Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results

Lucas Goncalves, Ali N. Salman, Abinay R. Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, Carlos Busso

Multimodal Signal Processing (MSP) Lab., Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
{goncalves, ali.salman, abinayreddy.naini, berrak.sisman, busso}@utdallas.edu
Johns Hopkins University
Center for Language and Speech Processing, Baltimore, Maryland, USA
{laureano, tthebaud1, lgarci27, ndehak3}@jhu.edu

Abstract

The Odyssey 2024 Speech Emotion Recognition (SER) Challenge aims to enhance innovation in recognizing emotions from spontaneous speech, moving beyond traditional datasets derived from acted scenarios. It offers speaker-independent training, development, and an exclusive test set, all annotated for the two tracks explored in this challenge: categorical and attribute SER tasks. This initiative promotes collaboration among researchers to develop SER technologies that perform accurately in real-world settings, encouraging researchers to explore innovative approaches that leverage the latest advancements in audio processing for SER. In this paper, we provide a detailed description of the baseline, leaderboard, evaluation of the results, and a discussion of the key findings. The competition website with leaderboards, links to baseline code, and instructions can be found here: https://lab-msp.com/MSP-Podcast_Competition/leaderboard.php

1. Introduction

The Odyssey 2024 Speech Emotion Recognition (SER) Challenge [1] represents a significant step forward in the pursuit of more sophisticated and real-world deployable SER systems. Central to this challenge is the objective to foster collaboration and innovation in SER research, particularly in naturalistic environments, where traditional datasets have often fallen short. Unlike conventional SER datasets that are predominantly derived from controlled, acted scenarios, the MSP-Podcast corpus—employed in this challenge offers a rich, diverse, and authentic compilation of naturalistic conversational data. With over 237 hours of annotated audio available in the entire corpus, we are able to provide for this challenge a speaker-independent publicly available train and development sets, and an exclusive non-public test set. The audios in this dataset are annotated by at least five annotators each. We also include gender, speaker id, and human transcriptions for sentences in the train and development sets, thereby offering a robust foundation for participants to develop, test, and benchmark novel SER methodologies.

This challenge underscores the critical need for advancements in SER that can operate effectively within the complexities of natural human communication, moving beyond the limitations of traditional acted datasets. By focusing on real-world conversational data, the Odyssey 2024 SER Challenge aims to address the nuanced dynamics of spontaneous speech, including *variability in emotion, intonation*, and *context*, which are often absent in controlled environments. Participants are encouraged to explore innovative approaches that leverage the latest advancements in machine learning, signal processing, and computational linguistics, with the goal of achieving higher levels of accuracy and reliability in emotion recognition.

Furthermore, the challenge facilitates an open forum for academic and industry professionals to share insights, methodologies, and breakthroughs. This collaborative environment is expected to not only propel the field of SER forward but also pave the way for practical applications in various domains where SER can be applied, such as mental health assessment, customer service optimization, and human-computer interaction. In essence, the Odyssey 2024 SER Challenge is not just a competition; it is a concerted effort to bridge the gap between theoretical research and practical, real-world applications of SER technology. Through this challenge, we aim to inspire a new wave of research that is firmly rooted in the complexities and richness of natural human communication, setting a new standard for future endeavors in the field.

2. Challenge Tracks

This challenge has two tracks: (1) Categorical Emotion Recognition and (2) Emotional Attributes Prediction.

Categorical Emotion Recognition: The task involves classifying each sample into one of eight emotional classes: anger, happiness, sadness, fear, surprise, contempt, disgust, and a neutral state. The ground truth for each sample is determined by applying the plurality rule among all annotations, ensuring that each sample is categorized into the emotional class it most closely aligns with. The challenge's test set is carefully constructed to ensure a balanced representation of all eight emotional categories. Performance in this track is evaluated using the Macro-F1 score as the primary metric.

Emotional Attributes Prediction: The task focuses on predicting a range of emotional attributes, including arousal (from calm to active), valence (from negative to positive), and dominance (from weak to strong). In this track, emotional attributes are quantified on a continuous scale from 1 to 7 across each

[1] https://www.odyssey2024.org/emotion-recognition-challenge

dimension, based on annotations provided by at least five annotators per sample. A ground truth value for each sample is established by averaging the scores reported by all annotators, facilitating a nuanced assessment of emotional states.

3. Dataset

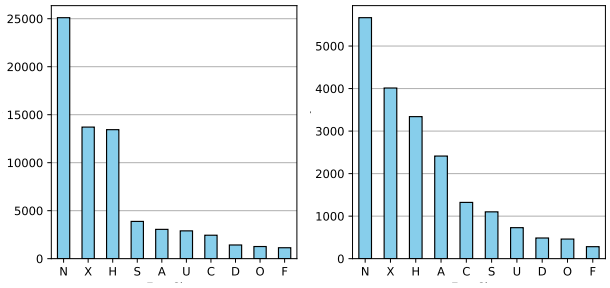
The MSP-Podcast corpus [1], contains spontaneous and diverse emotional speech samples collected from various podcast recordings, which are split into speaking turns to form a speech repository. Several SER algorithms are used to retrieve speaking turns that are expected to be emotional by using the approach presented in Mariooryad et al. [2]. The annotation process uses a crowdsourcing protocol inspired by the work of Burmania et al. [3]. Two types of evaluation setups are used during annotation: categorical and attributes Dimension. The perceptual evaluation for categorical emotions includes the *primary emotions* (P) and *secondary emotions* (S). The annotators choose a single primary emotion, but they can select multiple secondary emotions for each sample. The primary emotions contain nine options: anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other". The secondary emotions consist of the primary emotions and eight more classes: amusement, frustration, depression, concern, disappointment, excitement, confusion, and annoyance (17 options in total). Refer to Figure 1 for information about the emotional class distribution present in the train and development sets.

The perceptual evaluation for emotional attributes dimension consists of raters annotating the samples using a seven-point Likert scale where the attributes annotated are arousal (calm to active), dominance (weak to strong), and valence (negative to positive). We average the scores provided by raters for each sample to establish its ground truth values, the distributions for the emotional attribute in the train and development set are depicted in Figure 2 and3, respectively.

Each speaking turn is annotated by at least five different workers for both the categorical emotions and the attributes dimension emotions. The version of the MSP-Podcast corpus utilized in this study is a subset of release 1.11. The only distinction of this subset from the original release 1.11 is its exclusion of sentences without speaker ID information. Overall, the subset contains the following settings: The train set has 68,360 speaking turns from 1,405 speakers, the development set has 19,815 speaking segments from 454 speakers, and the test set comprises 2,347 unique segments from 187 speakers. The test set corresponds to the "test 3" set in the release 1.11 of the MSP-Podcast corpus. The annotations for this set have not been made publicly available. More details about the train and development set are shown in Table 1. The segments for the test set have been curated to maintain a balanced representation based on primary categorical emotions. Further details on the protocol used to collect this corpus are described in Lotfian and Busso [1].

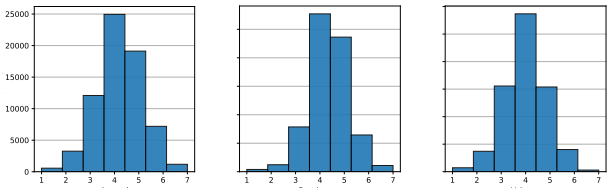
Table 1: Distribution of gender and number of speakers for speech samples across the Train and Development sets.

Set	Male	Female	Speakers
Train	37,370	30,990	1,405
Development	10,525	9,290	4,54



(a) Training Set (b) Development Set

Figure 1: Distribution of categorical emotions across training and development sets. A=Anger, C= Contempt, D= Disgust, F= Fear, H= Happiness, N= Neutral, S= Sadness, U= Surprise



(a) Arousal (b) Dominance (c) Valence

Figure 2: Overall emotional attributes distribution for training set.

4. Baseline

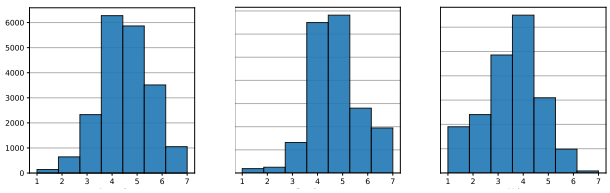
The overall structure of our baseline [2] is illustrated in Figure 4 and consists of two main modules. The Fine-Tuned (FT) module integrates components from a pre-trained Self-Supervised Learning (SSL) model, specifically WavLM [4].

In the final stage of our approach, termed the Head, we utilize pooling layers followed by Fully Connected (FC) layers for prediction. These pooling layers employ attentive statistics pooling [5], which utilize an attention mechanism to allocate varying weights to different frames. This technique enables the calculation of both weighted means and standard deviations for the frames coming from the SSL layers. The processed output from the pooling layer is subsequently fed into a series of FC layers for the final prediction.

4.1. Implementation Details

We use a pre-trained SSL model to conduct our experiments. Specifically, we are using the large version of WavLM [4].

[2]https://github.com/MSP-UTD/MSP-Podcast_Challenge



(a) Arousal (b) Dominance (c) Valence

Figure 3: Overall emotional attributes distribution for development set.

This model contains 24 transformer layers and is comprised of ∼310M parameters. We utilized the pre-trained off-the-shelf models from hugging face [6] "microsoft/wavlm-large" for WavLM. As evidenced in previous studies [7, 8, 9, 10] fine-tuning SER models from pre-trained SSL models can lead to a significant boost in performance. We fine-tune this model for 30 epochs, with a learning rate set to 1e-5, batch size of 32, and Adam as our optimizer. For emotion classification [3] (task 1), to address the issue of class imbalance, our training objective utilizes a weighted multi-class cross-entropy (CE) loss. We employ a weighted loss function to address the class imbalance problem (while the train and development set are unbalanced across emotions, the test set is balanced). This approach assigns more significance to the less frequent classes. Specifically, for tasks like ours that use the CE loss for classification, we adjust the weight parameter to reflect the inverse frequency of each class. This means assigning higher weights to less frequent classes, therefore, enhancing the model's sensitivity and performance on these classes. The weighted CE loss is defined as follows:

$$\mathcal{L}_{WCE} = -\sum_{c=1}^{M} w_c \cdot y_{o,c} \log(p_{o,c}) \quad (1)$$

where M represents the number of classes, log denotes the natural logarithm, c is the correct label for observation o, $p_{o,c}$ is the predicted probability that observation o belongs to class c, $y_{o,c}$ is 1 when observation o belongs to class c, and 0, otherwise, and w_c is the weight assigned to class c, reflecting its inverse frequency. This modification enhances the model's sensitivity and performance on less frequent classes by assigning higher weights to them.

For emotion attributes prediction (task 2) [4], we use a single task setup, where we train a separate regression model for arousal, valence, and dominance. For the loss on emotional attributes regression models, we use the *concordance correlation coefficient* (CCC), which measures the agreement between the true and predicted emotional attribute scores. Equation 2 illustrates the CCC measurements,

$$\mathcal{L}_{\text{CCC}} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2)$$

where μ_x and μ_y represent the means of the actual and predicted scores, respectively. Similarly, σ_x and σ_y denote the standard deviations of these scores. The term ρ corresponds to the Pearson correlation coefficient between the true and predicted scores. Our model's training objective is to optimize the CCC, aiming to achieve a high correlation between predicted and actual scores while minimizing prediction errors.

4.2. Baseline Results

Table 2 presents the results obtained on both tasks from the test set of our challenge with our proposed baseline. For categorical emotion recognition over the eight explored emotion classes, we obtain a F1-micro score of 0.311 and a F1-macro score of 0.327. For attribute emotion recognition we achieve CCC scores of 0.567, 0.607, and 0.424 for arousal, valence, and dominance respectively.

[3] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Categorical

[4] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Multi-Attributes

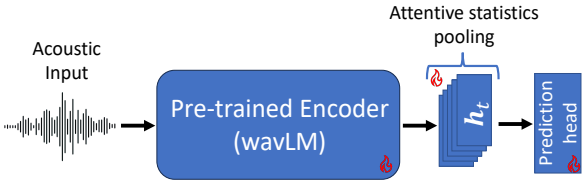


Figure 4: Baseline framework overview.

5. Results and Findings

This section presents a comprehensive overview of the strategies, technical advancements, and shared challenges encountered by participating teams. In total, 47 teams signed up for the challenge and requested the corpus. Out of these 47 teams 31 submitted to task 1 and 13 submitted to task 2. At the close of the challenge submission period, we reached out to the participants, inviting them to participate in a survey. The purpose was to gather detailed information about their implementations, allowing us to gain insights into their proposed methods. This will enable us to offer a comprehensive comparison of the various approaches. Table 3 shows a brief compilation of team's approaches who have responded to our survey. Teams that have responded to our survey focused on task1 for the most part with the exception of two teams focusing on both task 1 and task 2 and using the same approach for both tasks (NU and TalTech). Therefore, our analysis is focused on the categorical emotion recognition aspects of the models. This analysis presents descriptions of the approaches used by teams that have provided a description of their approaches and ranked prominently in both the categorical and emotional attributes recognition tasks, providing insights into the state-of-the-art in speech emotion recognition.

In assessing the performance of the models on the test set, we employ distinct metrics for each track. For categorical emotion recognition, we present both micro and macro F1-scores. The micro F1-score aggregates the counts of true positives, false negatives, and false positives across all categories, offering a metric that reflects sensitivity to class imbalance. In contrast, the macro F1-score individually computes the F1-score for each category before averaging them, ensuring equal contribution from each category to the overall score, regardless of its occurrence frequency. This approach offers a detailed perspective on the models' capacity for emotion classification, underlining their proficiency in identifying and accurately classifying a diverse array of emotional states. For the emotion attributes prediction track, we employ the CCC to measure the models' capability to predict the emotional dimensions of arousal, valence, and dominance.

To determine the final overall ranking of the models, we use the Macro F1 scores for the categorical emotion recognition and the average CCC values across arousal, valence, and dominance for the emotion attributes prediction track. This methodology ensures a comprehensive evaluation of the models' performance, emphasizing their ability to recognize and quantify a broad spectrum of emotional states. Results and rankings for teams in task 1 and task 2 are shown in Table 4 and Table 5, respectively. Task 1 top-3 scores (based on F1-Macro) were achieved by Sheffield-MINI with 0.3569 (.046 over Baseline), TalTech with 0.3543 (.043 over Baseline), and UPC-BSC with 0.3441 (.033 over Baseline). Task 2 top-3 scores (based on CCC

Table 2: Baseline performance on categorical emotion recognition and emotional attributes recognition.

Categorical Emotions

Model	F1-Macro	F1-Micro
Baseline	0.311	0.327

Emotional Attributes

Model	Arousal	Valence	Dominance
Baseline	0.567	0.607	0.424

Table 3: Summary of approaches.

Team	**Inputs**	**Features**	**Loss**	**Aug.**
Sheffield-MINI	Audio/Text	SSL-Based	Focal/CE	No
TalTech	Audio/Text	Fbank	CE/CCC	Yes
UPC-BSC	Audio/Text	Raw Audio	Focal/CE	Yes
THAU	Audio/Text	Fbank	CE	No
CONILIUM	Audio	Raw Audio	CE	Yes
L'antenne du Ventoux	Audio/Text	Raw Audio/text	Negative log likelihood / Jeffreys loss	Yes
NU	Audio/Text	SSL-Based	MSE/CE	No
Baseline	Audio	Raw Audio	CE/CCC	No
Team AGH	Audio/Metadata	SSL-Based	Log	No
Vivolab	Audio	SSL-Based	CE	Yes
VicomSpeech	Audio	Raw Audio	Hinge (SVM)	No

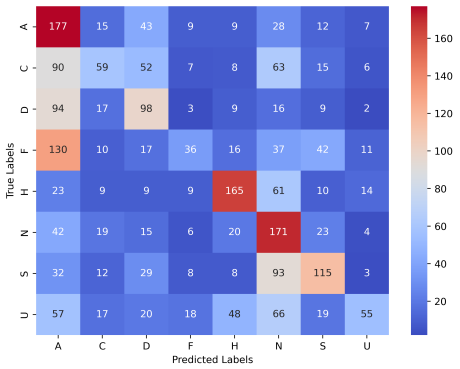
average over arousal, valence, and dominance CCC scores) were achieved by the Baseline with 0.5327, AIST-BahasaKita with 0.5297, and intema.ai with 0.529.

5.1. Model Architectures and Features

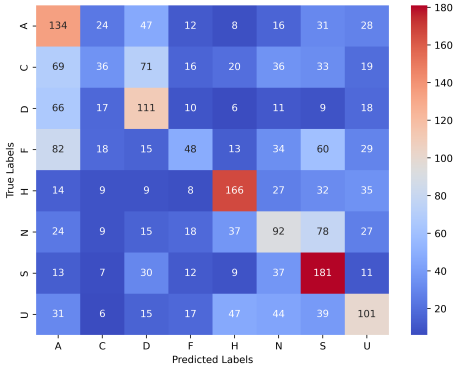
The challenge showcased a diversity in model architectures and feature sets. Sheffield-MINI utilized an ensemble approach combining self-attention layers and MLPs, employing features extracted from both audio and text using pre-trained models like Whisper [11], Wav2vec [12], HuBERT [13], WavLM [4], and RoBERTa [14]. Similarly, UPC-BSC, THAU, and NU integrated textual data with audio features, leveraging pre-trained models and fine-tuning strategies for enhanced emotion recognition capabilities. Team NU utilized SSL-based features from BERT [15] and Emotion2Vec [16] in addition to WavLM features. In contrast, teams like Vivolab, CONILIUM, and VicomSpeech focused solely on audio data, employing models like WavLM [4] with innovative architectures like second-order attention mechanisms [17] and *support vector machine* (SVM) [18] classifiers for emotion detection. CONILIUM utilized a structure that mirrored the baseline and made use of secondary annotations of all workers (annotators) in the weighed binary CE loss function. TalTech opted for features derived from mel-scale filters applied to the spectrogram, such as 80 Mel-frequency filter-banks as their acoustic inputs.

5.2. Data Modalities and Augmentation Techniques

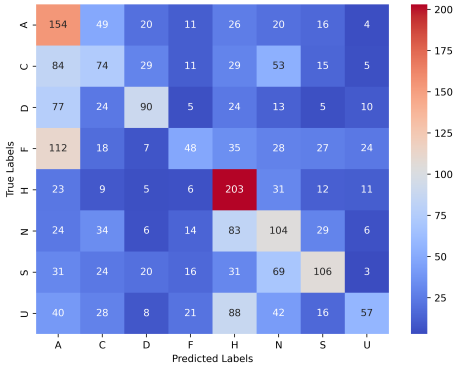
At least six teams utilized audio and textual input, with the transcriptions often derived from pre-trained speech models like Whisper [11]. The usage of text alongside audio indicates a trend towards exploiting multimodal data for improved emotion recognition. Data augmentation was also employed, with techniques ranging from speed perturbation and noise addition (TalTech, UPC-BSC, and Vivolab) sourced from datasets like



(a) Sheffield-MINI Model



(b) TalTech Model



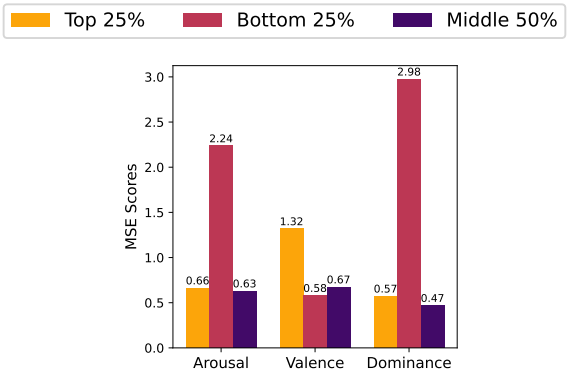
(c) UPC-BSC Model

Figure 5: Confusion matrix of top-3 performing categorical emotion recognition models. A=Anger, C= Contempt, D= Disgust, F= Fear, H= Happiness, N= Neutral, S= Sadness, U= Surprise

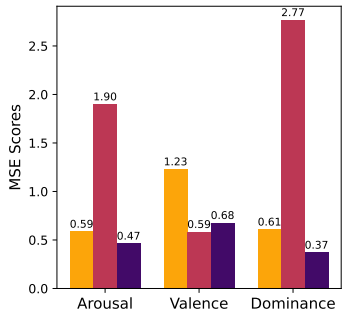
MUSAN [19] and RIR [20] to the approaches like redistribution of samples labeled 'X' using annotator rankings (L'Antenne du Ventoux) or focusing on high-quality training data determined by a basic analysis of each worker (annotator) performance compared to others (Team AGH). Some teams, like Sheffield-MINI, opted not to use augmentation, highlighting the variance in strategies to tackle overfitting and data imbalance.

5.3. Training Process and Optimization

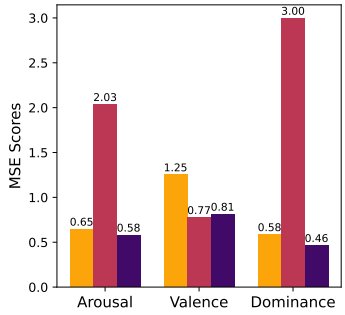
Training methodologies varied significantly, different loss functions from the ones used for the baseline were noted in some



(a) Baseline Model



(b) AIST-BahasaKita Model



(c) Intema Model

Figure 6: Mean Squared Error across the top, bottom, and middle percentiles of emotional attributes for top-3 models.

approaches. Specifically, teams Sheffield-MINI and UPC-BSC made use of Focal Loss [21] to address class imbalance and CE loss. NU optimized their models using CE loss and MSE loss. Team L'antenne du Ventoux utilized Negative log-likelihood and Jeffreys loss during training to jointly fine-tuned both speech and text encoders. TalTech's logistic regression-based fusion of audio and text-based models was implemented using *low-rank adaptation* (LoRA) [22] to finetune Llama 2 [23] and Wav2Vec-BERT [24]. TalTech then used the posterior probabilities of audio and text-based category prediction from the models and fused them using a separate logistic regression model, trained on development data representing another strategic approach to leveraging multimodal data.

Table 4: Task 1 categorical emotions ranking and results.

Pos.	Team	F1-Macro	F1-Micro
1	Sheffield-MINI	0.3569	0.3732
2	TalTech	0.3543	0.3703
3	UPC-BSC	0.3441	0.3562
4	THAU	0.3426	0.3528
5	IRIT/MFU	0.3367	0.3515
6	CONILIUM	0.3350	0.3473
7	VK Lab	0.3340	0.3502
8	L'antenne du Ventoux	0.3318	0.3473
9	thuhcsi	0.3303	0.3409
10	SUST-AI4AC-TEAM	0.3267	0.3430
11	NU	0.3256	0.3443
12	BIIC Lab	0.3142	0.3251
13	Baseline	0.3113	0.3272
14	Team AGH	0.3093	0.3247
15	AstroSpeech	0.2984	0.3132
16	Vivolab	0.2942	0.3076
17	VicomSpeech	0.2938	0.3247

Table 5: Task 2 emotional attributes ranking and results.

Pos.	Team	Val.	Aro.	Dom.	Avg.
1	Baseline	0.6069	0.5667	0.4244	0.5327
2	AIST-BahasaKita	0.6025	0.5821	0.4046	0.5297
3	intema.ai	0.5830	0.5846	0.4194	0.529
4	TJU_SER	0.5768	0.5823	0.4152	0.5248
5	NU	0.6014	0.5608	0.4028	0.5216
6	TalTech	0.6362	0.5417	0.3655	0.5144
7	SMILE	0.6197	0.554	0.3436	0.5058
9	sensein group	0.5332	0.5455	0.3706	0.4831
10	thuhcsi	0.5984	0.5057	0.3115	0.4719
11	mosaic	0.5966	0.5075	0.3032	0.4691
12	CASIA-MAIS	0.5507	0.495	0.2764	0.4407
14	dj-ulm	0.4550	0.4946	0.2909	0.4135
15	AstroSpeech	0.3327	0.4306	0.2263	0.3299

5.4. Challenges and Solutions

A common challenge across submissions was dealing with class imbalance and the general difficulty of emotion recognition from speech. Sheffield-MINI's focal loss and Vivolab's balance through sample equalization are direct responses to class imbalance. The challenge of accurately classifying emotions with subtle distinctions was tackled through innovative model architectures and feature sets, as seen in UPC-BSC's use of double multi-head attention component [25] and THAU's use of *large language models* (LLM) [26] for fusion. The computational resources required varied, with most teams employing high-end GPUs like the NVIDIA RTX 3090 or A100. This observation indicates the computationally intensive nature of training sophisticated models for speech emotion recognition, highlighting an entry barrier for researchers without access to such resources.

5.5. Top-3 Performing Models Evaluation

In Figures 5 and 6, we take a deeper look at the performance of the top3 performing models for each task to analyze with emotional classes or attribute regions are more challenging for these approaches.

5.5.1. Categorical Emotion Recognition Analysis

Figure 5 depicts the confusion matrices from the speech emotion recognition challenge for Task 1: categorical emotion recognition offering a deeper look into the performance of the top-3 models—Sheffield-MINI, TalTech, and UPC-BSC—to gain insights on their strengths and areas needing improvement in classifying various emotional categories.

The Sheffield-MINI Model demonstrates a strong ability to recognize 'Anger', 'Happiness', 'Neutral', and 'Sadness' emotions, as evidenced by the highest number of correct predictions in these categories. However, it appears to struggle with 'Fear', 'Contempt', and 'Disgust' emotions, where a substantial number of instances are misclassified as 'Anger', indicating a possible area of confusion between these categories. Additionally, 'Surprise', 'Contempt', 'Sadness', and 'Happiness' categories show a mix of misclassifications with a lot of 'Neutral' misclassifications, suggesting a challenge in distinguishing these emotions.

The TalTech Model shows a somewhat balanced classification ability but has notable confusion between 'Neutral' and 'Sadness', with many 'Neutral' emotions being incorrectly labeled as 'Sadness'. This model also seems to face difficulty accurately classifying 'Contempt' emotions, often mislabeling them as 'Disgust' or 'Anger', and to a larger extent, 'Fear' emotions being misclassified as 'Anger' or 'Sadness'.

The UPC-BSC Model, while proficient at identifying 'Happiness' emotions with a high number of true positives, has a prominent issue distinguishing between 'Anger' and other emotions like: 'Contempt', 'Disgust', and 'Fear', with a significant number of 'Contempt', 'Disgust', and 'Fear' labeled as 'Anger'. Moreover, this model also misclassifies a considerable amount of 'Surprise' emotions as 'Happiness', revealing weaknesses in recognizing 'Surprise.

From an overarching perspective, all three models exhibit a common challenge in correctly identifying 'Disgust', 'Fear', and 'Contempt' with 'Anger', which could be due to intrinsic similarities in the emotional expression of these categories that the models are unable to disentangle. Furthermore, 'Surprise' seems to be a difficult category for all models, with misclassifications scattered across other emotions, underscoring a widespread difficulty in distinguishing subtle nuances in speech that may signify 'Surprise'.

These insights reveal key challenges in the development of categorical emotion recognition systems, particularly in differentiating between emotions with closely related acoustic features. The confusion matrices highlight specific pairs of emotions where models are more likely to confuse, which can direct researchers toward targeted improvements, such as feature engineering to better capture the distinctive aspects of each emotional state or refining the training process to address these confusions. The results underscore the necessity of enhancing model sensitivity to the complex and nuanced nature of human emotional expression in speech.

5.5.2. Attributes Emotion Recognition Analysis

In the context of Task 2 of the speech emotion recognition challenge focused on attribute emotion recognition, an the results shown in 6 depicts results from our experiment conducted to assess the capability of models to predict emotional attributes—arousal, valence, and dominance—in speech focused at different ranges of attributes. We employed the *mean squared error* (MSE) to gauge the error of model predictions with the ground truth. We conducted these experiments to gain insights into the regions of emotional attributes that these models are struggling more. Here we segmented the emotional intensity for each attribute into the top 25% percentile, bottom 25% percentile, and the middle 50% attribute values based on the ground truth of the challenge's test set. All top-3 models essentially have the same patterns as the baseline model, displaying lower MSE scores with midrange intensity attributes and attributes in the top 25% while facing extreme difficulties with bottom 25% attribute intensity prediction for arousal and dominance. The AIST-BahasaKita Model emerged as more proficient in the midrange for Arousal and in the top 25% for Dominance. The Intema Model mirrored the baseline for the most part with slight lower MSE in the top 25% intensities for Dominance predictions.

As mentioned, a common trend across all models was their lower MSE scores in the middle 50% and top 25% intensity range, suggesting that models are better tuned to capture average and top emotional expressions rather than the lower extreme ones. The substantial challenges observed when predicting low-intensity dominance, along with all models registering higher MSE scores for Arousal at lower intensities, signals a clear direction for future refinement in this area.

The findings underscore the models' potential in identifying emotions in speech with varying degrees of success at different ranges of emotional attributes. This evaluation shows consistency in the difficulties faced across all models analyzed and it provides indication that while current models have made great improvements in understanding average emotional states and even model intense regions, the ability to detect and interpret the subtleties of human emotions, particularly at the lower end of the emotional intensity spectrum, requires further advancement.

6. Conclusion

The varied approaches and results presented in the challenge underscore the complexity of speech emotion recognition and the innovative efforts within the research community to address it. The fusion of audio and textual data, alongside the adoption of advanced neural network architectures and fine-tuning strategies, represents the cutting edge in this field. However, challenges such as data imbalance, emotional subtlety, and computational demand persist, inviting further research and development. The insights gained from this challenge not only demonstrate the current state of speech emotion recognition but also pave the way for future advancements in speech and multimodal emotion recognition systems.

7. References

[1] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[2] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[3] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[5] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 2252–2256.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. Lhoest amd A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[7] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Interspeech 2022*, Incheon, South Korea, September 2022, pp. 1168–1172.

[8] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, September 2023.

[9] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S.R. Jang, C.-C. Lee, and H.-Y. Lee, "EMO-SUPERB: An in-depth look at speech emotion recognition," *ArXiv e-prints (arXiv:2402.13018)*, pp. 1–10, February 2024.

[10] A. Reddy Naini, M.A. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, Seoul, Republic of Korea, April 2024, pp. 12031–12035.

[11] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of Machine Learning Research (PMLR 2023)*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202, pp. 28492–28518. PMLR, Honolulu, Hawaii, USA, July 2023.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3465–3469.

[13] W.-N. Hsu, Y.-H. H. Tsai B. Bolte, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in *Chinese National Conference on Computational Linguistics (CCL 2021)*, Huhhot, China, August 2021, pp. 1218–1227.

[15] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[16] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *ArXiv e-prints (arXiv:2312.15185)*, pp. 1–12, December 2023.

[17] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, June 2019, pp. 11057–11066.

[18] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July-August 1998.

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *ArXiv e-prints (arXiv:1510.08484)*, pp. 1–4, October 2015.

[20] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5220–5224.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, February 2020.

[22] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR 2022)*, Virtual conference, April 2022, pp. 1–13.

[23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P.S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *ArXiv e-prints (arXiv:2307.09288)*, July 2023.

[24] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2021)*, Cartagena, Colombia, December 2021, pp. 244–250.

[25] M. India, P. Safari, and J. Hernando, "Double multi-head attention for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, June 2021, pp. 6144–6148.

[26] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y.T. Lee, "Textbooks are all you need ii: **phi-1.5** technical report," *ArXiv e-prints (arXiv:2309.05463)*, pp. 1–16, September 2023.