TACKLING CROSS-LINGUAL GENERALIZATION IN SPEECH EMOTION RECOGNITION VIA LINGUISTIC GROUPING: A SPOTLIGHT ON NEO-LATIN LANGUAGES

Francesco Ardan Dal Rì, Nicola Garau, and Nicola Conci

DISI - Dept. of Information Engineering and Computer Science, University of Trento

ABSTRACT

While human languages share common prosodic traits to convey emotions, cultural and linguistic peculiarities may hinder robust cross-lingual generalization in Speech Emotion Recognition (SER). In this paper, we investigate SER from the point of view of linguistic lineage, with a specific focus on Neo-Latin languages (French, Italian, Portuguese, Romanian, and Spanish), evaluating direct transferability from and against languages from other families (Arabic, Bengali, Chinese, English, and Russian). Results show how linguistic relatedness can affects cross-lingual emotion recognition and provide insights for designing more robust SER systems.

***Index Terms*—** Speech Emotion Recognition, Generalization, Linguistic Groups.

1. INTRODUCTION

Speech Emotion Recognition (SER) aims at retrieving human emotional states from speech signals, usually through deep models, and finds applications in numerous fields such as affective computing, robotics, and healthcare [1]. Despite recent advancements in research, with many datasets and methodologies being introduced [1], SER still poses numerous challenges, such as robust generalization across speakers, emotions, and languages, a highly sought-after feature. While the majority of the literature focuses on monolingual assessment [1], there is a growing interest towards multi/cross-lingual scenarios, and many recent works have attempted to address this task. For instance, Sharma performed multilingual and multi-task training over 13 languages [2]; Lee investigated generalization possibilities across English and Japanese [3]; Neumann and Vu provided multilingual and cross-lingual experiments on English and French [4]; or Latif and colleagues compared transferability between Urdu and a selection of Western languages [5]. These works build on the assumption that human languages share common prosodic, rhythmic, and tonal inflections to convey emotions [6].

While evidence for universality in certain acoustic features of emotions exists, many studies also highlight differences shaped by cultural factors or specific linguistic structures, that may hamper generalization [7, 8]. Such differences are well recognized in computational linguistics, where phenomena such as semantic variation, colexification, or language-specific categorization are known to influence direct cross-linguistic transfer [9]. On the contrary, these specificities are often overlooked in SER literature. While several studies implicitly assumes a certain degree of compatibility between related languages (e.g., in [5], German and English indeed share a common origin within the Germanic branch), examining generalization constraints across linguistic topologies constitutes an overall underexplored research problem, with very few works explicitly taking such relations into account; recent examples being the work by Monisha and Sultana [10], comparing several Indo-Aryan languages against German ones, or by Lei [11], using specific encoders for different linguistic-geographical groups in a multilingual context.

To address this gap, we are interested in investigating the cross-linguistic generalization across individual languages, in specific relation to their linguistic lineage. For this purpose, we adopt a Wav2Vec [12] pipeline tailored to speech emotion classification. Wav2Vec has become a widely used backbone in recent SER research due to its ability to capture acoustic and prosodic features relevant to emotion recognition [2, 13, 14], making it a suitable framework for our cross-linguistic analysis. Building on this established setup, we propose a preliminary study focusing on Neo-Latin (Romance) languages - i.e. French, Italian, Portuguese, Romanian, and Spanish - and compare their generalization performance against an arbitrary selection of languages from other linguistic families, namely Germanic (English), Semitic (Moroccan Arabic), Slavic (Russian), Indo-Aryan (Bengali), and Sinitic (Mandarin Chinese).

2. METHOD

Following similar methods in the literature [2, 13, 15], we exploited a pretrained, 24-layers, XLSR-backbone Wav2Vec pipeline, tailored to SER tasks, with a randomly-initialized classifier head. Raw 16 kHz waveforms are processed by the Wav2Vec2 feature extractor and encoder to produce contextualized frame-level hidden states; utterance-level embeddings $\mathbf{z} \in R^{1024}$ are obtained by mean-pooling over frames. A linear projector reduces $\mathbf{z}$ to a 256-dimensional vector $\mathbf{h}$, which is then passed to a final classification head that outputs logits over emotion classes.

Training optimizes a joint objective combining a weighted Cross-Entropy term and a supervised contrastive term [16]:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{contr}, \quad (1)$$

where $\mathcal{L}_{contr}$ encourages label-consistent structure in the embedding space (pulling together same-label samples and pushing apart different labels). We set $\lambda = 0.1$ and use a contrastive temperature $\tau = 0.07$.

To mitigate class imbalance and cross-corpus dominance (see 4.4), we introduce a two-level balancing strategy acting at both class and dataset levels. First, the cross-entropy term is weighted inversely to class frequency, ensuring that minority emotions contribute proportionally to the optimization objective. Second, we compensate for discrepancies in dataset sizes by assigning each training sample a (normalized) weight, inversely proportional to the size of its respective corpus. Finally, mini-batches are constructed to balance contributions from different corpora during training, thus preventing large datasets from dominating. In addition, we also included preprocessing and augmentations: loudness normalization, per-utterance mean–variance waveform normalization, random gain scaling, and light additive noise.

An extensive analysis of the impact of each of these additional components is beyond the scope of this work; additionally, similar techniques are already documented in recent SER literature - e.g., [17, 18]. Therefore, for conciseness, results reported in Section 4 refer to the best model configurations, which include the aforementioned data balancing and augmentation strategies. Indeed, these combined strategies overall encourage the models to learn representations that are less dominated by corpora with substantially more training samples, while maintaining stable and comparable optimization dynamics across experimental settings.

Importantly, we deliberately chose such an overall simple pipeline over more recent and better-performing methods (e.g., [11, 19, 20]) to further highlight the main contribution of our work, which is not to provide a novel architecture or improve performances, but to rely on well-established baselines to verify our hypothesis and facilitate comparison against baselines.

3. EXPERIMENTAL SETUP

3.1. Datasets

We collect a number of datasets widely applied in the literature [15]. For languages other than Neo-Latin, we arbitrarily choose a single one per linguistic group, promoting diversity in accordance with the currently available datasets - Table 1. While most literature follows the discrete 7-classes model proposed by Ekman [21], not all datasets contain all of them: indeed, we choose to drop the *surprise* class as it was heavily underrepresented. As such, we train over six emotional classes: *anger* (ang), *disgust* (dis), *fear* (fea), *happiness* (hap), *neutral* (neu), and *sadness* (sad). In total, 10991 samples were collected for the Neo-Latin group, and 13946 samples for the other one. All files have been preprocessed by silence-trimming and resampled at 16kHz in order to be fed into Wav2Vec.

Table 1. Summary of considered datasets grouped by language and linguistic group.

Group	Language (Code)	Datasets
Neo-Latin	French (FRE)	Oreau [22]; CaFE [23]
	Italian (ITA)	Emozionalmente [24];
	Portuguese (POR)	emoUERJ [25]; EmoProsody [26]
	Romanian (ROM)	RED [27]
	Spanish (SPA)	MESD [28]; EmoFilm [29]
Semitic	Moroccan Arabic (ARA)	MDERD [30]
Indo-Aryan	Bengali (BEN)	SUBESCO [31]
Sino-Tibetan	Mandarin Chinese (CHI)	MCAESD [32]
Germanic	English (ENG)	CREMA-D [33]
Slavic	Russian (RUS)	RESD [34]

3.2. Training Procedure and Hyperparameters

Models are trained with AdamW (base learning rate 1e-4, weight decay 1e-5). We use mixed precision training with a GradScaler, gradient accumulation of 4 steps and a micro-batch size of 16 (effective batch size 64). Gradients are clipped with ℓ_2 norm 1.0 before the optimizer steps. Consistently with other works, e.g., [13], models were fine-tuned for 10 epochs, on a single NVIDIA GeForce RTX 4090 GPU.

We run three different sets of experiments to explore different settings and provide an evaluation baseline for our preliminary study, specifically:

- **Single-language:** train and evaluate separate models per language, using stratified, per-emotion, 70%/30% train/test splits; this serves as a baseline for within-corpus performance.
- **Multilingual:** train on combined corpora (i.e., individual linguistic groups and full data), and evaluate both within-group and on the whole corpus; splits are stratified per-emotion and per-language to preserve class/language proportions.
- **Cross-group:** leave one base language out per group during training, and evaluate on that held-out language (within-group) and on all languages from the other cluster (cross-group) to measure generalization on unseen languages.

Table 2. UAR / F1 single-language and multilingual models performance, with benchmark on individual languages (% ↑).

Train Set	FRE	ITA	POR	ROM	SPA	ARA	BEN	CHI	ENG	RUS
Individual	66.4 / 65.6	**81.0 / 81.1**	75.1 / 72.7	**98.3 / 98.3**	62.9 / 59.6	42.6 / 36.5	70.0 / 67.2	**80.7 / 80.9**	**78.8 / 78.7**	43.7 / 39.9
Neo-Latin	**78.4 / 78.1**	75.7 / 75.7	91.7 / 91.6	94.0 / 93.9	74.3 / 74.1	//	//	//	//	//
Others	//	//	//	//	//	91.0 / 89.3	**83.8 / 83.8**	74.8 / 74.7	70.7 / 70.1	**63.2 / 63.2**
All	77.4 / 76.2	73.8 / 73.7	**96.4 / 96.7**	96.3 / 96.4	**80.3 / 81.3**	**92.7 / 93.0**	81.1 / 81.2	72.8 / 73.3	69.1 68.4	57.3 / 55.1
Baseline (Wav2Vec)[1] [15]	41.4 / 40.2	56.7 / 56.6	//	//	62.9 / 62.8	//	51.2 / 50.9	//	61.9 / 61.7	52.8 / 52.9

Table 3. UAR / F1 cross-lingual generalization across unseen languages (% ↑).

Train Set	FRE	ITA	POR	ROM	SPA	ARA	BEN	CHI	ENG	RUS
ITA, POR, ROM, SPA	**58.4 / 57.4**	//	//	//	//	**37.4 / 37.3**	59.0 / 54.6	47.0 / 45.9	55.3 / **55.3**	31.3 / 28.4
FRE, POR, ROM, SPA	//	**62.3 / 61.3**	//	//	//	28.8 / 29.4	60.0 / 57.4	**51.1 / 51.4**	52.0 / 49.1	34.8 / 32.4
FRE, ITA, ROM, SPA	//	//	**62.8 / 67.6**	//	//	24.5 / 26.6	**63.8 / 62.1**	49.8 / 48.8	55.4 / 54.7	33.3 / 31.7
FRE, ITA, POR, SPA	//	//	//	**68.2 / 70.4**	//	27.6 / 27.9	62.5 / 60.8	49.3 / 48.7	51.4 / 51.1	34.0 / 29.7
FRE, ITA, POR, ROM	//	//	//	//	**43.0 / 41.8**	31.0 / 27.2	63.0 / 60.4	43.7 / 42.8	51.4 / 51.0	31.4 / 26.7
BEN, CHI, ENG, RUS	55.9 / 55.7	55.9 / 56.3	52.0 / 56.8	48.8 / 55.4	41.3 / 40.0	21.8 / 27.2	//	//	//	//
ARA, CHI, ENG, RUS	54.2 / 53.5	58.6 / 59.0	53.1 / 58.9	57.1 / 63.6	41.6 / 38.5	//	60.6 / 57.9	//	//	//
ARA, BEN, ENG, RUS	56.8 / 54.9	56.4 / 56.1	55.5 / 60.4	57.9 / 59.7	37.6 / 34.2	//	//	46.8 / 45.8	//	//
ARA, BEN, CHI, RUS	56.4 / 56.1	52.0 / 52.6	56.0 / 62.1	45.1 / 52.5	41.0 / 40.2	//	//	//	**55.6** / 53.8	//
ARA, BEN, CHI, ENG	58.2 / 57.1	**62.3** / 61.1	53.3 / 61.5	58.0 / 62.4	41.4 / 38.9	//	//	//	//	**35.7 / 33.5**

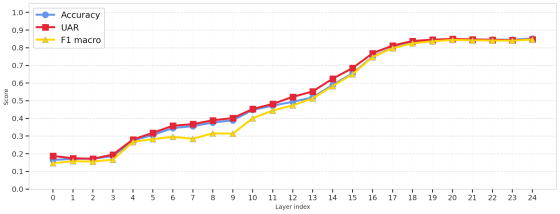

Fig. 1. Per-layer metrics analysis relative to the Multilingual Neo-Latin model.

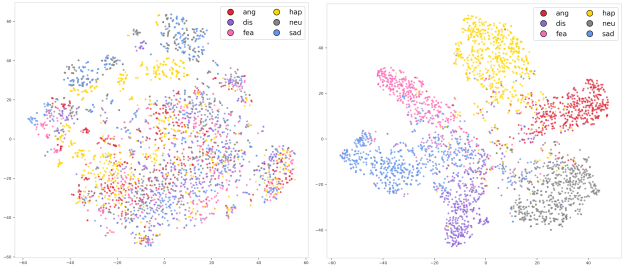

Fig. 2. Per-emotion t-SNE relative to the Multilingual Neo-Latin model, layer 4 (left) and layer 18 (right).

4. RESULTS

4.1. Model Analysis

As in related literature, e.g., [35], we perform a per-layer analysis in order to assess whether the model encodes prosodic and paralinguistic cues relevant for SER. Indeed, Fig. 1 shows that intermediate layers (approximately layers 13–19) yield the best metrics. This suggests that emotion-discriminative information is most accessible at intermediate-to-upper transformer blocks, where representations remain sensitive to prosodic variability. As shown in Fig. 2, indeed, early-layer representations exhibit substantial overlap between classes, suggesting that low-level acoustic features alone are insufficient for the task, while intermediate-to-late layers form more compact and well-separated clusters, indicating that these layers can encode higher-level paralinguistic information.

4.2. Single and Multilingual Experiments

Table 2 summarizes UAR / F1 metrics for the first two sets of experiments, single-language and multilingual. Overall, multilingual training improves performance over individual models, especially when languages are grouped by linguistic lineage. For the Neo-Latin group, training on the joint corpus consistently boosts performance for most languages. ROM and ITA represent partial exceptions, as both already achieve high results in the individual setting. In fact, ITA slightly drops in multilingual configurations, while ROM remains roughly stable. This ceiling effect likely reflects dataset characteristics and reduced benefit from additional cross-lingual information.

For the other group, multilingual training shows a similar trend. In particular, ARA yields a dramatic increase of roughly 50% when trained on all languages. In contrast, CHI and ENG perform best or near-best in the individual configuration, and slightly degrade in the fully multilingual setting, similar to ITA and ROM. This pattern further supports the hypothesis that dataset size and balance strongly influence the impact of multilingual training [17, 18]. Indeed, ITA, CHI, and ENG are the larger corpora in the selection (5916, 4006, and 7442 samples respectively), while ROM, with 2100 samples, contains only 3 classes; therefore, it is arguable that more data and a simpler task naturally increase performance.

[1] In case of multiple datasets, benchmarks accuracies are averaged.

The models proved robust in correctly classifying the emotion classes - Fig. 3. On average, we observe slightly better results for the Neo-Latin group model, reaching a peak of 84% on the *happiness* class, with the other yielding the worst result with a negative peak of 61% on *fear*. In both cases, *sadness* sometimes gets misclassified as *fear* and vice-versa. This is further highlighted by performing dimensionality reduction (t-SNE) on the projector's space, where a certain degree of overlapping between the two classes is observable - Fig. 4.

Finally, overall, our models consistently outperform the baselines reported by Ma *et al.* [15] on the common datasets, confirming the robustness of the proposed approach across both single-language and multilingual scenarios.

4.3. Cross-Group Experiments

Table 3 presents direct cross-lingual generalization results, where models are trained on individual linguistic except for one language, and evaluated on all remaining unseen languages. As expected, performance drops substantially compared to the multilingual in-language results of Table 2, with degradations observed across all languages. The gap is particularly pronounced for ARA, whose best cross-lingual result (37.4 / 37.3) is far below its multilingual in-language score (92.7 / 93.0). RUS also exhibits a noticeable drop, with best cross-lingual performance at 35.7 / 33.5 against 57.3 / 55.1 in the full-multilingual setting. Notably, such drops are however in line with previous research, e.g., [4, 5].

Within the Neo-Latin group, however, it is possible to observe a coherent pattern: each held-out language tends to achieve its best cross-lingual performance when trained on the remaining languages from the same linguistic lineage. In contrast, cross-group generalization appears more inconsistent, and no clear lineage-based pattern consistently explains the best scores across all languages. Similarly, for languages such as ENG and CHI, best cross-lingual results are scattered across different training combinations, highlighting slightly weaker transfer across distant linguistic families.

These results, supported by findings in e.g. [10, 11], suggest that linguistic proximity may indeed facilitate direct transfer, and that investigating linguistic properties could lead to more robust SER.

4.4. Weaknesses and Limitations

While results prove that our hypothesis holds, we believe it is fair to point out several weaknesses of our approach, many of which have already been highlighted by several works such as [17, 36, 37]. Specifically, datasets are often Western-centric and there is a lack of coverage for many underrepresented cultures; in addition, with many of them not being freely available, this constitutes an important limitation to the range of possible choices. Moreover, many of them contain few and/or unbalanced samples [15]: while this may be sufficient

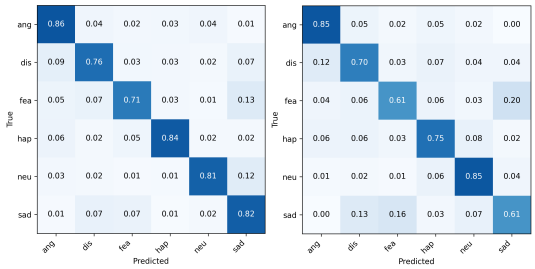

Fig. 3. Global, per-emotion Confusion Matrices for the Multilingual Neo-Latin model (left) and the Multilingual Others one (right).

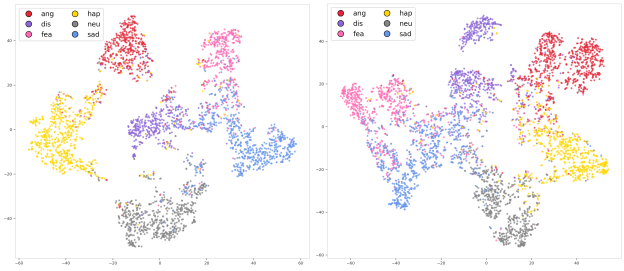

Fig. 4. Global, per-emotion t-SNE for the Multilingual Neo-Latin model (left) and the Multilingual Others one (right).

for generic convergence, also partially mitigated by data balancing strategies [17, 18], fully exploiting and understanding linguistic nuances in SER requires much larger, coherent, and linguistically-treated databases. Finally, as already mentioned, many corpora may implicitly inject biases due to differences in recording conditions, speaker demographics, word/sentence-based material, or annotation protocols: this indeed motivated our choice of losses and augmentations.

Overall, these limitations suggest that future progress in multilingual SER will strongly depend on the creation and sharing of larger and more diverse datasets.

5. CONCLUSION

In this work, we investigated cross-lingual generalization in Speech Emotion Recognition, explicitly focusing on linguistic lineage. By comparing Neo-Latin languages against a homogeneous set of unrelated languages, our contribution shows that models trained within the same language family achieve better transfer performance, whereas generalization across distant families is substantially weaker. While dataset issues may hamper direct generalization, we show how few-shot fine-tuning can help partially mitigate these gaps, especially within the same linguistic group. While an exhaustive analysis will require substantially larger and more uniform databases, results suggest that taking into account the linguistic lineage of different languages can improve generalization.

6. REFERENCES

[1] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[2] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP*. IEEE, 2022, pp. 6907–6911.

[3] S.-w. Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *ICASSP*. IEEE, 2019, pp. 5881–5885.

[4] M. Neumann et al., "Cross-lingual and multilingual speech emotion recognition on english and french," in *ICASSP*. IEEE, 2018, pp. 5769–5773.

[5] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *FIT*. IEEE, 2018, pp. 88–93.

[6] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[7] P. Van Rijn and P. Larrouy-Maestri, "Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody," *Nature Human Behaviour*, vol. 7, no. 3, pp. 386–396, 2023.

[8] S. Paulmann and A. K. Uskul, "Cross-cultural emotional prosody recognition: Evidence from chinese and british listeners," *Cognition & Emotion*, vol. 28, no. 2, pp. 230–244, 2014.

[9] X. Zhang, R. Mao, and E. Cambria, "Multilingual emotion recognition: Discovering the variations of lexical semantics between languages," in *IJCNN*. IEEE, 2024, pp. 1–9.

[10] S. T. A. Monisha and S. Sultana, "A deep learning approach toward analyzing the cross-lingual acoustic-phonetic similarities in multilingual speech emotion recognition," *Journal of Electrical and Computer Engineering*, vol. 2025, no. 1, pp. 4748790, 2025.

[11] Y. Lei, "Rlhf-powered multilingual audio understanding: A cross-cultural emotion analysis framework for international communication," *Journal of Sustainability, Policy, and Practice*, vol. 1, no. 4, pp. 66–79, 2025.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[13] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP*. IEEE, 2023, pp. 1–5.

[14] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech*, 2021, pp. 3400–3404.

[15] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," in *Interspeech*. ISCA, 2024, pp. 1580–1584.

[16] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Proc. Int. Conf. Advances in Neural Information Processing Systems*, vol. 33, 2020.

[17] S. Wang, J. Gunason, and D. Borth, "Learning emotional representations from imbalanced speech data for speech emotion recognition and emotional text-to-speech," in *Proc. Interspeech 2023*, 2023, pp. 351–355.

[18] J. Wu, Z. Wen, H. Huang, H. Su, F. Liu, H. Wang, Y. Ding, and Q. Wu, "A reweighting method for speech recognition with imbalanced data of mandarin and sub-dialects," *Service Oriented Computing and Applications*, vol. 18, no. 2, pp. 145–152, 2024.

[19] Y. Terraf and Y. Iraqi, "Lancet: Lightweight attention-enhanced network for robust speech emotion recognition," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025, pp. 371–375.

[20] T. Das, M. F. Islam, and N. Mamun, "Attention-based multi-level feature fusion for multilingual speech emotion recognition," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2025, pp. 1–6.

[21] P. Ekman, "Basic emotions," *Handbook of Cognition and Emotion*, p. 45, 2000.

[22] L. Kerkeni, C. Cleder, Y. Serrestou, and K. Raoof, "French emotional speech database - oréau," Dec. 2020.

[23] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proc. of the 9th ACM Multimedia Systems Conf.*, 2018, pp. 399–402.

[24] F. Catania, J. W. Wilke, and F. Garzotto, "Emozionalmente: A crowdsourced corpus of simulated emotional speech in italian," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[25] R. G. Bastos Germano, M. Pompeu Tcheou, F. da Rocha Henriques, and S. Pinto Gomes Junior, "emouerj: an emotional speech database in portuguese," September 2021.

[26] S. L. Castro and C. 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.

[27] T. Telembici, L. Muscar, L. Grama, and C. Rusu, "Emotion recognition audio database for service robots," in *ISETC*. IEEE, 2022.

[28] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The mexican emotional speech database (mesd): elaboration and assessment based on machine learning," in *EMBC*. IEEE, 2021, pp. 1644–1647.

[29] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs dimensional perception of italian emotional speech," in *Interspeech*, Hyderabad, India, 2018, pp. 3638–3642.

[30] M. A. Soumiaa, "Moroccan dialect emotion recognition dataset," 2024.

[31] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, pp. e0250173, 2021.

[32] B. Gong, N. Li, Q. Li, X. Yan, J. Chen, L. Li, X. Wu, and C. Wu, "The mandarin chinese auditory emotions stimulus database: A validated set of chinese pseudo-sentences," *Behavior Research Methods*, vol. 55, no. 3, pp. 1441–1459, 2023.

[33] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[34] I. Lubenets, N. Davidchuk, and A. Amentes, "Aniemore," `https://github.com/aniemore/Aniemore`, 2021.

[35] B. Maji, R. Guha, A. Routray, S. Nasreen, and D. Majumdar, "Investigation of layer-wise speech representations in self-supervised learning models: A cross-lingual study in detecting depression," in *Proc. Interspeech 2024*, 2024, pp. 3020–3024.

[36] F. A. Dal Rí, F. C. Ciardi, and N. Conci, "Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks," *IEEE Access*, vol. 11, pp. 116638–116649, 2023.

[37] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *ICASSP*. IEEE, 2018, pp. 5099–5103.