An Incremental Selection Method for Semi-Supervised Speaker Adaptation in Speech Emotion Recognition

Carlos Castorena, Maximo Cobos, *Senior Member, IEEE*, and Francesc J. Ferri, *Senior Member, IEEE*

***Abstract*—Adapting Speech Emotion Recognition (SER) to new, previously unseen speakers, remains a significant challenge due to the variability in emotional expression across speakers and the scarcity of labeled data for adaptation. This study introduces a novel incremental adaptation framework designed to address these challenges by leveraging a modified k-means algorithm to iteratively select representative samples in a latent space. By progressively refining the model in this way, the method facilitates the alignment of a given source domain to a new speaker, enhancing generalization. Experiments conducted using data from diverse datasets, under both balanced and unbalanced conditions, demonstrate that the proposed approach outperforms random selection and labeling, achieving comparable or superior results to state-of-the-art, non-incremental methods. These findings underscore the potential of the proposed incremental strategy for improving speaker adaptation in SER tasks, particularly in data-limited scenarios.**

***Index Terms*—Clustering, domain adaptation, deep learning, speech emotion recognition (SER).**

I. Introduction

SPEECH Emotion Recognition (SER) presents a critical challenge in natural language processing and human-machine interaction, as it seeks to identify diverse emotional states solely from speech signals [1], [2]. While contemporary systems excel in tasks like speech transcription and translation [3], their effectiveness diminishes significantly when applied to SER. This performance gap arises from several intrinsic challenges [4], including the complexity of modeling and capturing the subtle nuances of emotional expressions in speech, the tendency of models to develop undesired semantic dependencies during training [1], and the inherent imprecision in emotion labeling [5]. The latter is particularly problematic, as emotion annotations often suffer from subjectivity and inconsistencies, reflecting the intricate and multifaceted nature of human emotional expression.

A significant challenge in SER lies in the lack of model robustness when encountering unseen speakers [6], [7], [8], [9], [10]. Even models trained on large sets of speakers often struggle to generalize effectively to new speakers due to inherent variations in both the acoustic and semantic aspects of emotional expression. This discrepancy, known as a domain mismatch [11], arises from differences between the feature distribution of the training speakers, referred to as the Source Domain (SD), and that of the unseen speakers, referred to as the Target Domain (TD). Addressing this domain shift is fundamental to achieve different levels of speaker-independence in SER.

Generally, the differences between domains in SER are addressed through various domain adaptation strategies. Two common approaches are adversarial multitask training and domain discrepancy reduction. Adversarial multitask training involves training the model on multiple tasks simultaneously to learn more robust representations that minimize the gap between the source and target domains [12]. Domain discrepancy reduction methods, on the other hand, aim to directly minimize the differences between the feature distributions of the source and target domains [13]. While both techniques show effectiveness, they typically assume access to a significant amount of labeled data from the TD.

In the context of managing data from a new domain, prototype-based or diversity-based selection methods, such as Protodash [14], offer a promising approach in scenarios with limited labeled data. These methods, widely applied in classification tasks [15], focus on selecting a small set of representative samples that capture the distribution of the TD. Although these methods have not been widely explored in the context of SER, they present a novel approach for adapting a given model to a new, shared domain in SER tasks. By prioritizing the most informative samples, these methods help to improve the model's understanding of the TD.

In this work, we address the challenge of adapting trained SER models to a new speaker with limited labeled data. While some SER speaker adaptation methods exist, they have primarily focused on fully unsupervised settings [16], [17], [18]. However, unsupervised approaches may lead to unreliable cluster structures due to the absence of labeled guidance. In contrast, introducing a small number of strategically selected labels can help anchor the adaptation process, leading to more accurate domain alignment and better emotional class separation. Our main contribution is a novel semi-supervised approach that

Received 13 June 2025; accepted 24 June 2025. Date of publication 30 June 2025; date of current version 29 July 2025. This work was supported in part by Spanish Research Agency (AEI), in part by European Regional Development Fund (ERDF) under Grant TED2021-131003B-C21 and Grant PID2022-137048OB-C41, in part by MCIN/AEI/10.13039/501100011033, in part by the 'EU Union NextGenerationEU/PRTR', in part by *Generalitat Valenciana* through the *Santiago Grisolía* Program under Grant GRISOLIAP/2021/060, CPI-21-232, in part by the EU ERDF and Comunitat Valenciana through Artemisa Computer Resources, and in part by Technical Support. The associate editor coordinating the review of this article and approving it for publication was Prof. Sri Rama Murty Kodukula. *(Corresponding author: Carlos Castorena.)*

The authors are with Department d'Informàtica, Universitat de València, 46010 Valencia, Spain (e-mail: carlos.castorena@uv.es; maximo.cobos@uv.es; francesc.ferri@uv.es).

Digital Object Identifier 10.1109/LSP.2025.3584714

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

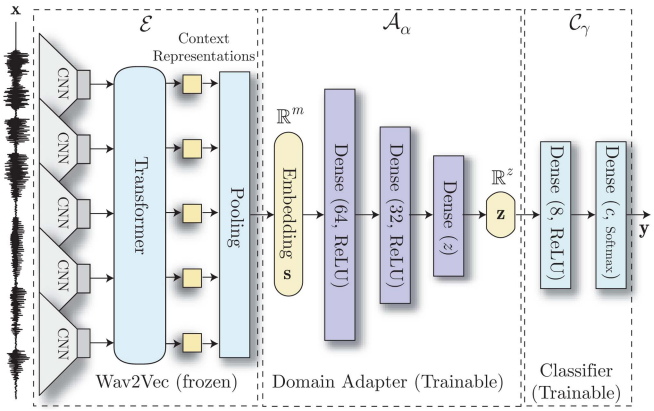


Fig. 1. Model architecture.

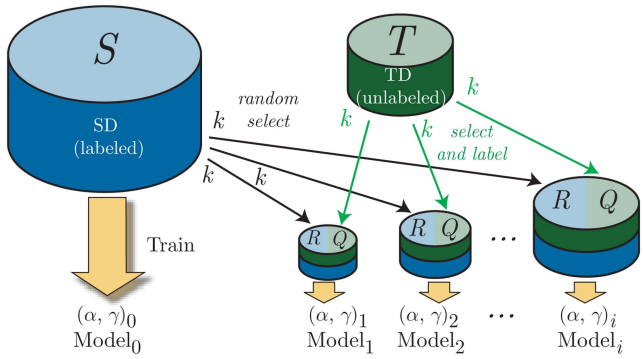


Fig. 2. Proposed method. Yellow arrows represent training steps. The initial model is fully trained on S. New adapted models result from incrementally selecting k new samples from both S and T at each step.

incrementally selects samples while disclosing only a few new labels to adapt the model. This incremental adaptation strategy helps align the source and target domains by creating a shared representation in the latent space, enhancing the model's robustness to speaker variability and minimizing the need for large labeled datasets.

II. Model Formulation

The architecture of the base model employed in this work is depicted in Fig. 1. The input speech signal, $\mathbf{x} \in \mathbb{R}^p$, is initially processed using Wav2Vec [19], a transformer-based feature extractor pre-trained on speech data. This processing step maps $\mathbf{x}$ into a time-independent embedding, $\mathbf{s} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^m$, using global average pooling, where $m = 1024$. Following the approach suggested by Wagner et al. [1], the representation $\mathbf{s}$ is further mapped into a prediction through a dense feed-forward network with a bottleneck, structured as an adaptor, $\mathcal{A}_\alpha$, and a classifier, $\mathcal{C}_\gamma$ to produce a one-hot encoded output, $\mathbf{y} = \mathcal{C}(\mathcal{A}(\mathbf{s})) \in \mathbb{R}^c_{[0,1]}$, where c is the number of classes or categories. The (learnable) parameters of this composite model are (α, γ) and will be omitted if they can be safely deduced from the context. The domain adaptation submodule, $\mathcal{A}$, aims to harmonize and adapt representations to a new speaker by obtaining meaningful latent representations $\mathbf{z} = \mathcal{A}(\mathbf{s}) \in \mathbb{R}^z$, where z is the latent space dimension.

We consider a composite loss which is defined on a given training set, $X \subseteq \mathbb{R}^m$, as

$$\mathcal{L}(X) = \mathcal{L}_{CCE}(\mathcal{C}(\mathcal{A}(X))) + \lambda \mathcal{L}_C(\mathcal{A}(X)),$$

where CCE refers to Categorical Cross Entropy and $\mathcal{L}_C$ is the modified center loss [20],

$$\mathcal{L}_C(Z) = \sum_{j=1}^{c} \sum_{\substack{\mathbf{z} \in Z \\ cl(\mathbf{z})=j}} ||\mathbf{z} - \boldsymbol{\mu}_j||^2 + \left(\sum_{j \neq j'} ||\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j'}||^2 \right)^{-1}.$$

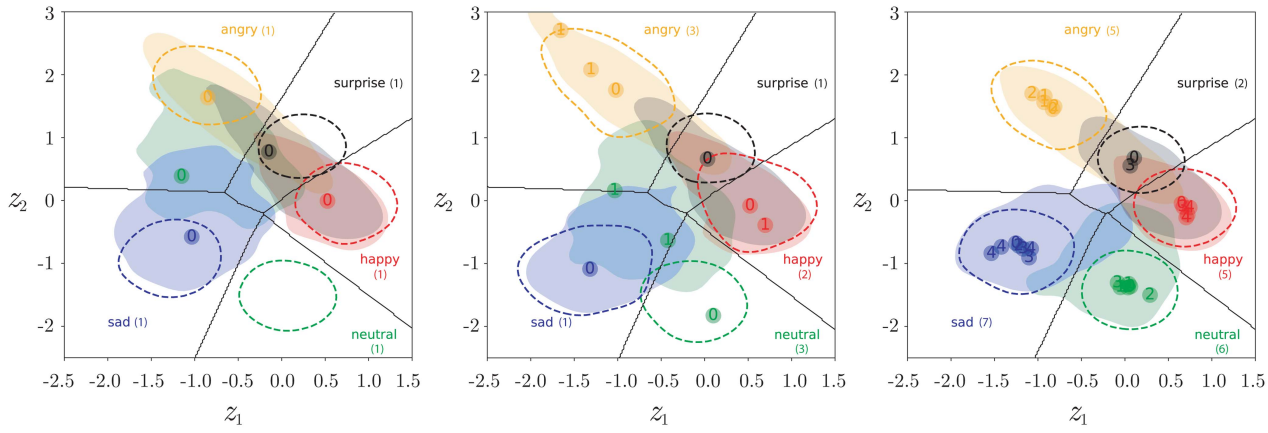
For each class j, $\boldsymbol{\mu}_j$ is the corresponding centroid (updated for each minibatch during training [21]) and $cl(\cdot)$ is the class label of a given element. The second term in the above equation accounts for interclass dispersion to enforce discrimination in the latent space as in previous works [22].

III. Proposed Method

We assume that our base model has been trained using a (large) SD dataset, S, and we are interested in adapting this model to a new TD corresponding to a new speaker using the dataset T. In contrast to S, the class labels in T are initially not made available to the algorithm but can be obtained on demand. The goal is to adapt the model to the TD using or even disclosing as few labels as possible. To this end, the subsets $R \subseteq S$ and $Q \subseteq T$ will be defined incrementally and used to retrain the initial model leading to progressive, more adapted models as illustrated in Fig. 2. As labels in S are available, elements in R will be obtained by stratified random sampling. The elements in Q will be selected by grouping the unlabeled elements in T and selecting the appropriate representatives.

As shown in Algorithm 1, in a given step, we have (labeled) subsets R and Q of the same size. The next k selected elements, B_k, from the remaining unlabeled TD samples, $T \backslash Q$ (with $\backslash$ denoting set subtraction), are obtained in the latent space using a modified $(\ell + k)$-means (Algorithm 2) in which ℓ cluster centers are given and remain fixed [24]. The new centers, B_k, represent the underlying distribution of unlabeled data constrained by previous selected samples in the current latent space. Once the (not necessarily balanced) labels for B_k are obtained and the subsets R and Q are updated, the model is retrained to update parameters (α, γ). It is important to note that, in contrast to [24], even though the ℓ previously labeled centers are the same, they are not strictly fixed in the latent space due to the retraining of the composite model (α, γ).

The proposed selection strategy, defined by Algorithm 2, is illustrated in Fig. 3, which depicts the distribution of source (dashed contours) and target (shaded areas) samples in the latent space for $z = 2$. The contour lines represent 90% of the data distribution, while the gray lines define the classification boundaries. Fig. 3(a) shows the latent space of the initial model and the selected key samples, labeled as "0", from the first iteration with $k = c = 5$. Notably, there are almost no TD points in the region corresponding to the neutral class from the SD. After the first iteration (Fig. 3(b)), the latent space evolves, and the source and target contours of the neutral class exhibit a small overlap. The previously selected samples ("0") move closer to their respective class centroids, and a new set of $k = 5$ samples is selected for labeling ("1"). Subsequent iterations progressively shift the target distributions toward



(a) Step 0 (b) Step 1 (c) Step 5

Fig. 3. SD (dashed lines) and TD (shaded areas) distributions across iterations for $z = 2$ and ESD speaker 0001. Different colors are used for each class. TD samples in Q are represented as circles where numbers indicate the step in which they were labeled. Per-class counts shown in parentheses.

Algorithm 1: Incremental Selection and Adaptation.

In: S, T //source and target datasets
$\mathcal{A}_\alpha, \mathcal{C}_\gamma$ //trainable mappings
k //elements to select per iteration
Out: $(\alpha, \gamma), Q$ //final model and selected set

$(\alpha, \gamma) \leftarrow \arg\min \mathcal{L}(S)$ //train initial model
$R, Q \leftarrow \emptyset$
foreach *iteration* **do**
  $R \leftarrow R \cup random(k, S \backslash R)$
  $B_k \leftarrow$ **select**$(k, \mathcal{A}_\alpha(Q), \mathcal{A}_\alpha(T \backslash Q))$
  $getLabels(B_k)$
  $Q \leftarrow Q \cup B_k$
  $(\alpha, \gamma) \leftarrow \arg\min \mathcal{L}(R \cup Q)$ //retrain

Algorithm 2: **select**(k, F, X) //modified fixed-centers kmeans.

In: k //number of new clusters/centers
F=$\{\mathbf{c}_1, .., \mathbf{c}_\ell\}$ //set of ℓ fixed, previous centers
X //unlabeled dataset
Out: $C = \{\mathbf{c}_{\ell+1}, .., \mathbf{c}_{\ell+k}\}$ //k added, new centers

$C \leftarrow k$ arbitrary[1] elements from X **repeat**
  Set clusters $\{X_j\}_{j=1}^{\ell+k}$ as

$$X_j \leftarrow \{\mathbf{x} : j = \arg\min_i ||\mathbf{x} - \mathbf{c}_i||\}$$

  Recompute k centers, $\{\mathbf{c}_j\}_{j=\ell+1}^{\ell+k}$, in C only as

$$\mathbf{c}_j \leftarrow \frac{1}{|X_j|} \sum_{\mathbf{x} \in X_j} \mathbf{x}$$

until C *converges*
Replace each $\mathbf{c} \in C$ by its nearest neighbor in X

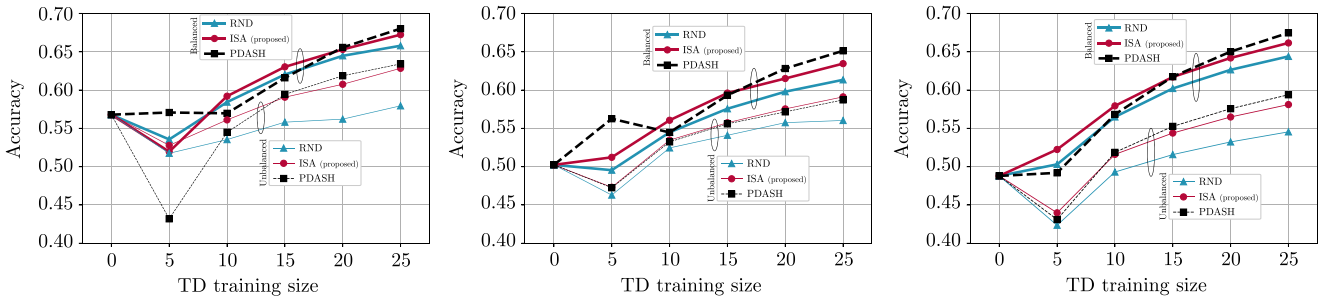
the SD regions, increasing their overlap. The final state after five iterations is shown in Fig. 3(c), where all selected target prototypes are displayed. Note that source and target distributions become notably aligned, particularly for the sad class.

[1] Centers can be randomly initialized or the well-known kmeans++ initialization [23] can be used.

IV. Experiments and Results

Several datasets widely employed in SER research have been considered for their diversity in characteristics such as language, number of samples, represented emotions, and participants. Among these are MSP-Podcast [25], with over 50,000 English-labeled samples based on valence, arousal, and dominance; IEMOCAP [26], comprising 10,039 samples of scripted and improvised dialogues; EMO-DB [27], containing 800 recordings in German expressing seven emotions; MESD [28], a Spanish dataset with 864 samples and six emotions; RAVDESS [29], featuring 7,356 recordings from 24 native English actors; and ESD [30], a multilingual dataset with 7,000 dialogues in English and Mandarin expressing five core emotions. Datasets MSP-Podcast and IEMOCAP are the ones used to train and validate the pretrained Wav2Vec model [1]. Consequently, we have used EMO-DB, MESD and RAVDESS to compose our SD dataset (that correspond to 37 speakers) and trained our base model using 70% of the data from each speaker during 200 epochs using Adam optimizer and no early stopping.

The TD in our experiments will be each one of the 20 speakers in ESD, separately. These included 10 who communicated in Mandarin and 10 in English, offering a multilingual setting with a diverse range of speakers to evaluate the proposed methodology. To ensure uniformity in the analysis, we homogenized the emotional labels across all datasets to $c = 5$ classes: Angry, Happy, Neutral, Sad, and Surprise. The data for each speaker-specific experiment was divided into training (70%) and testing (30%) sets in the same way as with the SD data. Both source and target training sets are given to our proposed selection strategy that will run for 5 iterations to simulate progressive data acquisition, reflecting real-world scenarios where data becomes available gradually. Model retraining at each iteration is done as in the base model but only for 20 epochs. The latent space dimension was fixed as $z = c - 1 = 4$ as in linear discriminant analysis [31]. Moreover, increasing this dimension did not lead to significant changes according to preliminary experiments.



(a) Speaker 0005 (Mandarin) (b) Speaker 0011 (English) (c) Overall (Avg. 20 speakers)

Fig. 4. Average accuracy at each iteration for balanced and unbalanced configurations, shown for two representative speaker adaptation experiments (a–b) and average across speaker-specific models on all ESD subjects (c).

We compared our proposed incremental selection and adaptation (*ISA*) approach against a basic random strategy (*RND*), where key samples from the unlabeled target set are selected at random. In order to assess the behavior of our incremental strategy at each iteration, we compared it also to an optimistic bound as follows. The non-incremental and state-of-the-art unsupervised selection algorithm Protodash [14] (*PDASH*), has been applied in each new latent space discovered by our ISA algorithm at each iteration to obtain an optimal subset of $\ell + k$ key samples. While this results in quadratic growth in the number of TD labels used, it provides a useful performance bound, particularly at the final iterations.

Two different experimental configurations of the SD were considered. In the first one (B), the emotional classes were perfectly balanced, ensuring an equal number of samples for each label. In the second configuration (U), the neutral class was predominant, with a 10:1 ratio compared to other classes. This strong imbalance simulates real-world conditions, where neutral emotions are far more common in practice.

Since all the alternatives are non deterministic (even when using PDASH which is deterministic) due to the use of random selections from the SD, each experiment for every speaker and configuration was repeated 100 times. Fig. 4(a) and (b) shows averaged performance results for two speakers (with different languages) from the ones considered as TD. Similar behavior was observed for the remaining speakers: the accuracy of the base model is relatively low and varies across speakers, with initial iteration results displaying noticeable fluctuations. From this point onward, all methods exhibit consistent improvements; however, the proposed approach clearly and significantly outperforms random selection.[2] Conversely, the bound achieved by Protodash becomes evident only in the final iterations, when the number of prototypes is sufficiently large, and the disclosed target labels are far greater than those required by incremental methods. Moreover, it is crucial to note that Protodash's selection process benefits from the progressively improved latent space established by the proposed method. This bound reflects an idealized and unfair scenario, combining a significantly higher reliance on labeled data with latent space enhancements not independently achievable by Protodash. Despite these advantages, the proposed method demonstrates superior performance across intermediate iterations, where realistic data limitations apply, underscoring its practicality and effectiveness in real-world scenarios.

[2] Significance was determined using rank-based Mann-Whitney tests applied to the non-paired results from 100 repetitions of each method.

TABLE I
RESULTS AT ITERATIONS 3 AND 5. # IS THE NUMBER OF TD LABELS USED. SYMBOLS $\gtrless$ INDICATE STATISTICAL SIGNIFICANT DIFFERENCES.

		init	RND	**ISA**	PDASH	RND	**ISA**	PDASH
	#	*0*	*15*	*15*	*30*	*25*	*25*	*75*
TD	(U)	0.488	0.516 $<$	0.544 $=$	0.553	0.545 $<$	0.581 $<$	0.594
	(B)		0.602 $<$	0.617 $=$	0.617	0.644 $<$	0.662 $<$	0.675
SD	(U)	0.841	0.689 $=$	0.692 $=$	0.694	0.733 $=$	0.733 $=$	0.733
	(B)		0.717 $<$	0.730 $>$	0.715	0.744 $<$	0.754 $>$	0.744

The overall average accuracies for all TD speakers are presented in Fig. 4(c) and summarized in Table I. In this case, we used a Wilcoxon test on the paired results per speaker to compare our proposal against both RND and PDASH. Similar trends were observed for both balanced and unbalanced configurations, with more pronounced differences between methods and lower absolute accuracies in the latter case. As seen in the last two rows of Table I, improvement on TD comes at the price of decreasing SD performance. According to the overall results and statistical evidence, the proposed method significantly outperforms random selection at a minimal cost. Furthermore, its accuracy results are equal to or better than the quasi-optimal PDASH bound in most iterations, except for the final iterations (and the second). This suggests that the proposal offers an effective trade-off between high performance and a minimal number of target labels.

V. CONCLUSION

In this work, we presented an incremental and adaptive approach for speaker adaptation in speech emotion recognition (SER), addressing the challenges posed by variability in emotional expression and limited labeled data. By leveraging a modified k-means algorithm, the proposed method incrementally selects representative samples to align source and target domains effectively. Extensive experiments demonstrated that the method consistently outperforms random selection, achieving comparable or superior results to an optimistic Protodash-based bound while using fewer target labels. In both balanced and unbalanced scenarios, the method achieved competitive performance, with pronounced differences between methods in the latter, more realistic case. These results suggest that the proposed approach offers an efficient trade-off between computational simplicity, data efficiency and performance.

REFERENCES

[1] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.

[2] F. Zhu-Zhou, R. Gil-Pita, J. García-Gómez, and M. Rosa-Zurera, "Robust multi-scenario speech-based emotion recognition system," *Sensors*, vol. 22, no. 6, 2022, Art. no. 2343.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, vol. 202, 2023, pp. 28492–28518.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[5] M. A. Islam, M. S. H. Mukta, P. Olivier, and M. M. Rahman, "Comprehensive guidelines for emotion annotation," in *Proc. 22nd ACM Int. Conf. Intell. Virtual Agents*, New York, NY, USA, 2022, pp. 1–8.

[6] C. Lu, Y. Zong, H. Lian, Y. Zhao, B. W. Schuller, and W. Zheng, "Improving speaker-independent speech emotion recognition using dynamic joint distribution adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 11696–11700.

[7] C. Lu et al., "Hierarchical distribution adaptation for unsupervised cross-corpus speech emotion recognition," in *Proc. Interspeech*, 2024, pp. 3739–3743.

[8] T. Rajapakshe, R. Rana, S. Khalifa, and B. W. Schuller, "Domain adapting deep reinforcement learning for real-world speech emotion recognition," *IEEE Access*, vol. 12, 2024, pp. 193101–193114.

[9] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1190–1194, 2021.

[10] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[11] E. Singer and D. A. Reynolds, "Domain mismatch compensation for speaker recognition using a library of whiteners," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2000–2003, Nov. 2015.

[12] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[13] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang, and B. W. Schuller, "Domain invariant feature learning for speaker-independent speech emotion recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2217–2230, 2022.

[14] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *Proc. IEEE Int. Conf. Data Mining*, vol. 2019, Nov. 2019, pp. 260–269.

[15] Y. Arslan, K. Allix, C. Lefebvre, A. Boytsov, T. F. Bissyandé, and J. Klein, "Exploiting prototypical explanations for undersampling imbalanced datasets," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl.*, 2022, pp. 1449–1454.

[16] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 5117–5120.

[17] J.-B. Kim and J.-S. Park, "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 126–134, 2016.

[18] M. Tran, Y. Yin, and M. Soleymani, "Personalized adaptation with pre-trained speech encoders for continuous emotion recognition," in *Proc. Interspeech*, 2023, pp. 636–640.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "WAV2VEC 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12449–12460, 2020.

[20] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A comprehensive study on center loss for deep face recognition," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 668–683, 2019, doi: 10.1007/s11263-018-01142-4.

[22] C. Qi and F. Su, "Contrastive-center loss for deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 2851–2855.

[23] D. Arthur and S. Vassilvitskii, "K-means: The advantages of careful seeding," in *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, vol. 7, Jun. 2007, pp. 1027–1035.

[24] M. Ay, L. Özbakir, S. Kulluk, B. Gülmez, G. Öztürk, and S. Özer, "FC-Kmeans: Fixed-centered K-means algorithm," *Expert Syst. Appl.*, vol. 211, 2023, Art. no. 118656.

[25] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.

[26] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.

[27] F. Burkhardt et al., "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[28] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody," *Data*, vol. 6, 2021, Art. no. 130.

[29] S. R. Livingstone and F. 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, May 2018, Art. no. e0196391.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. Proc.*, vol. 2021, Jun. 2021, pp. 920–924.

[31] P. Shen, X. Lu, L. Liu, and H. Kawai, "Local fisher discriminant analysis for spoken language identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5825–5829.