MULTI-LINGUAL MULTI-TASK SPEECH EMOTION RECOGNITION USING WAV2VEC 2.0

Mayank Sharma

Amazon Prime Video International Expansion
mysharm@amazon.com

ABSTRACT

Speech Emotion Recognition (SER) has several use cases for Digital Entertainment Content (DEC) in Over-the-top (OTT) services, emotive Text-to-Speech (TTS) engines and voice assistants. In this work, we present a Multi-Lingual (MLi) and Multi-Task Learning (MTL) audio only SER system based on the multi-lingual pre-trained wav2vec 2.0 model. The model is fine-tuned on 25 open source datasets in 13 locales across 7 emotion categories. We show that, a) Our wav2vec 2.0 single task based model outperforms Pre-trained Audio Neural Network (PANN) based single task pre-trained model by 7.2% (relative), b) The best MTL model outperforms the PANN based and wav2vec 2.0 based single task models by 8.6% and 1.7% (relative) respectively, c) The MTL based system outperforms pre-trained single task wav2vec 2.0 model in 9 out of 13 locales in terms of weighted F1 scores, and d) The MTL-MLi wav2vec 2.0 outperforms the state-of-the-art for the languages contained in the pre-training corpora.

***Index Terms*—** Multi-task Multi-lingual speech emotion recognition, Pre-trained wav2vec 2.0, PANN

1. INTRODUCTION

Speech Emotion Recognition (SER) has been an active area of research for decades. SER has several applications for OTT content providers such as Prime Video, Netflix, Hulu, etc., and other related domains such as emotive Text-to-Speech (TTS) and voice assistants. It has potentially wide applications in several multi-lingual human computer interaction systems, such as, the interface with robots, banking, call centers, car board systems, computer games etc. Moreover, providing emotional states for virtual voice assistants is becoming increasingly popular. However, manual evaluation of speech emotions is costly, time consuming and requires at-least bilingual expertise for several use-cases. Hence, we develop a multi-lingual model that can identify 7 major emotions across 13 locales. Here, we train and evaluate various pre-trained models using audio features only, as aforementioned applications require emotion classification using the audio modality.

Several open source datasets have been created for training and validating the SER models. With the exception of a few, majority of them are relatively small in size ($\approx$ 3K samples), usually contain few speakers, limited emotional categories and are laboratory simulated in nature. Hence, in this paper, we combine small scale SER datasets to create a relatively larger multi-lingual corpus to fine-tune and evaluate the models. This aggregated speech emotion corpus consists of 25 open source datasets in 13 locales across 7 emotion categories. The corpus contains both conversational (natural) and laboratory simulated (acted) human speech emotions.

2. RELATION TO PRIOR WORK

Recently, Multi-Task Learning (MTL) and Multi-Lingual (MLi) frameworks for SER have become popular [1, 2]. This paper amalgamates the MTL and MLi frameworks to generate a single SER model. The advantages of a single MTL-MLi model are: a) Better generalization capabilities and b) Reduction in deployment and maintenance expenses, rather than deploying multiple SER models across languages.

Despite aggregation, our corpus has 95K samples, which is small considering the model parameters of current SER systems. A common strategy to deal with such small datasets is to apply transfer learning techniques. Typically, a model is initially pre-trained on a different task for which large datasets are available. Such a model can further be used as a feature extractor or fine-tuned on small datasets. In line with transfer learning, MTL and MLi frameworks, we use the wav2vec 2.0 model [3] pre-trained on the multilingual XLSR-53 corpora with an audio transcription task as the base network for SER model. While wav2vec 2.0 has been used for emotion classification task [4] without MTL, and using only English corpora, our model is fine-tuned on the aggregated multilingual speech emotion corpus, allowing the model to generalize across languages. Moreover, the training of our model encompasses an MTL framework, with several auxiliary tasks, such as gender and language classification along with mean and standard deviation of F0 and energy and voice ratio prediction tasks, resulting in better emotion classification F1 score. A variation of similar auxiliary tasks has demonstrated better MOS scores for FastSpeech2 [5] than traditional neural TTS systems. We also compare pre-trained PANN for emotion classification [6].

In this paper, we empirically demonstrate the following: a) Our wav2vec 2.0 single task based model outperforms PANN

Table 1: Datasets used in the paper.

Dataset Name	Locales	Average Duration	Samples	Type	Reference
ESD-en	en-US	2.76	17,500	acted	[7]
ESD-zh	zh-Hans	3.22	17,500	acted	[7]
AESDD	el-EL	4.10	604	acted	[8]
CREMA-D	en-US	2.54	7,442	acted	[9]
SAVEE	en-US	3.84	480	acted	[10]
TESS	en-US	2.06	2,800	acted	[11]
RAVDESS	en-US	4.09	2,452	acted	[12]
IEMOCAP	en-US	4.56	7,529	acted	[13]
EMODB	de-DE	2.78	535	acted	[14]
MELD	en-US	3.27	12,983	natural	[15]
URDU-Dataset	ur-UR	2.50	400	natural	[16]
Emovo	it-IT	3.12	588	acted	[17]
ShEMO	fa-FA	4.12	2,992	natural	[18]
EmoV-DB	en-US	4.67	3,644	acted	[19]
SUBESCO	bn-BN	4.03	7,000	acted	[20]
Pavoque	de-DE	5.76	4,867	acted	[21]
ADEPT	en-US	1.95	50	acted	[22]
eNterface05	en-US	2.58	1,263	acted	[23]
Café-48k	fr-FR	4.44	936	acted	[24]
jl-corpus	en-US	2.09	1,200	acted	[25]
OreauFR-02	fr-FR	2.78	434	acted	[26]
VERBO	pt-BR	2.40	1,174	acted	[27]
MESD	es-MX	0.72	436	acted	[28]
polish-word-emotion	pl-PL	0.84	448	acted	[29]
CASTRO	pt-PT	1.46	368	acted	[30]

[6] based single task pre-trained (on Google Audioset) model by 7.2% (relative), b) The best MTL model outperforms the PANN and wav2vec 2.0 based single task models by 8.6% and 1.7% (relative) respectively, c) The MTL based system outperforms pre-trained single task wav2vec 2.0 model in 9 out of 13 locales in terms of weighted F1 scores, and d) The MTL wav2vec 2.0 outperforms the state-of-the-art for the languages which are contained in the pre-training corpora.

3. DATASETS

Table 1 describes the datasets and presents the locales, average duration of clips and dataset type (acted or natural). To fine-tune and evaluate the performance of the model, we combine the datasets as follows. First, we divide the individual datasets into train/test/validation (75%/15%/10% respectively) sets using stratified sampling across gender and emotions. We keep the emotions ratio similar across the three sets if the individual datasets do not contain predefined partitions, otherwise we use the predefined partitions of the individual datasets and second, we combine the respective sets across the 25 datasets.

We observe that only 3 of 25 datasets are conversational, while others are acted. Natural conversational datasets generally exhibit mild emotions in contrast to acted datasets. Thus, it is difficult for human annotators to judge the right emotion as described in [15] with audio modality only. Therefore, we will empirically demonstrate that for a natural dataset MELD, SER systems results in smaller F1 values compared to others. The total number of samples for emotion categories are as follows: *neutral*: 24,225, *angry*: 18,598, *happy*: 17,524, *sad*: 14,590, *surprise*: 11,268, *disgust*: 4,769 and *fear*: 4,651. We observe that neutral class has the largest number of samples, and fear has the smallest with an imbalance ratio of 5x. The locale distribution of the datasets is as follows: *en-US*: 57,343, *zh-Hans*: 17,500, *bn-BN*: 7,000, *de-DE*: 5,402, *fa-FA*: 2,992, *fr-FR*: 1,370, *pt-BR*: 1,174, *el-EL*: 604, *it-IT*: 588, *pl-PL*:448, *es-MX*:436, *ur-UR*:400, and *pt-PT*:368. We observe that the locale distribution is highly skewed towards en-US, as majority of publicly available, free-to-use, open source datasets are in English. The dataset contains roughly an equal proportion of male to female ratio with 49,817 *male* and 45,808 *female* speech clips. In the next section, we describe the model and present various auxiliary tasks.

4. METHOD

We choose wav2vec 2.0 [3] as our base model. The model contains a convolutional feature encoder that maps the raw audio sampled at 16 kHz to latent speech representations, followed by a transformer network that converts them to output representation. The model is pre-trained on multilingual speech specific corpora named XLSR-53, consisting of 53 languages. The resulting output audio embedding thus contains information about language, gender and pitch. Hence, wav2vec 2.0 acts as a suitable choice of a pre-trained model and feature extractor for emotion classification fine tuning task, rather than pre-trained CNN based models trained on ImageNet or Google Audioset and fine-tuned on emotion classification task with spectrogram as their inputs [6, 31].

Wav2vec 2.0 has been used in emotion classification task [4] and results in state-of-the-art results for IEMOCAP and RAVDESS datasets with a recall of 0.67 and 0.84 respectively. Here, the authors replace the final text prediction layer with the emotion prediction CNN, LSTM and global time dimension average pool layers. Instead, we perform a mean pool across the time dimension on the wav2vec 2.0 embedding layer, and add multiple dense layer heads for primary emotion classification and various other auxiliary tasks. Further, we freeze the CNN feature extractor and fine-tune the remaining transformer and dense layers. We experiment with several auxiliary tasks, such as: a) *Gender Prediction*: As shown in [1], gender classification shares mutual features such as pitch and MFCC with emotion classification, b) *Language Prediction*: Due to the multi-lingual nature of the classifier, we add language prediction task to help the model learn the intonation of individual languages. c) *F0 mean and standard deviation regression task*: Pitch (F0) is highly correlated with emotions. For example, high pitch is attributed with excited and shouting states, whereas low pitch is generally related to sad and pensive states. Hence, we use pYin's algorithm to first predict F0 contour, and then task the model with predicting the mean and standard deviation of F0 for a given clip, d) *Energy mean and standard deviation regression task*: Emotions can be categorized in three dimensions of valence, activation and dominance [32]. High energy is usually associated with positively dominant emotions, such as anger and excitement. Low energy is associated with negatively dominant fear and sadness. Thus, we use mean and standard deviation of clip energy prediction as an auxiliary task, and e) *Voice ratio re-*

Table 2: Mean of weighted F1 score, weighted precision, weighted recall across datasets.

Name	F1	Recall	Precision
PANN + Em	0.786	0.788	0.804
w2v + Em	0.845	0.847	0.855
w2v + Em + Ge	0.846	0.848	0.857
w2v + Em + Ge + La + F0-me + F0-st + En-me + En-st	0.850	0.854	0.861
w2v + Em + Ge + La + F0-me + F0-st + En-me	0.853	0.855	0.865
w2v + Em + Ge + La	0.854	0.856	0.864
w2v + Em + Ge + La + F0-me + F0-st	0.855	0.857	0.864
w2v + Em + Ge + La + F0-me + F0-st + En-me + En-st + Vr	0.857	0.859	0.867
w2v + Em + Ge + La + F0-me	**0.860**	**0.862**	**0.870**

Table 3: Per dataset weighted average recall of emotions.

Dataset Name	PANN + Em	w2v + Em	w2v + Em + Ge + La + F0-me
MELD	0.446	**0.498**	**0.498**
IEMOCAP	0.545	**0.641**	**0.641**
ESD-en	0.966	**0.973**	0.967
ESD-zh	0.958	**0.979**	0.977
CREMA-D	0.726	0.786	**0.802**
SUBESCO	0.941	0.956	**0.958**
Pavoque	0.976	0.978	**0.978**
EmoV-DB	0.835	0.966	**0.977**
ShEMO	0.807	**0.862**	0.859
TESS	**1.000**	**1.000**	**1.000**
RAVDESS	0.849	**0.874**	**0.874**
eNterface05	0.727	**0.894**	0.886
jl-corpus	0.942	0.979	**0.988**
VERBO	0.449	0.498	**0.566**
Café-48k	0.745	0.851	**0.899**
AESDD	0.835	0.893	**0.934**
Emovo	0.480	**0.713**	0.670
EMODB	0.907	0.953	**0.963**
SAVEE	0.698	0.823	**0.844**
MESD	0.878	**0.889**	0.878
OreauFR-02	0.535	0.609	**0.701**
polish-word-emotion	0.815	0.827	**0.901**
URDU-Dataset	**0.900**	0.763	0.838
CASTRO	0.932	**0.973**	**0.973**
ADEPT	0.800	**1.000**	**1.000**

gression task: As the clips are of variable length, we task the network to predict the voice ratio following the auxiliary tasks in [5]. For the three classification tasks, we use a weighted cross entropy loss, with weight equal to the inverse probability of the classes to account for the imbalance in their respective distributions and use L2 loss for regression tasks. In the next section, we present a comparative analysis of emotion classification between wav2vec 2.0 and PANN.

5. RESULTS

In this section, we describe the training methodology and results. We compare emotion classification using PANN [6] pre-trained on Google Audioset and wav2vec 2.0 pre-trained on XLSR-53 corpora. We remove the final audio classification layer in PANN and replace it with a dense layer for emotion classification task. Further, we fine-tune the wav2vec 2.0 and PANN for 10 and 150 epochs (no loss decrease post these epochs) respectively on the training set of the aggregated corpus. We tuned the weight decay and learning rate parameters from the set $\{0, 1e-05, 1e-04\}$ and $\{1e-04, 1e-03\}$ respectively and present the results with best hyperparameter settings.We use a maximum duration of 6 seconds (s) for the audio clip and discard any remaining duration. We pad the clips with duration < 6 s with zeros. *In all the tables, w2v*

Table 4: Mean of weighted F1 score per locale of emotions.

Locales	PANN + Em	w2v + Em	w2v + Em + Ge + La + F0-me
bn-BN	0.942	0.956	**0.958**
de-DE	0.942	0.966	**0.970**
el-EL	0.838	0.899	**0.937**
en-US	0.770	0.857	**0.859**
es-MX	0.877	**0.888**	0.876
fa-FA	0.802	**0.861**	0.857
fr-FR	0.639	0.731	**0.800**
it-IT	0.489	**0.710**	0.667
pl-PL	0.816	0.826	**0.901**
pt-BR	0.451	0.447	**0.534**
pt-PT	0.933	**0.973**	**0.973**
ur-UR	**0.899**	0.770	0.840
zh-Hans	0.958	**0.979**	**0.979**

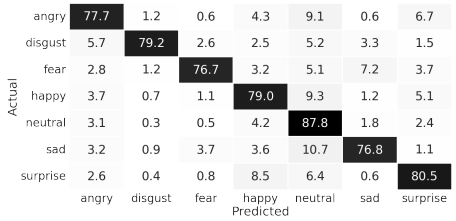

Fig. 1: Confusion matrix for emotion classification using best MTL-MLi model.

denotes wav2vec2.0.

First, we evaluate the performance of the wav2vec 2.0 model with PANN. We compute the weighted F1, precision and recall of emotion classes for each dataset and average across all the datasets as shown in the table 2. Here, *Em* denotes the Emotion classification, *Ge* denotes the Gender classification, *La* denotes Language classification, *F0-me* and *F0-std* denotes mean and standard deviation for F0 regression, *En-me* and *En-std* denotes mean and standard deviation energy regression, and *Vr* denotes the voice ratio regression auxiliary task. From the first two rows, we observe that single task wav2vec 2.0 outperforms the PANN by 7.2% (relative). This is because, (a) PANN is pre-trained on audio classification corpus consisting of speech and non-speech sounds. However, wav2vec 2.0 is trained on speech-only corpora of 53 languages. (b) PANN use spectrograms compared to wav2vec 2.0 which uses raw audio waveforms, leading to better generalization.

Second, we present comparison across models with various auxiliary classification and regression heads. We observe that the metrics for the MTL model are better than the pre-trained single task model. We empirically demonstrate that MTL with gender, language and mean F0 heads outperforms the pre-trained single task wav2vec 2.0 model by 1.7% (relative) in terms of weighted F1-scores. We observe a similar trend for weighted precision and recall scores. We observe that adding a voice ratio (Vr) head reduces the mean weighted F1 score slightly from the best performing model, as Vr does not provide any new information for the emotion classification task. From the table 3, we observe that best MTL wav2vec 2.0 model outperforms single head PANN and wav2vec 2.0 in 23 and 18 out of 25 datasets respectively in terms of weighted recall scores.

Third, we compare the weighted recall scores of several datasets from their recent state-of-the-art results. From the table 3, we observe that for IEMOCAP and RAVDESS datasets, our wav2vec 2.0 MTL model results in 0.640 and 0.874 weighted recall, which is similar and 3.5% better than recall for the wav2vec 2.0 model in [4] respectively, which was solely trained on English corpora. Moreover, even the PANN model fine-tuned on aggregated corpus results in 0.849 weighted recall for RAVDESS dataset, compared to 0.72 in [6]. For natural speech datasets, such as MELD, URDU-Dataset and ShEMO, the weighted recalls are 0.49, 0.84 and 0.86 respectively. For MELD, we observe a 20% relative improvement on the baseline results for weighted recall in the original paper [15]. The recall score is still smaller compared to other datasets, as MELD requires video, audio and text modalities for better emotion recognition due to mild emotional cues present in the dataset. For the URDU-Dataset, we observe a small (0.5% relative) improvement from the baseline [16] as URDU was not part of pre-training XLSR-53 corpora. For the ShEMO dataset, we observe a 32% relative increase in recall scores, compared to the baseline in [18]. This shows that the MTL wav2vec 2.0 outperforms the state-of-the-art for the locales contained in the pre-training corpora.

Fourth, we evaluate the performance across locales. To calculate the F1 across locales, we average F1 values for the datasets belonging to each locale. As shown in the table 4, we observe that MTL model outperforms the other variants in 9 out of 13 locales. Finally, we present the confusion matrix for the 7 emotion classes for the best MTL model in the figure 1. The confusion matrix is calculated using all the test data-points simultaneously, as opposed to averaged per-dataset metrics. We observe that, all the diagonal entries are $> 76\%$. The most confusing category for angry, happy and sad is neutral, which is due to mild emotional differences for certain utterances. Some (surprise, happy) and (disgust, angry) pairs are similar in pitch, thus have highest confusion among them. These can be resolved using valence, activation and dominance signals.

6. CONCLUSION

In this paper, we presented a MTL-MLi emotion recognition system based on wav2vec 2.0. We experimented with 25 datasets across 13 locales, comprising 7 major emotion categories. We showed that our MTL system outperforms single head PANN and wav2vec 2.0 models by 7.2% and 1.7% (relative) respectively. We empirically demonstrated that the MTL wav2vec 2.0 outperforms the state-of-the-art for the languages contained in the pre-training corpora. Our MTL model outperforms a single task wav2vec 2.0 in 9 out of 13 locales. Moreover, for natural datasets, MTL model outperforms the state-of-the-art baselines, demonstrating its effectiveness on mild emotions. However, we demonstrated that some datasets, require other modalities for better classification. Future works will investigate the use of combination of other modalities, such as video and text, addition of other auxiliary tasks, such as valence, arousal and dominance prediction to separate confusing classes along with classification into natural and acted voices.

7. REFERENCES

[1] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in *Interspeech*, 2019, pp. 2803–2807.

[2] Shi-wook Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5881–5885.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020, vol. 33.

[4] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[5] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2020.

[6] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[7] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[8] N Vryzas, A Liatsou, Rigas Kotsakis, Charalampos Dimoulas, and George Kalliris, "Augmenting drama: A speech emotion-controlled stage lighting framework," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, 2017, pp. 1–7.

[9] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[10] Sanaul Haq, Philip JB Jackson, and James Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, 2008.

[11] Kate Dupuis and M Kathleen Pichora-Fuller, "Toronto emotional speech set (tess)-younger talker_happy," 2010.

[12] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

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

[14] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al., "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[15] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[16] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2018, pp. 88–93.

[17] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco, "Emovo corpus: an italian emotional speech database," in *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association, 2014.

[18] Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami, "Shemo: a large-scale validated database for persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, 2019.

[19] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[20] Sadia Sultana, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, pp. e0250173, 2021.

[21] psibre, "Pavoque Corpus of Expressive Speech," https://github.com/marytts/pavoque-data/releases, 2016.

[22] Alexandra Torresquintero, Tian Huey Teh, Christopher GR Wallis, Marlene Staib, Devang S Ram Mohan, Vivian Hu, Lorenzo Foglianti, Jiameng Gao, and Simon King, "Adept: A dataset for evaluating prosody transfer," *arXiv preprint arXiv:2106.08321*, 2021.

[23] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[24] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre, "A canadian french emotional speech dataset," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 399–402.

[25] Jesin James, Li Tian, and Catherine Inez Watson, "An open source emotional speech corpus for human robot interaction applications.," in *INTERSPEECH*, 2018, pp. 2768–2772.

[26] Leila KERKENI, Catherine CLEDER, Youssef SERRESTOU, and Kosai RAOOF, "French emotional speech database - oréau," Dec. 2020.

[27] JRT Neto, GP Filho, LY Mano, and J Ueyama, "Verbo: voice emotion recognition database in portuguese language," *J Comput Sci*, vol. 14, 2018.

[28] D.I. Ibarra-Zarate M.M. Duville, L.M. Alonso-Valerdi, "Mexican emotional speech database (mesd): adult and child affective speech," *Data in Brief*, 2021.

[29] Marzena Mięsikowska and Dariusz Świsulski, "Emotions in polish speech recordings," 2020.

[30] São Luís Castro and César F Lima, "Recognizing emotions in spoken language: A validated set of portuguese sentences and pseudosentences for research on emotional prosody," *Behavior Research Methods*, vol. 42, no. 1, pp. 74–81, 2010.

[31] Margaret Lech, Melissa Stolar, Christopher Best, and Robert Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, pp. 14, 2020.

[32] Srinivas Parthasarathy and Carlos Busso, "Jointly predicting arousal, valence and dominance with multi-task learning.," in *Interspeech*, 2017, vol. 2017, pp. 1103–1107.