IRIT-MFU Multi-modal systems for emotion classification for Odyssey 2024 challenge

Adrien Lafore, Clément Pagés, Leila Moudjari, Sebastião Quintas, Hervé Bredin, Thomas Pellegrini, Farah Benamara, Isabelle Ferrané, Jérôme Bertrand, Marie-Françoise Bertrand, et al.

IRIT-MFU Multi-modal systems for emotion classification for Odyssey 2024 challenge

Adrien Lafore[1,2], *Clément Pagés*[1], *Leila Moudjari*[1], *Sebastião Quintas*[1]
Hervé Bredin[1], *Thomas Pellegrini*[1], *Farah Benamara*[1,3], *Isabelle Ferrané*[1]
Jérôme Bertrand[2], *Marie-Françoise Bertrand*[2], *Véronique Moriceau*[1], *Jérôme Farinas*[1]

(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) My Family Up, Toulouse, France
{first_name.last_name}@irit.fr
mf.bertrand@myfamilyup.com j.bertrand@myfamilyup.com
(3) IPAL, CNRS-NUS-A*STAR, Singapore

Abstract

In this paper, we present our contribution to emotion classification in speech as part of our participation in Odyssey 2024 challenge. We propose a hybrid system that takes advantage of both audio signal information and semantic information obtained from automatic transcripts. We propose several models for each modality and three different fusion methods for the classification task. The results show that multimodality improves significantly the performance and allows us surpassing the challenge baseline, which is an audio only system, from a 0.311 macro F1-score to 0.337.

1. Introduction

The work presented here details our contribution as part of our participation in the Task 1 of the Odyssey 2024 challenge on emotion recognition. Task 1 consists in classifying audio segments across 8 emotional classes. In this framework, we proposed 3 hybrid systems which aim to use state-of-the-art techniques to combine audio signals with text transcripts information.

Emotion detection is a well established task in Natural Language Processing (NLP) [1, 2], through many shared tasks such as SemEval [3]. The detection of emotion from text has mainly been tackled in the context of the detection of (1) specific discourses (hate speech, influence, politics [4, 5, 6]) in conversations and comments on social media, or (2) mental illnesses such as depression, eating disorders or bipolar disorders (see [7] for a review of existing datasets and tasks). The Computational Linguistics and Clinical Psychology (CLPsych) [8] or eRisk [9] evaluation campaigns focused in particular on the task of automatically detecting depression of social network users, with a focus on "early" detection. The learning models developed for automatic detection are either feature-based models for the most successful ones (with features such as use of personal pronouns, positive or negative sentiment, verb tense, etc.) [10, 11], rule-based or lexicon based methods or deep learning ones (see [12] for an overview of existing automatic methods for emotion analysis).

In Automatic Speech Processing (ASP), emotion recognition is closely linked to automatic prosody analysis. Indeed, non-verbal information is often the source of information for characterizing emotions. Several international competitions have been held in this field, such as The Interspeech 2009 Emotion Challenge and The Interspeech 2010 Paralinguistic challenge. Early systems were based on discriminative systems fed by numerous parameters extracted from the audio signal, such as the OpenSmile toolbox [13]. Current systems are based on deep neural network architectures and have achieved good performance in categorical emotion classification [14, 15, 16, 17]. The challenge now is to project emotions into a continuous space, enabling the study of cognitive mental states [18].

In the following sections, we first present the data and the classification task. We then present the architecture of our 3 systems - corresponding to our 3 submissions to Task 1 - highlighting the key features and underlying design choices that have contributed to their performances. Finally, we present and discuss the results, showing that multimodality improves the performance.

1.1. Data

The data made available for this challenge contains English recordings from the MSP-Podcast dataset [19], which contains audio segments from online podcasts. The speech turns were annotated by at least 5 annotators according to emotion categories and their respective dimensions. The annotated emotion categories are: Anger (A), Contempt (C), Disgust (D), Fear (F), Happiness (H), Neutral (N), Sadness (S), Surprise (U), Other (O), and No agreement (X). The dimensions for each emotion are *valence* (positive or negative state of the individual), *arousal* (activity or passivity of the individual) and *dominance* (weak to strong control). Each of these dimensions is annotated on a scale from 1 to 7.

The training and development sets consist of 68,360 and 19,815 annotated speech turns respectively (see Table 1). For each set, the transcripts are provided as well as the gender of the speakers. Transcripts were obtained using the crowd-source service rev.com, a platform that contains a marketplace for expert transcribers. Force alignment results of the train and development sets are also provided, obtained from an acoustic model trained on the Librispeech dataset. The train and development sets do not present a balanced distribution of all emotion classes.

The test set is composed of 2,347 speech segments from 187 individuals. No transcript is provided as well as no force alignment. In addition, the classes "Other" and "No agreement" have been removed, and the distribution of emotion categories

is balanced. Therefore, we have removed these two classes from the training and development sets. Given that the test set is balanced across classes, we similarly built randomly a balanced development set from the one provided by the challenge organizers: this crafted set contains 282 files per emotion which corresponds to the maximum file number possible for this balance. We hypothesize that a system optimization on the original development set and on the balanced one may help the generalization of our system on the final challenge test set.

1.2. Classification Task and Evaluation

The task we took part in the context of the Odyssey 2024 challenge was categorical emotion recognition (Task 1). In this context, 8 emotion categories are provided: Anger (A), Contempt (C), Disgust (D), Fear (F), Happiness (H), Neutral (N), Sadness (S), and Surprise (U). The challenge does not allow the use of any existing models trained for emotion detection, nor additional training data.

Participating systems were evaluated according to the classic measures of precision, recall, F1-score and accuracy. As the distribution of classes in the test dataset is balanced, macro-F1 is used to rank the systems ; macro-F1 being the average of the F1-scores for each of the 8 classes (see Formula (1)).

$$\text{Macro F1} = \frac{1}{8}\sum_{i=1}^{8} 2 \times \frac{\text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \quad (1)$$

2. Models

In this section, we first present the baseline provided by the challenge and the global architecture of our system. Afterwards, we provide a detailed presentation of each module of our system.

2.1. Baseline

A baseline is provided by the organizers [20] and presented in Figure 1. The system consists of a previously pre-trained large version of the self-supervised model WavLM [21] fine-tuned for the categorical emotion recognition task. A learning rate of $1e-5$, 20 epochs and a batch size of 32 were used. The system uses an attentive statistics pooling [22] approach, that provides a weighted mean and standard deviation pooling based on learned self-attention weights.

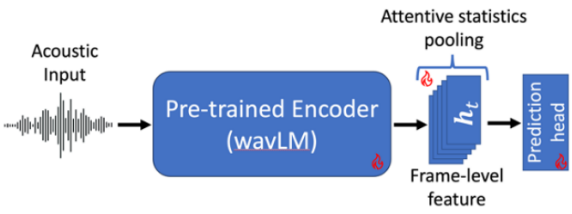

Figure 1: Baseline provided by Odyssey 2024 challenge.

2.2. Global Architecture

State-of-the-art studies showing the advantage of a multimodal (text and audio) emotion classification [23], we decided for this challenge to use a hybrid system that combines a text and an audio model in order to exploit the strengths of each model and compensate their weaknesses.

Generally speaking, these hybrid systems use text and audio embedding models, a separate data processing pipeline and operate a fusion of the two embedding layers, either through the means of an early fusion [24], or more typically a fusion at the end of the process [25]. In this context, we developed two separate models for audio processing as well as for text processing. Then, we propose several simple fusion methods in order to compare their performances. Figure 2 shows the global architecture of our hybrid system.

2.2.1. Automatic Transcription

Transcripts are provided for the training and development sets, but not for the test set. Thus, an ASR system is needed to get the transcripts for the test set. For automatic speech recognition, we made use of the Whisper-large-v3 model [26], a large speech model pre-trained on 680,000 hours of multilingual data collected on the Web, making it robust to different accents and acoustic conditions. Whisper provides transcripts with punctuation and capital letters for named entities, an aspect that differs from typical ASR systems. We assume that these aspects may further help the accuracy of a text-based emotion classifier. On the other hand, the introduction of punctuation can hurt WER performance, an expected consequence.

To evaluate the transcription performance, we computed a Word Error Rate (WER) between the reference transcripts (provided by the challenge organizers) and the Whisper-based outputs. A WER of about 32% was obtained on both the training and development sets. Furthermore, the WER reduces to about 13% on both sets if we introduce a normalization that removes punctuation and provides uncased transcripts. However, we decided not to apply normalization since punctuation may bring information about emotions, for example an exclamation mark may be a clue for surprise or anger.

2.2.2. Speech Model

The model used for emotion recognition from audio segments has been implemented and trained using the pyannote.audio [27, 28] toolkit, and the architecture is inspired by those of PyanNet [29] and SSeriouSS[1]. This system takes a 10 second-long audio chunk as input. Firstly, a feature extractor provides a sequence of frame-wise features from the audio chunk. Then, this sequence is fed to a stack of bidirectional LSTM (BLSTM) layers with hidden state size set to 256 (128 for each direction). These BLSTMs provide a sequence of embeddings used by a classification linear layer, inferring emotion probabilities at the frame level. Finally, we use pooling to aggregate this information at the chunk level, and a log-softmax activation function, assigning a probability to each emotion class.

The training phase was carried out using a batch size of 32, the same size as in the baseline. The learning rate was set to 0.001. A reduce-on-plateau scheduler was used to divided this learning rate by two every 10 epochs without any improvement on the macro F1-score, up to a value of $1e-8$. The loss function used is the Negative log likelihood, weighted with coefficients inversely proportional to the total number of samples for each emotion in the training set. Additionally, data augmentation using music (total of 42,5 hours of recordings) and noise (total of 6 hours) from MUSAN corpus [30] was performed on-the-fly

[1]The code for this architecture is available here: `https://github.com/pyannote/pyannote-audio/blob/develop/pyannote/audio/models/segmentation/SSeRiouSS.py`

Table 1: Category distribution in the training and development sets (*Other* and *No agreement* classes have been removed)

	Train			Development		
Category	**Number**	**%**	**Duration**	**Number**	**%**	**Duration**
Neutral (N)	25,106	36.72	39h43	5,667	28.60	08h47
Happiness (H)	13,440	19.66	22h09	3,340	16.86	05h14
Sadness (S)	3,882	5.68	06h13	1,101	5.56	01h44
Anger (A)	3,053	4.47	05h05	2,413	12.18	04h01
Surprise (U)	2,897	4.24	04h38	729	3.68	01h04
Contempt (C)	2,443	3.57	04h09	1 323	6.67	02h17
Disgust (D)	1,426	2.09	02h24	486	2.45	00h52
Fear (F)	1,139	1.67	01h46	282	1.42	00h26

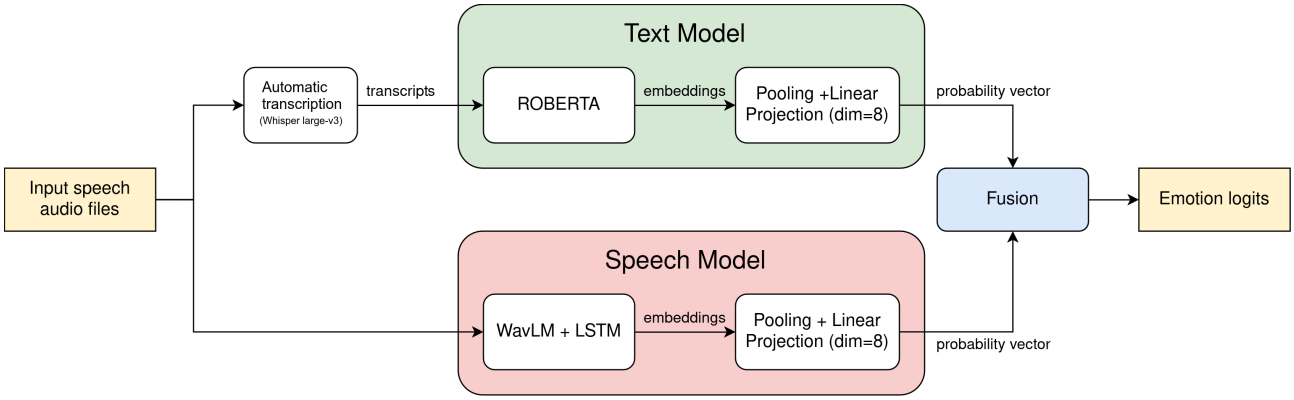

Figure 2: System global architecture.

to increase data variability. Indeed, the model tends to rapidly overfit on the training set. MUSAN data have been added to audio chunks with a SNR between 5 and 10 dB.

We have tested several features extractor modules, namely SincNet [31] and different versions of WavLM, pretrained on LibriSpeech and freezed during training. For these experiments, the number of BLSTM layers was set to 2, and we use mean pooling as pooling method. Results can be found in Table 2. These results show that the best performance in terms of macro F1-score are obtained using WavLM rather than SincNet, with a relative performance improvement of 40% and 45% using WavLM-base, on the development set and on our balanced development set respectively. The use of WavLM-large further increased the F1-score, with a relative improvement of between 7% and 9% on the development set and on the balanced one.

Table 2: Results of the audio system with different features extractors. Underlining indicates the audio system specifications used in the hybrid system.

Features extractors	**Macro F1-score**	
	Dev	**Balanced dev**
SincNet	0.179	0.175
WavLM-base	0.251	0.253
WavLM-large	**0.269**	**0.275**
WavLM-base-plus	0.247	0.243

We have also evaluated the pooling method used to aggregate frame-wise classification at the chunk level, and the number of BLSTM layers. Two methods have been experimented: mean and max pooling. In addition, to ensure that all the audio chunks seen by the model during training were 10 second long,

Table 3: F1-score of the audio system using different pooling methods, with a WavLM-Large as features extractor. Underlining indicates the audio system specifications used in the hybrid system.

Pooling method	**Zero-padding removal**	**BLSTM layers**	**Macro F1-score**	
			Dev	**Balanced dev**
mean	✗	2	0.269	0.275
mean	✓	1	0.276	0.304
mean	✓	2	0.290	**0.307**
mean	✓	3	**0.296**	0.303
max	✗	2	0.278	0.283
max	✓	2	0.288	0.289

the shorter audios were zero-padded at the beginning and end of the chunk. To remove the noise introduced by the addition of these virtual frames, we implemented zero-padding suppression at the pooling layer. Table 3 shows all the results obtained for these experiments, which were carried out using a WavLM-Large model as feature extractor.

The results demonstrate that suppressing zero-padding at the pooling level improves system performance by eliminating the noise introduced by these frames, whether with mean or max pooling. This is particularly true with mean pooling, where we observe a relative improvement in F1-score of 7.8%. In fact, removing the frames corresponding to zero padding prevents them from being taken into account when calculating the mean. The improvement is smaller when using max pooling, as this

method is less sensitive to these frames. We also observe that the best performance is obtained with 2 layers of BLSTM, with slightly lower performance when this number is both increased and decreased on our balanced development set. Finally, based on these results, we have decided to use a WavLM-large model with 2 BLSTM layers, and mean pooling with zero padding-removal in our hybrid system.

2.2.3. *Text Model*

Not being allowed to use specific emotion detection models, we trained and evaluated several state-of-the-art language models on the reference transcripts of the training and development (dev) sets as well as our balanced development set (b-test): several BERT versions (base, multilingual, large, etc.) [32] and RoBERTa-base [33]. The best performance was achieved with RoBERTa, that is why we chose this model for our hybrid system.

The RoBERTa-base model consists of 12 layers with a hidden size of 768 and an attention head count of 12. The intermediate feedforward layers have a size of 3,072. The model is built on a vocabulary of 50,265 words, including special tokens for the beginning and end of the sequence (CLS and SEP). Additionally, the model is case-sensitive and uses random masking tokens (MLM) for training.

We fine-tuned RoBERTa-base (125 million parameters) on the training data. For this, we used the Adam optimizer with a learning rate of 2.10^{-5} for 4 epochs. Batch sizes were experimentally set to 64, and we used a weighted cross-entropy loss function. Finally, a linear layer was added for the emotion classification task. Similar to the audio, the output of this layer is then subjected to a log-softmax activation function, assigning a probability to each class. We will refer to this configuration as "Base". We also tried to use the RoBERTa-base embeddings as an input to a CNN (RoBERTa+CNN) and an LSTM (RoBERTa+LSTM).

Furthermore, we conducted an additional experiment involving data augmentation (referred to as RoBERTa+DA). In this experiment, we aimed to address the scarcity of data in certain classes (C, D, and F) by generating similar texts. Using ChatGPT-3.5[2], we generated a minimum of 500 texts per class and added them to the training dataset. It is worth noting that the generated texts exhibit a less natural flow and adopt a more formal tone, for example:

Disgust reference example: "...this is disgusting. i sat down and started reading these articles and my".

Generated texts:

- The sight of their self-serving agenda is sickening.
- Their actions are morally abhorrent, they have no conscience.

Table 4 displays the top-performing results from each experiment set in terms of accuracy and macro F1-score, on the whole developement set and on the balanced one. The result analysis reveals that our Base configuration yielded the most optimal performance. Interestingly, the incorporation of data augmentation contributed to a marginal improvement of 0.03% on the development set and about 1% on the balanced one, albeit relatively negligible.

[2] https://openai.com/blog/chatgpt

Table 4: Results of language models on the development dataset. Underlining indicates the model specifications used in the hybrid system.

	Dev		**Balanced dev**	
	Acc	Macro F1	Acc	Macro F1
RoBERTa-base	44.15	**29.44**	29.03	**26.41**
RoBERTa+CNN	43.29	28.68	28.66	26.23
RoBERTa+LSTM	43.91	29.32	29.26	**26.51**
RoBERTa+DA	43.95	**29.47**	29.55	**27.40**

2.2.4. *Fusion*

Among all mono-modal models that we experimented, we choose the WavLM-large model with 2 BLSTM layers, and mean pooling with zero padding-removal for audio and RoBERTa-Base model for transcripts since they have the best results on our balanced development dataset.

After training models on audio segments and transcripts respectively, we worked on how getting a maximum benefit from both models. State of the art shows many ways to operate a fusion in a multi-modal system, some of them in the context of emotion recognition using hybrid attention mechanisms [23].

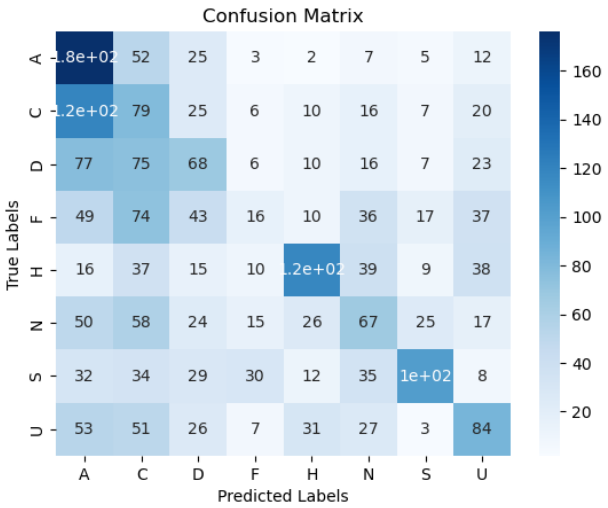

Figure 3: Confusion matrix for the audio system

In order to check if a fusion will improve the final predictions, we first analyzed the performance of both mono-modal models for each class. Figures 3 and 4 show that the audio system is better at recognising Anger, Happiness and Sadness - this is not surprising as these are emotions with recognisable prosody markers - whereas the text model has good performance on Happiness and Neutral, even if it tends to over-represent this class (contrary to the audio model). Considering these observations, we can expect some result improvements with a fusion of the outputs of both models. We propose here 3 simple fusion methods, corresponding to our 3 submissions to Task 1.

Fusion 1: Mean of probability predictions. We first experimented with a simple mean between probability predictions of our WavLM-large+BLSTM and RoBERTa-base models for each class. This allowed us to establish the efficiency of a

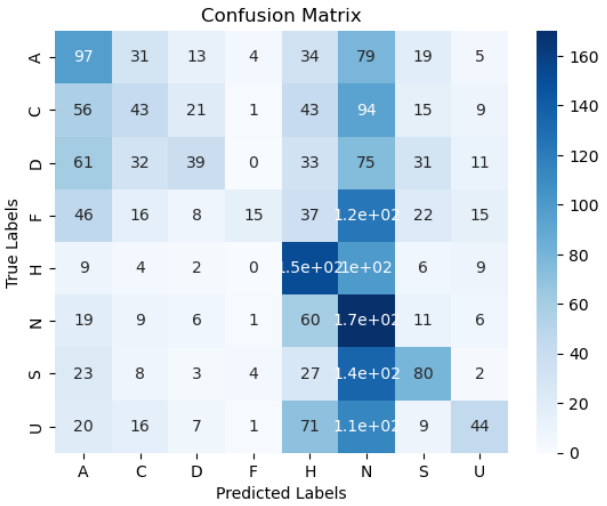

Figure 4: Confusion matrix for the text system

simple fusion method and later to compare with more advanced mechanisms.

Fusion 2: Multi Layer Perceptron As shown by state-of-the-art studies, it may be very efficient to operate the fusion of several modalities with a Multi Layer Perceptron (MLP) learning model [34]. For our learning fusion model, we were inspired by existing architectures composed of a hidden layer followed by a classification layer [35]: its inputs being the outputs of the two prediction systems for 8 classes, we concatenated both mono-modal outputs to fed our MLP fusion model. Its architecture is based on a hidden layer of size 16 followed by a size 8 classification head to return the final prediction. We added a 0.5 dropout and a sigmoid activation to avoid a too much linear behavior.

The whole train set being used for the training of both mono-modal models, we had to use our balanced development dataset (BDD) to train the MLP fusion model. We trained our model on 80% of BDD files and validated it on the other 20% to find the best training parameters. We obtained optimal values for all number epoch, batch-size and learning rate (respectively 100, 16, $1e{-}4$).

Once the best parameters obtained, we trained our fusion model on the totality of our balanced development dataset. The resulting model is then used to take both outputs of our mono-modal systems and predict emotion on our development and test datasets.

Fusion 3: Weighted fusion Finally, we experimented with a weighted fusion model: we attributed weights to the two mono-modal systems.

This model outputs depend on the outputs of our two models and a parameter alpha ($0 \leq \alpha \leq 1$) so that the probabilities outputs P_i for each emotion i are :

$$P_i = \alpha P_{i,audio} + (1 - \alpha) P_{i,text} \quad (2)$$

where $P_{i,audio}$ and $P_{i,text}$ are probabilities provided by the audio and text models for emotion i, respectively.

Then we analyzed the performance on the balanced development set with all possible alpha values between 0 and 1 (with a granularity of 0.01). The best results were obtained with $\alpha = 0.53$ (note that the outputs obtained with $\alpha = 0.5$ correspond to the outputs of our first fusion model where both audio and text models have the same weight). At first sight, this does not seem to be a big improvement compared to the mean fusion method but the results show that this weighted fusion enabled to gain few ranks in the challenge.

3. Results

Table 5 shows the results of all systems on the different datasets.
Mean Fusion. The results obtained with the mean fusion show clearly that the emotion recognition performance has been improved compared to audio and text models alone. We note that the performance for Anger, Happiness and Sadness reached by our audio system remained stable with this latter fusion process (see Figure 5). Furthermore, this fusion method provides the system a good capacity to detect neutrality which is the main weakness of our audio model. With these improvements, we were able to get a macro F1-score of 0.331 by combining two mono-modal systems which respectively obtained 0.307 and 0.264 on our balanced development set.

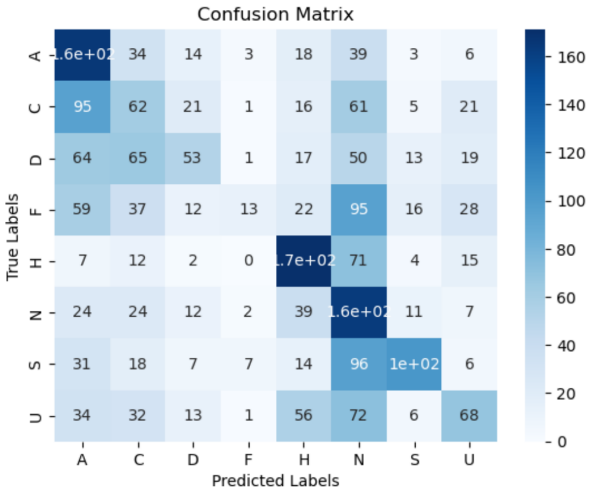

Figure 5: Confusion matrix of mean fusion system

MLP Fusion. The second fusion model shows a real performance improvement on the balanced development dataset. Indeed, the MLP fusion has a F1-score of 0.363 on the development dataset and 0.343 on our balanced version. Nonetheless, it seems that our training sample data was too small to avoid over-fitting because our second submission reveals poor results on the test set (macro F1-score = 0.18). The additional training on the 20% files may have been a mistake too, because, as the model is tiny and the quantity of data is small, learning parameters should have been adapted to this new data pool (but we would have faced a serious bias issue that could have decrease our performance even more).
Weighted Fusion. Our third system performs quite well compared to the baseline. With a macro F1-score of 0.346 and 0.337 on the challenge development and test datasets, we achieve an improvement with a tiny weighting optimisation.

In Table 5, we report the results of each of our mono-modal and multi-modal systems compared to the challenge baseline, either on the challenge development dataset, our balanced subset of the development set or the challenge test set. Table 5

Table 5: Results of all systems

Modality	Systems	Development		Balanced Development		Test	
		Accuracy	**Macro F1**	**Accuracy**	**Macro F1**	**Accuracy**	**Macro F1**
Audio	WavLM-Large + LSTM	0.331	0.290	0.314	0.307	N/A	N/A
Text	RoBERTa-base	0.441	0.294	0.290	0.264	N/A	N/A
Audio	Baseline	0.316	**0.409**	0.345	0.324	0.327	0.311
Audio + Text	Mean Fusion	0.488	0.345	0.353	0.331	0.351	0.333
Audio + Text	MLP-Fusion	0.395	0.363	0.368	**0.343**	0.237	0.181
Audio + Text	Weighted Fusion	0.486	0.346	0.351	0.328	0.352	**0.337**

shows that our systems are all less efficient on the development dataset than the baseline which has a macro F1-score of 0.409 (against 0.363 for our MLP fusion system). Nevertheless, we achieved better performance on our balanced development set and the test set than the baseline. Indeed, our MLP-Fusion system has a F1-score of 0.343 against 0.324 for the baseline. On the test set, our weighting fusion system has the better performance with a F1-score of 0.337 against 0.311 for the baseline.

4. Conclusion

For the Odyssey 2024 challenge, we proposed a hybrid system merging audio and text modeling. We succeeded in obtaining a macro F1-score of 0.3367, surpassing the proposed baseline system. The best audio system proposed is composed of a WavLM-Large parameter extraction module, followed by two BLSTM layers and a linear layer for emotion classification. The text model is based on a transcription performed by Whisper, then implements a RoBERTa model. In the framework of the challenge, we have only exprimented with linear combinations of the two systems but in the future, it will be interesting to try and take into account each audio and text emotion classifier and learn the best combinations of both, for example weighting them differently according to classes. It would also be interesting to merge several other models, both on audio and text sides, especially LLMs.

5. Acknowledgements

This work has been partially supported by the French national research and technology agency ANRT, and the Coram project, an Audio Mobility 2030 project funded by BPI France. It was also granted access to the HPC resources of IDRIS under the allocation AD011014274 made by GENCI.

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2024-AD011014274 on the supercomputer Jean Zay with the V100 partition.

6. References

[1] Carlo Strapparava and Rada Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008.

[2] Saif M. Mohammad, "Chapter 11 - Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text," in *Emotion Measurement (Second Edition)*, Herbert L. Meiselman, Ed. Woodhead Publishing, second edition edition, 2021.

[3] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal, "SemEval-2019 task 3: EmoContext contextual emotion detection in text," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[4] Ankita Bhaumik, Andy Bernhardt, Gregorios Katsios, Ning Sa, and Tomek Strzalkowski, "Adapting emotion detection to analyze influence campaigns on social media," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 2023, pp. 441–451.

[5] Pablo Sánchez-Núñez, Manuel J Cobo, Carlos De Las Heras-Pedrosa, José Ignacio Peláez, and Enrique Herrera-Viedma, "Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis," *IEEE Access*, vol. 8, pp. 134563–134576, 2020.

[6] Yan Sun, Changqin Quan, Xin Kang, Zuopeng Zhang, and Fuji Ren, "Customer emotion detection by emotion expression analysis on adverbs," *Information Technology and Management*, vol. 16, pp. 303–311, 2015.

[7] Keith Harrigian, Carlos Aguirre, and Mark Dredze, "On the state of social media data for mental health research," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik, Eds., Online, June 2021, pp. 15–24, Association for Computational Linguistics.

[8] Ayah Zirikly, Dana Atzil-Slonim, Maria Liakata, Steven Bedrick, Bart Desmet, Molly Ireland, Andrew Lee, Sean MacAvaney, Matthew Purver, Rebecca Resnik, and Andrew Yates, Eds., *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, Seattle, USA, 2022. Association for Computational Linguistics.

[9] Javier Parapar, David E. Martin-Rodilla, Patricia ans Losada, and Fabio Crestani, "Overview of eRisk at CLEF 2023: Early Risk Prediction on the Internet," in *CLEF - CEUR-WS Working Notes*, 2023.

[10] Yi Ji Bae, Midan Shim, and Won Hee Lee, "Schizophrenia detection using machine learning approach from social media content," *Sensors*, vol. 21, no. 17, pp. 5924, 2021.

[11] Antonio Molina, Xinhui Huang, Lluís-F. Hurtado, and Ferran Pla, "ELiRF-UPV at eRisk 2023: Early detection of pathological gambling using SVM," in *CLEF - CEUR-WS Working Notes*, 2023.

[12] Flor Miriam Plaza-del Arco, Alba Cercas Curry, and Dirk Hovy, "Emotion analysis in NLP: Trends, Gaps and Roadmap for Future Directions," *LREC-COLING Proceedings 2024*, February 2024.

[13] Florian Eyben, Martin Wöllmer, and Björn Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM), ACM*, 2010.

[14] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, arXiv preprint arXiv:1804.05788.

[15] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa, "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding," in *Proc. Interspeech 2018*, Hyderabad, India, 2018, pp. 3688–3692.

[16] Bagus Tris Atmaja, Kiyoaki Shirai, and Masato Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, 2019.

[17] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822–2826.

[18] B.T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Transactions on Signal and Information Processing*, vol. 9, no. 1, 2020.

[19] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, 2019.

[20] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L. Paola 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.

[21] 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, Oct. 2022, arXiv:2110.13900 [cs, eess].

[22] Koichi Shinoda Koji Okabe, Takafumi Koshinaka, "Attentive statistics pooling for deep speaker embedding," in *Proc. INTERSPEECH*, 2018.

[23] Shiqing Zhang, Yijiao Yang, Chen Chen, Ruixin Liu, Xin Tao, Wenping Guo, Yicheng Xu, and Xiaoming Zhao, "Multimodal emotion recognition based on audio and text by using hybrid attention networks," *Biomedical Signal Processing and Control*, vol. 85, pp. 105052, 2023.

[24] Thong Nguyen, Xiaobao Wu, Anh Tuan Luu, Zhen Hai, and Lidong Bing, "Adaptive contrastive learning on multimodal transformer for review helpfulness prediction," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, Eds., Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 10085–10096, Association for Computational Linguistics.

[25] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[27] Alexis Plaquet and Hervé Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. INTERSPEECH 2023*, Dublin, Ireland, 2023, pp. 3222–3226.

[28] Hervé Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, Dublin, Ireland, 2023, pp. 1983–1987.

[29] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

[30] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[31] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, arXiv:1907.11692.

[34] William C. Sleeman, Rishabh Kapoor, and Preetam Ghosh, "Multimodal classification: Current landscape, taxonomy and future directions," *ACM Comput. Surv.*, vol. 55, no. 7, dec 2022.

[35] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2017, ICMI '17, p. 569–576, Association for Computing Machinery.