TalTech Systems for the Odyssey 2024 Emotion Recognition Challenge

Henry Härm, Tanel Alumäe

Institute of Software Science
Tallinn University of Technology, Estonia
henry.harm@taltech.ee, tanel.alumae@taltech.ee

Abstract

Odyssey 2024 Emotion Recognition Challenge aims to compare different emotion recognition systems in two tasks: classifying speech across eight emotional classes and predicting emotional attributes for arousal, valence and dominance. This paper describes TalTech's systems prepared for the challenge that fuse the predictions of text and speech based emotion recognition models. The audio-based model adapts the Wav2Vec2-BERT model for emotion recognition, while the text-based model uses finetuned LLaMA2-7B as the backbone. The two models are combined for the classification task by training a multi-class logistic regression model, using the posteriors of the underlying models as input features. The model obtained a macro-F1 score of 0.354 on evaluation data and was ranked 2nd among all teams. The fusion model for the attribute prediction task achieved an average score of 0.5144 and was thereby ranked 6th among the teams.

1. Introduction

Speech, as the most instinctual mode of expression, conveys not only the intended message but also encodes rich information about the speaker's identity, emotional state, and language. Recognizing the emotional content in speech, or speech emotion recognition (SER), has been a field of study for over two decades [1]. This domain has broad applications ranging from improving human-computer interaction [2] to analyzing dialogues in call centers [3]. Despite its longstanding presence and wide applicability, emotion detection in speech presents significant challenges, primarily due to the subjective nature of emotions themselves. There is a notable lack of consensus on how emotions should be measured or categorized, which complicates the task of developing accurate and reliable SER systems [4].

SER systems must accommodate different speakers and languages while recognizing emotional aspects of the speech signal, while nullifying other aspects such as linguistic and cultural information [5]. Recent research have described the SER process to consist of three steps, which are data preprocessing, feature extraction, and classification of audio signals. Recently self-supervised transformer-based models have shown good performance in in different automatic speech processing tasks, with models such as wav2vec [6] and wav2vec2.0 [7] being proposed. The wav2vec2.0 framework enables self-supervised learning of speech representations by masking latent representations of the raw waveform and solving a contrastive task over quantized speech representations, achieving state of the art results in many speech processing benchmarks and evaluations, ranging from speech recognition [7] to spoken language identification [8]. It has also been successfully used for emotion recognition [9, 10, 11].

Alternatively, emotion recognition can be done in the text domain by firstly transcribing the speech. Emotion recognition from text is mainly based on identifying keywords and understanding the context, and consists of rule-based, classical learning-based, deep learning and hybrid methods. Pretrained language models (PLM) based on various architectures that have been trained on large unlabeled datasets, such as BERT [12], have shown good results in various NLP tasks. Scaling up language models has shown to be an effective way to improve performance in downstream tasks, leading to the introduction of large language models (LLMs). LLMs such as LLaMA 2 are pretrained on a large amounts of text using the next-word predicion task [13]. LLMs are characterized by their large parameter size which typically reaches tens of billions and possess stronger generalization capability across a wide range of downstream tasks. However, as LLMs are not designed for emotion understanding task, specific models can be trained as in [14], where an emotion and context knowledge enhanced LLM (DialogueLLM) was proposed. These models can be adeptly finetuned from publicly available LLMs using parameter-efficient methodologies such as low-rank adaptation (LoRA) [15], optimizing them for speech emotion recognition (SER) tasks.

This paper describes the systems built by the Tallinn University of Technology (TalTech) team for the Odyssey 2024 Emotion Recognition Challenge. Our models are based on two freely available foundation models: Wav2Vec2-BERT for speech and LLaMA2-7B for text. Both models are adapted to emotion recognition using additional output layers and the foundation model backbones are finetuned for emotion classification using LoRA. For the emotion classification task, the two models are combined using an additional logistic regression model that uses the posteriors of the underlying models as input. For emotional attribute prediction, we combine the models using attribute-specific interpolation weights. Our best systems were ranked second in the emotion classification task and sixth in the emotional attribute prediction task.

2. Data

The training, development and evaluation data for the challenge originates from the MSP-Podcast Corpus [16]. The training dataset encompasses 68,119 speaking turns, while the development set consists of 19,815 segments from 454 speakers. The test dataset includes 2,347 unique segments from 187 speakers, with labels that remain undisclosed to the public. The segments for the test set have been curated to maintain a balanced representation based on primary categorical emotions. The utterances in the corpus were selected from podcasts with spontaneous conversations. The average duration of utterances in the training set is 5.8 seconds. Emotional annotation of these utterances was conducted via crowdsourcing, where each utterance

Table 1: Different speech and text-based foundation models used in our experiments.

Model	Modality	#Parameters
Wav2Vec2-BERT	Speech	555M
SONAR-speech-encoder-eng	Speech	628M
RoBERTa-large	Text	355M
LLaMa2-7b	Text	6.74B
LLaMa2-70b	Text	69B

received multiple annotations to capture the perceived primary emotional category — such as anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise — as well as any applicable secondary categories. Additionally, annotators evaluated each utterance on three emotional attribute dimensions: valence (ranging from very negative to very positive), arousal (from very calm to very active), and dominance (from very weak to very strong), employing a seven-point Likert scale for each dimension. A singular, aggregate categorical label was assigned to each utterance based on the majority vote rule. The label "No Agreement" was reserved for utterances lacking a majority consensus. The aggregate mean values of emotional attributes are also provided for all labelled utterences.

The training and development datasets are also annotated by speaker identity and speaker gender. Human-made transcriptions and forced alignments between the transcript words and the audio is also provided.

Since there are no transcripts of the test data provided, we decided to automatically transcribe all data splits on our own, in order for the evaluation data transcripts to be consistent with training and development data. This was done using the NVIDIA NeMo Canary 1B[1] multilingual ASR model [17], using greedy decoding.

3. Methods

Our approach to emotion recognition is based on finetuning speech or text based foundation models. Table 1 provides a overview of the models with which we conducted our experiments.

3.1. Speech-based emotion recognition

For both emotion category and emotion attribute prediction directly from speech, we use a model based on the Wav2Vec2-BERT model[2] shared by the Seamless4MT project [18]. This model was pre-trained on 4.5M hours of unlabeled audio data covering more than 143 languages, using self-supervised loss. Wav2Vec2-BERT follows the same architecture as Wav2Vec2.0 [7], but replaces the attention-block with a Conformer-block as introduced in [19]. It also employs a causal depthwise convolutional layer and uses mel-spectrogram representation of the audio as input, instead of the raw waveform. Wav2Vec2-BERT uses Shaw-like position embeddings [20]. This particular Wav2Vec2-BERT model comprises 24 Conformer layers with approximately 600M parameters.

The Wav2Vec2-BERT model was adapted into an emotion classification model by aggregating its outputs with an attentive pooling layer, followed by a fully connected layer featuring ReLU activation and BatchNorm, and the final output layer, corresponding to the emotion categories of the training dataset. This model categorizes emotions into eight primary classes, along with "O" (Other) and "X" (No agreement) as additional distinct classes for training purposes. Training uses cross-entropy loss on random 2 to 4 second chunks of emotion-labeled utterances, employing consensus-based labels derived from plurality voting rather than directly utilizing individual annotator labels. To enhance model robustness, on-the-fly data augmentation was applied using point source noises and simulated room impulse responses (RIRs) from the MUSAN corpus. The model underwent a training regimen spanning 10 epochs, with optimization via the Adam optimizer, a peak learning rate of 10^{-4}, weight decay 0.001 and an effective batch size of 64. Additionally, speed perturbation was applied to half of the training batches to further diversify training data. LoRA was employed to fine-tune the pre-trained model, optimizing its performance with a configuration ($\texttt{rank} = 32$, $\alpha = 32$ and $\texttt{dropout} = 0.05$). Due to the use of LoRA, there are only around 8 million trainable parameters. The emotion classification model was not directly used for prediction but it served as an embedding extractor. Utterance embeddings were derived from the output of the first dense layer following the pooling layer. A logistic regression model was then trained on these embeddings, excluding data labeled as "X" and "O" with feature normalization and dimensionality reduction to 15 using LDA. The training data is highly unbalanced with regard to the emotion categories. However, the emotion distribution in evaluation data is uniform. Therefore, we post-processed the trained logistic regression model to use uniform prior over the 8 emotion categories, by appropriately modifying the biases of the softmax layer.

For audio-based emotion attribute prediction, the model structure mirrored that of the emotion classification model, but replaced the final softmax layer with a $\tanh$ nonlinearity, followed by an additional linear layer with three outputs. Training initially used 3 to 4-second audio chunks, with subsequent fine-tuning on 6 to 8-second segments, using training data where the proportion of different categories had been balanced. This adjustment was done because it was noticed that for the neutral category (which was dominating in training data), the mean value for most attributes was considerably lower than for other categories. As a loss function, we used negative average concordance correlation coefficient of the current batch.

3.2. Text-based emotion recognition

For attribute and category prediction from the text, an open-source LLaMA 2[3] LLM with 7 billion parameters is fine-tuned for the two tasks. In the category prediction task, the model is initialized with a sequence classification head, which is a linear layer using the last token in order to do the classification, similar to other causal models. The linear layer is initialized with 8 features according to the number of categories to be predicted. The model is trained on automatically generated utterance transcriptions. In order to tackle the attribute prediction task, the model is initialized with a sequence classification head configured with three output features corresponding to the arousal, valence, and dominance attributes. The model is fine-tuned using a regression task with a loss function based on the mean concordance correlation coefficient of the attributes. LoRA adapter is used ($\texttt{rank} = 256$, $\alpha = 256$ and $\texttt{dropout} = 0.1$) for efficient fine-tuning over 40 epochs, using AdamW optimizer, a learning

[1] https://huggingface.co/nvidia/canary-1b

[2] https://huggingface.co/facebook/w2v-bert-2.0

[3] https://huggingface.co/meta-LLaMA/LLaMA-2-7b-hf

Table 2: Concordance correlation coefficient (CCC) scores of various models on Task 2 balanced development and evaluation data.

		Dev_{bal}				Eval			
ID		Valence	Arousal	Dominance	Average	Valence	Arousal	Dominance	Average
	Official baseline					0.6069	0.5667	0.4244	0.5327
#1	Wav2Vec2-BERT	0.6063	0.5442	0.4430	0.5312				
#2	+ finetuning on balanced data	0.6295	0.5712	0.4772	0.5593				
#3	RoBERTa	0.5863	0.2676	0.2740	0.3760				
#4	LLaMA2-7b	0.5932	0.2634	0.2532	0.3699				
#5	Fusion #2 + #4	0.6771	0.5712	0.4772	0.5752	0.6362	0.5417	0.3655	0.5144

Table 3: Macro F1 scores of various models in Task 1.

ID	Model	Dev_{filt}	Dev_{bal}	Eval
	Official baseline			0.311
#1	Wav2Vec2-BERT	0.241	0.169	
#2	+ Balanced prior	0.283	0.297	
#3	RoBERTA	0.292	0.268	
#4	+ Balanced prior	0.273	0.307	
#5	LLaMA2-7b	0.281	0.248	
#6	+ Balanced prior	0.276	0.311	
#7	Fusion: interpolation (#2 + #6)	-	0.354	0.350
#8	Fusion: log. reg (#2 + #6)	-	0.381	0.354
Post-evaluation experiments				
#9	SONAR-speech-eng + PLDA		0.327	
#10	LLaMA2-70b, balanced prior		0.334	
#11	Fusion: interpolation (#9 + #10)		0.362	

rate of 2×10^{-5} and a batch size of 16. As LLaMA does not have a pad token we set it to be the same as EOS token.

As an alternative method, we train a model using RoBERTa[4] [21] as the backbone. The model utilizes the same architecture as with the LLM model and is finetuned with similar hyper-parameters, however it is not necessary to utilize a LoRa adapter, as parameters can be finetuned directly.

In order to rebalance the predictions of the text-based model to follow a uniform prior over the categories, we "fix" the conditional probability distribution $P(y|x)$ returned by the emotion identification model for input x to use the uniform prior:

$$P'(y|x) = \frac{\frac{P_u(y)}{P'(y)} \times P(y|x)}{Z}$$

where Z is a normalizing factor, $P_u(y)$ a uniform prior ($\frac{1}{8}$ for the 8 main categories, 0 for the pseudo-categories "X" and "O") and $P'(y)$ is the prior probability of emotion categories in training data.

3.3. Combining outputs from audio and text-based models

3.3.1. Emotion category identification

We tried two approaches for fusing predictions from audio and text-based models. In the first approach we simply linearly interpolate the posterior probabilities produced by different models, using an interpolation coefficient optimized on the balanced development data. The optimal interpolation coefficient is found using grid search from the interval of $[0, 1]$ using a step size of 0.05.

In the second approach, we train a multi-class logistic regression model, using the posteriors of the underlying models as input features. L2 penalty with the value of 1.0 is used for avoiding overfitting.

3.3.2. Emotion attribute prediction

In the emotion attribute prediction task, we used linear interpolation for fusing the predictions of two models. For each attribute dimension, an optimal interpolation coefficient was found.

4. Results

We tested our models on the official development split of the provided dataset as well as its several subsets. Development data contains items whose consensus-based emotion category labels is "X" or "O". As the evaluation data was guaranteed to contain only items labelled using the eight primary emotion categories, we created a dataset Dev_{filt} where items labelled as "X" or "O" are removed. However, Dev_{filt} is still highly unbalanced with regard to the emotion categories. To fix this, we created another subset Dev_{bal} that contains 300 randomly selected items for each primary category, making it more similar to the evaluation dataset that is balanced across category labels. This subset was used for optimizing all hyperparameters and for training fusion models.

4.1. Emotion category prediction

The macro-F1 scores of different models on development and evaluation data (where available) are shown in Table 3. Among our individual models, it is surprising to see that the F1 scores of the text-based models surpass those of the audio-based models, with the LLaMA2-based model with balanced priors being the most accurate. However, interpolating its predictions with those of a audio-based model provides substantial improvements both on development and evaluation data. The optimal interpolation coefficient of the text-based model was 0.45. Fusion using logistic regression gives noticeably larger improvements on development data than on evaluation data, suggesting moderate overfitting on development data. Our two fusion systems were ranked 2nd and 3rd among all submissions of Task 1 of the challenge.

[4] https://huggingface.co/FacebookAI/roberta-large

4.2. Emotion attribute prediction

Table 2 presents the CCC scores for various emotional attribute prediction models, evaluated on both the balanced development data and the evaluation dataset. The findings underscore the significance of fine-tuning the audio-based emotion model with balanced data, as this process notably enhances CCC scores across all evaluated attributes. Intriguingly, the text-based model demonstrates an impressive accuracy in predicting valence, with its optimal interpolation coefficient in the fusion identified at 0.55. However, for the other two categories — arousal and dominance — the text-based model's predictions do not contribute positively when optimized on development data, and the corresponding optimal interpolation coefficients were set to zero.

Despite these adjustments, our fusion model, which integrates both audio and text-based predictions, did not surpass the baseline model in terms of the average CCC metric on evaluation data. While our model achieved the highest accuracy in valence prediction among all submissions in the challenge, it significantly underperformed in predicting arousal and dominance. The discrepancy in performance across different attributes has yet to be fully understood and resolved. In terms of the average CCC score, our best submission was ranked 6th in the team ranking.

5. Post-evaluation experiments

In the post-evaluation phase, we experimented with some additional pretrained models. First, we finetuned the English SONAR speech encoder model[5] [22] for emotion identification. SONAR speech encoders are based on the finetuned Wav2Vec2-BERT model, with an additional sentence embedding layer that is trained to map spoken utterance embeddings into the same semantic space as the corresponding text-based sentence embeddings. The embedding layer uses simple mean pooling. We didn't use the SONAR model as-is, but finetuned it for emotion recognition using LoRA, using the same hyperparameters as when training the Wav2Vec2-BERT based model. Using SONAR embeddings instead of a self-supervised model is partly inspired by previous research that has shown that en emotion recognition model benefits from pretraining on an ASR task [11]. Row #9 in Table 3 shows that the SONAR-based model indeed outperforms Wav2Vec2-BERT based model. In this model, we also replaced logistic regression based classification of the utterance embeddings with PLDA-based scoring. PLDA is a generative model and does not depend on the prior probability of the different categories in the training data, eliminating thus the need for the prior balancing step.

Secondly, we finetuned the larger 70 billion parameter variant of the LLaMA2 model[6] for text based emotion classification. As the model is much larger, we utilized quantization to load the model weights and activations with lower 4-bit data types, which is a technique to reduce memory and computational cost. In addition we utilize nested quantization, which saves additional memory with minimal performance impact. Due to experiencing overflow issues while training the model with fp16 computational type, we selected bf16 for its larger dynamic range. These methods allowed us to finetune the model with our 80GB A100 GPUs. We finetuned our model for 4 epochs with a LoRa rank of 16, alpha of 16 and dropout of 0.05.

[5] https://github.com/facebookresearch/SONAR

[6] https://huggingface.co/meta-llama/Llama-2-70b-hf

Row #10 in Table 3 shows that the model outperforms our 7 billion parameter model wiht a F1 score of 0.334 on our balanced dataset. The model additionally outperformed the previously mentioned SONAR-based model. Row #11 shows that fusing the SONAR and larger LLaMA2 model outperforms previous interpolation of Wav2Vec2-BERT and LLaMA2-7b models results with an F1 score of 0.362.

6. Conclusion

Our results showed that combining state-of-the-art foundation models from the audio and text domains is efficient for emotion recognition, as it allowed us to obtain high-ranked results in both tasks in the Odyssey 2024 Emotion Recognition Challenge. Our experiments suggested that it is also important to tune the models' predictions to the expected prior probabilities of the evaluation data. The accuracy of our models in predicting emotional attributes however remained uneven, which suggests a possible direction for future work.

7. References

[1] Björn W Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] Mirosław Płaza, Robert Kazała, Zbigniew Koruba, Marcin Kozłowski, Małgorzata Lucińska, Kamil Sitek, and Jarosław Spyrka, "Emotion recognition method for call/contact centre systems," *Applied Sciences*, vol. 12, no. 21, pp. 10951, 2022.

[4] Mehmet Berkehan Akçay and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[5] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L Schneider, "Speech emotion recognition using machine learning—a systematic review," *Intelligent systems with applications*, p. 200266, 2023.

[6] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, 2020.

[8] Tanel Alumäe, Kunnar Kukk, Viet-Bac Le, Claude Barras, Abdel Messaoudi, and Waad Ben, "Exploring the impact of pretrained models and web-scraped data for the 2022 NIST language recognition evaluation," in *Proc. Interpseech*, 2023.

[9] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021.

[10] Li-Wei Chen and Alexander Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[11] Yuan Gao, Chenhui Chu, and Tatsuya Kawahara, "Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with ASR and gender pretraining," in *Proc. Interspeech*, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[14] Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin, "DialogueLLM: Context and emotion knowledge-tuned llama models for emotion recognition in conversations," *arXiv preprint arXiv:2310.11374*, 2023.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[16] 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, pp. 471–483, 2017.

[17] Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al., "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[18] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[20] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[21] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[22] Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot, "SONAR: sentence-level multimodal and language-agnostic representations," *arXiv preprint arXiv:2308.11466*, 2023.