

# Semantic Differentiation in Speech Emotion Recognition: Insights from Descriptive and Expressive Speech Roles

Rongchen Guo<sup>1\*</sup>, Vincent Francoeur<sup>2\*</sup>, Isar Nejadgholi<sup>1,3\*</sup>,  
Sylvain Gagnon<sup>2</sup>, Miodrag Bolic<sup>1†</sup>

<sup>1</sup>School of Electrical Engineering and Computer Science, University of Ottawa

<sup>2</sup>School of Psychology, University of Ottawa

<sup>3</sup>National Research Council Canada, Ottawa, Canada

{rongchen.guo, vfran022, sgagnon, Miodrag.Bolic}@uottawa.ca

isar.nejadgholi@nrc-cnrc.gc.ca

## Abstract

Speech Emotion Recognition (SER) is essential for improving human-computer interaction, yet its accuracy remains constrained by the complexity of emotional nuances in speech. In this study, we distinguish between *descriptive semantics*, which represents the contextual content of speech, and *expressive semantics*, which reflects the speaker’s emotional state. After watching emotionally charged movie segments, we recorded audio clips of participants describing their experiences, along with the intended emotion tags for each clip, participants’ self-rated emotional responses, and their valence/arousal scores. Through experiments, we show that descriptive semantics align with intended emotions, while expressive semantics correlate with evoked emotions. Our findings inform SER applications in human-AI interaction and pave the way for more context-aware AI systems.

## 1 Introduction

The ability to accurately detect and interpret emotions in speech is vital for developing intelligent systems capable of natural and empathetic human-computer interactions. Speech Emotion Recognition (SER) has gained significant traction in recent years, driven by applications ranging from virtual assistants to mental health monitoring (Ley et al., 2019; Rumpal et al., 2015). Despite these advancements, SER faces persistent challenges due to the complex and multi-dimensional nature of emotions, which often intertwine with contextual and speaker-specific factors.

Traditional approaches to SER have largely focused on acoustic features, such as pitch, energy, and spectral properties, to infer emotional states (Wu et al., 2011; Bitouk et al., 2010; Venkataramanan and Rajamohan, 2019; Likitha et al., 2017; Kwon

et al., 2003). While effective to some extent, these methods often overlook the semantic content of speech, which can provide crucial contextual information. With the advances in natural language processing, it has become increasingly feasible to analyze the semantic aspects of speech for emotion recognition (Tzirakis et al., 2021; Xu et al., 2021). However, the interplay between semantic roles and emotional expression remains underexplored. Specifically, the distinction between *intended emotions* elicited by a stimulus and *evoked emotions* experienced by the speaker is rarely addressed, leaving a critical gap in the field.

This paper introduces a novel framework to address this gap by distinguishing two types of semantic roles in speech. We hypothesize that *Descriptive semantics* captures scenario-specific content, such as the narrative or context described in the speech. In contrast, *Expressive semantics* reflects the speaker’s subjective emotional stance, shaped by their personal experiences and delivery style. In our framework, descriptive segments are expected to align with the intended emotion of the stimulus (the target emotion the video was designed to elicit), while expressive segments are expected to align with the evoked emotion (the participant’s self-reported experience). This mapping allows us to distinguish stimulus-driven affect from speaker-specific affect, thereby addressing a critical gap in prior SER research that often assumes a single ground-truth label. This semantic distinction is particularly important in settings where it is essential to understand not only *what happened* — the contextual content of speech — but also *how it was felt* — the speaker’s emotional state and tone. Such an understanding has practical implications for applications like emotion-aware AI systems, educational tools, and interactive entertainment, where both the content and emotional delivery of speech play key roles in creating engaging and effective

\* Equal contributions

† Corresponding author

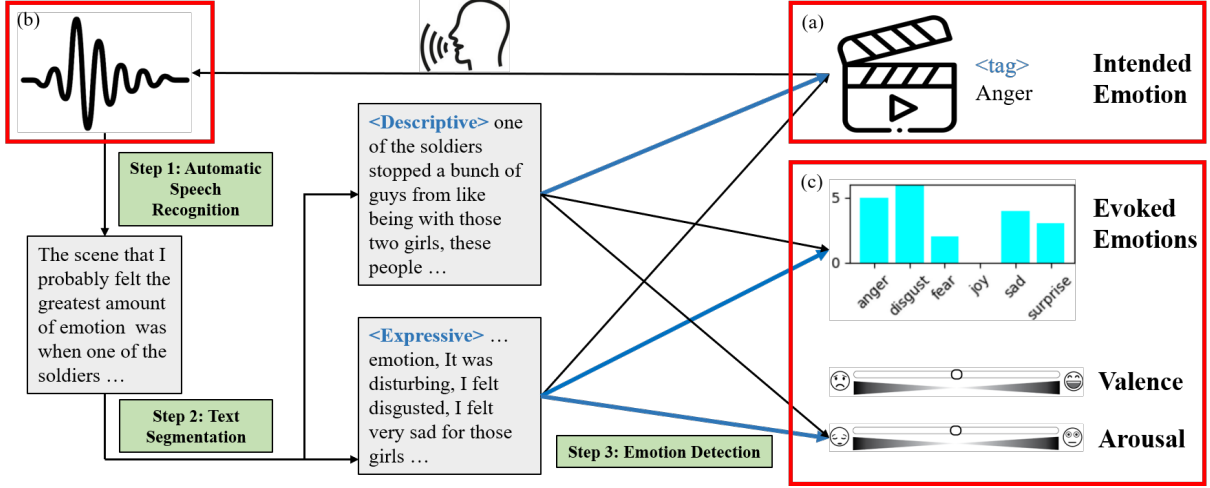


Figure 1: **Data Collection and Algorithm Workflow:** Participants watched six videos eliciting specific emotions and provided speech descriptions, emotion ratings, and valence/arousal scores. Speech data were transcribed, segmented into **descriptive** and **expressive** semantics, and used to train models for three tasks: predicting intended emotions (TASK-1), evoked emotions (TASK-2), and valence/arousal (TASK-3).

human-computer interactions.

To validate our hypothesis, we collected a dataset comprising emotionally evocative movie clips to elicit a specific emotion. Participants watched the videos and provided ratings for the actual evoked emotions, alongside valence and arousal scores, creating a robust foundation for analysis. Our methodology to uncover the distinct relationships between semantic roles and intended versus evoked emotions involves three key steps: speech transcription with automatic speech recognition (ASR), semantic segmentation with LLMs, and emotion prediction with fine-tuned text classifiers/regressors. This work makes the following contributions:

- First, we curated a SER dataset with 582 audio recordings spanning six emotion categories. Audio transcriptions are generated, and intended emotions, as well as evoked emotions, are measured in an experimental setup.
- Second, we implemented an LLM-based semantic segmentation approach to separate the expressive and descriptive parts of speech and validated that through human evaluation.
- Third, through experimentation, we show that descriptive semantics are more predictive of intended emotions, while expressive semantics are better aligned with evoked emotions.

Importantly, our work goes beyond simply predicting emotion labels from participants’ descriptions. By explicitly segmenting speech into descriptive

and expressive roles, we quantify how different semantic functions relate to stimulus-intended versus self-experienced emotions. This role-based separation provides a principled way to reconcile discrepancies between intended and evoked affect and offers interpretable insights that are not available from standard text-only or audio-only models.

Our findings have significant implications for designing more accurate and context-aware emotion recognition systems, with potential applications in virtual assistants, customer service, and mental health support. By bridging the gap between semantics and emotion, this research advances the state-of-the-art in SER and sets the stage for future exploration of semantic roles in emotional AI systems.

## 2 Related Work

Emotions are complex psychological and physiological responses to salient events, involving bodily sensations, expressive behaviors, and cognitive evaluations (Moors, 2024, 2009). Various linguistic features, including prosody, lexical choice, and sentence structure, play a role in the perception and expression of emotions (Mohammad and Turney, 2010; Barrett et al., 2007; Keltner et al., 2019). Speech emotion recognition (SER) models aim to detect emotional states from speech using acoustic, textual, or multimodal signals. With the advancement of LLMs and automatic speech recognitions (ASR), text-based emotion classification has seen improved accuracy (Hama et al.,

2024; Bekmanova et al., 2022; Bharti et al., 2022). Acoustic-based emotion detectors have also progressed using acoustic feature extractors, such as openSMILE (Eyben et al., 2010) or audio embedding models, such as wav2vec (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), which embed paralinguistic cues such as pitch, tempo, and energy into speech representations (Ulgen et al., 2024; Chakhtouna et al., 2024; Zhao et al., 2024; Dutta and Ganapathy, 2024; Ghosh et al., 2016). Multi-modal approaches, which combine speech, facial expressions, and physiological signals, have also become increasingly prominent in recent years (Cheng et al., 2024; Khan et al., 2024; Morency and Baltrušaitis, 2017; Yoon et al., 2018; Niu et al., 2023).

Emotion elicitation via multimedia stimuli (e.g. short film clips) is a common technique in SER to induce targeted emotions (e.g., sadness, joy, fear) (Li et al., 2021; Rumpa et al., 2015). These movie-based emotion elicitation techniques have applications in various fields, including e-health monitoring and human-computer interaction (Ley et al., 2019; Rumpa et al., 2015). The stimuli are selected and validated through self-report and physiological measures (Chen et al., 2021; Handayani et al., 2015; Soleymani et al., 2012). While these methods control for the *intended* emotional target, they do not always account for the *evoked* emotion the speaker experiences and expresses. Prior work such as Siedlecka and Denson (2019) reviewed these paradigms in detail, but focused primarily on affect induction rather than the emotional content of participants’ verbal responses. In our work, we analyze speech collected after stimulus exposure and study how intended and evoked emotions are reflected in participants’ spoken descriptions. In doing so, we explore a novel distinction between semantic roles in language—namely, whether a speaker is being descriptive (e.g., summarizing the movie) or expressive (e.g., conveying their own reaction)—and how these roles align with different emotion types.

Many SER datasets have been developed. In acted speech datasets, such as IEMOCAP (Busso et al., 2008) and SAVEE (Jackson and Haq, 2014), actors are recruited to read sentences or act in scenes that portray different emotions. In spontaneous speech datasets, such as MSP-Podcast (Lotfian and Busso, 2017), MSP-Conversation (Martinez-Lucas

et al., 2020), SAMAIN (McKeown et al., 2011), and RECOLA (Ringeval et al., 2013), and elicited speech datasets, such as LSSD (Fan et al., 2021), BAUM-1 (Zhalehpour et al., 2016), and eENTERFACE (Batliner et al., 2006), audios are recorded in a freely speaking environment or with emotion elicitation methods. Speech is then annotated by a third party (perception-of-other). However, these datasets focus on one emotion label per speech and do not distinguish different types of emotions. To this end, EMO-DB (Burkhardt et al., 2005) and IEMOCAP (Busso et al., 2008) analyzed emotional evocative sentences and perception-of-other in acted speech. Most similar to us, MuSE (Jaiswal and Bara, 2020) collects speech following emotional video stimuli and reports both self-reported and intended emotion annotations. While similar in structure, our work uniquely interprets the relationship between stimulus-intended and self-reported emotions through a *semantic* lens, enabling direct analysis of misalignment between the two emotion types.

Furthermore, some recent studies in NLP have explored emotion elicitation and manipulation in conversational settings (Gong et al., 2023; Ma et al., 2025; Qian et al., 2023; Meng et al., 2024). While our study does not model conversational interactions, our semantic framework may offer insights into these settings by helping to identify when emotional influence is being attempted or received. For example, expressive speech segments may signal internal affective states, while descriptive segments may reflect contextual awareness or narrative framing. These distinctions could inform models of emotion transfer and regulation in human-computer dialogue.

Our contribution lies in bridging the gap between stimulus-based emotion elicitation and the actual emotions conveyed by participants in speech. By segmenting utterances according to their semantic roles and analyzing how different roles align with either intended or evoked emotions, we propose a novel way to interpret emotional speech beyond traditional modality-based or label-based approaches. While prior SER studies have emphasized either acoustic or multimodal representations, our work suggests that semantic structure in language - accessible only through text - offers a distinct and interpretable signal for differentiating between types of emotion.

| Movie Clip                                     |             |  | Tag      | Scene Description                                                                                   | Duration | Validation Source            |
|------------------------------------------------|-------------|--|----------|-----------------------------------------------------------------------------------------------------|----------|------------------------------|
| The Blair Witch Project (Myrick et al., 1999)  | Blair Witch |  | Fear     | Final scene when screaming intensifies, man standing facing the wall and camera falls.              | 2:03     | Schaefer et al. (2010)       |
| The Conjuring (Wan, 2013)                      |             |  | Fear     | Girl gets out of bed at night and bags her head on a cupboard. Frantic scene.                       | 2:26     | İyilikci et al. (2024)       |
| American History X (Kaye, 1998)                |             |  | Anger    | Neo-Nazi kills a black man, smashing his head on the curb and then smiles after being arrested.     | 3:24     | Schaefer et al. (2010)       |
| Platoon (Stone, 1986)                          |             |  | Anger    | Villagers pushed around in burning village and soldier stops other soldiers from raping a child.    | 2:42     | Author tested in pilot.      |
| Baby laughing at ripping paper (YouTube, 2011) |             |  | Joy      | 8-month-old Micah (a boy) laughing hysterically while at-home daddy rips up a job rejection letter. | 1:44     | Author tested in pilot.      |
| Cats and Dog playing together (YouTube, 2022a) |             |  | Joy      | Dog lies peacefully on a large bed with kittens and adult cat moving around. With happy music.      | 1:53     | Author tested in pilot.      |
| One Day (Scherfig, 2011)                       |             |  | Surprise | Woman rides a bicycle; she gets hit by a truck.                                                     | 2:26     | Zupan and Eskritt (2020)     |
| Neighbors (Nicholas Stoller and O'Brien, 2014) |             |  | Surprise | Woman calls man about missing airbags Man is ejected to an office ceiling.                          | 1:07     | Author tested in pilot.      |
| Trainspotting (Boyle, 1996)                    |             |  | Disgust  | The main character enters "The worst toilet in Scotland" and later dives into a filthy toilet bowl. | 1:23     | Schaefer et al. (2010)       |
| Planet Terror (Rodriguez, 2007)                |             |  | Disgust  | Scene where man is examined by doctors in a hospital and exposes infected parts of his body.        | 2:01     | Michelini et al. (2019)      |
| Young impala and dead mother (YouTube, 2022b)  |             |  | Sadness  | Young impala finds adult impala lying down and apparently dead. Then lies by dead animal.           | 1:44     | Author tested in pilot.      |
| My Girl (Zieff and Elwhany, 1991)              |             |  | Sadness  | Funeral scene where girl cries and runs away after approaching the casket where a little boy lies.  | 1:39     | Gabert-Quillen et al. (2015) |

Table 1: Listing and information about the 12 movie clips used to elicit discrete emotions in the main study.

### 3 Dataset

The block diagram in Figure 1 summarizes our data collection, task definitions, and methodology, which we will elaborate on here and in the next section. Data collection was carried out in person at INSPIRE Laboratory of the School of Psychology at University of Ottawa. The experiment procedure was approved by the Research Ethics Board of University of Ottawa. The study included 97 student participants aged 18 to 27 ( $M = 19.9$ ,  $SD = 2.5$ ). The majority were women (81 women, 15 men, and 1 non-binary), and most participants were native English speakers (65 spoke English as their first language, 12 spoke French, and 20 spoke other languages). The sample was ethnically diverse, comprising 16 Asian, 20 Black/African, 7 Hispanic/Latino, 1 Indigenous, 15 Mixed/Multiple Ethnicities, 33 White/Caucasian, and 5 participants identifying as Other.

Our study focused on the six basic emotions identified by Ekman (1992) as the target emotions in our experimental setup: sadness, fear, joy, dis-

gust, surprise, and anger. Two movie clips for each emotion were sourced from film stimuli in the existing literature and validated in our pilot study. The twelve movie clips used in the study and their meta-information are listed in Table 1. We trimmed clips to ensure optimal emotional salience and duration. Their effectiveness was validated in a pilot study with 25 participants before the final data collection.

In the main study, participants watched six emotional video clips, one from each emotion category. To re-establish baseline levels of valence and arousal, the presentation of each emotional clip was preceded by a neutral video clip. To further mitigate potential carryover effects between conditions, a two-minute rest period was inserted between each neutral-emotional clip sequence, during which one of six still images was displayed on the computer screen. All video clips and still images were presented in random order to minimize potential sequence effects. The collected dataset consists of  $97 \times 6$  entries, with five elements: 1) **Speech**: a 30-second audio recording of the participant’s verbal response to the following instruction:

"You are asked to verbally describe the scene during which you felt the strongest emotion in the last film clip and say how it made you feel." 2) **Intended emotions**: Each video is expected to provoke a certain emotion. 3) **Evoked emotions**: the intensities at which each of the emotions (sadness, fear, joy, disgust, surprise, anger) was felt, as rated by the participants on a 7-point Likert scale going from *not at all* to *strongly*. 4) **Valence**: the extent to which the overall feeling of the participant was positive or negative. 5) **Arousal**: the intensity of the overall feeling of the participant while watching the video. Valence and arousal were measured on a validated sliding scale where each extreme was illustrated by an emoticon.

## 4 Tasks

We define three tasks to examine the relationship between semantic types and emotion recognition. To determine the most predictive semantic type for each task, we experimented with three different inputs: full transcriptions, descriptive semantic segments, and expressive semantic segments.

**TASK-1: Classification of Intended Emotion** involves classifying the intended emotion associated with each video based on participants' speech.

**TASK-2: Classification of Evoked Emotion** involves classifying participant-reported evoked emotions, which are subjective and may include multiple emotions simultaneously. While evoked emotions often include the intended emotion, individual differences can lead to variations. This task is framed as a multi-label classification problem, where each emotion (on a scale of 0 to 6) is binarized based on whether it is evoked or not.

**TASK-3: Regression of Valence and Arousal** predicts participants' self-reported valence and arousal ratings, which provide a two-dimensional representation of emotional states.

## 5 Methodology

As depicted in Figure 1, our methodology consists of three sequential steps: speech recognition, semantic segmentation, and emotion prediction.

**Step-1: Automatic Speech Recognition** - We used Whisper (Radford et al., 2023), an automatic speech recognition model, to transcribe the participants' speech data into text.

|                           | Descriptive semantics | Expressive semantics |
|---------------------------|-----------------------|----------------------|
| LLM & Annotator 1         | 0.71                  | 0.73                 |
| LLM & Annotator 2         | 0.84                  | 0.83                 |
| Annotator 1 & Annotator 2 | 0.77                  | 0.74                 |
| Random & Random           | 0.63                  | 0.64                 |

Table 2: Human evaluations of GPT-4o text segmentations. The agreement between two human annotators was comparable to human-LLM agreements.

**Step-2: Semantic Segmentation** - We used GPT-4o (OpenAI, 2023) to extract descriptive and expressive segments from the transcription obtained in step 1. The prompt is given in Table 3. We set the sampling temperature to 0 to make the process more deterministic. Overlapping phrases were allowed when semantic roles intersected, ensuring comprehensive representation.

**Step-3: Emotion Prediction** - The last step is to perform tasks described in Section 4 to study the relationship between semantic roles and emotion recognition. Each model is trained and evaluated on three input types: full transcriptions, descriptive segments, and expressive segments.

**Audio-Based Emotion Classification** - In addition to text-based models, we also experimented with audio-based models trained directly on the speech recordings. These included a HuBERT model (Hsu et al., 2021), a Wav2Vec2 model (Baevski et al., 2020), and a baseline MFCC (mel-frequency cepstral features) classifier. The audio classifiers were evaluated on TASK-1 and TASK-2 using the full utterance audio as input. However, all speech-based models performed significantly worse than text-based classifiers. Since semantic role segmentation (i.e., distinguishing between descriptive and expressive segments) is inherently a linguistic task and not inferable from acoustic signals alone, we prioritized text-based methods for the core analyses of this paper moving forward.

## 6 Experiments

### 6.1 Implementation Details

For Step-1, automatic speech recognition, we used 'whisper-large-V3'<sup>1</sup>, a state-of-the-art system ASR model known for its robustness across diverse accents and noise conditions. We manually reviewed transcriptions in the development set, consisting of  $33 \times 6$  audio transcriptions. Whisper achieved a

<sup>1</sup><https://huggingface.co/openai/whisper-large-v3>

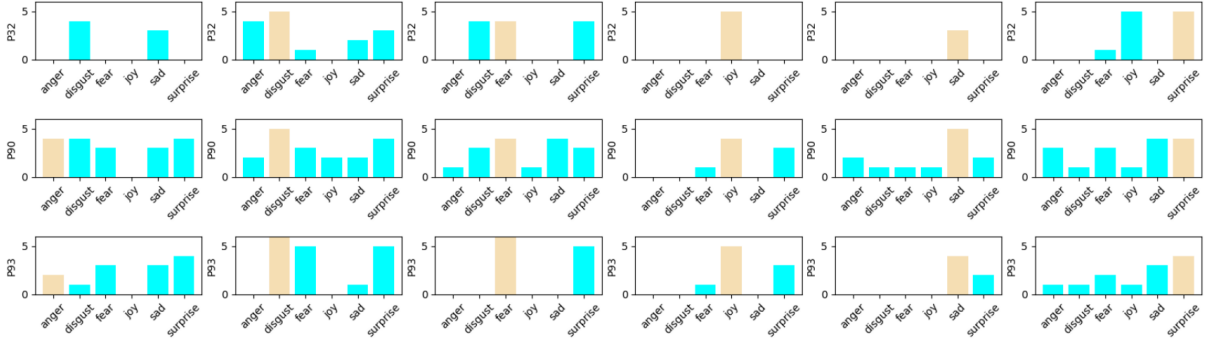


Figure 2: Examples of participants’ rated emotions. Each row represents a participant who watched six movie segments (6 columns) from each of the six emotional categories. The intended emotion tag associated with the video is plotted in a yellow bar. Other rated emotions are colored blue. The height of the bars represents the emotion ratings from participants. For example, in the second movie clip watched by Participant P93, the intended emotion was "disgust," as shown by the yellow bar. After watching the clip, P93 reported experiencing four emotions: disgust, fear, sadness, and surprise, indicated by the blue bars. Among these, “disgust” was the strongest emotion, receiving the highest score of 6.

The user will provide a paragraph describing their feelings towards a particular movie, delimited with “#####”.

Your task is to segment the paragraph into two parts according to the type of content: descriptive segments and expressive segments.

Descriptive segments refer to elements or clauses that provide factual or narrative information about the movie content without explicitly reflecting personal emotions or opinions.

Expressive segments refer to elements or clauses that convey personal feelings, attitudes, or opinions. These segments reflect individual reactions, emotions, and perceptions, or the intensity of these emotions.

The two parts (descriptive segments and expressive segments) can overlap, but all clauses of the given paragraph must be contained in at least one of the two parts.

Output your answer in the following format:

```
<answer>
  <descriptive> [descriptive segments] </descriptive>
  <expressive> [expressive segments] </expressive>
</answer>
```

Table 3: Prompt for extracting descriptive and expressive semantics from speech transcription.

4.13% word error rate, with errors mainly in unclear utterances at speech boundaries and between clauses.

For Step-2, we validated the effectiveness of GPT-4o segmentation again on the development set with  $33 \times 6$  transcriptions. Two authors of this paper, one from the Computer Science department and the other from the Psychology department, were given the same instructions as the LLM and inde-

pently performed the same segmentation task. To calculate the agreement between human annotators and also between LLM and annotators, we computed cosine similarities of the segments, using sentence-transformer embeddings<sup>2</sup>. From Table 2, the average agreement between two human annotators (0.76) was comparable to human-LLM agreement (0.73 and 0.83). Most discrepancies arose from minor conjunctions to make sentences more complete. As a baseline, two random text segmentations would result in a similarity score of 0.63 - 0.64. Overall, GPT-4o has an acceptable segmentation quality.

For Step-3, emotion prediction, we fine-tuned different classifiers/regressors, including BERT (Devlin, 2018), RoBERTa (Liu, 2019), and DeBERTa (He et al., 2020). Different text semantics identified in Step-2 are used as inputs to the models. For emotion classification (Tasks 1 and 2), we used the text embeddings from the models and applied a standard classification head with a softmax activation function to predict categorical emotions. For regression (Task 3), we modified the models by replacing the classification head with a fully connected layer that outputs a single continuous value, trained with mean squared error (MSE) loss to predict valence and arousal scores. This approach follows standard practice in adapting transformer encoders for regression tasks (Xin et al., 2021; Taha, 2024; Orso and Xie, 2008). Data are split on participants’ level, with 1/3 of partici-

<sup>2</sup><https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

|                                                                                   |        |
|-----------------------------------------------------------------------------------|--------|
| How often do participants experience the intended emotion conveyed by the videos? | 96.63% |
| How frequently do participants feel emotions other than the intended one?         | 89.39% |
| How often is the intended emotion rated as the highest by participants?           | 79.29% |
| Chippendale’s alpha coefficient between intended emotion and evoked emotions      | 0.1466 |

Table 4: Statistics of relationships between movie intended emotion tags and evoked emotions. Predicting the evoked emotions is a much more subjective task than predicting the intended emotion tag.

pants (33 participants) data used for training, 1/3 for validation, and the rest 1/3 for testing.

## 6.2 Comparative Analysis of Intended and Evoked Emotions

Table 4 shows the relationship between the intended and evoked emotions. While the participants experienced the intended emotion 96.63% of the time, they also reported other emotions 89.39% of the time. Surprisingly, more than 20% of the time, an emotion other than the intended one is experienced most. These results suggest that the experienced emotion is highly subjective and can deviate from the intended emotions.

To better quantify the subjectivity of Task-2, we calculate the Krippendorff’s alpha coefficient (Krippendorff, 2011, 2018) between the intended emotion and evoked emotions. We treat the agreement between intended emotion and evoked emotion as the agreement between two annotators performing multi-label annotations. Each annotator labels 594 data points, since there are  $99 \times 6$  speech. One annotator always label the intended emotion as *true* and other emotions as *false*. The other annotator labels the data with the participant’s evoked emotion ratings in a multi-label fashion. Krippendorff’s alpha coefficient is calculated as the inter-annotator agreement index on this multi-label annotation task with MASI distance (Passonneau, 2006) as the distance measurement between two sets of multi-label annotations. The low score of Krippendorff’s alpha coefficient shows the high subjectivity of task T2 and the high variation of evoked emotions with respect to the intended emotion.

Figure 2 gives examples of emotion ratings by three different participants in response to the six movie segments. Each row in the grid represents data from a different participant, while each column corresponds to one of the six movie segments. Within the bar charts, yellow bars indicate the intended emotion that the video clip aimed to elicit, while blue bars represent the emotions self-reported

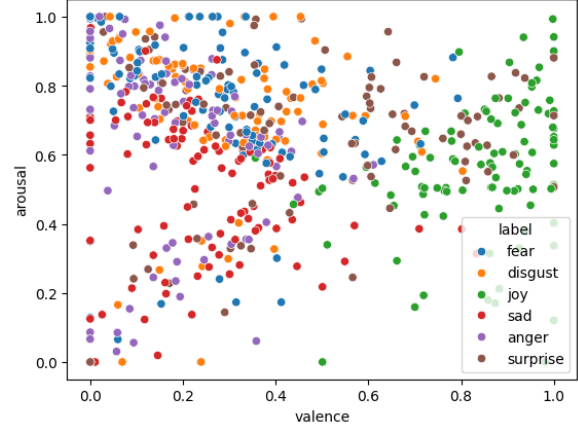


Figure 3: Valence and arousal ratings, colored by the intended emotion tags of movie segments.

by the participants after watching the clips. The height of the bars reflects the intensity of the rated emotions on a numerical scale. These examples highlight variability in participants’ emotional responses, often revealing discrepancies between the intended emotions and the emotions participants actually experienced.

## 6.3 Classification of Intended Emotion

Aligned with our hypothesis, the classification results for TASK-1 demonstrate a clear advantage of using descriptive semantics as input for predicting the intended emotions associated with each movie segment. Table 5 shows the classification accuracy for both semantic types across three different classifiers. Across all models, descriptive semantics consistently yield significantly higher accuracy in predicting the intended emotions.

## 6.4 Classification of Evoked Emotion

In TASK-2, we classified participant-reported evoked emotions, which are inherently subjective and may include multiple emotions simultaneously. Aligning with our hypothesis that expressive semantics better capture speaker-specific emotional experiences, results in Table 6 indicate that using expressive semantics as input achieves higher clas-

| Model   | Semantics   | Precision | Recall | F1   |
|---------|-------------|-----------|--------|------|
| BERT    | Descriptive | 0.83      | 0.81   | 0.81 |
|         | Expressive  | 0.68      | 0.65   | 0.65 |
|         | Full        | 0.89      | 0.88   | 0.88 |
| RoBERTa | Descriptive | 0.85      | 0.83   | 0.83 |
|         | Expressive  | 0.69      | 0.69   | 0.69 |
|         | Full        | 0.93      | 0.93   | 0.93 |
| DeBERTa | Descriptive | 0.81      | 0.81   | 0.81 |
|         | Expressive  | 0.65      | 0.64   | 0.64 |
|         | Full        | 0.91      | 0.90   | 0.90 |

Table 5: Model performances on classifying intended emotion associated with the movies.

| Model   | Semantics   | Precision | Recall | F1   |
|---------|-------------|-----------|--------|------|
| BERT    | Descriptive | 0.72      | 0.77   | 0.73 |
|         | Expressive  | 0.75      | 0.77   | 0.75 |
|         | Full        | 0.78      | 0.82   | 0.78 |
| RoBERTa | Descriptive | 0.73      | 0.76   | 0.73 |
|         | Expressive  | 0.74      | 0.82   | 0.77 |
|         | Full        | 0.76      | 0.82   | 0.77 |
| DeBERTa | Descriptive | 0.71      | 0.76   | 0.72 |
|         | Expressive  | 0.74      | 0.81   | 0.76 |
|         | Full        | 0.76      | 0.81   | 0.77 |

Table 6: Average model performances on classifying evoked emotions (std is always less than 0.1 over 5 run).

sification accuracy compared to using descriptive semantics. We also observe that even with full semantics, TASK-2 achieves significantly lower F-scores than Task-1, as expected due to the subjectivity of this task.

## 6.5 Discussion of Audio-Based Classifications

To assess the role of acoustic features in emotion recognition, we trained several audio-only classifiers, including models based on HuBERT<sup>3</sup>, Wav2Vec2<sup>4</sup>, and MFCC features, for both TASK-1 and TASK-2. Across all models, we observed consistently poor performance, with classifiers frequently defaulting to one or two majority emotion classes. This suggests that prosodic and paralinguistic cues in our dataset were not strongly indicative of emotional content. One likely explanation is that participants generally delivered their responses in a steady and emotionally neutral tone, which limited the expressiveness of acoustic features.

Moreover, unlike text-based inputs, speech signals do not easily lend themselves to semantic segmentation without speech recognition (Wang et al., 2003;

Ong and Herrera, 2005). Audio-based classifiers cannot distinguish between descriptive and expressive segments in an obvious way, making it difficult to explore the semantic roles that are central to our research questions. While acoustic features are valuable in many speech emotion recognition tasks, in our study design where subjective emotional experience is linked to semantic framing, textual cues proved more informative and interpretable.

## 6.6 Regression of Emotion Valence and Arousal:

Figure 3 shows the distributions of valence and arousal across different intended emotions, which exhibit high variability without clear patterns across emotions. Positive emotions, such as joy, correlates with higher valence, and negative emotions, such as fear, have lower valence and higher arousal. But there is no obvious clusters among the six emotions.

The results reported in Table 7 show that expressive semantics lead to more accurate predictions for both emotional valence and arousal compared to descriptive semantics. A statistical analysis in Table 8 shows that the differences in the prediction

<sup>3</sup>facebook/hubert-base-ls960

<sup>4</sup>facebook/wav2vec2-base

| Model   | Semantics   | Valence MSE | Valence MAE | Arousal MSE | Arousal MAE |
|---------|-------------|-------------|-------------|-------------|-------------|
| BERT    | Descriptive | 0.068       | 0.209       | 0.057       | 0.192       |
|         | Expressive  | 0.055       | 0.183       | 0.054       | 0.187       |
|         | Full        | 0.053       | 0.185       | 0.053       | 0.184       |
| RoBERTa | Descriptive | 0.050       | 0.184       | 0.055       | 0.184       |
|         | Expressive  | 0.037       | 0.151       | 0.051       | 0.182       |
|         | Full        | 0.034       | 0.146       | 0.051       | 0.182       |
| DeBERTa | Descriptive | 0.077       | 0.224       | 0.053       | 0.183       |
|         | Expressive  | 0.037       | 0.153       | 0.045       | 0.166       |
|         | Full        | 0.059       | 0.192       | 0.049       | 0.172       |

Table 7: Model performances on regression of emotion valence and arousal. Expressive semantics leads to smaller errors in estimating evoked valence and arousal. The difference is most pronounced for the DeBERTa-based model.

| Model   | Valence            |         | Arousal            |         |
|---------|--------------------|---------|--------------------|---------|
|         | Z                  | p       | Z                  | p       |
| BERT    | -1.74 <sup>a</sup> | 0.083   | -1.24 <sup>a</sup> | 0.215   |
| RoBERTa | -2.98 <sup>a</sup> | 0.003   | -0.16 <sup>a</sup> | 0.874   |
| DeBERTa | -5.86 <sup>a</sup> | < 0.001 | -3.70 <sup>a</sup> | < 0.001 |

Table 8: Wilcoxon signed-rank tests results to compare MSE between descriptive and expressive semantics for each model. <sup>a</sup>Based on positive ranks.

errors between descriptive and expressive semantics are statistically significant for valence under two of the three models and one model for arousal. The regression results are in line with the TASK-2 results and the statistical analysis partially supports the hypothesis that expressive semantics better capture subjective experience.

## 7 Conclusion

This study introduces a novel framework for Speech Emotion Recognition (SER) by distinguishing between semantic roles in speech. By leveraging LLMs’ zero-shot capabilities in text segmentation, we tackle a previously difficult challenge. To our knowledge, this is the first work to segment speech into two semantic roles, *expressive* and *descriptive* content, to enable more fine-grained and nuanced emotion detection.

Our findings reveal that descriptive semantics are more predictive of intended emotions, while expressive semantics are more closely aligned with evoked emotions and their valence and arousal dimensions. This differentiation can inform future research in emotion detection. In some contexts, it might be more useful to instruct users and guide them toward only one of these modes of expressing emotions. In other applications, it might be more suitable to leave it to the users to express their emo-

tions in a mixture of expressive and descriptive modes. The LLMs can then be used to segment the speech and use the segments depending on the predictive goals. This approach enhances the development of more accurate and context-aware emotion recognition systems, with applications in mental health, virtual assistants, and customer service.

## Limitations

This study, while providing valuable insights into the segmentation of speech for emotion recognition, has limitations. First, the dataset used in this research is curated from emotionally evocative movie clips, which, although varied, may not fully represent the broad diversity of real-world speech interactions. The emotional expressions captured in these clips might not encompass the full spectrum of spontaneous and everyday speech, which could limit the generalizability of the findings.

Second, although we included baseline speech-based emotion classifiers, their performance was substantially lower than that of text-based models. This gap likely stems from the emotional neutrality of the participants’ tone and the nature of the task. However, future work could explore whether jointly modeling text and acoustic features, perhaps guided by semantic segmentation, might uncover latent prosodic patterns aligned with specific se-

semantic roles.

Third, while the study distinguishes between descriptive and expressive semantics, it focuses primarily on self-reported emotional responses, which can be subjective and influenced by individual differences in emotional expression and perception. This subjectivity introduces variability in the emotional ratings, potentially affecting the accuracy and robustness of the regression models.

## Ethics Statement

This research was approved by the Research Ethics Board of University of Ottawa. All participants provided their informed consent prior to participating in the study. Participants had the option to withdraw from the experiment at any time and for any reason, including emotional distress. Data collected during the study were handled securely and used exclusively for research purposes. All personal data was anonymized.

In Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis, [Mohammad \(2022\)](#) provides a structured ethical framework for developing and deploying Automatic Emotion Recognition (AER) systems, along 50 ethical considerations. He specifically emphasizes on the risks of privacy violations, reinforcing biases, and potential misuse in surveillance or manipulation. This Ethics Sheet serves as a guide for responsible AER development, and encourages researchers to question *why they automate, whose interests are served, and how success is measured*.

Recognizing the ethical risks and potential misuse of SER technologies, we strongly caution against issues such as biases in emotion datasets, AI models enforcing rigid norms on emotional expression, and the exclusion of neurodiverse and marginalized groups. These concerns must be carefully addressed before deploying SER systems in real-world applications. We urge industries to adopt responsible, explainable, and inclusive AI development practices, ensuring that these technologies are fair, transparent, and beneficial to all users.

## Acknowledgments

We extend our gratitude to OrbMedic Inc. (Sonny Chaiwala), Ottawa, Canada, for their financial and technical support throughout this project. Their expertise and insights were pivotal in shaping the

direction of our research and ensuring its successful execution. Additionally, we acknowledge the support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Centres of Innovation (OCI).

## References

- Alexei Baeviski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Lisa Feldman Barrett, Batja Mesquita, Kevin N Ochsner, and James J Gross. 2007. The experience of emotion. *Annu. Rev. Psychol.*, 58(1):373–403.
- A Batliner, S Steidl, and E Nöth. 2006. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In *Programme of the Workshop on Corpora for Research on Emotion and Affect*, page 28.
- Gulmira Bekmanova, Banu Yergesh, Altynbek Sharipbay, and Assel Mukanova. 2022. Emotional speech recognition method based on word transcription. *Sensors*, 22(5):1937.
- Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa, and Amena Mahmoud. 2022. Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, 2022(1):2645381.
- Dmitri Bitouk, Ragini Verma, and Ani Nenkova. 2010. Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625.
- Danny Boyle. 1996. Trainspotting film. *United Kingdom: Polygram Filmed Entertainment*.
- Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.
- Adil Chakhtouna, Sara Sekkate, et al. 2024. Unveiling embedded features in wav2vec2 and hubert models for speech emotion recognition. *Procedia computer science*, 232:2560–2569.
- HongYi Chen, Kai Ling Chin, and Chrystalle B.Y. Tan. 2021. Selection and validation of emotional videos: Dataset of professional and amateur videos that elicit basic emotions. *Data in Brief*, 34:106662.

- Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Soumya Dutta and Sriram Ganapathy. 2024. Leveraging content and acoustic representations for efficient speech emotion recognition. *arXiv preprint arXiv:2409.05566*.
- Paul Ekman. 1992. Facial expressions of emotion: New findings, new questions.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.
- Wei-quan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. 2021. Lssed: a large-scale dataset and benchmark for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 641–645. IEEE.
- Crystal A Gabert-Quillen, Ellen E Bartolini, Benjamin T Abravanel, and Charles A Sanislow. 2015. Ratings for emotion film clips. *Behavior research methods*, 47:773–787.
- Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation learning for speech emotion recognition. In *Interspeech*, pages 3603–3607.
- Ziwei Gong, Qingkai Min, and Yue Zhang. 2023. Eliciting rich positive emotions in dialogue generation. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 1–8.
- Kenta Hama, Atsushi Otsuka, and Ryo Ishii. 2024. Emotion recognition in conversation with multi-step prompting using large language model. In *International Conference on Human-Computer Interaction*, pages 338–346. Springer.
- Dini Handayani, Abdul Wahab, and Hamwira Yaacob. 2015. Recognition of Emotions in Video Clips: The Self-Assessment Manikin Validation. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 13(4):1343.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Elvan Arıkan İyilikci, Merve Boğa, Elif Yüvrük, Yıldız Özkılıç, Osman İyilikci, and Sonia Amado. 2024. An extended emotion-eliciting film clips set (egefilm): assessment of emotion ratings for 104 film clips in a turkish sample. *Behavior Research Methods*, 56(2):529–562.
- Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.
- Mimansa Jaiswal and Cristian-Paul Bara. 2020. Muse: a multimodal dataset of stressed emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.
- Tony Kaye. 1998. American history x, new line cinema. *Cited on*, page 99.
- Dacher Keltner, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, 43:133–160.
- Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Sadik, and Soonil Kwon. 2024. Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, 245:122946.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Interspeech*, pages 125–128. Citeseer.
- Matthias Ley, Maria Egger, and Sten Hanke. 2019. Evaluating methods for emotion recognition based on facial and vocal features. In *AmI (Workshops/Posters)*, pages 84–93.
- Keding Li, Xunbing Shen, Zhencai Chen, Liping He, and Zhennan Liu. 2021. *Effectiveness of Emotion Eliciting of Video Clips: A Self-report Study*, pages 523–542. Springer International Publishing.
- MS Likitha, Sri Raksha R Gupta, K Hasitha, and A Upendra Raju. 2017. Speech based human emotion recognition using mfcc. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, pages 2257–2260. IEEE.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

- Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183.
- Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The msp-conversation corpus. *Interspeech 2020*.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.
- Tao Meng, Fuchen Zhang, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li. 2024. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *arXiv preprint arXiv:2404.17862*.
- Yanina Michelini, Ignacio Acuña, Juan Ignacio Guzmán, and Juan Carlos Godoy. 2019. Latemo-e: a film database to elicit discrete emotions and evaluate emotional dimensions in latin-americans. *Trends in Psychology*, 27(2):473–490.
- Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.
- Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.
- Agnes Moors. 2009. Theories of emotion causation: A review. *Cognition and emotion*, 23(4):625–662.
- Agnes Moors. 2024. An overview of contemporary theories of emotions in psychology. *Emotion Theory: The Routledge Comprehensive Guide: Volume I: History, Contemporary Theories, and Key Elements*.
- Louis-Philippe Morency and Tadas Baltrušaitis. 2017. Multimodal machine learning: integrating language, vision and speech. In *Proceedings of the 55th annual meeting of the association for computational linguistics: Tutorial abstracts*, pages 3–5.
- Daniel Myrick, Eduardo Sánchez, Heather Donahue, Michael Williams, and Joshua Leonard. 1999. *The Blair witch project*. Artisan Entertainment Santa Monica, CA.
- Andrew J. Cohen Nicholas Stoller and Brendan O’Brien. 2014. *Neighbors [Film]*. Universal Pictures.
- Minxue Niu, Amrit Romana, Mimansa Jaiswal, Melvin McInnis, and Emily Mower Provost. 2023. Capturing mismatch between textual and acoustic emotion expressions for mood identification in bipolar disorder. In *Interspeech*. Interspeech.
- Bee Suan Ong and Perfecto Herrera. 2005. Semantic segmentation of music audio. In *Proceedings of the International Computer Music Conference*, page 61.
- OpenAI. 2023. Models - openai api. <https://platform.openai.com/docs/models/gpt-4o>. Accessed: 2025-01-24.
- Alessandro Orso and Tao Xie. 2008. Bert: Behavioral regression testing. In *Proceedings of the 2008 international workshop on dynamic analysis: held in conjunction with the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2008)*, pages 36–42.
- Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *5th International Conference on Language Resources and Evaluation, LREC 2006*.
- Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.
- Robert Rodriguez. 2007. *Planet Terror [Film]*. Dimension Films.
- Lantana Dioren Rumpa, Adhi Dharma Wibawa, Mauridhi Heri Purnomo, and Harmelia Tulak. 2015. Validating video stimulus for eliciting human emotion: A preliminary study for e-health monitoring system. In *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, pages 208–213. IEEE.
- Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion*, 24(7):1153–1172.
- Lone Scherfig. 2011. *One Day [Film]*. Focus Features.
- Ewa Siedlecka and Thomas F Denson. 2019. Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, 11(1):87–97.
- M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.
- Oliver Stone. 1986. *Platoon [Film]*. Orion Pictures.

- Kamal Taha. 2024. Text regression analysis: A review, empirical, and experimental insights. *IEEE Access*.
- Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Björn W Schuller. 2021. Speech emotion recognition using semantic information. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE.
- Ismail Rasim Ulgen, Zongyang Du, Carlos Busso, and Berrak Sisman. 2024. Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12081–12085. IEEE.
- Kannan Venkataramanan and Haresh Rengaraj Rajamohan. 2019. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*.
- James Wan. 2013. *The Conjuring [Film]*. Warner Bros. Pictures.
- Dong Wang, Lie Lu, and Hong-Jiang Zhang. 2003. Speech segmentation without speech recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.
- Siqing Wu, Tiago H Falk, and Wai-Yip Chan. 2011. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785.
- Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. *BERxiT: Early exiting for BERT with better fine-tuning and extension to regression*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 91–104, Online. Association for Computational Linguistics.
- Xin Zhou Xu, Jun Deng, Nicholas Cummins, Zixing Zhang, Li Zhao, and Björn W Schuller. 2021. Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes. *IEEE Transactions on Multimedia*, 24:2752–2765.
- Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE spoken language technology workshop (SLT)*, pages 112–118. IEEE.
- YouTube. 2011. Baby laughing hysterically at ripping paper. <https://www.youtube.com/watch?v=RP4abiHdQpc>.
- YouTube. 2022a. Mom cat shows baby kittens that golden retriever is safe for them. <https://www.youtube.com/watch?v=N2hssAyomdU>.
- YouTube. 2022b. The saddest video ever captured. <https://www.youtube.com/watch?v=Xjwu2Auecko>.
- Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. 2016. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313.
- Huan Zhao, Nianxin Huang, and Haijiao Chen. 2024. Knowledge enhancement for speech emotion recognition via multi-level acoustic feature. *Connection Science*, 36(1):2312103.
- Howard Zieff and Laurice Elehwany. 1991. *My Girl [Film]*. Columbia Pictures.
- Barbra Zupan and Michelle Eskritt. 2020. Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion. *The Journal of Social Psychology*, 160(6):768–789.