The CONILIUM proposition for Odyssey Emotion Challenge Leveraging major class with complex annotations

Meysam Shamsi, Lara Gauder, Marie Tahon

The CONILIUM proposition for Odyssey Emotion Challenge Leveraging major class with complex annotations

Meysam Shamsi[1], *Lara Gauder*[2], *Marie Tahon*[1]

[1] Laboratoire d'Informatique (LIUM), Le Mans Universite, France
[2] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

meysam.shamsi@univ-lemans.fr, mgauder@dc.uba.ar, marie.tahon@univ-lemans.fr

Abstract

This paper describes the contribution of the CONILIUM team in the Odyssey Emotion Recognition Challenge. Our system focuses on predicting categorical emotions from speech recordings in the *MSP-Podcast* corpus. Focusing on the training protocol, we investigated several approaches to improve emotion recognition accuracy. Different pre-trained models (*WavLM-large*, *Wav2vec2-large*, *Hubert-large*) were evaluated as feature extractors. An agreement-aware loss functions based on all secondary annotations is proposed that consider the disagreement among annotators and the ambiguity of emotional labeling during training.

An idea of keeping only samples with high agreement annotation in the training process shows the benefit of using all annotations by all annotators. Our best performing system utilized *WavLM-large* as the upstream model, weighted binary cross-entropy with secondary labels as the loss function, and a post-processing step that adjusted the decision threshold. This model achieved an F1-Macro score of 0.361 on the development set, 0.335 on the test set, which is a significant improvement compare to the provided baseline. We also explore characteristics of *Easy* and *Difficult* samples based on their prediction performance consistency across different models.

1. Introduction

Speech emotion recognition (SER) aims at the identification of an emotion category or an affective dimension in a speech signal. According to Batliner et al. [1], obtaining large amounts of real-life data is one of the most important hurdles. Since many situations must be represented in real-life data, the collection of real-life emotions requires having large panels of speakers, several different acoustic environments, different emotional and social contexts, which are expensive to set up. Available corpora for emotion recognition in real tasks, currently contains more and more instances. However, in real-life contexts, emotions are quite sparse, neutral speech is predominant and emotions are shaded [2]. The proposed Odyssey Challenge aims at predicting the emotional category of a speech segment on MSP-Podcast dataset [3]. This dataset can be considered as ecological, and thus is biased towards the neutral state.

The collection and annotation of emotional databases are crucial issues in SER and many studies have already been realized in the framework of HUMAINE a decade ago [4]. While most of the databases are annotated with discrete categories or affective dimensions on a pre-segmented speech excerpt [3, 5], some authors claim for continuous detection of emotion with time [6]. In the first case, SER consists in the prediction of a single value or category at the segment level, while in the second case, a value is predicted at each time frame such as valence and dominance [7], or satisfaction [8]. The MSP-Podcast data falls in the first pre-segmented case.

As emotion perception is subjective by nature, many annotators are usually involved in the annotation process. To enable the annotator to capture the complex nature of emotion, emotional segments are evaluated along an annotation scheme inspired by the MECAS (Multi-level Emotion and Context Annotation Scheme) [2]. This scheme enables the representation of complex and realistic emotions using a primary and a secondary emotion label for each instance. Taking into account both primary and secondary labels can help in the definition of more robust target labels [9, 10] than the traditional majority vote.

In the past, most SER systems were based on the extraction of high-level prosodic and spectral features [11] such as the well-known eGeMAPS feature set [12]. Many recent studies [13, 14] have shown that self-supervised pre-trained models (such as *Wav2Vec* [15] or *HuBERT* [16]) are able to capture both non-verbal and linguistic information from the signal.

The CONILIUM team participation to the Odyssey Emotion Challenge addresses three main challenges. First, what is the best pre-trained model for SER regarding MSP-Podcast data. Second, how to deal with the serious imbalance of the emotional categories within the loss function. And finally, we propose to benefit from the secondary labels to strengthen the number of low represented categories. The present article details the three contributions we made for this challenge.

2. MSP-Podcast Dataset

The dataset utilized in this paper was provided by the organizer of the emotion recognition challenge. It consists of a subset derived from the MSP-Podcast corpus [3]. This corpus comprises segments extracted from podcast recordings, each annotated by external annotators. The annotations enclose both attribute-based descriptors, such as activation, dominance, and valence, as well as categorical labels indicating emotions like anger (A), happiness (H), sadness (S), disgust (D), surprise (U), fear (F), contempt (C), neutral (N), and other (O). The organizers of the challenge have included all annotations conducted by the annotators, who not only identify the primary emotion in each segment but also provide annotations for secondary emotions. In the latter case, annotators are permitted to select multiple emotions for a given segment. This paper will delve into the comprehensive analysis conducted using all annotations from each annotator for every sample in the dataset.

In this study, we excluded samples lacking a clear majority voting winner among the available emotions or identified with the primary emotion label as "other".

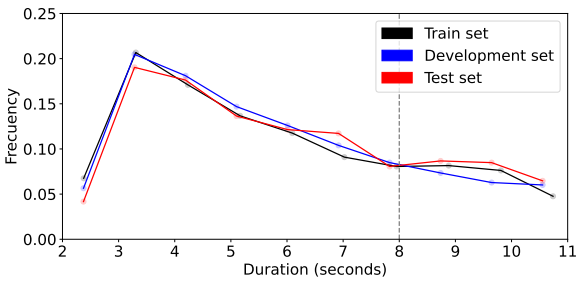


Figure 1: *Duration distribution of audio samples in seconds across each subset. All the results presented in this paper were obtained by truncating samples with a duration exceeding 8 seconds.*

The training set includes 68,360 samples with a mean duration of 5.8 seconds (standard deviation of 2.4 seconds). The development set comprises 19,815 samples, with an average duration of 5.76 seconds (standard deviation of 2.3 seconds). Additionally, the test set, which lacks labeled data, consists of 2,347 samples with a mean duration of 6 seconds (with a standard deviation of 2.34 seconds).

The baseline proposed by the challenge organizer [17] utilized a maximum audio duration of 12 seconds during the training process. As illustrated in Figure 1, they did not truncate any samples in the process. Due to resource constraints, specifically, the unavailability of a computer capable of executing experiments with audio files of this length, we opted for a more manageable approach. Previous research [18, 19] demonstrated that it humanly possible to detect the emotion in less than one second. While we did not find it imperative to significantly truncate the audio, we chose to work with only the initial 8 seconds of each sample.

3. Training Protocol

In this research, we utilized the baseline system provided by the organizers of the challenge. This system comprises a pre-trained upstream model, following by an attentive statistics pooling, along with a linear head for classification [1] [17]. The model takes the raw waveform of the given audio as input to predict 8 different emotional classes. Throughout all the experiments detailed in this study, we used AdamW as the optimizer, setting the learning rate to $1e-5$, and used a batch size of 16, with a maximum of 40 epochs. Subsequently, we selected the best checkpoint based on the development set in terms of macro average F1 score as our final model.

Our training was conducted on a machine equipped with a GPU RTX8000, which required approximately 2 days (variation may occur depending on the upstream model and training set size) for complete training on all data.

The primary focus of our team, CONILIUM, has been on exploring the training protocol to assess the effectiveness of various annotation and training samples. In the following, we provide details on several ideas and experiments which are explored:

- Comparison of using different pre-trained upstream models as feature extractors.
- Detailed proposition for changing the training objective in terms of loss function and desired output prediction during the training phase.
- Study of the problem of imbalanced classes (different frequency of training samples in each class) along with a solution of using a sampler.
- Examination of how annotator agreements can impact the quality of the training set, including testing the usage of only samples with high agreement as the training data.

3.1. Upstream model

While the baseline system [17] utilized the *WavLM-large*[2] [20] as its upstream model, we conducted a comparison of its performance (F1-scores on the development set) by replacing it with *Hubert-large*[3] [16], *Wav2vec2-base* [15] and *Wav2vec2-large*[4] [21], which is the robust large version of *Wav2vec2-base*.

Table 1 presents the comparison of the performance of the system using these different upstream models in terms of F1-Scores. Using *WavLM-large* outperformed both *Wav2vec2-large*, *Wav2vec2-base* and *Hubert-large*, leading to the decision to retain *WavLM-large* as the upstream model for the remainder of this study.

Table 1: F1-scores on development set for different upstream models.

Upstream	F1-Macro	F1-Micro
WavLM-large	**0.302**	0.402
Wav2vec2-large	0.282	0.374
Wav2vec2-base	0.272	0.390
Hubert-large	0.297	0.424

3.2. Data augmentation

A data augmentation process, which involved adding *MUSAN* noise [22] to randomly 80% of training samples, was implemented. However, contrary to expectations, this process did not enhance the performance of the three large models. Interestingly, it did lead to an improvement for the *Wav2vec2-base* model, with the F1-Macro score increasing from 0.272 to 0.286.

One potential explanation for this outcome could be that incorporating the augmentation process during the training of large models renders them inherently resilient to noise. Consequently, including additional data with added noise in the fine-tuning process of large models may not be relevant. However, the fact that the performance with *Wav2vec2-base* when data augmentation is applied, reached the same level as *Wav2vec2-large* underscores the importance of training data.

3.3. Agreement aware loss function

The baseline model used Weighted Cross-Entropy (L_{WCE}) as the loss function (Equation 1), where w_c is the weight of class c, $y_{i,c}$ denotes the reference that the model should predict for a given sample i, and $\hat{y}_{i,c}$ represents the corresponding predicted

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

[2] https://huggingface.co/microsoft/wavlm-large

[3] https://huggingface.co/facebook/hubert-large-ls960-ft

[4] https://huggingface.co/facebook/wav2vec2-large-robust

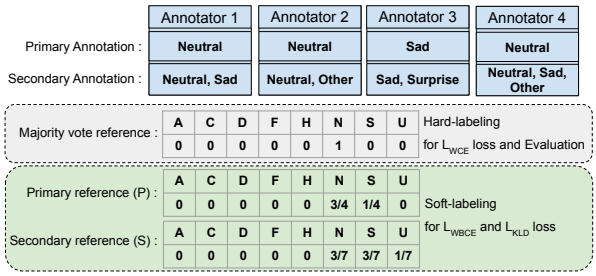


Figure 2: *An example of obtaining P and S as reference label for loss calculation.*

pseudo-probability. w_c is calculated as the inverse frequency of the samples in the class c.

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

Following the utilization of the aforementioned loss function as the training objective, the reference $y_{c,i}$ is a one-hot vector resulting from the majority vote (hard labels). This approach disregards the annotation agreement labeling during training.

Therefore, drawing inspiration from [23, 24], we propose the use of a Weighted Binary Cross-Entropy (L_{WBCE}), where $y_{i,c}$ captures the agreement between annotators for each class. For a given class c, the agreement is the number of annotators who perceived emotion c over all annotators. The L_{WBCE} loss function is expressed in Equation 2. Figure 2, displays an example for taking into account the disagreement of annotator for defining the reference label to calculate the loss during the training process. The "other" label is ignored in this calculation. The evaluation is always done on majority vote of primary labels (hard-labels).

$$L_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c\,(y_{i,c}\log(\hat{y}_{i,c}) + (1-y_{i,c})\log(1-\hat{y}_{i,c})) \tag{2}$$

Additionally, Equation 3 presents the Kullback-Leibler divergence (L_{KLD}) as an alternative loss function, with a similar definition of $y_{i,c}$ as in the L_{WBCE} formulation.

$$L_{KLD} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\frac{y_{i,c}}{\hat{y}_{i,c}} \tag{3}$$

We conducted two analyses. The first analysis conducted involves the primary (P) emotional labels provided by all annotators, which are regarded as pseudo-probabilities of classes. In this context, the training objective is to predict class labels while considering the disagreement among annotators. Additionally, as highlighted in [10], a single emotion often fails to adequately describe expressed speech, which may contain ambiguous emotions. As depicted in Figure 3, the samples are influenced by more than one label. Notably, it is evident that surprise, for instance, often encompasses another emotion, such as happiness, with a notable presence. Then we propose additionally to replace the primary label by the secondary labels when it is reasonable to infer that most frequent classes are associated with less frequent classes.

Therefore, more than one emotional labels can be incorporated into the training by utilizing all secondary (S) emotional

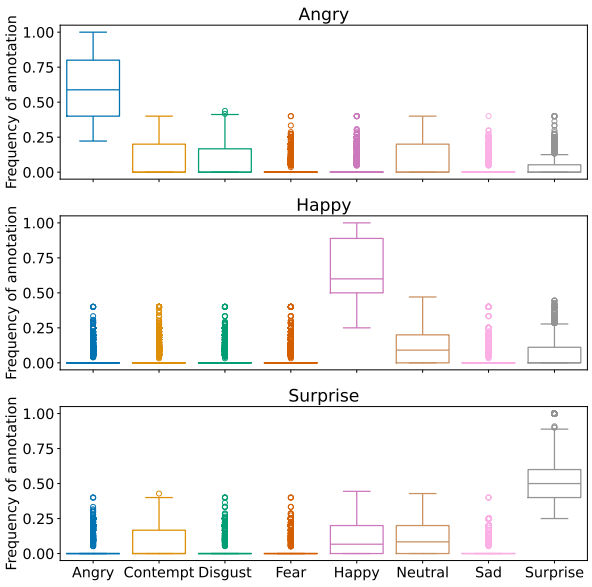


Figure 3: *In this analysis, the distribution of annotation pseudo-probabilities for each emotion is examined based on the primary emotion selected by the annotators at each sample.*

annotations, The secondary emotional labels are defined as all the emotional classes that have been perceived, including the primary emotions. In the preparation of the secondary label as the reference, all labels are treated equally without any weighting on primary emotions. As it can be observed in the example of Figure 2, the values of the reference vector in this case would lower due to the higher presence of different emotional labels in secondary annotations. It should be mentioned that while [24] employs both primary and secondary labels in multitasks learning, we utilize only secondary labels for training and subsequently evaluate the model on the hard labels resulting from the primary task, as per the challenge objective.

Table 2: F1-scores on development set for different loss functions, with the configuration of the best model in Table 1.

Loss	F1-Macro	F1-Micro
L_{WCE}	0.302	0.402
$L_{WBCE}(P)$	0.340	0.503
$L_{WBCE}(S)$	**0.350**	0.517
$L_{KLD}(P)$	0.322	0.481
$L_{KLD}(S)$	0.337	0.502

The results of training with different objective functions are presented in Table 2. Utilizing all primary (P) annotations leads to an improvement in performance compared to the hard-labeling objective function (L_{WCE} loss function). Defining secondary (S) labels as the training objective demonstrates the highest performance in terms of F1-scores even for predicting the majority votes (hard-labeling) of primary classes. This outcome reveals the importance of keeping all annotations of all annotators for training.

A significant difference between performances of L_{KLD} and L_{WBCE} loss functions is apparent. Unlike L_{WBCE}, the L_{KLD} loss function is implemented without weights calculated

from the class distribution in the training set. Therefore, the necessity of addressing class imbalance becomes evident when class weights are not considered in the loss calculation.

3.4. Balancing classes

The significant difference between F1-Macro and F1-Micro in development set, in all the previous results presented until now, confirm the impact of imbalanced number of samples in different classes. Given that the primary evaluation metric of the challenge is defined as macro average F1 score, there is a concerted effort to ensure equal importance of the model's performance across all classes. The implementation of class frequency as the weight in the loss calculation addresses this requirement. We observed a 1% degradation in F1-Macro when unweighted $L_{BCE}(P)$ is used as the loss function instead of $L_{WBCE}(P)$ in the *WavLM-large* upstream model.

To address this issue, specially in the case of L_{KLD}, we propose implementing a data loader sampler that provides samples with statistically balanced classes during training. This sampler assigns a weight of $1/f_{C_i}$ to each sample during its selection for the construction of the training set in each batch. Here, f_{C_i} represents the frequency of the class to which the sample belongs. In the case of L_{WBCE}, the weights would be ignored, since the frequency of classes is already accounted for by the sampler.

While the use of the sampler demonstrates a slight improvement in F1-Macro when employing the unweighted L_{BCE} (from 0.331 to 0.334 on development set), the performance of this approach is lower than L_{WBCE} for both P and S training sets. Conversely, the sampler proves beneficial in enhancing the performance of $L_{KLD}(S)$, achieving an F1-Macro score of 0.347. As a result, it has been determined to utilize the sampler in the case of L_{KLD} loss for the remainder of this study.

3.5. Annotation agreement

Another hypothesis that has been tested is the utility of samples with lower annotator agreement in the training process. To explore this, we propose two variations of the P and S sets, where only samples with high agreement in terms of annotation are retained. A sample set containing primary labels with higher consensus in annotation is denoted as P^*. The P^* is a subset of P comprising only samples with the same primary labels agreed upon by at least 60% of annotators. Similarly, S^* is a subset of S containing only samples with a secondary labels list, where at least 50% of the labels list are the same. This filtering of training samples reduces the number of samples in the training set from 54,651 in P (respective 53,523 in S) to 30,647 in P^* (respective 25,514 in S^*).

Table 3: F1-scores on development set, when only samples with high agreement annotation are used for training.

Loss	F1-Macro	F1-Micro
$L_{WBCE}(P^*)$	0.317	0.491
$L_{WBCE}(S^*)$	0.312	0.511
$L_{KLD}(P^*)$	0.314	0.483
$L_{KLD}(S^*)$	0.326	0.487

Table 3 presents the F1-scores obtained using samples with high annotation agreements. A comparison of these results with those in Table 2 reveals a degradation in F1-scores, particularly in F1-Macro. This decline can be attributed to the model's performance on less frequent classes e.g. disgust (D) and fear (F), due to the removal of more than three-quarters of samples with less frequent labels from the training set. Conversely, the deletion of samples in the most frequent classes, e.g. neutral (N), is less than one-third. From these observations, it can be concluded that in a naturalistic emotionally balanced corpus such as the MSP-Podcast, retaining all samples, even in cases of disagreement in their annotation, proves beneficial.

3.6. Ensemble voting

By comparing the predicted values on the development set obtained from different upstream models with various loss functions and training objectives, it is observed that different predictions can result from different upstream models. This discrepancy is more pronounced than the variations caused by loss functions or training objectives alone. Therefore, we propose employing a voting process to predict the final label based on the probability outputs of three upstream models with WBCE(P) as their loss function.

Following the ensemble voting of three models, different upstream architectures with the best configuration, resulted in F1-Macro of 0.338 and F1-Micro of 0.523. Compared to the best result (*WavLM-large*), a degradation of F1-Macro has been observed, regardless of whether the voting is weighted or not. This observation convinces us to retain only the *WavLM-large* model and not follow the ensemble voting process. The degradation of F1-Macro and improvement in F1-Micro in ensemble voting reveal that this process favors most frequent classes.

3.7. The best configuration

After conducting detailed experiments outlined in the previous sections, we propose the utilization of *WavLM-large* as the upstream model and WBCE(S) as the loss function. As a final step and post-processing measure, we suggest adjusting the decision threshold for converting output probabilities to hard labels.

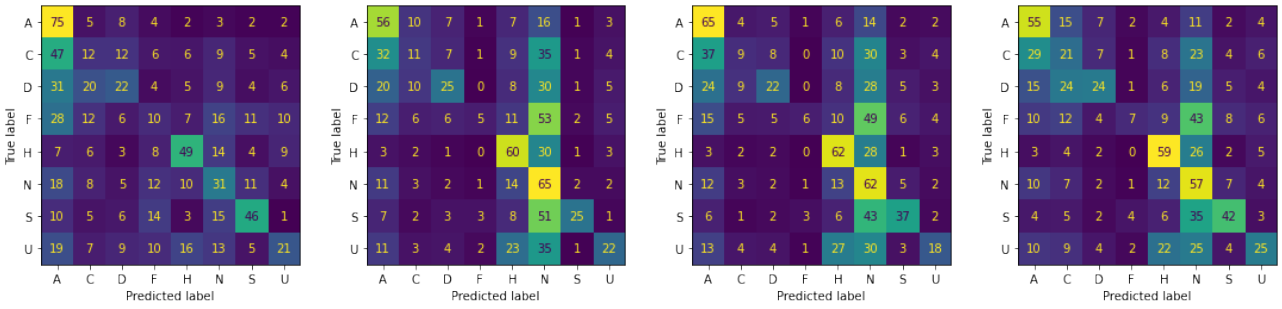
This adjustment is motivated by two key factors. Firstly, the main evaluation metric of the challenge is F1-Macro. Secondly, the equal scores of F1-Macro and F1-Micro indicate a balanced distribution of samples in the test set. Consequently, we empirically optimized the decision threshold on the development set to enhance F1-Macro, thereby increasing the likelihood of predicting less frequent classes. However, it's important to note that this adjustment results in a degradation of F1-Micro. Following this process, we successfully improved the performance of the best model with a F1-Macro score of 0.361 (F1-Micro score of 0.496) on development set.

Table 4: F1-scores of the best model at the two evaluation sets. The results using the test sets were extracted from the leaderboard provided by the organizers of the challenge.

Split	F1-Macro	F1-Micro
Development	0.361	0.496
Test	0.335	0.347

Comparison of this result with the F1-scores on test set has been displayed in the Table 4. A lower score of F1-Macro on test set comparing to the development set indicates our system is slightly over fitted on development set.

Figure 4 displays the confusion matrix on the development set under different configurations. Comparing the use of



(a) Baseline system
F1-Macro=0.30

(b) $L_{WBCE}(P)$
F1-Macro=0.34

(c) $L_{WBCE}(S)$
F1-Macro=0.35

(d) $L_{WBCE}(S)$ + post-processing
F1-Macro=0.36

Figure 4: Confusion Matrix of different steps of finding the best configuration on development set (values are in % of reference hard-labels).

$L_{WBCE}(P)$ and the Baseline system (Figure 4a and 4b), it is evident that the main improvement in F1-scores comes from enhanced performance on most frequent classes like happiness (H) and neutral (N), while the model's performance is diminished for the classes angry (A) and surprise (S). An enhancement of the model's performance on less frequent classes can be seen by comparing the impact of our threshold adjustment as post-processing on the usage of $L_{WBCE}(S)$ as loss function (Figure 4c and 4d).

4. Easy versus Difficult Samples

To further analyze the results and investigate the causes or correlations between the model's performance and sample characteristics, we examine the performance of our models on the samples in the development set in this section.

While that the predictions of three models with different upstream architectures (using their best configurations, see section 3) differ more than 30% of the time in the development set, their aggregation does not improve overall performance. Then, it can be assumed that there are some samples that are consistently predicted incorrectly regardless of the model used, while others are consistently predicted correctly. Therefore, we introduce the concept of *Easy* samples, which are predicted correctly by all three models, and *Difficult* samples, which are predicted wrongly by all three models. Out of the 15,341 samples in the development set, 5,179 samples can be considered as *easy* and 5,433 as *difficult*.

In the following we analyze the characteristics of these two types of samples in terms of length, inter-annotator agreement, and Speakers.

4.1. Samples duration

In terms of duration, it's observed that *Difficult* samples tend to be slightly longer than *Easy* samples. On average, *Easy* samples have a duration of 5.65 seconds ((standard deviation of 2.27 seconds), while *Difficult* samples have an average duration of 5.80 seconds ((standard deviation of 2.32 seconds). It means that *Difficult* samples tend to be slightly longer than *Easy* ones. Although this difference may not be significant, in the future study, it can be proposed to potentially increase the maximum duration of training samples. Such an adjustment might offer a more comprehensive representation of *Difficult* samples during training, potentially leading to improved model performance on longer-duration samples.

4.2. Inter-annotator agreement

The inter-annotator agreement of *Easy* samples is significantly higher than that of *Difficult* samples. The Krippendorff alpha value for *Easy* samples is 0.54, indicating a relatively high level of agreement among annotators. In contrast, the Krippendorff alpha value for *Difficult* samples is 0.38, suggesting a lower level of agreement among annotators. This discrepancy underscores the complexity and ambiguity present in *Difficult* samples, which pose challenges for consistent annotation across multiple annotators.

4.3. Diversity of speakers

In the dataset, speakers can be categorized based on the difficulty of their samples. When a speaker has more than 50% of their samples classified as *Easy* samples, they are considered an *Easy* speaker. Conversely, if more than 50% of their samples are classified as *Difficult* samples, they are labeled as a *Difficult* speaker. It's observed that *Easy* speakers appear, on average, 10 times in the development set, while *Difficult* speakers appear only around 2 times. This statistic shows, while the model does not have any information about speakers in development set, but still the rare speaker in development set are more difficult than speaker with higher frequency of appearance. Indeed, additional information concerning the selection process of speakers in the dataset can allow more logical explanation.

Moreover, when examining the gender distribution of speakers in the development set, with a balanced representation between genders (female: 7,619 samples, male: 7,421 samples), a comparison between the gender distributions of *Difficult* and *Easy* speakers reveals interesting insights. The prediction of samples with female speakers appears to be more challenging, with 2,586 *Difficult* samples of female speakers compared to 1,794 *Difficult* samples of male speakers. Conversely, for *Easy* speakers, there are 1,771 *Easy* samples of female speakers and 2,414 *Easy* samples of male speakers. This discrepancy in gender distribution further emphasizes the complexity and challenges associated with accurately predicting the emotional content of female speakers' samples in the MSP-Podcast corpus.

5. Conclusion

This paper is the description of our system named CONILIUM team in the Odyssey 2024 Emotion Recognition Challenge for

the classification task. In this contribution, we explored various aspects of training set design for speech emotion recognition, focusing on training objectives, class balancing techniques, annotation agreement, and ensemble voting. Our findings highlight the importance of considering the nuances of the dataset and the training process to optimize model performance. Is has been observed that utilizing all annotations, particularly secondary labels, as the training objective leads to improved performance across multiple upstream models. This implies that retaining all annotations from all annotators, regardless of any disagreement, proves advantageous for training a system tasked with predicting even the hard-label emotional class for a given audio file.

The best result on the development set was obtained by using the *WavLM-large* as the upstream model and profiting from all secondary labels in the training phase with a weighted binary cross-entropy (WBCE) loss. Our proposed system achieved F1-Macro of 0.361 and F1-Micro of 0.496, representing a 6% and 9% absolute improvement, respectively, compared to the baseline system on the development set. The submitted predictions, using our best model, on the test set achieved F1-Macro of 0.335 and F1-Micro of 0.347, placing our team in the 6th rank among all team participants.

Our analysis of *Easy* versus *Difficult* samples revealed insights into the characteristics of challenging samples, such as longer duration, lower inter-annotator agreement, and the samples with female speaker. Addressing these challenges, such as by incorporating longer-duration samples and exploring techniques to more focus on female (or rare) speakers, may further improve model performance.

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666 and the project PULSAR supported by the Region of Pays de la Loire, France grant agreement No 2022-09747. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565). The research reported here was conducted at the 2023 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Le Mans University (France) and sponsored by Johns Hopkins University.

7. References

[1] Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous, and Noam Amir, "Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech," *Computer Speech and Language*, vol. 25, no. 1, pp. 4, July 2010.

[2] Laurence Devillers, Laurence Vidrascu, and Lori Lamel, "2005 special issue: Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, may 2005.

[3] 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.

[4] Ellen Douglas-Cowie, C. Cox, J-C. Martin, L. Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, C. Pelachaud, C. Peters, O. Lowry, A. Batliner, and F. Hoenig, *Data and Databases: The HUMAINE database.*, pp. 243–286, Springer, 2011.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[6] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech*. Sept. 2008, pp. 597–600, ISCA.

[7] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau, "AlloSat: A New Call Center French Corpus for Satisfaction and Frustration Analysis," in *Language Resources and Evaluation Conference, LREC 2020*, Marseille, France, May 2020.

[9] Marie Tahon, Agnes Delaborde, and Laurence Devillers, "Real-life Emotion Detection from Speech in Human-Robot Interaction: Experiments across Diverse Corpora with Child and Adult Voices," in *Interspeech*, Firenze, Italy, 2011.

[10] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Interpreting ambiguous emotional expressions," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–8.

[11] Marie Tahon and Laurence Devillers, "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 16, 2016.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016, IEEE Transactions on Affective Computing.

[13] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6922–6926.

[14] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau, "On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous

Speech Emotion Recognition," in *IEEE Spoken Language Technology Workshop*, Virtual, China, Jan. 2021.

[15] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, 2021.

[17] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L.P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results," in *Odyssey 2024: The Speaker and Language Recognition Workshop)*, Quebec, Canada, June 2024, vol. To appear.

[18] Marc D Pell and Sonja A Kotz, "On the time course of vocal emotion recognition," *PlOS one*, vol. 6, no. 11, pp. e27256, 2011.

[19] Henrik Nordström and Petri Laukka, "The time course of emotion recognition in speech and music," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3058–3074, 2019.

[20] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu 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, 2022.

[21] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *CoRR*, vol. abs/2104.01027, 2021.

[22] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[23] Huang-Cheng Chou and Chi-Chun Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5886–5890.

[24] Huang-Cheng Chou, Wei-Cheng Lin, Chi-Chun Lee, and Carlos Busso, "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7717–7721.