Cross Culture Emotion Recognition

Avantika Singh, Bhargava Shrirama,
Kavya Gengedulla Yeranarasappa, Xiuyuan Guo

Viterbi School of Engineering, Department of Computer Science, USC
{singhava,shrirama,gengedul,guoxiuyu}@usc.edu

1 Problem Definition

The project is focused on developing an advanced multi-modal model designed for predicting emotion recognition across diverse cultural contexts. By integrating audio, visual, and textual data inputs, the model tries to capture the emotional expressions present in six distinct cultural backgrounds. The goal here is to find if there are significant differences in emotional expression across cultures or if emotions are universal.

2 Literature Review

The database utilized in this project was introduced in the paper by Kossaifi et al. (Kossaifi et al., 2021), aiming to offer a comprehensive audio-visual dataset portraying spontaneous human behavior observed in real-world settings. The SEWA DB comprises recordings of interactions stemming from diverse cultural backgrounds, carefully annotated with a plethora of behavioral cues and affective dimensions. Elfenbein and Ambady (Elfenbein H. A., 2002) conducted a meta-analysis on emotion recognition across cultures, delving into insights regarding the universality and cultural specificity of emotional expressions. Their findings suggest that while emotions are universally recognized at levels surpassing chance, nuances in recognition accuracy are influenced by cultural factors. They also indicate that individuals tend to comprehend emotions more accurately when expressed by members of their own cultural, ethnic, or regional group, pointing to an in-group advantage. Notably, cultural learning and expressive style emerge as significant determinants in shaping how emotions are both expressed and recognized across different cultures. The survey conducted by Lian et al (Lian et al., 2023) explores recent progress in deep learning-based multimodal emotion recognition, emphasizing the integration of speech, text, and facial cues. It discusses the progression of Multimodal Emotion Recognition from single-modality recognition to multimodal approaches and critically evaluates fusion methodologies, including early, late, hybrid, and intermediate layer fusion strategies, to enhance emotion recognition accuracy and effectiveness. (Jack et al., 2012) challenges the universality hypothesis of facial expressions, revealing cultural divergence in the mental representations of basic emotions. They find that Westerners depict each of the six basic emotions with distinct facial movements, contrasting with the less clear distinctions observed among Easterners.

Mani Kumar, Enrique Sanchez, et al. (T et al., 2021) propose a novel method that leverages temporal context in audio data, while also modeling culture-specific or person-specific factors drawn from a stochastic process. Their approach involves employing bidirectional GRU-RNNs for both the backbone model and the RNN encoder-decoder. Leena Mathur, Ralph Adolphs, et al. (Mathur et al., 2022) utilize the SEWA dataset to develop an Attention-Based Feature Selection (ABFS) approach within the temporal causal discovery framework. This framework aims to identify potential causal relationships between audio-visual features and valence/arousal labels within each source culture. They trained separate Temporal Convolutional Neural Networks for valence and arousal prediction, with each incorporating an attention mechanism to compute attention scores for each feature.

3 Data Description

The SEWA dataset provides a rich repository of annotated audio and 2D visual behavior data. With a comprehensive collection of 500,000 samples conducted over 199 sessions, the dataset encompasses recordings from 398 subjects representing different diverse cultural backgrounds and demographics who watches advertisements and engage

in discussions on it. Notably, the dataset features an equal gender distribution and spans an age range from 18 to 65 years. Culturally, the dataset covers six distinct backgrounds, including British, German, Hungarian, Greek, Serbian, and Chinese cultures, facilitating cross-cultural analyses of emotion recognition. The dataset has minimal missing values ensuring data integrity and reliability.

4 Method

In our study, Arousal, Valence, and Liking cues were integrated with text, audio, and video modalities. The experimental setup involved training on combined datasets of C1 (Chinese) and C4 (British) cultures, followed by testing on other cultures which are C2 (Hungarian), C3 (German), C5 (Serbian), C6 (Greek). Initially, binary classification tasks employed RNN for audio and video modalities, BERT for text data, followed by late fusion approach for all three modalities. Further, unimodal regression models were trained separately for video and audio, utilizing baseline models like RNN, LSTM (Hochreiter and Schmidhuber, 1997), LSTM-Attention (Wang et al., 2016), and BiGRU architectures predicting arousal and valence values, alongside a late fusion strategy. Inspired by the outstanding performance of transformers in sequential prediction(Luna-Jiménez et al., 2021), we propose a novel Channel-Attention Time Series Transformer architecture (CATS- Transformer). This architecture enables frame-by-frame continuous predictions of valence and arousal emotions.

4.1 Feature Extraction

Our study utilized pre-trained models to encode rich representations from text, audio and video modalities. The Whisper-large-v3 (Radford et al., 2022) model was used to extract text features from audio files in the original language and translated to English. The extracted sentences were tokenized using distilbert-base-uncased (Sanh et al., 2020), generating attention masks and standardized input IDs for further processing. Masked Autoencoder for facial video Representation LearnINg(MARLIN) (Cai et al., 2023) was initially used to extract features from the videos in the SEWA dataset, because of its ability to recognize complex facial expressions. But, for the regression task, we wanted to utilize all the data present in the dataset to build an accurate model. Since, we had access to frame-wise, arousal, and valence values we decided to extract features frame-wise from the available videos. However, MARLIN doesn't provide us with a way to extract features according to the frame rate of the video. Hence, we decided to extract all the frames from the video individually and then apply a state-of-the-art model to these frames to extract features and train them on the available valence and arousal label values. For the said purpose, we chose the VGGFace Model, which is based on VGG16(Simonyan and Zisserman, 2014) and is optimized for facial data.
For audio modality, Facebook's HuBert large model (Hsu et al., 2021), pre-trained on 960 hours of Librispeech data (Panayotov et al., 2015), was utilized for feature extraction. Librispeech is a corpus of read English speech data. Here both the HuBert model and the audio data in the SEWA dataset are sampled at 16kHz. The raw audio files were first cropped to their median length of 15 seconds. For each frame of audio, the pre-trained HuBert model extracted 1024 features. These features capture various aspects of the audio signal, such as frequency components, temporal patterns, and spectral characteristics. To reduce the dimensionality of the extracted features and retain the most relevant information, Principal Component Analysis was then applied. The resulting 128 features were used for further processing and analysis.

4.2 Experiments

4.2.1 Classification

Three binary classification tasks aimed at predicting high or low arousal, high or low valence, and liking or disliking responses were conducted for 40 epochs. The BERT (Devlin et al., 2019) classifier underwent training on text features, while recurrent neural network (RNN) was individually trained for video and audio features. For video data, two separate models were trained: one using Marlin features and the other with VGG-pooled features. Further late fusion was employed to combine results from audio, video, and text modalities using a majority vote approach.

4.2.2 Uni-Modal Regression

Video: We focused on training various models on frame-wise extracted features from VGGFace. Specifically, four models were utilized: RNN, LSTM, LSTM-Attention, and Bi-GRU. The features from the max-pooling layer of VGG16 was used as input. The shape of each vector was 7x7x512. Max pooling was applied to convert these

vectors to flatten them to obtain 512 features. Prior to model application, we performed pre-processing tasks on all videos in the SEWA dataset to find that the median video length was of 15 seconds, corresponding to 750 frames given a frame rate of 30 frames per second. If the video had more than 750 frames, we only used the first 750 frames for our analysis, while those with fewer frames were padded with zero vectors to standardize length. After all the pre-processing, for each frame in the video we had 512 features as input and for each video a feature vector of shape (750, 512) was obtained. Following data preprocessing, after vectors of equal length were obtained, model training and validation was performed.

Audio: HuBert features extracted from audio files were used to predict frame-wise valence and arousal values. SEWA dataset had manual annotations for every frame which were used as labels. Basline models like recurrent neural network, bi-directional LSTM and bi-directional GRU models were used for regression and performances were compared. Models were trained on the two largest cultures, Chinese and British, and evaluated on other four cultures.

4.2.3 Late Fusion with Audio and Video

Two fusing techniques, average and weighted fusion, were used to integrate predictions from separate audio and video modalities using the Bi-GRU model. Bi-GRU model was the best performing baseline model. **Average fusion** involves calculating the mean of the predictions from the two modalities, thereby treating audio and video data as equally influential. This method is straightforward and assumes that both modalities contribute equally to the overall prediction accuracy. **Weighted fusion**, on the other hand, assigns different weights to the predictions from each modality. In our experiment we assign 60 percent weight to video modality and rest 40 percent to audio modality. This is because visual cues provide richer information for accurate emotion state prediction, as seen in previous experiments. Weighted fusion aims to optimize the model's performance by leveraging the strengths of each data source according to its predictive power.

4.2.4 Channel attention Time-Series Transformer with early fusion

Channel attention Time-series Transformer(CATS-Transformer) consists of 3 modules: Channel attention module, Sequence to Sequence transformer and Multi-layer perceptron.

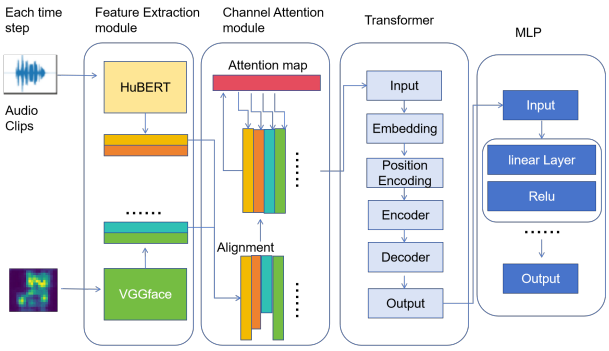

Figure 1: Pipeline of CATS-Transformer

a) Channel Attention module

In order to make model focus on the significant features. We applied channel attention(Woo et al., 2018) to fused features before feeding it to the transformer.

We first aligned the audio features(dimension D_1 = 128) and image features extracted from videos(dimension D_2 = 512). After alignment, extracted features were truncated and padded to an aligned length L (L=749). As a result, we obtained an audio feature representation (denoted as $AU_1...AU_N$ where each $AU_i \in R^{L\times D_1}$) and image feature representation (denoted as $VI_1...VI_N$ where each $VI_i \in R^{L\times D_2}$). We then concatenated AU_i and VI_i to obtain a fusion feature FU_i ($FU_i \in R^{L\times D_3}$, where D_3 = 640)

In channel attention module, the average and maximum features of each channel across the entire time sequence ($avp = \frac{1}{T}\sum_{t=1}^{T} FU_t$, $mxp = \max_{t=1}^{T} FU_t$) were computed. Followed by transforming these features using fully connected layers($x_{avg} = FC(avp)$, $x_{max} = FC(mxp)$) and combining the transformed features($x_{comb} = x_{avg} + x_{max}$) then normalizing the combined tensor using a sigmoid activation function to obtain attention ($W = \sigma(x_{comb})$). at the last step different channels with attention were attached:

$$X_i = FU_i \cdot W_i$$

b) Time-Series Transformer

In Time-Series Transformer (Wen et al., 2022), a linear layer as embedding layer was used to map inputs to features space ($X_d = L(X)$, $X_d \in R^{L\times D_s}$,D_s = 64), then X_d is applied to the position encoding($X_{pe} = PE(X_d)$).

Past-Focused Encoder. A subsequent triangular mask was employed to ensure that the model only

uses data from prior to the present location during self-attention computation in the Transformer encoder, preventing the usage of future information.

$$X_e = encoder(X_{pe})$$

Decoder and Teacher Forcing. The decoder accepts X_e as input and output embedding of last position Y_{p-1}. Similar to Encoder a subsequent triangular mask was applied to prevent cheating. During training to avoid error accumulation the teacher-forcing technique was applied (Figure 2). To be more precise, this means that during training, the model is fed the ground truth outputs from the previous time step as inputs, instead of its own predictions.

$$X_d = decoder(X_e)$$

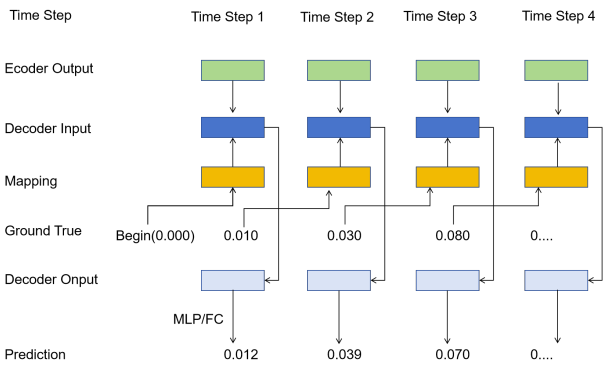

Figure 2: Training process of CATS-Transformer

MLP. After decoder, a Multi-layer perceptron was employed to get the final result.

$$X_o = mlp(X_d)$$

5 Results

5.1 Classification

Interestingly, when utilizing both Marlin and VGG features in video classification separately, no significant performance difference was observed. However, in single-modality classification, text consistently outperformed both video and audio, Table 1 demonstrates the performances after combing audio, video and text classification model's results by using majority vote in the late fusion technique. Figure 3 illustrates the confusion matrix resulting from three individual models: text, audio, and video which was tested on the Hungarian cultural dataset.

Culture	Valence High/Low	Arousal High/Low	Like/Dislike
C2	0.48	0.50	0.53
C3	0.52	0.46	0.52
C5	0.47	0.92	0.46
C6	0.43	0.36	0.62

Table 1: Classification accuracies

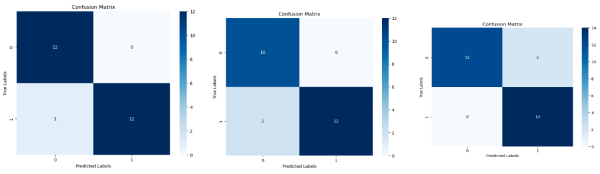

Figure 3: Confusion matrix for Hungarian culture

5.2 Uni-Modal Regression

5.2.1 Video

Table 2 shows RMSE and CCC values for baseline models like RNN, LSTM, LSTM-Attention and Bi-GRU. As we can see that CATS-Transformer dominates in comparison to other models with exceptional performance.

Model	Culture	Valence	Arousal
RNN	C2	0.4397 / 0.1053	0.4852 / 0.0423
	C3	0.3253 / 0.1027	0.2923 / 0.0387
	C5	0.4063 / 0.0600	0.2661 / 0.027
	C6	0.3295 / 0.1237	0.3928 / 0.0846
LSTM	C2	0.4444 / 0.1571	0.4686 / 0.0387
	C3	0.3234 / 0.1282	0.2953/ 0.0628
	C5	0.3884 / 0.0589	0.2622 / 0.0177
	C6	0.3222 / 0.1410	0.4152 / 0.0906
LSTM-Attention	C2	**0.4355** / 0.1187	0.4637 / 0.0065
	C3	0.3084 / 0.1026	0.2662 / 0.0502
	C5	0.3676/ 0.0332	0.2632 / 0.0013
	C6	0.3213 / 0.1792	0.4046 / 0.0794
Bi-GRU	C2	0.4480 / 0.1236	0.4867 / 0.0142
	C3	0.3511 / 0.1408	0.3229 / 0.0509
	C5	0.3793 / 0.0878	0.2896 / 0.0151
	C6	0.3098 / 0.1408	0.3975 / 0.0621
CATS-Transformer	C2	0.4471 / 0.0000	**0.1526** / 0.0009
	C3	**0.0767** / 0.0000	**0.0969** / 0.0001
	C5	**0.0733** / 0.0001	**0.1018** / 0.0001
	C6	**0.1019** / 0.0000	**0.1190** / 0.0000

Table 2: Arousal/Valence Prediction Results(RMSE/CCC) for Video

5.2.2 Audio

Table 3 shows the RMSE and CCC values for baseline models like RNN, Bi-LSTM and Bi-GRU. The performance of Bi-LSTM and Bi-GRU is much better than RNN model. Bi-LSTM and Bi-GRU have comparable performance with the Bi-GRU occasionally outperforming the Bi-LSTM. Also, if we compare baselines model performances on audio and video modality, models trained on video features have lesser RMSE and CCC values.

Model	Culture	Valence	Arousal
RNN	C2	0.500 / 0.1500	0.500 / 0.1000
	C3	0.400 / 0.1200	0.400 / 0.0800
	C5	0.520 / 0.1800	0.520 / 0.1400
	C6	0.450 / 0.1300	0.450 / 0.1100
Bi-LSTM	C2	0.410 / 0.1200	0.410 / 0.0500
	C3	0.380 / 0.1000	0.380 / 0.0600
	C5	0.460 / 0.1600	0.460 / 0.1200
	C6	0.390 / 0.1100	0.390 / 0.0900
Bi-GRU	C2	**0.400** / 0.1100	0.400 / 0.0400
	C3	0.370 / 0.0900	0.370 / 0.0500
	C5	0.450 / 0.1500	0.450 / 0.1100
	C6	0.380 / 0.1000	0.380 / 0.0800
CATS-Transformer	C2	0.5985 / 0.0015	**0.1983** / 0.0005
	C3	**0.1376** / 0.0014	**0.0890** / 0.0012
	C5	**0.0867** / 0.0021	**0.1035** / 0.0003
	C6	**0.1648** / 0.0018	**0.1369** /0.0009

Table 3: Arousal/Valence Prediction Results (RMSE/CCC) for Audio

5.3 Late Fusion with Audio and Video

Table 4 gives the RMSE and CCC values for the Bi-GRU model using both average and weighted fusion techniques. Weighted fusion has lower RMSE and CCC values indicating that giving higher weight to video modality can lead to better prediction of emotional states.

Model	Culture	Valence	Arousal
Average Fusion			
Bi-GRU	C2	0.420 / 0.115	0.420 / 0.115
	C3	0.360 / 0.115	0.360 / 0.115
	C5	0.415 / 0.120	0.415 / 0.120
	C6	0.345 / 0.120	0.345 / 0.120
Weighted Fusion			
Bi-GRU	C2	0.415 / 0.110	0.413 / 0.116
	C3	0.350 / 0.113	0.355 / 0.113
	C5	0.414 / 0.118	0.414 / 0.118
	C6	0.344 / 0.118	0.344 / 0.118

Table 4: RMSE/CCC values for Bi-GRU model after Average and Weighted Fusion

5.4 Channel attention Time-Series Transformer with early fusion

5.4.1 Uni-modal vs. Audio-video Model

In the comparison of uni-modal and Audio-video modal, as illustrated in Table 5, we observe that the video uni-modal model using CATS-Transformer consistently achieved the best performance when compared to audio

Culture	Audio	Video	Audio-video
C2	0.1983 / 0.0005	**0.1526** / 0.0009	0.1778 / 0.0000
C3	0.0890 / 0.0012	**0.0825** / 0.0001	0.1298 / 0.0000
C5	0.1035 / 0.0003	0.1018 / 0.0001	**0.0913** / 0.0001
C6	0.1369 / 0.0009	**0.1190** / 0.0000	0.1518 / 0.0000

Table 5: Performance of single/multi CATS-Transformer

5.4.2 Ablation Study

In this section, we conduct a series of experiments to evaluate the contributions of channel attention in the model and its impact on model performance. We trained the models for 50 iterations and optimized using Adam with 0.001 learning rate. To ensure fair comparison, we used the same hyperparameter settings and same random seed in all experiments. As shown in Table 6, after incorporating the channel attention mechanism, the model's performance (RMSE) typically experiences a slight improvement. This is attributed to the enhanced ability of the model to capture informative features from different channels.

Model Type	Culture	CATS-Transformer	Channel Attention
Valence			
Audio-Video	C2	**0.4105** / 0.0000	0.5552 / 0.0000
	C3	**0.0854** / 0.0000	0.0881 / 0.0000
	C5	**0.0707** / 0.0001	0.0836 / 0.0001
	C6	0.1600 / 0.0000	**0.1341** / 0.0000
Arousal			
Audio-Video	C2	**0.1778** / 0.0000	0.2267 / 0.0000
	C3	**0.1298** / 0.0000	0.1779 / 0.0000
	C5	0.1064/ 0.0001	**0.0913** / 0.0001
	C6	**0.1518** / 0.0000	0.2184 / 0.0000

Table 6: Arousal/Valence Prediction Results(RMSE/CCC) for CATS-Transformer

5.5 Channel Attention Analysis

In this section, we analyse the importance of channels in valence and arousal regression task across six different cultures.

Critical channels in Arousal & Valence regression. As shown in the figure 4, while the distribution of critical channels may vary, there were shared channels critical in both tasks. Specifically, channels 85, 29, 103, 82, and 57 were crucial in both arousal and valence regression prediction.

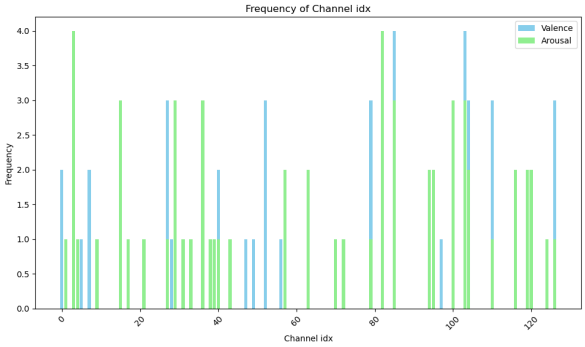

Figure 4: Frequency of 10 most important channels in arousal & valence regression across six cultures

Similarity of cultures. Based on the similarity of weights assigned to audio modality channels

across six languages, we can ascertain the similarity among these languages at this level.(As shown in Figure 5).

In conclusion, in arousal Regression task, the models between Hungarian and German(C2 and C3) rely on similar features; in valence Regression task, Serbian and Greek(C5 and C6) rely on similar features.

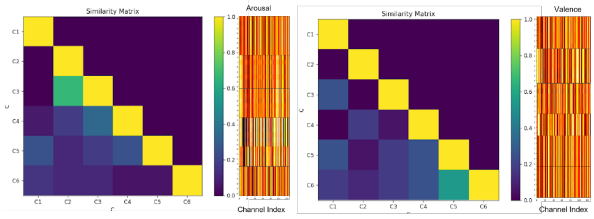

Figure 5: Similarity of weights assigned to audio modality channels across six languages

6 Conclusion and Lessons Learned

After conducting classification tasks, we found that arousal and valence prediction did not meet desired accuracy levels. The two video features (MARLIN and VGG) provided very similar results in classification models which gave us the confidence to use VGG for our regression tasks. There were quite a few challenges with aligning text with audio and video for CATS Transformer and we couldn't find a place for text in our regression model. The video modality seems to have a high impace when compared to other modalities, which was seen in unimodal results as well as CATS-Transformer results. However, our CATS-Transformer model surpassed baseline models (LSTM, RNN, Bi-GRU) due to the integration of self-attention and cross-attention mechanisms, effectively capturing long-term dependencies across modalities. The channel attention mechanism improved feature learning by prioritizing relevant information. Multimodal transformer model consistently outperformed unimodal ones by capturing complementary information from each modality. Interestingly, Furthermore, attention weight analysis revealed common features among certain languages, such as Hungarian and German, and Serbian and Greek. We can also observe that we obtained very low CCC scores in all our models. This indicates that the inter-culutural accuracies we obtained must be subject to more scrutiny. If we can get hold of bigger dataset for these cultures we might be able to conclude better on inter-cultural learning. We also need to work on creating models that perfrom well in minimizing regression error and maximizing correlation across cultures

7 Contributions

Avantika Singh Audio feature extraction using HuBERT and pre-processing including PCA, Audio uni-modal regression on baseline models (RNN, Bi-LSTM, Bi-GRU). Average and Weighted fusion of audio and video using Bi-GRU models.

Bhargava Shrirama: Video Feature Extraction using MARLIN, Frame Wise Image Extraction from Videos, Feature extraction using VGGFace (for each frame's image), Unimodal analysis of video data using RNN, LSTM, LSTM Attention Mechanism, Bi-GRU models, Literature Review.

Kavya Gengedulla Yeranarasappa Audio to text conversion using Whisper, language conversion to English for text tokenization for text using Bert, Classification tasks using Bert for text and RNN for audio and video, majority vote Late fusion, conducted regression experiments, assessing performance with RMSE and CCC, Literature review.

Xiuyuan Guo: CATS-Transformer model's design and implementation. Experiment of uni-model transformer and early fusion transformer model. Analysis and visualization of channel attention results. Write method, experiment and result part of CATS-Transformer model, Literature Review.

References

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. 2023. MARLIN: Masked Autoencoder for facial video Representation LearnINg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493–1504.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ambady N Elfenbein H. A. 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 12:203–235.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244.

Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjorn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 2021. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10).

Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M Montero, and Fernando Fernández-Martínez. 2021. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12(1):327.

Leena Mathur, Ralph Adolphs, and Maja J Matarić. 2022. Towards intercultural affect recognition: Audio-visual affect recognition in the wild across six cultures.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Mani Kumar T, Enrique Sanchez, Georgios Tzimiropoulos, Timo Giesbrecht, and Michel Valstar. 2021. Stochastic Process Regression for Cross-Cultural Speech Emotion Recognition. In *Proc. Interspeech 2021*, pages 3390–3394.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. pages 606–615.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.