IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

Prototypical Networks for Speech Emotion Recognition in Spanish

Adria Mallol-Ragolta[1,3], *Anika A. Spiesberger*[1,3], *Antonio Barba Salvador*[4], *Björn W. Schuller*[1,2,3,5]

[1] CHI – Chair of Health Informatics, MRI, Technical University of Munich, Germany
[2] MDSI – Munich Data Science Institute, Germany
[3] MCML – Munich Center for Machine Learning, Germany
[4] Science, Computing and Technology Department, Universidad Europea de Madrid, Spain
[5] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

adria.mallol-ragolta@tum.de

Abstract

We explore the utilisation of prototypical networks in the Speech Emotion Recognition (SER) problem, creating prototypical representations of the targeted emotions in the embeddings space. We hypothesise this technique can help to improve the performance and robustness of the models, in comparison to standard classification-based approaches. We investigate two approaches to train the prototypes: one optimising a triplet loss, and the other minimising a prototypical loss. To assess our hypothesis, we exploit the EmoMatchSpanishDB Corpus; a novel dataset for SER in Spanish, which includes speech samples conveying the six basic emotions defined by Paul Ekman, in addition to the neutral state. We methodologically split the available samples into three speaker-independent train, development, and test partitions. The proposed splitting is not only balanced in terms of the speakers' gender, but also homogenised in terms of their recognition difficulty. We analyse the performance of our models with a gender perspective. The models exploit the eGeMAPS and the wav2vec 2.0 feature representations extracted from the speech samples. We choose the Unweighted Average Recall (UAR) as the evaluation metric to assess the models' performance. The chance level UAR for a seven-class classification problem is 14.3 %. The models optimising the prototypical loss obtain the highest UAR scores on the test set, 52.0 % and 52.7 %, with the eGeMAPS and the wav2vec 2.0 representations, respectively. Nevertheless, the best performances are obtained with a Support Vector Classifier (SVC) implementing a radial basis function kernel, with a UAR of 54.4 % and 56.9 % when exploiting the eGeMAPS and the wav2vec 2.0 representations, respectively.
Index Terms: Affective Computing, Paralinguistics Analysis, Speech Emotion Recognition, Prototypical Networks.

1. Introduction

The *Human-Computer Interaction* (HCI) field has experienced revolutionising changes in the last few years by means of the *smart* devices that are now indispensable companions in our daily lives, or the robotic systems that support our daily activities, both at home and at work. Powering these systems with natural interaction capabilities will have the leading role in the next HCI revolution. Towards this goal, affective computing technologies [1] will contribute to automatically understand and synthesise human emotions, so that machines can interact with their users in an empathetic and affective manner [2, 3].

This paper focuses on the problem of *Speech Emotion Recognition* (SER), exploiting the paralinguistic information embedded in the human voice [4, 5]. Although the vast majority of the SER literature tackles the task as a classification problem [6], our main hypothesis relies on the utilisation of prototypical networks [7] to create prototypical representations of the considered, discrete emotions. At inference time, the model predictions are determined by computing the distances between the test samples and the obtained prototypes. We hypothesise this approach can help to create more robust representations of the targeted emotions, ultimately improving the performance of the classification-based approaches, which we include in our experiments for benchmarking purposes. Additionally, we add to our study a gender perspective to assess how the speakers' biological sex impacts the performance of the SER models.

The prototypical networks operate in the embeddings space. We compare the performance of two different losses to train our models: a triplet loss [8], and a prototypical loss [7]. The triplet loss requires triplets of samples and aims at minimising the distance between the representations of the same class, and maximising the distance with the representation from a different class. This loss is theoretically sound for binary problems [8]; nonetheless, we investigate how it performs in the multiclass, SER problem. The prototypical loss seems a more suitable option in this case, as it takes into account the distances between the current samples and all the prototypes in its formulation.

We assess our hypotheses on the EmoMatchSpanishDB Corpus [9], a novel Spanish SER dataset featuring elicited emotions. Despite Spanish being the second most widely spoken language by number of native speakers globally, there is a lack of resources for SER in this language. The EmoMatchSpanishDB Corpus aimed at filling this gap. We split the samples into three speaker-independent train, development, and test sets. The resulting partitions are not only balanced in terms of the speakers' gender, but also homogenised in terms of their recognition difficulty.

The rest of the paper is laid out as follows. Section 2 highlights related works in the field. Section 3 introduces the dataset explored, and Section 4 describes the methodology followed. Section 5 compiles and analyses the results obtained, and Section 6 concludes the paper.

2. Related Works

The application of SER spans various fields, such as customer service (e. g., sentiment analysis in call centres [10]), HCI (e. g., emotion-aware assistants in cars [11]), and healthcare (e. g., mental health monitoring systems [12]). However, the challenges attributed to its implementation include the need for appropriate datasets, the selection of relevant features, and the implementation of classification methods [13].

Cultural and linguistic differences in emotional expression underscore the need for culture- and language-specific datasets. The authors in [14] show that people are better at recognising emotions from people of the same culture than from a different cultural group. This is supported by a meta-analysis on cross-cultural emotion recognition and in-group advantages, which

10.21437/IberSPEECH.2024-45

indicates that there are culture-specific signals [15]. Despite efforts to demonstrate the feasibility of cross-lingual and cross-cultural classification, studies have shown higher accuracies in within versus cross-cultural conditions [16] and in within versus cross-lingual conditions [17], demonstrating the necessity for language and culture-specific datasets.

According to [13], most SER datasets used today focus on English [18], Chinese [19], or German [20], while datasets for other languages are scarce. Despite Spanish being the second most spoken native language globally, existing datasets are limited. The Spanish Emotional Speech database [21] contains four emotions plus a neutral version of different words, sentences, and paragraphs each performed by the same professional male actor. RekEmozio [22] used recordings of 17 actors (7 basques, 10 castilians), and for the EmoFilm [23] database, emotional content was extracted from 43 (originally English) movies.

Gender disparities in emotional expression pose additional challenges for SER. Even though these differences are attributed to socialisation effects, rather than biological factors [24], they influence the performance of SER algorithms. In regard to SER, studies have shown that performance increased when separate models were used for females and males [25]. It is noteworthy that in most studies, no matter if they were gender-agnostic or gender-specific, SER models performed better for males [26].

3. The EmoMatchSpanishDB Corpus

The dataset [9] contains 2 005 speech recordings from 50 speakers (20 f, 30 m) reading out loud 12 sentences and conveying the Ekman's six basic emotions [27]: anger, disgust, fear, happiness, sadness, and surprise; in addition to the neutral state. The selected sentences, which do not contain semantic connotations to avoid potential emotional influences on the speakers, represent the phonemes of the Spanish language spoken in Spain, as defined by the Spanish Royal Academy, and preserve the appearance frequency of the phonemes according to [28].

3.1. Data Preparation

We pre-process the available samples by removing the unvoiced frames, as these do not contain paralinguistic information related to the speakers' emotional state. For this, we implement a *Root-Mean-Square* (RMS)-based *Voice Activity Detector* (VAD) with a frame length of 64 ms and a 50 % overlap. We scale the computed RMS using min-max normalisation, and we empirically set a threshold for the RMS of 0.1 to differentiate the voiced from the unvoiced frames. To preserve the naturalness and intelligibility of the processed speech samples, only when the VAD detects at least 10 consecutive unvoiced frames, we remove the corresponding portion of the original signal. We normalise the resulting signal, so its amplitude ranges $\in [-1, 1]$.

3.2. Data Partitioning

We distribute the pre-processed speech samples into three – train, development, and test – speaker-independent partitions, so that all the samples corresponding to the same speaker are contained in the same partition. In an attempt to balance the recognition difficulty among the partitions, we follow a nested *Leave-One-Speaker-Out Cross-Validation* (LOSO-CV) approach, splitting the samples in the inner loop into five speaker-independent folds, to assess the performance of a preliminary model on each speaker separately. To this end, we train a linear *Support Vector Classifier* (SVC), exploiting the functionals of the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [29] extracted with OPENSMILE [30]. We normalise the extracted features before training the model. This preliminary model targets the recognition of the seven aforementioned emotions.

Table 1: *Summary with the number of speech samples in the EmoMatchSpanishDB Corpus per partition, emotion, and speakers' sex – F(emale), and M(ale).*

Emotion	Train			Devel			Test			Σ
	F	M	Σ	F	M	Σ	F	M	Σ	
Surprise	75	95	170	27	33	60	27	40	67	297
Happiness	79	81	160	25	22	47	24	25	49	256
Neutral	107	169	276	36	60	96	42	58	100	472
Disgust	25	38	63	20	8	28	13	24	37	128
Sadness	78	98	176	16	30	46	29	23	52	274
Fear	81	70	151	32	22	54	26	31	57	262
Anger	83	108	191	30	24	54	33	38	71	316
Σ	528	659	1 187	186	199	385	194	239	433	2 005

The C parameter in the SVC is the hyperparameter to optimise in this preliminary modelling. At each iteration of the LOSO-CV routine, with the speech samples corresponding to one speaker reserved for testing the optimal model, each $C \in [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]$ is utilised to iteratively train and test an SVC model following a 5-fold cross-validation approach, in accordance to the number of folds in which we distribute the data in the inner loop. We average the *Unweighted Average Recall* (UAR) scores obtained after testing the corresponding model on each one of the five iteratively excluded folds to assign an overall performance to each C parameter. The C value that obtains the highest averaged UAR is defined as the optimal hyperparameter in the current LOSO-CV iteration. The optimal C is employed to train the optimal SVC which we finally test on the speech samples corresponding to the initially excluded speaker. With this preliminary modelling, analysing the results obtained for each individual speaker globally, we achieve an averaged UAR of 44.6 % with a standard deviation of 11.0 %. The chance level in terms of the UAR is 14.3 %, as we tackle the task as a 7-class classification problem.

The resulting individual UAR scores help us to distribute the speakers among the train, development, and test partitions. Following the data partitioning procedure presented in [31], we select and sort the female speakers in descending order in terms of the obtained UAR score. Then, we follow a Round-robin fashion and distribute 60 %, 20 %, and 20 % of the female speakers among the train, development, and test partitions, respectively. An analogous procedure is followed to distribute the male speakers. With this approach, we not only guarantee that the dataset partitions are balanced in terms of the speakers' gender, but also that the 'difficulty of recognition' is homogenised across the partitions. The statistics of the partitioned EmoMatchSpanishDB Corpus are summarised in Table 1. The metadata associated with the data partitioning is available upon request.

4. Methodology

This section presents the methodology. Section 4.1 describes the feature representations of the speech samples explored, Section 4.2 details the SER models implemented, and Section 4.3 defines the parameters utilised for training the models.

4.1. Feature Extraction

We investigate the performance of hand-crafted and deep-learnt representations extracted from the pre-processed speech samples

(cf. Section 3). For the former, we extract the functionals of the eGeMAPS feature set [29]. For the latter, we extract the wav2vec 2.0 self-supervised learning representations [32], and average the extracted representations over time to obtain a single representation of each instance. Specifically, we use the wav2vec 2.0 base model pre-trained and fine-tuned on 960 h of speech data from LibriSpeech [33]. As a result, 88- and 512-dimensional feature vectors are extracted from the aforementioned representations. We normalise the extracted features before their modelling.

4.2. Models Description

The models implemented can be grouped into two categories.

4.2.1. Prototypical-Based Models

These models operate in the embeddings space and aim at creating a prototypical representation [7] of each emotion, $\mathbf{c}_e$, which can be mathematically defined as:

$$\mathbf{c}_e = \frac{1}{\mid S_e \mid} \sum_{(\mathbf{x}_i, y_i) \in S_e} f_\phi\left(\mathbf{x}_i\right), \tag{1}$$

where $f_\phi\left(\cdot\right)$ represents a neural network-based encoder. The goal of the encoder is to learn an 88-dimensional representation from the extracted features in the embeddings space. The dimensionality of the embedded representations was determined empirically with preliminary experiments not reported in this work. The encoder implements a single linear layer, and the number of input neurons depends on the dimensionality of the features to model. At inference time, the closest prototype to the embeddings extracted from the unseen speech samples determines the emotion assigned to the latter. The distances in the embeddings space are computed utilising the Euclidean distance. We explore three different encoder models, which we proceed to describe.

Naïve Prototypical Model (Naïve Proto.). This model computes the prototypes by directly averaging the extracted features from the samples corresponding to each emotion. It does not implement the previously described encoder and, therefore, requires no training. We include it as a baseline in our experiments to assess the feasibility of generating prototypical representations with the proposed encoder.

Triplet Loss-Based Model (Triplet Loss). To train this model, we feed into Siamese encoders the feature representation of the current sample, the anchor X^a, accompanied by two additional, randomly selected representations from the training data: one corresponding to a different speaker than the anchor but conveying the same emotion, X^+, and the other also belonging to a different speaker than the anchor but conveying a different emotion, X^-. We then compute the Euclidean distance between the embedded representations of X^+ and X^-, and the anchor; i. e.

$$d^+ = \mid\mid f_\phi\left(X^+\right) - f_\phi\left(X^a\right) \mid\mid_2^2, \tag{2}$$

and

$$d^- = \mid\mid f_\phi\left(X^a\right) - f_\phi\left(X^-\right) \mid\mid_2^2, \tag{3}$$

respectively. The triplet loss [8] that the training of this model minimises is mathematically defined as:

$$\mathcal{L} = max\{0, d^+ + \alpha - d^-\}, \tag{4}$$

where α is a hyperparameter, which we empirically set to 1. At each epoch, the prototypes are updated after updating the encoder weights, and the triplets of samples are reassigned.

Prototypical Loss-Based Model (Proto. Loss). The loss [7] this model minimises can be mathematically defined as:

$$\mathcal{L}\left(\theta\right) = -\log\left(\frac{\exp\left(-d\left(f_\phi\left(\mathbf{x}\right), \mathbf{c}_k\right)\right)}{\sum_{k'} \exp\left(-d\left(f_\phi\left(\mathbf{x}\right), \mathbf{c}_{k'}\right)\right)}\right). \tag{5}$$

In contrast to the triplet loss, this loss aims at minimising the distance between the embeddings of the current sample and its corresponding prototype, and maximising the distance with the other prototypes. We use the Euclidean distance as the distance function $d\left(\cdot\right)$ in Equation (5). We initialise the prototypes with the initial encoder weights, and we update them at each epoch, after updating the encoder weights. At inference time, the prototype that minimises Equation (5) determines the emotion to assign to the embeddings of the current unseen sample.

4.2.2. Classification-Based Models

These models receive the features extracted from the speech samples as input and produce a label at the output, indicating the emotion into which the models classify the input samples. We incorporate them in our analysis to benchmark the performance of the proposed prototypical-based models. The classification-based models implemented are described below.

Neural Network-Based Model (NN). This model is composed of an encoder and a classification block. While the former is responsible for learning deep-learnt representations from the input features, the latter performs the actual classification. The architecture of the encoder block is the same as the encoder utilised in the prototypical-based models (cf. Section 4.2.1). The classification block implements a single linear layer with 88 and 7 input and output neurons, respectively. The training of this model minimises the categorical cross-entropy loss.

SVC-Based Models. We finally include in our analysis SVCs with a linear (**LinearSVC**) and a *Radial Basis Function* (RBF) (**RbfSVC**) kernel, as these are standard machine learning techniques. We train these models optimising the C parameter $\in \left[10^{-2}, 10^{-1}, 1, 10^1, 10^2\right]$ on the development partition.

4.3. Networks Training

All the neural network-based models are trained under the exact same conditions for a fair comparison. Regardless of the loss function to minimise, we use Adam as the optimiser with a fixed learning rate of 10^{-2}. We select UAR as the evaluation metric and define $\mathcal{L}_{\text{UAR}} = 1 - \text{UAR}$ as the validation error to monitor during training. Network parameters are updated in batches of 64 samples and trained during a maximum of 150 epochs. We implement an early-stopping mechanism to stop training when the validation error does not improve for 20 consecutive epochs. With this training routine, we aim to determine the optimal number of training epochs that minimises the risk of overfitting. We seed the pseudorandom number generator at the models' initialisation.

5. Experimental Results

We investigate the models in four different training scenarios: i) utilising all the samples available, regardless of the speakers' gender (Gender-Agnostic); ii) with the samples corresponding to the female speakers only (Trained on Females); iii) with the samples corresponding to the male speakers only (Trained on Males); and iv) utilising the speakers' gender as *a priori* information (Gender-Specific). The Gender-Specific models train a

Table 2: *Performance summary of the SER models trained exploiting the **eGeMAPS features**, including the CI on the test partition, in terms of the UAR (%). The chance level for a seven-class classification problem is 14.3 %.*

Models	Gender-Agnostic			Trained on Females			Trained on Males			Gender-Specific		
	Devel	Test	CI on Test	Devel	Test	CI on Test	Devel	Test	CI on Test	Devel	Test	CI on Test
Naïve Proto.	36.0	40.1	(35.4 – 44.9)	34.6	41.2	(36.8 – 45.7)	35.1	31.9	(27.3 – 36.8)	45.9	45.6	(40.7 – 50.5)
Triplet Loss	54.4	47.3	(42.3 – 52.4)	42.4	**43.8**	(39.3 – 48.3)	44.0	35.7	(31.7 – 40.2)	55.6	47.9	(42.7 – 52.9)
Proto. Loss	49.9	**52.0**	(47.2 – 57.0)	41.7	41.0	(36.6 – 45.5)	39.8	**38.7**	(34.1 – 43.1)	52.9	**51.0**	(45.9 – 55.6)
NN	48.9	50.1	(45.3 – 54.9)	43.1	43.1	(38.8 – 47.4)	41.0	40.1	(36.5 – 43.8)	52.9	50.7	(46.1 – 55.5)
LinearSVC	48.2	51.0	(46.4 – 55.8)	39.9	43.3	(39.2 – 47.3)	41.1	38.4	(34.1 – 43.0)	49.7	49.2	(44.8 – 54.0)
RbfSVC	50.0	51.7	(46.6 – 57.1)	38.5	42.3	(37.8 – 46.7)	37.9	**41.1**	(36.7 – 46.1)	50.4	**54.4**	(49.7 – 59.5)

Table 3: *Performance summary of the SER models trained exploiting the **wav2vec 2.0 features**, including the CI on the test partition, in terms of the UAR (%). The chance level for a seven-class classification problem is 14.3 %.*

Models	Gender-Agnostic			Trained on Females			Trained on Males			Gender-Specific		
	Devel	Test	CI on Test	Devel	Test	CI on Test	Devel	Test	CI on Test	Devel	Test	CI on Test
Naïve Proto.	45.5	43.1	(38.2 – 47.6)	37.5	37.6	(33.6 – 41.5)	42.1	37.8	(33.2 – 42.4)	51.3	41.9	(37.0 – 46.4)
Triplet Loss	55.8	51.6	(46.7 – 56.5)	44.5	43.6	(39.2 – 48.0)	45.6	39.2	(34.7 – 43.9)	53.9	**50.4**	(45.8 – 54.9)
Proto. Loss	58.5	**52.7**	(48.2 – 56.8)	43.4	**48.1**	(43.6 – 52.3)	49.6	**46.1**	(41.2 – 50.8)	56.4	50.3	(45.8 – 55.0)
NN	58.1	48.3	(44.5 – 52.1)	44.8	44.8	(40.7 – 49.0)	46.1	38.5	(34.0 – 43.0)	56.9	48.8	(44.1 – 53.4)
LinearSVC	52.7	53.8	(48.9 – 58.0)	38.0	45.0	(40.7 – 48.8)	42.1	43.6	(38.8 – 48.4)	48.9	52.6	(47.9 – 57.2)
RbfSVC	58.3	**56.9**	(52.0 – 61.5)	43.2	**50.7**	(46.0 – 55.1)	43.6	44.4	(39.7 – 49.4)	54.4	**54.6**	(49.8 – 59.1)

female-specific and a male-specific submodel with the samples corresponding only to the female and only to the male speakers, respectively; at inference time, the gender of the speaker corresponding to the current sample determines the submodel that analyses the sample. For the Gender-Agnostic model, the Trained on Females model, and the Trained on Males model, the samples corresponding to both female and male speakers are processed by the same model. The results obtained in terms of the UAR exploiting the eGeMAPS and the wav2vec 2.0 features are compiled in Tables 2 and 3, respectively. As outlined, the chance level performance for a seven-class classification problem in terms of the UAR is 14.3 %. We include the 95 % *Confidence Intervals* (CI) on the test set, computed by repeating 1 000 times the bootstrapping approach.

Table 2 compiles the model performances when exploiting the eGeMAPS features. We observe that the best performance on the test set was obtained with the Gender-Specific RbfSVC model, scoring a UAR of 54.4 %. Focusing the analysis on the prototypical-based models, those optimising the prototypical loss achieved the highest UAR scores in three out of the four cases investigated: the Gender-Agnostic, the Gender-Specific, and the Trained on Males models scored a UAR of 52.0 %, 51.0 %, and 38.7 % on the test set, respectively. The Trained on Females model optimising the triplet loss obtained the highest UAR, 43.8 %. The results also indicate that the prototpyical-based models outperform the classification-based models for the Gender-Agnostic and the Trained on Females models.

Table 3 depicts the model performances when exploiting the wav2vec 2.0 features. The best model scored a UAR of 56.9 % on the test set and corresponds to the Gender-Agnostic model implementing the RbfSVC classifier. We observe that only the Trained on Males model optimising the prototypical loss outperformed the classification-based models. The results support the suitability of the RbfSVC classifier when exploiting the wav2vec 2.0 features, as this classifier also obtained the highest UAR scores in the Gender-Specific, and the Trained on Females models, 54.6 % and 50.7 %, respectively. It is worth highlighting that the prototypical-based models outperformed the NN classifiers in all the cases investigated. While the prototypical loss achieved the highest UAR scores in the Gender-Agnostic, the Trained on Females, and the Trained on Male models with a UAR of 52.7 %, 48.1 %, and 46.1 %, respectively, the Gender-Specific model optimising the triplet loss scored the highest UAR, 50.4 %.

6. Conclusions

We investigated the performance of prototypical networks for SER in Spanish. The implemented prototypical networks optimised a triplet loss and a prototypical loss. We exploited the EmoMatchSpanishDB Corpus and extracted the eGeMAPS and the wav2vec 2.0 feature representations from the available speech samples. Our experiments supported the utilisation of the proposed prototypical-based models above the neural network-based classifiers. The Gender-Agnostic model optimising the prototypical loss obtained the highest UAR scores on the test set, 52.0 % and 52.7 %, with the eGeMAPS and the wav2vec 2.0 features representations, respectively. Furthermore, the prototypical loss-based models achieved a better performance than the NN model in six out of the eight cases investigated. Nevertheless, the best performances were scored by the RbfSVC model, with a UAR of 54.4 % and 56.9 % when exploiting the eGeMAPS and the wav2vec 2.0 features with the Gender-Specific and the Gender-Agnostic models, respectively.

Follow-up studies could investigate the suitability of the proposed prototypical networks exploiting SER datasets of different languages to analyse the cross-lingual capabilities of the proposed approach. Further works could focus on advancing the current prototypical network method – for instance, introducing attention mechanisms – towards improving the robustness and the generalisation capabilities of the resulting prototypes.

7. Acknowledgements

This project has received funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), and from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101060660 (SHIFT).

8. References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Transactions on Affective Computing*, 2022, 13 pages.

[3] A. Mallol-Ragolta and B. Schuller, "Coupling Sentiment and Arousal Analysis Towards an Affective Dialogue Manager," *IEEE Access*, vol. 12, pp. 20 654–20 662, February 2024.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of INTERSPEECH*. ISCA, 2009, pp. 312–315.

[5] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*. ISCA, 2013, pp. 148–152.

[6] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, vol. 154, 2023, article ID, 102974.

[7] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*. Long Beach, CA, USA: NeurIPS, 2017, 11 pages.

[8] A. Mallol-Ragolta, F. B. Pokorny, K. D. Bartl-Pokorny, A. Semertzidou, and B. Schuller, "Triplet Loss-Based Models for COVID-19 Detection from Vocal Sounds," in *Proceedings of the 44th Annual International Conference of the Engineering in Medicine & Biology Society*. Glasgow, UK: IEEE, 2022, pp. 998 – 1001.

[9] E. Garcia-Cuesta, A. Barba Salvador, and D. Gachet Pãez, "EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database," *Multimedia Tools and Applications*, 2023, 20 pages.

[10] M. Płaza, R. Kazała, Z. Koruba, M. Kozłowski, M. Lucińska, K. Sitek, and J. Spyrka, "Emotion recognition method for call/contact centre systems," *Applied Sciences*, vol. 12, no. 21, 2022.

[11] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, jul 2020.

[12] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Speech emotion recognition using supervised deep recurrent system for mental health monitoring," in *Proceedings of the 8th World Forum on Internet of Things*. IEEE, 2022, pp. 1–6.

[13] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, 2022.

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

[15] P. Laukka and H. A. Elfenbein, "Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis," *Emotion Review*, vol. 13, no. 1, pp. 3–11, 2021.

[16] P. Laukka, D. Neiberg, and H. A. Elfenbein, "Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations." *Emotion*, vol. 14, no. 3, p. 445, 2014.

[17] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 125–131.

[18] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops*. IEEE, 2006, pp. 8–8.

[19] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, pp. 913–924, 2017.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Proceedings of INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.

[21] J. M. Montero, J. M. Gutiérrez-Arriola, J. Colás, J. M. Guarasa, E. Enríquez, and J. M. Pardo, "Development of an emotional speech synthesiser in spanish." in *Proceedings of EUROSPEECH*, 1999, pp. 2099–2102.

[22] J. M. López, I. Cearreta, N. Garay, K. López de Ipiña, and A. Beristain, "Creación de una base de datos emocional bilingüe y multimodal," in *Proceeding of the 7th Spanish Human Computer Interaction Conference*, vol. 6, 2006, pp. 55–66.

[23] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Proceedings of INTERSPEECH*, 2018, pp. 3638–3642.

[24] L. R. Brody, "Gender and emotion: Beyond stereotypes," *Journal of Social Issues*, vol. 53, no. 2, pp. 369–393, 1997.

[25] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using hierarchical models," *Procedia Computer Science*, vol. 151, pp. 37–44, 2019.

[26] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, "The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning," *Sensors*, vol. 22, no. 7, 2022.

[27] P. Ekman, "Expression and the Nature of Emotion," *Approaches To Emotion*, pp. 319–343, 1984.

[28] I. A. Rodríguez, "Cálculo de frecuencias de aparición de fonemas y alófonos en español actual utilizando un transcriptor automático," *Loquens: Spanish Journal of speech sciences*, vol. 3, no. 1, 2016, article ID e029, 29 pages.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proceedings of the 18th International Conference on Multimedia*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[31] A. Mallol-Ragolta, N. Urbach, S. Liu, A. Batliner, and B. Schuller, "The MASCFLICHT Corpus: Face Mask Type and Coverage Area Recognition from Speech," in *Proceedings of INTERSPEECH*. Dublin, Ireland: ISCA, August 2023, pp. 2358–2362.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings of the 34th Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc., 2020, pp. 12 449–12 460.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. South Brisbane, QLD, Australia: IEEE, 2015, pp. 5206–5210.