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: May 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: 2025-05-20

Abstract

This thesis investigates emotion recognition in Swedish speech through a multimodal approach using AI models. Combining speech-based and text-based analysis with self-assessed emotion scores from participants in semi-structured interviews, this study addresses three research questions: (1) How does AI-model for speech emotion recognition compare to research on vocal markers for emotions in Swedish speech?; (2) What similarities and differences emerge between emotions detected from audio features and from the textual transcripts of the same speech data?; (3) How do AI-generated emotion labels (speech & text-based) compare to self-reported emotions? To answer these questions, data was collected in form of spontaneous speech from interviews, resulting in a more naturalistic dataset than acted datasets which are largely used in research. The results from analysing the collected data revealed partial alignments between vocal features and the speech-based AI model, Hume AI, as well as strongly suggesting some emotions are more difficult to detect due. The text-based AI model, NLP Cloud, proved to better align with the self-assessed scores, indicating that the textual context gave important cues more consistently than vocal features alone. The results highlighted the importance of a multimodal approach to capture a wider range of emotional expressions. Contributing to the fields of affective computing and natural language processing particularly by using spontaneous speech over an acted dataset, this study gives a deeper understanding in emotion recognition applied to the Swedish language.

Keywords: Emotion recognition, text-based emotion detection (TBED), speech-based emotion recognition (SER), vocal markers, Swedish speech, self-assessed emotion scores, AI-based emotion detection

Acknowledgement

"The best way to predict the future is to create it."

— Alan Kay

"If we knew what it was we were doing, ..."

— Albert Einstein

"Will we go down? As far as we know, How do we know?"

— Anna Von Hausswolff

The following study is a bachelor thesis in Software Engineering at Jönköping University. As not uncommon for the female nature, behavioral patterns interests us, beyond doubt in combination with technology potentially holdning capacity to change not only society, but how we think. For the latter, it should be acknowledged, I surely hope not.

Thanks to Knowit, Jönköping, for supporting our project even if interdisciplinary to the general software development area. Thanks to our beloved computers for bringing some light into the dark night, consuming our eyes and sometimes let us forget when our brain fight. Thanks to all participants in this study, sharing your thoughts and a fragment of your minds, after all, it is what us all binds. Thanks to my thesis partner, and dear friend, Janna, you are worth more than a dime. Glad you read my rhyme, purhaps some joy is found when reading what is following, at least part-time.

Jönköping, May 2025

Sara Ljung

Contents

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

According to Oatley et al. (2019), 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 express them. Intonation 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 autociteOatley2019. Beyond spoken signals, researchers have also developed a set of Natural Language Processing (NLP) techniques to interfer emotional states and opinions directly from text, based on methods at the intersection of 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 raised. (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 NLP.

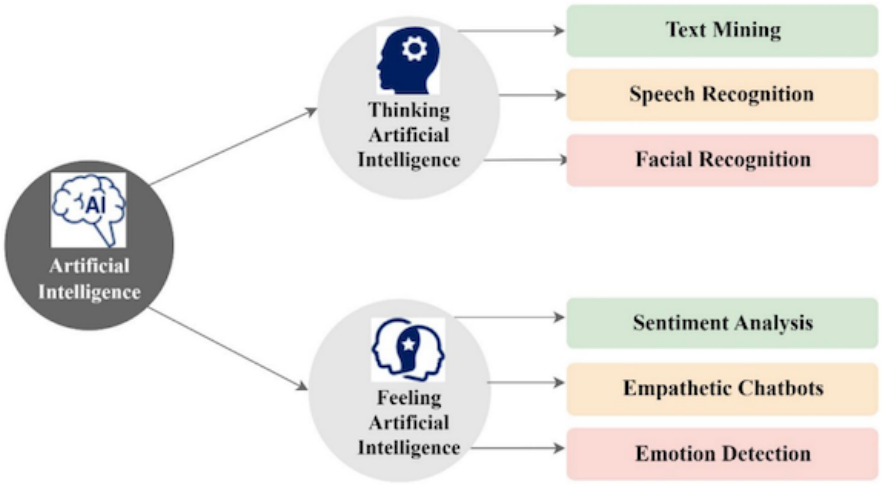


Figure 1.1: *Subdomains of NLP (S. Kusal et al., 2023).*

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 analysis of 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 detection is used with sentiment analysis (Madhuri & Lakshmi, 2021). With the evolution of deep learning networks, emotion detection has advanced (Safari & Chalechale, 2023). 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. The emotions categorization fluctuate depending on the research. These basic six categories were 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 there was no agreement about which 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 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 (76%) 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 (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. Those features offer invaluable insights into the subtle emotional expressions conveyed through speech, assisting the complicated process of emotion recognition (Lian et al., 2023). Several studies distinguish different emotions through vocal features. Already in 2005, automatic recognition of positive and negative emotions in spoken dialogs was investigated (C. M. Lee & Narayanan, 2005). In that study, acoustic, lexical, and discourse information were combined to enhance emotion detection and move beyond traditional acoustic-only ways. The authors analysed acoustic features, lexical features, and discourse features. Linear Discriminant Classifiers were used and resulted in good performance for acoustic and lexical information. A study by Bänziger et al. (2014) demonstrated that human listeners could reliably rate emotional expressions in acted voices. These human judgements had higher accuracy for detection of certain emotions, such as happiness, compared to technical analyses of acoustic measurements. According go Khalil et al. (2019), acoustic features enable emotion recognition through speech using deep learning, which offers many advantages over traditional sentiment-analysis methods. Deep learning models has the capability to automatically detect complex patterns and varying features without requiring manual feature extraction. The goal of speech emotion recognition (SER) is identification of emotions in speech, unrelated to the semantic content (S. D. Kusal et al., 2024). Figure 1.2 represents a SER system.

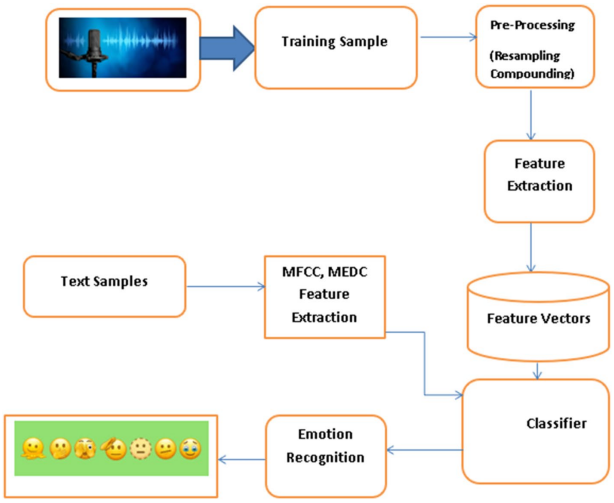


Figure 1.2: *Block diagram of SER (Tyagi & Szénási, 2024).*

In recent years, speech emotion recognition has emerged as a research area driven by its applications in human-computer interaction (S. Zhang et al., 2021). The advancements has led to the development of intelligent affective services in fields such as call centres, healthcare, surveillance, and affective computing. The accuracy of models tested in recent years have improved significantly (Adebiyi et al., 2024; Praseetha & Joby, 2022; Rahman et al., 2024). Several studies conducted in the last year's show emotion detection accuracy results over 90%. Juslin et al. (2018) concluded a study in 2018 analysing 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 underscore that posed and spontaneous emotions are not interchangeable and that SER datasets must distinguish between these sample types to build models that can generalize to real-world emotional speech accurately. One recent review of SER corpora and features (Rathi & Tripathy, 2024) shows that most studies still target only six emotions—happiness, anger, sadness, surprise, fear, and neutrality—even though narrower sets (e.g. anger, fear, happiness, sadness) dominate earlier work (K. R. Scherer et al., 2018). In contrast, GoEmotions is a large, detailed text database for 27 distinct emotions, a study by Demszky et al. (2020) obtained an average F of 0.46 (0.86 for gratitude, 0.00 for grief) and 0.64 when reduced to six labels. While this GoEmotions-study is included in the research behind the commercial system Hume.ai, and highlights the value of fine-grained emotion categories (Hume, n.d.-a), which is important to acknowledge since because of potential biases.

Datasets drive both speech- and text-based emotion models. Speech emotion recognition datasets are gathered in three ways, acted by performers, induced in controlled settings, or from natural conversations, affecting how expressive and realistic the recordings are. Rathi and Tripathy (2024) analysed 93 research papers where IEMOCAP and RAVDESS are among the most widely used datasets, chosen by 35.83% and 21.50% of researchers, respectively. They further state that dataset choice, recording conditions, and selected features (e.g. MFCCs, pitch, intensity, prosody) impact SER accuracy significantly, and that natural speech is more difficult to classify due to its high variability and background noise. The number of natural datasets is relatively limited (Cai et al., 2023), and many research papers test on acted datasets. For example, the empirical analysis Ahammed et al. (2024) demonstrates a high-accuracy SER system (100% accuracy, precision, and F1-score) on a combined RAVDESS, TESS, and SAVEE dataset. Each dataset includes posed or elicited emotions in English speech. Similarly, different models for SER achived over 94% accuracy for these same acted datasets (Alroobaea, 2024). However, spontaneos speech is not validated in these studies and depends on acted data. In contrast, Text-Based Emotion Detection (TBED) are driven by diverse text datasets, from six emotions to GoEmotions set with 27 emotions (S. Kusal et al., 2023). Researchers in TBED can use publicly available datasets with reliable annonating, for instanse derieved from stories, publications, news, social media texts, or reviews on movies. According to S. Kusal et al. (2023), 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. Over 3.5 milliom self-labeled posts on Twitter was used to train a model in S. J. Lee et al. (2023) , achieving up to 0.87 F1 on human-annonated sets and 0.79 F1 on self-reported hashtags. However, like SER, TBED is dominated by English and lacks large, natrualistic datasets in other languages.

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 (Abbaschian et al., 2021). It can be implemented in vehicles' safety structures to recognize the driver's emotional state and therefore prevent accidents. Several studies (DeSouza et al., 2021; Drougkas et al., 2024; Simcock et al., 2020; Singh, 2023) demonstrate the potential benefit of AI-based emotion recognition in mental health, investigating it can assist psychiatrics diagnosing and identifying potential mental illnesses. DeSouza et al. (2021) showed how leveraging speech and text analysis with NLP can help detect late-life depression and predict its severity with 86-92% accuracy. Drougkas et al. (2024) compared unimodal approaches, either text- or audio-based and combined audio-text models, resulting in text unimodal accuracy between 78% and 87% with F1 scores from 0.60 to 0.79, audio unimodal accuracy of 64%-72% with F1 values as low as 0.0 up to 0.46. Multimodal approaches, combining text and audio, showed similar accuracy (80% - 87%) and F1 scores (0.60-0.80) as text-unimodal approaches. The authors conclude that text model outperform the acoustic model in recognising mental health indicators, but that multimodal models can outperform unimodal techniques since positive F1 scores increase combining the models.

In summery, speech emotion recognition is proficient in capturing vocal cues, especially on acted datasets, while text-based emotion detection relies on transformer models trained on large text collections. However, most research is based on English and uses acted or social-media data, with few studies exploring natural, spontaneous speech or on other languages.

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 (K. R. 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 1.3 demonstrates the difficulty level for varying settings.

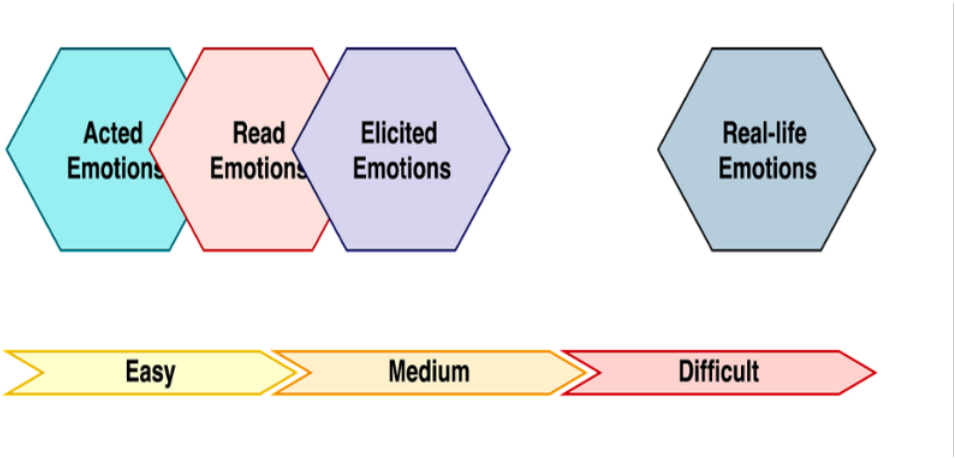


Figure 1.3: *Emotion recognition databases and their difficulty level (Khalil et al., 2019).*

Text-Based Emotion Detection (TBED) shows similar gaps regarding the data the majority of the researched models are trained and evaluated on. Although transformer models reach up to 76 % accuracy on English datasets (S. Kusal et al., 2023), they are heavily dependent on informal social-media or review texts. Moreover, TBED resources for other languages is limited, and challenges like sarcasm and cultural nuance affects the reliability of the models (S. J. Lee et al., 2023; Maruf et al., 2024).

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. Limitations as overreliance on acted English data, lack of natrual and non-English datasets, and modality-specific biases, are motivations for this thesis. We will compare SER and TBED in Swedish, using both speech and text from the same speakers and explore the alignment against their self-reported emotions to test real-world performance.

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 the dataset gap by investigating the performance of AI-driven emotion recognition systems in a specific 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. Although emotions are inherently difficult to define and can be challenging for individuals to self-assess, it is valuable to examine the alignment between model outputs and people's own perceieved emotions. Publicly available AI models and APIs, despite their use in real-world applications, are rarely compared agaisnt such subjective human data, making this comparative evaluation both novel and scientifically signficiant. The purpose of this thesis is therefore to explore how the Llama-3 model from NLP Cloud and Hume AI 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). 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 an 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] What similarities and differences emerge between emotions detected from audio features and from the textual transcripts of the same speech data?*

The perception of emotions is a complex field, with few studies made on the alignment between machine-labeled emotions and human-perceived 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

The scope 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 analyses emotions from about 15 Swedish-speaking participants through short interviews designed to evoke natural emotions. Particpants, both male and female, have varying age from 20-78 years. The study includes:

- **Vocal Extraction:** With Praat Parselmouth, a Python library for Praat software used for feature extraction from audio recordings (Jadoul et al., 2018).
- **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.
- **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 semantic orientations, one positive and one negative designed interview for each participant. Five emotions are derieved from the audio and are reported by the participants.

The analysis relies on existing AI tools and the API's Hume and NLP Cloud, as well as Praat 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. With only 15 participants, the dataset is limited, and the results may not apply to all Swedish speakers as well as the findings may not apply beyond Swedish. The interviews are designed to elicit emotions and may not fully capture natural emotional responses, since they are partially induced, and the very nature of the interview setting cannot be directly applicable to real-world environments. The design of the emotion-eliciting scenarios may not be optimal because of deficient psychological expertise, even if the scenarios are based on prior research. By the same reason, the composition by the self-reports could be a limitation in combination with the subjectivity of participants' emotion reports, that may be influenced by personal biases or recall inaccuracies. The selected emotion

categories, commonly used in prior research, include more negative (anger, fear, sadness) than positive (joy) oriented emotions, leading to potential limitations on self-reports. Focusing on these emotions may exclude other relevant emotional states, as Hume AI's output in fact cover several more emotion labels. Pre-trained AI models are utilized without modifying their algorithms, which may introduce biases. For vocal extraction, Praat Parselmouth is applied, which in our implementation, does not cover the full set of vocal features included in the Swedish research (Ekberg et al., 2023) used for comparison in RQ1.

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:

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.

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.

Results: Presents the collected data and analyses of the research questions presented with statistical analyses, supported by visualisations.

Discussion: Discussion for each research question, how the results compare with prior studies and the methodology impact on the results.

Conclusion: Overall conclusion of the study, with key findings, implications and future research recommendations.

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 analysed using a Python application. The elements of the Python application for analysing the data from the interviews will be comprehensively explained in this chapter.

2.1 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 behaviours and communication, the domain of affective computing is a field that merges the principles of artificial intelligence with insights from social and behavioural science (Tian et al., 2022).

2.2 Natural Language Processing and Emotion Recognition

The first English language lexical database was created in 1998 for Natural Language Processing (NLP) tasks, the term sentiment and emotional analysis came to practice in 2001 to predict the stock market, and in 2005 the first article was written on emotion and opinion detection from text (S. Kusal et al., 2023). 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 (Kansara et al., 2020). 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 preprocessing 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).

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).

2.3 Speech-Based Emotion Recognition

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). A typical SER system contains of three components: signal preprocessing, feature extraction, and classification (Sahoo et al., 2023).

Although there is a wide range of SER-algorithms, with 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 2.1, involving several steps as feature extraction, selection and classification.

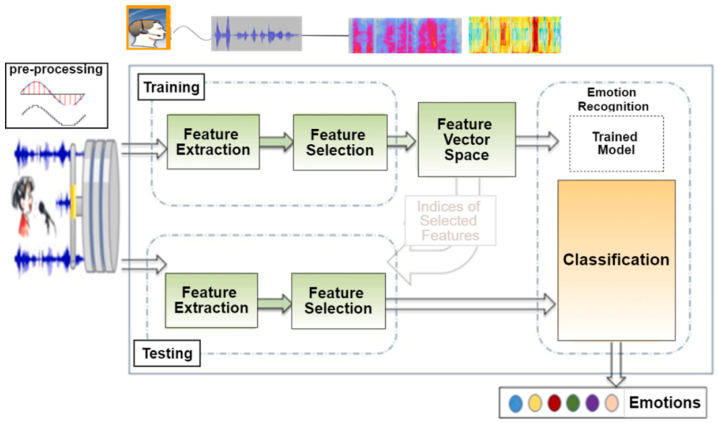


Figure 2.1: *An overview of the stages in SER to analyse speech data for emotion detection* (Sönmez & Varol, 2024)

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 and Varol (2024) there are several key benefits with SER, such as a limited amount of hardware needed for the capturing the vocal data which simplifies the process of the vocal data collection. Another benefit is that vocal data being less demanding in terms of storage, compared to video footage for example, and 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.

2.3.1 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 2.2, 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.

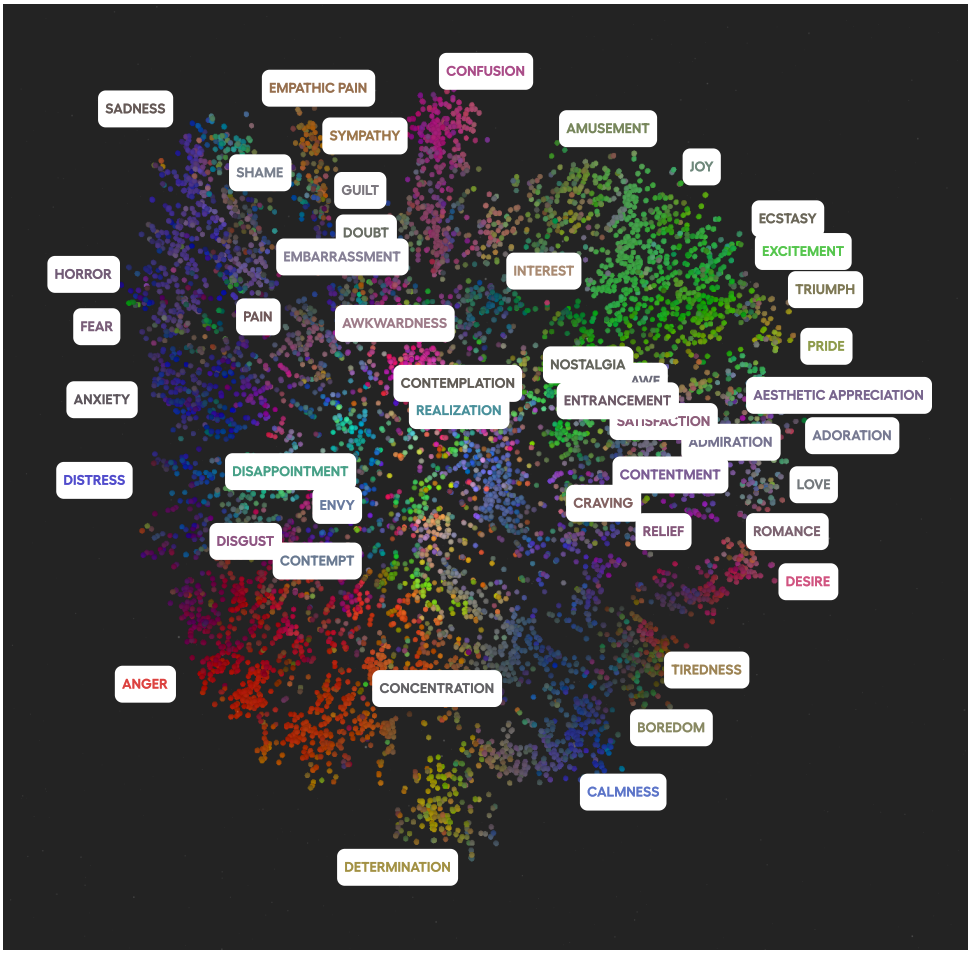


Figure 2.2: *Visual representation of Hume AI's speech prosody model* (AI, n.d.-a)

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 have received less attention 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 2.3 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.

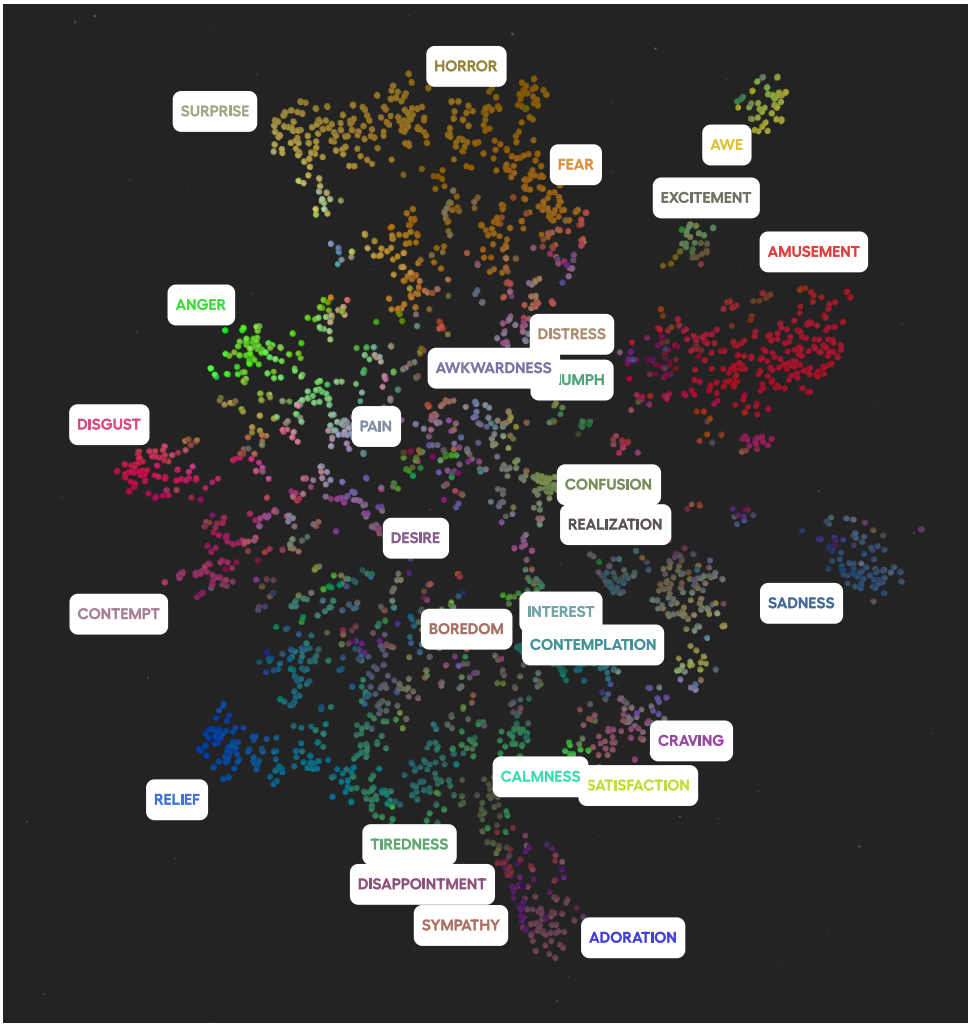


Figure 2.3: *Visual representation of Hume AI's vocal burst model* (AI, n.d.-b)

While there are other tools for emotion recognition, Hume AI is 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 there are other models, like OpenAI whisper, which is an extensively trained model on hundreds of thousands of hours on data, their main focus leans toward transcribing speech (OpenAI, 2022). This ultimately led to the decision to use Hume AI.

2.3.2 Praat Parselmouth

In the field of software for linguistic analysis, Praat is a well-established tool to analyse 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 analysing 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 analysing 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, there some main objectives in Parselmouth according to Jadoul et al. (2018). One objective is to 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, is also an important aspect, as well as 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 library following the principles of Python and 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 analysis still falls on Praat Parselmouth due to it being more fitting for the purpose of this thesis.

2.4 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 systems rely strongly on lexical cues, and research shows different types of words carries different levels of intensity (Chauhan et al., 2024). This highlights both a strength and a limitation of text-based emotion recognition, as emotional sentiment may go undetected unless it is verbalized, since emotions may not necessarily be expressed through text (Soleymani et al., 2017).

Text-based emotion detection relies on four main approaches, according to S. Kusal et al. (2023). The first approach is keyword-based, which matches words in a text with predefined emotion keywords from resources like WordNet, adjusting for intensity and negation. The second approach is rule-based, and this approach uses linguistic rules and probabilistic affinity to classify emotions after preprocessing. The third is machine learning-based which applies supervised or unsupervised models to classify emotions, extracting key features from preprocessed text. The fourth and last approach is deep learning-based and it 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 labelled 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).

2.4.1 NLP Cloud

There are limited publicly available APIs for text-based emotion detection (TBED). The decision to use NLP Cloud for this study consists of several reasons as following. NLP Cloud was selected for this study because of its extensive model offering, multilingual support, transparent documentation, and available support for data privacy and security (Cloud, n.d.). Compared to other alternatives such as Vern AI (AI, u.d.) or TwinWord (TwinWord, n.d.), both either lacked comprehensive documentation and information about the models the API used, NLP Cloud had technical transparency and many different API endpoints which was most compatible with this research. Additionally, NLP Cloud also provides speech to text transcription, 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 robustness and generalization (Radford et al., 2022). 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.

For text-based emotion recognition, NLP Cloud offers different models, including DistilBERT Base Uncased Emotion and Llama 3. DistilBERT is a compressed version of the original BERT (Bidirectional Encoder Representations from Transformers), developed to reduce model size and gain faster results while remaining the language understanding capacity. (Qazi et al., 2025). DistilBERT has demonstrated high accuracy, from 95.7% to 96.6% on benchmark datasets (Areshey & Mathkour, 2024). However, the model does not natively support Swedish and using it for this study would require translation of the transcripts, potentially leading to translation bias or loss of linguistic nuance.

Regarding multiple language performance, a study examined Llama 3 vs. State-of-the-Art Large Language Models on their ability to detect fake news (Repede & Brad, 2024). Two datasets were tested, one Romanian and one English. Their proposed Llama 3 model accomplished higher precision and accuracy across several metrics in fake news 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, which is noteworthy considering both Romanian and Swedish are underrepresented languages in fields of LLM's and NLP relative to English for example. The study by Repede and Brad (2024) 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.

Comparing these two alternatives for text-based emotion detection in Swedish, the fine-tuned Llama 3 model shows promise 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.

In the beginning of the data analysis, the sentiment analysis endpoint from NLP Cloud with the fine-tuned Llama 3-70b model was considered and applied for textual emotion detection. However, the model returns a wide range of emotion classes without sufficient control over which emotions should be included in the output. In this study, a set of five emotions – anger, joy, sadness, fear, and surprise – were used to align with previous research on vocal markers in emotion expressions explained in 2.5 Vocal Markers. Therefore, the decision to utilize NLP Cloud's text generation endpoint instead was made, using their fine-tuned Llama 3-70b model but with prompting instead of direct emotion classification output (NLP Cloud, 2025). This model operates as an instruction-following generative language model with ability to respond to natural language prompts and generate emotions classifications based on how the request is phrased. This provided more flexibility and allowed control of which emotions should be included in the output as well as the format by prompting with instructions.

For the aim of this study where emotion detection from Swedish interview transcripts is explored, the fine-tuned Llama 3 model through NLP Cloud's text generation endpoint was determined to be the most suitable approach due to its language support, flexibility through prompting, and ability to control the emotion categories included in the output.

2.5 Vocal Markers

Vocal features have a significant role in distinguishing emotion through speech. It has been demonstrated that a listener accurately can recognize different emotions based on vocal cues, suggesting that emotional vocal expression has different patterns (Banse & Scherer, 1996). Acoustic variables that are involved in signalling emotions vocally include the fundamental frequency (pitch), vocal energy (intensity, loudness), the location of frequency formants (F1, F2, F3) which is associated with how articulation is perceived, and speech tempo. High arousal states include increased pitch and intensity, are related to positive emotion states as happiness/joy. These features indicate the same in Swedish (Ekberg et al., 2023), where acoustic markers for five different emotions (anger, happiness, fear, sadness, surprise) were studied on fourteen acted sentences, each sentence articulated expressed with each emotion. Happiness presented highest fundamental frequency, increased loudness (intensity) and second highest harmony-noise-ratio (HNR). These results describe the largest pitch SD (6.25) for happiness, implying acoustic variability. Similar patterns are presented in a review on 108 studies (Kamiloğlu et al., 2020), including twenty-six that researched acoustic features on positive emotions. Pitch, loudness, and formant features revealed the strongest indicators of positive emotions when compared with neutral vocalizations. The review suggests candidates as HNR indicating happiness as well but explain that clear conclusions cannot be composed due to limited empirical evidence. Anger shares increased pitch and vocal energy with joy (Banse & Scherer, 1996). The Swedish study indicates this pattern as well, although anger is not captured by frequency features separately but by amplitude cues as HNR and intensity (Ekberg et al., 2023). HNR exhibited lower values than for happiness and fear, but higher than sadness and surprise. Sadness is predicted by lower intensity and pitch with slower speech rate, aligning with results on the Swedish language (Ekberg et al., 2023; K. Scherer, 2003). Vocalised fear is associated with panic in (Banse & Scherer, 1996), where pitch and frequency formants are heightened with increased rate of articulation. The Swedish report does not describe fear aligned with panic, although the opposite is not mentioned. The results showed second highest pitch, third highest loudness, and top HNR value. Frequency related jitter distinguished fear with the lowest value of the emotions. Jitter alongside amplitude associated shimmer characterised surprise in the Swedish results, with lower loudness and pitch than anger and happiness. Surprise is not included with listed values in the other studies referred to in this subsection.

The results from the Swedish study (Ekberg et al., 2023) on emotional acoustic features are presented in Figure 2.4 with mean and standard deviation values for the five emotions. Other studies mentioned (Banse & Scherer, 1996; Kamiloğlu et al., 2020; K. Scherer, 2003) and additionally Figure 2.5 (Khalil et al., 2019) demonstrate similar vocal expression patterns, but are mainly studied on English. Therefore, the results from the research on Swedish speech will serve as our reference when categorising emotions based on vocal markers in this study, 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 2.4: *Table comparing acoustic parameters between emotions* (Ekberg et al., 2023)

The mean loudness divergencies between different emotions is shown in table 2.4 come with a very small standard deviation (e.g. $\sigma = 0.66$ for anger mean = 7.16, happiness mean = 6.49). Contrasting, pitch differentiates between 2-6 σ. This suggests that loudness varies far less reliably than pitch across anger, happiness and fear. According to Banse and Scherer (1996) is the fundamental frequency (pitch) the most studied and perceptually prominent feature.

Happiness is one of the emotions that can be identified from pitch, yet both Banse and Scherer (1996) and Ekberg et al. (2023) shows this positive emotion a wide overall acoustic spread. Figure 2.5 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). Additionally, Figure 2.5 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 2.5: *Acoustic variations in different emotions* (Khalil et al., 2019).

These findings will serve as a reference point for the comparison between vocal markers and speech-based emotion recognition, not only for the categorisation function but also for exploring patterns between extracted vocal features and previous research.

2.6 The Experiment

An experiment was be conducted and will consist of short voluntary interviews. These interviews were recorded for data collection and used to extract emotions with a Python application and analysed with a speech-based and text-based AI model.

2.6.1 Python Application

The Python application serves as the central system for processing and analysing emotions in speech and will integrate several tools and frameworks to extract emotions. The overall structure is illustrated in Figure 2.6.

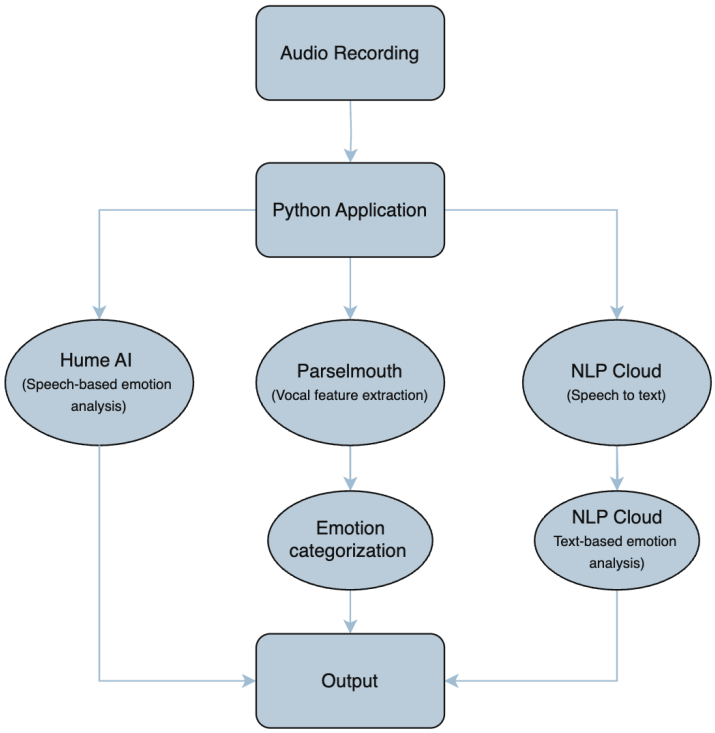


Figure 2.6: Overview of the Python application structure.

Recorded audio is processed through the application, including feature extraction with Praat Parselmouth, and manual categorisation of emotions based on these features, explained further in 3.3.1 Method. The audio is processed through both Hume AI and NLP Cloud, where prosody and vocal bursts are analysed with Hume to detect emotional cues from pitch, intonation and vocal bursts. NLP Cloud is used to transcribe the audio to text and analyse the textual content with a text classifier model fine-tuned for sentiment analysis, described in 2.4.1 NLP Cloud. The application includes statistical analysis for an simple interpretation of different measurements and output correlations, specified in 2.7 Statistical Analysis.

2.6.2 Interviews and Surveys

The semi-structured 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. As the format of the interviews are semi-structured and involve spontaneous speech unlike acted datasets, a well-known challenge emerges. Spontaneous vocal data tends to involve more neutral expressions, and research have shown some emotional classifiers to have a lower accuracy in detecting neutral statements (Cao et al., 2015). This is also supported by other research, stating that acted speech shows higher levels of intensity (Chakraborty et al., 2016). While spontaneous speech presents some documented challenges, it also reflects real word conditions. 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, known as “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 and align with the technical focus of the thesis. 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, joy, and one on the negative spectrum, anger.

2.7 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).

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.

3.1 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.

3.2 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.

16 Swedish speakers, primarily acquaintances to the researchers, are recruited via invitations. Interviews include 2 scenarios, each scenario includes 5-7 questions designed to elicit either positive or negative emotions, the participants select one of these questions for each semantic area, to maintain freedom in the speech as well as avoiding asking questions the participant are not comfortable to answer. The participants are not pre-informed about the feeling that are aimed to be provoked during the interviews. The scenarios have been pilot-tested for effectiveness. The interviews are recorded in a quiet room. Each scenario last between 2 and 4 minutes with breaks between them, the order of the scenarios varies to minimize affecting the

results because of the possibility of the order influencing the different emotions. Each recording have been edited to delete our questions and any longer silent pauses.

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 regarding their everyday lives or society, while the questions related to happiness explore moments of pride and unexpected joy through memories. The semi-structured format allows for follow-up questions based on participant responses, to bring out as much emotion as possible. The follow-up questions include "How did that make you feel?", "Can you evaluate on that specific situation?", and "What feelings did you experience?". See Appendix 7.2 for the full list of interviews.

Audio is recorded and pre-processed to reduce background noise and normalise volume. Vocal markers from each recording are extracted using Parselmouth, a Python library for vocal feature extraction, to answer RQ1. The audio is analysed simultaneously using two emotion recognition models: Hume.ai to extract speech-based emotional labels, and NLP Cloud to transcribe the speech and then analyse the textual content in terms of emotion scores. The same dataset is used for all research questions to ensure consistency.

The diagram in Figure 3.1 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 features extracting utilizing Praat Parselmouth are chosen are based on previous research, see 2.5, 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.

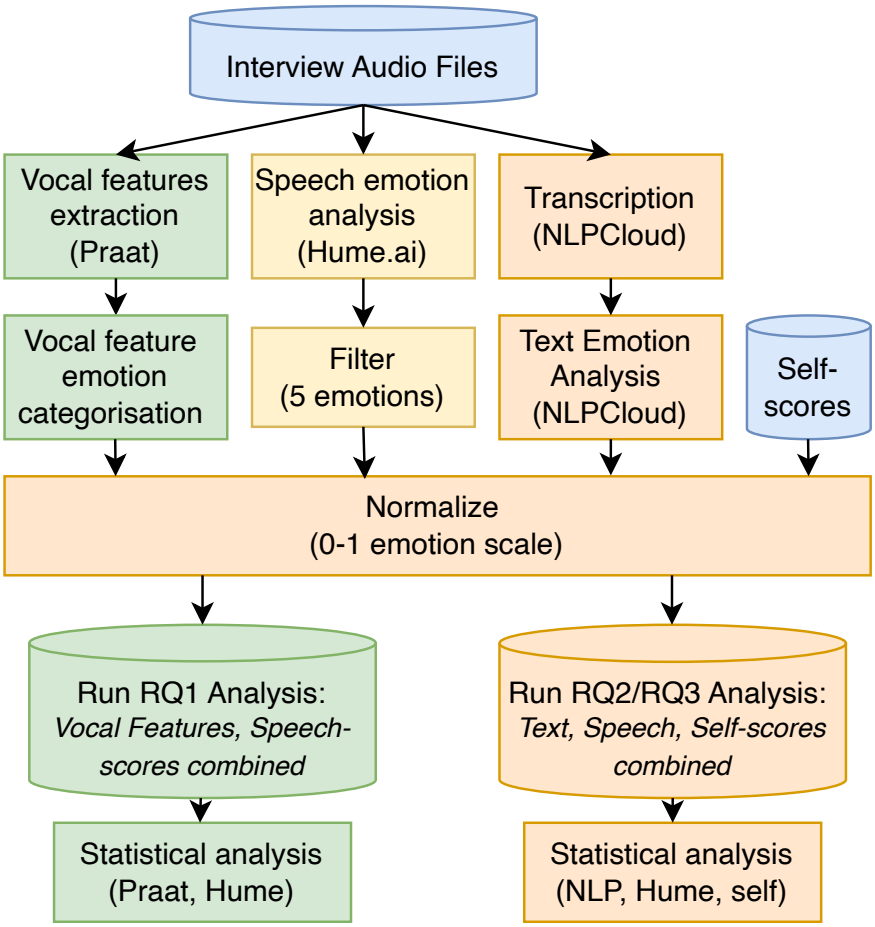


Figure 3.1: The multi-modal pipeline used in this study.

3.2.1 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).

3.2.2 Research Question 2

What similarities and differences emerge between emotions detected from audio features and from the textual transcripts of the same speech data?

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.

3.2.3 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.

3.3 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.

Visualised analysis is included as a complement to tables with data, for easy interpertation of how the different methods correlate and observe behavioral differences, or where the models agree. This includes correlation heatmaps, confusion matrices of top emotion labels, scatter plots illustrating two methods mean emotion probability, and bar charts. Individual clips are presented with bar charts and diagrams representing changes over time segments.

3.3.1 RQ1: Emotion Categorization from Vocal Markers

We analysed each of the 30 clips in several stages to answer the first research question. Extracted vocal features included mean pitch (st, Hz), mean intensity(dB), mean hnr (dB), F1, F2, F3 formants (Hz), jitter and shimmer see Theoretical Framework 2.5. First, acoustic features were categorized in five emotion groups based on standardized distances based on Table 2.4. Because of the method yielded near uniform emotion scoring (0.20), a rule-based function was developed and refined through four variants (V0-V3), where threshold adjustions and anchors was applied to each layer.

Standardized distance function

- 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.

Baseline for neutral speech was not described in the reffered research, therefore the baseline for this study is the average vocal features of all clips. Analysis of our data yielded every emotion probability around 0.18-0.22 for each clip, appearing creating arbitrary scores. This motivated the rule-based approach.

Rule-based method: V0 (Global K)

To mark outer values, 90% confidence interval were used to mark the outer population (Bruce & Bruce, 2017) yielding the critical value Z = ± 1.645 σ. Our threshold K_EXTREME was therefore initialized to 1.6 σ for all emotions. For the normal distribution range we used roughly 1.5 standard deviation from mean, as K_NEAR = 1.25 σ. Benchmarks from Ekberg et al. (2023) were applied:

- ”anger”: [(”hnr”,”below”), (”jit”,”below”), (”loud”, ”above”)]
- ”joy”: [(”pitch”,”above”), (”hnr”,”above”), (”loud”,”above”)]
- ”sadness”: [(”pitch”,”below”), (”hnr”,”below”), (”loud”, ”below”)]
- ”fear”: [(”hnr”,”above”), (”jit”,”below”)]
- ”surprise”: [(”jit”,”above”), (”shim”,”above”)]

If any emotion goes beyond the extreme threshold, the emotion is scored.

For all eight features, if the clip value is within the K_NEAR, 1.25 σ of an emotion's mean, one more point is added, to ensure typical matches are added even if no extreme cue is found.

After the functions loop, each emotion has a score and the highest score is assigned the clip's main emotion label.

V0 resulted in mainly anger and surprise scores.

Rule-based method: V1 (Per-emotion K)

K_EXTREME values were adjusted for all emotions. Justifications is described in Theoretical Framework 2.5 explaning wide spread acoustic pattern for happiness and similar vocal palette as anger. Therefore, joy was set to a smaller extreme value while heightened for anger. Joy = 0.5 σ Anger = 1.6 σ Others = 1.0 σ

V1 showed the same results as V0 with no distinct difference.

Rule-based method: V2 (+ Feature Weight)

V1 was extended with utilizing the standardised function 3.3.1 as fallback before returning the loaded dictionary with emotion scores. The returned values from the standardised function was multiplied with 0.25 and added to the rule-based versions result.

Initially, all vocal features were equally weighted as 1.0. For this version, loudness were decreased to 0.5 motivated by prior research showing that loudness is the least discriminative of the vocal features extracted for this study as explained in 2.5.

The approach yield more dispered emotion scores but still rated anger as top-label for more than 50% of the recordings.

Rule-based method: V3 (+ Benchmarks)

Benchmarks cues for anger, joy, and sadness was implemented to avoid overlabeling. To yield a score, anger are required to have at least two benchmark cues, ensuring at least two of the emotion-specific benchmarks (see V0) is fulfilled. Joy and sadness need one of their individual benchmark. This approach avoids false scoring.

Other emotions skip this gate and can only get their scores from their seperate benchmarks. Each score is worth 1.0 points.

This gating is motivated by data from Ekberg et al. (2023), see Theoretical Framework 2.5, showing that pitch and intensity are most distinctive for the two high-arousal emotions, while single universal cues caused possibly misjudged positives in the pilot runs Results 4.2.1. The affected clips were analysed seperately and compared to Table 2.4 for this conclusion.

In this version, the individual extreme values, k, was tested and changed to 1.0 for all except joy, changed to 0.7. This did not affect the results. Version 3 is the final method that is used for data analysis in results. The full code is listed in Appendix 7.1.

Analysis rule-based emotion categorisation and Hume AI

Alignment and divergence between the rule-based function and Hume AI emotion probability is analysed with confusion matrices including both sources top-label for each interview. Mean rating score for the full dataset is visualised with scatter plots followed by Pearson's correlation coefficients to assess relationships and corresponding significance.

3.3.2 RQ1: Comparison Speech-Based AI and Vocal Features

Analysis that are not including the emotion categorisation function includes the following vocal features: Pitch, Loudness/Intensity, HNR, Shimmer, and Jitter. To explore the relationship between Hume AI Emotion labels and vocal features, without feature based emotion categorisation, further analysis were conducted to assess correlation between each acoustic marker and Hume predefined emotions, presented as a heatmap. Further analyses include one-way ANOVA tests, see 2.7 for clarification. To track patterns over time and enable more detailed analysis, each audio recording were segmented into timeframes including Hume AI emotion scoring and vocal features for that specific segmment. These were analyzed by tracking Z-score variations, see 2.7, in key vocal features (pitch, intensity, jitter, and shimmer) throughout each recording. The general time-segment is set to 1.25 seconds. For case examples 4.2.7, different time-segments were tested to observe correlation differences depending on segment length for executed vocal analysis. Beside correlation measurements, top 30% and bottom 70% of emotion probability time-segments are compared to test wheather the mean z-scored feature differs between the high vs. low groups. A large t-statistic value indicate a reliable shift in that feature when Hume rates that emotion high. For the time-segmented analysis, the baseline was set to the average vocal value for our dataset.

Output from Hume AI include segmented results with individual time stamps, varying for each clip. Vocal feature extraction is set to a fixed 1.25s window, therefore, the segments are not fully aligned. By this reason, individual recordings from two interviews are included to visualise how pitch and intensity behaves for Hume labelled joy and sadness.

3.3.3 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. This includes Pearson correlation

coefficients and associated p-values, t-tests and calculations of Cohen's d to evaluate statistical significance and effect sizes, see Theoretical Framework 2.7.

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.

3.3.4 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.

3.3.5 Visual Analysis

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.

Visualizations were also 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.

3.4 Model Configuration

3.4.1 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 than their sentiment analysis endpoint, explained further in 2.4.1. Each text input uses the following prompt:

Listing 3.1: 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.
f"{transcription}"
)
```

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).

3.4.2 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.

3.5 Validity and Reliability

3.5.1 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).

3.5.2 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 same data is not analysed multiple times to check if the results are different. The prompt used for NLP Cloud is therefore zero-shot. 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 noise 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.

3.6 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.

Results

4.1 Presentation of Collected Data

4.1.1 Overview of Interviews

We conducted semi-structured interviews with 15 native Swedish speakers (9 M/6F, age 23-78), each lasting 1-3 minutes. Each participant was interviewed for two different scenarios, resulting in 30 different recordings. 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. Interview ID 003 is deleted from the data collection.

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
4	F	27	4	1	6	1	2	1	6	1	1	3
5	M	29	2	1	3	2	1	1	4	2	2	2
6	F	28	4	1	4	1	2	1	5	1	1	5
7	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 ID, gender (M/F) and age, with counts of Negative (A = Anger, J = Joy, Sad = Sadness, F = Fear, Sur = Surprise) and Positive labels per emotion.

4.1.2 Data Collection for RQ1: Vocal Features & Speech

The collected audio recordings from the interviews were processed for research questions 1 to specifically focus on vocal features and speech-based emotion recognition. See 3.3 for specification and data normalization.

Overview Collected Data

The extracted vocal features, custom emotion categorizations of vocal features, and Hume AI outputs were combined for each recording and stored in JSON format, to enable direct comparison and further analysis. All data is normalized to sum up to 1 before loaded into the JSON files. Table 4.2 includes the mean values and standard diversion for the analysed features (pitch, intensity, HNR, jitter, shimmer) separated by sentiment. Filtered Hume probabilities are presented in Table 4.3 as mean values with standardized diversion for positive and negative recordings. The custom categorization of emotion based on vocal features are presented as mean values including standard diversions for positive and negative sentiments in Table 4.4.

Positive Clips		
Feature	**Mean**	**Std**
mean_pitch_st	−0,3513	6,0752
mean_pitch_hz	155,7733	55,2791
mean_intensity_db	63,5547	3,0952
mean_hnr_db	5,6053	5,8954
jitter_local	0,0253	0,0048
shimmer_local	0,1278	0,0223
formant_F1_hz	592,0180	275,6199
formant_F2_hz	1726,4807	438,5605
formant_F3_hz	2856,0840	377,0989

Negative Clips		
Feature	**Mean**	**Std**
mean_pitch_st	−0,7607	5,8785
mean_pitch_hz	151,5587	51,8871
mean_intensity_db	62,9593	2,5110
mean_hnr_db	5,2880	5,6446
jitter_local	0,0250	0,0033
shimmer_local	0,1264	0,0253
formant_F1_hz	789,0987	344,7615
formant_F2_hz	2021,6813	592,0110
formant_F3_hz	3169,5640	383,1132

Table 4.2: Summary Statistics: Vocal Features by Sentiment

Positive Clips		
Metric	**Mean**	**Std**
hume_anger	0,2279	0,0600
hume_fear	0,1484	0,0530
hume_joy	0,3339	0,1321
hume_sadness	0,1637	0,0650
hume_surprise	0,1262	0,0217

Negative Clips		
Metric	**Mean**	**Std**
hume_anger	0,2768	0,0646
hume_fear	0,1568	0,0382
hume_joy	0,2748	0,1047
hume_sadness	0,1769	0,0673
hume_surprise	0,1147	0,0209

Table 4.3: Summary Statistics: Hume AI Probabilities by Sentiment

Positive Clips		
Metric	**Mean**	**Std**
custom_anger	0,2330	0,0571
custom_joy	0,2177	0,0836
custom_sadness	0,2356	0,0703
custom_fear	0,1163	0,0612
custom_surprise	0,1978	0,0610

Negative Clips		
Metric	**Mean**	**Std**
custom_anger	0,2527	0,0539
custom_joy	0,1916	0,0891
custom_sadness	0,2425	0,0895
custom_fear	0,1169	0,0560
custom_surprise	0,1960	0,0576

Table 4.4: Summary Statistics: Custom Emotion Categorization Scores by Sentiment

Segment-Level Data

Certain analyses in RQ1 rely on time-segmented data. For each recording, Hume AI returns emotion probabilities at regular time segments, an example of this is presented in Table 4.5. In the data analysis, we extract the same set of acoustic features from time-segments with Praat set to a 1.25s window for general analysis, enabling time-to-time comparisons across modalities.

time (s)	anger	fear	joy	sadness	surprise
1,47	0,2332	0,1590	0,4244	0,1214	0,0620
5,15	0,1469	0,0342	0,6693	0,0110	0,1387
8,27	0,0993	0,0259	0,7804	0,0184	0,0759
...	...	...	...	...	...
43,2342	0,1216	0,0837	0,5861	0,0500	0,1586

Table 4.5: Segment-Level Hume Probabilities for clip: id_001_neg

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 analysed with NLP Cloud (text-based), to extract emotion probabilities from the transcription. The same audio was analysed 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, each audio object consists of five emotion labels from each data type (Hume, NLP, Self).

Table 4.6 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.

All Recordings

Emotion	Self Mean	Hume Mean	NLP Mean	Self Std	Hume Std	NLP Std
Anger	0,210	0,260	0,200	0,124	0,072	0,223
Joy	0,312	0,302	0,396	0,200	0,117	0,351
Sadness	0,190	0,167	0,181	0,105	0,065	0,138
Fear	0,136	0,150	0,093	0,061	0,045	0,092
Surprise	0,149	0,118	0,129	0,082	0,022	0,089

Table 4.6: Means and standard deviations of self-reported, Hume, and NLP emotion intensities.

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 4.7 for positively oriented interviews, and in Table 4.8 for negatively oriented interviews. Each table displays the mean and the standard deviation for the respective AI model's emotion probability.

Positive Recordings

Emotion	Self Mean	Hume Mean	NLP Mean	Self Std	Hume Std	NLP Std
Anger	0,103	0,227	0,015	0,032	0,060	0,057
Joy	0,497	0,333	0,708	0,081	0,132	0,169
Sadness	0,117	0,163	0,067	0,059	0,065	0,062
Fear	0,108	0,148	0,040	0,033	0,052	0,067
Surprise	0,173	0,126	0,171	0,098	0,022	0,096

Table 4.7: Means and standard deviations of self-reported, Hume, and NLP emotion intensities for positive recordings.

Negative Recordings

Emotion	Self Mean	Hume Mean	NLP Mean	Self Std	Hume Std	NLP Std
Anger	0,305	0,289	0,363	0,092	0,072	0,183
Joy	0,148	0,275	0,121	0,106	0,098	0,205
Sadness	0,254	0,171	0,282	0,095	0,066	0,103
Fear	0,161	0,152	0,141	0,069	0,038	0,087
Surprise	0,128	0,112	0,092	0,059	0,021	0,064

Table 4.8: Means and standard deviations of self-reported, Hume, and NLP emotion intensities for negative recordings.

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 Evaluation of Emotion Categorisation on Vocal Features

Step-wise alternation of the rule-based function for vocal based emotion probability in Table 4.10 for the tested versions of the category function. Macro F1 scores and UAR (Unweighted Average Recall) uses Hume labels as a relative metrics are pretended in Table 4.10. This should not be seen as ground truth.

The standardised z-scores yielded higher F1 and UAR. However, the emotion probabilities were very similar, all around 0.2 points per emotion. By this reason, it is not used. First version of the rule-based function, V0 resulted in near uniform anger top-labels, with F1 score 0.107. Implementing individual extreme values in V1 resulted in minimal difference from V0. Decreasing intensity for version V2, see Method 3.3.1, while using the standardised z-scores

Variant	N	anger	joy	sadness	fear	surprise
SdZ-Scores	30	18	7	0	5	0
V0_globalK	30	29	1	0	0	0
V1_perK	30	27	1	2	0	0
V2_featW	30	16	8	3	1	2
V3_benchmark	30	15	8	4	1	2

Table 4.9: Emotion distribution on different versions of categorization function.

Variant	Macro–F1	UAR	Joy Rec	Anger Rec
SdZ-Scores	0,282	0,413	0,4	0,667
V0_globalK	0,107	0,183	0	0,917
V1_perK	0,103	0,167	0	0,833
V2_featW	0,204	0,197	0,4	0,583
V3_benchmark	0,208	0,197	0,4	0,583

Table 4.10: Performance metrics for each version of vocal-based emotion categorisation.

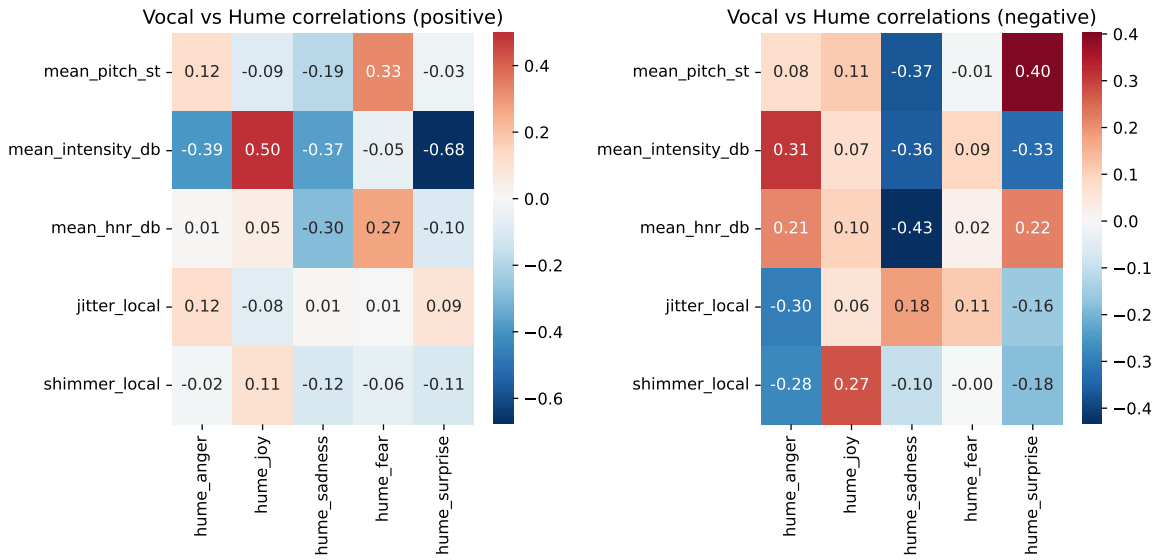
as a fallback minimised the overestimation of anger swapping to top-labelled joy (8), sadness (3), fear (1), and surprise (2). The dispersed emotion-labelling using V2 resulted in F1 = 2.04 and UAR = 1.97, both increased compared to V0 and V1. The benchmark extended V3 with increased feature weight for pitch yielded in slightly higher F1 score, from 0.204 to 0.208, still labelling surprise for two clips. This version is used for the further analysis.

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

Figure 4.1 demonstrates heatmaps of the Pearson correlation coefficients between selected vocal features and Hume AI emotion labels across positive clips in Figure 4.1a and negative clips in Figure 4.1b. The results show generally weak correlations, with slightly stronger correlations for the negative recordings.

Correlation values for positive interviews in Figure 4.1a ranging roughly between -0.7 and 0.5, most values suggest generally weaker correlations than the highest values. Stronger positive correlations suggests that certain feature is higher when Hume rates the correlating emotion. Negative correlations imply the opposite, low value of a certain feature for the correlated emotion. Mean intensity stands out from other vocal features with a moderate positive correlation with Hume Joy (r = 0.50), a moderate negative correlation with Hume Anger (r = -0.39), a moderate negative relationship with Hume sadness (r = -0.37) and a strong negative correlation with Hume Surprise (r = -0.68). Mean pitch shows strongest effect on Hume Fear (r = 0.30) and a moderate negative link with Hume Sadness (r = -0.21). Mean HNR has a moderate negative correlation with Hume Sadness (r = -0.30) as well it is moderately positively correlated with Hume Fear (r = 0.27). Jitter and shimmer remain near zero for most emotions in the positive clips, with none exceeding $|r| = 0.13$. This suggests that, in more positively oriented interviews, variation in pitch, intensity, and HNR capture the core emotion-related cues moderately, while jitter and shimmer have small predictive influence in semi-spontaneous speech during interview conversations.

Figure 4.1b illustrates Pearson correlation values for negative clips in the dataset, again presenting generally weak effects even if slightly stronger correlations occur, approximately ranging between -0.5 and 0.5. The strongest relationship appears for sadness, where mean



(a) Hume vs vocal features positive

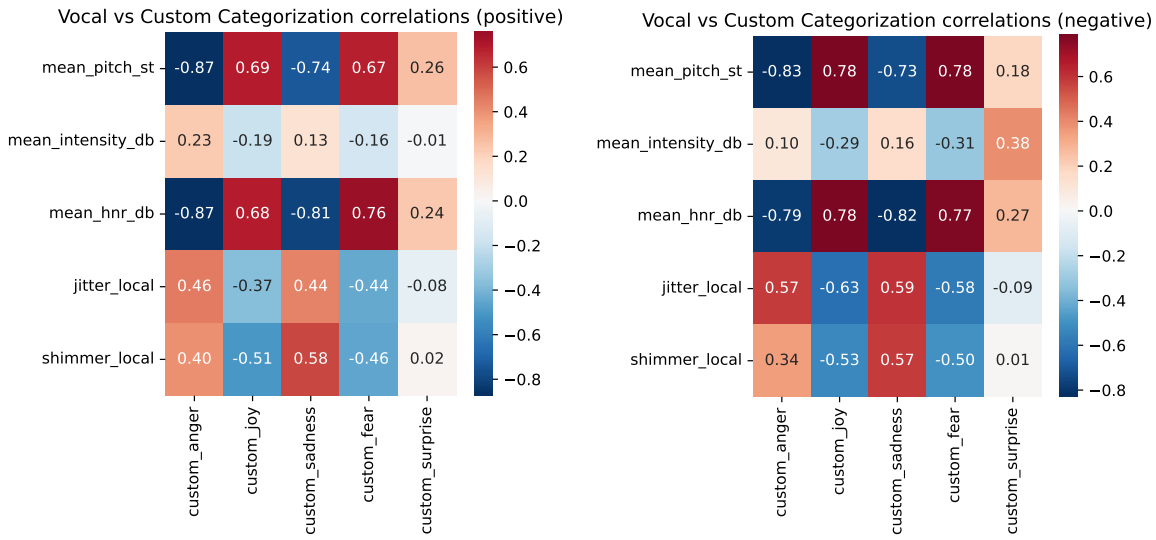
(b) Hume vs vocal features negative

Figure 4.1: Correlation Heatmaps between Hume AI and vocal features.

HNR shows a strong negative correlation (r = -0.43) and moderately negative correlated with pitch (r = -0.37) and intensity (r = -0.36). Anger had the highest positive correlation with intensity (r = 0.31) followed by a weak positive link with HNR (r = 0.21) and weak to moderate negative correlations with jitter (r = -0.30) and shimmer (r = -0.28). Correlations between vocal features and joy are all perceived weak, shimmer emerges as strongest (r = 0.27) compared to all other features (r < 0.11). Hume predicted fear presents coefficients indicating no linear correlation for all vocal features (r < 0.11). Surprise presents a moderate positive correlation with pitch (r = 0.36) and negative relationship with intensity (r = -0.33), other features are weakly correlated to the emotion. Jitter and shimmer show similar relationships to Hume Fear (r 0.2). Shimmer has strongest correlation with Hume Anger, and a moderate relationship with Hume Joy. Jitter is moderately correlated with Anger as well and have a moderate correlation with Sadness where shimmer has almost no correlation.

Overall, the negative and positive diversion suggests that intensity consistently reflect Hume AI's anger, sadness, and surprise predictions, and more prominent for joy in positive recordings and near zero for joy rated in negative contexts. Correlations between pitch and emotions predicted by Hume is generally weak and varying between the sentiment categories. Jitter and shimmer remain minor indicators for positive conditions, while having moderate correlations with Hume anger and joy in negative recordings.

4.2.3 Correlation with Rule-Based Emotion Scores



(a) Positive recordings.

(b) Negative Recordings

Figure 4.2: Correlation heatmaps for custom categorized emotion scores vs. vocal features.

Figure 4.2 presents heatmaps of Pearson's correlation r between selected vocal features and the emotion scores obtained from the custom emotion categorization function for Figure 4.2a positive oriented clips and Figure 4.2b negative clips. In contrast to the correlation results between Hume AI and vocal markers, these heatmaps present similar values and orientations for both negative and positive recordings, with values roughly ranging from -0.9 to 0.8 that implies strong correlations. Pitch, HNR, and shimmer has highly related patterns where the rule-based function shows a strong negative correlation for anger with pitch and HNR (r(≈ -0.87)), joy have significant relationships with pitch and HNR as well, slightly higher in negatives (r = 0.78). Additionally, these features show a strong negative link with sadness, approximately r = 0.81 for HNR and r = -0.73 for both sentiments. Strong positive correlations are found with fear that are corresponding in negatives ($r(\approx 0.77)$), but slightly more dispersed in positives (pitch r = 0.67, HNR r = 0.76). Shimmer has the strongest positive correlation with sadness for both sentiments (($r \approx 0.57$) respectively) and a strong negative relation with joy (($r \approx -0.57$)).

Intensity reveals weak correlation across all features in both sentiment contexts except for a moderate positive relation with surprise in the negative subset (r = 0.38). Beside intensity, appear only weak correlations for surprise in both subsets ($|r| < 0.27$). Jitter has somewhat stronger correlations in negative recordings, where sadness (r = 0.59) and joy (r = -0.63) is most distinct.

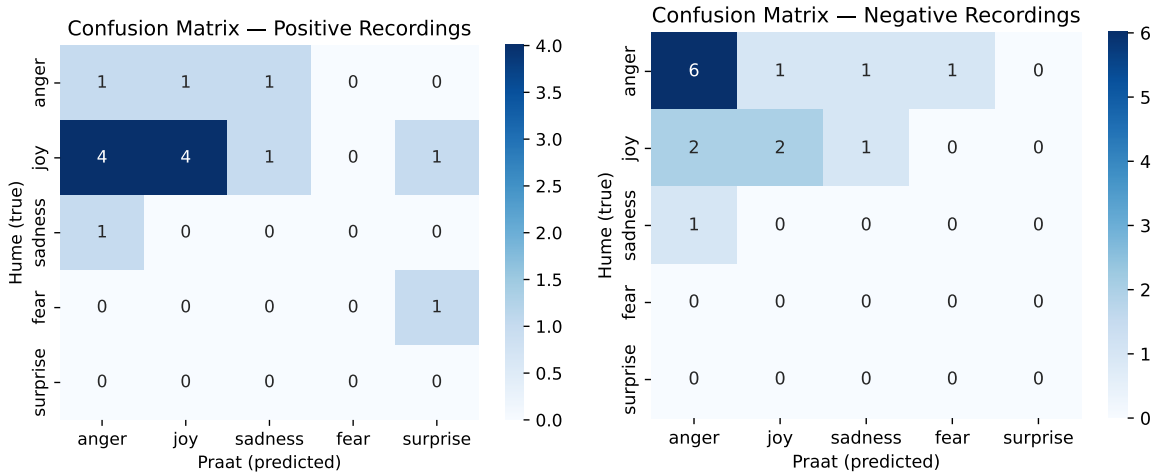
The similar values for pitch and HNR vocal features with strong correlation coefficients, implies that harmonic clarity is weighted almost identically to pitch in the categorization function, both with a dominant impact on the outcomes of the emotion categorisation.

4.2.4 Correlation Rule-Based Categorization and Hume AI Labels

To explore the alignment of top-labelled emotion by Hume AI and customised categorization based on vocal features, Figure 4.3 illustrates the correlations in a confusion matrix, treating Hume labels as ground truth and comparing these to emotions grouped by the custom function. For positive recordings, Hume and the custom function agree on anger for one clip, while Hume rates joy highest in four cases where the rule-based function selects anger. Both methods

rate joy as the top emotion for four positive oriented clips. Divergencies occur for fear-joy, sadness-anger, anger-joy, anger-sadness, and joy-sadness pairs.

In negative contexts, the methods agree on anger as the top emotion for six recordings and agree on joy for two clips. Beside these, the top-rated emotion is discrepant where Hume labels two clips as joy that the custom function rates as sadness, and two anger-rated clips by Hume are rated as joy by the customized categorization. Other divergences occur for joy-anger, sadness-anger, and anger-sadness in single-pairs.



(a) Positive Recordings

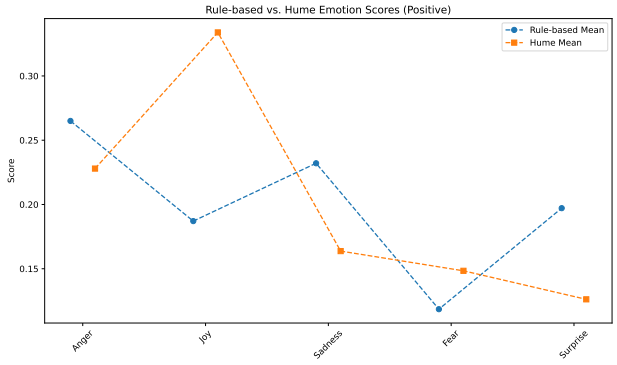
(b) Negative Recordings

Figure 4.3: Confusion matrix diagram for top label emotion between Hume and Custom Categorization.

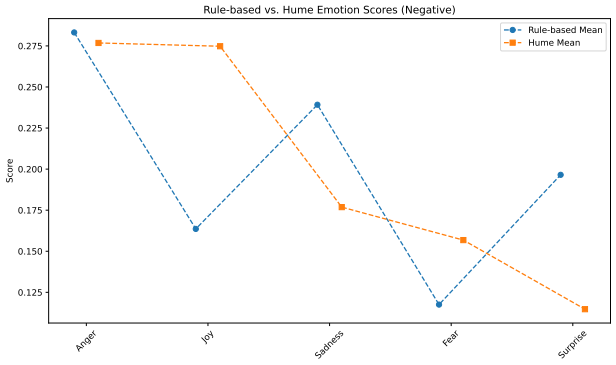
The mean rated scores for Hume versus custom categorized emotion labels for positive recordings are illustrated in Figure 4.4a, negative in Figure 4.4b. As shown, several ratings follow the same pattern while other emotions diverge, as Hume joy are significantly higher for both sentiment categories while the rule-based approach overestimated surprise compared to the Hume ranking.

Rule-based labelled joy, fear, and surprise are ranked similarly in both positive and negative contexts. This pattern aligns with Hume probabilities that are relatively similar for these emotions in both subsets, including sadness as well. Anger is the most divergent emotion label across sentiments, with higher ratings in negative clips that are relatively corresponding in the diagrams. Significant divergencies between the methods appear for joy, sadness and surprise, marginally wider dispersed for negative recordings.

Pearson's correlation coefficients between Hume and rule-based emotion labels are presented in Appendix Table 7.3, measuring the correlation between each emotion score. No significant correlation occurs for either of the sentiments. Moderate, yet unsignificant, correlation is revealed as diverged for anger ($r = -0.320$, $p = 0.2451$) where Hume predicts the emotion higher while rule-based results in lower score for anger in positive contexts. Sadness has a weak positive correlation between the sources for both subsets ($r = 0.265$, $r = 0.222$), suggesting that the output is correlated, yet unsignificant ($p > 0.2$). The correlations for the remaining emotions are perceived weak ($|r| < 0.24$).



(a) Mean Hume vs. Custom Emotion, positive



(b) Mean Hume vs. Custom Emotion, negative

Figure 4.4: Scatter Plots of Mean Hume and Custom Emotion Scores for positive and negative recordings.

4.2.5 Limitations of the Custom Vocal Emotion Categorization Method

The rule-based function, customised to group emotions by vocal features, assigns equal top scores to multiple emotions for the same clip in 27% cases. By contrast, Hume has one singular top label for 29 of 30 clips. Furthermore, the function cannot see a reliable measurement for comparison of Hume with vocal markers. Therefore, the following analysis consists of single vocal marker behaviour in contrast to Hume emotion probabilities.

4.2.6 ANOVA Tables of Vocal Features Across Emotions

An ANOVA was implemented to further examine whether essential vocal features varied across Hume-labelled emotions. This was conducted on mean values of full recordings for pitch, intensity, HNR, jitter, and shimmer. The results are summarised in Appendix, Table 7.4a for positive recordings and 7.4b for negative recordings, showing that none of the features showed statistically significant differences between the five Hume emotion categories (all p-values $>$ 0.13, ranging up to p 0.90). 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.

4.2.7 Time-to-Time Analysis

Time-to-Time Analysis: Full Dataset

To understand both if acoustic cues correlates with emotion probability by Hume and when they produce clear categorical shifts, two analyses were conducted at time segmented level. These effects were observed further with sentiment separations of all, negative, and positive contexts to see if the emotion-acoustic feature relationship are dependent on positive vs. negative sentiment. Table 4.11 presents the significant correlations between z-scored acoustic features and Hume emotions.

Feature	Emotion	Pearson's r	p-value	Significant
pitch	joy	0.065	0.0448	Yes
pitch	sadness	-0.230	0.0000	Yes
pitch	fear	0.082	0.0110	Yes
intensity	joy	0.164	0.0000	Yes
intensity	sadness	-0.142	0.0000	Yes
intensity	fear	-0.173	0.0000	Yes
intensity	surprise	-0.110	0.0006	Yes
HNR	sadness	-0.253	0.0000	Yes
HNR	fear	0.112	0.0005	Yes
jitter	sadness	0.084	0.0091	Yes

Table 4.11: Significant Pearson correlations for the full dataset.

Only significant correlation is included (all $p < 0.05$). 9 correlations out of 24 indicated significance, the full analysis is presented in Appendix [FIGURE REF]. All correlations are perceived as weak even if there is statistical significance. Pitch correlated positively with joy (r = 0.065), negatively with sadness (r = -0.230) and positive with fear (r = 0.082). Intensity had a increased relationship with joy (r = 0.164), but decreased with sadness (r = -0.142), fear (r = -0.173) and surprise (r = -0.110). HNR showed a weak negative correlation with sadness (r = -0.253) and positive with fear (r = 0.112). Jitter had a single significant, yet weak correlation with sadness (r = 0.084). Shimmer showed no significant correlations.

Table 4.12 compares the top 30% and bottom 70% of emotion-probability time-segments, and tests whether the mean z-scored feature differs between the high vs. low groups. A large t-statistic value indicates a reliable shift in that feature when Hume rates that emotion high.

Feature	Emotion	t-statistic	p-value	Significant
pitch *	anger	2.529	0.0116	Yes
pitch	joy	2.293	0.0221	Yes
pitch	sadness	-7.769	0.0000	Yes
pitch	fear	5.185	0.0000	Yes
intensity *	anger	1.975	0.0485	Yes
intensity	joy	4.602	0.0000	Yes
intensity	sadness	-3.981	0.0001	Yes
intensity	fear	-3.567	0.0004	Yes
intensity	surprise	-2.949	0.0033	Yes
HNR	anger	2.709	0.0069	Yes
HNR	sadness	-7.506	0.0000	Yes
HNR	fear	4.914	0.0000	Yes
HNR *	surprise	2.287	0.0224	Yes
jitter **	sadness	1.811	0.0705	No

Table 4.12: High-vs-low t-test results for significant acoustic features (full dataset)

High-anger predictions of segmented recordings shifted towards higher pitch (t = 2.529), intensity (t = 1.975) and HNR (t = 2.709), none of these associations (*) was significant in the data for Table 4.11. Joy segments had increased shifts in pitch (t = 2.293) and intensity (t = 4.602), while fear segments have lower intensity (t = -3.567) and increased pitch (t = 5.185) and HNR (t = 4.914). High-sadness segments showed notable decreased pitch (t = -7.769),

intensity (t = -3.981), HNR (t = -7.506), and a small, not significant increase in jitter (**) (t = 1.811, p = 0.0705) that showed a significant correlation in Table 4.11. Segments with high surprise predictions has the moderately low intensity (t = -2.949) and a modest positive differentiation for HNR (*) (t = 2.287) with no significant correlation in Table 4.11.

Time-to-Time Analysis by Sentiment

(a) Pitch and Anger (Pearson r)

Sentiment	r	p	Sign.
All	0.032	0.3284	No
Positive	-0.042	0.3772	No
Negative	0.101	0.0209	Yes

(b) Pitch and Anger (t-test)

Sentiment	t	p	Sign.
All	2.529	0.0116	Yes
Positive	1.392	0.1647	No
Negative	2.642	0.0085	Yes

Table 4.13: (a) Pearson correlations and (b) high-vs-low t-test results for pitch vs. anger by sentiment.

Table 4.13 presents an absent correlation between pitch and anger for the full dataset (r = 0.032, p = 0.3284), and positive clips (r = -0.042, p = 0.3772) but a significant, yet weak correlation in the negative subset (r = 0.101, p = 0.0209). The t-tests in 4.13b confirms this where high-anger segments have moderate higher pitch in the negative set (t = 2.642, p = 0.0085) and in for all clips (t = 2.529, p = 0.0116), but not in positive contexts. This implies that pitch is a considerable signal of anger when the overall context is negative. Table 4.12 presents a significant shift in HNR for Hume anger (t = 2.709, p = 0.0069), yet the correlation between them were not significant in Table 4.11. The t-tests reveal stronger differences in HNR for high vs low anger than both pitch, and intensity (t = 1.975) Together with pitch, HNR appears to be the most prevalent features correlated with anger in negative circumstances, even if the shifts in vocal features for this emotion is fairly weak.

(a) Intensity and Joy (r)

Sentiment	r	p	Sign.
All	0.164	0.0000	Yes
Positive	0.175	0.0002	Yes
Negative	0.152	0.0005	Yes

(b) Intensity and Joy (t-test)

Sentiment	t	p	Sign.
All	4.602	0.0000	Yes
Positive	3.343	0.0009	Yes
Negative	2.375	0.0179	Yes

Table 4.14: Intensity–Joy correlations and t-test results by sentiment.

Table 4.14 demonstrates the correlation between intensity and joy, the most prominent feature for the emotion. All contexts show a positive relationship, strongest for positive clips (r = 0.175, p = 0.0002) with significant mean differences where the full dataset has highest significance (t = 4.602, p < 0.001). Pitch and joy correlations is presented in Table 4.11, showing unlike anger, a weak correlation for the full dataset (r = 0.065, p = 0.0448). Pitch reaches a weak correlation with joy in positive recordings (r = 0.110, p = 0.0210) and no significance in the negative subset. The t-tests is only significant in the full dataset (t = 2.293, p = 0.0221), not for the positive clips (t = 1.861, p = 0.0701), implying a more context-specific and non-linear affect. No other features revealed significancy in either correlation or t-statistics. These results suggests that intensity is a reasonably stable cue to joy regardless of the overall sentiment context, even if higher correlation occurs for the full and positive sets than the negative subset (r = 0.152 and t = 2.375).

(a) Pitch and Sadness (r)

Sentiment	r	p	Sign.
All	-0.230	0.0000	Yes
Positive	-0.275	0.0000	Yes
Negative	-0.187	0.0000	Yes

(b) Pitch and Sadness (t-test)

Sentiment	t	p	Sign.
All	-7.769	0.0000	Yes
Positive	-6.332	0.0000	Yes
Negative	-4.980	0.0000	Yes

Table 4.15: Pitch–Sadness correlations and t-test results by sentiment.

Table 4.15 display sentiment correlations for pitch and Hume predicted sadness, with significant difference between high and low groups indicating prominent shifts in pitch when Hume rates sadness high. Correlations are weak, but significant for all sentiment contexts. Positive recordings show the largest negative correlation (r = -0.275, p < 0.001), although the full dataset reveals greater diverge in variations (t = -7.769, p < 0.001) which is the largest t-statistic for all emotion-feature groups. Intensity has a weaker shift than pitch for high-sadness segments (t = -3.981) while HNR demonstrates a similar pattern as pitch (t = -7.506). These prominent feature-shifts suggests that reduced pitch and HNR are indicators of sadness independent from sentiment orientation.

Case Examples

For a more concrete illustration of the prior tendencies, three interview recordings were analysed 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. The navy dashed trace represents the probability of the emotion estimated by Hume, scaled to 0-1 on the right y-axis. The solid blue and orange curves show mean vocal features for time-segments, plotted as Z-scores against the left y-axis. The x-axis represents time given in seconds.

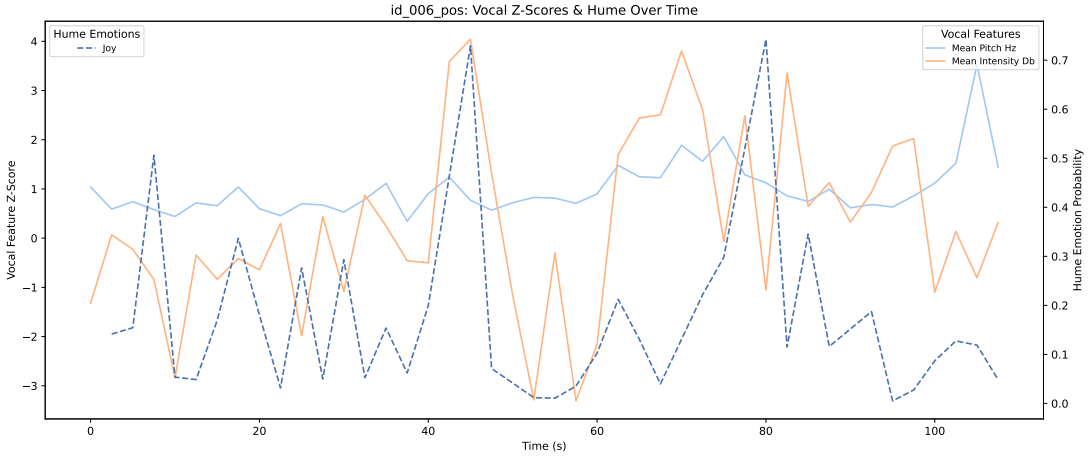


Figure 4.5: Pitch and Intensity vs. Hume Joy for single positive clip.

Figure 4.5 illustrates some similar patterns between high-predicted Hume joy, pitch and intensity that show similar magnitude as the emotion trace. Segment time for analysis was 1s. Corresponding low values occur around 10s, 50-55s, 75s, while aligned raises appear around 20s and 45s. Statistical tests in Appendix Table 7.7 and 7.8 confirm the visual trend, where peaks in joy trace often coincide with rises in the orange intensity curve. A significant, moderate correlation between joy and intensity is evident (r = 0.351, p = 0.0134), visualised in Figure 4.5,

aligning with the pattern for these curves. Pitch shows no significant correlation (r = 0.188, p = 0.1965) and is weaker than that with intensity, which is consistent with the figure. Table 7.8 in Appendix further suggests a moderate shift toward higher intensity when Hume labels joy with t = 2.718 for the top-30% joy frames. Therefore, t-tests reveals that the highest-joy segments are associated with higher intensity than the remaining segments of the clip.

The same clip with fixed time-segments for vocal analysis using 2 second frames yielded weaker correlation for pitch (r = 0.078, p = 0.6103), but higher for intensity (r = 0.329, p = 0.0272), listed in appendix Table 7.7.

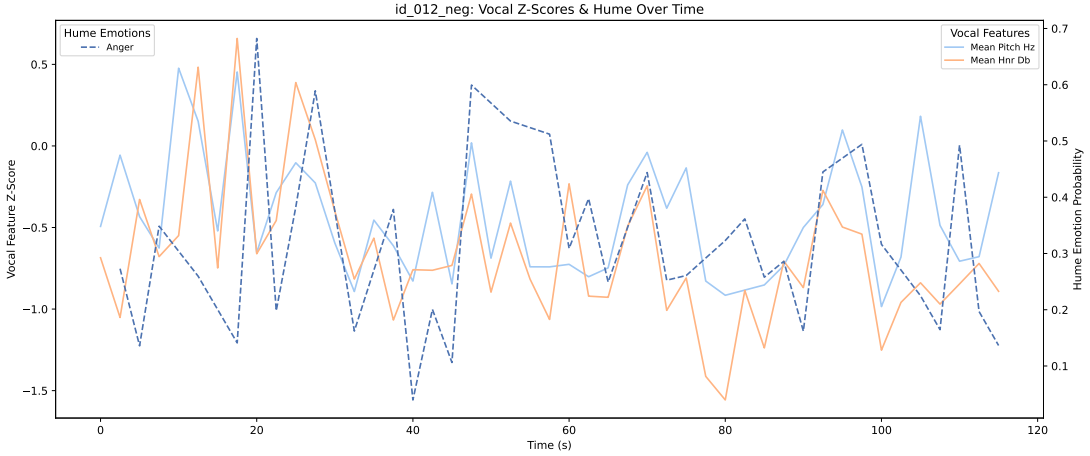


Figure 4.6: Pitch and HNR vs. Hume Anger for single negative clip.

Anger correlation with pitch and HNR is presented for one negative recording in Figure 4.6, vocal features segmented into 1.25s windows. The vocal features show a similar pattern as the emotion probability in certain time segments, where both pitch and HNR have relatively aligned magnitude and direction. Correlation values and t-statistics is listed in Appendix Table 7.9 and 7.10, no significant correlation is found for either feature.

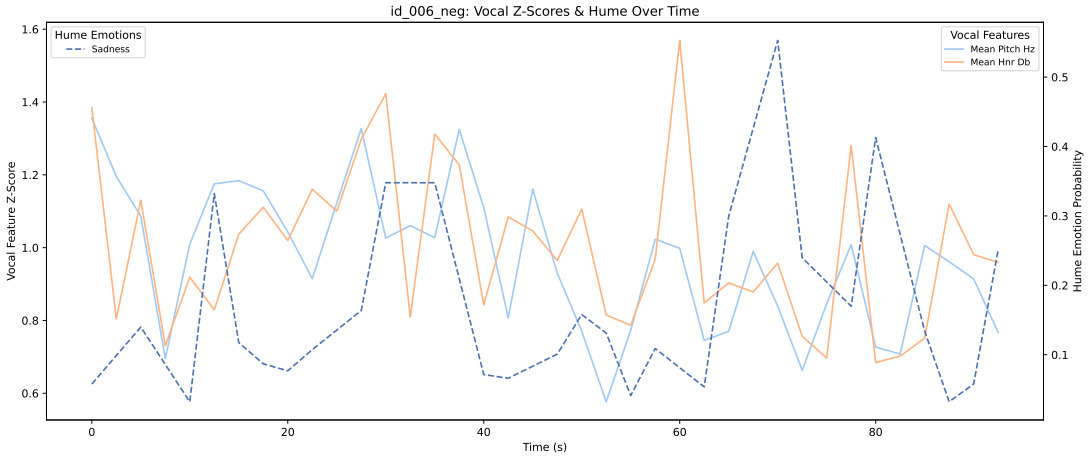


Figure 4.7: Pitch and HNR vs. Hume Sadness for single negative clip.

The opposing pattern for sadness shown in prior results, see Table 4.12, are demonstrated in Figure 4.7 with 2s segments. Clear fluctuations in magnitude where the Hume labelled emotion have adverse direction. Low sadness (dashed navy trace) occurs in time frames where both acoustic markers tend to raise (e.g. 5s, 20s, 45s, 60s, 90s). Reversed pattern with high sadness segments are resembled as well (e.g. 45s, 70s, 80s). Analysis data is listed in Appendix Table 7.11 and 7.12, sadness-pitch showed a negative moderate correlation (r =

-0.333, p = 0.072), yet unsignificant. These table includes differences when segmenting at 2s compared to 1s. T-statistic in pitch shifts for high-sadness showed significance (r = -2.203, p = 0.036), implying pitch was low in high-sadness segments. These statistics endorse the visual interpretation with negative correlation coefficients for both pitch and HNR even if statistical significance is absent.

4.2.8 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-labelled 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 the degree to which two modalities for AI-based emotion recognition systems - speech-based (Hume AI) and text-based (NLP Cloud) - agree or diverge when labelling emotional expressions in semi-structured interviews. We examine five target emotions (anger, joy, sadness, fear, surprise) across the full dataset, as well as positive and negative interviews separately. To acquire a detailed picture of how the models align, we compare their average emotion scores, measure Pearson correlations and paired t-tests with Cohen's d. This multimethod approach supports a comprehensive understanding of how the two modalities responds to the same emotional input, to find mutual strengths and diverse tendencies in how they classify emotions.

4.3.1 Comparative Overview of Model Outputs

As presented in Table 4.6, Table 4.7, and Table 4.8 (4.1.3 Data Collection), the mean emotion scores and standard deviations differ between the two models across the full dataset, including patterns within positive and negative interviews.

Figure 4.8 visualises these differences for positive and negatives recordings separately. As presented, anger in positive interviews was detected as significantly higher levels by Hume compared to NLP, that rated anger near zero. For the negative interviews, the rating was more aligned where NLP rated anger slightly higher. Joy is rates substantially high by NLP in the positive interviews, compared to both other emotions and Hume's probability. In contrast, Hume rates joy higher than NLP for negative recordings. Sadness and fear are both rated higher by Hume than NLP in positive contexts, while NLP rates sadness higher in negative contexts where fear has more aligned scoring by the systems. Surprise was detected at similar, low levels by both models for both sentiment categories. Highest contrast for surprise is found in positive interviews where NLP rated it slightly higher.

The differences in the average emotion scoring are presented further in Figure 4.9. Positive values indicate that Hume AI assigned higher scores for respective emotions, while negative values imply higher scores from NLP Cloud. As explained for Figure 4.8, the most evident

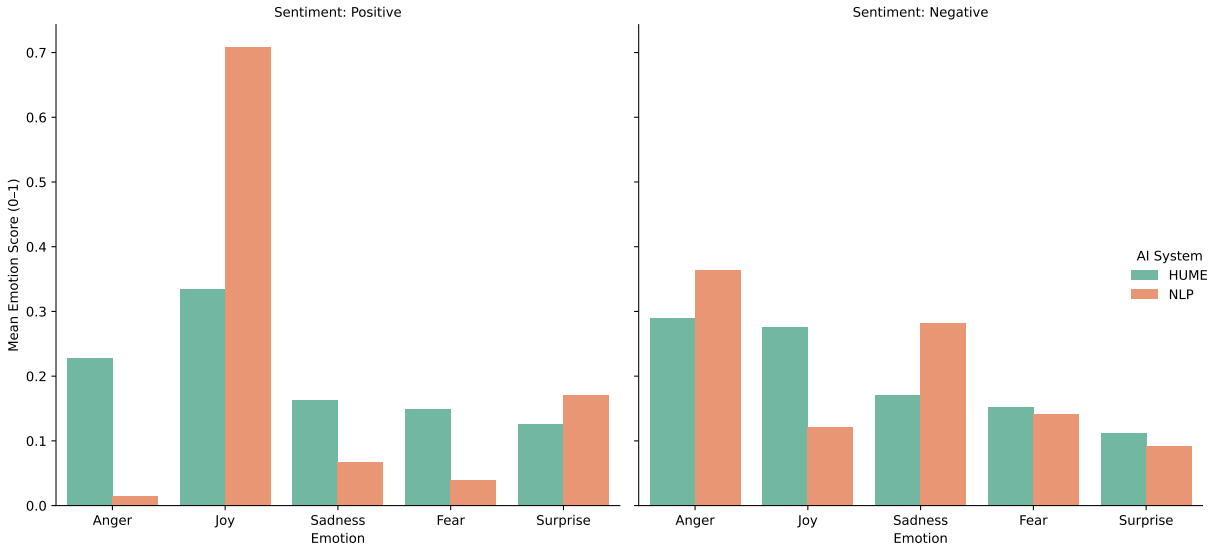
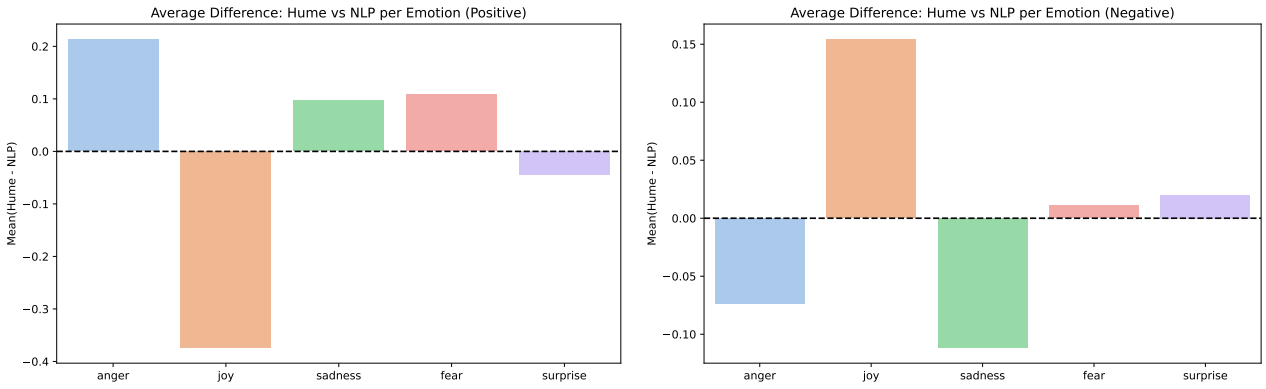


Figure 4.8: Average emotion score for Hume AI and NLP Cloud, seperated by positive and negative recordnigs.

difference was shown for joy in both sentiment contexts, where NLP rates it significantly higher in positive settings and Hume higher in negative settings. Differences for sadness and surprise were insignificant in negative interviews, aligned with surprise in positive interviews. In Figure 4.9, the divergence in rating of fear in negative contexts is obvious where NLP rated the emotion more frequent.



(a) Positive recordings.

(b) Negative recordings.

Figure 4.9: Average difference in emotions scores between Hume and NLP.

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 4.16 includes the full dataset (positive and negative recordings), presenting the correlation values as well as corresponding p-values to examine the statistical significance.

All recordings

Emotion	Pearson r	p-value	Significant
Anger	0.466	0.007	Yes
Joy	0.521	0.002	Yes
Sadness	0.167	0.362	No
Fear	0.171	0.348	No
Surprise	0.197	0.281	No

Table 4.16: Pearson Correlations Between NLP and Hume Emotion Scores (Full dataset)

This data demonstrates a reasonable positive correlation for Anger($r = 0.466$, $p=0.007$) and Joy ($r=0.521$, $p=0.0022$), implying that these emotions are relatively consistent identified throughout the AI systems on the full dataset. 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 when analysing the full dataset. 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.

Positive Recordings

Emotion	Pearson r	p-value	Significant
Anger	0.160	0.568	No
Joy	0.682	0.005	Yes
Sadness	0.546	0.035	Yes
Fear	0.098	0.729	No
Surprise	-0.050	0.860	No

Table 4.17: Pearson Correlations Between NLP and Hume Emotion Scores (Positive)

Table 4.17 presents the same data as Table 4.16, but for positive recordings separately. As for the full dataset, Joy shows a significant correlation ($r = 0.682$, $p = 0.005$) between the model's detection. In contrast, Anger has a lower correlation ($r = 0.160$, $p = 0.568$) in positive contexts and Sadness presents a significant correlation ($r = 0.549$, 0.035) distinct from the full dataset. Fear and Surprise has even lower correlations for positive recordings than the dataset combined, with p-values close to 1.

Negative Recordings

Emotion	Pearson r	p-value	Significant
Anger	0.260	0.313	No
Joy	0.556	0.020	Yes
Sadness	0.028	0.914	No
Fear	0.270	0.294	No
Surprise	0.209	0.422	No

Table 4.18: Pearson Correlations Between NLP and Hume Emotion Scores (Negative)

Table 4.18 summerizes the correlation coefficients for the negative recordings. Consistent with the full dataset and positive subset, Joy again demonstrated a significant correlation in negative contexts (r = 0.556, p = 0.020). All other emotions failed to reach significance, with values that markedly diverged from their corresponding values in the positive recordings: Sadness resulted r = 0.028 (p = 0.914) versus r = 0.546 (p = 0.035) for positives, and Fear showed r = 0.270 (p = 0.294) compared to r = 0.098 (p = 0.729) in the positive context.

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 4.19 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	0.094	No	-0.305
Sadness	-0.548	0.588	No	-0.097
Fear	3.341	0.002	Yes	0.591
Surprise	-0.657	0.516	No	-0.116

Table 4.19: 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, suggesting a systematic modality difference for this emotion. Although Anger, Joy, Sadness, and Surprise had some mean-score differences, none reached statistical significance (all p>0.05) and their effect sizes were small ($|d| < 0.03$). Apart from Fear, the two models demonstrated close agreement in recognizing these emotional expressions.

Positive Recordings

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	Yes	1.323
Surprise	-1.723	0.107	No	-0.445

Table 4.20: t-statistics, p-value with significance, and Cohen's d for positive interviews.

Table 4.20 demonstrates t-tests and Cohen's d for positive oriented interviews, where all emotions except for surprise (t = -1.723, p = 0.107, d = -0.445) shows significant differences (p $<$ 0.001) with certainly large effect sizes. Negative T-value and Cohen's d for Joy (t = -11.665, d = -3.012) indicates that NLP have the aspects of overestimating this emotion compared to Hume with large effect sizes, where Hume in contrast tends to overestimate Anger (t = 10.903, d = 2.815) in positive contexts. Hume rates Sadness and Fear more prominent than NLP, and Surprise remain inconsistent as previous results with no significant difference (t = -1.723, p = 0.107).

Negative Recordings

Emotion	t-statistic	p-value	Significant	Cohen's d
Anger	-1.702	0.108	No	-0.413
Joy	3.720	0.002	Yes	0.902
Sadness	-3.796	0.002	Yes	-0.921
Fear	0.536	0.599	No	0.130
Surprise	1.311	0.208	No	0.318

Table 4.21: t-statistics, p-value with significance, and Cohen's d for negative interviews.

Table 4.21 presents conducted t-tests and Cohen's d in negative interviews, with significant differences for Joy (t = 3.720, p = 0.002), where Hume rates it significantly higher than NLP. In contrast, NLP has clear higher scoring for Sadness with large effect size (t = -3.796, d = -0.921). However, the effect sizes are not as big as for the positive recordings. For example, the effect sizes for Joy (t = 0.902) are lower than Joy in positive contexts (t = -3.012) where NLP overestimated the emotion compared to Hume. 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.

Conclusion Statistical Analysis

Comparison of speech-based (Hume AI) and text-based (NLP Cloud) with statistical analysis demonstrates correlation particularly for clear expressed emotions as Anger and Joy when analysing the full dataset. However, anger shows no correlation between the models for either positive or negative recordings when separated. Joy shows a significant correlation throughout all sentiment contexts, where t-tests confirmed that NLP had higher predictions for joy in positive contexts and Hume in negative. Emotions that are more subtle like Sadness, Fear, and Surprise, revealed low correlations for all sentiment contexts except positive that showed a

strong correlation for sadness, indicating modality-specific distinctions. Paired t-tests strengthened this observation regarding the full dataset and negative subset, pointing out Fear as the only emotion with statistically significant divergence in the full dataset where speech-based analysis assigned higher scores consistently. However, negative recordings showed significant difference for joy and sadness, while positively oriented clips showed significant difference for all emotions except surprise.

Conclusion Sentiment-Based Analysis

In conclusion, Hume AI and NLP Cloud show moderate to strong agreement on anger (r = 0.47) and joy (r = 0.52) across the full dataset, but weak correlations on sadness, fear, and surprise. Paired t-tests showed that fear is the single emotion that exhibits a significant mean difference across the full interview set (Hume > NLP, d = 0.59), while joy and anger showed modality divergencies in positive and negative subsets where NLP overestimated joy in positive interviews (d = 3.01) and Hume overestimated anger (d = 2.82). These results suggest that text and speech modalities agree on certain emotions particularly when considering the full dataset. However, divergencies occur for sentiment-specific analyses, especially for positive interviews.

4.3.3 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) – aligns with participants' own emotion ratings, to evaluate the agreement and divergence in different interview sentiments (positive and negative). This section includes average emotion scores from self-reports, Hume AI, and NLP Cloud across all recordings and for each sentiment category. Linear agreements are quantified with Pearson correlations and mean-level differences are analysed with paired t-tests and Cohen's d for assess effect sizes. This approach allows to see the overall alignment between AI-models and participants own assessment as well as how it depends on the sentiment context.

4.4.1 Model Emotion Score and Self-Reports Comparison

An initial overview is summarised in Table 4.6, Table 4.7, and Table 4.8 (4.1.3 Data Collection), with average emotion scores across all 30 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.Table 4.7 includes the same values for positive recordings and Table 4.8 presents the data from negative recordings.

These differences are visualized in Figure 4.10, illustrating a bar chart that compares the average emotion scores defined by Hume AI, NLP Cloud, and participants self-assessment seperated by positive and negative oriented interviews.

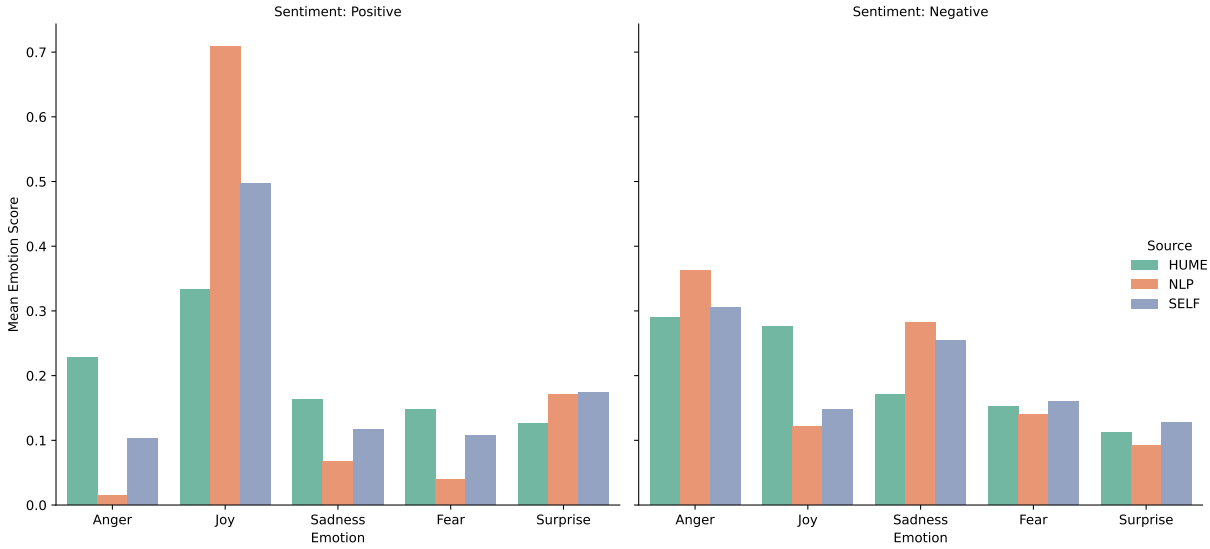


Figure 4.10: Comparison of emotional labels for Hume, NLP, and self-assessed.

For the positive recordings, Joy consistently had higher self-reported scores than other emotions. NLP rates Joy higher than the participants while Hume rates it lower. In contrast, Joy was markedly rated higher by Hume in negative contexts than NLP and self-reporting which rated the emotion equally. The negative related emotions (Anger, Sadness, Fear) were assessed at lower levels by participants in the positive interviews. Self-assessed scores generally matched Hume's higher detection of Sadness, Fear and Surprise than NLP's minor predictions. Anger had higher rating by Hume than both self-reports and NLP in positive contexts. Surprise had similar average score across all sources, slightly lower detection rate by Hume.

For negative recordings, Anger had similar rating across all sources, with slightly higher rating by NLP. Joy has markedly higher average score by Hume compared to the other sources, while the speech-model rates Sadness lower than the text-model and participants. Fear and surprise had aligned rating by all sources, with a similar pattern where both emotions are rated slightly higher by the participants, closely followed by Hume and lowest rating by NLP.

This comparison suggests that emotional ranking are more aligned in for the negative recordings, where Joy is the emotion most distinct in the rating by Hume. The positive oriented interviews have more varying results between the sources, Joy are significanly rated higher by NLP while the text-based model rates Anger close to zero compared to Hume that rates this emotion higher than all other emotions except for Joy.

The sentiment-based comparison clearly presents that emotional expression and self-awareness have a signficant variance between modalities and emotional contexts. 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.

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

All Hume

Emotion	Pearson's r	p-value	Significant
Anger	0,359	0,043	Yes
Joy	0,334	0,062	No
Sadness	0,050	0,784	No
Fear	−0,007	0,969	No
Surprise	0,088	0,631	No

Table 4.22: Pearson's r, p-values, and significance for all Hume recordings.

The correlation results for Hume AI predictions on all recordings in the dataset is demonstrated in Table 4.22, 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. Joy shows a moderate correlation but without statistical significance (r = 0.334, p = 0.062), other emotions, such as Fear (r = 0.007, p = 0.969), presents no relevant correlation.

(a) **Positive Recordings (Hume)**

Emotion	r	p	Sign.
Anger	0.404	0.136	No
Joy	0.401	0.138	No
Sadness	0.320	0.244	No
Fear	–0.027	0.924	No
Surprise	0.091	0.748	No

(b) **Negative Recordings (Hume)**

Emotion	r	p	Sign.
Anger	–0.105	0.690	No
Joy	0.127	0.627	No
Sadness	–0.146	0.576	No
Fear	–0.036	0.891	No
Surprise	–0.143	0.585	No

Table 4.23: Pearson's r, p-values, and significance for Hume AI vs. self (positive and negative).

Table 4.23a presents correlation coefficients for positive recordings with no significant agreement occurs between Hume predictions and self-reported emotions. Anger, sadness, and sadness show moderate correlations (r = 0.320-0.404) with no statistical significance (p = 0.136-0.244). Weak correlation appears for both fear and surprise with high p-values suggesting no convincing evidence for these correlations.

Negatively oriented interviews, Table 4.23b show similar results as for positive interviews where no correlations of significance are found (r = -0.146-0.127, p = 0.576-0.0.891). Four out of five emotion correlations are negative while joy has a weak positive relationship. Each correlation is considered weak without statistical significance, implying that Hume predicted emotions distinct from participants own evaluation.

NLP Cloud vs Self-Reported Emotions

All NLP

Emotion	Pearson's r	p-value	Significant
Anger	0,739	0,000	Yes
Joy	0,863	0,000	Yes
Sadness	0,710	0,000	Yes
Fear	0,669	0,000	Yes
Surprise	0,092	0,616	No

Table 4.24: Pearson's r, p-values, and significance for all NLP recordings.

Table 4.24 presents correlation coefficients between self-reported and NLP-predicted emotions for the full dataset. When analysing the full dataset, NLP Cloud demonstrated strong and statistically significant correlations with self-reporting for four of five emotions. Joy showed the strongest correlation ($r = 0.863$, $p < 0.001$), followed by Anger ($r = 0.739$, $p < 0.001$) and Sadness ($r = 0.710$, $p < 0.001$). Fear had a moderately strong correlation with high statistical significance ($r = 0.669$, $p < 0.001$). Surprise was the single emotion showing weak correlation with no statistical significance ($r = 0.092$, $p = 616$).

(a) **Positive Recordings (NLP)**

Emotion	r	p	Sign.
Anger	–0.199	0.477	No
Joy	0.622	0.013	Yes
Sadness	0.363	0.183	No
Fear	0.527	0.043	Yes
Surprise	0.011	0.969	No

(b) **Negative Recordings (NLP)**

Emotion	r	p	Sign.
Anger	0.286	0.266	No
Joy	0.366	0.149	No
Sadness	0.429	0.086	No
Fear	0.599	0.011	Yes
Surprise	–0.146	0.575	No

Table 4.25: Pearson's r, p-values, and significance for NLP Cloud vs. self (positive and negative)

Table 4.25a presents correlation data between self-reports and NLP Cloud for positive interviews, where lower alignments between self-reports and NLP is found compared to the full dataset. Only correlations for Joy ($r = 0.622$, $p = 0.013$) and Fear ($r = 0.527$, $p = 0.043$) are statistically significant. Sadness had a moderate correlation without statistical significance, Anger and Surprise presented weak correlations.

Correlation coefficients for negative interviews are presented in Table 4.25b, with similar results as for the positive interviews with weaker correlations compared to the full dataset. The single strong correlation with statistical significance is Fear ($r = 0.599$, $p = 0.011$). Moderate correlation is found for Joy ($r = 0.366$ $p = 0.149$) and Sadness ($r = 0.429$, $p = 0.086$), both with no statistical significance. As for the positive recordings, both Anger and Surprise had weak correlations between NLP and self-reporting.

Visual Correlation

Figure 4.11 illustrates the correlation between self-reported Anger scores and AI-labelled predictions for positive oriented interviews, while Figure 4.12 illustrates the correlated data for negative interviews. As shown, Hume shows a moderate positive correlation with no statistical

significance (r = 0.40, p = 0.136) where the data points have some spreading around the trend line. NLP Cloud shows a weaker correlation with self-reports for anger (r = -0.20, p = 0.477) than Hume in positive recordings, as demonstrated in the Figure 4.11 for NLP vs Self where data points are spread out vertically in line with the 0.0 axis.

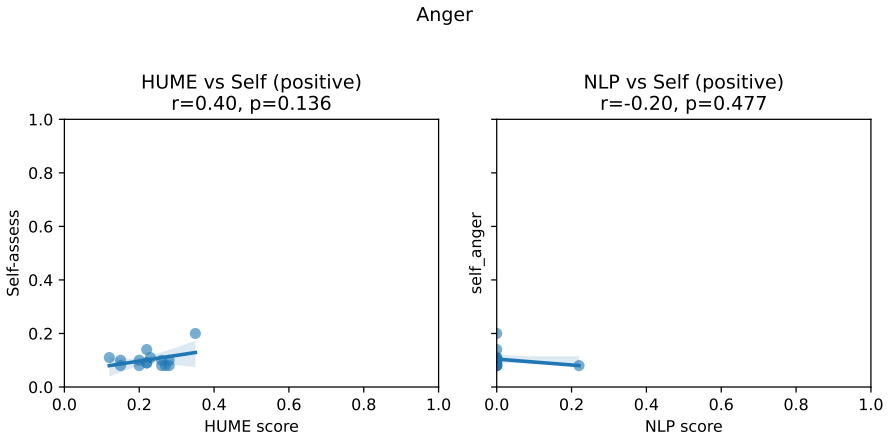


Figure 4.11: Scatter plot, Hume, NLP vs. Self for Anger.

The correlation coefficients remain low in Figure 4.12, where NLP presents a moderate positive correlation (r = 0.29, p = 0.266) with self-assessed anger in negative contexts, while the relationship with Hume is weaker (r = -0.10, p = 0.690) than in the positive interviews. The dispersed data points around the trend line visualises the divergence between the AI-systems and participants own judgement.

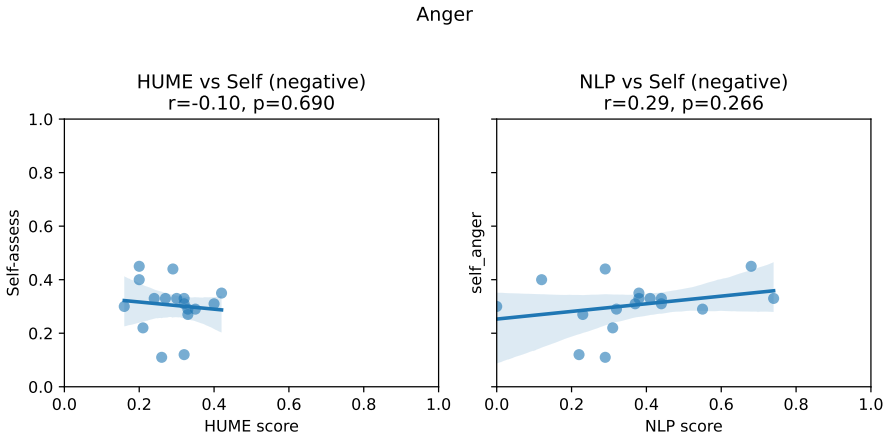


Figure 4.12: Scatter plot, Hume, NLP vs. Self for Anger.

Correlations for Joy is presented further in Figure 4.13 for positive recordings. Both Hume and NLP show a moderate to strong correlation with self-reported joy, NLP with the strongest correlation with statistical significance (r = 0.62, p = 0.013). This relationship is clearly presented with the data points being relatively close to the trend line for NLP, while the Hume diagram has more dispersed data points.

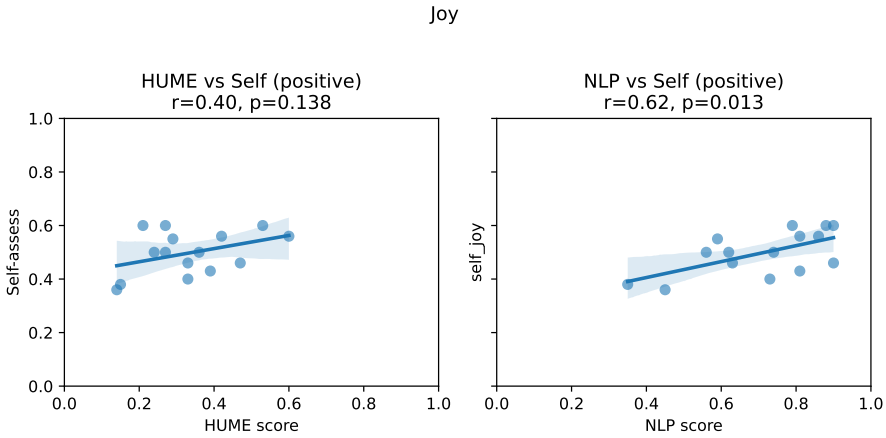


Figure 4.13: Scatter plot, Hume, NLP vs. Self for Joy.

Agreement between self-reported and AI-predicted joy is demonstrated for negative interviews in Figure 4.14. The trend where NLP has a higher correlation (r = 0.37, p = 0.149) remain, however the moderate relationship has no statistical significance. Hume shows a weak correlation, in contrast with the positive oriented interviews. Data points are more widespread for both Hume and NLP correlation with self-reported joy, suggesting varying rating of this emotion in negative contexts.

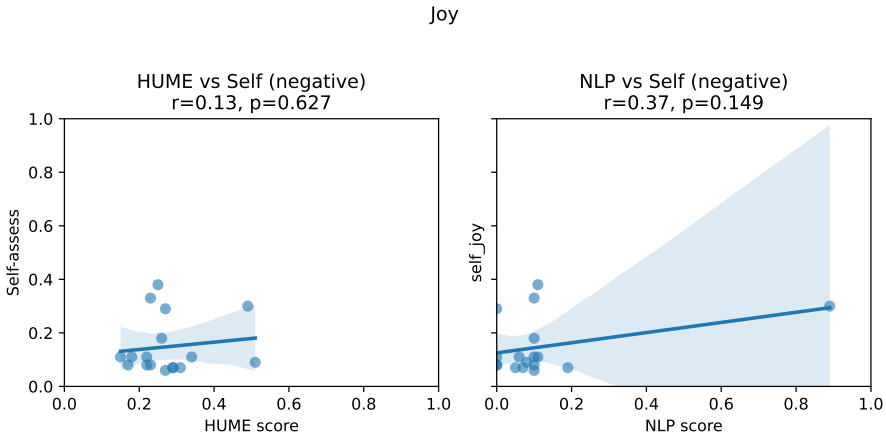


Figure 4.14: Scatter plot, Hume, NLP vs. Self for Joy.

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 for each sentiment. To evaluate the effect size of these differences, Cohen's d were calculated.

Hume vs Self-Reports

Table 4.26 presents the results on paired t-tests with Cohen's d to compare Hume AI's speech-based emotion scores to participants reports across the full dataset. As shown, Hume AI ratings on anger are higher than self-reports (t = 2.399, p = 0.023, d = 0.424), suggesting a moderate tendency for Hume overestimating anger compared to participants own perception. In contrast, Hume underestimate Surprise relatively to self-reports (t = -2.109, p = 0.043, t = -0.373). No significant differences are found for sadness, fear, and surprise ($p > 0.05$, $|d| < 0.20$).

All Recodings

Emotion	t-statistic	p-value	Significant	Cohen's d
Anger	2.399	0.023	Yes	0.424
Joy	–0.271	0.788	No	–0.048
Sadness	–1.069	0.293	No	–0.189
Fear	1.052	0.301	No	0.186
Surprise	–2.109	0.043	Yes	–0.373

Table 4.26: Hume AI vs. Self—All Recordings (paired t-test & Cohen's d)

Table 4.27 seperates these comparisons by sentiment. In positive interviews, four of five emotion comparisons show significant differences. Hume tends to remarkably overestimate anger (t = 8.776, p < 0.001, d = 2.266) and underestimate joy (t = -5.112, p < 0.001, d = -1.320), and assigning notable higher scores for sadness and fear compared to participants own evaluation. Surprise is the single emotion that remains unsignificant. In negative interviews, significant differences appears for joy (t = 3.878, p = 0.001, d = 0.941) where Hume predicts higher levels than self-reported scores, and for sadness (t = -2.890, p = 0.013, d = -0.677) that is rated higher by participants than Hume. Other emotions show no reliable difference.

(a) Positive Recordings

Emotion	t	p	Sign.	d
Anger	8.776	0.000	Yes	2.266
Joy	–5.112	0.000	Yes	–1.320
Sadness	2.451	0.028	Yes	0.633
Fear	2.463	0.027	Yes	0.636
Surprise	–1.855	0.085	No	–0.479

(b) Negative Recordings

Emotion	t	p	Sign.	d
Anger	–0.517	0.612	No	–0.125
Joy	3.878	0.001	Yes	0.941
Sadness	–2.790	0.013	Yes	–0.677
Fear	–0.454	0.656	No	–0.110
Surprise	–1.033	0.317	No	–0.250

Table 4.27: Paired t-test and Cohen's d for Hume AI vs. Self in positive and negative interviews.

These results suggests that Hume AI's speech-based assessments only have weak agreements with participant's self-reports, with varying alignment depending on emotion and sentiment context. In positive interviews, the model remarkably over- or underestimates anger and joy compared to self-reported emotions, while in negative interviews the only significant differences occur for joy and sadness.

NLP Cloud vs Self-Reports

Paired t-tests with Cohen's d for comparison of NLP Cloud's text-based emotion scores and participants' self-reports for all recordings are presented in Table 4.28. Significant differences are found for joy, where NLP rates it higher than self-reports (t = 2.331, p = 0.026, d = 0.412). In contrast, NLP tends to underestimate fear comparing to self-reports (t = -3.496, p = 0.001, d = 0.618). Anger, sadness, and surprise show now significant difference (p > 0.05, |d| < 0.20).

All Recodings

Emotion	t-statistic	p-value	Significant	Cohen's d
Anger	–0.373	0.711	No	–0.066
Joy	2.331	0.026	Yes	0.412
Sadness	–0.525	0.603	No	–0.093
Fear	–3.496	0.001	Yes	–0.618
Surprise	–1.011	0.320	No	–0.179

Table 4.28: Paired t-tests and Cohen's d for NLP Cloud vs. Self. All Recordings

Table 4.29 seperates the comparisons by sentiment. In positive interviews, NLP rates anger at significantly lower levels than participants (t = -4.853, $p < 0.001$, d = -1.253), while rating joy higher than self-reports (t = 6.066, $p < 0.001$, d = 1.566). Both sadness and fear have a significant difference where NLP tends to underestimate these emotions compared to self-reports. Surprise remains without significance. In negative oriented recordings, no significant difference between NLP and self-reports are found ($p > 0.05$, $|d| < 0.389$), suggesting closer alignment during negative contexts.

(a) Positive Recordings

Emotion	t	p	Sign.	d
Anger	–4.853	0.000	Yes	–1.253
Joy	6.066	0.000	Yes	1.566
Sadness	–2.852	0.013	Yes	–0.736
Fear	–4.603	0.000	Yes	–1.188
Surprise	–0.075	0.941	No	–0.019

(b) Negative Recordings

Emotion	t	p	Sign.	d
Anger	1.335	0.200	No	0.324
Joy	–0.578	0.571	No	–0.140
Sadness	1.073	0.299	No	0.260
Fear	–1.149	0.267	No	–0.279
Surprise	–1.604	0.128	No	–0.389

Table 4.29: Paired t-test and Cohen's d for NLP Cloud vs. Self in positive and negative interviews.

Overall, these results implies that NLP Cloud have higher agreement with self-reports in negative contexts, but in positive interviews notable divergencies are found for certain emotions, most appearing for anger, joy and fear.

4.4.4 Conclusion of RQ3 Data Analysis

The result for the full dataset shows that Hume AI only has two small significant divergencies from self-reports, where anger is overestimated (d = 0.42) and surprise underestimated (d = -0.37) compared to ratings by the participants, while all other emotions show no significance in t-tests and weak correlations. When separated by sentiment, Hume tends to overestimate anger and underestimate joy relatively to self-reports in positive contexts, in negative interviews it diverges on joy and sadness. In contrast, NLP Cloud are closely aligned with self-reports in negative contexts, with no significant differences, but in positive interviews NLP remarkedly underestimates anger (d = -1.25), rates fear at lower levels, and joy at higher levels compared to self-reports. compared to self-reports. Correlation analyses strengthen these patterns, where NLP correlations with self-assessed scores are strong for anger, joy, sadness, and fear in the full set, and Hume only shows a moderate correlation for anger and weaker correlations for other emotions. However, the correlations are not as strong when separating the interviews by sentiment. Overall, text-based emotion detection with NLP Cloud has higher agreement with

participants' self-assessments, especially in negative interviews, while speech-based detection with Hume have more variations between positive and negative contexts. These results shows that each modality captures distinct features when comparing with human-labelled rating of their own emotions, and the alignment is fluctuating depending on the interview sentiment.

Discussion

5.1 Result Discussion RQ1

For the first research question, this thesis investigated how an AI model 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 et al. (2023).

The attempts to categorise emotions based the results on this Swedish study cannot be directly evaluated in terms of how accurate the function yielded emotion labels. Even if the standardised function had better recall against Hume, the divergences between the scores were minimal. The rule-based version got higher recall in addition of feature-based adjusting compared to Hume, which is not a ground truth, still beneficial for relative comparison. The improvement is aligned to prior research (Banse & Scherer, 1996; Ekberg et al., 2023), minimising the extreme values for joy according to the wide acoustic markers while increasing the weight of pitch, which is stated to be the most prominent feature perceptually. Restricting anger to be required two of its studied characterises lead to slightly fewer recordings being top labelled as anger, which was added by the reason to separate anger from happiness. The final version presented some alignment between the top emotions, negative recordings yielded higher correlation where both models rated anger highest for six clips and sadness for two clips. More confusion occurred for the positive recordings. The interview setting is likely a contributor to this, where positive oriented questions can be answered with the same tone as negative evoked answers. Fear and surprise were rarely labelled by both methods, Hume distribution is presented in more detail for RQ2. However, even if the results yielded higher recall against Hume, it cannot be benchmarked that these alternations resulted in exactly the emotions that were expressed – a complex question to answer no matter what it is compared to, due to the nature of abstract perception of expressed emotional state, both internally and externally. When formulating the rule-based categorisation function, subjective opinions have an impact about fear and surprise rarely being expressed in the interviews overall, by the reason that the interviews were not designed to evoke these emotions which is complicated to recall in interview circumstances, especially for surprise.

Comparison of vocal features against Hume's emotions probability and our categorisation labelling demonstrates that the methods label emotions based on different data. Only sadness shows a similar vocal pattern between the sources, both negatively correlated with pitch and HNR, for other features, not as strong correlation similarities occur. No other emotion shows similarities, presumably by the reason that Hume has a much more advanced approach to recognising emotions than only using a few vocal features as indicators, as explained in Theoretical Framework 2.5. The fact that average values for a whole recording was used for these analyses, are most likely impacting the results due to low-expressed vs high-expressed segments even out. The subjective possible impact is a clear limitation as well as utilizing the Swedish research as a foundation for this categorisation since it is based on a small, acted dataset whereas this study analyses semi-structured interviews with spontaneous speech in conversational interview format. Limitations for the categorisation function also involves alternations that, even if motivated by prior research, might not yielding fully accurate results. Therefore, the remaining

data analysis include Hume's probability and vocal features, not categorised manually.

To answer the research question, Time-to-Time analysis reveal what occurred for sadness in the comparison between vocal features and both emotion categorisation methods. Highest correlation values together with shifts for high-emotion occurrence with certain features resulted in sadness being most prominent predicted when pitch and HNR was lower, but also when intensity decreased. All three relationships are aligned to the Swedish research, where these acoustic markers have lower mean than other emotions. Fear was the second emotion with higher correlations, for example being predicted when pitch and HNR was raised, also according to the utilized research results. Joy had the highest intensity shift, that is stated in the research. However, anger had lower results for intensity which contrasts with expectations. When interpreting these results, it is important to acknowledge that the Swedish research do not specify how strongly the emotions are expressed, it is hard to define if anger is perceived as screaming or as a negative tone. Again, the nature of emotions is complex, and it is not possible to compose if the results are good or bad, which was not the purpose of this study. Conducting time segmented analyses generated additional insights that reduced the limitation with using mean values from the recordings, revealed by observing feature shifts.

Despite that less than half of the feature-emotion group was significant, with most correlations being weak, certain patterns aligned with theoretical expectations was disclosed. The case examples provide a visualisation on both how certain vocal features align with Hume-predicted emotions, presenting similar patterns as prior research. It also shows how the fixed time-segments can have a negative impact on the correlation values, since Hume have clip varying time frames, the segments are not fully aligned. While correlations between AI-labelling and vocal markers are generally weak to moderate, it is important to acknowledge that Hume emotion scores should not be perceived as perfect estimates of true emotional states, even if it probably results in reasonable predictions. These analysed recordings consist of spontaneous and conversational speech which likely do not involve as strong emotional expressions as acted datasets, with tendency of more subtle and reduced level of vocal features. The simplicity regarding the number of applied vocal markers in contrast to the emotion-trained AI model reaching beyond the use of single acoustic features is a certain reason that this study should be interpreted as explanatory and not to benchmark the general efficiency of emotion recognition in Swedish speech.

The wide acoustic spread for certain emotions, as anger and joy, presented in both (Banse & Scherer, 1996; Ekberg et al., 2023) and was prominent when adjusting the rule-based emotion categorisation where alternations had to be included to separate them from each other. This imply that vocal-marker theory can be limited in conversational context and the intention of the speaker, such as sarcasm and genuine versus polite emotions.

Certain vocal patterns do in fact recognize expressed emotions, as we as humans can interpret certain emotions in others through their speech, algorithms and technology can do the same, at least to some extent. Even if some Hume-labelled emotions align with research on acoustic markers, more than these features are utilized to recognise emotions with AI. The fact that some alignment arises between the Swedish research and Hume AI, suggests that the language does not necessarily have an overly negative impact, yet our results are not comprehensive enough to confirm this.

5.2 Result Discussion RQ2

For the second research question in this thesis, the aim was to investigate the similarities and differences between the two AI models Hume AI, a speech-based model, and NLP Cloud, a text-based model. This was measured when labelling five different emotions in semi-structured

interviews.

The decision to shift from the initial sentiment analysis model to the text generation endpoint provided necessary control to target the five emotion categories of this study, explained in 2.4.1. The generative model relies on prompt phrasing, this may have introduced variability in the results, not only by the manual prompting but also because of utilizing a zero-shot prompt without thorough testing or evaluation of the outputs. Future studies may benefit from a more systematic testing of prompts and the stability through different analyses on the same input.

Using both descriptive statistics and visual analyses to calculate the differences, the system overall seemed to show some levels of agreement for certain emotions. In figure 4.8, the results of the entire dataset divided between the positive and negative recordings were presented and the results highlighted some key differences in the interpretation of the emotional content between NLP Cloud and Hume AI. NLP Cloud appeared to better capture the contextual nuances, as scores for the positive clips had higher scores for joy and surprise, while negative recordings had a higher score for anger and sadness. Hume on the other hand, showed potentially misleading results. A significantly higher emotion score for anger and a somewhat higher score for sadness and fear in the positive recordings, while the score for joy appeared significantly higher for the negative recordings. This may be due to some signals being misinterpreted when speaking of both positive and negative topics. For example, some participants might have talked about the negative topics in an ironic tone, with sarcasm, or expressed nervous laughter. This could be hard for a speech-based model to differentiate and pick up on without the textual context. Another possible explanation may be due to pitch variations, as earlier research found that prosodic features like pitch are informative for arousal detection (Soleymani et al., 2017). Pitch is one of the key features for emotion recognition in audio and it is possible participants of the interviews might have spoken with a high pitch even if they were not using a very descriptive language. Hume AI could have interpreted a high pitch as emotional intensity, possibly explaining the high score of joy in the negative recordings and the high score of anger in the positive recordings.

Joy being highly scored by NLP Cloud in the positive recordings indicates that joy may have been easier identified in textual context than in speech-based emotion analysis, as the textual context may have conveyed a more positive tone from the text than what appeared in the voice. This is further backed up by earlier research stating specific words like “amazing” holds more intensity for emotions than other words like “leaves” (Chauhan et al., 2024). The participants in the interviews may have used very descriptive language when describing positive experiences, which may be a reason for the high score for joy by NLP Cloud. In contrast, as anger and fear appeared to have been more consistently captured by the speech-based emotion detection in the positive clips, this possibly suggests 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 in table 4.16, showing the association between the text-based and speech-based emotion recognition models for all recordings, the strongest alignments were shown for joy and anger. As 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. These emotions being more complex and seemingly more difficult to detect, may have contributed to the lower consistency across the two models for these specific emotions. Research have found that acted speech involves a stronger degree of intensity than spontaneous speech, which makes it more difficult to recognize emotions in this format (Chakraborty et al., 2016). As spontaneous speech from interviews was used in this study, it is possible the models failed to detect the low-intensity emotions to an extent.

For the positive recordings only, there were two emotion showing significant correlations

between the two models, joy and sadness, whereas for the negative recordings the only emotion that showed a significant correlation was joy. These results suggests that joy is the easiest emotion to detect for the models regardless of the context, and that anger, fear and surprise may be too complex to detect, although there are several possible reasons as to why they have low or inconsistent correlations in addition to the difficulties of emotion detection in spontaneous speech. The way that the interviews are set up and the fact that the participants of the interviews talk about situations and feelings they have lived through in the past, may have resulted in emotions being expressed in a subdued way. Talking about a time where you felt fear or surprise might not be translated as strongly when time has passed, as it would in the moment when the emotions were felt. It is possible the interviewees did not feel or express strong emotions when speaking about different situations, and as neutral emotions have been found harder to detect according to earlier research (Cao et al., 2015), it is possible some emotions have remained undetected or incorrectly detected. With this in mind, the lack of correlations for more complex emotions suggest that the interview format may have been insufficient to draw out more nuanced emotional responses. Alternatively, the AI models used may have some limitations in detecting subtle 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 t-test indicated that while the systems do not align on detecting patterns for fear, Hume AI consistently rates 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.

T-test for the positively oriented interviews in comparison to the negative oriented interviews revealed notable findings. For the positive interviews, significant differences between the models were found for all emotions with the exception of surprise, while NLP on the other hand overestimated joy significantly. 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 voice could reveal more subtle cues in the tone, rhythm and pitch. For the positive interviews, the participants may have had a lower and more neutral tone and pitch than an actor acting out happiness, which could be one explanation for this result, which also can be explained by earlier research stating neutrality makes detection of emotions more difficult (Cao et al., 2015).

The T-tests for the negatively oriented interviews showed that 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 misalignment for the two models for the analyses made for the negatively oriented interviews. A possible explanation for this is the participants of the interviews using an overly positive tone of voice out of politeness due to the interview setting, even if the content of the words may have been negative. This could explain why Hume AI detected joy in interviews with a negative theme.

5.3 Result Discussion RQ3

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

In examining the alignment with the speech-based emotion labels from Hume AI and the text-based emotion labels from NLP Cloud with the self-assessed emotion scores, insightful findings revealed some levels of alignment dependent on both the model and emotion.

Figure 4.10 builds on figure 4.8, discussed for RQ2. Presenting mean emotion scores for Hume AI and NLP Cloud, figure 4.10 also introduced the self-assessed emotion scores. Key differences were found between the models in comparison to the scores obtained from the interviewees, as NLP Cloud rated joy much higher than both the other modalities. As earlier discussed in section 5.2, this could still be explained by the spontaneous interview format, as it may not encourage expressively conveying emotions, leading to the model interpreting textual language as joyful even though the tone may have been more neutral. Suprise presented almost identical values for the self-assessed scores in comparison to NLP Cloud, although for all other emotions (anger, sadness and fear) the participants of the interviews rated their emotions almost an average between the two models, with the scores not being quite as high as Hume AI, and not as low as NLP Cloud. In the case of Hume AI estimating anger substantially higher than NLP Cloud and somewhat higher than the interviewees themselves, there might have been misinterpreted signs of anger coming from pitch and intensity during the interviews. These results suggest that the only emotion detected somewhat similarly here compared to the self-assessed scores was surprise, whereas the self-scores landed on a middle ground between the two models for all other emotions in the positive oriented interviews. Higher ratings of anger by NLP Cloud compared to the two other sources was likely due to the context of negative wording in the negative interviews forwarding the emotions more than was both felt by the participants and detected trough the voice. NLP Cloud scored substantially closer scores to the self-assessed emotion scores for joy and sadness compared to Hume AI, suggesting the context of the words matched the emotions in the interviews better which indicates that the text-based model might have been better at capturing emotions from the vocal recordings. Hume may not have picked up vocal cues in the same capacity, likely due to the low expressions and more neutral speech during the interviews. This further reflects a limitation in emotion recognition from spontaneous speech, which also has been highlighted in earlier research. Research done by Cao et al. (2015), found even advanced ranking-based classifiers which had outperformed traditional models, to struggle with neutrality in spontaneous speech (Cao et al., 2015). Further analysis for the negative clips showed Hume AI rated the emotion joy extremely high in comparison to the two other modalities. This further confirms what was earlier explained in the discussion for RQ2 5.2, that Hume AI may have incorrectly interpreted certain vocal cues as joy, for example nervous laughter or sarcasm, which could be difficult for a speech-based AI to recognize.

Overall for the sentiment-based comparisons, the negative recordings showed a better alignment for all three modalities. For both the positive and negatively oriented recordings, the two models performed better for some emotions as the text-based AI NLP Cloud seemingly captured the context for each interview more consistently for some emotions than the speech-based model and vice versa. This underscores the limitation of relying exclusively on either speech-based or text-based emotion recognition, as the different models capture different emotions with varying degrees of accuracy. Using both models in comparison to the self-assessed scores gives a wider understanding of the performance for the text-based versus speech-based emotion recognition model and their different strengths and weaknesses. This aligns with earlier research which also have concluded that using more than one approach results in a better performance than only relying on an individual source (Cao et al., 2015).

When examining the positive recordings and the negative recordings individually for the correlation analysis for Hume AI, no significant correlations were found for any emotions, although the positive recordings showed moderate positive trends for anger and joy. These emotions did not reach any significant correlations, but moderate r-values suggests a possible relationship that may be of interest to explore in the future with a lager dataset or different methods, as the lack of statistically significant correlations indicated that the emotions captured in the interviews did not align closely with those in Hume AI. For the correlation analyses for NLP Cloud, surprise was poorly detected with low correlations, suggesting the model struggled

with the interpretation of the emotion surprise, possibly as a consequence of the complexity of the emotion and once again the nature of the spontaneous interview format.

While NLP Cloud showed stronger correlations with the self-reported emotions overall compared to the Hume AI model, results showed that the model performed inconsistently across the different contexts (negative and positive), suggesting that a more consistent recognition of emotions may demand more modalities for better accuracy. These results also point to the fact that surprise remains a complex emotion with more challenges to capture from the data used in this study, although there are additional reasons to this challenge. In 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 it's unlikely that the interviewees are able to genuinely experience the same amount of 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 possibly could have presented higher correlations for this emotion. As stated in earlier research, acted speech is an amplification of emotions and spontaneous speech may lack the level of intensity to be distinguished from different emotions (Chakraborty et al., 2016).

The statistical analysis and effect sizes showed to be consistent with earlier findings, further giving grounds to this discussion. The full dataset showed Hume to have a moderate tendency for overestimation of the emotion anger in comparison to what the participants of the interviews had reported themselves, which remains true for the positive recordings where anger was very highly detected by Hume. As Hume also tends to underestimate surprise compared to the self-assessed scores, it is further confirmed that surprise is a difficult and complex emotion to detect. This validates conclusions from earlier research stating that neutrality is difficult for a model to deal with (Cao et al., 2015). It is possible the neutrality of the speech often coming across in a spontaneous interview format may have been one of the reasons as to why emotions like surprise were detected at low levels. Hume AI may, as a speech-based emotion recognition model, not be capable of detecting subtle emotions fully and appears to struggle without the textual context as some vocal cues seems to have been misinterpreted. No significant differences were found for NLP Cloud in the negative context, which suggests a closer alignment for NLP Cloud and the self-assessed scores in the negative contexts. In the cases of where the alignment for the models and self-assessed scores did not align as well, further explanations can be drawn from earlier research stating sentiment do not inevitably display themselves in expressions or behaviors (Soleymani et al., 2017). In many cases both Hume AI or NLP Cloud overestimated or underestimated scores compared to the self-assessed values. This could likely be due to vocal expressions not always aligning with the internal affective states, as sentiment is not always fully articulated.

5.4 Method Discussion

The methodology to collect data through interviewing people with the intention to evoke emotions to analyse the recorded data have limitations, even if beneficial to the purpose of the study where real-world speech should be analysed in terms of expressed emotions. The interviews were designed to provoke one positive and one negative emotion, yet not directly oriented towards one of the three negatively oriented emotions. This structure was motivated by allowing the participants to talk freely about a subject they related to and felt comfortable to discuss. During the self-report after each interview, several participants raised confusion about how to interpret and rate surprise. This emotion showed most inconsistency throughout all research

questions and was not aimed to be induced through the interviews. By reducing this study to only focus on one positive and one negative emotion, both for interview design and in data analysis, a more comprehensive analyse of only these emotions could be conducted to yield a narrower, yet deeper analysis of two focus emotions.

Regarding the interviews, its setting cannot be perceived as a clear representation of real conversational speech, due to its reflective nature. The first research question where vocal markers based on previous research on Swedish vocal markers are based on based on pre-defined sentences, repeated by four actors, contrasting our analysed dataset. The spontaneous speech and large variety of interview questions combined with dataset size may indicate some limitations for the result. During the data analysis it was found that pitch and HNR are significantly diverged for male and females, see Appendix Table 7.1, pitch (Hz) differed roughly 100 Hz, HNR mean for females was ≈ 12 while male ≈ 1. This has without doubt impacted the results for all analyses conducted on the full dataset, by the reason that these features even out by the other gender and may create a normalised average even if it would be distinct if analysed separately. Additionally, it was nine male and six female participants, probably resulting in higher mean pitch and HNR values than if it would be even gender groups. It would be to advantage to analyse them separately or at least have a more equal diverged gender distribution. During the preprocessing of the recorded audio, loudness was normalised to ensure consistent volume levels across different recordings. This step was made to prevent significant volume differences between interviews that could have impact the analysis. However, at a later stage in this study it was realized that this normalization very likely has unintentionally affected the vocal feature extraction for intensity/loudness, which are both relevant vocal markers for emotion recognition and are a key descriptor for certain emotion expressions, as explained in 2.5 Theoretical Framework. Since intensity have showed weaker correlations in the analysis of this study, the preprocessing presumably has led to weakening the variability of this vocal marker and therefore impacted the results. Due to time constraints, it was not possible to repeat the full analysis with unprocessed recordings. This is a methodological limitation that should be addressed in future research.

Comparing the results with the prior study on Swedish vocal markers indicate some similarities and patterns providing valuable information to this study, even if several correlations and patterns remain weak. Mainly focusing on this study as reference may create bias, by the reason that there is vastly limited prior research on Swedish in this field. Additionally, this study employed a larger number of vocal features than extracted for these results. The three frequency formants were included in both categorisation functions yet excluded in the overall data analysis. Beyond these, the Swedish study included 14 additional features, some of these were not possible to extract with Praat Parselmouth and therefore excluded. Voiced- and unvoiced length were not relevant for this study since the recordings were edited, including deleting some silent moments. Not including the full set of features is a clear limitation for the results. 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. The interviews did not have a clear timeframe, resulting in varying length of the recordings even after editing. This could have been stricter and more planned to maintain consistency across all recordings. The initial idea was to analyse emotions in a clip in its entirety and find correlations, including average vocal markers and Hume probabilities, which due to differentials in expressions during an interview most likely contributed to even-out values. Time-segmenting proved to have a notable strength to execute the analysis on a segment level to capture emotional fluctuations in a more dynamic way. Output from Hume is pre-segmented with frequency variation of 1-4 seconds, single clip dependent. Vocal features were segmented into 2.5 second timeframes used for all clips in the general data analysis, leading to divergencies in segment length. Therefore, we tested how changing the segmented time with 1 second for

selected individual clips, yielding in both higher and lower results, depending on the clip. For more accurate results, it should have been adjusted separately for each clip before segmented analysis to match the time frames extracted from Hume. The methodological approach for the RQ1 was partially fulfilled by identifying some vocal fluctuations in emotions, while also revealing challenges in both segment-level and average values. Further insights could have been provided with utilizing the rule-based functionality for emotion grouping segment wise as well, to find if grouped features could provide higher correlation results for Hume labelling and acoustic markers. Hume AI was one of the models used and provided some advantages such as avoiding manual pre-training, 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.

Comparing our interview-based result with a larger and more controlled dataset with acted emotions could possibly have validated some observed patterns or strengthened the opposite whereas acted speech is expressed explicitly different from speech in real-world similar contexts. This is also relevant for the second research question extending to text-based emotion recognition with NLP Cloud that analysed the transcript from the same recordings as RQ1. However, collecting our own data was beneficial for the third research question where the AI-models output was compared to self-assessed emotion scores. Utilizing NLP Cloud have limitations. For the initial implementation for RQ2 and RQ3, the sentiment analysis API by NLP Cloud was tested. However, this method resulted in a broader range of emotions being returned than those analysed in this study. To address this limitation and maintain consistency across modalities, the approach for the analysis utilized NLP Cloud's text generation endpoint, applying the fine-tuned Llama 3 model. This allowed control of which emotion categories were extracted by a zero-shot prompt. This approach allowed processing without fine-tuning; however, no testing of the prompt was conducted prior to the analysis to discover potential output differences. Therefore, the accuracy and consistency of the NLP model's output may have been affected by the formulation of the prompt. It would have been beneficial to evaluate prompt variations and if one transcription could have varying results if analysed more than once, to verify stability of the emotion classifications and strengthen the reliability of the results.

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 potential limitations and 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. In this context, it could have been beneficial to include a larger number of emotions that are more relevant to interview circumstances. However, most research include around five to six emotions, including the selected for this study.

The multi-method approach combining speech- and text-based emotion recognition with self-reported emotion scores of the participants from the interviews contributed to insights into emotional expression in Swedish speech, despite several limitations. A narrower approach could have been constructive in terms of deeper analysis. It could also have been useful to study a larger dataset and compare speech from conversational interviews, actual real conversations and acted datasets to gain an understanding of how much acted datasets can impact speech recog-

nition overall. 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.

Conclusion

6.1 Summary of Key Findings and Answering Research Questions

This thesis explored AI-based emotion recognition in Swedish, comparing speech-based and text-based analysis as well as self-assessed emotion scores. The research addressed three research questions, the first aimed to investigate how speech-based AI emotion recognition (Hume AI) aligns with previous research on vocal markers in Swedish speech. The results showed some alignments exist between acoustic vocal features and Hume AI outputs, especially for certain emotions where sadness had most prominent correlations with prior research. However, many correlations were weak or moderate and the results showed some limitations, for example. the nature of spontaneous speech involved in interviews presented challenges for some of the emotions, while other emotions proved to be more relevant indicators of emotional states. Segment-level analysis contributed with valuable insights by illustrating fluctuations in emotional expression within recordings, implying that analysing speech over time in smaller segments may capture emotions more effectively than using average values.

For the second research question this study explored differences, similarities and correlations between speech-based emotion recognition through Hume AI and text-based emotion recognition through NLP Cloud. For some emotions more than others, some agreements were found between the models. Joy and anger showed better alignment whereas fear and surprise did not align as much. The overall cause for this most likely being the complex nature of some emotions where the vocal cues may not be as pronounced.

For the third and last research question, the comparisons between the models Hume AI and NLP Cloud in combination with self-assessed emotion scores from the interviews were investigated. These results showed there were some alignments between the models and self-assessed scores, though these alignments were stronger with NLP Cloud than with Hume AI. This gave insight into the fact that although some alignments were stronger with NLP Cloud than others, a multimodal approach that integrates several sources results in better detection of emotions, where only relying on either a speech-based model or a text-based model would not give as much insight in the results. Overall, the findings in this research provided answers to all research questions, contributing to a deeper understanding of emotion recognition in the Swedish language and highlighting that the different modalities capture different aspects of emotional expressions.

6.2 Contribution to the Field

This thesis contributes to the vocal and linguistic aspect of the growing fields affective computing and natural language processing (NLP). This study utilized a multimodal approach, using both text and audio. The results of this suggested that using more than one modality improves the accuracy when classifying emotions. While there is a lot of existing research focusing on vocal emotion recognition in affective computing and NLP, there was a noticeable gap

for research specifically in the Swedish language. Much of the existing research also used acted datasets, which was not used in this study. Instead, this study consisted of semi-structured interviews with a set of questions for the interviewees to choose from. With the spontaneous nature of the datasets from speech recorded from interviews, this offers valuable insight into how emotions can be recognized in a setting more similar to real life. This can also contribute to the development of more emotionally aware AI models and systems through the insights of the more subtle cues for different emotions in these settings. Possible areas where this can be applied is in human-computer interactions, virtual assistance, and mental health monitoring or similar.

6.3 Limitations of the Study

There were several limitations of significance throughout this thesis. The limited size of the dataset and the chosen emotions, as the acoustic features, may have impacted the results. A bigger dataset with more emotions and acoustic features could have given important insight and clearer results if included. Another important limitation to mention is the spontaneous nature of the speech gathered from the interviews. Using an acted dataset may have resulted in emotions being more accurately and stronger identified, whereas the interview setting may have led to some emotions being left undetected as they may have been expressed too subtly. This may be due to conversational speech being more subtle than acted speech, reducing levels of clear emotions and vocal markers. Worth noting is also that the Hume AI emotion scores used for comparisons, should not be considered ground truth, but more as a something to compare against. This study also used self-reported emotion scores as a comparison, which may have given a large variety across the dataset due to the participants own self-perception.

6.4 Future Research

There are several ways to build on this study for future research. Key areas for further advancement are a bigger dataset, a more diverse real-world settings, and a bigger set of emotions and vocal features. Using more vocal features and fine-tuning models for Swedish would possibly improve accuracy. To explore how the results differ between languages, Swedish recordings could be compared to an English dataset. Furthermore, multimodal emotion recognition pipelines combining speech, text, and facial expressions could offer more comprehensive emotion detection. Real-time emotion recognition in real world scenarios and fields as education, healthcare, or mental health monitoring is areas that could benefit from this kind of technology but requires clear ethical considerations.

6.5 Final Conclusion

This thesis investigated AI based emotion recognition in textual and vocal content in the Swedish language. The results showed some similarities with existing research on vocal markers in Swedish language and strongly demonstrated the importance of using more than one modality to improve the accuracy of the emotion recognition, especially for more complex emotions. The findings gathered from the research in this thesis help contribute to the growing fields of affective computing and natural language processing, highlighting opportunities and challenges of emotion recognition technologies in multilingual and spontaneous speech contexts. With technology becoming an increasingly larger part of society, understanding human emotions is

a step towards AI becoming more effective and empathetic, which can help mold the future of education, health care and human-computer interactions.

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

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636. https://doi.org/10.1037/0022-3514.70.3.614

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

Cao, H., Verma, R., & Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer Speech & Language*, *29*, 186–202. https://doi.org/10.1016/j.csl.2014.01.003

Chakraborty, R., Pandharipande, M., & Kopparapu, S. K. (2016). Spontaneous speech emotion recognition using prior knowledge. *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, 2866–2871. https://doi.org/https://doi.org/10.1109/ICPR.2016.7900071

Chauhan, K., Sharma, K. K., & Varma, T. (2024). Multimodal emotion recognition using contextualized audio information and ground transcripts on multiple datasets. *Arabian Journal for Science and Engineering (2011)*, *49*, 11871–11881. https://doi.org/10.1007/s13369-023-08395-3

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

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

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

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin and Review*, *27*, 237–265. https://doi.org/10.3758/s13423-019-01701-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

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

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

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

NLP Cloud. (2025). *Nlp cloud documentation: Text generation*. Retrieved June 15, 2025, from https://docs.nlpcloud.com/#generation

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.

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

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

Qazi, A., Goudar, R. H., Patil, R., Hukkeri, G. S., & Kulkarni, D. (2025). Leveraging bert, distilbert, and tinybert for rumor detection. *IEEE Access*, *13*, 72918–72929. https://doi.org/10.1109/ACCESS.2025.3563301

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. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227–256. https://doi.org/10.1016/S0167-6393-02-00084-5

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

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, *65*, 3–14. https://doi.org/10.1016/j.imavis.2017.08.003

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

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, 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

Appendix

(a) Summary statistics for female speakers

Feature	Mean	Std
mean_pitch_hz	215.168	12.744
mean_pitch_st	6.218	1.012
mean_intensity_db	62.419	3.022
mean_hnr_db	12.068	1.642
jitter_local	0.022	0.003
shimmer_local	0.110	0.022
formant_F1_hz	694.094	353.464
formant_F2_hz	1810.434	647.845
formant_F3_hz	3045.689	464.207

(b) Summary statistics for male speakers

Feature	Mean	Std
mean_pitch_hz	112.665	13.657
mean_pitch_st	–5.072	2.067
mean_intensity_db	63.816	2.550
mean_hnr_db	1.033	1.262
jitter_local	0.027	0.003
shimmer_local	0.138	0.017
formant_F1_hz	688.201	311.007
formant_F2_hz	1916.512	457.306
formant_F3_hz	2990.914	375.011

Table 7.1: Vocal feature means and standard deviations by gender

Table 7.2: Interview prompts for positive and negative scenarios

Negative:

- Think about a situation when something did not go as planned. What was it? What feelings did you experience?
- Is it something about society that makes you upset? What? Can you elaborate?
- Is it something in your everyday life that frustrates you? What? Can you elaborate?
- Think of a time when you lost your patience at something or someone. What was it triggered by?
- Can you describe a time where someone betrayed you or your trust? How did you manage that situation?
- Have you ever felt someone took credit for something you did or disregarded your efforts? How did it feel?
- Can you remember a time when you felt unfairly treated by someone? What happened and how did you react?

Positive:

- Describe a situation when something unexpected raised immense happiness in you. How did you feel in that moment?
- Think about a moment when you were very proud of yourself. What did you do and how did you feel?
- Think about a moment when you had an amazing experience with close friends or family. What happened and how did you feel at that time?
- Can you remember a time when you felt completely carefree and happy? What did you do and who were you with?
- Think of a moment when you got a compliment or acknowledgement that made you feel really good about yourself. How did it affect your mood?

```
BENCHMARKS = {
    "anger":    [("hnr","below"), ("jit","below"), ("loud","above")],
    "joy":      [("pitch","above"), ("hnr","above"), ("loud","above")],
    "sadness":  [("pitch","below"), ("hnr","below"), ("loud","below")],
    "fear":     [("hnr","above"), ("jit","below")],
    "surprise": [("jit","above"), ("shim","above")],
}

FEATURE_WEIGHTS = {
    "pitch": 1.3, "loud": 0.5, "hnr": 1.0,
    "jit":   1.0, "shim": 1.0,
    "f1":    1.0, "f2":   1.0, "f3":  1.0,
}
```

```
SD_KEY = {
    "pitch":"sd",    "loud":"sd_l", "hnr":"sd_h",
    "jit":"sd_j",    "shim":"sd_s",
    "f1":"sd_f1",    "f2":"sd_f2",  "f3":"sd_f3",
}

def categorise_emotion_all_scores(
    vf: dict,
    K_NEAR: float = 1.25,
    k_extreme: float = 1.6,
    K_EXTREME_PER_EMO: Dict[str, float] = None,
    use_bm_gate: bool = False,
) -> List[Tuple[str, float]]:
    # Extract features
    pitch    = vf.get("mean_pitch_st")
    loud     = vf.get("mean_intensity_db")
    hnr      = vf.get("mean_hnr_db")
    jitter   = vf.get("jitter_local")
    shimmer  = vf.get("shimmer_local")
    formants = vf.get("formants_hz", {})
    f1, f2, f3 = formants.get("F1"), formants.get("F2"), formants.get("F3
       ")

    # Reference means & SDs (Ekberg, 2018)
    M = {
      "anger": {
         "pitch":5.00, "sd":5.39, "loud":7.16, "sd_l":0.66,
         "hnr":2.36,  "sd_h":0.52, "jit":-0.13, "sd_j":0.38,
         "shim":-1.03,"sd_s":0.21, "f1":0.78,   "sd_f1":0.34,
         "f2":1.20,  "sd_f2":0.35, "f3":0.80,   "sd_f3":0.34
      },
      "fear": { … },
      "joy": { … },
      "sadness": { … },
      "surprise": { … }
    }

    k_ext_per_emo = K_EXTREME_PER_EMO or {}
    default_k     = k_extreme
    bm_gate       = use_bm_gate

    def near(val, mean, sd, k=K_NEAR):
        return val is not None and abs(val-mean) <= k*sd

    def extreme(val, mean, sd, dir, emo):
        if val is None: return False
        k = k_ext_per_emo.get(emo, default_k)
        if dir=="above":
            return val > mean + k*sd
        else:
            return val < mean - k*sd
```

```
    def feature_val(key):
        return {
          "pitch": pitch, "loud": loud, "hnr": hnr,
          "jit": jitter, "shim": shimmer,
          "f1": f1 and f1/1000, "f2": f2 and f2/1000, "f3": f3 and f3
             /1000
        }[key]

    cue_counts = {emo: 0.0 for emo in M}
    for emo, stats in M.items():
        hits = 0
        for feat_key, dir in BENCHMARKS[emo]:
            v = feature_val(feat_key)
            if extreme(v, stats[feat_key], stats[SD_KEY[feat_key]], dir,
               emo):
                hits += 1

        # Apply benchmark gate logic
        if bm_gate and emo=="anger" and hits>=2:
            cue_counts[emo] += hits
        elif bm_gate and emo in ("sadness","joy") and hits>=1:
            cue_counts[emo] += hits
        else:
            cue_counts[emo] += hits

        # Add ""near hits feature weighted
        for val, key in [
            (pitch,"pitch"), (loud,"loud"), (hnr,"hnr"),
            (jitter,"jit"), (shimmer,"shim"),
            (f1 and f1/1000,"f1"), (f2 and f2/1000,"f2"), (f3 and f3
               /1000,"f3")
        ]:
            if near(val, stats[key], stats[SD_KEY[key]]):
                cue_counts[emo] += FEATURE_WEIGHTS[key]

    prob = categorize_emotion_table(vf)
    combined = {e: cue_counts[e] + 0.3*prob.get(e,0.0) for e in
       cue_counts}
    return sorted(combined.items(), key=lambda kv: kv[1], reverse=True)
```

Listing 7.1: Emotion categorization code

Positive Recordings

Emotion	Pearson's r	p-value	Sign.
Anger	−0.320	0.2451	No
Joy	−0.004	0.9876	No
Sadness	0.265	0.3391	No
Fear	−0.022	0.9371	No
Surprise	−0.054	0.8476	No

Negative Recordings

Emotion	Pearson's r	p-value	Sign.
Anger	−0.157	0.5767	No
Joy	0.097	0.7319	No
Sadness	0.222	0.4268	No
Fear	0.216	0.4398	No
Surprise	−0.212	0.6081	No

Table 7.3: Pearson's r, p-values, and significance for positive vs. negative recordings.

Positive recordings		
Feature	**ANOVA P-value**	**Sign.**
Pitch	0.7595	No
Intensity	0.8627	No
HNR	0.6149	No
Jitter	0.9564	No
Shimmer	0.7828	No

(a) ANOVA: Positive recordings.

Negative recordings		
Feature	**ANOVA P-value**	**Sign.**
Pitch	0.5393	No
Intensity	0.1307	No
HNR	0.5142	No
Jitter	0.9066	No
Shimmer	0.6863	No

(b) ANOVA: Negative recordings.

Table 7.4: ANOVA for vocal features across emotions.

Feature	Emotion	Pearson r	p-value	Significant
mean_pitch_hz	anger	0.053	0.1094	No
mean_pitch_hz	joy	0.031	0.3503	No
mean_pitch_hz	sadness	–0.226	0.0000	Yes
mean_pitch_hz	fear	0.087	0.0091	Yes
mean_pitch_hz	surprise	0.060	0.0719	No
mean_intensity_db	anger	0.038	0.2548	No
mean_intensity_db	joy	0.156	0.0000	Yes
mean_intensity_db	sadness	–0.164	0.0000	Yes
mean_intensity_db	fear	–0.136	0.0000	Yes
mean_intensity_db	surprise	–0.094	0.0047	Yes
mean_hnr_db	anger	0.064	0.0557	No
mean_hnr_db	joy	0.061	0.0650	No
mean_hnr_db	sadness	–0.271	0.0000	Yes
mean_hnr_db	fear	0.079	0.0173	Yes
mean_hnr_db	surprise	0.030	0.3677	No
jitter_local	anger	0.033	0.3237	No
jitter_local	joy	–0.033	0.3219	No
jitter_local	sadness	0.040	0.2362	No
jitter_local	fear	0.000	0.9885	No
jitter_local	surprise	–0.042	0.2034	No
shimmer_local	anger	0.001	0.9794	No
shimmer_local	joy	0.057	0.0896	No
shimmer_local	sadness	–0.073	0.0283	Yes
shimmer_local	fear	–0.022	0.5001	No
shimmer_local	surprise	–0.018	0.5823	No

Table 7.5: Pearson correlations for all data in time-to-time analysis.

Feature	Emotion	t-statistic	p-value	Significant
mean_pitch_hz	anger	1.992	0.0467	Yes
mean_pitch_hz	joy	1.438	0.1508	No
mean_pitch_hz	sadness	–6.515	0.0000	Yes
mean_pitch_hz	fear	3.445	0.0006	Yes
mean_pitch_hz	surprise	0.770	0.4417	No
mean_intensity_db	anger	1.209	0.2271	No
mean_intensity_db	joy	4.000	0.0001	Yes
mean_intensity_db	sadness	–4.745	0.0000	Yes
mean_intensity_db	fear	–2.469	0.0138	Yes
mean_intensity_db	surprise	–2.324	0.0204	Yes
mean_hnr_db	anger	3.027	0.0025	Yes
mean_hnr_db	joy	2.619	0.0090	Yes
mean_hnr_db	sadness	–8.105	0.0000	Yes
mean_hnr_db	fear	3.713	0.0002	Yes
mean_hnr_db	surprise	1.263	0.2070	No
jitter_local	anger	0.867	0.3863	No
jitter_local	joy	–0.956	0.3396	No
jitter_local	sadness	1.092	0.2753	No
jitter_local	fear	0.388	0.6980	No
jitter_local	surprise	–1.117	0.2642	No
shimmer_local	anger	1.052	0.2933	No
shimmer_local	joy	0.983	0.3256	No
shimmer_local	sadness	–2.425	0.0155	Yes
shimmer_local	fear	1.125	0.2609	No
shimmer_local	surprise	–0.621	0.5350	No

Table 7.6: T-statistics for all data in Time-to-Time analysis.

T(s)	Clip	Feat	Emo	r	p	Sign
1.0	id_006_pos	mean_pitch_hz	joy	0.188	0.1965	No
1.0	id_006_pos	mean_intensity_db	joy	0.351	0.0134	Yes
1.0	id_006_pos	mean_hnr_db	joy	0.043	0.7674	No
1.0	id_006_pos	jitter_local	joy	0.033	0.8244	No
1.0	id_006_pos	shimmer_local	joy	0.002	0.9888	No
2.0	id_006_pos	mean_pitch_hz	joy	0.078	0.6103	No
2.0	id_006_pos	mean_intensity_db	joy	0.329	0.0272	Yes
2.0	id_006_pos	mean_hnr_db	joy	–0.014	0.9248	No
2.0	id_006_pos	jitter_local	joy	0.063	0.6803	No
2.0	id_006_pos	shimmer_local	joy	0.115	0.4528	No

Table 7.7: Clip 006 pos. Pearson correlation for Joy, 1 and 2s segment

T(s)	Clip	Feat	Emo	t	p	Sign
1.0	id_006_pos	mean_pitch_hz	joy	0.331	0.7419	No
1.0	id_006_pos	mean_intensity_db	joy	2.718	0.0092	Yes
1.0	id_006_pos	mean_hnr_db	joy	0.958	0.3427	No
1.0	id_006_pos	jitter_local	joy	0.062	0.9506	No
1.0	id_006_pos	shimmer_local	joy	0.178	0.8598	No
2.0	id_006_pos	mean_pitch_hz	joy	–0.237	0.8136	No
2.0	id_006_pos	mean_intensity_db	joy	1.162	0.2517	No
2.0	id_006_pos	mean_hnr_db	joy	–0.562	0.5768	No
2.0	id_006_pos	jitter_local	joy	1.210	0.2329	No
2.0	id_006_pos	shimmer_local	joy	1.918	0.0617	No

Table 7.8: Clip 006 pos. T-statistics for Joy, 1 and 2s segment.

T(s)	Clip	Feat	Emo	r	p	Sign
1.25	id_012_neg	mean_pitch_hz	anger	–0.097	0.5496	No
1.25	id_012_neg	mean_intensity_db	anger	0.257	0.1091	No
1.25	id_012_neg	mean_hnr_db	anger	–0.114	0.4846	No
1.25	id_012_neg	jitter_local	anger	0.173	0.2862	No
1.25	id_012_neg	shimmer_local	anger	0.284	0.0762	No
2.25	id_012_neg	mean_pitch_hz	anger	–0.217	0.1839	No
2.25	id_012_neg	mean_intensity_db	anger	0.094	0.5707	No
2.25	id_012_neg	mean_hnr_db	anger	–0.112	0.4955	No
2.25	id_012_neg	jitter_local	anger	0.221	0.1761	No
2.25	id_012_neg	shimmer_local	anger	0.180	0.2717	No

Table 7.9: Clip 012 neg. Pearson correlations for Anger, 1.25 and 2.25 s segments.

T(s)	Clip	Feat	Emo	t	p	Sign
1.25	id_012_neg	mean_pitch_hz	anger	–0.353	0.7259	No
1.25	id_012_neg	mean_intensity_db	anger	2.255	0.0300	Yes
1.25	id_012_neg	mean_hnr_db	anger	0.880	0.3844	No
1.25	id_012_neg	jitter_local	anger	0.223	0.8249	No
1.25	id_012_neg	shimmer_local	anger	1.839	0.0737	No
2.25	id_012_neg	mean_pitch_hz	anger	–1.498	0.1427	No
2.25	id_012_neg	mean_intensity_db	anger	0.705	0.4852	No
2.25	id_012_neg	mean_hnr_db	anger	0.656	0.5156	No
2.25	id_012_neg	jitter_local	anger	1.087	0.2841	No
2.25	id_012_neg	shimmer_local	anger	1.055	0.2982	No

Table 7.10: Clip 012 neg. T-statistics for Anger, 1.25 and 2.25 s segments.

T(s)	Clip	Feature	Emotion	Pearson r	p-value	Significant
1.0	id_006_neg	mean_pitch_hz	sadness	–0.069	0.7073	No
1.0	id_006_neg	mean_intensity_db	sadness	0.210	0.2495	No
1.0	id_006_neg	mean_hnr_db	sadness	0.365	0.0398	Yes
1.0	id_006_neg	jitter_local	sadness	–0.312	0.0816	No
1.0	id_006_neg	shimmer_local	sadness	–0.326	0.0683	No
2.0	id_006_neg	mean_pitch_hz	sadness	–0.333	0.0722	No
2.0	id_006_neg	mean_intensity_db	sadness	–0.213	0.2585	No
2.0	id_006_neg	mean_hnr_db	sadness	–0.265	0.1567	No
2.0	id_006_neg	jitter_local	sadness	0.092	0.6285	No
2.0	id_006_neg	shimmer_local	sadness	–0.050	0.7943	No

Table 7.11: Clip 006 neg. Pearson correlations for Sadness, 1.0 and 2.0 s segment.

T(s)	Clip	Feature	Emotion	t-statistic	p-value	Significant
1.0	id_006_neg	mean_pitch_hz	sadness	0.556	0.5826	No
1.0	id_006_neg	mean_intensity_db	sadness	0.818	0.4196	No
1.0	id_006_neg	mean_hnr_db	sadness	2.565	0.0156	Yes
1.0	id_006_neg	jitter_local	sadness	–1.891	0.0683	No
1.0	id_006_neg	shimmer_local	sadness	–1.663	0.1068	No
2.0	id_006_neg	mean_pitch_hz	sadness	–2.203	0.0360	Yes
2.0	id_006_neg	mean_intensity_db	sadness	–1.307	0.2018	No
2.0	id_006_neg	mean_hnr_db	sadness	–1.289	0.2079	No
2.0	id_006_neg	jitter_local	sadness	1.009	0.3215	No
2.0	id_006_neg	shimmer_local	sadness	0.157	0.8766	No

Table 7.12: Clip 006 neg. T-statistics for Sadness, 1.0 and 2.0 s segment.