Interspeech 2024
1-5 September 2024, Kos, Greece

SAMSEMO: New dataset for multilingual and multimodal emotion recognition

Paweł Bujnowski[1], *Bartłomiej Kuźma*[1], *Bartłomiej Paziewski*[1], *Jacek Rutkowski*[1], *Joanna Marhula*[1], *Zuzanna Bordzicka*[1], *Piotr Andruszkiewicz*[1,2]

[1]Samsung Research, Poland
[2]Warsaw University of Technology, Poland

{p.bujnowski,b.kuzma,b.paziewski,j.rutkowski2,j.marhula,
z.bordzicka,p.andruszki2}@samsung.com

Abstract

The task of emotion recognition using image, audio and text modalities has recently attained popularity due to its various potential applications. However, the list of large-scale multimodal datasets is very short and all available datasets have significant limitations. We present SAMSEMO, a novel dataset for multimodal and multilingual emotion recognition.

Our collection of over 23k video scenes is multilingual as it includes video scenes in 5 languages (EN, DE, ES, PL and KO). Video scenes are heterogeneous, they come from diverse sources and are accompanied with rich manually collected metadata and emotion annotations.

In the paper, we also study the valence and arousal of audio features of our data for the most important emotion classes and compare them with the features of CMU-MOSEI data. Moreover, we perform multimodal experiments for emotion recognition with SAMSEMO and show how to use a multilingual model to improve the detection of imbalanced classes.

Index Terms: multimodal emotion recognition, multilingual models, emotion recognition dataset

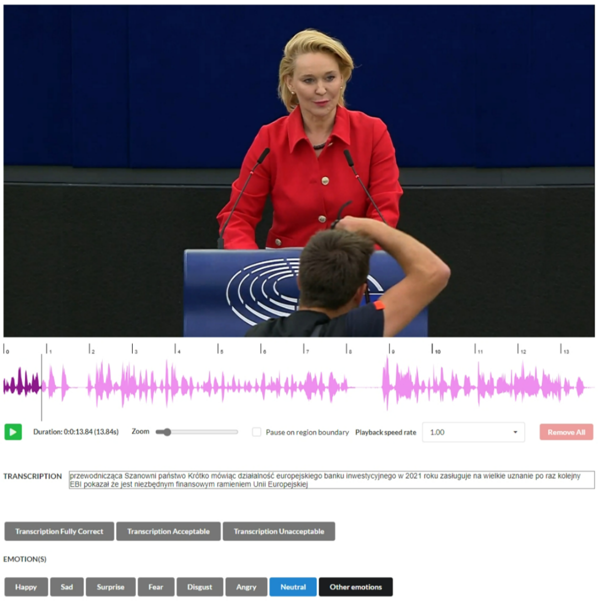
1. Introduction

Multimodal emotion recognition (MER) task has gained popularity due to the multifaceted analysis of human communication involving image, voice and text. Machine learning approach to MER can be compared to the medical research on human perception where human brain engages overlapping and non-overlapping neuroanatomical systems responsible for different emotional stimuli and affect production [1, 2]. For humans, processing of signals across different modalities enables holistic assessment of emotions.

Our research concerning MER was affected by the lack of diverse and multilingual datasets. To our knowledge, most of the accessible large-scale MER datasets are in English and have various drawbacks, which we point out in Section 1.2. In this paper, we present a new dataset for 5 languages, hoping that it can further advance the research on multilingual MER.

1.1. SAMSEMO overview

The main contribution of this paper is SAMSEMO: Samsung Multimodal and Multilingual Dataset for Emotion Recognition. The dataset contains 23,086 video scenes in 5 languages coming from approximately 1.4k speakers and a wide range of video genres (see Section 2.2). All video scenes were manually transcribed and annotated for emotions. Although the dataset has been designed primarily for multimodal emotion recognition, the presence of accompanying metadata (gender, language, indication if one or more faces are visible within a video scene etc.) makes it applicable to various tasks involving different combinations of text, vision and audio signals. SAMSEMO annotations and metadata are freely available to the scientific community under the CC BY-NC-SA 4.0 license.[1]



Duration: 0:0:13.84 (13.84s)

Playback speed rate 1.00

przewodnicząca Szanowni państwo Krótko mówiąc działalność europejskiego banku inwestycyjnego w 2021 roku zasługuje na wielkie uznanie po raz kolejny EBI pokazał że jest niezbędnym finansowym ramieniem Unii Europejskiej

Figure 1: *SAMSEMO building (here Europarl proceedings)*

The primary contribution is accompanied by our uni- and multimodal experiments with SAMSEMO: the results of emotion recogition for English, German, Spanish, Polish and Korean as well as the combination of these language subsets in a multilingual model, reused from [3], set a baseline for this task.

1.2. Related datasets

Table 1 presents SAMSEMO alongside previously published datasets for emotion recognition and sentiment analysis. Among them, CMU-MOSEI [4], a monolingual dataset for English, seems to be frequently exploited in MER studies [5, 6, 7]. However, despite its remarkable size and diversity of topics and speakers, we have identified two major shortcomings within the provided annotations: (a) an alarming number of utterances were labeled with contradictory emotions, e.g. Happiness and Sadness; (b) there is a striking imbalance in the number of videos per annotator, which results in a significant annotator

[1] The video materials are shared under their original licenses. Dataset link: `https://github.com/samsungnlp/samsemo`

10.21437/Interspeech.2024-212

Table 1: *Comparison of SAMSEMO and other multimodal emotion recognition and sentiment analysis datasets.*

Dataset	Samples	Speakers	Languages	Duration (h)	Modalities	Emotion/Sentiment
SAMSEMO	23086	∼1390	EN, DE, ES, PL, KO	38:34	l,v,a	e
CMU-MOSEI	23453	1000	EN	65:53	l,v,a	e,s
CMU-MOSEAS	40000	1654	DE, ES, FR, PT	68:49	l,v,a	e,s
MELD	13708	∼260	EN	∼13:00	l,v,a	e,s
IEMOCAP	10000	10	EN	11:28	l,v,a	e
AMMER	288	36	DE	01:18	l,v,a	e
CMU-MOSI	2199	98	EN	02:36	l,v,a	s
K-EmoCon	4159	32	KO	02:53	v,a	e
CH-SIMS	2281	474	ZH	∼02:30	l,v,a	s

bias (some annotators had a larger influence on the final labels than others, which we find especially problematic in the context of emotion annotation). IEMOCAP [8], whose considerable size of 10k scenes makes it also a popular monolingual MER dataset [7, 9, 10], is characterized by a rather low diversity of speakers (only 10 actors). The same limitations apply to MELD [11], a dataset containing scenes from the TV show *Friends*, where more than 80% of scenes come from 6 actors. The availability of multilingual MER and sentiment analysis datasets is very restricted, the only exception being CMU-MOSEAS [12] (currently inaccessible), and thus studies on multilingual data are rarely mentioned in the literature. The same issue applies to monolingual datasets for languages other than English, with a few exceptions of small-scale datasets for German [13], Korean [14] or Chinese [15].

2. SAMSEMO Dataset

2.1. Data collection, verification and annotation

SAMSEMO dataset collection process started with a manual search for videos with open licences (public domain, CC0, CC BY and CC BY-SA) available online. Wikimedia Commons and Vimeo platform became the main sources of such data for English, Spanish, German and Korean. For Polish, due to the low availability of videos on the aforementioned platforms, we used also the European Parliament proceedings.

As the second step, linguists worked on data collection. A group of at least C1 users of the target language were asked to watch the videos and manually perform a preliminary emotional video scene selection. The linguists defined the duration of the proposed video scenes (pointing at the scene beginning and ending) and annotated them for speaker's gender, language used and emotion(s) occurring in the scene. Next, the video scenes were cut out from the source videos, manually verified and, in some cases, edited to ensure the best picture and sound quality (9% of initial video scenes were rejected due to low picture/sound quality or unintelligible speech). Also, at this stage the linguists checked the correctness of the automatic transcription of each video scene and improved it where necessary.

All verified video scenes were annotated for emotions by three judges, all of them fluent speakers of the target language (initial emotion labels from the data collection process were discarded). To reduce annotator bias (control the impact of a few annotators across final annotation decisions), we assigned a similar number of video scenes to each annotator working within one language dataset. For emotion categories, we followed CMU-MOSEI and used Ekman emotions [16] of Happiness, Sadness, Anger, Fear, Disgust and Surprise. However, we extended the emotion labels by adding Neutral and Other categories. Thus, the annotators had three annotation options: they could (a) select maximally two Ekman labels for a given scene, (b) choose 'Neutral' to denote that the speaker does not express any emotions, (c) point at 'Other' where the Ekman basic six emotions were inadequate. Inter-Annotator Agreement, defined as the proportion of instances where all annotators agreed on at least one label, stands at 39.37%, indicating moderate agreement. For experiments with SAMSEMO (Section 3) we have chosen the majority voting scheme to select the final label for each video scene. This way, out of more than 23k of video scenes for all languages we receive 20,010 video scenes with clear emotion labels. Nevertheless, for further experiments with our dataset, we encourage exploring other emotion label aggregation strategies.

2.2. Quantitative analysis of dataset

Our dataset is composed of video scenes in 5 languages: English, German, Spanish, Polish and Korean. A detailed quantitative summary of the dataset for all languages is presented in Table 2. The scenes come from 14 different video genres, out of which debate, interview, documentary, movie, speech and vlog are the most frequent. The distribution of video genres per language differs, but each language dataset contains scenes from 10 to 12 video types (in English, Spanish and Korean interview is the most frequent video genre, in Polish - debate, German - movie). Average scene length varies between the languages from 5.65 seconds for English to 8.1 for Polish. Such difference may result from the specifics of each language, but also from video types dominating in each language subset.

SAMSEMO is also characterized by imbalanced proportions of emotion categories. Happiness, Anger and Neutral are more frequent in types of videos we have found, while Fear, Disgust or Surprise are observed very rarely. Thus 21% of all video scenes are labeled as Happiness, 41% as Neutral, 7% as Anger, 5% as Sadness, 3% as Surprise, 1% as Disgust, <1% as Fear and 5% as Other emotions (note that those values do not sum up to 100% as one scene can have multiple labels or no label whatsoever).

2.3. Valence and arousal model of audio data

According to emotion psychology research [17, 18], emotions can be predicted on the basis of 2 dimensions: valence and arousal. Valence relates to the positiveness of the emotion (from -1.5 to 1.5) and arousal refers to its intensity (from -1 to 1).

One of our experiments, using [19], focused on calculating both factors for separated audio tracks for every scene from our dataset. The results are shown in Figure 2. We observe more

Table 2: *Summary of SAMSEMO dataset statistics for all languages.*

Language	**EN**	**DE**	**ES**	**PL**	**KO**
Total number of video scenes	6924	4152	2674	7303	2033
Total number of source videos	413	271	250	672	212
Number of source video types	11	10	12	11	11
Approx. number of distinct speakers	390	260	240	300[*]	200
Average length of video scenes (in seconds)	5.65	5.75	6.74	8.10	6.48
Average length of video scenes (in words)	15	14	17	17	12
Total length of video scenes (in hours)	10:52	06:38	05:00	16:26	03:40
Total number of words in video scenes	101923	57849	44591	126619	23678
Vocabulary size (unique words)	12039	11540	7757	25208	9160
Speech rate (no. of words per second)	2.60	2.42	2.47	2.14	1.79

[*] Reduced speaker count attributed to European Parliament-sourced videos.

positive valence values (greater than zero) for scenes labeled as Happiness, while for the rest of the audio files the values lower than zero prevail, thus they can be predicted to contain negative emotions. Moreover, arousal for Happiness and Anger is much closer to 1, which indicates very intense emotions. For Sadness intensification is less expressive. These results imply that SAMSEMO is well-balanced and labels for the majority of the video scenes are correct. We also calculated valence and arousal for the CMU-MOSEI dataset to compare it with our data for English. Table 3 contains the results for the 4 largest classes - Anger, Happiness, Sadness and Neutral. We found 6 significant differences in mean values for both datasets (we marked them by a-f in Table 3 with bold for higher values) using Student t-test and Holm-Bonferroni correction for multiple comparisons (all p-values were lower than 0.01). We noticed the biggest difference for valence in results for Sadness: lower positiveness of SAMSEMO videos compared to CMU-MOSEI. The strongest dissimilarity in terms of arousal can be observed for Anger with higher expression in SAMSEMO scenes than in CMU-MOSEI. Results show that the combination of both datasets for audio or multimodal research, e.g. MER, could be challenging.

Table 3: *Comparison of the valence and arousal audio models of 4 biggest classes from 2 English MER datasets: CMU-MOSEI and SAMSEMO - here only for English data*

		SAMSEMO		**CMU-MOSEI**	
		VAL	**ARO**	**VAL**	**ARO**
Ang	*Mean*	-0.11	**0.38**^{a}	-0.11	0.20^{a}
	Median	-0.11	0.41	-0.10	0.24
	Std	0.51	0.28	0.51	0.30
Sad	*Mean*	-0.20^{b}	0.08^{c}	**-0.05**^{b}	**0.13**^{c}
	Median	-0.20	0.10	-0.04	0.16
	Std	0.58	0.33	0.53	0.30
Hap	*Mean*	0.00^{d}	**0.31**^{e}	**0.08**^{d}	0.25^{e}
	Median	0.02	0.33	0.11	0.27
	Std	0.54	0.31	0.53	0.29
Neu	*Mean*	0.00	**0.25**^{f}	0.01	0.19^{f}
	Median	-0.01	0.28	0.03	0.21
	Std	0.50	0.29	0.53	0.29

3. Numerical experiments with SAMSEMO

To test SAMSEMO in practice, we conducted experiments with multimodal emotion classification. The dataset can be used for studies concerning various aspects included in the metadata, e.g. gender, age, race, etc., similarly as in [20]. However, we eventually decided to focus only on the image, audio and textual layers of video scenes in our experiments. We performed a range of ablation studies by controlling the impact of modality signals (unique or combined) for emotion detection accuracy.

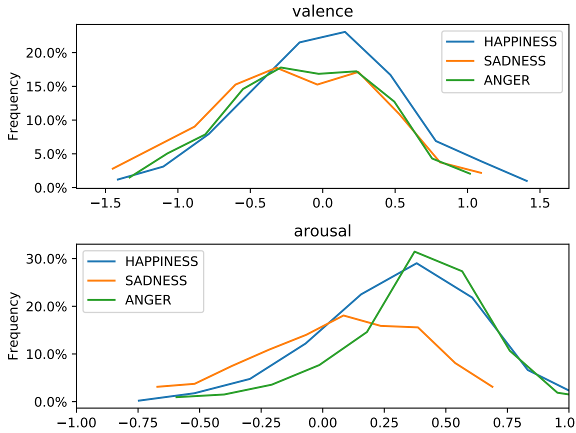


Figure 2: *Valence and arousal dimensions of the audio modality. Plots of Happiness, Sadness and Anger classes from the English data of the SAMSEMO corpus.*

3.1. End2End models for MER

As a benchmark model we used the system for MER [3] that was verified in our previous experiments and demonstrations in [21]. We marked it as $E2E$ model.

To reduce computations of the chosen baseline model, we introduced another simplified model (which we call *E2E-L*, *L* for light) with a new preprocessing phase. Such an approach was highly inspired by [22]. The preprocessing phase consisted of three steps: a) retrieval of a face from images; b) getting spectograms out of audio files; c) extracting BERT features from texts. These computations are performed once per every epoch in the *E2E* model and only once, before the training process, in the *E2E-L* model, which leads to an approximately 10x speedup during the training. In the *E2E-L* model, we also introduced some minor changes: dropout regularization, reduction of the number of images per scene to 16 and reduction of the number of spectrograms to 32. We made *E2E-L* model available as open source.[2]

3.2. Monolingual MER experiments

Typically, MER models are trained for separate languages [12]. We started with a similar approach and trained the models independently for 5 languages available in SAMSEMO. *E2E* and *E2E-L* models were trained for 70 epochs using various combinations of modalities in the input: VAT, VA, VT, AT, V, A, T (V-Video, A-Audio, T-Text). We split the data for train, validation and test subsets with respective ratio: 70%, 15%, 15%. For the light model *E2E-L* we used Optuna hyperparameter optimization framework [23], which slightly improved the results. For *E2E* training we used hyperparameters given in [3], our modification being the usage of "bert-base-multilingual-uncased" as a text model [24] to deal with various languages in the input.[3] For these experiments, we used a selection of videos featuring only one person throughout the video scene from SAMSEMO dataset, with the following volumes: EN: 3898, DE: 2374, ES: 1854, PL: 5119, KO: 1279. Due to frequency issues, we combined Anger and Disgust emotions into one class named Anger+ and removed Fear emotion category. Finally, we ran the multi-label classification of five emotions. Results of the experiments are presented in Table 4. Among the five languages, Polish ($F1 = 76.7$) achieved the best results, while German turned out to be the most challenging ($F1 = 65.3$). Polish SAMSEMO subset has the largest amount of data with the lowest speaker diversity among languages, which could have positively affected the final results.

Table 4: *Validation of MER models (with F1 measure in %) splitted for modality combinations. Abbreviations: EN-English, DE-German, ES-Spanish, PL-Polish, KO-Korean; ML - multilingual training with combined data (EN, DE, ES, PL and KO). E2E and E2E-L stand for benchmark MER models.*

MODEL	EN	DE	ES	PL	KO	ML
VAT$_{\text{E2E}}$	69.9	64.0	63.9	75.4	67.5	**69.0**
VAT$_{\text{E2E-L}}$	**71.5**	64.5	71.5	**76.7**	66.7	64.6
VA$_{\text{E2E}}$	68.6	64.4	67.0	74.5	65.8	68.6
VA$_{\text{E2E-L}}$	68.8	63.4	67.2	75.3	69.0	64.1
VT$_{\text{E2E}}$	68.5	63.0	66.4	74.2	**71.1**	67.5
VT$_{\text{E2E-L}}$	71.8	63.6	71.6	76.0	67.2	62.7
AT$_{\text{E2E}}$	57.3	49.5	58.3	66.1	58.8	59.9
AT$_{\text{E2E-L}}$	61.1	59.5	59.9	67.4	63.1	61.7
V$_{\text{E2E}}$	68.6	**65.3**	67.6	74.7	69.3	68.2
V$_{\text{E2E-L}}$	69.5	65.1	**74.0**	74.8	67.2	64.4
A$_{\text{E2E}}$	60.1	56.0	56.2	65.9	61.4	61.1
A$_{\text{E2E-L}}$	58.7	59.5	60.9	69.3	64.2	59.9
T$_{\text{E2E}}$	58.5	54.5	58.9	69.0	69.3	63.0
T$_{\text{E2E-L}}$	51.8	45.9	56.4	60.8	62.1	51.9

Among all verified modalities, we found that video was the most important signal. This observation stands in contrast to other studies where text [3] or audio models [4] for English data obtained the best scores or scores equal to other modality models. In our experiments, the video models achieved the best results for German and Spanish. Also in two cases, for EN and PL, the VAT$_{\text{E2E-L}}$ model performed best, and for Korean the leader was the VT$_{\text{E2E}}$ model. Improving the multimodal architecture to leverage better video, audio and text features remains a challenge. We suppose that the application of pre-trained audio models, like wav2vec 2.0 [25] or HuBERT [26], could improve the model, which can be verified in future studies.

3.3. Multilingual MER experiments

In Table 4 we show the results of the multilingual models (ML) that were trained on the combination of all monolingual datasets (the ML train/val/test data consist of corresponding parts of all the 5 languages). For the German testset, the ML VAT$_{\text{E2E}}$ model gained F1=65.5% versus 64.0% of the monolingual model, and for Spanish: 65.7% vs. 63.9%. These differences indicate the benefits of using ML models for difficult or small datasets.

Table 5: *Validation of VAT$_{\text{E2E}}$ models considering emotion classes. Anger+ (Ang+) includes both Anger and Disgust. The values shows F1 measure (%). Word abbreviations: EMO-Emotion, Hap-Happiness, Sad-Sadness, Sur-Surprise, Neu-Neutral, Ave-Avarage. Other shortcuts explained in Table 4.*

EMO	EN	DE	ES	PL	KO	ES-ML
Ang+	54.9	67.5	21.1	54.7	36.4	50.0
Hap	81.2	67.1	67.9	71.8	66.0	69.9
Sad	34.7	55.2	50.0	36.7	22.2	50.0
Sur	18.2	26.1	10.0	5.3	12.5	24.4
Neu	77.6	69.0	76.1	86.1	78.6	78.9
Ave	69.9	64.0	63.9	75.4	67.5	69.2

In our experiments, we also focused on emotion classes. In Table 5 we present the results for each emotion category for the VAT$_{\text{E2E}}$ model from Table 4. Lower results for some emotions show imperfections of the dataset such as the imbalance of emotion classes. To improve the results for Spanish, we built the multilingual model (ES-ML) with the original ES data combined with the training data from EN and DE for Anger+, Sadness and Surprise. Valid and test ES subsets were not changed. We observed a strong improvement for Anger+ and Surprise classes as well as in the average results.

4. Data limitations

SAMSEMO includes various video types with their different proportions in all languages. Also, the five languages constituting SAMSEMO embrace a range of their variants (e.g. in English dataset, speakers use British, American and other English varieties). Importantly, Fear and Disgust are represented by relatively few scenes compared to other emotions, which makes it difficult to build efficient models for all emotion categories in all languages. Another limitation is the potential for human error within some of the SAMSEMO labels, which we tried to reduce by selecting majority voting for the final label selection.

5. Conclusion

We have proposed the multilingual and multimodal SAMSEMO dataset designed primarily for the MER task but with possible extensions for other research areas. A few MER experiments were performed to create a benchmark for models and to indicate data challenges. Furthermore, we showed two examples of multilingual models that can be easily trained with our data and that can solve some dataset drawbacks, for example, emotion category imbalance. We hope this direction will be explored further in future MER studies.

[2] Code link: `https://github.com/samsungnlp/samsemo`

[3] https://huggingface.co/google-bert/bert-base-multilingual-uncased

6. References

[1] M. Klasen, Y.-H. Chen, and K. Mathiak, “Multisensory emotions: perception, combination and underlying neural processes,” *Reviews in the Neurosciences*, vol. 23, no. 4, pp. 381–392, 2012. [Online]. Available: https://doi.org/10.1515/revneuro-2012-0040

[2] A. Schirmer and R. Adolphs, “Emotion perception from face, voice, and touch: Comparisons and convergence,” *Trends in Cognitive Sciences*, vol. 21, 02 2017.

[3] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, “Multimodal end-to-end sparse model for emotion recognition,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 5305–5316. [Online]. Available: https://aclanthology.org/2021.naacl-main.417

[4] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://aclanthology.org/P18-1208

[5] A. Takashima, R. Masumura, A. Ando, Y. Yamazaki, M. Uchida, and S. Orihashi, “Interactive co-learning with cross-modal transformer for audio-visual emotion recognition,” in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 4740–4744. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-11307

[6] C. Peng, K. Chen, L. Shou, and G. Chen, “Carat: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition,” *ArXiv*, vol. abs/2312.10201, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266348665

[7] S. Hong, H. Kang, and H. Cho, “Cross-modal dynamic transfer learning for multimodal emotion recognition,” *IEEE Access*, vol. 12, pp. 14 324–14 333, 2024. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3356185

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[9] H. Dhamyal, B. Raj, and R. Singh, “Positional encoding for capturing modality specific cadence for emotion detection,” in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 166–170. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-11085

[10] 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*, vol. abs/2312.15185, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266551115

[11] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 527–536. [Online]. Available: https://doi.org/10.18653/v1/p19-1050

[12] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L. Morency, “CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 1801–1812. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.141

[13] D. Cevher, S. Zepf, and R. Klinger, “Towards multimodal emotion recognition in german speech events in cars using transfer learning,” in *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*, 2019. [Online]. Available: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_16.pdf

[14] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. J. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, “K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,” *CoRR*, vol. abs/2005.04120, 2020. [Online]. Available: https://arxiv.org/abs/2005.04120

[15] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, “CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 3718–3727. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.343

[16] P. Ekman, W. V. Freisen, and S. Ancoli, “Facial signs of emotional experience,” *Journal of Personality and Social Psychology*, vol. 39(6), pp. 1125–1134, 1980.

[17] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, “Speech emotion estimation in 3d space,” in *2010 IEEE International Conference on Multimedia and Expo*, 2010, pp. 737–742.

[18] J. Russell, “Core affect and the psychological construction of emotion,” *Psychological Review*, vol. 110, pp. 145–172, 01 2003.

[19] T. Giannakopoulos, “pyaudioanalysis: An open-source python library for audio signal analysis,” *PloS one*, vol. 10, no. 12, 2015.

[20] T.-W. Sun, “End-to-end speech emotion recognition with gender information,” *IEEE Access*, vol. PP, pp. 1–1, 08 2020.

[21] P. Bujnowski, B. Kuźma, B. Paziewski, J. Rutkowski, J. Marhula, Z. Bordzicka, and P. Andruszkiewicz, “”Select language, modality or put on a mask!” Experiments with Multimodal Emotion Recognition,” in *Proc. INTERSPEECH 2023*, 2023, pp. 672–673.

[22] J. Cheng, I. Fostiropoulos, B. Boehm, and M. Soleymani, “Multimodal phased transformer for sentiment analysis,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2447–2458. [Online]. Available: https://aclanthology.org/2021.emnlp-main.189

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, p. 2623–2631.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[25] A. Baevski, H. Zhou, A. Mohamed, and M. 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*, ser. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020, p. 12449–12460.

[26] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291