Multimodal Audio Emotion Recognition with Graph-based Consensus Pseudolabeling

**Gabriel Natal Coutinho[1] , Artur de Vlieger Lima[1] , Juliano Yugoshi[1] ,
Marcelo Isaias de Moraes Junior[1] , Marcos Paulo Silva Gôlo[1] ,
Ricardo Marcondes Marcacini[1]**

[1]Institute of Mathematics and Computer Sciences - University of São Paulo (USP)
São Carlos – São Paulo – Brazil

{gabriel.natal, vligmart, juliano.yugoshi, marcelo.junior}@usp.br,

marcosgolo@usp.br, ricardo.marcacini@icmc.usp.br

Abstract. *This paper presents a novel method called Multimodal Graph-based Consensus Pseudolabeling (MGCP) for unsupervised emotion recognition in audio. The goal is to determine the emotion of audio segments using the circumplex model of emotions. The method combines pre-trained unimodal models for audio and text and follows a three-step process. First, audio segments are represented using embeddings from unimodal models. Then, modality-specific graphs are constructed based on similarity and integrated into a multimodal graph. Finally, pseudolabels are generated by measuring consensus between modalities, and a graph regularization framework is introduced to estimate the final emotion coordinates. Experimental evaluation shows the effectiveness of the MGCP method, surpassing both unimodal and traditional multimodal models, enabling audio emotion recognition without labeled data specific to the target domain.*

1. Introduction

Emotion recognition from audio is a challenging study field with applications in various domains, including human-computer interaction, affective computing, and multimedia analysis [Saxena et al. 2020]. The main objective is to determine the emotion of audio segments, usually using the circumplex model of emotions, a two-dimensional space representing valence and arousal [Konar and Chakraborty 2015]. Figure 1 illustrates a basic example of the emotion circumplex model. In this case, an audio segment contains a valence coordinate, which represents the pleasantness or unpleasantness of the emotion, and an arousal coordinate, which represents the intensity or activation level of the emotion. From these coordinates, it is possible to infer a closer emotion, ranging from the six primary emotions (joy, sadness, anger, fear, surprise, and disgust) or neutral to more refined models encompassing approximately 27 emotions, as proposed by the GoEmotions dataset [Demszky et al. 2020].

Traditional methods for emotion recognition from audio have primarily focused on unimodal analysis, mainly exploiting the audio signal [Tomar et al. 2022]. However, recent advancements in the field have emphasized integrating multimodal information importance, particularly the textual modality derived from transcribed audio [Abdullah et al. 2021, Ezzameli and Mahersia 2023]. While the audio modality captures the "how" component of information expression, such as tone of voice and

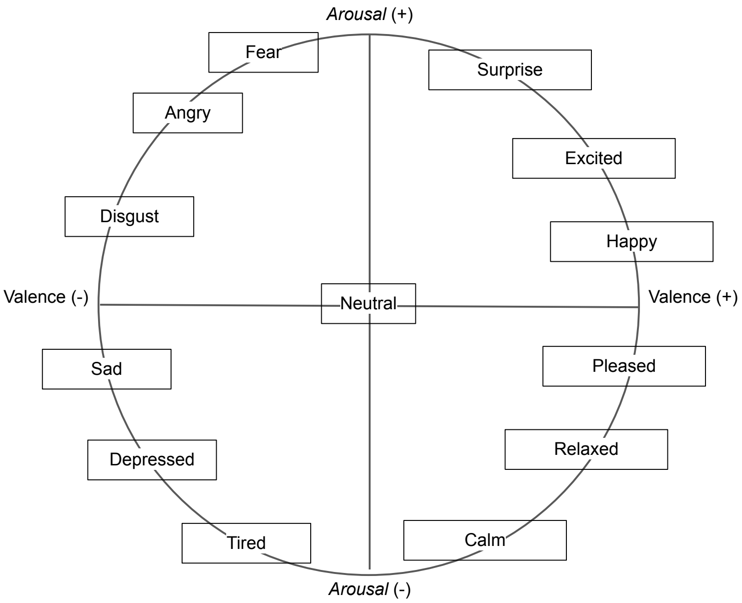


Figure 1. Example of the emotion circumplex model in two-dimensional space representing valence and arousal (Adapted from [Russell 1980]).

speech patterns, the textual modality provides insights into the “what” component, enabling the estimation of emotions based on the words and topics. By considering both modalities together, models can capture complementarity and, consequently, enhance the accuracy and robustness of emotion recognition systems [Priyasad et al. 2020, Siriwardhana et al. 2020, Krishna and Patil 2020]. However, this also presents additional challenges, as it requires appropriate preprocessing, structuring, and representation of each modality, as well as investigating techniques for aligning and extracting complementarity and redundancy between audio and texts [Baltrusaitis et al. 2019].

In unimodal emotion recognition, deep learning transformer-based models have emerged as the state-of-the-art approach [Siriwardhana et al. 2020, Hsu et al. 2021]. Wav2Vec [Schneider et al. 2019] and HuBERT [Hsu et al. 2021] have demonstrated superior performance for emotion recognition tasks in the audio modality. These models leverage large amounts of unlabeled audio data for pretraining a speech representation model using self-supervised representation learning [Schneider et al. 2019, Hsu et al. 2021]. Meanwhile, in the textual modality, language models such as BERT and GPT achieved state-of-the-art results [Adoma et al. 2020]. These pre-trained models are fine-tuned on emotion-labeled text data to capture contextual information specifically related to emotions [Kenton and Toutanova 2019]. However, both approaches require labeled data for fine-tuning the pre-trained models, which poses challenges in emotion recognition. Labeling data in a single modality is already a complex task, and this challenge becomes even greater when aligned labeled data is required for both audio and textual modalities. Therefore, we raise the following research question: ***How can we leverage and combine pre-trained unimodal emotion recognition models to develop a multimodal model capable of estimating audio emotion in the absence of labeled data?***

This paper presents a novel method called MGCP (Multimodal Graph-based Consensus Pseudolabeling) for unsupervised emotion recognition in audio that leverages the concept of pseudolabeling. Pseudolabeling is a technique where annotations or labels are generated based on certain dataset statistics instead of relying on human-labeled data [Lee et al. 2013, Arazo et al. 2020]. Our MGCP method consists of three steps. In the first step, each audio segment is represented in a latent vector space using embeddings obtained from the respective audio and text unimodal models, which capture the semantic meaning of the segments. Subsequently, we construct a graph for each modality, with segments as vertices and edges connecting the top-k most similar segments in the embedding space. In the second step, we combine the two modalities in a new unified graph that preserves the edge relations from the individual modality graphs while associating each vertex with the arousal-valence coordinates predicted by the unimodal models. To generate pseudolabels, our MGCP method selects segments with close arousal-valence coordinates in both modalities, thereby leveraging the consensus between the unimodal models to create labeled data. Finally, in the third step, we propose a novel multimodal regularization method for graph learning to estimate the final arousal-valence coordinates for all vertices based on the pseudolabels obtained in the previous step. As a result, our MGCP method enables emotion estimation of audio segments using only pre-trained unimodal models relevant to the application, eliminating the need for labeled emotion data. In summary, our paper presents three main contributions.

1. **Combination of pre-trained unimodal models**: We introduce a strategy to represent audio and text segments in a unified multimodal graph structure, leveraging any pre-trained unimodal models fine-tuned on other datasets. Thus our MGCP method is agnostic to the specific unimodal models used, allowing for easy updates and adjustments as new state-of-the-art unimodal approaches emerge.
2. **Consensus-based pseudolabeling**: By exploiting the consensus among the unimodal models, the MGCP method generates pseudolabels based on the distance of valence-arousal coordinates predicted by each unimodal model. Additionally, as the unimodal models have been pre-trained on other datasets, our approach can be interpreted as pseudolabeling through transfer learning.
3. **Regularization framework for learning final arousal-valence coordinates**: Our MGCP method extends a regularization framework for multimodal emotion recognition in graphs. To the best of our knowledge, this is the first graph learning method that utilizes consensus-based pseudolabels for emotion recognition.

We conducted an experimental evaluation on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which is an acted, multimodal and multispeaker dataset comprising approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions, and 7 emotion categories. By leveraging the consensus between the audio and text modalities, our method achieved significant performance improvements, surpassing not only the individual unimodal models but also the multimodal models based on the late-fusion strategy (decision level). These results highlight the robustness and generalization capability of our graph-based consensus pseudolabeling approach, as it effectively estimates emotions in the audio segments without relying on labeled data specific to the target domain.

We organized the remainder of this paper as follows. First, Section 2 presents the background and related work under multimodal emotion recognition for audio and text. Second, Section 3 presents the proposal MGCP: Multimodal Graph-based Consensus Pseudolabeling. Third, Section 4 presents the experimental evaluation and discusses the results. Finally, Section 6 presents our concluding remarks and future work.

2. Background and Related Work

Emotion recognition from audio has been a topic of extensive research in recent years. Early works in this field focused on using prosodic features, which capture acoustic characteristics such as pitch, intensity, and rhythm, to infer emotions. These features were derived from speech signals and used in classification algorithms to recognize different emotional states [Shah Fahad et al. 2021]. However, these approaches had limitations in capturing the complexities and nuances of emotions, as prosodic features alone may not fully capture the audio's semantic meaning and contextual information.

With the advent of deep learning and the success of transformer models in NLP [Kenton and Toutanova 2019] and computer vision [Kolesnikov et al. 2020], significant advancements have been made in audio representation learning. Methods leveraging pre-trained models, such as Wav2Vec and HuBERT, have achieved state-of-the-art results. Both models have been pre-trained on large audio corpora using self-supervised learning. Wav2Vec 2.0 [Baevski et al. 2020] employs a CNN feature encoder to transform audio waveforms into latent speech representations, followed by mask operations and a Transformer-based contextualized encoder. On the other hand, HuBERT shares a similar architecture but incorporates an offline clustering step for Masked Language Model (MLM) pretraining. During this process, masked portions of the audio are associated with their respective clusters, allowing the model to predict the cluster assignments of these masked parts. This clustering step aids in learning discriminative representations for different audio segments. HuBERT and Wav2Vec 2.0 have outperformed traditional Automatic Speech Recognition (ASR) models. These models can capture acoustic information and learn meaningful representations, making them promising options for audio emotion recognition tasks. Recent studies have evaluated the fine-tuning of Wav2Vec and HuBERT for audio emotion recognition, resulting in state-of-the-art performance in unimodal scenarios [Wang et al. 2021].

Furthermore, [Bagadi 2021, Deng et al. 2021] have explored the incorporation of textual information in audio emotion recognition tasks. Text can provide complementary information, such as the explicit expression of emotions and the topics discussed [Das and Singh 2023]. Automatic text extraction methods, such as automatic speech recognition systems, can transcribe the spoken content and provide textual representations of the audio. We can integrate multiple modalities through early fusion, combining the features from different modalities at the input level. In the absence of annotated data for multimodal training, early fusion methods may not be practical. However, leveraging pre-trained models allows for late fusion through the decision level, where the output probabilities of the models are combined to generate a consensus decision [Baltrusaitis et al. 2019].

Another recent technique explored in deep learning to address the challenge of limited labeled data is pseudolabeling [Lee et al. 2013]. These methods leverage the predictions made by pre-trained models on unlabeled data to generate pseudo labels, which are then used for additional training or fine-tuning. Two common forms of pseudolabeling with pre-trained models involve confidence-based and consensus-based approaches. In the confidence-based approach, the confidence of pre-trained models is used to filter predictions and generate pseudo labels only for samples where the model exhibits high confidence. In the consensus-based approach, multiple pre-trained models are employed to generate pseudo labels. These models are trained on different datasets or architectures, and the consensus is obtained by agreement among their predictions.

We emphasize that combining pseudolabeling with pre-trained unimodal models, leveraging the consensus among different modalities, can potentially lead to improved performance and address the limitations of the scarcity of labeled data. In this study, we specifically investigate this strategy in audio-based emotion recognition, particularly in obtaining pseudolabels based on arousal-valence predictions within the emotion circumplex model.

3. Multimodal Graph-based Consensus Pseudolabeling

In this section, we present our proposed Multimodal Graph-based Consensus Pseudolabeling (MGCP) method for multimodal emotion recognition. The MGCP method consists of three main steps: (1) representing audio segments using embeddings from unimodal models, (2) constructing modality-specific graphs based on similarity, and integrating them into a multimodal graph for (3) learning the final valence-arousal coordinates in the circumplex model of emotions.

In the first step, we aim to represent each audio segment in a latent vector space using embeddings obtained from audio and text unimodal models. To achieve this, we encode the audio segment through a pre-trained audio unimodal model, such as Wav2Vec or HuBERT, which has been fine-tuned for emotion recognition on a labeled dataset (different from the unlabeled target dataset) [Schneider et al. 2019, Hsu et al. 2021]. This process yields the corresponding audio embedding and arousal-valence coordinates mapped to the circumplex model of emotions. Similarly, the text associated with the audio segment is encoded using a unimodal text model, such as BERT or GPT, to obtain the text embedding and its respective arousal-valence coordinates [Kenton and Toutanova 2019, Brown et al. 2020].

Formally, let $S = \{s_1, s_2, ..., s_n\}$ be a set of n audio segments, where each segment $s_i = (\vec{a}_i, \vec{t_i}) \in S$ consists of a pair of acoustic signal embedding $\vec{a}_i$ and transcribed text embedding $\vec{t_i}$. Now, we construct two graphs, one for each modality. In these graphs, the vertices represent the segments, and the edges connect the most similar segments in each embedding space. This graph construction relies on a similarity measure between the embeddings, such as cosine similarity[1]. By connecting segments that exhibit high similarity, we capture the local structure of the data within each modality [Rossi 2016]. Let $G_a = (V, E_a)$ and $G_t = (V, E_t)$ denote the graphs for audio and text, respectively, where each segment s_i is mapped to a vertex $v_i \in V$. The edges E_a and E_t are obtained based on Equations 1 and 2:

$$E_a = \{(v_i, v_j) | \textit{cosine similarity}(\vec{a}_i, \vec{a}_j) > \textit{threshold}\} \quad (1)$$

[1] Cosine similarity is the similarity between two vectors defined through an inner product space.

$$E_t = \{(v_i, v_j) | \textit{cosine similarity}(\vec{t_i}, \vec{t_j}) > \textit{threshold}\} \tag{2}$$

In these equations, $\vec{a}_i$ and $\vec{a}_j$ represent the audio embeddings of segments s_i and s_j, respectively, while $\vec{t_i}$ and $\vec{t_j}$ represent the text embeddings. We calculated the cosine similarity between these embeddings and formed the edges between vertices that surpassed a certain predefined threshold.

In the second step of the MGCP method, we integrate the two modality-specific graphs to create a multimodal graph that maintains the edge relations from the individual audio and text graphs. Let $G_m = (V, E_m)$ denote the multimodal graph, where V represents the vertices shared by the audio and text graphs. In the case of audio emotion recognition, we assume that all segments have corresponding audio and text representations, making them common vertices for both modalities. However, the edges can vary depending on the latent space of each modality. There are three possible scenarios:

- There are no edges connecting two vertices v_i and v_j in either G_a or G_t: in this case, there will also be no edge in G_m.
- There is an edge connecting v_i and v_j, but only in a single modality. In this case, there will be an edge in G_m, but with reduced weight.
- There are edges connecting v_i and v_j in both modalities. In this case, there will be an edge in G_m with a strengthened weight.

The edges E_m in the multimodal graph are obtained by merging the edges from the audio graph E_a and the text graph E_t, using a simple average of their respective weight matrices to achieve these scenarios, as described in Equation 3,

$$E_m = \frac{1}{2}(E_a + E_t), \tag{3}$$

in which E_a represents the edge matrix[2] from the audio graph, and E_t represents the edge matrix from the text graph. Operation $\frac{1}{2}(E_a + E_t)$ calculates the average of the corresponding edge weights in the two matrices.

Now, we proceed to the third step of the MGCP method, where we learn the final arousal-valence coordinates from the multimodal graph and the predicted coordinates in both modalities. Specifically, we exploit the consensus among the unimodal models to generate pseudolabels based on the proximity of their coordinates in both audio and text modalities. Let $c_a(x, y)_{v_i}$ and $c_t(x, y)_{v_i}$ represent the arousal-valence coordinates of segment s_i represented by vertex $v_i \in V$, obtained from the audio and text modalities, respectively. In Equation 4, we propose the strength of a pseudolabel (SPL),

$$\textit{SPL}_{v_i} = 1 - \frac{\sqrt{(w_a c_a(x)_{v_i} - w_t c_t(x)_{v_i})^2 + (w_a c_a(y)_{v_i} - w_t c_t(y)_{v_i})^2}}{2\sqrt{2}}, \tag{4}$$

where SPL_{v_i} represents the strength of the pseudolabel for vertex v_i. The numerator calculates the Euclidean distance between the arousal and valence coordinates obtained

[2] We abuse the notation of edges and weights in a graph for simplicity in explaining the method.

from the audio and text modalities. The fraction's denominator represents a normalization by the maximum possible distance in the circumplex arousal-valence diagram. By subtracting the normalized distance from 1, we obtain the pseudolabel (SPL) strength, where a higher value indicates closer proximity between the coordinates in both modalities. Moreover, the SPL incorporates the importance of the audio and text modalities, denoted by w_a and w_t, respectively. Both weights must be defined in the range of [0,1], with the constraint that $w_a + w_t = 1$. Note that reducing the importance of a modality in the context of the circumplex model of emotions involves mapping its output to a neutral emotion. In the arousal-valence diagram, the neutral emotion is represented by coordinates close to the circle's center (0, 0). SPL quantifies the agreement between the audio and text modalities, allowing us to assign stronger pseudolabels to segments with more consistent arousal-valence coordinates across modalities and weaker pseudo labels to segments with larger discrepancies.

Finally, in the third step, we propose a novel multimodal graph regularization to learn the arousal-valence coordinates for all vertices in the graph based on the pseudolabels obtained in the previous step. The MGCP aims to satisfy two key assumptions: (i) the final valence-arousal coordinates of nearby vertices in the graph should be close in the emotion circumplex, and (ii) the higher the strength of the pseudolabel (SPL) assigned to a vertex, the closer the initial and final valence-arousal coordinates estimated by the model [Zhou et al. 2003, Rossi et al. 2014]. The general form of the regularization function in MGCP, which seeks to find a matrix $\mathbf{F}_{|V|\times 2}$ of final valence-arousal coordinates that minimize $Q(\mathbf{F})$, is presented in Equation 5. Each line i of the matrix $\mathbf{F}$ represents the arousal-valence coordinate $\mathbf{f}_{v_i}$ of the segment s_i, represented by the vertex v_i.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{v_i, v_j \in V} w_{v_i,v_j} \Omega(\mathbf{f}_{v_i}, \mathbf{f}_{v_j}) + \mu \sum_{v_k \in V} SPL_{v_k} \Omega(\mathbf{f}_{v_k}, \mathbf{y}_{v_k}) \tag{5}$$

In Equation 5, the first term measures the distance $\Omega(.)^3$ between the estimated valence-arousal coordinates of neighboring vertices, satisfying the first assumption. Here, $\mathbf{f}_{v_i}$ and $\mathbf{f}_{v_j}$ represent the valence-arousal coordinates estimated by MGCP, and w_{v_i,v_j} denotes the weight of the relationship between segments in the multimodal graph G_m.

The second term calculates the proximity between the estimated valence-arousal coordinates by MGCP, $\mathbf{f}_{v_i}$, and the average of the original valence-arousal coordinates obtained from the unimodal models, $\mathbf{y}_{v_i}$. We aim to minimize this term to satisfy the second assumption. We introduce the term SPL_{v_k} to give higher weight to segments with stronger pseudolabels, indicating greater consensus and a desire to preserve the coordinates obtained from the unimodal models. On the other hand, if the pseudolabel strength is low, the MGCP is more likely to seek new final valence-arousal coordinates since there will be less penalty during the minimization. Additionally, the parameter μ allows us to control the importance weight of the second assumption globally [do Carmo and Marcacini 2021].

The proposed MGCP minimizes Eq. 5 through iterative solvers based on label propagation [Zhu and Goldberg 2022], where segments gradually propagate their

[3] Our method uses an iterative version to minimize $Q(\mathbf{F})$, which $\Omega(.)$ is analogous to a Euclidean distance.

valence-arousal coordinates to neighboring segments considering their weights and SPL values. The process continues for a predefined number of iterations or until it reaches an equilibrium state in the graph, where there are no further changes in the values of $\mathbf{f}_v$ for each segment. Specifically, we adapted the GFHF (Gaussian Fields and Harmonic Functions) method proposed by [Zhu et al. 2003]. In the next section, we will present the details of the experimental evaluation of the MGCP method, including the dataset description, evaluation metric, and comparisons with other approaches.

4. Experimental Evaluation

4.1. Dataset Description

There is limited availability of labeled multimodal emotion audio datasets with access to raw data. Among the available datasets, we chose the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset for our experimental evaluation [Busso et al. 2008]. The IEMOCAP dataset consists of dyadic sessions where actors engaged in scripted and improvised scenarios while their audio signals were captured, thereby providing a collection of emotional expressions in speech.

The IEMOCAP dataset contains approximately 12 hours of data. We utilized 10,039 audio segments, which were manually labeled with nine emotion categories: ’sadness’, ’frustration’, ’neutral’, ’happiness’, ’excitement’, ’surprise’, ’anger’, ’fear’, and ’other’. Table 1 presents the distribution of the segments and the corresponding duration in hours for each emotion category.

Table 1. Distribution of Segments and Duration for Each Emotion Category

Emotion	**Segments**	**Duration (hours)**
Sadness	1250	1.84
Frustration	2917	3.64
Neutral	1726	1.87
Happiness	656	0.78
Excitement	1976	2.47
Surprise	110	0.10
Anger	1269	1.61
Fear	107	0.10
Other	28	0.03
Total	**10039**	**12.44**

4.2. Experimental Setup

This section presents the methods used for comparison with the proposed MGCP method. We provide an overview of each method, including unimodal and multimodal approaches.

4.2.1. Unimodal Methods

Audio Modality: We utilized the HuBERT-Emotion[4] model for the audio modality, which is a HuBERT pre-trained model fine-tuned on the CREMA-D dataset. The

[4] Model available at: `https://huggingface.co/Rajaram1996/Hubert_emotion`

CREMA-D dataset consists of over 7,000 labeled samples of emotional speech from various actors.

Text Modality: In the text modality, we employed the RoBERTa-GoEmotions[5] model, a pre-trained RoBERTa model fine-tuned on the GoEmotions dataset. The GoEmotions dataset contains a large collection of Reddit comments annotated with 27 emotion categories.

4.2.2. Multimodal Methods

Late Fusion (Decision Level): As a reference method for comparison, we adopted the late fusion strategy via decision level. In this approach, the unimodal models' output (arousal-valence coordinates) is merged through a weighted average. We varied the weights assigned to each modality, ranging from 0.1 to 0.9 during the late fusion process.

MGCP: We evaluate the proposed method using different weights for each modality, ranging from 0.1 to 0.9 (i.e. parameters w_a and w_t of the Equation 4). Additionally, we explored different parameter μ values (see Equation 5), ranging from 0.1 to 0.9. During minization of the regularization function using label propagation, we use a maximum of 15 iterations.

4.3. Evaluation Criteria

In order to assess the performance of the proposed MGCP method and compare it with other approaches, we employ the Mean Squared Error (MSE) as the evaluation metric.

Let $c(x)^*_{v_i}$ and $c(y)^*_{v_i}$ represent the predefined arousal-valence coordinates of segment s_i represented by vertex v_i (ground truth). Similarly, let $c(x)_{v_i}$ and $c(y)_{v_i}$ denote the estimated arousal-valence coordinates of the same segment obtained from the MGCP method. The MSE for segment s_i is then calculated according to Equation 6,

$$\mathit{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left((c(x)^*_{v_i} - c(x)_{v_i})^2 + (c(y)^*_{v_i} - c(y)_{v_i})^2\right) \tag{6}$$

where n represents the total number of segments in the dataset. The MSE measures the average of the squared differences between the predefined arousal-valence coordinates and the estimated coordinates for all segments in the dataset.

5. Results and Discussion

The experimental evaluation considers three aspects. Firstly, we assess the importance of each modality during fusion by analyzing the weights assigned to the textual and audio modalities. Secondly, we evaluate the significance of pseudolabeling in the MGCP method. Finally, we comprehensively compare the proposed MGCP method with multimodal late fusion (decision level) and the unimodal models.

In the first aspect, we analyze the weights assigned to each modality. We compare these weights in the proposed MGCP method and the multimodal late-fusion (decision

[5] Model available at: `https://huggingface.co/bsingh/roberta_goEmotion`

level) reference model. The results show that combining audio and text modalities leads to lower Mean Squared Error (MSE) values, indicating a substantial performance improvement. This finding aligns with the core principle of multimodal learning, wherein the exploitation of complementarity between audio and text enhances emotion recognition. Notably, the proposed MGCP method outperforms the reference model, reinforcing its efficacy in integrating multimodal information. Figure 2 presents these results for MGCP with $\mu = 1.0$. On the x-axis, we represent the level of multimodality importance, where values close to zero indicate a higher importance for the textual modality (i.e. $w_t \approx 1$), while values close to one indicate a higher importance for the audio modality (i.e. $w_a \approx 1$).

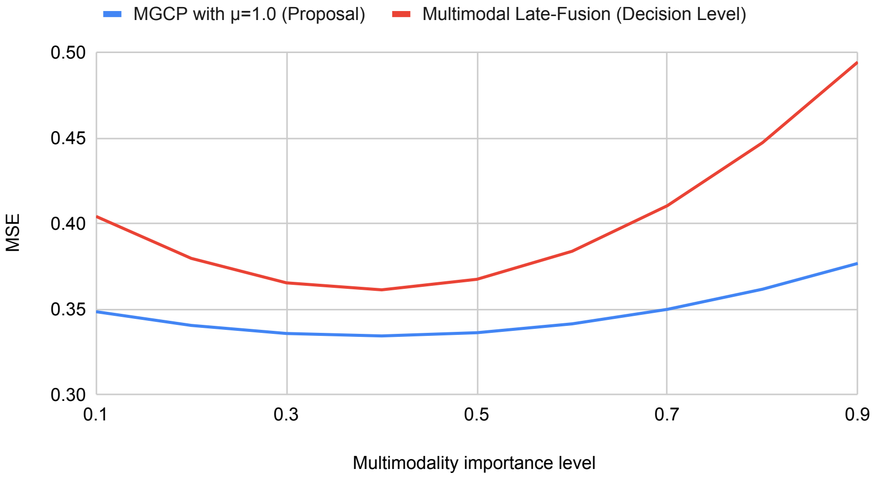


Figure 2. Combination of audio and text modalities leads to lower Mean Squared Error (MSE) values for MGCP and Multimodal Late-Fusion, indicating a substantial improvement in performance.

Regarding the second aspect (Figure 3), we explore the effect of Pseudolabel Strength (SPL) by thoroughly analyzing parameter μ of the MGCP method. To visually depict this analysis, we present a heatmap graph that correlates the μ parameter with the importance of each modality. Lower values of μ tend to prioritize the individual modalities' predictions, placing relatively less emphasis on the consensus of pseudolabels. On the other hand, higher values of μ assign greater significance to the consensus, enabling the model to harness the collective agreement between the modalities. However, higher values of the μ parameter in the MGCP method lead to results comparable to those of the multimodal reference model, as they essentially function as a weighted average of the outputs from each unimodal model. Thus, the strength of the MGCP lies in its ability to leverage the consensus of the unimodal outputs while providing flexibility to adapt to the graph's topology and data characteristics.

Through experimental analysis, we observed that the MGCP method achieves lower Mean Squared Error (MSE) values when the μ parameter is defined within the range of 0.1 to 0.5. This range allows the model to effectively exploit the consensus among the modalities while considering the variations and complexities within the data.

Next, in the third aspect, we compare the Mean Squared Error (MSE) values

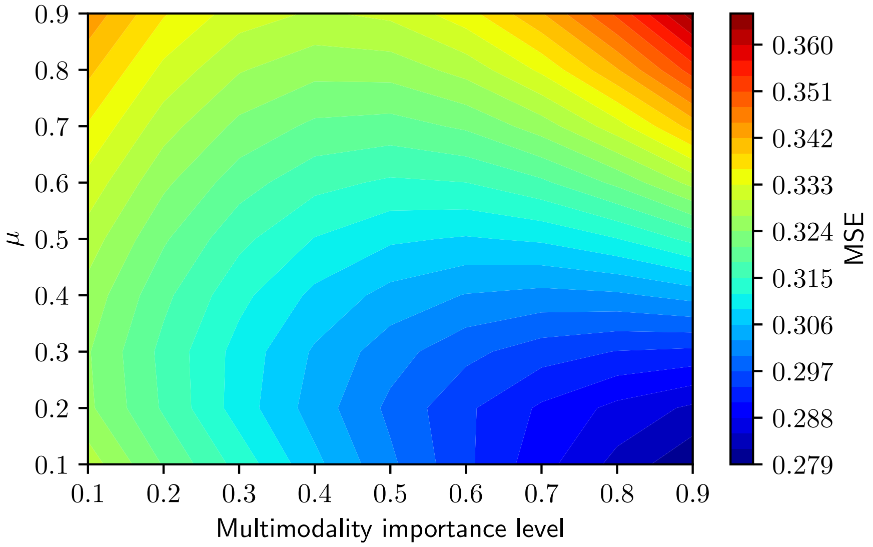


Figure 3. Heatmap graph that correlates MSE values considering the μ parameter and the importance of each modality in the MGCP proposed method.

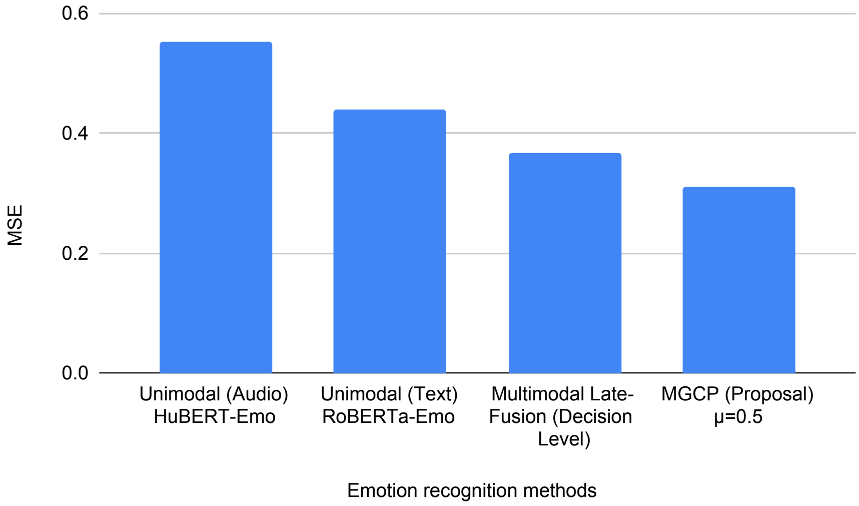
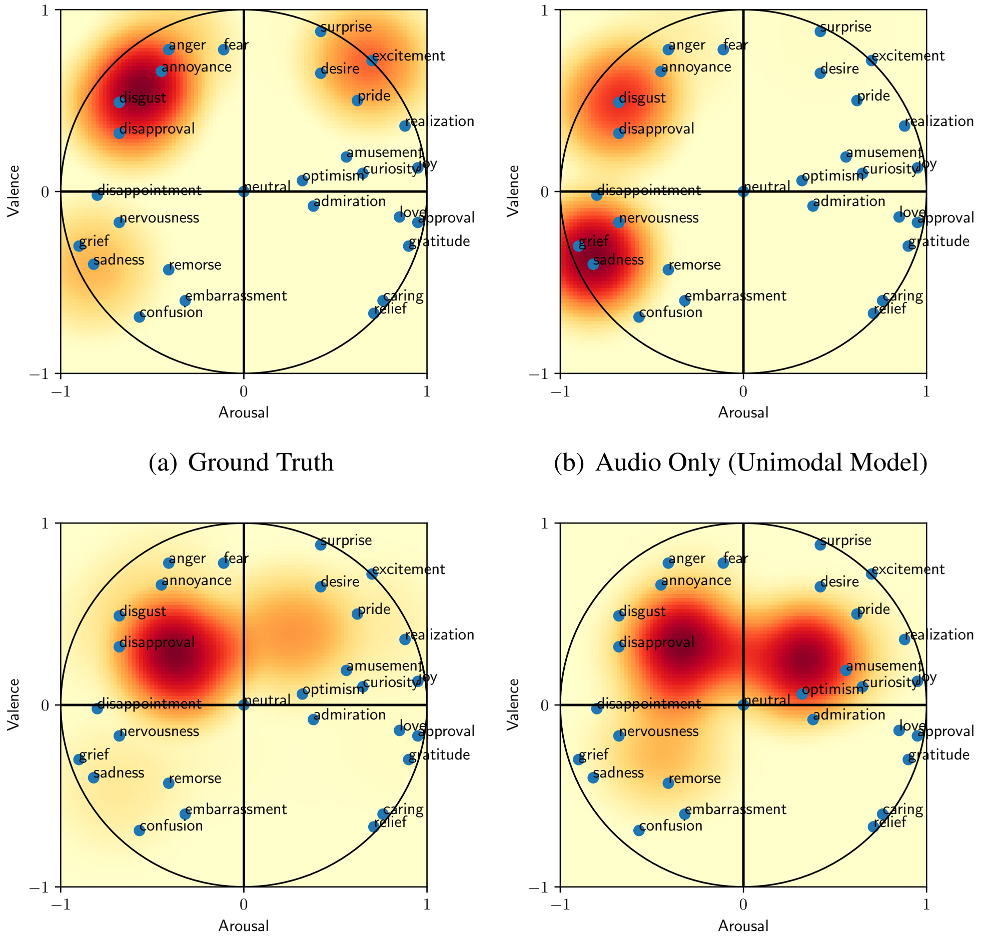


Figure 4. Comparison of the Mean Squared Error (MSE) values among various multimodal and unimodal approaches.

among various multimodal and unimodal approaches. This analysis is particularly challenging due to the absence of labeled data in the evaluation set, making it difficult to define a validation set for parameter estimation. Therefore, for this aspect, we rely on the default parameters of the MGCP method, namely $\mu = 0.5$, and equal weights of 0.5 assigned to each text and audio modality. Although these may not be the optimal experimental parameters for the MGCP method, we chose the defaults to ensure a fair comparison with other models.

The experimental evaluation shows that the MGCP method outperforms all other models, achieving the lowest MSE values, as illustrated in Figure 4. This finding high-



(a) Ground Truth

(b) Audio Only (Unimodal Model)

(c) Audio+Text (Multimodal Model)

(d) MGCP (Multimodal Model)

Figure 5. Heatmap visualization of the arousal-valence coordinates (circumplex model) obtained from (a) ground truth data, (b) audio-based unimodal model, (c) multimodal model (late fusion), and (d) the proposed MGCP method.

lights the effectiveness of the proposed method in accurately capturing and representing emotions. Furthermore, we provide a heatmap visualization of the arousal-valence coordinates (circumplex model of emotions) obtained by the unimodal audio-only model, multimodal audio+text model, the proposed MGCP method, and a reference heatmap generated using manually annotated data (ground truth). By comparing these heatmaps, it becomes apparent that the MGCP method produces results that are visually closer to the ground truth.

6. Concluding Remarks and Future Work

In this paper, we presented the MGCP (Multimodal Graph-based Consensus Pseudolabeling) method for emotion recognition in audio. Combining embeddings from pre-trained unimodal models and constructing a multimodal graph, the MGCP method effectively captures the semantic meaning of audio segments and their associated text. Our approach leverages pseudolabeling to generate arousal-valence data based on the consensus between audio and text modalities. Finally, we extended a regularization framework for graph learning to estimate the final arousal-valence coordinates for all vertices in the multimodal graph. This regularization approach, which considers the agreement between nearby nodes and the strength of the pseudolabels, enables accurate emotion recognition.

We analyzed to determine the significance of each modality in the fusion process, and our findings revealed that the inclusion of both audio and text modalities considerably enhances emotion recognition performance. Additionally, we observed further improvement by incorporating the consensus among the unimodal models through the pseudolabeling technique. However, the MGCP method does have certain limitations. One of the challenges lies in finding appropriate parameters for constructing the multimodal graph and determining the weights assigned to each modality. The default parameters of MGCP often yield satisfactory results, but additional research is needed to explore how to estimate these parameters in the absence of labeled data. This is particularly important when considering scenarios where labeled data is scarce or unavailable.

We propose several directions to advance multimodal emotion recognition further. Firstly, evaluating the MGCP method with additional datasets would provide a broader understanding of its generalization capabilities. Furthermore, exploring the visual modality and incorporating emotion recognition from facial expressions would enhance the multimodal framework. Lastly, analyzing multilingual models for textual and audio modalities would enable the extension of the MGCP method to different languages.

Finally, we would like to emphasize that the source code of the developed method, along with detailed experimental results and datasets, are publicly available on the project's GitHub repository at **`https://github.com/Deflyer/Multimodal-Emotion-Recognition-from-Videos`**.

Acknowledgments

This work was supported by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) grant numbers [88887.671481/2022-00, 88887.710875/2022-00] and São Paulo Research Foundation (FAPESP grant 2022/14903-1). The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the FAPESP grant 2019/07665-4 and from the IBM Corporation. Finally, this work was supported by *Ministério da Ciência, Tecnologia e Inovações*, with funds from Law 8248, of October 23, 1991, PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44.

References

Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., and Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58.

Adoma, A. F., Henry, N.-M., and Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bagadi, K. R. (2021). A comprehensive analysis of multimodal speech emotion recognition. In *Journal of Physics: Conference Series*, volume 1917, page 012009. IOP Publishing.

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Das, R. and Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends and challenges. *ACM Computing Surveys*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Deng, J. J., Leung, C. H., and Li, Y. (2021). Multimodal emotion recognition using transfer learning on audio and text data. In *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part III 21*, pages 552–563. Springer.

do Carmo, P. and Marcacini, R. (2021). Embedding propagation over heterogeneous event networks for link prediction. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4812–4821. IEEE.

Ezzameli, K. and Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, page 101847.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (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.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., and Zhai, X. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Konar, A. and Chakraborty, A. (2015). *Emotion recognition: A pattern analysis approach*. John Wiley & Sons.

Krishna, D. and Patil, A. (2020). Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Interspeech*, pages 4243–4247.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Priyasad, D., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2020). Attention driven fusion for multi-modal emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3227–3231. IEEE.

Rossi, R. G. (2016). *Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo.

Rossi, R. G., Lopes, A. A., and Rezende, S. O. (2014). A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In *Proceedings of the 29th annual acm symposium on applied computing*, pages 79–84.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Saxena, A., Khanna, A., and Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *Interspeech 2019*.

Shah Fahad, M., Ranjan, A., Yadav, J., and Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, 110:102951.

Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. (2020). Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition. *Proc. Interspeech 2020*, pages 3755–3759.

Tomar, P. S., Mathur, K., and Suman, U. (2022). Unimodal approaches for emotion recognition: A systematic review. *Cognitive Systems Research*.

Wang, Y., Boumadane, A., and Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003). Learning with local and global consistency. *Advances in neural information processing systems*, 16.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919.

Zhu, X. and Goldberg, A. B. (2022). *Introduction to semi-supervised learning*. Springer Nature.