 reveals a clear trend of increasing accuracy over time. A 2019 review (Khalil et al., 2019) reported significantly lower results for the SVM, with accuracy of 74% for anger, 70% for happiness, and 93% for sadness. However, the same study found that a Deep Convolutional neural network (DCNN), achieved substantially higher accuracy, with 99% for both anger and happiness, and 96% for sadness. Both Emo-DB and SAVEE datasets were included, as well as IEMOCAP which also is an acted dataset recorded in English.

Datasets for Text Based Emotion Detection (TBED) are discussed in (S. Kusal et al., 2023), where the authors state that researchers can use publicly available datasets or create their own. Publicly available, useful datasets with reliable annotating include several sets based on various data, from stories, publications, news, social media texts, to reviews on movies. The datasets emotion-classification reaches from the basic six emotions to GoEmotions set that include 27 emotions. Many datasets are based on social media, including casual writing style which is a big challenge. The use of short messages and informal language has limited research. Human emotion expressions and the texts conveying them are ambiguous and subjective, additionally, emotions are multifaceted with varying expressions. Therefore, the authors claim that human mapping is important. Self-labeled emotion datasets are tested in a study (S. J. Lee et al., 2023) that uses a Transformer Transfer Learning (TTL) model trained on a dataset of over 3.5 million tweets with self-reported emotion hashtags. The model is tested against 10 prior published datasets and achieved highest score on annotator-rated sets (0.87), it also performed well on self-reported sets (0.79) which demonstrates generalizability.

The promising development of emotion recognition has been adapted in research for other areas than computer and machine learning science. SER is beneficial in translating languages, interactive courses and tutorials held online where the student's emotional state can be understood to help the machine make decisions on how to present the course. It can be implemented in vehicles' safety structures to recognize the driver's emotional state and therefore prevent accidents (Abbaschian et al., 2021). The mental health sector has great potential to benefit from emotion recognition. Several studies (DeSouza et al., 2021; Drougkas et al., 2024; Simcock et al., 2020; Singh, 2023) have investigated how AI-based emotion recognition can be used to help therapists and psychiatrics diagnosing and identifying potential mental illnesses. One of the studies (DeSouza et al., 2021) demonstrates how leveraging speech and text analysis with NLP can help detect late-life depression. It showed consistent alterations in individuals with depression, including acoustic features such as pitch, speech rate, pause duration, and word choice. The authors suggest that automated speech analysis can identify late-life depression as well as predicting depression severity with accuracy of 86-92%. Multimodal machine learning with a combination of text and audio analysis has been explored to identify indicators for various mental health disorders. The research (Drougkas et al., 2024) compared unimodal text models, which showed strong results but was outperformed by the multimodal models, especially for identification of the presence of markers. The multimodal models achieved accuracy up to 86.73%. This study highlights the importance of machine learning integration in mental health diagnostics.

1.2 Problem Description

Despite significant progress in speech emotion recognition, there are limitations in current research. For instance, emotional voice samples are usually obtained from actors portraying emotions using scripted speech. These acted expressions tend to be more intense and exaggerated than naturally occurring emotions. However, this method risks overemphasizing obvious emotional cues while missing subtle ones. It is argued that such portrayals reflect social norms more than genuine physiological responses, although all public expressions may involve some degree of performance (Scherer et al., 2018).

The way emotional speech data is collected depends on the design and purpose of the SER system. As datasets shift from acted emotions to more spontaneous or real-life emotions, emotion recognition becomes more challenging. Many researchers prefer acted emotion datasets because they offer a wide range of emotions and large amounts of data (Rathi & Tripathy, 2024). Induced datasets are collected by constructing an artificial emotional situation, without the knowledge of the performer or speaker. This results in a more naturalistic database, but issues regarding ethics may apply, since the speaker should know they have been recorded for research (Khalil et al., 2019).

Estimation of emotions from spontaneous speech is a challenging task. Most studies test models on acted datasets (Ahammed et al., 2024; Alroobaea, 2024; Khalil et al., 2019; Praseetha & Joby, 2022). The primary reason for the concentration on acted SER tasks is that acted emotions can be easily performed in a controlled laboratory setting, often resulting in high SER accuracy. However, these emotions tend to be exaggerated and may not accurately reflect how emotions are expressed in real-world situations. Consequently, detecting spontaneous emotions in natural environments is significantly more complex and challenging compared to recognizing acted emotions (S. Zhang et al., 2021). Figure 1.4 demonstrates the difficulty level for varying settings.

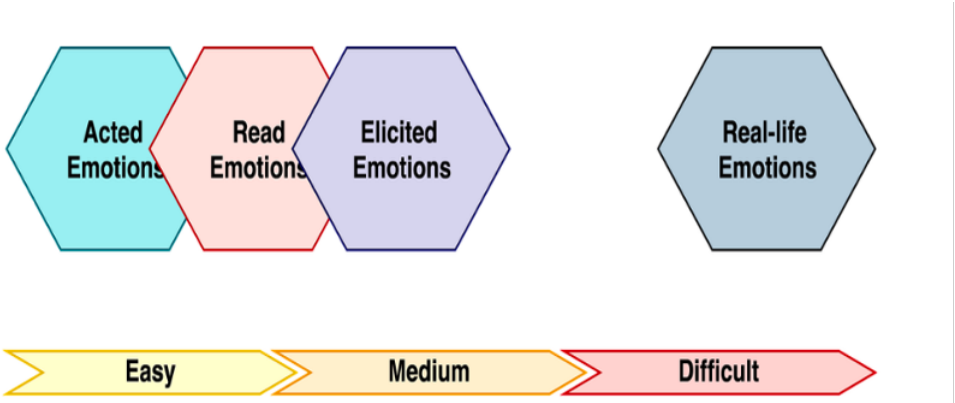


Figure 1.4: *Emotion recognition databases and their difficulty level (Khalil et al., 2019).*

One study from 2019 (Milner et al., 2019) using four English speech datasets, two acted (eNTERFACE and RAVEDESS), one elicited (IEMOCAP), and one natural (MOSEI), to see how emotions can be recognized across them. The researchers in this study tested different training methods to explore whether mixing acted and natural data works. They found that acted data does not easily help with natural emotions unless the system is tweaked. The differences in real and acted voices are shown in other studies as well (Juslin et al., 2018). Real ones sound more genuine and have some unique features, even if the basic emotion patterns are alike. When emotions are strong, real voices can show clear emotions just as well as acted ones, which supports an idea that strength matters more than whether it is real or fake.

Speech emotion recognition (SER) plays a crucial role in identifying emotions from speech. However, because emotions are complex and often overlap, extracting accurate emotional features from speech remains challenging. As SER research advances, cross-cultural emotion recognition is becoming increasingly important. While people from different countries and cultures have distinct ways of expressing emotions, they can still interpret tone and attitude even without understanding the spoken language (Cai et al., 2023). Behind the development of the SER model Hume.ai, a study (Brooks et al., 2023) on cultural differences in vocal bursts has been conducted. It is suggested that 24 acoustic dimensions of vocal expression have emotion-related meanings that distinguish them. With participants from China, South Africa, Venezuela, India, and the USA, the authors concluded that 79% of complex vocal modulations were preserved across these countries. Research shows that speech acoustics change based on emotion, intensity, and cultural background. Studies comparing tonal and non-tonal languages (Scherer et al., 2018) suggest that features like pitch and speech rate are influenced not only by physiology but also by cultural differences in emotional expression. It is debated whether emotional expressions are universal or culturally influenced is particularly relevant to vocal emotion recognition, as languages differ in phonemic structure, intonation, and rhythm. If emotion expression differs between languages, emotion recognition across cultures may also be affected. One study from 2001 is referred in (Scherer et al., 2018), where listeners from seven countries recognized emotions in German-accented speech. While accuracy averaged 66%, recognition rates varied significantly, from 74% in Germany to 52% in Indonesia, indicating that cultural and linguistic differences impact vocal emotion perception.

The Swedish language is not widely spoken and therefore very limited research has been concluded on the Swedish speech. One study (Ekberg et al., 2023) investigated Swedish emotion recognition through feature extraction and concluded that emotions in Swedish speech have unique sound patterns. The results from this study differed from previous research in some spots, which could be due to the language itself. The emotion results showed that surprise is a very distinct emotion, but happiness and anger sound alike, which could confuse listeners.

There are studies that explore SER for different languages. EMO-DB is a Berlin Database of Emotional Speech that is used in several studies with high accuracy (Alroobaea, 2024; Jahangir et al., 2022; Khalil et al., 2019; Zhao et al., 2019). SER across different cultures and languages are explored in an article (Pandey et al., 2023) that tackles the challenge of recognizing emotions in speech from five languages: English, German, Persian, Hindi, and Telugu, since emotions can sound different depending on the speaker and language. A language model was pre-trained on four languages to identify language-specific cues and applied to the remaining one, Persian. Two different methods were tested, one excelled with known languages (92.24% unweighted accuracy for English), and resulted in 63.09% unweighted accuracy for the new language. The second method had similar results, with up to 95% accuracy for pre-defined language but struggled with unseen languages. Emotional variation due to culture and speech is mentioned in this study as well, for example, Hindi anger might sound different than German anger.

Similarities as well as differences in cultural and linguistical vocal features affecting emotions are declared depending on the research. Studies using common databases on other languages than English, such as the German EMO-DB (Alroobaea, 2024; Jahangir et al., 2022; Khalil et al., 2019; Zhao et al., 2019), cannot generalize linguistic performance since many of the models in these studies are trained on these databases. The studies highlight that models trained on specific databases like the Berlin EMO-DB or pre-trained on a limited set of languages struggle to achieve high accuracy when applied to unseen languages, such as Persian, due to cultural and linguistic variations in emotional expression. Therefore, these models cannot be considered as generalizable to other languages, as their performance is dependent on the linguistic and cultural conditions of the data it was trained on.

1.3 Purpose and Research Questions

The advancement of artificial intelligence (AI) has significantly improved the ability to recognize human emotions, both through speech and text. This offers transformative potential across domains such as mental health, education, and human-computer interaction. Speech Emotion Recognition (SER) and Text-based emotion detection (TBED) have become key areas within the field of Natural Language processing (NLP), leveraging deep learning to interpret different emotional cues with increasing accuracy. However, despite these advancements, significant challenges remain in ensuring that emotion recognition systems are robust, culturally inclusive, and reflective of real-world emotional expressions. Much of the existing research relies on acted datasets, which may underperform when it comes to subtle, spontaneous emotions in everyday contexts, and there is a notable gap in understanding how these models perform across diverse linguistic and cultural situations, such as the Swedish language. The number of studied languages is not that broad, and the studies on accuracy for a new language implies that more research on the generalizability to other languages is essential. Furthermore, while speech and text offer complementary perspectives on emotions, their alignment with individuals own perceptions of their emotions remains unexplored.

This study aims to address these gaps by investigating the performance of AI-driven emotion recognition systems in a specific, but relevant, context: Swedish speech and its transcribed textual content. By focusing on Swedish – a language with limited prior research in SER – this thesis seeks to contribute to a broader understanding of how linguistic and cultural factors can influence emotion recognition, which is applicable to multilingual understanding for emotion recognition for different languages. Additionally, the integration of speech and text analysis gives an opportunity to explore multimodal approaches. The alignment between AI-generated emotion labels and self-reported emotions is an overlooked area. Even if emotions are hard to define, and might be difficult for some individuals to assess, it is interesting to investigate the alignment between them. Publicly available AI models and APIs, commonly used in real-world applications, are rarely tested against such subjective human data, which makes this comparison novel and interesting to investigate.

The purpose of this thesis is therefore to evaluate how an AI model recognizes emotions from Swedish speech, to assess whether its transcribed textual content can convey emotional states independently and compare these AI-generated labels with self-reported emotions from Swedish speakers. By addressing the specific challenge of emotion recognition in a less-studied language, the study contributes to the broader scientific discussion on emotion-recognitions generalizability. The study will provide insights into alignment between speech and text modalities, cultural emotional expression, and the alignment between AI outputs and human experience.

To explore speech emotion recognition for Swedish speech, vocal markers from Swedish speech recordings will be extracted and compared to a prior study (Ekberg et al., 2023) on Swedish speech vocal markers for emotions. With the usage of this research, the performance of an AI model for Swedish can be compared, and therefore the first research question of this study is:

> *[1] How does AI model for speech recognition compare to research on vocal markers for emotions in Swedish speech?*

Text-based emotion recognition is a commonly used research field, but mostly for English text. To address this, it is interesting to assess whether transcribed Swedish speech can reveal emotions independently, which leads to the second research question:

> *[2] Can we understand the emotions from the textual content of the speech, with the same data as in RQ1?*

The perception of emotions is a complex field, with few studies made on the alignment between machine-labeled emotions and human-percived emotions. To undertake this, its comparison will be explored in the third research question:

> *[3] How do AI-generated emotion labels (speech & text-based) compare to self-reported emotions?*

1.4 Scope and Limitations

This section defines what the study includes and excludes. It focuses on AI-based emotion recognition in Swedish speech and text, considering its constraints in design and resources. The study explores challenges like reliance on acted datasets, language differences, and the alignment between AI predictions compared to self-reported emotions. Since this is an exploratory thesis, some limitations are recognized but accepted for feasibility.

1.4.1 Scope

The study evaluates AI-driven emotion recognition in Swedish, a language with little prior research in this area. It analyzes emotions from about 15 Swedish-speaking participants through short interviews designed to evoke natural emotions. The study includes:

- **Speech-based analysis:** Using Hume.ai (Hume, n.d.-a), an API with AI-based emotion recognition in speech for AI-based speech emotion recognition.
- **Text-based analysis:** Using NLP Cloud (Cloud, n.d.), an API utilizing AI to transcribe speech and detect emotions from text, see Theoretical Framework.
- **Comparison with self-reports:** Participants rate their emotions on a scale of 1-6 (1 = very weak, 6 = very strong), compared to AI-generated labels.

To keep this study manageable, it focuses on two emotions in the interviews (one positive, one negative) and uses recordings made in a quiet environment using a microphone. The analysis relies on existing AI tools and API's (Cloud, n.d.; Hume, n.d.-a) and software for voice feature extraction, without developing new models. A mixed method is used, combining AI outputs with qualitative insights.

1.4.2 Limitations

Several factors limit the study's depth and generalizability:

- **Small dataset:** With only around 15 participants, results may not apply to all Swedish speakers.
- **Subjectivity of self-reports:** Participants' emotions may be influenced by personal biases or recall inaccuracies.
- **Limited emotion categories:** Focusing on only two emotions excludes other relevant emotional states. If other emotions occur, they will be acclaimed but distinguished.
- **Use of pre-trained AI models:** The study relies on third-party tools without modifying their algorithms, which may introduce biases.
- **Acted vs. Spontaneous emotions:** Interviews may not fully capture natural emotional responses, since they are partially induced.

- **Lack of psychological expertise:** The design of emotion-eliciting scenarios may not be optimal, even if it's based on research.
- **Language limitations:** Findings may not apply beyond Swedish, as most SER research is based on English datasets. It still measures generalizability.
- **No physiological measures:** The study does not include biometric data, which could provide additional insights.

These limitations are necessary compromises for feasibility within the study's timeframe and resource constraints. The study does not aim to develop new AI models or solve all SER challenges. Instead, it provides initial insights into Swedish emotion recognition, tests existing AI tools, and identifies areas for future research.

1.5 Disposition

From here, the report is structured as follows:

Method and Implementation: This section introduces explanatory mixed method, experimental approach used to answer the research question. It describes the experimental setup, data collection process, methods of analysis, and considerations regarding validity and reliability.

Theoretical Framework: This chapter explores the underlying theories relevant to this study. It provides an overview of Natural Language Processing (NLP), Speech Based Emotion Recognition (SER), Hume.ai, Praat Parselmouth, Text-Based Emotion Recognition (TBED), NLP Cloud, and theories behind vocal markers in speech. The experiment is explained with relevant research for the interviews used for this study.

Time Plan: This section outlines the remaining phases of this thesis, detailing the planned milestones and schedule.

Method and Implementation

This chapter outlines the work process for this study, designing a methodical approach to investigate emotions in Swedish speech using both AI-based analysis and self-reported data. The chapter describes the study's approach and design, justifies methodological decisions, provides details regarding data collection and analysis procedures, and addresses validity and reliability considerations.

2.1 Approach and design

This study adopts an explanatory sequential mixed method approach, which integrates both quantitative and qualitative approaches in a structured sequence. The study first collects and analyzes quantitative data, as AI-generated emotion labels and self-reported emotions, and then qualitative interprets the results to explore alignment and divergence. This approach ensures a systematic, layered analysis rather than pure comparison (Creswell & Creswell, 2023).

The study follows a deductive research approach, as it builds upon existing theories of emotional expression in text and speech. The AI models will be tested and compared to established findings. Instead of developing new theories, the study aims to evaluate whether AI-based emotion recognition methods align with each other, prior research on vocal emotion markers and self-reported emotions for Swedish speech. This is classified as an experimental study, as it involves a controlled setting where participants are asked questions on predefined emotional recall scenarios. It does not manipulate independent variables in a traditional experimental way (Creswell & Creswell, 2023), instead observes and analyzes the natural emotional responses provoked through structured questions (Bryman et al., 2022).

The study evaluates AI-generated emotion labels from speech compared with existing research on vocal markers, text-based emotion recognition and self-reported emotions. The self-reports serve as a reference point and not a ground objective truth, to acknowledge the subjective nature of emotional perception.

2.2 Data Collection

The study involves participants for semi-structured interviews where they respond to predefined scenarios to provoke emotions. Ethical considerations, such as informed consent and anonymization, are followed firmly to ensure participant well-being. In the first phase, interviews are collected for analysis.

Approximately 15 Swedish speakers, primarily acquaintances to the researchers, are recruited via invitations. Interviews take place in a quiet room and last for 5-10 minutes including 2-4 scenarios that have been pilot tested for effectiveness, and breaks between the scenarios. Participants complete a self-assessment after each scenario. To minimize acted behavior, the participants are not pre-informed about the feelings that are aimed to be provoked through the scenarios.

Participants are asked open-ended questions designed to bring out previously lived through personal experiences of anger and happiness. The questions about anger are focused on previous

experiences of unfair treatment and frustration, while the questions related to happiness explore moments of pride and unexpected joy. The semi-structured format allows for follow-up questions based on participant responses, to bring out as much emotion as possible. Although there will be a few different predefined scenarios for the participants in the interviews to choose from to maintain freedom in the participants emotional expression, two example questions for both emotions in the interview are:

- Can you remember a time when you were treated unfairly by someone? What happened and how did you react?
- Tell me about a moment when you were very proud of yourself. What did you do and how did you feel?

The study adopts a mixed-method interview approach (Bryman et al., 2022), where the qualitative data from speech is transformed into quantitative AI-generated emotion labels, to enable comparison in a structured way (Creswell & Creswell, 2023). Audio is recorded and analyzed using two emotion recognition models: Hume.ai to extract speech-based emotional labels, and NLP Cloud to transcribe the speech and extract text-based emotional labels. The recordings are preprocessed to reduce background noise. The same dataset is used for all research questions to ensure consistency and all participants remain anonymous.

The diagram in Figure 2.1 visualizes the multi-modal pipeline used in this study. The interview audio files are processed through three primary channels: speech-to-text transcription via NLP Cloud, Speech emotion recognition via Hume AI and acoustic feature extraction via Praat. These channels represent two main pipelines. The entities presented in yellow are prevalent in both pipelines, where the audio recordings are analyzed with Hume AI, the output is filtered to the 6 emotions analyzed in this study. The pipeline illustrated in green represents the analysis to answer research question 1. The vocal markers are extracted using Praat Parselmouth. The features that are chosen are based on previous research, see 3.5, Theoretical Framework - Vocal Markers, where pitch, intensity/loudness, formant frequencies (F1, F2, F3), HNR, jitter and shimmer have distinguished values for certain emotions. To compare the extracted data with Hume AI, these values are clustered into emotion groups. Data from speech analysis and vocal markers are combined to statistically analyze the results for RQ1. The pipeline used to answer the second and third research question is presented in orange. Interview audio is processed the same way as for RQ1 but extended with normalization for the Hume values to enable comparison with outputs from NLP Cloud and self-assessment. For text-based emotion recognition, the recording is transcribed before text-analysis is composed. Results from speech and text prediction are combined with the self-assessment scores to answer RQ2 and RQ3.

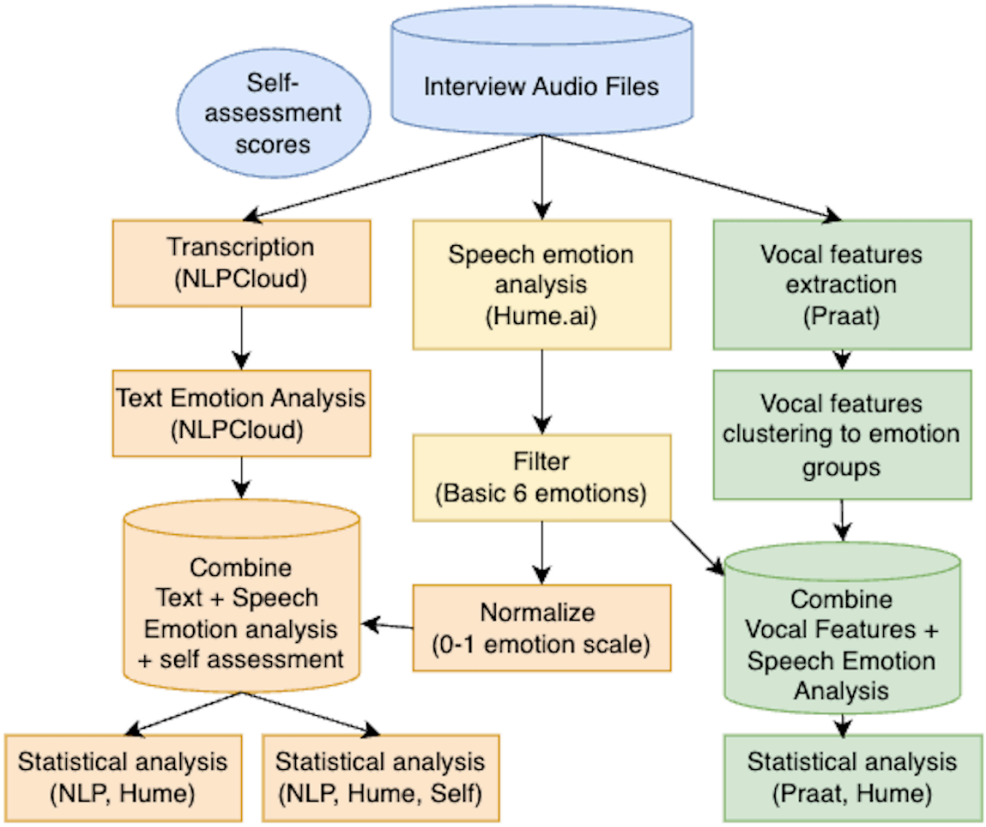


Figure 2.1: The multi-modal pipeline used in this study.

2.2.1 Research Question 1

How does AI-model for speech emotion recognition compare to research on vocal markers for emotions in Swedish speech?

To answer this question, speech recordings are collected from participants as they describe emotionally charged experiences. AI-based emotion recognition using Hume.ai, are used for AI-based Speech Emotion Recognition. Voice feature extraction from the recordings is made, to compare to AI-labeled emotions with known vocal markers from existing Swedish emotion research (Ekberg et al., 2023).

2.2.2 Research Question 2

Can we understand the emotions from textual content of the speech, with the same data as in RQ1?

To answer the second question, the recorded speech is transcribed and analyzed for emotion recognition using NLP Cloud's emotion recognition to assess the emotional content of speech transcripts. The text-based AI labels are compared with speech-based AI labels to determine whether emotion is preserved in textual content alone.

2.2.3 Research Question 3

How do AI-generated emotion labels (speech & text-based) compare to self-reported emotions?

For the third question, participants complete a self-assessment survey after each interview segment, where they rate their emotional state on a 1-6 scale (1 = very weak, 6 = very strong)

for relevant emotions. The self-reported emotions are compared with AI-generated labels from both speech and text models to analyze agreement and divergence. The results are clustered as agreements, partially agreements, and disagreements across methods.

2.3 Data Analysis

To systematically evaluate the agreement between different emotion detection methods, a combination of statistical analyses and visualizations was applied for speech-based AI, text-based AI, vocal markers, and self-reported emotions. The analysis aimed to assess the alignment with established vocal marker research and subjective human perception, where identification and categorization of differences where analyzed.

2.3.1 Statistical Analysis Approach

RQ1: Vocal Markers vs. Speech-Based Emotion Recognition

The first step involved applying a custom emotion categorization function based on standardized distances from the Swedish research on vocal markers (Ekberg et al., 2023), to cluster vocal features into emotional categories. This approach adapted absolute standardized deviations across key acoustic features, as detailed in section 3.5 and 3.7.

Following the initial clustering, correlations were calculated between:

- Individual vocal features and Hume AI emotion scores.
- The categorized vocal emotion scores and corresponding vocal features.
- The categorized vocal emotion scores and Hume AI emotion scores.

Limiations of the customized categorization method were identified during this phase and are discussed in the results.

For further exploration of how vocal features varied across AI-detected emotions, one-way ANOVA tests were conducted, follwed by Tukey's HSD analysis to evaluate specifc feature differences, see 3.7 Statistic Analysis for details.

To track changes over time, each audio recording were segmented into timeframes of around 5s including Hume AI emotion scoring and vocal features for that specific segmment. These were analyzed by tracking Z-score variations, see 3.7, in key vocal features (pitch, intensity, jitter, and shimmer) throughout each recording. This enabled observation of how vocal features shifted within a single clip and to see how these changes matched Hume's emotion predictions.

Custom Emotion Categorization Method

To complement the correlation analysis with Hume AI, a custom rule-based categorization method was integrated to group emotion probabilities from extracted vocal features. The approach was based on the Swedish research on vocal markers in emotions (Ekberg et al., 2023), where statistical patterns were identified in vocal markers across five emotions: anger, joy, sadness, fear, and surprise. The Swedish research did not state their used baseline for neutral speech, therefore the baseline for this study is the average vocal features of all clips.

The method functions as following:

- 1. Predefined means and standard deviations for each vocal features identified for each emotion were retrieved from the Swedish research (Ekberg et al., 2023). These features are stored in JSON format.

- 2. For every feature included in a recording, the function calculates the standardized distance between the measured vocal value and the mean for each emotion:

$$d_{\text{emo}} \mathrel{+}= \frac{|x - \mu|}{\sigma}$$

where x is the observed value, and μ, σ are the mean and standard deviation for the pair of feature and emotion.

- 3. To increase the functionality, distances are inverted so smaller distances can result in higher emotion scores. This is calculated with the function below, where ϵ is a small constant to avoid division with zero.

$$\text{score}_{\text{emo}} = \frac{1/(d_{\text{emo}} + \epsilon)}{\sum_e 1/(d_e + \epsilon)}$$

- 4. The output is a normalized probability value that is distributed across all five emotions.

The method allowed to map vocal data to emotion estimates in an interpretable way to enable structured comparison with speech-based emotion labels. However, the performance of this method for our dataset had limitations and gave uniform emotion labels.

2.3.2 RQ2 and RQ3: Speech-Based AI vs. Text-Based AI, AI-labels vs. Self-Assessed Emotions

For both RQ2 and RQ3, statistical analyses were used to evaluate the alignment between AI-generated emotion scores and self-reported emotions. The following methods were applied across both research questions:

- Pearson correlation coefficients and associated p-values for measurement both for single clips and the full dataset.
- T-tests and calculation of Cohen's d to evaluate statistical significance and effect sizes.

For all statistical tests, a p-value below 0.05 was considered as statistical significant, implying that the observed correlations or differences were unlikely to have occurred by chance. These statistical methods were applied to the entire dataset, and separately for negatively and positively oriented interview scenarios to identify potential contextual differences. For RQ2, comparisons focused on speech-based vs. text-based AI results, and RQ3 extended the comparison to include self-reported emotions as a subjective component.

2.3.3 Data Normalization

To enable direct comparison across different sources, all emotion scores were normalized to sum up to 1. The normalization included the following steps:

- 1. Surprise combination: Hume AI predicted two seperate labels for Surprise, one positive and one negative. These were merged into a single "surprise" by calculating the average.
- 2. Filtering and formatting: Filtering to only include the five target emotions (anger, joy, sadness, fear, surprise), since Hume predicted around 30 different emotions. All emotion labels were converted to lowercase.
- 3. Normalization: The emotion scores were normalized so the sum of all five target emotion values equals 1. This was done by dividing each score by the total sum of the emotion values. If the total sum was zero (no emotion detected), all scores were set to 0.

Normalization ensured consistent comparability between the sources, for both AI models and self-assessments, regardless of scale differences in raw scores.

2.3.4 Visual Analysis

RQ1:

Visualizations included:

- Heatmaps showing correlation matrices.
- Line plot diagrams for the full dataset.
- Bar chart diagrams for single clips.
- Composite correlation diagrams.
- Over-time diagrams comparing vocal features and Hume labels.

To evaluate the performance of the custom vocal feature-based emotion categorization method, line plots and bar charts were implemented to visualize differences between the generated scores and Hume AI's predictions. The line plot summerizes average emotion scores across the full dataset, while bar charts presented detailed comparisons for individual audio recordings. These diagrams emphasizes deviations and alignments between the categorized vocal marker method and AI-based emotion prediction.

Composite correlation diagrams were used to explore associations between single vocal features and AI-generated emotion scores. For these diagrams, Pearson correlation coefficients were calculated for each emotion and its relation to pitch and intensity, the results are visualized as grouped bar charts to easily identify positive or negative tendencies.

RQ2 and RQ3:

Visual methods included:

- Difference bar charts to illustrate the average difference in Hume AI and NLP Cloud's emotion scores for each emotion category, to provide a clear view of systematic deviations.
- Grouped bar charts to present average emotion scores from both AI systems, seperated by interview sentiment (positive vs. negative), to explore variations in emotion detection depending on the interview context.
- Scatter plots (RQ3) to explore detailed relationships between AI-based emotion scores and self-reported emotions for single emotions and recordings.

These visualizations were integrated to support the identification of alignment patterns between the AI systems, and contributed with insights into how the modality of the AI influences emotion recognition results. Python have been utilized to create all visualizations.

Given the limited dataset size and timeframe, a combination of statistical methods and visual analysis, was utilized to balance quantitative data with qualitative interpretation and support the exploratory nature of this study.

2.4 Model Configuration

2.4.1 NLP Cloud

The text-based emotion recognition is classified with NLP Clouds finetuned-llama-3-70b model through prompting, which allows a more flexible approach. Each text input uses the following prompt:

Listing 2.1: NLP Cloud configuration prompt.

```
        prompt = (
            "You␣are␣an␣emotion␣analysis␣system.␣"
            "Given␣a␣Swedish␣text,␣respond␣only␣
               with␣a␣JSON␣object␣using␣these␣
               emotion
␣␣␣␣␣␣␣␣␣␣␣␣␣labels:␣"
            "joy,␣surprise,␣fear,␣anger,␣sadness.␣
               Each␣value␣must␣be␣a␣float
␣␣␣␣␣␣␣␣␣␣␣␣␣between␣0.0␣and␣1.0.␣"
                "Respond␣with␣the␣JSON␣directly␣and
                   ␣nothing␣else.\n\n"
                f"{text}"
            )
```

The prompt that is used returns a JSON response with float values ranging from 0.0 to 1.0 for each of the emotions with the labels “joy”, “surprise”, “fear”, “anger”, “disgust” and “sadness”. This approach was chosen to ensure these specific emotions being analyzed due to them being the feelings of the basic six, which are the feelings used in the research about acoustic features in swedish speech done by Ekberg (Ekberg et al., 2023).

2.4.2 Hume AI

To ensure consistency across the different models used in this research, some changes have been made to adjust the output from the Hume AI model to better match the format used in NLP Cloud. Additionally, NLP Cloud has the feeling surprise while Hume has two different feelings for surprise, the two being positive surprise and negative surprise. Therefore, the scores of the two feelings of surprise from the Hume model have been combined in this research to give just one number that creates the average of the two to match the format.

2.5 Validity and Reliability

2.5.1 Validity

To ensure validity, the interview scenarios are pilot tested to ensure they provoke intended emotions (Bryman et al., 2022). The use of multiple AI models (speech- and text-based) allows for cross-validation of results. Standardized interview prompts ensure consistency across participants. Participant self-assessment serves as a secondary reference to evaluate AI-labeled emotions. Triangulation across AI, vocal markers, text analysis, and self-assessments enhance convergent validity (Creswell & Creswell, 2023).

2.5.2 Reliability

To ensure reliability, standardized equipment and scenario are used to ensure replicability. Hume.ai, NLP Cloud, and Praat provide consistent measures. The AI models used in the study (Hume.ai and NLP Cloud) are pre-trained and validated emotion recognition systems. Correlation will be determined and are used to quantify the reliability of AI models in detecting emotions. The study has a replicable experimental setup, with documentation supporting replication to allow researchers to reproduce similar evaluations.

Triangulation is achieved in the study through comparison of speech AI, text AI, and self-reports which improves creditability. Any discreteness will be analyzed qualitatively to contextualize potential biases rather than assuming errors. Reliability is ensured through standardization in data collection. All interviews are preprocessed to reduce background noise and normalize volume levels. The online tool Auphonic (Auphonic, n.d.) is used for this, due to its simple usability for noice reduction, ability to cut out pauses and limit loudness. The same data processing steps are applied consistently for all recordings, ensuring equality in analysis. The study has a replicable experimental setup, with usage of pre-trained, publicly available APIs, and documentation supporting replication to allow researchers to reproduce similar analyses. These measures ensure that our study is generalizable within the scope or automated emotion recognition for stress analysis.

2.6 Considerations

To consider the implications of this study, several factors must be recognized. To address ethical and privacy concerns, all participant data is anonymized and securely stored to ensure privacy. The participants provide informed consent before engaging in this study. The emotion-provoking scenarios are designed to minimize distress, focusing on natural, everyday emotions rather than triggering events. The participants will have scenarios to choose from, see 2.2 Data Collection.

Scientific considerations extend to emotion research to Swedish speech and AI tools. Findings in the study can inform future human-interaction research in emotion-based applications. Societal considerations include that the insights could enhance AI-driven mental health tools and future research, especially for Swedish language and real-world interviews.

Theoretical Framework

The following chapter will introduce the relevant theories and key concepts related to emotion recognition, such as speech-based and text-based models and technologies. This chapter will explain how vocal features and speech prosody can help to identify different emotions in spoken languages, using different AI tools and software. This study aims to compare accuracy and effectiveness of different approaches by conducting interviews and collecting data, which will be analyzed using a Python application. The elements of the Python application for analyzing the data from the interviews will be comprehensively explained in this chapter.

3.1 Affective Computing

Affective computing was introduced by Professor Rosalind Picard in the mid-1990s to early 2000s (Tian et al., 2022). By exploring the ways in which human emotions are recognized, understood and expressed through different forms of behaviors and communication, the domain of affective computing is a field that merges the principles of artificial intelligence with insights from social and behavioral science (Tian et al., 2022).

3.2 Natural Language Processing and Emotion Recognition

The first English language lexical database was created in 1998 (S. Kusal et al., 2023) for Natural Language Processing (NLP) tasks. The term sentiment and emotional analysis came to practice in 2001 (S. Kusal et al., 2023) to predict the stock market. In 2005, the first article was written on emotion and opinion detection from text. Concept-level sentiment analysis resources were publicly available in 2009. Word embedding is the term to represent words for NLP text analysis and was developed in 2013, the same year as neural network first was adopted in NLP tasks. The field had a massive upwelling when the transformers concept was published in 2017, followed by the evolution of BERT, a pretrained model that automated text analysis and classification in 2018 (S. Kusal et al., 2023).

Traditional approaches for sentiment analysis classification have been used since the past few decades, which rely on rule-based methods such as “bag of words” method to process text. The method represents text based on word frequency without consideration of word order. It can identify sentence structure, negation, emphasis, subjectivity and irony. Recent models leverage deep learning algorithms that process raw text by first cleaning and pre-processing it, including punctation, stop words and markups, as well as applying stemming (the process of reducing words to their root form by removing prefixes or suffixes to simplify text analysis in NLP) (Kansara et al., 2020).

Deep learning applicate artificial neural networks (ANN) to learn tasks using multiple layers of network. In traditional models only one or two layers could be used, but in deep learning much more learning power of artificial neural networks is exploited (L. Zhang et al., 2018).

Studies have shown consistently higher accuracy for sentiment analysis using deep learning algorithms compared to traditional machine learning algorithms (Kansara et al., 2020).

Traditionally, SER systems are comprised of three components: signal preprocessing, feature extraction, and classification (Sahoo et al., 2023).

Emotion recognition from speech is generated by other types of machine learning algorithms, some algorithms are used for both types of emotion recognition, where Support Vector Machine (SVM) have shown high accuracy for speech analysis.

SVM is a simple machine learning algorithm that is highly preferred, since it requires less computational power for producing substantial accuracy (S. D. Kusal et al., 2024).

3.3 Speech-Based Emotion Recognition

In the field of artificial intelligence, researching areas to be able to identify emotions is important to read humans better. Studies about speech-based emotion recognition (SER) have been ongoing since 1978 (Sönmez & Varol, 2024). SER identifies how something is being said without the context of the words spoken. These systems are used in many different areas, most often in areas of interactions between humans and machines (Zhang, 2025).

Although there is a wide range of SER-algorithms, some approaches using more complex setups that involve Convolutional Neural Networks (CNN) based SER algorithms among others (Ri et al., 2023), the process of a SER algorithm could look like figure 3.1, involving several steps as feature extraction, selection and classification.

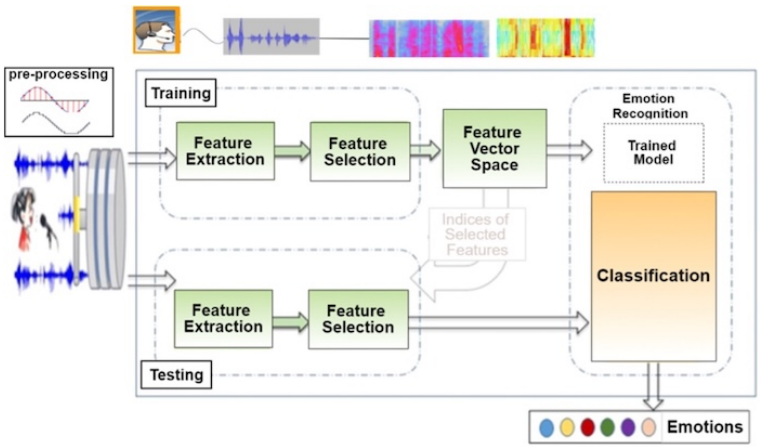


Figure 3.1: *An overview of the stages in SER to analyze speech data for emotion detection* (Sönmez & Varol, 2024)

While emotions can be recognized via many channels such as speech, facial expressions and text, speech signals are rapid and natural which makes vocal audio fitting for emotion recognition. According to (Sönmez & Varol, 2024) there are several key benefits with SER:

- There is only a small amount of hardware needed to capture vocal data.
- It is a simple process to collect vocal data, due to little hardware being needed.
- It is harder to mimic vocal data compared to facial data.
- Vocal data is less demanding in terms of storage than video footage for example.
- Participants in SER experiments may feel more comfortable with vocal analysis than face analysis in terms of confidentiality, resulting in datasets reflecting real emotions more accurately.

3.3.1 Hume AI

Hume is a technology company dedicated to advancing the field of emotion recognition.

Having conducted extensive psychology studies to explore human emotions and the way these emotions are expressed, Hume AI has used the research to develop advanced machine learning models (Hume, n.d.-b) as well as using deep learning for the research and development (Brooks et al., 2023).

The official website of Hume AI outlines several influences on their emotional mapping. Drawing influence from key figures such as David Hume, Charles Darwin and Paul Ekman, Hume AI's research is grounded in these foundational theories of emotions. Paul Ekman's "The Basic 6" is mentioned (Hume, n.d.-a) and remains relevant throughout this research.

One of the measurements used to recognize emotions in vocals with Hume AI in this research is speech prosody.

Speech prosody gives crucial insights into a speaker's purpose in their communication. Particular emotions and the intensity of those emotions are indicated with intonation, rhythm and pitch of the speaking voice (Thompson et al., 2004)(Tomasello et al., 2022). It simply refers to the patterns and tone in the speech that are not related to the actual words being spoken (Cowen et al., 2019).

Happiness and sadness show the opposite characteristics of each other, where happiness is linked to quicker tempo and higher pitch while sadness has the opposite, a slower tempo and lower pitch. The clear difference between the characteristics serves as the difference in the speech prosody of the two emotions (Thompson et al., 2004).

In figure 3.2, a visual presentation of Hume AI's speech prosody model is visible (AI, n.d.-a). Emotions are clustered with other similar emotions, one example being amusement and joy, or distress and anxiety.

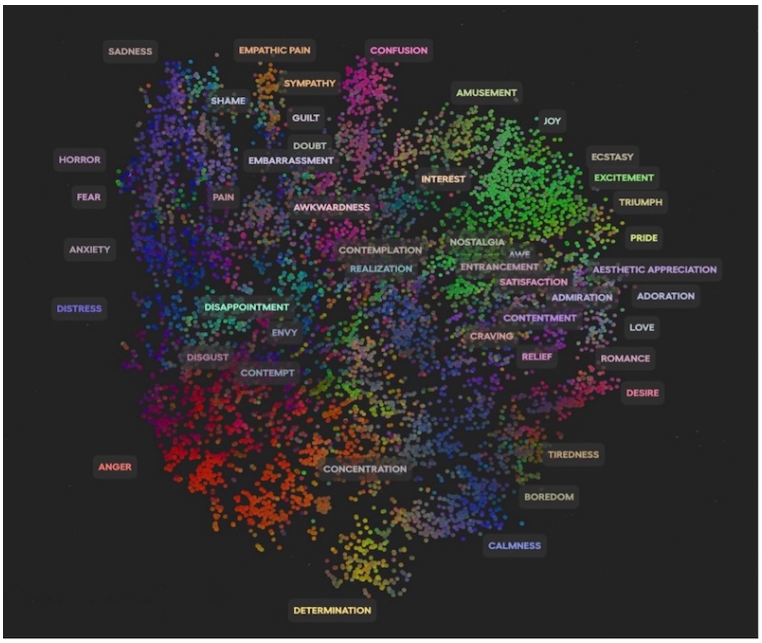


Figure 3.2: *Visual representation of Hume AI's speech prosody model* (AI, n.d.-a)

To ensure a broader range of emotion recognition with a more comprehensive analysis of human voices in this research, speech prosody is used in combination with another measurement, vocal bursts.

Vocal bursts play a key role in social communication between humans. They are short emotional sounds which occur naturally, some examples being laughs, sighs or cries (Brooks et al., 2023).

Vocal burst and voice overall have been relatively overlooked in the fields of machine learning and affective computing due to more focus being held on facial expressions. Even if speech prosody has been studied more extensively, there has been newer research showing that vocal bursts convey more than ten different emotions with consistency, being mostly consistent across different cultures as well (Baird et al., 2022).

Figure 3.3 shows Hume AI's mapping of non-verbal communication, vocal bursts (AI, n.d.-b). Emotions are shown and as well as in Hume AI's speech prosody model, the emotions are clustered indicating some emotions are more associated with each other.

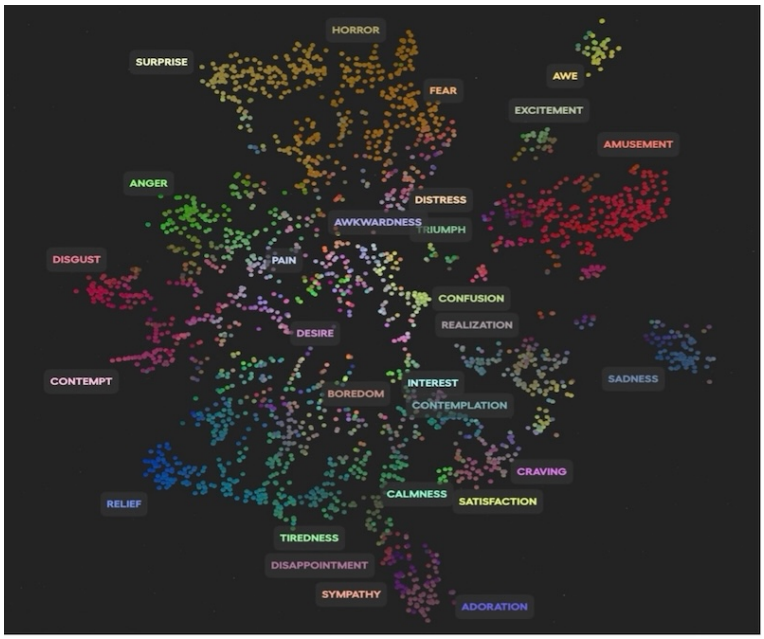


Figure 3.3: *Visual representation of Hume AI's vocal burst model* (AI, n.d.-b)

While there are other tools for emotion recognition, Hume AI seems to be among the few that are specifically designed for Voice AI while being able to recognize emotions through specifically speech prosody and vocal burst with no need to finetune it yourself. Using models needing finetuning would not fit the scope of this thesis given the limited timeframe, and while models like OpenAI whisper is an extensively trained model on hundreds of thousands of hours on data, their main focus leans toward transcribing speech (OpenAI, 2022).

3.3.2 Praat Parselmouth

In the field of software for analyzation in linguistics, Praat is a well-established tool to analyze different elements in speech. Being able to estimate elements such as fundamental frequency and intensity among others, Praat holds a broad spectrum of algorithms in acoustics, being a successful tool for analyzing acoustics (Jadoul et al., 2024).

Designed to provide efficient access to the core functionalities of Praat in Python for programmers, Parselmouth is an open-source Python library (Jadoul et al., 2018).

Python is widely used for data analysis, but it had been noted that there were challenges with analyzing acoustics in Praat, this due to the functionality often being missing or scattered across multiple incompatible libraries.

Parselmouth streamlines and optimizes workflows in a single programming environment by enabling a deeper integration of the capabilities of Praat in combination with other libraries (Jadoul et al., 2024). Not designed to replace Praat, but rather a way to enable users to access the functionality of Praat directly in Python, the main objectives in Parselmouth according to (Jadoul et al., 2018) are:

- Enable users already experienced with Praat to effectively incorporate its functionality with Python's scientific tools, being tools that are not obtainable in Praat.
- Providing Python users with the ability to utilize the functionality of Praat, even if they are not experienced users of it.
- Enhancing the optimal aspect of workflow for users preferring to conduct their work within a single programming language.

The benefits of Parselmouth both in terms of the usage for completing this thesis and overall, are it being open source and compatible with Python as Python is widely utilized and backed by a vast community of researchers and engineers, among others. Parselmouth integrates the different strengths of different approaches to provide a truly pythonic library, behaving consistently with other well-known Python libraries.

Parselmouth directly utilizes the official C/C++ source code of Praat instead of having to reconstruct its algorithms. This simplifies the process since it guarantees full consistency with Praat without the requirement of learning its scripting language (Jadoul et al., 2018).

There are other similar tools that essentially could accomplish the same task, like Librosa although it is more tailored for both audio and music analysis. It does have feature extraction (Babu et al., 2021), but the decision on which software to use for linguistic analyzation still falls on Praat Parselmouth due to it being more fitting for the purpose of this thesis.

3.4 Text-Based Emotion Recognition

In the field of NLP, the comprehension of the context behind words in text-form has gone from only being able to determine the tone in text to actually identifying the emotions behind them (Esfahani & Adda, 2024), recognizing these capabilities has valuable practical applications in enhancing different domains within human-computer interaction (Shelke et al., 2022).

Text-based emotion detection relies on four main approaches, according to (S. Kusal et al., 2023) written in 2023:

1. Keyword-based: Matches words in a text with predefined emotion keywords from resources like WordNet, adjusting for intensity and negation.
2. Rule-based: Uses linguistic rules and probabilistic affinity to classify emotions after preprocessing.
3. Machine learning-based: applies supervised or unsupervised models to classify emotions, extracting key features from preprocessed text.
4. Deep learning-based: utilizes neural networks to learn complex patterns from tokenized and embedded text data for emotion classification.

Machine learning classifiers are significantly used in text-based classification, since they use labeled datasets and are therefore data driven. Machine learning models are trained on large number of datasets and learn from experience, with classifiers that contain labels for input and desired output. Transformer-based models, such as BERT, are based on machine learning models which are trained on vast amounts of data and can be fined-tuned for specific tasks. BERT is a deep learning model based on attention processing. It gains a thorough text-understanding through considering left and right contexts equally. The model solves NLP issues and is used to train general language models on large datasets (S. D. Kusal et al., 2024).

3.4.1 NLP Cloud

There are limited publicly available APIs for text-based emotion detection. The decision to use NLP Cloud has several reasons. Compared to other available TBED APIs, that do not require pre-training, NLP Cloud has comprehensive documentation for both API and the models it is based on. The company has support as well as information about Data Privacy and Security (Cloud, n.d.), which also other APIs for TBED has. For example, Vern AI (AI, u.d.) provides customer support but has very limited information about the API or documentation that is easily accessible. TwinWord (TwinWord, n.d.) was another choice, also providing contact support and privacy information, but was deficient in comprehensive documentation and limited information about the API.

NLP Cloud was the obvious choice for several reasons. In addition to what is mentioned above, the company has solid customers like Zoom and collaboration with Nvidia. They provide information about what models they use for their API which can be downloaded and fine-tuned if desired. In contrast to the other APIs, NLP Cloud provide APIs for Speech-to-Text transcription, and they have an emotional analysis model supporting Swedish Language. Therefore, no translation is required beforehand. Their speech-to-text API is based on OpenAI's Whisper model (Cloud, n.d.). OpenAI provides a research report on the model, which is a speech recognition system designed to process and transcribe audio with remarkable robustness and generalization. Contrasting traditional models that heavily rely on unsupervised pre-training or dataset-specific fine-tuning, Whisper leverages large-scale weakly supervised training from over 600,000 hours of multilingual audio data. This includes 96 languages beyond English. Whisper handles several tasks, for instance speech recognition, language identification, and translation (Radford et al., 2022).

For sentiment and emotion analysis, NLP Cloud provides several options. Two of them are equally researched with relatively comparable results. DistilBERT Base Uncased Emotion is one option, with studies on DistilBERT, and Transformer-model which require finetuning and require less computational resources than traditional BERT (Bidirectional Encoder Representations from Transformers) models. The model was developed by Google AI Language researchers in October 2018. BERT has challenges regarding fixed input length and computational complexity, reasons that led to DistilBERT's development in October 2019. This pre-trained model uses technology that accomplishes to reduce a BERT model by 40% while preserving 97% of language understanding capabilities 60% faster. The model accuracy for sentiment analysis ranges from 95.7% to 96.6% on Yelp Open Dataset, which has been demonstrated as a valuable resource of sentiment-labeled review-data (Areshey & Mathkour, 2024). NLP Cloud provides a pre-fine-tuned version of DistilBERT. However, it does not support Swedish language and translation beforehand is necessary for this study. A fine-tuned version of Llama 3 is an option on NLP Cloud that support several languages including Swedish (Cloud, n.d.). Opposed as to DistilBERT, some studies on emotion detection have been conducted on Llama 3. Both models have predominantly been evaluated for sentiment analysis. Emotion identification using Llama 3 showed an F1 score of 0.48 as average for all tested emotions: anger, joy, sadness, surprise, fear, and love (J. Zhang & M, 2024). A fine-tuned version of Llama 3 was tested for sentiment analysis, which resulted in an accuracy increase from 0.333 to 0.923 and F1 score from 0.50 to 0.91. The authors of this study imply that these results are superior performance against other models that are included for comparison, including DistilBERT (Kumar & Singh, 2024). In a text classification study, mainly examines speed performance, DistilBERT had an accuracy between 0.94-0.96 for the Amazon Alexa Reviews dataset but 0.35-0.41 for the Brexit Blog Corpus dataset (Barbon & Akabane, 2022). As for many models for emotion recognition mentioned in this report, the dataset affects the results heavily.

Regarding multiple language performance, a study examined LLaMA 3 vs. State-of-the-Art Large Language Models on their ability to detect fake news. Two datasets were tested, one

Romanian and one English. Their proposed Llama 3 model accomplished higher precision and accuracy across several metrics in fake new detection. For the English dataset, the fine-tuned Llama 3 model had lower accuracy compared to ChatGPT 4 and Gemini. Yet, it outperformed these models for the Romanian dataset. The study also explored their fine-tuned Llama model compared to its base version. The fine-tuned model outperformed earlier models in distinguishing nuanced categories, particularly for the Romanian dataset where it achieved a remarkably high accuracy of 68% in one category (Repede & Brad, 2024).

Comparing these two alternatives for text-based emotion detection in Swedish, the fine-tuned Llama 3 model emerges as the most suitable choice. Although the exact fine-tuned version of the model available on NLP cloud has not been publicly researched, its built-in compatibility with Swedish, combined with research on a Romanian dataset, makes it a stronger candidate than DistilBERT. Both models have achieved high accuracy for TBED. Nevertheless, given this study is aimed to focus on emotion recognition models for Swedish speech, the Llama model without need for prior translation is a more valuable choice.

3.5 Vocal Markers

States of emotions are determined by many different factors, but in speech, different vocal markers such as prosody, pitch and loudness are among the telling features when it comes to emotions (Ekberg et al., 2023), where frequency is perceived as pitch, while amplitude is heard as loudness (Frühholz & Belin, 2019).

In this research vocal markers are crucial to understand the difference in certain emotions in voice.

Figure 3.4 (Ekberg et al., 2023) shows a table of comparisons with certain parameters of vocal markers measured for five different emotions in the Swedish language and some explanations for the different parameters. This is used in this thesis as a comparison for the vocal markers collected from the interview data in the conducted experiment, although this thesis does not use all acoustic features to measure emotions.

TABLE 2.
Comparisons of Acoustic Parameters Between Emotions

Acoustic features (parameters)	Anger	Happiness	Fear	Sadness	Surprise				Comparisons
	M (*SD*)	M (*SD*)	M (*SD*)	M (*SD*)	M (*SD*)	*F*	*P*	pEta2	Post hoc-tests (Bonferroni adjusted)
Frequency-related:									
pitch	5.00 (5.39)	7.18 (6.25)	5.81 (2.31)	3.99 (5.36)	3.56 (4.14)	2.77	0.029	.068	Happiness > Surprise ($P = 0.039$)
jitter	-0.13 (0.38)	0.58 (0.38)	-0.98 (0.41)	0.32 (0.39)	2.14 (0.39)	8.24	<0.001	.178	Surprise > Anger ($P < 0.001$) Surprise > Fear ($P < 0.001$) Surprise > Sadness ($P = 0.014$)
F1Frequency	0.78 (0.34)	1.75 (0.34)	1.47 (0.37)	0.12 (0.35)	0.57 (0.35)	3.60	0.008	.086	Happiness > Sadness ($P = 0.012$)
F2Frequency	1.20 (0.35)	1.94 (0.35)	1.75 (0.37)	0.23 (0.36)	1.03 (0.36)	3.53	0.009	.085	Happiness > Sadness ($P = 0.008$) Fear > Sadness ($P = 0.038$)
F3Frequency	0.80 (0.34)	1.59 (0.34)	0.88 (0.37)	-0.10 (0.35)	0.72 (0.35)	2.95	0.022	.072	Happiness > Sadness ($P = 0.008$)
F1Bandwidth	-1.05 (1.29)	-0.96 (0.95)	-0.44 (0.94)	-0.88 (1.35)	-0.82 (0.88)	1.38	0.244		
Amplitude-related:									
shimmer	-1.03 (0.21)	-1.02 (0.21)	-1.43 (0.23)	-1.02 (0.22)	0.13 (0.22)	7.12	<0.001	.158	Surprise > Anger ($P = 0.002$), Surprise > Fear ($P < 0.001$) Surprise > Happiness ($P = 0.002$) Surprise > Sadness ($P = 0.003$)
loudness	7.16 (0.66)	6.49 (0.66)	5.09 (0.71)	2.96 (0.68)	1.24 (0.68)	13.36	<0.001	.260	Anger > Sadness ($P < 0.001$) Anger > Surprise ($P < 0.001$) Fear > Surprise ($P = 0.001$) Happiness > Sadness ($P = 0.003$) Happiness > Surprise ($P < 0.001$)
HNR	2.36 (0.52)	3.99 (0.52)	4.83 (0.55)	2.16 (0.54)	1.31 (0.54)	7.09	<0.001	.157	Fear > Anger ($P = 0.014$) Fear > Sadness ($P = 0.007$) Fear > Surprise ($P < 0.001$) Happiness > Surprise ($P = 0.004$)
alphaRatio	2.52 (0.40)	2.15 (0.40)	1.14 (0.43)	1.95 (0.41)	0.48 (0.41)	4.05	0.004	.096	Anger > Surprise ($P = 0.005$) Happiness > Surprise ($P = 0.043$)
Hammarberg	-1.57 (0.28)	-1.19 (0.28)	-0.74 (0.30)	-1.4 (0.29)	-0.35 (0.29)	2.97	0.022	.072	Surprise > Anger ($P = 0.032$)
slopeV0V500	2.53 (0.43)	2.68 (0.43)	4.90 (0.46)	2.76 (0.44)	1.81 (0.44)	6.54	<0.001	.147	Fear > Anger ($P = 0.002$) Fear > Happiness ($P = 0.006$) Fear > Sadness ($P = 0.010$) Fear > Surprise ($P < 0.001$)
slopev500V1500	1.45 (0.32)	1.57 (0.32)	1.28 (0.34)	0.35 (0.33)	0.12 (0.33)	4.21	0.003	.100	Anger > Surprise ($P = 0.042$) Happiness > Surprise ($P = 0.019$)
F1Amplitude	-0.3 (0.22)	-0.31 (0.22)	-0.19 (0.24)	-0.49 (0.23)	-0.85 (0.23)	1.30	0.274		
F2Amplitude	0.32 (0.20)	0.43 (0.20)	0.21 (0.21)	0.10 (0.20)	-0.56 (0.20)	3.68	0.007	.088	Anger > Surprise ($P = 0.024$) Happiness > Surprise ($P = 0.007$)
F3Amplitude	0.34 (0.20)	0.46 (0.20)	0.24 (0.21)	0.14 (0.21)	-0.52 (0.21)	3.54	0.009	.085	Surprise < Anger ($P = 0.030$) Happiness > Surprise ($P = 0.008$)
H1H2	1.44 (0.24)	1.61 (0.24)	0.66 (0.25)	0.48 (0.25)	1.13 (0.25)	4.00	0.004	.095	Happiness > Sadness ($P = 0.012$)
H1A3	-0.91 (0.29)	-1.19 (0.29)	-1.61 (0.30)	-1.40 (0.29)	-0.83 (0.29)	1.23	0.301		
Temporal-related:									
loudnesspeaksRate	-1.79 (0.27)	-1.35 (0.27)	-0.71 (0.28)	-1.30 (0.27)	-0.13 (0.27)	5.65	<0.001	.129	Surprise > Anger ($P < 0.001$) Surprise > Happiness ($P = 0.016$) Surprise > Sadness (P=.029)
voicedLength	0.28 (0.19)	0.31 (0.19)	0.17 (0.20)	0.35 (0.19)	-0.40 (0.19)	2.68	0.034	.066	
unvoicedLength	0.15 (0.27)	-0.05 (0.27)	-0.17 (0.28)	0.42 (0.28)	0.22 (0.28)	0.69	0.598		
pseudoyllableRate	-0.34 (0.19)	-0.22 (0.19)	-0.26 (0.20)	-0.38 (0.19)	0.44 (0.19)	3.00	0.020	.073	Surprise > Anger ($P = 0.046$) Surprise > Sadness ($P = 0.034$)

Note: F1Frequency = Frequency- formant 1, F2Frequency = Frequency-formant 2, F3Frequency = Frequency-formant 2, F1Bandwidth = Formant 1 bandwidth, HNR = Harmonics-to Noise ratio, AlphaRatio = Alpha ratio, Hammar = Hammarberg index, v0v500 = Spectral Slope V 0-500 Hz, v500v1500 = Spectral slope V 500-1500 Hz, F1Amp = Formant 1 relative energy, F2Amp = Formant 2 relative energy, F3Amp = Formant 3 relative energy, H1H2 = Harmonic difference H1-H2, H1A3 = Harmonic difference H1-A3, LoudPeak = Rate of loudness peaks, Voice = Length of continuously voiced regions, Unvoice = The length of unvoiced regions, Pseudo = Pseudo syllable rate.

Figure 3.4: *Table comparing acoustic parameters between emotions* (Ekberg et al., 2023)

Figure 3.5 shows a table of variations in different emotions measuring the acoustic parameters pitch, intensity, speaking rate and voice quality which are often used to identify emotions (Khalil et al., 2019). Although figure 3.4 displays a wider variety of more specific acoustic features, figure 3.5 provides a foundational understanding of different acoustic features connected to the different emotions.

Emotions	**Pitch**	**Intensity**	**Speaking rate**	**Voice quality**
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

Figure 3.5: *Acoustic variations in different emotions* (Khalil et al., 2019).

3.6 The Experiment

An experiment will be conducted and will consist of short interviews with voluntary people. These interviews will be recorded for data collection and will be used to extract emotions with a Python application.

3.6.1 Python Application

The Python application serves as the central system for processing and analyzing emotions in speech and will integrate several tools and frameworks to extract emotions.

The steps for the process consist of:

1. Audio recording
2. Feature extraction

- Prosody and vocal bursts will be analyzed with Hume AI to detect emotional cues from pitch, intonation and vocal bursts using Hume AI's models.
- Extracting vocal features such as jitter, shimmer and frequency with Praat Parselmouth.
- Convert speech to text and analyze emotions in text with NLP Cloud.

3.6.2 Interviews and Surveys

The interviews involve voluntary participants engaging in short audio-recorded interviews, designed to draw out natural emotional responses. The participants will be asked questions to prompt them to recall and reflect on past experiences which encourages them to revisit emotions they felt at that time.

For ethical purposes the participants will be given a selection of topics to choose from, minimizing the risk of discomfort or distress. The audio recordings will be anonymous and recorded in a controlled acoustic environment to ensure minimal noise interference.

Questions asked during the interview follow one of many emotion induction techniques, "autobiographical recall". This is a method used to facilitate the re-experiencing of emotions felt in a previous moment (Siedlecka & Denson, 2019), which is what is intended for the interviews to be able to collect emotional data from vocal recordings. By letting the participant think and speak about a memory from the past, emotions felt in that moment reflect in their voice.

After the interview is done, the participants will answer a survey doing a self-assessment of their emotions felt during the interview. This will enable a comparison between emotion detection and the participants reported emotional experiences.

There are many different methods for self-assessment, and emotional self-assessment is linked to many different theories. Many are connected to emotional intelligence (EI), trait emotional intelligence (trait EI) and Core Self Evaluation (CSE) (Montasem et al., 2013), but rather than conducting a comprehensive exploration of different psychological theories in self-assessment, this research will use simplified surveys at a basic level for the purpose of fulfilling the technical objectives of the work.

The theories behind the interviews are stated in the possibility of detecting emotions in voices. While there are a lot of recognized emotions that can be detected in different AI models and software tools, the interview will focus on bringing out two different basic emotions to maintain a manageable scope, while ensuring ethical feasibility.

Research has stated that there are different levels of unique universal signs for different affective states and while there are evidence supporting the universality for certain emotions such as anger, fear, surprise, sadness, happiness and more, there are also emotions that do not include all characteristics that distinguish them from other mental states, two examples being guilt and shame (Ekman & Cordaro, 2011). Research of this nature supports the rationale for having the focus solely on two of the basic emotions for this thesis. The questions in the interviews will focus on bringing out two separate emotions, one on the positive spectrum, happiness, and one on the negative spectrum, anger.

3.7 Statistical Analysis

Pearson Correlation Coefficient

Pearson's r is a measurement of the strength and direction of a linear relationship between two variables. The value range is from -1 to 1, where positive values implies a positive correlation and negative values the opposite. Values close to +-1 indicate a strong correlation, values between +-0.30 and +-0.49 a moderate correlation, and values below +-0.29 are seen as a weak correlation. Values around 0 implies no linear correlation (Bruce & Bruce, 2017).

P-Value

A p-value indicates if the observed results have a probability of occurance by chance. A widely accepted threshold for statistical significance is $p < 0.05$, which means there is less than a 5% possibility that the observed effect is random (Bruce & Bruce, 2017).

Z-score Standardization

Acoustic features such as pitch, intensity, jitter and shimmer can have great variation. To ensure comparability in statistical analyses, features are often standardized using Z-score standardization (Ekberg et al., 2023). This method transform data to have a mean of zero and a standard deviation of one, ensuring meaningful comparisons across features. By this, a variable does not have an overly influence due to a scale of the measurement. The measurements are described as "standard deviations away from the mean". (Bruce & Bruce, 2017).

Standardized Distance for Emotion Categorization

Emotion categorization based on vocal features can be operated through standardized distance methods, where deviations from the baseline of acoustic profiles are quantafied. Using standardized differences allow an interpretable measure of how vocal features aligns with expected patterns for each emotion (Ekberg et al., 2023) (Bruce & Bruce, 2017). The categorization method used in this study is a custom method inspired by this standard practice.

ANOVA Tests

ANOVA (Analysis of Variance) is a standard statistical method used to determine if there are any significant differences in means across multiple groups Bruce2017. It is used to categorize grouping factors and are one method in the Swedish research for vocal markers Ekberg2023.

Tukey's HSD

When ANOVA presents significant differences between group means, Tukey's HSD test is incorporated as a post analysis to identify which groups are divergent from each other. This method controls for errors when making multiple comparisons (Bruce & Bruce, 2017).

T-Tests and Cohen's d

Paired T-tests are used to compare the means between two groups to determine statistical significance, while Cohen's d provides a standardized measure of the effect size which indicates the magnitude of the observed differences (Cohen, 1977) (Bruce & Bruce, 2017).

Results

4.1 Presentation of Collected Data

4.1.1 Overview of Interviews

We conducted semi-structured interviews with 16 native Swedish speakers (10 M/6F, age 23-78), each lasting 1-3 minutes. Each participant was interviewed for two different scenarios, resulting in 30 different recordings. All interviews were audio-recorded in a quiet room and elicited two target emotions – anger and happiness – via open-ended prompts (e.g. "Is there anything in society that makes you upset? What? How does that make you feel?"; "Can you remember one time you felt really proud of yourself?"). The participants rated their perceived emotions on a 1-6 scale immediately after each scenario. The rated emotions covered the basic 5 emotions mentioned in this report: anger, joy, sadness, fear, and surprise.

Table 4.1 presents the participants ID, gender, age, and self-assessed scores for their perceived emotions for respective interview scenario.

Participant			Negative					Positive				
ID	M/F	Age	A	J	Sad	F	Sur	A	J	Sad	F	Sur
1	M	23	5	1	3	1	1	1	6	1	1	4
2	M	26	6	1	3	4	1	1	6	1	2	1
3	F	27	4	1	6	1	2	1	6	1	1	3
4	M	29	2	1	3	2	1	1	4	2	2	2
5	F	28	4	1	4	1	2	1	5	1	1	5
6	M	25	2	2	1	1	1	1	3	1	1	1
8	M	27	3	1	2	1	2	1	5	1	1	1
9	F	26	3	1	3	1	1	1	5	1	1	1
10	F	78	5	1	3	2	4	1	6	4	1	1
11	F	27	3	3	2	1	1	1	6	1	1	1
12	M	58	1	3	1	2	1	1	6	1	1	3
13	F	54	4	1	4	3	1	1	6	1	1	1
14	M	20	1	3	1	2	2	1	4	1	1	3
15	M	30	3	2	2	3	1	2	5	1	1	1
16	M	25	4	1	2	1	1	1	6	1	1	1

Table 4.1: Participant table. A: Anger, J: Joy, Sad: Sadness, F: Fear, Sur: Surprise.

4.1.2 Data Collection for RQ1: Vocal Features & Speech

The collected audio recordings from the 32 interviews were processed for research questions 1 to specifically focus on vocal features and speech-based emotion recognition.

Vocal Feature Extraction (Praat Parselmouth)

Audio recordings were processed with Praat Parselmouth. Vocal parameters were extracted from each recording, which have been validatied by Swedish emotion research on vocal markers (Ekberg et al., 2023).

- Pitch: mean pitch in Hz and semitones (ST).
- Intensity: mean intensity measured in decibels (dB).
- Voice Quality Metrics: Harmonic-to-Noise Ratio (HNR), jitter (local frequency perturbation), shimmer (local ampliture perturbation).
- Formant frequencies: mean frequencies (Hz) of F1, F2, and F3.

These acoustic features were then categorized into discrete emotional labels (anger, joy, sadness, fear, surprise) based on thresholds and criteria defined in prior Swedish research (Ekberg et al., 2023).

Speech-Based Emotion Recognition (Hume AI)

The same audio were analyzed with Hume AI, that provided probability distributions across several emotions, where the five targeted emotions was filtered from. The results were normalized according to ADD THAT PART INTO METHOD FROM PDF REFER TO HERE.

Data Structure

The extracted vocal features, Praat-based emotion categorizations, and Hume AI outputs were matched and stored in JSON format, to enable direct comparison and further analysis. The Listening 4.1 presents the structured data.

Listing 4.1: Example of stored JSON structure for vocal features vs. Hume

```
    {
    "entry_id": "id_005_neg",
    "vocal_features": {
        "mean_pitch_st": -4.12,
        "mean_pitch_hz": 118.24,
        "mean_intensity_db": 58.15,
        "mean_hnr_db": -0.5,
        "jitter_local": 0.0261,
        "shimmer_local": 0.1096,
        "formants_hz": {
            "F1": 1118.56,
            "F2": 2623.03,
            "F3": 3611.26
        }
    },
    "praat_scores": {
        "anger": 0.208,
        "joy": 0.205,
        "fear": 0.199,
        "sadness": 0.19,
        "surprise": 0.198
    },
    "praat_label": "anger",
    "hume_probs": {
        "anger": 0.21,
        "fear": 0.16,
        "joy": 0.15,
        "sadness": 0.34,
        "surprise": 0.14
    },
    "hume_label": "sadness"
}
```

Segment-Level Data

text

4.1.3 Data Collection for RQ2 and RQ3: Text, Speech and Self-Assessment

The data collection for RQ2 and RQ3 is based on the same audio recordings as for RQ1. Each recording was transcribed and analyzed with NLP Cloud (text-based), to extract emotion probabilities from the transcription. The same audio was analyzed using Hume AI for speech-based emotion detection, resulting in paired emotion probability scores alongside self-reported emotion ratings. All scores were normalized for comparison.

The data was structured in JSON format as shown in Figure 4.1.3, each audio object consists of five emotion labels from each data type (Hume, NLP, Self).

Listing 4.2: Example of stored JSON structure for Hume, NLP, Self-labeling.

```
        {
            "id_013_neg": {
                "audio_file": "audio_use/negative
                   /13-neg.m4a",
                "nlp_emotions": {
                    "anger": 0.44,
                    "joy": 0.0,
                    "sadness": 0.31,
                    "fear": 0.21,
                    "surprise": 0.05
                },
                "hume_emotions": {
                    "anger": 0.32,
                    "fear": 0.13,
                    "joy": 0.22,
                    "sadness": 0.19,
                    "surprise": 0.13
                },
                "self_assessed": {
                    "anger": 0.31,
                    "joy": 0.08,
                    "sadness": 0.31,
                    "fear": 0.23,
                    "surprise": 0.08
                }
            }
    }
```

Table 4.2 summarize the average emotion scores and standard deviations for both speech-based (Hume AI) and text-based (NLP Cloud) models across all clips in the dataset.

Emotion	Self Mean	Hume Mean	NLP Mean	Self Std	Hume Std	NLP Std
Anger	0,21	0,26	0,2	0,124	0,072	0,223
Joy	0,312	0,302	0,396	0,2	0,117	0,351
Sadness	0,19	0,167	0,181	0,105	0,065	0,138
Fear	0,136	0,15	0,093	0,061	0,045	0,092
Surprise	0,149	0,118	0,129	0,082	0,022	0,089

Table 4.2: Mean and standard deviation for Hume, NLP, Self-labeling for full dataset.

The interviews were conducted with either a positive or negative orientation. Each recording was analyzed individually, and the data structure distinguishes between negative and positive audio files. The corresponding emotion scores from Hume AI and NLP Cloud are presented in Table 4.3 for negatively oriented interviews, and in Table 4.4 for positively oriented interviews. Each table displays the mean and the standard deviation for the respective AI model's emotion probability.

Negative				
Emotion	**Hume M**	**NLP M**	**Hume Sd**	**NLP Sd**
Anger	0,29	0,363	0,072	0,183
Joy	0,276	0,121	0,098	0,205
Sadness	0,171	0,282	0,066	0,104
Fear	0,152	0,141	0,038	0,087
Surprise	0,112	0,092	0,021	0,064

Table 4.3: Mean and standard deviation for Hume and NLP for negative interviews.

Positive				
Emotion	**Hume M**	**NLP M**	**Hume Sd**	**NLP Sd**
Anger	0,228	0,015	0,06	0,057
Joy	0,334	0,708	0,132	0,168
Sadness	0,164	0,067	0,065	0,062
Fear	0,148	0,04	0,053	0,068
Surprise	0,126	0,171	0,022	0,097

Table 4.4: Mean and standard deviation for Hume and NLP for positive interviews.

4.2 Data Analysis for RQ1: Vocal Features & Speech Emotion Recognition

The first research question explores how vocal features correlates with AI-based emotion detection in conversational Swedish speech. To analyse this, acoustic features such as pitch, intensity, jitter, shimmer and HNR were extracted using Praat Parselmouth and compared with emotion scores from the speech-based model Hume AI. A custom categorization method based on Swedish vocal emotion research (Ekberg et al., 2023) were tested for comparison. The goal with this analysis was to explore if these vocal markers could explain or predict how speech-based AI systems interpret emotional expressions in semi-structured, spontaneous speech in an interview setting.

4.2.1 Correlation Between Vocal Features and AI Emotion Scores (Hume AI)

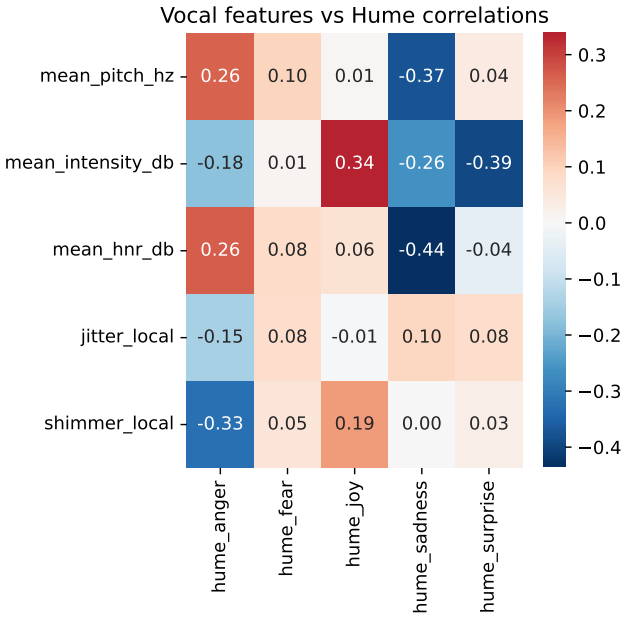


Figure 4.1: Heatmap of correlation between vocal markers and Hume labels.

Figure 4.1 demonstrates a heatmap of the Pearson correlation coefficients between selected vocal features and Hume AI emotion labels across all clips in the dataset. The results show generally weak correlations, with most values in the range of -0.4 and 0.3.

Key findings:

- Mean pitch (Hz) shows a moderate correlation with sadness (r = -0.37).
- Mean Intensity (dB) has a moderate positive correlation with joy (r = 0.34) and a moderate negative correlation with surprise(r = -0.39). These results align with the referred research indicating that vocal intensity tends to increase with positive arousal states (Ekberg et al., 2023).
- Mean HNR has a moderate negative correlation with sadness (r = -0.44), but shows weak correlations across all other emotions.
- Shimmer shows a moderate negative correlation with anger (-0.33).
- Jitter and shimmer showed overall weak correlations, which may reflect that these features are subtle in spontaneous speech contexts compared to controlled or acted settings.

Overall, the correlations with Hume suggest minor tendencies that reflect known vocal-emotion relationships, especially regarding pitch and intensity. The low degree of these correlations indicates that selected acoustic features alone did not strongly predict AI-detected emotions.

4.2.2 Correlation with Praat-Based Emotion Scores

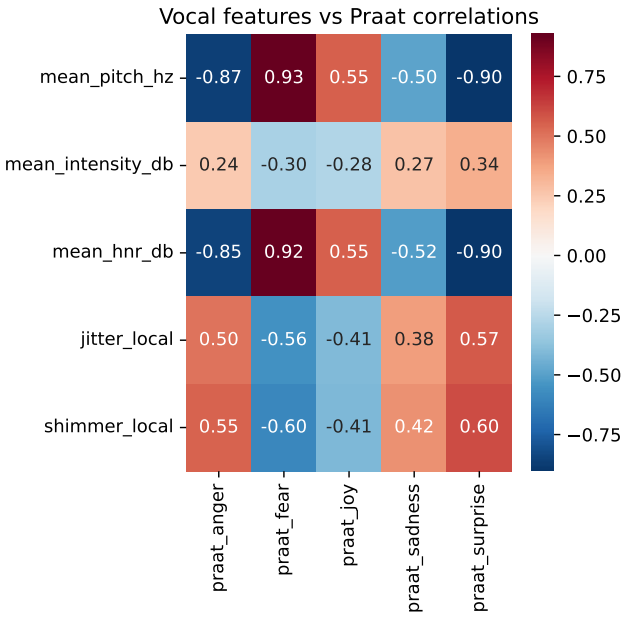


Figure 4.2: Heatmap of correlation between vocal markers and custom emotion categorization.

Figure 4.2 illustrates correlation between vocal features and the emotion scores obtained from the custom Praat-based categorization function. In contrast to the results for Hume, these correlations are notably higher. However, it shows patterns that are inconsistent with theoretical expectations. Key findings:

- Mean pitch (Hz) shows very strong correlations (r = -0.87) with anger and fear (r = 0.93), which suggests that pitch strongly impacted the outcomes of the categorization.
- Mean HNR (dB) revealed similar strong correlations as pitch, with high values for anger (r = -0.85) and fear (r = 0.92).
- Jitter and shimmer display moderate to strong correlations as well, with varying directions for different emotions.

The results suggest that the Praat-based function weighted some vocal features very heavily, especially pitch and HNR, resulting in inflated correlations that may be misleading. The absence of varying differences across the emotions suggest that the rule-based approach did not capture some expressions in spontaneous, interviewed, speech.

4.2.3 Limitations of the Custom Vocal Emotion Categorization Method

To evaluate the performance of our custom emotion categorization function, which was developed based on vocal markers reported in the Swedish study (Ekberg et al., 2023), we compared the emotion labels to the labels generated by Hume AI's speech-based emotion recognition model. This comparison included both individual clip level and across the full dataset.

However, the results revealed significant limitations in our approach. Regardless of the vocal input from our dataset, the categorization function consistently rendered near homogeneous emotion scores across all five emotions. This indicates that the function failed to capture emotional distinctions within spontaneous, conversational speech during interviews, regardless of theoretical relevance.

Figure 4.3 illustrates this issue across the full dataset, where the average scores assigned by our Praat-based categorization remain clustered around 0.2 across all emotions. Opposed to Hume that had greater variation through its labeling and reflects a more dynamic emotional detection.

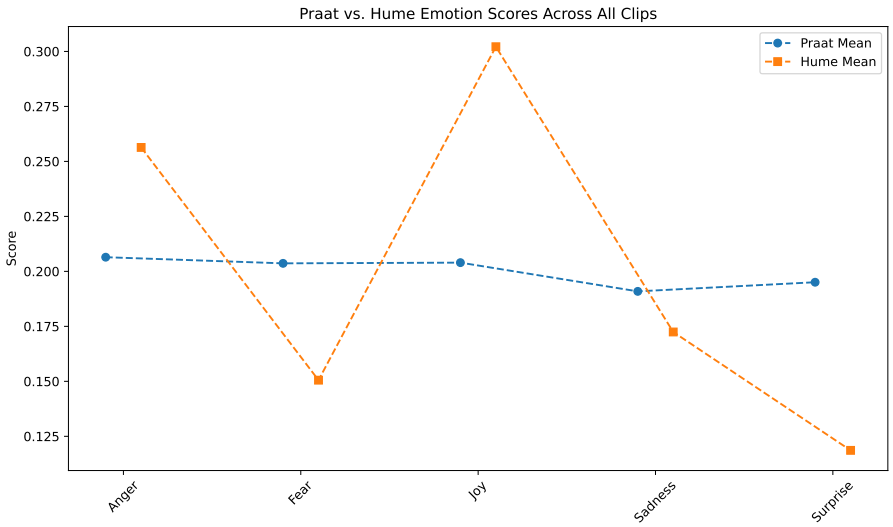


Figure 4.3: Average Score of vocal features-categorization and Hume labeling across all clips.

This is presented similarly in Figure 4.4, which presents a comparison for a single negative directed interview `id_008_neg`. In the same way as in Figure 4.3, the Praat-based scores are distributed very evenly across all emotions, while Hume assigns a higher probability to anger and lower for surprise, with joy, sadness, and fear clustered more closely. Despite the fact that the score diversity between the Hume-labeled emotions are relatively moderate - approximately 0.10 between anger and sadness/joy, and around 0.15 between surprise and sadness/joy - the probabilities are still more diverse and provide a more interpretable output. The variability could be considered more reflective of potential emotional nuances in the interview.

These findings demonstrate that our vocal feature-based categorization lacked sensitivity and adaptability when applied to our interview-based Swedish speech data. As a result, subsequent analyses focused on direct comparisons between raw vocal features and AI-predicted emotions, instead of relying on this flawed categorization method.

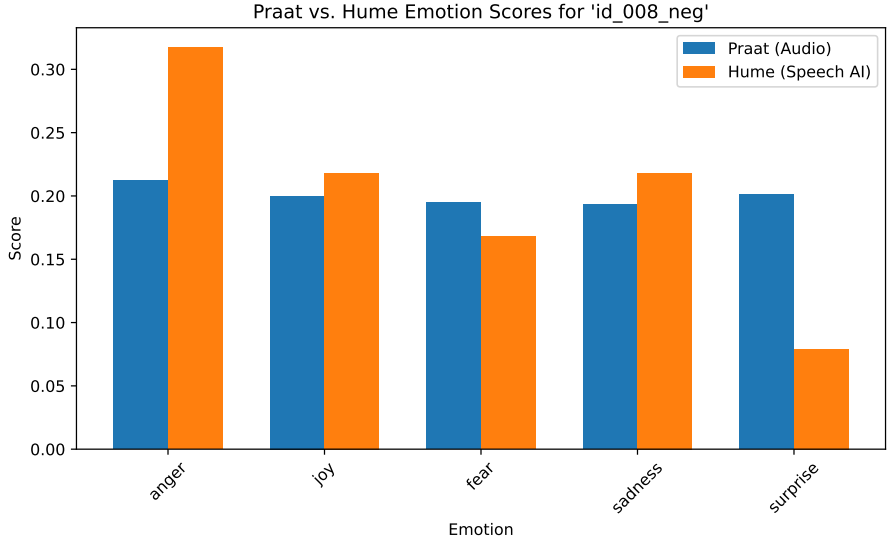


Figure 4.4: Vocal feature-categorization vs. Hume for single clip.

4.2.4 ANOVA Summery of Vocal Features Across Emotions

An ANOVA was implemented to further examine whether essential vocal features varied across AI-labeled emotions. This was conducted on pitch, intensity, HNR, jitter, and shimmer. The results are summarised in Table 4.5 and showed that none of the features showed statistically significant differences between the five Hume emotion categories (all p-values > 0.23). To confirm these findings, Tukey HSD tests were conducted and resulted in no pairwise comparisons between emotion labels with significant difference.

Feature	ANOVA p-value	Significant Differences
mean_pitch_hz	0,4435	No
mean_intensity_db	0,5793	No
mean_hnr_db	0,2327	No
jitter_local	0,7797	No
shimmer_local	0,385	No

Table 4.5: ANOVA table for vocal features variance across emotions

These results imply that within our dataset of spontaneous speech during interviews, the average values of the acoustic features did not systematically vary according to AI-labeled emotions. This could indicate that emotional expression in conversations during interview circumstances is either:

- More subtle than in controlled studies.
- Features like pitch and intensity fluctuate instead of differing consistently at the audio clip level.
- A deficient set of vocal features were extracted for these analyses.

The lack of variance in these results does not align with findings in controlled settings (Ekberg et al., 2023), where clear differences in vocal features were found between different emotional states.

4.2.5 Correlation Between Vocal Features and Hume AI Emotion Scores

Considering the limitations that had been identified in our vocal feature-based emotion categorization function, subsequent analyses shifted focus towards examining direct correlation between raw acoustic features and AI-predicted emotions. Instead of applying predefined vocal emotion mappings to rely on, essential vocal markers have been investigated to obtain an understanding of how these correlate with Hume AI's emotion scores across our dataset.

Composite Correlation Overview

Figure 4.5 illustrates a composite correlation analysis, including Pearson correlation coefficients (r) between two acoustic features - pitch and intensity - and the given emotion categories that are filtered from Hume AI. Even if the correlations are generally moderate, certain patterns appear which aligns with previous findings in vocal emotion research.

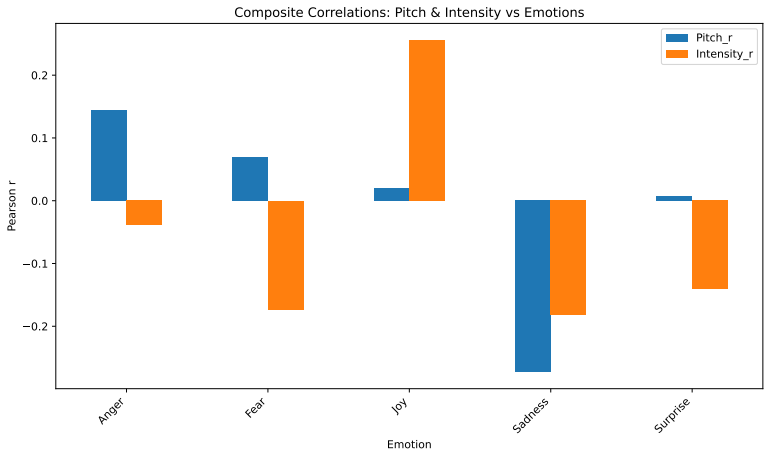


Figure 4.5: Composite correlation analysis

Particularly intensity shows a positive correlation with joy, and suggests that higher vocal intensity tends to co-occur with AI detected happiness. In contrast, intensity shows a negative correlation with emotions such as sadness and fear, which is aligned with the expectations that these emotions are generally expressed with lower vocal energy.

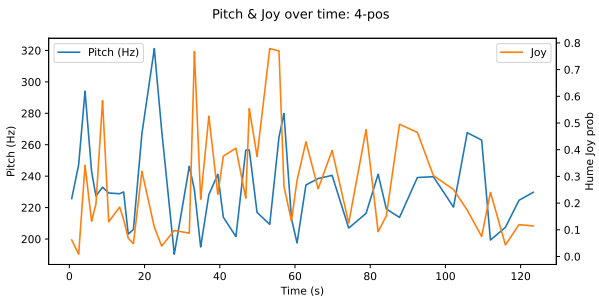
For pitch, a minor positive correlation is found in correlation with anger and fear, also reflecting expectations reported in prior research, where higher pitch is associated with heightened arousal states, for example anger. A negative correlation between pitch and sadness is shown, also supporting the prior findings where sadness is linked to lower pitch.

However, the correlation strength is weak across all emotions, without values that indicate a strong linear association. This indicates, as our previous results, that single acoustic features like pitch and intensity alone are insufficient markers of emotional states when detected by speech emotion recognition systems, in the context of conversational, but interviewed, speech.

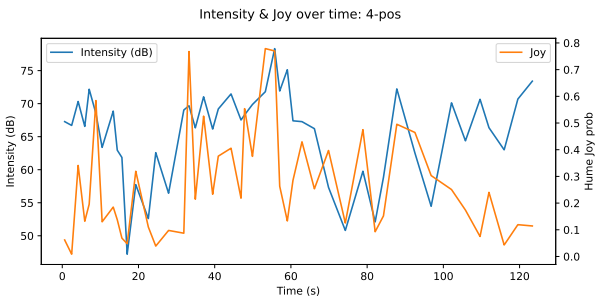
Supporting observations from individual clips

For a more concrete illustration of the prior tendencies stated, two interview recordings were analyzed in detail. The purpose was to examine whether emotional shifts become more apparent when evaluating shorter time segments within individual speakers, compared to the weaker correlations observed at the dataset level.

In Figure 4.6a and 4.6b, data is demonstrated from a positively directed interview `id_004_pos`, female, which shows that increases in pitch and intensity considerably often correlate with higher joy probabilities by Hume AI. While the correlation is not consistent throughout the recording, these moment-to-moment variations reflect the general expectation that higher vocal energy is associated with positive emotional expression.

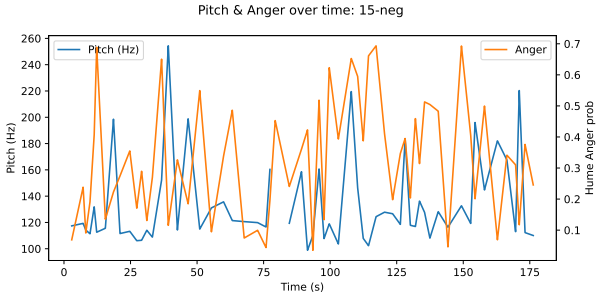


(a) Pitch(Hz) and Hume label joy over time. Clip 4-pos

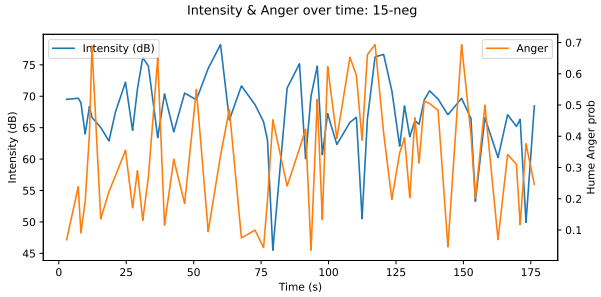


(b) Intensity(dB) and Hume label joy over time. Clip 4-pos

Similarly, Figure 4.7a and 4.7b and 4.8 present data from a negatively directed interview `id_015_neg`, male. Here, clear peaks in pitch and intensity correspond with increased anger probabilities. These results are partly aligned with prior research on vocal markers of high-arousal negative emotions, such as raised pitch and loudness during expressed anger.



(a) Pitch(Hz) and Hume label joy over time. Clip 15-neg



(b) Intensity(dB) and Hume label joy over time. Clip 15-neg

Additionally, Figure C illustrates z-score fluctuations of key vocal features for clip 13-neg. For this interview, several segments exceed +-1 standard deviation from the baseline, especially for pitch, intensity, and shimmer. These flagged moments align frequently with the emotion probabilities of Hume, which strengthens the link between vocal patterns and perceived emotional intensity.

The z-score fluctuation diagram Figure 4.8 illustrates the segments further where vocal features significantly deviate from baseline levels, in patterns aligned with associated high-arousal negative emotions such as raised pitch and loudness.

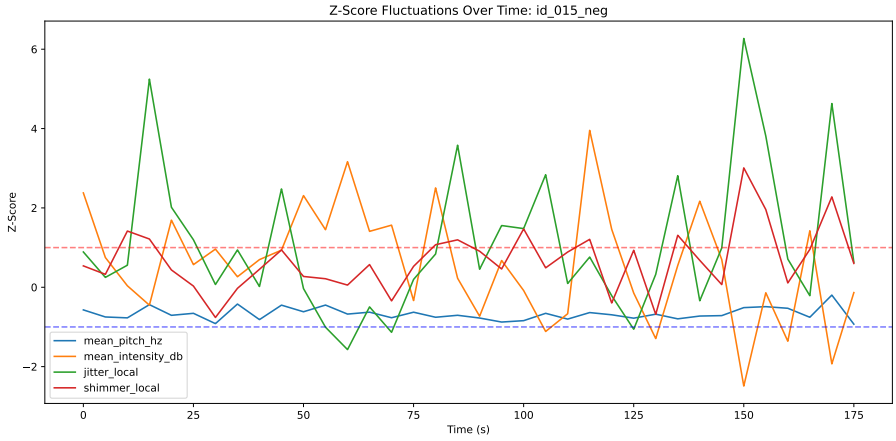


Figure 4.8: Z-score

Summery

This supports the idea that analyzing how vocal features change over time can provide more meaningful insights into emotional expression in conversational, partly spontaneous speech during interviews compared to only using overall clip-level statistics.

4.2.6 Conclusion RQ1 Data Analysis

The results revealed only weak to moderate correlations for the analysis between individual vocal features and how Hume AI predicted emotions, where intensity and pitch showed most patterns consistently. The custom vocal categorization method did not function well in this

context and resulted in very uniform results. This method was built on a basic group of vocal features which may overlooked important indicators for certain emotions. ANOVA tests found no significant differences in vocal features across AI-labeled emotions. However, examining pitch and intensity fluctuations over time segments in individual clips gave more promising results. This implies that dynamic changes in vocal features can offer more insights than static averages when analysing conversational, yet spontaneous speech during interviews.

4.3 Data Analysis for RQ2: Text and Speech Based Emotion Recognition

Research Question 2 explores how two modalities for AI-based emotion recognition systems - speech-based (Hume AI) and text-based (NLP Cloud) - aligns and diverge in their interpretation of emotional expressions for semi-structured interviews. The analysis evaluate the agreement between the models for five basic emotions by comparing average scores, measuring statistical correlation and significance, and exploring how the sentiment context (positive vs. negative) may impact detection. This multimethod approach support a comprehensive understanding of how the two modilities responds to the same emotional input, to find mutual strenghs and diverse tendencies in how they classifies emotions.

4.3.1 Overall Comparison of AI Systems

To compare the overall performance and certain tendencies of the two AI-based emotion recognition systems (Hume AI and NLP Cloud), both descriptive statistics and visual analyses were conducted to interpret the results through calculating the average differences.

As presented in Table 4.2, Table 4.3, and Table 4.4 (4.1.3), the mean emotion scores and standard deviations differ between the two models across the full dataset, including patterns within positive and negative interviews.

Figure 4.9 illustrates the average difference in emotion scores between Hume AI and NLP Cloud across the full dataset. Positive values indicate that Hume AI assigned higher scores for the respective emotion, while negative values implies higher scores from NLP Cloud. The most evident difference was shown for Joy, where NLP Cloud consistently provided higher scores compared to Hume AI. In contrast, Fear and Anger had a higher tendency to be detected by Hume AI. Differences for Sadness and Surprise were insignificant, which suggests general agreement between the systems for these emotions.

shows that Hume AI generally compose higher scores for Fear and Sadness, while NLP Cloud assigned higher values for Joy and Anger to some extent. Still, the differences were relatively small across the majority of emotions, except for Fear, where Hume showed a prominent higher pattern.

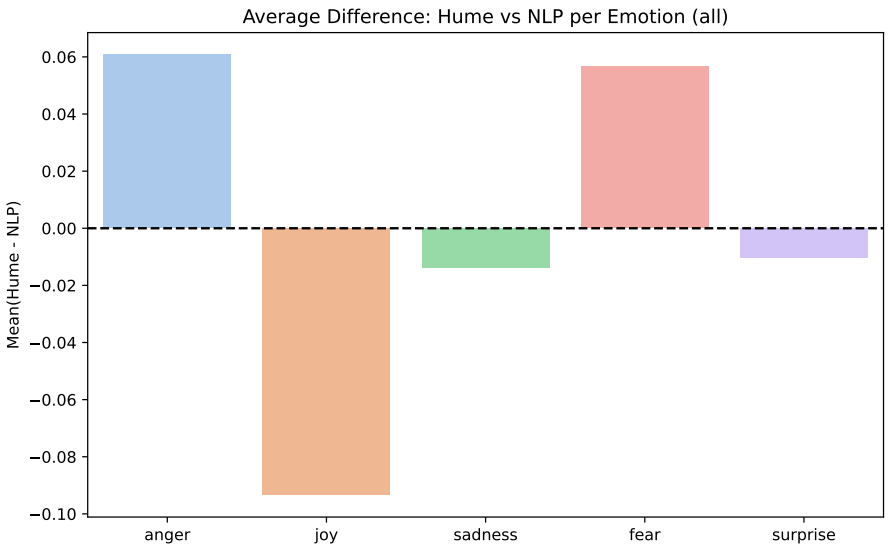


Figure 4.9: Average difference in emotion scores between Hume AI and NLP Cloud

4.3.2 Statistical Analysis

Correlation Analysis

To evaluate how text-based (NLP Cloud) and speech-based (Hume AI) emotion recognition aligns, Pearson correlation coefficients (r) were calculated for each emotion across all interview recordings. Table 4.6 presents the correlation values as well as corresponding p-values to examine the statistical significance.

Emotion	Pearson r	p-value
Anger	0,468	0,0069
Joy	0,521	0,0022
Sadness	0,166	0,3645
Fear	0,173	0,3439
Surprise	0,193	0,2889

Table 4.6: Pearson Correlation (r) and p-value for Hume AI and NLP across emotions.

This data demonstrates a reasonable positive correlation for Anger($r = 0.468$, $p=0.0069$) and Joy ($r=0.521$, $p=0.0022$), implying that these emotions are relatively consistent identified throughout the AI systems. The p-values ($p<0.05$) show a statistical significance and highlights a relevant relationship in how Anger and Joy are detected through different processes.

Sadness, Fear, and Surprise show contrasted results with weak correlations ($r<0.20$) where the p-values indicate no significancy with low agreement between the AI models for these emotions. This suggest that text-based and speech-based emotion analysis have a higher disagreement when detecting these emotion expressions.

Overall, some alignment for the more distinct emotions as Anger and Joy are declared through the correlation analysis, but some difficulties with consistent agreement are prominent for more nuances emotions as Sadness, Fear, and Surprise.

Paired t-Tests and Effect Sizes

To further explore alignment and differences between speech-based (Hume AI) and text-based (NLP Cloud) emotion recognition, paired t-tests and Cohen's d were conducted. Table 4.7 shows

the t-statistics, p-values, and Cohen's d for each emotion across the full dataset. Positive t-values implies that Hume rated that emotion more frequent than NLP, negative t-values suggest the opposite.

Full Dataset				
Emotion	**t-statistic**	**p-value**	**Significant**	**Cohen's d**
Anger	1,717	0,096	No	0,303
Joy	-1,726	0,0943	No	-0,305
Sadness	-0,548	0,5876	No	-0,097
Fear	3,341	0,0022	Yes	0,591
Surprise	-0,657	0,5158	No	-0,116

Table 4.7: t-statistics, p-value with significance, and Cohen's d for all clips.

Across all interviews, only Fear had statistically significant difference between the AI-models ($t = 3.341$, $p = 0.0022$), and had a medium effect size (Cohen's $d = 0.591$). Hume AI rated fear consistently higher than NLP Cloud, which suggest a systematic difference in expression of vocal and textual for Fear.

Even if differences were detected for Anger, Joy, Sadness, and Surprise, the variation was not statistically significant ($p > 0.05$), with small effect sizes (Cohen's $d < 0.3$). This indicates that regardless of minor divergences, the AI-models had relative alignment for recognizing these emotions, apart from Fear.

Conclusion Statistical Analysis

Comparison of speech-based (Hume AI) and text-based(NLP Cloud) with statistical analysis demonstrates notable alignment, particularly for clear expressed emotions as Anger and Joy, which is confirmed by moderate to high degree of correlations. Emotions that are more subtle like Sadness, Fear, and Surprise, yet revealed low correlations, indicating modality-specifc distinctions. Paired t-tests strengthened this observation, pointing out Fear as the only emotion with statistically significant divergence where speech-based analysis assigned higher scores consistently. Overall, the analysis confirm that the AI modalities align in explicit emotional expressions but diverge when capturing nuanced emotional states. This highlight the strengths of the complementary use of speech- and text-based emotion recognition, each revealing unique features of emotional detection.

4.3.3 Sentiment-Based Analysis

To explore how AI-based emotion recognition systems adapt to different emotional contexts, an sentiment-based analysis are conducted where the interviews are seperated by the design to provoke either positive or negative emotions. The interviews followed an autobiographical recall approach, see **??** Theoretical Framework, Interviews.

While the interviews were structured to evoke either positive or negative emotions through different scenarios, it is important to note that these categorizations do not serve as a ground truth. Emotional expression, certainly in a conversational interview setting, may not fully align with the intended sentiment through the whole recording. Additionally, Hume AI analyzed vocal characteristics (how something is said), and NLP Cloud focuses on the semantic content of the transcription of the recording (what is said), which can have an impact on how each system interprets emotional tone.

This analysis focuses on comparing Hume AI and NLP Cloud across positive and negative oriented interviews to investigate how each system acts to shifts in emotional context.

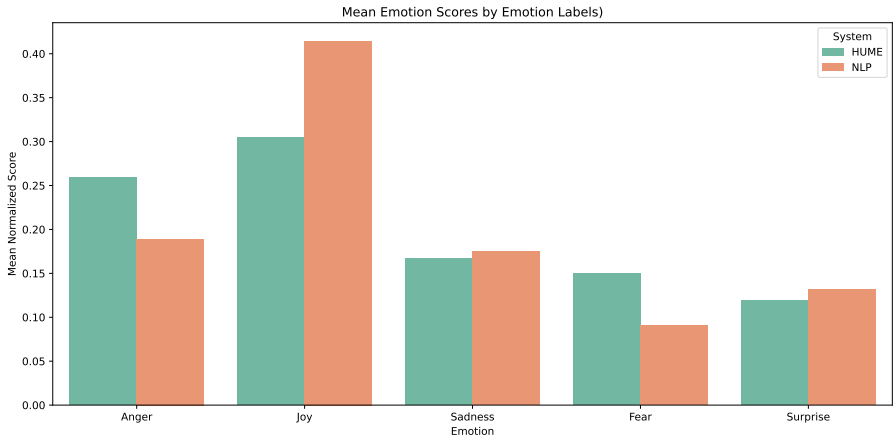


Figure 4.10: Sentiment comparison bar for NLP and Hume.

Positive Interviews Key Findings: Key Findings:

- Anger was detected at significantly higher levels by Hume compared to NLP cloud, which rated Anger near zero. This implies that Hume AI, that focuses on vocal tone, might pick up subtle vocal cues even when the participants are talking about positive memories. NLP Cloud only analyzes textual content and does not interpret Anger when negative language is missing.

- Joy is rated substantially high by NLP Cloud compared to both other emotions and Hume's probability. When participants discuss happy experiences, NLP detects the emotion in high levels. Hume's lower detection for Joy might indicate that the expressed vocal features in the interview setting are less distinct, which leads to the underestimation.

- Sadness and Fear are both rated higher than NLP, which gave these emotions minor scores. This aligns with that subtle negative undertones in speech, which could be due to reflective or serious tones, are picked up by Hume in positive contexts. NLP misses these cues since nuances are not present in the transcribed content.

- Surprise was detected at similar, low levels by both models. The emotion appears to be detected inconsistently, which may indicate that Surprise is difficult to capture, or rarely prominent in the recordings.

Negative Interviews Key Findings:

- Anger was somewhat higher rated by NLP than Hume. This indicate that NLP Cloud is sensitive to words expressed negatively, and may flag explicit language as anger. Hume focuses on vocal tone, rated Anger slightly less, which can be due to vocal expressions being more controlled in an interview setting.

- Joy was assigned higher by Hume than NLP in the negative interviews. Presumably due to the reason for the other emotions, where vocal patterns may be interpreted by Hume as positive even if its tied to a negative setting, compared to NLP that devalued this emotion in comparison, probably because of negative transcriptions.

- Sadness was detected at higher levels by NLP than Hume.

- Fear and Surprise had similar detection results, with very small variation. This suggest that the emotions were expressed more consistently across both models, and the low levels that it might not be expressed prominent in the interviews overall.

The text-based NLP Cloud is dependent on explicit emotional language, with highest rating for emotions that are clearly expressed in words, for example Joy in positive interviews, Anger and Sadness in negative interviews. The speech-based model Hume interprets vocal tones and may detect underlying emotional nuances, sometimes leading to unexpected results such as Joy in negative scenarios, Anger in positive scenarios.

Statistical Analysis

Table 4.8 demonstrates t-tests and Cohen's d for positive oriented interviews, where all emotions except for surprise shows significant differences with certainly large effect sizes. NLP have the aspects of overestimating Joy compared to Hume, where Hume in contrast tends to overestimate Anger in positive contexts. Hume rates Sadness and Fear more prominent than NLP, and Surprise remain inconsistent as previous results with no significant difference. The AI systems diverge significantly in positive interviews across almost all emotions, suggesting that text-based analysis can misinterpret emotional subtle cues, leading to inflating Joy. Hume may interpret vocal expressions in positive contexts as negative, probably due to the nature of the recordings, in agreement with previous results.

Positive Oriented Interviews				
Emotion	**t-statistic**	**p-value**	**Significant**	**Cohen's d**
Anger	10,903	0	Yes	2,815
Joy	-11,665	0	Yes	-3,012
Sadness	6,177	0	Yes	1,595
Fear	5,125	0,0002	Yes	1,323
Surprise	-1,723	0,1069	No	-0,445

Table 4.8: t-statistics, p-value with significance, and Cohen's d for positive interviews.

Table 4.9 presents conducted t-tests and Cohen's d in negative interviews, with significant differences for Joy, where Hume rates it significantly higher than NLP. In contrast, NLP has clear higher scoring for Sadness with large effects. Anger has a moderate difference, even if it is not statistically significant. No notable differences are detected for either Fear or Surprise. This implies that the AI systems strongly disagrees on Joy and Sadness detection in the negative contexts of the dataset.

Negative Oriented Interviews				
Emotion	**t-statistic**	**p-value**	**Significant**	**Cohen's d**
Anger	-1,702	0,108	No	-0,413
Joy	3,72	0,0019	Yes	0,902
Sadness	-3,796	0,0016	Yes	-0,921
Fear	0,536	0,5993	No	0,13
Surprise	1,311	0,2084	No	0,318

Table 4.9: t-statistics, p-value with significance, and Cohen's d for negative interviews

The emotional context impacts how the AI models detects emotions. Both systems presented larger differences in positive contexts, where vocal subtle cues and explicit language has different impact. NLP Cloud highly detects explicit emotions in text, especially Joy Anger and Sadness, but does not interpret subtle emotions such as Fear or Sadness in a positive context, since these emotions only might be expressed in nuanced vocal cues. Hume AI captures these vocal nuances, resulting in detecting unexpected emotional states, such as Anger in positive interviews and Joy in negative ones. Which is affected by how something is said, where the tone

can be opposing to the actual context. The speech-based model might overestimate negative emotions in reflective settings by this reason.

Positive oriented interviews presents a wider divergence between the systems, which is confirmed by the large effect sizes. The negative interviews had more consistent detection, except for Joy and Sadness that had significant differences. For surprise, detection challenges or lack of expression of that emotion in the interviews might impacting the results of no significant difference for both sentiment contexts.

Conclusion Sentiment-Based Analysis

In conclusion, the sentiment-based analysis shows that AI emotion detection in sensitive to both context and the modality. Text-based analysis has higher performance in identifying the explicit stated emotions, while speech-based analysis reveals deeper and probably more subtle emotional cues. These results suggests that none of the systems provides a universally accurate interpretation across all emotions, which indicates that text and speech-based emotion detection has limitations.

Case Example

single clip comparison

breifly illustrate how speech vs text differ in practice

4.3.4 Conclusion of RQ2 Data Analysis

The results of this research question show that even if Hume AI and NLP Cloud partially aligns in detecting emotions, certainly for clearly expressed emotions such as Anger and Joy, they diverge significantly in their predictions of more nuanced emotions such as Fear, Sadness, and Surprise. Statistical tests confirmed a significant difference for Fear. Sentiment-based analysis showed that emotional context have an impact on the results, when analysing five basic emotions, where positive scenarios had a larger model divergence. As discussed above, the interview setting and overall data collection may have different impacts on the results. Still, the findings highlights how speech- and text-based models are complementary, each with their own strenghts to capture different aspects of emotion expression, and indicate that relying on a single modality could have limitations for comprehensive emotion detection in speech.

4.4 Data Analysis for RQ3: AI and self-assessed emotion labels

The third research question explores how AI-generated emotion labels - from both speech-based (Hume AI) and text-based (NLP Cloud) analyses - aligns with self-assessed emotions. It is of importance to understand how these AI systems reflect humans perception of their own emotions and explore how the systems align with the participants own emotional viewpoint. To answer this question, participants own assessments are compared with AI-labels to gain an understanding if these systems are aligned with human self-perception of their emotions and how the modality affects the results. To explore alignment and divergence, the analysis include comparisons of average scoring, correlations, statistical analysis, and sentiment separation.

4.4.1 Descriptive Overview

For an initial overview, Table 4.10 summerizes the average emotion scores across all 32 interview recordings for each emotion category (Anger, Joy, Sadness, Fear, Surprise). The table presents mean values and standard deviation for self-resported scores aside both AI-systems.

Emotion	Self Mean	Hume Mean	NLP Mean	Self Std	Hume Std	NLP Std
Anger	0,21	0,26	0,2	0,124	0,072	0,223
Joy	0,312	0,302	0,396	0,2	0,117	0,351
Sadness	0,19	0,167	0,181	0,105	0,065	0,138
Fear	0,136	0,15	0,093	0,061	0,045	0,092
Surprise	0,149	0,118	0,129	0,082	0,022	0,089

Table 4.10: Mean and standard deviation for Hume, NLP, and Self-labels for the full dataset.

As shown in the table, Joy consistently has the highest average scores across all sources, in certain for NLP Cloud, which has a notable higher mean value (0.396) compared to self-reports (0.312) and Hume (0.302). In contrast, emotions as Fear and Surprise have a tendency for lower average scores, where both AI-systems generally assigns these emotion labels slightly lower for these emotions compared to the participants own scoring.

To visualize these differences, Figure 4.11 illustates a bar chart that compares the average emotion scores defined by participants self-assessment, Hume AI and NLP Cloud.

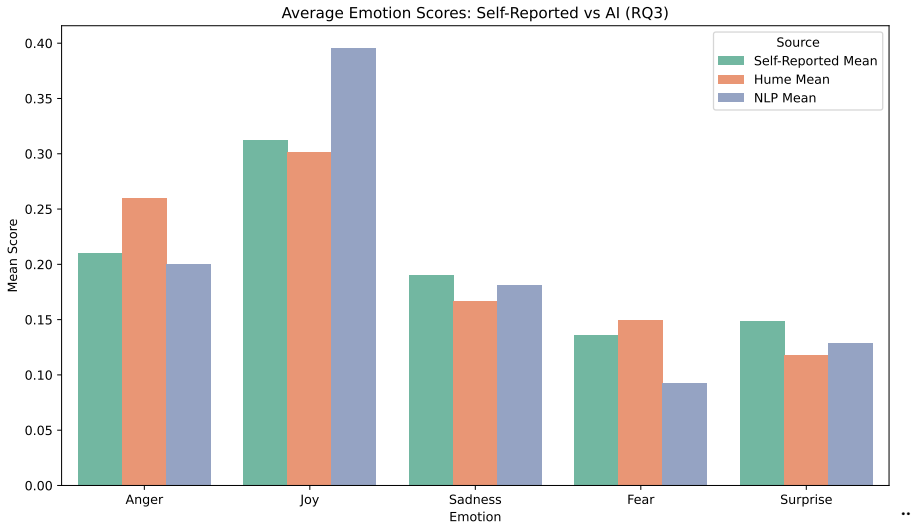


Figure 4.11: Comparison of emotional labels for Hume, NLP, and self-assessed.

Key patterns from bar chart 4.11:

- NLP Cloud tends to overestimate Joy compared to both self-reported and Hume's emotion labeling.
- For Anger, self and NLP-labeling are highly congurent, while Hume assigns higher average scores.
- Regarding Sadness and Surprise, the AI systems tends to report moderately lower scores than the participants, particularly for Surprise.
- Sadness, Fear, and Surprise are consistently rated lower across all sources, especially Fear and Surprise, where Hume tends to rate Fear more frequent than the other, and self-labels has a higher surprise score than the other sources.

This comparison suggests that even if general alignment in emotional ranking occur, where Joy is most prominent across all sources, both AI systems demonstrate differences compared to human self-perception. Particularly NLP Cloud that appears to be more prone to rating Joy, while Hume provides more moderate scores, although all sources tends gravitate towards high scores for emotions Anger and Joy.

These insights compose the framework for further statistical analysis, where correlation analysis and significance tests will be used to evaluate the strenghts and consistency of these patterns.

4.4.2 Correlation and Visual Analysis

To evaluate the alignment between AI-generated emotion scores and participants self-reported emotions, Pearson correlation analyses were conducted across the five emotion categories for both speech-based (Hume AI) and text-based (NLP Cloud) compared to self-reporting. With these measurements the relationship's strength and direction and the statistical significance can be reviewed.

Hume AI vs Self-Reported Emotions

The correlation results for Hume AI is demonstrated in Table 4.11, and indicate generally weak correlations across the majority of emotions. Anger is the only emotion showing a statistic significant correlation ($r = 0.359$, $p = 0.043$), which indicates a moderate alignment between Hume AI's speech based emotion detection and participants own perception for this emotion. Other emotions, such as Fear ($r = 0.007$, $p = 0.969$), presents no relevant correlation.

It is of importance to note that Hume AI analyzes vocal expression patterns, such as vocal bursts and porosody, rather than the semantic content of the recorded speech. Considering the interview setting that the recordings which complies the data collection, where participants recalled past emotional experiences in a controlled environment, the extent to which emotions were expressed vocally may vary, that potentially incluence the correlation results.

Hume vs Self-Assessed		
emotion	**pearson_r**	**p_value**
anger	0,359	0,043
joy	0,334	0,062
sadness	0,050	0,784
fear	-0,007	0,969
surprise	0,088	0,631

Table 4.11: Correlation and p-value for Hume AI and self reporting.

NLP Cloud vs Self-Reported Emotions

NLP Cloud demonstrated strong and statistically significant correlations for four of five emotions, see Table 4.12.

Key Findings:

- Joy showed a very strong correlation ($r = 0.863$, $p < 0.0001$).
- Anger and Sadness had very high correlation values (Anger: $r = 0.793$, Sadness: 0.710), with extremely low p-values for both emotions.

NLP vs Self-Assessed		
emotion	**pearson_r**	**p_value**
anger	0,739	0,00000
joy	0,863	0,00000
sadness	0,710	0,00001
fear	0,669	0,00003
surprise	0,092	0,61569

Table 4.12: Correlation and p-value for NLP Cloud and self reporting.

- Fear had a moderately strong correlation (r = 0.67, p = 0.00003), with considerably statistical significance.
- Surprise had no statistical significance and weak correlation (r = 0.092, p = 0.616).

Text-based emotion detection with NLP shows a explicit correlation with self-reported emotions for four out of five emotions. The emotion scoring certainly similar for Anger, Joy, Sadness and Fear. Surprise was the only emotion that did not show a significant correlation for NLP and self-perceived emotion labels. Surprise is the one emotion both AI-systems concurrently failed to show a significant correlation, which suggests that the models have difficulties in detect these emotions accurately. However, this inconsistency may not reflect limitations of the AI-systems. During data collection, several participants expressed challenges in assessing their level of Surprise and had difficulties understanding how to rate that emotion, this indicates both potential variablity and misunderstanding in self-reported scores for Surprise. Therefore, the low correlation for Surprise could be associated to unreliable self-assessments, rather than a underlying weakness in AI-based detection.

Visual Correlation

Figure 4.12 illustrates the correlation between self-reported Anger scores and AI-labeled predictions. As shown, Hume AI shows a weak to moderate positive tendency (r = 0.36, p = 0.043), still reaches statistical significance, even if the data points are more dispersed around the trend line. In contrast, NLP Cloud shows a strong and statistical significant correlation (r = 0.74, p < 0.0001), which is visually very noticable with the tightly clustered linear relationship.

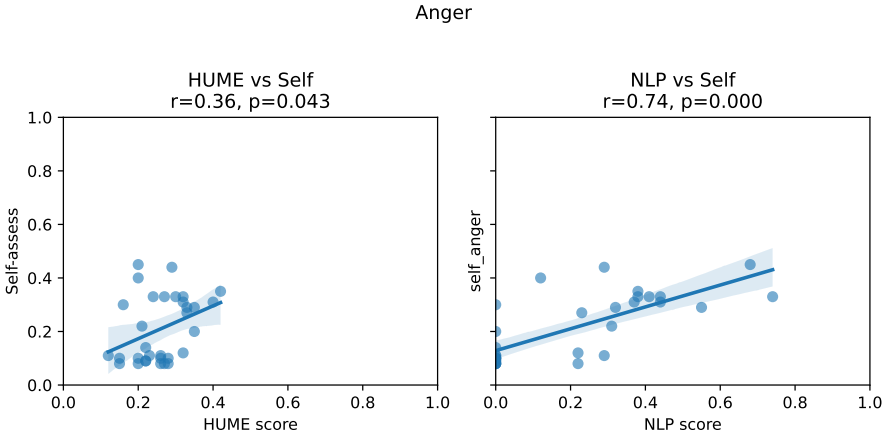


Figure 4.12: Scatter plot, Hume, NLP vs. Self for Anger.

Figure 4.13 demonstrates the Joy correlation between self-reports and AI-predictions. Similar to Anger, Hume AI presents a weak to moderate positive trend (r = 0.33, p = 0.062),

however it does not reach statistical significance. The diagram illustrates a disperse for the data points comparable to Anger. As stated in previous results, NLP Cloud reveal a prominent statistical correlation (r = 0.86, p < 0001), clearly presented in the diagram.

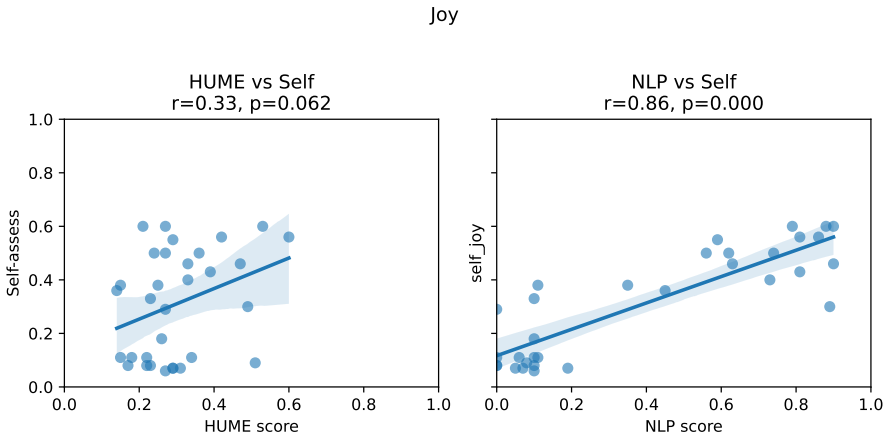


Figure 4.13: Scatter plot, Hume, NLP vs. Self for Joy.

Figure 4.14 presents the correlation results for Surprise. Both Hume AI and NLP Cloud has minor correlation with self-reported Surprise scoring (r = 0.09), the lack of alignment is reflected in the scattered plots where no direct clear linear trend is shown. As stated in 4.3.3 RQ2, Sentiment-Based Analysis, the inconsistent results may be due to Surprise being slighter expressed during the interviews. For this research question, it is important to note that several participants expressed confusion about how to interpret and answer their experienced Surprise, which can affect these inconsistent outcomes.

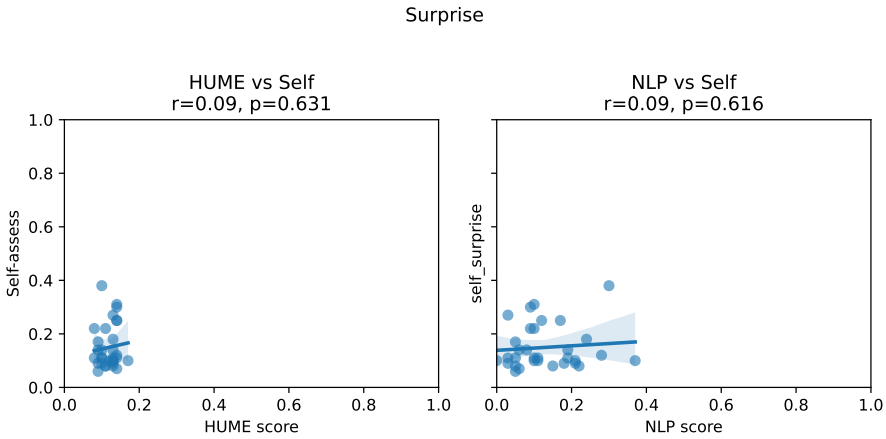


Figure 4.14: Scatter plot, Hume, NLP vs. Self for Surprise.

The scatter plots support previous presented statistical findings, highlighting the higher consistency of alignment between NLP Cloud and participant assessment for explicit expressed emotions as Joy and Anger. The shared difficulty of both models in detecting Surprise is presented further, possibly due to participant misunderstanding or that the interview scenarios were not oriented directly towards Surprise.

Conclusion Correlation and Visual Analysis

This section reveal a clear trend, both statistically and visually: NLP Cloud aligns strongly with self-reported emotions for four out of five emotions, Anger, Joy, Sadness, and Fear, while Hume

has weaker and non-significant correlations except for Anger. Both AI-models show weak levels of detecting Surprise, likely due to the interview setting and participant understanding of the question. These results emphasize the strength of text-based emotion analysis for recognition of explicit verbally emotional states, often aligned with self-reports, and implies that vocal-based interpretation within reflective interview setting may have difficulties, particularly in comparison to the participants own perceived emotion states.

4.4.3 Statistical Analysis and Effect Sizes

To explore if AI-generated emotion scores has a significant difference from self-reported emotions, paired t-tests were conducted for both Hume AI and NLP Cloud across each emotion. To evaluate the effect size of these differences, Cohen's d were calculated. These results are presented in Table 4.13.

System	Emotion	t-statistic	p-value	Significant	Cohen's d
HUME	Anger	2,399	0,023	Yes	0,424
NLP	Anger	-0,373	0,711	No	-0,066
HUME	Joy	-0,271	0,788	No	-0,048
NLP	Joy	2,331	0,026	Yes	0,412
HUME	Sadness	-1,069	0,293	No	-0,189
NLP	Sadness	-0,525	0,603	No	-0,093
HUME	Fear	1,052	0,301	No	0,186
NLP	Fear	-3,496	0,001	Yes	-0,618
HUME	Surprise	-2,109	0,043	Yes	-0,373
NLP	Surprise	-1,011	0,320	No	-0,179

Table 4.13: T-statistics, p-value and Cohen's d for AI-models and self-assessed emotions.

Key Findings:

- Hume AI:
  - Anger ($p = 0.023$) and Surprise ($p = 0.043$) showed significant differences.
  - Both Anger and Surprise showed small to moderate effect sizes (Cohen's $d > 0.4$), which suggests that Hume AI tends to overestimate Anger and underestimate Surprise compared to self-reported emotion scores. However, as declared in NLP vs. Self-Reported emotions 4.4.2, particpants stated confusion regarding assessing Surprise.
  - For the other emotions, Joy, Sadness, and Fear, no significant differences were found, suggesting closer average alignment for these emotions.
- NLP Cloud:
  - Joy ($p = 0.026$) and Fear ($p = 0.001$) presented significant differences.
  - Fear had a particularly evident difference, (Cohen's $d = -0.618$) states a medium-to-large effect size, indicating that NLP Cloud underestimated Fear compared to self-reported emotion scores.
  - Joy had a small-to-moderate effect size (Cohen's $d = 0.412$), suggesting that NLP overestimated Joy compared to the participants own perception.
  - Anger, Sadness, and Surprise showed no significant difference.

General Analysis:

Both AI systems demonstrates some alignment with self-reported emotions, the results indicate certain biases. Self reported emotion scores appears to be more prone to diverge from Hume AI's detection of Anger and Surprise. Regarding NLP Cloud, the emotion scores are distinct from self-assessed emotions for Fear and Joy, where Fear is most prominent. This could indicate either challenges in text-based detection of these subtle, negative emotions, or challenges for the participants to assess these emotions.

Even where statistical significance is found, the effect size implies that most differences are small to moderate, which suggests that even if deviations exists, they are not disturbingly large. It is no consistent significant difference for the AI-systems across all emotions, proposing that AI performance may vary depending on emotional category and the method (speech vs text, and different AI-models). However, it is important to emphasize that these results may not fully reflect the performance or accuracy of the AI models, as the analysis are based on a small dataset with semi-structured interviews. Furthermore, disparity and miscalculations may arise from challenges participants experienced in assessing their own emotions after each interview.

Overall, these findings suggest that AI systems can approximate human emotional self-assessment with partial alignment, some differences still appear depending on emotion and detection method. Hume AI had more diverse scores for Anger and Surprise, while NLP Cloud was more distinct for Fear and Joy. The differences may reflect some challenges in AI detection or difficulties for the participants to assess their emotions accurately.

Even if some differences were significant statistically, the effect sizes indicate that they are generally small to moderate. There is no consistent pattern across all emotions which indicates that the AI performance vary by emotion category and analysis method (speech vs text).

Considering the small dataset, the subjective nature of self-assessment, and the recording circumstances of the interview setting, these findings should be interpreted carefully and should be seen as exploratory and not conclusive.

4.4.4 Sentiment-Based Analysis

To explore how AI-generated emotion labels and self-reported emotions align, a sentiment-based analysis have been conducted through seperation of negative and positive interviews. Each participant reported scores (1-6 scale, normalized to 0-1) for all five emotions after each interview-scenario, to provide a subjective component to compare with AI predictions.

Figure 4.15 illustrate the average emotion scores for positive and negative interviews for Hume AI, NLP Cloud, and self-assessment. When interpreting these results, it is important to note that participants actual emotional expressions, particularly vocally, may not be fully aligned with the intention to evoke distinct emotional tones, due to the reflective and conversational nature of the setting.

Positive Interviews

- Joy was highly self-reported consistently, which aligns closely with NLP Cloud's text-based ratings, but considerably higher than Hume AI's vocal analysis. This alignment suggest that participants expressed positive emotions verbally, which was captured by NLP Cloud effectively. The lower Hume ratings for Joy indicate a divergence since vocal expressions of Joy in the reflective, interview setting may be less pronounced and therefore harder for the speech-based model to detect.

- The negative emotions (Anger, Sadness, Fear) were assessed at lower levels by participants, as expected for positive oriented interviews. Self-assessments generally matched Hume's higher detection of subtle negative emotion expressions better than NLP's minor

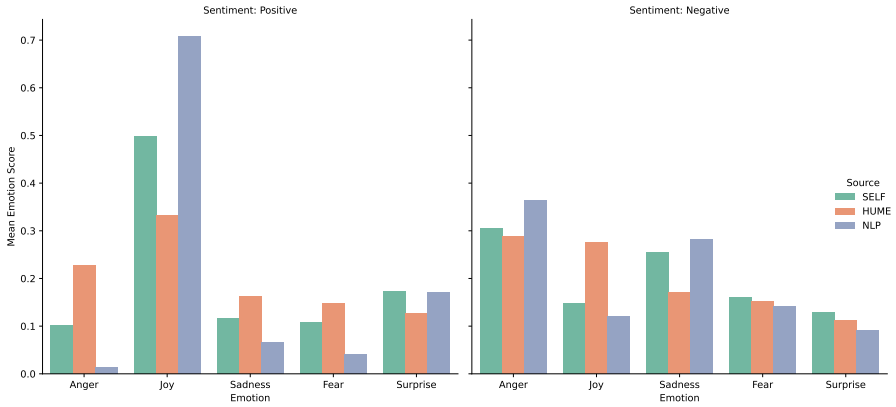


Figure 4.15: Emotion scores for all sources grouped by sentiment.

ratings. This aligns with the idea that participants may express suble negative tones unintentionally in their voice, even if the participants perceived their emotions as predonimantly positive.

- Surprise showed moderate self-report levels, more close agreed to NLP and slighly less alignment with Hume, reflecting inconsistency in vocal vs. textual emotion cues.

In positive interviews, NLP Cloud consistently rated Joy higher, in agreement with previous results. The difference from self-reported joy is similar, although NLP tends to rate it moderately higher. This reflect the strength of text-based detection for explicit positive sentiment when analyzed as a transcription.

Negative Interviews

- Anger and Sadness were reported at higher levels by participants, which reflects the intention of the interview of recalling negative emotions. Both AI models were aligned with self-reports for Anger, even if NLP rated it somewhat higher, probably due to explicit negative language in transcriptions. Sadness showed stronger alignment with NLP than Hume, indicating explicit expressed sadness-related emotions verbally.

- Joy was reported low by participants, but frequently assigned higher by Hume than the other sources. Suggesting that vocal features were interpreted as positive by Hume might not align with subjective emotional experiences, showing a clear divergence between vocal emotion cures and participants own perception.

- Fear and Surprise showed low levels consistently across self-reports and both AI-models, indicating that these emotions were less expressed or experiences, resulting in minimal discrepancies.

These results emphasize that emotional expression is complex in conversational settings, showing that textual anlyses track explicit articulated emotions, while speech-based analyses reveal less intentionally expressed emotional nuances, sometimes in disagreement with participants subjective assessments.

The sentiment-based comparison clearly presents that emotional expression and self-awareness have a signficant variance between modalities. Explicit emotions articulated in words are closely aligned between self-assessed rating and text-based analysis. while implicit or suble emotions expressed through vocal tone have a notable divergence. This highlights that AI modalities complement rather than duplicate emotional understanding. Speech-based models provide insights into emotional subtleties that might be unconscious to the participant, while text-based analyses capture verbalized emotions closely to subjective experiences.

4.4.5 Conclusion of RQ3 Data Analysis

The results show that both systems have partial alignment with self-reported emotions, however, the correlation varies depending on emotion and modality. Text-based (NLP Cloud) presented a strong alignment with participants own perception for clear verbally expressed emotions like Joy And Anger. Subtler vocal cues were interpreted by the speech-based model (Hume AI), which may not always reflect self-perceived emotions. Emotions like Fear and Surprise had most visiable differences for both AI models. Overall, the results implies that speech- and text-based emotion recognition comprehend different layers of expressed emotions, suggesting that complementary use of modalities can offer a more nuanced understanding.

Discussion

5.1 Result Discussion RQ1

For the first research question, this thesis investigated how AI models for speech recognition compare to existing research on vocal markers. More specifically, the goal is to assess whether the AI models align with the findings of the Swedish research on vocal markers done by Ekberg (Ekberg et al., 2023).

5.1.1 Interpretation of Results

Vocal Features and Hume AI Emotion Scores

Correlations between different vocal markers (pitch, intensity, HNR, jitter and simmer) and Hume AI emotion labels, were visualized on a heatmap using Pearson correlation coefficients. Overall, low to very low correlations were found here, with only a few instances of what could be considered moderate correlations. Where values closer to 1.0 for positive linear relationships or -1.0 for negative linear relationships would suggest a strong correlation, the heatmap showing the correlation between the vocal features and AI emotion scores from Hume AI primarily showed low numbers ranging from -0.44 as the lowest, to 0.34 as the highest. Some numbers appeared to have an extremely low correlation. For example, the correlation between sadness and shimmer was as low as 0.00, with other correlations hovering around being 0.01-0.04. This suggests a weak linear relationship between the vocal features mean pitch, mean intensity, mean HNR, jitter and simmer and the emotion labels from Hume AI, indicating that the AI model may not fully have captured the complex details from the vocal markers as well as the findings from Ekberg (Ekberg et al., 2023).

Despite this, some of the results align with the findings of Ekberg. For example, in this analysis mean pitch with anger shows the value r = 0.26 while mean pitch with sadness gave the value r = -0.37. Ekberg's research showed anger to have elevated pitch and while sadness had lower pitch than anger and happiness, so both studies show anger to be associated with elevated pitch and sadness to be associated with lower pitch. Looking at intensity, this analysis showed the value of joy to be r = 0.34, sadness r = -0.26, and surprise r = -0.39. In Ekberg's study, happiness (joy) showed higher intensity while sadness and surprise showed patterns of lower intensity. Although these emotions show matching patterns, other emotions show mismatches. Ekberg's study reported higher intensity for anger and fear, while the heatmap analysis in this study presents fear to have a minimal correlation (r = 0.01), while anger has a negative correlation (r = -0.18) with intensity, suggesting that intensity decreasing as anger increases.

The findings for HNR only partially matched with the findings of Ekberg's study which reported that fear and happiness were linked with higher HNR while sadness is associated to lower HNR. In this study sadness showed a moderate negative correlation with HNR (r = -0.44), which is aligns with the Ekberg study. However, correlations for fear (r = 0.08) and joy (r = 0.06) were very weak, differing from the associations observed by Ekberg. Jitter showed no moderate or strong correlations with any emotions in this study, and shimmer proved to be slightly positive for joy (r = 0.19) and moderately negative for anger (r = -0.33) but not for

surprise which was the only emotion in Ekberg's study which had higher shimmer.

Although some Pearson correlations presented in the heatmap showed consisting pattern with the Ekberg research and others diverged, it is important to note that there are methodological differences between this analysis and Ekberg's research. Ekberg employed a different statistical approach using both simple and multiple logistic regression models to predict the emotions from speech. While Pearson correlation which was used for this analysis is useful for detecting linear associations, it might not have captured the complex non-linear interactions that the logistic regression models are able to capture. The weak correlations found in the heatmap showing the correlation between the vocal features and AI emotion scores from Hume AI might also suggest that the vocal features used in this analysis were insufficient in predicting the emotions in spontaneous speech. Where the research done by Ekberg uses an acted dataset with repeated sentences, emotional expression in natural speech (e.g. interviews, which were used in this analysis) tends to be more subtle.

Praat-Based Emotion Scores

As for the second heatmap presented in the result, we see the correlations between vocal features and the emotion scores from the custom Praat-based categorization function. While the Pearson correlation values in this heatmap are seemingly stronger than with the Hume AI labels, these results may be misleading as they reflect an over-reliance on pitch and HNR, rather than authentic emotional differences. For instance, examining the highest correlations, pitch correlated strongly with fear (r = 0.93) and surprise (r = -0.90). HNR also demonstrated high correlations with fear (r = 0.92) and surprise (r = -0.90).

While the correlations here suggest a strong relationship between the emotional labels and the vocal features, this specific pattern indicates methodological limitations. The Praat-based categorization function appears to prioritize a narrow range of features (notably pitch and HNR), leading to inflated correlations that do not necessarily capture the complexity of emotional expression.

The failure of accurately categorizing vocal emotions is likely due to the spontaneous nature of the interviews which unlike acted datasets where emotions are exaggerated, result in subtle emotional expression. The limited number of vocal features may also be a contributing factor as to why full emotional complexity was not captured, together with variability in recordings which potentially diluted the emotional markers across time. Contrary to theoretical expectations, the categorization failed to distinguish between the emotional states in a meaningful way as indicated by the lack of clear emotional differentiations. These findings suggest that although some acoustic features were captured effectively, the Praat-based function did not categorize the emotions accurately, potentially due to the speech being in a spontaneous interview format instead of, for example, an acted dataset.

Custom Vocal Emotion Categorization Method

To assess the effectiveness of the custom emotion categorization function developed for this study, the outputs were compared to the Hume AI generated emotion scores.

Despite individual vocal features having high correlations, the Praat-based emotion categorization function reveals significant limitations. With an average score of 0.2, the Praat based function presented minimal variability, suggesting that the function was unable to differentiate emotional expressions within the spontaneous speech obtained from the interviews.

The Hume model demonstrated wider variations for the emotion scores, reflecting on a more nuanced detection of the different emotions. Several factors likely contributed to this outcome as these results highlight the challenge of categorization in spontaneous speech. In real world

emotional expressions, as opposed to acted ones, emotional signals are often more subtle and dependent on context.

The limited set of acoustic features may as well have excluded some important clues, further restricting the sensitivity of the function. All factors point to spontaneous speech potentially being too complex for this function.

ANOVA Vocal Features

To further investigate whether fundamental vocal features used in the previous analyses varied systematically across the five different AI-labeled emotions, an analysis of variance (ANOVA) was conducted. None of the results revealed any significant statistical differences for pitch, intensity, HNR, jitter or shimmer. This was further confirmed by Tukey HSD tests, suggesting that within the context of spontaneous speech obtained from the interviews, emotional states may not reliably be differentiated by average values of the core vocal features.

Clear differences in acoustic features between emotions were reported by Ekberg (Ekberg et al., 2023), which contrasts with these findings, possibly due to the controlled nature of Ekberg's acted dataset. The lack of variance for the present study likely reflects the subtle nature of emotional expressions in an interview format, being a natural conversational form of expression with some emotions possibly being interwoven with context and which may also have more variations from time to time.

Correlation Between Vocal Features and Hume AI Emotion Scores

Following the limitations identified with the rule-based emotion categorization, the analysis shifted to examining direct correlations between raw acoustic features and Hume AI's emotion scores. With a composite correlation analysis visualized, showing the two acoustic features pitch and intensity and the correlation with Pearson correlation coefficients, generally weak correlations were found across all emotions.

However, some patterns aligned with established findings of Ekberg (Ekberg et al., 2023), where intensity showed a negative correlation with fear, sadness and surprise which bears some resemblance with the results of Ekberg's study. Although pitch presented minor positive correlations with anger and fear, and happiness, these results alone suggest that average pitch and intensity alone are insufficient in capturing the complex nature of emotional expression made in spontaneous speech, as it remains too subtle. Although some expected relationships have been observed, emotions fluctuate dynamically within a clip, contributing to the difficulty of detecting consistent patterns through static averages of single features.

Observations from individual interviews

Analyses of individual interviews revealed results can give more clarity by analyzing a voice recording in its entirety combined with looking at peaks in specific moments. Where static measures often result in an average of emotions when analyzing an entire voice recording, segment analysis done from time to time shows more detailed emotional shifts. Although correlations are not entirely consistent throughout the recordings, the positive interview revealed increasing intensity corresponding with higher joy, whereas one of the negative voice recordings revealed peaks in pitch aligned with elevated anger scores. There are several possible factors that can account for this result, one being that the emotional expression in natural conversations such as the interview format for this study is context driven with emotions fluctuating throughout the entire conversation. Static measurements take an average of all the peaks and fluctuations, likely resulting in a somewhat neutral end result, where datasets with acted recordings

might not differ as much in one recording. These findings highlight the importance of analyzing emotions on a segment level rather than relying exclusively on an average for one whole clip. Looking at one clip at its entirety for a natural conversational clip gives the opportunity to inspect all emotional peaks and distinguish between different emotional states throughout the recording without emotions being averaged and neutralized. Furthermore some variability across individual speakers was observed demonstrating different vocal patterns suggesting that some personal vocal traits can impact the level of detectability of emotions. Some of these patterns could be due to factors such as gender or speaking style.

5.1.2 Limitations and Explanations

There are several limitations which should be taken into consideration when interpreting the results presented for the first research question.

Firstly, the dataset conducted and used for this study consist of spontaneous conversational semi-structured interviews. This likely has resulted in more subtle emotions than an acted dataset would contain with weaker emotional expressions reducing the detectability of different vocal markers. This study has also been restricted to a small set of acoustic features for the analysis. There are possibilities other acoustic features would have contributed with relevant aspects for the analyses if included. This study does not involve all vocal features that are included in the research by Ekberg (Ekberg et al., 2023) which potentially could have shown relevant information and possible alignments, however, due to the timeframe and scope of this study only five vocal markers were chosen.

Another important limitation to consider is the static averages of the vocal features. Despite effort where some clips have been visualized in its entirety, showing spikes in one specific emotion combined with one specific vocal feature from time to time, most clips involved in the process of answering the first research question are the result of an average per clip. This may have obscured some dynamic fluctuations of all emotions over time.

A further constraint worth noting is that the Hume AI emotion scores are not perfect estimates of true emotional states themselves and should not be considered ground truth in this study, only a way of comparison and finding potential similarities. Additionally, although Hume showed varied and appropriate emotion outputs, it is important to note that the Hume AI emotion outputs are based on soft scoring. This often produces mixed emotions rather than single emotional states, making comparisons more difficult.

Finally, the usage of the Swedish vocal data collected from semi-structured interviews. These interviews consist of spontaneous and conversational speech which likely do not involve as strong emotional expressions as acted datasets. Conversational speech tends to be more subtle and may have a reduced level of clear vocal markers.

Worth noting is that the interviews conducted for this study were context-dependent and influenced by topics selected by the participants of the interviews themselves, introducing potential additional variability from interview to interview. Contrary, the Swedish research by Ekberg consisted of 14 repeated sentences.

5.1.3 Conclusion for RQ1

The result for the first research question has investigated if AI-based emotion recognition models align with existing research on vocal markers with a focus on the Swedish language. Exploring vocal markers correlation with Hume AI emotion labels as well as the correlation between vocal markers and Praat, the overall result demonstrated limited strength. Overall the results showed weak or moderate correlations, although some relevant patterns aligning with the existing research on Swedish vocal markers (Ekberg et al., 2023) was found. While the values from

the Hume correlations appeared moderate at best, the Praat correlations overall showed weak correlations except some misleading numbers suggesting an over-reliance on the vocal markers pitch and HNR.

The segment level analysis gave important insight into the fluctuations in the emotions throughout entire clips compared to an average value of the different emotions.

The result indicates that while there are certain vocal features that remain relevant as indicators of emotional states, spontaneous speech presents challenges in emotion recognition. In comparison to acted datasets, emotions are more subtle with more variety and contextual dependence in spontaneous speech. Though there are challenges with spontaneous speech, the result suggest that AI-based emotion recognition systems such as Hume AI showed promise, demonstrating some flexibility and context-awareness.

Future work could benefit from incorporating a wider range of vocal features, emotions and a more dynamic approaches to capture the complexity of emotional expression.

5.2 Result Discussion RQ2 and RQ3

For the second research question in this thesis the aim was to investigate whether we could understand the emotions from textual content of the speech, with the same data as in RQ1. This was achieved by transcribing the vocal recordings and analyzing them using NLP Cloud's emotion recognition to detect emotions in the textual content of the speech.

For the third and final research question, the objective was to assess how the AI generated emotion labels obtained through the speech-based and text-based emotion recognition would compare to the self-reported emotions provided by the interviewees.

5.2.1 Interpretation of Results

5.2.2 RQ2: Speech-based AI vs Text-based AI

In the comparison between the speech-based emotion recognition model, Hume AI, and the text-based emotion recognition model NLP Cloud, the system overall seemed to show some levels of agreement for certain emotions. Using both descriptive statistics and visual analyses to calculate the differences, an overall comparison of both ai systems showed that the mean emotion scores differ across the two models. The average difference in the emotion scores showed values indicating that Hume AI obtained higher scores for the emotions anger and fear, while NLP Cloud proved to show higher scores for joy, sadness and surprise. Despite these findings, the score for sadness and surprise were sufficiently low, suggesting that the models were substantially aligned on specifically those emotions. Joy being highly scored by NLP Cloud indicates that joy may not have been as easily identified in speech-based emotion analysis, while the textual context may have conveyed a more positive tone from the text than appeared in the voice. In contrast, anger and fear appeared to have been more effectively captured by the speech-based emotion detection, possibly suggesting that someone might sound angry or fearful even though they may not be experiencing these emotions in the moment. Based on the Pearson correlation analysis showing the association between the text-based and speech-based emotion recognitions, the strongest alignments were shown for Joy ($r = 0.521$) and anger ($r = 0.468$). Joy and anger also showed statistically significant p-values, where joy had a p-value of 0.0069, and anger had a p-value of 0.0022. No further strong correlations or statistically significant p-values were found in the other emotions. Several factors may account for this result. For example, joy and anger are distinct emotions, while sadness, fear and surprise may likely involve more subtle cues and contextual factors. Being more complex to detect may have contributed to the lower consistency across the two models for these specific emotions.

For the full dataset, paired t-tests showed no significant differences for the mean score of the emotions across the dataset for all emotions except fear. Although the correlation between Hume AI and NLP Cloud showed non-significant scores for fear, the t-test indicated that while the systems do not align on detecting patterns for fear, Hume AI consistently rates the fear higher than NLP Cloud. Possible explanations for this result may reflect the differences in how emotions are conveyed and detected in the different models, whereas Hume AI possibly could have captured the more subtle vocal indicators that might not have been as easily expressed or detected in text.

Examining the t-tests for the positive oriented interviews in comparison to the negative oriented interviews, notable findings emerged. For the positive interviews, significant differences between Hume AI and NLP Cloud were found for all emotions with the exception of surprise, where Hume consistently detected higher levels of sadness and fear and NLP on the other hand overestimated joy and anger in comparison to Hume. This may be explained by the complexity of emotions and emotional expression. In the positive interviews, the participants discussed joyful topics, and while this may have been detected for the text-based emotion recognition, the vocal tone could reveal more subtle cues in the tone, rhythm and pitch. For a positive interview, the participant may have a lower and more neutral tone and pitch than an actor acting out happiness, which could be one explanation for this result. For the negative interviews, significant differences were only identified for joy and sadness, where Hume rated joy with a higher score, and NLP rated sadness higher. This indicates a better alignment for the different models for the analyses made for the negatively oriented interviews.

Possible explanations for these results are that people participating in the interviews may have used overly positive language out of politeness, even if the content of the words may have been negative.

Sentiment-Based Analysis RQ2

In comparing Hume AI and NLP Cloud, the sentiment-based analysis presented a distinct pattern in how the different AI models interpreted the positively oriented interviews versus the negatively oriented interviews, where some emotions were consistently rated higher than others.

NLP Cloud showed patterns of consistently rating joy higher than Hume AI, while Hume rated higher for negative emotions such as anger, sadness and fear in the positive interviews.

These results indicate that the subtle features such as pitch, loudness and more may have been interpreted by Hume as negative emotions even in positive conversations. Surprise remained a challenging emotion to detect, and NLP Cloud showed higher values of anger and sadness in the negative interviews, likely due to being able to better capture the negative context of the interviews through the text-based analysis. Hume rated joy unexpectedly high in the negative interviews, where a possible reason could be nervous laughter or other emotions that could have been misclassified. Overall, the results underscores that the two AI models differ in the job of emotion detection, possibly due to the vocal recordings involving subtly expressed emotions or possible irony or laughter that could have been incorrectly categorized as joy.

5.2.3 RQ3: AI vs Self-Assessed Emotions

For the third and final research question the alignment with the speech-based emotion labels from Hume AI and the text-based emotion labels from NLP Cloud in combination with self-assessed emotion scores were examined. Insightful findings revealed some levels of alignment dependent on both the model and emotion. With an analysis showing an average of the emotion scores across the entire dataset of interviews, joy emerged as the emotion with the highest average scores. Fear and surprise showed the lowest scores out of the emotions. A visualization

of the average emotion scores across all three channels shows NLP Cloud overestimating joy excessively, while Hume on the other hand overestimated anger to a certain degree. The rest of the emotions are relatively close in scores across the models and self-assessment scores, where sadness, fear and surprise were rated low for all channels. The low scores overall for sadness, fear and surprise could be explained by the spontaneous interview format in a calm setting, which may not encourage expressively conveying these emotions.

For the case of joy having a substantially higher score for NLP Cloud, the model may have interpreted language as joyful even though the tone was more neutral, also possibly missing out on cues such as irony. Hume estimated anger higher than NLP Cloud and the interviewees themselves, which may be due to misinterpreted signs of anger for example from pitch and intensity.

Hume AI and NLP Cloud vs Self-Reported Emotions

The correlations between Hume AI and the self-reported emotions indicated very weak to modest correlations for all emotions, where only anger showed a modest statistic significant correlation with a Pearson value of $r = 0.359$ and p-value $= 0.043$. This may be due to the nature of anger which often produces a distinct vocal change typically involving increased loudness and change of pitch, while emotions like fear and surprise may often be expressed with more subtle vocal expressions which may not have been captured as successfully. Once again, the calm setting in an interview environment might also have affected the results, further muting emotional expressions. This suggest that although Hume moderately detect some emotions based on voice, a multimodal approach with further analysis might be necessary for more complex emotion recognition.

The results for NLP Cloud presented statistically significant correlations for all emotions except surprise. This indicate a high degree of alignment between the self-reported emotions and the text-based emotion detection NLP Cloud, where the strongest correlations were presented in joy ($r = 0.863$, $p = 0.0000$), anger ($r = 0.739$, $p = 0.0000$), sadness ($r = 0.710$, $p = 0.0001$), and lastly fear ($r = 0.669$, $p = 0.0003$). This strongly indicates the effectiveness of NLP Cloud in capturing emotional content through text in combination with alignment with self-reported emotions gathered from the interviews.

Surprise being the only emotion not to show a strong correlation, emphasizes a consistent challenge observed over both models used in the study. During the self-evaluation segment of the interviews, multiple participants expressed certain confusion regarding the assessment of the emotion surprise. A large part of the interviews consisted of describing past emotional experiences which may have reduced the intensity of surprise. Typically, surprise is expressed as an immediate reaction to unexpected events and its unlikely that the interviewees are able to genuinely experience the same surprise felt in the original moment of the memory. This provides a possible explanation for why both AI models overall detected low levels of surprise, while an acted dataset could present higher correlations for this emotion. The statistical analysis made evident that both Hume AI and NLP Cloud showed partial alignment with the self-assessed values for some emotions. Hume AI showed a larger deviation for anger and surprise, whereas NLP Cloud deviated more for fear and joy. This suggest that the effectiveness of the AI models in emotion detection is not consistent across the emotions, although the challenging nature of self-assessing emotions, especially fear and surprise, retrospectively possibly complicates this process, effecting the results as well.

Sentiment-Based Analysis RQ3

In the sentiment-based analysis comparing Hume AI, NLP Cloud, and self-reported emotions insight was gained into how the AI models align with the personal perceptions of emotions.

In the positively oriented interviews, NLP showed higher rating of joy compared to the self-assessed scores, while Hume AI consistently rated joy lower than the self-assessed scores in combination with detecting higher levels of all negative emotions (anger, sadness and fear).

These results suggest that subtle vocal markers were captured by Hume that may not have matched the content. For surprise, NLP Cloud closely matched the scores of the self-assessment whereas Hume AI detected lower levels. This may reflect all challenges previously mentioned in regard to the emotion surprise.

In the negatively oriented interviews, both fear and surprise were relatively evenly rated across all three sources, with the self-assessed being the highest rated in both emotions.

The ratings for anger were high for all sources as well and fairly evenly matched between Hume and the self-assessed scores, while NLP Cloud rated anger higher. The higher rating by NLP Cloud was likely due to the context of negative wording in the negative interviews. Hume AI rated sadness low, while NLP Cloud was relatively close in score compared to the self-scores, suggesting the text-based model might have been better at capturing sad emotions from the vocal recordings. Hume may not have picked up the cues for sadness in the same capacity, likely due to the low expressions of sadness during the interviews. Further analysis showed Hume AI rating joy higher than both NLP Cloud and the self-reported emotions in the negative interviews, likely due to misclassifications of certain vocal cues that may have been subtle or complex, for example irony which could be difficult for a speech-based AI to recognize. Overall, the text-based AI NLP Cloud seems to align closer with the self-assessed scores rated by the participants of the interviews, possibly capturing the context for each interview more effectively. This underscores the limitations of relying exclusively on either speech-based or text-based emotion recognition.

5.2.4 Limitations and Explanations

This study has presented several important insights in emotion detection using AI models, although there are several limitations that should be noted. The spontaneous nature of the interviews remains a limitation through RQ2 and RQ3, as these vocal recordings may have given more subtle and muted emotional expressions compared to an acted dataset would. The calm setting of the interviews may also be an explanation to why fear and surprise especially was not detected to a high degree.

The self-reported emotions unavoidably involve subjective biases, which possibly could have resulted in some variety across interviews. The dataset size and the limited emotions remains a limitation, where a larger dataset and more emotions possibly could lead to broader findings and correlations.

5.2.5 Conclusion for RQ2 and RQ3

In answering RQ2 and RQ3, this study explored effectiveness of speech-based emotion recognition, Hume AI, text-based emotion recognition, NLP Cloud. These were later examined for potential alignments with self-assessed scores for emotions. For RQ2, partial agreements were found between the two AI models for some particular emotions such as joy and anger, though notable disagreements were present for fear and surprise.

This brought attention to the challenges of detecting the emotional cues for more complex emotions, where they might have more subtle cues for detection. In comparisons between the models, certain differences in how each model captured emotions were found.

For RQ3, the comparison of the models with the self-assessed emotion scores indicated a stronger consistency for NLP Cloud than it did for Hume AI. Overall, surprise was presented as an emotion consistently challenging to detect across the models, possibly due to the complexity

of the emotion combined with the nature of the interviews where this emotion may have been expressed the least.

The results for this study suggest that a multimodal approach integrating multiple sources could enhance the precision and reliability of the detection systems for emotions, as relying on only one type of analysis may not be enough for the complexity of emotions.

5.3 Method Discussion

5.3.1 RQ1 Methodological Considerations

To answer RQ1, the methodological approach involved analysis of emotional expression for vocal markers in Swedish speech in comparison to AI based emotion recognition models. The idea was to analyze emotions in a clip in its entirety and find correlations, which had some differing results, but it proved to be a notable strength to execute the analysis on a segment level to capture emotional fluctuations in a more dynamic way. While this offers another another perspective, this approach introduced challenges of its own in having some inconsistent emotions not aligning completely across the segments. Therefore the methodological approach was partially fulfilled for answering RQ1 by identifying some emotional fluctuations, while also revealing challenges in both the analyses for segment-leveled clips and full clips.

One of the studies chosen to compare the results with, being the existing Swedish emotion research by Ekberg (Ekberg et al., 2023) proved some similarities and patterns which provided valuable information to this study. However, the research used pre-defined sentences, repeated by actors, which may have given a more consistent result than the dataset used in this research which consisted of interviews capturing spontaneous speech. With 16 participants, the dataset resulted in a total of 32 recordings across a diverse group of participants consisting of men and women with ages ranging from the 23-78. The spontaneous speech and large variety of interview questions combined with dataset size may indicate some limitations for the result. While RQ1 was addressed, a larger and more controlled dataset with acted emotions along with repeated sentences, could possibly have validated some observed patterns, ensuring more consistent emotions throughout the recordings.

Hume AI was one of the models used and provided some advantages such as avoiding manual labeling and being pre-trained, although the Hume AI emotion scores had to be normalized and the emotions were filtered to use only the specific five emotions necessary for the comparisons in this research, which may have had some limitations on the model's capacity. Along with working well for the research's purpose, the model has some downsides. For example, there is limited publicly available information about functions of the model, making it difficult to fully assess possible limitations and biases.

Despite these limitations, Hume AI contributed with valuable insights in answering RQ1.

A set of basic vocal features which consisted of pitch, intensity, harmonic-to-noise ratio (HNR), jitter, shimmer, was extracted through Praat. These features are well established indicators of emotional expression but proved to be somewhat of a limitation which possibly could have been avoided by incorporating additional vocal features. Given that the dataset for this research consisted of interviews capturing spontaneous speech, a broader range of vocal features might have contributed to the detection of the complex vocal patterns and given a more nuanced understanding of the correlations for emotion recognition in the Swedish language.

While the selected vocal markings chosen for this study gave some insight into addressing RQ1, expanding the set of features could have helped address RQ1 more comprehensively.

5.3.2 RQ2 and RQ3 Methodological Considerations

The methodological approach to address RQ2 combines analysis of transcribed text in emotion recognition using NLP Cloud in order to assess the emotional content of speech transcripts in relation to speech-based AI models.

In addressing RQ3, the approach was to compare self-reported emotions with the AI-generated labels from both speech and text-based models to analyze potential alignments. This is a multi-modal approach with several methodological considerations, but also some methodological strengths.

The vocal recordings were transcribed and analyzed with the text-based emotion recognition tool NLP Cloud and self-assessments of emotions were collected after each interview, allowing a comparison between speech-based AI, text-based AI and self-assessment scores given by the participants in the interviews. The use of three different methods resulted in triangulation, which increased the flexibility and credibility in the findings. In addition to this, the usage of pre-trained AI models ensured consistent processing. However, some information loss was expected for the transcript text analyzed in NLP Cloud. When the model takes in what is said rather than how it is said, many important emotional cues such as intensity or pitch get lost. This could possibly have led to some emotions being misinterpreted or not catching the full complexity of the emotions expressed, based on only the text-based analysis.

While NLP Cloud contributed in addressing RQ2, some limitations in loss of prosodic information may have reduced the full emotional understanding. The self-reported emotions introduced a valuable reference point for this research. Some agreement was found between the AI models and self-reported emotions, but some of the self-assessed scores may also have been slightly exaggerated. The emotional memories and personal interpretations of emotions by participants can have influenced the self-assessed emotion scores. While the self-reported emotion scores have some limitations which likely contributed to some variability in the analyses, they helped valuably address RQ3.

The complexity of emotion detection across different modalities is highlighted by the AI models being able to capture some emotions in a quite robust way while struggling more with others. The few emotional categories used for this research may have limited the emotion recognition, where an implementation of more emotions and features possibly could have captured the emotions in a better way.

All methods used in answering RQ2 and RQ3 provided valuable information and findings, however the research could have benefited from an expansion of the emotion categories to help identify emotions in a more accurate and complex way.

5.3.3 Summary of Methodological Considerations

To address all research questions, this study utilized a multi-method approach combining speech-based and text-based emotion recognition with self-reported emotion scores of the participants from the interviews.

Although all methods contributed with important findings and significant insights into emotional expression in Swedish speech, a number of limitations emerged.

While the triangulation of speech, text and self-assessment scores contributed to the strength and credibility of the findings, size of dataset, model transparency and other limitations such as variabilities and inconsistency in having spontaneous interviews may have impacted the effectiveness of the findings. Although highlighting some areas for improvement for future studies, the methods chosen for this research overall contributed to answering the research questions in a comprehensive manner.

Conslusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6.1 Presentation of Collected Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6.2 Data Analysis

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Bibliography

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors Basel, Switzerland*, *21*, 1–27. https://doi.org/https://doi.org/10.3390/s21041249

Adebiyi, M. O., Adeliyi, T. T., Olaniyan, D., & Olaniyan, J. (2024). Advancements in accurate speech emotion recognition through the integration of cnn-am model. *Telkomnika*, *22*, 606–618. https://doi.org/https://doi.org/10.12928/TELKOMNIKA.v22i3.25708

Ahammed, M., Sheikh, R., Hossain, F., Liza, S. M., Rahman, M. A., Mahmud, M., Brown, D. J., Ahmed, M. R., Ben-Abdallah, H., Kaiser, M. S., & Zhong, N. (2024). Speech emotion recognition: An empirical analysis of machine learning algorithms across diverse data sets. In *Applied intelligence and informatics* (pp. 32–46). Springer. https://doi.org/https://doi.org/10.1007/978-3-031-68639-9_3

AI, H. (n.d.-a). Prosody. https://www.hume.ai/products/speech-prosody-model

AI, H. (n.d.-b). Vocal expression. https://www.hume.ai/products/vocal-expression-model

Alroobaea, R. (2024). Cross-corpus speech emotion recognition with transformers: Leveraging handcrafted features and data augmentation. *Computers in biology and medicine*, *179*, 108841. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108841

Areshey, A., & Mathkour, H. (2024). Exploring transformer models for sentiment classification: A comparison of bert, roberta, albert, distilbert, and xlnet. *Expert systems*, *41*. https://doi.org/https://doi.org/10.1111/exsy.13701

Auphonic. (n.d.). Features. https://auphonic.com/features

Babu, P. A., Nagaraju, V. S., & Vallabhuni, R. R. (2021). Speech emotion recognition system with librosa. *10th IEEE International Conference on Communication Systems and Network Technologies CSNT*, 421–424. https://doi.org/https://doi.org/10.1109/CSNT51715.2021.9509714

Baird, A., Tzirakis, P., Brooks, J. A., Gregory, C. B., Schuller, B., Batliner, A., Keltner, D., & Cowen, A. (2022). The acii 2022 affective vocal bursts workshop & competition. *10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos ACIIW*, 1–5. https://doi.org/https://doi.org/10.1109/ACIIW57231.2022.10086002

Bänziger, T., Patel, S., & Scherer, K. R. (2014). The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of nonverbal behavior*, *38*, 31–52. https://doi.org/https://doi.org/10.1007/s10919-013-0165-x

Barbon, R. S., & Akabane, A. T. (2022). Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: A case study. *Sensors (Basel, Switzerland)*, *22*, 8184. https://doi.org/https://doi.org/10.3390/s22218184

Brooks, J. A., Tzirakis, P., Baird, A., Kim, L., Opara, M., Fang, X., Keltner, D., Monroy, M., Corona, R., Metrick, J., & Cowen, A. S. (2023). Deep learning reveals what vocal bursts express in different cultures. *Nature human behaviour*, *7*, 240–250. https://doi.org/https://doi.org/10.1038/s41562-022-01489-2

Bruce, P., & Bruce, A. (2017). *Practical statistics for data scientists*. O'Reilly.

Bryman, A., Bell, E., Reck, J., & Fields, J. (2022). *Social research methods*. Oxford University Press.

Cai, Y., Li, X., & Li, J. (2023). Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. sensors. https://doi.org/https://doi.org/10.3390/s23052455

Cloud, N. (n.d.). Advanced ai platform. https://nlpcloud.com/

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (revised edition)*. Academic Press.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, *3*, 369–382. https://doi.org/https://doi.org/10.1038/s41562-019-0533-6

Creswell, J. W., & Creswell, J. D. (2023). *Research design : Qualitative, quantitative, and mixed methods approaches* (Fifth). SAGE.

Demszky, D., D, M.-A., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computional Linguistics*, 4040–4054. https://doi.org/https://doi.org/10.48550/arxiv.2005.00547

DeSouza, D. D., Robin, J., Gumus, M., & Yeung, A. (2021). Natural language processing as an emerging tool to detect late-life depression. *Frontiers in psychiatry*, *12*, 719125. https://doi.org/https://doi.org/10.3389/fpsyt.2021.719125

Drougkas, G., Bakker, E. M., & Spruit, M. (2024). Multimodal machine learning for language and speech markers identification in mental health. *BMC medical informatics and decision making*, *24*, 320–354. https://doi.org/https://doi.org/10.1186/s12911-024-02772-0

Ekberg, M., Stavrinos, G., Andin, J., Stenfelt, S., & Dahlström, Ö. (2023). Acoustic features distinguishing emotions in swedish speech. *Journal of voice*. https://doi.org/10.1016/j.jvoice.2023.03.010.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on psychological science*, *11*, 31–34. https://doi.org/https://doi.org/10.1177/1745691615596992

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, *3*, 364–370. https://doi.org/https://doi.org/10.1177/1754073911410740

Ermakova, T., Fabian, B., Golimblevskaia, E., & Henke, M. (2023). A comparison of commercial sentiment analysis services. *SN computer science*, *4*, 477–. https://doi.org/https://doi.org/10.1007/s42979-023-01886-y

Esfahani, S. H. N., & Adda, M. (2024). Classical machine learning and large models for text-based emotion recognition. *Procedia Computer Science*, *241*, 77–84. https://doi.org/https://doi.org/10.1016/j.procs.2024.08.013

Frühholz, S., & Belin, P. (2019). *The oxford handbook of voice perception*. Oxford University Press.

HappyPlanetIndex. (n.d.). What is the happy planet index? https://happyplanetindex.org/learn-about-the-happy-planet-index/

Hume, A. (n.d.-a). About hume. https://www.hume.ai/about

Hume, A. (n.d.-b). About the science. https://dev.hume.ai/docs/resources/science

Jadoul, Y., de Boer, B., & Ravignani, A. (2024). Parselmouth for bioacoustics: Automated acoustic analysis in python. *Bioacoustics Berkhamsted*, *33*, 1–19. https://doi.org/https://doi.org/10.1080/09524622.2023.2259327

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of phonetics*, *71*, 1–15. https://doi.org/https://doi.org/10.1016/j.wocn.2018.07.001

Jahangir, R., Teh, Y. W., Mujtaba, G., Alroobaea, R., Shaikh, Z. H., & Ali, I. (2022). Convolutional neural network-based cross-corpus speech emotion recognition with data aug-

mentation and features fusion. *Machine vision and applications*, *33*. https://doi.org/10.1007/s00138-022-01294-x.

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, *42*, 1–40. https://doi.org/https://doi.org/10.1007/s10919-017-0268-x

Kansara, D., Sawant, V., Shekokar, N., Vasudevan, H., Narvekar, M., & Michalas, A. (2020). Comparison of traditional machine learning and deep learning approaches for sentiment analysis. In *Advanced computing technologies and applications* (pp. 365–377). Springer. https://doi.org/https://doi.org/10.1007/978-981-15-3242-9_35

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE access*, *7*, 117327–117345. https://doi.org/https://doi.org/10.1109/ACCESS.2019.2936124

Kumar, S., & Singh, S. (2024). Fine-tuning llama 3 for sentiment analysis: Leveraging aws cloud for enhanced performance. *SN computer science*, *5*, 1161. https://doi.org/https://doi.org/10.1007/s42979-024-03473-1

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2023). A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *The Artificial intelligence review*, *56*, 15129–15215. https://doi.org/https://doi.org/10.1007/s10462-023-10509-0

Kusal, S. D., Patil, S. G., Choudrie, J., & Kotecha, K. V. (2024). Understanding the performance of ai algorithms in text-based emotion detection for conversational agents. *ACM transactions on Asian and low-resource language information processing*, *23*, 1–26. https://doi.org/https://doi.org/10.1145/3643133

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, *13*, 293–303. https://doi.org/https://doi.org/10.1109/TSA.2004.838534

Lee, S. J., Lim, J., Paas, L., & Ahn, H. S. (2023). Transformer transfer learning emotion detection model: Synchronizing socially agreed and self-reported emotions in big data. *Neural computing & applications*, *35*, 10945–10956. https://doi.org/https://doi.org/10.1007/s00521-023-08276-8

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy (Basel, Switzerland)*, *25*, 1440–. https://doi.org/https://doi.org/10.3390/e25101440

Livingstone, S. R., & Russo., F. A. (2019). Ravdess emotional speech audio dataset. https://doi.org/10.34740/KAGGLE/DSV/256618

Madhuri, S., & Lakshmi, V. (2021). Detecting emotion from natural language text using hybrid and nlp pre-trained models. *Turkish journal of computer and mathematics education*, *12*, 4095–4103.

Maruf, A. A., Khanam, F., Haque, M. M., Jiyad, Z. M., Mridha, M. F., & Aung, Z. (2024). Challenges and opportunities of text-based emotion detection: A survey. *IEEE access*, *12*, 18416–18450. https://doi.org/https://doi.org/10.1109/ACCESS.2024.3356357

Milner, R., Jalal, M. A., Ng, R. W. M., & Hain, T. (2019). A cross-corpus study on speech emotion recognition. In *Ieee automatic speech recognition and understanding workshop asru* (pp. 304–311). IEEE. https://doi.org/https://doi.org/10.1109/ASRU46091.2019.9003838

Montasem, A., Brown, S. L., & Harris, R. (2013). Do core self-evaluations and trait emotional intelligence predict subjective well-being in dental students? *Journal of Applied Social Psychology*, *43*, 1097–1103. https://doi.org/10.1111/jasp.12074

Núñez, A. Á., del C Santiago Díaz, M., Vázquez, A. C. Z., Marcial, J. P., & Linares, G. T. R. (2024). Emotion detection using natural language processing. *International Journal of*

Combinatorial Optimization Problems and Informatics, *15*, 108–114. https://doi.org/https://doi.org/10.61467/2007.1558.2024.v15i5.564

Oatley, K., Keltner, D., & Jenkins, J. M. (2019). *Understanding emotions fourth edition*. Blackwell.

of Surrey, U. (n.d.). Surrey audio-visual expressed emotion (savee). https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

OpenAI. (2022, September). Introducing whisper. https://openai.com/index/whisper/

Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2023). Multi-cultural speech emotion recognition using language and speaker cues. *Biomedical signal processing and control*, *83*, 104679–. https://doi.org/https://doi.org/10.1016/j.bspc.2023.104679

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (tess). https://doi.org/https://doi.org/10.5683/SP2/E8H2MF

Praseetha, V. M., & Joby, P. P. (2022). Speech emotion recognition using data augmentation. *International journal of speech technology*, *25*, 783–792. https://doi.org/https://doi.org/10.1007/s10772-021-09883-3

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *Robust Speech Recognition via Large-Scale Weak Supervision*. https://doi.org/https://doi.org/10.48550/arxiv.2212.04356

Rahman, M. M., Hossain, M. A., Hasan, T., Ahmed, M. K., Sultana, R., & Islam, M. S. (2024). Emotionnet: Pioneering deep learning fusion for real-time speech emotion recognition with convolutional neural networks. *2024 6th International Conference on Electrical Engineering and Information & Communication Technology ICEEICT*, 592–597. https://doi.org/https://doi.org/10.1109/ICEEICT62016.2024.10534404

Rathi, T., & Tripathy, M. (2024). Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review. *Speech communication*, *162*, 103102–. https://doi.org/https://doi.org/10.1016/j.specom.2024.103102

Repede, S. E., & Brad, R. (2024). Llama 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers (Basel)*, *13*, 292. https://doi.org/https://doi.org/10.3390/computers13110292

Ri, F. A. D., Ciardi, F. C., & Conci, N. (2023). Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. *IEEE Access*, *11*, 1. https://doi.org/https://doi.org/10.1109/ACCESS.2023.3326071

Safari, F., & Chalechale, A. (2023). Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope. *The Artificial intelligence review*, *56*, 3273–3297. https://doi.org/https://doi.org/10.1007/s10462-023-10603-3

Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using deep learning techniques: A comprehensive review. *International journal of multimedia information retrieval*, *12*, 41–. https://doi.org/https://doi.org/10.1007/s13735-023-00308-2

Scherer, K. R., Frühholz, S., & Belin, P. (2018). *Acoustic patterning of emotion vocalizations*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.4

Shelke, N., Chaudhury, S., Chakrabarti, S., Bangare, S. L., Yogapriya, G., & Pandey, P. (2022). An efficient way of text-based emotion analysis from social media using lra-dnn. *Neuroscience informatics*, *2*, 100048. https://doi.org/https://doi.org/10.1016/j.neuri.2022.100048

Siedlecka, E., & Denson, T. F. (2019). Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, *11*, 87–97. https://doi.org/https://doi.org/10.1177/1754073917749016

Simcock, G., McLoughlin, L. T., Regt, T. D., Broadhouse, K. M., Beaudequin, D., Lagopoulos, J., & Hermens, D. F. (2020). Associations between facial emotion recognition and mental

health in early adolescence. *International Journal of Environmental Research and Public Health*, *17*, 330. https://doi.org/10.3390/ijerph17010330

Singh, S. (2023). Emotion recognition for mental health prediction using ai techniques: An overview. *International Journal of Advanced Research in Computer Science*, *14*, 87–107. https://doi.org/10.26483/ijarcs.v14i3.6975

Sönmez, Y. Ü., & Varol, A. (2024). In-depth investigation of speech emotion recognition studies from past to present. the importance of emotion recognition from speech signal for ai. *Intelligent systems with applications*, 200351–. https://doi.org/https://doi.org/10.1016/j.iswa.2024.200351

Thaler, F., Haug, M., Gewald, H., Brune, P., Pennarola, F., Pallud, J., & Braccini, A. M. (2024). The context sets the tone: A literature review on emotion recognition from speech using ai. In *Technologies for digital transformation* (pp. 129–143, Vol. 64). Springer Nature Switzerland. https://doi.org/https://doi.org/10.1007/978-3-031-52120-1_8

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, *4*, 46–64. https://doi.org/https://doi.org/10.1037/1528-3542.4.1.46

Tian, L., Oviatt, S., Muszyński, M., Chamberlain, B. C., Healey, J., & Sano, A. (2022). *Applied affective computing*. Association for Computing Machinery.

Tomasello, R., Grisoni, L., Boux, I., Sammler, D., & Pulvermüller, F. (2022). Instantaneous neural processing of communicative functions conveyed by speech prosody. *Cerebral cortex*, *32*, 4885–4901. https://doi.org/https://doi.org/10.1093/cercor/bhab522

TwinWord. (n.d.). Just the best keywords. https://www.twinword.com/ideas/

Tyagi, S., & Szénási, S. (2024). Semantic speech analysis using machine learning and deep learning techniques: A comprehensive review. *Multimedia tools and applications*, *83*, 73427–73456. https://doi.org/https://doi.org/10.1007/s11042-023-17769-6

Zhang, J., & M, Z. (2024). Is llama 3 good at identifying emotion? a comprehensive study. *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence MLMI*, 128–132. https://doi.org/https://doi.org/10.1145/3696271.3696292

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Data mining and knowledge discovery*, *8*, 1253–n/a. https://doi.org/https://doi.org/10.1002/widm.1253

Zhang, S., Tao, X., Chuang, Y., & Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech communication*, *127*, 73–81. https://doi.org/https://doi.org/10.1016/j.specom.2020.12.009

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, *47*, 312–323. https://doi.org/https://doi.org/10.1016/j.bspc.2018.08.035